# Laboratorium 1

Wstęp do Analizy Danych | Politechnika Krakowska

Jakub Kapała

Numer albumu: 151885

Data: 29.03.2025

## Zadanie 1 - rozgrzewka, filtrowanie

**Treść** Zadeklarować wektor zawierający liczby nieparzyste z przedziału [-5, 15]. Wyświetlić jego podwektory:

1. całość
2. pierwszy element
3. elementy o indeksach od 1 do 3
4. wszystkie poza pierwszym
5. o indeksach wyłącznie 2, 4 i 7
6. elementy dodatnie
7. elementy podzielne przez 3.

**Rozwiązanie** Deklaracja wektora:

```
liczby <- seq(-5, 15, by = 2)
```

Wyświetlenie podwektorów:

1. Całość:

```
liczby
```

```
##  [1] -5 -3 -1  1  3  5  7  9 11 13 15
```

2. Pierwszy element:

```
liczby[1]
```

```
## [1] -5
```

3. Elementy o indeksach od 1 do 3

```
liczby[1:3]
```

```
## [1] -5 -3 -1
```

4. Wszystkie poza pierwszym

```
liczby[-1]
```

```
##  [1] -3 -1  1  3  5  7  9 11 13 15
```

5. O indeksach wyłącznie 2, 4 i 7

```
liczby[c(2, 4, 7)]
```

```
## [1] -3  1  7
```

6. Elementy dodatnie

```
liczby[liczby > 0]
```

```
## [1]  1  3  5  7  9 11 13 15
```

7. Elementy podzielne przez 3

```
liczby[liczby %% 3 == 0]
```

```
## [1] -3  3  9 15
```

## Zadanie 2 - rozgrzewka, malarze, boxplot

**Treść** Korzystając z danych `painters` biblioteki `MASS` wykreślić diagram zawierający wykresy pudełkowe (boxplot) dla poszczególnych szkół malarstwa.

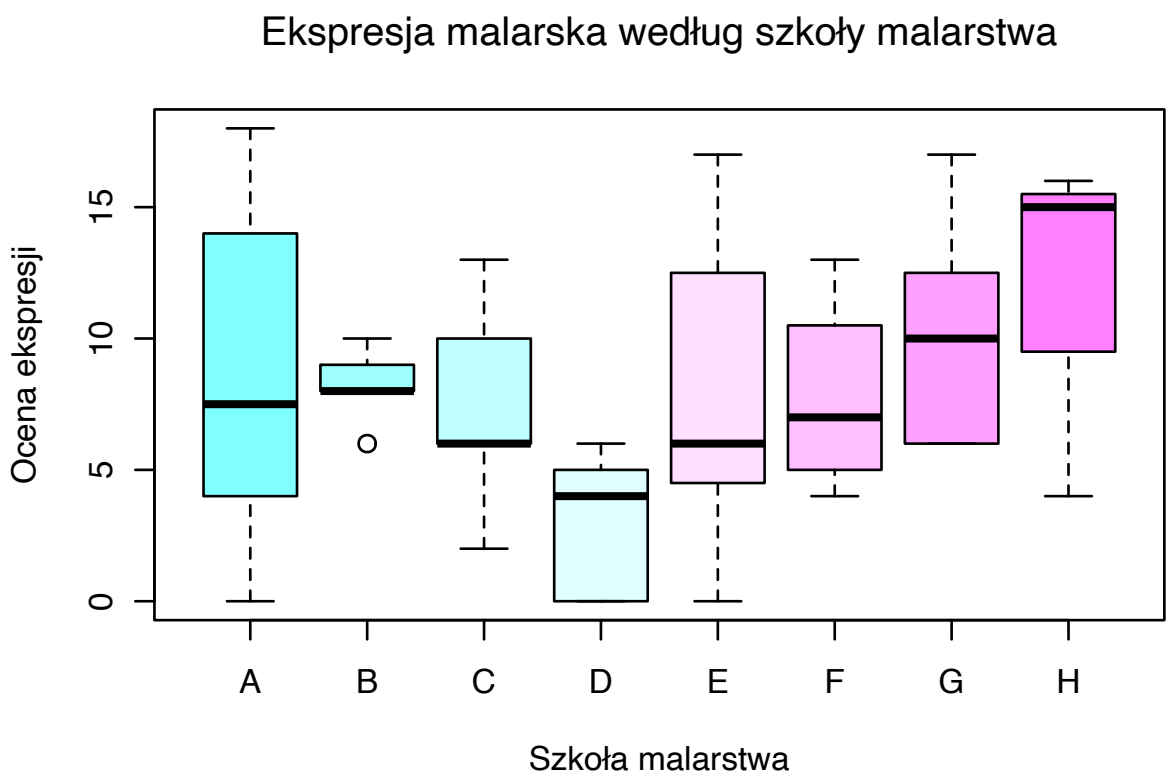
**Rozwiązanie** Wczytanie biblioteki `MASS` oraz danych `painters`:

```
library(MASS)
data(painters)
summary(painters)
```

```
##   Composition       Drawing          Colour        Expression         School
##  Min.   : 0.00   Min.   : 6.00   Min.   : 0.00   Min.   : 0.000   A      :10
##  1st Qu.: 8.25   1st Qu.:10.00   1st Qu.: 7.25   1st Qu.: 4.000   D      :10
##  Median :12.50   Median :13.50   Median :10.00   Median : 6.000   E      : 7
##  Mean   :11.56   Mean   :12.46   Mean   :10.94   Mean   : 7.667   G      : 7
##  3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:16.00   3rd Qu.:11.500   B      : 6
##  Max.   :18.00   Max.   :18.00   Max.   :18.00   Max.   :18.000   C      : 6
##                                                                   (Other): 8
```

