Miary to grupy statystyk próbkowych, które charakteryzują własności rozkładu empirycznego.

Wyróżniamy:

1) **Miary położenia** wskazują na centralne cechy w rozkładzie, jednak stosowane są różne kryteria w celu określania centralności. Bywają one także nazywane **miarami przeciętnymi**. Wśród nich wyróżnia się:
   a) **średnią arytmetyczną**, nazywaną także **średnią próbkową** , która jest zdefiniowana wzorem:

   dla szeregu szczegółowego:

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

   dla szeregu rozdzielczego:

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

   b) **miary pozycyjne**:

- **kwantyl rzędu** ρ (0 < ρ < 1) w rozkładzie empirycznym cechy to taka wartość cechy x_p , dla której dystrybuanta pierwsza spełnia warunek

$$\widehat{F_n}(x_p) \geq p$$

- **kwartyle**, czyli kwantyle rzędu ρ = 0,25; ρ = 0,5; ρ = 0,75 . Kwantyl q_1 rzędu 0,25 nazywany jest **pierwszym (dolnym) kwartylem** - stanowi taką wartość, że ¼ wartości w próbie jest od niego mniejsza, a ¾ większa. Kwantyl q_2 rzędu 0,5 nazywany jest **medianą z próby** (*Me*) i dla szeregu szczegółowego jest określany wzorem:

$Me = x_k$ dla $k = \frac{N+1}{2}$ gdzie *N - nieparzyste*

$Me = \frac{x_k + x_{k+1}}{2}$ dla $k = \frac{N}{2}$ gdzie *N - parzyste*

  Wartości mediany, posiadają tę własność, że dla danej próby dokładnie połowa elementów w próbie posiada wartości od niej mniejsze, a połowa większe. Jest to więc wartość środkowa w uporządkowanym niemalejąco zbiorze wartości cechy. Kwantyl q_3 rzędu 0,75 nazywany jest **trzecim (górnym) kwartylem** - stanowi on taką wartość, że ¾ wartości w próbie jest od niego mniejsza, a ¼ większa.

  Dla szeregu rozdzielczego wzór na dowolny kwartyl wygląda następująco:

$$Q_i = x_{0Q_i} + (N_{Q_i} - n_{isk-1}) \cdot \frac{h_{Q_i}}{n_{Q_i}}$$

gdzie:

x_{0Qi} - dolna granica przedziału zawierającego kwartyl

N_{Qi} - pozycja kwartyla

n_{isk-1} - liczebność skumulowana przedziału poprzedzającego liczebność skumulowaną kwartyla

h_{Qi} - rozpiętość przedziału zawierającego kwartyl

n_{Qi} - liczebność przedziału zawierającego kwartyl

- **dominanta**, która w rozkładzie empirycznym jest tą wartością cechy, która występuje w rozkładzie najczęściej, która dla szeregu rozdzielczego jest podana wzorem:

$$D = x_D + \frac{n_D - n_{D-1}}{(n_D - n_{D-1}) + (n_D - n_{D+1})} \cdot \Delta x_D$$

gdzie:

x_D - początek przedziału, w którym jest dominanta

n_D - liczebność przedziału, w którym jest dominanta

n_{D-1} - liczebność przedziału poprzedzającego przedział, w którym jest dominanta

n_{D+1} - liczebność przedziału następnego po przedziale, w którym jest dominanta

Δx_D - rozpiętość przedziału, w którym jest dominanta

2) **Miary zróżnicowania wartości w próbie**

a) **wariancja obciążona z próby (próbkowa)**, dla szeregu szczegółowego zdefiniowana wzorem:

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

dla szeregu rozdzielczego:

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

Wyróżniamy także **wariancję nieobciążoną**, która dla szeregu szczegółowego jest zdefiniowana wzorem:

$$s_*^2 = \frac{1}{n-1} \Sigma_i^n (x_i - \bar{x})^2$$

a dla szeregu rozdzielczego:

$$s_*^2 = \frac{1}{n-1} \Sigma_i^n (x_i - \bar{x})^2 n_i$$

b) **odchylenie standardowe (odchylenie próbkowe) obciążone**, zdefiniowane wzorem:

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

c) **odchylenie standardowe (odchylenie próbkowe) nieobciążone** jest zdefiniowane wzorem:

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

d) **współczynnik zmienności**, stanowi odniesienie wartości odchylenia standardowego do średniego poziomu cechy i jest względną, niemianowaną miarą rozrzutu wyników próby. Jest zdefiniowany wzorem:

$$v = \frac{s}{\overline{x}} \cdot 100\%$$

e) **odchylenie przeciętne** to średnia arytmetyczna bezwzględnych odchyleń wartości cechy od średniej arytmetycznej. Określa o ile jednostki danej zbiorowości różnią się średnio, ze względu na wartość cechy, od średniej arytmetycznej. Dla szeregu szczegółowego określane jest wzorem:

$$d = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x} \right|$$

a dla szeregu rozdzielczego:

$$d = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x} \right| \cdot n_i$$

f) **odchylenie ćwiartkowe** to parametr określający odchylenie wartości cechy od mediany. Mierzy poziom zróżnicowania tylko części jednostek; po odrzuceniu 25% jednostek o wartościach najmniejszych i 25% jednostek o wartościach największych. Jest ono określane wzorem:

$$Q = \frac{(Q_3 - Me) + (Me - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$$

g) **pozycyjny współczynnik zmienności** jest definiowany następującym wzorem:

$$V_Q = \frac{Q}{Me} \cdot 100\% \quad \text{gdzie } Me > 0$$

3) **Miary asymetrii rozkładu**

a) **skośność** jest najczęściej stosowana do charakteryzowania asymetrii układu. Przyjmuje wartości z przedziału <– 1, 1>, gdzie wartość 0 oznacza rozkład symetryczny, wartości dodatnie - rozkład o symetrii prawostronnej (prawoskośny), a ujemne - rozkład o symetrii lewostronnej (lewoskośny). Dla szeregu szczegółowego jest zdefiniowana wzorem:

$$as = \frac{\frac{1}{n} \sum_{i=1}^{n} (x - \overline{x})^3}{s^3}$$

a dla szeregu rozdzielczego:

$$as = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^3 n_i}{s^3}$$

4) **Miary koncentracji (skupienia)**

a) **kurtoza** - za jej pomocą wyraża się koncentrację poszczególnych obserwacji wokół średniej. Dla rozkładu normalnego wartość kurtozy wynosi 3. Dla szeregu szczegółowego jest zdefiniowana wzorem:

$$krt = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}$$

dla szeregu rozdzielczego:

$$krt = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4 \cdot n_i}{s^4}$$

b) **eksces** definiowany jest wzorem:

$$ex = krt - 3$$

Im większa jest jego wartość tym większa jest koncentracja wokół wartości oczekiwanej.

1. Ania

Szereg rozdzielczy, nazywany także **rozkładem empirycznym**, jest zbudowany z szeregu statystycznego szczegółowego poprzez grupowanie pomiarów x_1 , .., x_n w pewne klasy. Liczba podziałów (k) powinna zostać odpowiednio dobrana - często przyjmuje się, że jest ona proporcjonalna do pierwiastka kwadratowego z liczby przedziałów (n). Długości przedziałów są ustalane na jednakową wartość równą:

$$h = \frac{x_{max}-x_{min}}{k}$$
$$k = \sqrt{n}$$

gdzie:

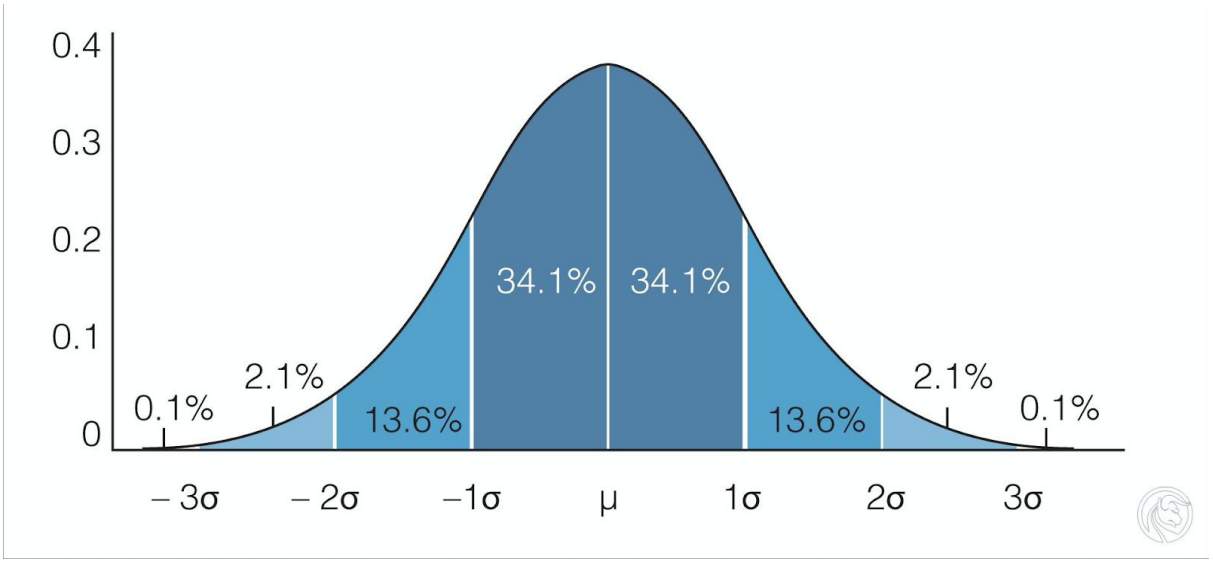
x_{max} - największa wartość cechy statystycznej

