MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

Juristische & Wirtschaftswissenschaftliche Fakultät
Lehrstuhl für Statistik

Prof. Dr. Claudia Becker

Statistik I

(Kapitel 1 – 5)

Statistik II

(Kapitel 6 – 10)

Wintersemester 2021 / 2022

Sommersemester 2022

Inhaltsverzeichnis

1 Beispiele und Grundbegriffe

1.1 Beispiele, Prinzipien statistischer Untersuchungen

Statistik:

- wissenschaftliche Methoden zur zahlenmäßigen Erfassung, Untersuchung und Darstellung von Massenerscheinungen und zufallsbehafteten Ereignissen

- insbes. im Umgangssprachlichen oft: (schriftlich) dargestelltes Ergebnis einer Untersuchung nach erfolgter statistischer Auswertung

Beispiel 1.1 *Politische Umfragen*

- *Befragungen zur Beliebtheit von Politikern, zur Beurteilung der wirtschaftlichen Lage, zur Entscheidung bei einer anstehenden Wahl ("Sonntagsfrage"), ...*

- *oft von Meinungsforschungsinstituten, z.B. Allensbach, Emnid*

- *es kann immer nur ein Teil der Bevölkerung befragt werden, das Ergebnis der Befragung soll aber möglichst die Meinung der gesamten Bevölkerung darstellen*
  → geeignete Auswahl einer **Stichprobe**
  → Beurteilung des **Stichprobenfehlers** *(wie groß ist die zufällige Schwankung der Aussagen gegenüber substantiellen Größen?)*
  → Schluss von Stichprobe auf **Grundgesamtheit** *möglich?*

- *Darstellung der Umfrageergebnisse, z.B. Kreisdiagramme, Säulendiagramme*

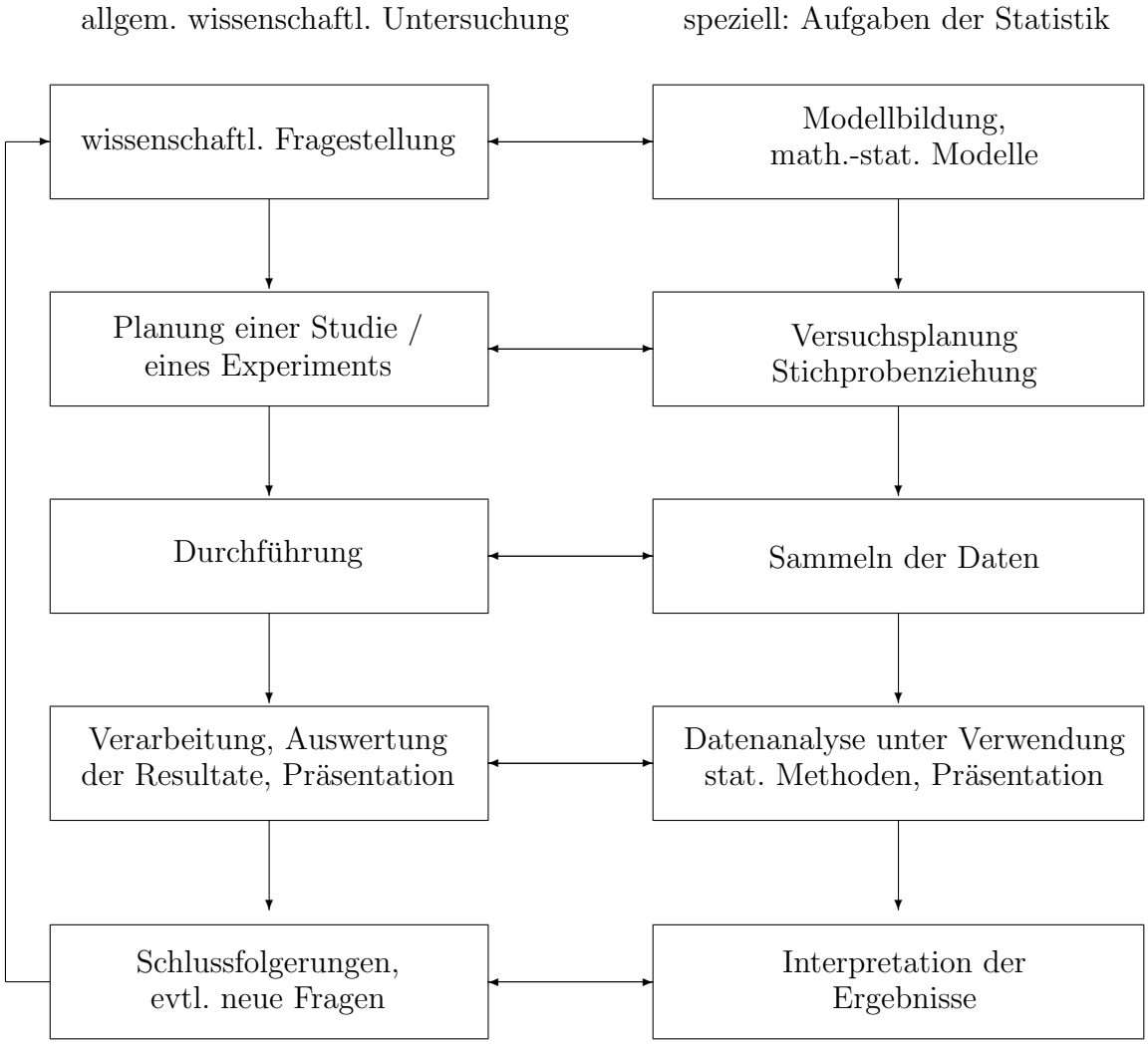
Beispiel 1.2 *Klinische Studien*

Bevor neues Medikament zur Verordnung zugelassen wird, muss seine Wirksamkeit und seine Unbedenklichkeit nachgewiesen werden (Studie an Patienten).

Typische Wirksamkeitsuntersuchung:

- *zwei Gruppen von Patienten*
- *Gruppe A erhält neues Medikament, Gruppe B erhält Placebo*
- *Erfassung der Wirkung, dabei Abhängigkeit von äußeren "Umständen" wie z.B. Geschlecht, allgem. Gesundheitszustand, Dosierung*
  → geeignete Auswahl von Patienten
  → geeignete Zuordnung der Behandlung (wer kommt in Gruppe A, wer in Gruppe B?)
  → Zeitraum der Studie
  → Auswahl der zu erfassenden **Merkmale**
  usw.

Ablauf einer statistischen Untersuchung



Beispiel 1.3 *Untersuchung zu Warteschlangen*

Hintergrund: Deutsche Bahn stellte in vielen Servicecentern das Warteschlangensystem um

von "eine Warteschlange pro Schalter"

auf "eine gemeinsame Warteschlange für alle Schalter"

Fragestellung: Welches System ist günstiger für die Wartezeit der Kunden?
Modell: sog. Warteschlangenmodell, stochastischer Prozess

Planung: Erhebung von Kundenankünften und Bearbeitungszeiten im Service-Zentrum eines Bahnhofs
Stichprobenziehung: Erhebung an ausgewählten Tagen (Stoßzeiten und ruhige Zeiten)

Durchführung: Ankünfte und Bedienzeiten von Kunden im Bahnhof erheben

Analyse: mit statistischer Theorie aus gemessenen Ankunfts- und Bearbeitungsraten für beide Warteschlangensysteme die sog. "erwartete" Wartezeit berechnen; auch: gemessene Daten darstellen

Schlussfolgerung: die "gemeinsame Warteschlange" ist günstiger, weil insbesondere die im schlimmsten Fall zu erwartende Wartezeit deutlich geringer ist.

Aus dem Ablauf einer wissenschaftlichen / statistischen Untersuchung ableitbar: drei Grundaufgaben der Statistik

- Beschreiben von Datenmaterial
  → **deskriptive Statistik**

  rein beschreibende Aufbereitung und Komprimierung umfangreicher Datensätze durch Tabellen, Graphiken, Kenngrößen (z.B. Mittelwert)

- Suchen nach Strukturen, Gewinnung neuer Forschungshypothesen
  → **explorative Statistik**

  weiterentwickelte deskriptive Statistik; typischerweise angewandt, wenn die zu untersuchende Fragestellung sehr vage oder die Auswahl eines geeigneten Modells nicht klar ist (moderne Weiterentwicklung: Data Mining)

- Schließen von Experiment / Stichprobe auf Grundgesamtheit
  → **induktive Statistik**

  auf Basis erhobener Daten wird versucht, allgemeine Schlussfolgerungen auf die Grundgesamtheit zu ziehen; dies erfordert

  - sorgfältige Versuchs- / Stichprobenplanung
  - deskriptive Analyse
  - Wahrscheinlichkeitsrechnung
  - Festlegung eines statistischen Modells
  - Schätz- und Testtheorie

1.2 Grundbegriffe

Untersuchungseinheiten, Grundgesamtheit und Stichprobe

Daten werden an gewissen Objekten beobachtet, z.B.

Wirksamkeit eines Medikaments	an	Patienten
Lebensdauern	an	elektronischen Geräten
Ankunftsraten	an	Bahnkunden
Einschätzung der wirtschaftlichen Lage	an	$\underbrace{\text{Personen aus der Bevölkerung}}$ sog. **Untersuchungseinheiten**

Definition 1.4 *Untersuchungseinheit*

Untersuchungseinheit = *Einzelobjekt einer statistischen Untersuchung*

Untersuchungseinheit ist Informationsträger

Dabei gilt: jede Untersuchungseinheit wird hinsichtlich des Untersuchungsziels durch sachliche, räumliche und zeitliche Kriterien **abgegrenzt** *bzw.* **identifiziert**.

Beispiel 1.5 *Konsumverhalten*

Umfrage zum Konsumverhalten der Bundesbürger

sachlich: volljährige Bürger

räumlich: z.B. Sachsen-Anhalt oder BRD

zeitlich: Tag der Umfrage

Die Menge aller Untersuchungseinheiten, über die man Aussagen gewinnen möchte, ist die sog. **Grundgesamtheit** (wichtig: muss klar umgrenzt sein). Problem: Untersuchung der kompletten Grundgesamtheit (**Vollerhebung**, z.B. alle Bundesbürger) oft nicht möglich, da zeitaufwändig und mit hohen Kosten verbunden (etwa: Volkszählung). Daher oft nur Teilerhebung, sog. **Stichprobenziehung**.

Definition 1.6 *Grundgesamtheit, Stichprobe, Erhebungsgesamtheit*

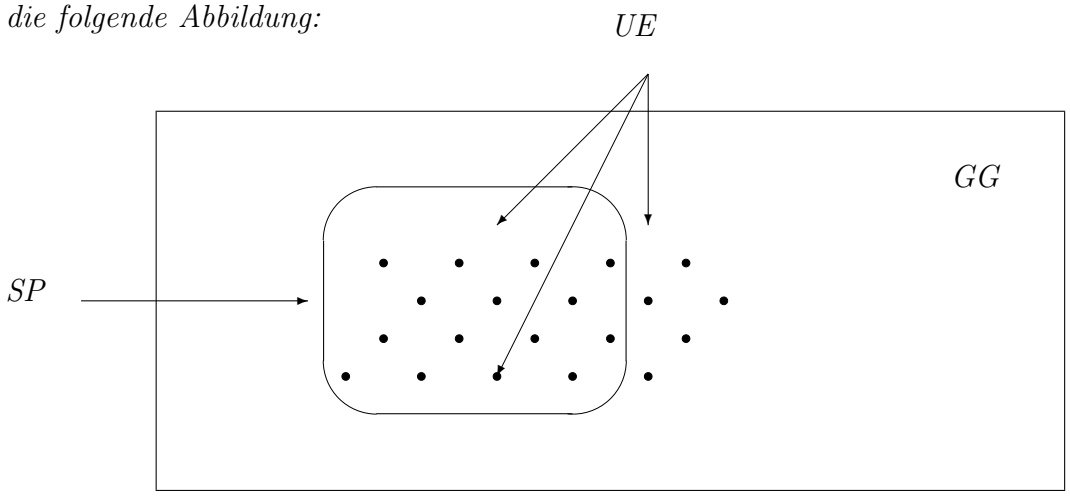
Grundgesamtheit *(GG): Gesamtheit von Untersuchungseinheiten mit übereinstimmenden Identifikationskriterien*

Stichprobe*: Bildung einer Teilmenge der Untersuchungseinheiten bei einer statistischen Untersuchung*

Erhebungsgesamtheit*: Menge der tatsächlich erhobenen Untersuchungseinheiten (*$E = GG \rightarrow$ *Vollerhebung,* $E \subset GG \rightarrow$ *Stichprobe)*

Bemerkung 1.7 *Grundgesamtheit, Untersuchungseinheiten, Stichprobe*

Zur Illustration der Zusammenhänge zwischen den Begriffen betrachte man die folgende Abbildung:

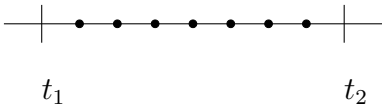


Verschiedene Arten von Grundgesamtheiten möglich, z.B.

Bestandsmasse ("stock")
– zu einem festen Zeitpunkt

z.B. Einwohner in Halle am ..., MLU-Studenten am 1.10.02

Bewegungsmasse ("flow")
– in einem Zeitraum



z.B. Zu- und Fortzüge Deutscher aus dem / ins Ausland im Jahr 2000, Geburten in Sachsen-Anhalt im Januar 2002

Merkmale und Merkmalstypen

Eigentlich interessieren nicht die Untersuchungseinheiten selbst, sondern bestimmte Eigenschaften der Untersuchungseinheiten → sog. **Merkmale** oder **Variablen** (z.B. interessiert nicht der Patient selbst, sondern man will wissen, ob das Medikament wirkt; es interessiert nicht der Passant, sondern seine Einschätzung der wirtschaftlichen Lage).

Definition 1.8 *Merkmal, Merkmalsausprägung*
Merkmal: *Größe oder Eigenschaft einer Untersuchungseinheit, die auf Grund der interessierenden Fragestellung erhoben bzw. gemessen wird*
Merkmalsausprägung: *möglicher Wert (Kategorie), den ein Merkmal annehmen kann*

Untersuchungseinheiten heißen auch **Merkmalsträger**.

Beispiel 1.9 *Mietspiegel*

Bestimmung der ortsüblichen Vergleichsmiete: "die üblichen Entgelte, die in der Gemeinde oder vergleichbaren Gemeinden für nicht preisgebundenen Wohnraum vergleichbarer Art, Größe, Ausstattung, Beschaffenheit und Lage in den letzten vier Jahren vereinbart oder, von Erhöhungen nach §4 MHG abgesehen, geändert worden sind."

→ *Nettomiete abhängig von Merkmalen wie*

Art:	*Altbau, Neubau*
Lage:	*Innenstadt, Stadtrand*
Größe:	*40m², 95m², ...*
$\underbrace{\textit{Baujahr:}}_{\textit{Merkmale}}$	$\underbrace{\textit{1932, 1965, 1983, 1995, ...}}_{\textit{Ausprägungen}}$

- In der Regel werden mehrere Merkmale an einem Merkmalsträger beobachtet
  z.B. "Wetter um 12 Uhr an der Wetterstation Brocken"
  Merkmale: Temperatur, Niederschlagsmenge, Luftdruck, Bewölkung, Luftfeuchtigkeit, Sicht, ...

- Merkmalsausprägungen müssen keine Zahlen sein
  z.B. Bewölkung: "wolkenlos", "heiter", "leicht bewölkt", "wolkig", "bedeckt"
  oder Autofarben: "rot", "grün", "schwarz", ...

- **Schreibweise:**
  Merkmale mit großen lateinischen Buchstaben, z.B. X, Y, Z
  Merkmalsausprägungen mit kleinen lateinischen Buchstaben, z.B. x_1, x_2, x_3, x, y, z

- Häufig: Unterscheidung von Merkmalen / Variablen in solche, die
  - beeinflussen → kontrollierbar → **Einflussgrößen**
    → nicht kontrollierbar → **Störgrößen**
  - beeinflusst werden → **Zielgrößen**

  z.B. Wirkung eines Medikaments: Heilungserfolg = Zielgröße, Dosierung = Einflussgröße, Geschlecht = Störgröße

An den Beispielen der Merkmale und Ausprägungen: man erkennt, dass sich die Merkmale in ihrem Informationsgehalt unterscheiden

Beispiel 1.10 *Merkmale*

Merkmal	*Ausprägungen*		*Art*
Geschlecht	*m / w*	→ 2 *Auspr., keine Ordnung*	*qualitativ*
Automarke	*Fiat, Mercedes, Toyota, ...*	→ *? Auspr., keine Ordnung*	*qualitativ*
Diplomnote	1, 2, 3, 4, 5	→ 5 *Auspr., Ordnung, Abst. nicht interpret.*	*Rangmerkmal*
Beliebtheit von Politikern	*sehr, mäßig, gar nicht*	→ 3 *Auspr., Ordnung, Abst. nicht interpret.*	*Rangmerkmal*
Anzahl Kinder in einer Familie	0, 1, 2, 3, ...	→ *? Auspr., Ordnung, Abstand interpretierbar, keine Auspr. zwischen zwei anderen möglich*	*quantitativ, diskret*
Regenmenge an einem Tag	20*mm*, 50*mm*, ...	→ *unendl. viele Auspr., Ordnung, Abstand interpretierbar, zwischen zwei Auspr. immer weitere möglich*	*quantitativ, stetig*

Definition 1.11 *Merkmale*

qualitatives Merkmal: *es gibt weder eine natürliche Ordnung der Ausprägungen, noch ist es sinnvoll, Abstände (Differenzen) oder Verhältnisse (Quotienten) der Ausprägungen zu betrachten*

Rangmerkmal: *es gibt eine natürliche Ordnung der Ausprägungen, aber es ist nicht sinnvoll, Abstände oder Verhältnisse zu betrachten*

quantitatives Merkmal: *es gibt eine natürliche Ordnung der Ausprägungen, Abstände oder Abstände und Verhältnisse sind interpretierbar*

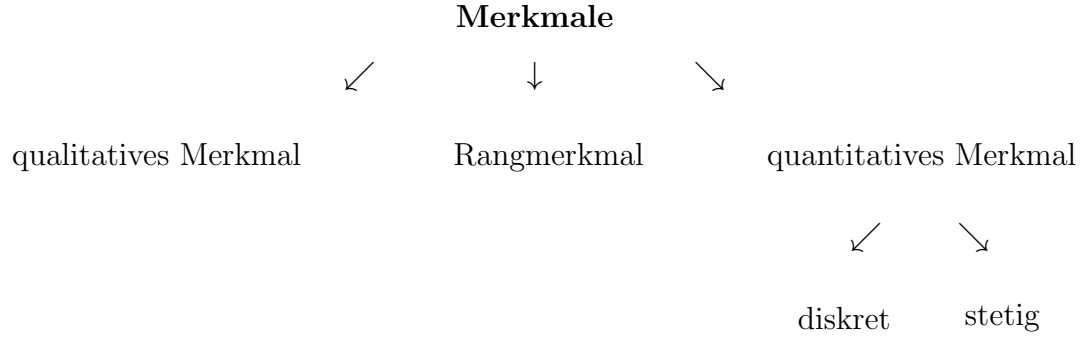
quantitative Merkmale werden unterschieden in diskrete und stetige

diskretes Merkmal: *Ausprägungen sind isolierte Zustände, Menge der möglichen Ausprägungen ist abzählbar*

stetiges Merkmal: *Ausprägungen liegen in einem Intervall, zwischen je zwei Ausprägungen ist stets eine weitere möglich*

Beachte: Jede praktische Messung bei stetigen Merkmalen ist – durch die jeweilige Grenze der Messgenauigkeit bedingt – diskret.

Klassifikation von Merkmalen



Bemerkung 1.12 *Skalen*

Die verschiedenen Merkmalstypen werden auf unterschiedlichen Skalen gemessen:

qualitatives Merkmal	$\rightarrow$	*Nominalskala*
Rangmerkmal	$\rightarrow$	*Ordinalskala*
quantitatives Merkmal	$\rightarrow$	*Kardinalskala / metrische Skala*

Höhere Skalen können stets (durch "Vergröbern") in niedrigere umgewandelt werden.

Zulässige bzw. sinnvoll interpretierbare Berechnungen für Daten aus den verschiedenen Messskalen:

Skala	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
metrisch	ja	ja	ja	nein
	ja	ja	ja	ja

2 Die Aufbereitung eindimensionaler Merkmale

2.1 Häufigkeitsverteilungen

2.1.1 Absolute und relative Häufigkeiten

Beispiel 2.1 *Absolventenstudie*

36 Absolventen des Soziologiestudiums an der LMU München wurden befragt nach den Merkmalen

G = Geschlecht (1 = weiblich, 2 = männlich)

S = Studiendauer in Semestern

E = fachliches Engagement im Studium (1 = sehr engagiert, ..., 5 = gar nicht)

D = Ausrichtung Diplomarbeit mit

1 = *empirisch (Primärerhebung)*

2 = *empirisch (Sekundäranalyse)*

3 = *empirisch (qualitativ)*

4 = *theoretisch (Literaturarbeit)*

N = Gesamtnote Diplomprüfung

Die Daten sind in der Tabelle auf der folgenden Seite dargestellt.

Frage: Wie können diese Daten zusammengefasst werden?

Daten aus der Münchner Absolventenstudie 1995 des Instituts für Soziologie der LMU München

Person i	*G*	*S*	*E*	*D*	*N*	*Person i*	*G*	*S*	*E*	*D*	*N*
1	1	12	1	3	2	21	1	13	3	4	2
2	1	13	3	4	2	22	2	13	4	3	3
3	1	12	5	4	3	23	1	15	1	4	2
4	1	12	2	3	3	24	1	13	3	2	2
5	1	9	3	4	2	25	2	15	4	4	3
6	1	12	2	1	1	26	1	12	2	4	2
7	2	14	5	3	3	27	1	14	1	3	2
8	2	10	1	4	2	28	1	10	2	4	2
9	1	18	3	3	1	29	1	12	3	3	2
10	2	10	3	4	3	30	1	17	2	3	2
11	1	13	4	4	3	31	1	11	1	4	2
12	1	15	4	3	2	32	1	14	3	2	3
13	2	13	2	2	2	33	1	11	2	1	2
14	1	16	3	3	2	34	2	13	2	4	3
15	1	14	3	4	2	35	2	11	3	4	3
16	1	13	2	3	2	36	2	7	1	4	2
17	1	13	2	4	2						
18	1	17	1	4	3						
19	2	12	2	2	2						
20	1	15	2	3	3						

Definition 2.2 *Absolute und relative Häufigkeit*

In einer Stichprobe vom Umfang n sei das Merkmal X erhoben worden.

Urliste (Rohdaten): *die beobachteten Werte von X in der erhobenen Reihenfolge, Bezeichnung:* $x_1, \ldots, x_n$

Merkmalsausprägungen: *Menge aller Ausprägungen von X, Bezeichnung:* $a_1, \ldots, a_k$

absolute Häufigkeit *der Ausprägung* a_i*:* $h_i = h(a_i) =$ *Anzahl der Fälle, in denen* a_i *auftritt, wobei* $\sum_{i=1}^{k} h_i = n$

relative Häufigkeit *der Ausprägung* a_i*:* $f_i = f(a_i) = \frac{h(a_i)}{n} =$ *Anteil der Untersuchungseinheiten, die Ausprägung* a_i *besitzen, wobei* $\sum_{i=1}^{k} f_i = 1$

absolute Häufigkeitsverteilung: $h_1, \ldots, h_k$

relative Häufigkeitsverteilung: $f_1, \ldots, f_k$

Oft fasst man die Häufigkeiten in einer Häufigkeitstabelle zusammen (**unklassierte Häufigkeitsverteilung**).

Beispiel 2.3 *Absolventenstudie*

Merkmal D = Ausrichtung der Diplomarbeit

Urliste: 3, 4, 4, 3, 4, 1, ..., 1, 4, 4, 4

Menge der Merkmalsausprägungen: 1, 2, 3, 4

(Unklassierte) Häufigkeitstabelle:

Ausrichtung a_i	1	2	3	4	$\sum$
$h(a_i)$	2	4	12	18	36
$f(a_i)$	$\frac{2}{36} = 0.056$	$\frac{4}{36} = 0.111$	$\frac{12}{36} = 0.333$	$\frac{18}{36} = 0.5$	1

Bei der Erstellung einer Häufigkeitsverteilung ist es oft sinnvoll oder sogar nötig, die Information aus der Urliste zu straffen, da

- Anzahl der Merkmalsausprägungen zu groß,
- stetiges Merkmal,
- Übersichtlichkeit leidet bei Erfassung aller Merkmalsausprägungen.

Ausweg: **Klassenbildung**

Benachbarte Merkmalsausprägungen werden zu einer **Klasse** oder **Gruppe** zusammengefasst. In der **klassierten Häufigkeitsverteilung** erscheinen nur noch die Gruppen mit der Häufigkeit aller Ausprägungen in der Gruppe.

