Wprowadzenie do zagadnień statystycznych

Jednym z podstawowych celów nauki jest wyjaśnianie i przewidywanie wyników obserwacji zdarzeń i relacji przyczynowych, jakie między nimi zachodzą. Pomocna w tych zagadnieniach jest statystyka. Na podstawie znajomości cech odpowiednio wybranej części elementów (próby) pewnej zbiorowości będziemy wysnuwali wnioski dotyczące rozważanych cech dla pozostałych, nieznanych elementów tej zbiorowości (populacji generalnej). Wykorzystamy w tym celu główne działy statystyki matematycznej:

- testowanie hipotez statystycznych
- szacowanie (estymacja) parametrów lub funkcji

Pierwszym krokiem, jaki musimy uczynić, jest zdefiniowanie populacji, na temat której chcemy formułować sądy. Należy wiec powiedzieć, co rozumiemy przez pojęcia **populacja** i **próba**. **POPULACJA** jest to zbiór wszystkich elementów, które podlegają badaniu z punktu widzenia różnych kryteriów badawczych, a **PRÓBA** jest podzbiorem wylosowanym z populacji w celu wnioskowania o zbiorze. Pamiętajmy, aby próba była reprezentatywna dla całej populacji. Możemy wówczas uogólnić wyniki badania przeprowadzone na próbie do wszystkich elementów populacji, które nie były badane. Ważna jest również odpowiednia liczebność próby zapewniająca odpowiednią moc przeprowadzanych analiz.

Cechy, którymi wyróżniają się jednostki wchodzące w skład badanej zbiorowości, nazywamy cechami statystycznymi. Zbiorowość ma dużo cech. Jednak do konkretnego badania wybieramy tylko te najważniejsze dla analizowanego problemu. Po wytypowaniu cech, które nas interesują musimy podjąć decyzję jak będziemy mierzyć wartości tych cech w trakcie obserwacji. Stevens w 1951 roku zaproponował następujące cztery skale pomiarowe:

- skala nominalna
- skala porządkowa
- skala interwałowa
- skala ilorazowa

SKALA NOMINALNA to jakby opis wyniku eksperymentu (badania) w terminach zdarzeń. Pozwala nam tylko na pogrupowanie badanych obiektów w rozłączne klasy względem kategorii skali. Np. płeć: samica, samiec; zamieszkanie: miasto, mała miejscowość lub wieś. Nie są możliwe działania arytmetyczne na danych opisanych na tej skali.

SKALA PORZĄDKOWA pozwala uporządkować obiekty wg wartości badanej cechy. Np. skala natężenia choroby. Podobnie jak dla skali nominalnej, nie są możliwe działania arytmetyczne na danych opisanych na skali porządkowej

SKALE INTERWAŁOWE pozwalają stwierdzić o ile jednostek natężenie (wielkość) badanej cechy dla obiektu A jest większe (mniejsze) od natężenia (wielkości) tejże cechy dla obiektu B. Np. temperatura ciała A równa 40°C jest większa o 10°C od temperatury ciała B równej 30°C. Działania arytmetyczne – dodawanie i odejmowanie są możliwe.

SKALE ILORAZOWE pozwalają ponadto na stwierdzenie krotności obserwowanej różnicy natężenia (wielkości) badanej cechy, np. wiek, waga oraz parametry biochemiczne itd.

Zestawienie własności skal pokazuje poniższa tabela

	Skala nominalna	Skala porządkowa	Skala przedziałowa	Skala ilorazowa
Czy obiekt X jest różny od obiektu Y	**TAK**	**TAK**	**TAK**	**TAK**
Czy obiekt X jest lepszy od obiektu Y	**NIE**	**TAK**	**TAK**	**TAK**
O ile obiekt X jest lepszy od obiektu Y	**NIE**	**NIE**	**TAK**	**TAK**
Ile razy obiekt X jest lepszy od obiektu Y	**NIE**	**NIE**	**NIE**	**TAK**

W konkretnym badaniu badacz musi określić, jakiego rodzaju skali będzie używał dla badanej cechy. Konsekwencją wyboru określonej skali jest stosowalność odpowiednich narzędzi statystycznych do analizy zgromadzonych danych.

W trakcie badania statystycznego powinniśmy uwzględnić następujące etapy:

I. Etap wstępny – określenie celu badania, rodzaju zbiorowości statystycznej i cech statystycznych, które będą badane.

II. Etap zbierania danych – obserwacja jednostek, wyniki badań lub też wywiad z właścicielem zwierzęcia.

III. Opracowanie surowego materiału statystycznego – kontrola zebranych danych, wyliczenie statystyk opisowych oraz prezentacja graficzna opracowanego materiału.

