Übungsblatt vom 06.01.2015

Aufgabenkomplex 1

Der Datensatz OECD enthält Variablen (Stand 2009), die das Wohlergehen von Kindern in den Mitgliedsstaaten messen sollen. Abgefragt wurde:

- Kontinent: Kontinent dem das Land zugehört
- Europa: 1=europäisches Land; 0=außereuropäisches Land
- Einkommen: das durchschnittliche Einkommen der Eltern in US Dollar
- Armut: der Anteil [in Prozent] an Kindern in einem armen Elternhaus
- Bildung: der Anteil [in Prozent] an Kindern, die ohne eine Grundausstattung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen
- WenigRaum: der Anteil [in Prozent] an Kindern, die auf zu wenig Raum wohnen
- Umwelt: der Anteil [in Prozent] an Kindern, die unter schlechten Umweltbedingungen leben
- Lesen: mittlerer PISA-Score zur Lesefähigkeit
- Geburtsgewicht: der Anteil [in Prozent] an Kindern, die bei der Geburt weniger als $2.5kg$ wiegen
- Säuglsterblichkeit: Säuglingssterblichkeit (<1 Jahr) [x in Tausend]
- Sterblichkeit: Sterblichkeit (<20 Jahre) [x in 100 000]
- Selbstmord: Selbstmord von Jugendlichen im Alter von 15 bis 19 [x in 100 000]
- Bewegung: der Anteil [in Prozent] an 11, 13 und 15-jährigen Jugendlichen, die sich regelmäßig bewegen
- Rauchen: der Anteil [in Prozent] an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen
- Alkohol: der Anteil [in Prozent] an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren
- Bullying: der Anteil [in Prozent] an Kindern, die angeben in der Schule bedroht zu werden
- Schule: der Anteil [in Prozent] an Kindern, die angeben die Schule zu mögen

(a) Lesen Sie den Datensatz *oecd.txt* mit der Funktion `data<-read.table(...)` ein und überprüfen Sie die Dimension der Daten.

(b) Berechnen Sie die Mittelwerte und Varianzen für geeignete Variablen mit einem `apply` Befehl.

(c) Überprüfen Sie, ob die Niederlande im der Länderliste des des Datensatzes auftaucht. Gibt es auch einen Eintrag für China? (Benutzen sie die `R`–Hilfe, um herauszufinden wie man auf die Ländernamen zugreifen kann.)

(d) In welchem Land waren die meisten Jugendlichen mindestens zweimal betrunken? Wie hoch ist der maximale Prozentsatz?

(e) In welchem Land ist die Säuglingssterblichkeit am geringsten? Wie hoch ist sie in diesem Land?

(f) In welchen Ländern ist der Prozentsatz an Jugendlichen, die sich regelmäßig bewegen kleiner als der Durchschnitt?

(g) In welchen Ländern werden besonders viele Kinder in der Schule bedroht? Als Indikator für "besonders viel" soll ein "Bullying"-Wert gelten, der mindestens eine Standardabweichung (*standard deviation*) vom Mittelwert aller Ländern entfernt ist.

(h) Erstellen Sie einen neuen Datensatz, der aufsteigend nach dem Einkommen geordnet ist. Speichern Sie diesen in einer neuen Datei ab.

(i) Erstellen sie die Variable Einkommen_binär. Diese erhält den Wert 0 wenn das Einkommen des Landes kleiner als das mediane Einkommen aller Länder ist und ansonsten 1. Füge die neue Variable in den Datensatz ein.

(j) Installieren sie das Paket `gmodels`. Informieren sie sich in der Hilfe über die Funktion CrossTable.

(k) Stellen sie den Zusammenhang der Variablen Einkommen_binär und Europa sowie Einkommen_binär und Kontinent mit CrossTable dar. Konfigurieren sie die Funktion um eine möglichst übersichtliche und ansprechende Darstellung zu erhalten.

(l) Stellen sie die Zusammenhänge graphisch dar. Versuchen sie hierbei die Graphikparameter der entsprechenden Funktionen an ihre Bedürfnisse anzupassen.

(m) Visualisieren Sie die Variable Lesen, getrennt nach dem Faktor Europe in einem vertikalen Stripchart.

(n) Erstellen Sie einen Boxplot für die Variable "Bildung". Was fällt Ihnen auf?

(o) Untermauern Sie die Beobachtung aus Aufgabe (a) durch Berechnung einiger Quantile mit Hilfe der Funktion `quantile()`.

(p) Stellen Sie zudem die aufsteigend geordneten Werte der Variable *Bildung* mit Hilfe der Funktion `plot()` als Kurve dar.

(q) Begründen Sie anhand Ihrer Beobachtungen, dass das 75% Quantil der Daten einen guten Trennpunkt zwischen Ländern mit "guter" und "schlechter" Grundausstattung für Bildung darstellt.

(r) Bestimmen Sie eine Liste mit Ländern, in denen es an der Grundausstattung für Bildung besonders fehlt.

Aufgabenkomplex 2

Aufgabe: Das Geburtstagsparadoxon

Die Wahrscheinlichkeit des Ereignisses A_n, dass unter einer Anzahl von n Leuten mindestens 2 Leute am gleichen Tag Geburtstag haben, beträgt:

$$P(A_n) = 1 - \frac{365!}{(365-n)!365^n}$$

Zur praktischen Berechnung verwendet man eine Umformung mittels Logarithmus und Exponentialfunktion:

$$P(A_n) = 1 - \exp\big[\log(365!) - \log[(365-n)!] - n \cdot \log(365)\big]$$

Den Logarithmus der Fakultät erhält man in `R` per `lfactorial()`–Befehl, die Exponentialfunktion per `exp()`.

