Lösungen zu Serie 2

Vorbesprechung: 27/28. Februar 2013

Aufgabe 1

```
> x<-c(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**) $\sum x_i =$

```
> sum(x)
```

```
[1] 269.1
```

$\sum x_i^2 =$

```
> sum(x^2)
```

```
[1] 5729.27
```

(**b**) Median:

```
> median(x,na.rm=FALSE)
```

```
[1] 10.8
```

(**c**) Mittelwert:

```
> round(mean(x),4)
```

```
[1] 12.8143
```

Standardabweichung:

```
> round(sd(x),4)
```

```
[1] 10.6793
```

Aufgabe 2

```
> x<-c(2,2,6,7,7,8,8,9)
> y<-c(11,14,14,16,27,27,27,38)
```

(**a**) $\sum x_i =$

```
> sum(x)
```

```
[1] 49
```

$\sum x_i^2 =$

```
> sum(x^2)
```

```
[1] 351
```

$\sum y_i =$

```
> sum(y)
```

```
[1] 174
```

$\sum y_i^2 =$

```
> sum(y^2)
```

```
[1] 4400
```

$\sum x_i \cdot y_i =$

```
> sum(x*y)
```

```
[1] 1209
```

(**b**) $s_{xx} =$

```
> n <- length(x)
> 1/(n-1)*sum((x-mean(x))^2)
```

```
[1] 7.267857
```

alternativ

```
> cov(x,x)

[1] 7.267857
```

$\frac{1}{n-1}\left(\sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2\right) =$

```
> n <- length(x)
> 1/(n-1)*(sum(x^2)-n*mean(x)^2)

[1] 7.267857
```

(**c**)

$$\begin{aligned}\sum_{j=1}^{n} (x_j - \overline{x})^2 &= \sum_{j=1}^{n} \left(x_j^2 - 2x_j\overline{x} + \overline{x}^2\right) \\ &= \sum_{j=1}^{n} \left(x_j^2 + \overline{x}^2\right) - 2\overline{x}\sum_{j=1}^{n} x_j \\ &= n\overline{x}^2 + \sum_{j=1}^{n} x_j^2 - 2\overline{x}n\overline{x} \\ &= \sum_{j=1}^{n} x_j^2 - n \cdot \overline{x}^2 \, .\end{aligned}$$

Aufgabe 3

Gegeben ist

$$F(a,b) \equiv \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \, .$$

Folgende zwei Bedingungen müssen erfüllt sein:

$$\frac{\partial}{\partial a} F(a,b) = \sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \stackrel{!}{=} 0 \tag{1}$$

$$\frac{\partial}{\partial b} F(a,b) = \sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \cdot x \stackrel{!}{=} 0 \tag{2}$$

Aus Gleichung (1) folgt:

$$\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) = 0$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^{n} a = na = \sum_{i=1}^{n} \left(y_i - bx_i\right)$$

$$\Longleftrightarrow$$

$$a = \frac{\sum_{i=1}^{n} \left(y_i - bx_i\right)}{n} = \overline{y} - b\overline{x}$$

Aus Gleichung (2) folgt:

$$\sum_{i=1}^{n} 2\left(y_i - a - bx_i\right) \cdot x = 0$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^{n} \left(y_i - \overline{y} + b\overline{x} - bx_i\right) \cdot x_i = 0$$

$$\Longleftrightarrow$$

$$b \cdot \sum_{i=1}^{n} (x_i - \overline{x})\, x_i = \sum_{i=1}^{n} (y_i - \overline{y}) \cdot x$$

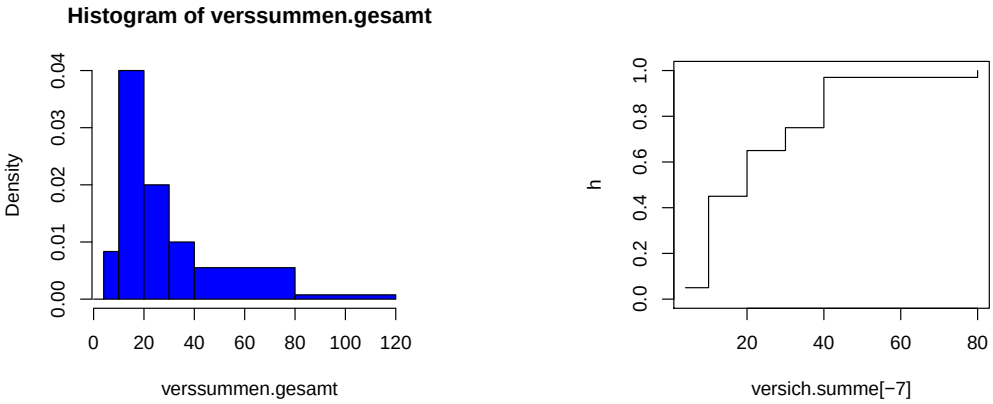
$$\Longleftrightarrow$$

$$b = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, x_i}{\sum_{i=1}^{n} (x_i - \overline{x})\, x_i} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})\, (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}\,.$$

Aufgabe 4