IV. Analiza opracowanego materiału – estymacja punktowa i przedziałowa, weryfikacja hipotez, a także meta-analiza.

W tym kursie traktujemy statystykę jako dyscyplinę zajmującą się metodami zbierania, opracowywania i analizy danych.

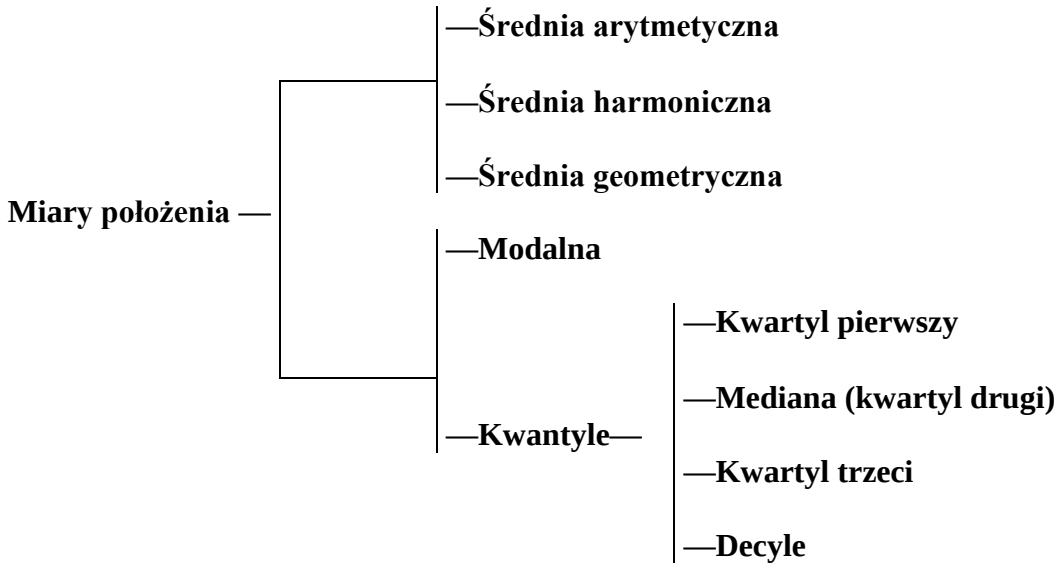
STATYSTYKI OPISOWE

Po zakończeniu etapu obserwacji tzn. zebraniu wszystkich interesujących wyników otrzymujemy „surowy" materiał statystyczny (duża liczba sprawozdań, indywidualnych danych itd.). Zebrany materiał powinien być usystematyzowany i odpowiednio opisany statystycznie. Oczywiście możemy oceniać „na oko" różnice pomiędzy rozkładami czy grupami wyników, ale zdecydowanie lepszym pomysłem jest posługiwanie pewnymi wielkościami, które służą nam do opisu charakterystyki rozkładu czy grupy wyników. **Opis statystyczny** to obliczenie pewnych charakterystyk liczbowych (**zwanych statystykami**) badanych cech. Stanowi on punkt wyjścia do wnioskowania w przypadku działania na grupie losowej.
Statystyki tak charakteryzują zbiorowość, że porównywanie różnych zbiorowości statystycznych można sprowadzić do ich porównań. Podstawowe zadania tych statystyk opisowych to:

- Określenie przeciętnego rozmiaru i rozmieszczenia wartości zmiennej. Dokonujemy tego przez obliczenie miar położenia.
- Określenie granic obszaru zmienności wartości zmiennej. Dokonujemy tego przez obliczenie miar zmienności.

- Określenie skupienia i spłaszczenia (w stosunku do kształtu krzywej normalnej) oraz stopnia zmiany od idealnej symetrii. Dokonujemy tego przez obliczenie miar asymetrii i koncentracji.

Od opisania miar położenia zaczniemy dokładniejsze poznanie rodzajów i sposobów obliczania statystyk opisowych. Schematyczny podział miar położenia przedstawia poniższy rysunek. Ich nazwa wywodzi się stąd, że wskazują miejsce wartości najlepiej reprezentującej wszystkie wielkości zmiennej. Miary przeciętne charakteryzują średni lub typowy poziom wartości zmiennej (cechy) czyli mówią o przeciętnym poziomie rozważanej cechy. Abstrahując od indywidualnych różnic pojedynczych jednostek, miary te podają za pomocą jednej liczby charakterystykę poziomu wartości zmiennej, czyli tzw. centralną tendencję. Dlatego miary te bywają też nazywane miarami tendencji centralnej. Na początku omówimy najczęściej stosowane miary przeciętne.



