

Applied Statistics for Data Science

Serie 4

Aufgabe 4.1 in R

Bestimmen Sie die Korrelationskoeffizienten der Aufgaben in Serie 3 und interpretieren Sie die Resultate und vergleichen Sie ihre Interpretation mit dem Streudiagramm.

Aufgabe 4.2 in R

- Erzeugen Sie den Vektor `t.x` mit den Werten $-10, -9, \dots, 9, 10$ und den Vektor `t.x1` mit den Werten $0, 1, \dots, 9, 10$. Erzeugen Sie dann die Vektoren `t.y` und `t.y1`, deren Elemente die Quadratwerte der entsprechenden Elemente von `t.x` bzw. `t.x1` enthalten.
- Zeichnen Sie die Streudiagramme `t.y` vs. `t.x` und `t.y1` vs `t.x1`. Benützen Sie die `R`-Funktion

`plot()`

- Berechnen Sie die Korrelationskoeffizienten zwischen `t.x` und `t.y` bzw. zwischen `t.x1` und `t.y1`. Benützen Sie die `R`-Funktion

`cor()`

Warum sind die beiden Korrelationen so verschieden?

Aufgabe 4.3

Wo steckt in den folgenden Aussagen der Fehler? Begründen Sie.

- Bei einer gezinkten Münze wurde festgestellt, dass $P(\text{Kopf}) = 0.32$ und $P(\text{Zahl}) = 0.73$.
- Die Wahrscheinlichkeit für einen Sechser im Zahlenlotto ist $-3 \cdot 10^{-6}$.
- Bei einer Befragung wurden die Ereignisse untersucht. Man findet $P(S) = 0.1$, $P(M) = 0.5$ und $P(S \cup M) = 0.7$

- S: Befragte Person ist schwanger.
M: Befragte Person ist männlich.

Aufgabe 4.4

Bei einem Zufallsexperiment werden ein roter und ein blauer Würfel gleichzeitig geworfen. Wir nehmen an, dass sie „fair“ sind, d. h. die Augenzahlen 1 bis 6 eines Würfels treten mit gleicher Wahrscheinlichkeit auf.

- Beschreiben Sie den Ereignisraum in Form von Elementarereignissen.
- Wie gross ist die Wahrscheinlichkeit eines einzelnen Elementarereignisses?
- Berechnen Sie die Wahrscheinlichkeit, dass das Ereignis E_1 „Die Augensumme ist 7“ eintritt.
- Wie gross ist die Wahrscheinlichkeit, dass das Ereignis E_2 „Die Augensumme ist kleiner als 4“ eintritt.
- Bestimmen Sie $P(E_3)$ für das Ereignis E_3 „Beide Augenzahlen sind ungerade“.
- Berechnen Sie $P(E_2 \cup E_3)$.

Aufgabe 4.5

Die Ereignisse A und B seien unabhängig mit Wahrscheinlichkeiten $P(A) = 3/4$ und $P(B) = 2/3$. Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

- Beide Ereignisse treten ein.  $P(A \cap B) = P(A) \cdot P(B)$
- Mindestens eines von beiden Ereignissen tritt ein.  $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$
- Höchstens eines von beiden Ereignissen tritt ein.  $1 - P(A \cap B)$
- Keines der beiden Ereignisse tritt ein.  $P(A \cup B) = 1 - P(A \cap B) = 1 - (P(A) + P(B) - P(A) \cdot P(B))$
- Genau eines der Ereignisse tritt ein.  $= P(A \cup B) - P(A \cap B)$
 $= P(A) + P(B) - 2 P(A) \cdot P(B)$

Aufgabe 4.6

Der Einsturz eines Gebäudes in Tokio kann durch zwei voneinander unabhängige Ereignisse verursacht werden.

- E_1 : ein grosses Erdbeben
- E_2 : ein starker Taifun

Die jährlichen Eintrittswahrscheinlichkeiten dieser beiden Ereignisse sind $P(E_1) = 0.04$ und $P(E_2) = 0.08$.

Berechnen Sie die jährliche Einsturzwahrscheinlichkeit des Gebäudes.

