Übungsaufgaben Kapitel 4
Lösungen

Prof. Dr. Torben Kuhlenkasper

Aufgabe 1

In der Fußballbundesliga erhielt die siegreiche Mannschaft ab der Saison 95/96 drei anstatt zwei Punkte für einen Sieg.

Die folgende Tabelle gibt für 10 Jahre die Anzahl der pro Saison geschossenen Tore an.

Saison	Anzahl Tore
90/91	886
91/92	994
92/93	892
93/94	895
94/95	918
95/96	831
96/97	911
97/98	883
98/99	866
99/00	885

1. Was sind hier die beiden Merkmale A und B?

   Merkmal A: Punkteregelung

   Merkmal B: Anzahl Tore

2. Welche Merkmalsausprägungen haben sie?

   Ausprägung A_1 = alte Punkteregelung (bis Saison 94/95)

   Ausprägung A_2 = neue Punkteregelung (ab Saison 95/96)

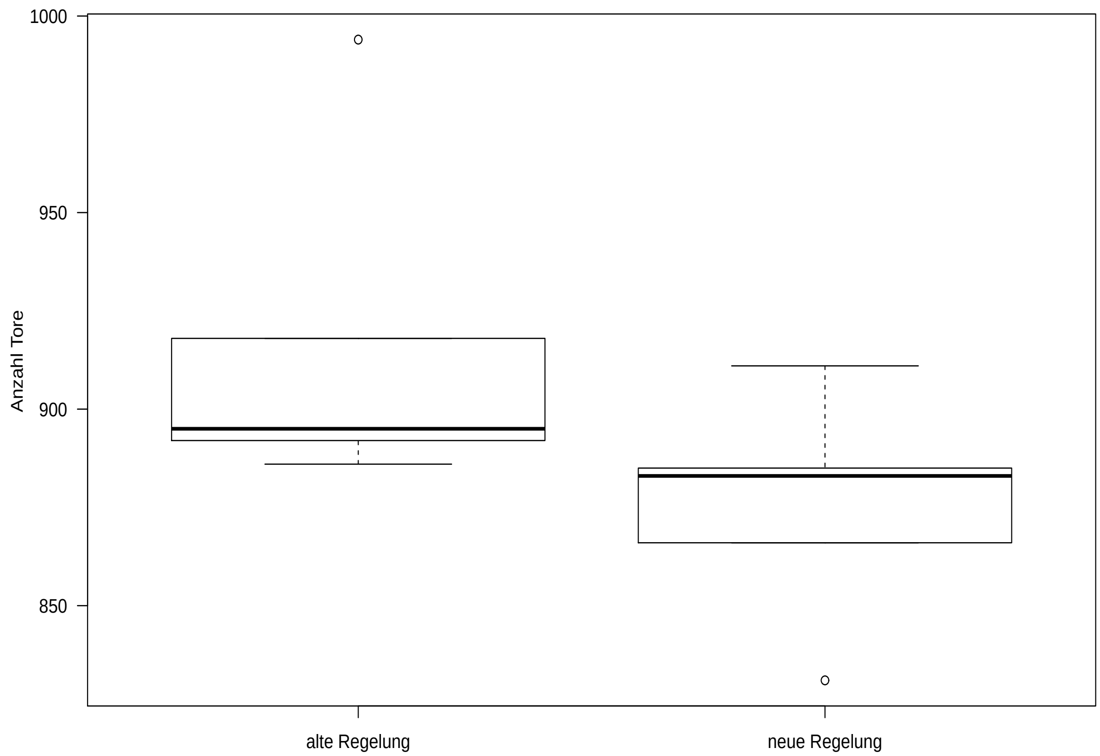
   Merkmal B ist ein quantitatives, diskretes Merkmal

3. Untersuchen Sie, ob die neue Regelung dazu geführt hat, dass sich die Verteilung der Tore verändert hat.

   zunächst: Vergleich der Kennzahlen für beide Gruppen von Merkmal A:

	Mittelwert	Median	Varianz
alte Regelung	917	895	2000
neue Regelung	875.2	883	869.2

Vergleich der beiden Boxplots:



Sowohl die Maßzahlen als auch die beiden Boxplots deuten daraufhin, dass die Anzahl der Tore durch die neue Regelung geringer geworden ist. Auch die Streuung ist geringer geworden.