Rys. 1 Miary położenia

Średnie

Średnią arytmetyczną znamy oczywiście wszyscy. Średnia arytmetyczna jest najlepszą miarą charakteryzującą rozkład cechy i dlatego jest miarą najczęściej używaną. Obliczanie jej opiera się na wszystkich obserwacjach i ma ogromne znaczenie teoretyczne i praktyczne.

Przykład

Niech x_i będzie masę ciała 8 zwierząt (w kg) leczonych w pewnej klinice

Pacjent	1	2	3	4	5	6	7	8
Masa ciała	49	65	80	48	56	74	90	85

Średnia masa ciała leczonych pacjentów jest równa: $\overline{X} = \frac{49+65+\mathrm{K}+85}{8} = \frac{547}{8} = 68{,}38\,\mathrm{kg}.$

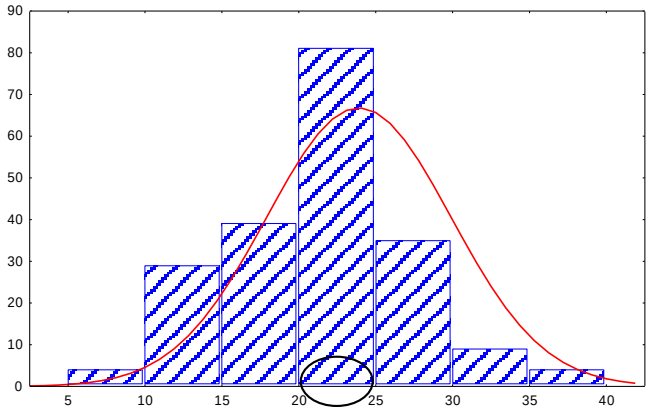
Używając analogii fizycznej średnią możemy uważać za środek ciężkości lub „punkt równowagi”. Otóż gdybyśmy wyniki odmierzyli na jakimś pręcie i w każdym punkcie odpowiadającym wynikowi zawiesili takie same odważniki, to średnia okazałaby się punktem, w którym należałoby pręt podeprzeć żeby zachować równowagę Taką interpretację zauważył belgijski astronom Lambret Quetelet nazywany ojcem statystyki. Przed nim statystyka to tylko staranne i systematyczne zapisywanie narodzin, zgonów i innych

obserwacji interesujących tylko urzędników. Quetelet pierwszy zauważył, że te monotonne liczby, są źródłem cennych informacji, gdy zinterpretujemy je zgodnie z prawami rachunku prawdopodobieństwa.
Oprócz średniej arytmetycznej można również wyróżnić inne rodzaje klasycznych miar tendencji centralnej w tym m. in. średnią geometryczną i średnią harmoniczną. Średnią geometryczną stosujemy, gdy zjawiska są ujmowane dynamicznie (np. średnie tempo zmian). Średnie te są rzadziej wykorzystywane w problemach statystycznych w biologii i medycynie.

Jedyną poważniejszą wadą średniej arytmetycznej jest to, że duży wpływ na nią wywierają najmniejsza i największa wartość badanego szeregu, czyli tzw. skrajne wartości cechy (niejednokrotnie przypadkowo włączone do próby). Przykładowo w pewnej firmie usługowej zarobki ośmiu zatrudnionych pracowników osiągnęły 500 zł miesięcznie. Księgowa i kierownik zespołu otrzymali po 2 000 zł, a właściciel wypłacił sobie 10 000 zł. Średnia zarobków w tej firmie wygląda zachęcająco (około 1455 zł). Jakże myląca i niepełna jest ta informacja. Sytuację tę lepiej opisują inne (poniżej opisane) miary.

Modalna

Modalna zwana również dominantą (moda, wartość najczęstsza) jest to wartość cechy statystycznej, która w rozkładzie empirycznym występuje najczęściej. Oznaczana jest symbolem Mo. Dla wspomnianych powyżej danych modalna wynosi 500 zł i lepiej odzwierciedla sytuację w firmie usługowej. Słowo „Moda” kojarzy się nam ze słowem „modna”. I to bardzo dobrze, bo słowo to dobrze określa, czym jest statystyczna modalna. Jest to po prostu wartość najbardziej „popularna” - występująca najczęściej. W szeregach szczegółowych i rozdzielczych jest to wartość cechy, której odpowiada największa liczebność. Znalezienie klasy o największej liczebności nie jest sprawą trudną, wyróżnia ją jeden wyraźny punkt (szczyt) reprezentujący największą liczbę obserwacji. Przykładowo na poniższym histogramie kółkiem zaznaczono przedział modalnej.



