

## Probeklausur Musterlösung

**Aufgabe 1** Auf der Homepage zur Veranstaltung finden Sie unter dem Punkt Klausurdaten eine Datei `elastic.csv`. Laden Sie diese herunter und speichern Sie sie an einem geeigneten Ort ab.

a) Mit welchem Befehl lesen Sie die Datei in eine Variable `elastic` ein?

```
elastic <- read.table(  
  "http://fawn.hsu-hh.de/~steuer/Klausur/elastic.csv",  
  header=TRUE, sep=";", strip.white=TRUE)  
str(elastic)
```

Die Datei enthält Beobachtungen eines Versuchs, bei dem für zwei verschiedene Materialien von Gummibändern gemessen wurde, bei welcher Auslängung des Bandes (in mm) dieses wie weit (in cm) fliegt.

**b)** Fertigen Sie einen Plot an, aus dem sie erkennen können, ob die Materialien sich ähnlich verhalten. Begründen Sie die Wahl des Plots und beschreiben Sie Ihre Beobachtungen! Notieren Sie die Befehle, die Sie zur Anfertigung des Plots genutzt haben.

```
plot(elastic[,1], elastic[,2], col=elastic[,3])  
boxplot(elastic[,2] ~ elastic[,3])
```

**c)** Fällt eine Beobachtung besonders auf? Wenn ja, entfernen Sie diese für die weitere Analyse aus den Daten. Mit welchem Befehl identifizieren Sie den auffälligen Punkt? Mit welchem Befehl entfernen Sie ihn aus den weiteren Analysen. Notieren Sie die nötigen Befehle.

```
which(elastic[,2] > 1000)  
elastic <- elastic[-7,]
```

**d)** Betrachten Sie ein lineares Regressionsmodell für den bereinigten Datensatz. Mit welchem Befehl berechnen Sie in R das Modell? Interpretieren Sie das Ergebnis. Sehen Sie ein inhaltliches Problem mit dem Ergebnis der Regression?

```
summary(lm(Distanz ~ Streckung, elastic))
Call: lm(formula = Distanz ~ Streckung, data = elastic)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.8805     19.6089  -3.564  0.00311 **
### Und was heisst das inhaltlich?

Streckung      5.7467      0.4239  13.557 1.92e-09 ***
---
Residual standard error: 15.27 on 14 degrees of freedom
Multiple R-squared:  0.9292, Adjusted R-squared:  0.9242
F-statistic: 183.8 on 1 and 14 DF,  p-value: 1.925e-09
```

**e)** Führen Sie die einfache lineare Regressionen für jedes der beiden Materiale durch. Was sind die Hauptunterschiede in den Regressionsergebnissen?

```
weich <- which(elastic$Material == "weich")
lm(Distanz ~ Streckung , elastic[weich,])
lm(Distanz ~ Streckung , elastic[-weich,])
plot()
```

**f)** Zeichnen Sie mit R in einen Scatterplot, der die Regressionsgeraden für beide Fälle enthält. Notieren Sie die Befehle. Fällt Ihnen bei Betrachtung der Residuen eine Idee zur Modellverbesserung ins Auge?

```
plot(elastic$Streckung, elastic$Distanz, col=elastic$Material)
abline(lm(Distanz ~ Streckung , elastic[weich,]),col=2)
abline(lm(Distanz ~ Streckung , elastic[-weich,]),col=1)
plot(lm(Distanz ~ Streckung , elastic[weich,]))
plot(lm(Distanz ~ Streckung , elastic[-weich,]))
### evtl leicht nicht-linear
```

**Aufgabe 2:** Erzeugen sie 100 Datenpunkte aus der Normalverteilung mit  $\mu = 11$  und  $\sigma = 3$ . Zeichnen Sie die empirische Verteilungsfunktion und die theoretische Verteilungsfunktion für diese Punkte in einen Plot. Notieren Sie die Befehle, die dazu nötig sind.

```
curve(pnorm(x, mean=11, sd=3), 4,17, main="Vergleich ecdf und Vtg")  
lines(ecdf(rnorm(100, mean=11, sd=3)), pch="+")
```

**Aufgabe 3: a)** Welche Wertemenge bezeichnet man als *five-number-summary* nach Tuckey?

```
?fivenum
```

```
Tukey Five-Number Summaries
```

```
Description:
```

```
Returns Tukey's five number summary (minimum, lower-hinge,  
median, upper-hinge, maximum) for the input data.
```

**b)** In welcher grafischen Darstellung spielen diese Werte eine große Rolle und was möchte man über die Daten aus dieser Darstellung ablesen?

**Lösung:** Boxplot, Schiefe/Symmetrie, evtl Lagevergleich bei parallelen Boxplots

**Aufgabe 4:** Es sollen drei Waschmittel auf Unterschiede in ihrer Waschkraft untersucht werden. Gemessen wird die Waschkraft als Anteil des reflektierten Lichts in Prozent nach einer Wäsche an der Reflektion des Lichtes an einem reinweissen Referenzstoffs. Die Verschmutzung wurde durch gleichmässige Einfärbung simuliert. Die Daten zu diesem Experiment finden Sie in der Datei `waschkraft.csv` im selben Verzeichnis wie die Daten aus Aufgabe 1.

a) Lesen Sie die Datei ein und speichern Sie die Daten in einer Variablen `waschen`.

```
waschen<- read.table(  
  "http://fawn.hsu-hh.de/~steuer/Klausur/waschkraft.csv",  
  header=TRUE,sep="," , strip.white=TRUE, dec=".")
```

b) Welches statistische Verfahren beantwortet Fragestellungen wie die vorliegende?

**Lösung:** ANOVA

c) Führen Sie die entsprechende Analyse durch. Interpretieren Sie das Ergebnis für die vorliegende Fragestellung.

```
str(waschen)
```

```
'data.frame': 60 obs. of 2 variables:
```

```
$ Mittel      : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
$ Reflektion: num  48.2 48.4 49.9 47.5 46.4 ...
```

```
waschen[,1] <- as.factor(waschen[,1])
```

```
> summary(aov(Reflektion ~ Mittel, waschen))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mittel	2	13.336	6.6679	5.8143	0.005035 **
Residuals	57	65.368	1.1468		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Erzeugen Sie parallele Boxplots für die Reflektion der verschiedenen Waschmittel. Welches Mittel unterscheidet sich augenscheinlich von den anderen? **Lösung:**

```
boxplot(waschen$Reflektion ~ waschen$Mittel)
### Mittel 1.
```

e) Wie viele Paarvergleiche wären nötig, um mittels statistischer Tests herauszufinden, welche Waschmittelpaare unterschiedlich sind?

**Lösung:**  $\binom{3}{2} = 3$

f) Welche Problematik tritt bei zahlreichen Paarvergleichen auf und wie kann man dem Problem entgegenwirken?

**Lösung:** Problematik des multiplen Testens, einfachste Gegenmaßnahme: Bonferroni.