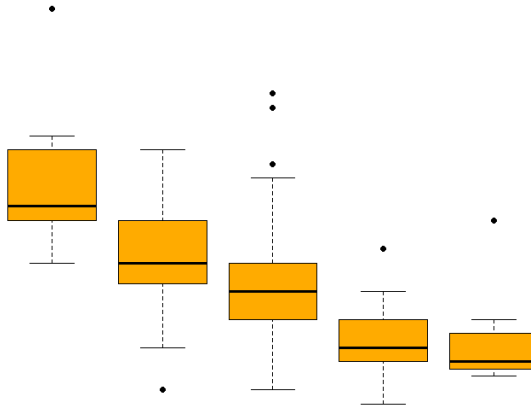


STATISTIK

Wintersemester 2016/2017

Vorlesungsfolien

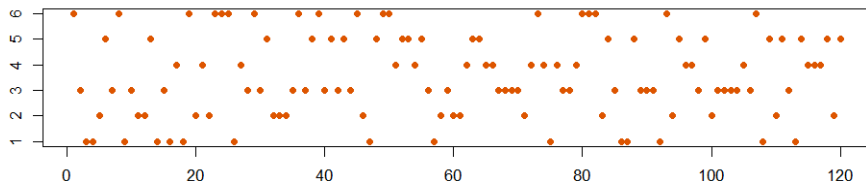


1.

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) Ω 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 Ω .
■ **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 Ω 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 :

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

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

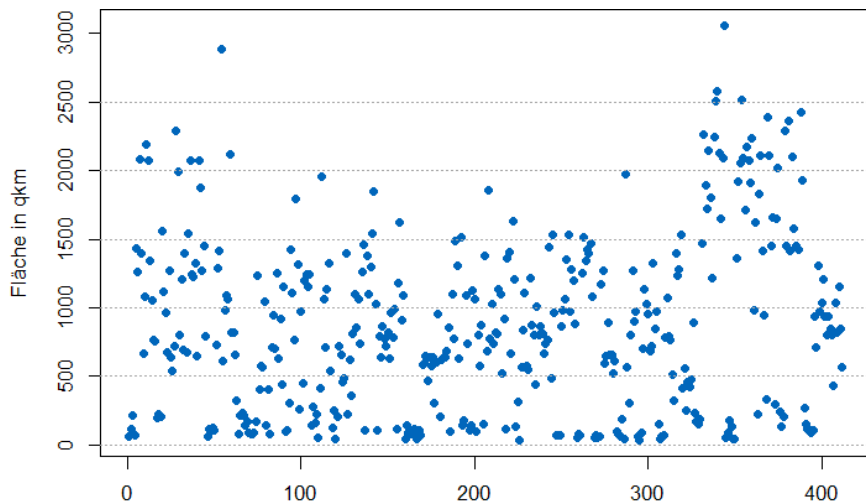
Man definiert Klassenhäufigkeiten als absolute/relative Häufigkeiten, summiert über alle Elemente der Klasse.

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 km^2 , Quelle: Stat. Bundesamt).



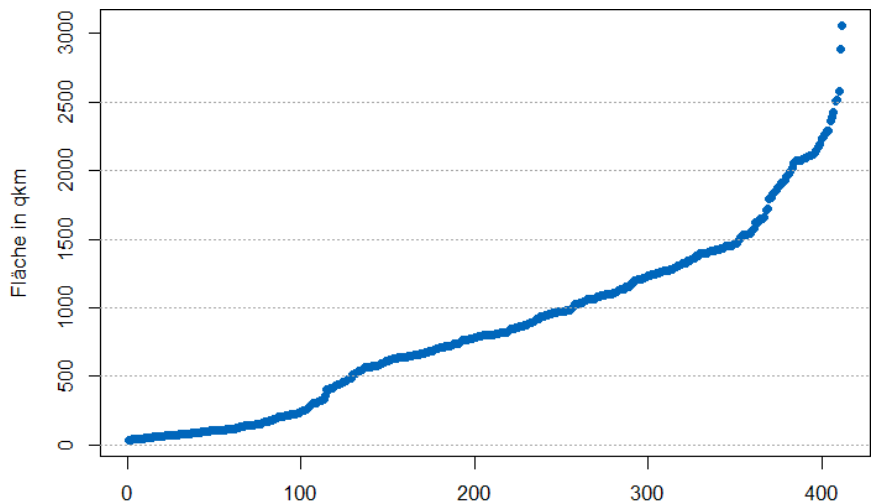
Wir teilen die Merkmalsausprägungen in Klassen ein:

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

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 $x_i \leq x$:

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

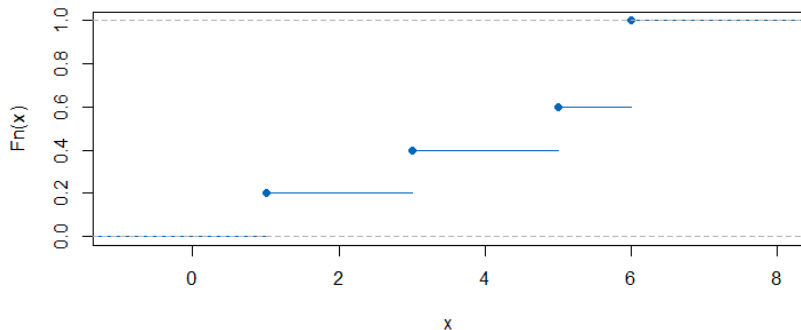
Es gilt:

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 ∞ strebt,
3. $F_n(x)$ ist dort konstant, wo keine Beobachtungswerte vorliegen.

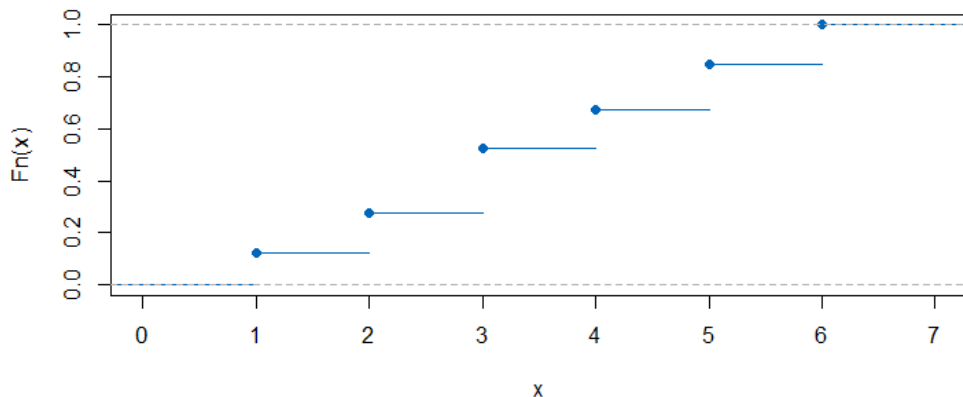
■ **Beispiel B2.21** \Rightarrow B2.18: Ein Würfel wird $n = 5$ Mal geworfen, die entsprechenden Beobachtungen sind:

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

Es ergibt sich folgende empirische Verteilungsfunktion:

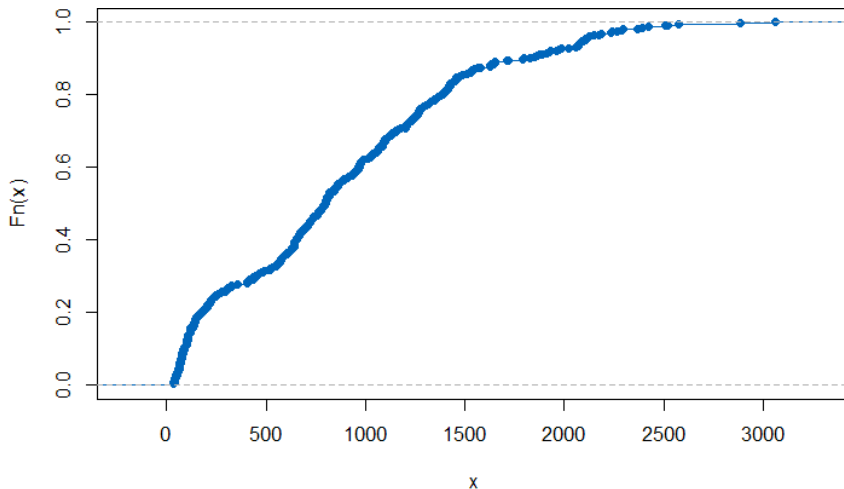


■ **Beispiel B2.22** $\Rightarrow_{B1.1}$: Im Eingangsbeispiel wurde ein Testwürfel 120 Mal geworfen. Es ergibt sich:



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:

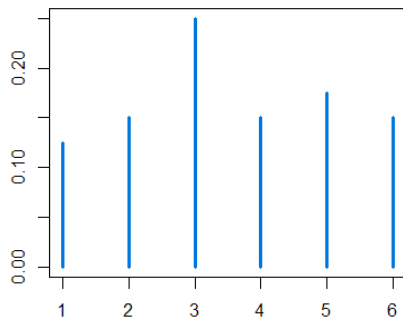
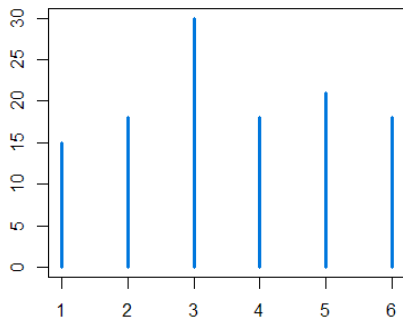


2.3. Diagramme und Grafiken

2.3.1. Stab- und Säulendiagramme

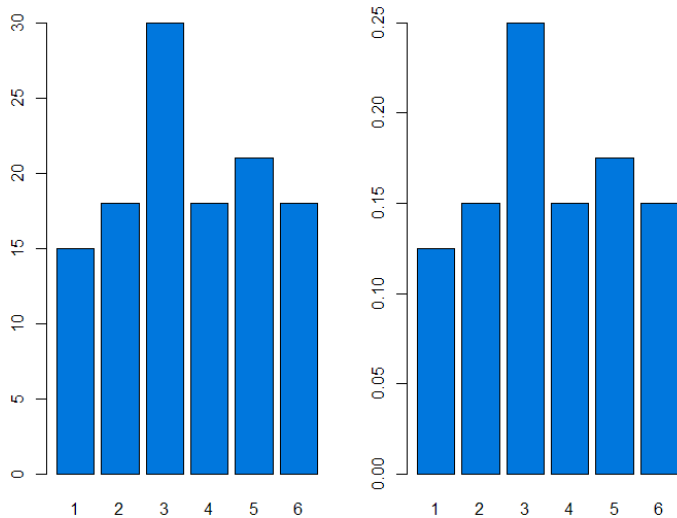
In Stabdiagrammen werden die relativen/absoluten Häufigkeiten als vertikale Linien dargestellt.

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



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

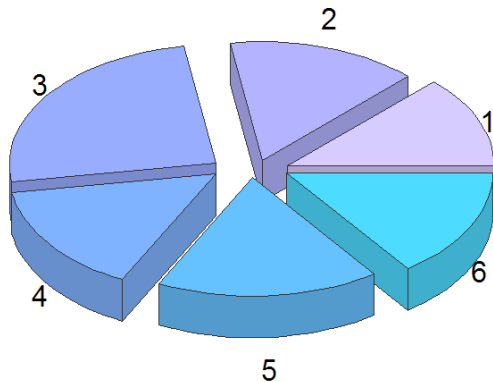
B2.25 \Rightarrow B1.1 :



2.3.2. Kreis- und Tortendiagramme

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

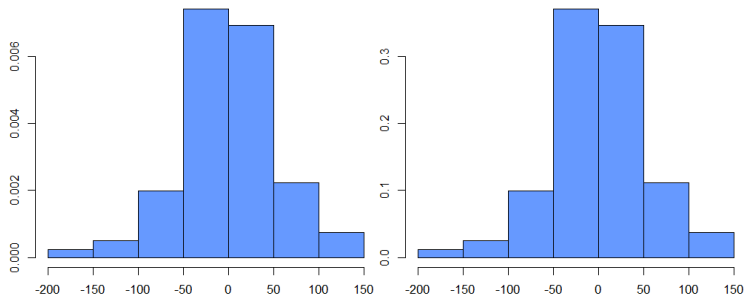
■ **Beispiel B2.26** \Rightarrow B1.1:



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.