Aufgabe 4.3

a) $P(\text{Kopf}) = 0.52$ und $P(\text{Zahl}) = 0.73$

$$P(\text{Kopf}) + P(\text{Zahl}) = 0.52 + 0.73 = 1.05 \neq 1$$

b) Würheit muss $P(A) \geq 0$

c) $P(S) = 0.1$, $P(M) = 0.5$

$$P(S \cup M) = 0.7 \neq 0.6 \quad \text{da} \quad P(S \cup M) = P(S) + P(M)$$

falls $P(S) \cap P(M) = \emptyset$, trifft zu da
Ereignisse stochastisch unabhängig sind

Aufgabe 4.4

a) $\Omega = \{(1,1), (1,2), \dots, (6,6)\}$ $|\Omega| = 36$

b) Laplace: $\frac{1}{|\Omega|} = \frac{1}{36}$

Anzahl günstige Fälle
1

c) $E_1 = \{(3,4), (4,3), (5,2), (2,5), (6,1), (1,6)\}$ $|E_1| = 6$

$$\frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$$

günstige
mögliche

d) $E_2 = \{(1,1), (2,1), (1,2)\}$ $|E_2| = 3$

$$\frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$$



e) E_3 - beide Augenzahlen ungerade

$$\frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$$

f) $P(E_2 \cup E_3) = P(E_2) + P(E_3) - P(E_2 \cap E_3)$ 3

$$= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} = \frac{11}{36}$$

Aufgabe 4.5

- a) $P(A) \cap P(B) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{2}{3} = \frac{6}{12} = \frac{1}{2}$
- b) $P(A) + P(B) - P(A \cap B) = P(A) + P(B) - (P(A) \cdot P(B))$
 $= \frac{3}{4} + \frac{2}{3} - \frac{1}{2} = \frac{11}{12}$
- c) $1 - P(A \cap B) = 1 - P(A) \cdot P(B)$
 $= 1 - \frac{1}{2} = \frac{1}{2}$
- d) $\overline{P(A \cup B)} = 1 - P(A \cup B) = 1 - \frac{11}{12} = \frac{1}{12}$
- e) $P(A \cup B) - P(A \cap B) = P(A) + P(B) - 2 P(A) \cdot P(B)$
 $= \frac{5}{12}$

Aufgabe 4.6

weil stochastisch unabhängig

$$P(E_1) + P(E_2) - \underbrace{P(E_1) \cdot P(E_2)}_{= 0.0082} = 0.04 + 0.08 - 0.0082 = \underline{\underline{0.1168}}$$

