

**LITERATURA:**

- [1] J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwa Naukowo-Techniczne, Warszawa, 2006.
- [2] A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the theory of statistics*, McGraw-Hill Publishing Company, 1983.
- [3] P. Biecek, *Przewodnik po pakiecie R*, Oficyna Wydawnicza GIS, Wrocław, 2008.
- [4] P. Dalgaard, *Introductory Statistics with R*, Springer, 2008

**Wykład 1: Analiza danych a wnioskowanie statystyczne**

Statystyka obejmuje dwa nurty:

1. analizę danych,
2. wnioskowanie statystyczne.

Celem **analizy danych** jest prezentacja konkretnego zbioru danych, w sposób ukazujący jego własności; w szczególności syntetyczny opis podstawowych jego cech. Otrzymujemy wówczas wnioski, które dotyczą **wyłącznie analizowanego zbioru danych**. Na przykład mamy zebrane informacje na temat słuchaczy studiów doktoranckich w PW, którzy rozpoczęli owe studia w 2018 roku. Dokładniej, mamy listę tych słuchaczy wraz z następującymi danymi:

- płeć słuchacza,
- czas (w miesiącach) kiedy słuchacz był doktorantem,
- wysokość pobranego stypendium,
- czy studia zakończyły się uzyskaniem dyplomu doktora.

Na podstawie tych danych możemy stwierdzić np.

- jaki procent słuchaczy, rozpoczynających studia doktorskie w PW w 2018 r., zakończył studia uzyskaniem dyplomu doktora;
- jaki procent owych słuchaczy stanowiły kobiety;
- jakie łączne wydatki poniesiono na stypendia studentów, którzy nie uzyskali dyplomu.

Otrzymamy wyniki **dokładne** i **pewne**, ale dotyczyć będą one **jedynie** słuchaczy studiów doktoranckich w PW, którzy rozpoczęli te studia w 2018 roku.

Teraz wyobraźmy sobie, że chcemy wiedzieć:

- jaki procent studentów doktorantów, studiujących w Polsce, to kobiety;
- ile wynosi średnie miesięczne stypendium doktorantów studiujących w Polsce.

Aby uzyskać dokładną i pewną odpowiedź na powyższe pytania, potrzebowalibyśmy zebrać dane dotyczące wszystkich doktorantów studiujących obecnie w Polsce. To trudne zadanie – czasochłonne i kosztowne, a nawet niekoniecznie wykonalne, bo niektóre jednostki mogą odmówić nam współpracy lub zwlekać z dostarczeniem stosownych danych. Pozostaje wtedy pójść na kompromis – zebrać dane dotyczące tylko wybranych studentów i na ich podstawie wyciągać wnioski o wszystkich studentach. Mamy wtedy do czynienia z **wnioskowaniem statystycznym**. Musimy w nim zwrócić uwagę na dwa aspekty.

1. Bardzo ważny jest odpowiedni wybór studentów do naszego badania - ogólniej – **odpowiedni wybór obserwacji do próby**, tak by dobrze reprezentowały one całą populację.

2. Jeśli tylko jako próby nie weźmiemy całej populacji (a tak we wnioskowaniu statystycznym postępujemy), to **uzyskane wyniki nie będą ani dokładne, ani pewne – pozostaną obciążone błędem.**

### Analiza danych

Jak już wspomnieliśmy, celem analizy danych jest opis podstawowych cech konkretnego zbioru danych. Często, aby taki opis uzyskać, musimy najpierw dane, zawarte w zbiorze, uporządkować i uprościć. Porządkowanie danych rozpoczynamy od ustalenia jakiego są one typu. Możemy mieć:

- **dane ilościowe**, czyli dane w postaci liczb; np. wysokości miesięcznego stypendium studentów w PLN, czas (w miesiącach) od rozpoczęcia studiów doktoranckich do ich zakończenia;
- **dane jakościowe** opisujące cechę jakościową, jak np. płeć, kolor oczu, zawód itp.

Do opisu danych ilościowych możemy użyć miar liczbowych.