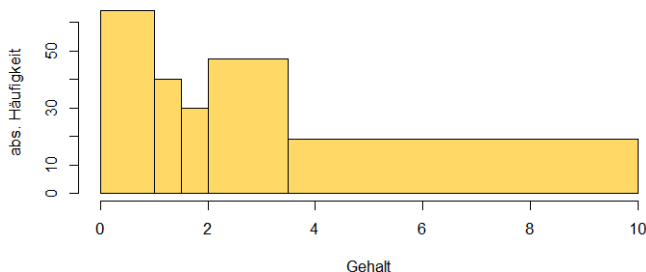
■ **Beispiel B2.27:** Tagesgewinne/-verluste des DAX vom 1. Januar bis 27. April 2011, in Punkten (Quelle: yahoo.com)



⚠ 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- ∞	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

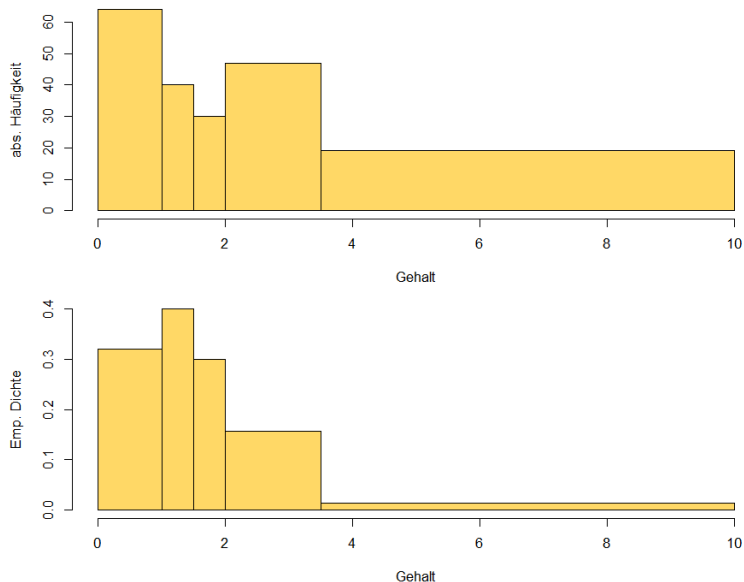
$$f_n(x) = \frac{h(K_i)}{b_i - a_i}, \quad x \in K_i. \quad (2.4)$$

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!

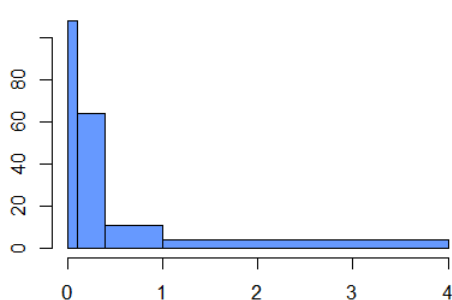
■ **Beispiel B2.29** \Rightarrow B2.28: Vergleich des klassischen Histogramms mit dem Diagramm für die empirische Dichte:



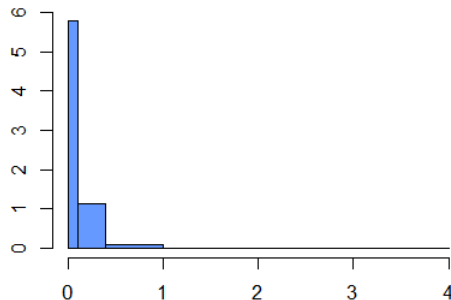
■ **Beispiel B2.30:** Einwohnerzahl 187 deutscher Städte am 31.12.2015 (Quelle: <http://www.citypopulation.de>, Angaben in Mill. Einwohnern). Wir definieren folgende Klassen der Form $(a, b]$ (in Mill. Einw.):

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:

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

Speziell gelten die Identitäten

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

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

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

■ **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:

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

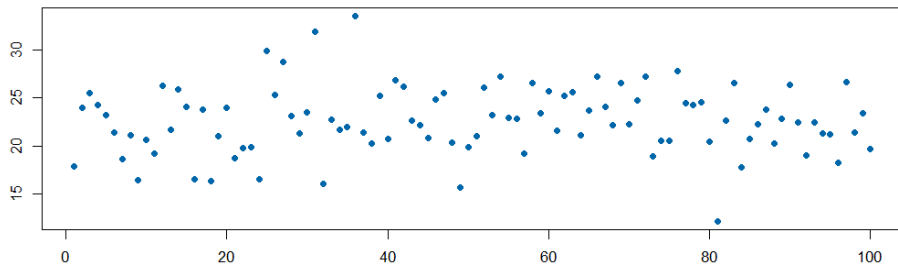
Vorteile:

⊕ Robust gegen Ausreißer.

Nachteile:

- ⊖ Einige Datenpunkte werden nicht verwendet.
- ⊖ Wahl von α beliebig. Missbrauch möglich.

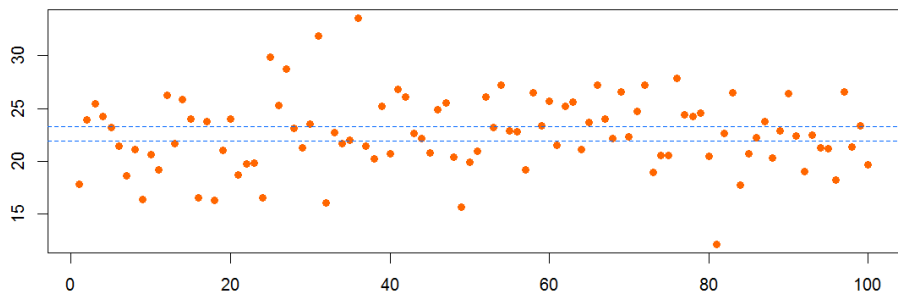
■ **Beispiel B2.32:** Dreißig Jahre lang wurde an einem Ort die Tageshöchsttemperatur am 1. September gemessen:



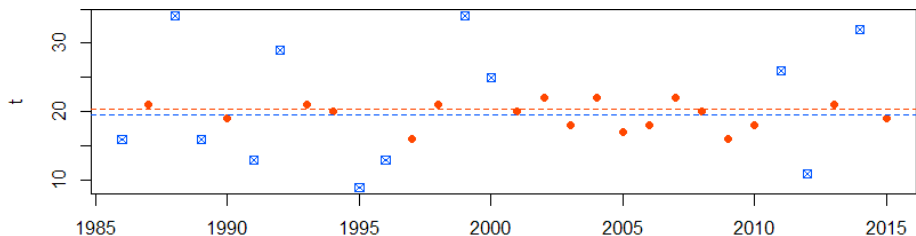
Es ergibt sich ein arithmetisches Mittel von

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

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 α -Quantil ist die kleinste Zahl \tilde{x}_α für die mindestens αn 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

$$\text{Var}(x) = \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n}$$

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

- $\text{Var}(x)$ 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:

$$\text{Var}(x) = \frac{1}{n} \sum_{m \in M_x} (m - \bar{x})^2 \cdot n(m).$$

- Meistens ist folgende alternative Formel leichter zu berechnen:

$$\text{Var}(x) = \overline{(x^2)} - (\bar{x})^2.$$

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

$$\text{Var}(ax + b) = a^2 \text{Var}(x).$$

Speziell ist die Varianz translationsinvariant.

Die Standardabweichung ist definiert als

$$\sigma(x) = \sqrt{\text{Var}(x)}.$$

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

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.

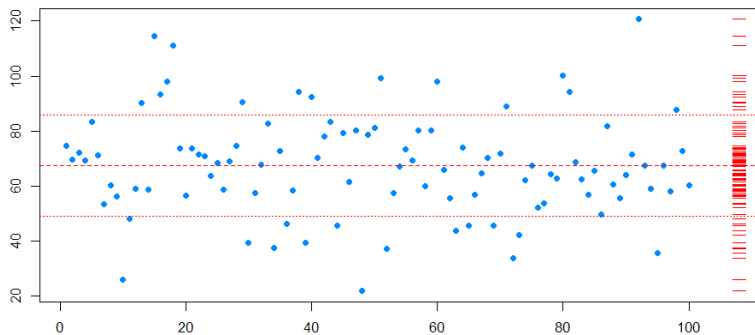
In der Statistik benötigt man neben der oben beschriebenen empirischen Varianz noch die Stichprobenvarianz (korrigierte Varianz) und die Stichprobenstandardabweichung (korrigierte Standardabweichung):

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

- Es gilt offenbar

$$\widehat{\text{Var}}(x) = \frac{n}{n-1} \text{Var}(x).$$

- 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.

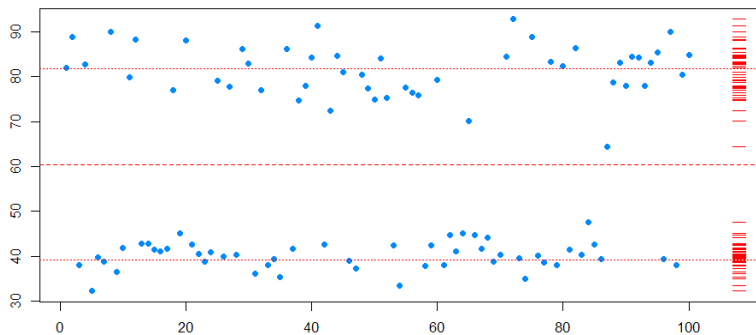
■ Beispiel B2.40:

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

$$\widehat{\text{Var}}(x) = 472.267,$$

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

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

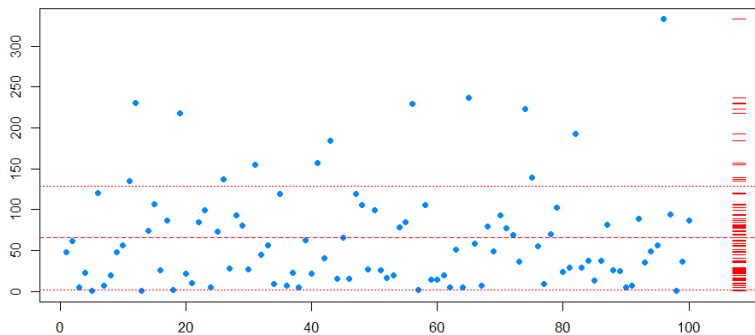
■ Beispiel B2.41:

$$\bar{x} = 60.44387$$

$$\widehat{\text{Var}}(x) = 452.3576,$$

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

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

■ Beispiel B2.42:

$$\bar{x} = 65.37265$$

$$\widehat{\text{Var}}(x) = 4082.81,$$

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

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

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 $\text{Var}(x_1), \text{Var}(x_2), \dots, \text{Var}(x_m)$ ergibt sich

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

(Varianzzerlegung).

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

	x_{ki}	n_k	\bar{x}_k	$\text{Var}(x_k)$
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: $\sum_{k=1}^m \frac{\text{Var}(x_k) \cdot n_k}{n} = 2.0$

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

Varianz: $\text{Var}(x) = 2 + 0.63 = 2.63\bar{8}$.

