DRS Spring Kurs: Biostatistik

Lorenz Gygax & Vitaly Belik

Institute for Veterinary Epidemiology and Biostatistics, FU Berlin

Mar 11, 2020

Schedule

URL: http://belik.userpage.fu-berlin.de/springschool

	Wednesday, 11th March	Thursday, 12th March
8.30 – 10.00	R Intro (VB)	Proportions, Crosstabs, RR/IR and OR + Exercise (LG)
Break		
10.30 – 12.00	Basics of statistical hypothesis testing, general aspects in choice of a model + Exercise (LG)	Linear regression models, regression parameters and fit, (G)LM LinRreg (with ANOVA table) + Exercise (LG)
Lunch		
13.00 – 14.00	Descriptive statistics. Association concept, Graphical representation + Exercise (VB)	Model diagnostics, fit, outliers, residuals, assumptions (transformations, alternative distributions) + Exercise (LG)
Break		
14:15-15:15	ANOVA. Simple linear models + Exercise (VB)	Exercise: Model diagnostics, fit, outliers, residuals, assumptions (transformations, alternative distributions) (LG)
		Odds Ratios from logistic regression (VB)
Break		
15:30 -16:30	Exercise (VB)	Exercise: Odds Ratios from logistic regression (VB)
Break		
17:00-18:00	Extended exercises, self-study	Extended exercises, self-study

Kontakt

Prof. Dr. Vitaly Belik

Fachbereich Veterinärmedizin

Institut für Veterinär-Epidemiologie und Biometrie

Juniorprofessor

Leitung Arbeitsgruppe Systemmodellierung

Adresse	Königsweg 67 Raum 104 14163 Berlin
Telefon	+49 30 838 61129
Fax	+49 30 838 4 61129
E-Mail	vitaly.belik@fu-berlin.de
Homepage	Working Group Modelling

Was ist Statistik / Biostatistik (Biometrie)?

- Was sind Ihre Erwartungen?

Was ist Statistik?

- Welche Daten soll man zur Beantwortung einer gegebenen Aufgabenstellung ermitteln?
- Wie viel Daten soll man ermitteln?
- Auf welche Art soll man das Untersuchungsmaterial auswählen?
- Wie soll mans eine Untersuchungsdaten ermitteln?
- Wie sollen die gewonnenen Daten geordnet werden?
- Wie sollen die Daten beschrieben und übersichtlich dargestellt werden?
- Wie wertet man die Daten aus?
- Welche Schlüsse lassen sich ziehen?
- Wie zuverlässig sind die getroffenen Aussagen?
- Welche weiterführenden Fragestellungen haben die Ergebnisse aufgeworfen?

Was ist Statistik? (1)

1. Aufgabenstellung. Nach präziser Formulierung der Fragestellung muss eine geeignete Wahl von Merkmalengetroffen,eine Mess-bzw. Beobachtungsmethode festgelegt und ein Versuchsplan aufgestellt werden.
2. Datengewinnung. Gewinnung des Untersuchungsmaterials (Ziehen der Stichprobe) und Ausführungder Messungen bzw. Beobachtungen an diesem Material.
3. Datenverarbeitung. Das gewonnene Datenmaterial muss graphisch und rechnerisch aufbereitet werden, dann sind Schlüsse von der Stichprobe auf die Grundgesamtheit zu ziehen; diese werden anschließend geprüft und interpretiert.

Was ist Statistik? (2)

Statistik

ist eine wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung von Methoden zur Datenerhebung, -beschreibung und -analyse sowie der Beurteilung der Ergebnisse ist. Dabei unterscheidet man:

Deskriptive (beschreibende) Statistik:

Methoden zur Auswertung und übersichtlichen Darstellung und Zusammenfassung von Daten.

Induktive (schliessende) Statistik:

Methoden zum Treffen von vernünftigen Entscheidungen im Falle von Unsicherheit bzw. Risiko. “Den Zufall in den Griff bekommen”. “Sicherheit über Unsicherheit gewinnen”.

Biostatistik oder Biometrie

Biostatistik (Biometrie)

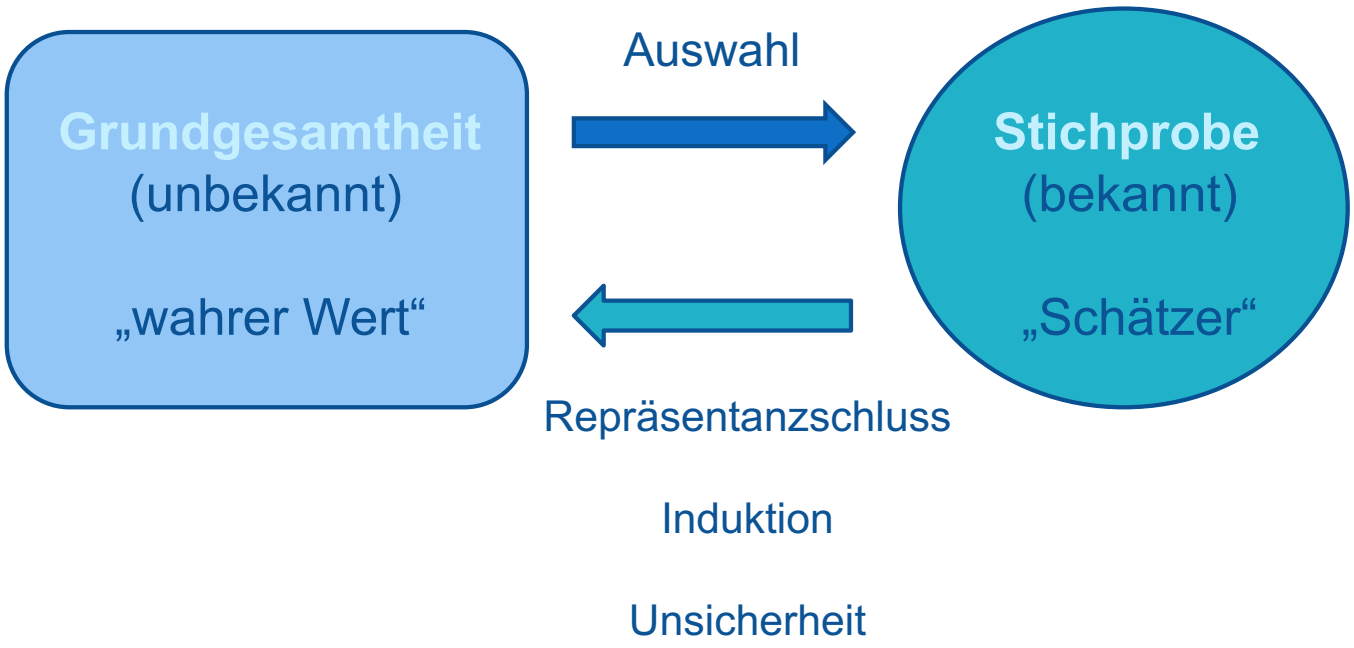
angewandte Statistik zur Beschreibung, Modellierung und Beurteilung biologisch-naturwissenschaftlicher Phänomene.

Beispiele

- Wie sicher ist das Ergebnis eines Diagnosetests zur Bestimmung einer Erkrankung?
- Wie viele Versuche müssen durchgeführt werden, um Verbesserung eines Produktes zu gewährleisten?

Deskriptive Statistik wird manchmal als *explorative* und Schliessende Statistik als *konfirmatorische Datenanalyse* bezeichnet.

Daten und Zufall



[Grafik: M. Doherr]

Statistische Schlüssweise

- Schätzen der unbekannten Parameter der Grundgesamtheit. “Finde eine Größe aus den Daten der Stichprobe, die”möglichst nah" an der unbekannten Wirklichkeit ist."
- Angabe von Konfidenzintervallen (Vertrauensbereichen). “Gebe basierend auf den Daten der Stichprobe ein Intervall an, das den wahren Wert (Populations-Parameter) mit einer gewissen Wahrscheinlichkeit überdeckt.”
- Entscheiden mittels eines statistischen Tests, ob anhand der Daten der Stichprobe eine Aussage über einen Parameter der Grundgesamtheit (bspw. Unterschied eines Mittelwertes zwischen Gruppen) wahr oder falsch ist.

Lernziele des Kurses

Ziel des Kurses is es Ihnen die wichtigsten statistischen Methoden zur Planung und Auswertung der Versuche und Daten aus wissenschaftlicher Studien zu vermitteln.

Sie sollen die Notwendigkeiten, Möglichkeiten und Grenzen grundlegender statistischer Analysen verstehen und selbst einfache statistische Berechnungen durchführen können.

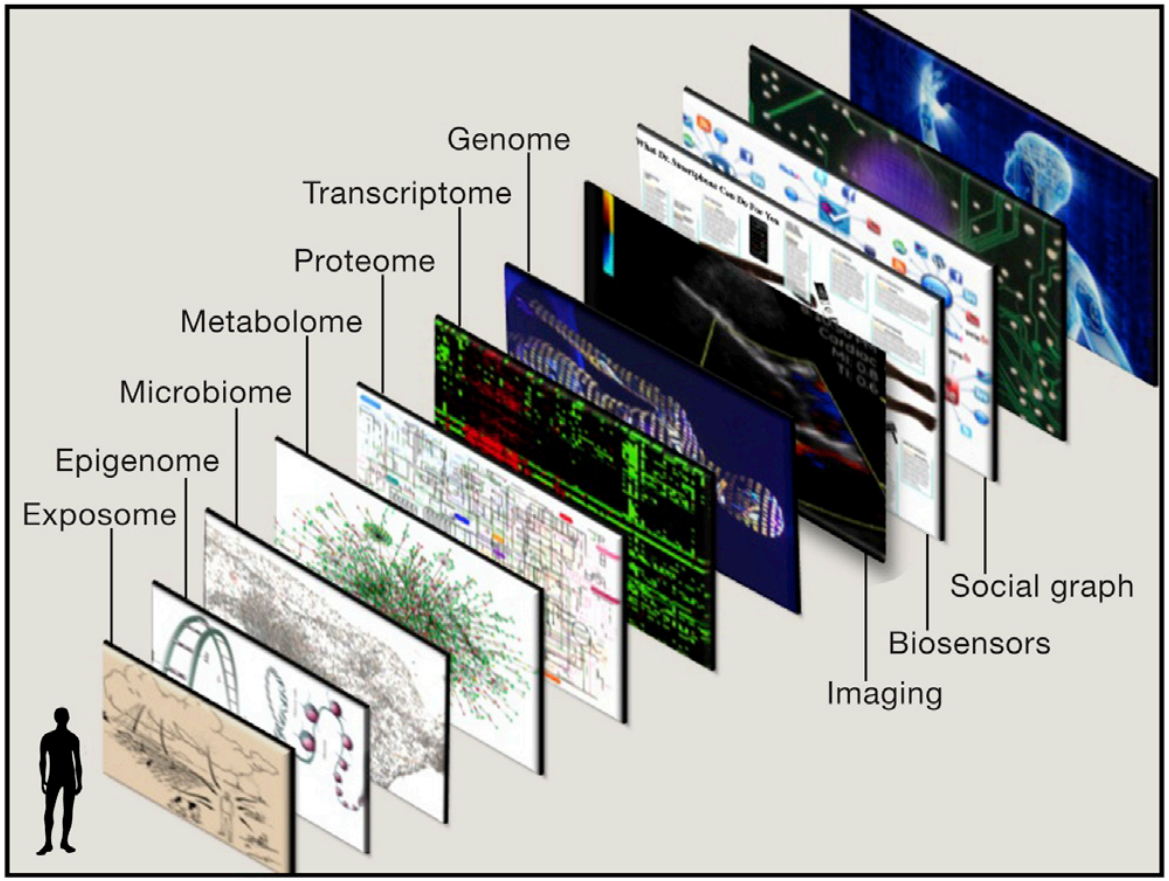
Falls nötig, sollen Sie in der Lage sein bei einer statistischen Beratung, Ihr Anliegen sicher zu kommunizieren.

Literatur

1. R. Kabacoff. *R in Action*
2. A. Field *Discovering Statistics Using R*
3. M. Crawley. *The R Book*
4. I. Dohoo et al. *Veterinary Epidemiologic Research*
5. A. Field. *An Adventure in Statistics: The Reality Enigma*

Daten

In letzter Zeit mit der rasanten Entwicklung der ausgefallenen Sensoren (IoT), Rechner- und Speicherkapazitäten werden sehr viele Daten prozudiert.

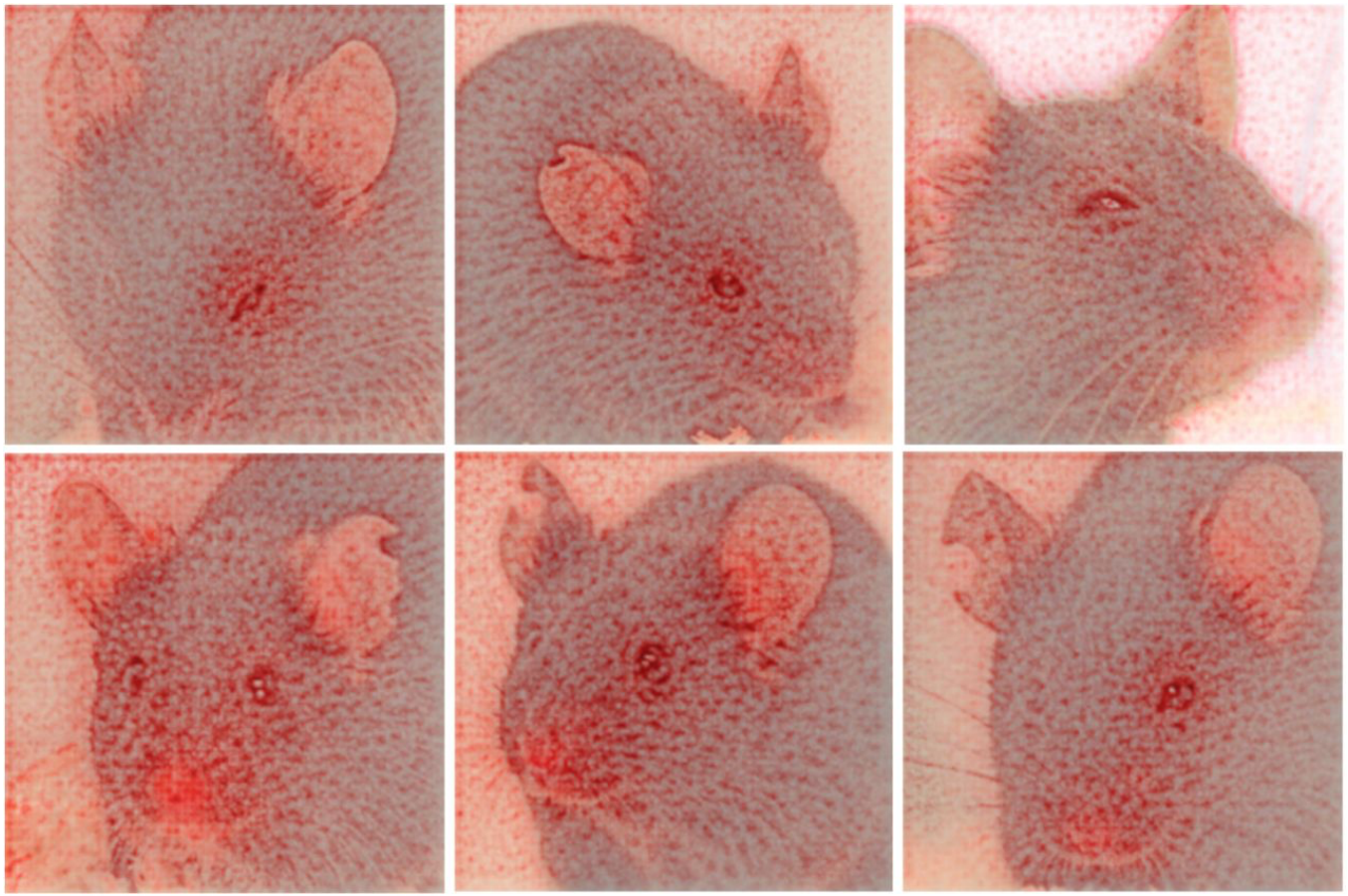


[Topol, 2014]

Daten sind heterogen

Bilder

Feature visualization von Mäuserbildern

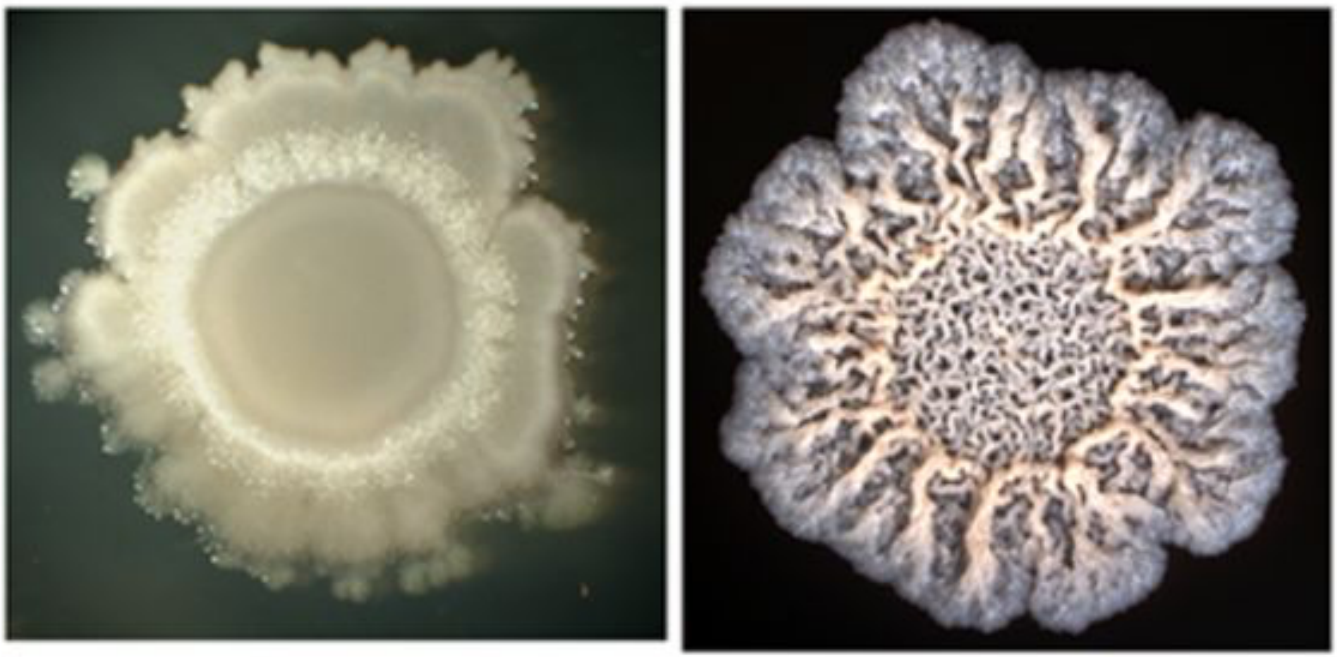


[doi:10.1101/582817]

Daten sind heterogen (1)

Bilder

Biofilm von *B. subtilus*

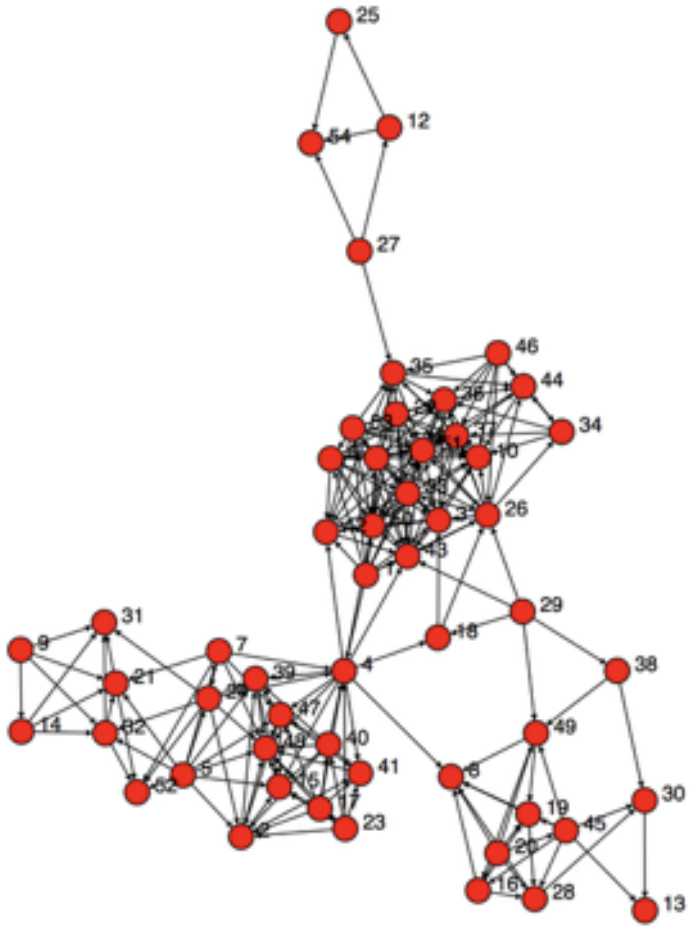


[doi:10.1128/JB.00028-13]

Daten sind heterogen (2)

Netzwerke

Kontaktnetzwerk von Tieren



[Daten: Thomas Selhorst]

Maschinelles Lernen

Große Mengen von heterogenen Daten (Big Data) verlangen nach entsprechenden Werkzeugen für die Datenanalyse. Dabei können, ausser klassischen statistischen Methoden, das *maschinelle Lernen* (z.B. *künstliche Neuronale Netze*) sehr hilfreich sein.

Es stellt sich sogar die Frage, ob sich die Versuchsplanung und Datenanalyse nicht von einer Maschine erledigen lässt.

