PPK – ĆWICZENIE PROJEKTOWE NR 2

WSKAZÓWKI I DEFINICJE

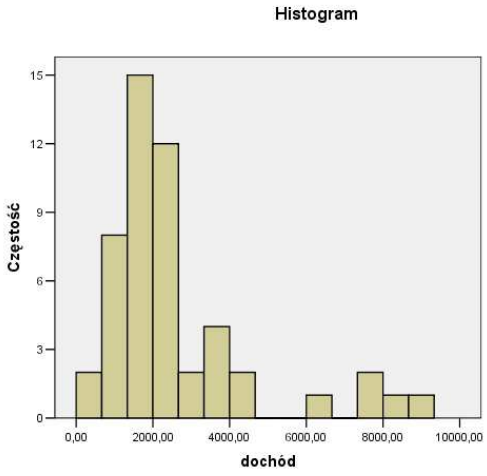
STATYSTYKA OPISOWA **-** Dział statystyki, w którym parametry populacji są liczone bezpośrednio ze zbioru danych, jakim dysponuje badacz. Nie ma za to obecnego we wnioskowaniu statystycznym przeniesienia wyników z próby na populację. Jednostki, których cechy są zarejestrowane w bazie danych są traktowane jako populacja, którą można scharakteryzować szeregiem miar i współczynników.

POPULACJA STATYSTYCZNA **-** Zbiór, na jakim prowadzi się badanie statystyczne. Populacja statystyczna jest zbiorowością trudną do zbadania w całości (z powodów technicznych, finansowych, czasowych), stąd wybiera się z niej odpowiednimi metodami przedstawicieli, czyli tworzy się próbę badawczą. Dąży się przy tym do zachowania reprezentatywności populacji statystycznej. Z populacją związane są cechy statystyczne podlegające badaniu.

PRÓBA STATYSTYCZNA **-** Podzbiór elementów wybranych z populacji i poddanych obserwacji/pomiarowi/ankiecie. Próba statystyczna może mieć charakter losowy lub celowy. Od próby wymaga się, aby była reprezentatywna dla populacji pod względem badanej cechy.

Wartości cechy możemy ustawić w dwojaki sposób albo pogrupować wyniki od najmniejszego do największego tworząc w ten sposób – ***Szereg punktowy*** lub przy dużej ilości wyników grupując te wyniki w ***Szereg rozdzielczy***.

SZEREG ROZDZIELCZY jest to tablica, która pozwala na grupowanie wyników w pewne ich klasy co w znacznym stopniu ułatwia posługiwanie się nimi. Szereg rozdzielczy służy do zbudowania histogramu. ***HISTOGRAM*** jest jednym z najbardziej popularnych wykresów statystycznych. Służy on do przedstawienia liczebności obserwacji, danych w zadanych przedziałach badanej zmiennej. Poniżej przedstawiono przykład **histrogramu** . Przedstawione na nim są wielkość wynagrodzenia wszystkich osób w firmie



Szereg rozdzielczy jest to tablica, która pozwala na grupowanie wyników w pewne ich klasy co w znacznym stopniu ułatwia posługiwanie się nimi. Podstawowy szereg ma następującą budowę:

Numer klasy	Od … do	Środek klasy	Liczebność
……	……	……	……
……	……	……	……

Podstawą do zbudowania takiego szeregu jest odpowiednie pogrupowanie wyników w klasy. W tym celu konieczne jest ustalenie dla danej próby: rozstęp próby, ilości klas, długości klas, początku klasy dolnej.

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

Ilość klas k jest uzależniona od n i najczęściej ustala się ją za pomocą wzorów:

$$k \approx \sqrt{n}$$

lub

$$k \approx 1 + 3{,}3221 \log n$$

Przyjmuję liczę całkowitą oszacowaną według jednego z powyższych wzorów!

Długość klasy b ustalamy wykorzystując następujący wzór :

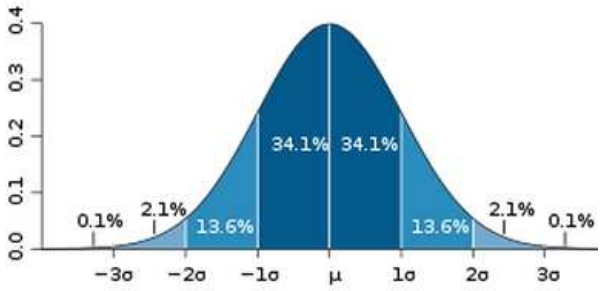
$$b \approx \frac{R}{k}$$

ROZKŁADY PRAWDOPODOBIEŃSTWA WYKORZYSTYWANE W TECHNICE:

ROZKŁAD NORMALNY

Rozkład normalny ma 2 parametry: μ oznacza wartość oczekiwaną, czyli średnią. **Czasem zamiast μ używa się literkę m,** σ oznacza odchylenie standardowe - im większe odchylenie standardowe tym częściej występują obserwacje bardziej oddalone od średniej.

