Statystyka opisowa

Wykład 3:
Statystyki opisowe
- miary położenia, miary zmienności, miary asymetrii

Wprowadzenie

W przypadku danych liczbowych do ich charakterystyki można wykorzystać tak zwane *STATYSTYKI OPISOWE*.

Za pomocą statystyk opisowych można w zwięzły sposób scharakteryzować rozkład wartości cechy liczbowej w badanej zbiorowości.

Najważniejsze rodzaje statystyk dotyczą:

- **przeciętnego poziomu cechy (miary położenia);**
- **rozproszenie danych (miary zmienności);**
- **asymetrię rozkładu danych (miary asymetrii).**

Podział statystyk opisowych ze względu na sposób wyznaczania

Miary klasyczne (średnia, odchylenie standardowe i inne) są wyznaczane na podstawie wszystkich obserwacji – w związku z tym są nieodporne na obserwacje odstające.

Miary pozycyjne (minimum, maksimum, mediana, kwartyle, centyle) są wyznaczane na podstawie pozycji zajmowanej przez odpowiednie obserwacje i w związku z tym nie są zależne od ewentualnych obserwacji ekstremalnych.

Miary położenia

Do najczęściej wyznaczanych miar położenia, zawierających informacje o przeciętnym poziomie wartości danych cechy w badanej zbiorowości, należą:

- **średnia arytmetyczna;**
- **inne rodzaje średnich (harmoniczna, geometryczna);**
- **wartość najmniejsza i największa (minimum i maksimum);**
- **wartość środkowa – mediana;**
- **wartość najczęstsza – moda;**
- **kwartyle;**
- **centyle.**

O niemal każdej z tych miar można by przeprowadzić osobny wykład…

Średnia arytmetyczna

Średnia arytmetyczna jest najbardziej popularną miarą przeciętnego poziomu cechy liczbowej. Poniżej opisano, powszechnie znany, sposób wyznaczania średniej dla szczegółowego szeregu statystycznego.

Przy okazji „przemycono" oznaczenia, które mogą się okazać przydatne, przy samodzielnym zgłębianiu podręczników do statystyki.

Wartość cechy (x_i)	15	10	11	9	7	4	15	13	14
	x_1	x_2	x_3	x_4	…	x_{n-3}	x_{n-2}	x_{n-1}	x_n

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{15+10+11+9+7+4+15+13+14}{9} = ?$$

Mediana – wartość środkowa

Alternatywną metodą opisu danych liczbowych jest wartość środkowa (mediana), która określa jednostkę znajdującej się „w środku” badanej zbiorowości, jeśli chodzi o poziom rozważanej cechy liczbowej.

Aby wyznaczyć medianę, szereg szczegółowy należy uporządkować (rosnąco) i wskazać wartość dla środkowego obiektu.

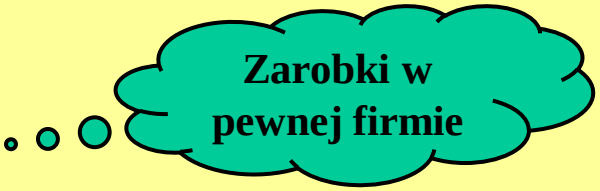
Wartość cechy (x_i)	4	7	9	10	11	13	14	15	15
	x_1	x_2	x_3	x_4	...	x_{n-3}	x_{n-2}	x_{n-1}	x_n

Me

Uproszczona definicja i interpretacja mediany to stwierdzenie, iż 50% pomiarów jest od niej mniejszych, a 50% pomiarów większych.

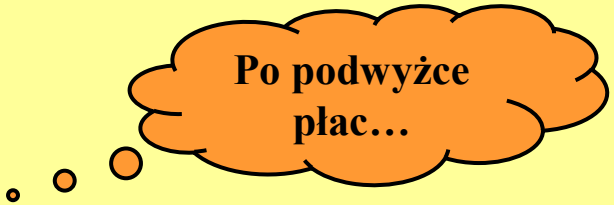
Co jest lepsze – mediana czy średnia?

Mediana i średnia mogą dla jednych danych mieć zbliżone wartości, a dla innych mogą się bardzo wyraźnie różnić. W praktyce zalecamy wyznaczanie obu tych statystyk jednocześnie i wyciąganie wniosków na podstawie ich jednoczesnego oglądu.



3000 zł 3200 zł 3400 zł 3700 zł 32000 zł

Średnia = 9060 zł **Mediana =3400 zł**



3000 zł 3200 zł 3400 zł 3700 zł 62000 zł

Średnia = 14060 zł **Mediana =3400 zł**

Nie zawsze średnie zarobki odzwierciedlają dobrze płace. Jeżeli tylko można dowiedz się także ile wynosi mediana.

Centyle – uogólnienie mediany

Mediana jest wartością, którą znajdujemy jako odpowiedź na pytanie: poniżej (powyżej) jakiej wartości sytuuje się 50% pomiarów.

