Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

MODULENDPRÜFUNG

STATISTICS FOR DATA SCIENCE HS18

DAUER: 90 MINUTEN

Name, Vorname: _____

Stick-Nummer: _____

Unterschrift: .....

Aufgabe	**max. Punktezahl**	**erreichte Punktzahl**
Aufgabe 1	22	
Aufgabe 2	17	
Aufgabe 3	24	
Aufgabe 4	27	
Total	90	

Viel Erfolg!
M. Birbaumer

HINWEISE

Erlaubte Hilfsmittel

- `Python`-Referenzkarte, resp. `R`-Referenzkarte
- eine eigenhändig handgeschriebene Zusammenfassung im Umfang von 12 einseitig beschriebenen A4-Seiten
- Laptop mit `Python`, resp. mit Statistiksoftware `R`
- Taschenrechner
- (leeres) Papier und Schreibzeug

Daten

- Die Daten zu den Aufgabe 1, 3 und 4 befinden sich im Verzeichnis `Austausch` des USB-Sticks.

Prüfung

- Nennen Sie die Python-Datei `name_vorname.py`, resp. die R-Datei `name_vorname.R` in Ihren Namen um, und speichern Sie diese im Verzeichnis `Austausch` mit den von Ihnen benützten `Python`, resp. `R`-Befehlen.
- Schreiben Sie alle Lösungen zu den Aufgaben vollständig auf Papier.
- Die Prüfung ist in sauberer Schrift und übersichtlicher Darstellung zu schreiben. Die Lösungswege müssen immer klar erkennbar und kontrollierbar sein, beispielsweise unterstützt mit Formeln, Stichworten oder Skizzen.
- Die Bestnote kann bereits ab 80 Punkten erreicht werden.

Aufgabe 1: Elektrische Widerstände ..... 22 Punkte

Zwei verschiedene Messmethoden für elektrische Widerstände sollen miteinander verglichen werden. Aus diesem Grund wurden an 30 Widerständen Parallelmessungen vorgenommen, die in der Datei **widerstaende.txt** im Verzeichnis **Austausch** enthalten sind. In der ersten Spalte befinden sich die Messwerte nach der Methode A, in der zweiten Spalte befinden sich die Messwerte nach der Methode B - die Einheit ist Ohm. Wir möchten überprüfen, ob die beiden Methoden gleichwertig sind.

a) (2 Punkte) Handelt es sich um einen gepaarten oder einen ungepaarten Test?

b) (3 Punkte) Geben Sie die Null- und die Alternativhypothese an.

c) (4 Punkte) Geben Sie eine Schätzung für den Mittelwert μ und die Varianz σ^2 der Differenzen an.

d) (9 Punkte) Führen Sie den geeigneten t-Test durch: Geben Sie die Teststatistik und deren Verteilung unter der Nullhypothese an, bestimmen Sie den (realisierten) Wert der Teststatistik T, den Verwerfungsbereich für T und den Testentscheid. Überprüfen Sie, ob ein t-Test angebracht ist. Welche weiteren statistischen Tests kämen in Frage, falls die Bedingungen für die Anwendung eines t−Tests nicht erfüllt sind? (Wenn Sie obige Aufgabe nicht lösen konnten, benutzen Sie im folgenden als Ersatzwert $\sigma^2 = 2.5$.)

e) (4 Punkte) Bestimmen Sie ein zweiseitiges 95%-Vertrauensintervall für μ_D, erklären Sie die Bedeutung dieses Intervalls, und fällen Sie den Testentscheid aus Teilaufgabe (d) mit Hilfe des von Ihnen ermittelten Vertrauensintervalls.

Aufgabe 2: Stochastischer Prozess ..... 17 Punkte

Betrachten Sie den stochastischen Prozess, der gegeben ist durch

$$X(t) = A \cdot \cos(\pi t)\,,$$

wobei A eine Exponential-verteilte Zufallsvariable ist, also $A \sim \text{Exp}(\lambda)$.

(a) (4 Punkte) Zum Zeitpunkt $t_0 = \frac{1}{3}$ wurde über sehr viele Realisierungen von $X(t)$ der empirische Scharmittelwert $\bar{x}(t = t_0) = 0.1$ ermittelt. Schätzen Sie λ.

(b) (4 Punkte) Wie gross ist die Wahrscheinlichkeit $P[X(t = t_0) > 0.5]$? Benützen Sie dazu die Schätzung $\widehat{\lambda}$ aus Teilaufgabe (a) (Wenn Sie obige Aufgabe nicht gelöst haben, rechnen Sie mit $\widehat{\lambda} = 1$).

(c) (4 Punkte) Berechnen Sie den Erwartungswert $\mu_X(t) = \text{E}[X(t)]$ zu einem beliebigen Zeitpunkt t.

(d) (4 Punkte) Wie gross ist die Varianz $\text{Var}[X(t)]$ zu einem beliebigen Zeitpunkt t?

(e) (1 Punkt) Handelt es sich bei $X(t)$ um einen stationären stochastischen Prozess?

Aufgabe 3: Versuchsplanung und Varianzanalyse ..... 24 Punkte

(b) (6 Punkte) Der Datensatz **stream** enthält die Zinkstufen (Variable **ZINC**) verschiedener Flüsse (Variable **STREAM**) und die entsprechende Biodiversität (Variable **DIVERSITY**). Zusätzlich kodiert die Variable **ZNGROUP** die verschiedenen Zink-Gruppen numerisch. Wir wollen untersuchen, ob eine signifikante Beziehung zwischen der Biodiversität und den Zink-Gruppen besteht.

Lesen Sie den in der Datei **stream.dat** gespeicherten Datensatz ein. Erstellen Sie einen Boxplot und Stripchart von **DIVERSITY** versus **ZNGROUP**, und kommentieren Sie die Graphik in Bezug auf Unterschiede und Ausreisser.

(c) (7 Punkte) Sie möchten feststellen, ob es einen signifikanten Unterschied der **DIVERSITY** für die unterschiedlichen Zink-Gruppen gibt. Formulieren Sie ein entsprechendes Modell, und führen Sie die entsprechende Varianzanalyse durch.

(d) (5 Punkte) Wie lauten die Schätzungen der Gruppenmittelwerte? Sind diese kompatibel mit Ihrer Beobachtung der Stripcharts?

(e) (6 Punkte) Nun möchten Sie überprüfen, ob die unterschiedlichen Flüsse **STREAM** neben der Zinkgruppe **ZNGROUP** einen Einfluss auf **DIVERSITY** haben. Führen Sie eine Zweiweg-Varianzanalyse durch. Wie interpretieren Sie die Variable **STREAM** in Bezug auf den Versuchsplan?

