Fachbereich Mathematik
Prof. Dr. Michael Kohler
Dominique Küpper

TECHNISCHE UNIVERSITÄT DARMSTADT

WS 09/10
17.11.2009

3. Übungsblatt zur „Mathematik und Statistik für Biologie“

Aufgabe 9 (3 Punkte)

Eine Messung der Kopf-Rumpf-Länge von Hausmäusen ergab folgende Messreihe (Ergebnisse in cm):

$$9.4, 7.2, 8.4, 8.3, 7.8, 8.5, 9.4, 8.8, 8.9, 9, 8.7, 8.3, 9.4, 8.6, 8.8, 9.1, 7.9, 8.1, 9.2, 8.9$$

(a) Bestimmen sie den Median, das empirische arithmetische Mittel und die empirische Varianz der Datenreihe.

(b) Stellen Sie die Daten als Boxplot dar.

Lösung: Wir bezeichnen die einzelnen Datenwerte mit $x_1, \ldots, x_{20}$. Die aufsteigend sortierten Werte mit $x_{(1)}, \ldots, x_{(20)}$.

(a) Da 20 Datenwerte gegeben sind, berechnet sich der Median durch

$$\frac{x_{(10)} + x_{(11)}}{2} = \frac{8.7 + 8.8}{2} = 8.75$$

Für das arithmetische Mittel gilt

$$\bar{x} = \sum_{i=1}^{20} x_i = \frac{1}{20}\left(9.4 + 7.2 + \cdots + 8.9\right) = \frac{172.7}{20} = 8.635$$

Um die empirische Varianz zu berechnen kann man entweder die Formel aus der Vorlesung verwenden, d.h.

$$s^2 = \frac{1}{19}\left((9.4 - 8.635)^2 + (7.2 - 8.635)^2 + \cdots + (8.9 - 8.635)^2\right)$$

oder man berechnet zuerst

$$\sum_{i=1}^{20} x_i^2 = 9.4^2 + \cdots + 8.9^2 = 1497.77$$

und verwendet dann die Formel

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right).$$

In beiden Fällen erhält man

$$s^2 = 0.3424.$$

(b) Um den Boxplot zu bestimmen, benötigen wir noch das 1. und das 3. Quartil. Unsere Messreihe hat 20 Datenpunkte. Für das 1. Quartil suchen wir diejenigen Werte aus der Datenreihe für die gilt:

i. mindestens 25 % der Datenpunkte sind kleiner oder gleich dem ausgewählten Wert und
ii. mindestens 75 % der Datenpunkte sind größer oder gleich dem ausgewählten Wert.

Das 1. Quartil liegt also zwischen $x_{(5)}$ und $x_{(6)}$ und ist somit der Mittelwert dieser beiden Werte (also gleich $\frac{8.3+8.3}{2} = 8.3$). Entsprechend geht man für das 3. Quartil vor. Hier suchen wir diejenigen Werte, für die gilt:

i. mindestens 75 % der Datenpunkte sind kleiner oder gleich dem ausgewählten Wert und
ii. mindestens 25 % der Datenpunkte sind größer oder gleich dem ausgewählten Wert.

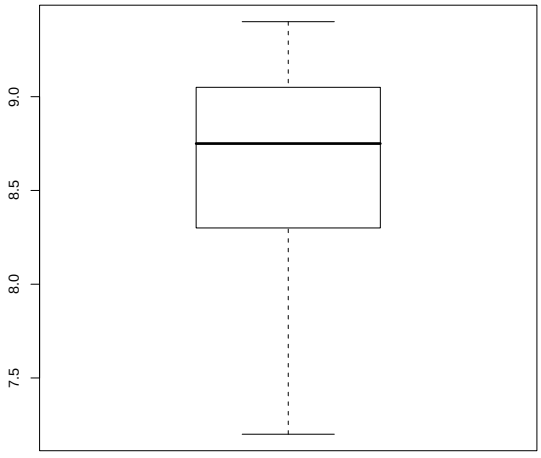
Das 3. Quartil liegt also zwischen $x_{(15)}$ und $x_{(16)}$ und ist somit der Mittelwert dieser beiden Werte (also gleich $\frac{9+9.1}{2} = 9.05$). Als Interquartilabstand erhalten wir also $9.05 - 8.3 = 0.75$.
Das oben beschriebene Verfahren zur Berechnung der Quartile entspricht der im Buch von Prof. Kohler beschriebenen Vorgehensweise. In der Vorlesung wurde dies etwas vereinfacht: Als 1. Quartil nimmt man

$$x_{(\lceil \frac{n}{4} \rceil)} = x_{(\lceil \frac{20}{4} \rceil)} = x_{(5)} = 8.3$$