```
> versich.summe <- c(4,10,20,30,40,80,120)
> mittlere.versich.summen <- c(7,15,25,35,60,100)
> anzahl.vertraege <- c(20,160,80,40,88,12)
> verssummen.gesamt <- rep(mittlere.versich.summen,anzahl.vertraege)
> brk <- c(0,4,10,20,30,40,80,120)
> h <- cumsum(anzahl.vertraege)/400

> hist(verssummen.gesamt,breaks=brk,col='blue',xlim=c(4,120))

> plot(versich.summe[-7],h,type="s",ylim=c(0,1))
```

(**a**) Histogramm und Summenkurve:



(**b**) Aus der empirischen Verteilungsfunktionskurve ergibt sich, dass 5% mit höchstens 18'000 Fr. versichert sind; die obersten 20% mit mind. 40'000 Fr. versichert. Der Median der Versicherungssumme ergibt

```
> median(verssummen.gesamt)

[1] 25
```

also einen Wert zwischen 20'000 und 30'000 Fr. Der Mittelwert der Versicherungssummen ist

```
> mean(verssummen.gesamt)

[1] 31.05
```

also ein Wert zwischen 30'000 und 40'000 Fr.

(**c**) Wir standardisieren den Vektor, der alle mittleren Versicherungssummen enthält, und lesen nur die Werte für jeden Verssicherungssummenbereich heraus

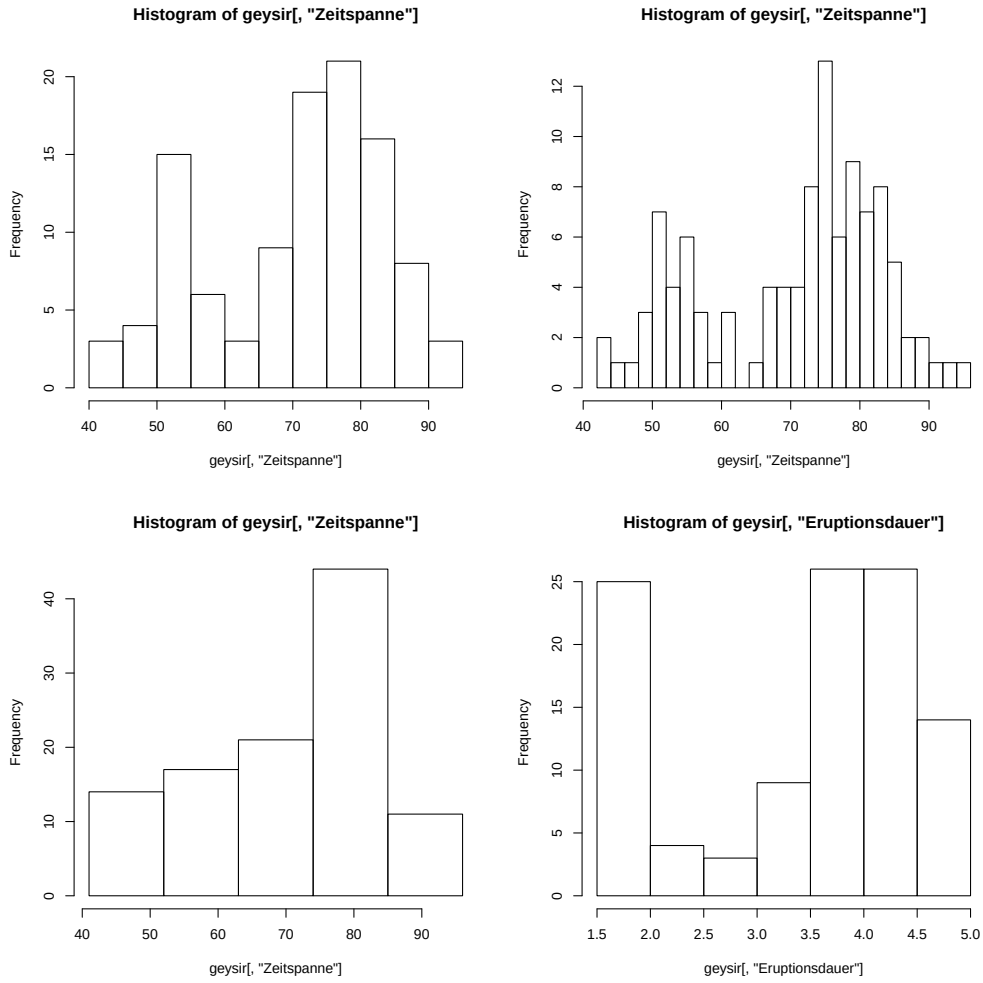
```
> unique(scale(verssummen.gesamt))