Aufgabe 4: Multiple Choice ..... 27 Punkte

*Es ist bei den folgenden Multiple-Choice-Aufgaben jeweils genau eine Antwort richtig. Kreuzen Sie also nur eine Antwort an. Eine korrekte Antwort gibt 3 **Plus**punkte und eine falsche Antwort einen **Minus**punkt. Minimal erhalten Sie null Punkte für Aufgabe 4.*

1) Sie haben eine Messreihe mit 20 Datenpunkten für den elektrischen Strom durch einen Kupferdraht. Der empirische Mittelwert ergab $I = 15.02\,\text{A}$ und für die empirische Standardabweichung fand man $s_I = 0.2203\,\text{A}$. Wie gross ist der relative Fehler?

   a) 0.003 %

   b) 4.9 %

   c) 0.05 %

   d) 0.3 %

   e) 0.05 A

2) Es sei $X \sim \mathcal{N}(\mu, \sigma^2)$ verteilt mit $\mu = 3$ und $\sigma^2 = 5$. Dann gilt

   a) $Y = 2X \sim \mathcal{N}(6, 10)$

   b) $Y = 2X \sim \mathcal{N}(3, 5)$

   c) $Y = 2X \sim \mathcal{N}(6, 20)$

   d) $Y = 2X \sim \mathcal{N}(3, 50)$

   e) keine der obigen Aussagen ist korrekt

3) Wird ein neuer Intelligenztest entwickelt, so muss er zuerst normiert werden. Dazu führen Psychologen den Test mit einer repräsentativen Gruppe aus der Bevölkerung durch. Die Psychologen ermitteln dann, wie viele Testaufgaben durchschnittlich gelöst wurden. Diesen Mittelwert definieren sie als IQ mit dem Wert 100. Als nächstes wird die Verteilung des IQ's so standardisiert, dass genau 34.1 Prozent der Getesteten einen IQ von über 115 haben. Die so resultierende Verteilung X des IQ's ist in guter Näherung $\mathcal{N}(\mu, \sigma^2)$ verteilt mit $\mu = 100$. Wie gross ist σ?

   a) 32.1 b) 36.6 c) 30.2 d) 37.9

4) Im Verzeichnis **`Austausch`** befindet sich die Datei **`rainDay.txt`**. Lesen Sie die Datei als Zeitreihe ein. Zeichnen Sie die Daten als Boxplots auf, wobei Sie als Gruppierungsvariable den Wochentag, Monat und das Quartal verwenden sollen. Welche Aussage ist korrekt?

   a) Am Mittwoch regnet es am stärksten.

   b) Im ersten Quartal hat es am wenigsten Regen.

   c) Von Mai bis September hat es am wenigsten Regen.

   d) Im Dezember hat es am wenigsten Regen.

5) Im Verzeichnis **`Austausch`** befindet sich die Datei **`rainDay.txt`**. Lesen Sie die Datei als Zeitreihe ein. Zeichnen Sie die täglichen Regenfälle auf, und zerlegen Sie die Zeitreihe in Trend, saisonale Effekt und Restterm. Welche der folgenden Aussagen ist korrekt?

   a) Es gibt einen steigenden Trend.

   b) Der saisonale Effekt ist approximativ konstant.

   c) Der Restterm bei einem additiven Modell zeigt keine saisonalen Effekte auf.

   d) Ein multiplikatives Modell wäre eher angebracht.

   e) Keine der obigen Aussagen ist korrekt.

6) Im Verzeichnis **`Austausch`** befindet sich die Datei **`rainDay.txt`**. Lesen Sie die Datei als Zeitreihe ein. Plotten Sie die relativen täglichen Zuwächse des Regenfalls. Welche der folgenden Aussagen ist korrekt?

   a) Es gibt einen steigenden Trend.

   b) Es gibt einen fallenden Trend.

   c) Es scheint einen Zusammenhang zu geben, so dass aus dem Vortag der Regenfall des Folgetages bestimmt werden kann.

   d) Es scheint keinen Zusammenhang zu geben, so dass aus dem Vortag der Regenfall des Folgetages nicht bestimmt werden kann.

   e) Keine der obigen Aussagen ist korrekt.

7) Angenommen, die Korrelation zwischen Einkommen und Weinkenntnisse ist 0.99. Wenn wir eine Person mit Weinkenntnissen kennenlernen, ist ihr Einkommen wahrscheinlich...

   a) klein

   b) gross

   c) Keine Aussage möglich

8) Angenommen, es stellt sich heraus, dass Personen mit grossem Einkommen auch grosse Weinkenntnisse haben, wenn die Korrelation der beiden Variablen gross ist. Sollte man also einen Kurs über Wein besuchen, um sein Einkommen zu verbessern?

   a) Ja, denn die grosse Korrelation beweist, dass grosse Weinkenntnisse ein grosses Einkommen verursachen.

   b) Es ist keine Aussage möglich. Die Weinkenntnisse könnten die Ursache für ein grosses Einkommen sein, aber das kann man mit der Korrelation nicht zweifelsfrei beantworten.

   c) Nein, denn die grosse Korrelation beweist, dass es keinen kausalen Zusammenhang zwischen grossen Weinkenntnissen und grossem Einkommen geben kann.

9) Betrachten Sie die Zahlen

$$1, 3, 4, 5, 6, 10, 23, 46$$

   Was ist das 20%-Quantil ($q_{0.2}$) dieser Zahlen

   a) $q_{0.2} = 3$

   b) $q_{0.2} = 12.25$

   c) $q_{0.2} = 5.5$

   d) $q_{0.2} = 1$

MEP Statistics for Data Science HS18

Musterlösungen

Lösung 1: ..... 22 Punkte

a) Gepaart. (**2 Punkte**)

b) $H_0: \mu_D = 0$ (**1 Punkt**)
$H_A: \mu_D \neq 0$ (**1 Punkt**) mit
$D_i = X_i - Y_i$, $\mu_D = \mu_X - \mu_Y$, (**1 Punkt**)
X_i : Messreihe nach Methode A,
Y_i : Messreihe nach Methode B.

c)

$$\overline{d}_n = \frac{1}{n}\sum_{i=1}^{n} d_i = -0.95 \quad \textbf{(2 Punkte)}$$

$$\widehat{\sigma}_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d}_n)^2$$
$$= 1.89 \quad \textbf{(2 Punkte)}$$

```
import pandas as pd
from pandas import DataFrame
import numpy as np
import scipy.stats as st
import warnings
warnings.filterwarnings("ignore")
widerstaende = pd.read_csv("./DatenHS18/widerstaende.txt", sep=" ", header=0)
methode_A = widerstaende["Methode A"]
methode_B = widerstaende["Methode B"]
d = methode_A - methode_B
d_mean = d.mean()
print(d_mean)

## -0.9502568696222127

d_var = d.var()
print(d_var)

## 1.8898564410608354
```

Der Code in **R** sieht wie folgt aus:

```
methode.a <- read.table(file = "./DatenHS18/widerstaende.txt",
    header = TRUE)[, 1]
methode.b <- read.table(file = "./DatenHS18/widerstaende.txt",
    header = TRUE)[, 2]
d <- methode.a - methode.b
mean(d)
```

```
## [1] -0.9502569

var(d)

## [1] 1.889856
```

d) **Teststatistik** Wir haben für $\overline{D}_n \sim \mathcal{N}(\mu_D, \frac{\sigma_D^2}{n})$ (Zentraler Grenzwertsatz). Dann ist die Teststatistik:

$$T = \frac{\overline{D}_n - \mu_0}{\widehat{\sigma}_D/\sqrt{n}} = \frac{\sqrt{n} \cdot \overline{D}_n}{\widehat{\sigma}_D} \quad \textbf{(1 Punkt)}$$

$$t = \frac{\sqrt{30} \cdot (-0.95)}{\sqrt{1.89}} = -3.786 \quad \textbf{(1 Punkt)}$$