x_{min} - najmniejsza wartość cechy statystycznej

Histogram rozkładu empirycznego, nazywany także **histogramem liczebności**, to graficzne przedstawienie liczebności wariantów lub przedziałów liczbowych.

Przedział ufności - przedział informujący o tym, w jakim zakresie na (1-α)*100% mieści się poszukiwana wartość parametru t zgodnie ze wzorem $P(L<t<P)=1-\alpha$ gdzie L, P to wartości krytyczne (krańcowe) przedziału. Przedział ufności jest przydatny w momencie w którym nie możemy przeanalizować całego zestawu danych aby znaleźć konkretną wartość liczbową parametru - wtedy estymuje

się ją na podstawie wybranej grupy danych podając w jakim przedziale i na ile % szukana wartość znajduje się w wyliczonym przedziale.
Rozkład normalny - jeden z najważniejszych rozkładów częstości występowania danej zmiennej. Według niego najczęściej występują wartości, które są bardzo bliskie średniej. Wykres rozkładu normalnego ma charakterystyczny kształt:



Cechy rozkładu normalnego:

- jest symetryczny
- średnia jest równa medianie i dominancie
- ponad 68% wyników leży w maksymalnej odległości jednego odchylenia standardowego od średniej
- wystąpienie wyników większych od średniej od 3 odchylenia standardowe jest niemal nieprawdopodobne

Do sprawdzenia czy rozkład jakiejś zmiennej jest normalny można użyć testu Kołmogorowa-Smirnowa lub Shapiro-Wilka.

Test dwóch średnich - test istotności służący do wnioskowania o równości dwóch średnich w dwóch populacjach. W przypadku takiego testu jako hipotezę zerową przyjmuje się równość średnich, natomiast jako hipotezę alternatywną ich nierówność. W zależności od posiadanych informacji do jego wykonania używa się trzech modeli:

- znane odchylenia standardowe
- nieznane odchylenia standardowe, ale wiadomo że $\sigma 1 = \sigma 2$
- nieznane odchylenia standardowe

Test ANOVA (analiza wariancji) - metoda statystyczna służąca do sprawdzenia czy czynnik (zmienna niezależna) ma wpływ na wyniki jednej zmiennej zależnej. Stosowana gdy zmienna niezależna ma 3 lub więcej poziomów. Polega ona na porównaniu wariancji międzygrupowej do wariancji wewnątrzgrupowej.

Funkcja regresji – matematyczna funkcja określonego typu, która jest przybliżeniem (aproksymatą) faktycznej zależności pomiędzy zmiennymi. W

zależności od rodzaju związku funkcja regresji może przybierać postać liniową i krzywoliniową.

Związek liniowy oznacza, że jednakowym przyrostom zmiennej niezależnej towarzyszą jednakowe co do kierunku (wzrost albo spadek) i siły zmiany zmiennej zależnej.
Regresja krzywoliniowa (funkcja kwadratowa, hiperboliczna, wykładnicza, potęgowa) jednakowym przyrostom zmiennej niezależnej odpowiadają różne co do siły i kierunku zmiany zmiennej zależnej.

Analiza regresji wykorzystywana jest do:

• rozpoznawania wielkości wpływu jednej cechy na drugą w związku przyczynowo – skutkowym (zmienna niezależna i zmienna zależna);

• objaśniania zmienności jednej zmiennej zmiennością drugiej zmiennej;

• szacowania nieznanych wartości jednej cechy na podstawie znanych lub założonych wartości drugiej zmiennej.

Regresja należy do zagadnień uczenia z nadzorem (supervised learning). Uczymy model tzn. na danych treningowych, ze znanymi wartościami Y. Model ma możliwie najtrafniej przewidywać wartości Y na nowych danych. W regresji Y nie jest zmienną kategoryczną, tylko liczbową. Może być wartością dyskretną, np. 1, 2, 3, ..., 10 (itd.), ale też może być wartością ciągłą, czyli dowolną wartością z pewnego przedziału.