W wielu sytuacjach analityka interesuje też kwestia *poniżej (powyżej) jakiej wartości znajduje się inna część pomiarów (1%, 5%, 10% czy 25%)*. Stwierdzenia te określają grupę miar zwanych *centylami*.

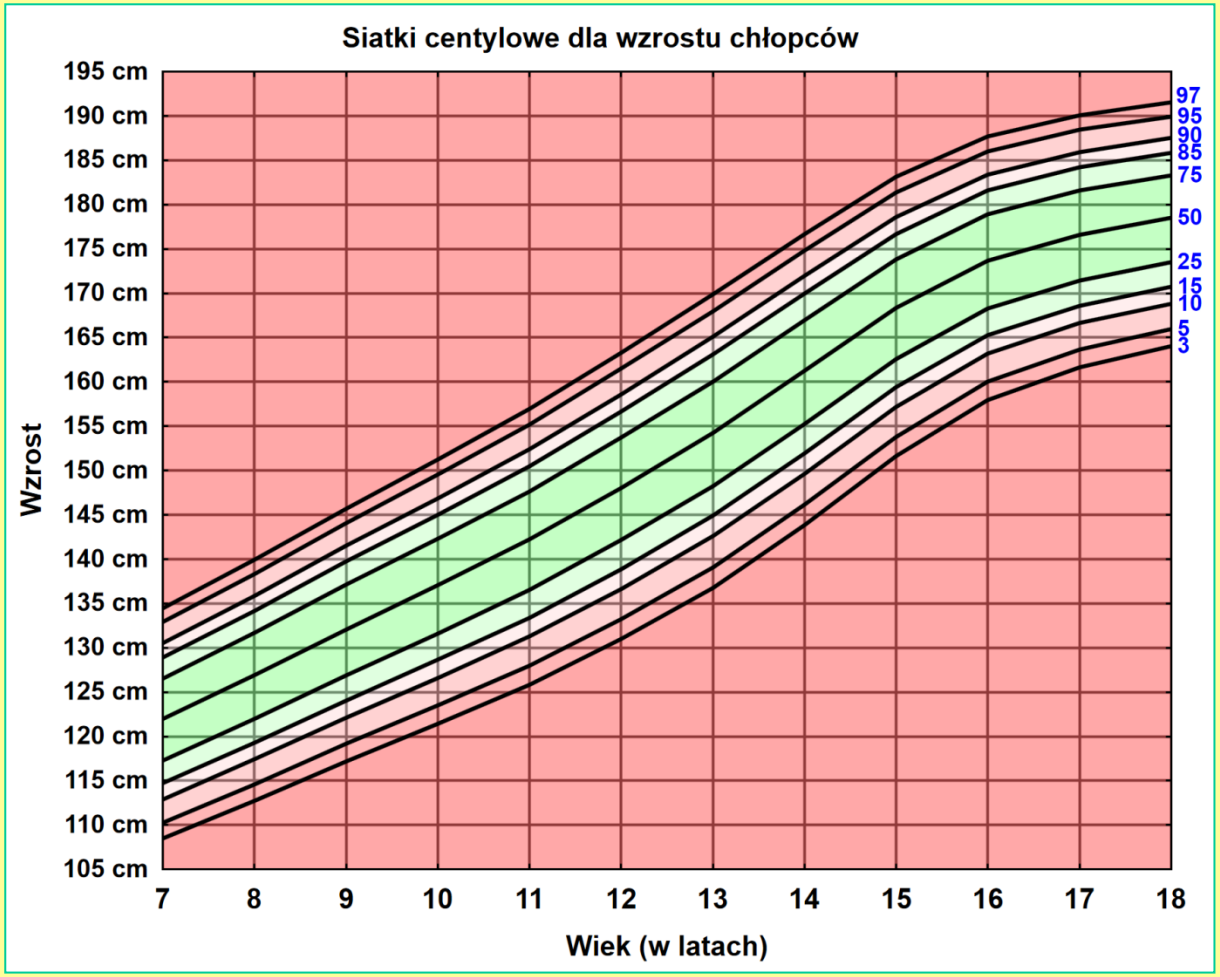
Centyl rzędu p (c_p) ($0 < p < 1$) jest to taka liczba, że poniżej niej znajduje się p-ta część pomiarów zaś powyżej $(1-p)$-ta część pomiarów. Wartość p jest też często podawana w procentach.

Niektóre centyle, z uwagi na popularność zastosowań mają swoje własne nazwy:

- c_{50} **to mediana;**
- c_{25} **to kwartyl dolny (Q_{25}) a** c_{75} **to kwartyl górny (Q_{75});**
- c_{10}, c_{20}, …, c_{90} **to tak zwane decyle (oznaczane też d_1, …, d_9).**

Siatki centylowe – praktyczne zastosowanie

Siatki centylowe znajdują szerokie zastosowanie np. w medycynie czy oświacie. Przedstawianie się w ten sposób rozkład wartości cech somatycznych czy też rozkład wyników matur lub egzaminów.