Beispiel 2.4 *Mietspiegel*

Merkmal: Nettomieten in €

Urliste für $n = 26$ Wohnungen

127.06	172	194.1	217.3	226.74
228.74	238.04	248.86	272.06	337.74
347.94	349.57	349.85	373.81	375.74
378.4	383.05	394.97	426.91	443.40
466.84	467.88	533.11	539.28	560.21
676.74				

klassierte Häufigkeitstabelle:

Klasse i	h_i	f_i
$100 < \ldots \leq 200$	3	$3/26 = 0.115$
$200 < \ldots \leq 300$	6	0.230
$300 < \ldots \leq 400$	9	0.346
$400 < \ldots \leq 500$	4	0.153
$500 < \ldots \leq 600$	3	0.115
$600 < \ldots \leq 700$	1	0.038

Bemerkung 2.5 *Hinweise zur Klassenwahl*

(a) Klasseneinteilung so wählen, dass Merkmalswerte möglichst gleichmäßig auf die Klassen verteilt sind ("Leerklasssen" vermeiden!)

(b) Falls n Werte $\Rightarrow \sqrt{n}$ *Klassen ("Daumenregel", nicht immer sinnvoll)*

(c) oft: Rechenvereinfachung bei äquidistanten Klassen, d.h. gleicher Klassenbreite

(d) Klassen müssen vollständig sein, d.h. jeder Merkmalswert liegt in genau einer Klasse

(e) offene Randklassen: im unteren und / oder oberen Bereich sind (halboffene) unbeschränkte Intervalle möglich

2.1.2 Tabellarische und graphische Darstellungen

Tabelle: dient der systematischen und übersichtlichen Zusammenfassung von Daten
zu beachten:

- jede Tabelle muss eine Überschrift haben, die den wesentlichen Inhalt in möglichst knapper Form kennzeichnet
- bei Zahlentabellen keine leeren Felder
  0.000 $\hat{=}$ fast Null, kleiner als kleinste Einheit
  x $\hat{=}$ keine Angabe möglich
  0 $\hat{=}$ genau Null
  ...
- beachte: Bezeichnungen können unterschiedlich sein; ggf. Erläuterungen erforderlich

Schema für eine Tabelle:

Überschrift

(Titel und wichtige Angaben)

Kopf zu Vorspalte	←	Tabellenkopf	→
↑			
Vorspalte			
↓			

← Zeilen ←

↑ Spalten ↑

Fußnoten (evtl. Erläuterungen)

- übersichtlich, leicht lesbar und unmissverständlich bezeichnet

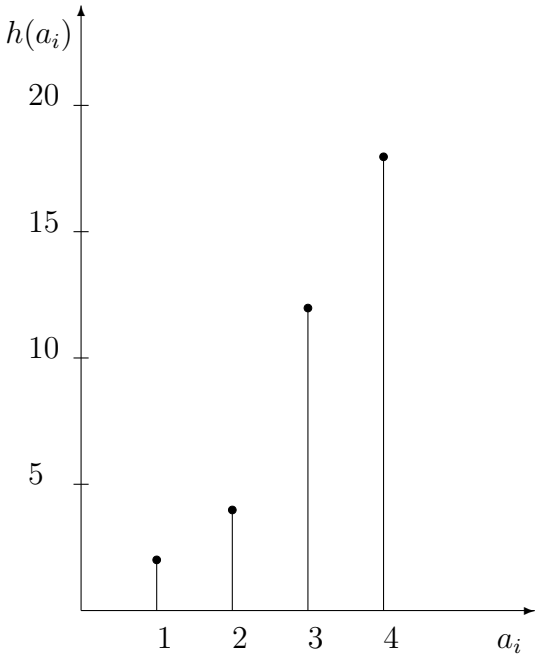
Graphische Darstellungen: Art hängt vom Merkmalstyp ab

(i) qualitatives Merkmal (nominal, ordinal)

- Stabdiagramm
- Säulen- oder Balkendiagramm
- Kreisdiagramm

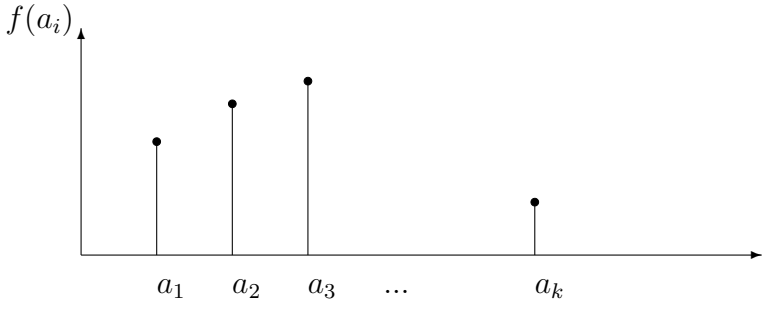
Beispiel 2.6 *Stabdiagramm*

Absolventenstudie, Merkmal D = Ausrichtung Diplomarbeit



hier: die Längen der Stäbe geben die absoluten Häufigkeiten an

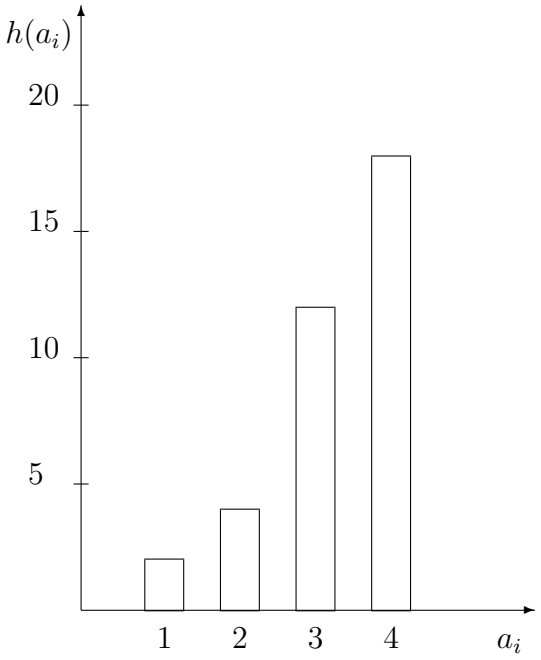
Auch möglich: Längen der Stäbe als relative Häufigkeiten



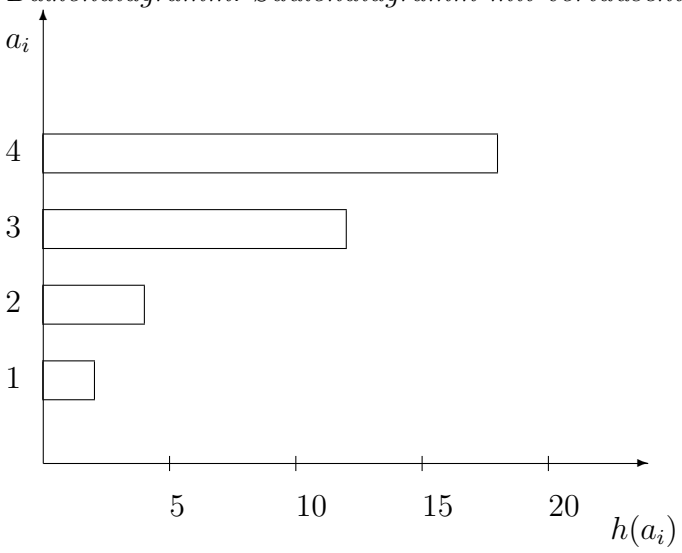
Beispiel 2.7 *Säulen- und Balkendiagramm*

Säulendiagramm: im Prinzip wie Stabdiagramm, nur ersetze die Stäbe durch Rechtecke gleicher Breite

Absolventenstudie, Merkmal D



Balkendiagramm: Säulendiagramm mit vertauschten Achsen



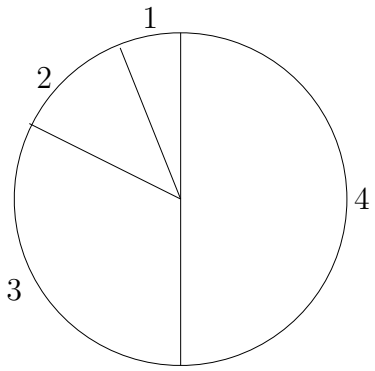
Beachte: bei Stab-, Säulen- und Balkendiagramm gilt das Prinzip der **Längentreue**, d.h. die Länge der Stäbe, Säulen, Balken ist proportional zur Häufigkeit.

Beispiel 2.8 *Kreisdiagramm*

Kreis wird in Segmente unterteilt; jedes Segment ist einer Merkmalsausprägung zugeordnet; der Winkel, der den Kreisausschnitt festlegt, ist proportional zur Häufigkeit der Ausprägung.

Es gilt: Winkel = rel. Häufigkeit $\cdot 360$ *Grad*

Absolventenstudie, Merkmal D



Legende:

1 = *empirisch (primär)*

2 = *empirisch (sekundär)*

3 = *empirisch (qualitativ)*

4 = *theoretisch*

(ii) quantitatives Merkmal (metrisch)

- Histogramm

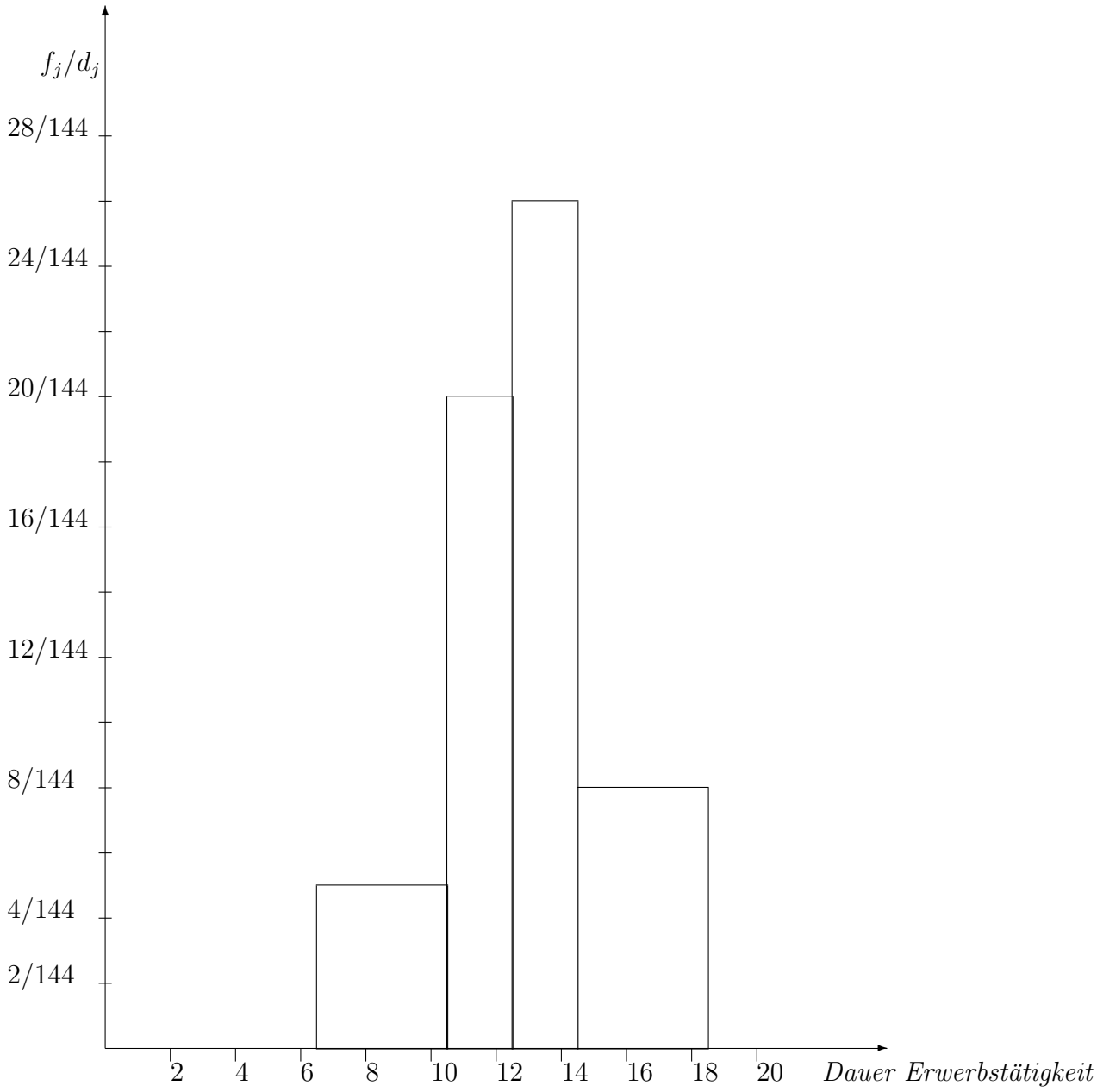
Beispiel 2.9 *Histogramm*

Histogramm: Darstellung ähnlich zu Säulendiagramm, aber:

- *Rechtecke stoßen direkt aneinander*
- *Rechtecke müssen nicht alle gleich breit sein, Breite entspricht Klassenbreite*
- *es gilt das Prinzip der* **Flächentreue**, *d.h. nicht die Höhe der Rechtecke ist proportional zur Häufigkeit, sondern die Fläche (= Höhe × Breite)!*

Absolventenstudie, Merkmal S = Studiendauer in Semestern
Merkmal liegt in vielen verschiedenen Ausprägungen vor → zunächst Zusammenfassung in Klassen, hier: 4 Klassen (willkürlich gewählt)
Zusammenstellung der absoluten und relativen Klassenhäufigkeiten:

Klasse j	h_j	f_j	*Klassenbreite* d_j	*Rechteckshöhe* f_j/d_j
$[6.5; 10.5)$	5	5/36	4	5/144
$[10.5; 12.5)$	10	10/36	2	20/144
$[12.5; 14.5)$	13	13/36	2	26/144
$[14.5; 18.5)$	8	8/36	4	8/144



Allgemein: Konstruktion eines Histogramms

- Ausprägungen in Klassen einteilen: Klassen stoßen ohne Lücke aneinander; Klassen umfassen insgesamt alle Ausprägungen
- relative Klassenhäufigkeiten bestimmen: zur j-ten Klasse bestimme f_j
- Klassenbreiten bestimmen: zur j-ten Klasse ist die Breite
  d_j = obere Klassengrenze - untere Klassengrenze

- im Koordinatensystem über jeder Klasse ein Rechteck zeichnen mit Rechteckshöhe der j-ten Klasse $= f_j/d_j$

Faustregeln:

- falls möglich, gleich breite Klassen wählen
- offene Randklassen vermeiden (immer endliche Begrenzung setzen)
- Klassenanzahl bei n Beobachtungen als $\sqrt{n}$ wählen

Verschiedene Klassenwahlen geben verschiedene optische Eindrücke (vgl. Übung)

2.1.3 Die empirische Verteilungsfunktion

Voraussetzungen: Merkmalsausprägungen lassen sich ordnen (also: Merkmal hat mindestens ordinales Niveau)
Oft von Interesse: Fragen wie z.B.

- Wieviel Prozent der rentenversicherungspflichtigen Arbeitnehmer verdienen zwischen 20 000 und 40 000 Euro im Jahr?
- Wie hoch ist der Anteil der Studierenden, die länger als 10 Semester studieren?

Definition 2.10 *Empirische Verteilungsfunktion*
Seien die Ausprägungen eines Merkmals X an n Objekten beobachtet, d.h. es liegen Beobachtungswerte $x_1, \ldots, x_n$ vor (mit x_i = Ausprägung von X am i-ten Objekt). Es sei $F(x)$ der Anteil der beobachteten Werte, die kleiner oder gleich x sind. Man nennt $F(x)$ die **empirische Verteilungsfunktion** *des Merkmals X.*
Formal:

$$F(x) = \sum_{j:a_j \leq x} f(a_j) = \sum_{j:a_j \leq x} f_j \text{ (kumulierte relative Häufigkeiten)}$$

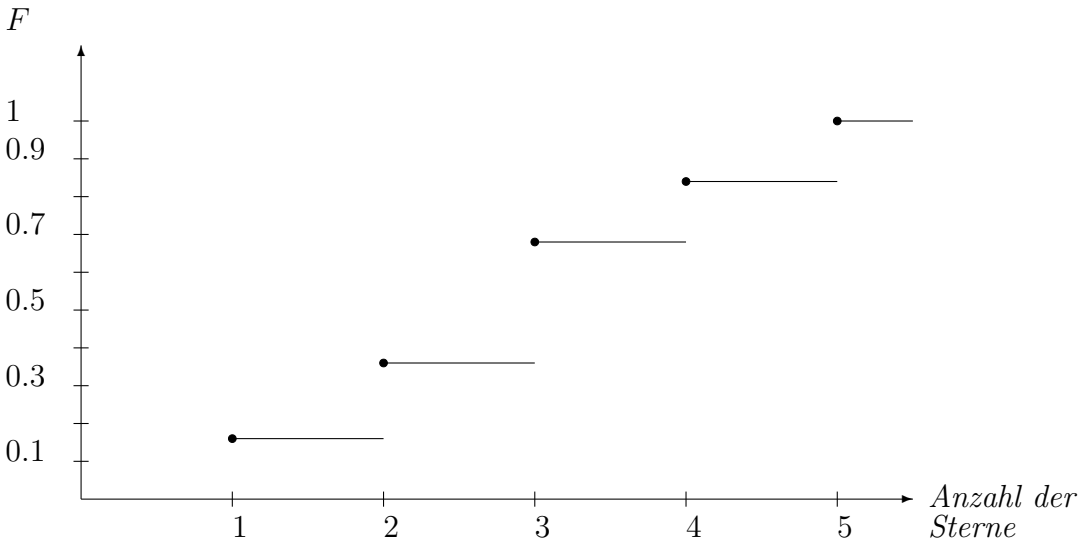
Eigenschaften:

- $F(x)$ ist monoton steigende Treppenfunktion
- $0 \leq F(x) \leq 1$
- $F(x)$ hat an den Stellen a_j Sprungstellen
- $F(x)$ = Anteil der Merkmalswerte, die x nicht übertreffen
- $1 - F(x)$ = Anteil der Merkmalswerte, die x übersteigen
- $F(x_o) - F(x_u)$ = Anteil der Merkmalswerte, die größer als x_u, aber höchstens so groß wie x_o sind

Beispiel 2.11 *Empirische Verteilungsfunktion*
Von 25 Hamburger Hotels wurde der Leistungsindex (= Anzahl der Sterne) erhoben: $3, 4, 5, 2, 2, 4, 3, 5, 2, 1, 4, 3, 3, 5, 3, 1, 3, 2, 3, 4, 2, 1, 1, 5, 3$

a_i	h_i	f_i	$F(a_i)$
1	4	$4/25 = 0.16$	0.16
2	5	0.2	0.36
3	8	0.32	0.68
4	4	0.16	0.84
5	4	0.16	1

$$\Rightarrow F(x) = \begin{cases} 0 & : \quad x < 1 \\ 0.16 & : \quad 1 \leq x < 2 \\ 0.36 & : \quad 2 \leq x < 3 \\ 0.68 & : \quad 3 \leq x < 4 \\ 0.84 & : \quad 4 \leq x < 5 \\ 1 & : \quad x \geq 5 \end{cases}$$



Mit Hilfe der empirischen Verteilungsfunktion kann man auch ausrechnen, wie groß der Anteil der Hotels mit mehr als zwei, aber höchstens vier Sternen ist:

Anteil Hotels mit höchstens vier Sternen $= F(4) = 0.84$

Anteil Hotels mit höchstens zwei Sternen $= F(2) = 0.36$

Damit: mehr als zwei und höchstens vier Sterne $= F(4) - F(2) = 0.84 - 0.36 = 0.48$

Auch für klassierte Daten kann die empirische Verteilungsfunktion bestimmt werden. Näheres hierzu wird in den Übungen besprochen.

2.2 Statistische Kenngrößen

Statistische Kenngrößen oder Maßzahlen: zur "zusammenfassenden" Beschreibung einer Verteilung

Es gibt u.a.

- Lagemaße (wo liegt Mehrzahl / Mitte / Schwerpunkt der beobachteten Merkmalswerte?)

- Streuungsmaße (über welchen Bereich erstrecken sich die Beobachtungen, wie stark schwanken sie?)
- Schiefemaße (wie ist das Symmetrieverhalten der Häufigkeitsverteilung?)
- Konzentrationsmaße (wie sind die Merkmalsausprägungen auf die Merkmalsträger verteilt?)

2.2.1 Lagemaße

Lagemaße beschreiben das Zentrum einer Verteilung → Beobachtungen werden zu einem Wert zusammengefasst → einfacher Vergleich verschiedener Beobachtungsreihen / Stichproben / Grundgesamtheiten anhand sogenannter "Mittelwerte"

Definition 2.12 *Arithmetisches Mittel*
Sei X ein kardinal skaliertes Merkmal, es seien $x_1, \ldots, x_n$ beobachtet. Das **arithmetische Mittel** $\overline{x}$ *der Beobachtungen $x_1, \ldots, x_n$ ist*

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

Liegt nur die unklassierte Häufigkeitsverteilung vor, das heißt, man kennt nur die Merkmalsausprägungen $a_1, \ldots, a_k$ und die zugehörigen relativen Häufigkeiten $f_1, \ldots, f_k$, so bestimmt man das arithmetische Mittel als

$$\overline{x} = \sum_{j=1}^{k} a_j \cdot f_j.$$

Beispiel 2.13 *Absolventenstudie*
Merkmal S = Studiendauer (→ kardinal skaliert)

$$\overline{x} = \frac{1}{36} \cdot \sum_{i=1}^{36} x_i = \frac{1}{36} \cdot \underbrace{464}_{\text{Summe aller Studiendauern}} = 12.89$$

oder auch:

$$\overline{x} = \underbrace{7}_{Semester} \cdot \underbrace{\frac{1}{36}}_{rel.\ Häufigkeit} +8 \cdot 0 + 9 \cdot \frac{1}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{3}{36} + 12 \cdot \frac{7}{36}$$
$$+13 \cdot \frac{9}{36} + 14 \cdot \frac{4}{36} + 15 \cdot \frac{4}{36} + 16 \cdot \frac{1}{36} + 17 \cdot \frac{2}{36} + 18 \cdot \frac{1}{36} = 12.89$$

Bemerkung 2.14 *Eigenschaften des arithmetischen Mittels*

(a) $\overline{x}$ *ist derjenige Wert, den jede Beobachtungseinheit annehmen würde, würde man die Gesamtsumme der Ausprägungen aller Beobachtungseinheiten gleichmäßig auf alle Einheiten verteilen.*

(b) Es gilt die sog. Schwerpunkteigenschaft oder Zentrierungseigenschaft:

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

d.h. die Summe der Abweichungen vom arithm. Mittel verschwindet.

(c) Sind neben $x_1, \ldots, x_n$ *weiterhin* $y_1, \ldots, y_n$ *beobachtet, so gilt für* $z_i = x_i + y_i$*:*

$$\overline{z} = \overline{x} + \overline{y}.$$

Etwa: z_i = *Bruttogewicht*, x_i = *Nettogewicht*, y_i = *Verpackungsgewicht*

} *damit* $z_i = x_i + y_i$, *und man kann das mittlere Bruttogewicht aus mittlerem Nettogewicht und mittlerem Verpackungsgewicht berechnen.*

Beispiel 2.15 *Zusammenfassung von Mittelwerten*

Eine Bausparkasse hat Marktforscher ausgeschickt, die bei zufällig ausgewählten, in Mietwohnungen lebenden Familien erfragen sollen, welchen Anteil ihres Einkommens sie für Miete ausgeben. Jeder Interviewer berichtet nur die Anzahl befragter Familien und den durchschnittlichen Anteil bei den von ihm befragten Familien.

Die Ergebnisse für drei Marktforscher:

Forscher 1	*Forscher 2*	*Forscher 3*
$n_1 = 4$	$n_2 = 4$	$n_3 = 8$
$\overline{x}_1 = 17.1(\%)$	$\overline{x}_2 = 17.125(\%)$	$\overline{x}_3 = 17.0625(\%)$

Die Bausparkasse möchte nun ermitteln, wie hoch der mittlere Anteil ist, wenn **alle** *befragten Familien gemeinsam betrachtet werden.*

Lösung:

$$
\begin{aligned}
\overline{x} &= \frac{1}{16} \cdot \sum_{i=1}^{16} x_i \\
&= \frac{1}{16} \cdot \left(\underbrace{\sum_{i=1}^{4} x_i}_{\textit{Forscher 1}} + \underbrace{\sum_{i=5}^{8} x_i}_{\textit{Forscher 2}} + \underbrace{\sum_{i=9}^{16} x_i}_{\textit{Forscher 3}} \right) \\
&= \frac{1}{16} \cdot \left(4 \cdot \underbrace{\frac{1}{4} \cdot \sum_{i=1}^{4} x_i}_{\overline{x}_1} + 4 \cdot \underbrace{\frac{1}{4} \cdot \sum_{i=5}^{8} x_i}_{\overline{x}_2} + 8 \cdot \underbrace{\frac{1}{8} \cdot \sum_{i=9}^{16} x_i}_{\overline{x}_3} \right) \\
&= \frac{1}{16} \cdot \left(\underbrace{4}_{n_1} \cdot \overline{x}_1 + \underbrace{4}_{n_2} \cdot \overline{x}_2 + \underbrace{8}_{n_3} \cdot \overline{x}_3 \right) \\
&= \frac{1}{16} \cdot (4 \cdot 17.1 + 4 \cdot 17.125 + 8 \cdot 17.0625) = 17.0875
\end{aligned}
$$

Beachte: dies ist nicht dasselbe wie $\frac{1}{3} \cdot (\overline{x}_1 + \overline{x}_2 + \overline{x}_3) = 17.0958$*!!!*

Dies führt zur folgenden Regel zur Zusammenfassung von Mittelwerten:

Bemerkung 2.16 *Mittelwerte aus Teilgesamtheiten*
Liegt ein Datensatz in r Teilgesamtheiten (sog. Schichten) vor und kennt man die Stichprobenumfänge n_j sowie die arithmetischen Mittel $\overline{x}_j$ pro Schicht, $j = 1, \ldots, r$, so lässt sich daraus das Gesamtmittel $\overline{x}$ berechnen als

$$\overline{x} = \frac{1}{n} \cdot \sum_{j=1}^{r} n_j \cdot \overline{x}_j$$

Dabei ist $n = \sum_{j=1}^{r} n_j$ der Gesamtstichprobenumfang.

