UE Statistik und Wahrscheinlichkeitstheorie

Abgabe 1

Volodymyr Yakovenko (Matr. Nr. 12329558)

22.04.2025

Erste Frage

Seed & dataset load:

```
load("Marathon.RData")
set.seed(12329558)
```

```
jahr <- sample(levels(mara[,4]),1)
```

Dieser Befehl zieht zufällig ein Jahr dieser Daten.

Führen Sie analog einen Befehl aus, um eine Altersgruppe zufällig auszuwähen.

```
altersgr <- sample(levels(mara[,3]), 1)
```

Nun selektieren Sie alle Daten von diesem Jahr und dieser Altersgruppe und speichern Sie die Daten in einem R Objekt ab.

```
df <- mara[mara[,4] == jahr & mara[,3] == altersgr, ]
```

Erstellen Sie weiters ein R Objekt mit den Daten vom gleichen Jahr aber von der nächsthöheren (oder nächstniedrigeren) Altersgruppe.

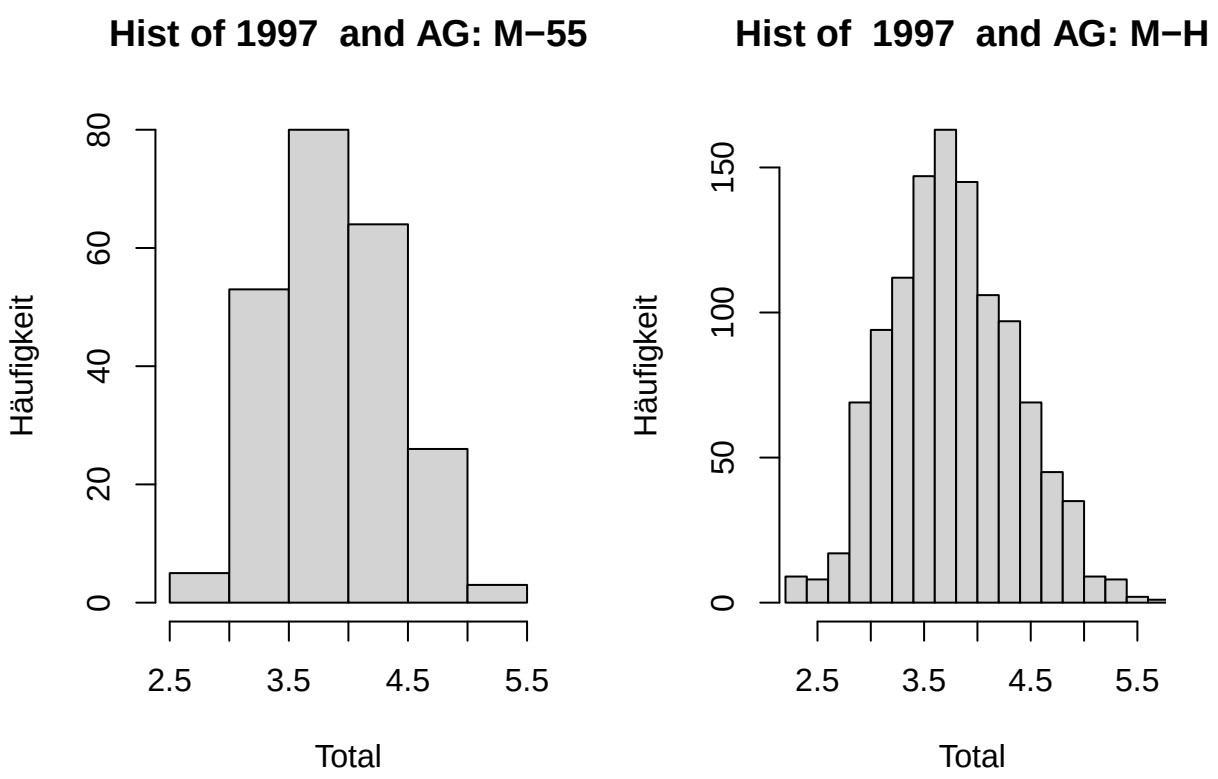
```
altersgr2 <- NA
pos <- which(altersgr == mara$Altersgr)
if(pos[1] == length(mara$Altersgr)){
  altersgr2 <- as.character(mara$Altersgr[pos[1]-1])
} else {
  altersgr2 <- as.character(mara$Altersgr[pos[1]+1])
}
df2 <- mara[mara[, 4] == jahr & mara[, 3] == altersgr2, ]
```

Im Folgenden interessieren wir uns für die Gesamtzeiten (Total) der Laufergebnisse.

