Anhang A

Übungen

A.1 R Einführung

A.1.1 Grundlagen

Starten Sie R und geben Sie die folgenden Befehle ein. Versuchen Sie zu verstehen, was passiert. Sind Sie sich unsicher, so versuchen Sie es mit eigenen Beispieln zu ergründen und fragen Sie den Übungsleiter.

```
> 2 * (3 + 4 * 5)
> sqrt(9)
> 6^2
> ls()
> x <- 3
> x^3
> ls()
> help(sqrt)
> help.start()
```

Memnontics: `ls` list, `sqrt` square root
Durch den Befehl `x <- 3` wird einer Variablen ein Wert zugewiesen.
Statt `x <- 3` kann man manchmal auch `x = 3` schreiben.

A.1.2 Lineare Algebra

A.1.2.1 Rechnen mit Vektoren

```
> x <- c(1, 2, 3)
> x
> y <- c(1/3, 1/3, 1/3)
> y
> 3 * y
> x + y
> x * y
> sum(x * y)
> sqrt(x)
> sum(x^2)
> sqrt(sum(x^2))
> length(x)
> x[2]
```

Memnontic: `c` concatenate, `sum` Summe
Welcher dieser Befehle ist

- Vektoradditon
- Skalare Multiplikation
- Skalarprodukt
- Zugriff auf die i-te Komponente des Vektors
- Elementweise Multiplikation
- Betragsquadrat
- Länge des Vektors
- Dimension des Vektors

R rechnet mit Vektoren in beliebiger Dimension. Es unterstützt auch komplexe Zahlen z.B. `3+4i`.

A.1.2.2 Rechnen mit Matrizen

```
> cbind(x, y, y, y)
> M <- rbind(x, x, y, y)
> M
> nrow(M)
> ncol(M)
> dim(M)
> matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9), nrow = 3)
> A <- matrix(c(1, 2, 3, 1, 1, 1, 1, 0, 0), ncol = 3)
> A
> B <- diag(c(1, 2, 3))
> B
> B %*% A
> solve(A, x)
> solve(A)
> solve(A) %*% A
> t(A)
> det(A)
> diag(A)
> sum(diag(A))
> eigen(A)
> sym <- t(A) %*% A
> chol(sym)
> B * A
> A[3, 1]
> A[, 1]
> A[2, ]
```

Memnontic: `cbind` column bind, `rbind` row bind, `nrow` number of rows, `ncol` number of columns, `diag`onale (zwei Bedeutungen), `solve` (engl. lösen), `t`ranspose, `det`erminante, `eigen`wertzerlegung, `%*%` inneres Produkt, `chol`elsky-Zerlegung
R legt die Einträge einer Matrix Spaltenweise fest.

A.1.3 Datensätze

In der Statistik bearbeitet man hauptsächlich Datensätze. Datensätze werden in einer Matrixartigen Struktur, der sogenannten Datenmatrix (`data.frame`) gespeichert. Es handelt sich allerdings nicht um eine Matrix im mathematischen Sinn, da auch qualitative Merkmale, wie z.B. Zeichenfolgen in den Daten enthalten sein können.

Daten werden normalerweise zunächst in einer ASCII-Datei (`.csv`-Format für Tabellencalculationen) auf der Platte gespeichert.

Legen Sie eine Datei des folgenden Inhalts auf der Platte an:

```
   MehrSchlaf Gruppe
1    0.7     Kontrolle
2   -1.6     Kontrolle
3   -0.2     Kontrolle
4   -1.2     Kontrolle
5   -0.1     Kontrolle
6    3.4     Kontrolle
7    3.7     Kontrolle
8    0.8     Kontrolle
9    0.0     Kontrolle
10   2.0     Kontrolle
11   1.9     Behandlung
12   0.8     Behandlung
13   1.1     Behandlung
14   0.1     Behandlung
15  -0.1     Behandlung
16   4.4     Behandlung
17   5.5     Behandlung
18   1.6     Behandlung
19   4.6     Behandlung
20   3.4     Behandlung
```

Die Daten entstammen einer (historisch bekannten) Studie zur Wirksamkeit eines Schlafmittels. Bei 20 Patienten wurde die Schlafdauer gemessen. In der ersten Nacht erhielt kein Patient ein Medikament. In der zweiten Nacht erhielten 10 zufällig ausgesuchte Patienten ein Placebo (also eine Tablette, die nur wie ein Schlafmittel aussieht) und 10 Patieten das zu testende Schlafmittel.

Die Spalte `Mehrschlaf` gibt die Anzahl Stunden an, welche die Patienten in der zweiten Nacht länger schliefen, als in der ersten.

Diese Datei läßt sich mit dem folgenden Befehl einlesen:

```
> X <- read.table("SchlafData.txt", sep = "", header = TRUE)
> X
```

Das Einlesen von Daten ist allgemein die erste ernsthafte Schwierigkeit. Je nach dem genauen Format der einzulesenden Daten können die `sep`peratorzeichen zur Spaltentrennung, mit `dec=","` das Dezimalkomma, und mit `header` die Anwesenheit einer ersten Zeile mit Spaltennamen festgelegt werden. Weiter Möglichkeiten findet man in der Hilfe.

Für in Deutschland aus Tabellenkalkulationen exportierte Daten eignet sich oft `read.csv2` statt `read.table`, da es die Defaultwerte für deutscher Spracheinstellung verwendet.

