

# STATISTIK

Wintersemester 2016/2017

---

*Vorlesungsfolien*

# Inhaltsverzeichnis

<b>1</b>	<b>Einführung</b>	<b>1</b>
1.1	Was ist Statistik?	3
1.2	R	5
<b>2</b>	<b>Deskriptive Statistik</b>	<b>6</b>
2.1	Ausgangspunkt	6
2.1.1	Die Grundgesamtheit	6
2.1.2	Stichproben	8
2.1.3	Merkmale	9
2.1.4	Klassifikation von Merkmalen	11
2.2	Kenngrößen univariater Daten	16
2.2.1	Stichproben	16
2.2.2	Häufigkeiten	17
2.2.3	Klassenbildung	21
2.2.4	Empirische Verteilungsfunktion	26
2.3	Diagramme und Grafiken	30
2.3.1	Stab- und Säulendiagramme	30
2.3.2	Kreis- und Tortendiagramme	32
2.3.3	Histogramm und empirische Dichtefunktion	33
2.4	Lagemaße	38
2.4.1	Arithmetisches Mittel	38

---

2.4.2	Arithmetisches Mittel für klassierte Daten . . . . .	43
2.4.3	Arithmetisches Mittel für gepoolte Daten . . . . .	44
2.4.4	Die Ordnungsstatistik . . . . .	46
2.4.5	Getrimmtes Mittel . . . . .	47
2.4.6	Median . . . . .	50
2.4.7	Quantile und Quartile . . . . .	53
2.4.8	Das geometrische Mittel . . . . .	54
2.4.9	Weitere Mittelwerte . . . . .	56
2.5	Streuungsmaße . . . . .	58
2.5.1	Varianz und Standardabweichung . . . . .	58
2.5.2	Varianz für gepoolte Daten (Varianzzerlegung) . . . . .	65
2.5.3	Spannweite und Interquartilsabstand . . . . .	67
2.5.4	Variationskoeffizient . . . . .	68
2.5.5	Weitere Streuungsmaße . . . . .	69
2.6	Boxplots . . . . .	80
2.7	Konzentrationsmaße . . . . .	83
2.7.1	Die Lorenz-Kurve . . . . .	83
2.7.2	Das Gini-Maß . . . . .	87
2.8	Bivariate Daten . . . . .	89
2.8.1	Häufigkeiten und Kontingenztabellen . . . . .	91
2.8.2	Unabhängige Merkmale . . . . .	95
2.8.3	Zusammenhangsmaße für nominale Daten . . . . .	97
2.8.4	Zusammenhangsmaße für metrische Daten . . . . .	102
2.8.5	Zusammenhangsmaße für ordinale Daten . . . . .	107

---

<b>3</b>	<b>Wahrscheinlichkeitsrechnung</b>	<b>111</b>
3.1	Ereignisse und Wahrscheinlichkeiten	113
3.1.1	Laplace-Experimente	118
3.1.2	Bedingte Wahrscheinlichkeiten	121
3.1.3	Unabhängigkeit	123
3.2	Kombinatorik	125
3.2.1	Permutationen	125
3.2.2	Variationen und Kombinationen	126
3.3	Zufallsvariablen und ihre Verteilungen	130
3.3.1	Zufallsvariablen	130
3.3.2	Verteilungsfunktionen	131
3.4	Erwartungswert und Varianz	136
3.5	Das Gesetz der großen Zahlen	139
3.6	Unabhängigkeit und Korrelation	142
3.7	Fünf wichtige Verteilungen	144
3.7.1	Die Bernoulli-Verteilung	144
3.7.2	Die Binomialverteilung	145
3.7.3	Die geometrische Verteilung	147
3.7.4	Die Multinomialverteilung	151
3.7.5	Die stetige Gleichverteilung	156
3.8	Die Normalverteilung und ihre Verwandten	159
3.8.1	Die Standardnormalverteilung	159
3.8.2	Tabellen und Quantile	161
3.8.3	Der zentrale Grenzwertsatz	164

---

3.8.4	Abschätzungen . . . . .	170
3.8.5	Die allgemeine Normalverteilung . . . . .	173
3.8.6	Rechenregeln und Transformationen für die Normalverteilung . . . . .	176
3.8.7	Die Chi-Quadrat-Verteilung . . . . .	178
3.8.8	Die t-Verteilung . . . . .	180
3.8.9	Die F-Verteilung . . . . .	182
3.8.10	Ein Beispiel zum Schluss . . . . .	184
<b>4</b>	<b>Induktive Statistik</b>	<b>189</b>
4.1	Punktschätzer . . . . .	189
4.1.1	Punktschätzer für den Erwartungswert . . . . .	193
4.1.2	Punktschätzer für die Varianz bei bekanntem Erwartungswert . . . . .	197
4.1.3	Punktschätzer für die Varianz bei unbekanntem Erwartungswert . . . . .	198
4.2	Intervallschätzer . . . . .	200
4.2.1	Intervallschätzer für den Erwartungswert bei bekannter Varianz . . . . .	200
4.2.2	Intervallschätzer für den Erwartungswert bei unbekannter Varianz . . . . .	205
4.2.3	Intervallschätzer für die Varianz bei bekanntem Erwartungswert . . . . .	209

---

---

4.2.4	Intervallschätzer für die Varianz bei unbekanntem Erwartungswert . . . . .	211
4.2.5	Schätzen „ohne Zurücklegen“ . . . . .	215
4.3	Hypothesentests . . . . .	216
4.3.1	Idee . . . . .	216
4.3.2	Wahl des Ablehnungsbereiches . . . . .	220
4.3.3	Vorgehensweise . . . . .	222
4.3.4	Die Gütefunktion . . . . .	226
4.3.5	Der p-Wert . . . . .	227
4.3.6	Einstichprobentests für den Erwartungswert bei normalverteilter Grundgesamtheit . . . . .	231
(1)	Test bei bekannter Varianz . . . . .	232
(2)	Test bei unbekannter Varianz (t-Test) . . . . .	234
4.3.7	Einstichprobentests für die Varianz bei normalverteilter Grundgesamtheit . . . . .	238
(1)	Test bei bekanntem Erwartungswert . . . . .	239
(2)	Test bei unbekanntem Erwartungswert . . . . .	244
4.3.8	Zweistichprobentest auf gleiche Erwartungswerte (t-Test) . . . . .	246
4.3.9	Zweistichprobentest auf gleiche Varianzen (F-Test) . . . . .	250
4.3.10	Chi-Quadrat-Anpassungstest . . . . .	254
4.3.11	Weitere Tests auf Normalität . . . . .	261
4.3.12	Q-Q-Plots . . . . .	262
4.3.13	Der Chi-Quadrat-Homogenitätstest . . . . .	269
4.3.14	Der Chi-Quadrat-Unabhängigkeitstest . . . . .	273

---

---

4.3.15	Test auf Ausreißer	277
4.4	Einfache lineare Regression	281
4.4.1	Die Kleinste-Quadrate-Methode	286
4.4.2	Prognosen	292
4.4.3	Standardbedingungen und Güte der Schätzer	294
4.4.4	Das Bestimmtheitsmaß	296
4.4.5	Intervallschätzer	301
4.4.6	Tests zur Anpassungsgüte	303
4.4.7	Beispielregression mit R	306
<b>A</b>	<b>Übungsaufgaben</b>	<b>313</b>
A.1	Aufgaben	313
A.2	Musterlösungen	345
<b>B</b>	<b>Anhang</b>	<b>352</b>
B.1	Kleine Formelsammlung	352
B.1.1	Notationen (Deskriptive Statistik)	352
B.1.2	Wahrscheinlichkeitstheorie	353
B.1.3	Schätzer und Konfidenzintervalle	355
B.2	Tabellen	356
B.2.1	Quantile $z_\alpha$ der Normalverteilung	356
B.2.2	Verteilungsfunktion $\Phi(\mathbf{x})$ der Normalverteilung	357
B.2.3	Quantile $t_{n,\alpha}$ der t-Verteilung	359
B.2.4	Quantile $\chi_{n,\alpha}$ der Chi-Quadrat-Verteilung	361

---

---

B.2.5	Quantile $F_{(n,m),\alpha}$ der F-Verteilung . . . . .	363
<b>C</b>	<b>Hinweise zur Klausur</b>	<b>368</b>
C.1	Hilfsmittel . . . . .	368
C.2	Welche Abschnitte und Gegenstände werden nicht abgefragt? . . .	369

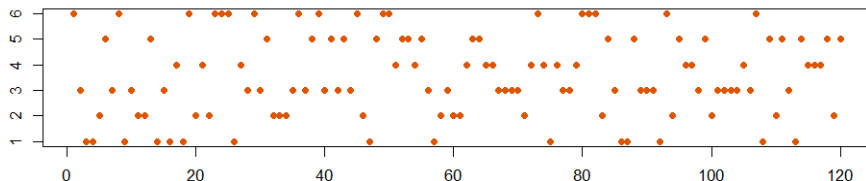


## 1.

# Einführung



■ **Beispiel B1.1:** Eine Firma stellt Spielwürfel her und überprüft von Zeit zu Zeit ihre Produkte, indem sie Stichproben zieht. Dazu wird ein Würfel ausgewählt und 120 Mal geworfen. Die Anzahl der Würfe für die verschiedenen Augenzahlen wird notiert.



Es ergibt sich folgende Häufigkeitstabelle:

Augenzahl:	1	2	3	4	5	6
Häufigkeit:	15	18	30	18	21	18

Wir können z.B. folgende Fragen stellen:

- Wie kann man die Daten grafisch darstellen?
- Wie häufig „sollten“ die Augenzahlen bei einem fairen Würfel vorkommen? (Ist so eine Frage überhaupt sinnvoll?)
- Welche Abweichungen sind noch akzeptabel?
- Kann man sagen, ob der vorliegende Würfel fair ist?  
Mit welcher Sicherheit ist eine solche Aussage zu machen?

## 1.1. Was ist Statistik?

- Erhebung, Erfassung, Darstellung/Präsentation, Analyse und Interpretation von Daten.

Man unterscheidet:

- Deskriptive/beschreibende Statistik: Reduktion von Datenmengen, Darstellung durch Tabellen und Diagramme, Ermittlung aussagekräftiger Kenngrößen (z.B. Mittelwert, Varianz)
- Induktive Statistik: Weitere Rückschlüsse durch mathematische Methoden aus der Wahrscheinlichkeitsrechnung (z.B. Schätzen des Erwartungswertes, Hypothesentests)

## Woher kommen die Daten?

Beispiele:

- Technische Messungen (z.B. in der Meteorologie)
- Umfragen (z.B. im Vorfeld von Wahlen oder zur Kundenzufriedenheit)
- Nutzerstatistiken (z.B. für Internetprovider)
- Patientendaten
- Zugverspätungen
- Jahresberichte von Konzernen
- Statistische Ämter
- Finanzdaten: z.B. via Yahoo-Finance
- ...

## 1.2. R

Die Grafiken/Analysen in diesem Skript wurden mit R, einer Programmiersprache, die primär für statistische Anwendungen geschaffen wurde, erstellt.

Begleitend zur Vorlesung kann optional R auf dem Rechner installiert werden (s. erste Übung). Das Erlernen von R ist nicht Gegenstand der Vorlesung und wird nicht von den Studierenden verlangt.

Gleichwohl ist ein begleitendes Lernen computergestützter Methoden mit R hilfreich für das Verständnis im Umgang mit Daten.

Links:

[The R Project for Statistical Computing](#)

[RStudio \(GUI\)](#)

## 2.

## Deskriptive Statistik

### 2.1. Ausgangspunkt

#### 2.1.1. Die Grundgesamtheit

Als Grundgesamtheit (Population)  $\Omega$  bezeichnet man eine Menge von sogenannten statistischen Einheiten  $\omega \in \Omega$ .

■ **Beispiel B2.1:** Beim einmaligen Würfeln kann man als Grundgesamtheit  $\Omega = \{1, 2, 3, 4, 5, 6\}$  wählen. Jede der sechs Elemente ist dann eine statistische Einheit.

■ **Beispiel B2.2:** Alle Studierenden der HTW Dresden werden im Rahmen einer Umfrage befragt. Wir wählen z.B.

$$\Omega = \{00000, \dots, 99999\}$$

und identifizieren die Studierenden mit ihrer fünfstelligen Matrikelnummer.

■ **Beispiel B2.3:** Ein Thermometer misst jeden Tag morgens um acht Uhr die Außentemperatur. Man kann das Intervall

$$\Omega = [-30, 50]$$

als Grundgesamtheit wählen.

## 2.1.2. Stichproben

Man unterscheidet bei der Datenerhebung zwischen:

- Vollerhebungen: Erfassung der gesamten Population  $\Omega$ .  
■ **Beispiel B2.4**  $\Rightarrow$  B2.2: Alle Studierenden der HTW werden befragt.
- Teilerhebungen: Erfassung einer Stichprobe  $S \subset \Omega$   
■ **Beispiel B2.5**  $\Rightarrow$  B2.2: Nur die Studierenden der Vorlesung Statistik werden befragt.  
Teilerhebungen sind kostengünstiger und weniger aufwendig, aber der Statistiker muss von der Stichprobe auf die Grundgesamtheit schließen.



### 2.1.3. Merkmale

Ein Merkmal ist eine Eigenschaft, die jede der statistische Einheiten aufweist.

■ **Beispiel B2.6**  $\Rightarrow$  **B2.2**: Studierende an der HTW werden in einer Umfrage befragt. Folgende drei Merkmale werden erfasst:

- das Semester,
- die gesammelten ECTS-Punkte,
- das Alter,
- mit Abitur?

Für jeden Studierenden ergibt sich für jedes dieser Merkmale jeweils eine Beobachtung, z.B. für den Studierenden mit der Matrikelnummer 60182, Semester=1, ECTS-Punkte=0, Alter=19.

Mathematisch kann man ein Merkmal  $X$  als Abbildungen aus der Menge  $\Omega$  in die Menge aller möglichen Merkmalsausprägungen  $M_X$  auffassen:

$$X : \Omega \rightarrow M_X.$$

■ **Beispiel B2.7**  $\Rightarrow$  B2.2: Das Merkmal  $X$  repräsentiere die Semesterzahl. Dann ist  $X$  eine Abbildung von

$$\Omega = \{00000, \dots, 99999\}$$

in die Menge der Merkmalsausprägungen

$$M_X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

### 2.1.4. Klassifikation von Merkmalen

Merkmale werden u.a. nach ihrem Skalenniveau eingeteilt:

- Nominalskala: Keine sinnvolle Anordnung der Ausprägungen.

■ **Beispiel B2.8**  $\Rightarrow$  **B2.2**: Das Merkmal  $Y$  nehme die beiden Werte „Ja“ oder „Nein“ an, je nachdem, ob der Studierende das Abitur besitzt oder nicht, es ist also  $M_Y = \{\text{Ja}, \text{Nein}\}$ . Dann ist  $Y$  ein nominales Merkmal, denn es gibt keine Reihenfolge unter den Ausprägungen.

■ **Beispiel B2.9**: An einer Autobahn werden die vorbeifahrenden Wagen notiert. Das Merkmal „Automarke“ ist ein nominales Merkmal.

- Ordinalskala: Die Ausprägungen lassen sich anordnen und die Anordnung macht Sinn. Es gibt eine ' $\leq$ '-Relation.
- **Beispiel B2.10**: Die Examensnote von Studierenden ist ein ordinales Merkmal.
- **Beispiel B2.11**: Die monatlichen Ausgaben eines Haushalts sind ein ordinales Merkmal.

- Intervallskala: Es macht außerdem Sinn von einem Abstand bzw. der Differenz zwischen den Ausprägungen zu sprechen. Kein sinnvoller Nullpunkt und keine Möglichkeit der Multiplikation.
- **Beispiel B2.12**: Eine gemessene Temperatur ist intervallskaliert. (Was ist mit dem Nullpunkt?)
- **Beispiel B2.13**: Das Merkmal Uhrzeit ist intervallskaliert.

- Verhältnisskala: Es macht Sinn von Verhältnissen zwischen den Ausprägungen zu sprechen. Multiplikation und Division machen Sinn, ein Nullpunkt ist vorhanden.
- **Beispiel B2.14**: Die Körpergröße von Befragten ist verhältnisskaliert.
- **Beispiel B2.15**: Das Merkmal „Preis“ für eine Ware ist verhältnisskaliert.

- Ein Merkmal ist diskret, wenn es nur abzählbar viele Werte annehmen kann.

□ (Abzählbar) Eine Menge  $A$  heißt abzählbar, wenn man ein Verfahren angeben kann, mit dem man an jedes Element in  $A$  eine eindeutige Nummer  $\in \mathbb{N}$  vergeben kann.

■ **Beispiel B2.16**  $\Rightarrow$  B2.2: Das Merkmal Lebensalter (angegeben in Jahren) ist ein diskretes Merkmal.

- Ein Merkmal ist stetig, wenn praktisch jeder Zahlenwert in einem Zahlenintervall als Ausprägung vorkommen kann.

■ **Beispiel B2.17:** Das Merkmal  $L$ , dass die Länge eines gefertigten Werkstücks bezeichnet, ist ein stetiges Merkmal.

## 2.2. Kenngrößen univariater Daten

Univariate Daten liegen vor, wenn nur ein Merkmal  $X$  untersucht wird.

### 2.2.1. Stichproben

Wir betrachten eine Stichprobe des Merkmals  $X$  vom Umfang  $n$ , also  $n$  Beobachtungen

$$x_1 = X(\omega_1), x_2 = X(\omega_2), \dots, x_n = X(\omega_n).$$

Wir schreiben dafür meistens einfach

$$x_1, x_2, \dots, x_n.$$

Es können natürlich verschiedene Beobachtungen denselben Werte besitzen.



### 2.2.2. Häufigkeiten

Es sei nun  $X$  zusätzlich diskret, d.h.

$$M_X = \{a_1, a_2, a_3, \dots\}$$

mit den Merkmalsausprägungen  $a_i$ ,  $i = 1, 2, 3, \dots$

□ (Mächtigkeit einer Menge) Wir schreiben  $\#A$  für die Anzahl der Elemente in einer Menge  $A$ , z.B.

$$\#\{1, 2, 3, 4, 5, 6\} = 6, \quad \#\{A, B, C\} = 3, \quad \#\mathbb{N} = \infty$$

Die absolute Häufigkeit der Ausprägung  $a_i \in M_X$  ist der Wert

$$\begin{aligned} n_i = n(a_i) &= \text{Anzahl der } x_j \text{ mit } x_j = a_i \\ &= \#\{j \in \{1, 2, \dots, n\} \mid x_j = a_i\}. \end{aligned}$$

■ **Beispiel B2.18:** Ein Würfel wird  $n = 5$  Mal geworfen. Das Merkmal  $X$  entspreche der Augenzahl, d.h.

$$M_X = \{1, 2, 3, 4, 5, 6\}, \quad a_1 = 1, a_2 = 2, \dots, a_6 = 6.$$

Die entsprechenden Beobachtungen seien

$$x_1 = 3, \quad x_2 = 6, \quad x_3 = 1, \quad x_4 = 5, \quad x_5 = 6.$$

Dann sind die absoluten Häufigkeiten der Merkmalsausprägungen gegeben durch

$$n_1 = n(1) = 1, \quad n_2 = n(2) = 0,$$

$$n_3 = n(3) = 1, \quad n_4 = n(4) = 0,$$

$$n_5 = n(5) = 1, \quad n_6 = n(6) = 2.$$

Die relative Häufigkeit der Ausprägung  $a_i \in M_X$  ist der Wert

$$h_i = h(a_i) = \frac{n_i}{n}.$$

Es gilt

$$0 \leq h_i \leq 1, \quad (2.1)$$

$$\sum_{i=1}^{\#M_X} n_i = n, \quad (2.2)$$

$$\sum_{i=1}^{\#M_X} h_i = 1. \quad (2.3)$$

Man drückt die relativen Häufigkeiten auch in Prozent aus: Einer relativen Häufigkeit von  $h_i$  entsprechen dann  $h_i \cdot 100\%$ .

Die kumulativen absoluten/relativen Häufigkeiten sind gegeben durch die Summen

$$N_i = N(a_i) = n_1 + n_2 + \dots + n_i = \sum_{k=1}^i n_k,$$
$$H_i = H(a_i) = h_1 + h_2 + \dots + h_i = \sum_{k=1}^i h_k.$$

■ **Beispiel B2.19**  $\Rightarrow$  *B2.18*: Im obigen Beispiel ergibt sich:

$i$	$n_i$	$h_i$	$N_i$	$H_i$
1	1	0.2	1	0.2
2	0	0.0	1	0.2
3	1	0.2	2	0.4
4	1	0.2	3	0.6
5	0	0.0	3	0.6
6	2	0.4	5	1.0

### 2.2.3. Klassenbildung

Ist die Anzahl der Ausprägungen eines Merkmals sehr groß oder sogar unendlich, so empfiehlt es sich, die Daten in Klassen einzuteilen.

Die Klassen müssen folgende Eigenschaften erfüllen:

- Jede Ausprägung muss in einer Klasse vorkommen,
- Keine zwei Klassen enthalten dieselbe Ausprägung.

Natürlich ist die Klasseneinteilung mit einem Informationsverlust verbunden.

Faustregeln für die Klassenanzahl  $m$ :

$$m \approx \sqrt{n}$$

$$m \approx 1 + \log_2(n). \quad (\text{Sturges})$$

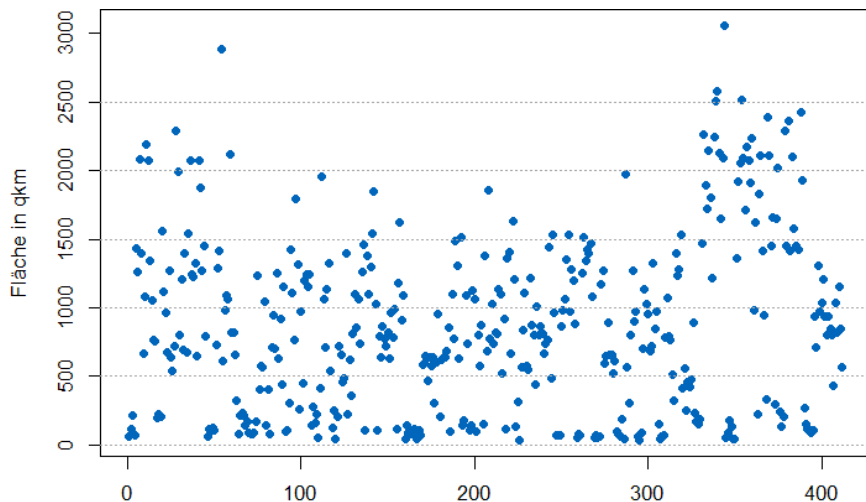
Man definiert Klassenhäufigkeiten als absolute/relative Häufigkeiten, summiert über alle Elemente der Klasse.

Für eine Klasse  $K \subseteq M_X$  ergibt sich also

$$n(K) = \sum_{a \in K} n(a),$$

$$h(K) = \sum_{a \in K} h(a).$$

■ **Beispiel B2.20:** Fläche von 407 bundesdeutschen Landkreisen (in  $\text{km}^2$ , Quelle: Stat. Bundesamt).



Wir teilen die Merkmalsausprägungen in Klassen ein:

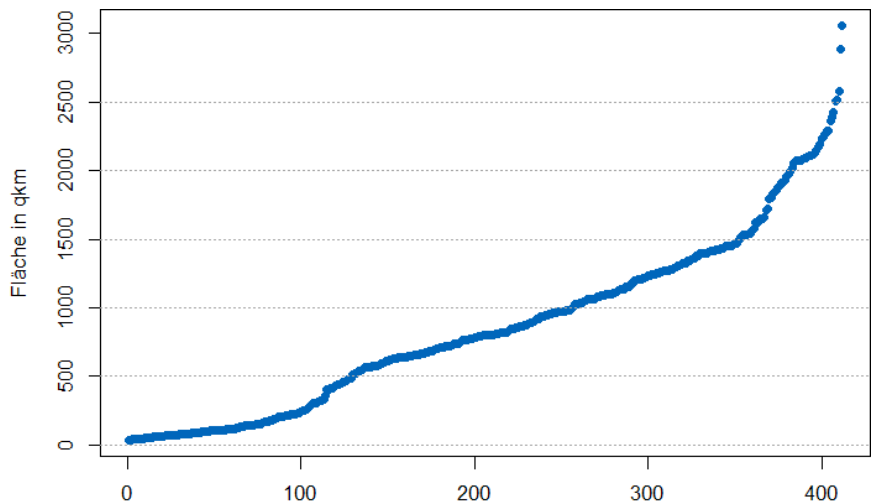
$$K_1 = (0, 500], K_2 = (500, 1000], K_3 = (1000, 1500], \\ K_4 = (1500, 2000], K_5 = (2000, \infty).$$

Absolute und relative Häufigkeiten:

$i$	$n(K_i)$	$h(K_i)$
1	129	0.317
2	127	0.312
3	96	0.236
4	30	0.074
5	30	0.074



Daten sortiert nach der Kreisgröße:



### 2.2.4. Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion beschreibt für jedes  $x \in \mathbb{R}$  die relative Anzahl von Beobachtungen  $x_i$  mit  $x_i \leq x$ :

$$F_n(x) = \frac{\#\{i \in \{1, 2, \dots, n\} | x_i \leq x\}}{n}.$$

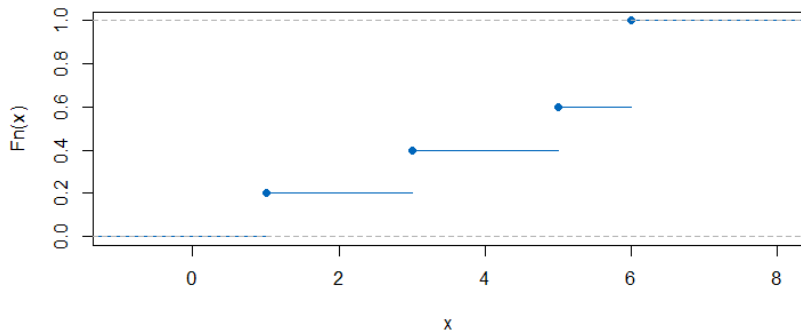
Es gilt:

1.  $F_n(x)$  ist monoton steigend (aber nicht streng monoton),
2.  $0 \leq F_n(x) \leq 1$ ,  $F_n(x)$  strebt gegen 0, wenn  $x$  gegen  $-\infty$  strebt,  $F_n(x)$  strebt gegen 1, wenn  $x$  gegen  $\infty$  strebt,
3.  $F_n(x)$  ist dort konstant, wo keine Beobachtungswerte vorliegen.

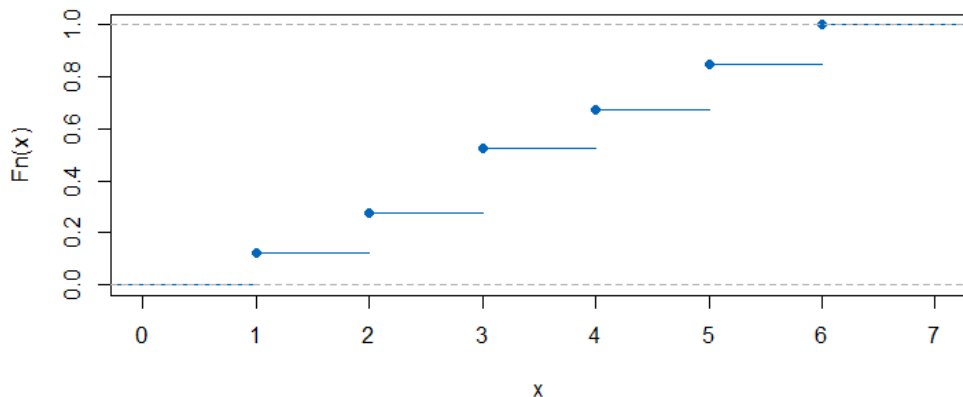
■ **Beispiel B2.21**  $\Rightarrow$  B2.18: Ein Würfel wird  $n = 5$  Mal geworfen, die entsprechenden Beobachtungen sind:

$$x_1 = 3, x_2 = 6, x_3 = 1, x_4 = 5, x_5 = 6.$$

Es ergibt sich folgende empirische Verteilungsfunktion:

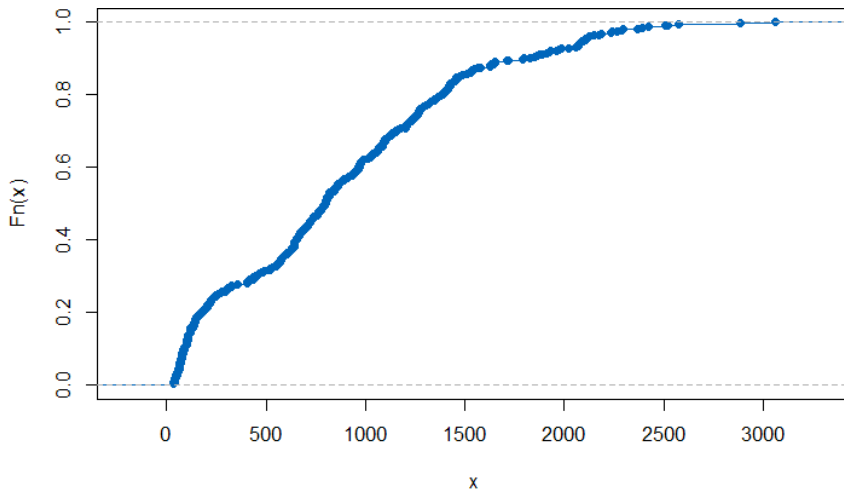


■ **Beispiel B2.22**  $\Rightarrow_{B1.1}$ : Im Eingangsbeispiel wurde ein Testwürfel 120 Mal geworfen. Es ergibt sich:



Wir werden später sehen, dass  $F_n(x)$  etwa der Verteilungsfunktion der Zufallsvariablen „Augenzahl“ entspricht.

■ **Beispiel B2.23**  $\Rightarrow$  B2.20: Für das Landkreisgrößen-Beispiel ergibt sich die folgende empirische Verteilungsfunktion:

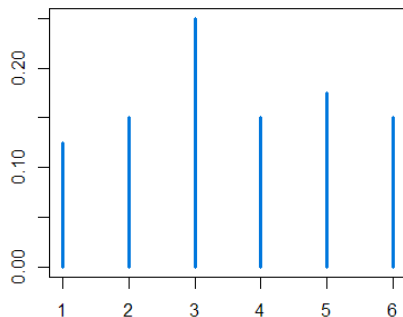
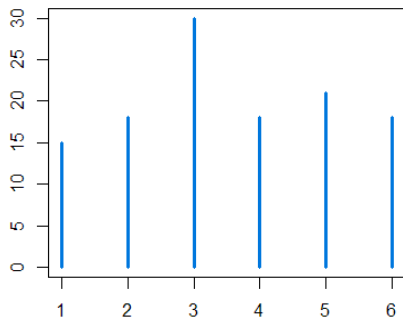


## 2.3. Diagramme und Grafiken

### 2.3.1. Stab- und Säulendiagramme

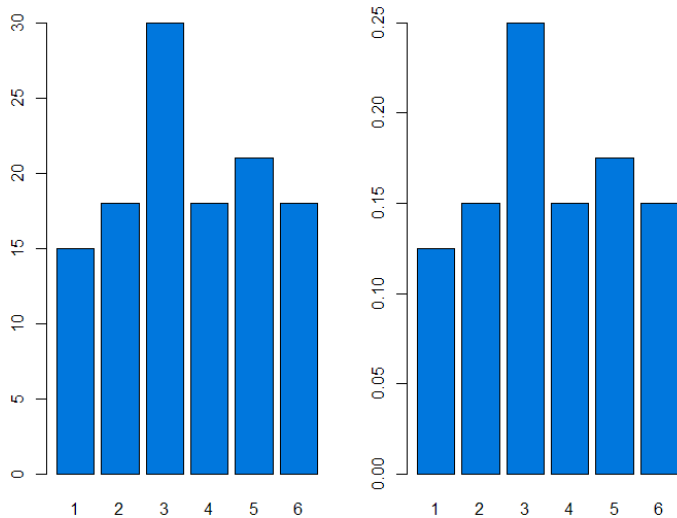
In Stabdiagrammen werden die relativen/absoluten Häufigkeiten als vertikale Linien dargestellt.

■ **Beispiel B2.24**  $\Rightarrow$  B1.1:



Im Balkendiagramm verwendet man stattdessen Balken. ■ **Beispiel**

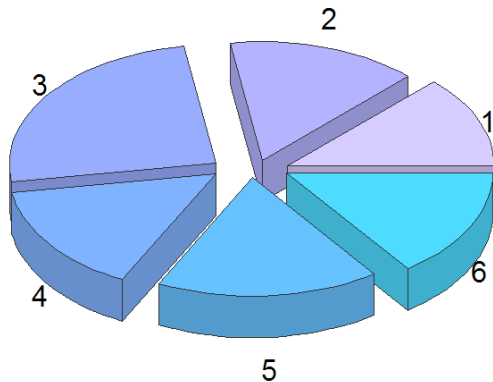
**B2.25**  $\Rightarrow$  B1.1 :



### 2.3.2. Kreis- und Tortendiagramme

Im Kreisdiagramm werden die relativen Häufigkeiten durch Kreissektoren beschrieben. Das Tortendiagramm ist eine dreidimensionale Variante.

■ **Beispiel B2.26**  $\Rightarrow B1.1$ :

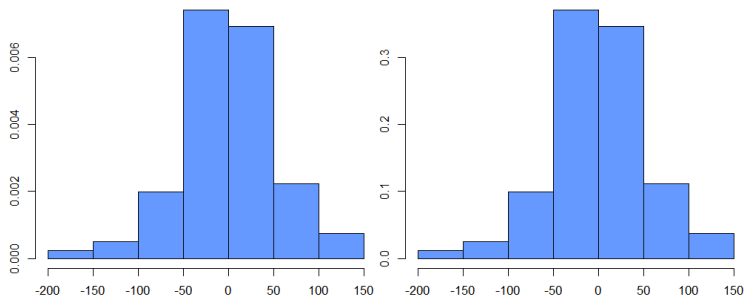




### 2.3.3. Histogramm und empirische Dichtefunktion

Klassierte Daten kann man übersichtlich in einem Histogramm darstellen. Dabei repräsentiert jeder Balken die absoluten Klassenhäufigkeiten der entsprechenden Klasse.

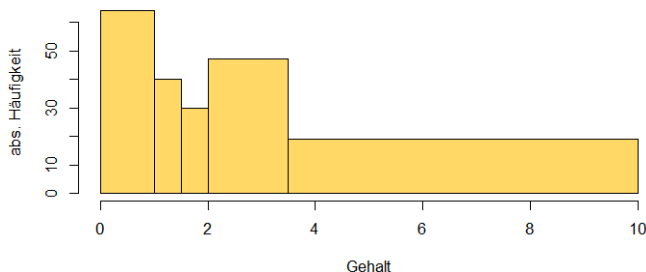
■ **Beispiel B2.27:** Tagesgewinne/-verluste des DAX vom 1. Januar bis 27. April 2011, in Punkten (Quelle: yahoo.com)



⚠ Wenn die Klassen nicht alle gleich groß sind, ist es nicht ratsam in Histogrammen absolute oder relative Häufigkeiten anzugeben.

■ **Beispiel B2.28:** 200 Besucher eines Einkaufszentrums werden befragt, über wieviel Geld sie im Monat verfügen (Nettogehalt). Die Befragung ergibt folgende Zahlen:

Klasse	$n(K)$
0-1000	64
1000-1500	40
1500-2000	30
2000-3500	47
3500- $\infty$	19



Die 70 Befragten mit Gehältern zwischen 1000 und 2000 Euro und die 19 Befragten über 3500 Euro scheinen in der Grafik unter- bzw. überrepräsentiert.

Die empirische Dichtefunktion ist im Falle von Klassenbildung mit Klassen  $K_i = (a_i, b_i]$  definiert als

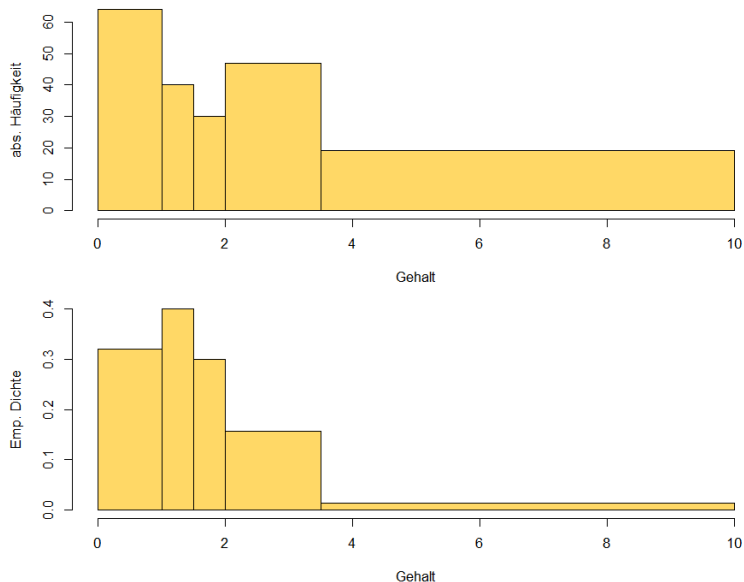
$$f_n(x) = \frac{h(K_i)}{b_i - a_i}, \quad x \in K_i. \quad (2.4)$$

Vorteil: Im Balkendiagramm ist die Gesamtfläche der Balken stets eins.

Im Diagramm entspricht nun die Balkenfläche der (geschätzten) Wahrscheinlichkeit dafür, dass das Merkmal einen Wert in der entsprechenden Klasse annimmt.

- Bei klassierten Daten mit unterschiedlich großen Klassen besser geeignet als das Standardhistogramm!

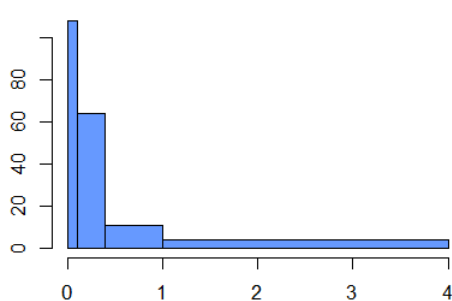
■ **Beispiel B2.29**  $\Rightarrow$  B2.28: Vergleich des klassischen Histogramms mit dem Diagramm für die empirische Dichte:



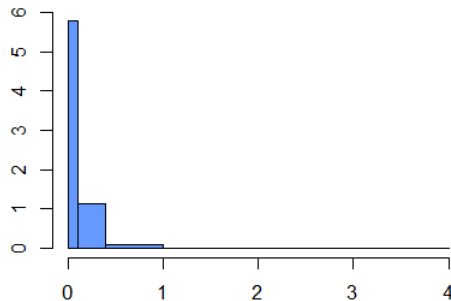
■ **Beispiel B2.30:** Einwohnerzahl 187 deutscher Städte am 31.12.2015 (Quelle: <http://www.citypopulation.de>, Angaben in Mill. Einwohnern). Wir definieren folgende Klassen der Form  $(a, b]$  (in Mill. Einw.):

$i$	$a$	$b$	$n(K_i)$	$h(K_i)$	$f_n(K_i)$
1	0	0.1	108	0.578	5.775
2	0.1	0.4	64	0.342	1.141
3	0.4	1.0	11	0.059	0.098
4	1.0	4.0	4	0.021	0.007

Es ergeben sich folgende Diagramme:



Histogramm



Emp. Dichte

## 2.4. Lagemaße

- Lagemaße sind im Allgemeinen für intervall- und verhältnisskalierte Daten (sog. metrische Daten) definiert.
- Lagemaße sollen einen ersten Eindruck über die „durchschnittliche Lage“ der Daten geben.