Rys. 2 Histogram z zaznaczonym przedziałem modalnej (oś x - czas działania leku)

Modalna taka prosta miara ma jednak swoje słabe strony. Jest niestety miarą niestabilną. Weźmiemy następujące liczby 2, 2, 4, 6, 8, 20. Modalna wynosi tu 2. Wystarczy jednak zmienić 2 na 20 i modalna przyjmuje nową wartość 20 („przeskoczy” na drugi koniec skali). Widać więc, że zmiana tylko jednej liczby może całkowicie zmienić wartość modalnej. Jest to duża wada w porównaniu z medianą. Dla niej bowiem zmiany liczb mogą mieć miejsce a nie wpływają na ich wartość tak dramatycznie.

Kwantyle

Kwantyle to wartości cechy badanej zbiorowości (przedstawionej w postaci szeregu statystycznego), które dzielą zbiorowość na określone części pod względem liczby jednostek. Części te pozostają do siebie w określonych proporcjach. Do najczęściej stosowanych kwantyli należą kwartyle (podział na 4 części) i decyle (podział na 10 części) oraz percentyle (podział na 100 części).

Omówimy najważniejsze kwartyle (wartości ćwiartkowe).

- Kwartyl pierwszy Q1 jest to wartość jednostki, która dzieli zbiorowość w ten sposób, że 1/4 (25%) jednostek ma od niej wartości nie większe, a 3/4 (75%) nie mniejsze.

- Kwartyl drugi (mediana, wartość środkowa, Me) to wartość jednostki położonej w zbiorowości w ten sposób, że dzieli zbiorowość na dwie równe części. Mediana została wprowadzona do praktyki statystycznej przez K. Pearsona w 1895 roku. Mediana obok średniej arytmetycznej jest najczęściej stosowanym parametrem statystycznym. Wartość mediany nie zależy od wartości krańcowych (odstających). Gdy wartości te pojawiają się (często jako błąd przy zbieraniu informacji) to średnia arytmetyczna może się zmienić znacznie a mediana się nie zmieni. Mediany używamy też do analizy cech porządkowych. Przykładowo mediana dla zarobków w firmie usługowej przedstawionej powyżej wynosi podobnie jak modalna 800 zł. Inny ciekawy przykład zastosowania mediany omówimy na przykładzie historycznego już dziś opracowywania danych eksperymentalnych (wg. E. V. Evarts, E. Bental, B. Bilhari P. R. Huttenlocher „Spontaneous discharge of single neurons during sleep and waking"). W omawianym eksperymencie mierzono częstość wyładowań wielu komórek nerwowych mózgu kota w czasie snu a także po przebudzeniu. Średnia i mediana częstości wyładowań dla 90 komórek nerwowych z obszaru wzrokowego przedstawiona jest w poniższej tabeli

	Średnia	Mediana
w czasie snu	7,3	5,20
po przebudzeniu	7,95	3,56

  Jak widać z tabeli średnia aktywności wzrasta po przebudzeniu kota, a mediana odwrotnie maleje. Wskazuje to na istnienie w mózgu neuronów dwóch różnych typów:

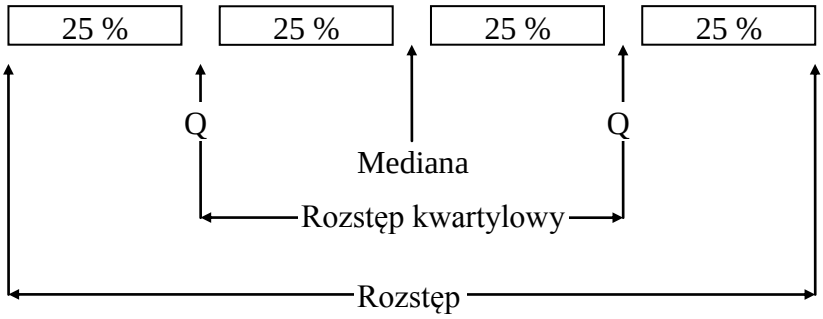
  * zwiększające wyładowania, gdy zwierzę czuwa
  * zmniejszający częstość wyładowań - liczniejszy (co zauważamy na podstawie mediany)

Wynikło stąd, że nie wszystkie komórki mózgu wykonują tę samą pracę (a tak uważano do momentu tego eksperymentu), ale że poszczególne komórki wyspecjalizowały się w różnych czynnościach.

