Applied Statistics for Data Science

Serie 2

Aufgabe 2.1

Im Wikipedia-Artikel

https://en.wikipedia.org/wiki/Heights_of_presidents_and_presidential_candidates_of_the_United_States

sind die Körpergrössen der US-Präsidenten und ihrer Herausforderer bei den Wahlen aufgeführt. Es wurde festgestellt, dass der grössere Präsidentschaftskandidat typischerweise die Wahlen gewinnt.

In dieser Übung untersuchen wir die Daten der Präsidentschaftswahlen, seit diese im Fernsehen übertragen werden. In Tabelle 1 sind die Körpergrössen aufgeführt.

Year	Winner	Height	Opponent	Height
2020	Joe Biden	183 cm	Donald Trump	191 cm
2016	Donald Trump	191 cm	Hillary Clinton	165 cm
2012	Barack Obama	185 cm	Mitt Romney	187 cm
2008	Barack Obama	185 cm	John McCain	175 cm
2004	George W. Bush	182 cm	John Kerry	193 cm
2000	George W. Bush	182 cm	Al Gore	185 cm
1996	Bill Clinton	188 cm	Bob Dole	187 cm
1992	Bill Clinton	188 cm	George H. W. Bush	188 cm
1988	George H. W. Bush	188 cm	Michael Dukakis	173 cm
1984	Ronald Reagan	185 cm	Walter Mondale	180 cm
1980	Ronald Reagan	185 cm	Jimmy Carter	177 cm
1976	Jimmy Carter	177 cm	Gerald Ford	183 cm
1972	Richard Nixon	182 cm	George McGovern	185 cm
1968	Richard Nixon	182 cm	Hubert Humphrey	180 cm
1964	Lyndon B. Johnson	193 cm	Barry Goldwater	180 cm
1960	John F. Kennedy	183 cm	Richard Nixon	182 cm
1956	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1952	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1956	Harry S. Truman	175 cm	Thomas Dewey	173 cm

Tabelle 1: Körpergrössen der Präsidenten und ihrer Herausforderer seit 1948

Wir bilden zwei Vektoren für die entsprechenden Körpergrössen

```
winner <- c(183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Bestimmen Sie die Länge der beiden Vektoren. So können Sie auch überprüfen, ob in beiden Vektoren gleich viele Einträge sind.

b) Bestimmen Sie die Einträge 6. bis 10. Eintrag des Vektors `winner`. Verwenden Sie dazu die Klammerschreibweise `winner[...]`.

c) Bestimmen Sie den 3., 5. und 10. bis 12. Eintrag.

d) Die Washington Post hat festgestellt, dass die Einträge für Bill Clinton (7. und 8. Eintrag) zu klein sind. Er misst 189 cm. Ändern Sie die Einträge im Vektor `winner` entsprechend ab und geben Sie den neuen Vektor nochmals aus.

e) Die Behauptung ist, dass die Gewinner grösser sind als die Herausforderer. Überprüfen Sie dies, indem Sie die Mittelwerte der Vektoren miteinander vergleichen.

f) Bestimmen Sie den durchschnittlichen Grössenunterschied.

g) Bestimmen Sie noch die Varianz s^2 und die Standardabweichung s des Vektors `winner`.

h) Bestimmen Sie diese Werte noch ohne die implementierten Funktionen zu verwenden (Übung im Umgang mit `R`). Die Varianz ist definiert durch

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

Aufgabe 2.2

In einer Klasse wurden in einer Statistik-Prüfung folgende Noten geschrieben:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

a) Ändern Sie drei Noten im Datensatz so ab, dass der Median gleich bleibt, aber der Mittelwert sich stark ändert.

   Verwenden Sie dazu den `sort(...)`-Befehl.

b) Erstellen Sie zu den beiden Datensätzen einen gemeinsamen Boxplot. Was erkennen Sie?

Aufgabe 2.3

Wir haben aus eigener Erfahrung das Gefühl, dass bei Ehepaaren der Mann eher älter als die Frau ist. Nun wollen wir statistisch untersuchen, ob dem so ist.

In einer Untersuchung in England wurden das Alter (in Jahren) und die Körpergrösse (in cm) von 170 Ehepaaren untersucht.

a) Lesen Sie die Datei **mannfrau.csv** ein.

b) Erstellen Sie einen Boxplot für die *Differenz* des Alters zwischen Ehemännern und Ehefrauen.

c) Interpretieren Sie im Boxplot den Median und die Quartile. Was können Sie über die Ausreisser sagen?

Aufgabe 2.4

In dieser Aufgabe geht es darum, dass Sie weitere R-Befehle kennenlernen und den Umgang mit R üben.

Wir verwenden den Datensatz **InsectSprays**, der in R enthalten ist.

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Dabei wurden 6 verschiedene Insektensprays verwendet, die auf verschiedenen Feldern versprüht wurden. Danach wurde die Anzahl Insekten gezählt, die sich auf dem entsprechenden Feld nach dem Besprühen befanden. (Beall, G., (1942) The Transformation of data from entomological field experiments, Biometrika, 29, 243–262.)

a) Wir wollen zunächst die Mittelwerte der einzelnen Sprays bestimmen. Dazu verwenden wir den R-Befehl **tapply(...)**

