# Analiza statystyczna z IBM SPSS Statistics

Justyna Wiktorowicz
Maria Magdalena Grzelak
Katarzyna Grzeszkiewicz-Radulska

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
UNIWERSYTET ŁÓDZKI
Unia Europejska Europejski Fundusz Społeczny

# Analiza statystyczna z IBM SPSS Statistics

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

# Analiza statystyczna z IBM SPSS Statistics

Justyna Wiktorowicz
Maria Magdalena Grzelak
Katarzyna Grzeszkiewicz-Radulska

Łódź 2020

Fundusze Europejskie
Wiedza Edukacja Rozwój

UNIWERSYTET ŁÓDZKI

Unia Europejska
Europejski Fundusz Społeczny

Publikacja opracowana w projekcie pt. „Doskonałość naukowa kluczem do doskonałości kształcenia" realizowanym w ramach Osi Priorytetowej III. Szkolnictwo wyższe dla gospodarki i rozwoju, Działania 3.5 Kompleksowe programy szkół wyższych w ramach Programu Operacyjnego Wiedza Edukacja Rozwój. Projekt realizowany przez Uniwersytet Łódzki w ramach konkursu Narodowego Centrum Badań i Rozwoju POWER.03.05.00-IP.08-00-PZ1/17 na podstawie umowy nr POWR.03.05.00-00--Z092/17-00 z dnia 28.06.2018 r.

Justyna Wiktorowicz, Maria Magdalena Grzelak – Uniwersytet Łódzki
Wydział Ekonomiczno-Socjologiczny, Instytut Statystyki i Demografii
Katedra Statystyki Ekonomicznej i Społecznej, 90-214 Łódź, ul. Rewolucji 1905 r. nr 41/43
Katarzyna Grzeszkiewicz-Radulska – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Instytut Socjologii, Katedra Metod i Technik Badań Społecznych
90-214 Łódź, ul. Rewolucji 1905 r. nr 41/43

RECENZENT
*Iwona Markowicz*

REDAKTOR INICJUJĄCY
*Beata Koźniewska*

REDAKTOR
*Monika Poradecka*

SKŁAD I ŁAMANIE
*Mateusz Poradecki*

KOREKTA TECHNICZNA
*Wojciech Grzegorczyk*

PROJEKT OKŁADKI
*Agencja Reklamowa efectoro.pl*
Zdjęcie wykorzystane na okładce: © Depositphotos.com/Megaloman1ac

Publikacja współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego
Publikacja dystrybuowana bezpłatnie

Wydrukowano z gotowych materiałów dostarczonych do Wydawnictwa UŁ



Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.10189.20.0.K

Ark. druk. 12,75

ISBN 978-83-8220-387-5, e-ISBN 978-83-8220-388-2

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. 42 665 58 63

# Spis treści

# Od autorek

Statystyka towarzyszy nam wszędzie. W mediach udostępniane są wyniki różnych badań prowadzonych za pomocą metod statystycznych. W szkole dowiadujemy się o tym, do czego wykorzystuje się i jak oblicza proste mierniki statystyczne (np. średnią arytmetyczną). Studiując na uczelni wyższej, przekonujemy się, że statystyka ma tak nieoczywiste zastosowania jak badania z dziedziny filologii, a w takich dziedzinach jak ekonomia, zarządzanie, gospodarka przestrzenna, socjologia czy psychologia stanowi nieodłączny element prowadzonych prac badawczych, a tym samym również programów studiów. Biorąc pod uwagę potrzeby pracodawców i rosnący popyt na pracowników wyposażonych choćby w podstawowe umiejętności analityczne, można oczekiwać, że również nasza książka trafi na podatny grunt i spotka się z zainteresowaniem studentów i praktyków.

Inspirację do napisania tej książki stanowiły dla nas doświadczenia, jakie wyniosłyśmy z pracy ze studentami różnych kierunków studiów, prowadzonych szkoleń oraz projektów badawczych. Niejednokrotnie przekonywałyśmy się, że pozornie proste metody mogą się wydawać niezrozumiałe, jeśli nie pokaże się ich praktycznej strony – zwłaszcza obecnie, gdy dostęp zarówno do danych, jak i oprogramowania statystycznego staje się coraz łatwiejszy i rzadko prowadzimy badania bez wykorzystania komputera. Z drugiej strony rodzi to ryzyko wpadnięcia w pułapkę zbytniego zmechanizowania, czasem bezrefleksyjnego stosowania oprogramowania i metod statystycznych, które w danej sytuacji nie powinny być wykorzystywane. A to z kolei prowadzi do nieuprawnionych wniosków, a tym samym błędnej oceny zjawisk, które badamy. W naszej książce starałyśmy się pogodzić obie te perspektywy, łącząc przy tym zagadnienia teoretyczne związane z wybranymi metodami statystycznymi z zaprezentowaniem sposobu wykorzystania w tym zakresie IBM SPSS Statistics (i szerzej – PS IMAGO).

Z założenia książka stanowi wprowadzenie do analizy statystycznej, z położeniem nacisku nie na matematyczne podstawy omawianych metod, a na warunki ich stosowania, ograniczenia i propozycje rozwiązań w przypadku, gdy z różnych powodów optymalna w danej sytuacji metoda nie może być zastosowana. Zdajemy sobie jednak sprawę, że bardziej wprawny Czytelnik może odczuwać potrzebę zgłębienia tajników wiedzy w konkretnym, wąskim obszarze. Mając to na uwadze, ważniejsze, aczkolwiek nie najłatwiejsze dla początkujących badaczy kwestie

uwzględniłyśmy w przypisach. Jednocześnie wskazałyśmy na literaturę przedmiotu, w której znaleźć można szersze ich omówienie. Zagadnienia przedstawiane są na konkretnych przykładach, opartych na realnych danych pochodzących z prowadzonych przez nas (i nie tylko) badań kwestionariuszowych. Rozwiązanie tych problemów badawczych z zastosowaniem IBM SPSS Statistics jest wyjaśniane krok po kroku – poczynając od przesłanek wyboru danej metody statystycznej, przez omówienie ścieżki postępowania w IBM SPSS Statistics, odniesienie do tabel wynikowych, na merytorycznej interpretacji wyników kończąc. Mamy nadzieję, że przyjęta formuła zostanie przez Państwa życzliwie przyjęta.

Niniejsza publikacja powstała w ramach projektu pn. „Doskonałość naukowa kluczem do doskonałości kształcenia", realizowanego przez Uniwersytet Łódzki w ramach środków Programu Operacyjnego Wiedza Edukacja Rozwój, Oś priorytetowa III Szkolnictwo wyższe dla gospodarki i rozwoju, Działanie 3.5 Kompleksowe programy szkół wyższych.

*Zachęcamy do lektury*

# 1. Wprowadzenie do IBM SPSS Statistics

**Kluczowe pojęcia:** Edytor danych, Edytor raportu, Edytor wykresów, tworzenie zbiorów danych, polecenie *Przekształcenia* – przekształcanie zmiennych (obliczanie wartości nowych zmiennych, rekodowanie, zliczanie wystąpień), polecenie *Dane* – operacje związane z obserwacjami (dzielenie na podzbiory, włączanie filtrów, ważenie)

## 1.1. Uwagi wstępne

IBM SPSS Statistics jest programem działającym w środowisku Microsoft Windows, jak również w systemach operacyjnych Linux i MAC OS. W podstawowych funkcjach i trybach pracy IBM SPSS Statistics jest bardzo podobny do innych programów pracujących w tych środowiskach. Główny człon nazwy – SPSS – to skrót od *Statistical Package for Social Sciences*, niemniej jednak jego zastosowanie znacznie wykracza poza zagadnienia sugerowane nazwą pakietu. W IBM SPSS Statistics jest uniwersalnym środowiskiem analizy danych ilościowych. Proponuje użytkownikowi szerokie spektrum technik analitycznych. Program ma budowę modułową, pozwalającą na rozszerzenie możliwości analitycznych wraz z każdym dodatkowym komponentem – od analiz eksploracyjnych, poprzez klasyfikację, grupowanie i redukcję danych, aż po prognozowanie. Wiele uczelni w Polsce wykorzystuje do analizy danych ilościowych PS IMAGO PRO®. To rozwiązanie przygotowane przez Predictive Solutions (dawniej SPSS Polska). PS IMAGO PRO to system analityczno-raportujący, którego najważniejszym elementem jest IBM® SPSS® Statistics.

PS IMAGO oznaczony jest następującą ikoną:  . PS IMAGO PRO® oferuje nie tylko wszystkie funkcje IBM SPSS Statistics, ale także szereg dodatkowych funkcjonalności, które dostarczają wsparcia etapu przygotowania danych do analizy, jak również dodatkowe metody statystyczne oraz nowe typy wizualizacji wyników badań.

Praca z pakietem IBM SPSS Statistics odbywa się w kilku rodzajach okien nazywanych edytorami. Są to przede wszystkim edytory: danych, raportów, wykresów i poleceń. Część edytorów może zostać zapisana jako samodzielny plik – dotyczy to edytora danych, raportów i poleceń. Inne edytory, na przykład wykresów czy tabel, otwierane są z poziomu raportu (dwukrotnie klikamy w dany obiekt, przechodząc w ten sposób do jego edycji, a po zakończeniu zmian zamykamy edytor, bez konieczności wykonania polecenia *Zapisz* w odniesieniu do tego „wewnętrznego" edytora – zmiany są zapisywane automatycznie).

Edytor danych (*Data Editor*), oznaczany ikoną  , służy do wpisywania surowych danych oraz przygotowania swego rodzaju „słowniczka" etykiet zmiennych i ich wartości. Ma on rozszerzenie .sav.

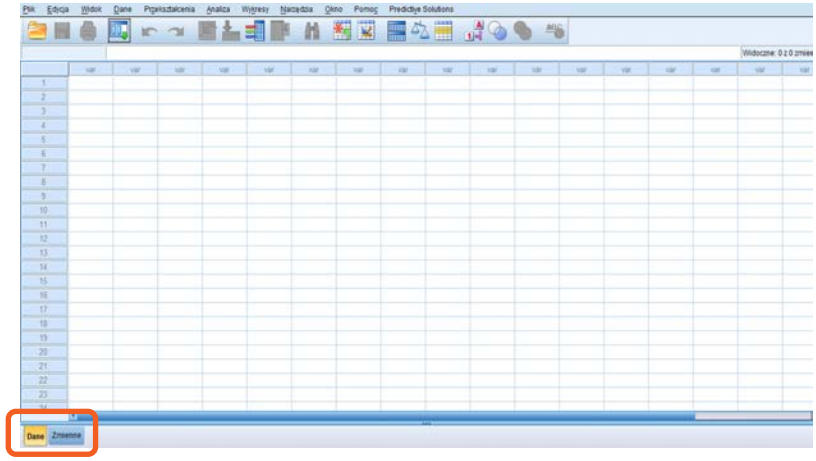
Edytor raportów (*Viewer*), oznaczony ikoną  , to okno, w którym pojawiają się wyniki prowadzonych analiz w postaci zestawień tabelarycznych i obiektów graficznych, a także – w języku poleceń – wykaz zrealizowanych komend (jest to zapisywane w Dzienniku). Edytor raportów ma rozszerzenie .spv (we wcześniejszych wersjach SPSS było to rozszerzenie .spo).

Edytor poleceń (*Syntax Editor*), oznaczony ikoną  , to okno, w którym możemy w języku programowania zapisać polecenia, które zwykle realizujemy za pomocą odpowiedniej opcji w menu. Jest on szczególnie użyteczny przy pracy z dużymi zbiorami danych, przy rozbudowanych analizach.

Praca w SPSS polega, ogólnie rzecz biorąc, na tym, że wydajemy określone polecenia, korzystając z menu lub Edytora poleceń, a wyniki prezentowane są w Edytorze raportu. Należy przy tym pamiętać, że – inaczej niż ma to miejsce przy analizach statystycznych z wykorzystaniem MS Excel – jakakolwiek zmiana wartości zmiennych w Edytorze danych nie przekłada się automatycznie na zmodyfikowanie tabel wynikowych w Edytorze raportu – konieczne jest ponowne wykonanie odpowiedniego polecenia. Menu jest jednak wspólne dla wszystkich trzech edytorów, tzn. polecenia *Analiza*, *Wykresy*, *Dane* czy *Przekształcenia* można wybrać zarówno z poziomu Edytora danych, jak i Edytora raportu czy Edytora poleceń i będą one realizowane zawsze na tych samych danych, zapisanych w Edytorze danych. W Edytorze danych, w przeciwieństwie do MS Excel, nie jest możliwe „przeciąganie" wartości czy nazw zmiennych, można to jednak zrobić, używając standardowej procedury *Kopiuj*/*Wklej* (dotyczy to również opisu wszystkich właściwości zmiennych w zakładce *Zmienne* – wrócimy do tego w dalszej części publikacji).

## 1.2. Edytor danych

Okno Edytora danych zaprezentowano na rysunku 1.1. Menu główne jest zbliżone do tego, jakie mamy w MS Office. Poza standardowymi poleceniami *Plik*, *Edycja*, *Widok*, *Okno*, *Pomoc* SPSS oferuje typowe narzędzia analityczne – polecenia *Analiza*, *Wykresy*. Różnych operacji na wierszach dokonujemy wykorzystując polecenie *Dane*, a operacji na kolumnach – wykorzystując *Przekształcenia*. Dodatkowo możliwe jest wybranie polecenia *Predictive Solutions*, w którym znaleźć można dodatkowe funkcjonalności, związane między innymi z wykresami i przekształceniami.

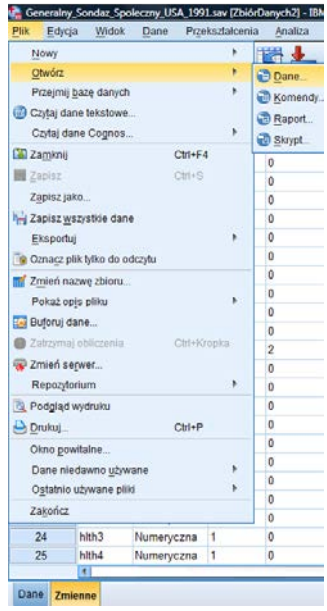



**Rysunek 1.1.** Okno Edytora danych

Istotne jest to, że w SPSS pracujemy na dwóch zakładkach – *Dane* i *Zmienne* (oznaczone na dole ekranu na rysunku 1.1). W zakładce *Dane* wprowadzamy dane liczbowe naszej bazy danych (wartości zmiennych dla poszczególnych jednostek badania), w zakładce *Zmienne* definiujemy zmienne. Sposób konstruowania takiego zbioru danych omówiony zostanie dalej. Menu główne Edytora danych obejmuje, jak wcześniej podkreślano, standardowe polecenia, analogiczne do MS Office. W tym miejscu omówione zostaną tylko te funkcjonalności, które będą szczególnie użyteczne w dalszych analizach.

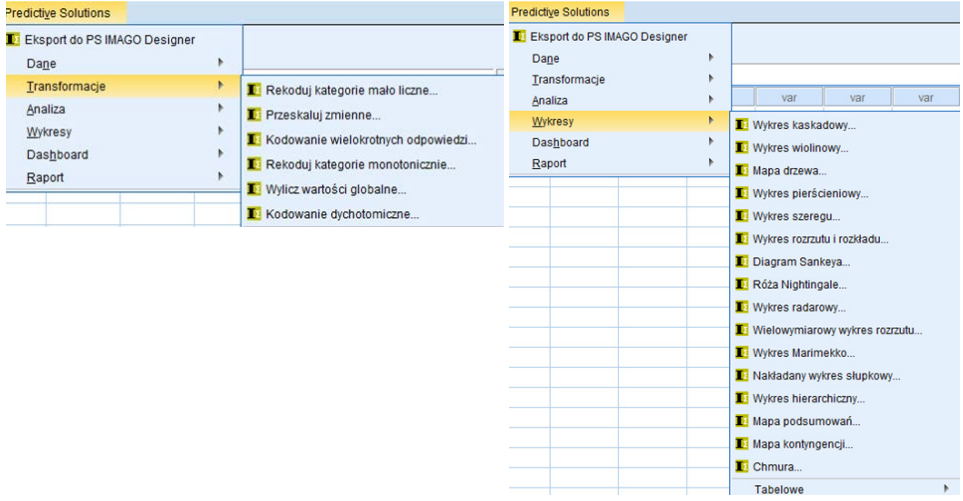
Otwierając zbiór danych w formacie .sav, postępujemy dokładnie tak jak w MS Office: klikamy w ikonę lub wybieramy w menu *Plik* polecenie *Otwórz*, a tam *Dane* (rysunek 1.2). Menu polecenia *Plik* zawiera analogiczne elementy jak w przypadku MS Office – otwieranie nowych plików, zapisywanie zmian, drukowanie itp. działają tak samo. Bardzo użyteczne jest polecenie *Zatrzymaj obliczenia* (zaznaczone na rysunku 1.2 jaśniejszą czcionką, bo nieaktywne). Z polecenia tego

korzystamy, gdy chcemy zatrzymać wykonywanie danego polecenia (np. konstruowanie rozbudowanej tabeli krzyżowej z dokładnym testem Fishera).




**Rysunek 1.2.** Otwieranie gotowych zbiorów danych

Ułatwiające pracę z danymi przekształcenia dostępne są w poleceniu *Predictive Solutions*.




**Rysunek 1.3.** Wybrane funkcjonalności polecenia *Predictive Solutions*

Korzystając z polecenia *Predictive Solutions*, można na przykład dokonać automatycznego przekształcenia zmiennej jakościowej w zmienne zero-jedynkowe (co jest szczególnie użyteczne na etapie konstruowania modeli regresji) – służy do tego opcja *Predictive Solutions → Transformacje → Kodowanie dychotomiczne*. Inne typy transformacji obrazuje rysunek 1.3. Z kolei wykorzystując polecenie *Wykresy*, można wygenerować obiekty graficzne niedostępne w standardowej wersji SPSS, na przykład wykres Marimekko, mapę kontyngencji, chmurę, różę Nightingale itp. Niektóre z nich są dostępne tylko w wersji PS IMAGO PRO[1].

## Tworzenie zbioru danych

W Edytorze danych znajduje się zbiór danych poddawanych analizie. Dane te można wprowadzać bezpośrednio w tym edytorze, można też jako plik .sav otworzyć zbiór utworzony w innym formacie – na przykład .xlsx. Należy pamiętać o tym, żeby – przygotowując zbiór danych – w kolumnach uwzględnić kolejne zmienne, a w kolejnych wierszach wprowadzić dane dla poszczególnych jednostek badania.

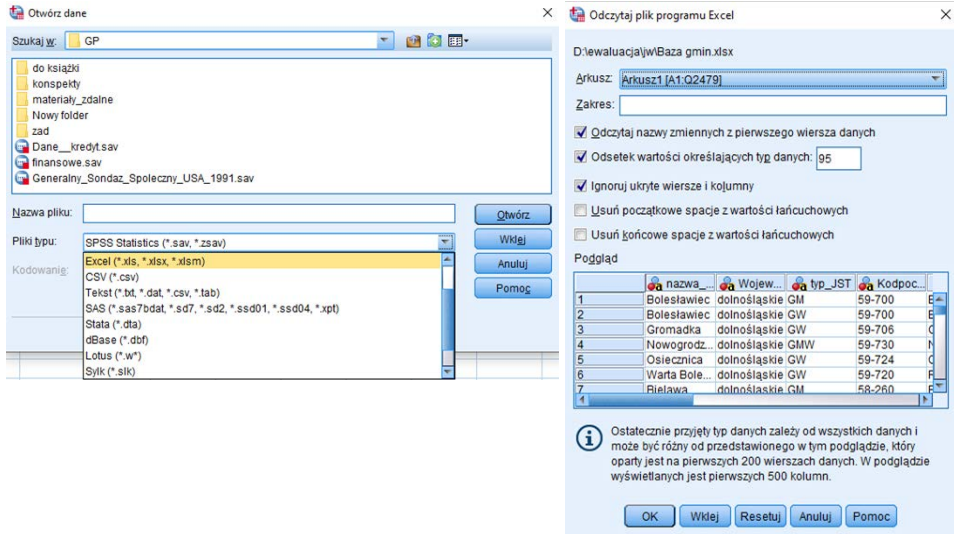
Wprowadzając dane w pierwszej kolumnie, należy zawsze podać kod (ID) jednostki badania (np. respondenta) – pozwoli to na identyfikację rekordu w zbiorze danych z kwestionariuszem, kartą badania itp. W kolejnych kolumnach wprowadzamy dalsze zmienne, zgodnie z kolejnością w wykorzystywanym narzędziu badawczym.

Tworząc zbiór danych, kodujemy je również w odniesieniu do wariantów odpowiedzi. W przypadku zmiennych mierzonych w sposób ilościowy wprowadzamy odpowiednie wartości, a w przypadku zmiennych, których warianty są opisowe, stosujemy ich kody liczbowe. Tym samym, jeśli korzystamy z ankiety on-line, z której uzyskujemy zbiór .xlsx, przed przystąpieniem do analizy najlepiej przekształcić w Excelu (używając polecenia *Zamień*) opisowe warianty na ich liczbowe kody. Używamy wówczas kolejnych numerów do oznaczenia kolejnych wariantów danej zmiennej, trzymając się ściśle kolejności, jaką przyjęliśmy w narzędziu badawczym (np. przy odpowiedziach *tak*/*nie*/*nie wiem*, odpowiedź *tak* kodujemy jako 1, *nie* jako 2, *nie wiem* jako 3). Jest to ważne, ponieważ (1) sprawia, że zbiór danych jest bardziej przejrzysty (bez utraty ważnych informacji, gdyż oznaczenia kodów liczbowych mogą i powinny zostać wyjaśnione w zakładce *Zmienne* – o czym za chwilę) oraz (2) wszystkie opcjonalności polecenia *Analiza* będą dostępne (z czym będą problemy – np. przy ANOVA – w przypadku zmiennych, których warianty są wprowadzone słownie). Podejście to należy uwzględnić, tworząc

1 Szerzej na ten temat w *PS IMAGO PRO…*, 2014.

zbiór danych bezpośrednio w Edytorze danych SPSS i w innych formatach (np. MS Excel).

Aby otworzyć w formacie SPSS inny typ pliku, powinno się wybrać polecenie *Otwórz*, a następnie po rozwinięciu *Pliki typu* (lewy panel na rysunku 1.4) należy wybrać właściwy format, w którym mamy zapisane dane (lub *Wszystkie pliki*) oraz właściwy plik z danymi. Przed utworzeniem pliku w formacie .sav pokaże się okno (prawy panel na rysunku 1.4) z podglądem nowego zbioru danych, w którym dodatkowo mamy możliwość wyboru arkusza pliku .xlsx, w którym znajdują się nasze dane. Jest to bardzo wygodne, bo pozwala nam na przykład w innym arkuszu zostawić dane oryginalne (o ile dokonywaliśmy jakichś przekształceń, np. zamiany wariantów opisowych na ich kody liczbowe) czy legendę do pliku. Tu też mamy oznaczone *Odczytuj nazwy zmiennych z pierwszego wiersza* (wyłączenie tej opcji oznaczać będzie, że to, co wpisano w pierwszym wierszu, będzie traktowane już jako wartości zmiennych) – zostawiamy ten domyślny wybór. Jeśli w arkuszu Excel ukryliśmy jakieś kolumny, które jednak też chcemy pobrać do zbioru, to należy wyłączyć opcję *Ignoruj ukryte wiersze i kolumny*. Jeśli to, co widzimy na podglądzie, odpowiada nam, zatwierdzamy zbiór przyciskiem *OK*. Podobnie jak w przypadku wprowadzania bezpośrednio w SPSS – wypełniony zostanie arkusz danych (zakładka *Dane*) oraz tylko w podstawowym zakresie zakładka *Zmienne*. Dalsza praca z danymi wymagać będzie doprecyzowania charakterystyk zmiennych (ich specyfikacji) w zakładce *Zmienne*.




**Rysunek 1.4.** Otwieranie zbioru danych

## Specyfikacja zmiennych

Specyfikacji zmiennych dokonuje się w zakładce *Zmienne*. Zmiany wprowadzamy wypełniając wprost dane pole (tak jest w przypadku *Nazwy*, *Etykiety*) lub posługując się strzałkami (np. *Szerokość*, *Dziesiętne*), bądź rozwijając okno i naciskając w prawym rogu komórki na szare pole. Należy pamiętać, że dany wiersz w zakładce *Zmienne* zarezerwowany jest dla konkretnej zmiennej (odpowiada określonej kolumnie w zakładce *Dane*), więc jeśli chcemy nadać etykiety wariantom zmiennej *płeć* (zapisać, że 1 oznacza kobietę, a 2 mężczyznę), powinniśmy ustawić się w wierszu, w którym mamy zmienną *płeć*. Poszczególne kolumny zakładki *Zmienne* można scharakteryzować w sposób następujący:

### *Nazwa*

Nazwy zmiennych powinny być stosunkowo krótkie (jeśli pracujemy na kwestionariuszu ankiety, najlepiej, gdyby odpowiadały numeracji pytań, np. P1, P2.4) i powinny być wprowadzone tylko w jednym – pierwszym – wierszu zbioru danych. Nazwy zmiennych w Edytorze danych nie mogą zawierać spacji i różnych znaków systemowych (!, %, ^, &, *, +, /, -, =, ?, ‘, „, (), : , :), nie powinny się też zaczynać od liczby. Kropka może pojawić się w środku nazwy zmiennej, ale nie na końcu. Nie można wpisać dwóch zmiennych o tej samej nazwie. Jeśli nie dostosujemy się do tych zaleceń, SPSS nie przyjmie proponowanej nazwy (pojawi się odpowiedni komunikat). Jednak jeśli w zbiorze danych utworzonym wcześniej w innym formacie (np. MS Excel) pojawią się wymienione elementy, to przy otwieraniu pliku automatycznie nazwy te zostaną skorygowane (np. przed liczbą na początku nazwy pojawi się @).

### *Typ*

Typ zmiennej może być numeryczny, łańcuchowy (tekstowy), czasu itp. Typ ten automatycznie dostosowuje się do sposobu oznaczenia wariantów zmiennej – jeśli użyliśmy symboli liczbowych zamiast opisowych wariantów zmiennej, przypisany zostanie typ numeryczny (i tak to zostawiamy). Typu łańcuchowego używamy tylko wtedy, gdy nie posługujemy się kafeterią odpowiedzi, gdy wprowadzamy odpowiedzi na pytania otwarte. Wybierając typ zmiennej, można również ustalić *Szerokość*, czyli liczbę znaków dopuszczalną dla danej zmiennej, a w przypadku zmiennych numerycznych również liczbę miejsc dziesiętnych. Służą do tego również dwie kolejne kolumny.

### *Szerokość*

Wprowadzamy dopuszczalną liczbę znaków – strzałkami lub ręcznie.

### *Dziesiętne*

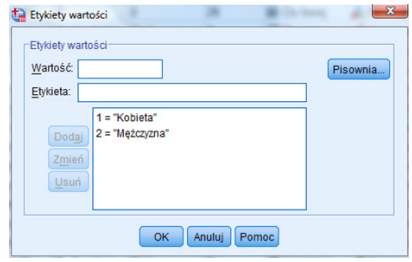
Ustalamy dopuszczalną liczbę miejsce dziesiętnych, zmniejszamy ją lub zwiększamy – w zależności od potrzeb.

### *Etykieta*

W polu tym wprowadzamy pełną nazwę zmiennej, wyjaśniamy, co dokładnie kryje się pod nią. Nie obowiązują nas tu żadne ograniczenia dotyczące formatu zapisu (dopuszczalna jest np. spacja). W przypadku badań kwestionariuszowych najlepiej wpisać lub przekopiować (standardowo – na zasadzie *Kopiuj/Wklej*) pełną treść pytania, dopisując na początku jego numer w kwestionariuszu, na przykład *M1. Płeć*.

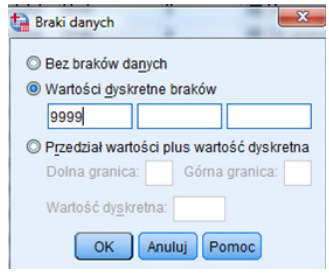
### *Wartości*

Pole to służy do wprowadzania pełnych oznaczeń wariantów zmiennych, jeśli wprowadzając dane, użyliśmy (zgodnie z zaleceniami) ich symboli liczbowych. Etykiet wartości nie wprowadzamy dla zmiennych ilościowych, gdy wariant zmiennej w bazie odpowiada rzeczywistej wartości zmiennej (np. dla zmiennej *Rok urodzenia*). Klikamy po prawej stronie, rozwijając okno *Etykiety wartości*. W polu *Wartość* wpisujemy używany symbol, a w polu *Etykieta* to, co się pod nim kryje. Po wprowadzeniu informacji dla jednego wariantu klikamy w *Dodaj* i wprowadzamy dane dla kolejnych. Na koniec klikamy w *OK*.




### *Braki*

W polu tym możemy wyłączyć pewne wartości zmiennej z analizy. Dotyczy to między innymi braków danych, które oznaczyliśmy na przykład przez 9999. Unikniemy wówczas chociażby zawyżenia wartości średniej arytmetycznej wieku – uwzględnienie tej wartości będzie prowadzić do rozbieżnych z rzeczywistością wyników. Należy wówczas wypełnić pola zgodnie z poniższym schematem. Można jednocześnie wprowadzić trzy wartości, które rozumiemy jako braki danych, lub wskazać przedział liczbowy, w którym się znajdują (plus – o ile jest taka potrzeba – jedną konkretną wartość zmiennej).



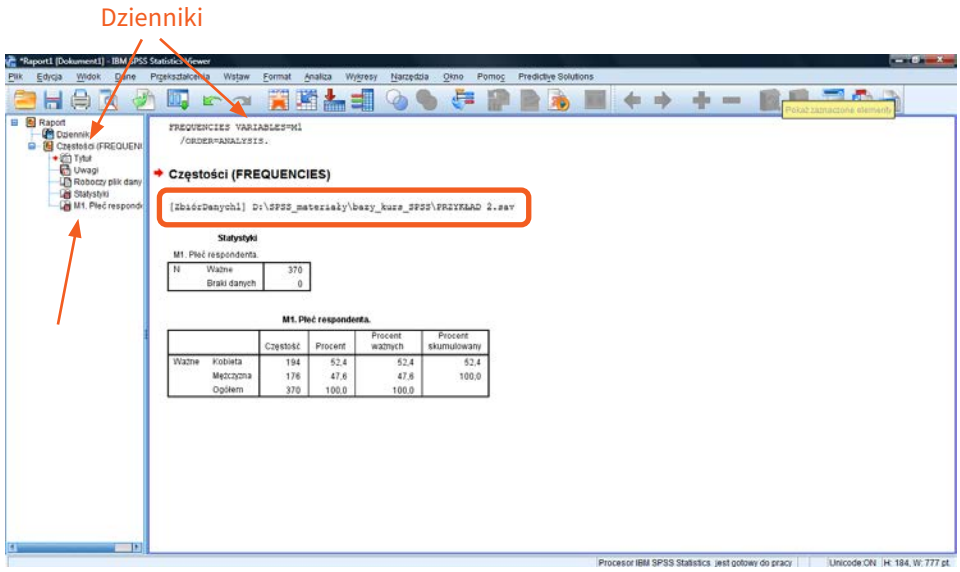

*Poziom pomiaru*

W polu tym ustalamy skalę pomiarową (poziom pomiaru) zmiennej (o skalach pomiarowych nieco więcej w dalszej części książki). Dopuszczalne są trzy opcje: *Nominalna*, *Porządkowa* i *Ilościowa*. Proszę zwrócić uwagę na oznaczenia poszczególnych skal pomiarowych – będą one widoczne przy wyborze danej zmiennej.


W przypadku większości funkcji nieprawidłowe oznaczenie poziomu pomiaru nie będzie miało znaczenia dla prowadzonych analiz, dla niektórych jednak uniemożliwi je. Dlatego warto uporządkować zbiór danych pod tym względem.

## 1.3. Edytor raportów

Okno Edytora raportów przechowuje wyniki analiz statystycznych w postaci tabel lub wykresów. Jego wygląd prezentuje rysunek 1.5. Po prawej stronie (oznaczenie strzałką) mamy możliwość prześledzenia struktury raportu – nazw tabel wynikowych, w tym zmiennych, dla których są budowane. Klikając w odpowiednie pole (np. *M1. Płeć*), przechodzimy do tabeli częstości dla tej zmiennej. Automatycznie po prawej stronie raportu przeniesieni zostaniemy do tego zestawienia wynikowego. Na górze okna pojawia się zapis Dziennika, a w nim – w języku programowania – wykaz zrealizowanych komend. Dodatkowo w oknie Edytora raportów pojawia się ścieżka dostępu Edytora danych, na podstawie którego wygenerowane zostały zestawienia (na rysunku 1.5 otoczona linią). Tabela częstości dla zmiennej *M1. Płeć* w strukturze raportu oznaczona została strzałką.



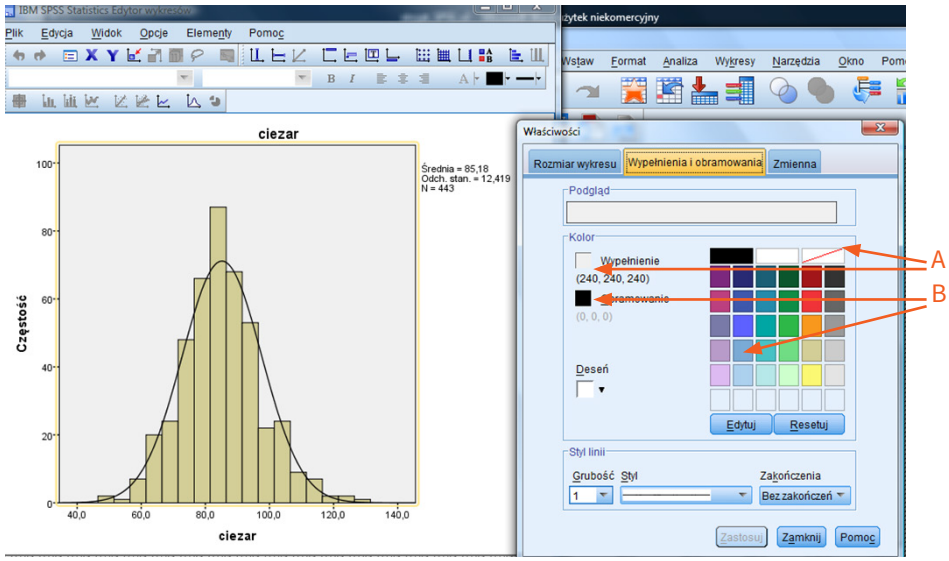

**Rysunek 1.5.** Okno Edytora raportów

## Edytowanie wykresów i tabel w raporcie

Poprzez dwukrotne kliknięcie w dany obiekt (wykres, tabelę) przechodzimy do ich edycji, co umożliwia nam na przykład zmianę kolorystyki wykresu, obramowania itd. (wykorzystując *Format*), ale również transponowanie tabel wynikowych (*Panel przestawiania* działa analogicznie jak w MS Excel). Dodatkowo po dwukrotnym kliknięciu w tabelę wynikową możemy najechać na daną liczbę i wówczas (na żółtym polu) pokazuje się ona z większą dokładnością (z większą liczbą miejsc po przecinku).

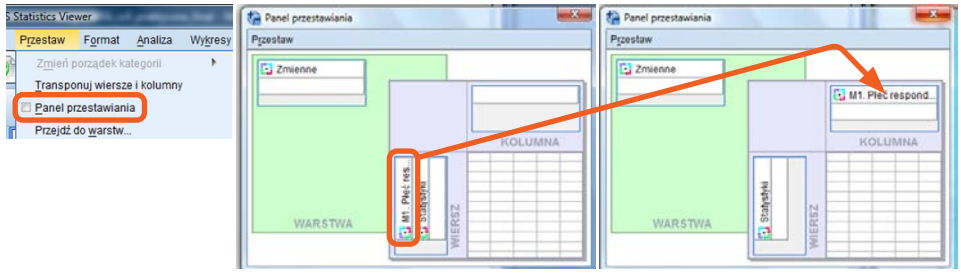
Dwukrotnie klikając w wykres (w Edytorze raportu), przechodzimy do Edytora wykresów, w którym możemy wykres edytować. Jest to ważne, gdyż po skopiowaniu wykresu na przykład do MS Excel lub do MS Word funkcjonuje on jako plik graficzny, którego nie można edytować bez programów graficznych. Wykres kopiowany do Worda powinien być zatem przygotowany wcześniej w Edytorze wykresów. Zależy nam zwykle na zmianie kolorystyki. W Edytorze wykresów należy dwukrotnie kliknąć w obszar wykresu, wówczas otworzy się okno po lewej stronie (rysunek 1.6). Aby wyłączyć kolor tła (na rysunku 1.6 w lewym panelu), klikamy w *Wypełnienie*, a następnie w przekreślony biały prostokąt (strzałka A). Aby zmienić kolorystykę słupków, wykonujemy analogiczne działania, wskazując na obramowanie, a potem odpowiedni kolor (strzałka B). Każdorazowo zmiany zatwierdzamy, wciskając *Zastosuj*. Klikając w dany element wykresu, możemy

poddać go edycji – wykasować lub zmodyfikować tytuły osi, zmienić liczbę miejsc po przecinku na osi x itd.




**Rysunek 1.6.** Wykorzystanie Edytora wykresów do zmiany kolorystyki wykresu

Jeśli układ tabeli nie do końca nam odpowiada, można dokonać jej przestawienia, używając *Panelu przestawiania*. Po dwukrotnym kliknięciu w tabelę przechodzimy do jej edycji (tabela będzie wówczas „otoczona" przerywaną linią). Pojawia się wówczas na pasku narzędzi polecenie *Przestaw*, które wcześniej nie było dostępne. Możemy w prosty sposób dokonać transpozycji tabeli (wybierając *Transponuj wiersze i kolumny*), ale możliwe jest również wykorzystanie *Panelu przestawiania* (rysunek 1.7).




**Rysunek 1.7.** Przykład wykorzystania tabeli przestawnej

Po wybraniu polecenia *Panel przestawiania* pokazuje się okno zaprezentowane w lewym panelu rysunku 1.7. W zaznaczonym polu wskazane jest, jaki jest obecnie

układ danych w tabeli. Jak widać, pierwszym kryterium dzielenia wierszy jest *Płeć*, potem – *Statystyki*. A więc najpierw podzielimy tabelę według płci i w ramach każdej płci będziemy wyznaczać te same statystyki. Jeśli chcielibyśmy mieć porównanie wartości statystyk dla kobiet i mężczyzn w sąsiednich kolumnach, musimy przenieść (przeciągnąć myszką) *Płeć* w nagłówki kolumn (tak jak to oznaczono w prawym panelu rysunku 1.7).

## Eksportowanie raportu

Tabele i wykresy wynikowe SPSS można przenieść do MS Word lub MS Excel, korzystając ze standardowych procedur *Kopiuj/Wklej*. Przy większych zbiorach wynikowych znacznie wygodniejsze jest wykorzystanie polecenia *Eksportuj*. Jest ono dostępne na trzy sposoby: (1) na pasku narzędzi (ikona  ), (2) *Plik* → *Eksportuj*, (3) prawy klawisz myszy → *Eksportuj*. Musimy przy tym znajdować się w odpowiednim Edytorze raportu (a nie Edytorze danych).

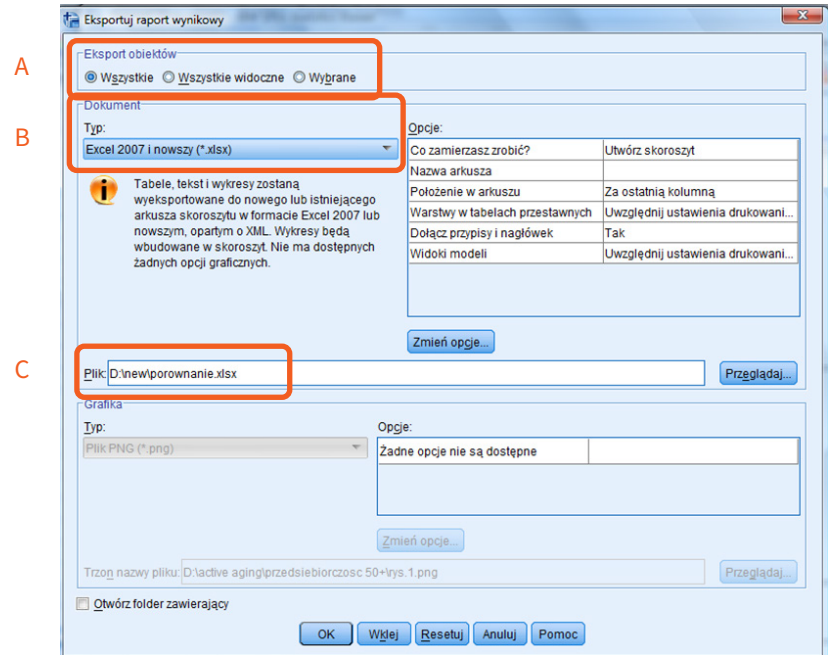
Po wybraniu polecenia *Eksportuj* wyświetla się okno zaprezentowane na rysunku 1.8. Procedura eksportu raportu do różnych programów MS Office jest analogiczna:

A. W polu *Edytor obiektów* wskazujemy, czy eksportujemy cały raport, czy może tylko wybrane zestawienia (muszą być one wcześniej wybrane – podświetlone – w raporcie: zaznaczamy je, klikając w nie w raporcie lub w strukturze raportu po prawej stronie okna edytora; możliwe jest przy tym wybranie jednego obiektu lub – z użyciem *Shift* – wielu).
B. W polu *Typ* wybieramy typ pliku, do którego dokonany zostanie eksport (do wyboru mamy m.in. Excel 2007 i nowszy, Word/RTF, PowerPoint).
C. W polu *Plik* wskazujemy, gdzie i pod jaką nazwą ma być zapisany wyeksportowany raport.

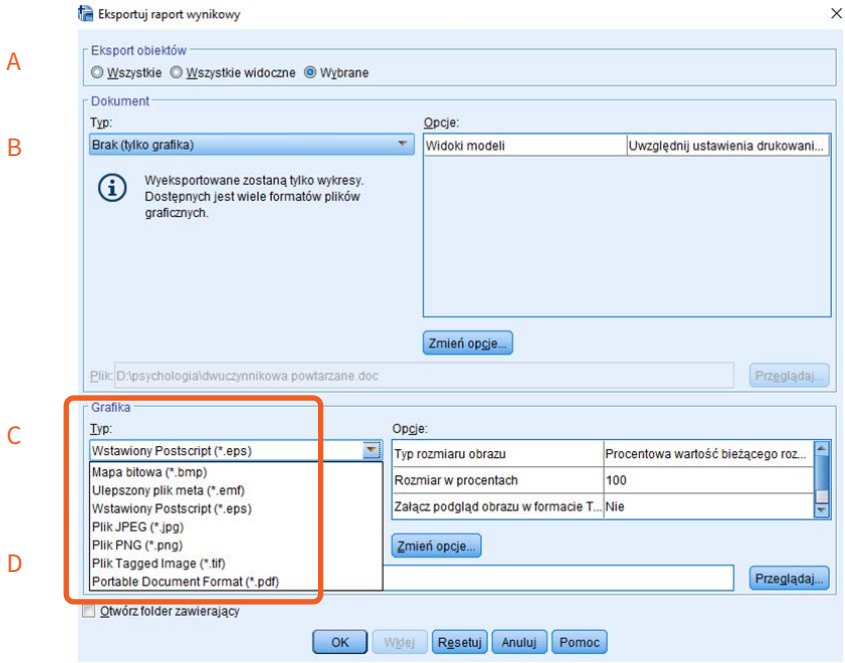
Procedura eksportu plików graficznych obejmuje następujące czynności (rysunek 1.9):

A. Wskazujemy, czy eksportujemy cały raport, czy może tylko wybrane zestawienia.
B. Wybieramy typ pliku, do którego dokonany zostanie eksport – *Brak (tylko grafika)*.
C. Na dole pojawia się nowe okno z listą formatów plików graficznych (*.tif, *.eps, *.jpg, *.png itd.) – wybieramy właściwy.
D. Wskazujemy, gdzie i pod jaką nazwą ma być zapisany wyeksportowany raport.

Każdy obiekt graficzny eksportuje się wówczas jako odrębny plik o nazwie podanej w polu *Plik* (z dopisanym numerem obiektu, tj. *równanie1*, *równanie2* itd.).



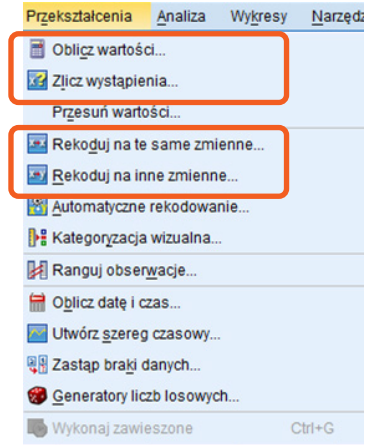

**Rysunek 1.8.** Eksportowanie zawartości raportu do formatu MS Excel lub MS Word




**Rysunek 1.9.** Eksportowanie plików graficznych

## 1.4. Polecenie *Przekształcenia*

Polecenie *Przekształcenia* umożliwia operacje na zmiennych (kolumnach). Szczególnie użyteczne i często wykorzystywane są cztery funkcjonalności: *Oblicz wartości*, *Zlicz wystąpienia*, *Rekoduj na te same zmienne*, *Rekoduj na inne zmienne*.



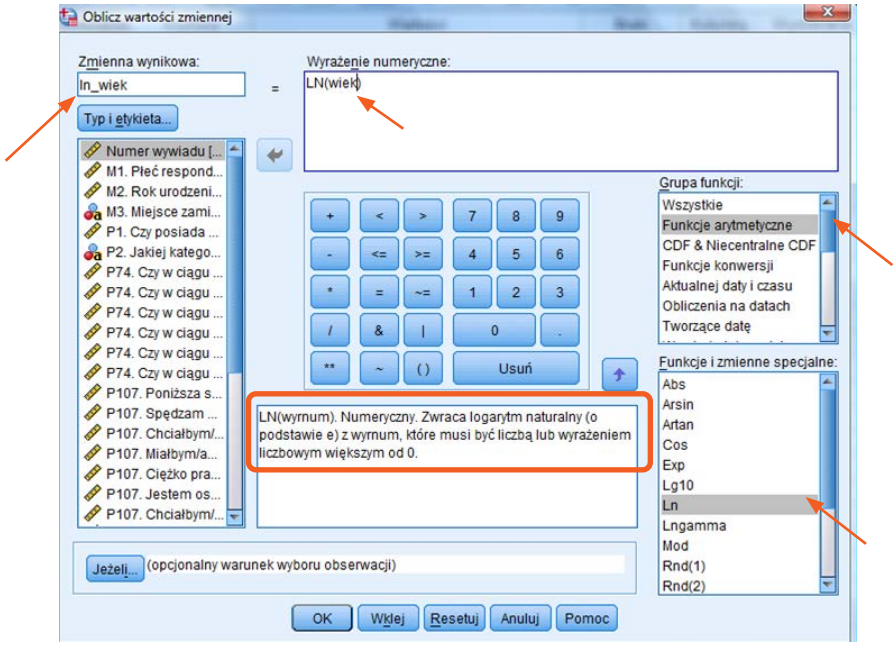

**Rysunek 1.10.** Polecenie *Przekształcenia*

### *Oblicz wartości*

Polecenie to pozwala na wyznaczenie nowej zmiennej na podstawie matematycznych przekształceń zmiennej wyjściowej – analogicznie jak przy wykorzystaniu „funkcji" w Excelu. Po wybraniu *Przekształcenia* → *Oblicz wartości* podajemy nazwę nowej zmiennej (w polu *Zmienna wynikowa*) oraz wskazujemy – w polu *Wyrażenie numeryczne* – jaka operacja matematyczna powinna być wykonana. W polu *Wyrażenie numeryczne* można też wybrać *Grupy funkcji* → *Funkcje arytmetyczne* → i na przykład *ln* (w ten sposób wyznaczymy logarytm zmiennej), potem pozostaje jeszcze przenieść (z listy po lewej stronie) w miejsce znaku zapytania zmienną, dla której liczone będzie wyrażenie – ten wariant jest zobrazowany na rysunku 1.11. Po wybraniu tych poleceń w białym polu pojawia się efekt takiej procedury, tj. zapisane jest, że liczymy LN(wiek), czyli logarytm naturalny ze zmiennej *wiek*.

W jakich sytuacjach zwykle korzysta się z tej funkcjonalności? Na przykład gdy, znając rok urodzenia badanych i wiedząc, że badanie było przeprowadzone w 2020 roku, ustalamy wiek badanych (należy dokonać następujących obliczeń: wiek = 2020 – rok urodzenia). Z polecenia tego korzystamy też wówczas, gdy

chcemy przekształcić logarytmicznie wartości wybranej zmiennej ilościowej (przykładowo dla wieku należy dokonać następujących obliczeń: ln_wiek = ln(wiek)).




**Rysunek 1.11.** Polecenie *Oblicz wartości*

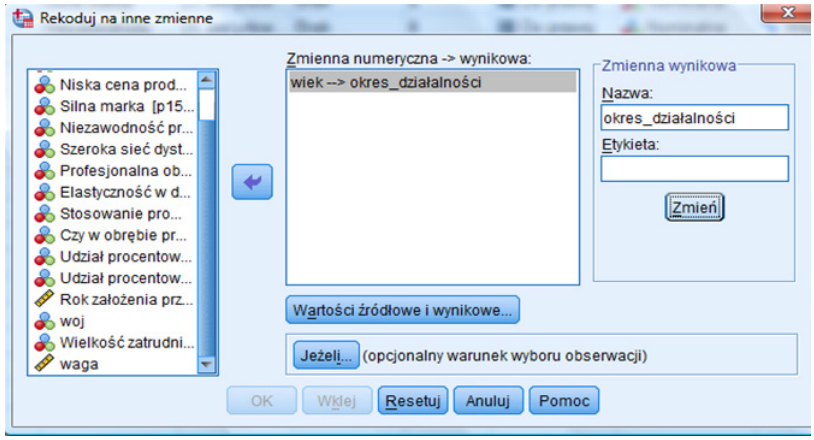
### *Rekoduj na inne zmienne | Rekoduj na te same zmienne*

Funkcje *Rekoduj na inne zmienne* i *Rekoduj na te same zmienne* mogą być wykorzystane jako krok do skonstruowania szeregu rozdzielczego przedziałowego na podstawie indywidualnych wartości cechy albo przy grupowaniu wariantów. Pierwsza z tych opcjonalności jest częściej polecana niż druga. Obie funkcje prowadzą do tego, że oryginalne wartości zmiennych przekształcane są w inne, zdefiniowane przez badacza, przy czym pierwsza z nich tworzy nowe, dodatkowe zmienne według zadanego sposobu, druga zaś powoduje „nadpisanie" nowych wartości na oryginalnych (nie ma więc powrotu do wartości oryginalnych).

W jakich sytuacjach po nie sięgamy? Na przykład wtedy, gdy na podstawie zmiennej *wiek* tworzymy zmienną *okres działalności*, dokonując grupowania wariantów do przedziałów klasowych: do 10 lat, 11–20 lat, 21 lub więcej (jak zostało zobrazowane na rysunku 1.12). Bardzo często po rekodowanie sięga się przy zmiennych mierzonych na skali Likerta. Przykładowo: gdy chcemy przekształcić zmienną mierzoną na pięciostopniowej skali Likerta na zmienną mierzoną

na trzystopniowej skali Likerta (łącząc warianty „zdecydowanie tak" i „raczej tak" oraz „zdecydowanie nie" i „raczej nie"), a także wtedy, gdy odpowiedź „trudno powiedzieć" chcemy przyjąć jako środek skali (np. dla zmiennej, której wyjściowe warianty są następujące: „zdecydowanie tak", „raczej tak", „raczej nie", „zdecydowanie nie", „trudno powiedzieć").

Co konkretnie powinniśmy zrobić w SPSS? Kolejne kroki zobrazowano na rysunkach 1.12 i 1.13. Po wybraniu polecenia *Przekształcenia → Rekoduj na inne zmienne* w pole *Zmienna numeryczna → Zmienna wynikowa* przenosimy zmienną poddawaną rekodowaniu (oryginalną zmienną w zbiorze danych), a w polu *Nazwa* wprowadzamy nazwę nowej zmiennej i koniecznie zatwierdzamy nazwę poprzez *Zmień* (niewykonanie tego kroku powoduje najczęściej pojawiający się problem – brak możliwości zatwierdzenia polecenia przyciskiem *OK*).




**Rysunek 1.12.** Polecenie *Rekoduj na inne zmienne*

Następnie wybieramy *Wartości źródłowe i wynikowe* i w nowym oknie definiujemy sposób rekodowania zmiennej (przekształcania jej wariantów). Podajemy przy tym wszystkie warianty zmiennej, również te, które pozostawiamy bez zmian (wskazujemy na przykład, że wartość źródłowa 1 pozostaje wartością wynikową 1). W ten sposób można przekształcić więcej niż jedną zmienną, przy czym każdej z nich trzeba nadać nową nazwę.

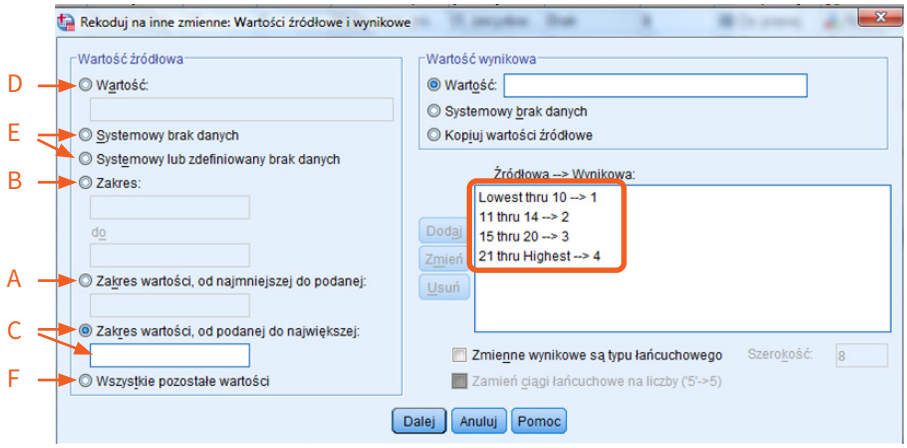
Okno *Wartości źródłowe i wynikowe* (rysunek 1.13) skonstruowane jest w ten sposób, że po lewej stronie podaje się oryginalne wartości zmiennej (*Wartość źródłowa*), a po prawej oznaczenie wariantu przekształconej zmiennej (*Wartość wynikowa*). Wartości źródłowe mogą być oznaczone jako:

- wartości nie wyższe niż podana (wybieramy *Zakres wartości, od najmniejszej do podanej*, oznaczone przez A na rysunku 1.13, a dalej wpisujemy górną granicę przedziału); na przykład w przypadku podanego na rysunku 1.13 okresu

działalności, jeśli dla pierwszego przedziału *do 10* wpiszemy w pole liczbę 10, wówczas po prawej stronie pojawia się zapis *Lowest thru 10 → 1*, a więc wartości „do 10" przekształcone zostaną w „1" (na dalszym etapie w zakładce *Zmienne* nadamy odpowiednie etykiety temu wariantowi zmiennej);

- wartości z określonego zakresu (wybieramy *Zakres*, oznaczony przez B na rysunku 1.13, a dalej wpisujemy dolną i górną granicę przedziału); na przykład dla przedziału *11–14* wpiszemy w pola liczby 11 i 14 – wówczas po prawej stronie pojawia się zapis *11 thru 14 → 2*, a zatem wartości „11–14" przekształcone zostaną w „2";
- wartości nie niższe od podanej (*Zakres wartości, od podanej do największej*, oznaczony przez C na rysunku 1.13, a dalej wpisujemy dolną granicę przedziału); wykorzystamy je na przykład dla przedziału *powyżej 20*, wpisując w pole liczbę 21 (20 nie wchodzi w ten przedział); wówczas po prawej stronie pojawia się zapis *21 thru Highest → 4*, wartości wyższe od 20 przekształcone zostaną zatem w „4";
- za pomocą konkretnej liczby (zaznaczamy pole *Wartość*, a następnie wpisujemy wybraną wartość zmiennej w pole oznaczone przez D na rysunku 1.13).

Można również przekształcić *Systemowy brak danych*, a także *Systemowy lub zdefiniowany brak danych* w dowolną liczbę (oznaczone przez E na rysunku 1.13), a po wybraniu interesujących nas wartości pozostałe przekształcić według określonego schematu (wybieramy *Wszystkie pozostałe wartości* – oznaczone przez F na rysunku 1.13).




**Rysunek 1.13.** Polecenie *Rekoduj na inne zmienne* – pole *Wartości źródłowe i wynikowe*

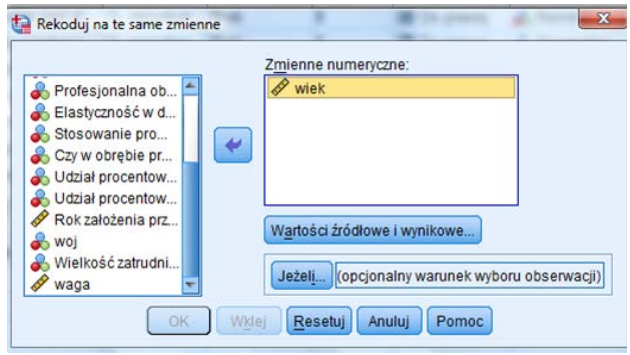
Wartości wynikowe najczęściej przyjmowane są za pomocą kolejnych liczb naturalnych, możliwa jest również zamiana wybranej wartości na *Systemowy brak*

*danych*. Jako wartości wynikowe nie wpisujemy oczekiwanych przedziałów klasowych, tylko ich oznaczenia liczbowe (kolejne liczby naturalne). Podobnie jak w przypadku innych zmiennych nadamy im ostatecznie etykiety zgodnie z zasadami przyjętymi w zakładce *Zmienne Edytora Danych*. Każdorazowo po wprowadzeniu wartości źródłowych i wynikowych zatwierdzamy wybór, wciskając *Dodaj*.

Analogicznie działa funkcja *Rekoduj na te same zmienne*. Różnice są dwie:

- nie nadajemy nazw zmiennych (bo nie tworzymy nowych zmiennych);
- można wskazać do rekodowania tylko wybrane warianty zmiennej.

Po wybraniu polecenia *Przekształcenia* → *Rekoduj na te same zmienne* w pole *Zmienne numeryczne* przenosimy zmienną (zmienne) poddawaną rekodowaniu (rysunek 1.14).

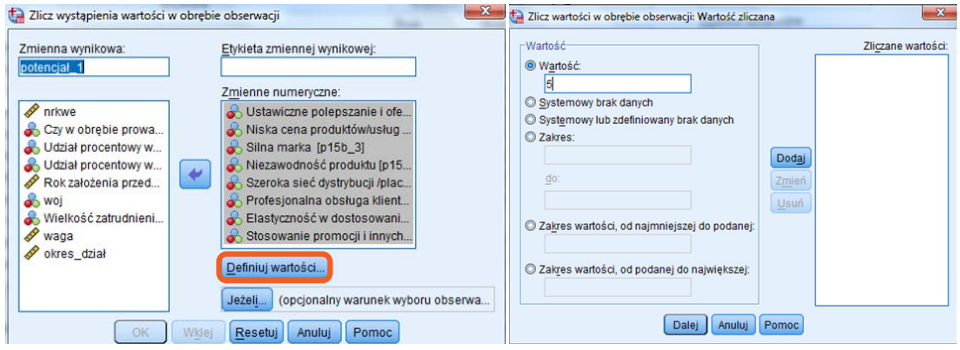



**Rysunek 1.14.** Polecenie *Rekoduj na te same zmienne*

Pole *Wartości źródłowe i wynikowe* definiujemy analogicznie jak na rysunku 1.13. Wskazujemy przy tym tylko te warianty, które chcemy przekształcić (przekształcone będą tylko wskazane warianty, a niewskazane pozostaną w przekształcanej kolumnie niezmienione). Funkcjonalność ta przydaje się zwłaszcza wtedy, gdy przekształcamy systemowe braki danych na przykład w zero albo na odwrót (gdy np. wprowadzając odpowiedzi na pytania wielokrotnego wyboru, nie kodowaliśmy odpowiedzi zero-jedynkowo, a używaliśmy tylko 1, zostawiając puste miejsce tam, gdzie nie wybrano danej odpowiedzi; jeśli dla takiej zmiennej chcemy zbadać zależność, nie będzie to możliwe – program będzie rozpoznawał tę zmienną jako stałą, z jedyną wartością „1", dlatego potrzebne jest rekodowanie systemowych braków danych, np. na 0).

### *Zlicz wystąpienia*

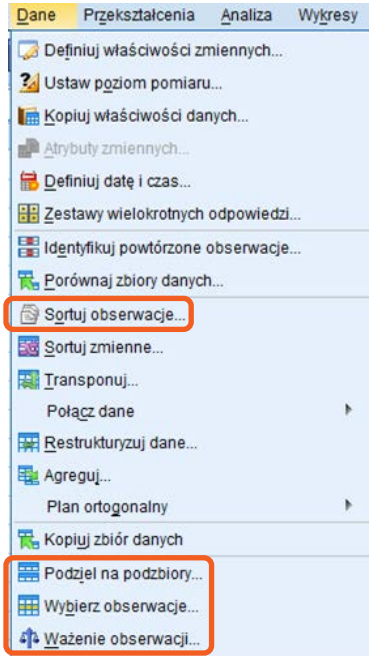
Polecenie to służy do zliczenia, ile razy w zestawie zmiennych (a nie obserwacji) pojawia się dana wartość lub dane wartości. Wymagane jest, aby zmienne poddawane zliczaniu miały typ numeryczny (mogą to być warianty zmiennej jakościowej, które oznaczone są numerycznie). Na przykład gdy uznajemy, że przedsiębiorstwo ma potencjał rozwojowy, jeśli spełniony jest przynajmniej jeden z ośmiu warunków (stanowiących odrębne zmienne), możemy sprawdzić, ile warunków zostało spełnionych, korzystając właśnie z tej funkcji. Jeśli wartość utworzonej nowej zmiennej jest większa od zera, tzn. że przynajmniej jeden z warunków został spełniony, a zatem przedsiębiorstwo ma potencjał rozwojowy. Na rysunku 1.15 przyjęto z kolei, że zliczamy, ile spośród zmiennych opisujących potencjał rozwojowy przedsiębiorstwa każdy z badanych ocenił na 5. Wykonanie tego polecenia wymaga wybrania *Przekształcenia* → *Zlicz wystąpienia*, a w polu *Zmienna wynikowa* nadajemy nazwę nowej zmiennej. W pole *Zmienne numeryczne* przenosimy zmienne, których dotyczy zliczanie (w omawianym przykładzie będą to zmienne wyrażające poszczególne warunki charakteryzujące potencjał rozwojowy przedsiębiorstwa). Z kolei w polu *Definiuj wartości* wskazujemy, jaka wartość (wartości) ma być zliczana, zatwierdzamy przez *Dodaj* (dla każdej wskazanej wartości), a po wprowadzeniu wszystkich kryteriów zatwierdzamy przez *Dalej*, a całość polecenia przez *OK*.




**Rysunek 1.15.** Wykonywanie polecenia *Zlicz wystąpienia*

## 1.5. Polecenie *Dane*

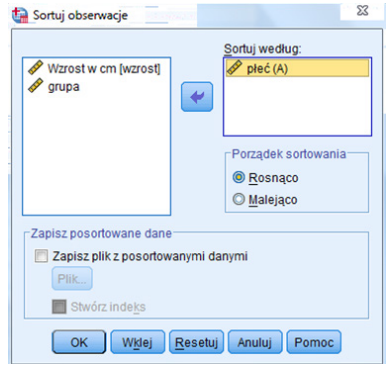
Polecenie *Dane* służy do wykonywania różnych operacji na obserwacjach (wierszach). W przypadku tego polecenia szczególnie ważne są następujące funkcjonalności: *Sortuj obserwacje*, *Podziel na podzbiory*, *Wybierz obserwacje* i *Ważenie obserwacji*.




**Rysunek 1.16.** Funkcjonalności polecenia *Dane*

## *Sortuj obserwacje*

Polecenie to nie wymaga większych wyjaśnień – należy jedynie wskazać, według jakiej zmiennej ma przebiegać sortowanie i w jakim porządku (rosnącym czy malejącym). Na listę można dodać jednocześnie kilka zmiennych, ich kolejność na niej decyduje o tym, według jakiego kryterium sortowanie będzie przebiegać w pierwszej, drugiej itd. kolejności. W tym celu wybieramy *Dane* → *Sortuj obserwacje* (rysunek 1.17). W pole *Sortuj według* przenosimy zmienną (jeśli chcemy sortować ze względu na jedno zjawisko) lub zmienne (jeśli chcemy sortować ze względu na kilka zjawisk). Ustalamy też *Porządek sortowania* – rosnący lub malejący (przy czym dla każdej wybranej zmiennej można wybrać dowolny, niekoniecznie ten sam porządek sortowania).