Mean Squared Error (MSE) - średnia kwadratów różnic między wartością rzeczywistą a predykcją modelu

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Root Mean Squared Error (RMSE) - jest tożsamy z MSW, ale ma lepszą interpretację (bo w jednostkach zmiennej Y)

$$RMSE = \sqrt{MSE}$$

Model prostej regresji liniowej (1 predyktor)
W modelu tym zakłada się, że związek między zmiennymi ma charakter liniowy tzn.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y - zmienna objaśniana, endogeniczna, predykowana
X - zmienna objaśniająca, egzogeniczna, predyktor
ε - element błędu, losowość
β_0 , β_1 - współczynniki modelu, ich wartości są nieznane i trzeba je oszacować, wyznaczając w ten sposób $\widehat{\beta_0}$ i $\widehat{\beta_1}$.

Wariancja składnika resztowego – miara wahań przypadkowych.

$$Se^2(Y) = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-k}$$

Odchylenie standardowe składnika resztowego – informuje o ile średnio wartości zaobserwowane różnią się od oszacowanych, średni błąd szacunku.

$$Se(Y) = \sqrt{Se^2(Y)} \quad Se(X) = \sqrt{Se^2(X)}$$

Współczynnik zbieżności (współczynnik indeterminacji):

$$\varphi^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Przyjmuje wartości z przedziału . Informuje jaka część zmienności zmiennej zależnej nie jest wyjaśniona zmianami zmiennej niezależnej. Im bliższy 0 tym funkcja regresji lepiej dopasowana do danych empirycznych.

Współczynnik determinacji – informuje, jaka część zmienności zmiennej zależnej jest wyjaśniana kształtowaniem się zmiennej niezależnej:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Pierwiastek ze współczynnika determinacji określany jest indeksem korelacji.

-odchylenie resztowe (residual standard error) - można traktować jak miarę niedopasowania modelu

$$RSE = \sqrt{RSS/(n-2)}$$

-błąd standardowy parametru β_1 :

$$SE(\beta_1) = \frac{RSE}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Wartość statystyki t:

$$t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}$$

Statystyka t, przy założeniu, że $\widehat{\beta_1}$ jest równe 0, ma rozkład t - Studenta z n-2 stopniami swobody w związku z tym jesteśmy w stanie wyznaczyć ρ - value.

Istotność parametrów modelu określa ρ - value. P value - określa prawdopodobieństwo wystąpienia wartości bardziej odległej od zera niż wartość t. Jeśli to prawdopodobieństwo jest odpowiednio małe mniejsze niż 0,05 to przyjmuje się, że parametr $\widehat{\beta_1}$ w istotny sposób różni się od 0. Dla regresji jednej zmiennej (tzn. z jednym predyktorem) współczynnik determinacji jest równy kwadratowi współczynnika korelacji liniowej Pearsona między X i Y.

Model regresji wielorakiej
Predyktorów może być więcej i wtedy zakłada się ich addytywność

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Założenia przy modelu regresji liniowej
-nieliniowość związku
-reszty nie mają rozkładu normalnego
-heteroskedastyczność (zróżnicowanie w zmienności reszt)
-odstające wartości zmiennej predykowanej (outliers)
-odstające wartości predyktorów
-współliniowość

Współczynnik korelacji liniowej Pearsona

$$r_{xy} = \frac{cov(x,y)}{s_x s_y}$$

Kowariancja – średnia arytmetyczna iloczynu odchyleń wartości zmiennych X i Y od ich średnich arytmetycznych.

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Ocena korelacji

$$\begin{array}{cc} |r_{xy}| \leq 0{,}3 & \text{korelacja jest słaba (niewyraźna)} \\ 0{,}3 < |r_{xy}| \leq 0{,}6 & \text{korelacja jest umiarkowana} \\ 0{,}6 < |r_{xy}| & \text{korelacja jest silna} \end{array}$$

Wyróżniamy dwie korelacje
-korelacja dodatnia (wartość współczynnika korelacji od 0 do 1) – informuje, że wzrostowi wartości jednej cechy towarzyszy wzrost średnich wartości drugiej cechy,
-korelacja ujemna (wartość współczynnika korelacji od -1 do 0) - informuje, że wzrostowi wartości jednej cechy towarzyszy spadek średnich wartości drugiej cechy.

-