Applied Statistics for Data Science

Musterlösungen zu Serie 4

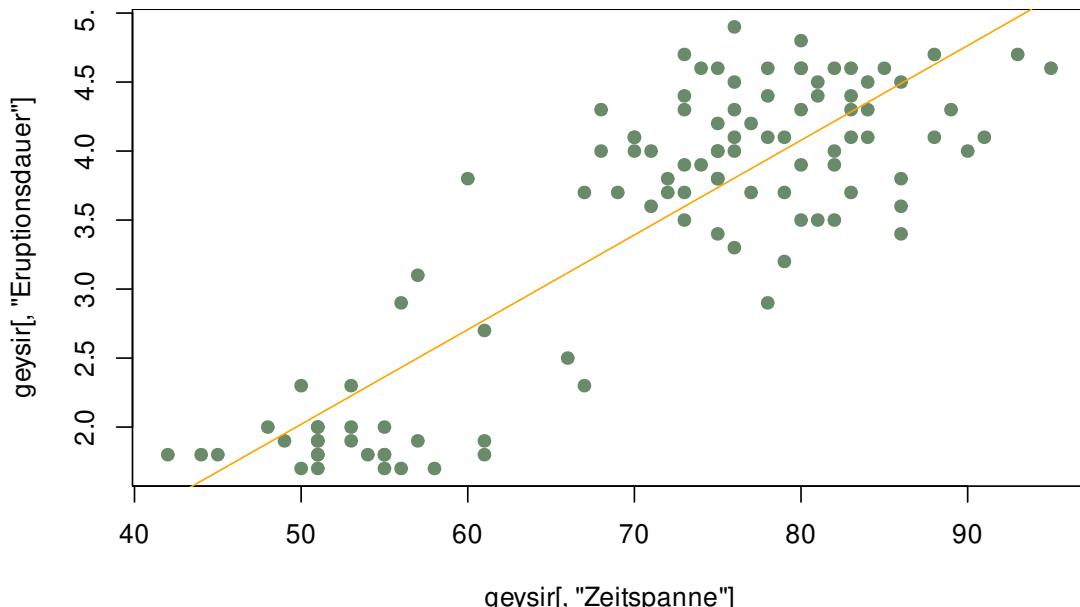
Lösung 4.1

a) Old Faithful

```
# Datensatz einlesen
geysir <- read.table("./Daten/geysir.dat", header = TRUE)

plot(geysir[, "Zeitspanne"], geysir[, "Eruptionsdauer"],
     pch=19,
     col="darkseagreen4")

abline(lm(geysir[, "Eruptionsdauer"] ~ geysir[, "Zeitspanne"]),
       col="orange")
```



```
cor(geysir[, "Zeitspanne"], geysir[, "Eruptionsdauer"])

## [1] 0.8584273
```

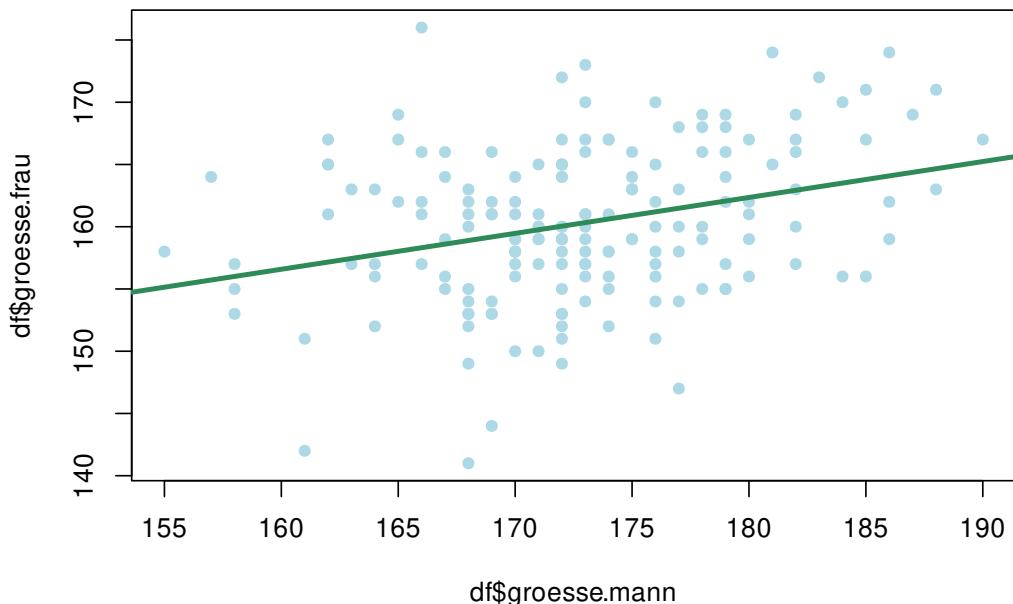
Der Korrelationskoeffizient ist mit 0.85 nahe bei 1. Somit ist die Punktwolke steigend und annähernd linear. Dies stimmt in etwa mit dem Streudiagramm überein.

b) Größenvergleich von Ehepaaren:

```
df <- read.csv("~/Dropbox/Statistics/Themen/Deskriptive_Statistik/Uebungen")

plot(df$groesse.mann, df$groesse.frau, col="lightblue", pch=16)

abline(lm(df$groesse.frau~df$groesse.mann), lwd=3, col="seagreen")
```



```
cor(df$groesse.mann, df$groesse.frau)
## [1] 0.3080731
```

Der Korrelationskoeffizient ist 0.308, somit positiv und die Punktwolke steigt auch. Allerdings ist er nicht eher nahe bei 0 und somit ist ein linearer Zusammenhang eher fraglich. Im Streudiagramm ist erkennbar, dass sehr verstreut ist und kein eindeutiges lineares Muster erkennbar ist.

c) Einkommen

```
income <- read.table(file="./Daten/income.dat", header=TRUE)

iq <- income[, 1]

anzahl.jahre.schule <- income[, 2]

einkommen <- income[, 3]

plot(iq,
     einkommen,
```