Verteilung von T unter H_0

$$T \sim t_{n-1} = t_{29}. \quad \textbf{(1 Punkt)}$$

Verwerfungsbereich zeigt in Richtung der Alternativhypothese

$$\begin{aligned} K &= (-\infty, -t_{n-1,1-\alpha/2}] \cup [t_{n-1,1-\alpha/2}, \infty) \quad \textbf{(2 Punkte)} \\ &= (-\infty, -t_{29,0.975}] \cup [t_{29,0.975}, \infty) = (-\infty, -2.045] \cup [2.045, \infty) \quad \textbf{(1 Punkt)} \end{aligned}$$

```
import scipy.stats as st
print(st.t.ppf(q=0.975, df=29))

## 2.04522964213
```

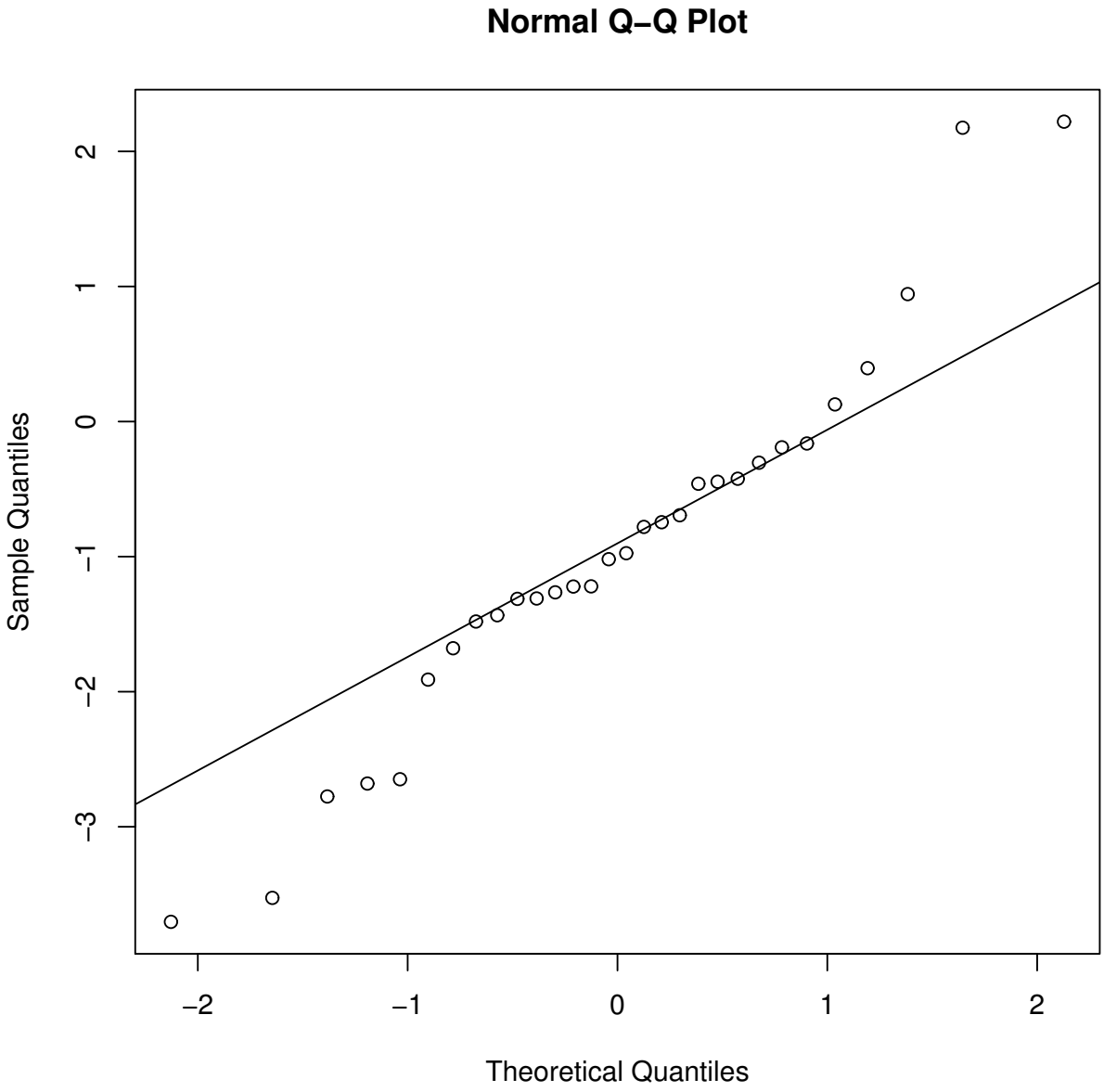
Der entsprechend **R**-Code lautet:

```
qt(0.975, 29)

## [1] 2.04523
```

Testentscheid $T \in K \Rightarrow$ H_0 wird verworfen. Es gibt einen statistisch signifikanten Unterschied zwischen den beiden Messmethoden (**1 Punkt**)

```
qqnorm(d)
qqline(d)
```



Aufgrund des QQ-Normal-Plots stellen wir fest, dass die Differenzen doch beträchtlich von einer Normalverteilung abweichen. Nun könnten entweder der Wilcoxon-Test oder der Vorzeichentest in Betracht gezogen werden. (**2 Punkte**) Alternativ kann der Testentscheid auch mit Hilfe des P-Wertes gefällt werden:

```
import scipy.stats as st
print(st.ttest_1samp(d, popmean=0).pvalue)

## 0.000712862938038
```

resp. in **R**

```
t.test(d, mu = 0, alternative = "two.sided")

##
##  One Sample t-test
##
## data:  d
```

```
## t = -3.7861, df = 29, p-value = 0.0007129
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.4635859 -0.4369278
## sample estimates:
##  mean of x
## -0.9502569
```

e) Das 95%-Vertrauensintervall (VI) beinhaltet mit 95%-Wahrscheinlichkeit den wahren Wert von μ_D und lautet

$$\begin{aligned} \text{VI} &= \left(\overline{d}_n - t_{n-1;1-\alpha}; \overline{d}_n + t_{n-1;1-\alpha}\frac{\widehat{\sigma}_D}{\sqrt{n}}\right] \quad \textbf{(2 Punkte)} \\ &= (-1.4636; -0.4369] \quad \textbf{(1 Punkt)} \end{aligned}$$

In **Python** können wir das Intervall direkt ermitteln mit Hilfe von

```
import scipy.stats as st
import numpy as np
print(st.t.interval(0.95, 29, loc = d.mean(), scale=d.std()/np.sqrt(d.size)))

## (-1.4635859390115671, -0.43692780023285827)
```

```
t.test(d, mu = 0, alternative = "two.sided")

##
##  One Sample t-test
##
## data:  d
## t = -3.7861, df = 29, p-value = 0.0007129
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.4635859 -0.4369278
## sample estimates:
##  mean of x
## -0.9502569
```