Krzywa Gaussa:
Krzywa Gaussa jest krzywą prezentującą rozkład prawdopodobieństwa rozkładu N(μ,σ)

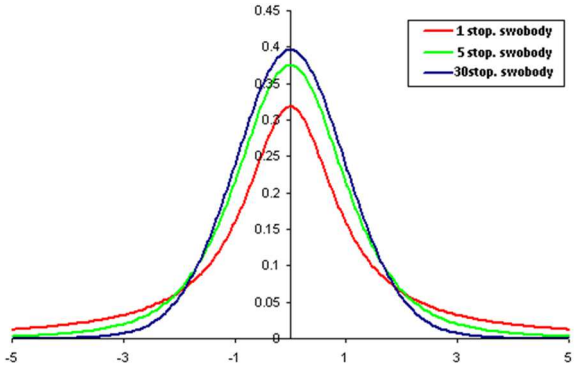


Przyczyną znaczenia rozkładu normalnego jest częstość występowania w naturze. Jeśli jakaś wielkość jest sumą lub średnią bardzo wielu drobnych losowych czynników, to niezależnie od rozkładu każdego z tych czynników jej rozkład będzie zbliżony do normalnego (centralne twierdzenie graniczne). Ponadto rozkład normalny ma interesujące właściwości matematyczne, dzięki którym oparte na nim metody statystyczne są proste obliczeniowo.

ROZKLAD STUDENTA

(nazywany również **rozkład t** albo **rozkład t-studenta**) to rozkład prawdopodobieństwa stosowany przy konstruowaniu przedziałów ufności, testowaniu hipotez statystycznych oraz do oceny błędów pomiaru. Do wyznaczania wartości rozkładu używa się tablicy rozkładu t - studenta .

Kiedy używamy rozkładu t studenta?
Rozkład t studenta **stosujemy tylko** w sytuacji gdy odchylenie standardowe populacji jest nieznane, a rozmiar próby(ilość obserwacji) jest mniejsza niż 30. W przypadku gdy rozmiar próby jest większy lub równy 30 wtedy zamiast brać rozkład t bierzemy rozkład normalny. Wynika to z faktu, że rozkład t studenta dla n≥30n≥30 jest bardzo podobny do rozkładu normalnego. Dla n < 30 rozkład studenta jest „szerszy", tzn. bardziej prawdopodobne są wartości mocno odbiegające od średniej niż w przypadku rozkładu normalnego.



Przedstawienie rozkładu t studenta dla 1, 5 i 30 stopni swobody

PODSTAWOWE PARAMETRY OPISU POPULACJI I PRÓBY:

Parametry statystyczne - są to wielkości liczbowe, które służą do opisu struktury zbiorowości statystycznej w sposób systematyczny

Miary położenia

Miary pozycyjne

modalna
kwantyle:
- kwartyl pierwszy
- mediana (kwartyl drugi)
- kwartyl trzeci
- decyle

Miary przeciętne

średnia arytmetyczna
średnia harmoniczna
średnia geometryczna
modalna

Miary zmienności (rozproszenia, dyspersji)

Miary klasyczne

wariancja
odchylenie standardowe
odchylenie przeciętne
współczynnik zmienności

Miary pozycyjne

rozstęp
odchylenie ćwiartkowe
współczynnik zmienności

Modalna **-** Mo (dominanta D, moda, wartość najczęstsza - jest to wartość cechy statystycznej, która w danym rozdziale empirycznym występuje najczęściej
EXCEL – **WYST.NAJCZĘŚCIEJ.WART()**
Zliczanie przy użyciu funkcji **LICZ.JEŻELI**, ile razy występuje określona wartość.