Wyświetlenie diagramu:

```
boxplot(
  painters$Expression ~ painters$School,
  main = "Ekspresja malarska według szkoły malarstwa",
  xlab = "Szkoła malarstwa",
  ylab = "Ocena ekspresji",
  col = cm.colors(length(unique(painters$School)))
)
```


Ekspresja malarska według szkoły malarstwa
Ocena ekspresji
0
5
10
15
A
B
C
D
E
F
G
H
Szkoła malarstwa

## Zadanie 3 - wzrost

**Treść** Otworzyć pakiet danych `survey` z biblioteki `MASS`. Przedstawić histogramy wzrostu studentów University of Adelaide osobno dla mężczyzn i kobiet. Za pomocą parametru breaks wypróbować różne szerokości przedziałów histogramu, w tym histogram z przedziałami o niejednakowej szerokości. Przedstawić histogram z podpisanymi osiami, podpisem całego histogramu oraz o wybranym kolorze słupków.

Znaleźć kwartyle/medianę/średnią/minimum/maksimum dla wzrostu mężczyzn oraz kobiet (funkcja summary()).

Przedstawić na wspólnym wykresie boxplot wieku mężczyzn oraz kobiet (analogicznie jak w poprzednim zadaniu).

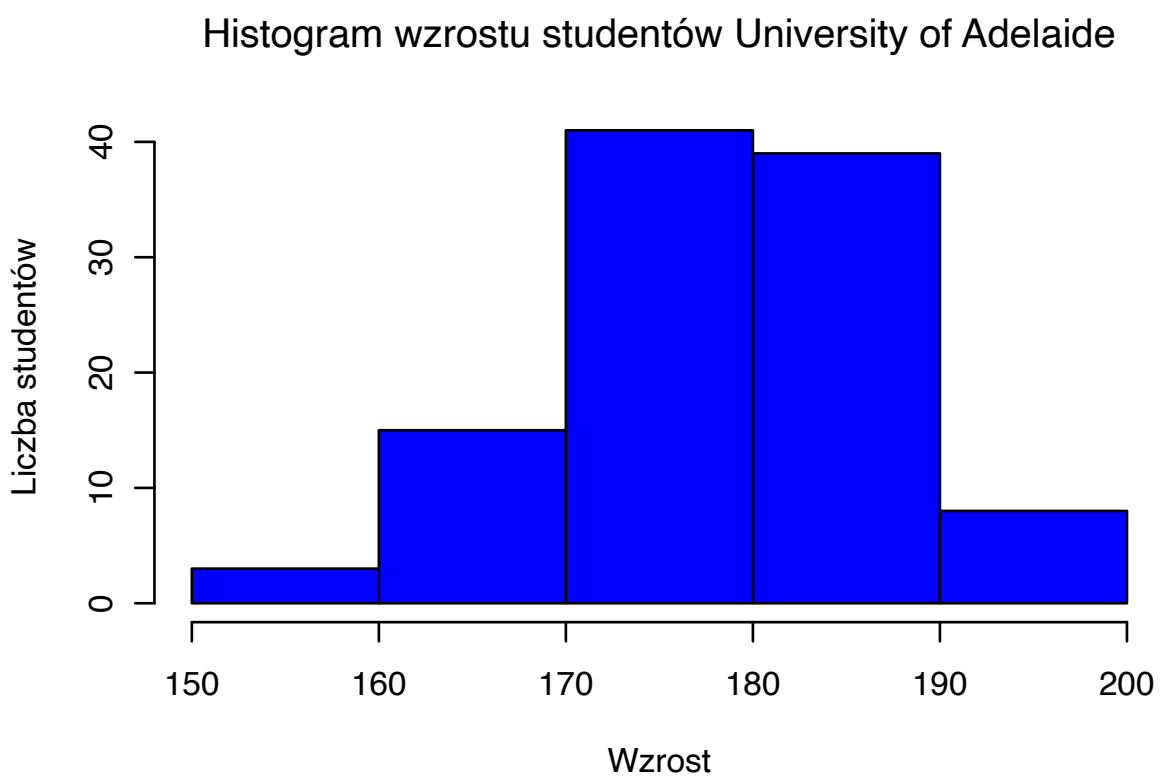
**Rozwiązanie** Wczytanie biblioteki `MASS` oraz danych `survey`:

```
library(MASS)
data(survey)
summary(survey)
```

