Analiza danych ilościowych dla politologów

Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP

Daniel Mider

Aneta Marcinkowska

Analiza danych ilościowych dla politologów

Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP

Wydanie drugie rozszerzone i poprawione

Warszawa 2013

Recenzent wydania drugiego:
prof. dr hab. Włodzimierz Okrasa

Recenzenci wydania pierwszego:
prof. dr hab. Wojciech Modzelewski
dr hab. Andrzej Wierzbicki

Redaktor:
Katarzyna Jaworska

Redaktor statystyczny:
Justyna Kaszek

Redaktor statystyczny:
Joanna Lewczuk

Skład i łamanie:
Mateusz Dworakowski
Skład okładki:
Małgorzata Kieliszek

Projekt okładki:
Łukasz Andrzejewski
Maciej Zaręba
z firmy Mantiz (www.mantiz.pl)

Daniel Mider - autor rozdziałów: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, Aneksu 1, Aneksu 2, Aneksu 4, Wstępu oraz współautor rozdziału 18.
Aneta Marcinkowska - autorka rozdziałów: 15, 16, Aneksu 3, współautorka rozdziału 18.

Wydanie drugie rozszerzone i poprawione

Warszawa 2013

Wskazówki dla bibliotekarzy:
1/ statystyka, 2/ politologia

ISBN: 978-83-62989-15-7

Druk: ACAD, Józefów, tel. (22) 789-4-789

W publikacji wykorzystano dane z *Polskiego Generalnego Studium Wyborczego 2007*, pod kierownictwem Radosława Markowskiego, afiliowane przy Instytucie Studiów Politycznych PAN, dofinansowane przez tę instytucję, oraz przez: Ministerstwo Nauki i Szkolnictwa Wyższego, Wissenschaftszentrum Berlin fur Sozialforschung (WZB), Polską Konfederację Pracodawców Prywatnych Lewiatan, Fundację Batorego, Instytut Filozofii i Socjologii PAN oraz instytucję badawczą realizującą sondaż - PBS DGA, w: http://www.ads.org.pl/opis-szczeg.php?id=72, dostęp: lipiec 2012.

Podziękowania

Książka powstawała w ramach Pracowni Metodologii Badań Politologicznych usytuowanej w Zakładzie Socjologii i Psychologii Polityki INP UW kierowanym przez **prof. dr hab. Jana Garlickiego**. Pomysł napisania pierwszego na polskim rynku wydawniczym podręcznika analizy ilościowej skierowanego do politologów powstał w toku zajęć *Statystyka i demografia* prowadzonych wspólnie z **dr hab. Andrzejem Wierzbickim** w roku akademickim 2011/2012 w Instytucie Nauk Politycznych Uniwersytetu Warszawskiego.

Żadne dzieło nie jest wytworem wyłącznie jego autorów. Chcielibyśmy gorąco podziękować wszystkim tym, którzy przyczynili się do powstania zarówno pierwszego, jak i drugiego wydania niniejszej publikacji. W pierwszej kolejności wyrazy wdzięczności należą się panom **Łukaszowi Andrzejewskiemu** i **Maciejowi Zarębie** z firmy Mantiz (www.mantiz.pl), którzy zadbali o graficzną oprawę książki. Niemniej serdeczne podziękowania składamy młodym i obiecującym (a już doświadczonym i cenionym w branży badawczej) badaczkom rynku i opinii specjalizującym się w badaniach ilościowych - paniom **Justynie Kaszek** oraz **Joannie Lewczuk**, które podjęły się zadania redaktorek statystycznych oraz pani **Katarzynie Jaworskiej**, która przyjęła na siebie trud redaktorki językowej. Podziękowania należą się również panu **Mateuszowi Dworakowskiemu,** który opanowawszy tajniki DTP dokonał łamania i składu niniejszej publikacji. Jego pomoc techniczna okazala się nieoceniona, a cierpliwość - niezmierzona. Dziękujemy również paniom **Jolancie Mider** i **Elżbiecie Mider**, **Agnieszce Maksimowicz, Patrycji Skierskiej, Marcie Świątkiewicz** oraz **Małgorzacie Kieliszek** za pomoc i wsparcie. Świadczona bezinteresownie pomoc wymienionych wyżej osób jest dla nas tym bardziej cenna, że rozumiały one szlachetne idee dydaktyczne przyświecające powstaniu tej publikacji.

Szczególne wyrazy wdzięczności należą się recenzentowi drugiego wydania publikacji - **prof. dr hab. Włodzimierzowi Okrasie**, którego wnikliwe uwagi i cenne wskazówki pozwoliły autorom uporządkować używaną terminologię i strukturę pracy. Dziękujemy serdecznie także recenzentom pierwszego wydania - **prof. dr hab. Wojciechowi Modzelewskiemu** i **dr hab. Andrzejowi Wierzbickiemu**.

Serdecznie dziękujemy również **dr hab. Wojciechowi Jakubowskimu**, prodziekanowi do spraw badań naukowych i współpracy z zagranicą - bez jego życzliwej pomocy niemożliwe byłoby zorganizowanie *Letnich warsztatów analizy danych ilościowych dla politologów 2012* dla studentów Instytutu Nauk Politycznych UW. Warsztaty te stały się swoistym poligonem dydaktycznym umożliwiającym przetestowanie przystępności materiału zawartego w niniejszej publikacji.

Niezależnie od wysiłku i wkładu wymienionych wyżej osób, które podjęły się trudu uczynienia tej książki lepszą, a którym jesteśmy głęboko wdzięczni, za wszystkie błędy odpowiadają sami autorzy.

Spis rozdziałów

Spis treści

Przedmowa do wydania drugiego

Dziwić może opublikowanie drugiego wydania w nieco ponad pół roku po ukazaniu się wydania pierwszego. Przyczyną stał się fakt błyskawicznego wyczerpania nakładu, co było dla autora tyleż miłe, co nieoczekiwane. Pojawił się jednak następujący dylemat: wznowić pierwsze wydanie bez zmian, czy rozpatrzeć postulaty i życzenia Czytelników oraz wziąć pod uwagę fakt, że wiele ważnych, statystycznych funkcji programu GNU PSPP pozostało bez omówienia, a także czy uwzględnić doświadczenia nabyte podczas prowadzenia *Letnich warsztatów analizy danych ilościowych dla politologów 2012* dla studentów Instytutu Nauk Politycznych UW. Wybór padł na realizację tego drugiego, bardziej pracochłonnego, lecz obiecującego większą satysfakcję scenariusza działań. Do współpracy nad drugą edycją książki dołączyła mgr Aneta Marcinkowska, doktorantka INP UW, współorganizatorka i współwykładowczyni rzeczonych warsztatów.

W efekcie podjętych prac autorzy przekazują Czytelnikowi materiał bogatszy w porównaniu z pierwszym wydaniem - dodano sto kilkadziesiąt stron (sześć nowych rozdziałów). Wprowadzone treści to wybrane średniozaawansowane metody analizy danych ilościowych z działu statystyki indukcyjnej. Modyfikacjom poddano również część stanowiącą pierwsze wydanie publikacji.

Wstęp

Celem niniejszej publikacji jest wprowadzenie Czytelnika nieposiadającego matematycznego przygotowania w podstawy analizy danych ilościowych - zarówno w jej podstawy teoretyczne, jak też praktyczne, w taki sposób, aby zainteresowany potrafił zastosować samodzielnie i ze zrozumieniem narzędzia analityczne służące do poznania zjawisk i procesów politycznych i społecznych. Książkę dedykujemy studentom nauk politycznych, jednak skorzystać z niej mogą adepci innych kierunków nauk społecznych oraz wszyscy ci, którzy zainteresowani są samodzielną analizą danych ilościowych.

Dokonanie pełnowartościowej analizy danych ilościowych wymaga przyswojenia wiedzy i umiejętności z zakresu statystyki, matematyki i informatyki; analiza danych ilościowych leży bowiem na przecięciu tych trzech dyscyplin. Po pierwsze, konieczne jest opanowanie teorii statystyki, a konkretnie wiedzy, w jaki sposób za pomocą liczb należy opisywać badane zbiorowości i zjawiska oraz, w jaki sposób i na jakich zasadach należy wyciągać wnioski o własnościach badanych grup wówczas, gdy dysponujemy informacjami o ich podgrupach nazywanych próbami. Po wtóre, niezbędne jest opanowanie matematycznych podstaw, na jakich opierają się miary statystyczne, co pozwala na ich zrozumienie, interpretowanie i adekwatne stosowanie. Po trzecie, przyswojenie umiejętności obsługi programów komputerowych służących do wykonywania analiz umożliwia zastosowanie wiedzy statystycznej i matematycznej w praktyce[1].

Niniejsza praca stawia sobie ambitny cel przełamania stereotypu statystyki jako przedmiotu odległego od praktyki, trudnego lub nieprzekładalnego na wnioski o otaczającej rzeczywistości. Z dydaktycznego punktu widzenia naukę analizy danych ilościowych studentom nauk społecznych znacznie utrudnia fakt, że jest ona uznawana za część matematyki. Jak pisze Hubert M. Blalock - na którego podręczniku

[1] Taki trójdyscyplinowy model kształcenia w zakresie analiz ilościowych stał się standardem w szkolnictwie zachodnim. Jego wprowadzenie na grunt polski proponuje prof. Walenty Ostasiewicz: W. Ostasiewicz, *Refleksje o statystyce wczoraj, dziś i jutro, Statystyka wczoraj, dziś i jutro. I Ogólnopolski Zjazd Statystyków z okazji 95-lecia Polskiego Towarzystwa Statystycznego i 90-lecia Głównego Urzędu Statystycznego*, „Biblioteka Wiadomości Statystycznych" 2008, t. 56, Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne, s. 15.

wychowały się pokolenia socjologów - jednym z najtrudniejszych zadań wykładowcy jest zachęcenie studenta do przezwyciężenia obaw przed matematyką i do zastosowania statystyki w dziedzinie jego badań[2]. Lęk przed liczbowym poznawaniem świata jest równie głęboko zakorzeniony, co irracjonalny. Już Józef Flawiusz twierdził, że nie kto inny, a właśnie Kain „wynalazłszy miary i wagi, zmienił ową niewinną i szlachetną prostotę, w jakiej żyli ludzie, (...) w życie pełne oszustwa". Z kolei w Czechach pod koniec XVIII wieku rozpowszechnione było przekonanie, że dziecko w wieku poniżej sześciu lat przestaje rosnąć, karłowacieje i staje się złowieszczym „mierzynem", jeśli ktoś zmierzy płótno przeznaczone na jego odzienie. Mieszkańcy XIX-wiecznej zachodniej Bułgarii wierzyli, że wprowadzanie metryk urodzenia, a tym samym próba liczenia dzieci, obraża Boga i zwiększa śmiertelność niemowląt. Twierdzono, że grzechem jest kontrolować Pana Boga. Jakikolwiek pomiar mógł się również negatywnie odbić na zdrowiu dorosłych: mieszkańcy Kujaw, w tym samym okresie, prosili farmaceutów, by lekarstwa do flaszek nalewać na oko, a nie na miarę, bowiem „jedynie niemierzone, szczodrą dłonią i ze szczodrego serca udzielone może wrócić choremu zdrowie". W rosyjskiej guberni włodzimierskiej w połowie XIX wieku chłopi z oporem reagowali na próby liczenia plonów używając przy tym uzasadnienia religijnego: „co Bóg da, to się i bez liczenia w sąsiekach znajdzie, a nam się sprawdzać wyroków Opatrzności nie godzi. (...) Grzeszy, kto oblicza zebrany z pola plon. Co nam potem?"[3].

W niniejszej publikacji zostaje podjęta próba przełamania (lub co najmniej naruszenia) obaw współczesnego adepta nauk politycznych wobec posługiwania się liczbami, następuje próba uleczenia „alergii na liczby". W tym kontekście należy przede wszystkim wskazać, że analiza danych ilościowych, a przede wszystkim jej najważniejsza część - statystyka - jest w swojej genezie i rozwoju bardzo ściśle związana z centralnym przedmiotem zainteresowań politologa - państwem. Statystyka powstała i przez większość czasu swego istnienia rozwijała się poza matematyką. Wyłoniła się z potrzeby organizacji życia we wspólnotach. Konstruowano i postrzegano ją jako ważny, a nawet nieodzowny filar sprawowania władzy; przez wieki była ona nierozdzielna z instytucjami państwa. Przekonuje o tym także źródłosłów tego pojęcia - „statystyka" pochodzi od łacińskiego *status,* co w sensie nabytym w średniowiecznym języku łacińskim oznaczało państwo lub polityczny stan rzeczy. Statystykę umieszczają współcześnie z powrotem na właściwym miejscu - to jest czynią zeń niezbędne narzędzie politologa - George H. Gallup i Saul F. Rae w wydanej w latach czterdziestych XX wieku wpływowej książce *The Pulse of Democracy*[4], a z rodzimych uczonych - Mirosław Szreder. Według wymienionych autorów statystyka umożliwia, by rządy demokratyczne sprawowane były „dla społeczeństwa i ze społeczeństwem". Jednorazowy akt wyborczy jest niewystarczający dla spełnienia tej formuły. Jak wskazuje M. Szreder - ważną, bo obecnie dobrze funkcjonującą (w przeciwieństwie do planowanych procesów konsultacji społecznych czy internetowej partycypacji politycznej) instytucją demokratyczną są sondaże. Służą one jako platforma komunikacji pomiędzy rządem a społeczeństwem, co więcej opierają się na mechanizmie naukowym, a nie mechanizmie w pewnym sensie ideologicznym i nieracjonalnym, bo zniekształcającym w istocie wolę społeczeństwa - jak... wybory. Ponadto są one pokojową metodą komunikowania i artykulacji interesów, w przeciwieństwie do takich form partycypacji politycznej jak demonstracje, strajki czy uczestnictwo w polityce z użyciem przemocy. W tym sensie analiza danych ilościowych stanowi swoisty „puls demokracji", który mogą i powinni badać wszyscy zainteresowani. Ważkimi przykładami pomiaru „pulsu demokracji" są badania

[2] H.M. Blalock, *Statystyka dla socjologów*, Państwowe Wydawnictwo Naukowe, Warszawa 1977, s. 9.

[3] Te zabawne, a jednocześnie jakże pouczające przykłady przytacza Witold Kula: W. Kula, *Miary i ludzie*, Wydawnictwo „Książka i Wiedza", Warszawa 2002, s. 19-21.

[4] G. Gallup, S. Rae, *The Pulse of Democracy,* Simon & Schuster, Nowy Jork 1940.

sondażowe i analiza wyników badań reakcji społeczeństw państw zachodnioeuropejskich na planowaną zbrojną interwencję w Iraku oraz reakcji społeczności międzynarodowej na atak terrorystyczny w Nowym Jorku 11 września 2001 roku[5].

W prezentowanej publikacji starano się przełamać negatywny stereotyp analizy danych ilościowych jako działania niepraktycznego i nieprzystępnego. Z całą mocą należy odrzucić oskarżenie o braku możliwości zastosowania jej w praktyce - analiza danych ilościowych jest narzędziem umożliwiającym wygenerowanie wartościowej, zrozumiałej i przekładalnej na praktyczne wnioski wiedzy o zjawiskach i procesach politycznych. Więcej nawet - jest ona narzędziem niezbędnym dla pełnowartościowego i pogłębionego poznawania procesów i zjawisk politycznych oraz społecznych. We współczesnej cywilizacji informacyjnej mamy bowiem do czynienia z zalewem danych. W tym właśnie kontekście Stanisław Lem - polski pisarz, filozof i futurolog - pisał o „bombie megabitowej". Wskazywał on, że ilość znajdujących się w obiegu informacji wielokrotnie przekroczyła pojemność umysłową jednostki ludzkiej[6]. Wtóruje S. Lemowi David Shenk, który wprowadził do publicystycznego i naukowego obiegu pojęcie „smogu informacyjnego" (*data smog*). Smog informacyjny jest wynikiem hiperprodukcji i hiperdystrybucji informacji wielokrotnie przekraczającym ludzkie możliwości percepcyjne[7]. Jeśli nie umiemy otaczających nas danych porządkować, agregować i przetwarzać na wiedzę, skazani jesteśmy na życie w chaosie informacyjnym lub manipulowanie przez tych, którzy przygotowują te informacje za nas. Umiejętność wyszukiwania, krytycznej selekcji i analizy danych jest warunkiem *sine qua non* rozumienia i wyjaśniania otaczającej nas rzeczywistości społecznej i politycznej w XXI wieku. Niniejsza książka powstała z myślą, by nauczyć Czytelnika podstaw samodzielnej analizy danych ilościowych. Dzięki owej samodzielności, uzyskujemy jeszcze jedną - wydaje się, że bezcenną umiejętność - uniezależniamy się od treści przetworzonych i preinterpretowanych przez media masowe. Umiejętność dokonywania analiz danych ilościowych jest jedyną chyba możliwością we współczesnej kulturze informacyjnej suwerenności jednostki, uwolnienia się jej od zmanipulowanych treści. Dysponując danymi źródłowymi zebranymi przez instytucje akademickie lub zgromadzonymi we własnym zakresie oraz posiadając umiejętność przetwarzania tych informacji, a więc zamiany ich w pełnowartościową wiedzę, uzyskujemy niezmanipulowany obraz rzeczywistości społecznej i politycznej. Wymiar praktyczny analizy danych ilościowych należy też rozumieć inaczej - niniejsza publikacja stanowi próbę pogodzenia potrzeb teoretycznego kształcenia akademickiego z praktycznymi potrzebami rynku pracy, starając się wpoić zarówno wiedzę i umiejętności wymagane w toku studiów, jak też nader przydatne w dziedzinie analiz ilościowych wykonywanych na potrzeby komercyjne.

Przyjmując, że Czytelnik może nie posiadać przygotowania matematycznego oraz informatycznego, starano się wyłożyć zagadnienia od podstaw, gwarantując otwartość warsztatu naukowego i informatycznego. Wszelkie treści, które mogłyby sprawiać problemy przedstawiono w sposób możliwie prosty, ale bez zbytnich uproszczeń. Chciano uniknąć zarzutu, że jest to jedynie „książka kucharska", techniczny instruktaż, w jaki sposób uzyskać wynik obliczeń za pomocą programu komputerowego, dlatego wszystkie miary zostały szczegółowo omówione, przedstawiona została ich geneza, a wzory wyprowadzone i szczegółowo wyjaśnione. Czytelnik nie powinien obawiać się również użycia komputera do wykonania obliczeń statystycznych - urządzenie to powinno stać się dla politologa narzędziem pracy i poznania jak mikroskop dla

[5] M. Szreder, *Statystyka w państwie demokratycznym*, „Wiadomości Statystyczne", 2009, 6, s. 6-7.

[6] S. Lem, *Bomba megabitowa*, Wydawnictwo Literackie, Kraków 1999, a także: T. Fiałkowski, S. Lem, *Świat na krawędzi*, Wydawnictwo Literackie, Kraków 2007 oraz S. Lem, *Cave Internetum*, w: http://www.szkolareklamy.pl/sections-viewarticle-398-str_w7-naj_w1.html, dostęp: czerwiec 2012.

[7] D. Shenk, *Data Smog. Surviving the information glut*, HarperSanFrancisco, Nowy Jork 1997, s. 27-28.

biologa. W związku z tym zadbano również o drobiazgowy techniczny instruktaż, tak by nawet niewprawny w obsłudze komputera użytkownik mógł posłużyć się z powodzeniem oprogramowaniem analitycznym. Starano się skonstruować maksymalnie przyjazny i przejrzysty podręcznik mogący służyć do samodzielnej nauki, jak i wspomagający proces nauczania akademickiego oraz jako kompendium na potrzeby praktycznej pracy analitycznej przypominające lub systematyzujące wiedzę badacza.

Warto jednak ostrzec, by nie popadać w „fetyszyzm statystyczny" - wyniki choćby najbardziej wysublimowanych obliczeń i testów statystycznych nigdy nie zastąpią logicznego rozumowania, zdrowego rozsądku i krytycyzmu oraz oparcia się badacza na solidnych faktach i obserwacjach. Ponadto analiza danych ilościowych jest tylko jednym z wielu narzędzi dostępnych politologowi, warto się każdorazowo zastanowić, czy dla rozwiązania określonego problemu badawczego adekwatne jest zastosowanie analiz ilościowych, a jeśli nawet tak, to czy może należałoby owe analizy pogłębić używając innych uznanych w politologii narzędzi badawczych. Przestrzegają przed takim fetyszyzowaniem analiz ilościowych dwaj amerykańscy autorzy Phillip I. Good i James W. Hardin, którzy pierwszą część wspólnego dzieła opatrzyli prowokującym mottem *Don't think - use the computer* („Nie myśl - używaj komputera")[8]. Programy komputerowe przygotowane dla potrzeb stosowania metod statystycznych pozwalają wykonywać obliczenia, przy których dawniej trzeba było spędzić wiele pracowitych godzin, a nawet tygodni lub miesięcy - lub też - w ogóle by ich ze względu na ogromną czasochłonność nie podejmowano. Niestety wraz z tym ułatwieniem pojawiło się niebezpieczeństwo bezmyślnego stosowania metod w sytuacji, gdy prawie wszystko daje się obliczyć. Mamy współcześnie do czynienia - niestety - z nadmiarem informacji i niedoborem wiedzy.

Komputerowym narzędziem analitycznym wykorzystanym do nauki analiz ilościowych na potrzeby niniejszej publikacji jest program GNU PSPP autorstwa amerykańskiego informatyka Bena Pfaffa. Wybór tego właśnie programu nie jest przypadkowy. Nazwa PSPP nawiązuje do nazwy SPSS (*Statistical Package for Social Sciences*). Program SPSS jest jednym z najpowszechniej używanych narzędzi służących do analizy danych ilościowych w zakresie nauk społecznych[9]. Jest on programem komercyjnym. Natomiast PSPP stanowi darmową i dostępną dla wszystkich zainteresowanych alternatywę dla programu SPSS. Program PSPP posiada większość istotnych i użytecznych funkcji SPSS[10]. Jest jego swoistym „klonem". Znajomość jednego z programów umożliwia swobodne korzystanie z drugiego, a drobne różnice pomiędzy nimi nie wpływają istotnie na komfort i jakość pracy. Ważką przesłanką wyboru GNU PSPP były techniczne walory tego narzędzia - pozwala ono na szybkie i sprawne przetwarzanie i analizowanie danych, umożliwia wykonywanie niemal dowolnych operacji na zbiorach danych i na zmiennych, jest przejrzysty, intuicyjny, ergonomiczny. Czyni go to doskonałym narzędziem do zastosowań zarówno akademickich, jak też biznesowych. Wybranie darmowego programu również nie jest przypadkiem - ma ono symboliczne znaczenie i ma wyrażać gorące poparcie autorów dla idei Wolnego Oprogramowania. Oprogramowanie komputerowe z punktu widzenia logiki, funkcjonalności i społecznego przeznaczenia należy traktować tak jak podstawowe twierdzenia fizyki teoretycznej lub matematyki i dlatego żadne z nich nie powinny podlegać ograniczeniom patentowym, nikt nie ma prawa zawłaszczać tego dorobku intelektualnego, w szczególności ze względów komercyjnych. Oprogramowanie jako byt wirtualny

[8] P.I. Good, J.W. Hardin, *Common Errors in Statistics (and How to Avoid Them)*, John Wiley and Sons, New Jersey 2003.

[9] Pomysłodawcą tego programu jest amerykański politolog Sidney Verba, który zadedykował go i tworzył z myślą o politologach.

[10] Warto zasygnalizować, że program SPSS powstał jako dedykowany produkt do obróbki statystycznej danych sondażowych i ta początkowa funkcjonalna specjalizacja pogłębiała się czyniąc z tego produktu (oraz z jego „klona" – programu PSPP) doskonałe narzędzie do analizy danych z wywiadów kwestionariuszowych i ankiet.

nie może podlegać mechanicznie zasadom prawa własności obiektów fizycznych. Traktowanie jako tożsamych produktów materialnych i niematerialnych jest logicznym, moralnym i prawnym nieporozumieniem, bowiem byty te nie podlegają takim fizycznym uwarunkowaniom jak przedmioty materialne - na przykład ich krańcowy koszt powielenia równy jest zero. Patentowe ograniczenia sprzyjają przede wszystkim krótkookresowemu zyskowi nielicznych komercyjnych podmiotów, a powodują negatywne skutki społeczne i ekonomiczne - obniża się poziom wykształcenia ogółu ze względu na spowolniony lub zatrzymany obieg informacji, hamowany jest rozwój myśli - głównie technologii, ale również kultury, w efekcie wychładzana jest gospodarka. Co więcej, patentowanie myśli programistycznej, wiązanie licencjami rozmaitych programów użytkowych jest uderzeniem w podstawową demokratyczną zasadę, jaką jest wolność słowa.

Oddana do rąk Czytelnika książka składa się z sześciu części, z których każda podzielona jest na rozdziały. W części pierwszej Czytelnik odnajdzie podstawowe informacje i nabędzie elementarnych kompetencji w zakresie analizy danych ilościowych. Zadbano tu o wprowadzenie historyczne - zarysowano genezę i rozwój analizy danych ilościowych. Wprowadzenie metodologiczne obejmuje rudymentarne wiadomości na temat istoty analizy danych ilościowych oraz uświadamia Czytelnikowi mechanikę procesu badawczego. Następnie omówiono historię narzędzia analitycznego, którym posługujemy się w niniejszej pracy - programu GNU PSPP, ze szczególnym uwzględnieniem formalno-prawnych aspektów jego użytkowania. Kolejne rozdziały części pierwszej wprowadzają w techniczne podstawy użytkowania programu oraz przekazują wiedzę na temat podstaw pracy ze zbiorami danych ilościowych.

Treść drugiej części ogniskuje się na przygotowaniu zbioru danych do analizy. Czynnościom tym poświęcono odrębną część, bowiem w naukach społecznych badane zjawiska nie są na ogół mierzone bezpośrednio tak, jak ma to miejsce w naukach przyrodniczych. Stąd konieczność kompleksowego przekształcania zbioru danych i zmiennych znajdujących się w nim. W programach komputerowych służących do analizy danych w naukach społecznych, moduły dedykowane do przetwarzania danych na potrzeby przygotowawcze, są bardzo rozwinięte - na ogół ta właśnie cecha odróżnia je od oprogramowania przeznaczonego na potrzeby nauk przyrodniczych. Ponadto, czynności te zajmują badaczowi w naukach społecznych często więcej czasu niż same analizy.

Trzecia część zawiera podstawy statystyki opisowej, począwszy od zasad sporządzania, prezentacji i interpretacji danych w tabelach, a skończywszy na analizie miar pozycyjnych (takich jak średnie, dominanta czy mediana) oraz miar rozrzutu (przede wszystkim wariancji i odchylenia standardowego, a także kwantylowych miar rozproszenia, rozstępu i współczynnika zmienności). W tej części umieszczono także statystyki, które mogą znaleźć ciekawe i płodne heurystycznie zastosowania w politologii – ryzyko względne i iloraz szans.

W czwartej części dokonano wprowadzenia do statystyki indukcyjnej. W części tej wyłożono analizę regresji - najszlachetniejszą metodę statystyczną, bowiem może ona posłużyć do predykcji cech statystycznych zjawisk oraz miary zależności pozwalające na zbadanie siły i kierunku związków pomiędzy badanymi zjawiskami.

Część piąta zawiera obszerny i przystępny wykład na temat wnioskowania statystycznego wykorzystującego testy istotności, w którym znalazły się przystępnie podane informacje na temat sposobów, zasad i ograniczeń wnioskowania o całości danej zbiorowości na podstawie niewielkiej jej części (próby). W części piątej zapoznajemy także Czytelnika z regułami porównywania zbiorowości (między innymi test t-Studenta, analiza wariancji ANOVA). Z kolei w części szóstej znalazło się wprowadzenie do analizy czynnikowej pozwalającej na odnajdywanie ładu w zbiorach danych.

Książkę zamykają aneksy zawierające między innymi wybór polecanej literatury do dalszej nauki, przegląd i ewaluację programów komercyjnych i niekomercyjnych służących do analiz statystycznych, a także omówienie wybranych zbiorów danych z badań społecznych dostępnych w formie elektronicznej.

W toku prac nad publikacją wykorzystano dostępne w języku polskim i angielskim najnowsze oraz klasyczne publikacje poświęcone podstawom statystyki. Wielokrotnie odwoływano się także do dzieł, w których po raz pierwszy ogłaszano odkrycia statystyczne, szczególnie w zakresie statystyki indukcyjnej. Dostępne na polskim rynku wydawniczym publikacje z zakresu analizy danych nader często obarczone są „grzechem niecałości" - nie zaopatrują Czytelników w zintegrowaną wiedzę statystyczną, metodologiczną i informatyczną. Przekazywana wiedza ma najczęściej charakter cząstkowy - obejmuje jedną, co najwyżej dwie wymienione dyscypliny. Lektura takiego podręcznika nie pozwala na osiągnięcie satysfakcjonującego i jednocześnie szybkiego efektu dydaktycznego. Klasyczne podręczniki statystyki nie umożliwiają zastosowania nabytej wiedzy w praktyce badawczej, a z kolei podręczniki zawierające instruktaż techniczny utrudniają pogłębione rozumienie miar statystycznych a w efekcie właściwe ich stosowanie. Problemem jest również to, że większość książek na polskim rynku wydawniczym omawia zastosowanie narzędzi komercyjnych (głównie programu SPSS), do którego użytkownik ma - ze względu na wygórowane ceny licencji - ograniczony dostęp[11].

Ważne źródło stanowiła dokumentacja techniczna programu GNU PSPP oraz innych programów służących do analizy danych. Kluczowe znaczenie w toku prac nad książką miała własna praktyka analityczna autorów - zarówno o komercyjnym, jak również akademickim profilu - w szczególności autorzy mieli zaszczyt analizować dane na potrzeby uczonych ze Szkoły Głównej Handlowej, Uniwersytetu Kardynała Stefana Wyszyńskiego oraz Uniwersytetu Warszawskiego.

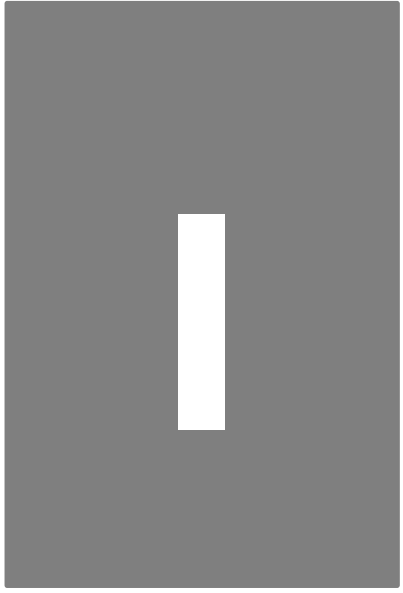
W celu ilustracji miar statystycznych omawianych w książce wykorzystano dane empiryczne pochodzące z przeprowadzonego w 2007 roku Polskiego Generalnego Studium Wyborczego (PGSW). PGSW stanowi odpowiednik rozpowszechnionych w starych demokracjach badań ogniskujących się na problematyce wyborów parlamentarnych. Badanie to stanowi - jak wskazują jego twórcy - „swoisty kanon w dziedzinie socjologii polityki i nauk o polityce". Jest ono afiliowane przy Instytucie Studiów Politycznych Polskiej Akademii Nauk, przeprowadzono je pod kierownictwem polskiego socjologa i publicysty Radosława Markowskiego. W Polsce wykonano je już pięciokrotnie - w latach 1997, 2000, 2001, 2005 i 2007. Dane zbierane są metodą standaryzowanego wywiadu kwestionariuszowego (*Face to Face*, F2F). Zbiór danych z 2007 roku obejmuje reprezentatywną próbę 1817 dorosłych Polaków[12]. Badanie to stanowi doskonały materiał do analiz politologicznych.

Niniejsza publikacja nie ma charakteru zamkniętego. Zastosowania analizy ilościowej w naukach politycznych nie ograniczają się wyłącznie do przedstawionej tutaj elementarnej statystyki opisowej

[11] Pojawiają się na szczęście szlachetne wyjątki. Na przykład kilka polskojęzycznych publikacji poświęcono darmowemu programowi R (jednakże jest on raczej przeznaczony dla użytkowników z przygotowaniem programistycznym): P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Wydawnictwo Naukowe PWN, Warszawa 2012; E. Gatnar, M. Walesiak, *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

[12] Informacje podawane za: *Polskie Generalne Studium Wyborcze 2007*, pod kierownictwem Radosława Markowskiego, afiliowane przy Instytucie Studiów Politycznych PAN, dofinansowane przez tę instytucję, oraz przez: Ministerstwo Nauki i Szkolnictwa Wyższego, Wissenschaftszentrum Berlin fur Sozialforschung (WZB), Polską Konfederację Pracodawców Prywatnych Lewiatan, Fundację Batorego, Instytut Filozofii i Socjologii PAN oraz instytucję badawczą realizującą sondaż - PBS DGA, w:, http://www.ads.org.pl/opis-szczeg.php?id=72, dostęp: lipiec 2012.

i indukcyjnej. Autorzy mają nadzieję rozwijać w przyszłości niniejszy podręcznik w kolejnych wydaniach, by objął również ponadelementarne treści ilościowej analizy danych.



Część I. Wprowadzenie do analizy danych ilościowych

1

Rozdział 1. Geneza i rozwój analizy danych ilościowych

Przez analizę danych ilościowych należy rozumieć zbiór reguł i zasad postępowania z danymi liczbowymi, w taki sposób, aby uzyskać wartościowe poznawczo informacje i użyteczne wnioski. Pozyskiwanie wiedzy z danych może się odbywać na różne sposoby - przede wszystkim wyróżnia się sposób klasyczny oraz nowy - eksplorację danych. Ta ostatnia nie jest przedmiotem rozważań niniejszej publikacji. Tak rozumiana analiza danych ilościowych jest obecnie klasyfikowana jako część działu matematyki i nosi nazwę statystyki. Pod względem funkcjonalnym na statystykę składają się wymienione wyżej procesy analizy danych, ich interpretacja, ale także metody ich gromadzenia. Analiza tych danych wymaga obecnie użycia specjalnych technik, w tym technik ich prezentacji. Klasyfikowanie statystyki jako działu matematyki jest zasadne ze względów metodologicznych, jednakże z funkcjonalnego i historycznego punktu widzenia statystyka znacznie wykracza poza matematykę.

Z **funkcjonalnego** punktu widzenia, statystyka ma charakter praktyczny, jest ona nauką pomocniczą, bogatym zestawem narzędzi umożliwiającym odnajdywanie i weryfikację prawidłowości we wszystkich dziedzinach współcześnie uprawnianych nauk: społecznych, biologicznych, matematycznych, fizycznych i chemicznych, rolniczych, leśnych i weterynaryjnych, medycznych, naukach o ziemi i górniczych. Znajduje ona zastosowanie wszędzie tam, gdzie mamy do czynienia ze zjawiskami masowymi dającymi wyrazić się za pomocą liczb. W politologii, zaklasyfikowanej do nauk społecznych, statystyki używamy do analizy szerokiego spektrum danych liczbowych odnoszących się do życia społecznego, politycznego, kulturalnego i gospodarczego.

Rozpatrując rzecz z perspektywy **historycznej,** statystyka została utworzona i przez większość czasu swego istnienia rozwijała się poza matematyką. Powstała ona przede wszystkim na potrzeby organizacji życia we wspólnotach, konstruowano i postrzegano ją jako ważny, a nawet nieodzowny filar państwa i władzy, a więc - jak dziś powiedzielibyśmy - stosowanej nauki o państwie i polityce. Przez wieki była ona właściwie nierozdzielna z instytucjami państwa. W 1581 roku G. Ghilini nazwał statystykę wiedzą

o przejawach życia w państwie, polityce i wojsku (wł. *civile politica, statistica e militare scienza*)[1]. Napoleon Buonaparte rzekł kiedyś: „*La statistique - c'est le budget des choses, et sans budget - point de salut publique",* słusznie wskazując, iż statystyka jest budżetem spraw, a bez owego budżetu niemożliwe staje się rozwiązanie spraw publicznych. W tym właśnie czasie zaczęto szukać pewnych prawidłowości społeczeństw ludzkich. Ten wymóg w coraz większym stopniu czynił ją nauką ścisłą. Wkład do statystyki innych dyscyplin badawczych, jak astronomia, czy biologia był ważny i trwały. Miał jednak znaczenie poboczne, a statystyka - zarówno w wymiarze teoretycznym, jak też instytucjonalnym rozwijała się jako służebne narzędzie władzy we wspólnotach, a przede wszystkim narzędzie wspomagające rządzenie państwem. Myśl, iż statystyka powstała i rozwijała się na potrzeby państwa i władzy, wymaga szerszego uzasadnienia.

Wyraźnie na tożsamość statystyki wskazuje źródłosłów jej nazwy. Słowa „statysta", „statystyk", „statystyczny" - pochodzą od łacińskiego *status,* co w sensie tego słowa nabytym w średniowiecznym języku łacińskim oznaczało państwo lub polityczny stan rzeczy. Pojęcia tego używano między innymi na określenie rady państwa (*statisticum collegium*), a także - w języku włoskim - na określenie polityka, męża stanu (*statista*). Słowo „statysta" odnajdujemy już w 1602 roku w *Hamlecie* (akt 5, scena 2) Williama Shakespeare'a, a także w *Cymbelinie* (akt 2, scena 4) z 1610 lub 1611 roku tegoż autora oraz w *Raju utraconym* Johna Miltona z 1671 roku. Z kolei najwcześniej słowa statystyka użyto - jak podają George U. Yule i Maurice G. Kendall - w trzytomowym dziele *The Elements of Universal Erudition* Barona J.F. von Bielfelda w 1770 roku. Jeden z rozdziałów tej książki zatytułowany został *Statystyka,* w którym zdefiniowano tytułowe pojęcie jako „naukę o politycznym układzie państw świata". W nieco szerszym znaczeniu pojęcie „statystyka" definiuje Eberhard A.W. von Zimmermann w 1787 roku[2]. Wcześniej pojęcie statystyki używane było w piśmiennictwie włoskim jako przymiotnik. Istotną kategorią staje się statystyka we włoskich państwach-miastach, zainteresowanie sprawami władzy, państwa, polityki. Rozwija się wówczas twórczość polegająca na opisach rozmaitych krajów. Pojawia się też po raz pierwszy pojęcie „statystyka", jednak jedynie jako przymiotnik. Statystykę utożsamiał z państwoznawstwem uczony niemiecki Gottfried Achenwall (1719-1772), wykładowca Uniwersytetu w Marburgu oraz Getyndze, tłumacząc, że jest to gromadzenie, przetwarzanie i wykorzystywanie danych przez instytucję państwa[3]. *W języku polskim jako pierwszy słowa „statystyka" użył Stanisław Staszic w 1809 roku w dziele O statystyce Polski - krótki rzut wiadomości, potrzebnych tym, którzy ten kraj chcą oswobodzić, i tym, którzy w nim chcą rządzić, używał go w rozumieniu Achenwallowskiego państwoznawstwa.*

Wskazuje się, że potrzeba uprawiania statystyki związana jest nierozerwalnie z powstaniem państwa, jednakże jej początków możemy doszukiwać się już znacznie wcześniej. Jest ona niezbędna w każdej świadomej działalności ludzkiej opartej na planowaniu i przewidywaniu, służy bowiem potrzebom najpierw mniejszych wspólnot i związków politycznych (na przykład związków plemiennych), a dopiero w późniejszym okresie - państw. Zadania statystyki wynikały z konkretnych problemów, jakie stawały przed władzami[4]. Warunkiem jej uprawiania był również wynalazek pisma, umożliwiający utrwalanie informacji.

[1] W. Ostasiewicz, *Refleksje o statystyce wczoraj, dziś i jutro*, w: *Statystyka wczoraj, dziś i jutro.* I Ogólnopolski Zjazd Statystyków z okazji 95-lecia Polskiego Towarzystwa Statystycznego i 90-lecia Głównego Urzędu Statystycznego, Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne, Warszawa 2008, s. 5.

[2] G.U. Yule, M.G. Kendall, *Wstęp do teorii statystyki*, Państwowe Wydawnictwo Naukowe, Warszawa 1966, s. 18-19.

[3] S. Szulc, *Metody statystyczne*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1961, s. 10.

[4] C.R. Rao, *Statystyka i prawda,* Wydawnictwo Naukowe PWN, Warszawa 1994, s. 53; W. Skrzywan, *Historia statystyki. Materiały do wykładów,* Państwowe Wydawnictwo Naukowe, Warszawa 1954, s. 8.

Właściwie od samego początku istnienia państw odnajdujemy potrzebę istnienia gromadzenia, analizy i interpretacji danych na potrzeby zarządzania nimi. Wszystkie cywilizacje starożytne korzystały z tego narzędzia prowadząc spisy ludności i innych zasobów państwa na cele administracji państwowej. Najwcześniejsze ślady spisów odnajdujemy już w okresie czterech tysięcy lat przed naszą erą, świadczą o tym sumeryjskie gliniane tabliczki klinowe. Spisy takie prowadzono również w Asyrii, Babilonie i Persji. W Egipcie prowadzono spisy już trzy tysiące lat temu - dynastia Menesa, która zjednoczyła Egipt, wykorzystywała drobiazgowe spisy - ewidencję statystyczną na potrzeby gospodarki planowej. Z kolei w Chinach spisy powszechne, jak podaje Konfucjusz, były wykonywane już w 2300 roku p.n.e. podczas panowania cesarza Yao[5]. Regularne spisy powszechne organizowała półlegendarna dynastia Xia (2100-1600 p.n.e.). Istnieją liczne świadectwa prowadzenia szczegółowych spisów ludnościowych za czasów dynastii Zhou (1122-256 p.n.e.), a następnie - znacznie później - dynastii Ming (1368-1644), która prowadziła je periodycznie. Dynastia Zhou powoływała *szih-su* (księgowych) odpowiedzialnych za wykonywanie prac statystycznych[6]. W Babilonie interesowano się zbiorczymi danymi na temat finansów, obrotów, płac i cen, z kolei w Egipcie i w Chinach państwo scentralizowane wymagało danych na temat zasobów demograficznych[7].

Liczne wzmianki na temat spisów ludności na potrzeby aparatu rządzenia odnajdujemy w Biblii w czwartej księdze *Starego Testamentu* - *Księdze Liczb* (*Numeri*), w której opisano zarządzony przez Mojżesza i Arona spis mężczyzn powyżej 20 roku życia, zdolnych do noszenia broni. Druga *Księga Samuela* (*Septuaginta*, Druga *Księga Królewska*) w 24 rozdziale wspomina o spisie powszechnym przeprowadzonym na rozkaz Dawida przez dowódcę wojsk Joaba. Liczbowe spisy zawierają także księgi Ezdrasza i Nehemiasza - prezentują one szczegółowe listy repatriantów przybyłych z niewoli babilońskiej, w którą wziął ich Nabuchodonozor.

W Indiach, już w starożytności, rozwinął się skomplikowany system sporządzania rejestrów statystycznych na potrzeby administracji. Posiadano wówczas metodologię zbierania tych danych, zawartą w tekstach pisanych oraz w postaci instytucji służących temu celowi[8]. Statystyka była związana z władzą również w państwach-miastach greckich. Była ona sporządzana na bieżące potrzeby administracyjne i wojskowe, nad wyraz rozwinięta, jednak pomiary dokonywane były nieregularnie. Pierwszy spis odbył się w 444 roku p.n.e., zarządził go Perykles, który na potrzeby wojenne nakazał obliczenie ludności Aten. Warto również wspomnieć o zaginionym dziele Arystotelesa *Politeiai*, które miało charakter państwoznawczy[9].

W starożytnym Rzymie spisy oraz rejestry administracyjne były organizowane stosunkowo systematycznie już w okresie Republiki; *Imperium Romanum* prowadziło spisy przez pół tysiąca lat[10]. Rzymskie spisy powszechne wprowadził i spopularyzował Serwiusz Tuliusz, król, władca Rzymu w VI wieku p.n.e. Nazywany jest on ojcem rzymskiego cenzusu i jednocześnie pierwszym, który stworzył swoisty, empiryczny schemat stratyfikacji społecznej. Wprowadził sześć warstw społecznych wyróżnionych

[5] W. Skrzywan, dz. cyt., s. 19.

[6] C.R. Rao, dz. cyt., s. 2.

[7] W. Skrzywan, dz. cyt., s. 19.

[8] C.R. Rao, dz. cyt., s. 53.

[9] S. Szulc, dz. cyt., s. 11.

[10] Tamże, s. 8.

ze względu na socjodemograficzne czynniki stratyfikacyjne. Szczegółowe informacje o spisach rzymskich podaje między innymi Dionizjusz z Halikarnasu[11]. Spisy powszechne oraz spisy majątku były powtarzane regularnie; ostatnie z nich przeprowadzono odpowiednio w 47 i 72 roku. W tym kontekście istotna wydaje się postać Oktawiana Augusta, który przeprowadził w czasie swojego panowania trzy spisy ludności rzymskiej w Galii - w 28 roku p.n.e., 8 roku p.n.e. oraz 14 roku n.e. Dokonywał również spisów ludności w Egipcie i na Sycylii (co czternaście lat). Ponadto Oktawian August napisał pamiętnik zawierający kompletny spis zasobów Rzymu. Wkład nie do pominięcia wniósł Ulpian Domicjusz, rzymski jurysta, który skonstruował na przełomie II i III wieku, prawdopodobnie dla celów aktuaryjnych, pierwsze w historii tablice trwania życia. Spisy rzymskie miały dopracowaną procedurę ich przeprowadzania, były drobiazgowo dokładne, nie miały charakteru anonimowego. Przeprowadzano je w odstępach pięcioletnich, a odpowiedzialni za ich przeprowadzenie byli urzędnicy zwani cenzorami[12]. W Rzymie rozwijała się również myśl statystyczna - Cyceron i Salustiusz podjęli refleksję nad zasadami dokonywania spisów powszechnych, rozważali, w jaki sposób należy owe spisy prowadzić, aby były one adekwatne do rzeczywistości i jak najbardziej dokładne. Można zatem mówić tu o pierwszych próbach budowy teorii statystycznej. Oprócz swoistych podstaw metodologii badań cenzusowych niektórzy rzymscy autorzy, na przykład Ulpian, rozważając kwestie związane z prawem spadkowym, przejawiali pewne intuicje z zakresu teorii prawdopodobieństwa - wykładając myśl, że tym wyższe ryzyko zgonu im wyższy wiek. Tę myśl podejmą dopiero uczeni w XVII i XVIII wieku[13]. Dziedzictwo starożytności przejęli w tej mierze i rozwinęli w kierunku statystyki finansowej i gospodarczej Arabowie.

W średniowieczu myśl statystyczna rozwijała się powoli, a spisy powszechne były na początku tego okresu rzadko przeprowadzane. Jako pierwszy zarządził taki spis Karol Wielki, który nakazał sporządzenie wykazów posiadania swych dóbr i dóbr wasalów w formie brewiarzy i kapitularzy. Spis ten obejmował poddanych powyżej 12 roku życia. Sposób przeprowadzania tego typu spisów Karol Wielki dał w *Capitulare Villis vel curtis imperii* w 812 roku. Drugim znanym spisem powszechnym był przeprowadzony w 1085 roku z rozkazu Wilhelma Zdobywcy, który zarządził szczegółowy opis Anglii w *Domesday Book* (*Księdze Sądu Ostatecznego*). Dokument ten zawiera kataster gruntowy, topograficzny opis Anglii, wskazane są granice poszczególnych dóbr, przedstawiony został także rejestr ludności z podziałem na stany, zawierający informację o dochodach mieszkańców. Ten dokument naśladowano w czasach późniejszych - uczynił to Waldemar II Duński (w 1231 roku), cesarz Fryderyk II (w 1241 roku na Sycylii) oraz książę Karol Kalabryjski (w 1327 roku)[14]. W XIII wieku zaczęli organizować spisy powszechne Tatarzy, a następnie wedle ustalonego wówczas wzorca - książęta moskiewscy - Dymitr Doński (1359-1389) oraz Wasyl Dymitrowicz (1389-1425). Spisy o charakterze cenzusów przeprowadzano rzadko, gromadzono natomiast chętnie dane opisowe, a czynili to kronikarze. Pierwszy informacje liczbowe na temat Polski zebrał Gall Anonim w swojej kronice (było to około 1113-1116 roku)[15].

Przełomowy dla statystyki był wiek XIV, gdy wraz ze zmianami społecznymi, politycznymi i technologicznymi zaczęto stawiać przed nią nowe zadania[16]. Tworzy się wówczas państwo scentralizowane

[11] W. Skrzywan, dz. cyt., s. 21.

[12] C.R. Rao, dz. cyt., s. 53.

[13] W. Skrzywan, dz. cyt., s. 25.

[14] Tamże, s. 25.

[15] J. Berger, *Spisy ludności na ziemiach polskich do 1918 roku,* „Wiadomości Statystyczne", 2002, 1, s. 12.

[16] C.R. Rao, dz. cyt., s. 53; W. Skrzywan, dz. cyt., s. 8.

i pojawia się konieczność większej ilości wiedzy na temat istotnych dziedzin życia. Tu pierwszeństwo należy się państwom włoskim, w szczególności Florencji[17]. Zbierane wówczas informacje miały charakter słowny, a nie liczbowy. Najbardziej aktywnymi ośrodkami były nie dwory, lecz zarządy prowincji, miast, a także hierarchia kościelna. Przykładami takich dokumentów są miedzy innymi księga krajowa Księstwa Wrocławskiego (1358-1367) oraz księga Nowej Marchii (1337), które były spisami ludności miejskiej, żywności i broni. Z kolei pierwszy liczbowy spis przeprowadziły władze Norymbergi w 1449 roku. W 1501 roku władze Augsburga zrealizowały pomysł spisu ruchu ludności - urodzeń, małżeństw i zgonów[18]. Systematyczne i dokładne instytucje spisowe zostały powołane w ramach kościoła katolickiego - obowiązek prowadzenia spisów na poziomie probostw pod groźbą kary ekskomuniki wprowadził w 1563 roku Sobór Trydencki.

W drugiej połowie XVI wieku pojawiła się nowa kategoria - doraźnie sporządzane „ankiety". Zawierały one zapytania o gospodarcze i demograficzne *status quo* danego regionu lub regionów. Pierwszą taką ankietę, składającą się z 75 zapytań, rozesłał król Hiszpanii i Portugalii Filip II do prałatów i corregiorów. Dane zebrane z okręgów posłużyły do opracowania systematycznych, a nie opisowych jak dotąd zestawień[19].

Nowożytna statystyka rodzi się wraz z państwoznawstwem, swoistą wąsko pojmowaną statystyką opisową zajmującą się badaniem państwa, pod względem takich zmiennych jak warunki fizyczne i geograficzne, ludność, ustrój i gospodarka. Tak rozumiane państwoznawstwo zapoczątkował włoski historyk Gioranni Botero (1544-1617), autor wydanego w 1595 roku *Relationi vniuersali di Giouanni Botero Benese diuise in quattro parti*, w której zawarł statystyczny opis krajów, między innymi Polski. Nurt ten był kontynuowany przez Niemców - Hermanna Conringa (1606-1681) i Gottfrieda Achenwalla (1719-1772). Prace G. Achenwalla miały charakter dobrze sporządzonych kompilacji, były wielokrotnie wydawane i tłumaczone, inspirowały one licznych uczonych. Wielu współczesnych i historycznych uczonych nazywa G. Achenwalla ojcem statystyki[20]. W miarę upływu czasu zaczęto coraz częściej posługiwać się danymi liczbowymi, ujmowanymi w postaci tabelarycznych zestawień. Jak wskazuje Paul F. Lazarsfeld nie do przecenienia jest wkład w statystykę H. Conringa jednego z największych polihistorów swojego czasu. Jego wkład polega na próbie usystematyzowania nauki o państwie, tak by była przyswajalna dydaktycznie i jednocześnie łatwa do przetwarzania dla instytucji rządowych z niej korzystających. Sporządzony przezeń projekt nauki o państwie obejmował następujące obszary badawcze: wiedzę o działalności państwa - a konkretnie jego elit i administracji (podejmowanych przez nich decyzjach politycznych i ich skutkach), wiedzę o zasobach ekonomicznych i demograficznych państwa oraz formalnoprawne jego fundamenty (w tym przede wszystkim konstytucję). Po raz pierwszy w twórczości H. Conringa koncepcja ta została zastosowana w studium Hiszpanii[21].

W opozycji do nurtu państwoznawców powstali tabelaryści pragnący zapewnić zestawieniom statystycznym bardziej standaryzowaną, a przez to bardziej czytelną formę. Pomiędzy przedstawicielami niemieckiego państwoznawstwa a tabelarystami wywiązała się zaciekła, acz krótka walka

[17] S. Szulc, dz. cyt., s. 9.

[18] W. Skrzywan, dz. cyt., s. 25.

[19] Tamże, s. 31.

[20] P.F. Lazarsfeld, *Notes on the History of Quantification in Sociology - Trends*, Sources and Problems, „Isis", 1961, 52 (2), s. 284.

[21] Tamże, s. 290-291.

(*Tabellenknechte*) o pierwszeństwo i naukowość stosowanych przez obie strony metod. Państwoznawcy krytykowali brak naukowości tabelarystów, jednakże bardzo szybko naukowość metod tych ostatnich stała się standardem. Tabele, jakimi posługujemy się obecnie, wynalazł duński badacz Johan J. Ancherson (1741). Wypełniał on je jednak treściami słownymi, a nie liczbami. Dopiero August F.W. Crome wprowadził do sporządzanych przez siebie tabel liczby. Podawał w nich między innymi dane dotyczące powierzchni, liczby ludności i gęstości zaludnienia. Było to pierwsze na świecie ujęcie danych liczbowych w tabele. Liczbowe opracowania tabelaryczne jako innowacyjny standard pracy administracji usiłował wprowadzać rosyjski polityk, geograf i kartograf Iwan K. Kirgiłow (1726-1727). Nurt ten zaznaczył się także w dydaktyce akademickiej europejskich uniwersytetów. Jako pierwszy, w 1660 roku, niemiecki polihistor H. Conring rozpoczął wykłady państwoznawcze[22].

Warto podkreślić polski wkład w rozwój państwoznawstwa. W pierwszej kolejności należy wymienić Jana Długosza, który napisał w latach 1470-1480 *Liber beneficiorum dioecesis Cracoviensis* (*Księgę beneficjów diecezji krakowskiej*) zawierającą rejestr dóbr i przywilejów kościelnych i prywatnych. Wartościowym dziełem jest także *Traktat o obojgu Sarmacjach, azjatyckiej i europejskiej* Macieja z Miechowa - rektora Uniwersytetu Jagiellońskiego, lekarza Zygmunta Starego. Jego dzieło zostało przetłumaczone na szereg języków europejskich, miało kilkanaście wydań. Czesław Domański kwalifikuje do polskiego nurtu państwoznawczego także *Kroniki wszystkiego świata* oraz *Kronikę polską* Marcina Bielskiego, żołnierza, historyka, pisarza i poety oraz *Polonia sive de situ, populis, moribus, magistratibus et re publica regni Polonici libri duo* znaną także pod tytułem *Polska* lub *Polonia* opublikowane w 1577 roku, a także *Polonia lub Opisanie Królestwa Polskiego za czasów Zygmunta III* z 1652 roku. Wymienić należy również liczne dzieła Macieja Stryjkowskiego (*Sarmatiae Europeae descriptio* z 1578 roku, *Kronikę Polską, Litewską, Żmudzką i wszystkiej Rusi* wydaną w 1582 roku oraz *Zwierciadło kroniki litewskiej* z 1577 roku). Warto zauważyć, że w średniowieczu zmieniła się jednostka analizy, ma ona charakter prywatno-gospodarczy - są to spisy dóbr króla, arystokracji i szlachty - a nie jednostkowy[23].

Kolejnym, jakościowym etapem rozwoju statystyki było pojawienie się arytmetyki politycznej. Wielu uczonych przekonuje, że wraz z narodzinami tego nurtu mamy prawo mówić o statystyce *sensu stricte*, bowiem dopiero wówczas zaczęto posługiwać się językiem liczb, miar i wag[24]. Do zaistnienia arytmetyki politycznej walnie przyczyniła się doktryna merkantylizmu w XVII wieku w Anglii, której założenia metodologiczne głosiły rozumowanie na podstawie liczb, umożliwiające wykrycie prawidłowości w zjawiskach masowych.

Zdaniem P.F. Lazarsfelda powstanie tego nurtu stanowiło fazę przygotowawczą, ukonstytuowało niezbędny fundament dla współczesnej statystyki. Centralną - i nie do przecenienia - postacią tego okresu jest H. Conring. Rozwój artymetyki politycznej był spowodowany nie tylko swoistym oświeceniowym „duchem epoki". Naturalnie początki kapitalizmu, intelektualny klimat i pierwsze sukcesy nauk przyrodniczych miały swój niebagatleny wpływ. Jednak zwraca uwagę inny, praktyczny czynnik tworzący bariery rozwoju statystyki: obawę mas społecznych przed spisami powszechnymi (co może być niekorzystne ze względu na zwiększenie obciążeń podatkowych) oraz obawami rządzących, że tego typu informacje posłużą wrogim siłom politycznym, mając wymierne militarne znaczenie. Z tego powodu - wskazuje

[22] W. Skrzywan, dz. cyt., s. 33-36.

[23] S. Szulc, dz. cyt., s. 8.

[24] Cz. Domański, *Zasłużeni statystycy dla nauki*,
w: http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Zasluzeni_statystycy_dla_nauki.pdf, dostęp: czerwiec 2012, s. 1-2.

P. Lazarsfeld - ciekawość uczonych dotycząca parametrów demograficznych mogła być zaspokojona w toku obliczeniowych spekulacji (na przykład obliczając liczbę „dymów" i mnożąc przez szacowaną, uśrednioną liczbę osób składających się na rodzinę)[25]. Tak rozumiana artmetyka polityczna narodziła się i rozwijała w Anglii. Wybitnym przedstawicielem arytmetyki politycznej był John Graunt, wcześniej kupiec, a następnie uczony i pierwszy spośród stanu mieszczańskiego członek Królewskiego Towarzystwa Naukowego (*Royal Society*). Wydał on w 1662 roku swoje jedyne dzieło *Naturalne i polityczne obserwacje poczynione nad biuletynami śmiertelności* (*Natural and Political Observations Upon the Bills of Mortality)*. Data publikacji jest uznawana przez specjalistów za narodziny statystyki jako nauki. Dzieło J. Graunta przyczyniło się do rozwoju sposobu myślenia o analizach i prowadzenia ich w praktyce. Praca ta zawiera studium śmiertelności i narodzin Londyńczyków. Źródła, z których skorzystał J. Graunt, były wysoce niedokładne (a były to rejestry zgonów gromadzone przez władze miejskie - *Bills of mortality* z lat 1604-1661), a wyniki nieadekwatne do rzeczywistości. Jednakże nie stan obliczeń jest tu ważny, lecz pewien standard i sam schemat odkrywania prawidłowości na podstawie liczb. Po raz pierwszy w dziele tym sformułowano prognozy dotyczące liczby ludności. Szacunek liczby ludności Londynu przeprowadził J. Graunt dwoma sposobami: pierwszy ze sposobów polegał na oszacowaniu na podstawie znanej liczby zgonów liczby narodzin, a następnie liczby matek, rodzin i - w efekcie - zbiorczej liczby mieszkańców. Drugie podejście polegało na analizie mapy miasta i szacowaniu opartym na przybliżonym założeniu, że na 100 jardach kwadratowych mieszkają 54 rodziny. Uznaje się, że z punktu widzenia metodologii praca J. Graunta stoi u progu badań współczesnych[26]. Praca J. Graunta opierała się na prostych względnych częstościach i wykorzystaniu wartości średnich; autor jako jeden z pierwszych dostrzegł zależność pomiędzy wielkością liczby analizowanych przypadków a dokładnością i stabilnością wartości średniej. Wykorzystywał on także, choć wówczas jeszcze dalece niedopracowaną, koncepcję prawdopodobieństwa[27]. Następcami J. Graunta byli William Petty, a następnie Gregory King i Charles Davenant. W. Petty'emu, lekarzowi i ekonomiście, który jako pierwszy obliczył dochód narodowy Anglii i Walii, zawdzięczamy wprowadzenie do obiegu naukowego pojęcia „arytmetyka polityczna"; wprowadził je w wydanym w 1679 roku dziele noszącym tytuł *Political Arithmetic*. Ponadto W. Petty we wcześniejszej pracy (1672) sformułował liczne, ciekawe i odkrywcze hipotezy na temat prawidłowości życia społecznego. Były one na nowo podejmowane przez badaczy dopiero w XIX i XX wieku[28]. Z kolei G. King usiłował w 1696 roku oszacować ludność Anglii, a Ch. Davenant jako pierwszy sformułował teoretyczne podstawy arytmetyki politycznej, definiując ją jako sztukę rozumowania za pomocą liczb o sprawach dotyczących rządzenia. Znaczny wkład w rozwój arytmetyki politycznej wniósł angielski astronom Edmund Halley (1656-1742) (*notabene:* jego nazwiskiem nazwano kometę). Rozwinął on stosowaną przez J. Graunta metodologię i na podstawie danych zawartych w księgach zgonów i urodzeń dla miasta Wrocławia stworzył pierwszą nowoczesną tablicę trwania życia. Prace te kontynuował szwajcarski matematyk, fizyk i astronom Johann A. Euler (1707-1783). W tym nurcie znajdują się także Christiaan Huygens (1629-1695) oraz Kasper Neumann (1648-1715). Pierwszy z nich analizował czas trwania życia oraz prawdopodobieństwo zgonu człowieka w poszczególnych latach życia. Jego dzieło kontynuował drugi z wymienionych, walnie przyczyniając się do zracjonalizowania przekonań o cyklu życia ludzkiego. W owych czasach powszechnie uważano, że każdy siódmy rok życia człowieka jest krytyczny, zaś

[25] P.F. Lazarsfeld, *Notes on the History...*, jw., s. 282-283.

[26] W. Skrzywan, dz. cyt., s. 47.

[27] *History of Statistics,* „Stochastikon", w: http://132.187.98.10:8080/encyclopedia/en/statisticsHistory.pdf, dostęp: czerwiec 2012.

[28] P.F. Lazarsfeld, *Notes on the History...*, jw. , s. 282-284.

szczególnie niebezpieczne są lata 49 i 63. Posługując się wynikami obliczeń K. Neumann udowodnił, że nie istnieją lata „klimakteryczne". Podobnie, obalił potoczne mniemanie o związku faz księżyca i zdrowia ludzkiego. Artymetyka polityczna rozwijałaa się również w Niemczech - P. Lazarsfeld przywołuje między innymi J.P. Suessmilcha (1707-1767), który studiował medycynę i zajmował się teologią. Jego prace stanowiły syntezę dotychczasowego dorobku arytmetyków politycznych, dodatkowo uczony ten zgłosił kilka ciekawych spostrzeżeń dotyczących powtarzalnych prawidłowości życia społecznego. P. Lazarsfeld pokazuje, że arytmetyka polityczna nie była tylko mniej lub bardziej wielowymiarowym monograficznym opisem społecznej i politycznej rzeczywistości, lecz również w jej ramach starano się formułować (często trafne) hipotezy dotyczące prawidłowości społecznych i politycznych[29]. Ważny wkład w niemiecką arytmetykę polityczną wniosła również szkoła getyńska (powstała na Uniwersytecie w Getyndze)[30]. Warto przytoczyć jako ciekawostkę, że pierwszy na gruncie polskim użył pojęcia arytmetyka polityczna, tłumacząc je na „kałkuł polityczny" Józef Wybicki w 1814 roku. Dwa przedstawione powyżej nurty myślenia - państwoznawstwo i arytmetyka polityczna na nowo wyznaczyły statystyce ważne zadanie - miała być ona „narzędziem rządzenia". Statystyka spełniła pokładane weń nadzieje - stała się jednym z instrumentów absolutyzmu oświeconego, a następnie państwa wczesnego kapitalizmu[31].

Zasady poznawania świata polityki za pomocą liczb rodziły się długo - istotny wkład miała również praktyka zbierania danych przez instytucje państwowe w toku spisów powszechnych. Kultywowana w starożytności tradycja scentralizowanych, zawiadywanych przez władzę państwową cenzusów odrodziła się w nowożytności. Pierwsze postulaty prowadzenia spisów powszechnych podnoszono już na początku XVI wieku (jako pierwszy uczynił to francuski prawnik i polityk Jean Bodin[32]). Sebastian le Pretre de Vauban napisał na przełomie XVII i XVIII wieku dzieło *Projekt d'une Dixe Royale*. Ukazało się ono w 1707 roku. Wskazał w nim potrzebę badania struktury społecznej ludności i nakreślił projekt organizacji służby statystycznej. Służba ta miałaby pozostawać w gestii monarchy, przedstawiać panującemu niezafałszowany obraz rzeczywistości. Z analogicznym pomysłem wystąpił Gottfried W. Leibniz (1646-1716), przedstawiając listę 56 szczegółowych zagadnień statystycznych i demograficznych, które należy poznać z punktu widzenia funkcjonowania państwa. Do dziś wszystkie z nich zrealizowano. Między innymi postulował on utworzenie pozostającej w gestii państwa instytucji zajmującej się statystyką[33]. Początkowo władze organizowały lub zlecały lokalne spisy. W końcu XVII wieku (1696) zaczęto wydawać w Anglii tak zwane *parliaments papers* - zestawienia statystyczne przedkładane w celach informacyjnych parlamentowi, a w 1719 roku Fryderyk Wilhelm I zarządził opracowywanie tabel statystycznych dotyczących ludności, jej struktury zawodowej i majątków oraz budżetów lokalnych jednostek administracyjnych - miast. Pierwszy nowożytny spis przeprowadziła w 1749 roku Szwecja. Obowiązek prowadzenia spisów powszechnych wprowadziły Stany Zjednoczone - jako pierwszy i jedyny kraj na świecie zawarował w konstytucji instytucję spisu powszechnego.

[29] P.F. Lazarsfeld, *Notes on the History...*, jw., s. 282-283.

[30] Tamże, s. 292.

[31] W. Skrzywan, dz. cyt., s. 36, 50-51.

[32] Tamże, s. 30.

[33] Tamże, s. 53.

Konstytucja amerykańska w paragrafie 2 artykułu I głosi:

Najbliższe obliczenie nastąpi w ciągu trzech lat od pierwszego zebrania się Kongresu Stanów Zjednoczonych, a następne spisy ludności dokonywać się będą co dziesięć lat w sposób określony ustawą.

Pierwszy spis powszechny w Stanach Zjednoczonych odbył się w 1790 roku, lecz dopiero w 1830 roku zaczęto się posługiwać standaryzowanymi, drukowanymi kwestionariuszami. Ważnym uregulowaniem prawa amerykańskiego był fakt, że państwo nie rościło sobie monopolu na zbieranie danych statystycznych, mogły to także czynić bez przeszkód podmioty prywatne. Z kolei Francja przeprowadzała spisy od 1801 roku co 5 lat, po wojnie co 7 lat, Wielka Brytania od 1801, co 10 lat, Norwegia rozpoczęła przeprowadzanie spisów powszechnych w 1815 roku, Holandia w 1829 roku, Dania w 1840 roku, a Belgia w 1846. Swój pierwszy spis powszechny przeprowadziły Niemcy w 1871 roku, a Rosja na przełomie 1896 i 1897 roku. Szczególne znaczenie ma spis belgijski w 1846 roku, bowiem spełnia on wszelkie współczesne kryteria metodologiczne naukowości. Regularne, rygorystyczne pod względem metodologicznym spisy ludności stały się powszechną praktyką, standardem działania wszystkich państw świata w XX wieku. Warto również wspomnieć ogłoszoną przez Karola Marksa we francuskim piśmie *Revue Socialiste* ankietę składającą się ze stu pytań. Skierowana ona została do robotników i miała ujawniać rzeczywisty poziom ubóstwa tej klasy społecznej. Wyniki zamierzano opublikować w wymienionym piśmie, a następnie odrębnej monografii na ten temat. Od przełomu XVIII i XIX wieku rozpoczęto tworzenie służby statystycznej - na kontynencie scentralizowanej, a w Wielkiej Brytanii w powstającej przy poszczególnych gałęziach administracji. W roku 1800 utworzono we Francji *Bureau de la Statistique Generale* (Biuro Statystyki Powszechnej). Zlikwidowano je jednak w 1812 roku, a ponownie otworzono w 1833 roku. Warto zwrócić uwagę, na fakt, że państwo wymagało usług statystycznych nie tylko w czasie pokoju, ale również podczas wojen - opracowania statystyczne sporządzano na przykład w czasie podbojów Napoleona. We Francji na początku XIX wieku zaczęto zbierać dane statystyczne w szerszym zakresie. Pojawia się wówczas pojęcie ankiety. Termin „ankieta" wywodzi się od francuskiego słowa *enquête* (z fr. dochodzenie, wywiad, śledztwo, badanie) i oznacza zbieranie materiałów w terenie. W XIX wieku tworzą się i rozwijają towarzystwa statystyczne - powstają różne specjalistyczne subdyscypliny w ich ramach: kryminologiczna, demograficzna, gospodarcza. W połowie XIX wieku podejmuje się intensywne prace w zakresie statystyki społecznej, analizuje się wymieralność w rozmaitych zawodach, stopę zgonów upośledzonych społecznie - ubogich i podrzutków, bada się zdrowie marynarzy i żołnierzy przebywających w krajach tropikalnych, a także zdrowotność grup etnicznych[34].

Warto przyjrzeć się bogatemu i wartościowemu metodologicznie dorobkowi Polski w tym zakresie. Pierwszy polski spis powszechny przeprowadzono w 1789 roku na mocy uchwały Sejmu Czteroletniego (1788-1792), miał on na celu oszacowanie wpływów podatkowych na potrzeby utrzymania armii. Inicjatorem tego spisu nazwanego *Lustracja dymów i podanie ludności* był poseł ziemi bracławskiej Fryderyk J. Moszyński[35], który zasłużył na miano pierwszego polskiego „analityka danych", bowiem to on sporządził tablice prezentujące strukturę ludności kraju na podstawie uzyskanych w toku przeprowadzonego spisu powszechnego wyników[36]. Wcześniej pojawiały się w Polsce spisy lokalne.

[34] Tamże, s. 118-119.

[35] J. Berger, dz. cyt., s. 13.

[36] Dzień, w których F.J. Moszyński wygłosił podczas posiedzenia Sejmu Czteroletniego mowę uzasadniającą przeprowadzenie spisu statystycznego na ziemiach polskich (stało się to 9 marca) został na zebraniu Komitetu Statystyki i Ekonometrii PAN 2 grudnia 2008 roku ogłoszony Dniem Statystyki Polskiej.

W 1765 roku odbył się spis ludności żydowskiej, a w 1777 roku - spis ludności miast. Polska ma też swoje tradycje pod względem szacunków liczby ludności. Wysiłek ustalenia liczby osób zamieszkujących Rzeczpospolitą podjął między innymi Józef Wybicki w 1777 roku, który obliczył, że żyje w niej nieco ponad pięć milionów osób. Szacowania liczby ludności dokonali również Aleksander Buschning w 1772 roku (8,5 miliona), Stanisław Staszic w 1785 roku (6 milionów) oraz Fryderyk Moszyński w 1789 roku (uznał, że w Polsce żyje nieco ponad 7 milionów ludzi)[37]. W czasie istnienia Księstwa Warszawskiego (1807-1815) przeprowadzono aż trzy spisy - w 1808, 1810 i 1812 roku. Zbieranie danych trwało wiele miesięcy, nie obejmowało wszystkich grup ludności, badano odrębnie ludność wsi i ludność miast. Przedmiotem zainteresowania badaczy były podstawowe zmienne socjodemograficzne ludności - wiek, płeć, wyznanie i sytuacja zawodowa. W 1810 roku powołano w Polsce jedną z pierwszych instytucji statystycznych w Europie - Biuro Statystyczne. W Królestwie Polskim jego funkcję pełniła Komisja Rządowa Spraw Wewnętrznych i Duchownych. Pod zaborami funkcjonowały instytucje statystyczne podległe państwom zaborczym. Najbardziej niezależną było galicyjskie Krajowe Biuro Statystyczne we Lwowie funkcjonujące w latach 1873-1918, którym przez wiele lat kierował Tadeusz Z. Piłat, uczony specjalizujący się w statystyce rolnej, pionier polskich badań statystycznych, współzałożyciel i pierwszy Polak pośród członków Międzynarodowego Instytutu Statystycznego (The International Statistical Institute, ISI). Pod zaborem rosyjskim przeprowadzono jedyny spis powszechny - w 1897 roku. Liczne spisy odbyły się w Galicji (od 1773 roku), zawierały one jednak poważne nieścisłości. Z kolei spisy pod zaborem pruskim zaczęto przeprowadzać od 1816 roku. Należy podkreślić, że publikacje statystyczne i dorobek polskich statystyków wydany w latach I wojny światowej (były to między innymi roczniki statystyczne oraz *Geograficzno-statystyczny atlas Polski*) stały się istotnym źródłem i argumentem w wyznaczaniu granic Polski w rokowaniach pokojowych w Paryżu w 1919 roku i w Rydze w 1921 roku[38].

Wiek XVII, XVIII i XIX to rozwój nie tylko statystyki praktycznej, spisowej, lecz również probabilistyki i miar analitycznych innych niż tabele częstości. Ich wartość jest nie do przecenienia, wywarły one decydujący wpływ na wszystkie dziedziny nauki. W rozwój tych miar zaangażowane były astronomia i biologia, ale także nauki społeczne[39]. Pierwsze miary tendencji centralnej (średnie) wynaleziono w XVII wieku. W 1635 roku zaczęto używać średniej - jako pierwszy uczynił to Henri Gellibrand, a średnią ważoną zastosował dopiero blisko sto lat później (w 1722 roku) Roger Cotes. Jednak kluczowymi odkryciami były rachunek prawdopodobieństwa i testy statystyczne. Za pierwszą, raczej symboliczną i anegdotyczną próbę zastosowania w praktyce zasad probabilistyki uznać można prowadzoną w wiekach średnich w Anglii swoistą kontrolę jakości monety bitej w mennicy. Za czasów panowania Henryka II (1154-1189) zapoczątkowano tradycję *Trial of the Pyx*. Słowo *pyx* w staroangielskim oznacza „skrzynię", a zatem należy tłumaczyć ten związek frazeologiczny jako próbę lub doświadczenie dokonane z użyciem skrzyni. Zabieg ów służył testowaniu jakości wytwarzanych monet, a polegał na zbieraniu w ciągu trzech lub czterech lat wybieranych losowo, każdego dnia pracy mennicy monet. Z punktu widzenia używanej współcześnie nomenklatury można określić ów zabieg mianem testu statystycznego średniej wartości w populacji.

W istocie, rachunek prawdopodobieństwa zaczął rozwijać się na przełomie XVII i XVIII wieku - jako prekursorów tej dziedziny wskazuje się Johna Arbuthnota, Giorlamo Cardano, Pierre'a de Fermata,

[37] Cz. Domański, dz. cyt., s. 2.

[38] Tamże, s. 3, 4, 8.

[39] S.E. Fienberg, *A Brief History of Statistics in Three and One-Half Chapters: A Review Essay*, „Statistical Science", 1992, 7 (2), s. 208-225.

Blaise'a Pascala, Christiana Huygensa. Istotny wkład wnieśli Jakob Bernoulii wydanym pośmiertnie w 1713 roku dziełem *Ars Conjectandi* oraz Abraham de Moivre *Doktryną szans* (*Doctrine of the Chances*) opublikowaną w 1718 roku.

Największe znaczenia mają prace Carla F. Gaussa, Pierre'a S. Laplace'a oraz Siméona D. Poissona. C.F. Gauss i P.S. de Laplace zebrali i rozwinęli teorię probabilistyki na przełomie XVIII i XIX wieku. P.S. de Laplace zawarł cały swój dorobek w *Theorie analytique des probabilite,* zadedykowanej Napoleonowi, którego był gorącym zwolennikiem. W pracy tej dał wyraz przekonaniu, że prawdopodobieństwo każdego zdarzenia może zostać wyrażone za pomocą liczb. Teorię tę stosował zarówno do badań statystycznych, jak i prawnych. W dziele tym opublikował wyniki pierwszego spisu opartego nie na badaniu całości populacji, lecz na jej części - próbie. Dobrał on w sposób losowy 30 departamentów, a w ramach każdego z nich, celowo pewną liczbę gmin, w których mer był „wystarczająco inteligentny i staranny". Idee te nie spotkały się jednakże ze zrozumieniem mu współczesnych i dopiero pod koniec XIX wieku uczeni zaczęli je wykorzystywać i rozwijać. Badacz ten wykrył wiele ciekawych zależności, na przykład stałość odsetka niedostarczonych przez pocztę listów. Jednym z istotniejszych jego odkryć jest wprowadzenie pojęcia przedziału ufności - oszacował on liczbę ludności Francji z pewnym liczbowo wyrażonym prawdopodobieństwem[40]. Właściwie współczesna statystyka opiera się na dokonaniach P.S. de Laplace'a; zostały one tylko rozwinięte i ulepszone. C.F. Gaussowi przypisuje się wprowadzenie rozkładu normalnego nazywanego też rozkładem Gaussa lub (od kształtu tego rozkładu) krzywą dzwonową. Jest to jeden z najważniejszych rozkładów prawdopodobieństwa odgrywający istotną rolę w opisie społecznych, gospodarczych, przyrodniczych i medycznych zjawisk losowych. Kształt krzywej dzwonowej przyjmuje na przykład rozkład wzrostu lub inteligencji w populacji ludzkiej. Równie istotna jest postać S.D. Poissona, który sformułował prawo wielkich liczb, a także przedstawił rozkład nazwany od jego nazwiska rozkładem Poissona, uważany za niemal równie ważny jak rozkład normalny. Rozkład ten był następnie wynajdywany niezależnie przez Williama Gosseta oraz Władysława Bortkiewicza, nazywany czasami „prawem zdarzeń rzadkich" lub „prawem małych liczb". Ważny wkład w rozwój statystyki wniosły prace Wilhelma Lexisa (1837-1914). W pracy zatytułowanej *O teorii zjawisk masowych w społeczeństwie ludzkim* wyłożył zasady rachunku prawdopodobieństwa i analizy ilościowej danych. Z kolei w dziele opublikowanym w 1879 roku *O teorii stałości szeregów statystycznych* wykazał, że rozmaite zjawiska zachodzące w społeczeństwie przyjmują rozkłady dwumianowy i rozkład Poissona. Zagadnieniami z tego zakresu - teorią liczb oraz teorią prawdopodobieństwa zajmował się Pafnutij L. Czebyszow. Jego uogólnione prawo wielkich liczb należy do kanonu matematyki. W Polsce probabilistykę popularyzował Jan Śniadecki. Wydał on w 1817 roku publikację *O rachunku losów,* w której omawiał między innymi dwumian Newtona (w 1790 roku napisał *Rachunek zdarzeń i przypadłości losu*, lecz rozprawa ta nie została opublikowana). Zasady rachunku prawdopodobieństwa zaczęto stosować w praktyce na początku XX wieku. Pierwsze badanie na próbie populacji przeprowadzono w Wielkiej Brytanii w 1912 roku. Dotyczyło ono warunków życia klasy robotniczej w miastach Stanley oraz Reading. Przeprowadzili je Bowley oraz Benett-Hurst (1915). Z kolei pierwsze rządowe badanie próbkowe w Wielkiej Brytanii przeprowadzono w 1937 roku dla brytyjskiego Ministerstwa Pracy. Skupiało się ono na wydatkach klasy robotniczej. Kluczowy jest także wkład amerykańskiego socjologa i statystyka George'a Gallupa, pioniera współczesnych badań opinii publicznej, który w 1940 roku w ramach Spisu Ludności (1940 Population Census) przeprowadził badania na 5-procentowej próbie amerykańskiej populacji. Badanie to stanowiło uzupełnienie cenzusu[41]. Analiza danych ilościowych rozwinęła się wydatnie wskutek upowszechnienia

[40] W. Skrzywan, dz. cyt., s. 65-66, 106.

[41] M. Szreder, *Statystyka w państwie demokratycznym*, „Wiadomości Statystyczne", 2009, 6, s. 6-7.

sondażowych badań opinii (*straw polls*, *opinion polls*). Za pierwsze tego typu badanie uważa się przeprowadzony lokalny sondaż w 1824 roku w stolicy stanu Pensylwania - Harrisburgu. Pomiar ten wskazał jako zwycięzcę wyścigu prezydenckiego Andrew Jacksona. Prognoza ta okazała się trafna i dlatego sondaże stopniowo zaczęły zyskiwać na popularności, a na przełomie wieków były już powszechnie stosowane przez przez lokalne i ogólnokrajowe amerykańskie czasopisma. Wymierny wkład w popularyzację tego typu badań wniosło również czasopismo „Literary Digest", które metodą ankiet pocztowych trafnie przewidywało wyniki kolejnych czterech prezydenckich wyborów (w 1920, 1924, 1929 i 1932 roku)[42]. Warto zaznaczyć również wkład innego pioniera badań sondażowych i marketingowych Elmo Burnsa Ropera, któremu udało się przewidzieć zwycięstwo F.D. Roosvelta w 1936 i 1940 roku oraz odsetek głosów oddanych na tego kandydata z dokładnością poniżej jednego procenta[43]. Pierwszym amerykańskim prezydentem powszechnie korzystającym z sondaży był F.D. Roosvelt, który sprawując urząd prezydenta powołał zespół prowadzący badania sondażowe na jego polityczne potrzeby. Zespołem tym zarządzali amerykański polityki fińskiego pochodzenia Emil Edward Hurja oraz profesor psychologii Hadley Cantril.

Nie do przecenienia jest wkład w rozwój statystyki Adolphe'a Quételeta, który poświęcił się badaniu prawdopodobieństwa, analizował też wyniki spisów powszechnych[44]. Uczony ten miał ambicje zostać Izaakiem Newtonem socjologii. Niektórzy uważają, że statystyka stała się nauką za sprawą A. Quételeta, nazywają go wręcz patriarchą statystyki[45]. Zasługą A. Quételeta jest wykorzystanie krzywej normalnej do analizy i prognozowania przestępstw kryminalnych. Uprawianą przez siebie dyscyplinę badacz nazwał statystyką moralną (*statistique morale*). Największą zasługą A. Quételeta było wprowadzenie pojęcia człowieka średniego (*l'homme moyen*), traktował je jak odpowiednik środka ciężkości ciał fizycznych; średnia ta dotyczyła zmiennych socjodemograficznych; badał on również zasady odchylania się od wartości średniej. Współcześni A. Quételetowi uznali pojęcie człowieka średniego za genialne, a inni uczynili je przedmiotem kpin. A. Quételet zauważał również, że zjawiska świata społecznego mają charakter przyczynowo-skutkowy oraz, że każdy skutek jest proporcjonalny do przyczyny (obecnie tę początkowo intuicyjną myśl ujęto w formalne ramy i występuje ona jako popularna między innymi w politologii analiza LISREL - *linear structural relation analysis*). Związki przyczynowo-skutkowe są jednak możliwe do obserwowania w sposób niedokładny - stąd zainteresowanie tego badacza prawem błędu[46]. Jego uwagę zwróciły również obserwacje skrajne - nazywał je „monstrualnymi". Sformułował zasadę, że im większa liczba pomiarów, w tym większym stopniu pomiary typowe dominują nad nietypowymi. Używając koncepcji

[42] W 1936 roku „Literary Digest" po raz pierwszy omylił się w swoich prognozach wskazując jako zwycięzcę Alfa Landona, który jednak przegrał z F.D. Roosveltem. Pomyłka polegała na tym, że dobór próby został zaburzony na skutek zmian socjodemograficznych wywołanych kryzysem ekonomicznym. Operat losowania do pocztowej ankiety pochodził z list abonentów telefonicznych oraz posiadaczy aut. W sytuacji kryzysu liczba posiadaczy telefonów i samochodów została ograniczona do najbardzej majętnych Amerykanów – a ci reprezentowali preferencje polityczne odmienne niż preferencje mas.

[43] J.M. Converse, *Survey Research in the United States: Roots and Emergence 1890-1960*, University of California Press, 1987, s. 113.

[44] Poglądy na temat wkładu A. Quételeta są podzielone. Wacław Skrzywan krytykuje A. Quételeta, pomniejszając jego dokonania i dorobek; według tego uczonego błędnie pojmował on prawo wielkich liczb. Uważał on, że A. Quételet był przede wszystkim popularyzatorem statystyki, a nie samodzielnym uczonym. W. Skrzywan, *Historia statystyki. Materiały do wykładów*, Państwowe Wydawnictwo Naukowe, Warszawa 1954, s. 109-111.

[45] Obszernie, acz w sposób zdystansowany charakteryzuje sylwetkę tego uczonego P.F. Lazarsfeld: P.F. Lazarsfeld, *Notes on the History...*, jw., s. 294 i n.

[46] W. Ostasiewicz, L.A.J. *Quetelet: patriarcha statystyki*, „Śląski Przegląd Statystyczny", 2006, 5 (11), s. 7, 12-13, 16.

przeciętnego człowieka, A. Quételet usiłował tłumaczyć takie zjawiska społeczne jak samobójstwa, przestępczość, małżeństwa. Dostrzegał, że prawa mają charakter masowy i żadna jednostka nie jest od nich wolna. Sugerował swoisty determinizm w życiu ludzkim, uważając, że i tak niezwykłe i nietypowe zjawiska rozmyją się w przeważającej masie przeciętności. Jako pierwszy zastosował tabele krzyżowe rozumiane jako zestawienie ze sobą wartości dwóch zmiennych - jednej zmiennej w wierszach, a drugiej zmiennej - w kolumnach[47]. Jego odkrycia miały charakter praktyczny - po przebadaniu wzrostu 100 000 rekrutów francuskich i próbie dopasowania do tych wyników krzywej rozkładu normalnego odkrył, że około 2000 mężczyzn uchyliło się od poboru, pozorując wzrost niższy od minimalnego wymaganego. Jako ciekawostkę można podać, że A. Quételet jest również autorem wskaźnika określania otyłości (BMI, *Body Mass Index,* wcześniej - Quételet Index)[48].

Podstawy współczesnej statystyki w zakresie testów stworzyli Francis Galton, August Bravais, Karl Pearson i William Gosset. F. Galton w 1869 roku opublikował pracę *O dziedziczności naturalnej,* w której ogłosił odkryte przez siebie prawo regresji, szerzej wyłożone w dalszych partiach niniejszej publikacji. Ponadto F. Galton a następnie August Bravais dali podstawy rachunku korelacyjnego (słynny wzór na korelację R nazywa się co prawda powszechnie od nazwiska Pearsona, lecz czasami także Bravaisa-Pearsona). A. Bravais zapoczątkował ponadto w 1846 roku teorię współzmienności. Najpoważniejsze zasługi dla statystyki ma Karl Pearson (wcześniej: Carl), profesor matematyki i eugeniki, działający na polu antropologii, historii i statystyki. Jest on autorem wzmiankowanego współczynnika korelacji R noszącego jego nazwisko oraz jednego z największych i szeroko stosowanych testów statystycznych - testu chi-kwadrat[49]. Teoria sprawdzania hipotez statystycznych została ugruntowana przez Jerzego Spławę-Neymana oraz K. Pearsona. Kluczową postacią współczesnej statystyki jest Ronald A. Fisher. Książka R.A. Fishera *The Design of Experiment* okazała się niezwykle wpływowa - wprowadził w niej testy statystyczne wraz z koncepcją hipotezy zerowej. Według zasad sformułowanych przez R.A. Fishera prowadzone są obecnie eksperymenty stochastyczne. W uproszczeniu eksperyment ten polega na losowym podziale testowanej grupy na dwie części: eksperymentalną i kontrolną. Na pierwszą z nich oddziałujemy właściwym czynnikiem, o którym przypuszczamy, że jest zmienną niezależną. Druga grupa otrzymuje bodziec neutralny. Następnie wyniki uzyskane w obydwu grupach porównuje się. Wielu statystyków, współczesnych R.A. Fisherowi kwestionowało tę zasadę, jednak nie zaproponowano dotąd żadnej innej metodologii[50]. R.A. Fisher wynalazł także analizę wariancyjną dla porównywania różnic pomiędzy grupami u wykazywania ich statystycznej istotności. Metody R.A. Fishera zostały przyjęte w nauce na całym świecie[51]. Nie do przecenienia dla rozwoju statystyki, a konkretnie dydaktyki metody badań próbkowych jest wkład polsko-amerykańskiego matematyka Jerzego Spławy-Neymana. Szczególne znaczenie ma napisana przezeń i wydana w Warszawie w 1933 roku przez Instytut Spraw Społecznych broszura *Zarys teorii i praktyki badania struktury ludności metodą reprezentacyjną*. Wartościową z dydaktycznego punktu widzenia częścią tej publikacji jest synteza dotycząca rachunku prawdopodobieństwa oraz teoretycznych i praktycznych aspektów doboru próby. Zagadanienia te wyłożył uczony w pierwszych

[47] V. Coven, *A History of Statistics on the Social Sciences*, „An Academic Journal on the Web", 2003, s. 1-5.

[48] W. Ostasiewicz, dz. cyt., s. 18-19.

[49] Friedrich R. Helmert odkrył w 1875 roku chi-kwadrat. Ponownie wynalazku tego dokonał niezależnie K. Pearson w 1900 roku oraz William Gosset w 1908 roku.

[50] C.R. Rao, dz. cyt., s. 30.

[51] J.C. Stanley, *The Influence of Fisher's 'The Design of Experiments' on Educational Research Thirty Years Later*, „American Educational Research Journal", 1966, 3 (3), s. 223-229.

czterech rozdziałach pracy. Dzieło zostało wznowione rok później w języku angielskim i w zakreślonych granicach uzupełnione[52]. Od tego czasu można mówić o statystyce uprawianej w sposób standaryzowany na poziomie akademickim.

Polska ma swoje ugruntowane tradycje badań ilościowych - pierwsze z nich zostały przeprowadzone w 1903 roku wśród studentów Politechniki Warszawskiej. Wysłano wówczas 1000 ankiet i otrzymano 140 zwrotów. W 1904 roku ukazała się na podstawie uzyskanych wyników nakładem „Naszej Księgarni" kilkudziesięciostronicowa broszura autorstwa Izabeli Moszczyńskiej zatytułowana *Czego nie wiemy o naszych synach: fakta i cyfry dla użytku rodziców*. Z kolei w Krakowie Roman Rybacki zainteresował się przyczynami braków odpowiedzi wśród studentów, którzy w formularzach przyjęcia na studia nie wpisali pochodzenia społecznego. Przeprowadzone przezeń badanie wykazało, że odmowy podania pozycji społecznej były istotnie częstsze u studentów pochodzących z wyższych warstw społecznych. W okresie międzywojennym badanie takie przeprowadził działacz socjalistyczny Antoni Zdanowski - interesowało go, ile osób sypia w jednym łóżku. Na początku lat dwudziestych XX wieku utworzono Instytut Gospodarstwa Społecznego, który od 1926 roku, zajmując się badaniem problemów społecznych (warunków bytowania ludności, przeludnienia wsi, bezrobocia), wypracował swoistą metodologię studiów przypadku, w której skład wchodziły także badania ankietowe. Z Instytutem związani byli wybitni polscy uczeni między innymi Ludwik Krzywicki, Ludwik M. Landau, Stefan Z. Czarnowski. Najczęściej przytaczanym przykładem badań okresu międzywojennego były przeprowadzone przez wybitnego polskiego socjologa, folklorystę i historyka kultury S.Z. Czarnowskiego badania na wyczerpującej próbie 44 tysięcy polskich studentów. Jednakże o wynikach tych badań nic nie wiadomo, bowiem ankiety zostały ukryte w podwarszawskiej stodole, która spłonęła w 1939 roku od zrzuconej niemieckiej bomby. Po II wojnie światowej badania ilościowe (ankietowe) stały się popularne dopiero po 1956 roku, szczególnie w latach 1948-1956 nie prowadzono badań sondażowych[53]. Zaprzestano szerokiego prowadzenia badań sondażowych po opublikowaniu w 1962 roku artykułu *Ankietomania w polityce* autorstwa Adama Schaffa. W latach siedemdziesiątych badania ilościowe w Polsce odradzają się na nowo - powołano Instytut Podstawowych Problemów Marksizmu-Leninizmu (IPPM-L) zwany przez pracowników naukowych i absolwentów pieszczotliwie *Marleną*. W 1982 roku z inicjatywy pułkownika Stanisława Kwiatkowskiego, doradcy generała Wojciecha Jaruzelskiego powołano Centrum Badania Opinii Społecznej, którego głównym zadaniem było prowadzenie publicznych badań społecznych. Warto zwrócić uwagę na polityczną funkcję tych przedsięwzięć: w latach siedemdziesiątych władza państwowa wyznaczyła sondażom opinii zadanie ostrzegania przed niekorzystnymi fluktuacjami nastrojów społecznych, a w latach osiemdziesiątych miały one wspomagać propagandę państwową. Pokładane w nich nadzieje okazały się płonne[54]. Po 1989 roku rynek badawczy w Polsce zaczął dynamicznie rozwijać się. Szacuje się, że w ciągu dwóch dekad trwania III Rzeczypospolitej przebadano ponad 100 milionów respondentów. W samym tylko 2010 roku przebadano ich ponad 7 milionów. Obecnie najpopularniejszą techniką zbierania danych ilościowych jest metoda telefonicznych wywiadów wspomaganych komputerowo - w 2010 roku jej udział w rynku wynosił 33,4 proc.

[52] J. Kordos, *Professor Jerzy Neyman - Some Reflections*, „Lithuanian Journal of Statistics", 2011, 50 (1), s. 117. Elektroniczna wersja wzmiankowanej pracy J.S. Neymana znajduje się w zasobach GUS: http://statlibr.stat.gov.pl/exlibris/aleph/a18_1/apache_media/EJMR2E4131GDGULJXFCY13MAY4AFKF.pdf, dostęp: październik 2012.

[53] M. Witaszek, *Miejsce i rola sondaży w badaniu opinii społecznej*, „Zeszyty Naukowe Akademii Marynarki Wojennej", 2007, 4 (171), s. 143.

[54] Tamże, s. 143-144.

Popularne są również badania przez Internet (18,3 proc. rynku) oraz wywiady *face to face* w domach realizowane ze wspomaganiem komputerowym (16,5 proc.)[55].

W XX i XXI wieku rozwinięto, zaczęto powszechnie stosować i nauczać na poziomie akademickim licznych metod służących do analiz statystycznych - analizy wariancji, korelacji, regresji, analizy dyskryminacyjnej i analizy szeregów czasowych oraz analizy kanonicznej. Statystyka stała się nie tylko narzędziem wspomagającym, ale niezbędnym warunkiem niemal każdej działalności ludzkiej. Powszechnie stosuje się ją już nie tylko na potrzeby administracyjne i polityki państwa, ale również w ekonomii, finansach i przemyśle. Wyodrębniły się liczne dyscypliny, których trzonem są metody statystyczne: biometria, demografia, ekonometria, fizyka statystyczna, termodynamika statystyczna, psychometria, socjologia statystyczna i statystyka gospodarcza. Nowy rozdział w historii statystyki otwiera upowszechnienie się komputerów i oprogramowania służącego do obliczeń statystycznych. Rewolucja informatyczna, umasowienie komputerów spowodowało, że dane liczbowe można uzyskiwać i przetwarzać znacznie łatwiej, taniej i szybciej niż kiedykolwiek w historii ludzkości. Ponadto powstają nowe techniki analiz danych, w tym również takie, które wcześniej byłyby niemożliwe do zastosowania ze względów praktycznych - złożoności obliczeń i czasu koniecznego na przeprowadzenie obliczeń (na przykład drążenie danych - *data mining*). Otwierają się nieobecne dotąd perspektywy gromadzenia i przetwarzania danych, a w efekcie uzyskiwania informacji i generowania wiedzy na ich podstawie.

[55] *PTBRiO, katalog 2011/12, Edycja XVI,* Polskie Towarzystwo Badaczy Rynku i Opinii, s. 44-45.

2

Rozdział 2. Analiza danych ilościowych jako część procesu badawczego

Dane z zakresu nauk społecznych, w tym i politologii, poddawane analizom statystycznym są zbierane najczęściej w toku standaryzowanego procesu badawczego. Proces ten ma swoją ugruntowaną metodologię. Dane, które zostały zebrane z jej naruszeniem, nie nadają się do analiz statystycznych lub nadają się tylko w ograniczonym, niewystarczającym stopniu. W niniejszym rozdziale omówiono podstawy metodologiczne takiego procesu badawczego oraz nakreślono schemat, według którego najczęściej jest on przeprowadzany. W pierwszej kolejności wyłożono listę konstytutywnych cech badań ilościowych w naukach społecznych, a następnie - algorytm przebiegu badania ilościowego ze szczególnym uwzględnieniem roli analityka.

2.1. Istota badań ilościowych

Większość danych, które są następnie poddawane analizom statystycznym jest w naukach społecznych gromadzona w toku badań ilościowych. Badania te rozumiane jako gromadzenie danych liczbowych według pewnego schematu odbywały się, jak opisano w rozdziale 1, już w starożytności. Jednakże dopiero rozwój metod statystycznych w ciągu ostatnich trzystu lat, a w szczególności w XIX i XX wieku oraz intensywna praktyka badawcza w XX wieku wniosły do badań ilościowych kluczowy wkład techniczny i metodologiczny, nadając im ostateczną formę.

Metodologiczny rozwój badań ilościowych wiąże się z nurtem pozytywistycznym, powstałym na przełomie XIX i XX wieku empiriokrytycyzmem (tzw. „drugim pozytywizmem") i zapoczątkowanym na gruncie nauk społecznych przez „ojca" socjologii Augusta Comte'a. Jego głównymi przedstawicielami byli Ernst Mach i Richard Avenarius, na czele z pozytywizmem logicznym (neopozytywizmem) najsilniej wyrażanym w Kole Wiedeńskim oraz jego następcą - logicznym empiryzmem. Wkład o charakterze technicznym wniosła przede wszystkim praktyka prowadzenia spisów powszechnych, a następnie i przede wszystkim badania prowadzone na części zbiorowości - próbach. Za patriarchów praktyki badawczej w ostatnim

z wymienionych zakresów uznać należy George'a Gallupa, który w 1935 roku założył obecnie chyba najbardziej znaną agencję badawczą na świecie - Instytut Gallupa (dawniej noszący nazwę American Institute of Public Opinion) oraz amerykańskiego socjologa austriackiego pochodzenia Paula F. Lazarsfelda.

Badania ilościowe są zatem nierozerwalnie związane z naturalistycznym (esencjalistycznym) stanowiskiem ontologicznym[1] oraz neopozytywistycznym stanowiskiem epistemologicznym[2]. Naturalizm ontologiczny zakłada, że istnieje rzeczywistość obiektywna, to jest niezależna od naszej wiedzy od niej, a świat nie jest li tylko konstruktem społecznym. Z kolei neopozytywizm w zakresie epistemologii oznacza, że owa obiektywna rzeczywistość jest poznawalna (przynajmniej częściowo), dostępna za pomocą narzędzi, w które wyposażony jest badacz. Celem metod ilościowych jest opis i wyjaśnienie badanych zjawisk społecznych. Wyjaśnianie jest swoistym sposobem poznawania rzeczywistości społecznej związanym z poszukiwaniem praw i regularności, ujmowaniem jej w kategoriach przyczynowo-skutkowych, a nie jego rozumienia, czyli pytania o jej sens. Stosowanie metod ilościowych implikuje całkowite lub przynajmniej częściowe uznanie postulatu głoszącego, że nauki społeczne można uprawiać na wzór nauk przyrodniczych.

Rudymentarne pytanie badawcze zadawane w badaniach ilościowych dotyczy częstości występowania określonych zjawisk w badanej grupie. Przedmiotem zainteresowania badacza są stosunki ilościowe badanych zjawisk, zależności pomiędzy zmiennymi oraz różnice między badanymi grupami pod względem ilościowym. W badaniach ilościowych narzędziem badawczym służącym do gromadzenia danych jest standaryzowany kwestionariusz lub ankieta. Standaryzacja oznacza jednolity dla każdej jednostki analizy układ kwestionariusza lub ankiety pod względem treści i kolejności pytań. Oznacza także formułowanie pytań w kwestionariuszu lub ankiecie w sposób rozstrzygający, a więc poprzez tworzenie zamkniętych list odpowiedzi na zadawane pytania. Dane zbierane są w toku interakcji, podczas której następuje analogiczna dla wszystkich badanych jednostek analizy sekwencja „pytanie-odpowiedź". Najczęściej lista możliwych odpowiedzi na pytanie jest z góry określona i prezentowana pod pytaniem jako tak zwana kafeteria.

W klasycznych badaniach ilościowych wykorzystuje się najczęściej dobór probabilistyczny, aby można było wnioskować o całej populacji na podstawie zbadanej próby. Możliwe jest wówczas stosowanie najbardziej wymagających miar analitycznych, ale również pozwalających na najbardziej pewne wnioskowanie statystyczne (za pomocą testów parametrycznych). Dobór probabilistyczny nie jest jednak wyznacznikiem - wykorzystywany jest również dobór nieprobabilistyczny umożliwiający stosowanie innych narzędzi analiz (testy nieparametryczne). Najczęściej zakłada się zbadanie liczby jednostek analizy umożliwiającą wnioskowanie na populację. Minimalne liczebności określa się w literaturze przedmiotu na 30 lub 120 jednostek analizy. Metody ilościowe stosowane są najczęściej na końcowym etapie rozwiązywania problemu badawczego. Do metod ilościowych należą takie, jak wywiad bezpośredni (*Face to Face Interview*) prowadzony z użyciem kwestionariusza papierowego (*Paper and Pencil Interviewing,* PAPI) lub wykorzystujący komputer bądź inne urządzenie mobilne (*Computer Assisted Personal Interviewing,* CAPI), wywiad telefoniczny wspomagany komputerowo (*Computer Assisted Telephone Interviews*, CATI), ankieta internetowa (*Computer Assisted Web Interviewing*, CAWI), rzadziej ankieta audytoryjna czy ankieta prasowa.

[1] W uproszczeniu ontologia (metafizyka) to dział filozofii badający strukturę rzeczywistości.

[2] Teoria poznania, jeden z działów filozofii badający relacje pomiędzy poznawaniem a rzeczywistością, rozważający istotę takich pojęć jak prawda.

2.2. Etapy badania ilościowego

Rola analityka danych ilościowych nie ogranicza się wyłącznie do wykonania obliczeń zebranych w badaniu danych, choć jest to zasadniczy etap jego działania. W mniejszym lub większym stopniu obecność analityka jest konieczna na każdym niemal etapie procesu badawczego. Oddajmy w tej sprawie głos Hubertowi M. Blalockowi, który stwierdza, że rozważania statystyczne pojawiają się zarówno na etapie analizy danych, jak i na początku, gdy opracowujemy plan analizy i losujemy próbę:

Badacz nie może zaplanować i przeprowadzić całego procesu badawczego bez wiedzy statystycznej, a następnie przekazać cały materiał statystykowi mówiąc <<Ja zrobiłem swoje. Teraz ty przeprowadź analizę>>. Jeśli tak postąpi wyniki badań będą wątpliwe, lub nawet zupełnie bezużyteczne. Oczywiste jest bowiem, że problemy, które pojawią się przy analizie danych, muszą być antycypowane we wszystkich poprzednich etapach procesu badawczego i w tym sensie statystyka przewija się przez całość badań[3].

W niniejszym podrozdziale przedstawiono autorską propozycję algorytmu procesu badawczego[4]. Dają się wyróżnić trzy zasadnicze, następujące po sobie fazy lub etapy każdego badania ilościowego: faza projektowania badania, faza realizacji badania oraz faza analizy wyników badania.

Każdy z wymienionych etapów rozgrywa się na czterech zależnych od siebie, lecz integralnych płaszczyznach: **formalno-prawnej** związanej z kwestiami finansowymi i prawnymi (podpisaniem umów z realizatorami badania, nadzorem finansowym i organizacyjnym nad procesem badawczym, rozliczeniem i wypłatą wynagrodzeń za wykonane badanie), **metodologicznej** - najistotniejszej, w ramach której zostaje zaprojektowany sposób i przebieg rozwiązywania postawionego problemu badawczego, **praktycznej** - w ramach której przeprowadzane są wszelkie czynności realizacyjne (na przykład jest to druk kwestionariuszy wywiadu, szkolenie ankieterów, następnie realizacja wywiadów przez ankieterów oraz sprawowanie nad nimi nadzoru, proces przetwarzania danych z otrzymanych od ankieterów wypełnionych kwestionariuszy wywiadu) oraz **analitycznej** - która obejmuje statystyczne wspomaganie rozwiązywania problemu badawczego.

Skrzyżowanie wymienionych etapów z płaszczyznami stanowi pełny, wieloaspektowy algorytm procesu badawczego. Został on przedstawiony w tabeli 1 w postaci złożonej macierzy. Najmocniej płaszczyzna analityczna wiąże się z płaszczyzną metodologiczną, w znacznie mniejszym stopniu - z płaszczyzną praktyczną, a funkcjonuje niemal niezależnie od płaszczyzny formalno-prawnej. Omówmy rolę analizy danych na każdym z trzech etapów procesu badawczego: przygotowania, przeprowadzenia zbierania danych, badania oraz ich analiz.

[3] H.M. Blalock, *Statystyka dla socjologów*, Państwowe Wydawnictwo Naukowe, Warszawa 1977, s. 17.

[4] Zaproponowany algorytm procesu badawczego stanowi twórcze rozwinięcie znakomitego schematu opracowanego przez Grzegorza Babińskiego w latach osiemdziesiątych XX wieku: G. Babiński, *Wybrane zagadnienia z metodologii socjologicznych badań empirycznych*, Uniwersytet Jagielloński, Kraków 1980, Rozdział II. *Etapy procesu badawczego,* s. 19-34.

Tabela 1. Algorytm procesu badawczego

	1. Faza przygotowania badania							2. Faza realizacji badania	3. Faza analizy wyników badania							
I. Aspekt metodologiczny	a. Sformułowanie problemu badawczego	b. Eksplikacja problemu badawczego	c. Operacjonalizacja problemu badawczego	d. Przygotowanie narzędzia badawczego	e. Pilotaż badania	f. Dobór próby	g. Konsultacje na potrzeby praktyczne (szkolenia)	h. Konsultacje na potrzeby praktyczne	i. Konsultacje na potrzeby analityczne							j. Raport z badania
II. Aspekt analityczny	a. Konsultacje na potrzeby metodologiczne					b. Testowanie trafności i rzetelności narzędzia badawczego	c. Konsultacje na potrzeby praktyczne (szkolenia)	d. Obliczenia cząstkowych wyników na potrzeby kontroli procesu badawczego oraz na potrzeby zamawiającego bada-nie\kierującego badaniem	e/f. Tworzenie bazy danych	f. Weryfikacja danych	g. Kodowanie	h. Agregacja danych (rekodowanie danych)	i. Ważenie danych	j. Tabulacja danych	k. Testowanie hipotez / zaawansowane analizy statystyczne	j. Raport z badania
III. Aspekt praktyczny	a. Rekrutacja ankieterów, kontrolerów, przygotowanie infrastruktury			d. Sporządzenie materiałów szkoleniowych dla ankieterów oraz materiałów szkoleniowych i wytycznych dla kontrolerów w procesie badania		c. Szkolenie kontrolerów	d. Szkolenie ankieterów	e. Kontrola ad hoc ankieterów, kontrola procesu badawczego i postępów realizacji próby (kontrola ilościowa i jakościowa ad hoc)	g. Konsultacje na potrzeby analityczne							h. Konsultacje na potrzeby przygotowania raportu
IV. Aspekt formalny	a. Umowa, zlecenie	b. Nadzór finansowo-organizacyjny nad procesem badawczym														c. Protokół zdawczo-odbiorczy, faktura, rozliczenie projektu, wypłaty

Źródło: Opracowanie własne.

2.2.1. Faza przygotowania badania

W fazie projektowania badania należy kolejno podjąć następujące czynności: sformułować problem badawczy, dokonać jego eksplikacji oraz operacjonalizacji, skonstruować narzędzie badawcze, przeprowadzić pilotaż badania, a następnie dokonać doboru próby. Konieczne jest także sformułowanie wynikających z poprzednich etapów, przystępnych wniosków i wskazówek na potrzeby szkolenia. Wymienione czynności wymagają krótkiego wyjaśnienia. Pierwszym etapem każdej czynności badawczej jest sformułowanie problemu badawczego. Zazwyczaj czyni się to w formie pytania odnoszącego się do faktycznego (a nie subiektywnego dla badacza) stanu niewiedzy, najlepiej wyrażonego w języku naukowym, ponadto sformułowanego tak, by wiadomo było, jakie konkretne dalsze czynności należy podjąć, aby odnaleźć odpowiedź na to pytanie. Rozumienie problemu badawczego jako zespołu pytań, na które w toku dalszego postępowania badacz będzie się starał uzyskać odpowiedzi, jest powszechne w literaturze przedmiotu[5]. Kolejnym cząstkowym etapem procesu badawczego jest eksplikacja i operacjonalizacja problematyki badawczej. Stanowią one elementy wspomagające formułowanie problemu badawczego. Eksplikacja polega na uszczegółowieniu lub „ustawieniu" problemu badawczego, to jest modyfikacji tematu poprzez jego zawężenie. Ten cząstkowy etap wymaga od badacza sformułowania hipotez rozumianych jako propozycje twierdzenia naukowego. Z kolei operacjonalizacja to zabieg wyrażenia użytych pojęć w kategoriach operacyjnych, nadania im sensu empirycznego, a najlepiej przedstawienia listy konkretnych czynności, jakie należy podjąć, aby potwierdzić lub odrzucić postawione hipotezy. Na tym etapie należy również opracować wskaźniki występowania danego zjawiska, wybrać badaną zbiorowość oraz podjąć decyzję odnośnie metod i technik badawczych. Kolejny cząstkowy etap stanowi przygotowanie narzędzi badawczych. W badaniach ilościowych pod pojęciem narzędzia badawczego rozumie się kwestionariusz wywiadu lub ankietę, a także materiały instruktażowe ułatwiające gromadzenie i przetwarzanie zbieranych informacji ankieterom, osobom nadzorującym ankieterów, koderom i analitykom. Narzędzie badawcze jest podporządkowane wybranej podczas operacjonalizacji technice badawczej, dostosowane do specyfiki badanej zbiorowości, a także do warunków techniczno-organizacyjnych prowadzonych badań. Po zbudowaniu narzędzia badawczego należy przeprowadzić pilotaż badania, polegający na przetestowaniu i ewentualnym zmodyfikowaniu przyjętych rozstrzygnięć przed przystąpieniem do fazy realizacji badania. Obejmuje ona sprawdzenie metazałożeń prowadzonego badania, to jest poprawności jego eksplikacji i operacjonalizacji. Jest on praktycznym testem możliwości rozwiązania problemu badawczego za pomocą przyjętych metod, wykorzystanych technik i dostępnych narzędzi. Ponadto umożliwia szczegółowe, techniczne sprawdzenie narzędzia badawczego, w tym między innymi zweryfikowanie ilości braków danych oraz oceny wiarygodności uzyskiwanych odpowiedzi. Technicznie rzecz ujmując, pilotaż służy również zamknięciu pytań otwartych, a także określeniu czasu przeprowadzenia pojedynczego wywiadu. G. Babiński sugeruje, by procesowi pilotażu poddać co najmniej 5 procent założonej próby badawczej[6]. W praktyce badawczej, okazuje się jednak, że zupełnie wystarczające jest przeprowadzenie kilkunastu wywiadów pilotażowych. Kolejnym krokiem jest dobór próby badawczej, to jest określenie liczby i zasad doboru jednostek ze zbiorowości, którą chcemy zbadać. Dobór próby powinien nastąpić po przygotowaniu programu badań, w tym stworzeniu narzędzia badawczego. G. Babiński uważa, że warunek ten jest nader często bagatelizowany - bez należytej refleksji proponuje się liczne, reprezentatywne próby i uzyskuje

[5] G. Babiński, *Wybrane zagadnienia z metodologii socjologicznych badań empirycznych*, Uniwersytet Jagielloński, Kraków 1980, s. 19-23; T. Pilch, *Zasady badań pedagogicznych*, 1998 s. 25 Wydawnictwo "Żak"; J. Sztumski, *Wstęp do metod i technik badań społecznych*, Wydawnictwo Naukowe „Śląsk", Katowice 2010, s. 38.

[6] G. Babiński, dz. cyt., s. 28.

błędne wyniki lub też ponosi zbędne wydatki[7]. Ponadto nawet najlepsze przygotowanie badania pod względem metodologicznym nie zapewni sukcesu badaczowi, jeśli nie zostanie ono przełożone na praktyczne wskazówki dla wykonujących badanie ankieterów, koderów i analityków. Dlatego też metodologicznym wymogiem każdego projektu badawczego powinno być przygotowanie instrukcji i wskazówek na potrzeby szkolenia wymienionych grup uczestników procesu badawczego na podstawie dotychczasowych rozstrzygnięć i postanowień metodologicznych. Z punktu widzenia wymogów praktycznych na etapie przygotowywania projektu badawczego należy zapewnić wystarczająco wydolną infrastrukturę organizacyjną - przede wszystkim przeprowadzić rekrutację ankieterów i kontrolerów, wyszkolić ich i zorganizować ich pracę, przygotować odpowiednie zasoby organizacyjne i finansowe. W tle wymienionych procesów toczą się czynności formalno-prawne związane z ustaleniem wynagrodzeń i podpisaniem umów ze wszystkimi uczestnikami procesu badawczego, zleceniem im odpowiednich prac oraz wyznaczeniem terminów i zasad kontroli oraz prowadzeniem bieżącego monitorowania finansów prowadzonego projektu. Na tym etapie praca analityków powinna być podporządkowana pracy projektujących badanie pod względem metodologicznym. Sformułowanie problemu badawczego oraz jego eksplikacja i operacjonalizacja powinny się odbywać z udziałem analityków - pełnią oni na tych cząstkowych etapach rolę doradców - umożliwiając wybór optymalnych narzędzi statystycznych służących do rozwiązania problemu badawczego oraz dopasowanych do planowanej techniki badawczej. Analityk powinien również odgrywać kluczową rolę podczas pilotażu badania. Jest on wówczas odpowiedzialny za testowanie rzetelności indeksów i skal skonstruowanego narzędzia badawczego, prognozuje możliwość dokonania określonych obliczeń i zasadności zastosowania w tym kontekście narzędzi badawczych. Ponadto głos analityka powinien zostać uwzględniony podczas szkolenia ankieterów i kontrolerów. Istotne znaczenie ma tu jego opinia na temat sposobu zbierania danych w toku badań. Winien on uświadomić uczestnikom procesu badawczego przeznaczenie danych oraz wyjaśnić, jakie ich cechy będą miały kluczowe znaczenie dla dokonania pełnowartościowych analiz. Rola analityka na tym etapie procesu badawczego jest w praktyce badawczej nader często lekceważona, co skutkuje zerwaniem ciągłości procesu badawczego, brakiem przełożenia pomiędzy etapem projektowania badania a analizowaniem zebranych danych.

2.2.2. Faza realizacji badania

Fazę realizacji badania podejmuje się w celu zebrania określonych w poprzedniej fazie informacji o ściśle określonej formie i treści. Etap ten w praktyce badawczej nazywany jest fazą terenową realizacji badania lub z języka angielskiego - *fieldwork.* Wysiłek prowadzenia badania na tym etapie spoczywa na ankieterach oraz osobach kontrolujących ich pracę. Podczas fazy realizacji projektu badawczego największe znaczenie ma aspekt praktyczny, a wszystkie pozostałe są mu podporządkowane. Nie może być on jednak prowadzony w oderwaniu od elementów metodologicznych i analitycznych. Osoby odpowiedzialne za merytoryczne przygotowanie badania oraz przeprowadzenie analiz powinny kontrolować proces zbierania danych, służyć radą i pomocą. Ponadto analitykowi przypadać powinna tu rola aktywnego podmiotu - często dokonuje on bowiem cząstkowych obliczeń wyników na potrzeby kontroli procesu badawczego oraz na potrzeby kierującego badaniem. Ma to znaczenie nie tylko informacyjne - pozwala również korygować wszelkie niedociągnięcia, odstępstwa od normy, wykrywać artefakty, naruszenia liczebności założonej próby, wszelkie problemy realizacyjne, a w efekcie - kierować badanie z powrotem „na właściwy kurs". Istotnym elementem *fieldworku* powinien być nieprzerwany nadzór finansowy nad procesem badawczym nastawiony na optymalizację kosztów, zapobiegający przekraczaniu ustalonego budżetu projektu

[7] Tamże, s. 29.

badawczego. Jego wartość na tym etapie jest nie do przecenienia - właśnie ta część procesu badawczego jest najbardziej kosztowna, a jednocześnie najbardziej niepewna.

2.2.3. Faza analizy wyników badania

Po zrealizowaniu części terenowej badacze przystępują do przetwarzania, a następnie do analizy zebranych wyników. W naukach społecznych badane zjawiska nie są na ogół mierzone bezpośrednio tak, jak ma to miejsce w naukach przyrodniczych. Stąd konieczność kompleksowego i czasochłonnego przygotowania zbioru danych do analizy. W programach komputerowych, służących do analizy danych w naukach społecznych moduły dedykowane do przetwarzania danych na potrzeby przygotowawcze są bardzo rozwinięte - na ogół ta właśnie cecha odróżnia je od oprogramowania przeznaczonego na potrzeby nauk przyrodniczych. Przygotowanie zbioru danych do analizy obejmuje dwie fazy: rekonfigurację zbioru danych oraz przekształcenia zmiennych w zbiorze danych. Czynności te mają kluczowe znaczenie w analizie danych w naukach społecznych, w tym w politologii, wobec czego w niniejszej publikacji poświęcono im dwa obszerne rozdziały. Rekonfiguracja zbioru danych obejmuje takie czynności jak dodawanie, modyfikowanie i usuwanie zmiennych i jednostek analizy w zbiorze danych, sortowanie i agregowanie zbioru danych, dzielenie go na podgrupy oraz selekcję jednostek do analizy. Wszystkie te czynności wykonywane są podczas pierwszego i drugiego cząstkowego etapu analiz: przygotowania zbioru danych do analizy i jego weryfikacji.

Tworzenie zbioru danych z zebranych w toku badania informacji oznacza nadanie zebranym informacjom ściśle standaryzowanej, możliwej do poddania dalszym analizom struktury. Zbiór danych występuje obecnie wyłącznie w formie elektronicznej (bazy danych). Ten cząstkowy etap polega przede wszystkim na dokonaniu czynności o charakterze technicznym. Papierowe (choć coraz częściej elektroniczne) kwestionariusze wywiadu lub ankiety wymagają przetworzenia na formę elektroniczną. Na ogół wykorzystuje się do tego celu skaner oraz odpowiednie oprogramowanie. Rzadziej dokonuje się tworzenia bazy danych ręcznie - wpisując poszczególne kwestionariusze lub ankiety. Coraz częściej dane w fazie realizacji badania są zbierane w formie elektronicznej z użyciem komputerów, laptopów lub innych urządzeń mobilnych i od razu występują w formie gotowej bazy danych. W tym przypadku cząstkowy etap tworzenia bazy danych ogranicza się do jej zapisania w formacie umożliwiającym analizę danych w programie do tego służącym.

Kolejnym cząstkowym etapem realizacji projektu badawczego jest weryfikacja zebranego materiału empirycznego. Obejmuje ona takie czynności, jak sprawdzenie poziomu realizacji próby, to znaczy zweryfikowanie, czy zrealizowano wystarczającą liczbę wywiadów oraz jakie są zniekształcenia próby, a więc czy jakieś kategorie badanych reprezentowane są w zbyt małym lub w zbyt dużym stopniu. Na tym cząstkowym etapie powinno się także wyeliminować materiały niepełne lub nieprawidłowe - na przykład przerwane przez respondenta wywiady, niewypełnione całkowicie ankiety oraz takie materiały badawcze, co do których analityk poweźmie przekonanie, że są one z jakichś powodów niewiarygodne lub niepełnowartościowe. Ponadto należy określić, czy i które wskaźniki, indeksy lub skale miały wady wykluczające je z dalszego postępowania badawczego bądź nakazują traktować zebrane dane z daleko idącą ostrożnością, względnie wymagają podjęcia czynności naprawczych.

Kolejny etap przygotowania danych do analizy stanowią przekształcenia zmiennych w zbiorze danych - ich kodowanie, agregacja oraz ważenie. Celem kodowania jest takie spreparowanie danych, które umożliwia poddanie ich analizom statystycznym. Kodowaniu poddawane są odpowiedzi na pytania

otwarte i półotwarte, czyli odpowiedzi, które zostały zapisane w formie tekstowej, opisowej. W toku tej procedury zostają one przekształcone w odpowiedzi zamknięte, co jest warunkiem podjęcia wobec nich analizy ilościowej. Kolejny cząstkowy etap to agregacja zebranych danych nazywana również rekodowaniem. Funkcja rekodowania polega na zmianie wartości istniejących zmiennych w celu ich agregacji, uporządkowania, uspójnienia lub uproszczenia (zmiany poziomu pomiaru zmiennej). Z kolei ważenie danych jest to zabieg statystyczny, którego celem jest uczynienie struktury zbadanej grupy tożsamej ze strukturą całej populacji pod względem wybranych cech. W efekcie pragniemy, by zbadana grupa stała się pomniejszonym odbiciem proporcji cech występujących w populacji. Istota zabiegu ważenia polega na obniżeniu rangi grup respondentów nadreprezentowanych w zbiorze danych lub podniesieniu rangi grup respondentów niedoreprezentowanych. Wszystkie wymienione powyżej czynności stanowią domenę analizy danych i zostały szczegółowo opisane w dalszych częściach pracy. Mają one kluczowe znaczenie dla pracy analityka i bez podjęcia tych czynności analiza danych nie jest na ogół możliwa. Należy również zauważyć, że czynności te są o tyle istotne, że są bardziej czasochłonne niż sam proces analizy. Wymagają - podobnie jak sam proces analizy statystycznej - odpowiedniego przygotowania merytorycznego badacza. W praktyce badawczej często czynności weryfikacji danych, kodowania i rekodowania powierzane są szeregowym, początkującym, często przypadkowym pracownikom, ze względu na to, że jest to czynność żmudna i czasochłonna. Choć z formalnego punktu widzenia czynności te mogą wydawać się poprawnie wykonane, to ciągłość procesu badawczego ulega zaburzeniu lub nawet zerwaniu i w efekcie wykonana praca może być niedopasowana lub całkiem nie odpowiadać potrzebom rozwiązania problemu badawczego.

Właściwa analiza następuje w dwóch etapach: pierwszym - tabulacji danych i drugim - statystycznego testowania postawionych hipotez. Pierwszy etap, tabulacji danych ma charakter elementarny, służy rozpoznaniu uzyskanych wyników. Dokonywany jest w toku jednego z dwóch typów tabulacji prostej, gdy przedmiotem obliczeń częstości występowania jest jedna zmienna lub - tabulacji złożonej, gdy jej przedmiotem jest więcej niż jedna zmienna. Etap ten obejmuje także obliczenie miar tendencji centralnej i miar dyspersji. Warto przestrzec przed ograniczaniem się wyłącznie do tego typu analiz. Zarówno w działalności naukowej, jak też biznesowej opieranie się na tego typu analizach jest zbytnim uproszczeniem, nie pozwala na wyciągnięcie wystarczająco pogłębionych i wieloaspektowych wniosków. W drugim etapie wkraczamy w bardziej zaawansowane analizy statystyczne polegające na przykład na badaniu związków między zmiennymi, różnic pomiędzy badanymi grupami, przewidywaniu zależności.

Na etapie analiz przydatna lub nawet konieczna może być konsultacja z wykonawcami poprzednich faz procesu badawczego. Odwołanie się do ankieterów, osób kontrolujących ankieterów lub - nawet - do samych respondentów stanowić może bezcenne źródło pozwalające na właściwą ocenę zebranych danych - przede wszystkim ich spójności i wiarygodności. Standaryzowana forma kwestionariusza, możliwość wprowadzenia z góry przewidzianych informacji jest z jednej strony niezbędnym wymogiem wszystkich badań ilościowych, lecz z drugiej - niestety utrudnieniem - siłą rzeczy niedostępne stają się dla innych uczestników procesu badawczego cenne informacje, które mogą być kluczowe dla pełnego, wieloaspektowego zrozumienia zebranych danych. Opisanemu wyżej etapowi analiz poświęcona jest niniejsza publikacja, przeprowadza ona szczegółowo pod względem teoretycznym i praktycznym przez wszystkie wymienione cząstkowe etapy. Ukoronowaniem analiz jest raport badawczy opierający się w zasadniczej mierze na obliczeniach wykonanych przez analityków. Podstawy tworzenia raportu z badań również omówiono w niniejszej publikacji.

Proces badawczy wymaga zaangażowania wielu osób, dlatego warto sformułować postulat zachowania jego ciągłości, to jest nieprzerwanej obecności (co najmniej w postaci konsultantów i obserwatorów) wszystkich jego uczestników w trakcie realizacji badania[8]. Tylko wówczas wykonane analizy będą prawidłowe, jeśli wszystkie poprzednie fazy badania były poprawne pod względem merytorycznym i technicznym. W znacznym stopniu może do tego przyczynić się analityk z zainteresowaniem śledzący i wspomagający proces badawczy na wszystkich jego etapach.

[8] Liczne wskazówki na temat funkcjonowania i zarządzania efektywnością zespołów badawczych można odnaleźć w: W. Okrasa, *Funkcjonowanie i efektywność zespołów badawczych*, Zakład Narodowy im. Ossolińskich, Wrocław 1987.

3

Rozdział 3. Wstępna charakterystyka programu PSPP

W tym rozdziale znalazła się krótka specyfikacja programu PSPP - przedstawiono jego historię, formalno-prawne aspekty jego użytkowania oraz naszkicowano jego możliwości na tle rodziny programów służących do analiz statystycznych.

Wprowadzenie do pracy z programem warto zacząć od wyjaśnienia skąd pochodzi jego nazwa. Jest ona nieprzypadkowa i wyraźnie wskazuje na istotne z punktu widzenia użytkownika własności tego programu. Nazwa PSPP nawiązuje do nazwy SPSS (*Statistical Package for Social Sciences*) i stanowi jej quasi-anagram[1]. Program SPSS jest jednym z najpowszechniej używanych narzędzi służących do analizy danych ilościowych w zakresie nauk społecznych. Jest on programem komercyjnym[2]. Natomiast PSPP stanowi darmową i dostępną dla wszystkich zainteresowanych alternatywę dla programu SPSS. Nazwa PSPP będąca quasi-anagramem ma podkreślać względną tożsamość i względną odrębność obu narzędzi. Z jednej strony program PSPP odzwierciedla większość istotnych i użytecznych funkcji SPSS. Jest jego swoistym „klonem" w zakresie statystycznym (zawiera większość tych samych testów), mechanicznym (wyglądu interfejsu graficznego oraz sposobu obsługi), a także informatycznym (języka poleceń). Znajomość jednego z programów umożliwia swobodne korzystanie z drugiego, a drobne różnice pomiędzy nimi nie wpływają istotnie na komfort i jakość pracy.

Program PSPP zawiera podstawowe narzędzia służące do rekonfiguracji zbiorów danych oraz ich analizy statystycznej. Możliwe jest także wykonywanie w tym programie prostych grafik prezentacyjnych - wykresów oraz map. Z praktycznego punktu widzenia program ten zawiera wszelkie funkcje pozwalające na samodzielne analizy badań własnych do prac promocyjnych (licencjackich, magisterskich, a nawet doktorskich), a także analizy dostępnych danych statystycznych. Program został napisany w języku C, co zapewnia względną szybkość jego działania. PSPP może być używany zarówno w linii poleceń

[1] Anagram powstaje przez przestawienie szyku liter w wyrazie z zachowaniem pierwotnej liczby liter; anagram niepełny (quasi-anagram) jest pojęciem szerszym, bowiem dopuszcza użycie nie wszystkich liter lub dodanie innych.

[2] Szerzej na temat programu SPSS w Aneksie 2.

(*Command-line Interface*, CLI) jak również w trybie okienkowym (*Graphic User Interface*). Aplikacja ta ma charakter uniwersalny - może być zainstalowana i uruchamiana na komputerach z różnymi systemami operacyjnymi: Debian, Fedora, FreeBSD, GNewSense, HP-UX, Mac OS X, OpenSUSE, Ubuntu oraz Windows. Dla systemu operacyjnego Windows istnieją wersje programu PSPP zarówno jedno- jak też wielostanowiskowe, a także wersje dla systemu Windows 32- i 64-bitowego, jak też wyłącznie 64-bitowego. Obecna wersja programu PSPP to 0.7.9 (ta właśnie edycja wykorzystana została w publikacji)[3].

3.1. Historia programu PSPP

Pomysłodawcą i programistą programu PSPP jest Ben Pfaff amerykański informatyk, absolwent wiodących w światowych rankingach uczelni - Uniwersytetu Stanforda oraz Uniwersytetu Michigan. B. Pfaff prace nad projektem PSPP rozpoczął w 1992 roku jako czternastolatek. Pobudki podjęcia trudu stworzenia programu stanowiącego, jak sam twórca wskazuje „klon programu SPSS", były przede wszystkim natury ideowej. Dominującym motywem B. Pfaffa - była i jest - przede wszystkim negatywna moralna ocena przedsiębiorstw udostępniających oprogramowanie na zasadach komercyjnych oraz chęć przyczynienia się do dobra wspólnego dzięki posiadanym umiejętnościom programistycznym. Przesłankami podjęcia wysiłku był również dostęp B. Pfaffa do programu SPSS, a także fakt, że oprogramowanie statystyczne znalazło się w owym czasie na liście zapotrzebowań Free Software Foundation. B. Pfaff pierwotnie nazwał swój projekt Fiasco, jednak zrezygnował wkrótce z tej nazwy, gdy okazało się, że autor innego projektu jako pierwszy użył tej nazwy[4]. Nazwa Fiasco miała być przewrotnym żartem - można ją interpretować jako wyzwanie rzucone komercyjnemu oprogramowaniu SPSS, który poniósł fiasko moralne, a wskutek rozwijania i coraz szerszego stosowania programu PSPP poniesie również fiasko finansowe.

B. Pfaff jest autorem licznych narzędzi informatycznych usprawniających multimedia w systemie Linux oraz interesujących narzędzi sieciowych (między innymi służących do programistycznej pracy zespołowej). Opublikował wiele artykułów z zakresu informatyki oraz skryptów na temat składu tekstu i programowania. Jest także projektantem i administratorem stron internetowych. Jest on członkiem Electronic Frontier Foundation, wspiera projekty GNU oraz pracuje nad rozwojem darmowego systemu operacyjnego Linux[5]. B. Pfaff ukończył Michigan University w 2001 roku jako inżynier elektryk, a w 2007 roku Uniwersytet Stanforda, uzyskawszy stopień doktora nauk informatycznych na podstawie

[3] Należy zauważyć, że program PSPP w wersji 0.7.9. dostępny jest w postaci tak zwanych binariów (do bezpośredniego zainstalowania) jedynie dla systemu Windows, dla pozostałych systemów jest to wersja 0.6.x. Wykorzystanie tej najnowszej wersji w komputerach pracujących pod kontrolą systemów operacyjnych innych niż Windows jest możliwe pod warunkiem, że podejmiemy dodatkowe czynności - na przykład skompilujemy samodzielnie kod źródłowy programu specjalnie dla używanego systemu lub (co prostsze) użyjemy w danym systemie programu symulującego środowisko Windows (np. darmowego programu Wine).

[4] Przywoływane informacje uzyskano w toku korespondencji z B. Pfaffem (*via* e-mail) w lutym 2012 roku.

[5] B. Pfaff, *prywatna strona internetowa*, w: http://benpfaff.org/, dostęp: luty 2012.

dysertacji *Improving Virtual Hardware Interfaces*[6]. Od 2007 roku pracuje nad oprogramowaniem sieci Ethernet - Open vSwitch[7].

Od momentu ukazania się w sierpniu 1998 roku pierwszej wersji programu PSPP 0.1.0. prace nad nim posuwały się w ciągu dwóch minionych dekad w zmiennym tempie. Za najsłabszy należy uznać okres od 2000 do 2004 roku, kiedy projekt rozwijał się najwolniej. Prace przyspieszyły w 2007 roku, a w latach 2010-2011 opublikowano aż pięć stabilnych wersji programu (dla porównania: w latach 1998-2009 opublikowano ich zaledwie siedem). Program PSPP jest obecnie przedsięwzięciem zespołowym, pracują nad nim programiści z całego świata. Od samego początku najbardziej aktywnym programistą jest pomysłodawca projektu B. Pfaff. Drugim, aktywnym współtwórcą programu jest programista-wolontariusz dr John Darrington - Brytyjczyk, absolwent Murdoch University i University of Western Australia. Jest on autorem ważnych statystycznych analiz w programie PSPP - między innymi testu t-Studenta oraz jednoczynnikowej analizy wariancji ANOVA oraz tłumaczenia programu na brytyjski angielski, a także twórcą graficznego interfejsu (okienek) tego programu. Wśród programistów należy także wspomnieć Jasona Stovera (wcześniej bardzo aktywnego, obecnie - okazjonalnie), który zaimplementował do programu funkcję regresji oraz Mehmeta Hakana Satmana - autora algorytmu centroidów (*K-Means*). Wśród przeszłych współpracowników można wymienić Johna Williamsa, który zaimplementował podstawy testu t-Studenta oraz Michaela Kiefte i Patricka Kobly'ego którzy usunęli liczne błędy i nieścisłości w programie, a także Roba van Sona, który wdrożył istotne elementy testów nieparametrycznych. Warto także wspomnieć o licznych tłumaczach, dzięki którym program PSPP dostępny jest w ich językach narodowych: Harry Thijssen (holenderski), Palmira Payá Sanchez, Javier Gómez Serrano, Francesco Josep Miguel Quesada (kataloński i hiszpański), Michel Almada de Castro Boaventura (brazylijski i portugalski), Eric Thivant (francuski) i Mindaugas Baranauskas (litewski). Postępy prac nad projektem można śledzić na bieżąco, właściwie każdego dnia pojawiają się nowe wpisy prezentujące kolejne zmiany[8].

Jak deklaruje B. Pfaff - stara się nad programem PSPP pracować co najmniej godzinę dziennie. W lutym 2012 roku wskazał, że najbliższe plany pracy nad rozwojem aplikacji dotyczą jądra programu, przede wszystkim poprawy funkcjonowania okna raportów programu, wyglądu interfejsu graficznego, a także importu i eksportu plików. A także - *last but not least* - zamierza stworzyć ułatwienia dla innych programistów pozwalające dodawać im w funkcjonalnym i ergonomicznym trybie testy statystyczne[9].

Warto zwrócić uwagę Czytelników na fakt, że program PSPP jako projekt otwarty umożliwia przyłączanie się doń programistów, którzy chcieliby rozwijać go, poprawiać wybrane przez siebie charakterystyki programu lub dodawać nowe. Osoby zainteresowane nie będące programistami mogą z kolei przyczyniać się do rozwoju programu za pośrednictwem wpłat lub wynajmując i opłacając informatyków, którzy wdrożą wskazane rozwiązania. Można również (jeśli nie dysponuje się umiejętnościami programistycznymi ani zasobami finansowymi) zamieścić swoje sugestie na mailingowej „liście życzeń", licząc na łut szczęścia, że wdrożenia wskazanej przez nas funkcjonalności podejmie się jakiś programista-wolontariusz (wzmiankowana lista znajduje się pod adresem: http://sv.gnu.org/p/pspp).

[6] B. Pfaff, *Improving Virtual Hardware Interfaces*, w: http://benpfaff.org/papers/thesis.pdf, dostęp: kwiecień 2012.

[7] Oprogramowanie autorstwa B. Pfaffa można pobrać ze strony internetowej: http://openvswitch.org, dostęp: kwiecień 2012.

[8] Patrz: http://git.savannah.gnu.org/gitweb/?p=pspp.git;a=shortlog;h=refs/heads/master, dostęp: kwiecień 2012.

[9] Przytaczne informacje uzyskano w toku korespondencji z B. Pfaffem (*via* e-mail) w lutym 2012 roku.

3.2. Formalno-prawne aspekty użytkowania programu PSPP

Program PSPP jest produktem wolnym, co wynika z pełnej nazwy programu - GNU PSPP. Zastosowana gra słów w nazwie programu stanowi swoisty powszechnik subkultury Internetu (konkretnie subkultury hakerskiej) i jednocześnie identyfikator przynależności programu do grupy tak zwanego Wolnego Oprogramowania[10]. Jakakolwiek wzmianka na temat Wolnego Oprogramowania musi rozpoczynać się od przybliżenia postaci Richarda Matthew Stallmana. W środowisku informatycznym jest to postać pierwszoplanowa i natychmiast rozpoznawana. R.M. Stallman jest bowiem jednym z pierwszych hakerów - można wręcz rzec, że stanowi archetyp hakera. Urodzony w 1953 roku amerykański informatyk jest autorem powszechnie znanych, używanych i cenionych przez środowisko informatyków programów takich jak Emacs, GNU Compiler Collection czy GNU Debugger. Przede wszystkim jednak jest aktywnym działaczem, pomysłodawcą i założycielem organizacji działającej na rzecz wolności intelektualnej - Fundacji Wolnego Oprogramowania (Free Software Foundation). Trudno w kilku zdaniach streścić całość poglądów R.S. Stallmana, nie unikając przy tym uproszczeń czy powierzchownej interpretacji jego przekonań. Przede wszystkim należy podkreślić, że R.S. Stallman uważa, że oprogramowanie jako twór niematerialny, składający się z algorytmów matematycznych, należy uznać za dobro powszechne całej ludzkości. Według tego działacza oprogramowanie z punktu widzenia logiki, funkcjonalności i społecznego przeznaczenia należy traktować tak jak podstawowe twierdzenia fizyki teoretycznej lub matematyki. Dlatego programy nie powinny podlegać ograniczeniom patentowym - nikt nie ma prawa zawłaszczać tego dorobku intelektualnego w szczególności ze względów komercyjnych. Oprogramowanie jako byt wirtualny nie może podlegać mechanicznie zasadom prawa własności obiektów fizycznych. Traktowanie jako tożsamych produktów materialnych i niematerialnych jest logicznym, moralnym i prawnym nieporozumieniem, bowiem byty te nie podlegają takim fizycznym uwarunkowaniom jak przedmioty materialne - na przykład ich krańcowy koszt powielenia równy jest zero. Według R.S. Stallmana mamy do czynienia ze ścieraniem się dwóch sił - ładu, w którym dominuje imperatyw moralny oraz porządku, gdzie przeważa motyw ekonomiczny (lecz niepraktyczny, bo krótkookresowy i przynoszący zysk niewielkiej liczbie podmiotów, kosztem dużej liczby podmiotów). Wprowadzone bariery rozpowszechniania myśli informatycznej i jej efektów - przyjmujących postać programów - generują, zdaniem R.M. Stallmana, negatywne zjawiska społeczno-ekonomiczne: obniża się poziom wykształcenia ogółu ze względu na spowolniony lub zatrzymany obieg informacji, hamowany jest rozwój myśli - głównie technologii, ale również kultury, w efekcie wychładzana jest gospodarka. Ponadto patentowanie myśli programistycznej, wiązanie licencjami rozmaitych programów użytkowych jest uderzeniem w podstawową demokratyczną zasadę, jaką jest wolność słowa. Wolne oprogramowanie nie jest koncepcją marketingową lub biznesową, ale ideologią moralną[11]. *Ergo* - stwierdza Piotr Gawrysiak, wykładając poglądy R.S. Stallmana i jego zwolenników - pobieranie opłat za oprogramowanie jest praktyką złą i naganną, bowiem imperatyw ekonomiczny nie powinien znosić imperatywu moralnego, a praktyki firm informatycznych są negatywnym rozstrzygnięciem ważkiej, postawionej przez Ericha Fromma kwestii: mieć czy być[12].

[10] Wolne Oprogramowanie oznaczane jest nazwami wykorzystującymi tak zwane akronimy rekurencyjne. Najbardziej znanym tego typu programem (właściwie systemem operacyjnym) jest GNU Linux. Jego rozwinięty akronim rekurencyjny brzmi: Linux Is Not Unix, a więc Linux nie jest Unixem – programem komercyjnym.

[11] P. Gawrysiak, *Cyfrowa rewolucja. Rozwój cywilizacji informacyjnej*, Wydawnictwo Naukowe PWN, Warszawa 2008, s. 355.

[12] Tamże, s. 355.

Patentowanie kodu programów doprowadziło zdaniem pomysłodawcy FSF do sytuacji patologicznej - jest ich już tak wiele, że współcześnie nie sposób napisać programu bez naruszania (świadomie lub nieświadomie) prawa własności. W związku z tym każda z firm pieczołowicie zbiera patenty po to, by kreować swoistą równowagę strachu (na zasadzie: ja naruszam patent, lecz sam posiadam patenty, które z kolei naruszają ci, których patenty ja naruszam). W efekcie obowiązującego prawa każdy z napisanych programów staje się dowodem przestępstwa. Skrajność poglądów w kwestii wolnego oprogramowania oraz konsekwencja w działaniu sprawiły, że sympatycy utytułowali R.S. Stallmana przydomkiem Świętego Ignucego (od GNU).

Poglądy społeczne i filozoficzne R.M. Stallmana doprowadziły do pragmatycznych efektów - w 1984 roku sformułowana została koncepcja Wolnego Oprogramowania (*Free Software*). Wolność (*free*) w nazwie postulowana przez R.S. Stallmana może być rozumiana na dwa sposoby. Po pierwsze, odpowiada ono pojęciu „darmowe". Po drugie, oznacza niepodleganie licencjonowaniu ani patentom, dlatego można z programu korzystać bez nakładanych przez prawo ograniczeń. Często w celu wytłumaczenia tych dwóch rodzajów wolności zwolennicy Wolnego Oprogramowania posługują się sformułowaniem: *free as in freedom and free as in beer*. Wolne oprogramowanie oznacza swobodę każdego użytkownika dotyczącą wykorzystania programu do dowolnych zastosowań (tak zwana „wolność 0"), a także do nieodpłatnego rozpowszechniania kopii programu („wolność 2") oraz jego analizowania („wolność 1") i modyfikowania („wolność 3"), co oznacza dostępność kodu źródłowego takiej aplikacji. Kategoria wolnego oprogramowania została sformalizowana przez powstałą w 1985 roku dzięki Richardowi Matthew Stallmanowi Fundację Wolnego Oprogramowania (FSF, Free Software Foundation, Inc.). Efektem tej formalizacji jest licencja GNU[13] General Public License (GNU/GPL). GNU stanowi najbardziej rozpowszechniony typ licencji GPL (czyli - Powszechnej Licencji Publicznej)[14]. Obowiązuje ona obecnie w wersji trzeciej opublikowanej 29 czerwca 2007 roku. Obejmuje ona cztery wyżej wymienione podstawowe wolności.

Ważnym pojęciem związanym z tą licencją jest pojęcie *copyleft*. Nazwa *copyleft* jest to swoista gra słowna mająca oznaczać jednocześnie antonim słowa *copyright* (często używa się znaku *copyleft* z dopiskiem *All rights reversed,* a więc „wszystkie prawa odwrócone" - zamiast oryginalnego dla *copyright* - *All rights reserved* („Wszystkie prawa zastrzeżone"). Jednocześnie *left* oznacza drugą i trzecią formę czasownika nieregularnego *leave* oznaczającego „porzucić". Taki źródłosłów wyraźnie daje sygnał odnośnie dowolności wykorzystania programu PSPP.

Warto podkreślić, że idea Wolnego Oprogramowania zyskała sobie szerokie grono zwolenników, którzy stworzyli i nadal tworzą liczne programy. Jakość i komfort działania tych programów nie ustępują komercyjnym, a nawet je przewyższają. Do szczególnie udanych przedsięwzięć wolnego oprogramowania należy zaklasyfikować liczne odmiany systemów operacyjnych GNU Linux (między innymi Debian, Fedora, Slackware, Ubuntu, polskie i spolszczone systemy jak PLD, Mint Remix, a wcześniej Aurox), oprogramowanie biurowe (na przykład Open Office, Libre Office, a także arkusz kalkulacyjny Gnumeric), czy potężne oprogramowanie analityczne (R, Gate, Weka czy Rapid Miner). Działalność internetowej społeczności (choć z punktu widzenia terminologii socjologicznej raczej należałoby mówić o ruchu

[13] Tu także mamy do czynienia z akronimem rekurencyjnym, bowiem akronim GNU rozwijamy jako GNU Is Not Unix. Sięgamy tutaj do początków idei Wolnego Oprogramowania, która zrodziła się jako sprzeciw pierwszego pokolenia hakerów przeciwko komercyjnemu systemowi operacyjnemu Unix.

[14] Należy zauważyć, że istnieją liczne licencje „wolnościowe" - zarówno zgodne z GPL jak i niezgodne z nią. Ponadto istnieją rozmaite rodzaje licencji komercyjnych, które nie są wolne. Ich listę wraz z opisem dostarcza FSF pod następującym adresem http://www.gnu.org/licenses/license-list.

społecznym) nie ogranicza się jedynie do projektów *stricte* programistycznych. Ruchowi temu zawdzięczamy między innymi największą sieciową, globalną encyklopedię Wikipedię[15] oraz „projekty wolnej kultury" - biblioteki cyfrowe. Największym tego typu przedsięwzięciem jest Projekt Gutenberg, który według danych z roku 2011 udostępnił ponad 36 000 wolnych książek elektronicznych (http://www.gutenberg.org/). Warto wspomnieć o wzorowanym na wyżej wymienionym projekcie wolnodostępnego repozytorium dzieł literatury i sztuki nordyckiej - Projekcie Runeberg http://runeberg.org/)[16], gdzie publikowane są wyłącznie teksty i ilustracje starsze niż 70 lat (liczone nie od daty ich powstania, lecz od śmierci twórcy), a więc niechronione już prawem autorskim majątkowym. Na uwagę zasługuje również repozytorium żydowskich dzieł literackich - projekt ben Jehuda (http://benyehuda.org/). Wszystkie wymienione są przedsięwzięciami podejmowanymi nie dla ekonomicznego zysku. W efekcie powstaje wartościowy produkt, który dorównuje produktom komercyjnym, a nawet je przewyższa. Często mechanizmy takie proliferują do świata polityki - jak wskazuje P. Gawrysiak przytaczając przykład funkcjonującego w brytyjskim parlamencie serwisu TheyWorkForYou, który umożliwia monitorowanie wyborcom wkładu pracy każdego z posłów[17].

Interpretacje polskich Izb i Urzędów Skarbowych pozwalają na wykorzystywanie programów na licencji GNU/GPL do celów komercyjnych i niekomercyjnych przez osoby prywatne oraz instytucje (pierwszego sektora - publiczne, drugiego sektora - nastawione na zysk i trzeciego sektora - fundacje i stowarzyszenia). Otwarte oprogramowanie zdecydowały się wykorzystywać niektóre kraje (uczyniła to między innymi Dania). W 2007 roku duński parlament uchwalił prawo zobowiązujące administrację publiczną do używania dokumentów o otwartym źródle (*Open Document Format* - ODF)[18].

[15] Idea ta również związana jest z nazwiskiem R.S. Stallmana. Powołał on w 1999 roku projekt o nazwie GNUPedia (Patrz: R.M. Stallman, *The Free Universal Encyclopedia and Learning Resource. Announcement of the Project*, „GNU Operating System", w: http://www.gnu.org/encyclopedia/free-encyclopedia.html, dostęp: kwiecień 2012). Szybko jednak zawieszono prace nad nim, a sam R.S. Stallman użyczył swojego wsparcia dla Wikipedii, która powstawała na bazie Nupedii Jimmy'ego Walesa i Larry'ego Sangera. Wikipedia powstała w 2001 roku, obecnie zgromadziła ona ponad 20 milionów haseł we wszystkich edycjach językowych, w tym ponad 3,7 mln haseł w wersji angielskiej oraz ponad 0,8 mln haseł w wersji polskiej. Wikipedia posiada swoje liczne odmiany funkcjonalne (projekty siostrzane): Wikiźródła (*Wikisource*), Wikicytaty (*Wikiquote*), Wikiksiążki (*Wikibooks*), Wikisłownik (*Wictionary*), Wikiwersytet (*Wikiversity*), Wikigatunki (*Wikispecies*) i Wikijunior, a także liczne tak zwane forki (alternatywne odmiany różnych projektów, na przykład hiszpański projekt autonomiczny wobec Wikipedii - *Enciclopedia Libre* utworzona w lutym 2002 lub zrywająca z zasadą neutralnego punktu widzenia *Wikiinfo*).

[16] Projekt został rozpoczęty w roku 1992 roku, jego nazwa pochodzi od nazwiska narodowego poety Finlandii Johana Ludviga Runeberga.

[17] P. Gawrysiak, *Cyfrowa rewolucja..*, dz. cyt., s. 329.

[18] D. Kustre, *Denmark waves Good Bye to Microsoft formats*, w: www.brightsideofnews.com/news/2010/2/3/denmark-waves-good-bye-to-microsoft-formats.aspx, dostęp: kwiecień 2012, 2 marca 2010.

4

Rozdział 4. Techniczne podstawy pracy z PSPP

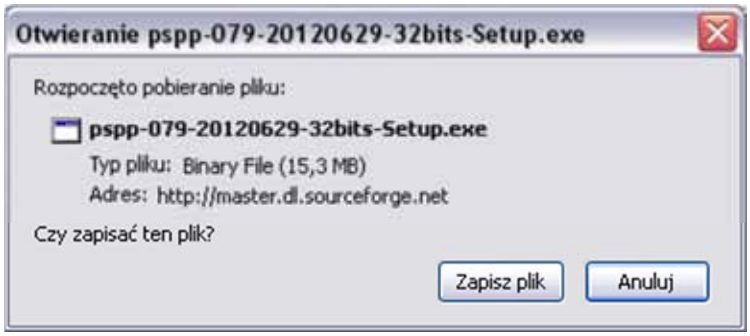
Poniżej omówiono techniczno-informatyczne podstawy pracy z PSPP: sposób odnalezienia programu w sieci WWW, zainstalowania i odinstalowania go, a także zasady uruchamiania i zamykania tej aplikacji oraz procedury otwierania, zamykania, importowania i eksportowania zbiorów danych przeznaczonych do analizy. Średniozaawansowani użytkownicy komputerów mogą pominąć punkty 4.1, 4.2. i 4.3 tego rozdziału - obsługa w zakresie czynności wdrożeniowych oraz podstaw działania jest analogiczna do obsługi innych programów użytkowych. Omówienia dokonano dla wersji przeznaczonej dla systemu MS Windows.

4.1. Pobieranie, instalowanie i deinstalowanie programu PSPP

Łączny czas pobrania i instalacji programu PSPP nie przekracza pięciu minut. Instalacja tego programu jest intuicyjna, prosta, analogiczna do instalacji innych programów w systemie operacyjnym MS Windows.

W celu pobrania programu PSPP przejdź do strony internetowej: http://www.gnu.org/software/pspp/ lub wyszukaj w wyszukiwarce alternatywną stronę www lub ftp, na której można pobrać PSPP; w wyszukiwarce możesz posłużyć się słowami kluczowymi: +PSPP +gnu +download. Na stronie http://www.gnu.org/software/pspp/ w menu (znajdującym się u góry strony) kliknij odnośnik 'Get PSPP'. Na otwartej stronie wybierz wersję dla MS Windows o nazwie 'Another Mingw-based version' (jest to wersja anglojęzyczna, wersja polskojęzyczna programu nie została jeszcze opracowana). Następnie kliknij na odnośnik z najnowszą wersją programu pod warunkiem; odnośnik ten ma postać: 'PSPP_0.7.9_2012-06-29_32bits'.

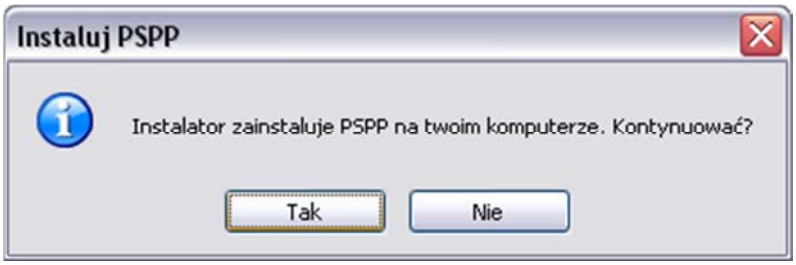
W ciągu kilku sekund pojawi się okno:



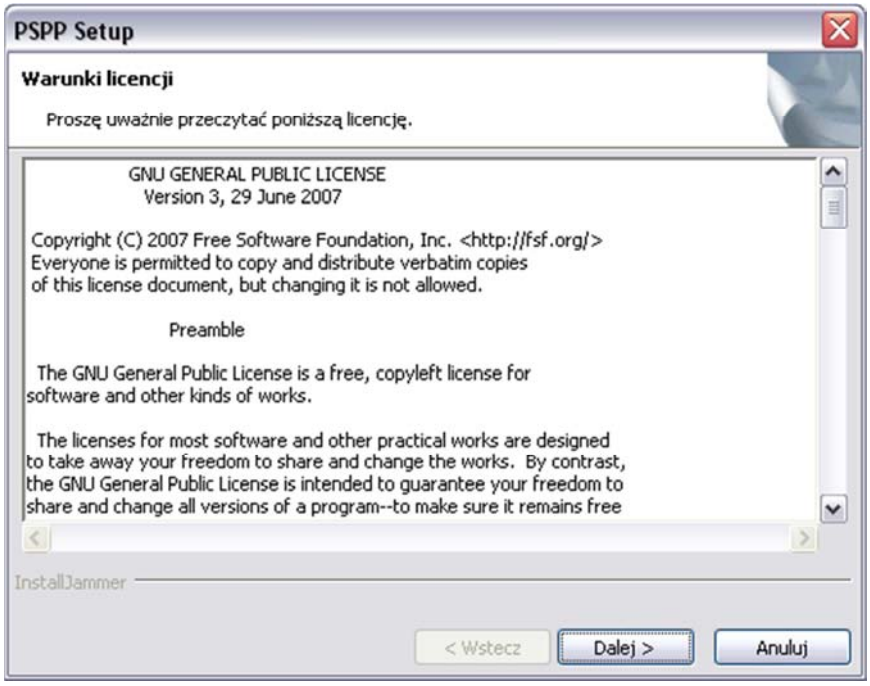
W celu zainstalowania PSPP potrzebujesz co najmniej 16 MB wolnego miejsca na twardym dysku. Potwierdź, aby zapisać plik. Po zapisaniu pliku na twardym dysku odnajdź go i kliknij dwukrotnie ikonę:

W oknie, które się wówczas pojawi wybierz język instalacji *Polski*:

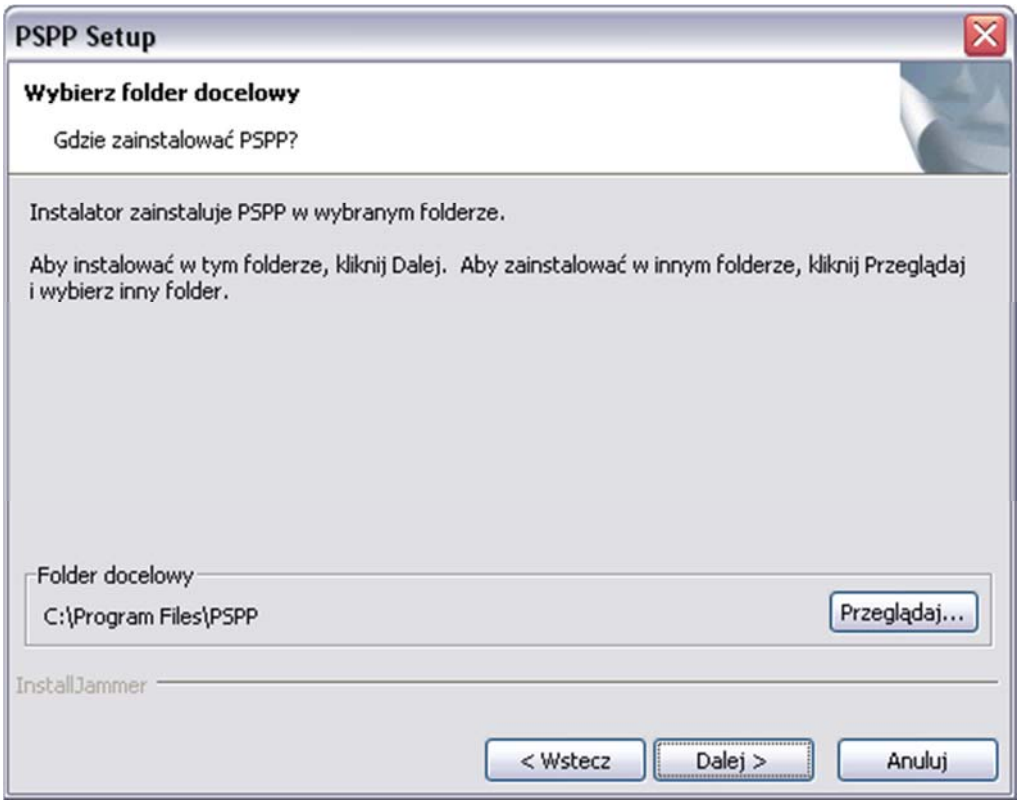
Kliknij *OK*:



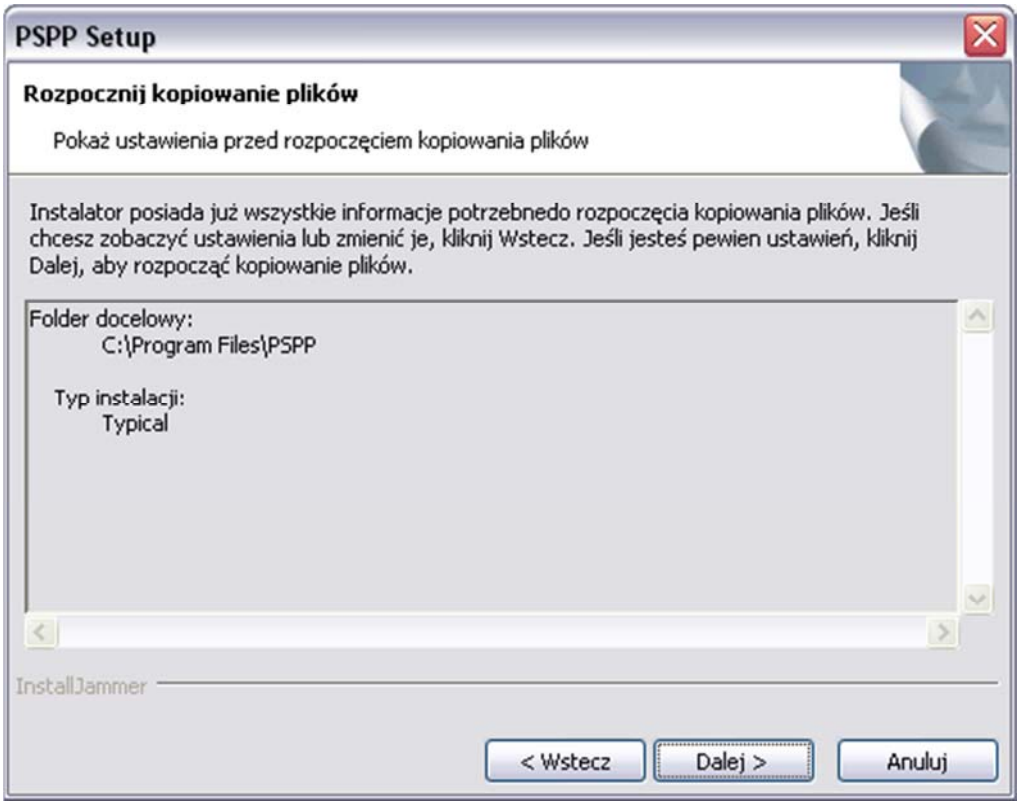
Kliknij *Tak*:



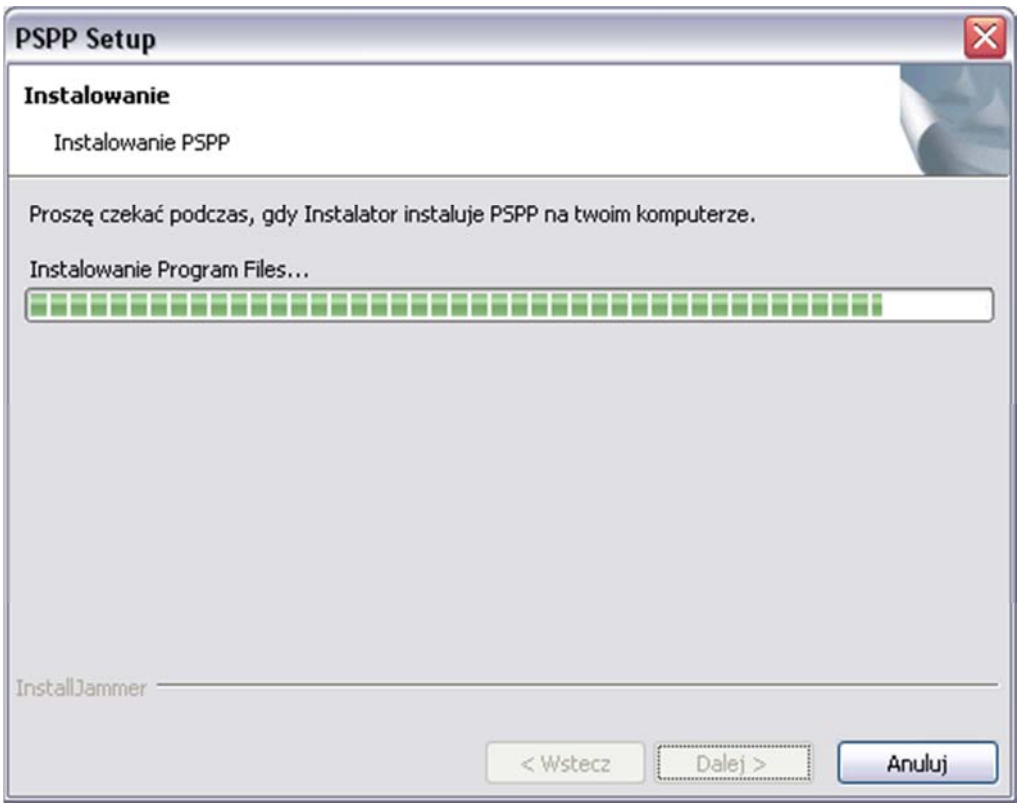
Zaakceptuj warunki licencji klikając *Dalej*.



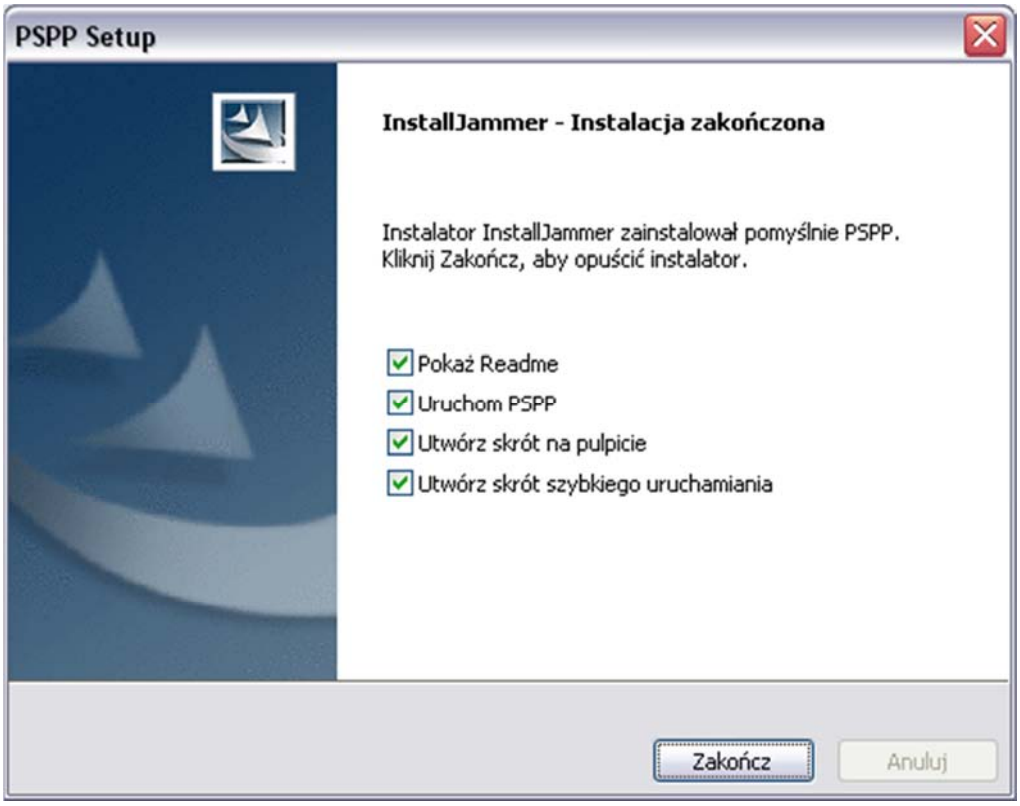
Wybierz folder docelowy instalacji PSPP (upewnij się, że masz wystarczającą ilość miejsca na twardym dysku), kliknij *Dalej*.



Sprawdź wybrane ustawienia i kliknij *Dalej*.

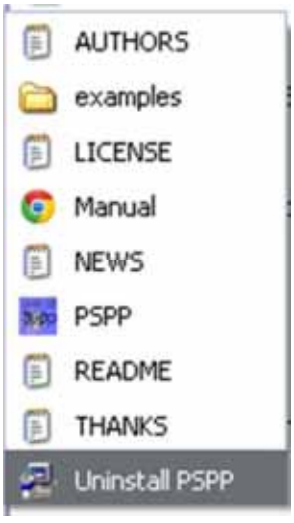


Rozpocznie się instalacja programu PSPP.



Wybierz w *check-boxach* operacje, których chcesz dokonać (najlepiej pozostaw wszystkie zaznaczone) i kliknij zakończ. Instalacja programu PSPP została zakończona, a on sam uruchomiony.

W celu odinstalowania programu PSPP wykonaj następującą sekwencję czynności. Po pierwsze, zamknij program PSPP. Upewnij się, że proces psppire.exe nie działa (CTRL + ALT + DEL Procesy psppire.exe). Następnie wejdź w Start Programy PSPP Kliknij na *Uninstall PSPP:*



Podążaj za wskazówkami (klikaj '*Dalej*'). Program PSPP zostanie usunięty z twardego dysku.

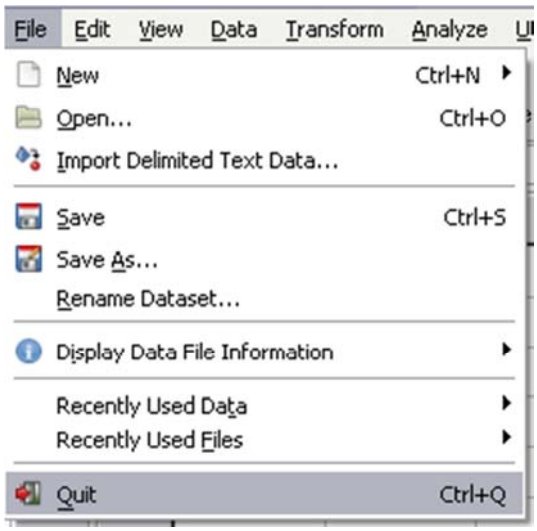
Pamiętaj, że w celu zainstalowania nowej wersji programu PSPP konieczne jest odinstalowanie poprzedniej wersji.

4.2. Uruchamianie i zamykanie programu PSPP

Uruchomienie programu odbywa się poprzez kliknięcie ikony programu PSPP na Pulpicie, pasku szybkiego uruchamiania lub w menu Start ⇨ Programy ⇨ Folder PSPP ⇨ PSPP. Ikona programu PSPP:

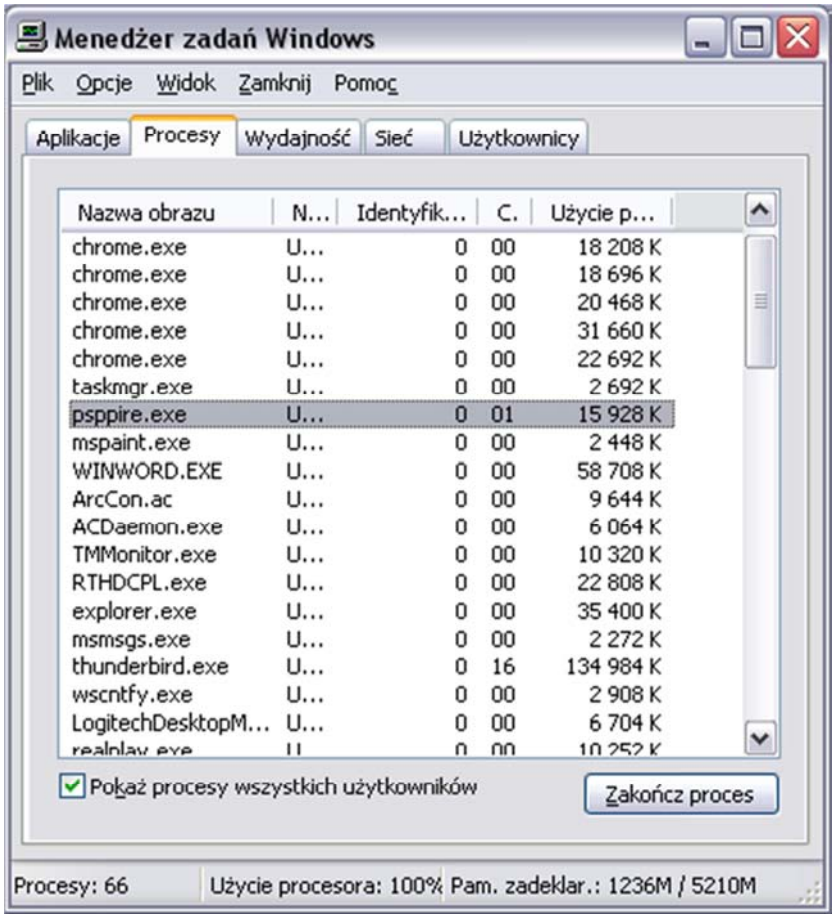
Po wykonaniu jednej z powyższych czynności program jest gotowy do pracy.

Zamykanie programu PSPP odbywa się poprzez wybranie w menu (menu tekstowe znajduje się u góry okna programu) '*File*', a następnie *'Quit'*:



Alternatywnie można skorzystać ze skrótu klawiszowego CTRL + Q (naciśnij jednocześnie klawisz *Control* oraz klawisz Q).

