Fachbereich Mathematik
M. Kohler
A. Fromkorth
J. Mehnert

TECHNISCHE UNIVERSITÄT DARMSTADT

SS 2009
8. Mai 2009

Lösungsvorschläge zum 2. Übungsblatt zur „Einführung in die Stochastik“

Lösung zur Aufgabe 6 (3 Punkte)

Datenpunkte der Größe nach aufsteigend geordnet (hier $n = 10 \hookrightarrow$ gerade):

$$1, 1.3, 2.3, 2.7, 3.3, 3.7, 4, 4, 5, 5$$

(a)

Mittel:	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\rightarrow\ \overline{x} = 3.23$
Median:	$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}\right), & \text{falls } n \text{ gerade.} \end{cases}$	$\rightarrow\ \tilde{x} = 3.5$
Spannweite:	$r := x_{max} - x_{min}$	$\rightarrow\ r = 4$
Varianz:	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	$\rightarrow\ s^2 = 1.9468$
Standardabweichung:	$s = \sqrt{s^2}$	$\rightarrow\ s = 1.3953$
Interquatilabstand:	$IQR = x_{(\lceil \frac{3}{4}n \rceil)} - x_{(\lceil \frac{1}{4}n \rceil)}$	$\rightarrow\ IQR = 1.7$

Lösung zur Aufgabe 7 (3 Punkte)

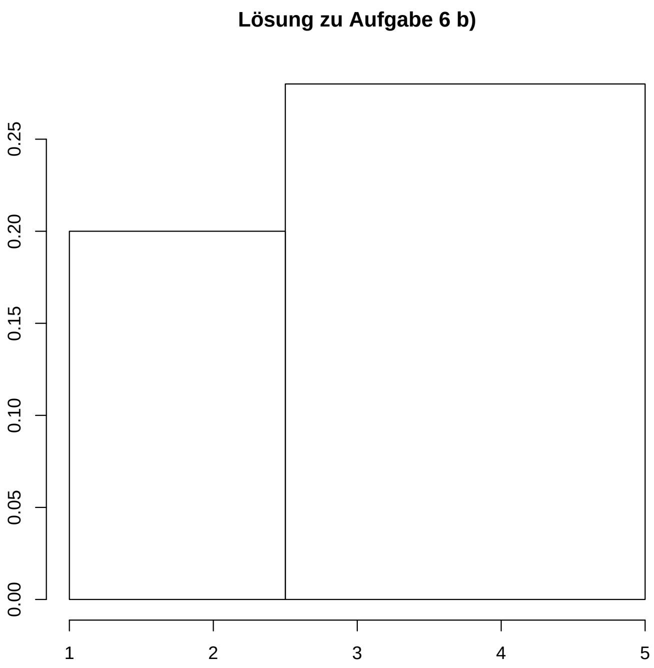
Wir verwenden die Notation aus der Vorlesung.

(a) Wir müssen die Anzahl der Studenten bestimmen, die zu den Intervallen $I_2 = [-20, 0)$ und $I_3 = [0, 10)$ gehören. D.h. wir suchen n_2 und n_3. Bezeichnen wir die Balkenhöhen über dem Intervall I_j mit y_j, für $j \in \{1, \ldots, 5\}$, so erhalten wir durch Umstellen der Formel aus der Vorlesung:

$$n_2 + n_3 = (y_2 \cdot \lambda(I_2) + y_3 \cdot \lambda(I_3)) \cdot n \approx (0.0165 \cdot 20 + 0.0075 \cdot 10) \cdot 40 = 16.2 \approx 16$$

(b)

$$\begin{aligned} \text{Median: } \tilde{x} &\approx -0.5 \\ \text{IQR} &\approx 45 - (-30) = 75 \end{aligned}$$



Lösung zur Aufgabe 8 (3 Punkte)

(b) Seien x_i die Ausgaben pro Student (in Euro) in Zeile i der Tabelle und y_i die Arbeitslosenquote (in Prozent) in Zeile i der Tabelle. Nach der Formel aus der Vorlesung hat die Regressionsgerade die Form

$$y = \hat{a}(x - \bar{x}) + \bar{y},$$

mit

$$\begin{aligned}
\hat{a} &= \frac{s_{xy}}{s_x^2}, \\
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \\
s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \\
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}$$

