

Katolicki Uniwersytet Lubelski Jana Pawła II

Dr Krzysztof Jurek

Praktyczne wykorzystanie IBM SPSS Statistics (wersja 21 PL)

Kurs dla użytkowników początkujących i
średniozaawansowanych

Lublin

SPIS TREŚCI

Rozdział I. Wprowadzenie do analizy danych ilościowych z użyciem SPSS

- 1.1. Analiza danych – podstawowe pojęcia i definicje
- 1.2. Charakterystyka programu SPSS for Windows
- 1.3. Struktura i organizacja zbioru danych w SPSS
- 1.4. Wprowadzanie danych do programu SPSS
 - 1.4.1. Zasady kodowania danych
 - 1.4.2. Pytania wielokrotnego wyboru
- 1.5. Wczytywanie i zapisywanie zbioru danych
- 1.6. Okno raportów

Rozdział II. Przygotowanie zbioru danych do analizy

- 2.1. Zarządzanie zbiorami danych
 - 2.1.1. Łączenie zbiorów danych
 - 2.1.2. Dodawanie zmiennych
 - 2.1.3. Agregacja zbiorów danych
 - 2.1.4. Analiza danych w podzbiorach
 - 2.1.5. Sortowanie obserwacji w zbiorze danych
 - 2.1.6. Wybór obserwacji i ich ważenie
- 2.2. Przekształcanie danych
 - 2.2.1. Obliczanie wartości zmiennej
 - 2.2.2. Rekodowanie wartości zmiennych
 - 2.2.3. Zliczanie wystąpień wartości
 - 2.2.4. Rangowanie wartości zmiennych
- 2.3. Analiza rzetelności skali metodą Alfa Cronbacha

Rozdział III. Analiza częstości występowania zjawisk

- 3.1. Tworzenie tabel częstości
 - 3.1.1. Tabele częstości dla jednej zmiennej
 - 3.1.2. Tabele częstości dla dwóch i więcej zmiennych
- 3.2. Rodzaje procentowania
- 3.3. Interpretacja i opis danych tabelarycznych

Rozdział IV. Analiza opisowa danych

- 4.1. Miary tendencji centralnej
 - 4.1.1. Średnia arytmetyczna
 - 4.1.2. Średnia ważona
 - 4.1.3. Średnia harmoniczna i geometryczna
 - 4.1.4. Średnia obciążona
 - 4.1.5. Mediana i pozostałe kwartyle
 - 4.1.6. Dominanta
 - 4.1.7. Porównanie miar tendencji centralnej
- 4.2. Miary zmienności
 - 4.2.1. Rozstęp
 - 4.2.2. Wariancja i odchylenie standardowe
 - 4.2.3. Rozstęp ćwiartkowy i odchylenie ćwiartkowe

4.2.4. Współczynnik zmienności

4.2.5. Porównanie miar zmienności

4.3. Miary asymetrii i kurtozy

4.4. Standaryzacja wyników

Graficzna prezentacja wyników

W opracowaniu poniższego kursu wykorzystano następującą literaturę:

S. Bedyńska, M. Cypryańska, *Statystyczny drogowskaz t. 1*, Warszawa 2013.

S. Bedyńska, M. Cypryańska, *Statystyczny drogowskaz t. 2*, Warszawa 2013.

J. Górniak, J. Wachnicki, *Pierwsze kroki w analizie danych, SPSS for Windows*, Kraków 2010.

T. Pavkov, K. Pierce, *Do biegu, gotowi – start!*, *Wprowadzenie do SPSS dla Windows*, Gdańsk 2005.

D. Mider, A. Marcinkowska, *Analiza danych ilościowych dla politologów Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP*, Warszawa 2013.

ROZDZIAŁ I.

WPROWADZENIE DO ANALIZY DANYCH ILOŚCIOWYCH Z UŻYCIEM SPSS

Przedmiotem tej części kursu jest prezentacja programu SPSS, opisanie jego struktury i podstawowych funkcji. Poza tym przedstawiono sposób wprowadzania danych do programu, ich zapis oraz wczytywanie. Poza tym omówiono podstawowe pojęcia i definicje z zakresu analizy danych, ta elementarna wiedza jest niezbędna do zrozumienia omawianych w dalszej części kursu zagadnień.

1.1. Analiza danych – podstawowe pojęcia i definicje

Analiza danych to tylko jeden z etapów procesu badawczego, który możemy przedstawić za pomocą wykresu:



SFORMUŁOWANIE PROBLEMU BADAWCZEGO

Na tym etapie należy ustalić przedmiot , cel, metodę, populację, jednostkę statystyczną. Określić cechy – czyli zmienne określające badanie

ZBIERANIE DANYCH

Stosując odpowiednie narzędzie badawcze np. ankietę. W ten sposób otrzymuje się **MATERIAŁ STATYSTYCZNY**

OPRACOWANIE I PREZENTACJA MATERIAŁU STATYSTYCZNEGO

Obejmuje grupowanie jednostek badawczych w szersze jednostki; zliczanie – liczenie ile osób znajdzie się w poszczególnych grupach, kategoriach, grupowanie typologiczne – wyodrębnienie na podstawie pewnych typów.

OPIS I WNIOSKOWANIE STATYSTYCZNE

Wnioskowanie statystyczne pomaga formułować wnioski na podstawie obserwacji, z reguły wnioski oparte są na badaniach na próbie, wnioskowanie oparte jest na metodach statystyki indukcyjnej. Opis statystyczny to analiza rozkładu cechy za pomocą określonych procedur statystycznych których efektem są pewne charakterystyki liczbowe np. średnia arytmetyczna jako właściwość badanej zbiorowości. Nie można przekształcić zmiennej na niższym poziomie pomiaru w zmienną na poziomie wyższym.

Podstawowe pojęcia i definicje:

ZBIOROWOŚĆ STATYSTYCZNA:

To skończony lub nieskończony zbiór jednostek w stosunku do których będziemy formułować wnioski w badaniu pełnym.

PRÓBA:

Jest częścią populacji, wybrana na podstawie pewnych kryteriów, zbieramy dane w próbie, ale wnioski odnosimy do całej populacji.

JEDNOSTKA STATYSTYCZNA:

Element zbiorowości statystycznej np. osoba, gospodarstwo domowe, student

CECHA STATYSTYCZNA:

Właściwości jednostek zbiorowości. Warianty (odmiany) cech mogą być opisane słownie i wtedy mówimy o cechach jakościowych np. wykształcenie podstawowe, kolor oczu niebieski. Z kolei cechy ilościowe to takie, które występują z różnymi natężeniami i te można je zmierzyć np. wzrost 178 cm, waga 65kg.

PARAMETR – wartość liczbową, która charakteryzuje populację

ESTYMATOR – wartość cechy uzyskana w toku badania

Wśród cech ilościowych możemy wyróżnić:

skokowe: wartości cechy zmieniają się co „skok” np. wiek (19, 20, 21, 22) , zbiór liczb całkowitych; ciągłe: mogą przyjmować każdą wartość z określonego przedziału liczbowego – czyli mogą być mierzone z różną dokładnością np. wiek (19, 19 i 3 miesiące, 19 i 5 miesięcy), zbiór liczb rzeczywistych.

ZMIENNE

To zjawiska, które się zmieniają i przyjmują różne wartości np. poparcie polityczne dla partii, wiek, płeć. To co badamy to JEDNOSTKA ANALIZY, to nie jest zmienna ale obiekt, który zmienną charakteryzuje np. osoba, obszar, instytucja.

HIPOTEZA

Wymienia dwie zmienne, które pozostają w związku i określa ten związek np. Im wyższe wykształcenie tym większe zarobki respondenta

ZMIENNA NIEZALEŻNA

Zmienną tą uważamy za przyczynę zjawiska, to zmienna wyjaśniająca.

ZMIENNA ZALEŻNA

Zmienną tą uważamy za skutek zjawiska, to zmienna wyjaśniana.

Badaniu statystycznemu podlegają cechy zmienne – różnicujące jednostki statystyczne. Wielkość, nasilenie badanej cechy określana jest w procesie POMIARU. Wyróżnia się różne poziomy pomiarów, którym odpowiadają określone SKALE POMIAROWE.

SKALA NOMINALNA

Pomiar nominalny to zbiór nazw lub określeń cech. W kategorii zmiennej nominalnej możemy o 2 osobach powiedzieć tylko tyle, że są takie same lub różne. Uzyskujemy informacje, które umożliwiają podstawowe rozróżnienia i podział obiektów ze względu na odmiany, warianty a następnie przypisanie ich do określonych kategorii, np. przynależność do partii politycznych. Szczególnym przypadkiem są zmienne dychotomiczne dychotomiczne to jest dwuwartościowe (binarne, zero-jedynkowe) o charakterze dopełnienia logicznego. Przykładami takich wartości zmiennej są: tak - nie, posiada - nie posiada.

SKALA PORZĄDKOWA

W kategoriach zmiennej porządkowej można nie tylko określić nie tylko czy dwie osoby są takie same, czy też różne, lecz również stwierdzić, że jedna z nich jest „bardziej” / „mniej” niż druga np. bardziej zadowolona.

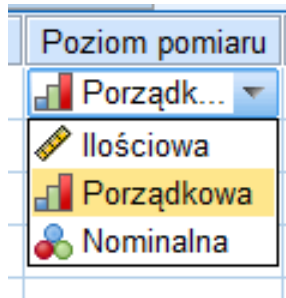
SKALA INTERWAŁOWA

Skala ta pozwala na wskazanie różnic między elementami - zaczyna się operowanie liczbami, możemy wykonywać operacje matematyczne. Możemy ocenić odległość między poszczególnymi wariantami poszczególnej cechy. Istotą pomiaru zmiennej na tym poziomie jest posiadanie przez nią umownego punktu zerowego, względem którego można zasadnie orzekać „o ile więcej” lub „o ile mniej”, lecz nie „ile razy więcej” lub „ile razy mniej”.

SKALA ILORAZOWA

To najwyższy (najsilniejszy) poziom pomiaru zmiennej, uzyskujemy największą możliwą do uzyskania ilość informacji na temat danego zjawiska. Na tym poziomie pomiaru najczęściej występują zmienne ciągłe. Można wykonywać wszystkie operacje matematyczne.

W programie SPSS skale ilorazową i interwałową określana są skalami ILOŚCIOWYMI.



ZASADA KUMULATYWNOŚCI

Każda kolejna skala ma wszystkie własności skal poprzednich. PRAKTYCZNE ZASTOSOWANIE SKAL POMIARU: pojawia się przede wszystkim na etapie analizy danych, ale należy je przewidywać na etapie tworzenia struktury projektu badawczego. Jeśli zmienna ma być używana na wiele sposobów wymagających różnych poziomów pomiaru badanie powinno być zaprojektowane tak, aby pozwoliło na osiągnięcie najwyższego wymaganego poziomu. Oto przykłady zmiany poziomu pomiaru zmiennej¹:

¹ D. Mider, A. Marcinkowska, *Analiza danych ilościowych dla politologów Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP*, Warszawa 2013, s. 83-84.

Lp.	Poziom ilorazowy pomiaru zmiennej	Poziom interwałowy pomiaru zmiennej	Poziom porządkowy pomiaru zmiennej	Poziom nominalny pomiaru zmiennej
1	Zmienna <i>temperatura</i> mierzona na poziomie ilorazowym - na skali bezwzględnej Kelwina: ... K	Zmienna <i>temperatura</i> mierzona na poziomie interwałowym - w stopniach Celsjusza: ...°C	Wartości zmiennej <i>temperatura</i> mierzona na poziomie przedziałowym: 1: niska temperatura 2: umiarkowana temperatura 3: wysoka temperatura	Wartości zmiennej <i>temperatura</i> mierzona na poziomie nominalnym: 1: temperatura zawiera się w przedziale bezpiecznym dla ludzkiego życia 2: temperatura nie zawiera się w przedziale bezpiecznym dla ludzkiego życia
2	Zmienna <i>wiek</i> mierzona na poziomie ilorazowym - wiek podawany w latach: 0: 0 lat 1: 1 rok ... 18: 18 lat 19: 19 lat 20: 20 lat ... n: 90 lat	Zmienna <i>wiek</i> mierzona na poziomie interwałowym - jako rok urodzenia: ... 1971 1972 1973 1974 1975 1976 ...	Wartości zmiennej <i>wiek</i> mierzona na poziomie porządkowym: 1: Młode pokolenie (od 18 do 35 lat) 2: Średnie pokolenie (powyżej 36 do 65 lat) 3: Starsze pokolenie (powyżej 65 lat)	Wartości zmiennej <i>wiek</i> mierzona na poziomie nominalnym: 1: Posiadamy informację na temat wieku respondenta 2: Nie posiadamy informacji na temat wieku respondenta
3	Zmienna <i>poparcie dla demokracji</i> mierzona na poziomie ilorazowym - na skali bezwzględnej od 0 do 100 proc.: ... %	Zmienna <i>poparcie dla demokracji</i> mierzona na poziomie interwałowym - na 10-punktowej od 1 do 10, gdzie: 1: popieram demokrację w najmniejszym stopniu 2: 3: 4: 5: 6: 7: 8: 9: 10: popieram demokrację w największym stopniu	Wartości zmiennej <i>poparcie dla demokracji</i> mierzona na poziomie porządkowym: 1: niskie poparcie dla demokracji 2: umiarkowane poparcie dla demokracji 3: wysokie poparcie dla demokracji	Wartości zmiennej <i>poparcie dla demokracji</i> mierzona na poziomie nominalnym: 1: popiera 2: nie popiera

Jednostki analizy i zmienne podlegające pomiarowi tworzą danych surowych. Ma ona postać tabeli. Powszechnie przyjmuje się, że w macierzy danych surowych zmienne umieszczane są w kolumnach, a jednostki analizy w wierszach. Innymi słowy liczba wierszy w tabeli odpowiada liczbie zbadanych przypadków, a liczba kolumn liczbie utworzonych zmiennych.

1.2. Charakterystyka programu SPSS for Windows

IBM SPSS Statistics Base to oprogramowanie do analiz statystycznych, które oferuje podstawowe funkcje potrzebne do przeprowadzenia procesu analitycznego od początku do końca. Jest łatwe w użyciu i oferuje szeroką gamę procedur i technik do zastosowań biznesowych i badawczych.

SPSS Statistics Base udostępnia niezbędne narzędzia statystyczne przydatne na każdym etapie procesu analitycznego.

- Wszechstronny zestaw procedur statystycznych umożliwiających prowadzenie dokładnych analiz.
- Wbudowane techniki przygotowywania danych do analizy.
- Wyrafinowane funkcje raportowania do sprawnego tworzenia wykresów.
- Zaawansowane możliwości wizualizacji pozwalające na przejrzyste przedstawienie istotnych spostrzeżeń.
- Obsługa wszelkiego typu danych, w tym bardzo dużych zbiorów danych².

IBM SPSS Statistics jest programem działającym w dominującym obecnie środowisku Microsoft Windows oraz w rzadziej spotykanych systemach operacyjnych Linux i MAC OS. W skład pakietu IBM SPSS Statistics wchodzi kilkanaście modułów: część z nich to moduły uniwersalne, takie jak IBM SPSS Statistics Base czy IBM SPSS Regression, inne są bardziej specjalistyczne i grupują statystyki najczęściej używane w pewnym obszarze wiedzy, np. IBM SPSS Direct Marketing czy IBM SPSS Forecasting.

W programie IBM SPSS Statistics dostępne są różne typy okien:

- Edytor danych. W tym oknie wyświetlana jest zawartość pliku danych. Korzystając z Edytora danych, można modyfikować pliki danych i tworzyć nowe. W razie otwarcia więcej niż jednego pliku danych, dla każdego pliku danych otwierane jest oddzielne okno Edytora danych.
- Edytor raportów. W Edytorze raportów wyświetlane są wszystkie wyniki statystyczne, tabele i wykresy. Wyniki można edytować oraz zapisywać. Edytor raportów otwiera się automatycznie, gdy po raz pierwszy zostaje wykonana procedura tworząca wyniki. Okno raportów ma własne okna edycyjne:
 - Edytor tabel przestawnych. W Edytorze tabel przestawnych wyniki wyświetlane w postaci tabel przestawnych mogą być modyfikowane na wiele różnych sposobów. Można edytować tekst, zamieniać miejscami dane w wierszach i kolumnach, dodawać kolory, tworzyć tabele wielowymiarowe oraz selektywnie ukrywać i wyświetlać wyniki.
 - Edytor wykresów. Edytor wykresów pozwala modyfikować wykresy wyświetlone w wysokiej rozdzielczości. Można zmieniać kolory, wybierać inne typy czcionek i ich wielkości, przełączać osie poziome i pionowe, obracać trójwymiarowe wykresy rozrzutu, a nawet zmieniać typy wykresu.
 - Edytor wyników tekstowych. Wyniki tekstowe, które nie są wyświetlane w tabelach przestawnych, mogą być modyfikowane w Edytorze wyników tekstowych. Można edytować wyniki i zmieniać właściwości czcionek (typ, styl, kolor i wielkość).
- Edytor komend. W oknie edytora komend można wklejać opcje wybrane w oknach dialogowych. Te opcje w Edytorze komend zostają wyświetlone w formie składni komend. Składnię komendy można następnie edytować w celu zastosowania różnych opcji, które nie są dostępne w oknach dialogowych. Komendy mogą być zapisywane w plikach, tak aby można było z nich korzystać w kolejnych sesjach.

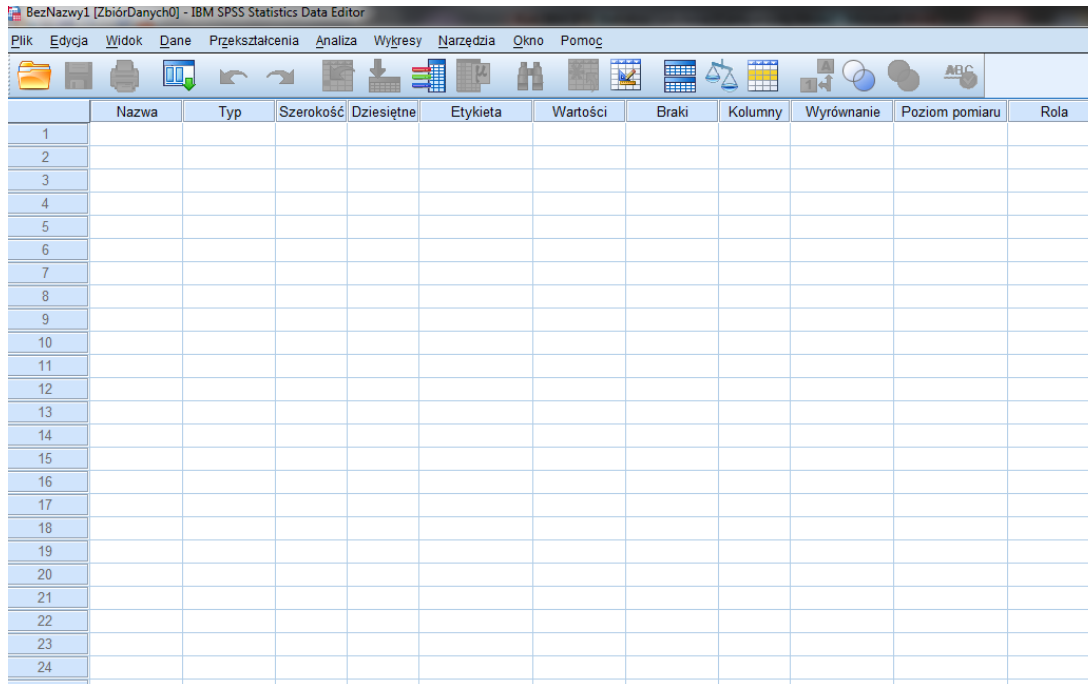
² <http://www-03.ibm.com/software/products/pl/spss-stats-base>

1.3. Struktura i organizacja zbioru danych w SPSS

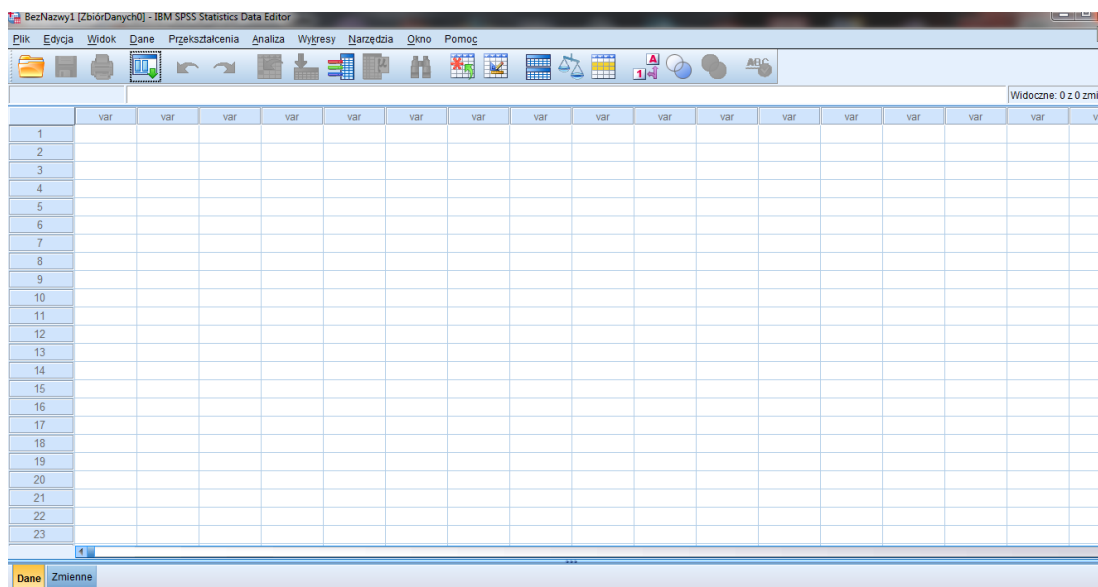
Struktura zbioru danych w edytorze danych:

ZMIENNE - użytkownik programu ma możliwość szczegółowego opisanie danych wprowadzanych do programu, nazw zmiennych i sposobu ich kodowania, możliwe opcje:

- **Nazwa zmiennej** - nie może przekraczać pewnej liczby znaków – wprowadzone nazwy muszą mieć długość 64 bajtów (zwykle to 64 znaki). W nazwie nie wolno stosować znaków specjalnych: spacji, oraz znaków: !, %, ^, &, *, +, /, -, =, ?, ' , ,, (). Nazwa zmiennej zawsze musi zaczynać się literą i nie może kończyć się kropką. Nie można wpisać dwóch zmiennych o tej samej nazwie.
- **TYP** - najczęściej stosowanym są zmienne liczbowe, które można opisać za pomocą **typu numerycznego**. Drugim typem są zmienne w postaci wpisanych do programu słów np. płeć, kolor oczu tj. zmienne jakościowe. Jeśli typ zmiennej został zdefiniowany jako **typ tekstowy**, to program SPSS nie jest w stanie przeprowadzić żadnych analiz (potrzebne są dane liczbowe).
- **SZEROKOŚĆ** - za pomocą strzałek z prawej strony okna możemy zwiększyć lub zmniejszyć szerokość danej zmiennej.
- **DZIESIĘTNE** - służy do ustawienia liczby miejsc dziesiętnych wyświetlanych w oknie DANE.
- **ETYKIETA** - nazwy zmiennych muszą być krótkie, w etykiecie nie ma tych ograniczeń. Tu można opisać badana zmienną.
- **WARTOŚCI** – ma istotne znaczenie dla czytelności wydruku, jaki pojawi się w Edytorze Raportów po wykonaniu analiz. Kolumna ta pozwala na zapisanie informacji o tym, jakimi wartościami są kodowane poszczególne opcje odpowiedzi. Wrócimy do tej opcji przy okazji omawiania procesu kodowania danych.
- **BRAKI** - podczas wprowadzania danych stykamy się z problemem braku odpowiedzi, powody mogą być różne np. przeoczenie, świadome opuszczenie odpowiedzi, udzielenie odpowiedzi niezgodnie z instrukcją. Przyjęło się oznaczać braki danych cyfrą 9 lub jej kombinacją, czyli 99, 999, 9999 itd. Trzeba pamiętać, że wartość oznaczająca brak danych musi wykraczać poza wartości zmiennej. Mamy różne sposoby wprowadzania braków danych: systemowe braki danych – automatycznie przypisywane przez SPSS pustym komórkom, zdefiniowane braki danych (badacz może podać przyczynę braków np. 98 to brak danych z powodu tego, że pytanie nie dotyczyło danego respondenta), przedział wartości plus wartość dyskretna
- **POZIOM pomiaru**. W SPSS mamy możliwość zdefiniowania 3 poziomów: NOMINALNY, PORZĄDKOWY i ILOŚCIOWY, deklaracja poziomu pomiaru ma znaczenie jedynie dla badacza. SPSS niestety nie potrafi rozpoznawać skal pomiarowych, niezależnie od tego, co wpisujemy w tej kolumnie.
- **ROLA** - do dyspozycji mamy następujące role: WEJŚCIE (dla zmiennej niezależnej lub predyktora), WYJŚCIE (dla zmiennej zależnej czy wyjaśnianej), ŁĄCZNIE (gdy zmienna może odgrywać obie role), BRAK (zmienna nie ma określonej roli), PODZIAŁ (zmienne definiujące podział na podzbiory), SEPARACJA (zmienne definiujące wybrane obserwacje). Domyślnie wszystkie zmienne mają zdefiniowaną rolę WEJŚCIE.



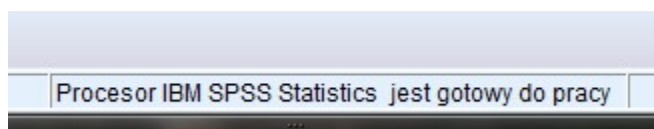
W zakładce **DANE** wpisujemy surowe dane, najlepiej zakodowane w postaci liczb.



W Edytorze Danych oprócz dwóch zakładki dane i zmienne znajdują się inne elementy tj.:

- Pasek stanu Jest to szary pasek z sześcioma okienkami, w którym znajdują się istotne informacje o tym, co aktualnie się dzieje w programie ('Procesor gotowy' oznacza, że SPSS nie wykonuje obliczeń; 'Wykonuje...' oznacza, że SPSS aktualnie realizuje nasze polecenie)
- STATUS OMS pokazuje informacje, czy został włączony system przekazywania wyników analiz do zewnętrznych plików
- Licznik przetwarzanych obserwacji - wskazuje ile obserwacji zostało już przeliczonych w danym momencie

- Status filtrowania obserwacji - informuje, czy analiza, którą wykonamy, zostanie przeprowadzona na wszystkich osobach badanych
- Informacja o ważeniu obserwacji
- Informacja o podziale zbioru danych na podgrupy



Pasek narzędzi w SPSS³:



1. OTWÓRZ PLIK. Funkcja ta pozwala otworzyć dowolny typ pliku IBM SPSS STATISTICS, a także pliki programów, z których możliwy jest import danych. Warto zwrócić uwagę na nazwy programów, które znajdują się w rozwijanej liście w okienku PLIKI TYPU. Możemy tam zobaczyć, że program IBM SPSS STATISTICS pozwala na import danych z Excela, danych tekstowych oraz innych programów statystycznych, takich jak Statistica lub SAS-a.
2. ZAPISZ PLIK. Gdy zapisujemy po raz pierwszy plik, ikonka ta działa jako ZAPISZ JAKO, a więc umożliwia zapisanie nowego pliku pod określoną nazwą w wybranej lokalizacji. Gdy plik ma już nadaną nazwę, funkcja ta spowoduje zapisanie ostatnich zmian.
3. DRUKUJ. Funkcja ta umożliwi nam wydrukowanie zawartości aktualnie otwartego okna. Możemy zatem na wydruku uzyskać surowe dane w formie tabeli, jeśli jesteśmy w zakładce DANE lub tabelę z opisem danych, jeśli aktywna jest zakładka ZMIENNE. Jeśli otwarty jest Edytor Raportów, ta funkcja pozwoli wydrukować wykonane obliczenia i wykresy.
4. PRZYWOŁAJ OKNO. Ta ikona pozwala na otwarcie ostatnio używanych funkcji i analiz. Docenimy ten przycisk, gdy okaże się, że musimy wykonać ponownie analizę, do której dostęp wymaga wielu kliknięć.
5. POWTÓRZ I COFNIJ. Tak jak w każdym edytorze tekstu również w programie IBM SPSS STATISTICS możliwe jest cofnięcie zmiany lub powrót do niej.
6. PRZEJDŹ DO OBSERWACJI. Gdy klikniemy ten przycisk, pojawi się okienko, w którym możemy wpisać numer poszukiwanej obserwacji. IBM SPSS STATISTICS przeniesie nas bezpośrednio do niej, tzn. umieści wiersz danych, w którym znajduje się ta obserwacja w pierwszym wierszu edytora danych widocznym na ekranie. Funkcja ta jest szczególnie przydatna, gdy mamy bardzo liczną próbę. Uwaga! Warto pamiętać, że numer, który mamy wpisać, jest numerem nadawanym przez IBM SPSS STATISTICS a nie dodaną przez nas zmienną „numer osoby badanej”.
7. PRZEJDŹ DO ZMIENNEJ. Ta funkcja umożliwi łatwe znalezienie odpowiedniej zmiennej – przydatna wtedy, gdy dane zawierają dużą liczbę zmiennych.
8. INFORMACJA O ZMIENNYCH. Kliknięcie tej ikonki spowoduje wyświetlenie opisu zmiennych, znajdujących się w pliku danych.
9. ZNAJDŹ. Funkcja ta pozwala na znalezienie szukanej wartości/litery/słowa w obrębie jednej zaznaczonej zmiennej. Najpierw musimy zaznaczyć zmienną klikając na jej nazwę (jeden klik), by została zaczerwniona, a szukana wartość zostanie pokazana na białym tle..

³ S. Bedyńska, M. Cypriańska, *Statystyczny drogowca t. 1*, Warszawa 2013, s. 64-66.

10. **WSTAW OBSERWACJĘ.** Dzięki tej opcji możemy wstawić nowy wiersz, czyli obserwację (osobę badaną). Jeśli zaznaczymy wiersz nr 9, to IBM SPSS STATISTICS wstawi tam nowy pusty wiersz, przesuwając poniższe obserwacje o jeden w dół. Przycisk jest nieaktywny, gdy znajdujemy się w zakładce ZMIENNE.
11. **WSTAW ZMIENNĄ.** Analogicznie do funkcji WSTAW OBSERWACJĘ przycisk ten dodaje nową zmienną. Aby wybrać miejsce pojawienia się nowej zmiennej, trzeba zaznaczyć odpowiednią kolumnę.
12. **PODZIEL DANE NA PODZBIORY.** Ikona ta otwiera okno umożliwiające podzielenie całej grupy osób badanych na mniejsze podgrupy na podstawie określonej zmiennej. Po wyborze tej funkcji na pasku stanu pojawi się informacja „Podziel według” i nazwa zmiennej określającej podział.
13. **WAŻENIE OBSERWACJI.** Funkcja umożliwia ważenie obserwacji
14. **WYBIERZ OBSERWACJE.** Ikona umożliwia szybki dostęp do okna dialogowego pozwalającego wybrać te obserwacje, które spełniają określone warunki, zdefiniowane na bazie zawartych w danych zmiennych.
15. **EYKIETY WARTOŚCI.** Kliknięcie na tę ikonę pozwala na wyświetlenie w Edytorze Danych w widoku Dane etykiet tekstowych zamiast wartości liczbowych.
16. **UŻYJ ZESTAWÓW.** Jeśli dane mają dużą liczbę zmiennych, można zdefiniować pewne ich części, nadając im określoną nazwę i tworząc zestaw. Dzięki temu można wygodniej orientować się w dużych plikach danych. Opcja UŻYJ ZESTAWÓW umożliwia przełączanie się między zdefiniowanymi wcześniej zestawami.
17. **POKAŻ WSZYSTKIE ZMIENNE.** Prezentuje listę wszystkich zmiennych
18. **SPRAWDZANIE PISOWNI.**

1.4. Wprowadzanie danych do programu SPSS

1.4.1. Zasady kodowania danych

Celem procesu kodowania jest takie spreparowanie danych, które umożliwia poddanie ich analizom statystycznym. Odpowiedzi na pytania, które zawarte są w ankietach (lub innych narzędziach badawczych) muszą być wprowadzone do SPSS. Pytania te mogą być otwarte tzn. zapisane w formie tekstowej, zamknięte, półotwarte.