Im Anschluß an das Einlesen sollte überprüft werden, ob die Daten richtig gelesen wurden:

```
> sapply(X, class)
```

Damit kann man überprüfen, ob die Daten in der richtigen Weise gelesen wurden. Numerische Daten sollten vom Typ `"numeric"` oder `"integer"` und kategorielle Daten vom Typ `"factor"` sein. In unserem Fall sollte die Aus

```
MehrSchlaf     Gruppe
 "numeric"   "factor"
```

Ist der Datensatz erst einmal geladen, so kann man in verschiedener Weise darauf zugreifen:

```
> X$MehrSchlaf
> X$Gruppe
> X[1, ]
> X[, 1]
> X[1, 1]
> X[[1]]
> sum(X$MehrSchlaf)/nrow(X)
> mean(X$MehrSchlaf)
> tapply(X$MehrSchlaf, X$Gruppe, mean)
> ifelse(X$Gruppe == "Behandlung", "Schlafmuetze", "Wachling")
```

A.1.4 Programmierung

R ist eine vollwertige interpretierte interaktive vektorisierte objektorientierte Programmiersprache. Hier finden sie die wichtigesten Kontrollstrukturen:

```
> f <- function(x, a = 1, b = 2) {
+     cat("Funktion f wurde mit Parameter x=", x, " a=",
+         a, " b=", b, " aufgerufen\n")
+     a * x + b
+ }
> f(3)

Funktion f wurde mit Parameter x= 3  a= 1  b= 2  aufgerufen
[1] 5

> f(c(1, 2, 3))

Funktion f wurde mit Parameter x= 1 2 3  a= 1  b= 2  aufgerufen
[1] 3 4 5

> 1:10

 [1]  1  2  3  4  5  6  7  8  9 10

> bubblesort <- function(x) {
+     needsMoreSort <- TRUE
+     cat("Sortierung Startet\n")
+     while (needsMoreSort) {
+         cat("Neuer Durchlauf der While-Schleife beginnt\n")
+         needsMoreSort <- FALSE
+         for (i in 1:(length(x) - 1)) {
+             if (x[i] > x[i + 1]) {
+                 cat("Tausche", x[i], " und ", x[i + 1],
+                   ".")
```

```
+                 tmp <- x[i]
+                 x[i] <- x[i + 1]
+                 x[i + 1] <- tmp
+                 needsMoreSort <- TRUE
+             }
+             else {
+                 "Nix zu tun (wird auch nicht ausgegeben)"
+             }
+         }
+         if (!needsMoreSort)
+             break
+         cat("\nAktueller Zustand:", x, "\n")
+     }
+     cat("Umsortierung abgeschlossen\n")
+     x
+ }
> bubblesort(c(1, 5, 2, 6, 7, 3, 2))

Sortierung Startet
Neuer Durchlauf der While-Schleife beginnt
Tausche 5  und  2 .Tausche 7  und  3 .Tausche 7  und  2 .
Aktueller Zustand: 1 2 5 6 3 2 7
Neuer Durchlauf der While-Schleife beginnt
Tausche 6  und  3 .Tausche 6  und  2 .
Aktueller Zustand: 1 2 5 3 2 6 7
Neuer Durchlauf der While-Schleife beginnt
Tausche 5  und  3 .Tausche 5  und  2 .
Aktueller Zustand: 1 2 3 2 5 6 7
Neuer Durchlauf der While-Schleife beginnt
Tausche 3  und  2 .
Aktueller Zustand: 1 2 2 3 5 6 7
Neuer Durchlauf der While-Schleife beginnt
Umsortierung abgeschlossen
[1] 1 2 2 3 5 6 7
```

A.2 Univariate Graphik

A.2.1 Graphikbefehle

Zunächst lernen wir die Graphikbefehle am Irisdatensatz kennen:

```
> data(iris)
> iris
> help(iris)
> dim(iris)
> names(iris)
> iris$Species
> levels(iris$Species)
> plot(iris$Petal.Width, iris$Petal.Length)
> plot(iris$Petal.Width, iris$Petal.Length, pch = 20, col = "red")
> plot(iris$Petal.Width, iris$Petal.Length, pch = 20, col = c("red",
+     "green", "blue")[iris$Species])
> pairs(iris)
> hist(Sepal.Length)
```

```
> boxplot(Sepal.Length)
> stripchart(Sepal.Length)
> stripchart(Sepal.Length, method = "jitter")
> stripchart(Sepal.Length, method = "stack")
> qqnorm(Sepal.Length)
> qqline(Sepal.Length)
> boxplot(iris$Sepal.Length ~ iris$Species)
> plot(iris$Species, iris$Sepal.Length)
> table(iris$Species)
> barplot(table(iris$Species))
> plot(iris$Species)
```

A.2.2 Ausblick auf anspruchsvolle Graphikgestaltung

Es gibt eine Reihe von Befehlen, mit denen man die Graphiken deutlich aufwerten kann, z.B. um verständliche Graphiken für Vorträge und Diplomarbeiten zu erzeugen.

```
> plot(iris[, 1], iris[, 2])
> plot(iris$Petal.Width, iris$Petal.Length)
> library(MASS)
> eqscplot(iris$Petal.Width, iris$Petal.Length)
> abline(lm(Petal.Length ~ Petal.Width, data = iris))
> title(main = "Iris Datensatz", "Anderson 1932")
> axis(side = 4, col = "red")
> "Der naechste Befehl schreibt die Art an mit der Maus angeklickte Punkte"
> "Man beendet diesen Modus mit der rechten Maustaste!!!!!"
> identify(Petal.Width, Petal.Length, labels = Species)
> text(2, 3, "Das ist der beruehmte \n Iris Blueten Datensatz",
+     cex = 2)
> dev.copy2eps(file = "MeineGraphikFuerTex.eps")
> dev.copy(jpeg, file = "MeineGraphikFuerWord.jpg")
> dev.off()
```