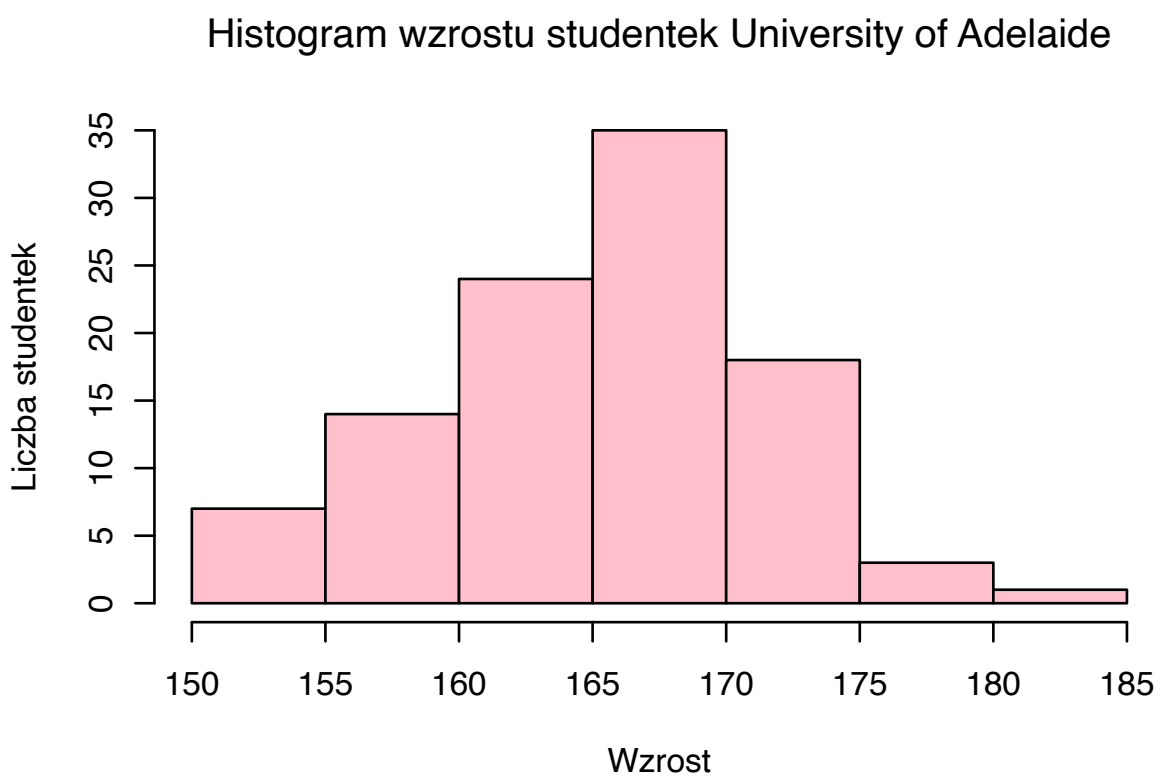
```
##      Sex          Wr.Hnd          NW.Hnd        W.Hnd          Fold
##  Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99
##  Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18
##  NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120
##               Mean   :18.67   Mean   :18.58
##               3rd Qu.:19.80   3rd Qu.:19.73
##               Max.   :23.20   Max.   :23.50
##               NA's   :1       NA's   :1
##      Pulse             Clap       Exer       Smoke         Height
##  Min.   : 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0
##  1st Qu.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0
##  Median : 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0
##  Mean   : 74.15   NA's   :  1              Regul: 17   Mean   :172.4
##  3rd Qu.: 80.00                            NA's :  1   3rd Qu.:180.0
##  Max.   :104.00                                        Max.   :200.0
##  NA's   :45                                            NA's   :28
##        M.I           Age
##  Imperial: 68   Min.   :16.75
##  Metric  :141   1st Qu.:17.67
##  NA's    : 28   Median :18.58
##                 Mean   :20.37
##                 3rd Qu.:20.17
##                 Max.   :73.00
##
```

Wyświetlenie histogramów wzrostu osobno dla męczyzn i kobiet:

```
hist(
  survey$Height[survey$Sex == "Male"],
  main = "Histogram wzrostu studentów University of Adelaide",
  xlab = "Wzrost",
  ylab = "Liczba studentów",
  col = "blue",
  breaks = 5
)
```




```r
hist(
  survey$Height[survey$Sex == "Female"],
  main = "Histogram wzrostu studentek University of Adelaide",
  xlab = "Wzrost",
  ylab = "Liczba studentek",
  col = "pink",
  breaks = 10
)
```




Wyświetlenie kwartyli/mediany/średniej/minimum/maksimum dla wzrostu mężczyzn oraz kobiet:

```
summary(survey$Height[survey$Sex == "Male"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   154.9   172.8   180.0   178.8   185.0   200.0      13
```

```
summary(survey$Height[survey$Sex == "Female"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   150.0   162.6   166.8   165.7   170.0   180.3      17
```