Statystyki opisowe w programie *STATISTICA*

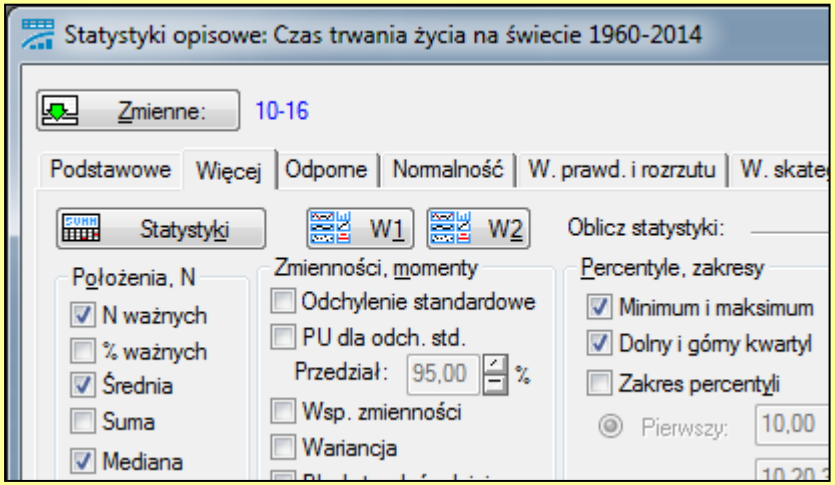
Statystyki opisowe w programie STATISTICA najlepiej wyznaczać za pomocą analizy wywoływanej za pomocą poleceń STATYSTYKA / STATYSTYKI PODSTAWOWE I TABELE / STATYSTYKI OPISOWE.

Po wybraniu zmiennej (lub wielu zmiennych) typu liczbowego, dla których chcemy wyznaczyć wartości statystyk opisowych należy ustalić listę wyliczanych parametrów (zakładka WIĘCEJ).

Przykład (1) – specyfikacja analizy

Analiza dotyczy zbioru danych *Czas trwania życia na świecie 1960-2014*. Celem analizy będzie opis rozkładu czasu trwania życia mężczyzn na świecie w przekroju międzynarodowym w latach objętych analizą.

W oknie analizy *STATYSTYKI OPISOWE* wybieramy zmienne o numerach 10-16, a w zakładce *WIĘCEJ* ustalamy zakres statystyk opisowych do wyliczenia (według poniższego rysunku).



Przykład (2) – wyniki

W tabeli wynikowej dokonujemy formatowania wyników (najlepiej ustalając taki format jak dla danych wejściowych), a następnie ich interpretacji.

Zmienna	Statystyki opisowe (Czas trwania życia na świecie 1960-2014)						
	Nważnych	Średnia	Mediana	Minimum	Maksimum	Dolny Kwartyl.	Górny Kwartyl.
Oczekiwany czas trwania życia mężczyzn 1960	187	52,0	52,5	27,4	71,5	42,0	62,3
życia mężczyzn 1970	189	56,0	58,4	31,5	72,2	47,8	65,6
życia mężczyzn 1980	190	59,4	61,5	25,1	73,6	52,4	67,6
życia mężczyzn 1990	193	62,3	64,9	31,0	75,9	56,9	69,1
życia mężczyzn 2000	198	64,6	67,0	37,7	78,0	58,5	71,9
życia mężczyzn 2010	197	67,9	69,6	46,1	80,3	62,3	74,4
Oczekiwany czas trwania życia mężczyzn 2014	196	69,0	70,6	48,8	81,2	63,7	75,2

Liczba analizowanych państw wynosiła od 187 do 198

Na podstawie wartości średniej stwierdzamy, iż oczekiwany czas trwania życia mężczyzn w państwach świata wykazuje tendencję rosnącą – średnia tego wskaźnika wzrosła z 52 lat w roku 1960 do 69 lat w roku 2014

Porównując wartość średniej i mediany stwierdzamy, iż rozkład oczekiwanego czasu trwania życia mężczyzn jest nierównomierny – średnia jest mniejsza od mediany, co oznacza, że są na świecie kraje odstające *in minus* pod tym względem. Jest to tak zwana asymetria lewostronna. Różnica między średnią i medianą uległa pewnemu zmniejszeniu w ostatnich latach.

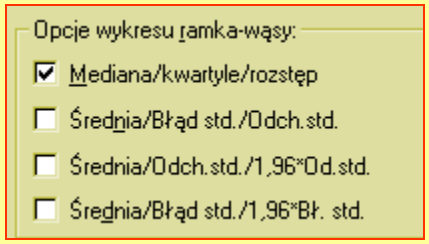
W 1960 w co czwartym państwie świata oczekiwany czas trwania życia mężczyzn nie przekraczał 42 lat, w 2014 roku próg ten wzrósł do poziomu 63,7 lat. Zmniejszyła się różnica między dolnym i górnym kwartylem – z ponad 20 lat w 1964 do niespełna 12 lat w 2014 roku.