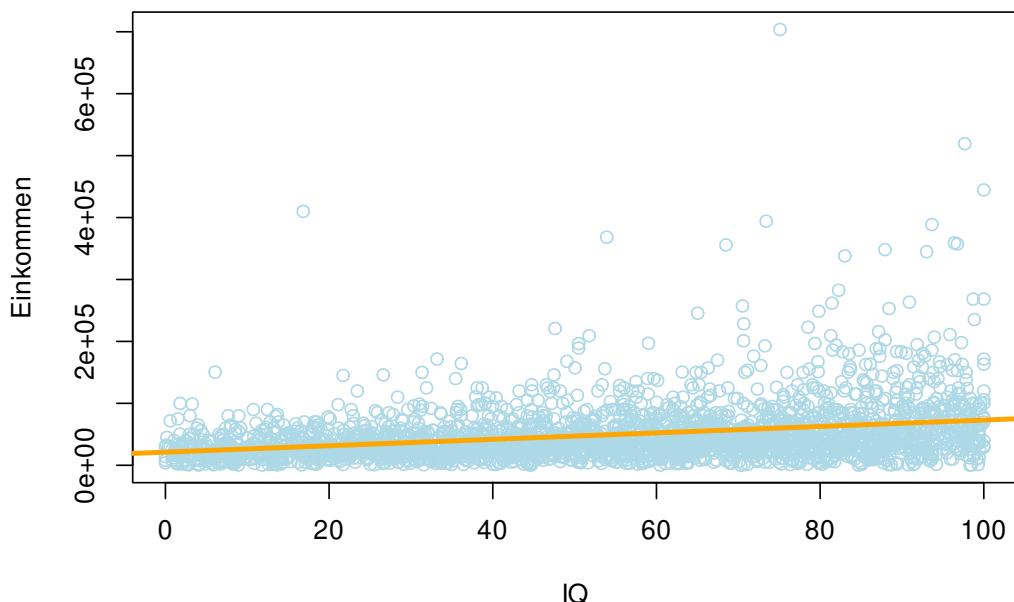
```

    type = "p",
    xlab = "IQ",
    ylab = "Einkommen",
    col = "lightblue"

)

abline(lm(einkommen ~ iq),
       col = "orange",
       lwd = 3)

```



```

cor(iq, einkommen)
## [1] 0.3081529

```

Der Korrelationskoeffizient ist 0.308, somit positiv und die Punktwolke steigt auch. Allerdings ist er nicht eher nahe bei 0 und somit ist ein linearer Zusammenhang eher fraglich. Im Streudiagramm ist erkenntlich, dass sehr verstreut ist und kein eindeutiges lineares Muster erkennbar ist.

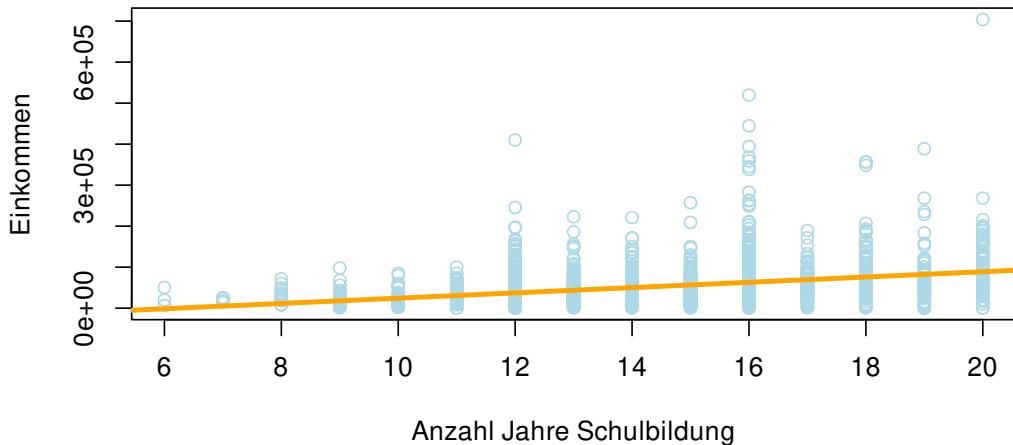
```

plot(anzahl.jahre.schule,
      einkommen,
      type= "p",
      xlab = "Anzahl Jahre Schulbildung",
      ylab= "Einkommen",
      col = "lightblue"

)

```

```
abline(lm(einkommen ~ anzahl.jahre.schule),
       col = "orange",
       lwd = 3)
```



```
cor(anzahl.jahre.schule, einkommen)
## [1] 0.3456474
```

Da der Korrelationskoeffizient relativ klein ist, scheint ein Modell beruhend auf einem linearen Zusammenhang zwischen Einkommen und Anzahl Jahre Schulbildung nicht angebracht zu sein.

