8. 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 studentów studiów 1-go stopnia na kierunku Informatyka na Wydziale MiNI PW, którzy rozpoczęli owe studia w 2018 roku. Dokładniej, mamy listę tych studentów wraz z następującymi danymi:

- data urodzenia,
- płeć,
- czy student obronił pracę inżynierską na MiNI.

Na podstawie tych danych możemy stwierdzić np.

- jaki procent studentów Informatyki, rozpoczynających studia 1-go stopnia na Wydziale MiNI PW w 2018 r., obronił na tym wydziale pracę inżynierską;
- jaki procent owych studentów stanowiły kobiety;
- jaki był średni wiek owych studentów w momencie, gdy rozpoczynali studia na Informatyce na MiNI?

Otrzymamy wyniki **dokładne** i **pewne**, ale dotyczyć będą one **jedynie** studentów 1-go stopnia Informatyki na Wydziale MiNI, którzy rozpoczęli te studia w 2018 roku.

Teraz wyobraźmy sobie, że chcemy wiedzieć:

- jaki procent studentów, rozpoczynających studia 1-go stopnia w Polsce na kierunku Informatyka, kończy te studia uzyskaniem dyplomu inżyniera;
- jaki procent studentów, rozpoczynających studia 1-go stopnia w Polsce na kierunku Informatyka, to kobiety;
- ile wynosi średni wiek osób, które rozpoczynają studia 1-go stopnia w Polsce na kierunku Informatyka.

Aby uzyskać dokładną i pewną odpowiedź na powyższe pytania, potrzebowalibyśmy zebrać dane dotyczące wszystkich osób, które rozpoczęły lub rozpoczną w Polsce studia 1-go stopnia na kierunku Informatyka. Jest to zadanie niewykonalne - teraz nie zdobędziemy danych dotyczących przyszłych studentów. Ponadto, nawet gdybyśmy zdecydowali się nasze pytania ograniczyć do studentów dotychczasowych, to zebranie odpowiednich danych byłoby trudne - czasochłonne i kosztowne, a nawet nadal niekoniecznie wykonalne, bo niektóre uczelnie 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 obarczone 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. czas (w miesiącach) od rozpoczęcia studiów do obronienia dyplomu; wiek (w latach); wysokość miesięcznego stypentium (w PLN);

- **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. <u>średnia</u> (mean): $\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n}$
     `> mean(wektor)` lub `> mean(wektor,na.rm=TRUE)` , gdy są obserwacje brakujące
  2. <u>mediana</u> (median) - wartość środkowa
     `> median(wektor)`
  3. <u>moda (dominanta)</u> (mode) - wartość najczęściej pojawiająca się w próbie;
- miary pozycji:
  1. <u>kwantyle</u> (quantiles): q_p, gdzie $p \in (0, 1)$
     q_p to wartość taka, że mniej-więcej $p \cdot 100\%$ obserwacji jest $\leqslant q_p$ i mniej-więcej $(1-p) \cdot 100\%$ obserwacji jest $\geqslant q_p$;
     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 niemalejącej;
  2. <u>dolny kwartyl</u> (lower quartile): $Q_1 := q_{0,25}$
     `> quantile(wektor,0.25,type=2)`
  3. <u>górny kwartyl</u> (upper quartile): $Q_3 := q_{0,75}$
     `> quantile(wektor,0.75,type=2)`
  4. <u>decyle</u> (deciles) k-ty decyl to kwantyl rzędu $k/10$ czyli $q_{k/10}$
  5. <u>percentyle</u> (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. <u>rozstęp</u> (range): $Max - Min$
   `> max(wektor)-min(wektor)`
2. <u>rozstęp międzykwartylowy</u> (interquartile range): $IQR := Q_3 - Q_1$
   `> IQR(wektor)`
3. <u>wariancja</u> (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 liczności 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"))`