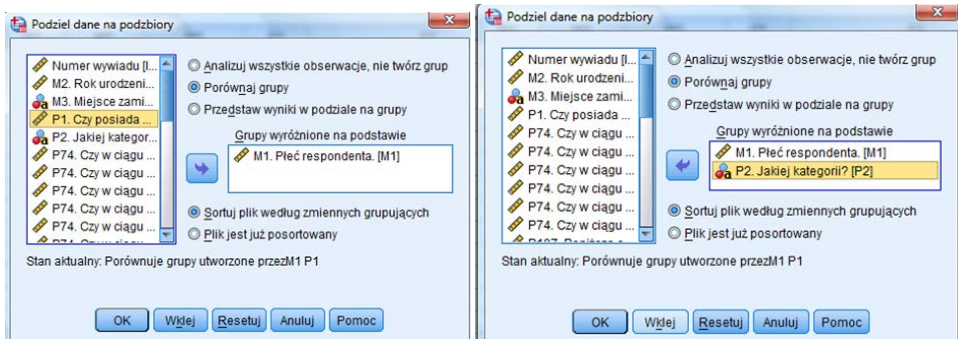

**Rysunek 1.17.** Wykonywanie polecenia *Sortuj obserwacje*

### *Podziel na podzbiory*

Polecenia to umożliwia dokonanie podziału zbiorowości na grupy na podstawie jednego lub kilku kryteriów. Kolejność wprowadzanych zmiennych decyduje o kolejności podziału na podzbiory (np. płeć na pierwszym miejscu, grupa krwi na drugim miejscu da podział na kobiety i mężczyzn, a wśród obu wariantów osobno – na podzbiory osób o konkretnej grupie krwi). Polecenie to stosujemy na przykład gdy:

- wyznaczamy statystyki opisowe dla zmiennej *wiek* osobno dla osób mieszkających na wsi i w miastach;
- zmienna *płeć* jest zmienną kontrolną dla relacji między zadowoleniem z życia a warunkami pracy (mierzonymi z różnych punktów widzenia), badamy więc zależność, konstruując modele regresji osobno dla kobiet i mężczyzn;
- oceniamy poparcie dla poszczególnych partii politycznych z uwzględnieniem przekroju płci i miejsca zamieszkania (miasto/wieś).

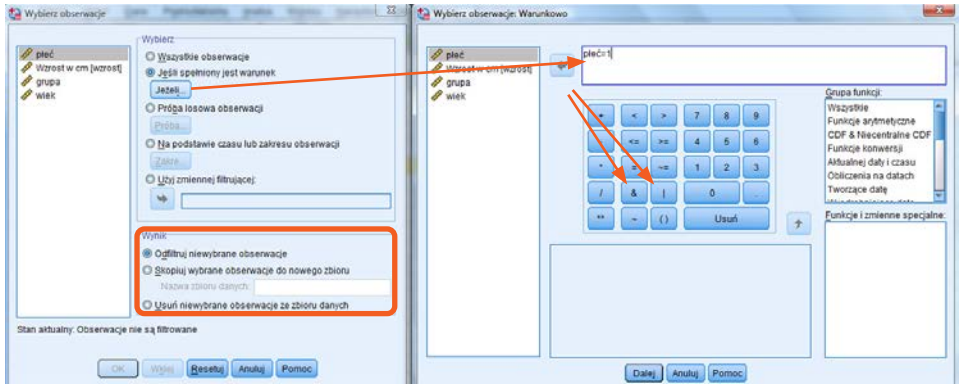
Aby dokonać podziału, wybieramy *Dane* → *Podziel na podzbiory*, a następnie *Porównaj grupy* i przenosimy zmienną (jeśli chcemy dzielić zbiorowość ze względu na jedną cechę) lub zmienne (jeśli chcemy dzielić zbiorowość ze względu na kilka cech – rysunek 1.18). Możliwe jest też wybranie opcji *Przedstaw wyniki w podziale na grupy* – wówczas najpierw pojawiają się wszystkie zestawienia dla jednej grupy (np. kobiet), a potem dla drugiej (w tym wypadku mężczyzn). Przy porównaniu dwóch zbiorowości znacznie wygodniejsza jest opcja *Porównaj grupy* (wtedy obok siebie możemy zestawić wyniki dla obu grup, co ułatwia analizę porównawczą). Należy pamiętać, że podział ten jest aktywny dopóty, dopóki go nie wyłączymy. Jeśli chcemy zrezygnować z podziału na podzbiory, wciskamy *Resetuj* lub *Analizuj wszystkie obserwacje, nie twórz grup*.




**Rysunek 1.18.** Wykonywanie polecenia *Podziel na podzbiory*

## *Wybierz obserwacje*

Polecenie to pozwala na prowadzenie analiz dla zawężonego zbioru, na przykład wyłącznie dla kobiet (m.in. gdy chcemy wyznaczyć statystyki opisowe dla zmiennej *wiek* dla osób mieszkających w miastach lub ocenić poparcie dla poszczególnych partii politycznych wśród kobiet mieszkających w miastach). W tym celu wybieramy *Dane* → *Wybierz obserwacje*, a następnie *Jeśli spełniony jest warunek* → *Jeżeli* i przenosimy zmienną oraz zapisujemy warunek, np. *płeć*=1 (gdzie *1* oznacza kobiety, więc będzie to skutkować wybraniem do analizy wyłącznie kobiet). Wybieramy też, w jaki sposób będzie „nakładany" filtr, poprzez przycisk *Wynik* – zwykle decydujemy się na *Odfiltruj niewybrane obserwacje*, dzięki czemu w każdej chwili możemy wrócić do analizy pełnego zbioru danych (wyłączyć filtr – por. lewy panel na rysunku 1.19).

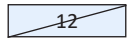


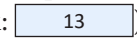

**Rysunek 1.19.** Wykonywanie polecenia *Wybierz obserwacje* – wybór podpróby według wskazanych kryteriów

Możemy równocześnie włączyć kilka kryteriów podziału, używając symboli & i | (oznaczono je strzałkami na prawym panelu rysunku 1.19). Przykładowo:

- jeśli chcemy wybrać kobiety (*płeć1=1*) z grupy pierwszej (*grupa=1*), kryterium będzie wyglądać w następujący sposób: *płeć=1* & *grupa=1* (& oznacza, że wybieramy część wspólną, „iloczyn" obu zbiorów);
- jeśli chcemy wybrać wszystkie kobiety i jednocześnie wszystkie osoby z grupy pierwszej (niezależnie od płci), kryterium wyboru będzie wyglądać tak: *płeć=1* | *grupa=1* (wybieramy zbiorowość będącą sumą obu zbiorów).

Istotne jest również to, jaki format wynikowy ma przyjąć takie ograniczenie zbiorowości. Mamy trzy opcje (obwiedzione na rysunku 1.19 linią, lewy panel):

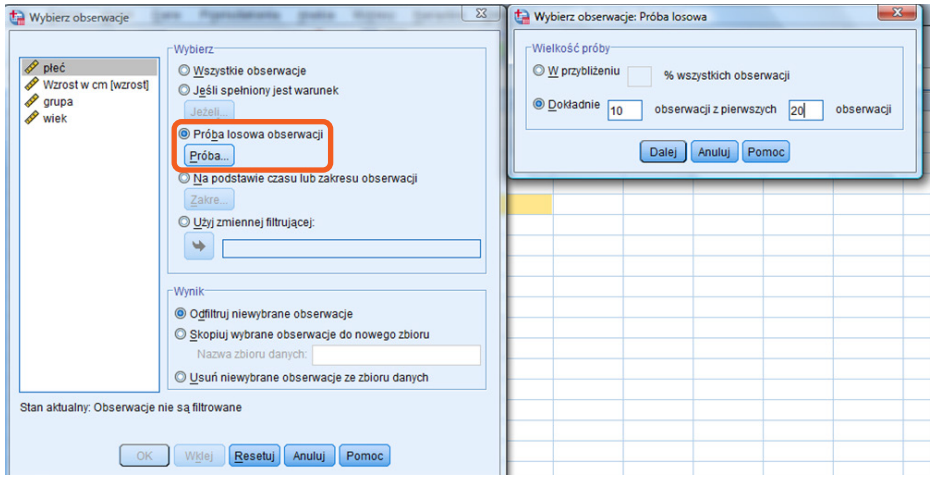
- *Odfiltruj niewybrane obserwacje* – żadna jednostka nie jest wówczas usuwana z bazy, jedynie chwilowo wyłączona z analizy, co wyróżnione jest w sposób następujący: 

(dla porównania, obserwacja brana pod uwagę w analizie oznaczona jest tak: 

);
- *Skopiuj wybrane obserwacje do nowego zbioru* – jednostki spełniające zadane kryteria są wybierane i tworzony jest nowy Edytor danych – zawierający tylko te jednostki; należy wpisać nazwę nowego zbioru, tyle że jest on tworzony tylko w pamięci wirtualnej komputera, a nie jest „fizycznie" zapisywany na dysku (jeśli chcemy to zrobić, trzeba wykorzystać standardową procedurę *Zapisz*);
- *Usuń niewybrane obserwacje ze zbioru danych* – jednostki spełniające zadane kryteria są wybierane, a pozostałe usuwane z aktualnego zbioru danych.

Wykorzystując polecenie *Wybierz obserwacje*, można też przeprowadzić proste losowanie jednostek. Wśród funkcjonalności tego polecenia mamy bowiem opcję *Próba losowa obserwacji* (lewy panel na rysunku 1.20).

W przypadku gdy wybierzemy opcję *Próba*, pojawia się okno zaprezentowane w prawym panelu rysunku 1.20. Mamy dwie możliwości – możemy przyjąć, że ma to być na przykład 10% wszystkich obserwacji albo że ma to być określona liczba spośród wszystkich obserwacji (przyjmijmy, że mamy ich 20). W pierwszym przypadku wybieramy *W przybliżeniu 10% wszystkich obserwacji* (wskazujemy, że wylosowana próba ma obejmować 10% wyjściowej zbiorowości), w drugim zaś wybieramy *Dokładnie 10 spośród 20 pierwszych obserwacji* (ta sytuacja jest zobrazowana na rysunku 1.20). Następnie ustalamy, jak ma być „realizowany" *Wynik* – zwykle w takiej sytuacji wybieramy *Skopiuj wybrane obserwacje do nowego zbioru*.

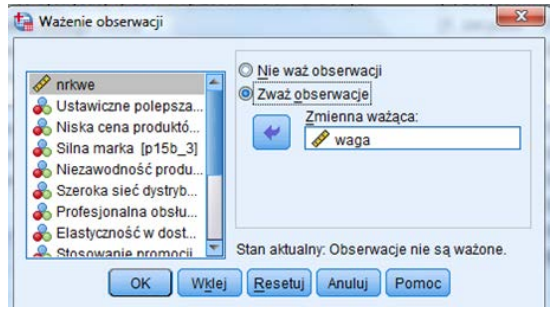



**Rysunek 1.20.** Wykonywanie polecenia *Wybierz obserwacje* – losowanie próby

### *Ważenie obserwacji*

Polecenie to wykorzystywane jest do wyrównania struktury próby do struktury populacji ze względu na cechy brane pod uwagę w celu zapewnienia reprezentatywności próby. Zwykle ma to miejsce, gdy próba dobrana została w sposób nieproporcjonalny albo gdy *response rate* (zwrotność ankiet) była zróżnicowana w różnych grupach, w związku z czym niektóre grupy są nadreprezentowane, a inne niedoreprezentowane. Na wartości zmiennych nakłada się wtedy wagi, zwiększające znaczenie grup zbyt mało licznie występujących w próbie oraz zmniejszające znaczenie grup nadreprezentowanych. Wagi te są wspólne dla całego zbioru danych i muszą zostać wprowadzone przez nas do zbioru danych (jako zmienna *waga*). Te same wagi nakładane będą na wszystkie zmienne (czyli badając jakąkolwiek zmienną, np. jej średnia liczona będzie jako ważona, tj. wartości zmiennych zostaną wcześniej przemnożone przez odpowiednie wagi). Co istotne, ważenie danych pozwala na wnioskowanie na podstawie wyników z próby na populację generalną w sytuacji nieproporcjonalnego doboru próby losowej. Przykładowo: w badaniu dotyczącym mieszkańców Polski na podstawie danych GUS zbadano strukturę kobiet i mężczyzn w pięcioletnich grupach wieku. Wyznaczono analogiczne odsetki (kobiet i mężczyzn z pięcioletnich grup wieku) w obrębie próby. Okazało się, że – pomimo założeń badania i starań ankieterów – kobiety z najstarszych grup wieku występują w próbie nieco częściej niż w populacji, niedoreprezentowani są zaś młodzi mężczyźni. Analiza statystyczna wyników powinna przewidywać wyrównanie struktury próby do struktury populacji ze względu na płeć i grupy wieku poprzez wagi analityczne wyznaczone dla każdej obserwacji (osoby) ze względu

na te dwie cechy jednocześnie. Aby to zapewnić, włączamy wagi – wybieramy *Dane* → *Ważenie obserwacji*. W pole *Zważ obserwacje* przenosimy zmienną, której wartości to wagi analityczne (w przykładzie jest to zmienna *waga*). Należy pamiętać, że *Ważenie obserwacji* jest aktywne dopóty, dopóki go nie wyłączymy. Jeśli chcemy zrezygnować z niego, wciskamy *Resetuj* lub *Nie waż obserwacji* (rysunek 1.21).




**Rysunek 1.21.** Wykonywanie polecenia *Ważenie obserwacji*

Omówione funkcjonalności polecenia *Dane* nie są oczywiście jedynymi dostępnymi w SPSS operacjami, jakie można wykonać na wierszach (dla obserwacji). Możliwe jest również wykonanie innych niż przedstawione operacji na kolumnach (zmiennych) z wykorzystaniem polecenia *Przekształcenia*. Z uwagi na ramy tego opracowanie nie będziemy ich jednak omawiać.

# 2. Metody statystyczne – podstawowe zagadnienia

**Kluczowe pojęcia:** skala pomiarowa, statystyka opisowa *versus* wnioskowanie statystyczne, populacja generalna *versus* próba, badania sondażowe *versus* badania eksperymentalne, estymator, test statystyczny, zasady weryfikacji hipotez

## 2.1. Uwagi wstępne

Prace badawcze podporządkowane są zawsze określonemu celowi. W zależności od tego, jak jest on sformułowany, dobieramy odpowiednie metody i techniki badawcze, w odpowiedni sposób pozyskujemy dane, opracowujemy je i analizujemy. Dane te mogą pochodzić ze źródeł pierwotnych – gromadzonych na potrzeby realizacji celów danego badania, lub wtórnych – pozyskanych w innych celach, które mogą zostać wykorzystane również dla realizacji naszych celów (Sobczyk, 1998, s. 15–16). W niniejszej publikacji skupiamy się na metodach analizy danych pochodzących ze źródeł pierwotnych, w tym zwłaszcza badań kwestionariuszowych, prowadzonych z wykorzystaniem wystandaryzowanego kwestionariusza badania. Z tego typu badań pochodzić będą analizowane przykłady. Pragniemy przy tym podkreślić, że omawiane tu metody znajdują zastosowanie również dla innych typów danych, także tych pochodzących ze źródeł wtórnych (choć rzecz jasna nie wszystkie w tych samych sytuacjach).

Każdy projekt badawczy, niezależnie od jego skali, wymaga odpowiedniego zaplanowania badania oraz opracowania i analizy jego wyników. Na każdym z tych etapów mają zastosowanie metody statystyczne. Wybór tych metod jest powiązany zarówno z podmiotem badania, jak i jego przedmiotem. Już planując badanie, określamy badaną zbiorowość (podmiot badania) i definiujemy badane zjawiska (przedmiot badania). W kolejnych punktach tego rozdziału omówione zostanie, w jaki sposób wpływa to na procedury statystyczne.

## 2.2. Zbiorowość statystyczna a wybór procedury statystycznej

Prowadząc badanie pierwotne na podstawie próby badawczej (badanie częściowe), ważnym zagadnieniem, które powinniśmy wziąć pod uwagę, jest reprezentatywność próby. **Próba reprezentatywna** powinna być dobrana w sposób losowy i być wystarczająco liczna (Wiktorowicz, 2004a, s. 34). Ustalenie minimalnej liczebności próby to jeden z pierwszych etapów badania, w których zastosowanie metod statystycznych jest konieczne. Z uwagi na ramy niniejszego opracowania nie będziemy szczegółowo zajmować się tymi zagadnieniami[2].

W zależności od tego, czy badamy populację generalną, czy próbę, a jeśli próbę – czy dobrana została w sposób pozwalający na uogólnienia (a więc w sposób zapewniający jej reprezentatywność), możemy zastosować metody statystyczne z innej grupy. Ogólnie rzecz biorąc, z formalno-statystycznego punktu widzenia metody statystyczne dzieli się na metody statystyki opisowej (opisu statystycznego) i statystyki matematycznej (wnioskowania statystycznego) (Wiktorowicz, 2004b, s. 23). **Metody statystyki opisowej** służą do opracowania danych z obserwowanej zbiorowości, a więc w przypadku badań częściowych tylko z próby, a dla badań całkowitych – oczywiście z całej populacji (Dobrowolska, Grzelak, Jarczyński, 2017). Oznacza to, że metody statystyki opisowej można zastosować tylko w odniesieniu do przebadanych jednostek, a więc w przypadku badań kwestionariuszowych (które w przeważającej mierze mają charakter częściowy) tylko do próby; nie pozwalają one na uogólnienia poza próbę (Szymczak, 2010, s. 92). Natomiast gdy wnioskujemy o populacji generalnej na podstawie wyników dla reprezentatywnej próby, stosujemy **metody statystyki matematycznej** (Domański i in., 2014, s. 11). O metodach tych będzie mowa w dalszej części tego rozdziału – omówienie wybranych z nich stanowi treść kolejnych części tej publikacji.

2 Zagadnienia metodologiczne związane z planowaniem badania, w tym doborem próby, zostały w przystępny sposób omówione między innymi w: Frankfort-Nachmias, Nachmias, 2001; Babbie, 2006; Nowak, 2007; Rószkiewicz i in., 2013. Szczegółowe wyjaśnienia w tym zakresie, uwzględniające mocną podbudowę statystyczną, znaleźć można na przykład w: Zasępa, 1972; Kordos, 1988; Steczkowski, 1995; Bracha, 1996; Sarndal, Swenson, Wretman, 1997; Domański, Pruska, 2000; Wywiał, 2003; 2010; Szreder, 2010a; 2010b. Warto dodać, że w badaniach sięga się również po metody repróbkowania statystycznego (*resampling*), a wśród nich zwłaszcza metody bootstrapowe i permutacyjne. Wnioskowanie statystyczne jest w ich przypadku przeprowadzane nie na pewnej skończonej próbie, ale na sztucznie utworzonej populacji generalnej, która powstaje poprzez repróbkowanie (wielokrotne losowanie, zaleca się nie mniej niż tysiąc powtórzeń) z tej skończonej, reprezentacyjnej próby. W ten sposób nie jest konieczne badanie założeń dotyczących populacji, gdyż otrzymana poprzez repróbkowanie próba traktowana jest jako populacja. W podejściu tym estymuje się między innymi przedziały ufności i na ich podstawie ocenia się zgodność parametrów w kilku populacjach czy też współczynniki regresji. Szerzej na ten temat por. na przykład Domański i in., 2014; Kończak, 2016.

Badania częściowe mogą mieć charakter badań sondażowych lub eksperymentalnych. Jak pisze Kowal (1998, s. 14–15), badania kwestionariuszowe (do których przede wszystkim odwołujemy się w tej książce) utożsamiane są z **badaniami sondażowymi (surveyowymi)**. Są to badania obserwacyjne, wykorzystywane zwłaszcza gdy chcemy poznać opinie na jakiś temat (aczkolwiek nie tylko), stąd określa się je też czasem jako badania jakościowe (bo dostarczają jakościowej oceny badanych zjawisk). Z uwagi na wysoki stopień standaryzacji badania tego typu (Krzewińska, Grzeszkiewicz-Radulska, 2013) dostarczają danych umożliwiających zastosowanie w ich analizie metod statystycznych (z natury opartych na procedurach matematycznych). Dlatego tak ważne jest odpowiednie opracowanie danych pochodzących z tego typu badań, w tym zwłaszcza właściwe utworzenie zbioru danych (o czym była mowa w poprzednim rozdziale). W przypadku tego typu badań bardzo ważne jest zapewnienie odpowiednich warunków jego przeprowadzania, tak aby spełnione były kryteria dobrego badania, tj. sprawdzalność, możliwość przenoszenia, niezawodność i wiarygodność (Brzezińska, Rycielski, Sijko, 2010, s. 42–72). Wymaga to dużej troski o przeprowadzenie badania, z zachowaniem ściśle określonych reguł gwarantujących wysoką jego jakość na każdym etapie.

**Badania eksperymentalne** z kolei prowadzone są zwłaszcza w celu wykazania wpływu jakiegoś czynnika (wrócimy do tej klasyfikacji w rozdziale szóstym). Badacz zmienia w nich warunki i obserwuje skutki, jakie to przynosi. W psychologii stosuje się (jako podstawowy) podział na badania eksperymentalne i korelacyjne. Mówiąc najprościej, w badaniu eksperymentalnym badacz sam tworzy „zmienność" między jednostkami, na jedną z grup działając bodźcem (np. podając określony lek), a na inną nie (stosując placebo zamiast leku), podczas gdy w badaniach korelacyjnych analizuje się istniejącą zmienność między jednostkami. Badania eksperymentalne pozwalają zatem na identyfikację związków między zjawiskami. W naukach społecznych (obejmujących ekonomię i finanse, nauki o zarządzaniu i jakości, geografię społeczno-ekonomiczną i gospodarkę przestrzenną, ale też np. nauki socjologiczne czy psychologię) znaczenie tych metod jest zróżnicowane. W ekonomii, socjologii czy gospodarce przestrzennej wykorzystuje się przede wszystkim badania sondażowe (które w swojej istocie odpowiadają korelacyjnym badaniom psychologicznym), podczas gdy w psychologii większe znaczenie przywiązuje się do badań eksperymentalnych. Zaletą badań eksperymentalnych jest to, że związki można interpretować w kategoriach relacji przyczynowo-skutkowych (jeśli uda się rzeczywiście zapewnić ścisłą kontrolę zmiennych sytuacyjnych, nie wymaga to budowania modeli, wystarczy zastosowanie prostych testów statystycznych). Z kolei w badaniach sondażowych (korelacyjnych) żadne pojedyncze kryterium obserwacyjne (czynnik) nie może być w pełni trafne, dla dokonania satysfakcjonującej oceny wpływu czynnika konieczne jest równoczesne uwzględnienie wielu kryteriów

(konstruujemy zatem model uwzględniający wiele czynników jednocześnie i mający podbudowę teoretyczną). W badaniach szuka się również rozwiązań pośrednich, łączących zalety obu podejść – proponuje się eksperymenty społeczne czy badania quasi-eksperymentalne (Jaworska, 2004, s. 126–147).

W badaniach sondażowych istnieje czasem potrzeba wyrównania struktury próby do struktury populacji, gdyż z punktu widzenia kluczowych zmiennych, z uwagi na które starano się zapewnić reprezentatywność próby, jej struktura różni się od struktury populacji. Jak podkreślano, dzieje się tak na przykład wówczas, gdy celowo dobieramy próbę w sposób nieproporcjonalny, aby zapewnić wystarczającą liczebność w każdej z wyróżnionych warstw (np. w badaniach przedsiębiorstw nadreprezentowane są duże podmioty), ale też wtedy, gdy w poszczególnych warstwach stopa zwrotu (*response rate*) jest różna (np. gdy w wylosowanej próbie osoby starsze rzadziej odmawiają udziału w badaniu niż osoby młodsze). W takiej sytuacji należy zastosować **wagi analityczne** i przed użyciem testów statystycznych uruchomić ważenie próby (Rószkiewicz i in., 2013).

Musimy mieć przy tym świadomość, że – jeśli badanie realizowane jest na podstawie próby, a nie populacji generalnej – z uwagi na prowadzenie analiz w warunkach niepewności „dowody empiryczne" dostarczają jedynie przesłanek za prawdziwością stawianych hipotez. Jeśli z kolei badamy populację generalną, nie jest nam potrzebne wnioskowanie statystyczne – analiza nie jest prowadzona w warunkach niepewności (przebadaliśmy wszystkie jednostki należące do populacji generalnej, możemy więc stawiać wnioski kategoryczne, a nie tylko przypuszczenia co do prawidłowości zachodzących w populacji generalnej) – stosuje się metody statystyki opisowej. Metody z tej grupy zastosujemy również wtedy, gdy próba nie spełnia warunków reprezentatywności (zwłaszcza gdy została dobrana w sposób celowy). Oznacza to jednak, że wnioski w takiej sytuacji można odnosić tylko do przebadanej próby (nie jesteśmy uprawnieni do uogólniania ich na populację generalną).

Następne ważne zagadnienie to schemat przeprowadzanego badania, a konkretnie to, czy pomiar jest niezależny, czy zależny. **Pomiar niezależny** dotyczy odrębnych podpopulacji. Przykładowo: gdy porównujemy kobiety i mężczyzn pod względem poziomu wynagrodzeń, mamy do czynienia z pomiarem niezależnym. Również gdy porównujemy grupę badaną z kontrolną, dokonywany jest niezależny pomiar w obu zbiorowościach. Z **pomiarem zależnym** mamy do czynienia głównie w badaniach eksperymentalnych, ale też w ewaluacji czy badaniach marketingowych. Dotyczy to zwłaszcza schematów typu pre-test – post-test, w których sprawdzamy skuteczność danej terapii, reklamy, wsparcia w ramach środków EFS itp. W badaniach tego typu dokonuje się pomiaru przed rozpoczęciem działania bodźcem (np. przed rozpoczęciem terapii), a następnie po zakończeniu działania bodźcem. Często wprowadza się dodatkowy, odroczony pomiar (np. w trzy

miesiące po zakończeniu terapii). Możliwe jest również przeprowadzenie analiz uwzględniających oba schematy (np. gdy sprawdzamy, czy poprawa stanu zdrowia oceniana w kolejnych dobach nastąpiła szybciej przy różnych dawkach leku).

## 2.3. Sposób pomiaru zjawisk jako kryterium wyboru metod statystycznych

Wybór metody statystycznej jest podyktowany również tym, jakie właściwości jednostek badania analizujemy. Przedmiotem analizy statystycznej są, rzecz jasna, te właściwości, które mają przynajmniej dwie różne wartości. Określamy je jako **cechy statystyczne** lub **zmienne**[3] (Bedyńska, Cypryańska, 2013a, s. 24).

W badaniach pierwotnych, zwłaszcza kwestionariuszowych, mamy wpływ na to, jakie zmienne uzyskamy. Tak więc już planując badanie, powinniśmy mieć świadomość, jakie metody statystyczne będą najbardziej adekwatne do stawianych problemów badawczych, i dostosować do nich sposób pomiaru zjawisk. Możemy na przykład zadać pytanie tak, aby uzyskać zmienną pozwalającą na wyznaczenie średniej arytmetycznej.

Sposób pomiaru zjawisk wiąże się z ich **skalą pomiarową (poziomem pomiaru)**, rozumianą jako wzorzec dokonywania pomiaru (Nawojczyk, 2002, s. 38). Skale pomiarowe dzieli się ogólnie na metryczne i niemetryczne. Najczęściej stosowany podział poziomów pomiaru rozróżnia skale: nominalną, porządkową, interwałową i ilorazową (Stevens, 1951). Należy pamiętać, że kolejność skal określa ich poziom (moc, siłę). Skala nominalna i porządkowa należą do skal niemetrycznych, a interwałowa i ilorazowa są metryczne. Powszechnie w badaniach obie skale metryczne traktuje się wspólnie jako skalę ilościową – tak też przyjęte jest w większości pakietów statystycznych, w tym w SPSS. W naukach eksperymentalnych zmienne mierzone na skali nominalnej i porządkowej określa się najczęściej jako dyskretne, a mierzone na skali ilościowej jako ciągłe.

**Skala nominalna** wykorzystywana jest do mierzenia zjawisk mających charakter jakościowy (np. płeć, opinia na temat partii politycznej, fakt posiadania dzieci itp.). Dokonując pomiaru na skali nominalnej, przypisujemy jednostki do określonej kategorii zmiennej, opisującej własności tej zmiennej. Kategorie te są opisane słownie, można je jednak wyrazić również za pomocą liczb. Liczby te odróżniają jedynie jedną kategorię od innej. Przy nominalnym poziomie pomiaru nie mogą być dokonywane inne operacje matematyczne ani logiczne (np. >, <, + itp.). Możliwe jest tylko porównywanie wariantów zmiennej (=, ≠).

3 Również w tej publikacji pojęcia te będą używane zamiennie.

W przypadku **skali porządkowej** porównujemy jednostki i oceniamy je, przyjmując za kryterium to, czy przejaw danej zmiennej w jednej jednostce jest większy, równy, czy mniejszy niż w innej. Nadajemy rangi, będące kolejnymi numerami jednostek w uporządkowanym szeregu. W przypadku gdy jednej kategorii odpowiada więcej niż jedna jednostka, mówimy o rangach wiązanych. Początek skali ustalamy wtedy arbitralnie. W odniesieniu do skali porządkowej wiadomo, że druga z kategorii przewyższa tę pierwszą pod względem nasilenia zmiennej, nie wiadomo jednak o ile. Nie można zatem określić różnicy między poszczególnymi wariantami zmiennej. Możliwe jest jedynie stosowanie operacji logicznych typu $>$, $<$, $=$, $\neq$.

Pomiar powinien być przeprowadzony na **skali przedziałowej** (interwałowej) lub **ilorazowej** (stosunkowej), jeśli jego celem jest nie tylko identyfikacja, stopniowanie, ale i pomiar poziomu zjawiska, które jest mierzalne. Na tych poziomach każdej jednostce badania przypisywana jest wartość liczbowa, a nie tylko własność (dająca się stopniować lub nie) określonej zmiennej. Skala przedziałowa pozwala na ustalenie odległości między wariantami zmiennej. W przeciwieństwie do skali ilorazowej nie ma jednak naturalnego (absolutnego, bezwzględnego) zera, przez co nie można ustalić, jaka jest absolutna wielkość poszczególnych punktów skali. Skala ilorazowa z kolei pozwala na ustalenie naturalnego punktu zerowego (Górniak, Wachnicki, 2008, s. 92–93). Liczby przypisane poszczególnym kategoriom zmiennej są proporcjonalne do stopnia, w jakim poszczególnym elementom tych kategorii przysługuje mierzona własność. Umożliwia to takie operacje matematyczne jak dzielenie, mnożenie czy pierwiastkowanie.

To samo zjawisko może być często zmierzone na różne sposoby, a tym samym na różnej skali pomiarowej, a to z kolei determinuje możliwości zastosowania określonych metod analizy statystycznej. Na przykład badając zaufanie, możemy zadać pytanie w jednej z trzech wersji, uzyskując zmienne mierzone na różnych skalach pomiarowych (tabela 2.1).

**Tabela 2.1.** Propozycje pomiaru zaufania a skala pomiarowa

| Wersja | Treść pytania | Warianty odpowiedzi | Źródło | Skala pomiarowa |
|---|---|---|---|---|
| v1 | Ogólnie rzecz biorąc, czy uważa Pan, że można ufać większości ludzi, czy też sądzi Pan, że w postępowaniu z ludźmi ostrożności nigdy za wiele? | 1. Większości ludzi można ufać.<br>2. Ostrożności nigdy za wiele. | *Diagnoza społeczna* | Nominalna |
| v2 | Ogólnie rzecz biorąc, czy uważa Pan, że można ufać większości ludzi? | 1. Zdecydowanie nie.<br>2. Raczej nie.<br>3. Ani tak, ani nie.<br>4. Raczej tak.<br>5. Zdecydowanie tak. | Nie dotyczy | Porządkowa |

| Wersja | Treść pytania | Warianty odpowiedzi | Źródło | Skala pomiarowa |
|---|---|---|---|---|
| V3 | Wskaźnik syntetyczny liczony na podstawie 15 itemów pytania:<br>Czy ma Pan zaufanie do:<br>2.1. banków komercyjnych?<br>2.2. Narodowego Banku Polskiego?<br>2.3. Sejmu?<br>……?<br>2.15. mediów (dziennikarzy)?<br>Na podstawie 15 itemów można utworzyć wskaźnik syntetyczny ogólnej oceny zaufania (jako sumę punktów uzyskanych dla poszczególnych 15 itemów). | 1. Tak, duże.<br>2. Tak, umiarkowane.<br>3. Nie. | *Diagnoza społeczna* | Ilościowa |

**Źródło:** opracowanie własne.

Przeanalizujmy inny przykład. Gdy pytamy o wiek (w latach), możemy poprosić o:

- podanie konkretnej liczby lat życia (uzyskujemy zmienną mierzoną na skali ilościowej, a konkretnie ilorazowej); na przykład porównując wiek osoby 20- i 60-letniej, jesteśmy wówczasw stanie wskazać, że jedna osoba jest od drugiej trzykrotnie starsza (wynik dzielenia ma interpretację merytoryczną);
- podanie roku urodzenia (również uzyskujemy zmienną mierzoną na skali ilościowej, ale tym razem przedziałowej); na przykład porównując wiek osób urodzonych w 2010 i 1990 roku, nie możemy wprost wskazać, ile razy jedna osoba jest starsza od drugiej – wynik dzielenia nie ma interpretacji merytorycznej, ale ma taką interpretację różnica tych dwóch wartości (jesteśmy w stanie ustalić, o ile jedna osoba jest starsza od drugiej);
- przypisanie się do konkretnego przedziału wieku (do 20 lat, 21–30, 31 lub więcej) – mierzymy wówczas wiek na skali porządkowej; ponownie, porównując wiek osoby, która zaznaczyła wariant „21–30" i osoby w wieku „do 20 lat", nie wiemy ile razy, ani o ile jedna osoba jest starsza od drugiej, wiemy jednak, że jest starsza – nie mamy więc już do czynienia ze skalą nominalną, ale też nie mamy jeszcze do czynienia ze skalą ilościową.

Rozróżnienie skal pomiarowych może przebiegać zatem w następujący sposób:

- porównując wartości zmiennej wyrażonej na skali nominalnej (np. płci), jesteśmy w stanie wskazać jedynie, czy dwie osoby mają ten sam wariant zmiennej, czy inny;
- jeśli dodatkowo możemy wskazać, która osoba ma wyższy wariant zmiennej (ale nie jesteśmy w stanie określić, o ile wyższy), mamy do czynienia

ze zmienną mierzoną na skali porządkowej (tak jest np. z poziomem wykształcenia czy cechą mierzoną na skali Likerta);
- jeśli możemy wskazać dodatkowo, o ile dany wariant jest wyższy czy niższy (odległości są ustalone), mamy do czynienia ze skalą ilościową.

Tak więc im wyższa skala pomiarowa, tym większa dokładność pomiaru, co z kolei umożliwia zastosowanie innych metod statystycznych (tabela 2.2).

**Tabela 2.2.** Skala pomiarowa a metody analizy statystycznej

| Skala pomiarowa | Analiza jednowymiarowa (najważniejsze statystyki) | Analiza dwuwymiarowa (przykłady) |
|---|---|---|
| Nominalna | Wskaźnik struktury (*w*), dominanta (*Do*) | Test niezależności chi-kwadrat, współczynnik V-Craméra |
| Porządkowa | Jw. + kwantyle, w tym zwłaszcza mediana (*Me*), a także kwartyle (*Q1*, *Q3*), decyle (*D1*, *D2*, …), percentyle (*P1*, *P2*, …) | Jw. + test Manna-Whitneya, test Kruskala-Wallisa, współczynnik rho Spearmana, współczynnik tau-Kendalla |
| Ilościowa | Jw. + średnia arytmetyczna ( $\overline{X}$ lub *M*), odchylenie standardowe (*S* lub *STD*), wariancja ($S^2$), współczynnik zmienności ($V_S$), współczynnik skośności ($W_S$), kurtoza (*K*) | Jw.+ test t-Studenta, ANOVA, współczynnik korelacji liniowej Pearsona (*r*) |

**Źródło:** opracowanie własne.

Ogólnie w przypadku skali wyższego rzędu dopuszczalne są metody możliwe do zastosowania przy skali niższej, aczkolwiek nie zawsze jest to wskazane. Przykładowo: choć jest możliwe zastosowanie testu niezależności chi-kwadrat przy skali ilościowej, wymaga to obniżenia poziomu pomiaru poprzez pogrupowanie wariantów zmiennej w przedziały klasowe, a więc tym samym mniejszej dokładności pomiaru. Podobnie możliwe jest zastosowanie współczynnika rho, ale tu z kolei zamiast wartościami zmiennej operuje się ich rangami, a więc wartościami liczbowymi przypisanymi pozycjom zajmowanym przez uporządkowane warianty zmiennej (tabela 2.3).

**Tabela 2.3.** Sposób przekształcenia wartości zmiennej w rangi

| Nr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *y* | 1 | 1 | 2 | 3 | 3 | 3 | 6 | 8 | 9 | 30 |
| Ranga | 1,5 | 1,5 | 3 | 5 | 5 | 5 | 7 | 8 | 9 | 10 |

**Źródło:** opracowanie własne.

Przechodząc na metody oparte na rangach, tracimy zatem część informacji, ignorujemy bowiem to, jak bardzo różnią się wartości zmiennej dla poszczególnych

jednostek badania, interesuje nas tylko ich hierarchia. Zauważmy, że niektórym z nich przypisana została taka sama ranga, stanowiąca średnią z zajmowanych przez nie pozycji (skoro wartości zmiennej są takie same, nie możemy nadać im innych rang) – mamy wówczas do czynienia z rangami wiązanymi. Różnicę między 8. i 9. jednostką traktujemy przy takim podejściu na równi z różnicą między 9. i 10. (choć w pierwszym przypadku wynosi ona 1, a w drugim aż 21). Przechodząc na test niezależności chi-kwadrat, tracimy jeszcze więcej informacji, bo do jednej grupy zaliczamy na przykład wszystkie jednostki o wartości zmiennej przynajmniej 5 (więc na równi z wartością 30 traktujemy 6, 8 i 9). Skoro tak, to – o ile to możliwe – najlepiej utrzymać jak największą dokładność pomiaru i stosować metodę dedykowaną danej skali pomiarowej. Nie jest jednak wykluczone zastosowanie metody zalecanej dla skali niższej. W niektórych sytuacjach będzie to nawet bardziej wskazane – tak jest również w prezentowanym w tabeli 2.3 przypadku. Wartość 30 wyraźnie zawyży średnią dla ogółu badanych, a tym samym lepiej będzie posłużyć się rangami, licząc się z obniżeniem dokładności pomiaru. Dlatego prowadząc analizę dla zmiennych mierzonych na skali ilościowej, sprawdza się normalność rozkładu zmiennej, a co za tym idzie – sprawdza się, czy średnia arytmetyczna poprawnie odzwierciedla przeciętny poziom zmiennej (nie zawyża lub nie zaniża oczekiwanego poziomu zmiennej).

## 2.4. Metody wnioskowania statystycznego – aspekty praktyczne

Jak podkreślano, w niniejszej publikacji główna uwaga skupiona jest na analizie wyników badań pierwotnych, zwłaszcza badań kwestionariuszowych, prowadzonych w sposób pozwalający na uogólnienie wyników z próby na populację generalną. W analizie wyników wykorzystuje się wówczas **metody wnioskowania statystycznego**, które – dzięki zapewnieniu obiektywnych zasad doboru próby – pozwalają na ustalenie prawdopodobieństwa błędu, z jakim uogólnienia te są dokonywane. Ponieważ analiza prowadzona jest na podstawie wycinka populacji generalnej, a więc w warunkach niepewności, posługujemy się pojęciem **zmiennej losowej** (w skrócie zmienną), czyli każdą mierzalną funkcją określoną na przestrzeni zdarzeń elementarnych $\Omega$ o wartościach w zbiorze liczb rzeczywistych (Szymczak, 2010, s. 41). Przez zmienną losową można zatem intuicyjnie rozumieć taką zmienną, która w wyniku doświadczenia może przyjąć wartość z pewnego zbioru liczb rzeczywistych i to z pewnym z góry określonym prawdopodobieństwem (Sobczyk, 1998, s. 83). Sposób przypisywania prawdopodobieństw poszczególnym wartościom dyskretnej zmiennej losowej oraz sposób

przypisywania prawdopodobieństw odcinkom na prostej w przypadku ciągłej zmiennej losowej określa się jako rozkład prawdopodobieństwa. Różne zmienne losowe będą generowały różne rozkłady prawdopodobieństwa. Z kolei dystrybuantą zmiennej losowej $X$ nazywamy funkcję $F(x)$ określoną wzorem:

$$F(x) = P(\omega \in \Omega : X(\varpi) < x).$$

Metody wnioskowania statystycznego obejmują dwa działy: estymację i weryfikację hipotez statystycznych (Malarska, 2005, s. 103).

**Estymacja**, czyli szacowanie wartości parametrów lub postaci rozkładu zmiennej losowej w populacji generalnej na podstawie rozkładu empirycznego uzyskanego z próby, pozwala na ustalenie przybliżonych wartości parametrów w populacji generalnej w sytuacji, gdy dysponujemy jedynie statystykami z próby. Pewna charakterystyka zbioru wartości, jakie może przybierać zmienna losowa, nazywa się parametrem tej zmiennej lub parametrem rozkładu zmiennej (Starzyńska, 2020). Największe znaczenie praktyczne mają dwie grupy parametrów: **wartość oczekiwana** (nadzieja matematyczna), oznaczana przez $E(X)$ lub $\mu$, reprezentująca średnią wielkość zmiennej losowej w populacji generalnej, oraz **wariancja**, oznaczana przez $D^2(X)$ lub $\sigma^2$, informująca o rozrzucie wartości zmiennej losowej. Pierwiastek kwadratowy z wariancji nosi nazwę odchylenia standardowego zmiennej losowej – $D(X)$ lub σ. Ważnym parametrem jest również frakcja (odsetek) w populacji – $p$.

Prowadząc badania na podstawie próby, nie znamy i nie jesteśmy w stanie wyznaczyć wartości parametrów w populacji. Możemy jedynie znaleźć wartości ich estymatorów. **Estymator** to wielkość (statystyka, charakterystyka) wyznaczona na podstawie próby losowej, służąca do oceny wartości nieznanych parametrów populacji generalnej. Aby funkcję zmiennych losowych można było uznać za estymator, musi mieć pewne własności. Dobry estymator powinien być (Hellwig, 1998: 194–197):

- zgodny – wraz ze wzrostem liczebności próby, na podstawie której obliczamy wartość estymatora, zmniejsza się liczba punktów, w których wartość estymatora różni się od prawdziwej wartości parametru;
- nieobciążony – estymator jest nieobciążony, jeśli dla każdej liczebności próby wartość oczekiwana estymatora jest równa wartości estymowanego parametru (przy wielokrotnym używaniu estymatora wielkości przeszacowań i niedoszacowań wielkości parametru będą się bilansowały i „średnio" oszacowanie będzie poprawne);
- najbardziej efektywny – estymator o najmniejszej wariancji.

Przykładowo: estymatorem wartości oczekiwanej jest średnia arytmetyczna, estymatorem odchylenia standardowego – odchylenie standardowe z próby.

Wartość estymatora przyjmujemy za oszacowanie nieznanego parametru (nie można zapominać, że nie należy tej decyzji traktować jako obiektywnej prawdy). Estymacji parametru dokonuje się na podstawie wyników z próby. Na tej podstawie, przy określonym poziomie ufności, oszacować można również błąd estymacji, określany też jako błąd szacunku lub błąd standardowy (oznaczany zwykle przez *SE*).

**Weryfikacja (testowanie) hipotez** oznacza sprawdzanie określonych przypuszczeń (założeń) wysuniętych w stosunku do parametrów (lub rozkładów) populacji generalnej na podstawie wyników z próby. Narzędziem wykorzystywanym do tego celu jest test statystyczny, tj. reguła postępowania, która na podstawie wyników z próby umożliwia **podjęcie decyzji** o prawdziwości bądź fałszywości sformułowanych hipotez statystycznych. Punktem wyjścia na etapie wyboru odpowiedniego testu statystycznego są nasze przypuszczenia dotyczące badanych zjawisk, jakie formułujemy na gruncie teoretycznym – na tej podstawie stawiamy określone hipotezy badawcze lub pytania badawcze. Przykładowo: jeśli badamy uwarunkowania aktywności zawodowej kobiet, możemy postawić następującą hipotezę badawczą: „Aktywność zawodowa jest wyższa w przypadku kobiet mieszkających w mieście niż na wsi". Przypuszczenie to dotyczy zatem pewnych prawidłowości, jakie występują w populacji generalnej. Tę hipotezę badawczą będziemy weryfikować na podstawie danych zgromadzonych dla próby stanowiącej reprezentatywny wycinek tej populacji. Stawianą hipotezę badawczą należy zatem sformalizować tak, aby można ją było zweryfikować, używając konkretnych procedur matematycznych, na których opierają się metody statystyczne. Dlatego też formułujemy hipotezy statystyczne – **hipotezę zerową ($H_0$)** i **hipotezę alternatywną ($H_1$)**. W hipotezie zerowej przyrównujemy do siebie: parametr w dwóch lub więcej niż dwóch populacjach, parametr do jakiejś wartości, rozkłady zmiennych itp. Hipoteza alternatywna jest przeciwna do hipotezy zerowej, zapisujemy więc w niej to, że parametr różni się w dwóch lub więcej niż dwóch populacjach, lub że parametr różni się od jakiejś wartości, że rozkłady są różne. A zatem jeśli w $H_0$ postawimy między parametrami czy rozkładami znak równości, to w hipotezie alternatywnej stanie znak „≠" (wykorzystywany test statystyczny określać będziemy jako dwustronny lub obustronny). Przekładając to na postawioną wcześniej hipotezę badawczą, mamy zatem:

$H_0$: Aktywność zawodowa kobiet mieszkających w miastach jest taka sama jak aktywność zawodowa mieszkanek wsi

$H_1$: Aktywność zawodowa kobiet mieszkających w miastach jest różna od aktywności zawodowej mieszkanek wsi.

Jak widać, hipoteza alternatywna nie oddaje wprost tego, co przyjęto w hipotezie badawczej, niemniej jednak na etapie dalszych procedur uda nam się sprawdzić

nie tylko to, czy aktywność zawodowa w tych dwóch populacjach się różni, ale również jak się różni[4]. Do zagadnień tych wrócimy w kolejnych rozdziałach.

Weryfikując hipotezy statystyczne w warunkach niepewności, możemy popełnić jeden z dwóch błędów – odrzucić prawdziwą hipotezę zerową (błąd pierwszego rodzaju) bądź przyjąć fałszywą hipotezę zerową (błąd drugiego rodzaju). Prawdopodobieństwo błędu pierwszego rodzaju określa się jako **poziom istotności** ($\alpha$), a prawdopodobieństwo błędu drugiego rodzaju to $\beta$ (powiązane z mocą testu $1 - \beta$) (Hellwig, 1998, s. 259). W praktyce najczęściej stosuje się **testy istotności**, w których kontrolowane jest prawdopodobieństwo błędu pierwszego rodzaju. Na podstawie wyników z próby wyznacza się wartość sprawdzianu testu (statystyki testu) według określonej dla danego testu formuły, a następnie przy właściwej liczbie stopni swobody (*df*) odczytuje się wartość krytyczną (z tablic odpowiedniego rozkładu) i porównuje ją z wartością sprawdzianu testu (statystyką testu). Korzystając z pakietów statystycznych, na poziomie danej procedury, przy określonej liczbie stopni swobody i wartości sprawdzianu testu, wyznacza się **prawdopodobieństwo w danym teście na podstawie wynik**ów z próby (oznacza się je przez *p*; w programach statystycznych oznaczone jest jako *Istotność* lub *p*-value) i porównuje się je z przyjętym poziomem istotności $\alpha$. Sprawdzamy więc, czy błąd odrzucenia prawdziwej hipotezy zerowej jest mniejszy od przyjętego progu ($\alpha$, który zwykle przyjmuje się jako 0,05). Jeśli tak, to możemy uznać hipotezę alternatywną za prawdziwą. Hipotezy zerowej nie możemy przyjąć w żadnej sytuacji, gdyż stosowany test istotności nie kontroluje błędu drugiego rodzaju (przypomnijmy – błędu przyjęcia fałszywej hipotezy zerowej). Tak więc:

- jeśli $p < \alpha$, $H_0$ odrzucamy, za prawdziwą uznajemy $H_1$ (uznajemy np. różnice między populacjami lub zależność za istotne statystycznie; stwierdzone na poziomie próby różnice można uogólnić na populację generalną, tzn. można przyjąć, że mają one miejsce w populacji, nie są jedynie efektem błędu losowego);

4 Oprócz testów dwustronnych wykorzystać można również testy jednostronne, tj. prawo- lub lewostronne. Rodzaj testu wiąże się z zapisem hipotezy alternatywnej. Jeśli w hipotezie alternatywnej postawimy znak „<", to test będzie lewostronny (np. wiek użytkowników smartfonów jest niższy niż wiek użytkowników telefonów tradycyjnych). Z kolei jeśli w hipotezie alternatywnej użyty będzie znak „>", test będzie prawostronny (np. wyższą cenę mają komputery marki X niż Y lub korelacja między długością serwisu pogwarancyjnego a ceną notebooka jest dodatnia, a więc droższe są notebooki o dłuższym okresie serwisu pogwarancyjnego). Jak podkreślano, w praktyce z wykorzystaniem oprogramowania statystycznego, w tym IBM SPSS Statistics, ograniczamy się do zweryfikowania hipotez w teście dwustronnym, a następnie – odwołując się do określonych statystyk wyznaczonych na poziomie próby – uszczegóławiamy nasze wnioski tak, aby odnieść się bardziej konkretnie do stawianej hipotezy badawczej.

- jeśli $p > \alpha$, nie mamy podstaw do odrzucenia $H_0$ (uznajemy, że różnice między populacjami czy zależność nie są istotne statystycznie; oznacza to np., że populacje można uznać za podobne – ale nie takie same).

Zasada ta ma zastosowanie w każdym z testów istotności. Dlatego tak ważne jest, aby świadomie dokonywać ich weryfikacji, a to wymaga świadomości stawianych hipotez statystycznych. Powszechnym błędem jest utożsamianie hipotezy badawczej z hipotezą zerową (odrzucamy $H_0$, więc uznajemy, że hipoteza badawcza nie jest prawdziwa). Na przykład: w hipotezie badawczej zakładamy, że zadowolenie z produktu jest powiązane z płcią jego nabywców. Tymczasem hipoteza zerowa w takim badaniu przewiduje brak związku między zmiennymi. Odrzucenie hipotezy zerowej nie podważa zatem stawianej hipotezy badawczej, wręcz przeciwnie – dostarcza argumentów za jej prawdziwością.

Należy również zwrócić uwagę na podział testów na **parametryczne** (służące do weryfikacji hipotez dotyczących parametrów rozkładu zmiennej losowej, wymagających określonej postaci rozkładu zmiennej) i **nieparametryczne** (służące do weryfikacji hipotez dotyczących rozkładów zmiennych losowych). Przykładowo: weryfikując hipotezę badawczą postaci: „Wyższe nakłady na inwestycje ponoszą duże podmioty niż MŚP" (mikro-, małe i średnie przedsiębiorstwa), hipotezy statystyczne można zapisać ogólnie:
$H_0$: Nakłady na inwestycje dużych pomiotów i MŚP są sobie równe
$H_1$: Nakłady na inwestycje dużych pomiotów i MŚP się różnią.

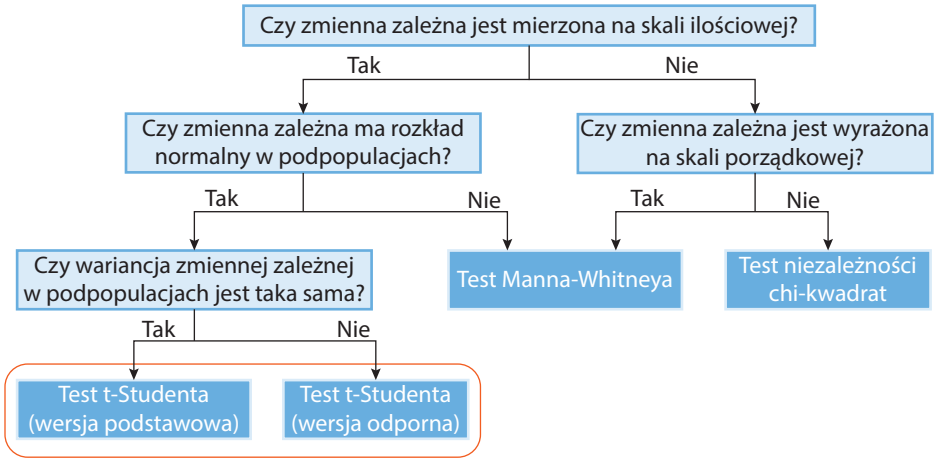
Z uwagi na ilościowy poziom pomiaru zmiennej zależnej (nakłady na inwestycje) można je weryfikować testem parametrycznym lub nieparametrycznym. W pierwszym przypadku porównuje się parametr rozkładu wyrażający przeciętny poziom tej zmiennej w populacjach (wartość oczekiwaną – $\mu$):
$H_0$: $\mu_1 = \mu_2$
$H_1$: $\mu_1 \neq \mu_2$.

W drugim przypadku porównuje się dystrybuanty rozkładu zmiennej ($F$) w populacjach:
$H_0$: $F_1 = F_2$
$H_1$: $F_1 \neq F_2$.

W przypadku zmiennych wyrażonych na skalach niemetrycznych można zastosować jedynie testy nieparametryczne.

Wybierając test statystyczny, bierzemy pod uwagę warunki wymienione wcześniej (hipotezę/pytanie badawcze i skalę pomiarową zjawiska, schemat badania – zależny/niezależny), a także liczbę porównywanych populacji oraz warunki, jakich spełnienia wymaga dana metoda. Najważniejsze rozróżnienie stanowi w tym przypadku poziom pomiaru zmiennej zależnej oraz liczba wariantów branego pod uwagę czynnika, która wyznacza liczbę porównywanych podpopulacji. Inne testy zastosujemy, porównując

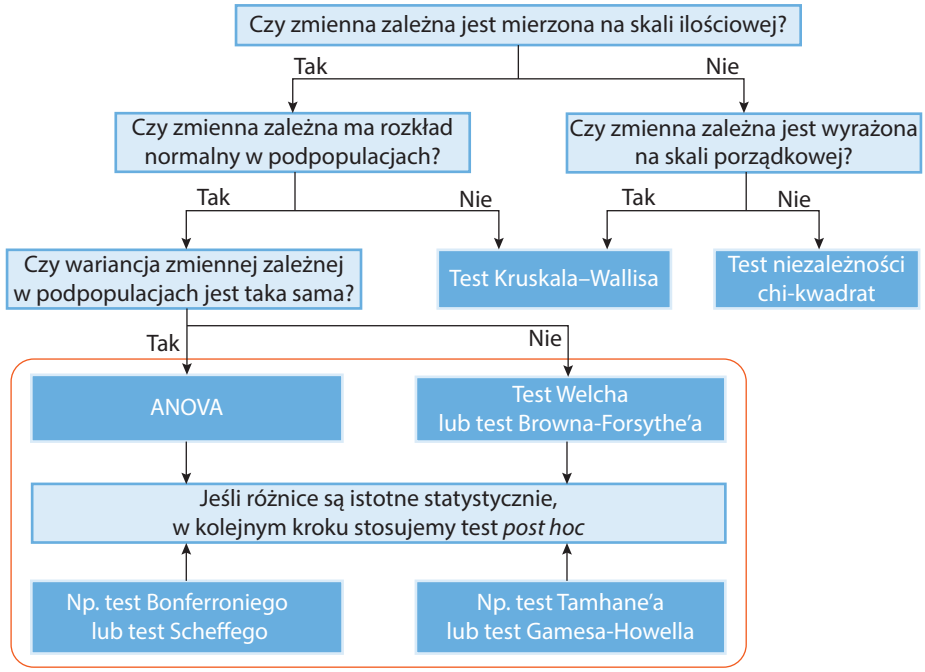
dwie populacje, a inne przy więcej niż dwóch populacjach (aczkolwiek testy z drugiej grupy można stosować również przy porównaniu dwóch populacji – są przewidziane dla „przynajmniej dwóch populacji"). Zastosowanie testów parametrycznych wchodzi w grę wtedy, gdy zmienna zależna jest mierzona na skali ilościowej, a jej rozkład jest normalny (a przynajmniej nie są obserwowane istotne odstępstwa w tym zakresie). Kryteria wyboru testów do porównania poziomu zmiennej w dwóch lub więcej niż dwóch populacjach zostały zobrazowane na rysunkach 2.1 i 2.2. Wymienione tu testy statystyczne zostaną szczegółowo omówione w kolejnych rozdziałach. Przyjęto przy tym zasadę, że testy stosowane dla skali ilościowej oraz te wykorzystywane dla skali porządkowej (w szczególnych przypadkach również zalecane dla zmiennych mierzonych na skali ilościowej) omówione zostały w tych samych rozdziałach – w rozdziale czwartym w odniesieniu do porównania dwóch populacji i w rozdziale piątym w odniesieniu do większej liczby populacji. Z kolei porównania odnoszące się do zmiennych mierzonych na skali nominalnej zostały zaprezentowane odrębnie, w rozdziale szóstym.




**Rysunek 2.1.** Schemat wyboru testu przy porównaniu dwóch populacji – pomiar niezależny

**Źródło:** opracowanie własne.

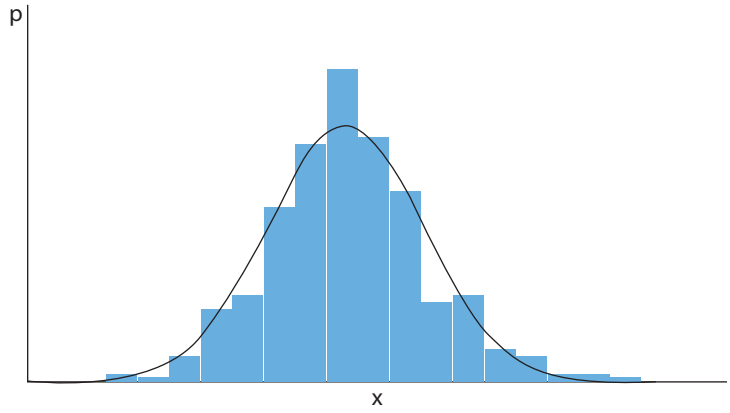
Testy, których nazwy zostały na rysunkach 2.1 i 2.2 obwiedzione pomarańczową linią, należą do testów parametrycznych. Zauważmy, że są one dość odporne na odstępstwa od normalności rozkładu, niemniej jednak gdy te odstępstwa są poważne, zastosowanie tych testów może prowadzić do błędnego osądu badanych zjawisk. Dlatego wstępnym etapem, poprzedzającym zastosowanie metod parametrycznych, powinna być analiza rozkładu zmiennej zależnej (w obrębie wyróżnionych podpopulacji) w celu oceny zgodności jej rozkładu z rozkładem normalnym.




**Rysunek 2.2.** Schemat wyboru testu przy porównaniu przynajmniej dwóch populacji – pomiar niezależny

**Źródło:** opracowanie własne.

**Rozkład normalny**, nazwany również rozkładem Gaussa, opisuje sposób, w jaki wartości zmiennych występują w populacji. Mówimy, że zmienna losowa $X$ ma rozkład normalny o parametrach $\mu$ oraz $\sigma$, co w skrócie zapisuje się jako: $X \sim N(\mu, \sigma)$ lub (rzadziej) $X \sim N(\mu, \sigma^2)$. Krzywa normalna jest krzywą w kształcie dzwonu, symetryczną względem prostej $x = \mu$. Ma ona jedno maksimum w punkcie, który jednocześnie odpowiada wartości oczekiwanej, dominancie i medianie rozkładu. Krzywa normalna ma dwa punkty przegięcia dla $x = \mu - \sigma$ oraz $x = \mu + \sigma$, a lewe i prawe ramię krzywej zbliżają się asymptotycznie do osi odciętych, przy czym dla $x < \mu - 3\sigma$ oraz dla $x > \mu + 3\sigma$ rzędne niewiele różnią się od zera. Obrazuje ją rysunek 2.3, na którym dodatkowo uwzględniono histogram (wyrażają go słupki na rys. 2.3).




**Rysunek 2.3.** Krzywa normalna

**Źródło:** opracowanie własne.

Punktem wyjścia jest tu leżące u podstaw statystyki przekonanie, że wszystkie zmienne ilościowe w populacji będą miały rozkład normalny – najwięcej będzie przypadków średnich, a znacznie mniej przypadków o wartościach bardzo niskich i bardzo wysokich (Bedyńska, Brzezicka, 2007, s. 103–104). Ze względu na nieskończoną liczbę kombinacji par parametrów $\mu$ i $\sigma$ istnieje nieskończona liczba krzywych normalnych. Konkretne wartości parametrów (wartości oczekiwanej i odchylenia standardowego) określają położenie krzywej normalnej w układzie współrzędnych. W praktyce duże znaczenie odgrywa szczególny przypadek rozkładu normalnego – rozkład normalny standaryzowany: $X \sim N(1, 0)$.

Jak podkreślano, **stosowanie testów parametrycznych wymaga, aby analizowane zmienne miały rozkład normalny**. Aby sprawdzić, czy ma to miejsce, wykorzystuje się zwykle **test Shapiro-Wilka**. Miarami wykorzystywanymi na etapie oceny normalności rozkładu są też współczynnik skośności i kurtoza. Warto również na poziomie próby (czyli analizując rozkład empiryczny zmiennej) porównać wartości średniej arytmetycznej, mediany, dominanty, średniej obciętej, M-estymatorów, przeanalizować histogram, wykres skrzynkowy, wykres normalny K-K. Ma to szczególne znaczenie przy dużych próbach. Zagadnienia te omówione zostaną w kolejnych rozdziałach.

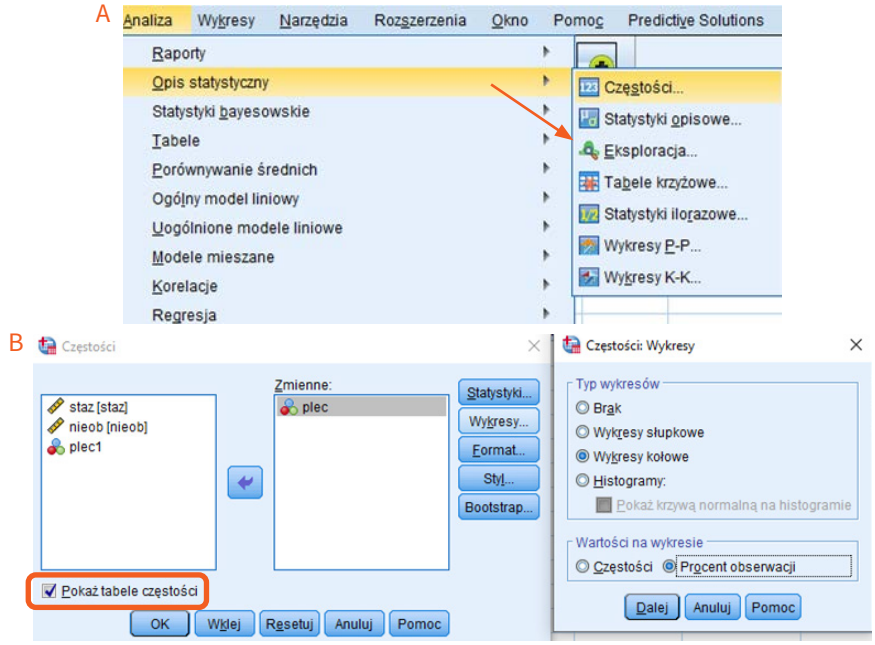
# 3. Statystyki opisowe w analizie rozkładu empirycznego zmiennej

**Kluczowe pojęcia:** rozkład częstości zmiennej, kategoria zmiennej, charakterystyki opisowe rozkładu, miary tendencji centralnej, miary położenia, średnia, dominanta, mediana, średnia obcięta, kwartyle, M-estymatory, miary rozproszenia, odchylenie standardowe, wariancja, skośność rozkładu, spłaszczenie, kurtoza

## 3.1. Rozkład częstości zmiennej

Przystępując do analizy statystycznej, należy w pierwszej kolejności zapoznać się z częstością występowania poszczególnych wariantów badanej zmiennej/cechy, czyli z rozkładem częstości tej zmiennej. Rozkład częstości to ujęcie danych w kategorie i wskazanie liczby obserwacji w obrębie każdej kategorii w badanej zbiorowości (Górniak, Wachnicki, 2000, s. 105). Znajomość rozkładu empirycznego zmiennej to znajomość częstości występowania – bezwzględnych ($n_i$) i/lub względnych ($w_i$) – poszczególnych kategorii zmiennej ($x_i$). Aby wyznaczyć częstości (niezależnie od skali pomiarowej zmiennej), korzystamy z polecenia *Analiza* → *Opis statystyczny* → *Częstości* (rysunek 3.1A).

Pojawia się okno dialogowe podzielone na dwie części (rysunek 3.1B). Po lewej stronie znajduje się wykaz wszystkich zmiennych, jakie mamy w naszym zbiorze danych. W pole *Zmienne* przenosimy zmienną (lub zmienne), dla których chcemy utworzyć *Tabelę częstości* – robimy to poprzez dwukrotne kliknięcie w nią, przeciągnięcie jej na prawą stronę albo po ustawieniu się na tej zmiennej wykorzystujemy strzałkę znajdującą się między oknami. Aby utworzyć taką tabelę, należy zaznaczyć *Pokaż tabele częstości* (opcja ta jest oznaczana domyślnie).




**Rysunek 3.1.** Wykonywanie polecenia *Częstości*

Dodatkowo można przedstawić rozkład zmiennej na wykresie – należy wybrać *Wykresy*, a następnie dobrać odpowiedni wykres (do wyboru mamy wykres kołowy, słupkowy i histogram, na którym można – jak na rysunku 2.3 – dodać krzywą normalną). Jeśli mamy do czynienia ze zmienną o niewielkiej liczbie wariantów (w tym zwłaszcza mierzoną na skali nominalnej lub porządkowej) i jednocześnie uwzględnione kategorie obejmują 100% jednostek należących do badanej zbiorowości (populacji lub próby), wygodny będzie wykres kołowy. Można na nim oznaczyć albo liczebności (opcja *Częstości*), albo procenty (opcja *Procent obserwacji*). Jeśli nie uwzględniamy 100% zbiorowości albo kategorii zmiennej jest dużo, lepszy będzie wykres słupkowy. Dla zmiennych mierzonych na skali ilościowej polecany jest histogram (warto wybrać dodatkowo *Pokaż krzywą normalną na histogramie* – rysunek 3.1).