Die Anwendung des arithmetischen Mittels ist nicht für alle Arten kardinaler Merkmale korrekt; bei Berechnung durchschnittlicher Wachstumsraten und durchschnittlicher relativer Änderungen braucht man stattdessen das sogenannte **geometrische Mittel**.

Definition 2.17 *Geometrisches Mittel*
Sei X ein kardinal skaliertes Merkmal mit nichtnegativen reellen Ausprägungen, es seien $x_1, ..., x_n$ beobachtet. Dann heißt

$$\overline{x}_{geom} = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

das **geometrische Mittel** *der x_i.*
Sind nicht alle Beobachtungswerte positiv (kommt z.B. bei Steigerungsraten vor), transformiert man die Beobachtungen bezüglich einer festen Basis (z.B. 100%), so dass man positive Werte erhält.

Beispiel 2.18 *Wachstumsraten*

Kurs X einer Aktie zu vier aufeinander folgenden Zeitpunkten, Beobachtungen x_1, x_2, x_3, x_4*:*

Zeitpunkt	t	*1*	*2*	*3*	*4*
Kurs	x_t	*10*	*12*	*15*	*10*
Wachtumsrate	r_t	/	*0.2*	*0.25*	-0.33
Wachstumsrate bezogen auf 100%	w_t	/	*1.2*	*1.25*	*0.67*

Dabei: Wachstumsrate r_t *= prozentuale Änderung des Werts von* x_{t-1} *auf* x_t

$$= \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{\textit{Kursendwert} \; - \; \textit{Kursanfangswert}}{\textit{Kursanfangswert}},$$

Wachstumsfaktor $w_t = 1 + r_t = \frac{x_t}{x_{t-1}}$.

Das heißt: $x_t = (1 + r_t) \cdot x_{t-1} = w_t \cdot x_{t-1}$.

Damit: durchschnittlicher Wachstumsfaktor vom Zeitpunkt 1 bis zum Zeitpunkt 4:

$$\overline{w}_{geom} = \sqrt[3]{1.2 \cdot 1.25 \cdot 0.67} \approx 1$$

(nicht exakt 1 wg. Rundung),

und durchschnittliche Wachstumsrate ist $\overline{w}_{geom} - 1 = 0.00$ *(dies ist auch sinnvoll, da Kurs zum Zeitpunkt 4 genauso hoch wie zum Zeitpunkt 1!)*

Ein Lagemaß, das sich bereits bei ordinal skalierten Merkmalen ermitteln lässt, ist der **Median**. Zu seiner Bestimmung müssen die Beobachtungen $x_1, \ldots, x_n$ zunächst der Größe nach geordnet werden.

Definition 2.19 *Median*

Sei $x_1, \ldots, x_n$ die Urliste der Beobachtungen des Merkmals X. Die der Größe nach geordneten Werte $x_{(1)} \leq \ldots \leq x_{(n)}$ heißen **Ordnungsstatistiken**.

Der **Median** *ist dann "die mittlere Ordnungsstatistik" bzw. derjenige Wert, der die geordnete Reihe der Beobachtungswerte in zwei gleiche Teile zerlegt. Man berechnet den Median als*

$$x_{med} = x_{\left(\frac{n+1}{2}\right)}, \quad \textit{falls } n \textit{ ungerade}$$

$$x_{med} = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right), \quad \textit{falls } n \textit{ gerade}$$

Beispiel 2.20 *Median*

Anzahl kariöser Zähne bei 20 Schulkindern

Urliste: 1, 0, 5, 0, 3, 1, 1, 4, 2, 7, 0, 0, 2, 1, 2, 1, 2, 6, 4, 1

$\rightarrow$ *Ordnungsstatistiken:*

0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 5, 6, 7

Hier: $n = 20$ *gerade* $\Rightarrow x_{med} = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) = \frac{1}{2} \cdot \left(x_{(10)} + x_{(11)} \right)$

$x_{(10)} = 1$, $x_{(11)} = 2 \Rightarrow x_{med} = \frac{1}{2} \cdot (1 + 2) = 1.5$

Zum Vergleich: es ist $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 2.12$

Beachte: In verschiedenen Lehrbüchern findet man auch abweichende Definitionen für den Median. Da der Median zunächst nur die "Mitte" der geordneten Beobachtungen wiedergeben soll, kommt für eine gerade Anzahl an Beobachtungen prinzipiell jeder Wert von $x_{\left(\frac{n}{2}\right)}$ bis $x_{\left(\frac{n}{2}+1\right)}$ als Median in Frage.

Der Median ist relativ stabil gegenüber einzelnen extrem kleinen oder extrem großen Beobachtungswerten, während das arithmetische Mittel $\overline{x}$ darauf stark reagieren kann: $\overline{x}$ ist empfindlich gegenüber sogenannten "Ausreißern".

Ein weiteres, sehr einfaches Lagemaß ist der **Modus** oder **Modalwert**. Dieser ist bereits sinnvoll für Beobachtungen auf Nominalskalenniveau.

Definition 2.21 *Modus, Modalitätsgrad*

Der **Modus** x_{mod} *ist definiert als die Ausprägung mit der größten relativen Häufigkeit. Besitzen mehrere Ausprägungen die gleiche größte Häufigkeit, so ist der Modus nicht bestimmbar.*

Der **Modalitätsgrad** *gibt einen Hinweis darauf, wie typisch der Modus für eine Häufigkeitsverteilung ist. Er entspricht der relativen Häufigkeit des Modus und wird in % angegeben: rel. Häufigkeit des Modus* $\cdot 100\%$.

Beispiel 2.22 *Modus*

Absolventenstudie

Merkmal Studiendauer: $x_{mod} = 13$, *Modalitätsgrad:* $\frac{9}{36} \cdot 100\% = 25\%$

Merkmal Geschlecht: x_{mod} =*weibl., Modalitätsgrad:* $\frac{26}{36} \cdot 100\% = 72.2\%$

Merkmal Diplomnote: $x_{mod} = 2$, *Modalitätsgrad:* $\frac{22}{36} \cdot 100\% = 61.1\%$

Bemerkung 2.23 *Lagemaße*

- x_{med} *und* $\overline{x}$ *können bei diskreten Merkmalen Werte annehmen, die mit keiner möglichen Ausprägung übereinstimmen (z.B.* 1.3 *Kinder pro Familie* $\rightarrow$ *auch für metrische Merkmale kann u.U. Modus sinnvoller sein)*

- *Manchmal: Beobachtungen werden transformiert; Frage: müssen Lagemaße neu berechnet werden?*
  Beispiel lineare Transformation: Beobachtungen $x_1, \ldots, x_n$*; betrachte* $y_i = a \cdot x_i + b, i = 1, \ldots, n$ *(dabei* a, b *Konstanten). Dann ist* $\overline{y} = a \cdot \overline{x} + b$*, d.h. das arithmetische Mittel macht diese Transformation mit. Für* $a \neq 0$ *werden auch Modus und Median entsprechend transformiert. Dies gilt aber nicht für jede Art von Transformation!*

Bemerkung 2.24 *Klassierte Daten*

Liegt von den Daten nur die klassierte Häufigkeitsverteilung vor (nur Klassengrenzen und Anzahl von Beobachtungen je Klasse bekannt), lassen sich die Lagemaße nicht mehr exakt bestimmen.

- *Für das arithmetische Mittel rechnet man als Näherung*

$$\overline{x}_{klass} = \sum_{j=1}^{k} m_j \cdot f_j$$

  Dabei bezeichnet k die Anzahl der Klassen, m_j die Klassenmitte und f_j die relative Häufigkeit der j-ten Klasse.

- *Für Modus und Median gibt es entsprechende Näherungsformeln. Näheres hierzu wird in den Übungen besprochen.*

Beispiel 2.25 *Arithmetisches Mittel bei klassierten Daten*

Absolventenstudie, klassierte Studiendauern

Klasse j	$[6.5, 10.5)$	$[10.5, 12.5)$	$[12.5, 14.5)$	$[14.5, 18.5)$
f_j	$5/36$	$10/36$	$13/36$	$8/36$
m_j	8.5	11.5	13.5	16.5

Damit

$$\begin{aligned} \overline{x}_{klass} &= \sum_{j=1}^{k} m_j \cdot f_j \\ &= 8.5 \cdot \frac{5}{36} + 11.5 \cdot \frac{10}{36} + 13.5 \cdot \frac{13}{36} + 16.5 \cdot \frac{8}{36} = 12.92 \end{aligned}$$

Zum Vergleich: es war $\overline{x} = 12.89$ (s. Bsp. 2.13) $\rightarrow$ man sieht, dass aus den klassierten Daten nur Näherung resultiert.

Durch den Median wird die geordnete Reihe der Beobachtungen in zwei gleich große Teile zerlegt. Entsprechend gibt es Punkte, die die Reihe in anderer Weise aufteilen (z.B. in zwei Teile mit 25% bzw. 75% der geordneten Beobachtungen) → sog. **p-Quantile**.

Definition 2.26 *Quantile*
Ein **p-Quantil** *teilt den Datensatz so in zwei Teile, dass ein Anteil p der Daten darunter und ein Anteil $1-p$ darüber liegt. Dabei ist $0 < p < 1$. Zur Berechnung eines p-Quantils unterscheidet man zwei Fälle:*

1. *$n \cdot p$ ist nicht ganzzahlig; dann ist das p-Quantil zu bestimmen als*

$$x_p = x_{([n \cdot p]+1)},$$

   wobei n die Anzahl der Beobachtungen ist und $[n \cdot p]$ die zu $n \cdot p$ nächstkleinere ganze Zahl.

2. *$n \cdot p$ ist ganzzahlig; dann ist das p-Quantil zu bestimmen als*

$$x_p = \frac{1}{2} \cdot \left(x_{(n \cdot p)} + x_{(n \cdot p+1)} \right).$$

Ähnlich wie beim Median, gibt es auch für die Quantile verschiedene Varianten der Definition. Beim Lesen in Lehrbüchern und bei der Verwendung statistischer Programmpakete ist daher darauf zu achten, mit welcher Definition gearbeitet wird, um Fehlinterpretationen zu vermeiden!

Beispiel 2.27 *Absolventenstudie*
Merkmal Studiendauer; $n = 36$

- 0.1-*Quantil:* $p = 0.1 \Rightarrow n \cdot p = 36 \cdot 0.1 = 3.6$ *nicht ganzzahlig*
  $\Rightarrow [n \cdot p] = 3$
  $\Rightarrow x_{0.1} = x_{([n \cdot p]+1)} = x_{(3+1)} = x_{(4)} = 10$

- 0.25-*Quantil:* $p = 0.25 \Rightarrow n \cdot p = 36 \cdot 0.25 = 9$ *ganzzahlig*
  $\Rightarrow x_{0.25} = \frac{1}{2} \cdot \left(x_{(9)} + x_{(10)}\right) = \frac{1}{2} \cdot (12 + 12) = 12$

Es gibt spezielle p-Quantile, die besonders häufig gebraucht werden. Dies sind

$x_{0.25}$, das sog. **untere Quartil**

$x_{0.75}$, das sog. **obere Quartil** und

$x_{0.5} = x_{med}$, der **Median** (für den Median gilt ja: 50% = 0.5 der Beobachtungen sind $\leq x_{med}$, 50% = 0.5 der Beobachtungen sind $\geq x_{med}$, d.h. Median = 0.5-Quantil)

Eine Zusammenfassung der Häufigkeitsverteilung eines Merkmals ist durch Angabe einiger weniger "Zerlegungspunkte" möglich.

Definition 2.28 *Fünf-Punkte-Zusammenfassung*
Die **Fünf-Punkte-Zusammenfassung** *für die Häufigkeitsverteilung eines quantitativen Merkmals teilt den Wertebereich in 4 Intervalle, die jeweils ca. ein Viertel der Beobachtungswerte enthalten.*
Sie ist gegeben durch

$x_{(1)}$	*kleinster Wert*
$x_{0.25}$	*unteres Quartil*
x_{med}	*Median*
$x_{0.75}$	*oberes Quartil*
$x_{(n)}$	*größter Wert*

Zur Fünf-Punkte-Zusammenfassung gibt es eine passende graphische Darstellungsart, den so genannten Boxplot (siehe Übung).

Bisher: Reduktion der Information aus einem Datensatz auf einige Lagemaße

→ gut für ersten Eindruck

→ aber: keine Aussage darüber, ob die beobachteten Merkmalswerte dicht bei den Lagemaßen liegen oder stark streuen

z.B.:
Datensatz 1: $7, 7, 8, 8, 8, 9, 9 \quad \Rightarrow \overline{x} = x_{mod} = x_{med} = 8$
Datensatz 2: $1, 1, 8, 8, 8, 15, 15 \quad \Rightarrow \overline{x} = x_{mod} = x_{med} = 8$

Frage: wie charakterisiert man den Unterschied zwischen den Datensätzen?

2.2.2 Streuungsmaße

Streuungsmaße geben Auskunft darüber, wie nahe die Beobachtungen zusammen liegen bzw. wie stark sie variieren. Ein Streuungsmaß hat prinzipiell einen umso kleineren Wert, je dichter die Beobachtungen beieinander liegen. Sind alle Beobachtungen gleich, so haben alle Streuungsmaße den Wert Null.

Definition 2.29 *Spannweite, Interquartilsabstand*
Sei X ein mindestens ordinal skaliertes Merkmal, es seien $x_1, \ldots, x_n$ beobachtet. Die **Spannweite** *oder der* **Range** *der Beobachtungen ist definiert als der Unterschied zwischen der größten und der kleinsten Beobachtung:*

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

Der **Interquartilsabstand** *ist die Differenz aus den beiden Quartilen:*

$$d_Q = x_{0.75} - x_{0.25}.$$

Die Spannweite ist extrem empfindlich gegenüber Ausreißern. Der Interquartilsabstand ist "robuster"; allerdings kann es passieren, dass d_Q den Wert Null annimmt, auch wenn die Beobachtungen nicht alle gleich sind.

Beispiel 2.30 *Spannweite, Interquartilsabstand*

Erhoben wurde das monatliche Nettogehalt X von 5 leitenden Angestellten:

3 180, 3 400, 3 660, 3 920, 5 140

Damit ergibt sich die Spannweite hier als $R = 5\,140 - 3\,180 = 1\,960$.

Für den Interquartilsabstand: zunächst die Quartile bestimmen

es ist $x_{0.25} = x_{(2)} = 3\,400$ *(da* $p = 0.25$, $n \cdot p = 5 \cdot 0.25 = 1.25$ *nicht ganzzahlig* $\Rightarrow x_p = x_{([n \cdot p]+1)}$*) und* $x_{0.75} = x_{(4)} = 3\,920$

damit: $d_Q = x_{0.75} - x_{0.25} = 3\,920 - 3\,400 = 520$

Spannweite und Interquartilsabstand basieren nur auf den Ordnungsstatistiken. Eine andere Idee für die Definition eines Streuungsmaßes ist, zu ermitteln, wie stark die Beobachtungen "im Mittel" um ein Lagemaß streuen.

Definition 2.31 *MAD*

Sei X mindestens ordinal skaliert, seien $x_1, \ldots, x_n$ beobachtet. Ein Maß für die Abweichung der Beobachtungen von ihrem Median ist die **mediane absolute Abweichung vom Median***, kurz* **MAD***:*

$$MAD = med\{|x_i - x_{med}|, i = 1, \ldots, n\}$$

Berechnung des MAD:

- bestimme aus den Beobachtungen $x_1, \ldots, x_n$ zunächst den Median x_{med}
- berechne die Hilfswerte $y_1 = |x_1 - x_{med}|, \ldots, y_n = |x_n - x_{med}|$
- bestimme jetzt den Median der $y_i, i = 1, \ldots, n$: $y_{med} = MAD$

Beispiel 2.32 *MAD*

Daten zum monatlichen Nettogehalt von 5 leitenden Angestellten (Bsp. 2.30)

Beobachtet: 3 180, 3 400, 3 660, 3 920, 5 140 $\Rightarrow x_{med} = 3\,660$

Arbeitstabelle: benötigte Größen zur Bestimmung des MAD

x_i	$x_i - x_{med}$	$y_i = \lvert x_i - x_{med} \rvert$
3180	−480	480
3 400	−260	260
3 660	0	0
3 920	260	260
5 140	1 480	1 480

Also ist zur Bestimmung des MAD der Median zu bilden aus $y_1, \ldots, y_n$*, d.h. von* 480, 260, 0, 260, 1 480

$\rightarrow$ *sortieren:* 0, 260, 260, 480, 1 480

$\Rightarrow MAD = 260$

Ähnlich zum MAD gibt es ein Streuungsmaß, das feststellt, wie stark die Beobachtungen um das arithmetische Mittel streuen. Dies ist das bekannteste Streuungsmaß, die sogenannte **(empirische) Varianz**.

Definition 2.33 *Varianz, Standardabweichung*

Sei X *ein kardinal skaliertes Merkmal, seien* $x_1, \ldots, x_n$ *beobachtet. Die Größe*

$$\widetilde{s}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

heißt die **(empirische) Varianz** *des Merkmals* X*.*

Die Größe $\widetilde{s} = \sqrt{\widetilde{s}^2}$ *heißt* **Standardabweichung** *von* X*.*

Liegt nur die unklassierte Häufigkeitsverteilung vor, man kennt also nur die Merkmalsausprägungen $a_1, \ldots, a_k$ und die zugehörigen relativen Häufigkeiten $f_1, \ldots, f_k$, so bestimmt man die Varianz als

$$\widetilde{s}^2 = \sum_{j=1}^{k} f_j \cdot (a_j - \overline{x})^2.$$

Beispiel 2.34 *Varianz, Standardabweichung*

Monatliches Nettogehalt von 5 leitenden Angestellten (vgl. 2.30)

Zunächst Berechnung des arithmetischen Mittels: es ist $\overline{x} = \frac{1}{5} \cdot 19\,300 = 3\,860$

Arbeitstabelle: benötigte Größen zur Bestimmung der Varianz

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
3 180	−680	462 400
3 400	−460	211 600
3 660	−200	40 000
3 920	60	3 600
5 140	1 280	1 638 400

Damit ist $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 462\,400 + 211\,600 + 40\,000 + 3\,600 + 1\,638\,400 = 2\,356\,000$ *und* $\widetilde{s}^2 = \frac{1}{5} \cdot 2\,356\,000 = 471\,200$*;*

weiterhin ist $\widetilde{s} = \sqrt{471\,200} = 686.44$

Bemerkung 2.35 *Verschiebungssatz*

- *Zur vereinfachten Berechnung der Varianz nutzt man den so genannten* **Verschiebungssatz***:*

$$\widetilde{s}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{x}^2.$$

- *In der induktiven Statistik verwendet man statt $\widetilde{s}^2$ die so genannte* **Stichprobenvarianz** s^2 *mit*

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Begründung: geänderter Vorfaktor $\frac{1}{n-1}$ bewirkt, dass s^2 günstigere statistische Eigenschaften besitzt als $\widetilde{s}^2$ (später genauer). Außerdem: falls n groß, unterscheiden sich $\widetilde{s}^2$ und s^2 kaum.

Auch für s^2 gibt es eine Variante des Verschiebungssatzes:

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2 \right).$$

Beispiel 2.36 *Verschiebungssatz*

Gehälter von 5 leitenden Angestellten: Berechnung der Varianz mit Hilfe des Verschiebungssatzes

Es war $\overline{x} = 3\,860$, und die Gehälter waren $x_1 = 3\,180$, $x_2 = 3\,400$, $x_3 = 3\,660$, $x_4 = 3\,920$, $x_5 = 5\,140$.

Damit:

$$\widetilde{s}^2 = \tfrac{1}{5} \cdot \underbrace{(3\,180^2 + 3\,400^2 + 3\,660^2 + 3\,920^2 + 5\,140^2)}_{\sum_{i=1}^{n} x_i^2} - \underbrace{3\,860^2}_{\overline{x}^2}$$

$$= \tfrac{1}{5} \cdot 76\,854\,000 - 14\,899\,600 = 471\,200$$

Bemerkung 2.37 *Varianz aus Teilgesamtheiten*

Ähnlich wie beim arithmetischen Mittel kann man auch die Varianz eines gesamten Datensatzes berechnen, wenn Aussagen über die Varianzen von Teilgesamtheiten vorliegen.

Liegt ein Datensatz in r Schichten vor und kennt man die Stichprobenumfänge n_j sowie die arithmetischen Mittel $\overline{x}_j$ und Varianzen $\widetilde{s}_j^2$ pro Schicht, $j = 1, \ldots, r$, so lässt sich daraus die Gesamtvarianz $\widetilde{s}^2$ berechnen als

$$\widetilde{s}^2 = \frac{1}{n} \cdot \Bigg(\underbrace{\sum_{j=1}^{r} n_j \cdot \widetilde{s}_j^2}_{\text{Variabilität innerhalb der Schichten}} + \underbrace{\sum_{j=1}^{r} n_j \cdot (\overline{x}_j - \overline{x})^2}_{\text{Variabilität zwischen den Schichten}} \Bigg)$$

Beispiel 2.38 *Zusammenfassung von Varianzen*

Befragung für die Bausparkasse

Die drei Marktforscher liefern insgesamt folgende Daten:

(aus Bsp. 2.15: $\overline{x} = 17.0875$)

	Forscher 1	*Forscher 2*	*Forscher 3*
n_j	4	4	8
$\overline{x}_j$	17.1	17.125	17.0625
$\widetilde{s}_j^2$	0.005	0.0119	0.0098

Damit:

$$\begin{aligned}
\widetilde{s}^2 &= \frac{1}{n} \cdot \left(\sum_{j=1}^{r} n_j \cdot \widetilde{s}_j^2 + \sum_{j=1}^{r} n_j \cdot (\overline{x}_j - \overline{x})^2 \right) \\
&= \frac{1}{16} \cdot \Bigg(\underbrace{4 \cdot 0.005 + 4 \cdot 0.0119 + 8 \cdot 0.0098}_{\sum_{j=1}^{r} n_j \cdot \widetilde{s}_j^2} + \underbrace{4 \cdot (17.1 - 17.0875)^2}_{n_1 \cdot (\overline{x}_1 - \overline{x})^2} \\
&\quad + \underbrace{4 \cdot (17.125 - 17.0875)^2}_{n_2 \cdot (\overline{x}_2 - \overline{x})^2} + \underbrace{8 \cdot (17.0625 - 17.0875)^2}_{n_3 \cdot (\overline{x}_3 - \overline{x})^2} \Bigg) \\
&= \frac{1}{16} \cdot \Big(\underbrace{0.146}_{\sum_{j=1}^{r} n_j \cdot \widetilde{s}_j^2} + \underbrace{0.01125}_{\sum_{j=1}^{r} n_j \cdot (\overline{x}_j - \overline{x})^2} \Big) \\
&= \frac{1}{16} \cdot 0.15725 = 0.0098
\end{aligned}$$

Bemerkung 2.39 *Eigenschaften der Varianz*

- *Werden Beobachtungen linear transformiert, d.h.* $x_i \mapsto y_i = a \cdot x_i + b$, $i = 1, \ldots, n$, *dann gilt:*

$$\widetilde{s}_y^2 = a^2 \cdot \widetilde{s}_x^2, \text{ und } \widetilde{s}_y = |a| \cdot \widetilde{s}_x.$$

- *Eine spezielle lineare Transformation ist die sog.* **Standardisierung**. *Beobachtungswerte* $x_1, \ldots, x_n$ *werden standardisiert, indem man* $\overline{x}$ *abzieht und das Ergebnis durch* $\widetilde{s}_x$ *teilt:*

$$z_i = \frac{x_i - \overline{x}}{\widetilde{s}_x}.$$

Dies ist eine lineare Transformation $x_i \mapsto z_i = \underbrace{\frac{1}{\widetilde{s}_x}}_{a} \cdot x_i \underbrace{- \frac{\overline{x}}{\widetilde{s}_x}}_{b}$.

Damit gilt für die z_i*:*

$$\overline{z} = a \cdot \overline{x} + b = \frac{1}{\widetilde{s}_x} \cdot \overline{x} + (-\frac{\overline{x}}{\widetilde{s}_x}) = 0$$

und

$$\widetilde{s}_z^2 = a^2 \cdot \widetilde{s}_x^2 = \frac{1}{\widetilde{s}_x^2} \cdot \widetilde{s}_x^2 = 1,$$

d.h. standardisierte Werte haben Mittelwert 0 und Varianz 1.