Sposób kodowania ustala badacz, przypisując odpowiedziom osoby badanej wartości liczbowe, gdy mamy do czynienia ze zmiennymi jakościowymi np. odpowiedzi na pytanie o płeć, wykształcenie. Zmienne ilościowe nie muszą być specjalnie kodowane, bo występują już w postaci liczbowej, którą wpisujemy do danych.

Spróbujmy zakodować następujące pytania:

- Proszę podać swoją płeć
 - a) **Kobieta**
 - b) **Mężczyzna**
- Ile ma Pan/Pani lat? **25**
- Czy jesteś zadowolony z życia?

- a) Zdecydowanie tak
- b) Tak**
- c) Ani tak, ani nie
- d) Nie
- e) Zdecydowanie nie

Proces kodowania powinien przebiegać dwuetapowo. Najpierw należy opisać zmienne, ich etykiety i wartości w zakładce *Zmienne*, a dopiero potem wypełnić okno *Dane*. Zanim jednak przystąpimy do tego działania warto ponumerować wprowadzane ankiety, jeśli popełnimy gdzieś błąd np. wpiszemy złe dane, zawsze możemy go skorygować, odszukując odpowiednią ankietę po zapisanym na jej rogu numerze. Aby to było możliwe należy wprowadzić zmienną np. Numer respondenta NR.

	Nazwa	Typ	Szerokość	Dziesiętne	Etykieta	Wartości	Braki	Kolumny	Wyrównanie	Poziom pomiaru	Rola
1	NR	Numeryczna	6	2	Numer kolejnego respondenta	Brak	Brak	6	Z prawej	Ilościowa	Wejście
2											

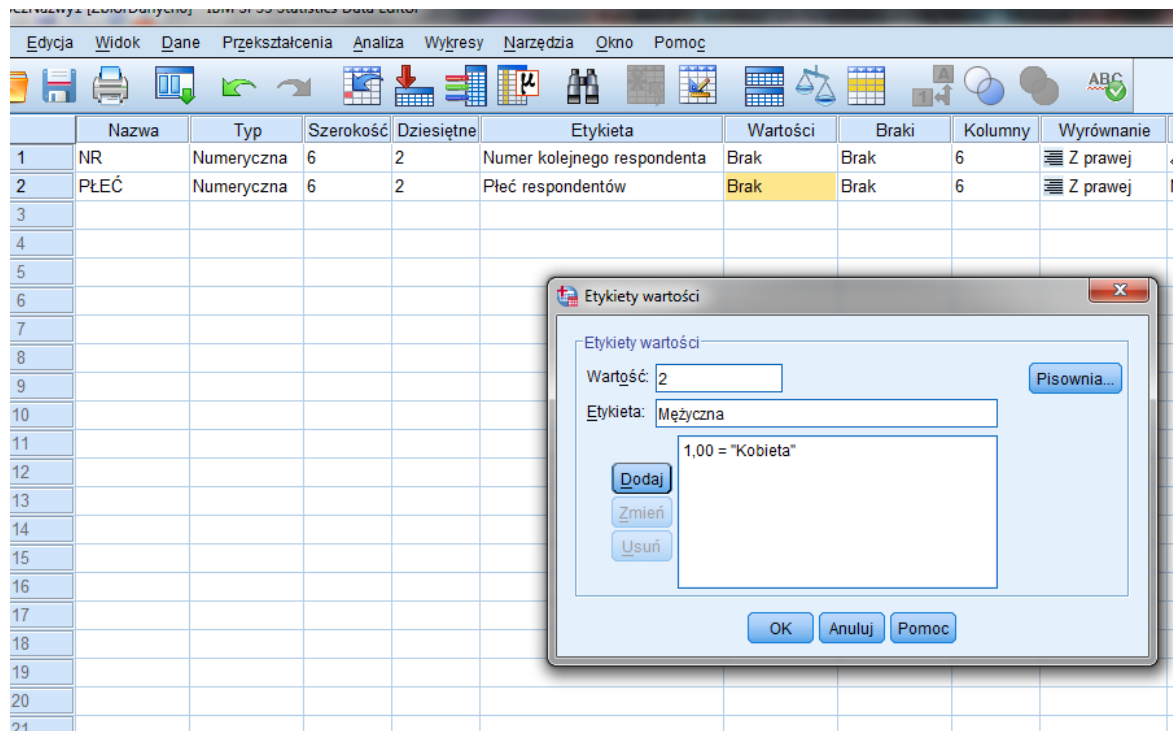
Wprowadzenie nazwy „włącza” pozostałe pola, które możemy zmieniać. Zmienna ta posłuży nam jedynie do uporządkowania danych i szybkiej identyfikacji błędów, zatem jedynie w polu etykieta wpiszemy co oznacza NR – Numer kolejnego respondenta. W oknie dane pojawiła się automatycznie kolumna z nazwą zmiennej. Tam będziemy wpisywać numery kolejnych ankiet: 1,2,3,4....

	NR	var	var
	1,00		
	2,00		
	3,00		

Kolejną zmienną, którą chcemy zdefiniować, jest płeć. Zmienna ta przyjmuje dwie możliwe wartości tj. kobieta i mężczyzna. Poszczególnym odpowiedziom musimy przyporządkować wartości liczbowe. Wartości te mogą być takie jakie chcemy, założymy, bez zbędnego kombinowania, że 1 to kobieta a 2 to mężczyzna. Po nadaniu etykiety, przypisane wartości wpisujemy w polu o tej właśnie nazwie.

Pamiętajmy, że zmienne jakościowe muszą zostać opisane za pomocą wartości, ponieważ wartości liczbowe są przypisane poszczególnym odpowiedziom arbitralnie. Przypisanie wartości powoduje, że mamy do czynienia ze zmienną typu numerycznego. Gdybyśmy

wybrali typ tekstowy zmiennej, i zamiast wartości tekstowych, posługiwali się słowami, możliwości SPSSa byłyby niewykorzystane. SPSS rozumie tylko cyfry i liczby.

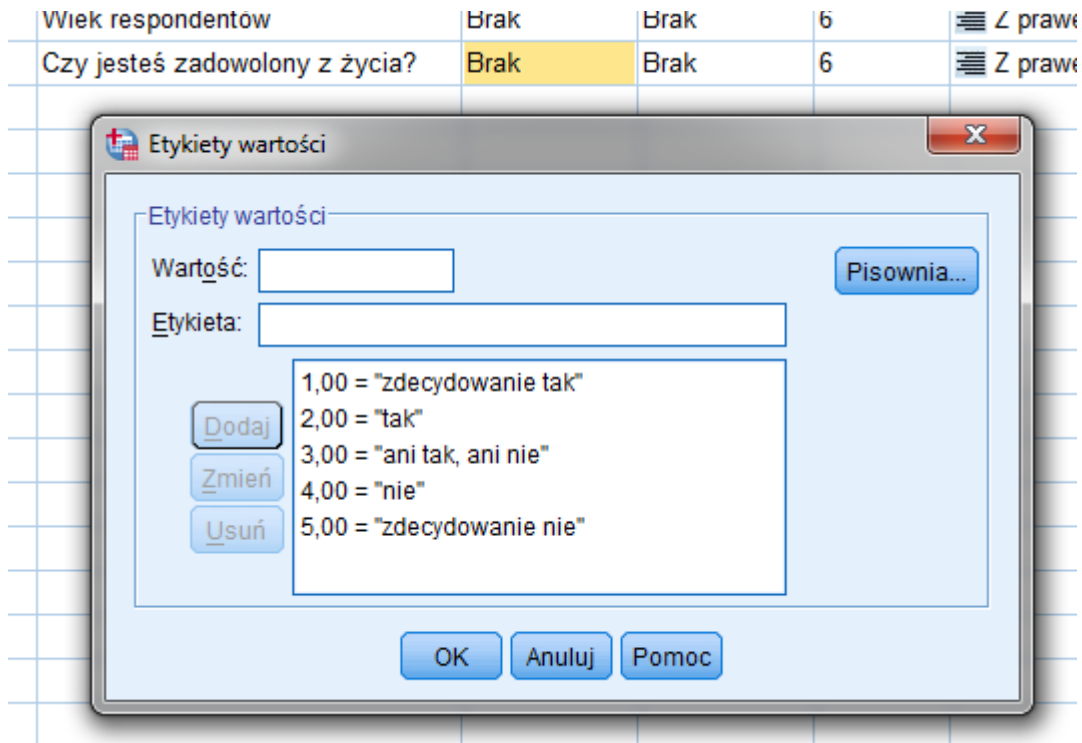


W oknie wartości posługujemy się intuicyjnym menu, w którym poza dodaniem wartości, przypisujemy jej etykietę. Pamiętajmy o przycisku dodaj. W każdym momencie możemy usunąć wartość lub zmienić jej etykietę.

Kolejne nasze pytanie dotyczy wieku. To pytanie otwarte, które kodujemy nieco inaczej. Zmienna ta nie musi mieć przypisanych wartości. Dlaczego? Sama dla siebie jest etykietą. Wpisywanie wartości i ich etykiet byłoby więc niepotrzebną pracą.

Nazwa	Typ	Szerokość	Dziesiętne	Etykieta	Wartości	Braki	Kolumny	Wyrównanie
NR	Numeryczna	6	2	Numer kolejnego respondenta	Brak	Brak	6	Z prawej
PŁEĆ	Numeryczna	6	2	Płeć respondentów	{1,00, Kobie...	Brak	6	Z prawej
WIEK	Numeryczna	6	2	Wiek respondentów	Brak	Brak	6	Z prawej

Ostatnie pytanie dotyczy poczucia zadowolenia, które jest oceniane na 5-stopniowej skali. Kodujemy je następująco: 1 – zdecydowanie tak, 2 – tak 3 – ani tak, ani nie, 4 – nie i 5 – zdecydowanie nie.



Po opisaniu wszystkich zmiennych przechodzimy do okna dane i wprowadzamy odpowiedzi z kwestionariusza.

	NR	PŁEĆ	WIEK	ZADOWOLENIE	v
1	1,00	1,00	25,00	2,00	
2					
3					

1.4.2. Pytania wielokrotnego wyboru

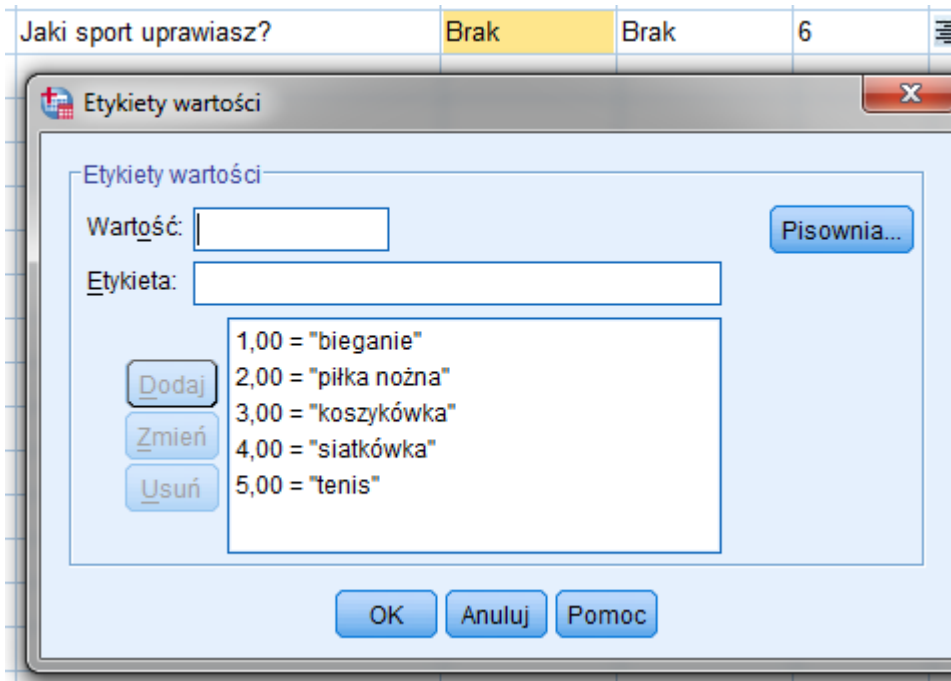
W praktyce badawczej mamy czasami do czynienia z pytaniami wielokrotnego wyboru, tzn. respondent mógł wybrać w danym pytaniu więcej niż 1 odpowiedź. Przykładowo mamy pytanie o uprawiany sport:

- Jaki sport Pan/Pani uprawia?
 - a) **Bieganie**
 - b) Piłka nożna
 - c) Koszykówka
 - d) Siatkówka
 - e) **Tenis**

W naszym przykładzie zaznaczone zostały odpowiedzi a i e.

Kodując odpowiedzi możemy postąpić dwojako:

Po pierwsze każdej dyscyplinie możemy nadać osobą wartość tzn. 1 to bieganie, 2 to piłka nożna....



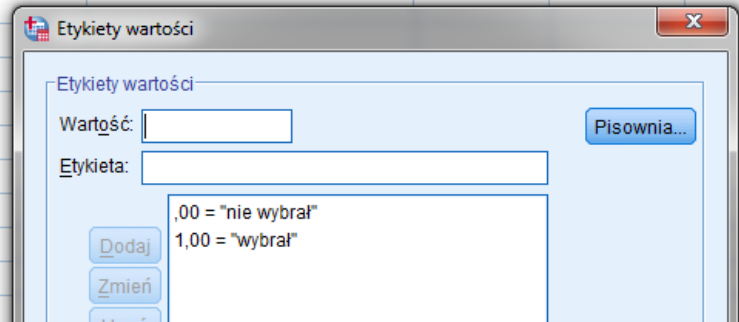
Należy jednak pamiętać, że każda odpowiedź musi być w osobnej kolumnie, a więc musi być tyle zmiennych ile wynosiła maksymalna liczba odpowiedzi. W naszym przypadku będą to dwie zmienne. W danych kodujemy odpowiedzi.

SPORT1	Numeryczna	6	2	Jaki sport uprawiasz?	{1,00, biega...	Brak	6
SPORT2	Numeryczna	6	2	Jaki sport uprawiasz?	{1,00, biega...	Brak	6

NR	PŁEĆ	WIEK	ZADOWOLENIE	SPORT1	SPORT2	var
1,00	1,00	25,00	2,00	1,00	5,00	

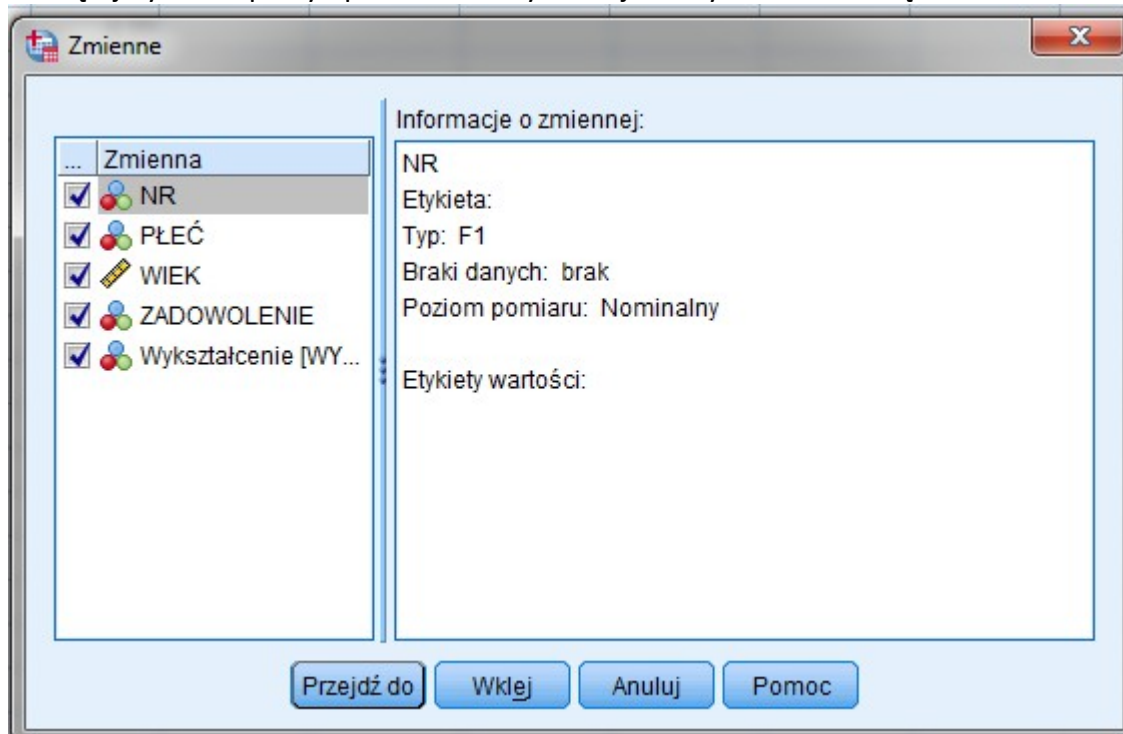
Drugi sposób to kodowanie dwoma wartościami gdzie jedna oznacza, że respondent wybrał daną odpowiedź, druga, że nie zrobił tego np. wartości 1 i 0. W tym przypadku musi być tyle zmiennych ile było możliwych odpowiedzi. W naszym przypadku 5.

SPORT1	Numeryczna	6	2	bieganie	Brak	Brak	6
SPORT2	Numeryczna	6	2	piłka nożna	Brak	Brak	6
SPORT3	Numeryczna	6	2	koszykówka	Brak	Brak	6
SPORT4	Numeryczna	6	2	siatkówka	Brak	Brak	6
SPORT5	Numeryczna	6	2	tenis	Brak	Brak	6

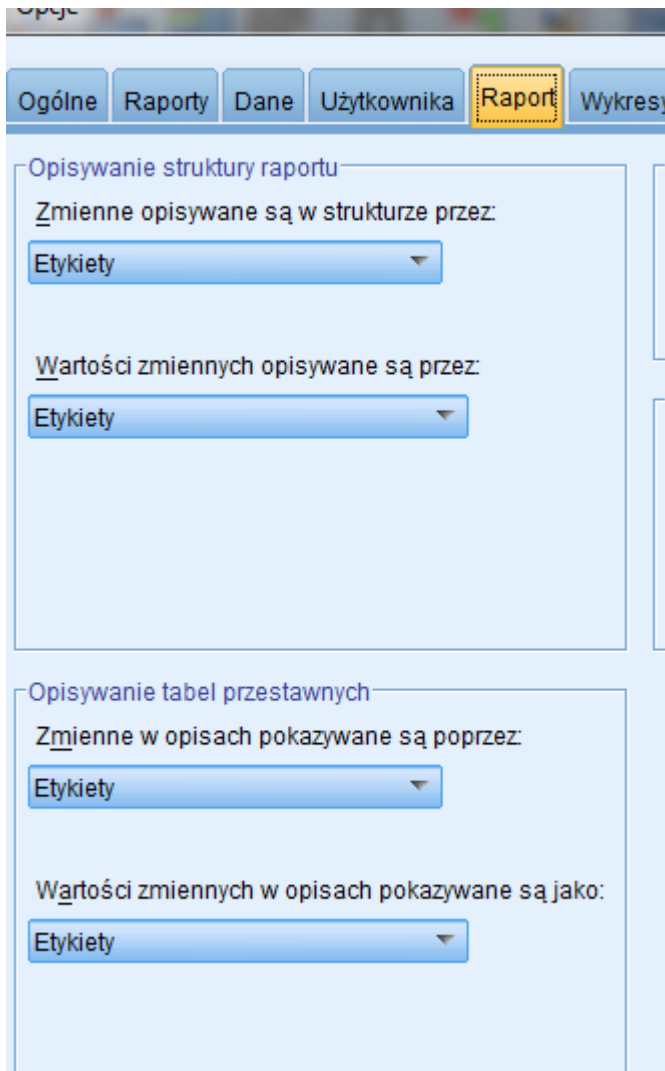


NR	PŁEĆ	WIEK	ZADOWOLENIE	SPORT1	SPORT2	SPORT3	SPORT4	SPORT5	var
1,00	1,00	25,00	2,00	1,00	,00	,00	,00	1,00	

Pamiętajmy też że pełny opis zbioru danych znajdziemy w menu *Narzędzia – zmienne*.



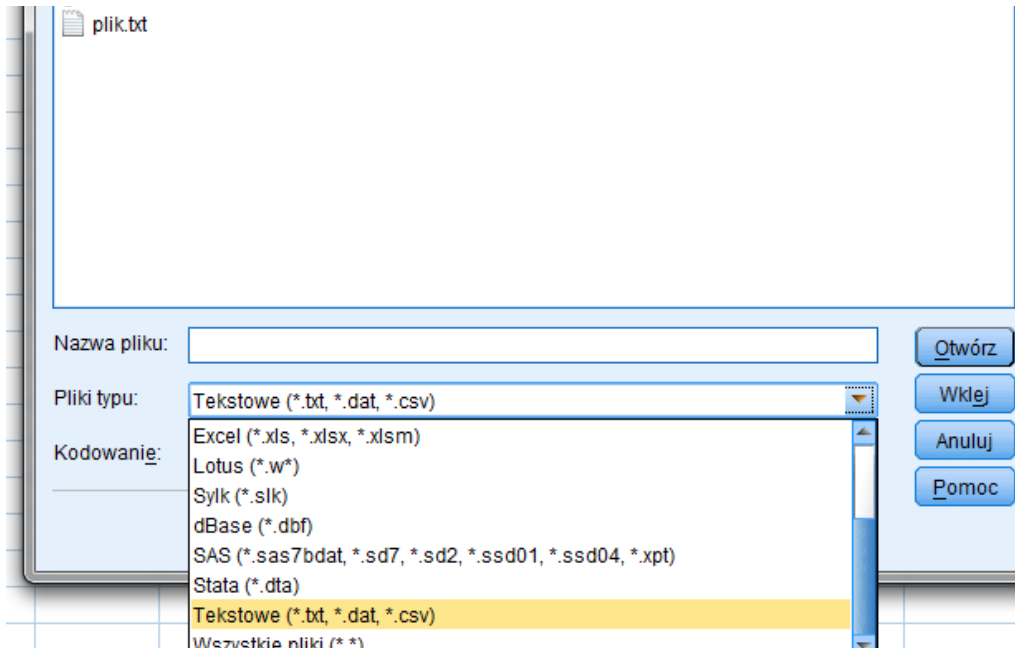
W SPSS zmienne mogą być różnie prezentowane zarówno w strukturze raportu np. z etykietami zmiennych oraz etykietami wartości lub bez tych etykiet. Opcje te znajdziemy klikając kolejno: Edycja – Opcje – Raport.



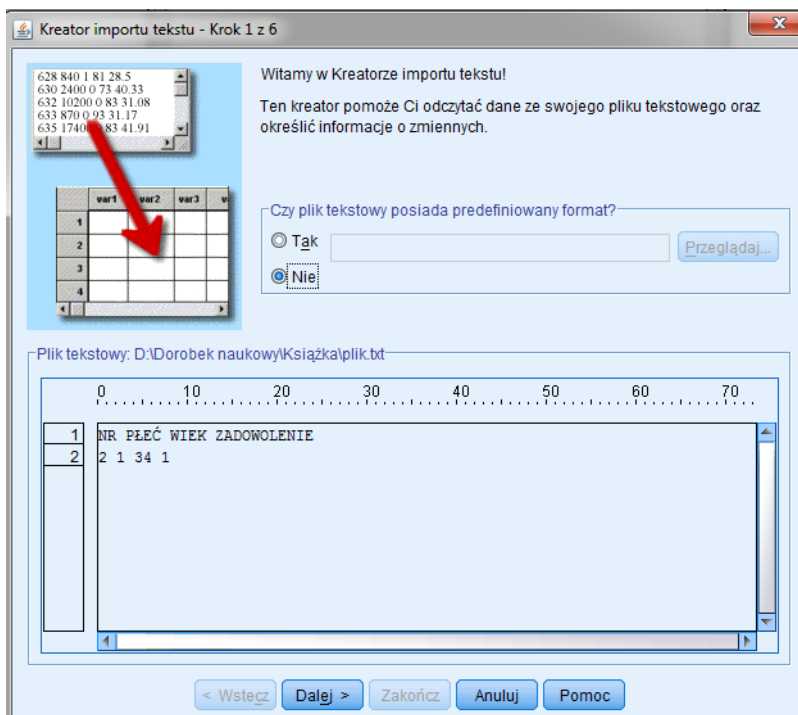
1.5. Wczytywanie i zapisywanie zbioru danych

W SPSS można zaimportować dane różnego formatu. Aby to uczynić klikamy kolejno *Plik-Otwórz – Dane*. Wybieramy plik, który chcemy zaimportować, a następnie z rozwijanej listy wybieramy typ pliku.

Pliki zapisane w programie SPSS mają rozszerzenie .sav, taki format najłatwiej otworzyć. Ale spróbujemy otworzyć plik tekstowy.

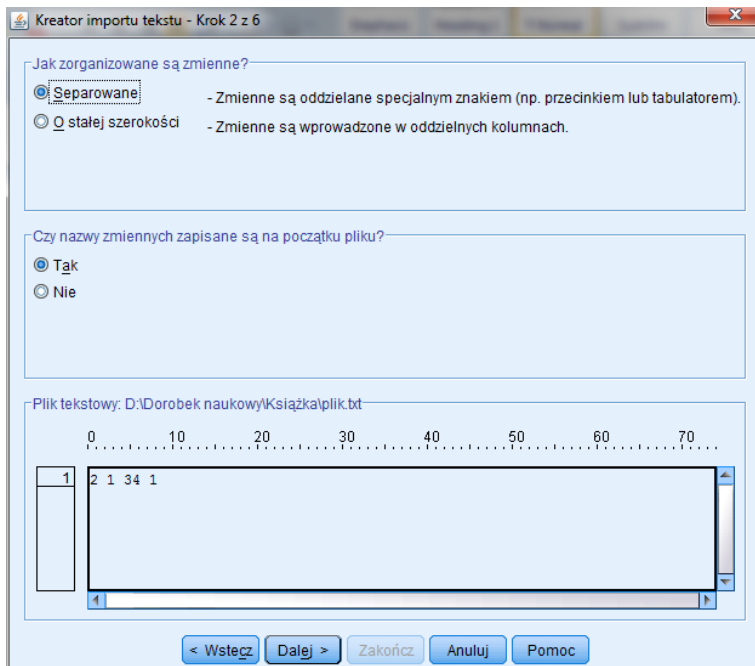


Pojawi się okno kreatora importu, które w łatwy sposób pozwoli nam przejść kolejne kroki w wczytywaniu danych:

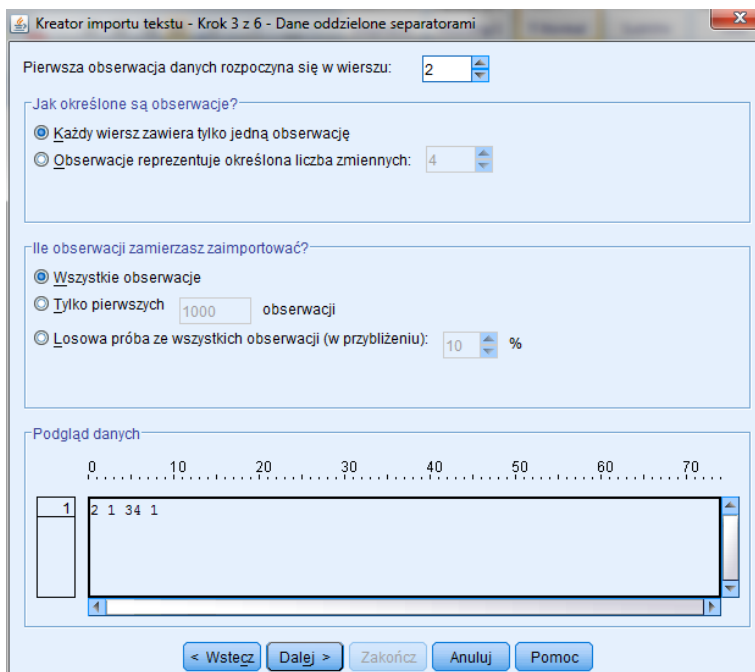


W pierwszym oknie zapyta nas, czy dane mają predefiniowany format oraz zaprezentuje fragment danych, które odczytał z pliku tekstowego. Nasze dane nie mają formatu predefiniowanego (nie mamy gotowego szablonu), dlatego klikamy *Dalej*. Następnie musimy określić, czy zmienne są separowane, czyli oddzielane jakimś znakiem albo spacją, czy też są wprowadzone w kolumnach o stałej szerokości. Nasze dane są oddzielone w pliku tekstowym

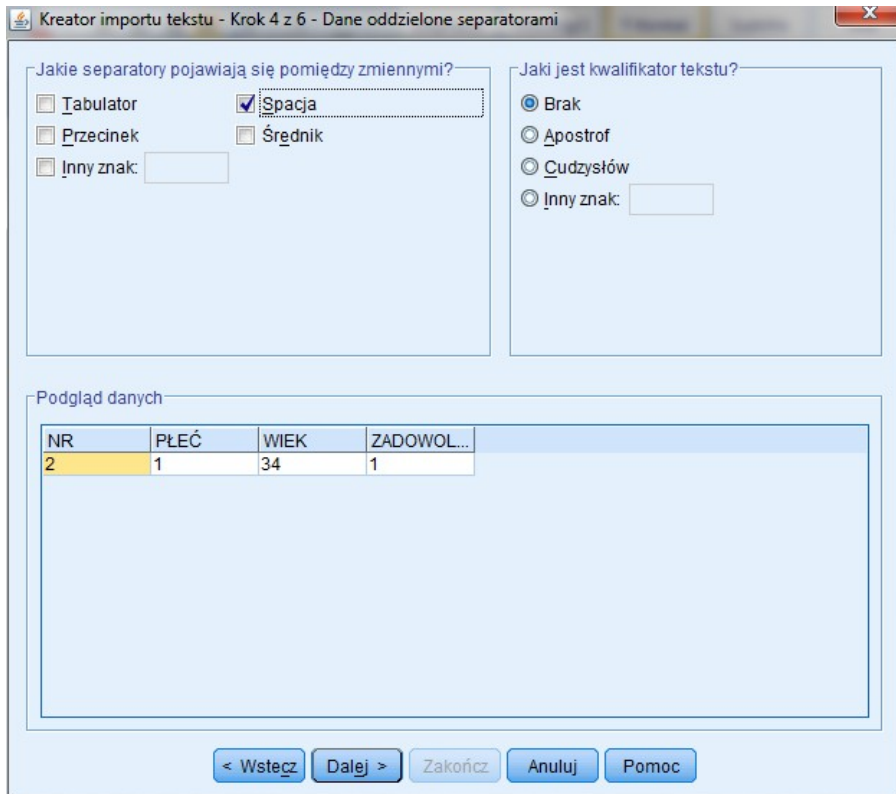
spacją. Program pyta nas czy w pliku zapisane są nazwy zmiennych na początku. W naszym przypadku tak właśnie jest.