W 1960 roku oczekiwany czas trwania życia mężczyzn w państwach świata wahał się od 27,4 lat (Mali) do 71,5 lat (Holandia). W 2014 roku wartość minimalna wzrosła o ponad 21 lat i wynosiła 48,8 (Republika Środkowoafrykańska), zaś wartość maksymalna wzrosła o ok. 10 lat i wynosiła 81,2 (Hong-Kong).

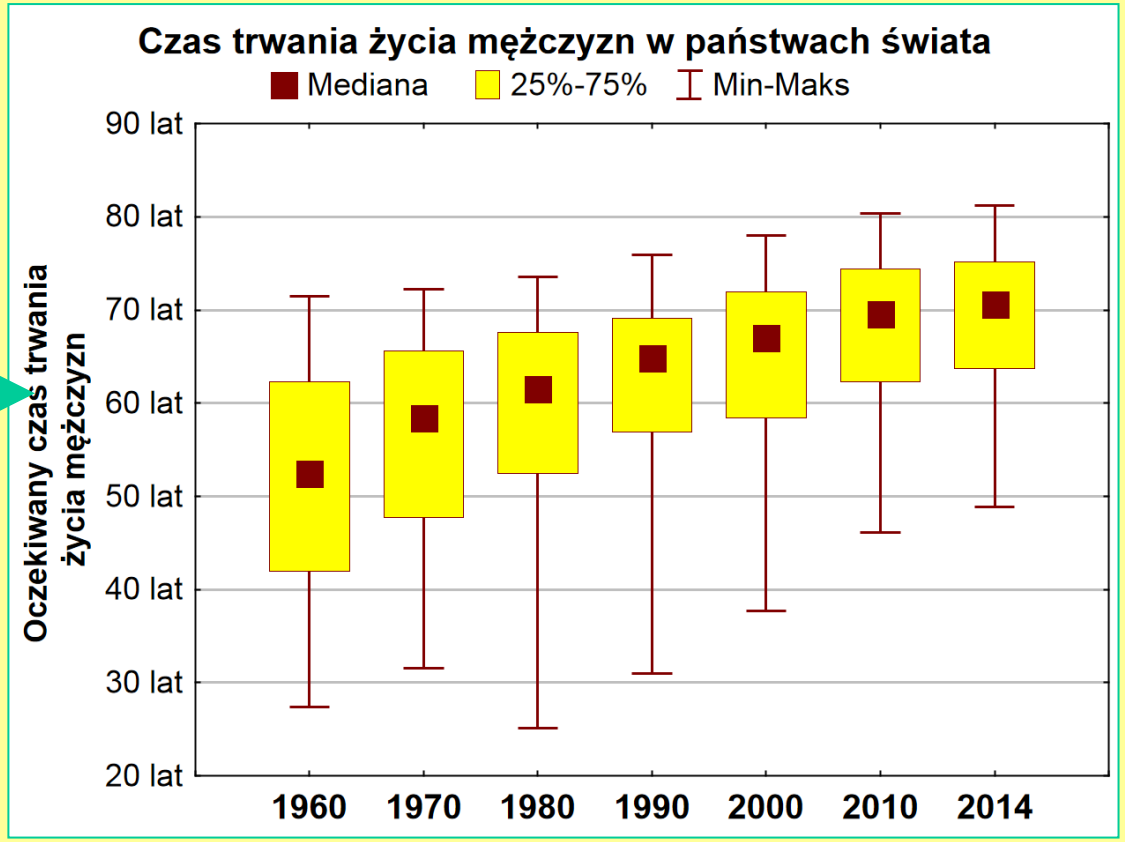
Ilustracja graficzna – wykres ramka-wąsy

Wartości statystyk opisowych można zilustrować za pomocą wykresu typu ramka-wąsy. Wykres ten w podstawowej formie można znaleźć w oknie analiz *STATYSTYKI OPISOWE*.

W zakładce *opcje* ustalamy typ wykresu ramka-wąsy:



W zakładce *PODSTAWOWE* wywołujemy wykres, który po sformatowaniu wygląda tak…



Miary zmienności

W wielu sytuacjach wyznaczenie samych miar położenia nie pozwala w wyczerpujący sposób opisać rozkładu wartości cechy liczbowej. W takiej sytuacji można dodatkowo wyznaczyć miary zmienności.

Miary zmienności pozwalają ocenić nie tylko przeciętny poziom danej cechy lecz także ich rozproszenie wokół wartości przeciętnej. Do najbardziej popularnych miar zmienności należą:

- **wariancja i odchylenie standardowe;**
- **współczynnik zmienności;**
- **rozstęp;**
- **rozstęp kwartylowy.**

Odchylenie standardowe

Odchylenie standardowe jest wyliczane jako przeciętne odchylenie pomiarów od wartości średniej. Poniżej opisano szczegółowo procedurę wyznaczania odchylenia standardowego.