Für eine bessere Übersichtlichkeit lohnt es sich of mehrere Graphiken in einem Fenster darzustellen.

```
> pairs(iris)
> par(mfrow = c(2, 2))
> hist(Sepal.Length)
> boxplot(Sepal.Length)
> qqnorm(Sepal.Length)
> qqline(Sepal.Length)
> stripchart(Sepal.Length, method = "jitter")
```

A.2.3 Selbständige graphische Analyse von Datensätzen

Suche Sie sich einen der folgenden Datensätze aus:

- Artenunterschiede bei Flohkäfern:
  `http://lib.stat.cmu.edu/DASL/Datafiles/FleaBeetles.html`
- Entwicklung der Form des Schädelknochens durch Migrationseinflüsse im alten Ägytpen:
  `http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html`

- Artenunterschiede bei Nordamerikanischen Eichen:
  `http://lib.stat.cmu.edu/DASL/Datafiles/Acorns.html`
- Historische Messungung der Lichtgeschwindigkeit:
  `data(morley)`
- Historische Messungen der Erddichte:
  `http://lib.stat.cmu.edu/DASL/Datafiles/Cavendish.html`

Laden Sie den Datensatz herunter und lesen Sie ihn in R ein. Verschaffen Sie sich einen Überblick über die Struktur des Datensatzes und analysieren Sie den Datensatz mit den bisher erlernten graphischen Methoden. Beschreiben Sie Ihre Beobachtungen. Versuchen Sie Fragen zu den Daten zu beantworten:

- Handelt es sich bei dem Datensatz um eine oder um mehrere Stichproben?
- Welche Grundgesamtheit gehört zu den Stichproben?
- Welche Skala haben die Daten?
- Beschreiben Sie die Verteilung der Daten (Schiefe, Modalwerte, Ähnlichkeit zur Normalverteilung)
- ...

A.3 Simulieren und Schätzen

Aufgabe 1: Poissonverteilung

In folgenden R-Code wird eine Stichprobe aus einer Poissonverteilten Grundgesamtheit realisiert und daraus der Parameter der Poissonverteilung geschätzt.

```
> SE <- function(theta, hatTheta) (theta - hatTheta)^2
> AE <- function(theta, hatTheta) abs(theta - hatTheta)
> hatLambdaML <- function(X) sum(X)/length(X)
> lambda <- rexp(1, 1/5)
> X <- function() rpois(10, lambda)
> x <- X()
> stripchart(x)
> (hatlambdaML <- hatLambdaML(x))

[1] 3.3

> simEst <- replicate(1000, hatLambdaML(X()))
> hist(simEst)
> mean(simEst)

[1] 3.7121

> sd(simEst)

[1] 0.5782542

> mean(SE(lambda, simEst))

[1] 0.334298

> mean(AE(lambda, simEst))
```

```
[1] 0.4580079

> mean(simEst) - lambda

[1] 0.01594936

> mean(log(simEst)) - log(lambda)

[1] -0.008048761
```

Aufgabe 2: *Schätzer berechnen*
Für diese Übung haben Sie heute, die ganze Woche Zeit. Die Ergebnisse stellen sie sich in der nächsten Übung gegenseitig vor. Suchen Sie sich eines der folgenden Verteilungsmodelle aus:

- $X_i \sim N(\theta, 3), \theta \in \mathbb{R} = \Theta$, wähle $p = 1$
- $X_i \sim N(0, \theta^2), \theta \in \mathbb{R}^+ = \Theta$, wähle $p = 2$
- $X_i \sim Exp(\theta), \theta \in \mathbb{R}^+ = \Theta$, wähle $p = 1$
- $X_i \sim Unif((0, \theta)), \theta \in \mathbb{R}^+ = \Theta$, wähle $p = 1$
- $X_i \sim Bi(p)$, $\theta \in [0, 1] = \Theta$, wähle $p = 1$
- $X_i \sim Bi(\frac{e^\theta}{1+e^\theta})$, $\theta \in \mathbb{R} = \Theta$, wähle $p = 1$
- $X_i \sim Geo\left(\frac{e^\theta}{1+e^\theta}\right)$ $\theta \in \mathbb{R} = \Theta$, wähle $p = 1$
- $X_i \sim Cauchy(0, \theta)$, $f(x) = \frac{1}{\pi} \cdot \frac{\theta}{\theta^2+(x-t_0)^2}$ mit $\theta \in \mathbb{R}^+ = \Theta$, und $t_0 = 0$ fest. (Die Aufgabe funktioniert für kein p, Wieso?)

(Der Schwierigkeitsgrad ist aufsteigend sortiert!).

(1) Simulieren sie einen Wert für den Parameter der Verteilung mit einer Gleichverteilung auf $(0.1, 10)$.

(2) Simulieren Sie einen Datensatz der Größe $n = 40$ Ihrer Verteilung und untersuchen den Datensatz mit statistischen Graphiken. Speichern Sie den Datensatz für die nächste Übung.

(3) Zeigen Sie, dass der Mittelwert der X_i^p ein Erwartungstreuer Schätzer für $E[X^p]$ ist.

(4) Schätzen sie die Varianz als $\hat{\sigma^2}(X) :=$`mean(X^2)`.

(5) Nennen wir den Parameter θ. Berechnen Sie die Funktion $s(\theta) := E_{P_\theta}[X_i^2] = E_{P_\theta}[\hat{\sigma^2}]$ und bestimmen Sie einen Momentenschätzer $\hat{\theta}_M$ für θ indem Sie $s(\hat{\theta}_M) = \hat{\sigma^2}(X)$ nach θ auflösen.

(6) Simulieren Sie jeweils 1000 mal die Schätzung mit $\hat{\theta}_M$) zu mehreren verschiedenen θ-Werten. Wie ist die Verteilung der Schätzwerte?