(a) Schreiben Sie eine `R`-Funktion, die die Wahrscheinlichkeit von A_n berechnet.

(b) Erstellen Sie eine Graphik, die diese Wahrscheinlichkeit von A_n in Abhängigkeit von n darstellt. Fügen Sie auch eine aussagekräftige Beschriftung hinzu.

HINWEIS: Sie können Ihrer Funktion mehrere n-Werte per Vektor übergeben. Dann erhalten Sie einen Vektor mit den jeweiligen Ergebnissen zurück. Dies ist z.B. bei der Erstellung des Graphen praktisch.

Aufgabe: Datensätze durchsuchen

(a) Schreiben Sie eine `R`-Funktion, die die Anzahl von Werten in einem Datensatz ermittelt, die in einem bestimmten offenen Intervall (a, b) liegen. Dabei sollten Sie auch auf fehlende Werte achten und Variablen ausschließen, die nicht numerisch sind (z.B. Faktoren oder Strings).

(b) Testen Sie die Methode anhand des *oecd*-Datensatzes.

Aufgabenkomplex 3

Aufgabe: Testen an PISA-Daten

Das *Programme for International Student Assessment*, kurz PISA, ist eine standardisierte Bewertung von (15 jährigen) Schülern unter den teilnehmenden Staaten. Ziel der Regierungen ist, eine Datenbasis zur länderübergreifenden Forschung zu ermöglichen. Im Datensatz *PISA.csv* finden Sie die Ergebnisse einiger ausgwählter OECD-Staaten, getrennt nach dem Geschlecht (Variable *sex*: 1 Female, 2 Male, Perc_Sex gibt den Anteil an). Folgende Variablen sind von Interesse:

- *R00* - *R06*: Mittlerer Score zur Lesekompetenz im Jahr 2000 bzw. 2006
- *M00* - *M06*: Mittlerer Score zur Kompetenz in der Mathematik im Jahr 2000 bzw. 2006
- *S00* - *S06*: Mittlerer Score in den Naturwissenschaften (*science*) im Jahr 2000 bzw. 2006

1. Laden Sie den Datensatz *PISA.csv* von der Homepage herunter und lesen sie ihn ein.
2. Untersuchen Sie deskriptiv, ob sich die drei PISA-Scores des Jahres 2006 im Vergleich zum Jahr 2000 verändert haben. (Gehen Sie hierbei und im weiteren nicht näher auf irgendwelche Geschlechtsunterschiede ein)
3. Untersuchen Sie mit einem geeigneten Test, ob sich die drei PISA-Scores signifikant verändert haben.

Aufgabe: Google-Suche in Deutschland

Im Datensatz "google.csv" finden Sie getrennt nach dem Bundesland die absolute Suchhäufigkeit nach:

- "Krise" im Jahr 2008 (Krise08)
- "Krise" im Jahr 2009 (Krise09)
- "Arbeitsamt" im Jahr 2010 (Arbeitsamt)

Dazu enthält der Datensatz einen Ost-West Indikator (OW) (Berlin wird von dieser Kategorisierung ausgenommen).

a) Lesen Sie den Datensatz ein und wandeln Sie die Variable "OW" in einen Faktor um, der fehlende Wert zulässt und mit 0 West und 1 Ost kodiert.

b) Führen Sie den geeigneten t-Test durch, um zu untersuchen, ob nach dem Begriff "Krise" im Jahr 2008 stärker gesucht wurde als im Jahr 2009. Erstellen Sie zusätzlich einen Boxplot und führen Sie den entsprechenden nonparametrischen Test durch. Interpretieren Sie die Ergebnisse.

c) Überprüfen Sie mittels des Ost-West Indikator, ob diese Abnahme der Suche nach dem Begriff "Krise" im Osten stärker ausgeprägt ist als im Westen.

d) Untersuchen Sie die Hypothese, dass im Osten häufiger nach dem Begriff "Arbeitsamt" gesucht wird als im Westen. Begründen Sie Ihre Entscheidung.

Aufgabe: Zu den Annahmen des t-Test

1. Ziehen Sie 100 exponential-verteilte Zufallszahlen mit dem Parameter $\lambda = 0.1$ und speichern Sie diese in dem Objekt `X1`. Erstellen Sie analog ein Objekt `X2` von 100 Zufallszahlen, für die Sie erneut 100 exponential-verteilte Zufallszahlen `HX2` mit dem Parameter $\lambda = 0.1$ ziehen und anschließend alle Elemente von 20 subtrahieren. Ein Element i des Vektors `X2` berechnet sich also als `20-HX2[i]`.

2. Zeichnen Sie für die beiden Objekte in einer Graphik je den Kerndichteschätzer.

3. Führen Sie den t-Test durch, um zu untersuchen, ob die beiden Objekte unterschiedliche Mittelwerte besitzen.

4. Führen Sie den Wilcoxen Rangsummen Test durch, um zu untersuchen, ob die beiden Objekte unterschiedliche Mediane besitzen.