```
tapply(InsectSprays[, "count"], InsectSprays[, "spray"], FUN = mean)
```

```
##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

Dieser Befehl wendet (apply) die Funktion (`FUN`) Mittelwert (`mean`) auf die Spalte `InsectSprays[,"count"]` an und ordnet diese nach der Spalte Spray (`InsectSprays[, "spray"]`). Das heisst, es werden die Mittelwerte für `count` für die Sprays $A, B, \ldots, F$ gebildet.

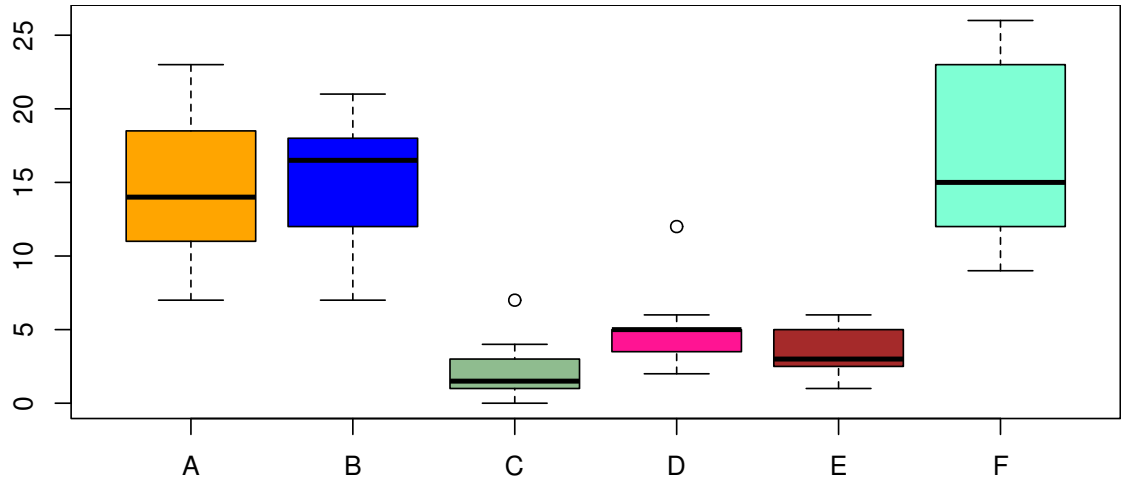
Die Mittelwerte sind sehr unterschiedlich. Die Sprays C, D und E sind scheinen wesentlich effizienter zu sein als die Sprays A, B und F.

Etwas einfacher geht auch folgende Schreibweise:

```
tapply(InsectSprays$count, InsectSprays$spray, mean)

##         A         B         C         D         E         F
## 14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

b) Wir wollen nun noch einen Boxplot der Daten machen. Da die Daten nach der Spalte `Spray` geordnet wird, verlangt `R` die Eingabe `boxplot(y ~ x)`, wobei `y` die Werte sind von denen `R` die Boxplot nehmen soll und `x` die Namen, nach denen die Werte geordnet werden sollen.

```
boxplot(count ~ spray,
        data = InsectSprays,
        col=c("orange", "blue", "darkseagreen", "deeppink",
              "brown", "aquamarine")
)
```



Auch hier ist offensichtlich, dass die Sprays C, D und E wesentlich effizienter erscheinen zu sein als die Sprays A, B und F.

Aufgabe 2.5

In der Datei `Diet.csv` sind 76 Personen aufgelistet, die jeweils einer der Diäten 1,2 oder 3 für 6 Wochen machten.