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:

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

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

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

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

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

Alternative Formel:

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

Für die Standardabweichung ergibt sich

$$\sigma(x) = \sqrt{\text{Var}(x)} \approx 1.92$$

Die Stichprobenvarianz ist entsprechend etwas größer als die empirische Varianz:

$$\widehat{\text{Var}}(x) = \frac{n}{n-1} \cdot \text{Var}(x) = \frac{22}{5} = 4.4$$

Dementsprechend ist

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

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:

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

MAD:

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

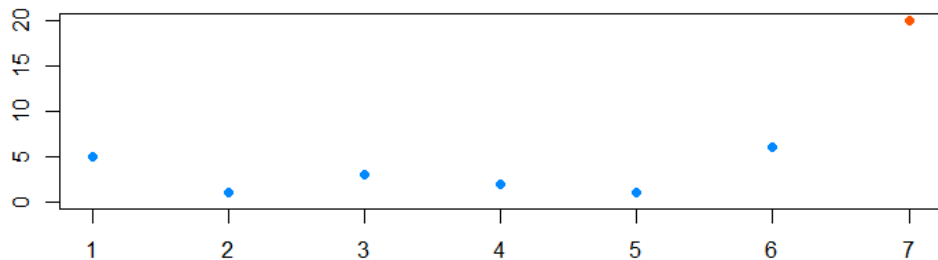
Mittlere absolute Abweichung vom Mittel:

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

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.



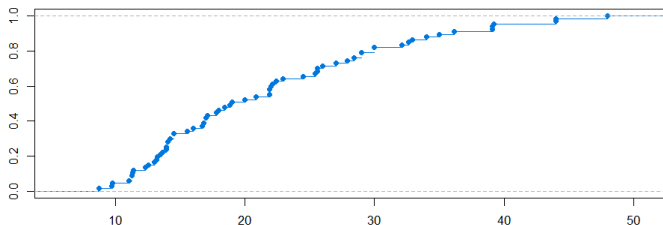
Nun ergibt sich:

	Alt	Neu
$\text{Var}(x)$	3.67	38.53
$\sigma(x)$	1.91	6.21
R_x	5	19
IQR_x	4	5
MAD_x	1.5	2

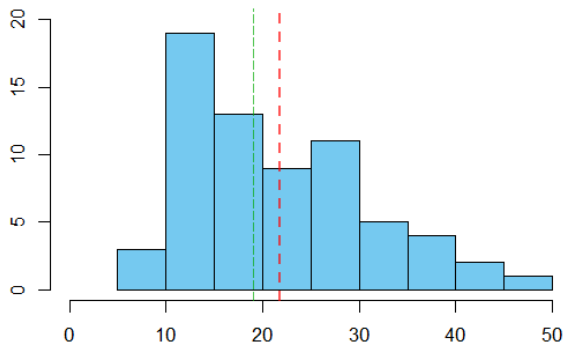
■ **Beispiel B2.45:** IT-Unternehmen in Österreich mit mehr als 99 Mitarbeitern (Quelle:<http://data.opendataportal.at>)

	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.



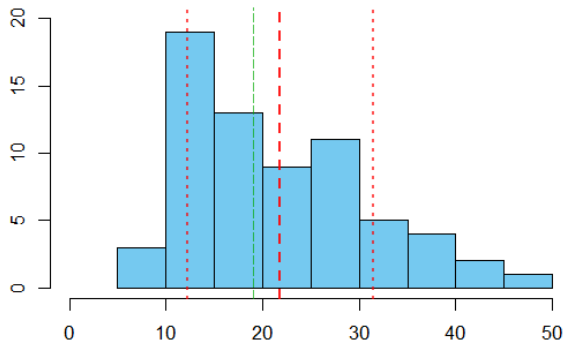
Histogramm:



Arithmetisches Mittel und Median:

$$\bar{U} = 21.8394$$

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



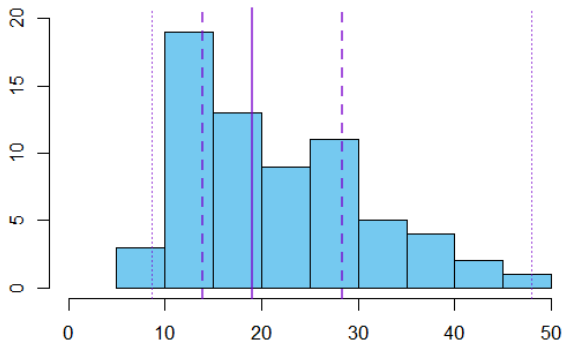
Varianz, Standardabweichung:

$$\text{Var}(U) = 90.90$$

$$\sigma(U) = 9.53$$

$$\widehat{\text{Var}}(U) = 92.28$$

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



Quartile:

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

Spannweite und Interquartilsabstand:

$$R_U = 48 - 8.7 = 39.3$$

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

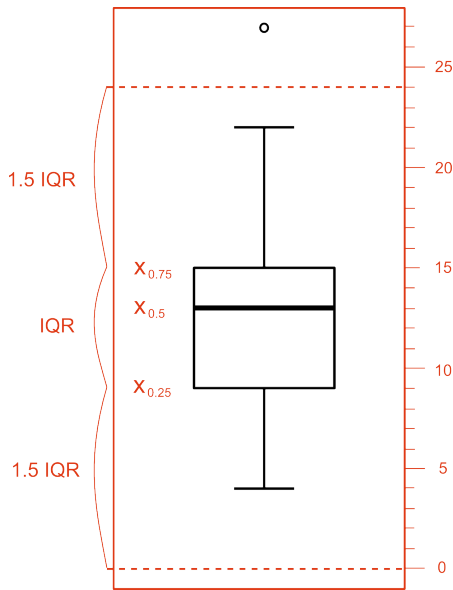
2.6. Boxplots

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

Vorgehensweise:

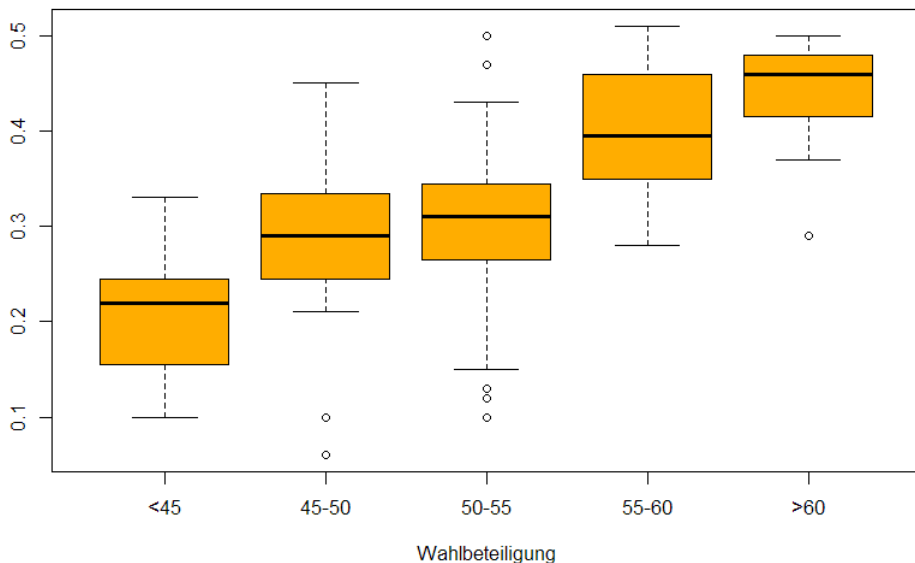
- Eine horizontale Linie wird auf der Höhe des Median eingezeichnet.
- Das obere und untere Quartil bestimmen die obere und untere Seite der „Box“.
- Die Länge der beiden Antennen (Whiskers) entspricht maximal dem 1.5-fachen des IQR (gerechnet vom oberen- bzw. unteren Quartil aus). Die Antennen enden 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)

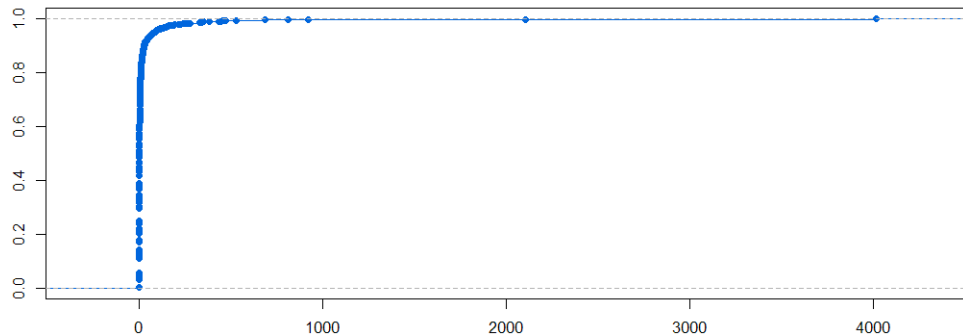
Stimmanteile für die CDU in den Wahllokalen



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.

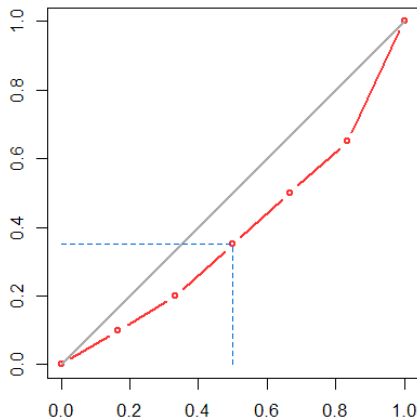
- Berechne zunächst für $i = 1, 2, \dots, n$ die Werte

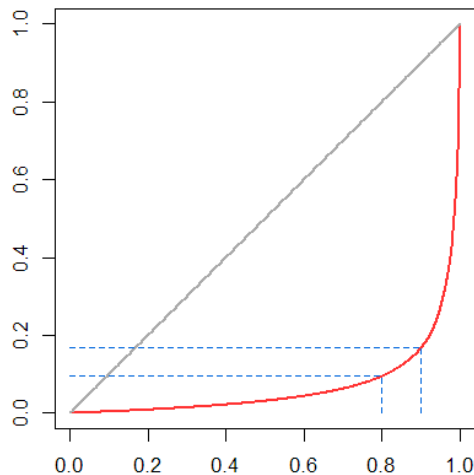
$$\begin{aligned} L_i &= \frac{\text{Summe der kleinsten } i \text{ Umsätze}}{\text{Gesamtsumme der Umsätze}} \\ &= \frac{\sum_{k=1}^i X_{(k)}}{\sum_{k=1}^n X_{(k)}}. \end{aligned}$$

- 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:

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

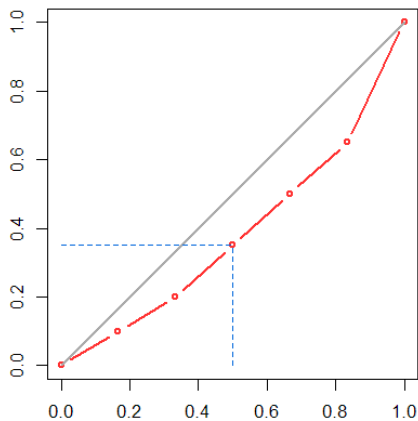
Das Gini-Maß entspricht der doppelten Fläche zwischen der Lorenz-Kurve und der Winkelhalbierenden.

- 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

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

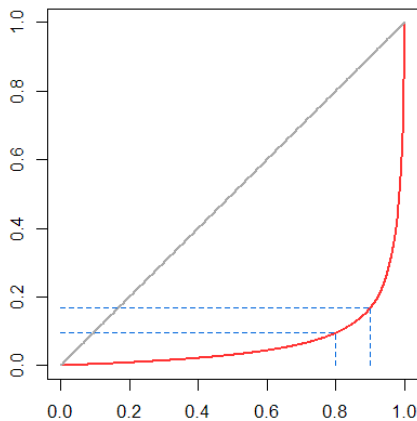
mit Werten im Intervall $[0, 1]$.

■ **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).