d) Anscombe:

```
data(anscombe)
reg <- lm(anscombe$y1 ~ anscombe$x1)
reg2 <- lm(anscombe$y2 ~ anscombe$x2)
reg3 <- lm(anscombe$y3 ~ anscombe$x3)
reg4 <- lm(anscombe$y4 ~ anscombe$x4)

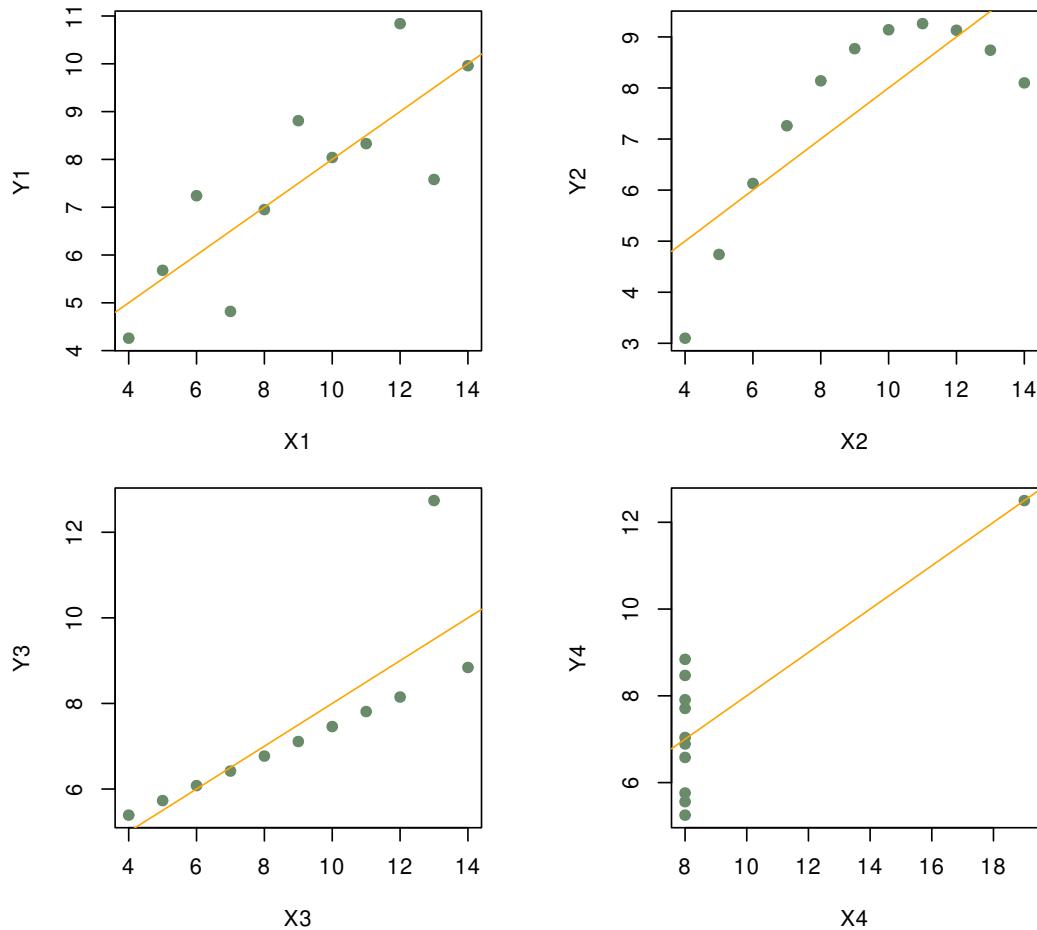
par(mfrow = c(2, 2))

plot(anscombe$x1, anscombe$y1,
      ylab = "Y1", xlab = "X1",
      col = "darkseagreen4", pch = 19)
abline(reg, col = "orange")

plot(anscombe$x2, anscombe$y2,
      ylab = "Y2", xlab = "X2",
      col = "darkseagreen4", pch = 19)
abline(reg2, col = "orange")
```

```
plot(anscombe$x3, anscombe$y3,
      ylab = "Y3", xlab = "X3",
      col = "darkseagreen4", pch = 19)
abline(reg3, col = "orange")

plot(anscombe$x4, anscombe$y4,
      ylab= "Y4", xlab= "X4",
      col = "darkseagreen4", pch = 19 )
abline(reg4, col = "orange")
```



```
cor(anscombe$x1, anscombe$y1)
## [1] 0.8164205

cor(anscombe$x2, anscombe$y2)
## [1] 0.8162365

cor(anscombe$x3, anscombe$y3)
```

```
## [1] 0.8162867
cor(anscombe$x4, anscombe$y4)
## [1] 0.8165214
```

Die Korrelationskoeffizienten sind bis auf die dritte Stelle nach dem Komma gleich und mit 0.816 recht nahe bei 1.