- *Standardabweichung und arithmetisches Mittel werden oft in Kombination benutzt, um* **Schwankungsbereiche** *anzugeben: ist die Variable X normalverteilt (später genauer; als Faustregel: wenn n sehr groß ist), dann liegen in*

$$\overline{x} \pm \widetilde{s} \quad \textit{ca. 68\% aller Beobachtungen}$$

$$\overline{x} \pm 2 \cdot \widetilde{s} \quad \textit{ca. 95\% aller Beobachtungen}$$

$$\overline{x} \pm 3 \cdot \widetilde{s} \quad \textit{ca. 99\% aller Beobachtungen}$$

Will man die Streuungen verschiedener Häufigkeitsverteilungen vergleichen, ist es sinnvoll, diese jeweils in Relation zum Mittelwert zu betrachten
→ Größenordnung der Daten wird mit einbezogen
→ führt zur Definition des **Variationskoeffizienten**

Definition 2.40 *Variationskoeffizient*
Sei $\overline{x} > 0$, dann ist der **Variationskoeffizient** *v definiert als*

$$v = \frac{\widetilde{s}}{\overline{x}}.$$

Der Variationskoeffizient ist maßstabsunabhängig und eignet sich damit insbesondere zum Vergleich zweier Datensätze, die prinzipiell dasselbe Merkmal messen, nur in unterschiedlichen Einheiten, z.B. pro-Kopf-Ausgaben für Tee in Pfund (GB) und Euro (Deutschl.).

Beispiel 2.41 *Mobilfunkunternehmen*
In einer Befragung ermittelt ein Mobilfunkunternehmen bei 200 seiner Kunden, wie hoch der Betrag der monatlichen Telefonrechnung X (in Euro) ist und wie viele Handys Y der Kunde gleichzeitig nutzt. Man erhielt folgende Tabelle:

	Anzahl der Handys Y				
Rechnungsbetrag X	1	2	3	4 *und mehr*	$\sum$
bis 20	14	10	5	1	30
über 20 *bis* 40	20	8	5	2	35
über 40 *bis* 80	22	24	12	7	65
über 80	16	26	16	12	70
$\sum$	72	68	38	22	200

Das Mobilfunkunternehmen hält einen maximalen Rechnungsbetrag von 200 Euro für möglich.

Gesucht: ein "mittlerer" Rechnungsbetrag für alle 200 Kunden
Problem: offene Randklasse bei Merkmal X (vergleiche Übung)
Da die anderen Klassen nicht alle gleich breit und kein Repräsentant für die letzte Klasse genannt → Berechnung von Median oder Modus möglich
laut Angabe wird die Randklasse durch einen maximalen Rechnungsbetrag begrenzt (Unternehmen geht von max. 200 Euro aus) → auch die Berechnung des arithmetischen Mittels zulässig

Betrachtet man nun die Gruppe derjenigen Kunden, die nur 1 Handy nutzen (Y=1), kann man als **mittleren Rechnungsbetrag** *aller Kunden in dieser Gruppe das arithmetische Mittel* $\overline{x}$ *und als* **Rechnungsbetrag des mittleren Kunden** *in dieser Gruppe den Median* x_{med} *bestimmen.*
Dazu:

j	X *in Euro*	n_j	f_j	m_j	$m_j \cdot f_j$	$m_j \cdot n_j$	*kum* f_j
1	*bis* 20	14	0.194	10	1.94	140	0.194
2	*über* 20 *bis* 40	20	0.278	30	8.34	600	0.472
3	*über* 40 *bis* 80	22	0.305	60	18.30	1 320	0.777
4	*über* 80 *bis* 200	16	0.222	140	31.08	2 240	1 *(Rundung!)*
	$\sum$	72	1		59.66	4 300	

Dann ist
$\overline{x} = \sum_{j=1}^{k} m_j \cdot f_j = 59.66$ *Euro*

Alternativ auch möglich: $\overline{x} = \frac{1}{n} \cdot \sum_{j=1}^{k} m_j \cdot n_j = \frac{1}{72} \cdot 4\,300 = 59.72$ *Euro*

Beachte: der Unterschied kommt durch die Rundung in den f_j *zustande!*

Außerdem:
$x_{med} = x_{0.5} = x_j^u + (0.5 - kum\, f_{j-1}) \cdot \frac{d_j}{f_j}$ *(vergleiche Übung)*
$kum\, f_j = 0.5$ *wird erstmals in Klasse* $j = 3$ *überschritten*
$\Rightarrow x_{med} = x_{0.5} = 40 + (0.5 - 0.472) \cdot \frac{40}{0.305} = 43.67$ *Euro*

Für die 4 Gruppen der Handynutzer liegen nun folgende Daten über den Rechnungsbetrag vor:

	Anzahl der Handys Y			
	1	2	3	4 *und mehr*
j	1	2	3	4
$\overline{x}_j$	59.72	79.71	83.16	98.64
$\widetilde{s}_j^2$	2 158.26	2 532.27	2 605.82	2 193.60
n_j	72	68	38	22
f_j	0.36	0.34	0.19	0.11

Es interessiert der Mittelwert des Rechnungsbetrags über alle 4 Gruppen sowie die Gesamtstreuung.

Nach den Rechenregeln zur Zusammenfassung von Mittelwerten (siehe 2.16) und Varianzen (siehe 2.37) ist

$$\begin{aligned}
\overline{x}_{Gesamt} &= \frac{1}{n} \cdot \sum_{j=1}^{k} n_j \cdot \overline{x}_j \\
&= \frac{1}{200} \cdot (72 \cdot 59.72 + 68 \cdot 79.71 + 38 \cdot 83.16 + 22 \cdot 98.64) \\
&= \frac{1}{200} \cdot 15\,050.28 = 75.25 \ \textit{Euro}
\end{aligned}$$

und

$$\widetilde{s}^2_{Gesamt} = \frac{1}{n} \cdot \left(\sum_{j=1}^{k} \widetilde{s}_j^2 \cdot n_j + \sum_{j=1}^{k} (\overline{x}_j - \overline{x})^2 \cdot n_j \right) = \widetilde{s}^2_{int} + \widetilde{s}^2_{ext}$$

mit

$$
\begin{aligned}
\widetilde{s}^2_{int} &= \frac{1}{n} \cdot \sum_{j=1}^{k} \widetilde{s}^2_j \cdot n_j \\
&= \frac{1}{200} \cdot (2\,158.26 \cdot 72 + 2\,532.27 \cdot 68 + 2\,605.82 \cdot 38 + 2\,193.60 \cdot 22) \\
&= \frac{1}{200} \cdot (155\,394.72 + 172\,194.36 + 99\,021.16 + 48\,259.20) \\
&= \frac{1}{200} \cdot 474\,869.44 = 2\,374.35 \\
\widetilde{s}^2_{ext} &= \frac{1}{n} \cdot \sum_{j=1}^{k} (\overline{x}_j - \overline{x})^2 \cdot n_j \\
&= \frac{1}{200} \cdot [(59.72 - 75.25)^2 \cdot 72 + (79.71 - 75.25)^2 \cdot 68 + (83.16 - 75.25)^2 \cdot 38 \\
&\quad + (98.64 - 75.25)^2 \cdot 22] \\
&= \frac{1}{200} \cdot [17\,365.03 + 1\,352.63 + 2\,377.59 + 12\,036.03] = \frac{1}{200} \cdot 33\,131.27 = 165.66
\end{aligned}
$$

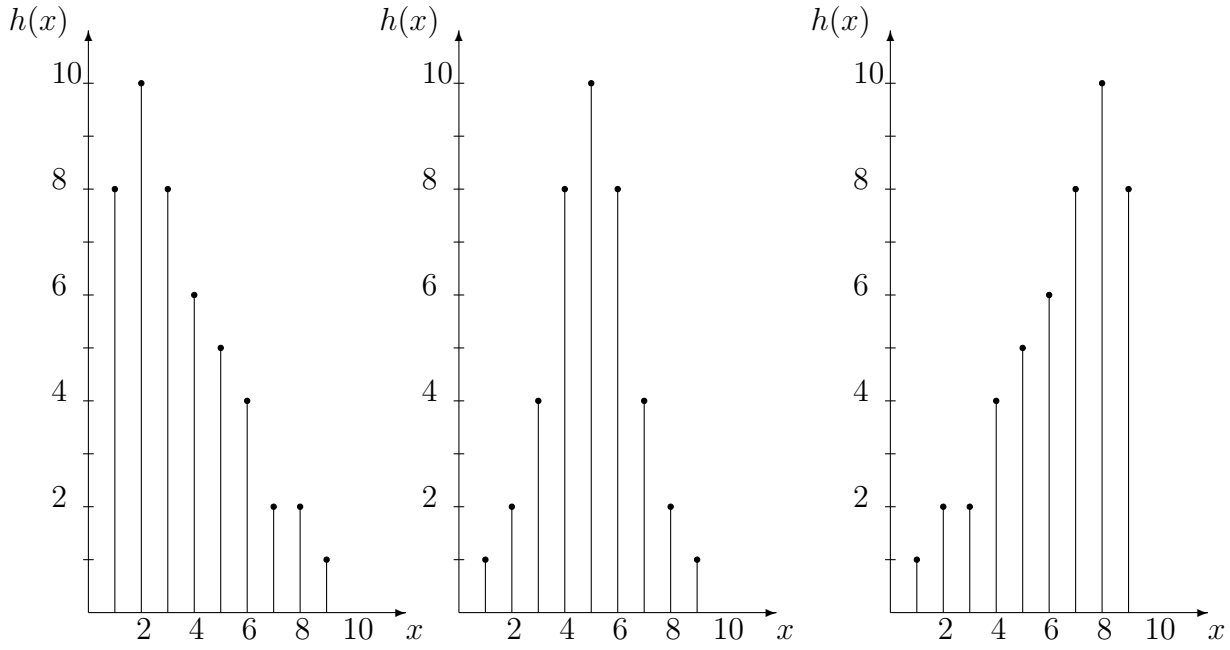
Damit

$$
\widetilde{s}^2_{Gesamt} = \widetilde{s}^2_{int} + \widetilde{s}^2_{ext} = 2\,374.35 + 165.66 = 2\,540.01
$$

2.2.3 Schiefemaße

Verteilungen unterscheiden sich nicht nur durch ihre Lage und ihre Streuung, sondern auch durch ihr Symmetrieverhalten.

Beispiel 2.42 *Symmetrie von Verteilungen*



Verteilung 1 *Verteilung 2* *Verteilung 3*

Verteilung 1: rechtsschief; es ist $\overline{x} = 3.57$, $x_{med} = 3$, $x_{mod} = 2$

$\rightarrow \overline{x} > x_{med} > x_{mod}$

Verteilung 2: symmetrisch; es ist $\overline{x} = 5$, $x_{med} = 5$, $x_{mod} = 5$

$\rightarrow \overline{x} = x_{med} = x_{mod}$

Verteilung 3: linksschief; es ist $\overline{x} = 6.43$, $x_{med} = 7$, $x_{mod} = 8$

$\rightarrow \overline{x} < x_{med} < x_{mod}$

Offensichtlich lässt sich die Schiefe einer Verteilung anhand der drei Lagemaße beurteilen.

Bemerkung 2.43 *Lageregeln*

Falls

- $\overline{x} = x_{med} = x_{mod} \rightarrow$ *Verteilung ist symmetrisch*
- $\overline{x} > x_{med} > x_{mod} \rightarrow$ *Verteilung ist rechtsschief*
- $\overline{x} < x_{med} < x_{mod} \rightarrow$ *Verteilung ist linksschief*

Eine erste Beurteilung der Schiefe kann also graphisch und an Hand der Lageregeln erfolgen. Zusätzlich gibt es Maßzahlen für die Schiefe.

Definition 2.44 *Schiefekoeffizient nach Pearson*

Eine Maßzahl für die Schiefe einer Verteilung ist der **Schiefekoeffizient nach Pearson** *(auch* **Momentenkoeffizient** *genannt):*

$$g_m = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^3}{\sqrt{\left(\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2\right)^3}}$$

Es gilt:

- $g_m = 0 \rightarrow$ symmetrisch
- $g_m > 0 \rightarrow$ rechtsschief
- $g_m < 0 \rightarrow$ linksschief

Beispiel 2.45 *Verteilungen 1 - 3*

In Beispiel 2.42 gilt für

Verteilung 1: $g_m = 0.72$

Verteilung 2: $g_m = 0$

Verteilung 3: $g_m = -0.72$

2.2.4 Konzentrationsmaße

Bisher: Verteilung der Beobachtungen auf den Wertebereich betrachtet.
Jetzt: betrachte die Verteilung der Merkmalssumme auf die Merkmalsträger.
Etwa: Wie teilt sich ein Markt auf die Anbieter auf (Marktkonzentration)? Konzentriert sich der Gesamtumsatz auf einige wenige Firmen, oder haben alle Firmen ungefähr gleichmäßig Teil daran?

Beispiel 2.46 *Interessierende Merkmale*

Typische Merkmale, für die Konzentrationen bestimmt werden, sind

Merkmal	*Merkmalsträger*
Jahreseinkommen	*Erwerbstätige*
Beschäftigtenzahl	*Industriebetriebe*
Jahresumsatz	*Betriebe*
Einwohnerzahl	*Gemeinden*
Jahresumsatz von Artikeln eines Sortiments	*Artikel*

Frage nach der Konzentration also z.B.: konzentriert sich der Hauptanteil des Jahresumsatzes auf einige wenige Artikel im Sortiment?

Generelle Idee bei der Konzentrationsmessung:

- keine Konzentration: Merkmalssumme ist gleichmäßig auf alle Elemente des Datensatzes verteilt
- höchste Konzentration: ein Element trägt die gesamte Merkmalssumme, die anderen "gehen leer aus"

Frage: wie kann man Konzentration messen?
Graphisches Hilfsmittel dazu ist die **Lorenzkurve**.

Definition 2.47 *Lorenzkurve*

Sei X ein kardinal skaliertes Merkmal mit nicht negativen Ausprägungen, seien $x_1, \ldots, x_n$ beobachtet, seien $x_{(1)}, \ldots, x_{(n)}$ die entsprechenden geordneten Beobachtungen. Jede einzelne Beobachtung x_i kommt also mit relativer Häufigkeit $f_i = \frac{1}{n}$ im Datensatz vor. Die **Lorenzkurve** *ist der Streckenzug durch die Punkte $(0,0)$, $(u_1, v_1), \ldots, (u_n, v_n) = (1,1)$ mit*

$$u_q = \sum_{i=1}^{q} f_i = \frac{q}{n} \qquad \textit{kumulierter Anteil der Merkmalsträger,}$$

$$v_q = \sum_{i=1}^{q} \widetilde{v}_i \qquad \textit{kumulierte relative Merkmalssumme.}$$

Dabei ist

$$\widetilde{v}_q = \frac{x_{(q)}}{\sum_{i=1}^{n} x_i}.$$

Beispiel 2.48 *Marktkonzentration in drei Städten*

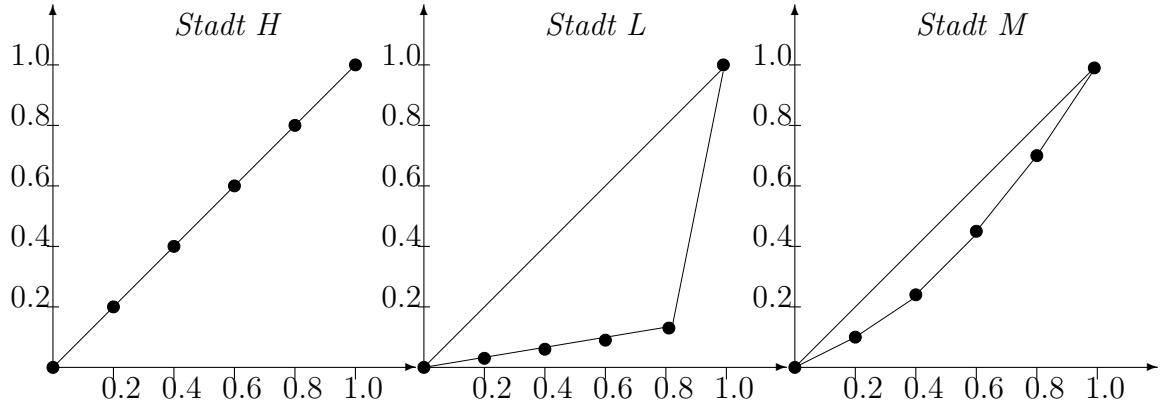
Monatlicher Umsatz (in 1 000 Euro) der Möbelbranche in drei Städten:

Einrichtungs-häuser	*Stadt* H	L	M
1	40	180	60
2	40	5	50
3	40	5	40
4	40	5	30
5	40	5	20

→ Gesamtumsatz in Stadt H gleichmäßig auf die fünf Anbieter verteilt (keine Konzentration), in L hat ein Anbieter fast Monopolstellung, in Stadt M variieren die Umsätze über die Anbieter

zugehörige Lorenzkurven: zunächst Arbeitstabelle der benötigten Werte

			H			L			M		
q	f_q	u_q	$x_{(q)}$	$\widetilde{v}_q$	v_q	$x_{(q)}$	$\widetilde{v}_q$	v_q	$x_{(q)}$	$\widetilde{v}_q$	v_q
1	0.2	0.2	40	0.2	0.2	5	0.025	0.025	20	0.10	0.10
2	0.2	0.4	40	0.2	0.4	5	0.025	0.050	30	0.15	0.25
3	0.2	0.6	40	0.2	0.6	5	0.025	0.075	40	0.20	0.45
4	0.2	0.8	40	0.2	0.8	5	0.025	0.100	50	0.25	0.70
5	0.2	1.0	40	0.2	1.0	180	0.900	1.000	60	0.30	1.00
$\sum$			200			200			200		



Ablesebeispiel:

Stadt L: auf 80% der Einrichtungshäuser entfallen lediglich 10% der Umsätze

Man sieht, dass die Lorenzkurve gleich der Winkelhalbierenden ist, wenn keine Konzentration vorliegt. Bei starker Konzentration (Stadt L im Beispiel) ist die Lorenzkurve sehr weit von der Winkelhalbierenden entfernt.

Bemerkung 2.49 *Eigenschaften*

- *Die Lorenzkurve ist* **monoton wachsend** *und* **konvex**, *d.h. sie wölbt sich nach unten.*

- *Die Lorenzkurve beschreibt die* **relative Konzentration***: sie setzt den Anteil der Menge (kumulierte relative Merkmalssumme) in Beziehung zum Anteil der Merkmalsträger (kumulierte relative Häufigkeit) (Fragestellungen wie "Wieviel Prozent der Marktteilnehmer teilen sich wieviel Prozent des Volumens?"). Zur Messung der* **absoluten Konzentration** *werden andere Maße benötigt (Fragestellungen wie "Wieviele Anbieter haben wieviel Prozent des Marktvolumens?")*

Bemerkung 2.50 *Lorenzkurve bei unklassierter und klassierter Häufigkeitsverteilung*

Sei X ein kardinal skaliertes Merkmal mit nicht negativen Ausprägungen.

1. *Kennt man nur die unklassierte Häufigkeitsverteilung, geht man zur Bestimmung der Lorenzkurve folgendermaßen vor:*
   bekannt sind die k Merkmalsausprägungen $a_1 \leq \ldots \leq a_k$ (der Größe nach geordnet!) und die zugehörigen relativen Häufigkeiten $f_1, \ldots, f_k = f(a_1), \ldots, f(a_k)$; die Lorenzkurve ist der Streckenzug durch die Punkte $(0,0)$, $(u_1, v_1), \ldots, (u_k, v_k) = (1,1)$ mit

   $$u_q = \sum_{j=1}^{q} f(a_j) = \sum_{j=1}^{q} f_j$$

   und

   $$v_q = \sum_{j=1}^{q} \widetilde{v}_j, \quad q = 1, \ldots, k.$$

   Dabei ist

   $$\widetilde{v}_q = \frac{f(a_q) \cdot a_q}{\sum_{j=1}^{k} f(a_j) \cdot a_j} = \frac{f_q \cdot a_q}{\sum_{j=1}^{k} f_j \cdot a_j}.$$

2. *Liegt nur eine klassierte Häufigkeitsverteilung vor, wobei die Klassengrenzen und die relativen Klassenhäufigkeiten bekannt sind, so ordnet man zunächst die k Klassen in natürlicher Weise an. Man unterscheidet dann zwei Fälle:*

(a) *Für die einzelnen Klassen sind die Merkmalssummen angegeben. Dann bezeichnet man die* **Merkmalssumme** *in der q-ten Klasse mit x_q und bestimmt die Lorenzkurve wie vorher als Streckenzug durch die Punkte $(0,0)$, $(u_1,v_1),\ldots,(u_k,v_k)=(1,1)$ mit*

$$u_q = \sum_{j=1}^{q} f_j,$$

$$v_q = \sum_{j=1}^{q} \widetilde{v}_j,$$

wobei

$$\widetilde{v}_q = \frac{x_q}{\sum_{j=1}^{k} x_j}.$$

(b) *Für die einzelnen Klassen sind die Merkmalssummen unbekannt. Man nimmt dann vereinfachend an, dass in den einzelnen Klassen keine Konzentration vorliegt, d.h. dass alle Beobachtungen denselben Wert, nämlich die Klassenmitte, annehmen; man bestimmt die Lorenzkurve in diesem Fall als Streckenzug durch die Punkte $(0,0)$, $(u_1,v_1),\ldots,(u_k,v_k)=(1,1)$ mit*

$$u_q = \sum_{j=1}^{q} f_j \quad \textit{und} \quad v_q = \sum_{j=1}^{q} \widetilde{v}_j,$$

wobei

$$\widetilde{v}_q = \frac{f_q \cdot m_q}{\sum_{j=1}^{k} f_j \cdot m_j},$$

k die Anzahl der Klassen bezeichnet und m_q die Klassenmitte der q-ten Klasse.

Wie gesehen, beschreibt bei der Lorenzkurve die "Entfernung" von der Hauptdiagonalen die Stärke der Konzentration. Eine Maßzahl für die Konzentration erhält man entsprechend, indem man diese Entfernung misst. Die Größe der

Entfernung der Kurve von der Hauptdiagonalen kann man messen durch die Fläche zwischen der Kurve und der Diagonalen.

Definition 2.51 *Gini-Koeffizient*

Der **Gini-Koeffizient** *ist definiert als*

$$G = \frac{\textit{Fläche zwischen Diagonale und Lorenzkurve}}{\textit{Fläche zwischen Diagonale und u-Achse}}.$$

Die Grundformel zur Berechnung des Gini-Koeffizienten lautet

$$G = \sum (u_{i-1} + u_i) \cdot \widetilde{v}_i - 1.$$

Dabei sind die u_i, $\widetilde{v}_i$ je nach Datenlage zu wählen.

1. *Liegen die Daten als Urliste vor, so sind u_i, $\widetilde{v}_i$ wie in Definition 2.47 und $u_0 = 0$, und es wird über alle n Beobachtungen summiert.*
   Alternativ kann man bei bekannter Urliste auch die Form
   $$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_{(i)}}{n \cdot \sum_{i=1}^{n} x_i} - \frac{n+1}{n}$$
   benutzen.

2. *Liegt nur die unklassierte Häufigkeitsverteilung vor (Merkmalsausprägungen $a_1 \leq \ldots \leq a_k$ der Größe nach geordnet und zugehörige relative Häufigkeiten $f_1, \ldots, f_k$), so sind u_q, $\widetilde{v}_q$ wie im ersten Teil von Bemerkung 2.50 zu wählen, also*
   $$u_q = \sum_{j=1}^{q} f_j,$$
   $$\widetilde{v}_q = \frac{f_q \cdot a_q}{\sum_{j=1}^{k} f_j \cdot a_j},$$
   so dass
   $$G = \sum_{j=1}^{k} (u_{j-1} + u_j) \cdot \widetilde{v}_j - 1.$$
   Auch hier ist $u_0 = 0$.

