Serie 2

Vorbesprechung: 27/28. Februar 2013

Aufgabe 1

Bei der Ermittlung der landwirtschaftlichen Nutzfläche von Bauernhöfen in einem Bezirk ergaben sich folgende Werte (in ha):

2.1 2.4 2.8 3.1 4.2 4.9 5.1 6.0 6.4 7.3 10.8 12.5 13.0 13.7 14.8 17.6 19.6 23.0 25.0 35.2 39.6

(**a**) Berechnen Sie die Summen $\sum x_i$ und $\sum x_i^2$.

(**b**) Bestimmen Sie den Median .

(**c**) Berechnen Sie den Mittelwert und die Standardabweichung .

Aufgabe 2

Gegeben sind die Datenpaare (x, y)

x	2	2	6	7	7	8	8	9
y	11	14	14	16	27	27	27	38

(**a**) Gesucht sind die Summen $\sum x_i,\ \sum x_i^2,\ \sum y_i,\ \sum y_i^2$ und $\sum x_i \cdot y_i$.

(**b**) Verifizieren Sie die Gleichheit der Formeln $s_{xx} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$ und $s_{xx} = \frac{1}{n-1}\sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2$ anhand der Zahlen in der obigen Tabelle .

(**c**) Beweisen Sie die Gleichheit von Teilaufgabe (b) allgemein .

Aufgabe 3

Zeigen Sie, dass

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

den kleinsten Wert annimmt, falls die Parameter

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ a &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

sind.

Aufgabe 4

Bei einer Firma werden in einem Monat 400 Lebensversicherungsverträge abgeschlossen. Nachstehend ist die klassifizierte Häufigkeitsverteilung für die Versicherungssummen angegeben.

Versicherungssumme (Tausend Fr) von... bis unter...	Anzahl der Verträge
4-10	20
10-20	160
20-30	80
30-40	40
40-80	88
80-120	12

(**a**) Man zeichne ein Histogramm und die Summenkurve für die relativen Häufigkeiten .

(**b**) Untersuchen Sie wieviel Prozent der Versicherten mit höchstens 18'000.- versichert sind, sowie mit welchem Betrag die 20% Personen, die am höchsten versichert sind, mindestens versichert sind. Bestimmen Sie zudem Median und Mittelwert der Verteilung .

(**c**) Berechnen Sie die Standardisierung der Daten .

(**d**) Wenn n Daten $(x_i)_{1\leq i\leq n}$ standardisiert sind (d.h. wenn $\overline{x}=0,\ s_x=1$ gilt), wie gross ist dann $\sum_{i=1}^{n} x_i^2$?

Aufgabe 5

Der Geysir Old Faithful im Yellowstone National Park ist eine der bekanntesten heissen Quellen. Für die Zuschauer und den Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse.
Im File `http://stat.ethz.ch/Teaching/Datasets/geysir.dat` sind die Messungen vom 1.8.1978−8.8.1978 in 3 Spalten abgelegt: "`Tag`", "`Zeitspanne`" und "`Eruptionsdauer`".

(**a**) Zeichnen Sie Histogramme von der Zeitspanne zwischen zwei Ausbrüchen:

```
> geysir <- read.table("http://stat.ethz.ch/Teaching/Datasets/geysir.dat",
+ header = TRUE) ## Datensatz einlesen
> par(mfrow = c(2,2)) ## 4 Grafiken im Grafikfenster
> hist(geysir[,"Zeitspanne"])
> hist(geysir[,"Zeitspanne"], breaks = 20)
> hist(geysir[,"Zeitspanne"], breaks = seq(41, 96, by = 11))
```

Was fällt auf? Was ist der Unterschied zwischen den drei Histrogrammen?
Bemerkung: Wenn man die Anzahl Klassen mit `breaks` = 20 vorgibt, so wird dies nur als "Vorschlag" interpretiert und intern unter Umständen abgeändert.

(**b**) Zeichnen Sie Histogramme (Anzahl Klassen variieren) von der Eruptionsdauer:

```
> hist(geysir[,"Eruptionsdauer"])
```

Was fällt auf? Vergleichen Sie mit der ersten Teilaufgabe.

Aufgabe 6

21 Labors bestimmten den Kupfergehalt von 9 verschiedenen Klärschlammproben. Die Daten stehen im Data Frame `klaerschlamm` zur Verfügung. Die erste Spalte bezeichnet das Labor, die restlichen 9 Spalten sind die verschiedenen Klärschlammprobe. Die Daten (in mg/kg) kann man mit dem Befehl

```
> url <- "http://stat.ethz.ch/Teaching/Datasets/klaerschlamm.dat"
> schlamm.all <- read.table(url, header = TRUE)
> schlamm <- schlamm.all[,-1] ## Labor-Spalte entfernen
```

einlesen.

(**a**) Erstellen Sie für jede Probe einen Boxplot, und berechnen Sie jeweils das arithmetische Mittel und den Median. Bei welchen Proben gibt es Ausreisser, und wo unterscheiden sich arithmetisches Mittel und Median wesentlich? Bei welchen der 9 Proben ist es plausibel, dass die wahre Konzentration unter 400 mg/kg liegt?
R-Hinweise: `summary(schlamm); boxplot(schlamm)` .

(**b**) Erstellen Sie für jedes Labor einen Boxplot der Messfehler. Unter dem Messfehler eines Labors bei einer Probe verstehen wir den gemessenen Wert minus den Median über alle Labors. Welche der 21 Labors haben systematische Fehler in ihrem Analyseverfahren? Welche haben grosse Zufallsfehler, und bei welchen Labors ist die Qualität der Analysen besonders gut?
R-Hinweise:

```
> ## Fuer jede Spalte Median berechnen
> med <- apply(schlamm, 2, median)
> ## Median von jeder *Spalte* abziehen
> schlamm.centered <- scale(schlamm, scale = FALSE, center = med)
> ## Boxplot zeichnen. Dazu zuerst data-frame transponieren
> boxplot(data.frame(t(schlamm.centered)))
```