Wartość cechy (x_i)	**15**	**12**	**12**	**13**	**8**
	x_1	x_2	x_3	x_4	x_5

$\longrightarrow \quad \bar{x} = 12$

Odchylenia od średniej	**3**	**0**	**0**	**1**	**-4**

Suma odchyleń od średniej zawsze wynosi 0

Kwadraty odchyleń od średniej	**9**	**0**	**0**	**1**	**16**

Średnie kwadratowe odchylenie od średniej nazywane jest wariancją (s^2) a jej pierwiastek odchyleniem standardowym (s).

$$s^2 = \frac{9+0+0+1+16}{5} = 5$$

$$s = \sqrt{\frac{9+0+0+1+16}{5}} = \sqrt{5} \approx 2{,}24$$

Właściwości odchylenia standardowego

Znajomość odchylenia standardowego i wartości średniej pozwala oszacować położenie większości pomiarów. Dla bardzo wielu danych (co wynika z odpowiednich twierdzeń matematycznych) są bowiem spełnione relacje.

Przedział $(\bar{x}-s, \bar{x}+s)$ **zwany typowym przedziałem zmienności zawiera zwykle ok. 68% pomiarów.**

Przedział $(\bar{x}-2s, \bar{x}+2s)$ **zwany rozszerzonym przedziałem zmienności zawiera zwykle ok. 95% pomiarów.**

Przedział $(\bar{x}-3s, \bar{x}+3s)$ **zawiera zwykle ok. 99,7% pomiarów, czyli niemal wszystkie wartości. Pomiary wykraczające poza ten zakres są często określane mianem obserwacji odstających (nietypowych) i niejednokrotnie eliminuje się je z analiz, gdyż mogą zaburzać badane relacje.**

Powyższe stwierdzenia są prawdziwe, gdy dane rozkładają się w sposób symetryczny (lub doń zbliżony) wokół wartości średniej. Dla tzw. rozkładów asymetrycznych, liczba obserwacji zawierających się w podanych wyżej przedziałach może być radykalnie inna.

Współczynnik zmienności

W przypadku porównywania zmienności wielkości wyrażonych w różnych jednostkach (na przykład dochody mieszkańców różnych państw) albo charakteryzujących się różnymi poziomami wartości średniej, konieczne jest wyznaczenie względnego poziomu zmienności.

W tym celu wyznacza się tzw. współczynnik zmienności *(V)*.

$$V = \frac{s}{\bar{x}} \cdot 100\%$$

Rozstęp kwartylowy

Kilka slajdów wcześniej pokazano jak na wartość średnią wpływa nawet jedna obserwacja nietypowa (odstająca). Również odchylenie standardowe, w przypadku występowania obserwacji nietypowych może przybierać bardzo duże wartości, a zakres typowego przedziału zmienności pozbawiony będzie sensu.

W takiej sytuacji wyznaczać można tzw. rozstęp kwartylowy, który definiowany jest jako różnica między kwartylem górnym i dolnym.

$$R_Q = Q_{75} - Q_{25}$$

Inne miary zmienności

Bardzo elementarną miarą zmienności, która jednakże bywa niejednokrotnie używana do opisu danych jest rozstęp, określany jako różnica pomiędzy wartością maksymalną i minimalną.

$$R = x_{\max} - x_{\min}$$

Inne miary zmienności (na przykład służące do badania zróżnicowania dochodów) są opierane na stosunku wybranych centyli.

c_{99} / c_1 — **Relacja zarobków 1% najbogatszych i 1% najbiedniejszych członków danego społeczeństwa**

$x_{\max} / x_{\min}$ — **Poziom zarobków w „najbogatszym" mieście wojewódzkim w Polsce do zarobków w mieście „najbiedniejszym"**

Miary zmienności w programie *STATISTICA*

Statystyki opisowe w programie STATISTICA najlepiej wyznaczać za pomocą analizy wywoływanej za pomocą poleceń STATYSTYKA / STATYSTYKI PODSTAWOWE I TABELE / STATYSTYKI OPISOWE.

Po wybraniu zmiennej (lub wielu zmiennych) typu liczbowego, dla których chcemy wyznaczyć wartości statystyk opisowych należy ustalić listę wyliczanych parametrów (zakładka WIĘCEJ).

Przykład (3)

Kontynuujemy analizy dotyczące czasu trwania życia mężczyzn w państwach świata.

Wznawiamy otwartą lub rozpoczynamy nową analizę *STATYSTYKI OPISOWE* wybieramy te same zmienne, co we wcześniejszej części przykładu i w zakładce *WIĘCEJ* wybieramy następujące statystyki.

- **średnią;**
- **medianę;**
- **minimum i maksimum;**
- **odchylenie standardowe;**
- **współczynnik zmienności;**
- **rozstęp kwartylowy.**

Przykład (4) – wyniki c.d.