Niekiedy zachodzi potrzeba (na przykład w sytuacji zawieszenia się programu) zamknięcia programu PSPP w trybie wymuszonym. Naciśnij jednocześnie klawisze CTRL + ALT + DEL (Control, Alt oraz Delete) lub w pole 'Uruchom' (Start ⇨ Uruchom) wpisz 'taskmgr'. Ukaże się wówczas Menedżer zadań Windows. Przejdź do zakładki 'Procesy', a następnie odnajdź proces o nazwie psppire.exe (procesy można sortować alfabetycznie klikając w pole 'Nazwa obrazu'), kliknij na niego i w dolnym prawym rogu okna Menedżera kliknij 'Zakończ proces':



4.3. Otwieranie i zamykanie zbiorów danych

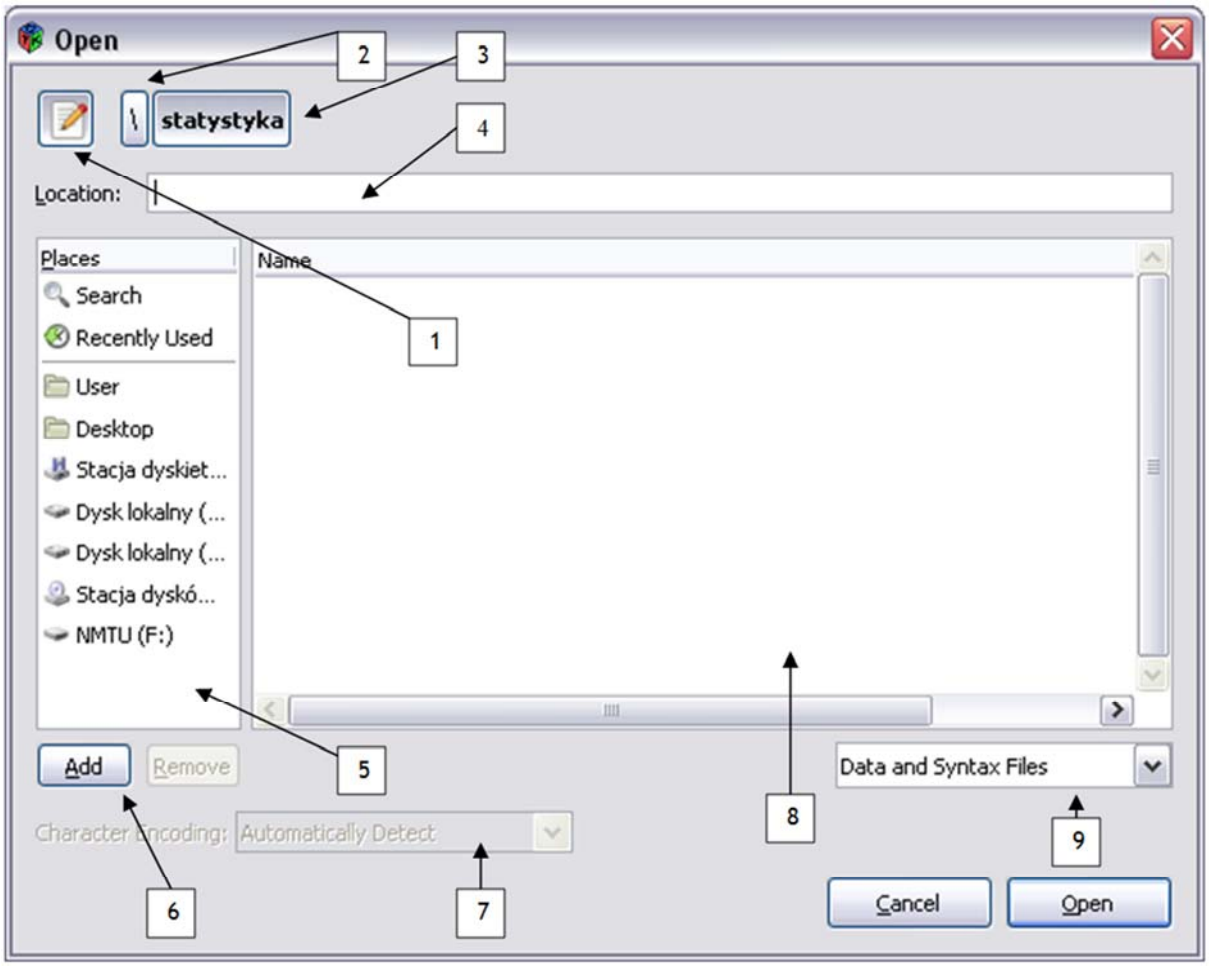
W programie PSPP bezpośrednio można otwierać, przeglądać i modyfikować następujące typy plików:

- **Pliki systemowe (*System Files*)** - są to zbiory danych. Jest to najpowszechniej używany format zapisu plików zarówno w programie PSPP, jak również w SPSS. Rozszerzenie tych plików nosi nazwę *.sav.

- **Pliki przenośne (*Portable Files*)** - pliki danych w formacie transportowym (przenośnym). Podobnie jak pliki *.sav zawierają zbiory danych. Możliwe jest odczytanie tych plików także w innych programach, na przykład w STATISTICA. Rozszerzenie tych plików to *.por.

- **Pliki składni (*Syntax Files*)** - są to pliki zawierające zbiory poleceń (programy) dla programu PSPP, które mają być wykonane na konkretnym zbiorze danych. Umożliwiają one automatyzację poleceń dotyczących analiz. Pliki te posiadają rozszerzenie *.sps.

Po wybraniu ikony „Open" z górnego menu lub *File* ⇨ *Open* ⇨ *Data* można otworzyć dwa pierwsze z wymienionych typów pliku (pliki ze zbiorami danych - *.sav i *.por). Z kolei pliki składni (*.sps) otwieramy wybierając odpowiednio z rozwijanego tekstowego menu *File* ⇨ *Open* ⇨ *Syntax*.



Okno '*Open*' zawiera następujące funkcje - podążając od lewego górnego rogu:

1/ Ikona ta umożliwia włączanie i wyłączanie znajdującego się poniżej paska *Location*, w którym można bezpośrednio wpisywać ścieżkę dostępu do pliku, który chcemy otworzyć,

2/ Prawy ukośnik (*slash*) umożliwia przejście do katalogu wyższego rzędu,

3/ Nazwy katalogów prezentowane w porządku hierarchicznym,

4/ Ścieżka dostępu do otwieranego pliku,

5/ Widok głównych katalogów, z których można korzystać w poszukiwaniu plików,

6/ *Dodawanie* (*Add*) umożliwiające zapamiętanie lokalizacji bieżącego katalogu umożliwiając nawigowanie. Znajdująca się obok opcja *Remove* pozwala na usunięcie zapamiętanej lokalizacji,

7/ Pole *Kodowanie znaków* (*Character Encoding*) (na rysunku nieaktywne, aktywizuje się po wybraniu pliku) umożliwiające otworzenie pliku w językach narodowych z zachowaniem liter diakrytycznych[1]. Domyślną opcją jest automatyczne wykrywanie (*Automatically Detected*). Funkcja ta umożliwia importowanie różnego rodzaju tekstu, na przykład polskich liter diakrytycznych lub cyrylicy. Rozwijana lista

[1] Litera diakrytyczna (z greckiego *diakritikós* - odróżniający) to taka, która dodatkowo zaopatrzona jest w znak graficzny (znak diakrytyczny). Dołączenie znaku diakrytycznego do danej litery alfabetu powoduje utworzenie nowej litery. W języku polskim istnieje dziewięć liter diakrytycznych: *ą, ć, ę, ń, ó, ś, ź, ż*, a także (choć nie wszyscy tak sądzą) litera *ł*.

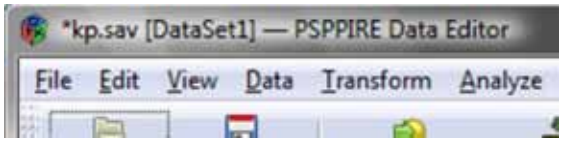
dostarcza licznych możliwości kodowania znaków - między innymi w językach takich jak chiński, rosyjski, hebrajski, turecki czy wietnamski,

8/ Pole nazwy (*Name*), w którym pojawiają się pliki do wyświetlenia w formatach akceptowanych przez program PSPP,

9/ Rozwijana lista umożliwiająca wybór rodzaju wyświetlanych plików. Domyślną wartość stanowi możliwość wyświetlania wszystkich odczytywanych przez program PSPP plików (*.sav, *.por, *.sps).

Zamykanie zbiorów danych w PSPP odbywa się za pomocą wybrania z menu, kliknięcia krzyżyka w prawym górnym rogu lub Ctrl + Q. Jeśli jakiekolwiek zmiany zostały dokonane w bazie danych lub w pliku składni, to niemożliwe jest bezpośrednie zamknięcie programu. Pojawia się wówczas komunikat podający ile czasu upłynęło od ostatniego zapisu. Do wyboru są trzy opcje: zamknięcie zbioru bez zapisywania zmian (*Close without saving*), pozostawienie otwartego zbioru bez zmian, to jest niezamykanie go i niezapisywanie (*Cancel*), zapisanie (*Save*).

Otwarty zbiór danych, w którym dokonano zmian, lecz ich nie zapisano, oznaczony jest symbolem * (gwiazdka) przed nazwą pliku podawaną w lewym górnym rogu okna programu.

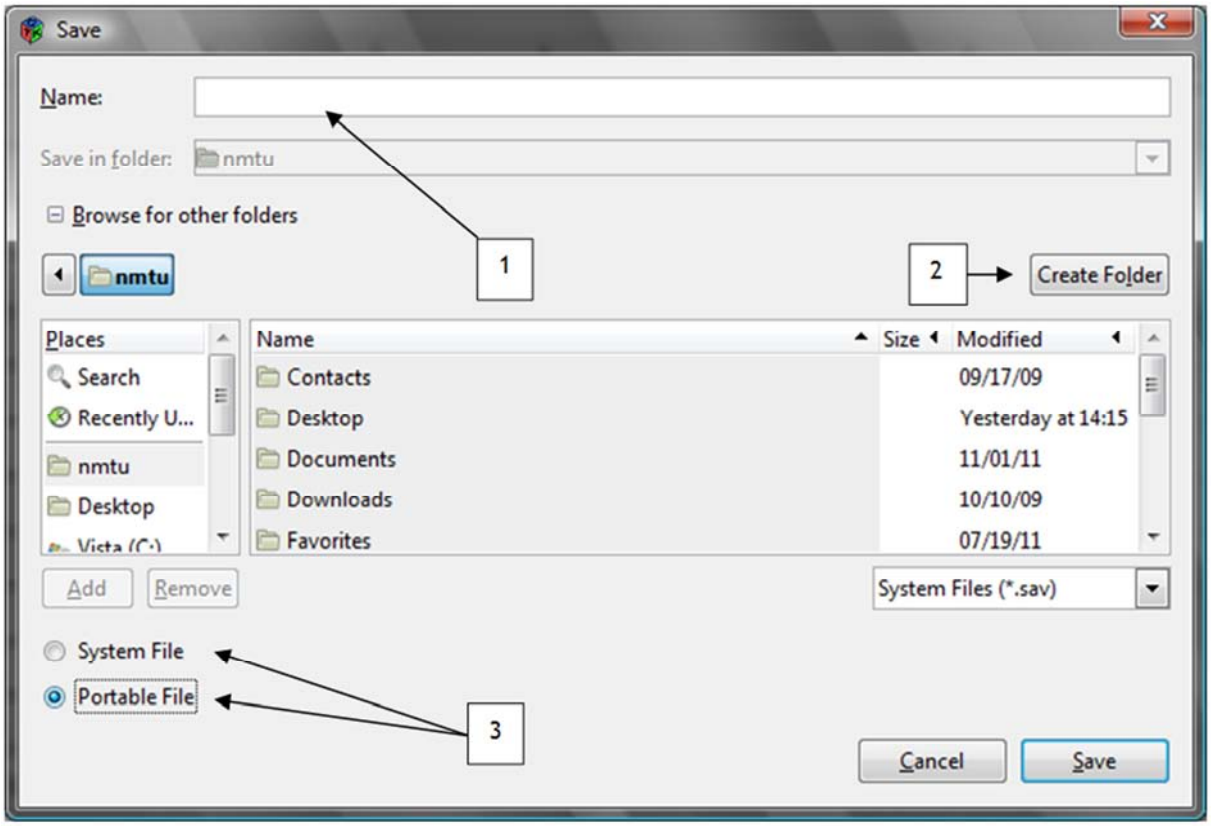


Po zapisaniu zmian w zbiorze danych symbol gwiazdki znika.

Na zakończenie kilka praktycznych uwag na temat pracy z danymi. Po pierwsze, nigdy nie pracujemy (tj. nie modyfikujemy, nie dokonujemy analiz) na oryginalnym zbiorze danych. Należy wykształcić w sobie nawyk pracy na kopii, podczas gdy oryginalną bazę danych mamy bezpiecznie zarchiwizowaną. Po drugie, poszczególne etapy pracy należy zapisywać w odrębnych zbiorach danych. Dokumentuje to postęp prac, a jednocześnie umożliwia cofnięcie się do zarchiwizowanego zbioru w przypadku popełnienia błędu. Im częściej będziemy zapisywać kolejne etapy pracy, tym więcej czasu zaoszczędzimy. Szczególnie początkujący analitycy. Po trzecie, należy wyrobić w sobie nawyk bieżącego zapisywania drobnych zmian i modyfikacji do zbioru danych. Liczne czynności analityczne, których dokonujemy trudno zapamiętać. W przypadku awarii programu będziemy dysponować aktualnym plikiem i nie trzeba będzie dochodzić, które zmiany zostały zapisane, a których się nie udało wprowadzić. Do bieżącego zapisywania można używać skrótu klawiszowego Ctrl + S (*Control* i *Save*).

4.4. Importowanie i eksportowanie zbiorów danych

Program PSPP umożliwia zapisywanie zbiorów danych tylko do dwóch rodzajów plików: systemowych (*.sav) oraz przenośnych (*.por). Domyślnie zbiór danych zapisywany jest do takiego formatu, w jakim został otworzony. Jeśli chcemy zmienić format zapisu lub nazwę pliku, należy z tekstowego menu wybrać *File* ⇨ *Save As* (zapisz jako). Pojawia się wówczas następujące okno:



Interpretujemy je analogicznie jak okno otwierania programu, poza następującymi, oznaczonymi elementami:

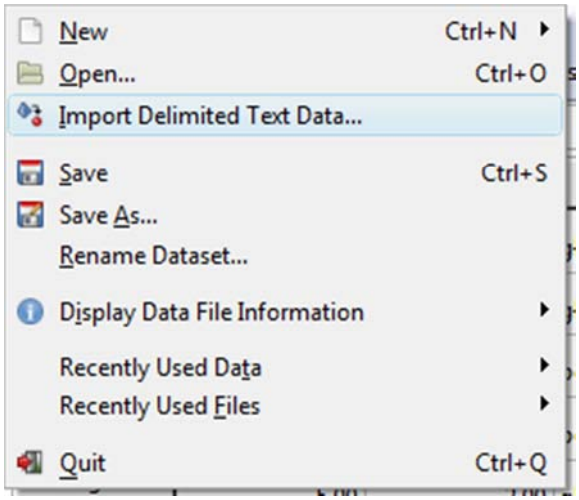
1/ nazwa pliku, którą nadajemy zapisywanemu zbiorowi danych,

2/ utworzenie dodatkowego katalogu, w którym możemy umieścić zbiór danych,

3/ wybór pomiędzy formatem zapisu - do pliku systemowego (*System File,* *.sav) lub do pliku przenośnego (*Portable File*, *.por).

Program PSPP umożliwia otwieranie plików nie zapisanych w formatach takich jak: *.sav, *.por, *.sps. Istnieje możliwość importowania zbiorów danych o innym formacie zapisu: plików z wartościami oddzielonymi przecinkami, tabulatorami lub innymi znakami (zazwyczaj ich końcówki mają następujące nazwy: *.csv, *.dat, *.txt).

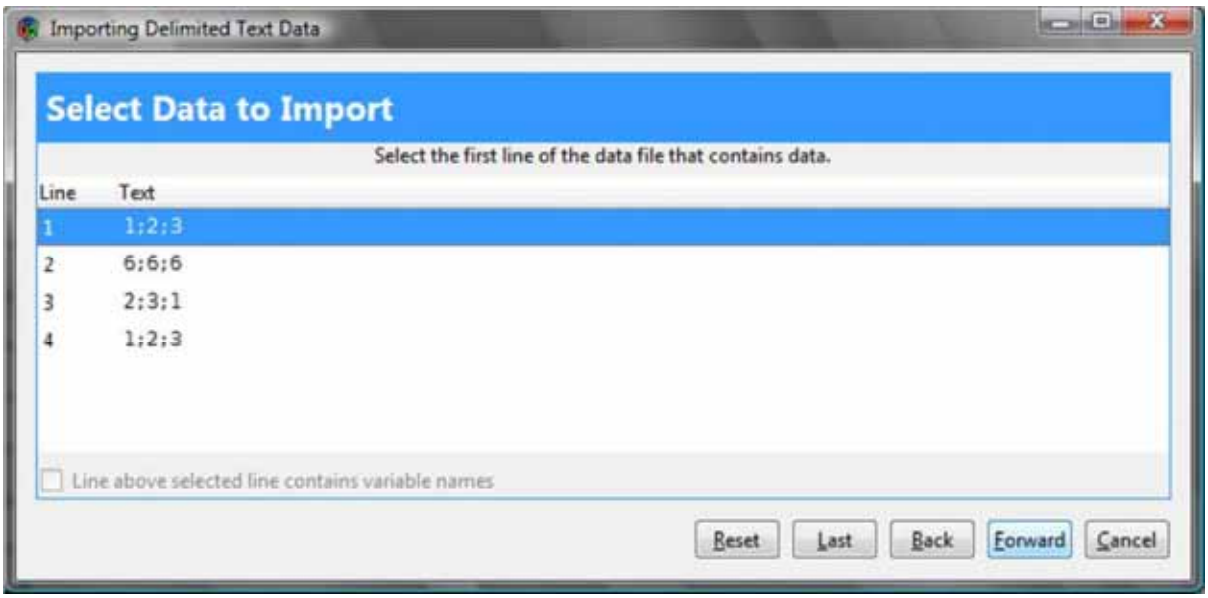
W celu importu plików tego typu do programu PSPP należy z tekstowego menu wybrać *File,* a następnie *Import Delimited Text Data:*



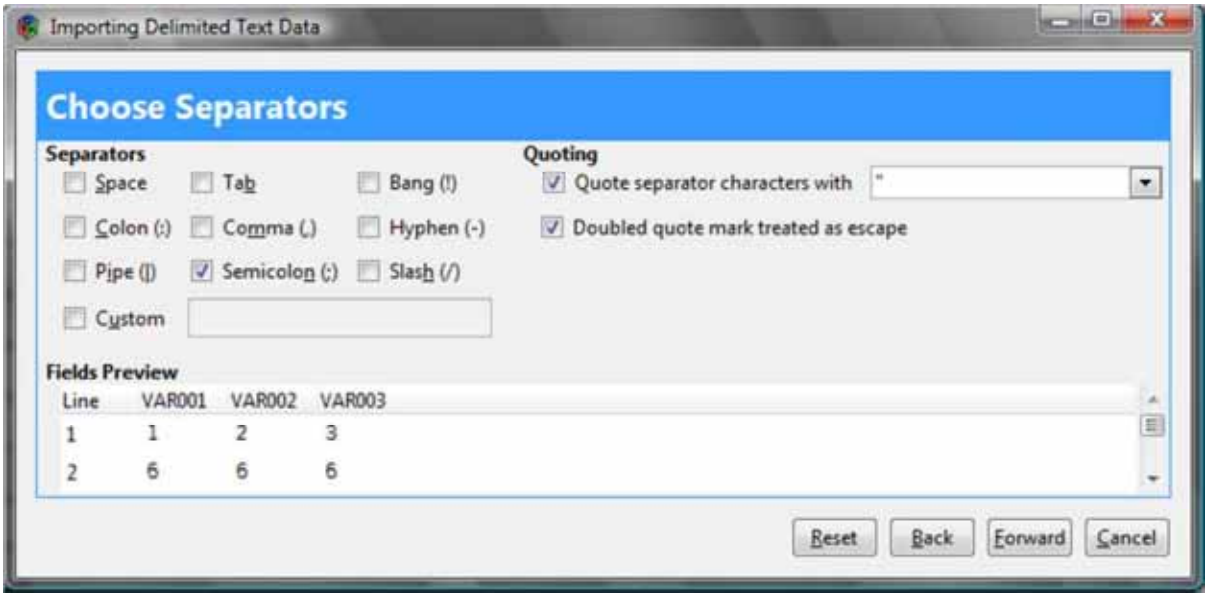
Pojawi się wówczas analogiczne okno do tego, które uzyskujemy w przypadku otwierania w ten sposób pliku systemowego lub przenośnego. Wybieramy plik, który chcemy zaimportować do programu PSPP, a następnie klikamy *OK*. Okno importera danych tekstowych przedstawiono poniżej:

Importer umożliwia nam wciągnięcie do programu PSPP całości zbioru danych (*All cases*), dokładnie określonej liczby kolejnych jednostek analizy (*Only the first ... cases*) lub określonej procentowo części zbioru (*Only the first ... % of file*). Importer wskazuje także informację o liczbie jednostek analizy oraz jaka ich liczba będzie wizualnie dostępna podczas procedury importu. Po wybraniu jednej z opcji (zazwyczaj interesuje nas całość zbioru danych) klikamy *Forward,* ukazuje się okno, w którym wybieramy pierwszą linię, zwierającą dane (domyślnie w pierwszej linii takiego tekstowego pliku z danymi znajdują się nazwy wszystkich kolejnych zmiennych - w takim przypadku zaznaczamy kliknięciem linię drugą).

W przedstawionym na poniższym rysunku przypadku plik nie zawiera linii definiującej zmienne, zatem pozostawiamy zaznaczenie w pierwszej linii:



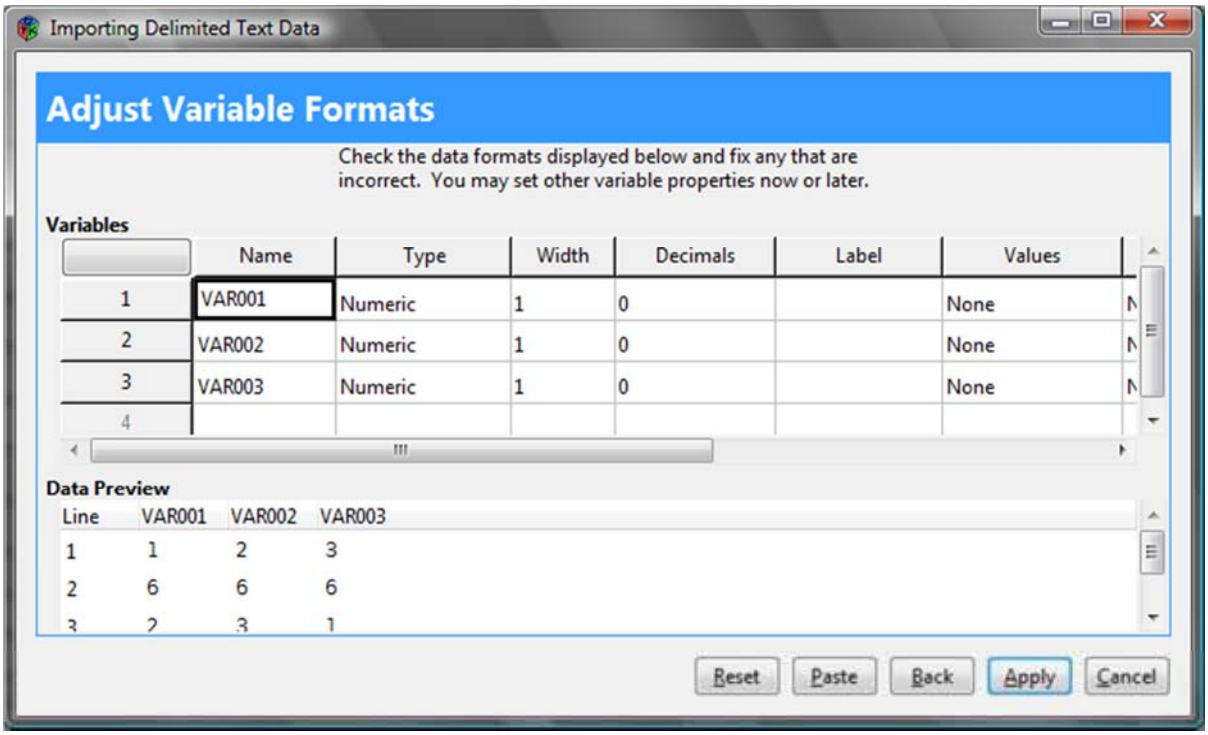
W następnym kroku (po kliknięciu *Forward*) możemy dokonać wyboru separatorów:



Separatory to umowne znaki oddzielające poszczególne wartości zmiennych. Taki zapis umożliwia odtworzenie macierzy danych surowych w PSPP z innych programów. Separatorami mogą być takie znaki jak: spacja (*space*), dwukropek (*colon*), średnik (*semicolon*), tabulator (*tab*), przecinek (*comma*), wykrzyknik (*bang*), linia środkowa pisma (*hyphen,* -), ukośnik prawy (*slash*, /). Jako separator można zdefiniować każdy inny dowolny znak wpisany w polu *custom.* Należy bezwzględnie przestrzegać zasady, że znaki użyte jako separatory nie mogą być wykorzystane w żadnym innym miejscu takiego tekstowego zbioru danych. Jeśli jednak tak się zdarza należy wyznaczyć inny rodzaj separatorów - tak zwane separatory tekstu - znaki cytowania (*quoting*). Wszystko co zostanie wpisane pomiędzy znakami cytowania będzie traktowane jak zwykły tekst i nie będzie interpretowane jako separator pól. Separatorem tekstu może być cudzysłów apostrofowy (podwójny - "), cudzysłów definicyjny (pojedynczy - ') lub oba na raz (- ', "). Wyboru dokonujemy w polu *Quote separator characters with.* Można także oznaczyć zakończenie zbioru

danych w polu *Double quote mark treated as escape.* Zawsze należy zwracać uwagę na stan pola nazwanego *Fields Preview.* Znajduje się tam podgląd fragmentu importowanego zbioru danych. W zależności od dokonywanych czynności na obszarach *Separators* i *Quoting* prezentowany układ danych w *Fields Preview* zmienia się. Należy kontrolować, czy wartości zmiennych są właściwie dzielone.

Ostatnim etapem importu jest opcja dostosowania formatu zmiennych (*Adjust Variable Format*):



Widok w *Variables* jest analogiczny dla okna *Variable View,* a widok *Data Preview* dla *Data View.* W tym etapie (pole *Variables*) formatujemy dane - nazywamy zmienne, jeśli nie chcemy nazw oryginalnych, zmieniamy wartości typu zmiennych, szerokości pól przeznaczonych dla zmiennych, etykiet i opisów danych. W polu *Data Preview* kontrolujemy dokonywane zmiany. Ten etap ma kluczowe znaczenie, bowiem jeśli niewłaściwie określimy parametry prezentowania zmiennych, zaimportujemy zbiór niepełnowartościowy, gubiąc część danych. Szczególne znaczenie ma to przy zmiennych tekstowych, ponieważ program PSPP domyślnie definiuje wszystkie zmienne jako numeryczne.

Po zaimportowaniu zapisujemy zbiór danych w formacie pliku systemowego PSPP.

Wyniki analiz w PSPP (jak je uzyskiwać wyjaśniają kolejne rozdziały) eksportujemy w celu zamieszczenia w raporcie badawczym. Program PSPP daje możliwość zapisu eksportowanych plików zapewniając kompatybilność między innymi z powszechnie używanymi edytorami tekstu, arkuszami kalkulacyjnymi, programami służącymi do sporządzania prezentacji multimedialnych, a nawet umożliwia zapis wyników w formacie pozwalającym na opublikowanie go w postaci strony internetowej.

Eksport wyników odbywa się w oknie widoku raportów (*Output Viewer*). W celu wykonania procedury eksportu wybieramy *File* ⇨ *Export.* Eksport jest możliwy do następujących formatów plików:

- **plik PDF** (*Portable Document Format,* rozszerzenie - *.pdf) - jest to powszechnie używany standard zapisu dokumentów tekstowych opracowany przez firmę Adobe Systems w 1993 roku;

- **plik hipertekstowy** (HTML - *Hyper Text Markup Language*, rozszerzenie - *.html) - podstawowy plik internetowy, zapisany w ten sposób dokument można zamieścić w Internecie jako stronę www;

- **plik pakietu OpenOffice** (*OpenDocument*, rozszerzenie - *.odt) - OpenOffice to darmowy, funkcjonalny odpowiednik pakietu Microsoft Office. Plik ze wskazaną końcówką odpowiada dokumentowi Word (*.doc);

- **prosty dokument tekstowy** (*Text,* rozszerzenie - *.txt) nie zawierający znaków specjalnych, ani formatowań. Zalecany, jeśli eksportowany jest głównie tekst. Eksportuje wyłącznie dane tekstowe, pomija grafikę (wykresy, grafy);

- **dokument PostScript** (*PostScript*, rozszerzenie - *.ps) - jest to uniwersalny standard zapisu poligraficznego, szeroko stosowany w tej branży, jest on językiem opisu tekstu i grafiki na stronie. *Notabene:* PS jest to również język programowania;

- **plik z wartościami oddzielonymi przecinkami** (*Comma-Separated Values*, rozszerzenie - *.csv) - jest to uniwersalny format baz danych. Umożliwia zaimportowanie danych do dowolnego arkusza kalkulacyjnego. W przypadku tabel plik wynikowy staje się nieczytelny. Jest nieprzydatny w przypadku eksportu rysunków.

Na zakończenie garść praktycznych porad. Zazwyczaj wyniki analiz publikowane są przez studentów w rozmaitych pracach promocyjnych (zaliczeniowa praca semestralna, projekt badawczy, praca licencjacka, magisterska). Wówczas najlepszym sposobem przeniesienia danych będzie eksport do formatu OpenOffice. Spośród wymienionych formatów zapisu zachowuje on najwięcej pierwotnych informacji (formatowań), a jednocześnie jest plikiem, który można swobodnie edytować, modyfikować. W tym programie pracuje się identycznie jak w edytorze Microsoft Word. Jeśli koniecznie chcemy pracować w MS Word wówczas możemy z pakietu OpenOffice wyeksportować taki plik bez przeszkód do formatu *.doc. Jeśli uzyskane dane wymagają dalszej obróbki statystycznej poza programem PSPP, najkorzystniej jest wyeksportować je do formatu *.csv, a następnie zaimportować do wybranego arkusza kalkulacyjnego.

5

Rozdział 5. Struktura i organizacja zbioru danych w PSPP

W niniejszym rozdziale przedstawiona została elementarna wiedza teoretyczna i praktyczne umiejętności dotyczące zbiorów danych ilościowych (baz danych). Zbiory danych opisywane są przez powszechnie używany zestaw pojęć, a ich struktura jest ściśle standaryzowana. Warunkiem koniecznym pracy z programem PSPP jest przyswojenie treści znajdujących się w poniższym tekście.

Pierwsza część rozdziału stanowi metodologiczne wprowadzenie do teorii pomiaru. Przede wszystkim omówiono podstawowe elementy składowe zbioru danych (bazy danych) - swoiste jej „cegiełki": zmienne i jednostki analizy. Następnie wyłożone zostały właściwości czterech podstawowych typów zmiennych: nominalnych, porządkowych, interwałowych i ilorazowych, a także zasady, wedle których jeden typ zmiennych można zamieniać na inny. Druga część rozdziału ma charakter praktyczny, zawiera podstawowe wskazówki odnośnie pracy ze zbiorami danych w programie PSPP.

5.1 Macierz danych surowych (zbiór danych): zmienne i jednostki analizy

Dane do analiz statystycznych gromadzone są według standaryzowanych zasad w specjalnie zorganizowanych zbiorach. Zbiory te zawierają informacje o tym, co podlegało pomiarowi i o tym, jaki wynik pomiaru uzyskano. Są to dane zarówno liczbowe, jak też tekstowe. Warunkiem brzegowym zrozumienia teoretycznych podstaw analizy danych jest znajomość podstawowych pojęć opisujących struktury zbioru danych i ich wzajemne zależności. W analizach ilościowych powszechnie używane są pojęcie zmiennej statystycznej, a także pojęcie jednostki analizy (przypadku, rekordu)[1]. Pojęcia te są używane wielokrotnie w dalszych partiach publikacji.

Zmienna (zmienna statystyczna) jest to pewna właściwość lub cecha, która może przyjmować co najmniej dwie wartości. Liczbę wartości, które przyjmuje zmienna, nazywamy zakresem zmiennej. Pomiar zmiennych dokonywany jest zazwyczaj na standaryzowanych skalach. Konkretne wartości zmiennej

[1] Pojęcie *jednostka analizy* ma proweniencję statystyczną i metodologiczną, *rekord* - informatyczną, a *przypadek* (ang. *case*) - pochodzi z praktyki analitycznej.

uzyskujemy w toku obserwacji i pomiaru; oznaczamy je przyporządkowując im liczby rzeczywiste[2]. Przykładem zmiennej jest na przykład udział obywatela w wyborach parlamentarnych. Zmienna ta może przyjmować trzy następujące wartości: uczestniczył w wyborach, nie uczestniczył, nie wiadomo czy uczestniczył. Jako inne przykłady zmiennych można wymienić: płeć, wiek, wykształcenie.

Jednostkami analizy nazywamy obiekty będące przedmiotem realizowanego pomiaru. Może to być osoba, gospodarstwo domowe, firma, instytucja publiczna, określony obszar (np. okręg wyborczy) lub zjawisko umiejscowione w czasie (np. wybory prezydenckie 1990 roku, 1995 roku, 2000 roku, 2005 roku i 2010 roku). Pojedyncza jednostka analizy reprezentuje zazwyczaj tylko jedną z wielu możliwych wartości danej zmiennej. Na przykład, gdy jednostką analizy jest osoba, wówczas zmienna płeć może przyjmować albo wartość 'mężczyzna', albo wartość 'kobieta'.

Jednostki analizy i zmienne podlegające pomiarowi tworzą **macierz danych surowych.** Ma ona postać tabeli. Powszechnie przyjmuje się, że w macierzy danych surowych zmienne umieszczane są w kolumnach, a jednostki analizy w wierszach. Liczba wierszy w tabeli odpowiada zatem liczbie zbadanych przypadków, a liczba kolumn odnosi się do liczby zebranych w badaniu zmiennych. Na przykład w sondażowym badaniu preferencji wyborczych obywateli, liczba wierszy będzie odzwierciedlała liczbę zbadanych osób, a liczba kolumn - liczbę zadanych pytań i innych rodzajów informacji.

Zmienna statystyczna jest funkcją, która każdej jednostce obserwacji (elementowi populacji) przyporządkowuje pewną liczbę rzeczywistą. Wartości zmiennej są przypisywane jednostkom obserwacji w wyniku pomiaru określonej własności[3].

5.2. Poziomy pomiaru zmiennych

Kluczową informacją opisującą zmienną jest tak zwany poziom pomiaru. Poziom pomiaru zmiennej determinuje istotne właściwości zmiennej: jakość danych i ich pojemność informacyjną, możliwe sposoby przetwarzania danych, tj. zakres i rodzaj konkretnych miar statystycznych, które możemy zastosować do jej analizy. W literaturze przedmiotu wyróżnia się cztery poziomy pomiaru zmiennych: nominalny, porządkowy (rangowy), interwałowy (przedziałowy) i ilorazowy (stosunkowy). Poziomy pomiaru zmiennych mają charakter hierarchiczny i determinują odmienne jakościowe charakterystyki. Mówi się o **mocy skali:** od najsłabszej do najsilniejszej. Jeśli to możliwe zawsze staramy się mierzyć zmienne na najwyższych możliwych poziomach - daje to większą swobodę przy wyborze metod analizy oraz możliwość zmiany poziomu mierzonych zmiennych w zależności od potrzeb.

Podział ten jest powszechnie stosowany w naukach społecznych i przyrodniczych[4]. Wprowadził go amerykański psycholog Stanley Smith Stevens (1906-1973), jedna z kluczowych postaci psychologii

[2] Jest to definicja uproszczona, bardziej sformalizowaną zawiera: G. Lissowski, J. Haman, M. Jasiński, *Podstawy statystyki dla socjologów,* Wydawnictwo Naukowe „Scholar", Warszawa 2008, s. 25.

[3] Tamże, s. 25.

[4] Należy wskazać na *votum separatum* niektórych uczonych. Otóż Grzegorz Lisowski, Jacek Haman i Mikołaj Jasiński wyróżniają aż pięć poziomów pomiaru: nominalny, porządkowy, przedziałowy, stosunkowy i absolutny. Zobacz: Tamże, s. 33-38.

eksperymentalnej w połowie XX wieku. Badacz ten zajmował się psychofizjologią ludzkiego słuchu oraz teorią pomiaru zjawisk psychologicznych[5].

Poziom nominalny. Jest to elementarna forma pomiaru opierająca się na dwuwartościowej logice predykatów. Polega na klasyfikowaniu obiektów wedle zasady posiadania określonej cechy lub jej braku, a więc orzekania, że porównywane obiekty są pod określonym przez badacza względem tożsame lub różne. Zbiór tak sklasyfikowanych obiektów ma charakter nieuporządkowany - nie można zasadnie orzekać, że pomiędzy obiektami z dwóch różnych klas zachodzą jakiekolwiek relacje metryczne: na przykład typu więcej - mniej. Cyfry wykorzystywane do oznaczenia poszczególnych wartości zmiennej mają charakter symboliczny, niematematyczny. Na przykład z faktu oznaczenia zmiennej *płeć* cyfrą 1 dla kobiet, a cyfrą 2 dla mężczyzn, nie można wnioskować odnośnie przypisywania przez badacza rang tym wartościom (a więc nie wynika z takiej klasyfikacji fakt, że mężczyźni są np. dwukrotnie lepsi od kobiet), a także dokonywać na nich jakichkolwiek operacji arytmetycznych. Liczby te zostały użyte tylko po to, by wskazać, że są to dwie różne kategorie. Zmienne mierzone na poziomie nominalnym muszą spełniać szereg warunków. Po pierwsze, konieczna jest rozłączność wartości zmiennej (wzajemne wykluczanie się). Jeśli danemu obiektowi przypisaliśmy jedną z wartości zmiennej, to nie możemy przypisać innej. Po drugie, ważną cechą jest wyczerpywalność wartości zmiennej: kategorie powinny zostać spreparowane tak, by wszystkie podlegające klasyfikacji jednostki analizy zostały przyporządkowane do stworzonych wcześniej kategorii. Tego typu zmienne nazywamy jakościowymi. Jest to najniższy (najsłabszy) poziom pomiaru zmiennej, uzyskujemy na temat zmiennej najmniejszą ilość informacji w porównaniu z innymi poziomami pomiaru.

Zmienne na poziomie nominalnym są zmiennymi dyskretnymi reprezentowanymi przez liczby całkowite lub kategorie pojęciowe, np. płeć lub zaufanie mierzone na pięciostopniowej skali. Mają one charakter niematematyczny. Poza zmiennymi dyskretnymi wyróżniamy również zmienne ciągłe, których wartościami mogą stać się liczby rzeczywiste z zakresu wyznaczanego przez dany zbiór. Przykładem takiej zmiennej może być waga. W programie PSPP zmienne nominalne oznaczane są nazwą *nominal* (nominalne). Dla tych zmiennych zasadnie możemy obliczać tylko rozkłady częstości, na przykład w zbiorze danych znalazło się 120 kobiet i 80 mężczyzn, a zatem: 60 proc. to kobiety, 40 proc. - mężczyźni. Przykładami zmiennych na poziomie nominalnym są na przykład: płeć, przynależność etniczna lub narodowa, przynależność do partii politycznej, województwo będące miejscem zamieszkania badanego, głosowanie w wyborach, wskazanie partii politycznej podczas głosowania. Szczególnym przypadkiem są zmienne **dychotomiczne** to jest dwuwartościowe (binarne, zero-jedynkowe) o charakterze dopełnienia logicznego. Przykładami takich wartości zmiennej są: tak - nie, posiada - nie posiada.

Poziom porządkowy (rangowy). Umożliwia nie tylko klasyfikację jak w przypadku zmiennych mierzonych na poziomie nominalnym, lecz również ich uporządkowanie ze względu na natężenie jakiejś cechy, a więc nadanie im określonej kolejności. O ile zmienne nominalne mają charakter nieuporządkowany, o tyle w przypadku zmiennych porządkowych tworzy się pewne kontinuum intensywności zmiennej. Na ogół zmienne mierzone na poziomie porządkowym są zmiennymi dyskretnymi reprezentowanymi przez liczby. Możliwe są zatem trzy typy relacji dla wartości zmiennych: równości, większości oraz mniejszości. Cyfry wykorzystywane do oznaczenia poszczególnych wartości zmiennej mają ograniczoną matematyczną interpretację na zasadzie mniejsze - równe - większe. Nieuprawnione jest jednak wyciąganie wniosków o równych odległościach pomiędzy poszczególnymi wartościami. Tego typu zmienne, są podobnie jak

[5] Klasyczny wykład na temat czterech poziomów pomiaru zawiera artykuł tegoż autora: S.S. Stevens, *On the Theory of Scales of Measurement*, „Science", 1946, 103 ss. 677-680.

nominalne, zmiennymi jakościowymi. Jest to niski (słaby) poziom pomiaru zmiennej - uzyskujemy na jej temat większą ilość informacji w porównaniu z poziomem nominalnym, lecz mniejszą w porównaniu z poziomem interwałowym i ilorazowym. W programie PSPP oznaczane są nazwą *ordinal* (porządkowe). Na tym poziomie zasadne jest stosowanie miar takich, jak mediana, percentyle oraz niektórych miar zależności (patrz tabela 2), a także wszystkich miar stosowanych na poziomie nominalnym. Przykładami tego typu zmiennych jest na przykład zaufanie do Sejmu na skali: od zdecydowanie nie ufam, raczej nie ufam, raczej ufam, do - zdecydowanie ufam lub też cztero- bądź sześciostopniowej skali ocen szkolnych.

Poziom interwałowy (przedziałowy). Istotą pomiaru zmiennej na tym poziomie jest posiadanie przez nią umownego punktu zerowego, względem którego można zasadnie orzekać „o ile więcej" lub „o ile mniej", lecz nie „ile razy więcej" lub „ile razy mniej". Na tym poziomie uprawniona jest więc ograniczona matematyczna interpretacja wartości liczbowych - dozwolone jest dokonywanie działań algebraicznych na poszczególnych wartościach zmiennych. Jest to zatem, w odróżnieniu do pomiaru na poziomie nominalnym lub porządkowym, pewna liczbowa jednostka miary. Liczby oznaczające poszczególne wartości zmiennej reprezentują tu konkretne wartości liczbowe, a nie zakodowane, symboliczne własności. Poszczególne wartości liczbowe zmiennej nie tylko porządkują, ale również określają dystanse pomiędzy nimi. Na tym poziomie pomiaru często występują zmienne ciągłe. Poziom ten posiada wszystkie własności pomiarów niższego poziomu: porządkowego i nominalnego. Pamiętać należy, że zarówno jednostki pomiaru, jak również punkt zerowy takiej skali ustalany jest arbitralnie. Ponadto należy zasygnalizować, że pomiar na tym poziomie jest w politologii rzadki i trudny do uzyskania. Jako egzemplifikacja tego typu zmiennych może posłużyć temperatura mierzona w stopniach Celsjusza (°C). Można co prawda orzekać o różnicach temperatur stwierdzając, że różnica pomiędzy 3°C i 13°C jest taka sama jak pomiędzy 50°C i 60°C, jednakże nie jest zasadne stwierdzenie, że pomiędzy 10°C i 40°C temperatura jest czterokrotnie większa. Innymi przykładami mogą być wskaźnik demokracji periodyku „The Economist" lub Indeks Percepcji Korupcji (*Corruption Perceptions Index*) Transparency International. Zmienne interwałowe należą do zmiennych ilościowych. Jest to wysoki (silny) poziom pomiaru zmiennej. Uzyskujemy na temat zmiennej mniejszą ilość informacji w porównaniu z poziomem ilorazowym, lecz większą w porównaniu ze zmiennymi jakościowymi: nominalnymi i porządkowymi. Tego typu zmienne oznaczane są w programie PSPP nazwą *scale* (skalarne).

Poziom ilorazowy (stosunkowy, proporcjonalny). Poziom ten kumuluje wszystkie właściwości charakterystyczne dla omówionych poziomów pomiaru. Jego cechą konstytutywną jest posiadanie naturalnego, a nie arbitralnie ustalonego punktu zerowego. Implikacją tej cechy jest możliwość stosowania wszystkich operacji matematycznych, a więc również dokonywania przekształceń zmiennych opartych na proporcjach - mnożenia i dzielenia. Wskazuje się, że poziom interwałowy i ilorazowy są niemal identyczne. Zmienne ilorazowe są to zmienne ilościowe. Jest to najwyższy (najsilniejszy) poziom pomiaru zmiennej, uzyskujemy największą możliwą do uzyskania ilość informacji na temat danego zjawiska. Na tym poziomie pomiaru najczęściej występują zmienne ciągłe. Tego typu zmienne oznaczane są w programie PSPP tak samo jak zmienne mierzone na poziomie interwałowym - *scale* (skalarne). Przykładem tego typu poziomu pomiaru jest wiek podawany w latach. Można nie tylko sensownie orzekać tu o różnicach, ale również wskazać, że dwudziestolatek jest dwukrotnie starszy od dziesięciolatka.

Warto zwrócić uwagę, że w literaturze przedmiotu używa się niekiedy pojęcia **absolutnego poziomu pomiaru,** w którym jednostki mają charakter naturalny, a nie dowolny i jednocześnie istnieje tu naturalny punkt zerowy jak w poziomie ilorazowym. Poziom absolutny pomiaru uzyskuje się w toku zliczania

obiektów - na przykład ludzi, partii politycznych w danym kraju, liczby urzędników, liczby głosujących, liczby zgonów podczas wojny.

Podsumowanie rozważań na temat poziomu pomiarów zmiennej zawiera tabela 2.

Tabela 2. Charakterystyka poziomów pomiaru zmiennych

Nazwa poziomu pomiaru zmiennej	Cechy definicyjne poziomu pomiaru zmiennej	Przykłady pomiaru zmiennej na danym poziomie	Stosowane miary pozycyjne	Stosowane miary rozrzutu	Stosowane pozostałe miary (przykłady)	Typy zmiennych	Hierarchia poziomów pomiaru	Nazwy poziomów pomiaru stosowane w programie PSPP
Nominalny	Tożsamość lub różnica (należy do zbioru - nie należy do zbioru)	Płeć Miejsce zamieszkania (np. województwo) Głosowanie lub niegłosowanie w wyborach Przynależność do partii politycznej Przynależność do grupy etnicznej lub narodowościowej	Dominanta		Chi-kadrat Phi Współczynnik kontyngencji V Kramera Lambda Tau Goodmana i Kruskala Współczynnk niepewności	Jakościowe	Najniższy (najsłabszy) poziom pomiaru - najmniejsza moc skali	*Nominal* (nominalny)
Porządkowy (rangowy)	Porządkowanie (mniejsze, tożsame, większe)	Zaufanie do Sejmu Skala ocen szkolnych Wielkość miejsca zamieszkania (w przedziałach)	Mediana	Percentyle	Gamma Tau-b Kendella Tau-c d-Somersa			*Ordinal* (porządkowy)
Interwałowy (przedziałowy)	Porównanie, pomiar dystansu (o ile większe o ile mniejsze)	Temperatura w stopniach Celsjusza (C) *Wiek* (rok urodzenia) Skala ilorazu inteligencji Stanforda-Bineta Wskaźnik demokracji *Economist intelligence Unit* Indeks Percepcji Korupcji (*Corruption Perceptions Index*)	Średnia arytmentyczna	Odchylenie standardowe Wariancja Rozstęp	Kurtoza Skośność Rho Spearmana Test t-Studenta ANOVA MANOVA	Ilościowe	Najwyższy (najsilniejszy) poziom pomiaru - największa moc skali	*Scale* (skalarny)
Ilorazowy (stosunkowy)	Porównywanie z użyciem wielkości absolutnych (ile razy większe, ile razy mniejsze)	Temperatura w Kelwinach (K) Zarobki Wiek w latach	Średnia geometryczna Średnia harmoniczna		R Pearsona Eta			

Źródło: Opracowanie własne.

5.3. Transformacja zmiennych pomiędzy różnymi poziomami pomiaru

Poziomy zmiennych zmierzonych na określonym poziomie możemy modyfikować, ale tylko w drodze ich redukcji, to jest utraty części posiadanych informacji o badanym zjawisku. Na przykład zmienna nominalna nie może być zmieniona na żadną inną, bowiem pomiar odbywał się na najniższym możliwym poziomie; zmienna porządkowa może być transformowana na zmienną nominalną; zmienna interwałowa - na zmienną porządkową lub nominalną; zmienna ilorazowa na wszystkie trzy pozostałe. Informacje na temat celowości redukcji danych zawierają kolejne rozdziały. Prześledźmy kolejne transformacje następujących trzech zmiennych mierzonych na poziomie ilorazowym: temperatury, wieku oraz poparcia dla demokracji. Proponujemy rozważać poniższe przykłady abstrahując od ich poprawności metodologicznej i logicznej - mają one tylko wyjaśniać procedury i konsekwencje zmiany poziomu pomiaru.

Przykład 1 - zmienna *temperatura*. Pomiaru temperatury na skali ilorazowej dokonuje się wykorzystując skalę Kelwina, bowiem skala ta spełnia konstytutywny wymóg tego poziomu pomiaru: posiada zero znaczące. Najniższa możliwa temperatura, jaką może przyjąć ciało to taka, w której ustały wszelkie drgania cząstek. Jest to temperatura równa 0 K (Kelwinów). Pomiar na tym poziomie umożliwia nam dokonywanie na zebranych danych wszystkich operacji matematycznych, a więc również mnożenia i dzielenia. Możemy zasadnie powiedzieć, że temperatura 20K jest pięciokrotnie wyższa od temperatury 4K. Przypuśćmy, że konieczna jest zamiana skali pomiaru temperatury na skalę Celsjusza: i tak zamiast 0K po transformacji zmiennych otrzymamy -273,15°C, zamiast 283K otrzymamy 10,15°C, itd. Pomiar na takiej skali jest dla nas co prawda łatwiejszy do zapercypowania i interpretacji, bowiem na co dzień posługujemy się skalą Celsjusza, jednak utraciliśmy ważną właściwość skali ilorazowej: możliwe operacje matematyczne jakich możemy dokonywać to tylko dodawanie i odejmowanie, lecz nie mnożenie i dzielenie. Nie możemy zasadnie powiedzieć, że 15°C to temperatura dwukrotnie mniejsza od temperatury 30°C. Możemy jednak twierdzić, że różnica temperatur pomiędzy parą: 10°C i 15°C jest taka sama, jak pomiędzy 20°C i 25°C. Dokonując dalszej redukcji danych, a więc przechodząc na poziom porządkowy, przyjmujemy na przykład, że temperatury poniżej 0°C uznamy za niskie, powyżej 0°C, lecz poniżej 25°C za umiarkowane, a powyżej 25°C - za wysokie. Na skutek takiego zabiegu zmienna *temperatura* zostaje uproszczona - ogranicza się zaledwie do trzech wartości, jednakże na tym poziomie pomiaru tracimy istotną właściwość - nie możemy na przykład zasadnie orzekać, o ile temperatura uznana za niską jest wyższa od tej uznanej za wysoką. Możemy natomiast stwierdzać, że istnieje określony porządek: temperatury umiarkowane są wyższe niż oznaczone jako niskie. Najniższy poziom pomiaru jeszcze bardziej redukuje posiadane informacje. Przypuśćmy, że uznaliśmy temperatury umiarkowane na skali porządkowej za optymalne i bezpieczne dla ludzkiego zdrowia, a temperatury oznaczone jako niższe lub wyższe - za takie, które nie zawierają się w granicach bezpiecznych dla człowieka. Otrzymujemy w ten sposób zmienną dwuwartościową: temperatur bezpiecznych (umiarkowane wartości) i temperatur stwarzających zagrożenie (zbyt niskie lub zbyt wysokie wartości temperatur). Nie możemy już orzekać, które z dwóch wartości zmiennych są niższe, a które wyższe. Zasadnie możemy tylko konstatować istnienie różnicy lub tożsamości wartości zmiennych.

Przykład 2 - *zmienna wiek*. Zmienną wiek redukujemy analogicznie jak zmienną temperatura. Poziom ilorazowy zmiennej to wiek podawany w latach: trzydziestolatek jest trzykrotnie starszy od dziesięciolatka i dwukrotnie młodszy od sześćdziesięciolatka. Zmienna wiek na poziomie interwałowym to rok urodzenia badanego. Punkt zerowy został wyznaczony arbitralnie - wedle kalendarza gregoriańskiego jest to rok urodzenia Jezusa. Redukując zmienną *wiek* do poziomu interwałowego możemy zaklasyfikować liczbę lat dla poszczególnych jednostek analizy do trzech ogólnych klas: młodego pokolenia obejmującego

osoby od 18 do 35 roku życia, pokolenia średniego, do którego zaliczono osoby powyżej 35, aż do 65 roku życia oraz pokolenia starszego, w którym pomieszczono tych, którzy przekroczyli wiek 65 lat. Wskutek takiego zabiegu tracimy możliwość arytmetycznych działań na zmiennych, pozostajemy przy możliwości stwierdzania, że poszczególne jednostki analizy są większe, mniejsze lub równe. Kolejny stopień redukcji - na zmienne nominalne mógłby polegać na podzieleniu jednostek analizy jak w tabeli 3 na takie, gdzie udało się ustalić wiek badanego i takie, gdzie się to nie powiodło.

Przykład 3 - zmienna *poparcie dla demokracji*. Zasady postępowania są tożsame z podanymi w powyższych przykładach. Zmienną tą na poziomie ilorazowym można mierzyć na skali procentowej od 0 do 100 procent, gdzie 0 oznacza brak poparcia dla demokracji, a 100 całkowite dla niej poparcie. Dzięki takiej skali możemy skonstatować, że poparcie dla demokracji na poziomie 20 proc. jest dziesięciokrotnie mniejsze, niż na poziomie 2 proc. Pomiar tej zmiennej na poziomie interwałowym mógłby odbywać się na skali od 1 do 10, gdzie 1 oznacza niskie poparcie dla demokracji, a 10 wysokie. Wartości zmierzone na takiej skali umożliwiają wykonanie przekształceń arytmetycznych. Na poziomie porządkowym poparcie dla demokracji moglibyśmy mierzyć na skali: niskie, umiarkowane i wysokie, a na poziomie nominalnym używając dwuwartościowej zmiennej popiera - nie popiera.

Tabela 3. Przykłady transformacji poziomu pomiaru zmiennych na skali S.S. Stevensa

Lp.	Poziom ilorazowy pomiaru zmiennej	Poziom interwałowy pomiaru zmiennej	Poziom porządkowy pomiaru zmiennej	Poziom nominalny pomiaru zmiennej
1	Zmienna *temperatura* mierzona na poziomie ilorazowym - na skali bezwzględnej Kelwina: ... K	Zmienna *temperatura* mierzona na poziomie interwałowym - w stopniach Celsjusza: ...°C	Wartości zmiennej *temperatura* mierzonej na poziomie przedziałowym: 1: niska temperatura 2: umiarkowana temperatura 3: wysoka temperatura	Wartości zmiennej *temperatura* mierzonej na poziomie nominalnym: 1: temperatura zawiera się w przedziale bezpiecznym dla ludzkiego życia 2: temperatura nie zawiera się w przedziale bezpiecznym dla ludzkiego życia
2	Zmienna *wiek* mierzona na poziomie ilorazowym - wiek podawany w latach: 0: 0 lat 1: 1 rok ... 18: 18 lat 19: 19 lat 20: 20 lat ... n: 90 lat	Zmienna *wiek* mierzona na poziomie interwałowym - jako rok urodzenia: ... 1971 1972 1973 1974 1975 1976 ...	Wartości zmiennej *wiek* mierzonej na poziomie porządkowym: 1: Młode pokolenie (od 18 do 35 lat) 2: Średnie pokolenie (powyżej 36 do 65 lat) 3: Starsze pokolenie (powyżej 65 lat)	Wartości zmiennej *wiek* mierzonej na poziomie nominalnym: 1: Posiadamy informację na temat wieku respondenta 2: Nie posiadamy informacji na temat wieku respondenta
3	Zmienna *poparcie dla demokracji* mierzona na poziomie ilorazowym - na skali bezwzględnej od 0 do 100 proc.: ... %	Zmienna *poparcie dla demokracji* mierzona na poziomie interwałowym - na 10-punktowej od 1 do 10, gdzie: 1: popieram demokrację w najmniejszym stopniu	Wartości zmiennej *poparcie dla demokracji* mierzonej na poziomie porządkowym: 1: niskie poparcie dla demokracji	Wartości zmiennej *poparcie dla demokracji* mierzonej na poziomie nominalnym: 1: popiera 2: nie popiera

Lp.	Poziom ilorazowy pomiaru zmiennej	Poziom interwałowy pomiaru zmiennej	Poziom porządkowy pomiaru zmiennej	Poziom nominalny pomiaru zmiennej
		2: 3: 4: 5: 6: 7: 8: 9: 10: popieram demokrację w największym stopniu	2: umiarkowane poparcie dla demokracji 3: wysokie poparcie dla demokracji	
Źródło: Opracowanie własne.				

W praktyce badawczej istnieją wyjątki od zasady kierunku transformacji poziomu zmiennych: a więc od mierzonych na silniejszych poziomach do mierzonych na poziomach słabszych. Jak łatwo zauważyć w niektórych przypadkach (np. wiek) łatwo przejść z poziomu niższego - interwałowego na poziom wyższy - ilorazowy. Jednak najistotniejszy z punktu widzenia praktyki analizy danych jest wyjątek obejmujący niektóre zmienne mierzone na poziomach porządkowych. Mniej rygorystyczni badacze dopuszczają uznanie ich za zmienne ilościowe - konkretnie interwałowe ze wszystkimi konsekwencjami dotyczącymi interpretacji i obliczeń statystycznych. Przykładem jest zmienna mierzona na poziomie przedziałowym z użyciem formatu skali Rensisa A. Likerta. Należy przyjrzeć się bliżej formatowi tego typu skali, ponieważ jest on bardzo często wykorzystywany w naukach społecznych. Badanemu prezentuje się poszczególne stwierdzenia, a on umiejscawia swoją postawę w danym punkcie kontinuum pomiędzy skrajnymi punktami: zgodą a odrzuceniem danego stwierdzenia. Swoją odpowiedź badany wybiera ze specjalnie spreparowanego zbioru możliwych odpowiedzi umieszczonych pod ocenianym stwierdzeniem. Taki zbiór możliwych wartości zmiennej - aby można było uznać go za element skali R.A. Likerta - musi uwzględniać cztery następujące zasady:

1/ zasadę symetrii - w zbiorze odpowiedzi musi znaleźć się taka sama liczba wartości zmiennej pozytywnych i negatywnych. Wartości o określonym natężeniu pozytywnym musi odpowiadać analogiczna (o przeciwnym znaku, lecz tożsamym natężeniu) wartość negatywna;

2/ zasadę punktu środkowego - w zbiorze odpowiedzi należy umieścić jedną i tylko jedną odpowiedź logicznie neutralną (indyferentną). Wartość ta oddziela wartości pozytywne od negatywnych;

3/ zasadę nieparzystości - wynikającej z powyższych stwierdzeń: sumy parzystej liczby wartości pozytywnych i negatywnych danej zmiennej oraz jednej wartości indyferentnej[6];

4/ zasadę uporządkowania - zbiór odpowiedzi musi być prezentowany respondentowi w sposób uporządkowany (od skrajnie negatywnej do skrajnie pozytywnej odpowiedzi lub odwrotnie), a nie przypadkowy.

[6] Niektórzy badacze - głównie psychologowie - stosują jako skalę R.A. Likerta kafeterie o parzystej liczbie itemów rezygnując z zasady nieparzystości. Zabieg taki ma na celu wymuszenie ujawnienia postawy respondenta. Jednakże rezygnacja z odpowiedzi neutralnej zaburza równość odstępów pomiędzy poszczególnymi itemami. Porównaj: R. Garland, *The Mid-Point on a Rating Scale: Is it Desirable?*, „Marketing Bulletin", 1991, 2, s. 66-70, S.L. Sclove, *Notes on Likert Scales*, 2001, w: http://www.uic.edu/classes/idsc/ids270sls/likert.htm, dostęp: wrzesień 2012.

Najczęściej stosowany w badaniach społecznych zbiór odpowiedzi jest pięciopunktowy:

1: zdecydowanie nie zgadzam się z danym twierdzeniem,
2: raczej nie zgadzam się z danym twierdzeniem,
3: ani się zgadzam, ani nie zgadzam z danym twierdzeniem,
4: raczej się zgadzam z danym twierdzeniem,
5: zdecydowanie zgadzam się z danym twierdzeniem.

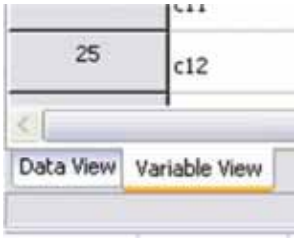
Stosowane są też, lecz rzadziej, listy wartości zmiennych trzy-, siedmio-, dziewięcio-, a nawet jedenastopunktowe[7].

Jest to typowa zmienna porządkowa (rangowa). Transformacja zmiennej utworzonej wedle formatu skali R.A. Likerta z poziomu niższego (porządkowego) na wyższy (interwałowy) odbywa się na podstawie hipotetycznego założenia badacza, że poszczególne wartości zmiennej dzielą przestrzeń własności na równe odcinki (interwały), na przykład odległość pomiędzy odpowiedzią 1 i 2 jest taka sama jak odległość między odpowiedzią 3 i 4. Zabieg ten umożliwia uznanie przypisanych poszczególnym wartościom zmiennych liczb rzeczywistych i podjęcie na nich pełni działań arytmetycznych dozwolonych na poziomie ilorazowym. Wyjątek ten ma niebagatelne znaczenie, umożliwiając stosowanie bardziej precyzyjnych miar w wielu przypadkach badawczych.

5.4. Anatomia zbioru danych w PSPP

W tym podrozdziale omówiono dwa podstawowe i najczęściej wykorzystywane elementy tego programu. Większość czasu analityk danych spędza operując na dwóch elementach: Widoku zmiennych (*Variable View*) oraz Widoku zbioru danych (*Data View*). Po uruchomieniu programu PSPP domyślnym widokiem jest Widok zmiennych. W tym widoku prezentowane są kompletne i standaryzowane informacje na temat wszystkich zmiennych znajdujących się w zbiorze danych. Z kolei w Widoku danych możemy odczytywać wartości poszczególnych zmiennych dla wszystkich jednostek analizy. W lewym dolnym rogu okna Programu PSPP możemy przełączać pomiędzy dwoma głównymi zakładkami programu: Widokiem zmiennych (*Variable View*) i Widokiem danych (*Data View*).

[7] Istotne wydaje się w tym miejscu wyjaśnienie związanego ze skalą R.A. Likerta powszechnego w praktyce badawczej nieporozumienia. Sformułowania „skala Likerta" używa się w dwojakim rozumieniu: wąskim i szerokim. W rozumieniu wąskim oznacza ono zbiór odpowiedzi na pytanie skonstruowany w sposób podany wyżej. Jest to swoisty skrót myślowy, właściwie powinno mówić się o „kafeterii jak w skali Likerta", nie zaś o „skali Likerta". We właściwym - szerokim - rozumieniu skala R.A. Likerta rozumiana jest jako przygotowana według określonych procedur wiązka (bateria) pytań wraz z odpowiednio przygotowanymi itemami oraz określony sposób analizy zebranych danych. W tym ostatnim znaczeniu pojęcia to używane jest w niniejszym artykule. Szczególnie wartą polecenia publikacją wyjaśniającą tę i inne metodologiczne i techniczne nieporozumienia i nadużycia dotyczące skali R.A. Likerta jest artykuł amerykańskich badaczy Jamesa Carifio i Rocco J. Perla: J. Carifio, R.J. Perla, *Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes*, „Journal of Social Sciences", 2007, 3 (3), s. 106-116. Artykuł ten dostępny jest w Internecie: http://www.comp.dit.ie/dgordon/Courses/ResearchMethods/likertscales.pdf, dostęp: wrzesień 2012. Patrz także: J.S. Uebersax, *Likert Scales: Dispelling the Confusion*, „Statistical Methods for Rater Agreement", 2006, w: http://john-uebersax.com/stat/likert.htm, dostęp: wrzesień 2012, strony nienumerowane.



5.4.1. Widok zmiennych (*Variable View*)

Widok zmiennych zawiera komplet informacji o wszystkich zmiennych znajdujących się w zbiorze: ich nazwie, opisie, typie, sposobie ich prezentowania, przyjmowanych przez nich wartościach oraz położeniu względem innych zmiennych. Jest to w istocie syntetyczne zestawienie wszystkich pytań kwestionariusza wraz z potencjalnymi odpowiedziami, na przykład:

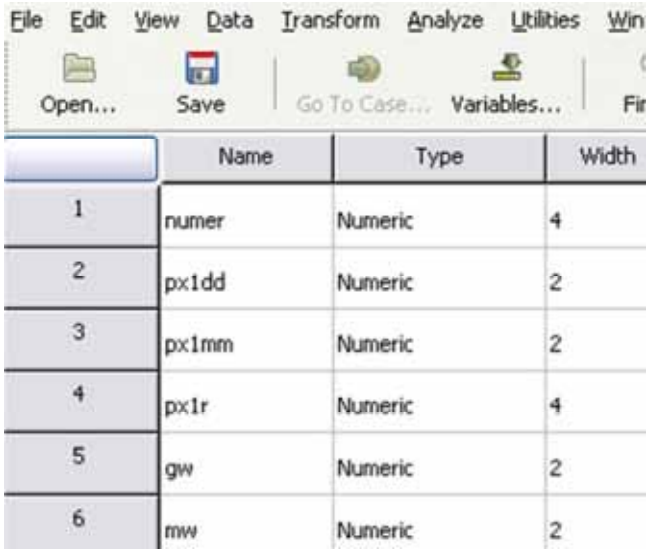
Gdyby wybory odbywały się w najbliższą niedzielę, to na którą partię oddał(a)by Pan(i) swój głos?

1 'Prawo i Sprawiedliwość'
2 'Platforma Obywatelska'
3 'Polskie Stronnictwo Ludowe'
4 'Polska Jest Najważniejsza'
5 'Ruch Palikota'
6 'Sojusz Lewicy Demokratycznej'
7 'Jeszcze nie wiem na kogo będę głosował(a)'
8 'Nie wezmę udziału w wyborach parlamentarnych'
9 'inna partia, jaka?'

Każda zmienna (pytanie kwestionariuszowe) oraz potencjalny zakres odpowiedzi na nie (w powyższym przykładzie 1-9), a także inne informacje prezentowane są w jednym i tylko jednym wierszu.

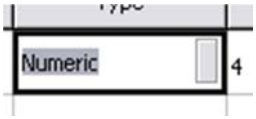
Po otworzeniu zbioru danych w programie PSPP dostrzeżemy zmienne uporządkowane w kolumnach. Każda z kolumn, oprócz pierwszej, posiada w pierwszym wyszarzonym wierszu, znajdującym się tuż pod paskiem menu programu, swoją nazwę. Kolejno są to kolumny zatytułowane: kolumna bez nazwy [numer zmiennej], nazwa zmiennej (*Name*), Typ zmiennej (*Type*), liczba pól przypisana wartości zmiennej (*Width*), liczba miejsc dziesiętnych wartości zmiennej (*Decimals*), etykieta zmiennej (*Label*), wartości zmiennej (*Values*), brakujące wartości (*Missing*), liczba widocznych kolumn (*Columns*), wyrównywanie danych (*Align*) i typ skali (*Measure*). Każdy zbiór danych posiada dokładnie tę jedenastkę kolumn w niezmienionej kolejności.

Kolumna 1. Bez nazwy (Numer). Numer porządkowy zmiennej. Jest to pierwsza kolumna w zakładce *Variable View.* Numer jest zafiksowany i niezmienny, określa porządkowo liczbę zmiennych w zbiorze danych. Kolejność zmiennych wyznacza porządek prezentowania danych w zbiorze: zmienna o numerze porządkowym jeden znajduje się w kolumnie pierwszej w zakładce Widok danych, zmienna o numerze drugim - w drugiej kolumnie Widoku danych, i tak dalej.

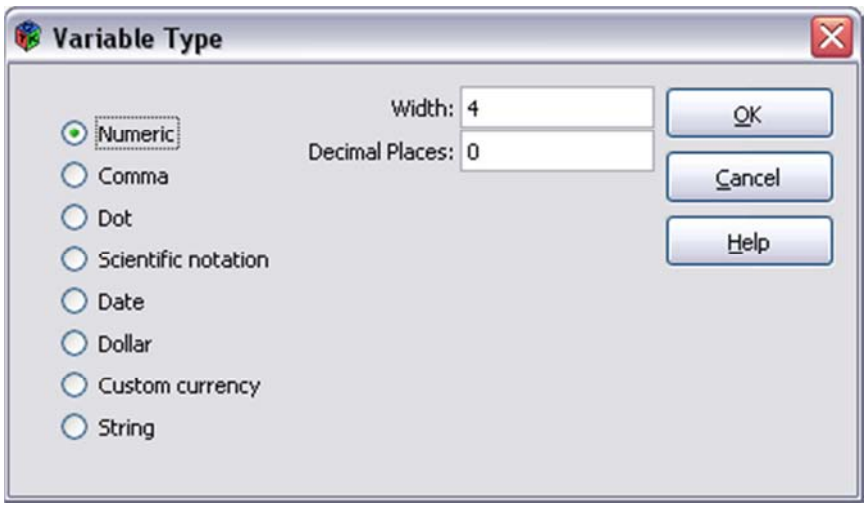


Kolumna 2. Nazwa zmiennej (*Name*). Jest to identyfikator zmiennej. Służy do oznaczania zmiennych podczas wykonywania na nich obliczeń. Identyfikator zmiennej może składać się z dowolnych liter lub cyfr, nie może jednak zaczynać się od cyfry. Sugeruje się, by zmienne nazywać jak najprościej i jak najkrócej tworząc z ich nazw logiczne ciągi, co pozwala na ich lepsze zapamiętanie i sprawniejsze korzystanie ze zbioru danych, na przykład: V1, V2, V3, P1, P2, M1, M2, nr, rok, wiek, płeć, itd. itp. Nazwy te mogą zawierać polskie znaki.

Kolumna 3. Typ zmiennej (*Type*). Najogólniej ujmując typ zmiennej określa czy dana zmienna jest zmienną tekstową czy liczbową. Standaryzuje również jej format - sposób prezentacji, maksymalną liczbę cyfr, z ilu może się składać oraz liczbę miejsc po przecinku. Określa, czy zmienna jest liczbą naturalną, datą lub zmienną walutową. **Najczęściej wykorzystywanymi typami zmiennych są *Numeric* (numeryczna) oraz *String* (ciąg znaków)**, znaczenie pozostałych jest marginalne. Zakładkę, umożliwiającą wybranie typu zmiennej, uzyskujemy klikając na daną komórkę w kolumnie *Type,* a następnie na szary prostokąt po prawej stronie komórki:



Po kliknięciu pojawi się okno (*pop-up window*), które pozwala na wybór formatu danych. W programie PSPP możliwe jest zapisanie zmiennej w następujących formatach:



Poniżej zostały scharakteryzowane pola z zakładki *Variable Type* (typ zmiennej):

- *Numeric* (numeryczna) - zmienna liczbowa (na przykład: 4, 7, 13). Jest to najczęściej używany w analizie danych politologicznych typ zmiennej;

- *Comma* - liczba z częścią dziesiętną oddzieloną przecinkiem (np. 4,57). W analizie danych jest to często wykorzystywany typ zmiennej;

- *Dot* - liczba z częścią dziesiętną oddzieloną kropką (np. 5.88);

- *Scientific notation* (notacja naukowa lub inaczej postać wykładnicza) - składająca się ona ze znaku liczby (-/+), liczby ułamkowej (znormalizowanej mantysy), małej lub wielkiej litery E, liczby całkowitej (wykładnika lub cechy dziesiętnej). Na przykład liczba 10 w zapisie naukowym ma postać: 1,0e1. Ten typ zmiennej nie jest na ogół wykorzystywany w politologicznej analizie danych;

- *Date* (data) - którą można zapisać w predefiniowanej postaci uwzględniającej sekundy, minuty, godziny, dni, miesiące i lata, dostępna w 18 postaciach;

- *Dollar* (dolar) - zapis walutowy, wartość poprzedzona znakiem $, dostępna w 12 predefiniowanych postaciach;

- *Custom currency* (własna waluta) - umożliwiająca zdefiniowanie własnych wartości pieniężnych;

- *String* (ciąg znaków) - dowolny zbiór znaków: cyfr, liter, znaków specjalnych, spacji. Nie jest interpretowany przez PSPP jako liczba. Służy do zapisywania wartości tekstowych na ogół pochodzących z pytań otwartych, gdzie respondent wypowiada się pełnym zdaniem lub podaje niekodowane nazwy. Czasami wykorzystywany w politologicznej analizie danych. W celu przeprowadzenia analizy ilościowej na ogół wymagane jest przeprowadzenie procesu kodowania danych.

Kolumna 4. Liczba pól przypisana wartości zmiennej (*Width*). Maksymalna liczba znaków (kolumn), z jakich może się składać dana zmienna. Należy mieć na uwadze, że zmniejszenie tej liczby powoduje automatyczne odcięcie (skrócenie) wartości zmiennych od prawej strony, bez możliwości przywrócenia wartości wyjściowych. Zmiana ta jest widoczna w Widoku danych.

Kolumna 5. Liczba miejsc dziesiętnych wartości zmiennej (*Decimals*). Określa dowolną liczbę miejsc po przecinku zmiennej numerycznej.

Kolumna 6. Etykieta zmiennej (*Label*). Jest to opis zmiennej. Może mieć postać pytania kwestionariuszowego lub - lepiej - jego równoważnika.

Kolumna 7. Wartości zmiennej (*Values*). Zakres potencjalnych wartości, jakie sensownie może przyjmować dana zmienna. Określenie wartości zmiennych ma szczególne znaczenie w przypadku zmiennych jakościowych: nominalnych i porządkowych. Zmienne tego typu mają o tyle sens, o ile poszczególne cyfrowe wartości zmiennej zostały opatrzone kategoriami pojęciowymi. Mniejsze znaczenie mają one w przypadku poziomów interwałowego i ilorazowego, bowiem w odniesieniu do tych zmiennych liczba jest jednocześnie jej znaczącym opisem i wówczas w tej kolumnie opisów nie stosuje się.

Kolumna 8. Brakujące wartości (*Missing*). W tej kolumnie oznacza się wartości wykluczone z udziału w analizach. Są to na ogół takie odpowiedzi badanych jak „nie wiem, trudno powiedzieć" lub odmowy podania odpowiedzi. Oznaczające je liczby wpisujemy w tę kolumnę. Rozważmy przykładową zmienną rok urodzenia, w której w większości przypadków uzyskaliśmy konkretne, liczbowe odpowiedzi respondentów, jednak w kilku przypadkach respondenci odmówili podania swojego wieku. Odpowiedzi te zakodowaliśmy jako 9998. Analizy takie jak obliczanie średniej wartości wieku wszystkich badanych z użyciem tej liczby byłyby nielogiczne. W związku z tym kod 9998 wpisujemy w tę kolumnę.

Kolumna 9. Liczba widocznych kolumn (*Columns*). Pełni funkcję estetyczną i wizualną. Jest to szerokość pola, w którym umieszczana jest zmienna w Widoku Danych (*Data View*). W przeciwieństwie do manipulacji czwartą kolumną dokonanie tu zmian nie powoduje utraty danych - manipulujemy wyłącznie widokiem danych.

Kolumna 10. Wyrównanie danych (*Align*). Wyrównanie zmiennej w Widoku Danych (*Data View*): *left* (do lewej), *center* (do środka), *right* (do prawej).

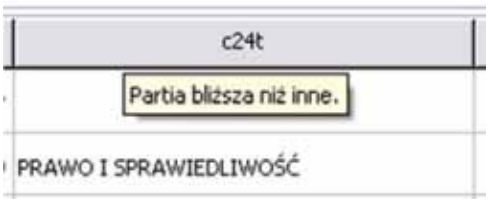
Kolumna 11. Typ skali (*Measure*). Poziom na jakim mierzona jest dana zmienna: *nominal* (nominalnym), *ordinal* (porządkowym), *scale* (ilorazowym lub interwałowym). Zagadnienie poziomu pomiaru szeroko opisano w poprzednich podrozdziałach. Patrz ⇨ Skale pomiaru zmiennych.

5.4.2. Widok zbioru danych (*Data View*)

Widok zbioru danych przedstawia macierz danych surowych. Znajdują się tu uporządkowane wszystkie jednostki analizy. Zwykle widoczny jest tylko fragment zbioru danych, niewidoczne fragmenty możemy obejrzeć używając poziomych i pionowych pasków przewijania.

	numer	px1dd	px1mm	px1r	gw	mw	c1t
20	1351	24	11	2007	14	20	RENTY I EMERYTURY
21	1532	24	11	2007	17	18	POPRAWA ŻYCIA CODZIENNEGO
22	1631	27	11	2007	-1	-1	ŻEBY IŚC NA WYBORY
23	117	12	11	2007	10	15	GOSPODARKA
24	273	19	11	2007	17	0	NIE CZUĆ SIĘ ZAGROŻONYM DZIAŁALNOŚCIĄ SŁUŻB
25	282	17	11	2007	13	20	ŻADNA
26	617	15	11	2007	19	55	ZAGŁOSOWANIE NA PO
27	618	15	11	2007	17	20	ŻEBY SIĘ W POLSCE ZMIENIŁO NA LEPSZE

Każdy wiersz zawiera wartości dla jednej i tylko jednej jednostki analizy. W kolumnach natomiast znajdują się poszczególne zmienne. Kolejność kolumn ze zmiennymi jest tożsama z kolejnością zmiennych w zakładce Widok zmiennych. Pierwszy wyszarzony wiersz znajdujący się pod menu zawiera nazwy zmiennych. Po najechaniu kursorem myszy na pole z nazwą zmiennej pokazuje się jako *yellow label* etykieta zmiennej:



Dwukrotne kliknięcie na to pole przenosi nas do Widoku zmiennych - konkretnie do pełnego opisu zmiennej, na której nazwę kliknęliśmy. W komórkach macierzy mogą znajdować się zarówno wartości liczbowe jak i tekstowe. Gwiazdki pojawiają się tam, gdzie wartość nie może być wyświetlona, ponieważ wprowadzone dane są zbyt długie. Pomimo zmiany liczby pól przypisanych wartości zmiennej (Kolumna 4), wprowadzone do komórki dane tracone są bezpowrotnie. Z kolei kropki w danej komórce oznaczają brak danych. Jest to uniwersalny sposób kodowania braków danych, nie tylko w programach, ale również w raportach - uważa się, że pola tabeli raportu, gdzie dane nie mogą się sensownie pojawić należy oznaczać kropką.

Po kliknięciu prawym przyciskiem myszy na komórkę macierzy pojawia się następujące *pop-up window*:

Wiersz *Input Methods* pozwala na wprowadzanie wartości zmiennych z użyciem znaków diakrytycznych w alfabetach łacińskich (np. cedylli używanej w francuskim, portugalskim, katalońskim, tureckim), a także innych alfabetów - między innymi amharskiego, cyrylicy, inuktitut, tigrigny, wietnamskiego, a także międzynarodowego zapisu fonetycznego (*International Phonetic Alphabet*, IPA). Jest to szeroko stosowany, blokowo implementowany element wielu aplikacji nazywany edytorem metody input (*Input Method Editor*), pozwalający wprowadzać znaki i symbole w różnych alfabetach za pomocą standardowej klawiatury. Jeśli nie analizujemy zbiorów danych w innych językach, opcje te mają marginalne znaczenie.

Z kolei opcje zawarte w wierszu *Insert Unicode Control Character* to znaki specjalne (tzw. kody sterujące) służące do wydawania komend urządzeniom peryferyjnym: drukarkom, terminalom lub modemom. Znaki te są powszechnie używane w programach służących do składu i obróbki publikacji (*Desktop Publishing,* DTP). Program PSPP oferuje następujące kody sterujące:

- LRM (*Left Right Mark*), RLM (*Right to Left Mark*), LRE (*Left to Right Embedding*), RLE (*Right to Left Embedding*) - które umożliwiają zmianę kierunku wprowadzania i wyświetlania tekstu w komórkach Widoku danych. Kody sterujące LRM i LRE przeznaczone są dla systemów zapisu znaków od lewej do prawej (język polski, angielski, rosyjski, itd.), z kolei RLM i RLE stosowane są w przypadku odwrotnego kierunku zapisu od prawej do lewej (język hebrajski, perski, arabski, itd.);

- ZWJ (*Zero Width Joiner*) - wstawia łącznik zamieniając odrębne litery na pismo ciągłe (łącząc je w wyraz). Używany jest w rodzinie języków abugida oraz arabskim, perskim, urdu;

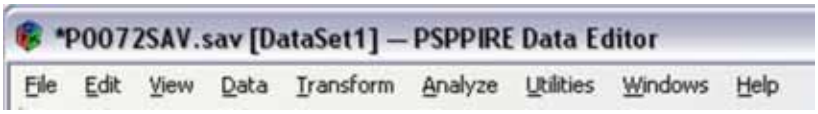
- ZWNJ (*Zero Width Non Joiner*) - komenda tworząca ligatury, a więc połączenia dwóch liter. Obecnie ligatury zanikły w językach narodowych, w których były używane, mają one znaczenie historyczne, pojawiają się w dawniejszych tekstach;

- ZWSP (*Zero Width Space*) - tworząca niewidoczne dla Czytelnika, lecz widoczne dla urządzenia odstępy (spacje) w tekście.

Kody sterujące są znakami niedrukowalnymi, należy je rozumieć raczej jako swoiste formatowania. Są one niewidoczne, dlatego zaleca się szczególną ostrożność przy ich używaniu.

5.5. Menu główne programu PSPP

Menu główne programu PSPP jest dostępne i identyczne w zakładkach Widoku zmiennych (*Variable View*) i Widoku danych (*Data View*). Składa się ono z dwóch elementów - znajdującego się na samej górze okna programu PSPP menu tekstowego:



oraz znajdującego się poniżej menu składającego się z ikon (menu ikoniczne) zawierającego najczęściej wykorzystywane funkcje programu PSPP:

Poniżej znajduje się omówienie umożliwiające ogólną orientację w możliwościach programu PSPP. Szerzej zostały omówione te funkcje, które nie są przedmiotem rozważań w pozostałych rozdziałach lub mają istotne znaczenie dla poznania podstawowej mechaniki programu. W menu tekstowym znajdują się następujące opcje:

1/ Plik (*File*) - najważniejsze elementy tej części menu zostały już omówione w poprzednich rozdziałach lub są samozrozumiałe. Ponadto warto zwrócić uwagę na następujące elementy, które mogą okazać się przydatne:

- *File* ⇨ *New* ⇨ *Syntax* - otwiera okno Edytora składni (*Syntax Editor*), umożliwiające wydawanie programowi PSPP poleceń w trybie tekstowym.

- *File* ⇨ *New* ⇨ *Data* - otwiera nowe okno główne programu PSPP z czystą macierzą danych.

- *File* ⇨ *Display Data File Information* - wyświetla komplet informacji na temat zbioru danych w Oknie raportów (*Output Viewer*). Może wyświetlać informacje o bieżącym, aktualnie otwartym zbiorze danych (*Working File*) lub wskazanych przez użytkownika (*External File*). Oto przykładowy fragment wyświetlanej po użyciu tej opcji tabeli:

c7	Czy to, kto rządzi, ma znaczenie? Format: F1.0 Measure: Scale Display Alignment: Right Display Width: 10	19
	1 To, kto rządzi ma duże znaczenie 5 To, kto rządzi nie ma żadnego znaczenia 7 Trudno powiedzieć	
c8	Czy to, na kogo się głosuje w wyborach może coś zmienić? Format: F1.0 Measure: Scale Display Alignment: Right Display Width: 9	20
	1 To na kogo się głosuje i tak niczego nie zmieni 5 To na kogo się głosuje może wiele zmienić 7 Trudno powiedzieć	

- *File* ⇨ *Recently Used Data* - umożliwia otworzenie ostatnio używanych plików zbiorów danych (*.sav, *.por). Wyświetla je w porządku chronologicznym.

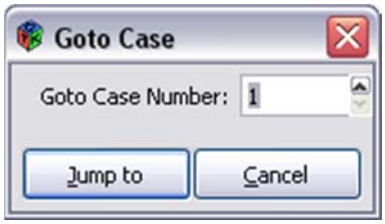
- *File* ⇨ *Recently Used Files* - umożliwia otworzenie ostatnio używanych plików składni (*.sps).

2/ Edycja (*Edit*) - zawiera podstawowe funkcje edycyjne, analogiczne do tych występujących w arkuszach kalkulacyjnych lub edytorach danych. Ponadto występują tu funkcje specyficzne dla PSPP:

- *Edit* ⇨ *Insert Variable* - umożliwia wstawienie pustej zmiennej (wiersza w Widoku zmiennych lub kolumny w Widoku danych). Opcja ta ma swój odpowiednik w postaci ikony.

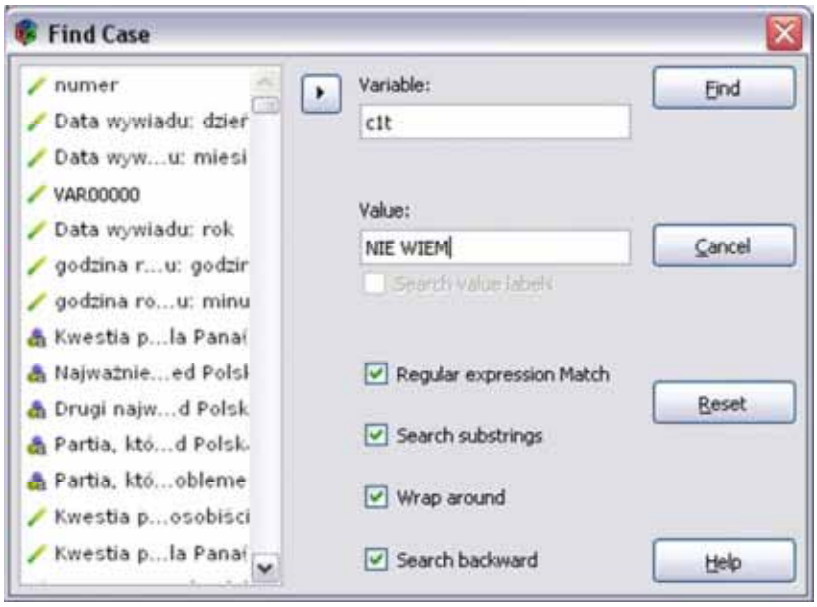
- *Edit* ⇨ *Insert Cases* - umożliwia dodanie pustego wiersza (jednostki analizy) w Widoku danych. Opcja ta ma swój odpowiednik w postaci ikony.

- *Edit* ⇨ *Go To Case* - pozwala na przejście do jednostki analizy o podanym numerze w zbiorze (w zakładce *Data View*):



Opcja ta ma swój odpowiednik w postaci ikony.

- *Edit* ⇨ *Find Case* - funkcja przeszukiwania zbioru danych. Możliwe jest odnajdywanie zarówno zmiennych liczbowych, jak również tekstowych. Zmienną, która ma być przeszukiwana dla każdej z jednostek analizy podajemy w polu *Variable,* a poszukiwaną wartość w polu *Value.* W przypadku zmiennych tekstowych możliwe są jeszcze następujące opcje: poszukiwania wyłącznie wprowadzonego tekstu (*Regular Expression Match*), poszukiwanie fragmentu tekstu (*Search substrings*), a także przeszukiwanie całego zbioru (program PSPP domyślnie przeszukuje od zaznaczonego miejsca aż do końca zbioru - „w dół") - *Wrap around* oraz przeszukiwanie wstecz (do początku zbioru, czyli „w górę"). W tym przypadku, co prezentuje poniższy rysunek, program PSPP będzie kolejno przeszukiwał i przechodził do komórek, które w zmiennej c1t przyjmują wartość dokładnie „NIE WIEM". Opcja ta ma swój odpowiednik w postaci ikony.



3/ Widok (*View*) - zawiera następujące opcje:

- Pasek stanu (*Status Bar*) - uaktywnia (jeśli zaznaczony) lub dezaktywuje tak zwany pasek stanu. Pasek stanu znajduje się na dole okna programu PSPP. Informuje on między innymi czy zbiór danych jest ważony czy nie (*Weight by ...* versus *Weight off*), czy zmienne są filtrowane (*Filter by...* versus *Filter off*) lub czy zbiór danych jest analizowany w grupach czy nie (*Split by...* versus *Split off*). Na pasek stanu należy zwracać szczególną uwagę podczas analizy danych. Brak tego nawyku skutkuje popełnianiem błędów podczas analiz i koniecznością powtarzania obliczeń.

- Czcionka (*Font*) - pozwalającą wybrać preferowany krój, rozmiar i formatowania czcionki dla zmiennych i jednostek analizy.

- Siatka podziału komórek (*Grid Lines*) - oddzielająca poszczególne komórki w widoku zmiennych i widoku danych. Ma znaczenie wyłącznie wizualne, wpływa na wygodę użytkowania programu.

- Widok wartości lub etykiet zmiennych (*Value Labels*). Domyślnym sposobem prezentacji wartości zmiennych w widoku danych są wartości liczbowe. Za pomocą opcji *Value Labels* można uzyskać widok etykiet zmiennych (pod warunkiem, że zostały one uprzednio zdefiniowane i włączone do zbioru danych). Opcja ta ma swój odpowiednik w postaci ikony.

- Dane (*Data*) - przełącza do Widoku danych (*Data View*)

- Zmienne (*Variables*) - przełącza do Widoku zmiennych (*Variable View*)

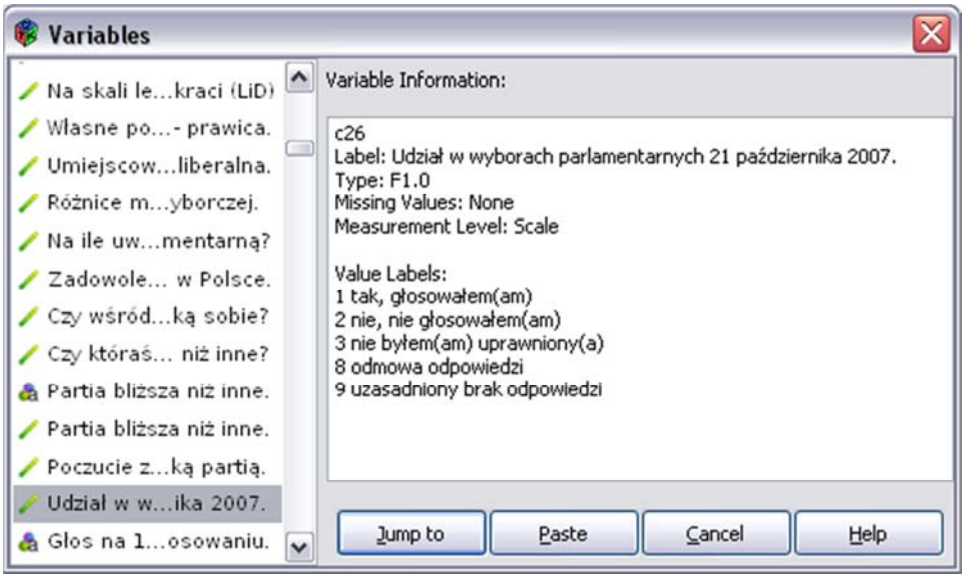
4/ Dane (*Data*) - jest to element zawierający procedury przekształcania danych. Szczegółowo omówiono go w dalszych częściach publikacji.

5/ Przekształcenia (*Transform*) - zawarte tu opcje umożliwiają przekształcenia zmiennych. Są one przedmiotem szczegółowych rozważań w kolejnych podrozdziałach.

6/ Analizy (*Analyze*) - kluczowy element menu. Za pomocą opcji zawartych w tej zakładce dokonujemy wszelkiego rodzaju analiz danych - od prostych zestawień tabelarycznych aż po średniozaawansowane analizy statystyczne.

7/ Narzędzia (*Utilities*) - zawierają dwie następujące opcje:

- Informacja o zmiennych (*Variables*) - umożliwia przeglądanie istotnych wartości definiujących poszczególne zmienne. Przeglądanie odbywa się w następującym oknie:



Opcja ta ma swój odpowiednik w postaci ikony.

- Informacja o zbiorze danych (*Data File Comments*) - umożliwia samodzielne wprowadzenie opisu dla zbioru danych. Opis ten może być wyświetlany w Oknie raportów, jeśli po wpisaniu komentarza zaznaczymy *Display comments in output.*

8/ Okna (*Windows*) - umożliwia zminimalizowanie wszystkich okien programu PSPP (*Minimize*), opcję ułatwionego przeglądania okien poprzez podzielenie ich z zakotwiczeniem widoku w pionie lub poziomie (*Split*), a także umożliwia przełączenie pomiędzy otwartymi oknami programu PSPP.

9/ Pomoc (*Help*) - zawiera informacje o wersji i Autorach programu PSPP oraz dane licencyjne. Umożliwia także wyświetlenie podręcznika programu PSPP (nie działa w prezentowanej wersji programu).

Menu tekstowe i menu ikoniczne nie wyczerpują możliwości programu PSPP. Komendy dla tego programu można wydawać również w formie tekstowej, używając do tego celu okna Edytor składni (*Syntax Editor*) i posługując się standaryzowanymi komendami. W Oknie składni można dokonywać tych samych czynności, co w menu, a dodatkowo wydawać mu polecenia niedostępne z poziomu menu tekstowego. Uzasadnieniem używania Okna składni jest także dokumentowanie całego przebiegu pracy nad zbiorem danych i w razie potrzeby zautomatyzowane odtworzenie całości lub części dokonywanych przekształceń.

5.6. Widok okna składni (*Syntax Editor*)

Operacje przetwarzania i analiz zbiorów danych w programie PSPP mogą być dokonywane nie tylko za pomocą menu, lecz również za pomocą Edytora składni (*Syntax Editor*). W oknie składni, uzyskiwanym po wybraniu w menu tekstowym *File* ⇨ *New* ⇨ *Syntax* lub po otworzeniu pliku składni *.sps, wpisujemy polecenia dla programu PSPP wydawane za pomocą języka skryptowego. Składnia poleceń PSPP umożliwia automatyzację pracy, jej dokumentowanie, używanie napisanych programów do innych zbiorów danych lub tych samych ponownie. Ułatwia i przyspiesza powtarzalną pracę. Część opcji programu PSPP niedostępna jest w menu i można z nich skorzystać wyłącznie za pomocą składni. Są to na

ogół zaawansowane analizy. Fragmenty składni można uzyskiwać za pomocą trybu okienkowego - w menu. Wystarczy w toku wykonywania jakiejś operacji kliknąć na przycisk Wklej (*Paste*). Wówczas kod zaplanowanej w trybie okienkowym operacji przenosi się w postaci składni do Edytora składni.

Program PSPP umożliwia uruchomienie wpisanych w Edytorze składni poleceń w rozmaitych trybach:

1/ po wybraniu z menu *Run* ⇨ *All* - całości znajdującego się tam zapisu,

2/ po wybraniu z menu *Run* ⇨ *Selection* - zaznaczonego fragmentu programu,

3/ po wybraniu z menu *Run* ⇨ *Current Line* - tej linii, w której aktualnie znajduje się kursor (skrót klawiszowy Ctrl + R),

4/ po wybraniu z menu *Run* ⇨ *To End* - partii od miejsca, w którym znajduje się kursor, aż do końca programu.

W Edytorze składni można stosować skróty poleceń, na przykład zamiast pełnego zapisu:

```
FREUENCIES /VARIABLES= VAR001.
```

możemy zastosować

```
FREQ /VAR= VAR001.
```

Każde polecenie należy kończyć kropką (.). Jest to znak oznaczający „wykonaj zmiany". Brak kropki nie skutkuje niewykonaniem polecenia, lecz zmiany mogą nie być dla użytkownika widoczne. Każdy program należy kończyć komendą:

```
EXECUTE.
```

Język programowania PSPP jak wszystkie języki programowania jest wrażliwy na składnię - trzeba zwracać szczególną uwagę na właściwy zapis komend. Szczególne znaczenie mają spacje. Nie ma natomiast znaczenia wielkość liter - wyżej przytaczana składnia zostanie wykonana, jeśli zostanie zapisana następująco:

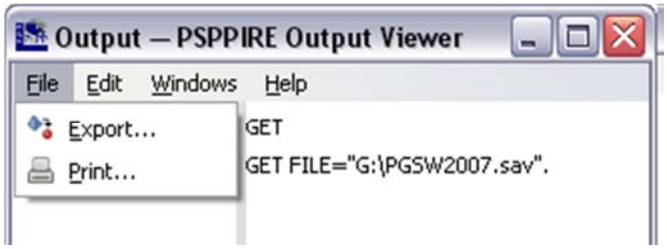
```
freq /var= VAR001.
```

W przypadku popełnienia pomyłki i próby wykonania wadliwego programu w Oknie raportów, pojawi się komunikat o błędzie. Zaleca się ćwiczenie od początku pracy z PSPP w trybie składni. Wprowadzanie w zasady posługiwania się składnią jest systematycznie prowadzone w dalszych częściach publikacji.

5.7. Widok okna raportów (*Output Viewer*)

W widoku okna raportów zamieszczane są wszelkie generowane przez użytkownika analizy oraz informacje o dokonanych przekształceniach zbioru danych lub zmiennych, a także komunikaty o błędach (na przykład o próbie wymuszenia wykonania nieprawidłowego polecenia). Widok okna raportów pojawia się jako samodzielne okno. Otwiera się on automatycznie w momencie wykonania pierwszej operacji na zbiorze danych - a więc przy jego otwarciu.

Wszystkie informacje wyświetlane w Widoku okna raportów są nieedytowalne - nie można wprowadzać w nich jakichkolwiek zmian, zaznaczać lub bezpośrednio kopiować. Można w nie ingerować dopiero po ich zapisaniu w odpowiednim formacie i otworzeniu w innym programie. W tym celu należy wybrać z menu tekstowego Widoku okna raportów *File* ⇨ *Export*:



Pojawi się okno, w którym można nadać nazwę zapisywanemu plikowi oraz wybrać format, w jakim ma być on wyeksportowany. Do wyboru pozostawiono następujące formaty zapisu:

- PDF (*.pdf) - w tym formacie na ogół zapisujemy plik w sytuacji, gdy nie zamierzamy edytować pliku lub chcemy swobodnie przenosić go (ze względu na mały rozmiar i uniwersalny format) pomiędzy różnymi platformami. Plik *.pdf można jednak edytować w programach typu DTP (*Desktop Publishing*), służących do składu publikacji, na przykład w darmowym programie Scribus. Domyślnie, na typowym komputerze zaopatrzonym w Microsoft Windows, pliki *.pdf będą otwierane w programie Acrobat Reader. Zapisany w formacie *.pdf plik zachowuje wizualną identyczność z Widokiem okna raportów (układ tekstu, tabel i wykresów),

- HTML (*.html) - zapisuje plik gotowy do opublikowania jako strona www. Pliki *.html mogą być edytowane w programach służących do projektowania stron internetowych w trybie WYSIWYG (z ang. *What You See Is What You Get* - To Co Widzisz Jest Tym Co Uzyskasz), na przykład WYSIWYG Web Builder lub Nvu (obydwa darmowe) lub w programach służących do edycji kodu HTML takich jak np. Ager Web Edytor (również na licencji GNU/GPL). Pliki te można także otwierać i modyfikować w edytorach tekstu takich jak Microsoft Word czy OpenOffice Writer. Zapisany w formacie *.html plik zachowuje formatowania, tabele i wykresy. Grafikę zapisuje w uniwersalnym formacie *.png, jednak zapis rysunków z programu PSPP wymaga jeszcze dopracowania (wersja 0.7.9.),

- OpenDocument (*.odt) - jest to tekstowy dokument OpenOffice Writer. Można go także otwierać w programie Microsoft Word po pobraniu odpowiedniego dodatku ze strony Microsoft. W tym formacie zapisywany jest tekst oraz tabele, lecz nie wykresy lub rysunki,

- Text (*.txt) - zwykły dokument tekstowy, zapisujący tekst bez formatowań, rysunków i wykresów,

- Postscript (*.ps) - do tego typu pliku odnoszą się uwagi sformułowane dla plików *.pdf. Dodatkowo do otwierania i edycji plików postscriptowych można używać darmowych programów takich jak na przykład Ghostscript i Evince,

- Comma-Separated Values (*.csv) - plik zapisujący wartości oddzielone przecinkami. Umożliwia przenoszenie pliku bez formatowań i rysunków.

Jeśli zależy nam na umieszczeniu takiego wykresu w raporcie, najlepiej posłużyć się skrótem klawiszowym *ALT + Prt Scr* (jednoczesne naciśnięcie klawisza ALT oraz Print Screen). Otrzymamy wówczas tak zwany zrzut ekranu, który należy przenieść do dowolnego programu graficznego (poprzez wklejenie za

pomocą skrótu klawiszowego (Ctrl + V), a następnie przyciąć i zapisać. Plik taki można zamieścić jako rysunek w tworzonym w edytorze tekstu raporcie.

6

Rozdział 6. Czy zmienne na skali w kwestionariuszu mierzą tak samo? Badanie rzetelności skali

Zdarza się, że pomiar za pomocą pojedynczej zmiennej jest dla badacza niewystarczający ze względu na złożoność analizowanego problemu. Konstruowane są wówczas składające się z wielu zmiennych wskaźniki badanych zjawisk. Mogą być to skale lub indeksy. Indeks jest prostą sumą poszczególnych zmiennych otrzymaną przez dodanie liczb wyrażających odpowiedzi respondentów na poszczególne pytania. Z kolei skala zbudowana jest ze zmiennych, którym badacz przypisuje nierówną wartość wskazując, że niektóre z nich są bardziej a inne mnie intensywnymi wskaźnikami badanego zjawiska; dokonuje zatem zabiegu ważenia poszczególnych zmiennych składających się na skalę. Zasady budowania indeksów i skal wykraczają poza zakres tematyczny niniejszej publikacji, zostały one opisane szeroko w literaturze przedmiotu[1].

Składające się na indeks lub skalę poszczególne zmienne powinny mierzyć to samo zjawisko, badać ten sam konstrukt teoretyczny, być maksymalnie spójne. Spójność indeksu lub skali nazywamy rzetelnością lub wiarygodnością. Im skala bardziej rzetelna, w tym większym stopniu możemy być przekonani, że daje ona takie same rezultaty dla kolejno dokonywanych pomiarów, jest on odporny na działanie czynników zewnętrznych. Nie należy oczywiście utożsamiać rzetelności z tym, że badacz rzeczywiście mierzy daną cechę. To odnosi się do pojęcia trafności danego testu. Rzetelność oznacza dokładność pomiaru oraz spójność pozycji wchodzących w skład danego konstruktu i nie należy jej utożsamiać z tym, czy rzeczywiście dana skala mierzy interesującą badacza cechę.

Ideę rzetelności wyjaśnić można na następującym przykładzie. Przypuśćmy, że badamy za pomocą ułożonej przez nas skali lub indeksu poziom autorytaryzmu grupy ludzi należących do ugrupowania skrajnie prawicowego. Jeśli wkrótce dokonamy kolejnego pomiaru na tej samej grupie lub też na grupie innych,

[1] Zainteresowanych tym zagadnieniem zachęcamy do zapoznania się z następującymi, podstawowymi pozycjami wprowadzającymi w tą problematykę: R. Mayntz, K. Holm, P. Hübner, *Wprowadzenie do metod socjologii empirycznej*, Wydawnictwo Naukowe PWN, Warszawa 1985, Rozdział II. *Pomiar,* A.N. Oppenheim, *Kwestionariusze, wywiady, pomiary postaw*, Wydawnictwo Zysk i S-ka, Poznań 2004, s. 203-239, E. Babbie, *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN, Warszawa 2004, s. 190-196.

losowo wybranych osób z tego lub innego skrajnie prawicowego ugrupowania powinniśmy wówczas uzyskać identyczne wyniki (zakładając, że w czasie pomiędzy pomiarami nie zaszła w badanych poważna duchowa zmiana).

Rzetelność mierzymy na wczesnym etapie przygotowywania badania, najczęściej obliczana jest ona na podstawie danych zebranych podczas pilotażu.

6.1. Przegląd sposobów badania rzetelności skali

Rzetelność indeksu lub skali możemy sprawdzić na kilka sposobów:

1/ **Metoda wielokrotnego testowania** polega na stosowaniu tego samego indeksu lub skali w tej samej grupie w różnych punktach czasu. Jeśli wyniki dwóch lub więcej pomiarów są tożsame lub zbliżone świadczy to o wysokiej rzetelności skali. Jeśli wyniki różnią się możemy na tej podstawie wnioskować, że skala jest nierzetelna lub, że pomiędzy pomiarami zaszły zjawiska zmieniające postawy badanych.

2/ **Metoda porównywania alternatywnych mierników** zjawiska polega na wykorzystaniu dwóch indeksów lub skal w dwóch różnych punktach czasowych i porównywaniu uzyskiwanych wyników. Metoda ta zapobiega wpływowi pierwszego pomiaru na drugi.

3/ **Metoda sędziów kompetentnych** polegająca na poddaniu ocenie skali grupie osób, które posiadają wystarczającą wiedzę w zakresie tematyki objętej badaniem. Ocena sędziów kompetentnych może mieć charakter jakościowy lub ilościowy.

4/ **Metoda badania korelacji pomiędzy poszczególnymi elementami składającymi się na indeks lub skalę**. Pomiar taki przeprowadza się używając współczynnika alfa opracowanego przez Lee J. Cronbacha, amerykańskiego psychologa edukacyjnego. Alfa Cronbacha stał się bardzo popularny ze względu na prostotę jego obliczania i uniwersalność[2]. Współczynnik ten jest elementarną i najbardziej rozpowszechnioną miarą. Istnieją również bardziej zaawansowane sposoby badania rzetelności skali: na przykład w celu sprawdzenia czy dwa zbiory zmiennych mierzą ten sam konstrukt lub pojęcie można użyć modelowania równań strukturalnych (*Structural Equation Modelling and Path Analysis*, SEPATH); wykorzystuje się także do tego celu analizę czynnikową.

Współczynnik alfa Cronbacha obliczamy następująco:

$$\alpha = \frac{k}{k-1} * \frac{s^2_{suma} - (s^2 i_1 + s^2 i_2 + \cdots + s^2 i_n)}{s^2_{suma}}$$

k - liczba pytań w indeksie lub skali
$s^2 i_1 + s^2 i_2 + \ldots + s^2 i_n$ - suma wariancji poszczególnych pytań
s^2_{suma} - suma wariancji wszystkich pytań

[2] L.J. Cronbach, *Coefficient alpha and the internal structure of tests*, „Psychometrika", 1951, 16, s. 297-334.

Spróbujmy zastosować współczynnik alfa Cronbacha w praktyce. W Tabeli 4 przedstawiono przykładowy indeks złożony z czterech zmiennych. Zakres każdej z tych zmiennych zawiera się pomiędzy 0 a 10 (dla uproszczenia wykorzystano wyłącznie liczby 5 i 10). Zbadano dwóch respondentów[3].

Tabela 4. Przykładowy indeks składający się z czterech zmiennych

	zmienna 1	**zmienna 2**	**zmienna 3**	**zmienna 4**
Respondent 1	5	5	5	5
Respondent 2	5	10	10	5

Wyznaczenie współczynnika alfa Cronbacha rozpoczynamy od obliczenia średnich i wariancji dla poszczególnych zmiennych.

Średnia dla zmiennej 1:

$$x_1 = \frac{5+5}{2} = 5$$

Średnia dla zmiennej 2:

$$x_2 = \frac{5+10}{2} = 7{,}5$$

Średnia dla zmiennej 3:

$$x_3 = \frac{5+10}{2} = 7{,}5$$

Średnia dla zmiennej 4:

$$x_4 = \frac{5+5}{2} = 5$$

Wariancja dla zmiennej 1:

$$\text{wariancja}_1 = \frac{(5-5)^2 + (5-5)^2}{2} = 0$$

Wariancja dla zmiennej 2:

$$\text{wariancja}_2 = \frac{(5-7{,}5)^2 + (10-7{,}5)^2}{2} = 6{,}25$$

Wariancja dla zmiennej 3:

$$\text{wariancja}_3 = \frac{(5-7{,}5)^2 + (10-7{,}5)^2}{2} = 6{,}25$$

[3] Jest to naturalnie zbyt mała liczba jednostek analizy dla zbadania rzetelności skali. Uproszczoną macierz wprowadzono w celach dydaktycznych. Minimalna liczba jednostek analizy określana jest na dwa sposoby. Niektórzy za wystarczające uznają co najmniej 30 zbadanych przypadków, inny postulują, by zbadać co najmniej 5 proc. zakładanej próby.

Wariancja dla zmiennej 4:

$$\text{wariancja}_4 = \frac{(5-5)^2 + (5-5)^2}{2} = 0$$

Na podstawie powyższych cząstkowych wyników obliczamy sumę wariancji poszczególnych pytań. Wynosi ona: $0 + 6{,}25 + 6{,}25 + 0 = 12{,}5$.

Następnie obliczamy sumę wariancji wszystkich pytań. Rozpoczynamy od zsumowania wskazań respondentów oraz średniej wskazań wszystkich respondentów.

Suma wskazań Respondenta 1:
$R_1 = 5 + 5 + 5 + 5 = 20$
Suma wskazań Respondenta 2:
$R_2 = 5 + 10 + 10 + 5 = 30$

Następnie obliczamy średnią wskazań wszystkich respondentów. Wynosi ona:

$$R_{\text{średnia}} = \frac{20 + 30}{2} = 25$$

Znając powyższe wartości możemy ostatecznie obliczyć sumę wariancji wszystkich pytań:

$$s^2_{\text{suma}} = \frac{(20-25)^2 + (30-25)^2}{2} = 25$$

Po obliczeniu powyższych wartości podstawiamy je do wzoru na alfa Cronbacha:

$$\alpha = \frac{4}{4-1} * \frac{25 - (25 - 12{,}5)}{25} = 1{,}3(3) * 0{,}5 = 0{,}67$$

Zasady interpretacji tego wyniku omówiono w dalszej części tekstu.

5/ **Metoda połówkowa** polegająca na podzieleniu indeksu lub skali na dwie równoważne połowy i zbadanie dwóch tożsamych grup w tym samym czasie. Jeśli wyniki są tożsame lub podobne można powiedzieć o skali, że jest rzetelna. W metodzie połówkowej używane są formuła Spearmana-Browna[4] oraz Guttmana[5] służące do ilościowej oceny rzetelności skali.

Współczynnik Spearmana-Browna obliczany jest następująco:

$$R = \frac{2R_p}{1 + R_p}$$

,gdzie R_p to współczynnik korelacji pomiędzy dwoma połówkami testu.

Z kolei współczynnik Guttmana wyrażany jest wzorem:

[4] Współczynnik ten wynaleźli niezależnie od siebie W. Brown i K. Pearson - W. Brown, *Some experimental results in the correlation of mental abilities*. „British Journal of Psychology", 1910, 3, s. 296-322, C. Spearman, *Correlation calculated from faulty data*. „British Journal of Psychology", 1910, 3, s. 271-295.

[5] L. Guttman, *A basic for analyzing test-retest reliability*, „Psychometrika", 145, 10, s. 255-282.

$$G = 1 - \frac{s^2 i_a + s^2 i_b}{s^2_{suma}}$$

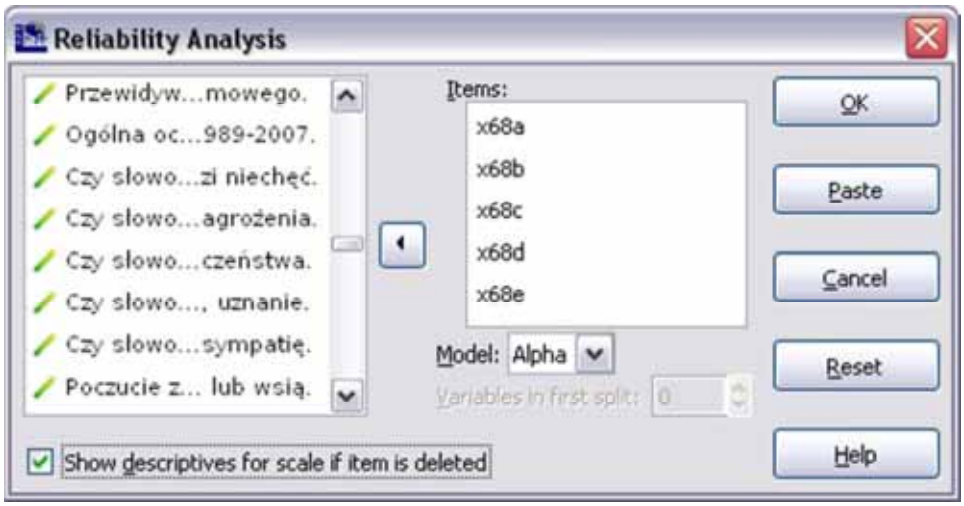
$s^2 i_a + s^2 i_b$ - suma wariancji podzielonych skal

s^2_{suma} - wariancja obu połówek skal

6.2. Obliczanie rzetelności skali w programie PSPP

Program PSPP umożliwia obliczenie rzetelności skali metodą badania korelacji pomiędzy poszczególnymi elementami składającymi się na indeks lub skalę oraz metodą połówkową.

Współczynnik alfa Cronbacha w programie PSPP obliczamy wybierając z menu tekstowego *Analyze ⇨ Reliability.* Pojawia się wówczas następujące okno dialogowe:



W oknie *Items* umieszczamy zmienne, które mają tworzyć skalę lub indeks. Współczynnik alfa Cronbacha oparty jest na korelacji R Pearsona i w związku z tym poszczególne zmienne wchodzące w skład skali lub indeksu powinny być mierzone na poziomach ilościowych. W ten sposób badamy skale lub indeksy składające się co najmniej z trzech zmiennych. Warto zwrócić uwagę, że im większa liczba zmiennych w skali, tym lepiej.

Za przykład obliczeń indeksu posłużą zmienne od x68a do x68e ze zbioru PGSW:

x68a - Niekiedy rządy niedemokratyczne mogą być bardziej pożądane niż demokratyczne
x68b - Dla ludzi takich jak ja nie ma znaczenia, czy rządy są demokratyczne
x68c - W demokracji za dużo jest niezdecydowania i gadania
x68d - Demokracja słabo radzi sobie z utrzymaniem porządku
x68e - W demokracji są problemy, ale jest to lepszy system rządzenia niż każdy inny

W celu obliczenia alfa Cronbacha w rozwijanej liście *Model* wybieramy *Alpha.* Oznaczamy również *Show descriptives for scale if item is deleted.*

Po zaakceptowaniu wybranych opcji w oknie raportu wygenerowane zostaną trzy tabele. Pierwsza z nich (zatytułowana *Case processing summary*) ma charakter podsumowania - podana jest w niej liczba zbadanych przypadków. Druga tabela (*Reliability Statistics*) zawiera najważniejszą

statystykę - wynik testu alfa Cronbacha. W drugiej kolumnie tabeli podawana jest liczba zmiennych w testowanym indeksie lub skali.

Reliability Statistics

Cronbach's Alpha	N of Items
,81	5,00

Wynik testu alfa Cronbacha został skonstruowany na podstawie współczynnika korelacji R Pearsona, zatem interpretujemy tę miarę analogicznie jak ten współczynnik. Alfa Cronbacha przybiera wartość od 0 do 1. Im wynik bliższy jest jedności tym większa jest zgodność poszczególnych składników indeksu lub skali. Szczegółowa interpretacja alfa Cronbacha jest następująca: współczynnik przyjmujący wartość poniżej 0,5 skłania nas do zdecydowanego odrzucenia indeksu lub skali ze względu na brak jej spójności. Z kolei wartości od 0,5 do 0,7 uznajemy za spełniające minimalne wymogi homogeniczności skali i skalę lub indeks warunkowo przyjmujemy jako rzetelny. Wynik powyżej 0,7 jest najczęściej przyjmowaną wartością jako zadowalającej w pomiarach rzetelności w naukach społecznych. O ile alfa zawierające się pomiędzy 0,5 i 0,7 można uznać za wątpliwe, o tyle w przypadku przekroczenia wartości 0,7 większość badaczy nie ma wątpliwości. Z kolei wyniki powyżej 0,8 należy uznać za bardzo dobre. Należy zastrzec, że im mniejsza liczba zmiennych, w tym mniej jesteśmy wymagający wobec współczynnika alfa Cronbacha - przy liczbie od trzech do pięciu zmiennych zadowalają nas niższe wartości alfa Cronbacha - pomiędzy 0,5, a 0,6.

Powyższy wynik wynoszący 0,82 każe uznać współczynnik alfa Cronbacha za więcej niż zadowalający. Składająca się na indeks piątka zmiennych jest między sobą silnie skorelowana, możemy ją nazwać rzetelną.

Druga tabela (zatytułowana *Item-Total Statictics,* a pojawiająca się wówczas, gdy zostanie oznaczona opcja *Show descriptives for scale if item is deleted*) zawiera wyniki dla poszczególnych zmiennych należących do testowanego indeksu lub skali.

Item-Total Statistics

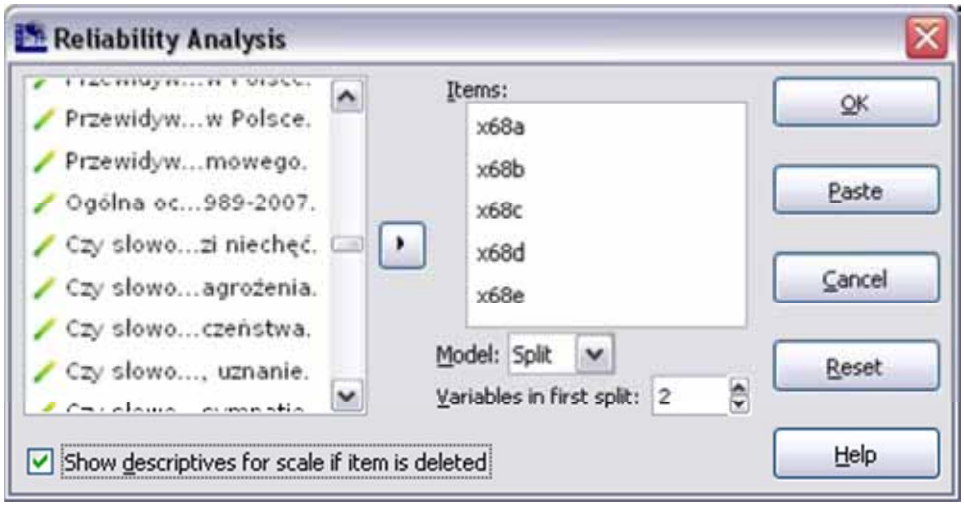
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Niekiedy rządy niedemokratyczne mogą być bardziej pożądane niż demokratyczne.	11,37	30,87	,62	,76
Dla ludzi takich jak ja nie ma znaczenia, czy rządy są demokratyczne.	11,63	33,69	,58	,78
W demokracji za dużo jest niezdecydowania i gadania.	12,28	32,50	,63	,76
Demokracja słabo radzi sobie z utrzymaniem porządku.	12,05	31,71	,68	,75
W demokracji są problemy, ale jest to lepszy system rządzenia niż każdy inny.	12,24	32,04	,49	,81

Dzięki tej tabeli możemy dowiedzieć się, które ze zmiennych nie pasują do indeksu lub skali i należy je wyłączyć z dalszych analiz. W celu zidentyfikowania zmiennych niepasujących lub najsłabszych odczytujemy wartości z ostatniej kolumny tabeli (*Cronbach's Alpha if Item Deleted*). Ze skali lub indeksu powinniśmy wyeliminować wszystkie te zmienne, których skasowanie podwyższa współczynnik alfa. W badanym przypadku żadna ze zmiennych nie jest typowana do usunięcia, a wyeliminowanie w szczególności zmiennych x68a, x68b, x68c i x68d wiązałoby się wręcz z obniżeniem wartości współczynnika.

W języku poleceń współczynnik alfa Cronbacha obliczamy następująco:

Składnia do wpisania w Edytorze	Opis działania składni
`RELIABILITY`	- przeprowadź test rzetelności
`/VARIABLES= x68a x68b x68c x68d x68e`	- przetestuj zmienne tworzące indeks lub skalę: x68a, x68b, x68c, x68d, x68e
`/MODEL=ALPHA`	- dokonaj testu rzetelności metodą alfa Cronbacha
`/SUMMARY = TOTAL.`	- wygeneruj tabelę Item-Total Statistics.

Drugim sposobem badania rzetelności w programie PSPP jest metoda połówkowa. Obliczeń dokonujemy wybierając z menu tekstowego *Analyze* ⇨ *Reliability,* a następnie oznaczając *Split* w rozwijanej liście *Model*. Z kolei w *Variables in first split* podajemy, w jaki sposób życzymy sobie, aby zbiory zmiennych zostały podzielone. Wybór ten ma znaczenie przy nieparzystej liczbie zmiennych. W poniższym przypadku, dzieląc pięć zmiennych, zdecydowaliśmy, że w pierwszej połówce znajdą się dwie zmienne (x68a i x68b), a w drugiej - pozostałe trzy.



Po zaakceptowaniu powyższego wyboru otrzymujemy dwie tabele zawierającą liczbę jednostek analizy (*Case Processing Summary*) oraz tabelę prezentującą wyniki testu rzetelności połówkowej (*Reliability Statistics*).

Reliability Statistics

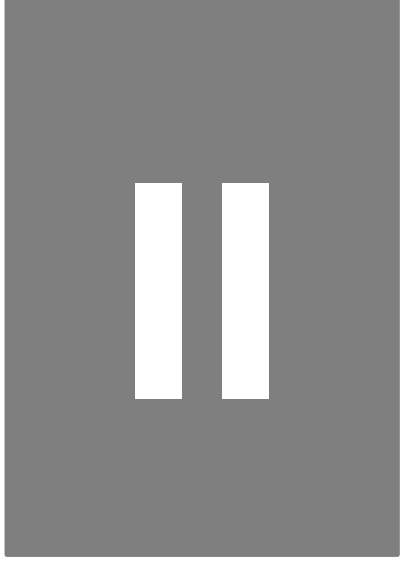
Cronbach's Alpha	Part 1	Value	,69
		N of Items	2,00
	Part 2	Value	,59
		N of Items	3,00
	Total N of Items		5,00
Correlation Between Forms			,56
Spearman-Brown Coefficient	Equal Length		,72
	Unequal Length		,72
Guttman Split-Half Coefficient			,72

Wiersze *Part 1* i *Part 2 Cronbach's Alpha* w tabeli *Reliability Statistics* pokazują współczynnik alfa w każdej z połówek. W *Part 1* jest to alfa dla zmiennych x68a i x68b, a w *Part 2* dla zmiennych x68c, x68d, x68e. Współczynniki te nie powinny się od siebie zbytnio różnić (nie więcej niż 0,2). Najważniejsze jednak dla interpretacji rzetelności są ostateczne wyniki testu połówkowego. Możemy wy-

bierać pomiędzy współczynnikiem Spearmana-Browna, a współczynnikiem Guttmana. Mierzą one korelację pomiędzy obiema połówkami testu. Wyniki pomiaru interpretujemy następująco: wartości poniżej 0,5 wskazują na niską rzetelność, a powyżej 0,8 pozwalają twierdzić, że jest ona nader zadowalająca. Uzupełniające miary odnajdujemy w *Item-Total Statistics.* Interpretujemy je analogicznie jak w przypadku alfa Cronbacha.

W Edytorze składni PSPP rzetelność połówkową obliczamy następująco:

Składnia do wpisania w Edytorze	Opis działania składni
`RELIABILITY`	- przeprowadź test rzetelności
`/VARIABLES= x68a x68b x68c x68d x68e`	- przetestuj zmienne tworzące indeks lub skalę: x68a, x68b, x68c, x68d, x68e
`/MODEL=SPLIT (2)`	- dokonaj testu rzetelności metodą połówkową, pierwsza połówka ma zawierać 2 zmienne
`/SUMMARY = TOTAL.`	- wygeneruj tabelę Item-Total Statistics.



Część II. Przygotowanie zbioru danych do analizy

7

Rozdział 7. Rekonfiguracja zbioru danych

Przygotowanie zbioru danych do analizy jest kluczowym zadaniem badacza. Często czynność ta zajmuje wielokrotnie więcej czasu niż same analizy. Jest ona ponadto warunkiem *sine qua non* wykonania poprawnych obliczeń i wyciągnięcia właściwych wniosków. Przez przygotowanie zbioru danych do analizy należy rozumieć wszystkie te czynności, które umożliwią uzyskanie pełnowartościowych danych poprzez wyeliminowanie błędów lub niepełnych danych, powstałych na różnych etapach przygotowania zbioru (od tworzenia kwestionariusza, aż do utworzenia zbioru danych w formacie PSPP) i takie ich przetworzenie, by umożliwiły osiągnięcie założonego celu analitycznego. Przygotowanie zbioru danych do analizy obejmuje dwie grupy czynności: po pierwsze przekształcenia całości zbioru danych oraz działania na samych zmiennych. Część dotycząca przygotowania zbioru danych do analizy stanowi techniczną instrukcję dokonywania modyfikacji, lecz jednocześnie ujawnia przesłanki, cele i prawidła dokonywania tych przekształceń.

7.1. Zasady rekonfiguracji zbioru danych

W niniejszym podrozdziale przedstawiono przesłanki oraz cele przekształceń zbiorów danych i zmiennych. Na początku rozważań należy podkreślić, że zastosowanie dużej części niżej przedstawionych zasad wymaga rozumienia i umiejętności zastosowania elementarnych analiz. Oznacza to, że proces rekonfiguracji zbioru danych odbywa się jednocześnie z ciągłymi, cząstkowymi analizami tego zbioru według sekwencji, której pierwszym członem jest modyfikacja, a drugim - ewaluacja skutków dokonanej ingerencji, to jest sprawdzenie, w jaki sposób dokonana zmiana wpływa na wyniki. Proces rekonfiguracji zbioru danych odbywa się przed rozpoczęciem właściwych analiz (choć w niektórych, rzadkich przypadkach cząstkowych rekonfiguracji dokonuje się także podczas prowadzenia obliczeń, ma to jednak charakter incydentalny). Proces rekonfiguracji zbioru danych można podzielić na następujące etapy:

Etap 1. Sprawdzenie kompletności i poprawności zbioru danych. Pierwszą czynnością, którą należy podjąć jest sprawdzenie, czy w zbiorze danych znajduje się wystarczająca i właściwa liczba jednostek analizy, a więc czy badanie zostało zrealizowane na właściwej liczebnie próbie. Tu także należy zdecydować, czy do dalszych analiz przyjmujemy tylko kompletne, czy także niekompletne jednostki

analizy (na przykład przerwane wywiady). Następnie należy sprawdzić, czy w zbiorze uwzględniono wszystkie żądane zmienne. Należy dokładnie sprawdzić liczbę niekompletnych jednostek analizy oraz niekompletnych zmiennych. Zjawisko braku wartości zmiennej lub zmiennych w zbiorze danych (wynikające na przykład z braku odpowiedzi respondenta, niewpisania jej przez ankietera do kwestionariusza, przypadkowego wyeliminowania jej przy wprowadzaniu danych) nazywamy brakami danych (*missing values*). Dopuszczalnym, typowym, przyjmowanym powszechnie w praktyce badawczej poziomem braków danych jest 5 proc., a nawet 10 proc.[1] (należy jednak pamiętać o specyfice poszczególnych zmiennych - badacz może zdawać sobie sprawę z potencjalnie wysokiego odsetka braków danych, jaki uzyska w badaniu, jednak mimo to świadomie może podjąć decyzję o zadaniu pytania drażliwego lub trudnego). Sytuacją, wymagającą interwencji badacza, są braki odpowiedzi na poziomie od 10 do 30 proc. Na ogół nie analizuje się zmiennych, gdzie braki danych sięgają powyżej 30 proc. (należy brać jednak pod uwagę jednostkowe przypadki, a podjęcie decyzji i poniesienie odpowiedzialności spada ostatecznie na badacza).

Naszej ocenie podlegać musi również poprawność reguł przejść pomiędzy poszczególnymi zmiennymi. Szczególne znaczenie ma to w przypadku badań z użyciem kwestionariuszy papierowych, gdzie ankieter jest odpowiedzialny za przestrzeganie reguł przejść. W przypadku zbierania danych z użyciem technik wspomaganych komputerowo, reguły przejścia są narzucone w programie i nienaruszalne. W analizie reguł przejść możemy stwierdzić dwa typy przypadków: brak danych w miejscu, gdzie powinny się one znaleźć i obecność danych w miejscu, w którym być nie powinny.

Istotne jest sprawdzenie, czy jednostki analizy nie powtarzają się, to jest czy w zbiorze danych nie ma dwóch lub więcej identycznych jednostek analizy. Praktyka badawcza pokazuje, że w toku elektronicznego przetwarzania danych tego typu problemy nie są rzadkie.

Po sprawdzeniu kompletności zbioru danych przystępujemy do oceny jego poprawności. Przede wszystkim należy sprawdzić, czy wartości zmiennych mieszczą się w przewidzianych dla nich zakresach. Jeśli na przykład badany mógł wybrać jedną opcję z sześciostopniowej skali (od 1 do 6), to wartości zmiennych 7, 8 czy 9 mogą być uznane za artefakty, powstałe chociażby w procesie wprowadzania danych - na przykład, jeśli dane były skanowane z papierowych kwestionariuszy, to program OCR *(Optical Character Renognition)* mógł nieprawidłowo zidentyfikować cyfrę 6 jako 8. Jeśli ustalimy, że wykraczające poza zakres wartości zmiennej poszczególne wartości zmiennej nie mają charakteru znaczącego (na przykład mogły zostać wpisane przez ankietera, ponieważ żadna z wartości zmiennych nie odzwierciedlała odpowiedzi respondenta), wówczas należy je usunąć lub zastąpić innymi wartościami. Ten cząstkowy etap można nazwać „sprawdzaniem kolumnami". Następnie ocenie poddajemy poszczególne jednostki analizy; ten zbiór czynności nazwijmy „sprawdzaniem wierszami". Rozpoznajemy tu takie sytuacje, gdy w danej jednostce analizy znajdują się sprzeczne, wykluczające się wartości zmiennych, na przykład osiemnastolatka deklaruje posiadanie wykształcenia wyższego lub zwolennik skrajnej lewicy wskazuje, że działa w organizacji skrajnie prawicowej. Tego typu przypadki należy traktować ostrożnie - analityk musi rozstrzygnąć czy mamy do czynienia z wiarygodnymi danymi odzwierciedlającymi jednak rzeczywistość, czy też do zbioru danych wkradł się błąd. W pierwszym przypadku takie jednostki analizy pozostawiamy, opatrując je w raporcie adekwatnym komentarzem, w drugim przypadku należy je zmodyfikować lub usunąć.

[1] Porównaj: P.L. Roth, F.S. Switzer II, *Missing Data: Instrument-Level Heffalumps and Item-Level Woozles*, w: http://division.aomonline.org/rm/1999_RMD_Forum_Missing_Data.htm, dostęp: luty 2012. Artykuł ten jest wart przeczytania - autorzy w zabawny i przystępny sposób wdrażają Czytelnika w elementarną wiedzę na temat procedur imputacji danych.

Etap 2. Eliminacja braków danych. Po dokonaniu rozpoznania zbioru danych badacz podejmuje decyzje odnośnie postępowania z brakami danych, obserwacjami niewiarygodnymi lub niesatysfakcjonującymi. Braki danych w zbiorach danych w naukach społecznych są zjawiskiem powszechnym[2]. Z drugiej strony część narzędzi statystycznych nie dopuszcza obecności braków danych. W związku z tym opracowano szereg procedur, spośród których badacz może wybrać właściwy sposób procedowania ze zbiorem danych.

Rozwiązanie 1. Pierwsze rozwiązanie ma charakter radykalny - polega na **usunięciu ze zbioru danych tych jednostek analizy**, w których znalazły się braki danych. Stosowane może być tam, gdzie nie dopuszcza się w danym typie analiz statystycznych braków danych oraz jeśli zbiór jest wystarczająco liczny, by zastosować w nim zamierzone analizy. Tego typu selekcja określana jest mianem wykluczania wierszami (*listwise deletion*). Usuwanie ze zbioru danych może też przyjmować charakter cząstkowy, bowiem możemy postępować elastycznie nie eliminując całych jednostek analizy, lecz tylko te, które w **zbiorze danych nie mają swojej pary dla danej analizy**. W takiej sytuacji niektóre analizy prowadzone są na większym zbiorze danych, a inne - na mniejszym w zależności od poziomu braków danych w poszczególnych zmiennych. Ten typ selekcji do analiz nazywamy wykluczaniem parami (*pairwise deletion*).

Różnice pomiędzy wykluczaniem wierszami i wykluczaniem parami można przedstawić odwołując się do przykładowej macierzy danych w tabeli 5. Jeśli zdecydowaliśmy się na wykluczanie wierszami (*listwise deletion*), wówczas jednostkami analizy zostaną te oznaczone numerami porządkowymi 1, 3 i 6. Z kolei wykluczanie parami (*pairwise deletion*), gdy analizujemy zmienne A i B, umożliwia nam włączenie do analiz jednostek analizy o numerach 1, 3, 4 oraz 6. Jeżeli przedmiotem analiz są zmienne B i C to wykluczamy parami jednostki 2, 4 i 5.

Tabela 5. Macierz danych zawierająca braki danych

Lp.	Zmienna A	Zmienna B	Zmienna C
Jednostka analizy 1	jest	jest	jest
Jednostka analizy 2	jest	brak danych	jest
Jednostka analizy 3	jest	jest	jest
Jednostka analizy 4	jest	jest	brak danych
Jednostka analizy 5	brak danych	brak danych	brak danych
Jednostka analizy 6	jest	jest	Jest

[2] Wyróżnić można następujące trzy rodzaje braków danych: losowe (jeśli nie są związane z innymi zmiennymi), częściowo losowe (braki danych nie są uzależnione od wartości zmiennej, do której należą, lecz są uzależnione od innych zmiennych) oraz nielosowe (zależne od wartości zmiennej, w której występują). Przykładem pierwszego typu braków danych jest sytuacja, w której odpowiedź na dane pytanie jest udzielana lub nieudzielana przez respondenta niezależnie od innych czynników (zmiennych). W przypadku braków danych częściowo losowych udzielenie lub nieudzielenie odpowiedzi na dane pytanie może zależeć od innych zmiennych. Obserwowaną zależnością tego typu jest nieudzielanie odpowiedzi na pytanie o wiek przez kobiety. Z kolei nielosowy brak danych to taki, w którym brak odpowiedzi na dane pytanie skorelowany jest z niektórymi poziomami zmiennej. Często występującą zależnością tego typu jest odmowa udzielenia odpowiedzi na pytanie o dochód przez badanych zarabiających znacznie poniżej i znacznie powyżej przeciętnej, a także na pytanie o wykształcenie wśród osób z wykształceniem zawodowym i niższym niż zawodowe. Badacz powinien dokonać interpretacji, z jakim typem braku danych ma do czynienia i w zależności od tego podejmować określone kroki eliminacji braków danych.

Rozwiązanie 2. Można również podjąć **próbę uzupełnienia braków danych poprzez ponowne dotarcie do osób, które uczestniczyły w procesie zbierania danych**: respondentów, ankieterów, operatorów wprowadzających dane (w tym także do źródła, z którego korzystali oni wprowadzając dane - mianowicie do kwestionariuszy). Nie zawsze jednak rozwiązanie to jest możliwe i efektywne (badacz musi przeprowadzić kalkulację zysków i strat oraz realnie ocenić prawdopodobieństwo zdobycia danych). Działanie to ma sens o tyle tylko, o ile nie korzystamy z danych wtórnych, lecz pierwotnych - zgromadzonych przez nas samych lub współpracujący zespół. Najlepszym rozwiązanie jest ponowne dotarcie do respondenta, jednak wiąże się to z kosztami, a także wysokim prawdopodobieństwem tego, że i tak nie uzupełnimy braku lub braków odpowiedzi, skoro nie udało się tego dokonać ankieterowi.

Można także podjąć próbę uzyskania brakujących informacji od zbierającego je ankietera. Z praktyki badawczej wynika, że próby takie są nieefektywne. Po upływie kilkunastu, a nawet kilku dni ankieter na ogół nie pamięta konkretnej sytuacji związanej z danym wywiadem, a także może udzielić fałszywej odpowiedzi w celu zadowolenia rozmówcy. Zalecane są natomiast konsultacje z operatorem wprowadzającym dane oraz - jeśli dane zostały zebrane w toku tradycyjnych standaryzowanych wywiadów kwestionariuszowych w formie papierowej - przeglądanie kwestionariuszy i porównywanie ich ze zbiorem danych. Z kolei w przypadku zbioru danych zebranych w toku wywiadów telefonicznych wspomaganych komputerowo zalecane jest w krytycznych przypadkach przeglądanie komentarzy wpisywanych przez ankieterów (większość systemów do badań telefonicznych posiada tego typu opcję). Należy podkreślić, że nie jest to samodzielne rozwiązanie, ma ono charakter uzupełniający w stosunku do rozwiązania pierwszego i trzeciego.

Rozwiązanie 3. Uzupełnianie braków danych. W literaturze przedmiotu szeroko mówi się o technikach uzupełniania brakujących danych (*Missing Data Techniques*, MDT) lub imputacji danych[3]. Jest to obszerny zbiór zagadnień, który został poniżej wyłożony w takim tylko zakresie, jaki konieczny jest by początkujący badacz poradził sobie w praktyce z przygotowaniem zbioru danych do analizy. Zastosowanie podanych rozwiązań wymaga znajomości podstawowych analiz w programie PSPP, a warunkiem zastosowania uzupełniania braków danych w pełnym zakresie jest opanowanie średniozaawansowanych i zaawansowanych analiz statystycznych. Wyróżniamy dwa typu imputacji danych: klasyczne oraz nowe.

Najprostszą klasyczną techniką uzupełniania braków danych jest **średnia z całości dostępnych danych**. Jest to najbardziej prymitywna z metod i powinna być stosowana wyłącznie w ostateczności, nazywa się ją metodą naiwną. Możemy również zastosować jej odmianę - tak zwaną **średnią lokalną**, to jest średnią z jakiejś wyodrębnionej podpróby ze zbioru danych (jednostek analizy tożsamych pod pewnym względem, istotnym z punktu widzenia badacza). Brakujące zmienne możemy zastępować również **medianą** oraz **dominantą** (zarówno globalną - dla całego zbioru danych, jak też lokalną - dla części zbioru). Należy pamiętać, że w toku zastępowania braków danych miarami pozycyjnymi następuje zmniejszenie się wielkości ich wariancji, a przez to redukcja w przypadku zastosowania niektórych miar, na przykład współczynnika korelacji. Ponadto zastępowanie braków danych miarami pozycyjnymi ma sens

[3] Anglojęzyczna literatura na ten temat jest stosunkowo obszerna. Zainteresowanych zachęcić można do zapoznania się w szczególności z następującymi pozycjami, które wprowadzają w zagadnienie i odsyłają do dalszych, bardziej zaawansowanych lektur: R.G. Downey, C.V. King, *Missing data in Likert ratings: A comparison of replacement methods*, „Journal of General Psychology", 1998, 125, s. 175-189; P.L. Roth, *Missing data: A conceptual review for applied psychologists*, „Personnel Psychology", 1994, 47, s. 537-560; G. Kalton, D. Kasprzyk, *The treatment of missing data*, „Survey Methodology", 1986, 12, s. 1-16; R.J.A. Little, D.B. Rubin, *Statistical analysis with missing data*, John Wiley and Sons, Nowy Jork 1987; R.J.A. Little, D.B. Rubin, *Statistical Analysis with Missing Data,* John Wiley and Sons, Nowy Jork 2002.

tylko wówczas, jeśli braki danych mają charakter losowy. Jeśli tak nie będzie, zastosowanie tej metody imputacji czyni wyniki mniej reprezentatywnymi dla populacji. Powszechnie stosuje się także mechanizm **regresji wielokrotnej**. W tym celu należy odnaleźć zmienne, które są na mocy przesłanek teoretycznych lub lokalnych empirycznych skorelowane ze zmienną zawierającą braki danych. Odnalezione zmienne traktujemy jak predyktory zmiennej zawierającej braki danych. Stosuje się także mechanizm **regresji stochastycznej** - tożsamy z regresją wielokrotną, lecz uzupełniony o zjawisko losowego rozrzutu danych. W liberalnej praktyce badawczej zdarza się uzupełnianie braków danych liczbami **losowymi** z zakresu wartości danej zmiennej. Nie zaleca się tej praktyki, choć może mieć ona uzasadnienie, gdy braki danych mają charakter całkowicie losowy. Wśród nieklasycznych metod można wyróżnić metodę **łączenia odpowiedników** (*pattern matching*), polegającą na zastępowaniu brakujących danych pochodzącymi z innych jednostek analizy, ale takich, które są pod jakimś względem lub względami tożsame z jednostką, w której braki danych uzupełniamy. Metodę tę stosuje się w sytuacji, gdy mamy do czynienia z niewielkimi liczebnie brakami danych w całym zbiorze. Ten typ imputacji nazywany jest także w literaturze imputacją typu *hot-deck.* Stosuje się także imputację typu *cold-deck,* polegającą na uzupełnianiu danych nie z bieżącego zbioru, lecz innych, dostępnych zbiorów danych lub nawet symulacji[4]. Do zaawansowanych metod uzupełniania braków danych należą między innymi Imputacja EM (*Expectation-Maximization Imputation*) lub metoda wielokrotnej imputacji (*Rubin's Multiple Imputation*)[5].

Należy pamiętać, że uzupełnianie braków danych jest metodą zastępczą, obniżającą jakość zbioru danych i powinna być stosowana tylko w ostateczności. Właściwe zaplanowanie i przeprowadzenie procesu zbierania danych pozwala na ograniczenie braków danych do akceptowalnego poziomu. Rozwiązanie uzupełniania braków danych stosujemy wówczas, gdy nie ma innego wyjścia. Należy wspomnieć, że niektórzy badacze postrzegają procedury imputacji jako swoiste fałszowanie danych, dlatego konieczna jest rzetelność i przejrzystość stosowania tych procedur. Badacz powinien w raporcie podać komplet informacji na temat rozmiarów zjawiska braków w zbiorze danych, wyjaśnić, a przynajmniej podzielić się przypuszczeniami skąd one wynikają, a także opisać czynności podjęte w celu wyeliminowania braków danych. Procedura uzupełniania braków danych musi być wyczerpująca, przejrzysta i zrozumiała dla czytelnika raportu.

Niezależnie od wad imputacji danych można wymienić jej następujące zalety: po pierwsze, umożliwia badaczowi stosowanie bardziej wyrafinowanych statystyk, po drugie, znacznie upraszcza pracę analityczną, po trzecie, wprowadza pewien porządek w raporcie z badań, bowiem wyniki badania prezentowane są na próbie jednej wielkości, nie zaś wielu, po czwarte, pozwala zachować kompletny zbiór danych.

W programie PSPP (wersja 0.7.9) brak jest jeszcze funkcji automatycznej imputacji danych, należy dokonywać ich ręcznie. W programie SPSS taka funkcja istnieje - należy spodziewać się, że z czasem to rozwiązanie pojawi się także w PSPP.

[4] Szerzej na ten temat: G. Schoier, *On partial nonresponse situations: the hot deck imputation method*, w: http://www.stat.fi/isi99/proceedings/arkisto/varasto/scho0502.pdf, dostęp: luty 2012; B.L. Ford, *An overview of hot-deck procedures*, w: *Incomplete data in sample surveys*, W.G. Madow, I. Olkin, D.B. Rubin (red.), Academic Press, Nowy Jork 1983, s.185-207.

[5] Zainteresowani mogą skorzystać z obszernego, wyczerpującego, lecz niezbyt przystępnego opracowania na temat metody imputacji EM: M.R. Gupta, Y. Chen, *Theory and Use of the EM Algorithm*, „Foundations and Trends in Signal Processing", 2010, 4 (3), s. 223-296. Z kolei metoda Rubina została omówiona w: R.J.A. Little, D.B. Rubin, dz. cyt.

Etap 3. Dostosowanie zmiennych do potrzeb analiz. W tym etapie dokonujemy wszystkich czynności objętych planem analiz, stosując tu w zależności od potrzeb głównie zabieg rekodowania oraz obliczania wartości zmiennych.

Etap 4. Dostosowanie zbioru danych do potrzeb analiz, w ramach którego podejmujemy głównie procedurę filtrowania, ważenia oraz agregacji danych zgodnie z założonym planem analiz.

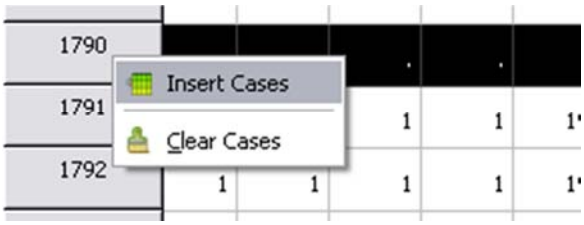
Po wykonaniu wyżej opisanych czterech etapów zbiór danych jest gotowy do analizy.

7.2. Dodawanie, modyfikowanie i usuwanie zmiennych i jednostek analizy w zbiorze danych

Rekonfiguracja zbioru danych polega na modyfikacji jednostek analizy: ich dodawaniu, usuwaniu i kategoryzowaniu. Kluczowe, najczęściej wykorzystywane opcje rekonfiguracji danych zostały opisane obszerniej w kolejnych podrozdziałach.

Dodawanie, modyfikowanie oraz usuwanie zmiennych i jednostek analizy w zbiorze danych jest elementarną czynnością wykonywaną przez badacza. Można tego dokonać na różne sposoby (zostały one omówione w dalszych lub - częściowo - we wcześniejszych partiach niniejszej publikacji). W tym podrozdziale przedstawiono najprostszy sposób dokonywania tych czynności - ręcznie, w sposób niezautomatyzowany.

W celu **dodania jednostki analizy** klikamy w Widoku danych prawym przyciskiem myszy po ustawieniu kursora w obrębie pola ID. Z rozwijanej listy wybieramy Wstaw jednostki analizy (*Insert Cases*).

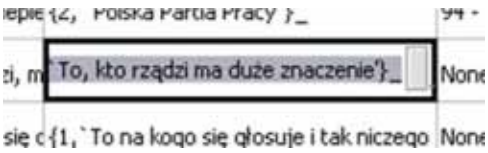


Dodania jednostki analizy można także dokonać z poziomu menu tekstowego (*Edit* ⇨ *Insert Cases*) lub po kliknięciu na ikonę Wstaw jednostki analizy:

Program PSPP wstawia wówczas pusty wiersz, w którym można umieścić wartości zmiennych. Analogicznie postępujemy, jeśli chcemy **dodać zmienną do zbioru danych**. Można tego dokonać zarówno w Widoku danych (klikając prawym przyciskiem myszy, gdy kursor znajdzie się w główce kolumny), jak też w Widoku zmiennych (gdy kursor umieścimy w kolumnie ID). Tego samego zabiegu dokonujemy także z użyciem menu ikon, jak też menu tekstowego.

Opis zmiennej modyfikujemy na dwa sposoby: w części pól w zakładce *Variable View* możemy wpisywać wartości, bezpośrednio klikając na danym polu lewym przyciskiem myszy i wprowadzając tekst lub liczby (dotyczy to przede wszystkim pól takich jak nazwa zmiennej (*Name*) lub etykieta zmiennej (*La-*

bel)). Część pól (przede wszystkim opis wartości zmiennej (*Value Labels*)) posiada rozwijane listy otwierane po kliknięciu na prostokąt znajdujący się po prawej stronie pola:



Opis wartości zmiennej modyfikujemy wpisując dodawaną wartość w polu *Value,* opis tej wartości w *Value Label* oraz klikając przycisk Dodaj (*Add*). W efekcie nowa wartość zmiennej zostanie dołączona do listy. Analogicznie postępujemy w przypadku, gdy chcemy usunąć wartość zmiennej - klikamy na przeznaczoną do usunięcia wartość z listy, a następnie na przycisk Usuń (*Remove*).

Zmienna może być także modyfikowana w trybie skryptowym. Po otworzeniu Edytora składni można dokonać modyfikacji istniejącej listy wartości zmiennych następującą komendą:

Składnia do wpisania w Edytorze	Opis działania składni
`ADD VALUE LABELS`	- dodaj wartości zmiennych
`/c2t`	- dodaj je do zmiennej o nazwie c2t
`3 'odmowa odpowiedzi'.`	- przypisz w zmiennej c2t liczbie 3 etykietę „odmowa odpowiedzi". Wykonaj program
`EXECUTE.`	- wykonaj polecenie.

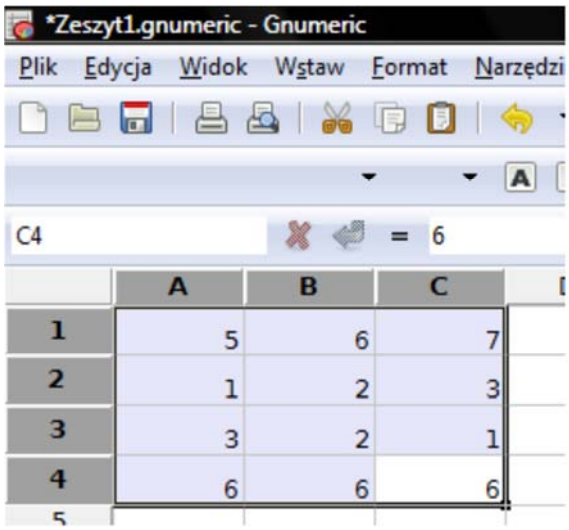
Z kolei pełna lista zmiennych może zostać stworzona za pomocą następującej składni:

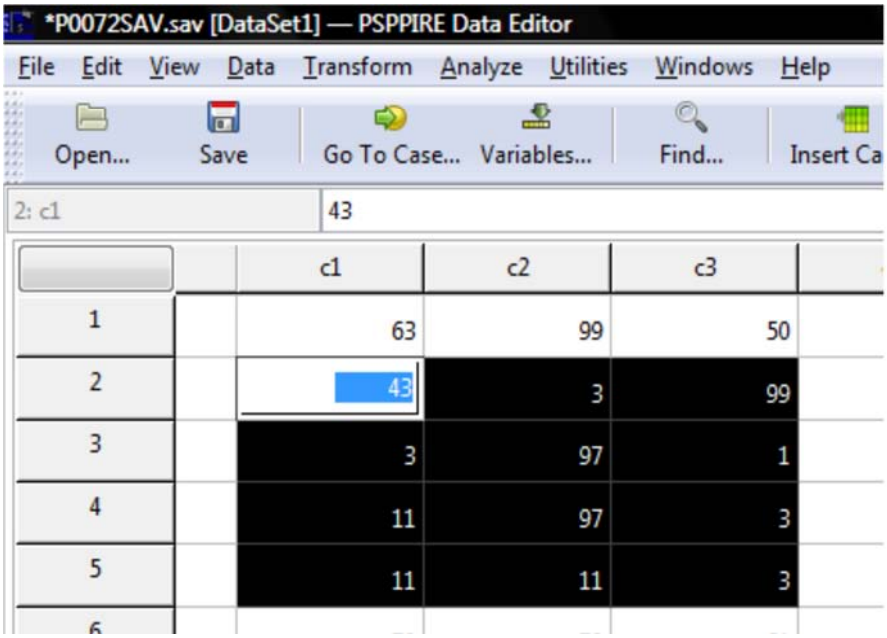
Składnia do wpisania w Edytorze	Opis działania składni
`VALUE LABELS`	- definiuj wartości zmiennych
`/c2t`	- definiuj je w zmiennej o nazwie c2t

`1 'tak'`
`2 'nie'`
`3 'odmowa odpowiedzi'.`
	- przypisz podanym liczbom wymienione etykiety. Wykonaj program

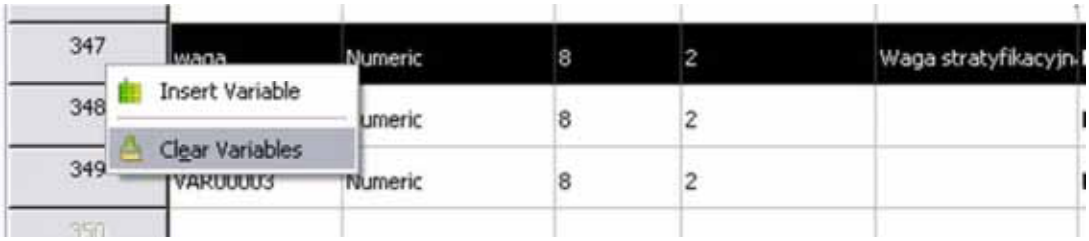
`EXECUTE.`	- wykonaj polecenie.

Modyfikacja wartości zmiennej następuje w Widoku danych - zaznaczamy dane pole, a następnie wpisujemy zadaną wartość. Wartości zmiennych mogą być modyfikowane również zbiorczo - w tym celu kopiujemy kilka wartości zmiennych na raz z innej macierzy (np. z arkusza Gnumeric, tabeli lub pliku tekstowego rozdzielanego tabulatorami), a następnie zaznaczamy analogiczny obszar w Widoku danych i wklejamy (posługujemy się klawiszami Ctrl + C i Ctrl + V). Poszczególne etapy tej czynności ilustrują dwa poniższe zrzuty ekranowe. Zwrócić należy uwagę na problem przy wklejaniu zmiennej znajdującej się w lewym górnym rogu tak zaznaczonego obszaru. Analogicznie można modyfikować zmienne znajdujące się w tym samym lub innym pliku PSPP.

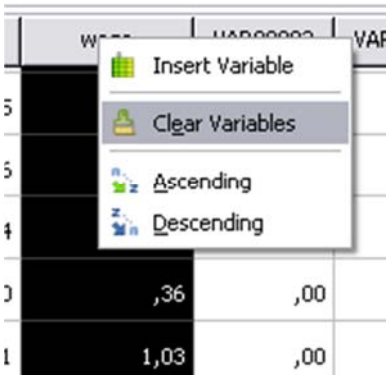




Zabiegu **całkowitego usunięcia zmiennych** dokonujemy w zakładce Widoku zmiennych (*Variable View*) poprzez kliknięcie prawym przyciskiem myszy na pole ID danej zmiennej (wiersz zostanie zaczerniony):



Z rozwijanej listy, która pojawi się po kliknięciu, wybieramy Usuń zmienne (*Clear Variables*). Zmienna zostanie bezpowrotnie usunięta. Usunięcia zmiennych można dokonać także w zakładce Widok danych. Prawym przyciskiem mysz klikamy na główkę kolumny, w której znajduje się zmienna i - analogicznie - wybieramy opcję usunięcia zmiennych.



Usuwanie **jednostek analizy** dokonuje się w sposób analogiczny jak usuwanie zmiennych. W Widoku danych wybieramy prawym przyciskiem myszy pole ID jednostki analizy, którą chcemy usunąć. Z rozwijanej list, która się wówczas pojawi, wybieramy Usuń obserwacje (*Clear Cases*). Pomimo tego, że możliwe jest zaznaczenie w programie PSPP kilku linii, kasowana jest tylko jedna z nich - zaznaczona

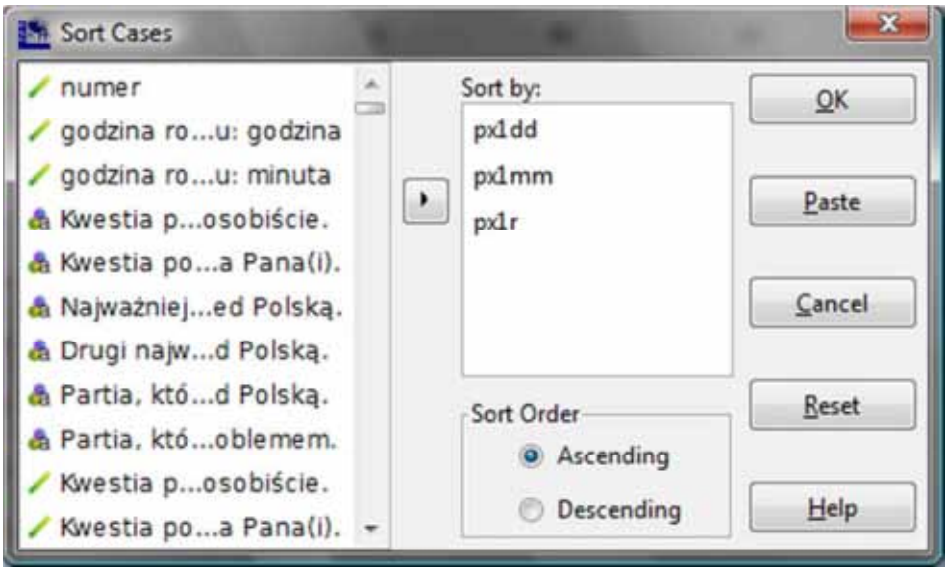
jako ostatnia. **Wartości zmiennych mogą być usuwane pojedynczo i zbiorowo.** W pierwszym przypadku przechodzimy do Widoku danych, ustawiamy kursor na polu, które chcemy usunąć, klikamy lewym przyciskiem myszy, a następnie naciskamy klawisz *Delete.* Analogiczny mechanizm działa w Widoku zmiennych. Jeśli chcemy usunąć za jednym razem większą liczbę wartości zmiennych w Widoku danych, zaznaczamy je kursorem, a następnie kasujemy (w istocie wycinamy) za pomocą skrótu klawiszowego Ctrl + X. Opcja ta nie działa w Widoku zmiennych.

99	99	PIERWSZY WALCOWNIK	ODMOW
99	99	STOLARZ	ROBIMY
99	99	SZWACZKA	SZYJE
99	99	SZWACZKA	SZYJE
99	99	PO SZKOLE ZASZŁA W CIĄŻE I NIE PRACUJE DO TEJ	
99	99	SPRZEDAWCA	OBSŁUG
99	99	PRACA W BIURZE	SKŁADA
99	99	PIELĘGNIARKA	ZAJMOV
99	99	PRACOWNIK UMYSŁOWY	PRACOV
99	99	SZWACZKA	SZYJĘ N

7.3. Sortowanie jednostek analizy w zbiorze danych

Sortowanie jednostek analizy polega na ich porządkowaniu rosnąco lub malejąco pod względem liczbowym, alfabetycznym lub liczbowo-alfabetycznym. Program PSPP umożliwia dokonanie sortowania na trzy następujące sposoby: w trybie okienkowym, z wykorzystaniem edytora składni lub ręczne.

Wykonanie sortowania w trybie okienkowym odbywa się po wybraniu z menu tekstowego *Data* ⇨ *Sort Cases*:



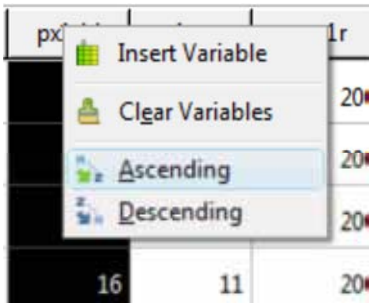
W lewej kolumnie okna znajdują się wszystkie zmienne ze zbioru danych. W prawym polu Sortuj według (*Sort by*) umieszczamy za pomocą strzałki zmienne, według których chcemy posortować zbiór danych. Sortowania można dokonać według porządku rosnącego (*Ascending)* lub malejącego (*Descending)*. Na przykład, w pierwszym przypadku najniższe wartości zmiennych będą znajdowały się na górze macierzy danych w Widoku danych, a najwyższe zostaną ulokowane na dole. Zbiór danych może być porządkowany według jednej lub wielu zmiennych. Kolejność zmiennych w polu Sortuj według wyznacza porządek sortowania. W powyższym przykładzie zmienne w pliku zbioru danych zostaną uporządkowane według następujących trzech zmiennych i kolejności: rok (px1r), miesiąc (px1mm) i dzień (px1dd) przeprowadzenia wywiadu. Za pomocą przycisku Wklej (*Paste*) możemy wygenerować komendy składni umożliwiające dokonanie sortowania.

Sortowania można dokonać również z użyciem Edytora składni. Komenda wpisywana w oknie Edytora składni, wywołująca analogiczne do powyższych skutki w zakresie sortowania powinna wyglądać następująco:

```
SORT CASES BY px1r px1mm px1dd(A).
```

Litera A umieszczona w nawiasach, oznacza sortowanie w porządku rosnącym (*Ascending*). Jeżeli zbiór danych chcemy sortować w porządku malejącym, to należy w nawiasach umieścić literę D (od anglojęzycznej nazwy *Descending)*.

Trzecim sposobem dokonania sortowania jest kliknięcie prawym przyciskiem myszy po umieszczeniu kursora w główce kolumny zmiennej przeznaczonej do sortowania:



W tym przypadku również możemy korzystać z sortowania w porządku rosnącym i porządku malejącym.

7.4. Transpozycja zbioru danych

Transpozycja jest zabiegiem polegającym na zamianie miejsc zmiennych i jednostek analizy. Wykonanie tej czynności powoduje, że dotychczasowe zmienne stają się jednostkami analizy, a dotychczasowe jednostki analizy stają się zmiennymi (prościej: wiersze zostają zamienione na kolumny i *vice versa*). Proces ten został wyłożony w tabeli 6 oraz tabeli 7. Pierwsza z tabel przedstawia stan początkowy, w którym jednostkami analizy są poszczególne głosowania ponumerowane od 1 do 5 w kolumnie ID, a zmiennymi - posłowie (Poseł 1, Poseł 2, Poseł n...). Każda ze zmiennych może przyjmować jedną z dwóch wartości: głosował za lub głosował przeciw. W tym stanie możemy prowadzić analizy dla poszczególnych posłów, a więc otrzymywać informacje o aktywności każdego z nich odrębnie.

Tabela 6. Macierz danych surowych przed transpozycją

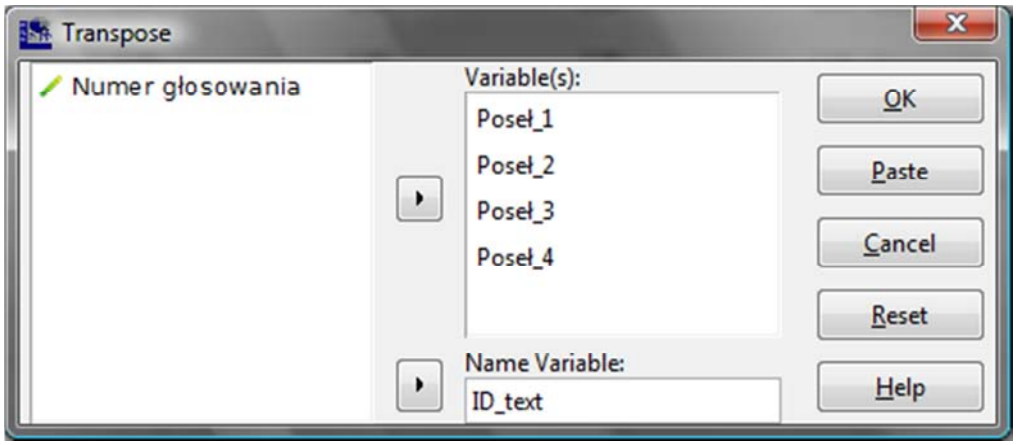
ID (numer głosowania)	Poseł 1	Poseł 2	Poseł 3	Poseł 4	ID tekstowe
1	za	za	za	za	Głosowanie 1
2	za	przeciw	za	przeciw	Głosowanie 2
3	przeciw	przeciw	przeciw	przeciw	Głosowanie 3
4	przeciw	za	przeciw	za	Głosowanie 4
5	przeciw	za	przeciw	przeciw	Głosowanie 5

Jeśli jednak potrzebujemy analiz w podziale na poszczególne głosowania, musimy dokonać transpozycji danych. Po dokonaniu tego zabiegu otrzymujemy dane ustrukturyzowane jak w tabeli 7 jednostkami analizy są posłowie, a zmiennymi - głosowania.

Tabela 7. Macierz danych surowych po dokonaniu transpozycji

ID (poseł) w PSPP - CASE_LBL	Głosowanie 1	Głosowanie 2	Głosowanie 3	Głosowanie 4	Głosowanie 5
Poseł 1	za	za	przeciw	przeciw	przeciw
Poseł 2	za	przeciw	przeciw	za	za
Poseł 3	za	za	przeciw	przeciw	przeciw
Poseł 4	za	przeciw	przeciw	za	przeciw

Transpozycji w programie PSPP dokonujemy wybierając *Data* ⇨ *Transpose:*



W polu po lewej stronie okna znajdują się wszystkie zmienne ze zbioru danych. Za pomocą strzałki umieszczamy w polu Zmienne (*Variable(s)*) te, które mają zostać transponowane. W polu Nazwa zmiennej (*Name Variable*) umieszczamy sztuczną, uprzednio stworzoną przez nas zmienną, której poszczególne wartości będą nazwami zmiennych po dokonaniu transformacji. Zmienna ta musi być tekstowa (*String*). W efekcie dokonania transpozycji utworzony zostanie w nowym oknie odrębny plik programu PSPP (należy go zapisać). W pliku tym w pierwszej kolumnie Widoku danych znajdą się dotychczasowe nazwy zmiennych (CASE_LBL).

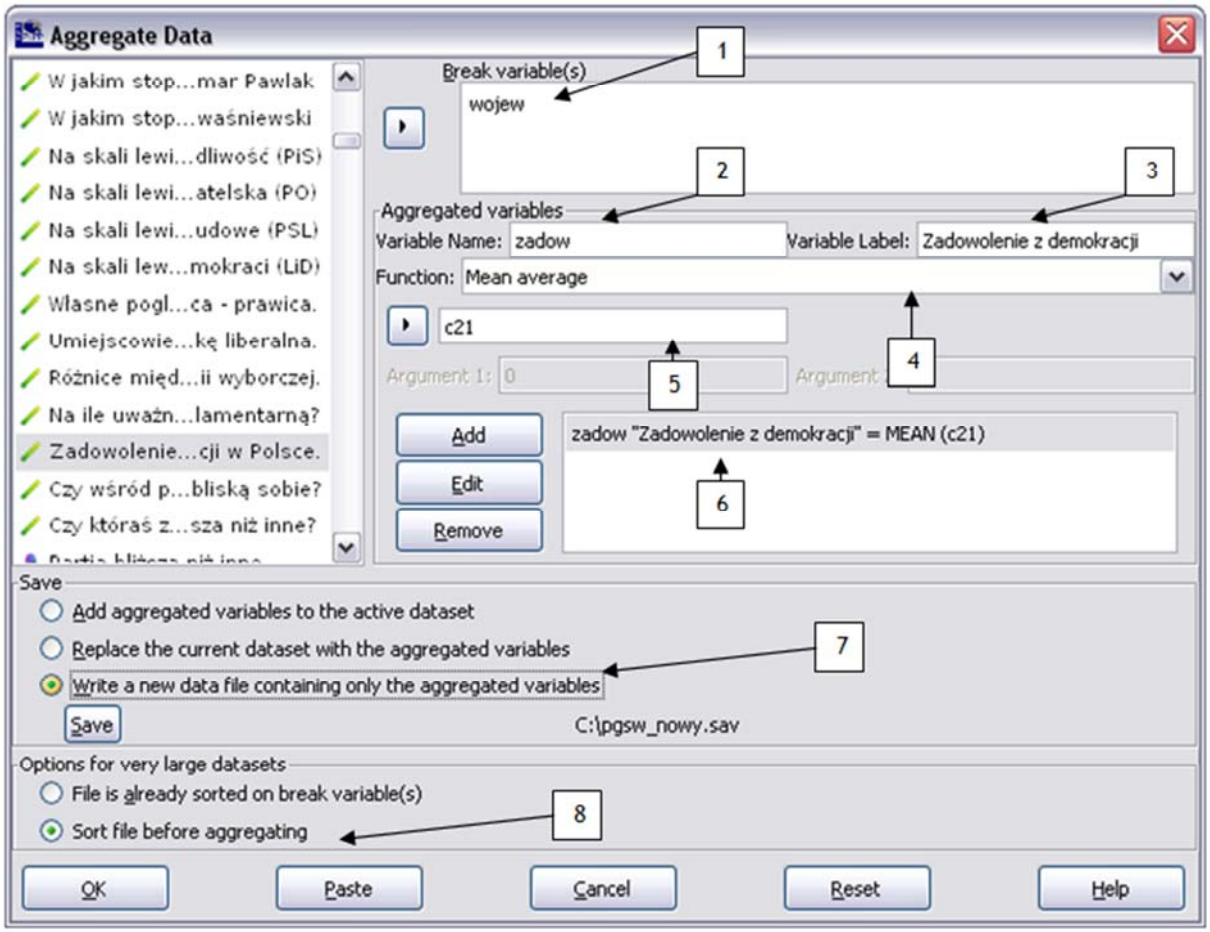
W celu dokonania powyżej przedstawionych czynności w edytorze składni należy wpisać i wykonać formułę:

```
FLIP /VARIABLES = posel_1 posel_2 posel_3 posel_4 /NEWNAME = ID_text.
```

7.5. Agregowanie zbioru danych

Agregowanie zbioru danych jest takim przekształceniem, które umożliwia zmianę jednostek analizy w inne. Nowe jednostki analizy tworzone są na podstawie syntezy z jednostek źródłowych. Syntetyzowanie może zostać przeprowadzone za pomocą rozmaitych miar położenia: miar pozycyjnych (między innymi modalnej, kwartyli, minimum i maksimum), miar przeciętnych (na przykład średniej arytmetycznej), miar dyspersji (np. odchylenia standardowego), czy innych miar opisujących zbiór danych (np. sumy).

Dla wyjaśnienia tej funkcji najlepiej posłużyć się egzemplifikacją. Przypuśćmy, iż zbadaliśmy za pomocą wywiadu kwestionariuszowego kilka tysięcy osób w całej Polsce zadając im pytanie, w jakim stopniu są oni zadowoleni z demokracji w naszym kraju. Nie interesują nas pojedyncze osoby, lecz chcemy analizować uzyskane wyniki ze względu na województwa. Oznacza to, że jednostką analizy pragniemy uczynić zmienną województwo w miejsce pojedynczego respondenta. Dzięki funkcji agregowania danych możemy uzyskać zbiór danych, w którym jednostkami analizy będą poszczególne województwa. Umożliwia to funkcja Agreguj dane (*Aggregate*) znajdująca się w menu tekstowym w zakładce Dane (*Data*).



Po jej wybraniu pokazuje się okno, którego kolejne elementy interpretujemy następująco:

1/ W polu *Break variable(s)* wybieramy tak zwane zmienne grupujące. Staną się one nowymi jednostkami analizy. W tym przypadku z bazy PGSW 2007 została wybrana zmienna województwo. W tym polu można umieścić więcej niż jedną zmienną grupującą.

2/ Pole Nazwa zmiennej (*Variable Name*) służy do umieszczenia nazwy zmiennej agregowanej, która zostanie utworzona. Użyto skrótu „zadow".

3/ W polu Etykieta zmiennej (*Variable label*) wpisujemy opisową etykietę zmiennej („Zadowolenie z demokracji").

4/ Rozwijana lista *Functions* umożliwia wybór sposobu przekształcania zmiennej agregowanej. Do wyboru pozostają miary pozycyjne takie jak modalna, n-tyle, minimum i maksimum, a także miary przeciętne i zróżnicowania. W tym przypadku wybór padł na średnią arytmetyczną. Oznacza to, że dla zmiennej „zadowolenie z demokracji", której pomiar odbywał się na skali od 1 do 5, zostaną policzone średnie wartości w podziale na zmienną „województwo".

5/ W miejsce tego pola należy dodać z listy znajdującej się po lewej stronie pola zmienną c21 - zadowolenie z demokracji.