Przedstawienie boxplotu wieku wg płci:

```
boxplot(
  survey$Height ~ survey$Sex,
  main = "Wiek osoby studenckiej według płci",
  xlab = "Płeć",
  ylab = "Wzrost",
  col = c("pink", "blue")
)
```


Wiek osoby studenckiej według płci
Wzrost
150
160
170
180
190
200
Female
Male
Płeć

## Zadanie 4

**Treść** Rozkład Gaussa. Jaki jest rozkład prawdopodobieństwa dla rzutu jedną kością?

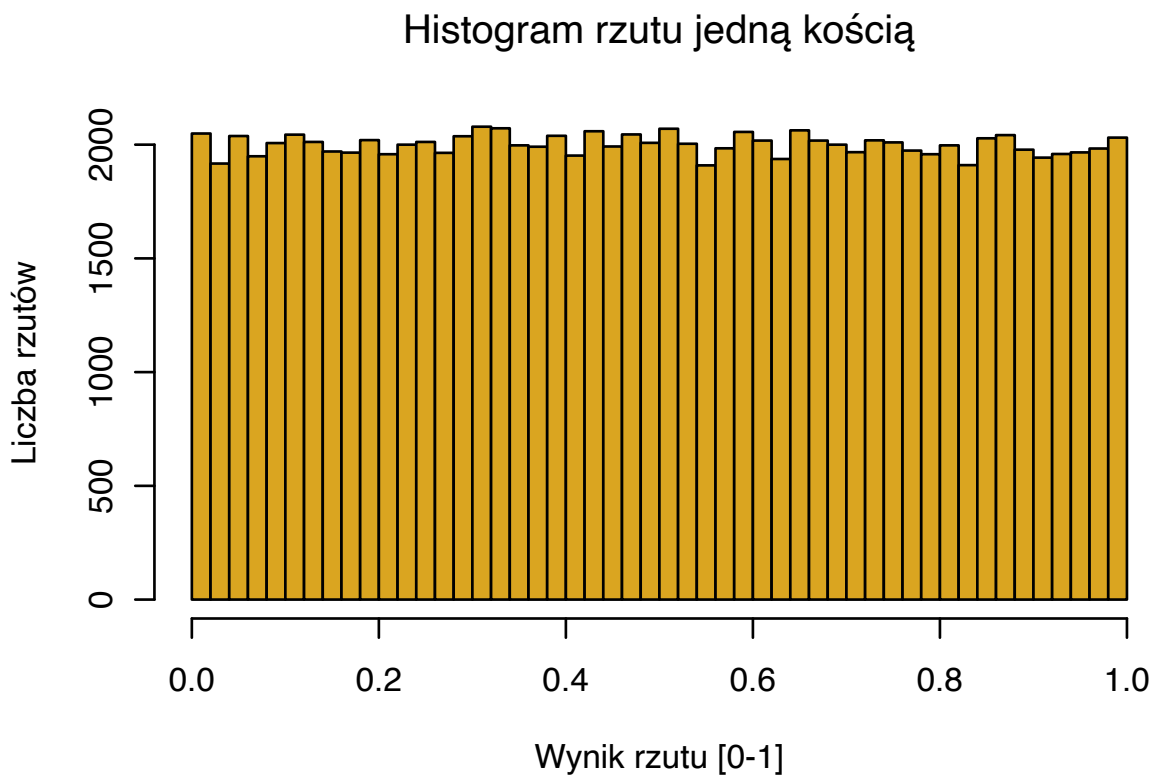
A dwiema (jak przy grze w Monopoly)?

Utworzyć wektor `single` zawierający 100 000 zmiennych losowych z przedziału `[0, 1]` (funkcja `runif()`). Przedstawić histogram wyników tych 100 000 "rzutów kością" – przyjąć szerokość przedziałów równą `0.02`. Opisać poziomą oś wykresu i podpisać cały wykres.

Podobne zadanie wykonać dla wektora `double`, którego każdy element jest średnią arytmetyczną dwóch niezależnych zmiennych losowych, oraz dla `five`, który jest średnią pięciu "rzutów kością".