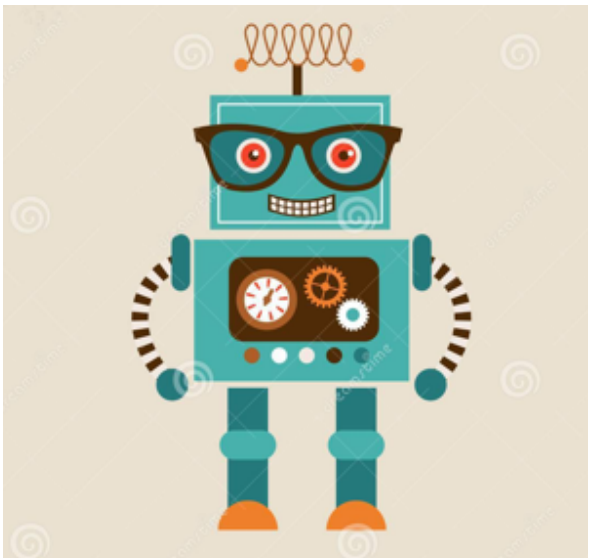
Roboscientist

SCIENCE VOL 324 3 APRIL 2009 **85**

The Automation of Science

Ross D. King,[1]* Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae* and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam's conclusions through manual experiments. To describe Adam's research, we have developed an ontology and logical language. The resulting formalization involves over 10,000 different research units in a nested treelike structure, 10 levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalization describes how a machine contributed to scientific knowledge.



Zurück zu eigentlichen Biostatistik!

Daten als Tabelle

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False
10	1	3	female	4.0	1	1	16.7000	S	Third	child	False	G	Southampton	yes	False
11	1	1	female	58.0	0	0	26.5500	S	First	woman	False	C	Southampton	yes	True
12	0	3	male	20.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
13	0	3	male	39.0	1	5	31.2750	S	Third	man	True	NaN	Southampton	no	False
14	0	3	female	14.0	0	0	7.8542	S	Third	child	False	NaN	Southampton	no	True
15	1	2	female	55.0	0	0	16.0000	S	Second	woman	False	NaN	Southampton	yes	True

Merkmale

Individuen oder Untersuchungsobjekte, die einer Erhebung / Untersuchung zu Grunde liegen, d.h. an / von denen Daten gesammelt werden, bezeichnet man als statistische Einheit, Merkmalsträger oder Untersuchungseinheiten.

Die Eigenschaften, die hinsichtlich des Untersuchungsziels an der statistischen Einheit untersucht werden, heißen *Merkmale*.

Studierendendaten

- Geschlecht, Körpergröße, Geburtsjahr
- Stadtnah oder ländlich aufgewachsen
- Wunsch, nach dem Studium in einem bestimmten Unternehmen zu arbeiten

Auswahl geeignete Merkmale

Objektivität

Die Ausprägung der zu ermittelneden Merkmales ist unabhängig von der Person des Auswerters eindeutig festszustellen.

Reliabilität

Das Merkmal gestattet reproduzierbare Mess- (bzw. Beobachtungs-) Ergebnisse, bei Wiederholung liegen also gleiche Resultate vor. Statt Reliabilität spricht man auch von “Zuverlässligkeit”.

Validität

Der Merkmal in seinen Ausprägungen spiegelt die für die Fragestellung wesentlichen Eigenschaften wider. Auch “Gültigkeit” oder “Aussagekraft” genannt.

Charakterisierung von Merkmalen

quantitative Merkmale:

Untersuchungseinheiten unterscheiden sich im absoluten (Zahlen-) Wert. - z.B. Alter, Gewicht, Temperatur, Anzahl Keime, Betriebsgröße, Schadstoffgehalt, ...

qualitative Merkmale:

Untersuchungseinheiten unterscheiden sich in ihrer Ausprägung (Art) - z.B. Geschlecht, Name, Befund, Rasse, Therapie, Haltungsform, Region, ...

Skalenniveaus von Merkmalen

nominale (qualitative) Skala:

die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar.

- z.B. Name, Geschlecht, Rasse, Haltungsform, Therapieform, pathologische Klassifikation

ordinale (qualitativ oder semiquantitative) Skala:

die Werte unterliegen einer Rangfolge, aber die Abstände zwischen den Werten der Skala lassen sich nicht interpretieren.

- z.B. Bewertung (Bonituren, Noten), Gesundheitszustand, Grad der Belastung mit Keimen (-, +, ++, +++)

metrische (quantitative) Skala:

die Werte unterliegen einer Rangfolge und die Abstände zwischen den Werten der Skala lassen sich interpretieren.

- z.B. Gewicht, Betriebsgröße, Keimzahlen, ...

Skalenniveaus von Merkmalen (1)

Es wird auch unterschieden zwischen

Intervallskala

Die Abstände zwischen Merkmalsausprägungen lassen sich vergleichen.
Die Skalla ist kontinuierlich.

- z.B. Temperatur in Grad Celsius

Verhältnisskala

Nicht nur die Differenz, sondern auch der Quotient aus zwei Messwerten darf verwendet werden.

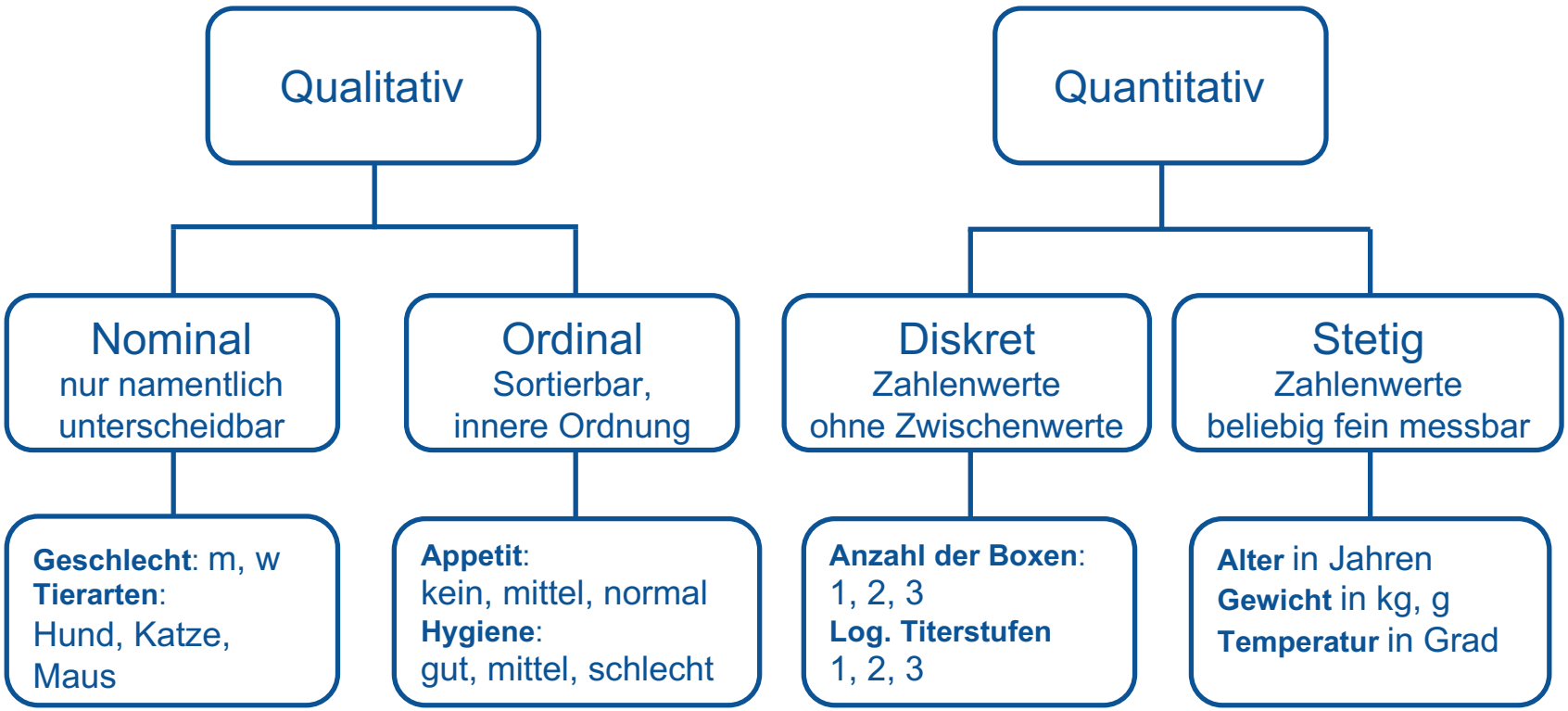
- z.B. Temperatur in Kelvin, Länge in Zentimetern

Skalenniveaus von Merkmalen (2)

Die statistischen Auswertungsmöglichkeiten sind vom Skalenniveau abhängig, weil auf höherem Niveau mehr Information festgehalten und ausgewertet werden kann, als bei niedrigeren Skalierungen.

Debei soll den Aufwand für den zusätlichen Informationgewinn berücksichtigt werden.

Skalenniveaus von Merkmalen (3)



Nicht jede Zahl ist eine Zahl.

Häufig werden Daten verschlüsselt, um die anschliessende Datenverarbeitung zu erleichtern

- Schulnoten: 1, 2, 3, 4, 5, 6 (ordinal)
- Testergebnis: 1, 0 (nominal)
- Kreiskennziffern: 3253, 3351 (nominal)
- Zuchtbuch-Nummern: 0511572 (nominal)

Descriptive Statistics

Deskriptive (beschreibende) Statistik(1)

Tabellen

Graphiken

Charakterischtische Maßzahlen

Tabellen

Table 1: Titanic dataset

Index	survived	pclass	sex	age	deck	fare	alone
0	0	3	male	22		7.2500	False
1	1	1	female	38	C	71.2833	False
2	1	3	female	26		7.9250	True
3	1	1	female	35	C	53.1000	False
4	0	3	male	35		8.0500	True
5	0	3	male	NA		8.4583	True
6	0	1	male	54	E	51.8625	True
7	0	3	male	2		21.0750	False
8	1	3	female	27		11.1333	False
9	1	2	female	14		30.0708	False

Urliste

Die ungeordnete Form von Messungen (Beobachtungen) einer Untersuchung, die der Reihe nach zusammengestellt ist.

Table 2: Urliste

	1	2	3	4	5	6	7	8	9	10
age	22	38	26	35	35	NA	54	2	27	14

->

Häufigkeitstabelle

Table 3: Häufigkeitstabelle

age	freq
0.42	1
0.67	1
0.75	2
0.83	2
0.92	1
1	7
2	10
3	6
4	10
5	4
6	3
7	3
8	4
9	8
10	2
11	4
12	1
13	2
14	6
14.5	1
15	5
16	17
17	13
18	26
19	25
20	15
20.5	1
21	24
22	27
23	15

Graphiken

Balkendiagramm

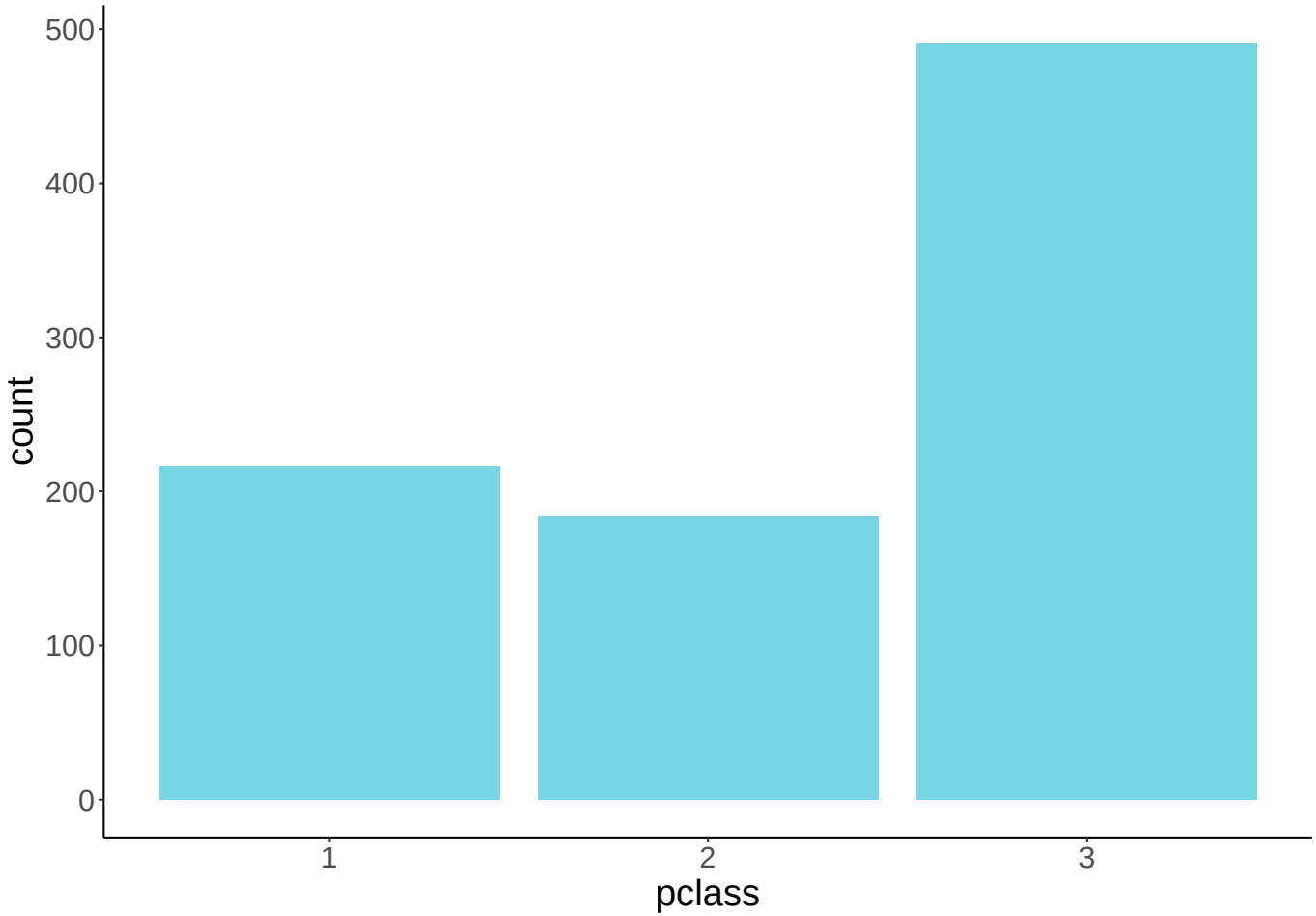


Figure 1: Balkendiagramm

Balkendiagramm (relative Häufigkeit)

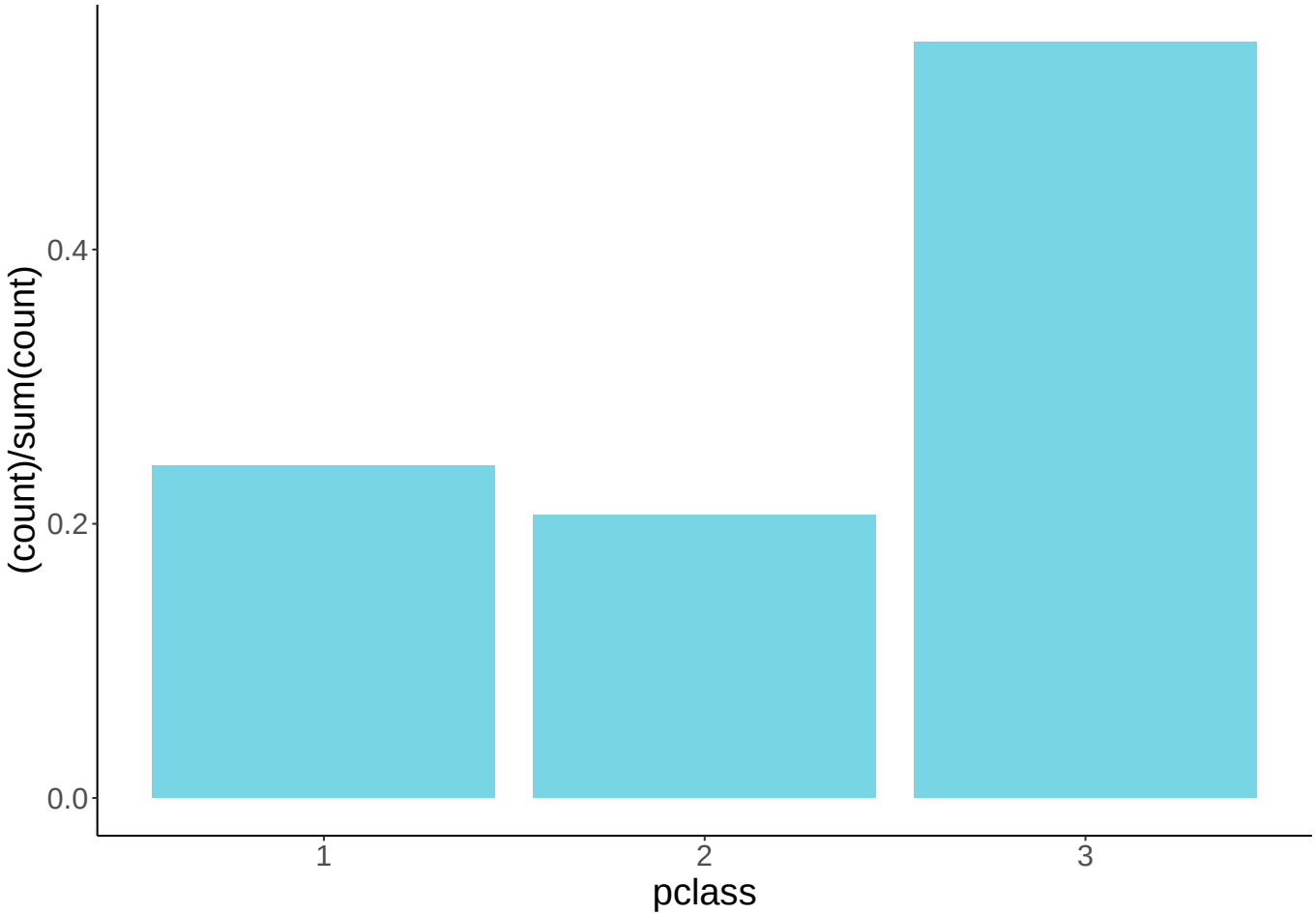


Figure 2: Balkendiagramm (relative Häufigkeit)

Komponenten-Balkendiagramm

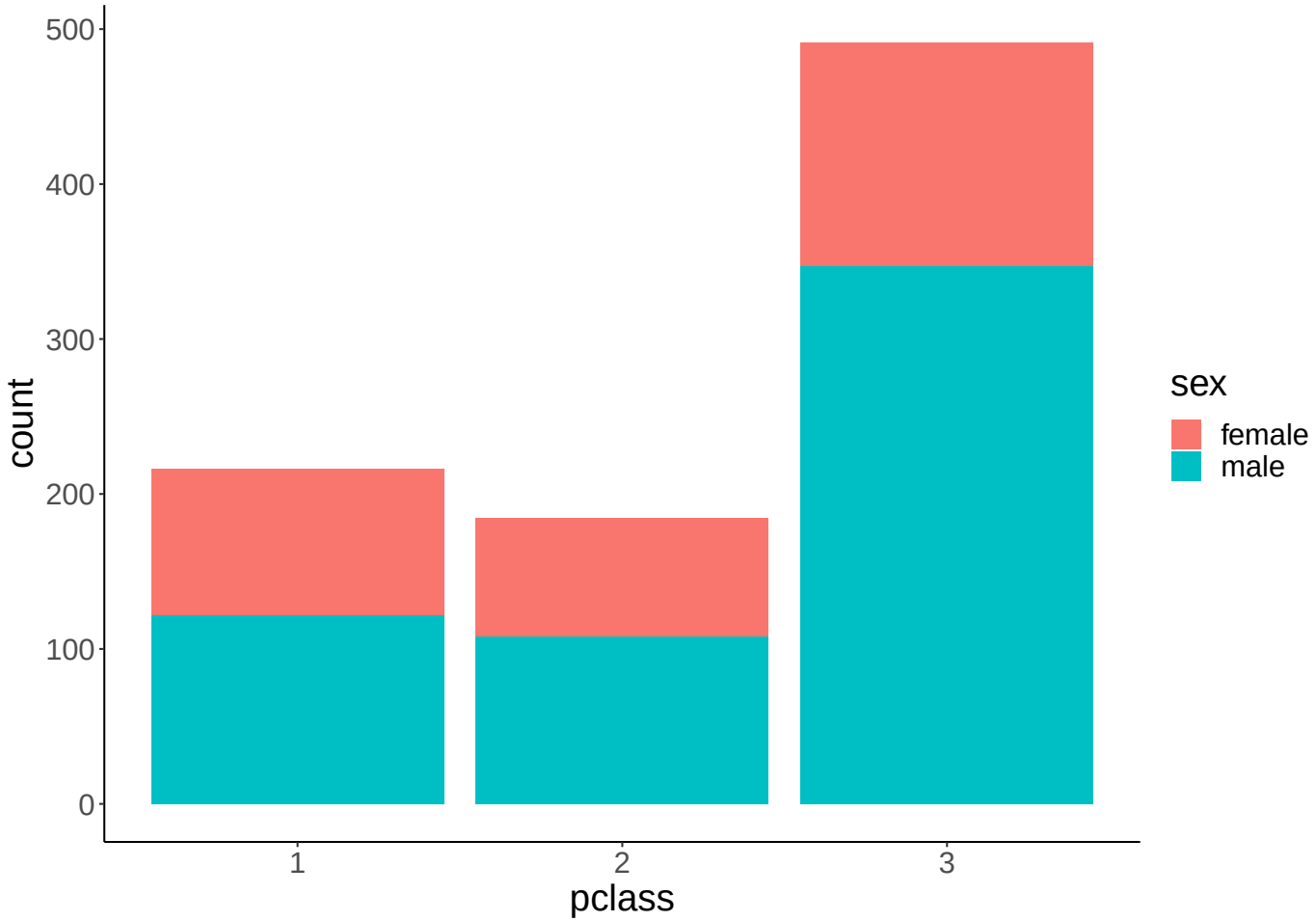


Figure 3: Komponenten-Balkendiagramm

Komponenten-Balkendiagramm (relative Häufigkeit)

