



## Lösungsblatt: Statistik mit R und RStudio

Musterlösungen zum Übungsdatensatz

<b>Hinweis</b>	Die folgenden Lösungen sind bewusst einsteigerfreundlich formuliert. Bei Diagrammen sind alternative, aber fachlich korrekte Lösungen ebenfalls möglich.
<b>Datensatz</b>	<b>Uebungsdatensatz_R_RStudio.csv</b> mit Semikolon als Trennzeichen

### Lösung 1: Daten importieren und Überblick

```
daten <- read.csv("Uebungsdatensatz_R_RStudio.csv", sep = ";", header = TRUE)
head(daten)
str(daten)
summary(daten)
```

#### Erwartete Kernergebnisse:

- 72 Zeilen und 14 Spalten

**Interpretation:** Der Datensatz enthält 72 Fälle und 14 Variablen. Die Struktur zeigt numerische und kategoriale Variablen.

### Lösung 2: Fehlende Werte erkennen

```
colSums(is.na(daten))
daten[!complete.cases(daten), ]
```

#### Erwartete Kernergebnisse:

- Fehlende Werte:
  - Lernzeit\_Std\_Woche = 2,
  - Motivation\_1\_10 = 2,
  - Koffein\_Tassen\_Tag = 1,
  - Teilnahmequote\_Prozent = 1,
  - Note = 2,
  - Kurszufriedenheit\_1\_10 = 2

**Interpretation:** Fehlende Werte treten in mehreren Variablen auf und müssen bei Berechnungen mit `na.rm = TRUE` oder `use = "complete.obs"` berücksichtigt werden.



### Lösung 3: Datentypen anpassen

```
daten$Geschlecht <- factor(daten$Geschlecht)
daten$Gruppe <- factor(daten$Gruppe)
daten$Vorwissen <- factor(daten$Vorwissen, levels = c("keins", "wenig", "mittel",
"gut"), ordered = TRUE)
daten$Abgabe_puenktlich <- factor(daten$Abgabe_puenktlich)
daten$Bestanden <- factor(daten$Bestanden)
str(daten)
```

**Interpretation:** Die kategorialen Variablen sind nun als Faktoren gespeichert und damit für Tabellen, Gruppierungen und Tests besser geeignet.

### Lösung 4: Deskriptive Statistik

```
mean(daten$Alter)
median(daten$Alter)
sd(daten$Alter)
range(daten$Alter)
mean(daten$Note, na.rm = TRUE)
median(daten$Note, na.rm = TRUE)
sd(daten$Note, na.rm = TRUE)
range(daten$Note, na.rm = TRUE)
table(daten$Geschlecht)
table(daten$Gruppe)
table(daten$Bestanden)
```

#### Erwartete Kernergebnisse:

- ▶ Alter: Mittelwert 23.32, Median 23, SD 4.30, Bereich 17 bis 32
- ▶ Note: Mittelwert 4.69, Median 4.6, SD 0.63, Bereich 3.1 bis 5.9
- ▶ Gruppe: A = 34, B = 38
- ▶ Bestanden: ja = 66, nein = 6

**Interpretation:** Die Noten liegen im Mittel im genügenden Bereich. Die Mehrheit der Teilnehmenden hat bestanden.



## Lösung 5: Gruppenauswertungen

```
aggregate(Note ~ Gruppe, data = daten, mean, na.rm = TRUE)
aggregate(Note ~ Geschlecht, data = daten, mean, na.rm = TRUE)
aggregate(Lernzeit_Std_Woche ~ Vorwissen, data = daten, mean, na.rm = TRUE)
```

### Erwartete Kernergebnisse:

- ▶ Durchschnittsnote Gruppe A = 4.61, Gruppe B = 4.77
- ▶ Durchschnittsnote Geschlecht: divers = 4.90, m = 4.55, w = 4.79
- ▶ Durchschnittliche Lernzeit nach Vorwissen: gut = 5.85, keins = 5.42, mittel = 5.35, wenig = 5.64

**Interpretation:** Gruppe B hat im Datensatz eine leicht höhere Durchschnittsnote als Gruppe A. Die Unterschiede sind jedoch klein.

## Lösung 6: Diagramme erstellen

```
barplot(table(daten$Bestanden), main = "Bestanden", ylab = "Anzahl")
hist(daten$Note, main = "Verteilung der Noten", xlab = "Note")
boxplot(Note ~ Gruppe, data = daten, main = "Noten nach Gruppe", xlab = "Gruppe",
        ylab = "Note")
plot(daten$Lernzeit_Std_Woche, daten$Note, main = "Lernzeit und Note", xlab = "Lern-
zeit pro Woche", ylab = "Note")
```

**Interpretation:** Im Histogramm ist eine Konzentration im Bereich um 4.5 bis 5.0 sichtbar. Das Streudiagramm zeigt einen positiven Zusammenhang zwischen Lernzeit und Note.

## Lösung 7: Korrelationen

```
cor(daten$Lernzeit_Std_Woche, daten$Note, use = "complete.obs")
cor(daten$Motivation_1_10, daten$Note, use = "complete.obs")
cor(daten$Fehlstunden, daten$Note, use = "complete.obs")
```

### Erwartete Kernergebnisse:

- ▶ Lernzeit und Note:  $r = 0.48$  (positiver, mittlerer Zusammenhang)
- ▶ Motivation und Note:  $r = 0.31$  (positiver, eher schwacher bis mittlerer Zusammenhang)
- ▶ Fehlstunden und Note:  $r = -0.34$  (negativer, mittlerer Zusammenhang)

**Interpretation:** Mehr Lernzeit und höhere Motivation gehen tendenziell mit besseren Noten einher. Mehr Fehlstunden gehen mit tieferen Noten einher.



## Lösung 8: Statistische Tests

```
t.test(Note ~ Gruppe, data = daten)
chisq.test(table(daten$Gruppe, daten$Bestanden))
```

### Erwartete Kernergebnisse:

- ▶ t-Test:  $p = 0.290$  -> kein signifikanter Unterschied zwischen Gruppe A und B
- ▶ Chi-Quadrat-Test:  $p = 0.776$  -> kein signifikanter Zusammenhang zwischen Gruppe und Bestanden

**Interpretation:** Beide Tests liefern in diesem Datensatz keinen statistisch signifikanten Unterschied bzw. Zusammenhang.

## Lösung 9: Lineare Regression

```
modell <- lm(Note ~ Lernzeit_Std_Woche, data = daten)
summary(modell)
```

### Erwartete Kernergebnisse:

- ▶ Regressionsgleichung:  $\text{Note} = 3.926 + 0.139 \cdot \text{Lernzeit\_Std\_Woche}$
- ▶ Bestimmtheitsmass  $R^2 = 0.228$

**Interpretation:** Mit jeder zusätzlichen Lernstunde pro Woche steigt die vorhergesagte Note im Modell leicht an.

## Lösung 10: Transferaufgabe

```
daten$Leistungsgruppe <- ifelse(daten$Note >= 5.0, "hoch", "normal")
table(daten$Leistungsgruppe)
```

### Erwartete Kernergebnisse:

- ▶ Leistungsgruppe hoch = 24, normal = 46
- ▶ Mögliche Erkenntnis 1: Die Erfolgsquote im Datensatz ist hoch.
- ▶ Mögliche Erkenntnis 2: Lernzeit hängt positiv mit der Note zusammen.

**Interpretation:** Die neue Variable erlaubt eine einfache Gruppierung nach Leistung und kann für weitere Vergleiche genutzt werden.