Da 0 nicht im Vertrauensintervall liegt, können wir schliessen, dass es einen signifikanten Unterschied zwischen den beiden Messmethoden gibt. (**1 Punkt**)

f) Das zweiseitige 95%-Vertrauensintervall (VI) bei bekannter Standardabweichung ermittelt sich wie folgt:

$$\begin{aligned} \text{VI} &= \left[\overline{d}_n - z_{1-\alpha}\frac{\sigma_D}{\sqrt{n}}; \overline{d}_n + z_{1-\alpha}\frac{\sigma_D}{\sqrt{n}}\right] \quad \textbf{(2 Punkte)} \\ &= [-1.442; -0.4583] \quad \textbf{(1 Punkt)} \end{aligned}$$

```
from scipy.stats import norm
print(norm.ppf(0.975))

## 1.95996398454
```

resp. in **R**

```
qnorm(0.975)

## [1] 1.959964
```

Das Vertrauensintervall wird kleiner, da die Unsicherheit aus der Schätzung von σ wegfällt. (**1 Punkt**)

g) Mit Hilfe von Bootstrapping können Vertrauensintervalle angegeben werden, ohne dass Annahmen zur Verteilung der Datenwerte getroffen werden müssen. Dazu werden aus dem Datensatz, hier bestehend aus jeweils 30 Werten, zufällig mit Zurücklegen gleich viele Werte gezogen, wie im Datensatz erhalten sind, also 11 Werte. Nun wird für ein solches Bootstrap-Sample der Mittelwert bestimmt. Das Verfahren wird nun 1000 Mal wiederholt, wobei für jedes Bootstrap Sample der Mittelwert bestimmt wird. Die Grenzen des 99 % Vertrauensintervalls erhält man nun, wenn man von den 1000 Mittelwerten das 0.5 %-Quantil als untere Grenze und das 97.5 %-Quantil als obere Grenze nimmt. Falls die beiden Vertrauensintervalle der beiden Altersgruppen überlappen, gibt es keinen signifikanten Unterschied und die Nullhypothese wird beibehalten.

Lösung 2: ..... 17 Punkte

(a) Da $\mathrm{E}[A] = \frac{1}{\lambda}$ für $A \sim \mathrm{Exp}(\lambda)$ ist, gilt:

$$\begin{aligned}\mathrm{E}[X(t = t_0)] &= \cos\left(\frac{\pi}{3}\right)\mathrm{E}[A] \quad (\textbf{1 Punkt}) \\ &= \frac{1}{2\lambda} \quad (\textbf{1 Punkt})\end{aligned}$$

Somit folgt mit der Momentenmethode

$$\widehat{\mathrm{E}}[X(t = t_0)] = \frac{1}{2\widehat{\lambda}} = 0.1 \quad (\textbf{1 Punkt})$$

also $\widehat{\lambda} = 5$ (**1 Punkt**).

(b) Es gilt

$$\begin{aligned}P[X(t = t_0) > 0.5] &= P\left[A \cdot \cos\left(\frac{\pi}{3}\right) > 0.5\right] \quad (\textbf{2 Punkte}) \\ &= P[A > 1] \\ &= 1 - P[A \leq 1] \quad (\textbf{1 Punkt})\end{aligned}$$

```
1 - pexp(1, rate = 5)

## [1] 0.006737947
```

Das heisst $P[X(t = t_0) > 0.5] = 0.006737947$ (**1 Punkt**).

(c) Wir finden für $\mu_X(t)$

$$\begin{aligned}\mathrm{E}[X(t)] &= \cos(\pi t) \cdot \mathrm{E}[A] \quad (\textbf{2 Punkte}) \\ &= \frac{\cos(\pi t)}{\lambda} \quad (\textbf{2 Punkte})\end{aligned}$$

(d) Für die Varianz von $X(t)$ finden wir

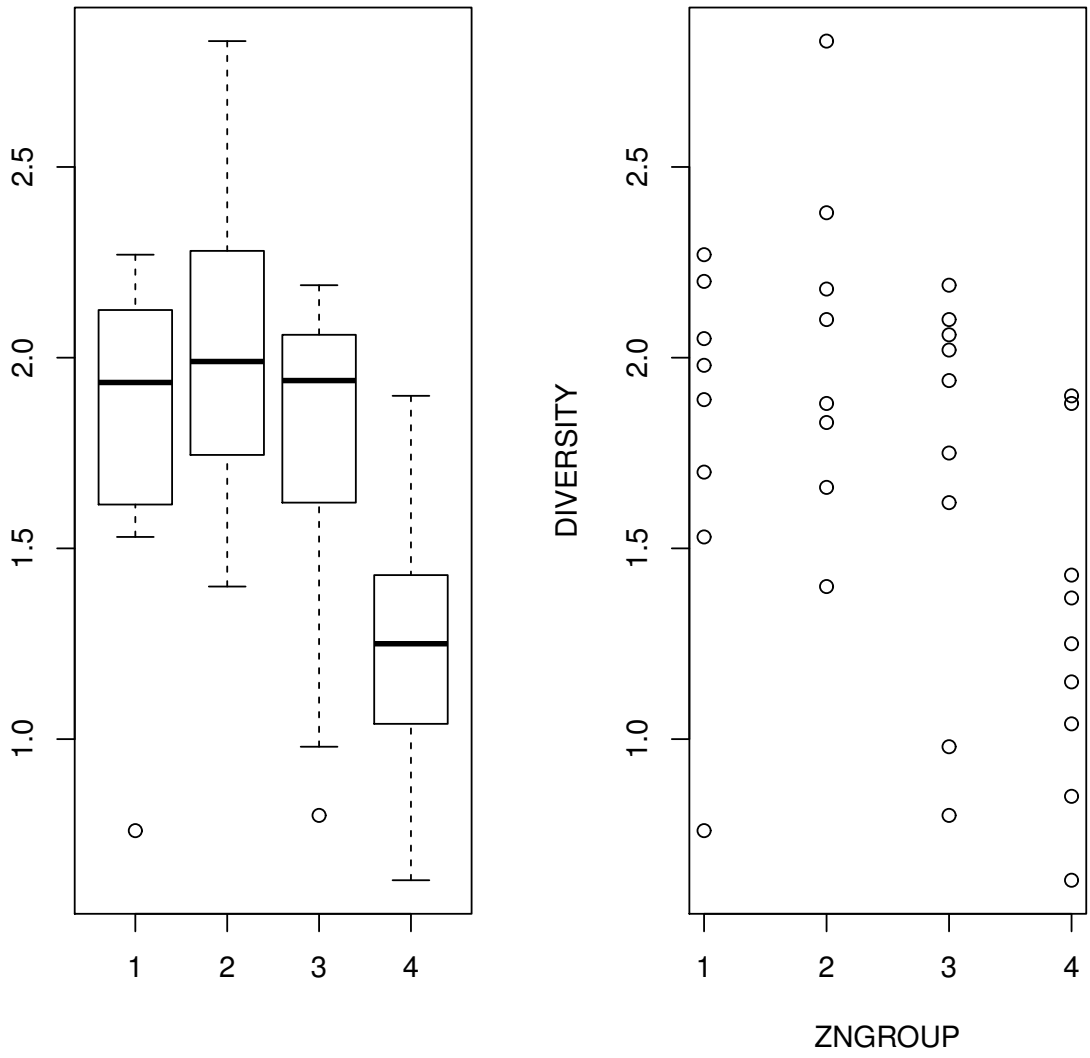
$$\begin{aligned} \text{Var}[X(t)] &= \cos^2(\pi t) \cdot \text{Var}[A] \quad (\textbf{2 Punkte}) \\ &= \frac{\cos^2(\pi t)}{\lambda^2} \quad (\textbf{2 Punkte}) \end{aligned}$$