6/ Po wykonaniu czynności zawartych w punktach 2-5 należy użyć przycisku *Add* i wówczas w polu pojawia się:

```
zadow "Zadowolenie z demokracji" = MEAN c21
```

7/ W tym miejscu ustalane są opcje zapisu nowoutworzonych zmiennych. Wybór może paść na jedną z trzech następujących opcji:

- *Add aggregated variables to active dataset* - dodanie syntetycznych zmiennych do bieżącego zbioru danych. Zmienna „zadow" zostanie dodana na końcu zbioru danych,

- *Replace the current dataset with the aggregated variables* - nadpisanie bieżącego zbioru danych nowoutworzonym. Pierwotny plik ulegnie nieodwracalnemu zniszczeniu,

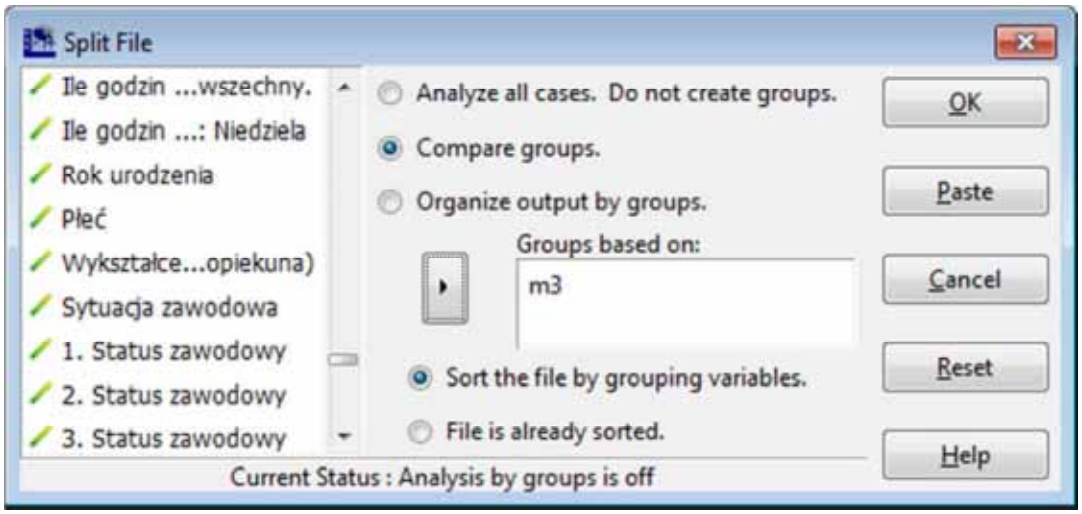
- *Write a new data file containing only the aggregated variables* - utworzenie nowego zbioru danych zawierającego syntetyczne wartości. Jest to opcja zalecana.

Przycisk Zapisz (*Save*) umożliwia wprowadzenie nazwy nowego pliku, w którym znajdą się zagregowane dane oraz ścieżki dostępu. Po wprowadzeniu tych danych obok przycisku pojawia się ścieżka dostępu i nazwa pliku, który ma zostać utworzony.

8/ Są to opcje dla bardzo dużych zbiorów danych - to jest liczących dziesiątki i setki tysięcy jednostek analizy - konieczne wówczas jest uprzednie posortowanie jednostek analizy. Możemy również wskazać, by dane zostały posortowane w toku agregacji.

7.6. Dzielenie zbioru danych na podgrupy

Ułatwieniem pracy analityka jest możliwość podzielenia zbioru danych na podgrupy. Służy to przyspieszeniu i uproszczeniu czynności analitycznych. Funkcja ta umożliwia otrzymywanie wyników analiz od razu w podziale na wyznaczone grupy, bez konieczności wykonywania dodatkowych czynności. Podział na podgrupy odbywa się poprzez wskazanie zmiennej wyznaczającej odrębne segmenty, na przykład kategorii wieku lub płci. Dzielenie zbioru danych na podgrupy analityczne odbywa się poprzez wybranie z menu tekstowego: *Data* ⇨ *Split File*:



W oknie podziału zbioru danych (*Split File*) możemy wybrać pomiędzy następującymi opcjami:

1/ Analizuj wszystkie jednostki analizy (*Analyze all cases. Do not create groups*) - poddaje analizie wszystkie jednostki analizy zbiorczo, a więc jest to podstawowy stan zbioru danych.

2/ **Porównaj grupy (*Compare groups*)** - zestawia dla porównania poszczególne grupy ze sobą w tabelach znajdujących się obok siebie. Opcja ta nie działa w wersji 0.7.9, a wyniki analiz po zaznaczeniu opcji 2 lub 3 nie różnią się.

3/ Podziel wyniki analiz na wyznaczone grupy *Organize output by groups* - dokonuje podziału według zmiennej grupującej lub zmiennych grupujących. Jeśli wybraliśmy jako zmienne grupujące płeć i wykształcenie, wówczas analizy będą kolejno prezentowane dla mężczyzn i kolejnych szczebli wykształcenia, a następnie kobiet i kolejnych szczebli wykształcenia.

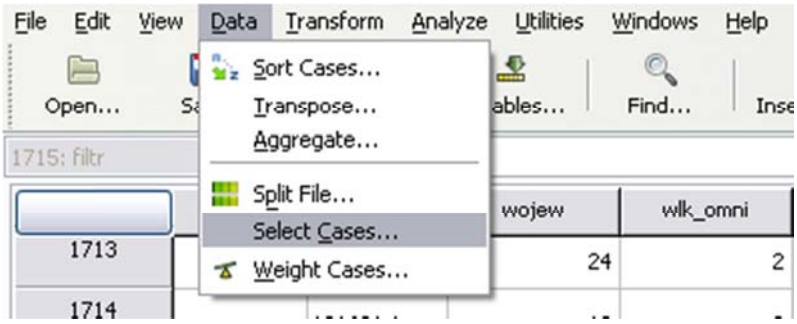
Po podjęciu decyzji odnośnie sposobu prezentacji danych wybieramy zmienne, które staną się zmiennymi podziału danych. Ponadto możemy wymusić prezentowanie wyników posortowanych według zmiennej grupującej (od najmniejszej do największej wartości) lub zachować istniejący porządek (odpowiednio: *Sort the file by grouping variables* oraz *File is already sorted*).

Informacja dotycząca włączonej opcji dzielenia na podgrupy jest widoczna na pasku stanu umiejscowionym w dolnej części okna programu PSPP. Wyłączenie dzielenia pliku następuje po wybraniu Analizuj wszystkie jednostki analizy (*Analyze all cases. Do not create groups*) w oknie *Split File* lub po wpisaniu w Edytorze składni:

```
SPLIT FILE OFF.
```

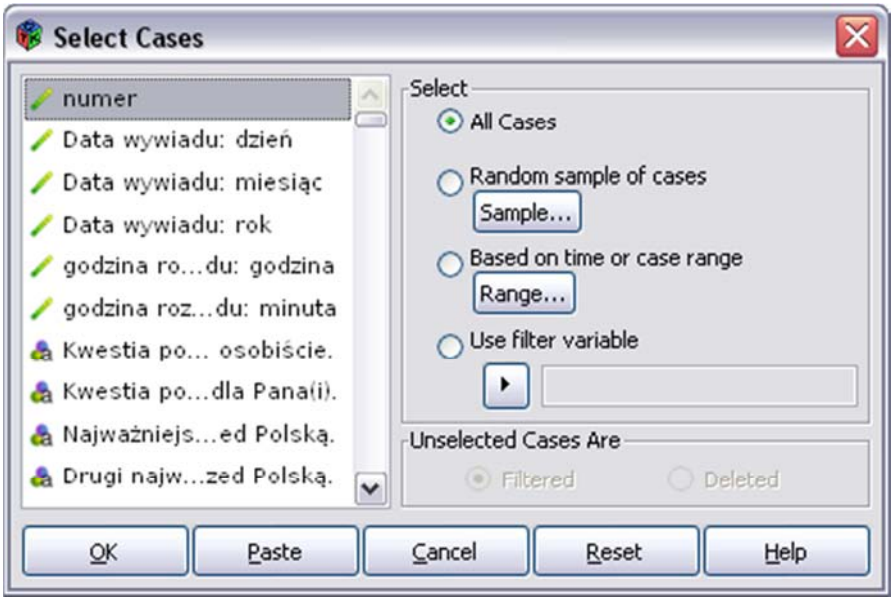
7.7. Selekcja jednostek analizy (*filter*)

W programie PSPP możemy poddawać analizom zarówno całość zbioru danych, jak również jego część - wybrany fragment lub fragmenty. Czynność tę nazywamy selekcją jednostek analizy lub w skrócie filtrowaniem danych. Do jej wykonania służy opcja w menu *Selekcja jednostek analizy* (*Data* ⇨ *Select Cases*):



Opcja ta służy wyodrębnieniu interesujących nas z punktu widzenia analizy danych grup. Na przykład chcemy skupić swoją uwagę wyłącznie na grupie kobiet znajdujących się w zbiorze danych. Musimy zatem wskazać, by mężczyźni nie byli brani pod uwagę podczas analiz.

Po wybraniu *Data* ⇨ *Select Cases* ukaże się następujące okno:

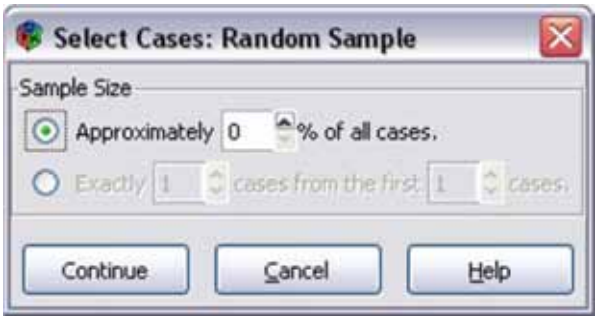


W lewej jego części wyświetlane są zmienne ze zbioru danych. Po prawej stronie umieszczono opcje służące do wykonywania operacji na zmiennych. Kolejno są to:

- Uwzględnienie wszystkich jednostek analizy ze zbioru (*All Cases*) - zaznaczenie tej opcji znosi wszystkie uprzednio nałożone filtry, wyłącza je. Alternatywnie możemy zastosować w oknie składni komendę wyłączającą filtr i przywracającą wszystkie jednostki analizy:

```
FILTER OFF.
```

- Selekcja losowej próbki jednostek analizy ze zbioru (*Random sample of cases*). Opcja ta umożliwia nam losowe wyodrębnienie ze zbioru danych jego części. Przesłankami do jej zastosowania może być testowanie zbioru danych lub chęć poddania go analizom wymagającym mniejszych próbek. Ta opcja może posłużyć także w procesie prowadzenia postępowania badawczego podczas doboru losowego próby z populacji.



Wielkość wylosowanego do analizy zbioru danych możemy zdefiniować dwojako - stosunkowo lub liczbowo. W pierwszym przypadku zaznaczamy wiersz *Approximately n% of all cases* i wpisujemy w nim jaką część z całości zbioru danych ma stanowić wylosowana próbka. Jeśli zbiór danych liczy 2000 przypadków, to jeśli wpiszemy 10 proc., otrzymamy jako wyjściowy zbiór danych około 200 przypadków. Możemy również bardzo precyzyjnie określić liczebność dobieranej próbki, jak również zawęzić obszar, z którego dobieramy. Służy do tego drugi wiersz zakładki. Po słowie *Exactly* wpisujemy, ile dokładnie jednostek analizy ma znaleźć się docelowo w zbiorze danych, a po słowach *cases from the first* umieszczamy informację z jakiej części bazy danych ma być dokonywane losowanie. Jeśli zbiór danych liczy 2000 przypadków, to zapis *Exactly 100 cases from the first 200 cases* będziemy interpretować jako komendę doboru dokładnie 100 jednostek analizy z pierwszych 200 rekordów w bazie danych.

- Selekcja oparta na zmiennej czasu lub na zakresie zmiennej (*Based on time or case range*) - pozwala na wybór wskazanego fragmentu zbioru danych. Jest to dobór nielosowy, celowy. Sensowne jest stosowanie tej opcji po uprzednim celowym uporządkowaniu bazy danych (sortowaniu).

Zakres wyboru jednostek analizy określamy posługując się numerami porządkowymi zbioru danych. Zapis: *Observation - First case 100; Last case 120* powoduje wybranie do analiz 21 jednostek analizy znajdujących się aktualnie na pozycjach od 100 do 120 w zbiorze danych.

- Selekcja z użyciem zmiennej filtrującej (*Use filter variable*) - umożliwia warunkowe selekcjonowanie jednostek analizy w zbiorze danych na podstawie wartości przyjmowanych przez wskazaną zmienną. Jest to funkcja najczęściej używana zarówno podczas przygotowania zbioru do analizy, jak również podczas samej analizy danych. Zmienną filtrującą tworzymy w procesie rekodowania w taki sposób, aby była zmienną dychotomiczną. Może przyjmować ona dwie i tylko dwie wartości - zero lub jeden. Nadanie jednostce analizy wartości zero spowoduje, że nie znajdzie się ona w selekcjonowanym zbiorze, a wartość jeden powoduje umieszczenie jej tam.

Ze względu na częstość używania tej opcji należy zapamiętać prosty zapis składni selekcji z użyciem zmiennej filtrującej:

```
FILTER BY filtr.
```

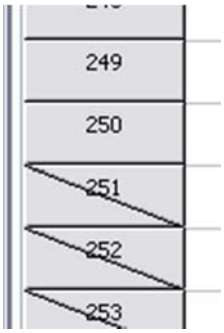
Filter by oznacza komendę dokonania selekcji, a *filtr* oznacza zmienną, na podstawie której dokonywane jest filtrowanie.

Po wybraniu opcji drugiej, trzeciej lub czwartej (selekcja losowej próbki jednostek analizy ze zbioru, selekcja oparta na zmiennej czasu lub na zakresie zmiennej, selekcja z użyciem zmiennej filtrującej) można zdecydować co ma się stać z jednostkami analizy, które nie zostały przez nas wybrane do analiz. Mogą one pozostać w zbiorze danych, lecz nie będą brane pod uwagę (jest to domyślne ustawienie) lub też mogą zostać usunięte ze zbioru danych. Pierwszy efekt uzyskujemy zaznaczając w *Unselected Cases Are* ⇨ *Filtered*, a drugi zaznaczając *Deleted.*

Zaleca się stosować tę drugą możliwość w sposób przemyślany, tylko wówczas gdy posiadamy kopię oryginalnego, pełnego zbioru danych, bowiem dane zostają skasowane nieodwracalnie.

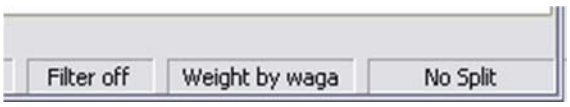
Utworzenie filtru powoduje utworzenie na końcu zbioru danych sztucznej zmiennej o nazwie *filter$*. Jest ona widoczna zarówno w Widoku zmiennych, jak i Widoku danych. Jednostki analizy oznaczone zerem nie będą uwzględniane w dalszych obliczeniach, a jednostki oznaczone jedynką - będą.

Zwróć uwagę na oznaczenia włączonego filtru. Po pierwsze, odfiltrowane jednostki analizy widoczne są w Widoku danych. Jednostki wyłączone z analiz oznaczane są tam w pierwszej zafiksowanej kolumnie (numer jednostki analizy) za pomocą przekreślenia:



Na rysunku widzimy, że jednostki analizy 249 i 250 będą uwzględnione w analizach, a 251, 252 i 253 - nie. Po drugie, fakt włączenia lub wyłączenia filtra odnotowany jest na tak zwanym pasku stanu znajdującym się w dolnej części okna programu PSPP:

Oznaczenie *Filter by filtr* oznacza, że aktualnie filtr jest włączony, a zmienna selekcjonująca jednostki analizy nosi nazwę *filtr*. Po wyłączeniu filtra uzyskujemy komunikat *Filter off:*



Częstym błędem popełnianym przez badaczy jest zapominanie o nałożonym filtrze. Każdorazowo przy wykonywaniu analiz należy zwracać uwagę na status nałożonego filtra.

8

Rozdział 8. Przekształcenia zmiennych w zbiorze danych

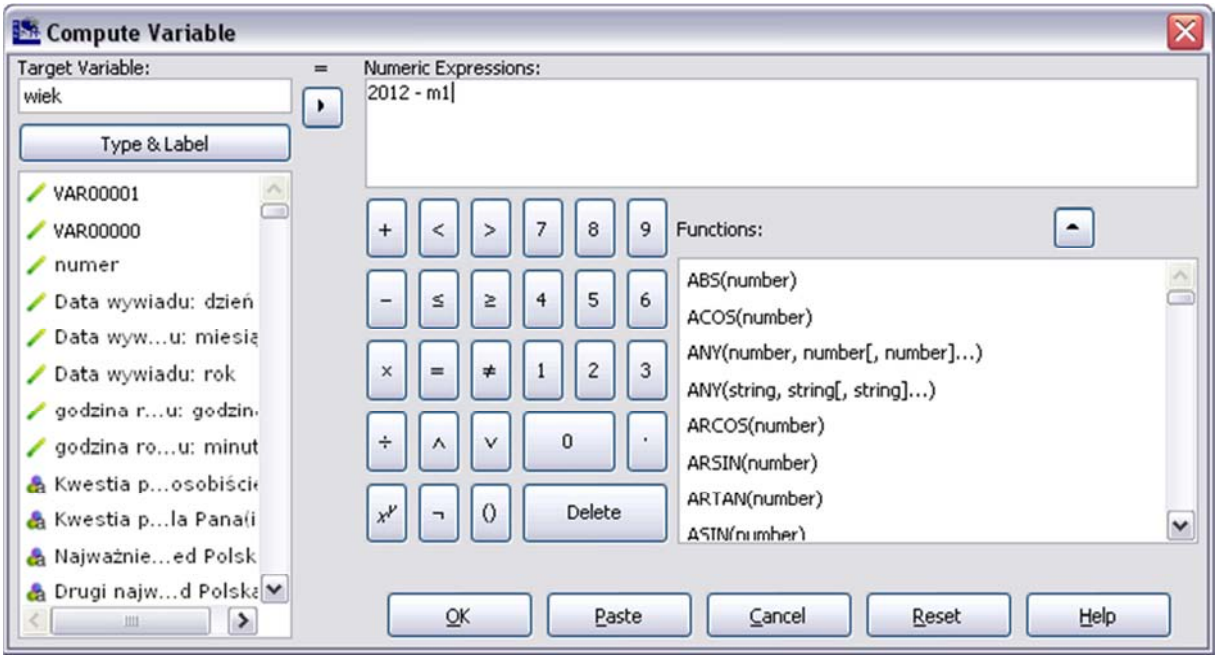
Kolejnym etapem przygotowań zbioru do analizy jest przekształcanie zmiennych. W bardzo rzadkich przypadkach badacz uzyskuje zbiór gotowy bezpośrednio do analizy. Zmienne należy przekształcić na potrzeby analityczne zgodnie z oczekiwaniami badacza wynikającymi z planu badawczego.

8.1. Obliczanie wartości zmiennych (*compute*)

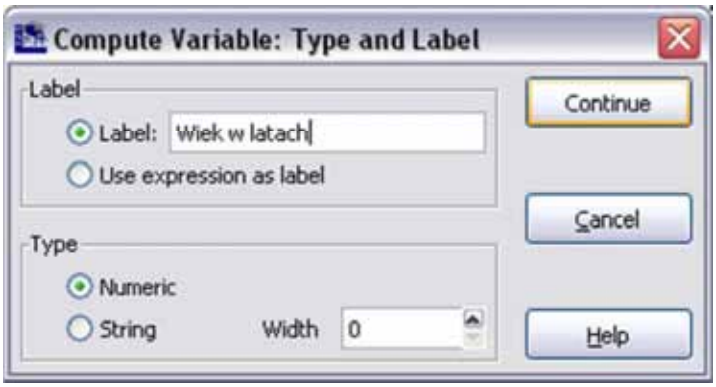
Funkcja obliczania wartości zmiennej umożliwia dokonywanie na niej matematycznych, statystycznych i logicznych operacji. Część z tych operacji może być dokonywana także na danych tekstowych. Wśród funkcji matematycznych znalazły się podstawowe przekształcenia arytmetyczne oparte na dodawaniu, odejmowaniu, mnożeniu, dzieleniu oraz potęgowaniu, a także inne operacje: pierwiastkowanie kwadratowe (SQRT), obliczanie wartości bezwzględnej (ABS), logarytmu naturalnego lub dziesiętnego (LN, LN10), zaokrąglanie do najbliższej liczby całkowitej (RND), itd. Funkcje statystyczne pozwalają między innymi na obliczenie sumy argumentów (SUM), średniej (MEAN), odchylenia standardowego lub wariancji (SD, VARIANCE). Z kolei funkcje logiczne umożliwiają przekształcenia z użyciem operatorów George'a Boole'a (1815-1864), operacje na zakresach danych (RANGE) lub konkretnych wartościach zmiennych (ANY). Program PSPP posiada także obszerny moduł funkcji służących do operowania na datach oraz zmiennych tekstowych. Zastosowanie większości z tych funkcji ma niewielkie znaczenie w naukach społecznych, odwołuje się do ponadelementarnej znajomości statystyki i matematyki, wykracza poza ramy niniejszej publikacji. Poniżej przedstawiamy kilka praktycznych przykładów najbardziej przydatnych i najczęściej stosowanych z nich.

Przykład 1. Proste arytmetyczne przekształcanie zmiennych. W badaniach empirycznych najczęściej informacje o wieku respondenta zbiera się pytając o jego rok urodzenia. Istnieje najmniejsze prawdopodobieństwo pomyłki przy zapisywaniu i przetwarzaniu danych w takim formacie. Następnie w toku rekonfiguracji zbioru przekształcamy tę zmienną na bardziej adekwatnie odpowiadającą potrzebom analitycznym, na przykład na wiek wyrażony w latach. W zbiorze PGSW zmienna zawierająca rok urodzenia respondenta oznaczona jest m1. W celu przekształcenia roku urodzenia na wiek wyrażony w latach należy wykonać działanie: aktualny rok - rok urodzenia.

W menu tekstowym należy wybrać *Transform* ⇨ *Compute*. Następnie wskazaną operację należy zapisać w polu *Numeric Expressions.* Będzie miała ona postać 2012 - m1 (gdzie m1 jest zmienną zawierającą rok urodzenia). W zmiennej docelowej (*Target Variable*) należy wpisać nazwę zmiennej docelowej, np. wiek. Tym sposobem zostanie utworzona nowa zmienna, dołączona na końcu zbioru.



Po kliknięciu na *Type & Label* można zdefiniować tekst etykiety nowej zmiennej wpisując go w pole *Label.* Jeśli wybierzemy *Use expression as label,* wówczas wpisana nazwa zmiennej stanie się jej etykietą. Funkcja *Type* pozwala określić, czy zmienna ma być numeryczna (*Numeric)* czy tekstowa (*String*). Jeśli ma być tekstowa należy określić w polu *Width* maksymalną liczbę znaków, jakie mogą się w niej znaleźć.



Przykład 2. Odnajdywanie zakresów zmiennych. Funkcja obliczania wartości zmiennej może także posłużyć do odnalezienia grupy lub grup zmiennych w zbiorze danych. Przypuśćmy, że interesują nas osoby w wieku od 20 do 30 lat i jednocześnie w wieku od 50 do 60 lat. Wówczas w polu wyrażenia liczbowego (*Numeric Expressions*) należy wprowadzić następującą formułę:

```
RANGE(m1,1952,1962,1982,1992)
```

W formule tej m1 oznacza nazwę zmiennej, a następnie wpisywane są pary liczb mające stanowić zakresy. W efekcie utworzona zostanie na końcu zbioru danych zmienna, która przyjmie wartość 1 dla

osób zawierających się w określonych przedziałach wiekowych i wartość 0 dla osób, które znajdą się poza tymi przedziałami.

Przykład 3. Warunkowe odnajdywanie zakresu zmiennych. Bardzo istotna jest dla analityka umiejętność posługiwania się wyrażeniami Boole'anowskimi. Przypuśćmy, że chcemy utworzyć zmienną, która wskazywałaby żonatych mężczyzn z wyższym wykształceniem, a jednocześnie nie mieszkających na wsi. Formalny zapis na podstawie zmiennych PGSW 2007 prezentuje się następująco:

```
m2=1&(m3=10|m3=11)&m24=1&~m33=1
```

Formalny zapis równania przekształcającego ma charakter standaryzowany - przed znakiem równości znajduje się żądana zmienna, a po znaku równości jej wartość. W tym miejscu należy wprowadzić charakterystykę operatorów logicznych:

& (logiczne *i,* AND) - zwraca prawdę (wartość 1) wtedy i tylko wtedy, gdy oba połączone nią elementy są prawdziwe.

| (logiczne *lub,* OR) - zwraca prawdę (wartość 1) wtedy i tylko wtedy, gdy jeden lub oba połączone nią elementy są prawdziwe.

~ (logiczna negacja, NOT) - zwraca prawdę (wartość 1) wtedy i tylko wtedy, gdy następująca po niej formuła jest fałszem.

Ponadto w tego typu zapisie można stosować inne niż znak równości operatory relacyjne: < (mniejsze niż), > (większe niż), <= (mniejsze bądź równe), >= (większe lub równe) i ~= (nierówne).

Przykład 4. Tworzenie prostego indeksu. Za pomocą obliczania wartości zmiennych można również tworzyć zmienne będące indeksami. Przyjmijmy, że potrzebujemy stworzyć syntetyczną wartość stanowiącą średnią poniższych wartości znajdujących się w zbiorze danych PGSW 2007:

- p51b - Stanowisko PO w kwestii prywatyzacji
- p51c - Stanowisko PO w kwestii dekomunizacji
- p51e - Stanowisko PO w kwestii systemu podatkowego
- p51f - Stanowisko PO w kwestii polityki wobec UE

Należy wówczas wpisać w polu wyrażenia liczbowego następującą formułę:

```
MEAN(p51b,p51c,p51e,p51f)
```

Funkcja MEAN oblicza średnią wartość dla czterech powyższych zmiennych. Zwróćmy uwagę, że funkcje mogą mieć charakter złożony. Jeśli chcemy nadać niektórym z elementów składający się na indeks większą wagę, wówczas możemy do funkcji MEAN wprowadzić działania arytmetyczne:

```
MEAN(p51b*2,p51c*2,p51e,p51f)
```

Przykład 5. Wyszukiwanie jednostek analizy z brakami danych. Przypuśćmy, że chcemy stworzyć jedną grupę jednostek analizy na podstawie faktu, że w pewnej zmiennej (na przykład stan cywilny) zawiera ona braki danych. Wówczas w polu wyrażenia liczbowego wpisujemy:

```
SYSMIS(m24)
```

W efekcie tworzona jest zmienna, w której 1 oznaczy jednostki z brakami danych, a 0 - bez braków danych.

8.2. Obliczanie występowania określonych wartości (*count*)

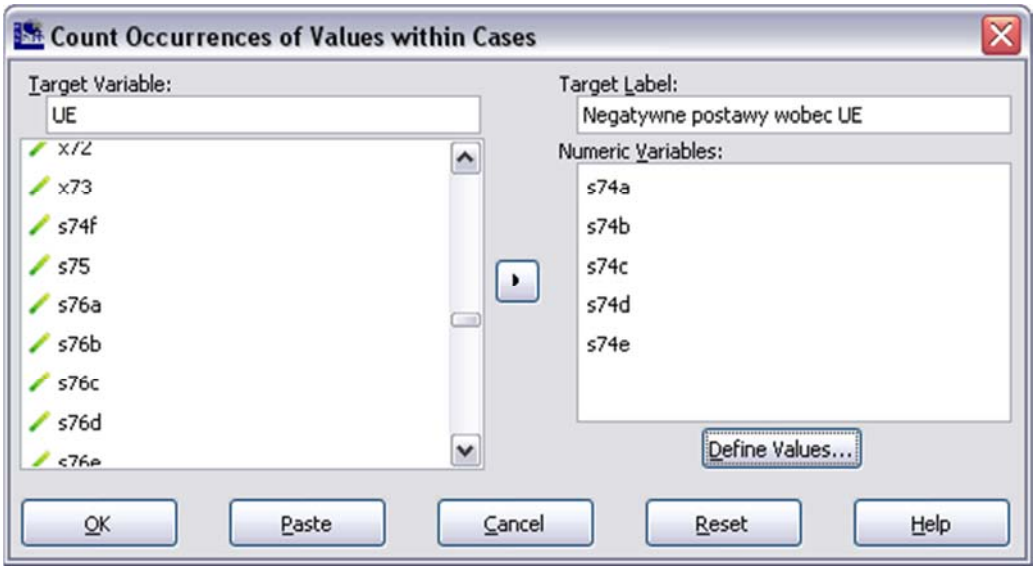
Funkcja ta służy do zliczania określonych wartości zmiennej lub zmiennych. Jest ona szczególnie przydatna w przypadku pytań dychotomicznych, które wymuszają odpowiedzi typu wskazał - nie wskazał, posiada - nie posiada. W wyniku działania tej funkcji tworzona jest zmienna na końcu zbioru, której wartości są zależne od liczby wystąpień określonej wartości. Zliczaniu mogą być poddawane pojedyncze wartości, grupy wartości (dla wielu zmiennych), braki danych, systemowe braki danych oraz przedziały wartości. Zliczania wartości zmiennych można dokonać poprzez wybranie w menu tekstowym *Transform* ⇨ *Count.*

Ilustracją działania tej funkcji jest poniższy przykład. Obliczamy wystąpienia negatywnych postaw wobec Unii Europejskiej na podstawie następujących zmiennych PGSW:

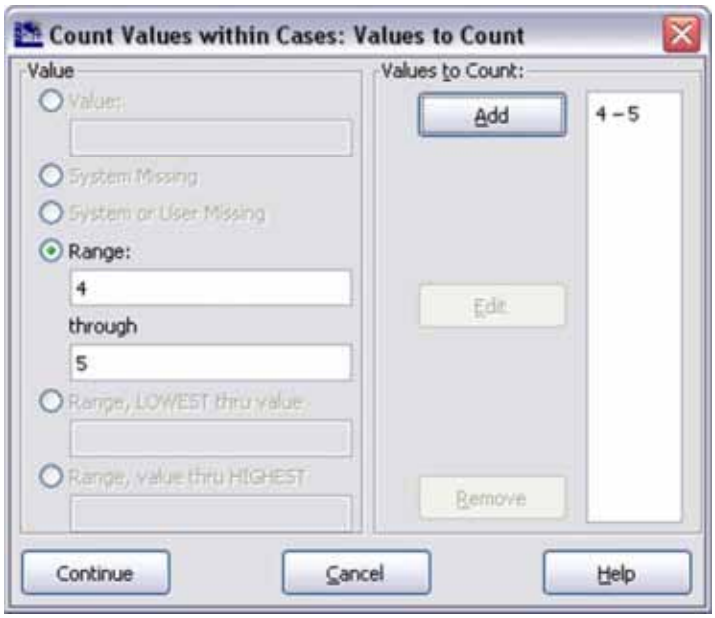
- s74a - Czy członkostwo Polski w UE jest korzystne osobiście?
- s74b - Czy członkostwo Polski w UE jest korzystne dla własnej rodziny?
- s74c - Czy członkostwo Polski w UE jest korzystne dla środowiska zawodowego?
- s74d - Czy członkostwo Polski w UE jest korzystne dla regionu, miejsca zamieszkania?
- s74e - Czy członkostwo Polski w UE jest korzystne dla kraju?

Zakres wartości wymienionych zmiennych jest tożsamy i zawiera się w wartościach od 1 do 5, gdzie 1 oznacza - zdecydowanie korzystne, 2 - raczej korzystne, 3 - ani korzystne, ani niekorzystne, 4 - raczej niekorzystne, 5 - zdecydowanie niekorzystne. Wskaźnikiem negatywnych postaw są liczby 4 i 5. Po wybraniu *Transform* ⇨ *Count* otwiera się okno, w którym wprowadzamy zmienne: s74a, s74b, s74c, s74d, s74e, s74f.

W celu zliczenia tych negatywnych wystąpień w polu *Target Variable* wpisujemy nazwę zmiennej, która ma zostać utworzona (UE). Następnie w pole *Target Label* wpisujemy etykietę zmiennej (negatywne postawy wobec UE). W polu *Numeric Variables* umieszczamy wyżej wymienione zmienne, klikamy przycisk definiowania wartości (*Define Values*).



W oknie definiowania wartości wybieramy zakres wartości (*Range*), wpisujemy wartości 4 i 5, klikamy przycisk *Add,* a następnie *Continue.* W polu tym można wybrać inne opcje. Opcja wartość (*Value*) pozwala na wskazanie konkretnej liczbowej wartości, która ma zostać poddana zliczaniu. Wtedy należy oddzielnie wpisać wartość 4, kliknąć przycisk *Add,* a następnie powtórzyć tą czynność dla wartości 5. Z kolei zaznaczenie opcji systemowych braków danych (*System Missing*) spowoduje, że zliczane będą te zmienne, w których nie figurują żadne wartości. Braki danych systemowe lub użytkownika (*System or User Missing*) zawiera szerszą formułę włączając systemowe braki danych, a jednocześnie te wartości, które zostały zdefiniowane jako braki danych. Z kolei opcja *Range, LOWEST thru value* (zakres od najmniejszej do x) umożliwia zliczenia wszystkich wystąpień zmiennej poniżej wartości liczby wpisanej w polu, a opcja *Range, value thru HIGHEST* (Zakres od x do największej) wszystkich wyższych wartości.



Efektem podjętych wyżej działań będzie utworzenie zmiennej UE, a następnie zliczenie przez program PSPP wszystkich wartości równych 4 lub 5 w zmiennych od s74a do s74e i umieszczenie wyniku obliczeń w nowoutworzonej zmiennej.

8.3. Rangowanie jednostek analizy (*rank cases*)

Rangowanie, czyli zabieg nadawania rang, polega na przyporządkowaniu poszczególnym jednostkom analizy określonych wartości ze względu na pewną cechę lub cechy hierarchicznego miejsca w zbiorze danych. Rangowanie ma zastosowanie wówczas, gdy chcemy uniezależnić się od rozkładu danej zmiennej, a także zapobiec zbytniemu wpływowi na analizy wystąpień obserwacji odstających. Efektem działania tej funkcji jest wygenerowanie zmiennej, której wartości porządkują jednostki analizy nadając im rangi. Zmienna ta umieszczana jest na końcu zbioru. Procedurę uruchamiamy wybierając z menu tekstowego *Transform* ⇨ *Rank Cases*. Pojawia się okno, w którego lewej części znajduje się lista wszystkich zmiennych ze zbioru danych. W oknie *Variable(s)* umieszczamy zmienne, które mają stać się podstawą procedury rangowania. W pełni uzasadnione jest poddawanie rangowaniu zmiennych mierzonych na poziomach ilościowych, a w szczególnych przypadkach mierzonych na poziomie porządkowym. Rangowania można dokonać dla całego zbioru danych lub w podgrupach. W tym ostatnim przypadku w polu *By* musi znaleźć się zmienna, którą wybraliśmy jako podstawę podziału na podzbiory.

Pole *Assign rank 1 to:* umożliwia wyznaczenie porządku rangowania. Zaznaczenie *Smallest Value* (opcja ta zaznaczona jest jako domyślna) powoduje, że najniższa wyznaczona ranga zostanie przyporządkowana najniższej wartości zmiennej będącej podstawą rangowania, a wybranie *Largest Value* odwrotnie - najniższej wartości zmiennej rangowanej przypisana zostanie najwyższa ranga. Zaznaczenie *Display summary tables* powoduje pojawienie się w oknie raportów tekstowej, opisowej informacji o dokonanym przekształceniu.

Przykładowo, na przedstawionym poniżej zrzucie ekranowym jako zmienną rangowaną wybrano liczbę głosów oddaną na posła, a jako podstawę podziału - numer listy, z której poseł startował w wyborach. W efekcie zostały utworzone i nadane rangi odrębne dla każdej z list. Zaznaczone *Largest Value* spowodowało, że posłom, którzy uzyskali największą liczbę głosów przyporządkowano najniższe liczby oznaczające rangi (a więc pierwsze miejsca), a tym którzy uzyskali najniższą liczbę głosów - liczby najwyższe, czyli miejsca końcowe.

Opcja *Rank Types* umożliwia wybranie sposobu nadawania rang. Domyślnym (wybieranym bez zaznaczania) typem rangi są tak zwane rangi proste (*Rank*). Przyporządkowują one liczby naturalne od 1 do n. Funkcję tę wykorzystano w powyższej egzemplifikacji. Ocena (punkty) Savage'a (*Savage score*) to mechanizm oparty na rozkładzie wykładniczym - różnice między kolejnymi nadawanymi rangami będą duże na jednym końcu przedziału i mniejsze (osiągnięcie *plateau*) na drugim. Rangę ułamkową (*Fractional rank*) uzyskujemy przez podzielenie rang regularnych przez liczbę obserwacji danej zmiennej. Przyjmijmy, że analizujemy zbiór zawierający 460 posłów i chcemy nadać każdemu z posłów rangę ułamkową na podstawie liczby głosów nań oddanych. W pierwszej kolejności program PSPP tworzy rangi regularne (od 1 do 460), a następnie przypisuje liczbę 1 posłowi, który otrzymał najniższą liczbę w zbiorze, a 460 temu, który otrzymał najwyższą (ta część procesu jest niewidoczna w programie PSPP). Następnie miejsce w danym rozkładzie dzielone jest przez liczbę jednostek analizy. A zatem poseł o najwyższej liczbie głosów otrzyma rangę 1, a o najniższej - 0,0022 (1/460). Wartości rang ułamkowych są większe od 0, lecz mniejsze lub równe 1. Po pomnożeniu takiej rangi ułamkowej przez 100 proc. otrzymujemy rangę ułamkową wyrażoną w procentach (*Fractional rank as %*). Rangi ułamkowe klasyczne i wyrażane w procentach pozwalają nam na lepszą porównywalność danych w zbiorze z brakami danych, ponadto zabiegi, jak te opisane powyżej, są pierwszym etapem analizy danych takiej jak korelacja rangowa.

Suma wag obserwacji (*Sum of case weights*). W tym przypadku ranga będzie równa liczbie obserwacji w zbiorze. Rangi dla wszystkich zmiennych mają jednakowe wartości. Z kolei w n-tylach (*Ntiles*) dokonywany jest podział zbioru na części w oparciu o wyznaczone grupy n-tyli. Jeśli przyjmiemy wartość 10 (decyle), wówczas zbiór jednostek analizy zostanie podzielony na 10 (w przybliżeniu równych) części.

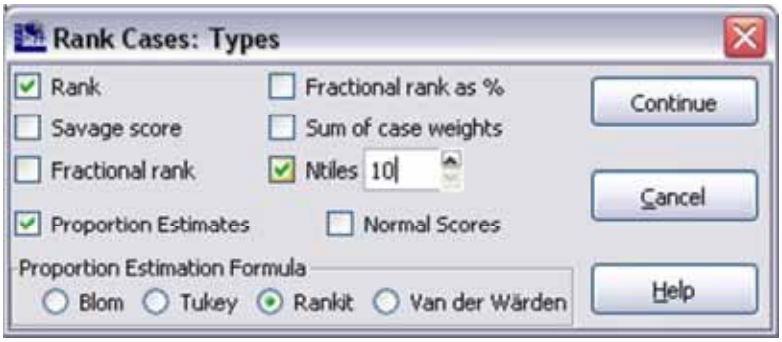
Pierwszej grupie zawierającej pierwsze 10 procent obserwacji zostanie przydzielona wartość rangi 1, piątej 5, a dziesiątej - 10. Program PSPP oferuje również bardziej zaawansowane procedury rangowania. Umożliwia on także tworzenie rang opartych na statystyce Z - wyniki (punkty) normalne (*Normal Scores*), które wyliczają wartości tak zwanej statystyki Z odpowiadające oszacowanemu udziałowi skumulowanemu. Oferowane są również formuły estymacji rozkładu (*Proportion Estimates*). Dla ocen częstości oraz wyników normalnych można wybrać następujące formuły estymacji rozkładu:

- transformacja Bloma dokonywana wedle wzoru: x - 0,375 / y + 0,25, gdzie x oznacza rangę danej jednostki analizy wyrażoną liczbą naturalną, a y liczbę jednostek analizy w zbiorze (tzw. sumę wag obserwacji).

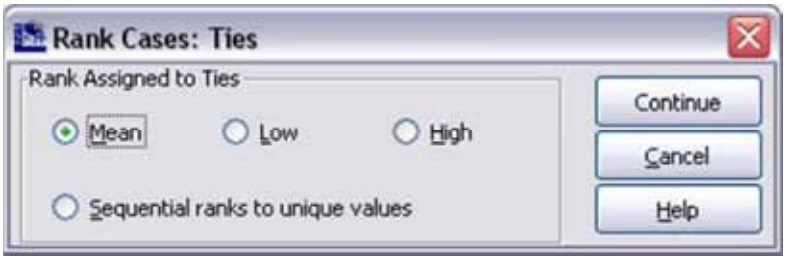
- transformacja Tukey'a wykonywana z wykorzystaniem wzoru: x - 0,3(3) / y + 0,3(3)

- rangowanie (*Rankit*), które funkcjonuje wedle formuły: x - 0,5 / y

- transformacja Van der Wärdena obliczana na podstawie wzoru: x / (y + 1)



Funkcja Wiązania (*Ties*) pozwala na wybór sposobu rangowania tych zmiennych, którym przypisano tożsame wartości rangowanej zmiennej.



Możliwe są następujące sposoby traktowania wartości tożsamych:

- Średnia (*Mean*) - średnia ranga z tożsamych (remisowych) wartości. Jeśli dwóm jednostkom analizy nadano rangę 4, zajęły one 4 i 5 miejsce. Wówczas dokonanie wiązania przypisze obu z nich wartości 4,5.

- Najniższa (*Low*) - obie analizowane jednostki analizy otrzymają rangę najniższą, a więc 4.

- Najwyższa (*High*) - obie analizowane jednostki analizy otrzymają rangę najwyższą, a więc 5.

- Sekwencyjne przypisywanie rang (*Sequential ranks to unique values*) - jednostka analizy, występująca bezpośrednio po jednostkach o rangach tożsamych, otrzyma rangę bezpośrednio następującą po nich:

Tabela 8. Rangowanie bez użycia sekwencyjnego przypisywania rang

Wartość rangowana	Ranga
60	1
66	2
66	2
66	2
70	5

Tabela 9. Rangowanie z sekwencyjnym przypisywaniem rang

Wartość rangowana	Ranga
60	1
66	2
66	2
66	2
70	3

W analizie danych najczęściej wykorzystuje się rangę regularną, a także rangi ułamkowe wyrażone w procentach oraz n-tyle. Pozostałe sposoby rangowania mają w analizie danych społecznych marginalne zastosowanie.

8.4. Rekodowanie wartości zmiennych (*recode*)

Funkcja rekodowania wartości zmiennych umożliwia zmianę wartości istniejących zmiennych w celu ich agregacji, uporządkowania, uspójnienia lub uproszczenia (zmiany poziomu pomiaru zmiennej). Funkcja rekodowania znajduje się w menu tekstowym w *Transform* i posiada ona dwie odmiany - rekodowanie w obrębie tej samej zmiennej (*Transform* ⇨ *Recode into Same Variables*), gdzie zmianom poddawana jest sama zmienna źródłowa oraz rekodowanie na inną, nowoutworzoną na końcu zbioru danych zmienną (*Transform* ⇨ *Recode into Different Variables*). W tym przypadku zmienna źródłowa zostaje zachowana w niezmienionym stanie. Zaleca się używać tej drugiej opcji ze względu na mniejsze prawdopodobieństwo przypadkowego uszkodzenia zbioru danych. Zalecane jest przy dokonywaniu rekodowania posługiwanie się Edytorem składni.

Poniżej przedstawiono przykłady zastosowań rekodowania wartości zmiennych.

Przykład 1. Agregacja zmiennej za pomocą rekodowania. Agregacja zmiennej w wąskim rozumieniu oznaczać będzie rekodowanie bez zmiany poziomu pomiaru (aczkolwiek w szerokim rozumieniu może ona obejmować przypadki zmiany poziomu pomiaru zmiennej, co zademonstrowano w kolejnym przykładzie). W PGSW pomiaru poglądów lewica - prawica dokonywano na skali 11-punktowej. Przypuśćmy, że potrzeba analityczna skłania nas do zastosowania skali pięciopunktowej. Przesłanką może być łatwiejszy odbiór tego typu skali. Zakres wartości oryginalnej zmiennej przedstawia się następująco (zwróćmy uwagę na brak etykiet w wartościach zmiennej z zakresu od 2 do 9, co niewprawnemu odbiorcy analiz może utrudniać interpretację):

-1 - nieuzasadniony brak odpowiedzi
0 - lewica
1 -
2 -
3 -
4 -
5 -
6 -
7 -
8 -
9 -
10 - prawica
97 - trudno powiedzieć

Docelowo należy uzyskać następujący zakres wartości zmiennej:

1 - skrajna lewica
2 - lewica
3 - centrum
4 - prawica
5 - skrajna prawica
6 - brak odpowiedzi

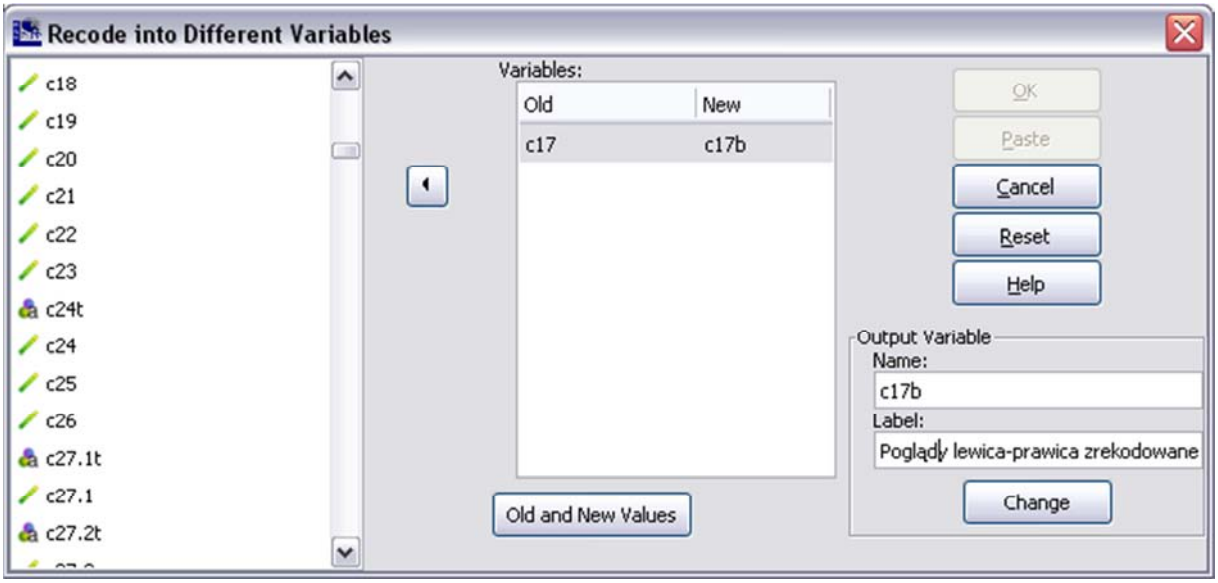
Dokonujemy zatem rekodowania czyli zabiegu zamiany według planu z tabeli 10.

Tabela 10. Przykład agregacji zmiennej z użyciem rekodowania

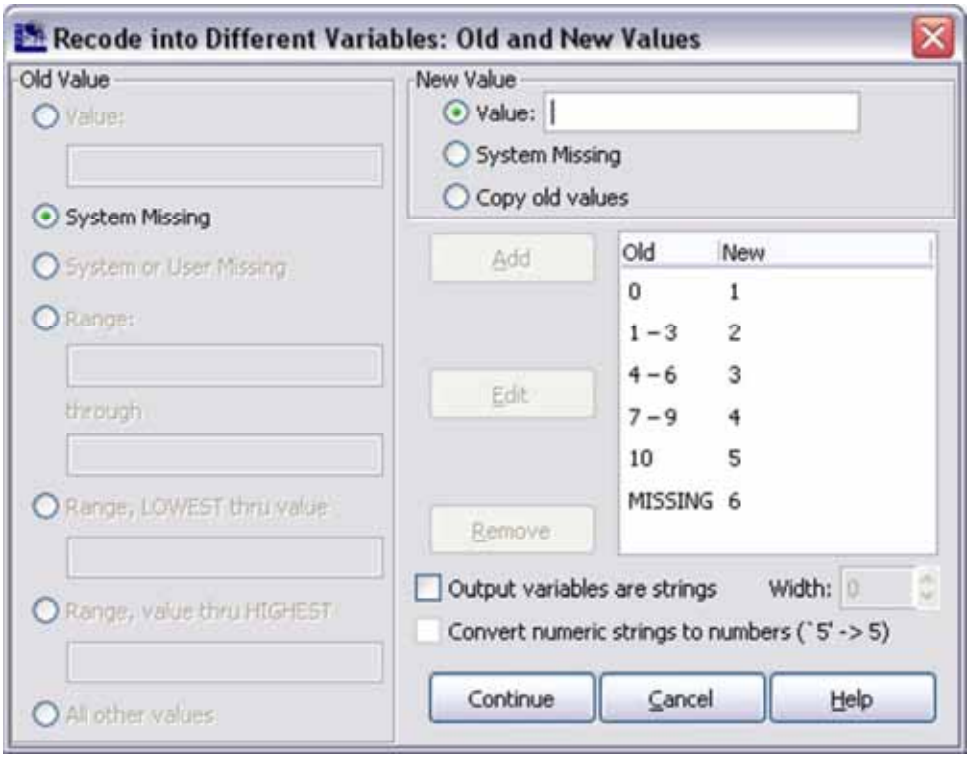
Zakres wartości zmiennej źródłowej	Zakres wartości zmiennej wynikowej
0 - lewica	1 - skrajna lewica
1 -	2 - lewica
2 -	
3 -	
4 -	3 - centrum
5 -	
6 -	
7 -	4 - prawica
8 -	
9 -	
10 - prawica	5 - skrajna prawica
97 - trudno powiedzieć	6 - brak odpowiedzi
-1 - nieuzasadniony brak odpowiedzi	

Przykład ten z merytorycznego punktu widzenia ma charakter umowny, do uzasadnienia są również inne konfiguracje rekodowania powyższych wartości. Chodzi tu o pewną egzemplifikację. W celu dokonania rekodowania według powyższego wzoru wybieramy *Transform* ⇨ *Recode into Different Variables*. Jako zmienną źródłową (*Old*) wstawiamy c17, czyli zmienną przekształcaną. Następnie musimy nazwać nową zmienną, którą chcemy uzyskać po zabiegu rekodowania. Należy pamiętać, aby nazwa była

unikalna i nie powtarzała się z inną nazwą zmiennej, którą mamy w zbiorze danych. W tym przypadku zmienną wynikową nazwano c17b (dodano również opis zmiennej w *Label:* Poglądy lewica-prawica zrekodowane). Następnie klikamy przycisk *Old and New Values,* gdzie będziemy zmieniać wartości zmiennej źródłowej na te, które chcemy uzyskać.



W oknie *Old and New Values* korzystamy z opcji *Old Value - Value* dla wartości zmiennych 0 oraz 10. W *New Value - Value* nadajemy im nowe wartości, odpowiednio: 1 i 5. Dla zakresów zmiennych 1-3, 4-6 oraz 7-9 korzystamy z opcji *Old Value - Range ... through ...* i nadajemy wymienionym zakresom odpowiednio wartości 2, 3 i 4. W celu zrekodowania braków danych użytkownika i systemowych braków danych w *Old Value* wybieramy *System or User Missing*, a następnie w *New Value* wpisujemy wartość 6. Po każdej wyżej wymienionej operacji klikamy przycisk *Add.*



Powyższy ciąg działań w Edytorze składni zapisujemy następująco:

Składnia do wpisania w Edytorze	Opis działania składni
RECODE	- zrekoduj
c17	- zmienną źródłową c17
(0=1) (MISSING=6) (10=5) (1 thru 3=2) (4 thru 6=3) (7 thru 9=4)	- zera zrekoduj na jedynki, wszelkie braki danych na szóstki, dziesiątki na piątki, wartości od jeden do trzy na dwójki ...
INTO c17b .	- to wszystko zapisz w zmiennej wynikowej c17b
VARIABLE LABELS c17b 'Poglądy lewica-prawica zrekodowane'.	- dodaj do zmiennej c17b etykietę o treści 'Poglądy lewica-prawica zrekodowane'
EXECUTE .	- wykonaj powyższą operację.

Przykład 2. Zmiana poziomu pomiaru zmiennej z użyciem rekodowania. Ten przykład wykorzystania rekodowania stanowi najczęstsze zastosowanie rekodowania wartości zmiennych. Z reguły zmienna *wiek* mierzona jest na poziomie interwałowym (rok urodzenia). Na ogół podczas analiz dokonuje się swoistego uproszczenia tej zmiennej, zabiegu przesunięcia jej „w dół skali", na poziom porządkowy w celu utworzenia ze zmiennej ciągłej zmiennej przedziałowej - w tym przykładzie trzech grup wiekowych - młodego pokolenia (od 18 do 35 lat), średniego pokolenia (powyżej 35 do 65 lat) i starszego pokolenia (powyżej 65 lat). W PGSW zmienna rok urodzenia zapisana jest pod nazwą m1. W Edytorze składni zapisujemy powyższą operację następująco:

Składnia do wpisania w Edytorze	Opis działania składni
RECODE	- zrekoduj
m1	- zmienną źródłową m1
(1947 thru 1976=2) (Lowest thru 1946=3) (1977 thru Highest=1) (MISSING=9998)	- urodzonych w latach 1947 - 1976, a więc w 2012 roku mających pomiędzy 26 a 65 lat zrekoduj na dwójki, urodzonych w 1946 roku i wcześniej zrekoduj na trójki, urodzonych w 1977 roku i później zrekoduj na jedynki, wszelkie braki danych zrekoduj na 9998 (domyślnie jest to kod odmowy podania odpowiedzi)
INTO m1b .	- to wszystko zapisz w zmiennej wynikowej m1b
VARIABLE LABELS m1b 'Wiek w latach - zrekodowany'.	- dodaj do zmiennej m1b etykietę o treści 'Wiek w latach - zrekodowany'
EXECUTE .	- wykonaj powyższą operację.

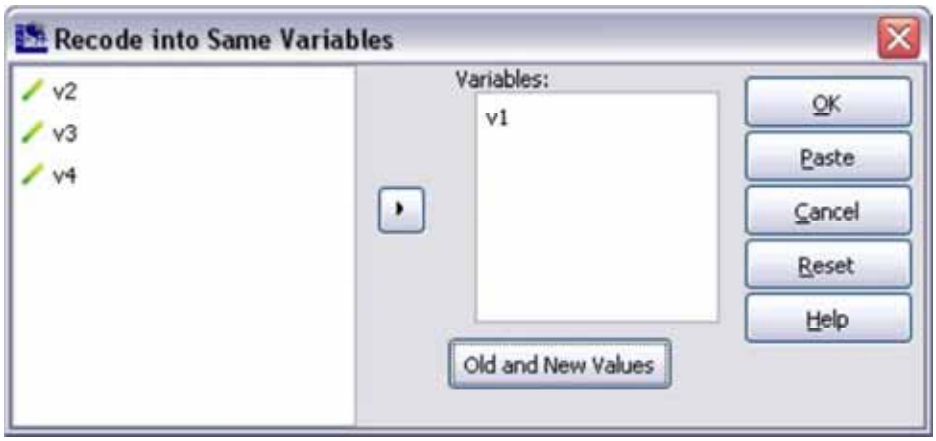
Przykład 3. Uporządkowanie zmiennej za pomocą rekodowania. W wielu sytuacjach analitycznych zachodzi konieczność dokonania zabiegu „odwrócenia skali". Z różnych powodów (często przyczyną jest błąd badacza lub konieczność zadania pytania w takiej formie) dane mogą być zmierzone na poziomie porządkowym z użyciem następujących wartości zmiennych:

1 - zdecydowanie tak
2 - raczej tak
3 - ani tak, ani nie
4 - raczej nie
5 - zdecydowanie nie

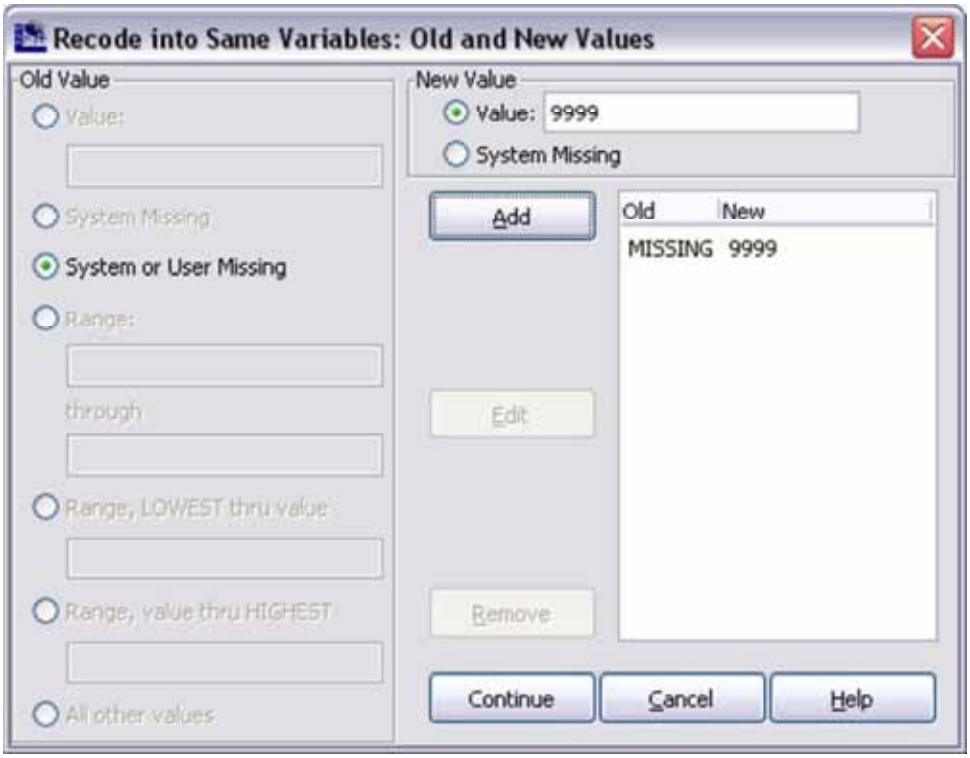
W takiej sytuacji nie możemy potraktować w analizach zmiennej porządkowej za interwałową. Zmienną interwałową uzyskamy wówczas, gdy odwrócimy skalę, a więc wartość najniższa stanie się najwyższą, a najwyższa - najniższą. Zmienną źródłową w tym abstrakcyjnym przykładzie nazwijmy V1, a zmienną wynikową V1b. Formuła takiej operacji w oknie składni jest następująca:

Składnia do wpisania w Edytorze	Opis działania składni
`RECODE`	- zrekoduj
`V1`	- zmienną źródłową V1
`(1=5) (2=4) (3=3) (4=2) (5=1)`	- jedynki ze zmiennej źródłowej zapisz w zmiennej wynikowej jako piątki, dwójki jako trójki, trójki jako trójki ...
`INTO V1b .`	- to wszystko zapisz w zmiennej wynikowej V1b
`VARIABLE LABELS V1b 'Zmienna z odwróconą skalą'.`	- dodaj do zmiennej V1b etykietę o treści 'Zmienna z odwróconą skalą'
`EXECUTE .`	- wykonaj powyższą operację.

Przykład 4. Uspójnienie zmiennej za pomocą rekodowania może polegać na zrekodowaniu systemowych braków danych na braki danych zdefiniowane przez użytkownika. Operacja ta pozwala uniknąć powszechnych pomyłek przy analizach danych. W przeciwieństwie do przykładów 1, 2 i 3 tu możemy posłużyć się zamianą na te same zmienne. Nie nastąpi tu utrata danych, lecz jedynie nadana zostanie spójność już istniejącym. W trybie okienkowym podejmujemy następujące działania: wybieramy *Transform* ⇨ *Recode into Same Variables*, a następnie umieszczamy w polu *Variables* zmienną, którą chcemy zrekodować (w tym przypadku V1).



Następnie wybieramy przycisk *Old and New Values* i w nowym oknie wybieramy w *Old Value* - *System or User Missing,* w *New Value* - *Value* wpisujemy kod 9999 i klikamy *Add.* W polu znajdującym się poniżej ukaże się zapis „MISSING 9999".



W Edytorze składni formuła w tym przypadku jest następująca:

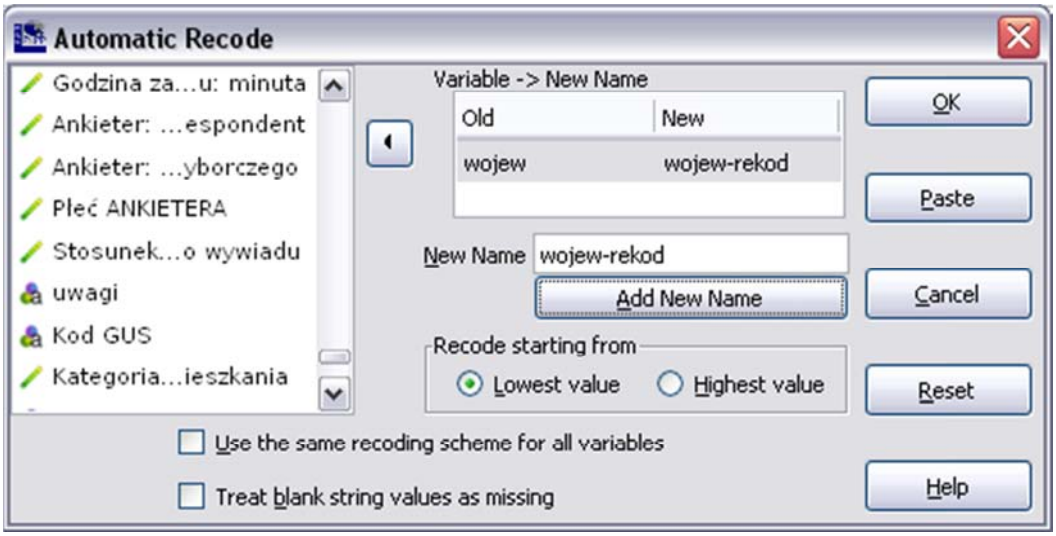
Składnia do wpisania w Edytorze	Opis działania składni
`RECODE`	- zrekoduj
`V1`	- zmienną V1
`(MISSING=9999) .`	- wszelkie braki danych zrekoduj na liczbę 9999
`EXECUTE .`	- wykonaj powyższą operację.

8.5. Automatyczne rekodowanie wartości zmiennych (*automatic recode*)

Automatyczne rekodowanie zmiennych pozwala na stworzenie kopii danej zmiennej z jednoczesnym uporządkowaniem zakresu wartości zmiennej lub zmiennych. Tworzona jest nowa zmienna, a wartości ze starej zmiennej mogą zostać skopiowane bez przekształceń (*Lowest Value*) lub zostać odwrócone (*Highest Value*), to jest najniższa wartość w zmiennej źródłowej będzie wartością najwyższą w zmiennej wynikowej i odwrotnie. Procedura ta powinna znajdować częste zastosowanie. Jeśli nie jesteśmy pewni wyników operacji na określonych zmiennych należy zostawiać ich oryginały w zbiorze danych, a zabiegów dokonywać na ich kopiach.

Automatyczne rekodowanie jednej zmiennej. Jest to najprostszy przypadek zastosowania funkcji automatycznego rekodowania. W poniższym przykładzie poddano rekodowaniu zmienną województwo. Z pola po lewej stronie zawierającego listę wszystkich zmiennych pobrano zmienną *wojew* do pola *Variable ⇨ New Name*. Po kliknięciu na zmienną *wojew* udostępnione zostaje pole *New Name* i przycisk *Add New Name.* W pole *New Name* wpisujemy nazwę zmiennej, którą chcemy utworzyć na podstawie zmiennej źródłowej i klikamy przycisk *Add New Name.* Wybieramy automatyczne rekodowanie bez

przekształceń (*Lowest Value*). Opcja *Treat blank string values as missing* powoduje, że wartości tekstowe typu spacja są kodowane jako braki danych.



Automatyczne rekodowanie wielu zmiennych. Opcja Użyj tego samego schematu rekodowania dla wszystkich zmiennych (*Use the same recoding scheme for all variables*) przeznaczona jest dla rekodowania więcej niż jednej zmiennej. Tworzy ona jednolity standard kodów dla wszystkich rekodowanych zmiennych. Sposób działania tej funkcji ilustrują poniższe pary tabel.

Tabela 11. Rekodowanie zmiennych według tego samego schematu rekodowania dla wszystkich zmiennych

Grupa zmiennych przed automatycznym rekodowaniem		
Zmienna 1	**Zmienna 2**	**Zmienna 3**
1	5	20
2	6	25
3	7	30
4	8	1

Grupa zmiennych po automatycznym rekodowaniu		
Zmienna 1a	**Zmienna 2a**	**Zmienna 3a**
1	5	9
2	6	10
3	7	11
4	8	1

Rekodowanie zmiennych według tego samego schematu rekodowania dla wszystkich zmiennych spowodowało utworzenie jednolitego ciągu kolejno po sobie następujących liczb. Trzy zmienne tworzą jednolitą, logiczną przestrzeń liczbowych własności. Zwróćmy również uwagę na opisy zmiennych w Widoku zmiennych. Opisy wartości zmiennych w każdej ze zmiennych (1a, 2a i 3a) są tożsame i zawierają się w zakresie od 1 do 11.

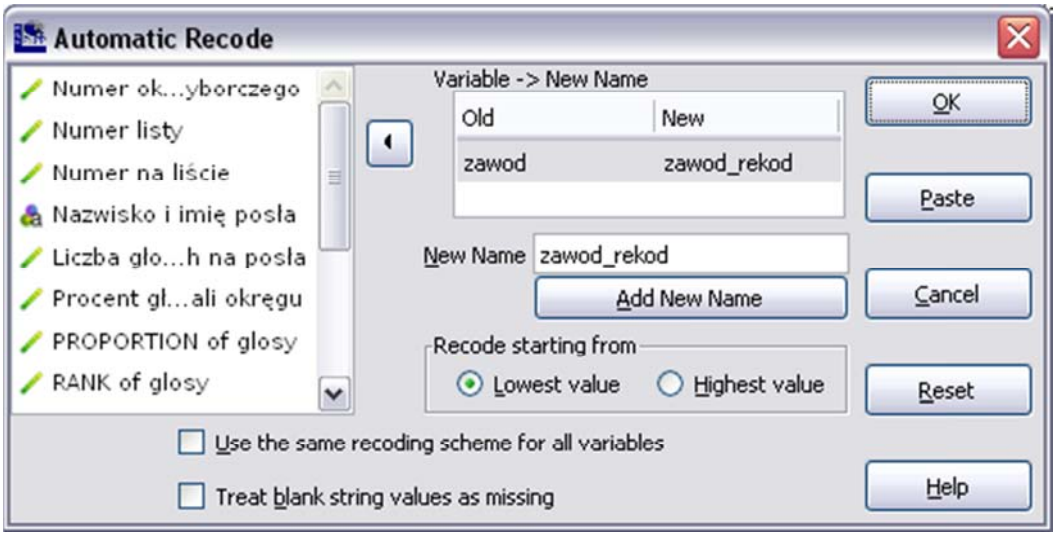
Tabela 12. Rekodowanie zmiennych bez użycia tego samego schematu rekodowania dla wszystkich zmiennych

Grupa zmiennych przed automatycznym rekodowaniem		
Zmienna 1	**Zmienna 2**	**Zmienna 3**
1	5	20
2	6	25
3	7	30
4	8	1

Grupa zmiennych po automatycznym rekodowaniu		
Zmienna 1a	**Zmienna 2a**	**Zmienna 3a**
1	1	2
2	2	3
3	3	4
4	4	1

Z kolei brak użycia funkcji rekodowania zmiennych z użyciem jednolitego schematu powoduje, że każda zmienna traktowana jest jako odrębna i porządkowana wyłącznie w swoim zakresie (np. patrz zmienna 3a). Opisy wartości zmiennych w Widoku zmiennych odnoszą się wyłącznie do samych zmiennych.

Automatyczne rekodowanie zmiennych tekstowych. Jednakże najważniejsze zastosowanie automatycznego rekodowania danych odnosi się do zmiennych, w których wartości nie mają charakteru liczbowego, lecz tekstowy. Funkcja ta umożliwia automatyczną zamianę danych tekstowych na dane liczbowe. Ten zabieg jest często warunkiem przystąpienia do analizy tych danych, ponieważ program PSPP nie wykonuje na ogół analiz na danych tekstowych. Ponadto obliczenia na tak przetworzonych danych są znacznie szybsze. W poniższym przykładzie poddano rekodowaniu zmienną tekstową - zawód posła. Warunkiem właściwego zastosowania procedury rekodowania danych tekstowych jest stworzenie danych w jednolitym formacie. Automatyczne rekodowanie zmiennych tekstowych odbywa się według tego samego schematu jak rekodowanie zmiennych liczbowych.



Efektem automatycznego rekodowania zmiennej tekstowej *zawod* jest zmienna liczbowa *zawod_rekod,* w której ciągi tekstowe ze zmiennej *zawod* stały się opisami wartości zmiennych - przypisano do nich wartości liczbowe według zasady, że tożsame ciągi znaków mają taką samą wartość liczbową. A zatem na przykład wartością 1 zostali oznaczeni wszyscy ci posłowie, u których w zmiennej *zawod* figurowała wartość 1, 2 - administrator, 3 - administratywista, 4 - adwokat, i tak dalej.

8.5.1. Zasady kodowania danych

Kodowanie jest szeroką kategorią pojęciową. W badaniach ilościowych kodowanie danych polega na kategoryzacji do reprezentacji ilościowej opisowych wartości zmiennych. Celem kodowania jest takie spreparowanie danych, które umożliwia poddanie ich analizom statystycznym. Kodowaniu poddawane są odpowiedzi na pytania otwarte i półotwarte. W toku tej procedury zostają one przekształcone w odpowiedzi zamknięte. W celu wyjaśnienia zagadnienia kodowania przedstawiamy podstawy merytorycznej klasyfikacji pytań na otwarte, półotwarte i zamknięte.

Pytanie zamknięte (skategoryzowane, pytanie-wybór, pytanie z wyborem) to takie pytanie, w którym respondent dokonuje wyboru spośród wyszczególnionych wariantów skończonej liczby odpowiedzi.

Tego typu pytań nie kodujemy. Z kolei **pytanie otwarte** to pytanie, w którym respondent udziela odpowiedzi spontanicznie, nie mając do dyspozycji gotowych jej wariantów, a ankieter tę odpowiedź zapisuje. Należy odróżnić pytania otwarte w sensie logicznym (tu liczba i rodzaj możliwych odpowiedzi jest nieskończona) od pytań otwartych w sensie gramatycznym (liczba możliwych odpowiedzi jest skończona i pomimo, że forma odpowiedzi pozostaje otwarta, to istnieje zbiór określonych odpowiedzi, stanowiący zakres danej zmiennej). Procedurze kodowania poddajemy przede wszystkim pytania otwarte w sensie logicznym. Różnice pomiędzy zamkniętością i otwartością w sensie logicznym i gramatycznym przedstawia tabela 13.

Tabela 13. Przykłady pytań otwartych i zamkniętych

		Logicznie	
		zamknięte	**otwarte**
Gramatycznie	**zamknięte**	Proszę powiedzieć, czy brał(a) Pan(i) udział w wyborach w 2011 roku? 1. Tak 2. Nie 3. Nie pamiętam 4. Odmowa odpowiedzi	Dlaczego nabył(a) Pan(i) ten właśnie produkt? 1. Nie miałem(am) wyboru 2. Ze względu na korzystną cenę 3. Z racji estetyki 4. Nie wiem / trudno powiedzieć 5. Odmowa odpowiedzi
	otwarte	Proszę powiedzieć, ile ma Pan(i) lat?	Proszę podać powód, dla którego zagłosował(a) Pan(i) na tego właśnie polityka?
Źródło: Opracowanie własne.			

Z kolei **pytaniem półotwartym** (półzamkniętym) nazywamy pytanie zawierające kilka wariantów odpowiedzi (jak w pytaniu zamkniętym), lecz oprócz tego respondent ma możliwość udzielenia odpowiedzi innej niż wyszczególnione (charakterystyczne dla pytania otwartego). Natomiast **pytanie prekategoryzowane** to takie, które jest zamknięte dla ankietera, lecz otwarte dla respondenta. Oznacza to, że ankieter klasyfikuje odpowiedź respondenta na podstawie uprzednio przygotowanej kafeterii. Wymienione rodzaje pytań lokują się pomiędzy pytaniami zamkniętymi i otwartymi, wymykając się jednoznacznej klasyfikacji. Pytania tego typu także poddajemy zabiegowi kodowania, aczkolwiek ograniczonemu przez zakres wyznaczony logicznym znaczeniem itemów poprzedzających w kafeterii.

Wyróżniamy dwa typy kodowania danych: **ilościowe** i **jakościowe**.

Ilościowe kodowanie danych polega na pogrupowaniu zmiennych w bardziej ogólne, abstrakcyjne kategorie odpowiadające celowi badawczemu. Badacz przygotowuje tak zwaną listę kodową składającą się z ograniczonej liczby kategorii, do których przyporządkowuje poszczególne wypowiedzi respondentów.

Zasady zamiany odpowiedzi otwartych, tekstowych, opisowych, słownych na zmienne liczbowe są następujące:

- Zasada redukcji danych. Kodowanie ilościowe ma na celu świadomą i kontrolowaną redukcję danych, pogrupowanie odpowiedzi respondentów i sprowadzenie ich do mniejszej liczby bardziej ogólnych kategorii, kosztem ich szczegółowości i bogactwa znaczeniowego.

- Zasada rozłączności kategorii. Żaden z utworzonych elementów ogólnej listy kategorii nie może pokrywać się znaczeniowo z innymi. Musi istnieć możliwość jednoznacznego rozstrzygnięcia, do której kategorii przyporządkować daną wypowiedź respondenta.

- Zasada wyczerpywalności kategorii. Przygotowana lista kategorii porządkujących musi umożliwić sklasyfikowanie wszystkich uzyskanych w badaniu odpowiedzi respondentów.

- Zasada wystarczającej liczebności poszczególnych kategorii. Zasada ta postuluje, by w efekcie procedury kodowania wszystkie wyodrębnione kategorie osiągnęły pewne minimum liczebności równe lub przekraczające 10 proc. wszystkich odpowiedzi. W przeciwieństwie do wyżej sformułowanych zasad nie jest to reguła, której należy przestrzegać bezwzględnie. Jest to tylko pewien postulat, który może zostać zniesiony przez badacza w zależności od jego potrzeb analitycznych.

Jakościowe kodowanie danych polega na kodowaniu dosłownym bez agregacji i redukcji danych. Nazywane jest anglojęzycznym słowem *verbatim* (ang. słowo w słowo, dosłownie). Po przeprowadzeniu takiego kodowania kodów liczbowych jest tyle co unikalnych odpowiedzi respondentów. W przypadku takiego kodowania rezygnujemy z agregacji i redukcji danych na rzecz zachowania całego opisowego bogactwa wypowiedzi badanych. Analiza tak zakodowanej zmiennej ma charakter jakościowy, a nie ilościowy. Tego typu kodowanie zapewniane jest przez przedstawioną w poprzednim podrozdziale procedurę automatycznego rekodowania zmiennych.

8.6. Definiowanie braków danych

Brak danych oznacza sytuację, w której pole wartości zmiennej nie zawiera informacji lub oznaczone jest jako informacja znajdująca się poza żądaną przez badacza skalą. Braki danych powstają na różnych etapach procesu badawczego i wynikają z rozmaitych przyczyn. Mogą być one efektem nieudzielenia informacji przez respondenta (odmowy podania odpowiedzi lub - w szerokim rozumieniu - udzielenia odpowiedzi „nie wiem"), braku wprowadzenia informacji lub błędnego jej wprowadzenia przez ankietera, kodera lub analityka danych. Możemy mówić o dwóch typach braków danych: systemowych brakach danych i brakach danych zdefiniowanych przez użytkownika.

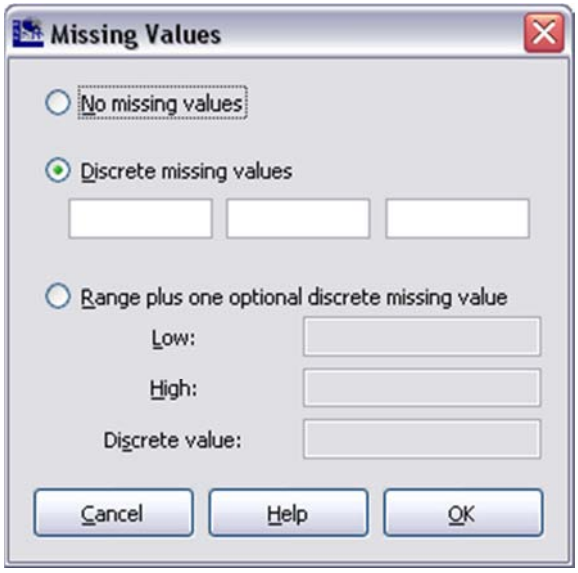
Systemowy brak danych oznacza całkowity brak obecności w polu wartości zmiennej. W Widoku danych braki danych oznaczane są kropką.

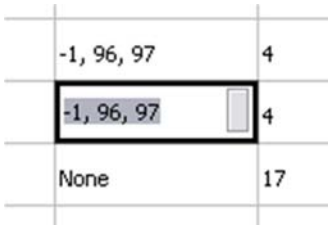
11	2007
11	.
11	.

Brak danych zdefiniowany przez użytkownika musi zostać przezeń oznaczony w Widoku zmiennych w kolumnie *Missing.* Po wybraniu zmiennej otwieramy formantowym przyciskiem komórkę w kolumnie. Do wyboru pozostają dwa typy oznaczania braków danych:

1/ Wartości dyskretne (*Discrete missing values*), gdzie w odpowiednich polach wpisujemy konkretne liczby, które będą oznaczały w zbiorze braki danych. Maksymalnie możemy wprowadzić trzy zmienne.

2/ Alternatywnie możliwe jest wprowadzenie zakresu wartości zmiennych, które uznajemy za braki danych oraz - dodatkowo - jednej zmiennej dyskretnej (*Range plus one optional discrete missing value*).





Nie istnieje w środowisku badaczy jednolity standard regulujący, jakimi liczbami oznaczać braki danych. Najczęściej można zetknąć się z oznaczaniem ich liczbą maksymalną dla danej zmiennej, np. 99, 999, 9999. W tym przypadku często różnicuje się odpowiedzi „nie wiem, trudno powiedzieć" (97, 997, 9997) oraz odmowy odpowiedzi (98, 998, 9998). Rzadziej braki danych oznaczane są kolejną liczbą spoza zakresu wartości zmiennej (np. jeśli zakres wartości zmiennej w danej zmiennej znajduje się w zakresie od 1 do 5, to brak danych oznaczony zostanie cyfrą 6) lub liczbami ujemnymi, np. -1.

8.7. Nadawanie wag jednostkom analizy (ważenie danych)

Nadawanie wag jednostkom analizy (ważenie danych, ważenie próby) to zabieg statystyczny, którego celem jest uczynienie struktury próby tożsamej ze strukturą całej populacji pod względem wybranych cech. W efekcie tego zabiegu zbadana próba ma stać się swoistym, pomniejszonym odbiciem proporcji cech występujących w populacji. Najczęściej w toku ważenia danych bierze się pod uwagę cechy socjodemograficzne takie jak płeć, wiek, wykształcenie, dochody lub wielkość miejsca zamieszkania. Rzadziej są to także rozmaite cechy związane z posiadaniem określonych dóbr lub cechy behawioralne.

Istota statystycznego zabiegu ważenia polega na obniżeniu rangi grup respondentów nadreprezentowanych w zbiorze danych lub podniesieniu rangi grup respondentów niedoreprezentowanych. Jeżeli domyślnie waga każdej jednostki analizy w idealnie dobranej próbie wynosi jeden (taką idealną próbę, gdzie waga dla każdej jednostki analizy wynosi jeden nazywamy **próbą samoważącą** (*self-weighted*)), to waga jednostek nadreprezentowanych musi zostań obniżona poniżej jedności, a jednostki niedoreprezentowane otrzymają wagi powyżej jedności[1]. Ważenie próby stanowi nieodłączny etap przygotowania danych do analizy, koryguje ono błąd systematyczny szacunku parametrów populacji; podwyższa reprezentatywność próby[2]. Opisany powyżej motyw tworzenia wag jest najpowszechniejszy. W literaturze przedmiotu zabieg taki nazywany jest **poststratyfikacją** (kalibracją lub nakładaniem wagi populacyjnej) i często jest on utożsamiany z ważeniem danych. Wiedzę o parametrach populacji koniecznych do ważenia próby można uzyskać między innymi z powszechnie dostępnych źródeł danych, między innymi z Głównego Urzędu Statystycznego (Narodowe Spisy Powszechne). Ze względu na cel dokonywania ważenia wyróżnia się jeszcze dwa typy nakładania wag, rzadziej jednak stosowane. Są to **wagi prawdopodobieństwa** (*sample design, probability weights*) używane wówczas, gdy zaistnieje konieczność zredukowania nierównego prawdopodobieństwa znalezienia się określonych jednostek badawczych w próbie oraz **wagi braków danych** (*non-response weights*), stosowane w celu zrównoważenia odmów udziału w badaniu (chodzi tu o odmowy uczestnictwa - *units non-response*, a nie odmowy odpowiedzi na poszczególne pytania - *items non-response*).

Z praktycznego punktu widzenia zabieg nakładania wag jest swoistym mnożeniem jednostek analizy przez liczbę naturalną lub ułamkową. Tak właśnie, technicznie rozumiane nakładanie wag może odbywać się na dwa sposoby: z wykorzystaniem **wag proporcji** (*proportional weighting*), które służą do uzyskania w zbiorze danych naturalnych wielkości populacji lub poprzez **wagi skali** (*scale weighting*) wyrównujące proporcje cech socjodemograficznych w zbiorze. Waga skali może być jednocześnie wagą proporcji.

Tworzenie wag proporcji. Rozważmy następujący przykład wag proporcji: zbadano 50 kobiet i 50 mężczyzn, jednak rozkład płci w populacji wynosi 60 proc. kobiet i 40 proc. mężczyzn. Oznacza to, że w przykładowej próbce kobiety są niedoreprezentowane, a mężczyźni nadreprezentowani. Wagę tych pierwszych trzeba zwiększyć, a tych drugich - zmniejszyć. Proporcje koniecznych zmian obliczamy według następującego wzoru:

$$\text{Waga proporcji} = \%\ \text{danej grupy w populacji} / \%\ \text{danej grupy w próbie}$$

Po podstawieniu do wzoru wartości dla mężczyzn uzyskujemy:

$$40 / 50 = 0{,}8$$

[1] Warto nadmienić, że nie czyni się tego zabiegu mechanicznie. Na przykład badacze prowadzący Diagnozę Społeczną ograniczają górny zakres wag wskazując, że zbyt duże ich zróżnicowanie jest niekorzystne dla wyników estymacji, gdyż zwiększa wariancję estymatorów. Uważają, że zakres zmienności wag powinien zawierać się dla tego badania w granicach od 0,3 do 3. Wartości wykraczające poza ten przedział przyjmują wielkości równe bliższej z granic tego przedziału. Patrz: *Diagnoza Społeczna 2009*, J. Czapiński, T. Panek (red.), Rada Monitoringu Społecznego Wyższa Szkoła Finansów i Zarządzania w Warszawie, Warszawa 2009, s. 34.

[2] Wyniki licznych badań wskazują, że powszechnie stosowane ważenie poststratyfikacyjne może być nieskuteczne w redukcji obciążeń estymacji o parametrach populacji na podstawie zbadanej próby. R.M. Groves, E. Peytcheva, *The Impact of Nonresponse Rates on Nonresponse Bias a Met-Analysis*, „Public Opinion Quarterly", 2009, 72 (2), s. 167-189. Należy sobie również zdawać sprawę, że fakt stosowania lub niestosowania wag ma istotny wpływ na uzyskane wyniki analiz. Zabieg ten musi być przemyślany, bowiem można zetknąć się w kontekście ważenia danych z zarzutami manipulacji wynikami.

i odpowiednio dla kobiet:

$$60 / 50 = 1{,}2$$

Oznacza to, że do każdej jednostki analizy oznaczonej jako kobieta powinniśmy przypisać wartość 1,2, a do każdej wartości w zbiorze danych oznaczonym jako mężczyzna 0,8.

Wagi proporcji można ustalać dla dwóch i więcej zmiennych jednocześnie. Uzupełnijmy powyższy przykład dotyczący płci o zagregowaną zmienną - miejsce zamieszkania. Dla uproszczenia przyjmijmy, że wyróżniamy tylko wieś i miasto. Tabela 14 pokazuje rozkład badanej populacji dla dwóch skrzyżowanych ze sobą cech socjodemograficznych - płci i miejsca zamieszkania, tabela 15 prezentuje rozkład w zbadanej próbie, natomiast tabela 16 - wynik zastosowania wyżej przytoczonego wzoru dla każdej z cząstkowych komórek tabeli.

Tabela 14. Hipotetyczny rozkład zmiennych socjodemograficznych płeć i miejsce zamieszkania wynikający z proporcji rozkładów w populacji

	Wieś	Miasto	Razem
Kobiety	20	40	60
Mężczyźni	20	20	40
Razem	40	60	100

Tabela 15. Hipotetyczny rozkład zmiennych socjodemograficznych płeć i miejsce zamieszkania w próbie

	Wieś	Miasto	Razem
Kobiety	20	30	50
Mężczyźni	30	20	50
Razem	50	50	100

Tabela 16. Wagi proporcji sporządzone na podstawie hipotetycznego rozkładu zmiennych socjodemograficznych płeć i miejsce zamieszkania w populacji i w próbie

	Wieś	Miasto
Kobiety	0,952381	1,37931
Mężczyźni	0,666667	1

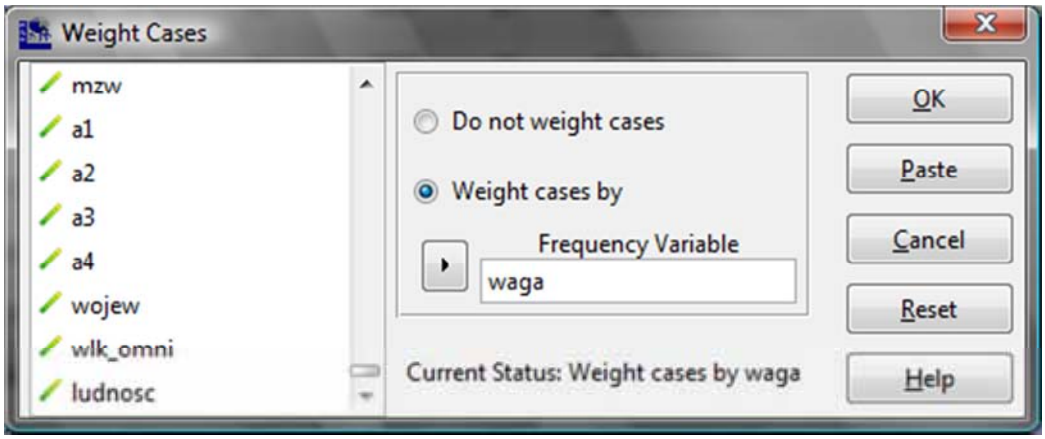
Uzyskane wagi powinny zostać umieszczone przy odpowiednich jednostkach analizy, na przykład wartość 0,666667 powinna znaleźć się przy mężczyznach zamieszkałych na wsi, wartość 1,37931 przy każdej jednostce analizy oznaczonej jako kobieta zamieszkała w mieście.

Tworzenie wag skali. Wagi skali służą do odzwierciedlenia rzeczywistych liczebności występujących w danej populacji. Przyjmijmy, że zbadaliśmy 120 osób, a liczebność danej populacji wynosi 250 000. Wówczas waga każdej jednostki analizy wyniesie 2083,33 (250 000 / 120 = 2083,33). Jest to najprostszy przykład wag skali. Mogą być one również tworzone dla wielu cech, a nie tylko dla całości populacji.

Tabela 17. Wagi skali sporządzone na podstawie hipotetycznego rozkładu zmiennych socjodemograficznych płeć w populacji i w próbie

	Liczebność w próbie	Liczebność w populacji	Waga
Kobiety	400	312 000	780
Mężczyźni	600	399 600	666

W programie PSPP wagi nadawane są ręcznie odrębnie dla grup zmiennych lub poszczególnych zmiennych. Czynność tę można zautomatyzować za pomocą funkcji rekodowania. Natomiast polecenie ważenia danych włączamy poprzez wybranie *Data* ⇨ *Weight Cases*.



W polu po lewej stronie okna wybieramy zmienną ważącą (tzw. wagę) uprzednio przez nas skonstruowaną. Zwyczajowo umieszcza się ją na końcu zbioru danych. Zaznaczamy przycisk radiowy *Weight cases by,* a następnie umieszczamy zmienną ważącą w polu *Frequency Variable.* Należy pamiętać o kontrolowaniu komunikatu na pasku stanu - *Weight by ...* oznacza włączoną wagę, a *Weight off* - wyłączoną.

Wagi można włączać i wyłączać za pomocą Edytora składni. Włączenie wagi o nazwie *waga* następuje poleceniem:

```
WEIGHT BY waga.
```

Wagę wyłączamy wpisując i uruchamiając polecenie:

```
WEIGHT OFF.
```

8.8. Automatyzowanie przekształceń zbioru danych (*do if, do repeat*)

Program PSPP pozwala na automatyzowanie wykonywania sekwencji komend. Oznacza to, że można wykonać więcej niż jedno przekształcenie naraz na podstawie rozmaitych szczegółowych warunków ich wykonywania. Jest to możliwe między innymi za pomocą komend *do if* oraz *do repeat*. Są to komendy, które same nie dokonują bezpośrednio zmian w zbiorze danych, a to oznacza, że muszą współwystępować z innymi. Komendy te w wersji 0.7.9. programu PSPP mogą być stosowane tylko w linii poleceń.

Komenda warunkowa „wykonaj jeżeli" (*do if*). Pozwala na podjęcie sekwencji działań w zależności od zdefiniowanych warunków. Formalna struktura tego polecenia przedstawia się następująco:

```
DO IF [warunek] .
[polecenia] .
ELSE IF [warunek] .
[polecenia] .
ELSE IF [warunek] .
[polecenia] .
...
END IF.
EXECUTE.
```

W warunkach można posługiwać się dowolnymi operatorami logicznymi i relacyjnymi, a w poleceniach - wszelkimi komendami przekształcania zbiorów danych lub zmiennych. W celu zrozumienia jej działania posłużmy się następującym przykładem pochodzącym z podrozdziału dotyczącego nakładania wag (8.7. Nadawanie wag jednostkom analizy...) w podziale na płeć oraz miejsce zamieszkania. Uzyskaliśmy tam następujące hipotetyczne wagi proporcji:

- dla kobiet zamieszkałych na wsi - 0,952381
- dla kobiet zamieszkałych w mieście - 1,37931
- dla mężczyzn zamieszkałych na wsi - 0,666667
- dla mężczyzn zamieszkałych w mieście - 1

W programie PSPP nie musimy ich wprowadzać ręcznie. Może posłużyć do tego celu polecenie *do if*. W zbiorze PGSW zmienna płeć oznaczona jest jako m1 (gdzie wartość 1 oznacza mężczyzna, a wartość 2 - kobieta), a zmienna miejsce zamieszkania zapisana jako wlk_omni zawiera trzy kategorie wielkości miast (1-3) oraz kategorię wieś, której przypisano wartość 4. W celu wprowadzenia wag do zbioru w sposób zautomatyzowany tworzymy zmienną waga2, do której zostaną wpisane żądane wagi. Następnie w edytorze składni wpisujemy następującą serię poleceń:

```
DO IF m1=1&(wlk_omni=1|wlk_omni=2|wlk_omni=3) .
RECODE waga2 (0 = 1) .
ELSE IF m1=1&wlk_omni=4.
RECODE waga2 (0 = 0.666667) .
ELSE IF m1=2&wlk_omni=4.
RECODE waga2 (0 = 0.952381) .
ELSE IF m1=2&(wlk_omni=1|wlk_omni=2|wlk_omni=3) .
RECODE waga2 (0 = 1.37931) .
END IF.
EXECUTE.
```

Zwróćmy uwagę na fakt posługiwania się kropkami przy używaniu liczb ułamkowych. Po uruchomieniu składni odpowiednie wagi proporcji zostaną przypisane do każdej jednostki analizy w nowej zmiennej waga2 .

Komenda warunkowa „powtarzaj sekwencję" (*do repeat*). Jest to elementarna pętla programistyczna. Określone przez użytkownika czynności wykonywane są we wskazanym przezeń zakresie. Komenda ta pozwala zaoszczędzić wiele pracy - zamiast dokonywać określonych zmian lub przekształceń dla każdej zmiennej odrębnie umożliwia zbiorcze przekształcenie grupy zmiennych. Składnia tego polecenia jest następująca:

```
DO REPEAT x=[lista zmiennych podlegających przekształceniom]
/y=[lista wartości zmiennych podlegających przekształceniom] .
[polecenia] .
END REPEAT.
EXECUTE.
```

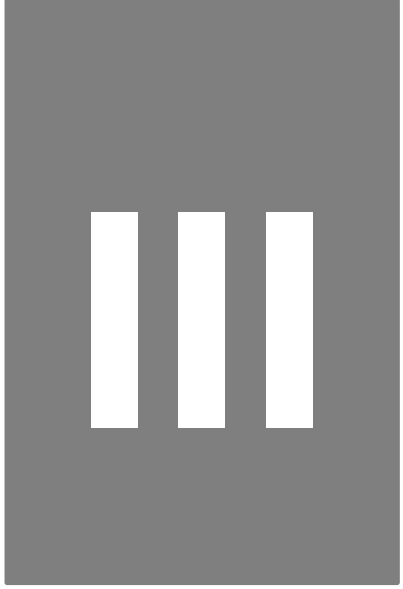
Polecenie to rozpoczynamy komendą DO REPEAT, a kończymy END REPEAT. W linii, w której znajduje się DO REPEAT należy umieścić nazwę sztucznej, pomocniczej zmiennej (w powyższym przykładzie reprezentowanej przez x), a następnie po znaku równości wpisać listę zmiennych. Mogą to być wartości nieciągłe, np. V1, V4, V10, ciągłe - zapisywane wówczas V1 TO V5 (w tym przypadku zmiany zostaną dokonane dla zmiennych V1, V2, V3, V4 i V5) lub obejmować wszystkie zmienne w zbiorze (wówczas wpisujemy parametr ALL). W drugiej linii polecenia deklarujemy sztuczny zakres wartości zmiennych

(y), a po znaku równości wpisujemy, które wartości zmiennych z wymienionych wyżej zmiennych będą podlegać zmianom. Stosujemy tu analogiczne polecenia jak w linii powyżej. W trzecim wierszu, oznaczonym jako [polecenia] wpisujemy komendy przekształceń. Będą one dotyczyły wyżej zdefiniowanych zmiennych i ich zakresów. W tym miejscu można używać dowolnych komend takich jak COMPUTE, COUNT czy RECODE. Sekwencję pętli należy zakończyć komendą END REPEAT.

Poniżej znajduje się przydatny przykład zastosowania komendy powtarzania sekwencji.

```
DO REPEAT var=ALL
/value=ALL.
RECODE var (MISSING=9999).
END REPEAT.
EXECUTE.
```

W efekcie jej wykonania wszystkie wartości braków danych w całym zbiorze danych zostaną zastąpione przez kod 9999 (informuje nas o tym fragment polecenia ALL, czyli zmiana dokonuje się na wszystkich zmiennych w bazie).



Część III. Analiza opisowa - elementarne metody analizy danych

9

Rozdział 9. Analiza częstości występowania zjawisk

Przedmiotem rozdziału są podstawy analizy danych w postaci tabelarycznej (tabel częstości). Zamieszczono w nim zbiór reguł prezentacji danych tabelarycznych, omówiono sposoby tworzenia tabel w programie PSPP, a także zaprezentowano wskazówki na temat zasad ich interpretacji. Podstawą wszelkiej analizy danych jest nabycie umiejętności interpretacji wyników przedstawionych w tabelach. Czynność tworzenia tabel nazywana jest tabulacją. Umożliwia ona zredukowanie ilości danych, obliczenie częstości występowania poszczególnych przypadków w zbiorze danych i przedstawienie wyników badania w czytelnej tabelarycznej formie. Należy pamiętać, że przed przystąpieniem do analizy danych w tabelach konieczne jest zrekonfigurowanie danych, w tym ich kodowanie i rekodowanie.

Wyróżniamy dwa typy tabulacji - **tabulację (tabelaryzację) prostą**, gdy przedmiotem powyższych czynności jest jedna zmienna oraz **tabulację (tabelaryzację) złożoną**, gdy jej przedmiotem jest więcej niż jedna zmienna.

9.1. Zasady prezentacji danych tabelarycznych

Tabele są podstawą analiz danych ilościowych. Wiele klasycznych dzieł, jak na przykład *Ruchliwość społeczna* Pitirima Sorokina, opiera się wyłącznie na analizie danych tabelarycznych, nie odwołując się przy tym do żadnych dodatkowych i wyrafinowanych statystyk. Warunkiem zrozumienia bardziej zaawansowanych analiz danych ilościowych jest dobre przyswojenie zasad tworzenia, prezentowania i interpretacji danych tabelarycznych. Tabele, nazywane także rozkładami marginalnymi, rozkładami brzegowymi lub w skrócie marginesami, posiadają swoistą strukturę.

Prezentuje ją poniższa przykładowa tabela:

Tytuł tabeli → Tabela 18. Podział badanych ze względu na płeć (N=1028)

Główka tabeli → **M1. Płeć respondenta** | **Wskazania respondentów**

Boczek tabeli → Kobiety

Komórki tabeli → 559 | 54,4

Mężczyźni | 469 | 45,6

Razem | 1028 | 100

Źródło: ...

Tytuł tabeli. Tabela zawsze powinna mieć numer porządkowy według kolejności jej występowania w tekście oraz tytuł adekwatny do swojej treści i zawartości. W tytule tabeli powinno podawać się informacje o liczbie jednostek analizy, czyli o liczebności próby badawczej. Oznaczamy ją literą N (prawidłowy zapis jak w tabeli 18). Czasami dla oznaczenia podprób - mniejszych grup wchodzących w skład całościowego badanego zbioru - używana jest mała litera n. Jeśli liczebność dla całości tabeli jest niższa niż 30 jednostek analizy (część badaczy wyznacza większą granicę - N=60, N=100 lub nawet N=120, lecz ekstremalnie rzadko bariera ta umieszczana jest niżej) wówczas na końcu tytułu tabeli, po informacji o liczebności próby, powinno zamieścić się odwołanie (oznaczone gwiazdką), a tuż pod tabelą informację następującej treści: „Liczebności są zbyt małe, aby uogólniać wyniki analiz na badaną populację i zostały podane jako orientacyjne". **Główka tabeli** powinna zawierać skrótowe powtórzenie informacji zawartych w tytule tabeli. Często na początku wpisywanej frazy dopisuje się kod składający się z liter i cyfr oznaczający nazwę zmiennej w zbiorze danych (M1). W raporcie z badania typowa tabela powinna zawierać dwa rodzaje danych: liczebności bezwzględne oraz frakcje, które najczęściej przedstawiane są jako wartości procentowe, rzadziej jako odsetki. Pierwszy rodzaj danych pozwala kontrolować, jak liczne są poszczególne wartości zmiennych - pozwala uniknąć błędu wnioskowania z liczebności niewielkich. Z kolei wartości frakcyjne oraz procentowe informują o udziale poszczególnych wartości zmiennych w stosunku do całkowitej liczebności zbioru. Standaryzują dane i umożliwiają porównywanie częstości występowania danych wartości zmiennej między sobą. **Boczek tabeli** zawiera poszczególne etykiety wartości zmiennych. Powinny być one ułożone w określonym porządku korespondującym z porządkiem tabel w całym raporcie. Jednym z rozwiązań jest ułożenie poszczególnych etykiet zgodnie z porządkiem występowania poszczególnych itemów w kwestionariuszu (czyni się to dla tabel, w których etykiety nadają porządek - na przykład w przypadku skali R. Likerta). Dane tabelaryczne można także porządkować wedle liczebności wskazań zgodnie z porządkiem malejącym - w górnych wierszach tabeli prezentowane są wartości zmiennej, które uzyskały najwyższą liczbę wskazań, a w dolnych - najniższą. W celu prezentowania niektórych zmiennych przyjmuje się niekiedy porządek rosnący. Dotyczy to na przykład zmiennej socjodemograficznej jak wiek podawany w przedziałach: w wierszu tuż pod główką tabeli znajdą się informacje o najmłodszej kategorii wiekowej, a na samym dole - o najstarszej. Czasami wartości zmiennej prezentowane są w porządku alfabetycznym według nazw etykiet. Przykładem może być zmienna województwo, począwszy od dolnośląskiego, a skończywszy na zachodniopomorskim.

W **komórkach tabeli** umieszczane są poszczególne wartości liczbowe, lecz nie znak procenta, ani innych jednostek miary, bowiem znajdują się one już w główce tabeli. Należy zdecydować, jaka liczba miejsc po przecinku będzie prezentowana w wartościach procentowych we wszystkich tabelach w raporcie

- wystarczające jest prezentowanie liczb dziesiętnych, dopuszcza się prezentowanie setnych, lecz już nie tysięcznych. Komórek tabeli, które nie mogą być wypełnione liczbami ze względu na brak wskazań, nie prezentuje się w przypadku listy zmiennych nominalnych (niektórzy badacze uznają, że w takich przypadkach wartości znacznie poniżej 5 procent lub nawet poniżej 10 proc. są bezużyteczne i nie prezentują ich - zależy to jednak od decyzji samego badacza), a w pozostałych przypadkach uzupełnia się zerami. W komórkach, gdzie mamy do czynienia z logicznym brakiem danych (odpowiedź nie mogła się pojawić) - wstawiamy znak kropki. W tabeli powinno występować podsumowanie - wartości procentowe muszą sumować się do 100, a wartości liczbowe - do liczby N podanej w główce tabeli. W przypadku pytań, na które respondent może udzielić więcej niż jednej odpowiedzi (tzw. pytania wieloodpowiedziowe), należy skorzystać z odwołania w tytule tabeli, a pod nią umieścić następujące zdanie: „Podane w tabeli wartości nie sumują się do 100 proc., ponieważ na pytanie respondent mógł udzielić więcej niż jednej odpowiedzi". W sytuacji, gdy po zaokrągleniu wartości procentowe nie sumują do 100 (zdarza się, że otrzymujemy wynik 99,9 proc. lub 100,1 proc.), stosujemy „zasadę świętego Mateusza"[1] - w pierwszym przypadku odejmujemy 0,1 od wartości zmiennej najniższej, a w drugim dodajemy 0,1 do najwyższej wartości zmiennej w tabeli, tak by otrzymać 100 proc.

Tuż poniżej tabeli podajemy **źródło** danych, chyba że jest ono jedno dla całości raportu i przywołaliśmy je już we wstępie do analiz. Źródło podajemy za każdym razem (za pierwszym razem w formie rozwiniętej, w kolejnych tabelach - w formie skróconej, zgodnie z zasadami tworzenia przypisów), gdy w tekście odwołujemy się do licznych zbiorów danych lub też przytaczamy dane pierwotne (zebrane w toku badań własnych). W tym ostatnim przypadku podajemy następującą informację: „Źródło: badania własne". Ważne jest, aby podkreślić wkład własnej pracy w analizę danych, jeśli cytujemy dane wtórne. Podajemy wówczas: „Źródło: opracowanie własne na podstawie...". Sposób cytowania danych wtórnych na ogół jest podawany przy zbiorach danych. Na przykład Polski Generalny Sondaż wyborczy na prośbę jego autorów przywołujemy następująco:

Polskie Generalne Studium Wyborcze 2007, pod kierownictwem Radosława Markowskiego, afiliowane przy Instytucie Studiów Politycznych PAN, dofinansowane przez tę instytucję, oraz przez: Ministerstwo Nauki i Szkolnictwa Wyższego, Wissenschaftszentrum Berlin fur Sozialforschung (WZB), Polską Konfederację Pracodawców Prywatnych Lewiatan, Fundację Batorego, Instytut Filozofii i Socjologii PAN oraz instytucję badawczą realizującą sondaż - PBS DGA.

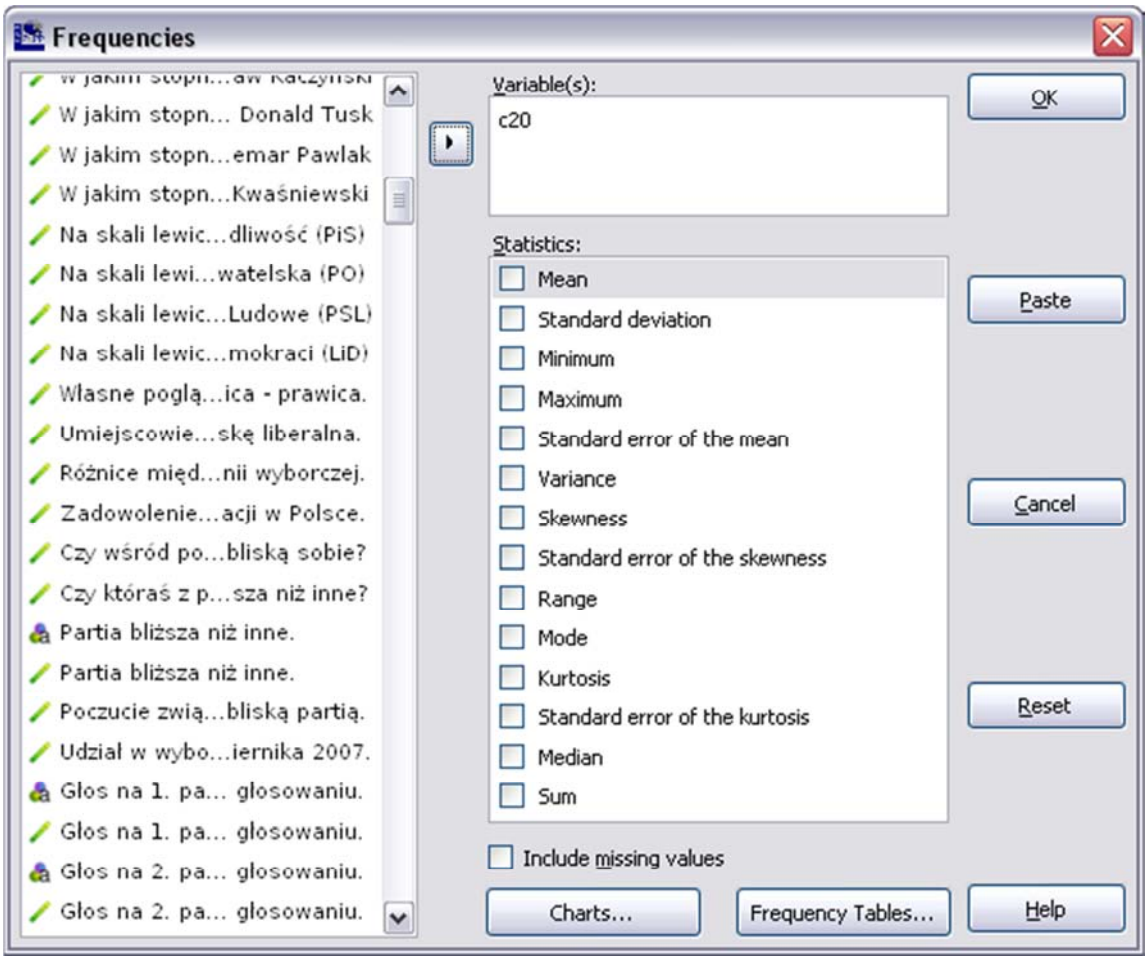
Nie musimy bezwzględnie stosować się do podanego przez autorów wzorca cytowania. Można i należy dostosować sposób cytowania do przyjętego we własnej pracy, jednak warunkiem jest zachowanie wszystkich istotnych podanych przez autorów informacji.

Wszystkie wartości w tabeli oraz w główce tabeli powinny zostać wyśrodkowane w pionie i poziomie. Z kolei w boczku tabeli należy środkować tekst tylko w pionie, a w poziomie przyciągnąć do lewej strony. Tabela powinna mieć czcionkę czytelną i jednolitą dla tabel (chyba że to niemożliwe, bowiem używamy obszernych, złożonych tabulacji) oraz dla całego tekstu. Szerokość tabeli nie może przekraczać szerokości kolumny, w której się znajduje. W miarę możliwości tabela nie powinna dzielić się na wiele kart raportu, a wiersze i kolumny tabeli powinny przyjmować (jeśli da się to osiągnąć) tę samą szerokość (w pionie i poziomie).

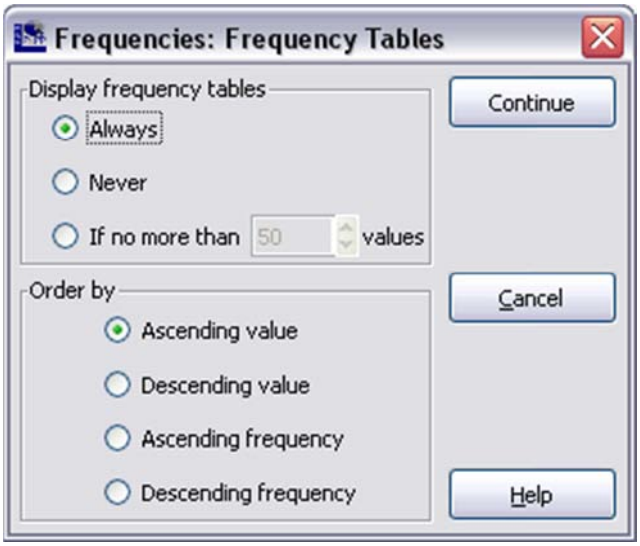
[1] Zasadę tę nazwano w praktyce badawczej odwołując się do chrześcijańskich pism religijnych, konkretnie do *Przypowieści o talentach* zapisanej w Ewangelii według Mateusza. U źródła czytamy: „Bo każdemu, kto ma, będzie dane i będzie miał w obfitości; a temu, kto nie ma, zabrane zostanie nawet to, co ma" (Mt 25, 29).

9.2. Tworzenie tabel dla jednej zmiennej (tabulacje proste)

W programie PSPP tworzenie tabel dla jednej zmiennej odbywa się poprzez wybranie w menu tekstowym *Analyze* ⇨ *Descriptives* ⇨ *Frequencies*. W oknie, z listy po lewej stronie, zawierającej pełny zestaw zmiennych ze zbioru danych wybieramy jedną lub więcej zmiennych do przedstawienia w formie tabelarycznej i umieszczamy ją lub je w oknie *Variable(s)*. Przedtem należy upewnić się, że zmienne zostały właściwie zrekodowane i opisane, a braki danych zlikwidowane lub oznaczone. Domyślnie braki danych nie powinny być uwzględniane w tabeli wynikowej - decyduje o tym odznaczony *check-box: Include missing values.* Zaznaczenie tej opcji powoduje, że braki danych są uwzględniane w obliczeniach. Zostają one włączone do wyników prezentowanych w tabelach i umieszczane na ich końcu z etykietą brak danych (*missing)*.



Funkcja *Frequency Tables* umożliwia zmodyfikowanie sposobu wyświetlania tabeli. Tabele mogą być wyświetlane (*Display frequency tables*) zawsze (*Always*), nigdy (*Never*) lub też dopiero wtedy, gdy osiągną wystarczającą liczebność założoną przez badacza (*If no more then ... values*). Zmienne w tabeli mogą być porządkowane według wartości przypisanych zmiennym (rosnąco - *Ascending value* lub malejąco - *Descending value*) oraz wedle liczby wskazań (rosnąco - *Ascending frequency* lub malejąco - *Descending frequency*).



Powyższych operacji można (i zaleca się) dokonywać w Edytorze składni. Typowy zestaw komend konieczny do stworzenia tabulacji prostej interpretujemy następująco:

Składnia do wpisania w Edytorze	Opis działania składni
FREQUENCIES	- stwórz tabelę lub tabele
/VARIABLES= m3	- ze zmiennej lub zmiennych (w tym przypadku jednej zmiennej - m3)
/FORMAT=AVALUE TABLE	- porządkuj w tabeli wedle wartości zmiennych
/STATISTICS=NONE.	- nie dołączaj do tabeli żadnych dodatkowych statystyk

W efekcie wykonania powyższych komend w Oknie raportów pojawia się tabela:

Wykształcenie

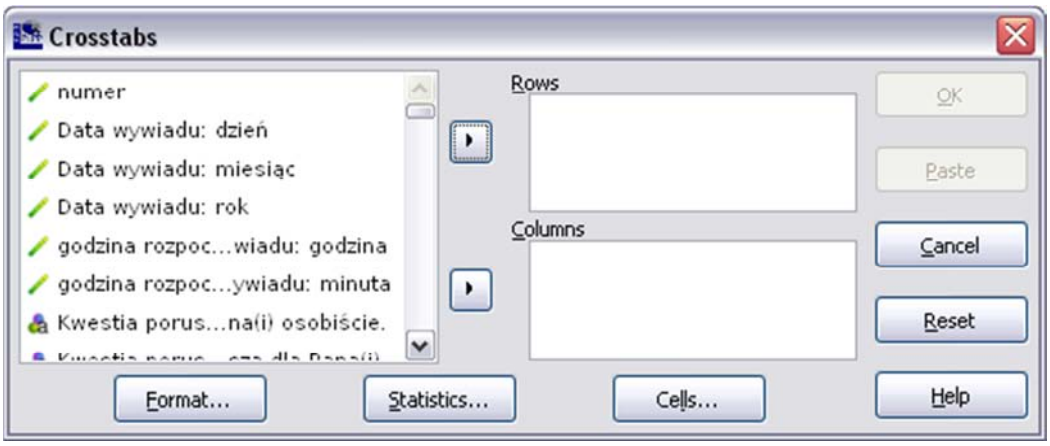
Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
nie ma żadnego wykształcenia	1	5	,28	,28	,28
podstawowe nieukończone	2	22	1,21	1,21	1,49
podstawowe	3	307	16,90	16,94	18,43
zasadnicze zawodowe	4	455	25,04	25,11	43,54
średnie nieukończone	5	33	1,82	1,82	45,36
średnie zawodowe	6	399	21,96	22,02	67,38
średnie ogólnokształcące	7	225	12,38	12,42	79,80
pomaturalne	8	66	3,63	3,64	83,44
wyższe nieukończone (6 semestrów lub więcej)	9	32	1,76	1,77	85,21
licencjat lub trzyletnie studia zawodowe	10	27	1,49	1,49	86,70
wyższe	11	241	13,26	13,30	100,00
nieuzasadniony brak odpowiedzi	-1	5	,28	Missing	
	Total	1817	100,0	100,0	

W tytule tabeli figuruje etykieta zmiennej, w boczku tabeli podane są poszczególne etykiety wartości zmiennych, a same wartości zmiennych w kolumnie *Value.* Kolumna Częstości (*Frequency*) przedstawia liczebności wskazań respondentów na poszczególne odpowiedzi. W kolejnej kolumnie figurują bezwzględne wartości procentowe włączające również braki danych. W kolumnie Procent ważnych (*Valid percent*) przedstawiane są wartości nieuwzględniające braków danych. Ostatnia z kolumn zawiera procent kumulatywny (*Cumulative percent*) - sumuje wynik dla danej wartości oraz wyniki dla wartości zmiennych poprzedzających w tabeli. Na przykład procent kumulatywny dla wykształcenia podstawowego składa się z trzech zsumowanych wartości: dla wykształcenia podstawowego, nieukończonego podstawowego oraz braku wykształcenia: 0,28 + 1,21 + 16,94 = 18,43.

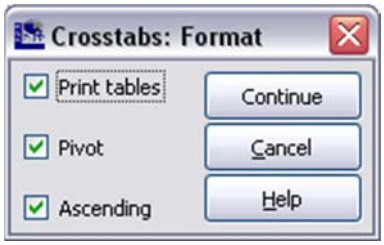
Funkcje graficznej prezentacji danych (wykresów kołowych oraz histogramów) nie są jeszcze dopracowane pod względem estetyki w programie PSPP (program SPSS także pozostawia w tym względzie wiele do życzenia). Nie zaleca się ich stosowania, lepszym rozwiązaniem wydaje się wykonanie rozmaitych wykresów w programach do tego przeznaczonych. Zadowalające pod względem estetycznym formy prezentacji danych można uzyskać na przykład w arkuszu kalkulacyjnym Gnumeric.

9.3. Tworzenie tabel dla dwóch i więcej zmiennych (tabulacje złożone)

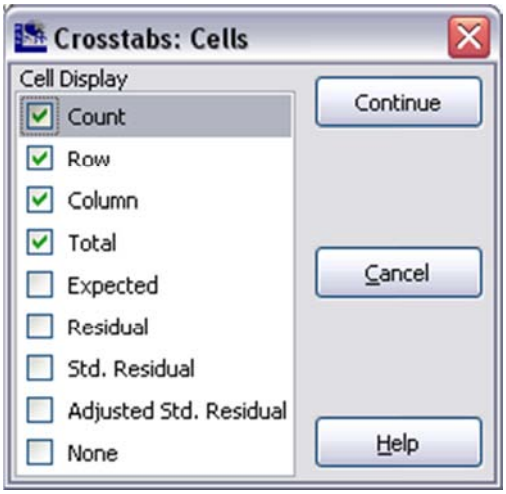
Tabulacje złożone to tabele dla więcej niż jednej zmiennej. Najprostszy przykład stanowi tabela dla dwóch zmiennych zwana tabelą krzyżową, lub nieprawidłowo tabelą kontyngencji lub korelacją. W takiej tabeli prezentowane są dane dla jednej zmiennej w wierszach, a dla drugiej w kolumnach. Istotne jest przemyślenie logicznego i estetycznego układu obu zmiennych. Tabele krzyżowe uzyskujemy wybierając *Analyze* ⇨ *Descriptives* ⇨ *Crosstabs*. W otrzymanym po wykonaniu tego polecenia oknie wstawiamy w polu Wiersze (*Rows*) te zmienne, które będą ukazywać się w wierszach i analogicznie w pole Kolumny (*Columns*) te zmienne, które będą tworzyły kolumny tabeli.



Funkcje ukryte pod przyciskiem formatowania tabel (*Format*) pozostawiamy zaznaczone. Drukuj tabele (*Print Tables*) powoduje wyświetlanie się tabeli krzyżowej w Oknie raportów (uzasadnieniem dla oznaczenia tej opcji jest chęć uzyskania samych dodatkowych statystyk zawartych w *Statistics*). Opcja tabel przestawnych (*Pivot*) pozwala na uzyskanie tabel, w których dodatkowo w boczku powinna być umieszczana etykieta zmiennej prezentowanej w wierszach, tak aby po wyeksportowaniu można było dokonać transpozycji. **Opcja ta nie działa w wersji 0.7.9 programu PSPP.** Z kolei opcja *Ascending,* gdy jest zaznaczona porządkuje wiersze tabeli wedle wartości zmiennych rosnąco, a gdy niezaznaczona - malejąco.



Następnie należy wybrać sposób procentowania w tabeli. Okno z wyboru pojawia się po wybraniu przycisku Komórki (*Cells*).



Można wskazać następujące opcje konfiguracyjne tabeli:

- Liczebność (*Count*) - powoduje wyłączenie liczebności; w komórkach tabeli podawane są wyłącznie wartości procentowe,

- Wiersz (*Row*) - jeśli zostanie zaznaczone sumowanie będzie następowało w wierszach,

- Kolumna (*Column*) - gdy zaznaczymy tę opcję, sumowanie będzie następowało w kolumnach,

- Razem (*Total*) - w tym przypadku sumowane będą ze sobą wszystkie komórki tabeli.

Oznaczenie opcji Wiersz, Kolumna lub Razem ma kluczowe znaczenie dla prezentacji, a następnie interpretacji wyników. Przyjrzyjmy się trzem kolejnym tabelom z oznaczonymi różnymi sposobami sumowania.

W tabeli 19 zaznaczono opcję sumowania w wierszach. W tym przypadku otrzymujemy informację o tym, jakie są proporcje płci w poszczególnych grupach wykształcenia. Sumowania do 100 proc. następują w ostatniej kolumnie. Dowiadujemy się na przykład, że największa dysproporcja dotyczy wykształcenia pomaturalnego. Spośród wszystkich osób, które posiadają takie wykształcenie zaledwie 24,2 proc. stanowią mężczyźni. Z kolei ostatni wiersz zatytułowany Razem (*Total*) mówi nam, że w całej badanej grupie (próbie) znalazło się 37,9 proc. mężczyzn i 60,3 proc. kobiet. Zwróćmy uwagę na informacje zawarte w tytule tabeli. Podawane są zmienne ze sobą krzyżowane oraz oznaczona przez użytkownika zasada procentowania i inne elementy konfiguracyjne. W tabeli 19 jest to sumowanie w wierszach (*row*) oraz podawanie liczebności bezwzględnych (*count*).

Tabela 19. Rozkłady zmiennych wykształcenie i płeć - sumowanie w wierszach (ROW)

Wykształcenie * Płeć [count, row %].

Wykształcenie	*Płeć* mężczyzna	kobieta	Total
nie ma żadnego wykształcenia	2,0 40,0%	3,0 60,0%	5,0 100,0%
podstawowe nieukończone	3,0 13,6%	19,0 86,4%	22,0 100,0%
podstawowe	107,0 34,9%	200,0 65,1%	307,0 100,0%
zasadnicze zawodowe	232,0 51,0%	223,0 49,0%	455,0 100,0%
średnie nieukończone	15,0 45,5%	18,0 54,5%	33,0 100,0%
średnie zawodowe	168,0 42,1%	231,0 57,9%	399,0 100,0%
średnie ogólnokształcące	57,0 25,3%	168,0 74,7%	225,0 100,0%
pomaturalne	16,0 24,2%	50,0 75,8%	66,0 100,0%
wyższe nieukończone (6 semestrów lub więcej)	11,0 34,4%	21,0 65,6%	32,0 100,0%
licencjat lub trzyletnie studia zawodowe	9,0 33,3%	18,0 66,7%	27,0 100,0%
wyższe	100,0 41,5%	141,0 58,5%	241,0 100,0%
Total	720,0 39,7%	1092,0 60,3%	1812,0 100,0%

Odmienną sytuację interpretacyjną stwarza wykonanie sumowania wartości procentowych w kolumnach. W tym przypadku odczytujemy jaki jest rozkład wykształcenia odrębnie dla poszczególnych kategorii płci. Otrzymujemy tu na przykład informację, że wśród wszystkich kobiet wykształcenie wyższe posiada blisko 14 proc. Umiejscowiona jako ostatnia po prawej kolumna Razem (*Total*) wskazuje nam, ilu respondentów znalazło się w danej kategorii wykształcenia. Z kolei w najniższym wierszu wartości sumują się do 100 proc. Zwróćmy uwagę, że pomimo faktu, iż liczebności w komórkach pozostały te same - zmieniły się procentowania.

Tabela 20. Rozkłady zmiennych wykształcenie i płeć - sumowanie w kolumnach (COLUMN)

Wykształcenie * Płeć [count, column %].

Wykształcenie	*Płeć* mężczyzna	kobieta	Total
nie ma żadnego wykształcenia	2,0 ,3%	3,0 ,3%	5,0 ,3%
podstawowe nieukończone	3,0 ,4%	19,0 1,7%	22,0 1,2%
podstawowe	107,0 14,9%	200,0 18,3%	307,0 16,9%
zasadnicze zawodowe	232,0 32,2%	223,0 20,4%	455,0 25,1%
średnie nieukończone	15,0 2,1%	18,0 1,6%	33,0 1,8%
średnie zawodowe	168,0 23,3%	231,0 21,2%	399,0 22,0%
średnie ogólnokształcące	57,0 7,9%	168,0 15,4%	225,0 12,4%
pomaturalne	16,0 2,2%	50,0 4,6%	66,0 3,6%
wyższe nieukończone (6 semestrów lub więcej)	11,0 1,5%	21,0 1,9%	32,0 1,8%
licencjat lub trzyletnie studia zawodowe	9,0 1,3%	18,0 1,6%	27,0 1,5%
wyższe	100,0 13,9%	141,0 12,9%	241,0 13,3%
Total	720,0 100,0%	1092,0 100,0%	1812,0 100,0%

Z kolei po zaznaczeniu opcji Razem (*Total*), tak jak prezentuje to tabela z podstawę procentowania stanowią wszyscy respondenci. Uzyskujemy informację o odsetku danej kategorii płci i wykształcenia do całości. W tym konkretnym przypadku możemy powiedzieć, że kobiety z wykształceniem wyższym stanowią 5,5 proc. populacji, podczas gdy mężczyźni - 7,8 proc.

Tabela 21. Rozkłady zmiennych wykształcenie i płeć - sumowanie dla całości (TOTAL)

Wykształcenie * Płeć [count, total %].

Wykształcenie	*Płeć* mężczyzna	kobieta	Total
nie ma żadnego wykształcenia	2,0 ,1%	3,0 ,2%	5,0 ,3%
podstawowe nieukończone	3,0 ,2%	19,0 1,0%	22,0 1,2%
podstawowe	107,0 5,9%	200,0 11,0%	307,0 16,9%
zasadnicze zawodowe	232,0 12,8%	223,0 12,3%	455,0 25,1%
średnie nieukończone	15,0 ,8%	18,0 1,0%	33,0 1,8%
średnie zawodowe	168,0 9,3%	231,0 12,7%	399,0 22,0%
średnie ogólnokształcące	57,0 3,1%	168,0 9,3%	225,0 12,4%
pomaturalne	16,0 ,9%	50,0 2,8%	66,0 3,6%
wyższe nieukończone (6 semestrów lub więcej)	11,0 ,6%	21,0 1,2%	32,0 1,8%
licencjat lub trzyletnie studia zawodowe	9,0 ,5%	18,0 1,0%	27,0 1,5%
wyższe	100,0 5,5%	141,0 7,8%	241,0 13,3%
Total	720,0 39,7%	1092,0 60,3%	1812,0 100,0%

Zwróćmy uwagę, że właściwą tabelę poprzedza każdorazowo tabela podsumowująca (*Summary*). Zawiera ona liczebności i odsetki przypadków ważnych i braków danych, a także informację które zmienne były ze sobą krzyżowane (w tym przypadku: wykształcenie i płeć).

Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Wykształcenie * Płeć	1812	99.7%	5	0.3%	1817	100.0%

Należy również zauważyć, że obliczenia dla tabel krzyżowych prowadzone są bez uwzględniania braków danych.

W przypadku umieszczenia w polach Wiersze i Kolumny więcej niż jednej zmiennej, będą kolejno generowane serie tabel krzyżujące parami podane zmienne ze sobą.

Generowanie tabel krzyżowych odbywa się w Edytorze składni za pomocą następującej serii komend:

Składnia do wpisania w Edytorze	Opis działania składni
CROSSTABS	- stwórz tabelę krzyżową lub tabele krzyżowe
/TABLES= m3 BY m2	- skrzyżuje ze sobą zmienne m3 i m2, przy czym m3 będzie prezentowana w wierszach, a m2 w kolumnach. Przed i po łączniku BY można wstawiać więcej niż jedną zmienną
/FORMAT=AVALUE TABLES PIVOT	- drukuj tabelę, porządkuj prezentowane w tabeli dane rosnąco wedle wartości zmiennych (jeśli malejąco, wówczas użyj DVALUE)
/CELLS=COUNT ROW.	- w komórkach tabeli sumuj w wierszach. Można użyć sumowania w kolumnach (COLUMN) lub sumowania dla całości (TOTAL)

9.4. Podstawy interpretacji danych tabelarycznych

Tabel pochodzących z programu PSPP nie prezentujemy bezpośrednio w raporcie z badań. Informacje wytwarzane przez program PSPP w Widoku raportów mają charakter redundantny - przeznaczone są do wstępnej, szerokiej oceny wyników, a w raportach z badań są niepotrzebne. Ponadto każdą z tabel należy opatrzyć odpowiednim opisem i interpretacją. Poniższe wskazówki mają charakter uniwersalny, można je odnosić do raportów z badań, jak też do danych ilościowych zawartych w pracach promocyjnych semestralnych lub rocznych, licencjackich, magisterskich i doktorskich.

Przeanalizujmy transformacje, jakim powinny zostać poddane dane z Widoku raportów, jeśli mają zostać włączone do raportu z badań.

Konstrukcja tabeli. Wygenerowana została prosta tabela prezentująca postawy respondentów wobec przestrzegania norm prawa:

```
FREQUENCIES

FREQUENCIES
    /VARIABLES= x84
    /FORMAT=AVALUE TABLE
    /STATISTICS=NONE.
```

Czy "Przepisy są po to, żeby je obchodzić"?

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
zdecydowanie się zgadzam	1	104	5,72	5,73	5,73
raczej się zgadzam	2	261	14,36	14,37	20,10
ani tak, ani nie	3	384	21,13	21,15	41,24
raczej się nie zgadzam	4	498	27,41	27,42	68,67
zdecydowanie się nie zgadzam	5	509	28,01	28,03	96,70
trudno powiedzieć	7	60	3,30	3,30	100,00
nieuzasadniony brak odpowiedzi	-1	1	,06	Missing	
	Total	1817	100,0	100,0	

W raporcie z badań z powyższej tabeli syntetyzujemy następujące informacje w formie jak niżej.

Tabela 22. Postawy badanych wobec prawa (N=1816)

x84. Czy „Przepisy są po to, żeby je obchodzić"?	Wskazania respondentów	
	N	%
Zdecydowanie się zgadzam	104	5,7
Raczej się zgadzam	261	14,4
Ani tak, ani nie	384	21,2
Raczej się nie zgadzam	498	27,4
Zdecydowanie się nie zgadzam	509	28,0
Nie wiem, trudno powiedzieć	60	3,3
Razem	1816	100,0

Źródło: *Polskie Generalne Studium Wyborcze 2007, pod kierownictwem Radosława Markowskiego, afiliowane przy Instytucie Studiów Politycznych PAN, dofinansowane przez tę instytucję, oraz przez: Ministerstwo Nauki i Szkolnictwa Wyższego,Wissenschaftszentrum Berlin fur Sozialforschung (WZB), Polską Konfederację Pracodawców Prywatnych Lewiatan, Fundację Batorego, Instytut Filozofii i Socjologii PAN oraz instytucję badawczą realizującą sondaż - PBS DGA.*

W powyższej tabeli nie zostały umieszczone dane z kolumn: *Value, Percent i Cumulative Percent.* Zwyczajowo nie umieszcza się tych informacji w tabeli. Braki danych nie zostały uwzględnione, jednak należy to pozostawić indywidualnej decyzji badacza. Wartości procentowe zostały zaokrąglone do jednego miejsca po przecinku.

Opis i interpretacja tabeli. Tabele zazwyczaj umieszcza się w odrębnych podrozdziałach raportu z badań. Tu warto zaznaczyć istotność struktury raportu z badań. Musi być ona przemyślana. Najprostszym rozwiązaniem jest ustrukturyzowanie raportu według kolejności pytań w kwestionariuszu. Należy nadmienić, że pytania o zmienne socjodemograficzne (tzw. pytania metryczkowe, takie jak płeć, wiek lub wielkość miejsca zamieszkania respondenta) powinny znaleźć się w końcowej części raportu. Z kolei pytania selekcyjne (screeningowe) - na podstawie których dokonywano podziału respondentów na tych, którzy wezmą lub nie wezmą udziału w badaniu - na początku.

Pojedyncza tabela umieszczona w odrębnym podrozdziale wymaga - jak już wskazano - odpowiedniego opisu. Zalecane jest, by w opisie tym formułować myśli stosując czas teraźniejszy i stronę czynną. Nad tabelą, a pod tytułem podrozdziału umieszczamy opis i okoliczności zadawania danego pytania. Wskazujemy, do jakiego bloku czy baterii pytań ono należało, przytaczamy je w całości, a także w wielu przypadkach prezentujemy zbiór odpowiedzi na nie (tak zwaną kafeterię). Jeśli na pytanie nie odpowiadali wszyscy respondenci, lecz jedynie jakaś ich grupa wyselekcjonowana na podstawie odpowiedzi na pytania poprzednie należy wówczas dołączyć informację komu i z jakich powodów było to pytanie zadawane. Z kolei pod tabelą umieszczamy jej interpretację. W wąskim rozumieniu interpretacja ta stanowi opisowy, syntetyczny równoważnik tabeli. W rozumieniu szerszym w interpretacji można zestawiać dane z tabeli z innymi dostępnymi danymi, faktami lub postawionymi wcześniej hipotezami roboczymi i rozważać, w jakim stopniu dane z badania potwierdzają je lub im zaprzeczają. W tekście powinniśmy podawać dane z tabeli, ale narracyjnie zagregowane. Sugeruje się posługiwanie się wartościami opisowymi, na przykład: co trzeci respondent, prawie połowa, czwarta część, większość, i tak dalej. Dopuszczalne jest podawanie

wartości całkowitych, bez liczb po przecinku. Należy się posługiwać skrótem „proc.", a nie znakiem %. Kluczowe wnioski płynące z interpretacji tabeli można zaznaczyć pismem grubym (*bold*). Te wyróżnione elementy są następnie zbierane i umieszczane na początku raportu z badań jako wnioski.

Warto wskazać pewien powszechnie popełniany błąd nazewnictwa wynikający ze skrótu myślowego. Należy bezwzględnie odróżniać pojęcia 'procent' i 'punkt procentowy'. '**Procent'** używa się do określenia operacji wykonywanych na wartościach liczonych od pewnej wartości bazowej. Ma charakter bezwzględny. Z kolei pojęcia **'punkt procentowy'** używa się w odniesieniu do operacji na wartościach wyrażonych w procentach; ma charakter względny. Te dwa rodzaje wartości są często ze sobą mylone. Ten powszechny błąd często powielają media posługując się na określenie 'punktu procentowego' skróconą, nieprawidłową w używanym przez nie kontekście nazwą 'procent'. Rozważmy przykład następującego komunikatu, z którym możemy zetknąć się w środkach masowego przekazu:

Poparcie dla partii politycznej wynosiło 15 procent, po czym nastąpił wzrost poparcia o 5 procent.

Stwierdzenie to może być interpretowane dwojako:

1/ Błędnie: poparcie wynosiło 15 proc., nastąpił wzrost o 5 proc., a zatem obecne poparcie wynosi 20 proc. Podana operacja będzie poprawna tylko wówczas, jeśli w powyższym zdaniu zamiast słowa 'procent' użyjemy pojęcia 'punkt procentowy'.

2/ Poprawnie: poparcie wynosiło 15 procent, nastąpił wzrost o 5 procent wyciągnięty z 15 proc. - ((15/100)*5) - a zatem o 0,75 proc. Poparcie dla partii wynosi więc: 15 proc.+ 0,75 proc. = 15,75 proc.

10

Rozdział 10. Miary tendencji centralnej (pozycyjne)

Bardziej zaawansowanym zabiegiem analitycznym w stosunku do opisu tabelarycznego zmiennych jest wyznaczanie **miar rozkładu**. Do miar tych należą miary tendencji centralnej (rozmaite typy średnich, mediana i dominanta), a także inne miary położenia jak N-tyle, miary asymetrii (współczynnik skośności, współczynnik asymetrii lub trzeci moment centralny), miary koncentracji (jak na przykład współczynnik Corrado Giniego i kurtoza), a także statystyki będące przedmiotem kolejnego rozdziału - rozmaite miary zróżnicowania zmiennych (dyspersji, rozrzutu). Miary tendencji centralnej lub inaczej zwane - pozycyjne - zaliczamy do tak zwanej **statystyki opisowej**. Przeznaczone są one do wyczerpującego statystycznego opisu jednej zmiennej.

10.1. Średnie

Wartości średnie stanowią zróżnicowaną rodzinę miar tendencji centralnej. Zaliczamy do nich między innymi: średnią arytmetyczną, średnią ważoną, średnią odciętą, średnią winsorowską, średnią kwadratową, geometryczną i harmoniczną. Oblicza się także przedział ufności dla średnich. Najczęściej używane w analizie statystycznej są: średnia arytmetyczna, średnia ważona oraz średnia odcięta, a także przedział ufności dla średniej. Pozostałe zaprezentowane w tym podrozdziale miary mają znaczenie marginalne, stosowane są w wyjątkowych przypadkach. Tylko część z opisywanych miar może być bezpośrednio obliczana w programie PSPP. W tych przypadkach, w których jest to niemożliwe, zostały podane sposoby ich obliczenia częściowo zautomatyzowanego za pomocą dostępnych w PSPP funkcji.

10.1.1. Średnia arytmetyczna

Średnią arytmetyczną obliczamy według następującego wzoru:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

gdzie

$x_1,\ldots,x_n$ – wartości zmiennej

n – liczba jednostek analizy dla tej zmiennej

Jest ona miarą przeciętności, wskazuje typowy poziom natężenia mierzonej cechy, zaciera różnice jednostkowe. Średnia arytmetyczna jest statystyką najistotniejszą, bowiem jej wartość jest liczona na podstawie wszystkich pomiarów w badanej grupie. Jest ona najczęściej używaną miarą tendencji centralnej. Liczne zaawansowane statystyki opierają się na wykorzystaniu rozmaitych średnich. Istotne jest w takim przypadku poznanie właściwości tej miary.

Średnią arytmetyczną możemy stosować tylko do zmiennych mierzonych na poziomach ilościowych: interwałowym i ilorazowym. W przypadku poziomu porządkowego jest to dopuszczalne w nielicznych przypadkach opisanych w rozdziałach poprzednich (skala R. Likerta). Dla większości zmiennych porządkowych oraz nominalnych obliczanie średniej arytmetycznej traci sens (wykorzystuje się wówczas dominantę i medianę). Należy zdawać sobie sprawę z faktu, że średnia arytmetyczna posiada szereg wad i ograniczeń.

Po pierwsze, średnia arytmetyczna jest miarą wrażliwą na wartości skrajne. W sytuacji, gdy liczebności nie skupiają się wokół wartości środkowych danego zakresu zmiennej, lecz skrajnych lub dodatkowo koncentrują się asymetrycznie przy wartościach największych lub najmniejszych, miara ta w dużym stopniu traci swoją wartość poznawczą. Wówczas zastępujemy średnią inną miarą tendencji centralnej – medianą. Należy mieć na uwadze, że im bardziej jednorodny (skupiony wokół wartość środkowej) rozkład danej zmiennej, tym bardziej adekwatnie charakteryzuje ona poziom badanej cechy.

Po drugie, w przypadku obliczania średniej dla zmiennych przedziałowych, gdy jeden z przedziałów klasowych jest otwarty (górny, dolny lub oba) wówczas istnieją przesłanki uniemożliwiające obliczanie tej statystyki[1]. W praktyce analitycznej przyjęło się, że jeśli liczebności przedziałów otwartych są nieliczne (tu badacze podają dwie wartości – rygorystyczną – do 1,5 proc. wszystkich jednostek analizy i liberalną – do 5 proc.) można przedziały takie zamknąć i arbitralnie ustalić środek przedziału. W przeciwnym przypadku nie możemy obliczać średniej arytmetycznej i należy stosować medianę.

Po trzecie, wynikiem średniej jest abstrakcyjny parametr w przypadku, gdy średnią obliczamy z użyciem liczb naturalnych. Jej wartość może być liczbą niewystępującą w zbiorze danych lub z punktu widzenia logiki niemożliwą (na przykład średnia liczba osób deklarujących działalność w organizacjach politycznych może być wartością zawierającą liczby ułamkowe).

[1] Przedział klasowy otwarty to taki, w którym jego dolna lub górna granica nie została określona liczbowo. Z kolei przedział klasowy zamknięty to taki, w którym zarówno jego dolna lub górna granica została liczbowo zdefiniowana. Przykład przedziału otwartego to item o treści „500 PLN i więcej".

Po czwarte, średnia obliczana ze zmiennych, w których rozkład wartości jest wielomodalny, jest bezwartościowa.

Ponadto średnia arytmetyczna posiada szereg właściwości, które nie są bezpośrednio istotne i stosowalne dla analiz danych ilościowych. Są to cechy typowo matematyczne, na przykład suma odchyleń poszczególnych wartości od średniej arytmetycznej jest równa zero lub suma kwadratów odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej jest najmniejsza w porównaniu z sumami kwadratów odchyleń od jakiejkolwiek innej liczby w szeregu.

Często w analizach danych, w konstruowaniu różnorodnych wskaźników, średnia arytmetyczna jest niewystarczającą miarą i dlatego wykorzystuje się inne miary tendencji centralnej.

Średnią arytmetyczną w programie PSPP możemy obliczyć na dwa różne sposoby - za pomocą *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics* lub poprzez *Analyze* ⇨ *Descriptive Statistics* ⇨ *Descriptives* ⇨ *Statistics*. W obu wymienionych przypadkach należy oznaczyć średnią (*Mean*). Pomocną opcję wbudowaną w program stanowi też porównywanie średnich (*Means*) znajdujące się w zakładce *Analyze* ⇨ *Compare Means*:

Funkcja ta umożliwia porównywanie średnich w jednej tabeli. W pole *Dependent List* wstawiamy zmienne, których średnie wartości mają być obliczone. Z kolei w *Independent List* - te zmienne, według których będzie dokonywane porównanie (podział). W powyższym przykładzie zaprezentowane zostaną średnie poparcia dla demokracji (c21) dla każdego z szesnastu województw.

10.1.2. Średnia ważona

Średnia ważona polega na tym, że poszczególnym jednostkom analizy przypisujemy wagi zwiększające, zmniejszające lub pozostawiające w nienaruszonym stanie ich udział w średniej. Można posługiwać się tu zarówno liczbami całkowitymi jak też ułamkowymi. Obliczanie średniej ważonej odbywa się według następującego wzoru:

$$x_w = \frac{(\text{wartość}_1 * \text{waga}_1) + (\text{wartość}_2 * \text{waga}_2) + (\text{wartość}_n * \text{waga}_n)}{\text{waga}_1 + \text{waga}_2 + \text{waga}_n}$$

Jeśli wagi są równe dla wszystkich jednostek analizy w zbiorze danych - średnia ważona równa jest średniej arytmetycznej.

Dla przybliżenia sposobu zastosowania średniej ważonej prześledźmy poniższy przykład. Przypuśćmy, że konstruujemy syntetyczny, porównawczy wskaźnik oceny wizerunku polityka. Pod uwagę bierzemy trzy wymiary: przygotowanie merytoryczne, kompetencje komunikacyjne oraz ocenę wyglądu zewnętrznego - prezencję polityka. Stwierdzamy przy tym, że najistotniejszą grupą cech są kompetencje komunikacyjne i postanawiamy nadać im wagę dwukrotnie większą niż pozostałym wymiarom wizerunku. Pozostałym wymiarom przypisujemy wagę 1. Otrzymane wyniki ocen to: dla wymiaru merytorycznego polityka - 4, dla wymiaru kompetencji komunikacyjnych polityka - 2, dla wymiaru prezencji polityka - 5.

Średnia arytmetyczna w tym przypadku wyniesie 3,6(6), a średnia ważona 3,25. Średnia arytmetyczna została obliczona następująco:

$$\bar{x} = \frac{4 + 2 + 5}{3} = 3{,}6(6)$$

Średnią ważoną obliczono podstawiając do wzoru:

$$x_w = \frac{(4 * 1) + (2 * 2) + (5 * 1)}{1 + 2 + 1} = \frac{13}{4} = 3{,}25$$

Obliczanie średniej ważonej w programie PSPP odbywa się między innymi po nałożeniu i włączeniu wag na jednostki analizy w zbiorze danych.

10.1.3. Średnia odcięta (obcięta, ucinana, trymowana)

Średnia ucinana, średnia obcięta lub średnia trymowana jest obok innych średnich, dominanty i mediany jedną z miar statystycznych tendencji centralnej. Średnia odcięta liczona jest identycznie jak średnia arytmetyczna, a różnica polega na liczeniu tej miary z niepełnego zakresu danych - wartości danej zmiennej zostają uporządkowane od najniższej do najwyższej, a następnie odcinane są wartości skrajne w równej proporcji z dołu i z góry zakresu wartości zbioru danych. Zazwyczaj odcina się po 5 proc. obserwacji najniższych i 5 proc. najwyższych. Dopuszcza się i stosuje także inne zakresy odcinania - często spotykane jest odcinanie po 10 proc. z każdej strony. Stosuje się również miary oparte na n-tylach, polegające na przykład na odrzuceniu pierwszego i ostatniego kwartyla. Przesłanką stosowania średniej odciętej jest swoiste „ustabilizowanie" średniej, poprzez pozbawienie jej wartości skrajnych, wpływających na jej podwyższenie lub obniżenie.

W programie PSPP zakres odcinanych wartości do wyliczania średniej odciętej ustawiony jest na stałe - program odrzuca symetrycznie górne i dolne 5 proc. wszystkich wartości. Jeśli chcemy obliczyć średnią odciętą można uczynić to wykorzystując na przykład rangowanie, rekodowanie i filtrowanie jednocześnie. Operację tę można także wykonać ręcznie po posortowaniu przypadków, zamieniając skrajne wartości na braki danych.

Średnią odciętą dla 5 proc. jednostek analizy z każdej strony można uzyskać poprzez wybranie *Analyze* ⇨ *Descritpive Statistics* ⇨ *Explore*, następnie wpisując do *Dependent list* zmienną, dla której wartości chcemy uzyskać, a po kliknięciu przycisku wybrając *Statistics* i tam zaznaczając *Descriptives*. Otrzymujemy w ten sposób tabelę zawierającą (pośród innych informacji) informację o wartości średniej odciętej.

10.1.4. Średnia winsorowska

Średnia winsorowska stanowi odmianę średniej odciętej, jej nazwa pochodzi od nazwiska Charlesa Winsora (1895-1951). Obliczana jest ona jak zwykła średnia arytmetyczna, jednak wybraną i równą liczbę najmniejszych i największych uporządkowanych w szeregu wartości zastępuje się przy jej obliczaniu wartościami bezpośrednio sąsiadującymi.

Przypuśćmy, że chcemy dokonać obliczenia średniej winsorowskiej ze zbioru danych składającego się z dziesięciu liczb. Decydujemy, że procesowi temu (określamy go mianem winsoryzacji) zostaną poddane dwie wartości z każdej strony przedziału. Oznacza to, że zmienna x_1 oraz x_2 zostaje zastąpiona przez wartość x_3, a x_9 i x_{10} przez wartość pochodzącą z x_8. Zbiór danych przed i po winsoryzacji przedstawia tabela 23. Po dokonaniu czynności zastąpienia wybranej liczby skrajnych wartości wartościami bezpośrednio sąsiadującymi dokonujemy obliczenia średniej arytmetycznej.

Tabela 23. Schemat winsoryzacji

Zbiór danych przed winsoryzacją									
x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	2	5	6	6	7	8	10	11	15
Zbiór danych po winsoryzacji									
x3	x3	x3	x4	x5	x6	x7	x8	x8	x8
5	5	5	6	6	7	8	10	10	10

Program PSPP nie posiada automatycznej możliwości obliczania średniej winsorowskiej, jednak łatwo to uczynić za pomocą posiadanych narzędzi służących do przekształceń zmiennych.

10.1.5. Średnia krocząca (średnia ruchoma)

Średnią kroczącą wylicza się dla pewnego wybranego okresu, a więc jakiejś liczby dokonanych pomiarów. Wyniki pojedynczych pomiarów mogą się wydawać chaotyczne - odchylając się w różne strony i uniemożliwiając odnalezienie jakiegokolwiek porządku. Średnia krocząca stanowi miarę pozwalającą na odnalezienie prawidłowości i trendów w serii pomiarów w czasie. Istnieją rozmaite odmiany średniej kroczącej.

Prosta średnia krocząca (*simple moving average,* SMA) to zwykła średnia arytmetyczna obliczana dla wybranej liczby następujących po sobie pomiarów. Jest ona najczęściej używana spośród średnich kroczących. Stosowana jest także **ważona średnia krocząca** (*weighted moving average*, WMA). Przypisujemy w niej wagi kolejnym pomiarom, które maleją w postępie arytmetycznym - pomiarowi najnowszemu przypisuje się wagę najwyższą (n), kolejnemu (n-1), a najstarszemu - najniższą (n-m, gdzie *m* oznacza liczbę wszystkich pomiarów). Z kolei **wykładnicza średnia krocząca** (*expotential moving*

average, EMA) w jeszcze większym stopniu (wykładniczo) umacnia pomiary najnowsze i osłabia najstarsze. Rzadziej stosowane są średnie kroczące szeregów czasowych oparte na regresji liniowej, **średnie kroczące trójkątne**, w których największe wagi otrzymują wartości środkowe, czy **zmienne średnie kroczące** reagujące wygładzaniem wykresu w okresach silnej zmienności czynnika pomiaru. Ponadto rozróżniać można średnie kroczące ze względu na okres czasu, w jakim je mierzymy - mówimy wówczas o średnich kroczących krótkoterminowych i długoterminowych. Średnie w zróżnicowany sposób reagują na zmiany wartości zmiennych - wrażliwsze są wykładnicza i ważona średnia krocząca, mniej wrażliwa jest z kolei prosta średnia krocząca.

Obliczanie średniej kroczącej może być prowadzone wedle następującego schematu. Załóżmy, że do obliczenia wyników będziemy brać pod uwagę trzy kolejne pomiary. Poniższy blokowy schemat ilustruje ideę średniej kroczącej dla trzech, czterech i pięciu dokonanych pomiarów:

Pomiar 1	Pomiar 2	Pomiar 3
Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej

Pomiar 1	Pomiar 2	Pomiar 3	Pomiar 4
Wyniki nie wchodzą w skład średniej	Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej

Pomiar 1	Pomiar 2	Pomiar 3	Pomiar 4	Pomiar 5
Wyniki nie wchodzą w skład średniej	Wyniki nie wchodzą w skład średniej	Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej	Wyniki wchodzą w skład średniej

Zaletą średnich kroczących jest możliwość przewidywania kierunków trendów. Natomiast problematyczne jest ich wykorzystanie w przypadku tak zwanych trendów bocznych (horyzontalnych), a więc niewystępujących, gdzie nie istnieje tendencja wzrostowa ani spadkowa. Średnia krocząca znajduje zastosowanie przede wszystkim w analizach ekonomicznych, głównie w analizie rynku walut, surowców i papierów wartościowych, czyli tak zwanej analizie technicznej. Zastosowanie odnajduje ona również na polu nauk społecznych w rozmaitych rankingach, gdzie wskaźnik powinien być stabilny, oparty nie tylko na pomiarze bieżącym, lecz także wybranych poprzednich[2].

Program PSPP nie generuje automatycznie tego typu średniej, jednak łatwo można ją uzyskać, wykorzystując funkcję obliczania wartości zmiennych (*compute*).

[2] Średnia krocząca została zastosowana na przykład w rankingu szkół wyższych przeprowadzanym na zlecenie Polskiej Konfederacji Pracodawców Prywatnych „Lewiatan" przez Pentor Research International.

10.1.6. Średnie potęgowe (kwadratowa i harmoniczna)

Średnia potęgowa jest to pierwiastek m-tego stopnia z sumy poszczególnych wartości zmiennych (n) podniesionych do potęgi o wykładniku m podzielonej przez liczbę wartości zmiennych (N):

$$x_p = \sqrt[m]{\frac{n_1^m + n_2^m + \cdots + n_n^m}{N}}$$

Średnia potęgowa obejmuje średnią kwadratową i średnią harmoniczną. **Średnią kwadratową** uzyskujemy, gdy m = 2, a **średnią harmoniczną** - gdy m = -1. Średnią kwadratową obliczamy wedle wyżej przytoczonego wzoru podstawiając jako wykładnik potęgi liczbę 2, natomiast średnią harmoniczną obliczamy wedle następującego wzoru:

$$x_h = \frac{N}{\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_m}}$$

Średnia harmoniczna jest więc odwrotnością średniej arytmetycznej z odwrotności wartości zmiennej. Jest to miara przydatna w przypadku konieczności przeliczania wartości cech na stałą jednostkę innej zmiennej, czyli w postaci wskaźników natężenia (na przykład w kilometrach na godzinę, sztukach na minutę, osobach na kilometr kwadratowy, kilogramach na osobę, itd.).

Program PSPP nie umożliwia bezpośredniego obliczenia tych statystyk.

10.1.7. Średnia geometryczna

Średnia geometryczna to iloczyn liczb, dla których chcemy obliczyć średnią następnie poddany pierwiastkowaniu o stopniu równym ilości liczb. Na przykład obliczenie średniej geometrycznej dwóch liczb polega na pomnożeniu dwóch liczb przez siebie, a następnie wyciągnięciu z iloczynu pierwiastka kwadratowego. Z kolei przy trzech liczbach wyciągamy pierwiastek sześcienny z wyniku mnożenia trzech liczb, itd. Średnia geometryczna jest miarą przydatną przede wszystkim w demografii - gdy poszczególne pomiary mają charakter miar względnych oraz w sytuacjach, gdy rozkład cechy badanej zmiennej jest asymetryczny oraz w sytuacji niemożności odcięcia wartości skrajnych. Średnia geometryczna wymaga ilościowego poziomu pomiaru.

W programie PSPP nie zaimplementowano bezpośredniego obliczania tej statystyki.

10.1.8. Przedział ufności dla średniej

Przedział ufności dla średniej mówi nam, jakie wartości średniej w populacji możemy przewidywać na podstawie średniej z próby. Dokonujemy zatem estymacji uzyskanego wyniku na populację generalną. Wartość tej statystyki zawiera się w przedziałach. Pierwszy z przedziałów podaje, jaką najniższą wartość może przybrać średnia w populacji, a drugi - jaką maksymalną wartość może mieć średnia. Statystyka ta jest rozmaicie obliczana dla małych prób (definiowane jako mniejsze niż 30 jednostek analizy) i dużych prób (powyżej 30 jednostek analizy). Odmienne sposoby obliczania stosujemy również przy znanych

i nieznanych parametrach populacji. W przypadku dużych prób i nieznanego parametru populacji dolny przedział ufności dla średniej obliczamy na podstawie wzoru:

$$\text{średnia} - \frac{1{,}96 * \text{odchylenie standardowe}}{\text{pierwiastek z liczebności próby}}$$

a górny przedział:

$$\text{średnia} + \frac{1{,}96 * \text{odchylenie standardowe}}{\text{pierwiastek z liczebności próby}}$$

Stała 1,96 jest to wartość ufności najczęściej przyjmowana w naukach społecznych - wynosi ona 95 procent. Jeśli chcielibyśmy przyjąć bardziej rygorystyczne założenia - a więc wartość ufności rzędu 99 procent, wówczas zamiast 1,96 wstawiamy w podanym wzorze 2,58.

Należy spodziewać się, że z czasem do programu PSPP zostaną wprowadzone bardziej wyrafinowane techniki „stabilizowania" średnich, swoistego uodparniania ich na wartości skrajne w postaci należących do tak zwanych statystyk odpornościowych (odpornych metod statystycznych) M-estymatorów (między innymi Tukey'a i Hubera) lub rzadziej stosowanych R-estymatorów i L-estymatorów.

Przedział ufności dla średniej uzyskujemy w Programie PSPP wybierając z menu tekstowego *Analyze* ⇨ *Descritpive Statistics* ⇨ *Explore*, wpisując do *Dependent list* zmienną, dla której tę statystykę chcemy uzyskać, a następnie po kliknięciu przycisku *Statistics*, zaznaczamy *Descriptives*. Program PSPP posiada możliwość automatycznego obliczania poziomu ufności tylko dla α=0,05 (czyli 5 proc.). Wynik obliczeń przedziału ufności dla średniej z programu PSPP interpretujemy następująco: „Wnioskując na podstawie zbadanej próby można orzec z 95 procentowym prawdopodobieństwem (lub inaczej: narażając się na ryzyko pomyłki rzędu 5 procent), że w populacji generalnej średnia będzie zawierała się między wartością x a wartością y".

10.2. Dominanta (moda)

Dominanta (moda, modalna, wartość najczęstsza, wartość typowa, wartość dominująca) wskazuje tę wartość zmiennej w rozkładzie, która jest wartością o najwyższym prawdopodobieństwie wystąpienia lub inaczej - argument, dla którego funkcja gęstości prawdopodobieństwa przyjmuje wartość najwyższą. Najprościej mówiąc, jest to ta wartość zmiennej, która występuje najliczniej. Dominantę zaliczamy do pozycyjnych miar położenia, podobnie jak n-tyle i oznaczamy literami Mo lub D. Miara ta ma zastosowanie wówczas, gdy nie można użyć średniej arytmetycznej lub jej odmian, a więc w sytuacji, gdy zmienne mierzone są na poziomie nominalnym. Jest to jedyna statystyka z rodziny miar przeciętnych, którą można wyznaczyć dla cech mierzonych na najniższym poziomie mocy skali.

Dominantę można obliczać także dla zmiennych mierzonych na poziomach porządkowym, interwałowym i ilorazowym. Często stosowanym wówczas zabiegiem jest grupowanie wartości zmiennej w przedziały - na przykład zmienną rok urodzenia zamyka się w przedziały wiekowe (np. od 25 do 35, powyżej 35 do 45, powyżej 45 do 55). W takiej sytuacji wyznaczyć możemy dwie wartości: po pierwsze wskazać przedział klasowy dominanty, po drugie wartość dominanty. **Przedział dominanty** wskazujemy analogicznie do wartości dominanty - jako ten, w którym wartości są najliczniejsze. Z kolei wartość dominanty dla szeregów rozdzielczych wyznacza się wedle wzoru:

$$M_0 = x_{0m} + \left[\frac{h_m * (n_m - n_{m-1})}{n_m - n_{m-1} + n_m - n_{m+1}}\right]$$

gdzie: m oznacza numer przedziału, w którym występuje modalna, x_{0m} - to dolna granica przedziału, w którym występuje modalna, n_m - liczebność przedziału modalnej, n_{m-1}, n_{m+1} - liczebność klasy poprzedzającej i następującej po przedziale, w którym wyznaczono dominantę, h_m - rozpiętość przedziału klasowego, w którym wyznaczono dominantę.

Obliczmy wartość dominanty dla przedziałów wiekowych, przyjmując, że najliczniejsza jest kategoria środkowa - powyżej 35 do 45 lat:

x_{0m} - 35 lat

n_m - przyjmujemy, że jest to 48

n_{m-1}, n_{m+1} - przyjmujemy, że jest to odpowiednio 10 i 8

h_m - 45 - 35 = 10

A zatem:

$$35 + \left[\frac{10 * (48 - 10)}{48 - 10 + 48 - 8}\right] = 39{,}87$$

Oznacza to, że wartość dominanty wynosi w przybliżeniu 40 lat.

W niektórych przypadkach dominanta w zbiorze danych może być niemożliwa do obliczenia. Dzieje się to wówczas, gdy żadna z wartości zmiennych nie posiada wyraźnie większej liczebności (a więc rozkład jest niezróżnicowany) lub też wartość taka istnieje, jednak jest to wartość skrajna w przypadku zmiennych zamkniętych w przedziały. Zmienna może posiadać więcej niż jedną dominantę - rozkład o dwóch dominantach nazywamy **bimodalnym**, a gdy posiada więcej niż dwie - **wielomodalnym.**

Wadą dominanty jest jej duża wrażliwość na losowość występowania rozkładu zmiennych - przy niewielkich różnicach w rozkładach wartości zmiennych dominanta w powtarzanych pomiarach może uzyskiwać i uzyskuje zróżnicowane wartości. Prowadzi to do jej ograniczonej możliwości stosowania w badaniach porównawczych. Ponadto dominantę sensownie można interpretować tylko wówczas, jeśli jednostki analizy w zbiorze danych są wystarczająco liczne. W przypadku, gdy mamy do czynienia ze zmienną o charakterze przedziałowym, lecz przedziały klasowe nie są równe, nie można wówczas obliczać dominanty.

Zaletą dominanty jest niewrażliwość na skrajne wartości występujące w danej zmiennej (w przeciwieństwie do średniej arytmetycznej) oraz łatwość jej interpretacji. Interpretacja dominanty jest prosta - pozwala ona na stwierdzenie tego co typowe, najczęstsze, najliczniejsze w danej zbiorowości.

Dominantę w programie PSPP możemy obliczyć na dwa różne sposoby - w menu *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics* oraz w menu *Analyze* ⇨ *Descriptive Statistics* ⇨ *Descriptives* ⇨ *Statistics*. W obu wymienionych przypadkach należy oznaczyć dominantę (*Mode*).

10.3. Mediana (wartość środkowa)

Medianą, czyli wartością środkową nazywamy tę wartość uporządkowanego rozkładu danej zmiennej, która dzieli rozkład na pół. Połowa wartości zmiennych przyjmuje wartości mniejsze (bądź równe medianie), a połowa - większe (bądź równe). Mediana może być mierzona dla zmiennych porządkowych, interwałowych i ilorazowych. Nazywa się ją inaczej kwartylem drugim, bowiem dzieli zbiór danych pod względem określonej zmiennej na dwie części. Oznaczamy ją Me lub Q_2 (od kwantyla (*quantile*) drugiego rzędu). Mediana obok średniej arytmetycznej stanowi najczęściej stosowaną miarę statystyczną.

Medianę obliczamy w zależności od tego, czy liczba jednostek analizy w zbiorze jest nieparzysta czy parzysta. W pierwszym przypadku, tj. dla nieparzystej liczby jednostek analizy, wzór na wartość środkową jest następujący:

$$\frac{n+1}{2}$$

gdzie n - oznacza liczbę obserwacji. Z kolei dla parzystej liczby jednostek analizy, wartością środkową jest średnia arytmetyczna pomiędzy dwoma środkowymi wartościami:

$$\frac{1}{2} * \left(\frac{n}{2} + \frac{n+2}{2}\right)$$

Medianę dla zbioru o nieparzystej liczbie wartości 1, 2, 3, 4, 5 obliczamy:

$$\frac{5+1}{2} = 3$$

W przypadku, gdy lista wartości zmiennych jest parzysta i obejmuje wartości 1, 2, 3, 4, 5, 6, mediana wyniesie:

$$\frac{1}{2} * \left(\frac{6}{2} + \frac{6+2}{2}\right) = \frac{1}{2} * \left(\frac{6}{2} + \frac{8}{2}\right) = \frac{1}{2} * \frac{14}{2} = 3{,}5$$

Mediana jest miarą statystyczną odporną na wartości skrajne - jest to zarówno jej zaleta, jak też wada. Z jednej strony otrzymujemy miarę „stabilną", z drugiej nawet obecność wartości skrajnych nie ma wpływu na jej wartość. Stosujemy ją wszędzie tam, gdzie nie jest możliwe wykorzystanie średniej arytmetycznej. Miarę tę należy stosować zamiast średniej arytmetycznej w przypadku dużej asymetrii rozkładu, statystyka ta wyraża wówczas znacznie lepiej tendencję centralną. Przeciwwskazaniem jest stosowanie tej miary w przypadku rozkładu wielomodalnego.

Warto zapamiętać, że pomiędzy dominantą, medianą i średnią arytmetyczną istnieje następujący związek:

$$Mo = 3Me - 2\bar{x}$$

W programie PSPP medianę obliczamy wybierając *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics* zaznaczając *Median*.

10.4. N-tyle (kwantyle)

N-tyle (kwantyle) stanowią tak zwane miary pozycyjne wyższych rzędów i charakteryzują rozkład zmiennej. Kwantyle dzielą uporządkowaną zbiorowość na zadaną liczbę równych sobie części. Kwantyle lub n-tyle są nazwą zbiorczą - każdy sposób podziału posiada swoje nazwy własne. Najczęściej używaną miarą spośród kwantyli są kwartyle i percentyle (centyle).

Kwartyle dzielą zbiorowość na cztery części, mówimy zatem, że jest to kwantyl rzędu 1/4. Kwartyl pierwszy (dolny) oznaczany Q_1 obejmuje 25 proc. zbioru danych o najniższych wartościach danej zmiennej. Z kolei kwartyl trzeci (górny) obejmuje ćwierć jednostek analizy, które posiadają najwyższe natężenie cechy, ze względu na którą następował podział. Jest to szczególnie przydatna miara przy poszukiwaniu różnic w zbiorowościach, jeśli w analizach autorzy ograniczają się do prostych statystyk opisowych[3].

Z kolei percentyle (centyle) to kwantyle rzędu 1/100. Dzielą one zbiorowość na sto części. Używane są na przykład w pediatrii w postaci siatek centylowych, umożliwiających obiektywną ewaluację cech fizycznych dziecka w toku jego rozwoju (uznaje się na przykład, że normą jest zawieranie się antropometrycznych cech dziecka pomiędzy 3 a 97 centylem).

Często wykorzystywanym n-tylem jest także opisana w poprzednim podrozdziale mediana. Należy wyraźnie wskazać, że jest ona kwantylem rzędu 1/2, bowiem dzieli zbiorowość na dwie części.

Istnieje szereg innych odmian n-tyli: n-tyle rzędu 1/3 to tertyle (T), rzędu 1/5 - kwintyle (QU), 1/6 - sekstyle (S), 1/20 - wigintyle (V), 1/1000 - promile (Pr).

W programie PSPP istnieje możliwość wygenerowania niektórych percentyli (5, 10, 25, 50, 75, 90, 95): *Analyze* ⇨ *Explore* ⇨ *Statistics* ⇨ *Percentiles*.

10.5. Skośność

Skośność jest miarą asymetryczności rozkładu wartości zmiennej. Jest to pomiar kształtu rozkładu danej zmiennej, należy do klasycznych miar rozkładu. Miarę tę najlepiej wyjaśnić posługując się wykresem rozkładu zmiennej, gdzie na osi rzędnych (oś Y) znajdują się liczebności (n) danej zmiennej, a na osi odciętych (oś X) - jej wartości (x). Rozkład symetryczny to taki, w którym wartość lub wartości środkowe danej zmiennej są najliczniej reprezentowane przez jednostki analizy, a skrajne wartości - najmniej licznie, przy czym od wartości środkowych do minimalnych oraz od wartości środkowych do maksymalnych obserwujemy identyczne tempo zmniejszania się liczebności jednostek analizy. Poprowadzenie linii prostopadłej do osi odciętych przez najwyższy punkt wartości liczebności (n) w rozkładzie symetrycznym dzieli krzywą rozkładu na dwie identyczne pod względem kształtu części. Innymi słowy trzy miary: średnia, mediana oraz dominanta są sobie równe.

W tym miejscu konieczne jest wprowadzenie ważnego pojęcia krzywej normalnej. Wprowadzenie ma charakter minimalistyczny, lecz jego znajomość jest niezbędna dla wykonania wielu istotnych testów

[3] Próbę zastosowania kwartyli dla badań porównawczych w ramach danej zbiorowości poczyniono w: D. Mider, A. Marcinkowska, *Przemoc w kulturze politycznej polskiego Internetu*, „Studia Politologiczne", t. 21, 2011, s. 239-296.

statystycznych. Nazwę *rozkład normalny* wprowadził brytyjski antropolog, genetyk, statystyk, pisarz i podróżnik Franciszek Galton (1822-1911) w 1889 roku. Swój wkład w badania nad fenomenem rozkładu normalnego miał także niemiecki matematyk, fizyk i astronom Karl Friedrich Gauss i dlatego równie często mówi się o tym zjawisku krzywa Gaussa. Rozkład normalny nazywany jest również krzywą dzwonową lub dzwonowatą (*bell curve*)[4]. Rozkład normalny obrazuje wartości, jakie przyjmuje zmienna losowa ciągła w sytuacji nieskończenie wielkiej liczby prób. Rozkład normalny posiada wiele ciekawych matematycznych własności - 68,3 proc. wyników skupia się w odległości jednego odchylenia standardowego od średniej arytmetycznej, 95,5 proc. wyników w odległości dwóch odchyleń standardowych od średniej, a 99,7 proc. - w odległości trzech odchyleń standardowych. Jest to właściwość określana w literaturze przedmiotu regułą trzech sigm (odchylenie standardowe oznaczamy grecką literą sigma - σ). Najistotniejszą z punktu widzenia wnioskowania statystycznego właściwością krzywej normalnej jest fakt, że wiele zachodzących na świecie zjawisk, w tym społecznych przyjmuje rozkład przybliżony do normalnego: wyniki pomiaru skupione są silnie i symetrycznie wokół wartości środkowej (średniej) i występują coraz mniej licznie im dalej od niej znajdują się, zachowując (w przybliżeniu) regułę trzech sigm. Rozkład przybliżony do rozkładu normalnego przyjmują takie cechy populacji ludzkich jak waga, wzrost, wykształcenie czy wysokość dochodów[5]. Nie każdy rozkład symetryczny jest rozkładem normalnym. Istnieją liczne sposoby przekonania się o tym, czy dany rozkład spełnia wystarczające warunki, by określić go tym mianem (m.in. wizualna ocena histogramu z nałożoną krzywą dzwonową, współczynniki skośności i kurtozy oraz test Kołmogorowa-Smirnowa).

Najprostszym sposobem oceny skośności rozkładu jest porównanie lokalizacji trzech miar tendencji centralnej: średniej arytmetycznej, mediany oraz dominanty, jednak stosuje się również dwie inne miary: wskaźnik skośności oraz współczynnik skośności.

Współczynnik skośności obliczany jest na podstawie rozmaitych wzorów. Najprostszym, klasycznym (lecz niezadowalającym wszystkich badaczy) sposobem obliczenia współczynnika skośności jest skorzystanie ze wzoru:

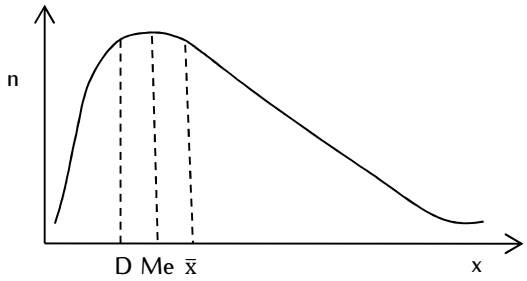
$$\text{współczynnik skośności} = \frac{\text{średnia} - \text{dominanta}}{\text{odchylenie standardowe}}$$

Program PSPP oblicza współczynnik skośności wedle innej niż powyższa, bardziej zaawansowanej i uodpornionej na zakłócenia formuły. Współczynnik skośności interpretujemy opierając się na dwóch elementach: wielkości wartości współczynnika oraz jego znaku. Jeśli rozkład zmiennej skupia się w obszarze niskich wartości, wówczas współczynnik skośności przyjmuje wartość dodatnią. Rozkład taki nazywamy **asymetrycznym prawostronnie**, prawoskośnym lub asymetrią dodatnią. W takiej sytuacji mediana (Me) jest większa niż średnia arytmetyczna (X), a ta jest większa niż dominanta (D). Fakt takiego rozkładu wnioskować można również z wartości poszczególnych kwartyli - Q3 - Q2 > Q2 - Q1.

[4] Porównaj: P. Francuz, R. Mackiewicz, *Liczby nie wiedzą skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów*, Wydawnictwo KUL, Lublin 2005, s. 186 i n.

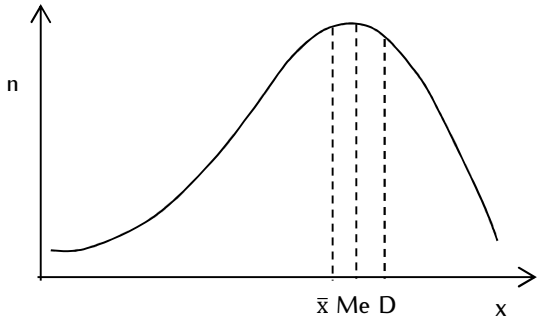
[5] Więcej na temat matematycznych własności rozkładu normalnego znajdziesz w: H.M. Blalock, *Statystyka dla socjologów,* Wydawnictwo Naukowe PWN, Warszawa 1977, Rozdział 7. *Rozkład normalny*, s. 92-104.