### 2.4.1. Arithmetisches Mittel

Das arithmetische Mittel (häufig einfach „Mittelwert“) einer Stichprobe  $x_1, x_2, \dots, x_n$  ist definiert als

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

- Das arithmetische Mittel ist eine gewichtete Summe mit jeweils identischen Gewichten  $1/n$ .
- Das arithmetische Mittel ist linear:

$$\overline{ax + b} = a\bar{x} + b, \quad a, b \in \mathbb{R}.$$

Speziell gelten die Identitäten

$$\begin{aligned} \overline{a\bar{x}} &= a\bar{x} \\ \text{und} \quad \overline{\bar{x} + \bar{y}} &= \bar{x} + \bar{y}. \end{aligned}$$

Beide Eigenschaften sind mehr oder weniger offensichtlich (Beweis in der Übung).

- Warnung: Es gilt i.A. keineswegs  $\overline{f(x)} = f(\bar{x})$ , z.B. ist  $\overline{(x^2)} \neq (\bar{x})^2$ .

■ **Beispiel B2.31**  $\Rightarrow$  B2.18: Ein Würfel wird  $n = 5$  Mal geworfen:

$$x_1 = 3, x_2 = 6, x_3 = 1, x_4 = 5, x_5 = 6.$$

Dann ergibt sich

$$\bar{x} = \frac{3 + 6 + 1 + 5 + 6}{5} = \frac{21}{5}.$$

Außerdem berechnet man leicht, dass

$$\overline{(x^2)} = \frac{9 + 36 + 1 + 25 + 36}{5} = \frac{107}{5} = 21.4$$

$$\text{aber} \quad (\bar{x})^2 = \left(\frac{21}{5}\right)^2 = \frac{441}{25} = 17.64$$

gilt.



- Die Summe der Abweichungen vom Mittelwert ist null:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

- Das arithmetische Mittel minimiert das mittlere Abweichungsquadrat:

$$\sum_{i=1}^n (x_i - c)^2$$

- Alternative Formeln:

$$\bar{x} = \frac{1}{n} \sum_{m \in M_X} m \cdot n(m)$$

oder auch 
$$\bar{x} = \sum_{m \in M_X} m \cdot h(m)$$

Vorteile des arithmetischen Mittels als Lagemaß:

- ⊕ Intuitive Formel, die leicht zu berechnen ist.

Nachteile:

- ⊖ Das arithmetische Mittel ist nicht robust, sondern reagiert empfindlich auf Ausreißer (s. Übung).
- ⊖ Manchmal ist die Interpretation als Mittelwert fragwürdig (s. geometrisches Mittel [2.4.8](#)).

### 2.4.2. Arithmetisches Mittel für klassierte Daten

Angenommen die Daten liegen in reduzierter Form in Klassen  $K_1, K_2, \dots, K_n$  vor. Dabei seien  $\mu_1, \mu_2, \dots, \mu_n$  die entsprechenden Klassenmittelwerte (z.B. die Intervallmitten).

Dann berechnen wir als arithmetisches Mittel

$$\bar{x} = \sum_{i=1}^n h(K_i) \cdot \mu_i.$$

- Offenbar haben wir dabei implizit vorausgesetzt, dass die Daten in ihren Klassen gleichverteilt sind.
- Der so ermittelte Mittelwert stimmt nicht mit dem arithmetischen Mittel der unklassierten Originaldaten überein.

### 2.4.3. Arithmetisches Mittel für gepoolte Daten

Angenommen es liegen mehrere Stichproben

Stichprobe 1:  $x_{11}, x_{12}, \dots, x_{1n_1}$

Stichprobe 2:  $x_{21}, x_{22}, \dots, x_{2n_2}$

$\vdots \quad \quad \quad \vdots$

Stichprobe m:  $x_{m1}, x_{m2}, \dots, x_{mn_m}$

mit verschiedenen Mittelwerten  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$  vor.

Dann kann man den Mittelwert der gepoolten Daten  $x_{11}, x_{21}, \dots, x_{mn_m}$  einfach berechnen, ohne die Daten selbst zu kennen:

$$\bar{x} = \sum_{k=1}^m \frac{\bar{x}_k \cdot n_k}{n}$$

(gepoolter Mittelwert).

Spezialfall: Möchte man zu einer Stichprobe

$$x_1, x_2, \dots, x_n$$

einen weiteren Datenpunkt  $x_{n+1}$  hinzufügen, so ergibt sich

$$\bar{x}_{neu} = \frac{n \cdot \bar{x}_{alt} + x_{n+1}}{n + 1} \quad (2.5)$$

als der neue Mittelwert.

Man erkennt, dass für sehr große Werte von  $n$  etwa

$$\bar{x}_{neu} \approx \bar{x}_{alt} + \frac{x_{n+1}}{n}$$

gilt, d.h. die Änderung des Mittelwertes ist etwa von der Größenordnung  $x_{n+1}/n$ .

### 2.4.4. Die Ordnungsstatistik

Gegeben seien ordinalskalierte Daten

$$x_1, x_2, \dots, x_n.$$

Als Ordnungsstatistik bezeichnet man die in aufsteigender Größe angeordneten Daten

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Dann ist z.B.

$$x_{(1)} = \min\{x_1, x_2, \dots, x_n\},$$

$$x_{(n)} = \max\{x_1, x_2, \dots, x_n\}.$$

### 2.4.5. Getrimmtes Mittel

Das arithmetische Mittel ist anfällig für Ausreißer. Das getrimmte Mittel ignoriert die  $\lfloor \alpha n \rfloor$  größten und kleinsten Beobachtungen:

$$\bar{X}_{(\alpha)} = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} X_{(i)}.$$

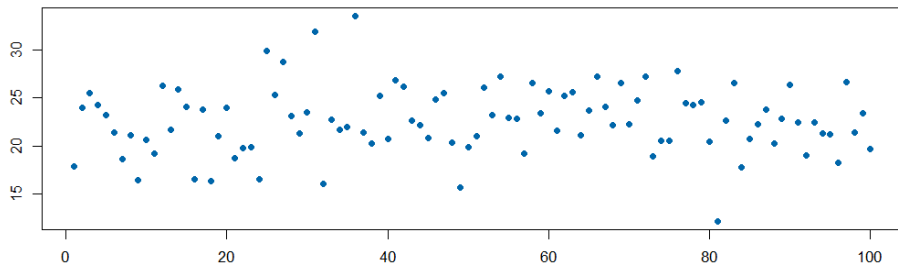
Vorteile:

⊕ Robust gegen Ausreißer.

Nachteile:

- ⊖ Einige Datenpunkte werden nicht verwendet.
- ⊖ Wahl von  $\alpha$  beliebig. Missbrauch möglich.

■ **Beispiel B2.32:** Dreiig Jahre lang wurde an einem Ort die Tageshchsttemperatur am 1. September gemessen:

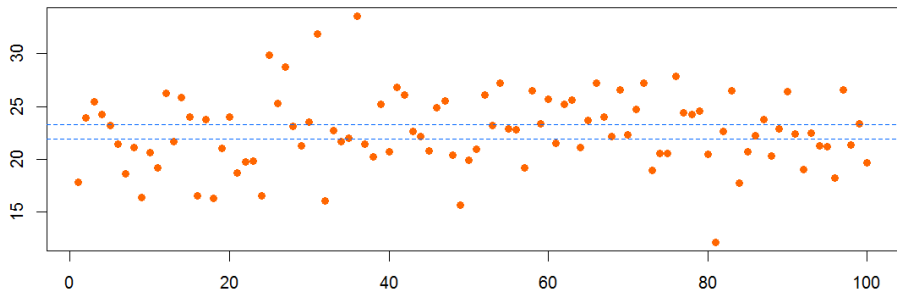


Es ergibt sich ein arithmetisches Mittel von

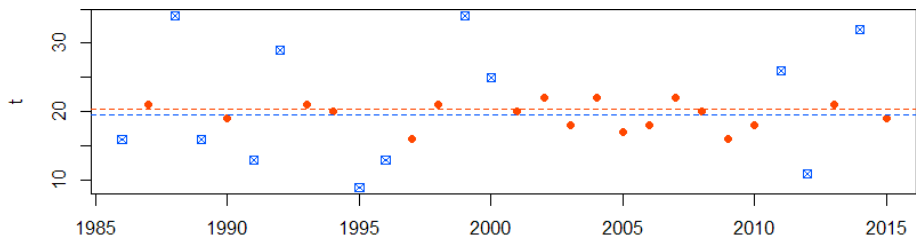
$$\bar{t} = 20.3^{\circ}\text{C}$$



Wir wählen  $\alpha = 0.1$



und  $\alpha = 0.2$ :



### 2.4.6. Median

Der (empirische) Median ist die kleinste Zahl  $\tilde{x}$ , für die mindestens die Hälfte der Beobachtungen  $\geq \tilde{x}$  ist und die andere Hälfte  $\leq \tilde{x}$  ist.

Genaue Definition:

$$\tilde{x} = \text{med}(x) = \begin{cases} x_{(\lfloor n/2 \rfloor + 1)} & ; n/2 \notin \mathbb{N} \\ \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) & ; n/2 \in \mathbb{N} \end{cases}$$

- Der Median minimiert den Abstand  $\sum_{i=1}^n |x_i - c|$ .

Vorteile des Median:

- ⊕ Robust gegen Ausreißer

Nachteile des Median:

- ⊖ Nicht alle Datenpunkte werden berücksichtigt.

■ **Beispiel B2.33**  $\Rightarrow$  B2.32: Ordnungsstatistik der Temperaturen:

9, 11, 13, 13, 16, 16, 16, 16, 17, 18,  
18, 18, 19, 19, **20, 20**, 20, 21, 21, 21,  
21, 22, 22, 22, 25, 26, 29, 32, 34, 34.

Da  $n = 30$  ist ergibt sich  $n/2 \in \mathbb{N}$ , also ist

$$\tilde{x} = \frac{x_{(15)} + x_{(16)}}{2} = \frac{20 + 20}{2} = 20.$$

■ **Beispiel B2.34**  $\Rightarrow$  B2.18: Ein Würfel wird  $n = 5$  Mal geworfen:

$$x_1 = 3, \quad x_2 = 6, \quad x_3 = 1, \quad x_4 = 5, \quad x_5 = 6.$$

Da  $n/2 \notin \mathbb{N}$  ergibt sich für den Median

$$\tilde{x} = x_{(3)} = 5.$$

### 2.4.7. Quantile und Quartile

Das  $\alpha$ -Quantil ist die kleinste Zahl  $\tilde{x}_\alpha$  für die mindestens  $\alpha n$  der Daten  $\leq \tilde{x}_\alpha$  sind:

$$\tilde{x}_\alpha = \begin{cases} x_{(\lfloor \alpha n \rfloor + 1)} & ; \alpha n \notin \mathbb{N} \\ \frac{1}{2} (x_{(\alpha n)} + x_{(\alpha n + 1)}) & ; \alpha n \in \mathbb{N} \end{cases}$$

- Der Median ist das 50%-Quantil.
- Die 25%- und 75%-Quantile heißen auch unteres und oberes Quartil.

■ **Beispiel B2.35**  $\Rightarrow$  B2.32: Ordnungsstatistik der Temperaturen:

9, 11, 13, 13, 16, 16, 16, **16**, 17, 18,  
 18, 18, 19, 19, 20, 20, 20, 21, 21, 21,  
 21, 22, 22, 22, 25, 26, 29, 32, 34, 34.

Dann ergibt sich für das untere Quartil

$$\tilde{x}_{0.25} = x_{(\lfloor 7.5 \rfloor + 1)} = x_{(8)} = 16.$$

### 2.4.8. Das geometrische Mittel

■ **Beispiel B2.36:** Ein Aktienindex steigt in drei Jahren zunächst um 15%, dann um 21% und sinkt schließlich um 12%. Wie groß ist das durchschnittliche Wachstum?

Insgesamt steigt der Index um den Faktor  $1.15 \cdot 1.21 \cdot 0.92 = 1.22452$ , also um knapp 22%.

Wie hoch müsste das Wachstum im Durchschnitt jährlich sein, um in drei Jahren insgesamt auf den Faktor 1.22452 zu kommen?

Wir suchen eine Lösung der Gleichung

$$x^3 = 1.22452,$$

also  $x = \sqrt[3]{1.22452} = 1.069848$ , das mittlere Wachstum beträgt also knapp 7%.

Das geometrische Mittel verwendet man, um Mittelwerte von relativen Wachstumszahlen zu berechnen:

$$\bar{x}_g = \sqrt[n]{\prod_{k=1}^n x_k}.$$

Liegen die Daten nahe bei eins, so gilt die Schätzung

$$\bar{x}_g \approx \bar{x}.$$

■ **Beispiel B2.37:** Es sei

$$x_1 = 1.1, \quad x_2 = 1.03, \quad x_3 = 0.99, \quad x_4 = 1.07.$$

Dann ist

$$\bar{x} = 1.0475, \quad \bar{x}_g = 1.046676.$$

### 2.4.9. Weitere Mittelwerte

Das harmonische Mittel ist gegeben durch die Formel

$$\bar{x}_h = \left( \frac{1}{n} \sum_{k=1}^n \frac{1}{x_k} \right)^{-1}.$$

Es entspricht also dem Kehrwert des arithmetischen Mittels der Datenkehrwerte.

■ **Beispiel B2.38:** Drei Autos legen eine Strecke von 100 km mit unterschiedlichen Geschwindigkeiten zurück (100 km/h, 150 km/h und 200 km/h). Wie ist ihre Durchschnittsgeschwindigkeit?

$$\bar{v}_h = \frac{300}{\frac{100}{100} + \frac{100}{150} + \frac{100}{200}} = \left( \frac{\frac{1}{100} + \frac{1}{150} + \frac{1}{200}}{3} \right)^{-1} = 138.4615 \text{ km/h.}$$



Der Modalwert (Modus)  $x_m$  ist bei diskreten Merkmalen die in der Stichprobe am häufigsten vorkommende Beobachtung. Bei klassierten Daten wählt man die Mitte der Klasse mit den meisten Beobachtungen.

- Der Modalwert ist nicht eindeutig.
- Modus und arithmetisches Mittel müssen keinesfalls nahe beieinander liegen.

■ **Beispiel B2.39**  $\Rightarrow$  **B2.32**: Im Beispiel **B2.32** wurden 30 Jahre lang Temperaturen gemessen:

9, 11, 13, 13, **16, 16, 16, 16**, 17, 18,  
18, 18, 19, 19, 20, 20, 20, **21, 21, 21**,  
**21**, 22, 22, 22, 25, 26, 29, 32, 34, 34.

Sowohl 16 als auch 21 sind Modi.

## 2.5. Streuungsmaße

In der Aufgabe 12 zeigte sich, dass sehr unterschiedliche Datensätze denselben Mittelwert aufweisen können. Um Daten adäquat mit wenigen Kennzahlen zu beschreiben, benötigen wir mindestens noch ein weiteres Maß für die Streuung der Daten um den Mittelwert.

### 2.5.1. Varianz und Standardabweichung

Die empirische Varianz ist durch

$$\hat{\sigma}_*^2(x) = \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n}$$

definiert, also durch die mittlere quadratische Abweichung der Datenpunkte von ihrem Mittelwert.

- $\hat{\sigma}_*^2(x)$  ist immer nicht-negativ und null nur dann, wenn alle  $x_k$  gleich sind.
- Wie schon im Falle des Mittelwerts gibt es eine oftmals kürzere Variante, die mit Hilfe der relativen Häufigkeiten formuliert wird:

$$\hat{\sigma}_*^2(x) = \frac{1}{n} \sum_{m \in M_X} (m - \bar{x})^2 \cdot n(m).$$

- Meistens ist folgende alternative Formel leichter zu berechnen:

$$\hat{\sigma}_*^2(x) = \overline{(x^2)} - (\bar{x})^2.$$

- Die emp. Varianz ist nicht linear, aber es gilt aber

$$\hat{\sigma}_*^2(ax + b) = a^2 \hat{\sigma}_*^2(x).$$

Speziell ist die Varianz translationsinvariant.

Die Standardabweichung ist definiert als

$$\hat{\sigma}_*(x) = \sqrt{\hat{\sigma}_*^2(x)}.$$

- Die Standardabweichung hat dieselbe Einheit, wie die Originaldaten.
- Es gilt die einprägsame Formel  $\hat{\sigma}_*(ax + b) = a\hat{\sigma}_*(x)$ .

Vorteile und Nachteile der Varianz (Standardabweichung) als Streuungsmaß:

- ⊕ Einleuchtende Interpretation.
- ⊕ Leicht zu berechnen und mathematisch handhabbar.
- ⊖ Anwendbar nur bei hinlänglich symmetrischen und möglichst „eingipfeligen“ Verteilungen der Daten.
- ⊖ Die emp. Varianz und die Standardabweichung reagieren empfindlich auf Ausreißer.

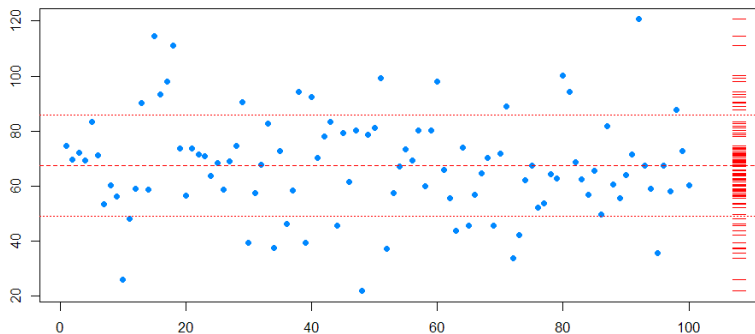
In der Statistik benötigt man neben der oben beschriebenen empirischen Varianz noch die Stichprobenvarianz (korrigierte Varianz) und die Stichprobenstandardabweichung (korrigierte Standardabweichung):

$$\hat{\sigma}^2(x) = \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n-1},$$
$$\hat{\sigma}(x) = \sqrt{\hat{\sigma}^2(x)}.$$

- Es gilt offenbar

$$\hat{\sigma}^2(x) = \frac{n}{n-1} \hat{\sigma}_*^2(x).$$

- Die Stichprobenvarianten der Varianz und der Standardabweichung werden in der Schätztheorie verwendet, weil sie sog. erwartungstreue Schätzer liefern.
- Für große Werte von  $n$  sind beide Varianten etwa gleich.

**Beispiel B2.40:**

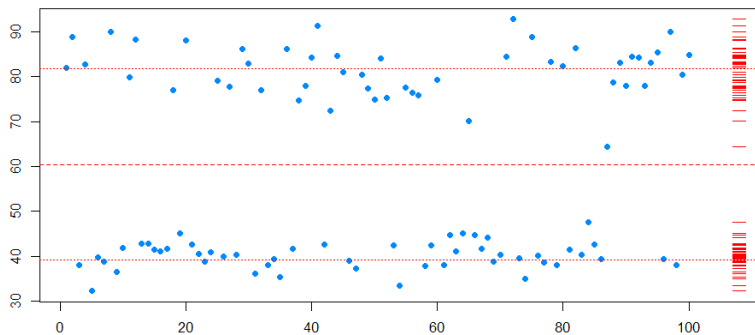
$$\bar{x} = 67.73633,$$

$$\hat{\sigma}^2(x) = 472.267,$$

$$\hat{\sigma}(x) = 21.73171,$$

$$F_n(\bar{x} + \hat{\sigma}_*(x)) - F_n(\bar{x} - \hat{\sigma}_*(x)) = 0.7$$

## ■ Beispiel B2.41:



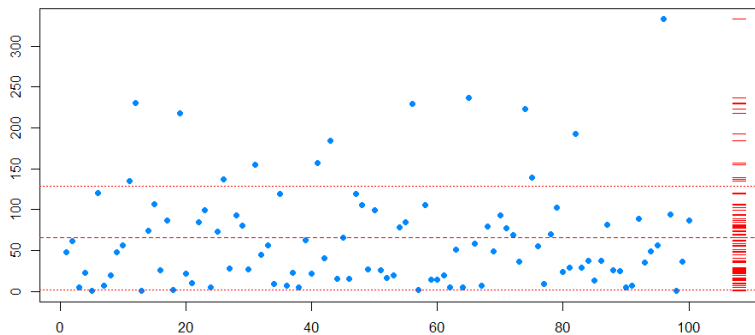
$$\bar{x} = 60.44387$$

$$\hat{\sigma}^2(x) = 452.3576,$$

$$\hat{\sigma}(x) = 21.2687,$$

$$F_n(\bar{x} + \hat{\sigma}_*(x)) - F_n(\bar{x} - \hat{\sigma}_*(x)) = 0.56$$

## ■ Beispiel B2.42:



$$\bar{x} = 65.37265$$

$$\hat{\sigma}^2(x) = 4082.81,$$

$$\hat{\sigma}(x) = 63.89687,$$

$$F_n(\bar{x} + \hat{\sigma}_*(x)) - F_n(\bar{x} - \hat{\sigma}_*(x)) = 0.84$$



## 2.5.2. Varianz für gepoolte Daten (Varianzzerlegung)

Bei mehreren Stichproben

Stichprobe 1:  $x_{11}, x_{12}, \dots, x_{1n_1}$

Stichprobe 2:  $x_{21}, x_{22}, \dots, x_{2n_2}$

$\vdots \quad \quad \quad \vdots$

Stichprobe m:  $x_{m1}, x_{m2}, \dots, x_{mn_m}$

mit verschiedenen Mittelwerten  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$  und Varianzen  $\hat{\sigma}_*^2(x_1), \hat{\sigma}_*^2(x_2), \dots, \hat{\sigma}_*^2(x_m)$  ergibt sich

$$\hat{\sigma}_*^2(x) = \underbrace{\sum_{k=1}^m \frac{\hat{\sigma}_*^2(x_k) \cdot n_k}{n}}_{\text{interne Varianz}} + \underbrace{\sum_{k=1}^m \frac{(\bar{x}_k - \bar{x})^2 \cdot n_k}{n}}_{\text{externe Varianz}}.$$

(Varianzzerlegung).

■ **Beispiel B2.43:** Gegeben seien die Stichproben

	$x_{ki}$	$n_k$	$\bar{x}_k$	$\hat{\sigma}_*^2(x_k)$
1	1,3,2,5,4	5	3.0	2.0
2	5,5,5	3	5.0	0.0
3	6,1,4,5	4	4.0	3.5

Gepoolter Mittelwert:  $\bar{x} = \frac{5 \cdot 3 + 3 \cdot 5 + 4 \cdot 4}{5 + 3 + 4} = 3.8\bar{3}$ .

Interne Varianz:  $\sum_{k=1}^m \frac{\hat{\sigma}_*^2(x_k) \cdot n_k}{n} = 2.0$

Externe Varianz:  $\sum_{k=1}^m \frac{(\bar{x}_k - \bar{x})^2 \cdot n_k}{n} = 0.63\bar{8}$ .

Varianz:  $\hat{\sigma}_*^2(x) = 2 + 0.63 = 2.63\bar{8}$ .

### 2.5.3. Spannweite und Interquartilsabstand

Als Spannweite bezeichnet man den Abstand zwischen Minimum und Maximum der Stichprobe:

$$R_x = x_{(n)} - x_{(1)}.$$

- ⊖ Nur wenige Daten fließen in die Berechnung ein.
- ⊖ Offenbar ist die Spannweite nicht robust gegenüber Ausreißern.

Der Interquartilsabstand misst den Abstand zwischen oberem und unterem Quartil:

$$\text{IQR}_x = \tilde{x}_o - \tilde{x}_u.$$

- ⊕ Robust in Bezug auf Ausreißer.

### 2.5.4. Variationskoeffizient

Der Variationskoeffizient setzt die durch die Standardabweichung gemessene Streuung ins Verhältnis zu ihrem Mittelwert:

$$V(x) = \frac{\hat{\sigma}_*(x)}{\bar{x}}$$

- Relatives Streuungsmaß
- Definiert für positive metrische Daten.
- Es gilt  $0 \leq V(x) \leq \sqrt{n}$ . Daher definiert man den normierten Variationskoeffizienten

$$V^*(x) = \frac{\hat{\sigma}_*(x)}{\sqrt{n} \cdot \bar{x}}$$

mit Werten im Intervall  $[0, 1]$ .

### 2.5.5. Weitere Streuungsmaße

Der Median der absoluten Abweichungen (MAD)

$$\text{MAD}_x = \text{med}(|x - \tilde{x}|)$$

ist unempfindlich in Bezug auf Ausreißer (viele Varianten).

Die mittlere absolute Abweichung vom Mittel

$$\overline{|x - \bar{x}|}$$

und die mittlere absolute Abweichungen vom Median

$$\overline{|x - \tilde{x}|}$$

sind weniger robust.

■ **Beispiel B2.44**  $\Rightarrow$  B2.18: Für sechs Monate wird die Anzahl der Unfälle an einer befahrenen Ausfahrtstraße in einer Statistik erfasst:

$$x_1 = 5, x_2 = 1, x_3 = 3, x_4 = 2, x_5 = 1, x_6 = 6$$

Es ist  $\bar{x} = 18/6 = 3$  und daher

$$\begin{aligned}\hat{\sigma}_*^2(x) &= \frac{(5-3)^2 + (1-3)^2 + \dots + (6-3)^2}{6} \\ &= \frac{4 + 4 + 0 + 1 + 4 + 9}{6} = \frac{22}{6} = 3.\bar{3}.\end{aligned}$$

Alternative Formel:

$$\begin{aligned}\hat{\sigma}_*^2(x) &= \overline{x^2} - (\bar{x})^2 \\ &= \frac{5^2 + 1^2 + 3^2 + 2^2 + 1^2 + 6^2}{6} - 3^2 \\ &= \frac{76}{6} - 9 = \frac{22}{6} = 3.\bar{3}.\end{aligned}$$

Für die Standardabweichung ergibt sich

$$\hat{\sigma}_*(x) = \sqrt{\hat{\sigma}_*^2(x)} \approx 1.92$$

Die Stichprobenvarianz ist entsprechend etwas größer als die empirische Varianz:

$$\hat{\sigma}^2(x) = \frac{n}{n-1} \cdot \hat{\sigma}_*^2(x) = \frac{22}{5} = 4.4$$

Dementsprechend ist

$$\hat{\sigma}(x) = \sqrt{4.4} \approx 2.1$$

Die Spannweite der Daten ist offenbar

$$R_x = 6 - 1 = 5.$$



Zur Berechnung des Interquartilabstands benötigen wir das untere und das obere Quartil. Es ist

$$x_{(1)} = 1, x_{(2)} = 1, x_{(3)} = 2, x_{(4)} = 3, x_{(5)} = 5, x_{(6)} = 6$$

Also ergibt sich

$$\begin{aligned}\widetilde{x}_{0.25} &= x_{(\lfloor 6/4 \rfloor + 1)} = x_{(2)} = 1, \\ \widetilde{x}_{0.75} &= x_{(\lfloor 18/4 \rfloor + 1)} = x_{(5)} = 5.\end{aligned}$$

Dann erhalten wir

$$\text{IQR}_x = 5 - 1 = 4.$$

Variationskoeffizient:

$$V(x) = \frac{\hat{\sigma}_*(x)}{\bar{x}} = \frac{\sqrt{22/6}}{3} \approx 0.64$$
$$V^*(x) = \frac{V(x)}{\sqrt{6}} \approx 0.26$$

MAD:

$$\text{MAD}_x = \text{med}(2.5, 1.5, 0.5, 0.5, 1.5, 3.5) = 1.5$$

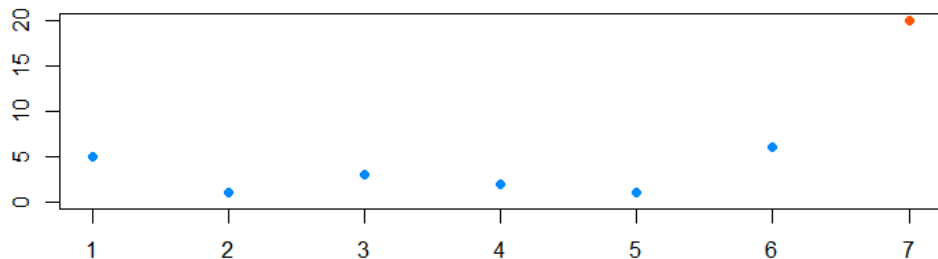
Mittlere absolute Abweichung vom Mittel:

$$\overline{|x - \bar{x}|} = \overline{(2, 2, 0, 1, 2, 3)} = \frac{10}{6} \approx 1.67$$

Mittlere absolute Abweichungen vom Median ( $\tilde{x} = 2.5$ ):

$$\overline{|x - \tilde{x}|} = \overline{(2.5, 1.5, 0.5, 0.5, 1.5, 3.5)} = \frac{10}{6}$$

Im siebten Monat geschehen 20 Unfälle.



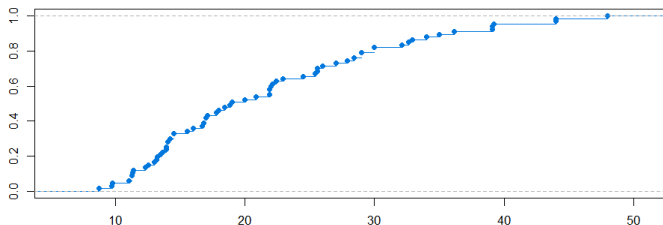
Nun ergibt sich:

	Alt	Neu
$\hat{\sigma}_*^2(x)$	3.67	38.53
$\hat{\sigma}_*(x)$	1.91	6.21
$R_x$	5	19
$IQR_x$	4	5
$MAD_x$	1.5	2

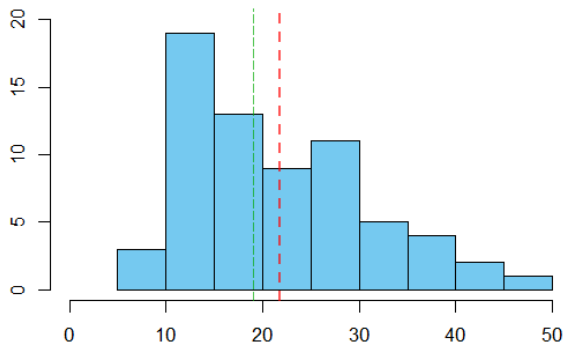
■ **Beispiel B2.45:** IT-Unternehmen in Österreich mit mehr als 99 Mitarbeitern (Quelle:<http://data.opendataportal.at>)

	Name	Umsatz	Mitarbeiter
1	A1 Telekom Austria AG	256	16240
2	Raiffeisen Informatik GmbH	172	3000
3	KAPSCH Group	361	5250

Wir betrachten die Umsatzwerte für 67 Firmen mit weniger als 50 Mio Euro Umsatz.



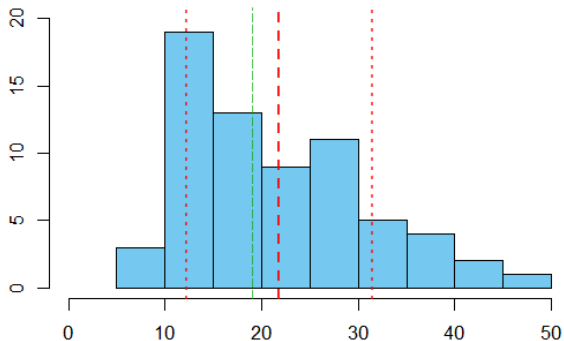
Histogramm:



Arithmetisches Mittel und Median:

$$\bar{U} = 21.8394$$

$$\tilde{U} = 19.$$



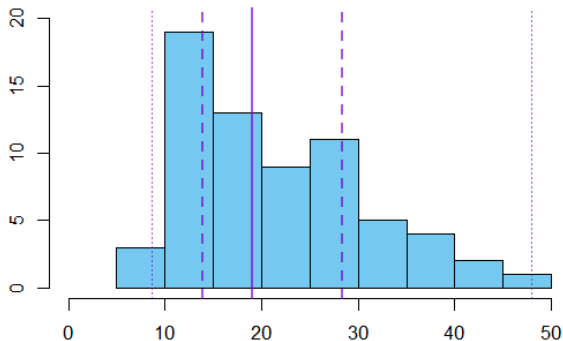
Varianz, Standardabweichung:

$$\hat{\sigma}_*^2(U) = 90.90$$

$$\hat{\sigma}_*(U) = 9.53$$

$$\hat{\sigma}^2(U) = 92.28$$

$$\hat{\sigma}(U) = 9.61$$



Quartile:

0%	25%	50%	75%	100%
8.70	13.93	19.00	28.40	48.00

Spannweite und Interquartilsabstand:

$$R_U = 48 - 8.7 = 39.3$$

$$\text{IQR}_U = 28.4 - 13.93 = 14.47$$

## 2.6. Boxplots

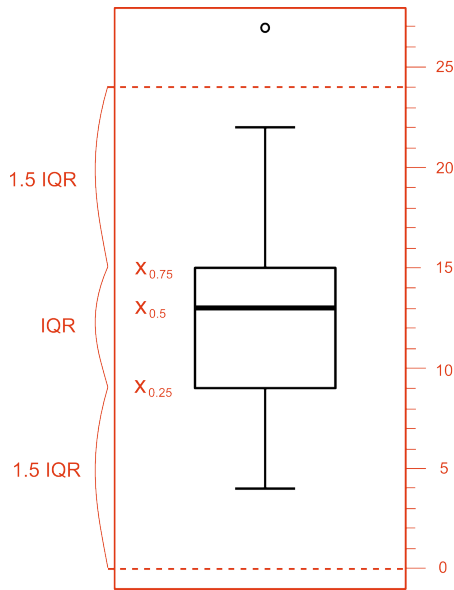
In einem Boxplot werden die wichtigsten Lage- und Streuungsmaße grafisch zusammengefasst.

Vorgehensweise:

- Eine horizontale Linie wird auf der Höhe des Median eingezeichnet.
- Das obere und untere Quartil bestimmen die obere und untere Seite der „Box“.
- Die Länge der beiden Antennen (Whiskers) entspricht maximal dem 1.5-fachen des IQR (gerechnet vom oberen- bzw. unteren Quartil aus). Die Antennen enden aber beim letzten tatsächlich vorliegenden Datenwert unter- bzw. oberhalb dieser Marke.
- Alle Datenpunkte außerhalb der Antennen werden als Ausreißer als Punkte eingezeichnet.

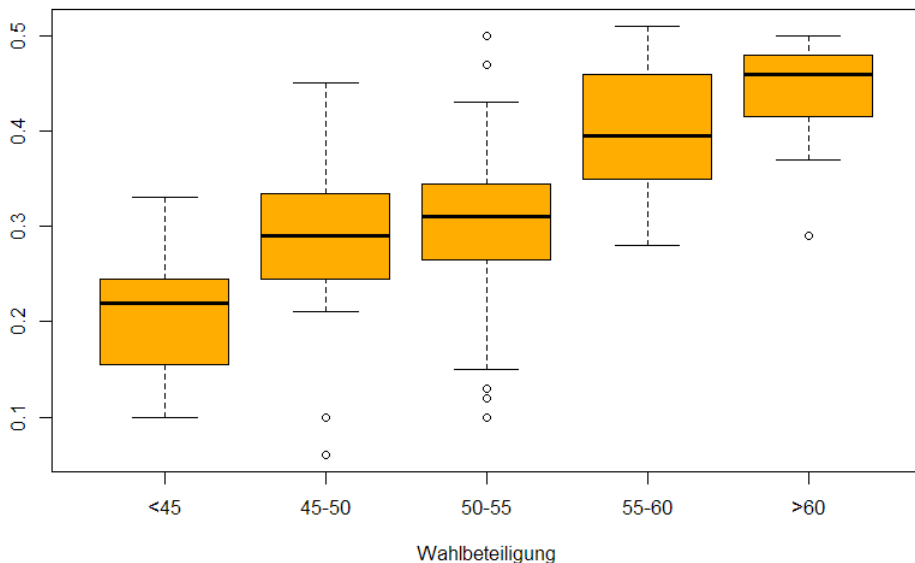


Beispiel:  $x = (4, 7, 9, 11, 12, 14, 14, 15, 22, 27)$ . Hier ist  $n = 10$ ,  $\tilde{x} = 13$ ,  $\tilde{x}_u = 9$ ,  $\tilde{x}_o = 15$ ,  $IQR_x = 6$  und  $1.5 \cdot IQR = 9$ .



**■ Beispiel B2.46:** Bürgerschaftswahlen in Hamburg (2009)

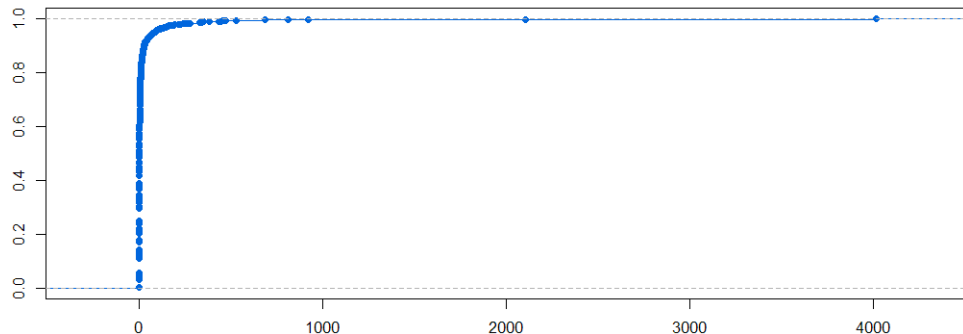
Stimmanteile für die CDU in den Wahllokalen



## 2.7. Konzentrationsmaße

### 2.7.1. Die Lorenz-Kurve

■ **Beispiel B2.47**  $\Rightarrow$  **B2.45**: Umsatz und Mitarbeiterzahl von österreichischen IT-Unternehmen. Empirische Verteilungsfunktion für die Umsätze im Beispiel **B2.45**:



- Ein relativ großer Teil der Umsatzgesamtsumme entfällt auf wenige Firmen (sog. Konzentration).

Um eine solche Konzentration grafisch darzustellen, verwendet man häufig die Lorenz-Kurve.

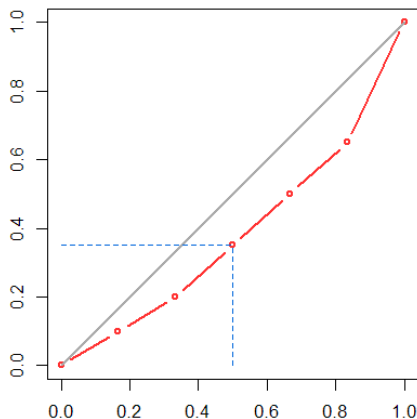
- Berechne zunächst für  $i = 1, 2, \dots, n$  die Werte

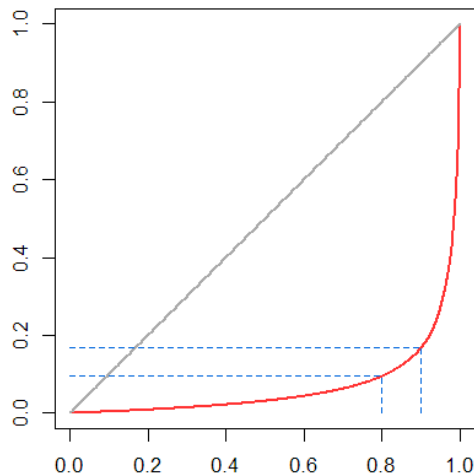
$$\begin{aligned} L_i &= \frac{\text{Summe der kleinsten } i \text{ Umsätze}}{\text{Gesamtsumme der Umsätze}} \\ &= \frac{\sum_{k=1}^i X_{(k)}}{\sum_{k=1}^n X_{(k)}}. \end{aligned}$$

- Interpretation:  $100 \cdot i/n$  Prozent der kleinsten Beobachtungen machen in der Summe  $100 \cdot L_i$  Prozent der Gesamtsumme der Beobachtungen aus.
- Zeichne dann eine Kurve, die im Einheitsquadrat die Punkte  $(i/n, L_i)$  miteinander verbindet (Polygonzug)

■ **Beispiel B2.48:** Sechs Mitarbeiter einer Firma haben folgende jährliche Gehälter (in tsd. Euro):

Gehalt:	30	20	30	70	30	20
Orderst.:	20	20	30	30	30	70
$L_i$ :	0.1	0.2	0.35	0.5	0.65	1.0
$i/n$ :	1/6	2/6	3/6	4/6	5/6	6/6



**■ Beispiel B2.49**  $\Rightarrow$  B2.45:

Interpretation:

- Auf die oberen 20% der Firmen entfallen etwa 90% der Umsätze

### 2.7.2. Das Gini-Maß

- Um eine Konzentration auch quantitativ zu erfassen, kann man das Gini-Maß berechnen:

$$G_x = \frac{\sum_{i=1}^n (2i-1)x_{(i)}}{n^2 \bar{x}} - 1.$$

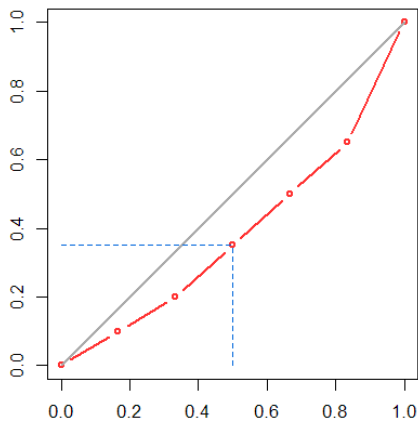
Das Gini-Maß entspricht der doppelten Fläche zwischen der Lorenz-Kurve und der Winkelhalbierenden.