- Kwartyl trzeci Q_3 jest to wartość jednostki, która dzieli zbiorowość w ten sposób, że 3/4 (75%) jednostek ma od niej wartości nie większe, a 1/4 (25%) nie mniejsze.

Z kwartylami związany jest też charakterystyka zwana rozstęp kwartylowy. Jest to różnica pomiędzy kwartylami trzecim i pierwszym. Rozstęp kwartylowy określa długość tej części przedziału zmienności cechy, w której znajduje się 50% „środkowych" obserwacji.

Wprowadzone pojęcia możemy graficznie podsumować na poniższym rysunku



Miary zmienności

W poprzednim artykule poznaliśmy liczby opisujące „środek zbioru" danych. Liczby te nie dają pełnego wizerunku naszego zbioru. Przykładowo w dwu grupach chorych zmierzono ciśnienie skurczowe krwi. Otrzymano następujące wyniki:

grupa I 45, 25, 30, 55, 40, 50, 35
grupa II 15, 40, 10, 80, 40, 65, 30

Po wykonaniu obliczeń otrzymujemy, że średnia i mediana jest taka sama w obu grupach i wynosi 40.

Wartości te nie opisują w pełni zbiorów danych. Patrząc na dane zauważamy bowiem, że pomiary w drugiej grupie są bardziej rozproszone niż w grupie pierwszej. Aby uzyskać lepsze wyobrażenie o naszych danych potrzebujemy drugi rodzaj „podsumowujących" liczb - miary zmienności (rozrzutu, dyspersji). Razem ze statystykami opisowymi dostarczają one bardzo zwięzłego opisu naszych danych. Występują trzy rodzaje miar rozrzutu.
Najprostszą miarą zmienności jest **rozstęp R = x_{max} - x_{min}** (wartość największa minus wartość najmniejsza). Jest to miara charakteryzująca empiryczny obszar zmienności badanej cechy. Omawiana miara nie jest jednak miarą najdoskonalszą. Można bowiem łatwo wyobrazić sobie dwa różne szeregi o jednakowych rozstępach.

W praktyce najczęściej stosuje się dwie miary wariancje i odchylenie standardowe. **Wariancją** zmiennej X nazywamy średnią arytmetyczną kwadratów odchyleń poszczególnych wartości zmiennej od średniej arytmetycznej całej zbiorowości:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

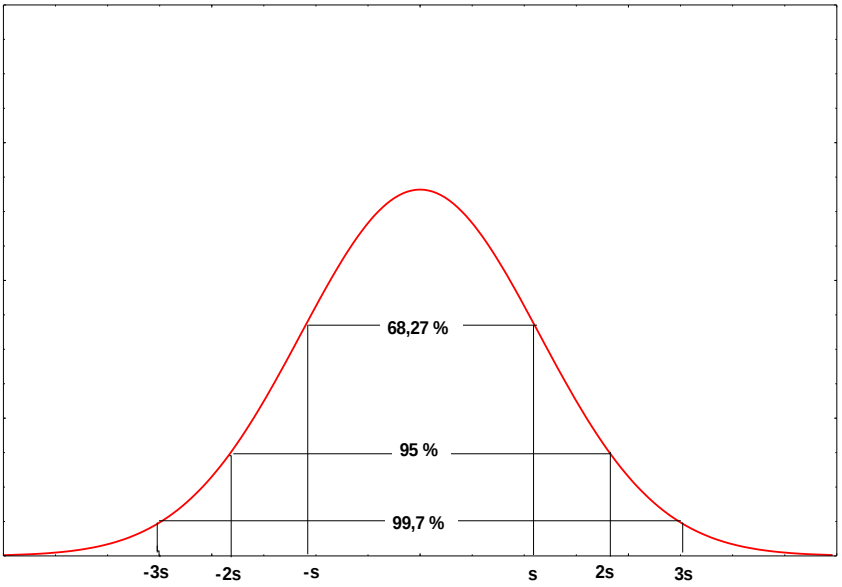
Jest to jedno z ważniejszych pojęć w statystyce i będziemy się z nim spotykać prawie we wszystkich rodzajach wnioskowań statystycznych. Pamiętajmy im wariancja większa, tym bardziej rozproszone są wyniki naszych danych.