(e) Nicht-stationär (**2 Punkte**)

Lösung 3: ..... 24 Punkte

(a)
```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import seaborn as sns
stream = pd.read_csv('./DatenHS18/stream.dat', sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
sns.stripplot(x="ZNGROUP", y="DIVERSITY", data=stream)
plt.xlabel("Zink-Gruppe")
plt.ylabel("Diversitaet")
sns.boxplot(x="ZNGROUP", y="DIVERSITY", data=stream)
plt.xlabel("Zink-Gruppe")
plt.ylabel("Diversitaet")
```

```
d.stream <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/stream.dat",
    header = TRUE)
d.stream[, "ZNGROUP"] <- as.factor(d.stream[, "ZNGROUP"])
par(mfrow = c(1, 2))
boxplot(DIVERSITY ~ ZNGROUP, data = d.stream)
stripchart(DIVERSITY ~ ZNGROUP, data = d.stream, vertical = TRUE,
    xlab = "ZNGROUP", pch = 1)
```



Der Boxplot weist zwei Ausreisser auf. Im Stripchart scheinen diese allerdings keine Ausreisser mehr zu sein (**2 Punkte**). Wenn nur wenige Datenpunkte vorhanden sind, so ist ein Boxplot nicht eine sehr vorteilhafte Visualisierungstechnik. Die Verteilung der Datenpunkte kann besser mit Stripcharts visualisiert werden. Zinkgruppe 4 weist eine signifikant tiefere Biodiversität auf.

(b)
```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
stream = pd.read_csv('./DatenHS18/stream.dat', sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit = ols("DIVERSITY ~ ZNGROUP", data=stream).fit()
print(anova_lm(fit))

##             df    sum_sq   mean_sq        F   PR(>F)
## ZNGROUP    3.0  2.566612  0.855537  3.93869  0.01756
## Residual  30.0  6.516411  0.217214      NaN      NaN
```

```
r.stream <- aov(DIVERSITY ~ ZNGROUP, data = d.stream)
summary(r.stream)

##             Df Sum Sq Mean Sq F value Pr(>F)
## ZNGROUP      3  2.567  0.8555   3.939 0.0176 *
## Residuals   30  6.516  0.2172
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Das Gruppenmittelmodell lautet

$$Y_i = \mu + \alpha_i + \varepsilon_i \quad (\textbf{1 Punkte})$$

wobei μ den (globalen) Mittelwert und α_i die Behandlungseffekte bezeichnen. Die Nullhypothese lautet :

$$\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0 \quad (\textbf{0.5 Punkte})$$

Die Alternativhypothese lautet, dass $\alpha_i \neq \alpha_j$ für mindestens zwei i j mit $i \neq j$ gilt (**0.5 Punkte**). Der F-Test ergibt ein signifikantes Resultat mit einem P-Wert von 0.018 (**4 Punkte**). Es gibt also einen signifikanten Unterschied der Biodiversität zwischen den unterschiedlichen Zinkgruppen (**2 Punkte**).

(c)
```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
stream = pd.read_csv('./DatenHS18/stream.dat', sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit = ols("DIVERSITY ~ ZNGROUP", data=stream).fit()
print(fit.summary())

##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:              DIVERSITY   R-squared:                       0.283
## Model:                            OLS   Adj. R-squared:                  0.211
## Method:                 Least Squares   F-statistic:                     3.939
## Date:                Tue, 26 Feb 2019   Prob (F-statistic):            0.0176
## Time:                        22:33:40   Log-Likelihood:                -20.159
## No. Observations:                  34   AIC:                             48.32
## Df Residuals:                      30   BIC:                             54.42
## Df Model:                           3
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [95.0% Conf. Int
## ------------------------------------------------------------------------------
## Intercept        1.7975      0.165     10.909      0.000         1.461     2.1
## ZNGROUP[T.2]     0.2350      0.233      1.008      0.321        -0.241     0.7
## ZNGROUP[T.3]    -0.0797      0.226     -0.352      0.727        -0.542     0.3
## ZNGROUP[T.4]    -0.5197      0.226     -2.295      0.029        -0.982    -0.0
## ==============================================================================
## Omnibus:                        2.067   Durbin-Watson:                   1.196
## Prob(Omnibus):                  0.356   Jarque-Bera (JB):                1.746
## Skew:                          -0.542   Prob(JB):                        0.418
```

```
## Kurtosis:                         2.757   Cond. No.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is co
```

```
dummy.coef(r.stream)

## Full coefficients are
##
## (Intercept):         1.7975
## ZNGROUP:                  1           2
##                  0.00000000  0.23500000
##
## (Intercept):
## ZNGROUP:                  3           4
##                 -0.07972222 -0.51972222
```

Das Gruppenmittelmodell lautet

$$Y_i = \mu + \alpha_i + \varepsilon_i$$

Die Ausgabe von **`dummy.coef(r.stream)`** ergibt, dass $\hat{\alpha}_1$ gleich null gesetzt wurde (**1 Punkt**). Somit sind $\hat{\mu}_1 = \hat{\mu} = 1.7975$ (**2 Punkte**), $\hat{\alpha}_2 = 0.235$, und somit ist $\hat{\mu}_2 = 1.7975 + 0.235 = 2.0325$ (**1 Punkt**), $\hat{\mu}_3 = 1.7975 - 0.0797 = 1.7178$ (**1 Punkt**) und $\hat{\mu}_4 = 1.7975 - 0.5197 = 1.2778$ (**1 Punkt**).

(d)
```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
stream = pd.read_csv('./DatenHS18/stream.dat', sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit2 = ols("DIVERSITY ~ ZNGROUP + STREAM", data=stream).fit()
print(anova_lm(fit2))

##              df    sum_sq   mean_sq         F    PR(>F)
## ZNGROUP    3.0  2.566612  0.855537  6.660454  0.001852
## STREAM     5.0  3.305153  0.661031  5.146196  0.002216
## Residual  25.0  3.211258  0.128450       NaN       NaN
```

Die Variable **`STREAM`** ist signifikant, man wird sie als Block-Variable interpretieren, da sie an und für sich für die Fragestellung nicht von Bedeutung ist (**3 Punkte**). Dennoch ist sie signifikant und sollte in der Analyse berücksichtigt werden (**2 Punkte**). Eine Interaktions-Analyse zwischen **`STREAM`** und **`ZNGROUP`** ergibt, dass es keine Interaktion zwischen den beiden Variablen gibt (**1 Punkt**).

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
stream = pd.read_csv('./DatenHS18/stream.dat', sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit3 = ols("DIVERSITY ~ ZNGROUP * STREAM", data=stream).fit()
print(anova_lm(fit3))
```

```
##                    df    sum_sq   mean_sq         F    PR(>F)
## ZNGROUP           3.0  2.566612  0.855537  7.118441  0.002642
## STREAM            5.0  3.305153  0.661031  5.500059  0.003430
## ZNGROUP:STREAM   15.0  2.232696  0.148846  1.238466  0.333107
## Residual         17.0  2.043163  0.120186       NaN       NaN
```