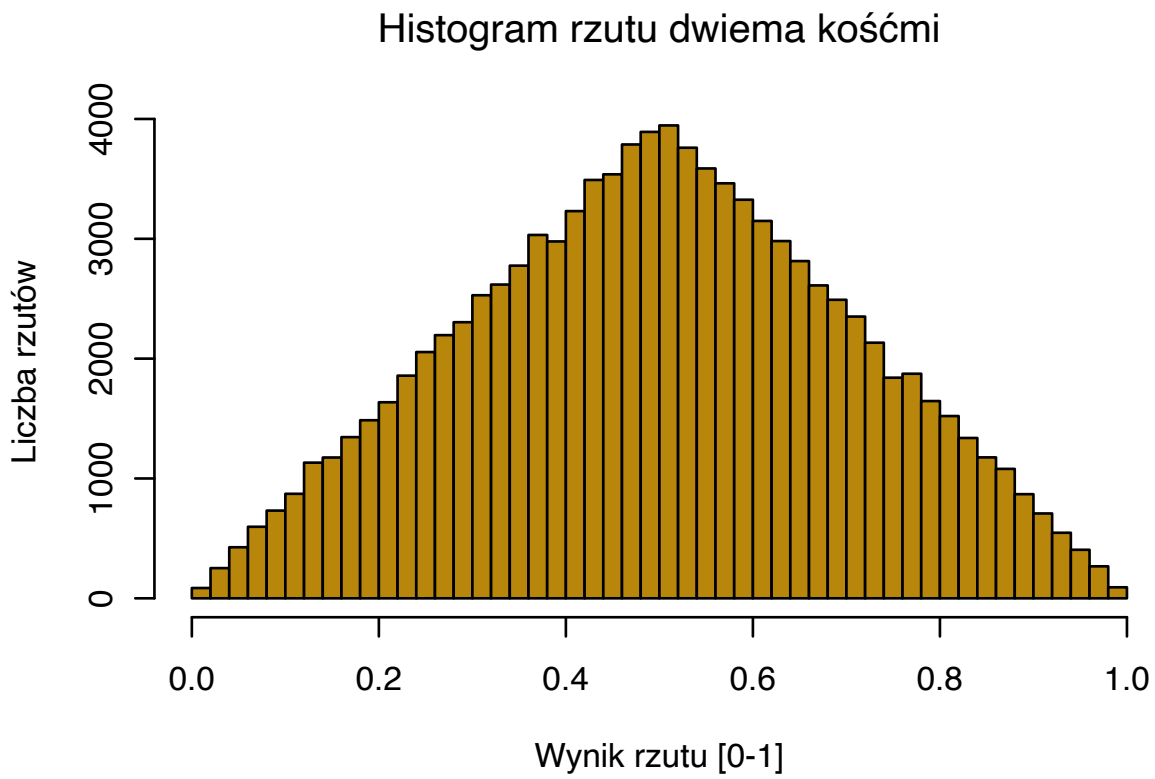
**Rozwiązanie** Wykonanie rzutów jedną kością - wektor `single`:

```r
single <- runif(100000, min = 0, max = 1)
hist(
  single,
  main = "Histogram rzutu jedną kością",
  xlab = "Wynik rzutu [0-1]",
  ylab = "Liczba rzutów",
  col = "goldenrod",
  breaks = seq(0, 1, by = 0.02)
)
```



