Übungen

Die Aufgaben können in **2er Gruppen** bearbeitet werden. Für jeden Aufgabenblock sollen Sie ein R Skript erstellen, das Sie als „`Blocknr.Vorname1.Vorname2.R`“ speichern, also z.B. „`3.Bernd.Ute.R`“ für das R Skript zum Block 3. Bitte Kennzeichnen Sie in ihren Skripten deutlich den Anfang einer neuen Aufgabe und eines neuen Aufgabenteile.
Abgabe: **02.04.2014 bis 20:00** Uhr per Email

5 Grundlegende Definitionen, Merkmalstypen

Grundlegende Definitionen

Aufgabe 5.1. (4 Punkte)
Füllen Sie den Fragebogen aus und überlegen Sie sich:

(a) Was sind hier die statistischen Einheiten?

(b) Was ist die Grundgesamtheit?

(c) Welche Stichprobe wird befragt?

(d) Welche Merkmale und zugehörige Merkmalsausprägungen gibt es?

Merkmalstypen

Aufgabe 5.2. (4 Punkte)
Diskutieren Sie die von Ihnen in Aufgabe 5.1 gefundenen Merkmale auf dem Fragebogen hinsichtlich ihres jeweiligen Skalenniveaus. Entscheiden Sie zudem, ob es sich um diskrete oder stetige, bzw. quantitative oder qualitative Merkmale handelt.

(Bitte wenden!)

6 Funktionsverläufe

Aufgabe 6.1. (4 Punkte)

(a) Skizzieren Sie die unten angegebenen Verteilungen. Zeichnen Sie dabei die verschiedenen Verteilungen einer Verteilungsfamilie in eine Grafik.

- $Exp(\lambda)$, $\lambda = 0.5, 1, 4$ (Exponentialverteilung)
- $\Gamma(\alpha, \beta)$, $\alpha = 0.5, 1, 4$, $\beta = 0.5$ (Gammaverteilung)

(Optional: Informieren Sie sich über die Funktion `legend`. Sie könnte helfen, die Grafiken übersichtlicher zu gestalten.)

(b) Skizzieren Sie die Binomialverteilung $B(n, p)$ mit $n = 1000$ und $p = 0.01$ auf einem geeigneten Intervall. Nutzen Sie dafür den 'plot type' `h`. Fügen Sie der Grafik die Punktwahrscheinlichkeiten der $Pois(10)$-Verteilung hinzu. Benutzen Sie dazu den 'plot type' `p`. Was fällt ihnen auf und aus welchem Grund tritt das beobachtete Phänomen ein?

Aufgabe 6.2. (4 Punkte)
Der Zentrale Genzwertsatz besagt folgendes: Sei $(X_i)_{i\geq 1}$ eine Folge unabhängig und identisch verteilter Zufallsgrößen mit Erwartungswert μ und positiver, endlicher Varianz σ^2. Dann gilt

$$\frac{\sum_{k=1}^{n}(X_k - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \quad n \to \infty.$$

Approximieren Sie die Standardnormalverteilung durch die $Exp(1)$-Verteilung wie folgt:

- Simulieren Sie 1000 mal die obige Summe im Zentralen Grenzwertsatz mit $n = 1000$ und $X_1 \sim Exp(1)$.
- Sortieren Sie den entspandenen Vektor und plotten Sie die empirische Verteilungsfunktion.
- Fügen Sie die Verteilungsfunktion der Standardnormalverteilung ins Diagramm ein.

Hinweis: Die Funktionen `colSums` und `table` könnten hilfreich sein.

(Bitte wenden!)

7 Darstellung univariater Daten

Grafische Darstellungen

Aufgabe 7.1. (4 Punkte)
Bei der letzten Landtagswahl in Nordrhein-Westfalen am 13. Mai 2012 ergab sich laut amtlichem Endergebnis folgende Stimmverteilung (bei 7.780.610 gültigen Erststimmen):

CDU	SPD	Grüne	Piraten	FDP	Andere
2.545.309	3.290.561	723.581	617.926	372.727	230.506

(a) Geben Sie die Daten als Vektor ein und ordnen Sie den Vektor absteigend. Berechnen Sie die zugehörigen prozentualen Anteile an den abgegebenen (und gültigen) Stimmen auf eine Nachkommastelle genau.

(b) Erzeugen Sie mit den Daten aus (a) ein mit den Parteinamen und den zugehörigen Prozentzahlen beschriftetes Kreissektorendiagramm (in den entsprechenden Parteifarben) (**Tipp:** Benutzen Sie zum zusammenfügen von Parteiname und Prozentzahl die `paste` Funktion)

(c) Erstellen Sie ein geordnetes Säulendiagramm in den entsprechenden Parteifarben.