- Stellen Sie beide Datensätze nebeneinander durch Histogramme dar (Grafik-Parameter mit par(mfrow=c(1,2)) anpassen). Wie kann die Klassengröße verändert werden?

```
par(mfrow = c(1, 2))
hist(df$Total,
     xlab = "Total",
     ylab = "Häufigkeit",
     main = paste("Hist of", jahr, " and AG:", altersgr ))
hist(df2$Total,
     xlab = "Total",
     ylab = "Häufigkeit",
```

```
     xlim = c(min(df$Total, df2$Total), max(df$Total, df2$Total)),
     main = paste("Hist of ", jahr, " and AG:", altersgr2 ))
```



Antwort: Wir können den Parameter breaks verwenden, um seinen eigenen Wert festzulegen. Laut der Dokumentation:

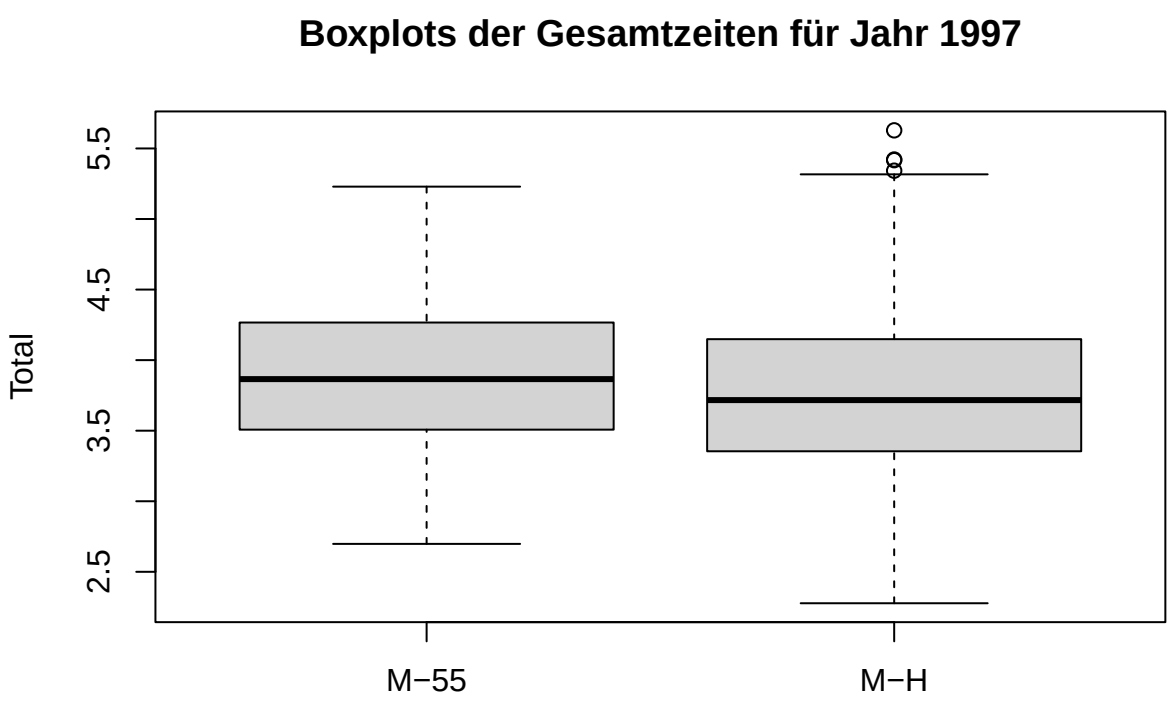
breaks one of:

1) *a vector giving the breakpoints between histogram cells,*
2) *a function to compute the vector of breakpoints,*
3) *a single number giving the number of cells for the histogram,*
4) *a character string naming an algorithm to compute the number of cells (see 'Details'),*
5) *a function to compute the number of cells.*

So können wir die Anzahl der Klassen und damit auch die Anzahl der zu verteilenden Werte ändern. Je kleiner die Anzahl der Klassen ist, desto größer ist die Anzahl der zu verteilenden Werte, je größer die Anzahl der Klassen ist, desto mehr Verteilungsmöglichkeiten gibt es und desto kleiner sind die Klassen selbst.