- Je größer  $G_x$  ausfällt, desto größer ist die Konzentration.
- Es gilt  $0 \leq G_x \leq (n-1)/n$ , daher berechnet man auch das normierte Gini-Maß

$$G_x^* = \frac{n}{n-1} \cdot G_x.$$

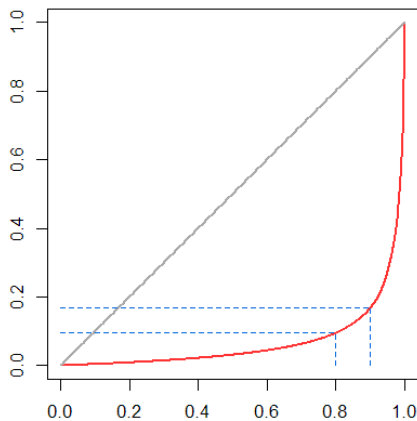
mit Werten im Intervall  $[0, 1]$ .

■ **Beispiel B2.50**  $\Rightarrow B2.48 \& B2.45$ :



$$G_x = 0.23,$$

$$G_x^* = 0.28.$$



$$G_x = 0.8645807,$$

$$G_x^* = 0.8654585.$$



## 2.8. Bivariate Daten

Häufig interessiert man sich in der Statistik gleichzeitig für mehrere Merkmale. Insbesondere versucht man etwas über die Abhängigkeit der Merkmale untereinander herauszufinden. Wir beschäftigen uns in diesem Paragraphen mit der Statistik bivariater Daten, also mit dem Fall zweier Merkmale.

Seien im Folgenden  $X$  und  $Y$  zwei Merkmale (definiert als Funktionen auf demselben Stichprobenraum/derselben Grundgesamtheit).

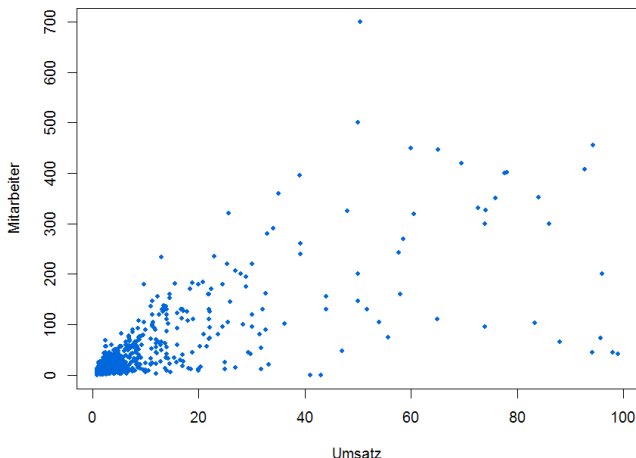
Die entsprechenden Merkmalsausprägungen seien

$$M_X = \{a_1, a_2, \dots\}$$

$$M_Y = \{b_1, b_2, \dots\}.$$

Bivariate Daten lassen sich besonders einfach im Streudiagramm darstellen.

■ **Beispiel B2.51**  $\Rightarrow$  B2.45: Umsatz und Mitarbeiterzahl von österreichischen IT-Unternehmen mit weniger als 100 Mill. Euro Umsatz (Quelle: <http://data.opendataportal.at>).



### 2.8.1. Häufigkeiten und Kontingenztabellen

Wir betrachten jetzt Stichproben der Form  $(x_i, y_i)$ , genauer

$$\{(x_i, y_i), i = 1, 2, \dots, n, X_i \in M_X, y_i \in M_Y\}.$$

Wie schon bei den univariaten Daten definieren wir die absolute bivariate Häufigkeit der Ausprägung  $(a_i, b_j)$ ..:

$$n_{ij} = n(a_i, b_j) = \#\{k : x_k = a_i, y_k = b_j\}.$$

Als absolute Randhäufigkeit bezeichnen wir die Werte

$$n_{i\bullet} = \#\{k : x_k = a_i\},$$

$$n_{\bullet j} = \#\{k : y_k = b_j\}.$$

Entsprechend ist

$$h_{ij} = \frac{n_{ij}}{n}$$

die relative bivariate Häufigkeit der Ausprägung  $(a_i, b_j)$  und

$$h_{i\bullet} = \frac{n_{i\bullet}}{n},$$
$$h_{\bullet j} = \frac{n_{\bullet j}}{n}$$

die relative Randhäufigkeit.

Im Falle endlich vieler Merkmalsausprägungen werden die bivariaten Häufigkeiten am übersichtlichsten durch sogenannte [Kontingenztafeln](#) bzw. [Kontingenztabelle](#)n dargestellt. Dort werden die bivariaten Häufigkeiten  $n_{ij}$  in der  $i$ -ten Zeile und  $j$ -ten Spalte eingetragen.

■ **Beispiel B2.52:** Für 40 Studierende werden das Geburtsjahr und der gewünschte Studienabschluss (B/M/D) ermittelt.

Kontingenztabelle mit absoluten Häufigkeiten:

Studienabschluss: Geburtsjahr	B	M	D	$n_{i\bullet}$
1990-1994	1	9	5	15
1995-1999	15	9	1	25
$n_{\bullet j}$	16	18	6	40

Kontingenztafel mit relativen Häufigkeiten:

Studienabschluss: Geburtsjahr	B	M	D	$h_{i\bullet}$
1990-1994	1/40	9/40	1/8	3/8
1995-1999	3/8	9/40	1/40	5/8
$h_{\bullet j}$	2/5	9/20	3/20	1

- Die relative Häufigkeit für die Ausprägung (1990 – 1994,  $D$ ) ist

$$h_{1,3} = 1/8 = 12.5\%$$

- Die relative Randhäufigkeit für den Bachelor-Studienabschluss ist

$$h_{\bullet 1} = 2/5 = 40\%.$$

## 2.8.2. Unabhängige Merkmale

Die Merkmale  $X$  und  $Y$  heißen unabhängig, wenn

$$h(a_i, b_j) = h(a_i, \bullet) \cdot h(\bullet, b_j)$$

für jede Kombination  $(a_i, b_j)$  mit  $a_i \in M_X$  und  $b_j \in M_Y$  gilt. Wir können das auch kurz als

$$h_{ij} = h_{i\bullet} \cdot h_{\bullet j}, \quad \forall i, j : 1 \leq i \leq k, 1 \leq j \leq l$$

oder

$$n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}, \quad \forall i, j : 1 \leq i \leq k, 1 \leq j \leq l$$

schreiben.

□ „ $\forall$ “ ist der sog. Allquantor und bedeutet „für alle“.

■ **Beispiel B2.53**  $\Rightarrow$  B2.52: Im obigen Beispiel,

Studienabschluss: Geburtsjahr	B	M	D	$h_{i\bullet}$
1990-1994	1/40	9/40	1/8	3/8
1995-1999	3/8	9/40	1/40	5/8
$h_{\bullet j}$	2/5	9/20	3/20	1

sind die Merkmale gewiss nicht unabhängig, denn es gilt z.B.

$$h_{1,2} = 9/40 \neq h_{1\bullet} \cdot h_{\bullet,2} = 3/8 \cdot 9/20 = 27/160.$$



### 2.8.3. Zusammenhangsmaße für nominale Daten

Die über alle Kombinationen von  $i$  und  $j$  summierte quadrierte Abstand

$$\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2$$

kann als Maß für die Unabhängigkeit der beiden untersuchten Merkmale gelten.

Um später entsprechende statistische Tests durchführen zu können, teilt man noch durch  $\frac{n_{i\bullet} n_{\bullet j}}{n}$  und definiert den [Chi-Quadrat-Koeffizienten](#) (auch einfach nur „Chi-Quadrat“) als:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

- Zwei alternative Formeln (häufig einfacher zu verwenden):

$$\chi^2 = n \cdot \left( \left( \sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} \right) - 1 \right)$$

und

$$\chi^2 = n \cdot \left( \left( \sum_{i=1}^k \sum_{j=1}^l \frac{h_{ij}^2}{h_{i\bullet} h_{\bullet j}} \right) - 1 \right).$$

- ⊕ Auch für nominalskalierte Merkmale definiert.
- ⊖ Schwer vergleichbar, da von der Dimension der Kontingenztafel abhängig.

- Korrektur: Der Pearsonsche Kontingenzkoeffizient ist gegeben durch

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

- Weitere Verbesserung: korrigierter Pearsonsche Kontingenzkoeffizient

$$C^* = \sqrt{\frac{\min\{k, l\}}{\min\{k, l\} - 1}} \cdot C.$$

Dann gilt

$$0 \leq C^* \leq 1.$$

■ **Beispiel B2.54**  $\Rightarrow$  B2.52: Gegeben Sei folgende Kontingenztafel:

	A	B	$n_{i\bullet}$
C	4	2	6
D	1	8	9
$n_{\bullet j}$	5	10	15

Wir tragen die Werte für  $\frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}}$  ein:

	A	B
C	8/15	1/15
D	1/45	32/45

$$\chi^2 = 15 \cdot \left( \frac{24 + 3 + 1 + 32}{45} - 1 \right) = 5.$$

Es ist

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{5}{20}} = \frac{1}{2}$$

und

$$C^* = \sqrt{\frac{\min\{k, l\}}{\min\{k, l\} - 1}} \cdot C = \sqrt{2} \cdot \frac{1}{2} = 0.7071$$

- Deutet eher auf einen stärkeren Zusammenhang der beiden Merkmale hin.

### 2.8.4. Zusammenhangsmaße für metrische Daten

Gibt es einen positiven Zusammenhang zwischen  $X$  und  $Y$ , so gilt:

- Ist  $(x_i - \bar{x})$  positiv, so gilt das häufig auch für  $(y_i - \bar{y})$ .
- Ist  $(x_i - \bar{x})$  negativ, so gilt das häufig auch für  $(y_i - \bar{y})$ .
- Also gilt für viele Datenpaare  $(x_i, y_i)$ :  $(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$ .

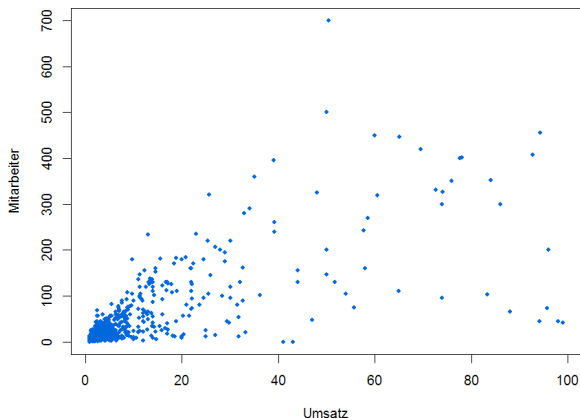
Daher wählt man als Maßzahl die empirische Kovarianz

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

bzw. die Stichprobenkovarianz

$$\hat{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

■ **Beispiel B2.55**  $\Rightarrow$  B2.45: Umsatz und Mitarbeiterzahl von österreichischen IT-Unternehmen mit weniger als 100 Mill. Euro Umsatz.



$$s_{xy} = 730.9737,$$

$$\hat{s}_{xy} = 731.7472.$$

- Alternative Berechnungsformel:

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}.$$

- Es gilt:

$$s_{xy} = s_{yx},$$

$$s_{(ax+b)(cx+d)} = a \cdot c \cdot s_{yx},$$

$$s_{xx} = \hat{\sigma}_*^2(x)$$

und die Cauchy-Schwarzsche Ungleichung:  $|s_{xy}| \leq \hat{\sigma}_*(x)\hat{\sigma}_*(y)$ .

- Man verwendet daher den (empirischen) Korrelationskoeffizienten (Bravais/Pearson)

$$r_{xy} = \frac{s_{xy}}{\hat{\sigma}_*(x)\hat{\sigma}_*(y)}$$

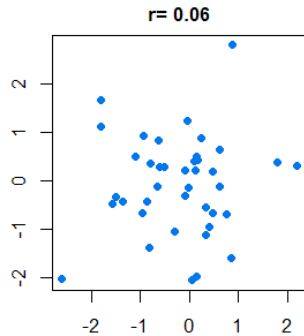
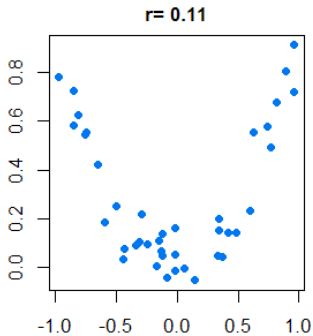
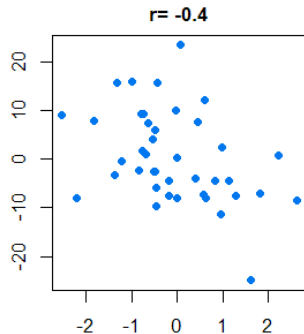
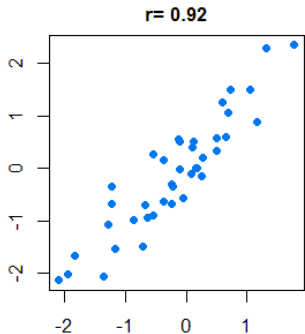
mit Werten im Intervall  $[-1, 1]$ .



- $r_{xy}$  kann als Maß für einen linearen Zusammenhang gelten:

$r_{xy}$	
$= 1$	$x = ay + b, a > 0$ , perfekte pos. Korrelation
$\in [0.5, 1)$	starke positive Korrelation
$\in [0, 0.5)$	schwache positive Korrelation
$\in [-0.5, 0)$	schwache negative Korrelation
$\in [-1, -0.5)$	starke negative Korrelation
$= -1$	$x = ay + b, a < 0$ , perfekte neg. Korrelation

- ⚠ Ein unmittelbarer kausaler Zusammenhang kann nicht erkannt werden.
- Wir werden später noch sehen, wie man einen möglichen linearen Zusammenhang genauer untersuchen kann (Abschnitt „Lineare Regression“)



### 2.8.5. Zusammenhangsmaße für ordinale Daten

■ **Beispiel B2.56:** Zehn Studierende werden nach ihrer Motivation  $Y$  ( $M_X = \{\ominus, \oplus\}$ ) und der Statistiklausurnote  $Y$  ( $M_Y = \{1, 2, \dots, 5\}$ ) gefragt.

Motivation:	$\ominus$	$\oplus$	$\oplus$	$\oplus$	$\ominus$	$\oplus$	$\oplus$	$\ominus$	$\oplus$	$\oplus$
Note:	4	4	2	3	5	1	3	4	1	5

Gibt es einen Zusammenhang?

Kontingenztafel:

	1	2	3	4	5	$\Sigma$
$\oplus$	2	1	2	1	1	7
$\ominus$	0	0	0	2	1	3
	2	1	2	3	2	10

- Der Rang  $R(x_i)$  einer Beobachtung  $x_1$  ist als die Zahl  $m$  definiert, für die  $x_{(m)} = x_i$  gilt.  
Ist der Rang nicht eindeutig (sog. Bindungen), so bildet man den Durchschnittswert der in Frage kommenden Ränge.

■ **Beispiel B2.57**  $\Rightarrow$  B2.56: Im obigen Beispiel ergeben sich die folgenden Ränge für die beiden Merkmale:

Motivation:	⊖	⊕	⊕	⊕	⊖	⊕	⊕	⊖	⊕	⊕
$R(x_i)$ :	2	7	7	7	2	7	7	2	7	7
Note:	4	4	2	3	5	1	3	4	1	5
$R(y_i)$ :	7	7	3	4.5	9.5	1.5	4.5	7	1.5	9.5

- Es gilt für den Mittelwert der Ränge

$$\bar{R} = \frac{n+1}{2}.$$

□ Gaußsche Summenformel:  $1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2}$

- Idee: Man verwendet die ermittelten Ränge um den sog. Rangkorrelationskoeffizienten (Spearman) zu berechnen:

$$R_{xy} = \frac{\sum_{k=1}^n R(x_i)R(y_i) - n\bar{R}^2}{\sqrt{\sum_{k=1}^n R(x_i)^2 - n\bar{R}^2} \times \sqrt{\sum_{k=1}^n R(y_i)^2 - n\bar{R}^2}}.$$

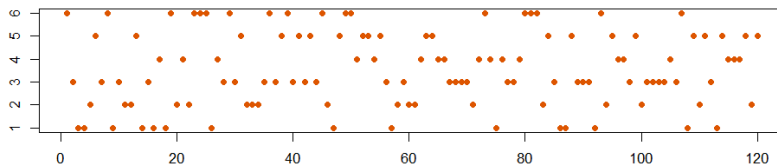
- Es gilt wieder  $R_{xy} \in [-1, 1]$ .

- Perfekter Zusammenhang, wenn  $|R_{xy}| = 1$  gilt, abnehmend mit abnehmendem Absolutbetrag des Koeffizienten.

## 3.

# Wahrscheinlichkeitsrechnung

■ **Beispiel B3.1**  $\Rightarrow_{B1.1}$ : Im Beispiel B1.1 wurde ein Spielwürfel 120 Mal gewürfelt. Es ergaben sich folgende Augenzahlen:



Häufigkeitstabelle:

Augenzahl:	1	2	3	4	5	6
Häufigkeit:	15	18	30	18	21	18

Neben den statistischen Fragestellungen, die unmittelbar die erhobenen Daten betreffen, können wir noch vom konkreten Experiment abstrahieren und uns allgemeinere Fragen stellen:

- Wie wahrscheinlich sind die verschiedenen Augenzahlen bei einem Würfelwurf?
- Wie wahrscheinlich sind die hier vorliegenden Augenzahlenhäufigkeiten bei 120 Würfeln?
- Was ist „Wahrscheinlichkeit“ überhaupt?

Frequentistische Interpretation: Die Wahrscheinlichkeit eines Ereignisses ist der Zahlenwert, gegen die relative Häufigkeit mit wachsendem Stichprobenumfang konvergiert.

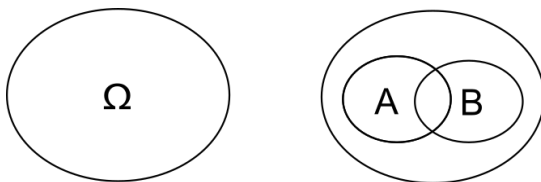


### 3.1. Ereignisse und Wahrscheinlichkeiten

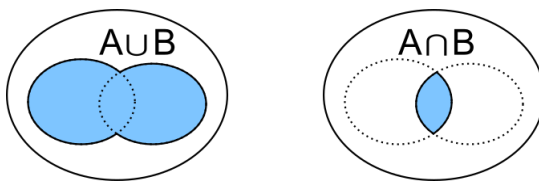
Die axiomatische Wahrscheinlichkeitstheorie lässt die philosophischen Fragen hinter sich und betrachtet Ereignisse und Wahrscheinlichkeiten als mathematische Objekte mit bestimmten Eigenschaften.

Das Grundgerüst kennen wir bereits aus der Statistik:

- Die Grundgesamtheit  $\Omega$  wird nun Wahrscheinlichkeitsraum genannt.
- Die Merkmale heißen nun Zufallsvariablen.
- Die Teilmengen von  $\Omega$  heißen Ereignisse.

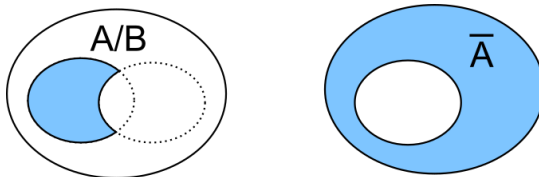


- Die gesamte Menge  $\Omega$  repräsentiert das sichere Ereignis, die leere Menge  $\emptyset$  das unmögliche Ereignis.
- Die Vereinigungsmenge  $A \cup B$  repräsentiert das Eintreten von  $A$  oder von  $B$  (dabei wird zugelassen, dass beide Ereignisse eintreten).
- Die Schnittmenge  $A \cap B$  repräsentiert das gleichzeitige Eintreten von  $A$  und  $B$ .



- Zwei Ereignisse  $A$  und  $B$  heißen unvereinbar, wenn  $A$  und  $B$  disjunkt sind, d.h. es gilt  $A \cap B = \emptyset$ .

- Die Differenzmenge  $A/B$  repräsentiert das Eintreten von  $A$  bei gleichzeitigem Nicht-Eintreten von  $B$ .
- Das Komplement  $\bar{A}$  repräsentiert das Nicht-Eintreten von  $A$ .



- Jedem Ereignis  $A \subseteq \Omega$  kann man eine Zahl  $P(A)$ , seine Wahrscheinlichkeit, zuordnen.

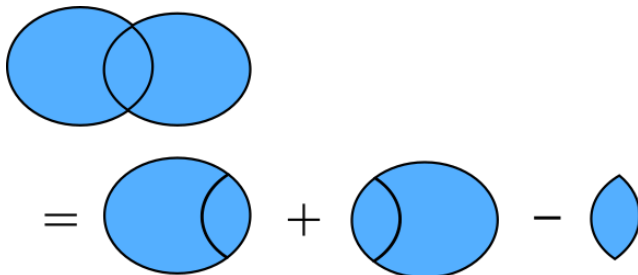
□ In der mathematischen Wahrscheinlichkeitstheorie stellt sich heraus, dass man nicht jedem Ereignis eine Wahrscheinlichkeit zuordnen kann. Das führt zu einigen Komplikationen, die wir hier ignorieren wollen ( $\rightarrow$  Vitali-Mengen, Banach-Tarski-Paradoxon).

- Das Wahrscheinlichkeitsmaß  $P$  muss dabei folgende Bedingungen erfüllen:
  1.  $P(\Omega) = 1$ ,
  2.  $P(A \cup B) = P(A) + P(B)$ , wenn  $A$  und  $B$  unvereinbar sind.

Folgende Regeln gelten dann automatisch:

- $P(A) = 1 - P(\overline{A})$ .
- $P(\emptyset) = 0$ .
- $P(A) \leq P(B)$  wenn  $A \Rightarrow B$ .
- Additionsregel:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



### 3.1.1. Laplace-Experimente

Wir sprechen von einem Laplace-Experiment, wenn  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$  endlich ist und

$$P(\omega_1) = P(\omega_2) = \dots = P(\omega_n) = \frac{1}{n}$$

gilt.

Bei Laplace-Experimenten kann man Wahrscheinlichkeiten „abzählen“:

**■ Satz 3.2 (Laplace-Experiment)**

Im Laplace-Experiment gilt für jedes Ereignis  $A \subseteq \Omega$

$$P(A) = \frac{\#A}{n}.$$

■ **Beispiel B3.2:** Ein Würfel wird geworfen. Es sei

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Dann handelt es sich um ein Laplace-Experiment mit

$$P(\omega) = \frac{1}{6}, \quad \forall \omega \in \Omega.$$

Es sei  $A = \{2, 4, 6\}$  das Ereignis, dass die Augenzahl gerade ist. Dann gilt

$$P(A) = \frac{3}{6} = \frac{1}{2}.$$

⚠ Liegt kein Laplace-Experiment vor, so gilt allgemein nur noch

$$P(A) = \sum_{\omega \in A} P(\omega).$$

■ **Beispiel B3.3:** Ein Würfel werde zweimal geworfen. Wir wählen

$$\Omega = \{(i, j) | i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

Dann handelt es sich um ein Laplace-Experiment mit

$$P(\omega) = \frac{1}{36}, \quad \forall \omega \in \Omega.$$

Es sei  $A = \{(i, j) \in \Omega | i < j\}$  das Ereignis, dass der zweite Wurf eine höhere Augenzahl anzeigt, als der erste Wurf. Dann ist

$$P(A) = \frac{5 + 4 + 3 + 2 + 1}{36} = \frac{15}{36} = \frac{5}{12}.$$



### 3.1.2. Bedingte Wahrscheinlichkeiten

Als bedingte Wahrscheinlichkeit bezeichnet man die Wahrscheinlichkeit eines Ereignisses  $A$ , unter der Voraussetzung, dass der Eintritt eines zweiten Ereignisses  $B$  (mit  $P(B) \neq 0$ ) schon bekannt ist:

$$P(A|B) = P(A, \text{ gegeben } B).$$

#### ■ Satz 3.3

Es gilt

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

Daraus ergibt sich unmittelbar

$$P(A) = P(A|B) P(B).$$

■ **Beispiel B3.4:** Es werde ein Würfel geworfen. Es sei

$A \stackrel{\sim}{=} \text{Die Augenzahl ist gerade} = \{2, 4, 6\},$

$B \stackrel{\sim}{=} \text{Die Augenzahl kleiner als 5} = \{1, 2, 3, 4\}.$

Dann gilt

$$P(A|B) = \frac{P(\{2, 4\})}{P(\{1, 2, 3, 4\})} = \frac{1}{2},$$

$$P(B|A) = \frac{P(\{2, 4\})}{P(\{2, 4, 6\})} = \frac{2}{3}.$$

### 3.1.3. Unabhängigkeit

Zwei Ereignisse  $A$  und  $B$  heißen stochastisch unabhängig, wenn

$$P(A \cap B) = P(A) P(B)$$

gilt.

- Die obige Bedingung ist gleichbedeutend mit

$$P(A|B) = P(A)$$

bzw.

$$P(B|A) = P(B).$$

- ⚠ Nicht mit Unvereinbarkeit verwechseln: Zwei unvereinbare Ereignisse sind fast immer abhängig.

■ **Beispiel B3.5**  $\Rightarrow_{B3.4}$ : Es sei wieder

$A \stackrel{\sim}{=} \text{Die Augenzahl ist gerade} = \{2, 4, 6\},$

$B \stackrel{\sim}{=} \text{Die Augenzahl kleiner als 5} = \{1, 2, 3, 4\}.$

Die beiden Ereignisse sind stochastisch unabhängig:

$$P(A \cap B) = P(\{2, 4\}) = \frac{1}{3} = \frac{3}{6} \cdot \frac{4}{6} = P(A) P(B).$$

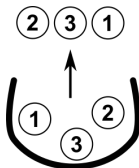
Die Ereignisse  $A$  und  $\bar{A}$  sind nicht unabhängig:

$$P(A \cap \bar{A}) = P(\emptyset) = 0 \neq \frac{1}{4} = P(A)^2.$$

## 3.2. Kombinatorik

### 3.2.1. Permutationen

Aus einem Gefäß mit  $n$  Kugeln werden alle Kugeln gezogen. Wieviele Möglichkeiten der Anordnung (sog. [Permutationen](#)) dieser gezogenen Kugeln gibt es?

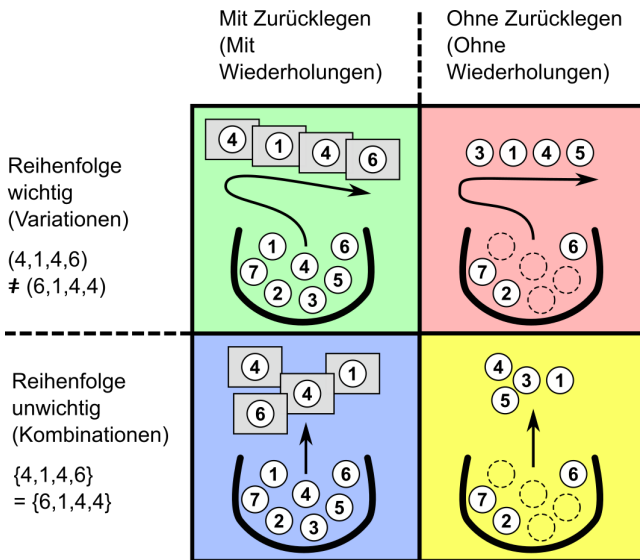


#### ■ Satz 3.4

Es gibt  $n!$  verschiedene Möglichkeiten  $n$  Objekte anzuordnen.

### 3.2.2. Variationen und Kombinationen

Als nächstes ziehen wir nur  $k$  der  $n$  Kugeln.



Unterscheidet man die Reihenfolge der gezogenen Kugeln, so spricht man von Variationen.

- Legt man die Kugeln nicht wieder zurück, so kommt man auf

$$n \cdot (n - 1) \cdots (n - k + 1) = \frac{n!}{(n - k)!}$$

Möglichkeiten.

- Legt man die Kugeln nach dem Ziehen jeweils wieder zurück, so ergeben sich

$$n \cdot n \cdots n = n^k$$

verschiedene Möglichkeiten.

Unterscheidet man die Reihenfolge der gezogenen Kugeln nicht, so spricht man von Kombinationen.

- Möglichkeiten ohne Zurücklegen:

$$\underbrace{\frac{n!}{(n-k)!}}_{\text{Variationen}} \times \underbrace{\frac{1}{k!}}_{\text{Anordnungen}} = \binom{n}{k}.$$

- Möglichkeiten mit Zurücklegen (ohne Beweis):

$$\binom{n+k-1}{k}.$$



**Zurücklegen  
Reihenfolge**

$$\overline{V}_n^k = n^k$$

**Ohne Zurücklegen  
Reihenfolge**

$$V_n^k = \frac{n!}{(n-k)!}$$

**Zurücklegen  
Ohne Reihenfolge**

$$\overline{C}_n^k = \binom{n+k-1}{k}$$

**Ohne Zurücklegen  
Ohne Reihenfolge**

$$C_n^k = \binom{n}{k}$$

## 3.3. Zufallsvariablen und ihre Verteilungen

### 3.3.1. Zufallsvariablen

Zufallsvariablen sind die wahrscheinlichkeitstheoretischen Pendanten metrischer Merkmale, also Abbildungen  $\Omega \rightarrow \mathbb{R}$ .

Wir unterscheiden wie bei den Merkmalen diskrete und stetige Zufallsvariablen.

- Eine Zufallsvariable ist diskret, wenn sie nur abzählbar viele Werte annehmen kann.
- Eine Zufallsvariable heißt stetig, wenn ihr Wertebereich ein Intervall oder die ganze Zahlengerade ist und eine weitere Bedingung erfüllt ist, die wir später betrachten.

Wir schreiben im Folgenden kurz  $P(X \leq x)$  an Stelle der korrekteren aber umständlicheren Schreibweise  $P(\{\omega \in \Omega | X(\omega) \leq x\})$ .

### 3.3.2. Verteilungsfunktionen

Die Verteilungsfunktion einer Zufallsvariablen  $X$  ist gegeben durch die Funktion

$$F_X(x) = P(X \leq x).$$

Wir schreiben kurz  $F$  statt  $F_X$ , wenn klar ist, welche Zufallsvariable gemeint ist.

- $F$  ist stets nicht-fallend,
- $F$  ist rechtsseitig stetig,
- $\lim_{x \rightarrow -\infty} F(x) = 0$ ,  $\lim_{x \rightarrow \infty} F(x) = 1$ .

Die stochastischen Eigenschaften einer Zufallsvariablen werden durch Angabe der Verteilungsfunktion vollständig beschrieben.

Mit Hilfe der Verteilungsfunktion kann man Wahrscheinlichkeiten berechnen:

$$P(X > x) = 1 - F(x)$$

$$P(y < X \leq x) = F(x) - F(y)$$

$$P(X = x) = F(x) - F(x-)$$

$$P(X < x) = F(x-)$$

$$P(X \geq x) = 1 - F(x-)$$

$$P(y \leq X \leq x) = F(x) - F(y-)$$

$$\vdots \qquad \qquad \vdots \qquad \qquad \vdots$$

□  $F(x-)$  bezeichnet den linksseitigen Grenzwert

$$F(x-) = \lim_{u \uparrow x} F(u).$$

Es gibt noch weitere Möglichkeiten die stochastischen Eigenschaften einer Zufallsvariablen zu beschreiben:

- Für eine diskrete Zufallsvariable  $X$  mit Werten  $M_X = \{x_1, x_2, \dots\}$  definiert man die Wahrscheinlichkeitsfunktion:

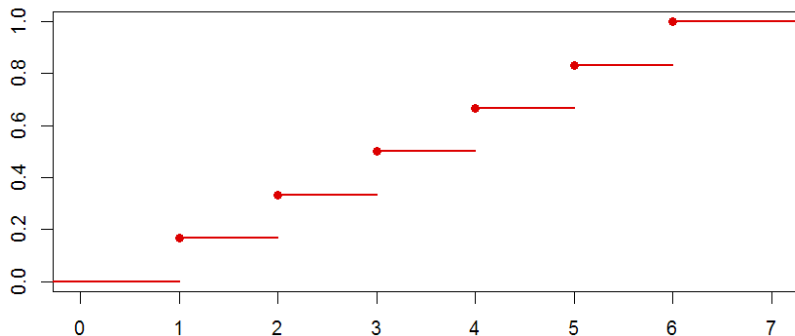
$$p(x) = P(X = x) = \begin{cases} 0 & ; x \notin M_X \\ P(X = x_i) & ; x = x_i \end{cases}$$

- Für stetige Zufallsvariablen fordern wir, dass  $F$  stetig und stückweise differenzierbar ist. Man definiert dann die Wahrscheinlichkeitsdichte als die Ableitung

$$f(x) = F'(x)$$

an den Stellen, wo  $F$  differenzierbar ist (an allen anderen Stellen kann man  $f(x)$  beliebig definieren).

■ **Beispiel B3.6**  $\Rightarrow_{B3.4}$ : Es sei wieder  $X$  die Augenzahl beim einmaligen Wurf mit einem fairen Würfel. Verteilungsfunktion:



Wahrscheinlichkeitsfunktion:

$$p(x) = \begin{cases} 0 & ; x \notin \{1, 2, 3, 4, 5, 6\} \\ 1/6 & ; x \in \{1, 2, 3, 4, 5, 6\} \end{cases}$$

- Diskreten und stetigen Zufallsvariablen ist also die Verteilungsfunktion

$$F(x) = P(X \leq x)$$

gemeinsam.

- Sie unterscheiden sich bei der Wahrscheinlichkeits- bzw. Dichtefunktion:

	W.-Funktion für diskrete ZV.	W.-Dichte für stetige ZV.
Symbol	$p(x) = P(X = x)$	$f(x)$
Nicht-Negativität	$p(x) \geq 0$ $p(x) = 0, \forall x \notin M_X$	$f(x) \geq 0$
Normierung	$\sum_{i=1}^{\infty} p(x_i) = 1$	$\int_{-\infty}^{\infty} f(x) dx = 1$
Wahrscheinlichkeiten	$P(A) = \sum_{x \in A \cap M_X} p(x)$	$P(A) = \int_{x \in A} f(x) dx$

### 3.4. Erwartungswert und Varianz

Der Erwartungswert ist das wahrscheinlichkeitstheoretische Gegenstück zum arithmetischen Mittel.

- Für diskrete Zufallsvariablen:

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot p(x_i).$$

- Für stetige Zufallsvariablen:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx.$$



Allgemeiner kann man den Erwartungswert von Funktionen  $g : \mathbb{R} \rightarrow \mathbb{R}$  einer Zufallsvariablen erklären:

- Für diskrete Zufallsvariablen:

$$E(g(X)) = \sum_{i=1}^{\infty} g(x_i) \cdot p(x_i).$$

- Für stetige Zufallsvariablen:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx.$$

Natürlich ist der Erwartungswert nur definiert, wenn die entsprechende Summe oder das entsprechende Integral definiert sind. Auf den Fall, wo diese Größen definiert aber unendlich sind, gehen wir hier nicht näher ein.

Die Varianz und die Standardabweichung einer Zufallsvariable sind definiert als

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2.$$

und

$$\hat{\sigma}_*(X) = \sqrt{\text{Var}(X)}.$$

Beide Größen beschreiben die Streuung der Zufallsvariablen  $X$ .

Es gelten die schon vom arithmetischen Mittel vertrauten Rechenregeln:

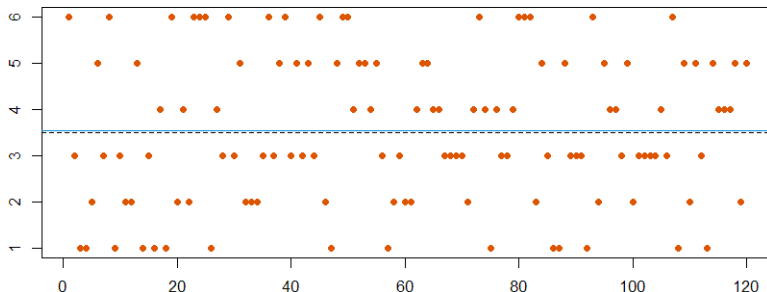
- $E(aX + b) = aE(X) + b,$
- $\text{Var}(aX + b) = a^2\text{Var}(X),$
- $\hat{\sigma}_*(aX + b) = a\hat{\sigma}_*(X),$
- $E(X + Y) = E(X) + E(Y).$

## 3.5. Das Gesetz der großen Zahlen

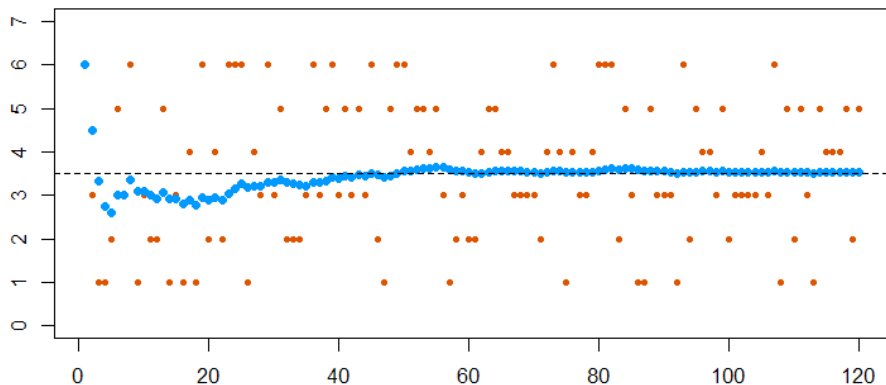
■ **Beispiel B3.7**  $\Rightarrow$  **B1.1**: Im Beispiel **B1.1** ergab sich ein arithmetisches Mittel von  $\bar{x} = 3.55$ . Das liegt verdächtig nahe beim theoretischen Erwartungswert

$$E(X) = 3.5$$

der Augenzahlen-Zufallsvariable  $X$ .



Wir betrachten den Mittelwert  $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$  der ersten  $n$  Würfe:



Man kann zeigen: Das ist kein Spezialfall, sondern einer der wesentlichen Grenzwertsätze der Wahrscheinlichkeitstheorie.

**■ Satz 3.6 (Das starke Gesetz der großen Zahlen)**

Es seien  $X_1, X_2, \dots$  unabhängige und identisch verteilte Zufallsvariablen mit dem gemeinsamen Erwartungswert  $\mu$  und

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

Dann ist die Wahrscheinlichkeit dafür, dass

$$\lim_{n \rightarrow \infty} \bar{X}_n = \mu$$

gilt, eins.

- $\bar{X}_n$  ist also bei großen Stichprobenumfängen ein guter Schätzer für den u.U. unbekannten Erwartungswert (ein sog. stark konsistenter Schätzer).

## 3.6. Unabhängigkeit und Korrelation

Zwei Zufallsvariablen  $X$  und  $Y$  heißen stochastisch unabhängig, wenn die gemeinsame Verteilungsfunktion

$$F_{X,Y}(x, y) = P(X \leq x \text{ und } Y \leq y) = P(X \leq x, Y \leq y)$$

die Produktgleichung

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

erfüllt.