Gdy chcemy uzyskać miarę zróżnicowania o mianie zgodnym z mianem zmiennej, obliczamy pierwiastek kwadratowy z wariancji. Zwany on jest **odchyleniem standardowym (s).** Odchylenie standardowe jest obok średniej najczęściej stosowaną statystyką o następujących podstawowych własnościach:

- Jest wielkością obliczaną na podstawie wszystkich obserwacji. Można je poddawać przekształceniom algebraicznym. Im zbiorowość jest bardziej zróżnicowana, tym większe jest odchylenie standardowe. W opisywanych powyżej dwu grupach chorych odchylenia standardowe wynoszą - w pierwszej grupie 10,8 a w drugiej grupie 25,33. Widać więc, że pomiary w drugiej grupie są bardziej rozproszone niż w grupie pierwszej.

- Odchylenie standardowe spełnia regułę trzech sigm, według której w przypadku rozkładu normalnego lub zbliżonego do normalnego
    1. blisko 31,73% wszystkich zaobserwowanych zmiennych różni się od średniej arytmetycznej więcej niż o ± s,
    2. tylko 5% obserwacji wykracza poza przedział ($\bar{x}$ - 2s, $\bar{x}$ + 2s),
    3. tylko 0,3% wszystkich obserwacji wykracza poza przedział ($\bar{x}$ - 3s, $\bar{x}$ + 3s).

Na rysunku poniższym mamy graficzną prezentację reguły trzech sigm.



Rys. 3 Graficzna prezentacja reguły trzech sigm

W wielu książkach statystycznych i pakietach komputerowych zauważymy dzielenie sumy kwadratów przez n - 1 (tzn. o jeden element mniej niż liczba naszych danych). Daje to w efekcie większe odchylenie standardowe. Jak więc należy liczyć? Jeżeli nasz zbiór danych jest tylko próbką to stosujemy wzór zawierający n - 1. Tak wyliczone odchylenie standardowe jest nieobciążonym estymatorem odchylenia w całej populacji. Gdy nasz zbiór danych jest populacją (nie służy do wnioskowania) stosujemy wzór zawierający n.

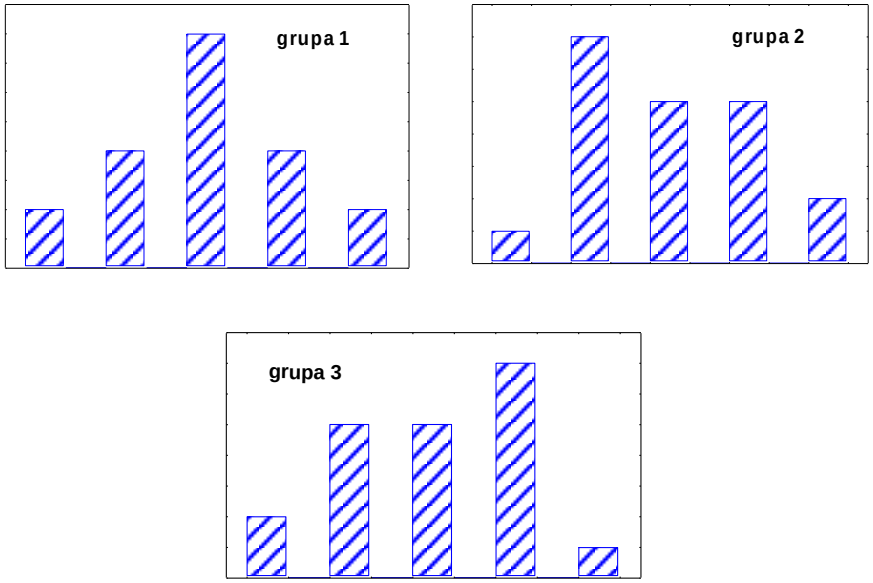
Omówione dotychczas miary zmienności służą do pomiaru absolutnej wielkości zróżnicowania i są liczbami mianowanymi, podobnie jak statystyki opisowe. Stwarza to trudności przy porównywaniu zmienności w dwu lub kilku grupach danych. Dla tego celu wprowadzono nową miarę zwaną **współczynnikiem zmienności.** Jest ona definiowana jako stosunek odchylenia standardowego do średniej arytmetycznej ($V = s/\bar{x} \cdot 100\%$). Dla naszych grup, w których badaliśmy ciśnienie, współczynniki zmienności wynoszą odpowiednio - 27% dla pierwszej grupy a 63,3% (prawie 2,5 razy większy) dla drugiej grupy.

