Übung 6 - Ausführliche Lösungen

Aufgabe 27: 14 Tage lang werden die Verkaufszahlen für ein Buch in einer Buchhandlung notiert: 7, 11, 12, 8, 10, 9, 9, 8, 0, 6, 13, 18, 5 und 11. Zeichnen Sie einen Boxplot für die Daten.

Lösung: Wie bestimmen zunächst die Ordnungsstatistik unserer Daten:

$$x_{(.)} = (0, 5, 6, 7, 8, 8, 9, 9, 10, 11, 11, 12, 13, 18)$$

Dann berechnen wir den Median, das obere, sowie das untere Quartil:

$$n = 14$$

$$n/2 = 7 \in \mathbb{N} \text{ (7 ist eine ganze Zahl)}$$
$$\widetilde{x} = \frac{x_{(7)} + x_{(8)}}{2} = \frac{9+9}{2} = 9$$

$$0.25 \cdot n = 3.5 \notin \mathbb{N} \text{ (3.5 ist keine ganze Zahl)}$$
$$\widetilde{x}_u = x_{(4)} = 7$$

$$0.75 \cdot n = 10.5 \notin \mathbb{N} \text{ (10.5 ist keine ganze Zahl)}$$
$$\widetilde{x}_o = x_{(11)} = 11.$$

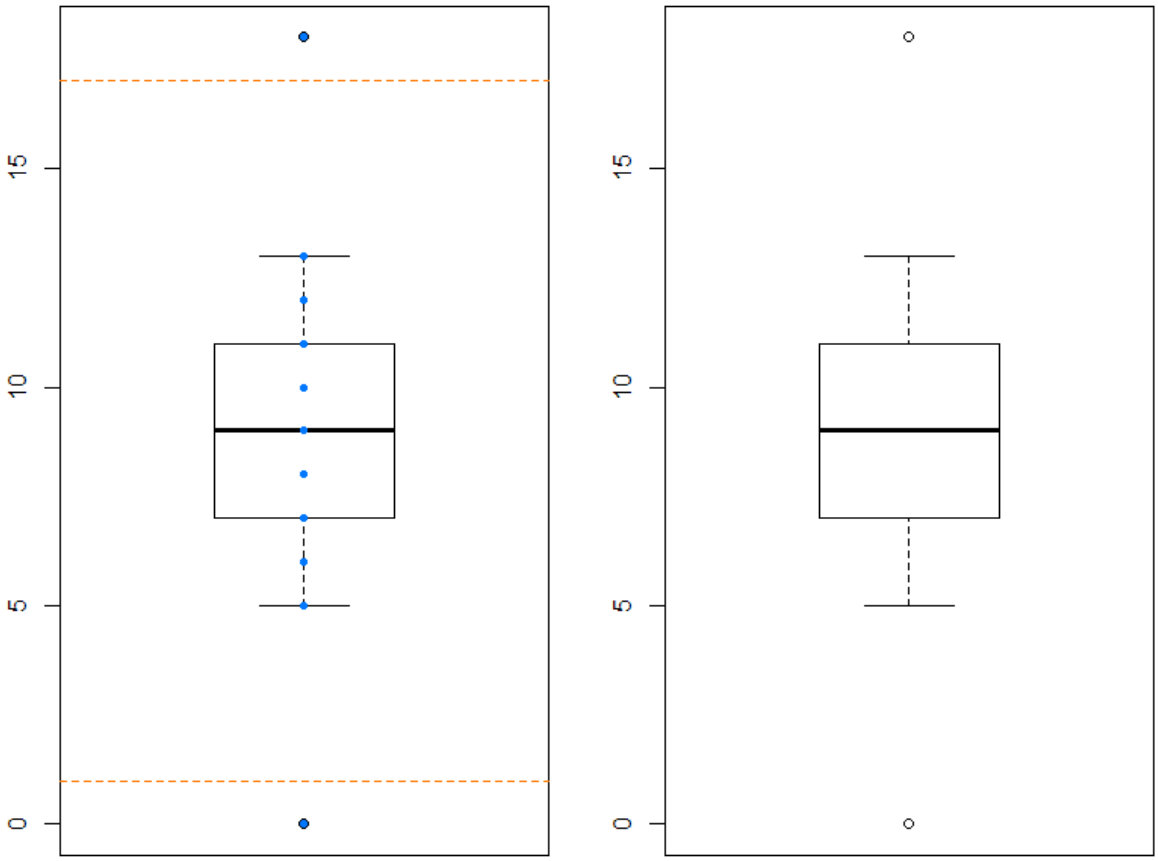
Daneben ist noch der Interquartilsabstand IQR_x gefragt. Wir geben zusätzlich noch Maximum und Minumum der Daten an, um die vertikale Größe des Plots zu ermitteln.

$$\mathrm{IQR}_x = 11 - 7 = 4, \qquad 1.5 \times \mathrm{IQR}_x = 6$$
$$\max x = 18, \qquad \min x = 0.$$

Jetzt zeichnen wir den Boxplot.