Für unabhängige Zufallsvariablen  $X$  und  $Y$  gilt

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

- Als Maß für den Zusammenhang zweier Zufallsvariablen kann die Kovarianz

$$\begin{aligned}\text{Cov}(X, Y) &= E((X - E(X)) \cdot (Y - E(Y))) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

verwendet werden.

- Der Korrelationskoeffizient

$$\varrho(X, Y) = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_*(X)\hat{\sigma}_*(Y)}$$

nimmt Werte im Intervall  $[-1, 1]$  an und gibt Auskunft über den linearen Zusammenhang der beiden Zufallsvariablen.

- Gilt  $E(XY) = E(X)E(Y)$ , so nennt man  $X$  und  $Y$  unkorreliert. Unabhängige Zufallsvariablen sind immer unkorreliert.

## 3.7. Fünf wichtige Verteilungen

### 3.7.1. Die Bernoulli-Verteilung

Eine Bernoulli-verteilte Zufallsvariable  $X$  nimmt nur die beiden Werte  $x_1 = 0$  („Misserfolg“) und  $x_2 = 1$  („Erfolg“) an. Sie ist dann das Ergebnis eines sog. Bernoulli-Experiments.

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

Offenbar gilt

$$E(X) = (1 - p) \cdot 0 + p \cdot 1 = p$$

und

$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= (1 - p) \cdot 0^2 + p \cdot 1^2 - p^2 = p(1 - p). \end{aligned}$$



### 3.7.2. Die Binomialverteilung

Werden  $n$  Bernoulli-Experimente unabhängig voneinander mit Ergebnissen  $X_1, X_2, \dots, X_n$  durchgeführt, so hat die Zufallsvariable

$$K \stackrel{\sim}{=} \text{Anzahl der Erfolge}$$

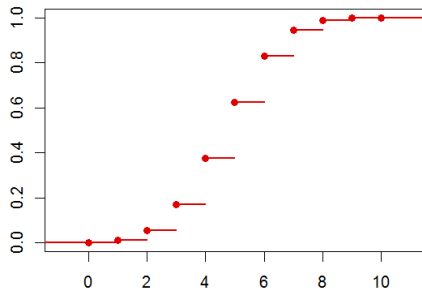
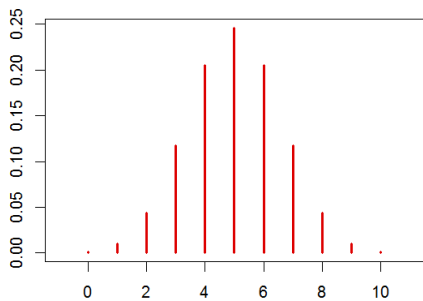
eine Binomialverteilung und es gilt

$$P(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

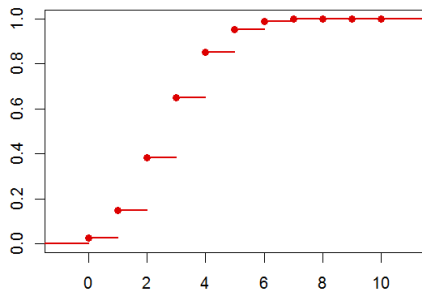
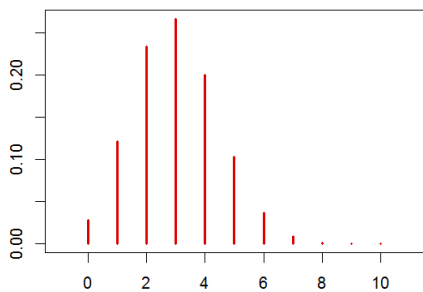
Dann ergibt sich

$$E(K) = nE(X_1) = np,$$

$$\text{Var}(K) = n\text{Var}(X_1) = np(1 - p).$$



$n = 10, p = 0.5$



$n = 10, p = 0.3$

### 3.7.3. Die geometrische Verteilung

Es werden Bernoulli-Experimente solange ausgeführt, bis zum ersten Mal Erfolg eintritt. Es sei  $Z$  der Index, für den zum ersten Mal  $X_Z = 1$  gilt. Dann hat  $Z$  eine geometrische Verteilung (Typ I):

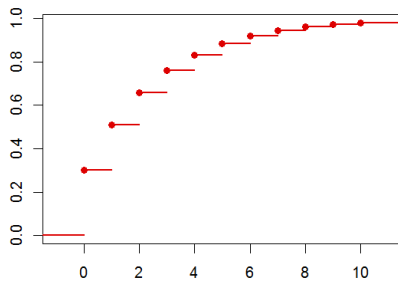
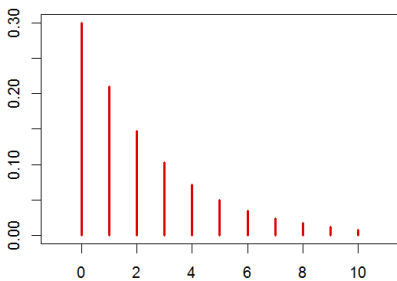
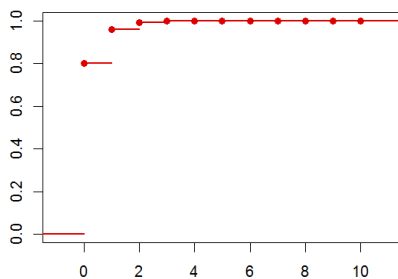
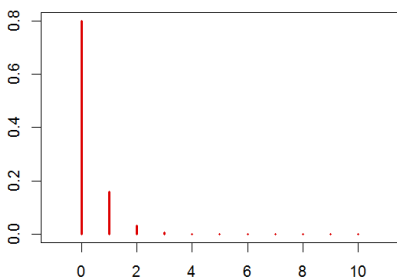
$$P(Z = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

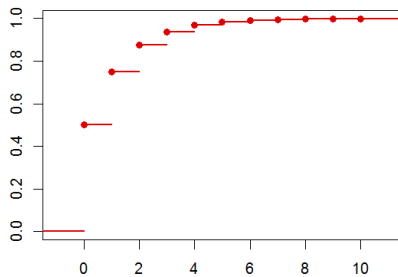
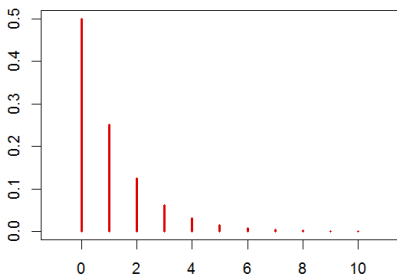
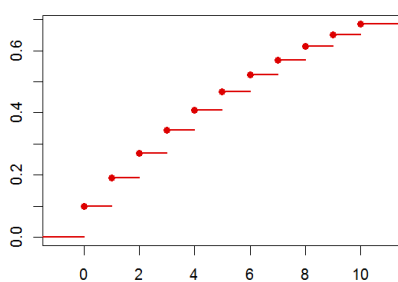
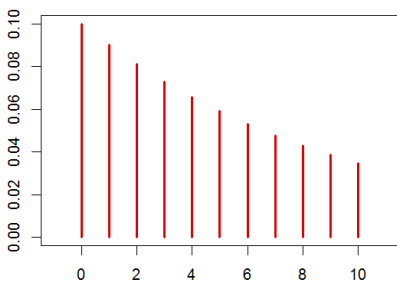
Die Anzahl der Misserfolge  $M = Z - 1$  hat eine geometrische Verteilung vom Typ II:

$$P(M = k) = (1 - p)^k p, \quad k = 0, 1, 2, 3, \dots$$

Es gilt

	$E(\cdot)$	$\text{Var}(\cdot)$
Typ I	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Typ II	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

 $p = 0.3$  $p = 0.8$

 $p = 0.5$  $p = 0.1$

## Übersicht:

Verteilung	Anzahl Experimente	Gefragt
Bernoulli	1	Ausgang (0=Misserfolg, 1=Erfolg)
Binomial	$n$	Anzahl der Erfolge
Geometrisch I	unbegrenzt	Index mit erstem Erfolg
Geometrisch II	unbegrenzt	Index mit letztem Misserfolg

### 3.7.4. Die Multinomialverteilung

Gegeben seien eine Folge diskreter Zufallsvariablen  $X_1, X_2, \dots, X_n$  mit Werten in der Menge  $\{x_1, x_2, \dots, x_m\}$  und jeweils gleicher Wahrscheinlichkeitsfunktion  $p$ . Es sei  $K_i$  die absolute Häufigkeit der  $X$ -Zufallsvariablen mit Wert  $x_i$ . Dann gilt für die gemeinsame Wahrscheinlichkeitsfunktion

$$\begin{aligned} P(K_1 = k_1, K_2 = k_2, \dots, K_m = k_m) \\ = \binom{n}{k_1 \ k_2 \ \dots \ k_m} p(x_1)^{k_1} p(x_2)^{k_2} \dots p(x_m)^{k_m}, \end{aligned}$$

wobei  $\sum_{i=1}^m k_i = n$  gelten muss.

□ (Multinomialkoeffizient)

$$\binom{n}{k_1 \ k_2 \ \dots \ k_n} = \frac{n!}{k_1! k_2! \dots k_n!}.$$

■ **Beispiel B3.8**  $\Rightarrow_{B1.1}$ : Es sei  $A_i$  die Augenzahl im  $i$ -ten Wurf mit einem fairen Würfels und

$$X_i = \begin{cases} 1 & ; A_i = 6, \\ 0 & ; A_i \neq 6. \end{cases}$$

Dann besitzen die  $X_i$  jeweils eine Bernoulli-Verteilung mit  $p = \frac{1}{6}$ , d.h.

$$E(X_i) = \frac{1}{6}, \quad \text{Var}(X_i) = p(1 - p) = \frac{5}{36}.$$

Es gilt z.B.

$$P(X_1 = 1, X_2 = 2, \dots, X_6 = 6) = \left(\frac{1}{6}\right)^6 = \frac{1}{46656}.$$



Es sei  $K$  die Anzahl der 6er bei 120 Würfeln. Dann ist  $K$  binomialverteilt, d.h.

$$P(K = k) = \binom{120}{k} (1/6)^k (5/6)^{120-k}.$$

Zum Beispiel ist

$$P(K = 18) = \binom{120}{18} (1/6)^{18} (5/6)^{102} \approx 0.09$$

und

$$P(K \leq 18) = \sum_{j=0}^{18} \binom{120}{j} (1/6)^j (5/6)^{120-j} = 0.3657$$

$$P(K \geq 30) = \sum_{j=30}^{120} \binom{120}{j} (1/6)^j (5/6)^{120-j} = 0.0129$$

Es sei  $B$  das Ereignis, dass folgende Häufigkeiten beobachtet werden:

Augenzahl:	1	2	3	4	5	6
Häufigkeit:	15	18	30	18	21	18

Dann ist

$$P(B) = \binom{120}{15 \ 18 \ 30 \ 18 \ 21 \ 18} \left(\frac{1}{6}\right)^{120} \approx 6 \cdot 10^{-7}.$$

Wollen wir die Wahrscheinlichkeit einer Abweichung von der „zu erwartenden“ Tabelle

Augenzahl:	1	2	3	4	5	6
Häufigkeit:	20	20	20	20	20	20

berechnen, müssen wir tiefer in die Trickkiste greifen. Mehr dazu später.

Wie lange dauert es im Mittel, bis eine 6 gewürfelt wird?

Die Zufallsvariable

$Z \approx$  # Versuche, bis eine 6 gewürfelt wird.

Dann hat  $Z$  eine geometrische Verteilung, d.h.

$$P(Z = k) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}, \quad k = 1, 2, 3, \dots$$

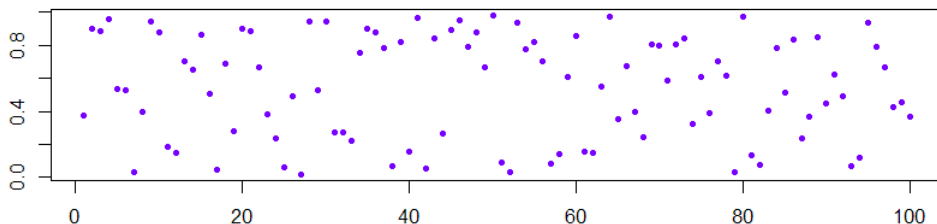
Als Erwartungswert erhalten wir

$$E(Z) = \frac{1}{p} = 6.$$

### 3.7.5. Die stetige Gleichverteilung

Ist  $X$  gleichverteilt auf dem Intervall  $[a, b]$ , so liegt  $X$  quasi „maximal zufällig“ verteilt in dem Intervall.

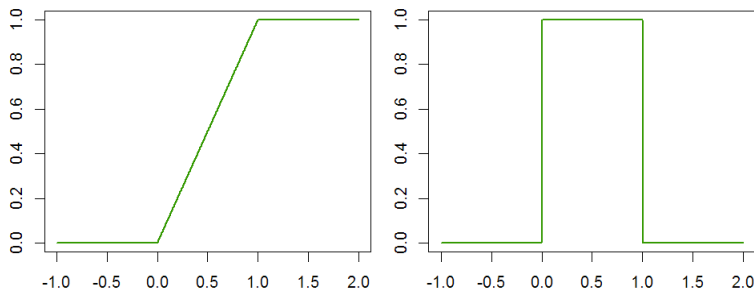
- Handelsübliche Taschenrechner verfügen über eine **RND**-Taste, die gleichverteilte Zufallszahlen erzeugt.
- Mit Hilfe gleichverteilter Zufallsvariablen kann man anders verteilte Zufallszahlen erzeugen (Inversionsmethode, Monte-Carlo-Simulation)



Verteilungs- und Dichtefunktion der stetigen Gleichverteilung sind gegeben durch

$$F(x) = \begin{cases} 0 & ; x < a \\ \frac{x - a}{b - a} & ; x \in [a, b) \\ 1 & ; x \geq b \end{cases}$$

$$f(x) = \begin{cases} 1 & ; x \in [a, b) \\ 0 & ; x \notin [a, b) \end{cases}$$



$$a = 0, b = 1$$

Es gilt für eine auf  $[a, b]$  gleichverteilte Zufallsvariable

$$E(X) = \frac{a + b}{2},$$

$$\text{Var}(X) = \frac{(b - a)^2}{12}.$$

## 3.8. Die Normalverteilung und ihre Verwandten

### 3.8.1. Die Standardnormalverteilung

Die wichtigste Verteilung der Statistik ist die [Standardnormalverteilung](#).

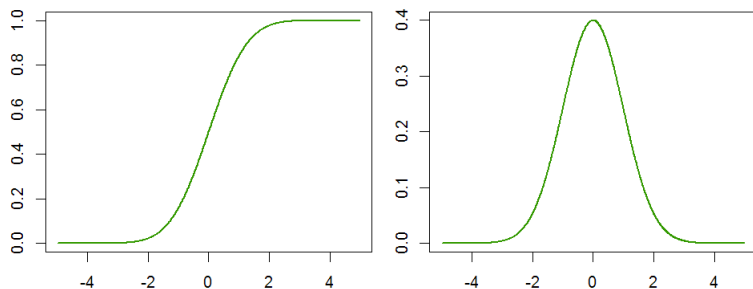
- Die Standardnormalverteilung besitzt die Dichtefunktion

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

- Die zugehörige Verteilungsfunktion lässt sich nicht in geschlossener Form angeben:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du.$$

Verteilungsfunktion  $\Phi(x)$  und Dichtefunktion  $\varphi(x)$ :



$$\mu = 0, \sigma = 1$$

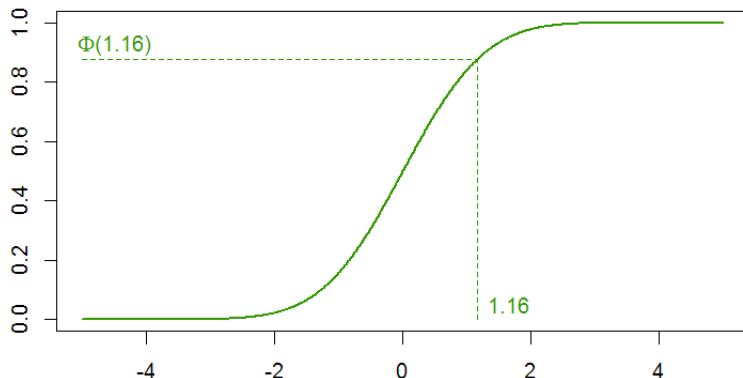
- Wir schreiben  $N(0, 1)$  für die Standardnormalverteilung und  $X \sim N(0, 1)$  für eine standardnormalverteilte Zufallsvariable.
- Für  $X \sim N(0, 1)$  gilt  $E(X) = 0$  und  $\text{Var}(X) = 1$ .



### 3.8.2. Tabellen und Quantile

- Die Werte  $\Phi(x)$  sind tabellarisch gegeben oder können mit Taschenrechnern und Computern abgerufen werden (s. Tabelle Seite ??).

Beispiel:  $\Phi(1.16) = 0.877$

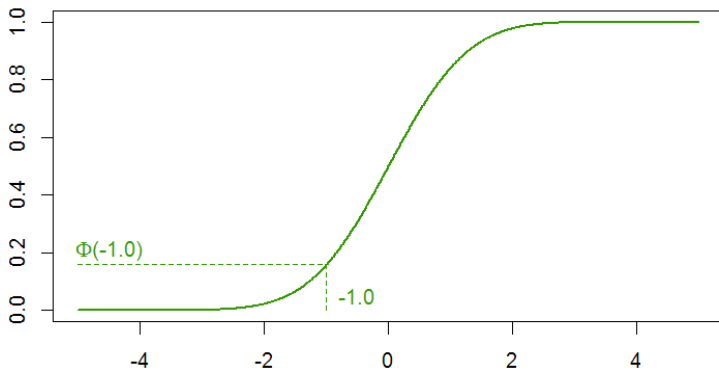


- Für negative Argumente kann man die Umformungsregel

$$\Phi(-x) = 1 - \Phi(x)$$

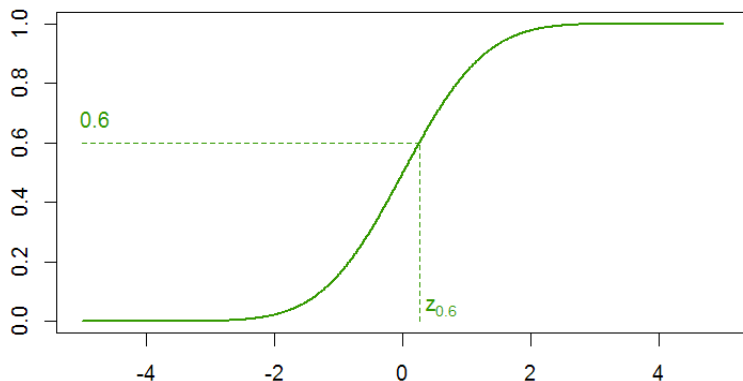
verwenden.

Beispiel:  $\Phi(-1.0) = 1 - 0.8413 = 0.1587$ ,



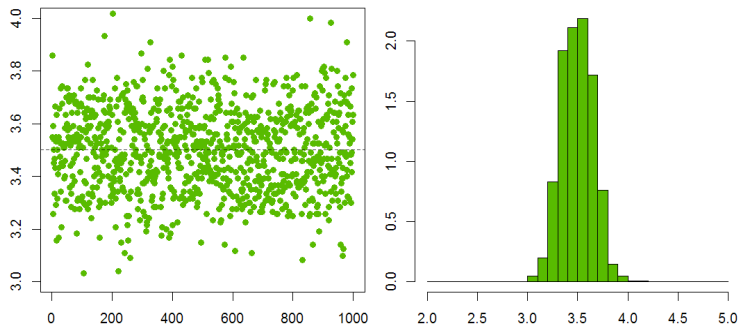
- Als  $\alpha$ -Quantil bezeichnet den Wert  $z$  für den  $\Phi(z) = \alpha$  gilt. Man verwendet die Bezeichnung  $z_\alpha$  für diesen Wert.
- Die Quantile kann man ebenfalls aus der Tabelle auf Seite ?? entnehmen.

Beispiel:  $z_{0.6} = 0.25$ .



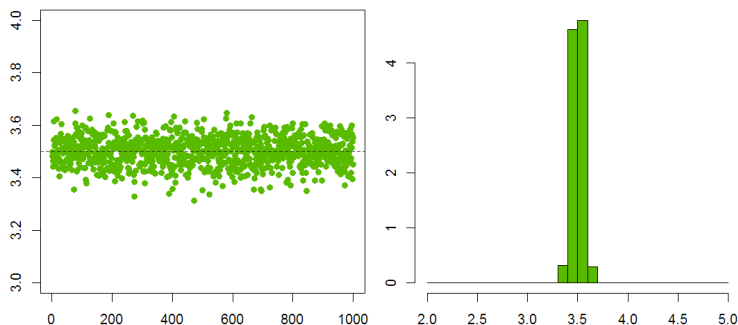
### 3.8.3. Der zentrale Grenzwertsatz

■ **Beispiel B3.9**  $\Rightarrow_{B1.1}$ : Wir wiederholen das Würfelexperiment aus dem Beispiel B1.1 eintausend Mal und betrachten für jeden Durchgang das arithmetische Mittel:



Standardabweichung dieser Mittelwerte: 0.159.

Wir würfeln nun  $n = 1000$  Mal und wiederholen das Experiment 1000 Mal:



Standardabweichung der Mittelwerte: 0.054.

- Wir beobachten: Die Standardabweichung wird mit wachsendem  $n$  immer kleiner.

Es seien  $X_1, X_2, X_3, \dots$  unabhängige und identisch verteilte Zufallsvariablen mit Erwartungswert  $\mu$  und Standardabweichung  $\sigma$  und

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

ihr arithmetisches Mittel.

Dann gilt

$$E(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu,$$

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n},$$

$$\hat{\sigma}_*(\bar{X}_n) = \sqrt{\text{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}.$$

**■ Satz 3.8**

Das arithmetische Mittel  $\bar{X}_n$  der Zufallsvariablen  $X_1, X_2, \dots$  besitzt den Erwartungswert  $\mu$  und die Standardabweichung  $\sigma/\sqrt{n}$ .

Es folgt, dass die standardisierte Zufallsvariable

$$X_n^* = \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}$$

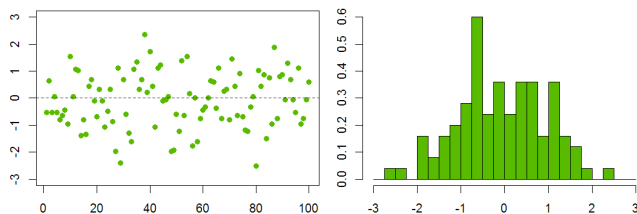
den Erwartungswert 0 und die Standardabweichung 1 besitzt.

Wir können auch mit  $n$  erweitern und schreiben:

$$X_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}.$$

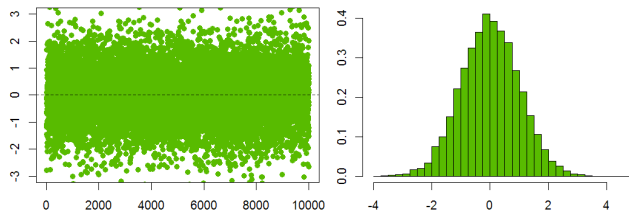
Welche Verteilung besitzt  $X_n^*$ ?

120 Würfe, 100 Mal wiederholt:



$$\mu = 0, \sigma = 1$$

10 000 Würfe, 10 000 Mal wiederholt:



$$\mu = 0, \sigma = 1$$

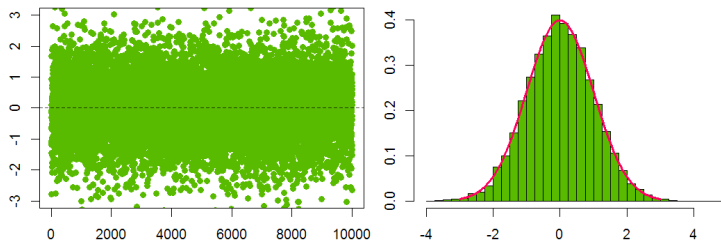


**■ Satz 3.9 (Zentraler Grenzwertsatz)**

Gegeben seien unabhängige und identisch verteilte Zufallsvariablen  $X_1, X_2, \dots$  mit Erwartungswert  $\mu$  und Varianz  $\sigma^2$ . Dann konvertiert die Verteilung der standardisierten Zufallsvariablen

$$X_n^* = \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}$$

für  $n \rightarrow \infty$  gegen die Standardnormalverteilung  $\Phi(x)$ .



$$\mu = 0, \sigma = 1$$

### 3.8.4. Abschätzungen

Mit Hilfe des zentralen Grenzwertsatzes können wir Wahrscheinlichkeiten für den Mittelwert und Summen von unabhängigen und identisch verteilten Zufallsvariablen abschätzen.

■ **Satz 3.10 (Zentraler Grenzwertsatz, Teil II)**

Für große Werte von  $n$  gilt

$$P\left(\sum_{i=1}^n X_i \leq x\right) \approx \Phi\left(\frac{x - n\mu}{\sqrt{n}\sigma}\right).$$

und

$$P(\bar{X}_n \leq x) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right).$$

■ **Beispiel B3.10**  $\Rightarrow_{B1.1}$ : War der gewürfelte Mittelwert im Beispiel B1.1 signifikant abweichend vom Erwartungswert?

Wie groß ist die Wahrscheinlichkeit, bei 120 Würfeln mit einem Spielwürfel, einen Mittelwert  $\bar{X}_n > 3.55$  zu erhalten?

$$\begin{aligned}
 P(\bar{X}_{120} > 3.55) &= 1 - P(\bar{X}_{120} \leq 3.55) \\
 &\approx 1 - \Phi\left(\frac{3.55 - 3.5}{\sqrt{\frac{35}{12}}/\sqrt{120}}\right) \\
 &= 1 - \Phi(0.3207135) \\
 &\stackrel{S.??}{=} 1 - 0.6255 \\
 &= 0.3745
 \end{aligned}$$

Die Wahrscheinlichkeit für einen Mittelwert über 3.55 beträgt bei 120 Würfeln etwa 37.5%.

■ **Beispiel B3.11:** Bei einem Spiel verliert der Spieler mit Wahrscheinlichkeit 0.7 fünf Euro und gewinnt mit Wahrscheinlichkeit 0.3 acht Euro. Es sei  $X_i$  der Gewinn bzw. Verlust im  $i$ -ten Spiel (sog. [Irrfahrt/Random Walk](#)). Wie groß ist die Wahrscheinlichkeit, dass der Spieler nach 30 Spielen einen (positiven) Gewinn verzeichnet?

Es gilt  $\mu = E(X) = -0.7 \cdot 5 + 0.3 \cdot 8 = -1.1$  und  $\text{Var}(X) = 0.7 \cdot 25 + 0.3 \cdot 64 - 1.1^2 = 35.49$ .

Damit erhalten wir

$$\begin{aligned} P\left(\sum_{k=1}^{30} X_k > 0\right) &= 1 - P\left(\sum_{k=1}^{30} X_k \leq 0\right) \\ &\approx 1 - \Phi\left(\frac{0 - 30 \cdot (-1.1)}{\sqrt{35.49 \cdot 30}}\right) \\ &= 1 - \Phi(1.011) \\ &= 1 - 0.8438 = 0.1562. \end{aligned}$$


### 3.8.5. Die allgemeine Normalverteilung

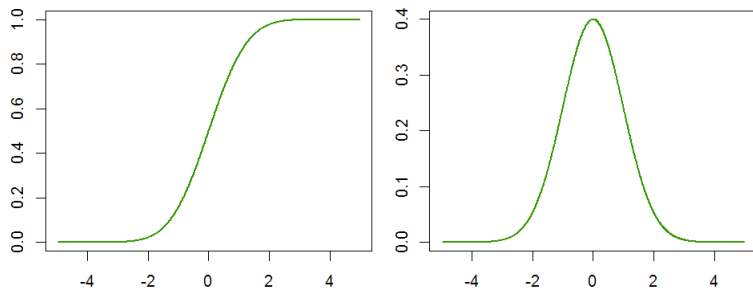
Wenn  $X \sim N(0, 1)$  gilt, dann besitzt  $\sigma X + \mu$  eine sog. [Normalverteilung](#).

- Die Normalverteilung besitzt die Dichtefunktion

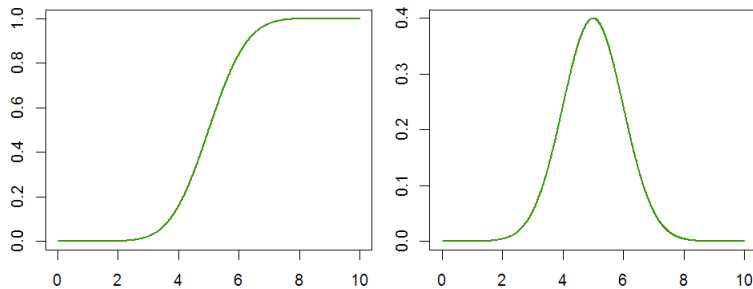
$$\varphi_{\mu, \sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}.$$

- Die zugehörige Verteilungsfunktion  $\Phi_{\mu, \sigma}$  lässt sich wieder nicht in geschlossener Form angeben.
- Wir schreiben  $N(\mu, \sigma)$  für die Normalverteilung.

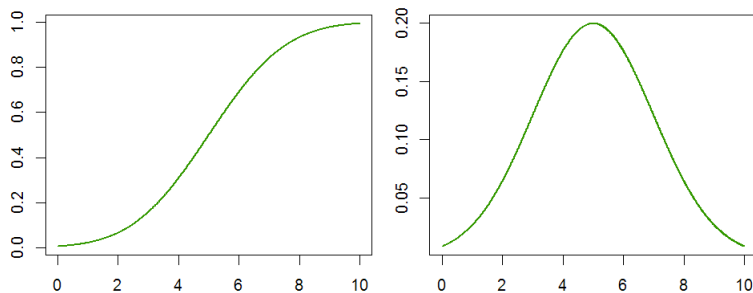
 In vielen Büchern bezeichnet  $N(\mu, s)$  eine Normalverteilung mit Erwartungswert  $\mu$  und Varianz  $s$ .



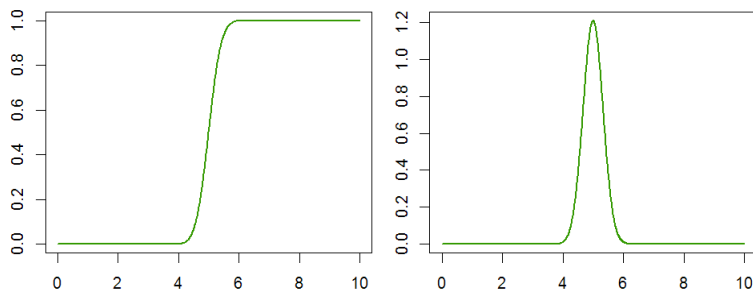
$$\mu = 0, \sigma = 1$$



$$\mu = 5, \sigma = 1$$



$$\mu = 5, \sigma = 2$$



$$\mu = 5, \sigma = 1/3$$

### 3.8.6. Rechenregeln und Transformationen für die Normalverteilung

- Angenommen  $X \sim N(\mu, \sigma)$ . Dann gilt

$$aX + b \sim N(a\mu + b, |a|\sigma).$$

Speziell erhalten wir, wenn wir  $a = \frac{1}{\sigma}$  und  $b = -\mu/\sigma$  wählen,

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

- Umgekehrt folgt aus  $X \sim N(0, 1)$

$$\sigma X + \mu \sim N(\mu, \sigma).$$



- Die Summe von zwei normalverteilten Zufallsvariablen ist wieder normalverteilt. Falls  $Y \sim N(\nu, \tau)$  und  $X \sim N(\mu, \sigma)$  unabhängig sind, gilt

$$X + Y \sim N(\mu + \nu, \sqrt{\sigma^2 + \tau^2}).$$

- Wenn  $X_1, X_2, \dots, X_n$  unabhängig sind und  $X_i \sim N(\mu, \sigma)$  gilt, so ergibt sich

$$\sum_{i=1}^n X_i \sim N(n\mu, \sqrt{n}\sigma)$$

und

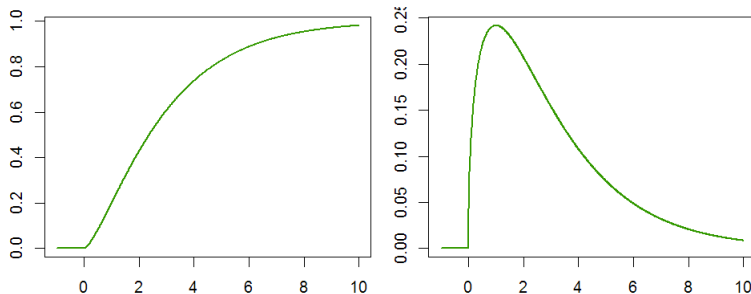
$$\bar{X}_n \sim N(\mu, \sigma/\sqrt{n}).$$

### 3.8.7. Die Chi-Quadrat-Verteilung

Wenn  $X_1, X_2, \dots, X_n$  standardnormalverteilte unabhängige Zufallsvariablen sind, so besitzt die Summe der Quadrate

$$\chi^2 = \sum_{i=1}^n X_i^2$$

eine sog. [Chi-Quadrat-Verteilung mit  \$n\$  Freiheitsgraden](#).



$n = 3$

- Das  $\alpha$ -Quantil  $\chi_{n,\alpha}$  der Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden ist der Werte  $z$  für den  $F(z) = \alpha$  gilt, wenn  $F$  die Chi-Quadrat-Verteilungsfunktion bezeichnet.
- Die Quantile sind aus der Tabelle auf Seite ?? zu entnehmen. Zum Beispiel ist

$$\chi_{6,0.99} = 16.81.$$

Das bedeutet, dass

$$P\left(\sum_{i=1}^6 X_i^2 \leq 16.81\right) = 0.99$$

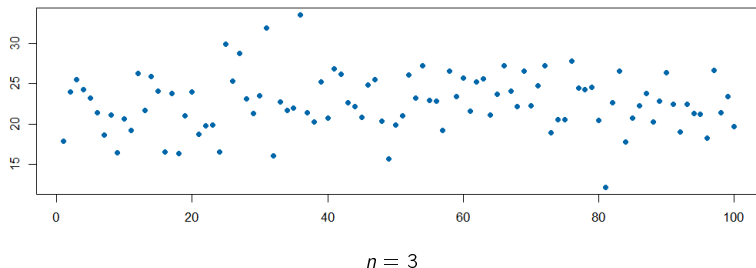
ist, wenn die  $X_i$  unabhängige standardnormalverteilte Zufallsvariablen sind.

### 3.8.8. Die t-Verteilung

Wenn  $X$  und  $X_1, X_2, \dots, X_n$  standardnormalverteilte unabhängige Zufallsvariablen sind, dann besitzt die Zufallsvariable

$$T = \frac{X}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}}$$

eine (Student)-t-Verteilung mit  $n$  Freiheitsgraden .



- Das  $\alpha$ -Quantil  $t_{n,\alpha}$  der t-Verteilung mit  $n$  Freiheitsgraden ist der Wert  $z$  für den  $F(z) = \alpha$  gilt, wenn  $F$  die t-Verteilungsfunktion bezeichnet.
- Die Quantile sind aus der Tabelle auf Seite ?? zu entnehmen.  
Beispielsweise ergibt sich

$$t_{20,0.9} = 1.325,$$

d.h.

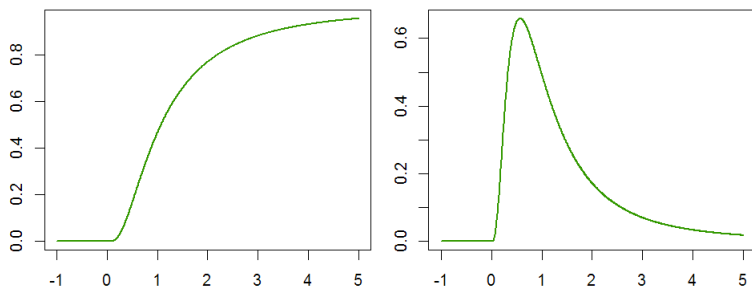
$$P(T \leq 1.325) = 0.9.$$

### 3.8.9. Die F-Verteilung

Es seien  $X_1$  und  $X_2$  zwei Chi-Quadrat-verteilte unabhängige Zufallsvariablen mit  $n$  bzw.  $m$  Freiheitsgraden. Dann hat die Zufallsvariable

$$F = \frac{X_1}{X_2}$$

eine F-Verteilung mit  $n$  und  $m$  Freiheitsgraden .



$n = 10, m = 5$

- Das  $\alpha$ -Quantil  $F_{(n,m),\alpha}$  der F-Verteilung mit  $n$  und  $m$  Freiheitsgraden ist der Werte  $z$  für den  $F(z) = \alpha$  gilt, wenn  $F$  die entsprechende Verteilungsfunktion bezeichnet.
- Die Quantile findet man in den Tabellen ab Seite ?? . Es ist z.B.

$$F_{(10,5),0.95} = 4.735,$$

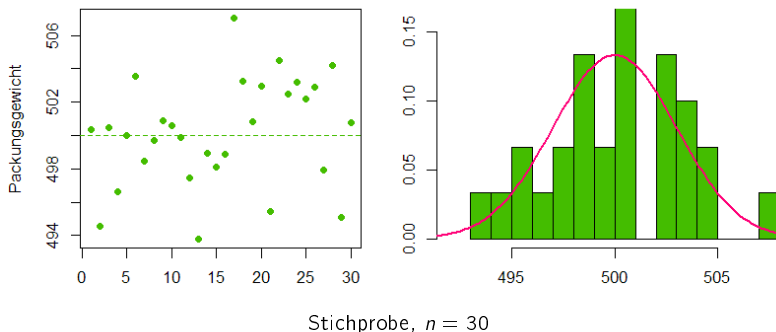
d.h.

$$P(F \leq 4.735) = 0.95.$$

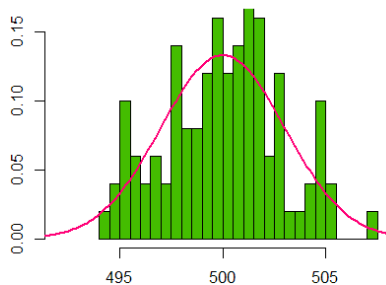
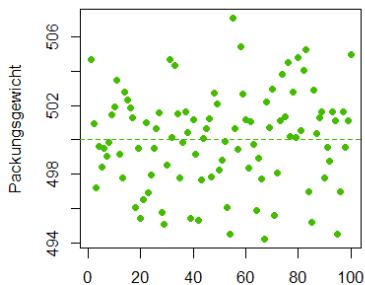
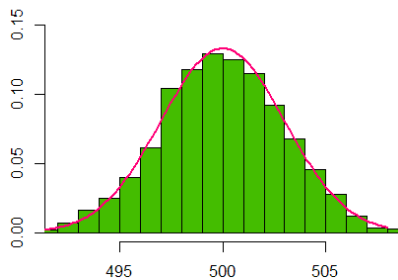
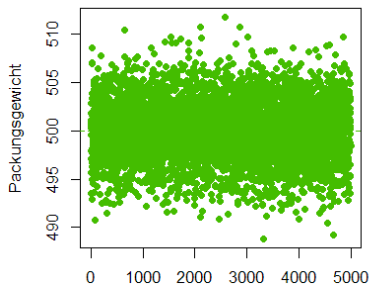
### 3.8.10. Ein Beispiel zum Schluss

■ **Beispiel B3.12:** In einer Fabrik wird Obst verpackt. Die Packungsgröße soll dabei jeweils 500g betragen, allerdings kommt es naturgemäß zu kleinen Schwankungen.