## 3.2. Statystyki opisowe rozkładu zmiennej

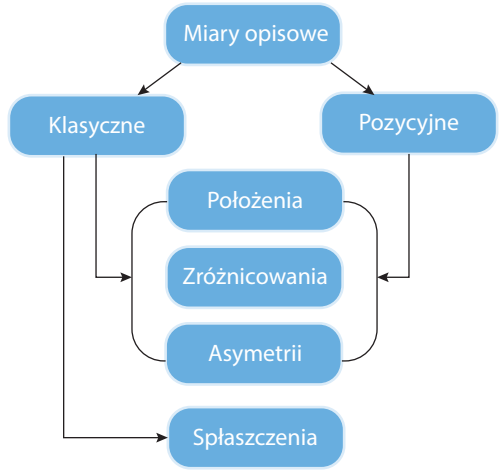
Należy pokreślić, że samo przedstawienie rozkładu częstości zmiennej nie wystarcza. Warto pójść dalej, aby uchwycić zasadnicze właściwości rozkładu i scharakteryzować go syntetycznie za pomocą niewielu liczb. Liczby, które sumarycznie i skrótowo opisują rozkład, nazywa się charakterystykami opisowymi rozkładu zmiennej (Lange, 1952, s. 46). Poznanie liczbowych charakterystyk rozkładu jest bardzo ważne, gdyż umożliwia porównania ilościowe rozkładów dwóch lub więcej

zmiennych. Jak podkreślano w rozdziale drugim, liczbowe charakterystyki rozkładu zmiennej losowej w populacji nazywa się parametrami, natomiast charakterystyki rozkładu wyznaczone na podstawie próby losowej – **statystykami opisowymi** (są one estymatorami konkretnych parametrów). Statystyk opisowych używa się również wtedy, gdy prowadzimy badanie całkowite, dotyczące populacji generalnej i obejmujące wszystkie jej jednostki.

Podstawowymi zadaniami statystyk opisowych są (Zając, 1994, s. 131–134):

- określenie przeciętnego poziomu (tendencji centralnej, położenia) wartości zmiennej;
- ocena zróżnicowania (rozproszenia) wartości zmiennej;
- określenie siły i kierunku asymetrii (skośności);
- ocena koncentracji (spłaszczenia rozkładu).

Liczbowymi charakterystykami syntetycznego opisu rozkładu cechy są cztery grupy mierników. Klasyfikację tych miar zaprezentowano na rysunku 3.2. Pierwszy podział na miary klasyczne i pozycyjne wynika z istoty problemu, jakiego dotyczą. Miary klasyczne są wyznaczane na podstawie wszystkich wartości zmiennej ($x_i$), natomiast miary pozycyjne na podstawie tylko niektórych obserwacji.




**Rysunek 3.2.** Klasyfikacja miar opisowych rozkładu zmiennej

**Źródło:** opracowanie własne.

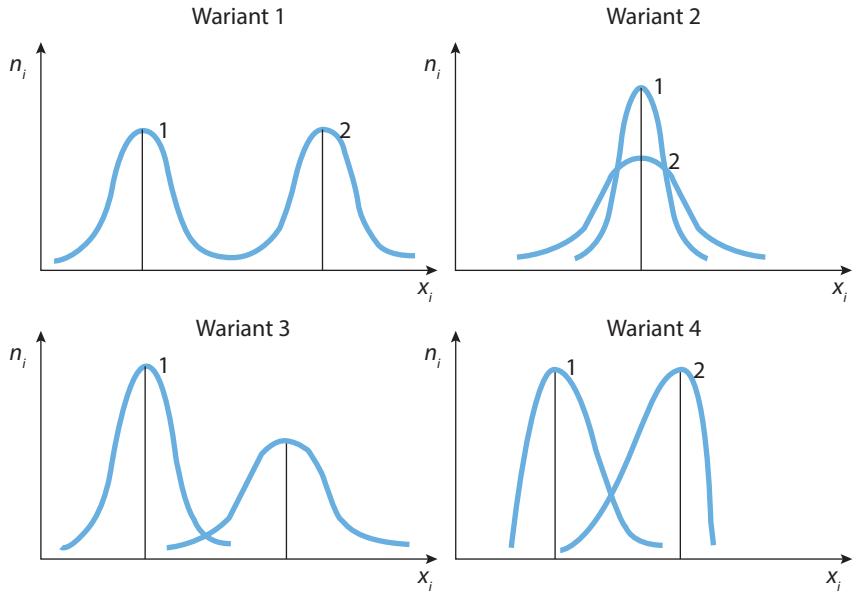
**W niniejszym podręczniku, ze względu na ograniczenia jego objętości, przedstawimy sposoby wyznaczania miar (położenia, rozproszenia, skośności i koncentracji), w tym wzory, jedynie dla danych indywidualnych (szeregów szczegółowych). Z tego samego powodu z grupy miar średnich klasycznych omówimy tylko średnią arytmetyczną.**

Porównując rozkład tej samej cechy (zmiennej) w różnych zbiorowościach, można stwierdzić, że różnice między nimi sprowadzają się do czterech charakterystycznych właściwości rozkładu (Zając, 1994, s. 133–134):

rozkłady mogą różnić się położeniem, tzn. wartością zmiennej, w której pobliżu skupiają się obserwacje (rysunek 3.3, wariant 1),

obserwacje mogą skupiać się wokół tej samej wartości, ale różnić się rozproszeniem oraz spłaszczeniem (rysunek 3.3, wariant 2),

rozkłady mogą wreszcie różnić się położeniem, rozproszeniem, spłaszczeniem oraz skośnością (rysunek 3.3, warianty 3 i 4).




**Rysunek 3.3.** Przykładowe rozkłady zmiennej różniące się położeniem i/lub rozproszeniem

**Źródło:** opracowanie własne.

Każdą z powyższych właściwości można rozważać oddzielnie.

## 3.2.1. Miary położenia

Nazwa miar położenia wynika z ich lokalizacji (miejsca położenia) na osi odciętych układu współrzędnych przedstawiającego rozkład zmiennej. Można wskazać miejsce, w którym leży wartość najlepiej reprezentująca wszystkie warianty zmiennej. Miary położenia dzielą się na: **przeciętne (tendencji centralnej)** i **kwantyle**.

**Miary przeciętne** informują o średnim lub **typowym** poziomie wartości cechy/zmiennej. Są wartościami, wokół których skupiają się pozostałe wartości

analizowanej zmiennej. Miary przeciętne, podobnie jak wszystkie pozostałe miary opisujące rozkład, dzielą się na dwie grupy: **klasyczne i pozycyjne**. Średnie klasyczne są wyznaczane na podstawie wszystkich wartości zmiennej badanych jednostek zbiorowości, podczas gdy miary pozycyjne wskazują określoną pozycję jednostek (np. środkową lub dominującą). Miary przeciętne są wielkościami mianowanymi, wyrażone są w jednostkach miary badanej zmiennej.

## Średnia arytmetyczna

Najpopularniejszą średnią klasyczną jest średnia arytmetyczna (najczęściej określana po prostu jako średnia). Średnia arytmetyczna ($\overline{x}$, $M$) jest sumą wartości zmiennej **mierzalnej**, podzieloną przez liczbę jednostek tej zbiorowości (przez jej liczebność):

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{N} x_i}{n}, \tag{1}$$

gdzie: $n$ – liczebność zbiorowości, $x_i$ – wartość zmiennej mierzalnej.

Tak obliczona średnia arytmetyczna nazywa się średnią prostą (nieważoną).

Średnia ma kilka własności – poniżej zaprezentowano te najpopularniejsze:

1. Średnia arytmetyczna jest wielkością mianowaną, czyli wyrażoną w tej samej jednostce miary co badana cecha.
2. Jako miara klasyczna jest wypadkową wszystkich wartości zmiennej, jest wielkością abstrakcyjną, ale zawsze spełnia warunek:

$$x_{\min} < \overline{x} < x_{\max}.$$

3. Suma odchyleń poszczególnych wartości cechy od średniej arytmetycznej równa się zero:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

4. Suma kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej jest minimalna:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \min.$$

5. Suma wartości cechy jest równa iloczynowi średniej arytmetycznej i liczebności zbiorowości:

$$n\overline{x} = \sum_{i=1}^{n} x_i \ .$$

Średniej arytmetycznej nie należy stosować do oceny przeciętnego poziomu zmiennej, gdy:

- w zbiorowości występują jednostki nietypowe;
- występuje silna skośność rozkładu;
- wartość średniej arytmetycznej znacznie różni się od wartości mediany, 5% średniej obciętej, M-estymatorów;
- rozkład jest silnie niejednorodny;
- rozkład jest wielomodalny.

## Dominanta

Dominanta (modalna, *D*) jest to kategoria/wartość zmiennej, która występuje najczęściej w badanej zbiorowości. Jest jedyną miarą tendencji centralnej w przypadku zmiennej nominalnej. I tak na przykład nie można obliczyć średniego zainteresowania w badanej grupie studentów, ale można wskazać, która kategoria zainteresowania występuje z największą częstością.

Niewątpliwą zaletą dominanty jest łatwość jej wskazania i interpretacji. Trzeba jednak pamiętać, że dominanta nie zawsze najlepiej wskazuje „typową" wartość dla danej zmiennej. Wartość informacyjna dominanty jest niewielka w sytuacji, gdy najczęściej występująca kategoria zmiennej nie występuje dużo częściej od innych kategorii. Może się również zdarzyć, że rozkład nie ma jednej, wyraźnej dominanty oraz może być rozkładem wielomodalnym.

## Mediana i inne percentyle

Mediana (*Me*) to wartość (kategoria) zmiennej środkowej jednostki zbiorowości w uporządkowanym rosnąco szeregu. Mediana, jako wartość środkowa lub kwartyl drugi, dzieli zbiorowość na dwie równe części w ten sposób, że połowa jednostek zbiorowości ma wartości niższe lub równe medianie, a połowa ma wartości równe lub większe od niej.

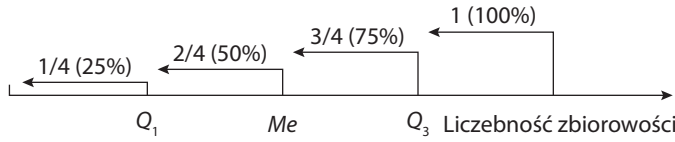
W przypadku parzystej liczby obserwacji mediana jest średnią dwu środkowych obserwacji.

Każdą zbiorowość można podzielić na części. **Kwantyle** to wartości cechy, które dzielą badaną zbiorowość na określone części pod względem liczby jednostek. Części te pozostają do siebie w określonych proporcjach. Do kwantyli zalicza się:

- kwartyle;
- decyle;
- percentyle.

**Kwartyle dzielą zbiorowość** na cztery części, **decyle** na dziesięć, a **percentyle** na sto. Wszystkie one są miarami położenia, ponieważ określają procent rozkładu liczebności poniżej lub równy wartości danej miary. Najpopularniejszymi kwantylami są kwartyle (w tym mediana) (Dobrowolska, Grzelak, Jarczyński, 2017, s. 38–40).

Sposób, w jaki kwartyle dzielą zbiorowość statystyczną na części, przedstawia rysunek 3.4.




**Rysunek 3.4.** Podział zbiorowości statystycznej na kwartyle

**Źródło:** opracowanie własne.

**Kwartyl pierwszy** ($Q_1$) jest to wartość zmiennej, którą ma jednostka znajdująca się na granicy pierwszej i drugiej ćwiartki zbiorowości. Dzieli więc zbiorowość na dwie części w ten sposób, że 25% ($^1/_4$) jednostek zbiorowości ma wartości niższe bądź równe wartości $Q_1$, a 75% ($^3/_4$) równe lub wyższe od wartości tego kwartyla.

**Kwartyl trzeci** ($Q_3$) to wartość zmiennej, którą ma jednostka znajdująca się na granicy trzeciej i czwartej ćwiartki. Dzieli zbiorowość na dwie części w ten sposób, że 75% ($^3/_4$) jednostek zbiorowości ma wartości cechy niższe bądź równe wartości $Q_3$, a 25% ($^1/_4$) równe lub wyższe od wartości tego kwartyla.

Kwartyle wyznacza się i interpretuje analogicznie jak medianę.

## Inne rodzaje średnich

Średnia arytmetyczna jest miarą zależną od wartości skrajnych. Obliczenie średniej płacy, na przykład w dziale marketingu, może prowadzić do dziwnych wyników, jeżeli w dziale tym pracuje kierownik, którego płaca jest kilka razy wyższa od płac pozostałych osób. Może się zdarzyć, że rzadko występujące w populacji wartości znajdą się przypadkowo w próbie i te wartości zniekształcą wartość średniej. Średnia nie będzie wtedy dobrą charakterystyką rozkładu zmiennej. W takich

sytuacjach lepszą miarą tendencji centralnej jest tzw. średnia obcięta, która jest obliczana po odrzuceniu 5% obserwacji o wartościach najniższych i 5% najwyższych.

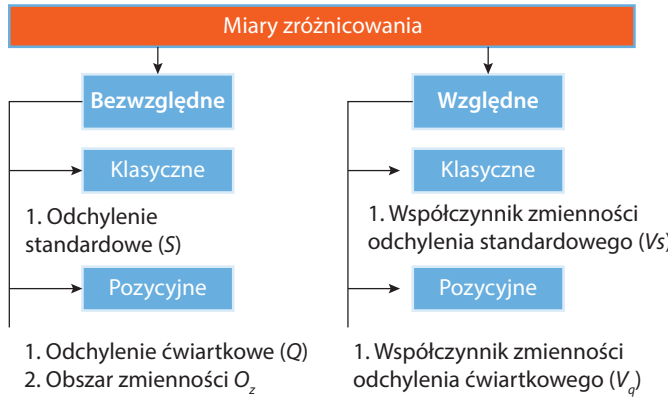
Kolejnym sposobem „niwelowania" wpływu wartości skrajnych na średnią są **M-estymatory** (Malarska, 2005, s. 21–22). Polegają one na różnych sposobach „ważenia" wpływu obserwacji skrajnych (odstających) na średnią. Specjalnie obliczone średnie służą do lepszego określania (szacowania) wartości tendencji centralnej w populacji na podstawie statystyk z próby. IBM SPSS Statistics oblicza cztery takie statystyki, nazwane nazwiskami autorów: Tukeya, Hubera, Andrew i Hampela.

### 3.2.2. Miary zróżnicowania

Opisując rozkład zmiennej, nie można poprzestać na ocenie średniego poziomu, na miarach tendencji centralnej. Porównując rozkład zmiennej w różnych zbiorowościach, zdarza się, że średni poziom zmiennej jest taki sam lub podobny, a rozkłady zmiennej w porównywanych zbiorowościach różnią się wyraźnie między sobą stopniem rozproszenia, co ilustruje rysunek 3.3.

Miary zróżnicowania (zmienności, rozproszenia, rozrzutu, dyspersji) informują przede wszystkim, jak duża jest różnica (odchylenie) wartości zmiennej od poziomu średniego. Stopień, w jakim poszczególne wartości odbiegają od średniej, czyli stopień zmienności, decyduje o znaczeniu danej średniej jako charakterystyki badanego rozkładu. Istnieje wiele miar zróżnicowania, ale poniżej zostaną opisane w sposób syntetyczny tylko najpopularniejsze.

W statystyce wyróżnia się wiele kryteriów i sposobów klasyfikacji miar zróżnicowania. Przyjętą w niniejszym opracowaniu klasyfikację prezentuje rysunek 3.5.




**Rysunek 3.5.** Klasyfikacja miar zróżnicowania

**Źródło:** opracowanie własne.

## Bezwzględne miary zróżnicowania

**Bezwzględne miary zróżnicowania** to wielkości mianowane, wyrażone w jednostkach miary analizowanej zmiennej. Wykorzystywane są w zasadzie do oceny zróżnicowania (rozproszenia) jednej zbiorowości pod względem jednej zmiennej. Porównywanie zróżnicowania danej zmiennej w różnych zbiorowościach za pomocą bezwzględnych miar jest uzasadnione tylko wtedy, gdy średni poziom zmiennej w tych zbiorowościach jest jednakowy lub bardzo podobny.

### Wariancja i odchylenie standardowe

Wariancja i odchylenie standardowe są bardzo ważnymi i popularnymi miarami statystycznymi. Należą do grupy miar klasycznych, co oznacza, że w obliczeniach uwzględnia się wszystkie wartości zmiennej. W przypadku wariancji najpierw oblicza się różnice między wartościami obserwacji a średnią, które następnie podnosi się do kwadratu. Po zsumowaniu tych kwadratów różnic i podzieleniu obliczonej sumy przez liczbę obserwacji uzyskuje się średni kwadrat odchylenia poszczególnych wartości zmiennej od średniej, który nazwano wariancją:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}. \tag{2}$$

Wariancja mierzy więc przeciętne odchylenie kwadratowe od średniej. Ze względu na fakt, że jednostki miary wariancji (kwadrat jednostki zmiennej) są nienaturalne, najczęściej używa się jako miary rozproszenia pierwiastka kwadratowego z wariancji, tj. odchylenia standardowego:

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n}}. \tag{3}$$

Odchylenie standardowe jest więc pierwiastkiem kwadratowym ze średniej arytmetycznej kwadratów odchyleń poszczególnych wartości zmiennej od średniej. Informuje, o ile przeciętnie różnią się wartości zmiennej poszczególnych jednostek od ich średniej. Odchylenie standardowe jest wyrażone, w przeciwieństwie do wariancji, w jednostkach miary analizowanej zmiennej.

### Obszar zmienności

**Obszar zmienności (rozstęp)** jest pozycyjną i jednocześnie najprostszą miarą zróżnicowania. Rozstęp to różnica między największą i najmniejszą wartością zmiennej:

$$O_z = x_{\max} - x_{\min}, \qquad (4)$$

gdzie: $x_{\max}$ – największa wartość cechy w badanej zbiorowości, $x_{\min}$ – najmniejsza wartość cechy w badanej zbiorowości.

Miara ta jest bardzo czuła na dwie skrajne wartości zmiennej, które mogą różnić się znacząco od wszystkich pozostałych wartości, a może się zdarzyć, że są wartościami nietypowymi dla badanej zbiorowości. Obszar zmienności jest więc miarą o małej wartości poznawczej.

Rozstęp to empiryczny obszar zmienności badanej cechy, nieinformujący jednak o zróżnicowaniu poszczególnych wartości w zbiorowości. Miara ta jest najczęściej wykorzystywana do wstępnej oceny zróżnicowania badanej zbiorowości.

### Odchylenie ćwiartkowe

Odchylenie ćwiartkowe ($Q$) wyznacza połowę rozpiętości przedziału, w którym znajduje się połowa obserwacji o wartościach najbliższych medianie:

$$Q = \frac{(Q_3 - Me) + (Me - Q_{1)}}{2} = \frac{Q_3 - Q_1}{2}. \qquad (5)$$

Z powyższego wzoru wynika bezpośrednio, że odchylenie ćwiartkowe to połowa różnicy między trzecim a pierwszym kwartylem.

Odchylenie ćwiartkowe, jako pozycyjna miara zróżnicowania, mierzy poziom zróżnicowania tylko części jednostek. Miara ta jest wykorzystywana wówczas, gdy do opisu tendencji centralnej zastosowano medianę. Określa odchylenie wartości cechy od mediany. Informuje, o ile przeciętnie wartości 50% środkowych jednostek zbiorowości różnią się od mediany. Tym samym odchylenie ćwiartkowe nie mierzy zróżnicowania całej zbiorowości, ale tylko 50% środkowych jednostek. Nie uwzględnia się 25% jednostek o najniższych i 25% jednostek o najwyższych wartościach zmiennej. Na wartość odchylenia ćwiartkowego nie mają wpływu skrajne, często przypadkowe wartości szeregu (Starzyńska, 2009, s. 149).

## Względne miary zróżnicowania (współczynniki zmienności)

W analizach porównawczych rozproszenia różnych zmiennych nie można stosować odchylenia standardowego (wariancji), ze względu na ich zależność od obszaru zmienności porównywanych zmiennych. Tak więc, czy odchylenie standardowe równe 50 jest duże, czy małe? Duże będzie w przypadku wieku osób (w latach), natomiast małe dla płac (w zł). W takich sytuacjach należy stosować względne miary zróżnicowania/rozproszenia, które są niezmienne względem skali. Względne miary zróżnicowania to wielkości stosunkowe, zwane współczynnikami zmienności ($V$), które mierzą stopień, skalę rozproszenia. Ich stosowanie jest niezbędne w porównaniach wielkości zróżnicowania. Wykorzystywane są do porównywania zróżnicowania kilku zbiorowości pod względem jednej cechy lub kilku cech jednej zbiorowości. Najczęściej wyrażone są w procentach. Wartości współczynników zmienności określają procentowy udział bezwzględnego odchylenia zmiennej w wartości miary tendencji centralnej.

**Współczynnik zmienności ($V$)** to stosunek bezwzględnej miary odchylenia (tj. odchylenia standardowego $S$ lub odchylenia ćwiartkowego $Q$) do średniej, wyrażony w procentach. Jak zauważono (rysunek 3.5), współczynniki zmienności możemy podzielić na klasyczne i pozycyjne.

### Współczynnik zmienności odchylenia standardowego

Najczęściej stosowanym klasycznym współczynnikiem zmienności jest współczynnik zmienności odchylenia standardowego:

$$V_S = \frac{S}{\overline{x}} \times 100 \,. \tag{6}$$

Informuje on o tym, jaki jest procentowy udział odchylenia standardowego w średniej arytmetycznej.

### Współczynnik zmienności odchylenia ćwiartkowego

Współczynnik ten należy do grupy miar pozycyjnych. Mierzy zróżnicowanie, podobnie jak odchylenie ćwiartkowe, tylko 50% środkowych jednostek zbiorowości. Wyznacza się go ze wzoru:

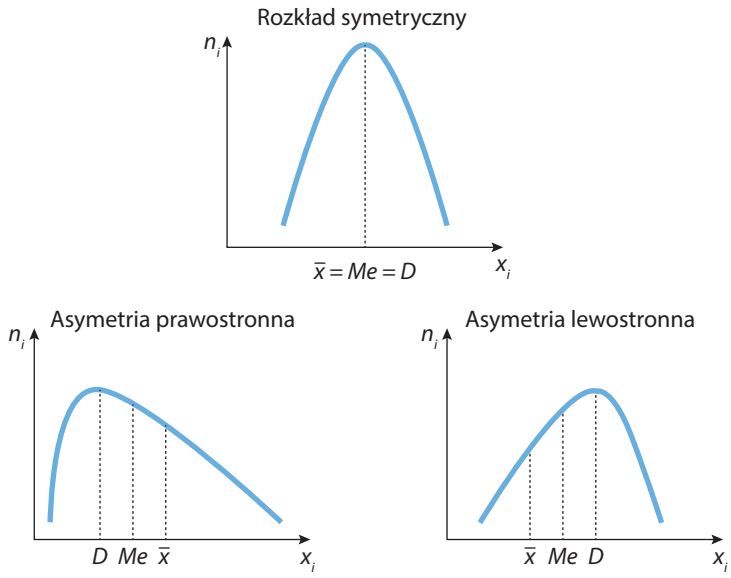
$$V_Q = \frac{Q}{Me} \times 100 \,. \tag{7}$$

Współczynnik zmienności odchylenia ćwiartkowego określa procentowy udział odchylenia ćwiartkowego w wartości mediany (analiza dotyczy 50% środkowych jednostek zbiorowości).

### 3.2.3. Miary skośności rozkładu

Kolejnym etapem analizy rozkładu jest badanie asymetrii (skośności) rozkładu. Ustalenie położenia i zróżnicowania rozkładu zmiennej nie daje jeszcze pełnego opisu. Zdarza się, że takie same wartości miar położenia i zróżnicowania mogą dotyczyć zasadniczo odmiennych rozkładów. Badanie średniego poziomu zmiennej oraz rozproszenia nie obrazuje dostatecznie istnienia różnic między rozkładami, a bardziej szczegółowa obserwacja wyklucza podobieństwo analizowanych rozkładów. W celu pełniejszego opisu rozkładu należy zastosować miary asymetrii (skośności). Dzięki tym miarom możemy się zorientować, czy odchylenia od wartości średniej (centralnej) w jedną stronę są mniej lub bardziej liczne od odchyleń w drugą stronę (Grzelak, 2009, s. 144–146). Analizując na przykład poziom płac w przedsiębiorstwie, obliczamy średnią płacę i chcemy ustalić, czy liczba pracowników, których płaca jest wyższa od płacy średniej, jest większa czy mniejsza od liczby pracowników, których płaca jest niższa od płacy średniej.

**Szereg symetryczny** to taki szereg, w którym liczebności rozkładają się w sposób identyczny po obu stronach dominanty. Zachodzi wówczas równość: $\bar{x} = Me = D_0$. W przypadku **asymetrii prawostronnej** mamy: $D_0 < \mathrm{Me} < \bar{x}$, a dla **asymetrii lewostronnej**: $\bar{x} < Me < D_0$. Relacje te zobrazowano na rysunku 3.6.




**Rysunek 3.6.** Asymetria (skośność) rozkładu zmiennej

**Źródło:** opracowanie własne.

Siłę i kierunek skośności (asymetrii) rozkładu mierzą współczynniki skośności. Poniżej scharakteryzowano jeden z nich – klasyczny współczynnik asymetrii (Pułaska-Turyna, 2005, s. 85–86), który w IBM SPSS Statistics nosi nazwę **Skośność**. Obliczany jest następująco**:**

$$W_{\mu_3} = \frac{\mu_3}{s^3} \tag{8}$$

gdzie $\mu_3$ to trzeci moment centralny.

Trzeci moment centralny (w szeregu szczegółowym) wyznaczany jest zgodnie z poniższą regułą:

$$\mu_3 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{n}. \tag{9}$$

Klasyczny współczynnik asymetrii (skośność) przyjmuje na ogół wartości z przedziału od –2 do +2[5]. W przypadku skrajnie silnej asymetrii współczynnik przekracza (wychodzi poza) powyższe wartości. Znak współczynnika informuje o kierunku asymetrii, a wartość bezwzględna o sile, którą określa się następująco (Pułaska-Turyna, 2005, s. 85–86):

- |0,0–0,4| – asymetria rozkładu bardzo słaba;
- |0,4–0,8| – asymetria rozkładu słaba;
- |0,8–1,2| – asymetria rozkładu umiarkowana;
- |1,2–1,6| – asymetria rozkładu silna;
- |więcej niż 1,6| – asymetria rozkładu bardzo silna.

Skośność jest miarą symetrii, miarą kształtu rozkładu. Jej wartość w przypadku rozkładu normalnego wynosi zero. Jeżeli wartość skośności jest większa od zera, oznacza to, że mamy do czynienia z asymetrią prawostronną (rozkład jest dodatnio skośny, prawoskośny, prawostronny – Bedyńska, Cypryańska, 2013a, s. 104), czyli taką, gdzie częstość występowania wyników niskich jest większa niż wyników wysokich. Jeżeli wartość skośności jest mniejsza od zera, mówimy wtedy o asymetrii

5 W literaturze przedmiotu można znaleźć również bardziej restrykcyjne kryteria (Bulmer, 1979, s. 63): jeśli współczynnik skośności jest większy od +1 lub mniejszy od –1, mamy do czynienia z silną asymetrią, jeśli kształtuje się między –0,5 a –1 lub między +0,5 a +1, asymetria jest umiarkowana, natomiast gdy jest (co do wartości bezwzględnej) mniejszy od 0,5, rozkład jest zbliżony do symetrycznego. Oznacza to nałożenie większych rygorów na wyniki, co ma znaczenie szczególnie w kontekście zgodności rozkładu z rozkładem normalnym.

lewostronnej (o rozkładzie ujemnie skośnym, lewoskośnym), w którym częstość występowania wyników niskich jest mniejsza niż wysokich (przeważają jednostki badania o wartościach zmiennej wyższych niż średnia).

### 3.2.4. Miary koncentracji rozkładu (spłaszczenia)

Omówione wcześniej miary średnie, zróżnicowania i skośności pozwalają już dość dokładnie opisać rozkład, ale do pełnego, wyczerpującego opisu brakuje jeszcze analizy koncentracji. Wyróżnia się dwa rodzaje koncentracji. W niniejszym opracowaniu rozważana będzie jedynie koncentracja rozumiana jako skupienie poszczególnych wartości zmiennej (cechy) wokół średniej. Ten rodzaj koncentracji można analizować tylko w przypadku rozkładów symetrycznych, a miarą skupienia poszczególnych obserwacji wokół średniej jest współczynnik spłaszczenia (skupienia) – kurtoza ($K$), obliczany następująco (Sobczyk, 2000, s. 64):

$$K = \frac{\mu_4}{S^4}, \tag{10}$$

gdzie $\mu_4$ – czwarty moment centralny, obliczany w następujący sposób:

$$\mu_4 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{n}. \tag{11}$$

Im większe jest skupienie (koncentracja) obserwacji wokół średniej, tym mniejsze jest zróżnicowanie rozkładu. Jeżeli zbiorowość ma rozkład normalny, to $K = 3$. Wartość współczynnika skupienia większa od 3 ($K > 3$) charakteryzuje rozkład bardziej wysmukły (skupiony) niż normalny. W przypadku gdy $K < 3$, spłaszczenie jest większe niż normalne.

W praktyce często do oceny skupienia wartości cechy wokół średniej stosuje się współczynnik ekscesu:

$$K' = \frac{\mu_4}{S^4} - 3. \tag{12}$$

Ze względu na wartość współczynnika ekscesu rozkłady dzieli się na:

- mezokurtyczne – $K' = 0$, spłaszczenie normalne;
- leptokurtyczne – $K' > 0$, wartości cechy są bardziej skoncentrowane wokół średniej niż w rozkładzie normalnym (rozkład jest bardziej smukły niż „odpowiedni" rozkład normalny, tj. rozkład normalny o konkretnych parametrach $\mu$ i $\sigma$);

- platykurtyczne – $K' < 0$, wartości cechy są mniej skoncentrowane wokół średniej niż w rozkładzie normalnym (rozkład jest bardziej spłaszczony niż „odpowiedni" rozkład normalny).

W IBM SPSS Statistics, podobnie jak w innych pakietach statystycznych czy Excelu, współczynnik ekscesu to po prostu kurtoza (dla rozkładu normalnego kurtoza jest równa zero).

**Przykład 3.1**

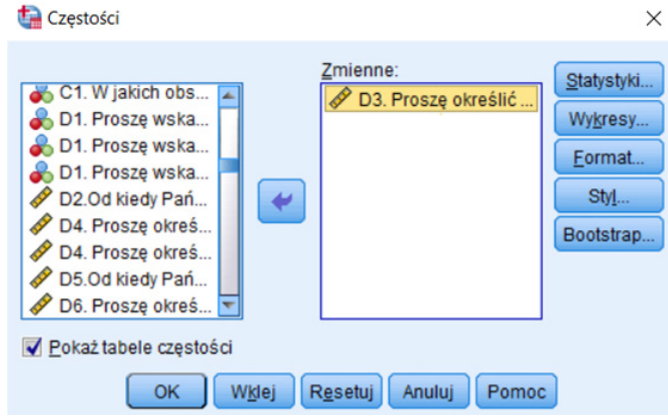
W 2019 roku przeprowadzono badanie 452 losowo wybranych przedsiębiorstw sektora kreatywnego w Polsce. Przedsiębiorstwa badano między innymi pod względem udziału sprzedaży przez Internet w pierwszym półroczu 2018 roku. Dane zapisano w pliku *Kreatywność_2018*. Na podstawie tych danych przeanalizujmy rozkład zmiennej *D3*, oznaczającej procent sprzedaży przez Internet w pierwszym półroczu 2018 roku.

**Rozwiązanie**

Badana zmienna *D3. Proszę określić procent sprzedaży przez Internet w pierwszym półroczu 2018* mierzona jest na skali ilościowej. W związku z tym można dla niej obliczyć wszystkie dostępne statystyki opisowe.

PS IMAGO, w tym jego „silnik" – IBM SPSS Statistics, jest wygodnym narzędziem do przeprowadzania analiz statystycznych. Dzięki niemu można w łatwy sposób bliżej przyjrzeć się danym oraz szybko obliczyć właściwe miary charakteryzujące rozkład zmiennej. Bez rzetelnej oceny rozkładu nie jest możliwe przeprowadzenie właściwych analiz statystycznych i tym samym dokonanie odpowiedniej interpretacji otrzymanych wyników. Punktem wyjścia każdej analizy danych jest zapoznanie się z rozkładem częstości badanej zmiennej.

Jak zaznaczono w punkcie 3.1, aby go sporządzić, wybieramy kolejno: *Analiza* → *Opis statystyczny* → *Częstości*. Na prawą stronę przenosimy zmienną *D3* (rysunek 3.7).




**Rysunek 3.7.** Definiowanie zmiennych poddawanych analizie częstości

Po zatwierdzeniu przyciskiem *OK* pojawia się raport z wynikami badań (rysunek 3.8). Widzimy, że w raporcie wyników znajdują się dwie tabele. Z pierwszej, zatytułowanej *Statystyki*, dowiadujemy się, ile mamy ważnych danych oraz braków danych (czasami są to informacje o badanych obiektach, których pytanie nie dotyczy). W rozważanym przykładzie 99 firm udzieliło odpowiedzi na pytanie, a 353 firmy nie udzieliły odpowiedzi na pytanie. Są to firmy, których pytanie to nie dotyczyło, nie mają bowiem sklepów internetowych.

Druga tabela (*Tabela częstości*) składa się z pięciu kolumn. W pierwszej (na szarym tle) mamy wymienione wartości, jakie przyjmuje badana cecha, czyli *D3. Procent sprzedaży przez Internet w pierwszym półroczu 2018.* W kolumnie drugiej, zatytułowanej *Częstość*, są przedstawione liczebności absolutne, czyli liczebności poszczególnych wartości zmiennej ($n_i$). Pierwszy otrzymany wynik należy odczytać w następujący sposób: pięć firm realizowało 10% swojej sprzedaży przez Internet w pierwszym półroczu 2018 roku. Z kolejnej kolumny, zatytułowanej *Procent*, możemy odczytać, jaki procent wszystkich badanych stanowi dana grupa. W naszym przypadku te 5 firm realizujących 10% swojej sprzedaży przez Internet w pierwszym półroczu 2018 roku stanowi 1,1% wszystkich (452) badanych przedsiębiorstw. Zwróćmy uwagę, że przy wyznaczaniu tego odsetka uwzględnione są również przedsiębiorstwa, których pytanie nie dotyczyło. Bardziej użyteczna byłaby w tej sytuacji informacja, jaki procent stanowią te firmy spośród 99 firm, których pytanie dotyczyło. I taką informację znajdziemy w kolumnie *Procent ważnych*. W naszym przykładzie te 5 firm realizujących 10% swojej sprzedaży przez Internet w pierwszym półroczu 2018 roku stanowi 5,1% spośród 99 przedsiębiorstw, których pytanie dotyczyło. W kolumnie *Procent ważnych* mamy zatem wartości procentowe dla poszczególnych grup lub wartości, gdzie punktem odniesienia są jedynie obserwacje niebędące brakami danych. Ostatnia kolumna w tabeli, tj. *Procent skumulowany*, pozwala szybko sprawdzić, jaki procent całej zbiorowości stanowi grupa badanych

jednostek o określonej wartości zmiennej lub wartościach niższych. Gdybyśmy chcieli ustalić, jaki procent stanowią firmy realizujące łącznie do 50% swojej sprzedaży przez Internet, to możemy odczytać, że tych firm w naszym badaniu było 42,4%. Odpowiedź odczytujemy z kolumny *Procent skumulowany*, z wiersza, gdzie jest wartość 50. Warto wspomnieć, że interpretacja procentu skumulowanego ma sens, jeżeli badana zmienna jest mierzona na co najmniej skali porządkowej.

**Statystyki**

D3. Proszę określić procent sprzedaży przez Internet w pierwszym półroczu 2018:

| | | |
|---|---|---|
| N | Ważne | 99 |
| | Braki danych | 353 |

**D3. Proszę określić procent sprzedaży przez Internet w pierwszym półroczu 2018**

| | | Częstość | Procent | Procent ważnych | Procent skumulowany |
|---|---|---|---|---|---|
| Ważne | 10 | 5 | 1.1 | 5.1 | 5.1 |
| | 15 | 2 | .4 | 2.0 | 7.1 |
| | 20 | 4 | .9 | 4.0 | 11.1 |
| | 25 | 2 | .4 | 2.0 | 13.1 |
| | 30 | 5 | 1.1 | 5.1 | 18.2 |
| | 35 | 3 | 0.7 | 3.0 | 21.2 |
| | 40 | 2 | 0.4 | 2.0 | 23.2 |
| | 50 | 19 | 4.2 | 19.2 | 42.4 |
| | 60 | 5 | 1.1 | 5.1 | 47.5 |
| | 65 | 9 | 2.0 | 9.1 | 56.6 |
| | 70 | 16 | 3.5 | 16.2 | 72.7 |
| | 75 | 7 | 1.5 | 7.1 | 79.8 |
| | 80 | 15 | 3.3 | 15.2 | 94.9 |
| | 85 | 2 | 0.4 | 2.0 | 97.0 |
| | 90 | 3 | 0.7 | 3.0 | 100.0 |
| | Ogółem | 99 | 21.9 | 100.0 | |
| Braki danych | Nie dotyczy | 353 | 78.1 | | |
| Ogółem | | 452 | 100.0 | | |

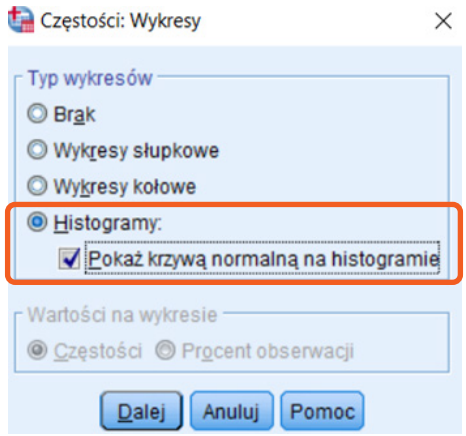
**Rysunek 3.8.** Tabela częstości dla zmiennej *D3*

Dzięki tabelom częstości uzyskujemy podstawowe informacje na temat badanej zmiennej, tj. rozkład liczebności absolutnych i procentowych. Jeżeli w badaniu pojawiłyby się jakieś dziwne wyniki, wtedy już na jego początkowym etapie jesteśmy

w stanie je wychwycić i skorygować. Badanie to pozwala odnaleźć ewentualne błędy popełnione w trakcie wpisywania danych.

Rozkład częstości można przedstawić również graficznie. W PS IMAGO wykorzystuje się w tym celu odrębną zakładkę, służącą do sporządzania różnorakich wykresów (menu główne → *Wykresy*). Jednak trzy najczęściej wykorzystywane typy wykresów, tj. słupkowy, kołowy lub histogram, można szybko sporządzić, wykorzystując opcję funkcji *Częstość* – klikamy kolejno w: *Analiza* → *Opis statystyczny* → *Częstości*, a następnie po prawej stronie wybieramy przycisk *Wykresy* i stosowny wykres.

Należy pamiętać o tym, aby w oknie dialogowym *Częstości* wybrać odpowiednią zmienną, czyli w naszym przykładzie *D3. Procent sprzedaży przez Internet w pierwszym półroczu 2018* (tak jak na rysunku 3.7). W naszym przykładzie zmienna jest mierzona na skali ilościowej – wybieramy zatem histogram (rysunek 3.9).

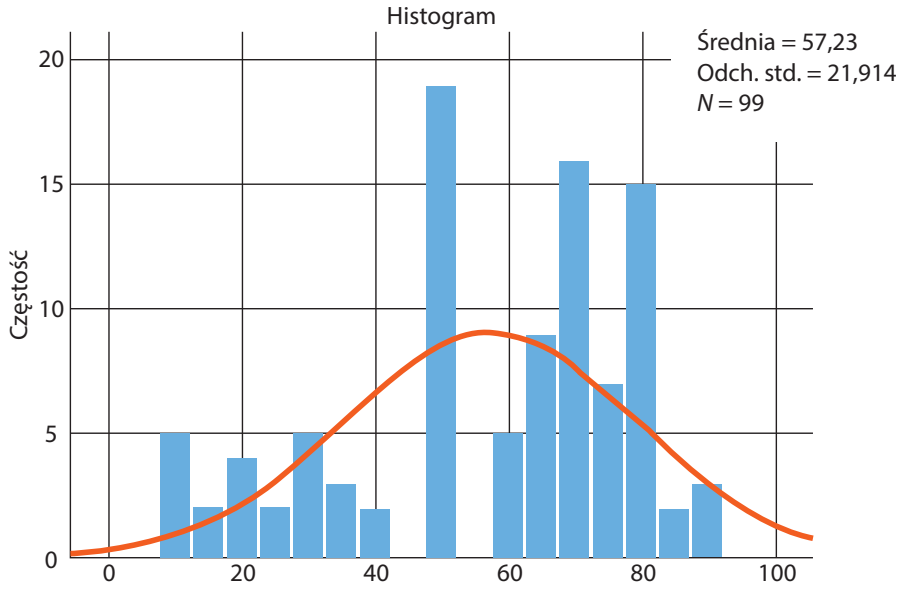



**Rysunek 3.9.** Definiowanie wykresów w oknie *Częstości*

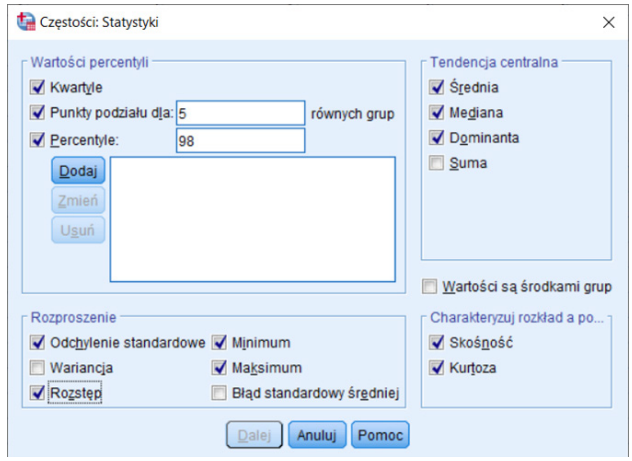
W efekcie otrzymujemy wykres (rysunek 3.10). Warto pamiętać, że na histogramie liczebności są pogrupowane w przedziały o równej rozpiętości.

Kolejnym etapem analizy wyników jest obliczenie podstawowych statystyk opisowych (inaczej – statystyk deskryptywnych), rozumianych jako liczbowe charakterystyki rozkładu wartości danej zmiennej.

Aby obliczyć statystyki opisowe, wybieramy kolejno: *Analiza* → *Opis statystyczny* → *Częstości*, a następnie po prawej stronie wybieramy przycisk *Statystyki*. Jak poprzednio, pamiętamy o tym, aby w oknie dialogowym *Częstości* wybrać odpowiednią zmienną (tu: *D3. Procent sprzedaży przez Internet w pierwszym półroczu 2018*). Następnie pojawia się okno z wykazem omawianych w tym rozdziale statystyk opisowych (rysunek 3.11). Wybieramy stosowne miary i wybór potwierdzamy przyciskami *Dalej* i *Ok*.




**Rysunek 3.10.** Histogram prezentujący rozkład zmiennej ilościowej *D3*




**Rysunek 3.11.** Statystyki dostępne w oknie *Częstości*

W raporcie wyników otrzymujemy wskazane statystyki opisowe (rysunek 3.12).

**Statystyki**

D3. Proszę określić procent sprzedaży odbywający się tym kanałem w pierwszym półroczu 2018:

| | | |
|---|---|---|
| N | Ważne | 99 |
| | Braki danych | 353 |
| Średnia | | 57.22 |
| Mediana | | 65.00 |
| Dominanta | | 50 |
| Odchylenie standardowe | | 21.914 |
| Skośność | | -.691 |
| Błąd standardowy skośności | | .243 |
| Kurtoza | | -.525 |
| Błąd standardowy kurtozy | | .481 |
| Rozstęp | | 80 |
| Minimum | | 10 |
| Maksimum | | 90 |
| Percentyle | 20 | 35.00 |
| | 25 | 50.00 |
| | 40 | 50.00 |
| | 50 | 65.00 |
| | 60 | 70.00 |
| | 75 | 75.00 |
| | 80 | 80.00 |
| | 98 | 90.00 |

**Rysunek 3.12.** Tabela miar statystyki opisowej dla zmiennej *D3*

Na podstawie uzyskanych wyników możemy powiedzieć, że w firmach mających sklep internetowy przeciętny udział sprzedaży przez Internet w sprzedaży ogółem wyniósł 57,22% (zob. wartość średniej arytmetycznej). Najwięcej firm charakteryzowało się pięćdziesięcioprocentowym udziałem sprzedaży przez Internet (zob. wartość dominanty). Jedna czwarta firm miała udział sprzedaży nieprzekraczający 50%, połowa firm osiągnęła sprzedaży nieprzekraczającą 65%, a trzy czwarte firm miało udział sprzedaży do 75% (zob. wartość percentyli – odpowiednio 25 (*Q1*), 50 (*Me*) i 75 (*Q3*)). Obliczone miary statystyki opisowej pozwalają powiedzieć, że w jednej czwartej firm procentowy udział sprzedaży przez Internet wyniósł co najmniej 75% (*Q3*). Obszar zmienności, tj. różnica między maksymalnym i minimalnym udziałem sprzedaży omawianym kanałem sprzedaży, wynosi 80 punktów procentowych. Procentowy udział sprzedaży przez Internet w poszczególnych firmach różni się przeciętnie od średniego udziału,

wynoszącego 57,22%, o +/–21,914 punktu procentowego (zob. wartość odchylenia standardowego).

Rozkład omawianej zmiennej (*procent sprzedaży przez Internet*), mierzony klasycznym współczynnikiem asymetrii (*Skośność*), charakteryzuje się słabą asymetrią ujemną (współczynnik na poziomie –0,691). Oznacza to, że więcej jest firm z udziałem sprzedaży powyżej niż poniżej średniej. Kurtoza, wynosząca –0,525, wskazuje, że rozkład udziału sprzedaży przez Internet jest platykurtyczny, a więc koncentracja wartości wokół średniej jest mniejsza niż w rozkładzie normalnym (rozkład jest bardziej spłaszczony niż w „odpowiednim" rozkładzie normalnym).

Dokonując kompleksowej oceny struktury zbiorowości, istotne jest, aby ocenić, czy do opisu zbiorowości możemy wykorzystać miary klasyczne i ewentualnie opis danej zmiennej wzbogacić o wyniki, jakie dają nam miary pozycyjne. Z taką sytuacją mieliśmy do czynienia w przykładzie pierwszym, rozkład zmiennej *D3. Procent sprzedaży przez Internet w I połowie 2018* charakteryzował się bowiem niewielką skośnością i wtedy średnia arytmetyczna (oraz inne miary klasyczne) jest dobrą charakterystyką takiego rozkładu. Może się jednak zdarzyć, że rozkład zmiennej będzie wyraźnie asymetryczny i wtedy do opisu takiej zmiennej musimy wykorzystać miary pozycyjne i inne charakterystyki rozkładu. Omówimy to zagadnienie na przykładzie 3.2.

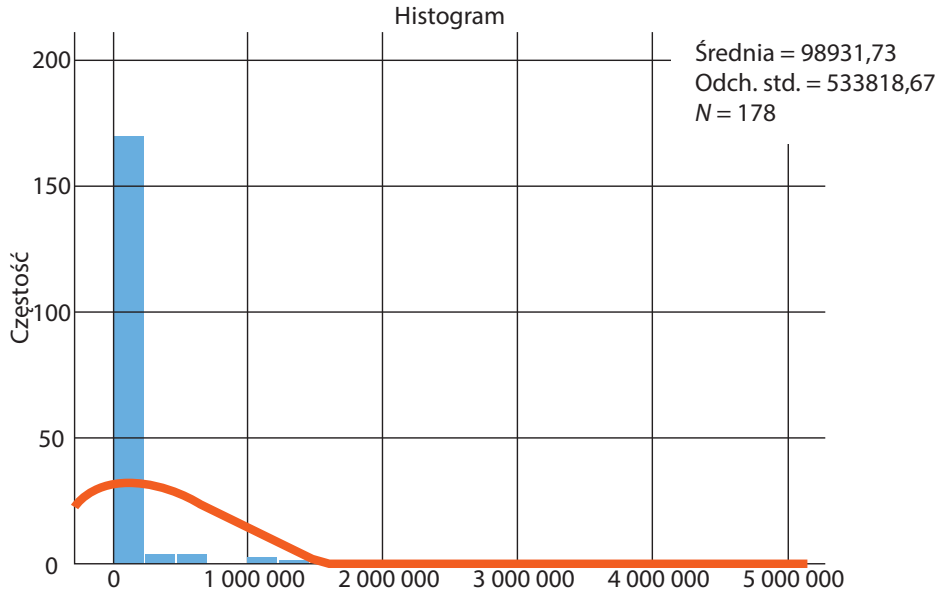
**Przykład 3.2**

W badaniu przedsiębiorstw zlokalizowanych w pewnej gminie wybrano losową próbę ($n$ = 178), w której oceniano między innymi przychody ze sprzedaży (w tys. zł) w I półroczu 2018 roku (zmienna *przychody ze sprzedaży*). Dokonajmy analizy rozkładu tej zmiennej w badanej zbiorowości.

**Rozwiązanie**

Badana zmienna mierzona jest na skali ilościowej. Ponownie zatem możliwe jest wyznaczenie wielu statystyk opisowych.

Zaczynamy analizę od sporządzenia wykresu, który pozwoli nam wstępnie ocenić rozkład zmiennej *przychody ze sprzedaży*. W tym celu wybieramy kolejno: *Analiza → Opis statystyczny → Częstości*, a następnie po prawej stronie przycisk *Wykresy*. Pamiętamy o tym, aby w oknie dialogowym *Częstości* wybrać odpowiednią zmienną, czyli w naszym przykładzie *przychody ze sprzedaży* (tak jak na rysunku 3.1). Wybieramy stosowny wykres (tu będzie to histogram – z uwagi na ilościowy poziom pomiaru zmiennej). W efekcie otrzymujemy wykres (rysunek 3.13).




**Rysunek 3.13.** Histogram prezentujący rozkład zmiennej ilościowej *przychody ze sprzedaży*

**Statystyki**

Przychody ze sprzedaży

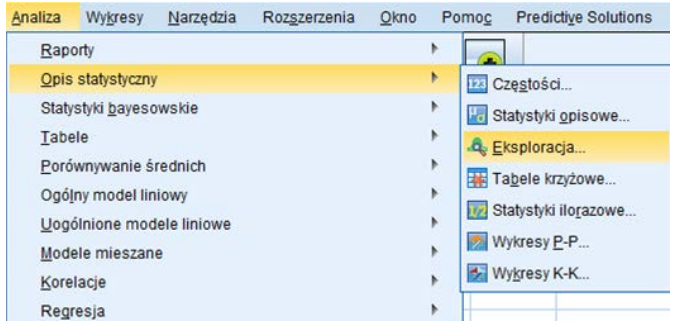
| | | |
|---|---|---|
| N | Ważne | 178 |
| | Braki danych | 0 |
| Średnia | | 98931.7334 |
| Mediana | | 1346.0320 |
| Dominanta | | 5526.00 |
| Odchylenie standardowe | | 533818.66987 |
| Skośność | | 6.826 |
| Błąd standardowy skośności | | .182 |
| Kurtoza | | 48.424 |
| Błąd standardowy kurtozy | | .362 |
| Rozstęp | | 4367080.85 |
| Minimum | | .00 |
| Maksimum | | 4367080.85 |
| Percentyle | 25 | 262.4475 |
| | 50 | 1346.0320 |
| | 75 | 5716.3320 |

**Rysunek 3.14.** Statystyki opisowej dla zmiennej *przychody ze sprzedaży*

Warto zwrócić uwagę na podstawowe charakterystyki, które pojawiają się na wykresie. Odchylenie standardowe ma wartość pięć razy większą niż średnia arytmetyczna. Rozkład jest silnie asymetryczny (dodatnio skośny) i w takiej sytuacji miary klasyczne nie będą właściwie charakteryzowały takiego rozkładu. Potwierdzają to obliczone statystyki opisowe (rysunek 3.14).

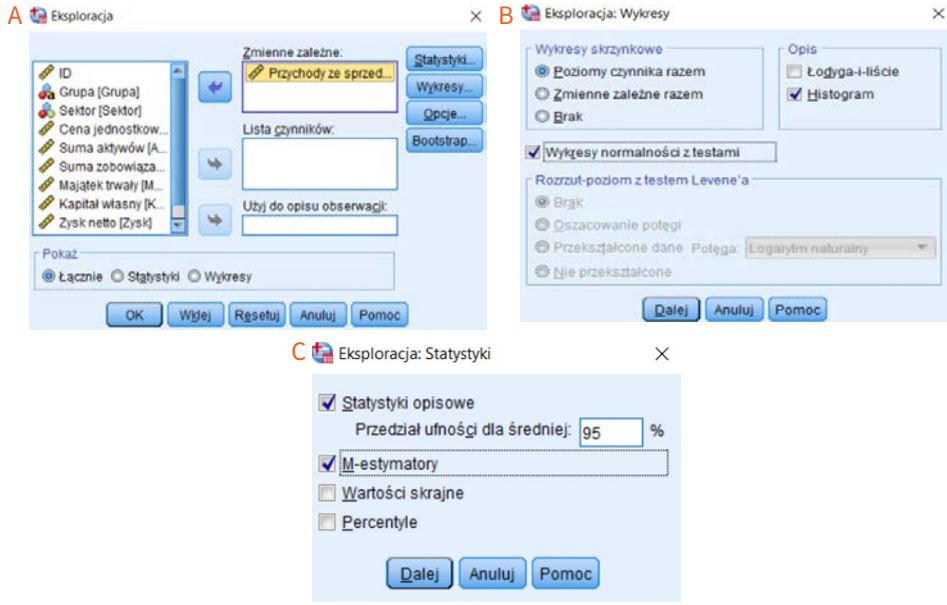
Trzeba zwrócić uwagę, że rozkład charakteryzuje się silną asymetrią, o czym świadczy bardzo wysoka wartość klasycznego współczynnika asymetrii (6,826), a także silną smukłością rozkładu, na co wskazuje wartość kurtozy (48,424). Ponadto wartość współczynnika skośności podzielonego przez błąd standardowy skośności (6,826/0,182) wynosi 9,5042. Mamy więc podstawy, aby uznać dany rozkład za odstający od rozkładu normalnego (stosunek ten przekracza bowiem 2). Potwierdza to również analiza relacji kurtozy do błędu standardowego kurtozy, również znacznie przekraczająca 2 (48,424/0,362 = 133,77).

Dodatkowo warto obliczyć inne pomocne do opisu rozkładu charakterystyki, tj. M-estymatory, przedziały ufności, 5% średnią obciętą, oraz sporządzić wykresy: skrzynkowy, K-K bez trendu i K-K z trendem. Analiza ta pozwoli również wstępnie ocenić, jak bardzo rozkład zmiennej jest podobny do rozkładu normalnego, którego własności omówiono w rozdziale drugim. Obliczenia tych charakterystyk rozkładu możemy łatwo przeprowadzić w PS IMAGO. W tym celu wybieramy: *Analiza → Opis statystyczny → Eksploracja* (rysunek 3.15).




**Rysunek 3.15.** Wykonywanie polecenia *Eksploracja*

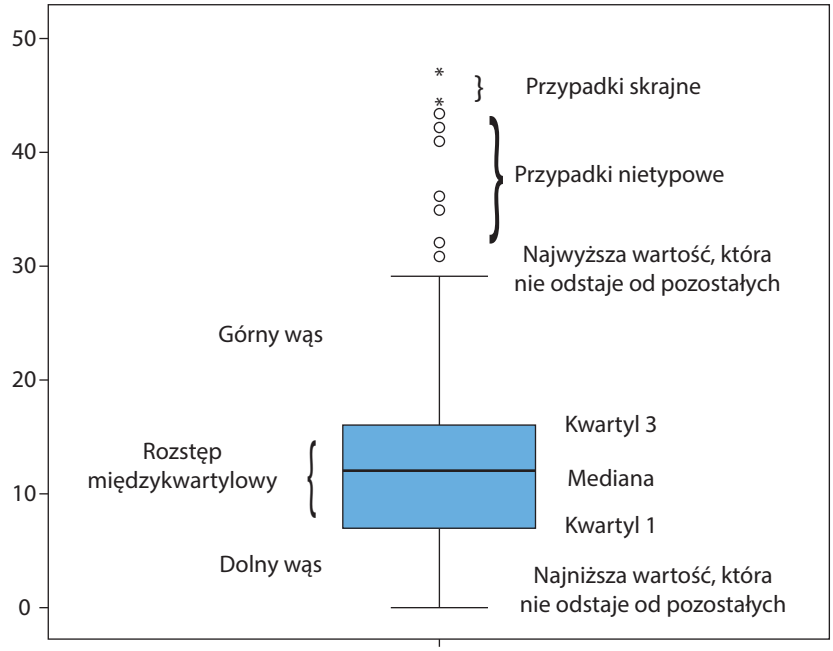
W kolejnym oknie dialogowym, które pojawia się w PS IMAGO, wskazujemy zmienną zależną, czyli w naszym przykładzie *przychody ze sprzedaży* (rysunek 3.16A). Po prawej stronie wybieramy przycisk *Wykresy*, gdzie wskazujemy: *Wykresy skrzynkowe → Poziomy czynnika razem, Opis → Histogram*, oraz zaznaczmy, aby program wykonał test normalności rozkładu (*Wykresy normalności z testami*) (rysunek 3.16B). Następnie w zakładce *Statystyki* wskazujemy *Statystyki opisowe* i *M-estymatory* (rysunek 3.16C).




**Rysunek 3.16.** Definiowanie procedury *Eksploracja* dla zmiennej *przychody ze sprzedaży*

W efekcie obliczeń otrzymujemy interesujące nas wyniki (ograniczymy się w tym miejscu do metod statystyki opisowej, pominiemy tym samym testy normalności rozkładu – wrócimy do nich w kolejnym rozdziale).

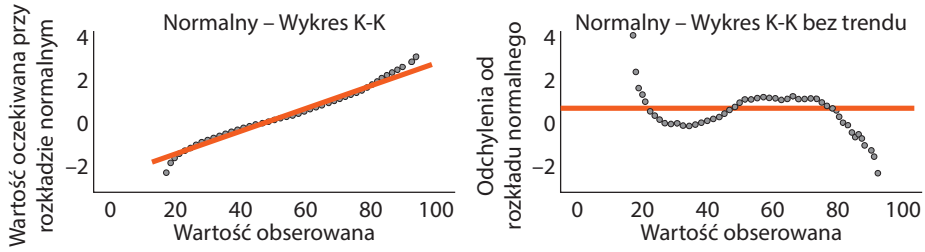
Warto w tym miejscu wyjaśnić, w jaki sposób interpretuje się wymienione wcześniej wykresy – skrzynkowy, K-K bez trendu, K-K z trendem. Wykres skrzynkowy (określany też jako wykres pudełkowy lub *box-plot*) interpretuje się zgodnie z rysunkiem 3.17. Linia w środku oznacza medianę. Góra i dół skrzynki (tzw. zawiasy) wskazują wartości pierwszego i trzeciego kwartyla (wyznaczają więc rozstęp ćwiartkowy, określany też jako rozstęp międzykwartylowy). Wąsy reprezentują rozstęp wartości tych obserwacji, które nie odstają od pozostałych. **Długość wąsów** interpretuje się jako miarę zróżnicowania – im są dłuższe, tym zróżnicowanie wyników jest większe. Obserwacje, których wartości leżą poza skrzynką i wąsami, traktujemy jako wartości odstające (*outliers*). Obserwacje te oznaczane są kółeczkiem, jeśli są odległe o 1,5–3 rozstępy ćwiartkowe od zawiasów, a gwiazdką, jeśli **są odległe o więcej niż** 3 rozstępy ćwiartkowe. Wydłużona górna część wykresu (dłuższy górny wąs i – ewentualnie – punkty wskazujące na wartości odstające) świadczy o występowaniu asymetrii prawostronnej, a wydłużona dolna część wykresu (dłuższy dolny wąs i – ewentualnie – punkty wskazujące na wartości odstające) o występowaniu asymetrii lewostronnej. Im bardziej wydłużona jest górna lub dolna część wykresu, tym siła skośności jest większa.




**Rysunek 3.17.** Schemat wykresu skrzynkowego

**Źródło:** Malarska, 2005, s. 26.

Z kolei tzw. wykresy normalności, czyli wykresy K-K (kwantyl – kwantyl) – zarówno z trendem (lewy panel rysunku 3.18), jak i bez trendu (prawy panel rysunku 3.18) – pozwalają na ocenę zbieżności rozkładu zmiennej z rozkładem normalnym.



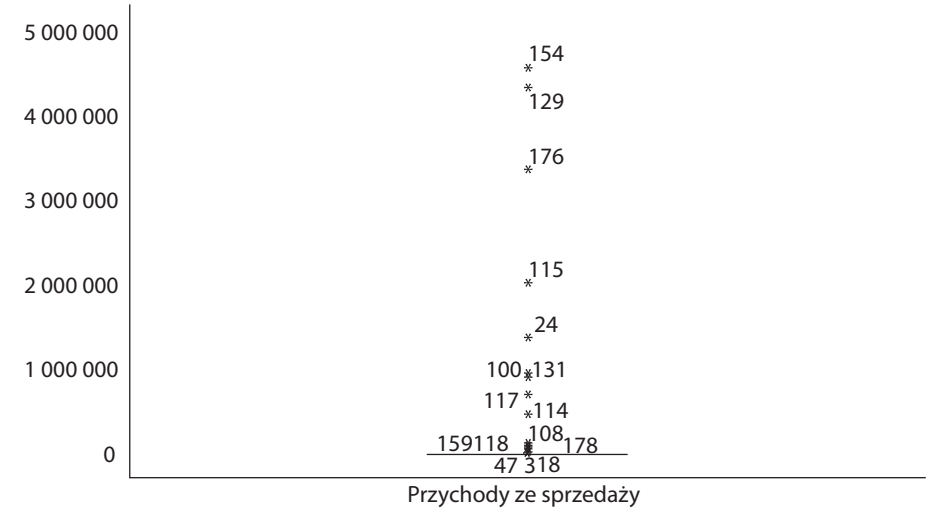

**Rysunek 3.18.** Wykresy normalności (K-K)

**Źródło:** opracowanie własne.

W przypadku wykresu normalności K-K z trendem dla rozkładu normalnego punkty układają się dokładnie na prostej, a zatem im bardziej od niej odbiegają, tym większe odstępstwa od normalności rozkładu. Z kolei wykres normalności K-K bez trendu czytelniej opisuje poziom odchyleń rozkładu empirycznego

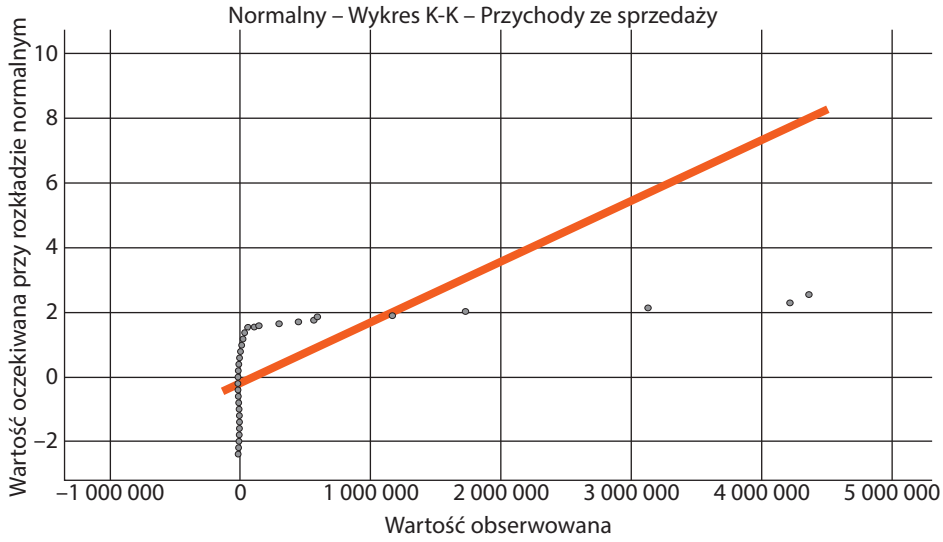
od hipotetycznego rozkładu normalnego. Na wykresie tym oczekiwane są jak najmniejsze dystanse wszystkich punktów empirycznych od linii regresji. W myśl teorii trzech sigm odstępstwa te nie powinny przekraczać +/–3 odchyleń standardowych (odczytujemy te wartości na osi y) (Malarska, 2005, s. 32–33). W przypadku danych zaprezentowanych na rysunku 3.18 na wykresie K-K (lewy panel) punkty układają się mniej więcej na prostej, a na wykresie K-K z trendem punkty układają się mniej więcej w przedziale od –0,6 do 0,6. Oba wykresy wskazują więc na to, że rozkład analizowanej zmiennej nie odbiega znacząco od rozkładu normalnego (można powiedzieć, że jest zbieżny z rozkładem normalnym).

W naszym przykładzie wykres skrzynkowy (rysunek 3.19) nawet nie przypomina charakterystycznej skrzynki, z jaką zwykle kojarzy się ten typ wykresu. Liczne gwiazdki na wykresie wskazują, że występuje wiele obserwacji oddalonych od skrzynki o więcej niż 3 rozstępy ćwiartkowe. Wyraźnie sugeruje to, że odstępstwa od normalności rozkładu są w przypadku naszej zmiennej bardzo wyraźne.

