Institut für Angewandte Mathematik
Sommersemester 2022
Andreas Eberle, Luis Aniello La Rocca

10. Übungsblatt „Einführung in die Statistik"

Abgabe bis Mittwoch 22.6., 16 Uhr.

---

1. (Robustheit und Bruchpunkt) Bestimmen Sie die asymptotischen Bruchpunkte der folgenden Statistiken von reellwertigen Stichproben $x_1, \ldots, x_n$:

a) Spannweite,

b) Stichprobenquantil $\hat{q}_\gamma$, $\gamma \in (0,1)$,

c) MAD.

Hinweis: Sie finden Aufgabenteil b) auch als Lemma 4.2 im Buch von Dümbgen. Sinnvoller ist es aber, wenn Sie sich selbst überlegen, wie man den Bruchpunkt bestimmt, und dies mit eigenen Worten erklären. Dabei dürfen Sie auch etwas anschaulich argumentieren (aber trotzdem möglichst präzise).

2. (Box-Plots, Mittelwerte und Quantile) Angenommen, Sie kennen von einem Beobachtungsvektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ nur die fünf Kenngrößen

$$Q_0 \;:=\; \min(\boldsymbol{X}), \quad Q_j \;:=\; \hat{q}_{j/4}(\boldsymbol{X}) \;\text{ für } j = 1,2,3 \quad \text{und} \quad Q_4 \;:= \max(\boldsymbol{X}).$$

Nicht einmal der Stichprobenumfang n sei Ihnen bekannt.

a) Zeigen Sie, dass

$$\frac{Q_0 + Q_1 + Q_2 + Q_3}{4} \;\le\; \overline{X} \;\le\; \frac{Q_1 + Q_2 + Q_3 + Q_4}{4}.$$

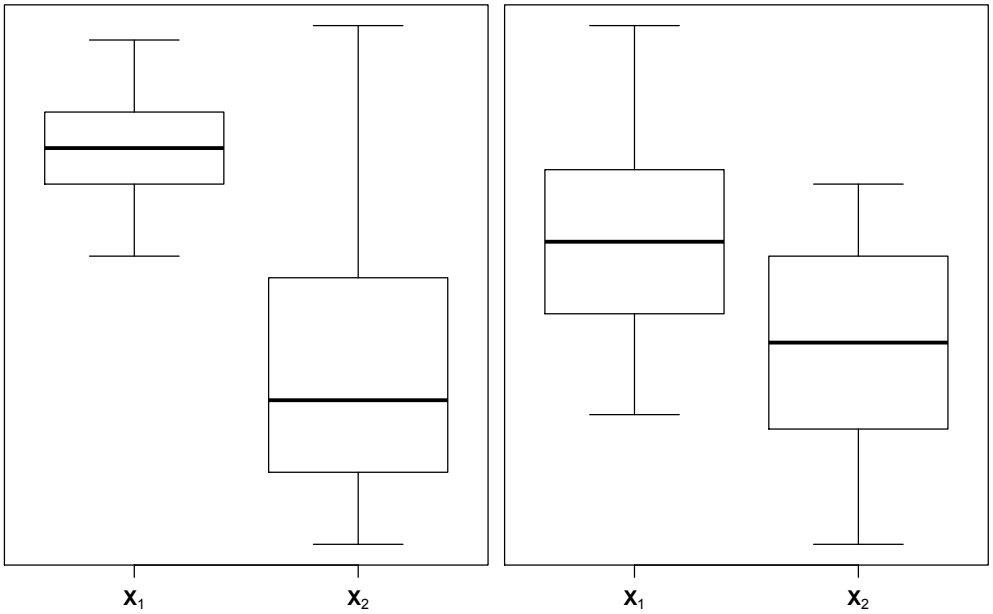
Hinweis: Die Größen $Q_0, \ldots, Q_4$ bleiben unverändert, wenn man jede Komponente von $\boldsymbol{X}$ durch k Kopien ersetzt und somit eine Stichprobe vom Umfang kn erhält. Sie dürfen also davon ausgehen, dass n ein beliebig großes Vielfaches von 4 ist.

b) Angenommen, $X_1, \ldots, X_n$ sind unabhängig und identisch verteilt mit stetiger Verteilungsfunktion. Mit welcher Wahrscheinlichkeit ist $[q_{1/4}, q_{3/4}] \subseteq [Q_0, Q_4]$? Wie groß muss n sein, damit diese Wahrscheinlichkeit mindestens 99% beträgt? Mit welcher Wahrscheinlichkeit ist $q_{1/2} \in \big[Q_1, Q_3\big]$, wenn n kein Vielfaches von 4 ist?

3. (Mann-Whitney-U-Statistik und Boxplots) Die Abbildungen unten zeigen Box-Plots zweier Stichproben $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ mit unbekannten Stichprobenumfängen n_1 bzw. n_2.

a) Bestimmen Sie aufgrund dieser Box-Plots untere und obere Schranken für die normierte sogenannte Mann-Whitney-U-Statistik

$$\hat{u} \;=\; \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_{1i}, X_{2j}) \quad \text{mit} \quad h(x,y) := 1_{[x>y]} + 1_{[x=y]}/2.$$



b) Welche anschauliche Interpretation hat $\hat{u}$? Für welche Kenngröße ist $\hat{u}$ ein Schätzwert, und für welche Testprobleme kann man $\hat{u}$ als Teststatistik verwenden?

4. (Datensätze mit R analysieren) Auf der webpage "ggobi.org/book" finden Sie unter "Data" diverse Datensätze im csv-Format, deren Zusammensetzung in der pdf-Datei "Data Descriptions" beschrieben ist. Einen solchen Datensatz können Sie in R mit dem Befehl x< −read.csv("http://.... .csv") einlesen und unter x abspeichern. Mit x$y können Sie dann auf den Teildatensatz zum Merkmal y zugreifen. Dies ist äquivalent zu x[[k]], wenn das Merkmal y in der k-ten Spalte steht.

a) Lesen Sie den Datensatz "Tips" ein, und lassen Sie sich diesen mit View(x) ausgeben. Erstellen Sie einen Scatterplot der Höhe des Trinkgelds (x$tip) in Abhängigkeit von der Höhe der Rechnung (x$totbill). Erstellen Sie auch Boxplots und Histogramme dieser Merkmale.

b) Es liegt nahe, dass die Höhe des Trinkgelds in vielen Fällen proportional zur Höhe der Rechnung ist. Daher betrachten wir das Verhältnis r< −xtip/xtotbill. Erstellen Sie auch hier einen Scatterplot und einen Boxplot.

c) Berechnen Sie Mittelwerte und getrimmte Mittelwerte für verschiedene Werte $\tau \in [0, 0.5]$. Wie kommt die Abhängigkeit der getrimmten Mittelwerte von τ zustande?

d) Erstellen Sie mithilfe von qqnorm(r) und qqline(r) einen Normal-Q-Q Plot. Was ist in einem solchen Plot dargestellt, und was können Sie daraus ablesen?

e) Berechnen Sie Konfidenzintervalle zum Niveau 95% für den Erwartungswert und den Median von r.

f) Untersuchen Sie die Abhängigkeit der Höhe der Gesamtrechnung von den anderen (kategoriellen) Merkmalen grafisch. Dies geht zum Beispiel mit dem Befehl boxplot(totbill∼sex+smoker+time,data=x).