Aufgabe 2

Die Richterin Bettina Freimund analysierte die Gefangenenakten von 103 Wirtschaftsstraftätern und 113 alternden Dieben, um herauszufinden, 'wie es um den Resozialisierungsgedanken im Strafvollzug bestellt ist'. Sie hatte diese Gruppen ausgesucht, weil 'sie sich gleichermaßen Delikte gegen das Vermögen' hatten zuschulden kommen lassen und zugleich 'absolute Gegenpole bezüglich ihrer Stellung in der Gesellschaft' repräsentierten.

Ein Ergebnis der Untersuchung war, dass von den 216 Personen 109 Freigänger waren, wobei 68 aus der Gruppe der Wirtschaftskriminellen kamen. (Quelle: DER SPIEGEL 43/1991, S.85-92)

1. Stellen Sie eine Kontingenztabelle auf!

   hier: Merkmal A: Art des Straftäters und Merkmal B: Freigänger?

	Freigänger	kein Freigänger	
Wirtschaftsstraftäter	68	35	103
Dieb	41	72	113
	109	107	216

2. Bestimmen Sie die korrespondierenden bedingten relativen Häufigkeiten und interpretieren Sie diese.

$$
\begin{aligned}
h_{B_1|A_1} &= \frac{68}{103} = 0.66 \\
h_{B_1|A_2} &= \frac{41}{113} = 0.36
\end{aligned}
$$

   Der Anteil an Freigängern (B_1) ist bei Wirtschaftsstraftätern (A_1) deutlich höher als bei Dieben (A_2). (eine umgekehrte Interpretation ($h_{A_1|B_1}$ etc.) macht hier inhaltlich wenig Sinn).

3. Wie stark ist der Zusammenhang?

   korrigierter Kontingenzkoeffizient! Zunächst: Tabelle mit erwarteten Häufigkeiten bei empirischer Unabhängigkeit (Spaltensumme 'mal' Zeilensumme 'geteilt durch' Gesamtsumme): $\tilde{n}_{ij}$

	Freigänger	kein Freigänger	
Wirtschaftsstraftäter	(109*103)/216 = 51.98	51.02	103
Dieb	57.02	55.98	113
	109	107	216

Daraus ergibt sich:

$$\begin{aligned}
X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}} \\
&= \frac{(68-51.98)^2}{51.98}+\frac{(35-51.02)^2}{51.02} \\
&+ \frac{(41-57.02)^2}{57.02}+\frac{(72-55.98)^2}{55.98} \\
&\approx 19.05
\end{aligned}$$

Daraus ergibt sich der zunächst der unkorrigierte Kontingenzkoeffizient:

$$\begin{aligned}
K &= \sqrt{\frac{X^2}{X^2+n}} \\
&= \sqrt{\frac{19.05}{19.05+216}} \\
&= 0.285
\end{aligned}$$

Daraus ergibt sich dann der korrigierte Kontingenzkoeffizient:

$$\begin{aligned}
K^* &= \frac{K}{\sqrt{\frac{\min\{I-1,J-1\}}{\min\{I,J\}}}} \\
&= \frac{0.285}{\sqrt{\frac{\min\{2-1,2-1\}}{\min\{2,2\}}}} \\
&= \frac{0.285}{\sqrt{\frac{1}{2}}} \\
&= 0.4
\end{aligned}$$

Es besteht ein mittelstarker empirischer Zusammenhang zwischen der Art der Straftat und ob die Person Freigänger sein darf.

Aufgabe 3

Von den Passagieren auf der Titanic waren 337 in der ersten Klasse, 285 in der zweiten Klasse und 721 in der dritten Klasse. Es waren 885 Besatzungsmitglieder an Bord.

Von den Passagieren der ersten Klasse wurden nach dem Unglück 135 vermißt, von denen der zweiten Klasse 160, von denen der dritten Klasse 541 und von der Besatzung 674.