- Stellen Sie beide Datensätze in parallelen Boxplots dar (eine gemeinsame Skalierung).

```
par(mfrow = c(1, 1))
 boxplot(df$Total, df2$Total,
          names = c(altersgr, altersgr2),
          ylab = "Total",
          main = paste("Boxplots der Gesamtzeiten für Jahr", jahr)
          )
```



- Welche Schlüsse ziehen Sie aus diesen Darstellungen?

Antwort: Histogramms: Die AG:M-H ist eindeutig größer, wie aus den Histogrammen und der Rangfolge der Werte auf der y-Achse hervorgeht, die M-H-Häufigkeit erreicht über 150 und die Werte und ihre Streuung sind einfach größer und breiter. Beide Verteilungen ähneln in gewisser Weise der Normalverteilung, obwohl sie andererseits etwas asymmetrisch (leicht nach links verschoben) sind. Der Spitzenwert wird bei etwa gleichen Werten auf der x-Achse erreicht.Die Anzahl der Klassen (auf der R-Seite definiert) ist im zweiten Diagramm ebenfalls größer, was ebenfalls darauf hindeutet, dass es mehr Werte in der Stichprobe gibt als im ersten Fall.

Boxplots: Beide haben einen Median von knapp 4, sind also ungefähr gleich groß. Die Streuung der Werte um diesen Wert (genauer gesagt von Q0,25 bis Q0,75) ist ungefähr gleich groß, aber auch hier ist der „Bereich“ im zweiten Diagramm etwas nach unten verschoben.

Ein signifikanter Unterschied ist bei der Betrachtung der Whisker (1,5*IQR-Abweichung) zu erkennen. Wenn die obere Grenze des Whiskers gleich ist, ist die untere Grenze im zweiten Diagramm um 1 niedriger als im ersten, was mit den in den Histogrammen beobachteten Ergebnissen übereinstimmt, dass es im zweiten Diagramm einfach mehr Werte gibt und diese „breiter“ verteilt sind. Es ist auch möglich, Ausreißer bei M-H durch charakteristische „Kreise“ oberhalb der Whisker zu erkennen.

- Berechnen Sie folgende Kenngrößen für beide Datensätze und vergleichen Sie die Ergebnisse: Minimum, Maximum, Mittelwert, Median, Quartile, Varianz und Standardabweichung.

```
stats_df <- c(Minimum = min(df$Total),
              Maximum = max(df$Total),
              Mittelwert = mean(df$Total),
              Median = median(df$Total),
              "Q" = quantile(df$Total, probs = 0.25),
              "Q" = quantile(df$Total, probs = 0.75),
              Varianz = var(df$Total),
```

```
                Standardabweichung = sd(df$Total))

stats_df2 <- c(Minimum_df2 = min(df2$Total),
                  Maximum_df2 = max(df2$Total),
                  Mittelwert_df2 = mean(df2$Total),
                  Median_df2 = median(df2$Total),
                  Quartile_025df2 = quantile(df2$Total, probs = 0.25),
                  Quartile_075df2 = quantile(df2$Total, probs = 0.75),
                  Varianz_df2 = var(df2$Total),
                  Std_df2 = sd(df2$Total))

stats_table <-data.frame(
  Statistik = names(stats_df),
  stats_M_55 = stats_df,
  stats_M_H = stats_df2,
  row.names = NULL,
  check.names = FALSE
)

print(stats_table)
```

```
##            Statistik stats_M_55 stats_M_H
## 1            Minimum  2.6977780 2.2772220
## 2            Maximum  5.2294440 5.6283330
## 3         Mittelwert  3.9101046 3.7662280
## 4             Median  3.8655560 3.7168055
## 5               Q.25%  3.5073610 3.3545833
## 6               Q.75%  4.2663890 4.1478470
## 7            Varianz  0.2380583 0.3404463
## 8 Standardabweichung  0.4879122 0.5834777
```

Antwort: M-H hat ein Minimum von weniger, ein Maximum von mehr. Der Durchschnittswert ist kleiner als der von M-55. Die Quantile entsprechen der grafischen Darstellung. Die 0,75 und 0,25 der zweiten sind verschoben, aber die Mediane sind ungefähr gleich.