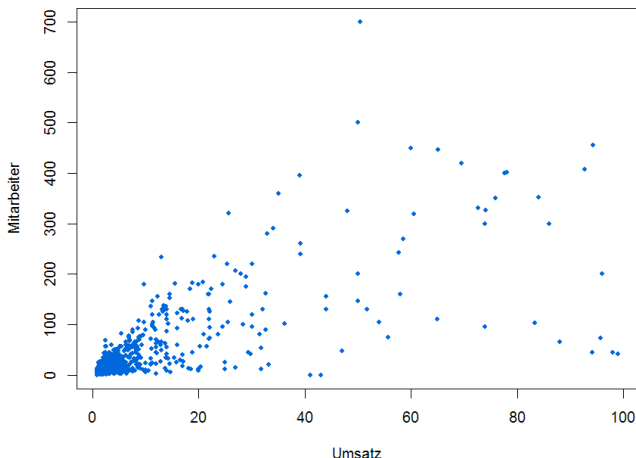
Die entsprechenden Merkmalsausprägungen seien

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

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

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) : $(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$.

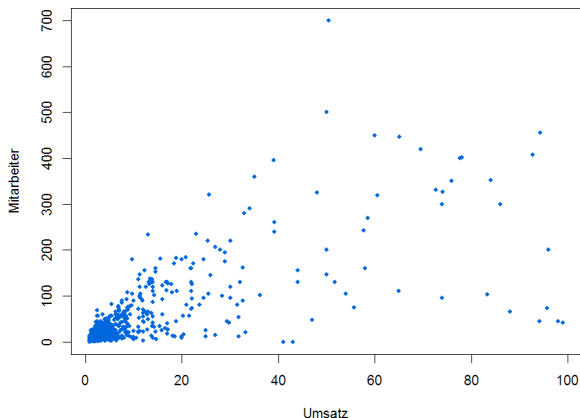
Daher wählt man als Maßzahl die empirische Kovarianz

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

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:

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

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

$$s_{xx} = \text{Var}(x)$$

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

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

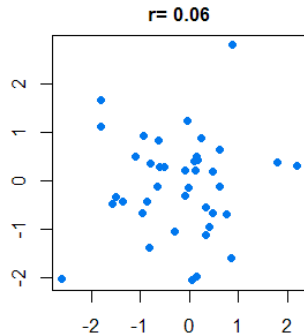
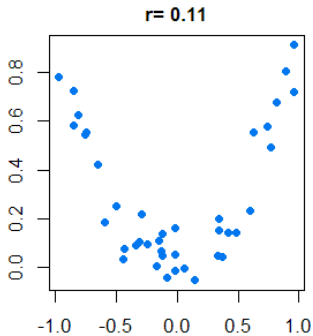
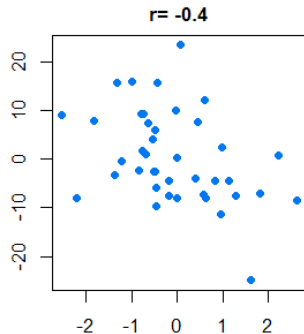
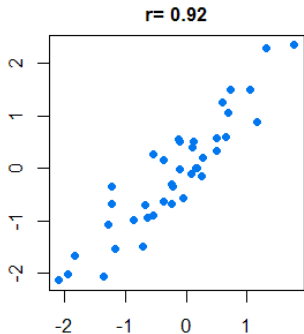
$$r_{xy} = \frac{s_{xy}}{\sigma(x)\sigma(y)}$$

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 Statistikklausurnote 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	Σ
\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:

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

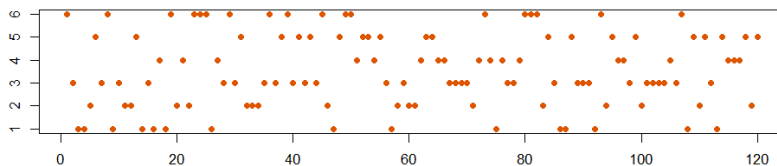
- Es gilt wieder $R_{xy} \in [-1, 1]$.

- 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?

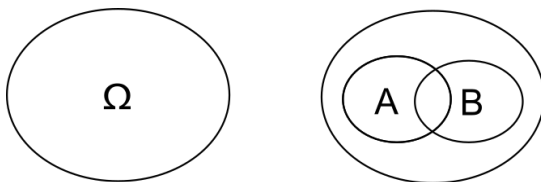
Frequentistische Interpretation: Die Wahrscheinlichkeit eines Ereignisses ist der Zahlenwert, gegen die relative Häufigkeit mit wachsendem Stichprobenumfang konvergiert.

3.1. Ereignisse und Wahrscheinlichkeiten

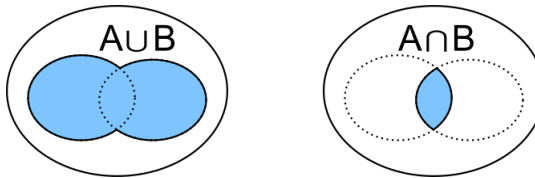
Die axiomatische Wahrscheinlichkeitstheorie lässt die philosophischen Fragen hinter sich und betrachtet Ereignisse und Wahrscheinlichkeiten als mathematische Objekte mit bestimmten Eigenschaften.

Das Grundgerüst kennen wir bereits aus der Statistik:

- Die Grundgesamtheit Ω wird nun Wahrscheinlichkeitsraum genannt.
- Die Merkmale heißen nun Zufallsvariablen.
- Die Teilmengen von Ω heißen Ereignisse.

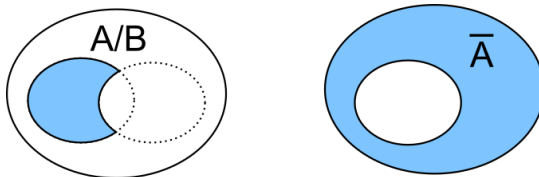


- Die gesamte Menge Ω repräsentiert das sichere Ereignis, die leere Menge \emptyset das unmögliche Ereignis.
- Die Vereinigungsmenge $A \cup B$ repräsentiert das Eintreten von A oder von B (dabei wird zugelassen, dass beide Ereignisse eintreten).
- Die Schnittmenge $A \cap B$ repräsentiert das gleichzeitige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

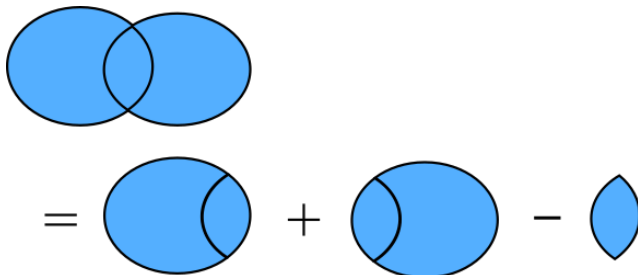
□ In der mathematischen Wahrscheinlichkeitstheorie stellt sich heraus, dass man nicht jedem Ereignis eine Wahrscheinlichkeit zuordnen kann. Das führt zu einigen Komplikationen, die wir hier ignorieren wollen (\rightarrow Vitali-Mengen, Banach-Tarski-Paradoxon).

- 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:

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

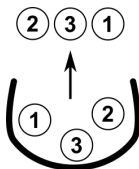
Die Ereignisse A und \bar{A} sind nicht unabhängig:

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

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?

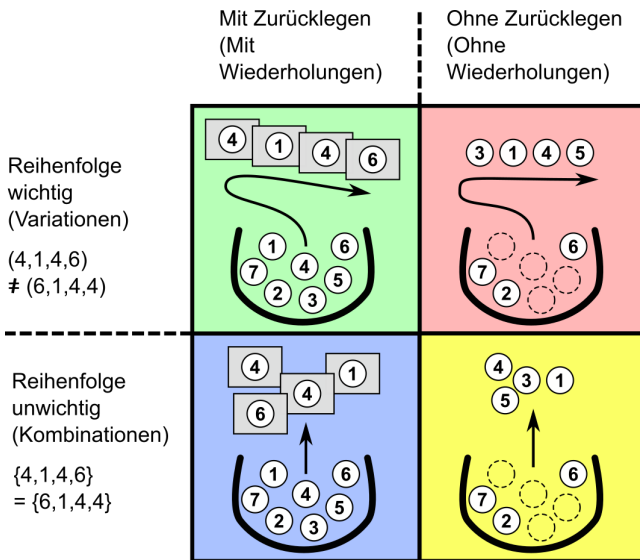


■ Satz 3.4

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:

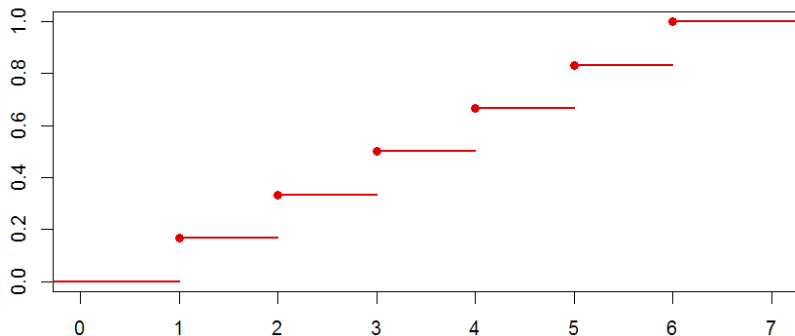
$$p(x) = P(X = x) = \begin{cases} 0 & ; x \notin M_X \\ P(X = x_i) & ; x = x_i \end{cases}$$

- 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

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

Beide Größen beschreiben die Streuung der Zufallsvariablen X .

Es gelten die schon vom arithmetischen Mittel vertrauten Rechenregeln:

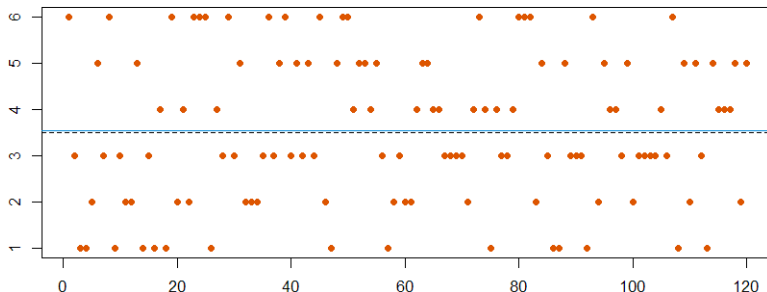
- $E(aX + b) = aE(X) + b,$
- $\text{Var}(aX + b) = a^2\text{Var}(X),$
- $\sigma(aX + b) = a\sigma(X),$
- $E(X + Y) = E(X) + E(Y).$

3.5. Das Gesetz der großen Zahlen

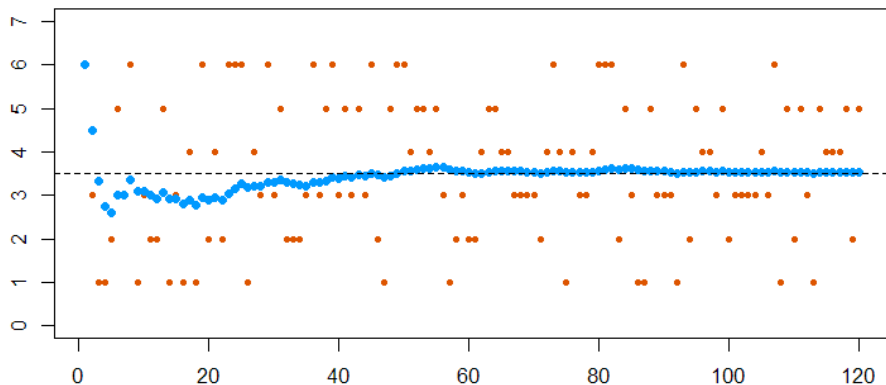
■ **Beispiel B3.7** \Rightarrow **B1.1**: Im Beispiel **B1.1** ergab sich ein arithmetisches Mittel von $\bar{x} = 3.55$. Das liegt verdächtig nahe beim theoretischen Erwartungswert

$$E(X) = 3.5$$

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 μ 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