Lösung 4: ..... 27 Punkte

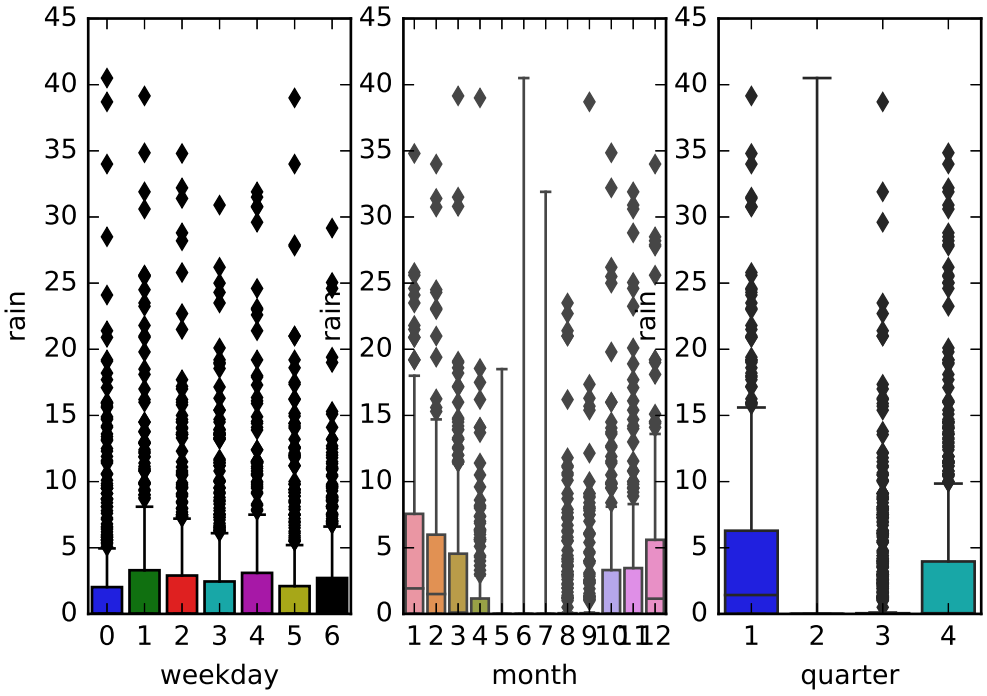
1.) (d) Der absolute Fehler ist $\Delta I = \frac{s_I}{\sqrt{n}} = \frac{0.2203}{\sqrt{20}}\,\mathrm{A} = 0.05\,\mathrm{A}$, der relative Fehler ist gegeben durch $(0.05\,\mathrm{A})/(15.02\,\mathrm{A}) \cdot 100\,\% = 0.3\,\%$.

2.) (c)

3.) (b)

4.) (c)

```
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame, Series
import seaborn
import numpy as np
rain = pd.read_csv('./Pruefungen/DatenFS18/rainDay.txt', sep=' ', header=None)
rain["Date"] = pd.DatetimeIndex(rain[1])
rain.set_index("Date", inplace=True)
rain["rain"] = rain[2] + 0.000001
rain_ts = Series(rain["rain"])
fig, axs = plt.subplots(ncols=3)
seaborn.boxplot(rain_ts.index.weekday, rain_ts, ax=axs[0])
axs[0].set(xlabel='weekday', ylabel='rain')
seaborn.boxplot(rain_ts.index.month, rain_ts, ax=axs[1])
axs[1].set(xlabel='month')
seaborn.boxplot(rain_ts.index.quarter, rain_ts,  ax=axs[2])
axs[2].set(xlabel='quarter')
```

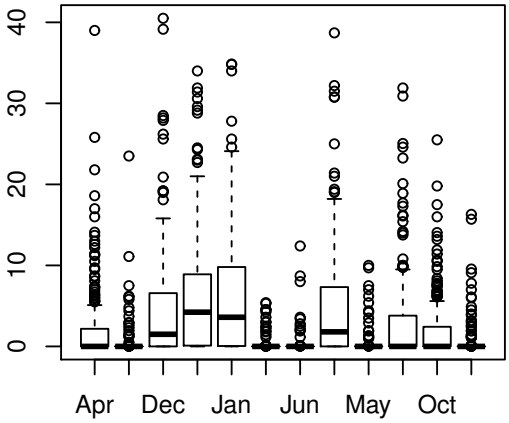


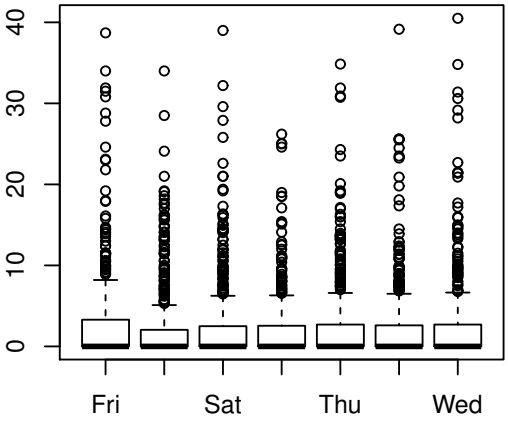
Der entsprechende **R**-Code sieht wie folgt aus:

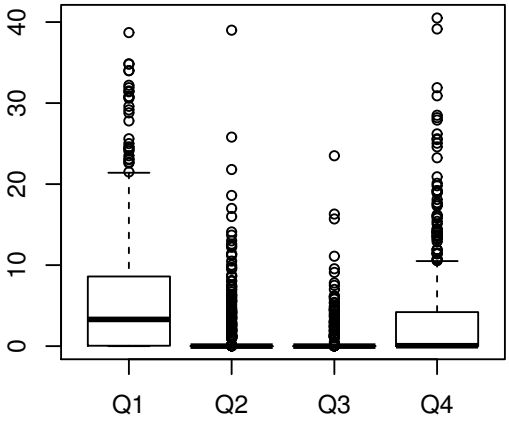
```
dd <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/rainDay.txt",
    header = T)
dd$DATE <- as.Date(dd$DATE, format = "%d.%m.%Y")
ts.dd <- ts(dd[, 2], start = 2000, freq = 365)
str(ts.dd)

##  Time-Series [1:2922] from 2000 to 2008: 0 12.9 0 0.05 3.55 2.05 3.5 7.65 1.1 9.8 ...

t.weekday <- factor(weekdays(dd$DATE, abbreviate = TRUE))  ##,levels=c('Mon','Tue','Wed','Thu','Fri','Sat','Sun'))
t.month <- factor(months(dd$DATE, abbreviate = TRUE))  ## levels=c('Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep',
t.quarter <- quarters(dd$DATE, abbreviate = FALSE)
dd.new <- data.frame(date = dd$DATE, rain = ts.dd,
    weekday = t.weekday, month = t.month, quarter = t.quarter)
par(mfrow = c(2, 2))
boxplot(dd$rain ~ dd.new$month)
boxplot(dd$rain ~ dd.new$weekday)
boxplot(dd$rain ~ dd.new$quarter)
```