Allerdings macht der Korrelationskoeffizient für die verschiedenen Streudiagramme ganz unterschiedlich interpretierbar.

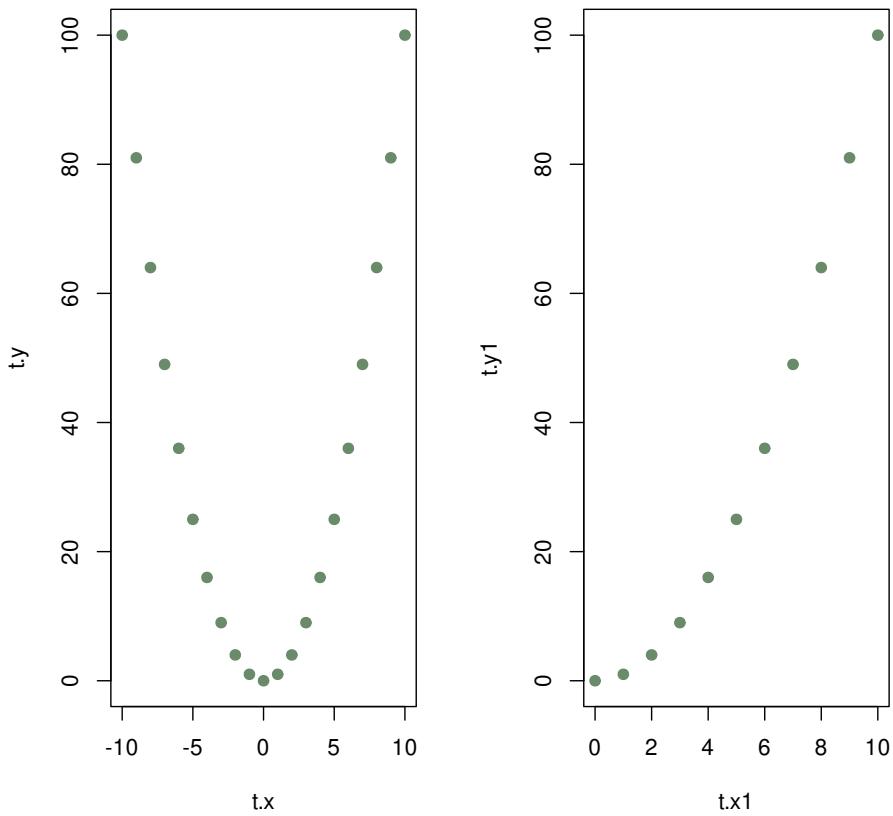
- Oben links: Das ist der „Normalfall“. Die Punkte folgen der Regressionsgerade gut und weichen nur wenig von dieser ab.
- Oben rechts: Hier haben wir eine klare Abhängigkeit von X2 und Y2, aber diese ist nicht linear, sondern quadratisch. Dies hat einen Einfluss auf den Korrelationskoeffizienten, der nur *lineare* Abhängigkeit erkennt.
- Unten links: Dieser Datensatz hat einen Ausreißer und alle anderen Punkte befinden sich auf einer Geraden. Dieser Ausreißer hat einen grossen Einfluss auf den Korrelationskoeffizienten.
- Unten rechts: Dies ist ein „Freak“-Datensatz. Der Korrelationskoeffizient ist hier nicht auch vernünftige Weise interpretierbar, er macht keinen Sinn.