1. Auf der Y-Achse reservieren wir einen Bereich von mindestens 0 bis 20.
2. Als erstes zeichnen wir den Median als dicke waagerechte Linie ein.
3. Die Box ergibt sich, wenn wir als obere Begrenzung das obere Quartil und als untere Begrenzung das untere Quartil wählen.
4. Vom **oberen Quartil** aus messen wir eine Strecke der Länge $1.5 \times \mathrm{IQR}_x = 6$ nach oben ab (hier $11 + 6 = 17$) und zeichnen beim letzten Datenpunkt unterhalb dieser Marke das Ende der oberen Antenne ein (hier $x_{(13)} = 13$).
5. Vom **unteren Quartil** aus messen wir eine Strecke der Länge $1.5 \times \mathrm{IQR}_x = 6$ nach unten ab (hier $7 - 6 = 1$) und zeichnen beim letzten Datenpunkt oberhalb dieser Marke das Ende der unteren Antenne ein (hier $x_{(2)} = 5$).
6. Zuletzt werden alle Datenpunkte, die oberhalb und unterhalb der Antennenenden liegen (die Ausreißer) als zusätzliche Punkte in die Grafik eingetragen (hier nur das Minimum $x_{(1)} = 0$ und das Maximum $x_{(14)} = 18$).

In der folgenden Grafik wurden zur Veranschaulichung auf der linken Seite zusätzlich alle Datenpunkte und zwei horizontale Linien in den Höhen $\widetilde{x}_o + 1.5 \times \text{IQR}_x$ und $\widetilde{x}_u - 1.5 \times \text{IQR}_x$ eingezeichnet. Die rechte Seite zeigt den fertigen Boxplot.

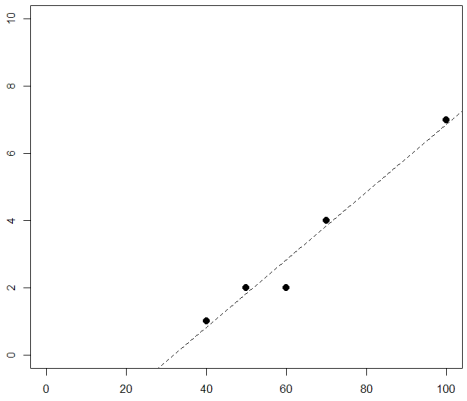


Aufgabe 28: Für sechs Straßen werden die Durchschnittsgeschwindigkeit und die Anzahl der Unfälle in einem Jahr angegeben:

Geschw.:	50	60	100	70	50	40
Unfälle:	2	2	7	4	2	1

Geben Sie die für die beiden Merkmale die empirische Kovarianz und den Korrelationskoeffizienten an und interpretieren Sie das Resultat.

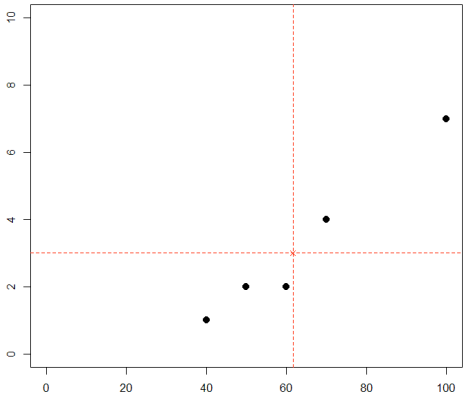
Lösung: Man sieht bereits im Streudiagramm, dass vermutlich ein starker linearer Zusammenhang der beiden Merkmale besteht (in der Grafik wurde mit Hilfe einer linearen Regression die sog. Regressionsgerade eingezeichnet).



Wir begründen unsere Vermutung zunächst mit der Berechnung der empirischen Kovarianz. Wir benötigen dazu die beiden Mittelwerte (es beschreibe x die Geschwindigkeiten und y die Unfallzahlen):

$$\overline{x} = 61\frac{2}{3}, \qquad \overline{y} = 3.$$

Wenn man auf Höhe der Mittelwerte vertikale und horizontale Linien einzeichnet, sieht man, dass alle Daten im ersten oder dritten Quadranten liegen:



Das gibt uns einen zweiten Hinweis auf eine hohe positive Korrelation.

Wir berechnen die Kovarianz:

$$\begin{aligned} s_{xy} &= \overline{x \cdot y} - \overline{x} \cdot \overline{y} \\ &= \frac{2 \cdot 50 + 2 \cdot 60 + \ldots + 1 \cdot 40}{6} - 3 \cdot 61\frac{2}{3} \\ &= 38\frac{1}{3}. \end{aligned}$$

Es liegt eine positive Korrelation vor. Allerdings ist $38\frac{1}{3}$ zunächst nur ein relativer Wert, der wenig über die Höhe der Korrelation aussagt.