(7) Bestimmen Sie nun den Maximum-Likelihood-Schätzer von θ indem sie

$$\hat{\theta}_{ML}(X) = \underset{\theta \in \Theta}{\operatorname{argmax}} \ln \prod_{i=1}^{n} f_{P_\theta}(X_i)$$

bestimmen. Prüfen Sie anhand der Ableitung, ob es tatsächlich ein Maximum ist. In welchen Fällen existiert kein Maximum?

(8) Implementieren Sie nun auch $\hat{\theta}_{ML}$ in R und führen Sie die gleichen Simulationen durch und beurteilen Sie die Verteilung des Schätzers.

(9) Welcher der Schätzer ist besser? Wählen Sie diesen Schätzer aus.

(10) Ist einer der beiden Schätzer erwartungstreu (d.h. ist sein Erwartungswert der wahre Wert von θ)?

(11) Schätzen Sie mithilfe ihres ausgewählten Schätzers den Parameter des ursprünglichen Datensatzes. Vergleichen Sie den Schätzwert mit dem wahren Wert. Hat das Verfahren gut funktioniert?

(12) Fügen Sie Ihrem Datensatz einen Ausreißer mit dem Wert 999 (einem häufig in Datensätzen anzutreffendem Fehlerwert) hinzu. Wie ändert sich dadurch die Schätzung des Parameters?

A.4 Konfidenzschätzung und t-Verteilung

Aufgabe 1: Ergebnisvorstellung
Stellen sie sich gegenseitig die Ergebnisse der letzten Übung vor:

- Wie erkennt man die von Ihnen gewählte Verteilung in statistischen Graphiken?
- Wie lautet der Momentenschätzer?
- Wie lautet der Maximum-Likelihood Schätzer?
- Welche mathematischen "Merkwürdigkeiten" sind Ihnen beim Ausrechnen begegnet?
- Wie sind die Schätzer verteilt?
- Kann man den Parameter mit 40 Beobachtungen verläßlich schätzen?

Aufgabe 2: Konfidenzintervall

(1) Implementieren Sie in R eine Routine `t.confidence(X)`, welche das t-Verteilungsbasierte Konfidenzintervall

$$[\bar{x} + \sqrt{\frac{1}{n}\hat{\sigma}^2}t_{n-1,\alpha/2}, \bar{x} + \sqrt{\frac{1}{n}\hat{\sigma}^2}t_{n-1,1-\alpha/2}]$$

(mit

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

und $t_{d,p}$ dem p-Quantil der t-Verteilung mit d Freiheitsgraden) für den Erwartungswert $E[X]$ berechnet und testen sie in einer Simulationstudie, ob mit verschiedenen n und verschiedenen α, ob tatsächlich der wahre Erwartungswert mit einer Wahrscheinlichkeit von $1 - \alpha$ in diesem zufälligen Intervall enthalten ist.

(2) Prüfen Sie, ob das auch mit exponentialverteilten Datensätzen funktioniert. (Tip: $E[X] = \frac{1}{\lambda}$)

A.5 Tests I

Aufgabe 1: Das Junggesellenkochbuch der Tests

```
data(iris)
plot(iris,col=c("red","green","blue")[iris$Species])
# Offenbar unterscheidet sich das Petal sehr deutlich
# Statistische Tests sind immer folgendermassen aufgebaut:
#
```

```
# Name: Irgend ein Name
# Voraussetzung: z.B. Unabhaengigkeit der Beobachtungen
# Hypothese: Irgend eine Vorstellung
#  gegen
# Alternative: Das Gegenteil dieser Vorstellung
# Befehl: irgend ein R-Befehl um den Test durchzufuehren
# Der Computer berechnet jeweils einen p-Wert. Liegt der p-Wert unter 0.05,
# und sind die Voraussetzungen erfuellt, so gilt die Hypothese als
# widerlegt und die Alternative als vermutlich wahr. 0.05 ist dann eine
# Grenze fuer die Irrtumswahrscheinlichkeit.

# Unser erster Test:
# Name: Kruskal Wallis Rang Summen Test
# Voraussetzungen: Unabhaengige Stichproben, stetige Verteilung
# Hypothese: Gruppen haben im Mittel gleich grosse Werte
#  gegen
# Alternative: Verteilungen der Gruppen sind gegegeneinander versetzt
kruskal.test(Sepal.Width~Species,data=iris)

# Aehnliche Dinge testen die folgenden Tests:
# Name: Varianzanalyse
# Voraussetzungen: Unabhaengig, Normalverteilt, gleiche Streuung
# Hypothese: Gruppen haben gleichen Mittelwert
#  gegen
# Alternative: Gruppen haben unterschiedlichen Mittelwert
# Besonderheit: der p-Wert heisst Pr(>F)
anova(aov(Sepal.Width~Species,data=iris))

# Name: Zwei Stichproben t-test
# Voraussetzungen: Zwei Gruppen, unabhaengig, normalverteilt, gleiche Streuung
# Hypothese: Gruppen haben gleichen Mittelwert
#  gegen
# Alternative: Gruppen haben unterschiedlichen Mittelwert
Setosa <- split(iris,iris$Species)[["setosa"]]
Virginica <- split(iris,iris$Species)[["virginica"]]
t.test(Setosa$Sepal.Width,Virginica$Sepal.Width)

# Name: Wilcoxen Rang Summen Test
# Voraussetzungen: Zwei Gruppen, unabhaengig, stetig verteilt
# Hypothese: Gruppen haben gleichen Median
#  gegen
# Alternative: Gruppen haben unterschiedlichen Median
wilcox.test(Setosa$Sepal.Width,Virginica$Sepal.Width)

# Test fuer die Streuung:
# Name: Fligner Killean Median Test auf gleiche Varianz
# Voraussetzungen: Unabhaengige Stichproben, stetige Verteilung
# Hypothese: Gruppen haben gleiche Streuung
#  gegen
# Alternative: Gruppen haben verschiedene Streuung
fligner.test(Sepal.Width~Species,data=iris)

# Test auf Normalverteilung
# Name: Shapiro-Wilk-Test
```

```
# Voraussetzungen:  Unabhaengige Stichprobe
# Hypothese: Daten sind normalverteilt
#  gegen
# Alternative: Daten sind nicht normalverteilt
shapiro.test(iris$Petal.Length)   # Mischung aus drei Normalverteilungen
shapiro.test(Setosa$Petal.Length)
shapiro.test(Virginica$Petal.Length)

# Test auf Unabh"angigkeit in kategorieller Merkmale
# Name: Chi-Quadrat-Test f"ur Kontingenztafeln
# Voraussetzungen:  Unabhaengige Stichprobe, kategorielle Daten
# Hypothese: Merkmale sind stochastisch unabh"angig
#  gegen
# Alternative: Merkmale sind stochastisch abh"angig
data(HairEyeColor)
ftable(HairEyeColor)
X <- apply(HairEyeColor,c(1,2),sum)
X
chisq.test(X)

# F"uhrt man im Rahmen der gleichen Studie mehrere Test durch, bei denen eine
# einzelne Signifikanz bereits zum ``Nachweis'' eines Effektes f"uhrt, so muss
# man die p-Werte korrigieren, da man ja bei jeder Testdurchf"uhrung erneut
# die Chance auf ein f"alschlich signifikantes Ergebnis hat.
Setosa <- split(iris,iris$Species)[["setosa"]]
Virginica <- split(iris,iris$Species)[["virginica"]]
Versicolor <- split(iris,iris$Species)[["versicolor"]]
t.test(Setosa$Sepal.Width,Virginica$Sepal.Width)
t.test(Setosa$Sepal.Width,Versicolor$Sepal.Width)
t.test(Versicolor$Sepal.Width,Virginica$Sepal.Width)
p.adjust(c(4.571e-09,2.484e-15,0.001819),method="bonferroni")
# Die ausgebenen adjustierten p-Werte sind immer noch <0.05 und damit
# koennen alle drei Test als signifikant angesehen werden, obwohl
# mehrere Tests durchgefuehrt wurden


# Bearbeiten Sie selbst:
# a) Koennte die Kelchblatt-Laenge bei den einzelne Irisarten normalverteilt sein?
# b) Haben die Arten virginica und versicolor unterschiedliche mittlere
#      Kelchblattlaengen
```