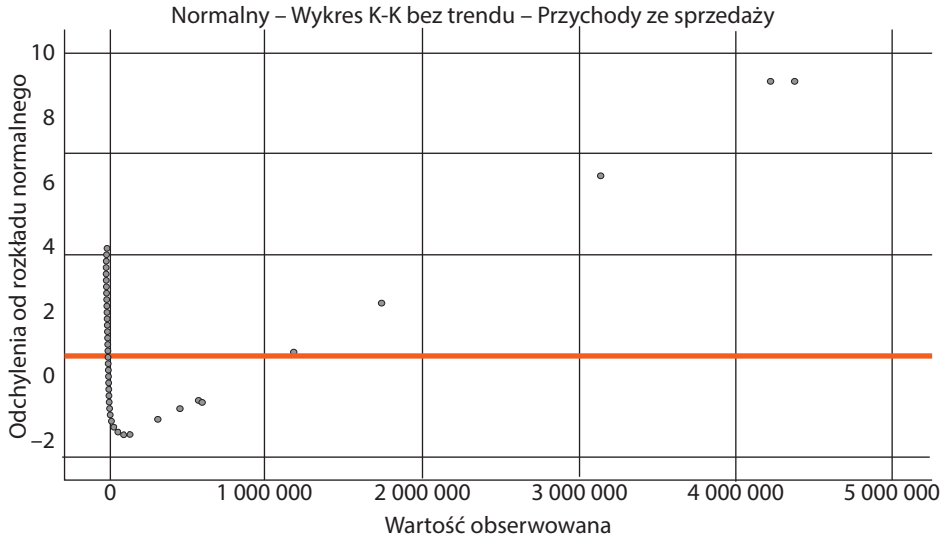



**Rysunek 3.19.** Wykres skrzynkowy dla zmiennej *przychody ze sprzedaży*

Również wykresy K-K z trendem i bez trendu potwierdzają, że odchylenia badanego rozkładu empirycznego od rozkładu normalnego są bardzo duże (rysunki 3.20 i 3.21). Na wykresie K-K z trendem większość punktów leży poza prostą. Wykres K-K bez trendu wskazuje z kolei, że odchylenia badanego rozkładu empirycznego od rozkładu normalnego mieszczą się w obszarze od –2 do 6 sigma (rysunek 3.21).




**Rysunek 3.20.** Wykres K-K z trendem dla zmiennej *przychody ze sprzedaży*




**Rysunek 3.21.** Wykres K-K bez trendu dla zmiennej *przychody ze sprzedaży*

W przypadku wstępnej oceny rozkładu zmiennej z punktu widzenia normalności rozkładu na podstawie statystyk opisowych zwracamy szczególną uwagę – oprócz miar symetrii rozkładu (skośności i kurtozy), o których już mówiliśmy wcześniej – na wartości M-estymatorów (rysunek 3.22).

M-estymatory to mocne estymatory wartości oczekiwanej, odporne na przypadki odstające. Osłabiają one wagę przypadków odstających i jeżeli takie przypadki

odstające występują, **wówczas M-estymatory różnią się znacząco od średniej. Dokładnie taka sytuacja ma miejsce w przypadku naszej zmiennej.**

**Statystyki opisowe (DESCRIPTIVES)**

| | | | Statystyka | Błąd standardowy |
|---|---|---|---|---|
| Przychody ze sprzedaży | Średnia | | 98931.7334 | 40011.40092 |
| | 95% przedział ufności dla średniej | Dolna granica | 19970.9461 | |
| | | Górna granica | 177892.5207 | |
| | 5% średnia obcięta | | 6643.1807 | |
| | Mediana | | 1346.0320 | |
| | Wariancja | | 284962372301.760 | |
| | Odchylenie standardowe | | 533818.66987 | |
| | Minimum | | .00 | |
| | Maksimum | | 4.37E+6 | |
| | Rozstęp | | 4367080.85 | |
| | Rozstęp ćwiartkowy | | 5453.88 | |
| | Skośność | | 6.826 | .182 |
| | Kurtoza | | 48.424 | .362 |

**M-estymatory**

| | Huber[a] | Tukey[b] | Hampel[c] | Andrew[d] |
|---|---|---|---|---|
| Przychody ze sprzedaży | 1669.4461 | 925.9489 | 1275.7647 | 917.6360 |

[a] Stała ważąca wynosi 1,339.

[b] Stała ważąca wynosi 4,685.

[c] Stałe ważące wynoszą 1,700, 3,400 i 8,500.

[d] Stała ważąca wynosi 1,340*pi.

**Rysunek 3.22.** Tabela miar statystyki opisowej i M-estymatorów dla zmiennej *przychody ze sprzedaży*

Wykorzystanie do obliczeń statystyk opisowych funkcji *Eksploracja* dostarcza nam dodatkowo informacji o: przedziale ufności, 5% średniej obciętej oraz rozstępie ćwiartkowym (różnicy między kwartylem trzecim a pierwszym). Przedział ufności zmiennej *przychody ze sprzedaży* charakteryzuje się dużą rozpiętością (157921,6). Ponadto 5% średnia obcięta (6643,2), do której obliczenia odrzucamy 5% najbardziej odstających obserwacji, różni się znacząco od średniej arytmetycznej (98931,7). Jak zaznaczono, również M-estymatory znacznie od niej odbiegają (rysunek 3.22).

Podsumowując, przeprowadzona wstępna analiza rozkładu zmiennej *przychody ze sprzedaży* wskazuje, że nie jest on zbliżony do rozkładu normalnego lub chociażby symetryczny, do oceny tej zmiennej możemy zatem wykorzystać tylko miary pozycyjne.

I tak powiemy, że badane firmy najczęściej osiągały przychody ze sprzedaży w pierwszym półroczu 2018 roku na poziomie 5526 tys. zł (zob. wartość dominanty na rysunku 3.14). Przychody ze sprzedaży nieprzekraczające 262,45 tys. zł osiągnęło 25% firm, połowa firm osiągnęła przychody ze sprzedaży nieprzekraczające 1346,03 tys. zł, a 75% firm osiągnęło przychody ze sprzedaży do 5716,33 tys. zł. (zob. wartość percentyli – odpowiednio 25, 50 i 75). Obszar zmienności, tj. różnica między maksymalnymi i minimalnymi przychodami ze sprzedaży, wynosił 4 367 080,85 tys. zł. Rozkład badanej zmiennej charakteryzuje się silną asymetrią, o czym mówiliśmy już wcześniej.

# 4. Porównanie dwóch populacji

**Kluczowe pojęcia:** porównanie dwóch populacji, próby niezależne i zależne, rozkład normalny, jednorodność wariancji, testy parametryczne i nieparametryczne, istotność różnic, test t-Studenta, test Shapiro-Wilka, test Levene'a, test Manna-Whitneya

## 4.1. Uwagi wstępne

Przedmiotem wielu badań jest porównanie dwóch populacji/podpopulacji na podstawie prób losowych (pomiarów wykonanych na próbach losowych pobranych z populacji będących przedmiotem zainteresowania). W badaniach sondażowych interesuje nas na przykład to, czy badani z dwóch populacji/podpopulacji/segmentów (np. mieszkańcy wsi i miast) różnią się pod względem badanej cechy lub czy określona zmienna, na przykład płeć, różnicuje postawy. Pojawia się wtedy problem, jak sprawdzić różnice między dwiema populacjami, który z testów należy zastosować. W literaturze przedmiotu wymienia się kryteria (warunki) wyboru właściwego testu statystycznego. Jak sygnalizowano w rozdziale drugim, to, który z testów należy zastosować, w pierwszej kolejności zależy od charakteru porównywanych prób (niezależne/zależne) i od skali pomiarowej zmiennej zależnej. Jeżeli zmienna zależna mierzona jest na skali ilościowej, wtedy do porównania dwóch populacji optymalnym rozwiązaniem jest test parametryczny. Warto w tym miejscu przypomnieć, iż testy parametryczne mają większą moc niż testy nieparametryczne (niższy błąd przyjęcia fałszywej hipotezy zerowej).

W rozdziale trzecim omówiono zastosowanie średniej arytmetycznej do syntetycznego opisu zmiennej ilościowej w badanej zbiorowości, a więc – w przypadku badania częściowego – w próbie. Porównanie rozkładu zmiennej ilościowej w dwóch populacjach można sprowadzić do porównania średnich w tych populacjach. Porównywanie dwóch populacji (generalnych) pod względem średniej wartości zmiennej nie oznacza oczywiście, że chcemy porównać średnie wartości obliczone na podstawie próby. Chodzi o porównanie średnich w populacjach/podpopulacjach, a dokładniej mówiąc – o porównanie wartości oczekiwanych,

natomiast wyniki uzyskane w próbach są podstawą porównań populacji. Na podstawie statystyk opisowych, takich jak średnia arytmetyczna, nie można jednoznacznie przesądzać, czy średnie te różnią się istotnie statystycznie. Zagadnienie porównywania dwóch średnich w populacjach to problem testowania właściwych hipotez statystycznych. Do testowania istotności różnic między dwiema średnimi służy **test t-Studenta** (w wersji dla prób zależnych lub niezależnych).

Test t-Studenta jest testem parametrycznym, co oznacza, że aby go stosować, rozkład zmiennej zależnej powinien spełniać pewne założenia (Bedyńska, Brzezicka, 2007, s. 163–164): (1) powinien być zbieżny z rozkładem normalnym, (2) wariancje zmiennej zależnej w podpopulacjach powinny być jednorodne.

W badaniach eksperymentalnych zakłada się dodatkowo, że liczebność prób powinna być podobna. W badaniach sondażowych założenie to się pomija, aczkolwiek im liczebności prób są bardziej podobne, tym metoda ta jest bardziej odporna na odstępstwa od założeń.

Przeanalizujmy na początek sposób weryfikacji założeń (1) i (2).

## 4.2. Sprawdzanie założeń testu t-Studenta

### Normalność rozkładu zmiennej zależnej w podpopulacjach

Sprawdzamy, czy rozkład może zostać uznany za nieróżniący się istotnie kształtem od wzorca (tj. rozkładu normalnego) testem Shapiro-Wilka (lub testem Kołmogorowa-Smirnowa – przy większych próbach). W obu testach hipotezy mają postać:

$H_0$: rozkład zmiennej zależnej jest normalny

$H_1$: $\neg H_0$.

Sprawdzian testu Shapiro-Wilka ma postać (Cieciura, Zacharski, 2007, s. 423):

$$W = \frac{\left[\sum_{i=1}^{B} a_{n-i+1}\left(x_{n-i+1} - x_i\right)\right]^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}, \tag{13}$$

gdzie: $x_{(1)}, \ldots, x_{(n)}$ – uporządkowana próba według wartości rosnących, $a_{n,i}$ – stablicowane współczynniki testu Shapiro-Wilka, $B = \left[\frac{n}{2}\right]$ – część całkowita z $\frac{n}{2}$.

Statystyka ta przy prawdziwości $H_0$ ma znany i stablicowany rozkład. Obszar krytyczny testu jest lewostronny. Wartości krytyczne testu i współczynników

$a_{n,i}$ zostały stablicowane przez Shapiro-Wilka (1965) dla $n \leq 50$. W 1995 roku Royston zaproponował algorytm, który pozwala na zastosowanie testu Shapiro-Wilka dla $3 \leq n \leq 5000$ (Royston, 1995)[6]. W praktyce stosowany jest w związku z tym przede wszystkim test Shapiro-Wilka. Rozkład zmiennej uznamy za zbieżny do rozkładu normalnego, jeśli $p > \alpha$. W przypadku gdy $p < \alpha$, odstępstwa od rozkładu normalnego uznajemy za statystycznie istotne.

W przypadku testu t-Studenta na mocy centralnych twierdzeń granicznych można pominąć ocenę normalności rozkładu, jeśli $n > 30$. Jeżeli dana próba liczy więcej niż 30 elementów, to można przeprowadzić procedurę testowania tak, jakby cecha miała rozkład normalny (Szymczak, 2010, s. 197). Niemniej jednak przy skrajnej asymetrii rozkładu podejście takie może prowadzić do zniekształconego obrazu badanych populacji. Im większa próba, tym odporność testu na odstępstwa od założeń jest większa, aczkolwiek i przy dużych próbach przy silnych odstępstwach od normalności rozkładu odpowiednim rozwiązaniem będzie zastosowanie testu nieparametrycznego – zaleca się test Manna-Whitneya. Możliwe jest też przeprowadzenie testu t-Studenta po wcześniejszym przekształceniu zmiennych (jednym z najczęściej stosowanych jest przekształcenie logarytmiczne, które sprawdza się zwłaszcza przy silnej asymetrii prawostronnej).

## Jednorodność wariancji zmiennej zależnej w podpopulacjach

**Jednorodność wariancji** zmiennej zależnej w podpopulacjach (albo krócej w populacjach – chodzi o podgrupy porównywane w badaniu) sprawdzić można testem Levene'a. Hipotezy są następując0 w indeksie dolnym

H0

:

$H_0$: wariancja zmiennej zależnej jest taka sama w porównywanych populacjach

$H_1$: ¬ $H_0$;

lub

6 W teście Kołmogorowa-Smirnowa z kolei sprawdzianem jest statystyka: $D_L = \max_{x_i} \left| F\left(x_i\right) - F^*_{N(\overline{x};S(x))}\left(x_i\right) \right|$, gdzie: $F(x_i)$ – dystrybuanta empiryczna zmiennej $X$, $F^*_{N(\overline{x};S(x))}\left(x_i\right)$ – dystrybuanta rozkładu normalnego zmiennej $X$ o średniej ( $\overline{x}$ ) i odchyleniu standardowym $S$ oszacowanych z próby. W 1967 roku Lilliefors zaprezentował tablice wartości krytycznych sprawdzianu $D_L$ Kołmogorowa-Smirnowa dla sytuacji, w której nie są znane średnia ($\mu$) i odchylenie standardowe ($\sigma$) dla populacji, z której pochodzi próba. W IBM SPSS Statistics dostępna jest wersja testu Kołmogorowa-Smirnowa z poprawką Lilieforsa.

$$H_0 : \sigma_1^2 = \sigma_2^2;$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Wartość statystyki testującej wyznaczana jest według wzoru (Rószkiewicz, 2011, s. 89–90):

$$\mathrm{F} = \frac{(n-2)\sum_{i=1}^{k} n_i \left(\overline{z}_i - \overline{z}\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \overline{z}\right)^2 n_i}, \tag{14}$$

gdzie $z_{ij} = \left|x_{ij} - \overline{x}_i\right|$, $\overline{z}_i = \frac{\sum_{j}^{n_i} z_{ij}}{n_i}$.

Statystyka ta ma rozkład *F* (określany też jako rozkład F Snedecora albo rozkład Fishera-Snedecora) o liczbie stopni swobody $v_1 = 1$ i $v_2 = n - 1$.

Hipotezę zerową można również sformułować następująco: „wariancje są jednorodne" albo „wariancje są homogeniczne". Założenie to uznamy za spełnione, jeśli $p > \alpha$. W przypadku gdy $p < \alpha$, wariancje uznajemy za niejednorodne. Jeżeli założenie nie zostało spełnione, to należy zastosować odporną („mocną") wersję testu t-Studenta.

## 4.3. Test t-Studenta

W przypadku testu t-Studenta (w obu wersjach – podstawowej i odpornej) testowane są te same hipotezy:
$H_0$: wartości oczekiwane w porównywanych populacjach są jednakowe
$H_1$: wartości oczekiwane w porównywanych populacjach nie są jednakowe;
lub
$H_0$: $\mu_1 = \mu_2$
$H_1$: $\neg H_0$.

Wartość sprawdzianu testu (statystyki) oblicza się na podstawie wzoru, w którego liczniku jest różnica średnich, a w mianowniku błąd standardowy tej różnicy.

Wzór na statystykę *t* dla jednorodnych wariancji (Greń, 1972, s. 66):

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_s S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{15}$$

Przy założeniu prawdziwości $H_0$ ma on rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

Wzór na statystykę *t* dla niejednorodnych wariancji (Greń, 1972, s. 65):

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \tag{16}$$

Statystyka ma rozkład t-Studenta o $\left(\frac{1}{2} + \frac{S_1^2 S_2^2}{S_1^2 + S_2^2}\right)\left(n_1 + n_2 - 2\right)$ stopniach swobody.

We wzorach 15–16 $\overline{x}_1$, $\overline{x}_2$ – średnie z prób; $S_1^2$, $S_2^2$ – wariancje z prób; $n_1$, $n_2$ – liczebności prób.

Różnice uznaje się za statystycznie istotne, jeśli $p < \alpha$. Wówczas kolejnym krokiem jest ustalenie, w której populacji wartość oczekiwana $\mu$ jest istotnie wyższa, co można zrobić, porównując wartości estymatorów wartości oczekiwanej (średnią arytmetyczną) w obu próbach.

## 4.4. Test Manna-Whitneya

W przypadku testu Manna-Whitneya testowane są hipotezy, które mogą być sformułowane w następujący sposób:

$H_0$: dwie niezależne próbki pochodzą z populacji o takim samym rozkładzie

$H_1$: nieprawda, że dwie niezależne próbki pochodzą z populacji o takim samym rozkładzie;

lub:

$H_0$: $F_1 = F_2$

$H_1$: $\neg H_0$,

gdzie $F_1$ i $F_2$ są dystrybuantami rozkładów prawdopodobieństwa badanej cechy w porównywanych populacjach (Szymczak, 2010, s. 198).

Statystyki stanowiące podstawę testu, gdy nie ma rang wiązanych, są wyrażone wzorami:

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot \left(n_1 + 1\right)}{2} - R_1, \tag{17}$$

$$U' = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2 , \tag{18}$$

gdzie: $R_1$ *i* $R_2$ – suma rang w poszczególnych próbach. Mniejsza z wartości $U$ i $U'$ odpowiada statystyce $W$ Wilcoxona.

W praktyce stosuje się funkcję $Z$ statystyki $U$, mającą w przybliżeniu rozkład normalny:

$$Z = \frac{U - \frac{1}{2} n_1 n_2}{\sqrt{\frac{1}{12} \cdot n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}} . \tag{19}$$

W przypadku gdy występują rangi wiązane, sprawdzianem testu Manna-Whitneya jest statystyka $Z$ również mającą w przybliżeniu rozkład normalny:

$$Z = \frac{U - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 \cdot n_2}{n \cdot (n-1)} \left[ \frac{n^3 - n}{12} - \sum T_i \right]}} , \tag{20}$$

gdzie: $n = n_1 + n_2$, a $T = \frac{\left(t_1^3 - t_1\right)}{12}$, $t$ – liczba obserwacji wiązanych daną rangą.

Statystyka $Z$ ma w przybliżeniu rozkład normalny (Szymczak, 2010, s. 198–200).

Różnice uznaje się za statystycznie istotne, jeśli $p < \alpha$. Odrzucamy wtedy $H_0$, za prawdziwą uznajemy $H_1$. Wnioskujemy, że rozkład zmiennej zależnej różni się w porównywanych populacjach w stopniu statystycznie istotnym. Próbki pochodzą z populacji o innym rozkładzie.

W tej sytuacji kolejnym krokiem jest ustalenie, w której populacji poziom zmiennej jest istotnie wyższy, co można zrobić, porównując mediany lub średnie rangi w obu próbach.

Jeśli $p > \alpha$, to nie ma podstaw do odrzucenia $H_0$, rozkład zmiennej zależnej nie różni się istotnie w porównywanych populacjach. Próbki pochodzą z populacji o podobnych rozkładach.

Zauważmy, że test ten zaleca się stosować nie tylko w przypadku odstępstw od normalności rozkładu w przynajmniej jednej z porównywanych podpopulacji, ale również wtedy, gdy badane zjawisko mierzone jest na skali porządkowej.

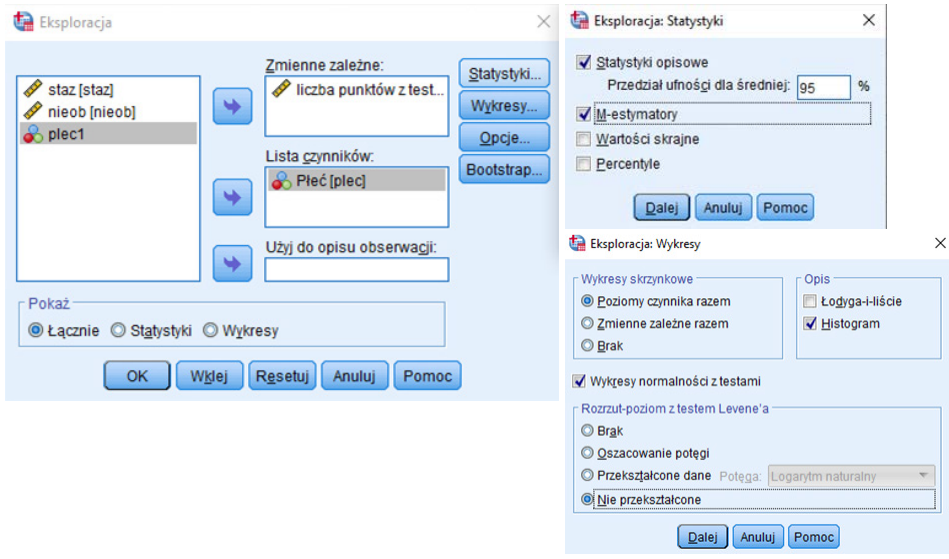
**Przykład 4.1**

Na podstawie danych dla losowo dobranej próby uczniów szkół średnich oceńmy, czy dziewczęta i chłopcy różnią się pod względem liczby punktów uzyskanych z testu wiedzy o Unii Europejskiej (zmienna *liczba punktów z testu wiedzy o UE*). Maksymalnie z testu można było uzyskać 100 punktów.

**Rozwiązanie**

Celem badania jest sprawdzenie, czy wyniki testu o UE różnią się w populacji dziewcząt i chłopców. Pośrednio chodzi o wykazanie, czy zmienna *płeć* różnicuje wyniki tego testu. W niniejszym badaniu zmienna zależna *liczba punktów z testu wiedzy o UE* jest mierzona na skali ilościowej, a próby są niezależne ($k = 2$). Najlepszym rozwiązaniem dla porównania dwóch populacji (dziewcząt i chłopców) byłoby zatem zastosowanie parametrycznego testu t-Studenta. Jednak wcześniej trzeba sprawdzić, czy spełnione są warunki stosowania tego testu. Częściowo można do tego wykorzystać polecenie *Częstości*, jednak bardziej użyteczna będzie *Eksploracja* (por. przykład 3.2).

W tym celu wybieramy *Analiza* → *Opis statystyczny* → *Eksploracja*. W oknie *Eksploracja* wskazujemy zmienną zależną, a w pole *Lista czynników* wprowadzamy *płeć*. Następnie wybieramy przycisk *Wykresy* (por. rysunek 4.1), gdzie wskazujemy *Wykresy skrzynkowe*, *Histogram*, *Wykresy normalności z testami*. Dodatkowo można z tego poziomu wykonać test Levene'a – wybieramy *Rozrzut-poziom z testem Levene'a* → *Nie przekształcone*.




**Rysunek 4.1.** Polecenie *Eksploracja* dla zmiennej *liczba punktów z testu wiedzy o UE*

Tabele wynikowe i wykresy polecenia *Eksploracja* pokazano na rysunku 4.2.

**Statystyki opisowe (DESCRIPTIVES)**

| | | | Statystyka Płeć | | Błąd standardowy Płeć | |
|---|---|---|---|---|---|---|
| | | | Dziewczyny | Chłopcy | Dziewczyny | Chłopcy |
| Liczba punktów z testu wiedzy o UE | Średnia | | 59.74 | 59.29 | 3.351 | 1.378 |
| | 95% przedział ufności dla średniej | Dolna granica | 52.84 | 56.57 | | |
| | | Górna granica | 66.64 | 62.01 | | |
| | 5% średnia obcięta | | 60.32 | 59.38 | | |
| | Mediana | | 57.46 | 59.31 | | |
| | Wariancja | | 291.914 | 282.943 | | |
| | Odchylenie standardowe | | 17.085 | 16.821 | | |
| | Minimum | | 22 | 13 | | |
| | Maksimum | | 84 | 98 | | |
| | Rozstęp | | 63 | 85 | | |
| | Rozstęp ćwiartkowy | | 29 | 19 | | |
| | Skośność | | -.163 | -.096 | .456 | .199 |
| | Kurtoza | | -.682 | .326 | .887 | .395 |

**M-estymatory**

| | Płeć | Huber | Tukey | Hampel | Andrew |
|---|---|---|---|---|---|
| Liczba punktów z testu wiedzy o UE | Dziewczyny | 59.85 | 60.04 | 60.31 | 60.03 |
| | Chłopcy | 59.25 | 59.10 | 59.38 | 59.06 |

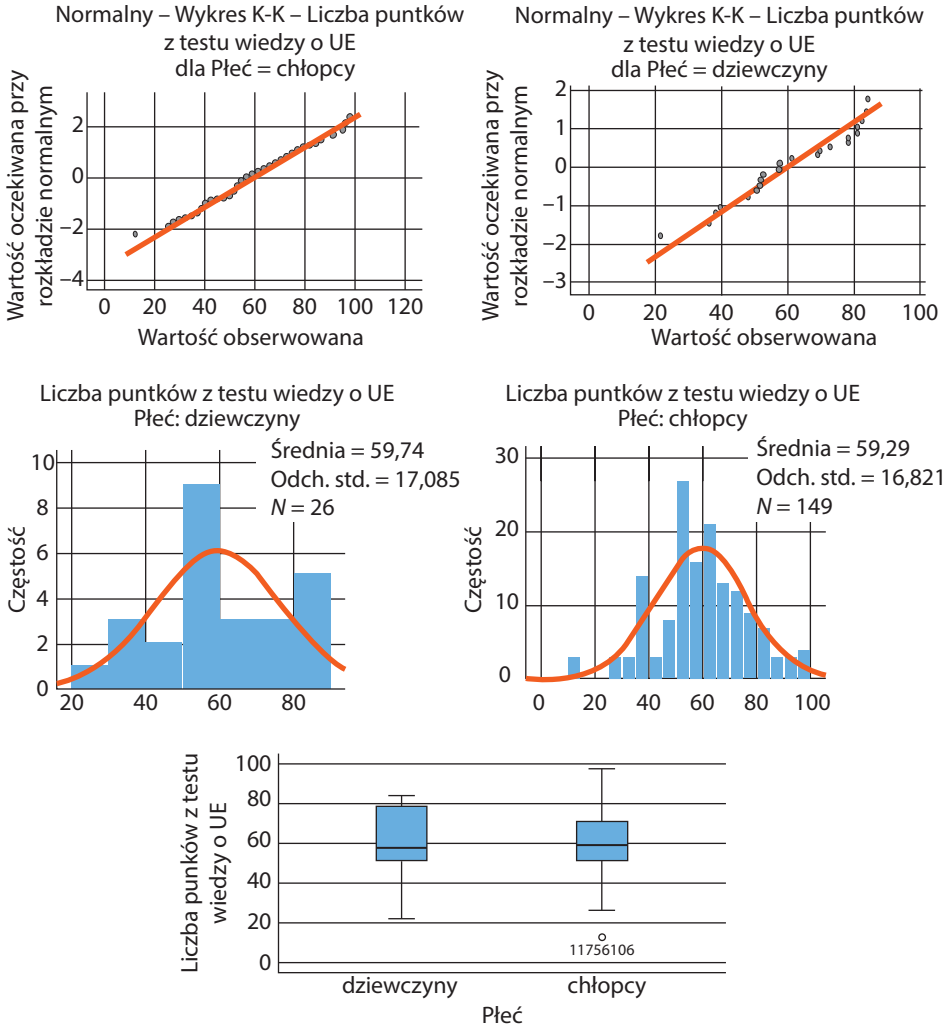
**Testy normalności rozkładu**

| | | Kołmogorow-Smirnow[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | Płeć | Statystyka | *df* | Istotność | Statystyka | *df* | Istotność |
| Liczba punktów z testu wiedzy o UE | Dziewczyny | 0.130 | 26 | .200* | .944 | 26 | .167 |
| | Chłopcy | 0.077 | 149 | .030 | .986 | 149 | .135 |

* Dolna granica rzeczywistej istotności.
[a] Z poprawką istotności Lillieforsa.

**Test jednorodności wariancji**

| | | Statystyka Levene'a | *df*1 | *df*2 | Istotność |
|---|---|---|---|---|---|
| Liczba punktów z testu wiedzy o UE | Bazując na średniej | .298 | 1 | 173 | .586 |
| | Bazując na medianie | .171 | 1 | 173 | .680 |
| | Bazując na medianie i skorygowanych *df* | .171 | 1 | 172.488 | .680 |
| | Bazując na średniej obciętej | .344 | 1 | 173 | .558 |




**Rysunek 4.2.** Wyniki *Eksploracji* dla porównania *liczby punktów z testu wiedzy o UE* według *płci*

W pierwszej kolejności należy ocenić, czy rozkład zmiennej zależnej (w obrębie wyróżnionych podpopulacji) jest zbieżny do rozkładu normalnego. Weryfikowanie normalności za pomocą testów statystycznych warto poprzedzić wstępną analizą rozkładu zmiennej ilościowej. W tym celu należy wykorzystać statystyki opisowe, można też sięgnąć po średnią obciętą, M-estymatory oraz wykresy (histogram, wykres skrzynkowe, wykres K-K) – analogicznie jak w przykładzie 3.2. Taka wstępna analiza rozkładu zmiennej – rozkładu empirycznego, bo niewychodząca jeszcze poza próbę – pozwoli na lepsze rozpoznanie badanych zjawisk i powinna stanowić etap poprzedzający zastosowanie bardziej zaawansowanych metod statystycznych (Wiktorowicz, 2017, s. 34).

Powyższe wyniki pozwalają przyjąć, że odstępstwa od normalności rozkładu są niewielkie. Średnia w grupie chłopców – 59,29 niewiele różni się od mediany – 59,31, podobnie jest w grupie dziewczyn, gdzie średnia wynosi 59,74, a mediana 57,46 punktu. Zróżnicowanie wyników w porównywanych grupach jest niewielkie. Potwierdza to współczynnik zmienności odchylenia standardowego, który wynosi 28,4% w grupie dziewczyn oraz 28,6% w grupie chłopców. Dodatkowo skośność i kurtoza osiągają wartości niewiele różniące się od tych mających miejsce przy rozkładzie normalnym. Również wykres skrzynkowy i wykres K-K potwierdzają, że w zbiorowości nie występują przypadki skrajne. Jedynie w zbiorowości chłopców ($n$ = 149, co nawet bez podania liczebności prób jesteśmy w stanie odczytać z tabeli *Testy normalności rozkładu*, gdyż *df* w obu testach odpowiada liczebności próby) obserwujemy trzy nietypowe wyniki. Wykres skrzynkowy wskazuje dodatkowo, że obie zbiorowości nie różnią się znacząco pod względem liczby punktów uzyskanych z testu wiedzy o UE.

Podstawową ocenę normalności rozkładu zapewniają testy statystyczne. Ocena skośności i kurtozy jest wtórna. Normalność rozkładu bada się zwykle testem Shapiro-Wilka, w którym hipotezy mają postać:

$H_0$: rozkład zmiennej zależnej jest normalny

$H_1$: $\neg H_0$.

W analizowanym przykładzie pierwsza próba jest mała – wynosi 26 dziewczyn. Druga próba jest duża – badaniem objęto 149 chłopców. W przypadku chłopców na mocy centralnych twierdzeń granicznych można zatem pominąć ocenę normalności rozkładu, postępując dalej tak, jak gdyby to założenie było spełnione (UWAGA: takie sformułowanie nie jest równoważne temu, że przyjmujemy/uznajemy normalność rozkładu zmiennej!). Dodatkowo przeprowadzona wcześniej wstępna analiza rozkładu zmiennej potwierdza, że nie mamy do czynienia z bardzo silną skośnością czy spłaszczeniem rozkładu – zasadniczo statystyki osiągają wartości analogiczne jak dla rozkładu normalnego (w uproszczeniu mówimy, że nie wskazują na znaczące odstępstwa od normalności rozkładu). W przypadku dziewcząt przy takiej wielkości próby wykorzystujemy test Shapiro-Wilka. W tablicach wynikowych *Eksploracji* znajdują się wyniki obu testów normalności (również testu Kołmogorowa-Smirnowa). Prawdopodobieństwo w teście Shapiro-Wilka dla dziewczyn wynosi

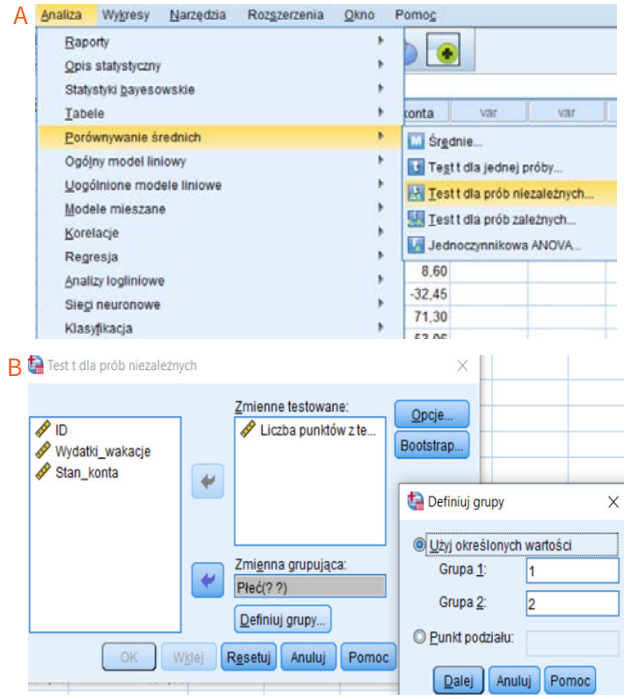
0,167 ($p$ = 0,167), nie ma zatem podstaw do odrzucenia hipotezy zerowej. Zauważmy, że dotyczy to zarówno dziewczyn, jak i chłopców ($p$ = 0,135) Rozkład zmiennej *liczba punktów z testu wiedzy o UE* w populacji dziewczyn i chłopców jest zbieżny do rozkładu normalnego (nie różni się od niego istotnie). Wyniki testu Shapiro-Wilka potwierdzają wcześniejsze oceny rozkładu zmiennej. Ponieważ pierwsze z założeń testu t-Studenta można uznać za spełnione, porównanie populacji dziewczyn i chłopców powinno być dokonane testem parametrycznym – testem t-Studenta właśnie.

W kolejnym kroku decydujemy, czy sięgniemy po jego podstawową wersję, czy też powinniśmy zastosować wersję odporną. W tym celu sprawdzimy teraz jednorodność wariancji zmiennej zależnej w obrębie porównywanych populacji – drugi warunek stosowania parametrycznego testu t-Studenta. Jednorodność wariancji weryfikujemy testem Levene'a. Hipotezy w tym teście są następujące:
$H_0$: wariancja zmiennej zależnej jest taka sama w porównywanych populacjach
$H_1$: $\neg H_0$.

Wyniki tego testu uzyskamy automatycznie, wybierając test t-Studenta. Dostępne są też (w nieco szerszym zakresie) z poziomu *Eksploracji*. Przyjrzyjmy się w pierwszej kolejności wynikom *Eksploracji*. Test Levene'a przeprowadzany jest tu na bazie trzech średnich – średniej arytmetycznej, średniej obciętej i mediany, a w ostatnim przypadku dodatkowo uwzględnia się też korektę liczby stopni swobody. Ponieważ uznaliśmy, że rozkład zmiennej nie różni się znacząco od normalnego, możemy posługiwać się średnią arytmetyczną, a zatem wybieramy test Levene'a bazujący na średniej. W teście tym $p$ = 0,586, jest więc wyższe od $\alpha$ = 0,05. Nie mamy więc podstaw do odrzucenia hipotezy zerowej, mówiącej, że wariancja zmiennej zależnej jest taka sama w porównywanych populacjach. Założenie o jednorodności wariancji można uznać za spełnione.

Do podobnych wniosków dojdziemy, korzystając z testu Levene'a bezpośrednio z poziomu tabel wynikowych testu t-Studenta. Aby je wygenerować, wybieramy kolejno: *Analiza → Porównywanie średnich → Test t dla prób niezależnych* (rysunek 4.3A), po czym pojawi się okno *test t dla prób niezależnych* (rysunek 4.3B). W oknie tym, w polu po lewej stronie, wyświetlane są wszystkie zmienne. Zmienną zależną (*liczba punktów z testu wiedzy o UE*) przenosimy w pole *Zmienne testowane*, natomiast zmienną niezależną (*płeć*) w pole *Zmienna grupująca* (rysunek 4.3B). UWAGA: nie jest możliwe wprowadzenie zmiennej grupującej, która w pliku ma typ „łańcuchowa"!

Następnie trzeba zdefiniować grupy, które chcemy porównać. Klikamy w przycisk *Porównaj grupy* (pasek *zmienna grupująca* musi być aktywny – należy kliknąć w tym miejscu tak, aby podświetlił się na żółto), pojawi się kolejne okno *Definiuj grupy* (rysunek 4.3B), w którym określamy wartości liczbowe, jakimi były kodowane poszczególne grupy. Po zdefiniowaniu grupy klikamy w *Dalej* i wracamy do poprzedniego okna *Test t dla prób niezależnych*. Na koniec wciskamy *Ok* i już pojawia się *Raport z wynikami* (rysunek 4.4).




**Rysunek 4.3.** Wykonywanie polecenia *Test t dla prób niezależnych*

Raport z wynikami testu t-Studenta jest pokazany na rysunku 4.4.

Test Levene'a prezentowany jest tu tylko w wersji *Bazując na średniej*. Tak jak z poziomu *Eksploracji*, możemy uznać założenie o jednorodności wariancji zmiennej zależnej w dwóch porównywanych populacjach za spełnione (w teście Levene'a $p = 0{,}586 > \alpha$). W tej sytuacji średnie wyników testu o UE w populacji dziewcząt i chłopców (a więc wartości oczekiwane tej zmiennej) porównujemy za pomocą podstawowej wersji testu t-Studenta. Wyniki tego testu odczytamy z pierwszego wiersza tabeli *Test t dla prób niezależnych* (*Założono równość wariancji*). W przypadku niejednorodnych wariancji korzystamy z odpornej wersji testu t-Studenta, wyniki znajdują się w drugim wierszu (*Nie założono równości wariancji*). Prawdopodobieństwo w odpowiedniej wersji testu t-Studenta odczytujemy z drugiej części drugiej tabeli (*Test t równości średnich*), z kolumny *Istotność dwustronna*.

**Statystyki dla grup**

| | Płeć | N | Średnia | Odchylenie standardowe | Błąd standardowy średniej |
|---|---|---|---|---|---|
| Liczba punktów z testu wiedzy o UE | Dziewczyny | 26 | 59.74 | 17.085 | 3.351 |
| | Chłopcy | 149 | 59.29 | 16.821 | 1.378 |

**Test dla prób niezależnych**

| | | Test Levene'a jednorodności wariancji | | Test t równości średnich | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% przedział ufności dla różnicy średnich | |
| | | F | Istotność | *t* | *df* | Istotność (dwustronna) | Różnica średnich | Błąd standardowy różnicy | Dolna granica | Górna granica |
| Liczba punktów z testu wiedzy o UE | Założono równość wariancji | .298 | .586 | .126 | 173 | .900 | .450 | 3.583 | -6.622 | 7.523 |
| | Nie założono równości wariancji | | | .124 | 34.008 | .902 | .450 | 3.623 | -6.912 | 7.813 |

**Rysunek 4.4.** Wyniki analiz testem t-Studenta dla prób niezależnych

Warto przypomnieć, że w teście t-Studenta (w obu wersjach – podstawowej i odpornej) testowane są hipotezy:

$H_0$: $\mu_1 = \mu_2$

$H_1$: $\neg H_0$.

W naszym przypadku $p = 0{,}900 > \alpha$. Nie mamy podstaw do odrzucenia hipotezy zerowej o równości wartości oczekiwanych – wartość oczekiwana wyników testu o UE dziewczyn i chłopców nie różni się istotnie. Dziewczyny w stopniu statystycznie istotnym nie różnią się od chłopców pod względem wyników z testu wiedzy o UE.

**Przykład 4.2**

Na podstawie danych dla losowo dobranej próby studentów UŁ oceńmy, czy studenci UŁ pochodzący z miast i ze wsi różnią się pod względem sytuacji finansowej, mierzonej stanem konta (w zł).

**Rozwiązanie**

Porównujemy dwie populacje – studentów UŁ pochodzących z miast i studentów UŁ pochodzących ze wsi. Zmienna zależna *stan konta* jest mierzona na skali ilościowej, próby są niezależne. Powstaje pytanie: „Czy populacje należy porównać testem parametrycznym, czy nieparametrycznym?".

Aby wybrać właściwy rodzaju testu, najpierw wstępnie analizujemy rozkład zmiennej ilościowej *stan konta* (postępujemy analogicznie jak w przykładzie 4.1).

Po pierwsze sprawdzimy, czy spełniony jest pierwszy warunek parametrycznego testu t-Studenta, czy rozkład zmiennej zależnej (w obrębie wyróżnionych podpopulacji) jest zbieżny do rozkładu normalnego, korzystając z polecenia *Eksploracja*. Tabele wynikowe i wykresy polecenia *Eksploracja* przedstawiono na rysunku 4.5. Następnie dokonamy oceny jednorodności wariancji zmiennej zależnej w porównywanych populacjach – wykorzystamy już bezpośrednio tabele wynikowe polecenia *Test t dla prób niezależnych* (rysunek 4.5).

**Informacja o analizowanych danych**

| | Miejsce zamieszkania studentów | Obserwacje Uwzględnione N | Uwzględnione Procent | Wykluczone N | Wykluczone Procent | Ogółem N | Ogółem Procent |
|---|---|---|---|---|---|---|---|
| Stan konta (zł) | Miasto | 27 | 100.0% | 0 | .0% | 27 | 100.0% |
| | Wieś | 154 | 100.0% | 0 | .0% | 154 | 100.0% |

**Statystyki opisowe (DESCRIPTIVES)**

| | | | Statystyka: Miejsce zamieszkania studentów: Miasto | Statystyka: Miejsce zamieszkania studentów: Wieś | Błąd standardowy: Miejsce zamieszkania studentów: Miasto | Błąd standardowy: Miejsce zamieszkania studentów: Wieś |
|---|---|---|---|---|---|---|
| Stan konta (zł) | Średnia | | 1430.7536 | 539.6011 | 1056.00747 | 352.62657 |
| | 95% przedział ufności dla średniej | Dolna granica | -739.9008 | -157.0446 | | |
| | | Górna granica | 3601.4080 | 1236.2467 | | |
| | 5% średnia obcięta | | 524.2484 | 66.4622 | | |
| | Mediana | | 332.8575 | 6.3317 | | |
| | Wariancja | | 30109097.988 | 19149207.196 | | |
| | Odchylenie standardowe | | 5487.17578 | 4375.98071 | | |
| | Minimum | | -3365.57 | -12320.58 | | |
| | Maksimum | | 28552.50 | 37239.00 | | |
| | Rozstęp | | 31918.07 | 49559.58 | | |
| | Rozstęp ćwiartkowy | | 396.39 | 146.31 | | |
| | Skośność | | 4.982 | 5.312 | .448 | .195 |
| | Kurtoza | | 25.575 | 39.813 | .872 | .389 |

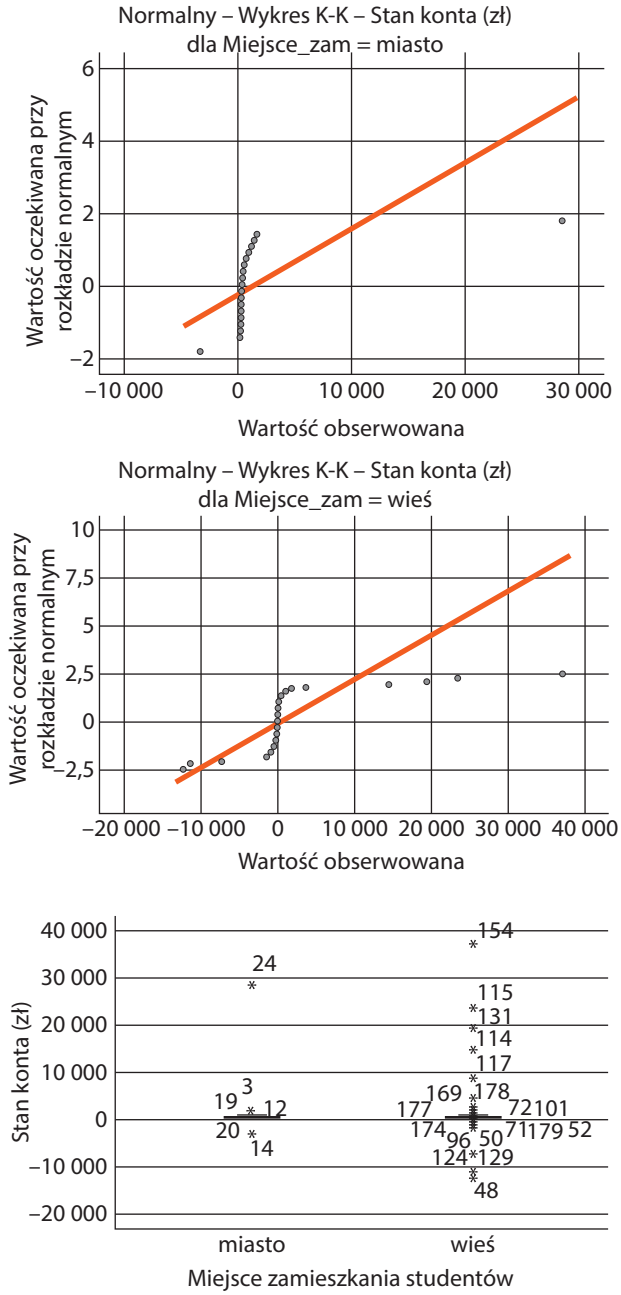
**M-estymatory**

| | Miejsce zamieszkania studentów | Huber | Tukey | Hampel | Andrew |
|---|---|---|---|---|---|
| Stan konta (zł) | Miasto | 356.3427 | 322.5025 | 330.2076 | 322.5085 |
| | Wieś | 21.0989 | 10.8063 | 18.1186 | 10.6790 |

**Testy normalności rozkładu**

| | Miejsce zamieszkania studentów | Kołmogorow-Smirnow[a] Statystyka | Kołmogorow-Smirnow[a] *df* | Kołmogorow-Smirnow[a] Istotność | Shapiro-Wilk Statystyka | Shapiro-Wilk *df* | Shapiro-Wilk Istotność |
|---|---|---|---|---|---|---|---|
| Stan konta (zł) | Miasto | .438 | 27 | .000 | .291 | 27 | .000 |
| | Wieś | .412 | 154 | .000 | .308 | 154 | .000 |

[a] Z poprawką istotności Lillieforsa.




**Rysunek 4.5.** Wyniki polecenia *Eksploracja* dla zmiennej *stan konta*

Uzyskane wyniki pokazują, że mamy istotne w sensie statystycznym odstępstwa od normalności rozkładu. Prawdopodobieństwo w teście Shapiro-Wilka jest dla obu populacji bliskie zera ($p < 0{,}001$), należy zatem odrzucić hipotezę o normalności rozkładu

zmiennej – zarówno dla studentów z miast, jak i studentów ze wsi. W przypadku oceny normalności rozkładu na podstawie statystyk opisowych zwracamy szczególną uwagę, oprócz miar symetrii rozkładu i koncentracji (skośności i kurtozy), na wartości M-estymatorów. W grupie studentów z miast M-estymatory wynoszą od 322,5 do 356,3 zł, wobec średniej 1430,7 zł, mediany 332,9 zł, średniej obciętej 524,2 zł. Różnice są zatem duże. Rozkład stanu konta studentów z miast charakteryzuje się także silną asymetrią, o czym świadczy wysoka wartość klasycznego współczynnika asymetrii (4,98). Podobna sytuacja ma miejsce w próbie studentów ze wsi. Wartości M-estymatorów (10,7–21,1 zł) zdecydowanie odbiegają od średniej (539,6 zł) i średniej obciętej (66 zł), a wysoka wartość współczynnika skośności (5,3) świadczy o silnej asymetrii rozkładu.

Na wykresie skrzynkowym (rysunek 4.5) nie jest nawet widoczna charakterystyczna dla tego rodzaju wykresu skrzynka. Liczne gwiazdki na wykresie wskazują, że występuje wiele obserwacji oddalonych od skrzynki o więcej niż 3 rozstępy ćwiartkowe. Wyraźnie sugeruje to, że odstępstwa od normalności rozkładu są w przypadku naszej zmiennej bardzo duże. Wyniki testu normalności Shapiro-Wilka oraz analiza statystyk opisowych i wykresów potwierdzają zatem, że w obu podpopulacjach występują istotne odstępstwa od rozkładu normalnego. Dodatkowo jedna z prób liczy tylko 27 elementów. Dlatego właściwą metodą jest w tej sytuacji test nieparametryczny – test Manna-Whitneya.

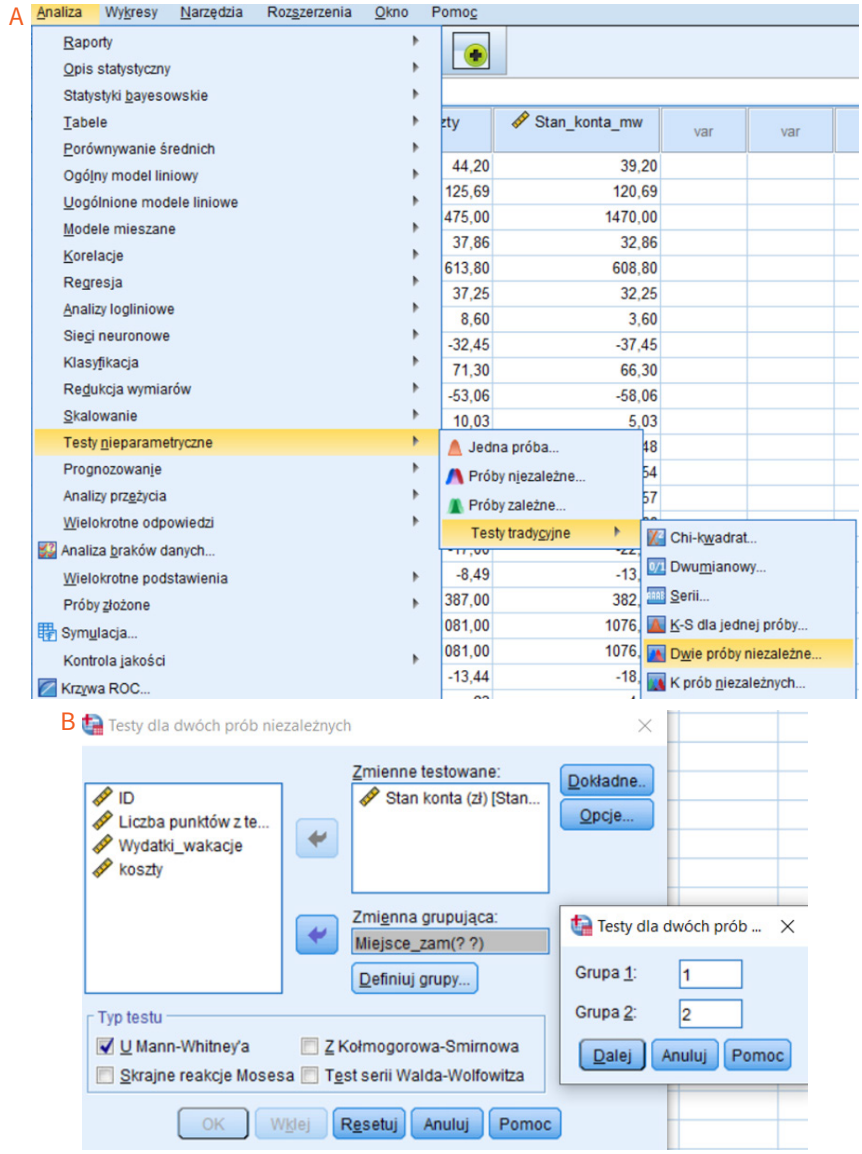
Hipotezy teście Manna-Whitneya są następujące:

$H_0$: dwie niezależne próbki pochodzą z populacji o takim samym rozkładzie

$H_1$: $\neg H_0$ (nieprawda, że dwie niezależne próbki pochodzą z populacji o takim samym rozkładzie).

W IBM SPSS Statistics test Manna-Whitneya uzyskać możemy na dwa sposoby.

Pierwszy z nich jest następujący: wybieramy *Analiza* → *Testy nieparametryczne* → *Testy tradycyjne* → *Dwie próby niezależne* (rysunek 4.6A). Po wyborze *Dwie próby niezależne* pojawia się okno *Testy dla dwóch prób niezależnych* (rysunek 4.6B), w którym w pole *Zmienne testowane* przenosimy zmienną *stan konta*, natomiast w polu *Zmienna grupująca* umieszczamy zmienną *miejsce zamieszkania*. Następnie trzeba zdefiniować grupy, które chcemy porównać. Klikamy w przycisk *Definiuj grupy* – pojawi się okno *Testy dla dwóch prób niezależnych: Definiuj grupy* (rysunek 4.6B). Tutaj, podobnie jak przy testach dla prób niezależnych, wpisujemy wartości liczbowe, które przypisaliśmy do poszczególnych grup. Nie jest możliwe wprowadzenie w tym miejscu oznaczeń tekstowych (zmienne typu łańcuchowego nie wyświetlają się w ogóle na liście dostępnych zmiennych). Po określeniu grup klikamy w *Dalej* i wracamy do poprzedniego okna *Testy dla dwóch prób niezależnych*. Na koniec musimy jeszcze w dolnej ramce zatytułowanej *Typ testu* zaznaczyć, który z testów chcemy wykorzystać. Zaznaczamy wariant *U Manna-Whitneya* i zatwierdzamy przyciskiem *Ok*. W efekcie pojawia się Edytor raportów z wynikami testu.




**Rysunek 4.6.** Wykonywanie polecenia *Testy nieparametryczne* → *Dwie próby niezależne* (ścieżka 1)

W Edytorze raportów pojawiają się dwie tabele: *Rangi* i *Wartość testowana* (rysunek 4.7).

**Rangi**

| | Miejsce zamieszkania studentów | N | Średnia ranga | Suma rang |
|---|---|---|---|---|
| Stan konta (zł) | Miasto | 27 | 145.11 | 3918.00 |
| | Wieś | 154 | 81.51 | 12553.00 |
| | Ogółem | 181 | | |

**Wartość testowana[a]**

| | Stan konta (zł) |
|---|---|
| U Manna-Whitneya | 618.000 |
| W Wilcoxona | 12553.000 |
| Z | –5.818 |
| Istotność asymptotyczna (dwustronna) | .000 |

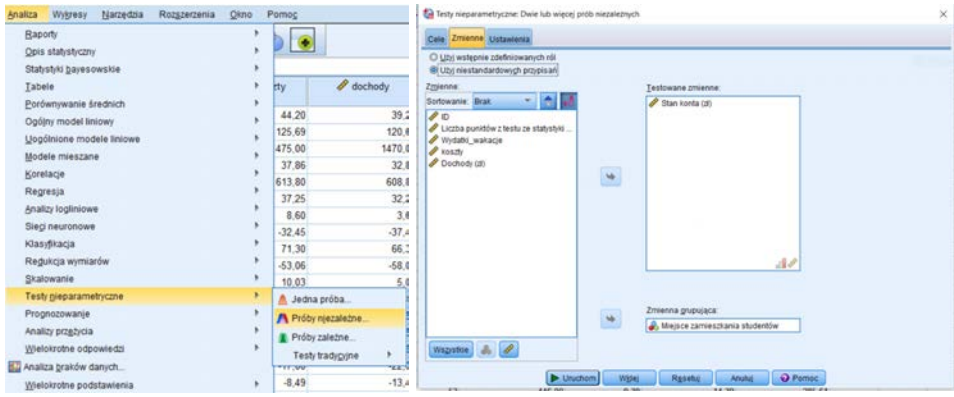
[a] Zmienna grupująca *miejsce zamieszkania studentów*.

**Rysunek 4.7.** Wyniki analiz testem Manna-Whitneya: porównanie *stanu konta* według *miejsca zamieszkania studentów* (ścieżka 1)

W tabeli *Rangi* (rysunek 4.7) znajdują się wartości średnich rang (odpowiedniki średnich arytmetycznych w teście t-Studenta) oraz sumy rang dla każdej grupy (miasto, wieś). Średnia ranga jest wyższa w grupie miasto. Z kolei suma rang jest wyższa w grupie wieś. Ta odwrotna relacja, którą obserwujemy, wynika z różnej liczebności porównywanych grup (miasto 27, wieś 154). Interpretując wyniki, zwłaszcza w przypadku różnej liczebności grup, odwołujemy się przede wszystkim do średnich rang.

Oceniając istotność różnic między populacjami, korzystamy z tabeli *Wartość testowana* (rysunek 4.7), gdzie zwracamy uwagę na wartość *U Manna-Whitneya* (618,000) oraz *istotność asymptotyczną* (prawdopodobieństwo w teście). Prawdopodobieństwo w teście Manna-Whitneya $p < 0{,}001$, a więc $p < \alpha$. Na poziomie istotności $\alpha = 0{,}05$ odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej – obie populacje (studenci z miast i ze wsi) różnią się zatem w stopniu statystycznie istotnym ze względu na stan konta. Porównując średnie rangi, można wskazać, że wyższym stanem konta dysponują studenci z miast.

Druga ścieżka jest w przypadku tego testu następująca: wybieramy *Analiza* → *Testy nieparametryczne* → *Próby niezależne*, po czym pojawia się okno *Testy nieparametryczne. Dwie lub więcej prób niezależnych*. W oknie tym w polu *Testowana zmienna* umieszczamy zmienną zależną *stan konta*, a w polu *Zmienna grupująca*: *miejsce zamieszkania studentów*. Zatwierdzamy wybór poprzez wciśnięcie *Uruchom* (rysunek 4.8).




**Rysunek 4.8.** Wykonywanie testu Manna-Whitneya (ścieżka 2)

Po zatwierdzeniu wyboru pojawia się w Edytorze raportów tabela wynikowa – *Podsumowanie testu hipotezy* (rysunek 4.9A). Klikając dwukrotnie w tę tabelę, przechodzimy do *Przeglądu modelu* (rysunek 4.9B).

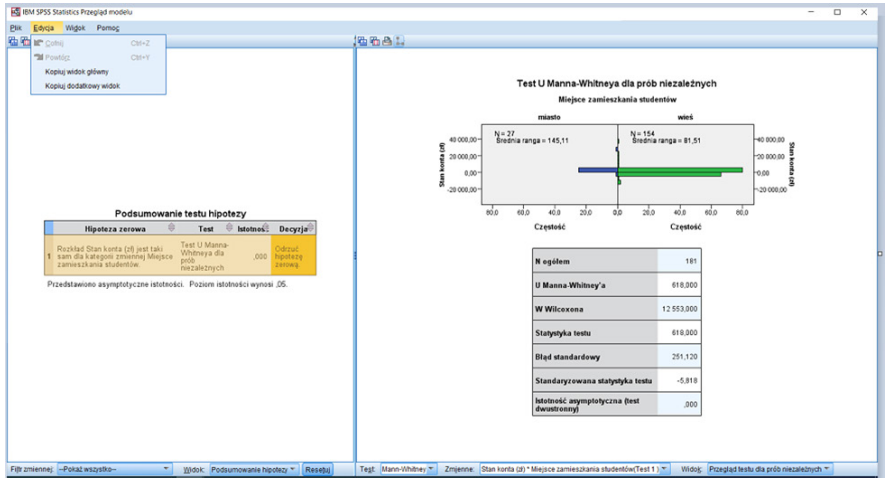
A

**Podsumowanie testu hipotezy**

| | Hipoteza zerowa | Test | Istotność | Decyzja |
|---|---|---|---|---|
| 1 | Rozkład Stan konta (zł) jest taki sam dla kategorii zmiennej Miejsce zamieszkania studentów. | Test U Manna-Whitneya dla prób niezależnych | ,000 | Odrzuć hipotezę zerową. |

Przedstawiono asymptotyczne istotności. Poziom istotności wynosi ,05.

B




**Rysunek 4.9.** Wyniki analiz testem Manna-Whitneya (ścieżka 2)

*Przegląd modelu* to okno, w którym po lewej stronie pojawia się ponownie tabela wynikowa *Podsumowanie testu hipotezy*. Z kolei po prawej stronie prezentowane

są średnie rangi i histogramy dla obu grup. Niżej znajduje się zestawienie, które pokazuje między innymi wartość *U* Manna-Whitneya (618,000), prawdopodobieństwo w teście $p < 0{,}001$ (istotność asymptotyczna), wartość *W* Wilcoxona (12553,000) oraz wartość statystyki *Z* (standaryzowana statystyka testu).

Wyniki te prowadzą rzecz jasna do tych samych wniosków, jakie przytaczaliśmy wcześniej – różnice między grupami uznamy za istotne statystycznie, hipotezę zerową należy bowiem odrzucić (por. *Podsumowanie testu hipotezy*) na rzecz hipotezy alternatywnej.

Na koniec warto nadmienić, że zawartość prawej części okna *Przegląd modelu* można kopiować – wybieramy *Edycja → Kopiuj dodatkowy widok* (obwiedzione pomarańczową linią na rysunku 4.9).

# 5. Porównanie więcej niż dwóch populacji

**Kluczowe pojęcia:** porównanie więcej niż dwóch populacji, próby niezależne, rozkład normalny, jednorodność wariancji, testy parametryczne i nieparametryczne, istotność różnic, test F, test Browna-Forsythe'a, test Welcha, test Levene'a, test Kruskala-Wallisa, siła efektu

## 5.1. Uwagi wstępne

Przy porównywaniu więcej niż dwóch populacji (a dokładniej – przynajmniej dwóch) dobieramy właściwą metodę analizy, kierując się analogicznymi przesłankami jak w przypadku porównania dwóch zbiorowości (rozdział czwarty). Najczęściej stosowaną w takiej sytuacji metodą jest analiza wariancji (ANOVA – *ANalysis Of VAriance*). Metoda ta jest wykorzystywana do rozstrzygania o istnieniu różnic między średnimi w kilku (dwóch lub więcej) populacjach. Jest to metoda parametryczna, służąca do porównania wartości oczekiwanych w kilku populacjach. Pod hasłem „analiza wariancji" kryje się *de facto* cała grupa metod pozwalających na dokonywanie tego rodzaju rozstrzygnięć, tymczasem my zajmiemy się najprostszą jej odmianą, tj. jednoczynnikową analizą wariancji (*One-Way ANOVA*). Pod uwagę weźmiemy przy tym jedynie analizę w schemacie międzygrupowym, a więc porównywać będziemy niezależne populacje.

Jednoczynnikowa analiza wariancji ma zastosowanie, gdy zmienna zależna mierzona jest na skali ilościowej[7], a zmienna niezależna jest zmienną jakościową/

7 W praktyce badawczej uznaje się, że pewne zmienne mierzone na skali porządkowej można niekiedy potraktować jak ilościowe. Dotyczy to na przykład zmiennych, dla których dane zbiera się za pomocą skal numerycznych o równo wyglądających interwałach, np. „Na skali od 0 do 10, gdzie 0 oznacza […], a 10 oznacza […], wskaż punkt, który odpowiada Twojej opinii w tym względzie" (zob. Agresti, Finlay, 2014: 13). Podobnie podchodzi się do zmiennych mierzonych na skali Likerta (przy spełnieniu określonych warunków, w tym: powinna być przynajmniej pięciostopniowa, rozkład odpowiedzi nie powinien być silnie symetryczny;

dyskretną (mierzoną na skali nominalnej lub porządkowej). W przyjętej w analizie wariancji terminologii zmienną niezależną nazywa się czynnikiem, a jej wartości poziomami. Konwencja ta odzwierciedla fakt, że ANOVA jest podstawową metodą analizy danych w badaniach eksperymentalnych.

Podajmy kilka przykładów pytań badawczych: „Czy średnie zarobków wśród osób z wykształceniem podstawowym, zasadniczym zawodowym, średnim, wyższym różnią się od siebie?", „Czy średnia liczba minut spędzonych dziennie w Internecie różni się w poszczególnych grupach wieku (18–24, 25–34, 35–50, 51 i więcej lat)?", „Czy metoda nauczania (A, B, C) różnicuje liczbę punktów uzyskanych w teście wiedzy?". W ostatnim przypadku zakładamy dodatkowo, że o tym, do której grupy dostała się każda z badanych osób, zdecydował mechanizm losowy.

W pierwszym przykładzie czynnikiem, czyli zmienną, która utworzyła grupy (dokładnie cztery), jest wykształcenie. Inaczej można powiedzieć, że czynnik wykształcenie występuje na czterech poziomach. W drugim przykładzie czynnikiem jest grupa wieku i zmienna ta również występuje na czterech poziomach. W ostatnim przykładzie czynnikiem o trzech poziomach jest metoda nauczania. W każdym z przykładów rozważana jest sytuacja, w której do objaśnienia zmienności zmiennej zależnej wykorzystuje się tylko jeden czynnik (stąd określenie „jednoczynnikowa").