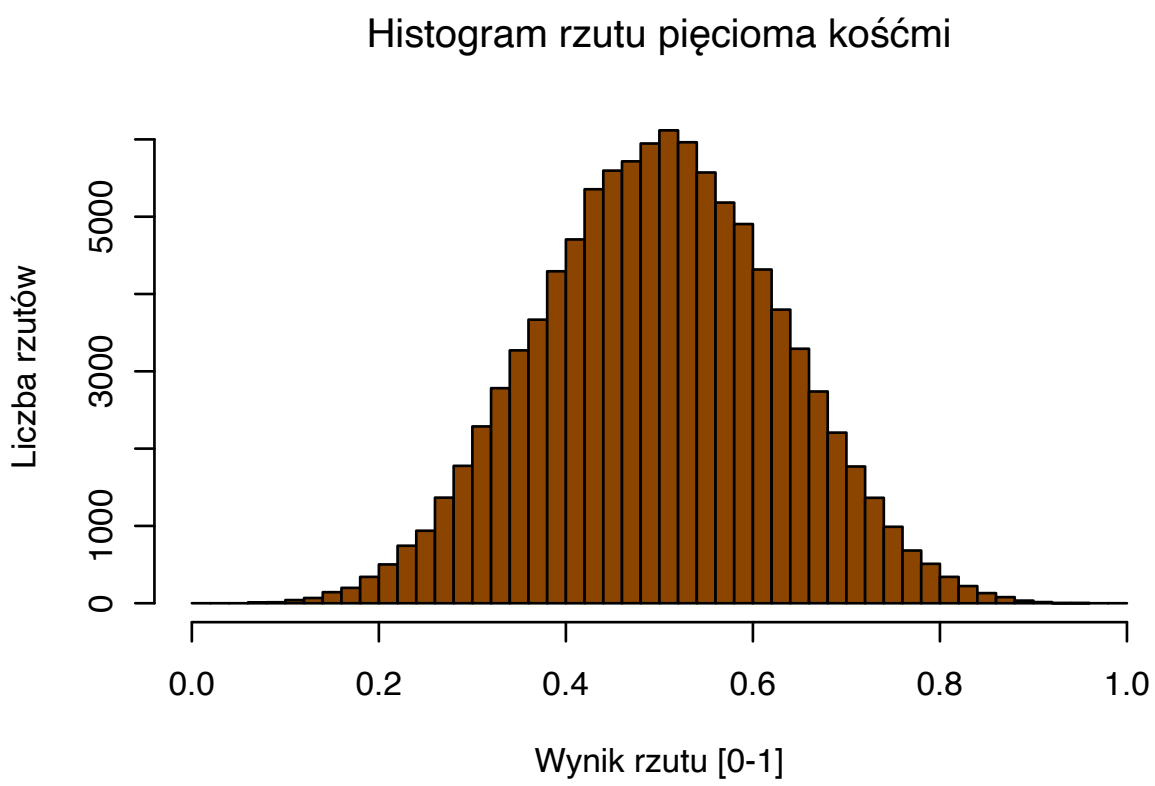

Wykonanie rzutów dwiema kośćmi - wektor `double`:

```r
double <- rowMeans(matrix(runif(100000 * 2, min = 0, max = 1), ncol = 2))
hist(
  double,
  main = "Histogram rzutu dwiema kośćmi",
  xlab = "Wynik rzutu [0-1]",
  ylab = "Liczba rzutów",
  col = "darkgoldenrod",
  breaks = seq(0, 1, by = 0.02)
)
```




Wykonanie rzutów pięcioma kośćmi - wektor `five`:

```r
five <- rowMeans(matrix(runif(100000 * 5, min = 0, max = 1), ncol = 5))
hist(
  five,
  main = "Histogram rzutu pięcioma kośćmi",
  xlab = "Wynik rzutu [0-1]",
  ylab = "Liczba rzutów",
  col = "darkorange4",
  breaks = seq(0, 1, by = 0.02)
)
```


Histogram rzutu pięcioma kośćmi
Liczba rzutów
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Wynik rzutu [0-1]

## Zadanie 3 - test SAT (IU)

**Treść** Wyniki egzaminu SAT Math dla studentów mają średnią 543 i standardowe odchylenie 110.

1. Oblicz Z-scores dla: 300, 400, 500, 600, 700, 800.
2. Oblicz wartości wyników SAT Math dla poszczególnych Z-scores: -2.09, -1.30, -0.39, 0.52, 1.43, 2.34.
3. Porównaj wyniki części (1) i (2).

**Rozwiązanie** Deklaracja zmiennych:

```
sd_sat <- 110
mean_sat <- 543
```

Obliczenie Z-scores:

```
sat_sample <- c(300, 400, 500, 600, 700, 800)
z_scores_from_sample <- sat_sample
z_scores_from_sample <- (z_scores_from_sample - mean_sat) / sd_sat
z_scores_from_sample
```

```
## [1] -2.2090909 -1.3000000 -0.3909091  0.5181818  1.4272727  2.3363636
```

Obliczenie wartości wyników SAT Math dla podanych Z-scores:

```
z_scores_sample <- c(-2.09, -1.30, -0.39, 0.52, 1.43, 2.34)
sat_scores_from_sample <- z_scores_sample
sat_scores_from_sample <- sat_scores_from_sample * sd_sat + mean_sat
sat_scores_from_sample
```

```
## [1] 313.1 400.0 500.1 600.2 700.3 800.4
```

Porównanie wyników:

```
comparison <- data.frame(
  sat_sample,
  z_scores_from_sample,
  z_scores_sample,
  sat_scores_from_sample
)
comparison
```

```
##   sat_sample z_scores_from_sample z_scores_sample sat_scores_from_sample
## 1        300           -2.2090909          -2.09                 313.1
## 2        400           -1.3000000          -1.30                 400.0
## 3        500           -0.3909091          -0.39                 500.1
## 4        600            0.5181818           0.52                 600.2
## 5        700            1.4272727           1.43                 700.3
## 6        800            2.3363636           2.34                 800.4
```

Jak widać, jedynie dla rzędu 1 (`sat_score = 300` -> `z_score = -2.2090909` oraz `z_score = -2.09` -> `sat_score = 313.1`) dane są różne, co wynika z tego, że dane wejściowe/wyjściowe różnią się miedzy soba.

Wszystkie inne wyniki są zgodne, ponieważ dane wejściowe do przykładu 2 są takie same jak dane wyjściowe z przykładu 1.

## Zadanie 6 - wzrost mężczyzn, analiza statystyczna (IU)

**Treść** Dla wzrostu mężczyzn z zestawu danych `survey` znaleźć:

1. Wartość średnią
2. Odchylenie standardowe
3. Z-score używając: (i) wzoru z-score, (ii) funkcji scale()
4. Pokazać z-score graficznie. Opisać oś poziomą oraz pionową.
5. Średnią oraz odchylenie standardowe otrzymanych z-scores. Zinterpretować wyniki.
6. Minimum oraz maksimum z-score. Co oznaczają te wartości?
7. Dla wybranych trzech z otrzymanych powyżej wartości z-score, użyj funkcję pnorm().

Zinterpretuj wynik.