            [,1]
[1,] -1.1105021
[2,] -0.7411043
[3,] -0.2793571
[4,]  0.1823902
[5,]  1.3367582
[6,]  3.1837471
```

(**d**) $n - 1$

Aufgabe 5

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

Die ersten drei Histogramme in der Abbildung unten zeigen die Intervalle zwischen zwei Ausbrüchen von Old Faithful. Aufallend ist, dass Zeitspannen um 55 Minuten aber auch zwischen 70 bis 85 Minuten häufiger vorkommen als andere Intervalle. So eine Verteilung mit zwei Gipfeln heisst auch *bimodal*.
Werden die Klassenbreiten ungeschickt gewählt, entdeckt man diese Besonderheit der Geysirdaten nicht. Das ist im dritten Histogramm passiert. Das Beispiel illustriert, dass die richtige Wahl der Klassenbreiten- bzw. grenzen wohlüberlegt sein mus.
Das vierte Histogramm schliesslich zeigt die Häufigkeiten verschiedener Eruptionsdauern. Hier sind die beiden Gipfel sehr deutlich erkennbar: "Entweder ist der Ausbruch sofort wieder vorbei, oder er dauert mindestens dreieinhalb Minuten". Ob die Dauer eines Ausbruchs aber etwas zu tun hat mit der Dauer des vorangegangenen Ruheintervalls (mit anderen Worten: ob die



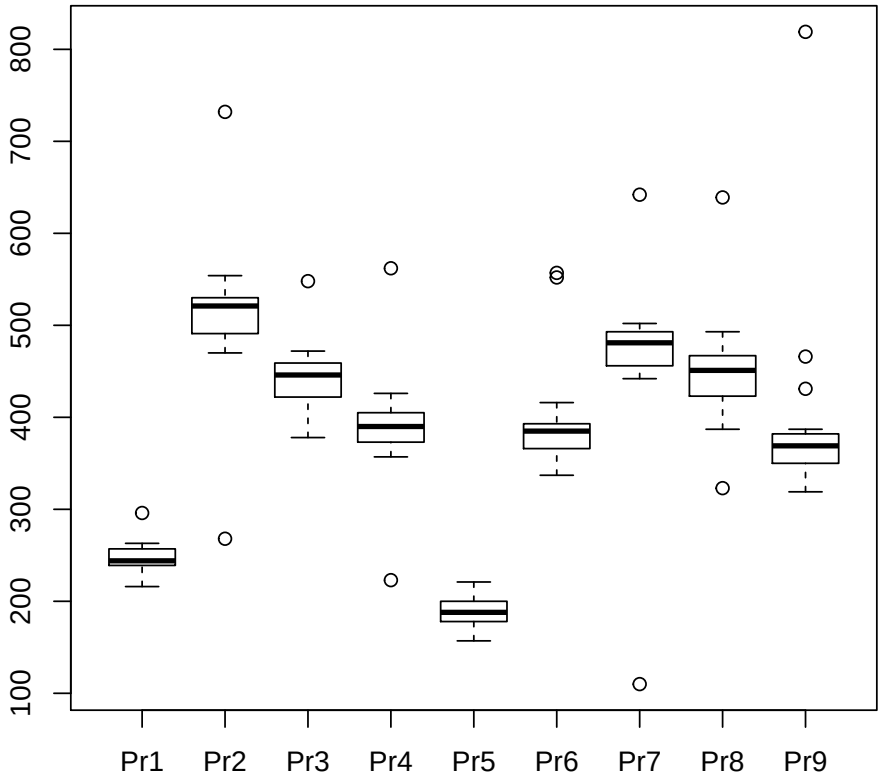
Gipfel des Histogramms aus Teilaufgabe b) den Gipfeln der Histogramme aus Teilaufgabe a) entsprechen), kann man aufgrund dieser Darstellungen nicht sagen.

Aufgabe 6

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

(**a**) Aus den Boxplots erkennen wir, dass es vor allem bei den Proben 2, 4, 6, 7, 8 und 9 Ausreisser gibt. Das arithmetische Mittel und der Median unterscheiden wesentlich bei den Proben 2, 6, 7 und 9.