3. *Liegt nur die klassierte Häufigkeitsverteilung vor (die k Klassen müssen in natürlicher Weise angeordnet sein!), sind wie bei der Lorenzkurve zwei Fälle zu unterscheiden:*

   (a) *Für die einzelnen Klassen sind die Merkmalssummen angegeben; dann ist wieder x_q als Merkmalssumme der q-ten Klasse zu wählen und*

$$u_q = \sum_{j=1}^{q} f_j,$$

$$\widetilde{v}_q = \frac{x_q}{\sum_{j=1}^{k} x_j},$$

   $u_0 = 0$ *(vgl. auch 2.50, Teil 2(a)).*

   (b) *Für die einzelnen Klassen sind die Merkmalssummen unbekannt; dann nimmt man wie bei der Lorenzkurve an, dass innerhalb der Klassen keine Konzentration vorliegt und benutzt*

$$u_q = \sum_{j=1}^{q} f_j,$$

$$\widetilde{v}_q = \frac{f_q \cdot m_q}{\sum_{j=1}^{k} f_j \cdot m_j},$$

   $u_0 = 0$, *wobei wieder m_i die Klassenmitte der i-ten Klasse bezeichnet (vgl. auch hierzu 2.50, Teil 2(b)).*

Beispiel 2.52 *Marktkonzentration in drei Städten*

Gini-Koeffizienten für die drei Städte: hier Rechnung über die alternative Formel aus Definition 2.51 (Teil 1)

Stadt H:

$$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_{(i)}}{n \cdot \sum_{i=1}^{n} x_i} - \frac{n+1}{n} = \frac{2 \cdot (1 \cdot 40 + 2 \cdot 40 + 3 \cdot 40 + 4 \cdot 40 + 5 \cdot 40)}{5 \cdot 200} - \frac{6}{5} = 0.$$

Stadt L:

$$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_{(i)}}{n \cdot \sum_{i=1}^{n} x_i} - \frac{n+1}{n} = \frac{2 \cdot (1 \cdot 5 + 2 \cdot 5 + 3 \cdot 5 + 4 \cdot 5 + 5 \cdot 180)}{5 \cdot 200} - \frac{6}{5} = 0.7.$$

Stadt M:

$$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_{(i)}}{n \cdot \sum_{i=1}^{n} x_i} - \frac{n+1}{n} = \frac{2 \cdot (1 \cdot 20 + 2 \cdot 30 + 3 \cdot 40 + 4 \cdot 50 + 5 \cdot 60)}{5 \cdot 200} - \frac{6}{5} = 0.2.$$

Bemerkung 2.53 *Eigenschaften*

- *Der Gini-Koeffizient entspricht zugleich dem Doppelten der Fläche zwischen Diagonale und Lorenzkurve.*
- *G kann als extreme Ausprägungen die Werte 0 (bei Nullkonzentration, d.h. alle x_i sind gleich) und $\frac{n-1}{n}$ (bei maximaler Konzentration, d.h. alle x_i bis auf eines sind Null) annehmen.*
- *Da die größtmögliche Ausprägung beim Gini-Koeffizient noch von der Anzahl n der Beobachtungen abhängt, geht man oft über zum* **normierten Gini-Koeffizienten** G^* *mit*

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

  Der normierte Gini-Koeffizient nimmt nur noch Werte zwischen 0 und 1 an und ist damit leichter zu interpretieren (0 = keine Konzentration, 1 = maximale Konzentration).
- *Bei der Interpretation des Wertes von G bzw. G^* ist zusätzlich die zugehörige Lorenzkurve zu betrachten, denn unterschiedliche Lorenzkurven können zum gleichen Gini-Koeffizienten führen.*

Bisher: relative Konzentration betrachtet; jetzt: absolute Konzentration

Beispiel 2.54 *Konzentration*

3 Firmen bringen dasselbe Produkt auf den Markt mit Umsätzen

$x_1 = 100$, $x_2 = 40$, $x_3 = 10$

in dieser Situation: Gini-Koeffizient $G = 0.4$

- *Kommt eine vierte (Schein-)Firma dazu mit Umsatz* $x_4 = 0$
  $\Rightarrow G = 0.55 \rightarrow$ *ökonomisch unlogisch, Konzentration ändert sich eigentlich nicht.*

- *Splittet man die 3 Firmen auf in 6 mit jeweils halbem Umsatz*
  $x_1 = x_4 = 50$, $x_2 = x_5 = 20$, $x_3 = x_6 = 5$
  $\Rightarrow G = 0.4 \rightarrow$ *ebenfalls unlogisch, hier ändert sich die Konzentration*

$\rightarrow$ *für solche Fälle scheint ein relatives Konzentrationsmaß nicht angebracht*

Definition 2.55 *Konzentrationsmaß nach Hirschmann / Herfindahl*
Sei X *ein Merkmal mit nicht negativen Ausprägungen, seien* $x_1, \ldots, x_n$ *(Urliste) beobachtet, und sei außerdem* $\sum_{i=1}^{n} x_i > 0$. *Dann heißt*

$$H = \sum_{i=1}^{n} \widetilde{v}_i^2$$

der **Index von Hirschmann / Herfindahl** *für die absolute Konzentration. Dabei ist*

$$\widetilde{v}_q = \frac{x_q}{\sum_{i=1}^{n} x_i}.$$

Bemerkung 2.56 *Eigenschaften*

- *Liegen die Daten nicht als Urliste vor, so gibt es alternative Berechnungsformeln für* H *(vergleiche Übung).*

- H *ist ein Maß für die absolute Konzentration; kommt zu den beobachteten Werten ein weiterer Wert mit Ausprägung "0" hinzu, ändert sich* H *nicht.*

- $\frac{1}{n} \leq H \leq 1$, *wobei die Extreme bei minimaler bzw. maximaler Konzentration angenommen werden, d.h.*

  $H = \frac{1}{n}$, *falls alle* x_i *gleich, und*

  $H = 1$, *falls ein* x_i *die gesamte Merkmalssumme trägt und alle anderen gleich Null sind.*

Beispiel 2.57 *Konzentration*

In der Situation aus Beispiel 2.54 ergibt sich

- *für die Ausgangssituation mit* $x_1 = 100$, $x_2 = 40$, $x_3 = 10$*:*
  $\sum_{i=1}^{3} x_i = 150$ *und* $H = \left(\frac{100}{150}\right)^2 + \left(\frac{40}{150}\right)^2 + \left(\frac{10}{150}\right)^2 = 0.52$

- *für die Hinzunahme einer Scheinfirma mit* $x_4 = 0$*:*
  $\sum_{i=1}^{4} x_i = 150$ *und* $H = \left(\frac{100}{150}\right)^2 + \left(\frac{40}{150}\right)^2 + \left(\frac{10}{150}\right)^2 + \left(\frac{0}{150}\right)^2 = 0.52$
  $\rightarrow$ *Konzentration ändert sich nicht*

- *für die Aufsplittung in 6 Firmen mit* $x_1 = x_4 = 50$, $x_2 = x_5 = 20$, $x_3 = x_6 = 5$*:* $\sum_{i=1}^{6} x_i = 150$ *und*

$$
\begin{aligned}
H &= \left(\frac{50}{150}\right)^2 + \left(\frac{50}{150}\right)^2 + \left(\frac{20}{150}\right)^2 + \left(\frac{20}{150}\right)^2 + \left(\frac{5}{150}\right)^2 + \left(\frac{5}{150}\right)^2 \\
&= 2 \cdot \left(\frac{50}{150}\right)^2 + 2 \cdot \left(\frac{20}{150}\right)^2 + 2 \cdot \left(\frac{5}{150}\right)^2 \\
&= \frac{2}{2} \cdot 2 \cdot \left(\frac{50}{150}\right)^2 + \frac{2}{2} \cdot 2 \cdot \left(\frac{20}{150}\right)^2 + \frac{2}{2} \cdot 2 \cdot \left(\frac{5}{150}\right)^2 \\
&= \frac{1}{2} \cdot 4 \cdot \left(\frac{50}{150}\right)^2 + \frac{1}{2} \cdot 4 \cdot \left(\frac{20}{150}\right)^2 + \frac{1}{2} \cdot 4 \cdot \left(\frac{5}{150}\right)^2 \\
&= \frac{1}{2} \cdot \left(\frac{2 \cdot 50}{150}\right)^2 + \frac{1}{2} \cdot \left(\frac{2 \cdot 20}{150}\right)^2 + \frac{1}{2} \cdot \left(\frac{2 \cdot 5}{150}\right)^2 \\
&= \frac{1}{2} \cdot \left(\underbrace{\left(\frac{2 \cdot 50}{150}\right)^2 + \left(\frac{2 \cdot 20}{150}\right)^2 + \left(\frac{2 \cdot 5}{150}\right)^2}_{0.52} \right) = 0.26
\end{aligned}
$$

  $\rightarrow$ *Konzentration halbiert sich*

3 Mehrdimensionale Merkmale

Im Rahmen einer Studie oder Umfrage: in der Regel mehr als ein Merkmal erhoben

Man spricht von **mehrdimensionalen (multivariaten)** Daten.

Hier: zweidimensionale Daten

→ insbesondere Interesse an **Zusammenhängen** zwischen den Merkmalen

3.1 Gemeinsame und bedingte Verteilung zweier Merkmale

Beispiel 3.1 *Autofarben*

Ein Autohändler hat bei $n = 36$ verkauften Neuwagen jeweils notiert, welche Farbe der Wagen hatte und ob ein Mann oder eine Frau Käufer war:

Merkmal X = Geschlecht mit Ausprägungen männlich, weiblich

Merkmal Y = Farbe mit Ausprägungen schwarz, weiß, rot, silbermetallic

Beobachtete Werte sind damit jetzt Paare: etwa für Kunde 1: (m, weiß), für Kunde 2: (w, rot), usw.

Darstellung möglich als Tabelle:

		Y					
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$	
X	*weiblich*	*2 (0.056)*	*2 (0.056)*	*10 (0.278)*	*12 (0.333)*	**26**	*Randverteilung des*
	männlich	*0 (0)*	*2 (0.056)*	*2 (0.056)*	*6 (0.167)*	**10**	*Geschlechts*
	$\sum$	**2**	**4**	**12**	**18**	*36*	
		Randverteilung der Farbe					

in Klammern: relative Häufigkeiten (kann man alternativ auch benutzen)

Das Innere der Tabelle gibt die **gemeinsame Verteilung** *der Merkmale X und Y in absoluten (relativen) Häufigkeiten an (z.B. 6 Autos (= 16.7%) waren silbern und gingen an männliche Käufer).*
Die rechte Randspalte gibt die **Randverteilung** *des Merkmals X in absoluten Häufigkeiten an (z.B. 26 weibliche Käufer).*
Die untere Randzeile gibt die **Randverteilung** *des Merkmals Y in absoluten Häufigkeiten an (z.B. 12 verkaufte rote Wagen).*

Bemerkung 3.2 *Formalisierung*

Die Größen aus Beispiel 3.1 lassen sich folgendermaßen formalisieren:

beobachtet wird

$$\left.\begin{array}{l} \text{Merkmal } X \text{ mit Ausprägungen } a_1, a_2 \\ \text{Merkmal } Y \text{ mit Ausprägungen } b_1, b_2, b_3, b_4 \end{array}\right\} \begin{array}{l} \text{beide Merkmale diskret oder} \\ \text{in Kategorien eingeteilt} \end{array}$$

Beobachtungswerte sind Paare $(x_1, y_1), \ldots (x_n, y_n)$ mit $n = 36$,
mögliche Merkmalskombinationen sind $(a_1, b_1), (a_1, b_2), \ldots, (a_2, b_4)$.
In der Tabelle wird eingetragen, wie oft welche Merkmalskombination auftritt
$\rightarrow$ absolute (relative) Häufigkeiten der (a_i, b_j), $i = 1, 2$, $j = 1, \ldots, 4$

		Y				
		b_1	b_2	b_3	b_4	$\sum$
X	a_1	h_{11}	h_{12}	h_{13}	h_{14}	$\sum_{j=1}^{4} h_{1j}$
	a_2	h_{21}	h_{22}	h_{23}	h_{24}	$\sum_{j=1}^{4} h_{2j}$
	$\sum$	$\sum_{i=1}^{2} h_{i1}$	$\sum_{i=1}^{2} h_{i2}$	$\sum_{i=1}^{2} h_{i3}$	$\sum_{i=1}^{2} h_{i4}$	$n = \sum_{i=1}^{2} \sum_{j=1}^{4} h_{ij}$

absolute Häufigkeit: $h_{ij} = h(a_i, b_j) =$ Häufigkeit der Kombination (a_i, b_j)
auch möglich: benutze relative Häufigkeit: $f_{ij} = f(a_i, b_j) = \frac{h_{ij}}{n}$

Tabelleninneres: gemeinsame Verteilung von X und Y
Tabellenränder: Randverteilungen von X bzw. Y

Definition 3.3 *Gemeinsame Verteilung, Randverteilung*

Gegeben seien zwei diskrete (oder kategorisierte) Merkmale X und Y mit Ausprägungen $a_1, \ldots, a_k$ (Merkmal X) und $b_1, \ldots, b_m$ (Merkmal Y). Seien $(x_1, y_1), \ldots, (x_n, y_n)$ beobachtet, und seien h_{ij} die absoluten Häufigkeiten der Ausprägungskombinationen (a_i, b_j), $i = 1, \ldots, k$, $j = 1, \ldots, m$.

(a) Die absoluten Häufigkeiten $h_{11}, \ldots, h_{km}$ heißen die **gemeinsame empirische Verteilung von X und Y in absoluten Häufigkeiten**.
Die relativen Häufigkeiten $f_{11} = \frac{h_{11}}{n}, \ldots, f_{km} = \frac{h_{km}}{n}$ heißen die **gemeinsame empirische Verteilung von X und Y in relativen Häufigkeiten**.

(b) Die aufsummierten Häufigkeiten
$h_{i\bullet} = \sum_{j=1}^{m} h_{ij}$ bzw. $f_{i\bullet} = \sum_{j=1}^{m} f_{ij} = \frac{h_{i\bullet}}{n}$ heißen **absolute** *bzw.* **relative Randhäufigkeiten von X** *($i = 1, \ldots, k$),*
$h_{\bullet j} = \sum_{i=1}^{k} h_{ij}$ bzw. $f_{\bullet j} = \sum_{i=1}^{k} f_{ij} = \frac{h_{\bullet j}}{n}$ heißen **absolute** *bzw.* **relative Randhäufigkeiten von Y** *($j = 1, \ldots, m$).*

(c) Die Folge der Randhäufigkeiten $h_{1\bullet}, \ldots, h_{k\bullet}$ ($f_{1\bullet}, \ldots, f_{k\bullet}$) heißt **empirische Randverteilung von X** *in absoluten (relativen) Häufigkeiten. Entsprechend heißt $h_{\bullet 1}, \ldots, h_{\bullet m}$ ($f_{\bullet 1}, \ldots, f_{\bullet m}$)* **empirische Randverteilung von Y** *in absoluten (relativen) Häufigkeiten.*

(d) Tabellen wie in Beispiel 3.1 und Bemerkung 3.2 heißen **Kontingenztafeln** *(oder* **-tabellen***). Sie können in absoluten oder relativen Häufigkeiten angelegt werden.*

Beispiel 3.4 *Zusammenhänge zwischen Merkmalen*

Im Beispiel des Autohändlers: insgesamt 36 Autos verkauft, davon nur zwei schwarze an weibliche Kunden

→ Verkaufen sich schwarze Autos nicht gut an Frauen?

→ Kann man so nicht beantworten. Man muss sich den Verkauf schwarzer Autos insgesamt anschauen.

→ Insgesamt 2 schwarze Wagen verkauft, davon 2 (=100%) an Frauen.

→ Schaut man sich nur die schwarzen Autos an, so verkaufen die sich besonders gut an Frauen.

Die Aussage wurde hier getroffen **unter der Bedingung***, dass man sich nur schwarze Autos anschaut*

→ sogenannte **bedingte Verteilung** *der Geschlechter, gegeben man beschränkt sich auf schwarze Autos*

dies kann man für jede Ausprägung des Merkmals "Farbe" machen, führt insgesamt zur **bedingten Verteilung des Merkmals "Geschlecht" (X), gegeben das Merkmal "Farbe" (Y)** *(immer in relativen Häufigkeiten):*

	schwarz	*weiß*	*rot*	*silbermetallic*
weiblich	*1*	*0.5*	**0.83***	*0.67*
männlich	*0*	**0.5****	*0.17*	*0.33*
$\sum$	*1*	*1*	*1*	*1*

**: 83% aller derjenigen, die ein rotes Auto kaufen, sind Frauen*

***: weiße Autos werden zur Hälfte von Männern gekauft*

(alle Aussagen beziehen sich auf den speziellen Händler)

Genauso: wenn man nur bei den Männern schaut, welches ist die beliebteste Autofarbe?

Dazu: **bedingte Verteilung des Merkmals "Farbe", gegeben das Merkmal "Geschlecht":**

	schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
weiblich	*0.08*	*0.08*	*0.38*	**0.46***	*1*
männlich	*0*	*0.2*	*0.2*	**0.6****	*1*

60% aller männlichen Kunden kaufen ein silbernes Auto (**),
ebenso 46% aller weiblichen Kunden (*)

Definition 3.5 *Bedingte Verteilung*
Seien h_{ij} bzw. f_{ij} die absoluten bzw. relativen Häufigkeiten der Merkmale X und Y ($i = 1, \ldots, k$, $j = 1, \ldots, m$).

(a) Für feste Wahl von j heißen $f_X(a_i|b_j) = \frac{h_{ij}}{h_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}$, $i = 1, \ldots, k$, **bedingte relative Häufigkeiten von X gegeben $Y = b_j$;**
$f_X(a_1|b_j), \ldots, f_X(a_k|b_j)$ heißt **bedingte Verteilung von X gegeben $Y = b_j$.**

(b) Für feste Wahl von i heißen $f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$, $j = 1, \ldots, m$, **bedingte relative Häufigkeiten von Y gegeben $X = a_i$;**
$f_Y(b_1|a_i), \ldots, f_Y(b_m|a_i)$ heißt **bedingte Verteilung von Y gegeben $X = a_i$.**

Bemerkung 3.6 *Eigenschaften*

- *Es gilt:*
  $\sum_{i=1}^{k} f_X(a_i|b_j) = 1$ für jedes feste j ($j = 1, \ldots, m$).
  $\sum_{j=1}^{m} f_Y(b_j|a_i) = 1$ für jedes feste i ($i = 1, \ldots, k$).

- *In Beispiel 3.4 ist etwa* $f_X(a_2|b_3) = f_X(männl.|rot) = 0.17$.
- *Die bedingten Häufigkeiten einer Variable (z.B.* X*) dienen zum Vergleich von Teilgesamtheiten, die durch die zweite Variable (*Y*) gebildet werden (z.B. Farbwahl in den durch die Geschlechter gegebenen Teilgesamtheiten).*
- *Kennt man alle bedingten und alle Randhäufigkeiten, so lassen sich daraus die gemeinsamen relativen Häufigkeiten* f_{ij} *rekonstruieren:*

$$f_{ij} = f_Y(b_j|a_i) \cdot f_{i\bullet} \qquad bzw. \qquad f_{ij} = f_X(a_i|b_j) \cdot f_{\bullet j}$$

3.2 Zusammenhangsanalyse in Kontingenztafeln

Ausgangspunkt: Kontingenztafel, kategorisierte oder diskrete Merkmale X und Y
gesucht: Maßzahl zur quantitativen Erfassung des Zusammenhangs zwischen X und Y

Idee:

- Nimm die Randverteilungen als gegeben an, d.h. die Verteilung von X (ohne Beachtung von Y) und die Verteilung von Y (ohne Beachtung von X).
- Nimm an, es gäbe keinen Zusammenhang zwischen X und Y. Wie müsste dann die gemeinsame Verteilung (also das "Innere" der Kontingenztafel) aussehen?
- Vergleiche diese hypothetische Kontingenztafel mit der, die tatsächlich vorliegt. Wie stark unterscheiden sich beide?
- Bei starkem Unterschied: schließe auf Abhängigkeit von X und Y.

Beispiel 3.7 *Autofarben*

Tabelle aus Beispiel 3.1 (nur absolute Häufigkeiten):

		Y				
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
X	*weiblich*	*2*	*2*	*10*	*12*	*26*
	männlich	*0*	*2*	*2*	*6*	*10*
	$\sum$	*2*	*4*	*12*	*18*	*36*

zunächst nur die Randverteilungen betrachten:

		Y				
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
X	*weiblich*					*26*
	männlich					*10*
	$\sum$	*2*	*4*	*12*	*18*	*36*

nehmen wir an, Geschlecht und gewählte Autofarbe wären unabhängig; dann müssten sich die 10 Männer und die 26 Frauen "in gleicher Weise" auf die vier Autofarben verteilen

→ (bedingte Verteilung auf die Autofarbe gegeben Geschlecht = männlich)

= (bedingte Verteilung auf die Autofarbe gegeben Geschlecht = weiblich)

Durch Einsetzen in die entsprechenden Formeln und Ausrechnen erhält man für die hypothetischen absoluten Häufigkeiten

$$e_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$$

Damit kann man das Tabelleninnere mit den hypothetischen Häufigkeiten füllen:

		Y				
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
	weiblich	$\frac{26 \cdot 2}{36} = 1.44$				**26**
X						
	männlich					*10*
	$\sum$	**2**	*4*	*12*	*18*	*36*

usw. für alle Zellen:

		Y				
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
	weiblich	*1.44*	*2.89*	*8.67*	*13*	*26*
X						
	männlich	*0.56*	*1.11*	*3.33*	*5*	*10*
	$\sum$	*2*	*4*	*12*	*18*	*36*

Dies ist nun zu vergleichen mit der Originaltabelle:

		Y				
		schwarz	*weiß*	*rot*	*silbermetallic*	$\sum$
	weiblich	*2*	*2*	*10*	*12*	*26*
X						
	männlich	*0*	*2*	*2*	*6*	*10*
	$\sum$	*2*	*4*	*12*	*18*	*36*

→ *Frage: wie kann man einen solchen Vergleich anstellen?*

Definition 3.8 *Chi-Quadrat-Koeffizient*

In einer Kontingenztafel seien h_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, m$, die beobachteten gemeinsamen absoluten Häufigkeiten zweier Merkmale X und Y sowie $h_{i\bullet}$, $h_{\bullet j}$ die entsprechenden Randhäufigkeiten. Mit $e_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$ seien die unter Unabhängigkeit von X und Y erwarteten gemeinsamen absoluten Häufigkeiten bezeichnet. Der sogenannte **Chi-Quadrat-Koeffizient (χ^2-Koeffizient)** *misst den Unterschied zwischen der beobachteten und der unter Unabhängigkeit erwarteten Tafel:*

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}}$$

Dabei bedeutet ein Wert von $\chi^2 = 0$, dass kein Zusammenhang zwischen den beiden Merkmalen vorliegt. Je größer χ^2, desto stärker ist der Zusammenhang.

Beispiel 3.9 *Autofarben*

Im Beispiel der Autofarben ergibt sich aus den in Beispiel 3.7 bestimmten Werten:

$$\begin{aligned} \chi^2 &= \frac{(2-1.44)^2}{1.44} + \frac{(2-2.89)^2}{2.89} + \frac{(10-8.67)^2}{8.67} + \frac{(12-13)^2}{13} \\ &\quad + \frac{(0-0.56)^2}{0.56} + \frac{(2-1.11)^2}{1.11} + \frac{(2-3.33)^2}{3.33} + \frac{(6-5)^2}{5} \\ &= 0.22 + 0.27 + 0.20 + 0.08 + 0.56 + 0.71 + 0.53 + 0.2 = 2.77 \end{aligned}$$

→ Frage: ist dieser Wert von 2.77 als "groß" oder eher als "klein" anzusehen?

→ offensichtlich bräuchte man Vergleichswerte, um dies beurteilen zu können

→ alternativ: Normierung herstellen

Zur Normierung von χ^2 berechnet man zunächst den sogenannten **Kontingenzkoeffizienten**

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

der Werte im Bereich von Null bis $\sqrt{\frac{M-1}{M}}$ annehmen kann, mit $M = \min\{k, m\}$. Indem man K durch den für K maximal möglichen Wert dividiert, erhält man den **korrigierten Kontingenzkoeffizienten**, der dann nur noch Werte im Bereich von Null bis Eins annimmt.

Definition 3.10 *Korrigierter Kontingenzkoeffizient*
Unter den gleichen Voraussetzungen wie in Definition 3.8 misst der **korrigierte Kontingenzkoeffizient**

$$K^* = \frac{\sqrt{\frac{\chi^2}{n+\chi^2}}}{\sqrt{\frac{M-1}{M}}} = \frac{K}{\sqrt{\frac{M-1}{M}}}$$

die Stärke des Zusammenhangs zwischen X und Y. K^* kann Werte zwischen 0 und 1 (jeweils einschließlich) annehmen, wobei $K^* = 1$ für den höchstmöglichen Zusammenhang steht und $K^* = 0$ bedeutet, dass kein Zusammenhang vorliegt.