**Rozwiązanie** Wczytanie biblioteki `MASS` oraz danych `survey`:

```
library(MASS)
data(survey)
summary(survey)
```

```
##      Sex          Wr.Hnd          NW.Hnd        W.Hnd          Fold
##  Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99
##  Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18
##  NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120
##               Mean   :18.67   Mean   :18.58
##               3rd Qu.:19.80   3rd Qu.:19.73
##               Max.   :23.20   Max.   :23.50
##               NA's   :1       NA's   :1
##      Pulse             Clap       Exer       Smoke         Height
##  Min.   : 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0
##  1st Qu.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0
##  Median : 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0
##  Mean   : 74.15   NA's   :  1              Regul: 17   Mean   :172.4
##  3rd Qu.: 80.00                            NA's :  1   3rd Qu.:180.0
##  Max.   :104.00                                        Max.   :200.0
##  NA's   :45                                            NA's   :28
##        M.I           Age
##  Imperial: 68   Min.   :16.75
##  Metric  :141   1st Qu.:17.67
##  NA's    : 28   Median :18.58
##                 Mean   :20.37
##                 3rd Qu.:20.17
##                 Max.   :73.00
##
```

Dane wzrostu mężczyzn:

```
male_heights <- survey$Height[survey$Sex == "Male"]
male_heights_clean <- na.omit(male_heights)
```

1. Wartość średnia:

```
mean_mh <- mean(male_heights_clean)
mean_mh
```

```
## [1] 178.826
```

2. Odchylenie standardowe:

```
sd_mh <- sd(male_heights_clean)
sd_mh
```

```
## [1] 8.380252
```

3. Z-score używając:

(i) wzoru z-score:

```
male_heights_z_scores <- male_heights_clean
male_heights_z_scores <- (male_heights_z_scores - mean_mh) / sd_mh
summary(male_heights_z_scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.8503 -0.7203  0.1401  0.0000  0.7367  2.5266
```

(ii) funkcji `scale()`:

```
male_heights_z_scores_scale <- scale(male_heights_clean)
summary(male_heights_z_scores_scale)
```

```
##        V1
##  Min.   :-2.8503
##  1st Qu.:-0.7203
##  Median : 0.1401
##  Mean   : 0.0000
##  3rd Qu.: 0.7367
##  Max.   : 2.5266
```

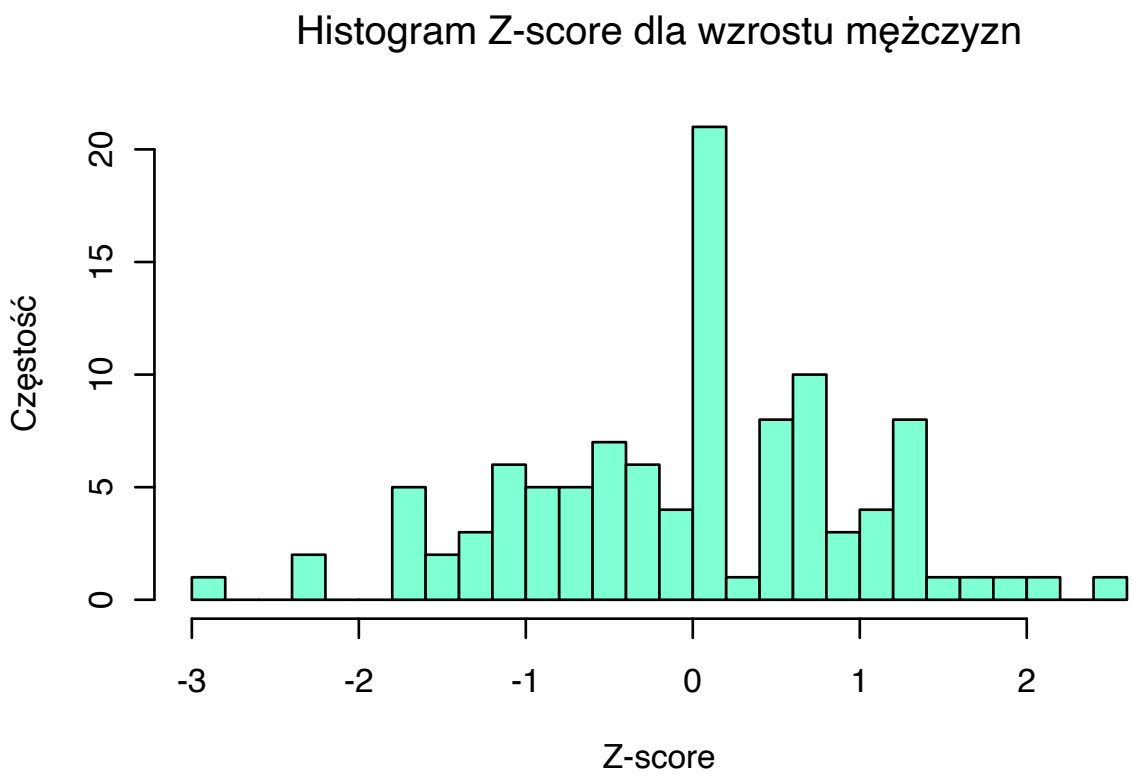
Sprawdzam czy wyniki są takie same:

```
all.equal(
  as.vector(male_heights_z_scores),
  as.vector(male_heights_z_scores_scale)
)
```

```
## [1] TRUE
```