```
> boxplot(schlamm)
```




```
> summary(schlamm)
```

Pr1	Pr2	Pr3	Pr4	
Min. :216.0	Min. :268.0	Min. :378.0	Min. :223.0	
1st Qu.:239.0	1st Qu.:491.0	1st Qu.:422.0	1st Qu.:373.0	
Median :244.0	Median :521.0	Median :446.0	Median :390.0	
Mean :246.1	Mean :511.4	Mean :443.4	Mean :389.2	
3rd Qu.:257.0	3rd Qu.:530.0	3rd Qu.:459.0	3rd Qu.:405.0	
Max. :296.0	Max. :732.0	Max. :548.0	Max. :562.0	
Pr5	**Pr6**	**Pr7**	**Pr8**	**Pr9**
Min. :157.0	Min. :337.0	Min. :110.0	Min. :323	Min. :319.0
1st Qu.:178.0	1st Qu.:366.0	1st Qu.:456.0	1st Qu.:423	1st Qu.:350.0
Median :188.0	Median :385.0	Median :481.0	Median :451	Median :369.0
Mean :188.2	Mean :394.9	Mean :465.5	Mean :450	Mean :388.9
3rd Qu.:200.0	3rd Qu.:393.0	3rd Qu.:493.0	3rd Qu.:467	3rd Qu.:382.0
Max. :221.0	Max. :557.0	Max. :642.0	Max. :639	Max. :819.0

Bei den Proben 1 und 5 ist es plausibel, dass die Konzentration unter 400 mg/kg liegt, während wir bei Probe 2, 3, 7 und 8 dazu tendieren, den Grenzwert 400 mg/kg als überschritten zu betrachten. Die übrigen Proben, Probe 4, 6 und 9 sind eher Grenzfälle. Die Konzentrationen scheinen zwar unter 400 mg/kg zu liegen, die drei Proben weisen jedoch jeweils extreme Ausreisser über dem Grenzwert auf.

(**b**) Als erstes stechen die Messungen der Labors 15 und 21 ins Auge. Beide haben sowohl eine grosse Standardabweichung als auch systematische Fehler. Die Labors 6 und 12 haben beide Ausreisser zu verzeichnen. Die Labors 1, 7, 12, 13, 14, 17, 18, 20 und 21 geben systematisch zu kleine Werte an, während die Labors 6, 8, 10 und 15 zu grosse Werte erhalten. Die Labors 2, 3, 4, 5 und 19 scheinen zuverlässige Untersuchungen durchzuführen. Sowohl systematische wie auch Zufallsfehler scheinen sich hier in Grenzen zu halten.

```
> ## 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)))
```