Orientierungshilfe für den ersten Hausaufgabenzettel

6. November 2012

Diese Zusammenfassung des ersten Übungszettels soll einen kleinen Überblick geben, wie die Aufgaben hätten bearbeitet werden können. Sie ist keine Musterlösung! Es gibt hin und wieder bei bestimmten Aufgaben auch alternative Lösungmöglichkeiten; es können jedoch nicht alle aufgezeigt werden. Im Folgenden werden manche Abschnitte etwas ausführlicher erklärt, manche nur grob skizziert; das hängt davon ab, an welchen Stellen es bei euch in den Aufgaben mehr Schwierigkeiten gab und an welchen weniger.
Bezüglich der Korrektur: Ff meint Folgefehler.

Aufgabe 2

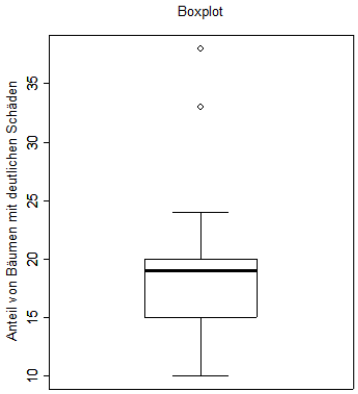
a)
Grundgesamtheit: Menge aller untersuchten Bäume
Untersuchungseinheit: ein untersuchter Baum
Merkmal: Umweltschaden
Ausprägung: deutlicher Umweltschaden ja oder deutlicher Umweltschaden nein

Man orientiere sich an der Abbildung $X : \Omega \to M$ aus dem Skript. Hier bildet die Funktion (Merkmal) Umweltschaden jedem Baum (Untersuchungseinheit) aus der Grundgesamtheit (Menge aller untersuchten Bäume) den Wert 'deutlicher Umweltschade' oder eben 'kein deutlicher Umweltschaden' (Ausprägung) zu. b)
Die Darstellung der Daten in Form eines Boxplots beinhaltet die folgenden wichtigen Werte:

1. Median (dicke Linie innerhalb der Box): hier bei 19
2. unteres Quartil (untere Begrenzung der Box): hier bei 15, da für $\alpha = 0.25$ mit $14 \cdot 0.25 = 3.5$ nach der Formel $x_\alpha = \min\{x : F(x) \geq \alpha\}$ der 4. Wert zählt.

3. oberes Quartil (obere Begrenzung der Box): hier bei 20, da für $\alpha = 0.75$ mit $14 \cdot 0.75 = 10.5$ nach der Formel $x_\alpha = \min\{x : F(x) \geq \alpha\}$ der 11. Wert zählt.

4. die sogenannten Whiskers und Ausreißer:
   berechnet nach der Formel: $x < x_{0.25} - 1.5QA$ bzw. $x > x_{0.75} + 1.5QA$.
   Bei einem Quartilabstand von $QA = x_{0.75} - x_{0.25} = 20 - 15 = 5$ sind die Ausreißer also bei weniger als 7.5 (hier also keine, da es keinen Wert unter 7.5 gibt) bzw. mehr als 27.5 (hier zwei: 33 und 38) einzuzeichnen. Die Whiskers werden bis zu den letzten Werten gezeichnet, die unter 27.5 bzw. ober 7.5 liegen.



Das arithmetische Mittel wird nach der Formel

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{16+15+20+20+10+10+19+14+38+19+19+33+24+19}{14} \\ &= \frac{276}{14} = 19.714\end{aligned}$$

berechnet.
Für den Median gilt, da eine gerade Anzahl an Daten vorliegt:

$$\begin{aligned}x_{med} &= \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \\ &= \frac{1}{2}(19+19) = 19.\end{aligned}$$

Folglich liegt der Median geringfügig unter dem arithmetischen Mittel.

Aufgabe 3

Bei insgesamt 11 Spielern, also einer ungeraden Anzahl, liegt der Median in diesem Beispiel bei 25 Jahren. Nachdem der 40 jähriger Torwart gegen einen 18 jährigen Torwart

ausgetauscht wurde, hat sich nichts am Median geändert. Zwar sind nun nicht mehr 4 sondern 5 Spieler jünger als 25 Jahre sind, immernoch 3 25 Jahre und nur noch 3 älter als 25, aber der 'Wert in der Mitte' also der 6. Wert ist immernoch 25 Jahre. Der neue Mittelwert (arithmetisches Mittel) muss wiederum berechnet werden. Ursprünglich waren es 11 Spieler mit einem Gesamtalter von 308 Jahren. Warum?