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

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

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

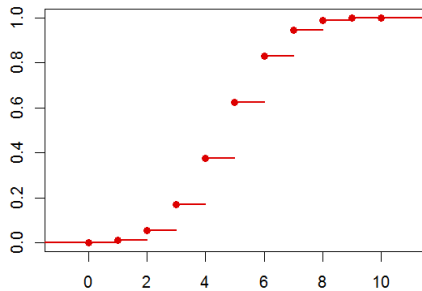
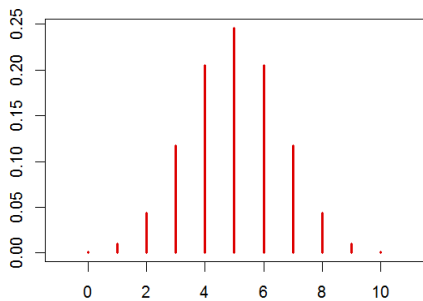
eine Binomialverteilung und es gilt

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

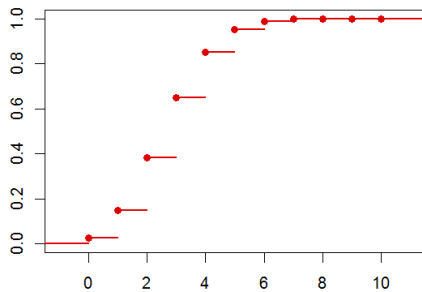
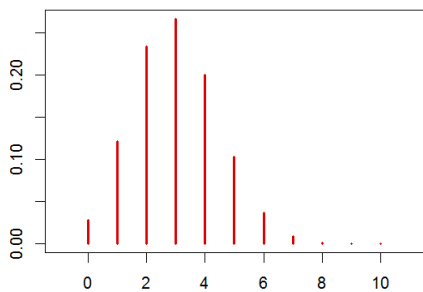
Dann ergibt sich

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

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



$n = 10, p = 0.5$



$n = 10, p = 0.3$

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):

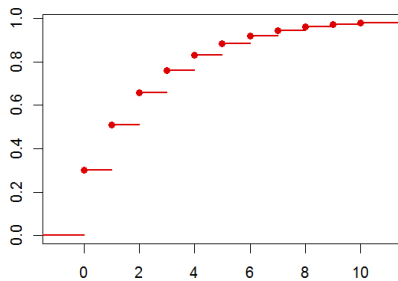
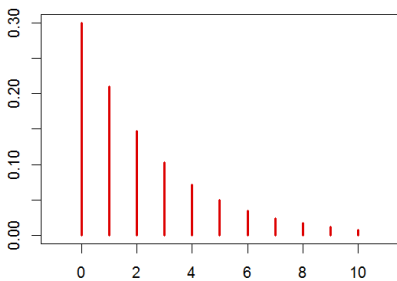
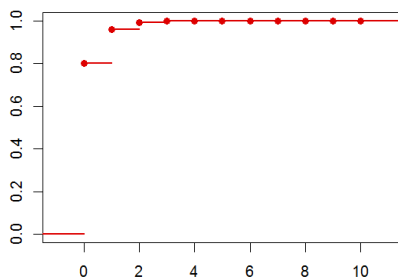
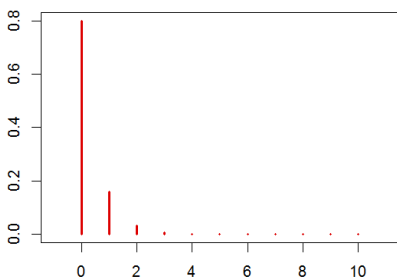
$$P(Z = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

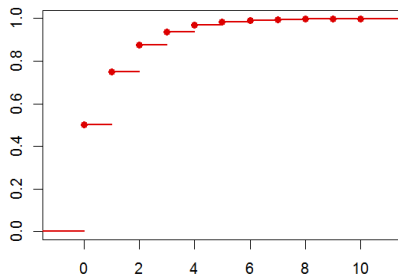
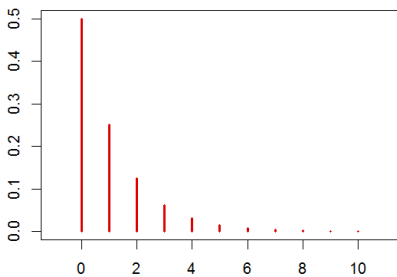
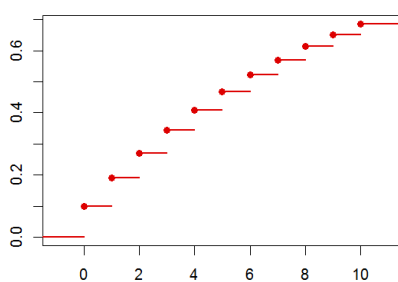
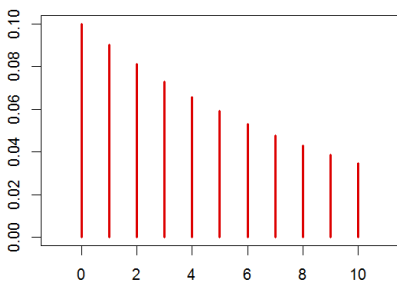
Die Anzahl der Misserfolge $M = Z - 1$ hat eine geometrische Verteilung vom Typ II:

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

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.

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

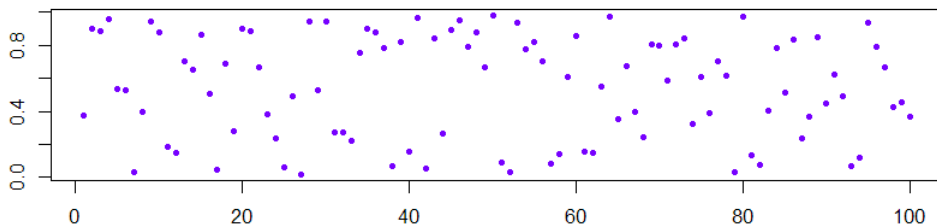
Als Erwartungswert erhalten wir

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

3.7.5. Die stetige Gleichverteilung

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

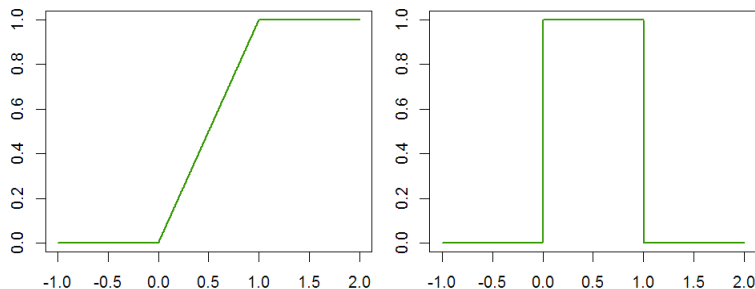
- Handelsübliche Taschenrechner verfügen über eine **RND**-Taste, die gleichverteilte Zufallszahlen erzeugt.
- Mit Hilfe gleichverteilter Zufallsvariablen kann man anders verteilte Zufallszahlen erzeugen (Inversionsmethode, Monte-Carlo-Simulation)



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](#).

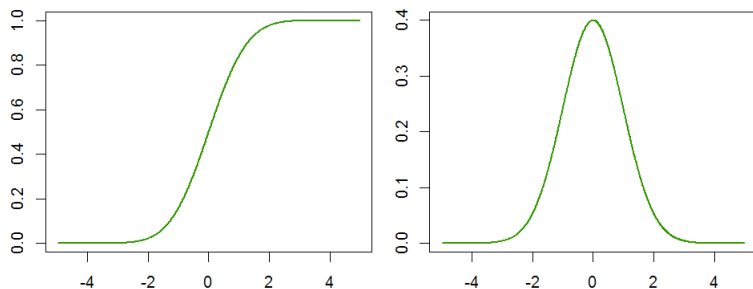
- Die Standardnormalverteilung besitzt die Dichtefunktion

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

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

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

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



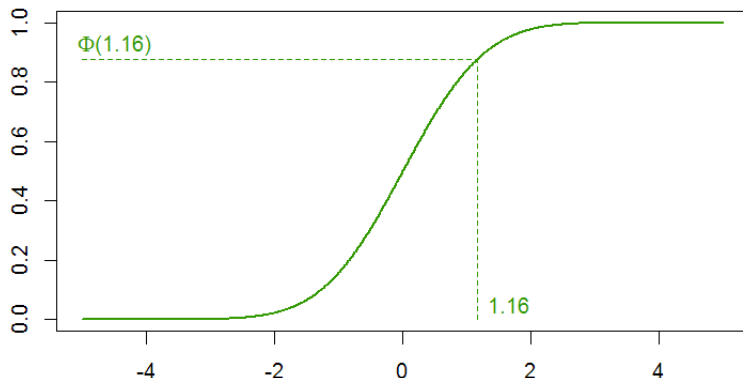
$$\mu = 0, \sigma = 1$$

- 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 256).

Beispiel: $\Phi(1.16) = 0.877$

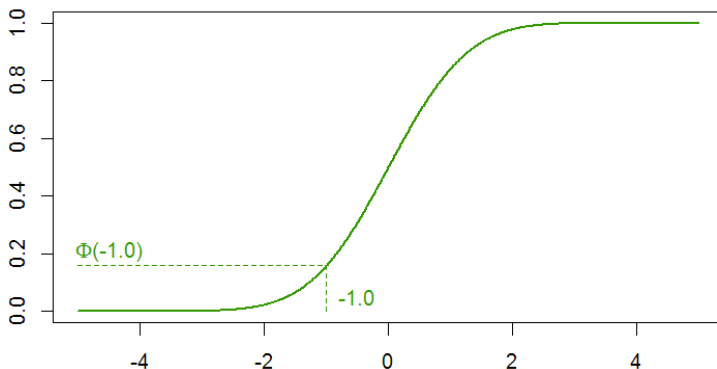


- Für negative Argumente kann man die Umformungsregel

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

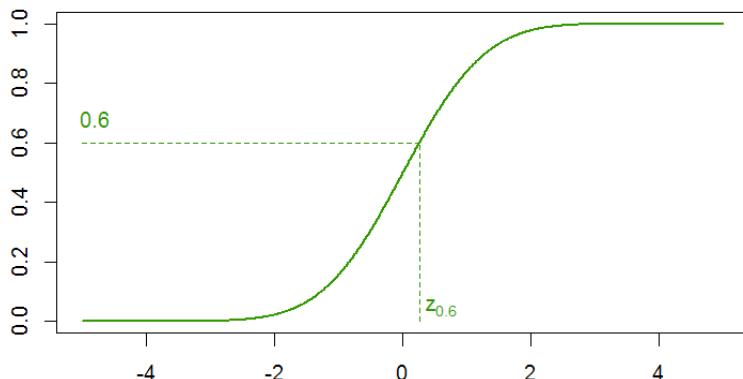
verwenden.

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



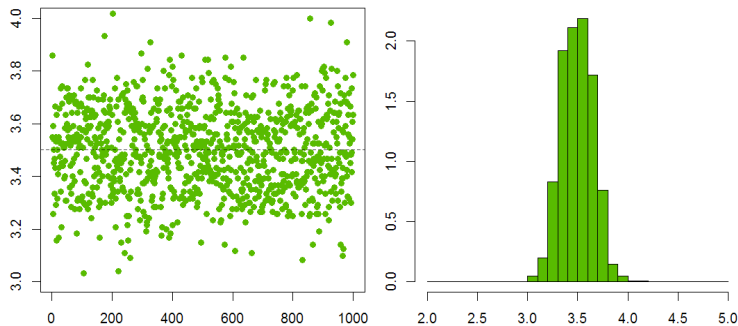
- Als α -Quantil bezeichnet den Wert z für den $\Phi(z) = \alpha$ gilt. Man verwendet die Bezeichnung z_α für diesen Wert.
- Die Quantile kann man ebenfalls aus der Tabelle auf Seite 256 entnehmen.

Beispiel: $z_{0.6} = 0.25$.