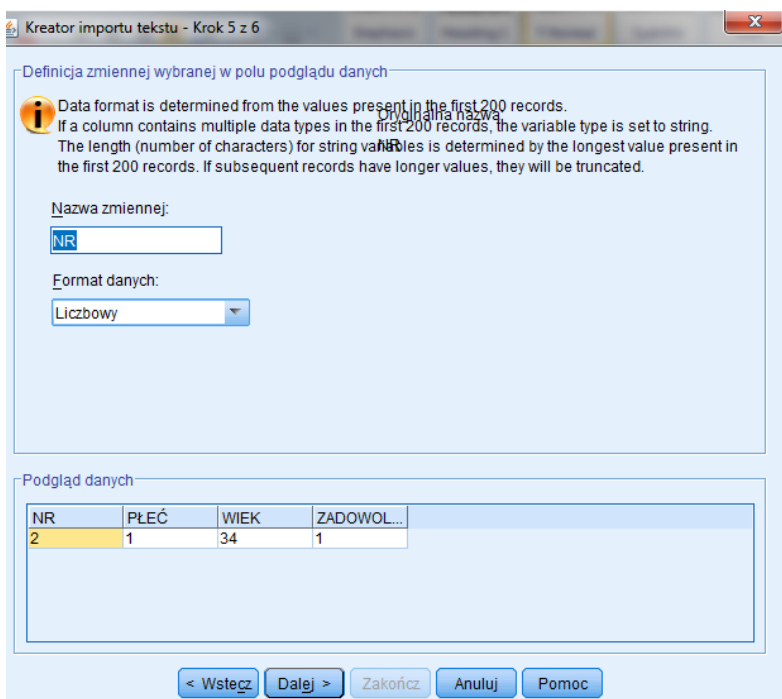
Teraz kreator importu pyta nas, jak określona jest każda obserwacja – czy wyniki jednej osoby mieszczą się w jednym wierszu, czy w wielu. U nas wiersz zawiera jedną obserwację. Chcemy też zaimportować wszystkie obserwacje, więc zostawiamy domyślnie ustawioną opcję i klikamy dalej.

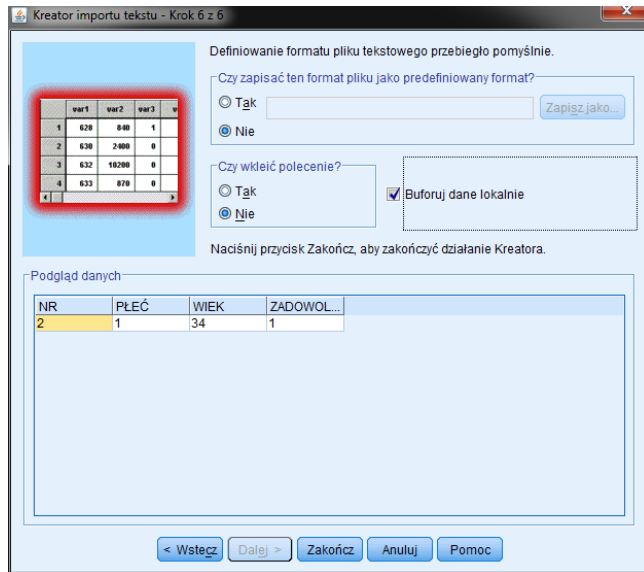


W czwartym kroku musimy określić, co stanowi separator poszczególnych zmiennych, u nas jest to spacja. Mamy podgląd naszych zmiennych.

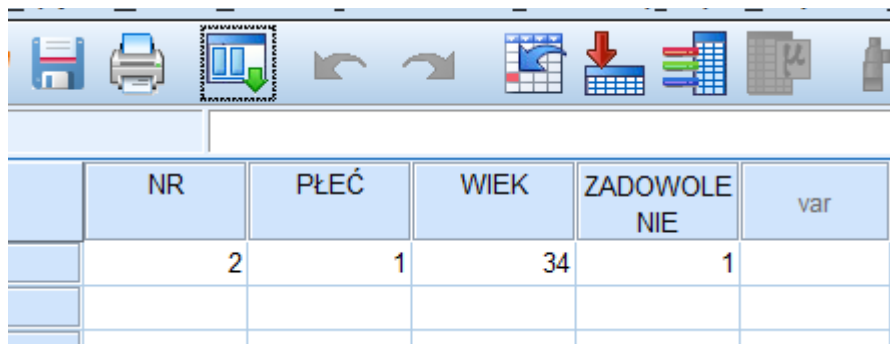


Kolejnym kroku program przeprowadza nas przez proces definicji formatu niektórych zmiennych – zwykle tych, które znajdują się na początku zbioru danych, w naszym przypadku jest to zmienna NR. W ostatnim kroku mamy możliwość zapisania tego sposobu importu jako predefiniowany format lub wklejenie całej procedury do pliku Języka Poleceń. Opcja ta jest przydatna, gdy będziemy chcieli importować większą liczbę plików z takimi danymi.

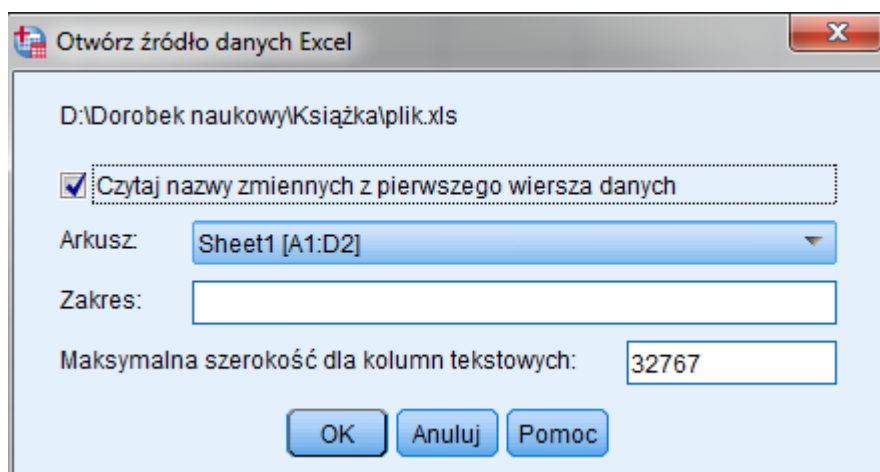




Klikamy *Zakończ* i gotowe!



Jeśli otwieramy plik z EXCELA procedura wygląda nieco inaczej.



Początek pracy jest taki samy, szukamy naszego pliku. Następnie otwiera się okno dialogowe, w którym musimy określić czy nazwy zmiennej są w pierwszym wierszu, w naszym pliku tak właśnie jest. Określamy też zakres danych w arkuszu. Klikamy OK i gotowe!

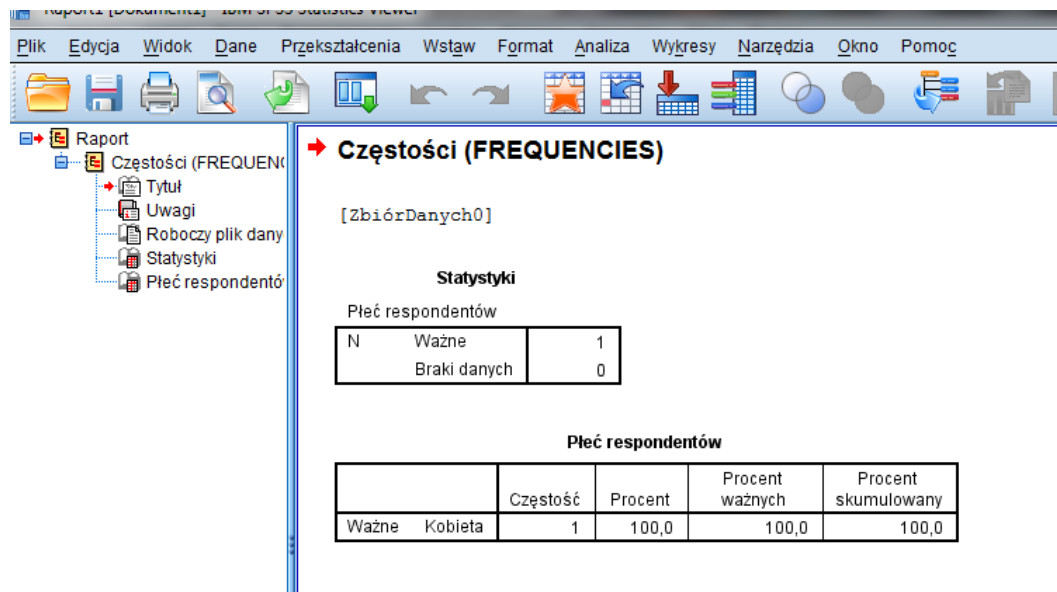
Zapiszmy oba zbiory jako ZBIÓR1 i ZBIÓR2.

Pracując z bazą danych warto ją co jakiś czas zapisywać. Pamiętajmy że operacja ta niesie za sobą konsekwencje, mianowicie: z jednej strony chroni nas przed utratą ostatnich zmian, ale z drugiej - Uniemożliwia powrót do poprzednich wersji pliku (plik stary zastępowany jest nowym) i przycisk Cofnij (strzałka w lewo) staje się nieaktywny.

1.6. Okno raportów

Edytor Raportów przechowuje wyniki analiz wykonanych przez SPSS. Okno Edytora Raportów składa się z dwóch elementów: spisu treści z lewej strony oraz okna z poszczególnymi tabelami i wykresami. Kliknięcie na pozycje spisu treści pozwala na szybkie odnalezienie interesującej nas analizy.

Dwukrotne kliknięcie na dowolny obszar tabeli pozwala nam dowolnie go modyfikować – wpisać zupełnie nowe treści lub edytować istniejące.



The screenshot shows the SPSS Report Editor interface. The left pane displays a tree view of the report structure, including 'Częstości (FREQUENCIES)', 'Tytuł', 'Uwagi', 'Roboczy plik danych', 'Statystyki', and 'Płeć respondentów'. The main pane displays the 'Częstości (FREQUENCIES)' report for the variable 'Płeć respondentów' from the dataset '[ZbiórDanych0]'. The report includes a 'Statystyki' section with a summary table and a detailed cross-tabulation table.

Statystyki

Płeć respondentów

N	Ważne	1
	Braki danych	0

Płeć respondentów

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	Kobieta	1	100,0	100,0	100,0

Statystyki

Płeć respondentów

N	Ważne	1
	Braki danych	0

		Częstość	Proc
Ważne	Kobieta	1	100

Co to jest?

- Wytnij Ctrl+X
- Kopiuj Ctrl+C
- Wklej Ctrl+V
- Usuń Delete
- Zaznacz
- Pokaż opis wymiaru
- Ukryj kategorię
- Grupuj
- Rozgrupuj
- Sortuj wiersze
- Utwórz wykres
- Właściwości tabeli...
- Właściwości komórki...
- Szablony Tabeli...

Opcja panel przedstawienia możliwość swobodnej edycji tabeli wydruku, na przykład zamiany wierszy i kolumn czy ukrycia pewnej części wydruku.

Wyniki z raportu mogą być eksportowane do innych programów np. WORD lub EXCEL.

ROZDZIAŁ II

PRZYGOTOWANIE ZBIORU DANYCH DO ANALIZY

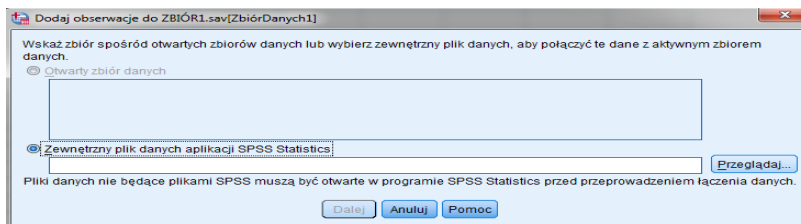
Zebranie i wprowadzenie danych do SPSS to dopiero początek pracy. Badacz musi następnie przygotować zbiór danych do analizy, niejednokrotnie etap ten okazuje się najbardziej żmudny i czasochłonny. Zbiór musi być „czysty”, pozbawiony błędów, na tym etapie pracy badacz tworzy też nowe zmienne z danych surowych.

2.1. Zarządzanie zbiorami danych

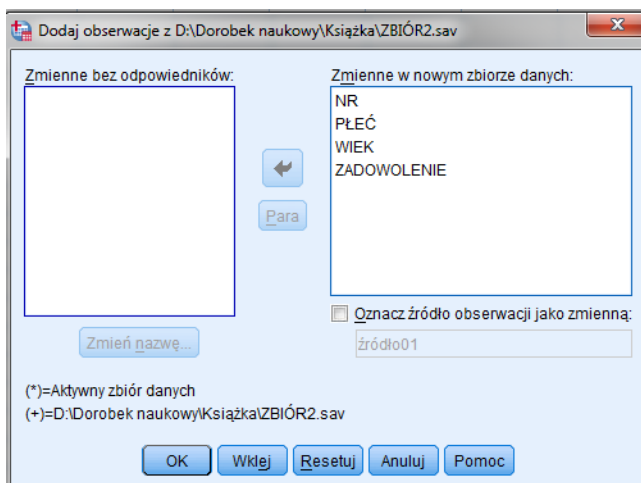
2.1.1. Łączenie zbiorów danych

SPSS pozwala nam połączyć nowe dane z wcześniej utworzonym zbiorem. W wcześniejszym etapie pracy otworzyliśmy dwa pliki z danymi, jeden z EXCELA, drugi z pliku tekstowego. Zapisaliśmy je jako ZBIÓR1 i ZBIÓR2. Teraz spróbujemy je połączyć.

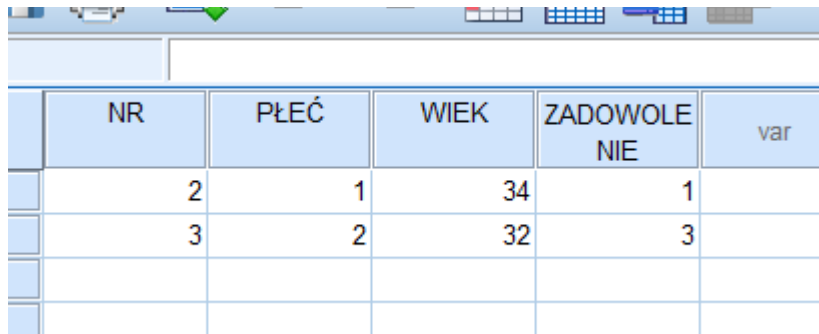
Otwieramy plik ZBIÓR1. Klikamy kolejno: *Dane – Połącz dane – Dodaj obserwacje*. Szukamy naszego pliku na dysku twardym komputera.



Widzimy okno z dwiema listami zmiennych: Zmienne bez odpowiedników tj. takie, które nie mają pary oraz zmienne w nowym zbiorze danych. Nasze zmienne pojawiły się w drugiej kategorii ponieważ takie same zmienne są w jednym i drugim zbiorze danych.



Po kliknięciu OK pojawi się nowy zbiór danych utworzony ze starych dwóch zbiorów.



	NR	PŁEĆ	WIEK	ZADOWOLENIE	var
	2	1	34	1	
	3	2	32	3	

Zapiszmy plik jako ZBIÓR3.

W przypadku, gdy zmienne różnią się jedynie nazwą i wiemy, że kodują to samo, możemy skorygować nazwę jednej ze zmiennych, by program mógł je połączyć w parę.

Pamiętajmy że przy łączeniu zbiorów danych, zmienne w tych zbiorach muszą być ustawione w tej samej kolejności.

2.1.2. Dodawanie zmiennych

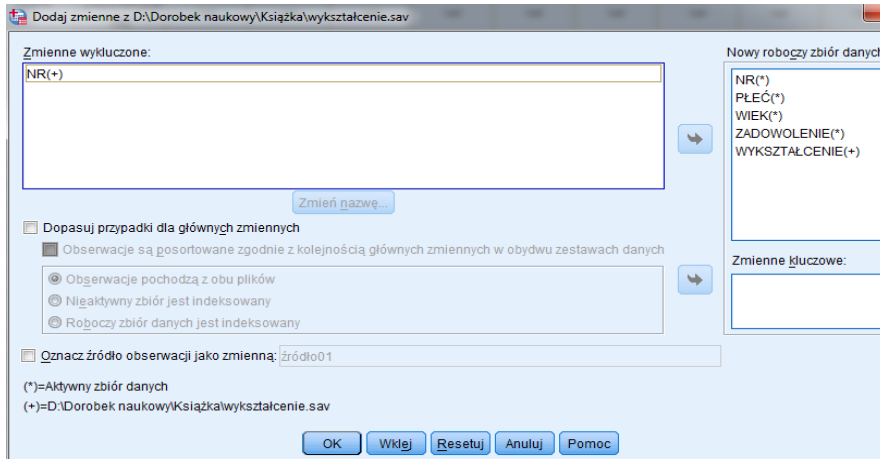
Jeśli chcemy dołączyć do zbioru dodatkowe zmienne, ważne jest, by oba pliki miały jakąś identyczną zmienną identyfikującą obserwacje, np. zmienna NR – numer respondenta. Będziemy pewni, że dane jednej osoby zostaną zapisane we właściwym wierszu.

Do naszego zbioru (ZBIÓR3) dołączmy zmienną WYKSZTAŁCENIE (plik wykształcenie).

- Jakie jest Pana/Pani wykształcenie?
 - a) Podstawowe
 - b) Zawodowe
 - c) Średnie**
 - d) Wyższe

Klikamy kolejno: *Dane – Połącz dane – Dodaj zmienne.*

Podobnie jak przy łączeniu zbiorów danych musimy określić plik z którego ma pochodzić nowa zmienna.



W oknie widzimy dwa zestawy zmiennych, pierwsze, po lewej, to te, które są zdublowane w obu plikach (zmiennie wykluczone) – tutaj jest to numer respondenta NR, a pozostałe zmienne zostały zamieszczone w okienku Nowy roboczy zbiór danych (lista zmiennych, które zostaną zamieszczone w nowym zbiorze danych). Zmienna wykształcenie ma plusik, tzn. ona zostanie włączona do tego zbioru. Możemy ustawić także zmienną kluczową, która ustala porządek w nowym zbiorze danych, to tzw. zmienna panująca.

Klikamy OK, a to efekt naszej pracy:

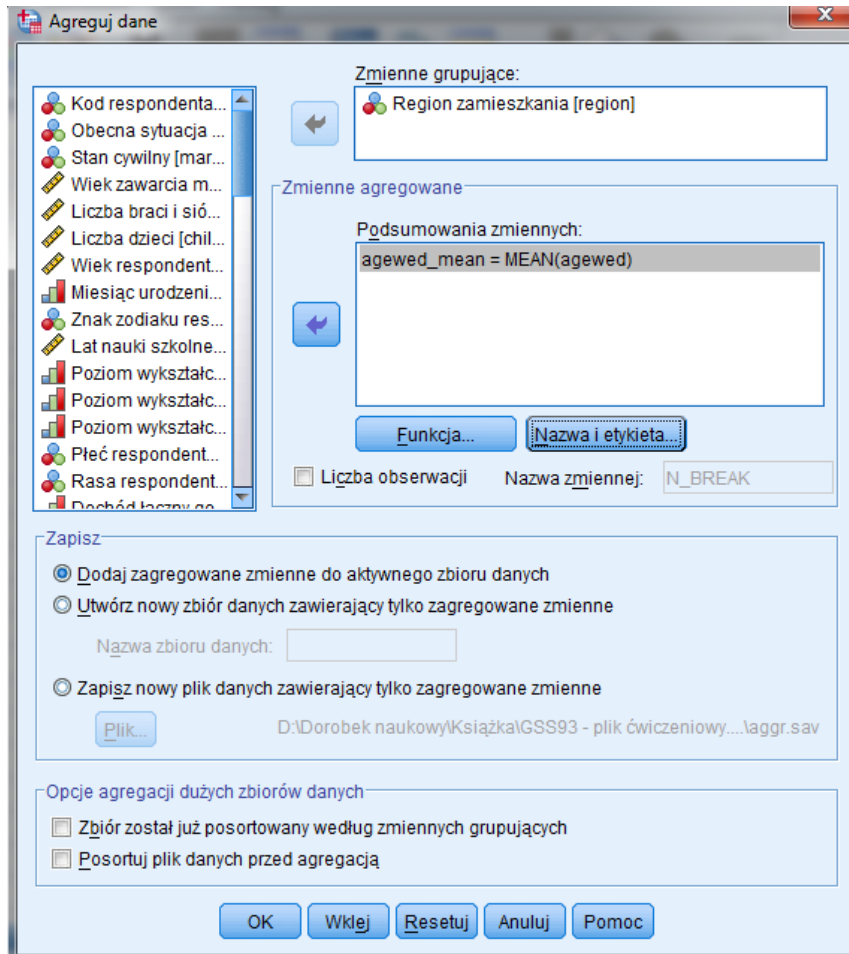
	NR	PŁEĆ	WIEK	ZADOWOLENIE NIE	WYKSZT AŁCENIE
	2	1	34	1	3,00
	3	2	32	3	4,00

2.1.3. Agregacja zbiorów danych

Agregowanie polega na wyliczeniu jednej lub wielu statystyk, dla grup obserwacji wyznaczonych przez zmienną grupującą. Pamiętajmy, że agregowanie zbioru danych jest takim przekształceniem, które umożliwia zmianę jednostek analizy w inne. Nowe jednostki analizy tworzone są na podstawie syntezy z jednostek źródłowych.

Spróbujmy wskazać (plik GSS93.sav) średni wiek zawarcia związku małżeńskiego, ale nie u pojedynczych osób, ale ze względu na region zamieszkania.

Klikamy *Dane – Agreguj*. W oknie dialogowym określamy zmienną grupującą region oraz podsumowania zmiennej wiek zawarcia pierwsze związku małżeńskiego. Określamy funkcję użytą do przekształceń, nas będzie interesować średnia arytmetyczna. Możemy także określić nazwę i etykietę nowej zmiennej. Zagregowane zmienne mogą być utworzone w nowym zbiorze danych lub dołączone zostaną do zbioru aktualnie używanego.

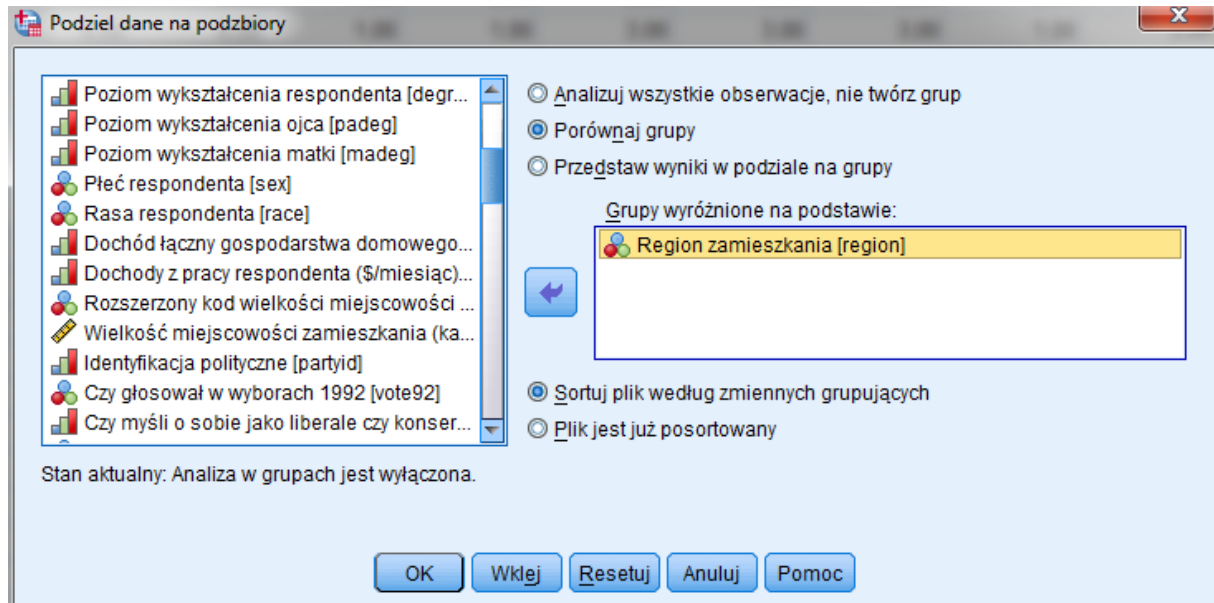


Po kliknięciu OK zostanie utworzona na końcu zbioru zmienna wynikowa. Teraz wiemy jaki jest średni wiek zawarcia małżeństwa respondenta w konkretnym regionie świata.

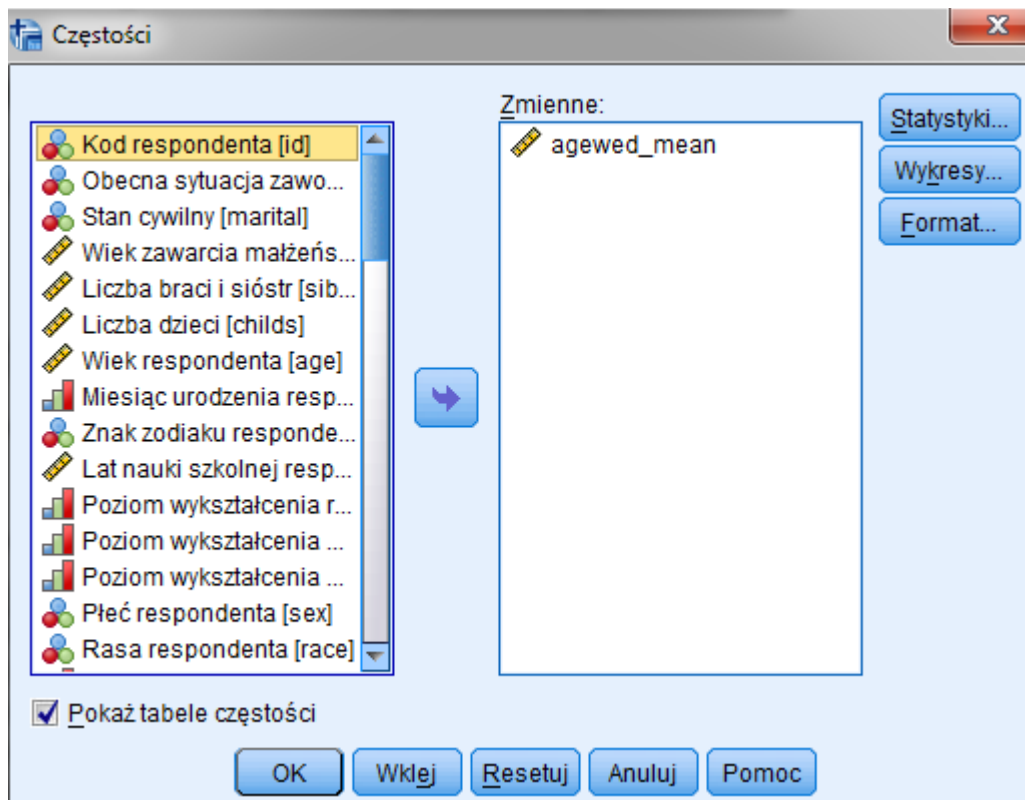
2.1.4. Analiza danych w podzbiorach

Opcja analiza danych w podzbiorach służy do rozdelenia każdej następnej analizy na podbiory ze względu na kategorie zmiennej. I tak wybranie zmiennej płeć spowoduje, że analizy będą dokonywane oddzielnie dla kobiet i dla mężczyzn. Gdy wprowadzimy więcej zmiennych dzielących na podgrupy, to będą one tworzone ze względu na kolejność wprowadzanych zmiennych.

Efekty naszego poprzedniego działania będą wyraźnie widoczne, gdy użyjemy opcji podziału danych na podbiory ze względu na region.



Teraz klikamy *Analiza – Opis statystyczny – Częstości*, a do analizowanych zmiennych przeliczamy zmienną utworzoną w procesie agregowania.



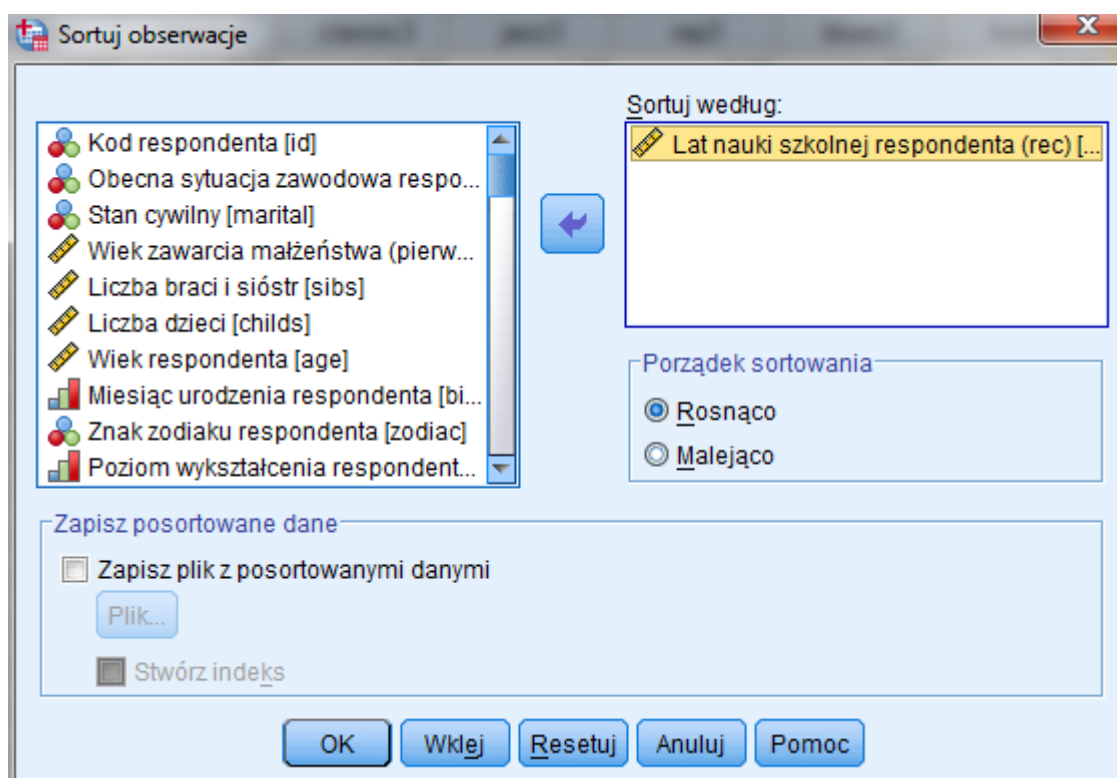
A to wynik naszego działania:

agewed_mean

Region zamieszkania			Częstość	Procent	Procent ważnych	Procent skumulowany
Nie przypisany	Ważne	22,66	743	100,0	100,0	100,0
Nowa Anglia	Ważne	26,22	31	100,0	100,0	100,0
Środkowy Atlantyk	Ważne	23,87	105	100,0	100,0	100,0
Płn-wsch. Centralny	Ważne	22,79	173	100,0	100,0	100,0
Płn-zach. Centralny	Ważne	21,72	48	100,0	100,0	100,0
Płd. Atlantyk	Ważne	22,11	123	100,0	100,0	100,0
Płd-wsch. Centralny	Ważne	22,38	56	100,0	100,0	100,0
Płd-zach. Centralny	Ważne	22,28	69	100,0	100,0	100,0
Góry	Ważne	23,85	35	100,0	100,0	100,0
Pacyfik	Ważne	23,22	117	100,0	100,0	100,0

2.1.5. Sortowanie obserwacji w zbiorze danych

Sortowanie ułatwia nam poruszanie się w zbiorze danych. Opcję tę znajdziemy w menu *Dane* – *Sortuj obserwacje*.