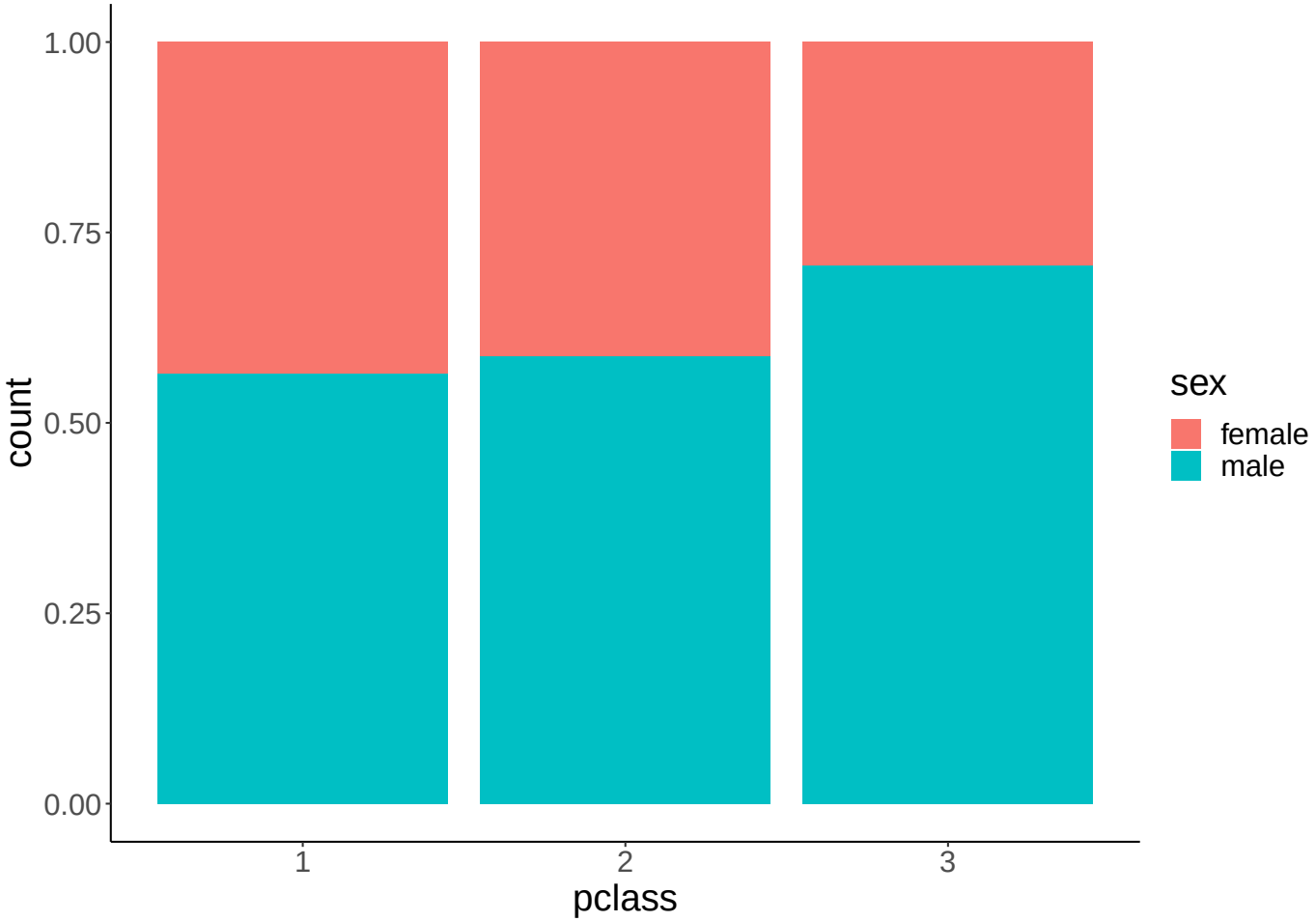


Figure 4: Komponenten-Balkendiagramm (relative Häufigkeit)

Komponenten-Balkendiagramm (relative Häufigkeit) (1)

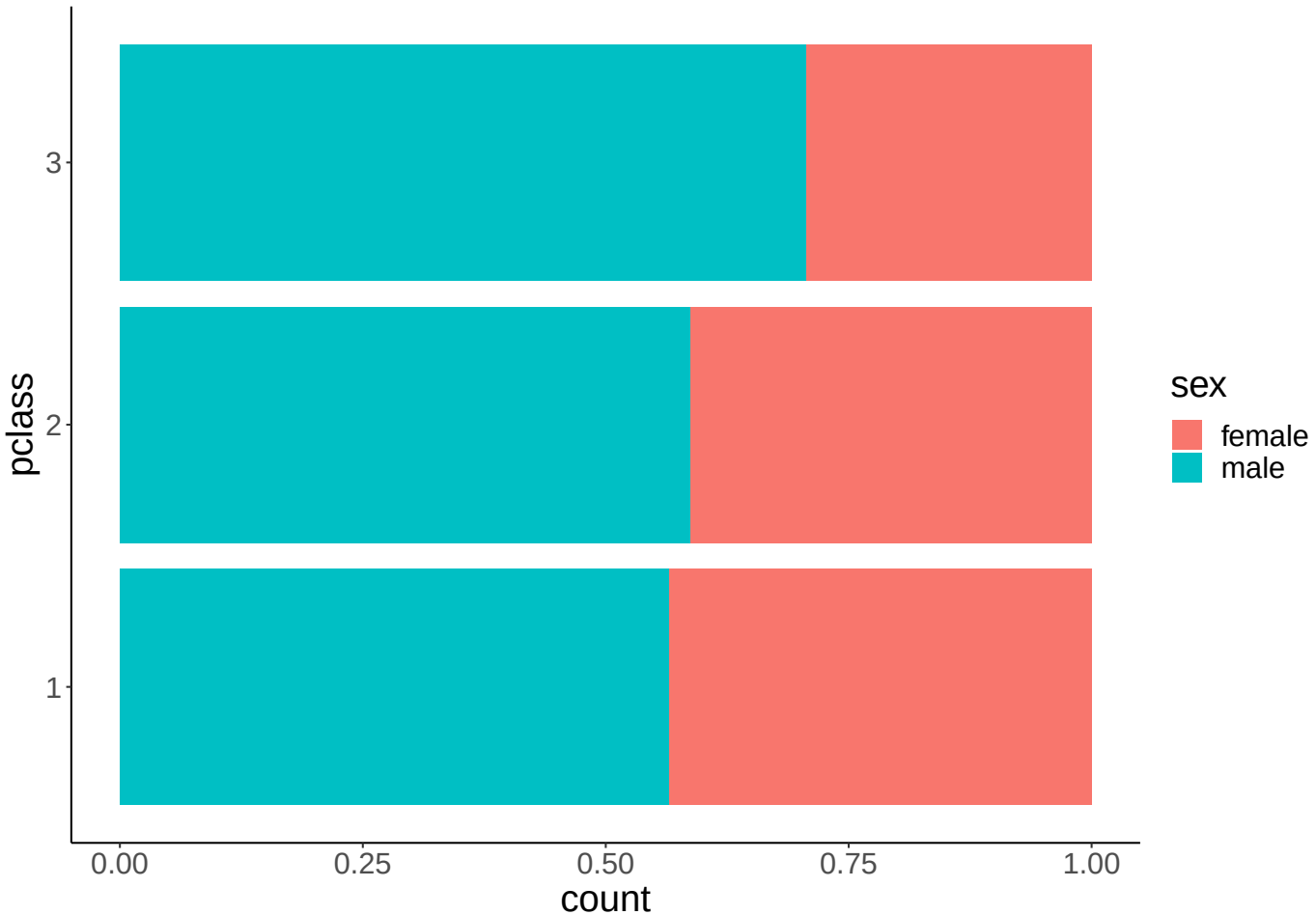


Figure 5: Komponenten-Balkendiagramm (relative Häufigkeit)

Daten auf der Zahlengerade

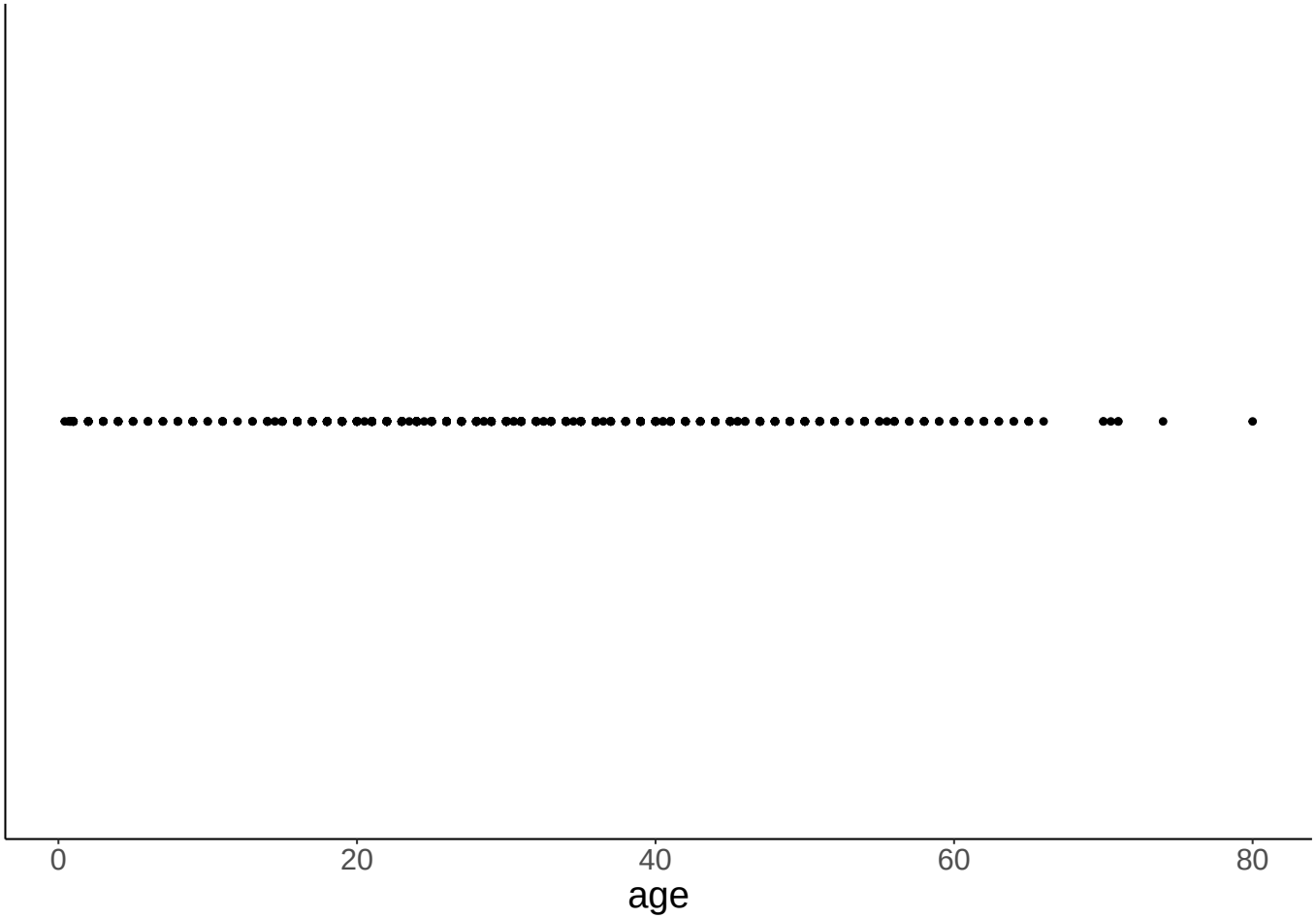


Figure 6: Daten auf der Zahlengerade

Histogramm

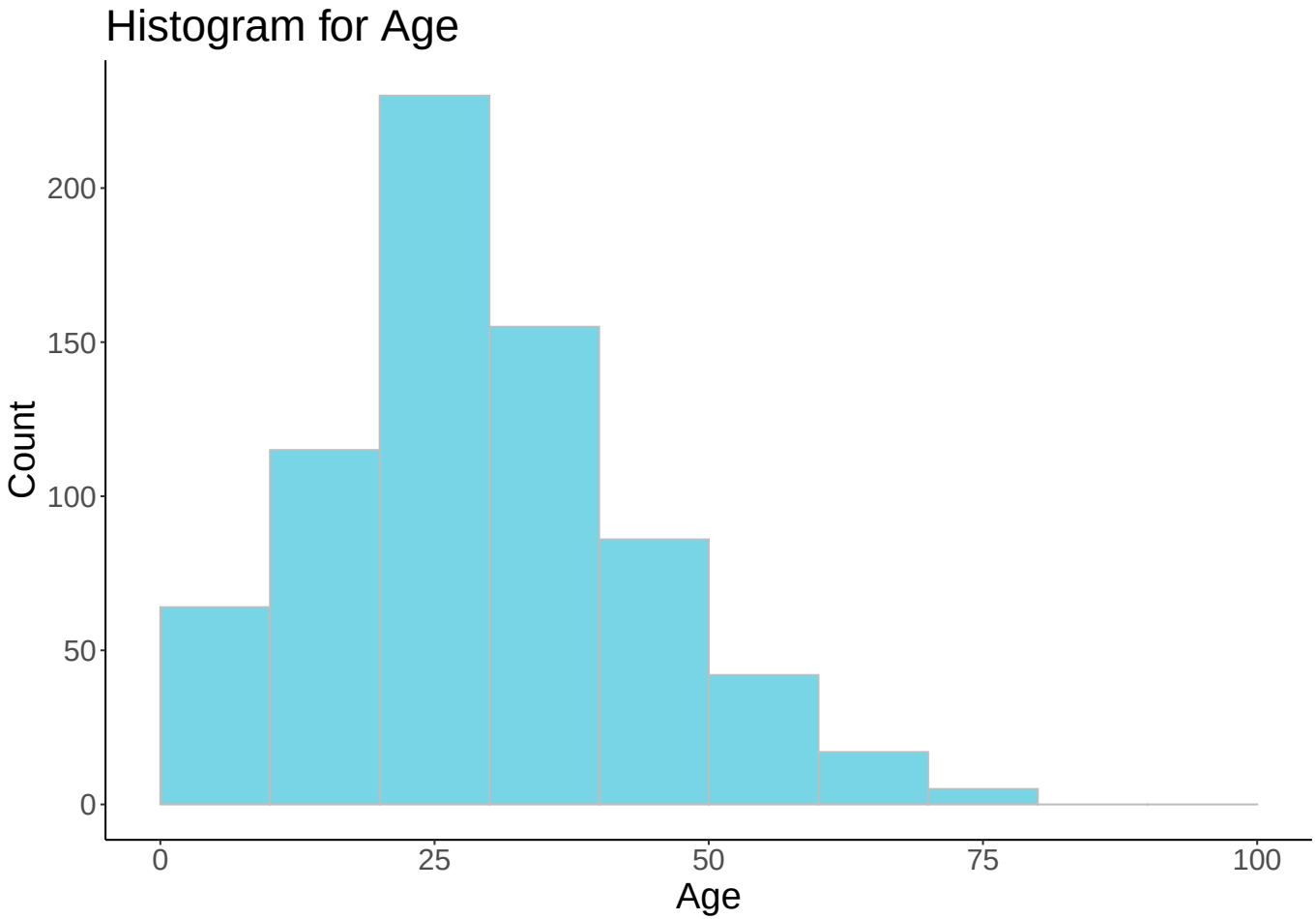


Figure 7: Histogramm

Histogramm (relative Häufigkeit)

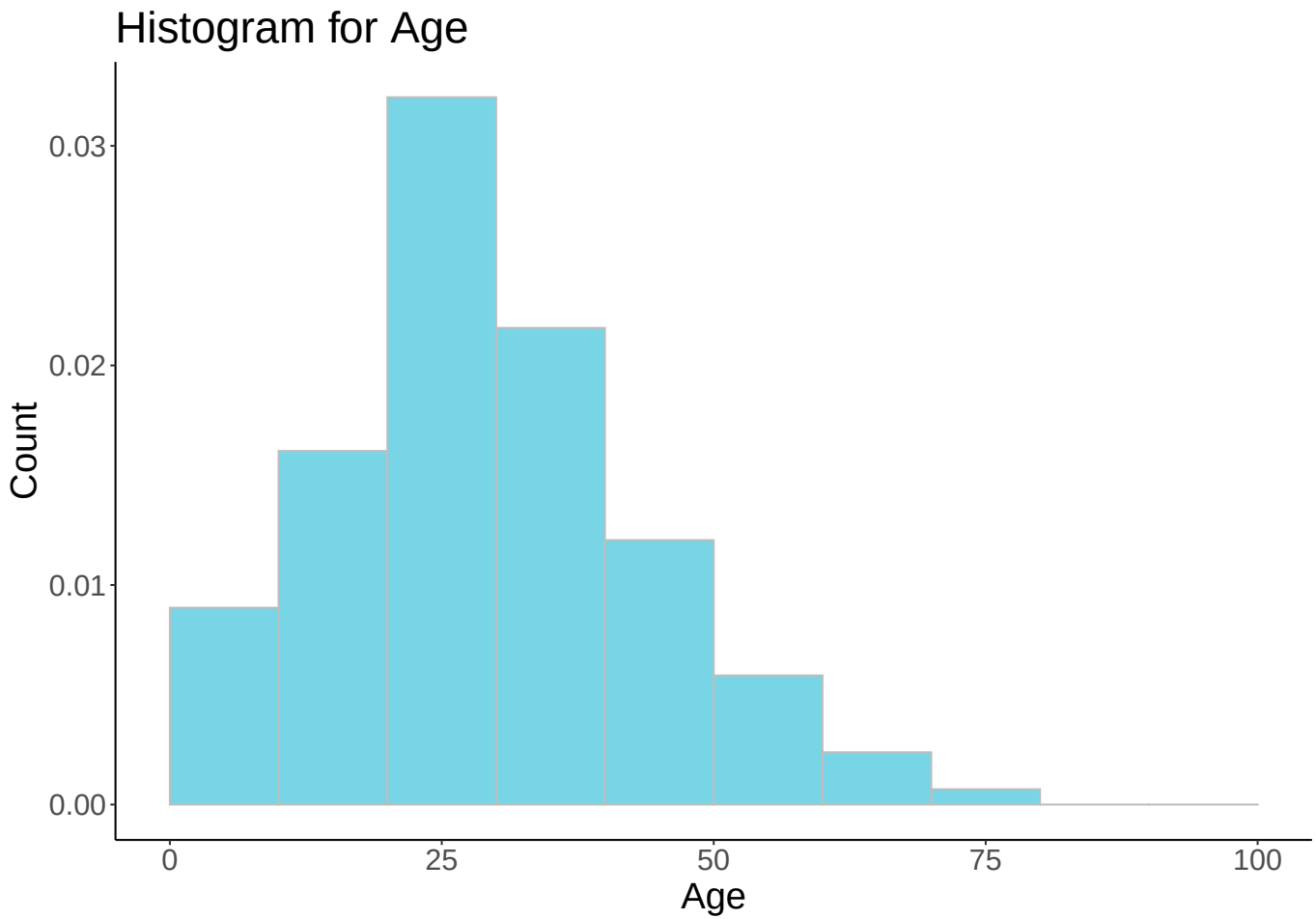


Figure 8: Histogramm (relative Häufigkeit)

Histogramm (relative Häufigkeit) (2)

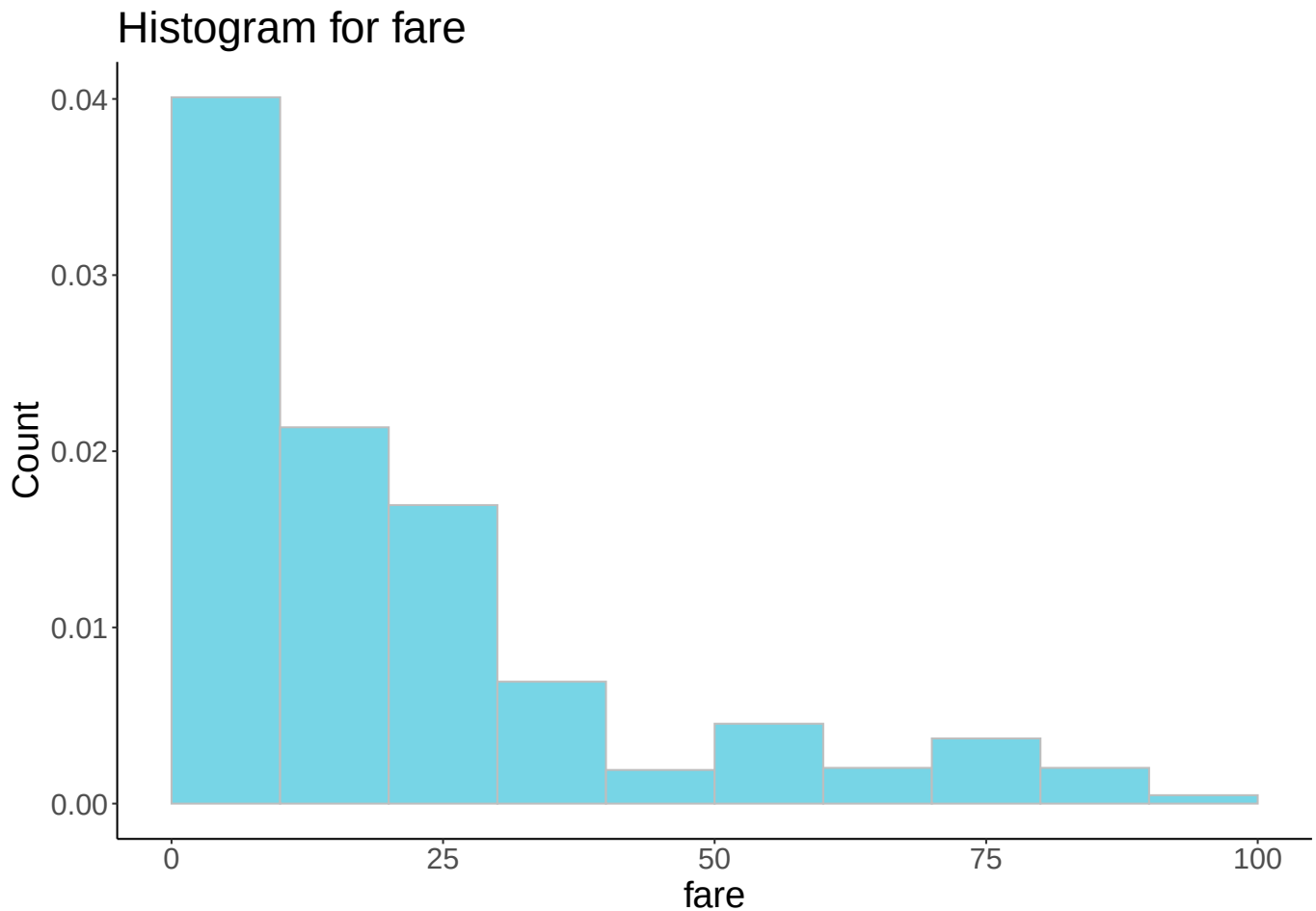


Figure 9: Histogram fare

Gruppenbildung

Klassenbreite b nach von Sturges

$$b = \frac{V}{1 + 3.32 \cdot \lg n} \approx \frac{V}{5 \lg n}$$

n - der Stichprobenumfang (Anzahl der Messwerte)