1. Erstellen Sie eine Kontingenztabelle.

	vermisst	überlebt	
erste Klasse	135	202	337
zweite Klasse	160	125	285
dritte Klasse	541	180	721
Crew	674	211	885
	1510	718	2228

2. Bestimmen Sie die korrespondierenden bedingten relativen Häufigkeiten und interpretieren Sie diese.

$$\begin{aligned} h_{B_1|A_1} &= \frac{135}{337} = 0.4 \\ h_{B_1|A_2} &= \frac{160}{285} = 0.56 \\ h_{B_1|A_3} &= \frac{541}{721} = 0.75 \\ h_{B_1|A_4} &= \frac{674}{885} = 0.76 \end{aligned}$$

Der Anteil an vermissten Passagieren (B_1) ist bei Passagieren der ersten Klasse (A_1) am geringsten, bei Passagieren der dritten Klasse und der Crew (A_3 bzw. A_4) am höchsten. (eine umgekehrte Interpretation ($h_{A_1|B_1}$ etc.) macht hier inhaltlich wenig Sinn).

3. Wie stark ist der Zusammenhang?

korrigierter Kontingenzkoeffizient! Zunächst: Tabelle mit erwarteten Häufigkeiten bei empirischer Unabhängigkeit (Spaltensumme 'mal' Zeilensumme 'geteilt durch' Gesamtsumme): $\tilde{n}_{ij}$

	vermisst	überlebt	
erste Klasse	228.4	108.6	337
zweite Klasse	193.15	91.84	285
dritte Klasse	488.65	232.35	721
Crew	599.8	285.2	885
	1510	718	2228

Daraus ergibt sich:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}} \\
&= \frac{(135-228.4)^2}{228.4}+\frac{(202-108.6)^2}{108.6} \\
&= \ldots + \frac{(211-285.2)^2}{285.2} \\
&= 182.07
\end{aligned}
$$

Daraus ergibt sich der zunächst der unkorrigierte Kontingenzkoeffizient:

$$
\begin{aligned}
K &= \sqrt{\frac{X^2}{X^2+n}} \\
&= \sqrt{\frac{182.07}{1182.07+2228}} \\
&= 0.27
\end{aligned}
$$

Daraus ergibt sich dann der korrigierte Kontingenzkoeffizient:

$$
\begin{aligned}
K^* &= \frac{K}{\sqrt{\frac{\min\{I-1,J-1\}}{\min\{I,J\}}}} \\
&= \frac{0.27}{\sqrt{\frac{\min\{4-1,2-1\}}{\min\{4,2\}}}} \\
&= \frac{0.27}{\sqrt{\frac{1}{2}}} \\
&= 0.38
\end{aligned}
$$

Es besteht ein mittelstarker empirischer Zusammenhang zwischen Klassenzugehörigkeit der Passagiere und ob die Person das Unglück überlebt hat.

Aufgabe 4

377 Personen wurden nach ihrem Einkommen befragt. Außerdem sollten sie angeben, wie zufrieden sie sind.

Es ergab sich folgende Tabelle:

	sehr unzufrieden	unzufrieden	zufrieden	sehr zufrieden
bis 3000 EUR	20	24	80	82
über 3001 EUR	7	18	54	92

Berechen Sie eine geeignete Maßzahl für den Zusammenhang und interpretieren Sie den Wert kurz.

Rechenhilfen:

$\tilde{n}_{11} = 14.75$ $\tilde{n}_{12} = 22.95$ $\tilde{n}_{13} = 73.22$ $\tilde{n}_{14} = 95.08$

$\tilde{n}_{21} = 12.25$ $\tilde{n}_{22} = 19.05$ $\tilde{n}_{23} = 60.78$ $\tilde{n}_{24} = 78.92$

$$
\begin{aligned}
\chi^2 &= \frac{(20-14.75)^2}{14.75} + \frac{(7-12.25)^2}{12.25} \\
&+ \frac{(24-22.95)^2}{22.95} + \frac{(18-19.05)^2}{19.05} \\
&+ \frac{(80-73.22)^2}{73.22} + \frac{(54-60.78)^2}{60.78} \\
&+ \frac{(82-95.08)^2}{95.08} + \frac{(92-78.92)^2}{78.92} \\
&= 9.58
\end{aligned}
$$

$$
\begin{aligned}
K &= \sqrt{\frac{\chi^2}{\chi^2 + n}} \\
&= \sqrt{\frac{9.58}{9.58 + 377}} \\
&= 0.157
\end{aligned}
$$

$$
\begin{aligned}
K^* &= \frac{K}{\sqrt{\frac{min(I-1,J-1)}{min(I,J)}}} \\
&= \frac{0.157}{\sqrt{\frac{1}{2}}} \\
&= 0.22
\end{aligned}
$$