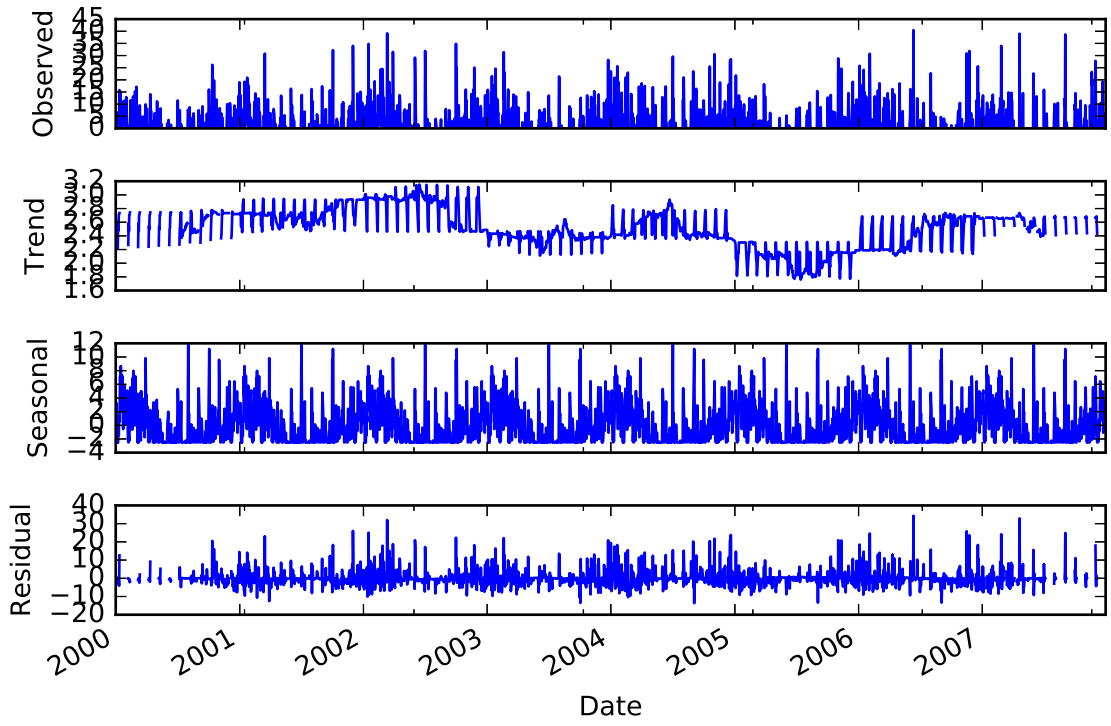


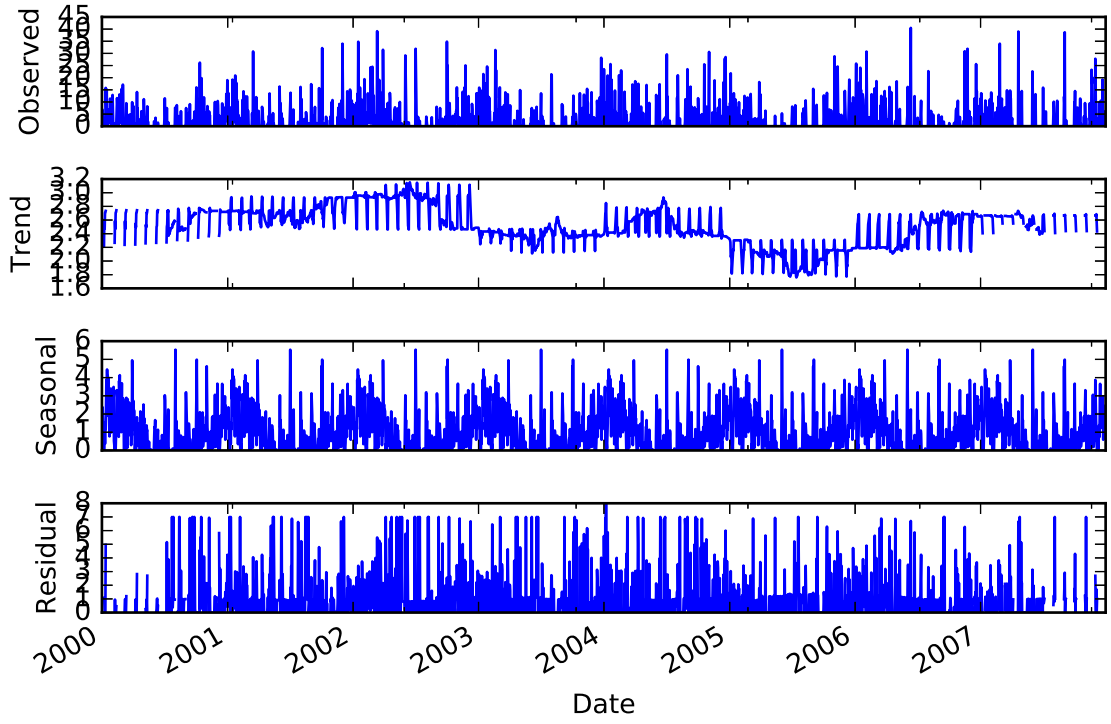


5.) (d)

```
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame
from statsmodels.tsa.seasonal import seasonal_decompose
rain = pd.read_csv('./Pruefungen/DatenFS18/rainDay.txt', sep=' ', header=None)
rain["Date"] = pd.DatetimeIndex(rain[1])
rain.set_index("Date", inplace=True)
# Wir wollen den Wert null vermeiden
rain["rain"] = rain[2] + 0.0000001
```

```
rain_ts = DataFrame(rain["rain"])
seasonal_decompose(rain_ts, model="additive", freq=365).plot()
seasonal_decompose(rain_ts, model="multiplicative", freq=365).plot()
```