Kwantyle **-** definiuje się jako 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:

Kwartyl pierwszy Q_1	dzieli zbiorowość na dwie części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q1, a 75% równe bądź wyższe od tego kwartyla
Kwartyl drugi (mediana Me)	dzieli zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od Me; stąd nazwa wartość środkowa
Kwartyl trzeci Q_3	dzieli zbiorowość na dwie części w ten sposób, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q3, a 25% równe bądź wyższe od tego kwartyla

Decyle	np. decyl pierwszy oznacza, że 10% jednostek ma wartości cechy mniejsze bądź równe od decyla pierwszego, a 90% jednostek wartości cechy równe lub większe od decyla pierwszego

EXCEL:

Rozstęp - różnica pomiędzy wartością maksymalną, a minimalną cechy - jest miarą charakteryzującą empiryczny obszar zmienności badanej cechy, nie daje on jednak informacji o zróżnicowaniu poszczególnych wartości cechy w zbiorowości.

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

EXCEL: **=MAX(); =MIN()**

Odchylenie standardowe (zróżnicowanie poszczególnych wartości od średniej)

- jest to pierwiastek kwadratowy z wariancji. Stanowi miarę zróżnicowania o mianie zgodnym z mianem badanej cechy, określa przeciętne zróżnicowanie poszczególnych wartości cechy od średniej arytmetycznej.

Jeżeli zbiorowość danych którą posiadamy jest tylko próbką z dużej zbiorowości odchylenie standardowe jest oszacowaniem (estymatorem), oznaczane jako „s" i liczone ze wzoru:

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

EXCEL: **ODCH.STANDARD.PRÓBKI()**

Jeżeli zbiorowość danych którą posiadamy jest całą populacją odchylenie standardowe jest dokładne, oznaczane jako „σ" i liczone ze wzoru:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

EXCEL: **ODCH.STAND.POPUL()**

Wariancja - jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej zbiorowości.

EXCEL:
Dla próbki: **WARIANCJA.PRÓBKI()**
Dla populacji: **WARIANCJA.POP()**

Współczynnik zmienności - jest ilorazem bezwzględnej mary zmienności cechy i średniej wartości tej cechy, jest wielkością niemianowaną, najczęściej podawaną w procentach.

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

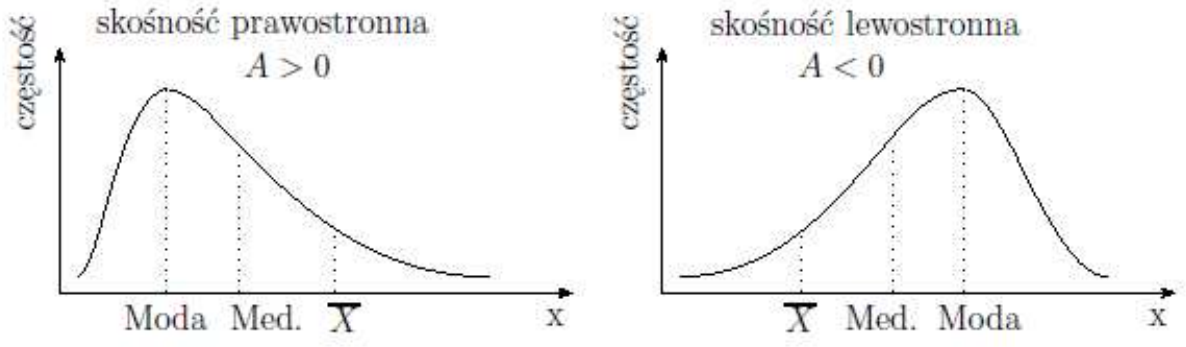
Współczynniki skośności (asymetrii) - są stosowane w porównaniach, do określenia siły oraz kierunku asymetrii, są to liczby niemianowane, im większa ich wartość tym silniejsza asymetria.

$$\gamma(A_s) = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)s^3}$$

EXCEL: **SKOŚNOŚĆ()**

Jeżeli As > 0 to jest asymetria prawostronna; Jeżeli As < 0 to lewostronna; Jeżeli As = 0 to mamy symetrię.

- Asymetria lewostronna - więcej jest wyników większych od średniej;
- Asymetria prawostronna - więcej jest wyników mniejszych od średniej;
- As = 0 mamy symetrię wyników względem średniej
-



Typowy obszar zmienności **-** jeżeli uwzględnimy 2/3 jednostek to wyniki typowe będą w przedziale:

Dla próbki: $(\bar{x} - s; \bar{x} + s)$

Dla populacji: $(\bar{x} - \sigma; \bar{x} + \sigma)$