Lösung 4.2

- a) Erzeugen der Vektoren:

```
t.x <- (-10):10
t.x1 <- 0:10
t.y <- t.x^2
t.y1 <- t.x1^2
```

- b)
- ```
par(mfrow=c(1,2)) # zwei Grafiken im Grafikfenster
plot(t.x, t.y, col = "darkseagreen4", pch = 19)
plot(t.x1, t.y1, col = "darkseagreen4", pch = 19)
```



```
c) cor(t.x, t.y)
[1] 0
cor(t.x1, t.y1)
[1] 0.9631427
```

Die Korrelation zwischen **t.x** und **t.y** ist 0, weil die Daten symmetrisch zur  $y$ -Achse liegen. Im zweiten Fall ist die Korrelation hoch (0.96), obwohl die Daten keine lineare Beziehung aufweisen. Der Grund dafür ist, dass  $x$  und  $y$  monoton steigen.

### Lösung 4.3

- Da Zahl und Kopf die möglichen Elementarereignisse sind, müsste die Summe deren Wahrscheinlichkeiten 1 sein. Dies ist hier aber nicht der Fall:  $P(\Omega) = P(\text{Zahl}) + P(\text{Kopf}) = 1.05$ . (Axiom 2 ist verletzt.)
- Die genannte Wahrscheinlichkeit ist negativ. (Axiom 1 ist verletzt.)

- c) Es gilt  $S \cap M = \emptyset$  und darum müsste  $P(S) + P(M) = P(S \cup M)$  wegen Axiom 3. Dies ist hier aber nicht erfüllt.

### Lösung 4.4

- a)  $\Omega = \{(1,1), (1,2), \dots, (1,6), (2,1), (2,2), \dots, (2,6), \dots, (6,6)\}, |\Omega| = 36$
- b)  $P(\text{Elementarereignis}) = \frac{1}{\Omega} = \frac{1}{36}$
- c)  $E_1 = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$  Anzahl günstige Fälle:  $|E_1| = 6$   
Anzahl mögliche Fälle:  $|\Omega| = 36$   $P(E_1) = \frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$
- d)  $E_2 = \{(1,1), (2,1), (1,2)\}; P(E_2) = \frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$
- e)  $E_3 = \{(1,1), (1,3), (1,5), (3,1), (3,3), (3,5), (5,1), (5,3), (5,5)\}; P(E_3) = \frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$
- f) Mit dem Additionssatz:

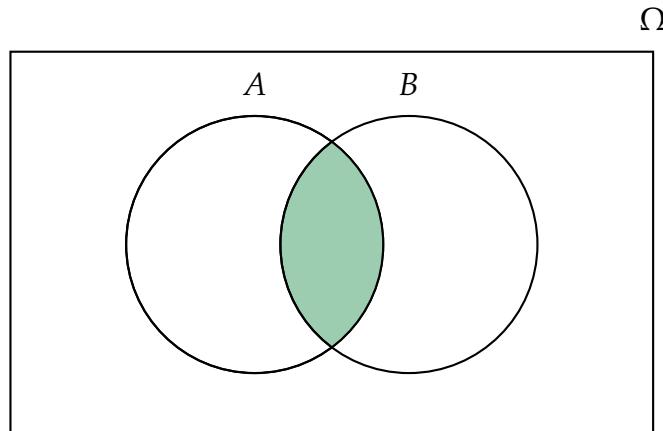
$$\begin{aligned}
 P(E_2 \cup E_3) &= P(E_2) + P(E_3) - P(E_2 \cap E_3) \\
 &= P(E_2) + P(E_3) - P(\{(1,1)\}) \\
 &= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} \\
 &= \frac{11}{36}.
 \end{aligned}$$

### Lösung 4.5

A  $\leftarrow 3/4$

B  $\leftarrow 2/3$

a)

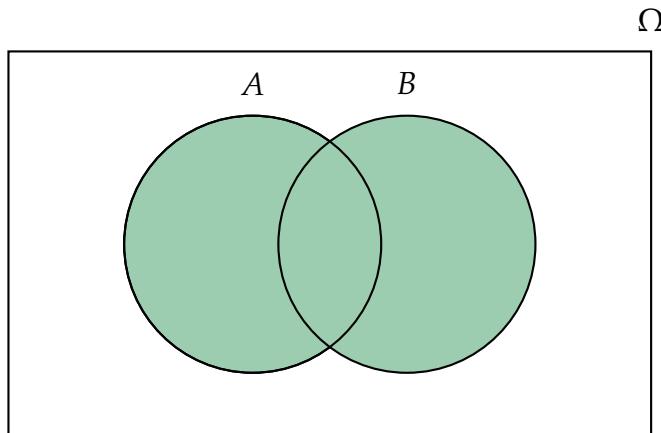


$$P(\text{beide Ereignisse}) = P(A \cap B) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{2}{3} =$$

```
library(MASS)
fractions(A * B)

[1] 1/2
```

b)