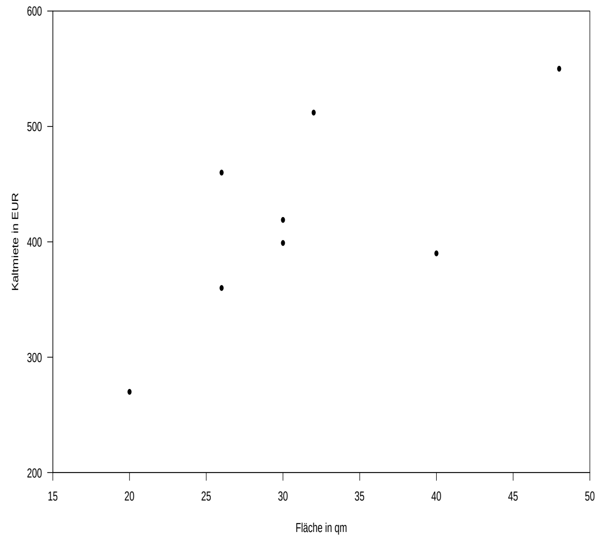
Ein eher schwacher Zusammenhang.

Aufgabe 5

Sie schlagen vor Beginn Ihres Studiums eine Pforzheimer Zeitung auf und suchen alle Einzimmerwohnungen heraus, die explizit in der Nähe der Hochschule liegen. Es sind acht. Die folgende Tabelle gibt neben der Fläche x_t in m^2 auch die Kaltmiete y_t in EUR der acht Wohnungen an.

t	x_t	y_t
1	20	270
2	26	460
3	32	512
4	48	550
5	26	360
6	30	399
7	30	419
8	40	390

1. Erstellen Sie das Streudiagramm der Daten.



2. Bestimmen und interpretieren Sie den Wert des Korrelationskoeffizienten von Bravais-Pearson.

   Es gilt:

$$r_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

   Es gilt hier $\overline{x} = 31.5$ und $\overline{y} = 420$

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
1	20	270	-11.5	-150	1725	132.25	22500
2	26	460	-5.5	40	-220	30.25	1600
3	32	512	0.5	92	46	0.25	8464
4	48	550	16.5	130	2145	272.25	16900
5	26	360	-5.5	-60	330	30.25	3600
6	30	399	-1.5	-21	31.5	2.25	441
7	30	419	-1.5	-1	1.5	2.25	1
8	40	390	8.5	-30	-255	72.25	900

$$r_{x,y} = \frac{3804}{\sqrt{542 \cdot 54406}} = 0.7$$

   Es besteht ein recht starker positiver linearer Zusammenhang.

3. Bestimmen und interpretieren Sie den Wert des Korrelationskoeffizienten von Spearman.

$$r_S = \frac{\sum\limits_{i=1}^{n} (r_i - \overline{r}) \cdot (s_i - \overline{s})}{\sqrt{\sum\limits_{i=1}^{n} (r_i - \overline{r})^2 \cdot \sum\limits_{i=1}^{n} (s_i - \overline{s})^2}} \tag{1}$$

Es gilt für die Ränge (mit Bindungen $\rightarrow$ Durchschnittsränge)

$\overline{r} = 4.5$ und $\overline{s} = 4.5$

i	$R(x) = r_i$	$R(y) = s_i$	$(r_i - \overline{r})$	$(s_i - \overline{s})$	$(r_i - \overline{r}) \cdot (s_i - \overline{s})$	$(r_i - \overline{r})^2$	$(s_i - \overline{s})^2$
1	1	1	-3.5	-3.5	12.25	12.25	12.15
2	2.5	6	-2	1.5	-3	4	2.25
3	6	7	1.5	2.5	3.75	2.25	6.25
4	8	8	3.5	3.5	12.25	12.25	12.25
5	2.5	2	-2	-2.5	5	4	6.25
6	4.5	4	0	-0.5	0	0	0.25
7	4.5	5	0	0.5	0	0	0.25
8	7	3	2.5	-1.5	-3.75	6.25	2.25

und somit

$$r_S = \frac{26.5}{\sqrt{41 \cdot 42}} = 0.64$$

Es besteht ein recht starker positiver monotoner Zusammenhang zwischen den beiden Merkmalen.