Miary asymetrii

Są sytuacje, w których badanie średniego poziomu zmiennej i rozproszenia jej wartości nie wskazuje na istnienie różnic między badanymi zbiorowościami. Obserwacja zaś rozkładów tych cech wyklucza podobieństwo rozważanych zbiorowości. Przykładowo badano czas reakcji na lek w trzech grupach 100 zwierząt. Dane przedstawiono w postaci poniższej tabeli:

Czas reakcji	Grupa 1	Grupa 2	Grupa 3
10 - 20	10	5	10
20 - 30	20	35	25
30 - 40	40	25	25
40 - 50	20	25	35
50 - 60	10	10	5

Średnia arytmetyczna i wariancja są jednakowe dla wszystkich grup i wynoszą odpowiednio $\bar{x} = 35$, $s^2 = 120$. Mimo to istnieją duże różnice. Widać to wyraźnie na poniższych histogramach.

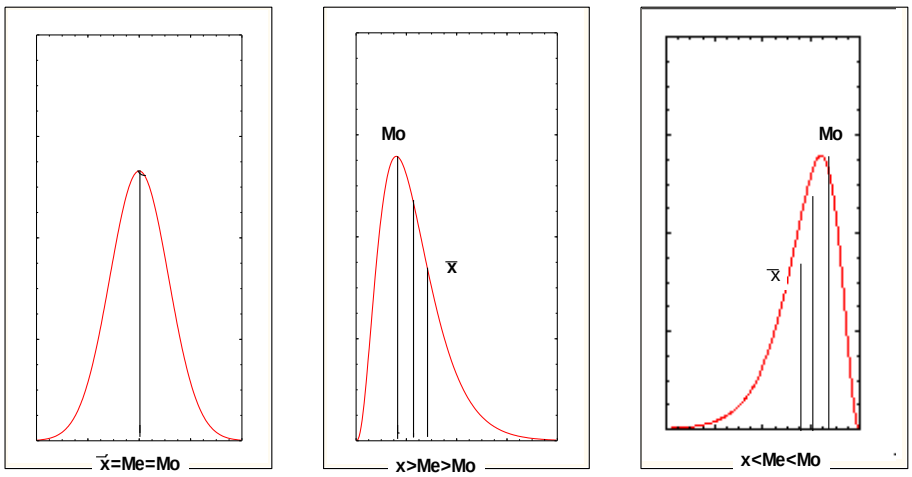


Rys. 4 Histogramy czasu reakcji na lek

Zauważyć można, że w grupie 2 u większości zwierząt czas reakcji na lek jest niższy od przeciętnego, natomiast w grupie trzeciej u większości zwierząt czas reakcji na lek jest wyższy od przeciętnego. Związane to jest z asymetrią rozkładu. Asymetrię można określić porównując średnią arytmetyczną z medianą i modalną. Wyróżnić można trzy przypadki:

- $\bar{x}$ = Me = Mo - dla rozkładu symetrycznego
- $\bar{x}$ > Me > Mo - dla rozkładu o asymetrii prawostronnej
- $\bar{x}$ < Me < Mo - dla rozkładu o asymetrii lewostronnej

Te trzy sytuacje przedstawione są na rysunku 5.