## MIARY LICZBOWE DLA DANYCH ILOŚCIOWYCH

### 1). Miary położenia:

- miary tendencji centralnej:
  1. średnia (mean):  $\bar{x} := \frac{\sum_{i=1}^n x_i}{n}$   
> `mean(wektor)` lub > `mean(wektor,na.rm=TRUE)` gdy są obserwacje brakujące
  2. mediana (median) - wartość środkowa  
> `median(wektor)`
  3. moda (dominanta) (mode) - wartość najczęściej pojawiająca się w próbie;
- miary pozycji:
  1. kwantyle (quantiles):  $q_p$ , gdzie  $p \in (0, 1)$   
 $q_p$  to wartość taka, że mniej-więcej  $p \cdot 100\%$  obserwacji jest  $\leq q_p$  i mniej-więcej  $(1 - p) \cdot 100\%$  obserwacji jest  $\geq q_p$ ;  
kwantyl rzędu  $p$  będziemy wyznaczać używając wzoru:

$$q_p = \begin{cases} \frac{x_{np:n} + x_{np+1:n}}{2} & \text{jeśli } np \text{ jest liczbą całkowitą} \\ x_{\lceil pn \rceil:n} & \text{jeśli } np \text{ nie jest liczbą całkowitą} \end{cases}$$

gdzie  $x_{i:n}$  to  $i$ -ta obserwacja po ustawieniu wszystkich obserwacji w kolejności nie-malejącej;

2. dolny kwartył (lower quartile):  $Q_1 := q_{0,25}$   
> `quantile(wektor, 0.25, type=2)`
3. górnny kwartył (upper quartile):  $Q_3 := q_{0,75}$   
> `quantile(wektor, 0.75, type=2)`
4. decyle (deciles)  $k$ -ty decyl to kwantyl rzędu  $k/10$  czyli  $q_{k/10}$
5. percentyle (percentiles):  $k$ -ty percentyl to kwantyl rzędu  $k/100$  czyli  $q_{k/100}$   
> `quantile(wektor, c(0.1, 0.99, 0.85), type=2)`  
Powyższa funkcja wyznacza pierwszy decyl, 99-ty percentyl i kwantyl rzędu 0,85.

### 2). Miary rozproszenia:

1. rozstęp (range):  $Max - Min$   
> `max(wektor)-min(wektor)`

2. rozstęp międzykwartyłowy (interquartile range):  $IQR := Q_3 - Q_1$   
> IQR(wektor)
3. wariancja (variance):  $S^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$   
> var(wektor)
4. odchylenie standardowe (standard deviation):  $S := \sqrt{S^2}$   
> sd(wektor).

### 3). Miary kształtu:

1. skośność (współczynnik asymetrii) (skewness):  $A := \frac{n}{(n-1)(n-2)S^3} \sum_{i=1}^n (x_i - \bar{x})^3$   
Jeśli obserwacje są symetrycznie rozłożone względem średniej (która w tej sytuacji równa się medianie), to  $A = 0$ .
2. kurtoza (współczynnik spłaszczenia) (kurtosis):  
$$K := \frac{n(n+1)}{(n-1)(n-2)(n-3)S^4} \sum_{i=1}^n (x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$
Wskazuje czy dane zawierają więcej i bardziej skrajne obserwacje odstające ( $K > 0$ ) czy ich mniej i mniej skrajne ( $K < 0$ ) niż byśmy oczekiwali od danych z rozkładu normalnego.

```
> install.packages("e1071")
> library(e1071)
> skewness(wektor)
> kurtosis(wektor)
```

## GRAFICZNA PREZENTACJA DANYCH ILOŚCIOWYCH

1. wykres skrzynkowy (wykres typu *skrzynka z wąsami*) (boxplot)  
> boxplot(wektor, range=1.5, horizontal=FALSE)
2. histogram licznosci i histogram częstości (histograms)  
> hist(wektor, freq=TRUE)      i      > hist(wektor, freq=FALSE)
3. jądrowy estymator gęstości (kernel density estimator) - wygładzona wersja histogramu częstości  
> plot(density(wektor))  
lub  
> lines(density(wektor))  
gdy jądrowy estymator gęstości chcemy nanieść na wcześniej sporządzony wykres, np. na histogram częstości.

## GRAFICZNA PREZENTACJA DANYCH JAKOŚCIOWYCH

1. wykres słupkowy (barchart, barplot)  
> barplot(licznosci, col=c("green", ..., "red"))
2. wykres kołowy (piechart)  
> pie(licznosci, col=c("blue", ..., "yellow"))