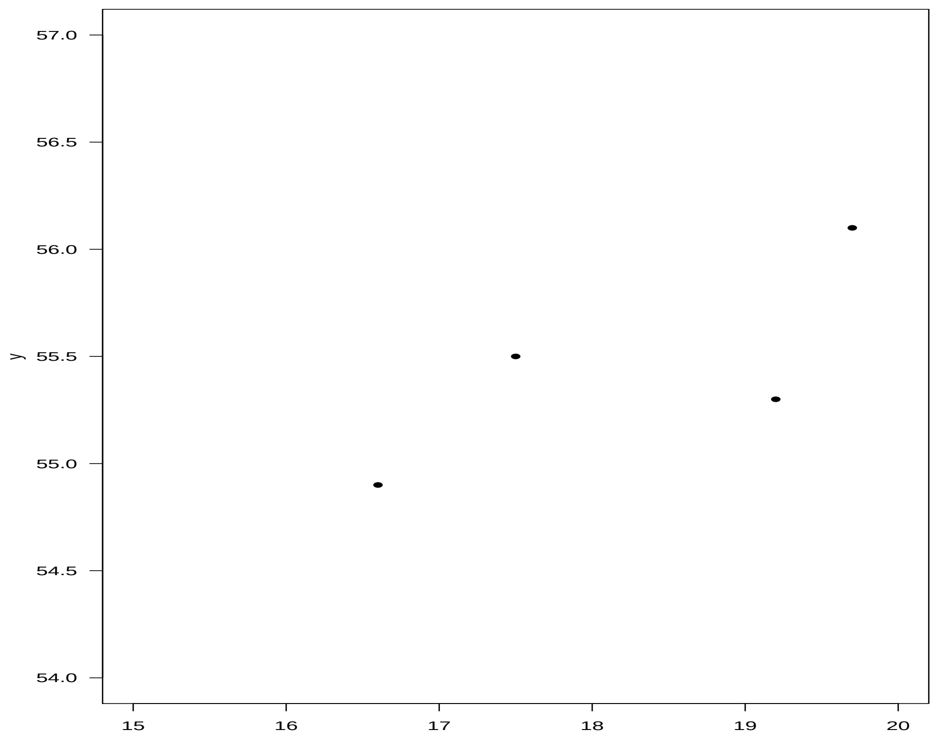
Aufgabe 6

Die folgende Tabelle enthält die Preise (in EUR) für 1 kg Schweinefleisch x und 1 kg Rindfleisch y in den Jahren von 2002 bis 2005 in Deutschland.

Jahr	Schweinefleisch	Rindfleisch
2002	19.2	55.3
2003	19.7	56.1
2004	17.5	55.5
2005	16.6	54.9

(Quelle: Statistisches Jahrbuch)

(a) Zeichnen Sie das Streudiagramm



(b) Wie stark ist der Zusammenhang zwischen dem Preis für Rindfleisch und dem Preis für Schweinefleisch?

Rechenhilfen: $\bar{x} = 18.25$, $\bar{y} = 55.45$, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 6.29$ und $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.75$

es gilt

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

mit

$$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = 1.67$$

ergibt sich

$$\begin{aligned} r_{x,y} &= \frac{1.67}{\sqrt{6.29 \cdot 0.75}} \\ &= 0.77 \end{aligned}$$

Es besteht ein recht starker Zusammenhang zwischen den Preisen.

(c) Wie ändert sich der Wert des Zusammenhangsmaßes aus Aufgabenteil (b), wenn Sie die Preise in US-DOLLAR umrechen? Dabei gilt: 1 EUR = 1.36 US-DOLLAR

wird kleiner	[ ]
bleibt gleich	[X]
wird größer	[ ]
kann man nicht sagen	[ ]

(d) (Hier liegen keine Bindungen vor, daher kann die schnellere Formel angewendet werden)

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}\,. \tag{2}$$

Dabei gilt $d_i = r_i - s_i$.