Po wywołaniu wyników i ich wstępnym sformatowaniu…

MIARY POŁOŻENIA **MIARY ZMIENNOŚCI**

Zmienna	Średnia	Mediana	Minimum	Maksimum	Kwartyl. Rozstęp	Odch.std	Wsp.zmn.
Oczekiwany czas trwania życia mężczyzn 1960	52,0	52,5	27,4	71,5	20,3	11,7	22,5%
Oczekiwany czas trwania życia mężczyzn 1970	56,0	58,4	31,5	72,2	17,8	10,7	19,1%
Oczekiwany czas trwania życia mężczyzn 1980	59,4	61,5	25,1	73,6	15,2	10,0	16,8%
Oczekiwany czas trwania życia mężczyzn 1990	62,3	64,9	31,0	75,9	12,2	9,4	15,1%
Oczekiwany czas trwania życia mężczyzn 2000	64,6	67,0	37,7	78,0	13,4	9,6	14,9%
Oczekiwany czas trwania życia mężczyzn 2010	67,9	69,6	46,1	80,3	12,1	8,5	12,5%
Oczekiwany czas trwania życia mężczyzn 2014	69,0	70,6	48,8	81,2	11,5	8,1	11,7%

Na podstawie wartości średniej i odchylenia standardowego można wyznaczyć typowy przedział zmienności, czyli zakres, w którym znajduje się ok. 70% państw:

- **1964: (52,0-11,7; 52,0+11,7) czyli (40,3; 63,7)**
- **2014: (69,0-8,1; 69,0+8,1) czyli (60,9; 77,1)**

Analiza względnej miary zmienności, jaką jest współczynnik zmienności pozwala stwierdzić, że zróżnicowanie w czasie trwania życia mężczyzn w krajach świata się zmniejsza:

- **1964: V = 22,5%**
- **2014: V = 11,7%**

Miary asymetrii rozkładu danych

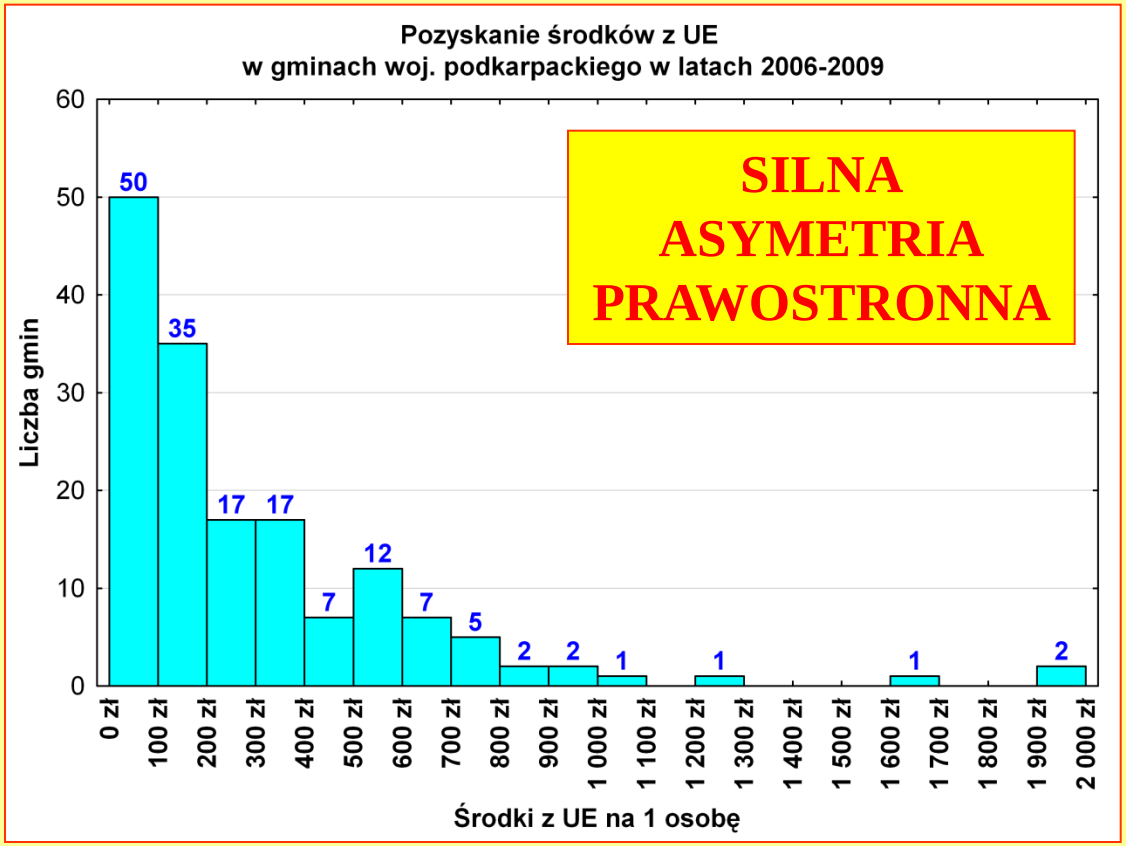
W analizie statystycznej istnieją pewne procedury, w których wymagane jest aby dane miały określony typ rozkładu (lub przynajmniej były doń zbliżone).

Na przykład wyznaczanie omówionego wcześniej typowego przedziału zmienności traci sens dla danych wykazujących bardzo dużą asymetrię. Dlatego też wskazana jest umiejętność oceny poziomu asymetrii za pomocą odpowiedniego współczynnika.