Die Cauchy-Schwartzsche Ungleichung gibt uns einen Maximalwert für die Kovarianz:

$$s_{xy} \leq \sigma(x)\sigma(y).$$

Wir berechnen die Standardabweichungen:

$$\begin{aligned} \sigma(x) &= \sqrt{\overline{x^2} - \overline{x}^2} = 2 \\ \sigma(y) &= \sqrt{\overline{y^2} - \overline{y}^2} = 19.50783 \end{aligned}$$

Also gilt

$$s_{xy} \leq 2 \cdot 19.50783 = 39.01567.$$

Wir sehen, dass unser berechneter Wert $38\frac{1}{3}$ sehr nahe beim Maximalwert 39.01567 liegt und können daher bereits jetzt von einem sehr starken positiven linearen Zusammenhang zwischen der Geschwindigkeit und den Unfallzahlen sprechen.

Wir berechnen zuletzt noch den Korrelationskoeffizienten:

$$r_{xy} = \frac{s_{xy}}{\sigma(x)\sigma(y)} = \frac{38\frac{1}{3}}{39.01567} = 0.9825113.$$

Der Wert ist in der Tat sehr hoch (r_{xy} liegt immer zwischen -1 und +1), es bestätigt sich also der starke positive lineare Zusammenhang.

Aufgabe 29: An zwei Hochschulen setzt man unterschiedliche Benotungssysteme ein. Während die Hochschule A die Benotungsskala $I \to II \to III \to IV$ verwendet, mit I als bester Note, ist an der Hochschule B die Skala $a \to b \to c$, mit a als bester Note, in Gebrauch. Für 20 Studierende, die von A nach B wechselten, wird die letzte Note an der Hochschule A mit der ersten Note an der Hochschule B verglichen:

A	I	I	I	I	I	I	I	II	II	II	II	II	II	III	III	III	III	IV	IV	IV
B	a	a	a	a	a	b	b	a	a	a	a	b	b	a	b	b	c	b	b	c

Berechnen Sie den Rangkorrelationskoeffizienten und interpretieren Sie das Ergebnis.

Lösung: Wir müssen zunächst die Ränge bestimmen. Die Ordnungsstatistik für das ordinalskalierte Merkmal A lautet:

I I I I I I I II II II II II II III III III III IV IV IV

Eigentlich müssten wir die Ränge 1 bis 20 verteilen, das wäre aber unfair, denn die ersten 7 Studierenden haben alle dieselbe Note bekommen, würden aber unterschiedliche Ränge (1-7) erhalten. Daher verteilen wir dort den gemeinsamen mittleren Rang

$$\frac{1+2+3+\ldots+7}{7} = \frac{1+7}{2} = 4.$$

Genauso verfahren wir mit den anderen Studierenden, jeweils für beide Hochschulen. Für A und B ergeben sich dann die Ränge:

$$4.0,\ 4.0,\ 4.0,\ 4.0,\ 4.0,\ 4.0,\ 4.0,\ 10.5,\ 10.5,\ 10.5,\ 10.5,$$
$$10.5,\ 10.5,\ 15.5,\ 15.5,\ 15.5,\ 15.5,\ 19.0,\ 19.0,\ 19.0$$

und für Hochschule B:

$$5.5,\ 5.5,\ 5.5,\ 5.5,\ 5.5,\ 14.5,\ 14.5,\ 5.5,\ 5.5,\ 5.5,\ 5.5,\ 14.5,$$
$$14.5,\ 5.5,\ 14.5,\ 14.5,\ 19.5,\ 14.5,\ 14.5,\ 19.5$$

Diese nunmehr metrischen Daten können wir verwenden, um den Korrelationskoeffizienten zu berechnen. Entweder, indem wir die klassische Formel für den Korrelationskoeffizienten verwenden, oder die etwas kürzere Formel aus dem Skript:

$$R_{xy} = \frac{\sum_{k=1}^{n} R(x_i)R(y_i) - n\overline{R}^2}{\sqrt{\sum_{k=1}^{n} R(x_i)^2 - n\overline{R}^2} \times \sqrt{\sum_{k=1}^{n} R(y_i)^2 - n\overline{R}^2}}.$$

Dabei bezeichnet $R(x_i)$ den i-ten Rang des ersten Merkmals und entsprechend $R(y_i)$ den i-ten für das zweite Merkmal. Außerdem gilt für den mittleren Rang

$$\overline{R} = \frac{1+2+\ldots+n}{n} \overset{\text{Gauß-Formel}}{=} \frac{n+1}{2} = \frac{21}{2} = 10.5.$$

Wir erhalten

$$\begin{aligned} R_{xy} &= \frac{4.0 \cdot 5.5 + 4.0 \cdot 5.5 + \ldots + 19.0 \cdot 19.5 - 20 \cdot 10.5^2}{\sqrt{4.0^2 + 4.0^2 + \ldots + 19.0^2 - 20 \cdot 10.5^2} \cdot \sqrt{5.5^2 + 5.5^2 + \ldots + 19.5^2 - 20 \cdot 10.5^2}} \\ &= \frac{315}{\sqrt{612.5} \cdot \sqrt{540}} = 0.5477226. \end{aligned}$$

Der Rangkorrelationskoeffizient ist größer als 0.5, wir können also einen starken positiven Zusammenhang zwischen der Hochschulnote A und der Hochschulnote B feststellen.