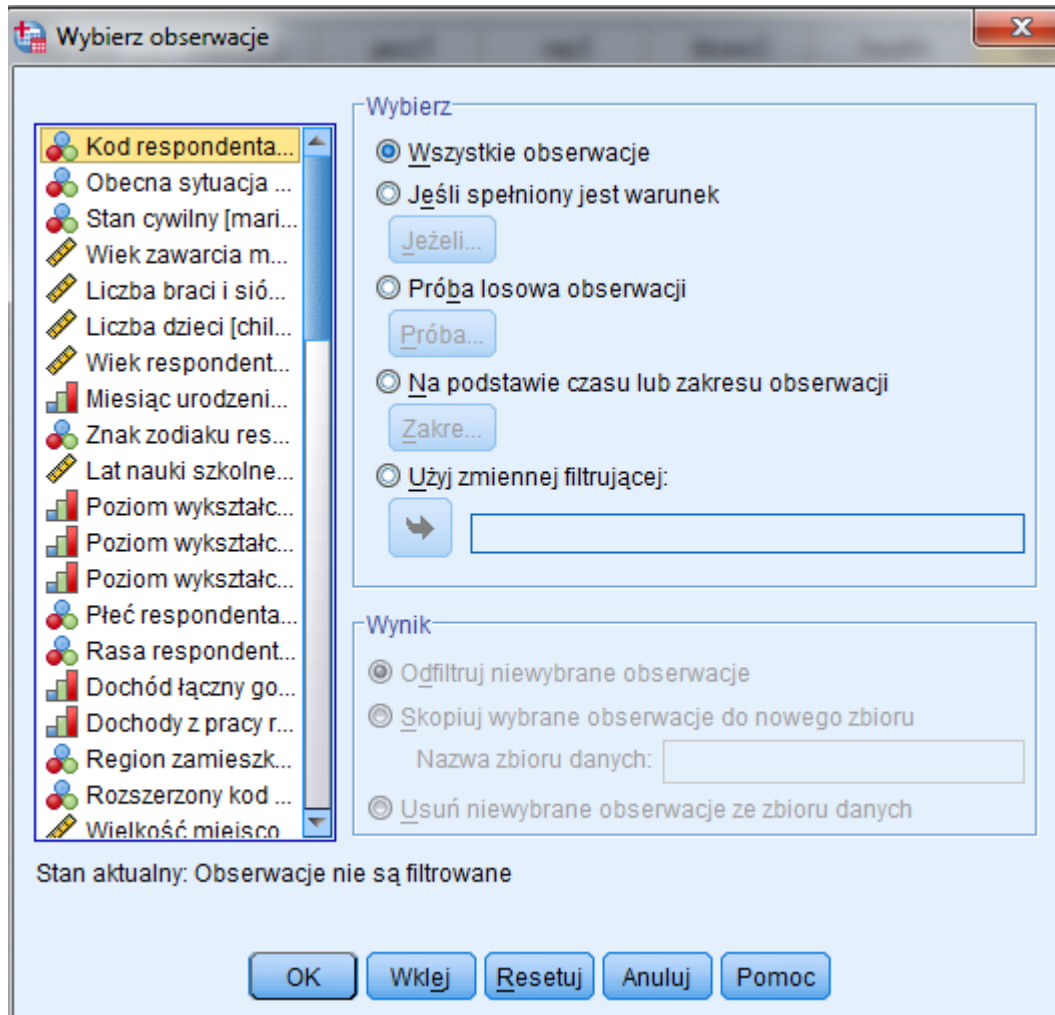


Sortowanie może być rosnące lub malejące według wskazanej przez nas zmiennej. Wprowadzenie wielu zmiennych spowoduje, że sortowanie będzie się odbywać według kolejności wprowadzanych zmiennych.

2.1.6. Wybór obserwacji i ich ważenie

Opcja wybór obserwacji umożliwia nam selekcję obserwacji do analizy. Możemy zdecydować, które obserwacje usunąć na stałe, chwilowo wyłączyć z analizy (odfiltrowanie). Wybór obserwacji automatycznie spowoduje utworzenie na końcu zbioru zmiennej filtrującej, z wartościami 1 i 0, gdzie 1 oznacza obserwację wyselekcjonowaną.

Opcję tę znajdziemy w menu *Dane – Wybierz obserwacje*.



Wybór obserwacji odbywa się na zasadzie:

- Warunku – np. wiek > 30
- Próby losowej obserwacji – określamy liczbę lub procent obserwacji włączony do analizy
- Na bazie czasu lub zakresu obserwacji np. od 100 do 200.
- Na bazie zmiennej filtrującej - umożliwia warunkowe selekcjonowanie jednostek analizy w zbiorze danych na podstawie wartości przyjmowanych przez wskazaną zmienną.



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

Ważenie obserwacji odbywa się poprzez wybór w menu *Dane – Ważenie obserwacji* i wskazanie zmiennej zawierającej wagi.

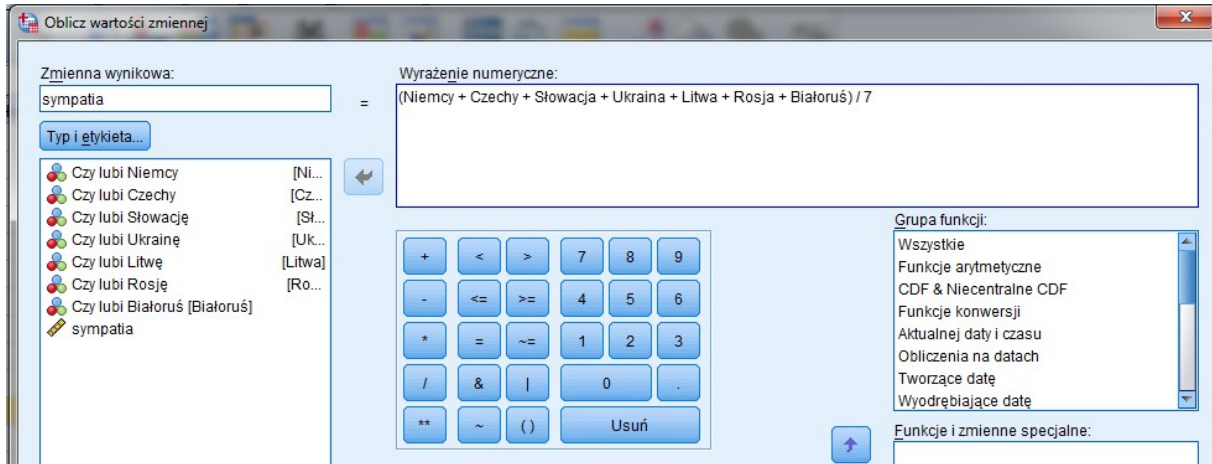
Ważenie obserwacji jest zabiegiem stosowanym w celu uzyskania takiej modyfikacji wyników uzyskanych w czasie wykonywania następných analiz, żeby wyniki były możliwie bliskie wynikom uzyskanym dla całej populacji, a nie wyłącznie dla wylosowanej próby. Np. wiemy że w populacji jest 100 000 kobiet i 150 000 mężczyzn. My zbadaliśmy 150 kobiet i 240 mężczyzn. Wagi w tej sytuacji wynoszą kolejno: $100\ 000/150=667$ i $150\ 000/240=625$.

2.2. Przekształcanie danych

2.2.1. Obliczanie wartości zmiennej

W praktyce badawczej można się spotkać z sytuacjami, kiedy to badacz musi obliczyć wartości zmiennej, wykorzystując związek tej zmiennej z danymi istniejącymi w zbiorze. Wówczas wykorzystujemy opcję *Przekształcenia – Oblicz wartości*.

Transformacja ta pozwala na tworzenie nowych zmiennych, np. w wyniku operacji arytmetycznych na istniejących zmiennych. Chociaż zakres funkcji dostępnych w SPSS jest bardzo szeroki. W celach ćwiczeniowych wykonamy prosty przykład (plik *Sympatia.sav*). Sprawdźmy jaki jest stopień nasilenia sympatii Polaków do sąsiadów. W tym celu otworzymy zmienną *sympatia*, na którą będzie się składać suma punktów (respondenci posługiwali się skalą od 1 do 10) podzielona przez liczbę państw. Zobaczmy:



Dodajemy do siebie zmienne (w oknie wyrażenia numeryczne) i dzielimy je przez 7. Używamy w tym celu kalkulatora z operatorami logicznymi.

	Niemcy	Czechy	Słowacja	Ukraina	Litwa	Rosja	Białoruś	sympatia
	3,00	5,00	8,00	5,00	3,00	3,00	9,00	5,14
	5,00	7,00	7,00	6,00	4,00	2,00	6,00	5,29
	4,00	6,00	8,00	6,00	4,00	3,00	8,00	5,57
	5,00	3,00	7,00	5,00	4,00	3,00	7,00	4,86
	3,00	4,00	7,00	6,00	4,00	3,00	6,00	4,71
	2,00	5,00	7,00	7,00	4,00	3,00	5,00	4,71
	6,00	7,00	7,00	7,00	3,00	2,00	6,00	5,43
	5,00	6,00	7,00	8,00	2,00	2,00	5,00	5,00
	7,00	4,00	5,00	4,00	2,00	3,00	5,00	4,29
	6,00	4,00	6,00	5,00	3,00	2,00	4,00	4,29

2.2.2. Rekodowanie wartości zmiennych

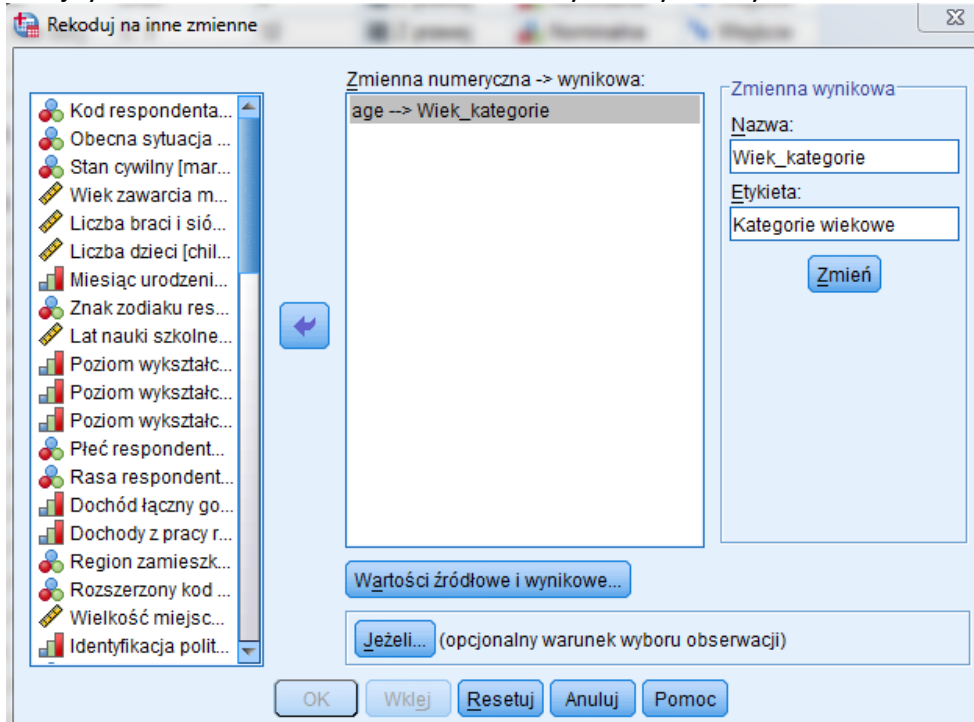
Dzięki rekodowaniu badacz może przekodować wartości zmiennych lub te wartości pogrupować (agregacja zmiennej). Umiejętność ta jest często wykorzystywana na etapie analizy (np. przy liczeniu chi-kwadrat niezależności). W procesie rekodowania powstaje nowa zmienna wynikowa, która tworzona jest za zmienną źródłową lub jako zmienna dodatkowa (opcje *Rekoduj na te same zmienne*, *Rekoduj na inne zmienne*).

Wykorzystując plik GSS93 spróbujemy przeprowadzić rekodowania. W tym celu wykorzystamy zmienną wiek. Zmienna ta jest ilościowa. My zrekodujemy ją na zmienną nominalną i wprowadzimy trzy kategorie respondentów: młodzi (do 35 lat), w średnim wieku (36-55 lat) i starsi (powyżej 55 lat). Zachowamy pierwotną zmienną.

Klikamy *Przekształcenia – Rekoduj na inne zmienne*.

W polu *Zmienna źródłowa->wynikowa* umieszczamy zmienną, którą będziemy rekodować. Następnie określamy nazwę i etykietę zmiennej wynikowej i klikamy *Zmień*.

Kolejny krok to określenie wartości źródłowych i wynikowych.

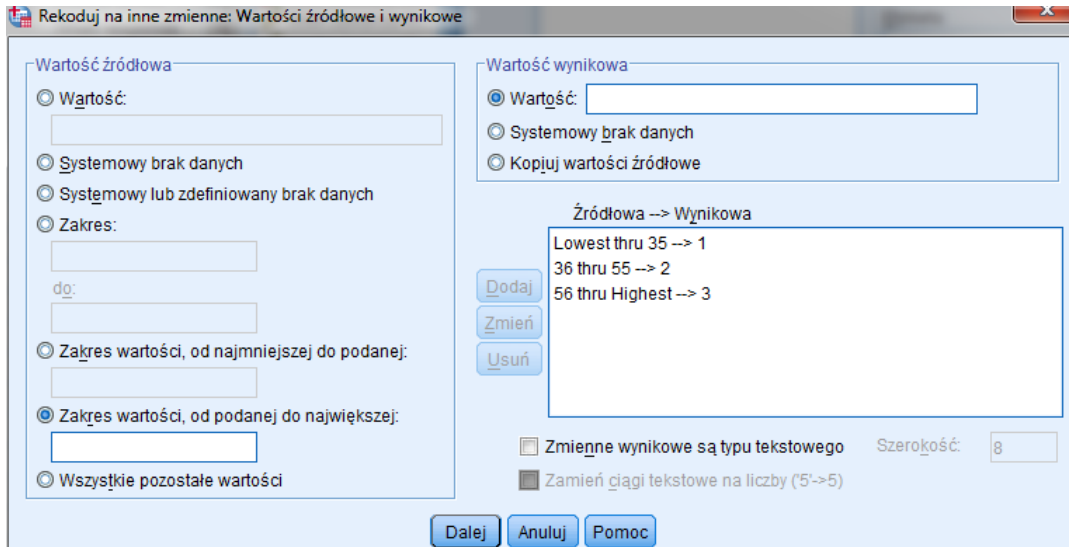


Rekodować można:

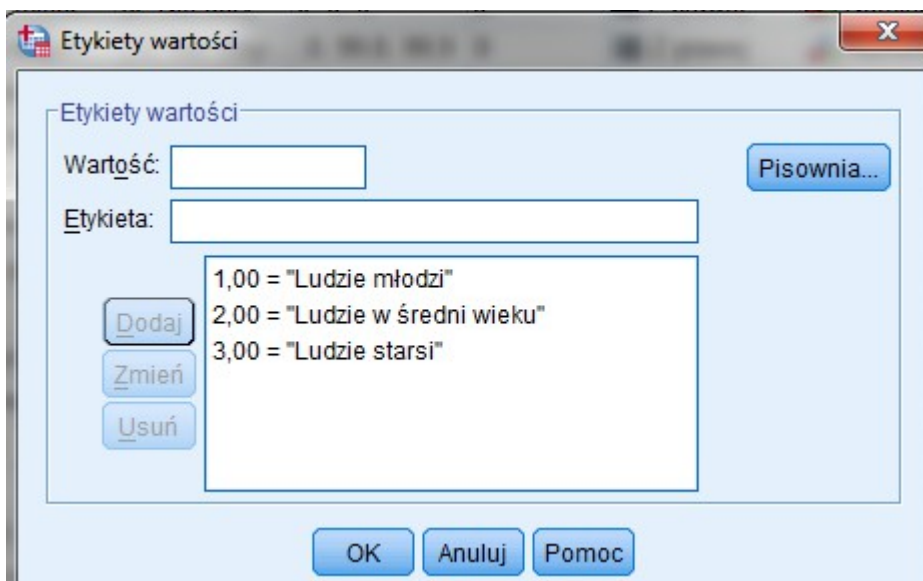
- Pojedynczą zmienną
- Zakres wartości od – do
- Przedział wartości jednostronnie otwarty

Zrekodować można także systemowe braki danych, systemowe lub zdefiniowane braki danych i wszystkie pozostałe wartości.

W naszym przykładzie mamy podane zakresy: od najmniejszego wieku do 35 lat, od 35 lat do 55 lat i od 55 lat do wartości największej. Zatem zakresy te przyjmą nowe wartości, kolejno 1, 2, 3.



Klikamy *Dalej* i *OK*. W ten sposób utworzyliśmy nową zmienną, a jej wartości nadajemy etykiety.

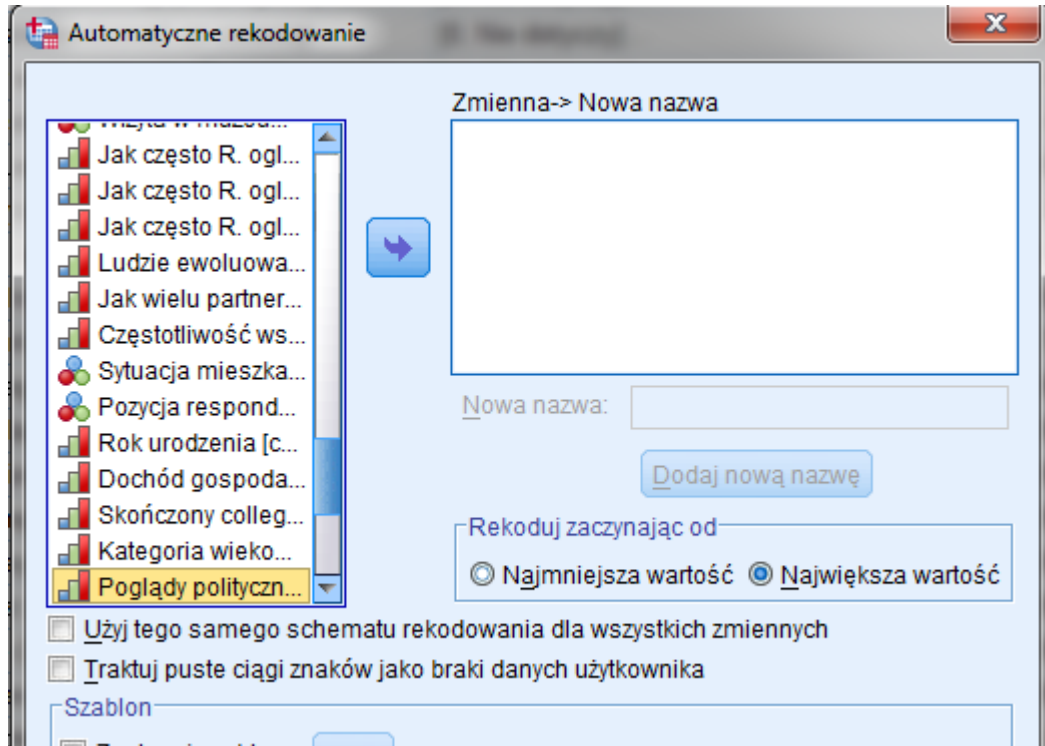


Kategorie wiekowe

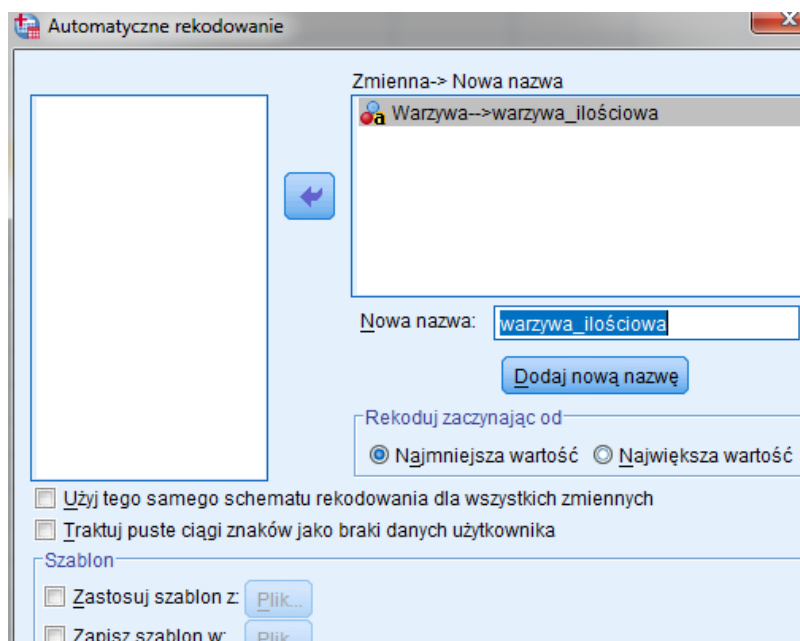
		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	Ludzie młodzi	473	31,5	31,5	31,5
	Ludzie w średnim wieku	592	39,5	39,5	71,0
	Ludzie starsi	435	29,0	29,0	100,0
	Ogółem	1500	100,0	100,0	

Rekodowanie automatyczne - pozwala na stworzenie kopii danej zmiennej z jednoczesnym

uporządkowaniem zakresu wartości zmiennej lub zmiennych. Tworzona jest nowa zmienna, a wartości ze starej zmiennej mogą zostać skopiowane bez przekształceń lub zostać odwrócone. Opcja ta przydatna jest w różnego rodzaju testach, gdzie niektóre pytania mogą mieć odwrócona skalę.



Najczęściej opcji tej używa się jednak do zamiany zmiennych jakościowych w zmienne ilościowe. Pamiętajmy, że w SPSS tylko zmienne ilościowe można poddać analizie. W naszym ćwiczeniu użyjemy zmiennej warzywa. Wprowadzamy nazwę nowej zmiennej i klikamy *Dodaj nową zmienną*.



Warzywa	warzywa_ilościowa
pomidor	4
fasola	2
burak	1
pomidor	4
fasola	2
burak	1
ogórek	3
ogórek	3
burak	1
fasola	2

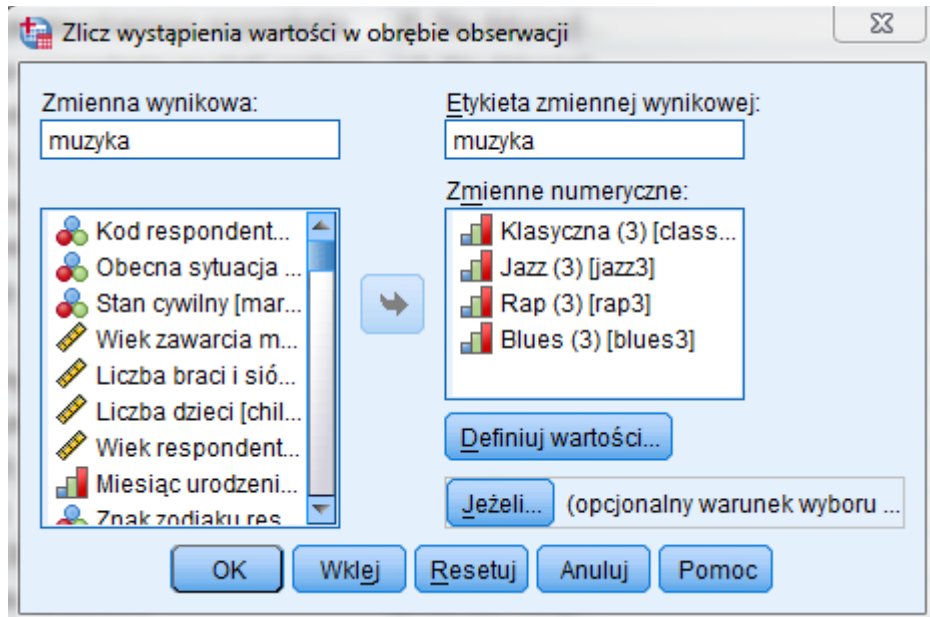
Zobaczymy, że wartość 1 została przypisana burakowi, ponieważ to pierwsze warzywo w spisie alfabetycznym naszych warzyw, ostatni jest pomidor, zatem jemu został przypisany kod 4. Kolejność mogliśmy odwrócić, zaznaczając w oknie dialogowym opcję *Największa wartość*.

2.2.3. Zliczanie wystąpień wartości

Jak sama nazwa wskazuje opcja ta służy do zliczania wystąpień określonych wartości w wybranym zestawie zmiennych. Tworzona jest nowa zmienna, która zawiera wartości mieszczące się w przedziale od 0 do wartości równej liczbie zadanych zmiennych.

Sprawdźmy (GSS93.sav) ile gatunków muzyki lubią respondenci spośród zestawu: klasyczna, jazz, blues, rap. Odpowiedź lubię została zakodowana jako 1, a więc ta wartość będzie zliczana. Skoro mamy cztery gatunki zakres zliczeń będzie wynosił 0-4, gdzie 0 oznacza, że respondent nie lubi żadnego z tych gatunków muzyki a 4, że lubi wszystkie.

W menu *Przekształcenia* klikamy *Zlicz wystąpienia*. Wprowadzamy nazwę i etykietę nowej zmiennej. Wprowadzamy również zestaw zmiennych.



W menu Definiuj wartości wpisujemy wartość (lub wartości), które mają być zliczane, podobnie jak w opcji rekodowania). U nas jest to 1. Po kliknięciu *dalej* i *OK* na końcu zbioru danych pojawi się nasza zmienna.

muzyka	var
1,00	
,00	
3,00	
3,00	
3,00	
3,00	
3,00	
3,00	
3,00	
3,00	
3,00	
2,00	
4,00	
1,00	
2,00	
2,00	
4,00	

2.2.4. Rangowanie wartości zmiennych

Rangowanie polega na przyporządkowaniu poszczególnym jednostkom analizy określonych wartości ze względu na pewną cechę lub cechy hierarchicznego miejsca w zbiorze danych. Dzięki temu zabiegowi możemy pozbyć się obserwacji odstających (niepasujących do reszty).

Efektom działania tej funkcji jest wygenerowanie zmiennej, której wartości porządkują jednostki analizy nadając im rangi.

Rangowanie znajdziemy w menu *Przekształcenia – Ranguj obserwacje*. W oknie dialogowym wybieramy zmienną, którą chcemy rangować. Rangowanie może być według innej zmiennej (rangowanie zmiennej wiek według płci, czyli osobne rangowanie dla kobiet i mężczyzn). Ranga 1 (początek rangowania) może być przypisana wartości najmniejszej lub wartości największej.

Najważniejsze typy rang:

- Ranga – przyporządkowanie liczb od 1 do n
- Ocena Savage'a – oparta na rozkładzie wykładniczym, różnice między nadawanymi rangami będą duże na jednym krańcu rozkładu i mniejsze na drugim
- Ranga ułamkowa i ranga ułamkowa w % – podzielenie rang regularnych przez liczbę obserwacji danej zmiennej, rangi ułamkowe, te klasyczne i te wyrażane w procentach pozwalają nam na lepszą porównywalność danych w zbiorze z brakami danych
- Suma wag obserwacji – ranga równa liczbie obserwacji w zbiorze
- Ntyle – obserwacje dzielone są na części, jeśli wpisujemy 5 do pierwsze 20% obserwacji będzie miało rangę 1, drugie 20% - 2, trzecie 20% - 3, czwarte 20% - 4 i ostatnie 20% - 5.

Oferowane są również formuły estymacji rozkładu. Dla ocen częstości oraz wyników normalnych można wybrać następujące formuły estymacji rozkładu⁴:

- transformacja Bloma dokonywana wedle wzoru: $x - 0,375 / y + 0,25$, gdzie x oznacza rangę danej jednostki analizy wyrażoną liczbą naturalną, a y liczbę jednostek analizy w zbiorze (tzw. sumę wag obserwacji).
- transformacja Tukey'a wykonywana z wykorzystaniem wzoru: $x - 0,3(3) / y + 0,3(3)$
- rangowanie (Rankit), które funkcjonuje wedle formuły: $x - 0,5 / y$
- transformacja Van der Wärdena obliczana na podstawie wzoru: $x / (y + 1)$

Wartości tożsame tworzą wiązania. Musimy określić jak mają być traktowane tzn. jakie rangi mają być im przypisane:

- Średnia (Mean) - średnia ranga z tożsamych wartości. Jeśli dwóm jednostkom analizy nadano rangę 1,5 to zajęły one 1 i 2 miejsce.
- Najniższa - obie analizowane jednostki analizy otrzymają rangę najniższą, patrząc na przykład wyżej byłoby to 1.
- Najwyższa - obie analizowane jednostki analizy otrzymają rangę najwyższą, a więc w przykładzie wyżej 2.
- Kolejne rangi dla niepowtarzalnych wartości przypisywanie rang - jednostka analizy,

⁴ D. Mider, A. Marcinkowska, *Analiza danych ilościowych dla politologów Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP*, Warszawa 2013, s. 135.

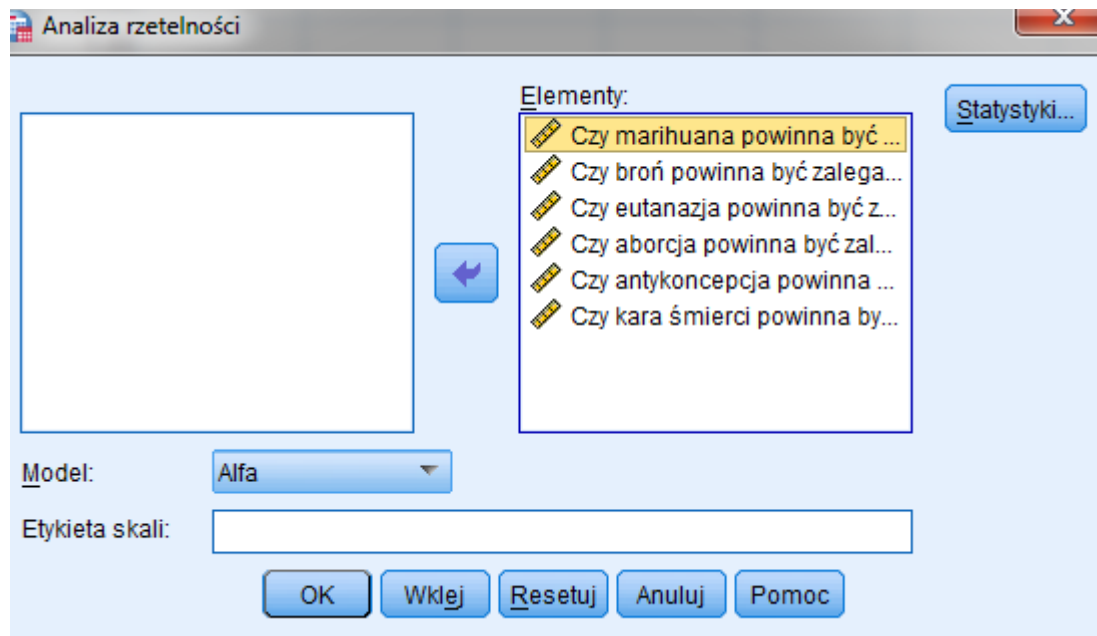
występująca bezpośrednio po jednostkach o rangach tożsamy, otrzyma rangę bezpośrednio następującą po nich

2.3. Analiza rzetelności skali metodą Alfa Cronbacha

W praktyce badawczej częstym zabiegiem jest konstruowanie wskaźników zjawisk, które składają się z wielu zmiennych. Mogą być to skale lub różnego rodzaju indeksy (np. skale psychologiczne). Zmienne je tworzące są bardziej lub mniej intensywnymi wskaźnikami badanego zjawiska. Innymi słowy niektóre zmienne są niezbędne w skali, inne powinny zostać usunięte ze względu na zaniżanie wartości poznawczej skali. Składające się na indeks lub skalę zmienne powinny mierzyć to samo, badać ten sam konstrukt teoretyczny, tzn. powinny być maksymalnie spójne. Ową spójność określa się jako rzetelność skali. Im skala bardziej rzetelna, w tym większym stopniu możemy być przekonani, że daje ona takie same rezultaty dla kolejno dokonywanych pomiarów. Warto zwrócić uwagę że przy narzędziach badawczych, standaryzowanych taka rzetelność jest określona, dzięki niej znamy „wartość” wykorzystywanego narzędzi.