Analizę wariancji można wykorzystywać do danych pochodzących nie tylko z badań eksperymentalnych, ale także obserwacyjnych, w tym sondażowych. Dwa pierwsze pytania dotyczą badań obserwacyjnych, trzecie – badań eksperymentalnych. Procedura obliczeniowa przebiega w ten sam sposób dla jednego i drugiego rodzaju danych. Typ danych ma znaczenie na poziomie interpretacji wyników. Jak podkreślano w rozdziale drugim, wnioski z badań eksperymentalnych można formułować w kategoriach przyczynowo-skutkowych, a zatem różnice między średnimi tłumaczyć można wpływem czynnika. Taka interpretacja jest uzasadniona, ponieważ każda z badanych jednostek jest losowo przypisana do jednej z porównywanych grup, w efekcie grupy różnią się – odwołując się do przykładu – jedynie metodą nauczania, jaką w nich zastosowano. W badaniach sondażowych o przynależności badanego do grupy, na przykład danej kategorii miejsca zamieszkania, nie decyduje mechanizm losowy, dlatego różnice w średnich poziomach badanej cechy między mieszkańcami – dajmy

zaleca się też nieparzystą liczbę wariantów, z uwzględnieniem tzw. środka skali – por. np. Olsson, 1979; Borgatta, Bohrnstedt, 1980; Lubke, Muthen, 2004; Wiktorowicz, 2016). W konsekwencji można spotkać w literaturze, w tym w podręcznikach, przykłady zastosowań analizy wariancji na zmiennych porządkowych potraktowanych jako quasi-ilościowe, o ograniczonych z dołu i z góry zakresach oraz dość wąskiej rozpiętości skali. Praktykuje się to przy dużych próbach (Rószkiewicz, 2011, s. 123; Agresti, Finlay, 2014, s. 371).

na to – wsi, małych miast i dużych miast, nie mogą być tłumaczone wpływem wielkości miejsca zamieszkania. Określenia *wpływ* czy *przyczyna* są tu za mocne. Mieszkańcy wsi, małych miast i dużych miast różnią się przykładowo pod względem zarobków, co nie oznacza, że miejsce zamieszkania determinuje poziom tych zarobków.

Analiza wariancji nie jest jedyną metodą wykorzystywaną przy porównaniu więcej niż dwóch populacji. Jak podkreślano, wymaga ona ilościowego pomiaru zmiennej zależnej. Co w takim razie zrobić, gdy poziom pomiaru jest niemetryczny – porządkowy lub nominalny? Należy wówczas sięgnąć po testy nieparametryczne – test Kruskala-Wallisa dla zmiennych mierzonych na skali porządkowej (będzie o nim mowa w dalszej części tego rozdziału) lub po test niezależności chi-kwadrat (który wykorzystuje się również do badania zależności między zmiennymi jakościowymi – zostanie on omówiony w rozdziale szóstym).

## 5.2. Porównanie średnich w populacjach

### Test F

W analizie wariancji wykorzystywany jest test F. Jest to test parametryczny, który – podobnie jak test t-Studenta – wymaga spełnienia założeń dotyczących rozkładu zmiennej zależnej w porównywanych populacjach:

- Zmienna zależna powinna mieć rozkład normalny w każdej porównywanej populacji. Niemniej jednak wraz ze zwiększaniem się liczebności próby rozkład statystyki *F* mniej zależy od rozkładu cechy w populacji. Ocenę rozkładów pod kątem zaburzeń normalności powinniśmy przeprowadzać zwłaszcza w sytuacji, gdy podpróby są małe (Agresti, Finlay, 2014, s. 401)[8]. Ocena ta przebiega zgodnie z procedurami opisanymi w poprzednich rozdziałach. Należy pamiętać, że wynik ANOVA może być niewiarygodny, gdy obserwujemy bardzo dużą skośność. Duża skośność stanowi zresztą problem z fundamentalnego powodu – przy znaczącej asymetrii średnia przestaje być dobrą charakterystyką rozkładu (Agresti, Franklin, 2013, s. 688).
- Wariancja zmiennej zależnej w porównywanych podpopulacjach (tj. rozproszenie wyników wokół średniej powinno być bardzo podobne w każdej

8 Jak piszą Keppel i Wickens (2004, s. 145) w odniesieniu do analizy wariancji: „[…] jeżeli próba liczy co najmniej kilkanaście przypadków, nie musimy martwić się o spełnienie założenia o normalności rozkładu". Z dyskusją dotyczącą założenia normalności rozkładów w analizie wariancji Czytelnik może się zapoznać w podręczniku Szymczaka (2018, s. 354–357).

podpopulacji) powinna być jednorodna (homogeniczna). Założenie to oceniamy za pomocą testu Levene'a (por. rozdział czwarty)[9].

- Próby powinny być niezależne i pobrane losowo, a w przypadku badań eksperymentalnych przyporządkowanie jednostek do grup powinno odbywać się na zasadzie losowej (poprzez randomizację).

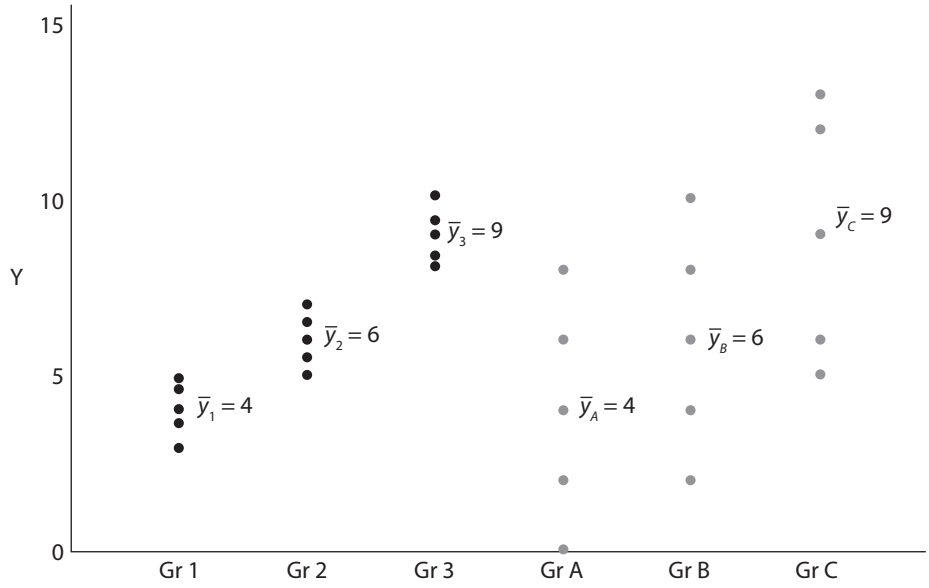
W analizie wariancji występuje następujący układ hipotez:
$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$ (wszystkie wartości oczekiwane w podpopulacjach są równe)
$H_1$: $\neg H_0$ (co najmniej dwie wartości oczekiwane nie są równe).

Aby lepiej zrozumieć logikę analizy wariancji oraz konstrukcję statystyki *F*, stanowiącej sprawdzian testu, odwołajmy się do dwóch hipotetycznych zestawów danych. Dane te zostały przedstawione na rysunku 5.1. Pierwszy zestaw stanowią dane dla grup 1, 2 i 3 (punkty zaznaczone czarnym kolorem). Przyjmijmy, że grupy te utworzył czynnik X. Drugi zestaw stanowią dane dla grup A, B i C (punkty zaznaczone szarym kolorem). Przyjmijmy, że grupy te utworzył czynnik Z.

Dane w każdym zestawie możemy przeanalizować pod kątem występowania różnic między grupami (wariancji międzygrupowej) oraz pod kątem występowania różnic wewnątrz grup (wariancji wewnątrzgrupowej) (Bedyńska, Cypryańska, 2013b, s. 15). Porównując teraz oba zestawy, zauważamy, że mają one tę samą wariancję międzygrupową (w przypadku obu zestawów średnie grupowe, oznaczone przez $\overline{y}$, wynoszą 4, 6 i 9), ale wariancja wewnątrzgrupowa jest większa w drugim zestawie (większe rozproszenie wyników wokół średnich grupowych) niż w pierwszym. Zastanówmy się teraz, który czynnik – X czy Z – lepiej objaśnia zmienność wyników. Czynnik, który dobrze objaśnia zmienność zmiennej zależnej, to taki, który powoduje, że obserwujemy dużą wariancję międzygrupową i jednocześnie małą wariancję wewnątrzgrupową. Gdyby wariancja wewnątrzgrupowa była zerowa, to wiedząc, do której grupy należy badana jednostka i znając średnią grupową, bez błędu odgadlibyśmy jej wynik. O takim czynniku powiedzielibyśmy, że objaśnia 100% zmienności wyników zmiennej zależnej. Przy dużej wariancji wewnątrzgrupowej (a więc różnic indywidualnych między jednostkami w poszczególnych grupach) znajomość grupy, do której należy badana osoba, i wiedza

9 Rozważając otrzymany wynik, pamiętajmy, że im większe podpróby, tym łatwiej o odrzucenie hipotezy zerowej, stąd – gdy mamy do czynienia z dużymi podporóbami – do oceny rozproszenia wyników w porównywanych grupach możemy wziąć pod uwagę wielkości odchyleń standardowych. Jak podają Agresti i Finlay (2014, s. 401), przy nierównolicznych podpróbach wynik testu F możemy uznać za wiarygodny, o ile stosunek największego odchylenia standardowego do najmniejszego nie jest większy niż 2. Jeżeli podpróby są równoliczne, test F jest odporny na złamanie drugiego i trzeciego założenia (Agresti, Franklin, 2013, s. 687–688). O równoliczność grup badacze mają możliwość zadbać w badaniach eksperymentalnych. Jednak w badaniu sondażowym realizowanym na próbie losowej, która stanowi przekrój populacji, jest to niemożliwe.

o średniej grupowej nie wystarczą, by bezbłędnie przewidzieć wynik tej osoby. Oznacza to, że zmienna zależna podlega dodatkowemu oddziaływaniu ze strony innych czynników, których nie uwzględniliśmy w badaniu. O wariancji międzygrupowej mówimy, że jest wariancją wyjaśnioną, bo jest to zmienność wynikająca z działania badanego czynnika, a o wariancji wewnątrzgrupowej mówimy, że jest wariancją niewyjaśnioną (wariancją błędu), bo jest to zmienność wynikająca z działania innych czynników. Stąd właśnie ta metoda, która rozstrzyga o istnieniu różnic między średnimi, nazywa się – na pozór myląco – analizą wariancji. Statystyka *F* zestawia ze sobą te dwa rodzaje zmienności:




**Rysunek 5.1.** Ilustracja idei analizy wariancji

**Źródło:** opracowanie własne.

$$F = \frac{\text{zróżnicowanie między grupami}}{\text{zróżnicowanie wewnątrz grup}}. \tag{21}$$

$H_0$ odrzucamy (i wnioskujemy, że różnica między średnimi okazuje się istotna statystycznie), jeśli zróżnicowanie między grupami jest większe niż zróżnicowanie wewnątrz grup, a zatem przy określonej liczbie stopni swobody ($df_1$ i $df_2$) wartość statystyki *F* jest odpowiednio duża, z pewnością większa od 1.

Od strony obliczeniowej statystyka *F* wygląda następująco (Agresti, Finlay, 2014, s. 372–374):

$$F = \frac{MS_{międzygr.}}{MS_{wewnątrzgr.}} = \frac{\frac{SS_{międzygr.}}{df_1}}{\frac{SS_{wewnątrzgr.}}{df_2}} = \frac{\frac{n_1\left(\overline{y}_1 - \overline{y}\right)^2 + \ldots + n_k\left(\overline{y}_k - \overline{y}\right)^2}{k-1}}{\frac{\left(n_1 - 1\right)s_1^2 + \ldots + \left(n_k - 1\right)s_k^2}{n-k}}, \tag{22}$$

gdzie: $n$ to całkowita liczebność próby, $k$ – liczba grup, $MS$ – wariancja (średni kwadrat odchyleń od średniej), $SS$ – suma kwadratów odchyleń od średniej, $s_1^2$ – wariancja w pierwszej próbie, którą obliczamy zgodnie ze wzorem:

$$s_1^2 = \frac{\Sigma\left(y - \overline{y}_1\right)^2}{n_1 - 1}, \tag{23}$$

i analogicznie do tego wariancje w pozostałych grupach.

W przypadku gdy prawdziwa jest hipoteza zerowa, statystyka $F$ ma rozkład $F$ o $df_1 = (k - 1)$ i $df_2 = (n - k)$ stopniach swobody.

Przyjmując $\alpha = 0{,}05$, wnioskowanie przeprowadzamy według reguły:

- jeżeli $p < 0{,}05$, to stwierdzamy, że są podstawy do odrzucenia hipotezy zerowej i przyjęcia hipotezy alternatywnej; istotny wynik testu $F$ uprawnia nas do uznania, że przynajmniej dwie wartości oczekiwane w podpopulacjach różnią się od siebie; w następnym kroku podejmuje się analizy mające na celu uzyskanie odpowiedzi na pytanie, między którymi wartościami oczekiwanymi te różnice występują; służą do tego testy *post hoc*;
- jeżeli $p > 0{,}05$, to stwierdzamy, że brak jest podstaw do odrzucenia hipotezy zerowej – różnice między wartościami oczekiwanymi w porównywanych podpopulacjach nie są statystycznie istotne.

## Wielkość efektu

Gdybyśmy chcieli ustalić wielkość efektu, możemy w tym celu wykorzystać miernik $\eta^2$ (czytaj: eta kwadrat). Wyznaczamy go zgodnie ze wzorem:

$$\eta^2 = \frac{SS_{międzygr.}}{SS_{całkowita}} = \frac{SS_{międzygr.}}{SS_{międzygr.} + SS_{wewnątrzgr.}}. \tag{24}$$

IBM SPSS Statistics nie raportuje tej miary na poziomie procedury *Jednoczynnikowa ANOVA*, ale ustalenie jej wielkości – jak widać ze wzoru – jest bardzo proste (możliwe jest też wyznaczenie cząstkowego eta kwadrat, korzystając z polecenia

*Ogólny model liniowy → Jednej zmiennej → Opcje → Ocena wielkości efektu*; dla modelu jednoczynnikowego cząstkowe eta-kwadrat ma taką samą wartość jak eta kwadrat).

Eta kwadrat informuje, jaką część zmienności zmiennej zależnej wyjaśnia czynnik. Przyjmuje wartości z zakresu [0; 1] (Gamst, Meyers, Guarino, 2008, s. 42). Może być stosowany zarówno wtedy, gdy próby są równoliczne, jak i wtedy, gdy nie są równoliczne[10].

Cohen (1988, s. 286–287) zaproponował następującą interpretację wartości $\eta^2$:
$\eta^2 = 0,0099 \approx 0,01$ – mała wielkość efektu,
$\eta^2 = 0,0588 \approx 0,06$ – średnia wielkość efektu,
$\eta^2 = 0,1379 \approx 0,14$ – duża wielkość efektu.

Rzecz jasna, ocena ta nie jest punktowa. Jeśli $\eta^2$ jest mniejsze od 0,01, przyjmuje się, że mamy do czynienia z brakiem efektu czynnika, a podczas gdy $\eta^2$ jest większe od 0,14, efekt uznaje się za silny.

## Procedura porównań wielokrotnych (testy *post hoc*)[11]

Do wykonywania porównań wielokrotnych (każda średnia w populacji z każdą) nie możemy posłużyć się testem t-Studenta. Przeprowadzenie wielu porównań testem t-Studenta grozi skumulowaniem się błędu I rodzaju ponad założony poziom $\alpha$[12].

IBM SPSS Statistics udostępnia całą gamę testów, które biorą poprawkę na liczbę porównań, tak aby prawdopodobieństwo błędu I rodzaju nie było większe od zakładanego poziomu $\alpha$. Testy te są polecane dla jednego z dwóch wariantów – inne zastosujemy, gdy uznamy założenie o jednorodności wariancji zmiennej

10 Dodajmy, że $\eta^2$ sprawdza się jako miara efektu w odniesieniu do danych z próby, natomiast jako miara efektu w populacji jest estymatorem nieznacznie obciążonym (Howell, 2010, s. 346).

11 W podręczniku tym zostanie omówiona jedynie procedura porównań wielokrotnych (z wykorzystaniem testów *post hoc*), którą należy odróżnić od procedury porównań (kontrastów) *a priori*. Porównania *a priori* wykorzystuje się przy konfirmacyjnym podejściu badawczym, a więc wtedy, gdy na podstawie literatury sformułowana została hipoteza mówiąca, jakiego konkretnie układu średnich należy się spodziewać. Jednocześnie porównania *a priori* są alternatywą dla testu *F* (nie wymagają jego przeprowadzenia). Czytelników zainteresowanych tym rodzajem porównań zachęcamy do lektury podręcznika Bedyńskiej i Cypryańskiej (2013b). Dodatkowo wśród testów *post hoc* oprócz wymienionych testów wielokrotnych porównań parami proponuje się też testy rozstępu. W ich przypadku szukamy populacji podobnych – na poziomie istotności *a* wskazujemy jednorodne populacje. W tej wersji (obok porównań wielokrotnych) opracowany został na przykład test Tukeya czy Scheffego.

12 Liczbę porównań wyznaczamy za pomocą wzoru (Agresti, Finlay, 2014, s. 377): $c = k(k-1)/2$. Przyjmując $a = 0,05$, wielkość tego błędu wyznaczymy za pomocą wzoru: Skumulowany błąd I rodzaju = $1 - (0,95)^c$ (Field, 2009, s. 349).

zależnej za spełnione (będzie to np. test Tukeya, Bonferroniego, Sidaka, Scheffego), inne – gdy założenie to nie będzie spełnione (wówczas sięgniemy np. po test T2 Tamhane'a, test Gamesa-Howella). Poszczególne testy różnią się też stopniem konserwatyzmu (przy teście konserwatywnym trudniej jest odrzucić hipotezę zerową niż przy teście liberalnym) czy możliwością zastosowania w zależności od tego, czy grupy są równoliczne, czy nie[13]. W wielu podręcznikach statystycznych wykorzystuje się lub wręcz poleca test Tukeya (np. Wieczorkowska, Wierzbiński, 2007; Szwed, 2008; Agresti, Finlay, 2014). Test ten należy do grupy konserwatywnych, a więc dobrze kontrolujących błąd I rodzaju.

## Test Browna-Forsythe'a i test Welcha

Jeżeli spełnione jest założenie o normalności rozkładu zmiennej zależnej w podpopulacjach, ale rozproszenie wyników wokół średnich jest wyraźnie różne (nie jest spełnione założenie o jednorodności wariancji), to zamiast testem F lepiej posłużyć się testem Browna-Forsythe'a lub testem Welcha. W niniejszej publikacji przedstawimy konstrukcję statystyki testowej tylko dla pierwszego z tych testów – jest ona następująca (Reed, Stark, 1988):

$$F_{BF} = \frac{SS_{międzygr.}}{S_1^2\left(1-\frac{n_1}{n}\right)+\ldots S_k^2\left(1-\frac{n_k}{n}\right)}. \tag{25}$$

Wyrażenie w mianowniku pokazuje pomysł na przezwyciężenie problemu nierównych wariancji – każdej wariancji nadaje się odpowiednią wagę, która jest adekwatna do liczebności podpróby, dla której została wyznaczona.

W przypadku gdy prawdziwa jest hipoteza zerowa, rozkład statystyki $F_{BR}$ jest dobrze przybliżony przez rozkład $F$ o $df_1 = (k - 1)$ i $df_2 = f'$ stopniach swobody, gdzie:

$$f' = \left[\sum_i \frac{c_i^2}{n_i - 1}\right]^{-1} \text{ oraz } c_i = \frac{\left(1-\frac{n_i}{n}\right)s_i^2}{\sum_i\left(1-\frac{n_i}{n}\right)s_i^2}.$$

Test Welcha jest polecany jako mający większą moc, ale powinien być stosowany przy równolicznych próbach oraz gdy nie ma podstaw, by kwestionować

13 Charakterystykę poszczególnych testów Czytelnik znajdzie w podręczniku A. Malarskiej (2005) oraz S. Bedyńskiej i M. Cypryańskiej (2013b).

normalność rozkładów (a więc również ich symetryczność). W pozostałych przypadkach zaleca się test Browna-Forsythe'a (Glantz, Slinker, Neilands, 2001, s. 525).

## 5.3. Test H Kruskala-Wallisa

Jeżeli złamane jest założenie o normalności rozkładów, to należy posłużyć się nieparametrycznym odpowiednikiem testu F, to jest testem H Kruskala-Wallisa. Zalecenie to dotyczy szczególnie sytuacji, w której podpróby są mało liczne. Test ten zastosujemy również wtedy, gdy zmienna zależna mierzona jest na skali porządkowej. Dodajmy jeszcze, że uzasadnienia dla wykorzystania testu H Kruskala-Wallisa są takie same co dla testu U Manna-Whitneya (por. rozdział czwarty) – jedyną różnicą jest to, że test H umożliwia dokonywanie porównań dla więcej niż dwóch grup.

Układ hipotez w teście Kruskala-Wallisa przyjmuje postać:

$H_0$: wszystkie niezależne próbki pochodzą z populacji o takim samym rozkładzie

$H_1$: nieprawda, że wszystkie próbki pochodzą z populacji o takim samym rozkładzie.

Stosując bardziej formalny zapis, hipotezy te możemy ująć w postaci:

$H_0$: $F_1 = F_2 = \ldots = F_k$

$H_1$: $\neg H_0$;

gdzie $F$ oznacza dystrybuantę rozkładu zmiennej zależnej (por. rozdział drugi).

W teście Kruskala-Walissa wartości zmiennej zależnej zastępowane są rangami, w konsekwencji możemy porównać nie średnie arytmetyczne wartości zmiennej, ale średnie rang.

Sprawdzianem testu jest statystyka $H$, która ma postać:

$$H = \frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(n+1), \tag{26}$$

gdzie $n$ to liczebność całej próby, $n_i$ liczebność danej podpróby, a $R_i$ to suma rang dla danej podpróby.

Jeżeli każda z podprób ma liczebność wynoszącą co najmniej 5, to rozkład statystyki $H$ jest dobrze przybliżony przez rozkład $\chi^2$ z $df = k - 1$ stopniami swobody.

Jeśli w teście Kruskala-Wallisa $p < \alpha$, odrzucamy $H_0$, za prawdziwą uznajemy $H_1$. Wnioskujemy na tej podstawie, że rozkład zmiennej zależnej w przynajmniej jednej z porównywanej podpopulacji istotnie się różni (albo inaczej: próbki nie pochodzą z populacji o takim samym rozkładzie). W przeciwnym wypadku ($p > \alpha$)

nie ma podstaw do odrzucenia $H_0$ – nie ma istotnych różnic między porównywanymi podpopulacjami z punktu widzenia badanego zjawiska (próbki pochodzą z populacji o podobnym rozkładzie). Podobnie jak w przypadku analizy wariancji, jeśli w teście Kruskala-Wallisa odrzucimy $H_0$, w kolejnym kroku należy zastosować testy *post hoc*. Przykładowo: dla trzech podpopulacji hipotezy są następujące:

I. $H_0$: $F_1 = F_2$ *versus* $H_1$: $F_1 \neq F_2$;
II. $H_0$: $F_2 = F_3$ *versus* $H_1$: $F_2 \neq F_3$;
III. $H_0$: $F_1 = F_3$ *versus* $H_1$: $F_1 \neq F_3$.

Sprawdzianem hipotezy zerowej będzie moduł z różnicy między średnimi rangami w porównywanych grupach:

$$D = \left| \overline{R}_i - \overline{R}_j \right|, \tag{27}$$

gdzie $\overline{R}_i$ to średnia ranga w *i*-tej grupie.

Testy przeprowadzane są poprzez porównanie wartości statystyki $D$ z wartością obliczoną dla punktu krytycznego $C_{KW}$ rozkładu $\chi^2$ o poziomie istotności $\alpha$. Wartość ta obliczana jest według wzoru (Aczel, 2000):

$$C_{KW} = \sqrt{\chi^2_{\alpha,k-1}\left[\frac{n(n+1)}{12}\right]}\left(\frac{1}{n_i}+\frac{1}{n_j}\right). \tag{28}$$

Porównując wartości $D$ oraz $C_{KW}$ dla każdej pary, prowadzimy porównania wielokrotne na łącznym poziomie $\alpha$, dla którego był przeprowadzany cały test Kruskala-Wallisa. Hipoteza zerowa zostanie odrzucona wtedy, gdy $D > C_{KW}$.

**Przykład 5.1**

Przedsiębiorstwo oferujące usługi medyczne prowadzi działalność w czterech filiach. Firma chce dowiedzieć się, czy poziom zadowolenia pacjentów z oferowanych usług jest taki sam w prowadzonych przez nią placówkach. Badanie przeprowadzono za pomocą ankiety wypełnianej przez 27 pacjentów, losowo wybranych w każdej filii. Na podstawie pytań ankiety zbudowano indeks zadowolenia z usług, przyjmujący wartości z zakresu [0, 100]. Porównajmy średni poziom zadowolenia (zmienna *zadowolenie*) w poszczególnych filiach (zmienna *filia*).

**Rozwiązanie**

W badaniu tym zmienną zależną jest *zadowolenie* (zmienna mierzona na skali ilościowej). Jej poziom porównujemy w czterech populacjach wyróżnionych na podstawie zmiennej *filia* ($k = 4$). Analizę zaczynamy od zbadania rozkładów zmiennej zależnej pod kątem założeń testu F.

Podpróby są małe, w każdej z nich musimy przyjrzeć się rozkładom zmiennej zależnej pod kątem zaburzeń normalności. W rozdziale czwartym szczegółowo opisano, jak powinna wyglądać diagnostyka w tym zakresie – tu ograniczymy się do podania najważniejszych jej ustaleń (osiągniętych za pomocą procedury *Eksploracja*). W teście Shapiro-Wilka prawdopodobieństwo testowe dla każdej filii jest większe niż założony poziom $\alpha = 0{,}05$ (dla filii A wynosi $p = 0{,}398$, dla B $p = 0{,}772$, dla C $p = 0{,}589$, a dla D $p = 0{,}632$ – rysunek 5.2). W żadnej z filii nie ma zatem podstaw do odrzucenia hipotezy zerowej, rozkład można uznać za zgodny z rozkładem normalnym. Wnioski te współgrają z analizą statystyk. Po pierwsze, współczynnik skośności w podpróbie dla filii A wynosi 0,343, dla filii B –0,163, dla filii C 0,200, a dla filii D –0,160. Wartości te świadczą o bardzo słabej skośności występującej w każdej z podprób (zob. rozdział trzeci). Niepokojące nie są także wartości kurtozy, które wynoszą odpowiednio –0,793; –0,714; –0,831; –0,726. Reasumując, uznajemy, że pierwsze założenie testu parametrycznego F zostało spełnione.

**Testy normalności rozkładu**

| | | Kołmogorow-Smirnow[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | filia | Statystyka | df | Istotność | Statystyka | df | Istotność |
| poziom zadowolenia z usług placówki | A | .089 | 27 | .200* | .961 | 27 | .398 |
| | B | .104 | 27 | .200* | .976 | 27 | .772 |
| | C | .100 | 27 | .200* | .970 | 27 | .589 |
| | D | .104 | 27 | .200* | .971 | 27 | .632 |

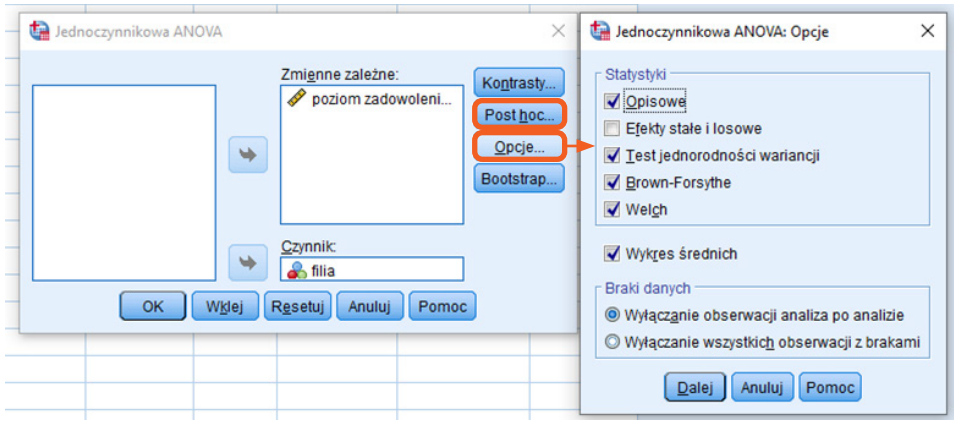
*. Dolna granica rzeczywistej istotności.

a. Z poprawką istotności Lilliefors

**Rysunek 5.2.** Wyniki testów normalności rozkładu zmiennej *zadowolenie* według zmiennej *filia*

Diagnostykę pod kątem spełnienia drugiego założenia, tj. jednorodności wariancji, przeprowadzimy już w ramach docelowej procedury, którą wywołujemy za pomocą polecenia *Analiza* → *Porównywanie* średnich → *Jednoczynnikowa ANOVA*. Zgodnie z rysunkiem 5.3 w oknie głównym tej procedury, w polu *Zmienne zależne* wprowadzamy zmienną informującą o poziomie zadowolenia klienta, a w polu *Czynnik* – zmienną informującą o filii, z której usług korzysta. Pod przyciskiem *Opcje* oznaczamy *Opisowe*, *Testowanie jednorodności*, *Wykres średnich*. Wybieramy również test Welcha i test

Browna-Forsythe'a na wypadek, gdyby okazało się, że wariancje nie są jednorodne i zamiast testem F musimy posłużyć się którymś z jego odpornych odpowiedników.



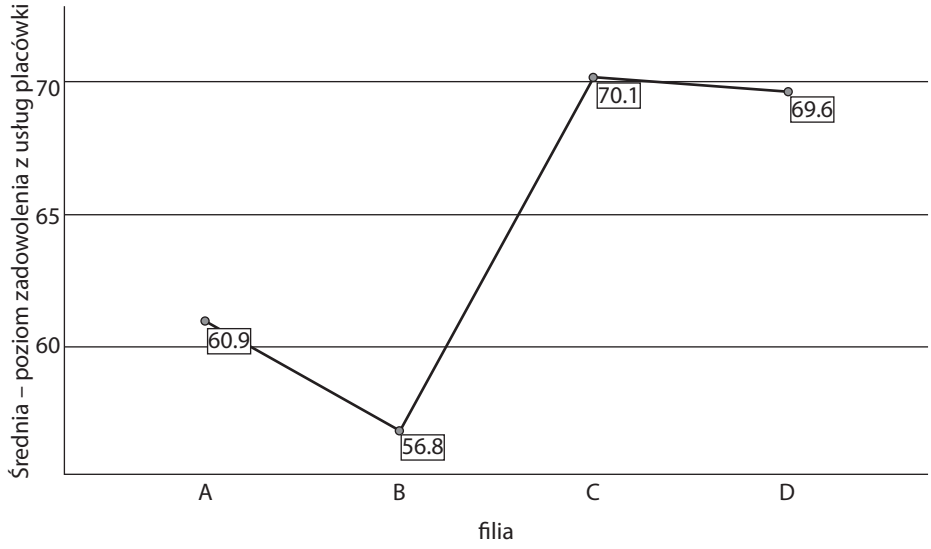

**Rysunek 5.3.** Wykonywanie polecenia *Porównywanie średnich → Jednoczynnikowa ANOVA*

Zgodnie z wynikami zaprezentowanymi na rysunku 5.4 wartości odchyleń standardowych są zbliżone. Sprawdźmy teraz wynik testu Levene'a bazujący na średniej. Skoro w teście tym $p = 0{,}220$ (jest wyższe od $\alpha = 0{,}05$), to nie ma podstaw do odrzucenia hipotezy zerowej mówiącej, że wariancja zmiennej zależnej jest taka sama w porównywanych populacjach. Założenie o jednorodności wariancji można uznać za spełnione, a więc do przeprowadzenia właściwej analizy możemy posłużyć się testem F.

**Statystyki opisowe**

poziom zadowolenia z usług placówki

| | N | Średnia | Odchylenie standardowe | Błąd standardowy | 95% przedział ufności dla średniej | | Minimum | Maksimum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Dolna granica | Górna granica | | |
| A | 27 | 60.89 | 8.724 | 1.679 | 57.44 | 64.34 | 48 | 79 |
| B | 27 | 56.81 | 6.264 | 1.205 | 54.34 | 59.29 | 44 | 69 |
| C | 27 | 70.15 | 7.609 | 1.464 | 67.14 | 73.16 | 57 | 84 |
| D | 27 | 69.59 | 8.924 | 1.717 | 66.06 | 73.12 | 50 | 84 |
| Ogółem | 108 | 64.36 | 9.706 | .934 | 62.51 | 66.21 | 44 | 84 |




**Test jednorodności wariancji**

| | | Test Levene'a | df1 | df2 | Istotność |
|---|---|---|---|---|---|
| poziom zadowolenia z usług placówki | Bazując na średniej | 1.495 | 3 | 104 | .220 |
| | Bazując na medianie | 1.438 | 3 | 104 | .236 |
| | Bazując na medianie i skorygowanych df | 1.438 | 3 | 99.688 | .236 |
| | Bazując na średniej obciętej | 1.502 | 3 | 104 | .218 |

**Jednoczynnikowa ANOVA**

poziom zadowolenia z usług placówki

| | Suma kwadratów | df | Średni kwadrat | F | Istotność |
|---|---|---|---|---|---|
| Między grupami | 3506.250 | 3 | 1168.750 | 18.488 | .000 |
| Wewnątrz grup | 6574.667 | 104 | 63.218 | | |
| Ogółem | 10080.917 | 107 | | | |

**Mocne testy równości średnich**

poziom zadowolenia z usług placówki

| | Statystyka[a] | df1 | df2 | Istotność |
|---|---|---|---|---|
| Welch | 21.597 | 3 | 57.199 | .000 |
| Brown-Forsythe | 18.488 | 3 | 97.653 | .000 |

a. Rozkład F asymptotyczny.

**Rysunek 5.4.** Statystyki opisowe, wyniki testu Levene'a, testu F i testów odpornych: porównanie zmiennej *zadowolenie* według *filia*

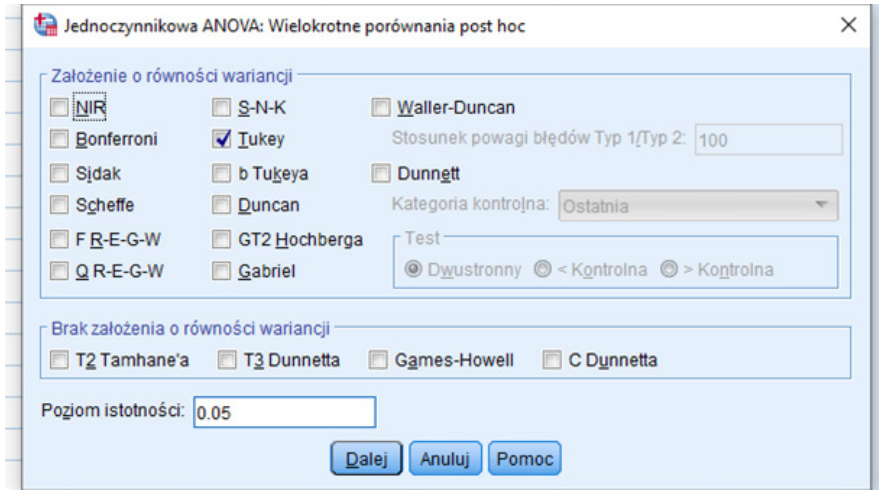
Uwzględniając analizowany przykład, układ hipotez możemy zapisać następująco:
$H_0$: $\mu_A = \mu_B = \mu_C = \mu_D$
$H_1$: $\neg (\mu_A = \mu_B = \mu_C = \mu_D)$.

Z tabeli wynikowej *Jednoczynnikowa ANOVA* odczytujemy wartość statystyki *F* oraz prawdopodobieństwo testowe: $F(3; 144) = 18{,}488$ (rysunek 5.4), co czytamy: „przy liczbie stopni swobody $df_1 = 3$ i $df_2 = 144$ statystyka *F* ma wartość 18,488". Prawdopodobieństwo w teście F jest bardzo małe (bliskie zera, $p < 0{,}001$), niższe od $\alpha$. Tym samym dostajemy argument na rzecz odrzucenia hipotezy zerowej oraz przyjęcia hipotezy alternatywnej. Tak więc między co najmniej dwiema średnimi (wartościami oczekiwanymi) różnica jest istotna statystycznie. W dalszej części za pomocą testów *post hoc* (porównań wielokrotnych) dowiemy się, między którymi konkretnie. Zanim jednak przejdziemy do tego etapu analizy, dokonamy oceny wielkości efektu. Do ustalenia wartości $\eta^2$ weźmiemy wartości sum kwadratów wyświetlane w tabeli *Jednoczynnikowa ANOVA*:

$$\eta^2 = \frac{SS_{międzygr.}}{SS_{całkowita}} = \frac{3506,250}{10080,917} = 0,3478 \approx 34,8\%.$$

Eta kwadrat można interpretować jako tę część całkowitej zmienności zmiennej zależnej, którą możemy przypisać oddziaływaniu czynnika. Nasz wynik informuje, że blisko 35% zmienności wyników reprezentujących poziom zadowolenia klientów badanej firmy pochodzi od czynnika *filia*. Efekt ocenimy jako duży. Pamiętajmy przy tym, że badanie miało charakter sondażowy, a nie eksperymentalny, i z tego względu nie możemy uznać, że za objaśnioną częścią zmienności stoi jedynie jakość usług dostarczanych w poszczególnych placówkach – filie mogą różnić się także pod innymi względami, chociażby strukturą pacjentów.

Skoro wynik testu F jest istotny statystycznie, analizę kontynuujemy. Wracamy do głównego okna procedury *Jednoczynnikowa ANOVA* (rysunek 5.3) i wybieramy przycisk *Post hoc*. Po otworzeniu się nowego okna (rysunek 5.5) zaznaczamy test *Tukey*. Zauważmy, że znajduje się on na liście testów mających zastosowanie w sytuacji, gdy spełnione jest założenie równości wariancji (a z taką sytuacją mamy tutaj do czynienia). Po dokonaniu wyboru testu deklarujemy również poziom $\alpha$, którego nie chcemy przekroczyć. Uruchamiamy wykonanie procedury.




**Rysunek 5.5.** Wykonywanie polecenia *Porównywanie średnich* → *Jednoczynnikowa ANOVA* → *Wielokrotne porównania post hoc*

Przed nami dokonanie sześciu porównań. Dla każdego porównania formułujemy hipotezę zerową i alternatywną:

I. $H_0 : \mu_A = \mu_B$ *versus* $H_1 : \mu_A \neq \mu_B$;
II. $H_0 : \mu_A = \mu_C$ *versus* $H_1 : \mu_A \neq \mu_C$;
III. $H_0 : \mu_A = \mu_D$ *versus* $H_1 : \mu_A \neq \mu_D$;
IV. $H_0 : \mu_B = \mu_C$ *versus* $H_1 : \mu_B \neq \mu_C$;
V. $H_0 : \mu_B = \mu_D$ *versus* $H_1 : \mu_B \neq \mu_D$;
VI. $H_0 : \mu_C = \mu_D$ *versus* $H_1 : \mu_C \neq \mu_D$.

Jak pokazują dane w tabeli *Porównania wielokrotne* (rysunek 5.6), wartość oczekiwana poziomu zadowolenia z usług medycznych dla klientów filii A nie różni się istotnie statystycznie od wartości oczekiwanej dla klientów filii B (w teście Tukeya $p > a$, a więc w przypadku porównania I nie ma podstaw do odrzucenia hipotezy zerowej). Innymi słowy, różnica w poziomie zadowolenia wynosząca 4,074 punktu, którą obserwujemy między losowymi podpróbami klientów filii A ($M = 60{,}89$; $S = 8{,}724$[14]) i filii B ($M = 56{,}81$; $S = 6{,}264$), mogłaby wystąpić, gdyby między populacjami klientów filii A i B różnicy w poziomie zadowolenia faktycznie nie było. Podobnie brak podstaw do odrzucenia hipotezy zerowej stwierdzamy w przypadku porównania VI. Obserwowaną różnicę między

14 Dane te odczytać możemy częściowo z wykresu średnich – są na nim zobrazowane średnie arytmetyczne, dostępne również w tabeli *Statystyki opisowe* (rysunek 5.3). Zwróćmy uwagę, że wykres średnich obrazuje średnie dla niezależnych prób. Choć linia łącząca punkty odnoszące się do tych średnich może sugerować ich zależność, na etapie interpretacji wyników należy pamiętać o tym, że próby są niezależne.

poziomem zadowolenia klientów filii C (*M* = 70,15; *S* = 7,609) i filii D (*M* = 69,59; *S* = 8,924), która sięga –0,056 punktu, można przypisać losowej zmienności próbek.

**Porównania wielokrotne**

Zmienna zależna: poziom zadowolenia z usług placówki

Test Tukey'a HSD

| (I) filia | (J) filia | Różnica średnich (I-J) | Błąd standardowy | Istotność | 95% przedział ufności | |
|---|---|---|---|---|---|---|
| | | | | | Dolna granica | Górna granica |
| A | B | 4.074 | 2.164 | .242 | -1.58 | 9.72 |
| | C | -9.259* | 2.164 | .000 | -14.91 | -3.61 |
| | D | -8.704* | 2.164 | .001 | -14.35 | -3.05 |
| B | A | -4.074 | 2.164 | .242 | -9.72 | 1.58 |
| | C | -13.333* | 2.164 | .000 | -18.98 | -7.68 |
| | D | -12.778* | 2.164 | .000 | -18.43 | -7.13 |
| C | A | 9.259* | 2.164 | .000 | 3.61 | 14.91 |
| | B | 13.333* | 2.164 | .000 | 7.68 | 18.98 |
| | D | .556 | 2.164 | .994 | -5.09 | 6.21 |
| D | A | 8.704* | 2.164 | .001 | 3.05 | 14.35 |
| | B | 12.778* | 2.164 | .000 | 7.13 | 18.43 |
| | C | -.556 | 2.164 | .994 | -6.21 | 5.09 |

*. Różnica średnich jest istotna na poziomie 0.05.

**Rysunek 5.6.** Wyniki porównań wielokrotnych średnich wartości zmiennej *zadowolenie* (w populacjach) pomiędzy filiami

W przypadku pozostałych porównań, czyli II, III, IV i V, prawdopodobieństwo w teście Tukeya $p < a$, co daje podstawę do odrzucenia hipotezy zerowej i przyjęcia hipotezy alternatywnej. Obserwowane wartości średnich (rysunek 5.4) sugerują, że klienci filii A (w obrębie podpróby *M* = 60,9; *S* = 8,724) są średnio mniej zadowoleni z usług medycznych niż klienci filii C (*M* = 70,15; *S* = 7,609), a także filii D (*M* = 69,59; *S* = 8,924). Podobnie pacjenci filii B są średnio mniej zadowoleni z poziomu świadczonych usług niż pacjenci filii C, a także filii D.

**Przykład 5.2**

W obecnej analizie skorzystamy z danych Europejskiego Sondażu Społecznego (ESS) zebranych w ósmej rundzie. Porównajmy średnią tygodniową liczbę godzin pracy (włączając w to nadgodziny) (zmienna: *wkhtot*) wśród pracujących mieszkańców Szwajcarii, Norwegii i Litwy (zmienna: *country*).

**Rozwiązanie**

W badaniu tym zmienną zależną jest *wkhtot* – tygodniowa liczba godzin pracy (zmienna mierzona na skali ilościowej). Jej poziom porównujemy w trzech populacjach wyróżnionych na podstawie zmiennej *country* ($k = 3$). Tym razem podpróby są duże (rysunek 5.7A), a więc do wyników diagnostyki pod kątem normalności rozkładów w podpróbach możemy podejść mniej restrykcyjnie. Dla porządku zreferujmy krótko jej wyniki. W przypadku każdego kraju wynik testu K-S nakazuje odrzucić hipotezę zerową mówiącą o normalności rozkładu zmiennej zależnej, ale wielkości współczynników skośności nie przekraczają 1 co do wartości bezwzględnej (tabele wynikowe nie są tu prezentowane). Kurtoza dla Szwajcarii wynosi 0,098, dla Litwy 6,711 i dla Norwegii 2,435. W tej sytuacji możemy bezpiecznie posłużyć się średnią i testem parametrycznym w celu porównania podpopulacji.

Odnieśmy się do drugiego założenia. Jak widać (rysunek 5.7C), wynik testu Levene'a nakazuje odrzucić hipotezę zerową o równości wariancji. Co więcej, analiza odchyleń standardowych wskazuje, że różnice w rozproszeniach wyników wokół średnich są duże – największe pojawiają się w przypadku Szwajcarii ($S = 16{,}816$) i Litwy ($S = 8{,}061$), a stosunek największego do najmniejszego odchylenia standardowego jest większy niż 2 (rysunek 5.7A). Zgodnie z wynikami mieszkańcy Litwy nie tylko średnio więcej pracują w porównaniu z mieszkańcami Szwajcarii, ale także ich zbiorowość jest znacznie mniej zróżnicowana, jeśli chodzi o liczbę godzin pracy. W tej sytuacji zasadne jest, aby hipotezę o równości średnich sprawdzić za pomocą testu Browna-Forsythe'a.

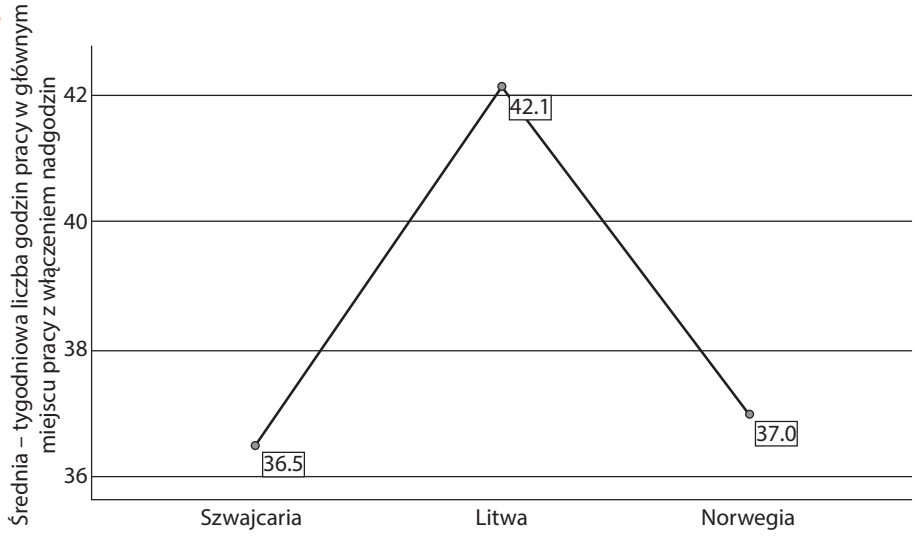
A

**Statystyki opisowe**

Tygodniowa liczba godzin pracy w głównym miejscu pracy z włączeniem nadgodzin

| | N | Średnia | Odchylenie standardowe | Błąd standardowy | 95% przedział ufności dla średniej | | Minimum | Maksimum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Dolna granica | Górna granica | | |
| Szwajcaria | 649 | 36.47 | 16.816 | .660 | 35.17 | 37.76 | 0 | 99 |
| Litwa | 179 | 42.13 | 8.061 | .603 | 40.94 | 43.32 | 3 | 100 |
| Norwegia | 402 | 36.97 | 13.375 | .667 | 35.66 | 38.28 | 0 | 126 |
| Ogółem | 1229 | 37.45 | 14.857 | .424 | 36.62 | 38.29 | 0 | 126 |

B




C

**Test jednorodności wariancji**

| | | Test Levene'a | df1 | df2 | Istotność |
|---|---|---|---|---|---|
| Tygodniowa liczba godzin pracy w głównym miejscu pracy z włączeniem nadgodzin | Bazując na średniej | 61.753 | 2 | 1226 | .000 |
| | Bazując na medianie | 38.226 | 2 | 1226 | .000 |
| | Bazując na medianie i skorygowanych df | 38.226 | 2 | 1114.351 | .000 |
| | Bazując na średniej obciętej | 60.073 | 2 | 1226 | .000 |

D

**Jednoczynnikowa ANOVA**

Tygodniowa liczba godzin pracy w głównym miejscu pracy z włączeniem nadgodzin

| | Suma kwadratów | df | Średni kwadrat | F | Istotność |
|---|---|---|---|---|---|
| Między grupami | 4634.706 | 2 | 2317.353 | 10.661 | .000 |
| Wewnątrz grup | 266491.786 | 1226 | 217.367 | | |
| Ogółem | 271126.492 | 1228 | | | |

E

**Mocne testy równości średnich**

Tygodniowa liczba godzin pracy w głównym miejscu pracy z włączeniem

| | Statystyka[a] | df1 | df2 | Istotność |
|---|---|---|---|---|
| Welch | 25.233 | 2 | 649.648 | .000 |
| Brown-Forsythe | 14.978 | 2 | 1181.134 | .000 |

a. Rozkład F asymptotyczny.

**Rysunek 5.7.** Statystyki opisowe, wynik testu Levene'a, testu F i testów odpornych: porównanie zmiennej *wkhtot* według *country*

Bardzo niskie prawdopodobieństwo dla statystyki $F_{BF}$ ($p < 0{,}01$) (rysunek 5.7E) daje nam mocny argument na rzecz odrzucenia hipotezy zerowej. Tak więc między co najmniej dwiema wartościami oczekiwanymi (średnimi w populacjach) występuje różnica

istotna statystycznie. Analizę będziemy kontynuować, wykorzystując test *post hoc*. Tym razem jednak wyboru dokonamy spośród testów odpowiednich do sytuacji, w której nie jest spełnione założenie o równości wariancji. Konkretnie wybierzemy test Gamesa-Howella – należy on do grupy testów liberalnych, mających zastosowanie w przypadku nierównolicznych prób.

**Porównania wielokrotne**

Zmienna zależna: Tygodniowa liczba godzin pracy w głównym miejscu pracy z włączeniem nadgodzin

Test Gamesa-Howella

| (I) kraj | (J) kraj | Różnica średnich (I-J) | Błąd standardowy | Istotność | 95% przedział ufności | |
|---|---|---|---|---|---|---|
| | | | | | Dolna granica | Górna granica |
| Szwajcaria | Litwa | -5.665* | .894 | .000 | -7.77 | -3.56 |
| | Norwegia | -.501 | .939 | .855 | -2.70 | 1.70 |
| Litwa | Szwajcaria | 5.665* | .894 | .000 | 3.56 | 7.77 |
| | Norwegia | 5.163* | .900 | .000 | 3.05 | 7.28 |
| Norwegia | Szwajcaria | .501 | .939 | .855 | -1.70 | 2.70 |
| | Litwa | -5.163* | .900 | .000 | -7.28 | -3.05 |

*. Różnica średnich jest istotna na poziomie 0.05.

**Rysunek 5.8.** Wyniki porównań wielokrotnych średnich wartości zmiennej *wkhtot* (w populacjach) pomiędzy krajami

Wyniki testu Gamesa-Howella (rysunek 5.8) pozwalają na wniosek, że wartości oczekiwane dla mieszkańców Litwy i Szwajcarii, a także Litwy i Norwegii różnią się istotnie statystycznie (dla każdej z tych par krajów $p < 0{,}001$). Wielkości średnich w próbach sugerują, że mieszkańcy Litwy ($M = 42{,}13$; $S = 8{,}061$) pracują tygodniowo więcej niż mieszkańcy Szwajcarii ($M = 36{,}47$; $S = 16{,}816$) oraz Norwegii ($M = 36{,}97$; $S = 13{,}375$). W przypadku porównania samych mieszkańców Szwajcarii i Norwegii wynik testu nie daje podstaw do odrzucenia hipotezy zerowej – $p = 0{,}855$ (liczba godzin pracy nie różni się więc istotnie w przypadku mieszkańców tych dwóch krajów).

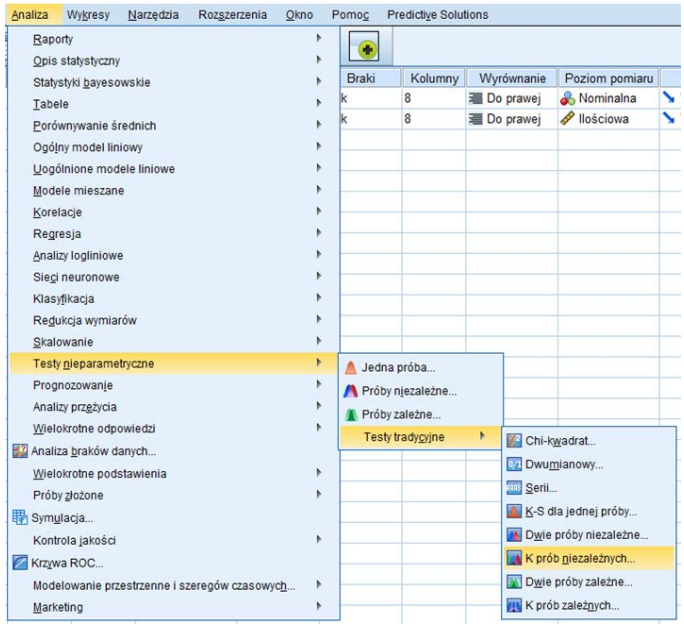
**Przykład 5.3**

Porównajmy firmy z trzech działów przemysłu lekkiego: włókienniczego, odzieżowego i skórzanego (zmienna: *przemysł*) pod kątem wielkości nakładów inwestycyjnych (w mln zł) (zmienna: *wsk*, etykieta zmiennej: współczynnik). Chcemy się dowiedzieć, czy w obrębie tych trzech gałęzi przedsiębiorstwa różnią się pod tym względem. Pobrano niewielkie próby firm w każdej z tych gałęzi. Analityk rynku sugeruje, że rozkład nakładów inwestycyjnych w każdej z trzech populacji firm odbiega od rozkładu normalnego.

### Rozwiązanie

W badaniu tym zmienną zależną jest *wsk* – nakłady inwestycyjne (zmienna mierzona na skali ilościowej). Jej poziom porównujemy w trzech populacjach wyróżnionych na podstawie zmiennej *przemysł* ($k = 3$). Problem badawczy może więc wskazywać na zastosowanie metody parametrycznej, niemniej jednak z uwagi na uszczegółowienie w treści zadania (wątpliwa normalność rozkładów w badanych populacjach) zastosujemy w tym przypadku test Kruskala-Wallisa. IBM SPSS Statistics udostępnia dwie procedury do przeprowadzenia tego testu.

Według pierwszego podejścia wybieramy *Analiza → Testy nieparametryczne → Testy tradycyjne → K prób niezależnych* (rysunek 5.9).




**Rysunek 5.9.** Wykonywanie polecenia *Testy nieparametryczne → K prób niezależnych* (ścieżka 1)

W oknie głównym tej procedury zmienną ilościową *wsk* umieszczamy w polu *Zmienne testowane*, a zmienną *przemysł* w polu *Zmienna grupująca*. Od użytkownika wymaga się jeszcze, aby zdefiniował zakres zmiennej grupującej. Ponieważ w naszym przypadku przemysł włókienniczy został oznaczony jako 1, odzieżowy jako 2, a skórzany jako 3, to zakres określamy jako 1–3. Akceptujemy domyślny wybór testu (*H Kruskala-Wallisa*). Uruchamiamy procedurę, klikając w *OK*.

W oknie raportowym (rysunek 5.10) mamy podane, jakie są wartości średnich rang w poszczególnych podpróbach (tabela *Rangi*), a także jaka jest wartość statystyki *H*

oraz prawdopodobieństwo testowe (tabela *Wartość testowana*). W naszym przykładzie będziemy się interesować rozstrzygnięciem:

$H_0$: $F_{włókienniczy} = F_{odzieżowy} = F_{skórzany}$

$H_1$: $\neg (F_{włókienniczy} = F_{odzieżowy} = F_{skórzany})$.

**Rangi**

| | przemysł | N | Średnia ranga |
|---|---|---|---|
| współczynnik | 1 włókienniczy | 8 | 10.19 |
| | 2 odzieżowy | 6 | 17.00 |
| | 3 skórzany | 6 | 4.42 |
| | Ogółem | 20 | |

**Wartość testowana**[a,b]

| | współczynnik |
|---|---|
| H Kruskala-Wallisa | 13.619 |
| df | 2 |
| Istotność asymptotyczna | .001 |

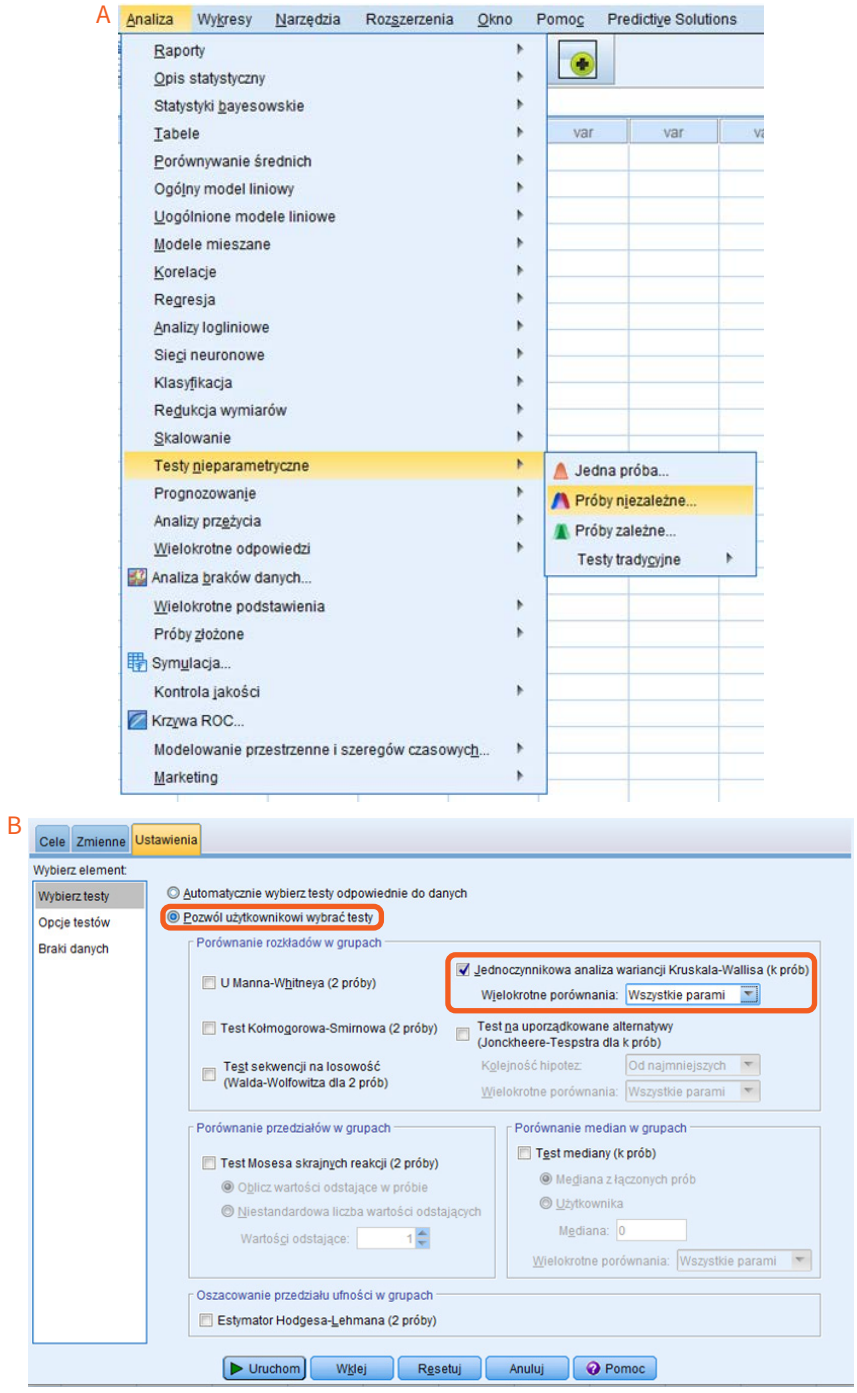
a. Test Kruskala-Wallisa

b. Zmienna grupująca: przemysł

**Rysunek 5.10.** Wynik testu Kruskala-Wallisa: porównanie poziomu zmiennej *wsk* według *przemysł* (ścieżka 1)

Widzimy, że prawdopodobieństwo testowe *p* jest bardzo małe – $p = 0{,}001 < \alpha$ (rysunek 5.10), stąd podejmujemy decyzję o odrzuceniu hipotezy zerowej i przyjęciu hipotezy alternatywnej. Różnice w podpopulacjach są zatem statystycznie istotne, niemniej nie wiemy jeszcze, między którymi działami przemysłu lekkiego można je stwierdzić. Aby się tego dowiedzieć, będziemy kontynuować analizy za pomocą wielokrotnych porównań. Patrząc na wartości średnich rang, możemy spodziewać się wystąpienia istotnej różnicy między firmami z przemysłu odzieżowego i skórzanego. Zobaczymy, czy pozostałe różnice również okażą się istotne. Aby uzyskać potrzebne wyniki, skorzystamy z drugiej ścieżki przeprowadzenia testu Kruskala-Wallisa w IBM SPSS Statistics.

Tym razem wybierzemy *Analiza → Testy nieparametryczne → Próby niezależne* (rysunek 5.11A). W zakładce *Cele* zaznaczamy *Analiza niestandardowa*. W zakładce *Zmienne* określamy, która zmienna jest zależna (testowana), a która jest czynnikiem (zmienną grupującą). W zakładce *Ustawienia* (rysunek 5.11B) wybieramy opcję *Pozwól użytkownikowi wybrać testy*, dalej zaznaczamy *Jednoczynnikowa analiza wariancji Kruskala-Wallisa (k prób)*, a linijkę niżej doprecyzowujemy, że interesują nas *Wielokrotne porównania* w opcji *Wszystkie parami*. Teraz możemy uruchomić procedurę.




**Rysunek 5.11.** Wykonywanie polecenia *Próby niezależne* (ścieżka 2)

Początkowy widok zawartości okna raportowego wygląda jak na rysunku 5.12.

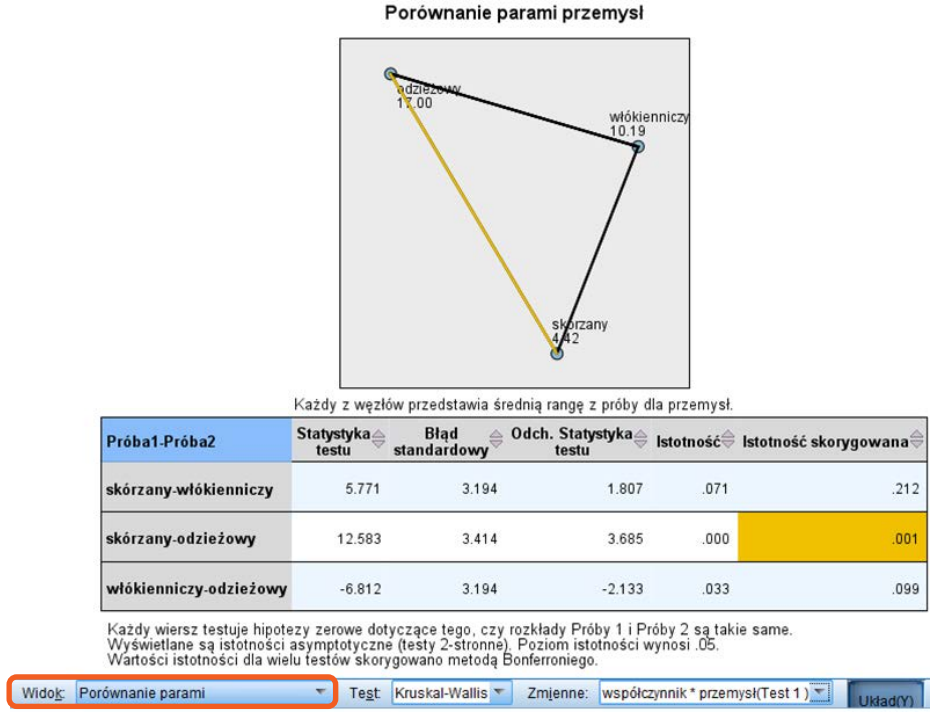
**Podsumowanie testu hipotezy**

| | Hipoteza zerowa | Test | Istotność | Decyzja |
|---|---|---|---|---|
| 1 | Rozkład współczynnik jest taki sam dla kategorii zmiennej przemysł. | Test Kruskala-Wallisa dla prób niezależnych | .001 | Odrzuć hipotezę zerową. |

Przedstawiono asymptotyczne istotności. Poziom istotności wynosi .05.

**Rysunek 5.12.** Wynik testu Kruskala-Wallisa: porównanie zmiennej *wsk* według *przemysł* (ścieżka 2)

W tabeli wyświetla się, podobnie jak przy ścieżce 1, informacja o prawdopodobieństwie testowym dla statystyki $H$, które wynosi $p$ = 0,001 (rysunek 5.12). Aby przejść do porównań wielokrotnych, musimy teraz dwa razy kliknąć w tabelę – w ten sposób przejdziemy do *Przeglądu modelu*. Raport otworzy się w nowym, interaktywnym oknie. Okno jest podzielone na dwie części. Lewa część zawiera to, co już widzieliśmy w oknie raportu. Przejdźmy do prawej części okna i paska na dole, by zmienić *Widok*. Z rozwijanej listy wybierzmy *Porównania parami* (rysunek 5.13).




**Rysunek 5.13.** Widok okna raportowego z wynikami porównań wielokrotnych: porównanie zmiennej *wsk* według *przemysł* (ścieżka 2)

Przed nami dokonanie trzech porównań. Dla każdego porównania formułujemy hipotezy:

I. $H_0$: $F_{włókienniczy} = F_{odzieżowy}$ *versus* $H_1$: $F_{włókienniczy} \neq F_{odzieżowy}$;
II. $H_0$: $F_{włókienniczy} = F_{skórzany}$ *versus* $H_1$: $F_{włókienniczy} \neq F_{skórzany}$;
III. $H_0$: $F_{odzieżowy} = F_{skórzany}$ *versus* $H_1$: $F_{odzieżowy} \neq F_{skórzany}$.

W kolumnie *Statystyka testu* wyświetlone są wartości *D*, a więc różnice między rangami. Dla porównania III (skórzany *versus* odzieżowy) w teście *post hoc* $p = 0{,}001 < a$ (bierzemy pod uwagę wersję z korektą Bonferroniego) różnica w rozkładach dla tych działów przemysłu jest istotna statystycznie. W przypadku porównań I i II nie ma podstaw do odrzucenia hipotezy zerowej.

Na dołączonym wykresie dostajemy również informację o wartościach średnich rang w porównywanych podpróbach. Zauważmy, że program IBM SPSS Statistics wspomaga użytkownika, wyróżniając dodatkowym kolorem (w oryginalnej tabeli wynikowej – kolorem żółtym) różnice istotne statystycznie.