Eine detailierte Auflistung der Standardtests (aus dem Kochbuch für Hausfrauen) finden Sie im Anhang B des Skripts.

Aufgabe 2: *Gütefunktion ausprobieren*

Mit dem folgenden R-Code werden wird der 2-Stichproben-t-Test auf jeweils 1000 Simulationen von zwei normalverteilten Stichproben mit jeweils 10 Beobachtungen simuliert. In den drei Fällen ist einmal die Hypothese und zweimal die Alternative zutreffend.

```
> p.Werte <- replicate(1000, t.test(rnorm(10, mean = 0),
+     rnorm(10, mean = 0))$p.value)
> hist(p.Werte)
> mean(p.Werte < 0.05)

[1] 0.05

> p.Werte <- replicate(1000, t.test(rnorm(10, mean = 0),
+     rnorm(10, mean = 2))$p.value)
> hist(p.Werte)
> mean(p.Werte < 0.05)

[1] 0.99

> p.Werte <- replicate(1000, t.test(rnorm(10, mean = 0),
+     rnorm(10, mean = 0.2))$p.value)
> hist(p.Werte)
> mean(p.Werte < 0.05)

[1] 0.071
```

(1) In welchem der drei Fälle sind die p-Werte gleichverteilt?

(2) Warum ist der Anteil der p-Werte kleiner als 0.05 im

- ersten Fall ungefähr 0.05?
- im zweiten Fall knapp unter 1?
- und im dritten Fall knapp über 0.05?

(3) Was ist die Wahrscheinlichkeit für einen β-Fehler in den drei simulierten Situationen? Der β-Fehler beschreibt, dass die Hypothese angenommen wurde, obwohl sie falsch ist.

Aufgabe 3: Legt der Kuckuck sein Ei passend zum Nest?
Laden Sie den Datensatz `Pipits.txt`. Der Kuckuck legt jeweils ein Ei in die Nester anderer Vögel, die selbst teiweise sehr verschieden große Eier legen. Die Frage ist nun, ob die vom Kuckuck gelegte Eigröße irgendwie abhängig vom Wirtsvogel (also dem Vogel in dessen Nest das Ei gelegt wird) variiert. Untersuchen Sie den Datensatz mit den erlernten Mitteln und versuchen Sie diese Frage zu beantworten.

A.6 Tests II

Aufgabe 1: Tests–Gemeinsamer Teil
Wir laden den Calcium Datensatz von statlib:

> *Datafile Name:* Calcium
> *Datafile Subjects:* Health , Medical
> *Story Names:* Calcium and Blood Pressure
> *Reference:* Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics. Original source: Lyle, Roseann M., et al., "Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men,"JAMA, 257(1987), pp. 1772-1776
> *Authorization:* contact authors
> *Description:* Results of a randomized comparative experiment to investigate the effect of calcium on blood pressure in African-American men. A treatment group of 10 men received a calcium supplement for 12 weeks, and a control group of 11 men received a placebo during the same period.