Informacja o rodzaju asymetrii jest też interesująca sama w sobie – pozwala lepiej zrozumieć zjawisko opisywane za pomocą cechy liczbowej.

Asymetria prawostronna cechuje na przykład rozkłady zarobków czy dochodów i jest wykorzystywana jako miara nierównomierności rozkładu dochodów w społeczeństwie.

Przykład asymetrii prawostronnej

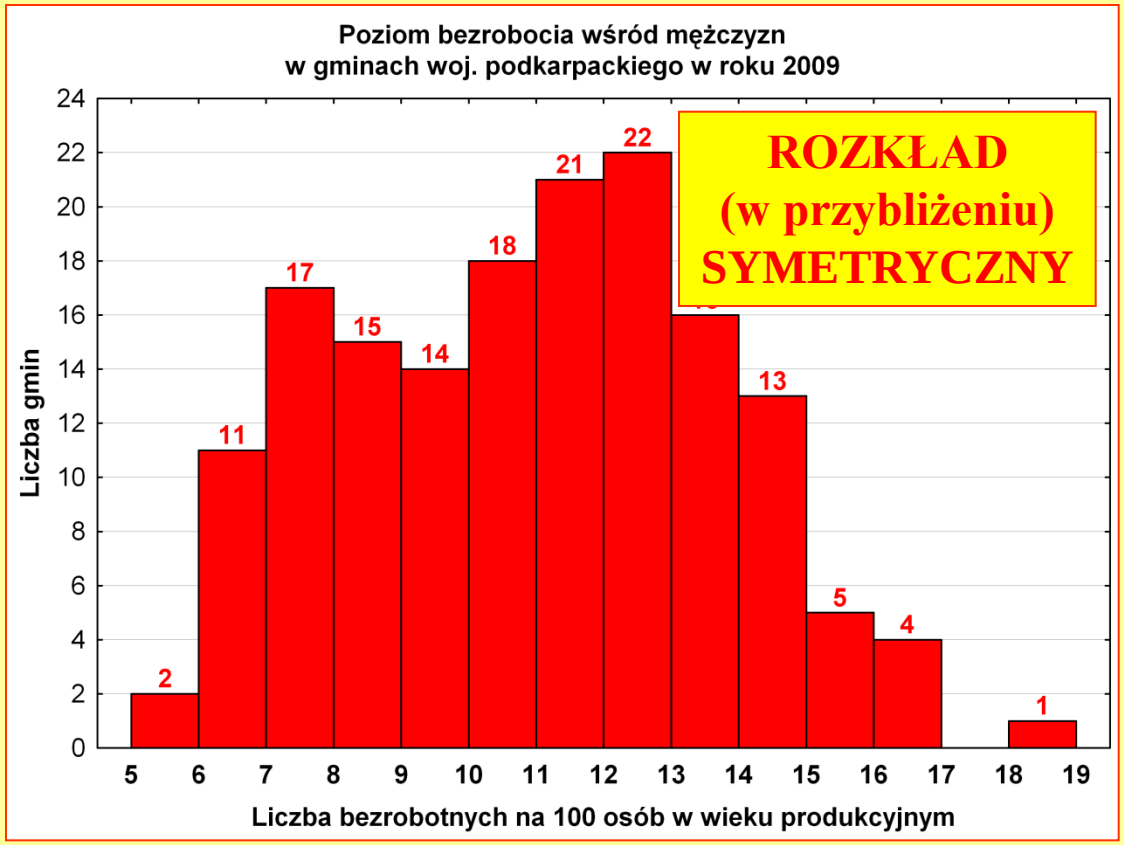


Średnia = 299 zł
Mediana = 181 zł
Skośność = 2,46

Miara asymetrii nazwana jest w programie *STATISTICA* skośnością i można ją wyznaczyć za pomocą analizy *STATYSTYKI OPISOWE*.

Rozkład wykorzystania środków z UE w gminach woj. podkarpackiego charakteryzuje się bardzo silną asymetrię prawostronną (jest „wydłużony” w prawą stronę). W praktyce oznacza to, że występują pojedyncze wartości wysokie i bardzo wysokie, nieliczne wartości na poziomie średnim i zdecydowana większość wartości na poziomie niskim i bardzo niskim (w większości gmin pozyskano niewiele środków z UE)