# 6. Ocena zależności między dwiema zmiennymi

**Kluczowe pojęcia:** tablica kontyngencji, tabela krzyżowa, niezależność cech, liczebności oczekiwane, test niezależności $\chi^2$, dokładny test Fishera, współczynniki zależności oparte na statystyce chi-kwadrat, współczynnik korelacji liniowej Pearsona, współczynnik korelacji rang Spearmana, współczynnik korelacji rang Kendalla

## 6.1. Badanie zależności między dwiema zmiennymi jakościowymi

### 6.1.1. Uwagi wstępne

W sytuacji gdy cechy X i Y są jakościowe (określane też jako dyskretne, tj. mierzone na skali nominalnej albo porządkowej, zwłaszcza o niewielkiej liczbie wariantów cechy), do badania zależności między nimi używa się tablicy kontyngencji. Tablice kontyngencji są częstym narzędziem analizy danych uzyskiwanych w badaniach sondażowych. Wynika to ze specyfiki tych danych, mają one bowiem najczęściej jakościowy charakter (większość zmiennych jest mierzona na skali nominalnej lub porządkowej).

Tablica kontyngencji nazywana jest w IBM SPSS Statistics tabelą krzyżową. Nazwa ta oddaje sposób budowy tabeli – w kolumnach umieszcza się kategorie (warianty) jednej zmiennej, a w wierszach kategorie drugiej zmiennej. W efekcie tabela ta pozwala – w uproszczeniu – „krzyżować" rozkład cechy Y z rozkładem cechy X, a także sprawdzić, czy przynależność obiektu do określonej kategorii jednej zmiennej jest związana z jego przynależnością do danej kategorii drugiej zmiennej.

I tak moglibyśmy dojść przykładowo do ustalenia, że przynależność do kategorii „głosował w ostatnich wyborach" częściej idzie w parze z przynależnością

do kategorii „ma wyższe wykształcenie" niż „nie ma wyższego wykształcenia". Ustalenie to świadczyłoby o istnieniu zależności między wykształceniem a udziałem w wyborach. Mówiąc inaczej, powiedzielibyśmy, że zmienne wykształcenie i udział w wyborach nie są wzajemnie niezależne. Innym razem odkrylibyśmy, dajmy na to, że przynależność do wybranej kategorii płci nie wiąże się z częstszą przynależnością do kategorii „głosował w ostatnich wyborach", skoro deklarację udziału w wyborach obserwowalibyśmy równie często wśród kobiet co wśród mężczyzn. Ustalenie to świadczyłoby o braku zależności między płcią a udziałem w wyborach – inaczej powiedzielibyśmy, że zmienne płeć i udział w wyborach są wzajemnie niezależne.

Przedstawmy inne przykłady problemów badawczych:

- Firma produkująca samochody przeprowadziła badanie marketingowe na losowej próbce swoich klientów. Każdemu nabywcy zadano pytanie o model kupionego samochodu (A, B, C, D) oraz pytanie, w którym proszono o wybranie jednej z czterech charakterystyk najlepiej opisujących go jako kierowcę. Badacz interesuje się odpowiedzią na pytanie, czy istnieje zależność między tym, jaki model samochodu nabywca wybiera, a tym, za jakiego kierowcę się uważa (Aczel, Sounderpandian, 2018).
- Czy istnieje zależność między rodzajem gminy (wiejska, wiejsko-miejska, miejska) a stosowaniem przez gminę instrumentów podnoszenia atrakcyjności osiedleńczej?

Dane jakościowe niekoniecznie muszą być zbierane na drodze badań sondażowych, mogą też pochodzić z dostępnych już źródeł informacji, zbieranych na przykład w ramach statystyki publicznej. To, co chcemy podkreślić, to fakt, że zarówno dane sondażowe, jak i jakościowe dane zastane są danymi obserwacyjnymi, a te zasadniczo różnią się od danych eksperymentalnych. Różnica ta dotyczy nie tyle sposobu analiz, który pozostaje taki sam dla obu typów danych, ile sposobu sformułowania wniosków. Rozważmy dwa sposoby pozyskania danych, które mogłyby posłużyć do rozstrzygnięcia, czy między obejrzeniem spotu wyborczego kandydata X (tak, nie) a deklaracją chęci oddania na niego głosu w wyborach (tak, nie) istnieje istotna statystycznie zależność. Badanie miałoby charakter obserwacyjny, gdybyśmy pobrali próbę i każdej osobie zadali pytanie o obejrzenie spotu kandydata oraz pytanie o chęć oddania na niego głosu. Badanie miałoby charakter eksperymentalny, gdybyśmy pobraną próbę podzielili losowo na dwie podpróby, po czym osobom znajdującym się w pierwszej z nich wyświetlili spot, a znajdującym się w drugiej nie, a następnie każdemu badanemu (z jednej i z drugiej podpróby) zadali pytanie umożliwiające ustalenie, czy zamierza oddać głos na X. Badania eksperymentalne mają tę przewagę nad obserwacyjnymi, że dają mocniejszą podstawę do mówienia o wpływie jednej zmiennej na drugą. O ile

stwierdzimy wystąpienie zależności o oczekiwanym kształcie, to możemy przyjąć, że różnice w liczbie chętnych poprzeć X w wyborach wywołał spot wyborczy. Takie rozumowanie jest zasadne, ponieważ porównywane grupy osób różniły się tylko obejrzeniem spotu wyborczego, poza tym grupy te były takie same (lub bardzo podobne) pod względem wszystkich pozostałych cech. Natomiast w badaniach obserwacyjnych ten warunek nie jest spełniony, dlatego wystąpienia zależności między obejrzeniem spotu a chęcią oddania głosu na występującego w nim kandydata nie możemy prosto tłumaczyć wpływem obejrzenia spotu. Nie możemy bowiem wykluczyć, że za wystąpienie zależności odpowiedzialne są inne czynniki, które różnicują osoby znające spot i nieznające go.

Podkreślenie ograniczeń, jakie powinniśmy sobie narzucić przy interpretacji wyników osiągniętych za pomocą badań obserwacyjnych, jest tym bardziej ważne, że w analizie tablic kontyngencji nierzadko praktykuje się wskazywanie jednej zmiennej jako niezależnej (wyjaśniającej), a drugiej jako zależnej (wyjaśnianej) – w celu łatwiejszego uchwycenia zależności dane przedstawia się w postaci procentowej, tak by pokazać, jak wyglądają rozkłady zmiennej zależnej w poszczególnych kategoriach zmiennej niezależnej (np. jak rozkłada się poparcie dla danej partii politycznej w poszczególnych grupach wieku) (zob. Szwed, 2008, s. 259; Agresti, Franklin, 2013: s. 90; Agresti, Finlay, 2014, s. 222). Także praktyki językowe narzucają nam myślenie, że „coś zależy od czegoś". Mówimy – przykładowo – że badamy zależność opinii od płci, ale już nie mówimy, że badamy zależność płci od opinii, bo taki kierunek oddziaływania nie jest możliwy. Z drugiej strony jednak, o ile zależność między płcią a opinią występuje, to ujawni się ona zarówno wtedy, gdy będziemy analizować procentowe rozkłady opinii w grupach płci, jak i wtedy, gdy będziemy analizować procentowe rozkłady płci w grupach opinii. Bywa też, że badając zależność, każdą zmienną z pary moglibyśmy sensownie rozpatrywać jako zmienną niezależną. I tak na przykład opinie jednostki mogą kształtować jej działania, ale i działania jednostki mogą kształtować jej opinie. To wszystko pokazuje, że analizując dane obserwacyjne, z ostrożnością powinniśmy posługiwać się pojęciem zmiennej wyjaśniającej i wyjaśnianej, pamiętając, że przypisywanie tych ról badanym cechom jest niekiedy wątpliwe, po części arbitralne, a z pewnością nie jest konieczne do ustalenia tego, czy cechy są ze sobą związane, czy nie – takiego rozróżnienia nie wymaga ani test $\chi^2$, ani miary siły związku oparte na $\chi^2$, które będziemy omawiać w tym rozdziale. Przeprowadzona analiza pozwoli nam w tym przypadku wnioskować o współwystępowaniu dwóch zjawisk lub o związku/zależności/relacji między dwiema zmiennymi (a nie wpływie/oddziaływaniu jednej zmiennej na drugą). Sugerujemy również nieużywanie w tym przypadku określeń „zmienna zależna" (można używać określeń „badane zjawisko") i „zmienna niezależna" („zmienna

traktowana jako czynnik"/„zmienna grupująca"). W końcu powinniśmy też pamiętać o ograniczeniach przedstawianych tu metod analizy, ponieważ przewidują one badanie zależności tylko między dwiema zmiennymi. Tymczasem metody bardziej zaawansowane, pozwalające badać powiązania między większą liczbą zmiennych, prowadzą niekiedy do ustalenia, że początkowo obserwowana zależność między dwiema zmiennymi znika, jeśli do analiz wprowadzimy dodatkowe zmienne (zagadnienia te będą przedmiotem ostatniego rozdziału).

### 6.1.2. Test niezależności $\chi^2$ a dokładny test Fishera

Test niezależności $\chi^2$ (czytaj: chi kwadrat) pozwala rozstrzygnąć, czy między dwiema zmiennymi istnieje zależność w populacji. Test ten zalecany jest (zgodnie z tym, co opisano w poprzednim podpunkcie) dla zmiennych jakościowych/dyskretnych.

Posłużenie się tym testem wymaga spełnienia następującego założenia: w każdej komórce tabeli krzyżowej liczebność oczekiwana ≥ 5.

Hipotezy w teście niezależności przyjmują następującą postać:
$H_0$: zmienne są niezależne
$H_1$: zmienne nie są niezależne.

W jaki sposób ustala się wystąpienie zależności? Zacznijmy od prostego spostrzeżenia, że przypadek braku zależności realizuje się w jeden sposób, a zależność może się realizować na wiele sposobów. Dlatego badając relację między zmiennymi, sprawdzamy, jak dalece otrzymane dane odbiegają od przypadku niezależności. W zgodzie z tą logiką przebiega również wnioskowanie statystyczne, w którym początkowo zakładamy, że prawdziwa jest hipoteza zerowa, by następnie sprawdzić, jak bardzo – przy tym założeniu – prawdopodobne jest uzyskanie takiej wartości sprawdzianu, jaką otrzymaliśmy. Jeżeli jest bardzo mało prawdopodobne, to tracimy zaufanie do tego, co głosi hipoteza zerowa, a nabieramy zaufania do tego, co głosi hipoteza alternatywna.

Przyjrzyjmy się dokładniej temu, jak przebiega analiza. Tabela 6.1 przedstawia ogólną postać tablicy kontyngencji. Jej wiersze odpowiadają wariantom jednej zmiennej, a kolumny wariantom drugiej zmiennej. Liczebność elementów w komórce $i$-tego wiersza (gdzie $i = 1, 2, \ldots, r$) i $j$-tej kolumny (gdzie $j = 1, 2, \ldots, c$) oznaczono jako $n_{ij}$. Określa się je jako liczebności warunkowe (liczebność warunkową rozumiemy jako liczebność danej grupy wyróżnionej przez warianty pierwszej zmiennej pod warunkiem, że jednostka ma określony wariant drugiej zmiennej). Z kolei sumę liczebności w poszczególnych wierszach i kolumnach określa się jako liczebności brzegowe.

**Tabela 6.1.** Układ tabeli kontyngencji $r \times c$ (tabeli o $r$ wierszach i $c$ kolumnach)

| | | Druga zmienna | | | | Ogółem w wierszach |
|---|---|---|---|---|---|---|
| | Warianty | 1 | 2 | … | $c$ | |
| Pierwsza zmienna | 1 | $n_{11}$ | $n_{12}$ | … | $n_{1c}$ | $R_1$ |
| | 2 | $n_{21}$ | $n_{22}$ | … | $n_{2c}$ | $R_2$ |
| | … | … | … | | … | … |
| | $r$ | $n_{r1}$ | $n_{r2}$ | … | $n_{rc}$ | $R_r$ |
| Ogółem w kolumnach | | $C_1$ | $C_2$ | | $C_c$ | $n$ |

**Źródło:** opracowanie własne.

Po przedstawieniu w tabeli liczebności empirycznych $n_{ij}$ kolejnym krokiem jest obliczenie dla każdej komórki tabeli tzw. liczebności oczekiwanych. Liczebności oczekiwane to te liczebności, których spodziewalibyśmy się w przypadku niezależności zmiennych. Liczebności oczekiwane obliczamy dla każdej komórki według wzoru (McClave, Sincich, 2018, s. 814):

$$\hat{O}_{ij} = \frac{R_i C_j}{n}. \tag{27}$$

Mniej formalnie możemy to przedstawić za pomocą formuły:

$$\hat{O}_{ij} = \frac{\text{Liczebność w danym wierszu ogółem} \times \text{liczebność w danej kolumnie ogółem}}{\text{Całkowita liczebność próby}}. \tag{28}$$

Tak więc liczebność oczekiwaną w komórce znajdującej się – powiedzmy – w pierwszym wierszu i drugiej kolumnie tabeli obliczymy w ten sposób: $\hat{O}_{12} = \frac{R_1 C_2}{n}$.

Statystyka chi-kwadrat ($\chi^2$) stanowi miernik, który w syntetyczny sposób zestawia informacje pochodzące ze wszystkich komórek tabeli i informuje o tym, jak bardzo liczebności empiryczne odbiegają od liczebności oczekiwanych. Przyjmuje ona postać (McClave, Sincich, 2018, s. 814):

$$\chi^2 = \sum \frac{\left[n_{ij} - \hat{O}_{ij}\right]^2}{\hat{O}_{ij}}, \tag{29}$$

a przy mniej formalnym zapisie może być ujęta jako suma obliczona dla każdej komórki tablicy kontyngencji z wyrażeń:

$$\frac{[\text{liczebność empiryczna w komórce} - \text{liczebność oczekiwana w komórce}]^2}{\text{liczebność oczekiwana w komórce}}.$$

W przypadku gdy prawdziwa jest hipoteza zerowa, statystyka $\chi^2$ ma rozkład zbliżony do rozkładu $\chi^2$ o $df = (r - 1) \times (c - 1)$ stopniach swobody.

Przyjmując $\alpha = 0{,}05$, wnioskowanie przeprowadzamy według reguły:

- jeżeli prawdopodobieństwo w teście $p < 0{,}05$, to stwierdzamy, że odrzucamy hipotezę zerową i uznajemy, że są podstawy do przyjęcia hipotezy alternatywnej; zależność uznajemy wtedy za statystycznie istotną;
- jeżeli $p > 0{,}05$, to stwierdzamy, że brak jest podstaw do odrzucenia hipotezy zerowej; zależność nie jest istotna statystycznie.

Odrzuceniu hipotezy zerowej sprzyjają wysokie wartości statystyki $\chi^2$, aczkolwiek bez wiedzy na temat liczby stopni swobody nie uprawniają one jeszcze do podjęcia decyzji o istotności związku.

Jeśli założenie dotyczące liczebności oczekiwanych w tablicy kontyngencji nie jest spełnione, a więc gdy zdarza się sytuacja, że w którejś komórce tablicy kontyngencji $\hat{O}_{ij} < 5$, co do zasady zaleca się stosować dokładny test Fishera. Taki problem dotyczy w szczególności małych prób. Gdy założenie nie jest spełnione, rozkład $\chi^2$ nie jest wystarczająco dobrym przybliżeniem rozkładu statystyki $\chi^2$, a co za tym idzie – prawdopodobieństwo testowe (oznaczone jako *Istotność asymptotyczna dwustronna*) może nie być wiarygodnym wynikiem. W takiej sytuacji, ustalając wartość $p$, zamiast metodą asymptotyczną należy posłużyć się metodą dokładną. Jak dodają Agresti i Finlay (2014, s. 228), dokładny test Fishera może być stosowany nie tylko przy małych próbach, ale i przy próbach o dowolnej liczebności[15]. Nie można jednak zapominać o tym, że iteracyjny sposób wyznaczenia $p$ powoduje, że przy dużych tabelach krzyżowych, a więc wielu wariantach zmiennych, test ten wymaga dużych mocy obliczeniowych komputera. W praktyce już przy tabelach o wymiarach $4 \times 4$ oszacowanie prawdopodobieństwa może być trudne.

Zauważmy, że jeśli uznamy, że zależność między zmiennymi jest istotna statystycznie, to *de facto* możemy też wnioskować, że podpopulacje, na jakie dzieli zbiorowość jedna ze zmiennych, różnią się istotnie. Przykładowo: jeśli istnieje statystycznie istotny związek między płcią a uczestnictwem w wyborach, to możemy powiedzieć, że kobiety i mężczyźni różnią się między sobą pod tym względem. Test niezależności chi-kwadrat i dokładny test Fishera proponuje się w związku z tym jako metody porównań między dwiema lub więcej niż dwiema populacjami w sytuacji, gdy badane zjawisko mierzone jest na skali nominalnej.

15 Nie będziemy tu przedstawiać sposobu obliczania poziomu istotności w dokładnym teście Fishera. Czytelnikom zainteresowanym tym zagadnieniem polecamy lekturę podręcznika Agrestiego (2007, s. 45–47).

**Przykład 6.1**

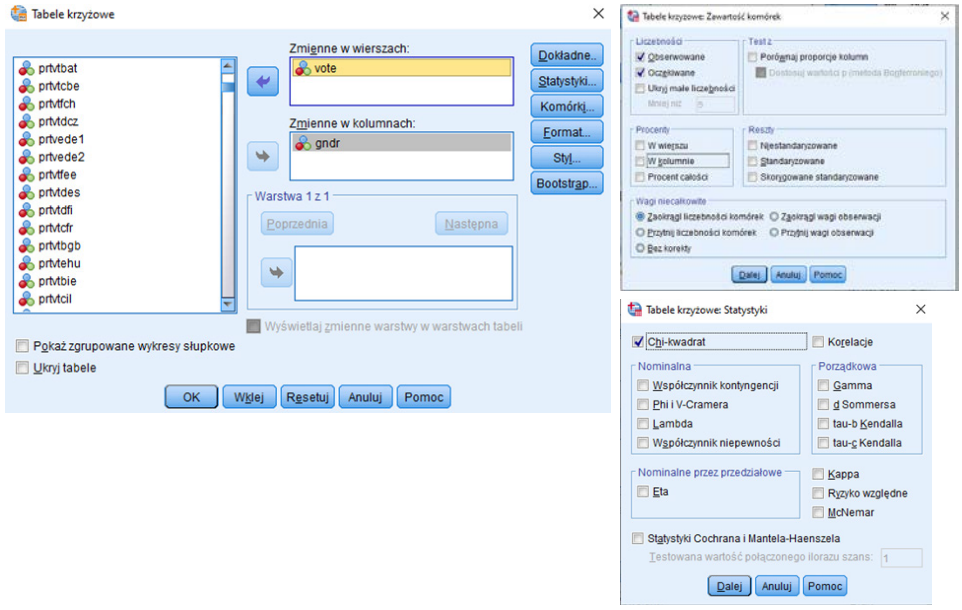
W ósmej rundzie badania ESS pytano między innymi, czy respondent brał udział w ostatnich krajowych wyborach. Dane, które poddamy analizie, dotyczą mieszkańców Polski uprawnionych do głosowania. Interesuje nas, czy między deklaracją uczestnictwa w wyborach (zmienna *vote*) a płcią (zmienna *gndr*) istnieje zależność istotna statystycznie.

**Rozwiązanie**

Obie z analizowanych tu zmiennych mierzone są na skali nominalnej. Każda z nich ma tylko dwie kategorie, otrzymamy zatem tabelę o wymiarach 2 × 2, która jest najprostszym wariantem tablicy kontyngencji. Aby skonstruować taką tabelę, a następnie na jej podstawie przeprowadzić test niezależności chi-kwadrat, w IBM SPSS Statistics wybieramy polecenie *Analiza → Opis statystyczny → Tabele krzyżowe*. Zasadniczo jest kwestią dowolną, którą zmienną potraktujemy jako wierszową, a którą jako kolumnową[16]. W tym podręczniku będziemy się trzymać reguły, by zmienną, która potencjalnie może warunkować drugą zmienną czy też stanowić zmienną grupującą (dzielącą zbiorowość na podpopulacje), umieszczać w kolumnach tabeli[17]. Tak więc zmienną informującą o płci respondenta wprowadzimy do kolumn, a do wierszy zmienną informującą o jego deklaracji uczestnictwa w wyborach (rysunek 6.1). Oprócz liczebności empirycznych SPSS może w tabeli również wyświetlić liczebności oczekiwane (a więc – przypomnijmy – liczebności spodziewane w sytuacji braku zależności). Aby skorzystać z tej możliwości, wystarczy pod przyciskiem *Komórki* zaznaczyć dodatkowo opcję *Oczekiwane*. Wyboru testu $\chi^2$ dokonujemy pod przyciskiem *Statystyki → Chi-kwadrat*. Z uwagi na fakt, że tabela krzyżowa ma wymiary 2 × 2, automatycznie wyświetlają się też wyniki dokładnego testu Fishera (dla większych tabel wyniki te nie będą generowane bez wybrania opcji *Dokładne* – więcej na ten temat w kolejnym podrozdziale).

16 Niekiedy można spotkać się z zaleceniem, by zmienną wyjaśniającą wprowadzać do wierszy, a zmienną wyjaśnianą do kolumn. Jak argumentuje Sawiński (2010, s. 75–76), taka konwencja rozmieszczenia cech w tabeli odpowiada naszym mechanizmom percepcyjnym, czyli ułatwia odbiór wyników.

17 Taka decyzja jest uzasadniona głównie tym, że w ramach procedury *Tabele krzyżowe* (oraz procedury *Tabele użytkownika*) program SPSS oferuje dodatkowo *Test proporcji kolumnowych z*. Zastosowanie testu *z* w SPSS zakłada, że zmienną wyjaśniającą umieścimy w kolumnie. Takie bowiem ustawienie umożliwia porównanie procentowych rozkładów zmiennej wyjaśnianej w poszczególnych kategoriach zmiennej wyjaśniającej. Czytelników zainteresowanych możliwością skorzystania z testu *z* w SPSS zachęcamy do lektury tekstu Saraty (b.r.).




**Rysunek 6.1.** Wykonywanie polecenia *Tabele krzyżowe*

Prześledźmy teraz otrzymane liczebności oczekiwane (rysunek 6.2). Czy w jakieś komórce ich wielkość jest mniejsza niż 5? Na szczęście taka sytuacja nie ma tu miejsca. Spełnione jest zatem założenie testu $\chi^2$. Zauważmy przy okazji, że pod tabelą zatytułowaną *Testy chi-kwadrat* SPSS wyświetla użytkownikowi komunikat, jaka część komórek nie spełnia omawianego założenia. W tym przypadku mamy komunikat „0,0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5", a więc w żadnej komórce tabeli krzyżowej założenie to nie zostało naruszone.

**Tabela krzyżowa Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? * Płeć**

| | | | Płeć | | |
|---|---|---|---|---|---|
| | | | Mężczyzna | Kobieta | Ogółem |
| Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? | Tak | Liczebność | 570 | 603 | 1173 |
| | | Liczebność oczekiwana | 559.8 | 613.2 | 1173.0 |
| | Nie | Liczebność | 185 | 224 | 409 |
| | | Liczebność oczekiwana | 195.2 | 213.8 | 409.0 |
| Ogółem | | Liczebność | 755 | 827 | 1582 |
| | | Liczebność oczekiwana | 755.0 | 827.0 | 1582.0 |

Wartość statystyki testowej (widniejącej w tabeli jako *Statystyka chi-kwadrat Pearsona*) wynosi 1,373 (przy $df = (2 - 1) \times (2 - 1) = 1$). To niewielka wartość, która świadczy o tym, że liczebności empiryczne nie różnią się zasadniczo od liczebności oczekiwanych. Z tym wynikiem koresponduje wysokie prawdopodobieństwo w teście niezależności

chi-kwadrat, które wynosi $p = 0{,}241$. W tej sytuacji formułujemy wniosek o braku podstaw do odrzucenia hipotezy zerowej, zależność nie jest zatem istotna statystycznie.

**Testy Chi-kwadrat**

| | Wartość | df | Istotność asymptotyczna (dwustronna) | Istotność dokładna (dwustronna) | Istotność dokładna (jednostronna) |
|---|---|---|---|---|---|
| Chi-kwadrat Pearsona | 1.373[a] | 1 | .241 | | |
| Poprawka na ciągłość[b] | 1.242 | 1 | .265 | | |
| Iloraz wiarygodności | 1.375 | 1 | .241 | | |
| Dokładny test Fishera | | | | .251 | .133 |
| Test związku liniowego | 1.372 | 1 | .241 | | |
| N ważnych obserwacji | 1582 | | | | |

a. 0.0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi 195.19.

b. Obliczone wyłącznie dla tabeli 2x2

**Rysunek 6.2.** Tabela krzyżowa i wyniki testu $\chi^2$: ocena związku między zmiennymi *vote* i *gndr*

Ponieważ niezależność oznacza identyczność proporcji (odsetków) ze względu na drugą zmienną (co można zobaczyć, oprocentowując liczebności oczekiwane „do wierszy" oraz „do kolumn"), to hipotezę zerową moglibyśmy też sformułować jako zdanie mówiące, że proporcje w porównywanych populacjach są równe[18]. Prześledźmy zatem rozkłady procentowe. W tym celu wróćmy do głównego okna procedury *Tabele*

18 Analizę zależności między dwiema zmiennymi, z których każda przyjmuje tylko dwie wartości, można przeprowadzić nie tylko za pomocą testu $\chi^2$, ale także za pomocą testu $z$, który służy porównaniu dwóch proporcji w grupach niezależnych. Między sprawdzianami w tych testach zachodzi następująca relacja: $\chi^2 = z^2$. Gdyby w naszym przypadku posłużyć się testem $z$, hipoteza zerowa mówiłaby, że proporcja deklarujących udział w wyborach jest taka sama wśród kobiet i wśród mężczyzn ($H_0$: $p_1 = p_2$). Test $z$ dopuszcza różne postacie hipotezy alternatywnej: bezkierunkową ($H_1$: $p_1 \neq p_2$) oraz jednokierunkową ($H_1$: $p_1 < p_2$ albo $H_1$: $p_1 > p_2$). Gdyby układ hipotez, jaki przyjmujemy w teście $\chi^2$ zapisać w sposób odpowiadający testowi $z$, to hipoteza alternatywna mogłaby brzmieć jedynie: $H_1$: $p_1 \neq p_2$, stąd też wartość $p$, którą odczytujemy z tabeli wynikowej SPSS, jest oznaczona jako „istotność dwustronna". Test dla proporcji można przeprowadzić również wtedy, gdy zmienna wierszowa ma więcej niż dwa warianty. Porównanie proporcji dokonywane jest wówczas dla każdego z wierszy osobno. Przykładowo: jeśli badamy zależność między preferowaną formą wypoczynku (zmienna wierszowa o wariantach: *wczasy*, *wycieczka objazdowa*, *camping*) a miejscem zamieszkania (zmienna kolumnowa o wariantach: *wieś*, *małe miasto*, średnie miasto, *duże miasto*), test $z$ pozwoli odpowiedzieć na pytanie: „Czy, a jeśli tak, to która populacja wyróżniona na podstawie miejscowości zamieszkania różni się istotnie od względem preferencji dotyczących wczasów?". Analogicznie dokonane zostanie porównanie dotyczące preferowania wycieczki objazdowej, a także – osobno – campingu. Aby przeprowadzić taką analizę, wybieramy *Analiza → Tabele krzyżowe → Komórki → Porównaj proporcje kolumnę → Dostosuj wartości p (metoda Bonferroniego)*. Ostatni element daje nam testowanie *post hoc*, jest zatem potrzebny, gdy zmienna kolumnowa ma więcej niż dwa warianty (jak w tym przykładzie).

*krzyżowe*. Pod przyciskiem *Komórki* wybierzmy *Procent → W kolumnie* i *Procent → W wierszu* (jeżeli chcemy mieć dwie oddzielne tabele, każdą z innym rodzajem oprocentowania, to procedurę przeprowadzamy „na dwa razy").

**Tabela krzyżowa Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? * Płeć**

% z Płeć

| | | Płeć | | |
|---|---|---|---|---|
| | | Mężczyzna | Kobieta | Ogółem |
| Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? | Tak | 75.5% | 72.9% | 74.1% |
| | Nie | 24.5% | 27.1% | 25.9% |
| Ogółem | | 100.0% | 100.0% | 100.0% |

**Rysunek 6.3.** Kolumnowo oprocentowane liczebności empiryczne w tablicy kontyngencji utworzonej dla zmiennych *vote* i *gndr*

Ponieważ zmienną *płeć* wprowadziliśmy do kolumn i zastosowaliśmy oprocentowanie „w kolumnie" („po kolumnach", „kolumnowe"), to w ostatniej kolumnie tabeli widzimy rozkład deklaracji uczestnictwa w ostatnich wyborach, obserwowany na poziomie całej próby, a w środku tabeli widzimy rozkłady tej zmiennej oddzielnie w grupie (próbie) kobiet i mężczyzn (rysunek 6.3). O rozkładach w grupie kobiet i w grupie mężczyzn powiemy, że są to rozkłady warunkowe zmiennej *vote* (przedstawiają one rozkład zmiennej *vote* w zależności od tego, jaką wartość przyjmie zmienna *gndr*).

W sytuacji niezależności obu zmiennych odsetek deklarujących uczestnictwo w wyborach powinien wynosić 74,1 zarówno w grupie kobiet, jak i w grupie mężczyzn, a więc tyle co w całej próbie. Innymi słowy, gdyby zmienne były wzajemnie niezależne, to rozkłady warunkowe byłyby takie same. Wyniki uzyskane w próbie pokazują, że odsetki głosujących nie są identyczne wśród kobiet i mężczyzn (tu mężczyźni nieznacznie częściej niż kobiety deklarują uczestnictwo). Niemniej jednak wynik testu $\chi^2$ okazał się nieistotny statystycznie, co oznacza, że obserwowana w próbie różnica 2,6 punktu procentowego mogłaby wystąpić, gdyby w populacji różnica ta faktycznie wynosiła zero. Innymi słowy, obserwowaną różnicę możemy przypisać losowej zmienności próbek.

**Tabela krzyżowa Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? * Płeć**

% z Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 201

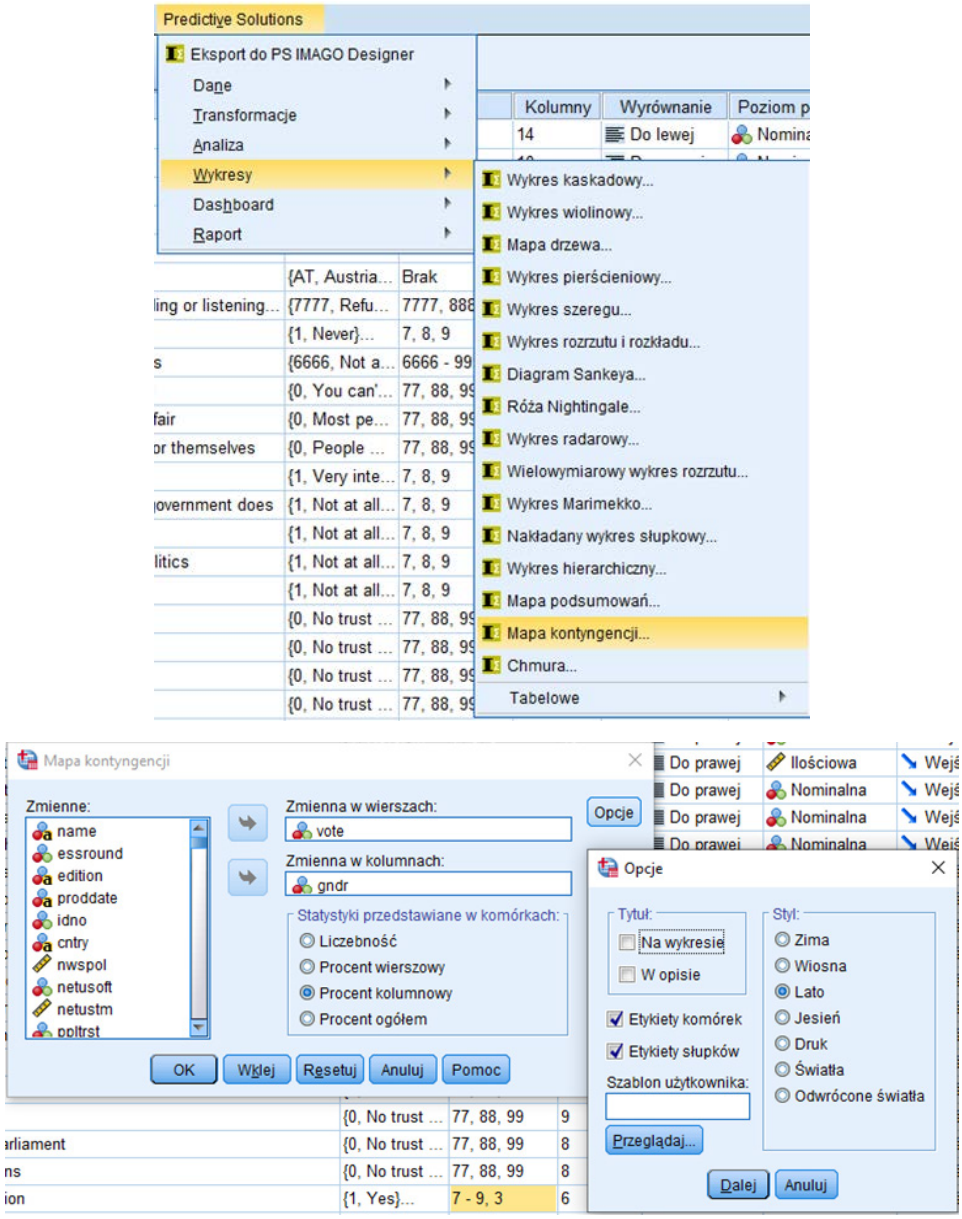
| | | Płeć | | |
|---|---|---|---|---|
| | | Mężczyzna | Kobieta | Ogółem |
| Czy brał(a) P. udział w ostatnich wyborach do Sejmu w październiku 2015 roku? | Tak | 48.6% | 51.4% | 100.0% |
| | Nie | 45.2% | 54.8% | 100.0% |
| Ogółem | | 47.7% | 52.3% | 100.0% |

**Rysunek 6.4.** Wierszowo oprocentowane liczebności empiryczne w tablicy kontyngencji utworzonej dla zmiennych *vote* i *gndr*

Podobnie moglibyśmy przeanalizować wyniki drugiej tabeli (rysunek 6.4), w której zastosowaliśmy procentowanie „do wierszy" („po wierszach", „wierszowe"). W ostatnim wierszu widzimy rozkład zmiennej *płeć* w całej próbie. W środku tabeli przedstawione są rozkłady warunkowe tej zmiennej – oddzielnie w grupie głosujących i w grupie niegłosujących. Widzimy, że udział mężczyzn jest nieco większy w grupie głosujących w porównaniu z grupą niegłosujących, co koresponduje z wcześniejszym ustaleniem. Różnica w odsetkach – jak już wiemy – jest nieistotna statystycznie.

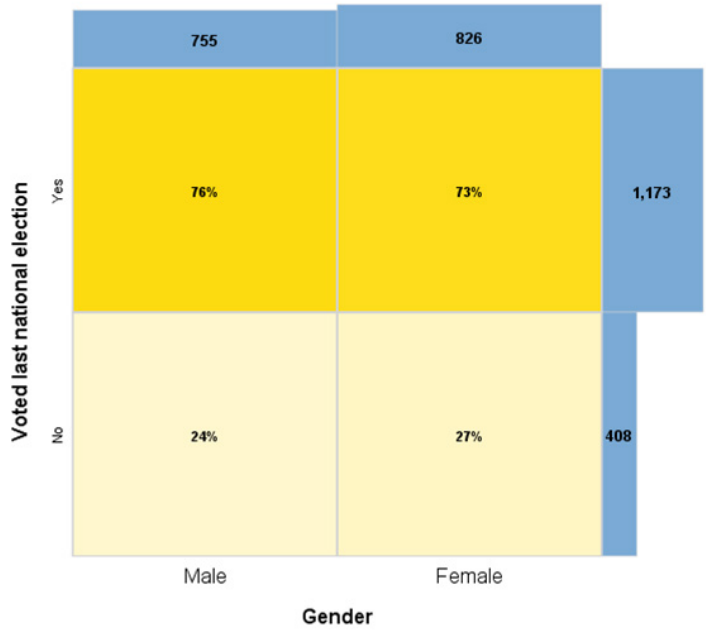
Przedstawianie wyników w postaci procentowych rozkładów warunkowych ma tę zaletę, że łatwo pozwala zorientować się w relacji, w jakiej pozostają ze sobą dwie zmienne. Daje też orientacyjny wgląd w siłę zależności między zmiennymi. Jeżeli chcemy prezentować wyniki w tej postaci, wystarczy, że skorzystamy z jednego sposobu oprocentowania. Jeżeli chcemy się dowiedzieć, czy rozkład deklaracji dotyczących uczestnictwa w wyborach zależy od płci (co w tym przypadku jest sensownym ujęciem tej zależności – płeć nie może przecież zależeć od uczestnictwa w wyborach), to powinniśmy przedstawić rozkłady zmiennej *vote* w grupach płci (rysunek 6.4). Ogólna zasada, którą możemy się kierować, mówi, że powinniśmy przedstawiać rozkłady badanego zjawiska w poszczególnych kategoriach zmiennej grupującej.

Jedna z nowszych funkcjonalności PS IMAGO pozwala zwizualizować wyniki tabeli kontyngencji w postaci tzw. mapy kontyngencji. W celu skorzystania z tej możliwości z menu wybieramy *Predictive solutions → Wykresy → Mapa kontyngencji* (rysunek 6.5).




**Rysunek 6.5.** Wykonywanie procedury *Mapa kontyngencji*

W efekcie wyboru opcji zgodnie z rysunkiem 6.5 otrzymamy wizualizację wyników odpowiadającą rysunkowi 6.6.




**Rysunek 6.6.** Mapa kontyngencji dla zmiennych *vote* i *gndr.* Wizualizacja odpowiada wynikom zaprezentowanym na rysunku 6.4.

Intensywność koloru w środku „wykresu" odpowiada wielkościom odsetków. Po bokach tabeli znajdują się słupki, które wysokością odpowiadają liczebnościom brzegowym tabeli. Na wykresie obecne są dodatkowo etykiety, które wprost informują o tych liczebnościach.

Wizualizacja odpowiada wynikom zaprezentowanym na rysunku 6.4. Intensywność koloru w środku „wykresu" odpowiada wielkościom odsetków. Po bokach tabeli znajdują się słupki, które wysokością odpowiadają liczebnościom brzegowym tabeli. Na wykresie obecne są dodatkowo etykiety, które wprost informują o tych liczebnościach.

**Przykład 6.2**

Posłużmy się ponownie danymi pochodzącymi z ósmej rundy badania ESS. Tym razem będziemy się interesować zależnością między wielkością miejsca zamieszkania jednostki (zmienna *domicil_recode*) a doświadczeniem bycia ofiarą włamania lub napadu w ciągu ostatnich pięciu lat przez jednostkę lub osobę z jej gospodarstwa domowego (zmienna *crmvct*). Zależność tę – podobnie jak poprzednio – będziemy analizować w odniesieniu do mieszkańców Polski.

### Rozwiązanie

Badamy zależność między dwiema zmiennymi jakościowymi:

- *crmvct* – bycia ofiarą włamania lub napadu w ciągu ostatnich pięciu lat przez jednostkę lub osobę z jej gospodarstwa domowego ($c = 2$); zmienna mierzona na skali nominalnej;
- *domicil_recode* – wielkość miejscowości zamieszkania ($k = 3$); zmienna mierzona na skali porządkowej.

Analiza powinna być zatem przeprowadzona z wykorzystaniem testu niezależności chi-kwadrat. Warunkiem będzie jednak to, że wszystkie liczebności oczekiwane są większe bądź równe 5. Analizując zapis pod tabelą *Testy chi-kwadrat* (rysunek 6.7), widzimy, że założenie to jest spełnione. Aby porównać mieszkańców różnych typów miejscowości pod względem doświadczania przemocy (*crmvct*), w tabeli krzyżowej oprócz liczebności warunkowej uwzględnimy procenty „w kolumnie".

**Tabela krzyżowa Czy P. lub ktoś z P. gospodarstwa domowego był ofiarą włamania lub napadu w ciągu ostatnich 5 lat? * miejsce zamieszkania**

| | | | miejsce zamieszkania | | | |
|---|---|---|---|---|---|---|
| | | | Duże miasto lub przedmieścia dużego miasta | średnie lub małe miasto | wieś | Ogółem |
| Czy P. lub ktoś z P. gospodarstwa domowego był ofiarą włamania lub napadu w ciągu ostatnich 5 lat? | Tak | Liczebność | 69 | 52 | 45 | 166 |
| | | % z miejsce zamieszkania | 16.9% | 9.9% | 6.0% | 9.9% |
| | Nie | Liczebność | 339 | 471 | 704 | 1514 |
| | | % z miejsce zamieszkania | 83.1% | 90.1% | 94.0% | 90.1% |
| Ogółem | | Liczebność | 408 | 523 | 749 | 1680 |
| | | % z miejsce zamieszkania | 100.0% | 100.0% | 100.0% | 100.0% |

**Testy Chi-kwadrat**

| | Wartość | df | Istotność asymptotyczna (dwustronna) |
|---|---|---|---|
| Chi-kwadrat Pearsona | 35.268[a] | 2 | .000 |
| Iloraz wiarygodności | 33.560 | 2 | .000 |
| Test związku liniowego | 34.344 | 1 | .000 |
| N ważnych obserwacji | 1680 | | |

a. 0.0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi 40.31.

**Rysunek 6.7.** Tabela kontyngencji i wyniki testu niezależności $\chi^2$ dla zmiennych *crmvct* i *domicil_recode*

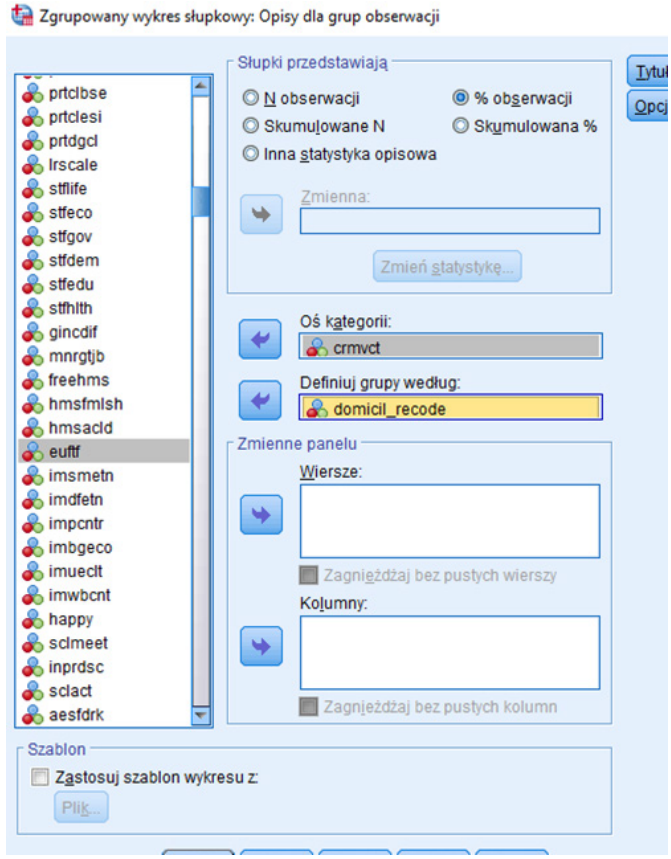
Zgodnie z wynikami (rysunek 6.7) odsetek osób, które doświadczyły włamania lub napadu (wobec siebie bądź osoby z gospodarstwa domowego), wynosi w całej próbie blisko 10%. Przechodząc teraz do analizy rozkładów warunkowych, widzimy, że odsetki osób z takim doświadczeniem różnią się w poszczególnych klasach miejscowości zamieszkania. Najwięcej osób w ten sposób pokrzywdzonych – według deklaracji samych badanych – jest wśród mieszkańców wielkich miast i osób żyjących na przedmieściach tych miast (17%). W przypadku osób zamieszkujących średnie i małe miasta odsetek ten jest mniejszy i odpowiada udziałowi, jaki obserwujemy na poziomie całego kraju (10%). Najmniejszy odsetek osób mających takie doświadczenie jest wśród mieszkańców wsi (6%). W celu zbadania, czy zmienne są ze sobą powiązane w populacji mieszkańców Polski, mamy prawo skorzystać z testu $\chi^2$, gdyż liczebności oczekiwane w każdej komórce tabeli są odpowiednio duże. Prawdopodobieństwo w teście niezależności chi-kwadrat jest bardzo niskie – $p < 0{,}001$, odrzucamy zatem $H_0$, a jednocześnie mamy bardzo mocne przesłanki, by przychylić się do hipotezy alternatywnej, głoszącej zależność. Wartości osiągniętych odsetków w próbie sugerują, że udział osób mających przykre doświadczenie bycia ofiarą włamania lub napadu – doznane osobiście lub przez domownika – zwiększa się wraz z wielkością miejscowości zamieszkania.

Zobaczmy teraz, jak moglibyśmy zilustrować uzyskane tu wyniki za pomocą wykresów słupkowych. Zaprezentujemy dwie możliwości – wykres słupkowy zgrupowany oraz wykres słupkowy zestawiony. Aby zaprezentować dostępne w IBM SPSS Statistics procedury, raz posłużymy się poleceniem *Wykresy tradycyjne*, a raz poleceniem *Kreator wykresów*.

Na początek z menu wybierzmy *Wykresy* → *Wykresy tradycyjne* → *Słupkowy*. Teraz musimy określić, jaki rodzaj wykresu słupkowego nas interesuje. Zgodnie z rysunkiem 6.8 wybierzmy opcję *Zgrupowany* oraz *Podsumowania dla grup obserwacji*.


**Rysunek 6.8.** Wywoływanie procedury *Wykresy słupkowe* za pomocą polecenia *Wykresy tradycyjne*

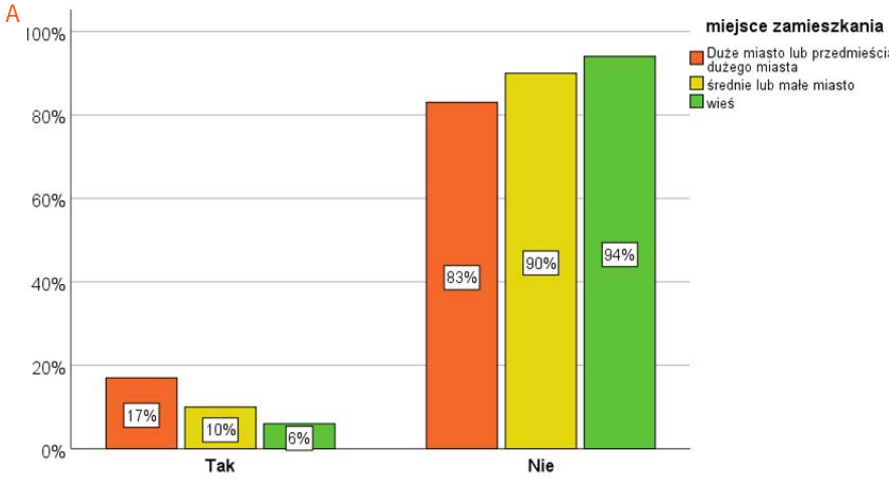
W kolejnym oknie (rysunek 6.9) definiujemy, jaką zmienną analizujemy w przekroju (według) której zmiennej. Ponieważ chcemy porównywać odsetki osób mających takie doświadczenie w poszczególnych klasach miejscowości zamieszkania, to zmienną *domicil_recode* – którą traktujemy jako zmienną grupującą – umieszczamy w polu *Definiuj grupy według*. Zmienną, którą traktujemy jako wyjaśnianą, czyli *crmvct*, wprowadzamy do pola *Oś kategorii*. Zaznaczamy jeszcze, że interesuje nas *% obserwacji*.



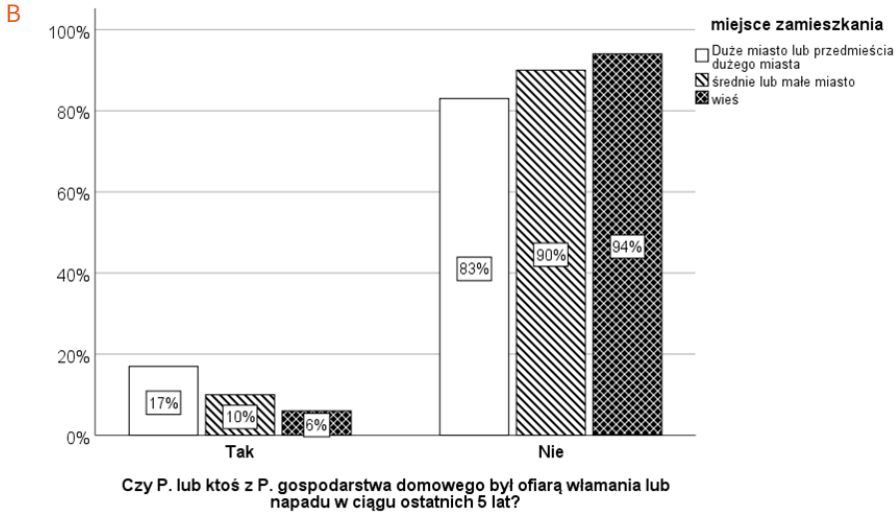

**Rysunek 6.9.** Główne okno procedury *Zgrupowany wykres słupkowy: Opisy dla grup obserwacji*

Wykres, który otrzymaliśmy w oknie raportowym, możemy poddać dalszej edycji w Edytorze wykresów, do którego uzyskujemy dostęp po dwukrotnym kliknięciu w wykres. Po naniesieniu na słupki etykiet informujących o wartościach odsetków (za pomocą *Tryb identyfikacji danych*) oraz po zmianie kolorów (deseni) słupków wykres zgrupowany może wyglądać w sposób prezentowany na rysunku 6.10 (wersja A lub B).

A




B

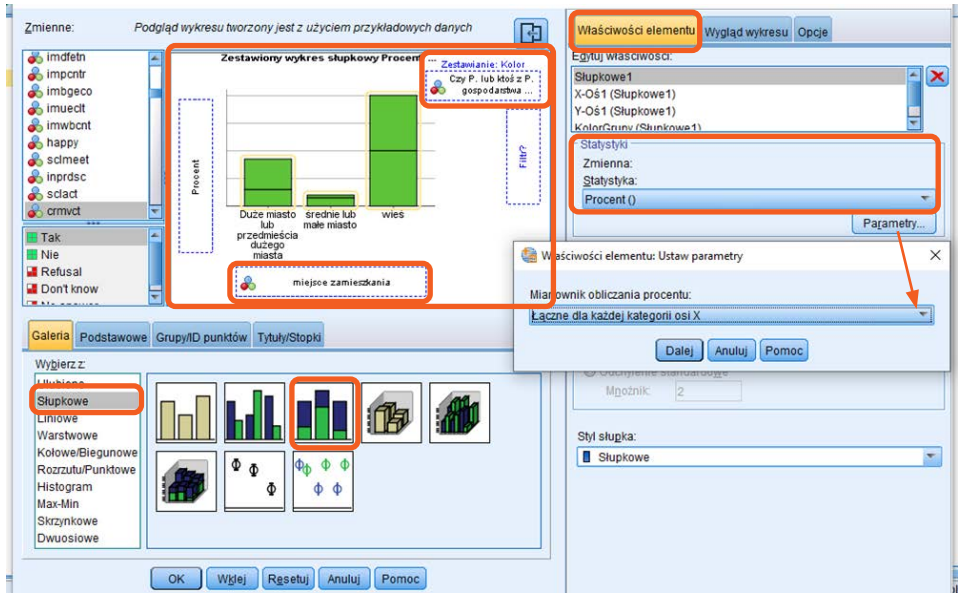



**Rysunek 6.10.** Wykres słupkowy zgrupowany dla zmiennych *domicil_recode* oraz *crmvct*

W celu skonstruowania wykresu słupkowego zestawionego wypróbujmy polecenie *Wykresy → Kreator wykresów*. Po wywołaniu tego polecenia dostaniemy komunikat, że zmienne, które chcemy użyć do budowy wykresu, powinny być poprawnie zdefiniowane, jeżeli chodzi o poziom pomiaru, gdyż od tego zależy, jakie opcje wykresu będziemy mieć udostępnione. W naszym przypadku obie zmienne zdefiniowano jako nominalne.

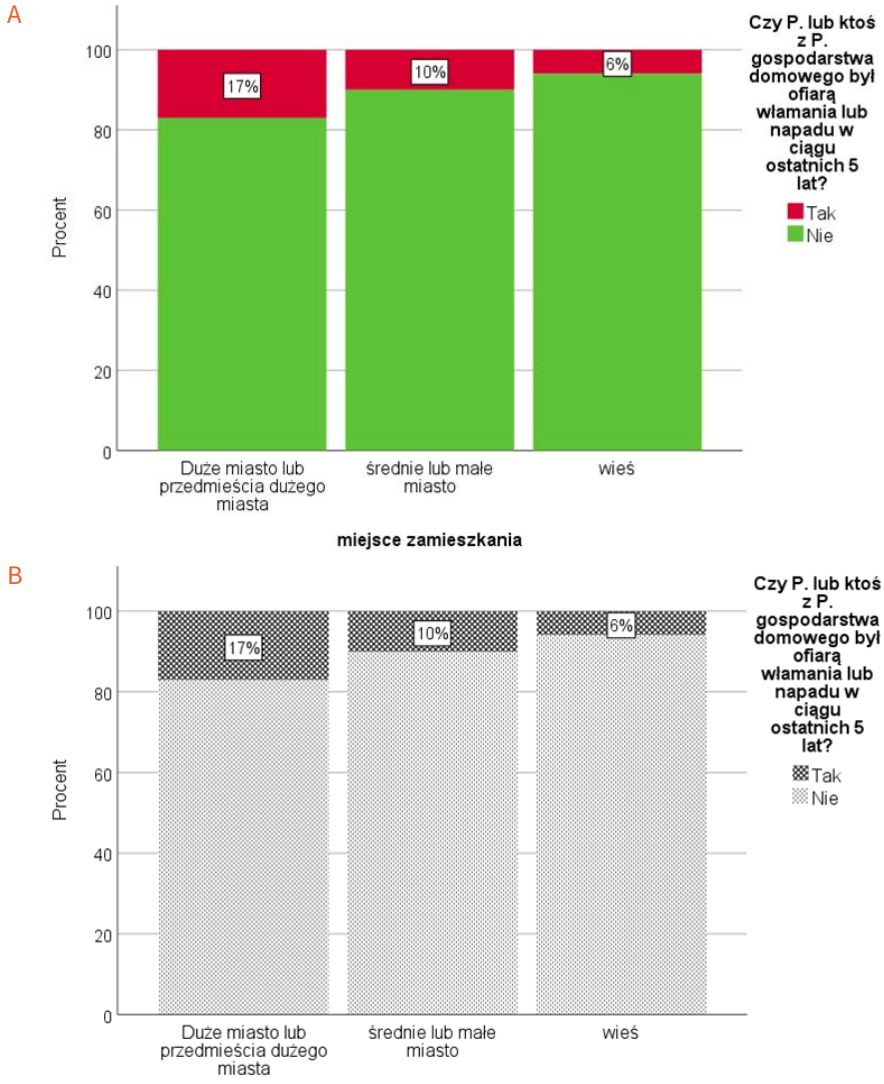
Po wyświetleniu się głównego okna procedury (rysunek 6.11) budowę wykresu rozpoczynamy od wskazania, że chcemy otrzymać wykres słupkowy (zaznaczamy tę opcję na liście znajdującej się w lewym dolnym rogu okna). Po dokonaniu tego wyboru

wyświetlą się ikony dostępnych wykresów słupkowych. Na ikony te możemy dodatkowo najeżdżać kursorem, by sprawdzić nazwy wykresów. Przypomnijmy, że nas interesuje wykres „zestawiony" (na rysunku 6.11 został oznaczony – na dole okna – pogrubioną linią). Po zlokalizowaniu ikony tego wykresu w galerii przeciągamy ją kursorem do obszaru roboczego wykresu (białe pole na górze). Z listy zmiennych znajdujących się w lewej górnej części okna wybieramy zmienną *domicil_recode* i kursorem przeciągamy ją do pola *Oś X*, a zmienną *crmvct* umieszczamy w polu *Zestawienie: kolor* (oba pola obwiedzione są owalnym zaznaczeniem). Teraz przechodzimy do ustawień w zakładce *Właściwości elementu*, którą znajdziemy w prawej górnej części okna. W panelu *Statystyki* wybieramy z listy opcję *Procent*, a pod przyciskiem *Parametry* dookreślamy, że właściwy *Mianownik obliczania procentu* to Łącznie dla każde kategorii osi X.




**Rysunek 6.11.** Główne okno procedury *Kreator wykresów*

Wykres, który dostaliśmy w oknie raportowym – podobnie jak poprzednio – możemy poddać dalszej edycji w *Kreatorze wykresów*. Efekt może być taki jak na rysunku 6.12 (A lub B).




**Rysunek 6.12.** Wykres słupkowy zgrupowany dla zmiennych *domicil_recode* oraz *crmvct*

**Przykład 6.3**

Przypuśćmy, że z bazy klientów firmy ubezpieczeniowej wybrano losowo niewielkie próby dla dwóch grup wieku (zmienna: *grupa_wieku*). Każdego klienta sklasyfikowano pod kątem tego, czy zgłaszał roszczenie do firmy (zmienna: *roszczenie*). Analiza ma na celu odpowiedzieć na pytanie, czy zgłaszanie roszczeń jest niezależne od grupy wieku.

**Rozwiązanie**

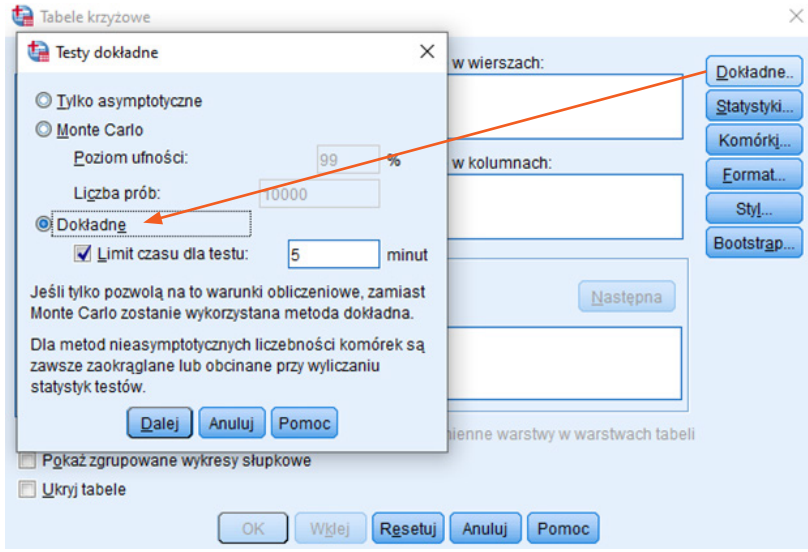
Badamy zależność między dwiema zmiennymi jakościowymi (dyskretnymi), przy czym jedna z nich – *grupa_wieku* – mierzona jest na skali porządkowej, a druga – *roszczenie* – na skali nominalnej. W szczególności interesuje nas rozstrzygnięcie, czy proporcja zgłaszających roszczenia jest taka sama, czy różna w grupie młodszych i starszych klientów, co zapiszemy w postaci:

$H_0$: $p_{do\ 30} = p_{30+}$

$H_1$: $p_{do\ 30} \neq p_{30+}$.

Sam problem badawczy nie wskazuje jeszcze jednoznacznie na zasadność wykorzystania dokładnego testu Fishera, niemniej jednak – jak się za chwilę przekonamy – faktycznie będzie to właściwa metoda analizy.

Początkowo różne pakiety statystyczne oferowały dokładny test Fishera tylko dla tabel 2 × 2, przez co jego zastosowanie ograniczało się w praktyce do tego typu tabel. Pakiet IBM SPSS Statistics standardowo udostępnia dokładny test Fishera dla tabel 2 × 2, a dzięki modułowi *Dokładne*, dostępnemu w głównym oknie procedury *Tabele krzyżowe*, możliwe jest przeprowadzenie tego testu dla każdej tabeli *r* × *c* (rysunek 6.13). Należy pamiętać, że wykonanie tego testu wymaga zaznaczenia również opcji *Chi-kwadrat* w polu *Statystyki*.




**Rysunek 6.13.** Wykonywanie dokładnego testu Fishera dla dowolnych tabel

**Tabela krzyżowa rozszczenie * grupa_wieku**

| | | | grupa_wieku | | |
|---|---|---|---|---|---|
| | | | do 30 lat | 30 lat i więcej | Ogółem |
| rozszczenie | tak | Liczebność | 5 | 0 | 5 |
| | | Liczebność oczekiwana | 3.0 | 2.0 | 5.0 |
| | | % z grupa_wieku | 41.7% | 0.0% | 25.0% |
| | nie | Liczebność | 7 | 8 | 15 |
| | | Liczebność oczekiwana | 9.0 | 6.0 | 15.0 |
| | | % z grupa_wieku | 58.3% | 100.0% | 75.0% |
| Ogółem | | Liczebność | 12 | 8 | 20 |
| | | Liczebność oczekiwana | 12.0 | 8.0 | 20.0 |
| | | % z grupa_wieku | 100.0% | 100.0% | 100.0% |

**Testy Chi-kwadrat**

| | Wartość | df | Istotność asymptotyczna (dwustronna) | Istotność dokładna (dwustronna) | Istotność dokładna (jednostronna) |
|---|---|---|---|---|---|
| Chi-kwadrat Pearsona | 4.444[a] | 1 | .035 | | |
| Poprawka na ciągłość[b] | 2.500 | 1 | .114 | | |
| Iloraz wiarygodności | 6.193 | 1 | .013 | | |
| Dokładny test Fishera | | | | .055 | .051 |
| Test związku liniowego | 4.222 | 1 | .040 | | |
| N ważnych obserwacji | 20 | | | | |

a. 50.0% komórek (2) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi 2.00.

b. Obliczone wyłącznie dla tabeli 2x2

**Rysunek 6.14.** Tablica kontyngencji oraz wyniki testów niezależności dla zmiennych *grupa_wieku* oraz *roszczenie*

Dane w tabeli kontyngencji (rysunek 6.14) wskazują, że udział zgłaszających roszczenia jest większy w grupie młodszych niż starszych klientów. Tymczasem IBM SPSS Statistics sygnalizuje nam, że w dwóch komórkach liczebność oczekiwana jest mniejsza niż 5. Gdybyśmy zlekceważyli fakt złamania założenia testu $\chi^2$ i mimo to posłużyli się nim, to uzyskalibyśmy $p = 0{,}035$ i w konsekwencji odrzucilibyśmy hipotezę zerową. Zobaczmy teraz, jakie rozstrzygnięcie przynosi dokładny test Fishera, którego zastosowanie jest poprawnym posunięciem. Skoro sformułowana hipoteza alternatywna jest bezkierunkowa ($\neq$), to interesuje nas wartość prawdopodobieństwa w teście dwustronnym. Wynosi ona $p = 0{,}055$, a więc przekracza założony poziom $\alpha = 0{,}05$, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej. Jak widać, niepoprawne posłużenie się testem $\chi^2$ prowadziłoby w tym przypadku do podjęcia

błędnej decyzji dotyczącej hipotezy zerowej. Zależność między badanymi zmiennymi nie jest statystycznie istotna. Można więc również powiedzieć, że odsetek klientów zgłaszających roszczenie nie różni się istotnie w grupach wieku.

W celach poglądowych jedynie przyjrzyjmy się, do jakiej konkluzji doprowadziłby nas wynik dokładnego testu Fishera, gdyby hipoteza alternatywna była jednokierunkowa, a konkretnie gdyby miała postać $H_1: p_{do\,30} > p_{30+}$. Uzyskalibyśmy wtedy $p = 0{,}051$, a ten wynik również nie uprawniałby nas do odrzucenia hipotezy zerowej (zauważmy, że w przypadku dokładnego testu Fishera $p$ w teście dwustronnym nie jest dwukrotnością $p$ w teście jednostronnym).

### 6.1.3. Miary siły zależności oparte na chi-kwadrat

Jak dotąd zajmowaliśmy się wnioskowaniem statystycznym – posługiwaliśmy się testem $\chi^2$ lub dokładnym testem Fishera, by ustalić, czy między badanymi cechami istnieje w populacji zależność. Powtórzmy przy okazji, że wnioskowanie statystyczne przeprowadzamy tylko wtedy, gdy o doborze jednostek do badania decyduje mechanizm losowy – tylko wtedy bowiem zasadne jest uogólnienie wyników poza przebadaną próbę. Jeżeli chodzi o kwestię tego, jak bardzo silna jest zależność między zmiennymi, to zwykle tym ustaleniem interesujemy się w drugiej kolejności (Szwed, 2008, s. 330–331). Nie trzeba jednak traktować tego jako wiążącej reguły, gdyż do mierników siły związku możemy też podejść jak do statystyk opisowych, traktując uzyskany wynik jako odnoszący się jedynie do przebadanych jednostek. Tylko do nich się odniesiemy, jeśli próba nie będzie pozwalała na uogólnienia na populację, albo gdy przebadano całą populację. Natomiast w tych przypadkach, w których przeprowadzenie wnioskowania statystycznego jest zasadne, zainteresowanie kwestią siły związku jest między innymi potrzebne dlatego, że może się okazać, że choć zależność jest istotna statystycznie, to związek jest bardzo słaby – różnice w proporcjach są na tyle małe, że pozbawione praktycznego znaczenia. Miary te odpowiadają więc (pod względem celu ich zastosowania) metodom oceny wielkości efektu w analizie wariancji.