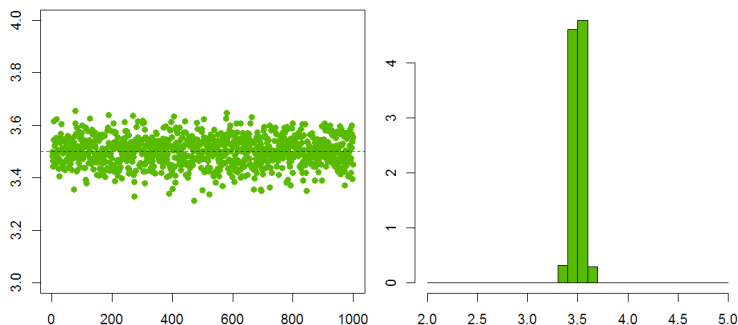
3.8.3. Der zentrale Grenzwertsatz

■ **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 μ und Standardabweichung σ und

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

ihr arithmetisches Mittel.

Dann gilt

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

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

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

■ Satz 3.8

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

Es folgt, dass die standardisierte Zufallsvariable

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

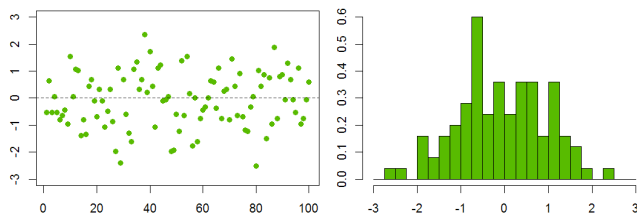
den Erwartungswert 0 und die Standardabweichung 1 besitzt.

Wir können auch mit n erweitern und schreiben:

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

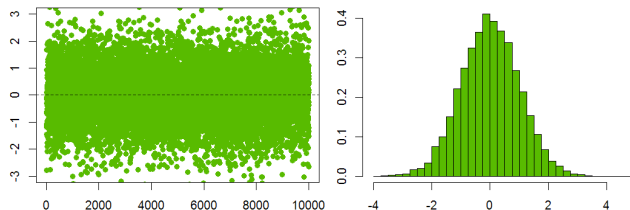
Welche Verteilung besitzt X_n^* ?

120 Würfe, 100 Mal wiederholt:



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

10 000 Würfe, 10 000 Mal wiederholt:



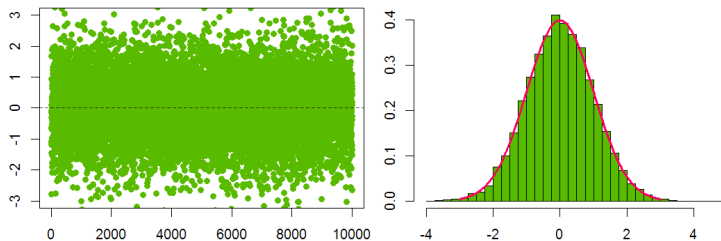
$$\mu = 0, \sigma = 1$$

■ Satz 3.9 (Zentraler Grenzwertsatz)

Gegeben seien unabhängige und identisch verteilte Zufallsvariablen X_1, X_2, \dots mit Erwartungswert μ und Varianz σ^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

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

■ **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?

$$\begin{aligned} P(\bar{X}_{120} > 3.55) &= 1 - P(\bar{X}_{120} \leq 3.55) \\ &\approx 1 - \Phi\left(\frac{3.55 - 3.5}{\sqrt{\frac{35}{12}}/\sqrt{120}}\right) \\ &= 1 - \Phi(0.3207135) \\ &\stackrel{S.256}{=} 1 - 0.6255 \\ &= 0.3745 \end{aligned}$$

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


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

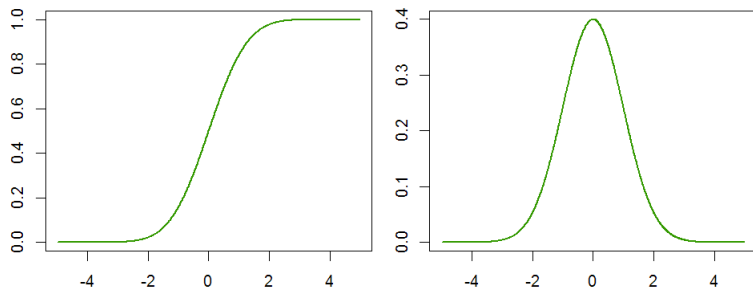
3.8.5. Die allgemeine Normalverteilung

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

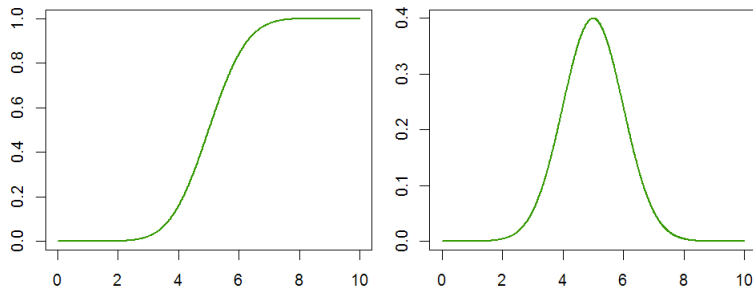
- Die Normalverteilung besitzt die Dichtefunktion

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

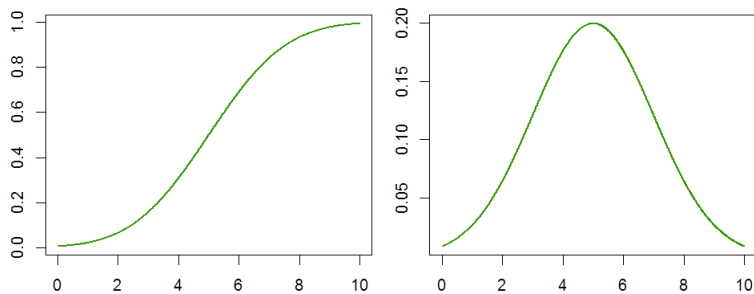
- Die zugehörige Verteilungsfunktion $\Phi_{\mu, \sigma}$ lässt sich wieder nicht in geschlossener Form angeben.
 - Wir schreiben $N(\mu, \sigma)$ für die Normalverteilung.
-  In vielen Büchern bezeichnet $N(\mu, s)$ eine Normalverteilung mit Erwartungswert μ und Varianz s .



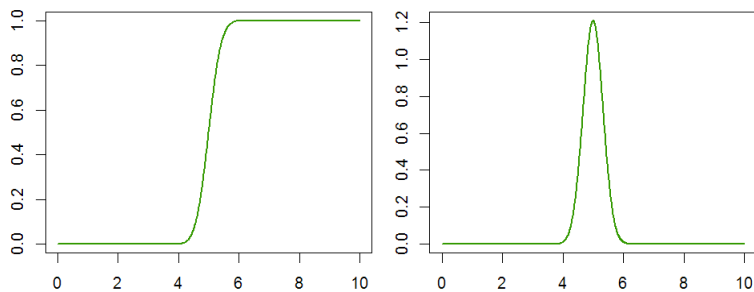
$$\mu = 0, \sigma = 1$$



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



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



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

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

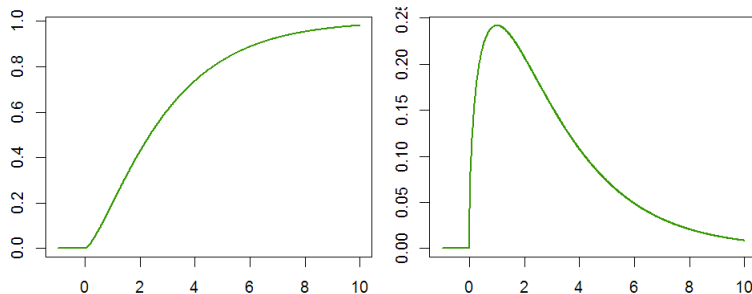
$$\bar{X}_n \sim N(\mu, \sigma/\sqrt{n}).$$

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 α -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 258 zu entnehmen. Zum Beispiel ist

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

Das bedeutet, dass

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

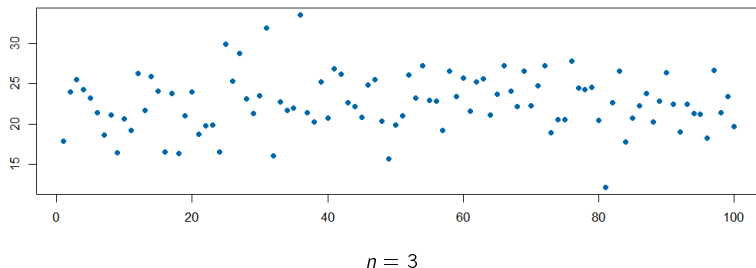
ist, wenn die X_i unabhängige standardnormalverteilte Zufallsvariablen sind.

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 α -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 [259](#) zu entnehmen.
Beispielsweise ergibt sich

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

d.h.

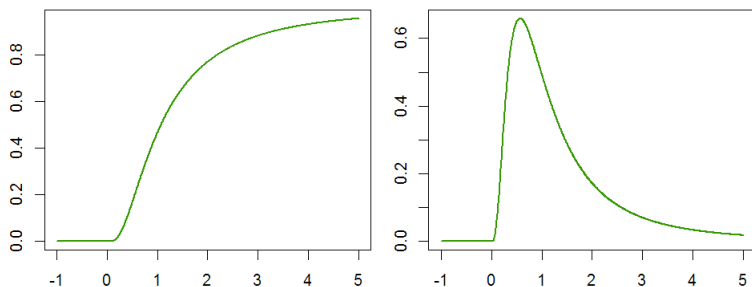
$$P(T \leq 1.325) = 0.9.$$

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 α -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 260. Es ist z.B.

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

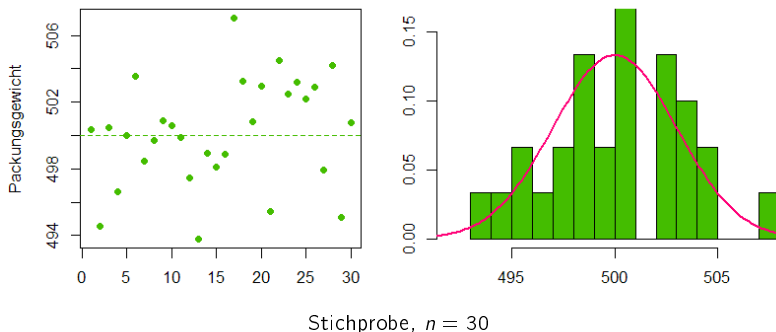
d.h.

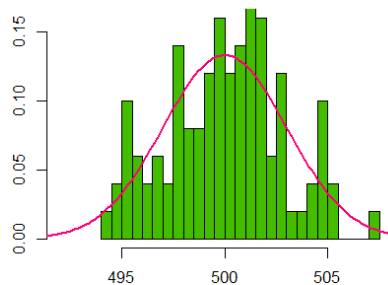
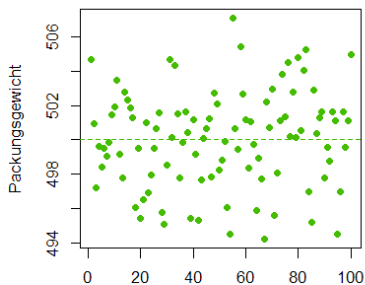
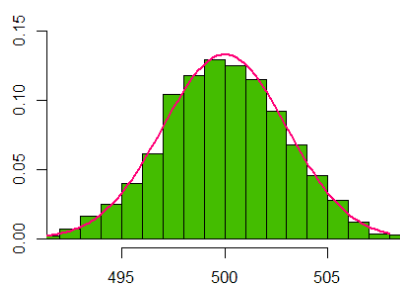
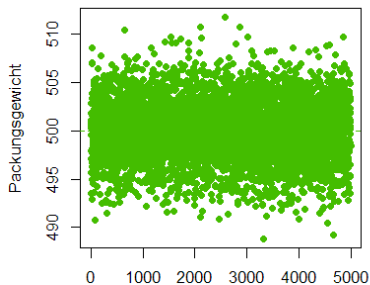
$$P(F \leq 4.735) = 0.95.$$

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.