Die Varianz der zweiten ist so viel größer, sogar viel größer als die der ersten. Dies wird folglich auch durch die Standardabweichung angezeigt, was das zuvor Gesagte bestätigt.

- Sind die Verteilungen linksschief, symmetrisch oder rechtsschief?

Antwort: Dies lässt sich mit Hilfe des Mittelwerts und des Medians ermitteln. Wenn der Median gleich dem Mittelwert ist, handelt es sich um eine symmetrische Verteilung.

Ist der Mittelwert größer als der Median, dann handelt es sich um eine rechtsschief, ansonsten um eine linksschief.

Also: In M-H und M-55 sind die Median- und Mittelwerte fast gleich, aber wenn man es genau nimmt(3.8655560<3.9101046 und 3.7168055<3.7662280), stellt sich heraus, dass beide rechtsschief sind.

- Gibt es Ausreißer in den untersuchten Daten?

Antwort: Laut BoxPlots sind sie im Fall von M-55 nicht vorhanden, aber im Fall von M-H sind sie deutlich im oberen Teil vorhanden und gehen über 1,5*IQR (oder einen Whisker) hinaus. Dargestellt in charakteristischen „Kreisen“.

- [1 Punkt] Kann man aufgrund des Boxplots im allgemeinen erkennen, ob es Ausreißer in den Daten gibt? Falls ja wie?

Antwort: Ja, mit dem Boxplot lassen sich Ausreißer leicht erkennen. Innerhalb des Rechtecks liegt der Interquartilsbereich (vom ersten Quartil Q1(oder Q0,25) bis zum dritten Quartil Q3(oder Q0,75)), und die „Whiskers“ reichen in der Regel bis 1,5 × IQR(oder Q1,5). In der Vorlesung haben wir uns speziell mit Q1,5 befasst, aber dieser Wert kann je nach Aufgabenstellung variiert und gewählt werden.

Werte, die über diese „Whisker“ hinausgehen, werden als Ausreißer betrachtet, die je nach Notation als „Kreise“ oder ähnlich bezeichnet werden, was wir brauchen, um Ausreißer zu identifizieren. Siehe Beispiel unten:

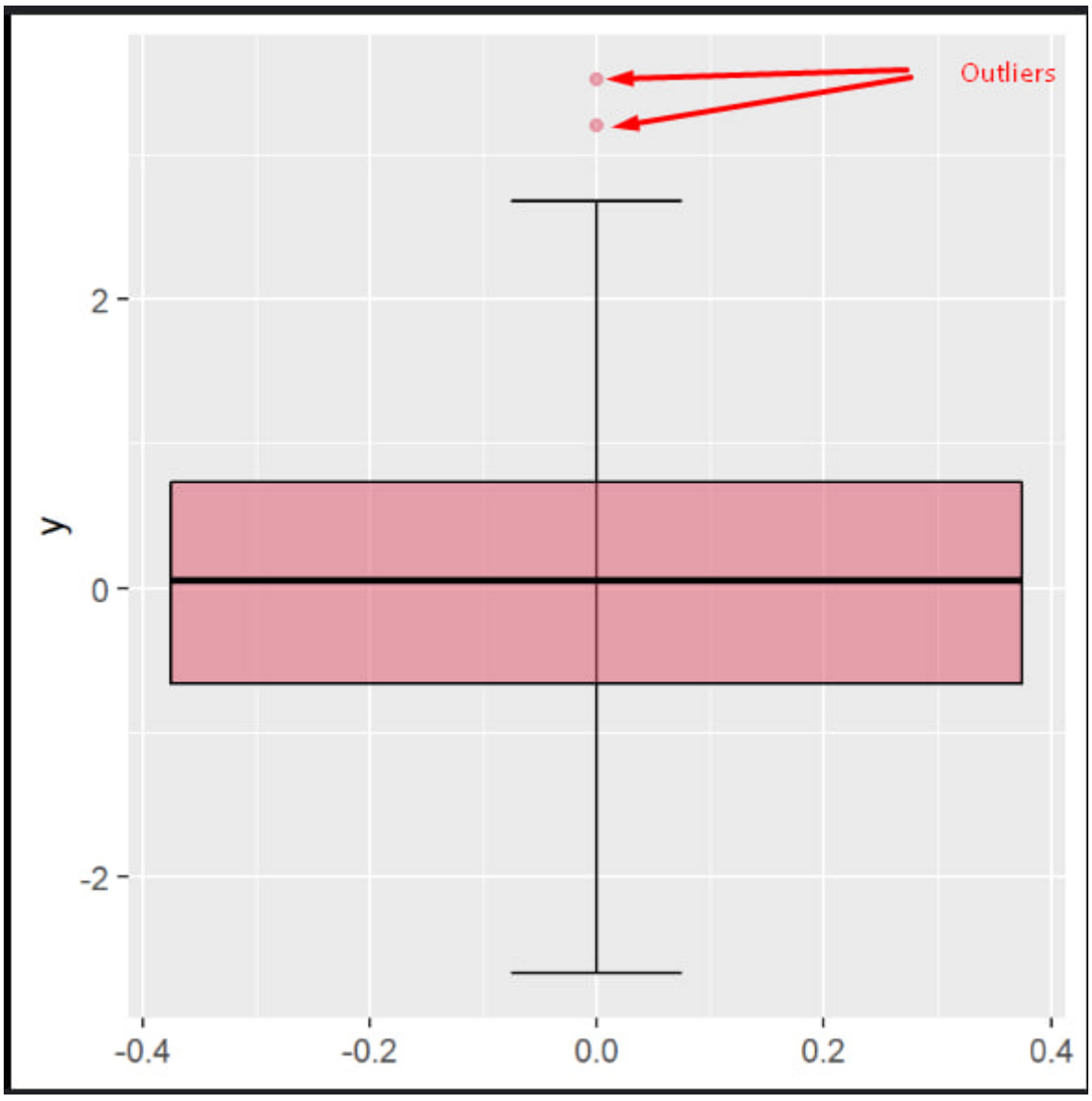


Figure 1: Quelle: Link

- [1 Punkt] Welche Funktion hat in R der Befehl head()?

Antwort: head()-Befehl (oft zusammen mit tail() betrachtet) - gibt die ersten n(default - 6) Zeilen eines Datensatzes, sprich den ersten Teil einer Matrix, eines Vektors, eines DataFrame oder einer Funktion zurück. Demonstrative Beispiel:

```r
df_test <- data.frame(
  Name = c("Volodymyr", "Max", "Otto", "Kevin", "John", "Mark", "Serhii", "Zoe"),
  City = c("Vienna", "Dnipro", "Innsbruck", "Budapest", "Bratislava", "Kyiv", "Lviv", "Praha"),
  Age = c(19, 32, 29, 45, 38, 41, 27, 36)
)
```

```r
head(df_test)
```

```
##        Name       City Age
## 1 Volodymyr     Vienna  19
## 2       Max     Dnipro  32
## 3      Otto  Innsbruck  29
## 4     Kevin   Budapest  45
## 5      John Bratislava  38
## 6      Mark       Kyiv  41
```

- [1 Punkt] Mit welchem Befehl kann in R die Schiefe berechnet werden? Was sagt sie aus?

Antwort: Ohne die Verwendung von Paketen und Bibliotheken kann die Schiefe direkt durch eine Formel unter Verwendung der Standardabweichungsfunktion in R berechnet werden, d.h. es gibt keine separate Funktion dafür:

```r
total_length <- length(df$Total)
total_mean <- mean(df$Total)
total_std <- sd(df$Total)
a <-(total_length * sum((df$Total - total_mean)^3))
b <-((total_length - 1) * (total_length - 2) * total_std^3)
total_skewness <- a/b
```

Ist die Schiefe = 0, ist die Verteilung symmetrisch. Wenn die Schiefe<0 ist, ist die Verteilung linksschief. Ist die Schiefe>0, ist die Verteilung rechtsschief.

Beispielpakete, die diese Funktion haben, sind `e1071` oder `moments`, aber sie müssen installiert werden.