Wykres 1. Wykres ilustrujący asymetrię prawostronną



Natomiast jeśli jednostki analizy będą się koncentrowały wokół wysokich wartości zmiennej, współczynnik skośności przyjmie wartość ujemną. Nazywamy go **asymetrycznym lewostronnie,** lewoskośnym lub asymetrią ujemną. W rozkładzie asymetrycznym lewostronnie mediana (Me) jest mniejsza niż średnia arytmetyczna (X), a ta jest mniejsza niż dominanta (D). Z kolei kwartyle spełniają wówczas następującą zależność: Q3 - Q2 < Q2 - Q1.

Wykres 2. Wykres ilustrujący asymetrię lewostronną



W interpretacji współczynnika skośności istotny jest nie tylko znak, ale również jego wielkość. Jeśli rozkład jest idealnie symetryczny, współczynnik skośności przyjmuje wartość zero. Im większa jego wartość (dodatnia lub ujemna), tym większa asymetria. Przyjmuje się jakościową formę interpretacji liczbowych standaryzowanych wartości współczynnika skośności jak w tabeli 24.

Tabela 24. Wartości współczynnika skośności i ich interpretacje

Wielkość współczynnika skośności (wartości mogą być dodatnie lub ujemne)	Interpretacja współczynnika skośności
0 - 0,30	Rozkład symetryczny (0) lub zbliżony do symetrycznego
0,31 - 0,65	Niska asymetria rozkładu
0,66 - 1,30	Umiarkowana asymetria rozkładu
1,31 - 1,60	Wysoka asymetria rozkładu
1,61 - 2,00	Bardzo wysoka asymetria rozkładu

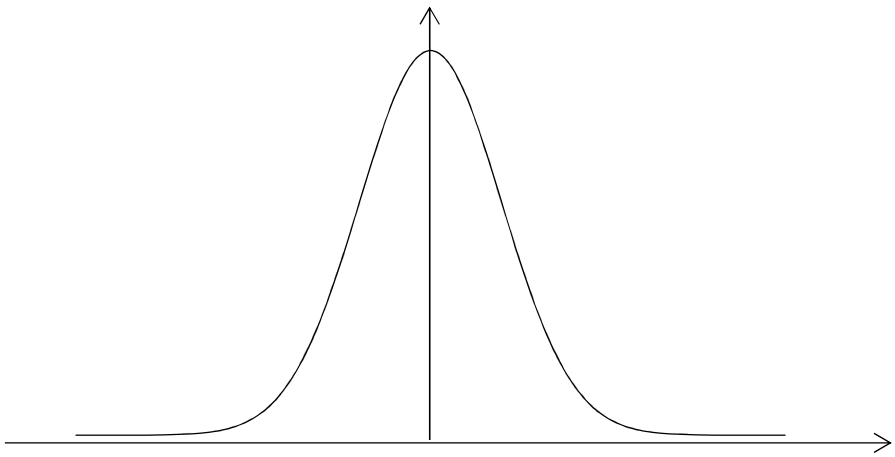
Z kolei **wskaźnik skośności** jest to różnica pomiędzy średnią a modalną. Wynik takiego obliczenia informuje, co prawda o symetrii lub asymetrii rozkładu, ale nie pozwala na porównanie miar asymetrii pochodzących z różnych zbiorów, ponieważ nie jest to wartość standaryzowana.

Współczynnik skośności można obliczyć w programie PSPP zaznaczając *Skewness* w zakładce *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics*.

10.6. Kurtoza (wskaźnik ekscesu)

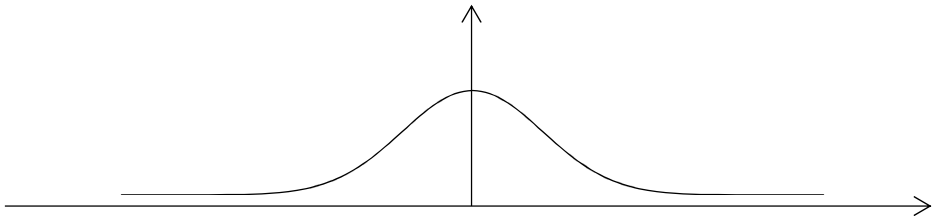
Kurtoza (od gr. *kurtos* - wydęty, wypukły) jest miarą koncentracji rozkładu (płaskości), to jest stopnia koncentracji wartości wokół średniej. Oznaczana jest literą K lub skrótem Kurt. Pojęcie to zostało po raz pierwszy użyte przez Karla Pearsona w 1905 roku. Rozkład normalny posiada kurtozę równą zero (jest to standaryzowana wartość, pierwotnie wynosiła 3); taki rozkład nazywamy **mezokurtycznym**. Tę standaryzowaną wartość powinno określać się właściwie mianem wskaźnika ekscesu. Jeśli kurtoza przyjmuje wartości dodatnie, świadczy to o większej koncentracji wartości niż przy rozkładzie normalnym. Kształt rozkładu jest wówczas bardziej wysmukły, a taki rozkład nazywamy **leptokurtycznym**. Natomiast jeśli kurtoza ma wartość ujemną, świadczy to o większym, niż przy rozkładzie normalnym rozproszeniu wartości cechy, a wykres jest bardziej spłaszczony niż przy rozkładzie normalnym. Wówczas mówimy o rozkładzie **platykurtycznym**. Kurtoza jest miarą bardziej wrażliwą niż wariancja - wydatniej reaguje na pojawianie się odchyleń od wartości oczekiwanej (ujemnych i dodatnich). Wskazówki odnośnie interpretacji kurtozy są w wielu źródłach ogólnikowe - podręczniki pozostają przy opisowym stwierdzeniu, że o normalności rozkładu nie możemy mówić, jeśli wyniki kurtozy są wyraźnie różne od zera[6]. Przyjmuje się, że jeśli kurtoza zawiera się w przedziałach od -1 do +1 wówczas możemy mówić o rozkładzie normalnym.

Wykres 3. Przykład rozkładu leptokurtycznego



[6] Pogłębiona interpretacja kurtozy została obszernie przedstawiona u Leszka Moszczyńskiego w: L. Moszczyński, *Interpretacja współczynnika kurtozy w analizie danych*, „Przegląd Elektrotechniczny", 2003, 79, 9 (1), s. 558-560.

Wykres 4. Przykład rozkładu platykurtycznego



W programie PSPP kurtozę otrzymujemy wybierając *Kurtosis* w zakładce *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics* lub *Analyze* ⇨ *Descriptive Statistics* ⇨ *Descriptives* ⇨ *Statistics*.

11

Rozdział 11. Miary rozrzutu (dyspersji)

Miary koncentracji wskazują typowość danej cechy, z kolei miary rozrzutu - jej rozproszenie wokół średniej. Wykorzystanie wyłącznie miar tendencji centralnej w analizie danych może prowadzić do błędów i nieporozumień. Można wyobrazić sobie sytuację, w której odmienne dwa rozkłady - platykurtyczny i leptokurtyczny mają tę samą średnią, medianę i modalną, a więc z punktu widzenia miar tendencji centralnej nie różnią się. Z tego powodu w opisie statystycznym powinno stosować się obok (lub w szczególnych przypadkach - zamiast) miary rozproszenia takie jak odchylenie standardowe, wariancja, współczynnik zmienności czy kwantylowe miary rozproszenia zmiennej.

11.1. Rozstęp (obszar zmienności)

Najprostszą miarę rozproszenia stanowi rozstęp (R). Jego zrozumienie wymaga wprowadzenia dwóch prostych parametrów - minimum i maksimum. **Minimum** to najmniejsza wartość, jaką może przyjąć zmienna z danego zakresu, natomiast **maksimum** stanowi wartość najwyższą. Z matematycznego punktu widzenia minimum i maksimum to odpowiednio najmniejsze z ograniczeń dolnych i najmniejsze z ograniczeń górnych (a więc kres lub kraniec dolny lub górny) skończonego zbioru danych. Rozstęp (*Range*) to różnica pomiędzy minimum i maksimum:

$$R = n_{max} - n_{min}$$

Wadą tej miary jest znaczna wrażliwość na obecność wartości skrajnych, a zaletą - prostota interpretacji.

Rozstęp może być obliczany po wybraniu w menu *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* ⇨ *Statistics* lub w menu *Analyze* ⇨ *Descriptive Statistics* ⇨ *Descriptives* ⇨ *Statistics*, a następnie poprzez zaznaczenie *Range*.

11.2. Odchylenie średnie (odchylenie przeciętne, średnie odchylenie bezwzględne)

Miara ta oznaczana literami OP lub - rzadziej AD (*Average Deviation*) lub D (*Deviation*) i należy do miar klasycznych. Jest to bezwzględna suma różnic pomiędzy poszczególnymi wartościami, a średnią podzielona przez liczbę jednostek analizy w zbiorze. Wyraża się ona wzorem:

$$OP = \frac{|n_1 - \bar{x}| + |n_2 - \bar{x}| + \cdots + |n_n - \bar{x}|}{n}$$

gdzie: $n_1,\ldots,n_n$ - wartości poszczególnych zmiennych, $\bar{x}$ - średnia arytmetyczna (lub - rzadziej - mediana i wówczas mówimy o odchyleniu przeciętnym od mediany), n - liczba obserwacji dla danej zmiennej w zbiorze.

Odchylenie średnie prezentuje średnią odległość wartości zmiennej od jej średniej. Użyteczność tej miary statystycznej jest ograniczona. Jest ona pewnym etapem obliczania wariancji i odchylenia standardowego. Nie została ona zaimplementowana do programu PSPP, jednak w łatwy sposób można ją obliczyć za pomocą funkcji *Compute.*

11.3. Wariancja

Wariancja jest klasyczną miarą zmienności i jedną z najpopularniejszych miar statystycznych. Wariancja z populacji oznaczana jest greckim znakiem sigmy o wykładniku potęgi równym 2 (σ^2), a wariancja z próby S^2. Intuicyjnie możemy interpretować ją jako miarę mówiącą o stopniu zróżnicowania wyników wokół średniej w danym zbiorze. Znajduje ona liczne zastosowania zarówno jako samodzielna miara, jak też jako część rozmaitych testów statystycznych. Miara ta wykorzystywana jest na przykład w jednoczynnikowej analizie wariancji - *Analysis of Variance,* ANOVA, wieloczynnikowej analizie wariancji *Multivariate Analysis of Variance,* MANOVA, czy teście t-Studenta. Wariancję obliczamy dla zmiennych na poziomie interwałowym lub ilorazowym, w uzasadnionych przypadkach także dla zmiennych porządkowych. W porównaniu z odchyleniem średnim zamiast modułu wykorzystuje się tu pierwiastek kwadratowy w celu wyeliminowania znaków ujemnych, które pojawiają się dla wartości zmiennej niższej niż średnia. Wariancję z próby definiuje się zatem jako sumę podniesionych do kwadratu różnic wartości pojedynczych pomiarów i średniej którą następnie dzieli się przez liczbę jednostek analizy w zbiorze danych:

$$S^2 = \frac{(n_1 - \bar{x})^2 + (n_2 - \bar{x})^2 + \cdots + (n_n - \bar{x})^2}{n}$$

Miara ta przyjmuje wartości od 0 do nieskończoności. Im wartość bliższa zeru, tym zróżnicowanie jest mniejsze. Wariancja o wartości 0 oznacza, że wyniki wszystkich pomiarów w analizowanej grupie są takie same. Wariancja jest wielkością mianowaną, to znaczy wyrażoną w jednostkach badanej cechy statystycznej. Jednak wartość ta wyrażona jest w jednostkach nienaturalnych dla danej zmiennej, bowiem podniesiona jest do kwadratu. Im zbiorowość statystyczna jest bardziej zróżnicowana, tym wartość wariancji jest wyższa. Związek frazeologiczny „brak wariancji" stanowi orzeczenie analityczne, które

mówi nam, że w ramach danej zmiennej nie występują żadne różnice, a więc nie podejmujemy dalszych kroków analitycznych.

Wariancję możemy również obliczać dla zmiennych dychotomicznych zerojedynkowych według następującego wzoru:

$$\sigma^2 = p * (p - 1)$$

gdzie p oznacza średnią arytmetyczną danej zmiennej (frakcję jedynek).

W programie PSPP wariancję obliczamy wybierając z zakładki *Analyze* ⇨ *Descriptive Statistics*, a następnie *Frequencies* lub *Descriptives* a w polu *Statistics* - wariancję (*Variance*). Funkcja wariancji jest także dostępna w *Analyze* ⇨ *Descriptive Statistics* ⇨ *Explore* ⇨ *Statistic* ⇨ *Descriptives*. Ta ostatnia opcja pozwala na porównywanie wariancji według zadanych podziałów w jednej tabeli.

11.4. Odchylenie standardowe

Odchylenie standardowe jest klasyczną miarą statystyczną najczęściej stosowaną obok średniej arytmetycznej. Określa ono, o ile wszystkie jednostki statystyczne danej zbiorowości różnią się przeciętnie od wartości średniej badanej zmiennej. Jak wskazywano przy opisie krzywej normalnej, miara ta posiada szereg istotnych właściwości z punktu widzenia probabilistyki. Odchylenie standardowe wykorzystywane jest do tworzenia typowego obszaru zmienności statystycznej. Dzięki niemu wiemy jakie typowe wartości występują w zbiorze, wartości z jakiego zakresu należy oczekiwać, a jakie są ekstremalnie mało prawdopodobne.

Możemy wyróżnić dwa typy obliczania odchylenia standardowego - odchylenie standardowe z populacji i odchylenie standardowe z próby. Odchylenie standardowe z populacji oznaczamy symbolem sigmy (σ):

$$\sigma = \sqrt{\frac{(n_1 - \bar{x})^2 + (n_2 - \bar{x})^2 + \cdots + (n_n - \bar{x})^2}{n}}$$

gdzie n_1-n_n - wartości poszczególnych zmiennych, $\bar{x}$ - średnia arytmetyczna, n - liczba jednostek analizy w zbiorze danych. Oznacza to, że odchylenie standardowe jest pierwiastkiem z wariancji:

$$S = \sqrt{\sigma^2}$$

Z kolei odchylenie standardowe z próby wyrażamy wzorem:

$$S = \sqrt{\frac{(n_1 - \bar{x})^2 + (n_2 - \bar{x})^2 + \cdots + (n_n - \bar{x})^2}{n - 1}}$$

Jest to tak zwany estymator nieobciążony. Poszczególne symbole wzoru interpretujemy analogicznie. Odchylenie standardowe wyrażane jest w jednostkach miary analizowanej zmiennej. Jeśli pomiaru dokonano w złotówkach, to również wartość tę można w takich jednostkach odczytywać. Odchylenie standardowe, podobnie jak wariancja, zawiera się w zakresie od 0 do nieskończoności. Im mniejsze odchylenie standardowe, tym mniejsze rozproszenie wartości danej zmiennej. Dla odchylenia standardowego równego 0 wszystkie wartości danej zmiennej są równe.

Należy podkreślić, że wadą odchylenia standardowego jest silna podatność na wartości skrajne danej zmiennej - skokowo różniące się od jej rozkładu. Zaleca się, by miarą tą posługiwać się po dokonaniu oceny, czy zmienna nie zawiera tego typu wartości.

W programie PSPP możemy tę miarę obliczyć wybierając z zakładki *Analyze* ⇨ *Descriptive Statistics*, a następnie *Frequencies* lub *Descriptives* a w polu *Statistics* - odchylenie standardowe (*Standard deviation*). Funkcja wariancji jest także dostępna w *Analyze* ⇨ *Descriptive Statistics* ⇨ *Explore* ⇨ *Statistic* ⇨ *Descriptives*. Ta ostatnia opcja pozwala na porównywanie wariancji według zadanych podziałów w jednej tabeli. Odchylenie standardowe obliczamy również w *Analyze* ⇨ *Compare Means* ⇨ *Means*.

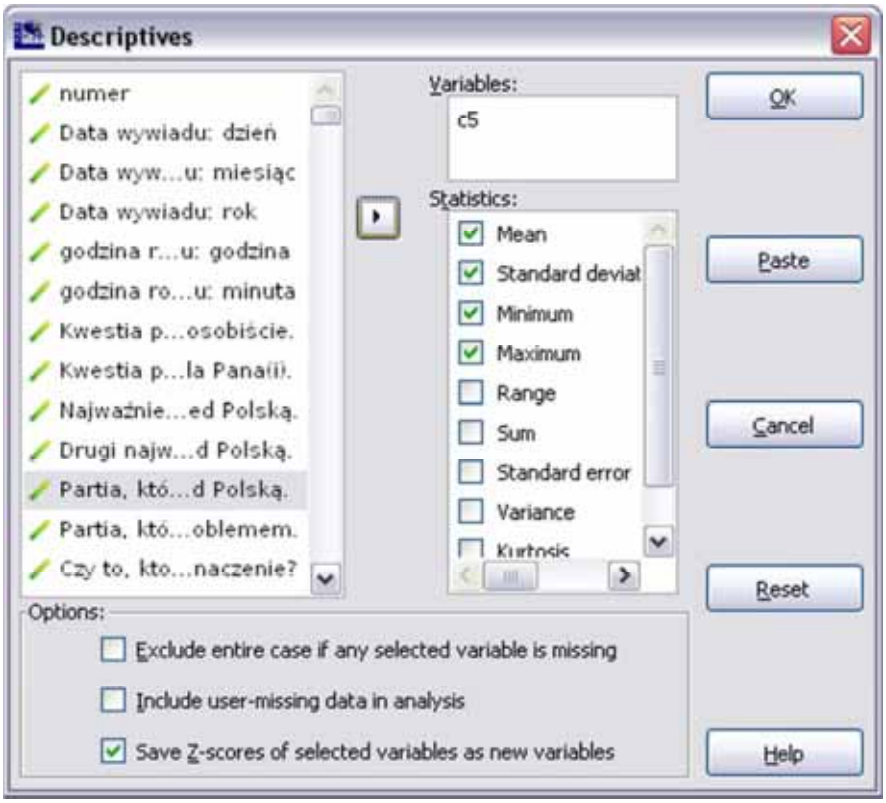
11.5. Standaryzacja empiryczna (standaryzacja typu Z)

Istnieją miary, które pozwalają standaryzować, a następnie porównywać zmienne mierzone na różnorodnych skalach. Zmienna, którą poddajemy standaryzacji posiada rozkład identyczny jak zmienna przed standaryzacją, lecz zmieniają się jej statystyki: średnia arytmetyczna równa jest 0, a odchylenie standardowe równe jest 1. Standaryzacji empirycznej każdej zmiennej dokonujemy za pomocą wzoru:

$$Z = n_n - \frac{\bar{x}}{s}$$

gdzie n_n - pojedyncza zmienna, $\bar{x}$ - średnia arytmetyczna, s - odchylenie standardowe.

Obliczenia miar standaryzowanych dokonujemy za pomocą opcji *Analyze* ⇨ *Descriptive Statistics* ⇨ *Descriptives*, a następnie oznaczamy pole *Zapisz wybrane standaryzowane empirycznie zmienne jako nowe zmienne* (*Save Z-Scores of selected variables as new variables*).



11.6. Kwantylowe miary rozproszenia

Do miar rozproszenia opartych na kwantylach zaliczamy rozstęp ćwiartkowy, odchylenie ćwiartkowe, współczynnik zróżnicowania decylowego oraz wskaźnik wahania decylowego. Miary te stosuje się głównie w przypadku zmiennych mierzonych na poziomach ilościowych, mogą też być stosowane do pomiaru zmiennych porządkowych.

Rozstęp ćwiartkowy (rozstęp kwartylny, *interquartile range,* IQR)**.** Jest to pozycyjny współczynnik zmienności. Używany jest w zastępstwie odchylenia standardowego i średniej arytmetycznej, gdy ze względu na właściwości rozkładu danej zmiennej nie jest wskazane ich obliczanie. Jest to obszar zmienności typowych, wskazujący, w jakim stopniu rozkład zmiennej odchyla się od mediany. Stanowi on różnicę pomiędzy czwartym i pierwszym kwartylem:

$$R_Q = Q_3 - Q_1$$

Pomiędzy trzecim, a pierwszym kwartylem mieści się połowa wszystkich jednostek analizy. Im większa jego szerokość tym większe zróżnicowanie zmiennej. Wskazaniem dla zastosowania tej miary jest otwarty szereg danej zmiennej lub rozkład obserwowanej zbiorowości jest niekompletny na krańcach szeregu, zmienna przyjmuje wartości znacznie odbiegające od średniej i możliwe jest zastosowanie jako miary pozycyjnej tylko mediany. Niekiedy miara ta poddawana jest w celach porównawczych standaryzacji (uniezależnienia od jednostek analizy) i wówczas wyrażana jest formułą:

$$R_{Q_s} = \frac{Q_3 - Q_1}{n}$$

gdzie n jest sumą liczebności jednostek analizy znajdujących się w pierwszym i trzecim kwartylu.

W programie PSPP rozstęp ćwiartkowy obliczany jest automatycznie w tabeli, obok innych miar, po wybraniu *Analyze* ⇨ *Descriptive Statistics* ⇨ *Explore* ⇨ *Statistic* ⇨ *Descriptives*.

Odchylenie ćwiartkowe. Jest to odmiana rozstępu ćwiartkowego. Mieści w sobie połowę jednostek analizy znajdujących się pomiędzy trzecim a pierwszym kwartylem. Miarę tę wyraża wzór:

$$OC = \frac{Q_3 - Q_1}{2}$$

Współczynnik zróżnicowania decylowego. Jest to iloraz dziewiątego i pierwszego decyla wartości danej zmiennej. Stosowany jest przy badaniu nierówności społecznych (wyrażanych zróżnicowaniem dochodów), posługuje się tą miarą między innymi Główny Urząd Statystyczny. Wyrażany jest wzorem:

$$D_{9/1} = \frac{D_9}{D_1}$$

Wartości współczynnika zróżnicowania decylowego mogą zawierać się w granicach od 1 do $+\infty$. Niezróżnicowanie, czyli skrajny egalitaryzm wyrażany jest wartością współczynnika równą 1. Im większe dysproporcje, tym współczynnik jest wyższy. Miarę tę czasami również wyraża się w postaci następującej:

$$ZD = \frac{D_9}{D_1 * 100}$$

Współczynnik zróżnicowania skrajnych części rozkładu. Miara ta stosowana jest jako modyfikacja powyższej. Jest to udział sumy wartości należących do pierwszej grupy decylowej w sumie wartości należących do najwyższej (dziesiątej grupy decylowej). Formuła współczynnika jest następująca:

$$K_{1/10} = \left(\frac{K_1}{K_{10}}\right) * 100$$

gdzie K_1 oznacza sumę wartości w przedziale wartości najniższych (dolne 10 proc.), K_{10} sumę dochodów w przedziale wartości najwyższych (górne 10 proc.). Współczynnik ten przyjmuje wartości z zakresu od 0 do 100 procent. Należy go interpretować jako udział grupy o najniższych wartościach zmiennej w grupie o najwyższych wartościach zmiennej. Im wartość jest niższa tym większe jest zróżnicowanie obu grup. Wartość 100 procent oznacza brak zróżnicowania. Im wartości są niższe, tym dysproporcje większe; wartość 0 oznacza, że dolna grupa decylowa nie posiada żadnych udziałów w grupie wyższej, a więc suma wartości w tej grupie równa jest zero. Podobnie jak współczynnika zróżnicowania decylowego miary tej używa się do pomiaru nierówności społecznych. Na dużą skalę wykorzystuje ją GUS.

Wskaźnik wahania decylowego jest wyrażonym w procentach stosunkiem różnic pomiędzy dziewiątym i pierwszym decylem do wartości mediany:

$$WD = \left(\frac{D_9 - D_1}{M(x)}\right) * 100$$

Wskaźnik ten przyjmuje wartości od zera do plus nieskończoności. Niska wartość wskaźnika oznacza małe zróżnicowanie. Im jest większa, tym zróżnicowanie jest bardziej pokaźne.

11.7. Współczynnik zmienności (klasyczny współczynnik zmienności)

Współczynnik zmienności to miara siły rozproszenia, która jest niezależna od skali na której mierzymy wartości, a także od zmian wartości zmiennych w kolejnych pomiarach. Oznaczany jest on literą v. Współczynnik zmienności umożliwia porównywanie zmiennych mierzonych na różnych skalach. Wskazaniem dla jego zastosowania jest pomiar zjawiska mierzonego w różnych jednostkach miary lub kształtowanie się go na niejednakowym poziomie.

Współczynnik zmienności pozwala także na ocenę wartości poznawczej średniej arytmetycznej. Ponadto jest cząstkową miarą między innymi współczynnika korelacji. Miara ta wyrażana jest ilorazem odchylenia standardowego w próbie i średniej arytmetycznej w próbie.

Zapisywany jest wzorem:

$$v = \frac{s}{\bar{x}}$$

gdzie s to odchylenie standardowe, a $\bar{x}$ to średnia. Czasami mnoży się wartość wynikową przez 100 proc. i podaje właśnie w takiej formie. Miarę tę interpretujemy jak w tabeli 25.

Tabela 25. Wartości współczynnika zmienności i ich interpretacje

Wielkość klasycznego współczynnika zmienności	Interpretacja współczynnika zmienności
od 0 do 20	nieznaczne zróżnicowanie wartości zmiennej – możemy mówić o względnej jednorodności zmiennej, średnia arytmetyczna jest adekwatną miarą charakteryzującą zmienną
powyżej 20 do 40	umiarkowane zróżnicowanie wartości zmiennej – zróżnicowanie zmiennej ma charakter przeciętny, średnia arytmetyczna jest akceptowalną miarą dla danej zmiennej
powyżej 40 do 60	silne zróżnicowanie wartości zmiennej – rozproszenie zmiennej jest znaczne, średnia arytmetyczna ma niewielką wartość eksplanacyjną
powyżej 60	bardzo silne zróżnicowanie wartości zmiennej – dyspersja jest bardzo duża, średnia arytmetyczna nie ma żadnej wartości poznawczej

12

Rozdział 12. Sztuka przewidywania zjawisk - ryzyko względne i iloraz szans

Ryzyko względne (ang. *relative risk, RR* lub inaczej współczynnik ryzyk, iloraz ryzyk (ang. *risk ratio*)) oraz iloraz szans są miarami prawdopodobieństwa wystąpienia zjawiska pod wpływem danego czynnika lub jego braku. W najprostszym rozumieniu miara ta służy nam do przewidywania jak bardzo prawdopodobne jest pojawienie się obserwowanego zjawiska jeśli oddziałuje jakiś czynnik w porównaniu z sytuacją, gdy dany czynnik nie wystąpił. Pomiar zakłada zestawienie dwóch badanych grup oraz dwóch stanów danego zjawiska (zjawisko wystąpiło vs. zjawisko nie wystąpiło). Stosując tę miarę możemy na przykład w zmatematyzowanej, liczbowej formie orzekać o ile większe jest prawdopodobieństwo spowodowania wypadku przez kierowcę pod wpływem alkoholu niż spowodowania wypadku przez kierowcę trzeźwego. Współczynniki mierzymy porównując ze sobą dwie dychotomiczne (dwuwartościowe) zmienne (a więc zmienne w tzw. tabeli 2x2, czyli czteropolowej). Miara ta jest szeroko stosowana w statystyce medycznej (w badaniach klinicznych i epidemiologicznych) oraz w innych naukach - w badaniach eksperymentalnych. Wywodzi się ona od narzędzi analitycznych projektowanych w branży ubezpieczeń (rynek finansowy). Zapoczątkowała je w XVIII wieku brytyjska firma Lloyd's, która wprowadziła ubezpieczenia statków płynących do Indii.

Wydaje się, że ta metoda analizy może być również pomocna i wartościowa poznawczo w politologii. Jest ona przydatna zarówno przy pomiarze danych zebranych w warunkach eksperymentalnych, jak również danych wtórnych, pochodzących z eksperymentów naturalnych, *ex post factum.*

Miarę tę charakteryzują trzy parametry: ryzyko względne, iloraz szans oraz przedział ufności (odrębnie dla ilorazu szans i ryzyka względnego). Kluczowe są parametry wymienione jako pierwszy i trzeci, drugi z nich ma charakter pomocniczy.

12.1. Ryzyko względne

Ryzyko względne oznaczamy z angielskiego skrótem RR lub w języku polskim - R_W. Mierzy ono ryzyko wystąpienia jakiegoś zjawiska w jednej grupie w porównaniu do wystąpienia tego zjawiska w innej grupie (w uproszczeniu: jest to porównanie stosunku częstości jakiegoś zdarzenia w dwóch grupach). Pojęcie ryzyka może być rozumiane potocznie, intuicyjnie. Miara ta, jak wskazano, pierwotnie stosowana była w medycynie. Klasycznym jej zastosowaniem na tym polu badawczym była ocena ryzyka zapadnięcia na jakąś chorobę (lub zgonu) pod wpływem jakiegoś czynnika zewnętrznego. Często przytaczanym przykładem wykorzystania tej metody jest szacowanie ryzyka zachorowania na nowotwór płuc przez osoby palące i osoby niepalące.

Sposób obliczania i interpretacji ryzyka względnego przeanalizujmy w następującym przykładzie. Przypuśćmy, że chcielibyśmy ocenić w jakiej sytuacji dochodzi do użycia przemocy przez protestujący tłum. Konkretnie pragniemy sprawdzić czy fakt obecności polityka, jego wyjście do protestujących, jeśli tłum tego żąda, ma znaczenie czy go nie ma. Mamy zatem do czynienia z dwiema dychotomicznymi zmiennymi: kontakt z politykiem (nastąpił vs. nie nastąpił) oraz użycie przemocy przez demonstrujących (nastąpiło vs. nie nastąpiło). Przypuśćmy, że zbadaliśmy 50 przypadków protestów społecznych, w których tłum zażądał kontaktu z politykiem. W niektórych przypadkach kontakt ten miał miejsce, a w innych - nie. Tłum w konsekwencji użył przemocy lub nie użył jej. Dane do obliczenia ryzyka względnego należy przedstawić jak w tabeli 26.

Tabela 26. Fikcyjne dane użycia lub braku użycia przemocy przez protestujących w sytuacji kontaktu lub braku kontaktu z politykiem (N=50)

Kontakt z politykiem	Użycie przemocy przez protestujących		Ogółem
	Tak	Nie	
Tak	8	17	25
Nie	17	8	25
Ogółem	25	25	50

Rozważając dane liczbowe z tabeli x dostrzegamy, że kontakt z politykiem sprzyja zachowaniu spokoju przez demonstrujących (tłum użył przemocy tylko w 8 przypadkach na 25), z kolei brak obecności polityka, jeśli tłum zażąda spotkania z nim sprzyja użyciu przemocy (aż 17 przypadków na 25). Można jednak te obserwacje przedstawić w bardziej standaryzowanej formie, a służy do tego ryzyko względne. Wprowadzona poniżej tabela 27 posłuży do wyjaśnienia zasad obliczania ryzyka względnego. Tabela ta stanowi uogólnioną postać tabeli 26.

Tabela 27. Poglądowy schemat tabeli służącej do obliczania ryzyka względnego

Zmienna 1 (oznaczająca badane grupy)	Zmienna 2 (oznaczająca wystąpienie lub brak spodziewanego zjawiska)		Ogółem
	Tak	Nie	
Tak	A	B	A + B
Nie	C	D	C + D
Ogółem	A + C	B + D	A + B + C + D

Obliczanie ryzyka względnego rozpoczynamy od wyliczenia prawdopodobieństwa wystąpienia zjawiska w obu badanych grupach (patrz: zmienna 1). Prawdopodobieństwo użycia przemocy w grupie, która miała kontakt z politykiem zapisujemy za pomocą wzoru A / A + B i analogicznie prawdopodobieństwo jej użycia w grupie, która nie miała kontaktu z politykiem: C / C + D.

Ryzyko względne jest to stosunek tych dwóch prawdopodobieństw:

$$R_w = \frac{A * (C + D)}{C * (A + B)}$$

Podstawmy do powyższego wzoru wartości z tabeli 26 obliczmy współczynnik i zinterpretujmy wyniki.

$$R_w = \frac{8 * (17 + 8)}{17 * (8 + 17)} = \frac{8 * 25}{17 * 25} = \frac{200}{425} = 0{,}47$$

Zasady interpretacji są następujące: wartość ryzyka względnego zawiera się w przedziale od 0 do nieskończoności. Jeśli współczynnik ten jest równy 1 oznacza to, że występowanie lub niewystępowanie danego czynnika nie ma wpływu na wystąpienie lub niewystąpienie badanego zjawiska. Wartość współczynnika ryzyka względnego powyżej jedności oznacza, że dana zmienna ma charakter stymulujący określone zjawisko, a poniżej jedności - że je destymuluje. Na przykład współczynnik o wartości 1,22 będziemy rozumieli następująco: ryzyko wystąpienia zjawiska w jednej grupie pod wpływem oddziaływania czynnika x jest o 22 proc. większe niż w grupie, gdzie czynnik x nie wystąpił. Jeśli natomiast mamy do czynienia ze współczynnikiem o wartości 0,52, oznacza to, że ryzyko wystąpienia zjawiska w grupie gdzie dany czynnik oddziaływał wynosi zaledwie 52 proc. w porównaniu z grupą, gdzie był on nieobecny (a więc ryzyko jest o 48 proc. mniejsze). W pierwszym przypadku ów czynnik x stanowi stymulantę zjawiska, w drugim - destymulantę.

Uzyskany na podstawie tabeli 26 wynik rozumiemy następująco: ryzyko użycia przemocy w czasie demonstracji, gdy nastąpi kontakt z politykiem wynosi zaledwie 47 proc. (gdy za 100 proc. przyjmiemy brak wyjścia polityka do protestujących). Zwróćmy uwagę, że interpretując wynik możemy mówić o ryzyku względnym w dwóch kontekstach: gdy zjawiska mają charakter negatywny uzyskany wynik określamy mianem ryzyka względnego, a gdy mają charakter pozytywny - o korzyści względnej. Uzyskany wyżej wynik można więc wyrazić innymi słowy: „korzyść względna kontaktu polityka z tłumem wynosi 47 proc." lub jeszcze inaczej: „ryzyko użycia przemocy przez tłum, gdy polityk odmówił z nim kontaktu jest 2,12 razy wyższe, niż gdy interakcja z tłumem nastąpiła" (bowiem: 100 / 47 = 2,12, gdzie 100 oznacza wartość bezwzględną - 100 proc.).

12.2. Iloraz szans

Iloraz szans (ang. *odds ratio*) oznaczamy skrótem OR lub polskojęzycznym IS. Pojęcie „szansa" rozumiana jest jako prawdopodobieństwo, że jakieś zdarzenie wystąpi do prawdopodobieństwa, że ono nie wystąpi. Obliczenie tego współczynnika jest dokonywane wedle następującego wzoru:

$$\mathrm{IS} = \frac{\mathrm{A} * \mathrm{D}}{\mathrm{C} * \mathrm{B}}$$

Interpretujemy współczynnik analogicznie do współczynnika ryzyka względnego: iloraz szans równy jedności wskazuje, że szanse porównywanych grup są tożsame, nie różnią się od siebie. Jeśli współczynnik jest większy od jedności wówczas oznacza to, że szansa wystąpienia danego zdarzenia jest większa w pierwszej grupie. Z kolei iloraz szans poniżej jedności wskazuje, że w pierwszej grupie szansa wystąpienia badanego zdarzenia zdrowotnego jest mniejsza niż w drugiej.

Podstawmy do wzoru dane z tabeli 26, obliczmy i zinterpretujmy iloraz szans:

$$\mathrm{IS} = \frac{8 * 8}{17 * 17} = \frac{64}{289} = 0{,}22$$

W przybliżeniu, prawdopodobieństwo użycia przemocy przez protestujących, z którymi porozmawiał polityk wynosi jedną piątą prawdopodobieństwa zachowania spokoju przez demonstrantów w grupie zdarzeń, gdzie polityk nawiązał z nimi kontakt. Innymi słowy: szansa, że wystąpi użycie przemocy w sytuacji, gdy polityk odmówi kontaktu wynosi (w przybliżeniu) 5:1.

Miara ta jest trudna do bezpośredniej interpretacji. Spowodowała ona liczne problemy interpretacyjne i nieporozumienia wśród autorów z dziedziny nauk medycznych (interpretowana jest jako ryzyko względne, co jest błędem)[1]. W związku z tym zaleca się, by miarę tę stosować tylko wówczas, gdy bezpośrednie obliczenie ryzyka względnego nie jest z jakichś powodów możliwe.

12.3. Obliczanie przedziału ufności dla ryzyka względnego

Przedział ufności obliczamy, gdy chcemy przenieść wynik z próby (zbioru, który posłużył nam do obliczeń) na całą badaną populację. Wynik takiej estymacji będzie zawierał się w przedziałach, nie możemy zgodnie z zasadami prawdopodobieństwa podać wyniku punktowego. Dla przyjmowanego powszechnie w naukach społecznych przedziału ufności równemu 95 proc. (α=0,05) wynik po uogólnieniu na populację będzie zawierał się dla ryzyka względnego pomiędzy:

$$\mathrm{R_w} * \mathrm{e}^{-1{,}96\sqrt{\mathrm{V}}} \text{ , a } \mathrm{R_w} * \mathrm{e}^{1{,}96\sqrt{\mathrm{V}}}$$

gdzie:

R_W - ryzyko względne obliczone według powyżej podanego wzoru,

[1] Na temat błędów interpretacyjnych współczynnika iloraz u szans wypowiadają się: W.L. Holcomb, T. Chaiworapongsa, D.A. Luke, K.D. Burgdorf, *An odd measure of risk: use and misuse of the odds ratio*, „Obstetrics and Gynecology", 2001, 98(4), s. 685-688.

e - to liczba stała (liczba Eulera lub liczba Nepera), jest to podstawa logarytmu naturalnego, która w przybliżeniu wynosi 2,718,

1,96 - wartość obliczona na podstawie tablic rozkładu normalnego dla poziomu ufności 95 proc. (dla poziomu ufności 99 proc. podstawilibyśmy do tego wzoru 2,58, a dla 90 proc. - 1,64),

z kolei V jest obliczane następująco:

$$V = \left(\frac{1-x}{A+B}\right) + \left(\frac{1-y}{C+D}\right)$$

gdzie

$$x = \frac{A}{A+B} \quad \text{zaś} \quad y = \frac{C}{C+D}$$

Po podstawieniu do wzoru otrzymamy dla 0,44 przedział ufności od 0,25 do 0,88. Oznacza to, że po uogólnieniu wyników na całą populację wyniki zgodnie z zasadami stochastyki będą zawierać się właśnie w tym zakresie, a więc od 25 procent do 88 procent. Zwraca uwagę szerokość tego przedziału - im większa liczba zbadanych przypadków (jednostek analizy), tym ten przedział węższy.

Przedział ufności dla ilorazu szans liczymy według analogicznego wzoru w miejsce R_W wstawiając wartość ilorazu szans, zaś V obliczając wedle wzoru:

$$V = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

12.4. Obliczanie ryzyka względnego i ilorazu szans w programie PSPP

Do programu PSPP dane, jak w podanym wyżej przykładzie w tabeli 26 można wprowadzić w następujący, ekonomiczny sposób z użyciem wag:

	V1	V2	waga
1	1	1	8
2	0	1	17
3	1	0	17
4	0	0	8

Ryzyko względne i iloraz szans obliczamy wybierając z menu tekstowego *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs* ⇨ *Statistics* i zaznaczamy *Risk.* Należy pamiętać o tym, aby w oknie *Crosstabs* w wierszach (*Rows*) umieścić zmienną zawierającą dwie badane grupy, a w kolumnach (*Columns*) zmienną oznaczającą wystąpienie lub brak danego zjawiska.

Risk estimate.

Statistic	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Uzycie przemocy (0 / 1)	,22	,07	,73
For cohort Kontakt z politykiem = 0	,47	,25	,88
For cohort Kontakt z politykiem = 1	2,13	1,13	4,00
N of Valid Cases	50		

Program PSPP oblicza trzy opisane wyżej parametry. Iloraz szans (*Odds Ratio*) znajduje się w pierwszym wierszu tabeli, ryzyko względne (są to wiersze rozpoczynające się od słów *For cohort...*) w drugim (dla jednej grupy) i w trzecim (dla drugiej grupy). W kolumnie *Value* podane są wartości współczynników. W kolejnych dwóch kolumnach podane są przedziały po ich estymowaniu na populację - odpowiednio dolny (*Lower*) i górny (*Upper*) przedział. Zwróćmy uwagę, że współczynniki te są automatycznie przez program obliczane na standardowym poziomie ufności 95 proc.

IV

Część IV. Badanie zależności między zmiennymi

13

Rozdział 13. Badanie zależności między zmiennymi - miary związku

Badanie zależności pomiędzy zmiennymi nosi nazwę teorii współzależności. Pomiar taki polega na sprawdzeniu, czy zmianom wartości jednej zmiennej towarzyszą zmiany wartości drugiej zmiennej. Jest to najprostsza sytuacja, gdy badamy związek pomiędzy dwiema zmiennymi. Takie właśnie elementarne zależności rozważane są w niniejszym rozdziale. Przykładowo, dzięki zastosowaniu miar zależności, możemy odpowiedzieć na następujące pytania:

Czy fakt bycia kobietą lub mężczyzną wpływa na podejmowanie aktywności politycznej lub społecznej? Czy płeć, wiek, wykształcenie lub wielkość miejsca zamieszkania pozwala przewidywać stopień poparcia dla demokracji lub Unii Europejskiej? Czy istnieje związek pomiędzy liczbą lat nauki a wysokością zarobków? Czy istnieje współzależność pomiędzy rodzajem protestującej grupy (pielęgniarki, górnicy) a stopniem natężenia przemocy przejawianej przez demonstrujących?

W zależności od tego, na jakich poziomach są mierzone badane zmienne (ilorazowym, interwałowym, porządkowym, czy nominalnym), jak liczebne są jednostki analizy oraz jaki typ rozkładu zmienne te reprezentują, dobieramy różne miary do orzekania, czy związek pomiędzy nimi występuje czy nie.

W literaturze przedmiotu stosuje się przede wszystkim podział na testy parametryczne i testy nieparametryczne. **Testy parametryczne** opierają się na średniej oraz na wariancji (a więc parametrach rozkładu zmiennej). Są one przeznaczone do badania zależności pomiędzy zmiennymi mierzonymi na poziomach silnych, ilościowych (ilorazowych i interwałowych). Są one też bardziej wymagające - żądają spełnienia rygorystycznych warunków, z których najważniejszy dotyczy odpowiedniego rozkładu obu badanych zmiennych (najczęściej jest to rozkład normalny), a także jednorodności ich wariancji. Testy te mają większą moc niż nieparametryczne, a dzięki temu z większą pewnością można dowieść postawionej hipotezy statystycznej o związku między badanymi zmiennymi. Ponadto dają one pełniejsze wyniki: mierzą nie tylko istotność statystycznego związku i jego siłę, ale także jego kierunek. Są bardziej precyzyjne: zależności nie muszą być znaczne, wystarczy jeśli są regularne, bo wówczas nawet niewielkie, ale systematyczne związki są wykrywane. Najważniejszym testem parametrycznym służącym do pomiaru zależności między zmiennymi jest współczynnik korelacji R Karla Pearsona.

Z kolei **testy nieparametryczne** służą do pomiaru związków dla zmiennych mierzonych na poziomach jakościowych. W tym przypadku nie można obliczyć parametrów takich jak średnia (stąd: testy nieparametryczne). Są one mniej dokładne, część z nich nie mierzy na przykład kierunku zależności, jednakże ich niewątpliwą zaletą jest fakt, że nie są one tak restrykcyjne jak testy parametryczne (nie jest na przykład wymagany rozkład normalny zmiennych). Do najważniejszych testów nieparametrycznych należą współczynnik korelacji rangowej rho (ρ) Spearmana oraz test zgodności chi-kwadrat (χ^2) Pearsona.

Z punktu widzenia ortodoksyjnej statystyki wyników testów nieparametrycznych nie wolno uogólniać na populację, jednakże w praktyce badawczej coraz częściej dokonuje się takich zabiegów; wskutek tego przedstawiony powyżej podział zaczyna się zacierać.

W naukach społecznych większość istotnych dla badaczy zmiennych jest mierzona na poziomach jakościowych: najczęściej nominalnych, rzadziej na poziomach porządkowych. W związku z tym, jako badacze zjawisk społecznych, jesteśmy niejako „skazani" na testy nieparametryczne. Tam, gdzie to możliwe i mamy wpływ na tworzenie narzędzia badawczego, powinniśmy starać się mierzyć zmienne na poziomach silnych.

W niniejszym rozdziale prezentowane są rozmaite, dostępne w programie PSPP miary zależności pomiędzy zmiennymi. Dobór odpowiednich miar powinien przede wszystkim opierać się na przesłankach merytorycznych. Faktycznie jednak często rządzi moda na stosowanie tych a nie innych miar - moda ta zmienia się w czasie. Dodatkowo, odrębne mody wytworzyły się w ramach różnych dyscyplin badawczych. Niniejszy podręcznik łamie te konwencje, dostarczając Czytelnikowi różnych narzędzi, które są niezależne od statystycznej mody.

Warto przedłożyć kilka rad dla początkujących analityków danych, którzy chcą wykorzystać statystyki służące do pomiaru zależności między zmiennymi. Jest to ostrzeżenie, by do liczb podchodzić krytycznie, nie popadać w *numerolatrię* (ubóstwienie liczb) lub „fetyszyzm statystyczny" (przedkładanie wyników liczbowych nad zdrowy rozsądek).

Po pierwsze, liczby nie mówią same za siebie. Fakt istnienia matematycznego związku pomiędzy zmiennymi nie oznacza faktycznego, rzeczywistego związku pomiędzy nimi w rozumieniu logicznym i społecznym. Korelacja w sensie matematycznym nie oznacza związku przyczynowo-skutkowego. Zanim skupimy się na statystycznej analizie zmiennych, konieczne jest uzasadnienie, dlaczego badamy związek właśnie tych, a nie innych zmiennych. Dobrym uzasadnieniem jest fakt odnotowywania istnienia takiego związku w literaturze przedmiotu, mniej satysfakcjonujące, lecz wystarczające jest powołanie się na intuicję badacza, według którego z jakichś uzasadnionych na gruncie zdroworozsądkowego rozumowania powodów związek taki prawdopodobnie istnieje.

Po drugie, konieczna jest odpowiednia doza podejrzliwości w stosunku do uzyskiwanych przez badacza wyników. Warto przestrzec przed dwiema sytuacjami. Pierwszą z nich można zilustrować przykładem następującym: badając populację dorosłych Polaków możemy odnotować istotny, silny związek pomiędzy wzrostem a intensywnością korzystania z solariów. Odkrywamy bowiem, że im niższy wzrost tym... częściej dana jednostka korzysta z usługi sztucznego opalania. Problem, o którym tu mowa nazywany jest w literaturze przedmiotu **korelacją pozorną** lub **iluzoryczną**. W istocie badacz uzyskujący taki wynik nie wziął pod uwagę złożoności badanej materii, nie dostrzegając oddziaływania innych wpływów niż tylko dwie spośród wybranych zmiennych. Mamy tu do czynienia z oddziaływaniem zmiennej ukrytej. Jest nią płeć. Kobiety mają w porównaniu z mężczyznami niższy wzrost, a jednocześnie częściej niż mężczyźni korzystają z solariów. Stąd wzięła się zależność pomiędzy wzrostem a intensywnością korzystania

z solariów. Drugie ostrzeżenie, skłaniające do ostrożnego traktowania zebranych danych dotyczy sytuacji, w której uzyskujemy istotny statystycznie wynik i korelację pomiędzy zmiennymi, które w istocie mierzą to samo zjawisko. Na przykład poparcie dla demokracji i niechęć do totalitarnych praktyk to w istocie dwie strony tego samego medalu. Tu konieczna jest szczególna ostrożność, bowiem w wielu przypadkach taka właśnie relacja pomiędzy dwoma zmiennymi nie jest aż tak oczywista i łatwo dostrzegalna jak w podanym przykładzie.

Zanim Czytelnik przystąpi do czytania niniejszego rozdziału konieczne jest przestudiowanie i zrozumienie rozdziału dotyczącego regresji liniowej. Ponadto przygodę z miarami korelacji sugeruje się rozpocząć od podrozdziału dotyczącego korelacji R Pearsona.

13.1. Miary związku dla zmiennych ilościowych

W niniejszym podrozdziale zostały omówione najbardziej popularne i najczęściej stosowane parametryczne miary zależności: współczynnik korelacji liniowej R Pearsona, stosunek korelacji nieliniowej eta oraz współczynnik zgodności kappa Cohena.

13.1.1. Współczynnik korelacji R Pearsona

Współczynnik korelacji liniowej według momentu iloczynowego (*Pearson product moment correlation,* PPMC) nazywany współczynnikiem R Pearsona lub – dla odróżnienia od innych miar związku – kowariancyjnym współczynnikiem korelacji. Jest jednym z największych odkryć metody statystycznej i najszerzej stosowaną ilościową miarą współzależności zmiennych. Nazwa współczynnika pochodzi od nazwiska jego głównego twórcy Carla (później Karla) Pearsona (1857-1936) angielskiego matematyka, twórcy pierwszego na świecie wydziału statystyki (Wydział Statystyki Stosowanej) w London University College. Swój udział w jego stworzeniu miał również francuski astronom Auguste Bravais (a niektórzy sądzą, że to właśnie jemu powinno się przyznać palmę pierwszeństwa za wynalezienie tej miary). Ideę korelacji, choć w niezmatematyzowanej formie, analizowali także Karol Darwin (w 1868 roku) oraz John Stuart Mill (w 1843 roku). Ten ostatni dał filozoficzne podstawy miary w słynnych schematach wnioskowania indukcyjnego nazywanych kanonami Milla (chodzi tu o metodę IV, a więc tak zwany kanon zmian towarzyszących). Do powstania tej miary walnie przyczynili się również Carl Friedrich Gauss i Franciszek Galton, przy czym ten pierwszy nie zdał sobie sprawy z istoty i potencjalnych zastosowań współczynnika. Niektórzy wskazują nawet, że idea współczynnika korelacji pojawiła się po raz pierwszy w dziejach ludzkości w świecie starożytnym za sprawą Arystotelesa, który w dziele *Historia animalium* wykorzystywał tę koncepcję w studiach nad klasyfikacją zwierząt. Powszechnie jednak za wynalazcę współczynnika korelacji uważa się Karla Pearsona[1]. Miara ta powstała na gruncie biometrii, gdy K. Pearson usiłował sformułować w formie matematycznego współczynnika prawo nauki w zakresie dziedziczenia cech biologicznych[2].

[1] J.L. Rodgers, W.A. Nicewander, *Thirteen Ways to Look at the Correlation Coefficient*, „The American Statistican", 1988, 42 (1), s. 59-66.

[2] J. Gayon, *Darwinism's Struggle for Survival. Heredity and the Hypothesis of Natural Selection*, Cambridge University Press, Cambridge 1998, s. 237-238.

Współczynnik korelacji stanowi składnik wielu bardziej zaawansowanych miar statystycznych, jak na przykład analizy czynnikowej czy modeli równań strukturalnych (LISREL). Ponadto jest on główną miarą współzależności - inne współczynniki takie jak rho Spearmana czy phi są w istocie wariacjami współczynnika R Pearsona i stanowią zaledwie jego adaptację do danych mierzonych na innych poziomach skal. Przede wszystkim należy podkreślić, że współczynnik korelacji jest obecnie nie tylko szeroko stosowany, lecz - jak wskazuje wielu autorów - w dużym stopniu nadużywany[3]. Niebezpieczeństwa związane z jego stosowaniem zostały sprecyzowane w dalszej części rozdziału.

Pojęcie „korelacja" wywodzi się ze średniowiecznej łaciny (*correlatio* oznaczające współzależność), skąd trafiło do europejskich języków narodowych i oznacza wzajemne powiązanie, wzajemną zależność przedmiotów, pojęć, zagadnień, zjawisk lub zmiennych matematycznych. W statystyce i matematyce oznacza ono związek między zmiennymi losowymi polegający na tym, że zmianie jednej zmiennej towarzyszy zmiana drugiej[4]. Korelacja może oznaczać związek oparty na założeniu, iż gdy jedna zmienna rośnie, to wartość drugiej zmiennej również wzrasta. Może też oznaczać sytuację, w której wzrostowi jednej zmiennej towarzyszy zmniejszanie się wartości drugiej. Przykładem korelacyjnego związku dwóch zmiennych jest na przykład wzrost i waga. Im wyższy wzrost, tym wyższa waga. Im wyższe wykształcenie, tym wyższe zarobki. Im wyższy staż pracy, tym wyższe zarobki. Im wyższy dochód, tym wyższe wydatki. Im wyższe zanieczyszczenie powietrza, tym większa liczba zachorowań na choroby układu krążenia. Pamiętajmy, że korelacją nazwiemy również takie sytuacje, gdy wartość jednej zmiennej rośnie, podczas gdy wartość drugiej - maleje.

Obliczanie współczynnika korelacji R Pearsona. Współczynnik korelacji R Pearsona jest na tyle istotnym narzędziem analityka, że warto zrozumieć i przyswoić sposób jego obliczania bez użycia komputera. Ponadto, oblicza się go banalnie łatwo. Wzór na obliczanie współczynnika korelacji R Pearsona jest następujący:

$$R = \frac{\text{kowariancja (A, B)}}{S_A * S_B}$$

gdzie kowariancję obliczamy według wzoru:

$$\text{kowariancja (A, B)} = \frac{(a_1 - A_{\text{średnia}})(b_1 - B_{\text{średnia}}) + \cdots + (a_n - A_{\text{średnia}})(b_n - B_{\text{średnia}})}{N}$$

Z kolei odchylenie standardowe dla $S_A * S_B$ obliczamy według wzoru:

$$S_A * S_B = \sqrt{\frac{(a_1 - A_{\text{średnia}})^2 + \cdots + (a_n - A_{\text{średnia}})^2}{n}} * \sqrt{\frac{(b_1 - B_{\text{średnia}})^2 + \cdots + (b_n - B_{\text{średnia}})^2}{n}}$$

W poniższej tabeli znajdują się przykładowe dane, na podstawie których zostanie obliczony współczynnik R Pearsona. W pierwszej kolumnie tabeli znajdują się liczby porządkowe; współczynnik zostanie obliczony na podstawie pięciu jednostek analizy. Kolumna druga i trzecia reprezentuje wartości poszczególnych zmiennych (a i b) dla pięciu przypadków. Pozostałe kolumny oraz dwa najniżej położone wiersze zawierają cząstkowe arytmetyczne obliczenia dokonywane w toku obliczania R Pearsona.

[3] D.D. Mari, S. Kotz, *Correlation and Dependence*, World Scientific Publishing, Londyn 2004, s. 29.

[4] Słownik Języka Polskiego, w: http://sjp.pwn.pl/haslo.php?id=2474043, dostęp: kwiecień 2012.

Tabela 28. Przykładowe dane do obliczenia współczynnika R Pearsona.

n	a	b	a*b	a2	b2
1	1	1	1	1	1
2	1	2	2	1	4
3	2	3	6	4	9
4	3	4	12	9	16
5	3	5	15	9	25
Suma	10	15	36	24	55
Średnia	2	3	-	-	-

Na podstawie tabeli i wyżej podanego wzoru obliczamy średnie dla obu zmiennych a i b:

$$A_{\text{średnia}} = \frac{1}{5} * 10 = 2$$

$$B_{\text{średnia}} = \frac{1}{5} * 15 = 3$$

Następnie obliczamy kowariancję tych zmiennych:

$$\text{kowariancja}\,(A, B) = \frac{(1-2)(1-3) + (1-2)(2-3) + (2-2)(3-3) + (3-2)(4-3) + (3-2)(5-3)}{5} = \frac{6}{5} = 1{,}2$$

Obliczamy odchylenie standardowe:

$$S_A = \sqrt{\frac{(1-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (3-2)^2}{5}} = \sqrt{\frac{4}{5}}$$

$$S_B = \sqrt{\frac{(1-3)^2 + (2-3)^3 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}} = \sqrt{\frac{10}{5}}$$

$$S_A * S_B = \sqrt{\frac{4}{5}} * \sqrt{\frac{10}{5}} = 1{,}264$$

Dokonujemy obliczenia współczynnika R:

$$R = \frac{1{,}2}{1{,}264} = 0{,}949$$

Interpretacja współczynnika korelacji R Pearsona. Współczynnik korelacji R Pearsona pozwala poznać nam **siłę** oraz **kształt** związku liniowego pomiędzy dwoma zmiennymi. Mierzy on stopień rozproszenia punktów wokół linii regresji, a więc stopień współzmienności dwóch zmiennych. Współczynnik korelacji R Pearsona obliczamy tylko pomiędzy zmiennymi ilościowymi (interwałowymi lub ilorazowymi); dla zmiennych mierzonych na poziomach jakościowych (jednej lub obu) stosujemy inne miary związku. Zmienne te powinny mieć także rozkład normalny (testowanie normalności rozkładu wyłożono między innymi w rozdziale dotyczącym regresji) lub zbliżony do normalnego. Nie jest to warunek konieczny, jednakże należy pamiętać o tym, że wartości skrajne mogą w sposób istotny zaburzyć wynik testu, ujawniając

korelację tam, gdzie jej nie ma[5]. Warunkiem przeprowadzenia korelacji jest przekonanie o tym, że wybrane zmienne mogą charakteryzować się współzmiennością. Przekonanie takie można powziąć na podstawie studium literatury przedmiotu, jak również przesłanką może być wnioskowanie lub przypuszczenia badacza. Wskazuje się, że istnieje pięcioprocentowa szansa, że wybrane zmienne są skorelowane ze sobą przez przypadek, tylko w wymiarze arytmetycznym, lecz nie logicznym i społecznym[6]. Należy podkreślić, że zależność arytmetyczna nie implikuje związku logicznego czy społecznego między cechami; innymi słowy zależności stwierdzonej testem R Pearsona nie wolno automatycznie uznawać za związek przyczynowo-skutkowy. Minimalna liczebność próby w teście R Pearsona wyznaczana jest przez działanie prawa wielkich liczb, czyli rozpoczyna się od $N \geq 30$. Warto mieć również na uwadze inny aspekt liczebności próby: nawet mała, niska korelacja może być istotna przy dużej liczbie przypadków, albowiem w rzeczywistości społecznej wszystkie cechy są ze sobą powiązane. Inaczej formułując: istotność w korelacji jest zależna od N: im większe N, tym większa istotność.

Współczynnik R Pearsona może przyjmować wartości od -1 (idealna korelacja ujemna) do +1 (idealna korelacja dodatnia). Pomiędzy członami korelacji, to jest zmienną zależną i zmienną niezależną mogą zachodzić trzy następujące rodzaje stosunków korelacyjnych:

1/ Stosunek korelacyjny zgodności (korelacja dodatnia, pozytywna) - zachodzi on wówczas, gdy wraz ze wzrostem wartości zmiennej niezależnej następuje również wzrost wartości zmiennej zależnej Współczynnik korelacji przyjmuje wówczas wartość większą od zera, lecz mniejszą lub równą jedności.

2/ Stosunek korelacyjny niezależności (brak korelacji) - zachodzi, gdy nie istnieje liniowy związek pomiędzy zmienną zależną, a zmienną niezależną. Współczynnik korelacji przyjmuje wówczas wartość równą zero.

3/ Stosunek korelacyjny przeciwieństwa (korelacja ujemna, negatywna) - zachodzi wtedy, gdy wraz ze wzrostem wartości zmiennej niezależnej następuje spadek wartości zmiennej zależnej. W tym przypadku wartość współczynnika jest mniejsza od zera, lecz nie mniejsza niż -1.

Sposób interpretacji poszczególnych wartości przybieranych przez współczynnik R szczegółowo ilustruje tabela 29. Warto nadmienić, że w praktyce badawczej reguły interpretacji współczynnika R Pearsona zmieniają się; do niedawna reguły te były bardzo rygorystyczne. Klasycznie uznaje się, że silna korelacja lokuje się powyżej 0,7. Jednak zmienne mierzące zjawiska społeczne, które ze sobą tak silnie korelują, traktowane są przez analityków bardzo podejrzliwie - uważa się, że w rzeczywistości nie tyle są ze sobą powiązane, lecz mierzą to samo zjawisko. Zjawiska społeczne korelują ze sobą na ogół na poziomie 0,6 lub 0,7 i obecnie te miary uważa się za korelacje wysokie. Warto również zauważyć, że dla współczynnika korelacji R Pearsona różne nauki przykładają różne miary. W naukach przyrodniczych, na przykład w fizyce współczynnik R Pearsona równy 0,9 oznacza korelację słabą[7].

Ważnym współczynnikiem przy interpretacji wyników korelacji jest **poziom istotności (p)** lub **współczynnik ufności (p)**, który informuje o prawdopodobieństwie, z jakim otrzymana korelacja może zdarzyć

[5] Pouczający, co do możliwych pułapek związanych ze stosowaniem korelacji, jest dostępny *online* program przygotowany przez Gary'ego McClellanda: http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/RangeRestrict.html, dostęp: kwiecień 2012.

[6] G.D. Garson, *Correlation*, Statistical Associates Publishing, Asheboro 2012, s. 6.

[7] J. Cohen, *Statistical power analysis for the behavioral sciences*, Routledge, Nowy Jork 1988.

się przypadkowo. Współczynnik ufności wyrażany jest przez liczbę z zakresu od zera do jedności. Im mniejsza jego wartość, tym mniejsze prawdopodobieństwo, że otrzymana korelacja zdarzyła się przypadkowo. W celu ustalenia, czy otrzymaną wartość współczynnika korelacji można uznać za istotną, porównujemy wartość poziomu istotności (p) z wartością poziomu ufności (α - alfa). **Poziom ufności (α)** jest to wartość przyjmowana przez badacza określająca maksymalne ryzyko błędu, jakie badacz jest skłonny zaakceptować. W naukach społecznych wartość poziomu ufności badacze na ogół ustalają się na poziomie 0,05 (5 proc.), rzadko - 0,01 (1 proc.)[8]. Jeśli przyjmiemy wartość poziomu ufności 0,05, wówczas możemy powiedzieć, że istnieje pięcioprocentowe prawdopodobieństwo ryzyka popełnienia błędu, tj. przyjęcia za prawdę wyniku korelacji, mimo że jest on błędny (korelacja przypadkowa).

Poziom istotności w korelacji R Pearsona może zostać obliczony na dwa sposoby: poprzez **test istotności jednostronnej** lub poprzez **test istotności dwustronnej**. Jeśli kierunek wpływu korelowanych zmiennych jest znany, wówczas możemy przeprowadzić test istotności jednostronnej. Natomiast test istotności dwustronnej przeprowadzamy wówczas, gdy kierunek wpływu nie jest znany. Jeśli uzyskane uprzednio wyniki analiz lub istniejące teorie stwierdzają o kierunku różnicy (a więc jesteśmy w stanie wskazać zmienną zależną i zmienną niezależną), wówczas mamy prawo postawić hipotezę jednostronną. Natomiast jeśli podjęty test R Pearsona ma charakter eksploracyjny i nie jesteśmy w stanie orzec o kierunku zależności, stawiamy hipotezę dwustronną. Należy zapamiętać, że test jednostronny łagodniej traktuje poziom istotności - potrzebuje on w porównaniu z testem dwustronnym mniejszej różnicy, by być istotnym statystycznie.

Ważną kwestią przy interpretowaniu współczynnika korelacji R Pearsona są zasady porównywania pomiędzy sobą poszczególnych wartości współczynnika korelacji. Współczynników nie można porównywać ze sobą bezpośrednio, gdy chcemy uzyskać odpowiedź na pytanie **o ile** jeden z nich jest silniejszy lub słabszy od drugiego. Bezpośrednie porównanie wartości liczbowych współczynnika R Pearsona informuje tylko o tym, że jedna wartość jest mniejsza, większa lub równa drugiej, nie informuje natomiast o ile. Nie jest więc tak, że różnica pomiędzy R=0,8 a R=0,6 jest taka jak pomiędzy R=0,4 a R=0,2. Nie można również mówić, że współczynnik R=0,6 jest dwukrotnie silniejszy od współczynnika R=0,3. Porównywanie współczynników korelacji, pozwalające na stwierdzenie, o ile jeden z nich jest większy lub mniejszy od drugiego, umożliwia **kwadrat współczynnika korelacji** (również określany mianem **współczynnika determinacji**, ang. *coefficient determination* lub potocznie R^2 (R kwadrat)). Współczynnik determinacji obliczamy przez podniesienie wartości R do kwadratu. Wartość współczynnika determinacji przyjmuje wielkości od zera do jedności i może być również wyrażana w procentach (R^2 * 100%). Wartość współczynnika determinacji informuje nas o tym, jaka część całkowitej zmienności zmiennej zależnej jest wyjaśniana przez zmienną niezależną. Tak przetworzone współczynniki korelacji można ze sobą bezpośrednio porównywać. Warto również wskazać, że istnieje miara niewyjaśnionej przez R^2 części zmienności obu zmiennych. Nosi nazwę **współczynnika alienacji** i wyrażana jest wzorem 1 - R^2.

[8] Por.: M. Nawojczyk, *Przewodnik po statystyce dla socjologów*, SPSS Polska, Kraków 2002, s. 152.

Tabela 29. Interpretacja współczynnika korelacji R Pearsona

Rodzaj stosunku korelacyjnego	Wartość współczynnika korelacji	Interpretacja współczynnika korelacji
Stosunek korelacyjny przeciwieństwa (korelacja negatywna, ujemna)	od -0,90 do -1,00	korelacja ujemna b. wysoka (zależność b. pewna)
	od -0,70 do -0,90	korelacja ujemna wysoka (zależność znaczna)
	od -0,40 do -0,70	korelacja ujemna umiarkowana (zależność istotna)
	od -0,20 do -0,40	korelacja ujemna niska (zależność wyraźna, lecz mała)
	do -0,20	korelacja ujemna słaba (zależność prawie nic nie znacząca)
Stosunek korelacyjny niezależności (brak korelacji)	0	brak liniowego związku pomiędzy zmiennymi
Stosunek korelacyjny zgodności (korelacja pozytywna, dodatnia)	do 0,20	korelacja dodatnia słaba (zależność prawie nic nie znacząca)
	od 0,20 do 0,40	korelacja dodatnia niska (zależność wyraźna, lecz mała)
	od 0,40 do 0,70	korelacja dodatnia umiarkowana (zależność istotna)
	od 0,70 do 0,90	korelacja dodatnia wysoka (zależność znaczna)
	od 0,90 do 1,00	korelacja dodatnia b. wysoka (zależność b. pewna)

Obliczanie i interpretacja współczynnika korelacji R Pearsona w programie PSPP. Przykład zastosowania korelacji z wykorzystaniem danych ze zbioru Polskiego Generalnego Studium Wyborczego (2007) można przedstawić jedynie z zastrzeżeniami. W zbiorze tym figurują przede wszystkim dane mierzone na poziomie porządkowym i nominalnym. Część tych zmiennych jakościowych została pomierzona z użyciem skali Rensisa A. Likerta. A zatem - jak wskazywano wcześniej - można przyjąć, że skala ta pozwala na ich zamianę na ilościowe. W celu dokonania egzemplifikacji wybrano ze zbioru PGSW dwie następujące zmienne: zainteresowanie polityką (zmienna p48) oraz miesięczny osobisty dochód netto respondenta w ostatnim miesiącu. Zainteresowanie polityką zmierzono na skali porządkowej, którą na mocy wyłożonych we wcześniejszych rozdziałach zasad można przekształcić na zmienną ilościową. Ostatecznie skala zmiennej dotyczącej zainteresowania przestawiała się następująco (odwrócono poprzez procedurę rekodowania pierwotną skalę tej zmiennej, aby uzyskać tożsamość obu zmiennych - im większy dochód, tym większe liczby oraz im większe zainteresowanie polityką również tym większe liczby): 1 - *żadne, praktycznie mnie to nie interesuje*, 2 - *nikłe, niewielkie, często umykają mojej uwadze nawet ważne wydarzenia polityczne*, 3 - *średnie, śledzę jedynie główne wydarzenia polityczne*, 4 - *duże, dość uważnie śledzę to, co się dzieje w polityce,* 5 - *bardzo duże, uważnie (szczegółowo) śledzę prawie wszystkie wydarzenia polityczne.* Przypomnijmy, że warunkiem badania korelacji między zmiennymi jest teoretyczne uzasadnienie ich zależności lub co najmniej silne przekonanie badacza o ich współzmienności. W tym przypadku mamy do czynienia z pierwszą sytuacją. W nauce o polityce funkcjonuje dobrze ugruntowana hipoteza o dodatniej korelacji między natężeniem partycypacji politycznej a poziomem zasobów społecznych i ekonomicznych obywateli. W amerykańskiej nauce o polityce zjawisko to nazywane jest hipotezą statusu socjoekonomicznego jednostki (SES - *socioeconomic status* lub SERL - *socioeconomic*

resource status)[9]. Herbert Tingsten sformułował tak zwane prawo dyspersji odnoszące się do korelacji tych dwóch zbiorów zmiennych. Głosi ono, że konsekwencją korelacji jest powstawanie istotnych nierówności poszczególnych grup społecznych w dostępie do władzy - lepiej wykształceni i bogaci mają dzięki swojej aktywności wpływ większy niż grupy ekonomicznie i społecznie upośledzone. Mamy zatem prawo oczekiwać, że pomiędzy wzrostem zarobków jednostki a jej elementarną aktywnością polityczną, a więc zainteresowaniem polityką, istnieje pewna dodatnia współzmienność.

Jak wcześniej wskazano przygotowując się do wykonania korelacji R Pearsona w pierwszej kolejności eliminujemy dane niepełne, to znaczy nieposiadające pary (stąd liczebność zaledwie N=1213), a także liczby skrajne (imputacja danych). Następnie sprawdzamy normalność rozkładu obu zmiennych oraz sporządzamy wykres rozrzutu, by przekonać się, czy mamy do czynienia z liniowym związkiem pomiędzy zmiennymi. Czynności te oraz sposób ich wykonania zostały drobiazgowo opisane rozdziale dotyczącym regresji, więc nie są tu ponownie przytaczane. Po sporządzeniu wykresu rozrzutu wybieramy z menu tekstowego *Analyze* ⇨ *Bivariate Correlation.*

Zmienne p48 oraz m21 (kolejność ich umieszczenia nie ma wpływu na wynik korelacji) z pola z lewej strony przenosimy do pola po prawej. W *Test of Significance* oznaczamy *One-tailed*, bowiem na podstawie naszej wiedzy teoretycznej możemy zakładać kierunek wpływu: wysokość dochodu determinuje stopień zainteresowania polityką, a więc im wyższy dochód tym większe nią zainteresowanie. Następnie oznaczamy *Flag significant correlations* (oznacz korelacje istotne), która to funkcja powoduje opatrzenie za pomocą gwiazdek korelacji istotnych (na jednym z dwóch poziomów ufności 0,01 lub 0,05).

Correlations

		Zainteresowanie polityką.	*Osobisty dochód netto w ostatnim miesiącu*
Zainteresowanie polityką.	*Pearson Correlation*	1,00	,16
	Sig. (1-tailed)		,00
	N	1213	1213
Osobisty dochód netto w ostatnim miesiącu	*Pearson Correlation*	,16	1,00
	Sig. (1-tailed)	,00	
	N	1213	1213

[9] W amerykańskiej nauce o polityce hipoteza SES nazywana jest wręcz żelaznym prawem: D. Rucht, *Rosnące znaczenie polityki protestu*, w: *Zachowania polityczne*, t. II, R.J. Dalton, H.-D. Klingemann (red.), Wydawnictwo Naukowe PWN, s. 359. Rozważania nad tym zagadnieniem mają bogatą tradycję w badaniach nad partycypacją polityczną. Na uwagę zasługuje praca Lestera W. Milbratha, który pierwsze swoje dzieło poświęcił próbom zidentyfikowania czynników skłaniających do zaangażowania obywateli w działalność polityczną (L.W. Milbrath, *Political Participation. How and Why Do People Get Involved in Politics?*, Rand McNally & Company, Chicago 1965, s. 39-141). Warto wymienić również prace Sidney'a Verby i jego współpracowników, które w istotnym zakresie przyczyniły się zoperacjonalizowania hipotezy SES: S. Verba, N.H. Nie, *Participation in America. Political Democracy and Social Equality*, The University of Chicago Press, Londyn 1972; S. Verba, K.L. Scholzman, H.E. Brady, *Voice and Equality. Civic Voluntarism in American Politics*, Harvard University Press, Cambridge, Londyn 1995.

Z tabeli stanowiącej efekt obliczeń odczytujemy w celu interpretacji wartość *Sig.* (*1-tailed*). Zwróćmy uwagę, że tożsama wartość znajduje się w dwóch wierszach tej tabeli, nie ma znaczenia skąd ją odczytamy. Wartość ta mówi o istotności testu korelacji. Im jest ona niższa, tym lepiej. Wykonany test statystyczny możemy uznać za istotny jeśli wartość ta jest niższa niż $p \leq 0{,}05$. Oznacza ona, że mamy 5 proc. szansę na popełnienie błędu I rodzaju posługując się otrzymanym wynikiem testu statystycznego. Błąd ten popełniany jest wówczas, gdy odrzucamy hipotezę zerową, która w istocie jest prawdziwa, a więc na podstawie wyników testu statystycznego twierdzimy, że jakiś fakt jest statystycznie istotny, natomiast w istocie jest on dziełem przypadku. Jeszcze lepiej, jeśli ta wartość jest niższa niż 0,01 (tak właśnie jest w przypadku wykonanego testu). Ryzyko popełnienia błędu I rodzaju wynosi zaledwie 1 proc. (Zwróćmy uwagę, że program PSPP nie zaokrągla tej wartości, podając po prostu wartość 0,00. Analityk w raporcie zapisuje jednak tę wartość następująco: $p \leq 0{,}01$). Po stwierdzeniu, że wynik testu statystycznego jest istotny odczytujemy wartość współczynnika korelacji z wiersza *Pearson Correlation*. Odczytujemy wartość na przecięciu dwóch różnych skorelowanych ze sobą zmiennych, czyli w przypadku niniejszej egzemplifikacji na przecięciu „Zainteresowanie polityką" i „Osobisty dochód netto w ostatnim miesiącu". Wynosi ona 0,16, co w raporcie zapisujemy jako R=0,16. (Nie interpretujemy natomiast tak zwanej autokorelacji, czyli skorelowania zmiennej z nią samą - wartość ta zawsze wynosi 1). Współczynnik ten interpretujemy na podstawie informacji zawartych w tabeli 29. Mamy zatem do czynienia ze stosunkiem korelacyjnym zgodności, a więc jest to korelacja dodatnia słaba (zależność prawie nic nie znacząca). Jeśli współczynnik R podniesiemy do kwadratu, wówczas uzyskamy wartość następującą: R^2=0,03. Oznacza to, że zaledwie 3 proc. wariancji zainteresowania polityką jest wyjaśniana przez wysokość dochodu respondenta.

Obliczenie współczynnika R Pearsona w Edytorze składni wymaga wpisania następujących komend:

```
CORRELATION
/VARIABLES = m21 p48
/PRINT = ONETAIL SIG.
```

ONETAIL oznacza test istotności jednostronnej, a TWOTAIL - dwustronnej. SIG wymusza opatrzenie gwiazdką istotnych wyników korelacji, a NOSIG wywołuje brak takiego oznaczenia. Jeśli po znaku równości, w linii, gdzie znajduje się komenda VARIABLES, umieścimy więcej niż dwie zmienne powstanie macierz korelacji krzyżująca wszystkie podane zmienne ze wszystkimi.

13.1.2. Stosunek korelacyjny eta (η)

Stosunek korelacyjny eta (współczynnik korelacji eta) jest miarą siły związku pomiędzy dwoma zmiennymi. Stosuje się go w dwóch modelowych przypadkach: po pierwsze, gdy jedna zmienna mierzona jest na poziomie ilościowym (ilorazowym lub interwałowym), a druga - na poziomie jakościowym (nominalnym lub porządkowym), a po drugie, gdy zachodzi podejrzenie, że związek pomiędzy dwiema zmiennymi ilościowymi ma charakter krzywoliniowy. W tych dwóch przypadkach miara ta zastępuje współczynnik korelacji R Pearsona. Ponadto jest on cząstkową składową innych miar, na przykład jednoczynnikowej analizy wariancji (ANOVA). Współczynnik eta jest miarą relacji, która mówi nam, jaka część wariancji zmiennej niezależnej jest wyjaśniana przez zmienną zależną. Wartość tę obliczamy następująco: całkowitą sumę kwadratów (patrz rozdział dotyczący regresji liniowej) dzieli się na dwie części - międzygrupową sumę kwadratów i wewnątrzgrupową sumę kwadratów. Międzygrupowa suma kwadratów to ta część wariancji, która jest wyjaśniana przez zmienną niezależną. Wewnątrzgrupowa suma kwadratów jest

z kolei tą częścią wariancji zmiennej zależnej, która nie jest wyjaśniana przez zmienną niezależną - odpowiadają za nią inne czynniki. Stosunek korelacyjny eta wyrażany jest następującym wzorem:

$$\eta_{xy} = \sqrt{\frac{\text{międzygrupowa suma kwadratów}}{\text{całkowita suma kwadratów}}}$$

Współczynnik obliczony według tego wzoru przyjmuje wartości od 0 do 1 (nigdy nie będzie on ujemny). Im wyższa wartość współczynnika eta, tym silniejszy związek pomiędzy zmiennymi. Stosunek korelacyjny eta jest współczynnikiem standaryzowanym w tym sensie, że można porównywać wyniki - po podniesieniu do kwadratu (η^2) mówi on o stopniu wariancji wyjaśnianej i zawiera się w przedziale od 0 proc. do 100 proc. Interpretujemy go zatem analogicznie jak współczynnik R lub R^2 Pearsona, jednak z pewnym zastrzeżeniem: wartość współczynnika eta będzie większa niż współczynnika R Pearsona w przypadku zależności nieliniowych, a równa w przypadku liniowej zależności pomiędzy zmiennymi. Jeśli zastosujemy łącznie obie miary w przypadku związku krzywoliniowego między zmiennymi, to ich różnica (eta - R) może stanowić pomocną interpretacyjnie miarę nieliniowości związku. Współczynnik eta jest testem kierunkowym. Oznacza to, że w zależności od tego, którą ze zmiennych wskażemy jako zależną, a którą jako niezależną uzyskamy odmienny wynik. W przypadku, gdy jedna ze zmiennych jest jakościowa jesteśmy poddani szczególnym rygorom. W takiej sytuacji zmienną niezależną musi być zmienna jakościowa. Ponadto istotne jest, aby zmienna ta miała jak największą liczbę kategorii. Warto również zwrócić uwagę na fakt, że test nie jest wrażliwy na kolejność kategorii zmiennej nominalnej. Ponadto dla obu typów zmiennych postuluje się, aby jednostki analizy były wystarczająco liczne. Warto podzielić się pewną wskazówką analityczną dla sytuacji pomiaru siły związku pomiędzy zmiennymi ilościowymi: sporządziwszy wykres rozrzutu oceniamy, czy punkty układają się liniowo czy nie. W tej drugiej sytuacji decydujemy się na użycie współczynnika korelacyjnego eta, a współczynnik korelacji R czynimy miarą pomocniczą obrazującą nieliniowość związku.

W programie PSPP współczynnik ten obliczmy wybierając z menu następującą sekwencję: *Analyze* ⇨ *Crosstabs* ⇨ *Statistics* ⇨ *Eta.* Z kolei w trybie Edytora składni należy wpisać następujące polecenia:

```
CROSSTABS
/TABLES= p48 BY m21
/FORMAT=AVALUE TABLES PIVOT
/STATISTICS=ETA CORR
/CELLS=COUNT ROW COLUMN TOTAL.
```

Przeanalizujmy raz jeszcze związek pomiędzy zainteresowaniem polityką a dochodem osobistym dochodem respondenta w ostatnim miesiącu.

Directional measures.

Category	Statistic Type	Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Nominal by Interval Eta	Zainteresowanie polityką. Dependent	,48			
	Osobisty dochód netto w ostatnim miesiącu Dependent	,18			

Zastosowanie miary eta umożliwi stwierdzenie, czy związek ten ma charakter krzywoliniowy. Zwróćmy uwagę, że program PSPP podaje dwie różne miary: dla zainteresowania polityką (0,48) oraz dla osobistego dochodu netto (0,18). Z tabeli odczytujemy wartość przyporządkowaną tej zmiennej, którą przyjęliśmy jako zmienną zależną, a więc - 0,18. Przypomnijmy, że współczynnik R Pearsona wyniósł zaledwie 0,16 co wyjaśniło 3 proc. całości wariancji. Wartość eta jest niewiele większa (w granicach błędu), a więc można wnioskować, że związek pomiędzy dwoma analizowanymi zmiennymi ma charakter prostoliniowy.

13.1.3. Współczynnik zgodności kappa (κ) Cohena

Współczynnik kappa służy do porównywania stopnia zgodności ocen obserwatorów wobec określonego obiektu lub obiektów. Może również posłużyć jako miara jednolitości ocen jednego obserwatora dokonującego pomiaru dwukrotnie w odstępie czasowym. Współczynnik zgodności przydatny jest wszędzie tam, gdzie oceniamy standaryzowane opinie (ekspertów, jury, specjalistów, nauczycieli i egzaminatorów, sędziów kompetentnych, a także opinii zwykłych obywateli). Efektem zastosowania współczynnika zgodności Cohena jest ilościowa miara stopnia zgody pomiędzy oceniającymi pozwalająca oszacować regularność i zbieżność wydawanych ocen. Ma on za zadanie subiektywne opinie oceniających uczynić intersubiektywnymi. Współczynnik został opracowany przez Jacoba Cohena na podstawie innych, niezadowalających jednak i wadliwych współczynników (między innymi współczynnika Williama Scotta)[10].

Współczynnik kappa przyjmuje wartości od -1 do 1. W praktyce badawczej posługujemy się tylko zakresem współczynnika zawierającym się od 0 do 1. Im współczynnik bliższy jedności, tym zgodność jest większa. Współczynnik o wartości 1 oznacza zgodność idealną. Współczynnik o wartości 0 oznacza zgodność na poziomie takim samym, jaki powstałby dla losowego rozłożenia danych w tabeli kontyngencji. Z kolei współczynnik poniżej 0 oznacza, że zgodność jest mniejsza niż powstała dla losowego rozłożenia danych w tabeli. Kappa jest dostępna tylko dla tabel kwadratowych (tabele dwóch zmiennych, mających taką samą liczbę kategorii). Do sprawdzania istotności współczynnika kappa Cohena służy test Z. Współczynnik ten jest wartością kappa z próby. W celu obliczenia współczynnika kappa z populacji należy odczytać wartość błędu standardowego kappa. Jakościowa interpretacja współczynnika jest rozbieżna u różnych badaczy[11]. Na przykład J.R. Landis i G.G. Koch przedstawiają następującą propozycję jego interpretacji[12]:

< 0,00 brak zgodności,
0,00 — 0,20 - zgodność słaba,
0,21 — 0,40 - zgodność średnia,
0,41 — 0,60 - zgodność umiarkowana,
0,61 — 0,80 - zgodność pokaźna,
0,81 — 1,00 - zgodność prawie perfekcyjna.

Z kolei amerykański biostatystyk Joseph L. Fleiss zaproponował bardziej restrykcyjną propozycję interpretacji:

< 0,40 - słaba zgodność,
0,40 — 0,74 - umiarkowana lub dobra zgodność,

[10] J. Cohen, *A Coefficient of Agreement For Nominal Scales*, „Educational and Psychological Measurement", 1960, 10 (3746), s. 37-46.

[11] Doskonałe, formalne opracowanie obliczania i interpretacji rozmaitych współczynników zgodności w tym współczynnika kappa Cohena przedstawia Joanna Jarosz-Nowak: J. Jarosz-Nowak, *Modele oceny stopnia zgody pomiędzy dwoma ekspertami z wykorzystaniem współczynników kappa*, „Matematyka stosowana", 2007, 8, s. 126-154, w: www.matstos.pjwstk.edu.pl/no8/no8_jarosz-nowak.pdf, dostęp: czerwiec 2012.

[12] J.R. Landis, G.G. Koch, *The measurement of observer agreement for categorical data*, „Biometrics", 1977, 33 (1), s. 159-174.

0,75 — 1,00 - perfekcyjna zgodność[13].

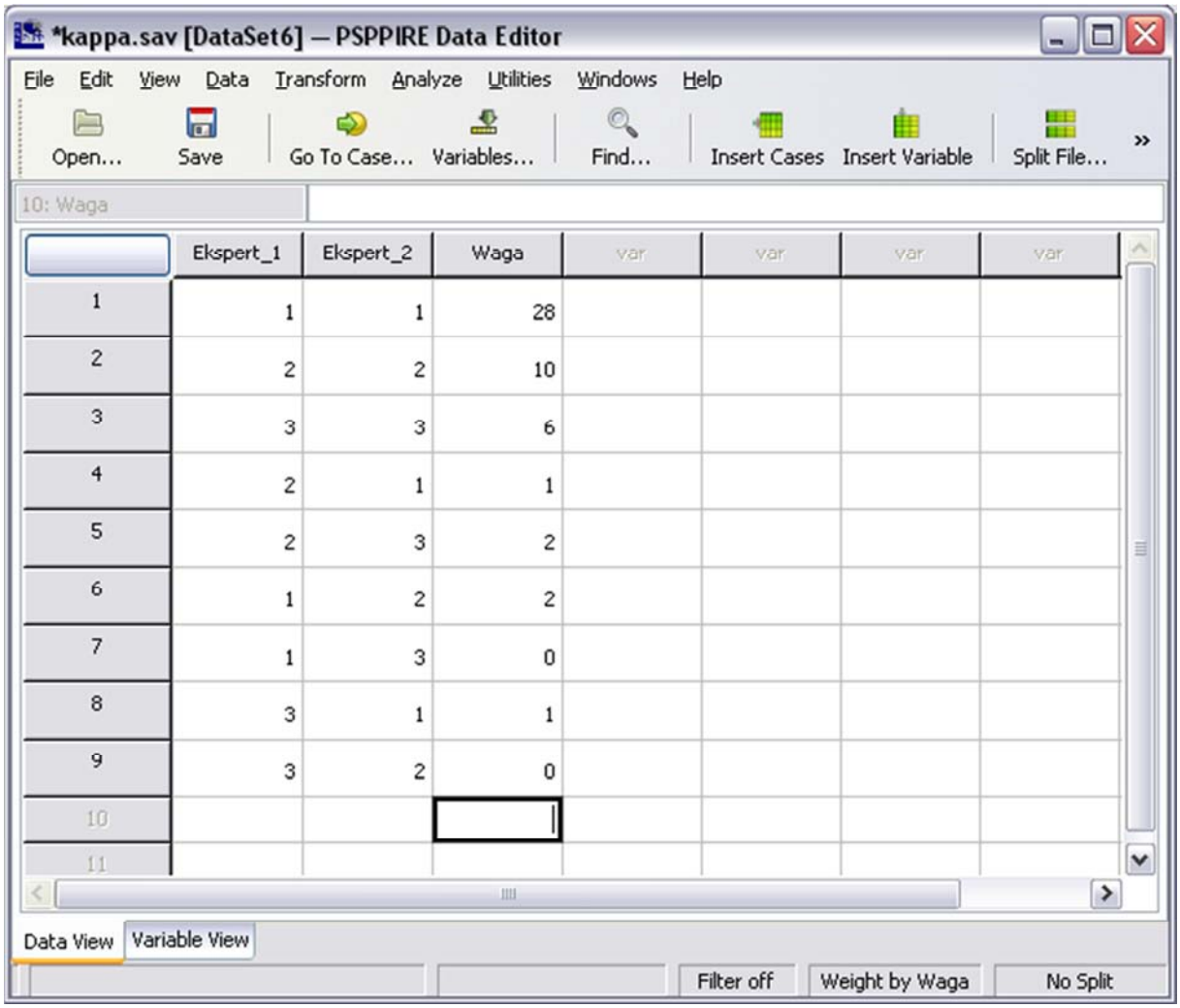
W celu wyłożenia technicznego aspektu wykorzystania tej miary przedstawiono poniżej przykładową sytuację, w której współczynnik kappa może zostać z powodzeniem wykorzystany. Dwóch ekspertów klasyfikowało pewną ilość przedłożonych im elementów systemu politycznego (dokładnie 50 elemntów), używając jednej z trzech kategorii: 1/ cecha systemu politycznego jest charakterystyczna dla systemów demokratycznych, 2/ cecha jest charakterystyczna dla systemów autorytarnych, 3/ cecha jest charakterystyczna dla systemów totalitarnych. Dokonaną przez nich klasyfikację przedstawia tabela 30.

Tabela 30. Tabela kontyngencji 3x3 liczbowo określająca stopień zgodności klasyfikacji cech systemu politycznego przez ekspertów

		Ekspert 2		
		Sklasyfikował jako demokratyczne	**Sklasyfikował jako autorytarne**	**Sklasyfikował jako totalitarne**
Ekspert 1	**Sklasyfikował jako demokratyczne**	28	2	0
	Sklasyfikował jako autorytarne	1	10	2
	Sklasyfikował jako totalitarne	1	0	6

W programie PSPP dane z powyższej tabeli powinny zostać zaimplementowane następująco: zmiennymi powinny stać się „pary wypowiedzi" ekspertów oraz zmienna wskazująca na częstotliwość tych wypowiedzi. Sposób przełożenia danych do programu ilustruje poniższy zrzut ekranowy.

[13] Interpretację zaproponowaną przez J.L. Fleissa oraz inne przedstawiają: D.V. Cicchetti, F. Volkmar, S.S. Sparrow, D. Cohen, *Assessing the Reliability of Clinical Scales When the Data Have Both Nominal and Ordinal Features: Proposed guidelines for neuropsychological assessments*, „Journal of Clinical and Experimental Neuropsychology", 1992, 14 (5), s. 673-686.



Zwróćmy uwagę na fakt, że trzecia zmienna stała się wagą dla pozostałych (*Weight by waga*). Współczynnik kappa dla powyższych danych wynosi 0,784. Oznacza to, iż eksperci byli w dużym stopniu zgodni w swojej opinii i w bardzo zbliżony sposób klasyfikowali elementy systemu politycznego.

Współczynnik kappa obliczamy w programie PSPP wybierając z menu tekstowego *Analyze* ⇨ *Crosstabs* ⇨ *Statistics* ⇨ *Kappa.* W wersji programu 0.7.9 współczynnik ten, pomimo że figuruje w menu, nie został jeszcze zaimplementowany. W celu obliczenia tego współczynnika można wykorzystać jeden z prostych w użyciu dostępnych w sieci kalkulatorów. Jeden z takich przyjaznych programów, prowadzący użytkownika krok po kroku, znajdziemy pod adresem: http://justusrandolph.net/kappa/[14].

13.2. Miary związku dla zmiennych jakościowych

Niniejszy podrozdział wprowadza w miary nieparametryczne dedykowane dla zmiennych jakościowych - nominalnych i porządkowych. Używane mogą być one także, gdy jedna zmienna mierzona jest na poziomie silnym (ilościowym), a druga - na poziomie słabym (jakościowym), a także w sytuacji, gdy zmienne ilościowe nie spełniają żądanych właściwości rozkładu (rozkładu normalnego) lub też wystarczającej liczby jednostek analizy.

[14] J.J. Randolph, *Online Kappa Calculator*. 2008, w: http://justus.randolph.name/kappa, dostęp: kwiecień 2012.

13.2.1. Miary związku dla zmiennych porządkowych

Podstawowym, zalecanym do opanowania jest współczynnik korelacji rangowej rho (ρ) Spearmana. Bardzo często używane są także gamma (γ) Leo A. Goodmana i Williama H. Kruskala oraz tau-B (τ-B) i tau-C (τ-C) Maurice'a G. Kendalla. Współczynnik d Roberta H. Somersa - choć pod względem jakości generowanych wyników nie ustępuje wymienionym - jest obecnie rzadziej stosowany.

13.2.1.1. Współczynnik korelacji rangowej rho (ρ) Spearmana

Współczynnik korelacji rang rho Spearmana został opisany przez Charlesa Spearmana w 1904 roku[15]. Właściwie, jak pisze William S. Gosset, C. Spearman rozwinął ogłoszone kilka lat wcześniej zalążkowe i niedopracowane pomysły dwóch francuskich uczonych - Alfreda Bineta i Victora Henriego[16]. Współczynnik ten mierzy siłę i kierunek korelacji dwóch zmiennych. Służy on do pomiaru siły zależności dwóch zmiennych mierzonych na poziomie porządkowym. Wykorzystuje się go również w przypadku, gdy pomiar został dokonany na poziomie interwałowym lub ilorazowym, lecz badana zbiorowość jest nieliczna (mniejsza od 30, jednak równa lub większa niż 5[17]). Wskazaniem do zastosowania tej miary są sytuacje następujące: po pierwsze, co najmniej jedna z dwóch zmiennych ilościowych, które chcemy poddać korelacji nie spełnia normalności rozkładu, po drugie, co najmniej jedna ze zmiennych mierzona jest na poziomie porządkowym (druga zmienna może być zmienną ilościową lub - jak pierwsza - porządkową). Współczynnik ten oznaczany jest grecką literą ρ (rho) lub małą literą r z literą s w indeksie dolnym (r_s).

Jest to współczynnik korelacji rangowej, a więc obie zmienne muszą być uporządkowane. Konieczne jest właściwe ich przetworzenie, to znaczy porangowanie. Rangowanie, w uproszczeniu, polega na przeliczeniu poszczególnych wartości uprzednio uporządkowanych zmiennych na kolejno następujące po sobie rosnące lub malejące wartości (rangi). Dopiero po takiej „standaryzacji" możliwe jest obliczenie współczynnika korelacji. Kierunek rang powinien być zgodny dla obydwu mierzonych zmiennych (a więc rosnący lub malejący), dla wygody interpretacji wyniku. Ponadto stymulanta (zmienna, o której sądzimy, że jej wzrost powoduje wzrost drugiej) powinna posiadać rangi rosnące, a destymulanta - malejące.

Współczynnik korelacji rangowej rho Spearmana obliczany jest według rozmaitych wzorów, różniących się od oryginalnej propozycji C. Spearmana. Najczęściej w dydaktyce podawany jest następujący uproszczony wzór:

$$\rho = 1 - \left[6 * \frac{(X_1 - Y_1)^2 + \cdots + (X_n - Y_n)^2}{n(n^2 - 1)}\right]$$

gdzie X_n - Y_n to różnice między parami rang dla każdej jednostki analizy, n to liczba jednostek analizy czyli próba.

[15] C. Spearman, *The Proof and Measurement of Association between Two Things*, „The American Journal of Psychology", 1904, 15 (1), s. 72-101.

[16] Student, *An Experimental Determination of the Probable Error of Dr Spearman's Correlation Coefficients*, „Biometrika", 1921, 13, 2-3, s. 263-282.

[17] Jeśli liczba jednostek analizy jest niższa niż pięć wówczas współczynnik ten można obliczać, aczkolwiek „jest on bardzo niepewny". G. Klaus, H. Ebner, *Podstawy statystyki dla psychologów i socjologów*, Państwowe Zakłady Wydawnictw Szkolnych, Warszawa 1972, s. 131.

Wzór ten jest obecnie rzadko stosowany, jednak podano go ze względu na fakt, że jest najprostszą formą, w jakiej można przedstawić ideę korelacji rangowej; może być on używany wówczas, jeśli nie występują pomiędzy wartościami zmiennych tak zwane rangi połączone (*ties*), to znaczy takie, które charakteryzują się dokładnie takimi samymi wartościami zmiennej[18].

Wykorzystajmy jednak podany wzór do obliczenia współczynnika rho Spearmana w następującym przykładzie. Rozważmy współzależność dwóch zmiennych: wielkości miejsca zamieszkania oraz udziału w lokalnym życiu politycznym. Pierwsza zmienna mogłaby być co prawda zmienną ilościową, jednakże zawierałaby zbyt duży zakres zmiennych (od kilkudziesięciu - reprezentujących małe wsie - aż do blisko dwóch milionów reprezentujących Warszawę). W takiej sytuacji najlepszym rozwiązaniem jest obniżenie poziomu pomiaru zmiennej do poziomu porządkowego. W przypadku drugiej zmiennej - przypuśćmy - mamy do czynienia ze zagregowanym indeksem licznych aktywności: od oddawania głosu w wyborach samorządowych aż do działania w stowarzyszeniach. Pierwszej zmiennej nadano zakres następujący: 1 - wieś, 2 - miasto do 20 tysięcy mieszkańców, 3 - miasto powyżej 20 do 50 tysięcy mieszkańców, 4 - miasto powyżej 50 tysięcy mieszkańców do 200 tysięcy mieszkańców, 5 - miasto powyżej 200 do 500 tysięcy mieszkańców, 6 - miasto powyżej 500 tysięcy mieszkańców. Druga zmienna została uporządkowana wedle schematu, w którym 1 oznacza najniższy poziom aktywności, a 6 oznacza aktywność o najwyższym natężeniu. Wartości przyjmowane przez te zmienne, przedstawiono w tabeli 31.

Tabela 31. Indeks poziomu aktywności ze względu na wielkość miejsca zamieszkania.

Wielkość miejsca zamieszkania (zmienna Xn)	Poziom aktywności obywateli (zmienna Yn)	Xn - Yn	(Xn - Yn)2
1 - wieś	5	-4	16
2 - miasto do 20 tysięcy mieszkańców	6	-4	16
3 - miasto powyżej 20 do 50 tysięcy mieszkańców	4	-1	1
4 - miasto powyżej 50 do 200 tysięcy mieszkańców	3	1	1
5 - miasto powyżej 200 do 500 tysięcy mieszkańców	1	4	16
6 - miasto powyżej 500 tysięcy mieszkańców	2	4	16
Suma (Xn - Yn)2		66	

Obliczone w tabeli cząstkowe wartości podstawiamy do wzoru:

$$\rho = 1 - \left[\frac{6 * 66}{6(6^2 - 1)}\right] = 1 - \frac{396}{210} = 1 - 1{,}886 = -0{,}886$$

[18] W sytuacji, gdy takie rangi tożsame wystąpią, liczy się dla nich średnie arytmetyczne. Jeśli na przykład w zbiorze znajdą się dwie jednostki o tej samej wartości, wówczas tym dwóm wartościom nadać kolejno następujące po sobie rangi, a następnie dodać do siebie i podzielić przez dwa obliczając ich średnią arytmetyczną.

Współczynnik korelacji rang zawiera się w zakresie od -1 do +1 i i**nterpretujemy go identycznie jak współczynnik korelacji R Pearsona**. Tożsame są przedziały przyjmowane przez obydwa współczynniki.

Współczynnik ten powstał po to, by zaradzić istotnej słabości współczynnika korelacji R Pearsona - dużej wrażliwości na obserwacje skrajne. Jego zastosowanie pozwala uniknąć problemu tzw. obserwacji odstających, to jest takich, które do danego modelu nie pasują, zaburzają go i obniżają wartość współczynnika korelacji. Należy podkreślić, że nawet jedna obserwacja może drastycznie zmienić wartość współczynnika korelacji. Rangowanie sprawia, że wartości odstające przestają zaburzać wynik korelacji. Warto zwrócić uwagę, że miara ta (podobnie zresztą jak R Pearsona) nie jest w stanie wykryć wszystkich możliwych typów zależności, na przykład nie wykrywa zależności okresowych noszących miano sezonowości.

W programie PSPP współczynnik ten obliczamy, wybierając w menu tekstowym *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs* ⇨ *Statistics* ⇨ *Corr.*

13.2.1.2. Współczynnik korelacji rangowej d Somersa

Współczynnik d został opracowany przez socjologa Roberta H. Somersa jako rozwinięcie współczynnika tau-B Kendalla. Współczynnik Somersa powstał, by modyfikować przeszacowanie współczynników występujące w szczególności w małych tabelach (gdy korelujemy zmienne o niewielkim zakresie). Współczynnik ten występuje w dwóch odmianach: symetrycznej i asymetrycznej. Należy jednak pamiętać, że współczynnik korelacji rangowej d Somersa pierwotnie został zaprojektowany jako miara asymetryczna.

Współczynnik korelacji rangowej d Somersa mierzy zarówno siłę, jak i kierunek związku dwóch zmiennych porządkowych. Przyjmuje on wartości od -1 do +1. Im liczby bliższe zera tym słabszy związek, a im bliższe jedności - tym związek pomiędzy zmiennymi jest silniejszy. Reguły interpretacji tego współczynnika są następujące: wartości do 0,1 oznacza słabą zależność, powyżej 0,1 do 0,2 - umiarkowaną zależność, o zależności umiarkowanie silnej mówimy, gdy wartość współczynnika przekroczy 0,2, lecz jest nie większa niż 0,3, a wartości powyżej 0,3 interpretujemy zależność jako silną. Wartości dodatnie bądź ujemne informują z kolei o kierunku zmienności. Jeżeli wartość współczynnika jest dodatnia to znaczy, że wraz ze wzrostem wartości zmiennej niezależnej rośnie wartość zmiennej zależnej. Przeciwną sytuację zaobserwujemy w przypadku wartości ujemnych - gdy wzrasta wartość zmiennej niezależnej, wartość zmiennej zależnej spada.

Współczynnik obliczamy wedle następującego wzoru:

$$d = \frac{Z - N}{Z + N + W}$$

gdzie Z oznacza liczbę zgodnych par zmiennych, N - liczbę niezgodnych par zmiennych, a W - liczbę par tworzących wiązania.

W programie PSPP współczynnik d Somersa uzyskujemy, wybierając w menu tekstowym *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs* ⇨ *Statistics* ⇨ *D.* Poddajemy interpretacji dwie wartości w wymienionej kolejności: poziom przybliżonej istotności oraz wielkość współczynnika d. Współczynnik jest

statystycznie istotny, jeśli jego wartość jest równa lub niższa niż 0,05. Wartość ta podawana jest w ostatniej kolumnie tabeli nazywanej *Approximate Significant.* Z kolei współczynnik korelacji rangowej odczytujemy z kolumny oznaczonej jako *Value.* Jeśli nie wyznaczyliśmy zmiennej zależnej i zmiennej niezależnej, wówczas odczytujemy wartość współczynnika d Somersa z wiersza oznaczonego *Symmetric.* Jeśli natomiast możemy wskazać zmienną wyjaśnianą i zmienną wyjaśniającą, wtedy odczytujemy wartość z jednego z dwóch kolejnych wierszy (a więc wybieramy miarę asymetryczną), w zależności od tego, którą zmienną uznaliśmy za zmienną zależną. Nazwa zmiennej zależnej jest wskazana w danym wierszu i oznaczona jako *Dependent.*

13.2.1.3. Współczynnik korelacji rangowej gamma (γ) Goodmana i Kruskala

Statystyka ta została zaproponowana w serii artykułów autorstwa Leo A. Goodmana i Williama H. Kruskala publikowanych w latach 1954-1972 w „Journal of the American Statistical Association". Jest to miara przeznaczona dla zmiennych mierzonych na poziomie porządkowym, dla dowolnej wielkości tabeli.

Jest to współczynnik symetryczny, to znaczy taki, gdzie nie zakładamy, która zmienna jest zmienną zależną, a która - niezależną. Współczynnik ten przyjmuje wartości od -1 do +1. Im wartość bliższa jedności, tym silniejsza zależność pomiędzy badanymi zmiennymi. Znak współczynnika ma również znaczenie. Dodatnia wartość współczynnika oznacza, że mamy do czynienia ze zgodnym uporządkowaniem par wartości obu zmiennych (wzrost wartości jednej zmiennej implikuje wzrost wartości drugiej). Z kolei ujemna wartość współczynnika oznacza, że wzrost wartości jednej zmiennej oznacza obniżanie się wartości drugiej. Należy podkreślić, że wartość współczynnika równa zero nie musi koniecznie oznaczać niezależności (chyba, że mamy do czynienia z tabelami 2x2). Obliczając współczynnik korelacji rangowej gamma interpretujemy dwa wyniki: wielkość współczynnika oraz poziom istotności. Jeśli wynik testu jest mniejszy niż 0,3 - oznacza to słaby związek. O związku umiarkowanym mówimy, gdy wartości współczynnika zawierają się powyżej 0,3, lecz poniżej 0,5. Test należy uznać za istotny statystycznie, jeśli poziom istotności będzie mniejszy od 0,05. W programie PSPP współczynnik gamma obliczamy wybierając z menu tekstowego *Analyze ⇨ Descriptive Statistics ⇨ Crosstabs ⇨ Statistics* i zaznaczamy pole *Gamma.*

13.2.1.4. Współczynnik korelacji rangowej tau-B (τ-B) i tau-C (τ-C) Kendalla

Współczynnik korelacji rangowej τ (od greckiej litery tau) jest nieco bardziej rozpowszechnionym w praktyce badawczej ekwiwalentem współczynnika korelacji rangowej rho Spearmana. Współczynnik ten zaproponował niemiecki fizyk i filozof Gustav T. Fechner w 1897 roku, a rozwinął i dopracował brytyjski statystyk Maurice G. Kendall w 1938 roku[19]. Niektórzy badacze wskazują, że miara ta została wynaleziona niezależnie przez B. Babingtona Smitha oraz Wilsona A. Wallisa w 1939 roku[20].

[19] M. Kendall, *A New Measure of Rank Correlation*, „Biometrika", 1938, 30 (1-2), s. 81-89.

[20] D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall CRC, Boca Raton, Londyn, Nowy Jork 2004, Test 31. Kendall's Coefficient of Concordance.

Tau Kendalla mierzy kierunek i siłę korelacji zmiennych porządkowych (rangowanych). Jest to nieparametryczna miara związku, dlatego normalność rozkładu korelowanych zmiennych nie musi być w tym przypadku zachowana.

Choć jest on ekwiwalentem rho Spearmana i tak jak on opiera się na analizie rang, to jednak logika jego obliczania jest różna, a różnice pomiędzy obu miarami wyraża zależność:

$$-1 \leq 3tau - 2r_s \leq 1$$

Istnieją trzy warianty tego współczynnika: tau-A, tau-B i tau-C. Tau-A to pierwotna miara związku nieodporna na tak zwane pary wiązane (te same wartości uporządkowanych zmiennych). Współczynnik tau-A nigdy jednak nie osiągał jedności, jeśli pojawiały się pary wiązane w zbiorze. W celu zaradzenia tej słabości wprowadzono tau-B i tau-C.

Tau Kendalla opiera się na pomiarze różnicy między prawdopodobieństwem tego, że dwie zmienne są ze sobą zgodne, a więc układają się w tym samym porządku (zmienna poprzedzająca i następująca), a prawdopodobieństwem, że ich uporządkowanie się różni. Współczynnik korelacji rangowej polega na zestawieniu uporządkowanych par wartości w trzech następujących grupach: 1/ par zgodnych - o tym samym kierunku zmian (rosnąco lub malejąco), oznaczamy je Z, 2/ par niezgodnych - o przeciwnym, niezgodnym kierunku zmian (N), 3/ par wiązanych, gdzie zmienne mają takie same wartości w dwóch kolejnych obserwacjach (W).

$$\tau = \frac{Z - N}{Z + N + W} = 2 * \left[\frac{Z - N}{N * (N - 1)}\right]$$

Minimalna liczba jednostek analizy konieczna do przeprowadzenia testu tau-B i tau-C Kendalla wynosi 10. Współczynnik przyjmuje wartości z zakresu od -1 do +1, jednak wartości -1 lub +1 mogą być uzyskane tylko dla tabel kwadratowych (tj. jeśli zmienne posiadają ten sam zakres, czyli taką samą liczbę rang). Znak współczynnika wskazuje na kierunek zależności, a jego wartość - na siłę związku. Im wartości bliższe jedności, tym silniejsza jest zależność.

Tau-B Kendalla stosujemy dla tabel o takiej samej liczbie kolumn i wierszy (tabel kwadratowych), a więc dla zmiennych, które mają taki sam zakres wartości. Z kolei tau-C przeznaczone jest dla tabel o różnej liczbie kolumn i wierszy (tabel prostokątnych), czyli stosujemy go wówczas, gdy korelowane zmienne mają różne zakresy wartości.

Wartość współczynnika powyżej 0,3 uznawana jest za korelację silną. Ponadto odczytujemy wartość istotności, którą interpretujemy identycznie jak w przypadku R Pearsona.

W programie PSPP współczynnik ten obliczamy wybierając z menu *Analyze ⇨ Descriptive Statistics ⇨ Crosstabs ⇨ Statistics*, a następnie zaznaczając pola BTau i CTau.

13.2.2. Miary związku dla zmiennych nominalnych

Najistotniejszą, konieczną do opanowania miarą, jest test zgodności chi-kwadrat (χ^2) autorstwa Karla Pearsona. Popularną miarą jest także V Haralda Craméra oraz fi (phi, ϕ) Udny Yule'a. Mniej popularne, rzadziej używane w praktyce analitycznej (choć obecne w programach statystyki na polskich uczelniach wyższych) jest współczynnik kontyngencji C Pearsona oraz współczynnik lambda (λ)

L.A. Goodmana i W.H. Kruskala. Rzadko używany w naukach społecznych jest z kolei współczynnik niepewności U Henri'ego Theila. Zamieszczenie go w podręczniku ma charakter promocyjny, ponieważ współczynnik ten może dostarczyć wiele wartościowych poznawczo wyników.

13.2.2.1. Test niezależności chi-kwadrat (χ^2) Pearsona

Test niezależności chi-kwadrat Pearsona jest jedną z najczęściej stosowanych statystyk, która stała się swoistym standardem w licznych dyscyplinach badawczych, jednym z najczęściej używanych testów statystycznych[21]. Autorem testu chi-kwadrat jest jeden z najznamienitszych statystyków Karl Pearson, o którym W.I. Lenin pisał jako o „najuczciwszym oponencie materializmu", trzeba bowiem wiedzieć, że zainteresowania tego uczonego daleko wykraczały poza statystykę, między innymi na obszar filozofii nauki[22]. K. Pearson ogłosił swoje odkrycie w 1900 roku, jednakże ze względu na zbyt skomplikowany sposób obliczania współczynnika nie znalazło ono oczekiwanego odzewu. W efekcie zaproponował prosty w obliczeniach test zgodności lub niezależności dla zmiennych mierzonych na poziomie nominalnym, który dziś powszechnie nazywa się testem niezależności chi-kwadrat Pearsona. Test chi-kwadrat Pearsona może posłużyć dwojako: po pierwsze, do sprawdzenia czy rozkład danej zmiennej różni się od rozkładu losowego (lub innego założonego przez badacza rozkładu), a po drugie, do orzeczenia o istnieniu lub nieistnieniu związku pomiędzy dwoma zmiennymi, co zostało opisane poniżej. Należy jednak pamiętać, że w przeciwieństwie na przykład do R Pearsona, chi-kwadrat nie służy do oceny siły i kierunku związku pomiędzy zmiennymi. Na podstawie uzyskanego wyniku orzekamy tylko czy związek jest obecny, czy go nie ma.

13.2.2.1.1. Zasady stosowania testu chi-kwadrat Pearsona

Test niezależności chi-kwadrat jest o wiele mniej wymagającą miarą, aniżeli współczynnik korelacji R, jednakże zmienne muszą spełnić szereg wymogów, by można było przeprowadzić ten test:

1/ Testy oparte na rozkładzie chi-kwadrat mogą być obliczane tylko dla wystarczająco licznych prób. Za taką próbę należy uznać posiadającą co najmniej 30 jednostek analizy ($N \geq 30$), choć niektórzy postulują, że minimalna liczba jednostek analizy powinna wynosić $N \geq 50$[23]. Im próba większa, tym mniejszy błąd w uogólnianiu wyników na populację.

2/ Liczba jednostek analizy, które znajdą się w tabeli krzyżowej przedstawiającej obie zmienne, na których chcemy przeprowadzić test chi-kwadrat, nie powinna być mniejsza niż 5 (choć jak wskazują dwaj znamienici brytyjscy statystycy George Udny Yule oraz Maurice George Kendall najlepiej, by nie schodziła poniżej 8[24] lub 10[25]). Jeśli w pojedynczej komórce tabeli krzyżowej powstałej z testowanych

[21] J. Ekström, *On Pearson-Verification And The Chi-Square Test*, w: http://preprints.stat.ucla.edu/625/Ekstro¨m%20-%20On%20Pearson-verification%20and%20the%20chi-square%20test.pdf, dostęp: maj 2012, s. 2.

[22] R.L. Plackett, *Karl Pearson and the Chi-Squared Test*, „International Statistical Review", 1983, 51, s. 59-72.

[23] G.U. Yule, M.G. Kendall, *Wstęp do teorii statystyki*, Państwowe Wydawnictwo Naukowe, Warszawa 1966, s. 471. Jednak autorzy sami przyznają, że „[...] trudno powiedzieć, kiedy liczebność jest duża" i traktują wskazaną liczbę jako orientacyjną.

[24] M. Sobczyk, *Statystyka,* Wydawnictwo Naukowe PWN, Warszawa 2002, s. 228.

[25] G.U. Yule, M.G. Kendall, dz. cyt., s. 471.

zmiennych jest zbyt mało wskazań, to problem ten można rozwiązać na dwa sposoby. Po pierwsze, agregując poszczególne wartości zmiennej w mniejszą liczbę kategorii. Na przykład zmienną PGSW 2007 s55b „Ważne dla bycia Polakiem: podzielanie polskich tradycji kulturowych", która zawiera wartości znaczące: *bardzo ważne, raczej ważne, niezbyt ważne, w ogóle nieważne* można zrekodować, uzyskując zmienną dychotomiczną o następujących wartościach: *uważa za ważne* oraz *uważa za nieważne*. W tej sytuacji liczba pól w poszczególnych komórkach zwiększy się, stanowiąc sumę liczebności komórek zagregowanych. Dla porządku należy wskazać, że scalaniu można poddać wartości zmiennej tylko wówczas, jeśli da się to sensownie uczynić. Drugim sposobem jest zastosowanie zmodyfikowanego testu opartego na rozkładzie chi-kwadrat - dokładnego testu R.A. Fishera dla tabel czteropolowych (2x2) (szczegóły jego wykorzystania podano w dalszej części tekstu). Współcześnie podaną wyżej zasadę obecności minimalnej liczby wskazań w każdej pojedynczej komórce traktuje się mniej restrykcyjnie. William Gemmell Cochran wykazał, że wartości w komórkach mogą być niższe, jednakże żadna z liczebności nie może zejść poniżej jedności oraz nie więcej niż 20 proc. wszystkich komórek może zawierać liczebności poniżej pięciu[26].

3/ Zmienne powinny być mierzone na skalach nominalnych (test niezależności chi-kwadrat można stosować także dla zmiennych porządkowych, jednakże wydaje się, że lepiej używać miar dedykowanych dla rangowego poziomu pomiaru[27]).

4/ Rozkład wartości zmiennych nie musi spełniać wymogu normalności rozkładu. Innymi słowy test niezależności chi-kwadrat stosujemy wówczas, gdy rozkład zmiennej nie przejdzie testu Kołmogorowa-Smirnowa lub wartość kurtozy i skośności wykraczać będzie poza wartości zakresu od -1 do +1.

5/ Zmienne wykorzystane w teście chi-kwadrat powinny być zmiennymi ciągłymi, to jest zakres zmiennej powinien zawierać jak najwięcej wartości. Jeśli tak się nie dzieje i mamy do czynienia ze zmiennymi dychotomicznymi, z których każda przyjmuje dwie wartości, wówczas wykorzystujemy test dokładny R.A. Fishera. Zwróćmy uwagę na fakt, że postulat ten może kolidować z wymogiem sformułowanym w punkcie 2/.

W zależności od liczby wartości zmiennych (liczba komórek w tabeli powstałej po skrzyżowaniu zmiennych) oraz liczebności (liczby wskazań, kategorii) w tych komórkach stosujemy różne sposoby zmodyfikowanego obliczania testu chi-kwadrat. Program PSPP wylicza test chi-kwadrat na cztery sposoby:

1/ Chi-kwadrat Pearsona *(Pearson chi-square)* - najszerzej używany test chi-kwadrat Pearsona. Jeśli wszystkie wyżej wymienione wymagania zostały spełnione, wówczas tego testu używamy do interpretacji wykonanych analiz. Jest on przybliżeniem rachunkowym ilorazu wiarygodności; oblicza się go prosto, jego wyliczenie nie wymaga użycia komputera. W tekście zaprezentowano właśnie ten uproszczony sposób obliczania chi-kwadrat.

2/ Iloraz wiarygodności (*likelihood ratio*), który jest klasycznym, pierwotnie (1900) zaproponowanym przez Karla Pearsona współczynnikiem chi-kwadrat. Jednak ze względu na trudności w jego obliczaniu

[26] W.G. Cochran, *The chi-square goodness-of-fit test,* „Annals of Mathematical Statistics", 1952, 23, s. 315-345.

[27] Jak podkreśla Hubert M. Blalock: „Skala pomiarowa może być oczywiście silniejsza. Stosuje się czasem test chi-kwadrat do skal porządkowych, a nawet interwałowych". H.M. Blalock, *Statystyka dla socjologów*, Państwowe Wydawnictwo Naukowe, Warszawa 1977, s. 244.

powszechnie stosowano uproszczoną miarę, znaną obecnie jako test chi-kwadrat Pearsona. Iloraz wiarygodności obliczany jest z użyciem regresji logistycznej; mówi nam, ile razy częściej wystąpiłby dany czynnik, gdyby prawdziwa była hipoteza alternatywna (H_1), niż gdyby prawdziwa była hipoteza zerowa (H_0). Interpretujemy go analogicznie jak inne odmiany chi-kwadrat i sam test chi-kwadrat. Warto go używać z przyczyn formalnych, choć przyjęło się, że najpowszechniej wykorzystuje się chi-kwadrat Pearsona.

3/ Dokładny test Ronalda Aymlera Fishera *(Fisher's exact test*). Przeznaczony jest dla tabel 2x2. Stosujemy go wówczas, gdy liczebności próby są małe (a więc liczba jednostek analizy jest niższa niż 30 lub 50) lub też, jeśli w którejkolwiek z komórek wartość jest poniżej 5[28].

4/ Poprawka ciągłości Franka Yatesa (C*ontinuity Correction*) - służy do obliczania testu chi-kwadrat dla zmiennych dwuwartościowych (tabel 2x2).

5/ Test związku liniowego (*Linear-by-Linear-Association*) - stanowi odmianę chi-kwadrat przystosowaną do pomiaru zmiennych na poziomie porządkowym. W praktyce rzadko stosowany, bowiem w przypadku zmiennych mierzonych na tym poziomie lepiej zastosować na przykład rho Spearmana lub inne miary przeznaczone dla porządku rangowego.

Należy wskazać, że powyższe testy (np. test dokładny Fishera) pojawiają się w odpowiedniej tabeli w oknie raportów PSPP wówczas, gdy liczebności zmiennych spełniają założone warunki.

13.2.2.1.2. Obliczanie testu chi-kwadrat Pearsona

Test chi-kwadrat jest na tyle ważną i popularną miara, że warto znać sposób jego obliczania. Istotą testu chi-kwadrat jest matematyczne, ilościowe porównanie współzmienności dwóch zmiennych nominalnych.

W teście chi-kwadrat, podobnie jak w innych tego typu testach, stawiamy dwie hipotezy: zerową i alternatywną:

H_0 - **nie istnieje** istotny statystycznie związek pomiędzy badanymi zmiennymi. Test **NIE JEST** statystycznie istotny;

H_1 - **istnieje** statystycznie istotny związek pomiędzy zmiennymi. Test **JEST** statystycznie istotny.

W toku przeprowadzania testu chcemy odrzucić hipotezę zerową, a więc stwierdzić, że związek między zmiennymi istnieje. Hipotezę zerową możemy odrzucić wówczas, gdy wartość współczynnika p jest mniejsza lub równa 0,05. Jeśli wartość p jest wyższa stwierdzamy, że nie ma podstaw do odrzucenie hipotezy zerowej, a w efekcie wnioskujemy, że związek pomiędzy zmiennymi nie istnieje. A zatem: jeśli p jest mniejsze niż 0,05 - wnioskujemy o zależności między zmiennymi, jeśli jest wyższe - o jej braku.

Współczynnik chi-kwadrat obliczamy porównując wartości oczekiwane (tzw. liczebności teoretyczne) z wartościami obserwowanymi (liczebnościami empirycznymi). Wartości oczekiwane to takie, jakie przyjmowałyby dwie badane przez nas zmienne, gdyby żaden związek pomiędzy nimi nie istniał. Z kolei liczebności obserwowane to te, które zmierzyliśmy.

[28] R.A. Fisher, *The logic of inductive inference,* „Journal of the Royal Statistical Society", 1935, 98, s. 39-54.

Przypuśćmy, że chcemy dowiedzieć się, czy istnieje związek pomiędzy płcią obywatela a poparciem dla pewnego polityka. W tym celu zbadaliśmy 60 osób - 30 mężczyzn i 30 kobiet. Zapytaliśmy ich czy zagłosowali w ostatnich wyborach na interesującego nas polityka X. Mogli oni wybierać pomiędzy dychotomiczną odpowiedzią *tak* lub *nie*. Wyniki tego fikcyjnego pomiaru przedstawia tabela 32.

Tabela 32. Rozkłady wartości obserwowanych dla zmiennej płeć i głosowanie (liczebności empiryczne).

Płeć	Głosował(a) na polityka X		Razem
	Tak	Nie	
Kobieta	21	9	30
Mężczyzna	11	19	30
Razem	32	28	60

Tak jak wskazano, istotę testu chi-kwadrat stanowi porównanie rozkładów wartości teoretycznych z wartościami empirycznymi. Liczebności empiryczne obliczamy na podstawie liczebności teoretycznych. Chodzi o uzyskanie rozkładów maksymalnie niezróżnicowanych, to jest takich, gdzie pomiędzy zmiennymi nie ma żadnej współzmienności. Obliczmy krok po kroku jedną przykładową liczebność teoretyczną.

Czynimy to wedle następującego wzoru:

$$E_{n,m} = \frac{O_n * O_m}{O}$$

gdzie:

$E_{n,m}$ - liczebność oczekiwana (z ang. *expected*) w konkretnej obliczanej przez nas komórce, to jest w wierszu n i kolumnie m,

O_n - suma liczebności obserwowanych (z ang. *observed*) w danym wierszu,

O_m - suma liczebności obserwowanych w danej kolumnie,

O - suma wszystkich liczebności obserwowanych.

Przyjmijmy, że chcemy obliczyć liczebność teoretyczną dla komórki kobiet głosujących na danego polityka. W tym przypadku suma liczebności obserwowanych w danym wierszu (O_n) wynosi:

O_n = 21 (kobiety głosujące na *tak*) + 9 (kobiety głosujące na *nie*) = 30

Z kolei suma liczebności obserwowanych w danej kolumnie dla tej komórki jest następująca:

O_m = 21 (kobiety głosujące na *tak*) + 11 (mężczyźni głosujący na *tak*) = 32

Natomiast suma wszystkich liczebności obserwowanych wynosi:

O = 21 (kobiety na *tak*) + 9 (kobiety na *nie*) + 11 (mężczyźni na *tak*) + 19 (mężczyźni na *nie*) = 60

A zatem możemy obliczyć liczebność teoretyczną dla wybranej komórki:

$$E_{n,m} = \frac{30 * 32}{60} = 16$$

Analogicznie obliczamy pozostałe trzy liczebności teoretyczne; wynik tych obliczeń prezentuje tabela 33.

Tabela 33. Rozkłady wartości oczekiwanych dla zmiennej płeć i głosowanie (liczebności teoretyczne).

Płeć	Głosował(a) na polityka X		Razem
	Tak	Nie	
Kobieta	16	14	30
Mężczyzna	16	14	30
Razem	32	28	60

Różnice pomiędzy wartościami obserwowanymi a wartościami oczekiwanymi nazywamy resztami (*residuals*). Na przykład dla mężczyzn głosujących na polityka X reszty przyjmują wartość ujemną i wynoszą: -5 (bowiem: 11-16=-5). Im większe wartości reszt, w tym większym stopniu możemy przypuszczać (ale tylko przypuszczać dopóki nie przeprowadzimy testu chi-kwadrat) o istnieniu zależności pomiędzy zmiennymi.

Po zestawieniu liczebności obserwowanych w tabelę i obliczeniu na ich podstawie liczebności oczekiwanych możemy przystąpić do obliczenia testu chi-kwadrat. Tu konieczne jest ważne zastrzeżenie - nie jest to oryginalny test zaproponowany przez K. Pearsona w 1900 roku, lecz uproszczony przezeń, stanowiący jego w miarę dokładne przybliżenie odpowiednik. Jednakże wzoru tego używa się najczęściej, ten oryginalny niemal całkowicie pomijając:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \cdots + \frac{(O_n - E_n)^2}{E_n}$$

gdzie:

O_1 - O_n - wartości obserwowane w poszczególnych komórkach.

E_1 - E_n - wartości oczekiwane w poszczególnych komórkach.

A zatem chi-kwadrat można określić jako sumę różnic wartości obserwowanych i oczekiwanych podniesionych do kwadratu do wartości oczekiwanej. Aby obliczyć chi-kwadrat należy podstawić do wzoru wartości z tabel 32 i 33:

$$\chi^2 = \frac{(21 - 16)^2}{16} + \frac{(9 - 14)^2}{14} + \frac{(11 - 16)^2}{16} + \frac{(19 - 14)^2}{14} = 6{,}696$$

Wartość tak obliczonego testu chi-kwadrat należy porównać z tablicami rozkładu chi-kwadrat. Takie tablice znajdują się w każdym niemal podręczniku statystyki, łatwo można również odnaleźć je w Internecie. Tablice takie zamieszczono także w formie aneksu na końcu niniejszej publikacji (tabela jest ograniczona do najczęściej występujących wartości).

Aby skorzystać z takiej tablicy, musimy jeszcze policzyć tak zwaną liczbę stopni swobody oraz wskazać akceptowany przez nas poziom ufności. Liczbę stopni swobody oznaczamy skrótem v i obliczamy następująco:

$$v = (\text{liczba wartości pierwszej zmiennej} - 1) * (\text{liczba wartości drugiej zmiennej} - 1)$$

a zatem:

$$v = (2 - 1) * (2 - 1) = 1$$

Z kolei poziom ufności oznacza przyjęte przez badacza ryzyko popełnienia błędu I rodzaju, a więc odrzucenie hipotezy zerowej, która w istocie jest prawdziwa. Wyznacza je sam badacz. W naukach społecznych umownie i powszechnie przyjmuje się, że zadowalającą wartością jest wartość mniejsza od 0,05. Oznacza to, że przyjmujemy ryzyko rzędu 5 procent, że popełnimy błąd I rodzaju.

Dysponując trzema wartościami: wynikiem testu chi-kwadrat równym 6,696, liczbą stopni swobody wynoszącą 1 oraz założonym przez nas poziomem ufności α=0,05, możemy odczytać wartość z tabeli wartości krytycznych rozkładu chi-kwadrat znajdującym się w aneksie na końcu publikacji. Właściwą liczbę odnajdujemy w tabeli wartości krytycznych na przecięciu poziomu ufności α (dla nas 0,05) oraz liczby stopni swobody (dla nas 1). Wartość ta wynosi: 3,841. Porównujemy ją z wynikiem testu chi-kwadrat. Jeśli wartość testu chi-kwadrat jest większa od wartości podanej w tabeli (a w naszym przykładzie tak jest) to odrzucamy hipotezę zerową. Innymi słowy: możemy zakładać, że pomiędzy faktem popierania polityka X w wyborach a przynależnością do określonej płci istnieje istotny statystycznie związek.

Zwróćmy jednak uwagę, że dla celów dydaktycznych przyjęliśmy nader uproszczony przykład, a mianowicie tabelę najprostszą z możliwych - czteropolową (tzw. tabelę 2x2). Oznacza to, że aby wykonać właściwie obliczenia powinniśmy użyć poprawki ciągłości Franka Yatesa (C*ontinuity Correction*) przeznaczonej dla tabel 2x2. Poprawka ta polega na odjęciu liczby 0,5 od modułu różnicy między liczebnościami obserwowanymi, a liczebnościami oczekiwanymi:

$$\chi^2 = \frac{(|O_1 - E_1| - 0{,}5)^2}{E_1} + \cdots + \frac{(|O_n - E_n| - 0{,}5)^2}{E_n}$$

$$\chi^2 = \frac{(|21 - 16| - 0{,}5)^2}{16} + \frac{(|9 - 14| - 0{,}5)^2}{14} + \frac{(|11 - 16| - 0{,}5)^2}{16} + \frac{(|19 - 14| - 0{,}5)^2}{14} = 5{,}424$$

Wynik testu chi-kwadrat z poprawką na ciągłość Yatesa również pozwala nam na stwierdzenie, że pomiędzy zmiennymi istnieje współzależność. Należy jednak zauważyć, że wartość testu chi-kwadrat po wprowadzeniu poprawki obniżyła się - poprawka Yatesa jest bardziej rygorystyczna, stawia bardziej wygórowane warunki.

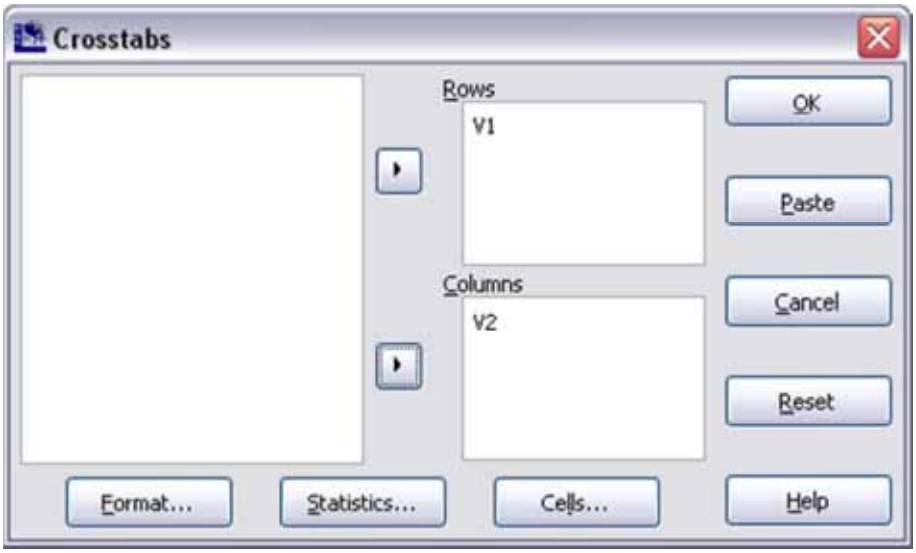
13.2.2.1.3. Obliczanie i interpretacja testu niezależności chi-kwadrat Pearsona w programie PSPP

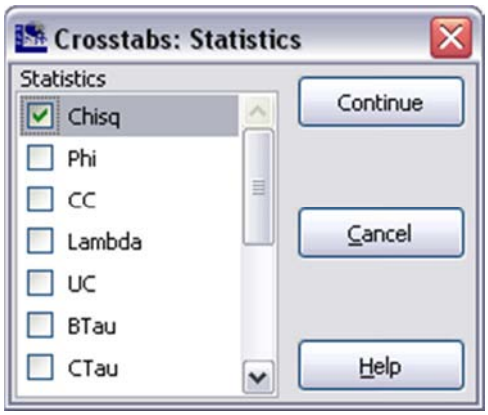
Jak wskazano we wstępie, test chi-kwadrat Pearsona może posłużyć w dwóch następujących sytuacjach: po pierwsze, do oszacowania czy rozkład badanej zmiennej różni się od rozkładu losowego lub innego wskazanego przez nas rozkładu (tak zwany test chi-kwadrat jednej zmiennej), po drugie, do zbadania istnienia zależności pomiędzy dwiema zmiennymi (test chi-kwadrat dwóch zmiennych). Poniżej przedstawiono przykłady drugiego z wymienionych zastosowań.

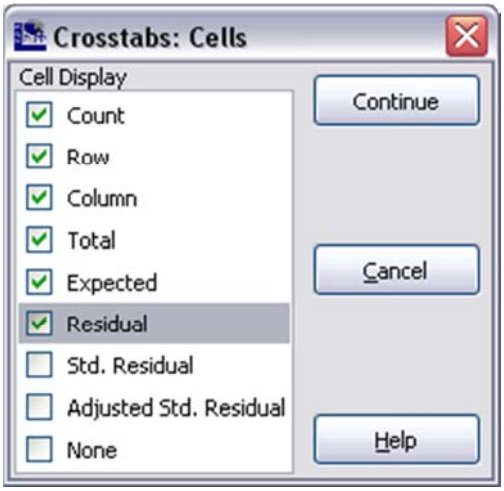
Test chi-kwadrat dla dwóch zmiennych służy do rozstrzygnięcia o istnieniu lub nieistnieniu związku pomiędzy tymi zmiennymi.

Przeprowadzamy go wybierając z menu tekstowego *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs.* W teście wykorzystane zostaną dwie zmienne ze zbioru PGSW 2007: płeć (m2) oraz pytanie o to czy rażą respondenta wypowiedzi Kościoła na tematy moralne i obyczajowe (p83e). Zakładamy przy tym, że pogląd na kwestie związane z „aktualizacją wartości", czyli swoistym, społecznym przypominaniem norm, będzie różny u mężczyzn i kobiet. W wiersze *(Rows)* wprowadzamy pierwszą zmienną, a w kolumny *(Columns)* - drugą. Nie ma znaczenia, w jakiej kolejności wpisujemy zmienne - test chi-kwadrat nie mierzy kierunku zależności, a wyniki testu, niezależnie od kolejności wprowadzenia zmiennych, są takie same. W *Statistics* domyślnie oznaczony jest test chi-kwadrat. Z kolei w *Cells* (komórki) oznaczamy, by program pokazywał wartości oczekiwane *(Expected)* oraz reszty (*Residuals*). Dzięki zaznaczeniu tych ostatnich dwóch opcji uzyskamy wyliczenie dla wartości oczekiwanych (*Expected*), a więc takich, które idealnie spełniają hipotezę zerową testu chi-kwadrat.

W efekcie podjętych działań uzyskujemy trzy kolejne tabele: tabelę podsumowania, tabelę krzyżową, uwzględniającą wartości obserwowane wraz z częstościami oraz wartości oczekine, a także wynik zestawu testów opartych na rozkładach chi-kwadrat. Pierwsza i druga tabela mają naczenie czysto orientacyjne.







Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Płeć * Czy rażą: wypowiedzi Kościoła na tematy moralne i obyczajowe.	1782,29	100.0%	,00	0.0%	1782,29	100.0%

Płeć * Czy rażą: wypowiedzi Kościoła na tematy moralne i obyczajowe. [count, row %, column %, total %, expected, residual].

	Czy rażą: wypowiedzi Kościoła na tematy moralne i obyczajowe.		
Płeć	rażą	nie rażą	Total
mężczyzna	306,2	542,9	849,1
	285,8	563,2	,0
	36,1%	63,9%	100,0%
	51,0%	45,9%	47,6%
	17,2%	30,5%	47,6%
	20,4	-20,4	,0
kobieta	293,8	639,5	933,2
	314,1	619,1	,0
	31,5%	68,5%	100,0%
	49,0%	54,1%	52,4%
	16,5%	35,9%	52,4%
	-20,4	20,4	,0
Total	600,0	1182,3	1782,3
	33,7%	66,3%	100,0%
	100,0%	100,0%	100,0%
	33,7%	66,3%	100,0%

Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*
Pearson Chi-Square	4,19	1,00	,04
Likelihood Ratio	4,18	1,00	,04
Continuity Correction	3,98	1,00	,05
Linear-by-Linear Association	4,18	1,00	,04
N of Valid Cases	1782,29		

Z pierwszej tabeli odczytujemy liczbę jednostek analizy, które poddano testowaniu (w analizowanym przypadku 1782,29 - wartość ta ma wartości ułamkowe, bowiem na zmienne PGSW nałożono wagi post-stratyfikacyjne). Druga tabela zawiera kolejno w komórkach wartości obserwowane, wartości oczekiwane, wartości procentowe w wierszach, wartości procentowe w kolumnach oraz wartości procentowe ogółem, aż wreszcie - reszty. Na podstawie danych odczytanych z tej tabeli orzekamy wstępnie o różnicach między zmiennymi. Zwróćmy uwagę, że w dwóch komórkach (po przekątnej) wartości reszt są dodatnie, a w dwóch - ujemne. Dobrze to rokuje dla ujawnienia istotnych statystycznie różnic pomiędzy zmiennymi. Zwróćmy uwagę, że mamy do czynienia z tak zwaną tabelą 2x2 (dwie wartości zmiennej płeć oraz dychotomiczne wartości zmiennej dotyczącej wypowiedzi Kościoła na tematy moralne i obyczajowe: *rażą* - *nie rażą*). Trzecia tabela zatytułowana Testy chi-kwadrat (*Chi-Square tests*) dostarcza najistotniejszych informacji. W pierwszej kolumnie znajduje się nazwa wykonywanego testu, w drugiej kolumnie figuruje jego obliczona wartość, trzecia kolumna zawiera liczbę stopni swobody, a czwarta - poziom ufności, którego wartość decyduje o przyjęciu lub odrzuceniu hipotezy zerowej.

W zależności od tego, jakie dane stanowiły przedmiot naszych analiz, odczytujemy i interpretujemy wartości różnych testów. Rozważmy: mamy do czynienia z tabelą 2x2 zawierającą duże liczebności w ogóle i w każdej komórce (przekraczające 10). W związku z tym niezbyt dobrą podstawę interpretacji stanowi test chi-kwadrat Pearsona (*Pearson Chi-Square*), a także iloraz wiarygodności

(*Likelihood Ratio*), które przewidziane są dla większej liczby wartości zmiennych niż dwie. Odrzucamy również test związku liniowego (*Linear-by-Linear Association*), ponieważ jest on przeznaczony dla zmiennych mierzonych na skali porządkowej. Pozostaje zatem odczytanie wartości poprawki ciągłości Yatesa (*Continuity Correction*) - miary dedykowanej dla tabel 2x2. Zwróćmy jednakże uwagę, że miary te różnią się od siebie nieznacznie i tylko w granicznych przypadkach wybór pomiędzy nimi może mieć istotne znaczenie. Wartość testu chi-kwadrat wynosi 3,98, a liczba stopni swobody - 1 (patrz: powyższe rozważania).

Liczba stopni swobody oraz wynik testu chi-kwadrat pozwala na odczytanie z tablic chi-kwadrat wartości właściwej dla tych dwóch parametrów i orzeczenie o istotności testu. Nie musimy jednak tego czynić - program PSPP sam podaje wynik tego porównania. Ostatnia kolumna tabeli mówi nam o tym, czy zaobserwowana różnica między wartościami zaobserwowanymi obu zmiennych a wartościami oczekiwanymi jest istotna statystycznie, pozwalając na przyjęcie lub odrzucenie hipotezy zerowej. Sformułujmy przyjęte dla przeprowadzonego testu hipotezy:

H_0 - **nie istnieje** istotna statystycznie zależność pomiędzy badanymi zmiennymi. Test **NIE JEST** statystycznie istotny;

H_1 - **istnieje** statystycznie istotna zależność pomiędzy zmiennymi. Test **JEST** statystycznie istotny.

Jeśli wartość podana w ostatniej kolumnie jest mniejsza lub równa standardowo przyjmowanego w naukach społecznych poziomowi 0,05 oznacza, że wynik jest istotny statystycznie, a zatem odrzucamy hipotezę zerową. Z kolei jeśli wartość p przekroczy 0,05 oznacza to, że nie ma podstaw, by hipotezę zerową odrzucić, a co za tym idzie stwierdzamy, że pomiędzy wartościami obserwowanymi a oczekiwanymi nie istnieje statystycznie istotna różnica. W przypadku danych analizowanych w niniejszym przykładzie wartość $p = 0,05$. Znalazła się ona na granicy istotności, jednakże pozwala na orzeczenie, że pomiędzy zmiennymi istnieje zależność: mężczyźni bardziej niechętnie reagują na publiczne propagowanie pewnych norm przez kler, a kobiety - bardziej przyzwalająco, konformistycznie. W raporcie z badań zapisujemy wynik testu statystycznego następująco:

$$\chi^2 (1, N = 1782,29) = 3,98; p \leq 0,05$$

W formalnym zapisie w pierwszej kolejności podajemy liczbę stopni swobody, następnie liczbę jednostek analizy oraz wynik testu chi-kwadrat, a po średniku prawdopodobieństwo prawdziwości hipotezy zerowej (jeśli wynik byłby statystycznie nieistotny, wówczas w miejsce $p \leq 0,05$ wstawiamy litery *ni* będące skrótem od *nieistotny statystycznie*). Podkreślmy jednocześnie, że integralną częścią raportu byłaby próba wyjaśnienia wydobytego zajwiska w świetle wiedzy naukowej (istniejących hipotez lub teorii) lub co najmniej na podstawie wiedzy pozaźródłowej (względnie: wykazanie artefaktualności zjawiska). Warto również przytoczyć za G.U. Yule'm i M.G. Kendallem ostrzeżenie dla początkujących analityków stosujących test niezależności chi-kwadrat: „[...] bardzo dobra zgodność jest zbyt dobra,aby mogła być prawdziwa"[29]. Poniżej zaprezentowany został sposób obliczania chi-kwadrat w języku skryp towym programu PSPP:

```
CROSSTABS
/TABLES= m2 BY p83e
/FORMAT=AVALUE TABLES PIVOT
```

[29] G.U. Yule, M.G. Kendall, dz. cyt., s. 472.

```
/STATISTICS=CHISQ
/CELLS=COUNT ROW COLUMN TOTAL EXPECTED RESIDUAL.
```

Rozkaz CHISQ powoduje wykonanie testu chi-kwadrat, a EXPECTED i RESIDUAL są odpowiedzialne za wywołanie odpowiednio wartości oczekiwanych i reszt.

13.2.2.2. Współczynnik kontyngencji C Pearsona

Współczynnik kontyngencji C został opracowany przez Karla Pearsona. Jest on miarą opartą na współczynniku chi-kwadrat i wykorzystuje się go do określania siły zależności między dwiema zmiennymi nominalnymi. Nie można natomiast za jego pomocą wnioskować kierunku tego związku. Współczynnik ten stosujemy do dowolnych tabel: od najmniejszych (2x2) do dowolnie dużych, zarówno do tabel symetrycznych (kwadratowych), jak też asymetrycznych (prostokątnych), to jest o nierównej liczbie wierszy i kolumn[30]. Niektórzy badacze sugerują, by wykorzystywać go raczej do tabel większych niż licząc dwa wiersze i dwie kolumny (tzw. tabel 2x2 lub inaczej czteropolowych)[31]. Jest on współczynnikiem niezbyt często używanym, choć promowanym w dydaktyce i komputerowych pakietach statystycznych.

Współczynnik kontyngencji C Pearsona obliczamy za pomocą wzoru:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

gdzie N oznacza liczbę jednostek analizy, a χ^2 wartość uprzednio obliczonego współczynnika chi-kwadrat.

Przeznaczeniem tego współczynnika jest standaryzacja chi-kwadrat, a więc umożliwienie porównywania współczynników chi-kwadrat pochodzących z obliczeń wykonanych na różnych pod względem liczby itemów zmiennych. Współczynnik kontyngencji C Pearsona przyjmuje wartości od 0 do 1, przy czym nigdy w praktyce nie osiąga jedności (jeśli liczba itemów, a więc zakres obu testowanych zmiennych byłby nieskończony, a związek pomiędzy nimi idealny, wówczas współczynnik ten osiągnąłby wartość 1). Jest on wrażliwy na liczbę kolumn i wierszy w tabeli (a więc na liczbę itemów w badanych zmiennych, czyli innymi słowy na zakres zmiennych). Im liczba ta większa, tym wyższa i bliższa jedności może być wartość współczynnika kontyngencji. Z kolei im liczba ta jest niższa, tym ten współczynnik niższy. A zatem: zwiększanie zakresu badanych zmiennych prowadzi do wyników stwierdzających zależność tych zmiennych, a zmniejszanie - do redukowania zależności. Na przykład wartość współczynnika kontyngencji przy idealnej, całkowitej współzależności pomiędzy zmiennymi dla tabeli 3x3 wynosi 0,816, a dla tabeli 4x4 już 0,866. W celu zaradzenia temu defektowi, a także po to by umożliwić porównywanie wyników pomiędzy tabelami o różnym zakresie wartości zmiennych (a więc wierszy i kolumn w tabeli), stosuje się tak zwane korekty tego współczynnika. Korektę tę obliczamy według wzoru dla tabel symetrycznych (kwadratowych):

[30] Tabelę, o której mowa tworzą dwie skrzyżowane ze sobą zmienne pomiędzy którymi badamy współzależność. Liczba wierszy zależy od liczby itemów (zakresu wartości) danej zmiennej, a liczba kolumn - drugiej zmiennej. Jeśli mamy do czynienia z równą liczbą itemów (równym zakresem zmiennej) dla obu zmiennych, wówczas mówimy o tabeli symetrycznej, a jeśli nie - to o asymetrycznej.

[31] G. Clauss, H. Ebner, *Podstawy statystyki dla psychologów, pedagogów i socjologów,* Państwowe Zakłady Wydawnictw Szkolnych, Warszawa 1972, s. 289.

$$C_{kor} = \sqrt{\frac{k-1}{k}}$$

gdzie k oznacza liczbę kolumn w tabeli (a tym samym wierszy).

Natomiast dla tabel asymetrycznych (prostokątnych) wartość korekty wynosi:

$$C_{kor} = \sqrt{\frac{1}{2} * \left(\frac{k-1}{k} + \frac{w-1}{w}\right)}$$

gdzie k i w oznaczają odpowiednio kolumny i wiersze.

Wartość skorygowanego współczynnika C obliczamy dzieląc uzyskaną wartość przez wartość korekty, jak następuje:

$$C_{skor} = \frac{C}{C_{kor}}$$

Tak obliczony współczynnik kontyngencji C jest bardziej standaryzowany, w tym sensie, że można porównywać i oceniać bardziej adekwatnie związki pomiędzy zmiennymi o różnych zakresach, a także lepiej interpretowalne są wartości pośrednie pomiędzy 0 a 1.

Współczynnik kontyngencji C Pearsona oznaczany jest w programie PSPP literami CC (nazwa ta to skrót od anglojęzycznego *contingency coefficient,* czyli współczynnik kontyngencji). W programie PSPP współczynnik obliczany jest bez poprawek. Powinien być on w programie PSPP obliczany po wyznaczeniu wartości chi-kwadrat (obliczamy chi-kwadrat dla interesującej nas pary zmiennych, a następnie sprawdzamy jego statystyczną istotność i dopiero wówczas decydujemy o obliczeniu lub zaniechaniu - w przypadku statystycznej nieistotności - współczynnika C Pearsona). Kolejność taka jest zalecana z dwóch powodów. Po pierwsze, dlatego, że współczynnik C jest miarą standaryzującą oraz uzupełniającą chi-kwadrat. Po drugie, dlatego, że w wersji 0.7.9 nie jest obliczany poziom istotności przy współczynniku kontyngencji C, w związku z czym należy wykorzystać ten umieszczany przy chi-kwadrat.

Współczynnik C jest dostępny w programie PSPP w *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs* ⇨ *Statistics* ⇨ CC. Z kolei w edytorze składni, aby obliczyć współczynnik kontyngencji C, wpisujemy następującą sekwencję komend:

```
CROSSTABS
/TABLES= V1 BY V2
/STATISTICS=CC.
EXECUTE.
```

13.2.2.3. Współczynnik fi (phi, φ) Yule'a

Współczynnik fi wprowadził brytyjski statystyk Udny Yule (wł. George Udny Yule). Jest to miara nieparametryczna dla dwóch zmiennych nominalnych, czyli tzw. tabel asocjacji 2x2. W tabelach 2x2 przyjmuje wartość od 0 do 1 i tu jest równy R Pearsona. Z kolei dla tabel większych (dowolnych) współczynnik przyjmuje wartości od 0 do nieskończoności, przy czym górna wartość wielkości tego współczynnika zależy od wielkości tabeli.

Współczynnik ten obliczamy według następującego wzoru dla tabel 2x2:

$$\varphi = \frac{(A * D) - (B * C)}{\sqrt{(A + B) * (C + D) * (A + C) * (B + D)}}$$

gdzie poszczególne litery oznaczają wartości znajdujące się w komórkach tabeli 34.

Tabela 34. Tabela asocjacji 2x2

	Nie (wartość 0)	**Tak (wartość 1)**
Tak (wartość 1)	A	B
Nie (wartość 0)	C	D

Z kolei dla tabel większych niż 2x2 stosujemy wzór:

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

gdzie chi-kwadrat to wartość przeprowadzonego testu chi-kwadrat, a N to liczba jednostek analizy. W istocie jest to wartość chi-kwadrat przypadająca na jednostkę analizy, z której to wartości wyciągnięto pierwiastek.

Przeanalizujmy przykładowe zastosowanie tego współczynnika. Doskonale nadaje się on do zbadania siły związku pomiędzy przynależnością do określonej płci a uczestnictwem w wyborach parlamentarnych. W zbiorze PGSW pierwsza z tych zmiennych oznaczona jest jako m2, a druga - c32. Zauważmy, że są to dwie zmienne nominalne, dwuwartościowe - m1 posiada wartość - kobieta, mężczyzna, a c32 wartość - tak, nie.

Podstawmy dane z PGSW do tabeli 35.

Tabela 35. Tabela asocjacji 2x2 z wartościami empirycznymi

	Nie brał(a) udziału w wyborach parlamentarnych 2005 roku	**Brał(a) udziału w wyborach parlamentarnych 2005 roku**
Kobieta	256	624
Mężczyzna	215	605

Pobieżna obserwacja tabeli wykazuje, że raczej nie istnieje związek pomiędzy płcią a uczestnictwem lub brakiem uczestnictwa w wyborach parlamentarnych. Przekonajmy się jednak o tym obliczając współczynnik fi:

$$\varphi = \frac{(256 * 605) - (624 * 215)}{\sqrt{(256 + 524) * (215 + 605) * (256 + 215) * (624 + 605)}} = 0{,}032$$

Współzależność wyrażaną przez ten współczynnik uznajemy za nikłą, jeśli jej wartość jest poniżej 0,1, za słabą, gdy zawiera się pomiędzy 0,1 a 0,3. Pomiędzy 0,3 a 0,5 jest to współzależność przeciętna, wysoka gdy współczynnik przyjmie wartości w zakresie 0,5 a 0,7 i bardzo wysoka powyżej 0,7. Oznacza to, że płeć nie determinowała udziału lub braku udziału w wyborach parlamentarnych. Ponadto ważne jest w tym teście - podobnie jak w innych - odczytanie wartości jego statystycznej istotności. By został on uznany za istotny statystycznie konieczne jest, by współczynnik istotności był równy lub mniejszy od 0,05. Warto jednocześnie zwrócić uwagę na pewną matematyczną właściwość współczynnika - może być on zawyżony, jeśli liczba wierszy i kolumn jest wyższa.

W programie PSPP obliczamy ów współczynnik oznaczając w *Analyze ⇨ Descriptive Statistics ⇨ Crosstabs ⇨ Statistics ⇨ Phi*. Zaznaczenie współczynnika Phi powoduje jednoczesne powoduje obliczenie V Craméra.

W trybie poleceń języka skryptowego PSPP współczynnik obliczamy następująco:

```
CROSSTABS
/TABLES= V1 BY V2
/FORMAT=AVALUE TABLES PIVOT
/STATISTICS= PHI
/CELLS=COUNT ROW COLUMN TOTAL.
```

13.2.2.4. Współczynnik lambda (λ) Goodmana i Kruskala

Współczynnik lambda (**λ**) autorstwa Williama H. Kruskala i Leo A. Goodmana jest miarą siły związku pomiędzy dwiema zmiennymi nominalnymi. Przyjmuje on wartość od 0 do 1, przy czym 0 nie wyklucza braku zależności (dzieje się tak w sytuacji, gdy zmienne przyjmą specyficzne wartości). Wartość 1 oznacza idealną zależność pomiędzy zmiennymi. Współczynnik lambda opiera się na mechanizmie nazywanym proporcjonalną redukcją błędu (*proportional reduction in error*). W uproszczeniu jest to sposób oszacowania wartości jednej zmiennej na podstawie drugiej. Znając wartość zmiennej niezależnej możemy przewidywać (a właściwie zgadywać) wartość nieznanej nam zmiennej zależnej. Im zmienne są ze sobą mocniej powiązane, tym mniej błędów popełnimy przewidując (odgadując) jedną na podstawie drugiej[32]. Miara ta może być obliczana jako symetryczna i jako asymetryczna. Miara symetryczna to taka, w której nie zakładamy, która zmienna jest zmienną zależną, a która - niezależną. Z kolei niesymetryczne miary zakładają wskazanie jednej zmiennej jako zależnej, a drugiej jako niezależnej. W tym przypadku zamiana roli zmiennych będzie skutkowała różnymi wynikami obliczeń.

W programie PSPP współczynnik lambda obliczamy wybierając z menu tekstowego *Analyze ⇨ Descriptive Statistics ⇨ Crosstabs ⇨ Statistics* i zaznaczamy *check box: Lambda.* Interpretacji poddajemy dwie wartości pochodzące z tabel wygenerowanych przez program PSPP. Po pierwsze, sprawdzamy poziom przybliżonej istotności (*approximate significance*); jeśli jest ona poniżej 0,05, wówczas wnioskujemy o fakcie istotności związku pomiędzy dwoma zmiennymi. W przeciwnym wypadku stwierdzamy, że związek jest nieistotny. Po wtóre, interpretujemy wartość współczynnika korelacji (pole *value)*. Jeśli jest on wyższy niż 0,3, mówimy o silnej korelacji pomiędzy zmiennymi. Jeśli w parze badanych zmiennych jesteśmy w stanie wskazać na podstawie wiedzy źródłowej lub pozaźródłowej zmienną niezależną i zmienną

[32] Szerzej na ten temat: J. Buttolph Johnson, H.T. Reynolds, J.D. Mycoff, *Metody badawcze w naukach politycznych*, Wydawnictwo Naukowe PWN, Warszawa 2010, s. 470 i n.

zależną, wówczas wartość współczynnika odczytujemy odpowiednio z drugiego lub trzeciego wiersza zatytułowanego nazwą zmiennej, którą uznajemy za zmienną zależną. Jeśli natomiast nie zakładamy, która ze zmiennych jest zależną, a która niezależną, wówczas wartość współczynnika korelacji odczytujemy z pierwszego wiersza zatytułowanego *symmetric*.

13.2.2.5. Współczynnik niepewności U Theila

Współczynnik niepewności U Theila nazywany jest również współczynnikiem entropii. Został opracowany przez duńskiego ekonometrę Henri'ego Theila. Jest to mniej znana alternatywa dla stosowanego na szerszą skalę współczynnika zaproponowanego przez włoskiego statystyka i demografa Corrado Gini'ego. Współczynnik ten wykorzystywany jest przede wszystkim w ekonometrii, aczkolwiek przy odrobinie pomysłowości i dobrej woli można próbować stosować go także w politologii. Miara ta służy do oceny stopnia dokładności poczynionych przewidywań. Na jej podstawie możemy stwierdzić czy przewidywania są miarodajne, czy nie. Porównujemy dwie zmienne: wartości przewidywane i wartości rzeczywiste. Współczynnik pozwala na standaryzowaną ocenę stopnia adekwatności przewidywania. Przykładowo, w politologii, współczynnik mógłby posłużyć do porównania, w jakim stopniu różne ośrodki badawcze szacują wyniki wyborów. Skupmy się na frekwencji wyborczej jako najprostszym przykładzie zastosowania tego współczynnika. Możemy porównać frekwencję wyborczą przewidywaną przez ośrodek badawczy w danych wyborach z faktyczną frekwencją wyborczą. Wyniki przewidywane i wyniki faktycznie odnotowane w danych wyborach będą tworzyły pary zmiennych. Zmienne mogą być zapisane zarówno jako zmienne nominalne (trafne przewidywanie vs. nietrafne przewidywanie), jak też mogą być mierzone na wyższych poziomach (np. odsetka frekwencji).

Współczynnik niepewności wskazuje, w jakim stopniu jedna zmienna może być wykorzystywana do redukcji błędu podczas przewidywania wartości drugiej zmiennej. Współczynnik przybiera wartości od zera do jedności. Wartość współczynnika niepewności równa zero oznacza, że jedna zmienna nie może posłużyć do wyjaśniania drugiej, a wartość równa jeden oznacza, że jedną zmienną możemy całkowicie przewidywać na podstawie drugiej. Na przykład, wartość współczynnika 0,20 oznacza, że wiedza o wartościach przyjmowanych przez jedną zmienną zmniejsza błąd w przewidywaniu wartości drugiej zmiennej o 20 proc. Interpretując współczynnik niepewności, bierzemy pod uwagę także istotności. Wynik jest istotny statystycznie, jeśli p jest mniejsze od 0,05.

W programie PSPP współczynnik niepewności U Theila obliczamy, wybierając z menu tekstowego *Analyze* ⇨ *Descriptive Statistics* ⇨ *Crosstabs* ⇨ *Statistics,* a następnie zaznaczamy UC.

13.2.2.6. Współczynnik V Craméra

Współczynnik V Craméra (współczynnik kontyngencji V Craméra lub phi Craméra) został nazwany od jego twórcy Haralda Craméra szwedzkiego statystyka i aktuariusza. Miara ta została ogłoszona po raz pierwszy w 1946 roku[33]. Współczynnik ten służy do pomiaru siły zależności pomiędzy dwiema zmiennymi jakościowymi mierzonymi na poziomach nominalnych (można używać tego współczynnika do pomiaru zależności między zmiennymi mierzonymi na wyższych poziomach, lecz lepiej jednak skorzystać

[33] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton 1946, s. 282.

z doskonalszych, dedykowanych miar dla konkretnych poziomów). Stanowi on uzupełnienie testu chi-kwadrat w dwojakim sensie: po pierwsze, bazuje na nim w obliczeniach, a po drugie, z wyliczonego chi-kwadrat czerpie informację, czy związek jest statystycznie istotny. Zatem obliczając chi-kwadrat, a następnie V Craméra badacz otrzymuje dwie ważne informacje: o istnieniu lub wykluczeniu istotnej statystycznie zależności między zmiennymi (chi-kwadrat) oraz sile tego związku (V Craméra).

Współczynnik ten obliczamy następująco (podany wzór jednocześnie ujawnia związek V Craméra z chi-kwadrat):

$$V = \sqrt{\frac{\chi^2 * N}{m - 1}}$$

gdzie m oznacza liczbę kolumn lub liczbę wierszy w tabeli (bierzemy pod uwagę mniejszą liczebność), a N oznacza liczbę wszystkich jednostek analizy. Wartość tak obliczonego współczynnika zawiera się pomiędzy 0 a 1. Współczynnik ten pozwala na interpretowanie wyników zarówno skrajnych, jak też pośrednich. Im wartość współczynnika jest niższa, tym siła zależności pomiędzy badanymi cechami jest mniejsza, a im bliższa jedności, tym siła zależności jest większa. Przyjmuje się następującą jakościową interpretację współczynnika V: jeśli wynik testu jest mniejszy niż 0,3 - oznacza to słabą zależność. O zależności umiarkowanej mówimy, gdy wartości współczynnika zawierają się powyżej 0,3, lecz poniżej 0,5. Z kolei wartość współczynnika powyżej 0,5 świadczy o silnej zależności. Pamiętajmy o ważnej zasadzie: sensowne jest obliczanie współczynnika V Craméra wtedy i tylko wtedy jeśli dzięki testowi chi-kwadrat stwierdzimy istnienie zależności pomiędzy badanymi zmiennymi.

Wykorzystajmy przykład przytoczony w podrozdziale 13.2.2.1.2. Wartość obliczonego testu chi-kwadrat dla zmiennej płeć oraz oddania głosu na polityka X wyniosła 5,424, a zatem:

$$V = \sqrt{\frac{5{,}424 * 60}{2 - 1}} = 0{,}3$$

W powyższym przypadku możemy mówić o zależności słabej (na granicy umiarkowanego) pomiędzy dwiema zmiennymi.

W programie PSPP współczynnika V Craméra obliczamy wykonując wszystkie konieczne kroki dla obliczenia statystyki chi-kwadrat, a dodatkowo oznaczając w *Analyze ⇨ Descriptive Statistics ⇨ Crosstabs ⇨ Statistics ⇨ Phi*. Zaznaczenie współczynnika Phi jednocześnie powoduje obliczenie V Craméra.

W trybie poleceń języka skryptowego PSPP współczynnik obliczamy następująco:

```
CROSSTABS
/TABLES= V1 BY V2
/FORMAT=AVALUE TABLES PIVOT
/STATISTICS=CHISQ PHI
/CELLS=COUNT ROW COLUMN TOTAL EXPECTED RESIDUAL.
```

Rozkaz PHI powoduje jednocześnie wyświetlenie współczynnika V Craméra, podobnie jak w trybie okienkowym.

14

Rozdział 14. Regresja liniowa - elementarna metoda predykcji statystycznej

W najprostszym ujęciu statystyczna metoda regresji umożliwia przewidywanie wartości jednej zmiennej (lub wielu zmiennych) na podstawie obserwacji innego zjawiska (pomiaru innych zmiennych). W tym sensie jest to najszlachetniejsza z metod analitycznych, bowiem umożliwia ona osiągnięcie tego, co jest jednym z celów wszystkich nauk a mianowicie - przewidywania.

14.1. Rys historyczny analizy regresji

Wynalezienie metody regresji liniowej zawdzięczamy przede wszystkim Francisowi Galtonowi (1822-1911) oraz Karlowi Pearsonowi (1851-1936). Pierwszy z nich opracował podstawową koncepcję tej metody oraz wykreślił linię regresji. Z kolei K. Pearson jako biograf F. Galtona (wydał poświęcone temu uczonemu czterotomowe dzieło) opisał zasadę regresji, a następnie sformalizował tę miarę oraz wyposażył ją w solidne podstawy matematyczne w postaci współczynnika korelacji momentu iloczynowego (*Pearson product moment correlation,* PPMC). Pod względem matematycznym wzbogacił tę metodę szkocki statystyk George Udny Yule (1871-1951). Podstawy metody matematycznej, na której opiera się regresja, powstały jednak znacznie wcześniej. Metoda najmniejszych kwadratów - bo ona stanowi jej istotę - została opisana przez Adriena-Marie Legendre'a (1752-1833) w 1805 roku oraz - niezależnie - przez Carla Friedricha Gaussa (1777-1855) w 1809 roku, a w ugruntowanej formie - przez ostatniego z wymienionych w 1821 roku. Użycie samego pojęcia „regresja" zawdzięczamy F. Galtonowi. Zostało ono jednak pierwotnie wykorzystane jako opis pewnego zaobserwowanego zjawiska, a nie określenie metody statystycznej podjętego przezeń badania. F. Galton odkrył, że powszechną zasadę dziedziczenia stanowi zmniejszanie się gabarytów potomstwa rodziców o wielkości ponadprzeciętnej do średnich wielkości charakterystycznych dla osobników danego gatunku. Zjawisko to powszechnie znane jest w badaniach przyrodniczych jako regresja do średniej (*regression toward the mean, regression to the mean*) lub powracanie do przeciętności (*regression to mediocrity*). W tym kontekście pojęcie regresji zostało użyte zgodnie z pierwotnym, łacińskim jego znaczeniem - „cofanie się" (od *regressio, regressus*) lub

„cofać się" (od *regredi*) lub anglojęzycznym *to regrees* lub *to revert*. Opis wyniku tego odkrycia został następnie zastosowany przez F. Galtona do nazwania samej procedury.

Warto przyjrzeć się okolicznościom powstania koncepcji regresji. Jak wspomniano pomysłodawcą tej miary statystycznej jest przede wszystkim F. Galton, który opracował ją na potrzeby badań eugenicznych, w których badał prawidłowości dziedziczenia cech; interesowało go jak silnie charakterystyki w jednym pokoleniu istot żywych manifestują się w kolejnych. Prowadzony przezeń eksperyment, mający na celu zbadanie zakresu dziedziczności cech, przyniósł oprócz potwierdzenia tej hipotezy również wynalezienie statystycznej miary regresji. F. Galton rozesłał do swoich przyjaciół paczki nasion groszku pachnącego (*Lathyrus odoratus* L.) zmierzywszy wcześniej rozmiary tych nasion i sklasyfikował je tak, by każdy z adresatów otrzymał nasiona jednorodnej klasy wielkości. Przyjaciele nasiona te zasiali, a potomne nasiona odesłali. F. Galton mierzył otrzymane nasiona i nanosił na dwuwymiarowy diagram wyniki - na osi rzędnych zaznaczał wielkość nasiona rodzica, a na osi odciętych - nasiona potomnego. Okazało się, że dadzą się wydobyć pewne prawidłowości - wyniki te układały się wzdłuż pewnej linii - nazwanej później linią regresji. Praca F. Galtona, zawierająca takie graficzne obrazowanie wyników, została opublikowana w 1887 roku[1]. Z kolei po raz pierwszy pojęcia regresji użył F. Galton w dziele z 1885 roku zatytułowanym *Regression towards Mediocrity in Hereditary Stature*.

14.2. Teoretyczne podstawy analizy regresji

Zrozumienie tej metody statystycznej wymaga opanowania jej matematycznych podstaw. Na wstępie konieczne wydaje się wprowadzenie pary pojęć: zmiennej zależnej i zmiennej niezależnej. Pojęcia te, w najprostszym rozumieniu, wpisują się w logiczny schemat rozumowania przyczynowo-skutkowego. Zjawisko uznawane za przyczynę nazywamy **zmienną niezależną** lub zmienną wyjaśniającą, a zjawisko uważane za skutek nazywamy **zmienną zależną** (wyjaśnianą). W metodzie regresji stosuje się terminy zastępcze dla zmiennej zależnej i zmiennej niezależnej. Tę pierwszą nazywa się zmienną przewidywaną, a drugą określa mianem predyktora[2].

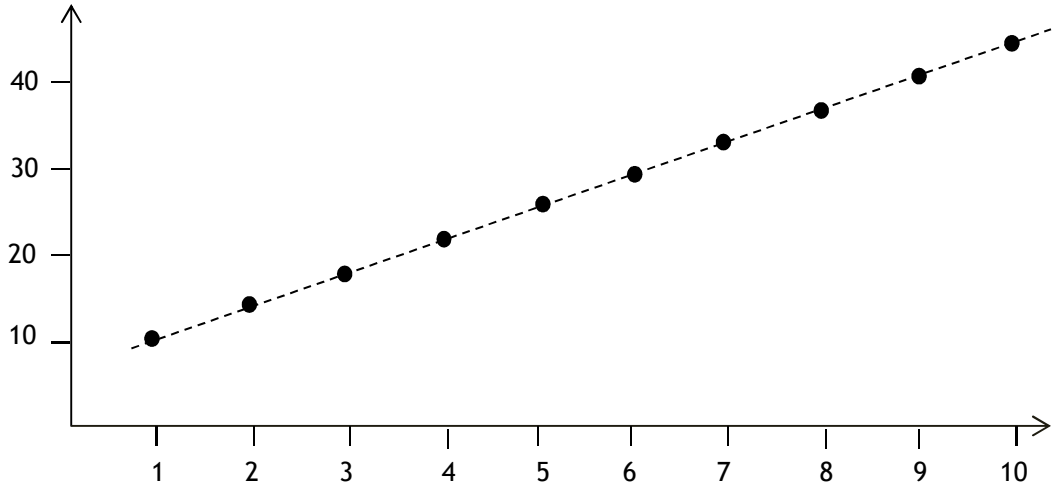
Analiza regresji oparta jest na bardzo prostym założeniu: na dwuwymiarowy diagram (zwany też wykresem korelacyjnym lub punktowym wykresem rozrzutu) na osi rzędnych umieszczamy skalę dla zmiennej zależnej, a na osi odciętych dla zmiennej niezależnej. Na wykres nanosimy odpowiednie wartości, jakie uzyskuje zmienna zależna przy kolejnych zmianach zmiennej niezależnej. W im większym stopniu zmiany zmiennej zależnej następują pod wpływem zmiennej niezależnej, w tym większym stopniu naniesione na diagram punkty układają się w liniowy wzór (rosnąco lub malejąco). Przyjmijmy, że zależność ta ma charakter liniowy, bowiem takie jest założenie najprostszej miary regresji - regresji liniowej. Jeśli na skutek zwiększania się zmiennej niezależnej mamy do czynienia ze zwiększaniem się zmiennej zależnej, mówimy, że jest ona **stymulantą**; istnieje wówczas dodatnia korelacja ze zmienną objaśnianą. Zależność tę przedstawiono na wykresie 5. Z kolei jeśli wzrost wartości zmiennej niezależnej prowadzi do zmniejszania się wartości zmiennej zależnej, mówimy, iż zmienna zależna jest **destymulantą**.

[1] Interesujące fakty na temat historii współczynnika korelacji i regresji podaje Jeffrey M. Stanton: J.M. Stanton, *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*, „Journal of Statistics Education", 2001, 9 (3).

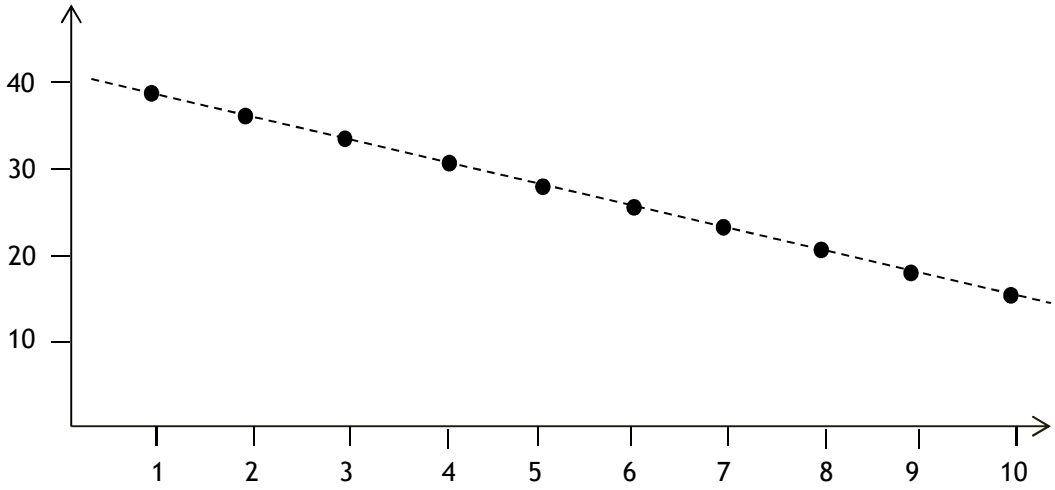
[2] Warto zwrócić uwagę, że w literaturze przedmiotu można zetknąć się również z innym jeszcze nazewnictwem: synonimem zmiennej wyjaśnianej są takie pojęcia jak zmienna endogeniczna lub wewnętrzna czy zmienna odpowiedzi, z kolei pojęciami tożsamymi z pojęciem zmiennej wyjaśniającej są: zmienna egzogeniczna oraz zewnętrzna.

Ten typ sytuacji przedstawia wykres 6. Zmienna wyjaśniająca (niezależna) może być również **neutralna,** jeśli nie wpływa ona na zmienną zależną. Idealnym przypadkiem takiej sytuacji byłaby jednolita wartość zmiennej zależnej niezależnie od zmian wartości zmiennej niezależnej (wówczas na diagramie pojawi się linia prosta równoległa do osi odciętych).