Opracowano wiele mierników siły związku dla zmiennych nominalnych oraz dla zmiennych porządkowych. Dużo możliwości w tym zakresie oferuje IBM SPSS Statistics. Wielość miar wynika po części z tego, że każda z nich ma swoje ograniczenia i nie spełnia wszystkich oczekiwań dobrego miernika siły związku. W tym podręczniku poprzestaniemy na przedstawieniu współczynników opartych na $\chi^2$, które stosuje się w badaniu zależności między dwiema zmiennymi mierzonymi na skali nominalnej bądź między zmiennymi, z których jedna mierzona jest na skali nominalnej, a druga na porządkowej[19].

19 Oprócz tych miar IBM SPSS Statistics udostępnia także mierniki proporcjonalnej redukcji błędu (PRE) – współczynnik tau Goodmana i Kruskala, lambda, współczynnik niepewności,

Jak już widzieliśmy, wysokie wartości statystyki $\chi^2$ sprzyjają niskim wartościom $p$ i odrzuceniu hipotezy zerowej. W takim razie może powstać pytanie, czy sam współczynnik $\chi^2$ mógłby informować o sile związku. Odpowiedź jest negatywna, gdyż wartość $\chi^2$ zależy nie tylko od siły związku, ale także od liczebności próby. Zwróćmy bowiem uwagę, że gdybyśmy porównywali proporcje 51/100 i 49/100, to $\chi^2 = 0{,}08$, a $p = 0{,}7773$. Natomiast dla pary proporcji 510/1000 i 490/1000 wartość $\chi^2 = 0{,}8$, a $p = 0{,}3711$. W obu przypadkach rozkłady procentowe są identyczne. Przykład pokazuje, że mimo iż siła związku jest taka sama, statystyka $\chi^2$ reaguje na wielkość próby, a co za tym idzie – zmienia się również wartość $p$. **Miara** $\phi$ [czytaj: fi] przezwycięża ten problem w następujący sposób:

$$\phi = \sqrt{\frac{\chi^2}{n}}. \tag{30}$$

To rozwiązanie powoduje, że wartości miernika leżą w zakresie [0, 1], niemniej jest to prawdą tylko dla tabel 2 × 2. W tabelach większych – $r \times c$ – miara może przekroczyć 1, co czyni ją nieużyteczną do stosowania. Miara $\phi$ ma tę ważną własność, że odpowiada modułowi wartości współczynnika korelacji $r$ Pearsona (korelacji między dwiema zmiennymi dychotomicznymi). Stąd $\phi^2$ będzie informować o tym, jaką część zróżnicowania jednej zmiennej można wyjaśnić za pomocą drugiej zmiennej (Blalock, 1975, s. 258).

**Miernik** $V$ **Cramèra** przyjmuje wartości z zakresu [0, 1] niezależnie od rozmiaru tabeli. Wzór na $V$ wygląda następująco:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}, \tag{31}$$

gdzie $k$ to liczba kolumn bądź wierszy, w zależności od tego, która jest mniejsza. Konstrukcja miary wykorzystuje fakt, że maksymalna wartość, jaką może przyjąć $\chi^2$, równa jest $n(k - 1)$, a zatem, jeśli $\chi^2 = n(k - 1)$, to $V = 1$ (Hershberger, Fisher, 2005, s. 1183–1192). Zauważmy, że dla tabel 2 × 2 $\phi = V$ .

---

jak również metody oceny siły związku dla skali porządkowej – gamma, współczynnik d-Sommersa, współczynnik tau-b Kendalla, współczynnik tau-c Kendalla. Z tego poziomu możliwe jest również wyznaczenie współczynnika eta czy też ryzyka względnego. Mierniki te nie zostaną w tym miejscu scharakteryzowane. Bardzo dobre i przystępne omówienie wszystkich miar proponowanych przez SPSS Czytelnik znajdzie w książce Górniaka i Wachnickiego (2000).

Jeszcze inny pomysł przezwyciężenia ograniczenia $\phi$ przedstawia **miernik** $C$, czyli **współczynnik kontyngencji**, który jest określony wzorem:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{32}$$

Rozwiązanie to powoduje, że wprawdzie w tabeli $r \times c$ miernik ten nie przekroczy 1, ale – jak pokazuje konstrukcja mianownika – nigdy też nie osiągnie 1, gdyż $n \neq 0$. W efekcie jego wartości leżą w zakresie [0, 1). Drugim ograniczeniem współczynnika kontyngencji jest to, że jego wartości zależą od liczby kolumn i wierszy w tabeli $r \times c$. Zgodnie ze wzorem:

$$C_{\max} = \sqrt{\frac{k-1}{k}} \tag{33}$$

maksymalna wartość $C$ w tabeli, w której minimum z liczby kolumn i wierszy wynosi 2 (a więc w tabeli 2 × 2, tabeli 2 × 3 itd.) to 0,707. Z kolei $C_{\max}$ w tabeli 3 × 3, tabeli 3 × 4 itd. wynosi 0,816. Tę niedogodność można przezwyciężyć, stosując następującą korektę:

$$C_{skoryg} = \frac{C}{C_{\max}}. \tag{34}$$

Korekta ta powoduje, że w przypadku idealnej zależności $C_{skoryg} = 1$ (Hershberger, Fisher, 2005, s. 1183–1192).

Dodajmy, że wszystkie omówione tu miary są symetryczne. To znaczy, że na ich wartość nie wpływa to, czy będziemy rozpatrywać wariant, w którym Y zależy od X, czy wariant, w którym X zależy od Y. Siła związku między X i Y jest taka sama jak między Y i X. Pozostaje jeszcze kwestia interpretacji otrzymanego rezultatu. Wprawdzie wiadomo, że zero świadczy o braku związku, jedność o związku pełnym, ale jak oceniać wartości pośrednie? Tu pomocnych wskazówek dostarczył Cohen (1988), który wprowadził $w$ jako uniwersalny miernik wielkości efektu, który również oparty jest na $\chi^2$. Cohen przypisał następujące interpretacje wartościom $w$:

$w = 0{,}1$ – mała wielkość efektu,
$w = 0{,}3$ – średnia wielkość efektu,
$w = 0{,}5$ – duża wielkość efektu.

Cohen zdefiniował relację między $w$ a każdą z miar opartą na $\chi^2$:

$$\text{dla } \phi: \; w = \phi, \tag{35}$$

$$\text{dla } C: \; w = \sqrt{\frac{C^2}{1 - C^2}}, \tag{36}$$

$$\text{dla } V: \; w = V \times \sqrt{k - 1}. \tag{37}$$

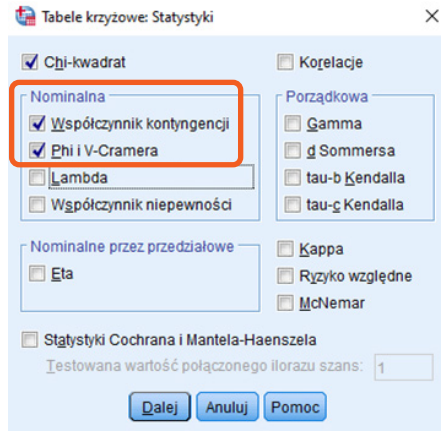
Maksymalna wartość, jaką może przyjąć $w$, wynosi $\sqrt{k-1}$, przy czym $k$ – przypomnijmy – to liczba wierszy bądź kolumn, w zależności od tego, która jest mniejsza (Cohen, 1988, s. 222–227). Powyższe przekształcenie pokazuje, że wartość współczynnika $V$ Cramèra rzędu – przykładowo – 0,2 w tabeli 2 × 2, 2 × 3, 2 × 4 itd. powinna być interpretowana inaczej niż dla tabel 5 × 5, 5 × 6, 5 × 7 itd. W pierwszym przypadku wielkość efektu jest bowiem równa $w = 0{,}2$, a w drugim przypadku $w = 0{,}4$ (efekt jest silniejszy). Dodajmy, że zaproponowane przez Cohena interpretacje wielkości efektu powinniśmy traktować jako orientacyjne, a ostateczną ocenę uzależnić od kontekstu. Ten kontekst może być nadany przez problem, który badamy, czy przez oczekiwania, jakie wynikają z poprzednich badań w danej dziedzinie.

**Przykład 6.4**

Wróćmy do przykładu, w którym interesowaliśmy się zależnością między deklaracją uczestnictwa w wyborach i płcią. Sprawdźmy siłę związku między tymi zmiennymi.

**Rozwiązanie**

Korzystając z polecenia *Tabele krzyżowe*, wybierzmy ponownie przycisk *Statystyki*, by zaznaczyć interesujące nas miary siły związku – w bloku *Nominalne* zaznaczamy *Współczynnik kontyngencji* i *Phi i V-Cramèra* (rysunek 6.15). Wyniki zestawiono na rysunku 6.16.




**Rysunek 6.15.** Wyznaczanie współczynników zależności opartych na statystyce chi-kwadrat

**Miary symetryczne**

| | | Wartość | Istotność przybliżona |
|---|---|---|---|
| Nominalna przez Nominalna | Phi | 0,029 | 0,241 |
| | V Kramera | 0,029 | 0,241 |
| | Współczynnik kontyngencji | 0,029 | 0,241 |
| N ważnych obserwacji | | 1582 | |

**Rysunek 6.16.** Wartości miar siły związku dla zmiennych *gndr* oraz *vote*

Zwróćmy na początek uwagę, że wartość *p* jest taka sama dla wszystkich mierników i odpowiada wielkością wartości *p*, jaką uzyskaliśmy, poddając testowi hipotezę o niezależności zmiennych (rysunek 6.3). Jest to zrozumiałe, bo wszystkie uwzględnione miary siły związku są oparte na $\chi^2$. Ta wysoka wartość *p* nie pozwoliła nam uznać, że w populacji istnieje związek między tymi zmiennymi. Przyjrzyjmy się teraz wynikom miar siły związku. Mamy do czynienia z tabelą 2 × 2, a więc każda z omówionych miar może być zastosowana. Każda zresztą daje ten sam rezultat. Wynik jest bliski zera, co świadczy o bardzo słabym związku. To ustalenie współgra z naszym wcześniejszym ustaleniem będącym efektem porównania rozkładów warunkowych (rysunek 6.4.). Jak pamiętamy, $\phi$ po podniesieniu do kwadratu umożliwia interpretację mówiącą, jaki procent zmienności pierwszej zmiennej można wyjaśnić zmiennością drugiej zmiennej. W naszym przypadku $\phi^2 = 0{,}029^2 = 0{,}000868$, co oznacza, że zmienna *płeć* wyjaśnia niespełna 0,1% zmienności zmiennej *vote*.

**Przykład 6.5**

Wróćmy do przykładu, w którym badaliśmy relację między wielkością miejscowości zamieszkania (trzy warianty) a doświadczeniem napadu lub włamania (dwa warianty). Sprawdźmy siłę związku między tymi zmiennymi.

**Rozwiązanie**

Dane analizowaliśmy w tabeli 2 × 3 (rysunek 6.9), właściwą miarą siły związku będzie zatem miernik *V* lub *C*. Wartości tych mierników są bardzo podobne (rysunek 6.17). Wybierając ostatecznie *V* jako miernik siły związku do analizowanych tu danych, otrzymujemy $w = V \times \sqrt{k-1} = 0{,}145 \times \sqrt{2-1} = 0{,}145$ , co pozwala nam zinterpretować efekt wpływu miejscowości zamieszkania na doświadczenie napadu lub włamania jako słaby.

**Miary symetryczne**

| | | Wartość | Istotność przybliżona |
|---|---|---|---|
| Nominalna przez Nominalna | Phi | .145 | .000 |
| | V Kramera | .145 | .000 |
| | Współczynnik kontyngencji | .143 | .000 |
| N ważnych obserwacji | | 1680 | |

**Rysunek 6.17.** Wartości miar siły związku dla zmiennych *crmvct* i *domicil_recode*

Odnosząc się jeszcze do przykładu 6.3, dodajmy, że problematyczne jest zastosowanie miar opartych na $\chi^2$ w sytuacji, gdy zależność bada się za pomocą dokładnego testu Fishera (Szymczak, 2018, s. 152).

## 6.2. Badanie zależności między dwiema zmiennymi ilościowymi

Zajmiemy się teraz badaniem zależności między dwiema zmiennymi $X$ i $Y$ mierzonymi na skali ilościowej. Popularnym miernikiem siły związku między takimi cechami jest **współczynnik korelacji liniowej Pearsona** (*r*). Jak sygnalizuje nazwa, posłużenie się nim zakłada, że zależność ma charakter liniowy. Liniowy związek oznacza sytuację, w której jednostkowym przyrostom jednej cechy towarzyszy – średnio rzecz biorąc – stały przyrost lub stały spadek wartości drugiej cechy (Starzyńska, 2009, s. 166). Podajmy przykłady. Wzrostowi odsetka użytkowników Internetu w danym kraju towarzyszy wzrost odsetka użytkowników Facebooka,

O takiej zależności powiemy, że jest dodatnia. Wzrostowi ciężaru roweru towarzyszy – średnio rzecz biorąc – spadek jego ceny. O takiej zależności powiemy, że jest ujemna.

Wartość współczynnika korelacji *r* możemy wyznaczyć za pomocą wzoru (Agresti, Franklin, 2013, s. 106):

$$r = \frac{1}{(n-1)} \sum \left( \frac{x - \overline{x}}{s_x} \right) \left( \frac{y - \overline{y}}{s_y} \right) = \frac{1}{(n-1)} \sum Z_x Z_y. \tag{38}$$

Według tej formuły dla każdej obserwacji sprawdzamy, o ile odchyleń standardowych jej wynik ze względu na cechę *X* odchyla się od średniej i – analogicznie – o ile odchyleń standardowych jej wynik ze względu na cechę *Y* odchyla się od średniej. Innymi słowy, dla każdej obserwacji uwzględniamy jej wyniki standaryzowane, czyli pozbawione mian, w jakich pierwotnie były wyrażone. Badając pod tym kątem każdą obserwację w zbiorze, jesteśmy później w stanie ustalić kształt ogólnej prawidłowości w danych – czy jest tak, że dodatnim odchyleniom od średniej dla cechy *X* towarzyszą dodatnie, czy może ujemne odchylenia od średniej dla cechy *Y*, a zatem czy zależność jest dodatnia, czy ujemna.

W podręcznikach statystyki znajdziemy też często następującą formułę (Sobczyk, 1998, s. 207–208; Starzyńska, 2009, s. 163):

$$r = \frac{\operatorname{cov}(x, y)}{S_x S_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}. \tag{39}$$

Najważniejsze własności współczynnika korelacji *r* Pearsona są następujące (Górniak, Wachnicki, 2000, s. 163; Starzyńska, 2009, s. 166–167):

- jego wartości leżą w zakresie [–1, 1];
- jest miarą niemianowaną; dzięki tej własności uprawnione jest porównywanie siły zależności dla różnych par zmiennych;
- pozwala na określenie siły związku – im *r* bliższe jest jedności co do wartości bezwzględnej, tym silniejszy związek; wartości bliskie zera świadczą o braku zależności (zakładając, że relację między zmiennymi opisuje linia prosta);
- pozwala na określenie kierunku związku – dodatnie wartości *r* świadczą o zależności dodatniej (wzrostowi poziomu jednej zmiennej towarzyszy – średnio rzecz biorąc – wzrost drugiej), a ujemne o zależności ujemnej (wzrostowi poziomu jednej zmiennej towarzyszy – średnio rzecz biorąc – spadek drugiej);

- jest miarą symetryczną – siła i kierunek zależności będą takie same, gdy będziemy rozpatrywać sytuację, w której *Y* zależy od *X*, jak i wtedy gdy będziemy rozpatrywać relację odwrotną;
- podniesiony do kwadratu, a więc $r^2$, informuje, jaka część zmienności jednej zmiennej jest wyjaśniona przez drugą zmienną (jaka część wariancji jest wspólnie dzielona przez zmienne);
- im bliżej linii regresji znajdują się dane, tym związek silniejszy i *r* zbliża się do jedności; im dalej dane leżą od linii regresji, tym słabszy związek i *r* zbliża się do zera;
- jest wrażliwy na obserwacje odstające, leżące daleko od pozostałych w zbiorze;
- jeżeli usuniemy ze zbioru obserwacje o najniższych lub najwyższych wartościach, które jednocześnie pasują do linii regresji i nie zmieniają jej nachylenia, to *r* zmniejszy swoją wartość.

Za Starzyńską (2009, s. 167) podajemy, jak orientacyjnie oceniać wartości pośrednie przyjmowane przez współczynnik korelacji. Zależność jest:

- niewyraźna, jeśli $|r| \leq 0,2$;
- wyraźna, ale niska, jeśli $0,2 < |r| \leq 0,4$;
- umiarkowana, jeśli $0,4 < |r| \leq 0,7$;
- znacząca, jeśli $0,7 < |r| \leq 0,9$;
- bardzo silna, jeśli $|r| > 0,9$.

Przedziały te trudno jednak odnosić do prób przekrojowych. Dla danych indywidualnych rzadko osiąga się współczynnik korelacji przekraczający (co do wartości bezwzględnej) 0,5, a tym samym jeśli jest bliski 0,5, mówi się już o dość silnym związku między zmiennymi (Cohen, 1988, s. 78)[20].

O ile jesteśmy zainteresowani rozstrzygnięciem, czy zależność między badanymi cechami występuje w populacji, a przy tym spełnione są założenia potrzebne do przeprowadzenia wnioskowania statystycznego (losowa próba, normalność rozkładów obu zmiennych), to współczynnik korelacji *r* wykorzystujemy jako estymator współczynnika korelacji w populacji $\rho$ (czytaj: ro).

Układ hipotez przyjmuje wtedy postać:
$H_0$: $\rho = 0$ (brak korelacji w populacji)
$H_1$: $\rho \neq 0$ (korelacja występuje w populacji).

Hipoteza alternatywna może też zostać doprecyzowana co do kierunku zależności – można sformułować wówczas hipotezę jednostronną $H_1$: $\rho < 0$ (występuje korelacja ujemna) albo $H_1$: $\rho > 0$ (występuje korelacja dodatnia). Podejście to jest jednak znacznie rzadziej stosowane.

20 Zobacz też propozycje Góralskiego (1974, s. 34 za Szwed, 2008, s. 313) oraz Bedyńskiej i Brzezickiej (2007, s. 96), które zostały przygotowane z myślą o wykorzystaniu w badaniach psychologicznych.

Sprawdzianem testu jest statystyka $t$ określona wzorem:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \tag{40}$$

która – gdy prawdziwa jest hipoteza zerowa – ma w przybliżeniu rozkład $t$ o $df = n - 2$ stopniach swobody (Agresti, Finlay, 2014, s. 282).

Przyjmując $\alpha = 0{,}05$, wnioskowanie przeprowadzamy według reguły:

- jeżeli $p < 0{,}05$, to odrzucamy hipotezę zerową i stwierdzamy, że są podstawy do przyjęcia hipotezy alternatywnej; korelacja jest zatem istotna statystycznie;
- jeżeli $p > 0{,}05$, to stwierdzamy brak podstaw do odrzucenia hipotezy zerowej; korelacja nie jest zatem istotna statystycznie.

W sytuacji, w której nie są spełnione założenia umożliwiające posłużenie się współczynnikiem $r$ Pearsona lub gdy poziom pomiaru przynajmniej jednej z dwóch zmiennych jest porządkowy, a nie ilościowy, należy posłużyć się **współczynnikiem korelacji rang Spearmana** ($r_s$ lub *rho*), który – w odróżnieniu od $r$ Pearsona – jest statystyką nieparametryczną. Statystyka $r_s$ jest obliczana na podstawie wzoru $r$ Pearsona, z tym że na danych, które zostały poddane rangowaniu (Field, 2009, s. 180). Nadmieniamy tylko, że w sytuacji, w której obliczenia wykonywane są samodzielnie, $r_s$ wyznacza się nieco inaczej, a do dyspozycji badacza są dwa wzory, z których jeden stosuje się w sytuacji braku rang wiązanych, a drugi, gdy takie wystąpiły[21].

Ważną charakterystyką współczynnika korelacji rang Spearmana jest to, że służy on do oceny zależności monotonicznej (zależność jest rosnąca, gdy wraz ze wzrostem wartości jednej cechy, rosną wartości drugiej cechy, zależność jest malejąca, gdy wraz ze wzrostem wartości jednej cechy, maleją wartości drugiej cechy). Zależność monotoniczna może się realizować jako zależność liniowa (średni przyrost $Y$ jest taki sam, gdy wartości $X$ są małe oraz wtedy gdy są duże) albo krzywoliniowa monotoniczna (przykładowo – średni przyrost $Y$ może być mały przy niskich wartościach $X$ i zwiększać się wraz ze wzrostem $X$). Warto pamiętać o następującej prawidłowości: jeżeli cechy są liniowo związane, to wartości $r$ i $r_s$ będą podobne. Jeżeli zaś cechy są związane krzywoliniowo, ale monotonicznie, to $r_s > r$ (Szymczak, 2018, s. 160–161).

Przeprowadzając wnioskowanie statystyczne, $r_s$ wykorzystujemy jako estymator korelacji rang w populacji $\rho_s$. Analogicznie jak w przypadku współczynnika korelacji liniowej Pearsona układ hipotez przyjmie postać:

21 Zainteresowanym polecamy lekturę podręcznika Szymczaka (2018).

$H_0$: $\rho = 0$
$H_1$: $\rho \neq 0$ (lub ewentualnie: $H_1$: $\rho < 0$ albo $H_1$: $\rho > 0$).

Sprawdzianem testu jest statystyka $t$ określona wzorem:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \tag{41}$$

która przy założeniu prawdziwości $H_0$ ma w przybliżeniu rozkład $t$ o $df = n - 2$ stopniach swobody (Szymczak, 2018, s. 159).

Drugą nieparametryczną statystyką jest **współczynnik korelacji rang Kendalla –** $\tau$ (tau). Jego zastosowanie jest polecane w przypadku małych prób, w których występuje duża liczba rang wiązanych (Field, 2009, s. 181), co – mówiąc prościej – oznacza sytuację, w której wiele obserwacji w zbiorze ma taki sam wynik. Efekt taki wystąpi, gdy – przykładowo – cecha będzie mierzona na pięciostopniowej skali porządkowej. Skrótowo przedstawiając tę statystykę, dodajmy jeszcze, że $\tau$ – tak jak $r$ oraz $r_s$ – przyjmuje wartości z zakresu [–1, 1]. Pamiętajmy także, że podczas kiedy $r$ i $r_s$ użyte do analizy tych samych danych ilościowych przyniosą bardzo podobny rezultat (zakładając liniowość związku), wynik $\tau$ będzie mniejszy o 66–75%. Jest to konsekwencja wynikająca z konstrukcji $\tau$ i trzeba o tym pamiętać, oceniając wielkość efektu (Field, 2009, s. 193).

**Przykład 6.6**

W przykładzie wykorzystamy dane *General Social Survey*, które zebrano na losowej próbie mieszkańców USA w 2018 roku. Rozpatrzmy zależność między statusem socjoekonomicznym (w skrócie SES) respondenta (zmienna *SEI10*) a statusem socjoekonomicznym współmałżonka (zmienna *SPSEI10*). Status socjoekonomiczny jest mierzony za pomocą indeksu, którego zakres wynosi [0, 100]. Zależność tę będziemy analizować dla osób młodych, mających co najwyżej 35 lat. Chcemy dowiedzieć się, czy zależność występuje w populacji młodych Amerykanów, a także jak jest silna.

**Rozwiązanie**

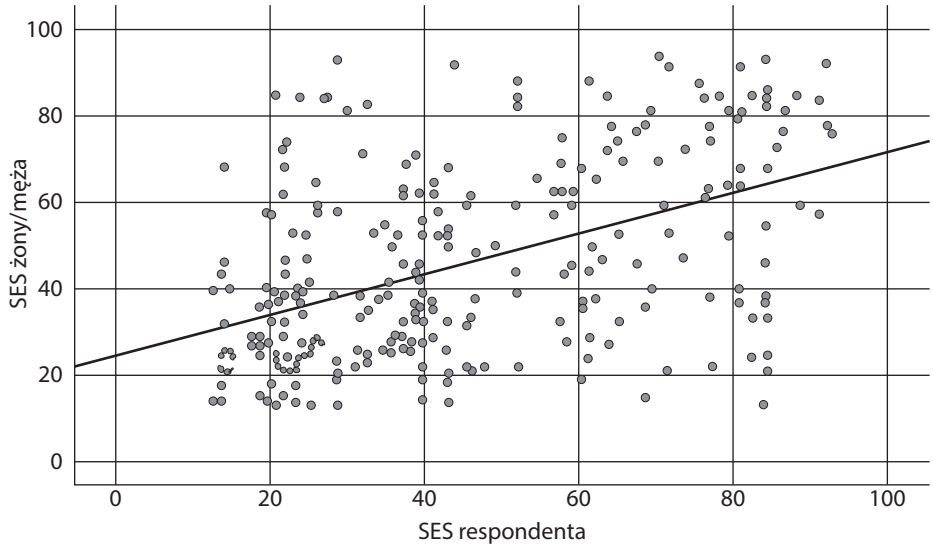
Badamy zależność między dwiema zmiennymi mierzonymi na skali ilościowej. Interesujemy się zatem rozstrzygnięciem, któremu odpowiada następujący układ hipotez: $H_0$: $\rho = 0$ *versus* $H_1$: $\rho \neq 0$. Hipotezy te najlepiej byłoby zweryfikować metodą parametryczną, tj. testem istotności dla współczynnika korelacji liniowej Pearsona. Analizę rozpoczynamy od diagnostyki (sprawdzenia, czy współczynnik $r$ może być tu zastosowany). To, czy zależność między interesującymi nas cechami rzeczywiście opisuje

linia prosta (a nie krzywa), można najprościej sprawdzić, wizualizując dane za pomocą wykresu rozrzutu. Wykres ten pozwoli również sprawdzić, czy w zbiorze występują obserwacje odstające oraz czy obserwacje są równomiernie rozłożone po obu stronach linii regresji (czyli czy *Y* ma podobną zmienność w poszczególnych wartościach *X*).

Z menu wybieramy *Wykresy* → *Wykresy tradycyjne* → *Rozrzutu* / *Punktowy* → *Prosty*. W głównym oknie procedury zmienna *SEI10* została umieszczona w polu *Oś X*, a zmienna *SPSEI10* w polu *Oś Y*.

Na wykresie rozrzutu (diagramie korelacyjnym) jedna kropka (punkt) reprezentuje jedną obserwację (jednego respondenta). Na uzyskanym diagramie (rysunek 6.18) widzimy, że punkty są dość mocno rozproszone, co świadczy o tym, że zależność między interesującymi nas zmiennymi nie jest silna, niemniej wyłania się ogólna prawidłowość, zgodnie z którą posiadanie niskiego SES sprzyja – średnio rzecz biorąc – posiadaniu niskiego SES przez małżonka, a posiadaniu wysokiego SES towarzyszy – średnio rzecz biorąc – wysoki SES małżonka. Prawidłowość ta nie jest wyraźnie liniowa, ale też nie jest wyraźnie krzywoliniowa.

Dołożenie prostej regresji do punktów jest możliwe w Edytorze wykresów. Przechodzimy do niego, dwukrotnie klikając w wykres. W menu Edytora wykresów wyszukajmy *Elementy* → *Linia dopasowania w ogółem*. W oknie, które się otworzy, wystarczy zaakceptować domyślny wybór *Regresja liniowa*.




**Rysunek 6.18.** Wykres rozrzutu/punktowy dla zmiennych *SEI10* i *SPSEI10*

Wykorzystajmy współczynnik *r* Pearsona do ustalenia siły związku i przeprowadźmy test istotności statystycznej. Aby dodatkowo upewnić się co do możliwości zastosowania tych metod (w kontekście niejednoznacznych wniosków dotyczących liniowości

związku), sprawdzimy też wartość współczynnika $r_s$ Spearmana. Jak podkreślano, jeśli przy nieznacznych odstępstwach od założeń obie metody dadzą podobne wyniki, wówczas wybierzemy bardziej adekwatną do zmiennych mierzonych na skali ilościowej metodę parametryczną.

Aby wyznaczyć obydwa współczynniki, wybieramy z *Analiza → Korelacje → Parami*. W głównym oknie procedury do pola *Zmienne* wprowadzamy *SEI10* oraz *SPSEI10*. Wybieramy interesujące nas miary i akceptujemy domyślne zaznaczenie *Test istotności – dwustronna*, gdyż ta opcja odpowiada rozpatrywanej hipotezie alternatywnej (dwustronnej). Wyniki są przedstawione na rysunku 6.19.

A

**Korelacje**

| | | SES respondenta | SES żony / męża |
|---|---|---|---|
| SES respondenta | Korelacja Pearsona | 1 | .496** |
| | Istotność (dwustronna) | | .000 |
| | N | 608 | 291 |
| SES żony / męża | Korelacja Pearsona | .496** | 1 |
| | Istotność (dwustronna) | .000 | |
| | N | 291 | 302 |

**. Korelacja istotna na poziomie 0.01 (dwustronnie).

B

**Korelacje**

| | | | SES respondenta | SES żony / męża |
|---|---|---|---|---|
| rho Spearmana | SES respondenta | Współczynnik korelacji | 1.000 | .466** |
| | | Istotność (dwustronna) | . | .000 |
| | | N | 608 | 291 |
| | SES żony / męża | Współczynnik korelacji | .466** | 1.000 |
| | | Istotność (dwustronna) | .000 | . |
| | | N | 291 | 302 |

**. Korelacja istotna na poziomie 0.01 (dwustronnie).

**Rysunek 6.19.** Wynik analizy korelacji dla zmiennych *SEI10* i *SPSEI10*

Na rysunku 6.19A zawarte są wyniki analizy dla metody parametrycznej, a na rysunku 6.19B nieparametrycznej. Z tabeli raportowej, tzw. macierzy korelacji (rysunek 6.19) wynika, że obydwa współczynniki mają zbliżone wartości. Odczytujemy, że *r* Pearsona między statusem socjoekonomicznym respondenta i jego małżonka wynosi 0,496, a wartość $r_s$ Spearmana jest nieco niższa niż *r* i wynosi 0,466. Podobne są również wyniki testu istotności współczynnika korelacji. Kierując się wynikami tego porównania (w sytuacji niejednoznacznej oceny liniowości związku), decydujemy o zastosowaniu współczynnika korelacji liniowej Pearsona.

Przed współczynnikiem stoi znak plus, korelacja jest więc dodatnia, a zatem wraz ze wzrostem SES respondenta wzrasta – średnio rzecz biorąc – SES jego małżonka. Oceniając siłę korelacji, powiemy, że jest ona umiarkowana. Wyniki te potwierdzają wstępne spostrzeżenia dokonane podczas analizy wykresu rozrzutu. Biorąc pod uwagę prawdopodobieństwo testowe *p*, które jest bardzo niskie ($p < 0{,}001$), odrzucamy hipotezę zerową, a za prawdziwą uznajemy $H_1$. Tym samym możemy sformułować wniosek, że zależność między analizowanymi cechami jest istotna statystycznie (istnieje w populacji młodych Amerykanów).

# 7. Wprowadzenie do regresji liniowej

**Kluczowe pojęcia:** regresja liniowa, klasyczna metoda najmniejszych kwadratów, dopasowanie modelu, współczynnik determinacji, ANOVA, średni błąd resztowy, test *t* w ocenie istotności parametrów strukturalnych, współczynnik regresji, standaryzowany współczynnik regresji, współliniowość zmiennych objaśniających, metoda krokowa, metoda wprowadzania

## 7.1. Uwagi wstępne

Omówione wcześniej metody umożliwiają jedno- lub dwuwymiarową analizę rozkładu zmiennej. Pozwalają zatem poznać kształtowanie się badanego zjawiska – jego poziom, zróżnicowanie, asymetrię i spłaszczenie rozkładu, a także zbadanie wzajemnych zależności między zmiennymi czy porównanie dwóch lub więcej populacji na podstawie wyników z próby. A co zrobić, jeśli chcemy ocenić związki przyczynowo-skutkowe, gdy poszukujemy odpowiedzi na pytanie o wpływ jakiegoś czynnika na badane przez nas zjawisko? W takiej sytuacji sięga się po modele. Są one różne, dostosowane do charakteru badanych zależności, zmiennych i innych określonych w przypadku danej metody warunków. Jednym z takich rozwiązań jest analiza regresji.

Istotą analizy regresji jest budowa modelu, który wyjaśnia mechanizm zmian zachodzących w badanym wycinku rzeczywistości. Nadrzędna zasada budowy modeli jest taka, że umożliwiają one ocenę relacji między badanym zjawiskiem a powiązanym z nim czynnikiem *ceteris paribus*, tj. zakładając, że pozostałe czynniki nie ulegają zmianie (a więc przy ustalonych wartościach pozostałych zmiennych objaśniających). Jak podkreślano w rozdziale drugim, samo porównanie populacji czy ocena współzależności nie dają jeszcze prawa do wnioskowania o związkach przyczynowo-skutkowych, nie można na ich podstawie powiedzieć, co wpływa na co. Jeśli zależy nam na tego typu wnioskach (a zwykle tak właśnie jest),

powinniśmy skonstruować model, w którym włączone zostaną oprócz interesującego nas czynnika również inne zmienne, które – w świetle przeglądu literatury przedmiotu – również mogą na badane zjawisko oddziaływać. Dzięki temu, że, konstruując model, wpływ tych innych czynników „zamrażamy", zakładamy, że skoro wszystkie jednostki badania mają te zmienne na takim samym poziomie, to nie oddziałują one na analizowane przez nas zjawisko (umownie, ich zmienność nie występuje, a tym samym nie może determinować zmienności badanej cechy). Dzięki takiemu podejściu można uniknąć związków pozornych, występowanie związku między zmiennymi nie musi bowiem jeszcze oznaczać, że kształtowanie się jednej z nich jest wynikową zmienności drugiej. Przykładowo: możemy zaobserwować zależność między wzrostem osoby a długością jej włosów – osoby wyższe mają średnio krótsze włosy. Jednak czy jest to efektem wzrostu? Im ktoś jest wyższy, tym szybciej rosną mu włosy, są zdrowsze itp., a w efekcie osoby te mają dłuższe włosy? Instynktownie odpowiemy, że takiego związku przyczynowo-skutkowego tu nie ma. A korelacja występuje. Ten prosty przykład pokazuje, że na podstawie korelacji nie można jeszcze wnioskować o „wpływie" jednej zmiennej na drugą. W badaniach związków przyczynowo-skutkowych powinno się przeprowadzić badanie eksperymentalne albo zbudować model, w którym – na podstawie teoretycznych przesłanek – oprócz wzrostu uwzględnimy również inne czynniki, mogące powodować, że długość włosów jest u różnych osób inna. W drugim z wariantów pokażemy, jak wzrost „oddziałuje" na długość włosów na tle innych czynników, określimy „czysty" efekt działania interesującego nas czynnika. Oczywiście czynników, które nas interesują, może być więcej, ale każdorazowo budując model regresji uwzględniający wiele czynników, będziemy mogli ocenić relację między każdą z tych zmiennych a badanym zjawiskiem, przy założeniu stałego poziomu pozostałych czynników. Model taki określa się jako model regresji wielorakiej (por. np. Nowak, 2001, s. 47; Walesiak, Gatnar, 2009, s. 128–129; Rószkiewicz, 2011, s. 244) lub regresji wielokrotnej (por. np. Larose, 2006, s. 83–84; Szymczak, 2010, s. 151; Bedyńska, Książek, 2012, s. 36), a czasem też jako model regresji wielozmiennowej (określenia te używane są m.in. w Szymczak, 2010; Bedyńska, Książek, 2012). Można, rzecz jasna, skonstruować również model uwzględniający tylko jeden czynnik (mamy wówczas do czynienia z regresją prostą), aczkolwiek taki model wyjaśni nam niewiele więcej niż współczynnik korelacji. Dlatego w praktyce badawczej konstruuje się przede wszystkim modele uwzględniające wiele potencjalnych determinant badanego zjawiska. W sensie statystycznym możemy w takim modelu wykazać nawet, że długość włosów wpływa na wagę czy wiek ludzi, ale oczywiście w sensie merytorycznym zależności takie nie mają sensu. Dlatego tak ważne jest dokonanie przeglądu literatury przedmiotu, zanim przystąpimy do budowy modelu, a nawet zanim przeprowadzimy badanie

kwestionariuszowe. Jeśli bowiem nie zapytamy o ważne z punktu widzenia badanego zjawiska charakterystyki, nie będzie możliwe ich uwzględnienie w modelu, a tym samym będzie miał on słabsze własności poznawcze.

Analizując procesy ekonomiczne, do badania zależności w opisanym powyżej ujęciu wykorzystuje się modele ekonometryczne, podczas gdy w innych obszarach mówi się raczej o analizie regresji. Jak definiuje Welfe (2009, s. 25): „modele ekonometryczne są to kwantyfikowalne relacje zapisane w postaci pojedynczych równań matematycznych lub ich układów, łączące w sposób zgodny z teorią ekonomii dane empiryczne dotyczące zjawisk gospodarczych". W obu przypadkach konstruowanie, weryfikacja i interpretacja wyników analizy przebiegają analogicznie, choć z uwagi na to, że procesy ekonomiczne zwykle analizuje się na podstawie prób czasowych danych makroekonomicznych, podczas gdy w innych badaniach raczej są to próby przekrojowe danych indywidualnych (dla pojedynczych osób, przedsiębiorstw, jednostek samorządu terytorialnego itp.), nie wszystkie metody znajdują zastosowanie w obu ujęciach. W niniejszym rozdziale skoncentrujemy się, jak dotychczas, na analizie danych indywidualnych, aczkolwiek większość z tych metod znajduje zastosowanie również dla prób czasowych (jeśli będzie inaczej, zwrócimy na to uwagę). W dalszej części rozdziału przedstawione zostaną zasady konstruowania modeli regresji i ich weryfikacji. Przykłady empiryczne dotyczyć przy tym będą nie tylko ekonomii, ale również zagadnień będących przedmiotem badań w innych dyscyplinach nauk społecznych. Jednocześnie pragniemy podkreślić, że rozdział ten nie wyczerpuje wszystkich zagadnień związanych z modelami regresji, wiele testów jest jedynie zasygnalizowanych. Zainteresowanych pogłębieniem wiedzy w tym zakresie odsyłamy do przywoływanych w pracy pozycji literatury[22].

## 7.2. Podstawowe założenia i etapy analizy regresji

Ogólnie rzecz ujmując, jednorównaniowy model regresji można zapisać jako:

$$Y = f\left(X_1, X_1, \ldots, X_k, \varepsilon\right), \tag{42}$$

gdzie:
$Y$ – zmienna objaśniana (zależna),
$X_k$ – zmienne objaśniające (niezależne, predyktory),
$\varepsilon$ – składnik losowy.

22 Szczegółowe informacje dotyczące sygnalizowanych tu zagadnień można znaleźć między innymi w: Welfe, 2003; 2009; Gajda, 2004; Gruszczyński, Podgórska, 2004; Larose, 2006; Kufel, 2007; Maddala, 2008; Goryl, Jędrzejczyk, Kukuła, 2009; Walesiak, Gatnar, 2009; Szymczak, 2010; Bedyńska, Książek, 2012; Gruszczyński, 2012; Sobczyk, 2013.

Uwzględnienie w modelu składnika losowego nadaje mu stochastyczny charakter, a włączenie więcej niż jednego $X_k$ daje możliwość weryfikacji opracowanego modelu teoretycznego opisującego wielowymiarową relację między zmiennymi objaśniającymi łącznie oddziałującymi na zmienną objaśnianą.

Najprostszy z modeli regresji to model jednorównaniowy, liniowy i tego typu modelem zajmiemy się w tym rozdziale.

W klasycznej analizie regresji liniowej przyjmuje się następujące założenia, określane jako założenia schematu Gaussa-Markowa (Welfe, 2009, s. 29–32):

- model jest niezmienniczy ze względu na obserwacje;
- model jest liniowy względem parametrów;
- zmienna objaśniająca jest nielosowa, jej wartości są ustalonymi liczbami rzeczywistymi (warunek identyfikacji);
- składnik losowy ma rozkład normalny;
- wartość oczekiwana składnika losowego jest równa zeru (występujące zakłócenia, które reprezentuje składnik losowy, mają tendencję do wzajemnej redukcji);
- składnik losowy jest sferyczny, tj. homoskedastyczny (ma stałą wariancję) i nie występuje autokorelacja składnika losowego;
- informacje z próby są jedynymi, na podstawie których estymuje się parametry strukturalne modelu.

Założenia te definiują model liniowy z jedną zmienną objaśniającą. W przypadku modeli z wieloma zmiennymi objaśniającymi formułuje się analogiczne założenia, przy czym z uwagi na występowanie kilku zmiennych objaśniających warunek trzeci zastępujemy przez dwa poniższe (Welfe, 2009, s. 60–61):

- macierz **X**, zawierająca wartości zmiennych objaśniających dla poszczególnych obserwacji, jest nielosowa, tzn. jej elementy są ustalone w powtarzalnych próbach (warunek identyfikacji);
- rząd macierzy **X** jest równy liczbie szacowanych parametrów, liczba obserwacji jest zatem co najmniej równa liczbie szacowanych parametrów oraz nie występuje współliniowość w zbiorze zmiennych objaśniających (Walesiak, Gatnar, 2009, s. 129).

Jeśli liczba stopni swobody (wyznaczana jako $df = n - k$, a więc liczba obserwacji pomniejszona o liczbę szacowanych parametrów strukturalnych) jest zbyt mała, oceny współczynników regresji mogą być bardzo niestabilne i będą się silnie zmieniać wraz ze wzrostem liczby przypadków. Wielu autorów zaleca, aby w przypadku prób przekrojowych uwzględnić w analizie przynajmniej około 10 do 20 razy więcej obserwacji niż zmiennych. Do kwestii współliniowości zmiennych objaśniających wrócimy w dalszej części rozdziału.

Analiza regresji przebiega według następujących etapów:
- specyfikacja zmiennych (opracowanie teoretycznych podstaw modelu);
- wybór analitycznej postaci modelu;
- estymacja parametrów strukturalnych modelu;
- weryfikacja statystyczna i merytoryczna modelu;
- praktyczne wykorzystanie oszacowanego modelu (opis zjawiska, prognozowanie jego poziomu poza próbą).

Pierwszy etap jest bardzo ważny i wymaga dobrego rozpoznania zjawisk będących przedmiotem analiz. Pod uwagę należy wziąć zwłaszcza następujące kwestie:
- dobierając zmienne objaśniające, kierujemy się teorią, do której odwołuje się model, wynikami innych badań w tym zakresie, procedurami statystycznymi;
- zmienne objaśniające powinny być jak najsilniej skorelowane ze zmienną objaśnianą, a słabo między sobą; zbyt silne skorelowanie zmiennych objaśniających prowadzić może do ich współliniowości;
- zmienne objaśniające powinny mieć wysoką zmienność;
- zmienne objaśniające powinny być jasno zdefiniowane i mieć wyraźną interpretację merytoryczną.

W modelu regresji liniowej zarówno zmienna objaśniana, jak i zmienne objaśniające mierzone są na skali ilościowej. Możliwe jest włączenie w roli zmiennej objaśniającej zmiennej jakościowej (co jest ważne zwłaszcza w modelach budowanych na podstawie próby przekrojowej), niemniej jednak należy je uprzednio sprowadzić do postaci quasi-ilościowej. Robi się to najczęściej poprzez przekształcenie zero-jedynkowe tych zmiennych (tworzy się zmienne instrumentalne, przyjmujące wartość 1 dla wybranego wariantu zmiennej, a 0 dla pozostałych). Przykładowo: badając uwarunkowania nakładów na innowacje na podstawie próby MŚP, w roli zmiennej objaśniającej można włączyć informację na temat tego, czy przedsiębiorstwo ma wyłącznie polski kapitał – przyjmijmy, że wprowadzamy zmienną *kapitał*. Zmienna ta ma wartość 1 lub 0, na przykład *kapitał* = 1 dla firm z udziałem kapitału zagranicznego, natomiast *kapitał* = 0 dla firm o wyłącznie polskim kapitale. Jeśli jakościowa zmienna objaśniająca ma więcej niż dwa warianty, wówczas włącza się $k - 1$ zmiennych zero-jedynkowych dla $k - 1$ wariantów zmiennej (nieuwzględniona w postaci zmiennej zero-jedynkowej wartość zmiennej stanowi kategorię odniesienia). Przykładowo: jeśli we wspomnianym badaniu chcemy uwzględnić wielkość przedsiębiorstwa (mikro, małe, średnie), możemy jako grupę odniesienia przyjąć na przykład średnie podmioty. Do modelu regresji włączamy wtedy dwie zmienne zero-jedynkowe:

*mikro* = 1 dla mikroprzedsiębiorstw
0 dla pozostałych MŚP

*małe* = 1 dla małych przedsiębiorstw
0 dla pozostałych MŚP

Podobnie włączając tego typu zmienne sztuczne do modelu regresji liniowej budowanego na podstawie próby czasowej (przy dezagregacji danych rocznych), można wyodrębnić wpływ sezonowości – wprowadza się wówczas zmienne zero-jedynkowe odpowiadające kwartałom.

W modelach regresji uwzględniana jest również zmienna czasowa. Podobnie jak zmienna zero-jedynkowa jest to zmienna sztuczna. Tworzy się ją, przypisując numery kolejnym okresom (lub momentom): $t = 1, 2, \ldots, n$. Jeśli w modelu regresji występuje wyłącznie zmienna czasowa, taki model określa się jako funkcję trendu. Może mieć ona (podobnie jak „tradycyjne" modele) różne postaci analityczne, a ich sposób estymacji i weryfikacji jest analogiczny jak dla „tradycyjnych" modeli. Zmienna czasowa może być też włączona do modelu regresji jako kolejna (obok „tradycyjnych" zmiennych $X$) zmienna objaśniająca. Szczególnym przypadkiem zmiennych objaśniających są zmienne autoregresyjne, czyli opóźnione wartości zmiennej objaśnianej. Modele autoregresyjne, uwzględniające takie podejście, nie będą w tym opracowaniu omawiane.

Dokonując wyboru postaci analitycznej modelu, decydujemy, czy badaną relację można opisać najprostszą funkcją liniową, czy też należy użyć funkcji nieliniowej (a jeśli tak, to jakiej). Jak pisze Sobczyk (2013), teoria ekonomii przeważnie nie dostarcza w tym zakresie gotowych rozwiązań. Dotyczy to również innych obszarów analiz. Pomocne są w związku z tym wyniki innych badań, intuicja, a także eksperymenty obliczeniowe, wykresy przebiegu zmiennej objaśnianej względem poszczególnych zmiennych objaśniających, własności funkcji matematycznych czy też testy statystyczne. Tak jak zaznaczono w tytule rozdziału, w tym miejscu zajmiemy się liniowym modelem regresji. Jeśli zależność nie ma charakteru liniowego, to można dokonać transformacji zmiennych, można też wykorzystać modele nieliniowe. Wśród modeli nieliniowych są takie, które są nieliniowe względem zmiennych, ale w wyniku prostych przekształceń można je sprowadzić do postaci liniowej względem parametrów, których parametry szacuje się tak samo jak dla modelu liniowego (więcej szczegółów – por. Welfe, 2009, s. 31). Kolejne dwa etapy analizy regresji są na tyle obszerne treściowo, że omówione zostaną w odrębnych podrozdziałach.

## 7.3. Linowy model regresji – estymacja parametrów

Charakter relacji zachodzących w populacji generalnej zapisać można następującym równaniem:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \tag{43}$$

gdzie: $\beta_j$ – parametr strukturalny modelu, w tym $\beta_0$ – wyraz wolny, $\beta_1$, $\beta_2$, …, $\beta_k$ – współczynniki regresji. W literaturze przedmiotu parametry strukturalne oznacza się też jako $\alpha$.

W zapisie macierzowym model ten ma postać (Kufel, 2007, s. 55):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{44}$$

gdzie: $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$; $\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{k1} \\ x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix}$; $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$; $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$.

W modelu regresji liniowej **współczynnik regresji** (czyli parametr strukturalny stojący przy zmiennej objaśniającej) informuje o tym, o ile przeciętnie zmieni się *y* wraz ze wzrostem *x* o jednostkę (przy założeniu *ceteris paribus*, tj. że pozostałe zmienne objaśniające nie ulegają zmianie). Współczynnik ten informuje o „bezpośrednim efekcie jednostkowej zmiany zmiennej objaśniającej" (Welfe, 2009, s. 64). Ujemna wartość współczynnika regresji wskazuje (*ceteris paribus*) na ujemną liniową relację między zmienną objaśnianą a zmienną objaśniającą (przy założeniu stałego poziomu pozostałych zmiennych objaśniających, im wyższy poziom *x*, tym średnio niższy poziom *y*). Z kolei dodatnia wartość współczynnika regresji wskazuje (*ceteris paribus*) na dodatnią liniową relację między zmienną objaśnianą a zmienną objaśniającą (przy założeniu stałego poziomu pozostałych zmiennych objaśniających, im wyższy poziom *x*, tym średnio wyższy poziom *y*). Współczynnik regresji równy zero wskazuje na brak liniowej zależności między zmienną objaśnianą i objaśniającą. W przypadku jakościowych zmiennych objaśniających współczynnik regresji stojący przy danej zmiennej zero-jedynkowej informuje (*ceteris paribus*), o ile przeciętnie *y* jest wyższy (dla $\beta > 0$) bądź niższy (dla $\beta < 0$) w grupie

badanej w porównaniu z grupą odniesienia. Z kolei współczynnik regresji stojący przy zmiennej czasowej *t* (współczynnik trendu) wskazuje, o ile przeciętnie z okresu na okres (np. z roku na rok) poziom zmiennej objaśnianej rośnie (dla $\beta > 0$) lub maleje (dla $\beta < 0$).

Parametry modelu regresji liniowej estymujemy na podstawie wyników z próby, uzyskując w efekcie równanie postaci:

$$\hat{y} = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k , \tag{45}$$

gdzie:

$\hat{y}$ – teoretyczne (oszacowane na podstawie modelu) wartości zmiennej objaśnianej,

$B_0, B_1, \ldots, B_k$ – oszacowania parametrów strukturalnych (estymatory współczynników regresji).

Do tego celu wykorzystywana jest najczęściej klasyczna metoda najmniejszych kwadratów (KMNK), która pozwala na uzyskanie takich oszacowań parametrów strukturalnych, przy których suma kwadratów reszt jest najmniejsza (Wątroba, 2011):

$$\sum e_i^2 = \sum \left( y_i - \hat{y}_i \right)^2 \rightarrow \min, \tag{46}$$

gdzie: $e_i$ – reszty modelu.

Reszty modelu wyrażają błąd, jaki popełniamy, szacując poziom *y* dla danej obserwacji (osoby, roku itp.) na podstawie wyznaczonego modelu. Estymator klasycznej metody najmniejszych kwadratów, oznaczany jako KMNK-estymator (*OLS – ordinary least squares*), zwany też krócej MNK-estymatorem, jest BLUE (*best linear unbiased estimator*), tzn. najlepszym nieobciążonym estymatorem w klasie estymatorów liniowych (Welfe, 2009, s. 67). Estymator ten jest zatem nieobciążony, tj. jego wartość oczekiwana jest równa poszukiwanemu wektorowi parametrów (ta własność najczęściej uznawana jest za najważniejszą), można go przedstawić jako liniową kombinację zaobserwowanych wartości zmiennej zależnej (jest liniowy) i ma najmniejszą wariancję.

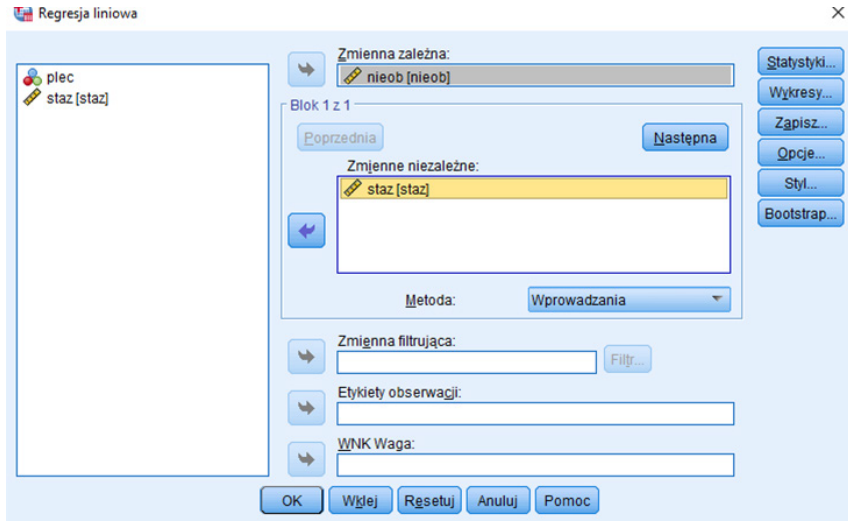
**Przykład 7.1**

Na podstawie danych dotyczących próby *n* = 120 pracowników średniego przedsiębiorstwa skonstruujmy model regresji liniowej opisujący absencję w pracy mierzoną liczbą dni nieobecności (*nieob*), przyjmując jako zmienną objaśniającą staż pracy pracowników (*staz*, w latach).

**Rozwiązanie**

Zdefiniujmy na początek problem badawczy. Zmienną objaśnianą jest *nieob*, a zmienną objaśniającą *staz*. Można to zapisać: $y = f(staz)$ (czyt. zmienna zależna *nieob* jest funkcją zmiennej *staz*). Obie zmienne mierzone są na skali ilościowej. Konstruując model, zakładamy liniowy związek między zmiennymi i estymujemy parametry liniowego modelu regresji.

Aby wykonać to polecenie, należy wybrać *Analiza* → *Regresja* → *Liniowa*. Następnie definiujemy *Zmienna zależna* (tu będzie to zmienna *nieob*) oraz *Zmienna niezależna* (tu *staz*) (rysunek 7.1). Już na tym etapie wygenerowane zostaną wyniki pozwalające na zapisanie oszacowanego równania regresji, ale też jego weryfikację statystyczną i merytoryczną.




**Rysunek 7.1.** Okno polecenia *Regresja liniowa* – estymacji modelu liczby dni nieobecności w pracy: $y = f(staz)$

Dodatkowo można wyznaczyć przedziały ufności dla poszczególnych parametrów strukturalnych (czyli przedziałowe estymatory parametrów – z określoną wiarygodnością $1 - \alpha$). W tym celu wybieramy polecenie *Analiza* → *Regresja* → *Liniowa*, a następnie *Statystyki* → *Współczynniki regresji* → *Przedziały ufności*.

Wartości ocen parametrów strukturalnych znajdziemy w tabeli *Współczynniki*, w kolumnie *Współczynniki niestandaryzowane B* (rysunek 7.2). Z wiersza *Stała* odczytujemy wartość oszacowania wyrazu wolnego, a z wiersza *staz* oszacowanie współczynnika regresji opisującego relację między stażem pracy a absencją w pracy.

**Współczynniki**[a]

| Model | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | t | Istotność | 95,0% przedział ufności dla B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Błąd standardowy | Beta | | | Dolna granica | Górna granica |
| 1 | (Stała) | 5.806 | .505 | | 11.500 | .000 | 4.814 | 6.798 |
| | staz | .251 | .006 | .898 | 42.770 | .000 | .239 | .262 |

a. Zmienna zależna: nieob

**Rysunek 7.2.** Wyniki estymacji modelu liczby dni nieobecności w pracy: $y = f(staz)$

Oszacowane równanie można zapisać:

$$\hat{y}_i = 5{,}806 + 0{,}251 \times staz\,.$$

$B = 0{,}251$, a więc wskazuje na dodatnią relację między stażem pracy a liczbą dni nieobecności w pracy – im dłuższy staż pracy, tym średnio wyższy poziom absencji. Jeśli staż jest dłuższy o rok, to liczba dni nieobecności w pracy jest wyższa średnio o 0,251 (z błędem średnio +/–0,006)[23].

Co zrobić, aby włączyć do modelu zmienne dyskretne? Prześledźmy to na przykładzie zmiennej *płeć* (przykład 7.2).

**Przykład 7.2**

Na podstawie tych samych co w poprzednim przykładzie danych skonstruujmy model regresji liniowej opisujący absencję w pracy mierzoną liczbą dni nieobecności (*nieob*), przyjmując jako zmienne objaśniające staż pracy pracowników (*staz*, w latach) i płeć (*plec*).

**Rozwiązanie**

Zdefiniujmy problem badawczy. Zmienną objaśnianą jest – jak poprzednio – *nieob*, a zmienne objaśniające to *staz* oraz *plec* (co można zapisać: $y = f(staz, plec)$). Zmienne *nieob* i *staz* mierzone są na skali ilościowej. Z kolei zmienna objaśniająca *plec* jest mierzona na skali nominalnej, a jej wartości wyrażone są liczbami 1 i 2.

23 Przedział ufności dla współczynnika regresji dla zmiennej *staz* można zinterpretować w następujący sposób: przedział o granicach (0,239; 0,262) pokrywa z wiarygodnością 95% szacowaną wartość parametru mierzącego zmiany liczby nieobecności pracy wraz ze wzrostem stażu pracy o rok. Dodajmy, że przy standardowej ocenie modelu interpretację przedziałów ufności zwykle się pomija. Przedziały te wykorzystuje się (zamiast testu t-Studenta) przy ocenie istotności współczynników regresji, zwłaszcza w przypadku zastosowania metod repróbkowania (np. metody bootstrapowej).

Na potrzeby modelu regresji zmienną jakościową *plec* należy przekształcić do postaci zmiennej zero-jedynkowej (*plec1*). Przyjmijmy, że grupę badaną stanowią mężczyźni (*plec1* = 1), a grupę odniesienia kobiety (*plec1* = 0). Wykorzystujemy polecenie: *Przekształcenia → Rekoduj na inne zmienne*. Możemy też wykorzystać polecenie *Predictive solutions → Rekodowanie dychotomiczne* (por. rozdział pierwszy) i włączyć do modelu jedną z wygenerowanych w ten sposób zmiennych.

W sytuacji gdy wszystkie zmienne są już ilościowe lub quasi-ilościowe, wykonujemy te same kroki co w poprzednim przykładzie: wybieramy *Analiza → Regresja → Liniowa*, a potem definiujemy *Zmienna zależna* (*nieob*) oraz *Zmienna niezależna* (tu *staz* i *plec1*). Wyniki zestawiono na rysunku 7.3.

**Współczynniki**[a]

| Model | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | | |
|---|---|---|---|---|---|---|
| | | B | Błąd standardowy | Beta | t | Istotność |
| 1 | (Stała) | 6.472 | .495 | | 13.083 | .000 |
| | staz | .267 | .006 | .955 | 43.352 | .000 |
| | plec1 | -2.153 | .337 | -.141 | -6.380 | .000 |

a. Zmienna zależna: nieob

**Rysunek 7.3.** Wyniki estymacji modelu liczby dni nieobecności w pracy: *y* = *f*(*staz*, *plec1*)

Oszacowane równanie można zapisać:

$$\hat{y}_i = 6,472 + 0,267^{*}staz - 2,153^{*}plec1 .$$

$B_1$ = 0,267, więc przy założeniu stałego poziomu pozostałych czynników, jeśli staż pracy jest dłuższy o rok, liczba dni nieobecności w pracy jest wyższa średnio o 0,267 (z błędem średnio +/–0,006). Z kolei $B_2$ = –2,153, więc w przypadku mężczyzn liczba dni nieobecności jest, *ceteris paribus*, średnio o 2,153 mniejsza niż w przypadku kobiet.

## 7.4. Weryfikacja modelu regresji

Oszacowany model regresji uznamy za dobry i wykorzystamy do celów praktycznych, jeśli będzie poprawny pod względem statystycznym i merytorycznym.

**Ocena merytoryczna modelu**, najprościej rzecz ujmując, polega na sprawdzeniu, czy oszacowane równanie jest zgodne z założeniami (przyjętymi na gruncie teoretycznym). Sprawdzamy też, czy model jest koincydentny, tj. czy znaki

parametrów są takie same jak przy współczynnikach korelacji między poszczególnymi zmiennymi objaśniającymi i zmienną objaśnianą[24]. Odnosimy się w tym miejscu do oszacowań współczynników regresji. Sprawdzamy również, czy wartość współczynnika regresji wskazuje na możliwą do wystąpienia skalę zmian *y* pod wpływem jednostkowej zmiany *x* (przykładowo: czy możliwe jest, że zatrudnienie kolejnego pracownika w małej firmie zwiększy wartość sprzedaży średnio o 400 tys. zł w miesiącu, chociaż przeciętnie przedsiębiorstwo osiąga wartość sprzedaży rzędu 200 tys. zł w miesiącu).

**Ocena statystyczna modelu** obejmuje przede wszystkim dwa podstawowe obszary – ocenę własności prognostycznych modelu oraz ocenę szacunków współczynników regresji (od strony statystycznej odnosimy się do relacji między zmiennymi objaśniającymi a zmienną objaśnianą). Rzecz jasna, model możemy ocenić pod tym kątem, jeśli spełnione zostały założenia, o których była mowa wcześniej, a więc należy również sprawdzić, czy faktycznie ma to miejsce. Najczęściej przyjmuje się, że nawet jeśli własności prognostyczne modelu nie są dobre, model regresji można wykorzystać praktycznie (innymi słowy, słabe własności prognostyczne modelu nie wykluczają wyjaśnienia na jego podstawie, „jak to działa").

Weryfikując model regresji liniowej od strony statystycznej, sprawdzamy:

1) własności prognostyczne modelu (jego dopasowanie do danych empirycznych) poprzez:
   1.1) sprawdzenie, czy współczynnik determinacji w populacji istotnie różni się od zera – stosujemy ANOVA,
   1.2) sprawdzenie stopnia dopasowania modelu do danych empirycznych w obrębie próby – wyznaczamy współczynnik determinacji z próby,
   1.3) sprawdzenie, o ile przeciętnie się mylimy, przewidując poziom *y* na podstawie modelu regresji – stosujemy średni błąd resztowy (średni błąd szacunku);
2) oceniamy relacje między poszczególnymi zmiennymi objaśniającymi a zmienną objaśnianą poprzez:
   2.1) ocenę istotności statystycznej parametrów strukturalnych – testem t-Studenta,
   2.2) odniesienie się do błędów szacunku parametrów strukturalnych – powinny być jak najniższe.

Odnosimy się też do merytorycznych aspektów związanych z parametrami strukturalnymi, tj.[25]:

24 Przyczyną braku koincydencji może być niewłaściwa postać analityczna modelu lub współliniowość zmiennych objaśniających.

25 W tym miejscu porównuje się też ewentualnie wartości współczynników korelacji cząstkowej (badanej przy założeniu, że liniowe efekty pozostałych zmiennych objaśniających

2.3) interpretujemy merytorycznie oszacowania współczynników regresji (wartości współczynników regresji z próby) – $B_j$,
2.4) interpretujemy współczynniki standaryzowane Beta (pozwalające na porównanie „ważności" zmiennych, tj. określenie, pod wpływem której zmiennej objaśniającej, *ceteris paribus*, badane zjawisko ulega największym zmianom).

Jak wspomniano, należy również sprawdzić, czy zostały spełnione założenia stawiane w regresji liniowej[26]. W tym miejscu odniesiemy się do współliniowości zmiennych objaśniających. Z uwagi na ramy tej publikacji, w tym zwłaszcza koncentrowanie się na szeregach przekrojowych (danych indywidualnych), oraz na ograniczeniaIBM SPSS Statistics dotyczące części z wymienionych powyżej metod, w rozdziale tym skupimy się przede wszystkim na zagadnieniach związanych z weryfikacją statystyczną modelu (pkt 1–2).

zostały wyeliminowane) ze współczynnikami korelacji rzędu zerowego (czyli „prostymi" współczynnikami korelacji liniowej Pearsona między daną zmienną objaśniającą a zmienną objaśnianą). Duże różnice między nimi mogą wskazywać na występowanie interakcji między zmiennymi objaśniającymi (tzn. współwystępowanie dwóch lub więcej czynników kształtuje poziom *y* i łączny efekt ich działania jest inny niż pojedyncze działanie każdego z czynników z osobna).

26 Ocena liniowości związku jest stosunkowo łatwa przy modelu z jedną zmienną objaśniającą, przy większej ich liczbie możliwe jest zastosowanie testów statystycznych – testu nieliniowości (w wersji „logarytmy" lub „kwadraty") bądź testu Ramsey'a RESET. Normalność rozkładu składnika losowego ocenić można standardowymi testami normalności – Shapiro-Wilka (mało wrażliwy na autokorelację i heteroskedastyczność składnika losowego), Jarque'a-Berry (tylko dla dużych prób), czy też Doornika-Hansena. Można dodatkowo wykorzystać wykresy reszt, w tym zwłaszcza histogram czy wykres normalności. Heteroskedastyczność składnika losowego ocenić można na przykład testem White'a czy Breuscha-Pagana. W tym przypadku pomocny będzie także wykres reszt względem wartości *y*. Dla modeli estymowanych na podstawie prób czasowych konieczne jest również zbadanie: (1) autokorelacji składnika losowego (najczęściej oceniana jest autokorelacja składnika losowego I rzędu – AR(1), tj. ocena, czy zakłócenia dla sąsiadujących okresów są ze sobą skorelowane), zwykle wykorzystuje się w tym celu test Durbina-Watsona, (2) stacjonarności szeregu czasowego, rozumianej jako niezmienność rozkładu prawdopodobieństwa zmiennej w czasie, a co najmniej niezmienność średniej danego procesu w czasie (np. test ADF), (3) stabilności parametrów w czasie (np. test Chowa).

## Ocena własności prognostycznych modelu

### Ocena istotności współczynnika determinacji ($R^2$)

Do oceny istotności współczynnika determinacji stosuje się ANOVA. Hipotezy mają postać (Szymczak, 2010, s. 159):
$H_0$: $R^2 = 0$
$H_1$: $R^2 > 0$.