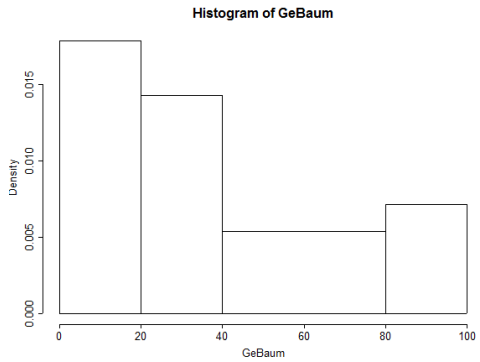
$$\bar{x}_{alt} = 28 = \frac{1}{11}\sum x_i$$
$$\Rightarrow \sum x_i = 28 \cdot 11 = 308$$

Nun werden die Torwärte ausgetauscht, d.h.

$$\bar{x}_{neu} = \frac{308 - 40 + 18}{11} = 26.$$

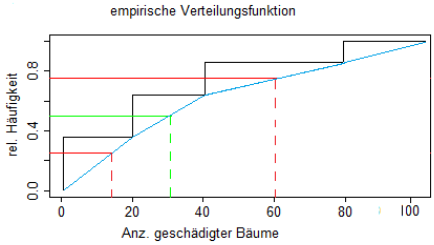
Aufgabe 6

a) Wichtig beim Histogramm ist die Tatsache, dass der Flächeninhalt der Anzahl n_j der Daten entspricht, die im Gruppenintervall $(g_{j-1}, g_j]$ liegen. Hier heißt dies besonders für die dritte Gruppe, da dort das Intervall doppelt so groß wie das der anderen Gruppen ist, dass die Höhe halbiert werden muss. Die Abbildung zeigt ein Histogramm mit relativer Häufigkeit, d.h. der Flächeninhalt aller Säulen zusammen ergibt genau 1. Das Histogramm mit absoluter Häufigkeit sieht genauso aus, lediglich die Skalierung der y-Achse (density) muss mit 20 multipliziert werden.



b)

c)
Die in rot eingezeichneten Linien geben die Quartile $x_{0.25}$ und $x_{0.75}$ an, während die grüne Linie den Median kennzeichnet. Für die graphische Bestimmung muss also nur noch 'abgelesen' werden, d.h. an welcher Stelle bzgl. der x-Achse trifft die (rote/grüne) Linie auf die Verteilungsfunktion. Demnach würde $x_{0.25}$ im Bereich $(0, 20]$, $x_{0.5}$ im Bereich $(20, 40]$ und $x_{0.75}$ im Bereich $(40, 80]$ liegen.



Besser ist jedoch der Weg über Interpolation, d.h. die 'unteren' Ecken der Verteilungsfunktion werden durch einen Linienzug verbunden. Dann können die Werte bezüglich der Interpolierten (blaue Linie) abgelesen werden. Die Werte sollten sich dann an den Werten aus Aufgabe d) orientieren.
Bezüglich der Achsenskalierung sollte man darauf achten, dass bei einer empirischen Verteilungsfunktion die y-Achse auf $[0, 1]$ skaliert ist, d.h. die relativen Häufigkeiten angibt.
d)
Wieder wird der Ansatz der Interpolation verwendet. Damit ergibt sich für das 25%-Quartil:

$$\begin{aligned} \alpha = 0.25 \Rightarrow N_{25} &= 70 \cdot 0.25 = 17.5 \\ \Rightarrow x_{0.25} &= \frac{N_{25} - 0}{25} 20 = 14, \end{aligned}$$

für den Median

$$\begin{aligned} \alpha = 0.5 \Rightarrow N_{50} &= 70 \cdot 0.5 = 35 \\ \Rightarrow x_{0.5} &= 20 + \frac{N_{50} - 25}{20} 20 = 30, \end{aligned}$$

und letztlich für das 75%-Quartil

$$\begin{aligned} \alpha = 0.75 \Rightarrow N_{75} &= 70 \cdot 0.75 = 52.5 \\ \Rightarrow x_{0.75} &= 40 + \frac{N_{75} - 25 - 20}{15} 40 = 60. \end{aligned}$$

e)
Nach d) also $[14, 60]$. Es wurde für diese Hausaufgabe aber auch $[0, 80]$ als richtig angesehen.
f)

Nach d) ergibt sich der Quartilsabstand durch $QA = x_{0.75} - x_{0.25} = 60 - 14 = 46$.
Der Mittelwert beträgt

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum x_i \\ &= \frac{25 \cdot 10 + 20 \cdot 30 + 15 \cdot 60 + 10 \cdot 90}{70} = 37.857.\end{aligned}$$

Die Varianz errechnet sich dann analog

$$\begin{aligned}var(x) &= \frac{1}{n}\sum (x_i - \bar{x})^2 \\ &= \frac{25 \cdot (10 - \bar{x})^2 + 20 \cdot (30 - \bar{x})^2 + 15 \cdot (60 - \bar{x})^2 + 10 \cdot (90 - \bar{x})^2}{70} = 788,265.\end{aligned}$$

Im Unterschied zu den ursprünglichen Daten liegen hier nur die gruppierten Werte vor; diese sind aber nicht so genau.