V - die Variationsbreite (Spannweite)

$\lg n$ - Zehnerlogarithmus von n

Histogramm (relative Häufigkeit) (3)

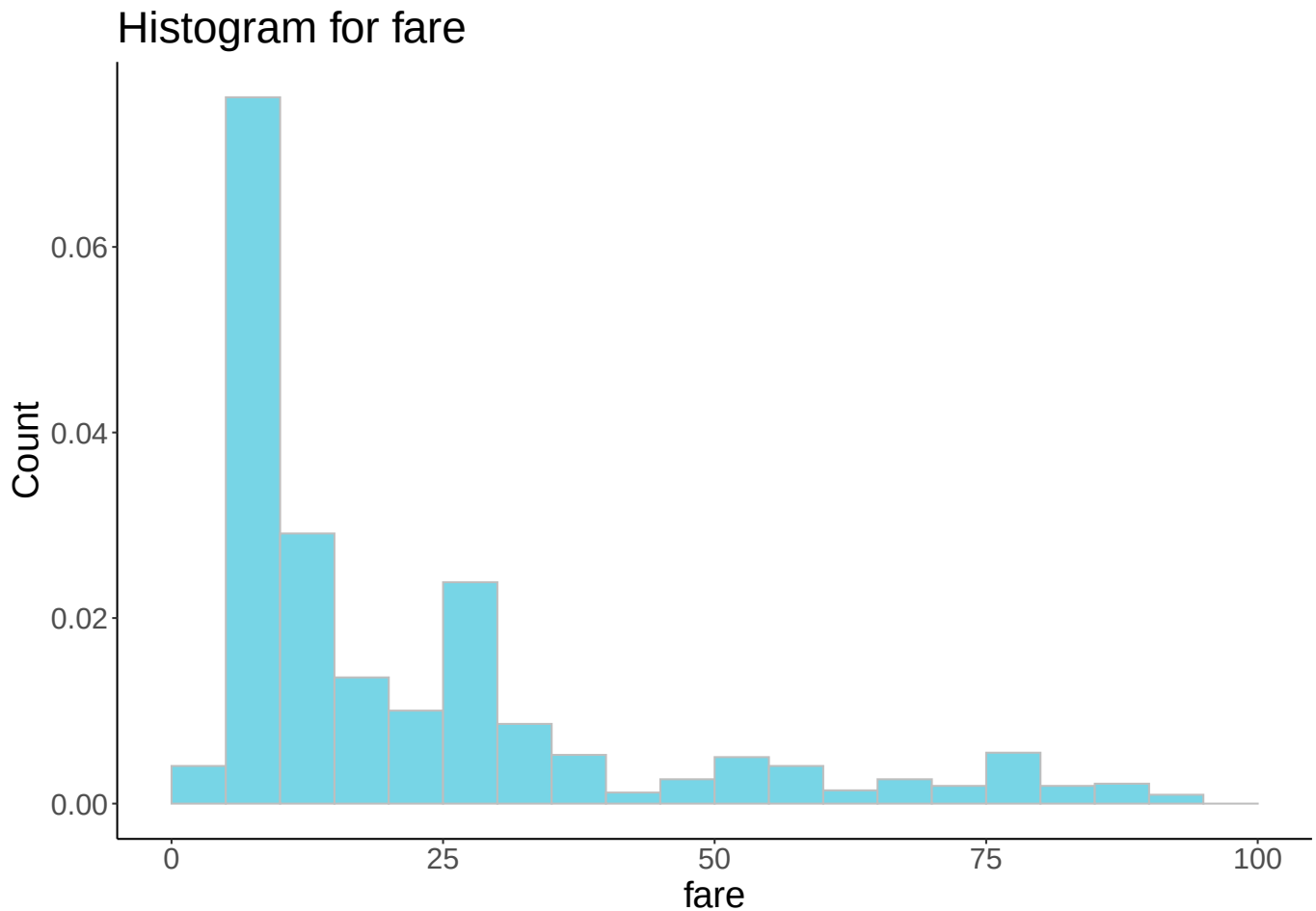


Figure 10: Histogram fare

Maßzahlen

Maßzahlen

Lageparameter

Streuungmaße

Lageparameter

Variationsbreite (Range)

(arithmetischer) Mittelwert

Median

Modus oder Modaler Wert

geometrischer Mittelwert

harmonischer Mittelwert

Streuungsmaße

Varianz und Standardabweichung

Quantile

Variationskoeffizient

Variationsbreite (Range)

$$V_x = \max(x) - \min(x)$$

(arithmetischer) Mittelwert

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Nutzliche Eigenschaft

$$\sum_{i=1}^{N} cx_i = c \left(\sum_{i=1}^{N} x_i \right)$$

Der Mittelwert ist sehr empfindlich was extreme Werte betrifft

Median

Zentralwert

Ungerade Anzahl der Beobachtungen

$\left(\frac{n+1}{2}\right).$ Beobachtungswert

Gerade Anzahl der Beobachtungen

Mittelwert von $\left(\frac{n}{2}\right).$ und $\left(\frac{n}{2}+1\right).$ Beobachtungswerten

Der Median ist hauptsächlich bestimmt durch die Werte in der Mitte der Stichprobe und ist wenige abhängig von den extremen Werten

Modus oder Modaler Wert

Dichtemittel

Der häufigste Wert. Wenn alle Werte nur einmal vorkommen, gibt es keinen Modus.

Geometrischer Mittelwert

$$G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln x_i}$$

wird z.B. für die Bestimmung der MIC benutz ($2^k c$, $k = 1, 2, \ldots$)

Harmonischer Mittelwert

$$H = \frac{1}{n}\left(\frac{1}{x_1} + \ldots + \frac{1}{x_n}\right)^{-1}$$

Vergleich vom Maßzahlen

Histogramm (relative Häufigkeit) (3)

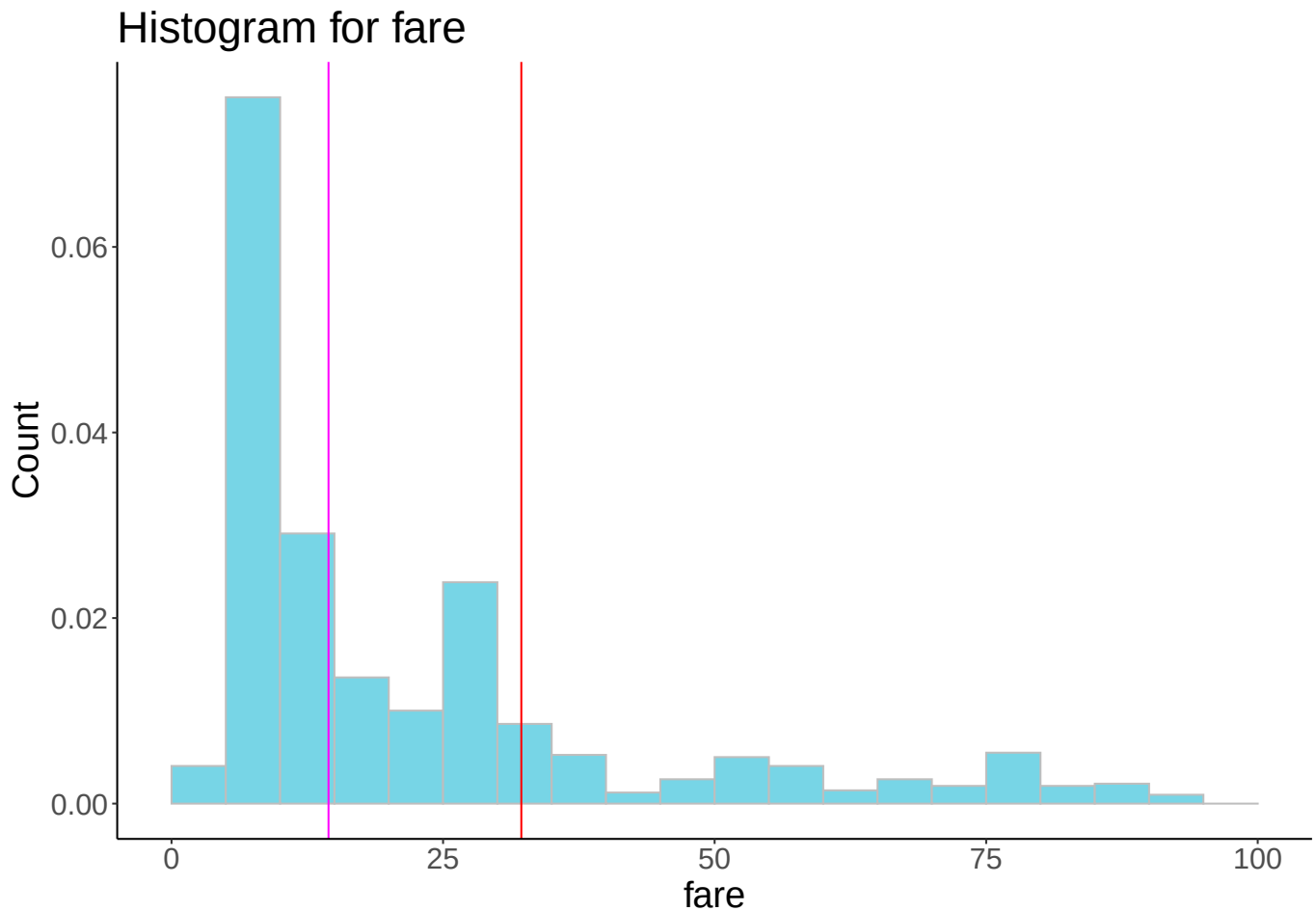


Figure 11: Histogram fare

Streuungsmaße

Varianz und Standardabweichung

Varianz

$$s_x^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2$$

$n - 1$ - Freiheitsgrad

Standardabweichung

$$s_x = \sqrt{\frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2}$$

Der mittlere Fehler des Mittelwertes

Streuung von $\bar{x}$ um den wahren Mittelwert μ der Grundgesamtheit

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Variationskoeffizient

Das Verhältnis der Standardabweichung zum Mittelwert

$$\mathrm{CV} = \frac{s}{|\bar{x}|}$$

Erlaubt unabhängig vom Mittelwert die Streuung der Daten zu Vergleichen

Quantile

p. Perzentil

- $(k+1)$. Datenpunkt wenn $\frac{np}{100}$ nicht ganzzahlig ist $(k < \frac{np}{100}, k \in \mathbb{N})$
- Durchschnitt von k. und $(k+1)$. Datenpunkt wenn $\frac{np}{100}$ ganzzahlig ist

1. Quantil (Q_1) - Der Datenpunkt wo 25% Messpunkte unterhalb und 75% oberhalb liegen
2. Quantil (Q_3) - Der Datenpunkt wo 75% Messpunkte unterhalb und 25% oberhalb liegen
3. Quantil (Q_2) - Median

Interquartilsabstand und MAD

(Inter)quartilsabstand

$$IQR = Q_{0.75} - Q_{0.25}$$

Median-Abweichung (Mean Abslolute Deviation)

$$|x_i - Q_{0.5}|$$

Kumulative Verteilung

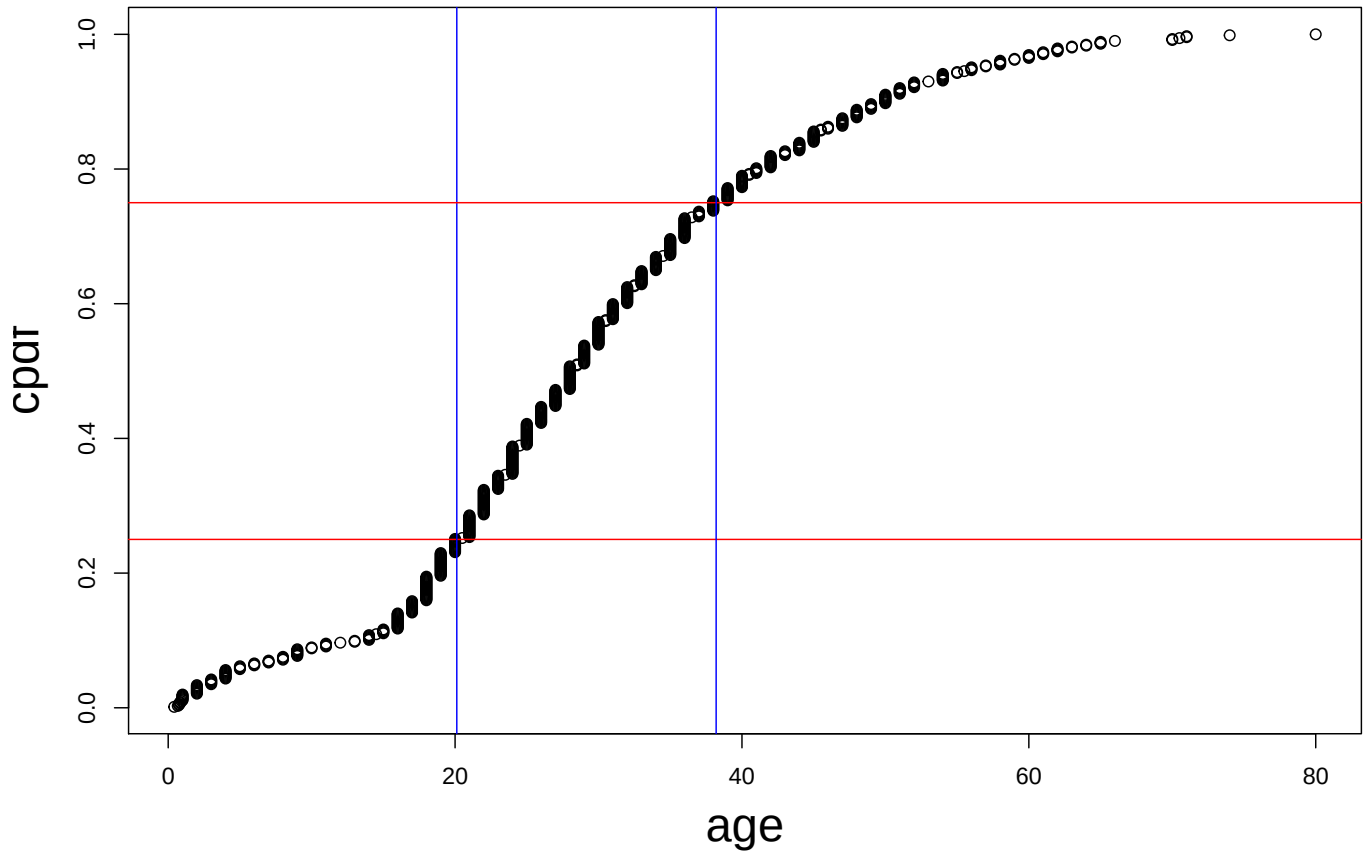
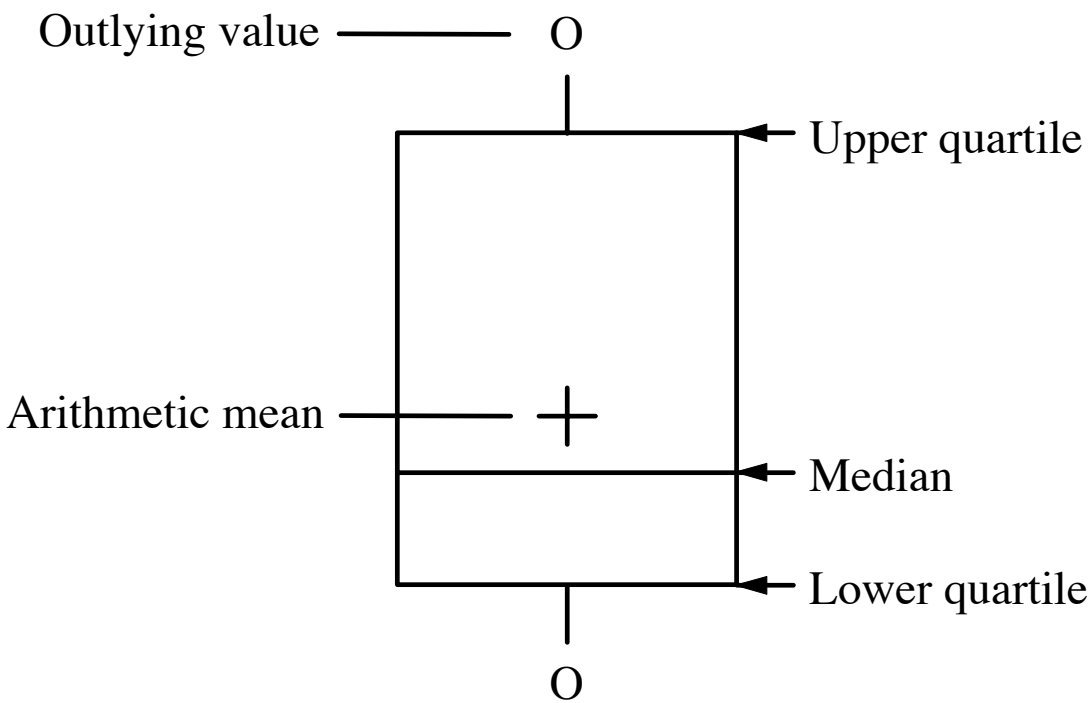