Wykres 5. Zmienna niezależna jako stymulanta zmiennej zależnej



Wykres 6. Zmienna niezależna jako destymulanta zmiennej niezależnej



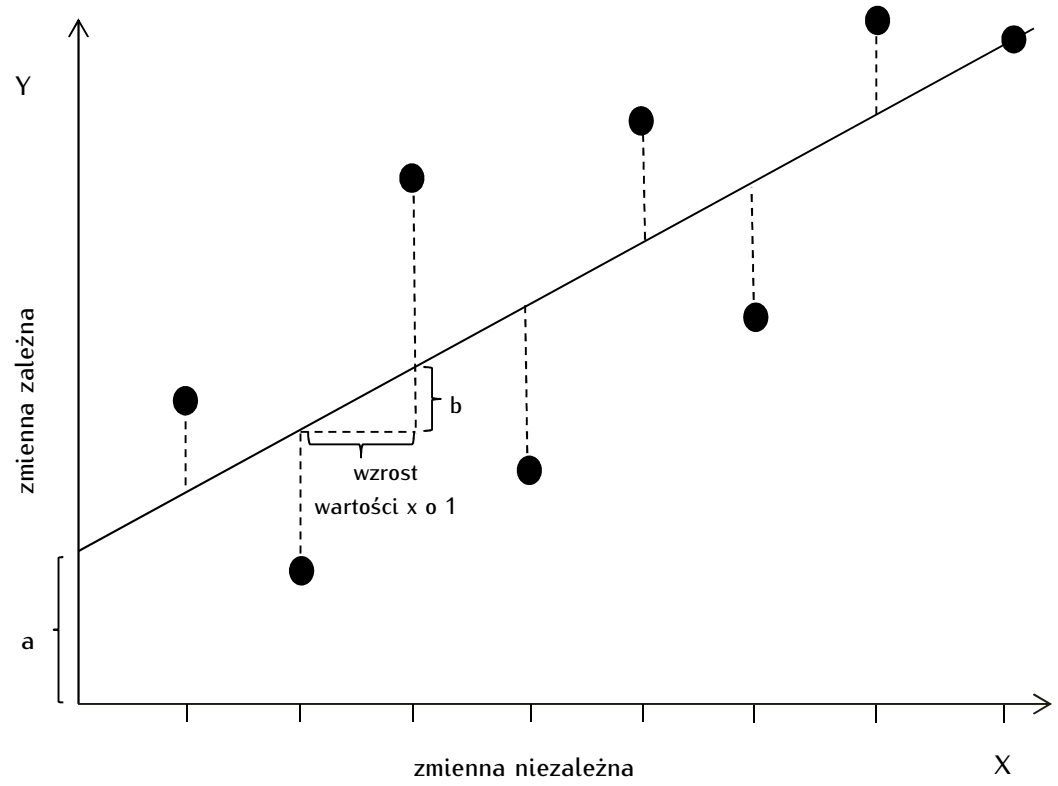
O przedstawionej wyżej zależności w języku matematyki mówimy, że zmienna zależna jest funkcją zmiennej niezależnej i wyrażamy następującym wzorem:

$$y = a + bx$$

gdzie y oznacza przewidywaną wartość zmiennej zależnej, a jest to współczynnik przesunięcia (*intercept*), czyli wartość jaką przyjmie y w momencie gdy x równe jest zero (a więc wartość y w punkcie przecięcia przez linię łączącą punkty osi rzędnych), b - współczynnik nachylenia (*slope*) lub współczynnik

kierunkowy, który wyraża zależność pomiędzy y a x, wskazując, o ile jednostek wzrośnie (lub zmaleje) wartość zmiennej zależnej, gdy wartość zmiennej niezależnej wzrośnie o jednostkę. Współczynnik nachylenia nazywany jest także współczynnikiem regresji i stanowi on najistotniejszy element omawianego równania. Wartości te zostały oznaczone na wykresie 7.

Wykres 7. Typowy, realny rozkład wyników dla zmiennej zależnej i zmiennej niezależnej wraz z linią regresji



W nauce, a w szczególności w naukach społecznych niemal nigdy nie mamy do czynienia z taką prostą zależnością jak przedstawione na wykresie 5 i wykresie 6. Zmienna zależna poddaje się innym niż zmienna niezależna czynnikom wpływu, pomiar nie jest i nie może być w pełni precyzyjny, a ponadto dają o sobie znać rozmaite zakłócenia i błędy. Typową, realną sytuację, w której wyniki pomiaru odchylają się od linii idealnej, obrazuje wykres 7. Linia prosta łącząca wszystkie punkty nie jest możliwa w takim przypadku do wykreślenia. Przewidywanie możliwe jest wtedy i tylko wtedy, gdy wykreślimy hipotetyczną linię prostą minimalizującą błąd pomiaru dla każdego z punktów. Błąd pomiaru to odchylenie wartości wyników pomiaru od hipotetycznej linii powstałych na skutek wymienionych wyżej czynników. Nazywa się je **błędami oszacowania** lub **resztami** (*residuals*). Z formalnego punktu widzenia oznacza to, że równanie regresji musimy uzupełnić o czynnik reszt. Przyjmie ono w związku z tym formę następującą:

$$y = a + bx \pm e$$

gdzie e oznacza błąd oszacowania (resztę), którego wartość oczekiwana jest równa zero (dążymy do jej minimalizacji). Na wykresie 7. wartości błędów obrazowane są przez przerywane linie równoległe do osi rzędnych biegnące pomiędzy punktem oznaczającym wynik pomiaru a hipotetyczną linią regresji. Zwróćmy uwagę, że na wykresie 7. tylko jeden z wyników pomiaru położony jest na tej hipotetycznej linii. Innymi słowy linia obrazuje nam typ idealny zależności. Linia ta powstaje na podstawie

rozrzuconych na wykresie realnych wyników pomiaru, które - jak wskazano - odbiegają od ideału. Nazywamy ją linią regresji (została ona naniesiona na wszystkie trzy prezentowane wykresy), a wykreśla się ją stosując matematyczną **metodę najmniejszych kwadratów.** Metoda ta pozwala na wykreślenie linii takiej, aby suma błędów oszacowania dla wszystkich wyników pomiaru była jak najmniejsza. Potęgowanie (a więc owe „kwadraty" w nazwie) bierze się stąd, że (jak widzimy na wykresie 7) część wyników pomiaru leży nad (a więc ma wartość dodatnią), a część pod linią regresji (zatem ich wartość jest ujemna). Zatem, gdyby je po prostu zsumować, wynik byłby równy zero. Stąd też zastosowano potęgowanie - dla usunięcia znaku poprzedzającego wartość błędu. Dzięki zastosowaniu tej metody można wykreślić zoptymalizowaną linię regresji dopasowaną do rozrzutu punktów. Wzór dla omawianej metody najmniejszych kwadratów dostępny jest w każdym standardowym podręczniku statystyki i ma postać następującą:

$$\chi^2 = \left(\frac{y_1 - ax_1 - b}{\sigma_1}\right)^2 + \left(\frac{y_2 - ax_2 - b}{\sigma_2}\right)^2 + \cdots + \left(\frac{y_n - ax_n - b}{\sigma_n}\right)^2$$

gdzie σ oznacza odchylenie standardowe (niepewność pomiaru danego punktu pomiarowego w zmiennej y).

Metoda najmniejszych kwadratów dla prostej regresji liniowej została szczegółowo opracowana i jest prezentowana w podręcznikach statystyki, jednak dla regresji krzywoliniowej nie ma takich standardowych wzorów i każdorazowo należy dokonywać dopasowania metodą prób i błędów.

Regresja cechuje się elegancją i prostotą. Ponadto jest ona metodą szczegółową i konkretną. Związek przyczynowy, który badamy za pomocą tej miary, daje się przedstawić za pomocą nieskomplikowanego równania matematycznego, tak jak ma to miejsce na przykład w fizyce, gdzie siła jest iloczynem masy i przyspieszenia danego ciała. Herbert M. Blalock mówi tu wręcz o wywodzeniu swoistych „praw" nauki na mocy równania regresji[3]. Zastrzec należy, że „prawa" te nie są tak precyzyjne w naukach społecznych jak w naukach przyrodniczych. Z kolei prostota tej miary bierze się stąd, że można czytelnie, nawet dla laika, przedstawić na dwuwymiarowym diagramie związek pomiędzy zmiennymi, dlatego interpretuje się ją niemal intuicyjnie. Z kolei szczegółowość objawia się tym, że na podstawie serii dokonanych pomiarów możemy obliczać wartości niepomierzone, a więc dokonywać ich predykcji, a co więcej czynić to w precyzyjnych kategoriach liczbowych. Ponadto możemy dowiedzieć się czy istnieje jakiś związek pomiędzy zmiennymi, a także jaki jest kierunek tego związku (dodatni jeśli zwiększaniu się zmiennej niezależnej towarzyszy zwiększanie się zmiennej zależnej czy ujemny - jeśli zwiększaniu się pierwszej towarzyszy zmniejszanie się wymienionej jako drugiej). Uzyskujemy również informację, w jakim stopniu jeden czynnik jest determinowany przez drugi, a w jakim zakresie wpływają nań inne czynniki. Wreszcie, wiemy jak wartość zmiennej zależnej będzie się zmieniała wraz ze zmianą zmiennej niezależnej.

Nie jest to jednak cudowna metoda, pozwalająca na odkrywanie prawidłowości życia społecznego i przewidywanie na podstawie jednego lub serii pomiarów. Pomiar w politologii nie jest jeszcze ugruntowany, nie wypracowano odpowiednich narzędzi pomiaru, brakuje rzetelnych i trafnych skal mierzących zjawiska społeczne na poziomie ilościowym. Jeśli nawet udaje się tego dokonać, to wyniki tego typu pomiarów nie są tak precyzyjne, jednoznaczne, uniwersalne i powtarzalne jak w przypadku pomiarów w fizyce.

[3] H.M. Blalock, *Statystyka dla socjologów,* Państwowe Wydawnictwo Naukowe, Warszawa 1977, s. 307.

Poczynić należy jeszcze jedno ważne zastrzeżenie - wykładana w tym rozdziale wiedza jest elementarna i stanowi ona zaledwie wprowadzenie do licznego zbioru metod regresji.

14.3. Analiza regresji liniowej (klasycznej) w programie PSPP

Przeprowadzenie analiz z użyciem metody regresji jest złożone, wymaga umiejętności oceny zmiennych pod kątem spełniania warunków dla przeprowadzenia takich analiz oraz interpretacji licznych statystycznych wskaźników. Analizę regresji liniowej przeprowadza się w dwóch następujących etapach:

1/ oceny zmiennych w zakresie spełniania warunków dla przeprowadzania analizy regresji obejmującej merytoryczne przesłanki wyboru zmiennych, testowanie normalności rozkładu wybranych zmiennych oraz poziomu ich pomiaru, a także wstępną wizualną ocenę modelu regresji,

2/ obliczenia regresji liniowej, na którą składają się: testowanie modelu regresji liniowej za pomocą jednoczynnikowej analizy wariancji (ANOVA), analiza parametrów modelu regresji liniowej oraz jego istotności statystycznej oraz analiza dopasowania modelu regresji liniowej.

14.3.1. Ocena zmiennych w zakresie spełniania warunków dla przeprowadzenia analizy regresji

Przeprowadzenie analizy regresji wymaga spełnienia szeregu warunków koniecznych. Zostały one przedstawione poniżej. Opisano ponadto, w jaki sposób należy przeprowadzić procedury testowania w programie PSPP.

14.3.1.1. Merytoryczne przesłanki wyboru zmiennych do analizy regresji

Badacz samodzielnie wybiera zmienne, które będą przedmiotem jego badań. W związku z tym musi być on w stanie uzasadnić, że pomiędzy wybranymi przezeń zmiennymi istnieje związek. Zależność, którą być może odkryje, nie może być li tylko zależnością czysto matematyczną, ale logiczną i społeczną. Nie należy popadać w „fetyszyzm statystyczny" - wyniki choćby najbardziej wysublimowanych testów statystycznych nigdy nie zastąpią logicznego rozumowania, zdrowego rozsądku i krytycyzmu oraz oparcia badacza na solidnych faktach i obserwacjach.

14.3.1.2. Testowanie normalności rozkładu zmiennych

Przede wszystkim mierzone zmienne muszą posiadać tak zwany rozkład normalny. O normalności rozkładu zmiennych możemy wnioskować na podstawie: 1/ wizualnej oceny histogramu z naniesioną krzywą normalną, 2/ pomiaru współczynników skośności i kurtozy oraz 3/ testu Kołmogorowa-Smirnowa. W praktyce badawczej rozkład normalny ocenia się na podstawie jednocześnie wizualnej oceny histogramu i analizy współczynników kurtozy i skośności lub też wyłącznie na podstawie testu Kołmogorowa-Smirnowa.

Wizualna ocena histogramu z krzywą normalną. Jest to wstępna ocena wizualna i jako taka nie powinna być rozstrzygającą zarówno negatywnie, jak i pozytywnie. Konieczne jest wykonanie histogramów z nałożoną krzywą normalną dla każdej z badanych zmiennych - niezależnej (lub niezależnych) oraz zależnej. W menu *Frequencies* ⇨ *Analyze* wybieramy zmienne, które będziemy testować. Następnie klikamy przycisk *Charts* umiejscowiony w *Analyze* ⇨ *Frequencies* i w otwartym nowym oknie zaznaczamy: wykreśl histogramy (*Draw histograms*) oraz nałóż krzywą normalną (*Superimpose normal curve*). W efekcie pojawią się histogramy - odpowiednio dla każdej ze zmiennych. Podkreślić należy, że opieranie się wyłącznie na ocenie wizualnej jest mylące i nie może stać się jedyną podstawą wnioskowania.

Analiza współczynnika kurtozy i skośności. Normalność rozkładu za pomocą miar kurtozy i skośności można przetestować, oznaczając w *Analyze* ⇨ *Frequencies*, a następnie wybierając *Kurtosis* i *Skewness.* Przyjmuje się umownie na potrzeby rozmaitych testów statystycznych, w których wymogiem jest normalność rozkładu testowanej zmiennej, że rozkład normalny powinien cechować się miarami kurtozy i skośności zawierającymi się pomiędzy +1 a -1. Warto także na tym etapie analizy zwrócić uwagę na skrajne wartości testowanych zmiennych - zaburzają one znacząco wyniki testów i naruszają normalność rozkładu. W tym momencie badacz powinien zdecydować o usunięciu, pozostawieniu lub zastąpieniu metodą imputacji wartości odstających.

Test normalności rozkładu Kołmogorowa-Smirnowa. Test ten powinien być wykonywany z uwzględnieniem tak zwanej poprawki Lilleforsa, która jest obliczana, gdy nie znamy średniej lub odchylenia standardowego całej populacji. Test Kołmogorowa-Smirnowa można wykonać w PSPP wybierając w zakładce *Analyze* ⇨ *Non-Parametric Statistics* ⇨ *1-Sample K-S*. W *Test Distribution* zaznaczamy, że chcemy porównywać tę zmienne z rozkładem normalnym (zaznaczamy zatem *Normal*).

Interpretacja testu Kołmogorowa-Smirnowa wymaga uwzględniania dwóch wartości: statystyki Z oraz poziomu istotności (p). Test Kołmogorowa-Smirnowa opiera się na następujących dwóch hipotezach - zerowej (H_0) i alternatywnej (H_1):

H_0 - rozkład badanej cechy w populacji jest rozkładem normalnym,

H_1 - rozkład badanej cechy w populacji jest różny od rozkładu normalnego.

Jeśli istotność jest niższa niż powszechnie zakładany w naukach społecznych poziom α (domyślnie $\alpha=0,05$), wówczas przyjmujemy H_1. Jeśli jest wyższa lub równa, wówczas nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że rozkład jest normalny.

14.3.1.3. Testowanie poziomu pomiaru zmiennych

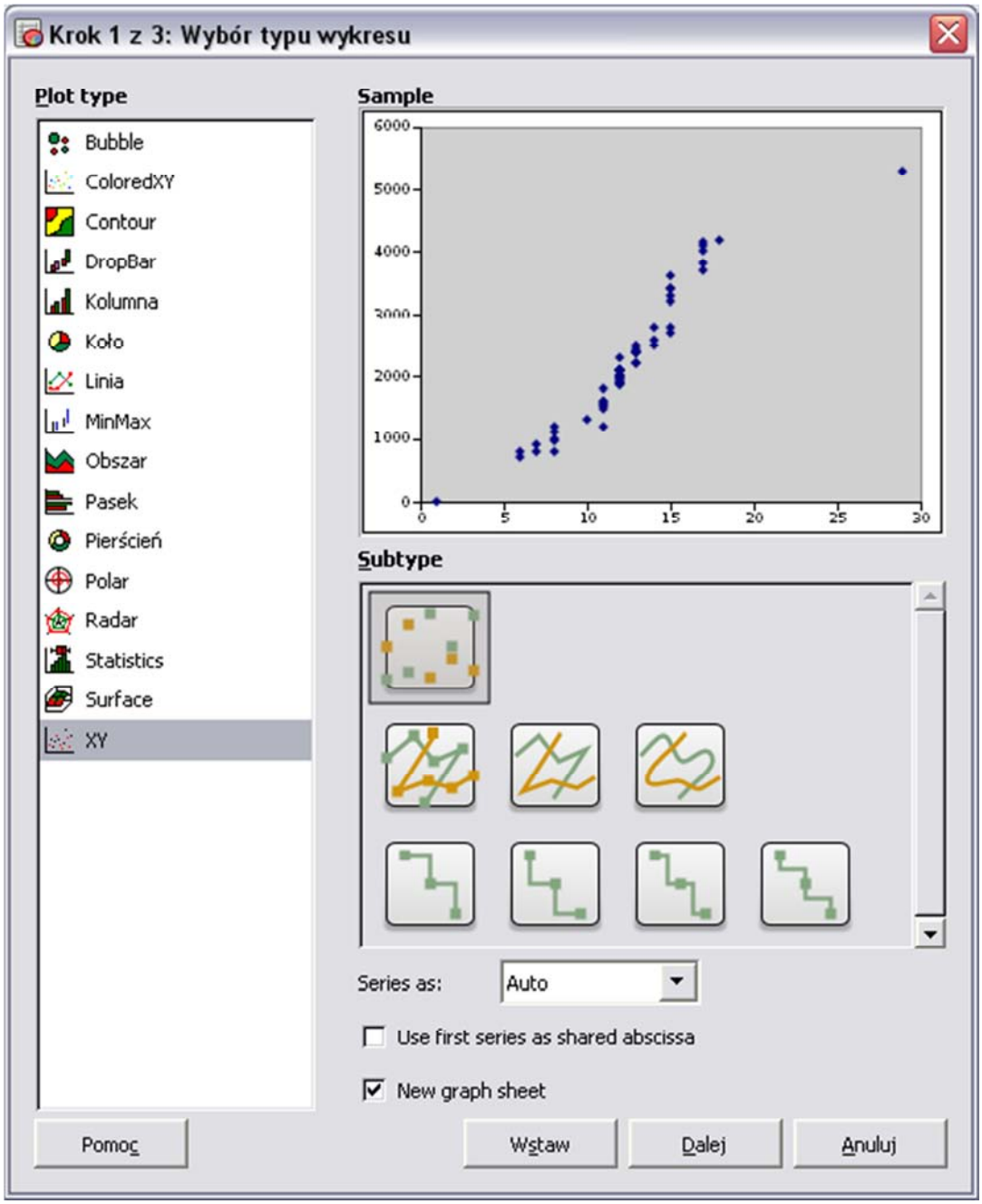
Koniecznym warunkiem dla przeprowadzenia analizy regresji jest wystarczający poziom pomiaru zmiennych. Obie zmienne - zależna i niezależna (lub niezależne w przypadku regresji wielozmiennowej) - muszą zostać zmierzone na poziomach silnych, a więc ilorazowym lub interwałowym. W wyjątkowych przypadkach - choć niektórzy będą traktowali to jako statystyczne nadużycie - stosuje się w regresji liniowej zmienne mierzone na poziomie porządkowym. Dopuszczalne jest także, aby jedna ze zmiennych była mierzona na poziomie nominalnym, ale tylko wówczas, jeśli jest ona zmienną dychotomiczną (a zatem zawiera dwie i tylko dwie kategorie oznaczające występowanie i niewystępowanie jakiegoś zjawiska lub cechy).

14.3.1.4. Minimalna liczebność próby w analizie regresji

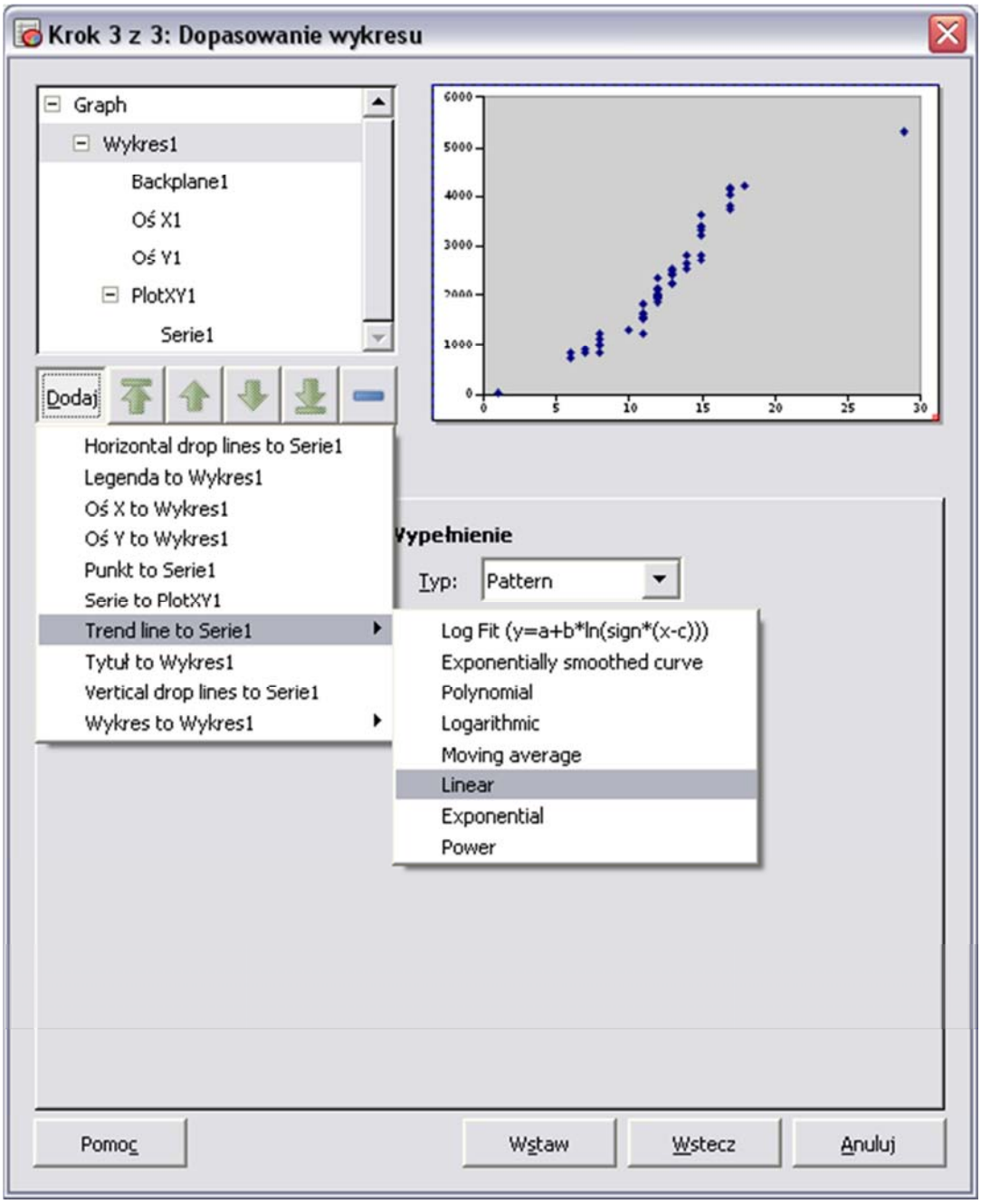
Minimalna liczebność w prostej regresji liniowej to 15 (wskazywana przez analityków wartość waha się między 10 a 20) jednostek analizy przypadających na jeden predyktor. W przypadku dwóch predyktorów minimalna liczebność to 30 jednostek, trzech - 45, i tak dalej. Niektórzy badacze wskazują, że minimalną liczebnością, wynikającą ze spełniania warunków centralnego twierdzenia granicznego, jest N=30. Od tej liczby rozkład danej zmiennej staje się rozkładem normalnym.

14.3.1.5. Wstępna wizualna ocena modelu regresji liniowej (ocena rozrzutu punktów wartości zmiennej zależnej i niezależnej na diagramie)

Ponadto należy poddać ocenie wizualnej diagram z rozrzutem wartości zmiennych. Ocena na podstawie współczynników jest niewystarczająca, ich wartości mogą być mylące. Program PSPP nie oferuje w wersji 0.7.9 możliwości tworzenia wykresów, jednak można je uzyskać na przykład w programie Gnumeric. Wizualna ocena rozkładu zmiennych jest nie do przecenienia - formalna poprawność wyników obliczeń dla regresji może być pozorna. Ponadto diagramy stanowią immanentną część raportu badawczego - powinny być w nim prezentowane. Z tego powodu wprowadzona zostaje instrukcja tworzenia wykresów z nałożoną krzywą regresji wraz ze wskazówkami interpretacyjnymi. Do stworzenia wykresów zostanie użyty program Gnumeric (wersja 1.10.12) będący rozbudowanym arkuszem kalkulacyjnym. Z programu PSPP kopiujemy kolumny ze zmienną zależną i niezależną wklejamy je do arkusza programu Gnumeric, a następnie wszystkie wklejone komórki zaznaczamy. Z menu tekstowego programu Gnumeric wybieramy *Wstaw* ⇨ *Wykres:*

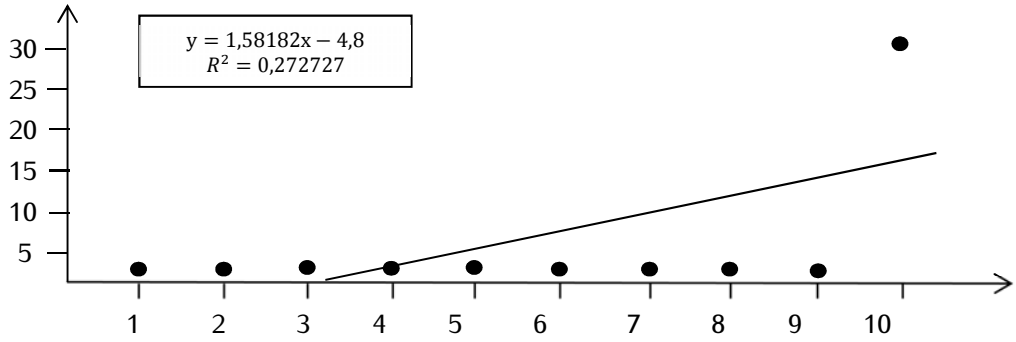


W oknie, które się wówczas pojawi należy wybrać ostatnią z opcji *Plot Type: XY.* Można również zaznaczyć, by wykres został wygenerowany w nowym arkuszu. Następnie klikamy przycisk *Dalej,* zaznaczamy *Wykres 1* ⇨ *Trend line to Serie1* ⇨ *Linear.* Wykonanie tej czynności nakłada linię regresji na wykres.



Po kliknięciu *Wstaw* generowany jest wykres z nałożoną krzywą regresji liniowej w nowej zakładce programu. Nieodzowność wizualnej weryfikacji rozrzutu zmiennych uzmysławiają przykłady zilustrowane wykresem 5 i 6. Większość wartości zmiennej zależnej (dziewięć na dziesięć) pozostaje bez zmian na poziomie 1. W jednym tylko przypadku zmienna ta przyjmuje wysoką wartość równą 30. Pomimo to linia regresji została nakreślona, a współczynnik dopasowania R^2 (patrz dalej) jest wysoki. Przewidywanie na podstawie tak skonstruowanego modelu byłoby błędne.

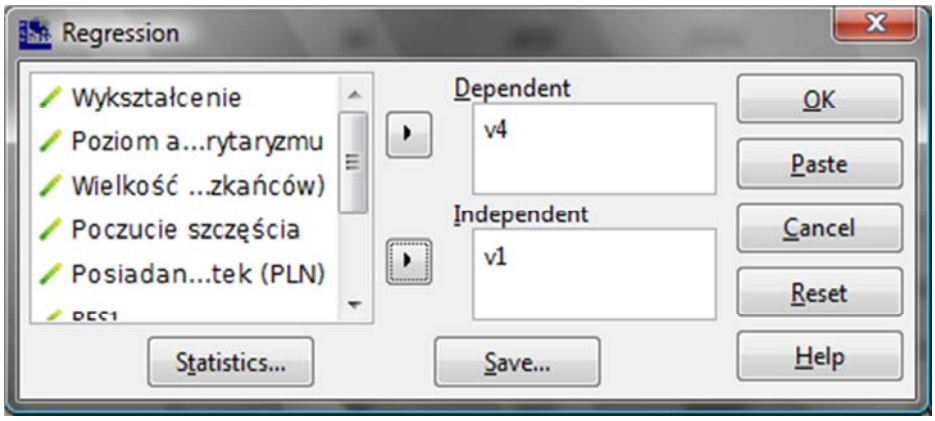
Wykres 8. Wadliwy model regresji liniowej



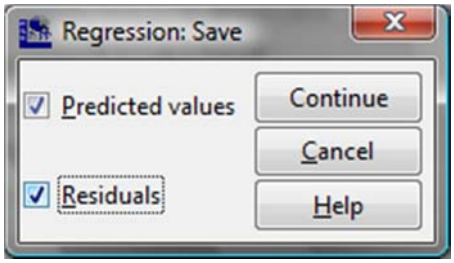
Niektórzy badacze badają również wariancję błędów dla poszczególnych poziomów zmiennej niezależnej. Wartość tej statystyki powinna być zbliżona do 2. Program PSPP w wersji 0.7.9. nie oferuje takiego testu.

14.3.2. Obliczanie i analiza regresji liniowej

W celu wykonania w programie PSPP analizy regresji wybieramy z zakładki *Analyze* ⇨ *Linear Regression*.



W polu *Dependent* umieszczamy zmienną zależną (wyjaśnianą, zmienną przewidywaną), a w polu *Independent* - zmienną niezależną (wyjaśniającą, predyktor). Jeśli w polu *Dependent* umieścimy więcej niż jedną zmienną, wówczas dla każdej z nich program wygeneruje komplet danych dla analizy regresji. Jeśli z kolei w polu *Independent* zostanie umieszczona więcej niż jedna zmienna, wykonana zostanie regresja wielozmiennowa. Przycisk *Save* umożliwia zapisanie wartości przewidywanych (*Predicted values*) oraz wartości reszt (błędów oszacowania) dla każdej zmiennej (*Residuals*). Wartości te zapisywane są na końcu zbioru danych jako odrębne zmienne.



Wykonanie analizy regresji może się odbyć także w trybie składni:

Składnia do wpisania w Edytorze	Opis działania składni
REGRESSION	- wykonaj analizę regresji
/VARIABLES= v1	- ustaw v1 jako zmienną niezależną
/DEPENDENT= v4	- ustaw v4 jako zmienną zależną
/STATISTICS=COEFF R ANOVA.	- wykonaj analizy konieczne dla analizy regresji: policz współczynniki parametrów, stopień dopasowania modelu R2 oraz wykonaj analizę wariancji ANOVA
/SAVE= PRED RESID.	- zapisz do zbioru danych wartości przewidywane oraz reszty (błędy oszacowania). Ta linia nie jest konieczna, można ją pominąć

Regresję liniową najprościej wyjaśnić, omawiając dwie proste zmienne: liczbę lat nauki (v1) oraz aktualne, comiesięczne zarobki podane w złotówkach (v4). Przyjmujemy, że zmienną niezależną (predyktorem) jest liczba lat nauki, a zmienną zależną (wyjaśnianą) są dochody. Przyjmujemy zatem jako naszą hipotezę, że wykształcenie istotnie różnicuje zarobki. Poniżej opisano trzy etapy interpretacji wyników analizy regresji dla tych dwóch zmiennych (przyjęto wartości fikcyjne dla potrzeb dydaktycznych).

14.3.2.1. Testowanie modelu regresji liniowej za pomocą analizy wariancji (ANOVA)

Analiza wariancji pozwala nam ocenić, jaka część zróżnicowania jest wyjaśniana w przyjętym liniowym modelu regresji, a jakiej przyjęta zmienna niezależna nie jest w stanie wyjaśnić (a więc w jakim stopniu na zmienną zależną wpływ mają inne czynniki niż predyktor). Im model regresji lepszy, to jest bardziej dopasowany, tym wariancja przypisana regresji powinna być większa, a wariancja przypisana resztom - mniejsza.

Wartości tej statystyki są prezentowane w tabeli zatytułowanej ANOVA:

ANOVA

	Sum of Squares	*df*	*Mean Square*	*F*	*Significance*
Regression	55428621	1	55428621	468,90	,00
Residual	6856172	58	118209,9		
Total	62284793	59			

W drugiej kolumnie (*Sum of Squares*) znajduje się suma wartości kwadratów wariancji wyjaśnionej (w pierwszym wierszu opisanym jako *Regression*) oraz wariancji niewyjaśnionej (w drugim wierszu - *Residual*). Liczba stopni swobody dla pierwszego wiersza to liczba predyktorów (w naszym przypadku 1), w drugim wierszu - liczba zmiennych obliczona na zasadzie n - 2, gdzie n oznacza liczbę jednostek analizy (a zatem: 60 - 2 = 58). Znajdujący się w trzeciej kolumnie średni kwadrat (*Mean of Squares*) został obliczony przez podzielenie wartości z kolumny drugiej przez liczbę stopni swobody w kolumnie trzeciej. Ta grupa statystyk służy nam tylko ogólnej ocenie: stwierdzamy, że wartości wyjaśniane przez model regresji są większe niż to, czego predyktor nie był w stanie wyjaśnić. W powyższym przykładzie tak właśnie jest. Następnie musimy ocenić, czy różnica ta jest statystycznie istotna. Oceniamy to na podstawie wartości znajdującej się w ostatniej kolumnie tabeli.

Interpretacja tej wartości opiera się na następujących dwóch hipotezach – zerowej (H_0) i alternatywnej (H_1):

H_0 – wariancja wyjaśniana przez model regresji **nie jest** istotnie pod względem statystycznym większa niż to, co jeszcze pozostało do wyjaśnienia (model nie może wyjaśnić poziomu zmiennej zależnej);

H_1 – wariancja wyjaśniana przez model regresji **jest** istotnie pod względem statystycznym większa niż to, co jeszcze pozostało do wyjaśnienia (model wyjaśnia poziom zmiennej zależnej).

Jeśli istotność jest niższa niż założony poziom α (domyślnie α=0,05), przyjmujemy H_1. Jeśli jest wyższa lub równa, to nie ma podstaw do odrzucenia H_0.

W przeprowadzanym teście wnioskujemy, że wariancja wyjaśniana przez model jest istotna statystycznie, model regresji z kolei tłumaczy zmienność analizowanej zmiennej zależnej.

14.3.2.2. Analiza parametrów modelu regresji liniowej oraz jego istotności statystycznej

W kolejnym kroku oceniamy parametry modelu regresji liniowej oraz jego istotność. W tabeli współczynników (*Coefficients*) interpretację możemy rozpocząć od sprawdzenia, czy znajdujące się w niej wyniki testów są istotne pod względem statystycznym. Jeśli w ostatniej kolumnie wartość poziomu istotności jest istotnie niższa niż zakładany w naukach społecznych poziom α= 0,05, wówczas stwierdzamy, że nie ma podstaw do odrzucenia twierdzenia o braku istotności statystycznej podanych w tabeli wyników. Tak właśnie jest w analizowanym przypadku. Zapis formalny poziomu istotności jest następujący p<0,001 (program PSPP nie pokazuje w *Significance* właściwej wartości).

Coefficients

	B	Std. Error	Beta	t	Significance
(Constant)	-916,31	151,47	,00	-6,05	,00
Liczba lat nauki	253,93	11,73	,94	21,65	,00

Następnie interpretujemy takie parametry równania regresji jak współczynnik przesunięcia (*intercept*), współczynnik nachylenia (*slope*) oraz ich błędy standardowe.

Współczynnik przesunięcia (w przytoczonym wyżej wzorze oznaczony literą a, natomiast w tabeli oznaczony jako B i znajdujący się w pierwszym wierszu tabeli) należy rozumieć następująco: przewidywana wartość zmiennej zależnej w momencie, gdy zmienna niezależna przyjmuje wartość równą zero wynosi -916,31. Jest to wartość nieinterpretowalna, bo ujemna. Nie możemy przecież powiedzieć, że na skutek braku wykształcenia (liczba lat nauki wynosi zero) będzie nam comiesięcznie odbierane 916,31 PLN. Wartość znajdująca się w analizowanej kolumnie w wierszu poniżej to z kolei współczynnik nachylenia (we wzorze został on oznaczony literą b). Jest to wartość, o jaką zwiększy się zmienna zależna, gdy zmienna niezależna wzrośnie o jedną jednostkę. Oznacza to, że wraz z każdym rokiem nauki można liczyć na wzrost zarobków o 253,93 PLN. Są to jednak wartości obarczone ryzykiem błędu. Potencjalną skalę pomyłki zawiera kolejna kolumna, w której podano błąd standardowy (*Standard Error*). Wartość w pierwszym wierszu oznacza błąd standardowy współczynnika przesunięcia, a w drugim – błąd standardowy współczynnika nachylenia. Interpretujemy go następująco: posiadając

jedynie wykształcenie podstawowe (8 lat nauki) i zawodowe (3 lata nauki) możemy liczyć na wynagrodzenie 1876,92 PLN (-916,31+(253,93*11)), jednak wynik ten obarczony jest ryzykiem błędu 151,47 PLN, co oznacza, że możemy spodziewać się dochodów w granicach od 1725,45 PLN do 2028,39 PLN. Podobnie rzecz ma się z przyrostem zarobków z każdym rokiem pobierania nauk - kwota ta może się potencjalnie wahać w granicach od 242,20 PLN do 265,66 PLN.

Podawany współczynnik Beta jest współczynnikiem nachylenia (znajdujący się w kolumnie drugiej, w drugim wierszu) wyrażonym w jednostkach odchylenia standardowego. Wartość tego współczynnika jest tożsama z wielkością korelacji pomiędzy zmienną zależną i niezależną. Ten współczynnik można interpretować analogicznie do R Pearsona lub rho Spearmana.

Współczynnika t znajdującego w przedostatniej kolumnie nie interpretujemy w linowej regresji z jednym predyktorem; ma on znaczenie przy regresji wielozmiennowej. Jest to wynik dzielenia współczynnika B przez błąd standardowy.

14.3.2.3. Analiza dopasowania modelu regresji liniowej

Najistotniejszym elementem regresji liniowej jest analiza dopasowania modelu. Rozpoczynamy ją od współczynnika R (tzw. korelacji wielokrotnej), który mówi nam, jak silnie zmienna niezależna związana jest ze zmienną zależną. Sposób jego rozumienia został wyłożony w rozdziale dotyczącym korelacji.

Model Summary

R	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
,94	,89	,89	343,82

Najważniejszy jest współczynnik R^2 (powstaje po podniesieniu współczynnika R do drugiej potęgi), nazywany współczynnikiem dopasowania modelu. Mówi on, jaki stopień wariancji jest wyjaśniany przez liniowy model regresji uzyskany metodą najmniejszych kwadratów. Przyjmuje się, że w naukach społecznych uwzględniamy i interpretujemy model, który wyjaśnia co najmniej 10 proc. wariancji. Współczynnik ten przyjmuje wartość z zakresu od 0 do 1, gdzie zero oznacza, że żadna część wariancji zmiennej zależnej nie jest wyjaśniana przez predyktor, a jeden - iż 100 proc. zmienności danej zmiennej wyjaśnia predyktor. Z kolei skorygowany R^2 (*Adjusted R Square*) odczytujemy zamiast poprzedniej kolumny w przypadku regresji wielozmiennowej. Przyjmuje on każdorazowo wartość niższą niż R^2. Ostatnia kolumna (*Standard Error of the Estimate*) wskazuje, jaki jest potencjalny błąd użycia uzyskanego modelu do celów prognostycznych. Przypuśćmy, że pragniemy sprawdzić, jakich zarobków można oczekiwać po 23 latach nauki (ukończenie szkoły podstawowej, pięcioletniego technikum, studiów wyższych, studiów doktoranckich oraz dwuletnich studiów podyplomowych). Do wzoru:

$$y = a + bx \pm e$$

podstawiamy wartości:

$$y = 916{,}31 + 253{,}93 * 23 \pm 343{,}82 = 6756{,}7 \pm 343{,}82$$

A zatem po 23 latach nauki możemy oczekiwać zarobków wahających się w granicach od 6412,88 PLN do 7100,52 PLN (tu ponownie należy zasygnalizować Czytelnikowi, że przytaczane dane są fikcyjne). Wykonana czynność nazywana jest predykcją punktową.

W statystyce używane jest czasami również pojęcie współczynnika zbieżności. Oznacza się go grecką literą ϕ^2 i oblicza następująco:

$$\phi^2 = 1 - R^2$$

Współczynnik ten przyjmuje wartości od zera do jedności i jako odwrotność współczynnika dopasowania modelu informuje, jaka część zaobserwowanej, całkowitej zmienności y nie została wyjaśniona przez model.

14.4. Regresja wielozmiennowa (wieloraka)

Regresja wielozmiennowa jest to postępowanie badawcze, polegające na włączeniu do analiz więcej niż jednego predyktora i pomiar stopnia, w jakim predyktory te łącznie wyjaśniają zmienność zmiennej niezależnej. W regresji wielozmiennowej uwidacznia się wyraźnie czynność budowania tak zwanego modelu statystycznego. Model statystyczny rozumiany jest jako hipoteza lub wiązka hipotez, sformułowanych wedle reguł statystycznych (w tym przypadku w postaci równania regresji), który ma za zadanie w sposób uproszczony (każdy model jest uproszczeniem rzeczywistości) obrazować i wyjaśniać zależności pomiędzy zjawiskami (zmiennymi), a także kwantyfikować czynnik losowy oraz czynnik nierozpoznany, a więc niewyjaśniany przez opracowany model (model wyraźnie wskazuje w formie liczbowej, w jakim stopniu wyjaśniane jest dzięki niemu dane zjawisko). Wszystkie te warunki spełnia analiza regresji (zarówno liniowa jak też wielozmiennowa). W regresji wielozmiennowej, jak wskazano, procedura ta uwidacznia się najpełniej - spośród wielu licznych zaobserwowanych zjawisk (zmiennych) selekcjonujemy te, o których wiemy, że wyjaśniają zmienność zmiennej zależnej, a więc innymi słowami konstruujemy swoisty model rzeczywistości.

Regresja wielozmiennowa przyjmuje postać następującą:

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$$

gdzie a oznacza współczynnik przesunięcia wskazujący wartość y, gdy x_1, x_2 oraz x_n będą równe zero. Wyrażenia b_nx_n to kolejne włączone do modelu predyktory. Współczynnikami regresji (nachylenia, kierunkowymi) są wartości b_1, b_2. i b_n.

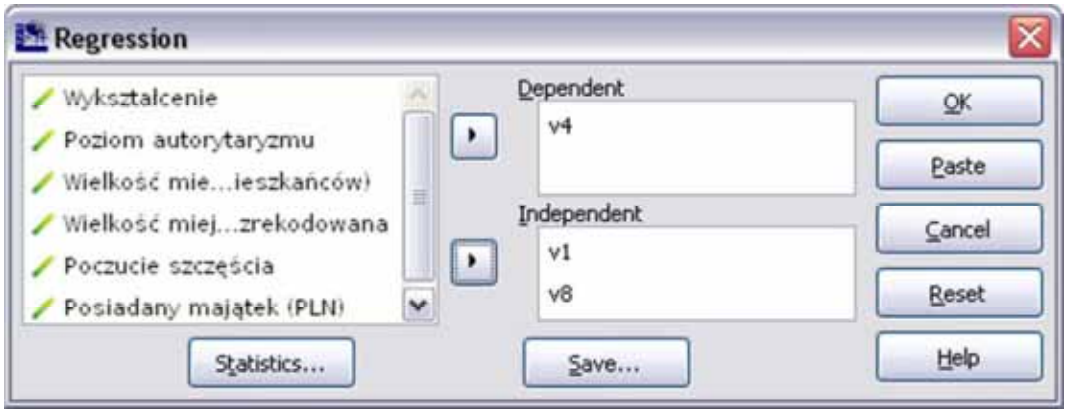
14.4.1. Warunki przeprowadzenia regresji wielozmiennowej

Analizę regresji wielozmiennowej prowadzimy wedle identycznych zasad co regresję jednoczynnikową (z jednym predyktorem). Analizowane zmienne muszą spełnić jednak dodatkowe kryteria. Po pierwsze, w regresji wielozmiennowej predyktory (zmienne niezależne) powinny być ze sobą nieskorelowane (sprawdzić można to za pomocą testu korelacji rho Spearmana). Ponadto postuluje się, by zmienne niezależne były mierzone na tożsamych skalach (chodzi o rzędy wielkości liczb) w celu uniknięcia zmarginalizowania zmiennych, których skale mają niski poziom wartości.

14.4.2. Obliczanie i analiza regresji wielozmiennowej

Przypuśćmy, że wykonaną w poprzednim podrozdziale regresję jednoczynnikową uznaliśmy za niewystarczającą. Chcielibyśmy lepiej wyjaśnić zróżnicowanie zarobków, rozumiemy bowiem, że ich wysokość nie jest prostą funkcją liczby lat nauki. Przewidujemy, że na poziom zarobków jednostki mogą mieć wpływ również inne zmienne.

Obliczanie regresji wielozmiennowej prowadzimy identycznie jak regresji jednoczynnikowej, z tą jednak różnicą, że włączamy do obliczeń więcej niż jedną niezależną:



Jako pierwszą włączamy zmienną, o której wiemy (na podstawie badań własnych lub na mocy rozstrzygnięć teoretycznych), że jest ona najsilniejszym predyktorem danego zjawiska. W rozważanym przykładzie będzie to liczba lat nauki oraz wiek respondenta w latach. W tym ostatnim przypadku możemy przewidywać, że im wyższy wiek i doświadczenie, tym wyższe zarobki lub odwrotnie - im niższy wiek, a co zatem idzie większa mobilność i lepsza adaptacja do rynku pracy, tym wyższe będą zarobki.

Uzyskujemy analogiczne statystyki jak w regresji jednoczynnikowej:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,95	,90	,90	324,52

ANOVA

	Sum of Squares	df	Mean Square	F	Significance
Regression	56281840	2	28140920	267,21	,00
Residual	6002953	57	105315,0		
Total	62284793	59			

Coefficients

	B	Std. Error	Beta	t	Significance
(Constant)	-108,16	296,70	,00	-,36	,72
Liczba lat nauki	230,66	13,09	,86	17,62	,00
Wiek w latach	-11,05	3,72	-,15	-2,97	,00

Interpretujemy je następująco: model regresji jest istotny statystycznie ($p<0,01$), a suma kwadratów dla części wyjaśnionej jest większa niż dla reszt (zwróćmy uwagę na różną, niż w przypadku regresji jednoczynnikowej, liczbę stopni swobody w wierszu pierwszym i drugim - włączyliśmy dwie zmienne i dlatego w pierwszym wierszu pojawia się liczba 2, a w drugim 57, obliczone na zasadzie n-3). Wiedząc, że model jest istotny przechodzimy do analizy tabeli zawierającej współczynniki. Widzimy, że z każdym kolejnym rokiem nauki możemy liczyć na dodatkowe 230,66 PLN miesięcznej pensji

(plus minus 13,09 PLN). Jednakże im starszy respondent, tym niższy jest poziom zarobków (każdy kolejny rok to utrata 11,05 PLN – plus minus 3,72 PLN). Zwróćmy uwagę na statystyki t – w tym miejscu otrzymujemy sygnał, że wiek w latach jest o wiele słabszym predyktorem niż liczba lat nauki (niska wartość statystyki t). Całość modelu (patrz: *Model Summary*) wyjaśnia aż 90 proc. wariancji (taką sytuację badawczą należałoby uznać w naukach społecznych za znacznie więcej niż zadowalającą). Należy zaznaczyć, że w przypadku regresji wielozmiennowej wartość współczynnika dopasowania odczytujemy z trzeciej kolumny – skorygowany współczynnik R^2 (*Adjusted R Square*). Przypomnijmy, że sama liczba lat nauki wyjaśniała 89 proc. wariancji, a zatem dodatkowa zmienna w niewielkim zaledwie stopniu zwiększa dokładność predykcji. (Uwaga! Program PSPP w wersji 0.7.9. nie umożliwia automatycznego porównywania zwiększania się lub zmniejszania współczynnika R^2 pod wpływem kolejnych predyktorów).

Istnieją różne metody budowania modelu. Zwykle nie poprzestajemy na dwóch zmiennych, lecz testujemy je w rozmaitych układach i konfiguracjach, poszukując najlepiej dopasowanego modelu wyjaśniającego dane zjawisko (zmienną). Istnieją liczne metody budowania modelu regresji, najpopularniejsze z nich to metoda krokowa i metoda hierarchiczna. Najprostszą metodą jest metoda krokowa. Występuje ona w dwóch odmianach: metody krokowej postępującej oraz metody krokowej wstecznej. Są to w istocie techniki selekcji predyktorów w modelu. Pierwsza z nich polega na dołączaniu kolejnych predyktorów do modelu. Analiza rozpoczyna się od utworzenia modelu z jednym – najsilniej wyjaśniającym daną zmienną – predyktorem, kolejno dołączane są kolejne dopóty, dopóki zostaną dołączone wszystkie zmienne ze zbioru posiadające poziom istotności poniżej 0,05 i jednocześnie najwyższy poziom współczynnika R^2. Z kolei metoda krokowa wsteczna rozpoczyna się od włączenia do modelu wszystkich zmiennych ze zbioru, a następnie odrzucania tych, które są najmniej istotne, aż do momentu, w którym w zbiorze pozostaną zmienne posiadające istotność niższą niż $p<0,1$.

Część V. Elementy wnioskowania statystycznego

15

Rozdział 15. Wprowadzenie do wnioskowania statystycznego

Myślą przewodnią niniejszego rozdziału jest zaznajomienie początkującego badacza z podstawami wnioskowania statystycznego. Wnioskowanie statystyczne może przybrać dwie formy - **estymację** oraz **weryfikację**. Estymacja pozwala nam, zbadawszy tylko część jakiejś grupy (próbę), wnioskować o charakterystykach jej całości (populacji). Z kolei weryfikacja umożliwia matematyczne potwierdzenie istotności naszych przypuszczeń (hipotez) o występowaniu określonych zjawisk w przyrodzie i społeczeństwie. Uprzednio zapoznajemy jednak Czytelnika z elementami teorii próbkowania oraz z pojęciem rozkładu zmiennej losowej. Zagadnienia te stanowią bowiem teoretyczną podstawę wnioskowania statystycznego. Ponadto prezentujemy Czytelnikowi w przystępnej formie podstawowe pojęcia oraz założenia teoretyczne, których poznanie jest niezbędne dla zrozumienia istoty wnioskowania statystycznego i wykorzystywania testów istotności do jego przeprowadzania.

Ważną część rozdziału stanowi wykład na temat wyznaczania minimalnej liczebności badanych jednostek, co umożliwia nam scharakteryzowanie całych grup po zbadaniu tylko ich wybranej części. Czytelnik znajdzie tutaj również metody określania wielkości ryzyka popełnienia błędu, gdy chcemy przenieść wyniki częściowe na pewną całościową zbiorowość.

Materiał niniejszego rozdziału stanowi niezbędne wprowadzenie teoretyczne do praktycznego używania programu PSPP w zakresie statystyki indukcyjnej, określanej również mianem statystki matematycznej, wykorzystywanej we wnioskowaniu statystycznym.

15.1. Elementy teorii próbkowania i pojęcie rozkładu zmiennej losowej

Pozyskiwanie danych ilościowych wiąże się z koniecznością ich uporządkowania, opisania i zinterpretowania za pomocą odpowiednich metod i narzędzi statystycznych. Szczególnie ważne są dane uzyskiwane w toku badania wybranej grupy ludzi. Są one źródłem wiedzy na temat społeczeństwa oraz zjawisk w nim zachodzących. Badanie bowiem całych zbiorowości ludzkich jest wręcz zadaniem niemożliwym do zrealizowania. Wynika to z wielu przesłanek, wymieniając chociażby wysokie koszty ponoszone na takie przedsięwzięcia, niemożność dotarcia z pytaniami do wszystkich interesujących nas osób czy też zwyczajnie - odmowy uczestniczenia w badanach i niechęć udzielania informacji na nurtujące badacza zagadnienia. Praktyka badawcza wskazuje, że zbadanie pewnej wybranej grupy jednostek jest wystarczające, aby pozyskać wiedzę na temat pozostałych jej elementów. Możliwe jest to jednak wraz ze spełnieniem szeregu wymogów metodologicznych i statystycznych, a także za pośrednictwem wykorzystania określonych metod i narzędzi analitycznych. Zagadnienia te mieszczą się w zakresie poniżej omawianych elementów teorii próbkowania oraz pojęcia rozkładu zmiennej losowej. Przyswojenie podstawowej wiedzy w tym aspekcie jest niezbędnym etapem kształcenia i warunkiem umożliwiających zastosowanie elementów wnioskowania statystycznego w praktyce.

15.1.1. Próba losowa a populacja

Badania całych zbiorowości nazywane badaniami zupełnymi lub pełnymi zastępowane są przez badania częściowe. Obejmują one wyłącznie pewien fragment owej zbiorowości. Kiedy będziemy chcieli dowiedzieć się, jak dorośli Polacy oceniają rządy obecnego premiera bądź też w jakim stopniu ufają konkretnemu politykowi wystarczy, iż zadamy takie pytania tylko wybranej ich grupie. Zbiorowość wszystkich dorosłych Polaków będziemy nazywać w badaniach społecznych **populacją**, natomiast wybraną z niej grupę osób - **próbą**. Zadeklarowane sądy tego wyselekcjonowanego podzbioru ludzi mają odzwierciedlać opinie, które podzielają pozostałe osoby, nie objęte bezpośrednio badaniem. Wybór jednak jednostek do próby badawczej nigdy nie ma charakteru dowolnego. Podlega on zawsze specjalistycznym procedurom kontroli ich doboru w celu zagwarantowania **reprezentatywności** wyników, czyli możliwości przeniesienia naszych obliczeń, analiz oraz wniosków na całą zbiorowość Polaków. W badaniach zawsze staramy się, aby przebadany przez nas zbiór jednostek był reprezentatywny. Osiągnięcie tego celu wymaga spełnienia dwóch warunków. Po pierwsze, próba badawcza musi obejmować jednostki populacji, które zostały do niej włączone w sposób losowy. Po drugie, liczba tych jednostek musi być odpowiednio duża. W tym przypadku mówimy o tzw. liczebności próby badawczej. Jeżeli te kryteria są spełnione, możemy podjąć się wyzwania przeniesienia wyników badania na całą interesującą nas zbiorowość. W tym celu przeprowadzamy omawiane w niniejszym rozdziale wnioskowanie statystyczne. Konieczne jest jednak uprzednie ustalenie jednej kwestii - jakie prawdopodobieństwo błędu będą miały nasze obliczenia? Poniżej Czytelnik odnajdzie objaśnienie tego zagadnienia.

15.1.2. Błąd oszacowania z próby losowej

Początkujący badacz musi wyzbyć się złudzenia, iż wszelkie analizy przeprowadzane na danych z próby badawczej są doskonałe. Prawdopodobieństwo, że średnia wieku bądź przeciętny poziom zaufania do polityka X czy też partii Y obliczony z badania będzie dokładnie odzwierciedlał średnią wieku bądź przeciętne zaufanie w populacji, jest niemal równe zero. Każdorazowo bowiem jesteśmy narażeni na po-

pełnienie szeregu błędów związanych chociażby z niewłaściwym doborem jednostek do próby badawczej. O ile podstawowym zadaniem badacza jest dążenie do minimalizacji tego błędu, jego obliczenie i podanie maksymalnego poziomu, jaki on może przyjąć, jest niezbędnym elementem prezentowania wszelkich wyników badań. Błąd ten nazywamy maksymalnym bądź też dopuszczalnym błędem oszacowania (lub szacunku) i oznaczamy symbolem d. W teorii jest definiowany jako różnica między wartością statystyki z próby a wartością rzeczywistą szacowanego parametru w populacji. Błąd z próby wyliczany jest według różnych wzorów, w zależności od obliczanego estymatora. Możemy zatem wyliczać maksymalny błąd szacunku dla średniej z próby, odchylenia standardowego z próby lub frakcji z próby. W dużej mierze zależy to od obliczanego parametru populacji oraz typu analizowanej zmiennej.

Praktycznym i powszechnym zabiegiem jest obliczanie ogólnego błędu szacunku dla całej próby, a tym samym dla wszystkich objętych badaniem zmiennych. Do tego celu wykorzystujemy założenie o normalności rozkładu zmiennych losowych. Przyjmujemy również, iż poszczególne cechy mogą wystąpić w populacji z równym prawdopodobieństwem wynoszącym p=0,5. Do obliczenia maksymalnego błędu oszacowania korzystamy z następującego wzoru:

$$d = Z_{\alpha} * \sqrt{\frac{p * (1 - p)}{n}}$$

Określenie maksymalnego błędu oszacowania wymaga również ustalenia wartości poszczególnych miar. Wartość Z_{α} to wielkość wyznaczana według ustalonego arbitralnie poziomu ufności (α). Najczęściej spotykanymi wartościami dla α jest 0,1, 0,05 oraz 0,01. Przy szacowaniu błędu zakładamy zawsze dwustronną postać obszaru krytycznego dla rozkładu normalnego. Zgodnie z tymi założeniami dla α=0,1 statystyki Z_{α} wynoszą: $Z_{0,1}$=1,64, dla α=0,05 - $Z_{0,05}$=1,96, zaś dla α=0,01 - $Z_{0,01}$=2,58. Jeżeli nie mamy żadnego rozeznania na temat rozkładu cech w populacji, przyjmujemy p=0,5. Możemy przystąpić do obliczenia błędu. Dokonamy tego dla próby liczącej N=1800 z satysfakcjonującym nas poziomem ufności wynoszącym 0,05. Obliczenia wyglądają następująco:

$$d = 1{,}96 * \sqrt{\frac{0{,}5 * (1 - 0{,}5)}{1800}} = 1{,}96 * \sqrt{\frac{0{,}25}{1800}} = 1{,}96 * \sqrt{0{,}000139} = 0{,}023099$$

Dla próby N=1800 maksymalny błąd oszacowania wynosi 2,31 proc. przy poziomie ufności α=0,05. Błąd szacunku przyjmuje wartości bezwzględne od 0 do 1. Standardem jest podawanie jego wielkości w procentach. Błąd szacunku informuje nas, o ile procent różni się wartość statystyki z próby z faktyczną jej wartością w populacji i pozwala określić przedział, w którym mieści się wartość rzeczywista. Jeżeli zatem uznaliśmy, iż 40 proc. dorosłych Polaków odda w nadchodzących wyborach głos na partię X, zaś błąd oszacowania wynosi ± 2,31 proc., to określenie zakresu maksymalnych rozbieżności wymaga następujacych obliczeń:

$$40\ proc. * 2{,}31\ proc. = 0{,}4 * 0{,}231 = 0{,}09$$

Uzyskana wartość w przeliczeniu na wartości procentowe daje w przybliżeniu 1 proc. Maksymalny błąd oszacowania określa dopuszczalny zakres rozbieżności, co oznacza, że wartość rzeczywista może być mniejsza (-) lub większa (+). W naszym przykładzie różnica wynosi 1 proc., czyli wartość rzeczywista poparcia dla partii X powinna mieścić się w przedziale od 39 proc. do 41 proc.

15.1.3. Wybrane rozkłady teoretyczne zmiennej losowej

Zachodzące wokół nas masowe zjawiska przyrodnicze i społeczne - choć na pierwszy rzut oka wydają się chaotyczne - posiadają jednak pewne ukryte wzory. Są one losowe, ale nie znaczy to, że przypadkowe. Odnalezienie wśród tego zamętu - lecz również pozornego - ukrytych wzorów i prawidłowości umożliwia matematyka i statystyka. Dyscypliny te dają nam sposobność poskromienia przyrodniczego i społecznego chaosu z pozoru mętnych i niezrozumiałych zdarzeń.

Zjawiska przyrodnicze i społeczne pojawiające się w zbiorowościach odznaczają się określonymi właściwościami i cechami. Wyrażenie ich przy pomocy liczb umożliwia wyznaczenie pewnych tendencji oraz wzorów ich występowania w rzeczywistości. Statystyka i matematyka wypracowały zestaw narzędzi służących do określania prawdopodobnych wartości przyjmowanych przez te cechy oraz sposobu wyznaczanie częstości ich występowania. Nazywamy je **rozkładami teoretycznymi** bądź **rozkładami prawdopodobieństwa**. Określają one szansę przyjęcia danych wartości przez badaną cechę w grupie. Wyznaczamy je na podstawie analizy wyników z próby badawczej. Warto przypomnieć, iż próbą nazywamy pewną wybraną grupę jednostek. Jednostki te są dobranymi w ściśle określony sposób reprezentantami (pod względem wybranych przez nas cech) szerszej zbiorowości, z której pochodzą. Większość badań zjawisk przyrodniczych i społecznych prowadzi się na części, a nie całości interesującego badacza zbioru jednostek lub zjawisk. Trudno jest bowiem dotrzeć do wszystkich elementów z interesującej nas zbiorowości, a co więcej, dzięki narzędziom statystycznym, okazuje się to niewymagane. Taką pełną zbiorowość zjawisk lub jednostek nazywamy populacją. Dzięki zastosowaniu odpowiednich metod możemy zbadać wybraną z populacji grupę, a mimo to pozyskać wiedzę o wszystkich, pozostałych przypadkach. Przykładem populacji jest zbiorowość wszystkich dorosłych Polaków, z kolei próbę stanowić będzie wybrana z niej (w ściśle określony sposób) grupa jednostek, licząca na przykład 1000 osób.

Jak pamiętamy z części pierwszej podręcznika, do zaprezentowania rezultatów badania wykorzystywaliśmy tabele częstości oraz tabele krzyżowe. Dodatkowo wyniki próby służyły nam do wyliczania miar statystycznych, przykładowo średniej arytmetycznej, wariancji, odchylenia standardowego, mediany i dominanty. Taką formę prezentacji rozkładów danych z próby nazywamy **rozkładem empirycznym zmiennej**. Celem natomiast wnioskowania statystycznego jest stworzenie możliwości przeniesienia owych wyników na populację. Wówczas przeprowadzamy odpowiednie obliczenia, sprawdzając przy pomocy właściwych metod, zasadność ich uogólnienia na zbiorowość generalną. Procedurę tą nazywamy testowaniem zgodności rozkładu z próby z rozkładem teoretycznym zmiennej. W parze bowiem ze statystykami z próby idą rozkłady oczekiwane cech w populacji. Określają one zbiór wartości liczbowych właściwych jednostkom badanej populacji, które występują w rzeczywistości z określonym prawdopodobieństwem. W przeprowadzaniu wnioskowania statystycznego odnosimy się do kilku klasycznych typów rozkładów. Należą do nich rozkład dwumianowy, rozkład Poissona, rozkład normalny, rozkład t-Studenta oraz rozkład chi-kwadrat. Zostały one omówione w poniższej części rozdziału. Uprzednio chcemy przypomnieć kilka kluczowych pojęć, na których opiera się konstruowanie oraz wykorzystywanie owych rozkładów. Włączamy do nich takie pojęcia jak zmienna losowa oraz jej dwa rodzaje - zmienna ciągła i zmienna skokowa (dyskretna).

We wnioskowaniu statystycznym przyjmujemy, że badana cecha dla pojedynczej jednostki analizy może przyjmować szereg wartości. Ich przypisanie zależy od wielu czynników. Związane jest to z nadaniem im szczególnej we wnioskowaniu statystycznym właściwości, jaką jest losowość zdarzeń. Wartości cech jesteśmy jednak w stanie przewidzieć na podstawie reguły prawdopodobieństwa.

W statystyce każdą badaną cechę nazywamy zmienną. We wnioskowaniu statystycznym szczególnie ważne jest pojęcie **zmiennej losowej**. Jest to zmienna obejmująca swym zasięgiem wiele jednostek analizy, którym przyporządkowane zostały konkretne wartości liczbowe, zaś dobór ten miał charakter losowy. Każda jednostka z grupy mogła zostać poddana badaniu. **Zmienną skokową** bądź **dyskretną** nazywamy taką zmienną, której wartości liczbowe są przeliczalne, czyli możemy określić pewną wartość liczbową, o jaką będą zmieniać się poszczególne warianty zmiennej. Przykładem zmiennej skokowej jest zatem liczba dzieci w gospodarstwie domowym, liczba studentów w konkretnej grupie ćwiczeniowej czy też liczba zaliczonych egzaminów. Natomiast **zmienna ciągła** przyjmuje wartości ze zbioru liczb rzeczywistych, które są nieskończone i nieprzeliczalne. Intuicyjnie wartości takie mogą przyjmować bardzo dokładne wartości z określoną liczbą miejsc po przecinku. Przykładem zmiennej ciągłej są wzrost, waga, wydatki ponoszone na daną usługę lub też uzyskiwane dochody netto. Istotne jest zapamiętanie, iż we wnioskowaniu statystycznym analizujemy wyłącznie dane, które składają się ze zmiennych losowych. Natomiast w zależności od typu przyjmowanych wartości liczbowych - ciągłych bądź skokowych - dobieramy odpowiednie statystyki dedykowane wnioskowaniu statystycznemu. Wartości zmiennej skokowej tworzą rozkład, który nazywamy **rozkładem skokowym** bądź **dyskretnym**. Rozkłady teoretyczne dla tego typu zmiennej określane są przez rozkład dwumianowy oraz rozkład Poissona. Z kolei dla zmiennych typu ciągłego przewidziane są **rozkłady ciągłe**, czego przykładem jest rozkład normalny, rozkład t-Studenta oraz rozkład chi-kwadrat. Zostaną one omówione w dalszej części niniejszego rozdziału.

15.1.3.1. Rozkład normalny

Gdy obserwujemy rzeczywistość przyrodniczą i społeczną, oszałamia nas mnogość i różnorodność zachodzących w niej zjawisk. Wydaje się, iż wszystko permanentnie zmienia się, pierwotne przejawy zjawisk zastępowane są przez ich nowe wersje, zaś tempo tych przeobrażeń jest nieuchwytne dla naszego oka.

W tym konglomeracie niezrozumiałych, wielobarwnych oraz niezliczonych zdarzeń możemy jednak odnaleźć porządek. Jako badacze społeczni możemy zawładnąć owym chaosem. Zachodzące w przyrodzie i przestrzeni społecznej zjawiska wykazują bowiem zadziwiającą prawidłowość. Pewne cechy osobników ludzkich, zwierzęcych oraz fenomenów przyrodniczych powtarzają się, przejawiają tożsame tendencje oraz jednolite schematy występowania. Prawidłowość tą nazywamy rozkładem normalnym. Dzięki niemu dowiadujemy się, jaka cecha charakteryzuje największą grupę jednostek w rzeczywistości przyrodniczej i społecznej. Rozkład normalny wskazuje bowiem na częstotliwość występowania danego zjawiska lub cech. Zaobserwujemy go w przypadku charakterystyk demograficznych, czego przykładem jest wiek, współczynnik umieralności lub narodzin, charakterystyk biologicznych - wagi i wzrostu osobników jednorodnych populacji ludzkich i zwierzęcych, a także w zjawiskach ekonomicznych i społecznych. Istotą takiego „ukrytego" porządku jest fakt, iż da się go bardzo ściśle opisać za pomocą liczb i wzorów matematycznych. Cechą charakterystyczną tego ujęcia jest wyznaczanie pewnej ogólnej, przeciętnej wartości cechy, którą przejawia większość członków danej grupy. Rozkład normalny pozwala również na zidentyfikowanie przypadków rzadkich, skrajnych, występujących sporadycznie w przyrodzie.

Powszechnie uznaje się, iż tego fenomenalnego odkrycia dokonał Carl F. Gauss, niemiecki matematyk, fizyk oraz astronom żyjący na przełomie XVIII i XIX wieku. Niekiedy za jego współautora uznaje się Pierre'a S. de Laplace'a, matematyka francuskiego pochodzenia, uznawanego za jednego z twórców teorii prawdopodobieństwa. O ile przyczynili się oni do rozwoju rozkładu normalnego oraz jego popularyzacji

na gruncie statystyki, to jednak faktycznym jego odkrywcą jest francuski XVIII-wieczny matematyk - Abraham de Moivre. Jako nauczyciel matematyki oraz doradca uczestników gier losowych w karczmach odkrył, iż pewne zjawiska podczas kolejnych powtórzeń przyjmują zbliżone wartości, natomiast zdecydowanie rzadziej posiadają one wartości skrajne, odbiegające od głównej prawidłowości. Jego rozważania opublikowane zostały w 1773 roku w niewielkiej broszurce, która zapomniana na niemal 150 lat, przypadkowo trafiła w 1924 roku do rąk Karla Pearsona. Założyciel czasopisma „Biometrika" w nocie historycznej poświęconej krzywej normalnej podał do publicznej wiadomości, iż jej rzeczywistym odkrywcą nie był C.F. Gauss oraz P.S. de Laplace, ale właśnie A. de Moivre[1]. Prawdą jest jednak, że P.S. de Laplace zastosował rozkład normalny do analizy błędów pojawiających się w eksperymentach. Natomiast C.F. Gauss przyczynił się do opracowania uzasadnień oraz założeń o normalności rozkładu błędów, które wykorzystał przy analizie danych astrologicznych w 1809 roku. Wraz ze swoimi współpracownikami zauważył, iż powtarzając pomiary dla tej samej badanej cechy uzyskuje się zbliżone wyniki. Algorytm opisujący ich występowania sprowadzał się do pewnego jednolitego wzoru:

aktualny pomiar = oczekiwana wielkość cechy + błąd pomiaru

W kolejnym etapie badacze nanosili wyniki uzyskane w poszczególnych pomiarach na układ współrzędnych. Zauważyli, że rozkład częstości dla odnotowywanych pomiarów tworzy symetryczną linię o kształcie dzwonu. Wyróżniała się na niej jedna wartość szczytowa, nazywana wartością oczekiwaną. Z kolei wyniki pojedynczych pomiarów nie były identyczne z tą wartością, ale bardzo do niej zbliżone. Wartości cechy, które były od niej bardzo oddalone, pojawiały się zdecydowanie rzadziej[2].

Odkrycie A. de Moivre'a łatwo zrozumieć, posiłkując się następującym przykładem. Zmierzywszy wzrost dużej grupy losowo wybranych ludzi okazałoby się, że po naniesieniu zaobserwowanych wartości na wykres, układ linii łączącej poszczególne przypadki miałby kształt dzwonu. Utworzona krzywa nazywana jest krzywą normalną, niekiedy również terminem krzywej dzwonowatej[3]. Z kolei nazwa „normalny" dla tego rozkładu została zaproponowana przez Charlesa S. Peirce'a, Francisa Galtona oraz Wilhelma Lexisa i w ostateczności przyjęta za nazwę powszechnie obowiązującą[4]. Związane jest to również z faktem, iż wiele zjawisk badanych w politologii, socjologii, ale również w ekonomii, fizyce, astronomii czy też medycynie przyjmuje rozkład częstości o takiej postaci. Należy do nich większość charakterystyk populacji ludzkich takich, jak wiek, waga, iloraz inteligencji lub dochody. Postać tego rozkładu prezentuje wykres 9.

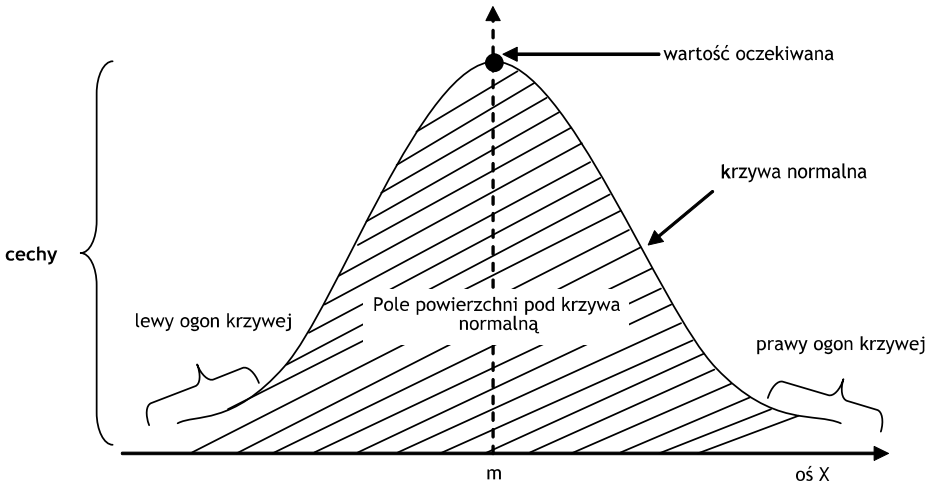
[1] K. Pearson, *Historical Note on the Origin of the Normal Curve of Errors,* „Biometrika", 1924, 16 (3-4), s. 402-404.

[2] A. Malarska, *Statystyczna analiza danych wspomagana programem SPSS*, SPSS Polska, Kraków 2005, s. 104.

[3] Nazwa „krzywa dzwonowata" wywodzi się od francuskiego matematyka Esprita Joufreta, który w 1872 roku użył terminu „powierzchnia dzwonowata" dla określenia kształtu, jaki przyjmuje dwuwymiarowy rozkład normalny z niezależnymi komponentami.

[4] Niektórzy badacze zwracają jednak uwagę na nietrafność używania tego określenia. W badaniach społecznych wręcz niespotykane są przypadki obserwowanych empirycznie zmiennych, których w idealnej postaci odzwierciedlałyby kształtem zamieszczoną powyżej prezentację rozkładu normalnego. Dla większości jednak cech o postaci zmiennej ciągłej ich rozkład zbliża się czy też upodabnia do modelowej postaci krzywej normalnej.

Wykres 9. Wykres rozkładu normalnego



Rozkładowi normalnemu przypisać możemy kilka własności. Po pierwsze, krzywa normalna jest krzywą **jednomodalną**. Oznacza to, że wartości uzyskiwane w kolejnych pomiarach są porównywane do pewnej najbardziej prawdopodobniej wartości cechy w populacji. Wartość taka nazywana jest wartością modalną, modą albo wartością oczekiwaną (jak na powyższym wykresie). W przypadku rozkładu normalnego miarą identyfikującą tą wartość jest średnia arytmetyczna (m) i jest jedną z dwóch parametrów odpowiadającą za kształt krzywej normalnej. Drugim parametrem jest natomiast odchylenie standardowe od średniej oznaczane symbolem σ. Kolejną własnością rozkładu normalnego jest **symetryczność** krzywej normalnej. Oznacza to, że obszar wykresu znajdujący się po lewej stronie od wartości oczekiwanej jest identyczny z kształtem znajdującym się z prawej strony - stanowi jego lustrzane odbicie. Trzecią cechą krzywej normalnej jest jej **asymptotyczność**. Istnieje nieskończona liczba wartości ekstremalnych, które mogą pojawić się na końcu prawej bądź lewej strony krzywej (czyli na tzw. ogonach krzywej). Istnieje bowiem niezerowe prawdopodobieństwo wystąpienia nowych wartości krańcowych. Dlatego znajdująca się na powyższym wykresie krzywa nigdy nie dotyka osi X. Przyjmuje się jednak, że **pole powierzchni** pod krzywą normalną jest skończone i równe **1**. Jest to kolejna, czwarta własność rozkładu normalnego. Piątą przypisywaną mu cechą jest uznanie, iż krzywa normalna prezentuje przybliżony rozkład częstości dla **zmiennych ciągłych i nieograniczonych**.

Rozkład normalny jest pewnym typem idealnym. W rzeczywistości jest on wręcz niespotykany. Występowanie zjawisk przyrodniczych oraz społecznych jednak w przybliżeniu odzwierciedla postać tego rozkładu. Dlatego jest on uznawany za użyteczne narzędzie statystyczne do porównywania i szacowania różnych wartości cech w populacji. W praktyce badawczej stosujemy ogólną, uproszczoną postać rozkładu normalnego z tzw. zestandaryzowną krzywa normalną. Stanowi ona płaszczyznę porównawczą i stały punkt odniesienia dla wszystkich obliczeń z próby. Dzięki temu możemy korzystać z jednolitych wzorów i jasnych procedur obliczeniowych. Największą zaletą standaryzowanego rozkładu normalnego jest możliwość przenoszenia wyników z próby na populację.

Istnieje wiele zastosowań dla rozkładu normalnego. Jednym z nich jest wnioskowanie statystyczne. Na podstawie statystyk rozkładu normalnego weryfikujemy hipotezy statystyczne oraz dokonujemy przenoszenia uzyskiwanych wyników z próby na populację. Metoda ta pozwala również na określenie tendencji w występowaniu pewnych cech w zbiorowościach generalnych. Tego rodzaju obliczenia opierają się na analizie wielkości pola powierzchni pod krzywą rozkładu normalnego i na porównywaniu średnich

ze sobą. W ten sposób możemy określić chociażby natężenie występowania danej cechy lub proporcję osób o określonych charakterystykach w populacji. Przyjmujemy bowiem, że to pole obejmuje wszystkie jednostki zbiorowości generalnej. Wielkość tego pola wynosi zatem 100 proc., co w przełożeniu na wartości bezwzględne równe jest 1. W określaniu tychże tendencji posługujemy się założeniem o symetryczności rozkładu. Wartość oczekiwana rozkładu jest środkiem rozkładu zmiennych, który dzieli całą badaną zbiorowość na dwie równe grupy - pierwszą grupę jednostek o wartościach cechy poniżej średniej (lewa strona rozkładu) oraz drugą grupę z wartościami powyżej średniej (prawa strona rozkładu). W zależności od tego, jaką wartość przyjmie poszczególna obserwacja, możemy określić prawdopodobieństwo procentowego udziału jednostek, które będą osiągały wyniki od niej wyższe bądź niższe. Wiąże się to z obliczeniem pola powierzchni pod krzywą normalną. Nie musimy jednak tego obliczać „ręcznie". Dokonują tego za nas odpowiednie programy statystyczne. Dla naszych potrzeb ważne jest nabycie umiejętności właściwego interpretowania statystyk rozkładu normalnego.

Wykorzystując rozkład normalny porównujemy wartość średniej w próbie badawczej z przewidywaną bądź zakładaną wartością średniej w populacji. Mówimy wówczas o szczególnym rodzaju rozkładu normalnego nazywanym **rozkładem średniej z próby**. Umożliwia on odnalezienie wartości średniej dla badanej zmiennej. Dokonujemy tego za pomocą statystyki Z, którą obliczamy ze wzoru:

$$Z = (\bar{x} - m) * \frac{\sqrt{N}}{\sigma}$$

W tym wzorze $\bar{x}$ to średnia z próby, m to średnia w populacji, natomiast $\frac{\sqrt{N}}{\sigma}$ to błąd standardowy średniej będący stosunkiem do odchylenia standardowego w populacji do pierwiastka kwadratowego liczby jednostek w próbie. Po obliczeniu tej statystyki określamy jej miejsce na osi X rozkładu normalnego. W tym celu wyznaczamy dwa obszary. Pierwszy z nich nazywany jest przedziałem ufności zaś drugi - obszarem krytycznym. Są one wydzielane na podstawie tzw. wartości krytycznej (Z_α), którą ustalamy arbitralnie na podstawie wybranego poziomu ufności (α). Służy ona do wyznaczenia przedziału ufności (1-α), czyli obszaru prawdopodobnych wartości cechy w populacji. Gdy obliczamy statystykę Z z próby, sprawdzamy czy jej wartość mieści się w tym obszarze. Powinna ona spełniać zatem następujący warunek:

$$-Z_\alpha \leq Z \leq Z_\alpha$$

Oznacza to, że średnia z próby nie różni się od założonej średniej w populacji. Sytuacja przeciwna ma miejsce w momencie, gdy zajdzie następująca relacja:

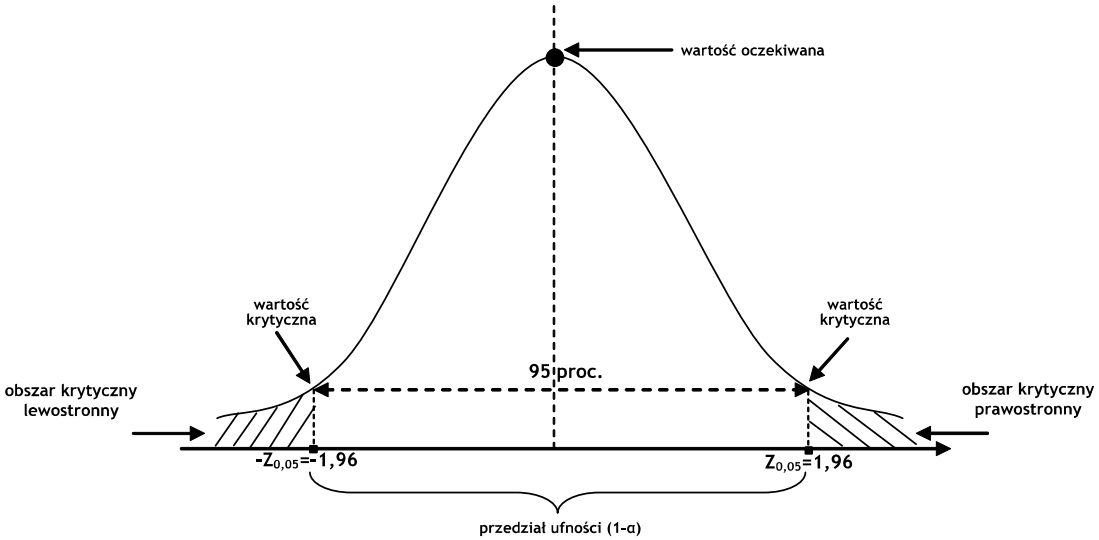
$$-Z_\alpha > Z \text{ lub } Z_\alpha < Z$$

Oznacza to, iż wartość obliczonej statystyki Z wykracza poza przedział ufności i mieści się w obszarach krytycznych rozkładu normalnego, czyli w tzw. „odciętych marginesach, ogonach" - po lewej lub prawej stronie rozkładu. Wnioskujemy, iż wartość statystyki z próby różni się istotnie od wartości oczekiwanej w zbiorowości generalnej.

Wracając jeszcze do wartości Z_α, powinniśmy zapamiętać, iż tradycyjnie jest ona wyznaczana dla trzech wartości przedziału ufności (1-α). W naukach społecznych powszechnie stosuje się wartość - 90 proc. (α=0,1) oraz 95 proc. (α=0,05), z kolei w naukach przyrodniczych najczęściej spotyka się 99 proc.

(α=0,01)[5]. Dla tych przedziałów wartość statystyki Z_α (dla przedziału obustronnego) będą wynosić odpowiednio: $|Z_{0,1}|$=|1,64|, $|Z_{0,05}|$=|1,96| oraz $|Z_{0,01}|$=|2,58|. Powyższe rozważania ilustruje wykres 10.

Wykres 10. Charakterystyka wykresu rozkładu normalnego



Przed przystąpieniem do omawiania kolejnych rozkładów teoretycznych chcielibyśmy przekazać wiedzę na temat minimalnej liczby jednostek analizy, od której zależy występowanie rozkładu normalnego. Wiąże się to z dwoma podstawowymi twierdzeniami, jakimi są założenie o normalności rozkładu zmiennej oraz centralne twierdzenie graniczne. Są to warunki podstawowe wymagane do spełnienia, w celu zastosowania wybranych testów oraz metod analizy danych omawianych w kolejnych rozdziałach niniejszego podręcznika.

Jak wiemy, empiryczne rozkłady częstości zmiennych dążą swym kształtem do modelu pewnego oczekiwanego rozkładu teoretycznego. Większość testów istotności wymaga jednak spełnienia **założenia o normalności rozkładu zmiennej.** Związane jest to z koniecznością sprawdzenia, czy rozkład badanej zmiennej przyjmuje w populacji postać rozkładu normalnego. Założenie o normalności rozkładu jest wymagane w testach parametrycznych. Bez spełnienia tego kryterium przeprowadzenie wnioskowania statystycznego jest nieuzasadnione, natomiast uzyskane wyniki mogą prowadzić do błędnych interpretacji.

Założenie o normalności rozkładu zmiennej zależne jest w dużej mierze od liczebności próby badawczej. Wraz ze wzrostem wielkości próby rozkład zmiennej zbliża się kształtem do postaci krzywej normalnej. Tendencja ta ma swoje odzwierciedlenie w twierdzeniu statystycznym nazywanym **centralnym**

[5] We wnioskowaniu statystycznym spotykamy się z dwoma parametrami – poziomem ufności (α) oraz poziomem istotności (p). Ważne jest umiejętne ich rozróżnienie, gdyż wielokrotnie w analizach ilościowych możemy spotkać się z ich utożsamieniem. Poziom ufności (α) jest miarą właściwą rozkładom teoretycznym i szacunkom dokonywam przy ich pomocy. Określa on maksymalne ryzyko błędu, jakie badacz jest skłonny zaakceptować. Jego wartość stanowi swoistą granicę wyznaczoną z góry, z którą porównywane są wartości właściwej mu statystki z próby. Tą wartością obliczaną dla poszczególnych wyników badania jest poziom istotności (p), który określa istotność wyników i poziom prawdopodobieństwa ich występowania w rzeczywistości, nie będących przy tym rezultatem przypadku. Poziom istotności (p) jest zatem miarą obliczaną indywidualnie dla wyników pojedynczego badania, z kolei poziom ufności jest wielkością ustalaną arbitralnie przez badania i właściwą dla danego rozkładu teoretycznego. W programie PSPP poziom istotności (p) jest obliczany dla poszczególnych typów testów statystycznych. Jeżeli jego wartość wynosić będzie p=0,00124, natomiast przyjęty poziom ufności (α) ustalimy na poziomie 0,05, to p jest mniejsze od α, zatem możemy mówić o istotności wyników z próby. Z kolei jeśli p byłoby większe od α, to nie możemy uznać wyników za istotne statystycznie i wnioskować na ich podstawie o cechach populacji generalnej.

twierdzeniem granicznym. Mówi ono, iż w sytuacji dobierania próby losowej o liczebności N z populacji o dowolnym rozkładzie wyrażonym za pomocą parametru m - średniej arytmetycznej i σ - odchyleniu standardowym, wraz ze wzrostem liczebności próby, rozkład średniej coraz bardziej przypomina rozkład normalny lub - jak wyrażają się statystycy - dąży do rozkładu normalnego.

Biorąc pod uwagę założenia centralnego twierdzenia granicznego uznaje się, iż rozkład zmiennych przyjmuje postać rozkładu normalnego, gdy liczebność próby jest większa niż 30 jednostek analizy. Najbezpieczniejszą granicą są jednak próby o minimalnej liczebności mieszczącej się w granicach od 100 do 120 jednostek. Dla prób mniejszych praktyka badawcza radzi skorzystać z rozkładu t-Studenta. Z kolei dla zmiennych mierzonych na poziomie nominalnym, statystyki powinny być obliczane na podstawie rozkładu chi-kwadrat. Rozkłady te omówiono poniżej.

15.1.3.2. Rozkład t-Studenta

Rozkład t-Studenta stanowi alternatywę względem wcześniej omawianego rozkładu normalnego. Jak pamiętamy, wiele zjawisk i cech populacji ludzkich oraz zwierzęcych przyjmuje postać rozkładu normalnego. Dzięki temu możemy określić średnią wartość badanej cechy lub odchylenie standardowe od tej miary. Wiemy bowiem, iż większość jednostek jest podobnych pod względem pewnych charakterystyk - wzrostu, wieku lub wagi. Problem pojawia się w sytuacji, gdy chcemy określić wartości tych cech na podstawie danych z próby badawczej. Bywa bowiem tak, iż nie dysponujemy wystarczającą liczbą jednostek analizy bądź nie znamy wartości określonych parametrów - średniej wartości cechy w populacji, a najczęściej - odchylenia standardowego. Niezmiernie pomocny w takiej sytuacji okazuje się rozkład t-Studenta. Dlatego też obliczenia dokonywane na jego podstawie są najczęściej spotykane w analizie danych ilościowych. Ponadto porównując go do rozkładu normalnego, jest on mniej wymagający w zakresie jakości danych potrzebnych do oszacowań. Jego dużą zaletą jest także możliwość zastosowania go dla obliczeń na małych próbach, gdzie $n \leq 30$. Poniżej przedstawiamy podstawowe założenia teoretyczne tego rozkładu. Poznanie ich jest o tyle istotne, iż w dalszej części podręcznika omawiamy testy statystyczne wykorzystujące statystyki rozkładu t-Studenta.

Twórcą rozkładu t-Studenta jest William S. Gosset, badacz jakości piwa w browarach Arthura Guinnessa. Odkrył on, że przez zastąpienie nieznanego odchylenia standardowego w populacji odchyleniem standardowym z próby, analizy prowadzone na małych liczebnościach prezentują nieprecyzyjne, „skrzywione" wyniki badania. Natomiast lepsze oszacowania nieznanych wartości otrzymuje się, gdy w obliczeniach zostanie wzięta pod uwagę wielkość próby badawczej. To spostrzeżenie posłużyło W.S. Gossetowi do opracowania rozkładu prawdopodobieństwa nazywanego rozkładem t-Studenta. Wyniki swojego odkrycia opublikował w 1908 roku w czasopiśmie „Biometrika" założonym i prowadzonym przez K. Pearsona. Nie podpisał się on jednak swoim imieniem i nazwiskiem, ale pseudonimem „Student". Anegdota głosi, iż uczynił to ze względu na odgórny zakaz korporacyjny, uniemożliwiający pracownikom publikowania wyników indywidualnych osiągnięć. Rzeczywistą intencją było jednak ukrycie przez browar Guinnessa faktu wykorzystywania tej statystycznej innowacji. Z tego względu omawiany poniżej rozkład teoretyczny nie jest nazywany rozkładem t-Gosseta, ale rozkładem t-Studenta.

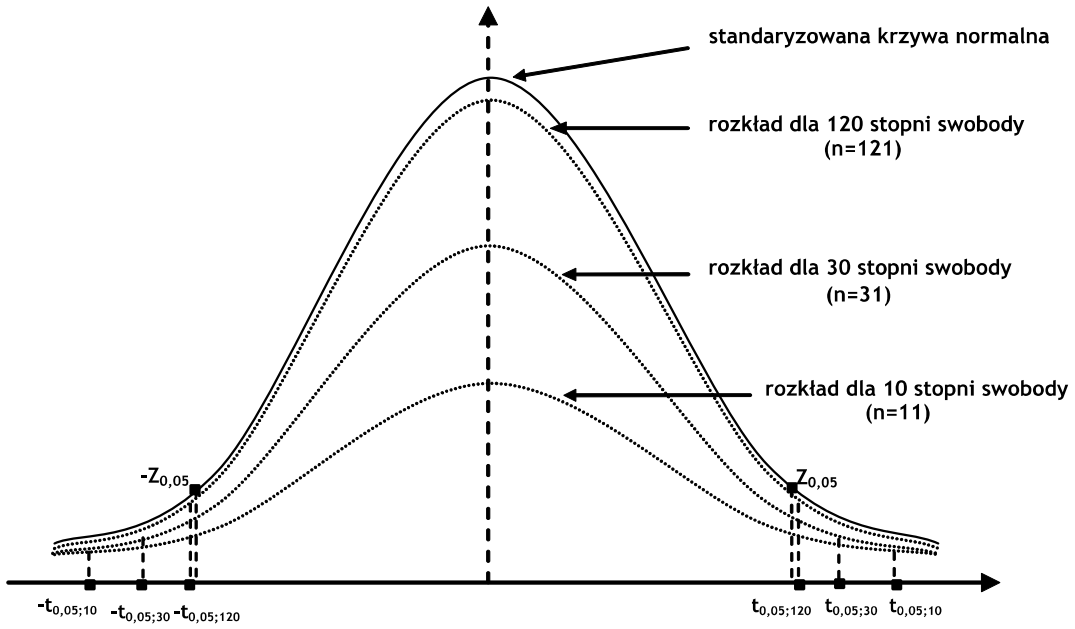
Rozkład t-Studenta stosujemy, gdy nie jesteśmy w stanie użyć rozkładu normalnego. Na kształt tego rozkładu wpływa jeden parametr - liczebność próby badawczej, a precyzując tzw. liczba stopni swobody definiowana równaniem $df=n-1$, gdzie n to liczba jednostek w próbie. Liczba stopni swobody określa maksymalną liczbę wartości, jakie mogą przyjąć jednostki niezależnie dobrane do próby badawczej.

Wyznaczenie liczby stopni swobody jest niezbędne do prowadzenia obliczeń na podstawie rozkładu t-Studenta.

Pod względem wielu charakterystyk rozkład t-Studenta jest podobny do rozkładu normalnego. Do jego opracowania W.S. Gosset wykorzystał bowiem główne założenia standaryzowanej krzywej normalnej. Z tego też względu do prezentacji rozkładu t-Studenta wykorzystuje się krzywą o postaci dzwonowatej. Jej właściwością jest symetryczny kształt. Oznacza to, iż lewa połowa rozkładu jest lustrzanym odbiciem jego prawej strony. Celem wyznaczenia takiego kształtu rozkładu jest poszukiwanie zakresu liczbowego dla szacowanego parametru populacji. Pole pod krzywą normalną obejmuje wszystkie wartości, jakie może przyjąć obliczana wielkość statystyki w zbiorowości generalnej. Mamy więc tożsamą sytuację z obserwowaną w przypadku rozkładu normalnego. Różnica między tymi rozkładami jest jednak zasadnicza. W rozkładzie normalnym korzystamy z jednej postaci krzywej nazywanej standaryzowaną krzywą normalną. Do niej przyrównujemy wszelkie obliczenia z próby oraz korzystamy ze stałych wartości Z_α. Z kolei w rozkładzie t-Studenta, za każdym razem mamy inny kształt krzywej, co jest zależne od liczebności próby badawczej. Im bowiem mniej jednostek zbadaliśmy, tym krzywa rozkładu jest bardziej spłaszczona. Z kolei im więcej jednostek zdołamy objąć badaniem, tym kształt krzywej jest bardziej wysmukły. Powoduje to, iż statystyka $t_{\alpha;n-1}$ właściwa temu rozkładowi, jest zawsze różna i zależa od liczebności próby badawczej, a dokładniej - od liczby stopni swobody. Zagadnienie to zostało omówione poniżej.

Podobnie, jak w przypadku rozkładu normalnego, właściwościami rozkładu t-Studenta są obszar krytyczny, wartości krytyczne ($t_{\alpha,n-1}$), poziom ufności (α) oraz przedział ufności ($1-\alpha$). W przypadku rozkładu normalnego wartości krytyczne Z_α są stałe i ustalane na podstawie przyjętego poziomu ufności (α). W rozkładzie t-Studenta wartości wyznaczające obszary krytyczne ($t_{\alpha,n-1}$) są zależne dodatkowo od liczby stopni swobody (df). Jeżeli porównamy wartości krytyczne dla tego samego poziomu ufności $\alpha=0,05$, ale dla prób o odmiennych liczebnościach, zauważymy, iż wartość $t_{\alpha,n-1}$ różni się względem wartości Z_α. Co więcej, wraz ze zmniejszeniem się liczebności próby, wartość $t_{\alpha,n-1}$ zwiększa się, zaś Z_α pozostaje na tym samym poziomie. Porównując wartości tych statystyk dla trzech wielkości prób $N_1=11$, $N_2=31$ oraz $N_3=121$ oraz dla poziomu ufności $\alpha=0,05$, uzyskujemy jedną wartość dla rozkładu normalnego dwustronnego, która wynosi $Z_\alpha=1,96$, zaś dla odpowiednich liczebności próby wartość t wyniesie - $t_{0,05;10}=2,228$, $t_{0,05;30}=2,042$, $t_{0,05;120}=1,965$. Jak możemy zaobserwować w próbach liczących więcej niż $N=120$ jednostek, wartość $t_{\alpha,n-1}$ zbliża się do wartości Z_α. Przy zachowaniu odpowiednich liczebności krzywa rozkładu t-Studenta zaczyna bowiem przypominać kształtem krzywą rozkładu normalnego. Wszelkich oszacowań możemy wówczas dokonywać w oparciu o statystykę Z. Zmniejszenie liczebności próby powoduje natomiast zwiększenie się wartości $t_{\alpha,n-1}$ względem Z_α, a co za tym idzie, obszar krytyczny odsuwa się od wartości średniej w populacji. W konsekwencji dochodzi do spłaszczenia pola powierzchni pod krzywą rozkładu, co powoduje zwiększenie się zakresu wartości, jakie może przyjąć średnia w populacji. Kształt rozkładu t-Studenta przyjmowany w zależności od wielkości próby, ilustruje wykres 11.

Wykres 11. Kształt rozkładu t-Studenta w zależności od wielkości próby



Dla usystematyzowania wiedzy powtórzmy, iż rozkład t-Studenta jest rozkładem zależnym od liczby stopni swobody (df=n-1), a tym samym od wielkości próby. Zastosujemy go zawsze, gdy analizujemy próby o liczebnościach mniejszych niż N=30. Należy jednak pamiętać, iż rozkład t-Studenta dąży do postaci rozkładu normalnego w dużych próbach, liczących więcej niż N=120 jednostek. Dlatego też większość testów parametrycznych dostępnych w programach statystycznych korzysta z tego rozkładu. Wynika to z jego elastyczności i możliwości zastosowania do prób o różnych liczebnościach.

15.1.3.3. Rozkład chi-kwadrat

Rozkład chi-kwadrat jest kolejnym rozkładem teoretycznym zmiennej losowej, z którym spotykamy się we wnioskowaniu statystycznym. Gdy chcemy porównać częstość pojawiania się danych cech empirycznie obserwowalnych z pewnym rozkładem ich występowania w całkowitych populacjach, zastosujemy statystykę opartą właśnie na rozkładzie chi-kwadrat. Rozkład ten jest swoistym modelem teoretycznym, według którego możemy prezentować, oceniać i analizować wszelkie częstości występowania jednostek o określonych charakterystykach. Stosujemy go w przypadku badania cech jakościowych, które najczęściej spotykamy w badaniach społecznych. Zagadnieniami badanymi są wówczas poglądy, preferencje, oczekiwania, sądy, a także charakterystyki socjodemograficzne - płeć, miejsce zamieszkania, rodzaj ukończonej szkoły średniej lub miejsce pracy. Formułujemy zawsze pewne przypuszczenie, a następnie oceniamy jego trafność i dopasowanie do oczekiwanego stanu w rzeczywistości. Możemy zakładać chociażby, iż wśród 20 noworodków, 10 z nich będzie płci męskiej. Okazuje się jednak, iż w rzeczywistości urodziło się wyłącznie 8 chłopców. Wykorzystując rozkład chi-kwadrat będziemy mogli sprawdzić, czy różnica między tymi częstościami jest efektem przypadku, czy też może jest wynikiem oddziaływania innych, znaczących czynników.

Twórcą rozkładu chi-kwadrat oraz testów opracowanych na jego podstawie, jest matematyk angielskiego pochodzenia i prekursor statystyki matematycznej - Karl Pearson[6]. Ian Hacking, kanadyjski matematyk zajmujący się filozofią nauki, uznał to osiągnięcie za jedno z dwudziestu największych odkryć dokonanych od 1900 roku w zakresie nauki i techniki[7]. Rozkład chi-kwadrat jest bowiem najczęściej wykorzystywanym rozkładem we współczesnych analizach ilościowych. Podstawowym zastosowaniem tego rozkładu jest wnioskowanie o wariancji w populacji. Z rozkładem chi-kwadrat mamy do czynienia wszędzie tam, gdzie ze względu na małą liczebność próby nie jesteśmy w stanie zaobserwować rozkładu normalnego. Poniżej w skróconej postaci przybliżamy istotę tego rozkładu.

Rozkład chi-kwadrat jest rozkładem asymetrycznym, który zbliża się swoim kształtem do postaci krzywej normalnej wraz ze wzrostem liczebności próby. Obliczenie statystyki chi-kwadrat zależne jest od parametru nazywanego liczbą stopni swobody, oznaczanego w przypadku tej miary symbolem *v*. Liczba stopni swobody mówi nam o wszystkich wartościach, jakie może przyjąć badana zmienna losowa. Zatem dla zmiennej ciągłej, która może przyjąć nieskończoną liczbę wartości, liczba stopni swobody wynosić będzie *v=n,* gdzie n oznacza liczebność próby badawczej.

Wartość statystyki chi-kwadrat związana jest z rozkładem prawdopodobieństwa sumy kwadratów niezależnych zmiennych losowych o standaryzowanym rozkładzie normalnym. Zależna jest ona od takich parametrów, jak wartość oczekiwana wyrażana za pomocą średniej arytmetycznej w populacji (m) oraz poprzez właściwe jej odchylenie standardowe (σ), a także wartość statystyk z n prób losowych - xn, gdzie n to liczba jednostek analizy w próbie. Wartość statystyki chi-kwadrat oznaczana jest symbolem χ^2 i wyrażana wzorem:

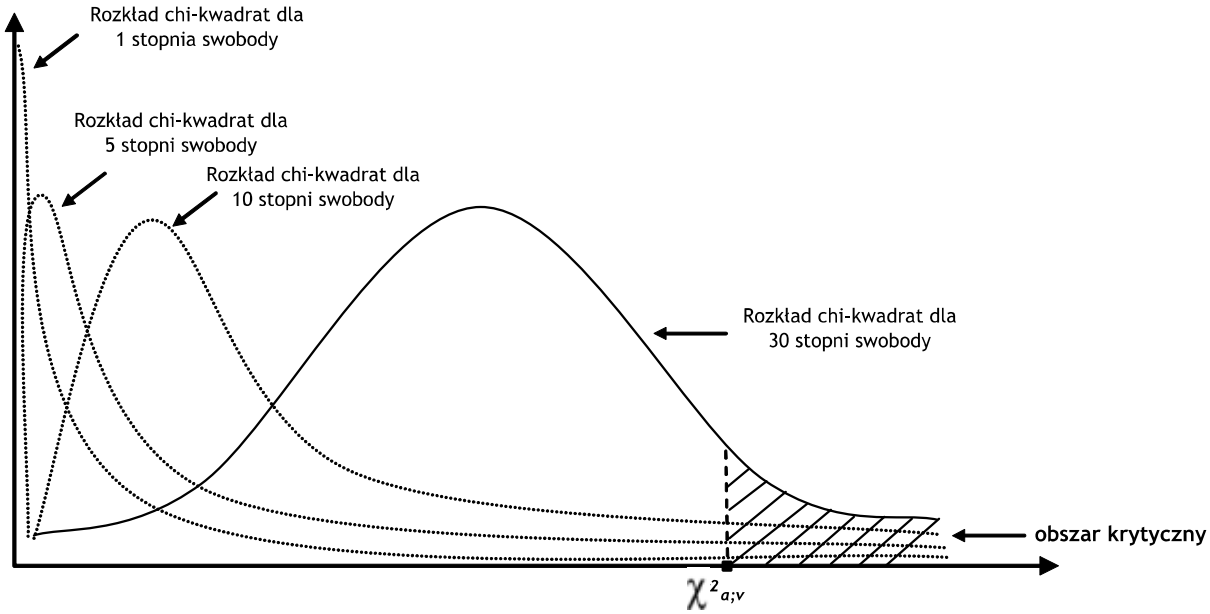
$$\chi^2 = \left(\frac{x_1 - m}{\sigma}\right)^2 + \left(\frac{x_2 - m}{\sigma}\right)^2 + \cdots + \left(\frac{x_n - m}{\sigma}\right)^2$$

Podobnie jak w przypadku rozkładu normalnego i rozkładu t-Studenta, do opisu rozkładu chi-kwadrat wykorzystujemy takiej pojęcia, jak wartość krytyczna ($\chi^2_{\alpha;v}$), obszar krytyczny, przedział ufności (1-α) i poziom ufności (α). Wartości krytyczne $\chi^2_{\alpha;v}$ dla tego rozkładu są stablicowane i zależne od liczby stopni swobody *v* oraz przyjętego poziomu ufności (α). Na tej podstawie wyznaczamy obszar krytyczny i określamy, czy analizowany rozkład jest tożsamy z przewidywanym rozkładem w populacji. Należy jednak zwrócić uwagę, iż rozkład chi-kwadrat jest rozkładem asymetrycznym prawostronnym, przyjmującym wyłącznie wartości dodatnie od 0 do +∞. W konsekwencji, jeżeli obliczona wartość χ^2 z próby spełnia relację $\chi^2 > \chi^2_{\alpha;v}$, gdzie α to przyjęty poziom ufności (z reguły przyjmujący wartość 0,1, 0,05 lub 0,01), zaś liczba stopni swobody wyznaczana jest przez liczebność próby, to stwierdzamy, iż dwa porównywane rozkłady różnią się między sobą. Natomiast jeżeli $\chi^2 \leq \chi^2_{\alpha;v}$ to przyjmujemy, że rozkład empiryczny jest tożsamy z zakładanym rozkładem w populacji. Postać rozkładu chi-kwadrat wraz ze wskazanymi powyżej parametrami prezentuje wykres 12.

[6] Ze szczegółowym opisem osiągnięcia K. Pearsona Czytelnik może zapoznać się w dziele: K. Pearson, *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, „Philosophical Magazine", 1900, 50 (302), s. 157-175.

[7] C.R. Roa, Karl Pearson *Chi-Square Test: The Dawn of Statistical Inference*, w: *Goodness-of-Fit Tests and Model Validity*, C. Huber-Carol (red.), Birkhauser, Boston 2002, s. 9.

Wykres 12. Wykres rozkładu chi-kwadrat w zależności od wielkości próby



5.1.3.4. Rozkład dwumianowy

Jednym z podstawowych rozkładów teoretycznych jest rozkład dwumianowy, nazywany również rozkładem binominalnym lub rozkładem Jakuba Bernoulliego. Pojawia się on zawsze tam, gdzie chcemy ustalić prawdopodobieństwo zajścia określonego zdarzenia. Spotkamy się z nim podczas analizy rzutu kostką bądź monetą, a także przy szacowaniu możliwości wygrania wyborów przez konkretną partię lub prawdopodobieństwo urodzenia dziecka płci męskiej bądź żeńskiej. Istnieje szereg zjawisk, których badanie wymaga porównywania do rozkładu dwumianowego. Podstawowe założenia tego rozkładu zostały przedstawione poniżej.

Wynalezienie tego rozkładu zawdzięczamy szwajcarskiemu matematykowi i fizykowi żyjącemu na przełomie XVII i XVIII wieku - Jakubowi Bernoulliemu. Jego osiągnięcia w dziedzinie badania rachunku prawdopodobieństwa do dzisiaj wykorzystywane są w statystyce matematycznej. Specjalizował się on w analizie twierdzeń o grach losowych z zastosowaniem permutacji i kombinatoryki w badaniu prawdopodobieństwa zajścia określonego zdarzenia. Rezultatem tego było opracowanie twierdzenia, które współcześnie znane jest pod nazwą schematu Bernoulliego bądź twierdzenia Bernoulliego. Jest ono podstawą dla konstruowania rozkładu dwumianowego[8]. Swoje osiągnięcie J. Bernoulli opisał w krótkim dziele

[8] W publikacjach polskojęzycznych najczęściej spotykaną nazwą na określenie tego rozkładu jest termin rozkład dwumianowy. Z kolei w publikacjach anglojęzycznych określa się go mianem rozkładu zero-jedynkowego lub rozkładu dwupunktowego. Zgodnie bowiem z założeniami jego autora, Jakuba Bernoulliego, służy on do opisywania rozkładu cech, które mogą przyjmować wyłącznie dwie wartości - 1 dla zajścia danego zdarzenia, oraz 0 dla zajścia zdarzenia alternatywnego bądź braku wystąpienia zdarzenia. Badane zmienne są zmiennymi dwuwartościowymi, dychotomicznymi i przyjmują wartości 0 bądź 1. Z tego też względu niektórzy autorzy uważają używanie terminu „rozkład dwumianowy" za błędne. Zob.: J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych,* Wydawnictwo Naukowo-Techniczne, Warszawa 2001, s. 105. W niniejszym podręczniku zachowujemy zgodność z tradycyjnie stosowaną terminologią właściwą dla publikacji polskojęzycznych i opisywany rozkład nazywamy rozkładem dwumianowym.

zatytułowanym *Ars Conjectandi*, które zostało opublikowane po śmierci uczonego w 1705 roku, natomiast faktycznie docenione 50 lat później[9].

Rozkład dwumianowy wykorzystywany jest do przewidywania tendencji, zachowań i zjawisk w badanych populacjach. W wymiarze teoretycznym opiera się on na założeniu o prawdopodobieństwie wystąpienia dwóch kategorii zdarzeń - sukcesu oznaczanego symbolem *p* oraz porażki *q=1-p* dla n niezależnych przypadków (warunkiem jest, aby *n* było większe niż 2). Za sukces uznajemy zajście danego zjawiska i przypisujemy mu wartość 1, zaś za porażkę traktujemy niewystąpienie żadnego zjawiska bądź zajście zjawiska alternatywnego, co oznaczamy jako 0. Po zliczeniu wszystkich przypadków, gdzie zaszło interesujące nas zdarzenie oraz gdzie się ono nie pojawiło, ustalamy wartość *p* oraz *q*. Przyjmują one wartości z zakresu od 0 do 1. Pojedynczą obserwację dla takiego zdarzenia, czyli określenie sukcesu oraz porażki dla zaistnienia danego zjawiska, nazywamy próbą Bernoulliego. Schemat Bernoulliego pozwala z kolei na wyznaczenie prawdopodobieństwa wystąpienia zjawiska, gdy analizujemy ciąg niezależnych sytuacji. Na tej podstawie określamy liczbę sukcesów w próbie, którą oznaczamy literą *k,* zaś literą *n* - liczebność próby. Posiadając takie dane, możemy określić prawdopodobieństwo zajścia zdarzenia w konkretnych warunkach. Wyrażane jest ono wzorem:

$$P_n(k) = \binom{n}{k} * p^k * q^{n-k}$$

Dla obliczenia $\binom{n}{k}$ stosujemy natomiast wzór:

$$\binom{n}{k} = \frac{n!}{k! * (n-k)!}$$

W celu zrozumienia zasady stosowania tego rozkładu wykorzystamy przykład. Przypuśćmy, iż interesuje nas szansa wygrania wyborów przez partię X wśród studentów I roku politologii. W tym celu przebadaliśmy grupę liczącą 20 osób. Z ogólnopolskich sondaży wiemy, iż poparcie dla tej partii jest duże i wynosi 60 proc. Prawdopodobieństwo wygrania przez tą partię wyborów wynosi zatem p=0,6, natomiast przegranej - q=0,4. Poproszono studentów o wskazanie na kartkach partii, na którą będą chcieli oddać głos w nadchodzących wyborach. Po zliczeniu uzyskanych wyników okazało się, że 10 studentów chce głosować na partię X, pozostali wybrali natomiast inne ugrupowania. Liczba sukcesów w tej grupie wynosi zatem k=10. Przystąpmy do obliczenia prawdopodobieństwa wygrania wyborów przez partię X wśród studentów I roku politologii:

$$P_{20}(10) = \binom{20}{10} * (0{,}6)^{10} * (0{,}4)^{10} = 0{,}117$$

Pomimo wysokich notowań w sondażach partii X, prawdopodobieństwo wygrania przez nią wyborów w grupie studentów I roku politologii nie jest zbyt wysokie. Wynosi ono zaledwie 11,7 proc. Należy jednak zwrócić uwagę, iż próba badawcza nie jest zbyt liczna, a przytoczony przykład dość abstrakcyjny. Rozkład dwumianowy ma jednak szerokie zastosowanie w statystyce. Spotkamy się z jego wykorzystaniem w teście dwumianowym, dostępnym również w programie PSPP pod nazwą *Test Binomial.* Należy on do rodziny testów nieparametrycznych i jest dedykowany zmiennym typu skokowego, przede wszystkim dychotomicznym (przyjmującym dwie wartości - 0 lub 1). Innym rozkładem teoretycznym

[9] G. Shafer, *The Significance of Jacob Bernoulli's Ars Conjectandi for the Philosophy of Probability Today*, „Journal of Econometrics", 1996, 75 (1), s. 15.

opracowanym dla zmiennych skokowych jest rozkład zaproponowany przez Simeona D. Poissona. Omawiamy go w kolejnej części podręcznika.

15.1.3.5. Rozkład Poissona

Rozkład S.D. Poissona jest wykorzystywany do określania właściwości zjawisk o charakterze dyskretnym (skokowym), które mogą nie wystąpić w rzeczywistości bądź zdarzyć się jeden, dwa, trzy aż do n-tego razu w danym przedziale czasowym lub obszarze występowania. Rozkład Poissona odnajduje zastosowanie, gdy prawdopodobieństwo zajścia danego zjawiska bądź zdarzenia jest bardzo małe, natomiast liczba jednostek, którym możemy przypisać ich wystąpienie, jest duża. Twórcą tego rozkładu, jak sama nazwa wskazuje, jest Simoen D. Poisson, francuski matematyk i fizyk żyjący na przełomie XVIII i XIX wieku. Prowadząc badania nad prawdopodobieństwem wydania danego orzeczenia sądowego w sprawach cywilnych i karnych, wyprowadził on teorię dla przewidywania określonego zdarzenia w wyznaczonym przedziale czasu[10]. Rozkład Poissona odnajduje zatem zastosowanie do opisu częstości występowania zjawisk rzadkich w określonym interwale czasowym, które są wręcz nieprawdopodobne, niespotykane i „niezwykłe". Ze względu na bardzo małą liczbę przypadków, które są przewidywane na jego podstawie, jest on określany również mianem prawa małych liczb Poissona[11]. Duży wkład do rozwoju tej teorii miał rosyjski matematyk polskiego pochodzenia - Władysław Bortkiewicz. Z tego też względu niekiedy w literaturze możemy spotkać się z nazwą prawa małych liczb Poissona-Bortkiewicza[12]. Badacz ten wykorzystał rozkład Poissona do określenia liczby zgonów wśród kawalerzystów pruskich korpusów, którzy zginęli na skutek śmiertelnego kopnięcia konia. Rozkład Poissona jest wykorzystywany do analizy wielu zdarzeń rzadkich występujących w przyrodzie oraz w populacjach ludzkich. Przykładem jest chociażby próba określenia liczby przypadków zachorowań na niespotykaną chorobę wśród ogółu dorosłych mężczyzn mieszkających w krajach skandynawskich lub prawdopodobieństwo wykolejenia się pociągu w okresie wakacyjnym. Przejdźmy do omówienia głównych założeń teoretycznych dla tego rozkładu.

Rozkład Poissona jest odpowiednikiem rozkładu dwumianowego. Jest on jednak przeznaczony do badania zjawisk o bardzo małym prawdopodobieństwie zajścia w rzeczywistości. Ich występowanie nazywamy sukcesem i oznaczamy symbolem p. Z reguły przyjmuje się, że jego wartość jest mniejsza bądź równa 0,02 ($p \leq 0{,}02$), a liczba przeprowadzonych obserwacji duża i wynosząca co najmniej $n=100$. Określenie prawdopodobieństwa zajścia zdarzenia przy ustaleniu takich warunków wyrażone jest wzorem:

$$P_n(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

[10] *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus system*, M. Denuit (red.), John Wiley and Sons, Chichester 2007, s. 3-4.

[11] J. Wawrzynek, *Metody opisu i wnioskowania statystycznego*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, Wrocław 2007, s. 56-57.

[12] Klasycznym przykładem wykorzystania rozkładu Poissona są badania W. Bortkiewicza prowadzone w celu określenia liczby przypadków śmierci kawalerzystów w korpusach pruskiej kawalerii na skutek kopnięcia w stajni przez konia. Do tego celu wykorzystał on dane zebrane dla 20 lat z 14 korpusów kawalerii i wykazał, że występowanie przypadków śmiertelnych dla tejże sytuacji jest zgodne z rozkładem Poissona. Tą metodę rosyjski badacz wykorzystał również do analizy innych zdarzeń, chociażby dla oszacowywania liczby samobójstw wśród kobiet w wybranych miastach niemieckich. Zob.: R. Łuczyński, *Badania stanów prawnych przy nabywaniu nieruchomości pod drogi krajowe*, „Przegląd Geodezyjny", 2007, 79 (3), s. 16.

Podobnie jak dla rozkładu dwumianowego, *k* oznacza liczbę sukcesów w próbie (liczbę pozytywnych wyników dla wszystkich badanych zdarzeń), natomiast n liczebność próby. W przypadku rozkładu Poissona pojawiają się dodatkowo dwa nowe parametry: λ oraz *e*, gdzie λ wyrażone jest wzorem $\lambda=n*p$, natomiast e przyjmuje stałą wartość wynoszącą w przybliżeniu2,718. Gdy znamy już te wielkości, możemy przystąpić do szacowania prawdopodobieństwa. Zasady jego obliczania są tożsame z przykładem omawianym dla rozkładu dwumianowego. Gdy znamy zatem wartości poszczególnych parametrów, po odpowiednim podstawieniu ich do wzoru, możemy obliczyć prawdopodobieństwa dla rozkładu Poissona.

15.2. Podstawy estymacji - szacowanie statystyki z próby

Badania mające na celu opisanie pewnej wybranej zbiorowości pod kątem interesującej badacza cechy są zwyczajowo prowadzone na odpowiednio dobranych próbach. Konsekwencją tego jest stosowanie właściwej nomenklatury opisującej z jednej strony - obliczenia prowadzone na próbach i z drugiej strony - wartości liczbowe charakteryzujące populację. Właściwe temu obszarowi są takie zagadnienia, jak parametr, estymator oraz estymacja, z którymi poniżej zapoznajemy początkującego badacza.

Parametrem nazywamy każdą rzeczywistą wartość cechy wyrażoną w liczbach, która charakteryzuje badaną populację generalną. Do opisu zbiorowości wykorzystujemy takie miary statystyczne, jak dla cech mierzalnych - średnią arytmetyczną, medianę, odchylenie standardowe lub wariancję oraz dla cech niemierzalnych, jakościowych - częstość występowania danej wartości zmiennej wyrażoną we frakcjach bezwzględnych (ułamkowych) lub procentowych. Z reguły jednak nie są znane wielkości parametrów, gdyż nie dysponujemy danymi od wszystkich elementów składających się na populację. Wtedy przeprowadzamy częściowe badanie statystyczne i szacujemy wartość parametru na podstawie danych z próby badawczej. Procedurę oszacowywania nieznanej wartości cechy w populacji nazywamy **estymacją**. Wartość konkretnej cechy uzyskana w toku badania nazywamy natomiast **estymatorem** określonego parametru populacji. Estymatorami nazywamy również wszelkie wyliczane statystki z próby. Estymację przeprowadzamy w dwóch trybach - po pierwsze, dla zmiennych ilościowych, czyli cech mierzalnych wyrażonych na skali interwałowej bądź ilorazowej, po drugie - dla zmiennych jakościowych określonych na nominalnym bądź porządkowym poziomie pomiaru. W pierwszym przypadku przeprowadzamy oszacowywania nieznanych wartości parametru w populacji na podstawie statystyk z próby; przykładowo - wyliczamy średnią arytmetyczną, odchylenie standardowe bądź wariancję. Ten tryb nazywamy **estymacją parametryczną**. Dla zmiennych jakościowych wyznaczamy natomiast rozkład badanych cech poprzez estymację frakcji, czyli określamy liczbę wystąpień danej wartości cechy. Tą procedurę nazywamy natomiast **estymacją nieparametryczną**. Poniżej przyjrzymy się bliżej kwestii przeprowadzania estymacji parametrycznej, w ramach której wyróżniamy **estymację punktową** oraz **estymację przedziałową.**

15.2.1. Estymacja punktowa

Estymacja punktowa jest metodą obliczania jednej, konkretnej wartości parametru populacji, przykładowo średniej wieku, średniego dochodu netto bądź średnich wydatków ponoszonych przez osoby kandydujące w wyborach parlamentarnych na prowadzenie kampanii. Obliczamy ją na podstawie danych uzyskanych od jednostek w próbie badawczej. Przykładem estymacji punktowej przeprowadzonej przy pomocy programu PSPP będzie każde zestawie statystyk opisowych dla zmiennej mierzonej na poziomie ilościowym. Jak pamiętamy, wymaga to wybrania w programie PSPP *Analyze* ⇨ *Descriptive Statistic*

⇨ *Frequencies*. Przykładowe wyniki dla estymacji punktowej dla zmiennej wiek w latach w badaniu PGSW prezentuje poniższy zrzut ekranowy.

wiek

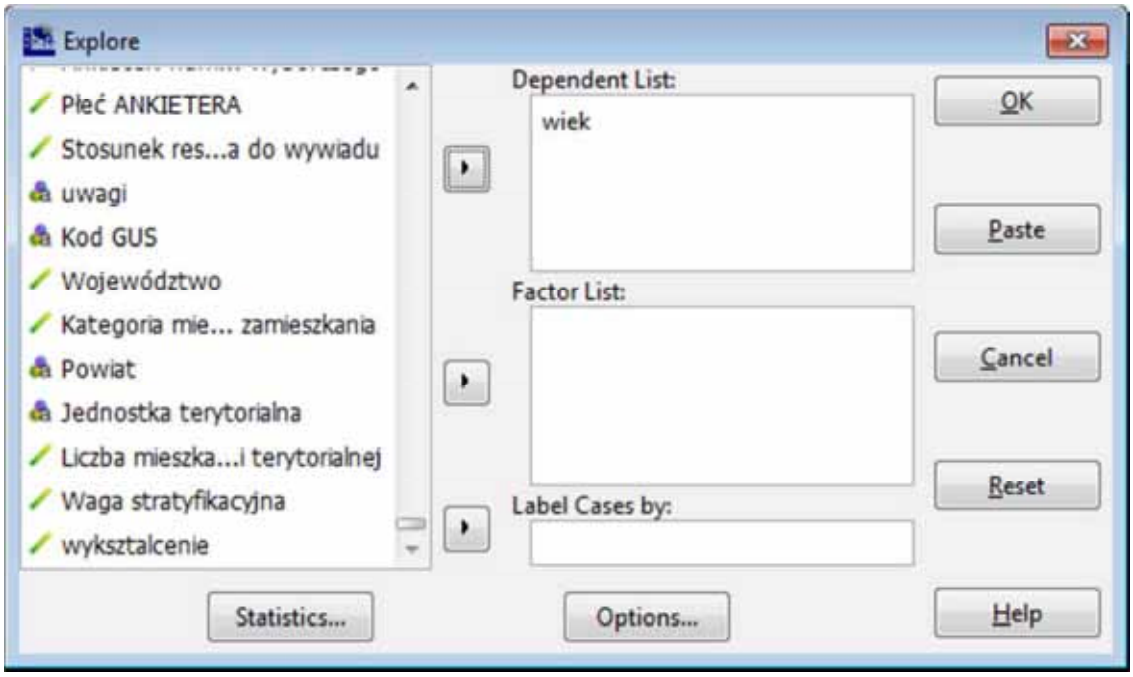
N	Valid	1797
	Missing	20
Mean		49.28
S.E. Mean		.39
Mode		53.00
Std Dev		16.70
Variance		278.84
Kurtosis		-.73
S.E. Kurt		.12
Skewness		.04
S.E. Skew		.06
Minimum		18.00
Maximum		92.00
Percentiles	50 (Median)	50.00

15.2.2. Estymacja przedziałowa

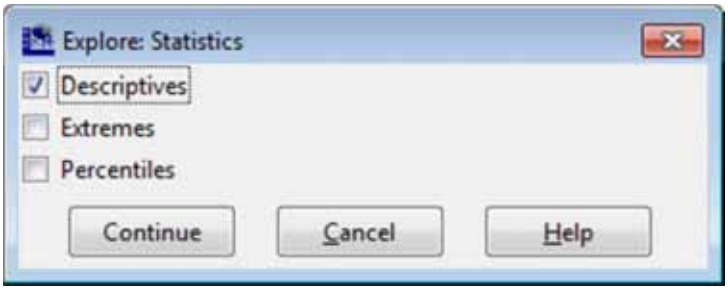
W praktyce prawdopodobieństwo, iż szacowany estymator punktowy przyjmie dokładną wartość parametru populacji, jest bliska zeru. Do wyliczenia parametru wykorzystuje się wyniki z próby losowej, które zawsze związane są z możliwością popełnienia błędu. Wartość parametru danej cechy w populacji różni się od jego wartości wyliczonej z danych uzyskanych od osób badanych. Z tego też względu powszechniejszą metodą wyznaczania jego wartości jest przeprowadzanie **estymacji przedziałowej**. Polega ona na wskazaniu przedziału liczbowego, w którym z określonym prawdopodobieństwem odnajdziemy rzeczywistą wartość parametru populacji. Jak dostrzegamy, prowadząc jakiekolwiek obliczenia na podstawie próby losowej, nie jesteśmy w stanie zagwarantować 100-procentowej dokładności obliczeń. Dlatego wskazujemy przedział liczbowy pokrywający wartość nieznanego parametru badanej zbiorowości generalnej. Ponadto, wyniki oszacowań prezentujemy za pomocą podania stopnia ich prawdopodobieństwa, czyli przekładalności na faktyczną wartość cechy. Zakres liczbowy, w którym mieści się wartość szacowanego parametru nazywamy przedziałem ufności. Jego pomysłodawcą jest statystyk polskiego pochodzenia - Jerzy Spława-Neyman. Do wyznaczenia prawdopodobieństwa wykorzystujemy założenia poszczególnych rozkładów teoretycznych dla zmiennych z próby. Szczególnie istotna dla przeprowadzenia estymacji przedziałowej jest wiedza o rozkładzie normalnym zmiennej. Poniżej wprowadzamy kilka kluczowych zagadnień, niezbędnych dla zrozumienia logiki przeprowadzania estymacji przedziałowej.

Przypuśćmy, iż interesuje nas wyznaczenie średniej wieku dla elektoratu pewnej partii. Pierwszym krokiem oszacowania tego parametru jest ustalenie współczynnika wiarygodności wyników. Jest to wspomniany wcześniej poziom prawdopodobieństwa pokrycia szacowanej wartości średniej w populacji, co wyrażane jest symbolem (1-α). W tym przypadku α oznacza poziom ufności i wskazuje na stopień ryzyka popełnienia błędu i nieprawidłowego ustalenia średniej wieku dla badanej zbiorowości. Najczęściej w analizach ilościowych spotykamy trzy wartości współczynnika ufności: 90 proc., 95 proc. i 99 proc. Oznacza to, iż w trakcie procedury ustalania przedziału ufności dla nieznanego parametru populacji na podstawie wyników z wielu prób o tej samej liczebności, mamy 90, 95 bądź 99 przypadków na 100, które posiadają wartość bliską nieznanej wartości średniej wieku. Jednakże, szacunki z próby zawsze są obciążone błędem. Odpowiednio, w przypadku przedziału ufności 90 proc., ryzyko pomyłki wynosić będzie 10 proc., natomiast dla wartości 95 proc. - 5 proc. oraz 99 proc. - 1 proc. W programie

PSPP z łatwością można wykonać estymację przedziałową dla średniej. W naukach społecznych przyjmuje się, iż współczynnik ufności wynosi 95 proc. Przyjmijmy, że interesuje nas ustalenie poziomu ufności dla średniej wieku w populacji dorosłych Polaków. W tym celu należy wybrać *Analyze* ⇨ *Descriptive Statistics* ⇨ *Explore*. Pojawia się wówczas następujące okno:



W polu o nazwie *Dependent List* wprowadzamy analizowaną zmienną wiek w latach, którą uprzednio spreparowaliśmy ze zmiennej m1 (rok urodzenia). Następnie w opcji *Statistics* zaznaczamy *Descriptives*, co przedstawia poniższy zrzut ekranowy:



Efektem wykonania tej czynności jest uzyskanie podstawowych miar statystycznych dla zmiennej wiek, które są publikowane w poniższej formie w oknie raportu:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
wiek	1797	100%	0	0%	1797	100%

Descriptives

		Statistic	Std. Error
wiek Mean		49.28	.39
95% Confidence Interval for Mean	Lower Bound	48.51	
	Upper Bound	50.05	
5% Trimmed Mean		49.12	
Median		50.00	
Variance		278.84	
Std. Deviation		16.70	
Minimum		18.00	
Maximum		92.00	
Range		74.00	
Interquartile Range		25.00	
Skewness		.04	.06
Kurtosis		-.73	.12

Do przedstawienia wyników estymacji przedziałowej dla zmiennej wiek niezbędne są cztery rodzaje danych z powyższej tabeli: średnia arytmetyczna (*Mean*), wartość dolnego krańca przedziału ufności (*95% Confidence Interval for Mean Lower Bound*), wartość górnego krańca przedziału (*Upper Bound*) oraz błąd oszacowania dla średniej (*Std. Error*). Z uzyskanych wyników dowiadujemy się, iż średnia wieku wynosi 49,3 lata. Jest to estymator punktowy. Dla estymacji przedziałowej wyznaczamy zaś zakres liczbowy, w którym mieści się wartość średniej w populacji, czyli dolny i górny kraniec przedziału ufności. W naszym przykładzie ostatecznie wnioskujemy, że z 95 proc. pewnością jesteśmy w stanie uznać, że nieznana średnia wieku w populacji mieści się w granicach przedziału od 48,51 lat do 50,05 lat, przy czym błąd oszacowania tego parametru wynosi d=0,39.

15.2.3. Wyznaczanie minimalnej liczebności próby

Jak wiemy, badania dotyczące sfery życia politycznego lub społecznego przeprowadzamy nie na całych zbiorowościach ludzkich, ale na pewnych ich wycinkach, czyli na próbach badawczych. We wnioskowaniu statystycznym niezmiernie ważne jest, aby dobór jednostek do owej puli miał charakter losowy. Każdy element składający się na badaną populację powinien posiadać jednakową szansę trafienia do próby badawczej. Przed przystąpieniem do bezpośredniego dobierania owych jednostek, istotne jest określenie minimalnej ich liczebności. Związane jest to z problemem dokładności oszacowań parametrów populacji - wartości średniej arytmetycznej, wariancji, odchylenia standardowego, a także częstości występowania poszczególnych wariantów badanych cech - frakcji. Jak wiemy, obliczenia prowadzone na próbach losowych nie odzwierciedlają idealnie rzeczywistej wartości cechy w populacji. Precyzja obliczeń zależy w dużej mierze od liczebności próby - im większą liczbę jednostek z populacji uda nam się przebadać, tym wartość obliczanego parametru będzie dokładniejsza i w mniejszym stopniu różnić się będzie od stanu faktycznego. Dążymy tym samym do minimalizacji poziomu błędu oszacowania. W niniejszej części rozdziału omawiamy procedurę obliczania minimalnej liczebności próby badawczej. O ile koncentrujemy się na dostarczeniu badaczowi praktycznych rad wyznaczenia owej liczebności, musimy wpierw poznać pewne teoretyczne założenia, które pozwalają na dokonywanie takich oszacowań.

Możemy spotkać się z trzema podstawowymi sposobami obliczania minimalnej liczebności próby. Każdy z nich wymaga wpierw dokonania przeglądu danych, jakimi dysponujemy na temat badanej populacji. Minimalną liczebność próby możemy określić bowiem na podstawie, **po pierwsze, znanej wielkości odchylenia standardowego dla szacowanej średniej, po drugie, gdy nie znamy wielkość odchylenia**

standardowego dla szacowanej średniej, po trzecie, za pomocą estymacji frakcji czyli rozkładu częstości występowania danej cechy w populacji. Ten ostatni sposób obliczania liczebności próby jest najczęściej praktykowany w badaniach o tematyce politycznej bądź społecznej. Zazwyczaj bowiem nie posiadamy wiedzy na temat wielkości odchylenia standardowego dla średniej. Ponadto, w badaniach społeczno-politycznych rzadko dążymy do uzyskania ściśle kwantytatywnych informacji mierzonych na skalach ilorazowych bądź interwałowych. Częściej interesują nas dane o charakterze jakościowym - sądy, opinie, poglądy czy też charakterystyki opisowe i ich rozkłady w populacji. W większości przypadków badamy częstość występowania danych zjawisk, rzadziej obliczamy średnie i odchylenia od tej miary. Dlatego minimalna liczebność próby wyznaczana jest głównie na podstawie frakcji.

Obliczanie minimalnej liczebności próby w każdym z tych trzech przypadków opiera się na założeniu, iż poddawane pomiarowi rozkłady zmiennych losowych będą lub powinny przybierać postać rozkładu normalnego. Ustalenie liczby jednostek, które powinny zostać objęte badaniem, dokonywane jest na podstawie kilku kluczowych miar tożsamych dla tego rozkładu. Wybór określonych wartości dla tych miar wpływa bowiem na precyzję oszacowywanych statystyk. Do tychże miar należy **przedział ufności (1-α), poziom ufności (α), statystyka Z nazywana dystrybuantą rozkładu normalnego** oraz **wielkość maksymalnego błędu oszacowania (d).** W zależności zaś od posiadanych informacji o populacji generalnej wybieramy odpowiednią metodę wyznaczania liczebności próby - ze znaną bądź nieznaną wartością odchylenia standardowego lub też w oparciu o zakładany rozkład frakcji.

O wartości przedziału ufności oraz wielkości maksymalnego błędu oszacowania decydujemy samodzielnie. W przypadku przedziału ufności najczęściej rozważa się trzy warianty jego zakresu - 0,90, 0,95 bądź 0,99. Jeżeli zdecydujemy się na pierwszą wersję, wówczas uzyskujemy 90 proc. pewności, iż uzyskane w badaniu wyniki będą odzwierciedlać wartości parametrów w populacji. W tym przypadku wartość dystrybuanty rozkładu normalnego wynosić będzie Z_{α}=1,64, natomiast przy 95-procentowej pewności - Z_{α}=1,96. Decydując się na przedział ufności o wartości 0,99, zwiększamy pewność oszacowań statystyk z próby do 99 proc. Dla tego wariantu wartość Z_{α} wynosi 2,58. Wartość maksymalnego błędu oszacowania wskazuje, na ile wyniki oszacowań z próby mogą odbiegać od faktycznej wartości w populacji. W tym przypadku również samodzielnie podejmujemy decyzję o tym, jaki stopień pomyłki akceptujemy. Maksymalny błąd oszacowania wyrażany jest w postaci liczb ułamkowych i mieści się w przedziale od 0 do 1. W praktyce badawczej i analitycznej umownie przyjmuje się, iż jego wartość nie powinna jednak przekraczać 0,1. Jeżeli przyjmiemy błąd na poziomie 0,05 oznacza to, że konkretne uzyskane w toku badania wyniki mogą odbiegać od ich wartości w populacji o ± 5 proc. Przypuśćmy, że w sondażu wyborczym poparcie dla partii X wynosi 30 proc. przy błędzie oszacowania równym 5 proc. Oznacza to, iż faktyczne poparcie dla partii może być mniejsze i wynosić 28,5 proc. bądź większe i sięgać 31,5 proc. Poniżej prezentujemy poszczególne metody obliczania minimalnej liczebności próby omawiając je na przykładach.

15.2.3.1. Wyznaczenie minimalnej liczebności próby dla estymacji średniej (m) w populacji przy znanym odchyleniu standardowym (σ)

Wyobraźmy sobie, iż nasze badanie wymaga określenia średniej wieku elektoratu partii X. Wiemy jednocześnie, że odchylenie standardowe dla średniej wynosi 1,3 lat. Chcemy mieć przy tym 95 proc. pewności (Z_{α} wynosi zatem 1,96), iż wyniki z próby badawczej będą odzwierciedlać średnią wieku w populacji i nie będą odbiegać od jej faktycznej wartości o więcej niż 5 punktów proc. (błąd

oszacowania wynosi zatem 5 proc.). W takiej sytuacji skorzystamy z następującego wzoru na minimalną liczebność próby:

$$n = \frac{Z_{\alpha}^2 * \sigma^2}{d^2}$$

W powyższym wzorze pod symbolem Z_{α} wstawiamy wartość 1,96, co jest właściwością przyjętego przedziału ufności wynoszącego 0,95. Symbolem σ oznaczamy odchylenie standardowe, które wynosi 1,3, natomiast symbolowi d przypisany jest maksymalny błąd oszacowania, który ustaliliśmy na poziomie 5 proc.; w wartościach bezwzględnych równy jest on 0,05. Podstawiające te wartości pod wzór uzyskujemy następujący wynik:

$$n = \frac{1{,}96^2 * 1{,}3^2}{0{,}05^2} = \frac{3{,}8416 * 1{,}69}{0{,}0025} = \frac{6{,}492304}{0{,}0025} = 2596{,}92$$

Obliczona minimalna liczebność próby dla estymacji średniej ze znanym odchyleniem standardowym dla powyższego przykładu wyniosła 2597 (uzyskaną wartość zawsze zaokrąglamy do góry do liczby całkowitej). Oznacza to, iż aby zachować 95 proc. pewności oszacowań średniej wieku elektoratu partii X mieszczącą się jednocześnie w granicy błędu wynoszącym 5 proc., musimy przebadać 2597 osób. Zauważmy jednak, iż z reguły nie znamy wartości odchylenia standardowego dla średniej w populacji. W takiej sytuacji korzystamy z drugiego sposobu wyznaczania minimalnej liczebności próby, który prezentujemy poniżej.

15.2.3.2. Wyznaczenie minimalnej liczebności próby dla estymacji średniej (m) w populacji z nieznanym odchyleniu standardowym (σ)

Gdy nie znamy odchylenia standardowego średniej w populacji, wówczas do wyliczenia minimalnej liczebności próby wykorzystujemy wyniki innych badań lub też przeprowadzamy wstępne badania pilotażowe w celu określenia odchylenia standardowego z próby losowej (s). W sytuacji, gdy żaden z parametrów właściwych rozkładowi normalnemu nie jest znany - czyli średnia (m) oraz odchylenie standardowe w populacji (σ) - do określenia minimalnej liczebności próby wykorzystujemy wartości właściwe rozkładowi t-Studenta dla n-1 stopni swobody. Przypuśćmy, iż celem naszego badania jest określenie przeciętnej liczby godzin spędzanej na czytaniu prasy politycznej w tygodniu. Z innych badań wiemy, iż Polacy poświęcają na ten cel ok. 3,5 godziny tygodniowo, przy czym wartość ta różni się o ± 1,2 godziny. Tyle wynosi bowiem odchylenie standardowe (s). Badanie zostało przeprowadzone na próbie n=20. Liczebność próby chcemy wyznaczyć dla poziomu ufności wynoszącego 0,05 (przedział ufności wynosi zatem 0,95) oraz dla błędu oszacowania wynoszącego d=0,25 godziny. Dla takiej sytuacji skorzystamy z poniższego wzoru na obliczenie liczebności próby badawczej:

$$n = \frac{t_{\alpha,n-1}^2 * S^2}{d^2}$$

Wartość statystyki t odnajdujemy w tablicach rozkładu t-Studenta. W naszym przypadku będziemy poszukiwać wartości dla poziomu ufności α=0,05 oraz dla df=n-1 stopni swobody, czyli df=20-1, gdyż próba badawcza, według której określiliśmy odchylenie standardowe (s) liczyła n=20. Wartość t dla tych

warunków będzie wynosiła 1,725. Posiadając wszelkie niezbędne dane możemy przystąpić do obliczenia liczebności próby dla powyższych założeń:

$$n = \frac{1{,}725^2 * 1{,}2^2}{0{,}25^2} = \frac{2{,}976 * 1{,}44}{0{,}0625} = \frac{4{,}285}{0{,}0625} = 68{,}57$$

W celu określenia średniej liczby godzin spędzanej na czytaniu prasy politycznej przez Polaków powinniśmy przebadać ± 69 osób. Jeżeli nie znamy zarówno wartości odchylenia standardowego w próbie bądź w populacji, wówczas należy skorzystać z metody wyznaczania minimalnej liczebności próby na podstawie frakcji. Czytelnik omówienie tej metody odnajdzie w kolejnym podrozdziale.

15.2.3.3. Wyznaczenie minimalnej liczebności próby dla estymacji frakcji (p)

Kolejna metoda wyznaczania minimalnej liczebności próby opiera się na estymacji frakcji. Frakcją nazywamy wyrażony w wartościach bezwzględnych (ułamkowych) udział lub częstość występowania danej cechy w populacji. Przykładem frakcji jest stosunek liczby mężczyzn względem liczby kobiet w populacji dorosłych Polaków, który może wynosić odpowiednio 0,45 oraz 0,55. Niekiedy zamiast pojęcia „frakcja" używa się terminu „proporcja". Współczynnik frakcji oznaczamy symbolem p. Przyjmuje on wartości z przedziału od 0 do 1. Przypuśćmy, iż naszym celem jest określenie udziału osób darzących zaufaniem polityka X. Zakładamy, że takich osób w populacji jest blisko 40 proc. (w wartościach bezwzględnych wartość ta wynosi 0,4). Dla oszacowań z próby chcemy zagwarantować 95-procentowy poziom pewności, natomiast błąd szacunku ma nie przekraczać 5 proc. Posiadając takie informacje, skorzystamy z następującego wzoru na obliczenie minimalnej liczebności próby:

$$n = \frac{Z_\alpha^2 * p * (1 - p)}{d^2}$$

Podstawiając wskazane wcześniej wielkości pod powyższy wzór, uzyskujemy następujący wynik:

$$n = \frac{1{,}96^2 * 0{,}4 * (1 - 0{,}4)}{0{,}05^2} = \frac{3{,}8416 * 0{,}4 * 0{,}6}{0{,}0025} = \frac{0{,}921984}{0{,}0025} = 368{,}79$$

W sytuacji znanej frakcji osób ufających politykowi X dla poziomu ufności 0,95 oraz dla maksymalnego błędu oszacowania równego 0,05, niezbędna liczebność próby wynosi 369 jednostek. Zauważymy jednak, iż projektując wszelkie badania, nie posiadamy praktycznie żadnych informacji o interesującej nas populacji. Korzystając z metody wyznaczania minimalnej liczebności próby na podstawie frakcji, nie jest konieczne posiadanie wiedzy w zakresie rozkładu badanych cech. Przyjmuje się, że p wynosi 0,5, gdyż wtedy wartość iloczynu p*(1-p) jest największa ze wszystkich możliwych kombinacji i równa 0,25. Jeżeli zatem nie znalibyśmy frakcji osób w populacji ufających politykowi X, przyjęlibyśmy następujący sposób obliczania minimalnej liczebności próby:

$$n = \frac{1{,}96^2 * 0{,}5 * (1 - 0{,}5)}{0{,}05^2} = \frac{3{,}8416 * 0{,}5 * 0{,}5}{0{,}0025} = \frac{0{,}9604}{0{,}0025} = 384{,}16$$

W tym przypadku próba badawcza wymagałaby objęcia 384 jednostek. Jak możemy zauważyć, brak znajomości rozkładu frakcji w populacji, powoduje zwiększenie liczebności próby względem sytuacji, gdy go znamy. Zauważmy ponadto, iż wszystkie trzy sposoby obliczania minimalnej liczebności próby

zakładają brak wiedzy o wielkości populacji. Jeżeli jednak populacja ma postać skończoną, tzn. znamy liczbę jednostek składających się na nią, zastosujemy niniejszy wzór na wyznaczenie liczebności próby:

$$n = \frac{p(1-p) * Z_{\alpha}^2 * N}{(d^2 * N) + [Z_{\alpha}^2 * p(1-p)]}$$

Dla przykładu podejmijmy się celu wyznaczenia minimalnej liczebności próby dla znanej nam populacji. Wyobraźmy sobie, iż chcemy przeprowadzić badanie wśród posłów na Sejm RP. Jak wiemy, liczba wszystkich posłów wynosi 460. Jest to więc populacja skończona. Obliczmy minimalną liczbę posłów, którzy powinni zostać objęci badaniem. Załóżmy, iż chcemy przyjąć poziom ufności wynoszący 0,95 oraz maksymalny błąd oszacowania 0,03. Nic nie wiemy o rozkładzie frakcji w populacji, zatem p=0,5. Procedura obliczeniowa przyjmie następującą postać:

$$n = \frac{0{,}5(1-0{,}5) * 1{,}96^2 * 460}{(0{,}03^2 * 460) + [1{,}96^2 * 0{,}5(1-0{,}5)]} = \frac{0{,}5^2 * 1{,}96^2 * 460}{(0{,}03^2 * 460) + (1{,}96^2 * 0{,}5^2)} = \frac{441{,}78}{1{,}37} = 321{,}46$$

Liczbę posłów, którą powinniśmy objąć badaniem wynosi 321. Niekiedy w takich sytuacjach musimy zastosować czynnik korekty liczebności próby dla małolicznych populacji. Stosujemy go wszędzie tam, gdzie badana próba stanowi ponad 5 proc. populacji. Jest to szczególnie uzasadnione w przypadku badania grup trudnodostępnych, jak chociażby osób publicznych czy też kadry menadżerskiej. Na podstawie współczynnika korygującego obliczamy wielkość próby rewidowanej, czyli nowej próby, która zostanie pomniejszona o odpowiednią liczbę jednostek względem próby pierwotnej. Skorygowaną liczebność próby wyliczamy z następującego wzoru:

$$n' = \frac{n * N}{N + n + s}$$

Symbol n' oznacza wielkość próby rewidowanej, n - liczebność próby pierwotnej wyznaczonej na podstawie wzoru dla minimalnej liczebności próby dla estymacji frakcji, N - wielkość badanej populacji, s - przedział ufności wyrażany wzorem: s = pierwiastek z p(1-p)/n, gdzie p w przedziale ufności ustalamy na poziomie najbardziej niekorzystnym, ale bezpiecznym dla badacza, czyli 0,5 (zakładamy największe zróżnicowanie badanej populacji). Dla powyższego przykładu uzyskujemy następującą liczebność próby skorygowanej:

$$n' = \frac{321 * 460}{321 + 460 + 0{,}0630} = 188{,}8$$

Uzyskana wartość wskazuje na minimalną liczebność próby po korekcie w celu zagwarantowania istotności wyników na poziomie 0,05. Badanie tak trudno dostępnej grupy, jaką są posłowie Sejmu RP nie wymaga zatem dotarcia do 321 osób piastujących mandat, ale do 189. Z tego względu współczynnik korygujący odnajduje praktyczne zastosowanie w projektowaniu badań empirycznych.

15.3. Weryfikacja hipotez statystycznych

Przeprowadzając analizy statystyczne często wykraczamy poza klasyczne metody statystyki opisowej i sięgamy po narzędzia opracowane przez dział statystyki matematycznej, precyzując - po testy

istotności. Ich zastosowanie umożliwia przekładanie wyników z próby na populację. W niniejszej części rozdziału zaznajamiamy Czytelnika z elementarną wiedzą w zakresie stosowalności tych metod oraz narzędzi. Na wstępie warto zasygnalizować, iż spektrum testów istotności stosowanych w statystyce jest bardzo szerokie. W zależności od podjętej problematyki badawczej i wyznaczanych przez nią celów analitycznych, dobieramy właściwe narzędzia i testy służące do prawidłowego uzasadnienia wyników badania. Poznamy je w dalszych częściach niniejszego podręcznika. W tym miejscu wprowadzamy badacza w pewne ogólne ramy stosowania testów istotności. Ich wykorzystanie wymaga uprzedniego poznania schematu formułowania i sprawdzania przypuszczeń o zjawiskach oraz cechach badanych zbiorowości. Procedurę tą nazywamy weryfikacją hipotez statystycznych. Ma ona postać ogólnych i powszechnych zasad, do których stosujemy się każdorazowo podczas wykorzystywania testów statystycznych. Ponadto zaznajamiamy Czytelnika z podstawowymi pojęciami, miarami oraz schematami związanymi z wykorzystaniem owych testów w praktyce.

15.3.1. Zasady weryfikacji hipotez statystycznych

Zadaniem testów istotności jest sprawdzenie naszego przypuszczenia o wartości bądź rozkładzie danej cechy w populacji. Takie przypuszczenie nazywamy hipotezą statystyczną, natomiast procedurę ich sprawdzania - weryfikacją. Czynność ta ogranicza się do kilku podstawowych i stosowanych każdorazowo podczas wykorzystywania testów istotności zasad.

Gdy zdecydujemy się na zastosowanie testów istotności, w pierwszym kroku zawsze sformułujemy przypuszczenia na temat charakterystyki populacji. Są to z reguły przypuszczenia o częstości występowania danego zjawiska lub wartości pewnej cechy. Następnie sprawdzamy ich słuszność w toku procedury nazywanej weryfikacją hipotez statystycznych. Hipotezy statystyczne występują w dwóch postaciach. Pierwsza hipoteza to hipoteza zerowa oznaczana symbolem H_0. Druga hipoteza to hipoteza alternatywna zwana również hipotezą badawczą bądź hipotezą roboczą. Oznaczamy ją H_1. Gdy chcemy sprawdzić, czy wyniki z próby odzwierciedlają cechy populacji, czyli czy są one statystycznie istotne, formułujemy hipotezy i poddajemy je weryfikacji za pomocą testów istotności. O ile w zależności od rodzaju testu porównujemy różne miary statystyczne ze sobą, to jednak schemat procedowania jest taki sam. Pierwszym krokiem jest postawienie hipotezy zerowej (H_0). W jej przypadku zakładamy, że pewna statystyka z próby, chociażby średnia arytmetyczna, mediana bądź odchylenie standardowe przyjmuje taką samą wartość jak w populacji.

Przykładowo, jeżeli będziemy chcieli sprawdzić ile wynosi średnia wieku w zbiorowości dorosłych Polaków, hipotezę zerową sformułujemy następująco:

H_0 - średnia wieku w próbie losowej jest równa średniej wieku w populacji. Zazwyczaj w podręcznikach statystycznych spotkamy się z następującym zapisem hipotezy zerowej - H_0: $\bar{x}$=m, gdzie $\bar{x}$ oznacza średnią wieku w próbie badawczej, natomiast m - średnią wieku w populacji. Hipoteza alternatywna, jak sama nazwa wskazuje, będzie przeciwieństwem hipotezy zerowej. W najprostszej postaci hipoteza alternatywna będzie mówiła, że statystyka z próby nie jest równa wartości parametru populacji.

Tak sformułowana hipoteza nazywana jest również hipotezą prostą i zapiszemy ją w następującej postaci:

H_1 – średnia wieku w próbie losowej nie jest równa średniej wieku w populacji. Hipoteza alternatywna prosta może również przyjąć następujący zapis – H_1: $\bar{x} \neq m$. Jej złożona postać może z kolei występować w dwóch wariantach. Po pierwsze, wartość statystki z próby może być większa od wielkości parametru populacji (H_1: $\bar{x} > m$), po drugie – wartość statystki z próby może być mniejsza od wielkości parametru populacji (H_1: $\bar{x} < m$). W przypadku badania hipotezy prostej zastosujemy dwustronny test istotności, z kolei w przypadku hipotezy złożonej, test lewo- bądź prawostronny. Schemat wyboru jednej z tych dwóch statystyk poznamy w dalszej części niniejszego rozdziału. Wymaga to bowiem wpierw zaznajomienia się z podstawowymi pojęciami i miarami niezbędnymi w interpretacji wyników testów istotności – przedziałem ufności oraz poziomem ufności.

Gdy wyznaczymy hipotezy statystyczne, w kolejnym etapie przystępujemy do ich weryfikacji. W najprostszym ujęciu oznacza to, iż dążymy do odrzucenia hipotezy zerowej i do przyjęciu hipotezy alternatywnej. W rzeczywistości bowiem jesteśmy zainteresowani statystycznym uzasadnieniem prawdziwości hipotezy alternatywnej, co możemy jedynie uzyskać poprzez przeprowadzenie dowodu pośredniego odrzucającego hipotezę zerową. Dokonujemy tego na podstawie testu istotności, interpretacji właściwych mu statystyk oraz założenia o istotności tych wyników. Pamiętajmy jednak, że w analizach statystycznych zawsze jesteśmy narażeni na popełnienie pewnych błędów obliczeniowych, natomiast nasze oszacowania są przyjmowane z określonym prawdopodobieństwem ich faktycznego obowiązywania.

W weryfikacji hipotez statystycznych mamy bowiem każdorazowo do czynienia z dwoma typami błędów. Nazywamy je **błędem I (α)** oraz **błędem II (β) rodzaju**. Rozważmy dwie sytuacje, w których one występują. Przypuśćmy, że wyniki analizy nakazują nam odrzucić hipotezę zerową. Musimy być jednak świadomi, iż mogliśmy popełnić błąd na różnych etapach badania, przez co wyniki analiz mogły stać się nieprawdziwe i nie prezentować tego, co faktycznie występuje w rzeczywistości społecznej. Konsekwencją tego jest odrzucenie hipotezy zerowej, która jednak mówiła prawdę o populacji. Tą sytuację nazywamy błędem I rodzaju i oznaczamy go symbolem α. Druga sytuacja stanowi przeciwieństwo pierwszej. Przyjmujemy bowiem hipotezę zerową, które *de facto* jest fałszywa i sprawia, że przyjmujemy nieprawdziwy ogląd na rzeczywistość społeczną. W tym przypadku mamy do czynienia z błędem II rodzaju, dla oznaczenia którego stosujemy symbol β. We wnioskowaniu statystycznym nigdy nie jesteśmy zatem pewni naszych oszacowań w 100 procentach. Zawsze weryfikację hipotez statystycznych przeprowadzamy z określonym prawdopodobieństwem popełnienia błędu. Każdorazowo dążymy jednak do jego minimalizacji. W praktyce statystycznej kluczową rolę odgrywa błąd I rodzaju. W dostępnych bowiem testach istotności przyjmuje się regułę arbitralnego określenia wielkości tego błędu, a związanego z prawdopodobieństwem odrzucenia hipotezy zerowej, która była hipotezą prawdziwą. W nomenklaturze statystycznej ten błąd nazywany jest **poziomem ufności (α)**. Jest to bowiem pewne ryzyko, które godzimy się zaakceptować w sytuacji, gdy w toku oszacowań odrzuciliśmy hipotezę prawdziwą. Wyznaczamy przy tym jednocześnie tzw. **przedział ufności (1-α).** Określamy pewien obszar wartości dla wyników testów istotności, którym najprościej mówiąc „ufamy" i które mówią nam, że zaobserwowane przez nas różnice w wynikach z próby to wyłącznie efekt oddziaływania przypadku i nie dają się one przełożyć na populację. Gdy zaś statystyka testu wykracza poza ten obszar, należy przypuszczać, iż w populacji doszło do pewnych zmian, o czym informują obliczenia poczynione na danych z próby losowej. Owym obszarem marginalnym jest interesujący nas w testach statystycznych poziom ufności (α). Ustalamy go arbitralnie. Z reguły przyjmujemy trzy jego wartości 0,1, 0,05 lub 0,01. W drugim przypadku,

który jest spotykany najczęściej, wyznaczamy 5-procentowy obszar dla wyników testu uzasadniający odrzucenie hipotezy zerowej i umożliwiający nam przeniesienie wyników z próby na populację. Jeżeli zaś statystyka testu będzie znajdowała się poza tą granicą, czyli w 95-procentowym przedziale ufności, wówczas przyjmujemy hipotezę zerową, gdyż nie mamy podstaw do jej odrzucenia. Jak można zauważyć, decydującym czynnikiem wnioskowania statycznego jest zidentyfikowanie, w którym z tych obszarów mieści się wynik obliczeniowy wykorzystanego przez nas testu istotności. Zbiór takich wartości dla każdego testu nazywamy obszarem krytycznym. W praktyce, program PSPP wylicza wszelkie niezbędne statystyki. Procedura zaś weryfikacji hipotez statystycznych wpisuje się w bardzo prosty schemat interpretacji wyników. Każdorazowo rozważamy wyłącznie dwie sytuacje, zwracając uwagę na wartość uzyskanego w teście poziomu istotności (p). Prezentują się one następująco:

Decyzja o odrzuceniu hipotezy zerowej - jeżeli wartość $p<0,05$ to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną. Oznacza to, że wyniki z próby losowej są istotne statystycznie i odzwierciedlają wartość badanej cechy w populacji. Wnioskujemy, iż wyniki z próby badawczej możemy przenieść na szerszą zbiorowość osób nieobjętych bezpośrednio badaniem.

Decyzja o przyjęciu hipotezy zerowej - jeżeli wartość $p\geq0,05$ to uznajemy, że nie mamy podstaw do odrzucenia hipotezy zerowej. Nie mamy bowiem przesłanek, aby sądzić, iż zaobserwowane różnice wartości estymatora z próby losowej występują w rzeczywistości. Najprawdopodobniej są one wynikiem przypadku bądź błędów, które mogliśmy popełnić w toku badania. Oznacza to, że wyniki z próby nie są istotne statystycznie i nie mogą być uogólnione na populację.

O ile udało nam się zapoznać z procedurą weryfikacji hipotez i zasadami wnioskowania statystycznego w oparciu o wyniki testów istotności, o tyle ważne jest również pamiętanie o dwóch typach testów - o teście dwustronnym oraz testach jednostronnych (prawo- bądź lewostronnym). Zwyczajowo w analizie danych najczęściej korzystamy z testów dwustronnych. W jego przypadku odrzucamy zarówno wartości ekstremalne niedoszacowane względem wartości oczekiwanej - znajdujące się po lewej stronie przedziału, bądź przeszacowane - znajdujące się po prawej stronie tego przedziału. Z reguły nie wiemy, jaką wartość może przyjąć statystyka z próby. Jeżeli uznany przez nas poziom ufności wynosi 5 proc. w teście dwustronnym odrzucamy 2,5 proc. przypadków z obu stron rozkładu zmiennej. Jeżeli jednak wiemy, iż możliwe wartość cechy mogą być wyłącznie większe od jej wartości oczekiwanej, to wybieramy test prawostronny. Z kolei test lewostronny wybierzemy dla sytuacji przeciwnej - gdy wartości cechy mogą być wyłącznie mniejsze od jej wartości przewidywanej w populacji.

15.3.2. Wprowadzenie do testowania hipotez parametrycznych i nieparametrycznych

We wnioskowaniu statystycznym mamy do czynienia z dwoma rodzinami testów statystycznych - testami parametrycznymi oraz testami nieparametrycznymi. Taki podział testów odnajdziemy również w programie PSPP. W zależności od właściwości analizowanych zmiennych oraz postawionych celów porównawczych dokonujemy wyboru odpowiedniej statystyki. W niniejszej części podręcznika dostarczamy badaczowi praktycznych rad w zakresie podejmowania decyzji o zastosowaniu odpowiedniego testu statystycznego. Testy parametryczne oraz nieparametryczne różnią się bowiem znacząco między sobą. Wybór niewłaściwej statystki rzutuje na wiarygodność przedstawianych wyników analiz. Za wybór właściwego testu statystycznego odpowiada wiele czynników. Należą do nich poziom pomiaru zmiennej,

rodzaj rozkładu teoretycznego zmiennej, rodzaj szacowanych parametrów, liczebność próby oraz cel ich zastosowania, co związane jest z zakresem dokonywanych porównań.

Testy parametryczne są stosowane przede wszystkim do badania zmiennych mierzonych na poziomie interwałowym bądź ilorazowym. Oznacza to, iż poddawana pomiarowi cecha jest wyrażona za pomocą wartości liczb rzeczywistych, czego przykładem jest wzrost, waga, wiek wyrażony w latach, uzyskiwany dochód netto. Tylko dla takich wartości jesteśmy w stanie obliczyć średnią arytmetyczną (m), odchylenie standardowe (σ) oraz wariancję (σ^2). Te miary są jednocześnie statystykami, które stanowią podstawę obliczania testów parametrycznych. Są one w ich ramach szacowane i porównywane do adekwatnie wyrażonych własności w badanych populacjach. Ponadto, zmienne badane za pomocą testów parametrycznych muszą spełniać założenie o normalności rozkładu. Jeżeli to kryterium nie jest spełnione, nie mamy prawomocności zastosowania testu parametrycznego. Z reguły rozkład zmiennej przyjmuje postać rozkładu normalnego, gdy liczebność próby wynosi n>30. Bezpieczniejszą granicą jest jednak sytuacja, gdy finalna liczba jednostek w próbie mieści się w przedziale od 100 do 120. Sprawdzenie założenia o normalności rozkładu zmiennej przeprowadza się również za pomocą odpowiednich metod i narzędzi analitycznych dostępnych w pakiecie PSPP. Do tego celu służy metoda oceny wizualnej histogramu z nałożoną krzywą normalną, ocena miar asymetrii - współczynnika skośności oraz kurtozy, a także testy statystyczne, w tym test Kołmogorowa-Smirnowa. Szczegółowe mówienie procedur weryfikacji założenia o normalności rozkładu z wykorzystaniem tychże metod, zostały wyłożono w kolejnych rozdziałach podręcznika. Spełnienie warunku o normalności rozkładu jest na tyle istotne, iż zmienne choć mierzone na skali interwałowej bądź ilorazowej, nie mogą być badane przy pomocy testów parametrycznych. W takiej sytuacji alternatywnie korzystamy z testów nieparametrycznych.

Testy nieparametryczne są testami stosowanymi w sytuacji niemożności wykorzystania testu parametrycznego. Ich cechą charakterystyczną jest liberalność, która objawia się w małej liczbie założeń koniecznych do spełnienia. Rzutuje to jednak na jakość dokonywanych obliczeń. Testy nieparametryczne charakteryzują się mniejszą dokładnością oszacowań oraz posiadają mniejszą mocą testów. Gwoli przypomnienia zaznaczmy, iż moc testów to prawdopodobieństwo popełnienia błędu II rodzaju (β), czyli szansy przyjęcia hipotezy zerowej, która jest hipotezą fałszywą. W przypadku testów parametrycznych ryzyko popełnienia tego błędu jest znacznie mniejsze niż w testach nieparametrycznych. Jednakże praktyka badawcza wskazuje, iż testy nieparametryczne są częściej stosowane w badaniach społecznych. W politologicznych oraz socjologicznych analizach danych ilościowych pomiarowi poddajemy bowiem zmienne o charakterze jakościowym - nominalne lub porządkowe. Testy nieparametryczne są głównie dedykowane tego rodzaju zmiennym. Ponadto w przypadku tychże testów nie ma konieczności spełnienia założenia o normalności rozkładu. Porównywanymi parametrami są z kolei mediana oraz rozkłady cech w zbiorowościach - liczebności oraz frakcje.

Kolejnym kryterium wyboru właściwego testu statystycznego jest ustalenie zakresu porównywanych rozkładów cech w zbiorowościach. Istotą bowiem wszelkich testów statystycznych jest szukanie istotnych statystycznie różnic między grupami. W tym celu porównujemy statystyki pochodzące z co najmniej dwóch grup nazywanych próbami. Z tego względu wyróżniamy trzy typy testów: dla jednej próby, dla prób niezależnych oraz dla prób zależnych. W przypadku testów dla jednej próby porównujemy cechy rozkładu empirycznego (z próby badawczej) z pewną zakładaną bądź hipotetyczną jej wartością w populacji. Testy dla prób niezależnych charakteryzują się tym, że porównywane grupy jednostek zostały pobrane z różnych, autonomicznych populacji, niezwiązanych ze sobą bezpośrednio (np. grupa kobiet i mężczyzn, zwolennicy poszczególnych ugrupowań politycznych, osoby o określonym wykształceniu).

Jednostki przynależące do różnych grup nie są połączone w pary. Odwrotną sytuację obserwujemy w przypadku testów dla prób zależnych. Jednostki z różnych grup tworzą bowiem pary, co oznacza, iż są one powiązane ze sobą i wzajemnie od siebie zależne.).

Program PSPP udostępnia szereg testów parametrycznych oraz nieparametrycznych. Do testów pierwszego rodzaju zaliczamy test t-Studenta dla jednej próby, dla dwóch prób niezależnych, dla dwóch prób zależnych oraz analizę wariancji ANOVA i MANOVA. Do rodziny testów nieparametrycznych zaliczmy z kolei test chi-kwadrat dla jednej próby Karla Pearsona, test Kołmogorowa-Smirnowa dla jednej próby, test U Mann-Whitney'a, test McNemara, test znaków, test rang Wilcoxona, test H Kruskala-Wallisa, test Friedmana, test W Kendalla oraz test Q Cochrana. W poniższej tabeli prezentujemy założenia oraz kryteria wyboru i zastosowania odpowiedniego testu. Procedurom ich zastosowania i obliczeniom za pomocą programu PSPP zostały poświęcone kolejne rozdziały podręcznika. W tabeli 36 uporządkowano rodzaje testów statystycznych i odpowiadające im poziomy pomiaru zmiennej.

Tabela 36. Testy porównań zmiennych a poziom ich pomiaru

Rodzaj testu	Poziom pomiaru zmiennej testowanej	Rozkład teoretyczny zmiennej testowanej	Rodzaj szacowanych parametrów	Rodzaje porównywanych prób				
				Porównywanie próby z hipotetyczną wartością statystyki w populacji	Próby niezależne		Próby zależne	
					Dwie próby	Trzy i więcej prób	Dwie próby	Trzy i więcej prób
TESTY PARAMETRYCZNE	interwałowy lub ilorazowy	zmienna MA rozkład normalny (spełnienie założenia o normalności rozkładu zmiennej)	średnia arytmetyczna bądź wariancja	test t dla jednej próby	test t dla dwóch prób niezależnych	analiza wariancji ANOVA	test t dla dwóch prób zależnych	analiza wariancji MANOVA
TESTY NIEPARAMETRYCZNE	interwałowy lub ilorazowy	zmienna NIE MA rozkładu normalnego (brak spełnienia założenia o normalności rozkładu zmiennej)	mediana	test Kołmogorowa-Smirnowa dla jednej próby	test U Manna-Whitney'a	test H Kruskala-Wallisa	test znaków, test rang Wilcoxona	test Friedmana,test W Kendalla, test Q Cochrana
	porządkowy	brak konieczności sprawdzenia założeń o normalności rozkładu zmiennej	mediana bądź rozkład frakcji dla poszczególnych wartości zmiennej					
	nominalny		rozkład liczebności bądź frakcji dla poszczególnych wartości zmiennej	test chi-kwadrat			test McNemara	

Źródło: opracowanie własne.

16

Rozdział 16. Badanie różnic między dwiema grupami - testy t-Studenta, test U Manna-Whitney'a, test McNemara, test znaków, test rang Wilcoxona i test chi-kwadrat dla jednej próby

Wielokrotnie w ilościowych badaniach politologicznych zachodzi potrzeba porównywania różnych grup społecznych między sobą. Głównym celem takiej analizy jest stwierdzenie, czy interesujące nas zbiorowości statystyczne są jednakowe pod względem badanej cechy, czy też różnią się. Poruszana w tym miejscu kwestia odnosi się do podstawowego zagadnienia charakterystycznego dla wszelkich analiz ilościowych, których istotą jest szukanie związków przyczynowo-skutkowych. Badanie związków dotyczy powszechnie znanej w statystce logiki zależności zmiennych. Zmienna niezależna, utożsamiana z przyczyną, jest rozumiana jako cecha badanej zbiorowości, która potencjalnie może wywoływać zmiany, różnicować opinie, przedstawiać przeciwstawne względem siebie tendencje lub opozycyjne poglądy. Wartości przyjmowane przez zmienną niezależną mają służyć do wyjaśniania skutków, które obserwowane są w różnicujących się wartościach przyjmowanych przez zmienną zależną, czyli tłumaczone przez nas zjawisko. Poszukiwanie różnic między grupami jest również poszukiwaniem związków i orzekaniem o istnieniu zależności między zmiennymi. W takim ujęciu porównywane grupy będą stanowić zmienną niezależną, czego przykładem mogą być cechy socjodemograficzne takie jak płeć, wiek, wykształcenie, ale również inne właściwości badanej zbiorowości - wybór danej partii politycznej, uczestnictwo w wyborach, preferowany kandydat na urząd prezydenta, poglądy polityczne. Poszczególne grupy wyznaczane są z kolei na podstawie wartości zmiennej niezależnej, gdzie przykładowo dla płci pierwszą grupą będą mężczyźni, a drugą kobiety. Uczestnictwo w wyborach możemy podzielić z kolei na grupę osób uczestniczących wyłącznie w wyborach prezydenckich oraz grupę obywateli biorących udział tylko w wyborach parlamentarnych. Te podziały mogą być różne. Mają one nam jednak pozwolić na porównywanie i stwierdzenie, czy w tych grupach reprezentowane są różne czy takie same poglądy, opinie,

oceny, właściwości, tendencje. Wiąże się to każdorazowo z koniecznością stawiania pytań, na przykład czy kobiety i mężczyźni różnią się w zakresie stopnia akceptacji zasad demokratycznych

lub czy osoby zorientowane prawicowo oceniają tak samo prowadzoną przez rząd politykę socjalną jak osoby deklarujące poglądy lewicowe.

Szereg analiz statystycznych umożliwia udzielenie odpowiedzi na tego rodzaju pytania. W niniejszym rozdziale wyjaśniamy, w jaki sposób porównywać wybrane grupy i orzekać na podstawie klasycznych, szeroko stosowanych testów statystycznych, czy różnią się one istotnie między sobą. Wyposażamy badacza w komplet narzędzi analitycznych, bowiem omawiamy sposoby procedowania ze wszystkimi rodzajami zmiennych: ilościowymi, gdzie wykorzystujemy test t-Studenta stworzony przez Williama S. Gosseta oraz jakościowymi, w których używa się testu U zaproponowanego przez Henry B. Manna i Donalda R. Whitney'a, jak również miary autorstwa Quinna McNemara oraz Franka Wilcoxona, a także odnosimy się do dedykowanego dla zmiennych nominalnych testu chi-kwadrat dla jednej próby opracowanego przez Karla Pearsona.

16.1. Test t-Studenta

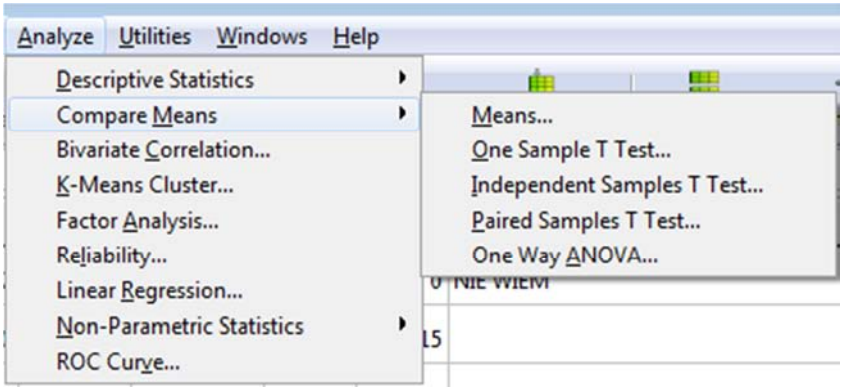
Przypuśćmy, że analizując dochód netto Polaków w 2007 roku, zaczęliśmy zastanawiać się, czy kobiety i mężczyźni różnią się między sobą pod względem jego przeciętnej wielkości. Wstępne wyliczenia przeprowadzone na podstawie danych z bazy PGSW pokazują, że faktycznie - między dochodem kobiet a mężczyzn istnieje różnica. Mężczyźni deklarują średni dochód miesięczny w wysokości 1311 PLN, z kolei kobiety - 949 PLN, czyli o 362 PLN mniej. Pamiętajmy jednak, że sformułowany przez nas wniosek został wywiedziony z przebadania pewnej tylko grupy osób, czyli próby badawczej liczącej n=562 w przypadku mężczyzn i n=667 dla kobiet. W tym miejscu powinniśmy zastanowić się nad kolejnym rodzącym się pytaniem - czy nasze twierdzenie o istnieniu różnicy w dochodach między przebadanymi mężczyznami a kobietami w ogólnej liczbie 1229 jednostek jest również zasadne dla wszystkich pozostałych polskich kobiet i mężczyzn, których w populacji generalnej jest znacznie więcej? Uzasadnienie takiego generalizującego wniosku wymaga zastosowania dodatkowych narzędzi i metod statystycznych, jakim są testy istotności. Jednym z klasycznych testów wykorzystywanych do porównywania grup na podstawie próby badawczej i odnoszenia naszych wyników do szerszej populacji jednostek jest test t-Studenta.