Das Gewicht  $X$  einer Obstpackung sei normalverteilt mit einem Mittelwert von  $\mu = 500g$  und einer Standardabweichung von  $\sigma = 3$ :

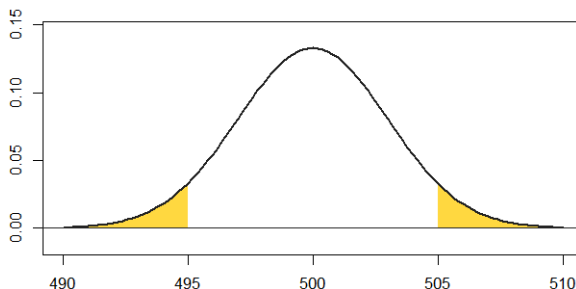




Stichprobe,  $n = 100$ Stichprobe,  $n = 5000$

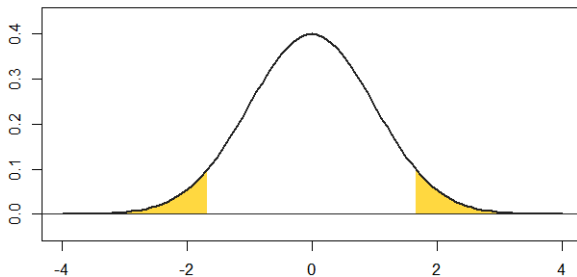
Nach einer Norm für den Obsthandel darf die Packungsgröße der Ware nicht um mehr als fünf Gramm vom angegebenen Gewicht abweichen.

Wie groß ist die Wahrscheinlichkeit einer solchen unzulässigen Abweichung?



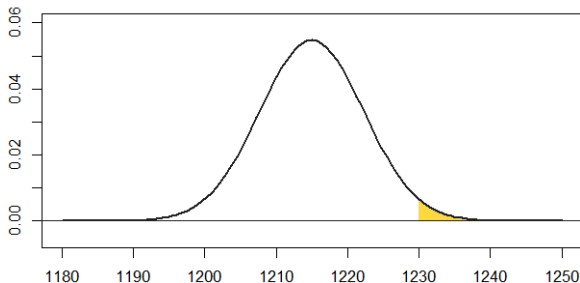
Wir transformieren  $X$  in eine standardnormalverteilte Zufallsvariable:

$$\begin{aligned} & P(X > 505 \text{ oder } X < 495) \\ &= 1 - P(X \in [495, 505]) \\ &= 1 - P\left(\frac{X - 500}{3} \in \left[\frac{495 - 500}{3}, \frac{505 - 500}{3}\right]\right) \\ &= 1 - P\left(\frac{X - 500}{3} \in [-5/3, 5/3]\right) \\ &= 1 - (\Phi(5/3) - \Phi(-5/3)) = 2(1 - \Phi(5/3)) = 0.075 \end{aligned}$$



In einem LKW sollen  $3 \cdot 3 \cdot 3 \cdot 90 = 2430$  der Obstpackungen transportiert werden, aber höchstens 1230 Kilogramm. Mit welcher Wahrscheinlichkeit ist das möglich?

Das Gesamtgewicht  $Y$  der 2430 Packungen ist normalverteilt mit  $E(Y) = 0.5 \cdot 2430 = 1215$  kg und  $\hat{\sigma}_*(Y) = 0.003 \cdot 2430 = 7.29$ .



$$\begin{aligned} P(Y \leq 1230) &= P\left(\frac{Y - 1215}{7.29} \leq \frac{1230 - 1215}{7.29}\right) \\ &= \Phi(2.058) = 0.98 \end{aligned}$$

## 4.

## Induktive Statistik

### 4.1. Punktschätzer

■ **Beispiel B4.1:** Bei einem Spiel ist dem Spieler die Wahrscheinlichkeit zu gewinnen nicht bekannt. In 20 Spielen hat er fünf Mal gewonnen. Wie kann der Spieler die Gewinnwahrscheinlichkeit schätzen?

■ **Beispiel B4.2:** In zehn Würfeln mit einem u.U. nicht fairen Würfel ist die Augensumme 41. Wie kann man den Erwartungswert der Augenzahl schätzen? Wie kann man die Varianz schätzen?

Gegeben seien unabhängige und identisch verteilte Zufallsvariablen

$$X_1, X_2, X_3, \dots, X_n,$$

eine sog. Stichprobe. Die gemeinsame Verteilung der  $X_i$  nennen wir auch Verteilung der Grundgesamtheit.

Wir schreiben

$$\mu = E(X_1)$$

für den gemeinsamen Erwartungswert und

$$\sigma^2 = \text{Var}(X_1)$$

$$\sigma = \hat{\sigma}_*(X_1)$$

für die Varianz und die Standardabweichung der Stichprobenelemente.

Eine Zufallsvariable  $S$ , die aus den Zufallsvariablen  $X_1$  bis  $X_n$  gebildet wird heißt Statistik.

Beispiele für Statistiken:

- $\sum_{i=1}^n X_i$ ,
- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,
- $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ ,
- $\frac{1}{n} \sum_{i=1}^n (X_i - E(X))^2$ ,
- $\min_{i=1,2,\dots,n} X_i$ ,
- $\max_{i=1,2,\dots,n} X_i$ .

Punktschätzer sind Statistiken, die geeignet sind, einzelne Parameter der zugrundeliegenden Verteilung zu schätzen.

Solche Parameter sind z.B.

- Die Erfolgswahrscheinlichkeit  $p$  der Bernoulli-Verteilung,
- $n$  oder  $p$  bei der Binomialverteilung,
- $p$  bei der geometrischen Verteilung,
- den Erwartungswert  $\mu$  oder die Varianz  $\sigma^2$ .

Wir schreiben  $\hat{\theta}$  für einen Punktschätzer des Parameters  $\theta$ , also z.B.  $\hat{\mu}$  für einen Punktschätzer des Erwartungswertes  $\mu$ , oder  $\hat{\sigma}$  für einen Punktschätzer der Standardabweichung.



### 4.1.1. Punktschätzer für den Erwartungswert

Es sei  $\mu$  der Erwartungswert der Zufallsvariablen  $X_1, X_2, X_3, \dots$

- Ein naheliegender Schätze für  $\mu$  ist der Mittelwert

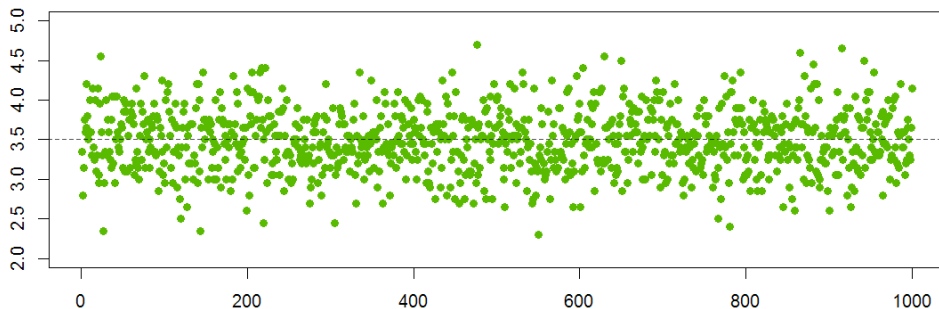
$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Dabei ist zu beachten, dass  $\hat{\mu}$ , im Gegensatz zur Zahl  $\mu$ , weiterhin eine Zufallsvariable ist, also eine Verteilung, einen Erwartungswert und eine Varianz besitzt.

- Wir haben schon früher den Erwartungswert der Zufallsvariablen  $\bar{X}$  berechnet. Es ergab sich

$$E(\hat{\mu}) = \mu.$$

Wir sagen:  $\hat{\mu}$  ist erwartungstreu, bzw. unverzerrt: Der geschätzte Wert ist im Mittel gleich dem zu schätzenden Wert.



Beispiel B1.1:  $\hat{\mu}$  für  $n = 20$ , 1000 Mal wiederholt.

- Es gilt ( $\Rightarrow$  Satz 3.8.3)

$$\hat{\sigma}_*(\hat{\mu}) = \frac{\sigma}{\sqrt{n}},$$

d.h. die Standardabweichung nimmt mit wachsendem  $n$  immer weiter ab

- Außerdem gilt

$$\lim_{n \rightarrow \infty} \hat{\sigma}_*(\hat{\mu}) = 0.$$

Wir sagen dann, dass  $\hat{\mu}$  ein konsistenter Schätzer ist.

- Im allgemeinen ist die Verteilung von  $\hat{\mu}$  nicht einfach zu beschreiben. Es gilt aber nach dem zentralen Grenzwertsatz

$$\hat{\mu} \stackrel{\text{annähernd}}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

für große Werte von  $n$ .

- Ist die Grundgesamtheit normalverteilt mit bekanntem  $\mu$  und bekanntem  $\sigma$ , dann ergibt sich, wie bereits oben gezeigt,

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

### 4.1.2. Punktschätzer für die Varianz bei bekanntem Erwartungswert

- Ist der Erwartungswert  $\mu$  bekannt, so ist die empirische Varianz

$$\hat{\sigma}_*^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

ein konsistenter und erwartungstreuer Schätzer, d.h.

$$E(\hat{\sigma}_*^2) = \text{Var}(X) \quad \text{und} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\sigma}_*^2) = 0.$$

- Ist die Grundgesamtheit normalverteilt, so besitzt die Zufallsvariable

$$n \cdot \frac{\hat{\sigma}_*^2}{\sigma^2}$$

hat eine Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden.

### 4.1.3. Punktschätzer für die Varianz bei unbekanntem Erwartungswert

Wenn man bei unbekanntem  $\mu$  den Ansatz

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

als Punktschätzer für die Varianz verwendet, so stellt sich heraus, dass der Erwartungswert dieses Schätzers  $\frac{n-1}{n}\sigma^2$  ist.

Um einen erwartungstreuen Schätzer der Varianz zu erhalten, müssen wir also den Schätzer

$$\hat{\sigma}^2 = \hat{\sigma}^2(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

verwenden.

- Dieser neue Schätzer ist erwartungstreu,

$$E(\hat{\sigma}^2) = \sigma^2,$$

und konsistent:

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\sigma}^2) = 0.$$

- Ist die Grundgesamtheit normalverteilt, so hat die Zufallsvariable

$$(n-1) \cdot \frac{\hat{\sigma}^2}{\sigma^2}$$

eine Chi-Quadrat-Verteilung mit  $(n-1)$  Freiheitsgraden.

## 4.2. Intervallschätzer

### 4.2.1. Intervallschätzer für den Erwartungswert bei bekannter Varianz

- Wir haben gesehen, dass der Mittelwert  $\hat{\mu}$  ein erwartungstreuer und konsistenter Schätzer für den Erwartungswert  $\mu$  ist.
- Es wäre interessant zu wissen, was man über die Abweichung  $|\mu - \hat{\mu}|$  sagen kann.
- Der Einfachheit halber gehen wir nun davon aus, dass
  1. die Grundgesamtheit normalverteilt ist, d.h. es gilt  $X_i \sim N(\mu, \sigma)$  und
  2. die Varianz  $\sigma^2$  bekannt ist.



- Dann ist  $\hat{\mu}$  normalverteilt mit Erwartungswert  $\mu$  und Standardabweichung  $\sigma/\sqrt{n}$ , d.h.

$$P\left(\hat{\mu} \leq \mu + c \frac{\sigma}{\sqrt{n}}\right) = P\left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq c\right) = \Phi(c)$$

für jedes Zahl  $c \in \mathbb{R}$ .

- Wenn wir  $c = z_{1-\alpha/2}$  (Quantil der Normalverteilung) wählen, so gilt

$$P\left(\hat{\mu} \leq \mu + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \frac{\alpha}{2}.$$

- Ebenso kann man zeigen:

$$P\left(\hat{\mu} \leq \mu - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = \frac{\alpha}{2}$$

- Es ergibt sich dann

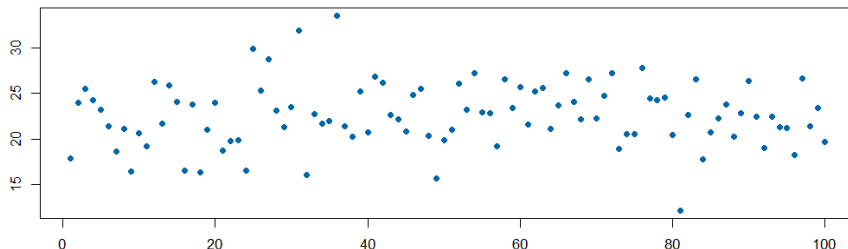
$$P\left(\hat{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- Das zufällige Intervall

$$\left[ \hat{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

heißt  $(1 - \alpha) \cdot 100\%$ -Konfidenzintervall. Es enthält (als Zufallsgröße verstanden, also solange es noch nicht konkret anhand vorliegender Daten ausgerechnet wurde) mit Wahrscheinlichkeit  $1 - \alpha$  den zu schätzenden Parameter  $\mu$ .

■ **Beispiel B4.3:** Die Temperaturen an einem Ort werden 100 Jahre lang jeweils am 1.Juni gemessen. Angenommen die Standardabweichung der Temperaturen betrage 4 Grad und die Temperaturen seien normalverteilt.



Es ergibt sich als Schätzer für den Erwartungswert der Temperatur


$$\hat{\mu} = 22.6$$

Als 95%-Konfidenzintervall ( $\alpha = 0.05$ ) erhalten wir dann

$$\begin{aligned} & \left[ \hat{\mu} - z_{0.975} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{0.975} \frac{\sigma}{\sqrt{n}} \right] \\ &= \left[ 22.6 - \frac{1.96 \cdot 4}{10}, 22.6 + \frac{1.96 \cdot 4}{10} \right] \\ &= [21.82, 23.38]. \end{aligned}$$

Als 90%-Konfidenzintervall ( $\alpha = 0.1$ ) berechnen wir

$$\begin{aligned} & \left[ \hat{\mu} - z_{0.95} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{0.95} \frac{\sigma}{\sqrt{n}} \right] \\ &= \left[ 22.6 - \frac{1.645 \cdot 4}{10}, 22.6 + \frac{1.645 \cdot 4}{10} \right] \\ &= [21.94, 23.26]. \end{aligned}$$

 Der Erwartungswert  $\mu$  liegt nicht mit 90% bzw. 95% Wahrscheinlichkeit in diesen Intervallen!  $\mu$  ist eine feste Zahl, keine Zufallsvariable.

### 4.2.2. Intervallschätzer für den Erwartungswert bei unbekannter Varianz

- Ist die Varianz unbekannt, so muss sie geschätzt werden:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Um allerdings

$$P\left(\hat{\mu} \leq \mu + c \frac{\hat{\sigma}}{\sqrt{n}}\right) = P\left(\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \leq c\right)$$

zu berechnen, benötigen wir die Verteilung der Zufallsvariablen

$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}.$$

- Man kann zeigen, dass  $T$  eine t-Verteilung mit  $(n-1)$ -Freiheitsgraden besitzt.

- Wenn wir  $c = t_{n-1, 1-\alpha/2}$  (Quantil der t-Verteilung) wählen, so gilt

$$P\left(\hat{\mu} \leq \mu + t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \frac{\alpha}{2}.$$

und

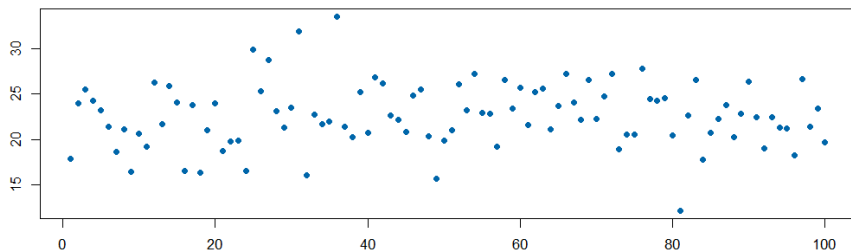
$$P\left(\hat{\mu} \leq \mu - t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}\right) = \frac{\alpha}{2}.$$

- Wir erhalten das  $(1 - \alpha) \cdot 100\%$ -Konfidenzintervall

$$\left[ \hat{\mu} - t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right],$$

dass den zu schätzenden Parameter  $\mu$  mit Wahrscheinlichkeit  $1 - \alpha$  enthält (solange noch kein konkretes Intervall berechnet wurde).

■ **Beispiel B4.4**  $\Rightarrow$  B4.3: Die Temperaturen an einem Ort werden 100 Jahre lang jeweils am 1. Juni gemessen. Angenommen die Temperaturen seien normalverteilt mit unbekanntem  $\mu$  und unbekanntem  $\sigma^2$ .



Die Punktschätzer für den Erwartungswert und die Varianz (Standardabweichung) der Temperatur sind

$$\begin{aligned}\hat{\mu} &= 22.6 \\ \hat{\sigma}^2 &= 12.25, \quad (\hat{\sigma} = 3.5)\end{aligned}$$

Als 95%-Konfidenzintervall ( $\alpha = 0.05$ ) erhalten wir dann

$$\begin{aligned} & \left[ \hat{\mu} - t_{99,0.975} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{99,0.975} \frac{\hat{\sigma}}{\sqrt{n}} \right] \\ &= \left[ 22.6 - \frac{1.984 \cdot 3.5}{10}, 22.6 + \frac{1.984 \cdot 3.5}{10} \right] \\ &= [21.91, 23.29]. \end{aligned}$$



### 4.2.3. Intervallschätzer für die Varianz bei bekanntem Erwartungswert

- Ist  $\mu$  bekannt, so ist  $\hat{\sigma}_*^2$  unser erwartungstreuer Schätzer für die Varianz und es gilt

$$P(\sigma^2 \leq c\hat{\sigma}_*^2) = P\left(n \cdot \frac{\hat{\sigma}_*^2}{\sigma^2} \geq \frac{n}{c}\right) = 1 - F(n/c),$$

wobei  $F$  die Verteilungsfunktion einer Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden bezeichnet.

- Wir setzen  $c = \frac{n}{\chi_{n,\alpha/2}}$  bzw.  $c = \frac{n}{\chi_{n,1-\alpha/2}}$  und erhalten

$$P\left(\sigma^2 \leq \frac{n\hat{\sigma}_*^2}{\chi_{n,\alpha/2}}\right) = 1 - \frac{\alpha}{2},$$
$$P\left(\sigma^2 \leq \frac{n\hat{\sigma}_*^2}{\chi_{n,1-\alpha/2}}\right) = \frac{\alpha}{2}.$$

- Dann ergibt sich

$$P\left(\frac{n\hat{\sigma}_*^2}{\chi_{n,1-\alpha/2}} \leq \sigma^2 \leq \frac{n\hat{\sigma}_*^2}{\chi_{n,\alpha/2}}\right) = 1 - \alpha.$$

- Wir erhalten das  $(1 - \alpha) \cdot 100\%$ -Konfidenzintervall

$$\left[ \frac{n\hat{\sigma}_*^2}{\chi_{n,1-\alpha/2}}, \frac{n\hat{\sigma}_*^2}{\chi_{n,\alpha/2}} \right].$$

#### 4.2.4. Intervallschätzer für die Varianz bei unbekanntem Erwartungswert

- Ist  $\mu$  unbekannt, so verwenden den Schätzer  $\hat{\sigma}^2$ .
- Es gilt dann, ganz ähnlich wie im Fall bekannten Erwartungswertes,

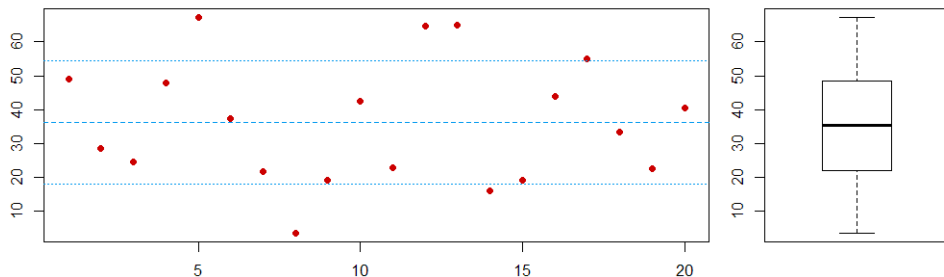
$$P(\sigma^2 \leq c\hat{\sigma}^2) = P\left((n-1) \cdot \frac{\hat{\sigma}^2}{\sigma^2} \geq \frac{n-1}{c}\right) = 1 - F((n-1)/c),$$

wobei  $F$  die Verteilungsfunktion einer Chi-Quadrat-Verteilung mit  $(n-1)$  Freiheitsgraden bezeichnet.

- Wie oben ergibt sich das  $(1-\alpha) \cdot 100\%$ -Konfidenzintervall

$$\left[ \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1, 1-\alpha/2}}, \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1, \alpha/2}} \right].$$

■ **Beispiel B4.5:** Es seien  $X_1, X_2, \dots, X_{20}$  die Ausgaben von zwanzig Kunden in einem bestimmten Supermarkt. Wir gehen von einer Normalverteilung  $X_i \sim N(\mu, \sigma)$  der Grundgesamtheit aus.



Die Punktschätzer für den Erwartungswert und die Varianz (Standardabweichung) sind:

$$\begin{aligned}\hat{\mu} &= 36.23 \\ \hat{\sigma}^2 &= 327.94 \quad (\hat{\sigma} = 18.11)\end{aligned}$$

Wir erhalten die Intervallschätzer ( $\alpha = 10\%$ )

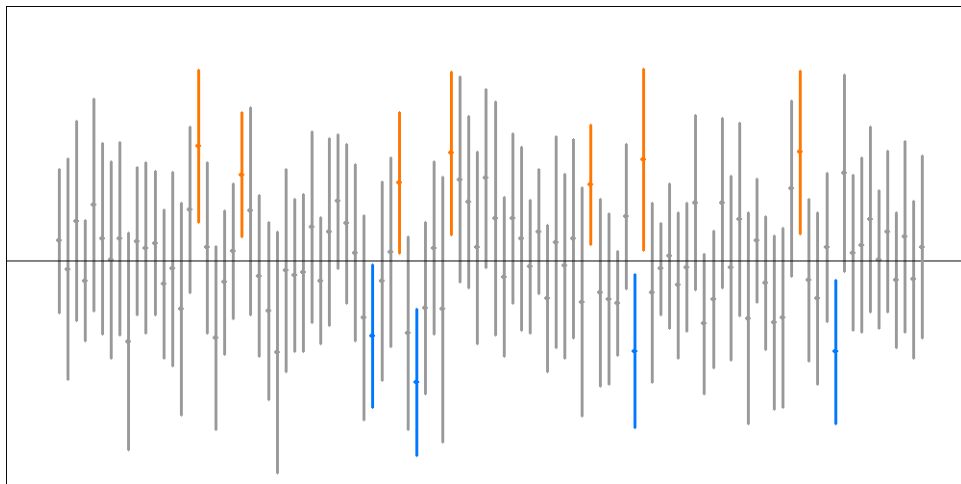
$$\begin{aligned} & \left[ \hat{\mu} - t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \\ &= [29.23, 43.23] \end{aligned}$$

für den Erwartungswert und

$$\begin{aligned} & \left[ \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1, 1-\alpha/2}}, \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1, \alpha/2}} \right] \\ &= [206.70, 615.87] \quad ([14.3, 24.82]) \end{aligned}$$

für die Varianz (bzw. Standardabweichung).

90%-Konfidenzintervalle für  $\mu$  für 100 Supermärkte:



### 4.2.5. Schätzen „ohne Zurücklegen“

- Wird eine Stichprobe ohne Zurücklegen aus einer endlichen Grundgesamtheit der Größe  $N$  gezogen, so sind die Zufallsvariablen  $X_1, X_2, \dots, X_n$  nicht mehr unabhängig.
- Der Mittelwert  $\hat{\mu} = \bar{X}$  ist weiterhin ein erwartungstreuer konsistenter Schätzer für den wahren Erwartungswert  $\mu$ .
- Allerdings ist der Schätzer für die Varianz nicht länger erwartungstreu. Ein erwartungstreuer und konsistenter Schätzer ist nun

$$\hat{\sigma}^2 = \frac{N-1}{N} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- Offensichtlich liegt der Korrekturfaktor  $(N-1)/N$  nahe bei eins, wenn  $N$  sehr groß ist.

## 4.3. Hypothesentests

### 4.3.1. Idee

- Bei einem statistischen Test versucht man anhand von Daten, den Wahrheitsgehalt von Hypothesen zu bestimmen.
- Meistens handelt es sich um Hypothesen, die die wahre Verteilung der Stichprobe betreffen, z.B. die Hypothesen
  - über den Erwartungswert,
  - über die Varianz,
  - über den Median oder Quartile,
  - über die Verteilung.

Es kann auch eine Hypothese über den Zusammenhang oder über Unabhängigkeit von Merkmalen getestet werden.



- Meistens wird zunächst eine Nullhypothese  $H_0$  formuliert, z.B., dass der Erwartungswert  $\mu$  einen bestimmten Wert  $\mu_0$  hat:

$$H_0 : \quad \mu = \mu_0.$$

- Eine einfache Hypothese liegt vor, wenn wir, wie im Fall oben, annehmen, dass ein Verteilungsparameter einen bestimmten Wert annimmt. Ansonsten ist die Hypothese zusammengesetzt.
- Die Alternative  $H_1$  beschreibt eine zweite Hypothese (die Gegenhypothese), die nur dann eintreten kann, wenn  $H_0$  nicht eintritt, z.B.

$$H_1 : \quad \mu > \mu_0$$

$$\text{oder} \quad H_1 : \quad \mu \neq \mu_0.$$

Häufig handelt es sich bei  $H_1$  um das logische Komplement von  $H_0$ .

Die generelle Vorgehensweise bei einem Hypothesentest ist:

1. Wir stellen eine Hypothese auf und formulieren sie mathematisch.
2. Wir finden eine passende Teststatistik  $T$ .
3. Wir finden einen sinnvollen Ablehnungsbereich  $A$  derart, dass wir die Hypothese dann ablehnen, wenn  $T$  nach Auswertung der Stichprobe in  $A$  liegt.

■ **Beispiel B4.6**  $\Rightarrow_{B1.1}$ : Wir haben den Verdacht, dass bei unserem Würfelexperiment zu Beginn der Vorlesung die Drei häufiger erschien, als gewöhnlich. Es sei  $X_1, \dots, X_{120}$  eine Stichprobe von Augenzahlen.

1. Es sei  $p$  die Wahrscheinlichkeit einer Drei. Dann stellen wir die Nullhypothese

$$H_0 : \quad p = 1/6.$$

auf. Die Alternative wäre  $H_1 : \quad p > 1/6$ .

2. Als Teststatistik wählen wir die Anzahl  $T$  der Dreier bei  $n$  Würfeln:

$$T = \#\{X_i | X_i = 3\}$$

3. und lehnen ab, wenn  $T > 20 + C$  ist, wobei wir  $C$  noch passend wählen müssen. Es ist also  $A = (20 + C, \infty)$ .

### 4.3.2. Wahl des Ablehnungsbereiches

- Es stellt sich die Frage, wie wir einen passenden und sinnvollen Ablehnungsbereich finden können.

Meistens ergeben sich aus der Hypothese bereits Ansatzpunkte, z.B., dass  $A$ , wie im obigen Beispiel, ein bestimmtes Intervall ist, bei dem noch die Intervallgrenzen zu bestimmen sind.

- Nach welchen Kriterien soll man  $A$  wählen?
- Wir überlegen uns, dass wir insgesamt zwei wichtige Fehler machen können:
  1. Fehler erster Art: Wir lehnen die Hypothese ab, obschon sie zutrifft.
  2. Fehler zweiter Art: Wir lehnen die Hypothese nicht ab, obschon sie nicht zutrifft.

- Üblicherweise wird nun bei einem statistischen Hypothesentest der Ablehnungsbereich  $A$  so festgelegt, dass die Wahrscheinlichkeit eines Fehlers erster Art eine bestimmte, vorher festgelegte Schwelle, das Signifikanzniveau  $\alpha$ , nicht überschreitet.
- Dazu benötigt man natürlich die Verteilung von  $T$  unter  $H_0$  (d.h. wenn  $H_0$  gilt).
- Warum sollte man nicht versuchen,  $A$  so festzulegen, dass die Wahrscheinlichkeit eines Fehlers erster Art minimal wird?

### 4.3.3. Vorgehensweise

1. Formulierung der Hypothese
2. Finden einer geeigneten Teststatistik  $T$ , deren Verteilung unter  $H_0$  bekannt ist.
3. Festlegen eines Signifikanzniveaus  $\alpha$ .
4. Angabe eines Ablehnungsbereiches mit

$$P(T \in A | H_0) = \alpha.$$

5. Konkrete Berechnung der Teststatistik  $t$  anhand der Daten.
6. Ablehnen der Hypothese genau dann, wenn  $t \in A$  gilt.

■ **Beispiel B4.7**  $\Rightarrow$  B1.1: Die Anzahl  $T$  der Dreier bei 120 Würfeln ist binomialverteilt mit Erfolgswahrscheinlichkeit  $p$ . Wir setzen  $\alpha = 0.01$ .

Wir lehnen die Hypothese  $p = 1/6$  ab, wenn  $T > 20 + C$  ist.

Die Wahrscheinlichkeit eines Fehlers erster Art ist:

$$P(T > 20 + C | H_0) = \sum_{k=20+C}^{120} \binom{120}{k} (1/6)^k (5/6)^{120-k}$$

Es ist sehr aufwendig  $C$  so zu bestimmen, dass

$$P(T > 20 + C | H_0) = 0.01$$

gilt.

Wir verwenden den zentralen Grenzwertsatz in folgender sehr bekannter Form:

■ **Satz 4.1 (Satz von Moivre-Laplace)**

Ist  $T$  binomialverteilt, so konvergiert die Verteilung von

$$\frac{T - np}{\sqrt{np(1-p)}}$$

für  $n \rightarrow \infty$  gegen eine Standardnormalverteilung.

Entsprechend haben wir die Näherung

$$P(T \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$



Also gilt

$$\begin{aligned} P(T > 20 + C | H_0) &\approx 1 - \Phi\left(\frac{C}{\sqrt{n\frac{1}{6}(1 - \frac{1}{6})}}\right) \\ &= 1 - \Phi\left(\frac{C}{\sqrt{100/6}}\right) \stackrel{!}{=} 0.01 \end{aligned}$$

genau dann, wenn

$$C = \sqrt{100/6} \cdot z_{0.99} = 4.0825 \cdot 2.3264 = 9.4973$$

ist, d.h. unser Ablehnungsbereich ist

$$A = (29.4973, \infty).$$

Bei 30 Dreiern, wie im Beispiel **B1.1**, würden wir also zum 1%-Niveau die Hypothese  $p = 1/6$  zu Gunsten der Alternative  $p > 1/6$  ablehnen!

#### 4.3.4. Die Gütefunktion

Angenommen unsere Hypothese beinhaltet einen Parameter  $\theta$  (z.B. den Erwartungswert  $\mu$  oder die Varianz  $\sigma^2$ ).

- Die Gütefunktion

$$G(x) = P(T \in A | \theta = x)$$

beschreibt die Wahrscheinlichkeit, die Hypothese abzulehnen, wenn  $\theta = x$  ist.

- Bei einem Signifikanzniveau  $\alpha$  gilt  $G(x) \leq \alpha$  wenn  $x$  in dem Bereich liegt, wo die Nullhypothese gilt.

### 4.3.5. Der p-Wert

Bei einem Hypothesentest beschreibt der p-Wert die Wahrscheinlichkeit, bei einer erneuten Stichprobe eine Teststatistik  $T$  zu beobachten, die unplausibler ist, als die konkret beobachtete Statistik  $t$ .

- Ist  $A = [a, \infty)$  (rechtsseitiger Test), so ergibt sich

$$p = P(T \geq t | H_0).$$

- Ist  $A = (-\infty, b]$  (linksseitiger Test), so ergibt sich

$$p = P(T \leq t | H_0).$$

- Ist  $A = (-\infty, b] \cup [b, \infty)$  (zweiseitiger Test), so ergibt sich

$$p = P(|T| \geq |t| | H_0).$$

- Ist der p-Wert klein, so ist der Wert  $t$  der Teststatistik als extrem anzusehen und daher die Nullhypothese abzulehnen.
- Ist der p-Wert groß, so ist der Wert  $t$  der Teststatistik als eher durchschnittlich anzusehen und daher die Nullhypothese nicht abzulehnen.
- Bei einem Signifikanztest zum Signifikanzniveau  $\alpha$  (vor dem Test festzulegen) lehnen wir die Nullhypothese genau dann ab, wenn

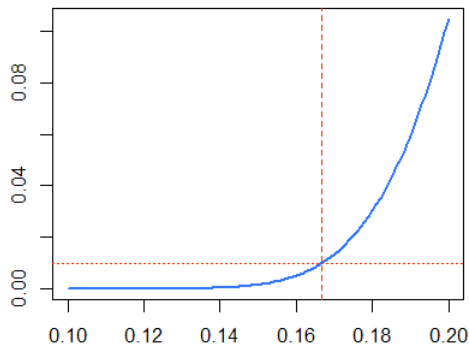
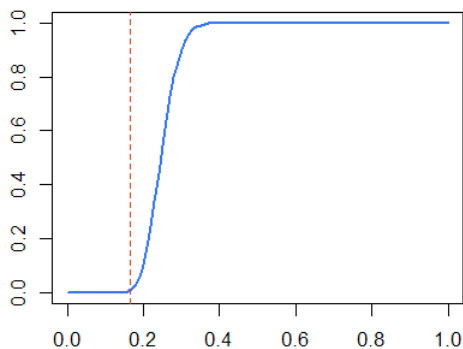
$$p \leq \alpha$$

ist.

- Computersoftware berechnet heute bei Hypothesentests immer auch den zugehörigen p-Wert. Eine Kenntnis des Wertes der Teststatistik und des Ablehnungsbereichs ist dann in der Regel nicht mehr notwendig.

■ **Beispiel B4.8**  $\Rightarrow_{B1.1}$ : Für das Würfelbeispiel B1.1 ergibt sich die Gütefunktion

$$G(x) = P(T > 29.4973 | p = x) \\ \approx 1 - \Phi\left(\frac{29.4973 - 120x}{\sqrt{120x(1-x)}}\right).$$



Unsere Teststatistik

$$T \stackrel{\sim}{=} \text{Anzahl der Dreier}$$

hatte den konkreten Wert  $t = 30$  angenommen.

Es ergibt sich der p-Wert

$$p = P(T > 30) \approx 1 - \Phi\left(\frac{10}{\sqrt{100/6}}\right) = 0.0072,$$

d.h. wir würden die Hypothese  $p = 1/6$  zu jedem Niveau  $> 0.72\%$  ablehnen.

#### 4.3.6. Einstichprobentests für den Erwartungswert bei normalverteilter Grundgesamtheit

Wir gehen wieder von einer normalverteilten Grundgesamtheit aus und wollen die Hypothese

$$\mu = \mu_0$$

gegen die Alternative

- $\mu \neq \mu_0$  (zweiseitiger Test) bzw.
- $\mu > \mu_0$  (rechtsseitiger Test) oder
- $\mu < \mu_0$  (linksseitiger Test) testen.

Dabei ist  $\mu_0$  ein fester vorgegebener Wert (der hypothetische Erwartungswert).

## (1) Test bei bekannter Varianz

- In dem eher unrealistischen Fall bekannter Varianz  $\sigma^2$  wählen wir als Teststatistik

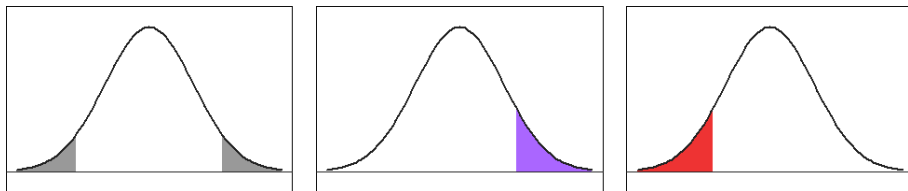
$$T = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \sim N(0, 1).$$

- Es ergeben sich die Ablehnungsbereiche

$$A = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty),$$

$$A = (z_{1-\alpha}, \infty),$$

$$A = (-\infty, -z_{1-\alpha}).$$





- Wir lehnen also in folgenden Fällen ab:

$$|T| > z_{1-\alpha/2}, \quad T > z_{1-\alpha}, \quad T < -z_{1-\alpha}.$$

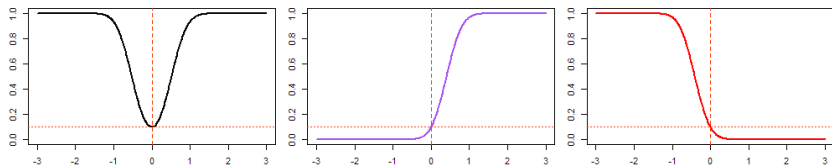
- Für die p-Wert ergibt sich

$$p = P(|T| > |t| | H_0) = 2(1 - \Phi(|t|)),$$

$$p = P(T > t | H_0) = 1 - \Phi(t),$$

$$p = P(T < t | H_0) = \Phi(t).$$

- Gütefunktion ( $\mu_0 = 0$ ,  $\sigma = 1$ ,  $\alpha = 10\%$ ):



## (2) Test bei unbekannter Varianz (t-Test)

Im Normalfall wird die Varianz, wie der Erwartungswert, nicht bekannt sein. In dem Fall schätzen wir  $\sigma^2$  durch den erwartungstreuen Schätzer  $\hat{\sigma}^2$  und verwenden die t-verteilte Teststatistik

$$T = \sqrt{n} \frac{\bar{X} - \mu_0}{\hat{\sigma}} \sim t(n-1).$$

- Es ergeben sich die Ablehnungsbereiche

$$A = (-\infty, -t_{n-1, 1-\alpha/2}) \cup (t_{n-1, 1-\alpha/2}, \infty),$$

$$A = (t_{n-1, 1-\alpha}, \infty),$$

$$A = (-\infty, -t_{n-1, 1-\alpha}).$$

- Wir lehnen also in folgenden Fällen ab:

$$|T| > t_{n-1, 1-\alpha/2},$$

$$T > t_{n-1, 1-\alpha},$$

$$T < -t_{n-1, 1-\alpha}.$$

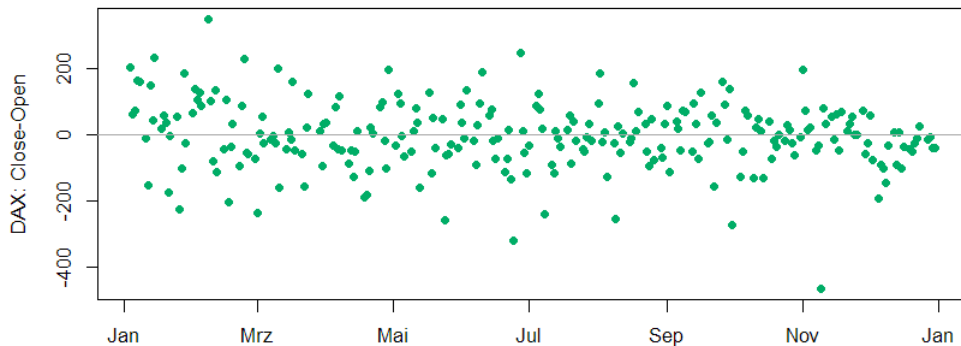
- Wir erhalten die p-Werte

$$p = P(|T| > |t| | H_0) = 2(1 - F_{n-1}(|t|)),$$

$$p = P(T > t | H_0) = 1 - F_{n-1}(t),$$

$$p = P(T < t | H_0) = F_{n-1}(t).$$

Hier bezeichnet  $F_{n-1}$  die Verteilungsfunktion der t-Verteilung mit  $(n-1)$  Freiheitsgraden.

**■ Beispiel B4.9:** Tägliche Renditen für den DAX, 2016 (Quelle: Yahoo)

Wir wollen zum Niveau 10% testen, ob  $\mu = 0$  gilt:

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0.$$

Es ergibt sich in diesem Fall

$$\begin{aligned} t &= \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\hat{\sigma}} \\ &= \sqrt{255} \cdot \frac{-3.085}{101.84} = -0.484, \end{aligned}$$

mit dem Schätzer für die Standardabweichung

$$\hat{\sigma} = \sqrt{\frac{\sum_{k=1}^n (x_k - \mu)^2}{n-1}} = 101.84.$$

Es ist  $t_{254,0.95} = 1.651$  also  $|t| < t_{254,0.95}$  d.h.  $H_0$  wird nicht abgelehnt.

Alternative: Als p-Wert ergibt sich

$$p = 2(1 - F_{254}(0.412)) = 0.629$$

so dass wir zu allen üblichen Signifikanzniveaus  $H_0$  nicht ablehnen.