```
##   Person gender Age Height pre.weight Diet weight6weeks
## 1     25      0  41    171         60    2         60.0
## 2     26      0  32    174        103    2        103.0
## 3      1      0  22    159         58    1         54.2
## 4      2      0  46    192         60    1         54.0
## 5      3      0  55    170         64    1         63.3
## 6      4      0  33    171         64    1         61.1
```

```
diet <- read.csv("...Diet.csv")

head(diet)
```

In der Datei ist das Gewicht `pre.weight` vor der Diät und das Gewicht `weight6weeks` nach 6 Wochen aufgeführt. Wir interessieren uns für den Gewichtsverlust. Dazu führen wir zu der Datei eine Spalte `weight.loss` hinzu. Dies geht folgendermassen:

```
diet$weight.loss <- diet$weight6weeks - diet$pre.weight

head(diet)


##   Person gender Age Height pre.weight Diet weight6weeks weight.loss
## 1     25      0  41    171         60    2         60.0         0.0
## 2     26      0  32    174        103    2        103.0         0.0
## 3      1      0  22    159         58    1         54.2        -3.8
## 4      2      0  46    192         60    1         54.0        -6.0
## 5      3      0  55    170         64    1         63.3        -0.7
## 6      4      0  33    171         64    1         61.1        -2.9
```

`R` versteht `diet$weight.loss` automatisch als neue Spalte und fügt die der Tabelle hinzu.

Führen Sie nun die Teilaufgaben in der Aufgabe vorher für `weight.loss` und `Diet` durch. Interpretieren Sie jeweils die Resultate.

Applied Statistics for Data Science

Musterlösungen zu Serie 2

Lösung 2.1

a) Wir bilden zwei Vektoren für die entsprechenden Körpergrössen

```
winner <- c(183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Länge der beiden Vektoren

```
length(winner)

## [1] 19

length(opponent)

## [1] 19
```

b) Einträge 6. bis 10. Eintrag des Vektors `winner`

```
winner[6:10]

## [1] 182 188 188 188 185
```

c) 3., 5. und 10. bis 12. Eintrag.

```
winner[c(3, 5, 10:12)]

## [1] 185 182 185 185 177
```

d) 6. und 7. Eintrag ändern:

```
winner[7] <- 189
winner[8] <- 189

# or winner[c(7,8)] <- 189

winner

##  [1] 183 191 185 185 182 182 189 189 188 185 185 177 182 182 193 183
## [17] 179 179 175
```

e) Mittelwerte der Vektoren miteinander vergleichen

```
mean(winner)
```

```
## [1] 183.8947

mean(opponent)

## [1] 181.0526
```

Der Gewinner der Wahl ist also durchschnittlich etwa 3.5 cm grösser.

f) Durchschnittliche Differenz

```
diff <- winner - opponent

mean(diff)

## [1] 2.842105
```

g) Varianz s^2 und die Standardabweichung s des Vektors `winner`.

```
var(winner)

## [1] 22.09942

sd(winner)

## [1] 4.701002
```

h) Ohne die implementierten Funktionen zu verwenden.

```
winner.var <- sum((winner - mean(winner))^2) / (length(winner)-1)

winner.var

## [1] 22.09942

winner.sd <- sqrt(winner.var)

winner.sd

## [1] 4.701002
```

Lösung 2.2

a) Der ursprüngliche Datensatz hat für den Median und Mittelwert folgende Werte:

```
noten_1 <- c(4.2,2.3,5.6,4.5,4.8,3.9,5.9,2.4,5.9,
             6,4,3.7,5,5.2,4.5,3.6,5,6,
             2.8,3.3,5.5,4.2,4.9,5.1)

median(noten_1)

## [1] 4.65

mean(noten_1)
```

```
## [1] 4.5125
```

Zuerst ordnen wir die Datenwerte der Grösse nach:

```
sort(noten_1)

##  [1] 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 4.2 4.2 4.5 4.5 4.8 4.9 5.0 5.0
## [17] 5.1 5.2 5.5 5.6 5.9 5.9 6.0 6.0
```