Beispiel 3.11 *Autofarben*
Im Beispiel der Autofarben war $\chi^2 = 2.77$, $n = 36$;
weiter ist $k = 2$, $m = 4 \Rightarrow M = \min\{2, 4\} = 2$
$\Rightarrow \sqrt{\frac{M-1}{M}} = \sqrt{1/2}$ *und*

$$K^* = \frac{\sqrt{\frac{2.77}{36+2.77}}}{\sqrt{1/2}} = \frac{0.27}{0.71} = 0.38$$

Interpretation: es besteht ein schwacher Zusammenhang zwischen dem Geschlecht und der Wahl der Autofarbe

Bemerkung 3.12 *Interpretation von K^**
Zur Interpretation von K^* benutzt man die folgende Faustregel:

$K^* \leq 0.2 \rightarrow$ *kein wesentlicher Zusammenhang zwischen X und Y*

$0.2 < K^* \leq 0.5 \rightarrow$ *schwacher Zusammenhang*

$0.5 < K^* < 0.8 \rightarrow$ *deutlicher Zusammenhang*

$0.8 \leq K^* \rightarrow$ *starker Zusammenhang*

3.3 Der Zusammenhang zwischen metrischen oder ordinalen Merkmalen

Bei zweidimensionalen, metrisch skalierten Merkmalen erstellt man häufig **Streudiagramme** zur Darstellung des Zusammenhangs

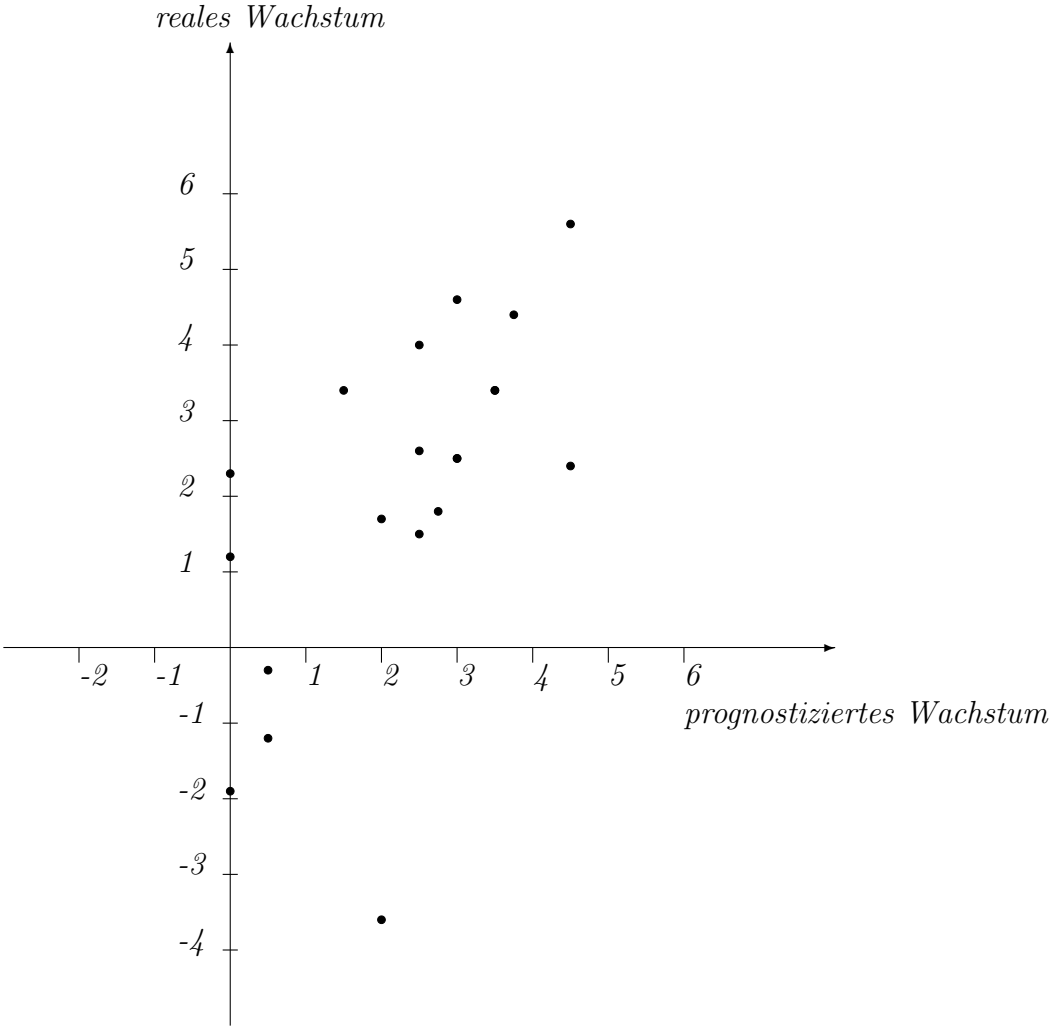
Merkmale: X, $Y \rightarrow$ Beobachtungen $(x_1, y_1), \ldots, (x_n, y_n)$
$\rightarrow$ trage Punktepaare in Koordinatensystem ein

Beispiel 3.13 *Wirtschaftswachstum*
X: vom Sachverständigenrat prognostiziertes Wachstum des Bruttoinlandsprodukts
Y: tatsächliches Wachstum
jeweils in Prozent

Jahr	*1975*	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*	*1982*	*1983*	*1984*
X	*2.0*	*4.5*	*4.5*	*3.5*	*3.75*	*2.75*	*0.5*	*0.5*	*1.0*	*2.5*
Y	*-3.6*	*5.6*	*2.4*	*3.4*	*4.4*	*1.8*	*-0.3*	*-1.2*	*1.2*	*2.6*

Jahr	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*
X	*3.0*	*3.0*	*2.0*	*1.5*	*2.5*	*3.0*	*3.5*	*2.5*	*0.0*	*0.0*
Y	*2.5*	*2.5*	*1.7*	*3.4*	*4.0*	*4.6*	*3.4*	*1.5*	*-1.9*	*2.3*

Streudiagramm:



Neben der graphischen Darstellung kann man den Zusammenhang zwischen metrischen Merkmalen auch messen. Man bestimmt den sogenannten **Korrelationskoeffizienten**.

Definition 3.14 *Korrelationskoeffizient nach Bravais-Pearson*
Gegeben seien zwei metrisch skalierte Merkmale X und Y, seien $(x_1, y_1), \ldots,$ (x_n, y_n) beobachtet. Weiter seien weder alle x_i identisch, noch seien alle y_i

identisch. Die Größe

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\widetilde{s}_X \cdot \widetilde{s}_Y}$$

heißt **Korrelationskoeffizient nach Bravais-Pearson**.

Der Ausdruck

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

heißt **empirische Kovarianz** *von X und Y.*

Bemerkung 3.15 *Eigenschaften*

- *Durch die Normierung mit $\widetilde{s}_X$, $\widetilde{s}_Y$ im Nenner von r_{XY} gilt stets:* $-1 \leq r_{XY} \leq 1$
- *r_{XY} misst, wie nahe die Punkte (x_i, y_i) entlang einer Geraden liegen (sog.* **linearer Zusammenhang***); der Wert von $|r_{XY}|$ ist umso näher bei 1, je näher die Punkte an einer Geraden liegen*
- $r_{XY} = 1 \quad (-1) \rightarrow$ *Werte liegen auf einer* **Geraden** *mit positiver (negativer) Steigung* $\rightarrow$ **linearer Zusammenhang**
- $r_{XY} = 0 \rightarrow$ *kein* **linearer** *(!) Zusammenhang; anderer Zusammenhang durchaus möglich*
- *Die Stärke des linearen Zusammenhangs wird durch die Größe von $|r_{XY}|$ charakterisiert. Man benutzt die gleiche Einteilung wie beim korrigierten Kontingenzkoeffizienten K^* (vergleiche Bemerkung 3.12)*
- *r_{XY} lässt sich alternativ berechnen als*

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2}}$$

Beispiel 3.16 *Wirtschaftswachstum*

Benutze zur Berechnung von r_{XY} die alternative Formel aus Bemerkung 3.15; bestimme dazu

$$\begin{array}{ll} \sum_{i=1}^{n} x_i = 46.5 & \overline{x} = 2.325 \\ \sum_{i=1}^{n} x_i^2 = 144.125 & \overline{x}^2 = 5.406 \\ \sum_{i=1}^{n} y_i = 40.3 & \overline{y} = 2.015 \\ \sum_{i=1}^{n} y_i^2 = 180.79 & \overline{y}^2 = 4.060 \\ \sum_{i=1}^{n} x_i \cdot y_i = 132.05 & \overline{x} \cdot \overline{y} = 4.685 \end{array}$$

Damit:

$$\begin{aligned} r_{XY} &= \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2}} \\ &= \frac{132.05 - 20 \cdot 4.685}{\sqrt{144.125 - 20 \cdot 5.406} \cdot \sqrt{180.79 - 20 \cdot 4.060}} = \frac{38.350}{59.881} = 0.640 \end{aligned}$$

Korrelation zwischen prognostiziertem und realem Wachstum beträgt also 0.640 $\rightarrow$ es besteht ein **deutlicher positiver linearer Zusammenhang** *zwischen Prognose des Sachverständigenrates und real eingetretenem Wachstum (Grad des Zusammenhangs genauso klassifiziert wie bei K^*)*

Neben dem Korrelationskoeffizienten nach Bravais-Pearson gibt es weitere Zusammenhangsmaße. Ein Maß, für das die beiden Merkmale X und Y lediglich ordinales Skalenniveau besitzen müssen, ist der **Rangkorrelationskoeffizient nach Spearman**.

Dieser basiert nicht auf den beobachteten Werten selbst, sondern auf ihren **Rängen**.

Definition 3.17 *Rang*

Sei X ein mindestens ordinal skaliertes Merkmal, seien $x_1, \ldots, x_n$ beobachtet. Seien weiter $x_{(1)}, \ldots, x_{(n)}$ die zugehörigen Ordnungsstatistiken. Dann ist der **Rang** *$rg(x_i)$ die Platznummer der i-ten Beobachtung x_i in der Reihe der Ordnungsstatistiken.*

Kommt eine Ausprägung mehrfach vor (sog. "Bindung"), so verwendet man das arithmetische Mittel der Platznummern als Rang.

Beispiel 3.18 *Rang*

6 Schüler haben folgende Noten in Englisch (Merkmal X):

x_1	x_2	x_3	x_4	x_5	x_6		$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$
3.0	*4.3*	*4.7*	*2.0*	*2.3*	*2.3*	$\rightarrow$ *ordnen:*	*2.0*	*2.3*	*2.3*	*3.0*	*4.3*	*4.7*
							x_4	x_5	x_6	x_1	x_2	x_3
						oder	x_4	x_6	x_5	x_1	x_2	x_3

$\rightarrow$ *Beobachtung x_1 steht an vierter Stelle der geordneten Reihe* $\Rightarrow rg(x_1) = 4$*;*

$\rightarrow$ *für Beobachtungen x_5 und x_6 ist keine eindeutige Anordnung möglich, beide könnten an zweiter oder dritter Stelle der geordneten Reihe stehen*

$\rightarrow$ *Bindung* $\Rightarrow rg(x_5) = rg(x_6) = (2+3)/2 = 2.5$

Ränge also:

Beob.	x_1	x_2	x_3	x_4	x_5	x_6
Wert	*3.0*	*4.3*	*4.7*	*2.0*	*2.3*	*2.3*
Rang	*4*	*5*	*6*	*1*	*2.5*	*2.5*

Definition 3.19 *Rangkorrelationskoeffizient nach Spearman*

Gegeben seien zwei mindestens ordinal skalierte Merkmale X und Y, seien $(x_1, y_1), \ldots, (x_n, y_n)$ beobachtet. Es seien weder alle x_i noch alle y_i identisch. Weiter seien $rg(x_1), \ldots, rg(x_n)$ die Ränge der x_i und $rg(y_1), \ldots, rg(y_n)$ die Ränge der y_i. Die Größe

$$\begin{aligned} r_{Sp} &= \frac{\sum_{i=1}^{n}(rg(x_i) - \frac{n+1}{2}) \cdot (rg(y_i) - \frac{n+1}{2})}{\sqrt{(\sum_{i=1}^{n}(rg(x_i))^2 - \frac{n\cdot(n+1)^2}{4}) \cdot (\sum_{i=1}^{n}(rg(y_i))^2 - \frac{n\cdot(n+1)^2}{4})}} \\ &= \frac{\sum_{i=1}^{n} rg(x_i) \cdot rg(y_i) - \frac{n\cdot(n+1)^2}{4}}{\sqrt{(\sum_{i=1}^{n}(rg(x_i))^2 - \frac{n\cdot(n+1)^2}{4}) \cdot (\sum_{i=1}^{n}(rg(y_i))^2 - \frac{n\cdot(n+1)^2}{4})}} \end{aligned}$$

heißt **Rangkorrelationskoeffizient nach Spearman**.

r_{Sp} ist vom selben Typ wie r_{XY}, nur unter Verwendung der Ränge. Daher reicht ordinales Skalenniveau.

Bemerkung 3.20 *Eigenschaften*

- **Falls keine Bindungen vorliegen**, *kann r_{Sp} einfacher berechnet werden:*

  $$r_{Sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

  wobei $d_i = rg(x_i) - rg(y_i)$.

- *Es ist $-1 \leq r_{Sp} \leq 1$*

- *r_{Sp} misst den Grad des* **monotonen Zusammenhangs**, *d.h. misst, inwieweit die y_i mit wachsenden x_i wachsen (bzw. fallen)*

- *$r_{Sp} = 1$: die Ränge der x_i stimmen mit denen der y_i voll überein ($rg(x_i) = rg(y_i)$ für alle i): wenn die x_i größer werden, werden auch die y_i größer*

- $r_{Sp} = -1$*: die Ränge der* x_i *sind zu denen der* y_i *gegenläufig* ($rg(x_i) = n + 1 - rg(y_i)$ *für alle* i*): wenn die* x_i *größer werden, werden die* y_i *kleiner*
- $r_{Sp} = 0$*: keine gemeinsame Tendenz der beiden Merkmale*
- *Die Stärke des Zusammenhangs wird durch die Größe von* $|r_{Sp}|$ *charakterisiert. Man benutzt die gleiche Einteilung wie beim korrigierten Kontingenzkoeffizienten* K^* *(wie bei* r_{XY}*, vergleiche auch Bemerkung 3.15)*

Beispiel 3.21 *Schulnoten in Englisch und Mathe*

Zu den Englischnoten aus Beispiel 3.18 werden auch die Mathenoten (Merkmal Y*) erhoben; es ist* $n = 6$*, damit* $(n+1)/2 = 3.5$ *und* $n \cdot (n+1)^2/4 = 73.5$

Zusammenstellung der für r_{Sp} *benötigten Größen:*

Schüler i	*1*	*2*	*3*	*4*	*5*	*6*	$\sum$
x_i	*3.0*	*4.3*	*4.7*	*2.0*	*2.3*	*2.3*	
y_i	*4.0*	*2.7*	*1.3*	*4.3*	*5.0*	*4.3*	
$rg(x_i)$	*4*	*5*	*6*	*1*	*2.5*	*2.5*	
$rg(y_i)$	*3*	*2*	*1*	*4.5*	*6*	*4.5*	
$rg(x_i) - \frac{n+1}{2}$	*0.5*	*1.5*	*2.5*	*-2.5*	*-1*	*-1*	
$rg(y_i) - \frac{n+1}{2}$	*-0.5*	*-1.5*	*-2.5*	*1*	*2.5*	*1*	
$(rg(x_i) - \frac{n+1}{2})$ $\cdot(rg(y_i) - \frac{n+1}{2})$	*-0.25*	*-2.25*	*-6.25*	*-2.5*	*-2.5*	*-1*	*-14.75*
$(rg(x_i))^2$	*16*	*25*	*36*	*1*	*6.25*	*6.25*	*90.5*
$(rg(y_i))^2$	*9*	*4*	*1*	*20.25*	*36*	*20.25*	*90.5*

$$\Rightarrow r_{Sp} = \frac{-14.75}{\sqrt{(90.5 - 73.5) \cdot (90.5 - 73.5)}} = \frac{-14.75}{\sqrt{17 \cdot 17}} = -0.87$$

Es besteht also ein starker negativer monotoner Zusammenhang zwischen der Englisch- und der Mathenote: je besser ein Schüler in Englisch, desto schlechter in Mathe (und umgekehrt).

Beachte: Vereinfachte Berechung von r_{Sp} hier nicht möglich, da Bindungen in den Daten vorliegen!

Bemerkung 3.22 *Korrelation und Kausalität, Scheinkorrelation, verdeckte Korrelation*

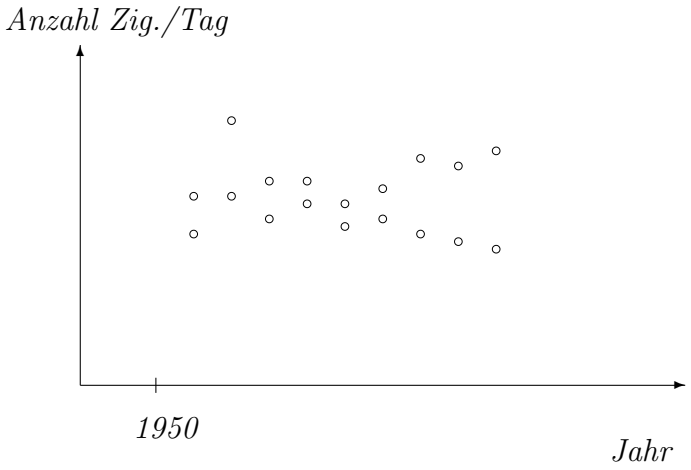
- *Korrelationskoeffizienten messen nur die Stärke des Zusammenhangs zwischen zwei Merkmalen, sagen aber nichts aus über Ursache und Wirkung. Kausalzusammenhänge können nicht allein an Hand von hohen Korrelationen begründet werden. Dazu müssen* **immer** *sachlogische Überlegungen herangezogen werden.*

- *Hängen zwei Merkmale X und Y über ein drittes (Z) miteinander zusammen und beachtet man Z nicht, so kann es zur* **Scheinkorrelation** *oder zur* **verdeckten Korrelation** *kommen.*

  Scheinkorrelation: *hohe Korrelation zwischen X und Y, obwohl sachlogisch kein Zusammenhang besteht.*
  Beispiel: X = Wortschatz von Kindern, Y = Körpergröße, $r_{XY} = 0.86$; Erklärung: sowohl Wortschatz als auch Körpergröße hängen zusammen mit Alter Z: $r_{YZ} = 0.99$, $r_{XZ} = 0.87$.

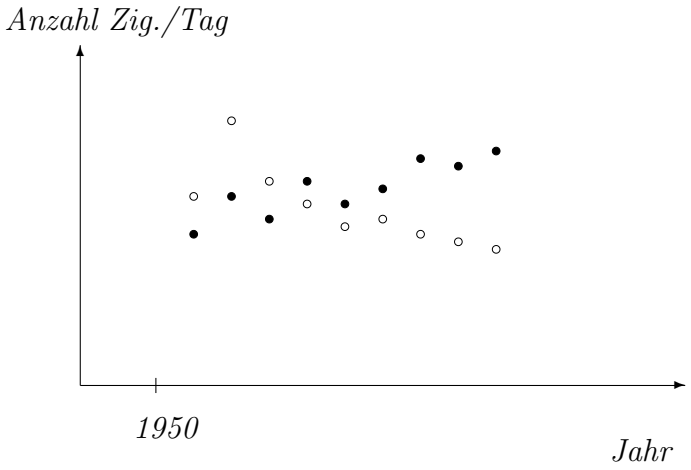
  Verdeckte Korrelation: *obwohl X und Y zusammenhängen, ist die Korrelation niedrig, oder es zeigt sich eine Korrelation mit anderem Vorzeichen als erwartet.*

Beispiel: X = Anzahl konsumierter Zigaretten pro Tag, Y = Jahr



→ *keine nennenswerte Korrelation*

→ *Aufsplittung nach Geschlecht Z: deutliche negative Korrelation bei den Männern ($\circ$), positive Korrelation bei den Frauen ($\bullet$).*



4 Regressionsanalyse

Die Regressionsanalyse dient zur Untersuchung von Zusammenhängen zwischen Merkmalen. Hier soll im Gegensatz zur Bestimmung der Korrelation aber nicht nur Typ und Stärke eines Zusammenhangs erfasst werden, sondern es soll der Zusammenhang durch eine Funktion beschrieben werden.

Oft: Zielgröße Y hängt ab von Einflussgrößen $X_1, \ldots, X_p$ über unbekannte Funktion g: $Y = g(X_1, \ldots, X_p)$

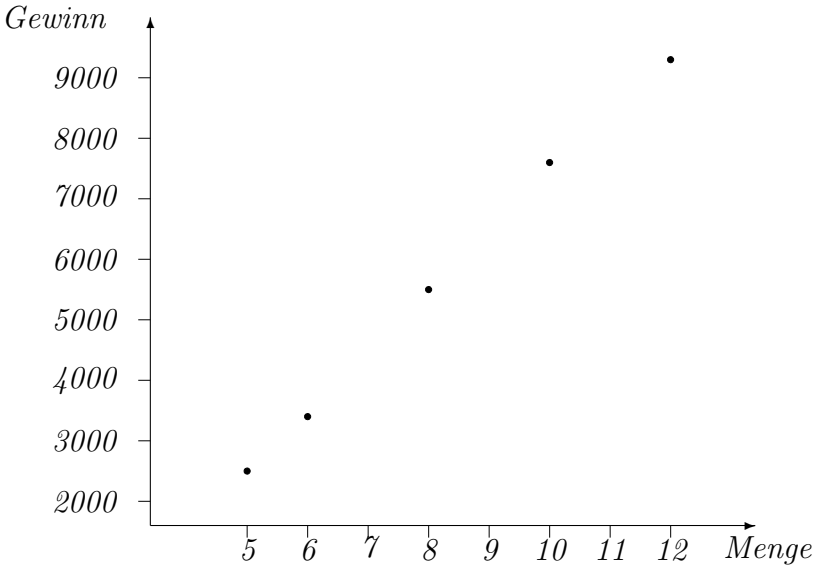
Ziel: Bestimmung der Funktion, um schließlich Werte von Y aus Werten für $X_1, \ldots, X_p$ vorherzusagen

Hier: Einfachster Fall, sogenannte **einfache lineare Regression**

Beispiel 4.1 *Einfache lineare Regression*

Ein Unternehmer beobachtet, welchen Gewinn er jeweils erwirtschaftet, wenn er eine bestimmte Menge seines Produkts herstellt. Er erhält folgende Daten:

Menge x_i (in 1000 Stück)	*5*	*6*	*8*	*10*	*12*
Gewinn y_i (in Euro)	*2600*	*3450*	*5555*	*7700*	*9350*



→ *Vermutung: Gewinn y hängt linear ab von produzierter Menge x; dabei ist der Zusammenhang nicht ganz exakt, da Zufallsschwankungen eine Rolle spielen (etwa schwankende Nachfrage)*

→ *Ansatz:* $y = \underbrace{a \cdot x + b}_{\text{linearer Zusammenhang}} + \underbrace{\varepsilon}_{\text{zufälliger Fehler}}$

Dabei sind a, b unbekannt. Die produzierte Menge x wird nicht als zufällig betrachtet, sondern ist vom Unternehmer deterministisch vorgegeben.
Möchte der Unternehmer nun wissen, mit welchem Gewinn er bei 9000 produzierten Stücken rechnen kann, wird er

- *graphisch: eine Ausgleichsgerade durch die beobachteten Punktepaare legen und deren Wert an der Stelle $x = 9$ ablesen;*
- *rechnerisch: a und b an Hand der Daten möglichst gut berechnen und $x = 9$ in den so berechneten Zusammenhang einsetzen.*

Definition 4.2 *Einfaches lineares Regressionsmodell*
Sei Y eine interessierende **Zielgröße** *und X eine deterministische* **Einflussgröße**. *Ein Modell der Form*

$$y = a \cdot x + b + \varepsilon,$$

das einen Zusammenhang zwischen y und x beschreibt, heißt **einfaches lineares Regressionsmodell**.
Dabei wird der Fehler ε als zufällige Störung des Zusammenhangs verstanden, ε enthält keine Systematik, und der mittlere Fehler ist Null. Dieser zufällige Fehler ε wird auch **Störgröße** *genannt.*
Die Konstanten a und b heißen **Regressionskoeffizienten**.

Bemerkung 4.3 *Bestimmung der Regressionskoeffizienten*
Betrachtet werden unabhängige Beobachtungen $y_1, \ldots, y_n$ von Y, zusammen

mit zugehörigen Werten $x_1, \ldots, x_n$ der Einflussgröße X, so dass alle Paare (x_i, y_i) dem gleichen einfachen linearen Regressionsmodell gemäß Definition 4.2 folgen.
Die Koeffizienten a und b der Modellgleichung werden aus den Beobachtungen bestimmt gemäß

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2},$$

$$\widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x}.$$

Die Schreibweise $\widehat{a}$, $\widehat{b}$ signalisiert dabei, dass es sich um die aus den Daten bestimmten Werte handelt.
Die Werte $\widehat{y}_i = \widehat{a} \cdot x_i + \widehat{b}$ werden **Vorhersagen** *oder* **Prognosen** *für die y_i genannt.*
Die Abweichungen $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ heißen **Residuen**.

Bemerkung 4.4 *Kleinste Quadrate (KQ) Methode*
Die berechneten Koeffizienten $\widehat{a}$, $\widehat{b}$ aus Bemerkung 4.3 sind so konstruiert, dass sie die Residuenquadratsumme minimieren:

$$\widehat{a}, \widehat{b} \text{ so, dass } \sum_{i=1}^{n}(y_i - a \cdot x_i - b)^2 \text{ minimal für } a = \widehat{a}, b = \widehat{b}$$

Daher heißt die Methode zur Bestimmung von $\widehat{a}$, $\widehat{b}$ auch die **Kleinste Quadrate (KQ) Methode**.