Prawidłowe stosowanie testu t-Studenta wymaga przyswojenia elementarnej wiedzy o charakterystyce tej statystyki. W literaturze statystycznej test t-Studenta jest zaliczany do rodziny testów parametrycznych. Informacja ta, choć być może niezrozumiała w tym momencie, jest pewną wskazówką o zasadach stosowania tego testu. Gdy porównujemy jakiekolwiek grupy między sobą, musimy zawsze posiadać pewien punkt odniesienia, czynnik lub właściwość, w oparciu o który stwierdzamy, że grupy są takie same lub różne. W przypadku testu t-Studenta tą porównywaną cechą jest średnia arytmetyczna. Z tego też względu jest on stosowany dla zmiennych mierzonych na poziomie interwałowym lub ilorazowym, w bardzo rzadkich przypadkach również na poziomie porządkowym. Test t-Studenta umożliwia wyłącznie porównywanie dwóch grup, czego przykładem są grupy wskazane we wstępie niniejszego rozdziału, chociażby mężczyźni i kobiety, uczestniczący i nieuczestniczący w wyborach, osoby o poglądach lewicowych oraz prawicowych. Jeżeli zaistnieje konieczność porównania więcej niż dwóch grup między sobą, wówczas stosuje się inne rozwiązania statystyczne takie, jak test analizy wariancji ANOVA. Należy również pamiętać, iż test t-Studenta służy do szukania różnic między grupami w oparciu o logikę formułowania hipotez statystycznych i ich weryfikację według zasady istotności wyników. Z tego też

względu jest on nazywany testem istotności. Umożliwia bowiem orzekanie o różnicach w zbiorowościach generalnych po przebadaniu jedynie pewnej części ich jednostek składowych.

Wynalezienie testu t-Studenta zawdzięczamy Williamowi S. Gossetowi, który stworzył go pracując w Arthur Guinness Son and Company od 1899 roku. Miara ta pierwotnie przeznaczona była do testowania jakości małych próbek piwa. Opracowana przezeń miara statystyczna została opublikowana w czasopiśmie założonym i prowadzonym przez Karla Pearsona w 1908 roku, gdzie W.S. Gosset podpisał się pseudonimem „Student". Uczynił to jednak nie dlatego, że polityka korporacyjna firmy zakazywała pracownikom publikowania własnych osiągnięć (jak głosi powszechnie znana anegdota), lecz z tego powodu, że browar Guinessa pragnął ukryć fakt istotnej innowacji - posługiwania się w ocenie jakości produktu testami statystycznymi oraz zatrudnianiem statystyków, co wówczas było nowością w sektorze przedsiębiorstw. W wyniku tej publikacji test, o którym mowa w niniejszym podrozdziale, nazywamy testem t-Studenta, a nie testem t-Gosseta[1].

Inspiracją dla W.S. Gosseta do opracowania własnej propozycji testu statystycznego była próba rozwiązania problemu wnioskowania na próbach o małych liczebnościach ($n \leq 30$). Test t-Studenta znalazł zatem zastosowanie przede wszystkim w porównywaniu grup, na które składała się niewielka zbiorowość jednostek. Do weryfikacji hipotez dla prób dużych ($n > 30$) stosuje się z kolei testy oparte na rozkładzie normalnym standaryzowanym. Ze względu jednak na fakt, iż krzywa rozkładu t przy liczebościach przekraczających 30, zaczyna zbliżać się kształtem do rozkładu krzywej normalnej, w testowaniu istotności różnic średniej w programach statystycznych wykorzystuje się test t-Studenta. Tą możliwość w programie PSPP odnajdujemy wybierając z zakładki *Analyze* funkcję *Compare Means,* czyli porównywanie średnich, gdzie w oparciu o statystykę t-Studenta można dokonywać porównywania grup. Ścieżkę wyboru tej opcji prezentuje poniższy zrzut ekranowy:



Test t-Studenta umożliwia przeprowadzenie trzech rodzajów testów, które są widoczne na powyższym rysunku:

1/ testu t-Studenta dla jednej próby (*One Sample T Test*),

2/ testu t-Studenta dla dwóch prób niezależnych (*Independent Samples T Test*),

3/ testu t-Studenta dla dwóch prób zależnych (*Paired Samples T Test*).

[1] Szerzej na ten temat: L. McMullen, *Student as a man*, „Biometrika", 1939, 30, s. 205-210 oraz E.S. Pearson, *Studies in the History of Probability and Statistics. XX: Some Early Correspondence Between W.S. Gosset, R.A. Fisher and K. Pearson With Notes And Comments*, „Biometrika", 1968, 55, s. 445-457.

Jak zasygnalizowaliśmy powyżej, test t-Studenta służy do porównywania średnich. Dokonujemy takiej analizy w dwóch celach. Po pierwsze, aby uzyskać wiedzę o populacji po zbadaniu jej wycinka, czyli próby badawczej, po drugie - aby dowiedzieć się czy dwie grupy, na które podzieliliśmy naszą populację, różnią się między sobą pod względem analizowanej cechy. W zależności od postawionego pytania badawczego, dokonujemy wyboru właściwego rodzaju testu t-Studenta. W dużym stopniu zależy to od formy wyznaczania dwóch porównywanych grup. W przypadku **testu t-Studenta dla jednej próby** będziemy porównywać wynik średniej obliczonej z danych w próbie badawczej z pewną przypuszczalną bądź zakładaną wartością średniej dla tej samej cechy w populacji. Przykładowo zestawiamy średnią wieku z próby, która wynosi 50 lat z założoną przez nas przeciętną wieku w populacji równą 45 lat. **Test t-Studenta dla dwóch prób niezależnych** zastosujemy do porównania średnich po uprzednim logicznym wyróżnieniu niepowiązanych i odmiennych grup jednostek w populacji, np. mężczyzn i kobiet, osób w wieku do 30 lat i osób powyżej 30 roku życia, zwolenników partii X oraz zwolenników partii Y. **Test t-Studenta dla dwóch prób zależnych**, chociaż służy do analizowania tej samej zbiorowości jednostek, jest pomocny w poszukiwaniu zmian, które mogły się dokonać w jej obrębie w perspektywie czasu i pod wpływem oddziaływania czynników zewnętrznych. Wówczas porównujemy wartość średniej danej cechy, przed i po poddaniu jej oddziaływaniu konkretnego bodźca, który mógł wywołać zmiany w stopniu występowania danej cechy, na przykład poparcie dla danej partii politycznej w trakcie i po zakończeniu kampanii wyborczej.

Każdy z tych testów został wyłożony w kolejnych podrozdziałach. Ponadto, omówiono proces testowania hipotez statystycznych, podstawowe pojęcia związane z testem t-Studenta oraz warunki konieczne do jego wykonania.

16.1.1. Procedura testowania hipotez przy pomocy testu t-Studenta

Zaznajomienie się z regułami weryfikacji hipotez statystycznych wymaga omówienia kilku zagadnień. Są one niezbędne dla przeprowadzenia właściwej procedury testowania przypuszczeń o równości średnich w grupach. Kolejne punkty wyznaczają kroki postępowania analitycznego:

1/ przełożenie problemu badawczego na logikę formułowania hipotez statystycznych. Związane jest to z określeniem hipotezy zerowej (H_0) oraz jej przeciwieństwa, czyli hipotezy alternatywnej, nazywanej również hipotezą roboczą (H_1). Hipoteza zerowa zawsze mówi, że średnie w dwóch grupach są równe, z kolei hipoteza alternatywna przeciwnie - że są różne;

2/ sprawdzenie i potwierdzenie warunków uprawniających do przeprowadzenia testu t-Studenta. W tym przypadku musimy zweryfikować takie kryteria, jak poziom pomiaru zmiennych, normalność rozkładu oraz - w przypadku testu dla dwóch prób niezależnych - założenie o równości wariancji w grupach na podstawie testu Levene'a;

3/ określenie poziomu ufności oraz wybór opcji testu t-Studenta (dla jednej próby, dla porównywania dwóch prób niezależnych lub dwóch prób zależnych);

4/ obliczenie testu t-Studenta;

5/ podjęcie decyzji na podstawie wyników testu t-Studenta o odrzucenia lub braku podstaw do odrzuceniu hipotezy zerowej;

6/ prezentacja wyników oraz ich słowna interpretacja w formie komentarza lub wniosków.

Poniżej zamieszczono omówienie tych punktów, które stanowią zbiór praktycznych rad, prowadzących badacza przez proces poszukiwania różnic między porównywanymi grupami.

16.1.1.1. Formułowanie hipotez statystycznych

Stosowanie testu t-Studenta łączy się każdorazowo z koniecznością ujęcia naszego problemu badawczego w postać hipotez statystycznych. Dokonujemy tego za pomocą dwóch rodzajów hipotez: **hipotezy zerowej**, oznaczanej symbolem H_0 oraz jej przeciwieństwa - **hipotezy alternatywnej** (H_1). Hipoteza jest zawsze pewnym oczekiwaniem, przypuszczeniem, sądem bądź wyobrażeniem na temat rzeczywistości społecznej. W określaniu hipotezy zerowej obowiązuje ogólna zasada najprostszego i jednoznacznego formułowaniem takich przypuszczeń. W jej przypadku zakładamy, że dwie grupy są takie same, czyli nie różnią się pod względem badanej cechy. Jak pamiętamy, w przypadku testu t-Studenta porównujemy wartość średniej arytmetycznej w dwóch grupach. Hipoteza zerowa będzie więc mówiła, że wartości średniej w jednej i drugiej grupie są identyczne, np. średnia wieku w grupie kobiet jest równa średniej wieku w grupie mężczyzn. Następnym krokiem jest sprawdzenie prawdziwości naszego przypuszczenia. We wszelkich testach statystycznych dążymy bowiem do potwierdzania bądź odrzucania hipotezy zerowej. Dokonujemy tego zawsze na podstawie zebranych danych ilościowych. Poza hipotezą zerową musimy również sformułować hipotezę alternatywną. Jest ona przeciwieństwem owej hipotezy. Jeżeli hipoteza zerowa mówi, że wartość średniej w pierwszej grupie jest równa wartości średniej w drugiej grupie, to w hipotezie alternatywnej będziemy twierdzić, że te wartości są różne, np. średnia liczba godzin poświęcana miesięcznie na czytanie prasy codziennej jest większa w grupie osób z wykształceniem wyższym niż wśród osób z wykształceniem średnim. W złożonej postaci możemy również sformułować hipotezę, iż wartość średniej w pierwszym przypadku jest większa bądź mniejsza niż w drugim.

W celu lepszego zrozumienia zasad formułowania hipotez statystycznych posłużymy się przykładem. Załóżmy, że interesuje nas średnia płaca uzyskiwana przez mężczyzn i kobiety. Biorąc pod uwagę powszechne opinie, przypuszczamy, że wynagrodzenie kobiet i mężczyzn zajmujących to samo stanowisko jest różne. Sprawdzenie naszej hipotezy badawczej wymaga jej przeformułowania w postać hipotezy statystycznej. W pierwszej kolejności określamy hipotezę zerową, pamiętając, iż zawsze jest ona wyrażeniem równości pewnych wartości. Hipotezy statystyczne zdefiniujemy następująco:

H_0 - średnie wynagrodzenia wśród mężczyzn **JEST RÓWNA** średniej wynagrodzenia wśród kobiet ($m_m = m_k$, gdzie m_m to średnie wynagrodzenie w grupie mężczyzn, natomiast m_k - średnia wynagrodzenia w grupie kobiet). W tym przypadku test **NIE JEST** statystycznie istotny ($p \geq 0,05$);

H_1 - średnie wynagrodzenia wśród mężczyzn **NIE JEST RÓWNA** średniej wynagrodzenia wśród kobiet ($m_m \neq m_k$). Test **JEST** statystycznie istotny ($p < 0,05$).

Sformułowanie hipotez według powyższego schematu jest punktem wyjścia dla dalszego procesu ich weryfikowania, a przede wszystkim - testowania istotności naszych wyników. Celem testu t-Studenta jest bowiem sprawdzenie, czy uzyskane przez nas wyniki z próby możemy przełożyć na szerszą zbiorowość nieobjętą badaniem, czyli populację. Weryfikowaną hipotezą jest zawsze hipoteza zerowa. Na postawie testu istotności dokonujemy jej przyjęcia bądź odrzucenia. W tym miejscu pojawia się kolejne

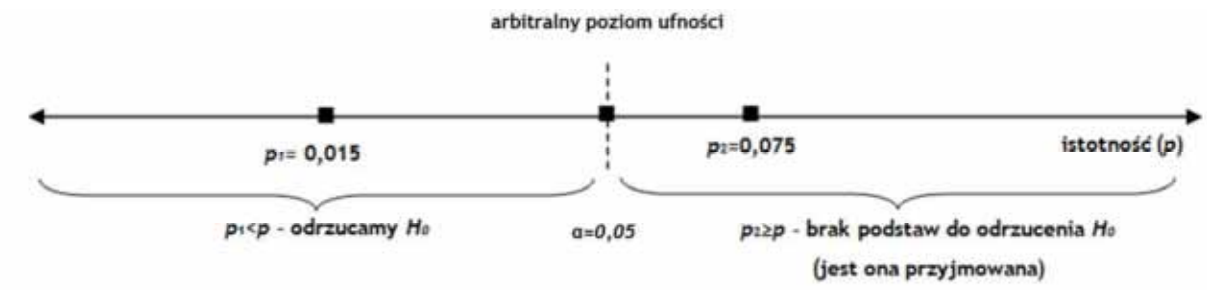
zagadnienie niezbędne do zgłębienia - kwestia poziomu ufności. Jego omówieniu został poświęcony poniższy podrozdział.

16.1.2. Weryfikacja hipotez statystycznych za pomocą testu istotności

W procesie weryfikacji hipotez statystycznych będziemy dążyć do potwierdzenia bądź odrzucenia hipotezy zerowej. Dokonujemy tego w oparciu o prawdopodobieństwo popełnienia tzw. **błędu I rodzaju**, który wyznaczany jest przez **poziom unfości** oznaczany symbolem α. Jego wyznaczenie jest tożsame z określeniem poziomu ryzyka, jaki jesteśmy skłonni przyjąć w naszym wnioskowaniu. Przy weryfikacji hipotez regułą jest stosowanie trzech wartości poziomu ufności - 0,1, 0,05 oraz 0,01. Zwyczajowo bowiem przyjmuje się, że α wynosi 0,05, które wyrażone w procentach wynosi - 5 proc. Oznacza to, że odrzucając hipotezę zerową mamy 5-procentową szansę na popełnienie pomyłki. Mówiąc inaczej, zakładamy, że nasze postępowanie jest słuszne w 95 proc., co wskazuje ustalony przez nas przedział ufności (1-α). Jednakże przyjmujemy przy tym, iż nasze szacunki mogą być błędne, z prawdopodobieństwem popełnienia pomyłki wynoszącym 5 proc. Istnieje zatem 5 proc. szans, że odrzuciliśmy hipotezę, która *de facto* była prawdziwa. Poziom ufności jest zawsze wyznaczany przez nas arbitralnie. Najczęściej przyjmuje się, że α wynosi 0,05. Ustalenie poziomu ufności o wartości wyższej wymaga zaś dodatkowego uzasadnienia.

Wraz z wyznaczaniem poziomu ufności, warto zapoznać się z regułą weryfikowania hipotez. Rozważymy w tym celu dwie sytuacje - pierwszą, odnoszącą się do możliwości przyjęcia hipotezy zerowej oraz drugą - dotyczącą sytuacji jej odrzucenia. Hipoteza zerowa jest przyjmowana bądź odrzucana w zależności od poziomu istotności (p) wyników naszego testu. Przy weryfikacji hipotezy zerowej należy pamiętać, iż zawsze dążymy do jej odrzucenia. Jej odrzucenie może być jednak uzasadnione jedynie w sytuacji osiągnięcia przez nas określonego stopienia pewności, który wyznacza przedział ufności i związany z nim poziom ufności (α). Jak zasygnalizowaliśmy powyżej, minimalny satysfakcjonujący nas rezultat to taki, który pozwoli nam z prawdopodobieństwem wynoszącym 95 proc. stwierdzić, iż faktycznie odrzucamy hipotezę nieprawdziwą. Jeżeli ten procent jest mniejszy i wynosi chociażby 90 proc., wówczas brakuje nam podstaw do odrzucenia hipotezy zerowej, gdyż ryzyko uznania jej za fałszywą jest zbyt duże i wynosi więcej niż zakładane przez nas 5 proc. W celu lepszego zrozumienia tego problemu rozważymy dwie sytuacje, gdzie w pierwszym przypadku nasz test wykazał wartość poziomu istotności p_1=0,015, drugi zaś p_2=0,075. Wyznaczyliśmy jednocześnie arbitralny poziom ufności α=0,05, gdyż uznaliśmy, że 5-procentowy margines błędu jest dla nas satysfakcjonujący i jesteśmy gotowi ponieść takie ryzyko pomyłki. Regułę odrzucania hipotezy zerowej prezentuje wykres 13.

Wykres 13. Reguły odrzucania hipotezy zerowej



W przypadku pierwszego wyniku testu, którego istotność wyniosła p_1=0,015, odrzucamy hipotezę zerową i przyjmujemy alternatywną. Prawdopodobieństwo popełnienia przez nas błędu jest bowiem niższe niż maksymalny przyjęty przez nas margines 5-procentowej szansy popełnienia pomyłki (p_1<0,05). W drugim wariancie ryzykujmy więcej niż 5 proc., dokładnie 7,5 proc. (p_2=0,075), iż nie przyjmiemy hipotezy prawdziwej, zatem nie mamy podstaw do jej odrzucenia i uznajemy ją za faktyczną.

Po poznaniu reguł formułowania hipotezy zerowej i alternatywnej, wyznaczeniu poziomu ufności (α) oraz zrozumieniu logiki weryfikacji hipotezy zerowej, możemy przystąpić do kolejnego etapu poszukiwania różnić między grupami przy pomocy testu t-Studenta. Wymaga to poznania założeń tego testu oraz sposobów ich sprawdzania i potwierdzania, co zastało omówione w kolejnym podrozdziale.

16.1.3. Założenia testu t-Studenta

Przeprowadzenie testu t-Studenta wymaga spełnienia kilka warunków niezbędnych do dalszego procedowania. W zależności od sformułowanego pytania badawczego, dokonujemy wyboru zmiennych oraz przeprowadzamy ocenę ich właściwości. Spełnienie określonych kryteriów, w szczególności w zakresie jej rozkładów, jest warunkiem, który uprawnia do użycia testu t-Studenta. Poniżej zostały omówione poszczególne kroki wraz z praktycznymi wskazówkami umożliwiającymi ich przeprowadzenie w programie PSPP. Sprawdzenia założeń dokonujemy przed przystąpieniem do obliczania statystyk t-Studenta.

16.1.3.1. Poziom pomiaru zmiennej

Test t-Studenta jest przeznaczony do poszukiwania różnic między grupami na postawie wartości średniej arytmetycznej. Z tego względu umożliwia on jedynie porównywanie rozkładów zmiennych mierzonych na poziomie interwałowym bądź wyższym - ilorazowym. Niekiedy badacze stosują test t-Studenta do określania różnic między zmiennymi mierzonymi na poziomie porządkowym. Przykładem tego zastosowania jest powszechnie wykorzystywana w badaniach sondażowych pięciostopniowa skala oceny wzorowana na klasycznej skali Rensisa A. Likerta. Jednakże testowanie zmiennych mierzonych na poziomie porządkowym przy pomocy rozkładów t-Studenta nie jest powszechnie praktykowane i jest uprawnione jedynie w wyjątkowych przypadkach. Do tego rodzaju porównań zaleca się stosowanie testów nieparametrycznych: testu Wilcoxona (dla dwóch prób zależnych), testu U Manna-Whitney'a (dla dwóch prób niezależnych), testu Friedmana (dla więcej niż dwóch prób zależnych) lub testu Kruskala-Wallisa (dla więcej niż dwóch prób niezależnych).

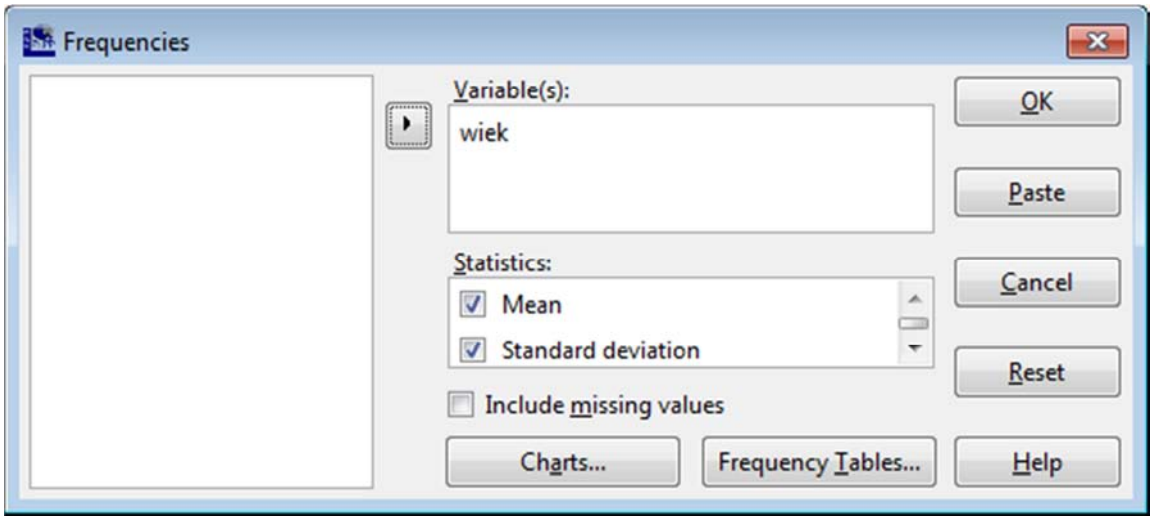
16.1.3.2. Założenie o normalności rozkładu zmiennych

Kolejnym etapem oceny wymagań stawianych przez test t-Studenta jest sprawdzenie, czy został spełniony warunek o normalności rozkładu zmiennej. Przyjmuje się, że krzywa rozkładu t-Studenta jest zbieżna z krzywą rozkładu normalnego w przypadku prób o dużych liczebnościach, wynoszących co najmniej n>30. Dla małych prób o liczebnościach n≤30 przed skorzystanie z testu t-Studenta wymagane jest sprawdzenie i potwierdzenie założeń o normalności rozkładu zmiennej.

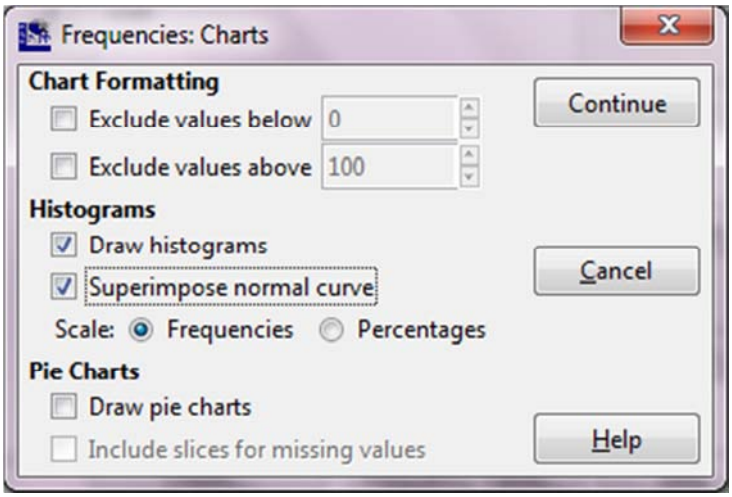
Weryfikacji tego warunku dokonujemy na podstawie wizualnej oceny rozkładu zmiennej przez analizę histogramu z nałożoną krzywą normalną lub przeprowadzenie testu normalności. Dodatkowo analizuje

się wyniki takich miar asymetrii rozkładu, jak kurtozy oraz współczynnika skośności. Jeżeli wyniki analiz wstępnych potwierdzą normalność rozkładu zmiennej, wówczas możliwe jest kontynuowanie analiz. W sytuacji, gdy założenie o normalności rozkładu zmiennej nie zostało spełnione, testowania hipotez dokonujemy w oparciu o testy nieparametryczne.

Wizualną ocenę rozkładu zmiennej dokonujemy za pomocą **analizy histogramu z nałożoną krzywą normalną**. Operację tą możemy wykonać przy pomocy programu PSPP. W tym celu należy wybrać opcję *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies*. W polu *Variable(s)* umieszczamy testowaną zmienną (dla przykładu wybraliśmy zmienną wiek).

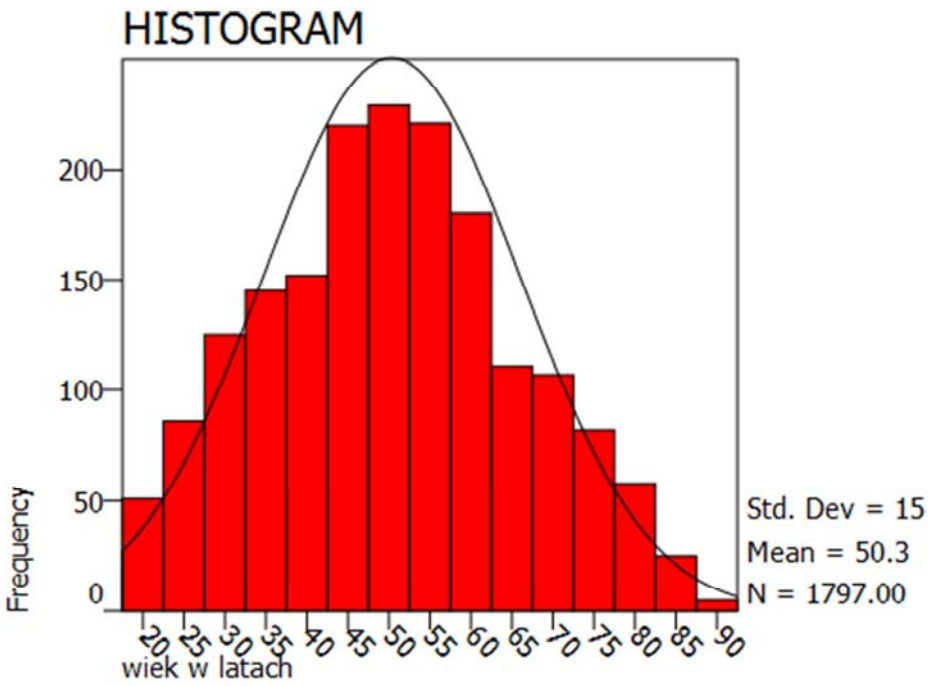


Następnie należy wybrać przycisk *Charts* umiejscowiony w dolnej części okna:



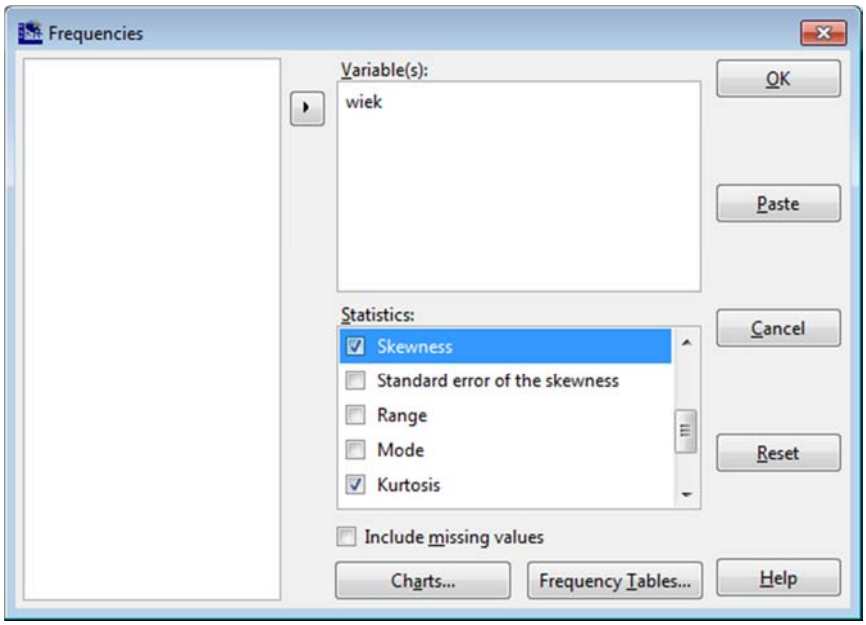
Należy zaznaczyć (jak na rysunku) powyżej opcję wykreślenia histogramu (*Draw histograms*) oraz opcję nałożenia krzywej normalnej (*Superimpose normal curve*). Efektem tej operacji jest wykres, który nazywamy histogramem.

Został on zaprezentowany na poniższym zrzucie ekranowym:



Na jego podstawie sprawdzamy, czy zmienna ma postać rozkładu normalnego. Dokonujemy tego poprzez analizę wykreślonej krzywej i ocenę jej kształtu. W naszym przypadku wykres wygląda obiecująco, bowiem przybiera kształt zgodny z krzywą normalną. Należy zaznaczyć, że opieranie się wyłącznie na ocenie wizualnej jest mylące i nie może stać się jedyną podstawą wnioskowania. Musimy przystąpić zatem do dalszych czynności, czyli obliczenia kurtozy oraz skośności, a także testu Kołmogorowa-Smirnowa.

Obliczenie kurtozy oraz skośności w programie PSPP wymaga wybrania powtórnie opcji *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies.* Następnie wymieramy z listy zmiennych znajdującej się po lewej stronie zmienną wiek, którą umieszczamy w polu *Variable(s)* oraz zaznaczmy w opcjach *Statistics* skośność (*Skewness*) oraz kurtozę (*Kurtosis*). Wykonanie tej czynności prezentuje poniższy zrzut ekranowy:



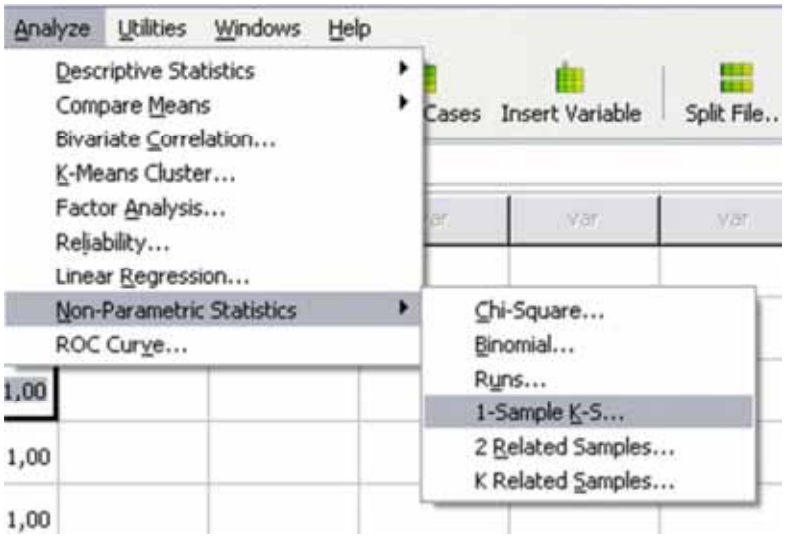
Efektem tej procedury jest poniższa tabela zawierająca obliczoną wartość kurtozy oraz skośności:

wiek w latach

N	*Valid*	1797
	Missing	0
Kurtosis		-.54
Skewness		.11

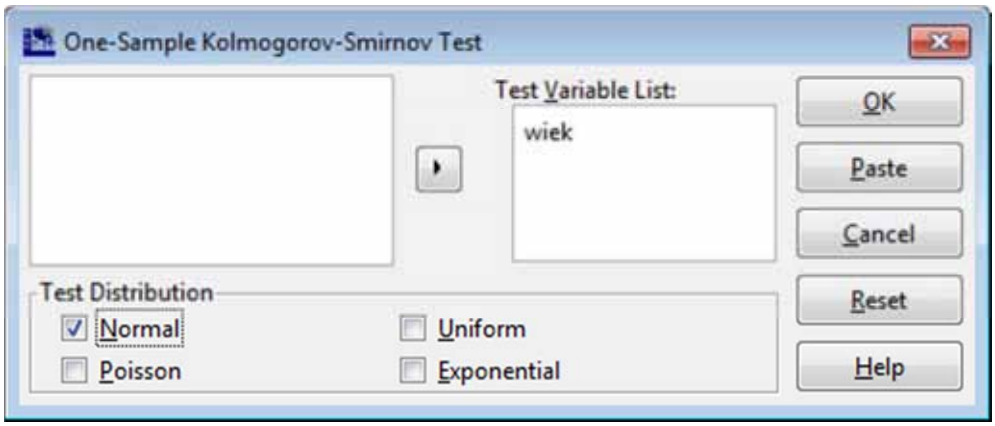
W przypadku analizowanej zmiennej wiek kurtoza wynosi -0,54. Znak ujemny oznacza spłaszczenie rozkładu w porównaniu z rozkładem normalnym, a więc rozkład jest platykurtyczny. Jednak ze względu na fakt, że wartość kurtozy zawiera się pomiędzy -1 a +1, nie odrzucamy twierdzenia o normalności rozkładu. Z kolei skośność, która jest miarą asymetrii rozkładu, wyniosła 0,11. Wartość tej statystyki świadczy o rozkładzie normalnym, jeśli wynosi dokładnie 0 lub o rozkładzie zbliżonym do normalnego, jeśli zawiera się w zakresie od -1 do +1. Wartość ujemna wskazuje na asymetrię lewostronną rozkładu, a wartość dodatnia - na asymetrię prawostronną. Rozkład wydłużony w prawą stronę to rozkład prawostronny, a rozkład posiadający tzw. „ogon" z lewej - lewostronny. Wartość skośności dla zmiennej wiek mieści się w tym przedziale, zatem jej rozkład jest zbliżony do normalnego. Zarówno miara kurtozy jak i skośności nie przekracza jedności, nie mamy więc podstaw do odrzucenia twierdzenia o normalności rozkładu.

Kolejnym (opcjonalnym) etapem jest sprawdzenie normalności rozkładu za pomocą testu Kołmogorowa-Smirnowa. Powinien być on wykonywany z uwzględnieniem tak zwanej poprawki Lilleforsa, która jest obliczana, gdy nie znamy średniej lub odchylenia standardowego dla całej populacji. Test Kołmogorowa-Smirnowa można wykonać w PSPP, wybierając w zakładce *Analyze* ⇨ *Non-Parametric Statistics* ⇨ *1-Sample K-S.*



Zmienną do testowania jest zmienna wiek, którą wybieramy z listy i przenosimy do pola o nazwie *Test Variable List.* W *Test Distribution* zaznaczamy, że chcemy porównywać tą zmienną z rozkładem normalnym (zaznaczamy zatem *Normal*).

Wykonanie tej operacji prezentuje następujący zrzut ekranowy:



Interpretacja testu Kołmogorowa-Smirnowa wymaga uwzględnienia dwóch wartości - statystyki Z oraz poziomu istotności (p). Test Kołmogorowa-Smirnowa opiera się na następującej hipotezie zerowej (H_0) i alternatywnej (H_1):

H_0 - rozkład badanej cechy w populacji jest rozkładem normalnym;

H_1 - rozkład badanej cechy w populacji jest różny od rozkładu normalnego.

Jeśli istotność dla tego testu jest niższa niż założony poziom α (domyślnie α=0,05) wówczas przyjmujemy H_1. Jeśli jest wyższa lub równa α=0,05, nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że rozkład jest normalny.

W tabeli, która jest efektem wykonania testu, dwie interpretowane wartości znajdują się w ostatnich dwóch wierszach tabeli. Wartość p jest wyższa niż 0,05, a zatem nie mamy podstaw do odrzucenia H_0. Zakładamy zatem, że rozkład jest zbliżony do normalnego. Im większa wartość Z, tym większe odchylenie danych empirycznych od teoretycznego rozkładu (w niniejszym przypadku rozkładu normalnego). W analizowanym przypadku wynosi ona 1,27.

One-Sample Kolmogorov-Smirnov Test

		wiek w latach
N		1797
Normal Parameters	*Mean*	50.33
	Std. Deviation	15.61
Most Extreme Differences	*Absolute*	.03
	Positive	.03
	Negative	-.03
Kolmogorov-Smirnov Z		1.27
Asymp. Sig. (2-tailed)		.06

16.1.3.3. Założenie o jednorodności wariancji

Kolejnym krokiem w sprawdzaniu założeń testu t-Studenta jest ocena homogeniczności wariancji w porównywanych grupach. Jest to istotne w przypadku porównywania dwóch grup niezależnych. Sprawdzenie tego warunku możliwe jest dzięki obliczeniu testu Levene'a jednorodności wariancji (*Levene's Test for Equality of Variances)*. Opiera się on na wartości statystyki F oraz wartości obliczonego poziomu istotności (p). Dla przypomnienia należy dodać, iż wartość p określa prawdopodobieństwo

popełnienia błędu I rodzaju, na podstawie którego weryfikuje się hipotezy statystyczne. Wartość tego testu umożliwia przyjęcie bądź odrzucenie hipotezy zerowej, mówiącej o równości wariancji w obu porównywanych grupach, i przełożenia wyników na badaną populację. W tym przypadku rozważa się dwie hipotezy:

H_0 - wariancje w pierwszej i drugiej porównywanej grupie **SĄ RÓWNE**. Test **NIE JEST** statystycznie istotny ($p \geq 0,05$);

H_1 - wariancja w pierwszej grupie **NIE JEST RÓWNA** wariancji w drugiej grupie. Test **JEST** statystycznie istotny ($p < 0,05$).

Wyniki uzyskane na postawie testu Levene'a wpływają na wybór właściwej statystyki t-Studenta. Wartości t-Studenta w programie PSPP obliczane są bowiem dla dwóch sytuacji: pierwszej - gdzie przyjmujemy założenie o równości wariancji w porównywanych grupach (w programie PSPP jest to określone terminem *Equal variances assumed*), oraz drugiej - gdzie nie zakładamy równości wariancji (*Equal variances not assumed*). W programie PSPP test Levene'a jest obliczany automatycznie wraz z wykonywaniem szacunków dla wybranej przez nas opcji testu t-Studenta. Wyniki obu testów pojawią się w tej samej tabeli, jednak testowi Levene'a odpowiadają dwie pierwsze kolumny umieszczone przed wartościami właściwymi statystyce t. Zawierają one dwie statystyki: wartość *F* oraz poziom istotności oznaczamy zawsze skrótem *Sig.,* który stanowi rozwinięcie anglojęzycznego terminu *Significance* (odpowiada mu poszukiwana wartość p). Na poniższym rysunku zamieszczono przykładowy wynik testu Levene'a, gdzie przy pomocy testu t dla dwóch prób niezależnych porównywano wiek wśród kobiet i mężczyzn. Przyjęto poziom ufności (α) wynoszący 0,05.

Independent Samples Test

	Levene's Test for Equality of Variances		
	F	*Sig.*	*t*
WIEKEqual variances assumed	10.44	.00	3.32
Equal variances not assumed			3.18

Wynik testu Levene'a pokazuje, iż $p < 0,05$. Oznacza to, że wariancje w porównywanych grupach różnią się istotnie. Nie przyjmujemy zatem założenia o równości wariancji, więc wyniki testu t-Studenta będziemy odczytywać dla drugiego wiersza tabeli (*Equal variances not assumed*). Jeżeli test Levene'a wykazałby, że wariancje w obu grupach są równe, czyli poziom istotności byłby większy bądź równy 0,05 ($p \geq 0,05$), wówczas wartości statystyki t odczytywane byłyby dla pierwszego wiersza (*Equal variances assumed*).

Na marginesie można wspomnieć, iż poza testem Levene'a, homogeniczność wariancji może być testowana przy pomocy innych rozwiązań takich, jak test Bartletta, test Cadwella, test Hartleya oraz test Coxa. Te statystyki nie są jednak dostępne w pakiecie PSPP. Wystarczające jest sprawdzanie założeń o jednorodności wariancji za pomocą testu Levene'a.

Sprawdzenie i potwierdzenie powyżej opisanych założeń testu t-Studenta, umożliwia przystąpienie do właściwej części jego obliczania i testowania hipotez przy jego pomocy. Wpierw konieczne jest jednak zapoznanie się z rodzajami testu t-Studenta i wybraniem adekwatnej opcji do naszych celów badawczych. Zostały one omówione w poniższym podrozdziale.

16.1.4. Zastosowanie testu t-Studenta

W programie PSPP możemy odnaleźć trzy sposoby porównywania grup w oparciu o statystykę t-Studenta: test t dla jednej próby, test t dla dwóch prób niezależnych oraz test t dla dwóch prób zależnych. Wybór właściwego rozwiązania zależy od przyjętych przez nas celów. Poniżej omówiono zastosowania każdego z testów.

16.1.4.1. Test t-Studenta dla jednej próby

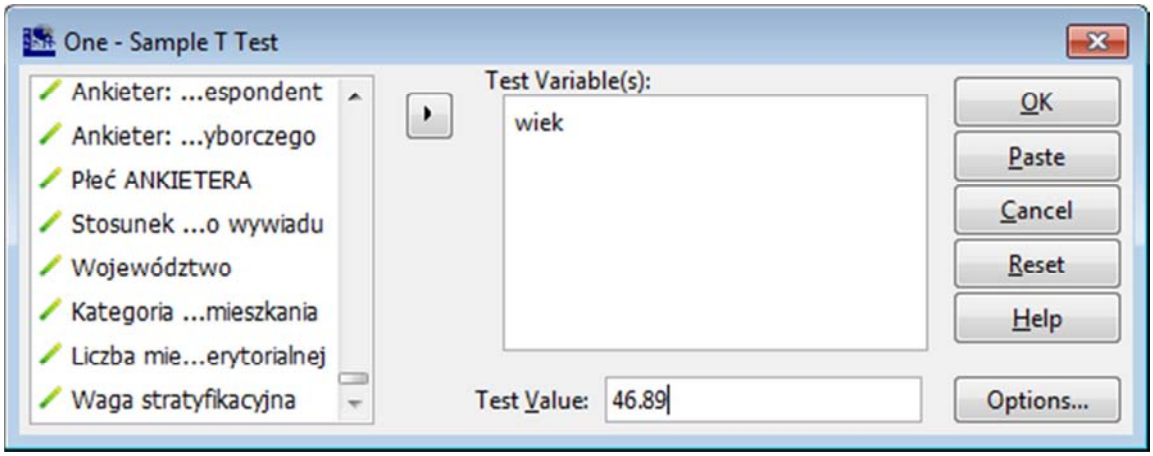
Na wstępie przypomnijmy, że test t-Studenta opiera się na tescie istotności. Służy on zatem do weryfikacji przypuszczeń o wartości średniej w populacji na postawie wyników uzyskanych po przebadaniu próby losowej. Celem testowania istotności jest możliwość sprawdzania i przenoszenia owych przypuszczeń na szerszą zbiorowość. Weryfikacja hipotez statystycznych stwarza nam zatem możliwość orzekania o interesującej nas populacji.

Jednym z testów t-Studenta, który pozwala nam na wnioskowanie na temat zbiorowości generalnej jest **test dla jednej próby** (*One Sample Test*), nazywany również **testem dla jednej średniej**. Jak pamiętamy, test t-Studenta umożliwia poszukiwania różnic między grupami w oparciu o porównywanie średnich (m). Test dla jednej próby służy do weryfikacji hipotezy zerowej o równości średniej wyliczonej z próby losowej z wartością średniej w populacji (m_0). Z reguły jest jednak tak, iż nie wiemy jaka jest wartość średniej w zbiorowości generalnej. Dlatego wartość, do której będziemy porównywać wyniki uzyskane z pojedynczej zmiennej losowej, ustalamy samodzielnie. Może to być wartość abstrakcyjna. Dobrą praktyką jest jednak testowanie parametrów o prawdopodobnych wartościach, o których wiedzę czerpiemy z innych źródeł, chociażby z wiedzy potocznej lub z innych badań, zestawień statystycznych, czego przykładem jest *Rocznik Statystyczny.* Przystępując do obliczania testu dla jednej próby za pomocą programu PSPP, posłużymy się przykładem. Załóżmy, że chcemy sprawdzić, czy średnia wieku w naszej próbie badawczej różni się od średniej wieku w populacji Polaków powyżej 18 roku życia. Do tego celu wykorzystamy dane PGSW z 2007 roku. W przypadku testu dla jednej próby, hipotezę zerową oraz alternatywną sformułujemy następująco:

H_0 - **NIE ISTNIEJE** istotna statystycznie **RÓŻNICA** między przyjętą średnią wieku w populacji dorosłych Polaków a średnią wieku w próbie badawczej ($m=m_0$, gdzie m to średnia wieku w próbie badawczej, a m_0 to średnia w populacji). Test **NIE JEST** statystycznie istotny ($p\geq0,05$);

H_1 - **ISTNIEJE** statystycznie istotna **RÓŻNICA** między przyjętą średnią wieku w populacji dorosłych Polaków a średnią wieku w próbie badawczej ($m\neq m_0$). Test **JEST** statystycznie istotny ($p<0,05$).

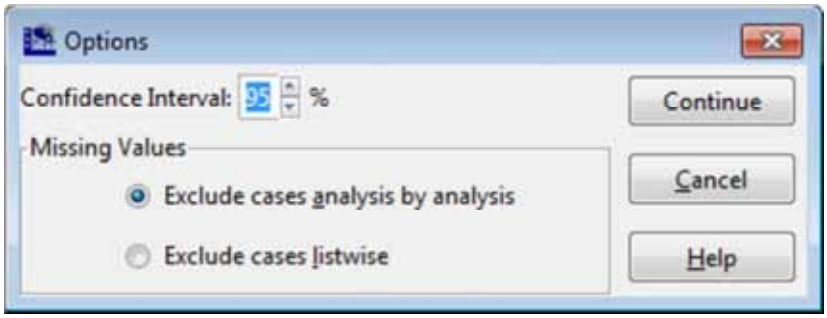
Zakładamy, że średnia w populacji Polaków powyżej 18 roku życia wynosiła 46,89 lat. Przystępujemy do testowania naszej zmiennej w programie PSPP. W tym celu wybieramy z zakładki *Analyze* ⇨ *Compare Means* ⇨ *One Sample T test.* Następnie pojawi się okno, jak na poniższym zrzucie ekranowym:



Przystępując do testowania zmiennej, wykonujemy następujące czynności. Odnajdujemy na liście zmiennych w lewej części okna interesującą nas zmienną wiek, która jest jednocześnie zmienną testowaną. Zmienna ta została wcześniej przygotowana przy pomocy funkcji *compute* i obliczona ze zmiennej m1 (rok urodzenia). Za pomocą strzałki przenosimy ją do okna *Test Variable(s).* Przyjmowaną przez tą zmienną wartość będziemy porównywać z ustaloną wcześniej wartością średniej w populacji (m_0=46,89).

Następnie w polu *Test Value* wprowadzamy hipotetyczną wartość średniej wieku w populacji dorosłych Polaków, którą ustaliliśmy na etapie formułowania hipotez statystycznych. Drogą przypomnienia - uznaliśmy, iż będzie ona wynosić m_0=46,89 lat. Program PSPP akceptuje jedynie wartości dziesiętne rozdzielone znakiem kropki (.), nie zaś przecinkiem (,). Należy o tym pamiętać podczas wpisywania wartości testowanej, gdyż w innym przypadku nie będziemy mogli uruchomić procesu obliczeniowego (krótko mówiąc - nie będziemy mogli kliknąć 'OK'). Należy zwrócić uwagę, iż program PSPP umożliwia wprowadzenie więcej niż jednej zmiennej testowanej.

W kolejnym etapie ustalamy poziom ufności. Dokonujemy tego za pomocą ikony *Options,* po czym pojawia się poniższe okno:



Pole *Confidence Interval* wymaga wprowadzenia przyjętego przez nas przedziału ufności. Wcześniej ustaliliśmy, że akceptowany przez nas poziom ufności wynosić będzie α=0,05. Przedział ufności (1-α) wynosi zatem 0,95, co w wartościach procentowych równe jest 95-procentom. Program PSPP ustala przedział ufności na poziomie 95 proc. jako domyślny. Możemy go dowolnie zwiększać bądź zmniejszać, w zależności od naszych preferencji. Warto pamiętać, że maksymalny poziom ufności, który możemy wprowadzić w tym oknie wynosi 99 proc. Testowanie zmiennych przy wyższym poziomach wynoszących np. 0,001 wymaga wykorzystania polecenia składni.

Zwróćmy jednocześnie uwagę na pole *Missing Values,* które odnosi się do wyłączania braków danych w procesie testowania zmiennych. W jego ramach pojawiają się następujące opcje:

1/ ***Exclude cases analysis by analysis - wyłączanie obserwacji analiza po analizie.*** W tym przypadku braki danych wyłączone są jedynie w analizie z udziałem tej zmiennej. Jest to opcja ustawiona jako domyślna. Wyklucza się wówczas brakujące wartości dla wybranej zmiennej wybranej;

2/ ***Exclude cases listwise - wyłączanie wszystkich obserwacji z brakami.*** W tym przypadku jednostki analizy z brakami danych będą wyłączane dla każdej zmiennej, gdzie są zdefiniowane jako braki danych (*Missing*) lub które stanowią tzw. systemowe braki danych. Wówczas usuwa się wartości brakujące dla każdej zmiennej uczestniczącej w analizie.

Tradycyjnie pozostawia się opcję zaznaczoną jako domyślną, czyli *Exclude cases analysis by analysis.* Po wybraniu zmiennej testowanej (bądź zmiennych testowanych), wprowadzeniu wartości testowanej oraz określeniu przedziału ufności, klikamy 'OK'. Korzystanie z edytora wymaga natomiast wprowadzenia następującej składni:

Składnia do wpisania w Edytorze	Opis działania składni
`T-TEST`	- wykonaj test t-Studenta dla jednej próby
`/TESTVAL = 46.89`	- jako wartość średniej testowanej przyjmij 46,89 lat
`/VARIABLES= wiek`	- porównaj wartość tej średniej ze średnią dla zmiennej *wiek*
`/MISSING=ANALYSIS`	- dla braków danych wybierz opcję *wyłączanie obserwacji analiza po analizie*
`/CRITERIA=CIN(0.95).`	- przyjmij przedział ufności 0,95, czyli poziom ufności wynosić będzie 0,05.

W oknie raportu pojawiają się następujące wyniki przeprowadzonego testu w postaci poniższych dwóch tabel:

One-Sample Statistics

	N	Mean	Std. Deviation	S.E. Mean
WIEK	1797	54.28	16.70	.39

One-Sample Test

	Test Value = 46.890000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
WIEK	18.76	1796	.00	7.39	6.62	8.16

Pierwsza tabela, zatytułowana *One-Sample Statistics,* zawiera podstawowe statystyki dla testowanej zmiennej wiek. Program PSPP dodatkowo wylicza średnią (*Mean*), odchylenie standardowe dla średniej (*Std. Deviation*) oraz błąd standardowy średniej (*S.E. Mean*). W przypadku weryfikacji hipotez przy pomocy testu dla jednej próby interesują nas wyniki zamieszczone w drugiej tabeli, zatytułowanej *One-Sample Test.* W wierszu wyznaczanym przez nazwę zmiennej wiek, zamieszczane są statystyki dla testu t-Studenta. Program PSPP wylicza wartość statystki t, poziom istotności, a dokładnie istotność asymptotyczną dla obustronnego obszaru krytycznego (*Sig. 2-tailed*), różnicę między średnią z próby i średnią zmiennej testowanej (*Mean Difference*), a także wartość nowej dla nas wielkości - liczbę stopni swobody. Warto w tym miejscu zaznaczyć, iż **liczba stopni swobody** oznaczana jest skrótem *df,* który pochodzi od anglojęzycznego terminu *Degree of Freedom.* Jest to parametr rozkładu zmiennej losowej,

który w najprostszej postaci oznacza liczbę możliwych zmiennych wartości, jakie może przyjąć zbiór wchodzących do próby jednostek analizy. W przypadku testu istotności dla jednej próby w rozkładzie t-Studenta liczba stopni swobody wyznaczana jest według wzoru:

$$df=n-1$$

Wyliczenie stopni swobody dla tego testu wymaga odjęcia wartości 1 od całkowitej liczebności próby badawczej (n). W teście t-Studenta liczba stopni swobody wraz z poziomem ufności (α) są wartościami podstawowymi i służą do odszukania w tablicach rozkładu t-Studenta wartości statystyki t. Pozwala to na określenie obszaru krytycznego, który jest niezbędny w toku weryfikacji hipotez statystycznych. W praktyce jednak program statystyczny automatycznie wylicza wszelkie potrzebne miary. Na tym etapie konieczne jest natomiast nabycie właściwej umiejętności odczytywania wyników oraz ich interpretacji.

Z tabeli *One-Sample Test* odczytujemy poziom istotności (*Sig.*)dla zmiennej testowanej wiek. Jest on mniejszy niż przyjęty przez nas *a priori* poziom ufności, gdyż $p<0,05$. Na podstawie przeprowadzonego testu dla jednej próby odrzucamy hipotezę zerową mówiącą o równości średnich. Słowna interpretacja wyników powinna przyjąć następującą postać: test t-Studenta dla jednej próby wykazał, że średnia wieku wśród respondentów PGSW różni się od przypuszczalnej średniej wieku dorosłych Polaków wynoszącej 46,89 lat, gdzie $t(1796)=18,76$ przy poziomie istotności $p<0,05$ (dla wartości t w nawiasach podajemy liczbę stopni swobody z tabeli wynikowej). Dodatkowo dodatnia wartość statystyki t pokazuje, iż wartość średniej wśród uczestników badania PGSW jest wyższa niż zakładana w populacji. Analogicznie, jeżeli wartość t byłaby ujemna, wówczas średnia wieku byłaby niższa od wartości testowanej.

16.1.4.2. Test t-Studenta dla dwóch prób niezależnych

Kolejnym testem dostępnym w pakiecie PSPP opierającym się na statystyce t-Studenta jest test dla dwóch prób niezależnych. Wykorzystywany jest on w celu porównywania dwóch grup nazywanych próbami. Charakterystyczną cechą analizowanych grup jest ich niezależność. Dobór jednostek analizy do jednej próby nie jest związany z doborem jednostek do drugiej próby i odbywa się niezależnie od pierwszej. Próby takie są tworzone na podstawie dzielenia całkowitej próby losowej uzyskanej w toku badania. Porównywane grupy można utożsamić z tzw. podpróbami, które składają się na pełną próbę badawczą. Jednocześnie pamiętajmy, że test t-Studenta umożliwia porównywanie wyłącznie dwóch grup, dlatego wynikiem dokonywanych przez nas kategoryzacji musi być zawsze wskazanie dwóch podprób. Tego rodzaju podziały tworzymy na podstawie **zmiennej grupującej**. Każda grupa wyróżniana jest na podstawie wartości przyjmowanych przez tą zmienną. Posługując się bazą danych PGSW, przytoczymy kilka przykładów jego zastosowania.

Najprostszy podział możemy stworzyć na podstawie zmiennej płeć (m2). Na jej podstawie z łatwością można wyróżnić dwie grupy. Pierwsza zostanie utworzona na podstawie wartości 1 i obejmować będzie mężczyzn, drugą natomiast wyróżnimy na podstawie wartości 2, która została przypisana kobietom. Innym przykładem może być zmienna posiadająca więcej wartości, jak zmienna c28 o treści *Na kandydata którego komitetu wyborczego głosował Pan(i) w wyborach do Sejmu 21.10.2007?* W tej zmiennej można wyróżnić dziesięć komitetów wyborczych, które mógł wskazać respondent, co daje możliwość stworzenia podziału próby badawczej na dziesięć grup. Pamiętając jednak, że w teście t-Studenta możemy porównywać wyłącznie dwie grupy, musimy dokonać ich wyboru. Załóżmy, że chcemy porównać osoby, które oddały swój głos na Platformę Obywatelską oraz na Prawo i Sprawiedliwość. Pierwszą grupę będą

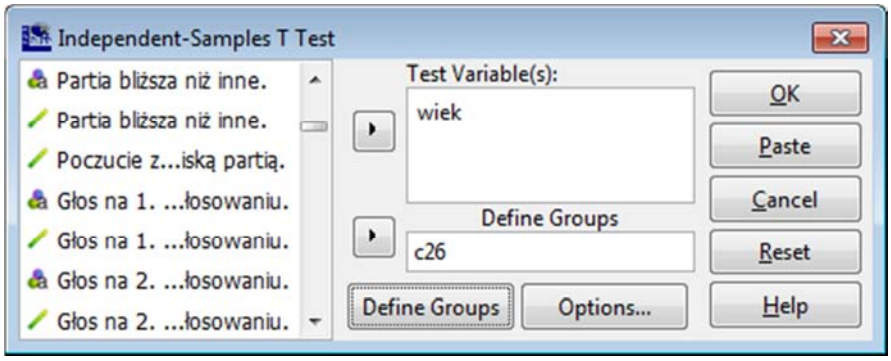
tworzyły jednostki analizy z wartością 4 (etykieta tej wartości to bowiem komitet wyborczy PO) w zmiennej c28, drugą grupę z kolei jednostki z przypisaną wartością 3 (czyli wybierający komitet wyborczy PiS). Tworzenie tego rodzaju podziałów może przyjąć prostą formę i wynikać z klasyfikacji wyznaczanej przez pojedyncze wartości zmiennej lub być stworzone samodzielnie, na drodze przekształcania zmiennych poprzez rekodowanie. Musi być ono jednak ściśle powiązane z postawionym problemem badawczym.

Procedurę obliczeniową należy rozpocząć od właściwego sformułowania hipotezy zerowej oraz hipotezy alternatywnej. Przypuśćmy, że chcemy porównać średnią wieku w grupie osób uczestniczących w wyborach parlamentarnych w 2007 roku oraz wśród osób, które nie brały w nich udziału. Zostały one utworzone na podstawie wartości zmiennej c26, gdzie 1 to *tak, głosowałem(am),* zaś 2 - *nie, nie głosowałem(am).* W oparciu o zasady opisane w podrozdziale 16.1.2. określamy hipotezę zerową mówiącą o równości średnich w obu grupach oraz hipotezę alternatywną o braku równości średnich w tychże grupach. Zapis w ten sposób sformułowanych hipotez będzie następujący:

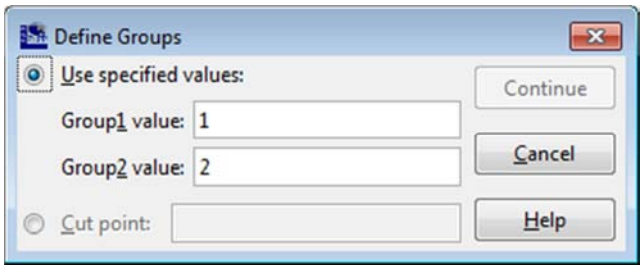
H_0 - średnia wieku w grupie osób głosujących w wyborach i średnia wieku w grupie osób niegłosujących w wyborach **SĄ RÓWNE** ($m_g = m_{ng}$, gdzie m_g to średnia wieku w próbie osób głosujących, a m_{ng} to średnia wieku w grupie osób niegłosujących). Test **NIE JEST** statystycznie istotny ($p \geq 0{,}05$);

H_1 - średnia wieku w grupie osób głosujących w wyborach i średnia wieku w grupie osób niegłosujących w wyborach **NIE SĄ RÓWNE** ($m_g \neq m_{ng}$). Test **JEST** statystycznie istotny ($p < 0{,}05$).

Przystępując do weryfikacji powyższej hipotezy w programie PSPP, należy wybrać *Analyze* ⇨ *Compare Means* ⇨ *Independent Sample T test.* Po wybraniu tej opcji pojawia się następujące okno:



W teście dla dwóch prób niezależnych wybieramy z listy zmiennych znajdującej się po lewej stronie zmienną testowaną. Wprowadzamy ją do pola o nazwie *Test Variable(s).* W naszym przypadku będzie to wcześniej utworzona zmienna wiek. Wybieramy następnie zmienną grupującą, na podstawie której wyznaczać będziemy porównywane grupy głosujących i niegłosujących w wyborach parlamentarnych w 2007 roku. Umieszczamy ją w polu *Define Groups.* Kolejnym krokiem jest wskazanie, po jakich wartościach zmiennej c26, na podstawie których wyróżnimy dwie próby. W tym celu skorzystamy z funkcji *Define Groups,* po wybraniu której pojawia się poniższe okno:



Istnieją dwa sposoby wyróżniania porównywanych prób. Pierwszy polega na wskazaniu wartości zmiennej grupującej dla pierwszej oraz dla drugiej grupy. Drugi sposób polega na wskazaniu punktu przecięcia (*Cut point*), czyli miejsca podziału. Jeżeli chcielibyśmy porównywać osoby do 50 roku życia i osoby powyżej 50 lat, wówczas punkt przecięcia określilibyśmy jako 50 w przypadku zmiennej wiek mierzonej na poziomie ilościowym. Dla naszego przykładu wykorzystamy opcję pierwszą, wpisując w polu *Group1 value* wartość 1 (w przypadku zmiennej c26 są to osoby głosujące w wyborach), natomiast w polu *Group2 value* wartość 2 (czyli osoby niegłosujące w wyborach). Kolejnym krokiem jest ustalenie przedziału ufności na standardowym poziomie 95 proc., który wprowadzamy po wprowadzeniu funkcji *Options.* Czynimy to w ten sam sposób, jak w przypadku testu dla jednej próby. Opcjonalnie, możemy wykonać tą operację za pomocą edytora składni:

Składnia do wpisania w Edytorze	Opis działania składni
`T-TEST`	- wykonaj test t-Studenta dla dwóch prób niezależnych
`/VARIABLES= wiek`	- porównaj rozkład średniej dla zmiennej wiek
`/GROUPS=c26 (1,2)`	- w grupach wyznaczonych przez wartości zmiennej c26 (uczestniczenie w wyborach parlamentarnych w 2007 roku), gdzie pierwsza grupa identyfikowana jest przez wartość 1 (uczestniczący w wyborach), natomiast druga grupa przez wartość 2 (nieuczestniczący w wyborach)
`/MISSING=ANALYSIS`	- dla braków danych wybierz opcję *wyłączanie obserwacji analiza po analizie*
`/CRITERIA=CIN(0.95).`	- przyjmij przedział ufności 0,95. Poziom ufności wynosić będzie 0,05.

Po wykonaniu tych czynności uzyskujemy w oknie raportu wyniki przeprowadzonego testu dla dwóch prób niezależnych:

Group Statistics

Udział w wyborach parlamentarnych 21 października 2007.	*N*	*Mean*	*Std. Deviation*	*S.E. Mean*
WIEK tak, głosowałem(am)	1274	55.10	16.11	.45
nie, nie głosowałem(am)	509	52.21	17.78	.79

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	*Sig.*	*t*	*df*	*Sig. (2-tailed)*	*Mean Difference*	*Std. Error Difference*	*Lower*	*Upper*
WIEK	Equal variances assumed	10.44	.00	3.32	1781.00	.00	2.89	.91	1.11	4.67
	Equal variances not assumed			3.18	859.29	.00	2.89	.91	1.11	4.67

W tabeli *Group Statistics* podane zostały wartości statystyk opisowych dla każdej z porównywanych grup oddzielnie. Interesujące nas statystyki testu dla dwóch prób zostały podane w tabeli *Independent Samples Test.* Przed przystąpieniem do interpretacji uzyskanych wyników statystyki t musimy wybrać właściwą opcję tego testu - zakładając równość wariancji w obu grupach lub też nie. W pierwszej kolejności musimy zatem przystąpić do analizy homogeniczności wariancji, co program PSPP wykonuje automatycznie. Jest nim opisany w podrozdziale 16.1.3.3. test Levene'a jednorodności wariancji. Wyniki tego testu wskazują, że $p<0,05$, co oznacza, iż wariancje w porównywanych grupach nie są równe. Obliczenia dla testu t będziemy zatem odczytywać dla drugiego wiersza, gdzie nie zakładamy uprzedniej hipotezy o równości wariancji (*Equal variances not assumed*). Poziom istotność dla statystki t okazał się być mniejszy niż przyjęte przez nas 0,05, w wyniku czego odrzucamy hipotezę zerową o równości średnich w porównywanych grupach. Interpretacja wyników będzie następująca: na podstawie testu t dla dwóch

prób niezależnych stwierdzamy, że średnia wieku osób uczestniczących w wyborach parlamentarnych w 2007 roku jest istotnie statystycznie różna od średniej wieku osób, które w nich nie uczestniczyły. Statystyka t wyniosła t(859)=3,18 przy poziomie istotności p<0,05. Wynik testu wskazuje, że osoby głosujące w wyborach były starsze niż osoby niegłosujące. Wskazuje na to wartość średniej wieku, która wynosiła odpowiednio 55,1 lat oraz 52,21 lat w poszczególnych grupach.

16.1.4.3. Test t-Studenta dla dwóch prób zależnych

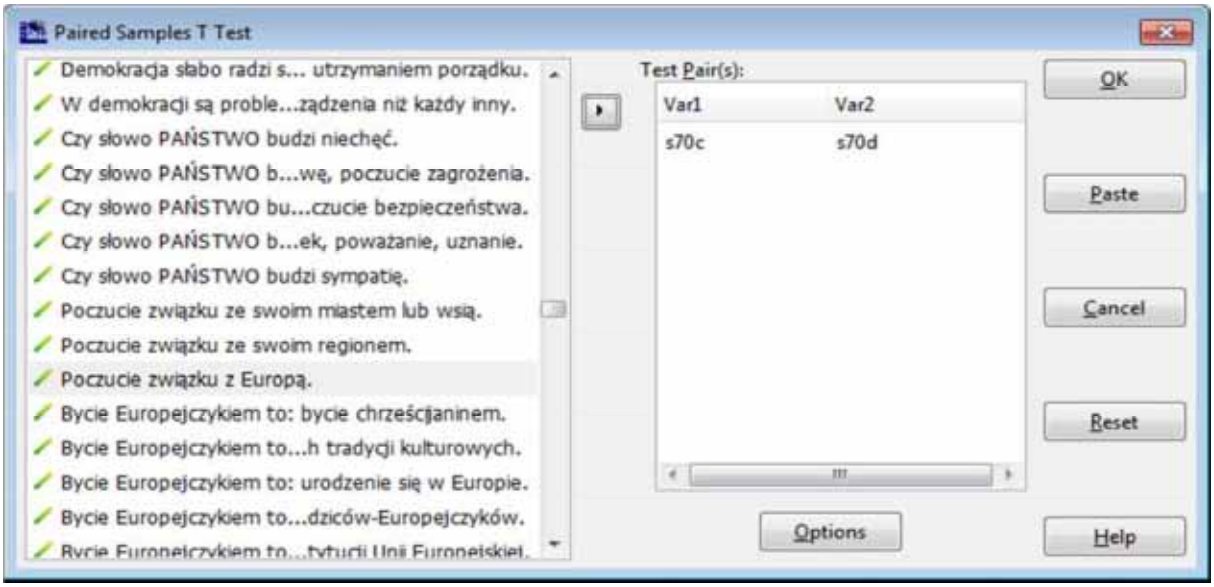
Kolejnym testem t-Studenta dostępnym w pakiecie PSPP jest test t dla dwóch prób zależnych. Zasady jego wykorzystania są tożsame z pozostałymi poznanymi przez nas testami. W przeciwieństwie jednak do nich, test t dla prób zależnych służy do porównywania średnich między grupami pochodzącymi z tej samej populacji bądź z powiązanych populacji. Z tego też względu, nazywany jest testem dla prób zależnych lub powiązanych. Wyznaczenie takich grup i zastosowanie tego rodzaju testu powinno wynikać zawsze powinno z postawionego problemu badawczego. Najprościej wykorzystanie testu t dla dwóch prób zależnych zrozumieć można, wyobrażając sobie sytuację eksperymentalną, kiedy chcemy sprawdzić reakcję danej grupy osób przed i po zaaplikowaniu pewnego bodźca. *De facto* będziemy porównywać ten sam zbiór jednostek analizy. Będą one jednak tworzyły dwie oddzielne grupy, gdyż będziemy chcieli porównywać reakcje, opinie, poglądy pewnej próby jednostek przed i po zdarzeniu, ponieważ mogły one ulec zmianie pod wpływem związanych z nim czynników zewnętrznych. Test t-Studenta dla dwóch prób zależnych służy zatem do porównywania średnich jednej zmiennej dla populacji tych samych jednostek. Jednak porównuje się rozkłady tej samej zmiennej w różnych odstępach czasu. Nie jest to jednak jedyny warunek zastosowania testu t-Studenta dla dwóch prób zależnych. Jeżeli zakładamy, że rozkłady porównywanych cech (zmiennych) są ze sobą powiązane, możemy również poszukiwać między nimi różnic z wykorzystaniem tego testu. Przykładem może być chociażby liczba godzin poświęcana na czytanie wybranych czasopism lub czas spędzany na oglądaniu określonych programów informacyjnych. Ważne jest, aby obie zmienne były mierzone na poziomie interwałowym lub ilorazowym, a także miały jednolitą skalę pomiaru. Przystąpmy do przeprowadzenia tego testu z użyciem programu PSPP. Podobnie jak w poprzednich podrozdziałach, dokonamy tego na przykładzie.

Przypuśćmy, że chcemy porównać uczestników badania PGSW pod względem stopnia poczucia związku z Polską (zmienna s70c) oraz poczucia związku z Europą (zmienna s70d). Obie zmienne są mierzone na jednakowej skali, gdzie 1 oznacza bardzo związany(a), 2 - raczej związany(a), 3 - raczej niezwiązany(a), 4 - w ogóle nie związany(a). Jest to co prawda skala porządkowa, jednak w tym przypadku może posłużyć nam również do porównywania średnich. Pozostałe odpowiedzi jak -1 - nieuzasadniony brak odpowiedz, 7 - trudno powiedzieć oraz 99 - uzasadniony brak odpowiedzi, zostały wykluczone z analiz. Przed przystąpieniem do analiz zakładamy, iż średnia poczucia związku z Polską jest różna od średniej poczucia związku z Europą. Przyjęty poziom ufności wynosi $\alpha=0,05$. Hipotezy statystyczne przyjmą następującą postać:

H_0 - średnia ocena poczucia związku z Polską **JEST RÓWNA** średniej ocenie poczucia związku z Europą *($m_p = m_e$*, gdzie m_p to średnia ocena poczucia związku z Polską, a m_e to średnia ocena poczucia związku z Europą). Test **NIE JEST** statystycznie istotny ($p \geq 0,05$);

H_1 - średnia ocena poczucia związku z Polską **NIE JEST RÓWNA** średniej ocenie poczucia związku z Europą *($m_p \neq m_e$)*. Test **JEST** statystycznie istotny ($p<0,05$).

W celu obliczenia testu t-Studenta dla dwóch prób zależnych, wybieramy w programie PSPP *Analyze ⇨ Compare Means ⇨ Paired Samples T Test,* po czym pojawia się poniższe okno:



Z listy zmiennych znajdującej się po lewej stronie wybieramy testowane zmienne (s70c oraz s70d). Zwróćmy uwagę na specyfikę pola *Test Pair(s)* znajdującego się po prawej stronie. Umieszczamy w nim pary zmiennych testowanych, gdzie pierwsza zmienna to Var1, natomiast druga - Var2. W *Options* wpisujemy właściwy przedział ufności, który w naszym przypadku wynosi 95 proc., gdyż przyjęliśmy poziom ufności α=0,05. Polecenie składni będzie miało następującą postać:

Składnia do wpisania w Edytorze	Opis działania składni
`T-TEST`	- wykonaj test t-Studenta dla dwóch prób zależnych
`PAIRS = s70c WITH s70d (PAIRED)`	- porównaj średnią ocen dla grupy osób pytanych o poczucie związku z Polską (zmienna s70c) ze średnią ocen dla grupy osób pytanych o związek z Europą (zmienna s70d)
`/MISSING=ANALYSIS`	- dla braku danych wybierz opcję *wyłączanie obserwacji analiza po analizie*
`/CRITERIA=CIN(0.95).`	- przyjmij przedział ufności 0,95; poziom ufności wynosić będzie 0,05.

Po wykonaniu tych operacji rozpoczyna się procedura obliczeniowa testu, którego wyniki publikowane są w raporcie wynikowym, prezentowanym na poniższym rysunku:

Paired Sample Statistics

		Mean	N	Std. Deviation	S.E. Mean
Pair 0	Poczucie związku z Polską.	1.42	1817	.66	.02
	Poczucie związku z Europą.	2.11	1817	1.14	.03

Paired Samples Correlations

		N	Correlation	Sig.
Pair 0	Poczucie związku z Polską. & Poczucie związku z Europą.	1817	.31	.00

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 0	Poczucie związku z Polską. - Poczucie związku z Europą.	-.69	1.13	.03	-.74	-.64	-26.20	1816	.00

Wyniki otrzymujemy w postaci trzech tabel. Pierwsza tabela *Paired Sample Statistics* zawiera zestawienie statystyk opisowych dla poszczególnych zmiennych testowanych. Druga tabela o nazwie *Paired Samples Correlations* podaje współczynnik korelacji R Pearsona między parami zmiennych. Istotność tej korelacji jest ważnym elementem interpretacyjnym. Jeżeli korelacja byłaby nieistotna ($p \geq 0,05$), wówczas między zmiennymi nie byłoby związku, a poszukiwana przez nas różnica między grupami byłaby nieuzasadniona. W naszym przykładzie korelacja jest dość słaba ($R=0,32$), ale istotna statystycznie. Między testowaną parą zmiennych istnieje zatem związek, więc słusznie przypuszczaliśmy, iż będą one tworzyły dwie próby zależne.

W kolejnym etapie przystępujemy do weryfikacji postawionej hipotezy zerowej o równości średnich w obu grupach. Do tego celu wykorzystujemy miary zamieszczone w tabeli *Paired Samples Test.* Podobnie, jak w przypadku pozostałych testów t-Studenta, w pierwszej kolejności sprawdzamy uzyskany poziom istotności. Jest on mniejszy niż przyjęte przez nas $\alpha=0,05$, zatem odrzucamy hipotezę zerową, stwierdzając, że średnia poczucia związku z Polską jest różna od średniej poczucia związku z Europą wśród badanej przez nas populacji. Wartość statystyki t wynosi -29,39 dla 1816 stopni swobody i jest wartością ujemną. Oznacza to, że średnia poczucia związku z Polską ma niższą wartość niż z Europą. Pamiętajmy jednak, że skala dla tych zmiennych miała postać odwróconą, gdzie 1 oznaczało *bardzo związany(a),* natomiast 4 - *w ogóle niezwiązany(a).* Im wartość średniej była bliższa wartości 1, tym większy stopień związku z Polską zaobserwujemy. Z przeprowadzonych analiz wynika, że badani czują większy stopień związania z Polską niż Europą, wyniki zaś są statystycznie istotne ($p<0,05$), co oznacza, iż wnioski uzyskane ze zbadania próby losowej możemy przenieść na badaną populację dorosłych Polaków.

16.2. Test U Manna-Whitney'a

W sytuacji badania dwóch grup niezależnych uznaliśmy, że najodpowiedniejszym rodzajem testu uzasadniającym istnienie między nimi różnic jest test t-Studenta dla dwóch prób niezależnych. Specyfika analizowanego przez nas zagadnienia nie zawsze jednak stwarza możliwość jego zastosowania. Jak pamiętamy, wykorzystanie testu t-Studenta wymaga spełnienia szeregu założeń. Kluczowe z nich to posiadanie przez zmienną rozkładu normalnego oraz interwałowego bądź ilorazowego poziomu pomiaru. Brak obecności tych wyróżników, nie pozwala na zastosowanie testu t-Studenta. Nie oznacza to jednak, że została zamknięta nam droga do porównywania dwóch grup i szukania między nimi różnic za pomocą testów istotności. Alternatywą wobec test t-Studenta jest test nieparametryczny nazywany testem U Manna-Whitney'a.

Nazwa testu U Manna-Whitney'a pochodzi od nazwisk jego twórców, amerykańskich matematyków i statystyków, Henry'ego B. Manna oraz jego studenta - Donalda R. Whitney'a, którzy jego ostateczną wersję zaproponowali w 1947 roku[2]. Wcześniej bowiem prace nad tym testem prowadzone były przez niemieckiego statystyka Gustava Deuchlera (1941), a następnie przez Franka Wilcoxona (1945), który opracował postać testu do porównywania wyłącznie prób równolicznych. H.B. Mann oraz D.R. Whitney dopracowali test o możliwość stosowania go dla grup o nierównych liczebnościach. W zaproponowanej przez nich postaci jest on nam współcześnie znany i stosowany.

[2] H.B. Mann, D.R. Whitney, *On a Test Of Whether One of Two Random Variables in Stochasticall Large Then The Other,* „The Annals of Mathematical Statistics", 1947, 18 (1), s. 50-60.

Test U Manna-Whitney'a służy do porównywania wyłącznie dwóch grup, które charakteryzują się niezależnością. Oznacza to, że dwie wyróżnione przez nas zbiorowości stanowią odrębne, niezwiązane ze sobą grupy pochodzące z różnych, odmiennych w sensie logicznym populacji. Ich wyszczególnienia dokonaliśmy sami, ze względu na interesujący nas problem badawczy. Przypominając sobie przykłady z wcześniejszych części niniejszego rozdziału, takimi dwiema grupami mogą być kobiety i mężczyźni, zwolennicy partii X i zwolennicy partii Y, uczestniczący i nieuczestniczący w wyborach parlamentarnych. Kolejnym krokiem jest sprawdzenie warunków umożliwiających zastosowanie tego testu.

Test U Manna-Whitney'a należy do dość liberalnych testów statystycznych. Jego zaletą są bowiem niewielkie wymagania. Jeżeli chcemy porównać dwie grupy niezależne, ale nasze zmienne nie spełniają warunków zastosowania test t-Studenta, alternatywnie skorzystamy z testu U Manna-Whitney'a. Wykorzystamy go w sytuacji, gdy analizowana zmienna, chociaż jest mierzona na poziomie interwałowym bądź ilorazowym, nie przyjmuje postaci rozkładu normalnego. Po drugie, sięgniemy po test U Manna-Whitney'a jeżeli testowana zmienna (zmienna zależna) jest mierzona co najmniej na poziomie porządkowym. To kryterium jest właściwie głównym warunkiem, który musimy spełnić. Niekiedy przy jego pomocy możemy również testować zmienne dychotomiczne (przyjmujące wartości 0 i 1), gdy zmienna nominalna po odpowiednim przekształceniu może przyjąć postać zmiennej porządkowej.

W przeciwieństwie do testu t-Studenta, który koncentruje się na porównywaniu średniej arytmetycznej w dwóch grupach, test U Manna-Whitney'a pod względem obliczeniowym opiera się na analizie rang. Z tego też względu niekiedy bywa on utożsamiany z testem sumy rang Wilcoxona. W najprostsze postaci polega to na zastąpieniu wartości zmiennej testowanej przypisanym im rangom, zaś samą procedurę nazywa się rangowaniem. Wartości zmiennej z dwóch grup są porządkowane rosnąco i kolejno numerowane za pomocą liczb naturalnych, z reguły poczynając od cyfry 1. Następnie wartości zmiennych w wydzielonych grupach zastępowane są przypisanymi im rangami. Dalsze obliczenia statystyczne przeprowadza się bowiem na rangach, które zastąpiły pierwotne wartości. Dla lepszego zrozumienia procedury rangowania oraz obliczania testu U Manna-Whitney'a wykorzystamy przykład.

Przypuśćmy, że chcemy sprawdzić, czy mężczyźni i kobiety różnią się pod względem stopnia zainteresowania polityką. W bazie PGSW zmienna ta występuje pod nazwą p48 i jest mierzona na poziomie porządkowym, o czym świadczy 5-stopniowa skala odpowiedzi, gdzie 1 oznacza *bardzo duże zainteresowanie*, 2 - *duże*, 3 - *średnie*, 4 - *nikłe* oraz 5 - *żadne*. Dla zrozumienia procedury rangowania będziemy analizować stanowisko dziesięciu kobiet (n=10) oraz dziesięciu mężczyzn (n=10). Tabela 37 zawiera informacje o rozkładzie odpowiedzi w tych grupach.

Tabela 37. Ocena zainteresowania polityką w grupie kobiet i mężczyzn

Kobiety (n=10)	Mężczyźni (n=10)
3	5
2	5
4	4
4	2
3	2
2	4
1	1
5	1
2	3
2	2

Przeprowadzenie procedury rangowania wymaga czterech czynności. W pierwszej kolejności łączymy odpowiedzi, czyli wszystkie wartości zmiennej dla poszczególnych jednostek analizy z grupy kobiet i mężczyzn włączamy do jednej tabeli. Następnie porządkujemy je rosnąco. W kolejnym kroku numerujemy odpowiedzi, zaczynając od 1. Efekt tych czynności prezentuje poniższa tabela.

Tabela 38. Porządkowanie i numerowanie wartości zmiennej dla poszczególnych jednostek analizy

Porządkowanie rosnące	Numerowanie
1	1
1	2
1	3
2	4
2	5
2	6
2	7
2	8
2	9
2	10
3	11
3	12
3	13
4	14
4	15
4	16
4	17
5	18
5	19
5	20

Trzecim krokiem naszego postępowania jest nadawanie rang. Jeżeli wartości zmiennej powtarzają się, stosuje się uśrednioną wartość rangi dla tych wartości. Przykładowo dla wartości 1 ranga będzie średnią z sumy przypisanych jej numerów porządkowych, czyli:

$$\frac{1+2+3}{3} = \frac{6}{3} = 2$$

Obliczone i przypisane rangi zostały zamieszczone w poniższej tabeli:

Tabela 39. Rangowanie wartości zmiennej dla poszczególnych jednostek analizy

Porządkowanie rosnące	Numerowanie	Obliczanie rang	Przypisanie rang
1	1	(1+2+3)/3=6/3=2	2
1	2		2
1	3		2
2	4	(4+5+6+7+8+9+10)/7=49/7=7	7
2	5		7
2	6		7
2	7		7
2	8		7
2	9		7
2	10		7
3	11	(11+12+13)/3=36/3=12	12
3	12		12
3	13		12
4	14	(14+15+16+17)/4=62/4=14,5	15,5
4	15		15,5
4	16		15,5
4	17		15,5
5	18	(18+19+20)/3=57/3=19	19
5	19		19
5	20		19

Ostatnią czynnością jest zastąpienie pierwotnych wartości zmiennej w poszczególnych grupach, czyli w próbie kobiet i mężczyzn, wyliczonymi rangami oraz ich zsumowanie. Dalsze wyliczenia oraz porównania będziemy bowiem prowadzić na rangach, co pokazuje poniższa tabela 40.

Tabela 40. Ocena zainteresowania polityką w grupie kobiet i mężczyzn po zastosowaniu procedury rangowania

	Kobiety		Mężczyźni	
	Pierwotne wartości zmiennej	Rangi	Pierwotne wartości zmiennej	Rangi
	3	12	**5**	19
	2	7	**5**	19
	4	15,5	**4**	15,5
	4	15,5	**2**	7
	3	12	**2**	7
	2	7	**4**	15,5
	1	2	**1**	2
	5	19	**1**	2
	2	7	**3**	12
	2	7	**2**	7
Suma rang		**104**		**106**

W przypadku testu U Manna-Whitney'a będziemy porównywać oraz poszukiwać różnic między grupą kobiet i mężczyzn na podstawie analizy rang. Uzasadnienie interpretacji naszego wyniku, podobnie jak w pozostałych testach, opierać będzie się na testowaniu ich istotności. Dla naszego przykładu przyjęliśmy poziom ufności wynoszący $\alpha=0,05$. Konieczne jest również postawienie właściwych hipotez statystycznych. Dla testu U Manna-Whitney'a będą brzmieć one następująco:

H_0 - kobiety i mężczyźni prezentują **RÓWNE (JEDNAKOWE)** zainteresowanie polityką. Test **NIE JEST** statystycznie istotny ($p\geq0,05$);

H_1 - kobiety i mężczyźni prezentują **RÓŻNE (NIEJEDNAKOWE)** zainteresowanie polityką. Test **JEST** statystycznie istotny ($p<0,05$).

Wszelkie procedury obliczeniowe dla testu U Manna-Whitney'a, w tym rangowanie, możemy przeprowadzić automatycznie w programie PSPP. Jego wykonanie wymaga jednak posłużenia się właściwym poleceniem w edytorze składni. Obecna wersja programu PSPP nie umożliwia jego przeprowadzenia w trybie okienkowym. Postać ogólna operacji w edytorze składni prezentuje się następująco:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/ MANN-WHITNEY =`	- wybierz statystykę testu U Manna-Whitney'a
`V1`	- oblicz ją dla zmiennej zależnej V1. Tutaj wpisujemy nazwę zmiennej testowanej, której wartości będziemy porównywać
`by V2 (group 1, group 2).`	- dokonaj tego w grupach wyróżnionych na podstawie zmiennej V2, gdzie w miejsce „group 1" wpisujemy wartość zmiennej identyfikującej grupę pierwszą, a w polu „group 2" wartość zmiennej dla drugiej porównywanej grupy
`EXECUTE.`	- wykonaj powyższą operację.

W naszym przykładzie składnia, którą wpiszemy w edytorze, przyjmie następującą postać:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/ MANN-WHITNEY =`	- wybierz statystykę testu U Manna-Whitney'a
`p48`	- w miejsce V1 wpisujemy nazwę zmiennej testowanej, czyli p48 (zainteresowanie polityką)
`by m2 (1, 2).`	- w miejsce V2 wpisujemy nazwę zmiennej identyfikującej grupę mężczyzn i kobiet, czyli m2 (płeć). Następnie w polu „group 1" wpisujemy wartość 1 (została ona przypisana na oznaczenie grupy mężczyzn), a w polu „group 2" wartość 2 (dla grupy kobiet)
`EXECUTE.`	- wykonaj powyższą operację.

Przed wykonaniem testu U Manna-Whitney'a należy pamiętać o zdefiniowaniu braków danych w przypadku zmiennej p48, czyli wykluczeniu takich wartości jak -1 - *nieuzasadniony brak odpowiedzi*, 7 - *trudno powiedzieć* oraz 99 - *uzasadniony brak odpowiedzi*.

Po wykonaniu powyższych operacji uzyskujemy wynik testu U Manna-Whitney'a, który prezentuje poniższy zrzut ekranowy:

Ranks

	N			Mean Rank		Sum of Ranks	
	mężczyzna	kobieta	Total	mężczyzna	kobieta	mężczyzna	kobieta
Zainteresowanie polityką.	721.00	1096.00	1817.00	799.98	980.72	576788.5	1074865

Test Statistics

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
Zainteresowanie polityką.	316507.5	576788.5	-7.60	.00

W przypadku testu U Manna-Whitney'a otrzymujemy wynik w postaci dwóch tabel. Pierwsza tabela *Ranks* zawiera statystki, które prezentowane są oddzielnie dla grupy kobiet i mężczyzn. Zawiera ona informacje o liczebnościach prób (*N*), wartość średniej rang (*Mean Rank*), która jest odpowiednikiem średniej arytmetycznej występującej w testach parametrycznych oraz sumę rang (*Sum of Ranks*). Wartość średniej rang służy do precyzyjnego określenia różnic między grupami, czyli stwierdzenia w której grupie natężenie danej cechy jest mniejsze bądź większe. W naszym przykładzie średnia rang okazała się być wyższa w grupie kobiet (980,72) niż mężczyzn (799,98). Pamiętajmy jednak, że skala ocen zainteresowań polityką przypisywała wartość 1 osobom najbardziej zainteresowanym polityką, z kolei wartość 5 - osobom w ogóle nią nieinteresujących się. Im wyższa zatem średnia ranga tym mniejszy stopień zainteresowania polityką. Wyniki powyższych obliczeń wskazują, iż kobiety są mniej zainteresowane polityką niż mężczyźni. Przyjęcie takiego wyniku wymaga sprawdzenia, czy jest on istotny statystycznie. W tym celu korzystamy z obliczeń zamieszczonych w kolejnej tabeli.

Najważniejsza dla interpretacji wyników testu U Manna-Whitney'a jest druga tabela zatytułowana (*Test Statistics*). Zawiera ona bowiem poziom istotności dla wyniku testu U Manna-Whitney'a, który jest niezbędny przy weryfikacji hipotez statystycznych. Jego wartość zamieszczona została w kolumnie zatytułowane *Asymp. Sig. (2-tailed),* co oznacza dokładnie istotność asymptotyczną dla obustronnego obszaru krytycznego. Program PSPP podaje dodatkowo dwie ważne dla nas wartości - statystykę dla testu U znajdującą się w kolumnie *Mann-Whitney U* oraz statystykę dla test Z (odpowiednio kolumna zatytułowana jako *Z*). Wartości te podawane są przy interpretacji wyników testu U Manna-Whitney'a. Ich wybór zależy od liczby jednostek analizy w obu grupach. Jeżeli liczebność próby wynosi $n \leq 20$, wybieramy wynik statystyki U. Gdy liczba jednostek przekracza 20 ($n>20$), wtedy podajemy wartość statystyki Z. Wynika to z faktu, iż przy dużych liczebnościach rozkład zmiennej testowanej przybliża się do rozkładu normalnego. Dla pełnej prezentacji wyników badania, dla prób większych niż 20 jednostek, zasadne jest jednak podawanie wartości obu statystyk.

Przystąpmy do interpretacji wyników obliczonych dla naszego przykładu. Na ich podstawie stwierdzamy, że mężczyźni i kobiety różnią się pod względem stopnia zainteresowania polityką. Wskazuje na to poziom istotności testu, który jest niższy niż przyjęte $\alpha=0,05$ dla $U=316507,5$ oraz $Z=-7,60$. Odrzucamy zatem hipotezę zerową mówiącą o braku różnic między grupami i przyjmujemy w jej miejsce hipotezę alternatywną. Ponadto, kobiety są w mniejszym stopniu zainteresowane polityką niż mężczyźni, na co wskazuje wartość średniej rang. Im wyższa bowiem wartość średniej rang (*Mean Rank*), tym mniejsze zainteresowanie polityką, gdyż w badaniu odpowiedź wyrażona była na skali 5-stopniowej, gdzie 1 oznaczało *bardzo zainteresowany(a)*, a 5 - *w ogóle niezainteresowany(a)*.

16.3. Test McNemara

Test McNemara jest kolejnym dostępnym testem statystycznym w pakiecie PSPP. Jego autorem, jak wskazuje nazwa, jest amerykański statystyk i psycholog - Quinn McNemara. Zapisał się on w historii nauki zrewidowaniem klasycznej skali pomiaru ilorazu inteligencji Stanforda-Bineta oraz wprowadzeniem w 1947 roku omawianego poniżej testu[3]. Opracowany przez niego test zaliczany jest do rodziny testów nieparametrycznych. Jest to klasa statystyk, które umożliwiają porównywanie różnic między grupami w przypadku zmiennych mierzonych na poziomie porządkowym lub nominalnym bądź ilościowym, ale których rozkład nie przyjmuje postaci rozkładu normalnego. Uznaje się, że testy nieparametryczne są bardziej liberalne i mniej wymagające względem restrykcyjnych założeń stawianych przez testy parametryczne.

Test McNemara jest testem stosowanym dla porównywania dwóch prób zależnych. Jest on przeznaczony wyłącznie dla zmiennych dychotomicznych, z reguły o wartościach zero-jedynkowych. Identyfikują one występowanie bądź niewystępowanie danej cechy (np. posiada poglądy prawicowe vs. nie posiada poglądów prawicowych), występowanie dwóch cech przeciwstawnych (np. posiada poglądy prawicowe vs. posiada poglądy lewicowe) lub wybranych dwóch wartości testowanych zmiennych (np. głosował na partię X vs. głosował na partię Y). Podobnie jak test t dla dwóch prób zależnych, wykorzystywany jest on do poszukiwania różnic między grupami w sytuacji eksperymentu - przed i po zaaplikowaniu bodźca. Jego celem jest szukanie różnic w reakcji grup na skutek oddziaływania tego impulsu. Przy jego pomocy analizuje się różnicę między rozkładem proporcji, inaczej frakcjami zmiennej, których łączna suma wynosi 1. Analizowane obserwacje zawsze muszą pochodzić z prób wyznaczonych z tej samej populacji bądź dwóch prób skorelowanych. W wymiarze technicznym, odnotowywanie różnic dokonuje się na podstawie porównywania rozkładów proporcji dwóch zmiennych w czteropolowej tabeli krzyżowej. Celem naszej analizy jest każdorazowo sprawdzenie, czy uzyskane przez nas wyniki z próby losowej są istotne dla populacji. Założenia tego testu najłatwiej zrozumieć na przykładzie.

Przypuśćmy, że chcemy porównać różnicę pomiędzy proporcją osób głosujących i niegłosujących w dwóch grupach - osób uczestniczących w wyborach parlamentarnych w 2005 roku (zmienna c32) oraz w 2007 roku (zmienna c26). W bazie PGSW zmienne te przyjmują wartość 1 dla *tak, głosowałem(am)* oraz 2 - *nie, nie głosowałem(am).* Pozostałe wartość, -1 - *nieuzasadniony brak danych,* 7 - *trudno powiedzieć,* 99 - *uzasadniony brak danych,* wykluczamy za pomocą funkcji *Missing* lub przekształcamy korzystając z funkcji *Recode into Different Variables.* W naszym przypadku bardziej korzystne jest zrekodowanie zmiennych za pomocą następującego polecenia:

```
RECODE
c32 c26
(1=1) (2=2) (ELSE=SYSMIS)
INTO V1 V2 .
VARIABLE LABELS V1 'Wybory w 2005 roku' V2 'Wybory w 2007 roku'.
EXECUTE .
VALUE LABELS
V1 TO V2
1 'głosujący'
2 'niegłosujący'.
EXECUTE.
```

[3] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, „Psychometrika", 1947, 12, s. 153-157.

Punktem wyjścia do zrozumienia idei zastosowania testu McNemara jest przyjrzenie się rozkładom częstości analizowanych zmiennych. Zostały one zaprezentowane w poniższych tabelach.

Tabela 41. Podział badanych ze względu na uczestnictwo w wyborach parlamentarnych w 2005 roku (N=1719)

V1. Udział w wyborach parlamentarnych w 2005 roku	Wskazania respondentów	
	N	%
Głosujący	1295	75,3
Niegłosujący	424	24,7
Razem	1719	100,0

Tabela 42. Podział badanych ze względu na uczestnictwo w wyborach parlamentarnych w 2007 roku (N=1719)

V2. Udział w wyborach parlamentarnych w 2007 roku	Wskazania respondentów	
	N	%
Głosujący	1233	71,7
Niegłosujący	486	28,3
Razem	1719	100,0

Po wstępnym porównaniu rozkładów tych dwóch zmiennych przypuszczamy, że dokonały się zmiany w preferencjach interesującej nas grupy badanych w zakresie udziału w wyborach parlamentarnych w 2005 roku i 2007 roku. Mianowicie, zwiększyła się liczba osób niegłosujących w wyborach w 2007 roku. W teście McNemara interesuje nas jednak konfiguracja decyzji o uczestniczeniu w wyborach w 2005 roku i w 2007 roku. Możemy bowiem wyszczególnić cztery możliwe sytuacje w dwóch wariantach: po pierwsze, osoby o trwałych decyzjach w podziale na grupę osób głosujących w jednych i drugich wyborach oraz grupę osób, które nie głosowały w żadnych wyborach, po drugie, osoby zmieniające decyzję, czyli grupę osób, które głosowały w wyborach w 2005 roku, ale nie głosowały w 2007 roku, oraz odwrotnie - osoby, które nie głosowały w 2005 roku, ale zdecydowały się pójść do urn w 2007 roku. Istotą testu McNemara jest poszukiwanie różnic w proporcjach kombinacji par wartości tych zmiennych oraz sprawdzeniu czy obserwacje poczynione na podstawie analizy próby losowej dają się przełożyć na populację generalną. Po zestawieniu zmiennej V1 oraz V2 w tabeli krzyżowej (przypomnijmy, że należy wybrać z zakładki *Descriptive Statistics ⇨ Crosstabs*) otrzymujemy w tabeli 43 następujący rezultat:

Tabela 43. Porównanie uczestnictwa w wyborach w 2005 roku i w 2007 roku

Wybory w 2005 roku (V1)	Wybory w 2007 roku (V2)		Ogółem (w wierszach)
	głosujący	niegłosujący	
głosujący	0,63 (A)	0,13 (C)	0,76
niegłosujący	0,09 (B)	0,15 (D)	0,24
Ogółem (w kolumnach)	0,72	0,28	1

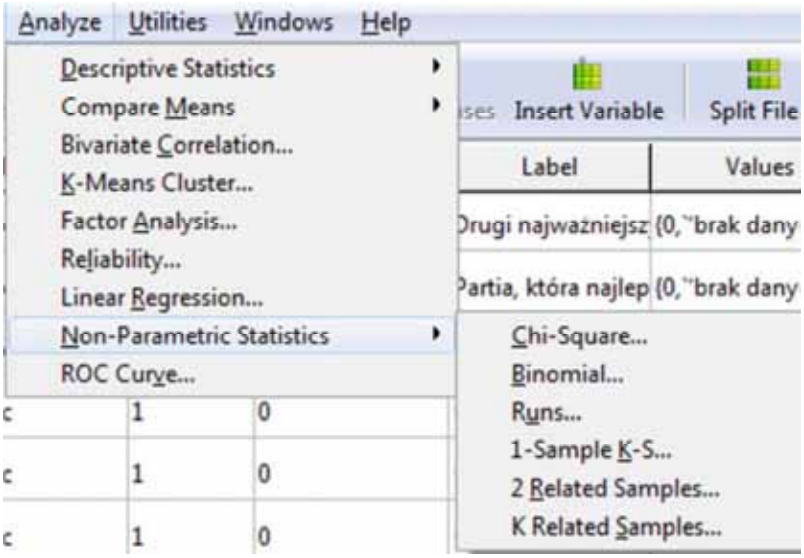
Tym sposobem skrzyżowaliśmy rozkład odpowiedzi dla dwóch prób: V1 - osób głosujących w wyborach w 2005 roku oraz V2 - osób głosujących w wyborach w 2007 roku, które będziemy porównywać. Na skrajach przecięcia głównej przekątnej tabeli, dostrzegamy frakcje świadczące o identyczności rozkładów

obu grup (są to sytuacje oznaczone literami A i D). Są to elementy porównawcze, które mówią o zgodności dwóch prób. Z kolei frakcje brzegowe, czyli komórki B oraz C, zawierają rozkłady świadczące o różnicach (braku zgodności) dwóch porównywanych przypadków (pomiarów). W potwierdzaniu istotności występowania tej różnicy w populacji ma nam pomóc test McNemara. Przed przystąpieniem do właściwej procedury jego obliczania w programie PSPP, musimy wpierw określić hipotezy statystyczne. Będą miały one następującą postać:

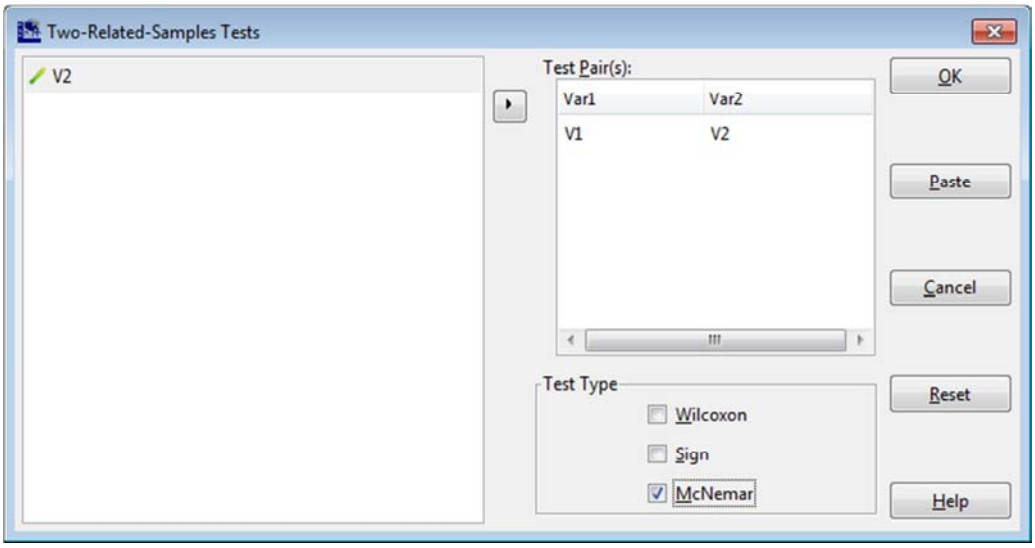
H_0 - rozkłady odpowiedzi dla grupy osób głosujących w wyborach w 2005 roku i 2007 roku **SĄ ZGODNE**; proporcje osób głosujących i nie głosujących w wyborach w 2005 roku **SĄ RÓWNE** proporcjom osób głosujących i niegłosujących w wyborach w 2007 roku. Test **NIE JEST** istotny statystycznie ($p \geq 0{,}05$).

H_1 - rozkłady odpowiedzi dla grupy osób głosujących w wyborach w 2005 roku i 2007 roku **NIE SĄ ZGODNE**; proporcje osób głosujących i niegłosujących w wyborach w 2005 roku **NIE SĄ RÓWNE** proporcjom osób głosujących i nie głosujących w wyborach w 2007 roku. Test **JEST** istotny statystycznie ($p < 0{,}05$).

Przystąpmy do weryfikacji naszych przypuszczeń za pomocą testu McNemara w programie PSPP. Jego wyboru dokonujemy poprzez *Analyze* ⇨ *Non-Parametric Statistics* ⇨ *2 Related Samples,* co zostało zaprezentowane na poniższym zrzucie ekranowym:



W efekcie wyboru tej opcji, wyświetli się poniższe okno:



Z listy zmiennych znajdującej się po lewej stronie okna wybieramy testowane zmienne V1 oraz V2, przenosząc je następnie do pola *Test Pair(s)* i ustalając właściwą ich kolejność - pierwszą zmienną V1 do pola *Var1* oraz drugą zmienną V2 w miejsce wskazywane przez *Var2*. Następnie na liście dostępnych testów (*Test Type*) zaznaczamy *McNemar*. W edytorze składni wykonywana operacja przyjmie postać:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/ MCNEMARA`	- wybierz statystykę testu McNemara
`V1`	- wybierz zmienną V1, która zawiera uczestnictwo w wyborach w 2005 roku
`WITH V2 (PAIRED).`	- wybierz zmienną V2, która zawiera uczestnictwo w wyborach w 2007 roku, a następnie porównaj je parami (*Paired*)
`EXECUTE.`	- wykonaj powyższą operację.

Efektem tych czynności jest następujący wynik w oknie raportu:

Wybory w 2005 roku & Wybory w 2007 roku

Wybory w 2005 roku	Wybory w 2007 roku	
	niegłosujący	głosujący
niegłosujący	265	159
głosujący	221	1074

Test Statistics

	N	*Exact Sig. (2-tailed)*	*Exact Sig. (1-tailed)*	*Point Probability*
Wybory w 2005 roku & Wybory w 2007 roku	1719	2.00	1.00	.00

Pierwsza tabela zatytułowana *Wybory w 2005 roku & Wybory w 2007 roku* podaje każdorazowo rozkład liczebności testowanych zmiennych w postaci czteropolowej tabeli krzyżowej. Do weryfikacji hipotezy zerowej potrzebujemy danych z drugiej tabeli o nazwie *Test Statistics.* Ważną dla nas wartością jest statystyka *Point Probability,* czyli stopień prawdopodobieństwa. Jest on tożsamy z wartością poziomu istotności (p). W przypadku testu McNemara nie ustalamy *a priori* przedziału ufności (jest on ustalony odgórnie jako). Arbitralnie uznajemy, iż dla $p<0,05$ zmiana rozkładów (np. opinii, oceny, reakcji) mierzona różnicą proporcji, jest różnicą istotną statystycznie. W analizowanym przykładzie wynik testu McNemara nakazuje odrzucić hipotezę zerową o równości rozkładów i stwierdzić, iż udział osób głosujących i niegłosujących w wyborach parlamentarnych w 2005 roku różni się od uczestniczących i nieuczestniczących w wyborach w 2007 roku. Są to wyniki statystycznie istotne dla poziomu $p<0,05$.

16.4. Test znaków

Innym rodzajem testu nieparametrycznego służącego do porównywania dwóch prób zależnych jest test znaków, nazywany również testem znaków rangowanych, a opracowany przez Franka Wilcoxona. Jest on stosowany dla zmiennych mierzonych co najmniej na poziomie porządkowym. Może być on również wykorzystany do porównywania grup zależnych do testowania zmiennych interwałowych i ilorazowych nieposiadających rozkładu normalnego.

Test znaków polega na porównywaniu wartości tej samej zmiennej w dwóch grupach zależnych na podstawie prawdopodobieństwa. W teście znaków przyjmujemy, że istnieje jednakowa szansa wystąpienia jednej z dwóch sytuacji, tj. 50 proc. szansy, że pierwsza próba będzie przejawiała mniejsze wartości zmiennej niż druga oraz 50 proc. szans, że w pierwszej grupie wartości zmiennej będzą wyższe w porównaniu do drugiej. Obliczenie testu znaków opiera się na zestawieniu wartości przypisanych dla tej samej jednostki analizy w jednej i drugiej grupie. Nazywamy to porównywaniem parami. Wymaga ona oznaczenia poszczególnych jednostek za pomocą znaku „+", jeżeli w drugiej grupie wartość zmiennej okazała się wyższa niż w pierwszej, lub znakiem „-", jeżeli była ona niższa. Celem takich obliczeń jest sprawdzenie, czy wartość pewnej cechy w jednej grupie jest większa bądź mniejsza niż w drugiej. Mówiąc prościej, czy wartość zmiennej wzrosła, spadła czy też została na tym samym poziomie, co we wcześniejszym badanym okresie. Przy weryfikacji wyników testu znaków również korzystamy z testu istotności oraz formułujemy stosowne hipotezy statystyczne, jednakże opierające się na regule prawdopodobieństwa. Zrozumienie zasad stosowania testu znaków umożliwi nam posłużenie się przykładem.

Załóżmy, że spytaliśmy się pewnej wylosowanej grupy Polaków równej N=200 o to, w jakim stopniu ufają politykowi X. Mogli oni udzielić odpowiedzi na pięciopunktowej skali, gdzie 1 oznacza *zdecydowanie nie ufam*, 2 - *raczej nie ufam*, 3 - *ani ufam, ani nie ufam*, 4 - *raczej ufam* oraz 5 - *zdecydowanie ufam*. Po pół roku zdecydowaliśmy się przebadać tą grupę pod tym samym kątem i zadaliśmy im identyczne pytanie. Zauważmy, że badaliśmy tą samą grupę osób, ale w różnych odstępach czasu, gdzie pomiar pierwszy oznaczymy jako X1, a pomiar drugi, który odbył się pół roku później - X2. Pewien wycinek uzyskanych odpowiedzi prezentuje tabela 44.

Tabela 44. Ocena zaufania do polityka X

Lp.	Pomiar X1	Pomiar X2 (pół roku później)
1	3	2
2	3	4
3	2	3
4	1	3
5	2	2
6	4	1
7	3	5
8	5	2
9	5	3
10	5	4

W kolejnym etapie musimy oznaczyć poszczególne jednostki analizy za pomocą symboli „+" i „-". Znak „+" przypiszemy obserwacją, które w drugim pomiarze uzyskały wyższą ocenę zaufania niż w pierwszym. Tego rodzaju przypadki nazywamy różnicami pozytywnymi. Analogicznie, znak „-" przyznamy jednostkom, które uzyskały niższą ocenę w drugim pomiarze w porównaniu do pierwszego. Będą to przypadki określone jako różnice negatywne. Tabela 45 prezentuje wynik procedury przyznawania znaków.

Tabela 45. Wynik procedury nadawania znaków dla oceny zaufania do polityka X

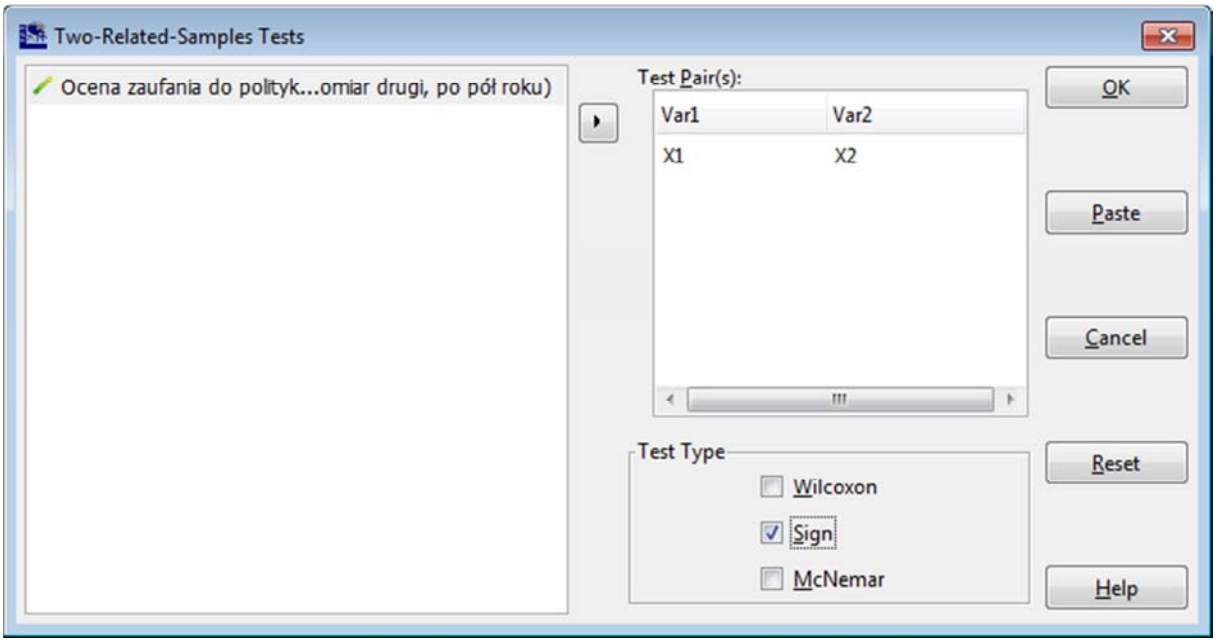
Lp.	Pomiar X1	Pomiar X2 (pół roku później)	Znaki
1	3	2	-
2	3	2	-
3	2	3	+
4	1	3	+
5	2	2	odrzucony
6	4	1	-
7	3	5	+
8	5	2	-
9	5	3	-
10	5	4	-

Zauważmy, że jednostce analizy o numerze porządkowym 5 przypisaliśmy status „odrzucony". Obserwacje, które nie wykazały różnicy w ocenie, czyli ocena zaufania w jednym jak i drugim pomiarze była taka sama, zostają bowiem wykluczone z całościowej próby badawczej. W kolejnych krokach obliczenia są dokonywane na próbach o liczebnościach pomniejszonych o jednostki nie wykazujące zmian w ocenie. W naszym przypadku wielkość próby została zmniejszona o 1 i wynosi ostatecznie N=9. W następnym etapie zliczamy liczbę znaków „+" i porównujemy ją z końcową liczebnością próby. Obliczenia wykazały, iż liczba znaków dodatnich wynosi 3, zaś całkowita liczebność próby 9. Znaków ujemnych jest zatem 6. Wynika z tego, że stopień zaufania do polityka X zmienił się na przestrzeni półrocznego okresu jego działalności. Co więcej, przewaga różnic negatywnych sugeruje, iż zaufanie do tego polityka spadło. Wniosek nasz musimy jednak uzasadnić w oparciu o test istotności. Wpierw jednak określmy hipotezy statystyczne, które będą brzmiały następująco:

H_0 - ocena zaufania do polityka X **NIE ZMIENIŁA SIĘ** w ciągu półrocznego okresu jego działalności. Test **NIE JEST** statystycznie istotny ($p \geq 0{,}05$);

H_1 - ocena zaufania do polityka X **ZMIENIŁA SIĘ** w ciągu półrocznego okresu jego działalności; jest ona obecnie lepsza bądź gorsza niż wcześniej. Test **JEST** statystycznie istotny ($p < 0{,}05$).

W programie PSPP test znaków możemy obliczyć poprzez wybranie *Analyze* ⇨ *Non-Parametric Statistics* ⇨ *2 Related Samples,* w wyniku czego pojawia się poniższe okno:



Z listy zmiennych znajdującej się po lewej stronie wybieramy dwa porównywane grupy, które tworzą dwie oddzielne zmienne: X1 dla pomiaru pierwszego oraz X2 dla pomiaru drugiego. Zmienną X1 wprowadzamy w miejsce pola *Var1,* z kolei zmienną X2 zawierającą ocenę polityka X pół roku później w pole oznaczone jako *Var2.* W ramce zatytułowanej *Test Type* zaznaczamy *Sign,* czyli test znaków.

W naszym przykładzie składnia, którą wpiszemy w edytorze przyjmie następującą postać:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/SIGN`	- wybierz statystykę testu znaków
`X1`	- wybierz zmienną X1, która zawiera wartości oceny zaufania w pomiarze pierwszym
`WITH X2 (PAIRED).`	- wybierz zmienną X2, która zawiera wartości oceny zaufania w pomiarze drugim, a następnie porównaj je parami (*Paired*)
`EXECUTE.`	- wykonaj powyższą operację.

Po wykonaniu tych operacji uzyskujemy następujące tabele:

Frequencies

		N
Ocena zaufania do polityka X (pomiar pierwszy) - Ocena zaufania do polityka X (pomiar drugi, po pół roku)	Negative Differences	62
	Positive Differences	112
	Ties	26
	Total	200

Test Statistics

	Ocena zaufania do polityka X (pomiar pierwszy) - Ocena zaufania do polityka X (pomiar drugi, po pół roku)
Exact Sig. (2-tailed)	.00
Exact Sig. (1-tailed)	.00
Point Probability	.00

Zamieszczony powyżej zrzut ekranowy zawiera dwie tabele. W pierwszej tabeli zatytułowanej *Frequencies,* podane zostały obliczenia dla procedury przypisywania znaków „+" oraz „-" jednostkom analizy podczas porównywania parami wyników pierwszego oraz drugiego pomiaru. Jak pamiętamy z wcześniejszej części rozdziału za różnicę dodatnią (znak „+") uznawaliśmy sytuację, gdy wartość zmiennej w drugiej grupie była większa niż w pierwszej. Różnicę ujemną (znak „-") przypisywaliśmy, jeżeli wartość ta okazała się niższa. W programie PSPP jednak pojawił się niewielki błąd. Liczebności dla różnic dodatnich oraz ujemnych zostały odwrotnie przypisane niż w rzeczywistości występują. Liczba różnic dodatnich jest bowiem podawana w wierszu zatytułowanym *Negative Differences,* czyli tam, gdzie powinna się znaleźć liczba różnic ujemnych. Natomiast liczbę znaków negatywnych odnajdziemy w wierszu *Positive Differences.* O ile wyliczenia są przeprowadzone prawidłowo, to błąd polega na odwróconym przypisaniu etykiet dla oznaczenia liczby znaków pozytywnych i negatywnych. Na tą pomyłkę należy zwrócić szczególną uwagę podczas interpretacji końcowych wyników testu. W wierszu *Ties* odnajdziemy z kolei liczbę jednostek analizy, w przypadku której nie odnotowaliśmy różnic w ocenie zaufania. Dla interpretacji wyników testu znaków ważna jest statystka o nazwie *Exact Sig. (2-tailed)* zamieszczona w kolejnej tabeli zatytułowanej *Test Statistics.* Jest ona odpowiednikiem poziomu istotności. W programie PSPP wszystkie testy nieparametryczne mają odgórnie ustalony przedział ufności dla testowania istotności wyników. Wynosi on 95 proc., dlatego poziom ufności każdorazowo przyjmujemy jako $\alpha=0,05$.

Przystąpmy do analizy wyników testu znaków dla naszego przykładu. Na postawie powyżej zamieszczonych obliczeń stwierdzamy, iż zaufanie Polaków do polityka X zmieniło się na przestrzeni półrocznego okresu jego działalności. Wynik testu jest istotny statystycznie ($p<0,05$). Ponadto, liczebności wykrytych różnic pozytywnych oraz negatywnych wskazują, iż zaufanie do tego polityka spadło. Większa bowiem grupa badanych zadeklarowała w drugim pomiarze niższy stopień zaufania (*Positive Differences = 112*) w porównaniu do grupy osób, u których zaufanie względem polityka X zwiększyło się (*Negative Differences = 62*).

16.5. Test rang Wilcoxona

Kolejnym testem nieparametrycznym służącym do porównywania dwóch prób zależnych jest test zaproponowany przez Franka Wilcoxona, znany pod nazwą testu par rangowanych znaków Wilcoxona lub testu par obserwacji Wilcoxona. Jest on stosowany do porównywania dwóch równolicznych prób pochodzących z tej samej populacji. Służy on do testowania zmiennych mierzonych na poziomie porządkowym, a także niekiedy nominalnym i interwałowym. Test Wilcoxona opiera się na poszukiwaniu różnic

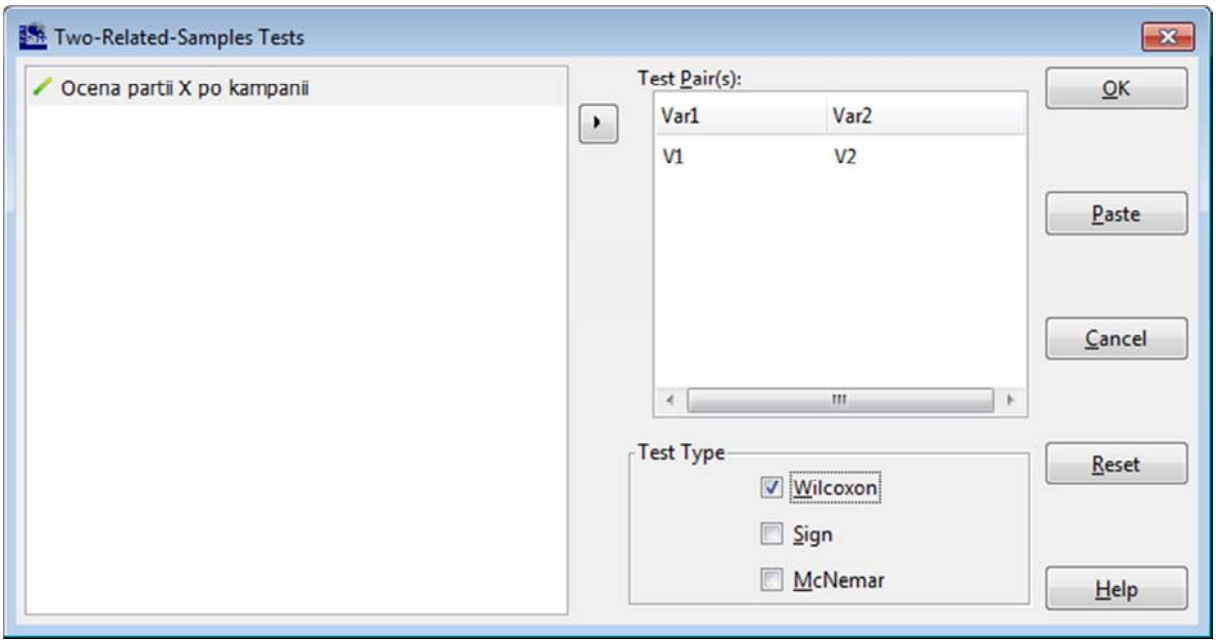
pomiędzy parami wartości zmiennych dla tej samej jednostki analizy. W przeciwieństwie do testu t-Studenta służącego do poszukiwania zmian na podstawie różnic wartości średnich arytmetycznych, test Wilcoxona polega na porównywaniu różnicy między medianami. Jest on przeznaczony do porównywania grup eksperymentalnych. Przy jego pomocy poszukuje się różnic między rozkładami tej samej cechy (zmiennej) w dwóch oddzielnych pomiarach (próbach), dążąc do zdiagnozowania zmian, które mogły zajść pod wpływem czynników na nie oddziałujących. W celu zrozumienia założeń zastosowania testu Wilcoxona, posłużymy się przykładem.

Przypuśćmy, iż chcemy porównać, jak Polacy oceniali partię X przed rozpoczęciem kampanii wyborczej i po jej zakończeniu. Do tego celu wykorzystaliśmy 5-stopniową skalę, gdzie 1 oznaczało *zdecydowanie źle*, 2 - *raczej źle*, 3 - *ani źle, ani dobrze*, 4 - *raczej dobrze*, 5 - *zdecydowanie dobrze*. Dokonaliśmy dwóch pomiarów na tej samej równolicznej próbie badawczej, pierwszy - przed rozpoczęciem kampanii wyborczej, drugi - tuż po jej zakończeniu. Zgodnie z logiką formułowania hipotez statystycznych, dla testu Wilcoxona hipotezę zerową oraz alternatywną określimy następująco:

H_0 - ocena partii X **JEST TAKA SAMA** przed jak i po zakończeniu kampanii wyborczej; mediany ocen w pierwszej i drugiej grupie **SĄ RÓWNE** ($Me_1 = Me_2$, gdzie Me_1 to mediana w próbie osób badanych przed kampanią wyborczą partii X, natomiast Me_2 jest medianą dla drugiej próby osób, czyli badanych po kampanii wyborczej partii X). Test **NIE JEST** istotny statystycznie ($p \geq 0{,}05$);

H_1 - ocena partii X **JEST RÓŻNA** przed jak i po zakończeniu kampanii wyborczej; mediany ocen w pierwszej i drugiej grupie **NIE SĄ RÓWNE** ($Me_1 \neq Me_2$, gdzie Me_1 to mediana w próbie osób badanych przed kampanią wyborczą partii X, natomiast Me_2 jest medianą dla drugiej próby osób, czyli badanych po kampanii wyborczej partii X). Test **JEST** istotny statystycznie ($p<0{,}05$).

Po tym etapie możemy przystąpić do obliczenia testu Wilcoxona w programie PSPP. Do tego celu posłużymy się danymi fikcyjnymi przygotowanymi dla próby N=230. W programie PSPP wybieramy kolejno *Analyze ⇨ Non-Parametric Statistics ⇨ 2 Related Samples,* po czym pojawia się poniższe okno:



Z listy zmiennych znajdującej się po lewej stronie wybieramy dwie porównywane zmienne - V1 dla zmiennej mierzącej poziom oceny partii X przed rozpoczęciem kampanii wyborczej oraz zmienną V2 - dla oceny partii X po zakończeniu kampanii. W następnej kolejności przenosimy je do pola *Test Pair(s).* W ramce zatytułowanej *Test Type* zaznaczamy *Wilcoxon.* Po wykonaniu tych czynności, klikamy OK.

Korzystając z edytora składni, polecenie wykonania testu Wilcoxona zapiszemy następująco:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/ WILCOXON`	- wybierz statystykę testu Wilcoxona
`V1`	- wybierz zmienną V1, która zawiera wartości oceny partii X przed kampanią wyborczą
`WITH V2 (PAIRED).`	- wybierz zmienną V2, która zawiera wartości oceny partii X po kampanii wyborczej, a następnie porównaj je parami (*Paired*)
`EXECUTE.`	- wykonaj powyższą operację.

Efektem tej operacji są następujące obliczenia:

Ranks

		N	Mean Rank	Sum of Ranks
Ocena partii X przed kampanią - Ocena partii X po kampanii	Negative Ranks	72	84.59	6090.50
	Positive Ranks	106	92.83	9840.50
	Ties	52		
	Total	230		

Test Statistics

	Ocena partii X przed kampanią - Ocena partii X po kampanii
Z	-2.76
Asymp. Sig. (2-tailed)	.01

W pierwszej tabeli zatytułowanej *Ranks* zamieszczono trzy ważne statystyki podane w kolumnie N. Dowiadujemy się z niej o liczbie przypadków, gdzie ocena partii X po kampanii wyborczej była niższa niż przed nią. Umieszczone jest to w wierszu o nazwie *Negative Ranks (Ujemne Rangi).* Drugi wiesz - *Positive Ranks (Dodatnie Rangi)* - informuje o liczbie ocen, które były większe w drugiej grupie niż w pierwszej. Natomiast wiersz *Ties (Wiązania)* podaje nam liczbę przypadków dla których wartość zmiennej w obu pomiarach się nie zmieniła. Druga tabela o tytule *Test Statistics* podaje wartość statystyki Z oraz asymptotyczny dwustronny poziom ufności. Na podstawie jego wartości ustalamy istotność uzyskanych wyników, a następnie przyjmujemy bądź odrzucamy hipotezę zerową. Przystąpmy do interpretacji wyników obliczonych dla naszego przykładu. Na ich podstawie stwierdzamy, że ocena partii X przed kampanią wyborczą oraz po jej zakończeniu uległa zmianie. Wskazuje na to poziom istotności testu, który jest niższy niż przyjęte α=0,05. Co więcej, ocena partii X po kampanii wyborczej wzrosła, na co wskazuje liczba rang dodatnich (n=106), która jest większa niż liczba rang ujemnych (n=72).

16.6. Test chi-kwadrat dla jednej próby

Test chi-kwadrat dla jednej zmiennej pozwala stwierdzić, czy rozkład badanej zmiennej istotnie różni się od rozkładu losowego lub innego rozkładu określonego przez badacza. Warunki zastosowania tego testu są następujące: przeznaczony jest on dla zmiennej mierzonej na poziomie nominalnym dla wystarczająco licznej próby. Za taką należy uznać próbę liczącą co najmniej 30 lub 50 jednostek analizy[4]. Ponadto, liczebności przypadające na poszczególne wartości zmiennej również powinny być wystarczające - na ogół podawane są liczby 5, 8 i 10 jednostek analizy[5]. Ponadto testowana zmienna nie musi spełniać warunku normalności rozkładu.

Baza danych PGSW nie dostarcza wdzięcznego (jednocześnie prostego i typowo przydatnego, bo politologicznego) przykładu dla wyjaśnienia idei tego testu. Posłużmy się zatem następującą, fikcyjną egzemplifikacją. Przypuśćmy, że przebadaliśmy grupę 100 studentów, pytając ich o poparcie dla ideologii populistycznych. Mogli oni w odpowiedzi wybierać pomiędzy alternatywami *popieram* i *nie popieram*. Załóżmy również, że skądinąd znamy poziom poparcia dorosłych Polaków dla ideologii populistycznych.

Chcielibyśmy się dowiedzieć, czy rozkłady cechy dla tych dwóch grup są ze sobą zgodne, a więc sprawdzić czy istnieje lub nie istnieje różnica pomiędzy zbadanymi przez nas studentami a dorosłymi Polakami. W tym celu można zastosować test chi-kwadrat dla jednej zmiennej. Tabela 46 prezentuje rozkłady mierzonej cechy dla obu grup - studentów i dorosłych Polaków (dla celów dydaktycznych przyjęto proste liczebności).

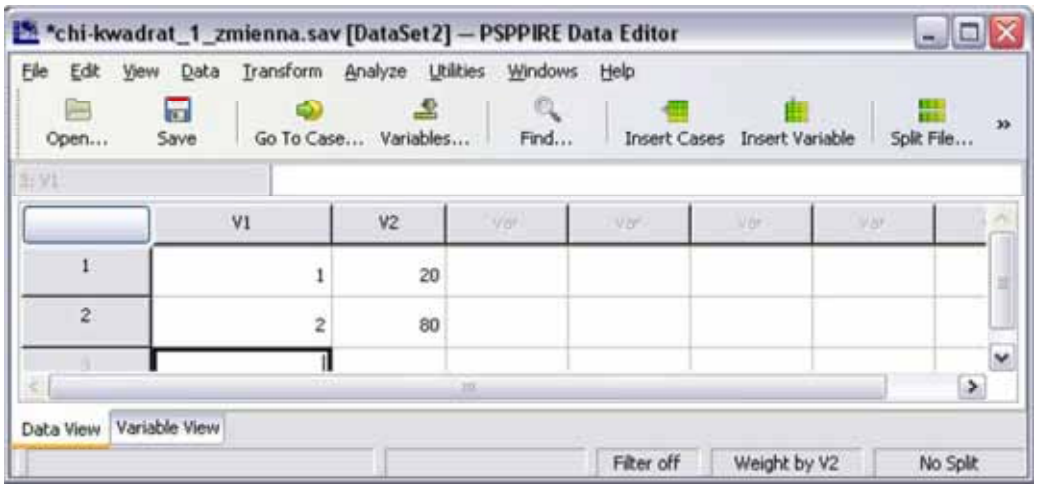
Tabela 46. Rozkład brzegowy mierzonej cechy w próbie studentów oraz populacji dorosłych Polaków

	Rozkład mierzonej cechy w grupie badanych studentów (N=100)		**Rozkład mierzonej cechy w populacji**
	N	**%**	**%**
popiera populizm	20	20	45
nie popiera populizmu	80	80	55

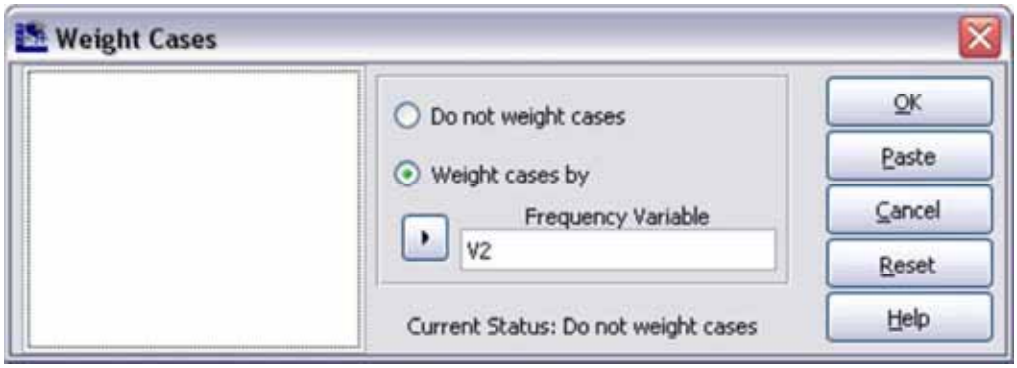
W celu rozwiązania wyłożonego wyżej problemu badawczego, wprowadzamy do programu PSPP zmienną dotyczącą odpowiedzi studentów w następującej konfiguracji: V1 oznaczać będzie poparcie (gdzie 1 - popiera, a 2 - nie popiera), a V2 - liczebność wskazanych przez studentów opcji.

[4] G.U. Yule, M.G. Kendall, *Wstęp do teorii statystyki*, Państwowe Wydawnictwo Naukowe, Warszawa 1966, s. 471.

[5] Porównaj: M. Sobczyk, *Statystyka,* Wydawnictwo Naukowe PWN, Warszawa 2002, s. 228 oraz G.U. Yule, M.G. Kendall, dz. cyt., s. 471.

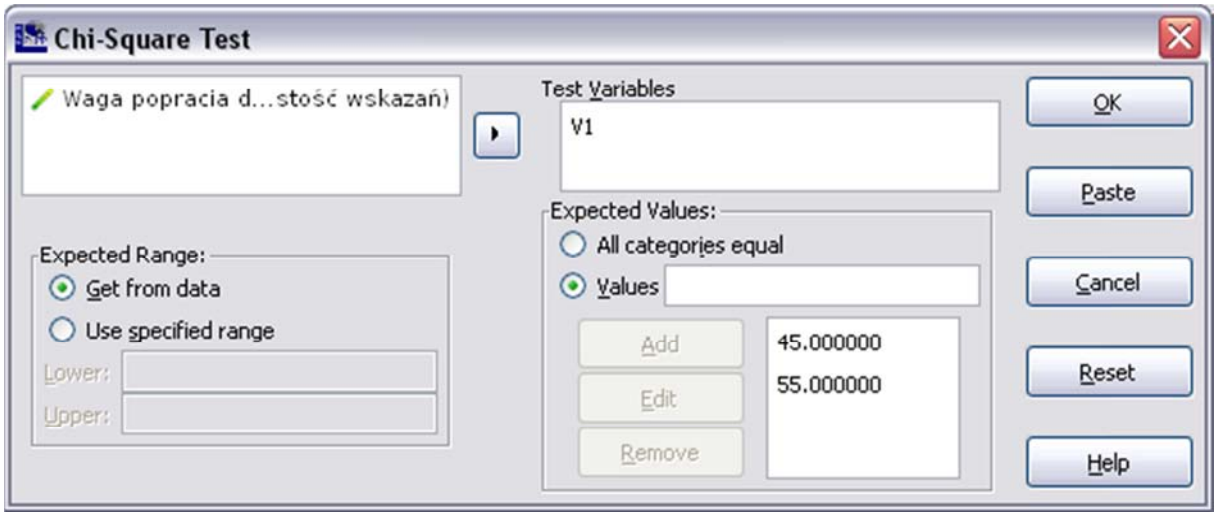


Następnie zmienną V1 ważymy liczbami zawartymi w zmiennej V2. Alternatywnie, moglibyśmy wykonać obliczenia posługując się tylko jedną zmienną liczącą 100 jednostek analizy, w tym 20 zmiennymi oznaczonymi jako *popiera* i 80 zmiennymi oznaczonymi jako *nie popiera.* Wybrany sposób zapisu jest jednak mniej czasochłonny.



Test chi-kwadrat dla jednej zmiennej obliczamy wybierając z zakładki *Analyze* ⇨ *Non-Paramteric Statistics* ⇨ *Chi-Square...* W oknie *Chi-Square Test* wprowadzamy w pole *Test Variables* zmienną V1, a w *Expected Values* w polu *Values* kolejno liczby 45 i 55. Ramka *Expected Values,* czyli wartości oczekiwane służy do wprowadzenia wartości zmiennej z rozkładem której porównujemy. W tym przypadku jest to rozkład odpowiedzi na pytanie o poparcia dla ideologii populistycznej dla populacji dorosłych Polaków. Wartości te muszą zostać wprowadzone w takiej kolejności, w jakiej ułożone są wartości testowane (zmienna V1).

Opcję *All categories equal* wybieramy wówczas, gdy oczekujemy, że rozkład, z którym porównujemy testowane zmienne jest losowy (a więc rozkład byłby w granicach 50/50 dla zmiennej dwuwartościowej, takiej, jak przez nas badana).



W edytorze, składnia dla testu chi-kwadrat przyjmuje następującą postać:

Składnia do wpisania w Edytorze	Opis działania składni
`NPAR TEST`	- wykonaj test nieparametryczny
`/CHISQUARE= V1`	- wybierz statystykę testu chi-kwadrat dla zmiennej testowanej V1
`/EXPECTED = 45 55.`	- porównaj z takimi wartościami oczekiwanymi, jak 45 oraz 55
`EXECUTE .`	- wykonaj powyższą operację.

W efekcie podjętych działań otrzymujemy następujące dwie tabele:

Poparcie dla ideologii populistycznych

	Observed N	Expected N	Residual
popiera	20	45,00	-25,00
nie popiera	80	55,00	25,00
Total	**		

Test Statistics

	V1
Chi-Square	25,25
df	1
Asymp. Sig.	,00

Wyniki interpretujemy zaczynając od tabeli zatytułowanej *Poparcie dla ideologii populistycznych,* w której porównywane są wartości obserwowane (kolumna oznaczona *Observed* - zmienna V2, czyli wynik badań studentów) z wartościami oczekiwanymi (kolumna oznaczona *Expected* - rozkład zmiennej w populacji dorosłych Polaków). Różnica pomiędzy wartościami obserwowanymi i oczekiwanymi została obliczona w ostatniej kolumnie zatytułowanej *Residuals* (Reszty). Tabeli tej używamy do wizualnej oceny różnic pomiędzy wartościami oczekiwanymi a wartościami obserwowanymi. Stosunkowo duża wartość ujemna w pierwszym wierszu tabeli, a wartość dodatnia w drugim, ujawnia znaczne różnice pomiędzy tym, co pomierzone, a tym co oczekiwane. Możemy przypuszczać (ale na podstawie samej tej tabeli - tylko przypuszczać), że istnieją istotne różnice pomiędzy studentami a populacją dorosłych Polaków.

Kluczowe wyniki dla interpretacji testu chi-kwadrat dla jednej zmiennej wyniki skupione są w tabeli drugiej. Wartość chi-kwadrat (*Chi-Square*) jest liczbą niestandaryzowaną (nieporównywalną z wynikami testów chi-kwadrat dla innych zmiennych), służy jedynie do porównywania jej z tablicami rozkładu

chi-kwadrat (program PSPP zrobił to już za nas prezentując efekt porównania w trzecim wierszu tabeli). Z kolei w drugim wierszu tabeli podano liczbę stopni swobody stanowiącą wynik działania: liczba wartości zmiennej - 1 (w tym przypadku mieliśmy dwie wartości zmiennej - *popiera* i *nie popiera,* co w efekcie daje wartość 1). Trzeci wiersz tabeli zawiera najważniejszy wynik analiz. Mówi nam o tym, czy zaobserwowana różnica między wartościami zaobserwowanymi a wartościami oczekiwanymi jest istotna statystycznie. Przyjmowane w prowadzonym teście hipotezy są następujące:

H_0 - **NIE ISTNIEJE** istotna statystycznie **RÓŻNICA** pomiędzy znanym rozkładem empirycznym populacji dorosłych Polaków, a zmierzonymi postawami studentów. Test **NIE JEST** statystycznie istotny ($p \geq 0{,}05$);

H_1 - **ISTNIEJE** statystycznie istotna **RÓŻNICA** pomiędzy rozkładem empirycznym populacji dorosłych Polaków, a zmierzonymi postawami studentów. Test **JEST** statystycznie istotny ($p < 0{,}05$).

Wartość $p<0{,}05$ oznacza, że wynik jest istotny statystycznie, a zatem odrzucamy hipotezę zerową. Z kolei jeśli wartość p przekroczy 0,05 oznacza to, że nie ma podstaw, by hipotezę zerową odrzucić, a co za tym idzie, stwierdzamy, że pomiędzy wartościami obserwowanymi a oczekiwanymi brak jest statystycznie istotnej różnicy.

W przypadku danych analizowanych w niniejszym przykładzie wartość $p<0{,}05$, a zatem z prawdopodobieństwem mniejszym niż pięć setnych odrzucamy hipotezę zerową. W praktyce oznacza to, że pomiędzy badanymi zmiennymi istnieje statystycznie istotna różnica: studenci w mniejszym stopniu popierają ideologie populistyczne niż ogół Polaków. W raporcie z badań zapisujemy wynik testu statystycznego następująco:

$$\chi^2(1, N = 100) = 25{,}25;\ p<0{,}05$$

W formalnym zapisie w pierwszej kolejności podajemy liczbę stopni swobody, następnie liczbę jednostek analizy oraz wynik testu chi-kwadrat, a po średniku prawdopodobieństwo prawdziwości hipotezy zerowej (jeśli wynik byłby statystycznie nieistotny, wówczas w miejsce $p<0{,}05$ wstawiamy litery *ni* będące skrótem od *nieistotny statystycznie*).

17

Rozdział 17. Badanie różnic między wieloma grupami - jednoczynnikowa analiza wariancji ANOVA

Analiza wariancji jest to metoda statystyczna, która umożliwia porównywanie wyodrębnionych przez badacza grup i orzekanie, czy pomiędzy nimi istnieją statystycznie istotne różnice. Istnieją liczne odmiany tej metody - jednoczynnikowe i wieloczynnikowe. W tym rozdziale wyłożono jednoczynnikową analizę wariancji. Nazwa ANOVA jest to akronim od anglojęzycznego **AN**alysis **O**f **VA**riance (wskazuje się, że bardziej adekwatną nazwą byłaby „analiza średnich"). Twórcą analizy wariancji jest genetyk i statystyk brytyjski Ronald Aymler Fisher (1890-1962). Wynalezione przezeń testy statystyczne (analiza wariancji, analiza dyskryminacyjna oraz metoda największej wiarygodności) są współcześnie powszechnie używane w naukach społecznych i przyrodniczych. R.A. Fisher objął po Karlu Pearsonie Katedrę Eugeniki w najstarszej angielskiej świeckiej uczelni - University College London[1]. Wkład tego uczonego w metody statystyczne zaczęto doceniać po II wojnie światowej - uczelnie z całego świata uhonorowały go siedmioma doktoratami *honoris causa*, a Richard Dawkins na łamach magazynu „Edge" nazwał go „ojcem współczesnej statystyki"[2].

Metoda analizy wariancji opiera się na porównywaniu miar wariancji w każdej z wyodrębnionych przez badacza grup. Na przykład za jej pomocą możemy dowiedzieć się, czy zwolennicy poszczególnych polskich partii politycznych różnią się istotnie pod względem wieku. ANOVA należy do testów parametrycznych co implikuje większą liczbę koniecznych warunków do spełnienia (ANOVA jest wymagającą metodą!). Charakteryzuje się ona za to większą mocą przeprowadzanych testów, dokładniejszym, bardziej wiarygodnym pomiarem oraz bardziej precyzyjną interpretacją wyników.

[1] Wiele ciekawych faktów z życia tego wybitnego uczonego znajdziemy w biografii opublikowanej przez jego córkę Joan Fisher Box. J. Fisher Box, *R.A. Fisher: the life of a scientist*, John Wiley, Nowy Jork 1978.

[2] R. Dawkins, *Who is the greatest biologist of all time. A talk with Armand Leroi*, „Edge", 15 maja 2011, w: http://www.edge.org/3rd_culture/leroi11/leroi11_index.html#dawkins, dostęp: styczeń 2012.

Jednoczynnikową analizę wariancji przeprowadzamy w trzech następujących etapach:

- po pierwsze, sprawdzamy, czy istnieją przesłanki do posłużenia się tą metodą, a więc czy zebrane dane odpowiadają określonym standardom;

- po drugie, wykonujemy test ANOVA i interpretujemy go;

- po trzecie, badamy, które z wyodrębnionych grup różnią się, a które nie.

Trzy wymienione etapy zostały szczegółowo opisane poniżej, bowiem procedura postępowania jest dość złożona.

17.1. Testowanie warunków koniecznych dla przeprowadzenia jednoczynnikowej analizy wariancji (Etap 1)

Etap ten polega na sprawdzeniu czy zbiór danych, który chcemy analizować spełnia warunki, by zastosować jednoczynnikową analizę wariancji. Istnieją cztery cząstkowe, warunki ustalające minimalny standard zbioru danych i parametrów analizowanych zmiennych: poziom pomiaru zmiennych, względna równoliczność wyodrębnionych do badania grup, wysoka homogeniczność wariancji w ramach badanych grup oraz by rozkład zmiennej objaśnianej był maksymalnie zbliżony do rozkładu normalnego. Zostały one szczegółowo opisane poniżej.

17.1.1. Sprawdzanie minimalnego poziomu pomiaru zmiennych

Zmienna zależna (wyjaśniana) musi być zmienną mierzoną co najmniej na poziomie porządkowym. Zmienna niezależna (wyjaśniająca) może być mierzona na dowolnej skali od nominalnej do ilorazowej, musi jednak przyjmować co najmniej dwie wartości. Ta zmienna wyznacza podział na badane grupy. Etap ten ogranicza się do sprawdzenia, czy zmienne, które chcemy analizować rzeczywiście posiadają wymienione wyżej właściwości, ewentualnie pogrupowaniu ich oraz zrekodowaniu. Czynności, które wówczas należy podjąć w programie PSPP zostały opisane w poprzednich rozdziałach.

Jak wskazano analiza wariancji ANOVA jest bardzo wymagającym testem statystycznym. Przeprowadzone analizy Polskiego Generalnego Studium Wyborczego 2007 wykazały, że żadna interesująca zmienna nie spełnia koniecznych dla wykonania analizy wariancji warunków (szczególnie normalności rozkładu oraz homogeniczności wariancji). Stworzono zatem fikcyjny zbiór danych zawierający dwie zmienne: poparcie dla demokracji mierzone na ilorazowej skali 11-punktowej (od 0 do 10) oraz wiek respondenta (zmienna porządkowa) zgrupowany według kryterium pokoleniowego: pierwsza grupa zawierała badanych należących do młodego pokolenia (od 18 do 35 lat), druga grupa należących do średniego pokolenia (powyżej 35 do 65 lat), a trzecia - należących do starszego pokolenia (powyżej 65 lat). Pierwszą z tych zmiennych uczyniono zmienną zależną, a drugą - zmienną niezależną.

17.1.2. Weryfikowanie liczebności jednostek analizy w grupach

Kolejnym koniecznym warunkiem przeprowadzenia jednoczynnikowej analizy wariancji jest uzyskanie względnej równoliczności wyodrębnionych do badania grup (wartości zmiennej niezależnej). Niektórzy badacze dopuszczają maksymalnie nie większe niż 10-procentowe różnice liczebności w badanych grupach, inni optują za całkowitą równolicznością. Na tym etapie koniecznymi czynnościami może okazać się uczynienie grup równolicznymi poprzez usunięcie części zmiennych lub uzyskanie w zbyt licznych grupach większej liczby przypadków danych do analizy (przeprowadzenie dodatkowych pomiarów jeśli to możliwe). Szczególnym sposobem uzyskania równoliczności grup może być imputacja danych w niepełnych jednostkach analizy.

Analizowany fikcyjny zbiór danych zawiera trzy równoliczne grupy reprezentujące trzy pokolenia: młode pokolenie (od 18 do 35 lat), średnie pokolenie (powyżej 35 do 65 lat) oraz starsze pokolenie (powyżej 65 lat). Każda z tych grup liczy 40 jednostek analizy (n=40, N=120). Informację o liczebności zbioru danych uzyskujemy tworząc prostą tabelę częstości.

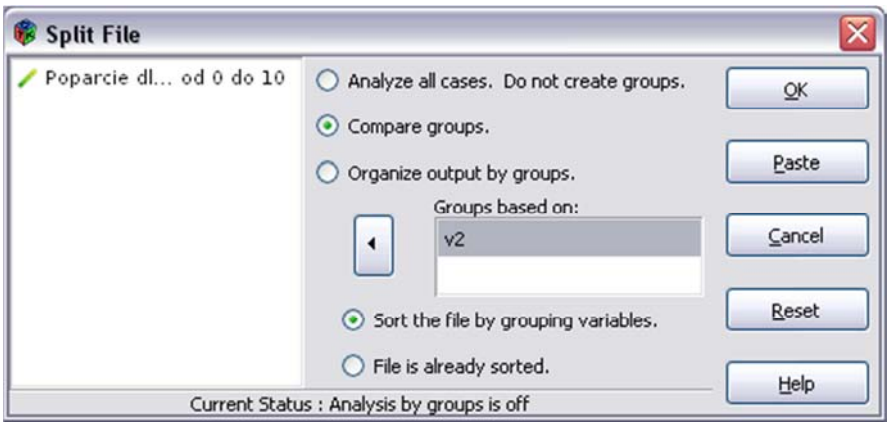
17.1.3. Testowanie normalności rozkładu zmiennej zależnej

Przetestowanie czy rozkład zmiennej zależnej zbliżony jest do tak zwanego rozkładu normalnego. Jest to warunek konieczny. Jeśli nie zostanie on spełniony nie możemy przeprowadzić jednoczynnikowej analizy wariancji. Najważniejszym narzędziem służącym do sprawdzenia, czy badana zmienna ma rozkład zbliżony do normalnego stanowi test Kołmogorowa-Smirnowa. Pomocne są także takie miary jak kurtoza, skośność oraz analiza histogramu.

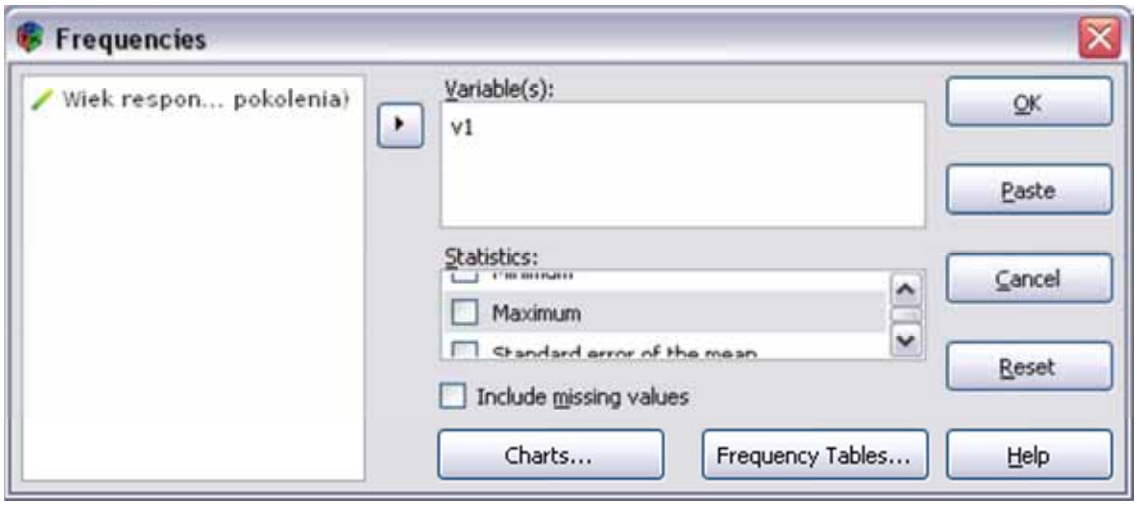
17.1.3.1. Analiza histogramu z nałożoną krzywą normalną

Ten etap ma charakter wstępnej oceny wizualnej. Negatywny wynik obserwacji nie wyklucza przeprowadzenia dalszych testów - analizy kurtozy i skośności oraz badania rozstrzygającego - testu Kołmogorowa-Smirnowa z poprawką Lilleforsa.

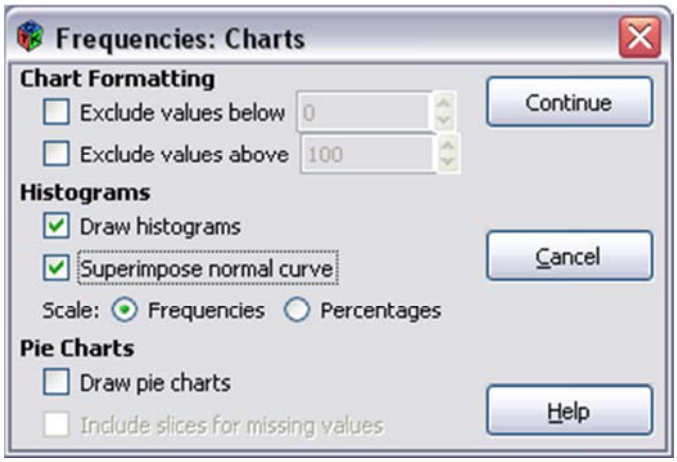
Konieczne jest wykonanie histogramów z nałożoną krzywą normalną dla każdej z wartości zmiennej niezależnej (każdej z pokoleniowych grup). W tym celu należy uruchomić opcję porównywania grupami: *Data* ⇨ *Split File* ⇨ *Compare Groups* i umieścić w polu *Groups based on:* zmienną niezależną.



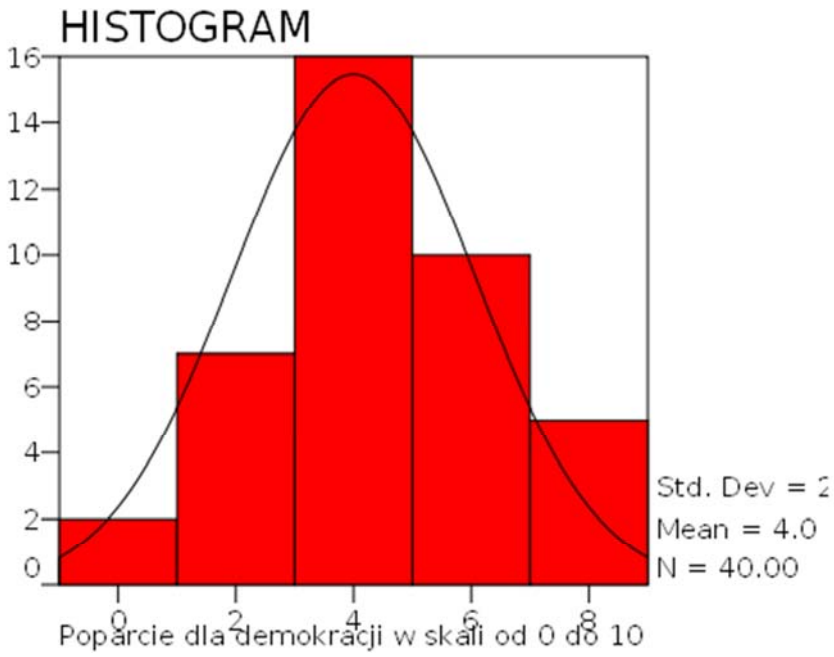
Po wykonaniu tej czynności przechodzimy do tabel częstości: *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies*. W polu *Variable(s)* umieszczamy zmienną zależną (w omawianym przypadku jest to stopień poparcia dla demokracji mierzony na 10-punktowej skali).



Następnie należy wybrać przycisk *Charts* umiejscowiony w dolnej przestrzeni okna:

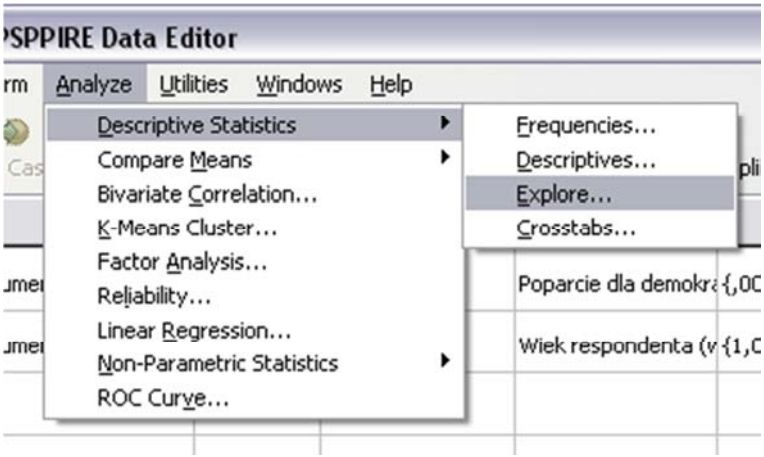


Należy zaznaczyć (jak na rysunku): wykreśl histogramy (*Draw histograms*) oraz nałóż krzywą normalną (*Superimpose normal curve*). W efekcie pojawią się trzy histogramy - odpowiednio dla każdej grupy pokoleniowej (zaprezentowany został tylko jeden z nich - dla starszego pokolenia). Jak wskazywano histogram podlega jedynie ocenie wizualnej, jest to wstępny element ewaluacji zbioru danych. Oceniamy w jakim stopniu wykres pasuje do krzywej normalnej. Należy zaznaczyć, że opieranie się wyłącznie na ocenie wizualnej jest mylące i nie może stać się jedyną podstawą wnioskowania. Na prezentowanym wykresie układ zmiennej wygląda obiecująco, bowiem przybiera kształt zgodny z krzywą normalną. Możemy przystąpić zatem do dalszych czynności. Podobny, pozytywny wynik uzyskano dla pozostałych dwóch analizowanych grup.

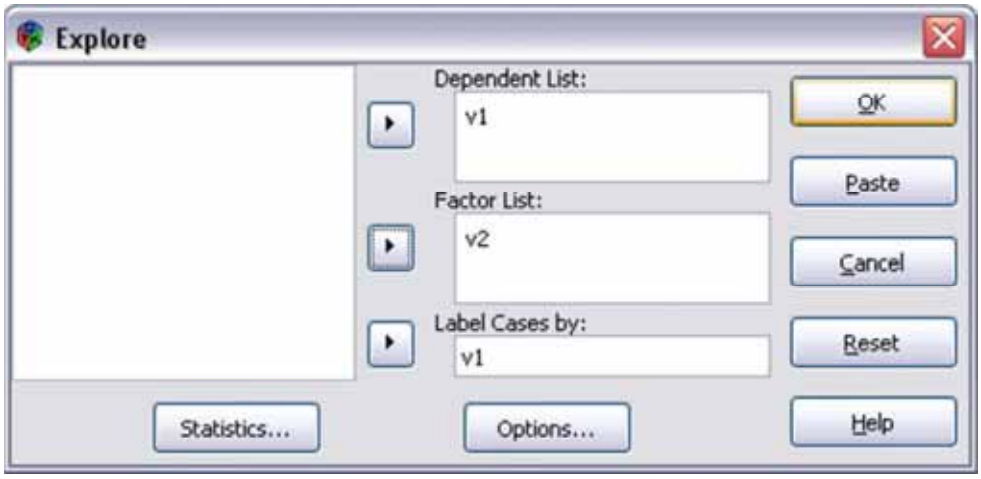


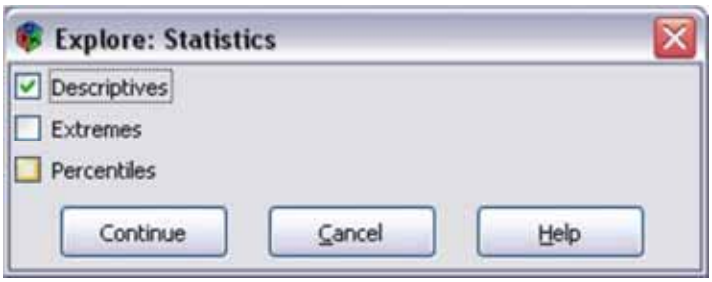
17.1.3.2. Analiza kurtozy i skośności

Na tym etapie dokonujemy analizy kurtozy i skośności. Ponadto niektóre wyniki uzyskiwane łącznie z wymienionymi mogą być pomocne w przygotowywaniu zbioru danych. Wybieramy opcję eksploracji danych w statystykach opisowych:



Jako zmienną zależną wstawiamy stopień poparcia dla demokracji, a jako czynnik (zmienną niezależną) - zmienną wiek w trzech pokoleniowych przedziałach. Jeśli chcemy uzyskać opisy zmiennych w tabelach, wówczas w *Label cases by* umieszczamy zmienną zależną. Otwieramy okno *Statistics* i zaznaczamy pierwszą opcję - statystyki opisowe (*descriptives*). Efektem tych działań są dwie następujące pary tabel: *Case Processing Summary* (podsumowanie) oraz *Descriptives* (statystyki opisowe) odpowiednio dla całości oraz dla trzech pokoleniowych grup.





W tabelach kontrolujemy następujące wyniki: kurtozę i skośność dla poszczególnych grup (w przypadku analizowanego przykładu grupy wiekowe) oraz dla całości.

W przypadku analizowanych zmiennych kurtoza wynosi odpowiednio: -0,71 dla młodego pokolenia, -0,46 dla średniego pokolenia i -0,42 dla starszego pokolenia. Znak ujemny oznacza spłaszczenie rozkładu w porównaniu z rozkładem normalnym a więc rozkład jest platykurtyczny, jednak ze względu na fakt, że wartość kurtozy w żadnym z przypadków nie jest mniejsza niż -1 nie odrzucamy twierdzenia o normalności rozkładu.

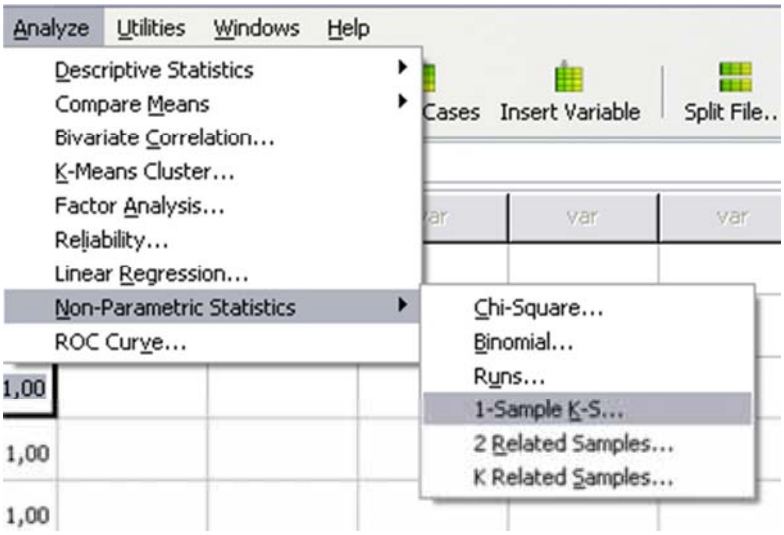
Z kolei wskaźnik asymetrii (skośność) jest miarą asymetrii rozkładu. Swiadczy o rozkładzie normalnym, jeśli wynosi dokładnie 0 lub o rozkładzie zbliżonym do normalnego jeśli zawiera się w zakresie od -1 do +1 (pominąwszy 0, które świadczy i idealnym rozkładzie normalnym). Wartość ujemna wskazuje na asymetrię lewostronną rozkładu, a wartość dodatnia na asymetrię prawostronną rozkładu. Rozkład wydłużony w prawą stronę to rozkład prawostronny, a rozkład posiadający tzw. „ogon" z lewej - lewostronny.

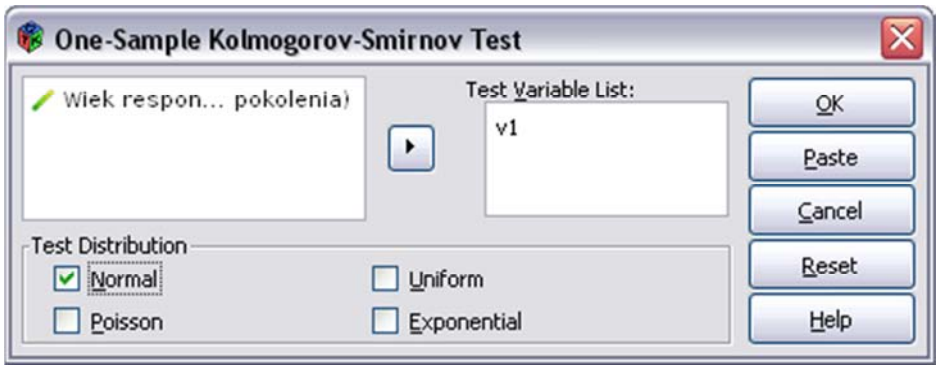
W analizowanym przypadku miary skośności (asymetrii) wynoszą odpowiednio: 0,16 dla młodego pokolenia, 0,24 dla pokolenia średniego oraz -0,06 dla pokolenia starszego. Dwie pierwsze wartości wykazują asymetrię prawostronną rozkładu, ostatnia - lewostronną. Żaden z wyników nie przekracza jedności, nie ma więc podstaw do odrzucenia twierdzenia o normalności rozkładu.

Ponadto warto zwrócić uwagę na wartości podane w tabeli *Extreme Values.* W przypadku wystąpienia wartości skrajnych można usunąć je i zastąpić średnimi dla danej grupy lub wykluczyć takie przypadki z dalszej analizy. Pomocniczą informacją mogą być także w minimum i maksimum wartości przyjmowanej przez zmienną, a także porównanie średniej z medianą lub pozostałe wartości skupienia i rozrzutu.

17.1.3.3. Test normalności rozkładu Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa (K-S) testuje normalność rozkładu. Powinien być on wykonywany z uwzględnieniem tak zwanej poprawki Lilleforsa, która jest obliczana, gdy nie znamy średniej lub odchylenia standardowego całej populacji, a tak jest w przypadku przykładu podanego w niniejszym rozdziale[3]. Test Kołmogorowa-Smirnowa można wykonać w PSPP wybierając w zakładce *Analyze* ⇨ *Non-Parametric Statistics* ⇨ *1-Sample K-S.*





Zmienną do testowania stanowi zmienna zależna. W *Test Distribution* zaznaczamy, że chcemy porównywać tę zmienną z rozkładem normalnym (zaznaczamy zatem *Normal*).

Interpretacja testu Kołmogorowa-Smirnowa wymaga uwzględniania dwóch wartości: statystyki Z oraz poziomu ufności (α). Test Kołmogorowa-Smirnowa opiera się na następujących dwóch hipotez zerowej (H_0) i alternatywnej (H_1):

H_0 - rozkład badanej cechy w populacji jest rozkładem normalnym,

[3] Brak informacji w dokumentacji technicznej PSPP, czy uwzględniono poprawkę Lilleforsa, każe ostrożnie traktować wyniki testu Kołmogorova-Smirnova. Niniejszy podrozdział należy traktować jako antycypujący rozwiązania, które dopiero zostaną wdrożone w programie. Należy przypuszczać, że w kolejnych wersjach test zostanie uzupełniony o poprawkę. W programie SPSS test K-S znajduje się on w *Analyze* ⇨ *Descriptive* ⇨ *Statistics* ⇨ *Explore* ⇨ *Plots* ⇨ *Normality plots with tests.* Innym polecanym, dysponującym większą mocą, lecz nie wdrożonym jeszcze do PSPP testem normalności rozkładu jest test Shapiro-Wilka. Jest on najbardziej polecany, jednakże ostrzega się, że może on dawać błędne wyniki dla jednostek analizy liczniejszych niż 2 000.

H_1 - rozkład badanej cechy w populacji jest różny od rozkładu normalnego.

Jeśli istotność jest niższa niż założony poziom α (domyślnie α=0,05) wówczas przyjmujemy H_1. Jeśli z kolei jest wyższa, wówczas nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że rozkład jest normalny.

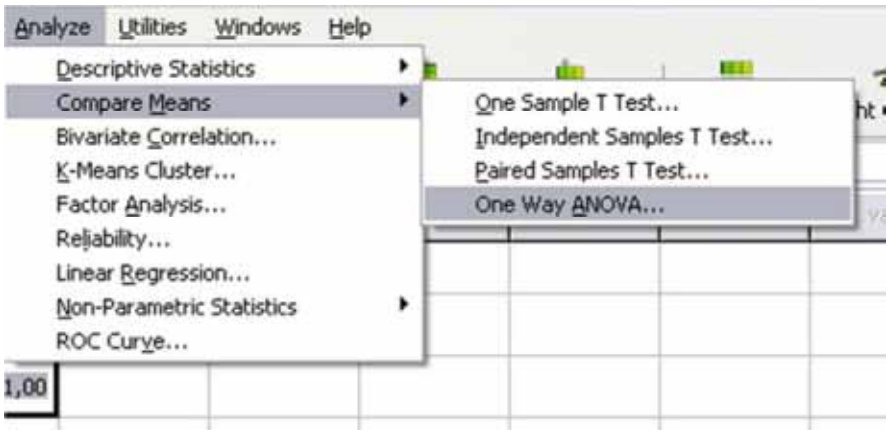
W tabeli, która jest efektem wykonania testu dwie interpretowane wartości znajdują się w ostatnich dwóch wierszach tabeli. Wartość α przekracza wartość 0,05, a zatem przyjmujemy H_0 i jednocześnie odrzucamy H_1. Zakładamy zatem, że rozkład jest zbliżony do normalnego. Im większa wartość Z tym większe odchylenie danych empirycznych od teoretycznego rozkładu (w niniejszym przypadku rozkładu normalnego). W analizowanym przypadku wynosi 1,23.

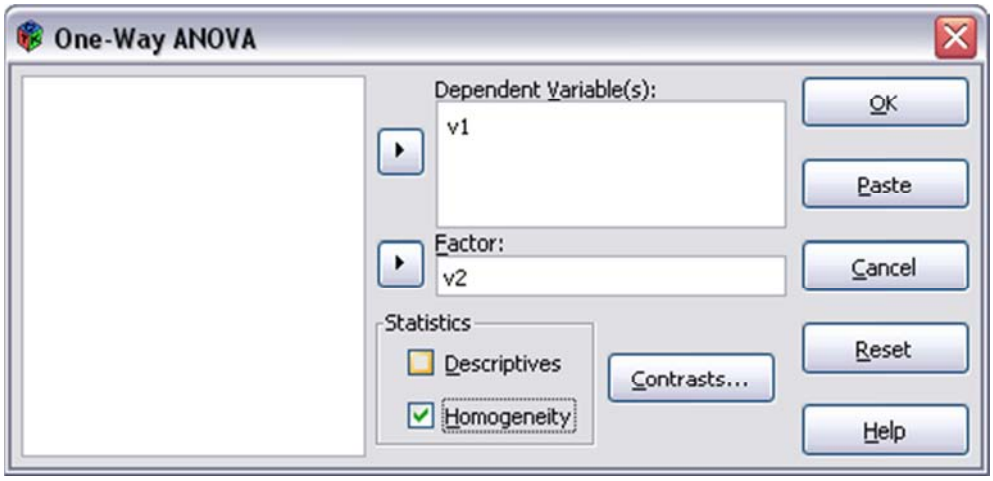
One-Sample Kolmogorov-Smirnov Test

		Poparcie dla demokracji w skali od 0 do 10
N		120
Normal Parameters	*Mean*	4,58
	Std. Deviation	2,22
Most Extreme Differences	*Absolute*	,11
	Positive	,11
	Negative	-,10
Kolmogorov-Smirnov Z		1,23
Asymp. Sig. (2-tailed)		,08

17.1.4. Testowanie homogeniczności wariancji zmiennej zależnej (test Levene'a)

Ten etap opisuje badanie czy rozkład zmiennej zależnej wykazuje jednorodność wariancji (brak wartości skrajnych). Jednorodność wariancji badamy testem homogeniczności wariancji Levene'a. Wykonujemy go jednocześnie z jednoczynnikową analizą wariancji ANOVA wybierając w *Analyze* ⇨ *Compare Means* ⇨ *One Way ANOVA,* a następnie zaznaczając w dolnej części okna *Homogenity.*





W teście jednorodności wariancji Levene'a posługujemy się następującą parą hipotez:

H_0 - różnica między wariancjami w badanych grupach jest jednorodna (lub zbliżona),

H_1 - wariancje w badanych grupach są różne.

Jeśli istotność jest niższa niż założony poziom α (domyślnie α=0,05) wówczas przyjmujemy H_1. Jeśli jest wyższa - nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że wariancja jest homogeniczna.

W przypadku analizowanych zmiennych poziom istotności p jest wyższy niż 0,05 i wynosi 0,92. Odrzucamy zatem hipotezę alternatywną i przyjmujemy hipotezę zerową stwierdzając, że różnica wariancji w testowanych grupach jest nieistotna statystycznie.

Test of Homogeneity of Variances

	Levene Statistic	*df1*	*df2*	*Significance*
Poparcie dla demokracji w skali od 0 do 10	,08	2	117	,92

17.2. Obliczenie i interpretacja jednoczynnikowej analizy wariancji ANOVA (Etap 2)

W etapie tym rozstrzygamy, czy badane grupy różnią się od siebie pod względem średnich czy nie. Jest to etap właściwy polegający na uzyskaniu statystyk jednoczynnikowej analizy wariancji ANOVA.

W jednoczynnikowej analizie wariancji formułujemy hipotezę statystyczną następująco:

H_0 - średnie w badanych grupach nie różnią się (są sobie równe),

H_1 - co najmniej jedna para średnich różni się od siebie (nie jest sobie równa).

Jeśli istotność jest niższa niż założony poziom α (domyślnie α=0,05) wówczas przyjmujemy H_1 stwierdzając, że co najmniej jedna para średnich różni się od siebie (nie jest sobie równa). Jeśli natomiast poziom α jest wyższy wówczas nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że badane grupy nie różnią się istotnie statystycznie pod względem średnich.