Współczynnik alfa Cronbacha oparty jest na korelacji R Pearsona i w związku z tym poszczególne zmienne wchodzące w skład skali lub indeksu powinny być mierzone na poziomach ilościowych.

Jedną z metod mierzenia rzetelności skali jest Alfa Cronbacha, to najbardziej rozpowszechniony i najbardziej popularny współczynnik. W SPSS znajdziemy go klikając Analiza – Skalowanie – Analiza rzetelności. Metodą ustawioną automatycznie jest właśnie Alfa Cronbacha. Przerzucamy interesujące nas zmienne (rzetelność.sav):



W menu *Statystyki* zaznaczamy pozycję *Skala przy wykluczeniu pozycji*.

Po kliknięciu *dalej* i *OK* wyświetli się raport. Alfa Cronbacha przybiera wartość od 0 do 1. Im wynik bliższy jest jedności tym większa jest zgodność poszczególnych składników indeksu lub skali.

Statystyki rzetelności

Alfa Cronbacha	Liczba pozycji
,921	6

Nasz wynik jest bardzo wysoki. Wynik powyżej 0,7 należy uznać za zadowalający w naukach społecznych. Skala jest rzetelna.

Dzięki drugiej tabeli możemy dowiedzieć się, które ze zmiennych nie pasują do indeksu lub skali i należy je wyłączyć z dalszych analiz. Pokazuje ona, co się dzieje z Alfa Cronbacha po wykluczeniu danej pozycji. Jeśli jego wartość znacznie wzrasta tzn. że dana pozycja nie pasuje do skali.

Statystyki pozycji Ogółem

	Średnia skali po usunięciu pozycji	Wariancja skali po usunięciu pozycji	Korelacja pozycji Ogółem	Alfa Cronbacha po usunięciu pozycji
Czy marihuana powinna być zalegalizowana	13,0000	27,333	,846	,899
Czy broń powinna być zalegalizowana	13,0000	22,444	,932	,885
Czy eutanazja powinna być zalegalizowana	13,1000	28,989	,683	,919
Czy aborcja powinna być zalegalizowana	12,3000	29,122	,624	,926
Czy antykoncepcja powinna być zalegalizowana	12,7000	25,122	,849	,897
Czy kara śmierci powinna być zalegalizowana	13,4000	28,044	,755	,910

Z naszej skali nie powinniśmy usuwać żadnej pozycji.

ROZDZIAŁ III

ANALIZA CZĘSTOŚCI WYSTĘPOWANIA ZJAWISK

Przedmiotem tej części kursu są podstawy analizy danych w postaci tabelarycznej. Poniżej omówiono kolejno: sposoby tworzenia tabel w programie SPSS, zasady prezentacji danych tabelarycznych, umiejętność interpretacji wyników przedstawionych za pomocą tabel (co wydaje się - na pierwszy rzut oka – czynnością mało skomplikowaną).

3.1. Tworzenie i interpretacja tabel częstości

Tabele posiadają swoją strukturę, którą należy brać pod uwagę przy ich tworzeniu. Oto przykładowa tabela:

Tabela 1. Płeć respondentów (N=100).

	Płeć respondenta	Wskazania respondentów	
		N	%
	Kobieta	50	50
	Mężczyzna	50	50
	Razem	100	100

← Główka tabeli

← Komórki tabeli

→ Boczek tabeli

Źródło: badania własne

Tytuł tabeli – powinien być adekwatny do treści i zawartości, tabela ma swój numer porządkowy, który wynika z miejsca (kolejności występowania) tabeli w tekście. W tytule powinna być informacja o liczebności próby badawczej, zapisana w nawiasie.

Główka tabeli (patrz przykładowa tabela – komórki cieniowane na szaro) – zawiera informację na temat analizowanej zmiennej, jej nazwę (np. płeć respondenta), która może być poprzedzić kodem oznaczającym nazwę zmiennej w zbiorze danych (np. V01). Typowa tabela zawiera dwa rodzaje danych tj. liczebności (N) oraz wartości procentowe lub odsetki. Prezentowanie tylko wartości procentowych może prowadzić do błędu, polegającego na wnioskowaniu z niewielkich liczebności. Wartości frakcyjne pozwalają na standaryzację danych, umożliwiają porównywanie częstości występowania danych wartości zmiennej między sobą.

Boczek tabeli – zawiera etykiety wartości zmiennych. Ich porządek musi podlegać jakiejś ustalonej regule np. zależny jest od porządku występowania danych wartości w kwestionariuszu. Dane mogą być również uporządkowane rosnąco (lub malejąco) tzn. w górnych wierszach prezentowane są wartości zmiennej, które uzyskały najwyższą liczbą wskazań (lub, gdy są uporządkowane malejąco, najniższą liczbę wskazań). Niekiedy, zwłaszcza w przypadku danych socjodemograficznych, porządek ten wyznacza rodzaj zmiennej np. w przypadku zmiennej wiek, utworzone przedziały, decydują o kolejności (rosnącej lub malejącej) prezentacji danych. W górnych wierszach będą kategorie wiekowe najmłodszych

lub najstarszych. Innym sposobem uporządkowania etykiet może być przyjęcie kryterium alfabetycznego. Dotyczy to szczególnie zmiennych jakościowych np. nazwa państwa.

Komórki tabeli – zawierają wartości liczbowe. Częstym sposobem prezentacji danych jest umieszczanie po wartości liczbowej jednostki miary. To błąd, określenie tej jednostki jest już bowiem w główce tabeli np. %. Kolejną ważną rzeczą jest określenie, jaka liczba miejsc po przecinku będzie prezentowana w tabeli, najczęściej są to liczby dziesiętne, rzadziej setne, unika się zaś tysięcznych (chyba, że chodzi nam o bardzo dużą szczegółowość prezentowanych danych np. szybkość reakcji na bodziec). Warto pamiętać o kilku wskazówkach. Gdy badacz nie jest w stanie wypełnić komórki liczbą (logiczny brak danych) wstawia kropkę. W przypadku zmiennych nominalnych, może być sytuacja, że liczba wskazań jest mniejsza niż 5%, niektórzy badacze uważają, że takie dane są bezużyteczne i nie warto ich prezentować.

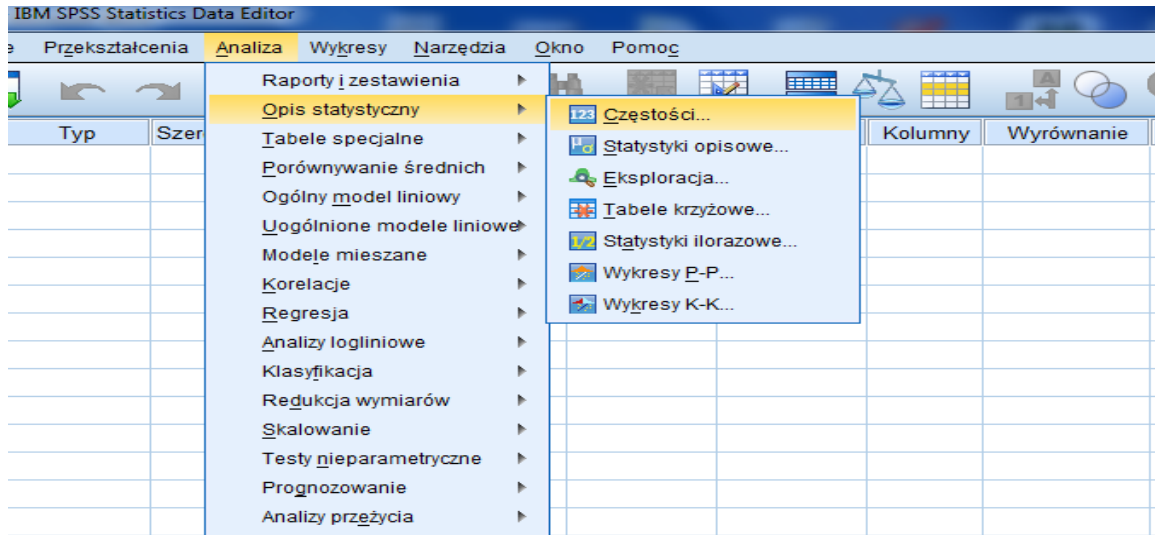
Zasada św. Mateusza – gdy po zaokrągleniu wartości nie sumują się do 100% (np. 99,9% lub 100,1%), odejmujemy 0,1 od wartości najniższej w tabeli lub dodajemy, 0,1 do najwyższej wartości zmiennej (tak by otrzymać 100%).

Źródło danych – badacz powinien wskazać pod każdą tabelą źródło danych, chyba że wszystkie prezentowane wyniki odnoszą się tylko do jednego źródła np. tego samego projektu badawczego (informujemy o tym we wstępie do raportu z badań). Powołując się na badania wtórne zazwyczaj stosujemy zapis taki jak w zwykłym przypisie, gdy dane są pierwotne (dane zebrane są w badaniach własnych) korzystamy z adnotacji „badania własne”. W praktyce badawczej bywa też tak, że dane wtórne poddaje się ponownemu opracowaniu lub powtórnej analizie. Wówczas podanie źródła można poprzedzić następującą informacją: „opracowanie własne na podstawie...”.

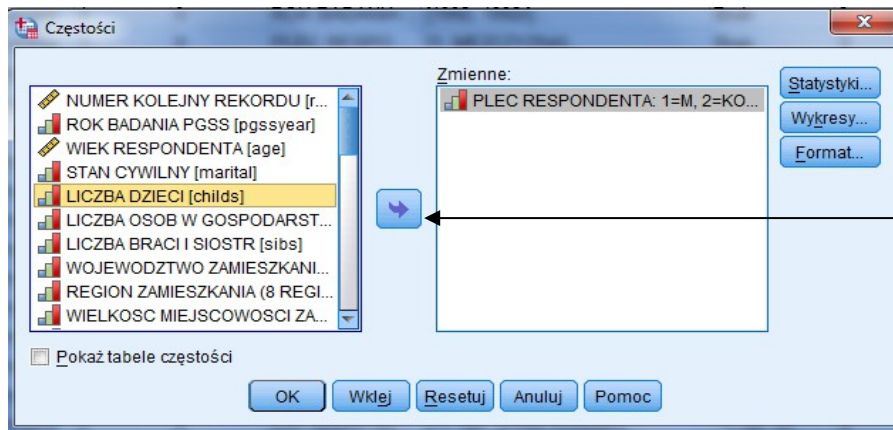
Czynność tworzenia tabel tj. tabulacja wymaga od badacza przygotowania zbioru danych do analizy. Tabulacja pozwala bowiem na zredukowanie nadmiernej ilości danych, pokazuje częstości występowania poszczególnych przypadków w zbiorze danych oraz pozwala na przedstawienie wyników w sposób czytelny i syntetyczny.

3.1.1. Tabele częstości dla jednej zmiennej

Tabela częstości to najprostszy sposób podsumowania zgromadzonych danych, poprzez ujęcie ich w kategorie i wskazanie liczby obserwacji tychże kategorii. Tworzenie tabel częstości dla jednej zmiennej odbywa się poprzez wybór następujących po sobie komend: *Analiza – Opis statystyczny – Częstości (GSS93.sav)*.



Główne okno dialogowe składa się z dwóch części. Po lewej stronie znajduje się lista zmiennych do wyboru. Z kolei po prawej stronie, w pustym okienku, powinny znaleźć się zmienne, które wybieramy do analizy. Z analizy wyłączamy braki danych, stąd też należy zwrócić uwagę czy proces kodowania (lub rekodowania) został prawidłowo przeprowadzony.



Zmienną do analizy wprowadzamy poprzez podwójne kliknięcie lub zaznaczenie zmiennej i kliknięcie strzałki, tu wybrano „płeć respondenta”

Efektom wskazanych czynności jest tabela w pliku raportu:

Statystyki

Płeć respondenta

N	Ważne	1500
	Braki danych	0

Płeć respondenta

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	Mężczyzna	641	42,7	42,7	42,7
	Kobieta	859	57,3	57,3	100,0
	Ogółem	1500	100,0	100,0	

Tabela *statystyki* informuje nas o liczbie braków danych, w tym przypadku jest ich 5, z kolei ważnych (użytych w analizie) obserwacji jest 1 500. Więcej informacji istotnych dla badacza wskazuje nam tabela druga. W główce tabeli umieszczona jest etykieta danej zmiennej. W boczku znajdują się etykiety wartości tej zmiennej. Kolumna *Częstość* zawiera liczebności wskazań respondentów na poszczególne odpowiedzi. Kolejna kolumna – *Procent* przedstawia wartości procentowe. Do 100% wliczane jest procent braków danych. Różnice pomiędzy tymi kolumnami widać wyraźnie w przypadku zmiennej z brakami danych:

Statystyki

Wydatki na zdrowie

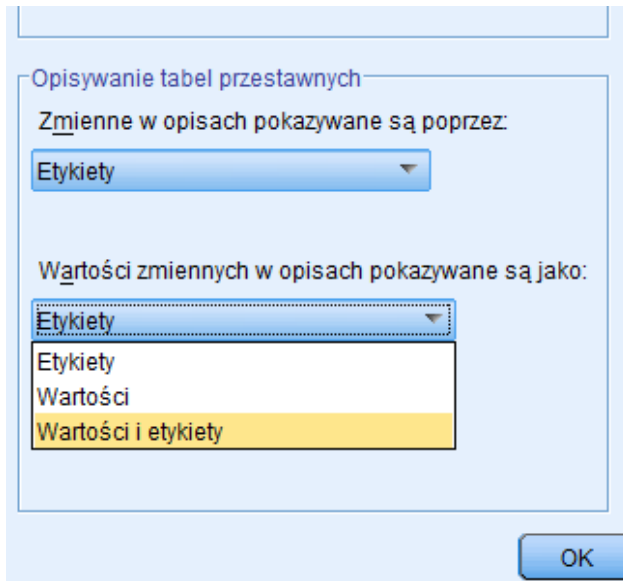
N	Ważne	1495
	Braki danych	5

Wydatki na zdrowie

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	Za mało	878	58,5	58,7	58,7
	W sam raz	384	25,6	25,7	84,4
	Za dużo	233	15,5	15,6	100,0
	Ogółem	1495	99,7	100,0	
Braki danych	Brak odpowiedzi	5	,3		
Ogółem		1500	100,0		

Braki danych, w przypadku analizowanej zmiennej „wydatki na ochronę zdrowia”, wynoszą 5, co stanowi 0,5% wszystkich odpowiedzi (kolumny *Częstość* i *Procent*). *Procent ważnych* nie uwzględnia braków danych. Z kolei *Procent ważnych*, który wskazuje kolejna kolumna wyłącza braki danych. *Procent skumulowany* sumuje wynik dla danej wartości oraz wyniki dla wartości poprzedzających tą wartość w tabeli. Przykładowo procent ten dla kategorii „w sam raz” wynosi 84,4%, składa się bowiem z dwóch zsumowanych wartości: dla „za mało” i dla „w sam raz”. Wynik tego działania to: $58,7\% + 25,7\% = 84,4\%$.

W programie SPSS można także określić sposób wyświetlania raportów. W tabeli powyżej mamy wskazane etykiety, ale możemy zlecić programowi, by prócz etykiet poszczególnych kategorii, widoczne były ich kody liczbowe. Wybieramy w menu kolejno: *Edycja – Opcje – Raport*. W wyświetlonym oknie interesuje nas pole *Opisywanie tabel przestawnych*:



W opcji *Wartości zmiennych w opisach...* zaznaczamy *Wartości i etykiety*

Nasz raport będzie wyglądał tak:

Statystyki

Wydatki na zdrowie

N	Ważne	1495
	Braki danych	5

Wydatki na zdrowie

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	1,00 Za mało	878	58,5	58,7	58,7
	2,00 W sam raz	384	25,6	25,7	84,4
	3,00 Za dużo	233	15,5	15,6	100,0
	Ogółem	1495	99,7	100,0	
Braki danych	9,00 Brak odpowiedzi	5	,3		
Ogółem		1500	100,0		

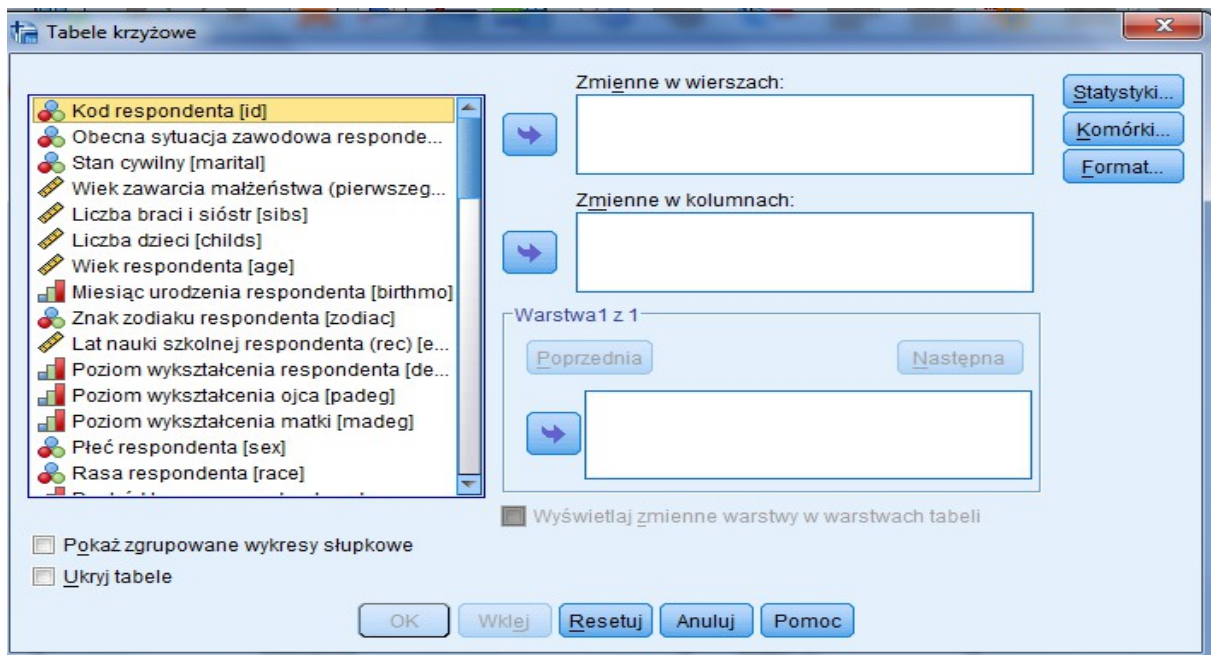
Wróćmy jeszcze na chwilę do głównego okna dialogowego (*Analiza – Opis statystyczny – Częstości*) badacz ma do dyspozycji trzy opcje: *Statystyki*, *Wykresy*, *Format*. *Statystyki* na razie pominiemy (wrócimy do nich przy okazji omawiania miar opisowych).

W menu *Format* badacz ma możliwość określenia uporządkowania danych. Może to zrobić według wartości, bądź według liczebności (według liczby wskazań), w obu przypadkach określając czy uporządkowanie ma być rosnące lub malejące. Gdy badacz wprowadzi wiele zmiennych do analizy, w menu *Format* może wybrać jak tabele będą prezentowane w raporcie, tzn. czy zmienne będą porównywane ze sobą, czy każda zmienna będzie przedstawiana oddzielnie.

3.1.2. Tabele częstości dla dwóch i więcej zmiennych

Najprostszy przykład tabeli złożonej stanowi tabela krzyżowa. Swoją nazwę zawdzięcza ułożeniu zmiennych: jedna znajduje się w wierszu, druga w kolumnie (zmienne krzyżują się). Czasami można spotykać się ze wskazówkami, które mówią, że w kolumnach powinna być zmienna niezależna. Tymczasem nie ma wytycznych, które określałyby tę kwestię, układ zmiennych to kwestia wyboru badacza. Pewne jest natomiast to, że tabela musi być czytelna, estetyczna i logiczna.

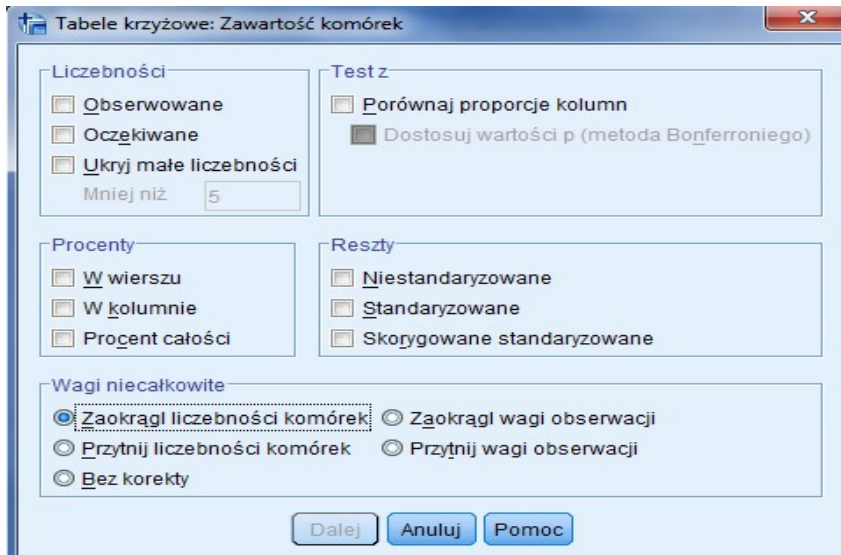
Tabele krzyżowe tworzymy wybierając kolejno: *Analiza – Opis statystyczny – Tabele krzyżowe*. W otworzonym oknie dialogowym, po lewej stronie, znajduje się lista zmiennych. Spośród nich wybieramy te, które będą w wierszach oraz te, które będą w kolumnach.



W oknie dialogowym mamy do wyboru trzy opcje: *Statystyki*, *Komórki* i *Format*. Podobnie jak poprzednio, na razie pominiemy statystyki (wrócimy do nich przy okazji badania zależności) i przejdziemy do opcji *Komórki*. Można tu wskazać szereg opcji konfiguracyjnych tabeli:

- W komórkach mogą być wyświetlane liczebności obserwowane (empiryczne, liczebności, które odpowiadają liczbie obserwacji) i oczekiwane (hipotetyczne liczebności, jakich oczekujemy w komórkach tabeli przy założeniu, że zmienne są niezależne). Ponadto można ukryć małe liczebności, ukryte wartości będą się wyświetlały jako <N, gdzie N jest określoną, przyjętą przez nas liczbą całkowitą.
- W komórkach mogą być wyświetlane procenty w wierszu, w kolumnie i procent całości. Oznaczenie którejs z opcji ma znaczenie i dla prezentacji wyników i dla ich interpretacji, o czym powiemy w dalszej części kursu.
- Reszta to różnica pomiędzy wartością obserwowaną a wartością oczekiwaną, Wartość oczekiwana jest liczbą obserwacji, której należałoby oczekiwać w komórce, gdyby nie istniał żaden związek pomiędzy dwiema zmiennymi. Reszta dodatnia wskazuje, że w komórce jest więcej obserwacji, niż powinno być, gdyby zmienne w kolumnie i w

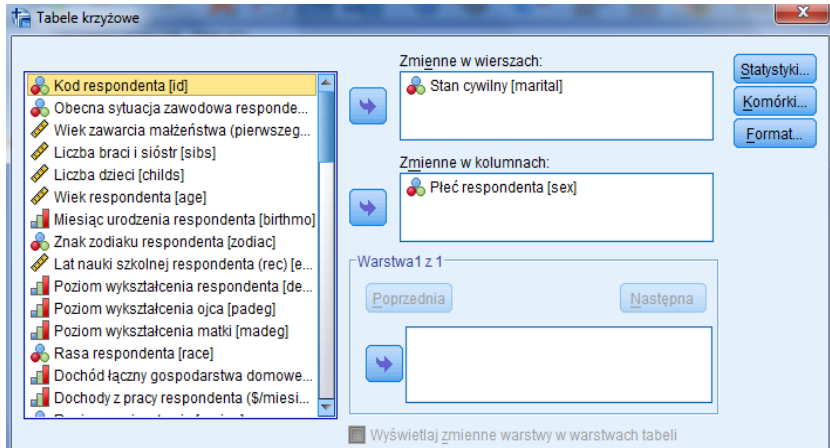
wierszu były niezależne. Reszty mogą być standaryzowane, jest to iloraz reszty i jej szacunkowego błędu standardowego. Reszta skorygowana standaryzowana z kolei to reszta dla komórki podzielona przez swój oszacowany błąd standardowy. Otrzymana reszta standaryzowana jest wyrażona w jednostkach odchylenia standardowego powyżej i poniżej średniej⁵.



Umieszczenie w polu *Kolumny* lub *Wiersze* kilku zmiennych powoduje, że w raporcie wyświetlonych zostanie kilka oddzielnych tabel, z kombinacją każdych dwóch zmiennych z tych list w roli kategorii wierszy lub kolumn. Aby utworzyć jedną tabelę z podziałem na więcej zmiennych w wymiarze wierszy, należy skorzystać z opcji *Warstwy*. Zmienna umieszczona w polu *Warstwy* zostanie uznana za nadrzędną w wymiarze wierszy, a jeśli będzie więcej warstw - zostanie nią ta z warstwy o najwyższym numerze (pamiętajmy jednak o ograniczeniach wymiarowych tabeli). Umieszczenie kilku różnych zmiennych w jednej warstwie spowoduje wyświetlenie kilku osobnych tabel.

Oto przykłady (GSS93.sav):

⁵ W menu komórki odjedziemy również opcję *wagi niecałkowite* - jeśli plik danych jest ważony przez zmienną ważącą zawierającą wartości ułamkowe, liczby komórek mogą być także wartościami ułamkowymi. Badacz ma kilka opcji w takiej sytuacji: *Zaokrąglaj liczby komórek*. Wagi obserwacji są używane bez zmian, ale skumulowane wagi w komórkach są zaokrąglane przed obliczaniem jakichkolwiek statystyk; *Obetnij liczby komórek*. Wagi obserwacji są używane bez zmian, ale skumulowane wagi w komórkach są przycinane przed obliczaniem jakichkolwiek statystyk; *Zaokrąglaj wagi obserwacji*. Wagi obserwacji są zaokrąglane przed użyciem; *Obetnij wagi obserwacji*. Wagi obserwacji są przycinane przed użyciem; *Bez korekty*. Używane są wagi obserwacji bez zmian oraz ułamkowe liczby komórek. Gdy jednak wymagane są dokładne statystyki (dostępne tylko z opcją Testy dokładne), skumulowane wagi w komórkach są obcinane lub zaokrąglane przed wyliczeniem dokładnych statystyk testowych. Zobacz więcej: [IBM SPSS Statistics Base 22, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/pl/client/Manuals/IBM_SPSS_Statistics_Base.pdf](http://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/pl/client/Manuals/IBM_SPSS_Statistics_Base.pdf). Copyright IBM Corp.

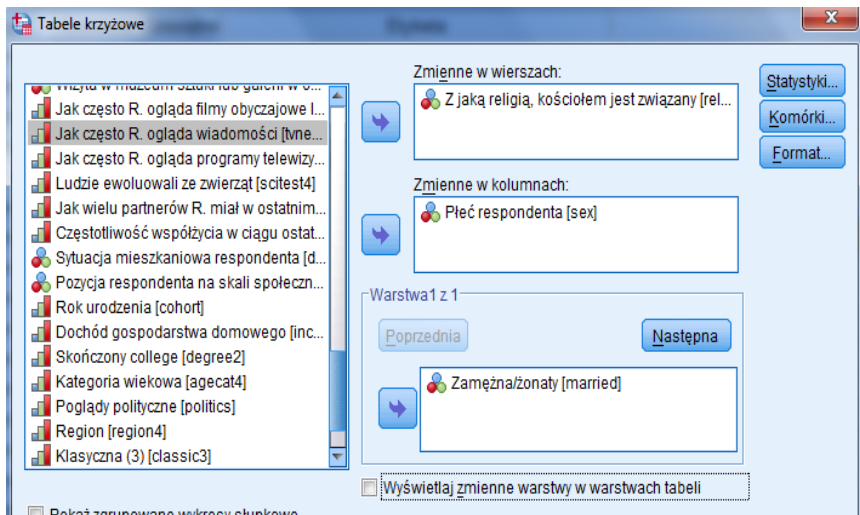


„Prosta” tabela krzyżowa:
kolumna – płeć, wiersz -
stan cywilny

Tabela krzyżowa Stan cywilny * Płeć respondenta

Liczebność

		Płeć respondenta		Ogółem
		Mężczyzna	Kobieta	
Stan cywilny	Żonaty/zamężna/KONK	383	412	795
	Wdowiec/wdowa	31	134	165
	Rozwiedziona/y	75	138	213
	Separacja	9	31	40
	Kawaler/panna	142	144	286
Ogółem		640	859	1499



„Złożona” tabela krzyżowa:
kolumna – płeć, wiersz –
identyfikacja religijna, warstwa –
zamężna/żonaty

**Tabela krzyżowa Z jaką religią, kościołem jest związany ^ Płeć respondenta ^
Zamężny/żonata**

Liczebność

Zamężny/żonata			Płeć respondenta		Ogółem
			Mężczyzna	Kobieta	
Tak	Z jaką religią, kościołem jest związany	Katolik	243	255	498
		Protestant	71	122	193
		Żyd	13	8	21
		Żadnym	40	17	57
		Innym	12	9	21
	Ogółem		379	411	790
Nie	Z jaką religią, kościołem jest związany	Katolik	145	309	454
		Protestant	59	81	140
		Żyd	6	4	10
		Żadnym	43	40	83
		Innym	3	11	14
	Ogółem		256	445	701

Opcja *Format* pozwala badaczowi uporządkować wiersze w kolejności rosnącej lub malejącej względem wartości zmiennej w wierszu.

3.2. Rodzaje procentowania

W programie SPSS mamy różne możliwości prezentacji wyników w tabelach krzyżowych za pomocą wartości procentowych. Program pozwala na użycie procentów w kolumnie, w wierszu i procenta całości. Wybór którejś z opcji, w sposób zasadniczy, wpływa na interpretację uzyskanych wyników. W pierwszym przypadku (procent w kolumnie) sumowanie do 100% następuje w ostatnim wierszu kolumny. Podstawą procentowania jest zatem zmienna w kolumnie. W naszym przykładzie jest to płeć. I tak oto z tabeli poniżej, dowiadujemy się m.in. że 13,1% mężczyzn oraz 6,7% kobiet nie jest związanych z żadną religią.