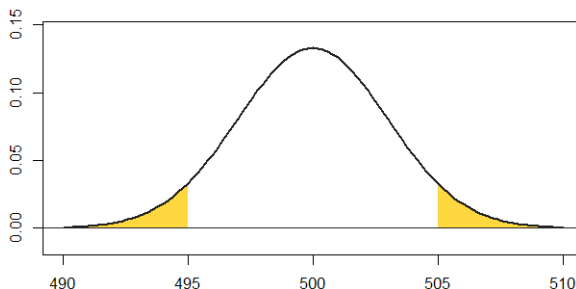
Das Gewicht X einer Obstpackung sei normalverteilt mit einem Mittelwert von $\mu = 500g$ und einer Standardabweichung von $\sigma = 3$:



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

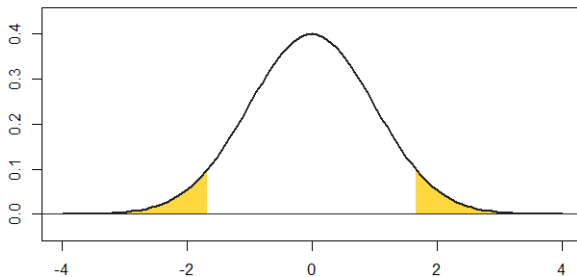
Nach einer Norm für den Obsthandel darf die Packungsgröße der Ware nicht um mehr als fünf Gramm vom angegebenen Gewicht abweichen.

Wie groß ist die Wahrscheinlichkeit einer solchen unzulässigen Abweichung?



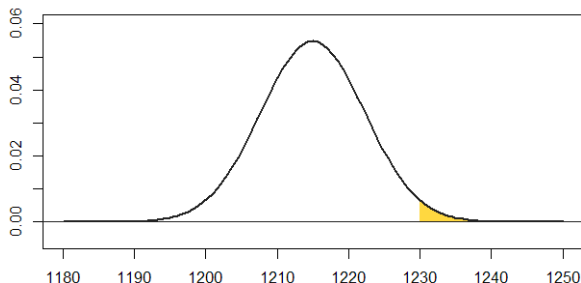
Wir transformieren X in eine standardnormalverteilte Zufallsvariable:

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



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 $\sigma(Y) = 0.003 \cdot 2430 = 7.29$.



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

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

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

$$\sigma = \sigma(X_1)$$

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 μ oder die Varianz σ^2 .

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

4.1.1. Punktschätzer für den Erwartungswert

Es sei μ der Erwartungswert der Zufallsvariablen X_1, X_2, X_3, \dots

- Ein naheliegender Schätze für μ ist der Mittelwert

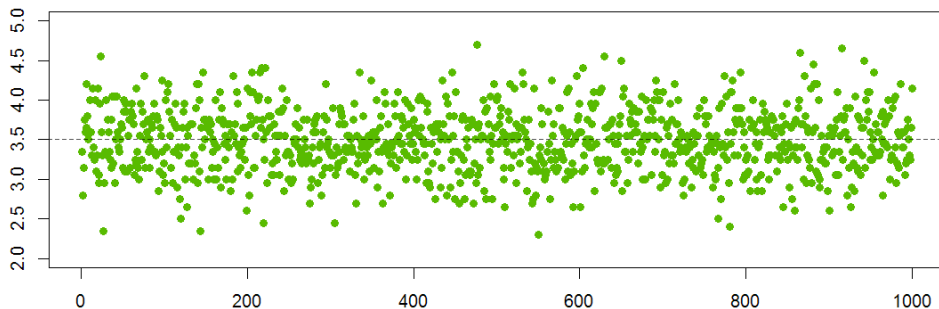
$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- Dabei ist zu beachten, dass $\hat{\mu}$, im Gegensatz zur Zahl μ , 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)

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

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

- Außerdem gilt

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

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 μ und bekanntem σ , 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 μ 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 μ 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

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

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 μ 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 σ^2 bekannt ist.

- Dann ist $\hat{\mu}$ normalverteilt mit Erwartungswert μ und Standardabweichung σ/\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

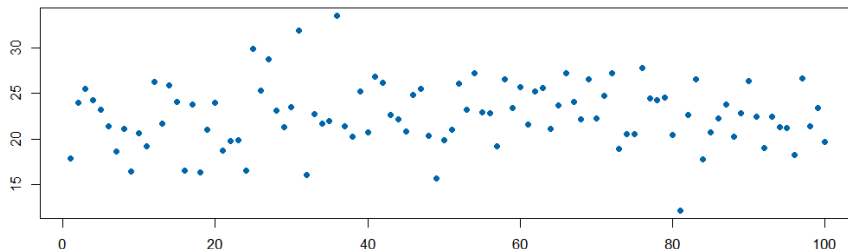
$$P\left(\hat{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- 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 μ .

■ **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 μ liegt nicht mit 90% bzw. 95% Wahrscheinlichkeit in diesen Intervallen! μ 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

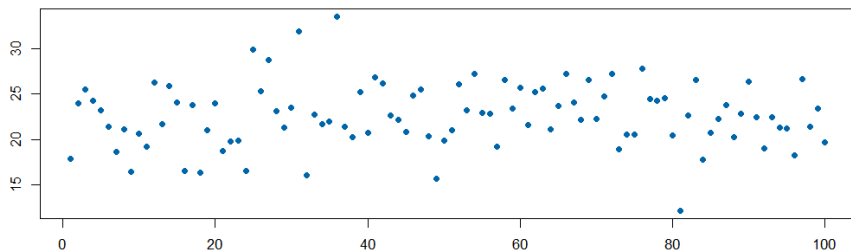
$$P\left(\hat{\mu} \leq \mu - t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}\right) = \frac{\alpha}{2}.$$

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

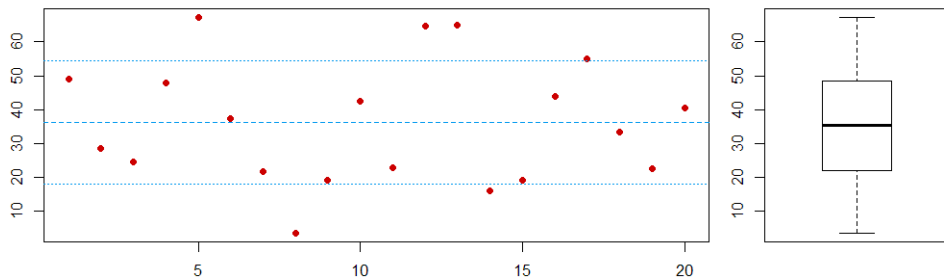
$$P(\sigma^2 \leq c\hat{\sigma}^2) = P\left((n-1) \cdot \frac{\hat{\sigma}^2}{\sigma^2} \geq \frac{n-1}{c}\right) = 1 - F((n-1)/c),$$

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 ($\alpha = 10\%$)

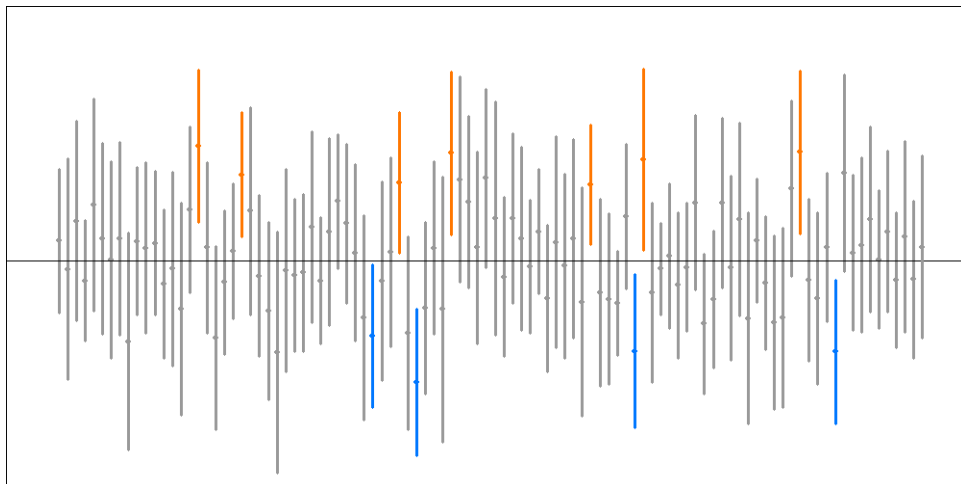
$$\begin{aligned} & \left[\hat{\mu} - t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + t_{n-1, 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \\ &= [29.23, 43.23] \end{aligned}$$

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 μ 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 μ .
- 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 μ einen bestimmten Wert μ_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 α , 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 α .
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

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

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!

Übung 1

Aufgabe 1: Es sei $x = (6, 1, 3, 4, 1)$. Berechnen Sie:

a) $\frac{1}{5} \sum_{k=1}^5 x_k$

b) $\sum_{l=1}^5 (x_l - 3)^2$

c) $\sum_{i=1}^5 i \cdot x_{6-i}$

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) $(x \cdot y)^b = x^b \cdot y^b$

b) $(x + y)^b = x^b + y^b$

c) $e^{(x^2)} = (e^x)^2$

d) $\sqrt{x^2} = |x|$

e) $\log(x + y) = \log(x) + \log(y)$

f) $\log(x \cdot y) = \log(x) \cdot \log(y)$

g) $\log(x \cdot y) = \log(x) + \log(y)$

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) $f(x) = 2x - 3$

b) $f(x) = \log(x)$

c) $f(x) = e^x$

d) $f(x) = e^{-x}$

e) $f(x) = e^{-x^2}$

f) $f(x) = e^{-(x-1)^2}$

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 Ω 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 Merkmalsskalen 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 Über-

nachtungszahlen in klassierter Form betrachtet:

Klasse:	0-100	100-500	500-2000	2000-5000
#Hotels:	20	90	40	50

- a) Zeichnen Sie ein Histogramm.
- b) Zeichnen Sie ein Diagramm, das die zugehörige empirische Dichte zeigt.
- c) 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 α -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 $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)]$?
- 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) $\bar{x} = 5$, $\sigma(x) = 1$,
- 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 3×2 -Kontingenztafel für zwei unabhängige Merkmale an.

Aufgabe 24: Ein neues Produkt kommt in drei Versionen 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 χ^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. Dabei sei $\sum_{i=1}^{120} x_i^2 = 1818$, wobei x_i die i-te Augenzahl bezeichnet.
- 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.

A.**Anhang****A.1. Kleine Formelsammlung****A.1.1. Wahrscheinlichkeiten**

$P(\overline{A}) = 1 - P(A)$	
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	
$P(A \cup B) = P(A) + P(B)$	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

A.1.2. Erwartungswerte

$E(aX + b) = aE(X) + b$	$a, b \in \mathbb{R}$
$E(X + Y) = E(X) + E(Y)$	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 $E(X_1) = E(X_1) = \dots = \mu$
$E(\bar{X}) = \mu$	falls $E(X_1) = E(X_1) = \dots = \mu$

A.1.3. Varianz und Standardabweichung

$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$	
$\sigma(X) = \sqrt{\text{Var}(X)}$	
$\text{Var}(aX + b) = a^2 \text{Var}(X)$	$a, b \in \mathbb{R}$
$\sigma(aX + b) = a \sigma(X)$	$a, b \in \mathbb{R}$
$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$	
$\sigma(X + Y) = \sqrt{\sigma(X)^2 + \sigma(Y)^2 + 2\text{Cov}(X, Y)}$	
$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$	falls X, Y unkorreliert
$\sigma(X + Y) = \sqrt{\sigma(X)^2 + \sigma(Y)^2}$	falls X, Y unabhängig

A.1.4. Tabellen

- a) Verteilungsfunktion der Standardnormalverteilung.
- b) Quantile der Standardnormalverteilung.
- c) Quantile der Chi-Quadrat-Verteilung.
- d) Quantile der t-Verteilung.
- e) Quantile der F-verteilung.