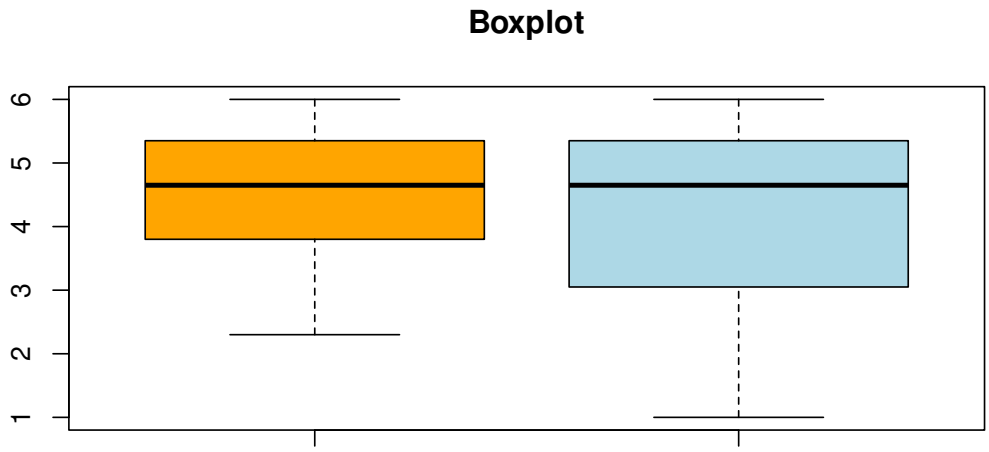
Da die Anzahl Noten gerade ist, wird der Median aus dem Mittelwert von $x_{(12)}$ und $x_{(13)}$ gebildet. Wenn wir also Noten kleiner als $x_{(12)}$ abändern, wird sich der Median nicht ändern. Dementsprechend ändern wir die Notenwert $x_{(9)}, x_{(10)}, x_{(11)}$ zu einer eins. Dies lässt den Median unverändert, lässt den Mittelwert aber maximal schrumpfen.

$$2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 1, 3.6, 5, 6, 2.8, 3.3, 5.5, 1, 4.9, 5.1$$

```
noten_2 <- c(1,2.3,5.6,4.5,4.8,3.9,5.9,
             2.4,5.9,6,4,3.7,5,5.2,1,3.6,5,6,
             2.8,3.3,5.5,1,4.9,5.1)

median(noten_2)

## [1] 4.65

mean(noten_2)

## [1] 4.1
```

b) Die Plots

```
boxplot(noten_1,
        noten_2,
        main = "Boxplot",
        col = c("orange", "lightblue"))
```



Der Median bleibt in der Tat gleich. Die Box wird breiter, da extreme Werte hinzukommen.

Lösung 2.3

a) Einlesen

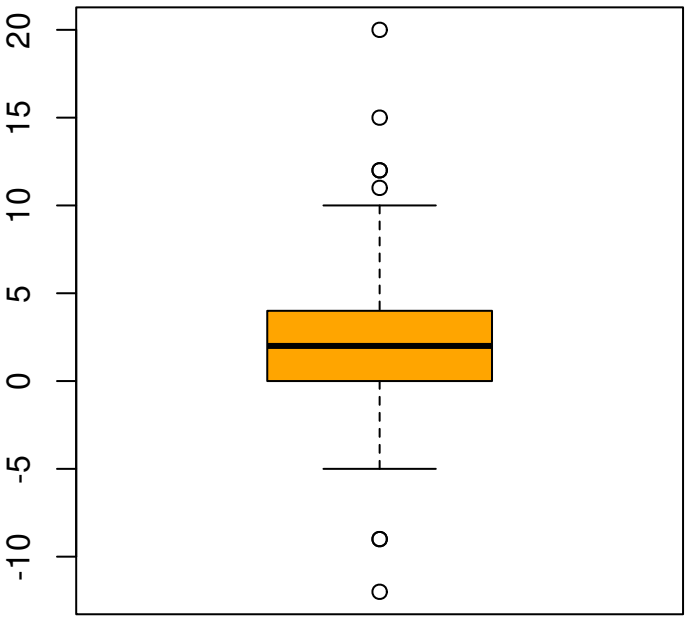
```
mannfrau <- read.csv(".../mannfrau.csv")
```