#### 4.3.7. Einstichprobentests für die Varianz bei normalverteilter Grundgesamtheit

Wir gehen von einer normalverteilten Grundgesamtheit aus und wollen die Hypothese

$$\sigma^2 = \sigma_0^2$$

gegen die Alternative

- $\sigma^2 \neq \sigma_0^2$  (zweiseitiger Test) bzw.
- $\sigma^2 > \sigma_0^2$  (rechtsseitiger Test) oder
- $\sigma^2 < \sigma_0^2$  (linksseitiger Test) testen.

Die hypothetische Varianz  $\sigma_0$  ist dabei ein fest vorgegebener Wert.

## (1) Test bei bekanntem Erwartungswert

- Bei bekanntem  $\mu$  wählen wir als Teststatistik

$$T = n \cdot \frac{\widehat{\sigma}_*^2}{\sigma_0^2} \sim \chi^2(n).$$

- Es ergeben sich die Ablehnungsbereiche

$$A = [0, \chi_{n,\alpha/2}) \cup (\chi_{n,1-\alpha/2}, \infty),$$

$$A = (\chi_{n,1-\alpha/2}, \infty),$$

$$A = [0, \chi_{n,\alpha/2}).$$

- Wir lehnen also in folgenden Fällen ab:

$$T < \chi_{n,\alpha/2} \text{ oder } T > \chi_{n,1-\alpha/2}$$

$$T > \chi_{n,1-\alpha},$$

$$T < \chi_{n,\alpha}.$$

- p-Werte:

$p =$  (komplizierter)

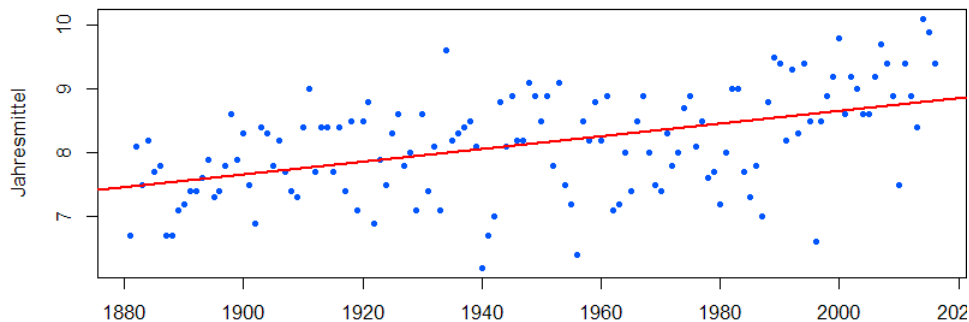
$$p = P(T > t | H_0) = 1 - F_n(t),$$

$$p = P(T < t | H_0) = F_n(t).$$

Hier bezeichnet  $F_n$  die Verteilungsfunktion der Chi-Quadrat-Verteilung mit  $n$  Freiheitsgraden.

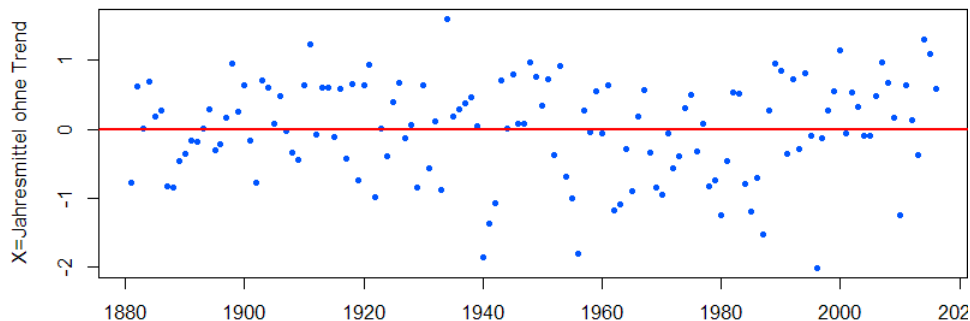


■ **Beispiel B4.10:** Jahresmitteltemperaturen in Sachsen, 1881-2016  
(Quelle: DWD):



Dies ist ein Beispiel für eine [Zeitreihe](#). Offenbar existiert ein gewisser Trend, den man mit Hilfe der Zeitreihenanalyse (Kleinste-Quadrate-Methode) herausrechnen kann.

Jahresmittelwerte ohne Trend ( $x_1, x_2, \dots, x_{136}$ ):



Wir wollen die Hypothese  $H_0 : \text{Var}(X) = \sigma^2 = 0.4$  (=Varianz der Daten für Bayern) mit einem zweiseitigen statistischen Test zum Signifikanzniveau  $\alpha = 5\%$  untersuchen. Dabei können wir für den Erwartungswert  $E(X) = \mu = 0$  annehmen.

Als erstes schätzen wir die Varianz mit Hilfe der empirischen Varianz

$$\begin{aligned}\hat{\sigma}_*^2 &= \frac{\sum_{k=1}^n (x_k - \mu)^2}{n} \\ &= \frac{\sum_{k=1}^{136} x_k^2}{136} = 0.495.\end{aligned}$$

Dann bestimmen wir den Wert der Teststatistik:

$$t = n \cdot \frac{\hat{\sigma}_*^2}{\sigma_0^2} = 168.2.$$

Für die beiden relevanten Quartile ergibt sich  $\chi_{136,0.025} = 105.61$  und  $\chi_{136,0.975} = 170.18$ . Da  $t \in [105.61, 170.18]$ , lehnen wir  $H_0$ , d.h. die Hypothese, dass die Varianz 0.4 ist, nicht ab.

## (2) Test bei unbekanntem Erwartungswert

- Bei nicht bekanntem  $\mu$  ergibt sich als Teststatistik

$$T = (n-1) \cdot \frac{\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-1).$$

- Es ergeben sich die Ablehnungsbereiche

$$A = [0, \chi_{n-1, \alpha/2}) \cup (\chi_{n-1, 1-\alpha/2}, \infty),$$

$$A = (\chi_{n-1, 1-\alpha/2}, \infty),$$

$$A = [0, \chi_{n-1, \alpha/2}).$$

- Wir lehnen also in folgenden Fällen ab:

$$T < \chi_{n-1, \alpha/2} \text{ oder } T > \chi_{n-1, 1-\alpha/2}$$

$$T > \chi_{n-1, 1-\alpha},$$

$$T < \chi_{n-1, \alpha}.$$

- p-Werte:

$p$  = (komplizierter)

$$p = P(T > t | H_0) = 1 - F_{n-1}(t),$$

$$p = P(T < t | H_0) = F_{n-1}(t).$$

Hier bezeichnet  $F_{n-1}$  die Verteilungsfunktion der Chi-Quadrat-Verteilung mit  $(n - 1)$  Freiheitsgraden.

#### 4.3.8. Zweistichprobentest auf gleiche Erwartungswerte (t-Test)

- Wir betrachten nun den Fall zweier normalverteilter unabhängiger Stichproben  $X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma)$  und  $Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma)$  mit gleichen, unbekannten Varianzen.
- Das Problem eines entsprechenden Tests mit möglicherweise ungleichen Varianzen ist schwerer zu lösen ([Behrens-Fisher-Problem](#), [Welch-Test](#)).
- Wir wollen also die Hypothese

$$H_0 : \mu_1 = \mu_2$$

gegen die Alternative

$$H_1 : \mu_1 \neq \mu_2$$

testen.

- Wir verwenden die Teststatistik

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

die unter  $H_0$  eine t-Verteilung mit  $(n+m-2)$  Freiheitsgraden besitzt.

- Ablehnungsbereich:

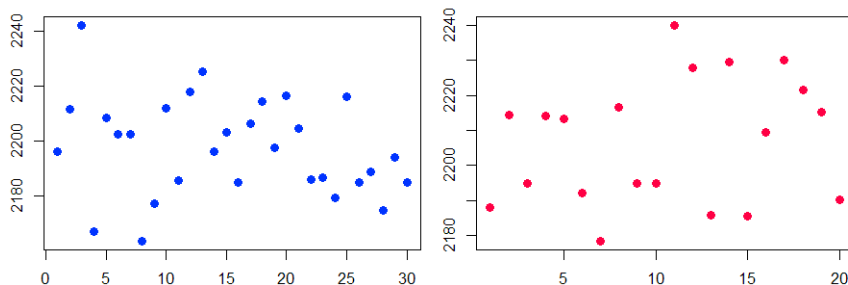
$$A = (-\infty, -t_{n+m-2, 1-\alpha/2}) \cup (t_{n+m-2, 1-\alpha/2}, \infty).$$

d.h. wir lehnen ab, falls  $|T| > t_{n+m-2, 1-\alpha/2}$  ist.

- P-Wert:

$$p = P(|T| > |t| | H_0) = 2(1 - F_{n+m-2}(|t|)).$$

■ **Beispiel B4.11:** Zwei Maschinen stellen Bauteile mit einem Gewicht  $X$  bzw.  $Y$  her (Angaben in Gramm). Es ist bekannt, dass beide Maschinen bei der Produktion Fehler mit derselben (unbekannten) Varianz  $\sigma^2$  machen. Es wird eine Stichprobe von 30, bzw. 20 Bauteilen untersucht.



Wir wollen zu einem Signifikanzniveau von 10% die Hypothese untersuchen, dass die Mittelwerte der Bauteilgewichte für beide Maschinen identisch sind.



Wir erhalten

$$\bar{x} = 2197.571, \quad \bar{y} = 2206.815$$

$$\hat{\sigma}_1^2 = 320.3355, \quad \hat{\sigma}_2^2 = 323.2014$$

$$t = \frac{-9.244066}{124.2198} \cdot \sqrt{24} = -1.786$$

$$p = 2 * (1 - F_{48}(1.786)) = 0.08042$$

$H_0$  wird zu jedem Niveau  $\alpha > 0.08042$  abgelehnt, also auch in unserem Fall.

#### 4.3.9. Zweistichprobentest auf gleiche Varianzen (F-Test)

- Wir betrachten den Fall zweier normalverteilter unabhängiger Stichproben  $X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1)$  und  $Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma_2)$ .
- Wir wollen nun die Hypothese

$$H_0 : \sigma_1^2 = \sigma_2^2$$

gegen die Alternative

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

testen.

- Die Teststatistik

$$T = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}$$

besitzt eine  $F$ -Verteilung mit  $n - 1$  und  $m - 1$  Freiheitsgraden.

- Ablehnungsbereich:

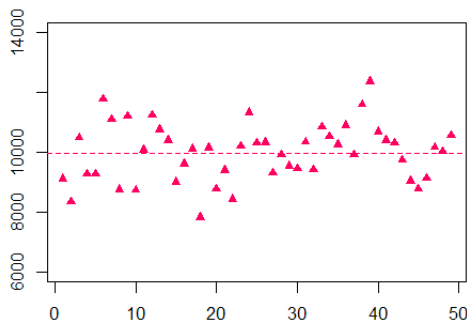
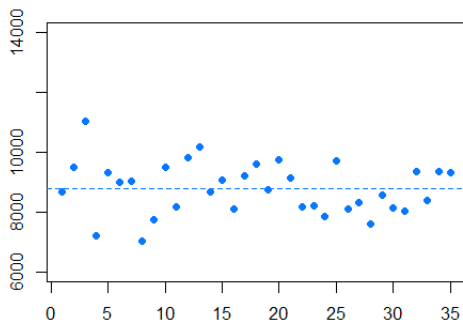
$$A = [0, F_{(n-1, m-1), \alpha/2}) \cup (F_{(n-1, m-1), 1-\alpha/2}, \infty).$$

- Wir lehnen ab, wenn

$$T < F_{(n-1, m-1), \alpha/2} \text{ oder } T > F_{(n-1, m-1), 1-\alpha/2}$$

ist.

■ **Beispiel B4.12:** Fünf bzw. sieben Wochen lang wird jeden Tag von 16 bis 17 Uhr die Verkehrsdichte (Fahrzeuge/h) an zwei Ausfahrtstraßen einer Großstadt aufgezeichnet.



Es ist

$$\hat{\sigma}_1^2 = 749347.3, \quad \hat{\sigma}_2^2 = 913983$$

Wir wollen zum Signifikanzniveau  $\alpha = 10\%$  testen, ob die Varianzen

gleich sind:

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Es ist

$$T = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = 0.82$$

und

$$F_{(34,48),0.05} = 0.582, \quad F_{(34,48),0.95} = 1.672.$$

Wir lehnen also die Hypothese nicht ab. Mit Hilfe von Statistiksoftware kann man den p-Wert  $p = 0.548$  berechnen.

#### 4.3.10. Chi-Quadrat-Anpassungstest

Wir wollen jetzt Hypothesen der Form  $H_0 : F = F_0$  testen. Dabei ist

- $F$  die (wahre und unbekannte) Verteilungsfunktion der Grundgesamtheit,
- $F_0$  unsere hypothetische Verteilungsfunktion.

Gegeben seien

- eine Stichprobe  $X_1, X_2, \dots, X_n$  von unabhängigen Beobachtungen,
- Klassen  $K_1, K_2, \dots, K_m$  (u.U. auch aus einzelnen Ausprägungen bestehend),
- absolute Häufigkeiten  $n(K_i)$  und zu erwartende Klassenhäufigkeiten für den Fall, dass  $H_0$  zutrifft,  $n_e(K_i) = n \cdot P(X \in K_i)$ .

■ **Beispiel B4.13**  $\Rightarrow_{B1.1}$ : Für unser ursprüngliches Würfelbeispiel ergibt sich, wenn unsere Hypothese die diskrete Gleichverteilung betrifft:

Augenzahl:	1	2	3	4	5	6
$n(K_i)$ :	15	18	30	18	21	18
$n_e(K_i)$ :	20	20	20	20	20	20

Als Testvariable könnten wir die absoluten Abstände

$$\sum_{k=1}^m |n(K_i) - n_e(K_i)|$$

verwenden. Es stellt sich heraus, dass eine etwas anders gewählte Statistik besser geeignet ist.

- Die Chi-Quadrat-Statistik ist gegeben durch

$$T = \sum_{k=1}^m \frac{(n(K_i) - n_e(K_i))^2}{n_e(K_i)}.$$

- $T$  besitzt unter  $H_0$  asymptotisch (also für  $n \rightarrow \infty$ ) eine Chi-Quadrat-Verteilung mit  $(m - 1)$  Freiheitsgraden. Für jede Schätzung eines weiteren Parameters verringert sich diese Zahl um eins.
- Wir lehnen die Hypothese ab, wenn  $T > \chi_{m-1, 1-\alpha}$  ist.
- Als p-Wert ergibt sich

$$p = P(T > t | H_0) = 1 - F_{m-1}(t),$$

wo  $F_{m-1}$  die Verteilungsfunktion der Chi-Quadrat-Verteilung ist.



■ **Beispiel B4.14**  $\Rightarrow_{B1.1}$ :

Augenzahl:	1	2	3	4	5	6
$n(K_i)$ :	15	18	30	18	21	18
$n_e(K_i)$ :	20	20	20	20	20	20

Es ist

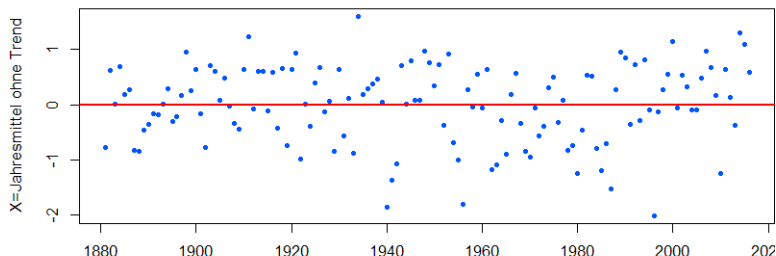
$$t = \frac{25 + 4 + 100 + 4 + 1 + 4}{20} = 6.9$$

und

$$p = 1 - F_5(6.9) = 0.2281843$$

Wir lehnen die Hypothese zu keinem vernünftigen Signifikanzniveau ab und gehen dementsprechend bis auf weiteres von einer diskreten Gleichverteilung (fairer Würfel) aus.

■ **Beispiel B4.15**  $\Rightarrow$  B4.10: Jahresmitteltemperaturen in Sachsen, 1881-2016, ohne Trend:



Einteilung in Klassen:

$(-3,-2]$	$(-2,-1]$	$(-1,0]$	$(0,1]$	$(1,2]$
1	11	50	69	5

Liegt eine Normalverteilung vor?

Es ist  $\hat{\mu} = 0$  und  $\hat{\sigma}_* = 0.703$ , also ergibt sich für unsere Hypothese

$$H_0 : X \sim N(0, 0.703).$$

$K_i$	$(-3,-2]$	$(-2,-1]$	$(-1,0]$	$(0,1]$	$(1,2]$
$n(K_i)$	1	11	50	69	5
$\Phi(a_i/\hat{\sigma}_*)$	0.001	0.023	0.159	0.500	0.841
$\Phi(b_i/\hat{\sigma}_*)$	0.023	0.159	0.500	0.841	0.977
$\Phi(b_i/\hat{\sigma}_*) - \Phi(a_i/\hat{\sigma}_*)$	0.021	0.136	0.341	0.341	0.136
$n_e(K_i)$	2.9	18.5	46.4	46.4	18.5

$$t = \sum_{k=1}^m \frac{(n(K_i) - n_e(K_i))^2}{n_e(K_i)} = 25.42$$

Die Teststatistik  $T$  hat etwa eine Chi-Quadrat-Verteilung mit  $5 - 1 - 1 = 3$  Freiheitsgraden (wir haben ja die Varianz geschätzt!).

Es ist

$$p = 1 - F_3(25.42) = 1.26 \cdot 10^{-5}.$$

Wir lehnen die Hypothese zu allen gängigen Signifikanzniveaus ab.

### 4.3.11. Weitere Tests auf Normalität

Es gibt noch eine Reihe weiterer Tests auf Normalität, für die allerdings die Anwendung von Statistiksoftware notwendig ist.

- Der Shapiro-Wilks-Test liefert für das obige Beispiel:

Test Name:	Shapiro-Wilk normality test
Data:	t
Test Statistic:	W = 0.9774639
P-value:	0.02354978

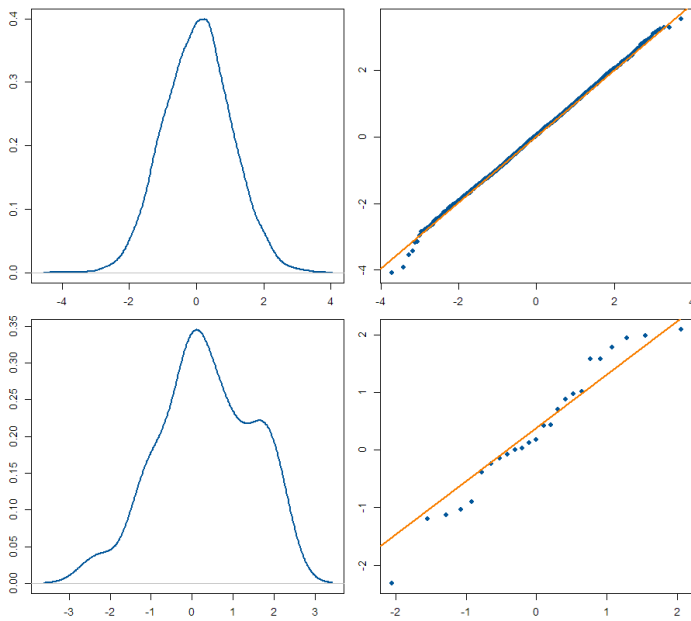
- Beim Lilliefors-Test ergibt sich:

Test Name:	Lilliefors (Kolmogorov -
+ Smirnov) normality test	
Data:	t
Test Statistic:	D = 0.07416185
P-value:	0.0641023

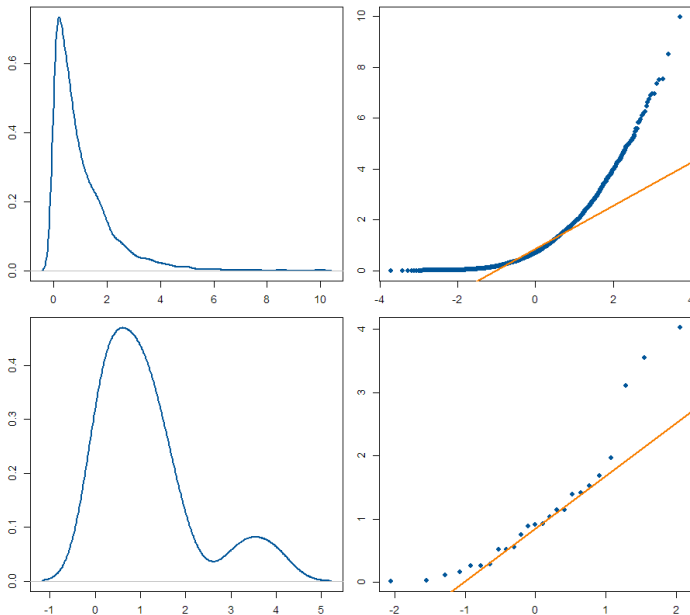
#### 4.3.12. Q-Q-Plots

- Optisch besteht die Möglichkeit sich mit Hilfe eines sog. Q-Q-Plots (Quantil-Quantil-Plot) von der Normalität der Daten zu überzeugen.
- Dabei werden die Quantile der Normalverteilung und die empirischen Quantile der vorliegenden Daten in einem Diagramm aufgetragen. Außerdem wird eine Hilfsgerade berechnet und aufgetragen.
- Im Fall einer vorliegenden Normalverteilung liegen die Punkte etwa auf der angegebenen Geraden.

- Etwa normalverteilte Daten (oben:  $n=5000$ , unten:  $n=25$ ):

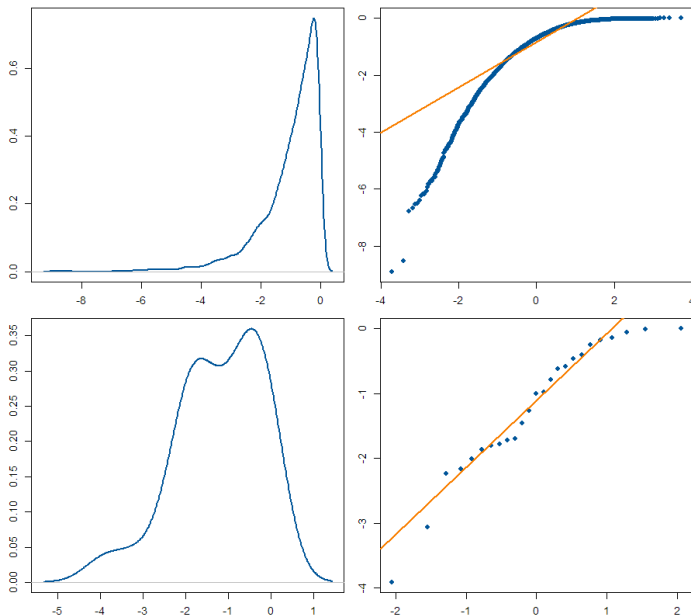


● Rechtsschiefe Daten:

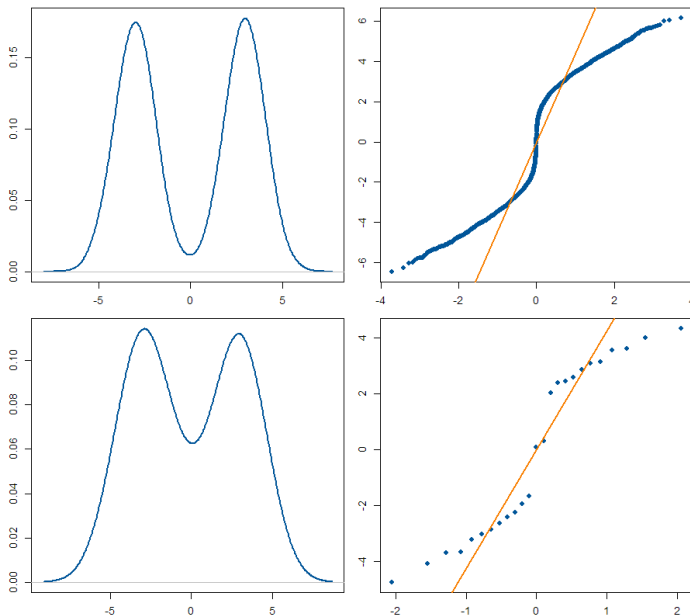




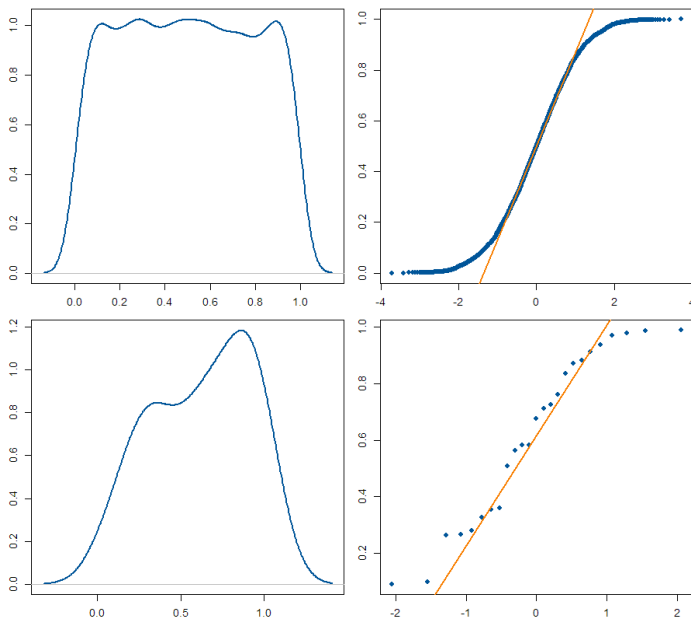
● Linksschiefe Daten:



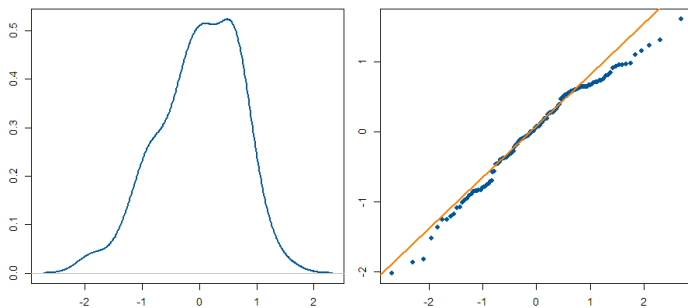
● Bimodale Daten:



● Beschränkter Träger:



- Wir erhalten im obigen Beispiel:



### 4.3.13. Der Chi-Quadrat-Homogenitätstest

Wir wollen jetzt testen ob zwei unabhängige Stichproben  $X_1, X_2, \dots, X_n$  und  $Y_1, Y_2, \dots, Y_m$  ein und dieselbe Verteilung besitzen:

$$H_0 : F_1 = F_2.$$

Wir verwenden folgende Größen:

- Klassen  $K_1, K_2, \dots, K_k$  (u.U. auch aus einzelnen Ausprägungen bestehend),
- absolute Häufigkeiten:

Klasse:	1	2	...	k	
X	$n_{1,1}$	$n_{1,2}$	...	$n_{1,k}$	$n_{1\bullet} = n$
Y	$n_{2,1}$	$n_{2,2}$	...	$n_{2,k}$	$n_{2\bullet} = m$
	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	$n + m$

- Als Teststatistik dient der Chi-Quadrat-Koeffizient

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n+m}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n+m}}$$

der etwa eine  $\chi^2$ -Verteilung mit  $k - 1$  Freiheitsgraden besitzt.

- Ablehnung der Hypothese, falls  $\chi^2 > \chi_{k-1, 1-\alpha}$  ist.
- P-Wert, falls  $\chi^2 = c$  ist:

$$p = P(\chi^2 > c) = 1 - F_{k-1}(c),$$

wobei  $F_{k-1}$  die entsprechende Chi-Quadrat-Verteilungsfunktion ist.

■ **Beispiel B4.16:** Die Besuchszahlen des Oktoberfestes werden für zwei Jahre ( $X, Y$ ) an jeweils 30 Tagen verglichen.

Klassen:	0-30	30-50	50-70	70-90	90-110	$n_{j\bullet}$
X	2	9	10	6	3	30
Y	0	1	12	15	2	30
$n_{\bullet j}$	2	10	22	21	5	60

Liegen für  $X$  und  $Y$  identische Verteilungen vor?

Wir testen bei einem Signifikanzniveau von  $\alpha = 0.01$ . Es ist

$$\chi^2 = 12.639$$

und

$$\chi_{4,0.99} = 13.2767$$

wir lehnen also  $H_0$  nicht ab.

Alternativ können wir den p-Wert berechnen und erhalten:

$$p = 1 - F_4(12.639) = 0.01318.$$



### 4.3.14. Der Chi-Quadrat-Unabhängigkeitstest

Wir wollen jetzt testen ob zwei Merkmale  $X$  und  $Y$  unabhängig sind:

$$H_0 : \quad X \text{ und } Y \text{ unabhängig}$$

Voraussetzungen:

- Stichproben  $X_1, X_2, \dots, X_n$  und  $Y_1, Y_2, \dots, Y_m$ ,
- Klassen oder Ausprägungen  $K_1, K_2, \dots, K_k$  und  $L_1, L_2, \dots, L_r$ ,
- absolute Häufigkeiten:

	$K_1$	$K_2$	$\dots$	$K_k$	
$L_1$	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,k}$	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$L_r$	$n_{r,1}$	$n_{r,2}$	$\dots$	$n_{r,k}$	$n_{r\bullet}$
$Y$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet k}$	$n + m$

- Als Teststatistik dient erneut der Chi-Quadrat-Koeffizient

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{\left(n_{ij} - \frac{n_{i\bullet} \cdot n_{\bullet j}}{n+m}\right)^2}{\frac{n_{i\bullet} \cdot n_{\bullet j}}{n+m}}$$

der etwa eine  $\chi^2$ -Verteilung mit  $\ell = (r-1) \cdot (k-1)$  Freiheitsgraden besitzt.

- Ablehnung der Hypothese, falls  $\chi^2 > \chi_{\ell, 1-\alpha}$  ist.
- P-Wert, falls  $c$  die berechnete Teststatistik ist:

$$p = P(\chi^2 > c) = 1 - F_{\ell}(c),$$

wobei  $F_{\ell}$  die entsprechende Chi-Quadrat-Verteilungsfunktion ist.

■ **Beispiel B4.17:** (Vergleiche mit Aufgabe 46) An einer Hochschule starten 140 Studierende ins erste Semester. Sie können zwischen 3 Studiengängen A,B,C und D wählen. Sind die beiden Merkmale  $X \stackrel{\sim}{=} \text{Studiengang}$  und  $Y \stackrel{\sim}{=} \text{Geschlecht}$  unabhängig?

	A	B	C	D	
m	10	30	10	5	55
w	20	20	40	5	85
	30	50	50	10	140

Wir testen zum Niveau  $\alpha = 0.1$ .

Wir erhalten

$$\chi^2 = 17.718$$

und

$$\chi_{3,0.9} = 6.251389.$$

Wir lehnen also ab.

In der Tat ist

$$p = 0.0005028544 < 0.01.$$

#### 4.3.15. Test auf Ausreißer

- Ein „Ausreißer“ ist ein Datenwert, der außergewöhnlich weit von den übrigen, bzw. von den meisten anderen Daten entfernt liegt. Es gibt keine genaue mathematische Definition.
- Der Grubbs-Test kann Ausreißer feststellen. Dazu wird angenommen, dass die Grundgesamtheit normalverteilt ist und die Teststatistik

$$T = \frac{\max_{i=1,\dots,n} |x_i - \bar{x}|}{\hat{\sigma}}$$

berechnet.

- Die Nullhypothese „es liegt kein Ausreißer vor“ wird abgelehnt, wenn

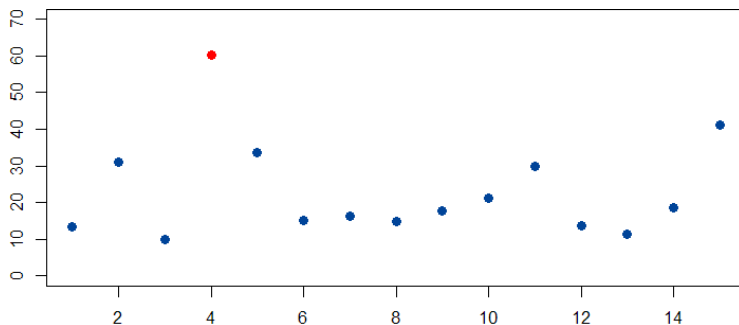
$$t > c^* = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{n-2, \alpha/2n}^2}{n-2 + t_{n-2, \alpha/2n}^2}}$$

ist.

- Wird die Hypothese abgelehnt, so kann man den verdächtigen Datenwert entfernen und einen neuen Test starten.
- Dieses Verfahren wird solange durchgeführt, bis kein Ausreißer mehr erkannt wird
- ⚠ Das Entfernen von Datenpunkten muss sich aus dem jeweiligen Zusammenhang rechtfertigen lassen. Im Normalfall dürfen keine Daten entfernt werden!

■ **Beispiel B4.18:** Ein handschriftlich notierter ursprünglich normalverteilter Datensatz weist u.U. Zahlendreher auf:

13.3, 31.1, 10.0, 60.2, 33.7, 15.2, 16.2,  
14.9, 17.7, 21.1, 29.8, 13.6, 11.4, 18.7, 41.1



Wir verwenden den Grubbs-Test zum Niveau 10%.

Es ist  $\bar{x} = 23.2$ ,  $\hat{\sigma}(x) = 13.748$  und

$$t = \frac{37}{13.748} = 2.691,$$
$$c^* = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{n-2, \alpha/2n}^2}{n-2 + t_{n-2, \alpha/2n}^2}} = 2.409.$$

Wir lehnen also die Nullhypothese ab.

Wir entfernen den Datenwert 60.2 und erhalten im zweiten Durchlauf

$$t = 2.157, \quad c^* = 2.372.$$

Wir lehnen die Nullhypothese nicht ab, belassen also alle übrigen Werte im Datensatz.



## 4.4. Einfache lineare Regression

In der einfachen linearen Regression versucht man lineare Zusammenhänge zwischen zwei Größen  $X$  und  $Y$  nachzuweisen. Dabei ist

- $X$  eine für uns nicht zufällige, also deterministische Größe (die erklärende Variable, exogene Variable oder Regressor), nach Ermittlung einer Stichprobe konkret gegeben durch Datenpunkte  $x_1, x_2, \dots, x_n$  und
- $Y$  eine zufällige Größe (die zu erklärende Variable, endogene Variable oder Regressand), konkret gegeben durch eine Stichprobe  $y_1, y_2, \dots, y_n$ .
- Zu jedem Datenelement  $x_i$  gehört eindeutig eine Stichprobe  $y_i$ .

- Idealerweise läge ein linearer Zusammenhang vor:

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

mit zwei unbekannten Regressionsparametern  $\beta_0, \beta_1$ .

- Tatsächlich werden allerdings noch gewisse Fehler- oder Störterme  $Z$  auftreten, so dass dann

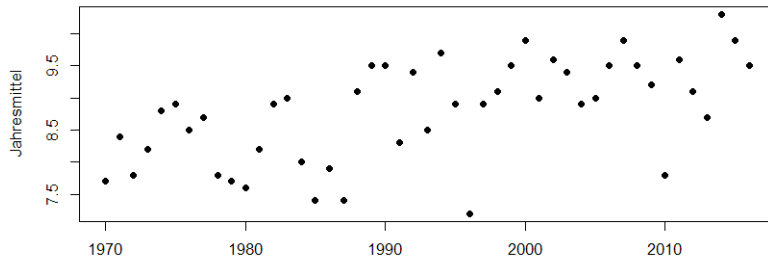
$$Y = \beta_0 + \beta_1 \cdot X + Z,$$

gilt.

- Wenn wir annehmen, dass  $E(Z) = 0$  ist, dann können wir auch schreiben:

$$E(Y|X = x) = \beta_0 + \beta_1 \cdot x.$$

■ **Beispiel B4.19:** Wir betrachten die Jahresmitteltemperaturen in Deutschland für den Zeitraum 1970-2016 (Quelle: DWD):

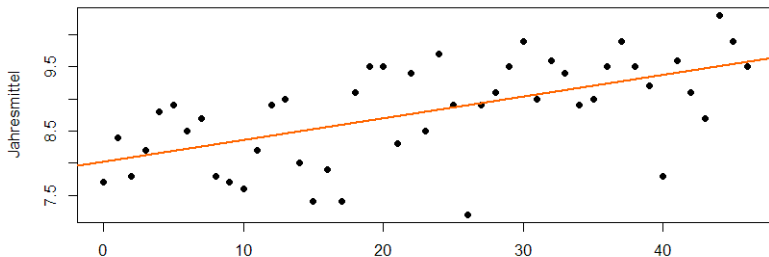


Es ist hier

$X \cong$  Zeit seit 1970 (on Jahren)

$Y \cong$  Jahresmitteltemperatur Deutschland

Wir nehmen an, es gäbe einen linearen Trend.



Mathematische Formulierung:

$$E(Y|X = x) = \beta_0 + \beta_1 \cdot x.$$

**⚠** Die beiden Regressionsparameter  $\beta_0$  und  $\beta_1$  sind prinzipiell unbekannt und können statistisch niemals mit 100%er Sicherheit ermittelt werden. Wir werden sie schätzen müssen...

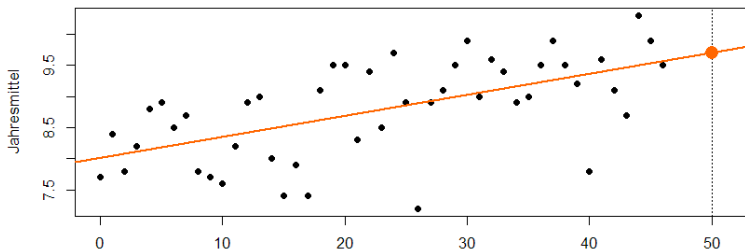
- In der Praxis liegen konkrete Daten  $x_1, x_2, \dots, x_n$  und  $y_1, y_2, \dots, y_n$  vor und es gilt i.A. nicht

$$y_k = \beta_0 + \beta_1 \cdot x_k,$$

sondern

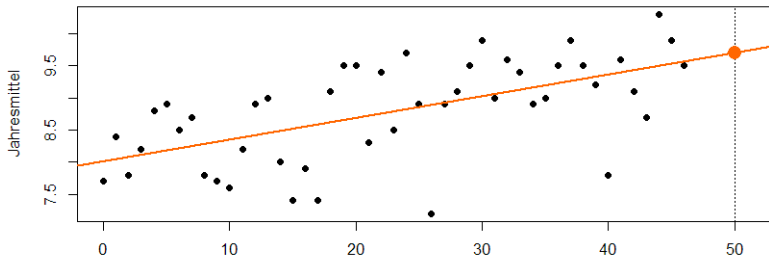
$$y_k = \beta_0 + \beta_1 \cdot x_k + z_k,$$

mit konkreten, aber prinzipiell unbekannten Fehlern  $z_1, z_2, \dots, z_n$ .



### 4.4.1. Die Kleinste-Quadrate-Methode

- Wie müssen die unbekannten Parameter  $\beta_0$  und  $\beta_1$  schätzen, also anhand der Daten möglichst gute Schätzer  $\hat{\beta}_0, \hat{\beta}_1$  berechnen.



- Die Ausgleichs- oder Regressionsgerade sollte so verlaufen, dass sie die Daten möglichst gut beschreibt.
- Was bedeutet „möglichst gut“?

- Wir versuchen die Regressionsparameter so zu wählen, dass der quadratische Fehler

$$Q^2 = \sum_{i=1}^n \left( y_i - (\hat{\beta}_0 + \hat{\beta}_1 x) \right)^2$$

möglichst klein wird.