Figure 12: Kumulative Verteilung

Boxplot


Outlying value
O
Upper quartile
Arithmetic mean
Median
Lower quartile
O

Boxplot

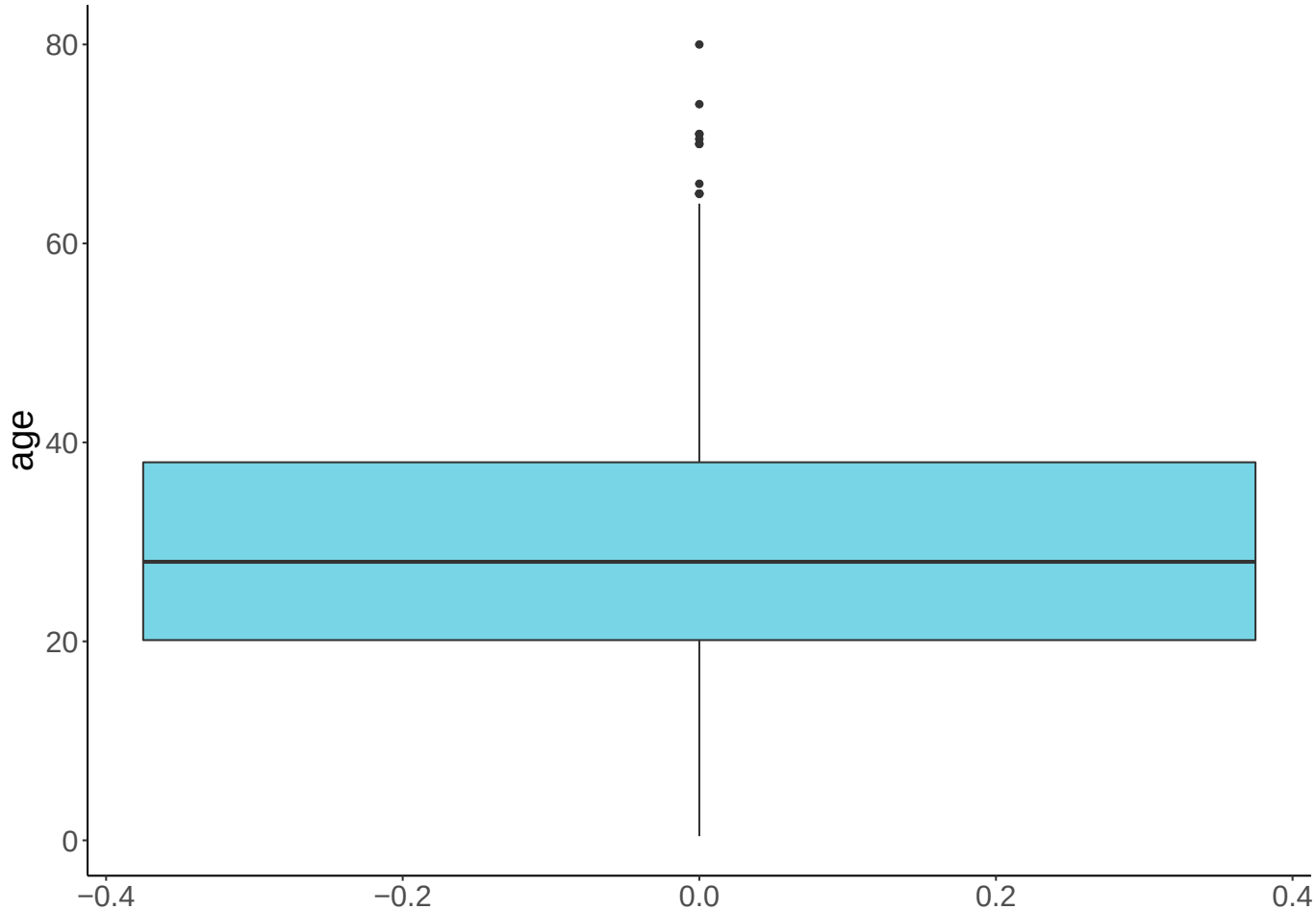


Figure 13: Boxplot

Boxplot

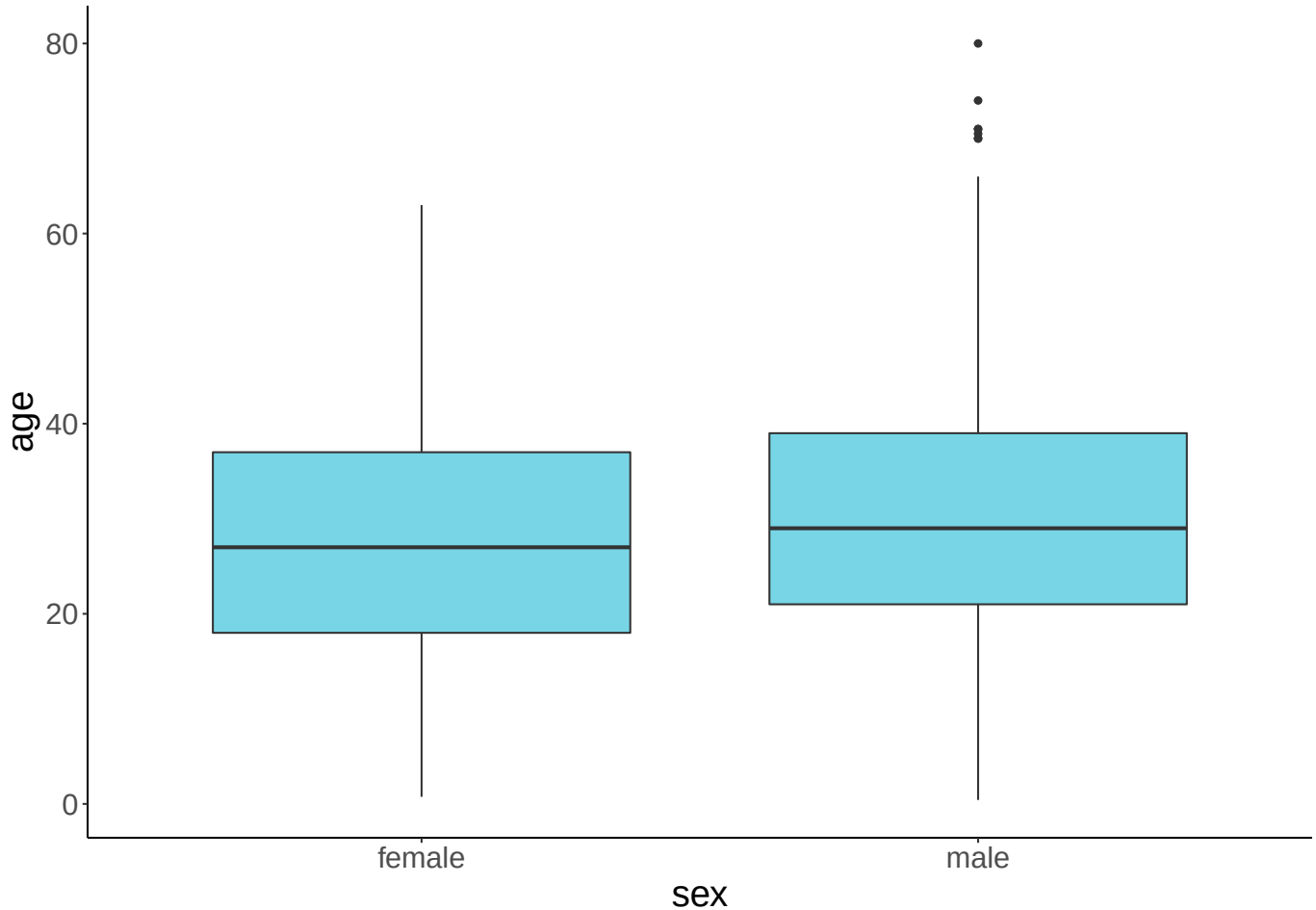


Figure 14: Boxplot

Diversitätsindex (Entropy)

$$H = \sum_i p_i \log p_i$$

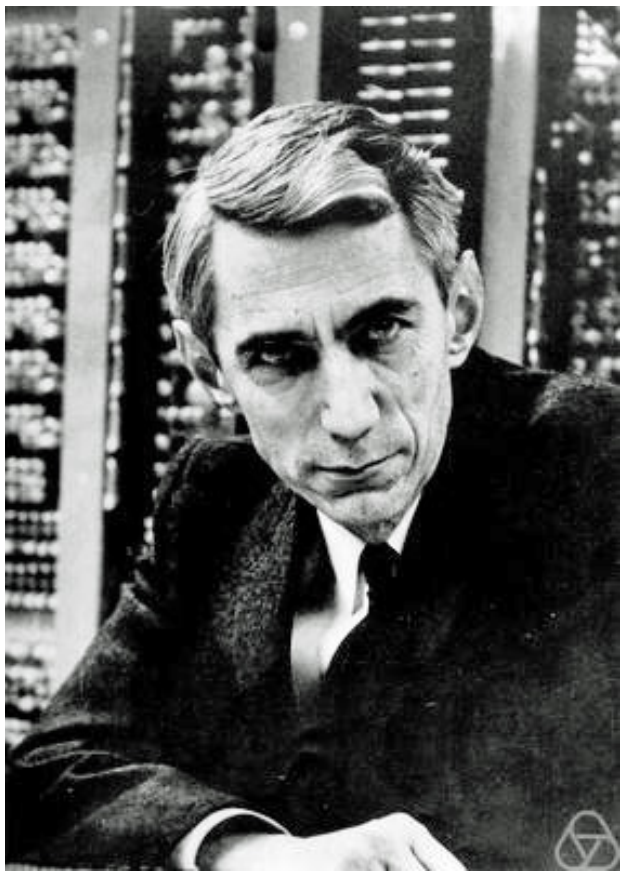
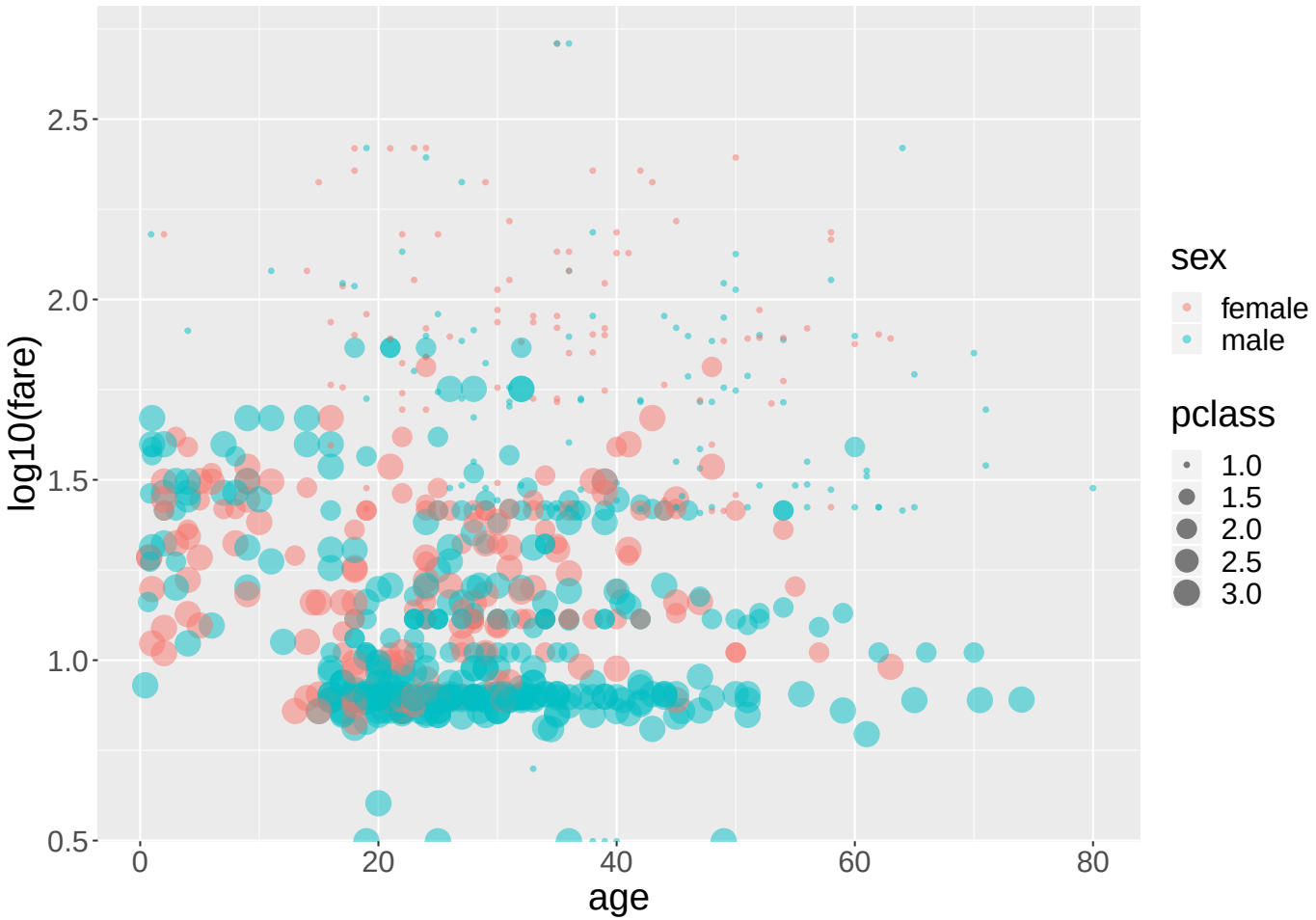


Figure 15: Claude Shannon (1910 — 2001)

Zusammenhänge in den Daten

Bivariate Daten



Zusammenhänge in kontinuierlichen Daten

Zusammenhänge in kontinuierlichen Daten

- Pearson-Korrelationskoeffizient
- Spearman-Rangkorrelationskoeffizient

Pearson-Korrelationskoeffizient

Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Covariance

$$\mathrm{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Pearson-Korrelationskoeffizient

$$r = \frac{\operatorname{cov}(x, y)}{s_x s_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1) s_x s_y}$$

Korregierter z-score

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Sein Standardfehler

$$SE_{z_r} = \frac{1}{\sqrt{n-3}}$$

Konfidenzintervalle

$$z_r \pm (1.96 \times SE_{z_r}), \text{ Transformation: } r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

Korrelationskoeffizient (Beispiel)

```
#fig.show = 'hide'
#results='hide'
library(ggplot2)
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/4_Statistische_Tests_and_Regression/Data/Data\ files/Album\ Sales\ 1.dat', s
cor(df)
```

```
##           adverts     sales
## adverts 1.0000000 0.5784877
## sales   0.5784877 1.0000000
```

```
cor.test(df$adverts,df$sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$adverts and df$sales
## t = 9.9793, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4781207 0.6639409
## sample estimates:
##       cor
## 0.5784877
```

Spearman-Korrelationskoeffizient (Beispiel)

```r
cor(df, method = "spearman")
```

```
##           adverts     sales
## adverts 1.0000000 0.5541557
## sales   0.5541557 1.0000000
```

```r
cor.test(df$adverts, df$sales, alternative = "greater", method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$adverts and df$sales
## S = 594444, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.5541557
```

Kendall-tau (Beispiel)

```r
cor(df, method = "kendall")
```

```
##           adverts     sales
## adverts 1.0000000 0.3985301
## sales   0.3985301 1.0000000
```

```r
cor.test(df$adverts, df$sales, alternative = "greater", method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$adverts and df$sales
## z = 8.2362, p-value < 2.2e-16
## alternative hypothesis: true tau is greater than 0
## sample estimates:
##       tau
## 0.3985301
```

Regression

Regression

$$Y = f(X_1, ..., X_n)$$

Regression ist ein weit gefasster Begriff für eine Reihe von Methoden, die zur Vorhersage einer *Antwort*-Variable Y (oder einer *abhängigen, resultierenden*) aus einer oder mehreren *Prädiktor*-Variablen X_i (*unabhängigen, erklärenden*) verwendet werden.

- Im Falle von *einer* Prädiktor-Variable spricht man von *einfacher* Regression
- Im Falle von *mehreren* Prädiktor-Variable spricht man von *multipler* oder *mehrfacher* Regression

Ziele der Regression

- *Bestimmung* der erklärenden Variablen, die sich auf die Antwortvariable beziehen
- *Beschreibung* der Form der Beziehung
- *Bereitstellen* einer Gleichung für die *Vorhersage* der Antwortvariablen aus den erklärenden Variablen

Aus der Regression konnte keine *Kausalität* direkt abgeleitet werden!

Regression, Beispiel (1)

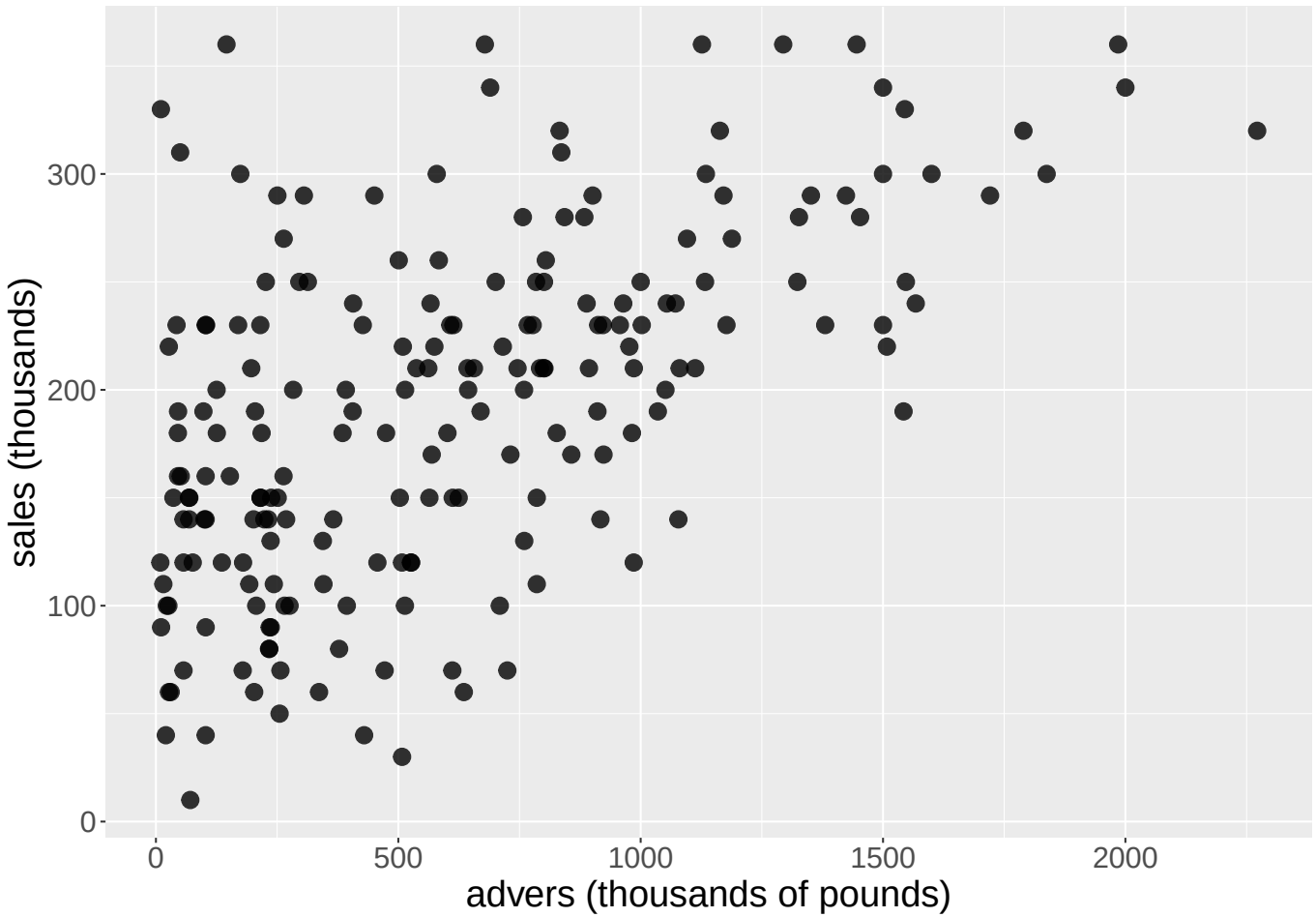


Figure 16: Auswirkung des Geldbetrags, der für Werbung im Albumverkauf ausgegeben wird. [A. Field et al.]

Regression, Beispiel (2)

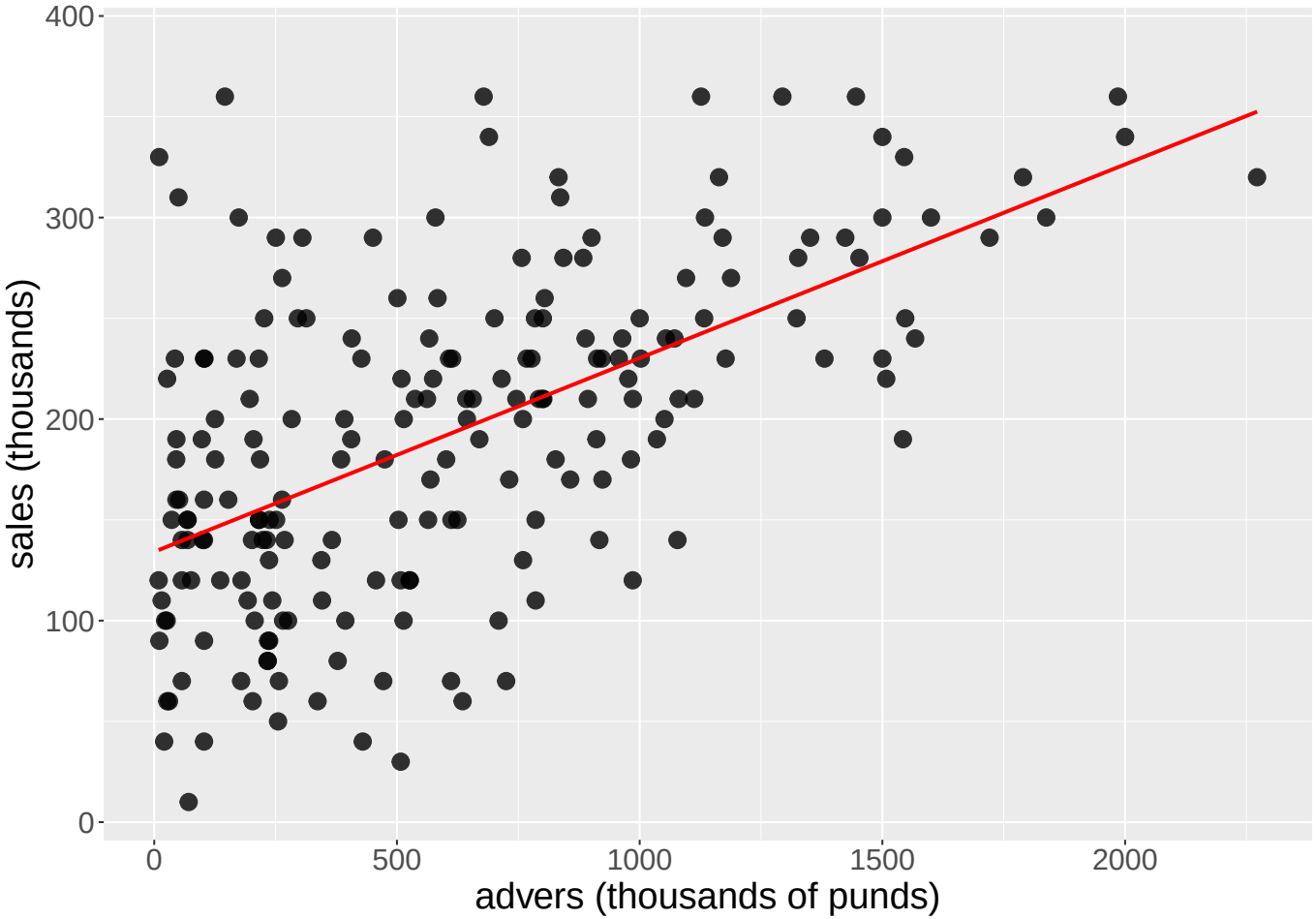


Figure 17: Auswirkung des Geldbetrags, der für Werbung im Albumverkauf ausgegeben wird. [A. Field et al.]

Regression, Beispiel (3)

```
library(ggplot2)
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/4_Statistische_Tests_and_Regression/Data/Data\ files/Album\ Sales\ 1.dat', s
ggplot(df, aes(x = adverts, y = sales)) +
theme(text = element_text(size = 20)) +
geom_point(alpha = 0.8, size = 4) +
labs(x = "advers (thousands of punds)", y = "sales (thousands)") +
geom_smooth(fill=NA,color="red",size=1,method="lm")
```

Regression, Beispiel: Modell-Ausgabe

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(sales~adverts, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ adverts, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Regression, Beispiel: ANOVA-Tabelle

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## adverts     1 433688  433688  99.587 < 2.2e-16 ***
## Residuals 198 862264    4355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA-Tabelle

$$\hat{Y} = \beta_0 + \beta_1 X$$

Um zu beurteilen, wie viel Information die Variable X über die Variable Y enthält, betrachten wir, wie viel von der *Summe der Quadrate* (SS) der Variablen Y durch die Kenntnis der Variablen X erklärt werden könnte.

Table 4: Zerlegung von Quadratsummen im Regressionsmodell mit k Prädiktorvariablen [Dohoo, p. 327]

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Model	$SSM = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$dfM = k$	$MSM = \frac{SSM}{dfM}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$dE = n - (k+1)$	$MSE = \frac{SSE}{dfE}$	
Total	$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$	$dfT = n - 1$	$MST = \frac{SST}{dfT}$	

$MSE = \sigma^2$ and σ is called *standard error of prediction*.