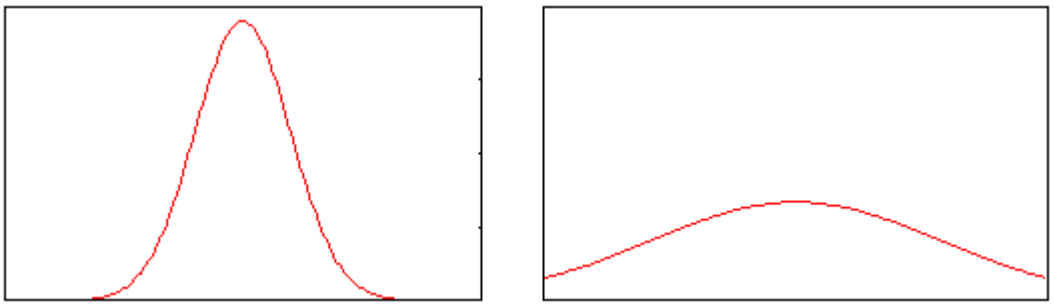


Rys. 5 Położenie miar przeciętnych dla szeregów o różnej asymetrii

W celu określenia kierunku i siły asymetrii wprowadzono **współczynnik asymetrii (skośność - skewness)** A_s. Wprowadzona miara jest cennym narzędziem analizy statystycznej. Sama średnia arytmetyczna mówi niewiele. Dopiero w połączeniu z miarą zmienności i miarą asymetrii otrzymujemy pełny obraz różnic, jakie istnieją między szeregami reprezentującymi zmienne. Współczynnik asymetrii równy zero wskazuje na symetrię rozkładu zmiennej. Wartość dodatnia oznacza asymetrię prawostronną (rozkład ma dłuższy prawy „ogon"), natomiast wartość ujemna oznacza asymetrię lewostronną (rozkład ma dłuższy lewy „ogon").
W naszym przykładzie obliczając mamy - $A_s = 0$ dla grupy 1 (rozkład symetryczny),
$A_s = 0{,}232$ dla grupy 2 (asymetria prawostronna),
$A_s = -0{,}232$ dla grupy 3 (asymetria lewostronna).

Miary koncentracji

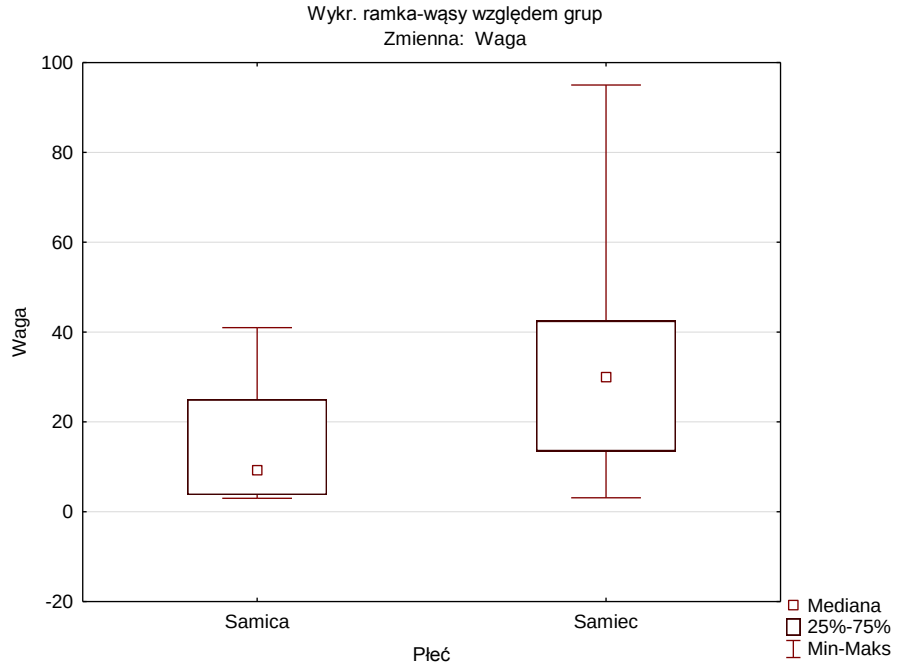
Miary koncentracji doskonale uzupełniają poznane dotychczas statystyki. Opisują one koncentrację wartości cechy wokół średniej. Najpopularniejszą miarą skupienia obserwacji wokół średniej jest **kurtoza (kurtosis)** - oznaczana dalej przez K. Im wyższa jest wartość K, tym bardziej wysmukła jest krzywa liczebności, a zatem większa koncentracja cechy wokół średniej. Jeżeli $K < 0$, to rozkład jest bardziej spłaszczony od normalnego, a jeżeli $K > 0$ to rozkład jest bardziej wysmukły niż normalny. Na rysunku 6 mamy dwie krzywe liczebności - dla pierwszej ma $K > 0$, natomiast dla drugie mamy $K < 0$.



Rys. 6 Krzywe liczebności z różną miarą koncentracji

Najbardziej popularną i syntetyczną interpretacją graficzną wyliczonych statystyk jest wykres ramka-wąsy (ang. *box and whiskey plot*). Wykres ten składa się z prostokąta z wewnętrznym

punktem oraz dwóch wąsów. Omawiane elementy mogą przedstawiać różne statystyki opisowe. Analizując wykres ramkowy możemy uzyskać wiele cennych informacji o rozkładzie badanej cechy. Przykładowy wykres ramka z wąsami pokazany jest na poniższym rysunku.



Rys. 7 Wykres ramka-wąsy dla wzrostu z rozbiciem na płeć zwierząt