4. Pokazanie z-score graficznie:

```r
hist(
  male_heights_z_scores,
  main = "Histogram Z-score dla wzrostu mężczyzn",
  xlab = "Z-score",
  ylab = "Częstość",
  col = "aquamarine",
  breaks = 20
)
```




5. Średnia i odchylenie standardowe z-scores:

```r
mean(male_heights_z_scores)
```

```
## [1] -1.645434e-15
```

```r
sd(male_heights_z_scores)
```

```
## [1] 1
```

Te wyniki oznaczają, że z-score jest znormalizowany i ma rozkład normalny (`mean ~ 0`, `sd == 1`).

6. Minimum i maksimum z-score:

```r
min(male_heights_z_scores)
```

```
## [1] -2.850277
```

```r
max(male_heights_z_scores)
```

```
## [1] 2.52665
```

Wartości skrajne są odpowiednio około 2.85 i 2.5 odchyleń standardowych od średniej.

7. Dla wybranych trzech z otrzymanych powyżej wartości z-score, użyj funkcję pnorm().

```r
male_heights_z_scores[1:3]
```

```
## [1] -0.1224352 -2.2464764 -1.6498356
```

```r
pnorm(c(
  male_heights_z_scores[1],
  male_heights_z_scores[2],
  male_heights_z_scores[3]
))
```

```
## [1] 0.45127718 0.01233675 0.04948828
```

Interpretacja:

- 1 wynik oznacza, że około 45.13% danych leży poniżej tej wartości, ponadto z-score jest bliski 0, co oznacza, że wartość jest bliska średniej.
- 2 wynik oznacza, że około 1.23% danych leży poniżej tej wartości, ponadto wskazuje to, że wartość jest znacznie poniżej średniej.
- 3 wynik oznacza, że około 4.95% danych leży poniżej tej wartości, ponadto ta wartość jest lekko powyżej poprzedniej wartości

## Zadanie 7 – percentyle, wiek studentów i studentek (IU)

**Treść** Określić:

(i) kwartyle,
(ii) percentyle (0.32, 0.48 i 0.86) oraz rozstęp ćwiartkowy

dla wieku studentów University od Adelaide (funkcja `quantile()`).

Przedstawić boxploty wieku mężczyzn oraz kobiet.

**Rozwiązanie** Wczytanie biblioteki `MASS` oraz danych `survey`:

```
library(MASS)
data(survey)
summary(survey)
```

```
##      Sex          Wr.Hnd          NW.Hnd        W.Hnd          Fold
##  Female:118   Min.   :13.00   Min.   :12.50   Left : 18   L on R : 99
##  Male  :118   1st Qu.:17.50   1st Qu.:17.50   Right:218   Neither: 18
##  NA's  :  1   Median :18.50   Median :18.50   NA's :  1   R on L :120
##               Mean   :18.67   Mean   :18.58
##               3rd Qu.:19.80   3rd Qu.:19.73
##               Max.   :23.20   Max.   :23.50
##               NA's   :1       NA's   :1
##      Pulse            Clap       Exer       Smoke         Height
##  Min.   : 35.00   Left   : 39   Freq:115   Heavy: 11   Min.   :150.0
##  1st Qu.: 66.00   Neither: 50   None: 24   Never:189   1st Qu.:165.0
##  Median : 72.50   Right  :147   Some: 98   Occas: 19   Median :171.0
##  Mean   : 74.15   NA's   :  1              Regul: 17   Mean   :172.4
##  3rd Qu.: 80.00                            NA's :  1   3rd Qu.:180.0
##  Max.   :104.00                                        Max.   :200.0
##  NA's   :45                                            NA's   :28
##        M.I           Age
##  Imperial: 68   Min.   :16.75
##  Metric  :141   1st Qu.:17.67
##  NA's    : 28   Median :18.58
##                 Mean   :20.37
##                 3rd Qu.:20.17
##                 Max.   :73.00
##
```

(i) Kwartyle:

```
quartiles <- quantile(na.omit(survey$Age))
quartiles
```

```
##     0%    25%    50%    75%   100%
## 16.750 17.667 18.583 20.167 73.000
```

(ii) Percentyle i rozstęp ćwiartkowy:

```
percentiles <- quantile(na.omit(survey$Age), probs = c(0.32, 0.48, 0.86))
percentiles
```

```
##      32%      48%      86%
## 17.91700 18.50000 21.90364
```

Rozstęp ćwiartkowy:

```
iqr <- IQR(na.omit(survey$Age))
iqr
```

```
## [1] 2.5
```

Boxploty wieku mężczyzn oraz kobiet:

```
boxplot(
  survey$Age ~ survey$Sex,
  main = "Wiek osoby studenckiej według płci",
  xlab = "Płeć",
  ylab = "Wiek",
  col = c("pink", "blue"),
  ylim = c(15, 30)
)
```