Beispiel 4.5 *Gewinn in Abhängigkeit der Menge*
Mit den Daten aus Beispiel 4.1 erhält man mit der KQ Methode für a und b:

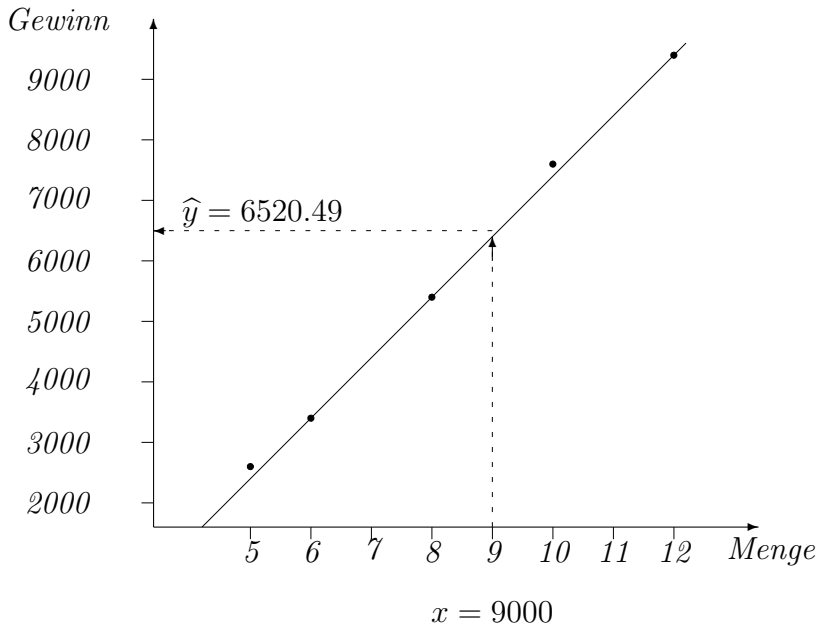
$$\widehat{a} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = 986.86,$$

$$\widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x} = -2361.25$$

Für eine produzierte Menge von 9000 Stück prognostiziert man also, dass ein Gewinn von

$$\widehat{y} = 986.86 \cdot 9 - 2361.25 = 6520.49 \text{ } \textit{Euro}$$

erwirtschaftet wird.



Bemerkung 4.6 *Bestimmtheitsmaß*

Im einfachen linearen Regressionsmodell wird die Güte der Anpassung der Daten durch die berechnete Gerade beurteilt durch das **Bestimmtheitsmaß** R^2 *mit*

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Das Bestimmtheitsmaß nimmt Werte von 0 bis 1 an, also $0 \leq R^2 \leq 1$. *Je näher* R^2 *an 1, desto besser ist die Anpassung.*

Beachte: Es gilt $R^2 = r_{XY}^2$, *d.h. das Bestimmtheitsmaß entspricht dem quadrierten Korrelationskoeffizienten.*

5 Analyse zeitlicher Verläufe

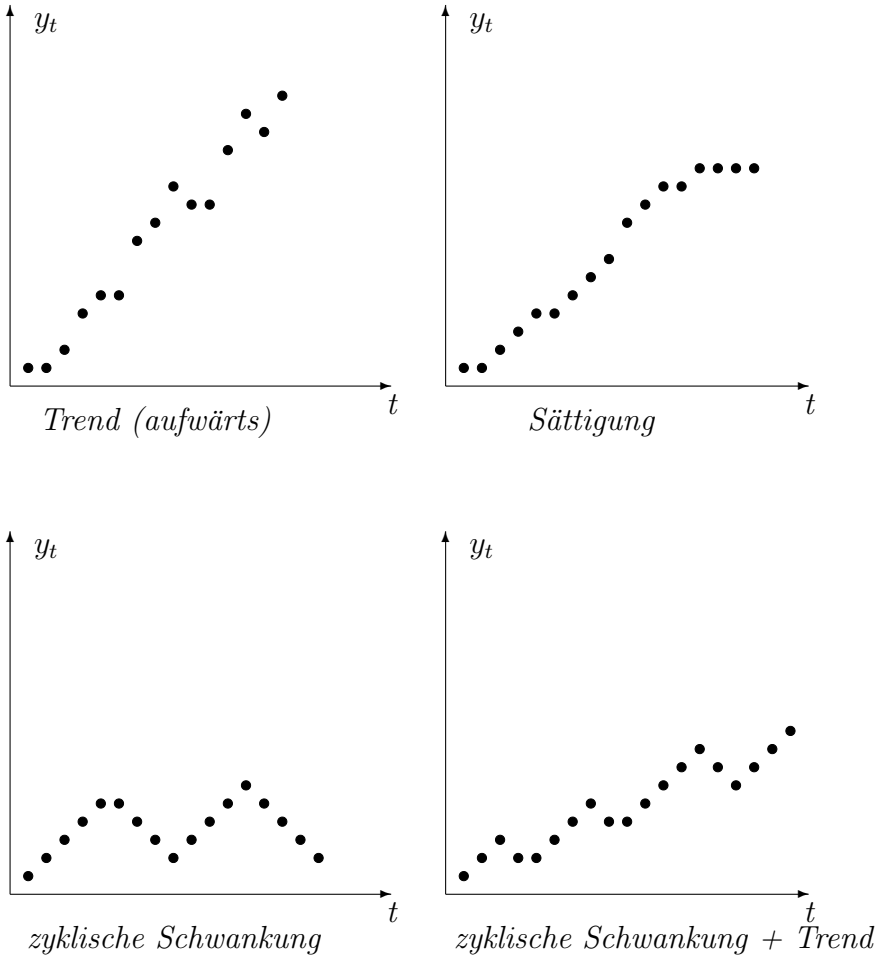
5.1 Zeitreihen

Definition 5.1 *Zeitreihe*

Wird ein Merkmal Y zu aufeinander folgenden Zeitpunkten oder Zeitperioden $t = 1, \ldots, T$ erfasst, so bilden die zugehörigen Beobachtungen $y_1, \ldots, y_T$ eine sogenannte (beobachtete) **Zeitreihe**.

Beispiel 5.2 *Zeitreihen*

Die folgenden Abbildungen zeigen Verläufe, wie sie bei Zeitreihen typischerweise beobachtet werden:



Bemerkung 5.3 *Komponentenmodelle für Zeitreihen*
Zur Beschreibung einer Zeitreihe unterstellt man oft, dass die Reihe sich zusammensetzt aus verschiedenen Komponenten, die sich überlagern. Das einfachste Komponentenmodell unterstellt, dass es eine **Trendkomponente**, *eine* **Saisonkomponente** *und eine* **irreguläre Komponente** *gibt. Dabei ist die Trendkomponente g verantwortlich für das langfristige Verhalten der Zeitreihe, die Saisonkomponente s für wiederkehrende zyklische Schwankungen und die irreguläre Komponente ε für den durch die beiden anderen nicht erklärten Rest.*
Man unterscheidet

- *das additive Modell:*

$$y_t = g_t + s_t + \varepsilon_t, \quad t = 1, \ldots, T.$$

  Hier überlagern sich die einzelnen Komponenten additiv.

- *das multiplikative Modell:*

$$y_t = g_t \cdot s_t \cdot \varepsilon_t, \quad t = 1, \ldots, T.$$

  Die multiplikative Überlagerung kann durch Logarithmieren in eine additive Überlagerung der logarithmierten Komponenten zurück geführt werden:

$$\log(y_t) = \log(g_t) + \log(s_t) + \log(\varepsilon_t), \quad t = 1, \ldots, T.$$

Nimmt man an, dass sich eine Zeitreihe durch ein solches Komponentenmodell darstellen lässt, so bestehen wichtige Ziele darin,

- die Trendkomponente zu schätzen
- die Reihe von der Saisonkomponente zu bereinigen

- die Zeitreihe zu glätten, d.h. gleichzeitig irreguläre Schwankungen auszuschalten und den Trend zu bestimmen.

Bemerkung 5.4 *Linearer Trend*
Für Zeitreihen ohne erkennbare Saisonkomponente, etwa Tagesdaten von Aktienkursen oder jährliche Preisindizes, unterstellt man oft ein reines Trendmodell der Form $y_t = g_t + \varepsilon_t$. Ist die Trendkomponente linear im zeitlichen Verlauf, so spricht man von einem **linearen Trendmodell***:*

$$y_t = a \cdot t + b + \varepsilon_t, \quad t = 1, \ldots, T.$$

In diesem Fall kann man zur Bestimmung der Trendkomponente auf die KQ-Methode des einfachen linearen Regressionsmodells zurück greifen.

Solche globalen Ansätze sind für längere Zeitreihen oft zu starr, da zeitlich sich verändernde Strukturen nicht berücksichtigt werden können. Möchte man daher statt einer so starren Modellierung über ein Komponentenmodell nur die hauptsächlichen Charakteristika des Verlaufs einer Zeitreihe heraus arbeiten, spricht man auch von der **Glättung** einer Zeitreihe. Dazu verwendet man Methoden, die die Zeitreihe in kleine Abschnitte zerlegen und nur lokal eine Anpassung vornehmen
$\rightarrow$ z.B. sogenannte **gleitende Durchschnitte**.

Bemerkung 5.5 *Einfacher gleitender Durchschnitt*
Betrachtet wird eine Zeitreihe $y_1, \ldots, y_T$, zu den Zeitpunkten $t = 1, \ldots, T$.
Die einfachste Möglichkeit, die Zeitreihe zu glätten, besteht darin, den Trend g_t zum Zeitpunkt t durch ein lokales arithmetisches Mittel der Zeitreihenwerte $y_{t-q}, \ldots y_{t+q}$ zu approximieren:

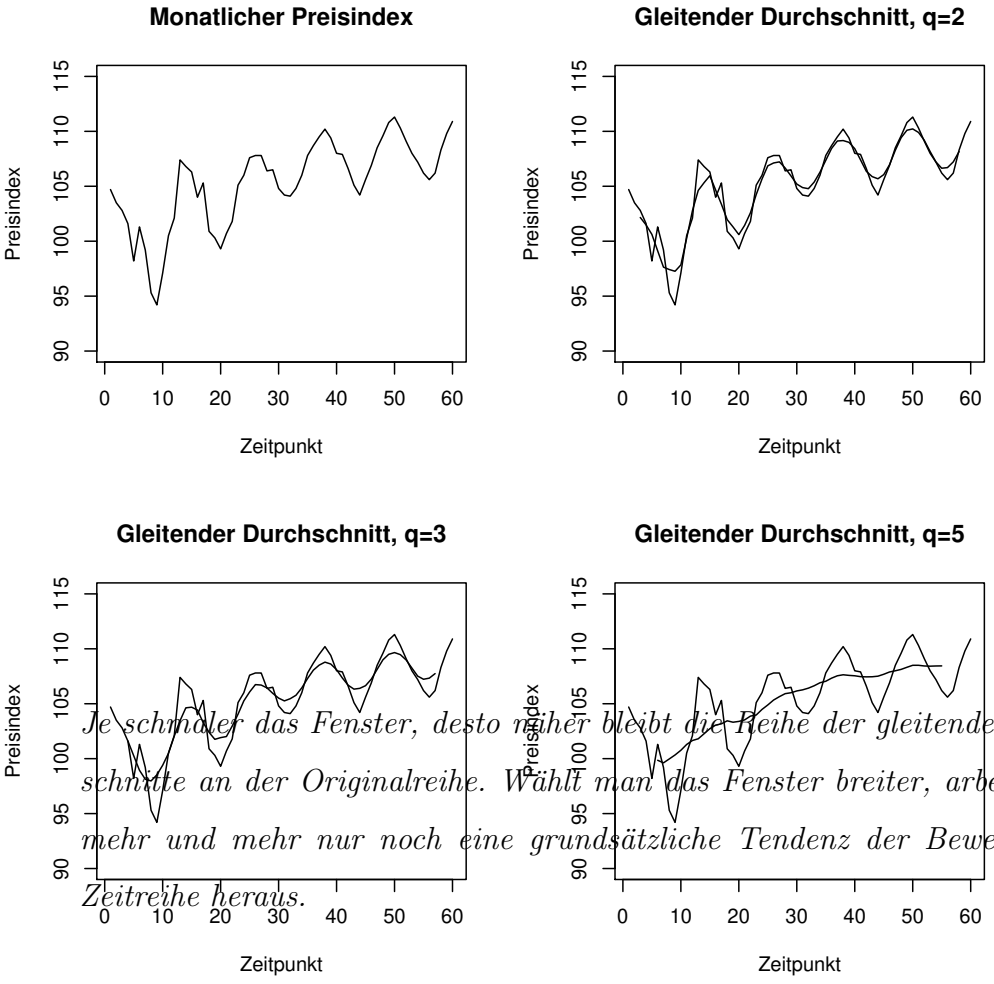
$$\widehat{g}_t = \frac{1}{2 \cdot q + 1} \sum_{j=-q}^{q} y_{t+j} = \frac{1}{2 \cdot q + 1} \cdot (y_{t-q} + \ldots + y_t + \ldots + y_{t+q}),$$

$t = q + 1, \ldots, T - q$.

Der Durchschnitt wird also aus einem "Fenster" aus q Zeitreihenwerten vor dem Zeitpunkt t, q Zeitreihenwerten nach dem Zeitpunkt t und dem Wert der Zeitreihe zum Zeitpunkt t selbst gebildet. Führt man dies für alle Zeitpunkte $t = q + 1, \ldots, T - q$ durch, so erhält man eine (verkürzte) Zeitreihe solcher Durchschnitte, die auf überlappenden Fenstern basieren. Diese geglättete Zeitreihe spiegelt den wesentlichen Verlauf der Ursprungszeitreihe wider.

Beispiel 5.6 *Gleitende Durchschnitte*

Für die Zeitreihe eines monatlichen Preisindex sind in der folgenden Abbildung drei Reihen von gleitenden Durchschnitten gezeigt, wobei die "Fenster" immer breiter werden. Im ersten Fall werden die Durchschnitte jeweils über 5 Beobachtungen gebildet, im zweiten Fall über 7 Beobachtungen, im dritten Fall über 11 Beobachtungen. Man sieht, dass durch die Glättung an den Enden der Reihe Werte verloren gehen. Die geglättete Reihe ist am Anfang und am Ende um jeweils q Werte kürzer als die Originalreihe.



Je schmaler das Fenster, desto näher bleibt die Reihe der gleitenden Durchschnitte an der Originalreihe. Wählt man das Fenster breiter, arbeitet man mehr und mehr nur noch eine grundsätzliche Tendenz der Bewegung der Zeitreihe heraus.

5.2 Indexzahlen

Indexzahlen vergleichen Maßzahlen desselben Merkmals zu verschiedenen Zeitpunkten, dem sogenannten Basiszeitpunkt (oder der Basisperiode) und

dem sogenannten Berichtszeitpunkt (oder der Berichtsperiode). Sie geben Auskunft darüber, wie sich die Maßzahl (z.B. Preis, Menge, Umsatz) zum Berichtszeitpunkt im Vergleich mit ihrem Wert zum Basiszeitpunkt verändert hat.

Beispiel 5.7 *Konsum*

Ein Konsument möchte wissen, wie sich seine Ausgaben für den Konsum gewisser "Luxusartikel" im Laufe der Jahre verändert haben. Dazu hat er die folgenden verbrauchten Mengen und zugehörigen Preise notiert:

	2000		*2005*	
	Preis	*Menge*	*Preis*	*Menge*
Zigaretten	*4*	*10*	*5*	*7*
Pizza	*5*	*4*	*6*	*3*
Kino	*8*	*2*	*12*	*1*
Bier	*0.6*	*10*	*1*	*8*

Damit kann er berechnen, wie hoch seine Gesamtausgaben in den betrachteten Jahren waren:

2000: 10 × Zigaretten zu 4 Euro + 4 × Pizza zu 5 Euro + 2 × Kino zu 8 Euro + 10 × Bier zu 0.6 Euro $=10 \cdot 4 + 4 \cdot 5 + 2 \cdot 8 + 10 \cdot 0.6 = 82$ *Euro*

2005: $7 \cdot 5 + 3 \cdot 6 + 1 \cdot 12 + 8 \cdot 1 = 73$ *Euro*

Damit kann er seine Ausgaben von 2005 relativ zu denen von 2000 betrachten:

$$\frac{\textit{Gesamtausgaben } 2005}{\textit{Gesamtausgaben } 2000} = \frac{73}{82} = 0.89,$$

d.h. seine Ausgaben haben 2005 nur 89% der Ausgaben von 2000 betragen.

Definition 5.8 *Umsatzindex*

Betrachtet wird ein Warenkorb aus n Gütern. Zu einer Basiszeit 0 haben die Güter die Preise $p_0(1), \ldots, p_0(n)$*, und der Warenkorb enthält die Mengen* $q_0(1), \ldots, q_0(n)$ *der Güter. Zu einer Berichtszeit t gelten entsprechend die Preise* $p_t(1), \ldots, p_t(n)$ *und Mengen* $q_t(1), \ldots, q_t(n)$*.*

Der **Umsatzindex** *ist definiert als*

$$U_{0,t} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)} \cdot 100\ (\%).$$

Beispiel 5.9 *Konsum: Fortsetzung*

Möchte der Konsument aus Beispiel 5.7 nun bestimmen, wie teuer das, was er 2000 verbraucht hat, 2005 gewesen wäre, so rechnet er:

10 × Zigaretten zu 5 Euro + 4 × Pizza zu 6 Euro + 2 × Kino zu 12 Euro + 10 × Bier zu 1 Euro $= 10 \cdot 5 + 4 \cdot 6 + 2 \cdot 12 + 10 \cdot 1 = 108$ *Euro*

D.h.: er nimmt die Mengen von 2000, aber die Preise von 2005. Den berechneten Wert kann er in Beziehung setzen zu den tatsächlichen Gesamtausgaben im Jahr 2000:

$$\frac{\textit{Hypothetische Ausgaben } 2005 \textit{ bei Mengen von } 2000}{\textit{Tatsächliche Ausgaben } 2000} = \frac{108}{82} = 1.32 = 132\%.$$

Die in der Basiszeit (2000) verbrauchten Waren wären demnach in der Berichtszeit (2005) um 32% teurer.

Anders herum könnte sich der Konsument auch fragen, wie hoch seine Ausgaben 2005 gewesen wären, hätten die Waren noch die gleichen Preise wie 2000 gehabt. Dann rechnet er:

7 × Zigaretten zu 4 Euro + 3 × Pizza zu 5 Euro + 1 × Kino zu 8 Euro + 8 × Bier zu 0.6 Euro $= 7 \cdot 4 + 3 \cdot 5 + 1 \cdot 8 + 8 \cdot 0.6 = 55.80$ *Euro*

D.h.: es werden die Mengen von 2005, aber die Preise von 2000 genommen. Dazu setzt man die tatsächlichen Ausgaben von 2005 in Beziehung:

$$\frac{\textit{Tatsächliche Ausgaben } 2005}{\textit{Hypothetische Ausgaben } 2005 \textit{ bei Preisen von } 2000} = \frac{73}{55.80} = 1.31 = 131\%.$$

Die in der Berichtszeit verbrauchten Waren sind um 31% teurer, als wenn man sie in der Basiszeit verbraucht hätte.

Die beiden im Beispiel gesehenen unterschiedlichen Ansätze, einen hypothetischen Wert für die Ausgaben in Beziehung zu setzen zu den tatsächlichen Ausgaben, stehen für den Preisindex nach Laspeyres und den Preisindex nach Paasche.

Definition 5.10 *Preisindex nach Laspeyres*
Unter den gleichen Voraussetzungen wie in Definition 5.8 ist der
Preisindex nach Laspeyres *definiert als*

$$P_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_0(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_0(i)} \cdot 100\ (\%).$$

Er ist der Quotient aus den hypothetischen Gesamtausgaben für den Warenkorb zur Berichtszeit t bei Verwendung der Mengen aus der Basiszeit 0 und den tatsächlichen Gesamtausgaben zur Basiszeit 0.

Definition 5.11 *Preisindex nach Paasche*
Unter den gleichen Voraussetzungen wie in Definition 5.8 ist der
Preisindex nach Paasche *definiert als*

$$P_{0,t}^{P} = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_t(i)} \cdot 100\ (\%).$$

Er ist der Quotient aus den Gesamtausgaben zur Berichtszeit t und den hypothetischen Gesamtausgaben für den Warenkorb der Berichtszeit zu Preisen der Basiszeit 0.

Bemerkung 5.12 *Preisindizes: Eigenschaften*

- *Grundsätzlich: Für die Basiszeit selbst als Berichtszeit haben die Indizes den Wert 100.*
- *Problem des Preisindex nach Laspeyres: Da der Warenkorb aus der Basiszeit betrachtet wird, müssen zum Zeitpunkt t (Berichtszeit) eventuell*

Güter mit Preisen bewertet werden, die zu dieser Zeit gar nicht mehr auf dem Mark sind (etwa: 5 1/4 Zoll Disketten im Jahr 2002)
→ veralteter Warenkorb.

- *Problem des Preisindex nach Paasche: Da der Warenkorb aus der Berichtszeit betrachtet wird, müssen zum Zeitpunkt* 0 *(Basiszeit) eventuell Preise für Güter angesetzt werden, die es zu dieser Zeit noch nicht gab (etwa: Taschenrechner im Jahr 1960)*

Preisindizes beschreiben, wie sich unter Berücksichtigung der nachgefragten Mengen die Preise entwickelt haben. Im Unterschied dazu geben Mengenindizes die durchschnittliche Mengenänderung unter Berücksichtigung der entsprechenden Preise an. Auch hier unterscheidet man nach den Definitionen von Laspeyres und Paasche.

Definition 5.13 *Mengenindizes*
Mit den Bezeichnungen aus Definition 5.8 ist der
Mengenindex nach Laspeyres *definiert durch*

$$Q_{0,t}^{L} = \frac{\sum_{i=1}^{n} q_t(i) \cdot p_0(i)}{\sum_{i=1}^{n} q_0(i) \cdot p_0(i)} \cdot 100\,(\%).$$

und der **Mengenindex nach Paasche** *durch*

$$Q_{0,t}^{P} = \frac{\sum_{i=1}^{n} q_t(i) \cdot p_t(i)}{\sum_{i=1}^{n} q_0(i) \cdot p_t(i)} \cdot 100\,(\%).$$

6 Wahrscheinlichkeiten

6.1 Der Wahrscheinlichkeitsbegriff

Bisher: deskriptive Statistik, Beschreibung von Datenmaterial

Jetzt: Grundlagen, um statistische Schlüsse zu ziehen

Basis: sog. Zufallsexperiment.

Hier zunächst: Einschränkung auf Experimente mit **endlich vielen** Ausgängen.

Definition 6.1 *Zufallsexperiment*

Ein **Zufallsexperiment** *ist ein wohldefinierter Vorgang mit mehreren möglichen Ergebnissen, bezeichnet mit* $\omega_1, \ldots, \omega_n$*; die möglichen Ergebnisse sind bekannt, aber es ist im konkreten Fall nicht klar, welches der Ergebnisse eintritt.*

Die Menge aller möglichen Ergebnisse wird mit Ω *bezeichnet (*$\Omega = \{\omega_1, \ldots, \omega_n\}$*).*

Beispiel 6.2 *Zufallsexperiment*

- *Würfeln mit einem Würfel; Ergebnismenge:* $\Omega = \{1, 2, 3, 4, 5, 6\}$
  vor konkretem Würfelwurf ist nicht bekannt, welches Ergebnis eintritt
  $\rightarrow$ *zufälliger Ausgang*

- *Schreiben einer Klausur; Ergebnismenge:* $\Omega = \{$*0 Punkte, ..., 100 Punkte*$\}$

Definition 6.3 *Ereignis*

Die Zusammenfassung von Ergebnissen eines Zufallsexperiments nennt man **Ereignis**. *Ein Ereignis* A *ist also eine Teilmenge der Ergebnismenge* Ω*:* $A \subseteq \Omega$*. Die einelementigen Teilmengen* $\{\omega_1\}, \ldots, \{\omega_n\}$ *heißen* **Elementarereignisse**.

Das Ereignis A *tritt ein, falls das realisierte Ergebnis* ω *des Zufallsexperiments in der Menge* A *enthalten ist:* $\omega \in A$*.*

Beispiel 6.4 *Ereignisse*

- *Würfeln mit einem Würfel; Ergebnismenge:* $\Omega = \{1, 2, 3, 4, 5, 6\}$
  Ereignis A: es wird eine Zahl größer als 4 gewürfelt $\rightarrow A = \{5, 6\} \subseteq \Omega$
  A tritt ein, wenn eine 5 oder eine 6 geworfen wurde, d.h. wenn das realisierte Ergebnis $\omega = 5$ *oder* $\omega = 6$ *lautet.*
  Ereignis B: es wird eine Eins oder eine Drei gewürfelt $\rightarrow B = \{1, 3\}$
  Falls also $\omega = 2 \Rightarrow B$ *tritt nicht ein.*

- *Schreiben einer Klausur; Ergebnismenge:* $\Omega = \{$*0 Punkte, ..., 100 Punkte*$\}$
  Ereignis C: die Klausur ist bestanden $\rightarrow C = \{$*50 Punkte, ..., 100 Punkte*$\}$.

Ordnet man Ereignissen Einschätzungen für ihr Eintreten zu, so spricht man von **Wahrscheinlichkeiten**.

Definition 6.5 *Wahrscheinlichkeit*
Für jedes Ereignis $A \subseteq \Omega$ *wird die Chance für das Eintreten von A bewertet, indem man A eine Zahl* $P(A)$ *zuordnet, die* **Wahrscheinlichkeit von** A. *Eine solche Wahrscheinlichkeitsabbildung P muss die sogenannten* **Kolmogoroff'schen Axiome** *erfüllen:*

(K1) $P(A) \geq 0$

(K2) $P(\Omega) = 1$ *(sicheres Ereignis)*

(K3) Falls $A \cap B = \emptyset$, *so ist* $P(A \cup B) = P(A) + P(B)$

Oft werden Wahrscheinlichkeiten mit relativen Häufigkeiten gleichgesetzt.