Überprüfen mit `head(...)`, ob Datei richtig eingelesen wurde:

```
head(mannfrau)

##   alter.mann groesse.mann alter.frau groesse.frau
## 1         49          180         43          159
## 2         25          184         28          156
## 3         40          165         30          162
## 4         52          177         57          154
## 5         58          161         52          142
## 6         32          169         27          166
```

b) Code:

```
alter.mann <- mannfrau[, 1]
alter.frau <- mannfrau[, 3]

boxplot(alter.mann - alter.frau,
        col = "orange")
```



c) Der Median ist etwa 2. Somit ist die Alterdifferenz bei der Hälfte der Ehepaaren kleiner als zwei und die andere Hälfte grösser als 2.

Das untere Quartil ist bei ungefähr 0, d. h. bei 25 % aller untersuchten Ehepaare ist die Frau älter als der Mann.

Das obere Quartil bei 5, d. h. bei 25 % aller untersuchten Ehepaare ist der Mann mehr als 5 Jahre älter als die Frau.

Die Hälfte aller Ehepaare hat also ein Altersunterschied (Mann älter als Frau) zwischen 0 und 5 Jahren.

Der maximale Unterschied ist 20 Jahre und das Minimum -10. In einem Fall war also die Frau mehr als 10 Jahre älter als der Mann.

Lösung 2.4

Lösung 2.5

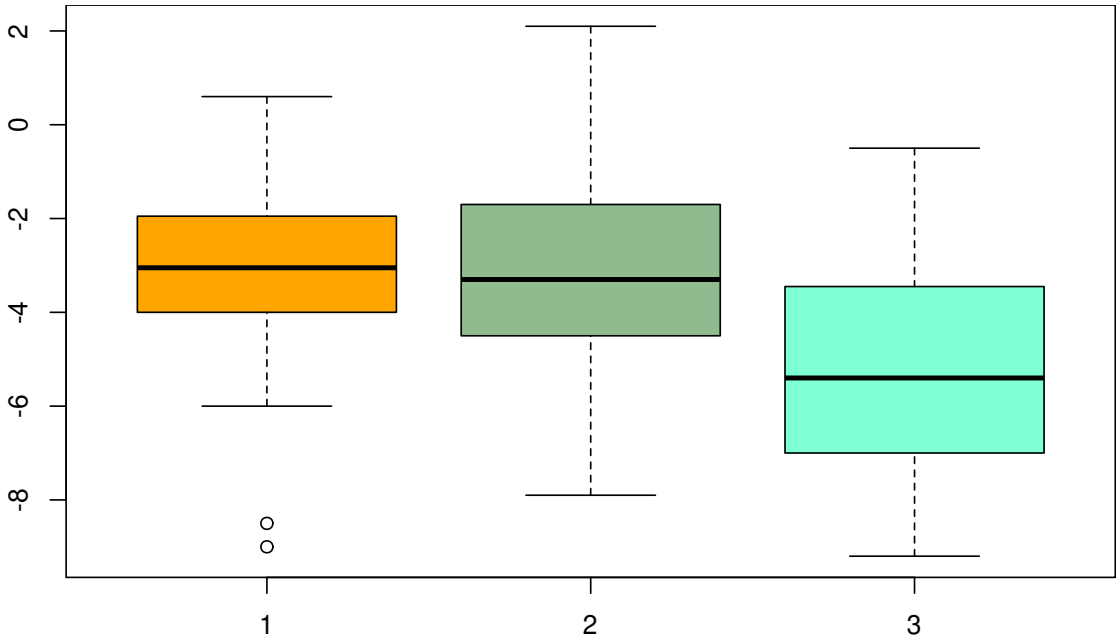
a) Gruppenmittel:

```
tapply(diet$weight.loss, diet$Diet, mean)

##         1         2         3
## -3.300000 -3.025926 -5.148148
```

Die Diäten 1 und 2 führen zu einem durchschnittlichen Gewichtsverlust von etwa 3 Kilogramm. Diät 5 hingegen ist der durchschnittliche Gewichtsverlust 5 Kilogramm.

b) Graphische Darstellung durch Boxplot:

```
boxplot(weight.loss ~ Diet,
        data = diet,
        col = c("orange", "darkseagreen", "aquamarine")

)
```



Boxplot bestätigt Vermutung aus Teilaufgabe a).