**Tabela krzyżowa Z jaką religią, kościołem jest związany ^ Płeć
respondenta**

% z Płeć respondenta

		Płeć respondenta		Ogółem
		Mężczyzna	Kobieta	
Z jaką religią, kościołem jest związany	Katolik	61,2%	65,9%	63,9%
	Protestant	20,4%	23,7%	22,3%
	Żyd	3,0%	1,4%	2,1%
	Żadnym	13,1%	6,7%	9,4%
	Innym	2,4%	2,3%	2,3%
Ogółem		100,0%	100,0%	100,0%

Oczywiście badacz ma możliwość zawarcia w tabeli zarówno procentów jak i liczebności poszczególnych kategorii. Nieco inaczej sytuacja wygląda w przypadku procentowania w

wierszu. Podstawą procentowania (sumowanie do 100%) jest zmienna, która znajduje się w wierszu. Interpretacja tabeli wygląda więc inaczej. Co wie badacz na podstawie poniższej tabeli? Choćby to, że spośród protestantów 39% to mężczyźni a 61% to kobiety. Sumowania do 100% następuje w ostatniej kolumnie.

Tabela krzyżowa Z jaką religią, kościołem jest związany * Płeć respondenta

% z Z jaką religią, kościołem jest związany

		Płeć respondenta		Ogółem
		Mężczyzna	Kobieta	
Z jaką religią, kościołem jest związany	Katolik	40,8%	59,2%	100,0%
	Protestant	39,0%	61,0%	100,0%
	Żyd	61,3%	38,7%	100,0%
	Żadnym	59,3%	40,7%	100,0%
	Innym	42,9%	57,1%	100,0%
Ogółem		42,6%	57,4%	100,0%

Z kolei po zaznaczeniu opcji *Procent całości* (tak jak prezentuje to tabela poniżej) z podstawę procentowania stanowią wszyscy respondenci. Uzyskujemy informację o odsetku danej kategorii płci i identyfikacji religijnej. Wiemy, że kobiety – katoliczki stanowią 37,8% populacji.

Tabela krzyżowa Z jaką religią, kościołem jest związany * Płeć respondenta

% z Ogółem

		Płeć respondenta		Ogółem
		Mężczyzna	Kobieta	
Z jaką religią, kościołem jest związany	Katolik	26,1%	37,8%	63,9%
	Protestant	8,7%	13,6%	22,3%
	Żyd	1,3%	0,8%	2,1%
	Żadnym	5,6%	3,8%	9,4%
	Innym	1,0%	1,3%	2,3%
Ogółem		42,6%	57,4%	100,0%

Warto zwrócić uwagę, że nad tabelą krzyżową - w raporcie SPSS- wyświetla się również informacja o analizowanych danych: ile obserwacji zostało wziętych do analizy, ile obserwacji jest wykluczonych (braki danych) a także informacja o krzyżowanych zmiennych. Obliczenia dla tabel krzyżowych prowadzone są po wyłączeniu braków danych. Tabela krzyżowa to dobre narzędzie opisu statystycznego i uporządkowania danych, ale samo jej sporządzenie nie dostarcza nam formalnych informacji o zależności zmiennych, o tym powiemy w dalszej części kursu.

3.3. Interpretacja i opis danych tabelarycznych

Zaprezentowane powyżej tabele krzyżowe nie mogą być automatycznie przeniesione do raportu z badań. Nie wszystkie informacje, które są generowane w raportach SPSS są na tyle ważne, aby koniecznie zamieszczać ja w oficjalnych (np. raporty, ekspertyzy) dokumentach.

Służą one jedynie do oceny wyników lub wstępnej analizy wyników. Ważną czynnością jest transformacja tabel, tzn. dostosowanie ich do wymogów sporządzania końcowych raportów z badań ilościowych (pisanych na podstawie raportów SPSS).

Oto przykładowy raport zawierający zmienną wiek (kategorie wiekowe):

Statystyki

Kategoria wiekowa

N	Ważne	1500
	Braki danych	0

Kategoria wiekowa

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	18-29	279	18,6	18,6	18,6
	30-39	352	23,5	23,5	42,1
	40-49	307	20,5	20,5	62,5
	50+	562	37,5	37,5	100,0
	Ogółem	1500	100,0	100,0	

W raporcie z badań pomijamy *Procent* i *Procent skumulowany*. Zazwyczaj „ostateczna” tabela nie zawiera również braków danych (choć to kwestia dyskusyjna, można się spotkać z opiniami, że jest to konieczne, ale równie dobrze można tę informację zamieścić pod tabelą, w tekście lub w przypisie). Przy tworzeniu tabeli stosujemy się do zasad wskazanych na początku rozdziału.

Tabela. Kategorie wiekowe respondentów.

Kategorie wiekowe	Wskazania respondentów	
	N	%
18-29	279	18,6
30-39	352	23,5
40-49	307	20,5
50+	562	37,5
Razem	1500	100

Źródło: badania własne.

Wartości liczbowe winny być zaokrąglone do jednego lub dwóch miejsc po przecinku. Warto pamiętać, że procedurą tą można przeprowadzić edytując raport SPSS poprzez podwójne kliknięcie na tabelę. Następnie zaznaczamy interesujący nasz zakres wartości i klikamy prawy przycisk myszy. Z listy komend (i opcji) wybieramy *Właściwości komórki*. W wyświetlonym oknie dialogowym szukamy *Wartość formatu*, gdzie mamy możliwość określenia miejsc dziesiętnych.

		Kategoria wiekowa	
		Częstość	Procent
Ważne	18-29	279	18,6
	30-39	352	23,5
	40-49	307	20,5
	50+	562	37,5
	Ogółem	1500	100,0

Kilka zasad opisu i interpretacji danych tabelarycznych:

- Tabele opisujemy w czasie teraźniejszym np. tabela przedstawia, opisuje itd.
- Opis tabeli może być skrócony - dokonujemy syntetycznej interpretacji tabeli (opisy takie charakterystyczne są dla sprawozdań z badań lub komunikatów z badań) lub rozszerzony – porównujemy uzyskane wyniki z innymi tego typu badaniami, danymi wtórnymi, komentujemy podobieństwa, różnice, opis taki ma charakter szczegółowy, uwzględnia postawione hipotezy.
- Opis powinien mieć swoją narrację np. co trzeci respondent, większość badanych, niewiele ponad połowa itp.
- W raporcie posługujemy się słowami nie znakami np. %, dopuszczalne jest stosowanie skrótów np. proc.
- Należy odróżnić pojęcia *procent* od *punkt procentowy*; terminu procent używa się na określenie operacji wykonywanych na wartościach liczonych do określonej wartości bazowej. Ma charakter bezwzględny; pojęcia punkt procentowy używa się w odniesieniu do operacji na wartościach wyrażonych w procentach; ma charakter względny⁶.

⁶ D. Mider, A. Marcinkowska, *Analiza danych ilościowych dla politologów Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP*, Warszawa 2013, s. 166.

Warto pamiętać, że w programie SPSS badacz ma do dyspozycji panel, w którym samodzielnie, od podstaw może tworzyć pożądane tabele. W tym celu należy wykonać następującą procedurę: *Analiza - Tabele specjalne – Tabele użytkownika*.

Tytuły – to zakładka, w której badacz może określić tytuł tabeli, czy stopkę (gdzie można podać np. źródło danych).

Statystyki podsumowujące – badacz ma możliwość wyboru sposobu prezentacji danych (liczebność, procent w wierszu, w kolumnie itd.) Okno jest aktywne dla zmiennej, którą określimy w polu *Źródło!*

Kategorie i podsumowania – badacz określa m.in. sposób sortowania kategorii, określa umiejscowienie podsumowań.

W oknie po lewej stronie znajduje się lista zmiennych. Ich wybór odbywa się poprzez przeciągnięcie zmiennej do okna po prawej stronie. Tu określamy, czy ma być ona wyświetlona w wierszu czy kolumnie. Badacz może wybrać więcej niż dwóch zmiennych (można tworzyć warstwy).

ROZDZIAŁ IV

ANALIZA OPISOWA DANYCH

W niniejszym dziale badacz pozna sposoby wyznaczania miar rozkładu. Do miar tych należą: miary tendencji centralnej, miary zmienności, miary asymetrii oraz miary koncentracji. Wpisują się one w dział statystyki opisowej, której celem, poza projektowaniem badań, czy sposobami gromadzenia informacji, jest przede wszystkim prezentacja i sumaryczny opis wyników. Sumaryczny, a zatem ogólny, zwięzły, za pomocą wartości liczbowych (pojedynczych liczb), które charakteryzują cały zbiór obserwacji.

4.1. Miary tendencji centralnej

Miary tendencji centralnej, określane są także jako miary przeciętne, miary pozycyjne lub miary położenia. Pierwszorzędne znaczenie mają miary przeciętne, które za pomocą liczb wyrażają prawidłowości niezauważalne w pojedynczych obserwacjach⁷.

4.1.1. Średnia arytmetyczna

Średnia arytmetyczna jest sumą wartości wszystkich wyników podzieloną przez liczbę elementów w zbiorze wyników. Stosujemy ją do zmiennych ilościowych (poziom pomiaru interwałowy lub ilorazowy)⁸. Obliczanie jej dla zmiennej nominalnej (liczona jest na podstawie kodów liczbowych) nie ma żadnego sensu.

$$M = \frac{x_1 + x_2 + x_n}{n}$$

Oto kilka cech średniej arytmetycznej:

- Średnia jest miarą stosowaną najczęściej, jej wartość liczona jest na podstawie wszystkich pomiarów.
- Średnia wskazuje typowy poziom natężenia mierzonej cechy.
- Na jej podstawie obliczane są liczne, bardziej zaawansowane statystyki.
- Wartość średniej zazwyczaj zmienia się najmniej od próby do próby, losowanych z tej samej populacji.

Średnia arytmetyczna ma również swoje ograniczenia:

- Jest miarą wrażliwą na wartości skrajne. W przypadku rozkładów skrajnie asymetrycznych traci swoje walory poznawcze. Innymi słowy, im rozkład jest bardziej jednorodny, tym wartość średniej jest bardziej adekwatna w opisie statystycznym.

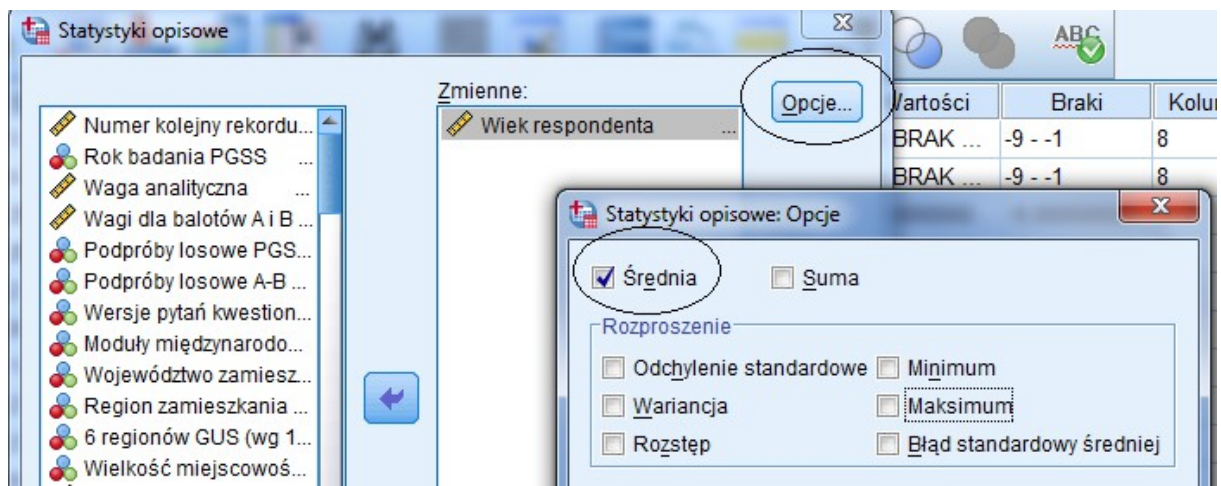
⁷ R. Szwed, *Metody statystyczne w naukach społecznych. Elementy teorii i zadania*, Lublin 2009, s. 52.

⁸ Wyjątkiem jest skala Likerta.

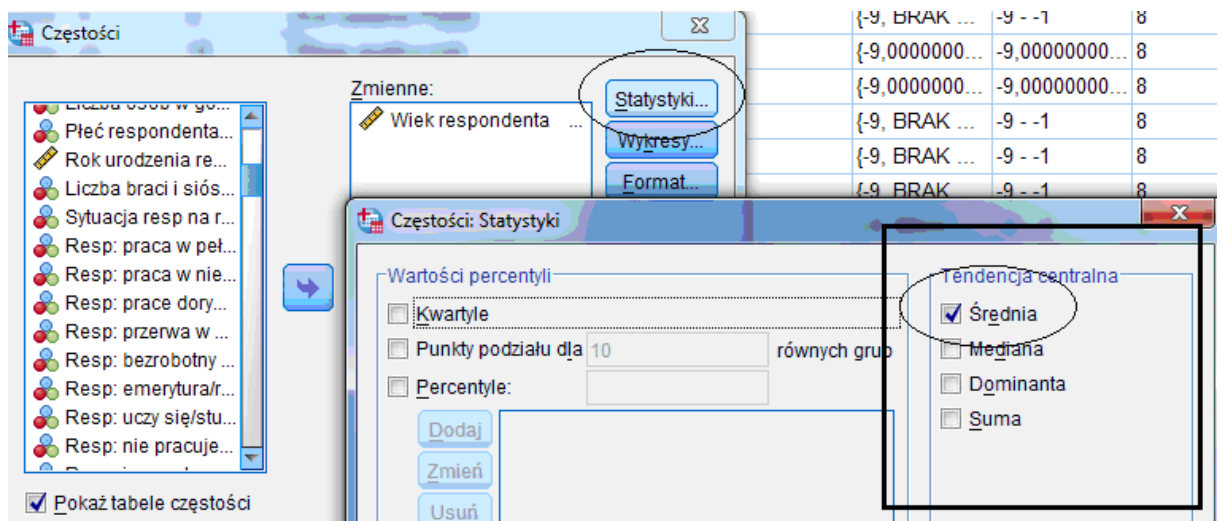
- Ograniczeniem średniej jest niemożność jej obliczenia w przypadku, gdy przedziały klasowe (pierwszy lub ostatni) są otwarte. Zazwyczaj w takich sytuacjach badacz domyka przedziały arbitralnie, ustalając jednocześnie środki przedziałów⁹. Nie można tego uczynić, gdy udział liczebności przedziału otwartego jest znaczny w porównaniu z liczebnością ogólną.
- Średnia jest wartością teoretyczną, co oznacza, że jej wartość nie musi się pokrywać z jakąkolwiek wartością w danym zbiorze. Średnia może zawierać liczby ułamkowe.

Średnią arytmetyczną w programie SPSS można obliczyć na kilka sposobów. Najprościej zrobić to za pomocą *Analiza – Opis statystyczny – Statystyki opisowe* lub poprzez *Analiza – Opis statystyczny - Częstości*.

W pierwszym przypadku należy zaznaczyć *średnia* w menu *Opcje* (PGSS2008.sav).



W drugim przypadku *średnią* zaznaczamy w menu *Statystyki*, w polu *Tendencja centralna*.



⁹ Przyjmuje się, że przedział ten powinien zawierać od 1,5 proc. do maksymalnie 5 proc. wszystkich jednostek analizy.

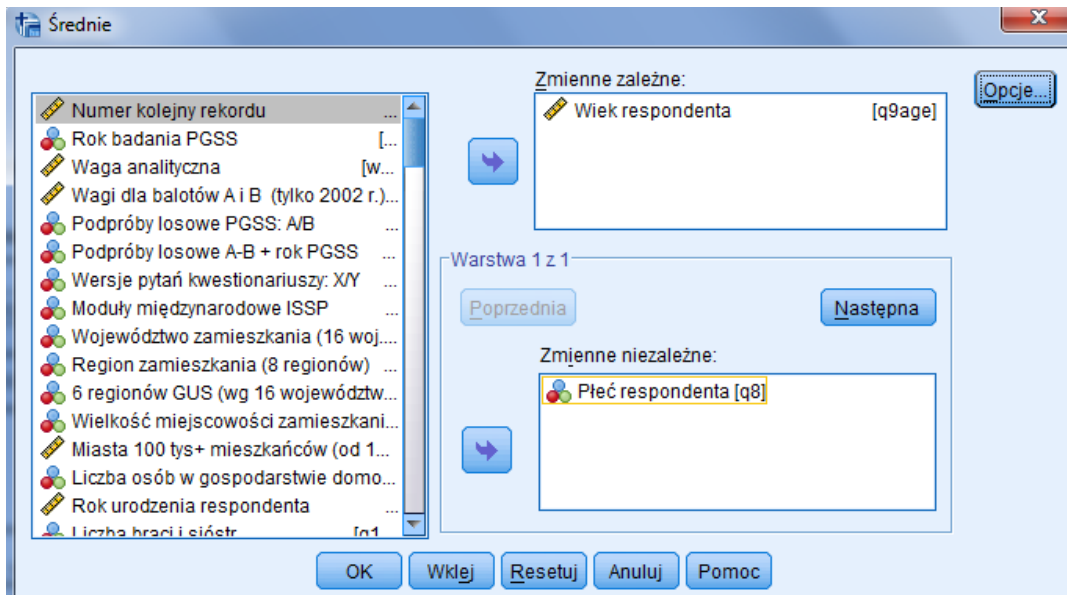
W naszym przykładzie do analiz użyliśmy zmiennej „wiek respondenta”. Pierwsza tabela w raporcie zatytułowana „Statystyki”, zawiera średnią arytmetyczną.

Statystyki

Wiek respondenta

N	Ważne	1495
	Braki danych	5
Średnia		46,23

Jeśli poza średnią ogólną (liczoną dla wszystkich jednostek analizy) interesuje nas średnia dla jakiejś jednej kategorii wówczas używamy *Analiza – Porównanie średnich – Średnie*. Załóżmy że chcemy znać średni wiek kobiet i mężczyzn. W oknie dialogowym w polu zmienna zależna (zm. ilościowa) wstawiamy zmienną wiek, zaś w polu zmienne niezależne, zmienną – płeć. Średnia wieku będzie obliczona dla każdej kategorii zmiennej niezależnej.



W wyniku tej procedury uzyskamy następujący raport:

Raport

Wiek respondenta

Płeć respondenta	Średnia	N	Odchylenie standardowe
Mężczyzna	45,34	640	16,949
Kobieta	46,89	855	17,742
Ogółem	46,23	1495	17,418

4.1.2. Średnia ważona

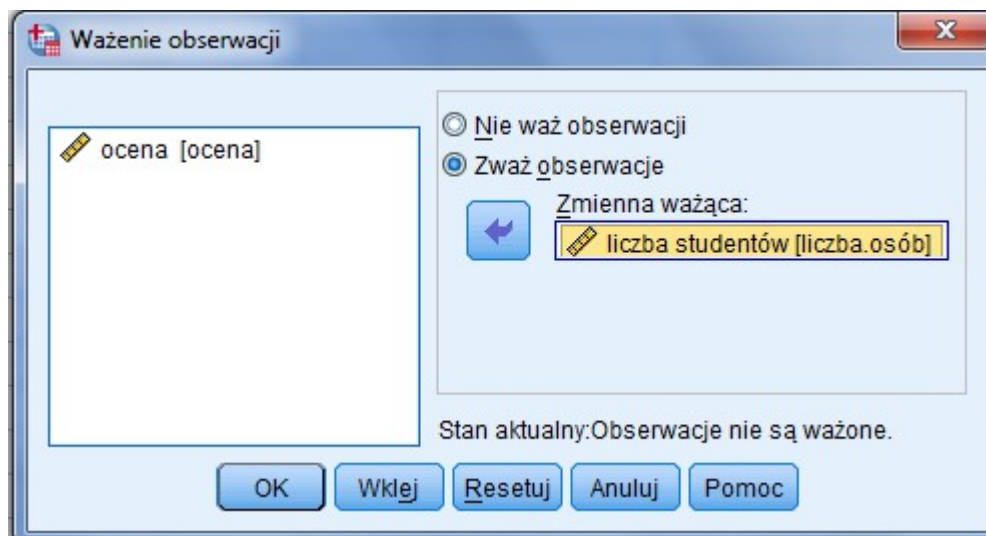
Średnia ważona może być różnie rozumiana. Po pierwsze jest ona średnią wyliczoną po uwzględnieniu liczebności każdej z uśrednianych wartości. Przykładowo chcemy wyliczyć średnią z egzaminu ze statystyki na III roku socjologii. Oto uzyskane wyniki:

ocena	liczba.osób
3,00	30,00
3,50	5,00
4,00	10,00
4,50	20,00
5,00	15,00
2,00	10,00

Obliczenie średniej ważonej odbywać się będzie według formuły:

$$M_w = \frac{(wartość_1 * waga_1) + (wartość_2 * waga_2) + (wartość_n * waga_n)}{waga_1 + waga_2 + waga_n}$$

W SPSS użyjemy procedury *Dane – Ważenie obserwacji*. W oknie dialogowym wybieramy zmienną ważącą. W naszym przypadku będzie to liczba studentów (którzy uzyskali określoną ocenę).



W raporcie wyświetli się informacja, że procedura ważenia przez zmienną „liczba studentów” jest uruchomiona (należy o tym pamiętać, ponieważ przed wykonaniem następnych analiz bez ważenia przypadków należy wyłączyć tą komendę). Aby zobaczyć efekty naszych działań, należy postępować dokładnie tak jak w przypadku obliczania średniej arytmetycznej. A więc klikamy kolejno *Analiza – Opis statystyczny – Statystyki opisowe* lub poprzez *Analiza – Opis statystyczny – Statystyki opisowe*. Oto nasze wyniki:

Statystyki

ocena

N	Ważne	90
	Braki danych	0
Średnia		3,6944

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	2,00	10	11,1	11,1	11,1
	3,00	30	33,3	33,3	44,4
	3,50	5	5,6	5,6	50,0
	4,00	10	11,1	11,1	61,1
	4,50	20	22,2	22,2	83,3
	5,00	15	16,7	16,7	100,0
	Ogółem	90	100,0	100,0	

Wyobraźmy sobie jednak inną sytuację. Mianowicie nauczyciel wystawia oceny ze statystyki. Każdy student ma ocenę z zaliczenia (ocena1), z samodzielnego projektu badawczego (ocena2) i pracy na zajęciach (ocena3). Oto oceny uzyskane przez jednego ze studentów:

statystyka	ocena
zaliczenie kolokwium	3,00
projekt	4,00
praca na zajęciach	5,00

Średnia arytmetyczna ocen to 4 ($12/3=4$). Nauczyciel postanowił jednak, że oceną końcową będzie średnia ważona z uzyskanych stopni, uznał bowiem, że ocena z kolokwium zaliczeniowego jest najważniejsza. I tak oto do kolokwium przypisano wagę 2, a do pozostałych ocen wagę 1. Średnia studenta w takim przypadku będzie niższa i wyniesie 3,75 (pomnożyliśmy kolejne oceny przez przypisane im wagi i podzieliliśmy przez sumę wag, a więc $15/4=3,75$). W tym przypadku średnia została wyliczona po uwzględnieniu ważności każdej z wartości. Jak możemy to zrobić w SPSS? Należy utworzyć nową zmienną *Waga*, a następnie użyć jej do ważenia obserwacji.

statystyka	ocena	waga
zaliczenie kolokwium	3,00	2,00
projekt	4,00	1,00
praca na zajęciach	5,00	1,00

Teraz wystarczy tylko wprowadzić komendy: *Analiza – Opis statystyczny – Statystyki opisowe*, użyć zmiennej ocena i otrzymamy raport z naszą średnią ważoną:

Statystyki opisowe

	N	Minimum	Maksimum	Średnia	Odchylenie standardowe
ocena	4	3,00	5,00	3,7500	,95743
N Ważnych (wyłączanie obserwacjami)	4				

Średnią ważoną należy obliczyć wtedy gdy¹⁰:

- dane są pogrupowane
- w próbie jest więcej obserwacji z jakiejś kategorii
- gdy chcemy obliczyć średnią zbiorczą dla kilku średnich

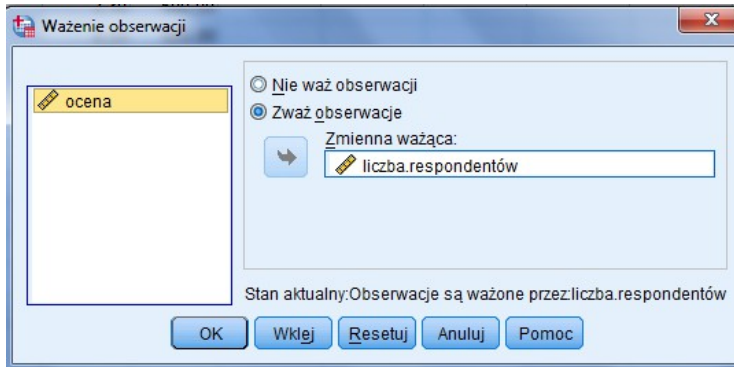
Na koniec przyjrzymy się bliżej temu ostatniemu przypadkowi. Załóżmy, że interesuje nas ocena polityki zagranicznej Stanów Zjednoczonych (wyrażona na skali od 1 do 10, gdzie – 1 to ocena zdecydowanie negatywna, a 10 – ocena zdecydowanie pozytywna). Wyniki uzyskane w poszczególnych państwach zestawiono ze sobą:

państwo	ocena	liczba.respondentów
Polska	5,60	500,00
Rosja	2,30	500,00
Ukraina	4,50	600,00
Wielka Brytania	3,70	800,00
Niemcy	4,60	800,00
Francja	3,80	500,00
Hiszpania	4,90	500,00
Norwegia	4,80	500,00
Czechy	5,70	500,00
Austria	4,10	500,00

Licząc średnią ważoną musimy wziąć pod uwagę liczbę respondentów w poszczególnych krajach

Dzięki komendzie *Ważenie obserwacji*, którą omówiliśmy wcześniej, jesteśmy w stanie bez problemu obliczyć średnią ocenę polityki zagranicznej USA wśród wszystkich wymienionych krajów.

¹⁰ J. Górniak, J. Wachnicki, *Pierwsze kroki w analizie danych, SPSS for Windows*, Kraków 2010, s. 140.



Następnie klikamy:
*Analiza – Opis
 statystyczny – Statystyki
 opisowe*, zaznaczamy
 zmienną ocena



Statystyki opisowe

	N	Minimum	Maksimum	Średnia	Odchylenie standardowe
ocena	5700	2,30	5,70	4,3754	,90101
N Ważnych (wyłączenie obserwacjami)	5700				

4.1.3. Średnia harmoniczna i średnia geometryczna

Średnia harmoniczna jest odwrotnością średniej arytmetycznej z odwrotności wartości zmiennej. Miara ta jest przydatna w przypadku przeliczania wartości cech na stałą jednostkę innej zmiennej np. kilometry na godzinę.

Obliczamy ją z formuły:

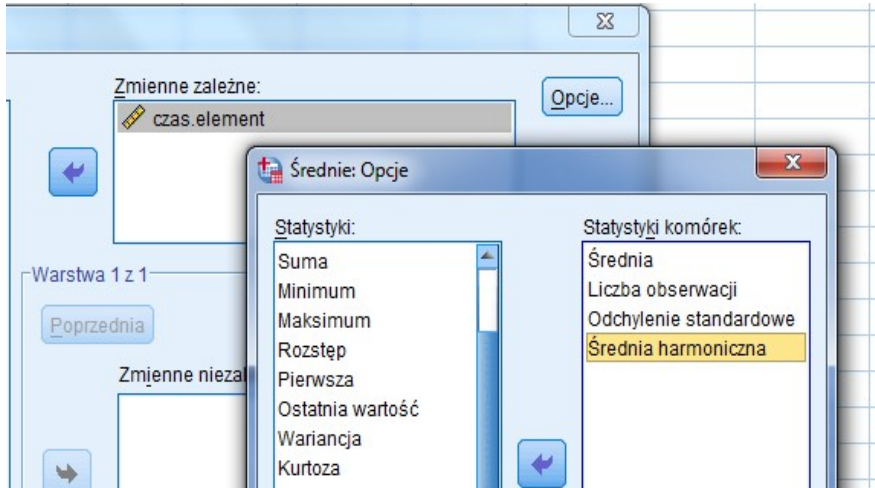
$$M_h = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_n}}$$

Rozpatrzmy następujący przykład. Pięciu pracowników pracuje w fabryce, w której produkuje się elementy składowe maszyn. Każdy z nich wykonuje tą samą pracę. W tabeli poniżej zebrano informacje na temat ich wydajności (czas, wyrażony w minutach, na wykonanie jednego elementu).

pracownik	czas.element
pracownik A	4,00
pracownik B	6,00
pracownik C	12,00
pracownik D	8,00
pracownik E	11,00

Ile wynosi średnia harmoniczna wydajności tzn. ile elementów w ciągu zmiany (8h pracy) wykonują wszyscy pracownicy? Statystykę tą możemy wygenerować klikając *Analiza-*

Porównanie średnich – Średnie. Następnie po naciśnięciu *Opcje* możemy wybrać, jakie statystyki wyświetli nam SPSS. W ten sposób możemy obliczyć nie tylko średnią arytmetyczną ale i średnią harmoniczną dla dowolnej zmiennej, wystarczy, że umieścimy ją na liście *Zmienne zależne* i nie wybierzemy *Zmiennych niezależnych*.



Nasz raport wygląda tak:

Raport

czas.element

Średnia	N	Odchylenie standardowe	Średnia harmoniczna
8,2000	5	3,34664	6,9841

Średnia geometryczna to iloczyn liczb, dla których chcemy obliczyć średnią, poddany pierwiastkowaniu o stopniu równym ilości liczb. Wzór na średnią geometryczną ma postać następującą:

$$M_g = \sqrt[n]{x_1 * x_2 * x_n}$$

A zatem obliczenie średniej geometrycznej z dwóch liczb polega na pomnożeniu tych liczb przez siebie, a następnie wyciągnięciu z otrzymanego iloczynu pierwiastka kwadratowego. Średnia ta jest szczególnie przydatna do obliczania zjawisk zmieniających się w czasie. Przykładowo interesuje nas zysk z lokaty, którą założyliśmy. Lokata ma zmienne oprocentowanie, po pierwszym roku wynosi ono 4%, po drugim 5% i po trzecim 6%. Jakie jest zatem średnie oprocentowanie? Aby odpowiedzieć na to pytanie posłużymy się właśnie średnią geometryczną. Średnią geometryczną w SPSS obliczamy dokładnie tak samo jak robiliśmy to w przypadku średniej harmonicznnej. We wskazanym przykładzie wynosi ona 4,93%.

Średnią geometryczną i harmoniczną znajdziemy również w menu *Analiza – Raporty i zestawienia – Podsumowania obserwacji...* (PGSS2008.sav)

Przeciągamy interesującą nas zmienną w pole *Zmienne*

Klikamy *Statystyki* i odnajdujemy interesujące nas średnie.

Jeśli chcemy aby w raporcie wyświetliły się same statystyki odznaczamy opcję *Pokaż obserwacje*.

4.1.4 Średnia obcięta

Średnia obcięta inaczej nazywana średnią odciętą liczona jest podobnie jak średnia arytmetyczna, z tym, że z niepełnego zakresu danych, ponieważ odcięte zostają wartości skrajne, w równej proporcji - z góry i z dołu zakresu wartości danego zbioru danych. Zazwyczaj odcina się 5% wartości najwyższych i 5% wartości najniższych. W programie SPSS zakres ten jest stały i wynosi dokładnie 5% (5% „górne” i 5% „dolne”). Średnią odciętą można uzyskać poprzez wybranie *Analiza – Opis statystyczny – Eksploracja*.