ANOVA

		Sum of Squares	df	Mean Square	F	Significance
Poparcie dla demokracji w skali od 0 do 10	Between Groups	160,87	2	80,43	22,08	,00
	Within Groups	426,30	117	3,64		
	Total	587,17	119			

W prezentowanym przykładzie wartość istotności wynosi $p \leq 0{,}001$, a zatem przyjmujemy hipotezę alternatywną konstatując, że co najmniej jedna para średnich różni się istotnie statystycznie od siebie. Prawidłowy zapis wyniku jednoczynnikowej analizy wariancji powinien zawierać następujący formalny zapis:

$$F\ (2,\ 117) = 22{,}08;\ p \leq 0{,}001$$

Symbolem F oznaczamy statystykę F (od nazwiska R.A. Fishera). Wyraża ona stosunek wariancji międzygrupowej do wariancji wewnątrzgrupowej. A zatem jest wynikiem dzielenia wartości znajdujących się w kolumnie zatytułowanej *Mean Square* (80,43/3,54=22,08). Liczby podane w nawiasie to tak zwane stopnie swobody - oznaczane literami df. W pierwszym wierszu tabeli widnieje liczba stopni swobody pomiędzy grupami. Obliczamy je: df = n-1, gdzie n oznacza liczbę grup. W drugim wierszu widnieje liczba stopni swobody w grupach. Obliczamy je analogicznie według wzoru df = n-1 (dla każdej z grup), a więc df = 120 - 1 = 119. Symbol n oznacza tu liczbę jednostek analizy, a więc badaną próbę (N=120).

17.3. Test *posthoc:* identyfikacja grup podobnych i różnych od siebie (Etap 3)

Procedury opisane w tym etapie podejmujemy opcjonalnie - tylko wówczas, gdy w Etapie 2 ustalono, że istotne statystycznie różnice pomiędzy grupami istnieją. Dzięki jednoczynnikowej analizie wariancji posiadamy informację na temat tego czy różnice istnieją czy nie - dodatkowe testy umożliwią rozpoznanie pomiędzy którymi grupami występują różnice. Służą do tego między innymi testy: Boneferroni, GH (Gamesa-Howella), LSD, Scheffe, Sidak, Tukey. Testy dzieli się na liberalne, a więc takie, które nie wymagają restrykcyjnych założeń, by wykazać istotną statystycznie różnicę pomiędzy średnimi w grupach i konserwatywne - wymagające takich założeń. Testem liberalnym jest na przykład LSD - Najmniejszej Istotnej Różnicy (*Least Significant Difference*). Z kolei do testów konserwatywnych zaliczane są test Boneferroniego, Tukey'a i najbardziej konserwatywny - Scheffego. Testy te posiadają także szereg innych właściwości, na przykład wskazaniem do zastosowania testu Gamesa-Howella jest nierównoliczność grup, a testu Tukey'a - duża liczebność grup. Najczęściej stosowanymi w testach *posthoc* są testy Tukey'a i Bonefferoniego.

W opisanej w niniejszej publikacji wersji programu nie zaimplementowano możliwości przeprowadzenia tych testów w trybie okienkowym, można jednak wykonać je używając okna składni.

Należy wpisać następujące komendy (należy zwrócić baczną uwagę na spacje oraz odpowiednio ich brak w niektórych miejscach składni - PSPP jest na nie wrażliwy):

Składnia do wpisania w Edytorze	Opis działania składni
`ONEWAY`	- wykonaj jednoczynnikową analizę wariancji ANOVA
`V1 BY V2`	- używając zmiennych V1 i V2
`/STATISTICS DESCRIPTIVES HOMOGENEITY`	- dołącz statystyki opisowe oraz test homogeniczności Levene'a
`/MISSING ANALYSIS`	- uwzględnij dane niepełne
`/POSTHOC = TUKEY ALPHA(.05).`	- wykonaj test *posthoc* Tukey'a, przy współczynniku α=0,05 (typowy poziom ryzyka popełnienia błędu I rodzaju w przyjmowany w naukach społecznych)

Wskutek wykonania komend zawartych w ostatniej linii wykonany zostaje test *posthoc* Tukey'a:

Multiple Comparisons

	(I) Wiek respondenta (w podziale na pokolenia)	(J) Wiek respondenta (w podziale na pokolenia)	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	młode pokolenie (18-35 lat)	średnie pokolenia (powyżej 35 do 65 lat)	-2,65	,43	,00	-3,66	-1,64
		powyżej 65 lat	-,45	,43	,54	-1,46	,56
	średnie pokolenia (powyżej 35 do 65 lat)	młode pokolenie (18-35 lat)	2,65	,43	,00	1,64	3,66
		powyżej 65 lat	2,20	,43	,00	1,19	3,21
	powyżej 65 lat	młode pokolenie (18-35 lat)	,45	,43	,54	-,56	1,46
		średnie pokolenia (powyżej 35 do 65 lat)	-2,20	,43	,00	-3,21	-1,19

W teście *posthoc* Tukey'a hipotezy zerowa i alternatywna brzmią następująco:

H_0 - porównywane dwie grupy różnią się pod względem średnich,

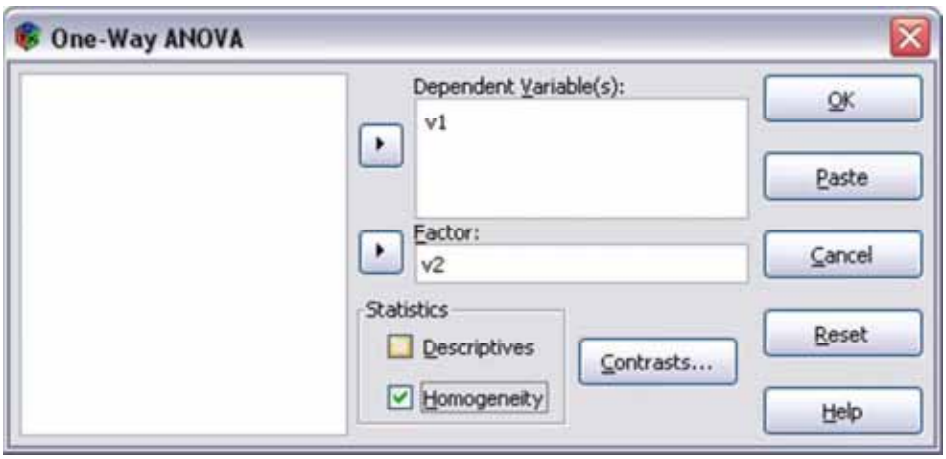
H_1 - porównywane dwie grupy nie różnią się istotnie statystycznie pod względem średnich.

Jeśli istotność jest niższa niż założony poziom α (założyliśmy w teście α=0,05) wówczas przyjmujemy H_1 stwierdzając, że porównywane grupy nie różnią się istotnie od siebie. Natomiast gdy poziom α jest wyższy lub równy wówczas nie ma podstaw do odrzucenia H_0, a więc przyjmujemy, że badane grupy różnią się istotnie statystycznie pod względem średnich. W analizowanym przypadku młode pokolenie istotnie statystycznie różni się od starszego na poziomie 0,43. Nie odnotowano natomiast różnic pomiędzy pokoleniem średnim i młodym oraz średnim i starszym (istotność na poziomie 0,01).

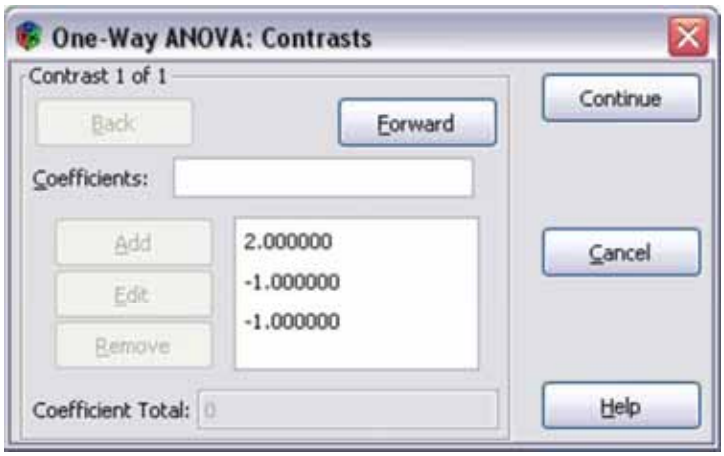
17.4. Testy *a priori:* poszukiwanie różnic w ramach zagregowanych grup (kontrasty)

Jest to możliwość dodatkowej eksploracji danych metodą jednoczynnikowej analizy wariancji ANOVA. Wskazaniem do zastosowania testu *a priori* jest teoretycznie ugruntowane założenie o istotnych różnicach pomiędzy określonymi grupami. Testy *a priori* pozwalają na przeciwstawienie sobie dowolnie zagregowanych grup. W przytaczanym przykładzie analizowano trzy grupy: młode pokolenie, średnie pokolenie i starsze pokolenie. Możemy jednak w toku analiz uznać, że potrzebujemy sprawdzić czy młode pokolenie istotnie różni się od pokolenia średniego i pokolenia starszego razem wziętych. Innym zastosowaniem kontrastów jest nadawanie grupom wag (tzw. współczynników kontrastów) - możemy je dowolnie zmniejszyć lub zwiększyć na przykład dokonując przeważenia na podstawie liczebności przedstawicieli

poszczególnych pokoleń w populacji dorosłych Polaków. Do tych właśnie celów służą testy *adhoc* zawarte w *Analyze* ⇨ *Compare Means* ⇨ *One Way ANOVA* ⇨ *Contrasts.*



Agregacji pokolenia średniego i pokolenia starszego dokonujemy następująco: w pole *Coefficients* wpisujemy kolejno liczby (tzw. wagi), po wprowadzeniu każdej z nich klikając *Add*. Będą się one pojawiały w polu poniżej, jak na załączonym rysunku. Kolejność tych liczb musi odpowiadać kolejności wartości zmiennej niezależnej (grup wiekowych). Zatem jeśli jako pierwsze w zmiennej niezależnej jest młodsze pokolenie (oznaczone liczbą jeden), wówczas waga podana w pierwszym wierszu będzie dotyczyć właśnie młodszego pokolenia. Drugie miejsce odnosić będzie się do średniego pokolenia, a trzecie do pokolenia starszego. Przypuśćmy, że chcemy zagregować średnie i starsze pokolenie i przeciwstawić je pokoleniu młodszemu, nadając jednocześnie nowoutworzonym dwóm grupom równe wagi. Wówczas w pierwszej linii umieszczamy liczbę 2, a w dwóch kolejnych wartości -1 (jak na rysunku).



Nadawanie wag wymaga przestrzegania następujących zasad:

1/ suma wag musi być równa zero (na analizowanym przykładzie: 2 + (-1) + (-1) = 0),

2/ grupy przeciwstawne oznaczane są przeciwnymi znakami (plus jest domyślny, natomiast minusa należy używać), wartości posiadające współczynniki dodatnie tworzą jedną grupę, a wartości posiadające współczynniki ujemne - drugą grupę,

3/ liczba wartości wpisana w pole *Coefficients* nie może przekraczać wartości zmiennej niezależnej.

W efekcie tych analiz otrzymujemy raporty tabelaryczne charakterystyczne dla jednoczynnikowej analizy wariancji ANOVA: test homogeniczności wariancji i tabelę z wynikiem analizy ANOVA. Interpretujemy je analogicznie do wyłożonych w niniejszym rozdziale zasad. Dodatkowo uzyskujemy tabele: *Contrast Coefficients* oraz *Contrast Tests*. Pierwsza z tabel zawiera dane podsumowujące dla przeprowadzonego testu: informuje o agregacjach, które przeprowadziliśmy oraz wagach, które nadaliśmy poszczególnym grupom. Druga tabela informuje nas o wynikach przeprowadzonego testu w zagregowanych grupach.

Jeśli wyniki testu Levene'a wykazały, że wariancje są homogeniczne wówczas odczytujemy wynik z pierwszego wiersza tej tabeli. Jeśli nie wykazały - wartości odczytujemy z wiersza drugiego. Wynik testu Levene'a w analizowanym przykładzie wynosi 0,92, a zatem stwierdzamy, że wariancja w badanych grupach jest nieistotna statystycznie. W konsekwencji z pierwszego wiersza tabeli odczytujemy wynik testu t:

$$t(117) = 4{,}19;\ p \leq 0{,}001$$

Interpretujemy zapis następująco: młode pokolenie różni się istotnie od pokoleń starszych (średniego i starszego).

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Significance
Poparcie dla demokracji w skali od 0 do 10	,08	2	117	,92

ANOVA

		Sum of Squares	df	Mean Square	F	Significance
Poparcie dla demokracji w skali od 0 do 10	Between Groups	160,87	2	80,43	22,08	,00
	Within Groups	426,30	117	3,64		
	Total	587,17	119			

Contrast Coefficients

		Wiek respondenta (w podziale na pokolenia)		
		młode pokolenie (18-35 lat)	średnie pokolenia (powyżej 35 do 65 lat)	powyżej 65 lat
Contrast	1	2	-1	-1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Poparcie dla demokracji w skali od 0 do 10	Assume equal variances	1	-3,10	,74	4,19	117	,00
	Does not assume equal	1	-3,10	,72	-4,33	85,12	2,00

Część VI. Odnajdywanie ładu w zbiorach danych

18

Rozdział 18. Poszukiwanie zmiennych ukrytych - analiza czynnikowa

Analiza czynnikowa (inaczej: eksploracyjna) jest metodą statystyczną powstałą na potrzeby psychologii. Jej wynalazcą jest Charles Spearman, a rozwinął ją Raymond Cattell, amerykański psycholog, pionier w dziedzinie badań nad osobowością. Pierwotnie wykorzystywana była w psychometrii - w badaniach nad inteligencją. Wydaje się ona obiecującą metodą również na gruncie nauk politycznych. Analiza czynnikowa jest techniką redukcji danych, podobnie jak najbliższe jej metody - analiza korespondencji i analiza skupień. W metodach tego typu doprowadza się do świadomego uporządkowania danych, co skutkuje zwiększeniem klarowności wyników i pozwala na wydobycie niewidocznych dotąd związków między danymi. Analiza czynnikowa opiera się na korelacji R Pearsona oraz analizie wariancji (zmienności). Za pomocą tego współczynnika mierzone są związki pomiędzy poszczególnymi zmiennymi, a zmienne są na tej podstawie porządkowane w grupy. Owe grupy nazywane są czynnikami lub - rzadziej - składowymi. Ich wyodrębnienie jest podstawowym celem analizy czynnikowej.

Analiza czynnikowa znajduje liczne zastosowania. Po pierwsze, używana jest w celach eksploracyjnych. Umożliwia ona rozpoznanie struktury zbioru danych, odnalezienie w chaosie różnorodnych zmiennych niedostrzegalnego na pierwszy rzut oka ładu. Po drugie, dzięki analizie czynnikowej można zmierzyć homogeniczność (jednorodność) skali. Pozwala ona wyodrębnić podskale lub podindeksy, to znaczy dowiedzieć się, czy interesujące badacza zjawisko jest jednorodne, czy też posiada wiele różnych wymiarów. Ponadto, wskazuje, które zmienne nie pasują do indeksu lub skali i należy je usunąć. To zastosowanie pozwala nam więc na konstruowanie narzędzi pomiarowych - kwestionariuszy zawierających indeksy i skale. W tym sensie analiza czynnikowa stanowi bardziej zaawansowany matematycznie zamiennik analizy rzetelności. Po trzecie, analiza czynnikowa jest przydatna wszędzie tam, gdzie badacz nie zakłada *a priori* modelu badanego zjawiska. Metoda ta pozwala na poznanie istoty związków pomiędzy badanymi danymi, bez uprzedniej wiedzy i koncepcji, jak struktura tych związków mogłaby wyglądać.

Po czwarte analiza czynnikowa może być wstępem do zastosowania odkrytych czynników w dalszych analizach wielowymiarowych[1].

18.1. Typy analizy czynnikowej

Analiza czynnikowa obejmuje kilka (na pierwszy rzut oka nierozróżnialnych) metod statystycznych. Najczęściej stosowanymi są właściwa analiza głównych składowych (*Principal Components Analysis*) oraz analiza głównych osi wymiarów (*Principal Axis Factoring*). Obie metody są względnie tożsame. Są one również dostępne w programie PSPP. Różnice między nimi mają charakter matematyczny. Analiza głównych osi wymiarów obliczana jest na macierzy korelacji, gdzie na przekątnej znajdują się wartości własne każdego czynnika, nie zaś korelacje zmiennej z nią samą (jak w przypadku analizy głównych składowych)[2]. Metody te dają bardzo zbliżone wyniki w sytuacji, gdy liczba badanych zmiennych jest duża[3]. Wskazuje się, że badacze niezbyt konsekwentnie stosują obie odmiany analizy czynnikowej[4]. Częściej stosuje się tę pierwszą, choć druga z metod jest ceniona przez polskich statystyków, bowiem z matematycznego punktu widzenia wydaje się lepiej uzasadniona. Istnieją również inne metody: najmniejszych kwadratów (*Unweighted Least Squares*), uogólnionych najmniejszych kwadratów (*Generalized Least Squares, GLS*), największej wiarygodności (*Maximum Likelihood, ML*) czy trójkątnej dekompozycji (*Triangular Decomposition*) lub alfa (*Alpha Factoring*).

Z punktu widzenia taksonomii funkcjonalnej metody analizy czynnikowej należy rozróżnić na **eksploracyjną** analizę czynnikową i **konfirmacyjną** analizę czynnikową. Pierwszą z nich stosujemy, gdy budujemy model i nie znamy wzajemnych zależności pomiędzy badanymi zmiennymi. Analiza konfirmacyjna jest stosowana wówczas, gdy posługujemy się istniejącymi teoriami na temat konstrukcji danego zjawiska.

18.2. Warunki wykonania analizy czynnikowej

Zbiór danych, który chcemy poddać analizie czynnikowej powinien spełniać warunek wystarczającej liczebności zmiennych oraz poziomu ich pomiaru. Odnośnie liczby jednostek analizy należy wskazać na dwa następujące wymogi: po pierwsze, powinna być ona wyższa od 30, po drugie - co najmniej półtorakrotnie przekraczać liczbę pytań w danej skali lub indeksie. Niektórzy uważają, że najlepsze wyniki uzyskuje się, jeśli liczba jednostek analiz jest co najmniej dziesięciokrotnie większa niż liczba zmiennych. Nie bez znaczenia jest również liczba czynników - uważa się, że na każdy wyodrębniony czynnik powinno przypadać co najmniej 10 jednostek analizy. Zmienne składające się na skale lub indeksy powinny być mierzone na poziomie porządkowym lub - najlepiej - interwałowym bądź ilorazowym. Istotne jest także,

[1] G. Wieczorkowska, G. Król, *O typowym zastosowaniu analizy czynnikowej i skalowania w badaniach społecznych*, „Zeszyty Metodologiczne", 1997, 1, s. 3-5.

[2] S. Bedyńska, *Clooney? Brzydal! Di Caprio? Brzydal! Analiza czynnikowa - metody wyodrębniania czynników*, w: http://www.predictivesolutions.pl/EKSPRESS/Analiza_danych_w_dzialaniu/Techniki_analityczne/Analiza_czynnikowa/Clooney_Brzydal_Di_Caprio_Brzydal_Analiza_czynnikowa_metody_wyodrebniania_czynnikow_cz2_SBedynska.pdf, dostęp: lipiec 2012, s. 2.

[3] B. Thompson, S.A. Vidal-Brown, *Principal Components versus Principle Axis Factors: When Will We Ever Learn?*, referat zaprezentowany na Annual Meeting of the Southwest Educational Research Association, Nowy Orlean 2001.

[4] *Statystyczny drogowskaz. Praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii*, S. Bedyńska, A. Brzezicka (red.), Wydawnictwo „Academica", Warszawa 2007, s. 136.

aby zmienne były wyrażone na podobnych skalach, o jednolitych kierunkach[5]. Należy również odpowiednio zdefiniować braki danych.

18.3. Etapy przeprowadzania analizy czynnikowej

Wykonanie analizy czynnikowej jest zadaniem złożonym. Poniżej prezentowane są cztery etapy jej przeprowadzenia: ocena zmiennych, na których chcemy wykonać analizę czynnikową, zidentyfikowanie optymalnej liczby czynników, dopasowanie czynników i ich jakościowa analiza.

1/ Sprawdzenie właściwości zmiennych poddanych analizie czynnikowej. Ocena zmiennych polega na stwierdzeniu występowania współzależności między nimi. Jest to nieodzowna właściwość zmiennych, będąca jednocześnie podstawowym warunkiem umożliwiającym przeprowadzenie analizy czynnikowej. Celem analizy czynnikowej jest bowiem zredukowanie wybranego skorelowanego zbioru zmiennych (określanego mianem zbioru zmiennych pierwotnych) do postaci odrębnych i nieskorelowanych grup (czynników).

W pierwszej kolejności należy sprawdzić poziom zróżnicowania zmiennych, przede wszystkim wartość odchylenia standardowego. Należy zadbać, by statystyka ta była jak najmniej zróżnicowana w ramach każdej zmiennej, a ponadto, by żadna ze zmiennych nie miała odchylenia równego zero. Jeśli odchylenie standardowe przyjmie wartość zero, wówczas analiza czynnikowa nie zostanie w programie PSPP wykonana. Po drugie, konieczne jest wykonanie macierzy korelacji i określenie, czy zmienne są w wystarczającym stopniu ze sobą skorelowane. Wykorzystuje się do tego celu trzy sposoby: wyznacznik macierzy korelacji, miarę adekwatności doboru próby KMO oraz test sferyczności Bartletta. Wyznacznik powinien utrzymywać się na jak najniższym poziomie (być bliski zeru), co oznacza, że zmienne ze sobą korelują. Nazwa KMO to skrót utworzony z pierwszych liter nazwisk jej twórców: Henry'ego F. Kaisera, Kennetha Meyera oraz Ingrama Olkina. Wynik tego testu przyjmuje wartości od 0 do 1. Wykonanie analizy czynnikowej możliwe jest, gdy jego wartość przekracza 0,5. Na wynik testu ma wpływ również liczba badanych (jednostek analizy) - im jest ich mniej, tym niższego wyniku KMO można się spodziewać. Alternatywną miarą dla KMO jest test sferyczności Bartletta. Jego nazwa pochodzi od nazwiska jego twórcy - brytyjskiego statystyka Maurice'a S. Bartletta. Test ten oparty jest na statystyce rozkładu chi-kwadrat. Weryfikuje on hipotezę zerową o braku występowania współzależności między zmiennymi. Odrzucenie hipotezy zerowej dokonuje się na podstawie wyników testu istotności. Jeżeli poziom istotności jest mniejszy niż przyjęty 0,05, należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną mówiącą o istnieniu korelacji między zmiennymi. Wyłącznie taki wynik testu jest dla nas satysfakcjonujący i umożliwia przeprowadzenie analizy czynnikowej.

2/ Wyodrębnienie liczby czynników na podstawie kryterium Kaisera lub wykresu osypiska. Na tym etapie zmienne poddane analizie czynnikowej dzielone są na grupy. Są one tworzone przez

[5] Niekiedy badacze włączają do analizy różne typy zmiennych, począwszy od konstruowanych na klasycznych skalach (skali Rensisa A. Likerta, dyferencjale semantycznym, skalach punktowych – na przykład od 0 do 10 bądź od 0 do 100), aż do zmiennych o charakterze ciągłym – wiek, waga, dochód. Choć z punktu widzenia teoretycznego nie ma przeciwwskazań włączania do analizy różnych typów zmiennych, to jednak w praktyce analitycznej przeprowadza się analizę czynnikową na porównywalnych zmiennych, mierzonych w ten sam sposób. Gdy zachodzi potrzeba badania zmiennych mierzonych na różnych typach skal, zalecane jest zastosowanie odpowiednich procedur standaryzacji bądź normalizacji zmiennych

najsilniej skorelowane ze sobą zmienne. Grupy te nazywane są **czynnikami** lub - rzadziej - **składowymi**. Badacz może wybierać spośród trzech następujących kryteriów wyodrębniania czynników:

a/ W programie PSPP czynniki można wyodrębnić na podstawie kryterium opracowanego przez Henry'ego F. Kaisera[6] (w skrócie nazywanym kryterium Kaisera lub niekiedy kryterium Guttmana-Kaisera[7]). Do tego celu wykorzystywana jest statystyka o nazwie wartość własna, która określa część wariancji wyjaśnionej przez każdą z grup zmiennych (czynników). Istotnymi czynnikami są tylko te, których wartość własna jest wystarczająco wysoka. Według kryterium Kaisera za istotne czynniki uznawane są te grupy, które uzyskały wartość własną większą niż 1. Czynniki takie wyjaśniają dużą część łącznej wyjaśnianej wariancji zmiennych pierwotnych.

b/ Alternatywną metodą dla kryterium Kaisera jest **wykres osypiska** (nazywany też wykresem kolankowym), którego koncepcja została opracowana przez Raymonda B. Cattella[8]. Nazwa dla tej statystyki nawiązuje do pojęcia „osypisko", funkcjonującego w geologii i służącego do określenia usypanej powierzchni gruzu gromadzącej się w dolnych partiach urwisk skalnych. Kryterium to nie posiada jednoznacznej, ilościowej interpretacji. Ocena ma w tym przypadku charakter czysto obserwacyjny, jakościowy - badacz odrzuca te czynniki, które znalazły się poza „stromizną" wykresu, oddzielając informację użyteczną od szumu informacyjnego. Informacje użyteczne tłumaczą dużą część wariancji, co jest identyfikowane przez duże spadki krzywej wykresu osypiska. Natomiast szumy informacyjne wyjaśniają mały zakres całkowitej wariancji, o czym świadczy niskie nachylenie krzywej względem osi odciętych. Analiza wykresu osypiska wymaga zidentyfikowania punktu, od którego nie obserwujemy gwałtownych spadków krzywej. Ten punkt wyznacza jednocześnie liczbę istotnych czynników.

c/ Swoistym sposobem wyboru czynników jest analiza wielkości wyjaśnianej wariancji. Niektórzy badacze dobierają czynniki na tej podstawie, kierując się zasadą, by czynników było co najmniej tyle, aby wyjaśnić 50 proc. wariacji całkowitej.

Spośród trzech scharakteryzowanych kryteriów najczęściej stosowanym w praktyce badawczej jest kryterium Kaisera i wykres osypiska, natomiast analiza wielkości wyjaśnianej wariancji ma charakter wspomagający. Wśród badaczy brakuje jednoznacznego stanowiska uznającego, która z metod jest najwłaściwsza. Kryterium Kaisera umożliwia wyznaczenie większej liczby czynników niż metoda wykresu osypiska. Utrudnia to niekiedy ich interpretację. Mniejsza liczba czynników wyróżnianych za pomocą wykresu osypiska powoduje z kolei zmniejszenie się poziomu wyjaśnianej wariancji. Większość analityków korzysta z kryterium Kaisera ze względu na precyzję - obliczeniowy sposób wyznaczania czynników.

3/ Maksymalizacja dopasowania i koncepcyjna analiza wyodrębnionych czynników. Na tym etapie wyodrębniamy i analizujemy zmienne tworzące poszczególne, wyróżniane w etapie drugim czynniki. Pojawia się tu pojęcie **ładunków**. Jest to poziom korelacji pomiędzy zmiennymi należącymi do danego czynnika. Tylko wyodrębnione czynniki biorą udział w tej procedurze. Etap ten składa się z dwóch części: 1/ uzyskania najlepszego dopasowania czynników, czyli jednoznacznego zakwalifikowania

[6] H.F. Kaiser, *The application of Electronic Computers of Factor Analysis,* „Educational and Psychological Measurement", 1960, 20, s. 141-151.

[7] K.A. Yeomans, P.A. Golder, *The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors*, „Statistician", 1982, 31 (3), s. 221.

[8] R.B. Cattell, *The scree test for the number of factors*, „Multivariate Behavioral Research", 1966, 1, s. 629-637.

zmiennej do niezależnej, nieskorelowanej z innymi czynnikami grupy oraz 2/ jakościowej syntezy czynników polegającej na umieszczeniu ich na szerszym tle poznawczym, teoretycznym.

W analizie czynnikowej do uzyskania najlepszego dopasowania czynników wykorzystujemy procedurę nazywaną **rotacją** wymiarów. Rotacja jest to czynność „obracania" układu współrzędnych (wymiarów), w taki sposób by badacz mógł odnaleźć związki między wieloma zmiennymi i pogrupować je w odpowiednie czynniki. Istnieją rozmaite sposoby rotacji czynników. Zostały one opracowane ze względu na rozbieżne poglądy na temat zasad upraszczania czynników. Wśród badaczy ścierają się dwa następujące oczekiwania: po pierwsze, uzyskania prostej w interpretacji struktury wyodrębnianych czynników, po drugie - pogrupowanie zmiennych w jednoznaczne grupy, mające silne związki z konkretnymi czynnikami. Opracowano dwie główne metody rotacji: ortogonalną (to jest zachowująca kąty proste podczas obracania osi współrzędnych) oraz nieortogonalną (ukośną, dostosowującą kąt obrotu układu współrzędnych do zbieżności z wieloma zmiennymi jednocześnie i ich krzyżowania się). W programie PSPP dostępne są następujące metody rotacji ortogonalnej: *Varimax*, *Equimax* i *Quartimax*. Ponadto istnieje szereg innych metod rotacji czynników, między innymi *Orthomax*, *Parsimax* i *Oblimin*. Najczęściej używanym sposobem rotacji jest *Varimax,* a w drugiej kolejności *quartimax.* Metoda *varimax* minimalizuje liczbę zmiennych, które mają wysokie ładunki czynnikowe. Dzięki temu łatwiej można wyodrębnić i zinterpretować czynnik. Z technicznego punktu widzenia metoda ta koncentruje się na upraszczaniu kolumn macierzy ładunków. Jest ona używana w badaniach, w których chcemy ujawnić, w jaki sposób grupowanie poszczególnych zmiennych mierzy tą samą koncepcję, np. typów osobowości. Z praktycznego punktu widzenia metoda ta upraszcza interpretację czynników dzięki minimalizacji liczby zmiennych koniecznych do wyjaśnienia danego czynnika. Używamy jej przede wszystkim, gdy chcemy uzyskać zmienne połączone w najbardziej niezależne od siebie grupy; może być ona bardzo pomocna przy tworzeniu typologii. Z kolei *Quartimax* minimalizuje liczbę czynników potrzebnych do wyjaśnienia danej zmiennej. Z technicznego punktu widzenia *Quartimax* koncentruje się na upraszczaniu wierszy macierzy ładunków czynnikowych. Dzięki tej metodzie rotacji możemy łatwiej interpretować zmienne w kontekście czynników (odwrotnie niż w przypadku *Varimax*). Metoda *Equimax* jest połączeniem obu powyższych. Ułatwia ona interpretację zarówno zmiennych jak i czynników. Należy pamiętać, iż wybór metody rotacji może istotnie zmienić wartość ładunku każdego z czynników. W praktyce możemy stosować różne metody rotacji, aż do uzyskania najbardziej klarownego podziału zmiennych na czynniki, a więc sekwencyjnie wybierając kolejne metody.

4/ Etap jakościowej syntezy wyników polega na nadaniu wyodrębnionym czynnikom adekwatnych nazw. Dokonujemy tego na podstawie informacji, jakie niosą ze sobą zmienne składające się na owe czynniki. Czynność ta ma charakter pozastatystyczny, czysto interpretacyjny i wynikać powinna z teoretycznego kontekstu prowadzonego badania.

18.4. Analiza czynnikowa w programie PSPP

Wykonanie analizy czynnikowej w programie PSPP rozpoczynamy od przeglądu bazy danych oraz wybrania odpowiednich zmiennych. Przypomnijmy, iż ich poziom pomiaru powinien być co najmniej porządkowy. Zasadne jest, aby zmienne były wyrażone na podobnych skalach o jednakowych kierunkach. Liczba jednostek analizy przypadająca na daną zmienną powinna być wystarczająco wysoka i wynosić więcej niż 30 oraz być co najmniej półtora razy większa niż liczba analizowanych zmiennych.

Zmienne powinny posiadać również zdefiniowane braki danych[9]. Wybór zmiennych nie powinien być przypadkowy. Praktyka badawcza podpowiada, aby wybierane zmienne tworzyły pewne logiczne powiązania. Pamiętać należy, że celem analizy czynnikowej jest ich pogrupowanie w czynniki, którym później musimy nadać odpowiednie i charakterystyczne nazwy[10]. Dlatego początkującemu badaczowi radzimy analizowanie zmiennych, które odnoszą się do podobnych zagadnień bądź problemów.

Analizę czynnikową w programie PSPP przeprowadzany w czterech krokach. Po pierwsze, sprawdzamy właściwości wybranych zmiennych, które chcemy poddać analizie czynnikowej. W drugim kroku dokonujemy wyodrębnienia liczby czynników za pomocą kryterium Kaisera bądź przy użyciu wykresu osypiska. Trzeci etap polega na odpowiednim dopasowaniu czynników oraz czwarty - na ich syntetycznej interpretacji jakościowej. Całość niniejszego rozdziału będziemy chcieli zakończyć omówieniem procedury wykonywania analizy czynnikowej w edytorze składni.

Do przeprowadzenia analizy czynnikowej w programie PSPP wykorzystamy zbiór danych pochodzący z badania Polskie Generalne Studium Wyborcze 2007. Dla przykładu, eksploracji poddamy dwadzieścia dwie zmienne określające intensywność słuchania danej stacji radiowej bądź czytania określonego czasopisma. W zakres analizowanych mediów wchodzą: 1/ Polskie Radio Program 1 (t98a), 2/ Polskie Radio Program 2 (t98b), 3/ Polskie Radio Program 3 (t98c), 4/ RMF FM (t98d), 5/ Radio ZET (t98e), 6/ TOK FM (t98f), 7/ Radio Maryja (t98g), 8/ Gazeta Wyborcza (t99a), 9/ Rzeczpospolita (t99b), 10/ Dziennik (t99c), 11/ Fakt (t99d), 12/ Nasz Dziennik (t99e), 13/ Trybuna (t99f), 14/ Super Express (t99g), 15/ Polityka (t100a), 16/ Wprost (t100b), 17/ Newsweek (t100c), 18/ Nie (t100d), 19/ Przekrój (t100e), 20/ Gazeta Polska (t100f), 21/ Tygodnik Powszechny (t100g) oraz 22/ Niedziela (t100h). Respondentom zadano pytanie o następującej treści: „Ile godzin spędzasz na słuchaniu odpowiedniej stacji radiowej/ czytaniu gazety/ czytaniu czasopisma ... ?". Odpowiedzi udzielano na pięciostopniowej skali, gdzie 1 oznaczało *nigdy nie czytam/słucham*, 2 - *mniej niż jedną godzinę w tygodniu*, 3 - *jedną-dwie godziny w tygodniu*, 4 - *dwie do trzech godzin w tygodniu* oraz 5 - *więcej niż trzy godziny w tygodniu*.

[9] Wykluczenia odpowiedzi w programie PSPP dokonujemy w dwojaki sposób. Po pierwsze, definiujemy braki danych w zakładce *Variable View* w kolumnie *Missing*, gdzie wprowadzamy zakres wyłączanych odpowiedzi. Do tego celu można również wykorzystać polecenie w edytorze składni o postaci MISSING VALUES 'nazwa zmiennej' ('wartości wykluczanych odpowiedzi'). Najbezpieczniejszą metodą wykluczania braków danych jest odpowiednie przekształcenie zmiennych za pomocą opcji rekodowania na tą samą zmienną (*Recode into Same Variables*) lub na inną zmienną (*Recode into Diferrent Variables*). Wybrane wartości zmiennych przekształcamy na systemowe braki danych (*System Missing*). Dzięki temu zabiegowi mamy pewność, iż niepożądany zakres wartości zmiennych nie zostanie uwzględniony w dalszych analizach.

[10] W praktyce badawczej funkcjonuje dość powszechna tendencja do obejmowania analizą czynnikową kilkudziesięciu bądź nawet kilkuset zmiennych, niekiedy wszystkich (posiadających odpowiednie właściwości) znajdujących się w kwestionariuszu badania. Stwarza to możliwość odkrycia niewidocznych prawidłowości między cechami, które jedynie pozornie nie wykazują zależności. Należy jednak wystrzegać się tego sposobu badania niedostrzegalnych struktur w zbiorach danych. Konstruowany model czynnikowy jest bowiem „wrażliwy" na obecność każdej pojedynczej zmiennej i jej związków z pozostałymi zmiennymi. Oznacza to, iż włączenie bądź wyłączenie jakiejkolwiek zmiennej z pierwotnego zakresu rzutuje na całą strukturę wyodrębnianych czynników. Dlatego sugerujemy, aby badacz uważnie dobierał zmienne w analizie czynnikowej.

18.4.1. Sprawdzenie właściwości zmiennych poddanych analizie czynnikowej

Pierwszym etapem analizy czynnikowej jest ocena właściwości eksplorowanych zmiennych. Sprawdzenia tego aspektu dokonujemy dwoma sposobami. Po pierwsze sprawdzamy poziom zróżnicowania zmiennych na podstawie analizy ich odchylenia standardowego. Po drugie, obliczamy odpowiednie statystyki, stwierdzające występowanie zależności między zmiennymi. Do tego celu wykorzystujemy miarę KMO, test sferyczności Bartletta, macierz korelacji oraz właściwą jej statystkę - wyznacznik macierzy korelacji. Wykonanie tychże czynności prezentują niniejsze podrozdziały.

18.4.1.1. Sprawdzenie poziomu zróżnicowania zmiennych za pomocą odchylenia standardowego

W programie PSPP odchylenie standardowe dla zmiennych poddawanych analizie czynnikowej obliczamy dwoma sposobami. Po pierwsze, możemy tą miarę obliczyć wybierając z zakładki *Analyze* ⇨ *Descriptive Statistics* ⇨ *Frequencies* lub *Descriptives,* następnie wybrać z listy interesujące nas zmienne i w polu *Statistics* zaznaczyć odchylenie standardowe (*Standard deviation*). Drugi sposób polega na skorzystaniu z opcji obliczenia statystyk opisowych dedykowanych wyłącznie analizie czynnikowej. W tym celu wykorzystujemy odpowiednie polecenie w edytorze składni o następującej postaci:

Składnia do wpisania w Edytorze	Opis działania składni
`FACTOR`	- wykonaj analizę czynnikową
`/ VARIABLES = t98a t98b t98c t98d t98e t98f t98g t98a t99a t99b t99c t99d t99e t99f t99g t100a t100b t100c t100d t100e t100f t100g t100h`	- dla wybranych zmiennych (po poleceniu VARIABLE wpisujemy wszystkie nazwy zmiennych ze zbioru danych)
`/ MISSING =LISTWISE`	- dla braków danych wybierz opcję *wyłącz wszystkie obserwacje z brakami danych.*
`/ PRINT = UNIVARIATE.`	- w oknie raportu pokaż obliczone statystki opisowe - średnią arytmetyczną, odchylenie standardowe oraz liczebności - dla wybranego zakresu zmiennych
`EXECUTE.`	- wykonaj obliczenia.

Rezultatem skorzystania z tego polecenia jest poniższa tabela:

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.	1.68	1.18	1686
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.	1.20	.55	1686
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.	1.40	.90	1686
Ile godzin w ciągu tygodnia słucha: RMF FM.	2.19	1.40	1686
Ile godzin w ciągu tygodnia słucha: Radio ZET.	1.99	1.32	1686
Ile godzin w ciągu tygodnia słucha: TOK FM.	1.16	.57	1686
Ile godzin w ciągu tygodnia słucha: Radio Maryja.	1.29	.87	1686
Ile godzin dziennie spędza na czytaniu gazet: Gazeta Wyborcza.	1.51	.93	1686
Ile godzin dziennie spędza na czytaniu gazet: Rzeczpospolita.	1.22	.62	1686
Ile godzin dziennie spędza na czytaniu gazet: Dziennik.	1.39	.83	1686
Ile godzin dziennie spędza na czytaniu gazet: Fakt.	1.48	.91	1686
Ile godzin dziennie spędza na czytaniu gazet: Nasz Dziennik.	1.09	.40	1686
Ile godzin dziennie spędza na czytaniu gazet: Trybuna.	1.05	.28	1686
Ile godzin dziennie spędza na czytaniu gazet: Superexspress.	1.27	.71	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Polityka.	1.23	.64	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Wprost.	1.21	.58	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Newsweek.	1.27	.67	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Nie.	1.07	.39	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Przekrój.	1.10	.43	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Gazeta Polska.	1.05	.31	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Tygodnik Powszechny.	1.05	.27	1686
Ile godzin w tygodniu spędza na czytaniu czasopism: Niedziela	1.08	.40	1686

Zawiera ona obliczone statystyki dla każdej zmiennej poddawanej analizie czynnikowej - średnią arytmetyczną, odchylenie standardowe oraz liczbę jednostek analizy. W pierwszej kolejności sprawdzamy, czy któraś ze zmiennych nie ma odchylenia standardowego równego zero. Jeżeli którakolwiek ze zmiennych posiadałaby odchylenie wynoszące zero, należałoby ją wykluczyć z dalszych analiz. Jej obecność uniemożliwiłaby przeprowadzenie analizy czynnikowej w programie PSPP. W naszym przykładzie wszystkie zmienne posiadają odchylenie standardowe większe niż zero. Żadna ze zmiennych nie wykazuje również zbyt wysokiego zróżnicowania ze względu na tą miarę. Został zatem spełniony pierwszy warunek umożliwiający przeprowadzenie analizy czynnikowej.

18.4.1.2. Ocena właściwości zmiennych za pomocą miary KMO, testu sferyczności Bartletta, wyznacznika macierzy korelacji

W kolejnym etapie dokonujemy oceny zestawu zmiennych za pomocą odpowiednich testów oraz miar statystycznych. W programie PSPP dostępne są cztery sposoby sprawdzania właściwości zmiennych. Należą do nich macierz korelacji (*Correlation Matrix*), wyznacznik macierzy korelacji (*Determinant of Correlation Matrix*), miara KMO oraz test sferyczności Bartletta. Przeprowadzenie tych obliczeń wymaga skorzystania z edytora składni. Na obecną chwilę ich wykonanie w programie PSPP nie jest możliwe w trybie okienkowym.

Korzystanie z edytora wymaga wprowadzenia składni o następującej postaci:

Składnia do wpisania w Edytorze	Opis działania składni
FACTOR	- wykonaj analizę czynnikową
/ VARIABLES = t98a t98b t98c t98d t98e t98f t98g t98a t99a t99b t99c t99d t99e t99f t99g t100a t100b t100c t100d t100e t100f t100g t100h	- dla wybranych zmiennych (po poleceniu VARIABLE wpisujemy wszystkie nazwy zmiennych ze zbioru danych)
/ MISSING =LISTWISE	- dla braków danych wybierz opcję *wyłącz wszystkie obserwacje z brakami danych*
/ PRINT = KMO DET CORRELATION.	- w oknie raportu pokaż obliczenia dla miary KMO oraz wyniki testu sferyczności Bartletta (identyfikowane jest to za pomocą polecenia 'KMO'), a także oblicz wyznacznik macierzy korelacji (czemu właściwa jest formuła 'DET' w poleceniu) oraz stwórz macierz korelacji (te statystyki są właściwe komendzie 'CORRELATION')
EXECUTE.	- wykonaj obliczenia.

Rezultatem wykonania tej czynności są dwie tabele publikowane w oknie raportu. Pierwsza z nich, zatytułowana *KMO and Bartlett's Test*, zawiera statystyki - obliczoną miarę KMO oraz test sferyczności Bartletta. W analizowanym przykładzie obliczenia dla tych miar prezentują się następująco:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.73
Bartlett's Test of Sphericity	*Approx. Chi-Square*	4358.01
	df	153
	Sig.	.00

Miarę KMO odnajdujemy w pierwszym wierszu o nazwie *Kaiser-Meyer-Olkin Measure of Sampling Adequacy.* Statystyka ta określa adekwatność doboru zmiennych do analizy czynnikowej. Jak pamiętamy z wcześniejszej części rozdziału, miara ta przyjmuje wartości z przedziału od 0 do 1. Za dolną akceptowalną granicę dla wyników tego testu przyjmuje się 0,5, co jednocześnie uprawnia do przeprowadzenia analizy czynnikowej. Jeżeli obliczona miara KMO wynosi mniej niż 0,5, wtedy dokonywanie analizy czynnikowej nie jest właściwe. W naszym przykładzie wynosi ona 0,73. Oznacza to, iż prawidłowo dobraliśmy wyboru zakres zmiennych i możemy kontynuować procedurę wyodrębniania czynników.

Test sferyczności Bartletta jest następną statystyką wykorzystywaną do oceny właściwości zmiennych. Wyniki tego testu znajdują się w trzech kolejnych wierszach tabeli, zaczynając od wiersza zatytułowanego *Bartlett's Test of Sphericity.* Obliczenia tego testu dokonywane są na podstawie statystyki testu chi-kwadrat. Służy on do zweryfikowania hipotezy zerowej mówiącej o występowaniu jednostkowej macierzy korelacji. Potwierdzenie tego przypuszczenia byłoby równoznaczne z brakiem współzależności między analizowanymi zmiennymi, czyli niemożności spełnieniu podstawowego warunku analizy czynnikowej. Podobnie jak w pozostałych testach statystycznych, kluczową miarą jest poziom istotności (p). Jeżeli jest on mniejszy niż przyjęty poziom ufności α arbitralnie ustalany na poziomie 0,05, to odrzucamy hipotezę zerową. Stwierdzamy, iż w testowanym zbiorze danych istnieje przynajmniej jeden wspólny czynnik dla badanych zmiennych. Analogicznie, jeżeli $p \geq 0{,}05$, należałoby przyjąć hipotezę zerową mówiącą o braku związku między zmiennymi, czego konsekwencją byłby brak możliwości

przeprowadzenia analizy czynnikowej. W naszym przykładzie wynik testu Bartletta jest istotny statystycznie ($p<0,05$), zatem dla badanego zbioru danych można wyodrębnić wspólne czynniki. Umożliwia to kontynuowanie analizy czynnikowej.

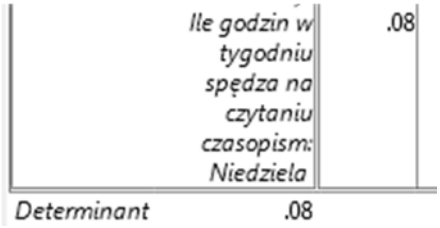
Macierz korelacji oraz przypisana jej miara - **wyznacznik macierzy korelacji** (*Determinant of Correlation Matrix*) są dodatkowymi statystykami weryfikującymi właściwości badanych zmiennych. Wyniki tych obliczeń służą do oceny właściwości całościowego zbioru danych i możliwości wykonania dla nich analizy czynnikowej. Dobrą praktyką jest dokładne przyjrzenie się zawartości macierzy korelacji w celu zidentyfikowania zmiennych słabo skorelowanych z pozostałymi włączonymi do zbioru zmiennymi. Macierz korelacji jest syntetyczną analizą zależności opierającą się na wyliczeniu współczynnika R Pearsona dla każdej pary zmiennych. Wyniki dla tych obliczeń prezentowane są w zbiorczej tabeli krzyżowej, gdzie główną przekątną tworzą wartości korelacji poszczególnej zmiennej z nią samą (dlatego zawsze posiadają one wartość 1), zaś odpowiednio w wierszach jak i kolumnach tabeli - wartości korelacji danej zmiennej z pozostałymi zmiennymi. Macierz korelacji zazwyczaj zawiera dużą liczbę zmiennych oraz bogaty zbiór danych. Poniżej prezentujemy fragment tejże macierzy:

Correlation Matrix

		Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.	*Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.*	*Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.*	*Ile godzin w ciągu tygodnia słucha: RMF FM.*	*Ile godzin w ciągu tygodnia słucha: Radio ZET.*
Correlations	*Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.*	1.00	.37	.14	-.12	-.10
	Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.	.37	1.00	.39	.06	.08
	Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.	.14	.39	1.00	.06	.04
	Ile godzin w	-.12	.06	.06	1.00	.45

Analizowanie oraz prezentowanie pełnej macierzy korelacji jest mimo wszystko zabiegiem dość trudnym. Jednak za pomocą wyznacznika macierzy korelacji możemy dokonać wstępnej ewaluacji współzależności zmiennych. Przyjmuje on wartości z przedziału od 0 do 1, przy czym 1 oznacza brak jakiejkolwiek korelacji między zmiennymi. Pozytywny wynik dla tej miary powinien być jak najniższy i bliski wartości zero. Jest to sygnał, iż w zbiorze danych znajduje się wiele istotnych korelacji, co potwierdza możliwość przeprowadzenia analizy czynnikowej. Im wartość wyznacznika macierzy korelacji jest większa i w znacznym stopniu bliska wartości 1, tym zmienne są słabiej skorelowane, zaś procedura wyodrębniania czynników może zakończyć się niepowodzeniem.

Wartość tej miary podawana jest pod macierzą korelacji, co prezentuje poniższy zrzut ekranowy:

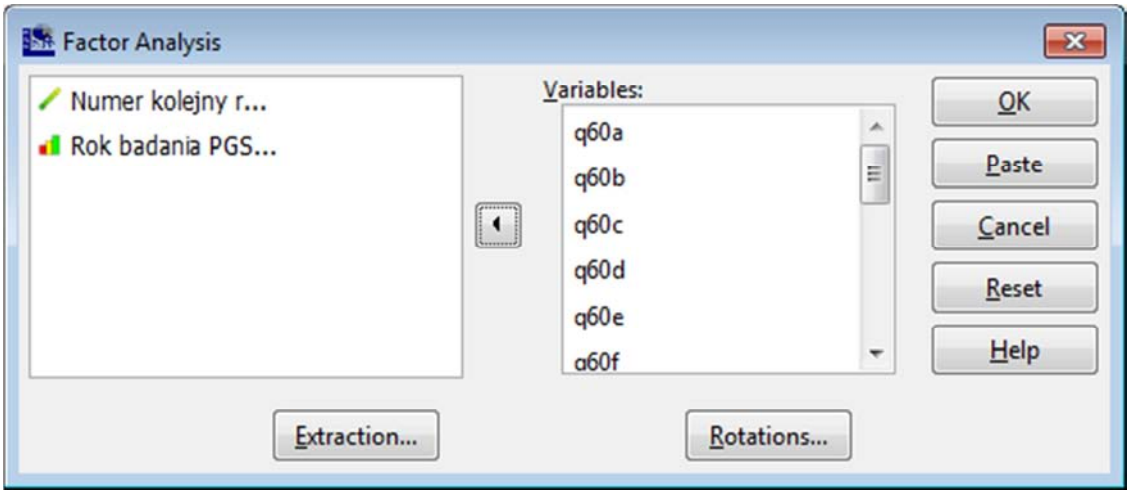


W analizowanym przykładzie wartość wyznacznika macierzy korelacji (*Determinant*) wynosi 0,08. Oznacza to, iż między dwudziestoma dwiema zmiennymi z badanego zbioru istnieje silna współzależność. Można zatem przypuszczać, iż zakres zmiennych pierwotnych jest adekwatnie dobrany i można go zredukować do szerszych kategorii czynnikowych.

Statystyczna ocena właściwości zbioru danych wykazała, iż między testowanymi zmiennymi istnieją współzależności. Sprawdzenie tychże założeń uprawnia do kontynuowania analizy czynnikowej i przejścia do etapu zasadniczego, polegającego na wyodrębnianiu odpowiednich czynników. Omówienie tej procedury Czytelnik odnajdzie w kolejnym podrozdziale niniejszej części podręcznika.

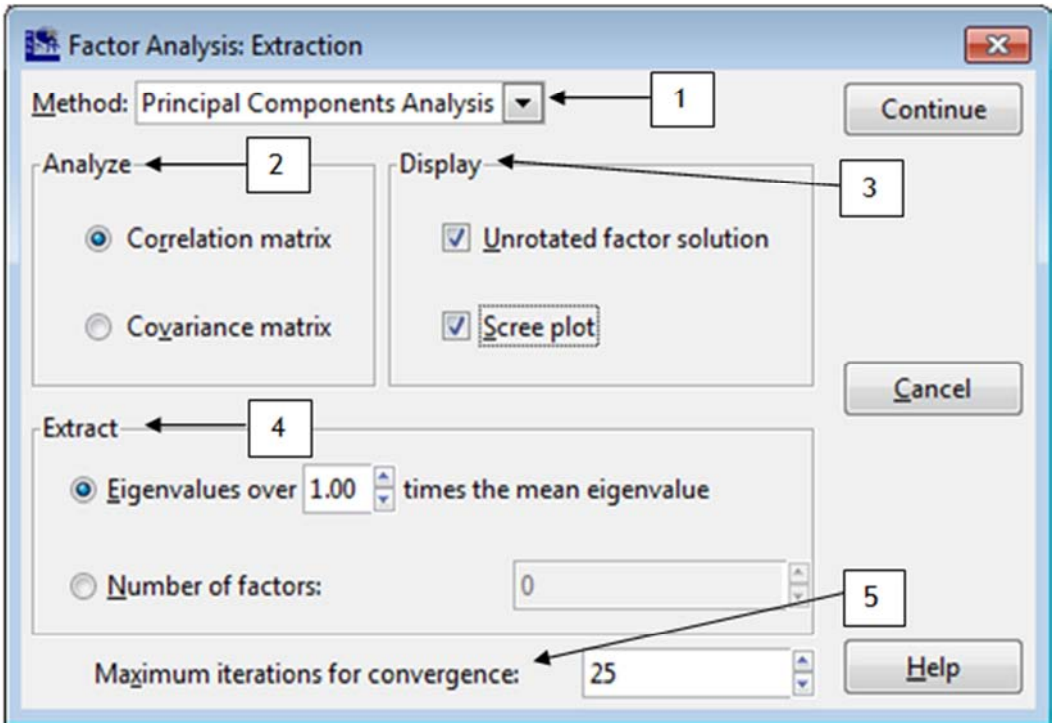
18.4.2. Wyodrębnianie liczby czynników na podstawie kryterium Kaisera lub wykresu osypiska

Analizę czynnikową w programie PSPP wykonujemy, wybierając w menu tekstowym opcję *Analyze* ⇨ *Factor Analysis*, efektem czego jest pojawienie się następującego okna:



Po lewej stronie okna odnajdujemy zbiór zmiennych, który po należy wybrać i przenieść do sąsiedniego pola o nazwie *Variables.* U dołu okna odnajdziemy dwie możliwości - wyodrębnianie (*Extraction...*) oraz rotacje (*Rotations...*). Są to odpowiednie opcje statystyczne, które należy ustawić w celu przeprowadzenia analizy czynnikowej. Następnie ustalamy odpowiednie warunki dla opcji *Extraction*.

Jej wybranie powoduje pojawienie się nowego okna, który prezentuje poniższy zrzut ekranowy:



Zamieszczone powyższej numery pól wyznaczają kolejne kroki analizy czynnikowej (służą one określeniu wstępnej liczby czynników, do których zostanie zredukowany zbiór zmiennych pierwotnych).

W polu o nazwie *Method* (oznaczonym numerem 1) z rozwijanej listy wybieramy odpowiednią metodę wyodrębniania czynników. W programie PSPP mamy do wyboru dwie opcje. Pierwsza z nich to metoda głównych składowych (*Principal Components Analysis, PCA*). W programie PSPP jest ona ustawiona jako domyślna. Drugim rozwiązaniem udostępnionym przez twórców programu jest metoda głównych osi wymiarów (*Principal Axis Factoring, PAF*). W analizowanym przykładzie w procesie wyodrębniania czynników wykorzystamy pierwszą metodę, czyli metodę głównych składowych. Jest to najprostszy i najmniej czasochłonny sposób zmiennych wyodrębniania czynników. Metoda ta pozwala bowiem na przekształcenie całej zawartości danych i dokonanie wstępnego rozeznania w liczbie składowych.

W kolejnym kroku dokonujemy wyboru odpowiedniego typu analizy, rozpatrując dwie możliwości: macierz kowariancji (*Covariance Matrix*) bądź też macierz korelacji (*Correlation Matrix*). W polu o nazwie *Analyze* jako domyślna jest ustawiona opcja macierzy kowariancji. Jak można zaobserwować na powyższym rysunku, została ona zmieniona na macierz korelacji. Praktyka badawcza podpowiada, aby procedurę wydobywania czynników przeprowadzać za pomocą tejże statystyki.

Trzeci etap wymaga odpowiednich ustawień w polu *Display* (numer 3), co w dosłownym tłumaczeniu oznacza „Pokaż". W tym miejscu określamy zakres oraz sposób prezentowania wyników w oknie raportu. Odnosi się on do dwóch aspektów. Pierwszy z nich wyznacza zakres prezentowanych analiz w tabeli dla całkowitej wariancji wyjaśnionej (*Total Variance Explained*). Omówienie tej tabeli oraz jej interpretację Czytelnik odnajdzie w dalszej części niniejszego podrozdziału. Należy pamiętać, że w podstawowym formacie zawiera ona oszacowanie tzw. sum kwadratów ładunków po rotacji (*Rotation Sums of Squared Loading*) dla poszczególnych czynników. Zaznaczając opcję *Unrotated factor solution,* która jest również ustawiona jako domyślna, w końcowej tabeli uzyskamy dodatkowe obliczenia określające sumy kwadratów

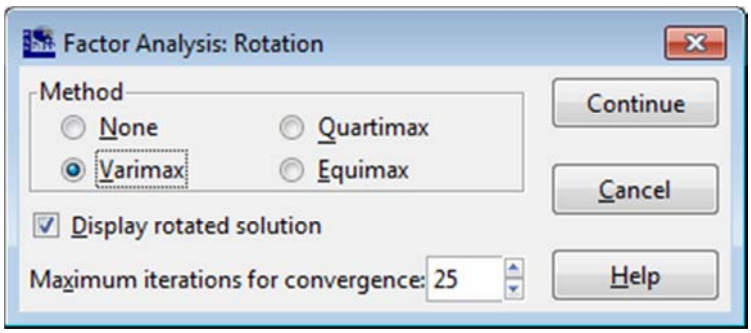
ładunków po wyodrębnieniu, ale bez rotacji (*Extaction Sums of Squared Loadings*). Z kolei zaznaczenie drugiej opcji - *Scree plot,* pozwala na stworzenie wykresu osypiska, nazywanego również wykresem kolankowym. Jego interpretacja pozwala na określenie wymaganej liczby czynników. Jest on publikowany w oknie raportu programu PSPP. Analizą konkretnego przykładu zajmiemy się w dalszej części podręcznika. Ważne jest bowiem poznanie kolejnych kryteriów wyodrębniania czynników.

Opcje znajdujące się w polu *Extract* (oznaczonym numerem 4), co również należy tłumaczyć jako „Wyodrębnianie", określają warunki brzegowe dla wydobywanych czynników. W tym miejscu wyznaczamy podstawowe kryteria dotyczące przekształcenia pierwotnego zakresu zmiennych i ich redukcji do odpowiedniej liczby składowych. W programie PSPP zawsze odnajdziemy zaznaczoną opcję *Eigenvalues over '1.00' times the mean eigenvalue.* Określamy w tym miejscu dolną, akceptowalną granicę wartości własnej, nazywaną również wartością charakterystyczną (*Eigenvalues*). Na ich podstawie wyznaczamy liczbę czynników dla przekształcanego zbioru zmiennych według kryterium Kaisera. Wartość własna jest miarą obliczaną dla każdego czynnika i pozwala ona na określenie stopnia wyjaśnianej wariancji przez dany czynnik. Program PSPP domyślnie ustala jej marginalną wartość na poziomie 1. Według kryterium Kaisera, które w tym miejscu jest wykorzystywane, w dalszej analizie będą brane pod uwagę wyłącznie te czynniki, których wartość własna jest większa niż 1. W praktyce analitycznej założenie to uznawane jest za standardowe. Przyjmujemy, że analizą zostaną objęte wyłącznie te czynniki, które tłumaczą nie mniej wariancji niż pojedyncza zmienna pierwotna. Mówiąc prościej - oznacza to, iż do poszczególnego wyodrębnionego czynnika zostanie przypisany co najmniej jeden pełny zakres informacji z pojedynczej zmiennej z oryginalnego zbioru danych. Obliczone wartości własne prezentowane są w oknie raportu w tabeli zatytułowanej *Total Variance Explained.* Zanim jednak przejdziemy do jej interpretacji, omówmy kolejną opcję w polu *Extract,* która nazywa się *Number of factors.* Umożliwia ona ręczne, odgórne określenie liczby wyodrębnianych czynników. W początkowym etapie analizy czynnikowej nie należy jej jednak wyznaczać i polegać na domyślnym wyodrębnianiu składowych przez program statystyczny. Arbitralne określenie liczby czynników rzutuje bowiem na strukturę redukcji zmiennych pierwotnych i ich klasyfikację do odpowiednich składowych. Zaleca się zatem pozostawienie tej opcji niezaznaczonej, przynajmniej na wstępnym etapie wyznaczania modelu przekształceń.

Ustalenie ostatniego kryterium dla procedury wyodrębniania czynników wymaga określenia maksymalnej liczby iteracji dla uzyskania zbieżności między danymi wejściowymi. Ta opcja w programie PSPP nazywana jest *Maksimum iterations for covergence* i została oznaczona na powyższym zrzucie ekranowym numerem 5. W tym miejscu określa się maksymalną liczbę powtórzeń tej samej procedury poszukiwania odpowiedniego dopasowania i odnajdywania zbieżności między zmiennymi. Jako wartość domyślna ustawiona jest liczba 25 iteracji. Maksymalną wartość, jaką może przyjąć ten współczynniki jest 100 powtórzeń. W analizie danych arbitralnie uznaje się, iż 25 iteracji jest wartością wystarczająca do wyznaczenia liczby czynników. Jeżeli okaże się ona zbyt mała, wówczas program PSPP poinformuje nas o tym w oknie raportu. Przy zaistnieniu takiej sytuacji należy zwiększać liczbę iteracji systematycznie o pięć bądź dziesięć jednostek, aż do uzyskania wyniku analizy. W początkowej fazie analizy czynnikowej nie jest jednak konieczne manipulowanie tym współczynnikiem i należy pozostawić go na poziomie domyślnym - 25 iteracji.

Dla wykonania analizy czynnikowej niezbędne jest ustalenie kolejnego, ostatniego kryterium - **metody rotacji**. Najczęściej stosowaną metodą jest *Varimax*. Ułatwia ona interpretację czynników dzięki minimalizacji liczby zmiennych o dużym poziomie ładunku w ramach pojedynczej składowej.

W programie PSPP wyboru tej metody dokonujemy w opcji *Rotation*s, co demonstruje poniższy zrzut ekranowy:



W programie PSPP mamy do wyboru trzy ortogonalne metody rotacji - *Varimax*, *Quatrimax* oraz *Equimax*. Dopuszczalne jest również niewybranie żadnej z metod rotacji (*None*), jednakże zalecane jest zastosowanie procedury rotowania osi dla uzyskania pełniejszych wyników analizy czynnikowej. W naszym przypadku dokonaliśmy wyboru opcji *Varimax*, którą jednocześnie polecamy początkującym badaczom. Jest to również metoda arbitralnie ustawiona przez twórców programu PSPP. Konieczne jest również zaznaczenie opcji wyświetlania rozwiązań rotowanych (*Display rotated solution*). Liczbę iteracji pozostawiamy na domyślnym poziomie 25 powtórzeń.

Ustalenie powyższych warunków umożliwia przystąpienie do interpretacji wstępnych wyników analizy czynnikowej. Kliknięcie 'OK' w oknie *Factor Analysis* powoduje opublikowanie właściwych statystyk w oknie raportu. Interpretację wyników należy rozpocząć od drugiej wyświetlanej tabeli, zatytułowanej *Total Variance Explained*. Przyjmuje ona następującą postać:

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.70	16.84	16.84	3.70	16.84	16.84	2.73	12.41	12.41
2	2.03	9.24	26.08	2.03	9.24	26.08	1.60	7.26	19.67
3	1.57	7.15	33.23	1.57	7.15	33.23	1.55	7.05	26.72
4	1.49	6.78	40.01	1.49	6.78	40.01	1.71	7.76	34.48
5	1.17	5.34	45.35	1.17	5.34	45.35	1.45	6.59	41.07
6	1.04	4.74	50.09	1.04	4.74	50.09	1.98	9.02	50.09
7	.96	4.36	54.44						
8	.90	4.09	58.53						
9	.83	3.79	62.32						
10	.83	3.77	66.09						
11	.78	3.54	69.63						
12	.77	3.48	73.11						
13	.74	3.34	76.46						
14	.69	3.13	79.59						
15	.65	2.94	82.53						
16	.64	2.91	85.44						
17	.62	2.81	88.25						
18	.58	2.62	90.87						
19	.55	2.50	93.37						
20	.52	2.38	95.75						
21	.49	2.24	97.99						
22	.44	2.01	100.00						

Wyniki zamieszczone w powyższej tabeli pozwalają na ustalenie minimalnej liczby czynników dla konstruowania zredukowanego modelu zbioru danych. Decyzję o liczbie nieskorelowanych składowych podejmujemy na podstawie obliczonych wartości własnych zamieszczonych w tabeli *Initial Eigenvalues*

w kolumnie *Total.* Wartości własne są szacowane dla maksymalnej liczby możliwych składowych (*Component*). Określają one wielkość wariancji danego czynnika przy założeniu, że wielkość wariancji standaryzowanej zmiennej pierwotnej wynosi 1. Jak można dostrzec, wartości własne odpowiadają liczbie wszystkich zmiennych włączonych do analizy czynnikowej. W naszym przypadku ich liczba wynosi 22 zmienne. W najmniej optymistycznym wariancie, pojedyncze zmienne nie wykazywałby żadnej zależności z pozostałymi zmiennymi i tworzyłyby nieskorelowane czynniki. W takiej sytuacji przeprowadzenie analizy czynnikowej byłoby nieuzasadnione. Po sprawdzeniu jednak wstępnych właściwości zmiennych wiemy, iż istnieją między nimi współzależności, co umożliwia zredukowanie pewnej liczby zmiennych do agregującego je czynnika.

Do ustalenia liczby czynników wykorzystujemy stosunek wartości własnej danego czynnika do sumy wartości własnych wszystkich składowych możliwych do wyodrębnienia. Określamy tym sposobem poziom całkowitej wariancji wyjaśnionej dla danego zbioru przez poszczególny czynnik. Mówiąc prościej, wyznaczamy zakres informacji zawartych w zmiennych pierwotnych, który możemy wytłumaczyć za pomocą jednego zbiorczego czynnika. Wielkość tejże wariancji dla poszczególnego czynnika odnajdujemy w kolejnej kolumnie zatytułowanej *% of Variance*. Jak możemy dostrzec, część całkowitej wariancji wyjaśnianej wyrażona jest w wartościach procentowych. Jest ona wyliczana według prostego schematu ilorazowego. Przyjmuje się w tym miejscu, że każda zmienna pierwotna ma charakter standaryzowany i wartość jej wariancji wynosi 1. Suma wartości własnych wszystkich czynników odpowiada zatem liczbie zmiennych włączonych do analizy czynnikowej. W naszym przykładzie dysponujemy 22-elementowym zbiorem, gdzie wielkość wariancji dla wystandaryzowanej zmiennej obliczymy według wzoru:

$$(1/22)*100 = 4{,}55$$

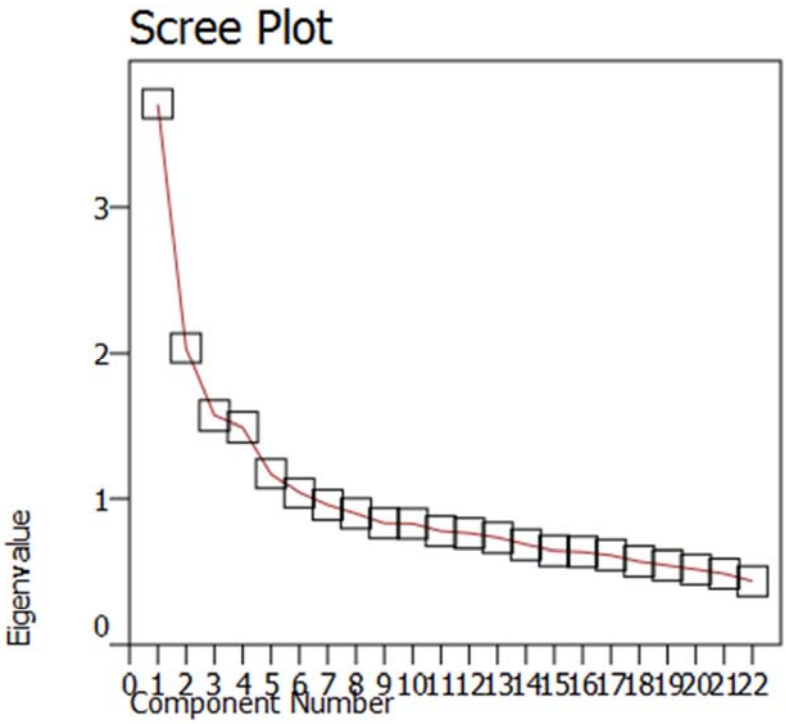
Pojedyncza zmienna w zoptymalizowanym wariancie wykazuje 4,55 proc. całkowitej wariancji wyjaśnianej. Gdy jednak odkryjemy ścisłe zależności między tymi zmiennymi i przekształcimy je do postaci czynników, zwiększamy tym samym zakres wyjaśnianej wariancji. Pierwszy wyodrębniony czynnik wyjaśnia 16,84 proc. wariancji wprowadzonych zmiennych, czyli niemal czterokrotnie (dokładnie 3,70 razy) więcej niż każda z nich. Tą miarę obliczamy według wzoru:

$$(\text{wartość własna czynnika}/\text{liczba zmiennych})*100$$

Przystąpmy jednak do wyznaczenia liczby czynników dla naszego przykładu. Dokonujemy tego na podstawie analizy wartości własnych wyodrębnionych składowych zamieszczonych w kolumnie *Total* w tabeli *Initial Eigenvalues*. W dokonywaniu wyboru odpowiedniej liczby czynników wykorzystujemy metodę zaproponowaną w 1960 roku przez H.F. Kaisera, nazywaną również kryterium Kaisera bądź niekiedy kryterium Guttmana-Kaisera. Służy ona do określania istotnej w sensie statystycznym liczby czynników. Metoda ta zakłada wyznaczenie dolnej granicy wartości własnej (*eigenvalue*) na poziomie większym niż 1. Oznacza to, że wyodrębniane czynniki powinny wyjaśniać więcej wariancji niż pojedyncza zmienna standaryzowana, której wartość właściwa wynosi 1. Program PSPP automatycznie selekcjonuje zakres istotnych czynników, co prezentują wyniki w tabelach: *Extracion Sums of Squared Loadings* (suma ładunków po wyodrębnieniu, ale przed rotacją) oraz *Rotation Sums of Squared Loadings* (suma ładunków po wyodrębnieniu i po zastosowaniu rotacji metodą *Varimax*). W analizowanym przykładzie na podstawie kryterium Kaisera możemy wyróżnić sześć czynników głównych. Wyróżnione czynniki są również posortowane według najwyższego poziomu wyjaśnianej wariancji. Pierwszy z nich odpowiada za największy zakres wyjaśnianej wariancji, tłumacząc 16,84 proc. zmienności całego zbioru danych. Własność własna tego czynnika wynosi 3,70. Wyjaśnia on zatem 3,7 razy więcej wariancji niż pojedyncza

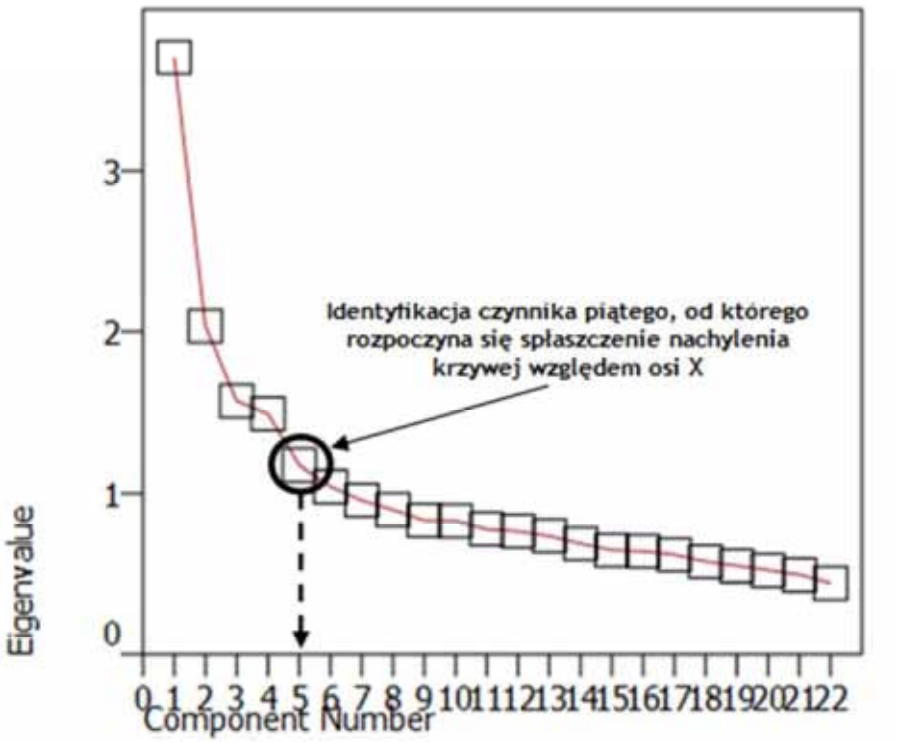
zmienna. Kolejne czynniki obejmują mniejszy zakres pozostałej wariancji wyjaśnianej, jednakże wciąż są istotne ze statystycznego punktu widzenia.

W program PSPP oprócz metody Kaisera, dostępna jest również metoda analizy wykresu osypiska (*Scree Plot*). Dla badanego przykładu został on zaprezentowany na poniższym zrzucie ekranowym:



Wykres osypiska, nazywany również wykresem kolankowym, jest metodą na wyznaczanie liczby czynników. Została zaproponowana przez R.B. Cattella - brytyjskiego psychologa i twórcę szesnastoczynnikowego modelu badania osobowości i pomysłodawcę kwestionariusza nazwanego *16PF Questionnaire.* Wykorzystując analizę czynnikową we własnych badaniach, zaproponował on graficzną metodę wyznaczania liczby głównych składowych określaną mianem testu osypiska. Wykres osypiska jest prostym wykresem liniowym prezentującym w układzie współrzędnych dane z tabeli *Total Variance Explained,* gdzie oś X określa kolejne wyróżniane czynniki, natomiast oś Y - przypisane im wartości własne. Analiza wykresu osypiska ma charakter interpretacji wizualnej. Na tej podstawie określamy optymalną liczbę czynników, do których zostanie zredukowany zbiór danych pierwotnych. Zamieszczona na wykresie linia przedstawia, w jak sposób zmienia się poziom wyjaśnianej wariancji przez kolejne czynniki. Jeżeli kolejne czynniki prezentują niską wielkość wariancji, ale jednocześnie podobną pod względem wartości, linia na wykresie ulega spłaszczeniu, ma małe nachylenie, zaś tendencja spadku zanika i wielkość wyjaśnianej wariancji wyrównuje się. Identyfikacja tego punktu jest równoznaczna z wyznaczeniem liczby wyodrębnianych czynników. W naszym przykładzie początek spłaszczenia obserwowany jest przy piątym czynniku.

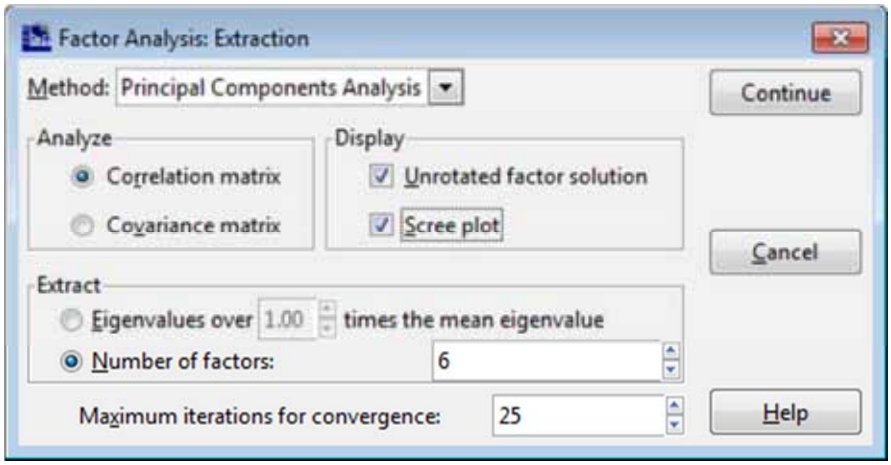
Punkt ten został zaznaczony na poniższym wykresie:



Na podstawie kryterium Cattella i po graficznej interpretacji wykresu osypiska, stwierdzamy, iż kolejne etapy analizy czynnikowej powinniśmy zostać przeprowadzone dla czterech czynników. Począwszy od piątego czynnika obserwujemy bowiem coraz mniejsze nachylenie krzywej, co oznacza, że pierwsze cztery czynniki skupiają odpowiedni łączny zakres wyjaśnianej wariancji. Zawierają one największą ilość właściwych informacji do zredukowania oryginalnego zbioru zmiennych, nie tracąc przy tym zbyt wiele z ich pierwotnej postaci.

Zauważmy, iż na podstawie kryterium Kaisera wyznaczyliśmy 6 czynników, natomiast za pomocą wykresu osypiska - 4 czynniki. Często jest tak, że metoda wartości własnych wyodrębnia większą liczbę czynników niż metoda Cattella. W praktyce analitycznej brakuje jednoznacznego stanowiska stwierdzającego, która z tych metod jest najwłaściwsza. Uznaje się, że metodą Kaisera uzyskuje się zbyt dużą liczbę czynników, co może utrudnić ich interpretację. Z kolei za pomocą wykresu osypiska wyznacza się mniejszą liczbę składowych, upraszczającą analizę i interpretację wyników. Argumenty przemawiające za wykorzystywaniem kryterium Kaisera ogniskują się na bardziej precyzyjnym i jednoznacznym wyznaczaniu czynników. Dlatego też jest ona powszechniej stosowana. Wyznaczenie liczby czynników zależy jednak od badacza i ma w dużym stopniu charakter umowny i arbitralny. Powszechną praktyką jest testowanie różnych struktur czynnikowych, obejmujących zarówno mniejszą jak i większą liczbę komponentów. W ostateczności wybierana jest ta wielkość, która pozwala na sensowną interpretację wyników. Poszczególne zmienne pierwotne klasyfikowane są bowiem do szerszych kategorii czynnikowych. Jeżeli uznamy, iż ciekawszą kompozycję uzyskamy po wyodrębnieniu 5 czynników, dalszą analizę przeprowadzamy dla tejże liczby składowych. W naszym przykładzie wybraliśmy sześcioczynnikowy model redukcji danych. Wymaga to wprowadzenia wartości '6' w opcji *Extraction* w polu *Number of factors* i powtórnego wykonania analizy czynnikowej dla wcześniej określonego zbioru zmiennych.

Ten aspekt zaprezentowaliśmy na poniższy zrzucie ekranowym:



Po wybraniu powyższych kryteriów oraz ustaleniu liczby wyodrębnianych czynników, możemy przejść do kolejnego etapu analizy czynnikowej. Polega ona na klasyfikacji poszczególnych zmiennych pierwotnych do poszczególnych kategorii czynnikowych. Wymaga to jednak odpowiedniego ich dopasowania, czemu poświęcony został kolejnych podrozdział.

18.4.3. Maksymalizacja dopasowania i koncepcyjna analiza wyodrębnionych czynników

Kolejnym etapem analizy czynnikowej jest przeprowadzenie procedury dopasowania zmiennych pierwotnych do struktury zredukowanego modelu czynnikowego. Dokonujemy tego na podstawie analizy ostatniej tabeli generowanej w oknie raportu programu PSPP, będącej macierzą rotowanych składowych (*Rotated Component Matrix).* Jej zawartość prezentuje poniższy zrzut ekranowy:

Rotated Component Matrix

	Component					
	1	2	3	4	5	6
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.	-.05	-.27	.12	.64	.14	-.02
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.	-.03	.14	.07	.80	.11	.14
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.	.22	.08	-.09	.67	-.09	.13
Ile godzin w ciągu tygodnia słucha: RMF FM.	.06	.78	.08	-.03	-.11	.05
Ile godzin w ciągu tygodnia słucha: Radio ZET.	.12	.80	.07	-.01	-.02	.03
Ile godzin w ciągu tygodnia słucha: TOK FM.	.23	.39	.15	.34	.11	-.07
Ile godzin w ciągu tygodnia słucha: Radio Maryja.	-.10	-.13	-.05	.20	.70	-.03
Ile godzin dziennie spędza na czytaniu gazet: Gazeta Wyborcza.	.60	.12	.11	.07	-.10	.05
Ile godzin dziennie spędza na czytaniu gazet: Rzeczpospolita.	.59	.13	.03	.07	.09	.11
Ile godzin dziennie spędza na czytaniu gazet: Dziennik.	.33	.07	.53	.01	.11	.00
Ile godzin dziennie spędza na czytaniu gazet: Fakt.	-.02	.07	.72	.05	-.06	.07
Ile godzin dziennie spędza na czytaniu gazet: Nasz Dziennik.	.08	.04	.30	-.01	.43	.43
Ile godzin dziennie spędza na czytaniu gazet: Trybuna.	.06	.08	.19	.00	.06	.69
Ile godzin dziennie spędza na czytaniu gazet: Superexspress.	-.01	.08	.69	.06	-.01	.24
Ile godzin w tygodniu spędza na czytaniu czasopism: Polityka.	.72	-.01	-.05	.09	-.05	.17
Ile godzin w tygodniu spędza na czytaniu czasopism: Wprost.	.74	.04	-.01	.01	.03	.09
Ile godzin w tygodniu spędza na czytaniu czasopism: Newsweek.	.69	.08	.09	-.07	.00	.05
Ile godzin w tygodniu spędza na czytaniu czasopism: Nie.	.18	-.05	.21	.00	-.18	.52
Ile godzin w tygodniu spędza na czytaniu czasopism: Przekrój.	.40	-.09	.07	.03	-.09	.39
Ile godzin w tygodniu spędza na czytaniu czasopism: Gazeta Polska.	.09	-.02	.05	.12	.09	.50
Ile godzin w tygodniu spędza na czytaniu czasopism: Tygodnik Powszechny.	.09	.09	-.15	.08	.18	.68
Ile godzin w tygodniu spędza na czytaniu czasopism: Niedziela	.03	.01	-.01	-.03	.77	.15

Na podstawie tej tabeli dopasowujemy zmienne do poszczególnych czynników. Mówiąc prościej, ustalamy, która zmienna daje się jednoznacznie przypisać do konkretnej grupy. W pierwotnej postaci dysponowaliśmy 22 zmiennymi, które chcemy zakwalifikować do odpowiednich sześciu wcześniej wyodrębnionych czynników. Dokonujemy tego na podstawie analizy wartości ładunków czynnikowych po zastosowaniu rotacji *Varimax*, które zamieszczone są w powyższej macierzy. Ładunki czynnikowe wyliczane są dla każdej zmiennej i dla każdego wyróżnionego czynnika. Określają one siłę korelacji R Pearsona, jaka zachodzi między zmienną a wyabstrahowanym czynnikiem. Przyjmuje się, iż minimalna wartość ładunku czynnikowego zmiennej musi wynosić 0,30. Przyjrzyjmy się wynikom rotowanej macierzy składowych i dokonajmy klasyfikacji odpowiednich zmiennych do poszczególnych grup. Na poniższym rysunku zostały zaznaczone odpowiednie ładunki spełniające warunek minimalnego poziomu zależności wynoszącego 0,30.

Rotated Component Matrix

	Component					
	1	2	3	4	5	6
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.	-.05	-.27	.12	.64	.14	-.02
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.	-.03	.14	.07	.80	.11	.14
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.	.22	.08	-.09	.67	-.09	.13
Ile godzin w ciągu tygodnia słucha: RMF FM.	.06	.78	.08	-.03	-.11	.05
Ile godzin w ciągu tygodnia słucha: Radio ZET.	.12	.80	.07	-.01	-.02	.03
Ile godzin w ciągu tygodnia słucha: TOK FM.	.23	.39	.15	.34	.11	-.07
Ile godzin w ciągu tygodnia słucha: Radio Maryja.	-.10	-.13	-.05	.20	.70	-.03
Ile godzin dziennie spędza na czytaniu gazet: Gazeta Wyborcza.	.60	.12	.11	.07	-.10	.05
Ile godzin dziennie spędza na czytaniu gazet: Rzeczpospolita.	.59	.13	.03	.07	.09	.11
Ile godzin dziennie spędza na czytaniu gazet: Dziennik.	.33	.07	.53	.01	.11	.00
Ile godzin dziennie spędza na czytaniu gazet: Fakt.	-.02	.07	.72	.05	-.06	.07
Ile godzin dziennie spędza na czytaniu gazet: Nasz Dziennik.	.08	.04	.30	-.01	.43	.43
Ile godzin dziennie spędza na czytaniu gazet: Trybuna.	.06	.08	.19	.00	.06	.69
Ile godzin dziennie spędza na czytaniu gazet: Superexspress.	-.01	.08	.69	.06	-.01	.24
Ile godzin w tygodniu spędza na czytaniu czasopism: Polityka.	.72	-.01	-.05	.09	-.05	.17
Ile godzin w tygodniu spędza na czytaniu czasopism: Wprost.	.74	.04	-.01	.01	.03	.09
Ile godzin w tygodniu spędza na czytaniu czasopism: Newsweek.	.69	.08	.09	-.07	.00	.05
Ile godzin w tygodniu spędza na czytaniu czasopism: Nie.	.18	-.05	.21	.00	-.18	.52
Ile godzin w tygodniu spędza na czytaniu czasopism: Przekrój.	.40	-.09	.07	.03	-.09	.39
Ile godzin w tygodniu spędza na czytaniu czasopism: Gazeta Polska.	.09	-.02	.05	.12	.09	.50
Ile godzin w tygodniu spędza na czytaniu czasopism: Tygodnik Powszechny.	.09	.09	-.15	.08	.18	.68
Ile godzin w tygodniu spędza na czytaniu czasopism: Niedziela	.03	.01	-.01	-.03	.77	.15

Większość zmiennych mierzących czas poświęcany na słuchaniu radia bądź czytaniu określonego czasopisma wykazuje jednoznacznie silną korelację z konkretnym czynnikiem. Pojawiły się jednak zmienne, które charakteryzują się zależnością przekraczającą 0,30 w ramach co najmniej dwóch wyodrębnionych czynników. Sytuację taką obserwujemy w przypadku radia TOK FM, gazety „Dziennik", gazety „Nasz Dziennik" oraz czasopisma „Przekrój". Oznacza to, iż zmienne właściwe tym mediom nie są jednoznacznie związane z danym czynnikiem. Dochodzi zatem do zaburzenia struktury modelu czynnikowego i zakwestionowania podstawowego założenia analizy czynnikowej, jaką jest redukowanie pierwotnego zbioru danych do nieskorelowanych składowych. Oznacza to również, iż powyższe zmienne nie pasują do ustalonej struktury czynnikowej i wymagają wyłączenia z analizy. Identyfikacja takich zmiennych wymaga

powtórnego przeprowadzenia analizy czynnikowej dla wybranego zbioru danych z wykluczeniem niejednoznacznie skorelowanych elementów, czyli zmiennej t98f (TOK FM), t99c (Dziennik), t99e (Nasz Dziennik), t100e (Przekrój). Procedurę określania zakresu redukowanych zmiennych pierwotnych nazywamy **maksymalizacją dopasowania czynników**. Jej celem jest ustalenie takiego zakresu czynników, który pozwoli na jednoznacznie rozłączne zakwalifikowanie danej zmiennej. Efekt powtórnego przeprowadzenie analizy czynnikowej dla sześciu czynników wynikowych prezentuje poniższa macierz rotowanych składowych:

Rotated Component Matrix

	Component					
	1	2	3	4	5	6
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 1.		.66				
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 2.		.82				
Ile godzin w ciągu tygodnia słucha: Polskie Radio Program 3.		.68				
Ile godzin w ciągu tygodnia słucha: RMF FM.			.82			
Ile godzin w ciągu tygodnia słucha: Radio ZET.			.82			
Ile godzin w ciągu tygodnia słucha: Radio Maryja.					.71	
Ile godzin dziennie spędza na czytaniu gazet: Gazeta Wyborcza.	.61					
Ile godzin dziennie spędza na czytaniu gazet: Rzeczpospolita.	.60					
Ile godzin dziennie spędza na czytaniu gazet: Fakt.						.82
Ile godzin dziennie spędza na czytaniu gazet: Trybuna.				.74		
Ile godzin dziennie spędza na czytaniu gazet: Superexspress.						.75
Ile godzin w tygodniu spędza na czytaniu czasopism: Polityka.	.74					
Ile godzin w tygodniu spędza na czytaniu czasopism: Wprost.	.76					
Ile godzin w tygodniu spędza na czytaniu czasopism: Newsweek.	.71					
Ile godzin w tygodniu spędza na czytaniu czasopism: Nie.				.55		
Ile godzin w tygodniu spędza na czytaniu czasopism: Gazeta Polska.				.49		
Ile godzin w tygodniu spędza na czytaniu czasopism: Tygodnik Powszechny.				.70		
Ile godzin w tygodniu spędza na czytaniu czasopism: Niedziela					.82	

Powyższa tabela prezentuje ostateczną postać wyników dla procedury wyodrębniania czynników i maksymalizacji ich dostosowania do właściwości zmiennych pierwotnych. Dla przejrzystości publikowanej tabeli skorzystano z dodatkowej opcji formatowania umożliwiającej wyświetlanie wartości ładunków czynnikowych większych niż 0,30. W programie PSPP wykorzystanie tej opcji wymaga wydania odpowiedniego polecenia w edytorze składni, z czym zapoznajemy Czytelnika w podrozdziale 18.4.4.

Przystąpmy do interpretacji uzyskanych wyników, które będziemy prezentować w końcowym raporcie z badania. Analizą czynnikową zostały objęte 22 zmienne identyfikujące ilość czasu poświęcanego w tygodniu na słuchaniu określonej stacji radiowej bądź na czytaniu określonego czasopisma. Ostatecznie do modelu czynnikowego zostało zakwalifikowanych 18 zmiennych, na podstawie których wyodrębniono sześć niezależnych czynników. Wyznaczona struktura tychże czynników wyjaśnia 56,12 proc. wariancji całkowitej pierwotnego zbioru zmiennych (wartość tą odczytujemy z tabeli *Total Variance Explained* z kolumny *Cumulative %* i odnajdujemy ją w wierszu właściwym dla ostatniego, szóstego wyodrębnionego czynnika). Dla czynnika pierwszego pięć zmiennych uzyskało wysoki poziom ładunku czynnikowego. Należą do nich „Gazeta Wyborcza", „Rzeczpospolita", „Polityka", „Wprost" oraz „Newsweek". Ze względu na dość charakterystyczne rodzaje czasopism, które ten czynnik grupuje, możemy określić go jako czynnik odzwierciedlający „polską prasę wiodącą". Są to bowiem czasopisma najczęściej i najchętniej czytane przez Polaków. Do drugiego czynnika zakwalifikujemy Program 1, Program 2

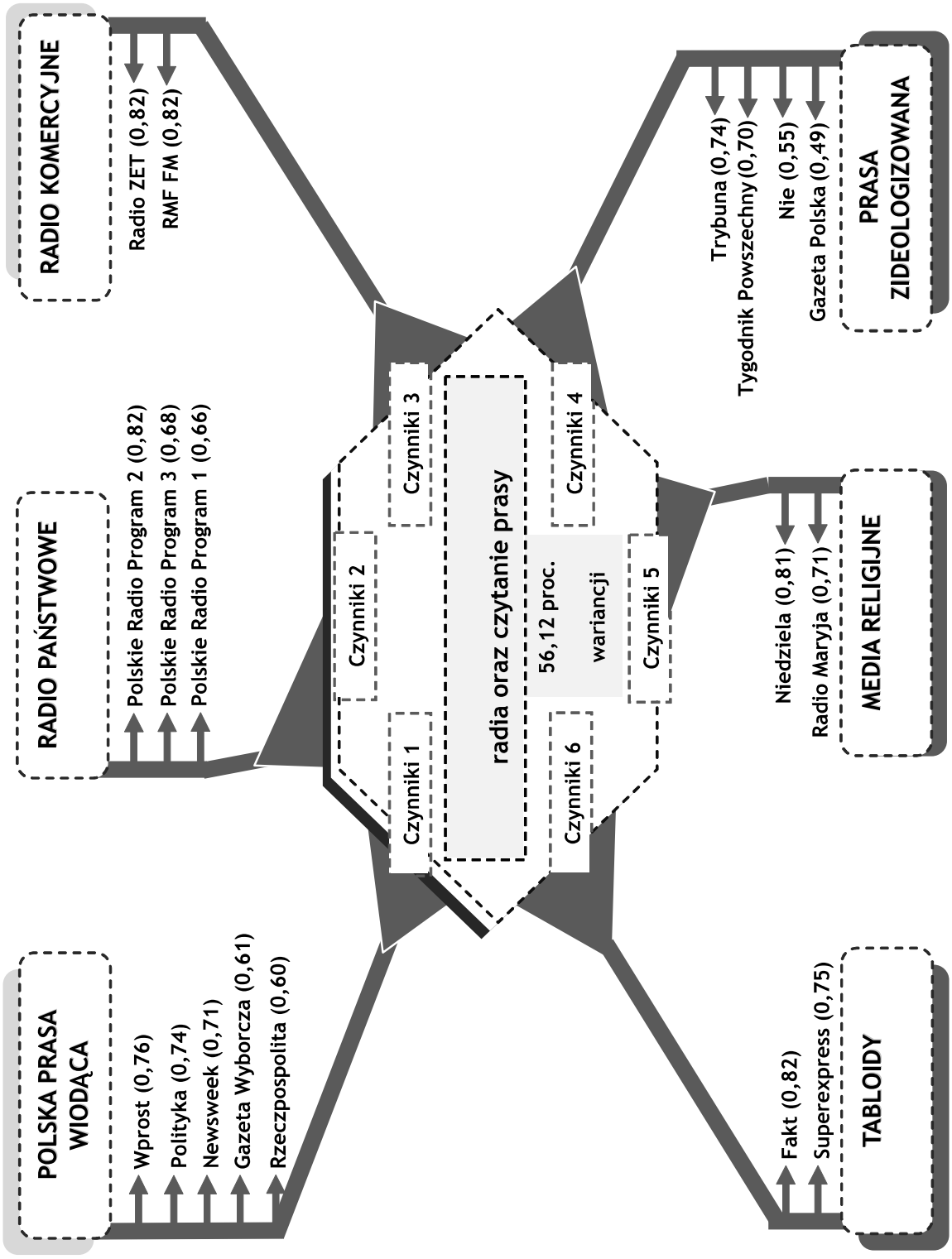
oraz Program 3 Polskiego Radia i określimy go mianem „państwowe stacje radiowe". Trzeci czynnik nazwiemy „komercyjne stacje radiowe", do którego zaliczmy takie stacje jak RMF FM oraz Radio ZET. Kolejny, czwarty czynnik obejmuje swoim zakresem cztery czasopisma - „Trybunę", tygodnik „Nie", „Gazetę Polską" oraz „Tygodnik Powszechny". Ze względu na dość charakterystyczne treści publikowane na łamach tychże czasopism ten czynnik określimy mianem prasy „zideologizowanej". Do piątego czynnika zaliczymy Radio Maryja oraz czasopismo „Niedziela" i przypiszemy mu ogólną nazwę „media religijne". Ostatni, szósty czynnik obejmować będzie gazetę „Fakt" oraz „Super Express". Na jego określenie użyjemy terminu „tabloidy".

W końcowym raporcie zasadne jest zaprezentowanie ostatecznej postaci macierzy rotowanych składowych, którą zamieściliśmy powyżej. Należy również podać rodzaj zastosowanej metody wyodrębniania czynników, czyli metodę głównych składowych (*Principal Components Analysis, PCA*) oraz wykorzystaną metodę rotacji - *Varimax.* Poza podaniem wielkości wyjaśnionej wariacji przez model wyodrębnionych czynników, dobrą praktyką jest zaprezentowanie rezultatów zastosowanych testów sprawdzających właściwości zmiennych pierwotnych. Powinny one zawierać wyniki obliczenia miary KMO, testu sferyczności Bartletta oraz wartość wyznacznika macierzy korelacji (zaprezentowano je w tabeli 47).

Tabela 47. Rezultat sprawdzenia właściwości zmiennych poddawanych analizie czynnikowej

Rodzaj testu weryfikującego właściwości zmiennych dla przeprowadzenia analizy czynnikowej	Wartości statystyk
Miara KMO	0,73
Test sferyczności Bartletta	χ^2 (153)=4358,01, p<0,05
Wyznacznik macierzy korelacji	0,08

Do prezentacji wyników analizy czynnikowej można wykorzystać również różnego rodzaju diagramy i tabele. Pewną propozycję przedstawiamy poniżej:



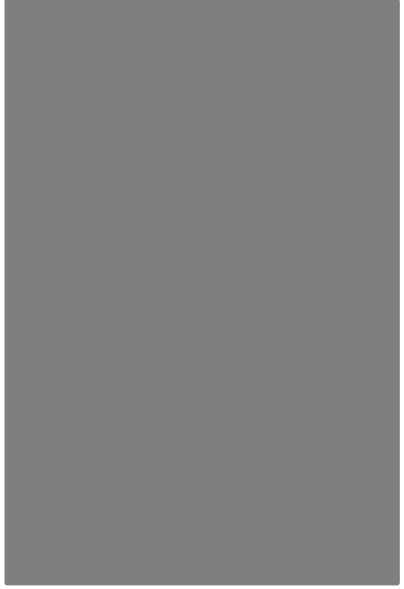
18.4.4. Analiza czynnikowa w edytorze składni

W programie PSPP analizę czynnikową najefektywniej przeprowadzamy za pomocą edytora składni. W trybie okienkowym wiele dodatkowych opcji jest niedostępnych. Należą do nich testy sprawdzające właściwości analizowanych zmiennych (miara KMO, test sferyczności Bartletta, wyznacznik macierzy korelacji), a także ustawienia formatu wyświetlanych współczynników i sposobów definiowania braków danych. Polecenie w edytorze składni posiada następującą postać ogólną:

Składnia do wpisania w Edytorze	Opis działania składni
FACTOR	- wykonaj analizę czynnikową
/ VARIABLES = 'lista zmiennych'	- dla wybranych zmiennych (po poleceniu VARIABLE wpisujemy wszystkie nazwy zmiennych ze zbioru danych)
/ METHOD = CORRELATION *lub* COVARIANCE	- wybierz odpowiednią metodę analizy czynnikowej opierającą się na macierz korelacji (CORRELATION) lub na macierzy kowariancji (COVARIANCE)
/ EXTRACTION = PC lub PAF	- wykonaj procedurę wyodrębnienia czynników z wykorzystaniem metody głównych składowych (*Principal Components* - oznaczamy symbolem PC) lub za pomocą metody głównych osi wymiaru (*Principal Axisnents* - oznaczamy symbolem PAF)
/ ROTATION = VARIMAX *lub* EQUAMAX *lub* QUARTIMAX *lub* NOROTATE	- wybierz odpowiednią metodę rotacji: *Varimax, Equimax lub Quartimax,* bądź zaniechaj procedury rotowania (NOROTATE)
/ PRINT = INITIAL *lub/i* EXTRACTION *lub/i* ROTATION *lub/i* UNIVARIATE *lub/i* CORRELATION *lub/i* COVARIANCE *lub/i* DET *lub/i* KMO *lub/i* SIG *lub* ALL *lub* DEFAULT	- wyświetl w oknie raportu odpowiednie statystki. Należą do nich: **1. INITIAL** - początkowe wartości współczynników zmiennych pierwotnych **2. EXTRACTION** - wartości współczynników obliczone po zastosowaniu procedury wyodrębniania czynników **3. ROTATION** - wartości współczynników obliczone po zastosowaniu procedury rotacji **4. UNIVARIATE** - statystyki opisowe obejmujące średnią arytmetyczną, odchylenie standardowe oraz liczebności obliczone dla każdej analizowanej zmiennej **5. CORRELATION** - obliczenie macierzy korelacji **6. COVARIANCE** - obliczenie macierzy kowariancji **7. DET** - obliczenie wyznacznika macierzy korelacji **8. KMO -** obliczenie miary KMO oraz testu sferyczności Bartletta **9. SIG -** obliczenie macierzy istotności korelacji (jest to dodatkowa metoda określania właściwości zmiennych pierwotnych i umożliwiającą ocenę związku konkretnej zmiennej z pozostałymi). **10. ALL -** opcja ta umożliwia obliczenie wszystkich powyższy statystyk **11. DEFAULT -** umożliwia obliczenie jedynie podstawowych statystyk z opcji INITIAL oraz EXTRACTION
/ PLOT = EIGEN	- wykonaj wykres osypiska dla wybranego zestawu zmiennych
/ FORMAT = SORT *lub/i* BLANK(*n*) *lub* DEFAULT	- wyznacz format wyświetlanych wyników w oknie raportu, biorąc pod uwagę takie możliwości, jak: **1. SORT -** funkcja ta umożliwia sortowanie malejące wartości wszystkich obliczanych współczynników np. wartości własnych lub wartości ładunków czynnikowych

Składnia do wpisania w Edytorze	Opis działania składni
/ FORMAT = SORT *lub/i* BLANK(*n*) *lub* DEFAULT	**2. BLANK (n)** - opcja ta powoduje niewyświetlanie w macierzy głównych składowych oraz w rotowanej macierzy składowych o wartości ładunków czynnikowych niższych niż zdefiniowana dolna granica (n). Z reguły przyjmujemy się, iż wynosi ona 0,30. W takiej sytuacji należy w edytorze składni umieścić następujące polecenie - BLANK (.3) **3. DEFAULT** - ta opcja powoduje brak zastosowania powyższych sposobów formatowania wyświetlanych wyników
/ CRITERIA = FACTORS(*n*) *lub/i* MINEIGEN(*l*) *lub/i* ITERATE(*m*) *lub/i* ECONVERGE (*delta*) *lub* DEFAULT	- wyznacz dodatkowe kryteria dla przeprowadzanej analizy czynnikowej. Za pomocą tego polecenie określamy: **1. FACTORS(n)** - liczbę *n* czynników, które chcemy wyodrębnić w przypadku analizowanego zbioru. Jeżeli będziemy chcieli przeprowadzić procedurę dopasowania dla 6 czynników, to liczbę tą wpiszemy w polu (n). Wówczas polecenie przyjmie postać FACTORS(6); **2. MINEIGEN(l) -** minimalną wartość własną *l* dla wyodrębnianych czynników. Według kryterium Kaisera wartość ta wynosi l=1,00. W takiej sytuacji polecenie przyjmie postać MINIEIGEN(1); **3. ITERATE(m)** - liczba *m* iteracji (powtórzeń) procedury wyodrębniania czynników lub przeprowadzania rotacji. Domyślnie jest ona ustalona na poziomie m=25, czego efektem jest następująco postać polecenia - ITERATE(25); **4. ECONVERGE(delta)** - wartość współczynnika delta wyznaczanego w sytuacji wykorzystania metody głównych osi wymiaru (*Principal Axis Factoring, PAF*). Wartość tego parametru jest określana w sytuacji zastosowania algorytmu iteracyjnego dla ekstrakcji czynników. Współczynnik delta określa maksymalną wartość bezwzględnych szacunków wspólnych między ostatnią a poprzedzającą ją iteracją. Jeżeli zostanie uzyskana wartość mniejsza niż wyznaczona przez współczynnik delta, to kolejne iteracje zostaną zaprzestane. Domyślnie wartość delta ustalona jest na poziomie 0,001 i w poleceniu zapisana jako ECONVERGE (.001); **5. DEFAULT** - ta opcja powoduje brak zastosowania powyższych kryteriów. Jako opcja domyślna ustalona jest jedynie liczba iteracji wynosząca 25 powtórzeń - ITERATE(25)
/ MISSING = LISTWISE lub/i PAIRWISE lub/i INCLUDE lub EXCLUDE	- określ sposób definiowania braków danych i włączania ich w zakres analizy. W ramach tego podpolecenia możliwe jest: **1. LISTWISE -** wyłącz wszystkie obserwacje z brakami danych **2. PAIRWISE -** wyłącz obserwacje parami, czyli przypadki, gdzie brakuje jednej z wartości dla danego współczynnika bądź zestawu zmiennych **3. INCLUDE -** włącz do analizy wartości zmiennych, które zostały uprzednio zdefiniowane jako braki danych, ale wyklucz wszystkie systemowe braki danych **4. EXCLUDE -** wyłącz wartości zmiennych zdefiniowanego jako braki danych oraz systemowe braki danych. Jest to opcja ustawiona jako domyślna
EXECUTE.	- wykonaj obliczenia.

Istnieje wiele przesłanek przemawiających na korzyść wykorzystania edytora składni w przeprowadzaniu analizy czynnikowej. Pierwszą z nich jest przede wszystkim efektywność i oszczędność czasu. Zauważmy, iż prawidłowe wykonanie analizy czynnikowej wymaga wielokrotnego jej powtórzenia: po pierwsze - dla różnego zakresu zmiennych, po drugie - dla różnych kryteriów, wymieniając chociażby liczbę wyodrębnianych czynników lub też liczbę wykonywanych iteracji. W edytorze składni łatwo można manipulować tymi kryteriami. Ponadto, część opcji właściwych analizie czynnikowej, możliwe jest do wykonania jedynie w trybie poleceń. Należą do nich testy weryfikujące zasadność przeprowadzenia analizy czynnikowej dla określonego zestawu zmiennych, a także możność ustalenia kryteriów prezentowanych wyników. Z tych względów zalecamy początkującemu badaczowi rozpoczęcie nauki analizy czynnikowej od jej wykonywania za pomocą edytora składni w programie PSPP.



Aneksy

Aneks 1. Zalecana literatura przedmiotu

Poniżej Czytelnik odnajdzie wybór i omówienie literatury z zakresu analizy danych ilościowych, z którą szczególnie warto się zapoznać, by ugruntować lub rozwinąć wiedzę i umiejętności nabyte podczas lektury i ćwiczeń z niniejszym podręcznikiem. Jest to wybór autorski, dedykowany szczególnie studentom nauk politycznych. Omówione publikacje są cennym uzupełnieniem dla tych, którzy realizują samodzielne projekty na etapie pierwszego i drugiego stopnia studiów, a wydają się niezbędne dla studentów studiów trzeciego stopnia.

* * *

Babbie Earl, *Badania społeczne w praktyce*, tłum. W. Betkiewicz, M. Bucholc, P. Gadomski, J. Haman, A. Jasiewicz-Betkiewicz, A. Kloskowska-Dudzińska, M. Kowalski, M. Mozga, Wydawnictwo Naukowe PWN, Warszawa 2004.

Nie jest to książka do nauki statystyki. Jest to przede wszystkim doskonały (acz podstawowy) podręcznik metodologii badań w naukach społecznych. Stanowi cenne i niezbędne uzupełnienie lektury każdego analityka danych. Szczególnie warto polecić początkującym badaczom następujące rozdziały dzieła: *Rozdział 2. Paradygmaty, teoria i badania społeczne*, *Rozdział 3. Pojęcie przyczynowości w badaniach społecznych*, całą Część 2 zatytułowaną *Struktura procesu badawczego* składającą się z pięciu rozdziałów. Warto zapoznać się również z rozdziałami 16 i 17 (*Model analizy rozbudowanej, Statystyka w naukach społecznych*). Praktyczną wiedzę dla analityka zawierają również zamieszczone na końcu książki załączniki: *C. Raport z badań, H. Szacunkowy błąd z próby oraz I. Podręcznik SPSS.*

* * *

Blalock Hubert M. Jr., *Statystyka dla socjologów*, tłum. M. Tabin, I. Topińska, K. Starzec, Państwowe Wydawnictwo Naukowe, Warszawa 1977.

Statystyka dla socjologów jest podręcznikiem wydanym po raz pierwszy w latach sześćdziesiątych XX wieku. Wychowały się na nim pokolenia polskich (i nie tylko) socjologów. Podręcznik ten jest dziełem amerykańskiego socjologa, profesora socjologii Uniwersytetu Waszyngtońskiego, prezesa Amerykańskiego Towarzystwa Socjologicznego, uważanego za jedną z postaci, które ukształtowały amerykańską socjologię.

Dzieło warto polecić przede wszystkim ze względu na obszerność i szczegółowość wykładu. Wydaje się obowiązkową lekturą dla każdego, kto chciałby pogłębić swoją wiedzę z zakresu statystyki. Można je traktować jako kolejny stopień wtajemniczenia po lekturze niniejszego podręcznika. Książka, co prawda, wymaga od czytelnika pewnego przygotowania matematycznego, jednakże wszystkie przykłady, na których poszczególne miary statystyczne są omawiane pochodzą z nauk społecznych. Fakt ten nie tylko ułatwia proces dydaktyczny, lecz również może stanowić źródło inspiracji i pomysłów wykorzystania rozmaitych miar przez początkujących badaczy.

Podręcznik składa się z trzech części: wprowadzenia, statystyki opisowej i statystyki indukcyjnej. We wprowadzeniu omówione zostały pokrótce funkcje statystyki oraz podstawowe pojęcia: teorii, hipotezy i poziomu pomiaru. Część dotycząca statystyki opisowej została podzielona na rozdziały poświęcone: graficznej prezentacji danych, miarom pozycyjnym oraz miarom dyspersji. Pierwszą część zamyka rozdział wprowadzający Czytelnika do zagadnienia rozkładu normalnego. Najobszerniejsza jest część trzecia, w której H.M. Blalock omawia zagadnienia statystyki indukcyjnej. W kilkunastu rozdziałach omawiane są miary wykorzystujące testy istotności. W pierwszych partiach tej części zostały wyczerpująco omówione kwestie związane z rachunkiem prawdopodobieństwa i testowaniem hipotez oraz estymacją. Kolejnych kilka rozdziałów dotyczy testów nieparametrycznych (tu należy się zastrzeżenie, że kwestie ich interpretacji nie są już obecnie tak rygorystycznie przestrzegane, jak postrzegał je H.M. Blalock) oraz testowaniu różnic między grupami (analiza wariancji). Szczególnie wiele miejsca autor podręcznika poświęcił miarom korelacji i analizie regresji. Wszystkich, którzy są zainteresowani tymi miarami lub pragną stosować bardziej zaawansowane, a oparte na nich sposoby należy zachęcić do zapoznania się z tą częścią podręcznika.

Szczególnie dzieło to warto polecić tym wszystkim, którzy - siłą rzeczy - wiedzę statystyczną zaprezentowaną w niniejszym podręczniku uznali za niewystarczającą. W przypadku prac promocyjnych na trzecim stopniu studiów znajomość tego podręcznika wydaje się nieodzowna.

* * *

Górniak Jarosław, Wachnicki Janusz, *Pierwsze kroki w analizie danych. SPSS PL for Windows*, SPSS Polska, Kraków 2000.

Książka stanowi abecadło analizy danych. W pierwszej części wprowadza do mechaniki i obsługi programu SPPS, a w drugiej - do analizy danych (również z użyciem programu SPSS). Pierwszą część podręcznika można polecić początkującym badaczom, drugą - średniozaawansowanym. Szczególnie dobrze zostały opisane zagadnienia związane z przekształcaniem danych (część I, rozdział 6) oraz dotyczące pomiaru siły związku pomiędzy zmiennymi w tabelach kontyngencji (rozdział 5).

* * *

Malarska Anna, *Statystyczna analiza danych wspomagana programem SPSS*, SPSS Polska, Kraków 2005.

Książkę tą należy polecić gorąco wszystkim tym, którzy pragną kontynuować swoją przygodę z analizą danych. Zaprezentowano w niej w sposób nieco bardziej sformalizowany - przez co trudniejszy, ale jednocześnie bardziej pogłębiony i drobiazgowy - niektóre miary statystyczne: analizę wariancji, analizę czynnikową, analizę wariancji analizę korelacji i regresji. Warto sięgnąć po tę pozycję dla ugruntowania i uzupełnienia zdobytej wiedzy.

* * *

Nawojczyk Maria, *Przewodnik po statystyce dla socjologów,* SPSS Polska, Kraków 2002.

Publikacja przeznaczona dla początkujących analityków. Została doskonale przygotowana pod względem dydaktycznym - prostoty i zrozumiałości wykładu. Gorąco zachęcam wszystkich początkujących badaczy do zapoznania się z nią. Zawiera nader przyjazne wprowadzenie do zagadnień statystycznych. Każdy rozdział zaopatrzono w propozycje praktycznych ćwiczeń z użyciem programu SPSS, co pozwala na ugruntowanie nabytej w toku lektury wiedzy. Lektura i nauka na jej podstawie statystyki nie wymaga przygotowania matematycznego. Wszelkie kwestie wyłożone są w sposób przystępny. Publikacja obejmuje zagadnienia statystyki opisowej oraz podstawy statystyki indukcyjnej (w tym: test t-Studenta, analizę wariancji, test chi-kwadrat oraz regresję liniową). Początkujący analityk odnajdzie również w pierwszych rozdziałach książki elementarne wprowadzenie metodologiczne dotyczące metody naukowej, testowania hipotez, typów zależności i związku przyczynowo-skutkowego oraz poziomów pomiaru.

* * *

Pavkov Thomas W., Pierce Kent A., *Do biegu, gotowi - start! Wprowadzenie do SPSS dla Windows*, tłum. J. Buczny, Gdańskie Wydawnictwo Psychologiczne, Gdańsk 2005.

Jest to podręcznik programu SPSS w wersji 11.0 na poziomie elementarnym. Instruktaż ma charakter czysto techniczny. Obejmuje zasady wyboru określonego testu, procedurę obsługi programu oraz - w większości przypadków - sposób interpretacji uzyskanych wyników. W publikacji znajdziemy wprowadzenie do takich testów statystycznych, jak test t-Studenta, analizę wariancji ANOVA, analizę regresji, korelacji R Pearsona oraz test chi-kwadrat. Podręcznik zawiera również opis mechaniki programu - menu i okien oraz rozdział na temat graficznej prezentacji danych. Ze względu na liczne uproszczenia i ograniczenie się autorów wyłącznie do elementów czysto „technicznych", publikację należy traktować jako swoiste, praktyczne repetytorium procedur wykonywania testów. Stanowi dobry, choć grzeszący zbytnimi uproszczeniami, wstęp do analizy statystycznej dla każdego, kto chciałby analizy wykonywać poprawnie pod względem „technicznym", lecz niekoniecznie ze zrozumieniem stosowanych miar.

* * *

Rao Calyampudi Radhakrishna, *Statystyka i prawda*, tłum. M Abrahamowicz, M. Męczarski, Wydawnictwo Naukowe PWN, Warszawa 1994.

Nie jest to typowy podręcznik, lecz książka w błyskotliwy sposób popularyzująca statystykę, napisana z pasją przez indyjskiego matematyka i statystyka, który poświęcił całe swoje życie zgłębianiu tych dyscyplin. Warto polecić ją wszystkim tym, których statystyka zafascynowała lub chociaż zainteresowała. Liczne, przykłady, opisy, paradoksy, anegdoty, towarzyszą Czytelnikowi w całym tekście książki. Książka ta umożliwia zrozumienie różnorodnych zastosowań i funkcji statystyki.

* * *

Shively W. Philips, *Sztuka prowadzenia badań politycznych,* tłum. E. Hornowska, Wydawnictwo Zysk i S-ka, Poznań 2001.

Książka, z którą powinien zapoznać się każdy student politologii. Zawiera przejrzyście i ciekawie podaną wiedzę dwojakiego rodzaju: z zakresu metodologii badań politologicznych oraz statystyki. Książka ta powstała na kanwie zajęć, jakie autor prowadził ze studentami na Uniwersytecie Yale. Zagadnienia metodologiczne obejmują typologię badań politologicznych, zagadnienia teorii politologicznych oraz wprowadzenie do badań

eksperymentalnych. Z kolei na omówione przez autora zagadnienia statystyczne, złożyła się szczegółowa i nader dobrze wyłożona analiza korelacji i regresji, w tym analiza *logit* i *probit* oraz testu chi-kwadrat. W podręczniku znalazł się również wykład na temat rzetelności i trafności pomiaru oraz liczebności próby. Publikacja ta w kontekście analizy danych ilościowych umożliwi pogłębione rozumienie wymienionych testów statystycznych.

* * *

***Statystyczny drogowskaz. Praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii*, Sylwia Bedyńska, Aneta Brzezicka (red.), Wydawnictwo „Academica", Warszawa 2007.**

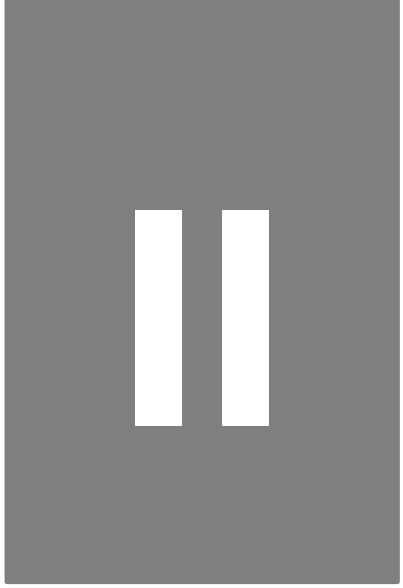
Jest to jedna z lepszych na polskim rynku wydawniczym książek przeznaczona do nauki obsługi programu SPSS (wersja 14.0). Publikacja została nagrodzona na targach „Academica 2008" jako najlepsza książka naukowa i akademicka. Jest to dzieło zbiorowe napisane siłami psychologów ze Szkoły Głównej Psychologii Społecznej w Warszawie. Obejmuje ono trzy części: przygotowanie danych do analizy, analizę danych oraz integrację zagadnień. W pierwszej części czytelnik zapoznaje się z podstawami metodologii (hipoteza, problem badawczy) oraz mechaniką pracy programu SPSS. Część ta obejmuje także analizę rzetelności oraz analizę czynnikową oraz wprowadzenie do analizy korelacji i regresji. W części poświęconej analizie statystycznej danych omówiono test chi-kwadrat oraz zastosowania testu t-Studenta (dla jednej próby, dla prób zależnych i dla prób niezależnych). Szczególnie dobrze została wyłożona część dotycząca porównywania wielu grup: jednoczynnikowa (ANOVA) i wieloczynnikowa (MANOVA) analiza wariancji. Część dotycząca integracji zagadnień zawiera pomocny początkującym badaczom poradnik obejmujący takie zagadnienia jak schemat procesu badawczego, wskazówki dotyczące wyboru właściwego testu statystycznego oraz tworzenia raportu.

Pomimo, że publikacja przeznaczona jest dla psychologów, warto ją polecić także politologom. Do pewnego stopnia może stanowić alternatywę dla niniejszego dzieła, szczególnie wśród tych politologów, których interesuje przede wszystkim pomiar ilościowy na obszarze psychologii polityki i psychologii społecznej.

* * *

Lissowski Grzegorz, Haman Jacek, Jasiński Mikołaj, *Podstawy statystyki dla socjologów,* Wydawnictwo Naukowe „Scholar", Warszawa 2011.

Podstawy statystyki dla socjologów są doskonałym, wyczerpującym podręcznikiem statystyki. Biblia każdego analityka danych napisana przez najlepszych polskich specjalistów. Na rynku wydawniczym dostępna jest edycja trzytomowa oraz jednotomowa (ta ostatnia wydana w serii *Wykłady z socjologii*). W edycji trzytomowej pierwszy tom został zatytułowany *Opis statystyczny,* drugi - *Zależności statystyczne,* a trzeci - *Wnioskowanie statystyczne.* Lektura dzieła wydaje się nieodzowna przy samodzielnie prowadzonych projektach badawczych.



Aneks 2. Przegląd i ewaluacja programów do analiz danych ilościowych

W niniejszym rozdziale zamieszczono autorski przegląd programów służących do analizy danych ilościowych, w tym zaprezentowano techniczne i merytoryczne charakterystyki tych aplikacji. Dokonano również ewaluacji tych programów pod względem zastosowań do celów naukowych i biznesowych. Badania na potrzeby nauki i na potrzeby biznesu różnią się w zakresie wymagań stawianych oprogramowaniu statystycznemu. W przypadku badań naukowych na ogół jeden projekt badawczy realizowany jest w dłuższym czasie, dokonywana jest wielostronna i pogłębiona analiza zbioru danych, stosowane są klasyczne, średniozaawansowane i zaawansowane testy statystyczne. Analizy biznesowe (jak wynika z wieloletnich doświadczeń autora aneksu) cechują się przede wszystkim (choć naturalnie bywają tu wyjątki) dynamiką i prostotą. Analizy wykonywane są najczęściej pod presją czasu i nie są one - nader często - zbyt wysublimowane pod względem statystycznym - wystarczające dla analityka są na ogół informacje ujęte w zestawieniach tabelarycznych, a największy nacisk kładzie się na możliwość rekonfiguracji i przekształceń zbioru danych.

Oprogramowanie służące do wykonywania analiz ilościowych można podzielić na komercyjne i niekomercyjne. Oprogramowanie komercyjne w założeniu przeznaczone jest dla biznesu (choć zaspokoi - nader dobrze - również potrzeby naukowe). Z kolei oprogramowanie niekomercyjne znajduje przede wszystkim zastosowanie w badaniach naukowych (może być jednak z powodzeniem stosowane również w biznesie). Przedstawione w rozdziale syntezy mają pomóc w wyborze oprogramowania do ewentualnych, przyszłych zapotrzebowań, wytworzyć u Czytelnika „mapę" oprogramowania służącego do analiz statystycznych, a przede wszystkim zachęcić do samodzielnych poszukiwań, porównywania oraz testowania dostępnych programów analitycznych.

2.1. Przegląd i ewaluacja oprogramowania niekomercyjnego

Spośród bogatej oferty niekomercyjnego, to jest wolnego oprogramowania wybrano przede wszystkim te aplikacje, które najlepiej wydają się odpowiadać badaczom społecznym. Pod uwagę brano funkcjonalność, elastyczność, łatwość obsługi oraz dostosowanie do potrzeb badaczy, co obejmuje nie tylko dostępność określonych testów statystycznych, lecz również opcje rekonfiguracji, formatowania i opisu danych. Lista analizowanych programów nie jest kompletna, ma ona charakter autorskiego przeglądu. Można jednak powiedzieć, że w dużym stopniu wyczerpuje ona zagadnienie dając przegląd i rozeznanie w zakresie Wolnego Oprogramowania.

2.1.1. GNU R

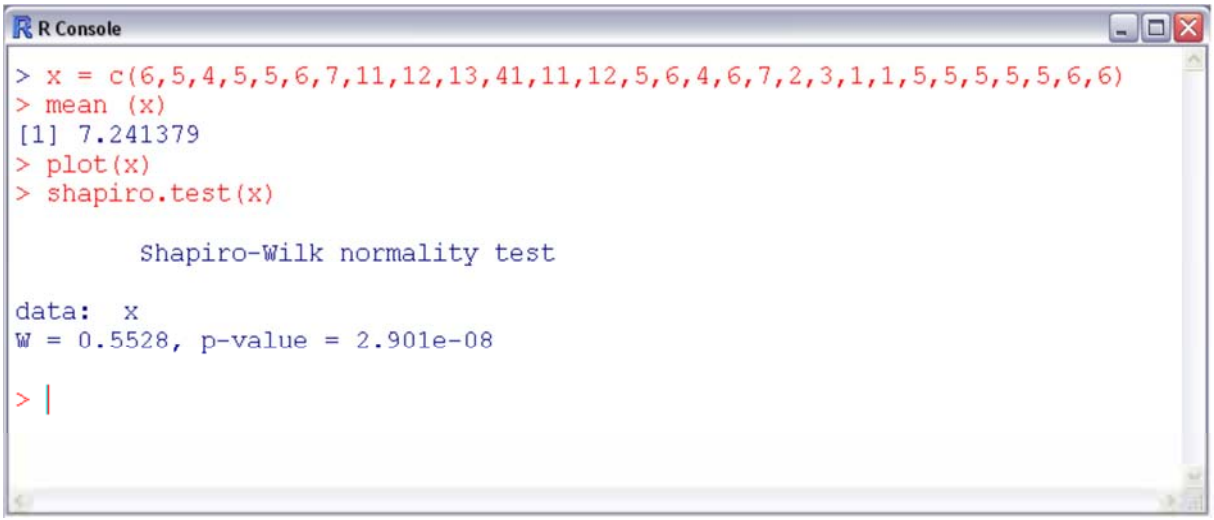
Program ten korzystnie wyróżnia się spośród innych programów komercyjnych i niekomercyjnych. Jest bogaty w rozmaite moduły statystyczne (zwane pakietami - *packages*), dynamicznie rozwijany, a w kształceniu uniwersyteckim cieszy się ogromnym poważaniem; na zachodnioeuropejskich uczelniach stał się standardowym narzędziem informatycznym wspierającym nauczanie statystyki na poziomie podstawowym i średniozaawansowanym. Autorami programu są Robert Gentleman (kanadyjski statystyk i bioinformatyk pracujący dla sektora prywatnego) oraz Ross Ihaka (nowozelandczyk, pracownik naukowy Uniwersytetu w Auckland, zajmujący się w pracy naukowej rozwojem oprogramowania służącego do analiz statystycznych). Pomysł stworzenia programu R pojawił się w 1993 roku, a pierwsza stabilna, oficjalna wersja 1.0.0. ukazała się w lutym 2000 roku. Najnowszą edycją jest opublikowana w marcu 2012 roku wersja 2.15.0. Program R jest obecnie częścią projektu GNU prowadzonego przez Free Software Foundation. Zajmuje się nim (od sierpnia 1997 roku) większe grono programistów-wolontariuszy noszące nazwę R Development Core Team, a wspiera go finansowo i organizacyjnie Fundacja R (The R Foundation for Statistical Computing) z siedzibą w Austrii na Wiedeńskim Uniwersytecie Ekonomicznym (Wirtschaftsuniversität Wien). Nazwa własna programu - „R" - jest efektem swoistej, charakterystycznej dla środowiska programistów gry słownej - w tym przypadku - kalamburyzacji. Program R powstał jako darmowy odpowiednik środowiska S i Scheme. Równie często jego nazwa wywodzona jest od inicjałów imion obydwu twórców programu (nazywanych również często R & R).

Program R zawiera blisko trzy tysiące rozmaitych pakietów statystycznych[1]. W podstawowej wersji programem tym steruje się za pomocą poleceń tekstowych (języka R) wpisywanych w tak zwanej konsoli (linii poleceń), brak jest graficznego interfejsu użytkownika (menu). Program R pozwala na wykonywanie statystyk podstawowych, jak też zaawansowanych. Konsola pozwala nie tylko na wprowadzanie poleceń dotyczących obliczeń, lecz również na administrowanie programem (pobieranie z Internetu i instalowanie nowych lub dodatkowych pakietów statystycznych). W linii poleceń program R może pracować w trybie interaktywnym lub w trybie wsadowym.

Tryb interaktywny polega na prowadzeniu dialogu z programem - po wpisaniu pojedynczego polecenia otrzymujemy od programu odpowiedź (potwierdzenie wykonania wraz z opisem lub zgłoszenie błędu). Z kolei istotą trybu wsadowego (przetwarzania wsadowego) jest wykonywanie już gotowych, napisanych uprzednio w języku R grup poleceń (programów, podprogramów). Pierwszy sposób pracy pozwala zacho-

[1] Porównaj: E. Gatnar, M. Walesiak, *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009, s. 13.

wać kontrolę nad pojedynczymi wykonywanymi operacjami, zaletą drugiego jest szybkość. Pracę w trybie interaktywnym przedstawia załączony zrzut z ekranu pokazujący proste funkcje języka R: utworzenie wektora danych, policzenie średniej, wykonanie wykresu oraz obliczenie testu Shapiro-Wilka.



* * *

W celu ułatwienia pracy z językiem R powstały programy wspomagające, organizujące przestrzeń roboczą i umożliwiające pracę w bogatszym środowisku graficznym, aniżeli tylko z użyciem składni. Zaawansowanymi projektami tego typu są RStudio i Tinn-R.

2.1.1.1. Rstudio

Tworzenie aplikacji rozpoczęto w 1997 roku. Wspomaga pracę ze składnią programu R, wyświetla macierz danych w układzie przypominającym PSPP i SPSS, umożliwia administrację programem R w trybie graficznym. Niewątpliwym ułatwieniem jest monitorowanie (w formie poleceń języka R) wszystkich wykonanych w zbiorze danych operacji - zarówno analitycznych, jak też administracyjnych. Nakładka współpracuje z programem R w wersji 2.11.1 lub nowszej. Można ją uruchomić w systemie operacyjnym Windows, Linux i Mac OS, a także jako usługę sieciową na serwerze www (zaimplementowano RStudio Server). Ułatwiono również pracę z grafiką statystyczną (wykresami, diagramami i grafami) - prosto eksportuje się je do formatu rysunkowego lub PDF. Wszystkie te udogodnienia przyspieszają i usprawniają pracę z programem GNU R. Głównym sposobem pracy w programie pozostaje jednak programowanie w języku R. Nakładkę można pobrać pod adresem http://rstudio.org/.

2.1.1.2. Tinn-R Editor (Tinn Is Not Notepad)

Jest to edytor kodu dla środowiska R (umożliwia jednak programowanie między innymi w odmianach języka C, Fortranie oraz HTML). Jest to projekt autorstwa brazylijskiego uczonego José Cláudio Faria z Universidade Estadual de Santa Cruz oraz jego współpracowników Philippe Grosjeana i Enio Jelihovschi wspomaganych przez zespół programistów Tinn-R Team. Program ten, wedle dokumentacji technicznej, umożliwia komunikację z R (przesyłanie instrukcji i odbieranie wyników), kolorowanie składni dla kilkunastu języków programowania (z możliwością dostosowania kolorowania), pracę na zakładkach,

wstawianie formuł LaTeX, nagrywanie makr, oznaczanie bloków i zarządzanie pakietami. Program ten nadaje się do zastosowań typowo naukowych i właściwie jest to produkt dedykowany programistom, w szczególności programistom języka R.

* * *

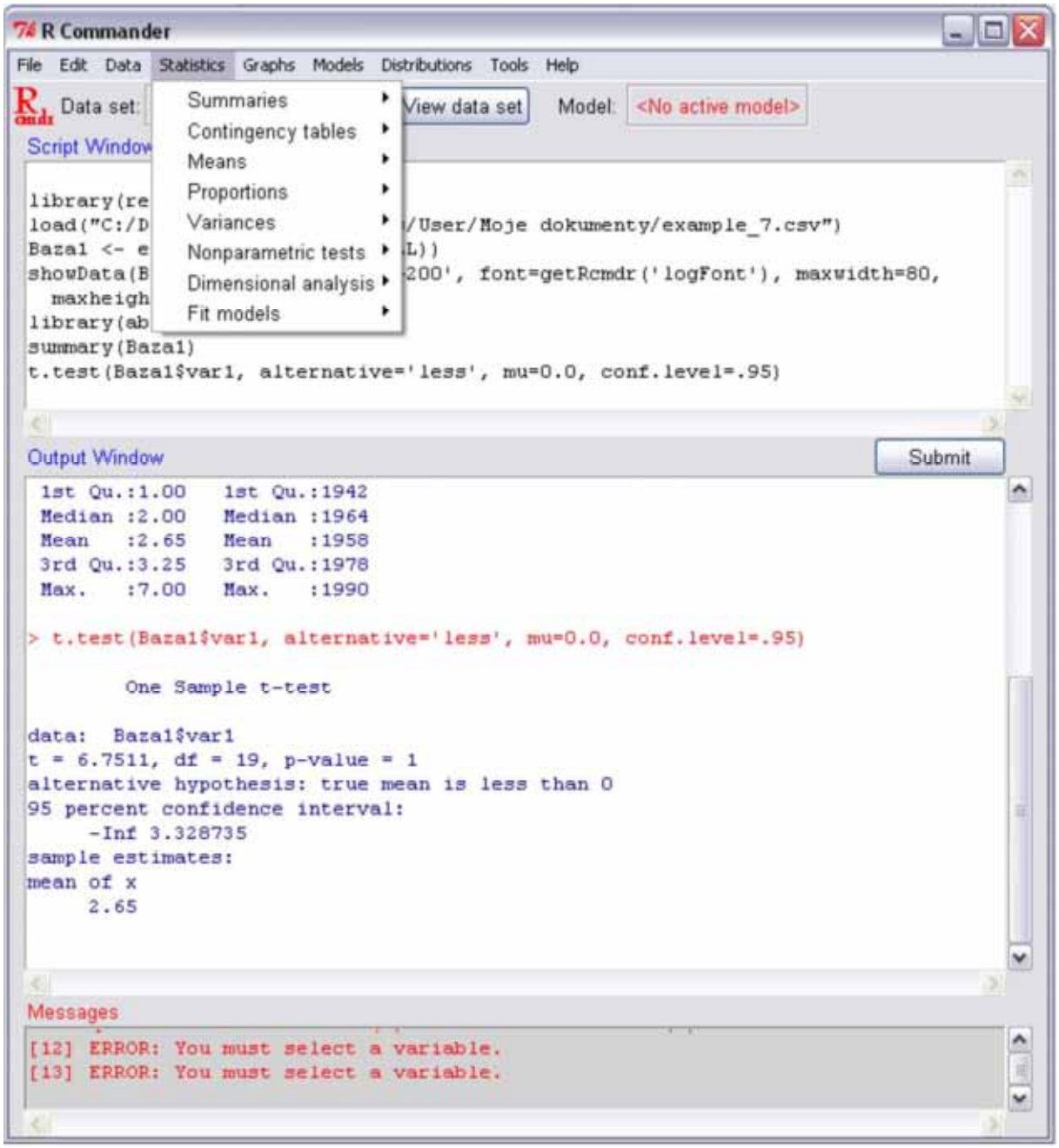
W związku z faktem, że użytkowanie programu wymaga nauczenia się języka programowania R, program ten uważany jest za trudny (niesłusznie zresztą, bowiem przezwyciężenie przyzwyczajenia pracy w trybie graficznym i nabycie umiejętności operowania w linii poleceń pozwala użytkownikowi uzyskać ogromną swobodę w zakresie projektowania analiz). Szczególnie jest to barierą dla zainteresowanych wykształconych w zakresie nauk społecznych. Dla tych właśnie użytkowników opracowano specjalny tryb pracy programu R nazywany „nadzorowanym". Są to nakładki umożliwiające pracę z wykorzystaniem pełnego interfejsu graficznego to jest okien i rozwijanych menu. Poniżej zaprezentowano znane nakładki na program GNU R.

2.1.1.3. Rcmdr (R Commander)

Jest to nakładka na program R umożliwiająca pracę z tym programem w trybie okienkowym. Instaluje się ją paroma prostymi poleceniami w języku R i uzyskuje środowisko pracy jak na zrzucie ekranowym znajdującym się poniżej. Jak wskazuje twórca aplikacji, amerykański socjolog i statystyk społeczny John Fox, idea tego programu wiąże się z jego potrzebami dydaktycznymi - nauczaniem studentów podstaw statystyki[2]. R Commander umożliwia rekodowanie, standaryzowanie, dodawanie i kasowanie jednostek analizy. Zmienne można obserwować, wprowadzać i modyfikować w trybie graficznym imitującym okno arkusza kalkulacyjnego. Nakładka umożliwia także generowanie tabel dla jednej i dla dwóch zmiennych oraz wykonywanie podstawowych statystyk parametrycznych i nieparametrycznych, a także badanie rozkładów. Można również za jej pomocą wygenerować wykresy, grafy i diagramy. Nakładka współpracuje z popularnymi programami służącymi do przechowywania lub analizy danych. Do programu można importować dane w trybie tekstowym (*.csv), a także formatów jak: Access, dBase, Excel, Minitab, PSPP i SPSS oraz Stata[3]. Program wydaje się użyteczny do zastosowań edukacyjnych, a także nieskomplikowanych zastosowań typowo naukowych. W ograniczonym stopniu może być przydatny w analizach biznesowych. Program można pobrać ze strony: http://cran.r-project.org/web/packages/Rcmdr/index.html, z kolei pod adresem: http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/ można odnaleźć instrukcję instalacji i obsługi oraz dokumentację techniczną programu.

[2] A. Ohri, *Interview Professor John Fox Creator R Commander*,w: http://www.decisionstats.com/interview-professor-john-fox-creator-r-commander/, dostęp: kwiecień 2012, 14 września 2009.

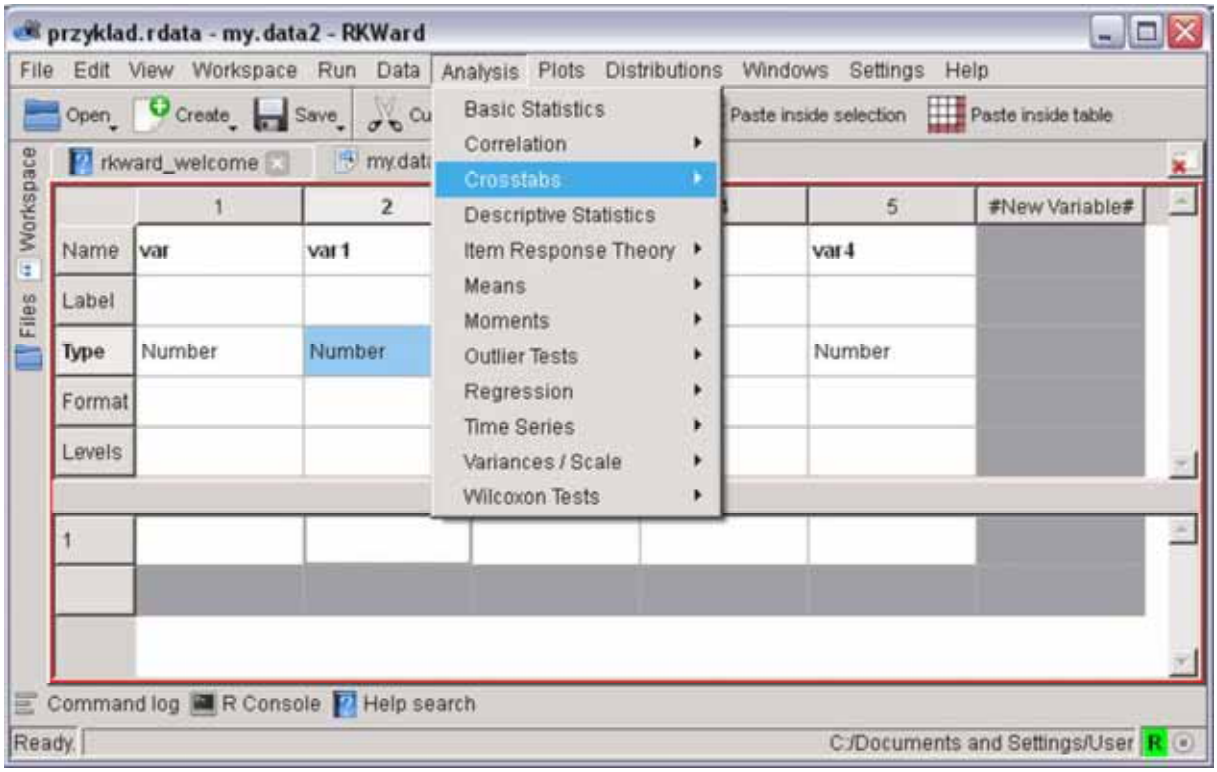
[3] Szerzej na temat tej nakładki na program GNU R: J. Fox, *The R Commander: A Basic Statistics Graphical User Interface to R,* „Journal of Statistical Software", 2005, 14 (9), s. 1-42; J. Fox, *Extending the R Commander by 'plug in' packages,* „R News", 2007, 7(3), s. 46-52.



2.1.1.4. RKWard

Program RKWard należy do rodziny Wolnego Oprogramowania (licencja GNU / GPL v. 2). Zapewnia on dużą wygodę i łatwość pracy nawet początkującemu użytkownikowi. Aplikacja ta pracuje zarówno pod kontrolą systemu operacyjnego Windows, jak również Linux. Instalacja w obu rodzajach systemów operacyjnych jest prosta, zautomatyzowana. W Linuxie aplikacja ta dodawana jest do powszechnie dostępnych repozytoriów (np. do Linuxa Ubuntu). Program uruchamiany jest odrębnie, niezależnie od R (stanowi to niewątpliwą zaletę, bowiem większość nakładek na program R uruchamiana jest w linii poleceń). W Linuxie działa w środowisku graficznym KDE. Obecnie (2012 rok) dostępny jest w wersji 4.7.0. Autorami programu są Thomas Friedrichsmeier (lider projektu), Pierre Ecochard (programista), Stefan Roediger (programista, tłumacz, specjalista w zakresie marketingu) oraz Prasenjit Kapat

(programista). Ponadto nad projektem pracują liczni ochotnicy z całego świata. Mankamentem tego programu jest fakt, że nie zawiera jakichkolwiek funkcjonalności służących do przetwarzania danych. Możliwe jest jedynie sortowanie oraz generowanie danych losowych. Program jest kompatybilny z PSPP i SPSS oraz STAT, umożliwia również importowanie danych w prostym zapisie tekstowych (*.dat, *.csv). Niewątpliwą zaletą jest fakt, że nakładka została napisana w języku C, co umożliwia szybsze działanie tego programu. Jest to dobrze dopracowany projekt, umożliwiający podstawowe i średniozaawansowane obliczenia statystyczne. Można polecić go do badań *stricte* naukowych, dydaktyki analiz ilościowych, spełni on również swoją rolę w analizach biznesowych. Program, dokumentację oraz listy dyskusyjne, a także pomoc użytkowników można uzyskać na stronie: http://rkward.sourceforge.net/. Poniżej znajduje się okno główne nakładki RKWard z rozwiniętym elementem analitycznym menu (*Analysis*).



2.1.1.5. Deducer (Java GUI For R, JGR)

Jest to najbardziej zaawansowana nakładka na program R, autorstwa Markusa Helbiga, Simina Urbanka i Iana Fellowsa z Uniwersytetu w Augsburgu. Została napisana w języku Java. Program może działać pod kontrolą systemów operacyjnych Windows, Linux oraz Mac OS X. Obecnie (tj. 2012) dostępna jest wersja 1.7-9 tego programu, rozwijany jest on od 2003 roku. Zapewnia on podstawowe funkcjonalności zarówno w zakresie przygotowania zbioru danych do analizy (między innymi: rekodowanie, transpozycję i łączenie danych), jak również analiz: proste tabele częstości, tabele krzyżowe, testy statystyczne (porównywanie średnich, regresję i korelację, analizę skupień), a także możliwość przedstawiania danych w estetycznej formie graficznej dzięki narzędziu Plot Builder (rysowanie rozmaitych wykresów dwuwymiarowych, grafów i diagramów). Praca w Deducer odbywa się w dwóch oknach: konsoli (*Console*) oraz przeglądarce danych (*Data Viewer*). Ten ostatni widok zamieszczono na zrzucie ekranowym. Pod względem wyglądu i funkcjonalności nakładka ta imituje program PSPP i SPSS i dla użytkownika korzystającego z wymienionych nie będzie problemem skorzystanie z programu Deducer.

Możliwe jest importowanie i praca na plikach pochodzących z programów Excel, Minitab, PSPP, SPSS, a także Stata. Obszerne instrukcje dotyczące instalacji oraz używania programu oraz sam program w wersji instalacyjnej znajdują się pod adresem http://www.deducer.org/pmwiki/index.php. Intuicyjna obsługa, możliwość wykonywania zautomatyzowanych zadań, a także ergonomia tego interfejsu graficznego sprawiają, że można używać go do zastosowań zdefiniowanych wyżej jako biznesowe, a w ograniczonym stopniu (obecność wyłącznie statystyk podstawowych i niektórych średniozaawansowanych) sprawiają, że można polecić go również do niektórych zastosowań naukowych. Do programu napisano wiele przydatnych wtyczek (*plug-ins*). Automatycznie instalowany jest wraz z programem Deducer (moduł DeducerExtras) zawierającym liczne średniozaawansowane testy statystyczne. Z kolei wtyczka DeducerPlugInScaling pozwala na wykonywanie testów rzetelności oraz analizy czynnikowej. Wymienić można także DeducerMMR (analiza regresji wielorakiej), DeducerSpatial (służący do wizualizacji danych). Na uwagę zasługuje eksperymentalna i jeszcze nie dołączana do pakietu wtyczka (*plug-in*) DeducerText umożliwiając dokonywanie drążenia tekstu (*text mining*).

	numer	px1dd	px1mm	px1r	gw	mw	c1t	c2t
1	192.0	16.0	11.0	2007.0	14.0	10.0	ŻADNA	ŻAD
2	203.0	16.0	11.0	2007.0	17.0	0.0	WYWIĄZANIE SIĘ Z OBIETNIC	PRO
3	320.0	15.0	11.0	2007.0	16.0	30.0	ŻADNA	-
4	328.0	17.0	11.0	2007.0	11.0	10.0	TRUDNO POWIEDZIEĆ	
5	359.0	19.0	11.0	2007.0	13.0	55.0	NIE INTERESOWAŁAM SIĘ TYM	NIE
6	424.0	18.0	11.0	2007.0	15.0	0.0	NIE WIEM	NIE
7	455.0	13.0	11.0	2007.0	14.0	0.0	WYŻSZE CENY ŻYWCA	POD
8	713.0	20.0	11.0	2007.0	11.0	20.0	NIE WIEM	NIE
9	772.0	15.0	11.0	2007.0	18.0	20.0	NIE BYŁO TAKIEJ KWESTII, NIE INTERESUJĘ SIĘ POL...	NIE
10	960.0	16.0	11.0	2007.0	16.0	0.0	NIE WIEM	NIE
11	1011.0	19.0	11.0	2007.0	18.0	45.0	SZACUNEK	
12	1063.0	24.0	11.0	2007.0	15.0	5.0	NISKIE ZAROBKI	
13	1100.0	21.0	11.0	2007.0	14.0	35.0	NIE WIEM	NIE
14	1168.0	18.0	11.0	2007.0	16.0	15.0	POMOC SAMOTNYM MATKOM	NIC
15	1479.0	23.0	11.0	2007.0	17.0	10.0	NIE WIEM	NIE

2.1.1.6. Rattle (R Analytical Tool To Learn Easily)

Nakładka ta jest najbardziej zaawansowana z opisywanych pod względem możliwości statystycznych obliczeń. Służy ona do drążenia danych (*data minig*)[4] i oferuje wszystkie grupy metod stosowane w eksploracji danych: wyszukiwanie asocjacji, klasyfikacja (wartości dyskretne), predykcja (wartości ciągłe), grupowanie (*clustering*), eksploracja złożonych typów danych, budowanie i ewaluacja modeli. Autorem programu jest Graham Williams pracownik naukowy University of Canberra i Australian National University. Należy podkreślić, że Rattle stanowi produkt unikalny - jest potężnym narzędziem statystycznym nie tylko dorównującym, ale przewyższającym komercyjne i niekomercyjne narzędzia eksploracji danych. Jak podkreśla G. Williams w jednym z wywiadów, Rattle oferuje wszystkie algorytmy drążenia danych wykorzystywane w komercyjnym oprogramowaniu, jak również o wiele więcej[5]. Niewątpliwą zaletą nakładki na program R jest również jej prostota. Znajomość zaimplementowanych do programu testów (wiedza statystyczna na ten temat) jest absolutnie wystarczająca, by w ciągu paru-dziesięciu minut opanować zasady działania programu i posługiwać się nim. Zdecydowanie można polecić go do zastosowań edukacyjnych, naukowych i biznesowych. Wykorzystują go między innymi Australijski Urząd Podatkowy, Australijski Departament Imigracyjny, Toyota Australia, Ulster Bank oraz United States Geological Survey.

Rattle działa w środowisku Windows i Linux, jego instalacja jest prosta, odbywa się w konsoli programu R. Współpracuje z bazami danych w formacie tekstowym (*.txt i *.csv), plikami Excela, bazami danych używanymi przez program Weka ARFF (*Atribute Relationship File Format*), plikami otwartych łączy baz danych (*Open DataBase Connectivity*, ODBC) oraz rodzimymi plikami programu R (R Dataset, RData File). Zarówno zbiór danych jak też wyniki obliczeń można wyeksportować z programu w uniwersalnym formacie zapisu. Do programu (w zakładce *Help*) załączono obszerny podręcznik zawierający techniczny opis wszystkich funkcji nakładki. Więcej informacji na temat Rattle zawiera strona http://rattle.togaware.com/. Można tam odnaleźć podręczniki użytkownika programu, wskazówki dotyczące instalacji i radzenia sobie z mogącymi wystąpić wówczas błędami.

[4] Drążenie lub eksploracja danych jest to zespół metod statystycznych służących do badania dużych zbiorów danych (kilkanaście, a nawet kilkadziesiąt terabajtów). Metody te są stosowane odmiennie niż w klasycznej statystyce opisowej - w drążeniu danych hipotezy się formułuje, a nie testuje, nie znamy bowiem i nie zakładamy postaci zależności. Dzięki zastosowaniu algorytmów drążenia danych odkrywamy nieznany i niezałożony z góry wzorzec zależności. Elementarną orientację odnośnie tej fascynującej grupy metod daje artykuł: T. Demski, *Od pojedynczych drzew do losowego lasu*, w: *Zastosowania statystyki i data mining w badaniach naukowych*, StatSoft Polska, Kraków 2011, s. 65-75. Doskonałe wprowadzenie do metody data mining znajduje się w: P. Biecek, K. Trajkowski, *Na przełaj przez Data Mining*, w: http://www.biecek.pl/NaPrzelajPrzezDataMining, 2011, dostęp: kwiecień 2012.

[5] A. Ohri, *Interview: dr Graham Williams*, w: http://www.decisionstats.com/interview-dr-graham-williams/, dostęp: kwiecień 2012, 13 stycznia 2009.

2.1.2. SOFA (Statistics Open For All)

Program ten rozwijany jest od 2009 roku przez Paton-Simpson & Associates Ltd. Aplikację wydano na licencji AGPL (tj. Affer GPL - uzupełniającej oryginalną licencję GPL o rozstrzygnięcia dotyczące programów internetowych uruchamianych w systemie serwer - klient). Aplikacja funkcjonuje pod kontrolą Windows, Linux (dedykowana wersja dla Ubuntu) oraz Mac OS. Program jest zaawansowany pod względem interfejsu graficznego w porównaniu z innymi omawianymi aplikacjami. Z przejrzystego menu głównego (pierwszy z prezentowanych zrzutów ekranowych) możemy wybierać pomiędzy:

1/ *Get Started* - samouczkiem przeprowadzającym użytkownika przez podstawy obsługi programu,

2/ *Enter/Edit Data* - umożliwiającym ręczne tworzenie (lub modyfikowanie zaimportowanego) zbioru danych,

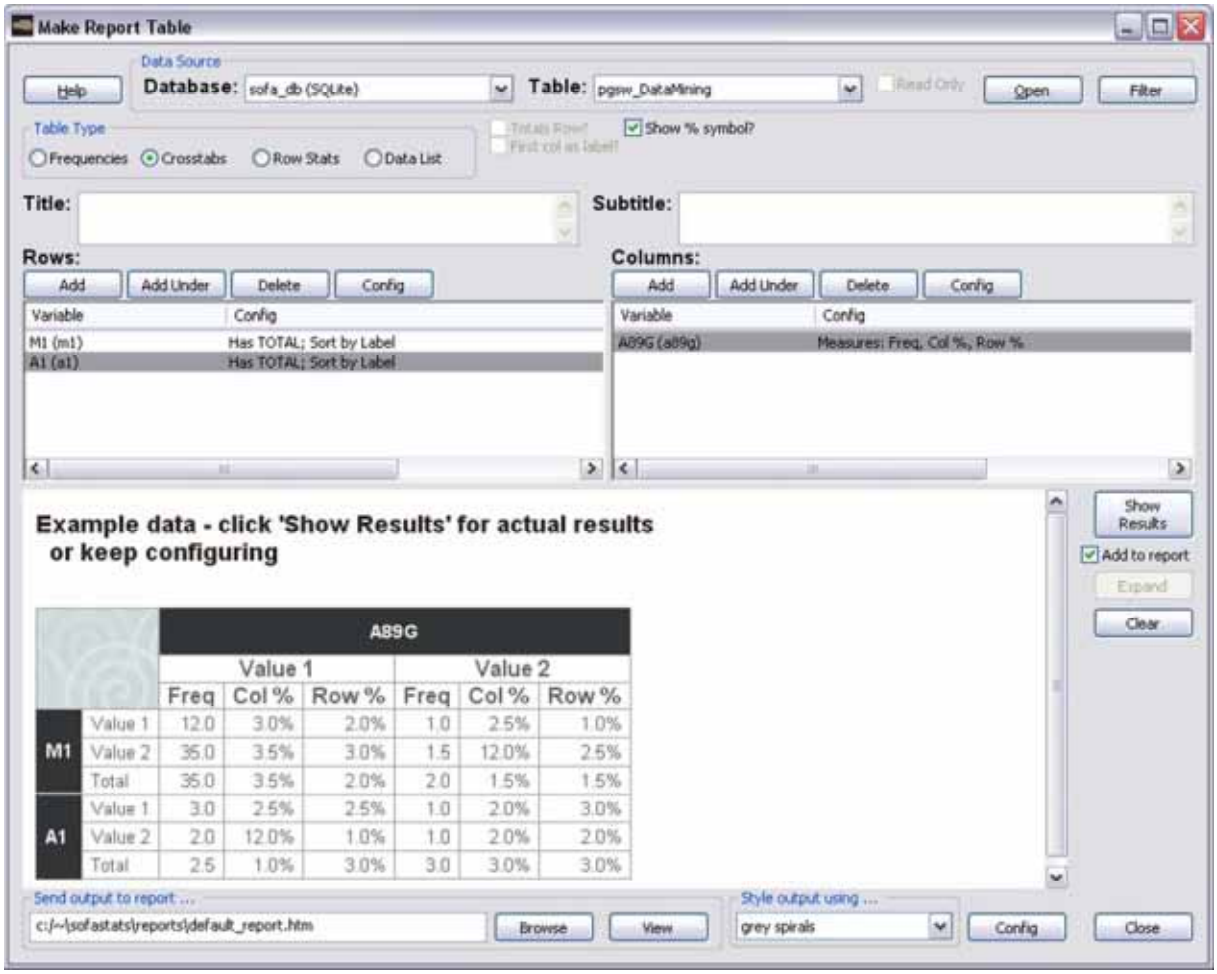
3/ *Import Data* - umożliwiającym importowanie danych do programu (z uniwersalnych formatów obsługiwanych jest *.csv oraz szereg formatów aplikacji bazodanowych: MySQL, Microsoft Access, SQLite, PostgreSQL, Microsoft SQL Server oraz format Google Docs),

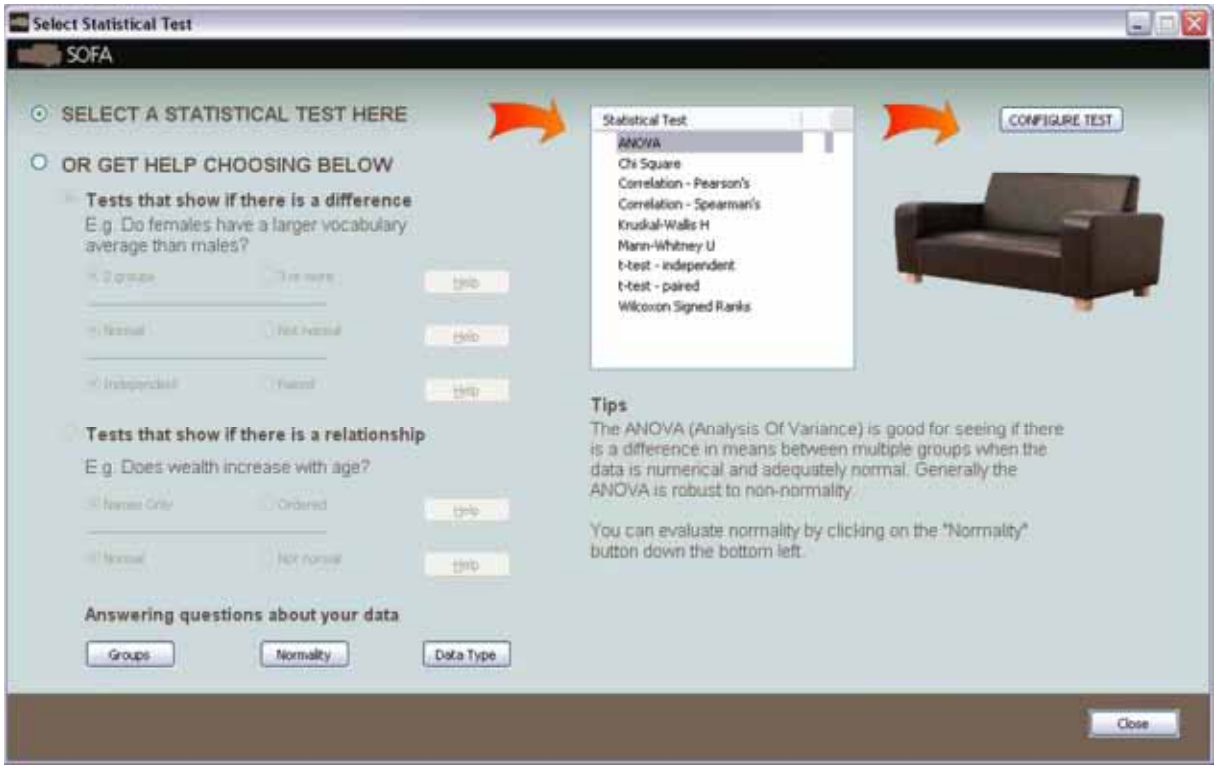
4/ *Report Tables* - pozwalającym na wykonywanie tabel z jedną zmienną i wieloma zmiennymi, wraz z możliwością filtrowania jednostek analizy,

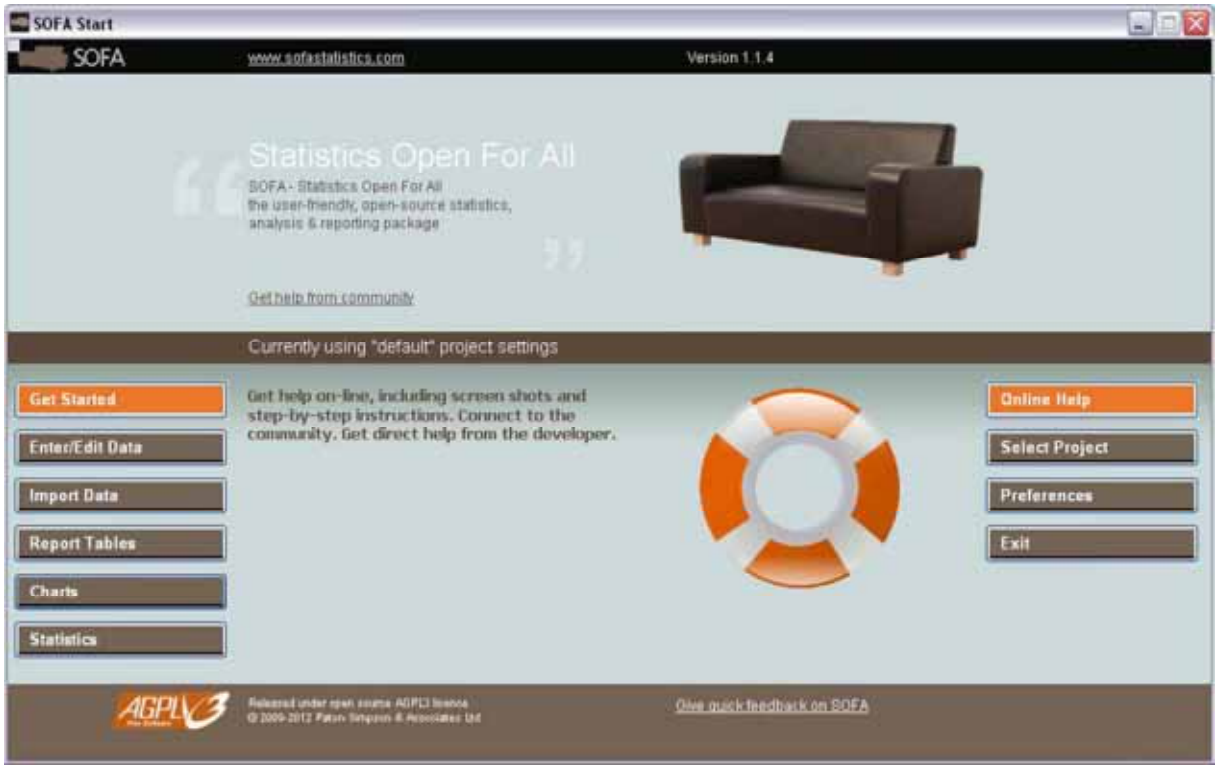
5/ *Charts* - zawierającym funkcje generowania diagramów, wykresów i grafów, wraz z ograniczoną modyfikacją ich wyglądu. Wykresy zawierają JavaScript tworzący ruchome części rysunków, co czyni je atrakcyjnymi jeśli zamieszczane są na stronach www,

6/ *Statistics*, w których zaimplementowano podstawowe statystyki parametryczne i nieparametryczne (zakładka *Statistics* jest widoczna na ostatnim z prezentowanych zrzutów ekranowych).

Warto polecić ten program, bowiem umożliwia analitykowi dynamiczną (i przyjemną ze względu na estetyczną prezentację wyników) pracę. Obsługa programu jest intuicyjna, rozwiązania mechaniki pracy cechują się wysoką ergonomią. Instalacja programu wymaga zainstalowania środowiska R dedykowanego specjalnie dla programu SOFA, a następnie samego programu. Program dostępny jest na stronie: http://www.sofastatistics.com/. Z tej samej strony można pobrać także dokumentację techniczną podręcznik, a także skorzystać z dobrze rozbudowanej bazy pomocy (dużym ułatwieniem jest między innymi instruktaż w formie kilkuminutowych filmów), w tym pomocy *online*.

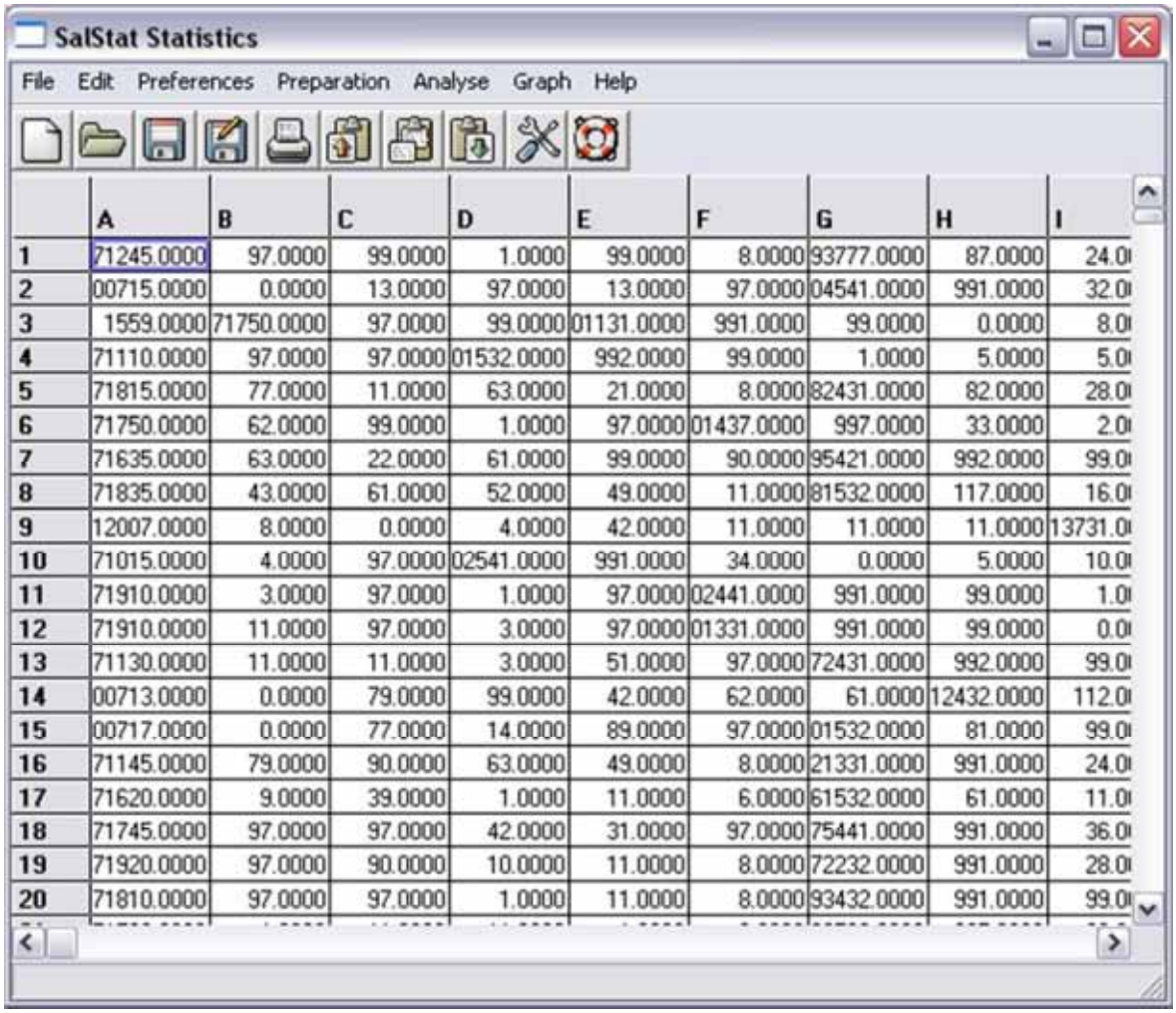



Select Statistical Test
SOFA
SELECT A STATISTICAL TEST HERE
OR GET HELP CHOOSING BELOW
Tests that show if there is a difference
E.g. Do females have a larger vocabulary average than males?
Tests that show if there is a relationship
E.g. Does wealth increase with age?
Answering questions about your data
Groups
Normality
Data Type
Statistical Test
ANOVA
Chi Square
Correlation - Pearson's
Correlation - Spearman's
Kruskal-Wallis H
Mann-Whitney U
t-test - independent
t-test - paired
Wilcoxon Signed Ranks
CONFIGURE TEST
Tips
The ANOVA (Analysis Of Variance) is good for seeing if there is a difference in means between multiple groups when the data is numerical and adequately normal. Generally the ANOVA is robust to non-normality.
You can evaluate normality by clicking on the "Normality" button down the bottom left.
Close


SOFA Start
SOFA
www.sofastatistics.com
Version 1.1.4
Statistics Open For All
SOFA - Statistics Open For All
the user-friendly, open-source statistics, analysis & reporting package
Get help from community
Currently using "default" project settings
Get Started
Enter/Edit Data
Import Data
Report Tables
Charts
Statistics
Get help on-line, including screen shots and step-by-step instructions. Connect to the community. Get direct help from the developer.
Online Help
Select Project
Preferences
Exit
AGPL

2.1.3. SalStat

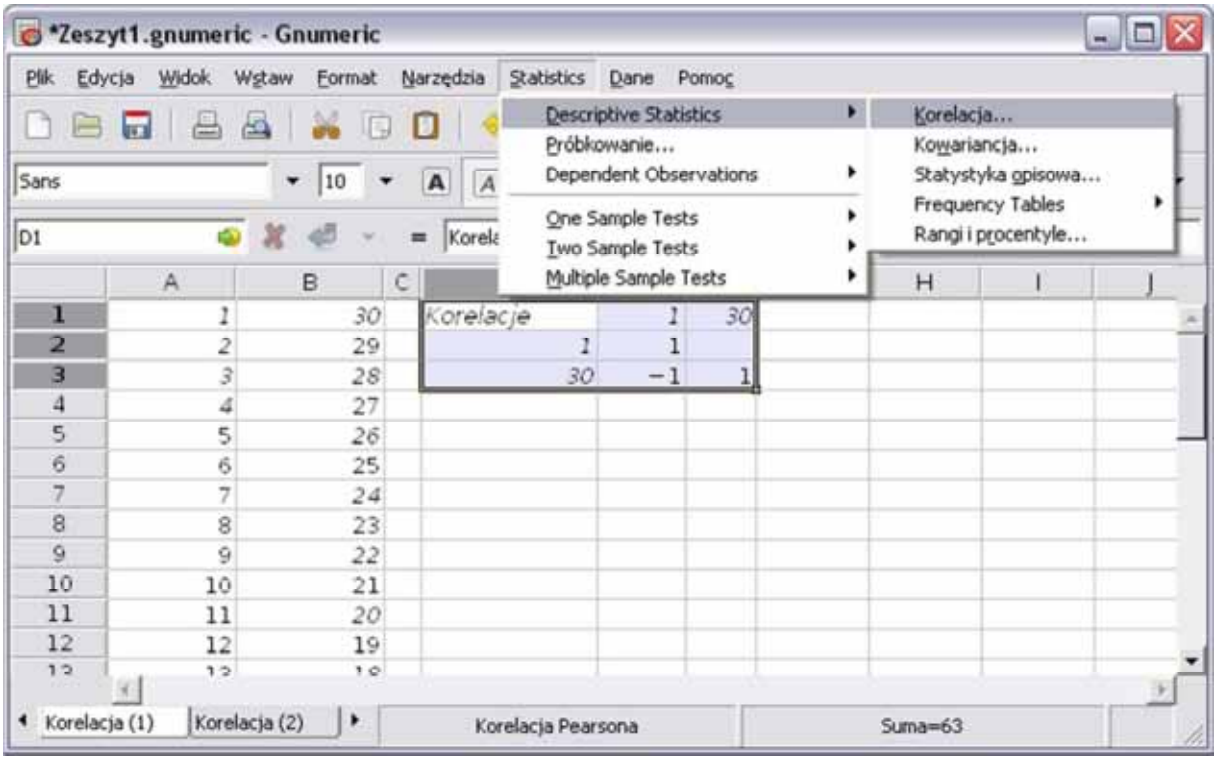
Pomysłodawcą tej aplikacji służącej do prostych analiz statystycznych jest Alan James Salmoni - doktor nauk informatycznych, Brytyjczyk zamieszkały na Filipinach. Pomysł napisania programu zrodził się w 2002 roku, gdy A.J. Salmoni nauczał studentów psychologii statystyki praktycznej i pragnął to czynić za pomocą prostej, niewymagającej zasobów aplikacji posiadającej podstawowe testy. Program został napisany w językach Python, wxPython, Numeric, SciPy i został wykonany na wzór SPSS, choć nie jest jej klonem w takim stopniu w jakim jest nim PSPP. Jest on bardzo uproszczony, nie zawiera funkcji zaawansowanego importu, eksportu i rekonfiguracji danych, pozwala jedynie na możliwość ich analizy za pomocą kilku prostych testów statystycznych (parametrycznych i nieparametrycznych). Wyniki przeprowadzonych testów można eksportować do formatu *.html. SalStat współpracuje z Windowsem, Linuxem i Mac OS. Program nadaje się do nauk podstaw statystyki, nie można niestety polecić go do typowych zastosowań biznesowych ze względu na brak możliwości prezentacji zestawień tabelarycznych. Program wraz dokumentacją można pobrać pod adresem: http://salstat.sourceforge.net/.