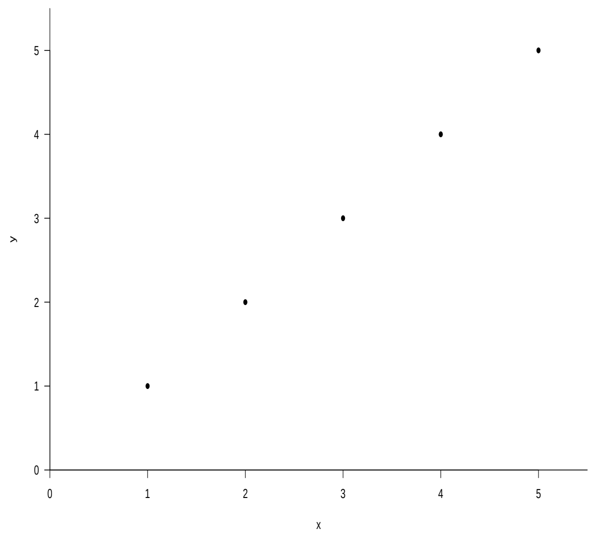
Jahr	2002	2003	2004	2005
r_i (Schwein)	3	4	2	1
s_i (Rind)	2	4	3	1
d_i	1	0	-1	0

Es gilt $\sum_{i=1}^{n} d_i^2 = 2$

$$r_S = 1 - \frac{6 \cdot 2}{4 \cdot (4^2 - 1)} = 0.8 \tag{3}$$

Aufgabe 7

Die folgende Grafik zeigt das Streudiagramm zwischen zwei Merkmalen y und x. Insgesamt sind fünf Punkte abgebildet, die aller auf einer Geraden liegen.



1. Welchen Wert nehmen $r_{x,y}$ und r_S an?

   $r_{x,y} = r_S = 1$

2. Wie ändern sich $r_{x,y}$ und r_S, wenn

   (a) alle $y_i (i = 1, ...5)$ quadriert werden?

   $r_{x,y}$ bleibt positiv, wird aber kleiner, da der perfekte lineare Zusammenhang durch das Quadrieren verloren geht. r_S ändert sich nicht, da die y_i lediglich monoton transformiert werden (die Ränge ändern sich nicht).

   (b) der Punkt $(x, y) = (5, 5)$ ersetzt wird duch $(8, 5)$, d.h. der Punkt nach rechts in der Abbildung verschoben wird?

   $r_{x,y}$ wird etwas kleiner (bleibt aber positiv), da kein perfekter linearer Zusammenhang mehr besteht. Da sich die Ränge durch die Verschiebung des Punktes nicht ändern, gilt weiterhin $r_S = 1$.

   (c) alle $y_i (i = 1, ...5)$ mit -1 multipliziert werden?

   $r_{x,y}^{neu} = r_S^{neu} = -1$

(d) alle y_i und $x_i (i = 1, ...5)$ mit -1 multipliziert werden?

beide Werte bleiben unverändert 1, da sich die Vorzeichenänderungen gegenseitig aufheben.