Przykład rozkładu symetrycznego

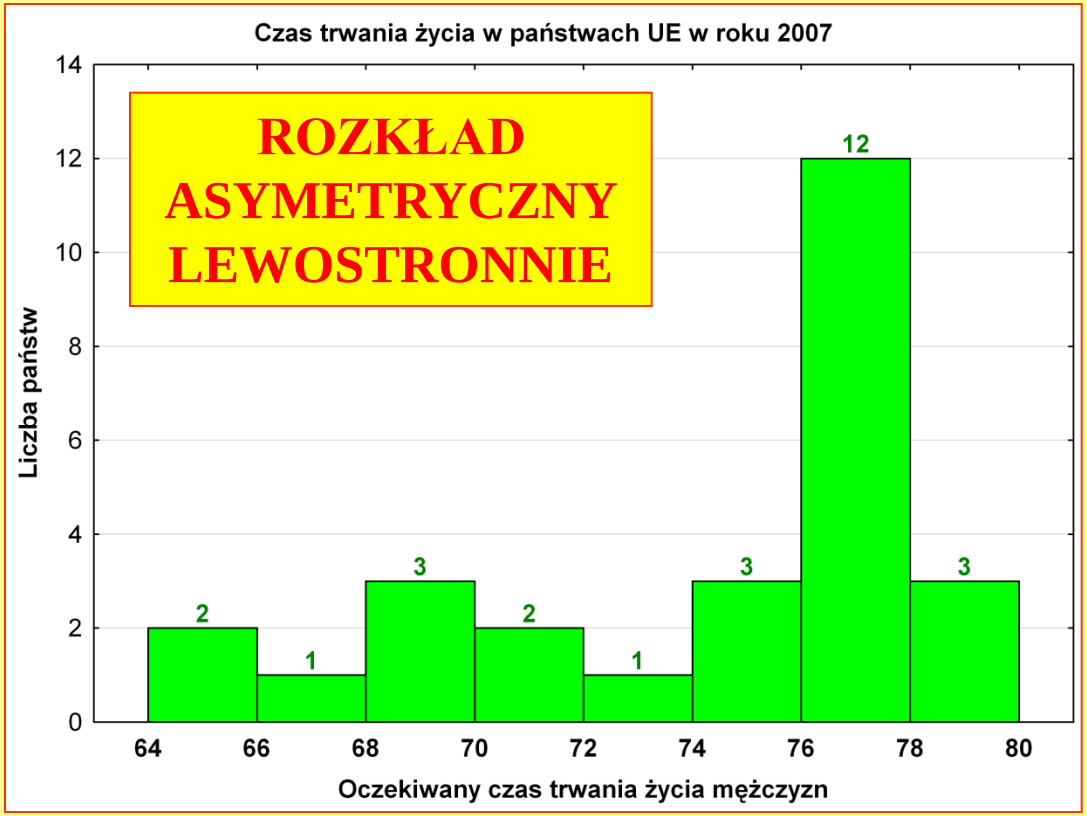


Średnia = 11,0
Mediana = 11,1
Skośność = 0,08

Miara asymetrii nazwana jest w programie *STATISTICA* skośnością i można ją wyznaczyć za pomocą analizy *STATYSTYKI OPISOWE*.

Rozkład wskaźnika bezrobocia wśród mężczyzn w gminach woj. podkarpackiego jest bardzo zbliżony do symetrycznego. Średni wskaźnik bezrobocia i wartość środkowa są niemal identyczne. Podobna liczba gmin charakteryzuje się wysokim i niskim bezrobociem.

Przykład asymetrii lewostronnej



Średnia = 74,5
Mediana = 76,7
Skośność = -1,02

Miara asymetrii nazwana jest w programie *STATISTICA* skośnością i można ją wyznaczyć za pomocą analizy *STATYSTYKI OPISOWE*.

Rozkład oczekiwanego czasu trwania życia mężczyzn z państwach UE w 2007 roku charakteryzował się asymetrią lewostronną – w większości państw wskaźnik ten jest na wysokim bądź bardzo wysokim poziomie a w nielicznych był na poziomie średnim bądź niskim.

Interpretacja wskaźnika skośności

A ≈ 0

Współczynnik skośności zbliżony do 0 pozwala stwierdzić, iż mamy do czynienia z symetrycznym rozkładem danych. Wtedy średnia i wartość środkowa (mediana) są do siebie zbliżone i można je stosować zamiennie.

$$\bar{x} \approx Me$$

A > 0

Współczynnik skośności większy od 0 oznacza asymetrię prawostronną. O silnej asymetrii prawostronnej będziemy mówić, gdy $A > 1$. Wartość średnia jest wyższa niż mediana.

$$\bar{x} > Me$$

A < 0

Współczynnik skośności mniejszy od 0 oznacza asymetrię lewostronną. O silnej asymetrii lewostronnej będziemy mówić, gdy $A < -1$. Wartość średnia jest niższa niż mediana.

$$\bar{x} < Me$$

Formatowanie statystyk opisowych

Na zajęciach (i EGZAMINIE) będziemy stosować następujące zasady formatowania statystyk opisowych:

- **miary położenia i bezwzględne miary zmienności (odchylenie standardowe, rozstępy) wyrażone są w tych samych jednostkach co dane wejściowe i można je formatować z taką samą dokładnością (ewentualnie dodając jedno miejsce po przecinku);**
- **współczynnik zmienności podajemy w procentach, z dokładnością do jednego miejsca po przecinku;**
- **współczynnik skośności (asymetrii) podajemy z dokładnością do dwóch miejsc po przecinku.**