F-Test zur Beurteilung der Modellgüte

H_0: $\beta_1 = \beta_2 ... \beta_k = 0$, $\beta_0 \neq 0$

H_1: zumindest manche von Koeffizienten $\beta_i \neq 0$, $i \neq 0$

- Variablen werden so asugewählt, dass die F-Statistik maximiert wird
- F-Test hat eine einfache Bedeutung, wenn unabhängige Variablen in einem kontrollierten Experiment manipuliert werden
- Vorsicht wird bei Beobachtungsvariablen geboten (beeinflusst von der Anzahl der Variablen, ihren Korrelationen und der Stichprobengröße)

t-test (mit $n - (k + 1)$) Freiheitsgraden (1)

H_0: $\beta_i = \beta^* \; e.g. \beta^* = 0$

H_1: $\beta_i \neq \beta^* \; e.g. \beta^* = 0$

$$t = \frac{\beta_i - \beta^*}{SE(\beta_i)}$$

$SE(\beta_i)$ ist Standardfehler vom geschätzen Koeffizienten (*standard error of the esimated coefficient*)

Im Falle eines einzelnen Prädiktors

$$SE(\beta_1) = \sqrt{\frac{MSE}{SSX(\beta_1)}},$$

wobei $SSX_1 = \sum_{i=1}^{n} \left(X_{1,i} - \bar{X_1}\right)^2$ ist Summe der Quadrate von Variablen X_1 (*sum of square of the variable X_1*).

t-test (mit $n - (k + 1)$) Freiheitsgraden (2)

- Variablen werden so ausgewählt, dass die t-Statistik maximiert wird und ihre Signifikanz gewährleistet ist
- t -Test hat nur dann eine einfache Bedeutung, wenn unabhängige Variablen in einem kontrollierten Experiment manipuliert werden
- Vorsicht ist geboten, wenn Beobachtungsvariablen angewendet werden (beeinflusst von der Anzahl der Variablen, ihren Korrelationen und der Stichprobengröße).

Intervalle für die Vorhersage (*prediction interval*) (1)

Zwei Ursachen für Abweichungen von Vorhersageschätzungen:

- aus der Schätzung der Regressionsparameter (SE)
- aus der mit einer neuen Beobachtung x^* verbundenen Unsicherheit (Variation über die Regressionsgleichung für den Mittelwert)

Standardfehler (*Error of the mean*)

Für einfache Regression (einzelner Prädiktor) der *Fehler des Mittelwerts* einer großen Anzahl neuer Beobachtungen

$$SE_{\text{mean}}(Y|x^*) = \sigma\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

Intervalle für die Vorhersage (1)

Standardfehler aus einer neuen Beobachtung

Für einfache Regression (einzelner Prädiktor)

$$SE_{\text{obs}}(Y|x^*) = \sigma\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

95% Konfidenzintervall (*Confidence interval*)

$$95\%CI = Y \pm t_{0.05}(SE)$$

Bestimmtheitsmaß (*coefficient of determination*), R^2

- R^2 beschreibt den Betrag der Varianz in der Ergebnisvariablen, der durch die Prädiktorvariablen erklärt wird.
- R^2 ist ein quadratischer *Korrelationskoeffizient* zwischen den vorhergesagten und beobachteten Y -Werten

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

Angepasstes R^2

Da R^2 immer mit der Anzahl der Variablen zunimmt, wird *angepasstes* R^2 verwendet

$$R^2_{\text{adjusted}} = 1 - \frac{MSE}{MST}$$

Akaike Information Criterion (AIC)

Bei der Schätzung des Informationsverlusts eines Modells berücksichtigt AIC den Kompromiss zwischen der Anpassungsgüte des Modells und der Einfachheit des Modells. Mit anderen Worten, AIC befasst sich sowohl mit dem Risiko einer Überanpassung als auch mit dem Risiko einer Unteranpassung.

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2k$$

n - Anzahl von Beobachtungen

k - Anzahl von Prädiktor-Variablen

- Man sollte ein Modell mit kleinerem AIC bevorzugen

```
model1 <- lm(milk120~cc, data = daisy2)
AIC(model1)
```

```
## [1] 112227.5
```

```
model2 <- lm(milk120~parity, data = daisy2)
AIC(model2)
```

```
## [1] 104745.6
```

Einfache Regression, Beispiel

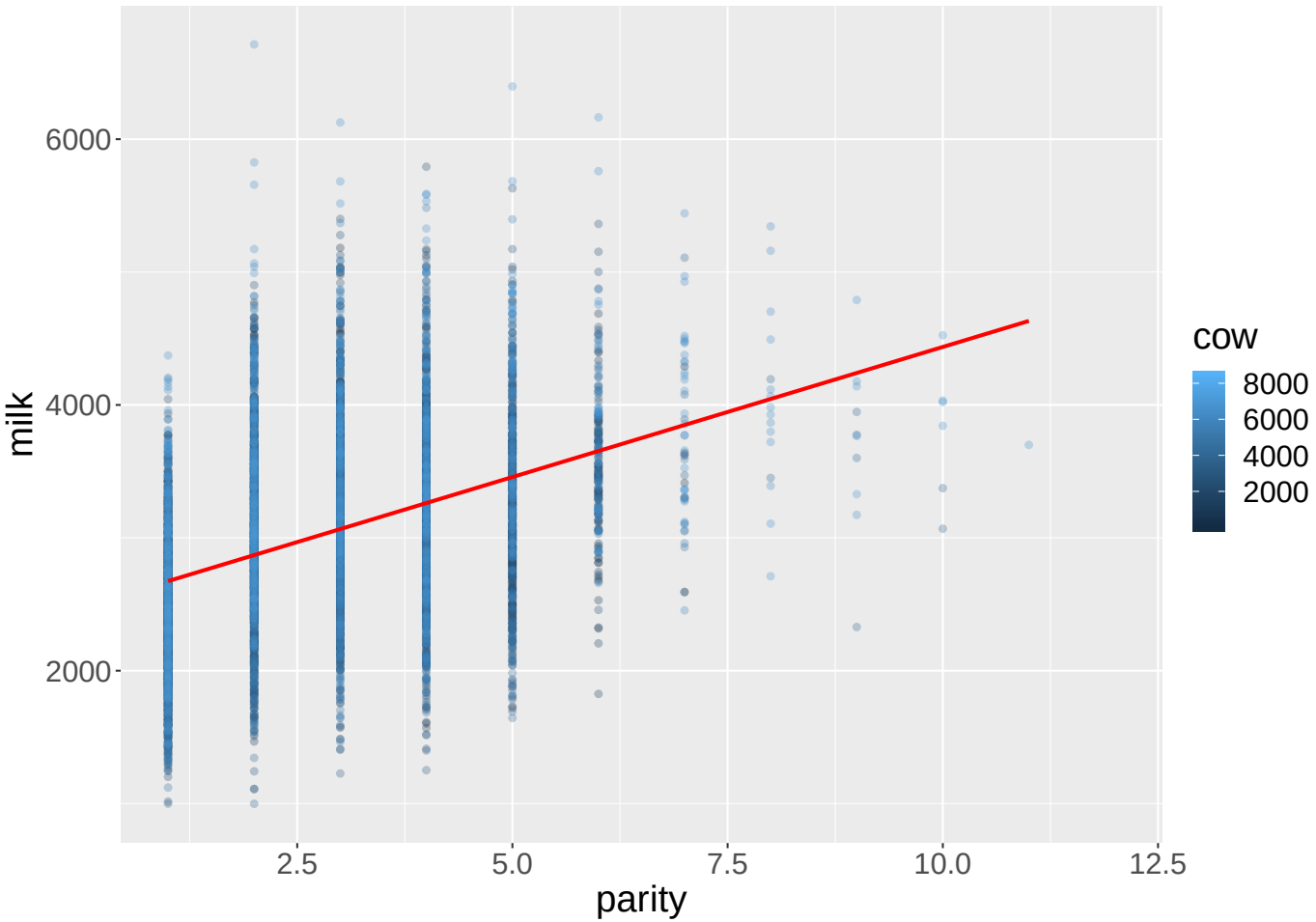


Figure 18: Impact of parity on Milk volume. 9383 lactation records in 42 year-round calving herds. http://projects.upei.ca/ver/data-and-samples/.

Einfache Regression, Beispiel: Modell-Ausgabe

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(milk120~parity, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = milk120 ~ parity, data = daisy2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2010.3  -454.6   -36.3   413.5  3842.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2478.286     17.176  144.28   <2e-16 ***
## parity       195.807      5.385   36.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.8 on 6608 degrees of freedom
##   (2773 observations deleted due to missingness)
## Multiple R-squared:  0.1667, Adjusted R-squared:  0.1666
## F-statistic:  1322 on 1 and 6608 DF,  p-value: < 2.2e-16
```

Einfache Regression, Beispiel: ANOVA-Tabelle

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: milk120
##             Df     Sum Sq   Mean Sq F value    Pr(>F)
## parity       1  589631568 589631568  1322.1 < 2.2e-16 ***
## Residuals 6608 2947090043    445988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Einfache Regression, Beispiel: Vorhersage-Intervall (1)

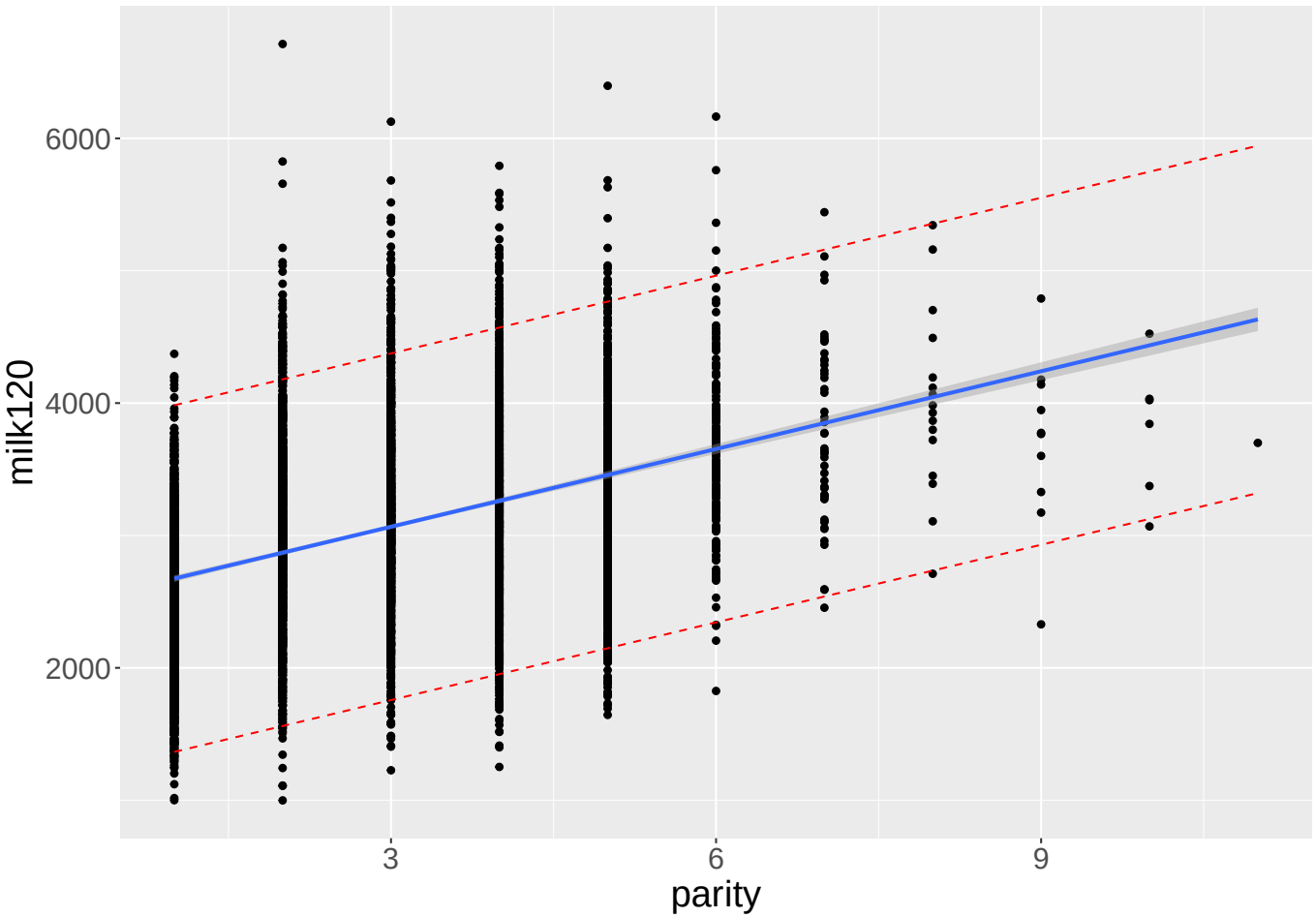


Figure 19: Prediction interval for the regression milk volume on parity.

Einfache Regression, Beispiel: Vorhersage-Intervall (2)

```
temp_var <- predict(model, interval="prediction")
new_df0 <- na.omit(data.frame("milk120" = daisy2$milk120, "parity" = daisy2$parity))
new_df <- cbind(new_df0, temp_var)
ggplot(new_df, aes(x=parity, y=milk120))+
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)
```

ANOVA

ANOVA

Falls die Antwortvariable (*response*) kontinuierlich ist, und die unabhängigen Variablen (*Faktoren*) kategoriell, macht es Sinn die Unterschiede zwischen den verschieden Gruppen von Faktorenniveaus oder Faktorenabstufugen zu betrachten.

Die statistischen Techniken für den Vergleich zwischen den Mittelwerten mehreren Gruppen bezeichnet man als **ANOVA**.

ANOVA kann und soll als GLM (veralgemeinerte *Regression*) betrachtet werden.

Terminologie

- Einfache ANOVA: eine erklärende (kategorielle) Variable (Faktor). Vergleiche zwischen den Niveaus von diesem Faktor.
- Zweifache (faktorielle) ANOVA: zwei erklärende (kategorielle) Variable (Faktor). Vergleiche zwischen den Niveaus von diesen Faktoren.
- Multiple ANOAVA – **MANOVA**: mehrere unabhägige Variable.
- ANCOVA (Analysis of covariances): zusätzlich zur ANOVA-Situtation kommt noch eine Kovariate für die abhängige Variable hinzu.
- Repeated measures (Messwiederholungen): bei den gleichen Objekten wurden die Messungen mehrmals vorgenommen.

–>

Einfache ANOVA

ANOVA Beispiel

```
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/Data files/Viagra.dat', sep = '', header = TRUE)
knitr::kable(df, caption = "Viagra-Datensatz [A. Field]")
```

Table 5: Viagra-Datensatz [A. Field]

person	dose	libido
1	1	3
2	1	2
3	1	1
4	1	1
5	1	4
6	2	5
7	2	2
8	2	4
9	2	2
10	2	3
11	3	7
12	3	4
13	3	5
14	3	3
15	3	6

ANOVA Beispiel (1)

```
library(gplots)
  plotmeans(df$libido ~ df$dose)
```

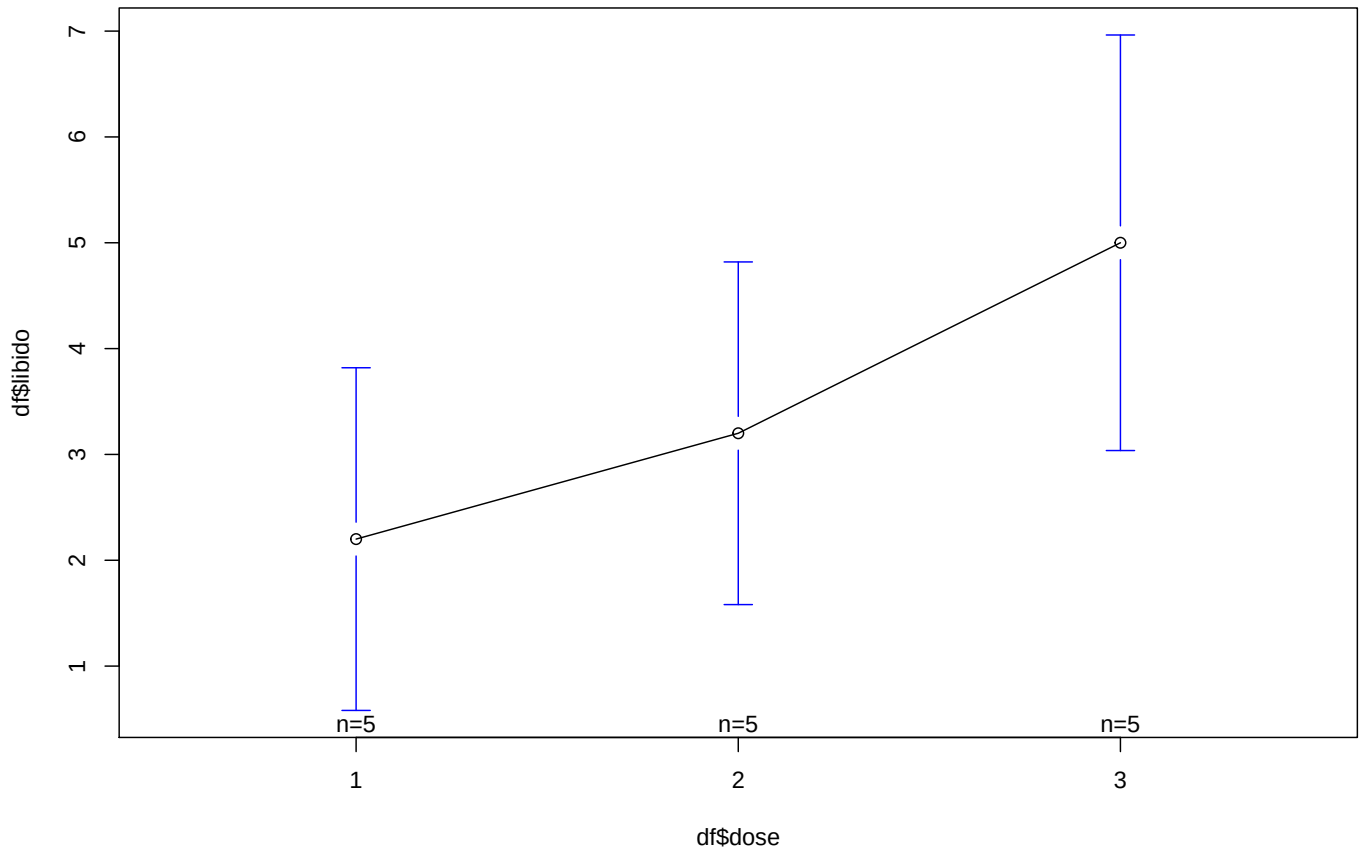


Figure 20: Viagra-Datensatz mit Konfidenzintervallen

ANOVA Beispiel (2)

```
library(pastecs)
by(df$libido, df$dose, stat.desc)
```

```
## df$dose: 1
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    1.0000000    4.0000000    3.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   11.0000000    2.0000000    2.2000000    0.5830952    1.6189318    1.7000000
##      std.dev     coef.var
##    1.3038405    0.5926548
## ------------------------------------------------------------
## df$dose: 2
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    2.0000000    5.0000000    3.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   16.0000000    3.0000000    3.2000000    0.5830952    1.6189318    1.7000000
##      std.dev     coef.var
##    1.3038405    0.4074502
## ------------------------------------------------------------
## df$dose: 3
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    3.0000000    7.0000000    4.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   25.0000000    5.0000000    5.0000000    0.7071068    1.9632432    2.5000000
##      std.dev     coef.var
##    1.5811388    0.3162278
```

ANOVA Beispiel (Levene's test)

```
library(pastecs)
library(car)
leveneTest(df$libido, df$dose, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.1176   0.89
##       12
```

If the *Levene's test* is non-significant, the data variances are very similar.

Falls nicht signifikant, kann man *Welch's F* oder robuste ANOVA benutzen.

ANOVA Beispiel (2)

$$\text{libido}_i = \text{dose}_i + \text{error}$$

ANOVA Beispiel (*lm*)

```r
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/Data files/Viagra.dat', sep = '', header = TRUE)
df$dose <- as.factor(df$dose)
library(ggplot2)
viagraModel <- lm(libido ~ dose, data = df)
summary(viagraModel)
```

```
##
## Call:
## lm(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2000     0.6272   3.508  0.00432 **
## dose2         1.0000     0.8869   1.127  0.28158
## dose3         2.8000     0.8869   3.157  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

ANOVA Beispiel (*aov*)

```
ViagraModel <- aov(libido ~ dose, data = df)
summary(ViagraModel)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dose         2  20.13  10.067   5.119 0.0247 *
## Residuals   12  23.60   1.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Test prüft, ob Unterschiede zwischen den Gruppen vorliegen.

F-Test sagt uns nicht zwischen welchen Gruppen die Unterschiede vorliegen.

ANOVA Beispiel (*aov* 1)

```
library(car)
qqPlot(lm(libido ~ dose, data = df), simulate = TRUE, main = 'Q-Q Plot', labels = FALSE)
```

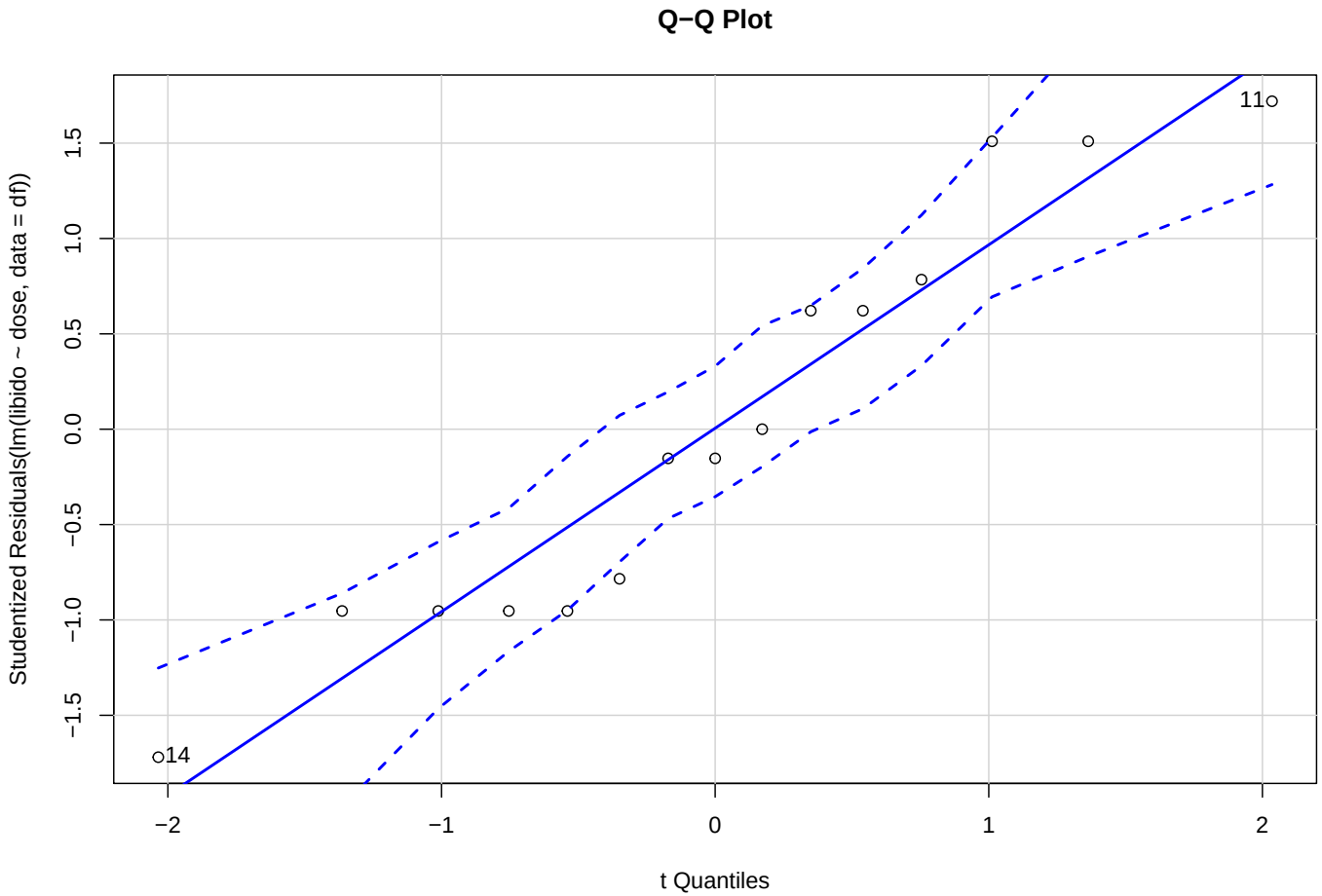


Figure 21: Q-Q Plot.