All subjects had their blood pressure tested before and after the 12-week period.
Number of cases: 21
Variable Names:

1.Treatment: Whether subject received calcium or placebo
2.Begin: seated systolic blood pressure before treatment
3.End: seated systolic blood pressure after treatement
4.Decrease: Decrease in blood pressure (Begin - End)

```
> CaBp <- read.table("CalciumData.txt", header = TRUE)
> CaBp

   Treatment Begin End Decrease
1    Calcium   107 100        7
2    Calcium   110 114       -4
3    Calcium   123 105       18
4    Calcium   129 112       17
5    Calcium   112 115       -3
6    Calcium   111 116       -5
7    Calcium   107 106        1
8    Calcium   112 102       10
9    Calcium   136 125       11
10   Calcium   102 104       -2
11   Placebo   123 124       -1
12   Placebo   109  97       12
13   Placebo   112 113       -1
14   Placebo   102 105       -3
15   Placebo    98  95        3
16   Placebo   114 119       -5
17   Placebo   119 114        5
18   Placebo   112 114        2
19   Placebo   110 121      -11
20   Placebo   117 118       -1
21   Placebo   130 133       -3

> plot(Begin ~ Treatment, data = CaBp)
> plot(End ~ Treatment, data = CaBp)
> plot(Decrease ~ Treatment, data = CaBp)
> groups <- split(CaBp, CaBp$Treatment)
> groups$Calcium

   Treatment Begin End Decrease
1    Calcium   107 100        7
2    Calcium   110 114       -4
3    Calcium   123 105       18
4    Calcium   129 112       17
5    Calcium   112 115       -3
6    Calcium   111 116       -5
7    Calcium   107 106        1
8    Calcium   112 102       10
9    Calcium   136 125       11
10   Calcium   102 104       -2

> groups$Placebo
```

```
   Treatment Begin End Decrease
11   Placebo   123 124       -1
12   Placebo   109  97       12
13   Placebo   112 113       -1
14   Placebo   102 105       -3
15   Placebo    98  95        3
16   Placebo   114 119       -5
17   Placebo   119 114        5
18   Placebo   112 114        2
19   Placebo   110 121      -11
20   Placebo   117 118       -1
21   Placebo   130 133       -3
```

(1) Beschreiben sie das statistische Modell für die Daten.

(2) Überprüfen Sie die Normalverteilung der 3 stetigen Variablen in den 2 Gruppen.

```
> qqnorm(groups$Calcium$Begin)
> shapiro.test(groups$Calcium$Begin)

        Shapiro-Wilk normality test

data:  groups$Calcium$Begin
W = 0.8778, p-value = 0.1231
```

Sind die Variablen normalverteilt?

(3) Überprüfen Sie, ob die drei Variablen in beiden Gruppen die gleiche Varianz haben. Welchen der beiden Tests würden Sie bei den gegeben Voraussetzungen verwenden?

```
> var.test(groups$Calcium$Begin, groups$Placebo$Begin)

        F test to compare two variances

data:  groups$Calcium$Begin and groups$Placebo$Begin
F = 1.4423, num df = 9, denom df = 10, p-value = 0.5749
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3816782 5.7172721
sample estimates:
ratio of variances
          1.442348

> fligner.test(list(groups$Calcium$Begin, groups$Placebo$Begin))

        Fligner-Killeen test of homogeneity of variances

data:  list(groups$Calcium$Begin, groups$Placebo$Begin)
Fligner-Killeen:med chi-squared = 0.0272, df = 1, p-value =
0.8689
```

Haben die Variablen haben in beiden Stichproben die gleiche Streuung?

(4) *Unterschiede zwischen Gruppen feststellen*

Überprüfen Sie, ob sich die verschiedenen Merkmale in den Gruppen unterscheiden. Welcher der folgenden Tests ist unter den gegebenen Voraussetzungen der geeignetste?

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin, var.equal = TRUE)
```

```
        Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3753, df = 19, p-value = 0.7116
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.447946 10.702492
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin, var.equal = FALSE)
```

```
        Welch Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3719, df = 17.621, p-value = 0.7144
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.579519 10.834064
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin)
```

```
        Welch Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3719, df = 17.621, p-value = 0.7144
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.579519 10.834064
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

```
> wilcox.test(groups$Calcium$Begin, groups$Placebo$Begin)
```

```
        Wilcoxon rank sum test with continuity correction

data:  groups$Calcium$Begin and groups$Placebo$Begin
W = 53.5, p-value = 0.9436
alternative hypothesis: true location shift is not equal to 0
```

Konnte ein Unterschied statistisch nachgewiesen werden?

(5) *Bessere Ausnutzung der Daten*
Gemeinsam: Jemand vermutet eine Manipulation und möchte beweisen, dass Leute mit geringerem Blutdruck für die Calciumgruppe ausgewählt wurden. Welcher der folgenden Tests ist dafür am geeignesten?

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin, alternative = "greater")
```

```
        Welch Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3719, df = 17.621, p-value = 0.3572
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
```

```
 -5.968982       Inf
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin, alternative = "less")

        Welch Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3719, df = 17.621, p-value = 0.6428
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 9.223527
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

```
> t.test(groups$Calcium$Begin, groups$Placebo$Begin, alternative = "two.sided")

        Welch Two Sample t-test

data:  groups$Calcium$Begin and groups$Placebo$Begin
t = 0.3719, df = 17.621, p-value = 0.7144
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.579519 10.834064
sample estimates:
mean of x mean of y
 114.9000  113.2727
```

Konnte eine Manipulation nachgewiesen werden?

Einzeln: Die Pharmaindustrie will ihre Daten natürlich optimal ausnutzen. Man ist von vornherein davon überzeugt, dass Calcium, wenn es überhaupt eine Wirkung hat, den Blutdruck senken würde und ist daher ausschließlich daran interessiert das zu beweisen. Welcher Test wäre hierfür geeignet?

Konnte ein Unterschied nachgewiesen werden?

(6) *Qualitätskontrolle*
Können Sie nachweisen, dass die Behandlungsgruppe und Kontrollgruppe vor der Behandlung in Bezug auf Ihre Blutdruckverteilung gleich waren?