Sprawdzianem testu jest statystyka $F$ postaci (Rószkiewicz, 2011, s. 244):

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{\frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}. \tag{47}$$

Statystyka ta (przy prawdziwości $H_0$) ma rozkład Fishera-Snedecora z $k$ i $n - k$ stopniami swobody (gdzie $k$ – liczba szacowanych parametrów, $n$ – liczebność próby). Jest ona liczona jako stosunek wariancji regresyjnej (zmienności zjawiska opisanej wyznaczonym modelem regresji) i wariancji resztowej (niewyjaśnionej przez model).

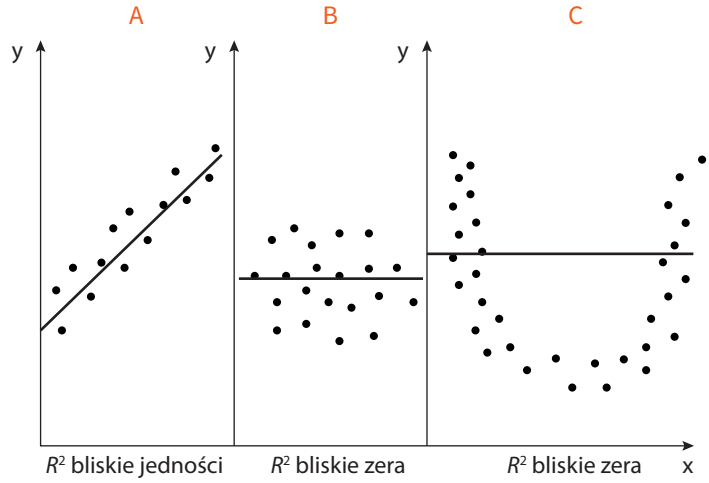
Jeśli $p < \alpha$, wnioskujemy, że współczynnik determinacji w populacji istotnie różni się od zera. Jeśli $p > \alpha$, model regresji nie ma własności poznawczych (nie pozwala wyjaśnić kształtowania się zmiennej objaśnianej).

## Interpretacja współczynnika determinacji z próby

Współczynnik determinacji obliczany na podstawie wyników z próby i informuje, w jakim stopniu zmiany $y$ zostały wyjaśnione za pomocą modelu (zmianami zmiennych objaśniających). Dla modelu liniowego z wyrazem wolnym współczynnik ten może przyjmować wartości z przedziału [0; 1]. Im jego wartość jest bliższa 1, tym lepsze własności prognostyczne modelu. Wyznacza się go ze wzoru (Sobczyk, 1998, s. 234):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2}. \tag{48}$$

Współczynnik determinacji wyznaczany jest na podstawie reszt modelu, a więc im są one mniejsze (im lepsze dopasowanie funkcji regresji liniowej do danych empirycznych), tym wyższa wartość współczynnika regresji z próby. Na rysunku 7.4 zobrazowano przebieg zależności między $x$ a $y$.




**Rysunek 7.4.** Charakter zależności a wartość współczynnika determinacji

**Źródło:** Welfe, 2009, s. 42.

W przypadku rysunku 7.4A regresja liniowa dobrze odwzorowuje charakter zależności między zmiennymi (linia regresji przebiega blisko punktów rozrzutu empirycznego) – w takiej sytuacji współczynnik determinacji jest bliski 1. Rysunek 7.4B obrazuje słabą zależność między zmiennymi – linia regresji liniowej nie odwzorowuje dokładnie tej relacji, niemniej jednak nie wynika to z nieodpowiedniej postaci analitycznej modelu regresji. W takiej sytuacji wartość $R^2$ jest niska. Współczynnik determinacji może być bliski zera również w sytuacji przedstawionej na rysunku 7.4C, gdy w przypadku wyraźnie nieliniowej zależności próbuje się dopasować liniową funkcję regresji (trafność przewidywań poziomu $y$ będzie wtedy niska).

W przypadku modelu regresji liniowej z jedną zmienną objaśniającą współczynnik determinacji jest równy kwadratowi współczynnika korelacji liniowej Pearsona ($r$). Dla modelu z wieloma zmiennymi objaśniającymi jest on z kolei równy kwadratowi współczynnika korelacji wielorakiej ($R$). Porównując dopasowanie kilku modeli regresji liniowej z różną liczbą zmiennych objaśniających, należy użyć skorygowanego współczynnika determinacji (uwzględniającego korektę ze względu na liczbę stopni swobody)[27].

Współczynnik determinacji jest najważniejszą miarą dopasowania modelu. Jak pisze Welfe (2009, s. 43–44), nie można jednoznacznie określić, co to jest

27 Alternatywą dla skorygowanego $R^2$ (bardziej uniwersalną, bo nie tylko dla KMNK) jest na przykład kryterium Akaike'a – *AIC* (lepszy jest ten model, dla którego *AIC* jest najmniejsze), kryterium informacyjne Hannana-Quinna lub bayesowskie kryterium informacyjne Schwarza (*BIC*).

„wysokie $R^2$". Dla modeli, których parametry estymowane są na podstawie szeregów czasowych, często $R^2$ są rzędu 0,90 i więcej. Z kolei w przypadku modeli opartych na danych przekrojowych otrzymywane w praktyce $R^2$ są zwykle znacznie niższe.

## Średni błąd resztowy

Średni błąd resztowy ($S_e$) określany jest również jako średni błąd szacunku, średni błąd oceny lub odchylenie standardowe składnika resztowego. Wyznacza się go według wzoru (Kufel, 2007, s. 57):

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1}}, \tag{49}$$

gdzie: $k$ - liczba zmiennych objaśniających.

W praktyce jest on znacznie rzadziej stosowany niż $R^2$. Informuje, o ile przeciętnie mylimy się, przewidując poziom $y$ na podstawie oszacowanego modelu. Jest on wielkością mianowaną i wyrażony jest w jednostkach zmiennej zależnej. Jeśli odniesiemy wartość tego błędu do średniego poziomu zmiennej objaśnianej, uzyskamy względną miarę dopasowania, łatwiejszą do porównań.

## Ocena relacji między zmiennymi objaśniającymi a zmienną objaśnianą

### Ocena istotności parametrów strukturalnych

Istotność poszczególnych parametrów strukturalnych, w tym zwłaszcza współczynników regresji (w populacji), ocenia się za pomocą testu t-Studenta. Hipotezy mają postać:

$H_0$: $\beta_j = 0$

$H_1$: $\beta_j \neq 0$.

Sprawdzianem testu t-Studenta jest statystyka $t$ postaci (Sobczyk, 1998, s. 257):

$$t = \frac{B}{S(B)}, \tag{50}$$

gdzie: $B$ – estymator współczynnika regresji, $S(B)$ – błąd szacunku współczynnika regresji.

Statystyka ta przy prawdziwości $H_0$ ma rozkład t-Studenta z $n - k$ stopniami swobody. Statystyka $t$ liczona jest jako stosunek oszacowania parametru strukturalnego i jego błędu szacunku (a więc im większy błąd, tym niższa wartość statystyki $t$, a tym samym – przy określonym $df$ – większe prawdopodobieństwo w teście t-Studenta, co z kolei przekłada się na nieistotność danego współczynnika regresji).

Jeśli w teście t-Studenta $p < \alpha$, parametr strukturalny uznajemy za istotny statystycznie; zmienna objaśniająca jest istotnie (w sensie statystycznym) powiązana ze zmienną objaśnianą. Jeśli $p > \alpha$, dany parametr strukturalny nie jest istotny statystycznie (a zatem zmienna objaśniająca nie jest istotnym predyktorem badanego zjawiska).

## Porównanie standaryzowanych współczynników regresji

Po dokonaniu oceny istotności statystycznej parametrów strukturalnych interpretuje się merytorycznie współczynniki regresji z próby ($B$), a także standaryzowane współczynniki regresji z próby (standaryzowane *Beta*). Standaryzacja współczynnika regresji polega na przekształceniu (Szymczak, 2010, s. 169):

$$Beta_j = B_j \cdot \frac{S_j}{S_y}, \tag{51}$$

gdzie: $S_j$ – odchylenie standardowe $j$-tej zmiennej objaśniającej, $S_y$ – odchylenie standardowe zmiennej objaśnianej, $B_j$ – oszacowanie parametru (współczynnika regresji) $\beta_j$.

Standaryzowane współczynniki regresji przyjmują wartości z przedziału [–1; 1]. Są one niezależne od zakresu zmiennych i uwolnione od ich jednostek. Możliwe jest zatem porównywanie (*ceteris paribus*) siły związku między poszczególnymi zmiennymi objaśniającymi a zmienną objaśnianą. Im większa jest wartość bezwzględna standaryzowanego współczynnika regresji, tym silniejszy wpływ zmiennej objaśniającej na zmienną objaśnianą.

## Ocena współliniowości zmiennych objaśniających

Biorąc pod uwagę założenia KMNK, ograniczymy się w tej publikacji do sprawdzenia współliniowości zmiennych objaśniających. Wiąże się ona ze zbyt silnym stopniem skorelowania zmiennych objaśniających. Dokładna współliniowość występuje rzadko, niemniej jednak możemy mieć do czynienia z przybliżoną współliniowością zmiennych objaśniających, która powoduje między innymi nieprawidłowy pomiar siły oddziaływania zmiennych objaśniających na zmienną objaśnianą

(Walesiak, Gatnar, 2009, s. 134). Do pomiaru współliniowości wykorzystuje się najczęściej statystykę *VIF* – czynnik inflacji wariancji estymatora parametru (*variance inflation factors*). Na problemy ze współliniowością wskazują wartości $VIF > 10$ (Kufel, 2007, s. 63). Sprawdzenie pozostałych założeń KMNK zostanie w tej publikacji pominięte[28].

**Przykład 7.3**

Wykorzystując dane dostępne w dokumentach kadrowych pracowników pewnego dużego przedsiębiorstwa, wylosowano próbę prostą ($n = 473$). Przyjęto, że analizie poddani zostaną mężczyźni ($n = 257$). Na podstawie tych danych skonstruowano liniowy model wynagrodzeń (w $). Jako zmienne objaśniające przyjęto: wiek pracowników, ich wykształcenie, staż pracy oraz przynależność do mniejszości. Dokonajmy interpretacji uzyskanych wyników.

28 Jak wcześniej podkreślano, normalność rozkładu składnika losowego można ocenić, stosując na przykład test Shapiro-Wilka. W IBM SPSS Statistics wśród podstawowych zestawień nie ma wyników testów pozwalających na ocenę w tym zakresie. Normalność rozkładu składnika losowego można jednak zbadać dzięki resztom modelu wyznaczonym na podstawie polecenia *Zapisz → Reszty → Niestandaryzowane*. Dla utworzonej w ten sposób zmiennej *RES_1* można przeprowadzić test Shapiro-Wilka z poziomu *Eksploracji*. Można również przeanalizować wykres normalności oraz histogram dla standaryzowanych reszt (*Regresja → Liniowa → Wykresy → Wykresy reszt standaryzowanych → Histogram / Normalny wykres prawdopodobieństwa*). Autokorelację składnika losowego bada się przede wszystkim dla prób czasowych. Do tego celu wykorzystuje się najczęściej test Durbina-Watsona (DW). Aby wyznaczyć statystykę DW, wybieramy: *Regresja → Liniowa → Statystyki → Reszty → Durbin-Watson*). Z kolei aby odnieść się do heteroskedastyczności (niejednorodności) składnika losowego, można narysować wykres rozrzutu, na którym zestawimy reszty i rzeczywiste wartości *y* (wykres ten może być wykreślony z poziomu *Regresja liniowa → Wykresy → Rozrzut 1 z 1 → Y: DEPENDENT, X: *ZRESID*). Założenie dotyczące jednorodności wariancji nie jest naruszone, jeśli rozrzut punktów nie jest systematycznie niejednakowy w pionie. Z kolei jeśli nie istnieje oczywista krzywizna na wykresie, można wnioskować, że nie jest naruszone założenie o liniowości modelu. W modelach regresji istotna jest również kwestia wartości odstających (o nietypowo wysokich wartościach zmiennej objaśnianej i/lub objaśniających). Dla tych wartości reszty są wysokie, a tym samym pogarszają one dopasowanie modelu. Bardziej groźne dla modelu są wartości wpływowe, które powodują przesunięcie linii regresji, a więc błędne szacunki parametrów strukturalnych. Identyfikacja takich przypadków może być dokonana na podstawie reszt modelu (w różnych ujęciach), jak również miar odległości i statystyk wpływu. W IBM SPSS Statistics można je wyznaczyć z poziomu *Regresja liniowa → Zapisz*. W zbiorze danych pojawiają się wówczas nowe zmienne. Ich wartości nietypowe względem pozostałych wskazują na występowanie takich obserwacji. Szczegółowe informacje dotyczące sygnalizowanych tu zagadnień można znaleźć między innymi w: Welfe, 2003; 2009; Gajda, 2004; Gruszczyński, Podgórska, 2004; Larose, 2006; Walesiak, Gatnar, 2009; Szymczak, 2010; Gruszczyński, 2012.

**Rozwiązanie**

Model skonstruowany został na podstawie próby przekrojowej. Stopień zróżnicowania zbiorowości jest w takiej sytuacji znacznie większy niż przy próbie czasowej, co z kolei wiąże się z niższym dopasowaniem do danych empirycznych.

W modelu dysponujemy danymi dotyczącymi następujących zmiennych:

- *Wynagr* – bieżące wynagrodzenie roczne (w $);
- *Wykszt* – liczba lat nauki szkolnej;
- *Wiek* – wiek (w latach);
- *Staz* – staż pracy (w miesiącach);
- *Mniejsz* – przynależność do mniejszości etnicznych (przy czym wartość 1 oznacza, że dany pracownik należy do mniejszości etnicznej, a 0, że nie należy do mniejszości etnicznej).

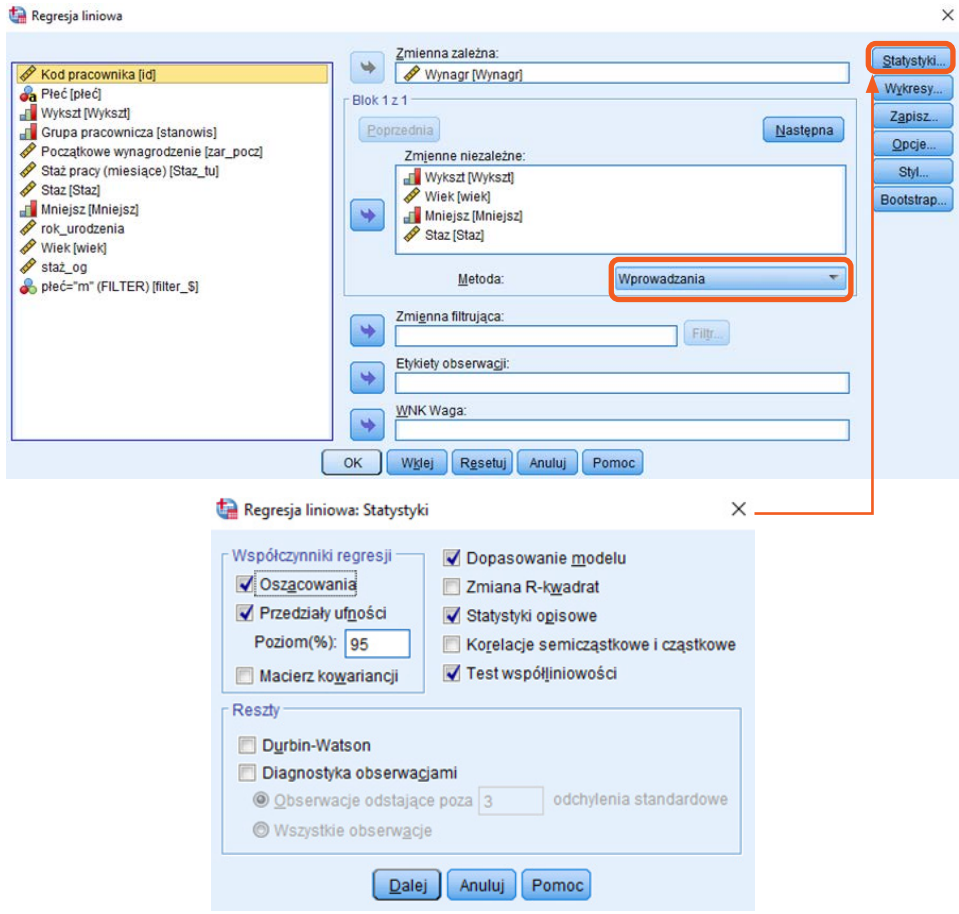
Zwróćmy uwagę na to, że wykształcenie badane jest nie poprzez określenie poziomu wykształcenia, tylko poprzez liczbę lat nauki szkolnej. Jest to sposób bardzo często wykorzystywany w badaniach społecznych. Co istotne, takie podejście zapewnia ilościowy pomiar tej jakościowej cechy. Z kolei zmienna *Mniejsz*, która jest jakościowa, jest zmienną zero-jedynkową, a więc jej pomiar jest quasi-ilościowy, co umożliwia jej włączenie do modelu regresji.

Problem badawczy można zatem opisać w sposób następujący:

- zmienna objaśniana: *y – Wynagr* (zmienna mierzona na skali ilościowej);
- zmienne objaśniające:
  - $x_1$ – *Wykszt (zmienna mierzona na skali ilościowej)*;
  - $x_2$ – *Wiek (zmienna mierzona na skali ilościowej)*;
  - $x_3$ – *Staz (zmienna mierzona na skali ilościowej)*;
  - $x_4$ – *Mniejsz (zmienna zero-jedynkowa, quasi-ilościowa)*.

Zakładając linową relację między analizowanymi zjawiskami, z wykorzystaniem KMNK można skonstruować liniowy model regresji opisujący poziom wynagrodzeń.

Aby skonstruować liniowy model regresji, w IBM SPSS Statistics wybieramy: *Analiza → Regresja → Liniowa*. W polu *Zmienna zależna* wprowadzamy zmienną *Wynagr*, a w polu *Zmienna niezależna* wszystkie cztery zmienne objaśniające: *Wykszt*, *Wiek*, *Staz*, *Mniejsz* (w dowolnej kolejności). Aby możliwe było dokładniejsze zbadanie własności modelu, wybieramy również: *Statystyki → Oszacowania*, *Przedziały ufności (95%)*, *Dopasowanie modelu*, *Test współliniowości* (rysunek 7.5). Zauważmy, że została wybrana metoda wprowadzania (*Metoda → Wprowadzania*), a więc wszystkie wyżej wymienione zmienne objaśniające zostaną włączone do budowanego modelu (alternatywnie można wybrać metody krokowe – szerzej na ten temat w dalszej części rozdziału).




**Rysunek 7.5.** Okno polecenia *Regresja liniowa* – estymacji modelu wynagrodzeń: *y* = *f*(*Wykszt, Wiek, Mniejsz, Staz*)

Wyniki zestawiono na rysunku 7.6.

A

**Statystyki opisowe**

| | Średnia | Odchylenie standardowe | N |
|---|---|---|---|
| Wynagr | 40922.30 | 17879.378 | 257 |
| Wykszt | 14.42 | 2.984 | 257 |
| Wiek | 38.1128 | 9.93096 | 257 |
| Mniejsz | .25 | .433 | 257 |
| Staz | 111.84 | 109.849 | 257 |

B

**Korelacje**

| | | Wynagr | Wykszt | Wiek | Mniejsz | Staz |
|---|---|---|---|---|---|---|
| Korelacja Pearsona | Wynagr | 1.000 | .673 | -.127 | -.280 | -.231 |
| | Wykszt | .673 | 1.000 | -.322 | -.287 | -.418 |
| | Wiek | -.127 | -.322 | 1.000 | .121 | .942 |
| | Mniejsz | -.280 | -.287 | .121 | 1.000 | .150 |
| | Staz | -.231 | -.418 | .942 | .150 | 1.000 |
| Istotność (jednostronna) | Wynagr | . | .000 | .021 | .000 | .000 |
| | Wykszt | .000 | . | .000 | .000 | .000 |
| | Wiek | .021 | .000 | . | .026 | .000 |
| | Mniejsz | .000 | .000 | .026 | . | .008 |
| | Staz | .000 | .000 | .000 | .008 | . |
| N | Wynagr | 257 | 257 | 257 | 257 | 257 |
| | Wykszt | 257 | 257 | 257 | 257 | 257 |
| | Wiek | 257 | 257 | 257 | 257 | 257 |
| | Mniejsz | 257 | 257 | 257 | 257 | 257 |
| | Staz | 257 | 257 | 257 | 257 | 257 |

C

**Model - podsumowanie[b]**

| Model | R | R-kwadrat | Skorygowane R-kwadrat | Błąd standardowy oszacowania | Statystyka Durbina-Watsona |
|---|---|---|---|---|---|
| 1 | .693[a] | .480 | .471 | 12999.475 | 1.154 |

a. Predyktory: (Stała), Staz, Mniejsz, Wykszt, Wiek

b. Zmienna zależna: Wynagr

D

**Anova[a]**

| Model | | Suma kwadratów | df | Średni kwadrat | F | Istotność |
|---|---|---|---|---|---|---|
| 1 | Regresja | 3.925E+10 | 4 | 9812877971 | 58.069 | .000[b] |
| | Reszta | 4.258E+10 | 252 | 168986340.1 | | |
| | Ogółem | 8.184E+10 | 256 | | | |

a. Zmienna zależna: Wynagr

b. Predyktory: (Stała), Staz, Mniejsz, Wykszt, Wiek

E

**Współczynniki[a]**

| Model | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | | | 95,0% przedział ufności dla B | | Statystyki współliniowości | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Błąd standardowy | Beta | t | Istotność | Dolna granica | Górna granica | Tolerancja | VIF |
| 1 | (Stała) | -33651.540 | 7881.713 | | -4.270 | .000 | -49173.963 | -18129.117 | | |
| | Wykszt | 3833.955 | 317.914 | .640 | 12.060 | .000 | 3207.848 | 4460.062 | .734 | 1.363 |
| | Wiek | 676.014 | 250.000 | .375 | 2.704 | .007 | 183.658 | 1168.371 | .107 | 9.338 |
| | Mniejsz | -3970.278 | 1958.926 | -.096 | -2.027 | .044 | -7828.232 | -112.325 | .916 | 1.091 |
| | Staz | -49.210 | 23.558 | -.302 | -2.089 | .038 | -95.605 | -2.816 | .099 | 10.145 |

a. Zmienna zależna: Wynagr

**Rysunek 7.6.** Wyniki estymacji modelu wynagrodzeń: *y* = *f*(*Wykszt, Wiek, Mniejsz, Staz*)

Tabele *Statystyki opisowe* (rysunek 7.6A) i *Korelacje* (rysunek 7.6B) pokazują się po zaznaczeniu opcji *Statystyki opisowe*. Na ich podstawie możemy przeanalizować między innymi proste związki między poszczególnymi zmiennymi objaśniającymi a zmienną objaśnianą oraz pomiędzy poszczególnymi zmiennymi objaśniającymi. Powinniśmy

włączyć do modelu zmienne objaśniające, które są jak najsilniej powiązane ze zmienną objaśnianą oraz jak najsłabiej między sobą. Współczynniki odczytane z pierwszego wiersza tabeli *Korelacje* mierzą zależność między wynagrodzeniami a poszczególnymi zmiennymi objaśniającymi.

Widzimy, że wszystkie zmienne objaśniające są istotnie powiązane ze zmienną objaśnianą ($p < 0{,}05$), a siła związku jest dość duża dla zmiennej *Wykszt*, niezbyt wysoka dla *Mniejsz* i *Staz* i niska dla *Wiek*. Również pomiędzy poszczególnymi zmiennymi objaśniającymi zależności są statystycznie istotne. Analizując te współczynniki korelacji, zauważamy jednak problem – siła korelacji między dwiema zmiennymi objaśniającymi – *Wiek* i *Staz* – jest bardzo wysoka: $r = 0{,}942$ ($p < 0{,}001$). Może to skutkować współliniowością zmiennych objaśniających. I rzeczywiście, analizując wyniki zamieszczone w ostatniej kolumnie tabeli E na rysunku 7.6, widzimy, że statystyka *VIF*, służąca do oceny współliniowości, ma – zgodnie z oczekiwaniami – wartość znacznie mniejszą od 10 dla dwóch zmiennych: 1,363 dla *Wykszt*, 1,091 dla *Mniejsz*. Jednak dla zmiennych *Wiek* i *Staz VIF* jest wysokie, a dla *Staz* – większe od 10. Oznacza to, że w modelu mamy do czynienia ze współliniowością zmiennych objaśniających.

Z uwagi na współliniowość zmiennych objaśniających model ten nie może zostać wykorzystany do celów praktycznych. Co należy zrobić w takiej sytuacji? W modelu nie powinny jednocześnie znaleźć się zmienne objaśniające, które są ze sobą zbyt silnie skorelowane, dlatego konstruujemy go ponownie, pomijając jedną z nich (tu – zmienną *Staz*, dla której *VIF* ma najwyższą wartość). Używamy tej samej ścieżki co poprzednio, tym razem jednak w modelu uwzględniamy trzy zmienne objaśniające: *Wykszt*, *Mniejsz* i *Wiek*. Wyniki zestawiono na rysunku 7.7.

A

**Statystyki opisowe**

| | Średnia | Odchylenie standardowe | N |
|---|---|---|---|
| Wynagr | 40922.30 | 17879.378 | 257 |
| Wykszt | 14.42 | 2.984 | 257 |
| Mniejsz | .25 | .433 | 257 |
| Wiek | 38.1128 | 9.93096 | 257 |

B

**Korelacje**

| | | Wynagr | Wykszt | Mniejsz | Wiek |
|---|---|---|---|---|---|
| Korelacja Pearsona | Wynagr | 1.000 | .673 | -.280 | -.127 |
| | Wykszt | .673 | 1.000 | -.287 | -.322 |
| | Mniejsz | -.280 | -.287 | 1.000 | .121 |
| | Wiek | -.127 | -.322 | .121 | 1.000 |
| Istotność (jednostronna) | Wynagr | . | .000 | .000 | .021 |
| | Wykszt | .000 | . | .000 | .000 |
| | Mniejsz | .000 | .000 | . | .026 |
| | Wiek | .021 | .000 | .026 | . |
| N | Wynagr | 257 | 257 | 257 | 257 |
| | Wykszt | 257 | 257 | 257 | 257 |
| | Mniejsz | 257 | 257 | 257 | 257 |
| | Wiek | 257 | 257 | 257 | 257 |

C

**Model - podsumowanie[b]**

| Model | R | R-kwadrat | Skorygowane R-kwadrat | Błąd standardowy oszacowania | Statystyka Durbina-Watsona |
|---|---|---|---|---|---|
| 1 | .686[a] | .471 | .464 | 13085.605 | 1.138 |

a. Predyktory: (Stała), Wiek, Mniejsz, Wykszt

b. Zmienna zależna: Wynagr

D

**Anova[a]**

| Model | | Suma kwadratów | df | Średni kwadrat | F | Istotność |
|---|---|---|---|---|---|---|
| 1 | Regresja | 3.851E+10 | 3 | 1.284E+10 | 74.974 | .000[b] |
| | Reszta | 4.332E+10 | 253 | 171233050.2 | | |
| | Ogółem | 8.184E+10 | 256 | | | |

a. Zmienna zależna: Wynagr

b. Predyktory: (Stała), Wiek, Mniejsz, Wykszt

E

**Współczynniki[a]**

| Model | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | | | 95,0% przedział ufności dla B | | Statystyki współliniowości | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Błąd standardowy | Beta | t | Istotność | Dolna granica | Górna granica | Tolerancja | VIF |
| 1 | (Stała) | -23784.907 | 6351.497 | | -3.745 | .000 | -36293.449 | -11276.365 | | |
| | Wykszt | 4064.090 | 300.191 | .678 | 13.538 | .000 | 3472.898 | 4655.282 | .834 | 1.199 |
| | Mniejsz | -4025.721 | 1971.725 | -.098 | -2.042 | .042 | -7908.806 | -142.637 | .916 | 1.091 |
| | Wiek | 185.994 | 87.022 | .103 | 2.137 | .034 | 14.614 | 357.375 | .896 | 1.117 |

a. Zmienna zależna: Wynagr

**Rysunek 7.7.** Wyniki estymacji modelu wynagrodzeń: *y* = *f*(*Wykszt, Wiek, Mniejsz*)

Oszacowane równanie można zapisać:

$$\widehat{Wynagr}_i = -23784{,}9 + 4064{,}1 * Wykszt - 4025{,}7 * Mniejsz + 186{,}0 * Wiek.$$

Ocenę modelu zaczynamy od ANOVA (rysunek 7.7D). Statystyka $F$ przy $df_1 = 3$ i $df_2 = 253$ wynosi 74,974, co zapisujemy: $F(3; 253) = 74{,}974$. Prawdopodobieństwo

w ANOVA jest bliskie zera ($p < 0{,}001$), a więc jest niższe niż $\alpha = 0{,}05$ ($p < \alpha$). Odrzucamy więc $H_0$ (przypomnijmy, $H_0$: $R^2 = 0$, $H_1$: $R^2 > 0$), a skoro tak, to współczynnik determinacji w populacji istotnie różni się od zera (na czym nam zależy).

Obliczony dla próby współczynnik determinacji ma wartość 0,471 (*R-kwadrat*, rysunek 7.7C), zmiany wynagrodzeń zostały zatem wyjaśnione modelem (a konkretnie – zmiennością zmiennych objaśniających) w 47,1%. W przypadku prób przekrojowych rzadko uzyskuje się współczynnik determinacji przekraczający 0,5, a tym samym, jeśli jest on bliski tej wartości, dopasowanie modelu uznaje się za dość dobre. Tak też można przyjąć w tym przypadku. Niemniej jednak 52,9% zmienności wynagrodzeń wynika z innych czynników niż uwzględnione w modelu, prognozowanie poziomu wynagrodzeń na podstawie tego modelu nie powinno być zatem prowadzone.

Średni błąd szacunku wynosi 13085,6$ (*Błąd standardowy oszacowania*, rysunek 7.7C), a więc przewidując poziom wynagrodzeń na podstawie tego modelu, mylimy się średnio o 13085,6$, co przy średniej wynagrodzeń na poziomie 40922,3$ (rysunek 7.7A) stanowi około 32%. Błąd ten często pomija się w analizach, jako kluczowy wskazując współczynnik determinacji.

Na rysunku 7.7C zamieszczono również wartość statystyki Durbina-Watsona (*DW* = 1,138), niemniej jednak w tym modelu ją pomijamy (z uwagi na charakter próby – próba przekrojowa).

Na podstawie wyników zaprezentowanych w tabeli *Współczynniki* (rysunek 7.7E) dokonamy oceny relacji między poszczególnymi zmiennymi objaśniającymi a zmienną objaśnianą. Poszczególne wiersze odpowiadają konkretnym zmiennym objaśniającym.

Po pierwsze, stosując test t-Studenta, oceniamy istotność statystyczną parametrów strukturalnych (interesują nas współczynniki regresji). Przypomnijmy, hipotezy są postaci: $H_0$: $\beta_j \neq 0$, $H_1$: $\beta_j \neq 0$.

Dla zmiennej *Wykszt* $p < 0{,}001$ (a więc $p < \alpha$), odrzucamy zatem $H_0$, a za prawdziwą uznajemy $H_1$. Współczynnik regresji w populacji istotnie różni się od zera. Można zatem wnioskować, że zmienna *Wykszt* jest istotnie (w sensie statystycznym) powiązana z poziomem wynagrodzeń. Z kolei dla zmiennej *Mniejsz* $p = 0{,}042$, a dla *Wiek* $p = 0{,}034$. W przypadku obu zmiennych $p < \alpha$, a więc również te zmienne są istotnie (w sensie statystycznym) powiązane z poziomem wynagrodzeń.

W kolejnym kroku interpretujemy współczynniki regresji z próby (*B*) (*Współczynniki niestandaryzowane B* – rysunek 7.7E).

Dla zmiennej *Wykszt B* = 4064,1, a więc w przypadku osób, których liczba lat nauki jest o rok większa, poziom wynagrodzeń jest wyższy średnio o 4064,1$ (przy założeniu stałego poziomu pozostałych zmiennych objaśniających).

Dla zmiennej *Wiek B* = 186,0, a więc w przypadku osób starszych o rok poziom wynagrodzeń jest wyższy średnio o 186$ (przy założeniu stałego poziomu pozostałych zmiennych objaśniających).

Dla zmiennej *Mniejsz B* = –4025,7, a więc w przypadku osób należących do mniejszości etnicznych poziom wynagrodzeń jest niższy średnio o 4025,7$ niż dla pozostałych pracowników (przy założeniu stałego poziomu pozostałych zmiennych objaśniających).

Zwróćmy uwagę, że błąd szacunku parametrów dla zmiennych (*Współczynniki niestandaryzowane*, *Błąd standardowy* – rysunek 7.7E) jest stosunkowo niski.

Która z tych zmiennych jest najważniejszą determinantą wynagrodzeń w tej firmie? Ustalimy to, porównując wartości standaryzowanych współczynników regresji (*Współczynniki standaryzowane Beta* – rysunek 7.7E). Biorąc pod uwagę wartość bezwzględną z *Beta*, można wskazać, że największe znaczenie ma w tym zakresie wykształcenie (dla *Wykszt Beta* = 0,678). *Ceteris paribus*, wpływ wieku i przynależności do mniejszości jest analogiczny, choć kierunek tej relacji jest odwrotny.

Odnosząc się do współliniowości zmiennych objaśniających, w tym modelu nie mamy z nią do czynienia (dla wszystkich zmiennych objaśniających *VIF* < 10 – kolumna *VIF* na rysunku 7.7E).

Podsumowując powyższą analizę, zaznaczmy, że prezentując wyniki badania prowadzonego z zastosowaniem analizy regresji, zwykle przedstawia się je w formie tabelarycznej (tabela 7.1).

**Tabela 7.1.** Wyniki estymacji modelu wynagrodzeń: $y = f(Wykszt, Wiek, Mniejsz)$

| Wyszczególnienie | *B* | *S*(*B*) | *Beta* | *t* | *p* | *VIF* |
|---|---|---|---|---|---|---|
| Stała | –23784,9 | 6351,5 | | –3,745 | < 0,0001* | |
| Wykszt | 4064,1 | 300,2 | 0,678 | 13,538 | < 0,0001* | 1,199 |
| Mniejsz | –4025,7 | 1971,7 | –0,098 | –2,042 | 0,042* | 1,091 |
| Wiek | 186,0 | 87,0 | 0,103 | 2,137 | 0,034* | 1,117 |

$F(3; 253) = 74,974$, $p < 0,001$; $R^2 = 0,471$; $S_e = 13085,6$.

*B* – współczynnik regresji z próby, *S*(*B*) – błąd szacunku współczynnika regresji, *Beta* – standaryzowany współczynnik regresji, *t* – statystyka t-Studenta, *p* – prawdopodobieństwo w teście t-Studenta, * – zależność istotna statystycznie ($\alpha = 0,05$), *VIF* – statystyka współliniowości, $R^2$ – współczynnik determinacji.

**Źródło:** opracowanie własne.

Podsumowując wyniki, można powiedzieć, że wszystkie z analizowanych cech – wiek, wykształcenie i przynależność do mniejszości etnicznych – są istotnie powiązane z poziomem wynagrodzeń w tym przedsiębiorstwie. *Ceteris paribus*, wynagrodzenia są wyższe w przypadku osób o wyższym poziomie wykształcenia i starszych (rosną z wiekiem i liczbą lat edukacji), niższe zaś w przypadku mniejszości niż pozostałych pracowników. Największe znaczenie odgrywa przy tym wykształcenie. Wartość współczynnika determinacji wskazuje, że poziom wynagrodzeń w podobnym stopniu kształtowany jest przez te trzy charakterystyki pracowników i przez inne czynniki, nieuwzględnione w tym modelu.

## 7.5. Selekcja zmiennych objaśniających w modelach regresji

W kontekście istotności statystycznej parametrów strukturalnych modelu pojawia się pytanie, czy za dobry uznać model zawierający również nieistotne statystycznie zmienne objaśniające, czy też należy je wyłączyć z modelu. Są tu dwa podejścia:

- pierwsze zakłada, że skoro weryfikujemy konkretny model teoretyczny, to wszystkie zmienne objaśniające (niezależnie od tego, czy w zestawieniu z innymi zmiennymi są istotne, czy też nie) należy uwzględnić w modelu;
- drugie zakłada, że nieistotne zmienne objaśniające należy z modelu wyłączyć; w sytuacji, gdy dana zmienna objaśniająca nie jest istotna statystycznie, to błąd jej szacunku jest relatywnie wysoki (a więc ocena relacji między tą zmienną objaśniającą a zmienną objaśnianą jest silnie nieprecyzyjna), a jej włączenie do modelu jedynie niepotrzebnie go rozbudowuje (niepotrzebnie, gdyż rezygnujemy w ten sposób z uproszczenia opisu relacji, a i tak zmienna ta w niewielkim tylko stopniu poprawia stopień wyjaśnienia zmiennej objaśnianej); jak podkreśla na przykład Maddala (2008), nadmierna liczba zmiennych objaśniających jest niekorzystna z uwagi na:
  - ryzyko nadmiernej współliniowości zmiennych objaśniających i związanych z tym problemów;
  - wprowadzenie do modelu niepotrzebnej informacji (szumu) i niecelowej utraty stopni swobody, czego skutkiem jest zwiększona wariancja parametrów modelu (pomimo małego obciążenia);
  - trudności w interpretacji najbardziej znaczącego wpływu zmiennych objaśniających na zmienną objaśnianą.

Przyjmując drugie z powyższych podejść, można zastosować metody pozwalające na zautomatyzowanie procedury konstruowania modelu „optymalnego", wykorzystując wybraną metodę krokową, na etapie której wyłączane (albo włączane) są krok po kroku kolejne zmienne objaśniające. Kolejność ich włączania/wyłączania określana jest zwykle przez prawdopodobieństwo w teście istotności poszczególnych parametrów strukturalnych (poza wyrazem wolnym, rzecz jasna).

Jak to zrobić? IBM SPSS Statistics przewiduje w tym zakresie pięć możliwości. **Metoda wprowadzania** oznacza (jak już wiadomo), że wszystkie zmienne w określonym bloku są jednocześnie wprowadzane do modelu (uzyskujemy więc model zgodny z pierwszym podejściem, w którym znajdą się wszystkie wprowadzone potencjalne predyktory badanego zjawiska). Na drugim biegunie mamy **metodę usuwania**, po której zastosowaniu wszystkie zmienne są jednocześnie usuwane z modelu, uzyskujemy model zawierający wyłącznie stałą (wyraz wolny). Dodatkowo SPSS umożliwia wybór metody:

- eliminacji wstecznej – po wprowadzeniu wszystkich zmiennych usuwana jest zmienna spełniająca kryteria usunięcia, aż do wyczerpania się zmiennych spełniających kryteria; wychodzimy zatem od pełnego zestawu zmiennych objaśniających, które na początku założyliśmy (uwzględniliśmy w naszym modelu teoretycznym), w kolejnym kroku usuwana jest zmienna najsłabiej powiązana z badanym zjawiskiem, w kolejnym kroku następna zmienna objaśniająca, dla której istotność związku jest słaba itd., aż do momentu, gdy w modelu nie znajdą się wyłącznie istotne zmienne; raz usunięta zmienna w kolejnych krokach nie wraca już do modelu;
- selekcji postępującej – wprowadzanie do modelu kolejno zmiennych spełniających kryteria wprowadzenia, zaczynając od zmiennej, która w najwyższym stopniu spełnia przyjęte kryterium, aż do wyczerpania się zmiennych spełniających kryteria; wychodzimy zatem od modelu uwzględniającego najważniejszą determinantę badanego zjawiska – najważniejszą spośród tych zmiennych objaśniających, które na początku założyliśmy (uwzględniliśmy w naszym modelu teoretycznym), w kolejnym kroku włączana jest zmienna, której siła powiązania z badanym zjawiskiem jest nieco niższa od „głównego predyktora" (plasująca się na drugim miejscu), w kolejnym kroku włączana jest następna zmienna objaśniająca, dla której istotność związku jest nieco mniejsza itd., aż do momentu, gdy włączenie kolejnej zmiennej oznaczałoby uwzględnienie nieistotnego czynnika; raz wprowadzona zmienna w kolejnych krokach zostaje już w modelu;
- krokowej – zarówno metoda eliminacji wstecznej, jak i metoda selekcji postępującej to metody krokowe; wybór tej opcji wiąże się z tym, że program sam wybierze jeden z wariantów (albo metodę selekcji postępującej, albo eliminacji wstecznej), dodatkowo zezwalając na włączenie/wyłączenie zmiennej, która w poprzednich krokach była już usunięta/wprowadzona (pracujemy w każdym kroku na pełnym zestawie zmiennych).

**Przykład 7.4**

Wykorzystajmy dane użyte w przykładzie 7.3, przy czym zamiast zmiennej *Staz* włączamy zmienną *Staz_akt*. Stosujemy tym razem metodę krokową. Podobnie jak poprzednio dokonajmy interpretacji wyników.

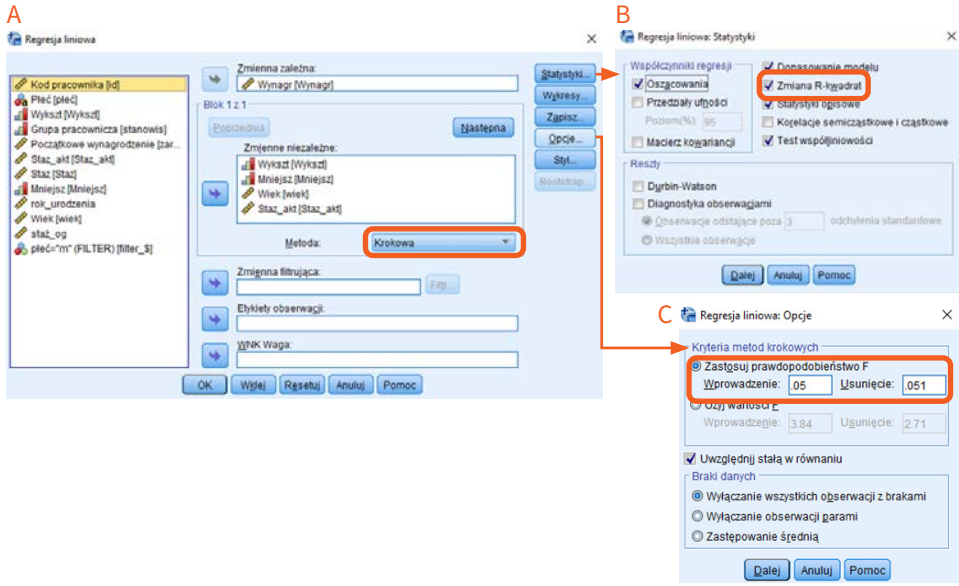
**Rozwiązanie**

W modelu dysponujemy danymi dotyczącymi następujących zmiennych:

- *Wynagr* – bieżące wynagrodzenie roczne (w $), skala ilościowa;
- *Wykszt* – liczba lat nauki szkolnej, skala ilościowa;

- *Wiek* – wiek (w latach), skala ilościowa;
- *Staz_akt* – staż pracy w aktualnym miejscu pracy (w miesiącach), skala ilościowa;
- *Mniejsz* – przynależność do mniejszości etnicznych (1 – pracownik należy do mniejszości etnicznej, 0 – dla pozostałych).

Podobnie jak w przykładzie 7.3 wybieramy: *Analiza* → *Regresja* → *Liniowa*. W polu *Zmienna zależna* wprowadzamy zmienną *Wynagr*, a w polu *Zmienna niezależna* – *Wykszt*, *Wiek*, *Staz_akt*, *Mniejsz*. Następnie przechodzimy do pola *Statystyki* → *Oszacowania*, *Przedziały ufności (95%)*, *Dopasowanie modelu*, *Test współliniowości*, dodatkowo zaznaczamy *Zmiana R-kwadrat* (rysunek 7.8B). Wybieramy metodę krokową (*Metoda* → *Krokowa*). Z uwagi na zastosowanie metody krokowej, która może (przy automatycznym wyborze) polegać na usuwaniu kolejnych zmiennych objaśniających lub na wprowadzaniu kolejnych, warto ujednolicić kryteria ich wprowadzania do modelu. W tym celu wybieramy *Opcje* → *Kryteria metod krokowych*, a następnie zmieniamy alfa na jak najbardziej zbliżone – w polu *Usunięcie* wpisujemy 0,051 (program nie przyjmuje 0,05, co stanowiłoby naturalne ujednolicenie kryteriów włączania zmiennych objaśniających do modelu przy metodzie selekcji postępującej i eliminacji wstecznej) (rysunek 7.8C).




**Rysunek 7.8.** Okno polecenie *Regresja liniowa* → *Krokowa*

Wyniki zestawiono na rysunku 7.9. Model zbudowany został metodą selekcji postępującej. W pierwszym kroku włączona została zmienna najsilniej powiązana ze zmienną objaśnianą – *Wykszt* (współczynnik korelacji $r = 0{,}673$, rysunek 7.9B), w drugim dołączona została zmienna *Wiek*, a w trzecim – zmienna *Mniejsz.* Postępowanie zakończyło się na trzecim kroku, do modelu nie została włączona zmienna *Staz_akt*. Interpretując

wyniki analizy, odnosimy się tylko do ostatniego kroku (stosowne elementy zaznaczono na rysunku 7.9 obramowaniem).

A

**Statystyki opisowe**

| | Średnia | Odchylenie standardowe | N |
|---|---|---|---|
| Wynagr | 40922.30 | 17879.378 | 257 |
| Wykszt | 14.42 | 2.984 | 257 |
| Mniejsz | .25 | .433 | 257 |
| Wiek | 38.1128 | 9.93096 | 257 |
| Staz_akt | 81.78 | 10.324 | 257 |

B

**Korelacje**

| | | Wynagr | Wykszt | Mniejsz | Wiek | Staz_akt |
|---|---|---|---|---|---|---|
| Korelacja Pearsona | Wynagr | 1.000 | .673 | -.280 | -.127 | .035 |
| | Wykszt | .673 | 1.000 | -.287 | -.322 | -.033 |
| | Mniejsz | -.280 | -.287 | 1.000 | .121 | -.017 |
| | Wiek | -.127 | -.322 | .121 | 1.000 | .055 |
| | Staz_akt | .035 | -.033 | -.017 | .055 | 1.000 |
| Istotność (jednostronna) | Wynagr | . | .000 | .000 | .021 | .291 |
| | Wykszt | .000 | . | .000 | .000 | .298 |
| | Mniejsz | .000 | .000 | . | .026 | .395 |
| | Wiek | .021 | .000 | .026 | . | .191 |
| | Staz_akt | .291 | .298 | .395 | .191 | . |
| N | Wynagr | 257 | 257 | 257 | 257 | 257 |
| | Wykszt | 257 | 257 | 257 | 257 | 257 |
| | Mniejsz | 257 | 257 | 257 | 257 | 257 |
| | Wiek | 257 | 257 | 257 | 257 | 257 |
| | Staz_akt | 257 | 257 | 257 | 257 | 257 |

C

**Model - podsumowanie**

| Model | R | R-kwadrat | Skorygowane R-kwadrat | Błąd standardowy oszacowania | Statystyki zmiany | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Zmiana R-kwadrat | F zmiany | df1 | df2 | Istotność F zmiany |
| 1 | .673 [a] | .453 | .451 | 13250.491 | .453 | 211.102 | 1 | 255 | .000 |
| 2 | .680 [b] | .462 | .458 | 13166.973 | .009 | 4.245 | 1 | 254 | .040 |
| 3 | .686 [c] | .471 | .464 | 13085.605 | .009 | 4.169 | 1 | 253 | .042 |

a. Predyktory: (Stała), Wykszt
b. Predyktory: (Stała), Wykszt, Wiek
c. Predyktory: (Stała), Wykszt, Wiek, Mniejsz

D

**Anova[a]**

| Model | | Suma kwadratów | df | Średni kwadrat | F | Istotność |
|---|---|---|---|---|---|---|
| 1 | Regresja | 3.706E+10 | 1 | 3.706E+10 | 211.102 | .000 [b] |
| | Reszta | 4.477E+10 | 255 | 175575519.5 | | |
| | Ogółem | 8.184E+10 | 256 | | | |
| 2 | Regresja | 3.780E+10 | 2 | 1.890E+10 | 109.017 | .000 [c] |
| | Reszta | 4.404E+10 | 254 | 173369178.8 | | |
| | Ogółem | 8.184E+10 | 256 | | | |
| 3 | Regresja | 3.851E+10 | 3 | 1.284E+10 | 74.974 | .000 [d] |
| | Reszta | 4.332E+10 | 253 | 171233050.2 | | |
| | Ogółem | 8.184E+10 | 256 | | | |

a. Zmienna zależna: Wynagr
b. Predyktory: (Stała), Wykszt
c. Predyktory: (Stała), Wykszt, Wiek
d. Predyktory: (Stała), Wykszt, Wiek, Mniejsz

E

**Współczynniki**[a]

| Model | | Współczynniki niestandaryzowane | | Współczynniki standaryzowane | | | Statystyki współliniowości | |
|---|---|---|---|---|---|---|---|---|
| | | B | Błąd standardowy | Beta | t | Istotność | Tolerancja | VIF |
| 1 | (Stała) | -17250.126 | 4088.214 | | -4.219 | .000 | | |
| | Wykszt | 4032.995 | 277.576 | .673 | 14.529 | .000 | 1.000 | 1.000 |
| 2 | (Stała) | -26908.091 | 6202.876 | | -4.338 | .000 | | |
| | Wykszt | 4226.100 | 291.314 | .705 | 14.507 | .000 | .896 | 1.115 |
| | Wiek | 180.322 | 87.519 | .100 | 2.060 | .040 | .896 | 1.115 |
| 3 | (Stała) | -23784.907 | 6351.497 | | -3.745 | .000 | | |
| | Wykszt | 4064.090 | 300.191 | .678 | 13.538 | .000 | .834 | 1.199 |
| | Wiek | 185.994 | 87.022 | .103 | 2.137 | .034 | .896 | 1.117 |
| | Mniejsz | -4025.721 | 1971.725 | -.098 | -2.042 | .042 | .916 | 1.091 |

a. Zmienna zależna: Wynagr

F

**Zmienne wykluczone**[a]

| Model | | Beta w modelu | t | Istotność | Korelacja cząstkowa | Statystyki współliniowości | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Tolerancja | VIF | Minimalna tolerancja |
| 1 | Mniejsz | -.094[b] | -1.961 | .051 | -.122 | .917 | 1.090 | .917 |
| | Wiek | .100[b] | 2.060 | .040 | .128 | .896 | 1.115 | .896 |
| | Staz_akt | .057[b] | 1.232 | .219 | .077 | .999 | 1.001 | .999 |
| 2 | Mniejsz | -.098[c] | -2.042 | .042 | -.127 | .916 | 1.091 | .834 |
| | Staz_akt | .053[c] | 1.144 | .254 | .072 | .997 | 1.003 | .895 |
| 3 | Staz_akt | .050[d] | 1.093 | .276 | .069 | .996 | 1.004 | .833 |

a. Zmienna zależna: Wynagr

b. Predyktory w modelu: (Stała), Wykszt

c. Predyktory w modelu: (Stała), Wykszt, Wiek

d. Predyktory w modelu: (Stała), Wykszt, Wiek, Mniejsz

**Rysunek 7.9.** Wyniki estymacji modelu wynagrodzeń metodą krokową

Oszacowane równanie można zapisać:

$$\widehat{Wynagr}_i = -23784{,}9 + 4064{,}1 * Wykszt - 4025{,}7 * Mniejsz + 186{,}0 * Wiek\,.$$

Jak widać, model ten jest identyczny z omawianym w przykładzie 7.3, nie będziemy go zatem szczegółowo analizować. Odnieśmy się tylko do oceny zmiany współczynnika determinacji (*Statystyki zmiany* rysunek 7.9C). W kolumnie *Zmiana R-kwadrat* podano, o ile zmienia się współczynnik determinacji w porównaniu z poprzednim krokiem, a następnie oceniono istotność tej zmiany (zmiany są istotne statystycznie, jeśli *p* w teście *F* – odczytywane z kolumny *istotność F zmiany* – jest niższe od $\alpha$). Jak widać, współczynnik determinacji w kolejnych krokach rósł i zmiana ta była statystycznie istotna (model zbudowany w drugim kroku jest lepszy niż pierwszym itd.).

Podsumowując, tak jak podkreślono wcześniej, niniejszy rozdział stanowi wprowadzenie do analizy regresji. Oprócz sygnalizowanych zagadnień, tj. sprawdzenia własności składnika losowego oraz identyfikacji przypadków odstających i wpływowych, analizie poddaje się również efekty pośrednie zmiennych objaśniających (efekt moderacji, mediacji czy supresji[29]). Zagadnienia te zostaną w tym miejscu pominięte.

29 Szerzej na ten temat na przykład w: Solecki, b.r.; Szymczak, 2010; Bedyńska, Książek, 2012; Hayes, 2013.

# Bibliografia


Aczel A.D. (2000), *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa.

Aczel A.D., Sounderpandian J. (2018), *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa.

Agresti A. (2007), *An introduction to categorical data analysis*, Wiley, Hoboken.

Agresti A., Finlay B. (2014), *Statistical methods for the social sciences*, Pearson, London.

Agresti A., Franklin Ch. (2013), *Statistics. The art and science of learning from data*, Pearson, Boston.

Babbie E. (2006), *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN, Warszawa.

Bedyńska S., Brzezicka A. (red.) (2007), *Statystyczny drogowskaz. Praktyczny poradnik analizy danych w naukach społecznych na przykładach z psychologii*, Wydawnictwo Szkoły Wyższej Psychologii Społecznej „Academica", Warszawa.

Bedyńska S., Cypryańska M. (red.) (2013a), *Statystyczny drogowskaz 1. Praktyczne wprowadzenie do wnioskowania statystycznego*, Wydawnictwo Akademickie SEDNO, Szkoła Wyższa Psychologii Społecznej, Warszawa.

Bedyńska S., Cypryańska M. (red.) (2013b), *Statystyczny drogowskaz 2. Praktyczne wprowadzenie do analizy wariancji*, Wydawnictwo Akademickie SEDNO, Szkoła Wyższa Psychologii Społecznej, Warszawa.

Bedyńska S., Książek M. (2012), *Statystyczny drogowskaz 3. Praktyczny przewodnik wykorzystania modeli regresji oraz równań strukturalnych*, Wydawnictwo Akademickie SEDNO, Szkoła Wyższa Psychologii Społecznej, Warszawa.

Blalock H.M. (1975), *Statystyka dla socjologów*, Państwowe Wydawnictwo Naukowe, Warszawa.

Borgatta E.F., Bohrnstedt G.W. (1980), *Level of measurement – Once over again*, „Sociological Methods and Research", no. 9, s. 147–160.

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, Państwowe Wydawnictwo Naukowe, Warszawa.

Brzezińska A.I., Rycielski P., Sijko K. (2010), *Wyzwania metodologiczne. Diagnoza potrzeb i ewaluacja wsparcia wśród osób z ograniczeniami sprawności*, Wydawnictwo Naukowe Scholar, Warszawa.

Bulmer M.G. (1979), *Principles of statistics*, Dover Publications Inc., New York.

Cieciura M., Zacharski J. (2007), *Metody probabilistyczne w ujęciu praktycznym*, VIZJA PRESS & IT, Warszawa.

Cohen J. (1988), *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum Associates, New York.

Diagnoza Społeczna (b.r.), http://www.diagnoza.com/ (dostęp: 15.05.2019).

Dobrowolska B., Grzelak M.M., Jarczyński J. (2017), *Praktyczne aspekty analizy danych w biznesie*, Wydawnictwo Biblioteka, Łódź.

Domański Cz., Pruska K. (2000), *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Domański Cz., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014), *Testy statystyczne w procesie podejmowania decyzji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Field A. (2009), *Discovering statistics using SPSS*, Sage, Los Angeles.

Frankfort-Nachmias Ch., Nachmias D. (2001), *Metody badawcze w naukach społecznych*, Zysk i S-ka, Poznań.
Gajda J.B. (2004), *Ekonometria*, Wydawnictwo C.H. Beck, Warszawa.
Gamst G., Meyers L.S., Guarino A.J. (2008), *Analysis of variance designs. A conceptual and computational approach with SPSS and SAS*, Cambridge University Press, Cambridge.
Glantz S.A., Slinker B.K., Neilands T.B. (2001), *Primer of applied regression & analysis of variance*, McGraw-Hill Education, New York.
Goryl A., Jędrzejczyk Z., Kukuła K. (2009), *Wprowadzenie do ekonometrii*, Wydawnictwo Naukowe PWN, Warszawa.
Góralski A. (1974), *Metody opisu i wnioskowania statystycznego w psychologii*, Państwowe Wydawnictwo Naukowe, Warszawa.
Górniak J., Wachnicki J. (2000), *Pierwsze kroki w analizie danych. SPSS PL for Windows*, SPSS Polska, Kraków.
Górniak J., Wachnicki J. (2008), *Pierwsze kroki w analizie danych. SPSS for Windows*, SPSS Polska, Kraków.
Greń J. (1972), *Modele i zadania statystyki matematycznej*, Państwowe Wydawnictwo Naukowe, Warszawa.
Gruszczyński M., Podgórska M. (2004), *Ekonometria*, Oficyna Wydawnicza SGH – Szkoła Główna Handlowa w Warszawie, Warszawa.
Gruszczyński M. (red.) (2012), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Oficyna a Wolters Kluwer business, Warszawa.
Grzelak M.M. (2009), *Zróżnicowanie, asymetria i koncentracja*, [w:] W. Starzyńska (red.), *Podstawy statystyki*, Wydawnictwo Difin, Warszawa, s. 127–163.
Hayes A.F. (2013), *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*, Guilford Press, New York.
Hellwig Z. (1998), *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, Wydawnictwo Naukowe PWN, Warszawa.
Hershberger S.L., Fisher D.G. (2005), *Measures of association*, [w:] B.S. Everitt, D.C. Howell (red.), *Encyclopedia of Statistics in Behavioral Science*, vol. 3, John Wiley & Sons, Chichester.
Howell D.C. (2010), *Statistical Methods for Psychology*, Wadsworth Cengage Learning, Belmont.
Jaworska A. (2004), *Główne nurty w metodologii badań nad skutecznością psychoterapii – w poszukiwaniu „trzeciej drogi"*, [w:] J. Brzeziński (red.), *Metodologia badań psychologicznych. Wybór tekstów*, Wydawnictwo Naukowe PWN, Warszawa, s. 116–147.
Keppel G., Wickens T. D. (2004), *Design and analysis. A researcher's handbook*, Pearson Prentice Hall, New Jersey.
Kończak G. (2016), *Testy permutacyjne. Teoria i zastosowania*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.
Kordos J. (1988), *Jakość danych statystycznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.
Kowal J. (1998), *Metody statystyczne w badaniach sondażowych rynku*, Wydawnictwo Naukowe PWN, Warszawa.
Krzewińska A., Grzeszkiewicz-Radulska K. (2013), *Klasyfikacja sondażowych technik otrzymywania materiałów*, „Przegląd Socjologiczny", nr 62(1), s. 9–31.
Kufel T. (2007), *Rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa.
Lange O. (1952), *Teoria statystyki*, cz. I, Polskie Wydawnictwa Gospodarcze, Warszawa.
Larose D.T. (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Lubke G.H., Muthen B.O. (2004), *Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons*, „Structural Equation Modeling", no. 11, s. 514–534.
Maddala G.S. (2008), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.
Malarska A. (2005), *Analiza statystyczna wspomagana programem SPSS*, SPSS Polska, Kraków.
McClave J.T., Sincich T. (2018), *Statistics*, Pearson, New York.
Nawojczyk M. (2002), *Przewodnik po statystyce dla socjologów*, SPSS Polska, Kraków.
Nowak E. (red.) (2001), *Metody statystyczne w analizie przedsiębiorstwa*, Polskie Wydawnictwo Ekonomiczne, Warszawa.
Nowak S. (2007), *Metodologia badań społecznych*, Wydawnictwo Naukowe PWN, Warszawa.
Olsson U. (1979), *On the Robustness of Factor Analysis against Crude Classification of the Observations*, „Multivariate Behavioral Research", vol. 14(4), s. 485–500.
*PS IMAGO PRO. Podręcznik użytkownika* (2014), Predictive Solutions, Kraków.
Pułaska-Turyna B. (2005), *Statystyka dla ekonomistów*, Wydawnictwo Difin, Warszawa.
Reed J.F., Stark D.B. (1988), *Robust alternative to traditional analysis of variance*, „Computer Methods and Programs in Biomedicine", vol. 26, s. 233–238.
Royston P. (1995), *A Remark on Algorithm AS 181: The W-test for Normality*, „Applied Statistics", no. 44, s. 547–551.
Rószkiewicz M. (2011), *Analiza klienta*, SPSS Polska, Kraków.
Rószkiewicz M., Perek-Białas J., Węziak-Białowolska D., Zięba-Pietrzak A. (2013), *Projektowanie badań społeczno-ekonomicznych. Rekomendacje i praktyka badawcza*, Wydawnictwo Naukowe PWN, Warszawa.
Sarata J. (b.r.), *Co buduje związek? O testach z dla proporcji kolumnowych*, e-biuletyn Predictive Solutions, https://support.predictivesolutions.pl/index.php?/Knowledgebase/Article/View/547/0/co-buduje-zwizek-o-testach-z-dla-proporcji-kolumnowych (dostęp: 15.05.2019).
Sarndal C.E., Swenson B., Wretman J. (1997), *Model Assisted Survey Sampling*, Springer, New York.
Sawiński Z. (2010), *Zastosowania tablic w badaniach zjawisk społecznych*, Wydawnictwo Instytutu Filozofii i Socjologii Polskiej Akademii Nauk, Warszawa.
Shapiro S.S., Wilk M.B. (1965), *An Analysis of variance test for normality (Complete samples)*, „Biometrika", no. 52, s. 591–611.
Sobczyk M. (1998), *Statystyka*, Wydawnictwo Naukowe PWN, Warszawa.
Sobczyk M. (2000), *Statystyka. Podstawy teoretyczne. Przykłady – zadania*, Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.
Sobczyk M. (2013), *Ekonometria*, Wydawnictwo C.H. Beck, Warszawa.
Solecki P. (b.r.), *Mediator czy moderator – trzecia zmienna w analizie*, Predictive Solutions blog, https://predictivesolutions.pl/mediator-czy-moderator-trzecia-zmienna-analizie (dostęp: 30.06.2020).
Starzyńska W. (2020), *Statystyka praktyczna*, Wydawnictwo Naukowe PWN, Warszawa.
Starzyńska W. (red.) (2009), *Podstawy statystyki*, Wydawnictwo Difin, Warszawa.
Steczkowski J. (1995), *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych*, Państwowe Wydawnictwo Naukowe, Warszawa – Kraków.
Stevens S.S. (1951), *Mathematics, Measurement and Psychophysics*, [w:] S.S. Stevens (red.), *Handbook of Experimental Psychology*, John Wiley, New York, s. 1–49.
Szreder M. (2010a), *Losowe i nielosowe próby w badaniach statystycznych*, „Przegląd Statystyczny", R. LVII, z. 4, s. 168–174.
Szreder M. (2010b), *Metody i techniki sondażowych badań opinii*, Polskie Wydawnictwo Ekonomiczne, Warszawa.
Szwed R. (2008), *Metody statystyczne w naukach społecznych. Elementy teorii i zadania*, Wydawnictwo Katolickiego Uniwersytetu Lubelskiego, Lublin.

Szymczak W. (2010), *Podstawy statystyki dla psychologów*, Wydawnictwo Difin, Warszawa.
Szymczak W. (2018), *Podstawy statystyki dla psychologów*, Wydawnictwo Difin, Warszawa.
Walesiak M., Gatnar E. (red.) (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa.
Wątroba J. (2011), *Prosto o dopasowaniu prostych, czyli analiza regresji liniowej w praktyce*, Statsoft Polska, https://media.statsoft.pl/_old_dnn/downloads/analiza_regresji_liniowej_w_praktyce.pdf (dostęp: 25.06.2020).
Welfe A. (2003), *Ekonometria*, Polskie Wydawnictwo Ekonomiczne, Warszawa.
Welfe A. (2009), *Ekonometria. Metody i ich zastosowanie*, Polskie Wydawnictwo Ekonomiczne, Warszawa.
Wieczorkowska G., Wierzbiński J. (2007), *Statystyka. Analiza badań społecznych*, Wydawnictwo Naukowe Scholar, Warszawa.
Wiktorowicz J. (2004a), *Organizacja badania statystycznego*, [w:] W. Starzyńska (red.), *Podstawy statystyki*, Wydawnictwo Difin, Warszawa, s. 26–46.
Wiktorowicz J. (2004b), *Wiadomości wstępne*, [w:] W. Starzyńska (red.), *Podstawy statystyki*, Wydawnictwo Difin, Warszawa, s. 47–85.
Wiktorowicz J. (2016), *Międzypokoleniowy transfer wiedzy a wydłużanie okresu aktywności zawodowej*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
Wiktorowicz J. (2017), *Analiza statystyczna wyników badań naukowych – możliwości zastosowania rozwiązania PS IMAGO/PS IMAGO PRO (opartego na IBM SPSS Statistics)*, [w:] J. Wiktorowicz, A. Kubiak, A. Krzewińska, *Wprowadzenie do projektowania i analizy wyników badań naukowych*, materiały powielone, opracowanie na potrzeby warsztatów metodycznych zorganizowanych przez Katedrę Statystyki Ekonomicznej i Społecznej Uniwersytetu Łódzkiego i Predictive Solutions, Łódź.
Wywiał J.L. (2010), *Wprowadzenie do metody reprezentacyjnej*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice.
Wywiał J.L. (red.) (2003), *Metoda reprezentacyjna w badaniach ekonomiczno-społecznych*, t. I i II, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice.
Zając K. (1994), *Zarys metod statystycznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.
Zasępa R. (1972), *Metoda reprezentacyjna*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.

# Spis przykładów

# Spis rysunków

# Spis tabel