LITERATURA:

[1] P. Dalgaard, *Introductory Statistics with R*, Springer, 2008
[2] J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwa Naukowo-Techniczne, Warszawa, 2006
[3] P. Biecek, *Przewodnik po pakiecie R*, Oficyna Wydawnicza GIS, Wrocław, 2008

1. WSTĘPNA ANALIZA DANYCH

Dane ze względu na ich charakter dzielimy na:

- *dane ilościowe*, czyli dane opisujące cechę mierzalną jak np. długość, waga, temperatura;
- *dane jakościowe*, czyli dane opisujące cechę niemierzalną jak np. kolor, płeć, zawód.

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) (moda) - wartość najczęściej pojawiająca się w próbie;
- miary pozycji:
  1. dolny kwartyl (lower quartile): Q_1
     > quantile(wektor,0.25)
  2. górny kwartyl (upper quartile): Q_3
     > quantile(wektor,0.75)
  3. decyle, percentyle i kwantyle (deciles, percentiles and quantiles): q_p
     > quantile(wektor,c(0.1,0.99,0.85))
     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 międzykwartylowy (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 devriance): $S := \sqrt{S^2}$
   > 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))

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"))

SPRAWDZANIE NORMALNOŚCI ROZKŁADU

Chcemy sprawdzić, czy rozkład, z którego pochodzi prosta próba losowa, jest rozkładem normalnym.

- Wykres normalności (wykres kwantylowy, Q-Q plot: quantile versus quantile plot).

  Jeśli próba losowa pochodzi z rozkładu normalnego $\mathcal{N}(\mu, \sigma)$, to wykres kwantylowy jest zbiorem punktów leżących mniej-więcej na prostej $y = \sigma x + \mu$.

  > qqnorm(wektor.danych)

  > qqline(wektor.danych)

  Pierwsza z powyższych komend rysuje w R wykres kwantylowy; druga nanosi na ten wykres linię przechodzącą przez kwartyle.

- Test Shapiro-Wilka

  Zaproponowany w 1965 r. jest to dziś uznawany za najlepszy test uniwersalny normalności rozkładu. Konstrukcja tego testu opiera się na wykresie kwantylowym. Dokładniej, wyznacza się linię, która jest możliwie najlepiej dopasowana do punktów tego wykresu (mówiąc precyzyjniej, wyznacza się tzw. *prostą regresji*) i następnie bada się stopień dopasowania tych punktów do owej prostej.

  H_0 : rozkład, z którego pochodzi badana próba losowa, jest normalny

  H_1 : rozkład, z którego pochodzi badana próba losowa, nie jest normalny

  > shapiro.test(dane)

  Niech α oznacza poziom istotności testu. Jeśli p-value $\leq \alpha$, to odrzucamy H_0. W pozostałych sytuacjach, tzn. gdy p-value $> \alpha$, nie mamy podstaw do drzucenia H_0 (przyjmujemy H_0).