und als 3. Quartil

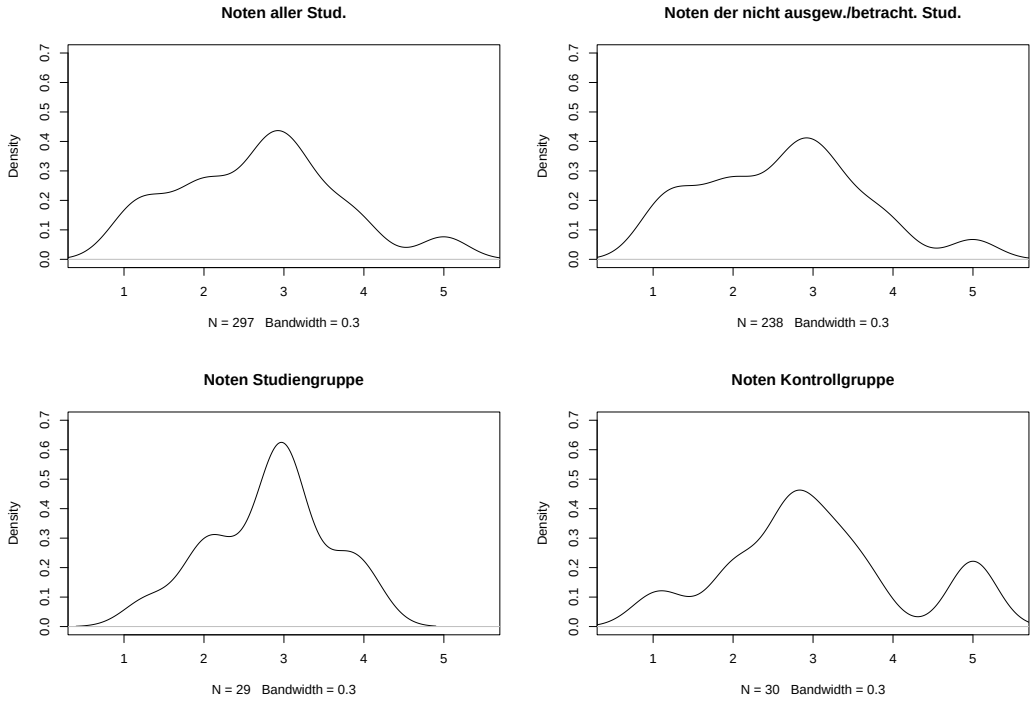
$$x_{(\lceil \frac{3 \cdot n}{4} \rceil)} = x_{(\lceil \frac{3 \cdot 20}{4} \rceil)} = x_{(15)} = 9.$$

Damit erhält man einen IQR von $9 - 8.3 = 0.7$. Datenpunkte, deren Abstand nach oben bzw. unten vom 3. bzw. 1. Quartil größer als $1,5 \cdot IQR$ ist, werden als Ausreißer betrachtet und durch Kreise gesondert dargestellt. Folgt man der Vorgehensweise aus dem Buch ergibt sich $1,5 \cdot IQR = 1.125$ und somit enthält der Datensatz keine Ausreißer. Geht man vor, wie in der Vorlesung beschrieben, erhält man $1,5 \cdot IQR = 1.5 \cdot 0.7 = 1.05$, d.h. in diesem Fall wäre $8.3 - 1.05 = 7.25$ und somit der Datenpunkt 7.2 ein Ausreißer.
In der Klausur werden beide Vorgehensweisen akzeptiert (es sollte aber aus der Lösung ersichtlich werden, welche der beiden Definitionsmöglichkeiten verwendet wird).
Als Boxplot für die Vorgehensweise aus dem Buch ergibt sich



Aufgabe 10 (3 Punkte)

Im Rahmen einer Diplomarbeit wurde an der Universität Stuttgart versucht, ein Verfahren zu entwickeln, welches in der Lage ist, StudentInnen zu indentifizieren, die voraussichtlich Probleme mit dem Bestehen einer Statistik-Prüfung haben werden. Dieses Verfahren wurde im Rahmen der Vorlesung *Statistik II für WirtschaftswissenschaftlerInnen* überprüft. Dazu wurden durch Anwendung diese Verfahrens aus den ca. 300 Teilnehmern an der Klausur 60 ausgewählt, und zufällig in zwei gleich große Gruppen, SG und KG, unterteilt. Die StudentInnen in der SG wurden vor der Prüfung schriftlich zu einem ca. sechsstündigen Zusatzkurs zur Klausurvorbereitung eingeladen. In der unten stehenden Abbildung sind Kern-Dichteschätzer angewandt auf die Noten aller StudentInnen, bzw. der StudentInnen in der SG, bzw. der StudentInnen in der KG, bzw. der StudentInnen, die weder in der SG noch in der KG waren, abgebildet. Wie können sie durch Vergleich



dieser Kern-Dichteschätzer feststellen, ob

(a) das Verfahren wirklich vor allem durchfallgefährdete StudentInnen ausgewählt hat ?

(b) das Anbieten des Zusatzkurses zu einer Verringerung der Durchfallquote bei den als durchfallgefährdet eingestuften StudentInnen geführt hat ?