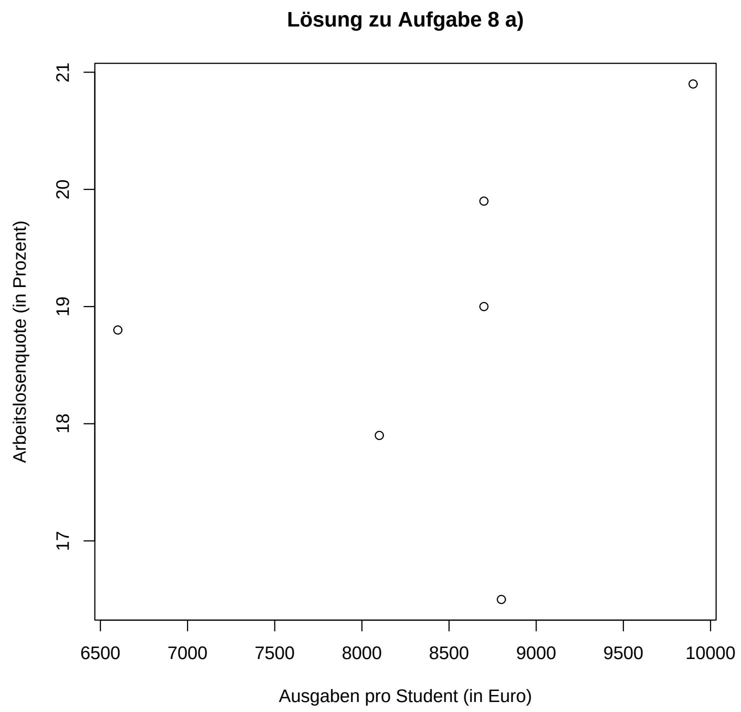


Abbildung 1: Aufgabe 8

und $n = 6$. Einsetzen der Werte ergibt

$$
\begin{aligned}
\bar{x} &= \frac{1}{6}(8100 + 6600 + 8700 + 8700 + 9900 + 8800) \approx 8466.667 \\
\bar{y} &= \frac{1}{6}(17.9 + 18.8 + 19.6 + 19 + 20.9 + 16.5) \approx 18.7833 \\
s_{xy} &= \frac{1}{5}((8100 - 8466.667) \cdot (17.9 - 18.7833) + \cdots + (8800 - 8466.667) \cdot (16.5 - 18.7833)) \\
&\approx 561.33 \\
s_x^2 &= \frac{1}{5}((8100 - 8466.667)^2 + \cdots + (8800 - 8466.667)^2) \approx 118667 \\
\hat{a} &\approx 0.000476
\end{aligned}
$$

und damit

$$y = 0.000476 \cdot (x - 8466.667) + 18.7833.$$

(c) Lässt man den Sachsen-Anhalt Datenpunkt weg, so wird die Steigung der Regressionsgeraden negativ (was auch zu einer Änderung des y-Achsenabschnitts führt).

$$
\begin{aligned}
\bar{x} &= 8180 \\
\bar{y} &= 18.42 \\
\hat{a} &= -0.0001977830
\end{aligned}
$$

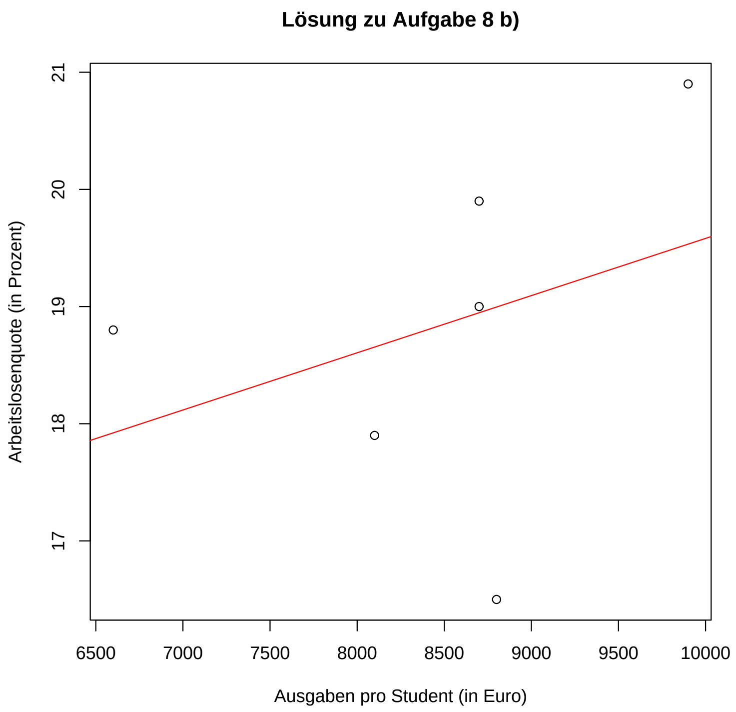


Abbildung 2: Aufgabe 8

Lösung zur Aufgabe 9 (3 Punkte)