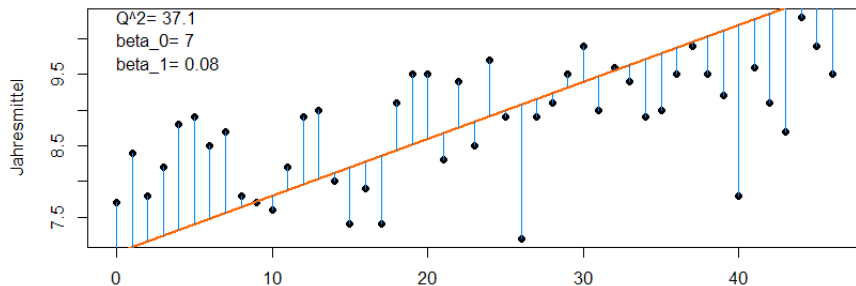
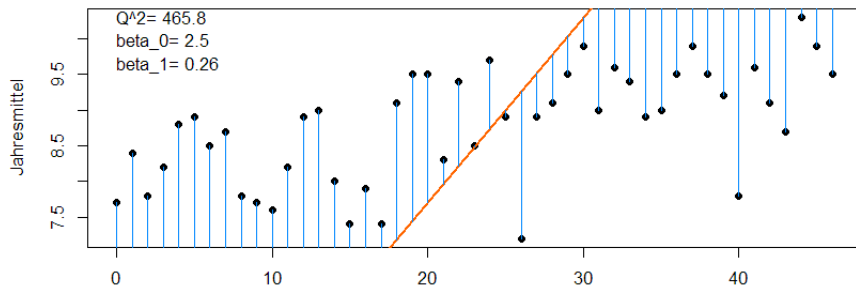
- Die auf diese Art und Weise minimierten Fehler

$$\hat{z}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x) \quad (4.1)$$

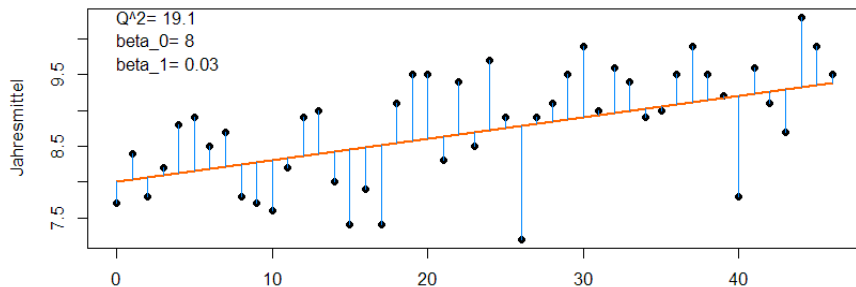
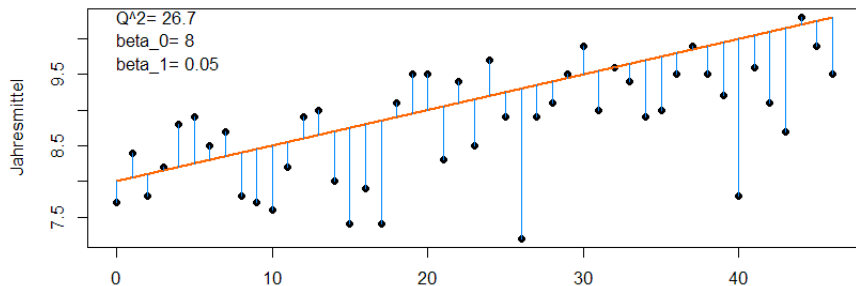
nennen wir Residuen.

- Wir minimieren also die Summe der Residuenquadrate:

$$Q^2 = \sum_{i=1}^n \hat{z}_i^2.$$







- Mit Hilfe der Analysis (Extremwertbestimmung bei Funktionen mit mehreren Variablen, s. Mathe-Vorlesung) kann man die Funktion  $Q^2(\hat{\beta}_0, \hat{\beta}_1)$  minimieren.
- Es ergibt sich dann für die Steigung der Regressionsgeraden

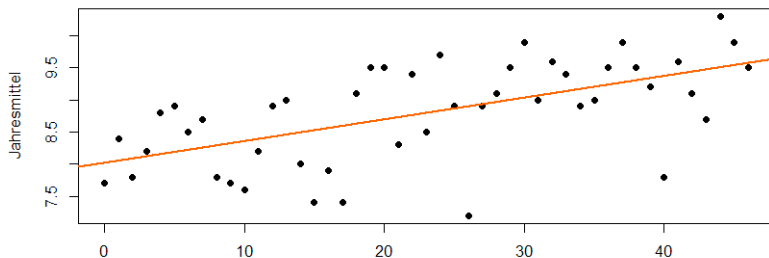
$$\hat{\beta}_1 = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{\hat{\sigma}_*^2(x)}$$

und für den Achsenabschnitt ([Intercept](#))

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- Speziell liegt der Schwerpunkt  $(\bar{x}, \bar{y})$  immer auf der Regressionsgeraden.

■ **Beispiel B4.20**  $\Rightarrow$  B4.19:



$$\overline{x \cdot y} = 208.464, \quad \bar{x} = 23, \quad \bar{y} = 8.794$$

$$\overline{x \cdot y} - \bar{x} \cdot \bar{y} = 6.211, \quad \overline{x^2} - \bar{x}^2 = 184$$

$$\hat{\beta}_1 = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = 0.0338, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8.017.$$

Interpretation: Die Temperatur steigt mit jedem Jahr um 0.0338 Grad.

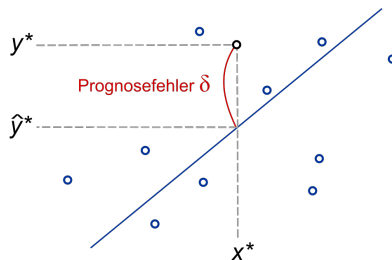
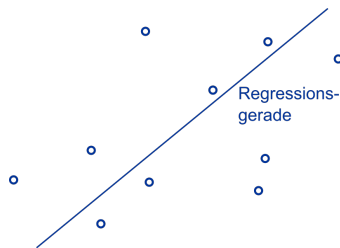
### 4.4.2. Prognosen

- Mit Hilfe der K-Q-Schätzer für das lineare Modell können wir für ein beliebiges  $x_*$  einen Schätzer  $\hat{y}_*$  für das unbekannte zugehörige  $y_*$  berechnen:

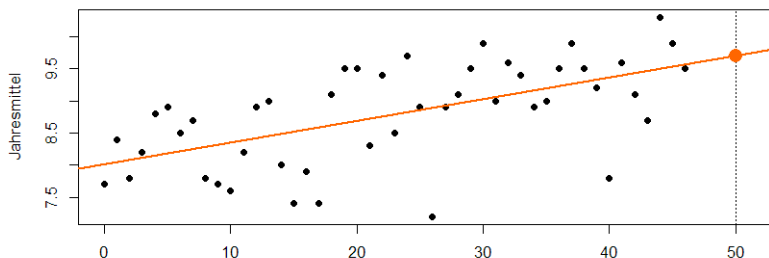
$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_*.$$

- Dabei machen wir naturgemäß einen Fehler, den Prognosefehler

$$\delta = \hat{y}_* - y_*$$



■ **Beispiel B4.21**  $\Rightarrow_{B4.19}$ : Jahresmitteltemperaturen in Deutschland, 1970-2016:



Für das obige Beispiel ergibt sich für die Jahresmitteltemperatur des Jahres 2020:

$$\hat{y} = 8.017 + 0.0338 \cdot 50 = 9.705$$

also knapp 9.7 Grad Celsius.

### 4.4.3. Standardbedingungen und Güte der Schätzer

- Normalerweise fordert man von den Residuen folgende Eigenschaften:

- $Z_i \sim N(0, \sigma_{\text{res}})$  (Normalverteilung der Störterme, mit Erwartungswert null und Homoskedasitizität, d.h. identische Varianzen  $\sigma_{\text{res}}^2$ ),
- $r_{Z_i Z_j} = 0$  für  $i \neq j$  (keine Autokorrelation).

Wir wollen das ab jetzt voraussetzen.

- Unter diesen Bedingungen sind  $\hat{\beta}_0$  und  $\hat{\beta}_1$  jeweils normalverteilt:

$$\hat{\beta}_0 \sim N \left( \beta_0, \sqrt{\sigma_{\text{res}}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n\hat{\sigma}_*^2(x)} \right)} \right),$$

$$\hat{\beta}_1 \sim N \left( \beta_1, \sqrt{\frac{\sigma_{\text{res}}^2}{n\hat{\sigma}_*^2(x)}} \right).$$

- ⚠ Die beiden Koeffizienten sind nicht stochastisch unabhängig!

**■ Satz 4.2**

Unter den genannten Voraussetzungen sind die beiden K-Q-Schätzer  $\hat{\beta}_0$  und  $\hat{\beta}_1$  erwartungstreu und konsistent, d.h.

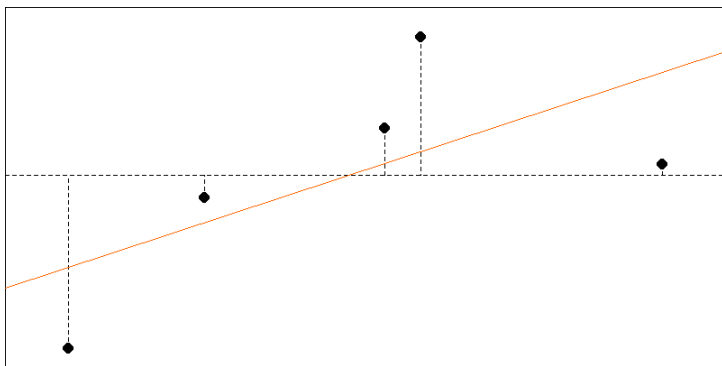
$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0, & \lim_{n \rightarrow \infty} \text{Var}(\hat{\beta}_0) &= 0, \\ E(\hat{\beta}_1) &= \beta_1, & \lim_{n \rightarrow \infty} \text{Var}(\hat{\beta}_1) &= 0. \end{aligned}$$

Außerdem besitzen sie die sog. BLUE-Eigenschaft, d.h. die Varianzen der beiden Schätzer sind jeweils kleiner als die Varianzen aller anderen linearen erwartungstreuer Schätzer (die Schätzer sind effizient).

#### 4.4.4. Das Bestimmtheitsmaß

- Die  $y$ -Datenwerte besitzen für die verschiedenen  $x_i$  jeweils unterschiedliche Werte. Die resultierende Streuung um den Mittelwert wird durch die Stichprobenvarianz beschrieben:

$$\hat{\sigma}^2(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

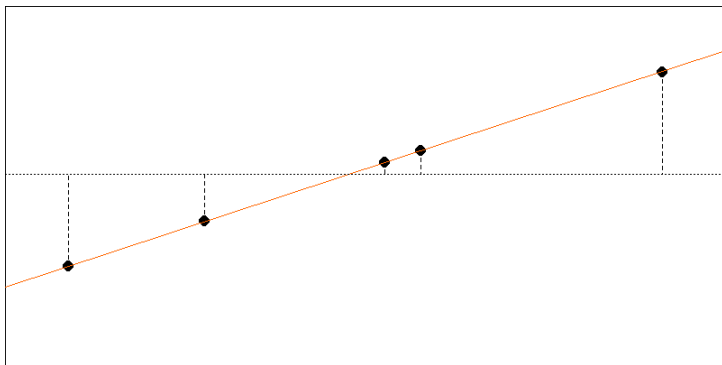




- Die „erklärte Varianz“

$$\hat{\sigma}_e^2(y) = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

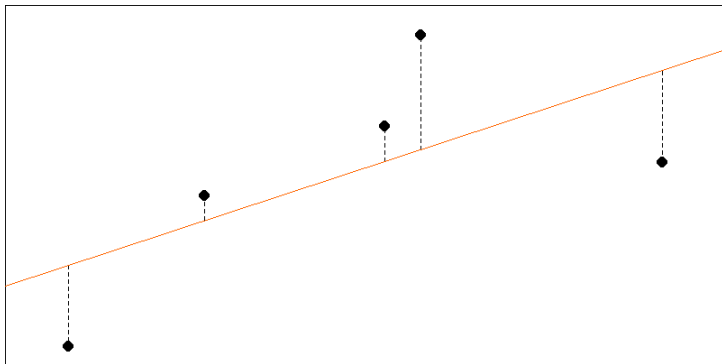
misst Abweichungen der Schätzungen vom y-Mittelwert.



- Die „nicht erklärte“ Varianz der Residuen

$$\hat{\sigma}_u^2(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y})^2$$

misst die Streuung um die Regressionsgerade.



**■ Satz 4.3 (Varianzzerlegung)**

Es gilt

$$\hat{\sigma}^2(y) = \hat{\sigma}_e^2(y) + \hat{\sigma}_u^2(y).$$

- Je höher der Anteil der erklärten Varianz an der Gesamtvarianz ausfällt, desto besser ist unser Modell angepasst.
- Der Anteil

$$R^2 = \frac{\hat{\sigma}_e^2(y)}{\hat{\sigma}^2(y)}$$

der erklärten Varianz an der Gesamtvarianz von  $y$  ist ein Maß für die Güte des Modells. Man nennt  $R^2$  das Bestimmtheitsmaß.

- Je höher  $R^2$  ausfällt, desto besser ist das Modell an die vorliegenden Daten angepasst, d.h. desto besser erklärt  $x$  die Variable  $y$ .
- Es gibt keine generelle Richtlinie, wie hoch  $R^2$  ausfallen muss, damit von einer guten Anpassung geredet werden kann. Werte  $< 0.3$  deuten allerdings eine schlechte Anpassung an.
- $R^2$  nimmt zu, wenn weitere erklärende Variablen hinzugezogen werden, auch wenn sich das Modell durch die Hinzunahme nicht verbessert.

In diesem Fall verwendet man auch das [korrigierte/adjustierte Bestimmtheitsmaß](#)

$$R_*^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1},$$

mit  $k$  = Anzahl der erklärenden Variablen.

### 4.4.5. Intervallschätzer

- Mit  $\hat{\beta}_0$  und  $\hat{\beta}_1$  besitzen wir zwei Punktschätzer für die unbekannten Regressionsparameter.
- Wie kennen, unter den Standardbedingungen, sogar ihre Verteilung:

$$\hat{\beta}_0 \sim N \left( \beta_0, \sqrt{\sigma_{\text{res}}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n\hat{\sigma}_*^2(x)} \right)} \right),$$
$$\hat{\beta}_1 \sim N \left( \beta_1, \sqrt{\frac{\sigma_{\text{res}}^2}{n\hat{\sigma}_*^2(x)}} \right).$$

- Allerdings muss vorher noch die Varianz  $\sigma_{\text{res}}^2$  der Residuen geschätzt werden. Wir verwenden den erwartungstreuen Schätzer

$$\hat{\sigma}_{\text{res}}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- Damit lässt sich problemlos ein  $(1 - \alpha) \cdot 100\%$ -Konfidenzintervall für  $\beta_0$

$$I_0 = \left[ \hat{\beta}_1 - t_{n-2, 1-\alpha/2} \hat{\sigma}(\hat{\beta}_0), \hat{\beta}_1 + t_{n-2, 1-\alpha/2} \hat{\sigma}(\hat{\beta}_0) \right].$$

und für  $\beta_1$

$$I_1 = \left[ \hat{\beta}_0 - t_{n-2, 1-\alpha/2} \hat{\sigma}(\hat{\beta}_1), \hat{\beta}_0 + t_{n-2, 1-\alpha/2} \hat{\sigma}(\hat{\beta}_1) \right].$$

bestimmen.

- Dabei benutzen wir die Schätzer

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{\frac{\hat{\sigma}_{\text{res}}^2}{n\hat{\sigma}_*^2(x)}}, \quad \hat{\sigma}(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\text{res}}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n\hat{\sigma}_*^2(x)} \right)}.$$

#### 4.4.6. Tests zur Anpassungsgüte

- Wenn wir die Güte unserer Schätzungen beurteilen wollen, können wir entsprechende Hypothesentests verwenden.
- Als Hypothese bietet sich an, jeweils die Nullhypothesen

$$H_0 : \beta_0 = 0, \quad H_1 : \beta_0 \neq 0$$

und

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

zu testen.

- Werden die Hypothesen abgelehnt, so spricht das für unser lineares Modell. Anderenfalls muss ggf. über ein anderes Modell nachgedacht werden.

- Wir wissen bereits, dass unsere Schätzer unter den Standardannahmen normalverteilt sind, d.h. unter der Hypothese  $\beta_0 = 0$  bzw.  $\beta_1 = 0$  gilt

$$\hat{\beta}_0 \sim N \left( 0, \sqrt{\sigma_{\text{res}}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{n\hat{\sigma}_*^2(x)} \right)} \right),$$
$$\hat{\beta}_1 \sim N \left( 0, \sqrt{\frac{\sigma_{\text{res}}^2}{n\hat{\sigma}_*^2(x)}} \right).$$

- Dementsprechend können wir die ersten beiden Hypothesen mit dem uns bekannten t-Test testen (s. Abschnitt (2)).



- Zum testen der Hypothese  $\beta_i = 0$  ( $i \in 0, 1$ ) verwenden wird die Teststatistik

$$T = \frac{\hat{\beta}_i}{\widehat{\sigma}(\hat{\beta}_i)}$$

und lehnen ab, wenn

$$|T| > t_{n-2, 1-\alpha/2}$$

ist.

- Als p-Wert ergibt sich also

$$p = P(|T| > |t|) = 2(1 - F_{n-2}(|t|)),$$

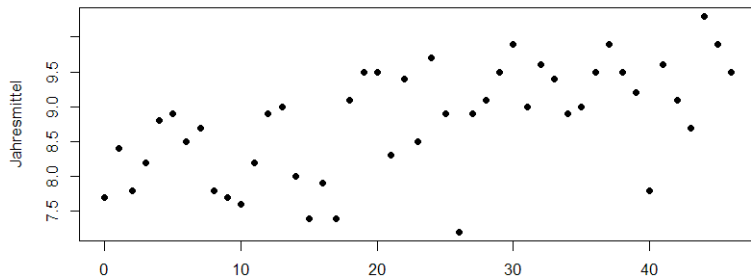
mit der Verteilungsfunktion  $F_{n-2}$  der t-Verteilung mit  $(n - 2)$  Freiheitsgraden.

### 4.4.7. Beispielregression mit R

■ **Beispiel B4.22**  $\Rightarrow$  B4.19: Wir betrachten wieder die Jahresmitteltemperaturen in Deutschland für den Zeitraum 1970-2016 (Quelle: DWD):

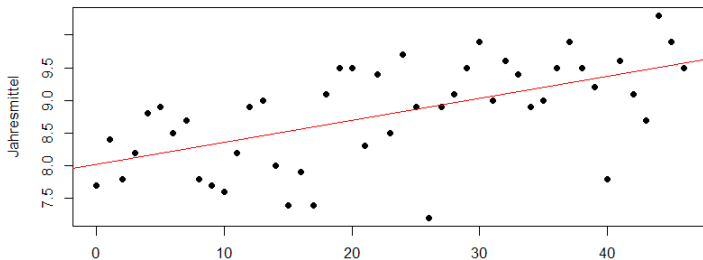
```
> tb=read.table("DWD.txt",sep=";",dec=".",header=T,fill=T)
> tb=tb[90:136,]
> x=tb$Jahr-1970
> t=tb$Deutschland
> x
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
+ 20 21 22 23 24
[26] 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
+ 45 46
> t
[1] 7.7 8.4 7.8 8.2 8.8 8.9 8.5 8.7 7.8 7.7 7.6 8.2
+ 8.9 9.0 8.0
[16] 7.4 7.9 7.4 9.1 9.5 9.5 8.3 9.4 8.5 9.7 8.9 7.2
+ 8.9 9.1 9.5
[31] 9.9 9.0 9.6 9.4 8.9 9.0 9.5 9.9 9.5 9.2 7.8 9.6
+ 9.1 8.7 10.3
[46] 9.9 9.5
```

```
> plot(x,t,col=col,pch=20,cex=1.4,ylab="Jahresmittel")
```



```
> cor(x,t)
[1] 0.5895415
```

```
> lin=lm(t~x)
> abline(lin,col="red")
```

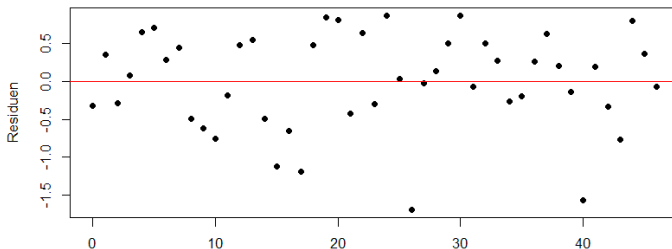


```
> lin

Call:
lm(formula = t ~ x)

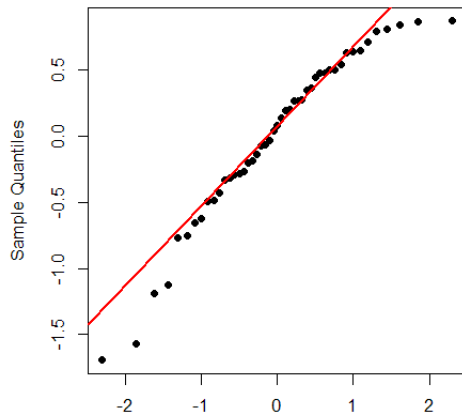
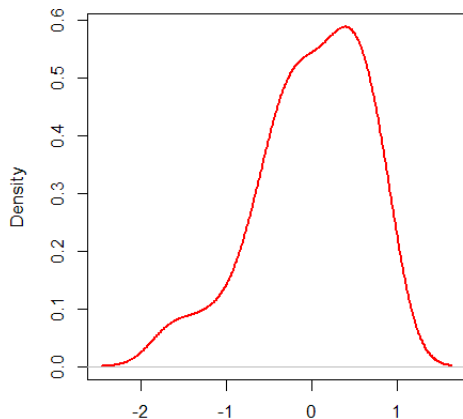
Coefficients:
(Intercept)          x
    8.01729      0.03375
```

```
> plot(x,lin$residuals,col=col,pch=20,cex=1.4,ylab="Residuen")  
> abline(h=0,col="red")
```



```
> mean(lin$residuals)  
[1] -2.406021e-17  
  
> sd(lin$residuals)  
[1] 0.6340938  
  
> summary(lin$residuals)  
   Min.   1st Qu.   Median     Mean   3rd Qu.    Max.     
-1.69500 -0.32610  0.08145  0.00000  0.49010  0.87260
```

```
> par(mar=c(2,4,1,1),mfrow=c(1,2))  
> plot(density(lin$residuals),main="",lwd=2,col="red")  
> qqnorm(lin$residuals,pch=16,main="")  
> qqline(lin$residuals,col="red",lwd=2)
```



```
> summary(lin)

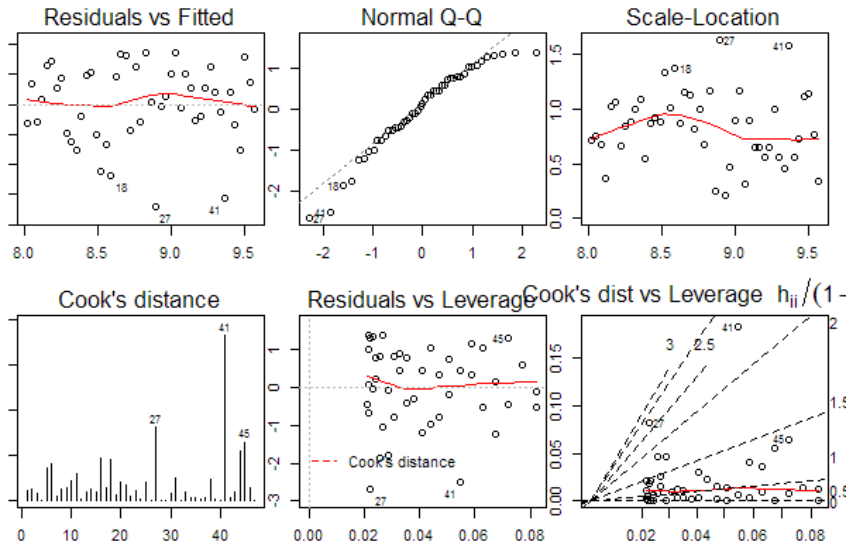
Call:
lm(formula = t ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-1.69488 -0.32611  0.08145  0.49014  0.87263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.017287   0.184083  43.553  < 2e-16 ***
x             0.033753   0.006894   4.896  1.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6411 on 45 degrees of freedom
Multiple R-squared:  0.3476,    Adjusted R-squared:  0.3331
F-statistic: 23.97 on 1 and 45 DF,  p-value: 1.299e-05
```

```
> par(mfrow=c(2,3),mar=c(3,3,3,3))
> for (i in 1:6) plot(lin,which=i)
```





## A.

# Übungsaufgaben

## A.1. Aufgaben

### Übung 1

**Aufgabe 1:** Es sei  $x = (6, 1, 3, 4, 1)$ . Berechnen Sie:

a)  $\frac{1}{5} \sum_{k=1}^5 x_k$

c)  $\sum_{i=1}^5 i \cdot x_{6-i}$

b)  $\sum_{l=1}^5 (x_l - 3)^2$

d)  $\prod_{j=1}^5 (-1)^{x_j}$

**Aufgabe 2:** Die Gaußklammer  $\lfloor x \rfloor$  ist als die größte ganze Zahl, die kleiner oder gleich  $x$  ist, definiert. Es sei  $n = 8$ . Geben Sie  $\lfloor \alpha n \rfloor$  für  $\alpha = 0.1, 0.4, 0.7$  an.

**Aufgabe 3:** Berechnen Sie  $\binom{6}{2}$ .

**Aufgabe 4:** Gelten die folgenden Rechenregeln?

a)  $(x \cdot y)^b = x^b \cdot y^b$

e)  $\log(x + y) = \log(x) + \log(y)$

b)  $(x + y)^b = x^b + y^b$

f)  $\log(x \cdot y) = \log(x) \cdot \log(y)$

c)  $e^{(x^2)} = (e^x)^2$

g)  $\log(x \cdot y) = \log(x) + \log(y)$

d)  $\sqrt{x^2} = |x|$

h)  $\sum_{k=1}^n a_k = \sum_{k=0}^{n-1} a_{k+1}$

**Aufgabe 5:** Vereinfachen Sie:

a)  $3^a \cdot 3^b \cdot 3^c$

b)  $a^3 \cdot b^3 \cdot c^3$

**Aufgabe 6:** Skizzieren Sie die folgenden Funktionen:

a)  $f(x) = 2x - 3$

e)  $f(x) = e^{-x^2}$

b)  $f(x) = \log(x)$

f)  $f(x) = e^{-(x-1)^2}$

c)  $f(x) = e^x$

d)  $f(x) = e^{-x}$

g)  $f(x) = e^{-\frac{(x-2)^2}{4}}$

## Übung 2

**Aufgabe 7:** Im Rahmen einer Wahlumfrage wird für 700 am Telefon Befragte das Alter und die bevorzugte Partei (A,B,C oder D) ermittelt. Geben Sie ein passendes  $\Omega$  an und beschreiben Sie die Merkmale mathematisch durch Angabe der Merkmalsausprägungen.

**Aufgabe 8:** Geben Sie für das Beispiel **B1.1** eine Tabelle an, die die relativen und absoluten Häufigkeiten, sowie die kumulativen relativen und kumulativen absoluten Häufigkeiten enthält.

**Aufgabe 9:** Warum gelten die Gleichungen 2.1–2.3?

**Aufgabe 10:** Geben Sie jeweils ein weiteres Beispiel für die besprochenen vier Merkmalskalen an.

**Aufgabe 11:** Auf der Straße werden 20 erwachsene Passanten im Rahmen einer Umfrage befragt. Eines der erfassten Merkmale ist die Kinderzahl  $K$ . Folgende Beobachtungen werden notiert:

1, 2, 0, 0, 2, 0, 0, 2, 1, 0, 3, 1, 0, 0, 0, 1, 1, 1, 0, 1

- a) Geben Sie die Menge der Merkmalsausprägungen für das Merkmal  $K$  an.
- b) Stellen Sie eine Tabelle auf, die die relativen und absoluten Häufigkeiten, sowie die kumulativen relativen und kumulativen absoluten Häufigkeiten enthält.

- c) Zeichnen Sie die empirische Verteilungsfunktion.

## Übung 3

### Aufgabe 12:

- a) Berechnen Sie das arithmetische Mittel der folgenden drei Datenreihen.
- (i) 4, 6, 9, 10, 13, 18    (ii) 0, 2, 2, 3, 3, 50    (iii) 1, 2, 3, 17, 18, 19
- b) Worin unterscheiden sich die Datensätze hinsichtlich der Lage der Datenwerte in Bezug auf ihren Mittelwert?

**Aufgabe 13:** Zeichnen Sie ein Histogramm für das Beispiel [B2.20](#).

**Aufgabe 14:** Für 200 Hotels in Sachsen werden die monatlichen Übernachtungszahlen in klassierter Form betrachtet:

Klasse:	0-100	100-500	500-2000	2000-5000
#Hotels:	20	90	40	50

- Zeichnen Sie ein Histogramm.
- Zeichnen Sie ein Diagramm, das die zugehörige empirische Dichte zeigt.
- Berechnen Sie das arithmetische Mittel für die klassiert vorliegenden Übernachtungszahlen.

**Aufgabe 15:** Wann wird das arithmetische Mittel bei Hinzunahme eines weiteren Datenpunktes größer? Argumentieren Sie unter Zuhilfenahme von Gleichung (2.5).

**Aufgabe 16:** Betrachten Sie die Daten aus Aufgabe 11.

- a) Zeichnen Sie ein Balkendiagramm und ein Kreisdiagramm.
- b) Berechnen Sie das arithmetische Mittel der Kinderzahl.
- c) Geben Sie die Ordnungsstatistik an.
- d) Berechnen Sie den Median.
- e) Berechnen Sie das  $\alpha$ -getrimmte Mittel für  $\alpha = 0, 1$ .
- f) Geben Sie das obere Quartil an.

**Aufgabe 17:** Zeigen Sie, dass die Formel  $\overline{ax + b} = a\bar{x} + b$  für beliebige Zahlen  $a, b \in \mathbb{R}$  gilt (Linearität des arithmetischen Mittels).

## Übung 4

**Aufgabe 18:** Auf einer Insel werden drei Jahre lang Erdbeben und ihre Stärke registriert. Dabei werden folgende Jahresmittelwerte und Varianzen beobachtet.

Jahr	# Beben	$\bar{x}$	$\text{Var}(x)$
2012	6	2	1
2013	3	4	4
2014	7	3	2

Berechnen Sie den gepoolten Mittelwert und die gepoolte Varianz der Erdbebenstärken.



**Aufgabe 19:** Betrachten Sie die Daten aus dem Beispiel B1.1.

- a) Berechnen Sie die Varianz und die Standardabweichung des beobachteten Merkmals Augenzahl.
- b) Wieviele Daten liegen im Intervall  $[\bar{x} - \hat{\sigma}_*(x), \bar{x} + \hat{\sigma}_*(x)]$ ?
- c) Berechnen Sie den Median, die Quartile und den IQR.

**Aufgabe 20:** Entwerfen Sie eine Stichprobe von  $n = 6$  Daten mit folgenden Anforderungen:

- a)  $\bar{x} = 0$ ,
- b)  $\bar{x} = 5$ ,  $\hat{\sigma}_*(x) = 1$ ,
- c)  $\tilde{x}_{.25} = -3$ ,  $\tilde{x}_{.75} = 4$
- d)  $\tilde{x} = 7$ ,  $R_x = 10$ .

**Aufgabe 21:** Gegeben seien die folgenden Schlusskurse des DAX an sieben aufeinander folgenden Tagen.

Tag	Schlusskurs
2016-10-26	10710
2016-10-25	10757
2016-10-24	10761
2016-10-21	10711
2016-10-20	10701

- a) Berechnen Sie die Stichprobenvarianz und die Stichprobenstandardabweichung der Schlusskurse.
- b) Geben Sie die Spannweite, den IQR, sowie den Variationskoeffizienten an.
- c) Berechnen Sie den MAD.

## Übung 5

**Aufgabe 22:** Sind alle Werte in einer Kontingenztafel eindeutig bestimmt, wenn nur die absoluten Randhäufigkeiten angegeben sind?

**Aufgabe 23:** Geben Sie fiktive absolute Häufigkeiten für eine  $3 \times 2$ -Kontingenztafel für zwei unabhängige Merkmale an.

**Aufgabe 24:** Ein neues Produkt kommt in drei Varianten I, II und III auf den Markt. Es ergeben sich an einem Tag an drei verschiedenen Standorten A, B und C in Deutschland folgende Verkaufszahlen:

	I	II	III
A	8	8	4
B	10	20	5
C	22	32	11

a) Geben Sie die relativen Häufigkeiten und die Randhäufigkeiten an.

- b) Sind die beiden Merkmale Version und Standort unabhängig?
- c) Berechnen Sie  $\chi^2$  und beide Varianten des Pearsonschen Kontingenzkoeffizienten.
- d) Interpretieren Sie das Ergebnis.

**Aufgabe 25:** In einem Land besitzen die fünf größten Städte 3 000 000, 1 000 000, 500 000, 250 000 und 250 000 Einwohner. Zeichnen Sie eine Lorenz-Kurve und geben Sie den Gini-Koeffizienten an.

**Aufgabe 26:** Warum ist der größtmögliche Wert des Gini-Maßes  $\frac{n-1}{n}$ ?

## Übung 6

**Aufgabe 27:** 14 Tage lang werden die Verkaufszahlen für ein Buch in einer Buchhandlung notiert: 7, 11, 12, 8, 10, 9, 9, 8, 0, 6, 13, 18, 5 und 11. Zeichnen Sie einen Boxplot für die Daten.

**Aufgabe 28:** Für 6 Straßen werden die Durchschnittsgeschwindigkeit

und die Anzahl der Unfälle in einem Jahr angegeben:

Geschw.:	50	60	100	70	50	40
Unfälle:	2	2	7	4	2	1

Geben Sie die für die beiden Merkmale die empirische Kovarianz und den Korrelationskoeffizienten an und interpretieren Sie das Resultat.

**Aufgabe 29:** An zwei Hochschulen setzt man unterschiedliche Benotungssysteme ein. Während die Hochschule A die Benotungsskala  $I \rightarrow II \rightarrow III \rightarrow IV$  verwendet, mit  $I$  als bester Note, ist an der Hochschule B die Skala  $a \rightarrow b \rightarrow c$ , mit  $a$  als bester Note, in Gebrauch.

Für 20 Studierende, die von A nach B wechselten, wird die letzte Note an der Hochschule A mit der ersten Note an der Hochschule B verglichen:

A	I	I	I	I	I	I	I	II	II	II
B	a	a	a	a	a	b	b	a	a	a
A	II	II	II	III	III	III	III	IV	IV	IV
B	a	b	b	a	b	b	c	b	b	c

Berechnen Sie den Rangkorrelationskoeffizienten und interpretieren Sie das Ergebnis.

## Übung 7

**Aufgabe 30:** Ein Würfel wird dreimal geworfen. Bestimmen Sie die Wahrscheinlichkeit,...

- a) ..., dass keine Sechs fällt,
- b) ..., dass die Augenzahlen gleich sind,
- c) ..., dass die Augensumme 8 ist,

- d) . . . , dass die Augensumme 8 ist, gegeben, dass keine Sechs fällt.
- e) . . . , dass genau zwei Sechsen fallen.

**Aufgabe 31:** In einem Raum befinden sich 12 Stühle. Fünf Personen kommen in den Raum, wählen sich zufällig einen Stuhl aus und setzen sich.

- a) Wie groß ist die Wahrscheinlichkeit, dass fünf vorher ausgewählte Stühle besetzt sind?
- b) Wie groß ist die Wahrscheinlichkeit, dass die vorher ausgewählten Stühle mit vorher genau benannten Personen besetzt sind?

**Aufgabe 32:** Eine Zufallsvariable  $X$  nimmt die Werte  $-2, -1, 0, 1$  und  $2$  mit den Wahrscheinlichkeiten  $0.2, 0.1, 0.4, 0.1, 0.2$  an. Zeichnen Sie die Wahrscheinlichkeitsfunktion und berechnen Sie  $P(X \leq 0.7)$ ,  $E(X)$ ,  $\text{Var}(X)$  und  $E(|X|)$ .

## Übung 8

**Aufgabe 33:** Die Zufallsvariable  $X$  beschreibe die Dauer zwischen zwei aufeinanderfolgenden Ankünften von Kunden in einer Bank (Einheit: Minuten).  $X$  besitze die Verteilungsfunktion

$$F(x) = \begin{cases} 0 & ; x < 0 \\ 1 - e^{-x/2} & ; x \geq 0 \end{cases}$$

- a) Zeichnen Sie die Verteilungsfunktion.
- b) Geben Sie die zugehörige Dichtefunktion an und zeichnen Sie sie.
- c) Wie groß ist die Wahrscheinlichkeit, dass zwischen zwei Kundenankünften weniger als fünf Minuten vergehen?
- d) Ein Kunde erreicht die Bank um 12 Uhr. Wie groß ist die Wahrscheinlichkeit, dass der nächste Kunde nach 12:01 Uhr, aber vor 12:03 ankommt?



- e) Berechnen Sie den Erwartungswert für die Zwischenankunftszeiten.
- f) Mit welcher Wahrscheinlichkeit ist eine Zwischenankunftszeit länger als der oben berechnete Erwartungswert?

**Aufgabe 34:** Angenommen zehn Prozent aller Autos seien weiß, 60 Prozent schwarz und 30 Prozent besäßen eine andere Lackierung.

- a) Auf einem Parkplatz stehen 30 Autos. Wie groß ist der Erwartungswert der Anzahl weißer Autos?
- b) Wie groß ist die Wahrscheinlichkeit, dass unter den Wagen auf dem Parkplatz weniger als drei weiße Autos sind?
- c) Wie groß ist die Wahrscheinlichkeit, dass an einer Kreuzung erst 15 nicht-weiße Autos vorbeifahren, bevor schließlich ein weißes Auto vorbeikommt?

- d) Wie lange muss man im Durchschnitt auf ein weißes Auto warten?
- e) Wie groß ist die Wahrscheinlichkeit unter zehn Autos zwei weiße, fünf schwarze und drei andersfarbige Wagen zu finden?

## Übung 9

**Aufgabe 35:** Angenommen  $X$  besitze eine Standardnormalverteilung. Berechnen Sie die folgenden Wahrscheinlichkeiten.

- a)  $P(X \leq 1)$ ,
- b)  $P(-1 \leq X \leq 1)$ ,
- c)  $P(X > 2)$ ,
- d)  $P(X > 2 \text{ oder } X < -2)$ .

Welche Verteilung besitzen die folgenden Zufallsvariablen?

- e)  $-X/10$ ,
- f)  $3 \cdot X + 2$ ,
- g)  $5 \cdot (X - 6)$ .

**Aufgabe 36:** Der jährliche Gewinn  $X$  einer Firma sei normalverteilt mit Erwartungswert 70 Mill. Euro und Standardabweichung 12 Mill. Euro. Berechnen Sie die Wahrscheinlichkeit, dass der Gewinn

- a) größer als 80 Millionen Euro ist,
- b) kleiner als 50 Millionen Euro ist,
- c) zwischen 50 und 80 Millionen liegt.

Eine zweite Firma macht  $Y \sim N(40, 5)$  Millionen Euro Gewinn.

- d) Wie groß ist die Wahrscheinlichkeit, dass die Summe der Gewinne beider Firmen die 100-Millionen-Euro-Marke überschreitet?

**Aufgabe 37:** Es gelte  $X \sim N(\mu, \sigma)$ . Wie groß sind folgende Wahrscheinlichkeiten?

- a)  $P(X > \mu + \sigma)$ ,
- b)  $P(X \leq \mu - \sigma)$ ,
- c)  $P(X \in [\mu - \sigma, \mu + \sigma])$ ,

Für welchen Wert  $x$  gilt

g)  $P(X > \mu + x\sigma) = 0.1,$

h)  $P(X \leq \mu - x\sigma) = 0.1,$

i)  $P(X \in [\mu - x\sigma, \mu + x\sigma]) = 0.9 ?$

## Übung 10

**Aufgabe 38:** Das Einkommen von Arbeitern in einem Land sei normalverteilt mit  $\mu = 3.5$  und  $\sigma = 0.8$  (tsd.Euro monatlich).