Standardnormalverteilung $\Phi(x)$

	0.	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1.	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2.	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

Quantile der Standardnormalverteilung

α	0.8	0.9	0.95	0.975	0.99	0.995	0.999
z_α	0.842	1.282	1.645	1.960	2.326	2.576	3.090
α	0.2	0.1	0.05	0.025	0.01	0.005	0.001
z_α	-0.842	-1.282	-1.645	-1.960	-2.326	-2.576	-3.090

Quantile der Chi-Quadrat-Verteilung (vertikal: Freiheitsgrad)

	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	7.88	6.63	5.02	3.84	2.71	0.02	0.	0.	0.	0.
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.34	9.35	7.81	6.25	0.58	0.35	0.22	0.11	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41
6	18.55	16.81	14.45	12.59	10.64	2.2	1.64	1.24	0.87	0.68
7	20.28	18.48	16.01	14.07	12.02	2.83	2.17	1.69	1.24	0.99
8	21.95	20.09	17.53	15.51	13.36	3.49	2.73	2.18	1.65	1.34
9	23.59	21.67	19.02	16.92	14.68	4.17	3.33	2.7	2.09	1.73
10	25.19	23.21	20.48	18.31	15.99	4.87	3.94	3.25	2.56	2.16
11	26.76	24.72	21.92	19.68	17.28	5.58	4.57	3.82	3.05	2.6
12	28.3	26.22	23.34	21.03	18.55	6.3	5.23	4.4	3.57	3.07
13	29.82	27.69	24.74	22.36	19.81	7.04	5.89	5.01	4.11	3.57
14	31.32	29.14	26.12	23.68	21.06	7.79	6.57	5.63	4.66	4.07
15	32.8	30.58	27.49	25.	22.31	8.55	7.26	6.26	5.23	4.6
16	34.27	32.	28.85	26.3	23.54	9.31	7.96	6.91	5.81	5.14
17	35.72	33.41	30.19	27.59	24.77	10.09	8.67	7.56	6.41	5.7
18	37.16	34.81	31.53	28.87	25.99	10.86	9.39	8.23	7.01	6.26
19	38.58	36.19	32.85	30.14	27.2	11.65	10.12	8.91	7.63	6.84
20	40.	37.57	34.17	31.41	28.41	12.44	10.85	9.59	8.26	7.43
21	41.4	38.93	35.48	32.67	29.62	13.24	11.59	10.28	8.9	8.03
22	42.8	40.29	36.78	33.92	30.81	14.04	12.34	10.98	9.54	8.64
23	44.18	41.64	38.08	35.17	32.01	14.85	13.09	11.69	10.2	9.26
24	45.56	42.98	39.36	36.42	33.2	15.66	13.85	12.4	10.86	9.89
25	46.93	44.31	40.65	37.65	34.38	16.47	14.61	13.12	11.52	10.52
26	48.29	45.64	41.92	38.89	35.56	17.29	15.38	13.84	12.2	11.16
29	52.34	49.59	45.72	42.56	39.09	19.77	17.71	16.05	14.26	13.12
39	65.48	62.43	58.12	54.57	50.66	28.2	25.7	23.65	21.43	20.
49	78.23	74.92	70.22	66.34	62.04	36.82	33.93	31.55	28.94	27.25
59	90.72	87.17	82.12	77.93	73.28	45.58	42.34	39.66	36.7	34.77
69	103.	99.23	93.86	89.39	84.42	54.44	50.88	47.92	44.64	42.49
79	115.12	111.14	105.47	100.75	95.48	63.38	59.52	56.31	52.72	50.38
89	127.11	122.94	116.99	112.02	106.47	72.39	68.25	64.79	60.93	58.39
99	138.99	134.64	128.42	123.23	117.41	81.45	77.05	73.36	69.23	66.51
149	197.21	192.07	184.69	178.49	171.51	127.35	121.79	117.1	111.8	108.29
199	254.14	248.33	239.96	232.91	224.96	173.9	167.36	161.83	155.55	151.37
499	584.13	575.42	562.79	552.07	539.89	458.97	448.2	439.	428.46	421.38

Quantile der t-Verteilung

	0.8	0.9	0.95	0.975	0.99	0.995	0.999
1	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	1.386	1.886	2.92	4.303	6.965	9.925	22.327
3	1.25	1.638	2.353	3.182	4.541	5.841	10.215
4	1.19	1.533	2.132	2.776	3.747	4.604	7.173
5	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	1.134	1.44	1.943	2.447	3.143	3.707	5.208
7	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	1.108	1.397	1.86	2.306	2.896	3.355	4.501
9	1.1	1.383	1.833	2.262	2.821	3.25	4.297
10	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	1.083	1.356	1.782	2.179	2.681	3.055	3.93
13	1.079	1.35	1.771	2.16	2.65	3.012	3.852
14	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	1.071	1.337	1.746	2.12	2.583	2.921	3.686
17	1.069	1.333	1.74	2.11	2.567	2.898	3.646
18	1.067	1.33	1.734	2.101	2.552	2.878	3.61
19	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	1.063	1.323	1.721	2.08	2.518	2.831	3.527
22	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	1.06	1.319	1.714	2.069	2.5	2.807	3.485
24	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	1.058	1.316	1.708	2.06	2.485	2.787	3.45
26	1.058	1.315	1.706	2.056	2.479	2.779	3.435
29	1.055	1.311	1.699	2.045	2.462	2.756	3.396
39	1.05	1.304	1.685	2.023	2.426	2.708	3.313
49	1.048	1.299	1.677	2.01	2.405	2.68	3.265
59	1.046	1.296	1.671	2.001	2.391	2.662	3.234
69	1.044	1.294	1.667	1.995	2.382	2.649	3.213
79	1.043	1.292	1.664	1.99	2.374	2.64	3.197
89	1.043	1.291	1.662	1.987	2.369	2.632	3.184
99	1.042	1.29	1.66	1.984	2.365	2.626	3.175
149	1.04	1.287	1.655	1.976	2.352	2.609	3.146
199	1.039	1.286	1.653	1.972	2.345	2.601	3.132
499	1.038	1.283	1.648	1.965	2.334	2.586	3.107

Quantile der F-Verteilung (horizontal/vertikal: 1. und 2.Freiheitsgrad), $\alpha = 0.95$

	1	2	3	4	5	6	7	8	9
1	161.448	199.5	215.707	224.583	230.162	233.986	236.768	238.883	240.543
2	18.513	19.	19.164	19.247	19.296	19.33	19.353	19.371	19.385
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999
5	6.608	5.786	5.409	5.192	5.05	4.95	4.876	4.818	4.772
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099
7	5.591	4.737	4.347	4.12	3.972	3.866	3.787	3.726	3.677
8	5.318	4.459	4.066	3.838	3.687	3.581	3.5	3.438	3.388
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.23	3.179
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.02
14	4.6	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646
19	4.381	3.522	3.127	2.895	2.74	2.628	2.544	2.477	2.423
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131
49	4.038	3.187	2.794	2.561	2.404	2.29	2.203	2.134	2.077
74	3.97	3.12	2.728	2.495	2.338	2.224	2.136	2.066	2.009
99	3.937	3.088	2.696	2.464	2.306	2.192	2.103	2.033	1.976
999	3.851	3.005	2.614	2.381	2.223	2.108	2.019	1.948	1.889

	10	14	19	29	39	49	74	99	999
1	241.882	245.364	247.686	249.951	251.062	251.723	252.595	253.028	254.187
2	19.396	19.424	19.443	19.461	19.47	19.475	19.482	19.486	19.495
3	8.786	8.715	8.667	8.62	8.596	8.582	8.563	8.554	8.529
4	5.964	5.873	5.811	5.75	5.719	5.701	5.677	5.664	5.632
5	4.735	4.636	4.568	4.5	4.466	4.446	4.419	4.405	4.369
6	4.06	3.956	3.884	3.813	3.777	3.755	3.727	3.712	3.673
7	3.637	3.529	3.455	3.381	3.343	3.321	3.29	3.275	3.234
8	3.347	3.237	3.161	3.084	3.046	3.022	2.991	2.975	2.932
9	3.137	3.025	2.948	2.869	2.829	2.805	2.772	2.756	2.712
10	2.978	2.865	2.785	2.705	2.664	2.639	2.606	2.589	2.543
14	2.602	2.484	2.4	2.314	2.27	2.243	2.206	2.188	2.136
19	2.378	2.256	2.168	2.077	2.03	2.001	1.961	1.941	1.884
29	2.177	2.05	1.958	1.861	1.809	1.777	1.733	1.71	1.645
39	2.084	1.954	1.86	1.759	1.704	1.67	1.623	1.598	1.526
49	2.03	1.899	1.803	1.699	1.643	1.607	1.557	1.531	1.453
74	1.961	1.828	1.729	1.621	1.561	1.524	1.469	1.44	1.352
99	1.928	1.793	1.693	1.582	1.521	1.482	1.425	1.394	1.297
999	1.84	1.702	1.597	1.479	1.412	1.367	1.3	1.261	1.11

Quantile der F-Verteilung (horizontal/vertikal: 1. und 2.Freiheitsgrad), $\alpha = 0.99$

	1	2	3	4	5	6	7	8
1	4052.181	4999.5	5403.352	5624.583	5763.65	5858.986	5928.356	5981.07
2	98.503	99.	99.166	99.249	99.299	99.333	99.356	99.374
3	34.116	30.817	29.457	28.71	28.237	27.911	27.672	27.489
4	21.198	18.	16.694	15.977	15.522	15.207	14.976	14.799
5	16.258	13.274	12.06	11.392	10.967	10.672	10.456	10.289
6	13.745	10.925	9.78	9.148	8.746	8.466	8.26	8.102
7	12.246	9.547	8.451	7.847	7.46	7.191	6.993	6.84
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467
10	10.044	7.559	6.552	5.994	5.636	5.386	5.2	5.057
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.14
19	8.185	5.926	5.01	4.5	4.171	3.939	3.765	3.631
29	7.598	5.42	4.538	4.045	3.725	3.499	3.33	3.198
49	7.182	5.066	4.208	3.728	3.416	3.195	3.028	2.898
99	6.898	4.826	3.986	3.515	3.208	2.99	2.825	2.696
999	6.66	4.626	3.801	3.338	3.036	2.82	2.657	2.529

	9	10	14	19	29	49	99	999
1	6022.473	6055.847	6142.674	6200.576	6257.053	6301.231	6333.79	6362.679
2	99.388	99.399	99.428	99.447	99.465	99.479	99.489	99.498
3	27.345	27.229	26.924	26.719	26.517	26.359	26.241	26.137
4	14.659	14.546	14.249	14.048	13.85	13.694	13.578	13.475
5	10.158	10.051	9.77	9.58	9.391	9.242	9.131	9.031
6	7.976	7.874	7.605	7.422	7.24	7.096	6.988	6.891
7	6.719	6.62	6.359	6.181	6.003	5.862	5.756	5.66
8	5.911	5.814	5.559	5.384	5.209	5.07	4.964	4.869
9	5.351	5.257	5.005	4.833	4.66	4.521	4.416	4.321
10	4.942	4.849	4.601	4.43	4.258	4.12	4.015	3.92
14	4.03	3.939	3.698	3.529	3.359	3.219	3.113	3.015
19	3.523	3.434	3.195	3.027	2.855	2.714	2.603	2.501
29	3.092	3.005	2.767	2.599	2.423	2.276	2.159	2.047
49	2.793	2.706	2.469	2.299	2.118	1.963	1.835	1.708
99	2.592	2.505	2.267	2.094	1.908	1.743	1.601	1.449
999	2.425	2.339	2.099	1.923	1.729	1.55	1.385	1.159

Inhalt

1	Einführung	1
1.1	Was ist Statistik?	3
1.2	R	5
2	Deskriptive Statistik	6
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
3	Wahrscheinlichkeitsrechnung	111
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
4	Induktive Statistik	189
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

A	Anhang	252
A.1	Kleine Formelsammlung	252
A.1.1	Wahrscheinlichkeiten	252
A.1.2	Erwartungswerte	253
A.1.3	Varianz und Standardabweichung	254
A.1.4	Tabellen	255