Darstellung und Kenngrößen quantitativer Merkmale

Aufgabe 7.2. (4 + 2Punkte)
Lesen Sie den Datensatz `nettomieten.csv` ein. Die erste Spalte mit dem Namen `nm` gibt das Mermal "Nettomiete" wieder. Lesen Sie diese in einen Vektor ein.

(a) Nehmen Sie eine geeignete Klasseneinteilung vor und listen Sie die absoluten Häufigkeiten auf (**Tipp:** Nutzen Sie den Befehl `table`).

(b) Zeichnen Sie ein Säulendiagramm und ein Histogramm mit der in (a) gewählten Klasseneinteilung des Datensatzes und einer Einteilung ihrer Wahl.

(c) Beurteilen Sie die Schiefe der Verteilung.

(d) Berechnen Sie arithmetisches Mittel, Median, Varianz und Spannweite des Merkmals Nettomiete.

(e*) Berechnen Sie das arithmetische Mittel, den Median und den Modus der in (a) gruppierten Daten. Was fällt Ihnen auf?

(Bitte wenden!)

Quantile, Boxplots und Normal-Quantil-Plots

Aufgabe 7.3. (4 Punkte)

Betrachten Sie wieder die Daten mit Merkmal Nettomieteäus `nettomieten.csv` .

(a) Zeichnen Sie einen Boxplot des Nettomietendatensatzes.

(b) Zeichnen Sie den NQ-Plot des Nettomietendatensatzes. Zeichnen Sie anschließend zum Vergleich ein NQ-Plot eines mit der Normalverteilung generierten Datensatzes. Wählen Sie dazu Anzahl der Zufallsdaten, sowie Mittelwert und Varianz entsprechend zu den Daten aus dem Nettomietendatensatz. Ist der Nettomietendatensatz normalverteilt?

(c) Vergleichen Sie den Nettomietendatensatz mit Hilfe von Histogrammen, Box- und QQ-Plots mit den Verteilungen $B(57000, 0.01)$-, der $Pois(570)$- bzw. der $\Gamma(5.7, 0.01)$. Mit welcher Verteilung stimmt der Nettodatensatz am ehesten überein und warum? (**Optional:** Sie können die Funktion `par(mfrow=..)` benutzen um mehrere Plots nebeneinander zu zeichnen.)

8 Zusatzaufgaben

Aufgabe 8.1. * (4 Punkte) Seien $(X_{n,k})_{n,k\geq 0}$ unabhängige und identisch verteilte Zufallsgrößen mit werten in $\mathbb{N}_0$. Sei weiter $Z_0 = 1$ und

$$Z_{n+1} \;=\; \sum_{k=1}^{Z_n} X_{n,k}$$

for $n \geq 1$. Ein solcher stochastischer Prozess $(Z_n)_{n\geq 0}$ heißt *Galton-Watson-Prozess* und beschreibt die Generationsgröße einer Population, in welcher die Individuen unabhängig und mit der gleichen Verteilung Nachkommen zeugt.

(a) Informieren Sie sich über den Galton-Watson-Prozess. Was beschreiben die Zufallsvariablen $(X_{n,k})_{n,k\geq 0}$?

(b) Simulieren Sie den Prozess $(Z_n)_{n\geq 0}$ für ein geeignetes n und $X_{1,1} \sim Pois(\lambda)$ mit $\lambda = 0.8, 1$ und 1.3. Stellen Sie den Populationsverlauf graphisch dar.

(c) Was fällt Ihnen auf? Interpretieren Sie das Ergebnis.

Aufgabe 8.2. * (4 Punkte)

In den Jahren 1931 und 1932 wurden in Minnesota auf sechs verschiedenen Farmen (`site`) jeweils zehn verschiedene Gerstesorten (`variety`) angebaut und die Ernteerträge (`yield`) gemessen. Diese Daten sind in `barley` im Paket `lattice` enthalten. Pakete lädt man in `R` mit dem Befehl `library`.

(a) Erstellen Sie für beide Jahre eine Zusammenfassung (`summary`) der Daten mit Hilfe der Funktion `by`, deren Verwendung im Hilfetext erläutert wird. Erstellen Sie ebenso Zusammenfassungen der Ertragsdaten `yield` aufgeteilt nach Gerstesorten.

(b) Berechnen Sie jeweils für beide Jahre die 0.9-Quantile der Ertragsdaten. Welche Farmen bilden die ca. zehn Prozent mit höheren Erträgen?

(c) Plotten Sie die Art gegen den Ertrag in Abhängigkeit von Ort und Jahr (`variety` $\sim$ `yield | site+year`) mit Hilfe der Funktion `dotplot`. Die entstandene Grafik ist ein Trellis-Display. Was lässt sich daraus ablesen? Erstellen Sie ein weiteres Trellis-Display mit einer anderen Aufteilung. Was lässt sich aus diesem ablesen?