W oknie dialogowym w polu *Zmienne zależne* przeciągamy interesującą nas zmienną. Pozostałe pola pozostawiamy puste. We wskazanym przypadku wybraliśmy zmienną wiek respondenta. Zwróćmy uwagę, że w raporcie otrzymujemy informację nie tylko o średniej

odciętej ale również o innych miarach opisowych (to pokazuje, że wygenerowanie statystyk opisowych w SPSS możliwe jest na kilka sposobów).

Statystyki opisowe (DESCRIPTIVES)

		Statystyka	Błąd standardowy	
Wiek respondenta	Średnia	46,23	,450	
	95% przedział ufności dla średniej	Dolna granica	45,34	
		Górna granica	47,11	
	5% średnia obcięta	45,64		
	Mediana	43,00		
	Wariancja	303,386		
	Odchylenie standardowe	17,418		
	Minimum	18		
	Maksimum	89		
	Rozstęp	71		
	Rozstęp ćwiartkowy	27		
	Skośność	,500	,063	
	Kurtoza	-,700	,126	

Jeśli w głównym oknie dialogowym nie zaznaczyliśmy aby program pokazywał tylko statystyki, wyświetlą nam się również wykresy (więcej na ich na temat w rozdziale dotyczącym graficznej prezentacji wyników w SPSS).

Szczególnym przypadkiem średniej odciętej jest tzw. średnia winsorowska, w przypadku której wartości najmniejsze i największe (po tyle samo) zastępujemy wartościami bezpośrednio z nimi sąsiadującymi. Czynność tą przyjęło się określać winsoryzacją. To badacz decyduje ile wartości ma być zastąpionych.

wartości	winsoryzacja
1,00	5,00
5,00	5,00
5,00	5,00
6,00	6,00
6,00	6,00
7,00	7,00
7,00	7,00
8,00	8,00
8,00	8,00
15,00	8,00

Winsoryzacja w tym przypadku polegać będzie na zastąpieniu wartości skrajnych tj. 1 i 15 nowymi wartościami, z nimi sąsiadującymi, a są to kolejno 5 i 8. Dopiero po tej czynności dokonujemy obliczenia średniej arytmetycznej. Średnia w naszym przypadku wynosi 6,5 (przed winsoryzacją wynosi 6,8).

Program SPSS umożliwia nam również obliczenie statystyk, które poprzez uwzględnienie rozkładu danych, „uodporniają” średnią na wartości skrajne. Statystyki te służą do lepszego szacowania wartości tendencji centralnej w populacji na podstawie statystyk z próby (stanowią alternatywę dla mediany i średniej). SPSS dostarcza nam cztery możliwości, które

określamy jako M-estymatory, są to statystyki: Tukeya, Hubera, Andrew i Hampela. Opierają się one na różnych sposobach ważenia obserwacji, o czym informuje nas adnotacja pod tabelą w raporcie. M - estymatory stosowane są gdy rozkład zmiennej jest asymetryczny lub symetryczny lecz z długimi ogonami po lewej i prawej stronie. Odnajdziemy je w menu *Analiza – Opis statystyczny – Eksploracja*. W oknie dialogowym klikamy opcję *Statystyki* i zaznaczamy *M-estymatory*. W raporcie jedna z tabelek będzie wyglądać następująco:

M-estymatory

	Huber ^a	Tukey ^b	Hampel ^c	Andrew ^d
Wiek respondenta	44,13	43,76	44,72	43,77

- 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.

Obliczenie średniej przy pomocy pakietu SPSS jest bardzo proste. Problemem badacza jest przede wszystkim odpowiedź na pytanie, którą średnią obliczyć? Oto kilka wskazówek:

- gdy rozkład jest symetryczny (zobacz miary skośności) najlepiej jak zastosujemy średnią arytmetyczną
- gdy występują wartości odstające, które mogą zawyżać średnią arytmetyczną, lepsza będzie średnia odcięta lub M-estymatory.
- gdy interesuje nas zmienność zjawiska w czasie użyjemy średnią geometryczną
- średnia harmoniczna z kolei, jest przydatna do przeliczania wartości cech na stałą jednostkę innej zmiennej np. sztuki na minutę.

4.1.5. Mediana

Przy okazji omawiania średnie arytmetycznej wspominaliśmy już o medianie. Teraz powiem coś więcej. Mediana to inaczej kwartył drugi, który dzieli zbiorowość na połowę. Innymi słowy 50% obserwacji jest poniżej mediany i 50% powyżej mediany w uporządkowanym zbiorze danych. Obok średniej arytmetycznej stanowi jedną z najczęściej stosowanych miar tendencji centralnej. Medianę można obliczyć dla zbioru o parzystej liczbie elementów, wówczas stosujemy formułę:

$$Me = \frac{1}{2} * \left(\frac{n}{2} + \frac{n+2}{2} \right)$$

W przypadku zbioru o nieparzystej liczbie elementów korzystamy z formuły:

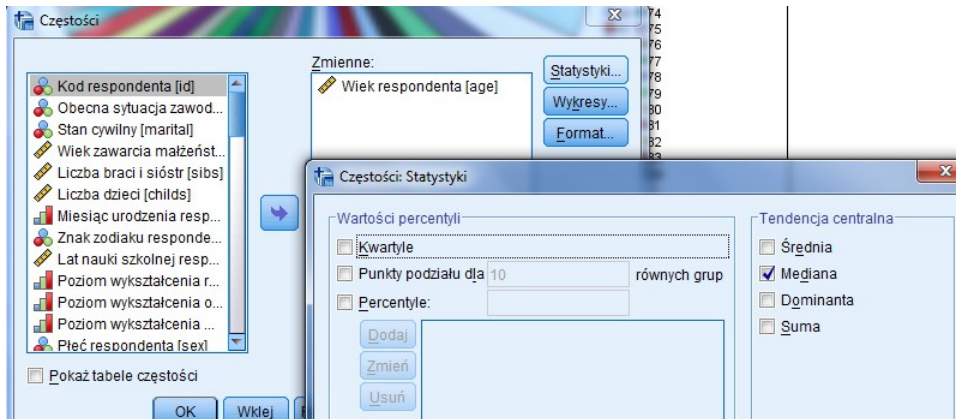
$$Me = \frac{n + 1}{2}$$

Mediana jest całkowicie odporna na wartości skraje, stąd też jest lepszą, niż średnia arytmetyczna, miarą tendencji centralnej w przypadku rozkładów asymetrycznych. Bezużyteczna jest natomiast w przypadku rozkładów wielomodalnych.

W programie SPSS istnieje kilka możliwości wygenerowania mediany. Większość z nich poznaliśmy, przy okazji obliczania średniej. Możemy użyć komend:

- *Analiza – Porównanie średnich – Średnie* (w menu *Opcje* zaznaczamy *mediana*)
- *Analiza – Opis statystyczny – Częstości* (w menu *Statystyki* zaznaczamy *mediana*)
- *Analiza – Opis statystyczny – Eksploracja*

Oto przykład (GSS93.sav):



Statystyki

Wiek respondenta

N	Ważne	1495
	Braki danych	5
Mediana		43,00

Mediana równa 43 oznacza, że 50% respondentów ma mniej niż 43 lata i równocześnie 50% badanych ma więcej niż 43 lata.

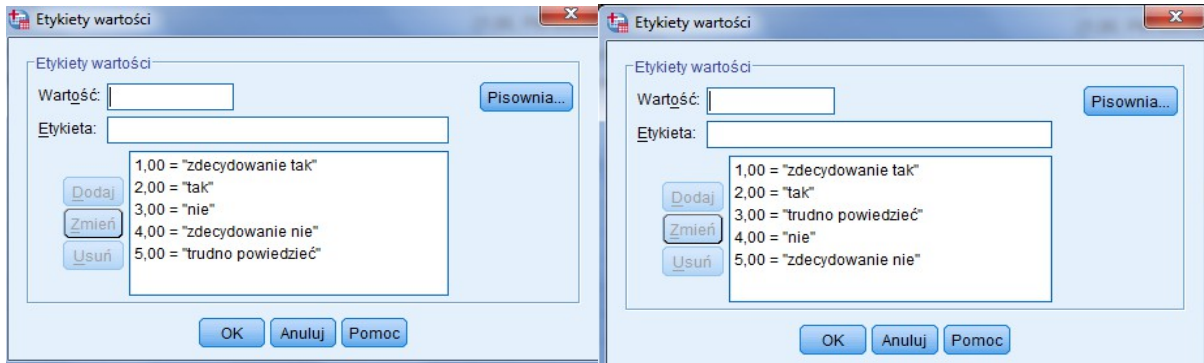
Medianę można obliczyć na skali porządkowej i ilościowej, z tym, że na skali porządkowej nie ma ona sensu liczbowego. Wskazuje ona jedynie do której kategorii, w uporządkowanym pod względem nasilenia cechy zbiorze, należy środkowa obserwacja. W przypadku zmiennych mierzalnych jest inaczej, mediana ma swój sens liczbowy. Wskazuje nam wartość obserwacji, która dzieli zbiorowość na pół. Innymi słowy mediana wyznacza środkową obserwację, w odniesieniu do której możemy powiedzieć, że połowa pozostałych obserwacji ma wartości nie większe niż wartość mediany i równocześnie połowa obserwacji ma wartości nie mniejsze.

Inne użyteczne cechy mediany to:

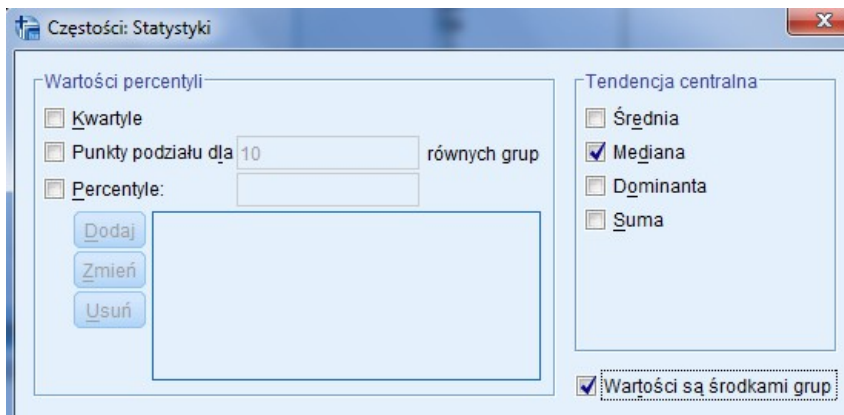
- obliczamy ją gdy interesuje nas, czy obserwacje przypadają w dolnej czy górnej połowie rozkładu
- należy ją oszacować gdy dany jest niepełny rozkład, można ją obliczyć gdy krańce rozkładu są otwarte
- w przypadku rozkładów bardzo skośnych jest lepsza niż średnia arytmetyczna.

Medianę można obliczyć tylko w zbiorze uporządkowanym. Szczególną uwagę należy zwrócić na zmienne typu porządkowego. Otóż, często sytuacją jest umieszczanie w kwestionariuszu

odpowiedzi „trudno powiedzieć” lub „ani tak, ani nie” na ostatnim miejscu w kafeterii. Zazwyczaj w procesie kodowania nadaje się jej ostatni kod cyfrowy. Jeśli jednak odpowiedź „trudno powiedzieć” lub „ani tak, ani nie” ma wyrażać śródkowe nasilenie cechy to przed wyznaczeniem mediany w SPSS, należy dokonać rekodowania, tak aby przypisać jej stosowną rangę. Obrazują to poniższe rysunki:

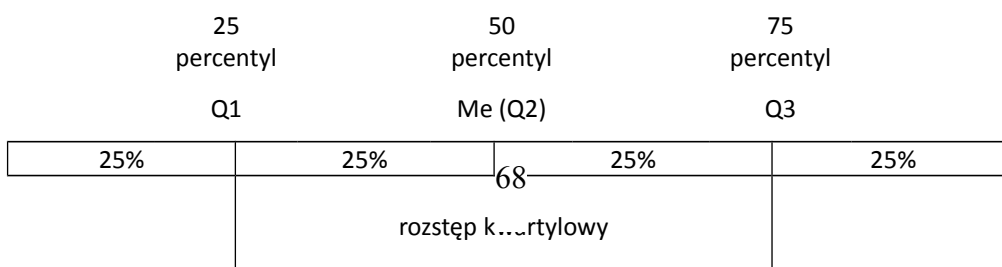


Kilka słów należy poświęcić również pogrupowanej zmiennej ilościowej. Aby wyznaczyć medianę - w takim przypadku - najpierw należy zakodować przedziały za pomocą wartości śródkowej w przedziale. Następnie korzystamy z komend *Analiza - Opis statystyczny - Częstości*. W wyświetlonym oknie dialogowym klikamy *Statystyki* a następnie poza *Medianą* zaznaczamy opcję *Wartości są środkami grup*.

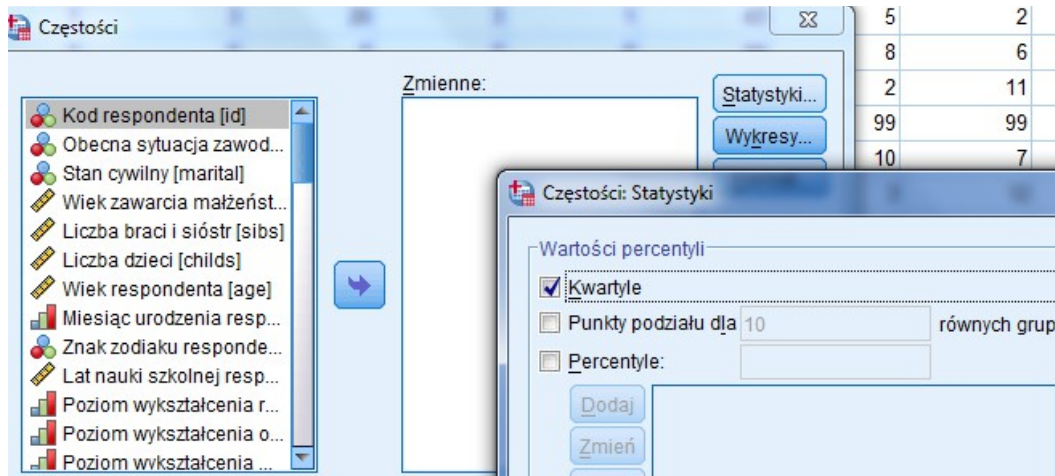


Medianę dla zmiennych pogrupowanych uzyskamy również klikając kolejno: *Analiza - Raporty i zestawienia - Podsumowanie obserwacji* a następnie zaznaczając w menu *Statystyki* opcję *Mediana z danych pogrupowanych*.

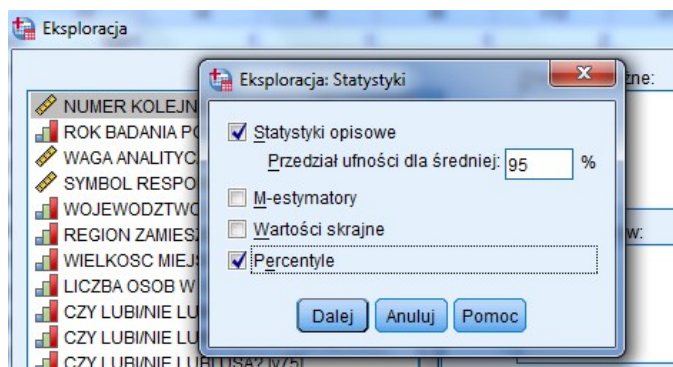
Miary, które mają zbliżony sens do mediany mają pozostałe kwartyle (to rodzaj kwantyli), choć nie mierzą one tendencji centralnej. Ich wyznaczenie polega na podziale zbiorowości na określone części. Pierwszym kwartylem jest wartość zmiennej poniżej której leży 25% obserwacji w uporządkowanym szeregu, zaś kwartyl trzeci wyznacza wartość poniżej której leży 75% obserwacji.



Kwartale można obliczyć dzięki procedurze *Analiza - Opis statystyczny – Częstości*. W menu *Statystyki* zaznaczamy opcję *Kwartyle*. Przed przystąpieniem do obliczeń należy zwrócić uwagę, czy w opcji *Format* porządkowanie ustawione jest według wartości (jest to ustawienie domyślne) a nie według liczebności.



Zwróćmy uwagę, że wskazany podział nie jest jedynym. W oknie dialogowym mamy również możliwość elastycznego definiowania kwantyli tzn. możemy określić na ile równych części ma być podzielona zbiorowość. Z kolei w polu *percentyle* badacz wyznacza dowolny kwantyl, dopuszczalne są wartości z zakresu 0-100. Ponadto percentyl 5, 10, 25, 50, 75, 90 i 95 możemy wyznaczyć poprzez menu *Eksploracja – Statystyki* (procedurę poprzedzamy poleceniami *Analiza – Opis statystyczny*).



Wówczas otrzymamy następujący raport:

		Percentyle						
		Percentyle						
		5	10	25	50	75	90	95
Przeciętne ważone (Definicja 1)	Wiek respondenta	22,00	25,00	32,00	43,00	59,00	73,00	78,00
Zawiasy Tukey'a	Wiek respondenta			32,50	43,00	59,00		

Warto pamiętać że w dużych grupach ($n > 100$) wyznaczenie kwartyli jest dosyć proste. W przypadku małych grup wartości te muszą być interpolowane, wówczas korzystamy z zawiasów Tukeya (na wartościach tych opierają się wykresy skrzynkowe), w przypadku których wartość kwartyli pierwszego i trzeciego mogą się różnić od tych wskazanych w wierszu *Przeciętne ważone*.

4.1.6. Dominanta

Dominanta, inaczej moda, modalna, to mówiąc najprościej wartość zmiennej, która występuje najczęściej. Dominanta jest miarą tendencji centralnej w tym sensie, że wskazuje, która kategoria jest typowa.

Dominantę można obliczyć na każdym poziomie pomiaru, z tym, że w przypadku zmiennej pogrupowanej najpierw musimy wskazać przedział klasowy dominanty, by dopiero później obliczyć jej wartość z formuły¹¹:

$$D = x_d + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} * i_d$$

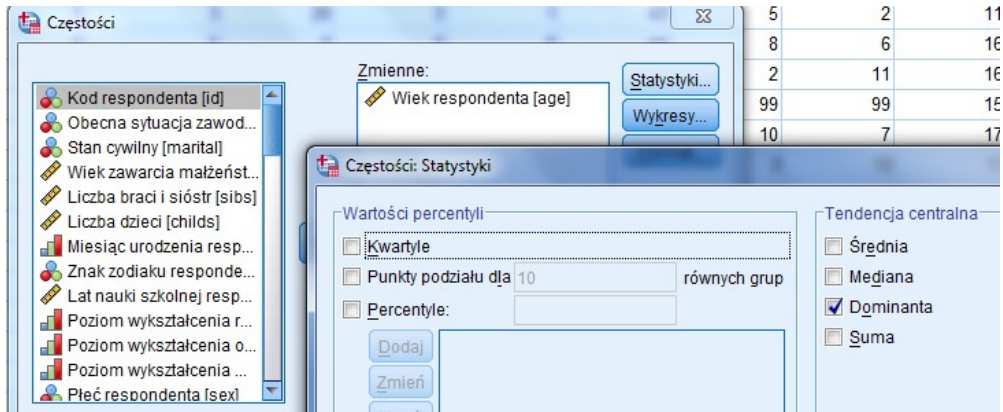
x_d - to dolna granica przedziału, w którym jest dominanta; n_d - to liczebność przedziału dominanty, n_{d+1} - liczebność przedziału następującego po przedziale dominanty, n_{d-1} - liczebność przedziału poprzedzającego przedział dominanty; i_d - interwał przedziału dominanty

Wyznaczenie dominanty jest niemożliwe, gdy dwa przedziały sąsiadujące z przedziałem dominanty nie są równe oraz gdy szereg jest skrajnie asymetryczny lub ma otwarty przedział dominujący (ostatni lub pierwszy w szeregu). Dominanta poza łatwością obliczenia (wyznaczenia) i interpretacji ma swoje wady:

- po pierwsze może zafałszowywać rzeczywistość, Dzieje się tak, gdy kategoria występująca najczęściej, nie występuje dużo częściej niż pozostałe kategorie. Wartość informacyjna dominanty jest wówczas bardzo mała
- po drugie dominanta może być bezużyteczna, np. wtedy gdy rozkład ma więcej niż jedną modę (rozkłady bimodalne, wielomodalne) lub nie ma jej w ogóle (rozkłady o równomiernej liczebności poszczególnych kategorii)
- ponadto należy pamiętać, że w przypadku modyfikacji danych pierwotnych np. w procesie rekodowania, modalna może ulec „wymuszonej” przez badacza zmianie.

W SPSS (GSS93.sav) dominantę znajdziemy w menu *Analiza – Opis statystyczny – Częstości (szukamy jej w Statystykach)*.

¹¹ W programie SPSS, w przypadku danych pogrupowanych, dominantą jest cały przedział klasowy zmiennej, występujący z największą częstotliwością.



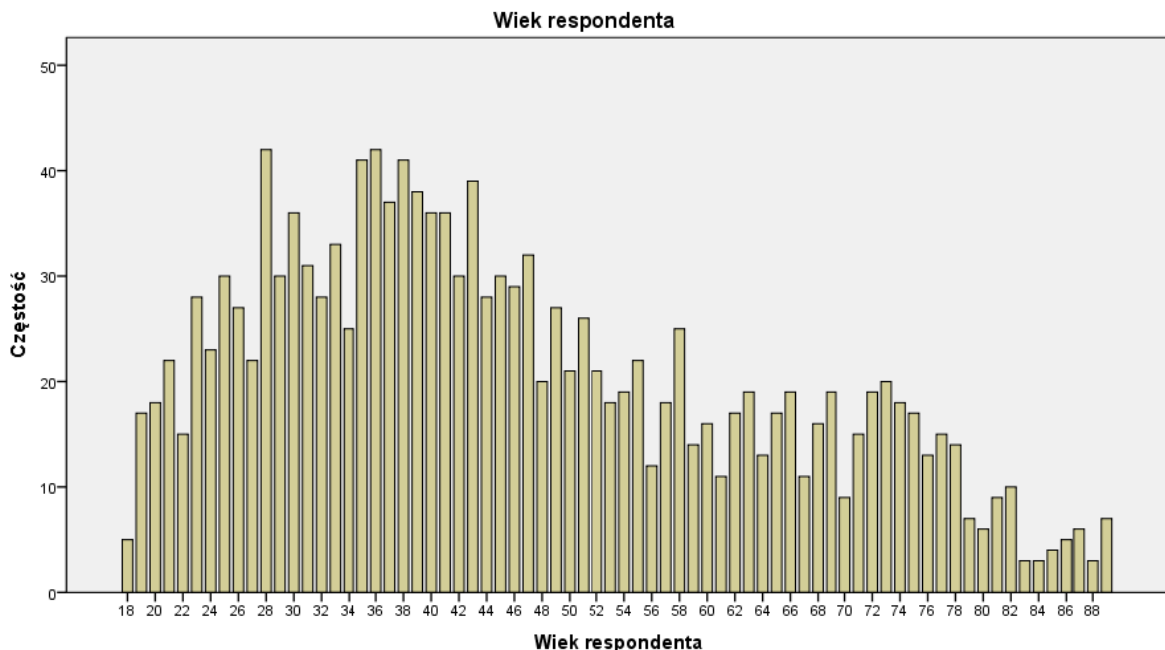
Statystyki

Wiek respondenta

N	Ważne	1495
	Braki danych	5
Dominanta		28 ^a

a. Istnieje wiele wartości modalnych. Podano wartość najmniejszą.

Jeśli w zbiorze jest więcej niż jedna dominanta, a tak jest w naszym przypadku, to SPSS informuje nas o tym pod tabelką ze statystykami. Wartość - 28 lat to tylko jedna z modalnych, program automatycznie generuje wartość najniższą. Warto zatem rzucić okiem na rozkład częstości zmiennej lub na jej wykres.

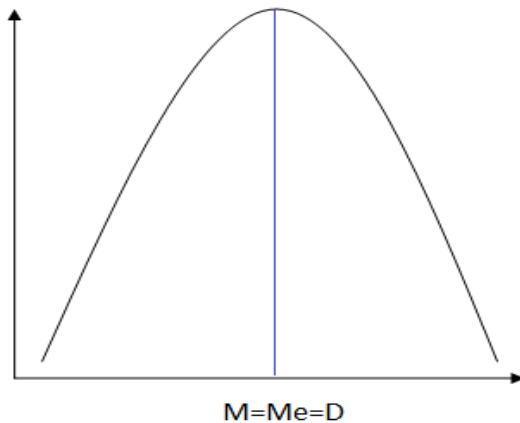


Na wykresie widać wyraźnie, że są wartości występujące równie często co 28 lub niewiele rzadziej. Dominanta w tym przypadku nie będzie najlepszą miarą tendencji centralnej.

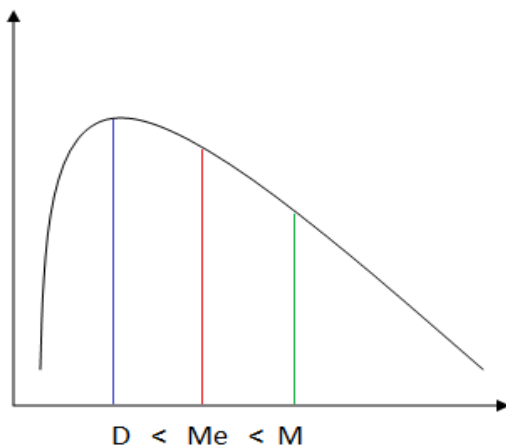
4.1.7. Porównanie miar tendencji centralnej

Każda z omówionych wcześniej miar jest właściwa dla określonego poziomu pomiaru: średnia dla poziomu ilościowego, mediana dla poziomu porządkowego, dominanta dla poziomu nominalnego. Należy jednak pamiętać, że poza poziomem pomiaru badacz musi przeanalizować rozkład danej zmiennej. I tak np. średnia arytmetyczna zawodzi w przypadku rozkładów skrajnych, lepsza jest wtedy mediana. Z kolei, dominanta jest bezużyteczna w rozkładzie wielomodalnym.

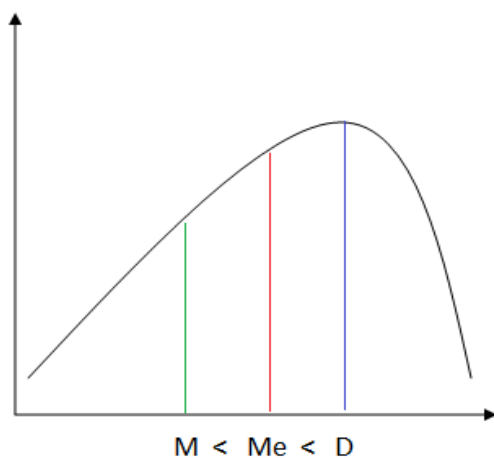
Pomiędzy średnią, medianą a dominantą zachodzą określone relacje.



W przypadku rozkładu symetrycznego wszystkie statystyki są tą samą wartością.



W przypadku rozkładu prawoskośnego (skośnego dodatnio) pomiędzy miarami zachodzi relacja: $D < Me < M$



W przypadku rozkładu lewoskośnego (skośnego ujemnie) pomiędzy miarami zachodzi relacja: $D > Me > M$

Miary tendencji centralnej pozostają również w zależności matematycznej, którą określamy jako wzór Pearsona:

$$D = 3 * M - 2 * Me$$

4.2. Miary zmienności

Opieranie się na miarach tendencji centralnej może prowadzić do wielu błędów i nadużyć w opisie badanej zbiorowości. Załóżmy że analizujemy wyniki 5 uczniów z klasówki. W klasie A uczniowie uzyskali następujące oceny: 1, 2, 3, 4, 5 a zatem średnia ocen to 3. W klasie B wszyscy uczniowie dostali te same oceny: 3, 3, 3, 3, 3, średnia w tym przypadku również wynosi 3. Czy możemy powiedzieć, że poziom uczniów w obu klasach jest taki sam? Zanim odpowiemy na to pytanie przyjrzyjmy się bliżej miarom zmienności (określane też jako miary dyspersji, rozproszenia lub rozrzutu).

4.2.1. Rozstęp

Rozstęp to najprostsza miara zmienności. Jej obliczenie opiera się na wartości najwyższej w zbiorze danych, którą określamy jako maksimum i wartości najniższej, czyli minimum. Rozstęp to różnica pomiędzy tymi wartościami:

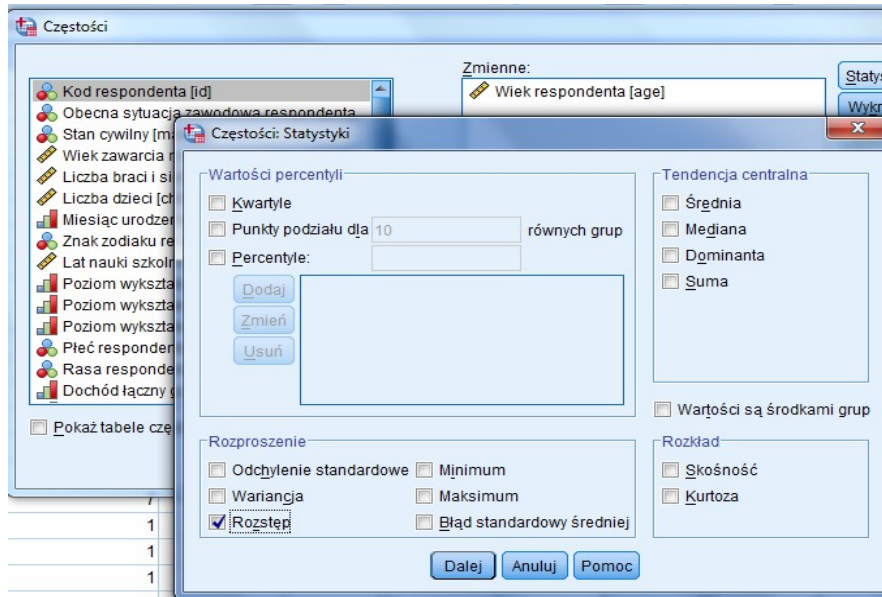
$$R = X_{max} - X_{min}$$

Mimo swoich zalet takich jak łatwość interpretacji, czy łatwość obliczenia, rozstęp ma więcej wad. Najpoważniejszą wydaje się duża wrażliwość na wartości skrajne. Ponadto rozstęp opiera się tylko na dwóch wartościach w zbiorze.

Wyznaczenie rozstępu, podobnie jak innych miar zmienności, przy pomocy SPSS możliwe jest na kilka sposobów, trzy z nich to:

- *Analiza – Opis statystyczny – Częstości* (w menu *Statystyki* zaznaczamy *rozstęp*)
- *Analiza – Opis statystyczny – Statystyki opisowe* (w menu *Opcje* zaznaczamy *rozstęp*)
- *Analiza – Opis statystyczny – Eksploracja*

Wykorzystamy pierwszy sposób *Analiza – Opis statystyczny – Częstości* (GSS83.sav):



Statystyki

Wiek respondenta

N	Ważne	1495
	Braki danych	5
	Rozstęp	71

4.2.2. Wariancja i odchylenie standardowe

Wariancją to klasyczna miara zmienności, która poza samodzielnym zastosowaniem, używana jest przy bardziej zaawansowanych analizach statystycznych m.in. w jednoczynnikowej analizie wariancji (ANOVA), wieloczynnikowej analizie wariancji, czy testowaniu hipotez testami t. Z matematycznego punktu widzenia wariancja to suma podniesionych do kwadratu różnic wartości pojedynczych pomiarów i średniej którą następnie dzieli się przez liczbę jednostek analizy w zbiorze danych. Pierwiastek kwadratowy stosuje się w celu wyeliminowania znaków ujemnych, które pojawiają się dla wartości zmiennej niższej niż średnia:

$$s^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \dots + (x_n - M)^2}{n}$$

Wariancja przyjmuje wartości od 0 do nieskończoności. Im wartość bliższa zero, tym zróżnicowanie jest mniejsze. Należy pamiętać, że wartość wariancji wyrażona jest w jednostkach nienaturalnych dla danej zmiennej, ze względu na podniesienie różnic kolejnych pomiarów i średniej do kwadratu. Odchylenie standardowe to klasyczna miara zmienności, którą stosujemy najczęściej. Określa ona o ile wszystkie badane jednostki statystyczne zbiorowości różnią się przeciętnie od średniej arytmetycznej. Dzięki odchyleniu możemy określić typowe wartości występują w zbiorze, tzn. jakich wartości należy oczekiwać, a jakie wartości są ekstremalne.