(a)

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\\
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \frac{1}{n}\sum_{i=1}^{n}\left(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i-\bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y}\\
&= \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}
\end{aligned}
$$

(b) Die empirische Korrelation ist definiert als

$$r_{xy}=\frac{s_{xy}}{s_x\cdot s_y}.$$

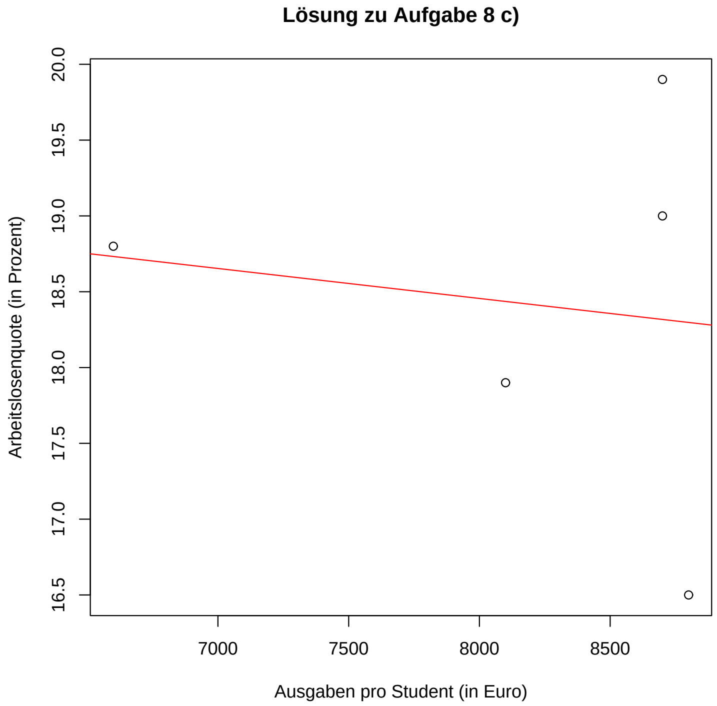


Abbildung 3: Aufgabe 8

Die Werte von s_{xy} und s_x wurden schon in Aufgabe 8 berechnet. Wegen

$$s_y^2 \approx 2.350667$$

folgt dann

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} \approx 0.3456429.$$

(c) Da das Vorzeichen der empirischen Korrelation mit dem Vorzeichen der Steigung der Regressionsgeraden übereinstimmt, ist im vorliegenden Fall die Steigung der Regressionsgeraden positiv, da $r_{xy} = 0.3456429 > 0$.

(d) Das Umrechnen der Einheiten kann man als Multiplikation mit einer positiven Konstanten realisieren. Anstelle der Daten $(x_1, y_1), \ldots, (x_n, y_n)$ betrachten wir also die Datenpunkte $(z_1, w_1), \ldots, (z_n, w_n)$ mit $(z_i, w_i) = (ax_i, by_i)$ für $i = 1, \ldots, n$. Dann gilt:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} ax_i = a\frac{1}{n}\sum_{i=1}^{n} x_i = a \cdot \bar{x}$$

und genauso

$$\bar{w} = b \cdot \bar{y}.$$

Damit erhalten wir

$$
\begin{aligned}
s_{zw} &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})\cdot(w_i-\bar{w}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})\cdot(by_i-b\bar{y}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}a(x_i-\bar{x})\cdot b(y_i-\bar{y}) \\
&= abs_{xy}, \\
s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i-a\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}a^2(x_i-\bar{x})^2 \\
&= a^2s_x^2.
\end{aligned}
$$

und

$$
\begin{aligned}
s_w^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(w_i-\bar{w})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(by_i-b\bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}b^2(y_i-\bar{y})^2 \\
&= a^2s_y^2.
\end{aligned}
$$

Das bedeutet für die empirische Korrelation

$$
r_{zw} = \frac{s_{zw}}{\sqrt{s_z^2s_w^2}} = \frac{abs_{xy}}{\sqrt{a^2s_x^2b^2s_y^2}} = r_{xy},
$$

d.h. die empirsiche Korrelation ändert sich durch die Umrechnung nicht.