```
## [1] 11 14
```

ANOVA Beispiel: Bartlett-Test

```r
bartlett.test(libido ~ dose, data = df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  libido by dose
## Bartlett's K-squared = 0.1853, df = 2, p-value = 0.9115
```

ANOVA Beispiel: Aussreißer

```
outlierTest(ViagraModel)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 11  1.71959            0.11348           NA
```

ANOVA Beispiel (*aov* 3)

```
par(mfrow=c(2,2))
plot(ViagraModel)
```

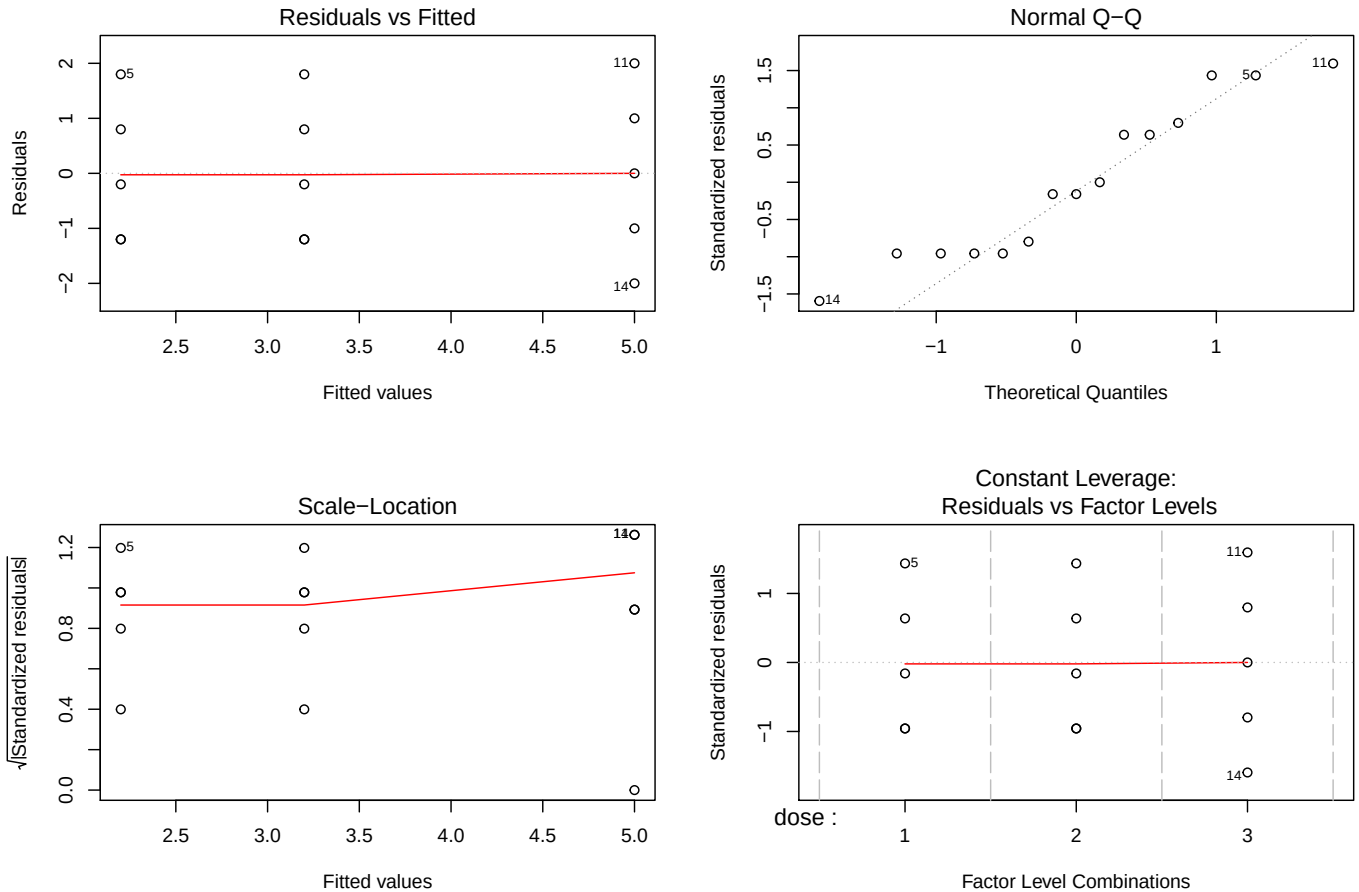


Figure 22: Ausgabe von aov-Funktion

Welch's F

Welch's F-Test berücksichtigt die vorhandenen Unterschiede in den Gruppenvarianzen.

```
oneway.test(libido ~ dose, data = df)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  libido and dose
## F = 4.3205, num df = 2.0000, denom df = 7.9434, p-value = 0.05374
```

In unserem Beispiel gibt's keine Unterschiede in den Gruppenvarianzen!

Robuste ANOVA

```
viagraWide <- unstack(df, libido ~ dose)
viagraWide1 <- data.frame(data = viagraWide)
colnames(viagraWide1) = c('Placebo','Low.Dose','High.Dose')
knitr::kable(viagraWide1, caption = "Viagra-Datensatz im breiten Format [A. Field]")
```

Table 6: Viagra-Datensatz im breiten Format [A. Field]

Placebo	Low.Dose	High.Dose
3	5	7
2	2	4
1	4	5
1	2	3
4	3	6

Robuste ANOVA (1)

```
library(WRS2)
t1way(libido ~ dose, data = df, tr = 0.1)
```

```
## Call:
## t1way(formula = libido ~ dose, data = df, tr = 0.1)
##
## Test statistic: F = 4.3205
## Degrees of freedom 1: 2
## Degrees of freedom 2: 7.94
## p-value: 0.05374
##
## Explanatory measure of effect size: 0.71
```

```
med1way(libido ~ dose, data = df) # median
```

```
## Call:
## med1way(formula = libido ~ dose, data = df)
##
## Test statistic F: 4.7829
## Critical value: 5.473
## p-value: 0.07
```

```
t1waybt(libido ~ dose, data = df)
```

```
## Call:
## t1waybt(formula = libido ~ dose, data = df)
##
## Effective number of bootstrap samples was 384.
##
## Test statistic: 3
## p-value: 0.09115
## Variance explained 0.623
## Effect size 0.789
```

Alle Tests sind nicht signifikant und somit hat die Dosis keine Aswirkung auf Libido.

Kontraste

Kontraste

```
summary.lm(viagraModel)
```

```
##
## Call:
## lm(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2000     0.6272   3.508  0.00432 **
## dose2         1.0000     0.8869   1.127  0.28158
## dose3         2.8000     0.8869   3.157  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Kontraste

```
contrasts(df$dose)
```

```
##   2 3
## 1 0 0
## 2 1 0
## 3 0 1
```

Geplante Kontraste

Sei gegeben die Variablen oder derer Statistiken $\theta_1, \theta_2, ..., \theta_k$ und die Konstanten $a_1, a_2, ..., a_k$.

Die lineare Kombination $\sum_i a_i \theta_i$ heißt *Kontrast* if $\sum_i a_i = 0$.

Zwei Kontraste $\sum_i a_i \theta_i$ und $\sum_i b_i \theta_i$ heißen *orthogonal* wenn $\sum_i a_i b_i = 0$.

Geplante Kontraste (1)

Contrast	Description
`contr.helmert`	Contrasts the second level with the first, the third level with the average of the first two, the fourth level with the average of the first three, and so on.
`contr.poly`	Contrasts are used for trend analysis (linear, quadratic, cubic, and so on) based on orthogonal polynomials. Use for ordered factors with equally spaced levels.
`contr.sum`	Contrasts are constrained to sum to zero. Also called *deviation contrasts*, they compare the mean of each level to the overall mean across levels.
`contr.treatment`	Contrasts each level with the baseline level (first level by default). Also called *dummy coding*.
`contr.SAS`	Similar to `contr.treatment`, but the baseline level is the last level. This produces coefficients similar to contrasts used in most SAS procedures.

Figure 23: Verschiedene eingebaute Kontraste [Quelle: A. Kabacoff]

```
contrasts(df$dose) <- contr.helmert(3)
df$dose
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
## attr(,"contrasts")
##   [,1] [,2]
## 1   -1   -1
## 2    1   -1
## 3    0    2
## Levels: 1 2 3
```

Geplante Kontraste (3)

```
contrasts(df$dose) <- cbind(c(-2,1,1),c(0,-1,1))
df$dose
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
## attr(,"contrasts")
##   [,1] [,2]
## 1   -2    0
## 2    1   -1
## 3    1    1
## Levels: 1 2 3
```

- Bei den geplanten Kontrasten vergleichen wir die Gruppen mit dem positiven Vorzeichen gegen die Gruppen mit dem negativen.

- Im Kontrast 1 vergleichen wir Placebo-Gruppe und die zwei Versuchsgruppen
- Im Kontrast 2 vergleichen wir niedrige Dosis-Gruppe und die höhe Dosis-Gruppe

Geplante Kontraste (4)

```r
ViagraModel3 <- aov(libido ~ dose, data = df)
summary.lm(ViagraModel3)
```

```
##
## Call:
## aov(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4667     0.3621   9.574 5.72e-07 ***
## dose1         0.6333     0.2560   2.474   0.0293 *
## dose2         0.9000     0.4435   2.029   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Geplante Kontraste (5)

- Weil wir davon ausgehen, dass Libido durch die Gabe des Medikaments steigt, können wir den zweiseitigen Test durch den einseitigen ersetzen und die entsprechenden p-Werte halbieren.
- Ohne die Hypothese, dass die Wirkung mit der Dosis steigt, wäre unser Ergebnis nicht signifikant
- Das zeigt, dass es wichtig ist, die Hypothese zu haben, bevor wir die Daten sammeln.

Post hoc-Tests

Post hoc-Tests

- *F*-Test sagt uns nur, dass es die Unterschiede vorhanden sind, aber nicht zwischen welchen Gruppen.
- Um die paarweisen Vergleiche durchzuführen, brauchen wir *post hoc*-Tests.

Bonferroni und Benjamini-Hochberg *post hoc*-Tests

```r
pairwise.t.test(df$libido, df$dose, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$libido and df$dose
##
##   1     2
## 2 0.845 -
## 3 0.025 0.196
##
## P value adjustment method: bonferroni
```

```r
pairwise.t.test(df$libido, df$dose, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$libido and df$dose
##
##   1     2
## 2 0.282 -
## 3 0.025 0.098
##
## P value adjustment method: BH
```

Tukey HSD paarweise Vergleiche

```
TukeyHSD(ViagraModel3)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## $dose
##     diff         lwr      upr     p adj
## 2-1  1.0 -1.3662412 3.366241 0.5162761
## 3-1  2.8  0.4337588 5.166241 0.0209244
## 3-2  1.8 -0.5662412 4.166241 0.1474576
```

Tukey HSD paarweise Vergleiche (1)

```
par(las=2)
par(mar=c(5,8,4,2))
plot(TukeyHSD(ViagraModel3), cex = 2, cex.main=2, cex.lab=2, cex.axis=2)
```

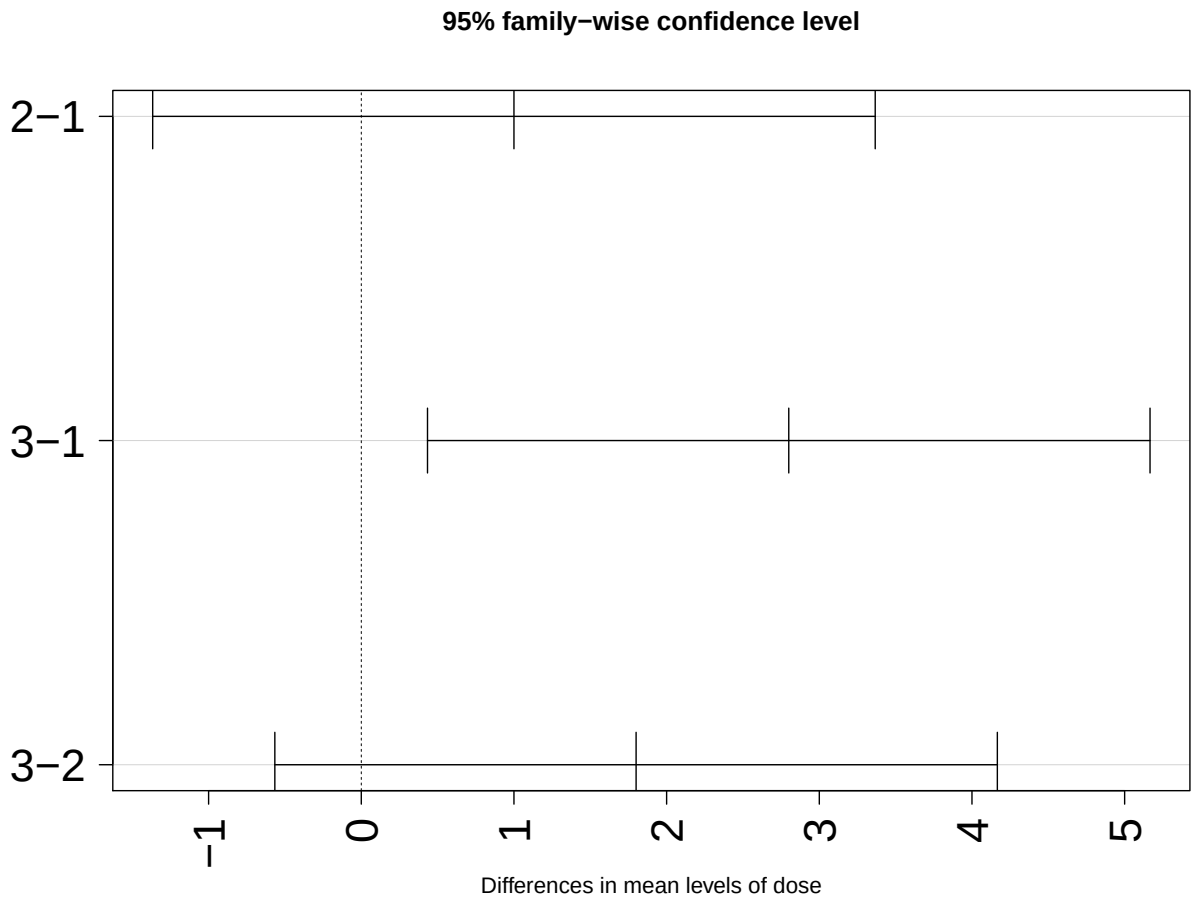


Figure 24: Tukey HSD paarweise Vegleiche.

Tukey HSD paarweise Vergleiche (2)

```r
library(multcomp)
par(mar=c(5,4,6,2))
tuk <- glht(ViagraModel3, linfct = mcp(dose = "Tukey"))
summary(tuk)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0   1.0000     0.8869   1.127   0.5163
## 3 - 1 == 0   2.8000     0.8869   3.157   0.0209 *
## 3 - 2 == 0   1.8000     0.8869   2.029   0.1475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Tukey HSD paarweise Vergleiche (3)

```r
confint(tuk)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## Quantile = 2.6658
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##            Estimate lwr     upr
## 2 - 1 == 0  1.0000  -1.3644  3.3644
## 3 - 1 == 0  2.8000   0.4356  5.1644
## 3 - 2 == 0  1.8000  -0.5644  4.1644
```

Für signifikante Unterschiede dürfen die Konfidenzintervalle die Null nicht durchkreuzen!

Tukey HSD paarweise Vergleiche (4)

```
par(mar=c(5,8,4,2))
plot(cld(tuk, level = .05, col = "lightblue"), cex = 2, cex.main=2, cex.lab=2, cex.axis=2)
```

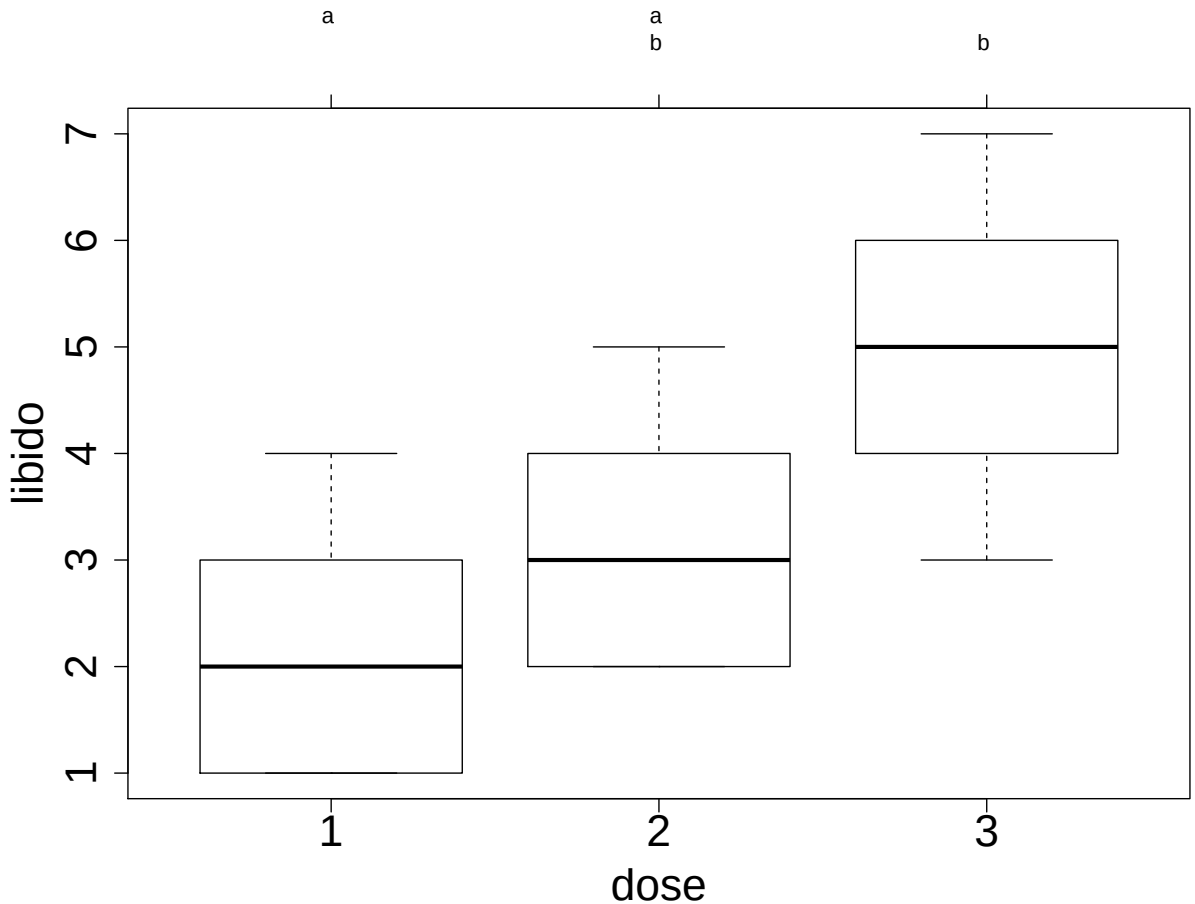


Figure 25: Tukey HSD paarweise Vegleiche. Gruppen, die die gleichen Buchstaben oben zu stehen haben, sind nicht signifikant unterschiedlich.