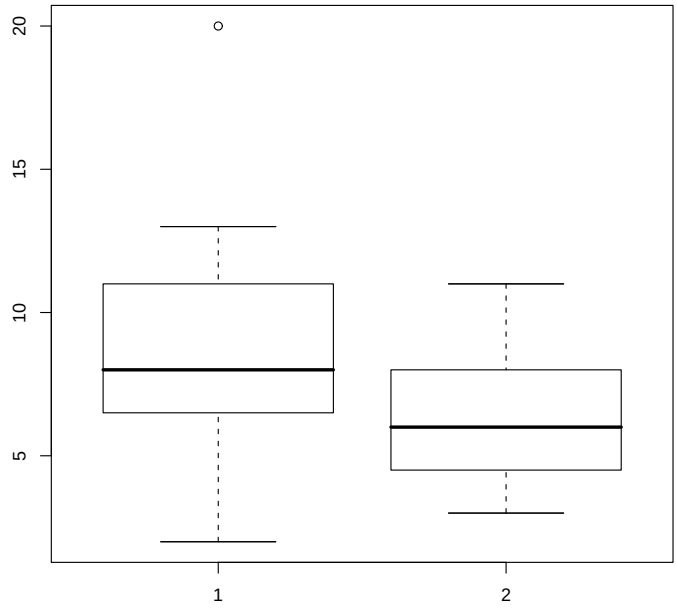
Lösung:

(a) Wird in der Abbildung die rechte obere mit der rechten unteren Graphik verglichen, dann ist festzustellen, dass der relative Anteil bei den Noten zwischen 4 und 6 in der Kontrollgruppe deutlich höher ist, als in der nicht ausgew./betracht. Stud., d.h. die betrachtete Fläche ist größer. Somit lässt sich die Aussage feststellen.

(b) Betrachtet man die Graphik "Noten Studiengruppe", dann ist der relative Anteil bei den Noten zwischen 4 und 6 deutlich niedriger als bei der Graphik "Noten Kontrollgruppe" (im gleichen Notenintervall). Daher lässt sich folgern, dass das Anbieten des Zusatzkurses zu einer Verringerung der Durchfallquote bei den als durchfallgefährdet eingestuften StudentInnen geführt hat.

Aufgabe 11 (3 Punkte)

Horst-Uwe und Klaus-Rüdiger haben montags morgens eine Vorlesung in der Stadtmitte. Beide fahren mit dem Auto zur Uni und parken im Martinsviertel. Im Wintersemester 2008/09 haben beide in den 15 Semesterwochen jeweils die Zeit gestoppt, die sie brauchten, um einen Parkplatz zu finden. Die Boxplots repräsentieren jeweils die Messreihen von Horst-Uwe (1) und Klaus-Rüdiger (2). Vergleichen Sie beide Plots. Welche Aussagen können Sie aufgrund der Boxplots machen? Wie könnten die Unterschiede zustande gekommen sein?



Lösung: Horst-Uwe braucht eher länger als Klaus-Rüdiger, um einen Parkplatz zu finden: Das 25%-Quantil liegt bei Horst-Uwe über dem 50%-Quantil (also dem Median) von Klaus-Rüdiger. Außerdem liegt das 75%-Quantil von Klaus-Rüdiger unter dem Median von Horst-Uwe. Weiterhin benötigte Klaus-Rüdiger für die Parkplatzsuche maximal 11 Minuten, wohingegen Horst-Uwe bis zu 20 Mnuten lang nach einem Parkplatz suchte. Mögliche Ursachen für diese Unterschiede wären z. B.:

- Klaus-Rüdiger ist ein geübterer Einparker als Horst-Uwe,
- Horst-Uwe fährt einen VW-Bus und Klaus-Rüdiger einen Smart
- etc.

Der Ausreißer bei Horst-Uwe könnte durch ein Hindernis an diesem Tag hervorgerufen worden sein, z.B. durch einen vor ihm fahrenden Müllwagen.

Aufgabe 12 (3 Punkte)

Ein Lehrer einer 4. Grundschulklasse bildet Messreihen über die Fehleranzahl in Diktaten, getrennt nach Rechts- und Linkshändern.

Linkshänder:

$$3, 8, 0, 12, 14, 7, 6, 2, 1$$

Rechtshänder:

$$4, 5, 2, 2, 0, 8, 11, 9, 7, 7, 0, 2, 2$$

Vergleichen Sie die Schwankungen der beiden Messreihen um ihren Mittelwert.

Lösung: Wir berechnen die empirische Mittel $\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{53}{9}$, $\bar{y} = \frac{1}{13}\sum_{i=1}^{13} y_i = \frac{59}{13}$ der Links- und Rechtshänder. Damit berechnen wir die Varianzen mit

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx 23,86$$

und

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \approx 12,78.$$

die Variationskoeffizienten sind $V_x = \frac{s_x}{\bar{x}} \approx 0,83$ und $V_y = \frac{s_y}{\bar{y}} \approx 0.79$. Sowohl die empirische Varianz als auch der Variationskoeffizient sind im ersten Fall größer.