Serie 7

1. In dieser Aufgabe wollen wir in einer Simulation die Überdeckungswahrscheinlichkeit von Vertrauensintervallen untersuchen.

Wir betrachten hierzu eine Binomialverteilung mit $n = 50$. Wählen Sie selber eine Erfolgswahrscheinlichkeit π.

a) Simulieren Sie 20 Realisationen von obiger Binomialverteilung und bestimmen Sie für jede Realisation das 95%-Vertrauensintervall für die Erfolgswahrscheinlichkeit π. Wie oft erwarten Sie, dass der wahre Wert im Vertrauensintervall liegt? Wie oft liegt er tatsächlich drin?
R-Hinweise:

```
## 20 Werte simulieren
p <- ...
x <- rbinom(20, 50, p)

## Grenzen der Intervalle in Matrix speichern
## 1. Spalte ist untere Grenze, 2. Spalte obere
confint.bound  <- matrix(0, nrow = 20, ncol = 2)
contains.truth <- logical(20)
## Alle 20 Faelle untersuchen und Grenzen speichern
for(i in 1:20){
  test <- binom.test(...) ## Setzen Sie die richtigen Argumente!
  confint.bound[i,] <- test$conf.int
  contains.truth[i] <-
    (p >= confint.bound[i,1]) & (p <= confint.bound[i,2])
}
sum(contains.truth)
```

b) Stellen Sie das Resultat aus a) geeignet dar.
R-Hinweise:

```
## Relative Haeufigkeiten plotten
plot(x / 50, 1:20, xlim = c(0, 1), xlab = "Probability",
     ylab = "Simulation Number")
## Vertrauensintervalle als Liniensegemente plotten
for(i in 1:20){
  segments(confint.bound[i,1], i, confint.bound[i,2], i)
}
## Wahrer Wert als vertikale Linie einzeichnen
abline(v = ...)
```

2. 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
## Histogramme zeichnen
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 diesen drei Histogrammen?

Bemerkung:
Wenn man die Anzahl Klassen mit `breaks = 20` vorgibt, so wird dies nur als "Vorschlag" interpretiert und intern unter Umständen noch abgeändert.

b) Zeichnen Sie Histogramme (Anzahl Klassen variieren) von der Eruptionsdauer.
`hist(geysir[,"Eruptionsdauer"], ...)`
Was fällt auf? Vergleichen Sie mit der ersten Teilaufgabe.

c) Zeichnen Sie nun noch ein Streudiagramm von Eruptionsdauer und Zeitspanne. Was schliessen Sie daraus?

3. 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 Klärschlammproben.

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)`

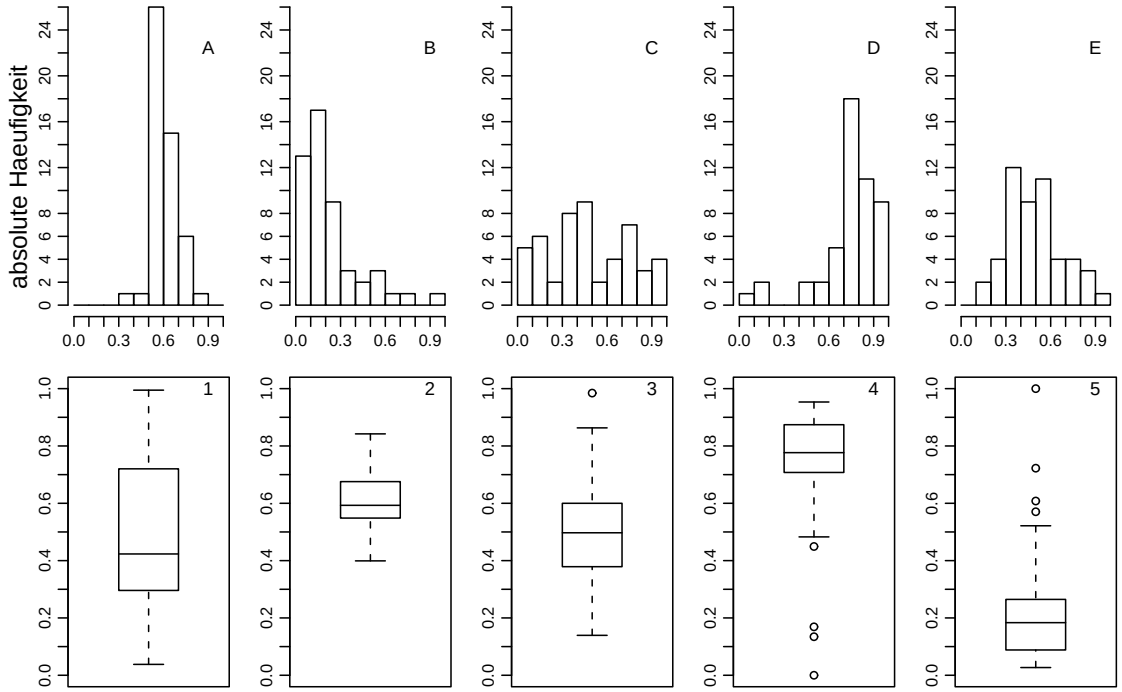
b) Erstellen Sie für jedes Labor einen Boxplot der Messfehler. Unter dem Messfehler eines Labors bei einer Probe verstehen wir den gemessenen Wert minus den Median über alle Labors. Welche der 21 Labors haben systematische Fehler in ihrem Analyseverfahren? Welche haben grosse Zufallsfehler, und bei welchen 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)))
```

4. Für fünf Stichproben vom Umfang $n = 100$ wurden je ein Histogramm und ein Boxplot gezeichnet. Ordnen Sie die fünf Boxplots den entsprechenden Histogrammen zu. Geben Sie für jede Zuordnung eine kurze Begründung!



Besprechung: Donnerstag, November 03.

Abgabe: Übung nicht abgeben - wird nicht korrigiert.