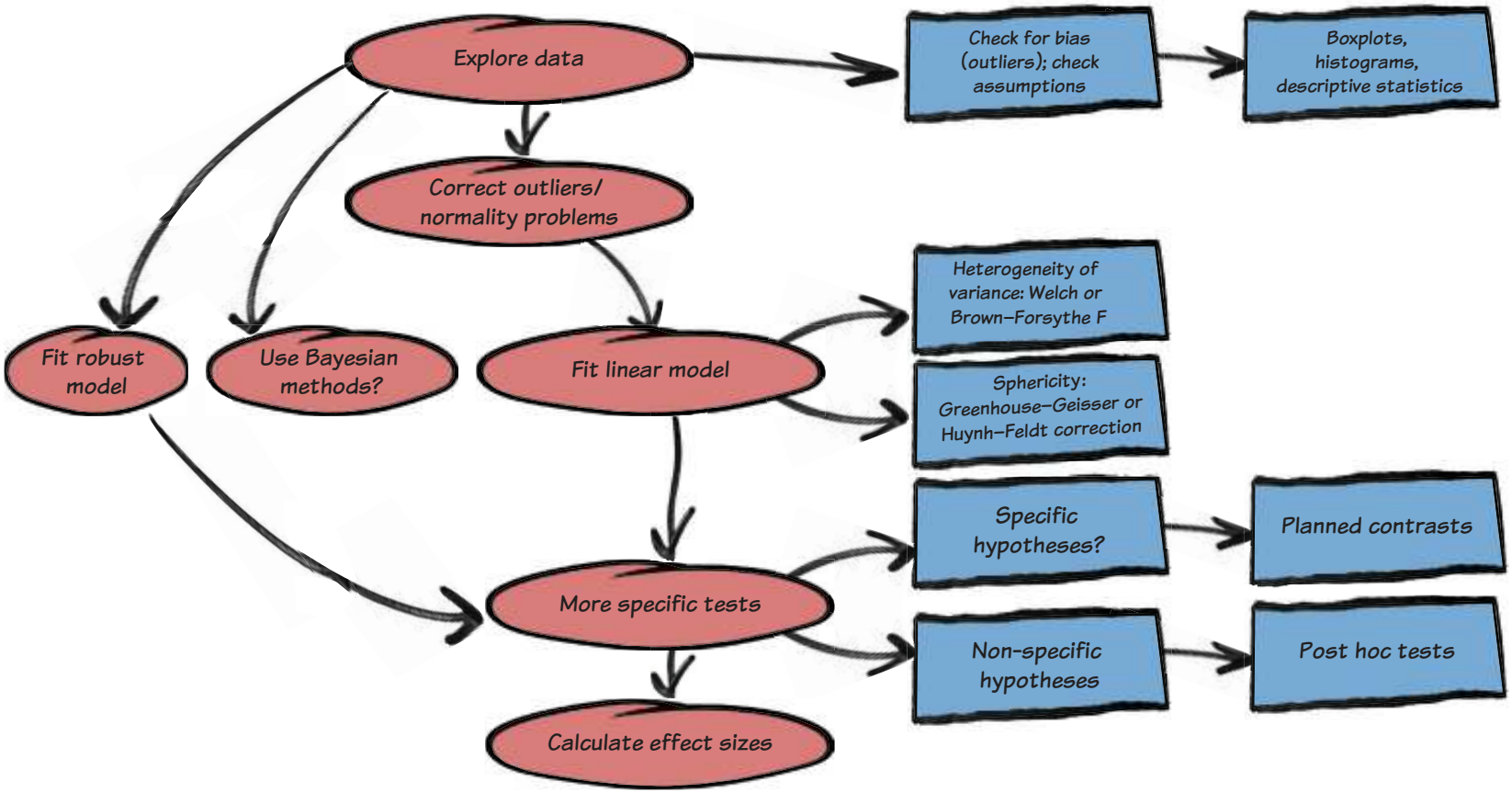


Figure 26: Vergleich mehrerer Mittelwerte [Quelle: A. Field]

Effect size

$$R^2 = \frac{SSM}{SST}$$

Effektgröße ist gegeben durch R (0.68)

Anstatt r^2 benutzt man in ANOVA η^2.

$$\omega^2 = \frac{SSM - (df_M)MSE}{SST + MSE}$$

$\omega = 0.60$

- $\omega^2 = 0.01$ (small effects)
- $\omega^2 = 0.06$ (medium effects)
- $\omega^2 = 0.14$ (large effects)

Logistic regression

Generalized Linear Models

Generalized Linear Models

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i,\ y \sim \mathcal{N}(\mu, \sigma)$$

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. - Linearität nur in den Koeffizienten

$$g(y) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Generalized Linear Models

- binomiale (dichotomische) logistische Regression
- multinominale logistische Regression

Table 7: Link functions

Datentyp	Transformation	Verteilung
kontinuierlich	$\log(x)$	Log-normal
Anzahl	$\sqrt{x}$ oder $\log(x + 0.5)$	Poisson, Negative binomiale, ...
Verhältnis	$\arcsin\sqrt{x}$ oder logit $= \log\frac{x}{1-x}$	Bernoulli / binomiale, Beta binomiale, ...

Logistische Regression

Es gibt Situationen wann die *Antwortvariable* nicht normal verteilt ist. Z.B. kann sie kategoriell und *binomial* oder *multinomial* sein.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

Dabei ist $\pi = \mu_Y$ ein bedingter Mittelwert (d.h. die Wahrscheinlichkeit, dass $Y = 1$ vorausgesetzt die vorhandenen X-Werte).

$\frac{\pi}{1-\pi}$ ist das Odds-Ratio, dass $Y = 1$.

$\log\left(\frac{\pi}{1-\pi}\right)$ ist *log odds* oder *logit*.

Logistische Regression: Beispiel

Welche persönliche, demographische, und Beziehungsvariablen können Untreue vorhersagen?

Table 8: Auszug aus dem Datensatz über Untreueverhalten [nach Kabacoff / Green&Fair]

	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
4	0	male	37	10.00	no	3	18	7	4
5	0	female	27	4.00	no	4	14	6	4
11	0	female	32	15.00	yes	1	12	1	4
16	0	male	57	15.00	yes	5	18	6	5
23	0	male	22	0.75	no	2	17	6	3
29	0	female	32	1.50	no	2	17	5	5
44	0	female	22	0.75	no	2	12	1	3
45	0	male	57	15.00	yes	2	14	4	4
47	0	female	32	15.00	yes	4	16	1	2
49	0	male	22	1.50	no	4	14	4	5
50	0	male	37	15.00	yes	2	20	7	2
55	0	male	27	4.00	yes	4	18	6	4
64	0	male	47	15.00	yes	5	17	6	4
80	0	female	22	1.50	no	2	17	5	4
86	0	female	27	4.00	no	4	14	5	4

Logistische Regression: Beispiel(1)

```
summary(Affairs)
```

```
##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation       rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

Logistische Regression: Beispiel(1)

```
knitr::kable(table(Affairs$affairs))
```

Var1	Freq
0	451
1	34
2	17
3	19
7	42
12	38

Logistische Regression: Beispiel(2)

- Transformation zu binären Variablen

```
Affairs$ynaffair[Affairs$affair > 0] <- 1
Affairs$ynaffair[Affairs$affair == 0] <- 0
Affairs$ynaffair <- factor(Affairs$ynaffair,
                           levels = c(0,1),
                           labels = c("No","yes"))
knitr::kable(table(Affairs$ynaffair))
```

Var1	Freq
No	451
yes	150

Logistische Regression: Beispiel(3)

```r
fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children + religiousness + education + occupation + rating,
                data = Affairs, family = binomial() )
summary(fit.full)
```

```
##
## Call:
## glm(formula = ynaffair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5713  -0.7499  -0.5690  -0.2539   2.5191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.37726    0.88776   1.551 0.120807
## gendermale     0.28029    0.23909   1.172 0.241083
## age           -0.04426    0.01825  -2.425 0.015301 *
## yearsmarried   0.09477    0.03221   2.942 0.003262 **
## childrenyes    0.39767    0.29151   1.364 0.172508
## religiousness -0.32472    0.08975  -3.618 0.000297 ***
## education      0.02105    0.05051   0.417 0.676851
## occupation     0.03092    0.07178   0.431 0.666630
## rating        -0.46845    0.09091  -5.153 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.51
##
## Number of Fisher Scoring iterations: 4
```

Beispiel: reduziertes Modell

```r
fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
                   data = Affairs, family = binomial() )
summary(fit.reduced)
```

```
##
## Call:
## glm(formula = ynaffair ~ age + yearsmarried + religiousness +
##     rating, family = binomial(), data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6278  -0.7550  -0.5701  -0.2624   2.3998
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.93083    0.61032   3.164 0.001558 **
## age           -0.03527    0.01736  -2.032 0.042127 *
## yearsmarried   0.10062    0.02921   3.445 0.000571 ***
## religiousness -0.32902    0.08945  -3.678 0.000235 ***
## rating        -0.46136    0.08884  -5.193 2.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 615.36  on 596  degrees of freedom
## AIC: 625.36
##
## Number of Fisher Scoring iterations: 4
```

Beispiel: Modellvergleich (χ^2)

```r
anova(fit.reduced, fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: ynaffair ~ age + yearsmarried + religiousness + rating
## Model 2: ynaffair ~ gender + age + yearsmarried + children + religiousness +
##     education + occupation + rating
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       596     615.36
## 2       592     609.51  4   5.8474   0.2108
```

Beispiel: Interpretation der Koeffezienten

Regressionskoeffizienten geben die Veränderung (in *log(odds)*) in der Antwortvariable, wenn alle weiteren Variablen konstant bleiben.

```
coef(fit.reduced)
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##    1.93083017   -0.03527112    0.10062274   -0.32902386   -0.46136144
```

```
exp(coef(fit.reduced))
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##     6.8952321     0.9653437     1.1058594     0.7196258     0.6304248
```

```
exp(confint(fit.reduced))
```

```
##                    2.5 %     97.5 %
## (Intercept)   2.1255764 23.3506030
## age           0.9323342  0.9981470
## yearsmarried  1.0448584  1.1718250
## religiousness 0.6026782  0.8562807
## rating        0.5286586  0.7493370
```

Overdispersion

$$\sigma^2 = n\pi(1-\pi).$$

```
fit <-  fit.reduced
fit.od <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
                data = Affairs, family = quasibinomial() )
pchisq(summary(fit.od)$dispersion*fit$df.residual, fit$df.residual, lower = F)
```

```
## [1] 0.340122
```

Wahrscheinlichkeitsverteilungen

Normalverteilung oder Gauß-Verteilung

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \mathcal{N}(\mu, \sigma)$$

μ — Mittelwert

σ — Standardabweichung

Normalverteilung oder Gauß-Verteilung (1)

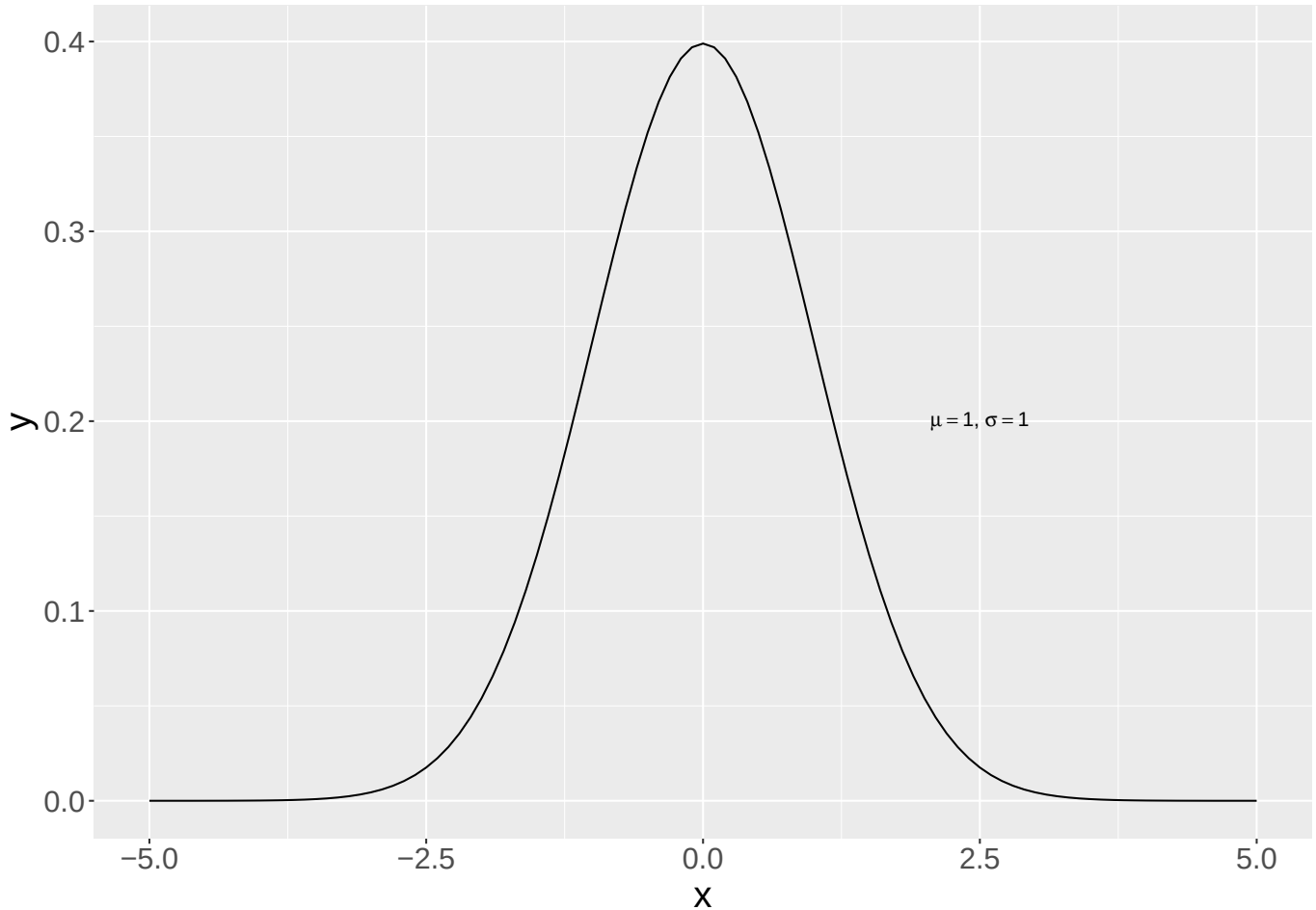


Figure 27: Normalverteilung

Binomiale Verteilung

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^n - k$$

n - Anzahl der Versuche

p - Erfolgs- oder Trefferwahrscheinlichkeit

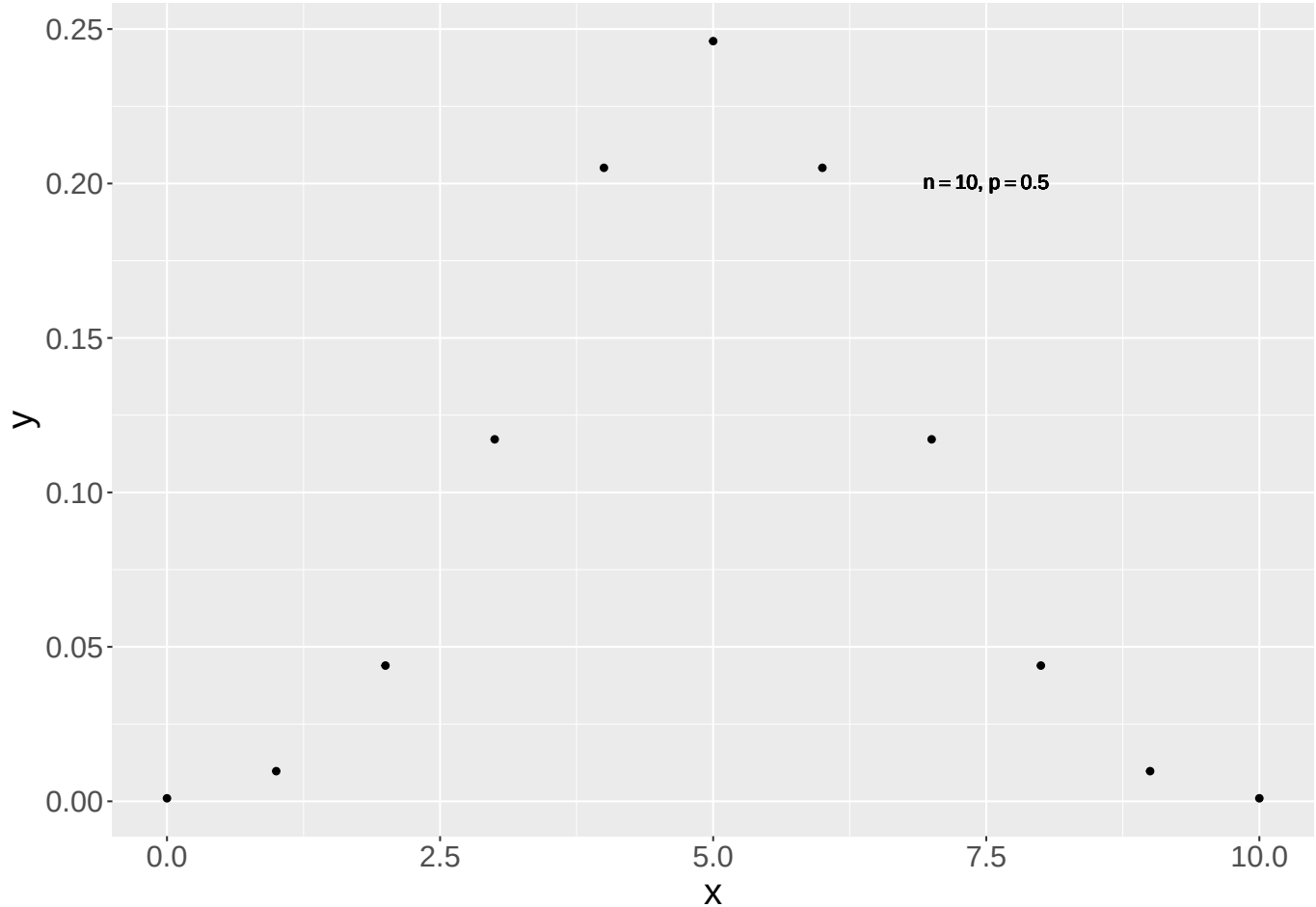


Figure 28: Binomiale Verteilung

Poisson-Verteilung

$$P_\lambda(k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}$$

λ Erwartungswert und gleichzeitig die Varianz der Poisson-Verteilung

Poisson-Verteilung (1)

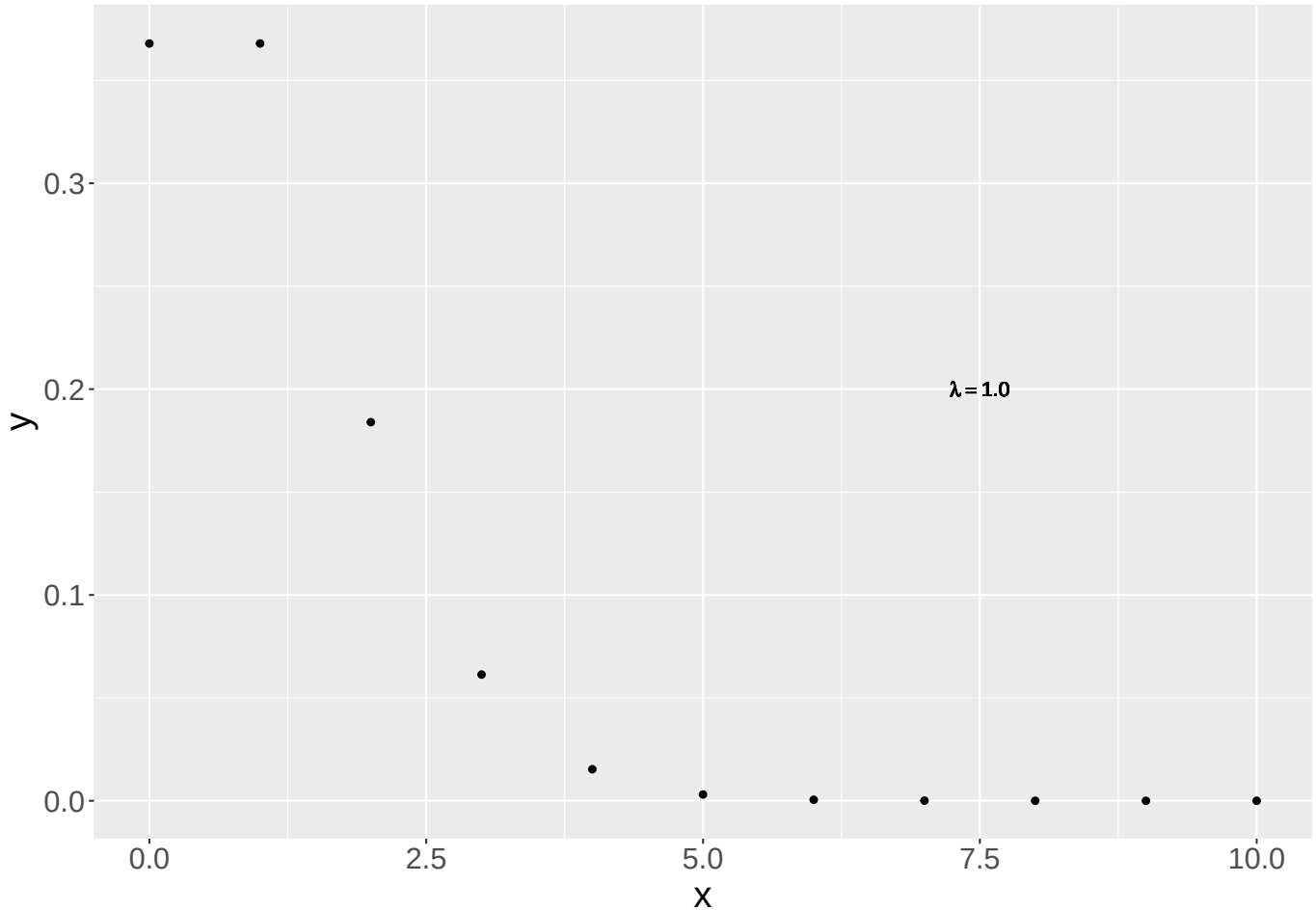


Figure 29: Poisson-Verteilung