	A	B	C	D	E	F	G	H	I
1	71245.0000	97.0000	99.0000	1.0000	99.0000	8.0000	93777.0000	87.0000	24.0
2	00715.0000	0.0000	13.0000	97.0000	13.0000	97.0000	04541.0000	991.0000	32.0
3	1559.0000	71750.0000	97.0000	99.0000	01131.0000	991.0000	99.0000	0.0000	8.0
4	71110.0000	97.0000	97.0000	01532.0000	992.0000	99.0000	1.0000	5.0000	5.0
5	71815.0000	77.0000	11.0000	63.0000	21.0000	8.0000	82431.0000	82.0000	28.0
6	71750.0000	62.0000	99.0000	1.0000	97.0000	01437.0000	997.0000	33.0000	2.0
7	71635.0000	63.0000	22.0000	61.0000	99.0000	90.0000	95421.0000	992.0000	99.0
8	71835.0000	43.0000	61.0000	52.0000	49.0000	11.0000	81532.0000	117.0000	16.0
9	12007.0000	8.0000	0.0000	4.0000	42.0000	11.0000	11.0000	11.0000	13731.0
10	71015.0000	4.0000	97.0000	02541.0000	991.0000	34.0000	0.0000	5.0000	10.0
11	71910.0000	3.0000	97.0000	1.0000	97.0000	02441.0000	991.0000	99.0000	1.0
12	71910.0000	11.0000	97.0000	3.0000	97.0000	01331.0000	991.0000	99.0000	0.0
13	71130.0000	11.0000	11.0000	3.0000	51.0000	97.0000	72431.0000	992.0000	99.0
14	00713.0000	0.0000	79.0000	99.0000	42.0000	62.0000	61.0000	12432.0000	112.0
15	00717.0000	0.0000	77.0000	14.0000	89.0000	97.0000	01532.0000	81.0000	99.0
16	71145.0000	79.0000	90.0000	63.0000	49.0000	8.0000	21331.0000	991.0000	24.0
17	71620.0000	9.0000	39.0000	1.0000	11.0000	6.0000	61532.0000	61.0000	11.0
18	71745.0000	97.0000	97.0000	42.0000	31.0000	97.0000	75441.0000	991.0000	36.0
19	71920.0000	97.0000	90.0000	10.0000	11.0000	8.0000	72232.0000	991.0000	28.0
20	71810.0000	97.0000	97.0000	1.0000	11.0000	8.0000	93432.0000	991.0000	99.0

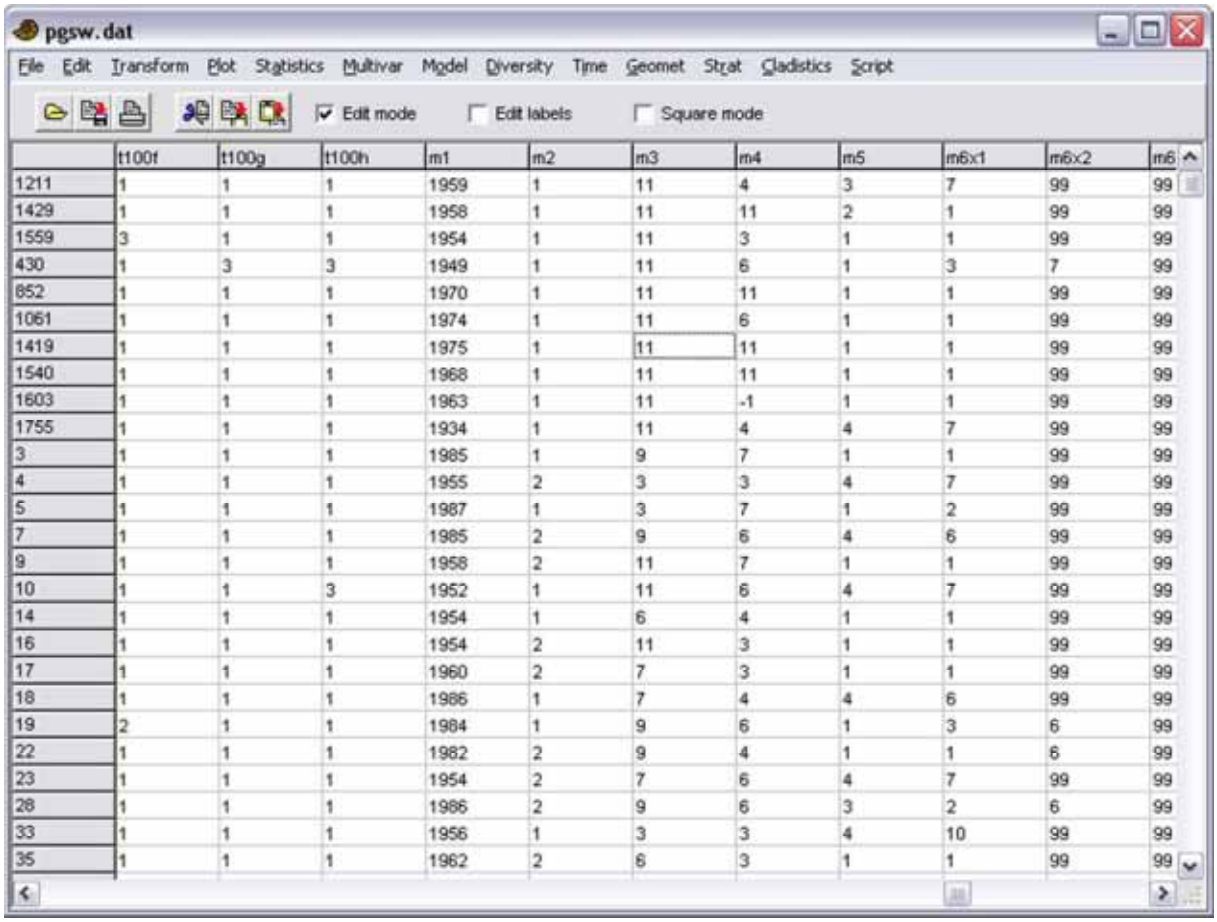
2.1.4. Gnumeric

Program ten jest rozwijany od 1998 roku; powstał on w celu zastąpienia komercyjnych arkuszy kalkulacyjnych ich darmowym odpowiednikiem. Rezultaty prac nad tą aplikacją są obecnie na tyle zadowalające, że w większości zastosowań warto zastanowić się na zastąpieniem komercyjnego arkusza kalkulacyjnego tym właśnie programem. Obecnie (2012) jest on dostępny w wersji 1.11.2 *aka* TBD. Program ten pozwala na otwieranie plików zapisanych we wszystkich niemal formatach arkuszy kalkulacyjnych i bazodanowych: MS Excel, Lotus 1-2-3, Applix, OpenOffice.org, Psion, Sylk, XBase, Oleo, PlanPerfect, Quattro Pro, a także HTML. Należy podkreślić, że dostępny jest w 46 wersjach językowych, w tym i polskojęzycznej. Gnumeric może posłużyć jako prosty program do zastosowań analitycznych. W zakładce *Statistics* znajdują się podstawowe statystyki - między innymi porównywanie średnich oraz podstawowa analiza korelacji i regresji, a także możliwość generowania prostych tabel dla jednej i wielu zmiennych. Gnumeric posiada również możliwość generowania zaawansowanej grafiki statystycznej. Więcej na temat arkusza kalkulacyjnego Gnumeric można odnaleźć na stronie: http://projects.gnome.org/gnumeric/, w tym także sam program, dokumentację oraz pomoc online (lista mailingowa oraz IRC). Zamieszczony poniżej zrzut ekranowy prezentuje możliwości analityczne arkusza Gnumeric - prostą korelację.



2.1.5. PAST (PAleontological STatistics)

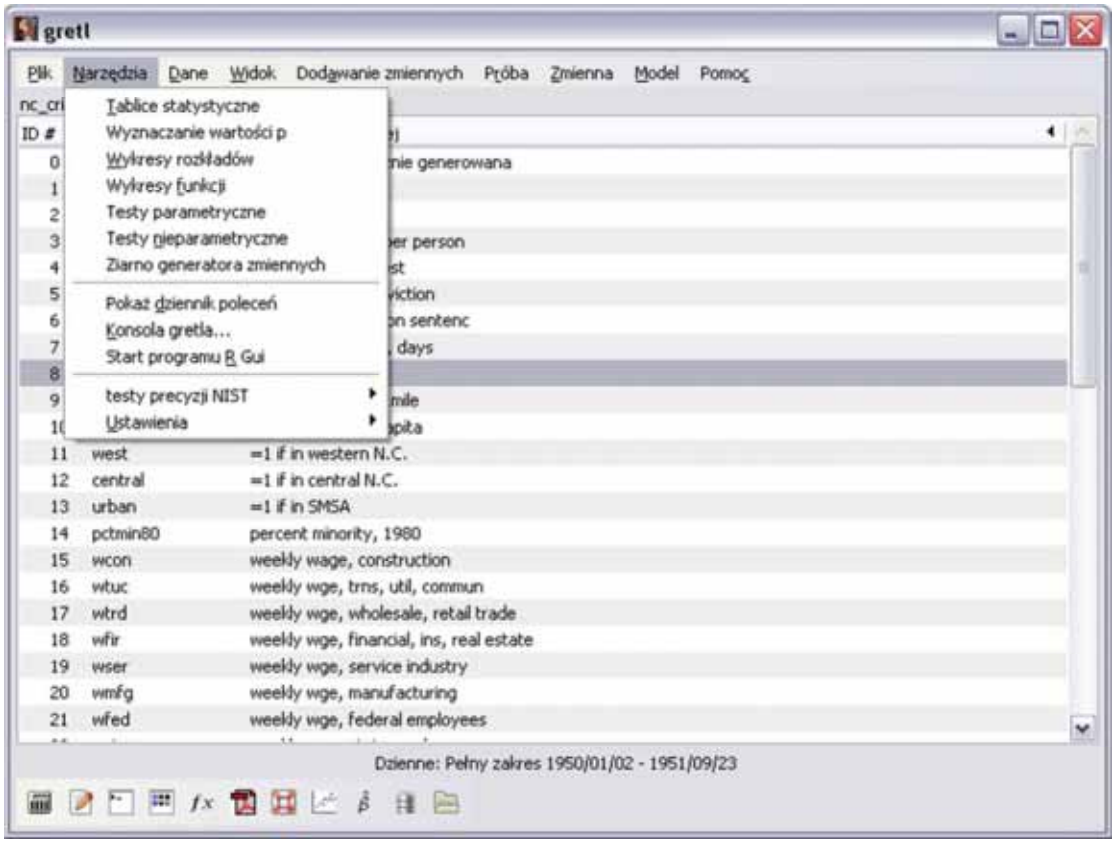
Pakiet statystyczny dedykowany paleontologom i ekologom, lecz pomimo tego warty polecenia również badaczom społecznym. Autorami programu są Øyvind Hammer z Muzeum Paleontologicznego na Uniwersytecie w Oslo, David A. T. Harper - pracownik Muzeum Geologicznego w Kopenhadze oraz Paul D. Ryan z National University of Ireland. Program działa bez konieczności instalacji, napisano go pod Windows, lecz pod kontrolą Wine (programu imitującego system Windows) może również działać w systemach Linux oraz Mac OS. Importować można do niego pliki danych takie jak *.dat *.nex *.nts oraz *.phy. Zadziwiają jego niewielkie rozmiary - w zaledwie 4 MB pomieściło się kilkadziesiąt rozmaitych testów statystycznych, w tym nader rzadko używanych w badaniach społecznych: analiza wielowymiarowa, dopasowanie krzywej, szeregi czasowe, algorytmy analizy filogenetycznej. Pakiet zawiera czternaście studiów przypadku (pliki danych oraz załączone do nich ćwiczenia), które stanowią kurs statystyki paleontologicznej, a jednocześnie są doskonałym instruktażem działania programu. Zaprezentowane poniżej okno programu przypomina PSPP lub SPSS, jednak z jednej strony zawiera o wiele więcej testów statystycznych, a z drugiej nie posiada funkcjonalności (dotyczących mechaniki interakcji z użytkownikiem) typowych dla tych programów. Program oraz podręczniki w języku angielskim można pobrać na stronie: http://nhm2.uio.no/norlex/past/.



	t100f	t100g	t100h	m1	m2	m3	m4	m5	m6x1	m6x2	m6
1211	1	1	1	1959	1	11	4	3	7	99	99
1429	1	1	1	1958	1	11	11	2	1	99	99
1559	3	1	1	1954	1	11	3	1	1	99	99
430	1	3	3	1949	1	11	6	1	3	7	99
852	1	1	1	1970	1	11	11	1	1	99	99
1061	1	1	1	1974	1	11	6	1	1	99	99
1419	1	1	1	1975	1	11	11	1	1	99	99
1540	1	1	1	1968	1	11	11	1	1	99	99
1603	1	1	1	1963	1	11	-1	1	1	99	99
1755	1	1	1	1934	1	11	4	4	7	99	99
3	1	1	1	1985	1	9	7	1	1	99	99
4	1	1	1	1955	2	3	3	4	7	99	99
5	1	1	1	1987	1	3	7	1	2	99	99
7	1	1	1	1985	2	9	6	4	6	99	99
9	1	1	1	1958	2	11	7	1	1	99	99
10	1	1	3	1952	1	11	6	4	7	99	99
14	1	1	1	1954	1	6	4	1	1	99	99
16	1	1	1	1954	2	11	3	1	1	99	99
17	1	1	1	1960	2	7	3	1	1	99	99
18	1	1	1	1986	1	7	4	4	6	99	99
19	2	1	1	1984	1	9	6	1	3	6	99
22	1	1	1	1982	2	9	4	1	1	6	99
23	1	1	1	1954	2	7	6	4	7	99	99
28	1	1	1	1986	2	9	6	3	2	6	99
33	1	1	1	1956	1	3	3	4	10	99	99
35	1	1	1	1962	2	6	3	1	1	99	99

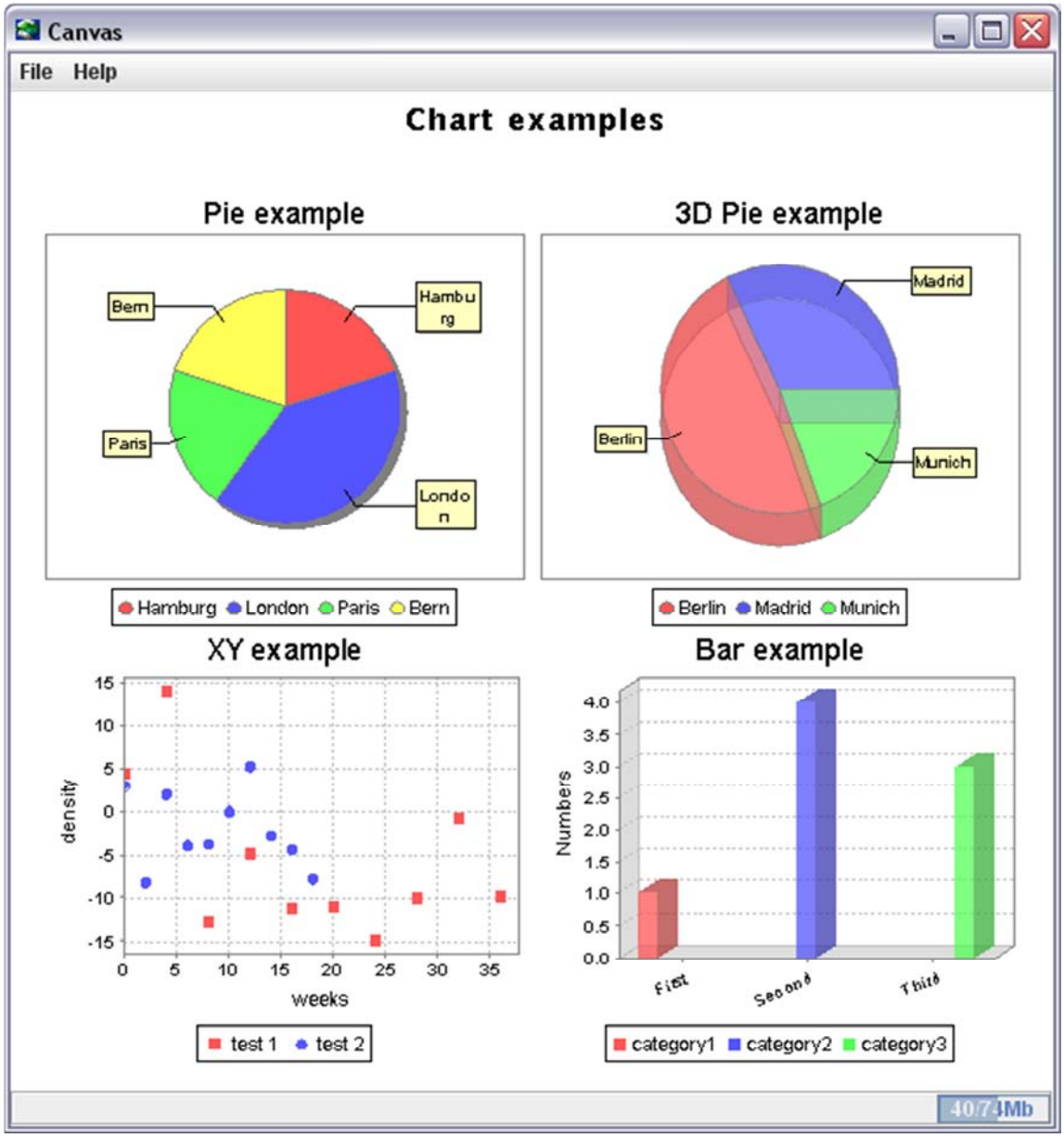
2.1.6. Gretl

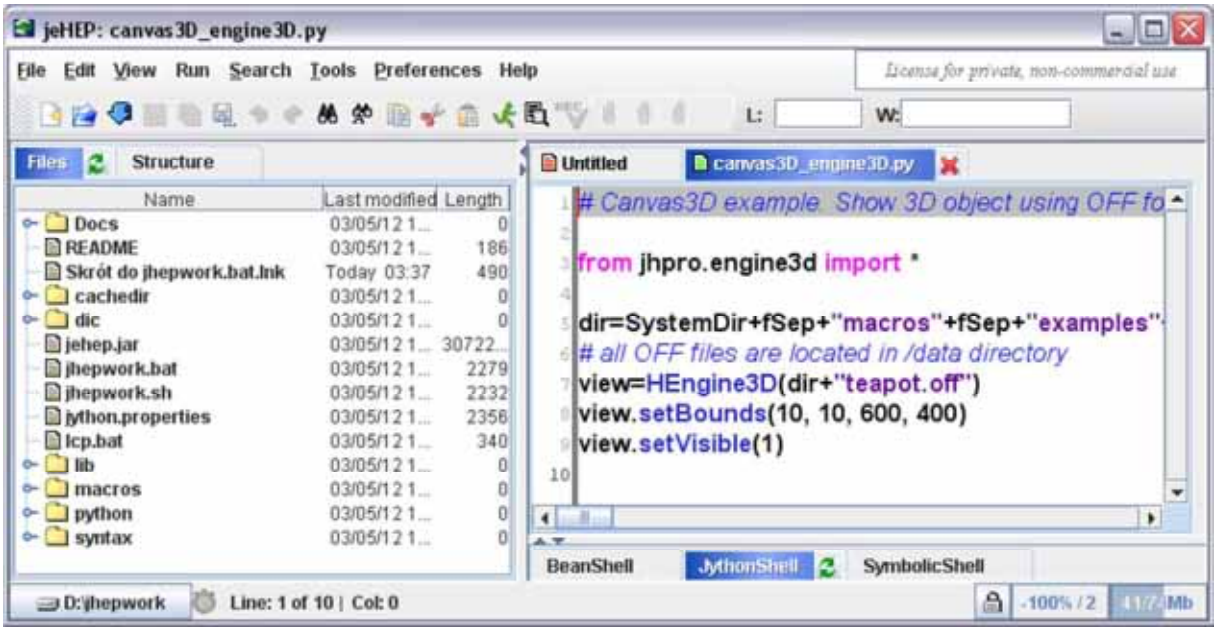
Autorami tego programu są dwaj ekonometrzy - Allin Cottrell z Uniwersytetu Wake Forest w Północnej Karolinie w Stanach Zjednoczonych i Ricardo „Jack" Luchetti z Università Politecnica delle Marche w Anconie we Włoszech. Jest to program przeznaczony przede wszystkim do zastosowań ekonometrycznych. Program współpracuje z językiem R, może być używany zarówno przez interfejs graficzny z przejrzystym i estetycznym menu okienkowym jak również w linii poleceń (to jest z użyciem konsoli). Intuicyjny interfejs graficzny, przejrzysty i łatwy do wyeksportowania tryb prezentacji wyników sprawiają, że pomimo typowo ekonometrycznego przeznaczenia programu warto go polecić również badaczom społecznym. Godnym pozazdroszczenia, wartościowym pod względem dydaktycznym rozwiązaniem jest dołączenie do programu licznych zbiorów danych analizowanych w podręcznikach ekonometrycznych, co umożliwia samodzielną analizę danych omawianych w literaturze przedmiotu. Program został bogato wyposażony w rozmaite funkcje szeregów czasowych i metody regresyjne. W programie można wygenerować estetyczne wykresy, grafy i diagramy. Szczególny nacisk położono na możliwość współpracy z różnymi formatami danych (tekstowymi, Eviews, Excel, Gnumeric, JMulti, Open Document, Stata, PSPP i SPSS, XML). Zaletą tego programu jest wielojęzyczność - jest on dostępny w języku angielskim, francuskim, hiszpańskim i niemieckim, a nawet w chińskim i albańskim. Opracowano także (Tadeusz Kufel i Paweł Kufel z Uniwersytetu Mikołaja Kopernika w Toruniu) wersję polskojęzyczną, co niewątpliwie jest atutem tego programu i szczególnie ważnym elementem decydującym o wyborze szczególnie dla początkującego użytkownika. Aplikacja współpracuje z programem R. Zwrócić uwagę należy także na kompaktowość tego programu - zajmuje on zaledwie 14 MB. Programu można używać w systemach operacyjnych Windows, Linux i Mac OS X. Komplet informacji, podręczników oraz sam program jest dostępny na stronie: http://gretl.sourceforge.net/, a także (w wersji polskojęzycznej): http://www.kufel.torun.pl/.



2.1.7. jHepWork (jWork)

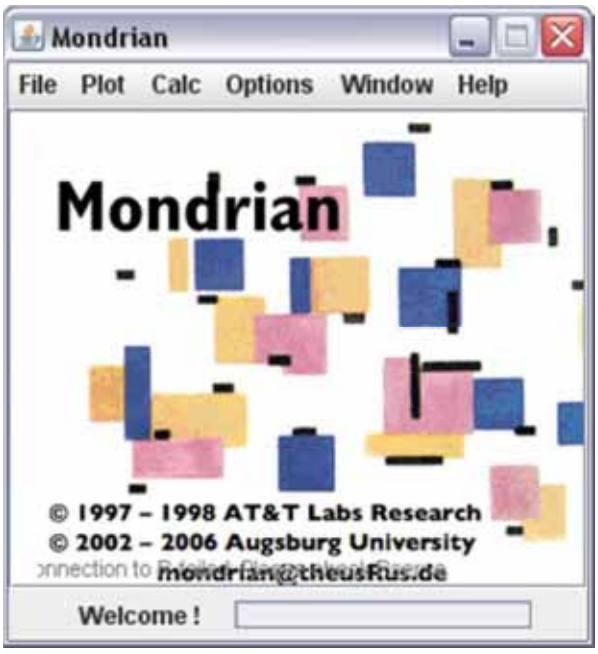
Program jest przeznaczony do analizy i wizualizacji danych. Jest to program na licencji częściowo wolnej, właściciele dopuszczają wykorzystywanie programu do celów dydaktycznych, lecz nie komercyjnych. W tym drugim przypadku program jest odpłatny. Uwaga dla niewtajemniczonych w informatyczne zawiłości Czytelników: program jHepWork instalujemy (po rozpakowaniu), uruchamiając dwukrotnym kliknięciem plik wsadowy DOS jhepwork.bat. Jest to program przeznaczony dla zaznajomionych ze sztuką programowania, nie polecany początkującym użytkownikom. Obszerna dokumentacja projektu znajduje się pod adresem: http://jwork.org/. Poniżej na zrzutach ekranowych zaprezentowano przykładowy efekt pracy programu w postaci wykresów oraz konsolę programu wraz z fragmentem kodu.

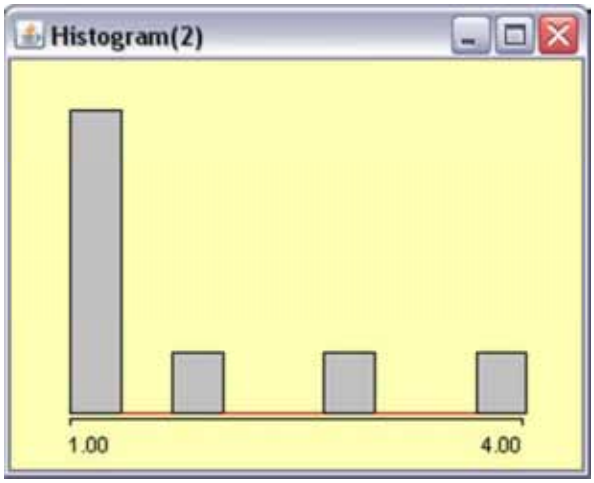


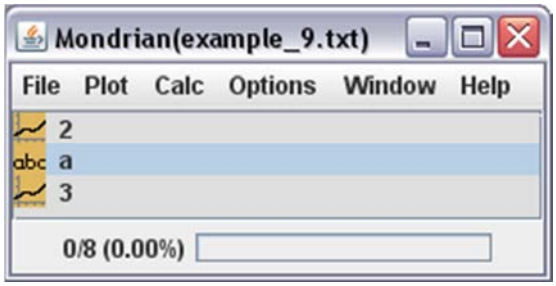


2.1.8. Mondrian

Niewielki, łatwy w instalacji, pracujący na dowolnej platformie systemowej program służący do wykonywania prostych wykresów, grafów i diagramów. Generowane przez program wizualizacje wykresów są konfigurowalne. Rozwijany był w latach 1997-1998 przez AT&T Labs Research a w latach 2002-2006 przez Uniwersytet w Augsburgu. Przydatność programu jest ograniczona - zawiera on wyłącznie funkcje graficzne, a wykonane wizualizacje można eksportować do innych aplikacji jako zrzuty ekranowe.

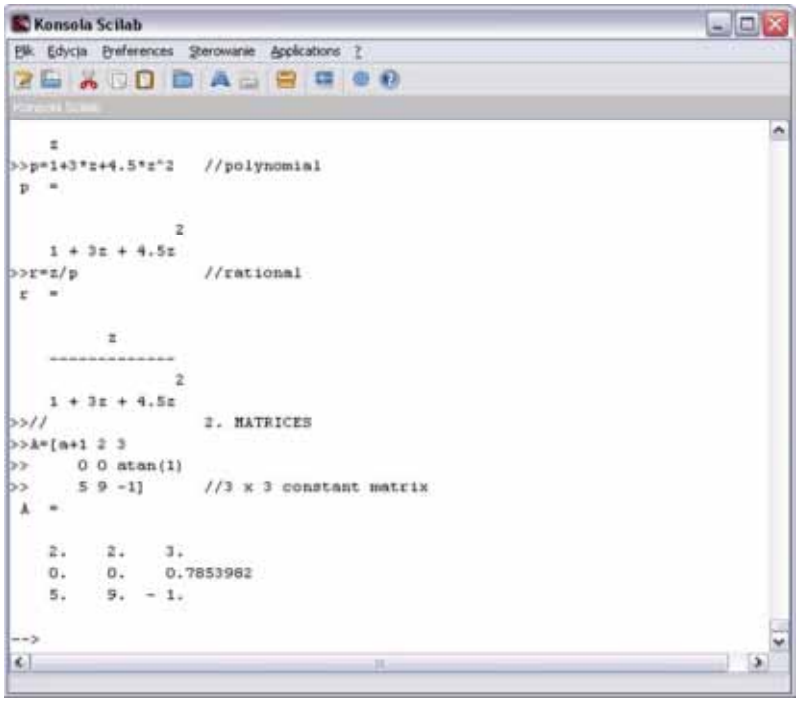


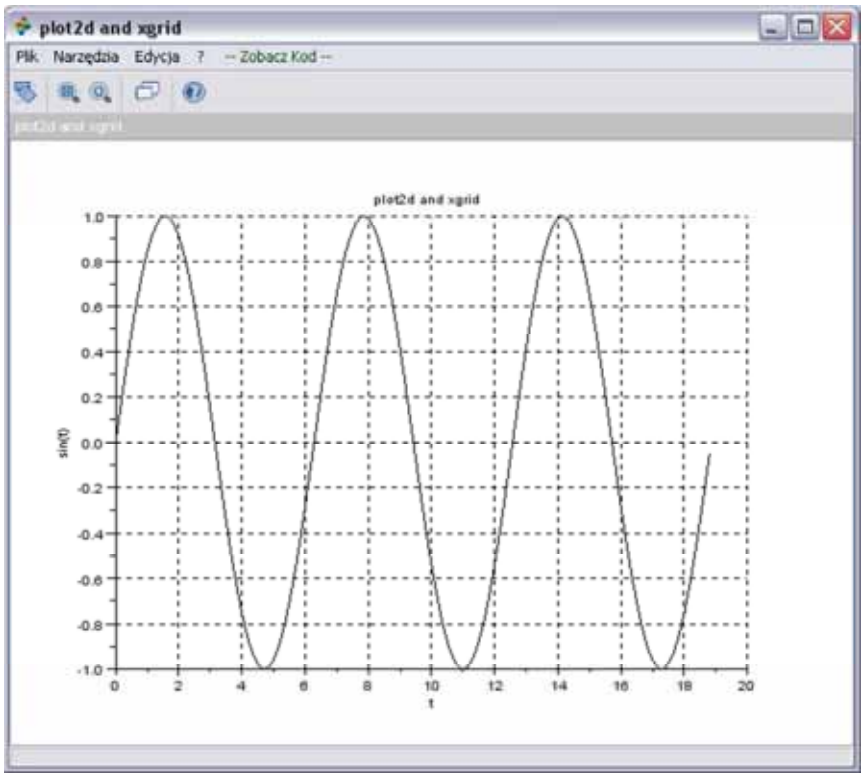




2.1.9. Scilab

Program SciLab powstał w 1989 roku obecnie rozwijany jest przez Scilab Consortium. Funkcjonuje na licencji CeCILL (*CEA CNRS INRIA Logiciel Libre*) zgodnej z licencją GNU/GPL. Jest to oprogramowanie przeznaczone przede wszystkim do zastosowań inżynierskich, wymaga co najmniej podstaw umiejętności programowania, bowiem aplikacja może być obsługiwana wyłącznie w trybie linii poleceń. Więcej na temat tego programu: http://www.scilab.org/. Poniżej zaprezentowano konsolę programu oraz efekt wykonanych w programie analiz.





* * *

Darmowe oprogramowanie wydaje się przede wszystkim użyteczne w zastosowaniach naukowych i dydaktycznych. Jego rola na tym obszarze jest nie do przecenienia. Szczególne znaczenie w obecnej sytuacji (gdy do zastosowań biznesowych używane jest głównie oprogramowanie komercyjne) mają wolne programy stanowiące klony aplikacji odpłatnych. Dzięki nim można niewielkim kosztem samodzielnie lub w sposób zinstytucjonalizowany przyswoić zasady działania programów, których znajomość jest przydatna i coraz częściej wymagana na rynku pracy. Warto również zauważyć, że niektóre z tych produktów już przewyższyły programy komercyjne - szczególnie dobrym przykładem jest Rattle, inne dorównały lub niedługo im dorównają, jak na przykład SOFA, czy PSPP.

2.2. Przegląd i ewaluacja oprogramowania komercyjnego

Oprogramowanie komercyjne wydaje się lepiej dostosowane do zastosowań biznesowych. Może być również użyte z powodzeniem do analiz typowo akademickich w pracach badawczych i promocyjnych, jednakże nie jest to w pełni zasadne rozwiązanie przede wszystkim z przyczyn finansowych. W niniejszym podrozdziale omówiono najpopularniejsze oprogramowanie komercyjne. W zestawieniu uwzględniono firmy, które udostępniają wersje ewaluacyjne programów. Pominięto jednocześnie oprogramowanie specjalistyczne, przeznaczone na przykład dla przyrodników (QSB+ i STORM) czy przedstawicieli nauk ścisłych (komercyjny MATHEMATICA, MATLAB czy ich wolne odpowiedniki Maxima, Ocatve, Sage), pomimo faktu, że ograniczone zastosowanie dla nauk społecznych miałby każdy z tych pakietów. Marginalnie potraktowano oprogramowanie do wizualizacji danych analitycznych - przedstawiono tylko przykładowe aplikacje, których (zarówno komercyjnych, jak też niekomercyjnych) jest znacznie więcej. Nie brano pod uwagę również wszystkich dostępnych na rynku arkuszy kalkulacyjnych (m. in. Quattro Pro arkusz kalkulacyjny rozwijany przez firmę Corel, Lotus 1-2-3 firmy IBM), pomimo że pozwalają one dokonywać analiz statystycznych. Tę funkcję arkuszy kalkulacyjnych zasygnalizowano w postaci wzmianki o Analysis ToolPak - dodatku do arkusza kalkulacyjnego Microsoft Excel. Pominięcie arkuszy kalkulacyjnych tłumaczone jest faktem, że ich stosowanie w pracy badawczej może być tylko tymczasowe i są one zdecydowanie niewystarczające do wykonania profesjonalnych analiz. Mogą one jednak pełnić z powodzeniem rolę wspomagania dla analizy i rekonfiguracji danych.

2.2.1. IBM SPSS Statistics

SPSS to akronim *Statistical Package for the Social Sciences,* co wyraźnie definiuje jego przeznaczenie i posiadane funkcje. Aktualnie oferowany jest w wersji 20. Obecna, pełna nazwa tego programu dla jego wersji 19. i 20. po przejęciu produktu przez firmę IBM brzmi: IBM SPSS Statistics Base. Poprzednie wersje - 15. i 16. nazywały się po prostu SPSS, wersja 17.0.1 - SPSS Statistics, a wersje - 17.0.3 i 18 - PASW Statistics. Program SPSS rozwijany jest od 1968 roku. Jako ciekawostkę warto wskazać, że inicjatorem powstania tego programu jest amerykański politolog Norman H. Nie, który jako jeden z pierwszych wraz z zespołem prowadzonym przez swojego mentora Sidneya Verbę badał zagadnienia partycypacji politycznej. Program dostępny jest w kilku językach europejskich i azjatyckich, w tym także w języku polskim. Przedmiot niniejszego omówienia stanowi wersja podstawowa (Base), jednak warto zwrócić uwagę na oferowane moduły dodatkowe programu takie jak: SPSS Statistics Developer umożliwiający programistom R i Python dostosowanie programu do potrzeb, SPSS Amos umożliwiający

modelowanie równań strukturalnych (metoda analizy wielozmiennowej włączająca liczne inne statystyki - między innymi konfirmacyjną analizę czynnikową, modele regresji i modele struktury kowariancyjnej i korelacyjnej), SPSS Text Analytics for Surveys służący do drążenia danych tekstowych, SPSS Sample Power oferujący zaawansowane wspomaganie doboru próby oraz SPSS Visualization Designer służący do tworzenia złożonej i dopracowanej grafiki statystycznej. Wszystkie te moduły można pobrać w wersji testowej 14-dniowej ze strony: http://www-01.ibm.com/software/analytics/spss/downloads.html. Moduł podstawowy programu SPSS Statistics działa w systemach operacyjnych Windows, Linux (Red Hat) oraz Mac OS X (nie wszystkie jednak zaawansowane moduły dostępne są dla systemów Linux i Mac.). W dwóch pierwszych wymienionych systemach możliwa jest instalacja wersji zarówno *desktop* jak też sieciowej. Program jest bardzo ciężki - SPSS 20 zajmuje w wersji instalacyjnej aż 919 MB.

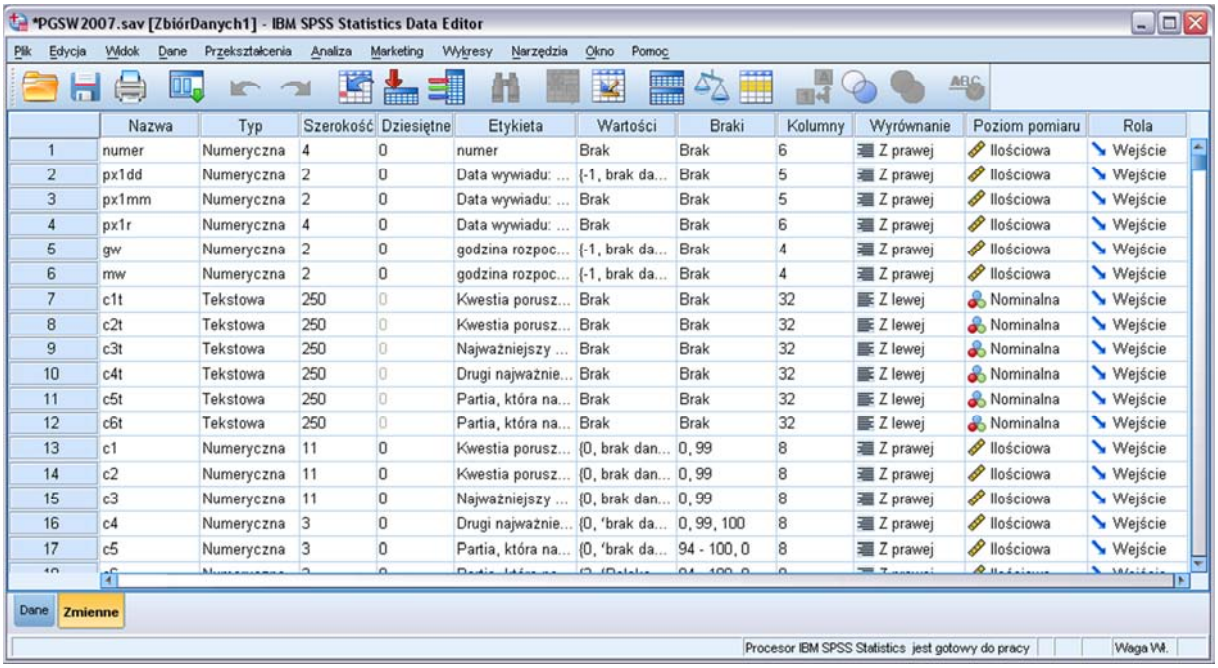
Program IBM SPSS Statistics Base umożliwia zaawansowaną pracę z bazami danych (rekonfigurację), bardzo bogate analizy (standardowe statystyki opisowe, statystyki parametryczne i nieparametryczne, a także sieci neuronowe, prognozowanie, analizę przeżycia. Ponadto dostępne są moduły analizy braków danych, imputacji danych, doboru próby oraz kontroli jakości. Na uwagę zasługuje także pozycja w menu zatytułowana *Marketing* - jest to ukłon w kierunku badaczy i praktyków marketingu bezpośredniego. Zakładka zawiera segmentację, profilowanie i scoring. Program ten może pracować zarówno w trybie okienkowym (wybór opcji odbywa się z menu i wywoływanych okienek), jak też w trybie składni (*syntax*). Ten ostatni sposób jest o wiele wydajniejszy i dlatego polecany. Analityk pracuje w czterech głównych oknach: widoku zmiennych i widoku danych (te dwa okna połączone są ze sobą) oraz okna raportu i (opcjonalnie) okna składni. Program współpracuje z najbardziej rozpowszechnianymi formami przechowywania danych: dBase, Excel, Lotus, SAS, Stata, Sylk, Systat, może pobierać także dane z plików tekstowych i hurtowni danych.

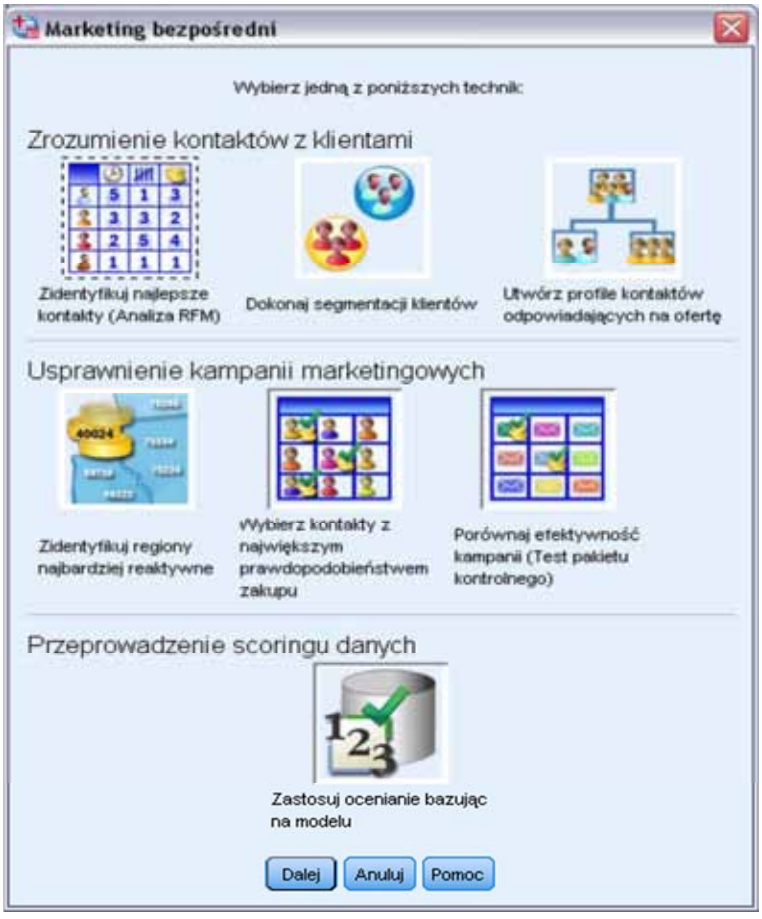
Program umożliwia szybką pracę, dobrze dostosowany jest zarówno do potrzeb biznesowych, jak też naukowych. Mankamentem jest brak współpracy programu z edytorami tekstu (niekompatybilność stwarza konieczność ponownego formatowania tabel).

Program ten jest szeroko rozpowszechniony w Polsce, w rozmaitych prywatnych i publicznych instytucjach (używany między innymi przez Główny Urząd Statystyczny, Departament Informacji Finansowej Ministerstwa Finansów oraz Komendę Główną Policji), a także powszechnie nauczany w szkolnictwie wyższym (między innymi w Uniwersytecie Warszawskim i Szkole Wyższej Psychologii Społecznej). Korzystają zeń liczne korporacje - między innymi Lloyds TSB, Spaarbeleg (członek AEGON Group), Standard Life, Wachowia (dawniej First Union).

Istnieją liczne podręczniki statystyki praktycznej wydanej pod patronatem SPSS Polska, powszechna jest również praktyka udostępniania licencji tego programu przez biblioteki uniwersyteckie, które dają możliwość instalacji oprogramowania na prywatnym komputerze studenta. Kwota rocznej (12 miesięcy) licencji najtańszej wersji programu przekracza nieco 4 000 PLN.

Poniżej zamieszczono widok zmiennych głównego okna programu SPSS oraz kartę tytułową modułu *Marketing bezpośredni.*



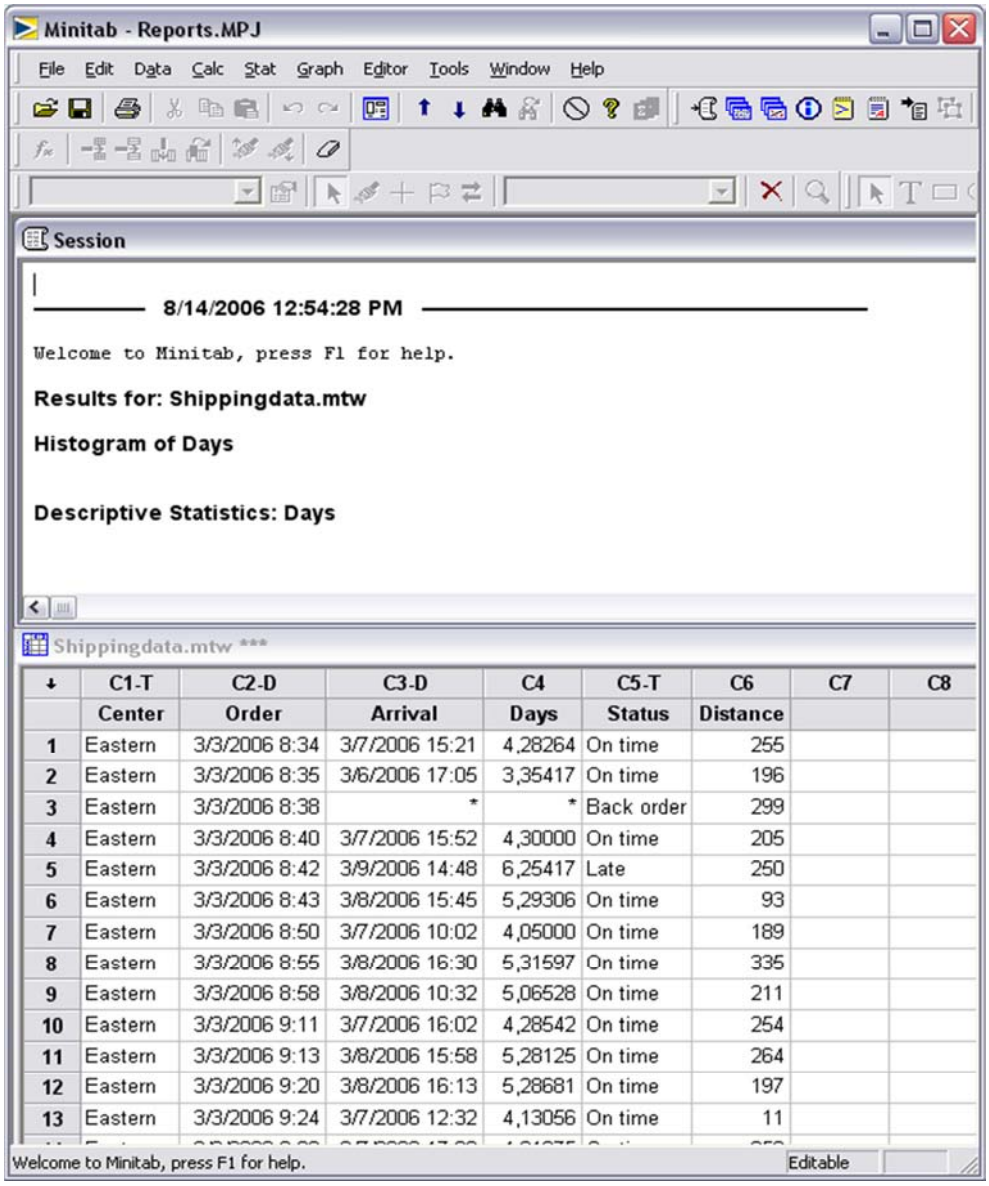


2.2.2. Minitab (Minitab Statistical Software)

Pakiet statystyczny Minitab jest produktem przedsiębiorstwa Minitab Inc. stanowiącego własność Pensylwania State University. Prace nad aplikacją rozpoczęta została w 1972 roku przez Barbara F. Ryan, Thomas A. Ryan Jr. i Brian L. Joiner. Obecnie program dystrybuowany jest na całym świecie w siedmiu wersjach językowych (w tym anglojęzycznej). Jest to najpopularniejszy pakiet statystyczny dla analityków *Six Sigma*[6].

W programie zaimplementowano bogaty zestaw rozmaitych funkcji statystycznych (statystyki opisowej, parametrycznej i nieparametrycznej, szeregów czasowych, a także obliczania wielkości próby oraz skalowania wielowymiarowego), dodatkowo pakiet zawiera starannie dopracowane opcje rekonfiguracji baz danych (zaawansowanego rekodowania i transpozycji) oraz rozbudowany moduł grafiki statystycznej. Obecnie (2012) dystrybuowana jest wersja 16 tego programu. W celu zaznajomienia się z aplikacją możliwe jest pobranie jej w pełni funkcjonalnej 30-dniowej wersji testowej (*trial*). Minitab został zaprojektowany dla środowiska Windows, możliwe jest jednak zainstalowanie go w Linux lub Mac OS z użyciem Wine. W programie dostępne są obszary robocze: widoku zbioru danych, wyników oraz komunikatów. Można również wywołać okno języka poleceń. Minitab jest zintegrowany z programami Windows: Eksplorerem, Notatnikiem i Kalkulatorem. Do Minitab można importować pliki popularnych formatów danych jak Excel, Quattro Pro, Lotus dBase ODBC, a także tekstowe (*.csv, *.dat, *.txt) oraz pliki XML. Dołączony przewodnik zawiera elementarne informacje na temat funkcji znajdujących się w programie oraz ich interpretacji. Niewątpliwą wadą programu jest trudność eksportu wyników - program nie współpracuje z bardziej zaawansowanymi edytorami tekstu, nie generuje też wyników w kodzie HTML. Zestawienia mają charakter tekstowy, należy włożyć nieco pracy, by dostosować je do minimalnego standardu estetycznego raportów analitycznych. Program dostępny jest na rynku w dwóch wersjach licencyjnych - akademickiej i biznesowej. Różnią się one jedynie kwotą opłaty licencyjnej. Roczna licencja dla pracowników nauki wynosi niecałe 100 PLN (23,37 Euro), a licencja wieczysta to koszt nieco ponad 300 PLN (77,92 Euro). W wersji biznesowej koszt licencji bezterminowej to 6051,42 PLN (przy większej liczbie jednoczesnego zakupu licencji uwzględniane są zniżki), a koszta *upgrade*'u do nowej wersji to 2421,82 PLN. Podręcznik *Minitab Handbook* pod redakcją Jona Cryera nie jest udostępniany bezpłatnie i kosztuje 333,39 PLN. Program (w tym 30-dniową wersję ewaluacyjną), dokumentację, podręczniki można pobrać ze strony http://www.minitab.com. Na stronie tej znajduje się również sklep, gdzie można dokonać zakupu oprogramowania. Program nie zawiera prostych funkcjonalności takich jak wyszukiwanie zmiennych z listy za pomocą liter lub przeciągania kolumn i wierszy w oknie widoku zmiennych. W przypadku pracy z dużymi zbiorami danych może być to czynnik utrudniający pracę.

[6] Sześć Sigma (Six Sigma) - jest to metoda zarządzania jakością. Celem jej wdrażania jest osiągniecie jakości „sześć sigma" (a więc wartości sześciu odchyleń standardowych). Została ona wprowadzona w korporacji Motorola w połowie lat 80. przez Boba Galvina (syna założyciela firmy) oraz Billa Smitha. Cyt. za: *Sześć sigma*, Encyklopedia Wikipedia, w: http://pl.wikipedia.org/wiki/Sześć_sigma, dostęp: kwiecień 2012.



	C1-T	C2-D	C3-D	C4	C5-T	C6	C7	C8
	Center	Order	Arrival	Days	Status	Distance		
1	Eastern	3/3/2006 8:34	3/7/2006 15:21	4,28264	On time	255		
2	Eastern	3/3/2006 8:35	3/6/2006 17:05	3,35417	On time	196		
3	Eastern	3/3/2006 8:38	*	*	Back order	299		
4	Eastern	3/3/2006 8:40	3/7/2006 15:52	4,30000	On time	205		
5	Eastern	3/3/2006 8:42	3/9/2006 14:48	6,25417	Late	250		
6	Eastern	3/3/2006 8:43	3/8/2006 15:45	5,29306	On time	93		
7	Eastern	3/3/2006 8:50	3/7/2006 10:02	4,05000	On time	189		
8	Eastern	3/3/2006 8:55	3/8/2006 16:30	5,31597	On time	335		
9	Eastern	3/3/2006 8:58	3/8/2006 10:32	5,06528	On time	211		
10	Eastern	3/3/2006 9:11	3/7/2006 16:02	4,28542	On time	254		
11	Eastern	3/3/2006 9:13	3/8/2006 15:58	5,28125	On time	264		
12	Eastern	3/3/2006 9:20	3/8/2006 16:13	5,28681	On time	197		
13	Eastern	3/3/2006 9:24	3/7/2006 12:32	4,13056	On time	11		

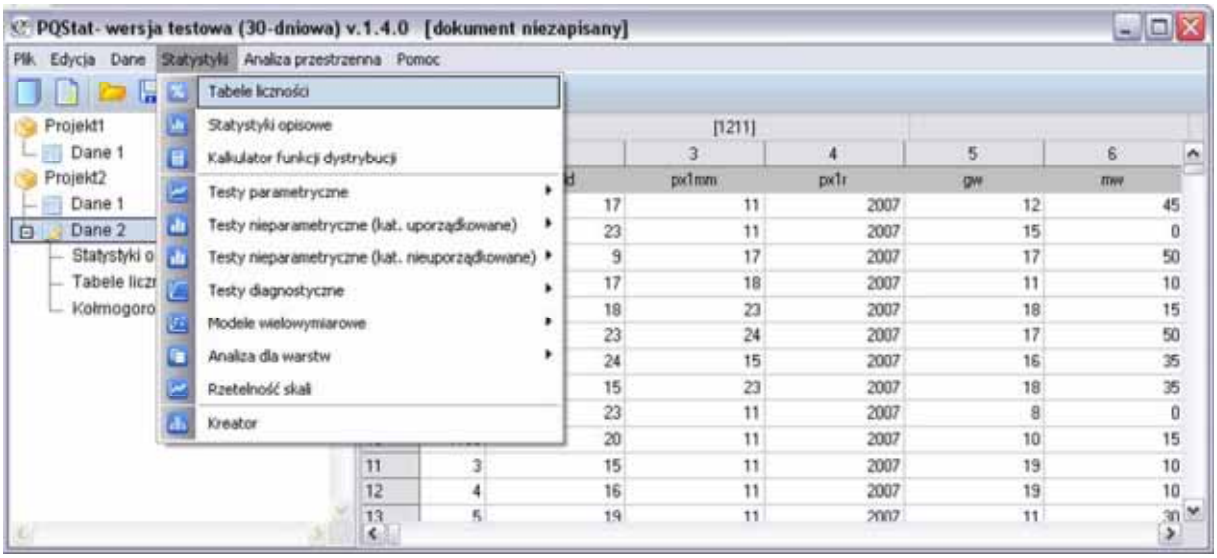
2.2.3. PQStat

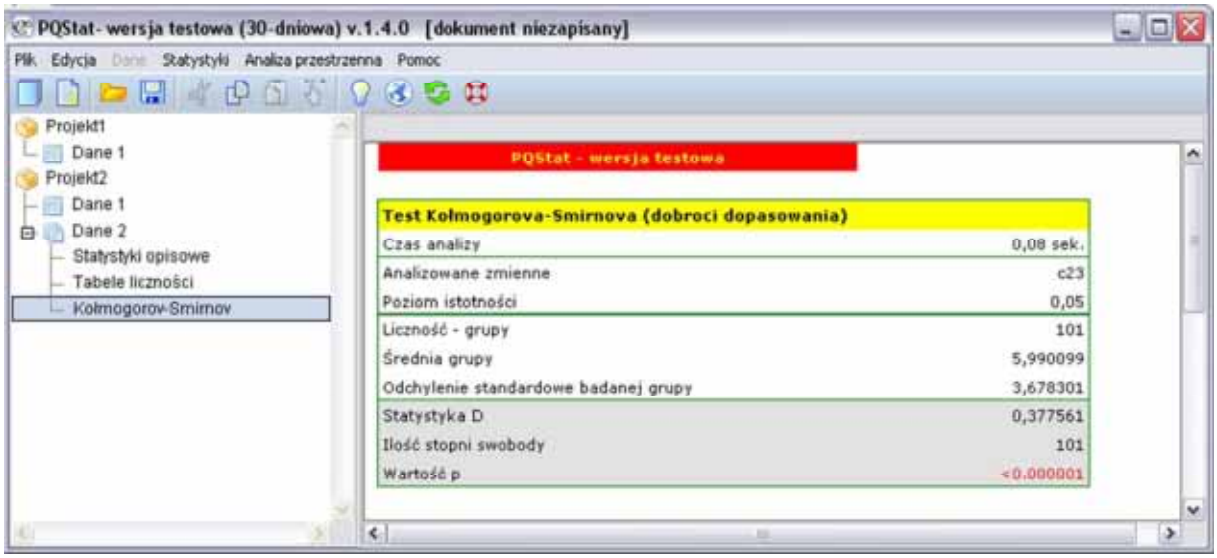
PQStat jest stosunkowo nowym produktem na rynku statystycznym, powstał w 2009 roku. Jest to produkt polski, wykonała go firma Tomasza Więckowskiego PQStat Software. Zaletą programu jest jego dostępność w wersji polskojęzycznej, a także uporządkowanie wszystkich dostępnych testów w przejrzystej formie (co utrwala właściwą systematykę testów statystycznych u początkującego użytkownika). Niedoświadczony użytkownik może skorzystać z zaimplementowanego do programu kreatora testów statystycznych - zostanie pouczony o zasadach ich wykonywania w przystępnej formie. Zwrócić należy również uwagę na rozbudowany system komunikatów uniemożliwiający wykonanie niezgodnych z zasadami testów statystycznych. Orientację pracującemu w programie ułatwia drzewo projektu umieszczone w lewym panelu programu dzięki któremu można łatwo przełączać się pomiędzy dostępnymi widokami: zbioru danych oraz raportów (każdy z wyników analiz prezentowany jest jako odrębny liść drzewa projektu). Do programu PQStat zaimplementowano możliwość generowania tabel częstości, statystyk opisowych, obliczania funkcji dystrybucji, przeprowadzanie testów parametrycznych i nieparametrycznych oraz diagnostycznych. Obecne są modele wielowymiarowe oraz udostępniono możliwość badania rzetelności skal.

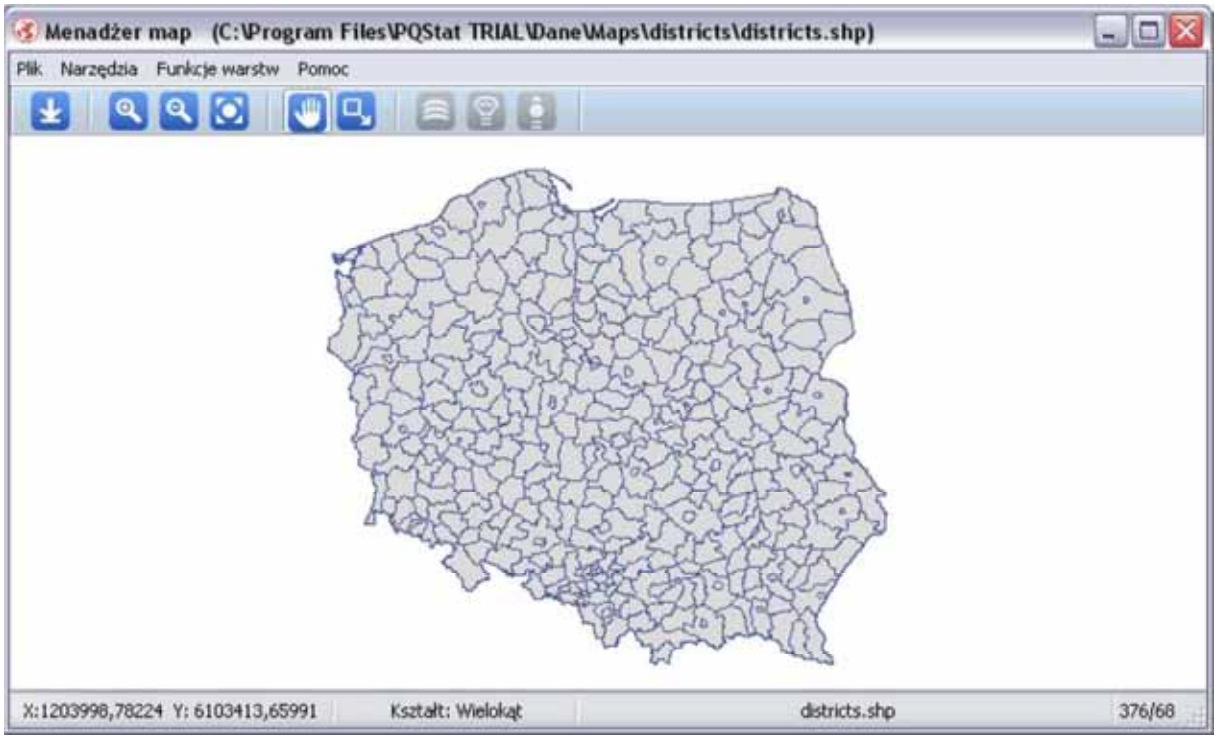
Opracowany został także moduł analizy przestrzennej - istnieje możliwość łączenia zebranych danych z koordynatami geograficznymi i nakładanie tych danych na mapy (praca odbywa się w odrębnym module programu Menedżerze map). Drobiazgowy opis (w języku polskim) elementów statystycznych i technicznych programu znajduje się na stronie internetowej (patrz: http://pqstat.pl/o_programie). Strona ta jest starannie przygotowana, bogata w informacje i często aktualizowana (można się z niej między innymi dowiedzieć, jakie ulepszenia planowane są w kolejnych wersjach programu i jakich postępów dokonano w pracy nad nimi). Program dystrybuowany jest w dwóch wersjach: edukacyjnej i komercyjnej, różniących się cenami. Jednostanowiskowa, wieczysta licencja edukacyjna to koszt 990 PLN brutto, a komercyjna - 1784 PLN. Aktualizacja wersji programu to odpowiednio 495 PLN dla wersji edukacyjnej i 892 PLN dla komercyjnej.

Program odczytuje pliki pochodzące z następujących źródeł: Excel, ESRI Shapefile, dBase, FoxPro oraz zwykłe pliki tekstowe (oddzielane znakami lub tabulatorami). Eksport wykonanych analiz może się odbywać wyłącznie w postaci danych tekstowych, co utrudnia dalszą pracę z wynikami (konieczność importowania tabel oraz formatowania ich). W zastosowaniach biznesowych utrudnieniem może być brak licznych drobnych funkcjonalności przyspieszających pracę analityka. Jest to między innymi brak możliwości ingerencji w zmienne na zasadzie ich przeciągania oraz niedostępne sortowanie w kolumnach po kliknięciu prawym przyciskiem myszy.

Do programu załączono obszerny podręcznik prezentujący nie tylko techniczne podstawy pracy z programem, ale również wprowadzający w podstawy statystyki. Program można pobrać w wersji testowej lub zakupić za pośrednictwem strony internetowej (http://pqstat.pl/wersja_testowa). Program jest dostępny w wersji dla Windows i Linux. Poniżej znajdują się zrzuty ekranów. Pierwszy z nich prezentuje widok zmiennych, drugi wynik raportów, a trzeci ukazuje menedżera map podczas pracy.





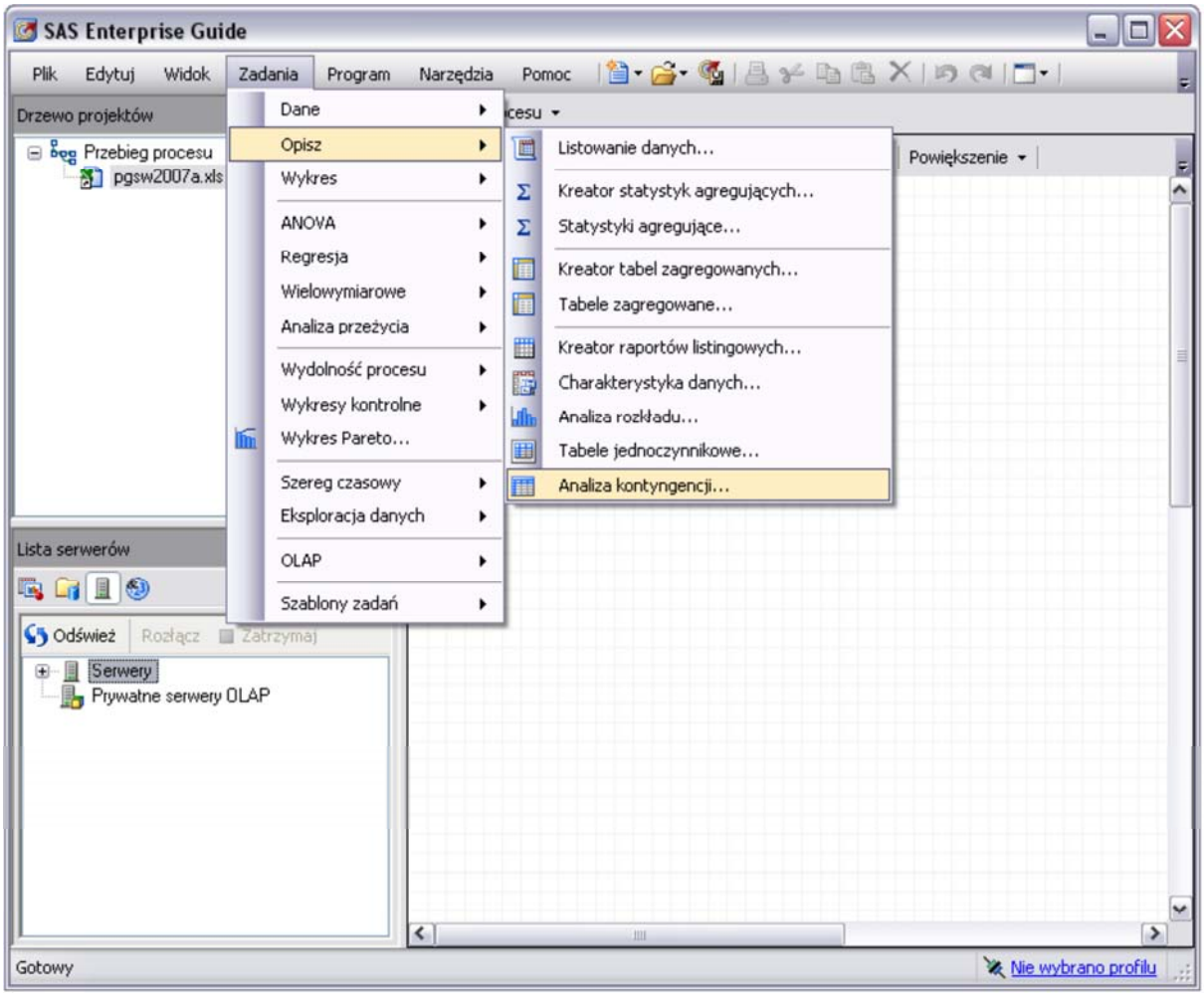


2.2.4. SAS Analytics Pro (Statistical Analysis System Analytics Pro)

Pakiet statystyczny SAS obecny jest na rynku już od 1976 roku. Początkowo rozwijano go na Uniwersytecie Stanowym w Północnej Karolinie w ramach grantów naukowych. Jest to właściwie grupa pakietów statystycznych: SAS Base, SAS/STAT oraz STAT/Graph. Trzy wymienione tworzą SAS Analytics Pro. Bardziej zaawansowaną opcję stanowi SAS Visual Data Discovery obejmujący dodatkowo SAS Enterprise Guide oraz JMP służące do zaawansowanej, interaktywnej wizualizacji danych. Podstawowym pakietem statystycznym jest SAS Enterprise Guide. Pakiet statystyczny SAS jest jednym z najbardziej rozbudowanych programów służących do analizy danych, podkreśla się jego innowacyjność i jakość[7].

[7] Patrz na przykład: *Materiały pomocnicze do studiowania statystyki. Wspomaganie komputerowe*, I. Kasperowicz-Ruka (red.), Szkoła Główna Handlowa w Warszawie, Warszawa 2004, s. 29.

SAS dostępny jest w kilku wersjach językowych - angielskiej, francuskiej, niemieckiej, hiszpańskiej, japońskiej, chińskiej i koreańskiej, a także polskiej. Instalacja programu SAS odbywa się w trybie okienkowym, jest dość złożona, należy zdefiniować liczne parametry konfiguracyjne, a jego pliki instalacyjne zajmują aż 13,2 GB. Base SAS nie zawiera procedur statystycznych - są one dostępne w SAS/STAT, SAS/QC, SAS/ETS. Dodatkowo dostępny jest SAS Enterprise Miner do data mining, SAS Forecast Server do automatyzacji prognozowania dużej ilości szeregów czasowych, SAS/IML - język macierzowy do tworzenia własnych procedur. Stanowi jeden z licznych produktów programistycznych amerykańskiej firmy. Aplikacja SAS Analytics Pro zawiera standardowe procedury statystyczne oraz moduł imputacji danych, a także zaawansowane możliwości prezentacji grafiki statystycznej (por.: http://www.sas.com/technologies/analytics/statistics/stat/index.html#section=3). Pod względem informatycznym oferta jest starannie przygotowana - program może pracować w systemach operacyjnych (wolnych i nie) wywodzących się z rodziny Unix, a także w systemie operacyjnym Windows. Licencja 12-miesięczna wersji jednostanowiskowej SAS Analytics Pro to koszt 22 764 PLN; odnowienie licencji to koszt zakupu pomniejszony o 28 procent. Na potrzeby edukacyjne i administracji publicznej ustalane są odrębne, bardziej przystępne cenniki.

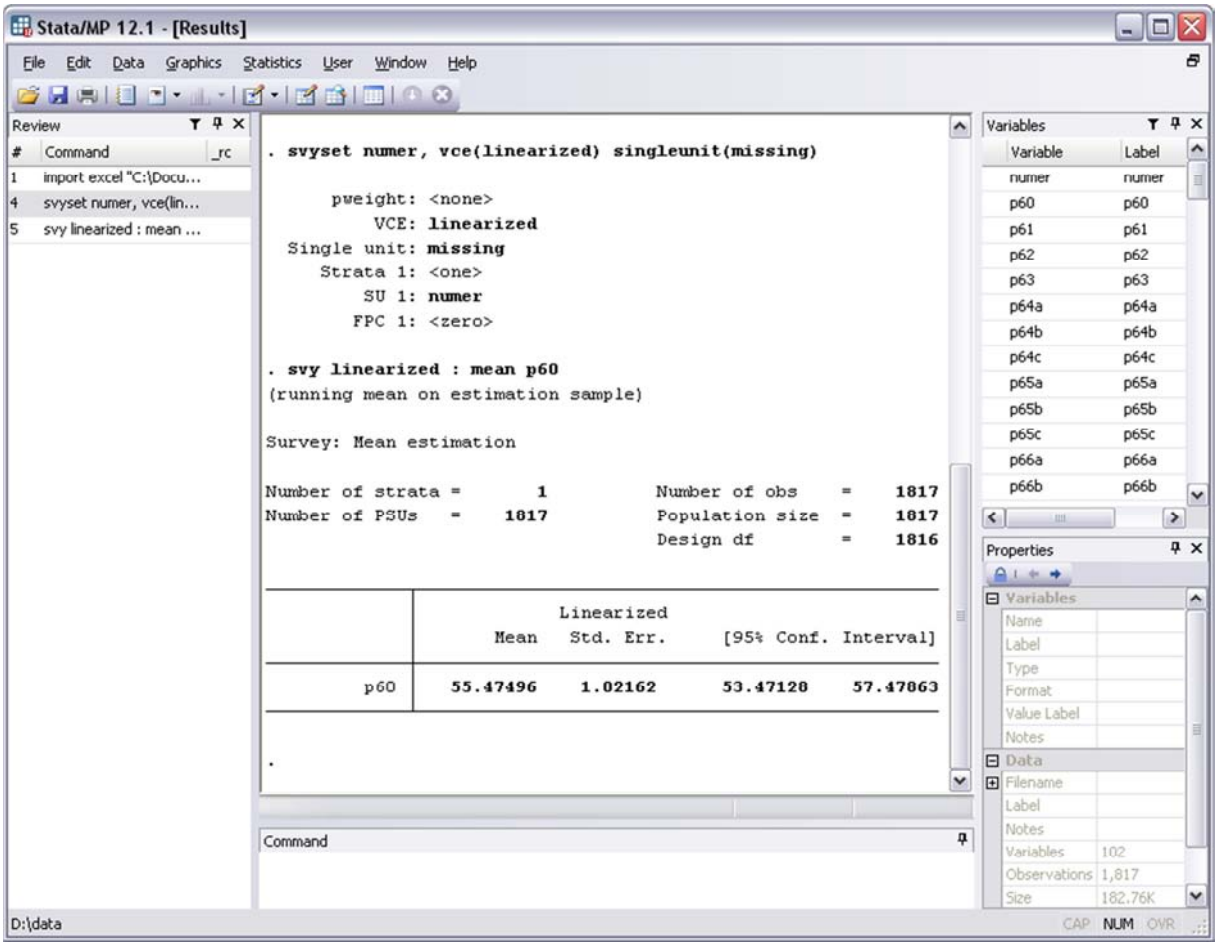


2.2.5. Stata

Program STATA (od 2011 roku w wersji 12) rozwijany jest od 1985 roku; jest on wielofunkcyjnym pakietem statystycznym przeznaczonym do zastosowań w ekonomii, naukach społecznych oraz epidemiologii. Nazwa tego programu to *portmanteau* dwóch anglojęzycznych słów: *statistics* (statystyki) i *data* (dane). Program Stata jest kompletnym pakietem statystycznym integrującym obliczenia i grafikę statystyczną. Pakiet zawiera setki narzędzi statystycznych od zaawansowanych, takich jak dynamiczna regresja danych panelowych (DPS), modele przeżycia, czy wielokrotna imputacja danych, aż do standardowych - jak na przykład statystyki opisowe, analizy ANOVA/MANOVA lub ARIMA (autoregresyjny zintegrowany proces średniej ruchomej). Program pozwala na zaawansowane zarządzanie zmiennymi, między innymi grupowe zmiany nazw zmiennych. Cieszy również funkcjonalny podział menu: wszystko co potrzebne do analizy danych sondażowych znalazło się w jednej zakładce. Ciekawym rozwiązaniem jest zgromadzenie w jednej zakładce wszystkich funkcji przydatnych w analizie danych panelowych lub trackingowych. Należy również podkreślić, że generowana przez program grafika ma standard nadający się do profesjonalnych biznesowych i naukowych publikacji.

Program - jak większość pakietów statystycznych - pracuje w dwóch trybach: okienkowym oraz języka składni (prostego, intuicyjnego, opatrzonego solidnym podręcznikiem). Odrębnymi oknami programu są menedżer zmiennych oraz edytor danych, a głównym oknem jest okno komunikatów programu będące jednocześnie miejscem, gdzie prezentowane są wyniki (widoczne na załączonym zrzucie z ekranu). W oknie tym pojawiają się odnośniki, co umożliwia na przykład szybkie dotarcie do odpowiedniej strony dołączonego do systemu podręcznika w celu uzyskania porady.

Program funkcjonuje pod kontrolą systemów Windows Linux i Mac, a jego instalacja jest intuicyjna, szybka i bezproblemowa. Stata występuje w kilku odmianach: Stata/MP przeznaczonej dla komputerów wieloprocesorowych, Stata/SE dedykowanej do obsługi wielkich baz danych oraz odmiana standardowej - Stata/IC i studenckiej - Small Stata. Koszt jednostanowiskowej licencji programu Stata/MP w wersji komercyjnej wynosi około 7500 PLN w opcji bezterminowej lub około 4 000 w opcji licencji rocznej. Program Stata jest tańszy na potrzeby edukacyjne; jego koszt wynosi odpowiednio około 4000 PLN i około 2000 PLN. Warto nadmienić, że Stata prowadzi także rozwiniętą działalność publicystyczną - wydawany jest kwartalnik „Stata Journal", a wydawnictwo Stata Press oferuje liczne pozycje interesujące dla analityków używających nie tylko programu Stata.

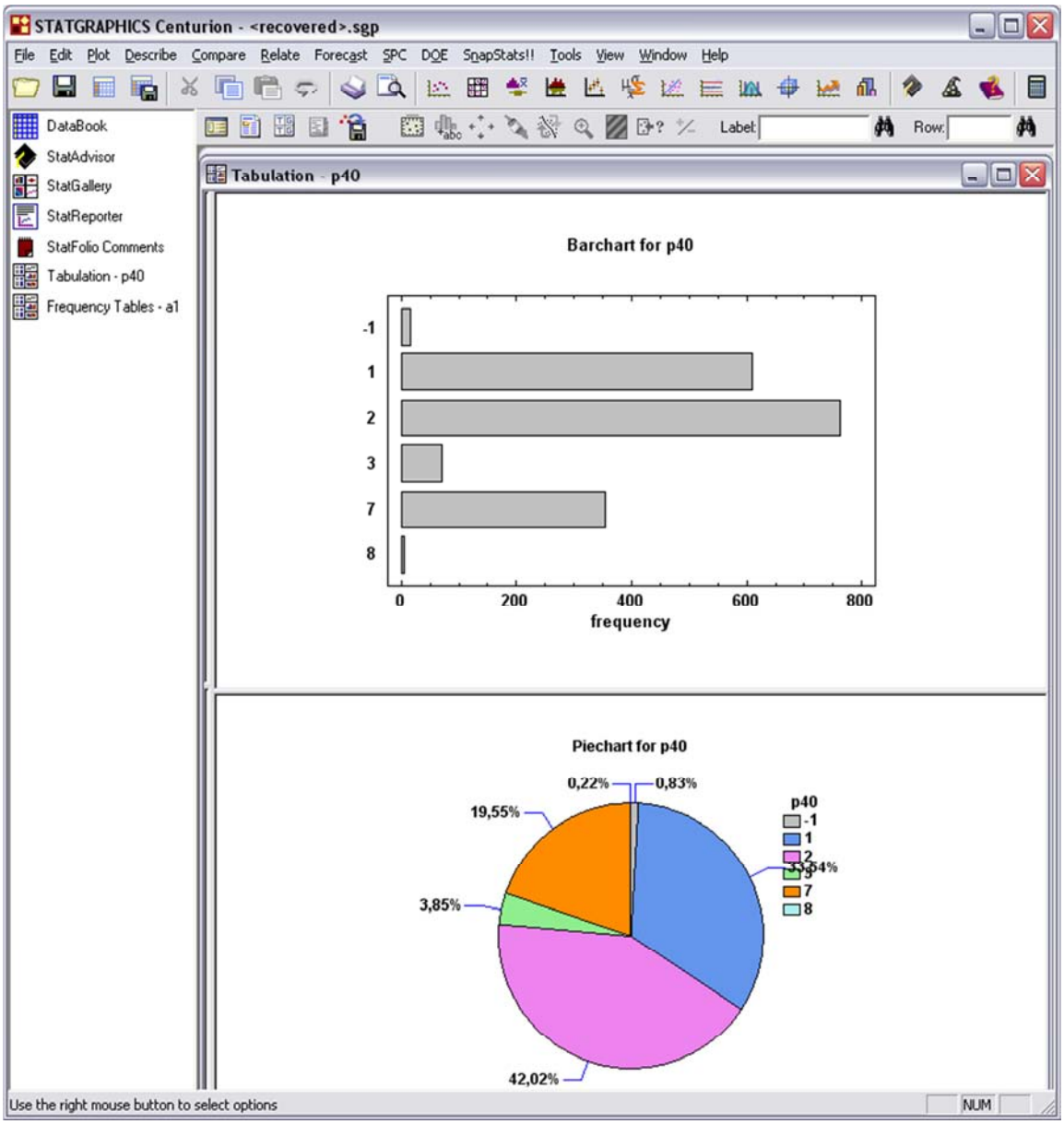


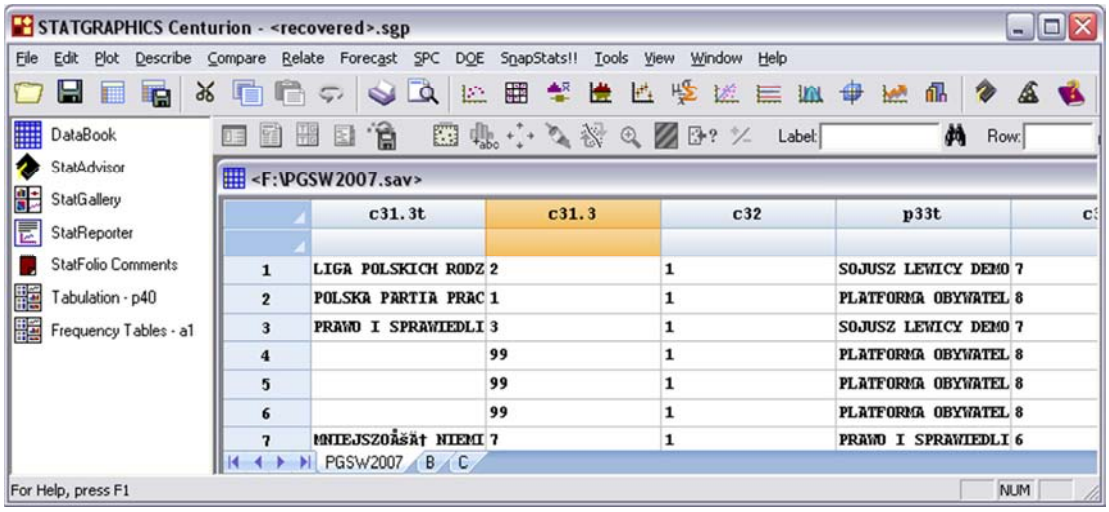
2.2.6. Statgraphics Centurion (Statistical Graphics System Centurion)

Jest to pakiet statystyczny rozwijany od 1980 roku przez StatPoint Technologies Inc. Program udostępniany jest w pięciu wersjach językowych: angielskiej, francuskiej, hiszpańskiej, niemieckiej i włoskiej, obecnie w wersji XVI. Program dostępny jest wyłącznie dla systemu Windows, dodatkowym atutem firmy jest wersja tego programu dla urządzeń mobilnych (Statgraphics Mobile). Można go używać w dwóch wersjach menu: standardowego dla pakietów statystycznych lub przeznaczone do analiz *Six Sigma*.

Program umożliwia współpracę z macierzami danych zapisanymi w niemal dowolnym formacie - między innymi - Excela, Lotusa, Matlab, SAS, SPSS, S-Plus, a także R. Zaimplementowano przydatne i oryginalne funkcje rekonfiguracji bazy danych (na przykład możliwości łatwego i dowolnego porządkowania kolumn zbioru danych), jednak zrezygnowano z innych, na przykład filtrowania całości zbioru danych, jego dzielenia i dołączania zmiennych. Ciekawym rozwiązaniem jest prezentowanie zbioru danych w formie arkusza kalkulacyjnego o wielu zakładkach. Umożliwia to dzielenie zbioru danych na mniejsze części oraz uporządkowaną pracę na zróżnicowanych danych. Wadą programu jest przystosowanie do analiz przede wszystkim na danych o charakterze liczbowym, wartość zmiennej nie jest zespolona z jej etykietą, konieczne jest zatem generowanie opisów zmiennych jako kolejnych zmiennych w arkuszu danych. Jest to niewygodne, a ponadto zbiór danych wskutek takiego zabiegu przestaje być przejrzysty i łatwy do zarządzania. Program zawiera wiele funkcji, przede wszystkim jednak ekonometrycznych. Oprogramowanie Statgraphics mieści się w średnim przedziale cenowym - wersja akademicka aplikacji kosztuje 2 210 PLN (32-bitowa) lub 2 655 PLN (64-bitowa). Aktualizacja programu do nowej wersji to

wydatek około tysiąca złotych. Program na potrzeby biznesowe kosztuje od 3896 PLN wzwyż. Nieco wyższe (kilkaset złotych) opłaty pobierane są za dwu- lub wielojęzyczne wersje programu. Stosunkowo tanie są licencje dla posiadających prawa studenckie - licencja sześciomiesięczna to wydatek rzędu około 100 PLN, a licencja obejmująca cały okres studiów kosztuje około 300 PLN. Jest to program potężny, dopracowany, jednak niedopasowany do potrzeb badaczy społecznych ani do prostych analiz w badaniach opinii, rynku i marketingowych - jest to zaawansowane narzędzie przeznaczone przede wszystkim dla analityków biznesowych. Więcej informacji na jego temat można znaleźć na stronie: http://www.statgraphics.com/. Program ten w latach dziewięćdziesiątych był względnie popularny na polskim rynku analitycznym, ukazało się wówczas kilka publikacji zwartych na temat pracy z pakietem Statgraphics.





2.2.7. Statistica

Pakiet północnoamerykańskiej korporacji StatSoft Inc. rozwijany od 1989 roku, obecnie (2012) dystrybuowany w wersji 9, przeznaczony jest wyłącznie dla systemu operacyjnego Windows. Program dostępny jest w kilkunastu wersjach językowych, w tym polskiej i rosyjskiej. Pakiet zawiera zestaw statystyk wymagany przez badaczy społecznych, ponadto statystyki przemysłowe i biomedyczne, a także dopracowany i wielofunkcyjny moduł grafiki statystycznej. Należy podkreślić, że program jest dopracowany pod względem technicznym i graficznym. W programie rozwinięto liczne mikrofunkcjonalności pozwalające pracować ze zbiorem danych na etapie przygotowania go do analiz. Ułatwiają one pracę analitykowi i przyspieszają ją. Są to przede wszystkim okienkowe funkcje porządkowania kolejności zmiennych w bazie danych. Zarządzanie projektem badawczym w programie jest przejrzyste - użytkownik operuje w dwóch oknach: zbioru danych oraz raportów. Wyniki analiz tabelarycznych można pobrać i umieścić w raportach tekstowych w formie statycznych obrazów (nadają się one do profesjonalnego druku, bowiem można ustawić w programie opcje rozdzielczości koniecznych dla drukarni).

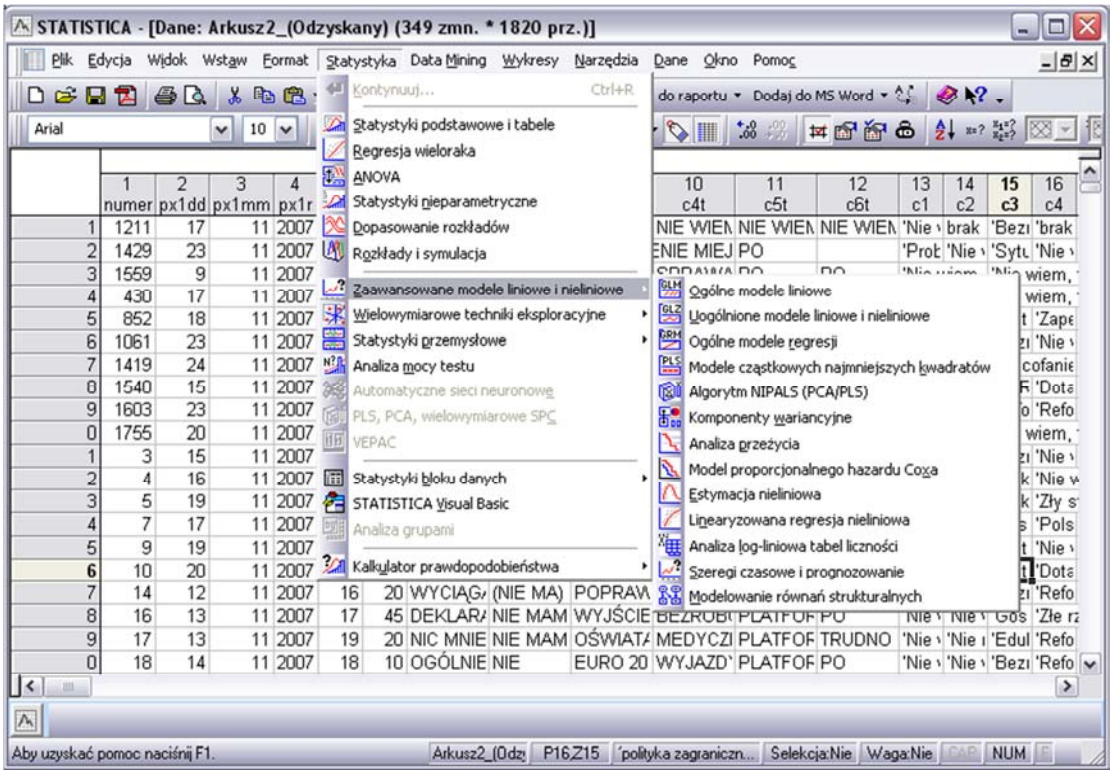
Zwracają również uwagę dodatkowe (konieczność poniesienia odrębnych opłat licencyjnych) moduły pakietu. Zawierają one oprogramowanie służące do drążenia danych liczbowych i tekstowych oraz ilościowej eksploracji stron internetowych. Wszystkie te własności czynią z programu unikalny produkt dla badaczy społecznych. W programie Statistica możliwa jest pełna integracja ze środowiskiem programu R za pomocą zautomatyzowanego darmowego programu STATCONN DCOM. W programie dostępna jest dokumentacja techniczna opisująca zasady pracy tych dwóch pakietów statystycznych. Ponadto możliwe jest programowanie w języku Visual Basic oraz R. Pakiet Statistica wyróżnia się także możliwością wbudowywania jego funkcji w inne aplikacje (za pomocą Java lub C++), a także bardzo duża precyzja obliczanych danych.

StatSoft oprócz dystrybucji programu równolegle prowadzi działalność szkoleniową i konsultingową. Ponadto, na zamówienie firmy powstają opracowania dotyczące analiz statystycznych z użyciem pakietu Statistica (także w języku polskim). Są one przeznaczone dla użytkowników na różnych poziomach znajomości statystyki i samego programu, pełnią rolę edukacyjną i promocyjną. Bogata oferta publicystyczna autorstwa wieloletnich praktyków i uczonych umożliwia samodzielną naukę pakietu oraz statystyki.

Warto zwrócić również uwagę na dostępny *online* rozbudowany i wartościowy merytorycznie *Elektroniczny Podręcznik Statystyki* dostępny pod adresem: http://www.statsoft.pl/textbook/stathome.html[8].

Pakiet Statistica w wersji podstawowej na licencji nielimitowanej w czasie kosztuje blisko 4 000 PLN brutto, a w rozszerzonej - niecałe 7 000 PLN brutto. Obok tego świadczona jest odpłatna usługa serwisowania.

Przykładowy zrzut ekranu z rozwiniętym menu *Statystyka* zaprezentowano poniżej.



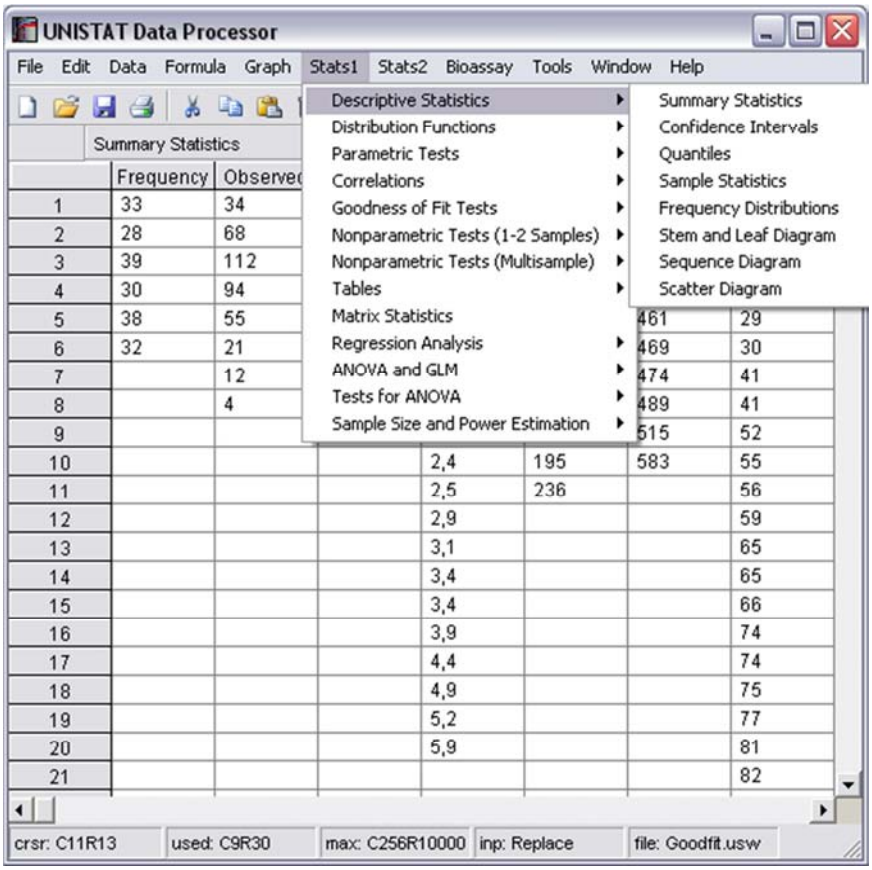
2.2.8. Unistat Statistical Package

Unistat jest brytyjskim, rozwijanym od 1984 roku pakietem statystycznym, obecnie oferowany w wersji 6.0 (edycja z 2011 roku). Najbardziej istotną charakterystyką programu jest fakt, że blisko współpracuje z arkuszem kalkulacyjnym Excel (wydania z 1997-2003 oraz 2007-2010). Program zawiera standard pakietów statystycznych: statystyki parametryczne, nieparametryczne, opisowe, rozwinięty moduł regresji, analizę wielowymiarową, analizę przeżycia oraz szeregi czasowe i kontrolę jakości. Aplikację można pobrać do przetestowania w wersji próbnej, lecz nie jest ona w pełni funkcjonalna można używać jedynie danych dołączonych do programu, nie jest możliwe importowanie i obróbka własnych danych. Unistat układem menu przypomina SPSS, jednak arkusz programu posiada odmienną logikę - ułożenie danych następuje w formie szeregu statystycznego, a nie macierzy danych jak w PSPP lub SPSS (choć i tak dane można układać). Opcje rekonfiguracji zbioru danych są nader uproszczone - jest to poważny mankament programu. Możliwa jest współpraca aplikacji z makrami w Visual Basic

[8] StatSoft, *Elektroniczny Podręcznik Statystyki PL*, Kraków 2006, w: http://www.statsoft.pl/textbook/stathome.html, dostęp: kwiecień 2012.

na zasadach identycznych jak w pakiecie Microsoft Office lub Libre Office. Jest to program działający wyłącznie w systemie Windows.

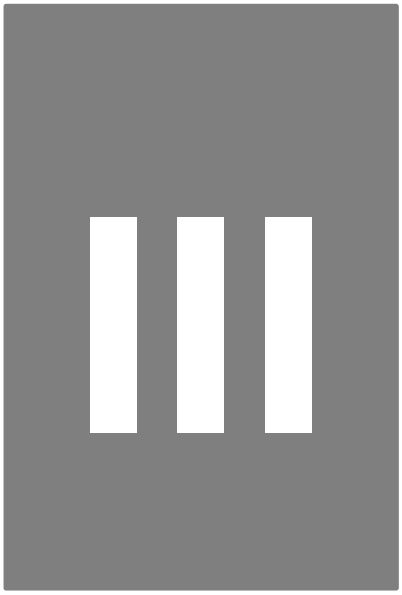
Program oferowany jest w dwóch edycjach: *Light* oraz *Standard.* Wersja *Light* nie zawiera modułów: skalowania wielowymiarowego oraz szeregów czasowych. Nie pozwala również na programowanie w Visual Basic. Ponadto w wersji tej maksymalna wielkość macierzy danych jest ograniczona. Jednostanowiskowa wersja tego programu kosztuje w wersji *Light* nieco ponad 1 000 PLN, zaś w wersji *Standard* - niecałe 4 000 PLN. Dla potrzeb edukacyjnych ceny ustalane są indywidualnie. Szersze informacje na temat programu znajdują się na stronie: http://www.unistat.com/download.



2.2.9. Excel Analysis ToolPak

Analysis ToolPak nie jest typowym pakietem statystycznym, jest to dodatek do programu Microsoft Office Excel. Zestaw narzędzi statystycznych jest udostępniany obecnie bezpłatnie przez Microsoft Excel. Z pomocą pakietu możliwa jest analiza wariancji ANOVA, regresja i korelacja, kowariancja, statystyki opisowe, wygładzanie wykładnicze, test F, t oraz z, analiza Fouriera, histogram, średnia ruchoma, generowanie liczb losowych, rangi i n-tyle, próbkowanie. Jest to stosunkowo najtańsze rozwiązanie, bowiem licencja wieczysta na arkusz statystyczny Excel 2010 kosztuje nieco ponad 800 PLN. Należy jednak pamiętać, że wielokrotnie zgłaszano zastrzeżenia odnośnie algorytmów liczenia statystyk w Excelu (nie odnaleziono jednak krytyki tego rodzaju odnoszącej się do Excela 2010)[9].

[9] Patrz na przykład: B.D. McCullough, D.A. Heiser, *On the accuracy of statistical procedures in Microsoft Excel 2007*, „Computational Statistics & Data Analysis", 2008, 52 (10), s. 4570-4578.



Aneks 3. Przegląd dostępnych zbiorów danych statystycznych

W niniejszym aneksie Czytelni odnajdzie przegląd dostępnych zbiorów danych statystycznych. Zbiory te - dostępne publicznie i nieodpłatnie - są cennym źródłem danych zastanych. Stanowią one alternatywę względem danych uzyskiwanych w toku badania własnego, przeprowadzanego na potrzeby konkretnego zagadnienia. W archiwach danych zastanych można odnaleźć szereg zbiorów pierwotnych, które po odpowiednim spreparowaniu posłużyć mogą jako materiał do stawiania i wyjaśniania własnych problemów badawczych.

Treść przeglądu obejmuje źródła pochodzenia poszczególnych zbiorów danych, informacje o instytucjach oraz osobach prowadzących te badania, formach finansowania tych przedsięwzięć oraz metodach pozyskiwania danych. Ponadto przedstawiony został opis danych zawartych w bazach wraz z informacjami dotyczącymi metod doboru próby, liczby jednostek analizy, zasięgu badania oraz sposobu realizacji pomiarów. Omówiona została również problematyka badawcza oraz kluczowe cele poszczególnych projektów. Podano również informacje techniczne - przedstawiono warunki korzystania ze zbiorów danych, formaty w jakich są one dostępne oraz konkretne adresy internetowe, pod którymi są zdeponowane (zostały one zamieszczone w końcowej tabeli). Przegląd obejmuje trzynaście zbiorów danych w podziale na zbiory pochodzenia polskiego oraz zagranicznego.

3.1. Przegląd polskich zbiorów danych statystycznych

Przegląd polskich zbiorów danych statystycznych obejmuje cztery archiwa. Trzy z nich, omówione w pierwszej kolejności, to zbiory pochodzące z ogólnopolskich badań opinii publicznej. Należą do nich Polskie Generalne Studium Wyborcze (PGSW), Diagnoza Społeczna oraz Polski Generalny Sondaż Społeczny (PGSS). Czwarty zbiór obejmuje dane spisowe, pochodzące z administracyjnych rejestrów urzędowych. Jest nim Bank Danych Lokalnych (BDL) stanowiący uporządkowane i kompleksowe źródło informacji na temat wszystkich jednostek samorządu terytorialnego w Polsce.

3.1.1. Polskie Generalne Studium Wyborcze (PGSW)

Polskie Generalne Studium Wyborcze jest pionierskim i największym pod względem zasięgu przedsięwzięciem badawczym poświęconym analizie postaw i zachowań wyborczych Polaków. Inspiracją dla jego realizacji były badania empiryczne prowadzone w większości zachodnich państw demokratycznych, inicjowane przy okazji wyborów ogólnokrajowych. Zamysł ten stał się przyczynkiem do przeprowadzenia tożsamych badań w Polsce. Umożliwia to dokonanie empirycznej eksploracji oraz porównywania tendencji w zakresie poglądów, postaw, zachowań politycznych Polaków, ujawniających się w kontekście szczególnego wydarzenia politycznego, jakim są wybory powszechne. Po decyzji Polskiego Komitetu Badań Naukowych w 1997 roku, Polska dołączyła do międzynarodowego kręgu krajów prowadzących badania naukowe nazywane studium wyborczym (National Election Study). Zbiór PGSW stanowi cenne źródło danych dla politologów i socjologów.

Polskie Generalne Studium Wyborcze jest cyklicznym badaniem empirycznym, którego pierwszy pomiar zainicjowany został w 1997 roku, a następnie powtórzony w 2001, 2005 oraz 2007 roku. Dzięki temu zabiegowi zebrane dane umożliwiają analizę trendów oraz uchwycenie dynamiki zmian w postawach i zachowaniach Polaków dokonujących się na przestrzeni czasu, będących wynikiem wydarzeń i przeobrażeń w sferze życia publicznego w Polsce i na świecie. Jest on projektem afiliowanym przez Instytut Studiów Politycznym Polskiej Akademii Nauk. Nad jego realizacją czuwa Radosław Markowski z tegoż ośrodka. Fundusze na realizację tego przedsięwzięcia pochodzą z różnych źródeł finansowania. Należą do nich fundusze Instytutu Studiów Politycznych PAN, granty badawcze uzyskane w drodze konkursów organizowanych przez Ministerstwo Nauki i Szkolnictwa Wyższe, a także środki przekazane przez Wissenschaftszentrum Berlin fur Sozialforschung, Polską Konfederację Pracodawców Prywatnych Lewiatan, Fundację Batorego, Instytut Filozofii i Socjologii PAN oraz ośrodki realizujące badanie - Centrum Badania Opinii Społecznej oraz PBS DGA.

Zbiory danych PGSW pochodzą z badań pierwotnych realizowanych na próbach ogólnopolskich, obejmujących populację Polaków powyżej 18 roku życia. Zastosowaną metodą badawczą jest bezpośredni wywiad kwestionariuszowy. Próby badawcze zostały dobrane metodą doboru warstwowo-losowego, gdzie poszczególne warstwy stanowiły regiony terytorialne wyznaczone według podziału regionalnego Głównego Urzędu Statystycznego. W trakcie realizacji projektu PGSW przeprowadzono cztery pomiary dotyczące wyborów parlamentarnych - w 1997, 2001, 2005 oraz 2007 roku, oraz trzy badania uzupełniające w wymiarze skróconym, poświęcone wyborom prezydenckim w 2000 i 2005 roku oraz wyborom do Parlamentu Europejskiego w 2004 roku. Ogólnodostępnymi zbiorami danych statystycznych są dane pochodzące z pierwszych czterech badań oraz dane z 2000 roku. Liczebności prób badawczych różniły się w zależności od pomiaru. W pierwszym badaniu, które miało miejsce w 1997 roku, zrealizowano N=2003 wywiady, przy czym próba zakładana miała liczyć N=2808. W 2001 roku próba zrealizowana wyniosła N=1794, natomiast zakładana - N=3240. W 2005 roku z przyjętych 2100 wywiadów przeprowadzono 1201, zaś w 2007 roku wynikowa próba badawcza wyniosła N=1817 przy założeniu zrealizowania 3600 wywiadów.

Zbiory danych surowych, pochodzące z powyżej wskazanych pomiarów, są przekazane do ogólnego dostępu przez wirtualny bank danych - Archiwum Danych Społecznych (ADS). Jest to instytucja powołana przez Instytut Studiów Społecznych Uniwersytetu Warszawskiego oraz Instytut Filozofii i Socjologii Polskiej Akademii Nauk. Ideą tego przedsięwzięcia jest gromadzenie i udostępnianie zbiorów

danych ilościowych poświęconych różnym aspektom życia społecznego. Archiwum Danych Społecznych działa w ramach umowy o wspólnym prowadzeniu programu badawczego w zakresie archiwizacji danych społecznych, zawartej między Uniwersytetem Warszawskim a Instytutem Filozofii i Socjologii Polskiej Akademii Nauk, z dnia 8 grudnia 2003 roku. ADS jest pośrednikiem - udostępnia dane, które pochodzą z oficjalnych stron poszczególnych badań i projektów. Dostęp do wszystkich zebranych zbiorów danych możliwy jest pod adresem www.ads.org.pl, po uprzednim wypełnieniu formularza rejestracyjnego oraz zalogowaniu się. Archiwum udostępnia surowe bazy danych z poszczególnych badań (w formacie *.sav lub *.por), a także pochodzące z nich wyniki analiz - statystyki opisowe, rozkłady częstości, raporty końcowe przekazywane w formacie PDF. Na portalu odnajdziemy również zbiory danych statystycznych właściwe dla poszczególnych pomiarów realizowanych w ramach cyklicznego projektu PGSW.

Poszczególne pomiary badania PGSW różnią się w zakresie liczby badanych zagadnień, na co wskazuje odmienna liczba zmiennych w bazach danych. Wszystkie one obejmowały jednak pewne stałe zestawy pytań, które zadawane w każdym pojedynczym pomiarze umożliwiają śledzenie trendów i określenie dynamiki zachodzących zmian w perspektywie longitudinalnej. Do jednych ze stałych modułów należy moduł CSES (Comparative Study of Electoral Systems). Jest to międzynarodowy projekt badawczy prowadzony w ponad pięćdziesięciu krajach świata. Jego celem jest przeprowadzenie transnarodowego studium porównawczego systemów wyborczych. Do badanej problematyki wchodzą takie zagadnienia, jak identyfikacja preferencji wyborczych, ocena komitetów wyborczych, ocena poszczególnych kandydatów, a także - samego systemu wyborczego i politycznego. Dodatkowymi badanymi zagadnieniami jest retrospektywna ewaluacja własnej sytuacji ekonomicznej oraz informacje o cechach socjodemograficznych badanych zbiorowości.

Po pobraniu właściwych plików z bazami danych PGSW, badacz ma sposobność dokonania przeglądu badanych zagadnień szczegółowych oraz oceny ich przydatności dla eksploracji własnych problemów badawczych. Dla przykładu omówimy zbiór danych, który możemy odnaleźć w najaktualniejszym pomiarze PGSW, pochodzącym z 2007 roku.

W badaniu PGSW z 2007 roku poddano ewaluacji przebieg parlamentarnej kampanii wyborczej. Respondenci zostali poproszeni zarówno o ocenę poszczególnych komitetów, jak i zagadnień poruszanych w trakcie kampanii, a także stosunku partii do najważniejszych problemów stojących przed Polską. Badano poparcie dla partii, stopień zainteresowania wyborami, a także sam przebieg kampanii wyborczej. Badanie poświęcone zostało również poglądom politycznym mierzonym na skali lewica-prawica oraz Polska solidarna-Polska liberalna. Istotnym aspektem było zagadnienie frekwencji wyborczej oraz udziału w wyborach parlamentarnych i prezydenckich w 2005 roku. Ponadto, poruszone zostały takie kwestie, jak: stanowisko wobec strategii walki z przestępczością, dekomunizacji, prywatyzacji, systemu podatkowego, polityki wobec UE, a także częstotliwość korzystania z różnych źródeł informacji.

3.1.2. Diagnoza Społeczna

Diagnoza społeczna jest jednym z większych projektów badawczych dotyczących oceny stanu społeczeństwa polskiego. Wyniki tego badania stanowią uzupełnienie tradycyjnych analiz wskaźników makroekonomicznych, makrospołecznych oraz statystyk publicznych. Nadzór merytoryczny nad realizacją projektu sprawuje Rada Monitoringu Społecznego powstała przy Wyższej Szkole Finansów i Zarządzania w Warszawie. Do jej gremium wchodzą przedstawiciele różnych dyscyplin naukowych - ekonomiści, demografowie, psycholodzy, socjolodzy oraz statystycy. Zakres badanej problematyki oraz skład

naukowców różnych specjalności, nadaje całemu projektowi szczególny, interdyscyplinarny charakter. Pomysł projektu zrodził się w 1999 roku i był inicjatywą polskiego socjologa i statystyka - Wiesława Łagodzińskiego. Realizacja projektu rozpoczęła się w 2000 roku. Obecnie jest on realizowany pod kierownictwem Janusza Czapińskiego. W skład zespołu badawczego wchodzą takie postaci polskiego świata nauki, jak Tomasz Panek, Wiesław Łagodziński, Dominik Batorski, Piotr Białowolski, Izabela Grabowska, Irena Kotowska, Paweł Strzelecki, Antoni Sułek, Tadeusz Szumlicz, Dorota Węziak-Białowolska. Każdy z tych badaczy odpowiedzialny jest za realizację innego modułu tematycznego. Diagnoza Społeczna jest finansowana ze środków pozyskanych od różnych instytucji, wymieniając chociażby: Ministerstwo Pracy i Polityki Społecznej oraz Centrum Rozwoju Zasobów Ludzkich czerpiącego fundusze z Europejskiego Funduszu Społecznego w ramach Projektu Operacyjnego Kapitał Ludzki, Narodowy Bank Polski, Ministerstwo Nauki i Szkolnictwa Wyższego, a także Telekomunikację Polską (obecnie Orange Polska), Centertel, Kancelarię Prezesa Rady Ministrów, Bank Zachodni WBK, BRE Bank oraz Główny Inspektor Sanitarny.

Diagnoza Społeczna jest badaniem cyklicznym, prowadzonym na ogólnopolskiej próbie gospodarstw domowych. Projekt jest badaniem panelowym, realizowanym metodą bezpośrednich wywiadów kwestionariuszowych. Uczestnikami badania są wszyscy członkowie gospodarstw domowych, którzy ukończyli 16. rok życia. Gospodarstwa domowe dobierane są metodą losowania warstwowego dwustopniowego. Jednostką losowania pierwszego stopnia są rejony i obwody statystyczne wyznaczone według podziału Głównego Urzędu Statystycznego. W losowaniu drugiego stopnia zastosowano metodę doboru systematycznego na podstawie listy mieszkań uporządkowanej losowo.

W ramach projektu przeprowadzono sześć pomiarów. Pierwszy odbył się w 2000 roku, następny trzy lata później - w 2003 roku, natomiast kolejne pomiary w odstępach dwuletnich - w 2005, 2007, 2009 oraz 2011 roku. Wszystkie pomiary realizowane były w marcu, w celu zagwarantowania porównywalności sezonowej wyników badania. Zrealizowane próby badawcze dla poszczególnych pomiarów miały różne liczebności końcowe. W 2000 roku przebadano 3005 gospodarstw domowych i zrealizowano 9995 wywiadów z jego członkami. W 2003 roku badaniem objęto 3961 gospodarstw domowych, zaś przeprowadzono 9587 wywiadów. Dla kolejnych lat liczebności te prezentują się następująco: 2005 rok - 3851 gospodarstw domowych oraz 8820 wywiadów, 2007 rok - 5532 gospodarstw domowych i 12645 wywiadów, 2009 rok - 12381 gospodarstw domowych oraz 26178 wywiadów, a także 2011 rok - 12386 gospodarstw domowych oraz 26453 wywiady.

Kwestionariusz wywiadu składa się z pięciu modułów. Pierwszy z nich dotyczy warunków życia gospodarstw domowych, gdzie badane zagadnienia dotyczyły dochodów, wyżywienia, stanu materialnego, warunków mieszkaniowych, edukacji, kultury i wypoczynku, opieki zdrowotnej, rynku pracy oraz niepełnosprawności. Drugi moduł odnosi się do indywidualnej oceny jakości życia, w ramach którego poruszane są takie kwestie, jak ogólny stan psychiczny, zadowolenie z poszczególnych aspektów życia, zaufanie do instytucji finansowych, zdrowie, stres w życiu, strategia radzenia sobie z problemami, cechy osobowościowe i styl życia. Trzeci blok tematyczny porusza problematykę stanu polskiego społeczeństwa obywatelskiego. Czwarty moduł jest poświęcony sposobom korzystania z technologii informatyczno-komunikacyjnych, natomiast ostatni blok porusza kwestie dotyczące wykluczenia społecznego - ubóstwa, nierówności społecznych i dochodowych, bezrobocia oraz dyskryminacji społecznej.

Projekt badawczy Diagnoza Społeczna ma w całości charakter publiczny, niekomercyjny oraz ogólnodostępny. Wszyscy zainteresowani mogą otrzymać wyniki oraz zbiory danych pochodzące z poszczególnych pomiarów. Są one dostępne nieodpłatnie na stronie internetowej projektu

www.diagnoza.com, jak i na portalu Archiwum Danych Społecznych. Można z nich pobrać komplet tabel z rozkładami odpowiedzi, a także surowe bazy danych dla sześciu pomiarów łącznie w formacie *.sav.

3.1.3. Polski Generalny Sondaż Społeczny (PGSS)

Polski Generalny Sondaż Społeczny jest ogólnopolskim badaniem sondażowym realizowanym od 1992 roku. Problematyka badawcza projektu obejmuje pomiar postaw, wartości, orientacji i zachowań społecznych Polaków, a także kwestie zróżnicowania społecznego, demograficznego, zawodowego, edukacyjnego i ekonomicznego w poszczególnych grupach i warstwach społecznych w Polsce. Cykliczny charakter projektu pozwala na analizę trendów oraz na badanie dynamiki zmian obserwowanych w społeczeństwie polskim na przestrzeni czasu. Projekt PGSS jest jednym z największych przedsięwzięć badawczych w polskich naukach społecznych i obejmuje szerokie spektrum zagadnień z różnych sfer życia społecznego.

Polski Generalny Sondaż Społeczny jest jednym z głównych programów badań statusowych Instytutu Studiów Społecznych Uniwersytetu Warszawskiego. Rada tegoż programu zrzesza wybitnych przedstawicieli różnych dyscyplin nauk społecznych, wymieniając chociażby Duana F. Alwina, Janusza Czapińskiego, Bogdana Macha, Mirosławę Marody, Andrzeja Rycharda, Kazimierza M. Słomczyńskiego, Andrzeja P. Wejlanda oraz Wojciecha Zaborowskiego. Całemu zespołowi przewodniczy polski socjolog - Antoni Sułek. Badania prowadzone w ramach projektu finansowane są ze środków Ministerstwa Nauki i Szkolnictwa Wyższego przyznawanych przez Komitet Badań Naukowych.

W ramach projektu zrealizowano dziewięć cyklicznych pomiarów począwszy od 1992 roku, przy czym do 1995 roku włącznie były one przeprowadzane corocznie. Następny pomiar odbył się w 1997 roku, zaś od 1999 roku w odstępach trzyletnich - w 2002, 2005 oraz 2008 roku. Dane z tychże pomiarów pochodzą z badań sondażowych z wykorzystaniem bezpośrednich wywiadów kwestionariuszowych. Zrealizowane zostały one na próbach ogólnopolskich. Docelową grupę respondentów stanowili Polacy powyżej 18 roku życia. Metoda doboru próby była zmieniana w trakcie projektu. Do 2002 roku badanie realizowano na losowej próbie adresowej, dobieranej z operatu Głównego Urzędu Statystycznego. Następnie, spośród dorosłych członków każdego z gospodarstw domowych, ankieter wybierał losowo jedną osobę jako respondenta. Dla pomiarów z 2005 i 2008 roku próba badawcza losowana była ze zbiorów Powszechnego Elektronicznego Systemu Ewidencji Ludności PESEL. W ramach wszystkich dziewięciu pomiarów zrealizowano następujące liczby wywiadów: 1992 rok - 1637 wywiadów, 1993 rok - 1649 wywiadów, 1994 rok - 1609 wywiadów, 1995 rok - 1603 wywiady, 1997 rok - 2402 wywiady, 1999 rok - 2282 wywiady, 2002 rok - 2473 wywiady, 2005 rok - 1277 wywiadów oraz 2008 rok - 1293 wywiady.

Projekt badawczy PGSS obejmuje wiele modułów tematycznych. Jednym z nich, realizowanym od początku istnienia programu, jest moduł badań porównawczych International Social Survey Programme, skupiającego obecnie przedstawicieli z ponad 40 krajów. Zaznaczyć należy, iż pomiar tego aspektu dokonywany jest po przeprowadzeniu wywiadu głównego; respondentowi wręczana jej wówczas ankieta ISSP, którą wypełnia on samodzielnie w obecności ankietera. Badane zagadnienia odnoszą się do różnorodnych sfer życia społecznego i obejmują szerokie spektrum problemowe. Przykładowo, w ramach badania systemu stratyfikacyjnego oraz problemu nierówności społecznych, poddano pomiarowi takie kwestie, jak subiektywna ocena pozycji zawodowej, autoidentyfikacja klasowa i warstwowa, ocena własnego statusu zawodowego, edukacyjnego i materialnego, ocena uwarunkowań odpowiedzialnych za osiągnięcie sukcesu zawodowego, a także ewaluacja czynników oraz mechanizmów wpływających na powstawanie

nierówności społecznych, rozwój biedy i możliwość bogacenia się. W aspekcie zagadnień interesujących politologów odnaleźć można informacje o zachowaniach i preferencjach wyborczych, ocenę efektywności systemu politycznego, orientacje polityczne i ideologiczne, zainteresowanie polityką i życiem publicznym, opinie na temat komunizmu i socjalizmu, czy też ocenę własnych poglądów politycznych. Konglomerat badanych zagadnień jest bardzo szeroki, natomiast możliwość porównywania wyników i odnotowywania zmian w ich zakresie sprawia, że dane z projektu PGSS są cennym źródłem wiedzy na temat przeobrażeń życia społecznego Polaków.

Istotnym walorem projektu jest jego niekomercyjny charakter. Zbiory danych oraz dokumentacje metodologiczne pochodzące z poszczególnych fal projektu są ogólnodostępne. Dodatkowym atutem jest możliwość bezpośredniego porównywania danych i wskaźników z wynikami badań przeprowadzonymi w innych krajach. Surowe bazy danych, pochodzące ze wszystkich pomiarów, dostępne są w formacie *.sav w portalu Archiwum Danych Społecznych. Można je pobrać wraz z dokumentacją, statystykami, rozkładami częstości (format PDF) oraz bibliografią (format MS Word). Oficjalne informacje o projekcie PGSS odnajdziemy pod adresem www.pgss.iss.uw.edu.pl.

3.1.4. Bank Danych Lokalnych (BDL)

Bank Danych Lokalnych jest zbiorem danych statystycznych o bieżącej sytuacji gospodarczej, demograficznej, społecznej, politycznej, a także o stanie środowiska naturalnego poszczególnych jednostek samorządu terytorialnego w Polsce. Zakres danych obejmuje podział na województwa, powiaty oraz gminy jako podmiotów systemu organizacji państwowej.

Bank Danych Lokalnych stanowi zbiór informacji pozyskanych drogą spisów powszechnych, administracyjnych systemów informacyjnych, a także z programów badań statystyki publicznej. Jednostką odpowiedzialną za przygotowanie zbiorów danych jest Główny Urząd Statystyczny. Całość przedsięwzięcia finansowana jest z budżetu państwa. Projekt został zainicjowany w 1999 roku. Aktualizacja danych dokonywana jest w trybie długookresowym - corocznie, oraz krótkookresowym - kwartalnie oraz miesięcznie. Zbiory danych rocznych obejmują sześć obszarów tematycznych: informacje na temat gospodarstw domowych, a także kwestie dotyczące nauki i techniki (działalność badawczo - rozwojowa, działalność innowacyjna, ochrona własności przemysłowej w Polsce, społeczeństwo informacyjne), rynku pracy (aktywność ekonomiczna ludności, bezrobocie, dojazd do pracy, pracujący i zatrudnieni w przedsiębiorstwach), samorządu terytorialnego, sektora non-profit (aktywne fundacje, stowarzyszenia i inne organizacje społeczne) oraz ludności (liczba gospodarstw domowych, stan ludności i ruch naturalny). W ramach danych krótkookresowych odnajdziemy tu informacje na temat stanu ludności, a także o aktualnej sytuacji na rynku pracy.

Zbiory danych BDL mają w całości charakter publiczny. Wszyscy zainteresowani mają możliwość zapoznania się z nimi na stronie projektu www.stat.gov.pl/bdl. Stworzona została również możliwość przeprowadzania analiz online. Wyniki prezentowane są w postaci wykresów, które następnie możemy wyeksportować do formatu PDF.

3.2. Przegląd zagranicznych zbiorów danych statystycznych

Zagraniczne zbiory danych mają zasięg zarówno globalny jak i regionalny, w głównej mierze europejski. We wszystkich omawianych poniżej zbiorach można odnaleźć informacje na temat Polski. Pozwala to na stworzenie płaszczyzny porównawczej i odniesienie wyników przypisanych Polsce do wyników uzyskanych przez inne kraje świata. Omawiane zbiory danych pochodzą z takich projektów badawczych, jak Eurobarometr, Europejski Sondaż Społeczny (ESS), International Social Survey Programme (ISSP), World Values Survey (WVS) oraz z archiwów Banku Światowego.

3.2.1. Eurobarometr

Eurobarometr jest cyklicznym badaniem opinii publicznej prowadzonym we wszystkich państwach członkowskich Unii Europejskiej, a także w krajach kandydujących. Jest to projekt międzynarodowy, o długiej tradycji, gdyż został zainicjowany w 1973 roku. Poruszane w nim zagadnienia cechują się wysokim stopniem różnorodności. Głównym celem projektu jest przeprowadzenie ewaluacji sytuacji politycznej oraz gospodarczej obywateli krajów europejskich, ale także stosunku do instytucji UE oraz nastrojów społecznych funkcjonujących wśród mieszkańców państw kandydujących.

Eurobarometr jest projektem realizowanym na zlecenie Komisji Europejskiej. Został on zainicjowany przez Jacques-Rene Rabiera, który uzyskał polityczne poparcie Parlamentu Europejskiego dla swojego przedsięwzięcia. Środki przeznaczone na finansowanie projektu pochodzą głównie z funduszy UE. Prekursorką postacią Eurobarometru były badania opinii publicznej przeprowadzone w latach 1970-1971 w sześciu krajach Wspólnoty Europejskiej - Belgii, Francji, Holandii, Luksemburgu, NRD oraz we Włoszech. Następne badanie odbyło się w latach 1972-1973, do którego dołączyły takie kraje, jak Dania, Irlandia i Wielka Brytania. W ramach projektu możemy wyróżnić dwa główne jego rodzaje. Pierwszy to Candidate Eurobarometer (CC-EB), początkowo funkcjonujący pod nazwą Applicant Countries Eurobarometer (AC-EB). Jest to badanie opinii publicznej, realizowane w krajach kandydujących do Unii Europejskiej. Pierwszy pomiar tego badania został zrealizowany w październiku w 2001 roku i obejmował trzynaście krajów ubiegających się o wstąpienie do wspólnoty - Bułgarię, Czechy, Cypr, Estonię, Węgry, Łotwę, Litwę, Maltę, Polskę, Rumunię, Słowację, Słowenię i Turcję. Drugi rodzaj badania nosi nazwę Central and Eastern Eurobarometer (CEEB). Przeprowadzony został w latach 1991-1997 i obejmował wyłącznie kraje Europy Centralnej oraz Wschodniej, w tym Polskę. Dodatkowo możemy wyróżnić Flash Eurobarometer. Są to badania typu *ad hoc*, odpowiadające na bieżące potrzeby Komisji Europejskiej, prowadzone poza ramowym programem Eurobarometru. Ich cechą charakterystyczną jest względna szybkość uzyskiwanych wyników. Przykładem Flash Eurobarometru jest badanie przeprowadzone w kwietniu w 2004 roku na temat wprowadzenia waluty euro w nowych państwach członkowskich.

Eurobarometr jest regularnym badaniem przeprowadzanym dwa razy w roku, w ramach wiosennej oraz jesiennej fali pomiarowej. Próba badawcza dla poszczególnych krajów wynosi 1000 respondentów. Wyjątkiem jest Wielka Brytania, gdzie badanie realizowane jest na próbie N=1300 oraz Luksemburg - N=500. Jednostkami analizy są obywatele poszczególnych państw Europy. Na stronie oficjalnej projektu www.ec.europa.eu/public_opinion istnieje możliwość zapoznania się z wynikami poszczególnych fal badania począwszy od 1974 roku. Są one udostępnione przede wszystkim w formie raportów końcowych opracowanych w trzech językach - angielskim, francuskim oraz niemieckim. W Archiwum Danych Społecznych odnajdziemy natomiast zbiory danych dla Eurobarometru Europy Środkowschodniej.

Obejmują one zagregowane dane dla wszystkich pomiarów odbywających się corocznie między 1990 a 1997 rokiem. Są one udostępnione w postaci surowych baz danych (format *.sav oraz *.por), tabel frekwencji oraz statystyk opisowych (format ASCII), a także opisów metodologicznych w formacie PDF.

3.2.2. Europejski Sondaż Społeczny (ESS)

Europejski Sondaż Społeczny (ESS - *European Social Survey*) jest międzynarodowym przedsięwzięciem badawczym stanowiącym zbiór empirycznych danych na temat postaw, przekonań oraz zachowań społecznych obywateli krajów europejskich. Jednym z powodów zainicjowania tego projektu było wzrastające zapotrzebowanie na transnarodowe badania porównawcze o tematyce społecznej. Projekt realizuje dwa cele. Pierwszym z nich jest zebranie wiedzy o społecznym i politycznym wymiarze życia Europejczyków. Drugi cel związany jest z opracowaniem standardów metodologicznych dla prowadzenia wiarygodnych badań porównawczych zarówno w wymiarze czasowym jak i w zakresie możliwości porównywania wyników między poszczególnymi krajami europejskimi. Realizacja tego celu została uznana za tak istotną, iż powołano w ramach projektu specjalny Międzynarodowy Komitet Metodologiczny składający się z przedstawicieli wszystkich wiodących europejskich ośrodków naukowych, którzy czuwają nad jakością realizowanego badania.

Nad realizacją Europejskiego Sondażu Społecznego czuwa Europejska Fundacja Nauki, która w dużej mierze jest fundatorem tegoż przedsięwzięcia. Projekt powstał w 2001 roku z inicjatywy Rogera Jowella - brytyjskiego statystyka i pomysłodawcy wielu projektów badawczych, w tym założyciela ośrodka Social nad Community Planning Research, obecnie znanego jako Narodowe Centrum Badań Społecznych (NCSR - National Center for Social Research). R. Jowell był koordynatorem ESS od początku istnienia projektu, aż do swojej śmierci w grudnia 2011 roku. Aktualnie projekt koordynowany jest przez Core Scientific Term pod kierownictwem Rory Fitzgerald z Centre for Comparative Social Surveys City University w Londynie. W Polsce projekt realizowany jest przez Instytut Filozofii i Socjologii Polskiej Akademii Nauk pod kierownictwem Pawła Sztabińskiego, natomiast przedstawicielem w Międzynarodowym Komitecie Metodologicznym jest Henryk Domański. Europejski Sondaż Społeczny, poza Europejską Fundacją Nauki, jest finansowany przez Program Ramowy Komisji Europejskiej, a także przez środki poszczególnych krajów.

Europejski Sondaż Społeczny jest badaniem cyklicznym, realizowanym w odstępach dwuletnich. Każdy pojedynczy pomiar nazywany jest rundą. Dotychczas zrealizowano pięć rund tego badania - w 2002, 2004, 2006, 2008 oraz 2010 roku. Obecnie prowadzone są przygotowania do kolejnego pomiaru. Polska uczestniczy w projekcie począwszy od pierwszej rundy. Badanie realizowane jest na ogólnokrajowych próbach imiennych. Respondentami są obywatele krajów europejskich, którzy ukończyli 15 rok życia. Na chwilę obecną, projektem objętych jest ponad trzydzieści krajów Europy. W trakcie kolejnych rund badania zrealizowano różne liczebności prób badawczych. Dla Polski te wielkości w kolejnych latach prezentują się następująco: 2002 rok - 2110 wywiadów, 2004 rok - 1716 wywiadów, 2006 rok - 1721 wywiadów, 2008 rok - 1619 wywiadów oraz 2010 rok - 1751 wywiadów.

Kwestionariusz badania ESS składa się z dwunastu modułów tematycznych. Są one powtarzane w kolejnych rundach projektu. Dotyczą one takich kwestii, jak zaufanie do władzy, polityków i instytucji, zainteresowanie polityką i uczestnictwo w życiu politycznym, orientacje społeczno-polityczne, ocena władzy i jej skuteczności na poziomie krajowym i międzynarodowym, wartości moralne, polityczne i społecznej, integracja i wykluczenie społeczne, tożsamość narodowa, etniczna i religijna, poczucie

materialnego dobrobytu, stan zdrowia i poczucie bezpieczeństwa, skład demograficzny gospodarstw domowych, wykształcenie i zawód respondenta, współmałżonka i rodziców, warunki materialne, a także warunki życiowe gospodarstwa domowego. Zakres poruszanych problemów jest bardzo szeroki i różnorodny. Sprawia to, iż wyniki badań ESS są cennym źródłem wiedzy dla przedstawicieli różnych dyscyplin nauk społecznych, w tym politologów.

Projekt ESS ma charakter niekomercyjny. Z tego względu wyniki badań są dostępne nieodpłatnie, po uprzednim zarejestrowaniu się na oficjalną stronę projektu – www.europeansocialsurvey.org. W tym miejscu odnajdziemy wszelkie dane oraz informacje dotyczące poszczególnych rund badania. W ich zakres wchodzi dokumentacja metodologiczna projektu (dane na temat badanych populacji, standardy kodowania, lista zmiennych oraz pytań, dobór próby, konstrukcja wag), dokumentacja z terenowej fazy realizacji badania, informacje na temat błędów oraz problemów zaistniałych w trakcie realizacji, metody konstruowania wskaźników oraz indeksów. Odnajdziemy również surowe zbiory danych dostępne w formacie *.sav oraz dla programu SAS. Ponadto istnieje możliwość analizy danych online – tworzenia tabel krzyżowych z wybranych zmiennych, dla każdego dostępnego zbioru z poszczególnych rund.

3.2.3. International Social Survey Programme (ISSP)

International Social Survey Programme jest międzynarodowym przedsięwzięciem badawczym zrzeszającym ponad 48 krajów świata w celu realizacji badań porównawczych o *stricte* socjologicznym charakterze. Cechą charakterystyczną tego projektu jest dążenie do opracowania jednolitych standardów metodologicznych oraz zestawów pytań o jednakowej treści, bez względu na istniejące różnice językowe oraz kulturowe. Należy zaznaczyć, iż tematyka poszczególnych pomiarów jest różna. Każdy pomiar jest jednak poświęcony jednej, konkretnej problematyce. W przypadku badania ISSP nie dokonuje się mieszania zagadnień tak, jak ma to miejsce w przypadku pozostałych badań o charakterze międzynarodowym.

Projekt ISSP powstał w 1984 roku z inicjatywy czterech instytucji naukowych – National Opinion Research Center (NORC) przy Uniwersytecie Chicago, Research School of Social Sciences (RSSS) istniejącej w ramach Australian National University, Social and Community Planning Research (SCPR), organizacji mającej swoją siedzibę w Londynie oraz Zentrum für Umfragen, Methoden, und Analysen (ZUMA) mieszczącej się w Mannheim. Wszelkie przedsięwzięcia badawcze są finansowane z funduszy własnych ISSP. Pierwszy pomiar miał miejsce w 1985 roku. Polska przystąpiła do niego w 1991 roku. Osobami odpowiedzialnymi za jego realizację w kraju jest Bogdan Cichomski oraz Marcin Zieliński, zaś placówką naukową czuwającą nad całością przedsięwzięcia jest Instytut Studiów Społecznych Uniwersytetu Warszawskiego.

ISPP jest międzynarodowym badaniem opinii publicznej realizowanym wśród respondentów indywidualnych. Jest ono badaniem cyklicznym, powtarzanym w odstępach rocznych. Pierwszy pomiar odbył się w 1985 roku i został poświecony zagadnieniu roli władz państwowych w życiu obywateli. Badanie tej kwestii zostało powtórzone w kolejnych latach – w 1990, 1996 oraz 2006 roku. Innymi badanymi zagadnieniami są więzi, relacje i stosunki społeczne (1986, 2001), nierówności społeczne (1987, 1992, 1999, 2009), rodzina i zmiana ról płciowych (1988, 1994, 2002, 2012 – w trakcie realizacji), praca, wykonywany zawód oraz bezrobocie (1989, 1997, 2005), religia i jej wpływ na zachowania społeczne, moralne i polityczne (1991, 1998, 2008), środowisko naturalne (1993, 2000, 2010), tożsamość narodowa oraz związane z tym zagadnienia takie, jak nacjonalizm, patriotyzm, lokalizm, globalizm, imigracja (1995, 2003), społeczeństwo obywatelskie (2004), czas wolny, sport i rekreacja (2007), a także zdrowie (2011).

Obecnie w przygotowaniach są kolejne fale projektu; w 2013 roku planowane jest powtórzenie badania poświęconego tożsamości narodowej, zaś w 2014 roku - społeczeństwu obywatelskiemu.

Wyniki pochodzące z poszczególnych pomiarów mają charakter ogólnodostępny. Zbiory danych surowych, w których uczestniczyła Polska, są dostępne na stronie Archiwum Danych Społecznych. Badacz ma możliwość pobrania piętnastu zbiorów danych pochodzących z lat 1991-2008. Są one udostępnione w formacie *.sav oraz *por. Dla publicznego użytkownika zostały również przekazane rozkłady częstości, statystyki opisowe oraz dokumentacja projektu, w formatach *.pdf lub ASCII. Istnieje również możliwość zapoznania się z danymi pochodzącymi z lat wcześniejszych za pośrednictwem oficjalnej strony projektu www.issp.org. Portal stwarza również możliwość przeprowadzenia analiz online za pomocą tabel częstości oraz tabel krzyżowych, po uprzednim wybraniu interesującego zakresu zmiennych.

3.2.4. Wybrane zbiory danych statystycznych Banku Światowego

Jedną z największych instytucji o zasięgu transnarodowym, zajmującą się badaniem niemal wszystkich krajów świata, jest Bank Światowy z siedzibą w Waszyngtonie. Rozpoczął on swoją działalność na mocy postanowień konferencji w Bretton Woods z lipca 1944 roku. Do jego głównych celów statutowych należy pomoc w odbudowie zniszczonych po II wojnie światowej krajów europejskich oraz Japonii, a także udzielanie wsparcia dla rozwijających się krajów Afryki, Ameryki Łacińskiej oraz Azji. Obecnie instytucja ta zrzesza 187 krajów świata. Jej podstawową działalnością jest udzielanie wsparcia finansowego w postaci długoterminowych pożyczek o preferencyjnym oprocentowaniu dla najbardziej potrzebujących krajów członkowskich. Bank Światowy prowadzi jednak szeroko zakresową działalność badawczą. Wszelkie przedsięwzięcia o tym charakterze stanowią cenne źródło danych empirycznych na temat globalnego obrazu społeczeństwa, dostarczając przy tym bogatych i doskonale opracowanych wskaźników makroekonomicznych oraz makrospołecznych. Dlatego też, wszelkie zbiory danych powstałe z inicjatywy Banku Światowego, mogą stanowić wartościowe źródło wiedzy dla politologów, socjologów oraz innych przedstawicieli nauk społecznych.

Kolekcje zbiorów danych udostępnione przez Bank Światowy charakteryzują się wysokim stopniem różnorodności. Odnajdziemy w nich ponad czterdzieści projektów badawczych. Każdy z nich koncentruje się na odrębnym zagadnieniu, w głównej mierze o charakterze globalnym, ale również regionalnym lub dedykowanym konkretnemu krajowi. Większość zbiorów zawiera zestawienia obejmujące konkretne informacje dla poszczególnych krajów członkowskich. Dane prezentowane są głównie w postaci opracowanych wskaźników makrostrukturalnych i odnoszą się one do szerokiego zakresu problemów, wymieniając chociażby kwestie dotyczące gospodarki, edukacji, środowiska naturalnego, sektora finansowego, zdrowia, infrastruktury, pracy i polityki socjalnej, ubóstwa, działalności sektora prywatnego oraz publicznego. Zebrane dane są podzielone ze względu na kraj pochodzenia oraz lata, w których dana wartość wskaźnika została odnotowana. Większość zbiorów danych zawiera informacje począwszy od 1960 roku. Dokonując wyboru zbiorów danych Banku Światowego kierowaliśmy się przede wszystkim ich użytecznością poznawczą dla badaczy nauk społecznych, w szczególności politologów. Pominięte zostały zbiory obejmujące wskaźniki makroekonomiczne, głównie finansowe oraz rankingi giełdowe.

3.2.4.1. Atlas of Social Protecions: Indicators of Resilience and Equity (ASPIR)

Badanie ASPIR jest projektem Banku Światowego poświęconym zagadnieniu polityki socjalnej. Obejmuje on dane z 56 krajów świata, głównie z krajów rozwijających się, w tym również Polski. Ten projekt realizowany jest w ramach programu Social Protection and Labor (SPL) prowadzonego przez Bank Światowy. Badane zagadnienia odnoszą się do kwestii statusu społecznego i ekonomicznego ludności świata, walki z ubóstwem i nierównościami społecznymi, a także do oceny sytuacji gospodarstw domowych. Ewaluacji poddawana jest również działalność programowa SPL, mająca na celu przezwyciężenie problemu biedy, wykluczenia i nierówności w społeczeństwach na całym świecie.

ASPIR jest badaniem cyklicznym prowadzonym dwa razy w roku, począwszy od 2005 roku. Dane zbierane są w toku badań sondażowych realizowanych na próbach gospodarstw domowych. Wyniki badania oraz wskaźniki zostały opracowane w oparciu o informacje zebrane od 1,3 mln gospodarstw domowych, w tym niemal 5 mln respondentów indywidualnych, pochodzące z blisko 69 zestandaryzowanych badań gospodarstw domowych z całego świata. Zbiory danych są ogólnodostępne pod adresem internetowym www.data.worldbank.org/data-catalog/atlas_social_protection. Można je pobrać w formacie *.csv bądź *.xls. Ponadto istnieje możliwość tworzenia analiz online.

3.2.4.2. Education Statistics (EdStats)

Education Statistics jest jednym ze stałych projektów badawczych realizowanych przez Bank Światowy. Poświęcony jest problemowi edukacji w wymiarze globalnym. Celem badania jest zebranie porównywalnych wskaźników w zakresie rozwoju edukacji, wzrostu poziomu wykształcenia mieszkańców poszczególnych krajów świata, a także wydatków ponoszonych na ten cel. Dodatkowo badaniu podlegają posiadane oraz nabywane umiejętności przez beneficjentów systemów oświatowych, począwszy od okresu przedszkolnego, aż po szkolnictwo wyższe. Cennym uzupełnieniem zbiorów danych są wyniki międzynarodowych ocen kształcenia.

Dane znajdujące się w zbiorach EdStats pochodzą z badań prowadzonych wśród gospodarstw domowych. Jest to projekt cykliczny, realizowany raz w roku, począwszy od 1970 roku w ponad 180 krajach członkowskich. Jego zakończenie przewidziane jest na 2050 rok. Zagadnienia szczegółowe, które odnajdziemy w zbiorach EdStat, to chociażby wskaźniki rekrutacji na studia wyższe, rodzaje ukończonych kierunków studiów, liczba nauczycieli, poziom analfabetyzmu wśród dzieci oraz dorosłych, długość nauki w latach dla poszczególnych etapów edukacji, a także szacunki dla liczby osób posiadających określony poziom wykształcenia.

Wyniki badania EdStats są ogólnodostępne na stronie oficjalnej projektu - www.data.worldbank.org/data-catalog/ed-stats. Istnieje możliwość nieodpłatnego pobrania surowych baz danych w formatach .csv oraz .xls, które mogą z łatwością zostać zaimportowane do odpowiednich programów statystycznych. Dane można analizować również online. W tym celu należy wybrać interesujący nas kraj, wskaźnik oraz rok pochodzenia danych. Uprzednio zdefiniowany zakres analiz jest prezentowany w postaci tabel częstości bądź tabel krzyżowych, a także w formie graficznej - wykresów lub map natężenia zjawisk.

3.2.4.3. Gender Statistics (GenderStats)

Gender Statistics jest przedsięwzięciem Banku Światowe poświęconym kluczowym zagadnieniom związanym z płcią. Zbiory danych zawierają szereg wskaźników, które podzielono na pięć modułów tematycznych, związanych kolejno z demografią, edukacją, zdrowiem, pracą, a także formami partycypacji politycznej kobiet i mężczyzn. Stanowią one agregację informacji dla krajów z poszczególnych regionów świata. Dane są aktualizowane raz w roku, począwszy od 1960 roku. Zestawienia statystyczne odnoszą się do wszystkich krajów członkowskich. Zbiory danych są ogólnodostępne pod adresem www.data.worldbank.org/data-catalog/gender-statistics. Podobnie, jak w pozostałych badaniach Banku Światowego, istnieje możliwość pobrania danych w formacie .csv oraz .xls, a także przeprowadzenia analiz online.

3.2.4.4. Global Bilateral Migration Database (GBMD)

Przedsięwzięcie badawcze Banku Światowego o nazwie Global Bilateral Migration Database (GBMD) jest pierwszym kompleksowym projektem poświęconym zagadnieniu migracji ludności w wymiarze globalnym. Na podstawie zagregowanych danych pochodzących z ponad tysiąca spisów ludności oraz rejestrów publicznych krajów członkowskich, opracowano obraz dwukierunkowej migracji w skali światowej dla ostatnich 50 lat. Jednostkami analizy są emigranci i imigranci z całego świata. Zbiory danych obejmują informację od 1960 roku i są prezentowane w cyklach dziesięcioletnich - dla 1960, 1970, 1980, 1990 oraz 2000 roku. Wyniki tego badania pokazują, iż na przestrzeni lat aktywność migrantów w skali globalnej wzrosła z poziomu 92 mln do 165 mln. Zbiory danych są dostępne wyłącznie online, w podziale na lata pochodzenia wyników badania, kraj oraz płeć. Można z nimi zapoznać się pod adresem internetowym www.data.worldbank.org/data-catalog/global-bilateral-migration-database.

3.2.4.5. World Development Indicators (WDI)

World Development Indicators jest jednym z największych zbiorów danych udostępnionych przez Bank Światowy. Zawiera on szereg danych oraz wskaźników rozwoju poszczególnych krajów świata, które opracowane zostały na postawie oficjalnie uznawanych źródeł międzynarodowych. Przedstawia on najbardziej aktualne i dokładne dane na temat wielowymiarowego rozwoju społeczeństw światowych, zarówno w wymiarze globalnych, regionalnych, jak i krajowych oszacowań. Poszczególne wskaźniki są opracowywane dla takich kategorii, jak edukacja, środowisko naturalne, finanse, zdrowie, infrastruktura czy też działalność podmiotów sektora prywatnego oraz publicznego. Ponadto, bazy danych WDP od 2011 roku zawierają dane na temat konfliktów światowych, wojen domowych, terroryzmu oraz kwestii bezpieczeństwa w ujęciu globalnym.

Informacje o wskaźnikach zbierane są regularnie, począwszy od 1960 roku. Są one na bieżąco aktualizowane w odstępach kwartalnych. Pierwsza aktualizacja przypada na kwiecień, kolejne na lipiec oraz wrzesień, zaś ostatnia na grudzień. Dane z tego projektu mają w całości charakter publiczny. Wszyscy zainteresowani mogą z nimi zapoznać się nieodpłatnie na stronie internetowej www.data.worldbank.org/data-catalog/world-development-indicators. Można je również pobrać w formacie .csv oraz .xls. Istnieje także możliwość przeprowadzenia analiz online oraz eksportu uzyskanych wyników do takich formatów, jak *.xls, *.csv, *.tabbed oraz *.sdmx.

3.2.5. World Values Survey (WVS)

World Values Survey jest jednym z największych światowych projektów badawczych i głównym źródłem danych empirycznych na temat wartości, postaw, przekonań i ich wpływu na życie społeczne i polityczne społeczeństwa globalnego. Projekt realizowany jest w niemal stu krajach, przy udziale sieci naukowców z całego świata.

Badania w ramach projektu realizowane są przez organizację non-profit - World Values Survey Association mającą swoją główną siedzibę w Sztokholmie. Całość przedsięwzięcia została zainicjowana przez obecnego kierownika projektu - Ronalda Ingleharta, profesora nauk politycznych na Uniwersytecie Michigan. W Polsce prace nad badaniem koordynowane są przez dwie osoby - Renatę Siemieńską oraz Aleksandrę Jasińską-Kania z Instytutu Studiów Społecznych Uniwersytetu Warszawskiego. Badania realizowane w poszczególnych krajach świata finansowane są ze środków pozyskanych przez krajowe zespoły badawcze. Z reguły pochodzą one z lokalnych funduszy. W przypadku zaistnienia problemów z finansowaniem, istnieje możliwość pozyskania funduszy ze środków centralnych World Values Survey. Działalność samej organizacji i pełnionych przez nią funkcji wykonawczych, finansowana jest przez Bank Sweden Tercentennary Foundation, a także z innych źródeł takich, jak National Science Foundation, Swedish International Development Cooperation Agency (SIDA), Volkswagen Foundation, holenderskie Ministerstwo Spraw Zagranicznych.

Badanie World Values Survey zostało zainicjowane w 1981 roku. Obecnie projekt jest globalnym przedsięwzięciem w dziedzinie nauk społecznych, w ramach którego przeprowadzane są cykliczne, reprezentatywne oraz porównawcze badania empiryczne. Obejmuje on ponad 90 proc. populacji świata. Do 2012 roku przeprowadzono łącznie sześć fal badania. Pierwsza została zrealizowana w 1981 roku, druga w latach 1990-1991, trzecia w latach 1995-1997, czwarta w latach 1999-2001, piąta w latach 2005-2007 oraz szósta, ostatnia - w latach 2011-2012. Ostatnimi trzema cyklami badania objęto kraje Europy Środkowschodniej, w tym również Polskę.

Badania Wolrd Values Survey przeprowadzane są na próbach ogólnokrajowych, gdzie respondentami są obywatele poszczególnych krajów, posiadający ukończony 18 rok życia. W realizacji poszczególnych pomiarów wykorzystuje się metodę mieszaną. Podstawową metodę realizacyjną są bezpośrednie wywiady kwestionariuszowe. W przypadku badań prowadzonych na odległych terenach oraz w lokalizacjach trudnodostępnych wykorzystuje się wywiad telefoniczny wspomagany komputerowo (*Computer Assisted Telephone Interviewing, CATI*). Minimalna liczebność próby dla każdego kraju wynosi 1000 wywiadów.

Podstawowa problematyka badawcza projektu World Values Survey koncentruje się na zagadnieniu wartości oraz przekonań społecznych ludności z całego świata. Cyklicznie przeprowadzane pomiary stwarzają możliwość analizowani trendów oraz umożliwiają zaobserwowanie zmian, jakie dokonują się na przestrzeni czasu. W tym względzie badane są takie aspekty życia, jak rodzina, przyjaciele, czas wolny, polityka oraz religia. Politologów w szczególności zainteresują zagadnienia dotyczące poparcia dla demokracji, tolerancji wobec cudzoziemców i mniejszości etnicznych, stosunku do równości płci, czy wpływu globalizacji na poglądy i wartości. Wyniki tego badania charakteryzują się wysoką użytecznością dla władz państwowych, które mogą odnaleźć w nich cenne informacje dla rozwijania koncepcji społeczeństwa obywatelskiego w ojczystych ośrodkach. Ponadto, projekt ma na celu przeprowadzenie bieżącej ewaluacji kluczowych wydarzeń na świecie i poznanie opinii obywateli poszczególnych krajów na ich temat. W wynikach badania odnajdziemy bowiem ocenę takich wydarzeń, jak sytuacja na Bliskim Wschodzie oraz w Ameryce Północnej w latach 2010-2011, odniesienie się do protestów we Francji w 2005

roku, czy też do wojny w Jugosławii w 1990 roku, bądź dramatycznych zbrodni ludobójstwa mających miejsce w Rwandzie w 1994 roku.

Zbiory danych z poszczególnych fal badania są dostępne bez ograniczeń do celów niekomercyjnych na oficjalnej stronie projekt www.wvsevsdb.com. Surowe bazy danych można pobrać w odpowiadającym właściwemu programowi statystycznemu formacie - dla SPSS, SAS oraz STATA. Ponadto, realizatorzy projektu stworzyli możliwość analizy danych online. W tym celu wybieramy interesujący nas pomiar, kraj oraz dział, z którego wynikami chcemy się zapoznać. W tym aspekcie do wyboru są takie zakresy zagadnień, jak ocena jakości życia, środowisko, praca, rodzina, polityka, społeczeństwo, religia i moralność, tożsamość narodowa oraz dane socjodemograficzne. Platforma online umożliwia stworzenie tabel częstości, tabel krzyżowych, a także wykresów słupkowych, kołowych oraz liniowych.

Tabela 48. Adresy stron internetowych z omawianymi zbiorami danych[1]

Lp.	Nazwa zbioru danych statystycznych	Adres internetowy udostępnionego zbioru danych
Polskie zbiory danych statystycznych		
1	Polskie Generalne Studium Wyborcze (PGSW)	**PGSW 2007:** www.ads.org.pl/pobieranie-zbioru-danych.php?id=72
		PGSW 2005: www.ads.org.pl/pobieranie-zbioru-danych.php?id=33
		PGSW 2001: www.ads.org.pl/pobieranie-zbioru-danych.php?id=23
		PGSW 2000: www.ads.org.pl/pobieranie-zbioru-danych.php?id=22
		PGSW 1997: www.ads.org.pl/pobieranie-zbioru-danych.php?id=21
2	Diagnoza Społeczna	www.ads.org.pl/pobieranie-zbioru-danych.php?id=58
3	Polski Generalny Sondaż Społeczny (PGSS)	www.ads.org.pl/pobieranie-zbioru-danych.php?id=53
4	Bank Danych Lokalnych (BDL)	www.stat.gov.pl/bdl/app/strona.html?p_name=indeks
Zagraniczne zbiory danych statystycznych		
5	Eurobarometr	www.ec.europa.eu/public_opinion/archives/eb_arch_en.htm
		zbiór danych Eurobarometru Europy Środkowo-wschodniej: www.ads.org.pl/pobieranie-zbioru-danych.php?id=14
6	Europejski Sondaż Społeczny (ESS)	www.ads.org.pl/pobieranie-zbioru-danych.php?id=15
7	International Social Survey Programme (ISSP)	www.issp.org/page.php?pageId=4
		zbiory danych z lat 1991-2006 dostępne są w: www.ads.org.pl/lista-zbiorow-danych.php
8	Atlas of Social Protecions: Indicators of Resilience and Equity (ASPIR)	www.data.worldbank.org/data-catalog/atlas_social_protection
9	Education Statistics (EdStats)	www.data.worldbank.org/data-catalog/ed-stats
10	Gender Statistics (GenderStats)	www.data.worldbank.org/data-catalog/gender-statistics

[1] Dostęp do baz danych jest aktualny na lipiec 2012.

Lp.	Nazwa zbioru danych statystycznych	Adres internetowy udostępnionego zbioru danych
11	Global Bilateral Migration Database (GBMD)	www.data.worldbank.org/data-catalog/global-bilateral-migration-database
12	World Development Indicators (WDI)	www.data.worldbank.org/data-catalog/world-development-indicators
13	World Values Survey (WVS)	www.wvsevsdb.com/wvs/WVSData.jsp

IV

Aneks 4. Tablica wartości krytycznych rozkładu chi-kwadrat

Liczba stopni swobody	Poziom ufności α		
	0,1	0,05	0,01
1	2,706	3,841	6,635
2	4,605	5,991	9,21
3	6,251	7,815	11,345
4	7,779	9,488	13,277
5	9,236	11,07	15,086
6	10,645	12,592	16,812
7	12,017	14,067	18,475
8	13,362	15,507	20,09
9	14,684	16,919	21,666
10	15,987	18,307	23,209
11	17,275	19,675	24,725
12	18,549	21,026	26,217
13	19,812	22,362	27,688
14	21,064	23,685	29,141
15	22,307	24,996	30,578
16	23,542	26,296	32
17	24,769	27,587	33,409
18	25,989	28,869	34,805
19	27,204	30,144	36,191
20	28,412	31,41	37,566
21	29,615	32,671	38,932
22	30,813	33,924	40,289
23	32,007	35,172	41,638
24	33,196	36,415	42,98
25	34,382	37,652	44,314
26	35,563	38,885	45,642
27	36,741	40,113	46,963
28	37,916	41,337	48,278
29	39,087	42,557	49,588
30	40,256	43,773	50,892

Bibliografia

Babbie E., *Badania społeczne w praktyce*, tłum. W. Betkiewicz, M. Bucholc, P. Gadomski, J. Haman, A. Jasiewicz-Betkiewicz, A. Kloskowska-Dudzińska, M. Kowalski, M. Mozga, Wydawnictwo Naukowe PWN, Warszawa 2004.

Babiński G., *Wybrane zagadnienia z metodologii socjologicznych badań empirycznych*, Uniwersytet Jagielloński, Kraków 1980.

Bedyńska S., *Clooney? Brzydal! Di Caprio? Brzydal! Analiza czynnikowa - metody wyodrębniania czynników*, w: http://www.predictivesolutions.pl/EKSPRESS/Analiza_danych_w_dzialaniu/Techniki_analityczne/Analiza_czynnikowa/Clooney_Brzydal_Di_Caprio_Brzydal_Analiza_czynnikowa_metody_wyodrebniania_czynnikow_cz2_SBedynska.pdf, dostęp: lipiec 2012.

Berger J., *Spisy ludności na ziemiach polskich do 1918 roku*, „Wiadomości Statystyczne", 2002, 1, ss.12-19.

Biecek P., Trajkowski K., *Na przełaj przez Data Mining*, w: http://www.biecek.pl/NaPrzelajPrzezDataMining, 2011, dostęp: kwiecień 2012.

Blalock H.M., *Statystyka dla socjologów*, Państwowe Wydawnictwo Naukowe, Warszawa 1977.

Brown W., *Some experimental results in the correlation of mental abilities*, „British Journal of Psychology", 1910, 3, ss. 296-322.

Buttolph J., Reynolds H.T., Mycoff J.D., *Metody badawcze w naukach politycznych*, Wydawnictwo Naukowe PWN, Warszawa 2010.

Carifio J., Perla R.J., *Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes*, „Journal of Social Sciences", 2007, 3 (3), w: http://www.comp.dit.ie/dgordon/Courses/ResearchMethods/likertscales.pdf, dostęp: wrzesień 2012.

Cattell R.B., *The scree test for the number of factors*, „Multivariate Behavioral Research", 1966, 1, ss. 245-276.

Cicchetti D.V., Volkmar F., Sparrow S.S., Cohen D., *Assessing the Reliability of Clinical Scales When the Data Have Both Nominal and Ordinal Features: Proposed guidelines for neuropsychological assessments*, „Journal of Clinical and Experimental Neuropsychology", 1992, 14 (5), ss. 673-686.

Clauss G., Ebner H., *Podstawy statystyki dla psychologów, pedagogów i socjologów*, Państwowe Zakłady Wydawnictw Szkolnych, Warszawa 1972.

Cochran W.G., *The chi-square goodness-of-fit test*, „Annals of Mathematical Statistics", 1952, 23, ss. 315-345.

Cohen J., *A Coefficient of Agreement For Nominal Scales*, „Educational and Psychological Measurement", 1960, 10 (3746), ss. 37-46.

Cohen J., *Statistical power analysis for the behavioral sciences*, Routledge, Nowy Jork 1988.

Coven V., *A History of Statistics in the Social Sciences*, „An Academic Journal on the Web", 2003, ss. 1-5.

Cramér H., *Mathematical Methods of Statistics,* Princeton University Press, Princeton 1946.

Cronbach L.J., *Coefficient alpha and the internal structure of tests*, „Psychometrika", 1951, 16, ss. 297-334.

Dawkins R., *Who is the greatest biologist of all time. A talk with Armand Leroi*, „Edge", 15 maja 2011, w: http://www.edge.org/3rd_culture/leroi11/leroi11_index.html#dawkins, dostęp: styczeń 2012.

Demski T., *Od pojedynczych drzew do losowego lasu*, w: *Zastosowania statystyki i data mining w badaniach naukowych*, StatSoft Polska, Kraków 2011.

Actuarial modelling of claim counts: risk classification, credibility and bonus-malus system, M. Denuit (red.), John Wiley and Sons, Chichester 2007.

Diagnoza Społeczna 2009, J. Czapiński, T. Panek (red.), Rada Monitoringu Społecznego Wyższa Szkoła Finansów i Zarządzania w Warszawie, Warszawa 2009.

Domański Cz., *Zasłużeni statystycy dla nauki*, w: http://www.stat.gov.pl/cps/rde/xbcr/gus/POZ_Zasluzeni_statystycy_dla_nauki.pdf, dostęp: czerwiec 2012.

Downey R.G., King C.V., *Missing data in Likert ratings: A comparison of replacement methods*, „Journal of General Psychology", 1998, 125, ss. 175-189.

Ekström J., *On Pearson-Verification And The Chi-Square Test*, w: http://preprints.stat.ucla.edu/625/Ekström%20-%20On%20Pearson-verification%20and%20the%20chi-square%20test.pdf, dostęp: maj 2012.

Fienberg S.E., *A Brief History of Statistics in Three and One-Half Chapters: A Review Essay*, „Statistical Science", 1992, 7 (2), ss. 208-222.

Fisher Box J., *R.A. Fisher: the life of a scientist*, John Wiley and Sons, Nowy Jork 1978.

Fisher R.A., *The logic of inductive inference*, „Journal of the Royal Statistical Society", 1935, 98, ss. 39-54.

Ford B.L., *An overview of hot-deck procedures*, w: *Incomplete data in sample surveys*, W.G. Madow, I. Olkin, D.B. Rubin (red.) , Academic Press, Nowy Jork 1983, ss. 185-207.

Fox J., *Extending the R Commander by 'plug in' packages*, „R News", 2007, 7 (3), ss. 46-52.

Fox J., *The R Commander: A Basic Statistics Graphical User Interface to R*, „Journal of Statistical Software", 2005, 14 (9), ss. 1-42.

Francuz P., Mackiewicz R., *Liczby nie wiedzą skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów*, Wydawnictwo KUL, Lublin 2005.

Gallup G., Rae S., *The Pulse of Democracy, Simon & Schuster*, Nowy Jork 1940.

Garland R., *The Mid-Point on a Rating Scale: Is it Desirable?*, „Marketing Bulletin", 1991, 2, ss. 66-70.

Garson G.D., *Correlation*, Statistical Associates Publishing, Asheboro 2012.

Gatnar E., Walesiak M., *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

Gawrysiak P., *Cyfrowa rewolucja. Rozwój cywilizacji informacyjnej*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Gayon J., *Darwinism's Struggle for Survival. Heredity and the Hypothesis of Natural Selection*, Cambridge University Press, Cambridge 1998.

Good P.I., Hardin J.W., *Common Errors in Statistics (and How to Avoid Them)*, John Wiley and Sons, New Jersey 2003.

Górniak J., Wachnicki J., *Pierwsze kroki w analizie danych. SPSS PL for Windows*, SPSS Polska, Kraków 2000.

Groves R.M., Peytcheva E., *The Impact of Nonresponse Rates on Nonresponse Bias a Met Analysis*, „Public Opinion Quarterly", 2009, 72 (2), ss. 167-189.

Gupta M.R., Chen Y., *Theory and Use of the EM Algorithm*, „Foundations and Trends in Signal Processing", 2010, 4 (3), ss. 223-296.

Guttman L., *A basic for analyzing test-retest reliability*, „Psychometrika", 1945, 10 (4), ss. 255-282.

History of Statistics, „Stochastikon", w: http://132.187.98.10:8080/encyclopedia/en/statisticsHistory.pdf, dostęp: czerwiec 2012.

Holcomb W.L., Chaiworapongsa T., Luke D.A., Burgdorf K.D., *An odd measure of risk: use and misuse of the odds ratio*, „Obstetrics and Gynecology", 2001, 98 (4), ss. 685-688.

Jarosz-Nowak J., *Modele oceny stopnia zgody pomiędzy dwoma ekspertami z wykorzystaniem współczynników kappa*, „Matematyka stosowana", 2007, w: www.matstos.pjwstk.edu.pl/no8/no8_jarosz-nowak.pdf, dostęp: czerwiec 2012.

Kaiser H.F., *The application of Electronic Computers of Factor Analysis,* „Educational and Psychological Measurement", 1960, 20, ss. 141-151.

Kalton G., Kasprzyk D., *The treatment of missing data*, „Survey Methodology", 1986, 12, ss. 1-16.

Kendall M., *A New Measure of Rank Correlation*, „Biometrika", 1938, 30 (1-2), ss. 81-89.

Klaus G., Ebner H., *Podstawy statystyki dla psychologów i socjologów*, Państwowe Zakłady Wydawnictw Szkolnych, Warszawa 1972.

Kordos J., *Professor Jerzy Neyman - Some Reflections*, „Lithuanian Journal of Statistics", 2011, 50 (1), ss. 114-122.

Koronacki J., Mielniczuk J., *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwo Naukowo-Techniczne, Warszawa 2001.

Kula W., *Miary i ludzie*, Wydawnictwo „Książka i Wiedza", Warszawa 2002.

Kustre D., *Denmark waves Good Bye to Microsoft formats*, w: www.brightsideofnews.com/news/2010/2/3/denmark-waves-good-bye-to-microsoft-formats.aspx, dostęp: kwiecień 2012.

Landis J.R., Koch G.G., *The measurement of observer agreement for categorical data*, „Biometrics", 1977, 33 (1), ss. 159-174.

Lazarsfeld P.F., *Notes on the History of Quantification in Sociology - Trends, Sources and Problems*, „Isis", 1961, 52 (2), ss. 277-333.

Lem S., *Bomba megabitowa*, Wydawnictwo Literackie, Kraków 1999.

Lem S., *Cave Internetum*, w: http://www.szkolareklamy.pl/sections-viewarticle-398-str_w7-naj_w1.html, dostęp: czerwiec 2012.

Lem S., Fiałkowski T., *Świat na krawędzi*, Wydawnictwo Literackie, Kraków 2007.

Lissowski G., Haman J., Jasiński M., *Podstawy statystyki dla socjologów*, Wydawnictwo Naukowe „Scholar", Warszawa 2008.

Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data*, John Wiley and Sons, Nowy Jork 2002.

Łuczyński R., *Badania stanów prawnych przy nabywaniu nieruchomości pod drogi krajowe*, „Przegląd Geodezyjny", 2007, 79 (3), ss. 14-16.

Malarska A., *Statystyczna analiza danych wspomagana programem SPSS*, SPSS Polska, Kraków 2005.

Mann H.B., Whitney D.R., *On a Test Of Whether One of Two Random Variables in Stochasticall Large Then The Other*, „The Annals of Mathematical Statistics", 1947, 18 (1), ss. 50-60.

Mari D.D., Kotz S., *Correlation and Dependence*, World Scientific Publishing, Londyn 2004.

Materiały pomocnicze do studiowania statystyki. Wspomaganie komputerowe, I. Kasperowicz-Ruka (red.), Szkoła Główna Handlowa w Warszawie, Warszawa 2004.

Mayntz R., Holm K., Hübner P., *Wprowadzenie do metod socjologii empirycznej*, Wydawnictwo Naukowe PWN, Warszawa 1985.

McClelland G., *Range Restrict,* w: http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/RangeRestrict.html, dostęp: kwiecień 2012.

McCullough B.D., Heiser D.A., *On the accuracy of statistical procedures in Microsoft Excel 2007*, „Computational Statistics & Data Analysis", 2008, 52 (10), ss. 4570-4578.

McMullen L., *Student as a man*, „Biometrika", 1939, 30, ss. 205-210.

McNemar Q., *Note on the sampling error of the difference between correlated proportions or percentages*, „Psychometrika", 1947, 12, ss. 153-157.

Mider D., Marcinkowska A., *Przemoc w kulturze politycznej polskiego Internetu*, „Studia Politologiczne", 2011, 21, ss. 239-298.

Milbrath L.W., *Political Participation. How and Why Do People Get Involved in Politics?*, Rand McNally & Company, Chicago 1965.

Moszczyński L., *Interpretacja współczynnika kurtozy w analizie danych*, „Przegląd Elektrotechniczny", 2003, 79, 9 (1), ss. 558-560.

Nawojczyk M., *Przewodnik po statystyce dla socjologów*, SPSS Polska, Kraków 2002.

Ohri A., *Interview Professor John Fox Creator R Commander*, w: http://www.decisionstats.com/interview-professor-john-fox-creator-r-commander/, dostęp: kwiecień 2012.

Ohri A., *Interview: dr Graham Williams*, w: http://www.decisionstats.com/interview-dr-graham-williams/, dostęp: kwiecień 2012.

Okrasa W., *Funkcjonowanie i efektywność zespołów badawczych*, Zakład Narodowy im. Ossolińskich, Wrocław 1987.

Oppenheim A.N., *Kwestionariusze, wywiady, pomiary postaw*, Wydawnictwo Zysk i S-ka, Poznań 2004.

Ostasiewicz W., *L.A.J. Quetelet: patriarcha statystyki*, „Śląski Przegląd Statystyczny", 2006, 5 (11), ss. 5-24.

Ostasiewicz W., *Refleksje o statystyce wczoraj, dziś i jutro, Statystyka wczoraj, dziś i jutro. I Ogólnopolski Zjazd Statystyków z okazji 95-lecia Polskiego Towarzystwa Statystycznego i 90-lecia Głównego Urzędu Statystycznego*, „Biblioteka Wiadomości Statystycznych", 2008, t. 56, Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne.

Pavkov T.W., Pierce K.A, *Do biegu, gotowi - start! Wprowadzenie do SPSS dla Windows*, tłum. J. Buczny, Gdańskie Wydawnictwo Psychologiczne, Gdańsk 2005.

Pearson E.S., *Studies in the History of Probability and Statistics. XX: Some Early Correspondence Between W.S. Gosset, R.A. Fisher and K. Pearson With Notes And Comments*, „Biometrika", 1968, 55, ss. 445-457.

Pearson K., *Historical Note on the Origin of the Normal Curve of Errors*, „Biometrika", 1924, 16 (3-4), ss. 402-404.

Pearson K., *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, „Philosophical Magazine", 1900, 50 (302), ss. 157-175.

Pfaff B., *Improving Virtual Hardware Interfaces*, w: http://benpfaff.org/papers/thesis.pdf, dostęp: kwiecień 2012.

Pfaff B., *prywatna strona internetowa*, w: http://benpfaff.org/, dostęp: luty 2012.

Pilch T., *Zasady badań pedagogicznych*, Wydawnictwo „Żak", Warszawa 1998.

Plackett R.L., *Karl Pearson and the Chi-Squared Test*, „International Statistical Review", 1983, 51, ss. 59-72.

Polskie Generalne Studium Wyborcze 2007, w:, http://www.ads.org.pl/opis-szczeg.php?id=72, dostęp: lipiec 2012.

PTBRiO, katalog 2011/12, Edycja XVI, Polskie Towarzystwo Badaczy Rynku i Opinii, ss. 45-46.

Randolph J.J., *Online Kappa Calculator*, 2008, w: http://justus.randolph.name/kappa, dostęp: kwiecień 2012.

Rao C.R., *Statystyka i prawda*, Wydawnictwo Naukowe PWN, Warszawa 1994.

Roa C.R., *Karl Pearson Chi-Square Test: The Dawn of Statistical Inference*, w: *Goodness-of-Fit Tests and Model Validity*, C. Huber-Carol (red.), Birkhauser, Boston 2002.

Rodgers J.L., Nicewander W.A., *Thirteen Ways to Look at the Correlation Coefficient*, „The American Statistican", 1988, 42 (1), ss. 59-66.

Roth P.L., Switzer F.S., *Missing Data: Instrument-Level Heffalumps and Item-Level Woozles*, w: http://division.aomonline.org/rm/1999_RMD_Forum_Missing_Data.htm, dostęp: luty 2012.

Roth P.L., *Missing data: A conceptual review for applied psychologists*, „Personnel Psychology", 1994, 47, ss. 537-560.

Rubin D.B., Little R.J.A., *Statistical analysis with missing data*, John Wiley and Sons, Nowy Jork 2002.

Rucht D., *Rosnące znaczenie polityki protestu*, [w:] *Zachowania polityczne*, t. II, R.J. Dalton, H.-D. Klingemann (red.), Wydawnictwo Naukowe PWN, Warszawa 2010.

Schoier G., *On partial nonresponse situations: the hot deck imputation method*, w: http://www.stat.fi/isi99/proceedings/arkisto/varasto/scho0502.pdf, dostęp: luty 2012.

Sclove S.L., *Notes on Likert Scales*, 2001, w: http://www.uic.edu/classes/idsc/ids270sls/likert.htm, dostęp: wrzesień 2012.

Shafer G., *The Significance of Jacob Bernoulli's Ars Conjectandi for the Philosophy of Probability Today*, „Journal of Econometrics", 1996, 75 (1), ss. 15-32.

Shenk D., *Data Smog. Surviving the information glut*, HarperSanFrancisco, Nowy Jork 1997.

Sheskin D.J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall CRC, Boca Raton, Londyn, Nowy Jork 2004.

Shively W. P., *Sztuka prowadzenia badań politycznych*, tłum. E. Hornowska, Wydawnictwo Zysk i S-ka, Poznań 2001.

Skrzywan W., *Historia statystyki. Materiały do wykładów*, Państwowe Wydawnictwo Naukowe, Warszawa 1954.

Słownik Języka Polskiego, w: http://sjp.pwn.pl/haslo.php?id=2474043, dostęp: kwiecień 2012.

Sobczyk M., *Statystyka*, Wydawnictwo Naukowe PWN, Warszawa 2002.

Spearman C., *Correlation calculated from faulty data,* „British Journal of Psychology", 1910, 3, ss. 271-295.

Spearman C., The *Proof and Measurement of Association between Two Things*, „The American Journal of Psychology", 1904, 15 (1), ss. 72-101.

Stallman R.M., *The Free Universal Encyclopedia and Learning Resource. Announcement of the Project*, „GNU Operating System", w: http://www.gnu.org/encyclopedia/free-encyclopedia.html, dostęp: kwiecień 2012.

Stanley J.C., *The Influence of Fisher's 'The Design of Experiments' on Educational Research Thirty Years Later*, „American Educational Research Journal", 1966, 3 (3), ss. 223-229.

Stanton J.M., *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*, „Journal of Statistics Education", 2001, 9 (3), ss. 1-10.

StatSoft, *Elektroniczny Podręcznik Statystyki PL*, Kraków 2006, w: http://www.statsoft.pl/textbook/stathome.html, dostęp: kwiecień 2012.

Statystyczny drogowskaz. Praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii, S. Bedyńska, A. Brzezicka (red.), Wydawnictwo „Academica", Warszawa 2007.

Stevens S.S., *On the Theory of Scales of Measurement*, „Science", 1946, 103, ss. 677-680.

Student, *An Experimental Determination of the Probable Error of Dr Spearman's Correlation Coefficients*, „Biometrika", 1921, 13 (2-3), ss. 263-282.

Sześć sigma, Encyklopedia Wikipedia, w: http://pl.wikipedia.org/wiki/Sześć_sigma, dostęp: kwiecień 2012.

Szreder M., *Statystyka w państwie demokratycznym*, „Wiadomości Statystyczne", 2009, 6, ss. 6-13.

Sztumski J., *Wstęp do metod i technik badań społecznych,* Wydawnictwo Naukowe PWN, Warszawa 1984.

Szulc S., *Metody statystyczne*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1961.

Thompson B., Vidal-Brown S.A., *Principal Components versus Principle Axis Factors: When Will We Ever Learn*, Nowy Orlean 2001, ss. 1-13.

Uebersax J.S., *Likert Scales: Dispelling the Confusion*, „Statistical Methods for Rater Agreement", 2006, w: http://john-uebersax.com/stat/likert.htm, dostęp: wrzesień 2012.

Wieczorkowska G., Król G., *O typowym zastosowaniu analizy czynnikowej i skalowania w badaniach społecznych,* „Zeszyty Metodologiczne", 1997, 1, ss. 1-29.

Witaszek M., *Miejsce i rola sondaży w badaniu opinii społecznej*, „Zeszyty Naukowe Akademii Marynarki Wojennej", 2007, 4 (171), ss. 141-172.

Verba S., Nie N.H., *Participation in America. Political Democracy and Social Equality*, The University of Chicago Press, Londyn 1972.

Verba S., Scholzman K.L., Brady H.E., *Voice and Equality. Civic Voluntarism in American Politics*, Harvard University Press, Cambridge, Londyn 1995.

Wawrzynek J., *Metody opisu i wnioskowania statystycznego*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, Wrocław 2007.

Wieczorkowska G., Król G., *O typowym zastosowaniu analizy czynnikowej i skalowania w badaniach społecznych*, „Zeszyty Metodologiczne", 1997, 1, ss. 1-29.

Yeomans K.A., Golder P.A., *The Guttman-Kaiser Criterion as a Predictor of the Number of Common Factors,* „Statistician", 1982, 31 (3), ss. 221-229.

Yule G.U., Kendall M.G., *Wstęp do teorii statystyki*, Państwowe Wydawnictwo Naukowe, Warszawa 1966.