Przypomnienie ważnych pojęć w analizie statystycznej

Oznaczamy obserwowane wartości zmiennej X przez $x_1, x_2, \ldots, x_n$.

Miary położenia

Dla zmiennych wyrażonych w skali interwałowej i ilorazowej **klasycznymi miarami tendencji centralnej** to najczęściej **średnie**, które informują o przeciętnym poziomie cechy, nie odzwierciedlając różnic pomiędzy poszczególnymi jednostkami.
W zależności od postaci wartości zmiennej stosujemy:

- średnią arytmetyczną (gdy wartości zmiennej można dodawać),
- średnią geometryczną (gdy wartości zmiennej można mnożyć),
- średnią harmoniczną (gdy wartości zmiennej można dodawać).

Wartość średniej wyznaczamy jeśli wartości zmiennej są jednorodne.

Średnia arytmetyczna

Średnia arytmetyczna równa się sumie wszystkich wartości zmiennej podzielonej przez ich liczbę.
Dla zmiennej, która przyjmuje wartości $x_1, x_2, \ldots, x_n$ średnia arytmetyczna $\bar{x}$ wynosi:

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

5% średnia ucięta - średnia wyznaczona z wartości zmiennej , z których wyeliminowano 5% największych i 5% najmniejszych wartości.
Wartość 5% średniej uciętej wyznacza się gdy chcemy aby zmienne nietypowe nie zakłócały wartości średniej.

Średnia geometryczna

Średnia geometryczna $\bar{x}_g$ jest pierwiastkiem *n* - tego stopnia iloczynu *n* wartości zmiennej. Stosuje się ją głównie przy badaniu zmian tempa zjawisk . Średnia geometryczna w mniejszym stopniu niż średnia arytmetyczna odzwierciedla wpływ wartości ekstremalnych na przeciętny poziom zmiennej. Średnia geometryczną wyznacza się ze wzoru:

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n}.$$

Z definicji wynika, że średnią geometryczną możemy wyznaczać tylko wtedy, gdy wartości obserwacje są liczbami dodatnimi i różnymi od zera.

Średnia harmoniczna

Średnią harmoniczna $\bar{x}_h$ (dla liczb różnych od zera) nazywamy odwrotność średniej arytmetycznej z odwrotności wartości zmiennej. Oblicza się ją, gdy wartości zmiennej są podane w jednostkach względnych. Średnia harmoniczną wyznacza się ze wzoru:

$$\bar{x} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}, \text{ przy czym: } \sum_{i=1}^{n}\frac{1}{x_i} \neq 0.$$

Dla wszystkich zmiennych, wyrażonych co najmniej na skali porządkowej, można wyznaczać **nieklasyczne miary tendencji centralnej**. Należą do nich:

- mediana,
- dominanta (moda),
- kwantyle.

Mediana (zwana też wartością środkową) to w wartość w szeregu uporządkowanym, powyżej i poniżej której znajduje się jednakowa liczba wartości zmiennej.
Dominanta (moda) - to najczęściej występująca wartość zmiennej.
Kwantylem rzędu *p* (K_p), gdzie 1 > p > 0, nazywamy każdą liczbę x_p przed, którą znajduje się 100*p*% wartości zmiennej. Kwantyle dla $p = 0{,}25$, $p = 0{,}5$, $p = 0{,}75$ nazywany **kwartylami**.
Gdy: $p = 0{,}25$ – kwartyl dolny (inaczej kwartyl rzędu 1 oznaczany przez Q_1, percentyl 25),
$p = 0{,}5$ - mediana (inaczej kwartyl rzędu 2, percentyl 50),
$p = 0{,}75$ – kwartyl górny (inaczej kwartyl rzędu 3 oznaczany przez Q_3, percentyl 75).

Miary zmienności (rozproszenia, dyspersji)

Miary zmienności dzielimy na:

Miary klasyczne:

- wariancja (dla zmiennych, które można mnożyć),
- odchylenie standardowe(dla zmiennych, które można mnożyć),
- odchylenie przeciętne (dla zmiennych, które można dodawać),
- współczynnik zmienności (dla zmiennych, które można mnożyć i dzielić),

Miary pozycyjne:

- rozstęp (dla zmiennych, które można dodawać),
- odchylenie ćwiartkowe (dla zmiennych, które można dodawać),,
- współczynnik zmienności.

Wariancję S_x^2 wyznaczamy ze wzoru:

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

odchylenie standardowe:

$$S_x = \sqrt{S_x^2} \ .$$

Odchylenie standardowe informuje o ile średnio odchylają się wartości zmiennej od wartości średniej $\bar{x}$. Im mniejsza wartość odchylenia tym wartości zmiennej są bardziej skupione wokół średniej.

Rozstęp *R* to wartość bezwzględna (moduł) różnicy pomiędzy wartością maksymalną i minimalną badanej zmiennej.

$$R = \left|\max(x_1, x_2, ..., x_n) - \min(x_1, x_2, ..., x_n)\right|$$

Odchylenie ćwiartkowe Q (rozstęp międzykwartylowy) - jest to wielkość określająca odchylenie wartości zmiennej od mediany. Mierzy poziom zróżnicowania tylko części jednostek; po odrzuceniu jednostek o wartościach nie większych niż Q_1 oraz jednostek o wartościach nie mniejszych niż Q_3. Im większa szerokość rozstępu ćwiartkowego, tym większe zróżnicowanie wartości zmiennej.

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

Współczynnik zmienności wyznacza się ze wzoru $ZM_x = \frac{S_x}{\bar{x}}, \ \bar{x} \neq 0$.

Miary asymetrii

Istnieje wiele miar służących do wyznaczania asymetrii rozkładu, do najczęściej stosowanych należy **trzeci moment centralny** , który wyznacza się ze wzoru:

$$M_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 ,$$

lub **współczynnik skośności** $WM_3 = \frac{M_3}{S_x^3}$.

Współczynnik skośności przyjmuje wartość zero dla rozkładu symetrycznego, wartości ujemne dla rozkładów o lewostronnej asymetrii (wydłużone lewe ramię rozkładu) i wartości dodatnie dla rozkładów o prawostronnej asymetrii (wydłużone prawe ramię rozkładu).

Miary koncentracji

Miary koncentracji mierzą koncentrację wartości zmiennej wokół średniej. Do najczęściej stosowanych współczynników koncentracji należy **kurtoza** Definiuje się ją następującym wzorem:

$$Kurt = \frac{M_4}{S_x^4} - 3 ,$$

gdzie M_4 nazywane **czwartym momentem centralnym** wyznacza się ze wzoru:

$$M_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \ .$$

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).

Histogram i poligon – różnica (poligon na poniższym wykresie obrazują czarne linie z niebieskimi punktami)