Odchylenie standardowe to pierwiastek wyciągnięty z wariancji:

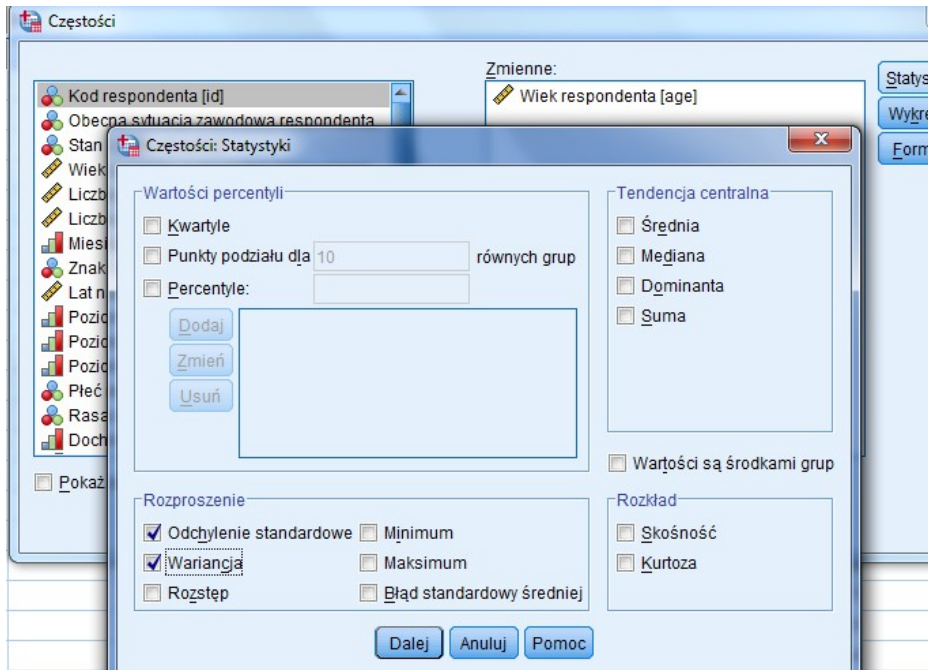
$$s = \sqrt{s^2}$$

Odchylenie standardowe - podobnie jak wariancja - zawiera się w przedziale od 0 do nieskończoności. Im wartość odchylenia bliższa zero, tym zróżnicowanie jest mniejsze. W odróżnieniu od wariancji, wynik odchylenia standardowego wyrażony jest w jednostkach miary analizowanej zmiennej. Oznacza to, że jeśli analizowana zmienną jest np. wiek, to badacz może określić o ile lat przeciętnie różnią się wszystkie jednostki od średniego wieku tychże jednostek (zobacz przykład i porównaj wynik wariancji i odchylenia standardowego).

Wariancję i odchylenie standardowe w SPSS można wyznaczyć wykorzystując różne komendy, m.in.:

- *Analiza – Opis statystyczny – Częstości* (w menu *Statystyki* zaznaczamy *wariancję* lub *odchylenie standardowe*)
- *Analiza – Opis statystyczny – Statystyki opisowe* (w menu *Opcje* zaznaczamy *wariancję* lub *odchylenie standardowe*)
- *Analiza – Opis statystyczny – Eksploracja*
- *Analiza – Raporty zestawienia – Podsumowania obserwacji* (w menu *Statystyki* zaznaczamy *wariancję* lub *odchylenie standardowe*)
- *Analiza-Porównanie średnich – Średnie* (w menu *Opcje* zaznaczamy *wariancję* lub *odchylenie standardowe*)

Poniżej zastosowano procedurę *Analiza – Opis statystyczny – Częstości* (GSS83.sav):



Statystyki

Wiek respondenta

N	Ważne	1495
	Braki danych	5
Odchylenie standardowe		17,418
Wariancja		303,386

4.2.3. Rozstęp ćwiartkowy i odchylenie ćwiartkowe

Rozstęp ćwiartkowy i odchylenie ćwiartkowe zaliczamy do miar rozproszenia opartych na kwantylach. Miary te stosujemy na poziomie ilościowym lub porządkowym.

Rozstęp ćwiartkowy to różnica pomiędzy kwartylem trzecim i pierwszym:

$$R_Q = Q_3 - Q_1$$

Pomiędzy tymi kwartylami mieści się połowa wszystkich jednostek analizy. Im większy zakres tego przedziału, tym większe zróżnicowanie zmiennej. Rozstęp ćwiartkowy stosujemy wtedy, gdy rozkład zmiennej jest niepełny lub występuje szereg otwarty.

Rozstęp ćwiartkowy możemy policzyć po wyznaczeniu kwartyli lub poprzez procedurę *Analiza – Opis statystyczny – Eksploracja* (GSS83.sav).

The screenshot shows the SPSS 'Eksploracja' (Explore) dialog box with 'Wiek respondenta [age]' selected as the dependent variable. Below the dialog box, the 'Statystyki opisowe (DESCRIPTIVES)' output window is visible, showing a table of descriptive statistics for 'Wiek respondenta'.

		Statystyka	Błąd standardowy
Wiek respondenta	Średnia	46,23	,450
	95% przedział ufności dla średniej	Dolna granica Górna granica	45,34 47,11
	5% średnia obcięta	45,64	
	Mediana	43,00	
	Wariancja	303,386	
	Odchylenie standardowe	17,418	
	Minimum	18	
	Maksimum	89	
	Rozstęp	71	
	Rozstęp ćwiartkowy	27	
	Skośność	,500	,063
	Kurtoza	-,700	,126

Odchylenie ćwiartkowe (rozstęp międzykwartylowy) mieści w sobie połowę jednostek analizy znajdujących się pomiędzy trzecim a pierwszym kwartylem:

$$Q = \frac{Q_3 - Q_1}{2}$$

4.2.4. Współczynnik zmienności

Współczynnik zmienności jest miarą niezwykle przydatną do porównywania rozproszenia wyników w dwu lub więcej grupach, a także w sytuacji, gdy poddajemy analizie jedną zbiorowość, ale uwzględniamy różne jej cechy. Współczynnik zmienności umożliwia porównywanie zmiennych mierzonych na różnych skalach. Miara ta wyrażana jest ilorazem odchylenia standardowego i średniej arytmetycznej:

$$V = \frac{s}{M}$$

Współczynnik zmienności, który wyrażamy w procentach, zwykle waha się w granicach 15-35%. Można przyjąć, że gdy jego wielkość jest większa niż 60%, to zmienność jest bardzo duża, a badana grupa jest względnie niejednorodna z punktu widzenia badanej cechy¹²:

Wartość współczynnika	Interpretacja
0-20%	nieznaczne zróżnicowanie wartości zmiennej, grupa jest względnie jednorodna, średnia arytmetyczna jest adekwatną miarą charakteryzującą grupę
20-40%	umiarkowane zróżnicowanie wartości zmiennej, średnia arytmetyczna jest akceptowalną miarą dla danej zmiennej
40-60%	silne zróżnicowanie wartości zmiennej - rozproszenie zmiennej jest znaczne, średnia arytmetyczna ma małą wartość poznawczą
od 60%	bardzo silne zróżnicowanie wartości zmiennej, grupa jest niejednorodna, arytmetyczna nie ma żadnej wartości poznawczej

Niestety, poza językiem poleceń (który polecamy bardziej zaawansowanym użytkownikom) programu SPSS, nie można automatycznie wygenerować współczynnika zmienności. Aby go wyznaczyć należy dokonać samodzielnego, „tradycyjnego” obliczenia, korzystając z wygenerowanych w SPSS, koniecznych do tego celu miar.

4.2.5. Porównanie miar zmienności

Wybór odpowiedniej miary zmienności, podobnie jak w przypadku miar tendencji centralnej, podyktowany jest, po pierwsze tym co badacz „chce wiedzieć” o uzyskanych wynikach, po drugie to z jakim rozkładem ma do czynienia. Oto kilka praktycznych wskazówek:

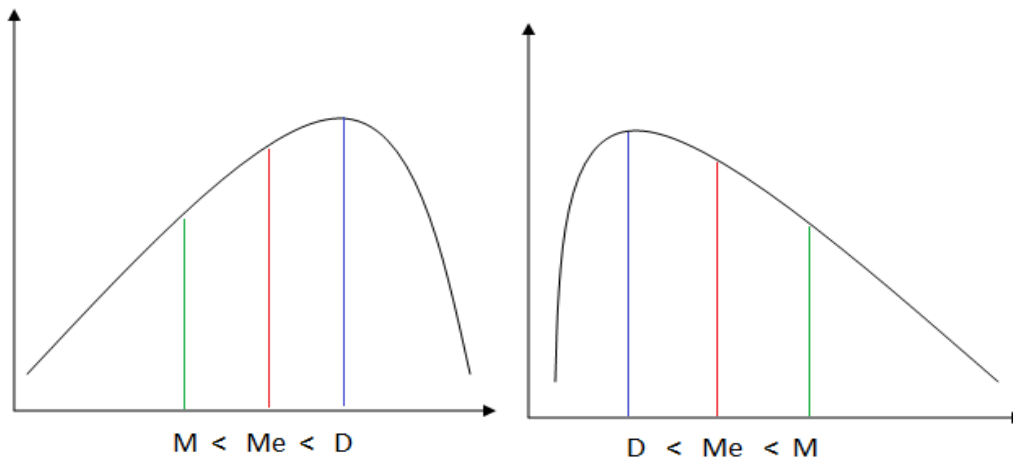
- rozstępu użyjemy, gdy potrzebna nam jest informacja na temat skrajnych pomiarów w zbiorze danych lub musimy szybko (i w łatwy sposób) ocenić rozproszenie wyników

¹² K. Zajac, *Zarys metod statystycznych*, Warszawa 1988, s. 209.

- odchylenie ćwiartkowe jest najbardziej „przydatne” w przypadku rozkładów niekompletnych lub ściętych w jednym z końców; miara ta pozwala również ocenić faktyczne granice wynikowe środkowych 50% obserwacji
- odchylenie standardowe to miara najbardziej rzetelna, używamy jej w sytuacji, gdy pożądane są interpretacje wiążące z krzywą normalną.

4.3. Miary asymetrii i kurtozy

Miary asymetrii służą do oszacowania, czy odchylenia od wartości centralnej grupują się z jednej, czy z drugiej strony rozkładu empirycznego danych. W przypadku nierównomierności rozproszenia, wartości średniej, mediany i dominanty nie pokrywają się, co obrazują poniższe wykresy:



Asymetria informuje nas zatem o tym jak wyniki dla danej zmiennej kształtują się wokół średniej. Czy większość zaobserwowanych wyników jest z lewej strony średniej, blisko wartości średniej czy z prawej strony średniej?

Najprostszym wskaźnikiem asymetrii wskazującym jej kierunek (asymetria lewostronna lub prawostronna) jest wskaźnik asymetrii, który określa różnica pomiędzy średnią a dominantą. Wskaźnik większy od 0 wskazuje na rozkład prawostronny, wynik mniejszy od zera opisuje rozkład lewostronny.

$$W = M - D$$

Wskaźnik ten można obliczyć również w oparciu o kwartyle. Przy rozkładzie symetrycznym różnica pomiędzy kolejnymi rozstępami międzykwartyłowymi równa się zero.

$$(Q_3 - Q_2) - (Q_2 - Q_1) = 0$$

Bardziej przydatne są jednak miary asymetrii, dzięki nim badacz określa nie tylko kierunek ale i siłę asymetrii. Umożliwia to porównanie asymetrii różnych rozkładów. Współczynnik asymetrii obliczany jest na podstawie różnych wzorów, klasycznym sposobem jego wyznaczenia jest podzielenie różnicy średniej i dominanty przez odchylenie standardowe:

$$A = \frac{M - D}{s}$$

Można również wykorzystać wartość momentu centralnego trzeciego rzędu, współczynnik asymetrii jest wówczas stosunkiem tej wartości do sześciastu odchylenia standardowego:

$$A = \frac{M_3}{s^3}$$

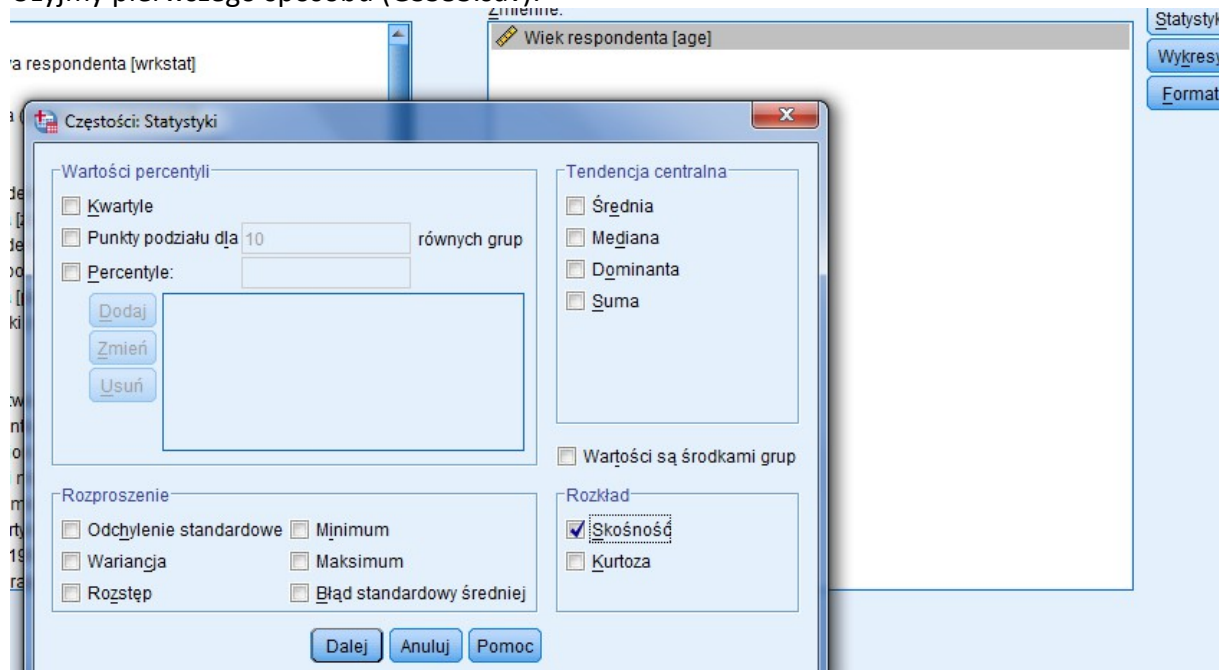
Trzeci moment centralny to suma trzecich potęg odchylen wartości cechy statystycznej od wartości średniej arytmetycznej, podzielona przez liczbę obserwacji.

Wartość miary asymetrii	Interpretacja
0,0 – 0,3	słaba asymetria
0,3 – 0,6	umiarkowana asymetria
0,6 – 1,0	silna asymetria

Obliczenie miar przy użyciu SPSS jest proste i możliwe dzięki kilku procedurom m.in.

- *Analiza – Opis statystyczny – Częstości* (w menu *Statystyki* zaznaczamy *Skośność*)
- *Analiza – Opis statystyczny – Statystyki opisowe* (w menu *Opcje* zaznaczamy *Skośność*)
- *Analiza – Opis statystyczny – Eksploracja*
- *Analiza – Raporty zestawienia – Podsumowania obserwacji* (w menu *Statystyki* zaznaczamy *Skośność*)
- *Analiza-Porównanie średnich – Średnie* (w menu *Opcje* zaznaczamy *Skośność*)

Użyjmy pierwszego sposobu (GSS83.sav):



Statystyki

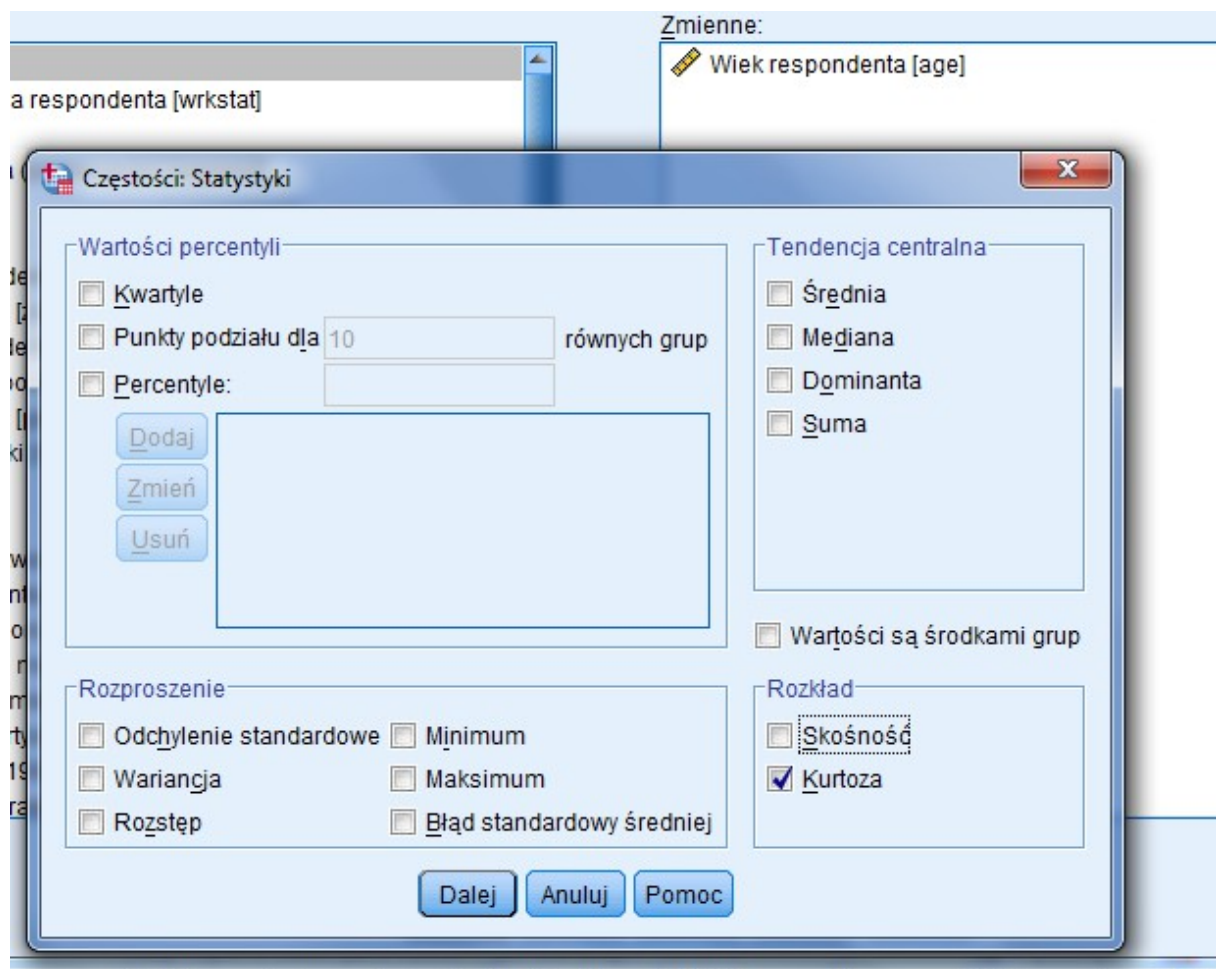
Wiek respondenta

N	Ważne	1495
	Braki danych	5
Skośność		,500
Błąd standardowy skośności		,063

Wynik, który uzyskaliśmy w naszym przykładzie to 0,500. Oznacza on, że rozkład jest umiarkowanie prawostronny.

O ile asymetria to miara pokazując skośność rozkładu to kurtoza wskazuje na jego spłaszczenie.

Aby poznać kurtozę rozkładu należy postępować podobnie jak w przypadku obliczania asymetrii.



Statystyki

Wiek respondenta

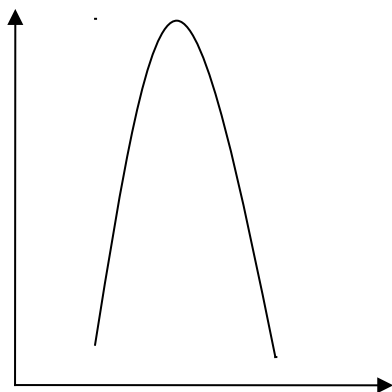
N	Ważne	1495
	Braki danych	5
Kurtoza		-,700
Błąd standardowy kurtozy		,126

Kurtoza ujemna oznacza, że rozkład jest platykurtyczny.

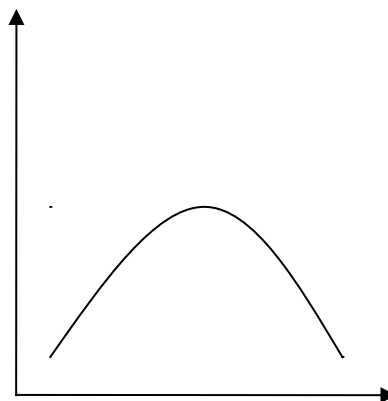
Wartości błędów skośności i kurtozy mają interpretację, jeśli badane obserwacje traktowane są jako próba z populacji. Przyjmuje się, że jeśli błąd standardowy skośności jest większy, bądź równy wartości bezwzględnej współczynnika skośności to w badanej populacji nie występuje asymetria. Z kolei, jeśli wartość błędu standardowego kurtozy jest większa lub równa wartości współczynnika kurtozy, to przyjmuje się, że w badanej populacji badana zmienna ma rozkład mezokurtyczny.

Rozkłady zmiennych można podzielić ze względu na wartość kurtozy na rozkłady:

- mezokurtyczne - wartość kurtozy wynosi 0, spłaszczenie rozkładu jest podobne do spłaszczenia rozkładu normalnego (dla którego kurtoza wynosi dokładnie 0)
- leptokurtyczne - kurtoza jest dodatnia, wartości cechy bardziej skoncentrowane niż przy rozkładzie normalnym (wykres wysmukły)
- platokurtyczne - kurtoza jest ujemna, wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym (wykres spłaszczony).



Rozkład leptokurtyczny



Rozkład platokurtyczny

4.4. Standaryzacja wyników

Standaryzacja wyników pozwala badaczowi na porównywanie zmiennych mierzonych na różnych skalach. Często bywa bowiem tak, że porównanie dwóch lub więcej zmiennych wymaga wyeliminowania wpływu jednostek miary na rozkład zmiennej, przy zachowaniu wzajemnych proporcji wartości. Standaryzacji dokonujemy na podstawie formuły:

$$Z = \frac{x_i - M}{s}$$

Standaryzacja typu Z polega zatem na odjęciu od każdej wartości zmiennej X średniej arytmetycznej i podzieleniu wyniku przez odchylenie standardowe.

Standaryzację zmiennej można wykonać w SPSS za pomocą procedury *Analiza – Opis statystyczny – Statystyki opisowe*. Dla każdej zmiennej umieszczone w oknie *Zmienne*, zostaną obliczone wartości standaryzowane. Wyniki zostaną zapisane w postaci oddzielnej zmiennej w zbiorze danych (nazwa zmiennej poprzedzona jest literą Z).

Zage
-,18525
-,12784
-,18525
-,07043
1,82416
2,11122
,50369
1,65193
-,87420
,44628
-,98902
-1,33349
,84816
,96299
-,75937
5,8744

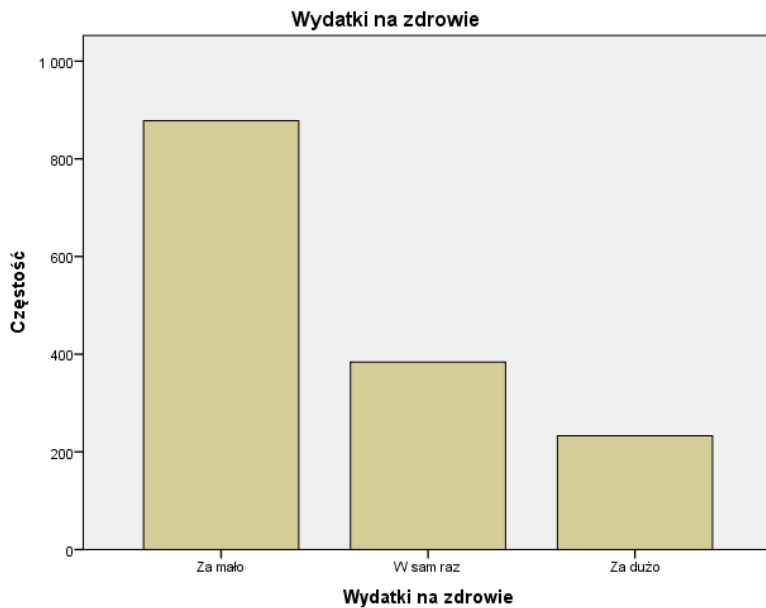
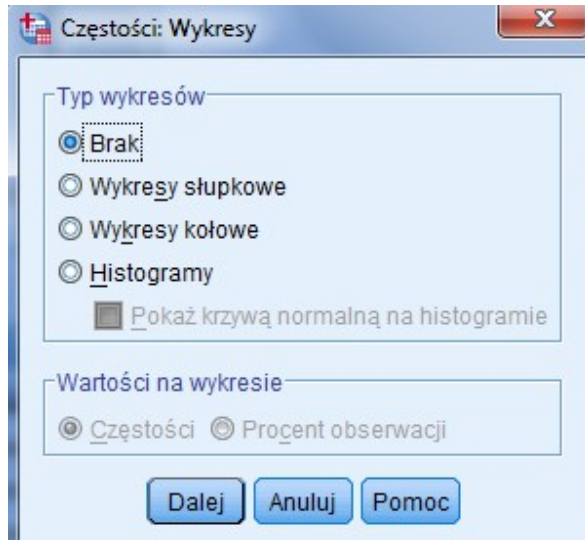
Wzrost	Wiek	Sex	Age	Zage
170	69	M	69	-1,18525
170	70	M	70	-1,12784

GRAFICZNA PREZENTACJA WYNIKÓW

Wykresy generowane przez SPSS pozostawiają wiele do życzenia, są niezbyt estetyczne, stąd warto wykorzystać do tego celu inne programy np. Excel. Efekty będą nieporównywalne.

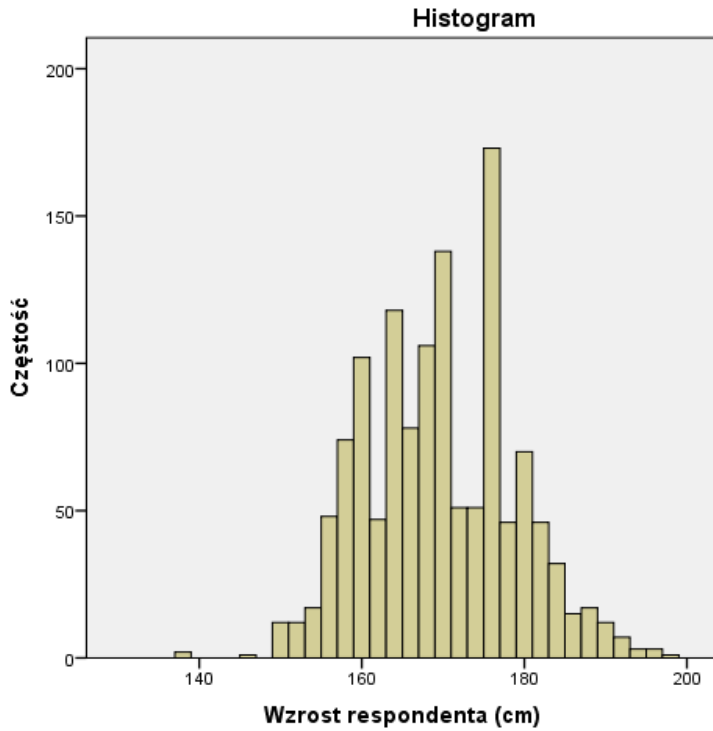
W menu *Analiza – Opis Statystyczny – Częstości – Wykresy* mamy do wyboru 3 rodzaje wykresów: słupkowy, kołowy, i histogram. Do porównania częstości występowania poszczególnych kategorii dobrym wyborem jest wykres słupkowy, idealnie nadaje się do zmiennych skokowych, słupki są oddzielone od siebie. Histogram jest odpowiedni do zmiennych ciągłych, w wykresie tym słupki są ze sobą połączone. Dodatkowo badacz ma możliwość naniesienia na wykres krzywej normalnej. Z kolei wykres kołowy służy do prezentacji części (udziałów, procentu) pewnej całości.

O wyborze wykresu decyduje również poziom pomiaru zmiennej. Wykres słupkowy nadaje się do graficznego przedstawienia cech wyrażonych na skali słabej (nominalnej). Cechy typu ilościowego można prezentować i na wykresie słupkowym i na histogramie¹³.



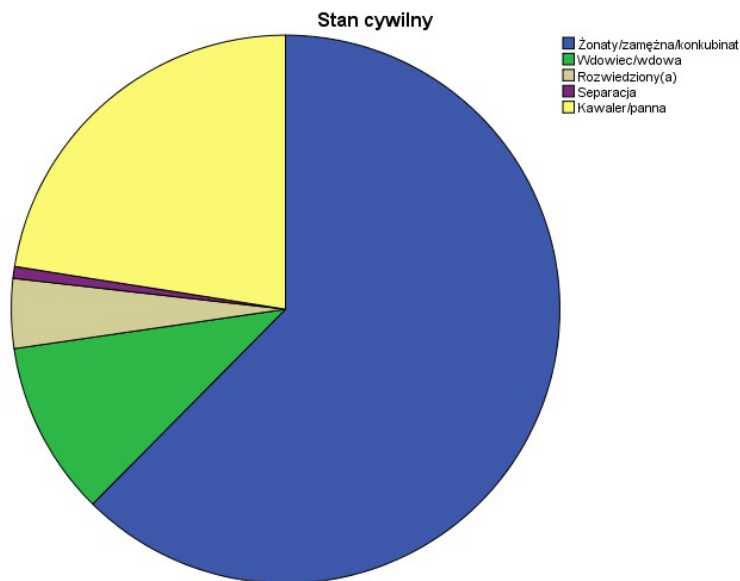
Wykres słupkowy

¹³ Więcej na ten temat: A. Malarska, *Statystyczna analiza danych wspomagana programem SPSS*, Kraków 2005, s. 24-39.



Histogram - obok wykresu podana jest średnia i odchylenie standardowe oraz liczebność próby.

W przypadku wykresu słupkowego i histogramu na osi odciętych znajdują się wartości cechy, na osi rzędnych albo tylko częstości (histogram), albo częstości i procentowo wyrażone wskaźniki struktury (wykres słupkowy).



Wykres kołowy - w przypadku zmiennych o wielu kategoriach traci swoje walory informacyjne.

Wykresy mogą zawierać różne elementy objaśniające: tytuł, opis osi, legendę, przypisy czy komentarze.

SPSS umożliwia nam również kreowanie wykresów wedle potrzeb, przy użyciu opcji Wykresy.