(7) *Ergebnis*
Senkt Calciumgabe den Blutdruck?
Welchen Test haben wir für die Antwort auf diese Frage effektiv verwendet?

(8) *Test mit weniger Information*
Angenommen es würde keine Kontrollgruppe geben. Dann würde man versuchen die Wirksamkeit durch einen Vergleich "Vorher" – "Nachher" zu bestätigen.

```
> boxplot(groups$Calcium$Begin - groups$Calcium$End)
> qqnorm(groups$Calcium$Begin - groups$Calcium$End)
> shapiro.test(groups$Calcium$Begin - groups$Calcium$End)

        Shapiro-Wilk normality test

data:  groups$Calcium$Begin - groups$Calcium$End
W = 0.8953, p-value = 0.1944
```

```
> t.test(groups$Calcium$Begin, groups$Calcium$End, alternative = "less",
+     paired = TRUE)
```

```
        Paired t-test

data:  groups$Calcium$Begin and groups$Calcium$End
t = 1.8084, df = 9, p-value = 0.948
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 10.06830
sample estimates:
mean of the differences
                      5
```

```
> wilcox.test(groups$Calcium$Begin, groups$Calcium$End,
+     alternative = "less", paired = TRUE)
        Wilcoxon signed rank test

data:  groups$Calcium$Begin and groups$Calcium$End
V = 41, p-value = 0.92
alternative hypothesis: true location shift is less than 0
```

- Welcher der Tests würde für die Überprüfung der Voraussetzung verwendet werden?
- Welcher der Tests würde für die Überprüfung der Wirksamkeit verwendet werden?
- Warum sollte man trotzdem eine Kontrollgruppe verwenden?

(9) *Test mit noch weniger Information*
Im Prinzip könnten auch nur die Änderungen im Blutdruck erhoben worden sein. In diesem Fall benötigen wir einen Ein-Stichproben-t-Test, der überprüft, ob eine Änderung nachgewiesen werden kann.

```
> t.test(groups$Calcium$Decrease, mu = 0, alternative = "greater")
        One Sample t-test

data:  groups$Calcium$Decrease
t = 1.8084, df = 9, p-value = 0.052
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -0.06829851         Inf
sample estimates:
mean of x
        5
```

- Warum unterscheidet sich das Ergebnis vom vorherigen Test nicht?
- Warum sollte man doch besser "Vorher" – "Nachher" Werte erheben, statt nur der Differenzen? (Wie ist es z.B. wenn Ca-Gabe nur bei Hypertonikern wirkt?)

Aufgabe 2: Ändert sich der Gleichgewichtssinn?
Laden Sie `http://lib.stat.cmu.edu/DASL/Datafiles/Balance.html` herunter und stellen Sie fest, ob es älteren Personen schwerer fällt das Gleichgewicht zu halten, wenn sie sich konzentrieren. (Tip: Unterstriche durch anderes Zeichen ersetzten)

(1) Welche Tests führen Sie durch und welche Graphiken sehen Sie an, um Voraussetzungen für weitere Tests zu überprüfen.

(2) Wodurch unterscheidet sich dieses Situation von der Vorherigen?

(3) Es muß eine p-Wert Korrektur durchgeführt werden, da sowohl bei Seitenschwankungen als auch bei Vorwärts-Rückwärtsschwankungen eine entsprechende Änderung als nachgewiesen angesehen werden würde.

(4) Was sind die Ergebnisse der Untersuchung?

A.7 Tests und Graphiken selbständig anwenden

Aufgabe 1: Steregramme
Auf dem Statlib Server findet man einen Datensatz (3Dview.txt) von Cleveland, W. S. (1993). Visualizing Data. Original source: Frisby, J. P. and Clatworthy, J.L., Learning to see complex random-dot steregrams,Perception, 4, (1975), pp. 173-178, der ein Experiment beinhaltet, in dem untersucht werden sollte, ob es beim Stereosehen hilft, wenn man eine Vorstellung hat, was man sieht. Dazu wurden zwei Gruppen (NV und VV) von Probanden künstliche Stereobilder gezeigt. Es wurde jeweils die Zeit gemessen, welche die Probanden benötigten um mit einem Stereoskop einen 3-Dimensionalen Eindruck zu erhalten. Dazu bekam die Gruppe VV vorher ein Bild gezeigt auf dem das 3D Objekt dargestellt war. Beiden Gruppen wurde gesagt, dass eine Diamantform zu sehen ist.

Abbildung A.1: Ein Steregram

(1) Werten Sie die Daten so gut wie möglich aus. Versuchen Sie die These der Wissenschaftler, dass graphische Vorinformationen beim 3D-sehen hilft an mit diesem Datensatz zu beweisen.

Aufgabe 2: Suchen Sie sich einen der folgenden Datensätze aus und versuchen daran die gestellte Frage zu beantworten.

(1) **Fütterungsstudie bei Hühnern** `? checkwts`: Was sollte man füttern? Welche Unterschiede kann man nachweisen, welche nicht?

(2) **Wirksamkeitsstudie eines Medikaments** `? sleep`: Wirkt das Schlafmittel?

(3) **Der Untergang der Titanic** `? Titanic`: Frauen und Kinder zuerst? Hat man die Leute im Unterdeck tatsächlich nicht hinausgelassen?

(4) **Der Old-Faithful Geiser** `? faithful` und/oder `require(MASS);? geyser` Wie funktioniert dieser Geiser? Sind die Erruptionen unabhängig? Sind die Wartezeiten und Erroptionslängen unabhängig?