- a) Wie groß ist die Wahrscheinlichkeit, dass ein Arbeiter mehr 3500, aber weniger als 5000 Euro verdient?
- b) Ein Arbeiter sagt, 80% seiner Kollegen verdienen mehr als er. Wieviel zusätzliches Gehalt müsste er bekommen, damit nur noch 50% der Kollegen mehr verdienen?
- c) Wie groß ist der Erwartungswert und die Standardabweichung des arithmetischen Mittels von 100 zufällig ausgewählten Arbeitern?

**Aufgabe 39:** Ein Würfel werde 120 Mal gewürfelt.

- a) Geben Sie ein genähertes Intervall an, in dem die Augensumme mit 90% Wahrscheinlichkeit liegt.
- b) Wir betrachten das konkrete Beispiel B1.1. Geben Sie Schätzer für den Erwartungswert und die Varianz der Augenzahlen an.
- c) Schätzen Sie die Standardabweichung des Schätzers für den Erwartungswert.

**Aufgabe 40:** Wir betrachten das Beispiel B4.1. Stellen Sie einen geeigneten Schätzer auf und überlegen Sie, ob der Schätzer erwartungstreu und konsistent ist.

## Übung 11

**Aufgabe 41** : Eine Firma verkauft in 6 Monaten 18,17,19,10,14 und 15 Fahrzeuge. Bestimmen Sie

- a) das arithmetische Mittel,
- b) die Stichprobenvarianz,
- c) den Median,
- d) das 0.2-Quantil und
- e) den IQR.

**Aufgabe 42:** Geben Sie für die Daten in Aufgabe 41 ein 99%-Konfidenzintervall für den Erwartungswert und die Varianz an. Gehen Sie von normalverteilten Daten aus.

**Aufgabe 43:** Berechnen Sie für das Beispiel B1.1 ein genähertes 95%-Konfidenzintervall für den Erwartungswert und für die Varianz.

**Aufgabe 44:** Ein Spieler gewinnt einen Euro, wenn er bei einem Münzwurf die richtige Seite vorhersagt, ansonsten verliert er zwei Euro. Der Spieler startet mit einem Guthaben von 40 Euro.

- a) Geben Sie ein genähertes Intervall an, in dem das verbliebene Guthaben des Spielers nach 100 Spielen mit 99% Wahrscheinlichkeit liegt.
- b) Geben Sie eine genäherte Wahrscheinlichkeit dafür an, dann noch ein positives Guthaben aufzuweisen.

## Übung 12

**Aufgabe 45** : Für zwei Studiengänge A und B werden 2016 an einer Hochschule insgesamt 1000 Studenten eingeschrieben. Davon entfallen auf die verschiedenen Studiengänge und Geschlechter:

	A	B
m	250	100
w	450	200

- Sind die beiden Merkmale Studiengang ( $=X$ ) und Geschlecht ( $=Y$ ) unabhängig?
- Berechnen Sie den Pearsonschen Kontingenzkoeffizienten.
- Interpretieren Sie das Ergebnis.



**Aufgabe 46:** Die Anzahl der Studierenden in der Vorlesung „Statistik“ sei normalverteilt. In 10 Jahren ergeben sich folgende Studierendenzahlen: 88, 75, 72, 87, 99, 80, 70, 59, 69, 84.

- a) Geben Sie Schätzer für den Erwartungswert und die Standardabweichung an.
- b) Die wahre Standardabweichung sei von nun an  $\sigma = 10$ . Geben Sie ein 90%-Konfidenzintervall für den Erwartungswert an.
- c) Jemand stellt die Hypothese auf, dass  $\mu = 80$  ist. Diese Hypothese wird zugunsten der Alternative  $\mu < 80$  abgelehnt, wenn  $\hat{\mu} < D$  ist. Bestimmen Sie die Konstante  $D$  so, dass der Fehler erster Art kleiner als 5% wird.
- d) Der wahre Erwartungswert sei in der Tat  $\mu = 80$ . Wie groß ist die Wahrscheinlichkeit, dass ein Raum mit 90 Sitzplätzen zu klein für die Vorlesung ist?

## Übung 13

**Aufgabe 47** : Die Körpergröße der Bevölkerung sei in Deutschland normalverteilt mit Erwartungswert  $\mu = 170$  cm und Standardabweichung  $\sigma = 10$  cm.

- a) Wie groß ist die Wahrscheinlichkeit dafür, dass eine zufällig ausgewählten Person über 190 cm groß ist?
- b) Wie groß ist die Wahrscheinlichkeit unter 50 zufällig ausgewählten Probanden weniger als zwei mit einer Körpergröße über 190 cm zu finden?
- c) Für einen Film wird ein Statist mit einer Größe zwischen 190cm und 195cm gesucht. Wie viele zufällig ausgewählte Kandidaten muss man im Durchschnitt einladen, bis ein passender Kandidat gefunden ist?

**Aufgabe 48:** Gegeben seien folgende Daten aus einer normalverteilten Grundgesamtheit: 12, 6, 8, 15, 14, 10, 25, 11, 10, 9.

- a) Testen Sie zum Niveau 10% die Hypothese  $H_0 : \mu = 11$  gegen die Alternative  $H_1 : \mu > 11$ .
- b) Testen Sie zum Niveau 10% die Hypothese  $H_0 : \sigma^2 = 30$  gegen die Alternative  $\sigma^2 \neq 30$ .

Für eine zweite normalverteilte Stichprobe vom Umfang  $m = 10$  ergibt sich ein arithmetisches Mittel  $\bar{y} = 10$  und eine Stichprobenvarianz von  $\hat{\sigma}^2(y) = 25$ .

- c) Testen Sie zum Niveau 10% die Hypothese gleicher Erwartungswerte.
- d) Testen Sie zum Niveau 10% die Hypothese gleicher Varianzen.

**Aufgabe 49** : Der Preis  $W$  eines Produktes sei normalverteilt mit Erwartungswert  $\mu = 120$  Euro und Varianz  $\sigma^2 = 100$  Euro.

Bestimmen Sie die Wahrscheinlichkeiten für folgende Ereignisse:

a)  $W > 120$

d)  $W < 130$

b)  $W < 120$

e)  $110 < W < 130$

c)  $W > 130$

f)  $W > 140$  oder  $W < 100$

## Übung 14

**Aufgabe 50** : Der Preis  $W$  eines Produktes sei normalverteilt mit Erwartungswert  $\mu = 120$  Euro und Varianz  $\sigma^2 = 100$  Euro. Geben Sie im folgenden jeweils eine passende Zahl  $z$  an.

a)  $P(W > z) = 0.1,$

c)  $P(W < z) = 0.99,$

b)  $P(W < z) = 0.05,$

d)  $P(|W - 120| > z) = 0.2.$

**Aufgabe 51:** Angeblich wählen 30% aller Wähler eines Landes die Partei A, 20% Partei B, 20% Partei C und 15% die Partei D (die übrigen Wähler sind Nichtwähler). Eine Umfrage mit 80 Befragten ergibt folgende Häufigkeiten:

A	B	C	D	N
20	14	11	19	16

Testen Sie mit einem Signifikanztest zum Niveau 10%, ob die obige Aussage plausibel ist.

**Aufgabe 52:** An vier Standorten A,B,C und D einer Lebensmittelkette werden drei verschiedene Varianten (I,II,III) eines Nahrungsmittels verkauft. An einem Wochenende ergeben sich folgende Verkaufszahlen.

	A	B	C	D
I	10	34	40	25
II	25	29	37	39
III	27	25	26	40

- a) Testen Sie zum Niveau 5% die Unabhängigkeit der beiden Merkmale Standort und Variante.
- b) Testen Sie zum Niveau 10% die Hypothese, die drei Nahrungsmittelvarianten würden im Verhältnis 3:5:4 verkauft.

**Aufgabe 53:** Für 10 Studierende wird die Statistiknote  $X$  und die Mathematiknote  $Y$  verglichen (Noten: 0 bis 15).

$x_i$	3	7	12	11	15	14	11	13	5	7
$y_i$	5	9	11	7	11	12	11	13	6	8

Zeichnen Sie ein Streudiagramm.

## Übung 15

**Aufgabe 54:** Bei einem Hypothesentest zum Signifikanzniveau 10% der Nullhypothese  $H_0 : \mu = 0$  gegen die Alternative  $H_1 : \mu < 0$  wird für die Teststatistik  $T = -8$  berechnet. Was genau bedeutet der p-Wert in Höhe von 0.07?

**Aufgabe 55** : (s. Aufgabe 53) Für 10 Studierende wird die Statistikknote  $X$  und die Mathematiknote  $Y$  verglichen (Noten: 0 bis 15).

$x_i$	3	7	12	11	15	14	11	13	5	7
$y_i$	5	9	11	7	11	12	11	13	6	8

- a) Berechnen Sie für die Daten in Aufgabe 55 bei Annahme eines linearen Modells die Schätzer für die Regressionskoeffizienten an.
- b) Geben Sie auch die Residuen an.

**Aufgabe 56** : Angenommen die Grundgesamtheit in Aufgabe 55 bestehe eine Normalverteilung.

- a) Geben Sie für die  $x$ -Daten ein 99%-Konfidenzintervall für den Erwartungswert an.
- b) Testen Sie für die  $x$ -Daten zum Signifikanzniveau 10% die Hypothese  $H_0 : \mu = 12$  gegen  $H_1 : \mu > 12$ .



## A.2. Musterlösungen

### Lösung 41:

a)  $\bar{x} = \frac{18+17+\dots+15}{6} = 15.5$

b) Zwei mögliche Rechenwege:

$$\hat{\sigma}^2(x) = \frac{6}{5} \left( \frac{18^2 + 17^2 + \dots + 15^2}{6} - 15.5^2 \right) = 10.7$$

$$\hat{\sigma}^2(x) = \frac{(18 - 15.5)^2 + \dots + (15 - 15.5)^2}{5} = 10.7$$

c) Ordnungsstatistik: 10, 14, 15, 17, 18, 19

$$\tilde{x} = \frac{x_{(3)} + x_{(4)}}{2} = 16.$$

$$d) x_{(0.2)} = x_{(\lfloor 1.2 \rfloor + 1)} = x_{(2)} = 14$$

$$e) x_{(0.25)} = x_{(\lfloor 1.5 \rfloor + 1)} = x_{(2)} = 14$$

$$x_{(0.75)} = x_{(\lfloor 4.5 \rfloor + 1)} = x_{(5)} = 18$$

### Lösung 45:

Es ergeben sich folgende Randhäufigkeiten:

	A	B	
m	250	100	350
w	450	200	650
	700	300	1000

a) Nein, denn es ist z.B.

$$h_{11} = 0.25 \neq 0.35 \cdot 0.7 = 0.245 = h_{1\bullet} \cdot h_{\bullet 1}$$

b) Wir berechnen zunächst den Chi-Quadrat-Koeffizienten:

$$\chi^2 = n \cdot \left( \left( \sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} \right) - 1 \right) = 0.5232862$$

Damit ergibt sich

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = 0.02286947$$

und dann der Pearsonsche Kontingenzkoeffizient:

$$C^* = \sqrt{\frac{\min\{k, l\}}{\min\{k, l\} - 1}} \cdot C = \sqrt{\frac{2}{1}} \cdot 0.02286947 = 0.03234231.$$

c) Da  $C^*$  sehr nahe bei 0 liegt, können wir von einer weitgehenden Unabhängigkeit der beiden Merkmale ausgehen.

**Lösung 47:**

a) Es sei  $X$  die Körpergröße einer zufällig ausgewählten Person. Dann ist

$$\begin{aligned} P(X > 190) &= P\left(\frac{X - 170}{10} > \frac{190 - 170}{10}\right) \\ &= P(X^* > 2) = 1 - \Phi(2) = 0.0228 \end{aligned}$$

b) Es sei  $N$  die Anzahl von Kandidaten mit einer Körpergröße über 190cm.  $N$  besitzt eine Binomialverteilung mit  $p = 0.0228$  und  $n = 50$ . Es gilt also

$$\begin{aligned} P(N < 2) &= P(N = 0) + P(N = 1) \\ &= \binom{50}{0} p^0 (1 - p)^{50} + \binom{50}{1} p^1 (1 - p)^{49} \\ &= 0.6847559. \end{aligned}$$

c) Es sei  $M$  die Anzahl der Kandidaten, die eingeladen werden müssen. Dann ist

$M$  geometrisch verteilt mit Erfolgswahrscheinlichkeit

$$\begin{aligned} q &= P(190 < X < 195) \\ &= P\left(\frac{190 - 170}{10} < X^* < \frac{196 - 170}{10}\right) \\ &= \Phi(2) - \Phi(-2) = 0.01654047. \end{aligned}$$

Es gilt (s. Seite 287)  $E(M) = 1/q = 60.45779$ .

### Lösung 49:

- a)  $P(W > 120) = 0.5$  (Symmetrie der Normalverteilung)
- b)  $P(W < 120) = 0.5$  (Symmetrie der Normalverteilung)
- c)  $P(W > 130) = P\left(\frac{W-120}{10} > \frac{130-120}{10}\right) = P(W^* > 1) = 1 - \Phi(1) = 0.1586553$
- d)  $P(W < 130) = 1 - P(W > 130) = 1 - 0.158655 = 0.8413447$
- e)  $P(110 < W < 130) = P(-1 < W^* < 1) = \Phi(1) - \Phi(-1) = 0.6826895$
- f)  $P(W > 140 \text{ oder } W < 100) = P(W > 140) + P(W < 100) = P(W^* > 2) + P(W^* < -2) = 2(1 - \Phi(2)) = 0.04550026$

## Lösung 50:

a)

$$\begin{aligned} P(W > z) &= 0.1 \\ \Leftrightarrow P\left(W^* > \frac{z - 120}{10}\right) &= 0.1 \Leftrightarrow \frac{z - 120}{10} = z_{0.9} \\ \Leftrightarrow z &= 120 + z_{0.9} \cdot 10 = 132.8155. \end{aligned}$$

b)

$$\begin{aligned} P(W < z) &= 0.05 \Leftrightarrow P\left(W^* < \frac{z - 120}{10}\right) = 0.05 \\ \Leftrightarrow \frac{z - 120}{10} &= z_{0.05} \Leftrightarrow z = 120 + z_{0.05} \cdot 10 = 103.5515. \end{aligned}$$

c)

$$\begin{aligned} P(W < z) &= 0.99 \\ \Leftrightarrow P\left(W^* < \frac{z - 120}{10}\right) &= 0.99 \\ \Leftrightarrow \frac{z - 120}{10} = z_{0.99} &\Leftrightarrow z = 120 + z_{0.99} \cdot 10 = 143.2635. \end{aligned}$$

d)

$$P(|W - 120| > z) = 0.2$$

$$\Leftrightarrow P(W < 120 - z) + P(W > 120 + z) = 0.2$$

$$\Leftrightarrow P\left(W^* < \frac{-z}{10}\right) + P\left(W^* > \frac{z}{10}\right) = 0.2$$

$$\Leftrightarrow 2P\left(W^* > \frac{z}{10}\right) = 0.2 \Leftrightarrow P\left(W^* > \frac{z}{10}\right) = 0.1$$

$$\Leftrightarrow \frac{z}{10} = z_{0.9} = 1.281552$$

$$\Leftrightarrow z = 12.81552.$$

**B.****Anhang****B.1. Kleine Formelsammlung****B.1.1. Notationen (Deskriptive Statistik)**

$\bar{x}$	Arithmetisches Mittel
$\hat{\sigma}_*^2(x), \hat{\sigma}_*^2$	Empirische Varianz (früher $\text{Var}(x)$ )
$\hat{\sigma}^2(x), \hat{\sigma}^2$	Stichprobenvarianz (früher $\widehat{\text{Var}}(\bar{x})$ )
$\hat{\sigma}_*(x), \hat{\sigma}_*$	Empirische Standardabweichung (früher $\sigma(x)$ )
$\hat{\sigma}(x), \hat{\sigma}$	Stichprobenstandardabweichung
$s_{xy}$	Empirische Kovarianz
$r_{xy}$	Empirischer Korrelationskoeffizient



## B.1.2. Wahrscheinlichkeitstheorie

$P(\overline{A}) = 1 - P(A)$	
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	
$P(A \cup B) = P(A) + P(B)$	falls $A, B$ unvereinbar
$P(A \cap B) = P(A) \cdot P(B)$	falls $A, B$ unabhängig
$P(A B) = P(A \cap B) / P(B)$	
$P(A B) = P(A)$	falls $A, B$ unabhängig

$E(aX + b) = aE(X) + b$	$a, b \in \mathbb{R}$
$E(X + Y) = E(X) + E(Y)$	$X, Y$ nicht notw. unabhängig
$E(X \cdot Y) = E(X) \cdot E(Y)$	falls $X, Y$ unkorreliert
$E(\sum_{i=1}^n X_i) = n\mu$	falls $E(X_1) = E(X_1) = \dots = \mu$
$E(\overline{X}) = \mu$	falls $E(X_1) = E(X_1) = \dots = \mu$

$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$	
$\hat{\sigma}_*(X) = \sqrt{\text{Var}(X)}$	
$\text{Var}(aX + b) = a^2 \text{Var}(X)$	$a, b \in \mathbb{R}$
$\hat{\sigma}_*(aX + b) =  a  \hat{\sigma}_*(X)$	$a, b \in \mathbb{R}$
$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$	
$\hat{\sigma}_*(X + Y) = \sqrt{\hat{\sigma}_*(X)^2 + \hat{\sigma}_*(Y)^2 + 2\text{Cov}(X, Y)}$	
$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$	falls $X, Y$ unkorreliert
$\hat{\sigma}_*(X + Y) = \sqrt{\hat{\sigma}_*(X)^2 + \hat{\sigma}_*(Y)^2}$	falls $X, Y$ unabhängig

### B.1.3. Schätzer und Konfidenzintervalle

Schätzer für..		Eigenschaften
$\mu$	$\hat{\mu} = \bar{X}$	erwartungstreu, konsistent
$\sigma$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	erwartungstreu, konsistent

KI für..	KI	Voraussetzung
$\mu$	$\left[ \hat{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$	$X_i$ normalverteilt, $\sigma$ bekannt
$\mu$	$\left[ \hat{\mu} - t_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right]$	$X_i$ normalverteilt, $\sigma$ unbekannt
$\sigma^2$	$\left[ \frac{n\hat{\sigma}^2}{\chi_{n-1,1-\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi_{n,\alpha/2}} \right]$	$X_i$ normalverteilt, $\mu$ bekannt
$\sigma^2$	$\left[ \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1,1-\alpha/2}}, \frac{(n-1)\hat{\sigma}^2}{\chi_{n-1,\alpha/2}} \right]$	$X_i$ normalverteilt, $\mu$ unbekannt

## B.2. Tabellen

### B.2.1. Quantile $z_\alpha$ der Normalverteilung

$\alpha$	0.8	0.9	0.95	0.975	0.99	0.995	0.999
$z_\alpha$	0.842	1.282	1.645	1.960	2.326	2.576	3.090
$\alpha$	0.2	0.1	0.05	0.025	0.01	0.005	0.001
$z_\alpha$	-0.842	-1.282	-1.645	-1.960	-2.326	-2.576	-3.090

Es gilt  $\Phi(x) = \alpha$  genau dann, wenn  $x = z_\alpha v$  ist.

Beispiel: Eine Zufallsvariable  $X$  besitze eine Standardnormalverteilung.  
Dann gilt

$$P(X > z) = 5\% \quad \Leftrightarrow \quad \Phi(z) = 0.95 \quad \Leftrightarrow \quad z = z_{0.95} = 1.645.$$

### B.2.2. Verteilungsfunktion $\Phi(x)$ der Normalverteilung

Angegeben sind die Werte für die Verteilungsfunktion, z.B.

$$\Phi(1.46) = 0.928$$

Beispiel:  $X$  besitzt eine Normalverteilung mit Erwartungswert  $\mu = 10$  und Standardabweichung  $\sigma = 8$ . Mit welcher Wahrscheinlichkeit ist  $X$  positiv?

Es gilt

$$\begin{aligned} P(X > 0) &= P\left(\frac{X - 10}{8} > \frac{0 - 10}{8}\right) \\ &= P(X^* > -5/4) = 1 - \Phi(-1.25) \\ &= \Phi(1.25) = 0.894. \end{aligned}$$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536	0.540
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575	0.579
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614	0.618
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652	0.655
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688	0.691
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722	0.726
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755	0.758
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785	0.788
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813	0.816
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839	0.841
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862	0.864
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883	0.885
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901	0.903
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918	0.919
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932	0.933
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944	0.945
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954	0.955
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963	0.964
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977	0.977
2	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.997
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999	0.999
3	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

**B.2.3. Quantile  $t_{n,\alpha}$  der t-Verteilung**

Eine Zufallsvariable  $T$  hat eine t-Verteilung mit 10 Freiheitsgraden. Wir wollen  $z$  so bestimmen, dass

$$P(T < z) = 0.9$$

ist.

Es gilt

$$P(T < z) = 0.9 \quad \Leftrightarrow \quad z = t_{10,0.9} = 1.372.$$

	0.001	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.999
2	-22.327	-6.965	-4.303	-2.920	-1.886	1.886	2.920	4.303	6.965	22.327
3	-10.215	-4.541	-3.182	-2.353	-1.638	1.638	2.353	3.182	4.541	10.215
4	-7.173	-3.747	-2.776	-2.132	-1.533	1.533	2.132	2.776	3.747	7.173
5	-5.893	-3.365	-2.571	-2.015	-1.476	1.476	2.015	2.571	3.365	5.893
6	-5.208	-3.143	-2.447	-1.943	-1.440	1.440	1.943	2.447	3.143	5.208
7	-4.785	-2.998	-2.365	-1.895	-1.415	1.415	1.895	2.365	2.998	4.785
8	-4.501	-2.896	-2.306	-1.860	-1.397	1.397	1.860	2.306	2.896	4.501
9	-4.297	-2.821	-2.262	-1.833	-1.383	1.383	1.833	2.262	2.821	4.297
10	-4.144	-2.764	-2.228	-1.812	-1.372	1.372	1.812	2.228	2.764	4.144
19	-3.579	-2.539	-2.093	-1.729	-1.328	1.328	1.729	2.093	2.539	3.579
20	-3.552	-2.528	-2.086	-1.725	-1.325	1.325	1.725	2.086	2.528	3.552
29	-3.396	-2.462	-2.045	-1.699	-1.311	1.311	1.699	2.045	2.462	3.396
30	-3.385	-2.457	-2.042	-1.697	-1.310	1.310	1.697	2.042	2.457	3.385
39	-3.313	-2.426	-2.023	-1.685	-1.304	1.304	1.685	2.023	2.426	3.313
40	-3.307	-2.423	-2.021	-1.684	-1.303	1.303	1.684	2.021	2.423	3.307
49	-3.265	-2.405	-2.010	-1.677	-1.299	1.299	1.677	2.010	2.405	3.265
50	-3.261	-2.403	-2.009	-1.676	-1.299	1.299	1.676	2.009	2.403	3.261
59	-3.234	-2.391	-2.001	-1.671	-1.296	1.296	1.671	2.001	2.391	3.234
60	-3.232	-2.390	-2.000	-1.671	-1.296	1.296	1.671	2.000	2.390	3.232
69	-3.213	-2.382	-1.995	-1.667	-1.294	1.294	1.667	1.995	2.382	3.213
70	-3.211	-2.381	-1.994	-1.667	-1.294	1.294	1.667	1.994	2.381	3.211
79	-3.197	-2.374	-1.990	-1.664	-1.292	1.292	1.664	1.990	2.374	3.197
80	-3.195	-2.374	-1.990	-1.664	-1.292	1.292	1.664	1.990	2.374	3.195
89	-3.184	-2.369	-1.987	-1.662	-1.291	1.291	1.662	1.987	2.369	3.184
90	-3.183	-2.368	-1.987	-1.662	-1.291	1.291	1.662	1.987	2.368	3.183
99	-3.175	-2.365	-1.984	-1.660	-1.290	1.290	1.660	1.984	2.365	3.175
100	-3.174	-2.364	-1.984	-1.660	-1.290	1.290	1.660	1.984	2.364	3.174
109	-3.167	-2.361	-1.982	-1.659	-1.289	1.289	1.659	1.982	2.361	3.167
110	-3.166	-2.361	-1.982	-1.659	-1.289	1.289	1.659	1.982	2.361	3.166
119	-3.160	-2.358	-1.980	-1.658	-1.289	1.289	1.658	1.980	2.358	3.160
120	-3.160	-2.358	-1.980	-1.658	-1.289	1.289	1.658	1.980	2.358	3.160



### B.2.4. Quantile $\chi_{n,\alpha}$ der Chi-Quadrat-Verteilung

Im Rahmen eines Chi-Quadrat-Tests ergibt sich für die Teststatistik

$$T = 19.$$

$T$  hat eine Chi-Quadrat-Verteilung mit 59 Freiheitsgraden, d.h. als kritischen Wert für den Test erhalten wir bei einem Niveau von 5%

$$\chi_{59,0.05} = 42.339.$$

Da  $T < \chi_{59,0.05}$  lehnen wir die Hypothese nicht ab.

---

	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
38	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162
39	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
48	28.177	30.755	33.098	35.949	60.907	65.171	69.023	73.683
49	28.941	31.555	33.930	36.818	62.038	66.339	70.222	74.919
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
58	35.913	38.844	41.492	44.696	72.160	76.778	80.936	85.950
59	36.698	39.662	42.339	45.577	73.279	77.931	82.117	87.166
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
68	43.838	47.092	50.020	53.548	83.308	88.250	92.689	98.028
69	44.639	47.924	50.879	54.438	84.418	89.391	93.856	99.228
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
98	68.396	72.501	76.164	80.541	116.315	122.108	127.282	133.476
99	69.230	73.361	77.046	81.449	117.407	123.225	128.422	134.642
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

---

### B.2.5. Quantile $F_{(n,m),\alpha}$ der F-Verteilung

$T$  besitze eine F-Verteilung mit 5 und 9 Freiheitsgraden. Es soll  $P(T > z) = 0.05$  gelten. Wie groß muss man  $z$  wählen?

$$P(T > z) = 0.05 \quad \Leftrightarrow \quad z = F_{(5,9),0.95} = 3.482.$$

 In der Tabelle stehen in der ersten Spalte die Werte von  $n$ , in der ersten Zeile die Werte für  $m$ .

$\alpha = 0.005$ 

	2	3	4	9	15	19	29	39	49	69	99
2	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
3	0.020	0.021	0.022	0.023	0.023	0.023	0.024	0.024	0.024	0.024	0.024
4	0.038	0.041	0.043	0.047	0.049	0.049	0.050	0.051	0.051	0.051	0.051
5	0.055	0.060	0.064	0.073	0.076	0.077	0.079	0.080	0.080	0.081	0.081
7	0.081	0.092	0.099	0.117	0.126	0.128	0.132	0.135	0.136	0.137	0.139
9	0.099	0.115	0.126	0.153	0.166	0.171	0.177	0.181	0.183	0.186	0.188
14	0.126	0.150	0.167	0.212	0.235	0.245	0.258	0.265	0.270	0.276	0.280
19	0.141	0.169	0.190	0.247	0.279	0.291	0.310	0.321	0.328	0.336	0.343
24	0.150	0.181	0.205	0.271	0.308	0.323	0.347	0.361	0.369	0.380	0.389
29	0.156	0.190	0.215	0.287	0.329	0.347	0.374	0.390	0.400	0.413	0.424
34	0.161	0.196	0.222	0.299	0.345	0.364	0.395	0.413	0.425	0.440	0.452
39	0.164	0.200	0.228	0.309	0.357	0.378	0.411	0.431	0.444	0.461	0.474
44	0.167	0.204	0.232	0.316	0.367	0.389	0.424	0.446	0.460	0.478	0.493
49	0.169	0.207	0.236	0.322	0.375	0.398	0.435	0.458	0.473	0.493	0.509
54	0.171	0.209	0.239	0.327	0.382	0.406	0.445	0.468	0.485	0.505	0.523
59	0.172	0.211	0.241	0.332	0.388	0.413	0.453	0.477	0.494	0.516	0.535
64	0.174	0.213	0.243	0.335	0.393	0.418	0.460	0.485	0.503	0.526	0.545
69	0.175	0.214	0.245	0.338	0.397	0.423	0.466	0.492	0.510	0.534	0.554
74	0.176	0.216	0.247	0.341	0.401	0.427	0.471	0.498	0.517	0.541	0.563
79	0.176	0.217	0.248	0.344	0.404	0.431	0.476	0.503	0.523	0.548	0.570
84	0.177	0.218	0.249	0.346	0.407	0.434	0.480	0.508	0.528	0.554	0.577
89	0.178	0.219	0.250	0.348	0.410	0.437	0.483	0.512	0.533	0.559	0.583
94	0.178	0.219	0.251	0.349	0.412	0.440	0.487	0.516	0.537	0.564	0.588
99	0.179	0.220	0.252	0.351	0.414	0.443	0.490	0.520	0.541	0.569	0.593
104	0.179	0.221	0.253	0.352	0.416	0.445	0.493	0.523	0.545	0.573	0.598
109	0.180	0.221	0.254	0.354	0.418	0.447	0.495	0.526	0.548	0.577	0.602
114	0.180	0.222	0.254	0.355	0.420	0.449	0.498	0.529	0.551	0.580	0.606
119	0.180	0.222	0.255	0.356	0.421	0.450	0.500	0.531	0.554	0.583	0.610
124	0.181	0.223	0.256	0.357	0.423	0.452	0.502	0.534	0.556	0.587	0.613

$\alpha = 0.05$ 

	2	3	4	9	15	19	29	39	49	69	99
2	0.053	0.052	0.052	0.052	0.051	0.051	0.051	0.051	0.051	0.051	0.051
3	0.105	0.108	0.110	0.113	0.115	0.115	0.116	0.116	0.117	0.117	0.117
4	0.144	0.152	0.157	0.167	0.171	0.172	0.174	0.175	0.175	0.176	0.177
5	0.173	0.185	0.193	0.210	0.217	0.219	0.222	0.224	0.225	0.226	0.227
7	0.211	0.230	0.243	0.272	0.285	0.289	0.296	0.299	0.301	0.304	0.305
9	0.235	0.259	0.275	0.315	0.333	0.339	0.349	0.353	0.357	0.360	0.363
14	0.267	0.299	0.321	0.378	0.406	0.417	0.432	0.441	0.446	0.452	0.457
19	0.284	0.320	0.345	0.413	0.448	0.461	0.481	0.493	0.500	0.508	0.515
24	0.294	0.332	0.360	0.435	0.474	0.490	0.514	0.528	0.536	0.547	0.555
29	0.301	0.341	0.370	0.450	0.493	0.511	0.537	0.553	0.563	0.575	0.585
34	0.305	0.347	0.377	0.461	0.507	0.526	0.555	0.572	0.583	0.596	0.607
39	0.309	0.351	0.383	0.469	0.518	0.538	0.569	0.587	0.599	0.614	0.626
44	0.312	0.355	0.387	0.476	0.526	0.547	0.580	0.599	0.612	0.627	0.641
49	0.314	0.358	0.390	0.481	0.533	0.555	0.589	0.609	0.622	0.639	0.653
54	0.316	0.360	0.393	0.486	0.539	0.561	0.596	0.617	0.631	0.649	0.664
59	0.317	0.362	0.396	0.490	0.544	0.566	0.603	0.624	0.639	0.657	0.673
64	0.318	0.364	0.398	0.493	0.548	0.571	0.608	0.630	0.645	0.665	0.681
69	0.320	0.365	0.399	0.495	0.551	0.575	0.613	0.636	0.651	0.671	0.688
74	0.320	0.367	0.401	0.498	0.554	0.578	0.617	0.640	0.656	0.677	0.694
79	0.321	0.368	0.402	0.500	0.557	0.581	0.621	0.645	0.661	0.682	0.700
84	0.322	0.369	0.403	0.502	0.560	0.584	0.624	0.648	0.665	0.686	0.705
89	0.323	0.369	0.404	0.503	0.562	0.587	0.627	0.652	0.669	0.691	0.709
94	0.323	0.370	0.405	0.505	0.564	0.589	0.630	0.655	0.672	0.694	0.714
99	0.324	0.371	0.406	0.506	0.565	0.591	0.632	0.657	0.675	0.698	0.717
104	0.324	0.371	0.407	0.507	0.567	0.593	0.634	0.660	0.678	0.701	0.721
109	0.325	0.372	0.407	0.508	0.568	0.594	0.636	0.662	0.680	0.704	0.724
114	0.325	0.373	0.408	0.509	0.570	0.596	0.638	0.664	0.682	0.706	0.727
119	0.325	0.373	0.409	0.510	0.571	0.597	0.640	0.666	0.685	0.709	0.730
124	0.326	0.373	0.409	0.511	0.572	0.598	0.641	0.668	0.687	0.711	0.732

$\alpha = 0.95$ 

	2	3	4	9	15	19	29	39	49	69	99
2	19.000	9.552	6.944	4.256	3.682	3.522	3.328	3.238	3.187	3.130	3.088
3	19.164	9.277	6.591	3.863	3.287	3.127	2.934	2.845	2.794	2.737	2.696
4	19.247	9.117	6.388	3.633	3.056	2.895	2.701	2.612	2.561	2.505	2.464
5	19.296	9.013	6.256	3.482	2.901	2.740	2.545	2.456	2.404	2.348	2.306
7	19.353	8.887	6.094	3.293	2.707	2.544	2.346	2.255	2.203	2.145	2.103
9	19.385	8.812	5.999	3.179	2.588	2.423	2.223	2.131	2.077	2.019	1.976
14	19.424	8.715	5.873	3.025	2.424	2.256	2.050	1.954	1.899	1.838	1.793
19	19.443	8.667	5.811	2.948	2.340	2.168	1.958	1.860	1.803	1.739	1.693
24	19.454	8.639	5.774	2.900	2.288	2.114	1.901	1.800	1.742	1.676	1.628
29	19.461	8.620	5.750	2.869	2.253	2.077	1.861	1.759	1.699	1.632	1.582
34	19.466	8.606	5.732	2.846	2.227	2.050	1.832	1.728	1.667	1.599	1.548
39	19.470	8.596	5.719	2.829	2.208	2.030	1.809	1.704	1.643	1.573	1.521
44	19.473	8.588	5.709	2.815	2.192	2.014	1.792	1.686	1.623	1.552	1.499
49	19.475	8.582	5.701	2.805	2.180	2.001	1.777	1.670	1.607	1.536	1.482
54	19.477	8.577	5.694	2.796	2.170	1.990	1.766	1.658	1.594	1.521	1.467
59	19.479	8.573	5.689	2.789	2.162	1.981	1.756	1.647	1.583	1.509	1.454
64	19.480	8.569	5.684	2.782	2.154	1.974	1.747	1.638	1.573	1.499	1.443
69	19.481	8.566	5.680	2.777	2.148	1.967	1.740	1.630	1.565	1.490	1.433
74	19.482	8.563	5.677	2.772	2.143	1.961	1.733	1.623	1.557	1.482	1.425
79	19.483	8.561	5.674	2.768	2.138	1.956	1.727	1.617	1.551	1.475	1.417
84	19.484	8.559	5.671	2.765	2.134	1.952	1.722	1.611	1.545	1.469	1.411
89	19.484	8.557	5.668	2.761	2.130	1.948	1.718	1.607	1.540	1.463	1.405
94	19.485	8.556	5.666	2.759	2.127	1.944	1.714	1.602	1.535	1.458	1.399
99	19.486	8.554	5.664	2.756	2.124	1.941	1.710	1.598	1.531	1.453	1.394
104	19.486	8.553	5.663	2.754	2.121	1.938	1.707	1.595	1.527	1.449	1.389
109	19.487	8.552	5.661	2.752	2.119	1.935	1.704	1.591	1.524	1.445	1.385
114	19.487	8.551	5.660	2.750	2.117	1.933	1.701	1.588	1.520	1.442	1.381
119	19.487	8.550	5.658	2.748	2.114	1.931	1.699	1.585	1.517	1.439	1.378
124	19.488	8.549	5.657	2.746	2.113	1.929	1.696	1.583	1.515	1.436	1.375

$$\alpha = 0.995$$

	2	3	4	9	15	19	29	39	49	69	99
2	199.000	49.799	26.284	10.107	7.701	7.093	6.396	6.088	5.915	5.727	5.592
3	199.166	47.467	24.259	8.717	6.476	5.916	5.276	4.995	4.838	4.667	4.545
4	199.250	46.195	23.155	7.956	5.803	5.268	4.659	4.392	4.243	4.081	3.966
5	199.300	45.392	22.456	7.471	5.372	4.853	4.262	4.004	3.860	3.703	3.592
7	199.357	44.434	21.622	6.885	4.847	4.345	3.775	3.526	3.387	3.237	3.130
9	199.388	43.882	21.139	6.541	4.536	4.043	3.483	3.239	3.102	2.955	2.850
14	199.428	43.172	20.515	6.089	4.122	3.638	3.088	2.848	2.713	2.568	2.464
19	199.447	42.826	20.210	5.864	3.913	3.432	2.885	2.645	2.510	2.364	2.260
24	199.458	42.622	20.030	5.729	3.786	3.306	2.759	2.519	2.384	2.236	2.131
29	199.465	42.487	19.911	5.639	3.701	3.221	2.674	2.433	2.296	2.148	2.041
34	199.470	42.392	19.826	5.575	3.639	3.160	2.612	2.369	2.233	2.083	1.975
39	199.474	42.320	19.763	5.527	3.593	3.114	2.565	2.321	2.183	2.033	1.923
44	199.477	42.265	19.713	5.489	3.557	3.077	2.527	2.283	2.145	1.993	1.882
49	199.479	42.221	19.674	5.459	3.528	3.048	2.497	2.252	2.113	1.960	1.849
54	199.481	42.185	19.642	5.435	3.504	3.024	2.473	2.227	2.087	1.933	1.820
59	199.483	42.155	19.616	5.414	3.484	3.004	2.452	2.205	2.065	1.910	1.797
64	199.484	42.130	19.593	5.397	3.467	2.987	2.434	2.187	2.046	1.890	1.776
69	199.485	42.108	19.574	5.382	3.452	2.972	2.419	2.171	2.029	1.873	1.758
74	199.486	42.089	19.557	5.369	3.440	2.959	2.405	2.157	2.015	1.858	1.742
79	199.487	42.073	19.542	5.358	3.429	2.948	2.394	2.145	2.002	1.845	1.728
84	199.488	42.058	19.530	5.348	3.419	2.938	2.383	2.134	1.991	1.833	1.716
89	199.488	42.045	19.518	5.339	3.410	2.929	2.374	2.124	1.981	1.822	1.705
94	199.489	42.034	19.508	5.331	3.402	2.922	2.366	2.116	1.972	1.812	1.695
99	199.489	42.024	19.499	5.324	3.395	2.914	2.358	2.108	1.964	1.804	1.685
104	199.490	42.014	19.490	5.317	3.389	2.908	2.351	2.100	1.956	1.796	1.677
109	199.490	42.006	19.483	5.311	3.383	2.902	2.345	2.094	1.950	1.789	1.669
114	199.491	41.998	19.476	5.306	3.378	2.897	2.339	2.088	1.943	1.782	1.662
119	199.491	41.991	19.470	5.301	3.373	2.892	2.334	2.082	1.938	1.776	1.656
124	199.492	41.984	19.464	5.297	3.369	2.887	2.329	2.077	1.932	1.770	1.650

**C.****Hinweise zur Klausur****C.1. Hilfsmittel**

Als Hilfsmittelsind zugelassen:

- a) Taschenrechner
- b) Eine gedruckte Formelsammlung
- c) Das komplette Vorlesungsskript, auch mit handschriftlichen Notizen, allerdings nicht mit Notizen zu den Lösungen, die wir in der Übung erarbeitet haben.

Generell nicht zugelassen sind Aufzeichnungen aus den Übungen.



## **C.2. Welche Abschnitte und Gegenstände werden nicht abgefragt?**

- a) Kombinatorik (3.2)
- b) Schätzen „ohne Zurücklegen“ (4.2.5)
- c) Beispielregression mit R (4.4.7)
- d) Der Satz von Moivre-Laplace (S. 224)