Universität Bremen

FB3, A. Mändle WS 2019/20, 18.11.2019

Einführung in R
6. Aufgabenblatt

Präsenzaufgabe 1

Ziehen Sie zufällig (`sample()`) ohne Zurücklegen eine Stichprobe vom Umfang $n = 1000$ aus dem Datensatz *bilirubin.txt*.

Hinweis: Verwenden Sie `hist()` mit der Option `freq=F`.

a. Veranschaulichen Sie die Verteilung der Variablen `Alter` und `Wert` durch Histogramme und Boxplots. Zeichnen Sie in Ihre Histogramme Dichteschätzungen hinein.

b. Erstellen Sie zwei Histogramme (mit Dichteschätzungen) für die Variable `Alter` für Männer und Frauen getrennt.

c. Veranschaulichen Sie die Verteilung der Bilirubin-Werte gruppiert nach dem Geschlecht mit einem gruppierten Boxplot.

Präsenzaufgabe 2

Laden Sie den Datensatz `iris`.

a. Veranschaulichen Sie die Verteilung der metrisch-skalierten Merkmale durch Histrogramme und Boxplots.

b. Veranschaulichen Sie die Verteilung der metrisch-skalierten Merkmalen gruppiert nach dem Merkmal `Species` mit den gruppierten Boxplots.

Präsenzaufgabe 3

Laden Sie den Datensatz `airquality`. Wir betrachten die Merkmale `Ozone`, `Solar.R`, `Wind` und `Temp` in diesem Datensatz.

a. Veranschaulichen Sie die Verteilung von `Ozone`, `Solar.R`, `Wind` und `Temp`.

b. Veranschaulichen Sie den Zusammenhang je zweier Merkmale durch Streudiagramme.

c. Bestimmen Sie die Kovarianz- und die Korrelationsmatrizen dieser Merkmale (nach *Pearson*-, *Spearman*- und *Kendall*-Methode).

Präsenzaufgabe 4

Den Datensatz `m111survey` aus dem Paket `tigerstats` haben wir in Blatt 5 untersucht. Wir möchten den Zusammenhang zwischen den Variablen `fastest` und `sex` mit Hilfe von gruppierten Boxplots veranschaulichen.

Präsenzaufgabe 5

a. Setzen Sie den Seed mittels `set.seed(123)` und generieren Sie anschließend 1000 Daten aus einer $N(0,1)$-Verteilung (`rnorm(1000)`).

b. Plotten Sie das Histogramm der Daten und beschriften Sie die Achsen und den Titel.

c. Schätzen Sie eine geglättete Dichtefunktion (`density()`) für die Daten und fügen Sie diese Dichtefunktion in das Histogramm ein.

d. Fügen Sie die Dichtefunktion der $N(0,1)$-Verteilung dem oben-erstellten Bild hinzu. (*Hinweis:* `curve(dnorm(x), add=TRUE)`).

Hausaufgabe 1 (2+2 Punkte)

Importieren Sie den Datensatz `gewicht.txt` (aus *StudIP*).

a. Veranschaulichen Sie die Verteilung des Geburtsgewichts durch ein Histogramm und einen Boxplot.

b. Veranschaulichen Sie die Verteilung des Gewichts gruppiert nach dem Rauchverhalten der Mutter mittels gruppierter Boxplots.

Hausaufgabe 2 (2+6(+4) Punkte)

Installieren Sie das R-Package `DAAG`. Der Datensatz `cuckoos` in diesem Paket enthält die Länge und Bereite gelegter Eier von verschiedenen Kuckucksvögeln.

a. Berechnen Sie Mittelwert und Standardabweichung der Variablen `length` und `breadth` für jede Vogelart aus dem Datensatz.

b. Veranschaulichen Sie durch gruppierte Boxplots die Verteilung von `length` und `breadth` für die verschiedenen Vogelarten. Welche Vogelart legt das kleinste Ei? Entspricht dies Ihrer Erwartung im Hinblick auf die Ergebnisse aus a.?

c. **Zusatzaufgabe:** Wir wollen den Unterschied der Größen der Vogeleier auf Signifikanz testen. Dies soll einmal mit Hilfe der *Varianzanlyse* und einmal mit dem *pairwise t-test* geschehen. Interpretieren Sie die Ergebnisse.
Hinweise:

   – Für die Varianzanalyse soeichern Sie zunächst mit der Funktion `lm()` das lineare Modell für den Zusammenhang `length~species` in der Variablen `Model1`. Im Anschluss beurteilen das Modell mithilfe von `anova(Modell)`.

   – Für den *pairwise t-test* wenden Sie die Funktion `pairwise.t.test()` an.

Hausaufgabe 3 (8 Punkte)

Die Funktion *scatterplot* enthalten im Paket *car* ist eine erweiterte Funktion für Streudiagramme. Mittels `scatterplot(y ~ x | z , ...)` kann man ein Streudiagramm für die Punkte (x, y) gruppiert nach z erstellen. Zusätzlich wird dabei eine glatte Funktion durch die Punktwolken gelegt, die mittels dem sogenannten *LOESS*-Verfahren (*locally weighted polynomial regression*) bestimmt wurde.

a. Erstellen Sie mit der Funktion `scatterplot` ein nach der Variable `Species` gruppiertes Streudiagramm für die Merkmale `Sepal.Length`, `Sepal.Width` aus dem `iris`-Datensatz. Verwenden Sie die Option `reg.line=F`.

b. Die Funktion `smooth.spline(x,y)` (aus dem Paket *stats*) schätzt eine andere glatte Funktion, einen sogenannten *smooth spline* zu den Punktwolken (x, y).
Erstellen Sie zunächst mit der Funktion `scatterplot` ein (diesmal *nicht* gruppiertes) Streudiagramm für `Sepal.Length` und `Sepal.Width`. Schätzen Sie dann dazu passend einen *smooth spline* für die Merkmale `Sepal.Length` und `Sepal.Width` und speichern Sie das Ergebnis in einer Variablen *smspline*. Fügen Sie die geschätze Kurve (also den Smooth Spline) dann dem Bild hinzu (Hinweis: `lines()` auf den Smooth Spline anwenden.).

Abgabe der Lösungen: bis ***Montag 25.11.2019,***

maendle@uni-bremen.de