A.8 Lineare Modelle

Aufgabe 1: Pearson und Spearman Korrelation: Transmissivität eines Aquifersystems
Im Zeisigwald in der Nähe von Chemnitz wurden an Bohrlöchern in unterschiedlicher Teufe die Transmissivität des angebohrten Grundwasserleiters bestimmt. Für die Simulation des Grundwasserflusses soll nun festgestellt werden, ob man im Zeisigwald

die Transmissivität des Grundwasserleiters aus Teufe und/oder Grundwasserleitertyp (Porenleitung, Kluftleitung) berechnen kann. Als Daten stehen die gemessene Transmissivität, die logarithmierte gemessene Transmissivität (logT), die Teufe und der vermutete Grundwasserleitertyp (Porenleitung, Kluftleitung).

```
Aqui <- read.table("Aquifers.txt")
Aqui$Transmissivitaet <- exp(Aqui$logT)
Aqui
attach(Aqui)
cor.test(Teufe,Transmissivitaet,method="pearson")
cor.test(Teufe,Transmissivitaet,method="spearman")
```

Welche Art von Zusammenhang besteht zwischen Teufe und Transmissivität?

Aufgabe 2: Lineare Modelle
In dieser Aufgabe sollen sie die R-Befehle für Lineare Modell kennenlernen und die Interpretation linearer Modelle erlernen.

```
R2  <- function(X){var(X$fitted.values)/var(X$fitted.values+X$residuals)}
Hebelwirkung <- function(Modell) {lm.influence(Modell)$hat}
Modell <- aov(logT~Teufe,data=Aqui)

Modell
summary(Modell)
anova(Modell)
coef(Modell)
plot(Modell$fitted.values,resid(Modell))
R2(Modell)
boxplot(Hebelwirkung(Modell))
qqnorm(resid(Modell))
```

Andere Modelle:

```
Modell <- aov(logT~Type,data=Aqui)
Modell <- aov(logT~Type+Teufe,data=Aqui)
Modell <- aov(logT~Teufe+Type,data=Aqui)
Modell <- aov(logT~Type*Teufe,data=Aqui)
```

(1) Welches der aufgeführten Modell hat die größte Erklärungskraft? .....

(2) Könnte man die Transmissivität allein aus der Tiefeninformation (ohne den Grundwasserleitertyp) gut vorhersagen? ..... Wieso?

---

(3) Können Sie nachweisen, daß sich die Transmissivität mit der Tiefe ändert? ..... Woran erkennt man das?

---

(4) Welches ist das beste dieser Modelle? ..... Wieso?

---

(5) Gibt es Ausreißer oder Besonderheiten in dem Datensatz? Beschreiben Sie kurz.

Aufgabe 3: Flohkäfer
An dieser Aufgabe soll die Interpretation linearer Modell selbsttätig geübt werden. Informieren Sie sich in der zugehörigen Datei über die Daten. (a) Haben verschiedene Arten von Flohkäfern ein verschieden breites aedeagus? (b) Ist der Winkel an der vorderen Spitze des aedeagus bei verschiedenen Arten unterschiedlich? (c) Sind die erkannten Unterschiede zwischen allen Arten ausgeprägt? (d) Gibt es Bedenken gegen die Anwendung der Varianzanalyse zum beantworten dieser Fragestellungen? Wie schwerwiegend sind diese Bedenken? (e) Denken sie, daß es möglich wäre nur von der Breite des aedeagus schon auf die Art des Flohkäfers zu schließen? (f) Wir wollen die Breite des aedeagus durch den Winkel an der Spitze und die Art des Flohkäfers erklären. Vergleichen sie folgende Modelle:

1. Width= a1 * Angle + b + e
2. Width= a1(Species) +b +e
3. Width= a1*Angle + a2(Species) + b + e
4. Width= a1(Species)+ a2*Angle + b + e
5. Width= a1*Angle + a2(Species) + a12 (Species) *Angle + b + e

Welches Modell hat die größte Erklärungskraft? Welche der Modelle hat nur signifikante Parameter? Welches der dieser Modelle hat die größte Erklärungskraft? Was ist der Unterschied zwischen Modell 3 und 4? Welches der Modelle ist das „richtige“? Warum?

Die hier aufgeführten Befehle sollen als Anregung dienen.

```
fb <- read.table("FleaBeetles.html",header=T,sep="\t",skip=26)
fb$Species
R2 <- function(X){var(X$fitted.values)/var(X$fitted.values+X$residuals)}
qqnorm(split(fb$Angle,fb$Species)$Con)
shapiro.test(split(fb$Angle,fb$Species)$Con)
hist(split(fb$Angle,fb$Species)$Con)

qqnorm(split(fb$Angle,fb$Species)$Hei)
shapiro.test(split(fb$Angle,fb$Species)$Hei)
hist(split(fb$Angle,fb$Species)$Hei)

qqnorm(split(fb$Angle,fb$Species)$Hep)
shapiro.test(split(fb$Angle,fb$Species)$Hep)
hist(split(fb$Angle,fb$Species)$Hep)
```

```
summary(aov(Angle~Species, data=fb))
boxplot(aov(Angle~Species, data=fb)$residuals ~ fb$Species )

summary(aov(Angle~Species, data=fb[fb$Species!="Hep",]))
R2(aov(Angle~Width+Species, data=fb))
boxplot(Angle~Species, data=fb)

summary(aov(Angle~Width, data=fb))

boxplot( aov(Angle~Width, data=fb)$residuals ~ fb$Species )
summary(aov(Angle~Width+Species, data=fb))
R2(aov(Angle~Width+Species, data=fb))
boxplot( aov(Angle~Width+Species, data=fb)$residuals ~ fb$Species )

summary(aov(Angle~Species+Width, data=fb))
R2(aov(Angle~Species+Width, data=fb))
plot( aov(Angle~Species, data=fb)$residuals ~ fb$Width )
```