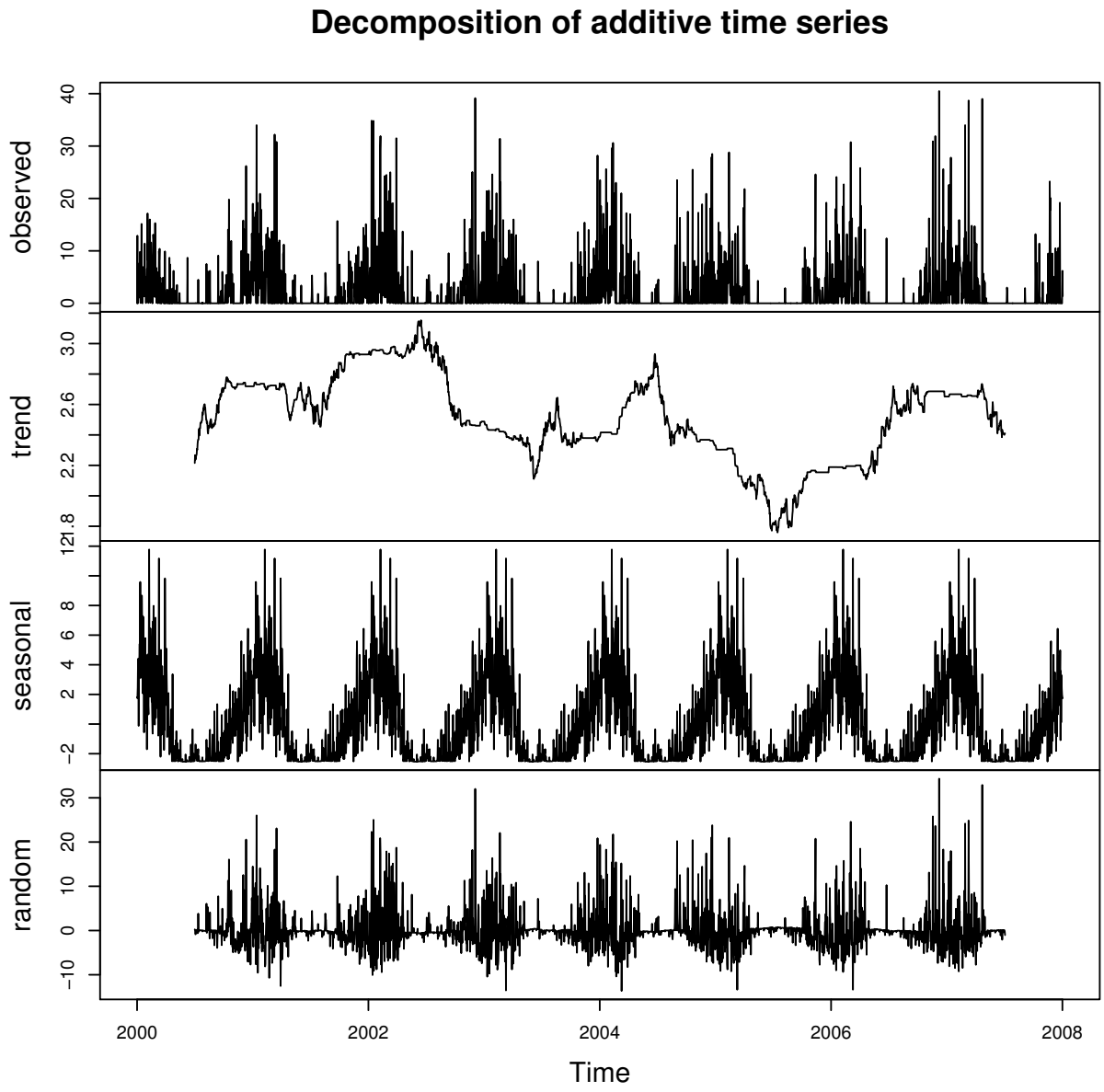




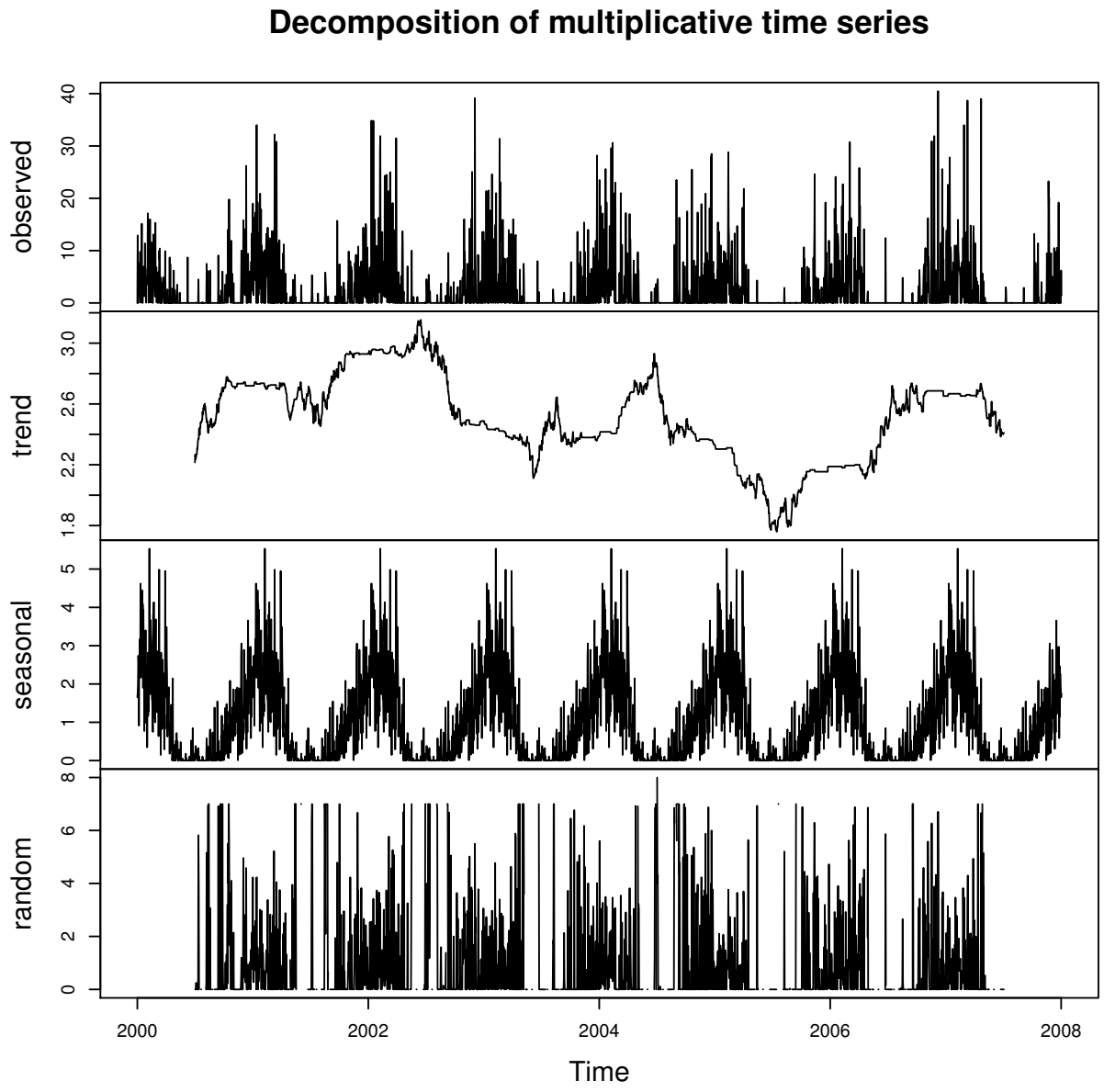
Der entsprechende **R**-Code sieht wie folgt aus:

```
dd <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/rainDay.txt",
    header = T)
dd$DATE <- as.Date(dd$DATE, format = "%d.%m.%Y")
ts.dd <- ts(dd[, 2], start = 2000, freq = 365)
```

```
decompose_rain_add <- decompose(ts.dd, "additive")
decompose_rain_mult <- decompose(ts.dd, "multiplicative")
par(mfrow = c(2, 2))
plot(decompose_rain_add)
```




```
plot(decompose_rain_mult)
```



Da der Restterm beim multiplikativen Modell beinahe keine saisonalen Effekte aufweist, scheint das multiplikative Modell eher zu passen.

6.) (d)

```
import matplotlib.pyplot as plt
import pandas as pd
from pandas import DataFrame
import numpy as np
from statsmodels.tsa.seasonal import seasonal_decompose
rain = pd.read_csv('./Pruefungen/DatenFS18/rainDay.txt', sep=' ', header=None)
rain["Date"] = pd.DatetimeIndex(rain[1])
rain.set_index("Date", inplace=True)
rain_rel_zuwachs = np.log(rain) - np.log(rain.shift(-1))
rain_rel_zuwachs.plot()
```

7.) (b)

8.) (b)

9.) (a)