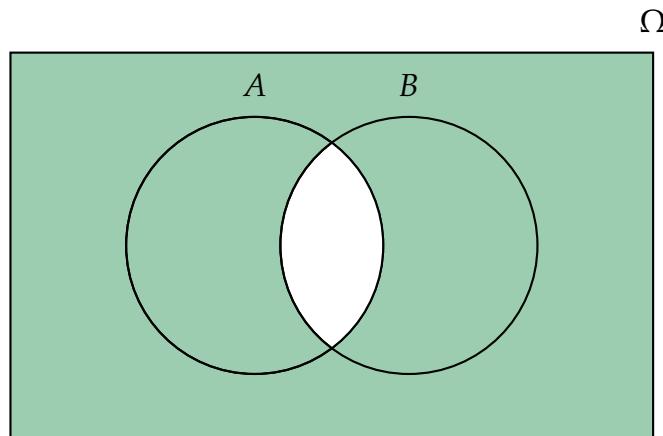


$$\begin{aligned} P(\text{mindestens eines}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A) \cdot P(B) \\ &= \end{aligned}$$

```
fractions(A + B - A*B)

[1] 11/12
```

c)

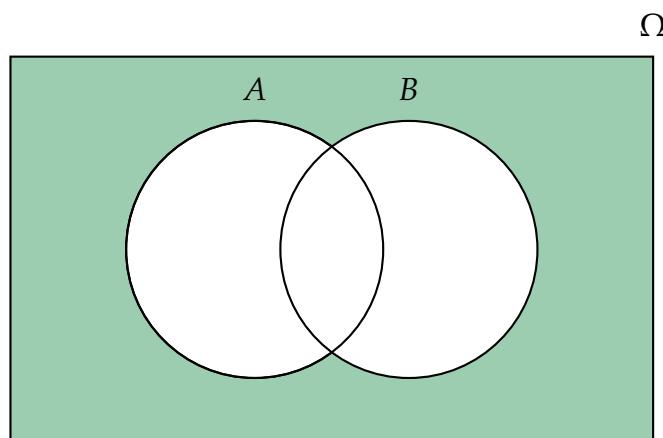


$$P(\text{höchstens eines}) = 1 - P(A \cap B) = 1 - P(A) \cdot P(B)$$

```
fractions(1 - A*B)
```

```
[1] 1/2
```

d)

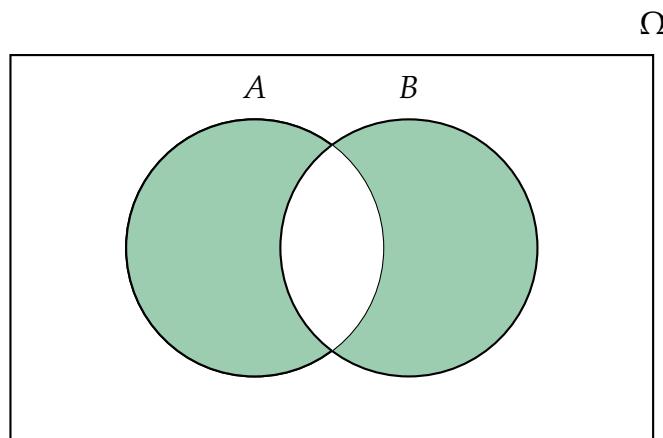


$$\begin{aligned} P(\text{kein Ereignis}) &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B) - P(A) \cdot P(B)) \\ &= \end{aligned}$$

```
fractions(1 - (A + B - A*B))
```

```
[1] 1/12
```

e)



$$\begin{aligned}
 P(\text{genau ein Ereignis}) &= P(A \cup B) - P(A \cap B) \\
 &= P(A) + P(B) - 2P(A) \cdot P(B) \\
 &=
 \end{aligned}$$

```
fractions (A + B - 2*A*B)
```

```
[1] 5/12
```

## Lösung 4.6

Die jährliche Einsturz wahrscheinlichkeit  $E_1 \cup E_2$  lässt sich wie folgt berechnen:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.04 + 0.08 - 0.04 \cdot 0.08 = 0.1168,$$

wobei  $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$ , da  $E_1$  und  $E_2$  unabhängig sind.