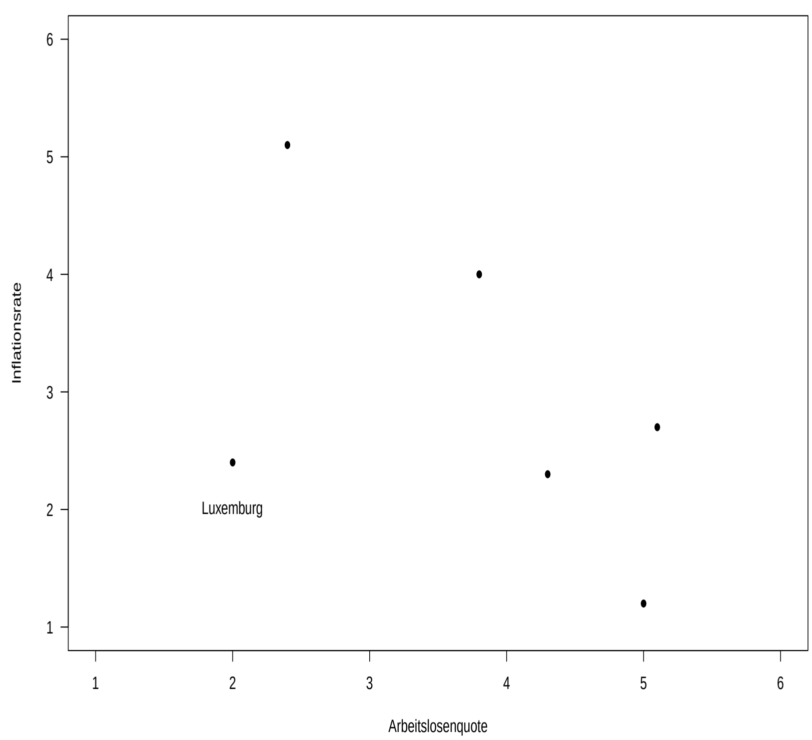
Aufgabe 8

Die folgende Tabelle enthält die Arbeitslosenquote und Inflationsrate in Ländern der EU im Jahr 2001.

Land	Arbeitslosenquote	Inflationsrate
Dänemark	4.3	2.3
Irland	3.8	4.0
Niederlande	2.4	5.1
Schweden	5.1	2.7
England	5.0	1.2

(Quelle: Statistisches Jahrbuch 2002 für das Ausland)

1. Zeichnen Sie das Streudiagramm



2. Bestimmen Sie den Wert des Korrelationskoeffizienten von Spearman.

   Es gilt für die Ränge:

Land	R(x)	R(y)	d_i	d_i^2
1	3	2	1	1
2	2	4	-2	4
3	1	5	-4	16
4	5	3	2	4
5	4	1	3	9

   Es ergibt sich $\sum d_i^2 = 34$
   und somit

$$
\begin{aligned}
r_S &= 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)} \\
&= 1 - \frac{6 * 34}{5 \cdot (25 - 1)} \\
&= 1 - 1.7 \\
&= -0.7
\end{aligned}
$$

   Es besteht es recht starker negativer Zusammenhang zwischen den beiden Merkmalen.

3. In Luxemburg betrug die Arbeitslosenquote im Jahr 2001 2 Prozent. Die Inflationsrate lag bei 2.4 Prozent.

   (a) Berücksichtigen Sie diesen Wert im Streudiagramm.

   (b) Wie ändert sich der Wert des Korrelationskoeffizienten von Spearman, wenn Sie den Wert von Luxemburg noch berücksichtigen?

   wird kleiner [ ]

   bleibt gleich [ ]

   wird größer [X]

   kann man nicht sagen [ ]

Aufgabe 9

Im Rahmen eines Experiments zur Entwicklung einer Spielekonsole für eine bestimmte Zielgruppe untersuchen Sie den Zusammenhang von drei Variablen:

- das Körpergewicht einer Person (G)
- die manuelle Geschicklichkeit der Person (M)
- das Alter der Person (A)

Sie haben herausbekommen, dass Gewicht und manuelle Geschicklichkeit mit einem Wert von 0.45 korreliert sind. Das Gewicht einer Person ist mit dem Alter wiederum mit 0.85 korreliert. Die manuelle Geschicklichkeit und das Alter sind mit einem Wert von 0.6 korreliert. Zum weiteren Testen müssen Sie folgende Frage beantworten: Wie ist das Gewicht einer Person mit der manuellen Geschicklichkeit korreliert, wenn Sie den Einfluss des Alters herausrechnen?

gesucht ist der partielle Korrelationskoeffizient. Hier:

$$r_{GM.A} = \frac{r_{GM} - r_{GA} \cdot r_{MA}}{\sqrt{(1 - r_{GA}^2) \cdot (1 - r_{MA}^2)}} \tag{4}$$

hier:

$$r_{GM.A} = \frac{0.45 - 0.85 \cdot 0.6}{\sqrt{(1 - 0.85^2) \cdot (1 - 0.6^2)}} = -0.14 \tag{5}$$

Wird der Einfluss des Alters ausgeschaltet (bzw. dafür kontrolliert), dann beträgt die Korrelation zwischen Körpergewicht und manueller Geschicklichkeit -0.14. Sie ist leicht negativ.