- Bei endlichen Grundgesamtheiten kann man relative Häufigkeit als Wahrscheinlichkeit interpretieren: ist etwa der Anteil an Frauen in der

BRD 51%, der Anteil an Männern 49%, so ist ein zufällig ausgewählter Bürger mit Wahrscheinlichkeit 0.51 weiblich

- Man führt ein Zufallsexperiment oft durch und notiert, wie oft insgesamt ein interessierendes Ereignis A eingetreten ist. Will man für ein erneutes Experiment einschätzen, mit welcher Wahrscheinlichkeit A eintreten wird, benutzt man die relative Häufigkeit des Eintretens von A aus den früheren Experimenten als Einschätzung für $P(A)$

Solche Überlegungen führen zur Bestimmung der **Laplace-Wahrscheinlichkeiten**

Definition 6.6 *Laplace-Wahrscheinlichkeiten*
Es werden Zufallsexperimente betrachtet, bei denen alle Elementarereignisse gleich wahrscheinlich sind, d.h. für $\Omega = \{\omega_1, \ldots, \omega_n\}$ *ist* $P(\{\omega_i\}) = \frac{1}{n}$, $i = 1, \ldots, n$. *Betrachtet man ein interessierendes Ereignis* $A \subseteq \Omega$, *das sich aus* m *Elementarereignissen zusammensetzt, so gilt:*

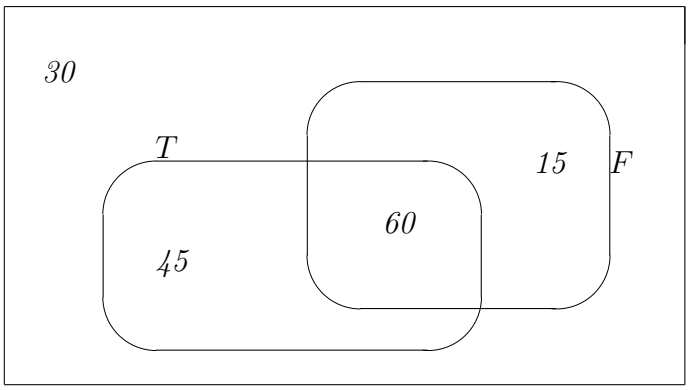
$$P(A) = \frac{m}{n}.$$

Solche Wahrscheinlichkeiten heißen **Laplace-Wahrscheinlichkeiten.**
(Faustformel: Anzahl der für A günstigen durch Anzahl der insgesamt möglichen Elementarereignisse.)

Beispiel 6.7 *Laplace-Wahrscheinlichkeit*
Razzia in einer berüchtigten Spielhölle von Las Vegas: insgesamt 150 Spieler versammelt, davon 70% Trickbetrüger (Poker), 50% Falschspieler (Black Jack) und 40% sowohl Trickbetrüger als auch Falschspieler
also Anzahl Trickbetrüger: $0.7 \cdot 150 = 105$,
Anzahl Falschspieler: $0.5 \cdot 150 = 75$ *und*
Anzahl derer, die beides sind: $0.4 \cdot 150 = 60$

damit: $105 - 60 = 45$ *reine Trickbetrüger,* $75 - 60 = 15$ *reine Falschspieler und* $150 - 60 - 45 - 15 = 30$*, die weder das eine noch das andere sind*

$\rightarrow$ *Veranschaulichung im sogenannten* **Venn-Diagramm**



Ein Spieler wird zufällig ausgewählt und festgenommen. Wahrscheinlichkeit, dass dieser Trickbetrüger oder Falschspieler ist: rechne mit Laplace

günstige Ergebnisse: der Festgenommene ist reiner Trickbetrüger, reiner Falschspieler oder beides $\Rightarrow$ *45 + 60 + 15 = 120 Personen erfüllen dies*
mögliche Ergebnisse: es gibt insgesamt 150 mögliche Festzunehmende
$\Rightarrow P($*Falschspieler oder Trickbetrüger*$) = \frac{120}{150} = 0.8$

Wahrscheinlichkeit, dass er weder Trickbetrüger noch Falschspieler, also unschuldig ist: 30 Personen erfüllen dies, also $P($*unschuldig*$) = \frac{30}{150} = 0.2$

Mit Ereignisschreibweise: sei T *das Ereignis, dass eine Person Trickbetrüger ist,* F *das Ereignis, dass sie Falschspieler ist; dann ist*
$P($*Falschspieler oder Trickbetrüger*$) = P(F \cup T)$ *(Vereinigungsmenge der beiden Ereignisse)*

und $P($*unschuldig*$) = P((F \cup T)^C)$ *(Komplementärereignis zu "Falschspieler oder Trickbetrüger")*

Allgemein benötigt man zur Darstellung von Ereignissen und deren Wahrscheinlichkeiten Mengenschreibweisen und Mengenoperationen.

Bemerkung 6.8 *Mengenoperationen*
Seien A und B Teilmengen einer Menge Ω: $A, B \subseteq \Omega$.

- **Schnittmenge:** $A \cap B = \{x : x \in A \text{ und } x \in B\}$
- **Vereinigungsmenge:** $A \cup B = \{x : x \in A \text{ oder } x \in B\}$ *(Achtung: "oder" im einschließenden Sinn; Elemente, die zu A und B gehören, gehören ebenfalls zur Vereinigungsmenge!)*
- **Differenzmenge:** $A \backslash B = \{x : x \in A \text{ und } x \notin B\}$ *("A ohne B")*
- **Komplementärmenge** *oder* **Komplement:** $A^C = \Omega \backslash A = \{x : x \notin A\}$ *(bezieht sich immer auf eine Grundmenge Ω)*
- $|A|$ *bezeichnet die* **Anzahl der Elemente** *von A*

Bemerkung 6.9 *Rechenregeln für Mengenoperationen*

- **Kommutativgesetz:**
  $A \cap B = B \cap A$, $A \cup B = B \cup A$
- **Assoziativgesetz:**
  $(A \cap B) \cap C = A \cap (B \cap C)$, $(A \cup B) \cup C = A \cup (B \cup C)$
- **Distributivgesetz:**
  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
- **de Morgan'sche Regeln:**
  $(A \cap B)^C = A^C \cup B^C$, $(A \cup B)^C = A^C \cap B^C$
- $A \subseteq B \Rightarrow B^C \subseteq A^C$
- $A \backslash B = A \cap B^C$

Bemerkung 6.10 *Rechenregeln für Wahrscheinlichkeiten*
Für eine Wahrscheinlichkeitsabbildung P und Ereignisse A, B, $A_1, \ldots, A_k$ sowie eine Grundmenge Ω von Ergebnissen gilt

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- *Falls* $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A^C) = 1 - P(A)$
- *Sind $A_1, \ldots, A_k$ paarweise disjunkt (d.h. keine zwei dieser Mengen besitzen gemeinsame Elemente), dann gilt:*
  $P(A_1 \cup \ldots \cup A_k) = P(A_1) + \ldots + P(A_k)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- *Ist Ω endlich mit Elementarereignissen $\{\omega_1\}, \ldots, \{\omega_n\}$, dann ist* $P(A) = \sum_{\omega_i \in A} P(\{\omega_i\})$,
  d.h. bei endlicher Grundgesamtheit muss man nur die Wahrscheinlichkeiten aller Elementarereignisse kennen. Daraus kann man alle anderen Wahrscheinlichkeiten berechnen.

6.2 Bedingte Wahrscheinlichkeiten und unabhängige Ereignisse

Analog zu den bedingten relativen Häufigkeiten gibt es auch bedingte Wahrscheinlichkeiten. Man bestimmt nicht mehr die einfache Wahrscheinlichkeit für das Eintreten eines Ereignisses A, sondern man betrachtet ein zweites

Ereignis B, das mit A zusammenhängt. Hat man ein Eintreten (oder Nicht-Eintreten) von B bereits beobachtet, beeinflusst dies die Wahrscheinlichkeit für das Eintreten von A.

Beispiel 6.11 *Bedingte Wahrscheinlichkeit*

Einmaliges Würfeln mit einem Würfel, Ergebnismenge $\Omega = \{\omega_1, \ldots, \omega_6\} = \{1, \ldots, 6\}$*, jedes Ergebnis hat die gleiche Wahrscheinlichkeit,*
d.h. $P(\{\omega_i\}) = 1/6$*,* $i = 1, \ldots, 6$*.*
Interessierendes Ereignis A: "gerade Zahl", d.h. $A = \{2, 4, 6\}$ *und* $P(A) = 3/6 = 0.5$

Jetzt bekommt man Zusatzinformation: Ereignis B: "Zahl kleiner gleich 3" $(B = \{1, 2, 3\})$ *ist schon eingetreten*
$\Rightarrow$ *P(gerade Zahl, wenn Zahl* ≤ 3 *gewürfelt wurde)* $= P(A|B) = 1/3$*, denn: es sind nur noch 3 Zahlen möglich (1, 2, 3), und unter den Zahlen von 1 bis 3 ist nur noch eine gerade Zahl (2)*

Man kann diese Wahrscheinlichkeit auch anders aufschreiben:
es ist $P(A \cap B)$ *= P(gerade Zahl, die zugleich* ≤ 3 *ist)* $= 1/6$ *(nur die 2 erfüllt diese Bedingung von allen 6 möglichen Ergebnissen)*
außerdem ist $P(B)$ *= P(Zahl* ≤ 3*)* $= 3/6$
zusammen ist

$$\frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = 1/3 = P(A|B)$$

Ebenso: betrachte $A^C = \{1, 3, 5\}$*, dann ist* $P(A^C) = 3/6 = 0.5$*, aber* $P(A^C|B)$ *= P(ungerade Zahl, wenn Zahl* ≤ 3 *geworfen wurde)* $= 2/3 = \frac{2/6}{3/6} = \frac{P(A^C \cap B)}{P(B)}$

Definition 6.12 *Bedingte Wahrscheinlichkeit*
Betrachtet werden zwei Ereignisse A und B, und es sei die Wahrscheinlichkeit des Eintretens von B größer als Null, also $P(B) > 0$. Die **bedingte Wahrscheinlichkeit von A gegeben B** *ist definiert als*

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

Für eine feste Wahl von B ist die Abbildung, die jedem Ereignis $A \subseteq \Omega$ die bedingte Wahrscheinlichkeit $P(A|B)$ zuordnet, wieder eine Wahrscheinlichkeitsabbildung, die die Kolmogoroff'schen Axiome erfüllt.

Die obige Formel für die bedingte Wahrscheinlichkeit kann man auch umstellen und nach $P(A \cap B)$ auflösen. Damit kann man die Wahrscheinlichkeit für das gleichzeitige Eintreten zweier Ereignisse A und B berechnen aus der Wahrscheinlichkeit für das bedingte Ereignis und der Wahrscheinlichkeit für die Bedingung.

Bemerkung 6.13 *Produktsatz für Wahrscheinlichkeiten*
Seien A, B Ereignisse und $P(B) > 0$. Dann gilt

$$P(A \cap B) = P(A|B) \cdot P(B).$$

Beispiel 6.14 *Bedingte Wahrscheinlichkeit und Produktsatz*
In der Situation von Beispiel 6.7 (Razzia in Las Vegas) gab es folgende Anzahlen:

		Falschspieler (F)		
		ja	*nein*	$\sum$
Trick-	*ja*	*60*	*45*	*105*
betrüger (T)	*nein*	*15*	*30*	*45*
	$\sum$	*75*	*75*	*150*

Wahrscheinlichkeit, bei zufälligem Herausgreifen eines Trickbetrügers einen Falschspieler zu erwischen:

$$P(F|T) = \frac{P(F \cap T)}{P(T)} = \frac{60/150}{105/150} = \frac{0.4}{0.7} = 0.57$$

Kennt man andererseits nur $P(F|T) = 0.57$ *und* $P(T) = 0.7$*, so berechnet man*

$$P(F \cap T) = P(F|T) \cdot P(T) = 0.57 \cdot 0.7 = 0.40 \quad \textit{(mit Rundung).}$$

Eine Erweiterung des Produktsatzes ergibt sich, wenn man nicht nur eine Bedingung B betrachtet, sondern k verschiedene Bedingungen $B_1, \ldots, B_k$, die zusammen genommen die Grundmenge Ω ergeben.

Beispiel 6.15 *Erweiterung des Produktsatzes*
Die heimische Fußballmannschaft "Rot-Weiß" muss im nächsten Qualifikationsspiel gegen einen von drei möglichen Gegnern antreten. Der Gegner wird dabei ausgelost. Falls der Gegner "MFC" lautet, haben die Rot-Weißen eine Gewinnchance von 0.6 (d.h. sie gewinnen mit Wahrscheinlichkeit 0.6). Müssen sie gegen "FCL" antreten, gewinnen sie mit Wahrscheinlichkeit 0.8. Nur gegen den Angstgegner "SVD" haben sie lediglich eine Chance von 0.1. Gesucht: Wahrscheinlichkeit, dass der Fußballclub "Rot-Weiß" die Qualifikation schafft.

Grundmenge = mögliche Gegner	$\Omega = \{$ *"MFC", "FCL", "SVD"* $\}$
interessierendes Ereignis	$A = \{$ *"Rot-Weiß" gewinnt* $\}$
Bedingungen:	$B_1 = \{$ *Gegner ist "MFC"* $\}$
	$B_2 = \{$ *Gegner ist "FCL"* $\}$
	$B_3 = \{$ *Gegner ist "SVD"* $\}$

Bekannte Wahrscheinlichkeiten:
$P(B_1) = P(B_2) = P(B_3) = 1/3$ *(da Gegner ausgelost wird)*

$P(A|B_1) = 0.6$, $P(A|B_2) = 0.8$, $P(A|B_3) = 0.1$ *(bekannte Gewinnchancen, wenn der Gegner bekannt ist)*

Das Ereignis, dass "Rot-Weiß" gewinnt, lässt sich nun aufsplitten in
Entweder "Rot-Weiß" gewinnt und der Gegner ist "MFC"
oder "Rot-Weiß" gewinnt und der Gegner ist "FCL"
oder "Rot-Weiß" gewinnt und der Gegner ist "SVD",

formal:

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$$

Damit

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)) \\ &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \end{aligned}$$

(da $(A \cap B_1), (A \cap B_2), (A \cap B_3)$ *disjunkte Ereignisse sind).*

Anwendung des Produktsatzes liefert

$$\begin{aligned} P(A) &= P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) + P(A|B_3) \cdot P(B_3) \\ &= 0.6 \cdot \frac{1}{3} + 0.8 \cdot \frac{1}{3} + 0.1 \cdot \frac{1}{3} = 0.5 \end{aligned}$$

Die "Rot-Weißen" qualifizieren sich also mit Wahrscheinlichkeit $1/2$.

Bemerkung 6.16 *Satz von der totalen Wahrscheinlichkeit*
Sei $B_1, \ldots, B_k$ *eine disjunkte Zerlegung von* Ω*, d.h. je zwei Mengen* B_i *und* B_j *haben keine gemeinsamen Elemente (für* $i \neq j$*), und die Vereinigung aller* B_i*,* $i = 1, \ldots k$*, ergibt die Menge* Ω*. Dann gilt:*

$$P(A) = \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i).$$

Umgekehrt kann man sich auch für die Wahrscheinlichkeit einer Teilmenge B_i aus einer solchen Zerlegung interessieren, wenn ein anderes Ereignis A eingetreten ist.

Beispiel 6.17 *Erkennung von Fälschungen*

Eine Firma hat ein Gerät auf den Markt gebracht, das Fälschungen von Euro-Scheinen mit großer Zuverlässigkeit erkennen soll.

Zu betrachten sind hier die folgenden Ereignisse:

A = *Gerät erkennt auf Fälschung*

B_1 = *Schein ist tatsächlich eine Fälschung*

$B_2 = B_1^C$ = *Schein ist tatsächlich echt*

(Beachte: B_1 *und* B_2 *bilden eine disjunkte Zerlegung der Menge aller Euro-Scheine)*

Die Firma gibt folgende Daten über die Zuverlässigkeit ihres Gerätes an:

$P(A|B_1)$= *P(Gerät erkennt auf Fälschung, falls Schein tatsächlich falsch) = 0.99*

$P(A|B_2)$= *P(Gerät erkennt auf Fälschung, falls Schein tatsächlich echt) = 0.05*

Weiterhin weiß man, dass Fälschungen nicht so häufig passieren:

$P(B_1)$ = *P(Schein ist gefälscht) = 0.002*

Nach den Daten der Herstellerfirma macht das Gerät einen zuverlässigen Eindruck. Für den Benutzer ist nun interessant: falls das Gerät eine Fälschung meldet, wie groß ist dann die Wahrscheinlichkeit, dass es sich tatsächlich um eine solche handelt?

D.h. gesucht ist $P(B_1|A)$.

Nach Definition der bedingten Wahrscheinlichkeit bestimmt man $P(B_1|A)$ *als*

$$P(B_1|A) = \frac{P(B_1 \cap A)}{P(A)}$$

Zunächst kann $P(B_1 \cap A)$ *mit dem Produktsatz bestimmt werden:*

$$P(B_1 \cap A) = P(A \cap B_1) = P(A|B_1) \cdot P(B_1) = 0.99 \cdot 0.002 = 0.00198,$$

so dass

$$P(B_1|A) = \frac{P(A|B_1) \cdot P(B_1)}{P(A)} = \frac{0.00198}{P(A)}.$$

Für $P(A)$ nutze den Satz von der totalen Wahrscheinlichkeit (da B_1 und B_2 eine disjunkte Zerlegung bilden (s.o.)):

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i) = P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) \\ &= P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot (1 - P(B_1)) \quad (da\ B_2 = B_1^C) \\ &= 0.99 \cdot 0.002 + 0.05 \cdot 0.998 = 0.05188. \end{aligned}$$

Insgesamt also:

$$P(B_1|A) = \frac{P(A|B_1) \cdot P(B_1)}{P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2)} = \frac{0.00198}{0.05188} = 0.038.$$

Das heißt: nur in 3.8% der Fälle, in denen das Gerät einen gefälschten Euro-Schein meldet, handelt es sich tatsächlich um Falschgeld! Oder mit anderen Worten: in 96.2% der Fälle handelt es sich um eine Fehldiagnose.

Die Herleitung der gesuchten Wahrscheinlichkeit aus dem Beispiel führt gerade zum sog. **Satz von Bayes**.

Bemerkung 6.18 *Satz von Bayes*
Sei $B_1, \ldots, B_k$ eine disjunkte Zerlegung von Ω, und es gebe mindestens ein j, so dass $P(B_j) > 0$ und $P(A|B_j) > 0$. Weiterhin werde ein zusätzliches Ereignis A betrachtet. Dann gilt für jedes der Ereignisse B_j, $j = 1, \ldots, k$:

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{i=1}^{k} P(A|B_i) \cdot P(B_i)} = \frac{P(A|B_j) \cdot P(B_j)}{P(A)}.$$

Bei der Betrachtung bedingter relativer Häufigkeiten wurde durch den Vergleich verschiedener bedingter Verteilungen auf den Zusammenhang zwischen Merkmalen geschlossen. Ähnlich ist man auch für Ereignisse daran interessiert, ob eine Abhängigkeit besteht.

Definition 6.19 *Unabhängige Ereignisse*
Zwei Ereignisse A und B heißen **(stochastisch) unabhängig**, *wenn*

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

Falls $P(B) = 0$, so nennt man A und B stets unabhängig.

Äquivalent zu Definition 6.19 kann man die Unabhängigkeit zweier Ereignisse auch über die bedingte Wahrscheinlichkeit formulieren: A und B sind unabhängig, falls $P(A|B) = P(A)$, d.h. falls die Information über das Eintreten von B nichts an der Wahrscheinlichkeit für das Eintreten von A ändert.

Beispiel 6.20 *Unabhängige Ereignisse*
Zweimaliges Würfeln mit einem Würfel
A = eine Eins beim ersten Wurf
B = eine Eins beim zweiten Wurf
$\Rightarrow P(A) = P(B) = 1/6$, *also* $P(A) \cdot P(B) = 1/36$

Andererseits:
Ergebnismenge beim zweimaligen Würfeln ist $\Omega = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$ *mit* $|\Omega| = 36$*; jedes Ergebnis kommt mit der gleichen Wahrscheinlichkeit* $1/36$ *vor*
$\Rightarrow P(A \cap B) = P(\{(1,1)\}) = 1/36$

Also: $P(A \cap B) = P(A) \cdot P(B)$, d.h. die Ereignisse sind unabhängig (auch intuitiv klar: Ereignis, das nur den ersten Wurf betrifft, ist unabhängig von Ereignis, das nur den zweiten Wurf betrifft).

6.3 Zufallsstichproben

In statistischen Analysen ist man oft nicht in der Situation, dass man die komplette Grundgesamtheit kennt und die Wahrscheinlichkeiten für interessierende Ereignisse exakt berechnen kann. In diesem Fall greift man auf Stichproben zurück und "schätzt" die benötigten Wahrscheinlichkeiten durch relative Häufigkeiten.
Man modelliert die Stichprobenziehung aus einer endlichen Grundgesamtheit abstrakt als das Ziehen von Kugeln aus einem Gefäß (sog. Urnenmodell). Betrachte dazu die Grundgesamtheit als N durchnummerierte Kugeln. Prinzipiell unterscheidet man zwei Möglichkeiten, an eine Stichprobe zu kommen: das **Ziehen mit Zurücklegen** und das **Ziehen ohne Zurücklegen**.

Ziehen mit Zurücklegen: dem Gefäß eine Kugel entnehmen, Nummer notieren, Kugel zurücklegen; Vorgang n-mal wiederholen
zwei Varianten möglich:

- Reihenfolge der Nummern interessiert
- es interessiert nur, welche Nummer wie oft gezogen wurde

Ziehen ohne Zurücklegen: dem Gefäß eine Kugel entnehmen, Nummer notieren, Kugel zur Seite legen; Vorgang n-mal wiederholen
gleiche zwei Varianten möglich:

- Reihenfolge der Nummern interessiert
- es interessiert nur, welche Nummern gezogen wurden

Von einer **einfachen Zufallsstichprobe** spricht man, wenn (in einem gewünschten Ziehungsmodell) jede mögliche Stichprobe vom Umfang n aus der Grundgesamtheit dieselbe Wahrscheinlichkeit hat, realisiert zu werden. Um dies zu

kontrollieren, muss man wissen, wie viele mögliche Stichproben es unter den vier obigen Ziehungsmodellen gibt.

Definition 6.21 *Fakultät und Binomialkoeffizient*
Die **Fakultät** *einer natürlichen Zahl k ist definiert als*

$$k! = 1 \cdot 2 \cdot \ldots \cdot (k-1) \cdot k,$$

wobei per Definition $1! = 1$ und $0! = 1$ gesetzt wird.
Der **Binomialkoeffizient** *aus zwei natürlichen Zahlen m und k ist definiert als*

$$\binom{m}{k} = \frac{m!}{k! \cdot (m-k)!}, \quad \text{falls } m \geq k.$$

Falls $m < k$, wird festgelegt, dass $\binom{m}{k} = 0$ gilt.
Sprechweisen: $k!$ = "k Fakultät"; $\binom{m}{k}$ = "m über k".

Bemerkung 6.22 *Anzahl möglicher Stichproben*
Gegeben sei eine Grundgesamtheit vom Umfang N, und es soll eine Stichprobe vom Umfang n daraus gezogen werden. Dann gibt es folgende Anzahlen solcher möglichen Stichproben:

	ohne Zurücklegen	*mit Zurücklegen*
mit Beachtung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
ohne Beachtung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

Modellhaft betrachtet man die Merkmale der n Stichprobenelemente als Ausgänge von n Wiederholungen eines Zufallsexperiments. Man betrachtet also die folgende Situation:

Ein Zufallsexperiment wird n-mal unabhängig wiederholt; ein interessierendes Ereignis A besitzt bei jeder Durchführung wieder dieselbe Wahrscheinlichkeit des Eintretens.

Bezeichne mit $f_n(A)$ die relative Häufigkeit des Eintretens von A bei n Wiederholungen des Experiments.
Mit wachsender Anzahl von Versuchen stabilisiert sich erfahrungsgemäß die Folge der relativen Häufigkeiten $f_n(A)$ um den Wert $P(A)$. Empirisch gilt also:

$$f_n(A) \underset{n\to\infty}{\longrightarrow} P(A)$$

Das heißt: man kann den Grenzwert der relativen Häufigkeiten eines Ereignisses A als Wahrscheinlichkeit für das Eintreten von A interpretieren. Dabei geht die Zahl der Wiederholungen gegen unendlich. In der Praxis hat man natürlich nur die Möglichkeit für endlich viele Wiederholungen. Man nutzt dann relative Häufigkeiten zur **Schätzung** der unbekannten Wahrscheinlichkeiten.

7 Zufallsvariablen

7.1 Eindimensionale diskrete und stetige Zufallsvariablen

Beispiel 7.1 *Funktionen auf dem Ergebnisraum*
Aus der Grundgesamtheit der Vorlesungsteilnehmer wird zufällig eine Person ausgewählt $\rightarrow$ *Zufallsexperiment mit* $\Omega = \{$ *Kai, Maria, Stefan, ...*$\}$, $|\Omega| = N$ *und* $P(\{\omega\}) = 1/N$ *für alle* $\omega \in \Omega$.
Die ausgewählte Person selbst interessiert aber eigentlich nicht, sondern sie wird nach ihrem Alter X *und ihrer Größe* Y *befragt* $\rightarrow$ *jedem Ergebnis* $\omega \in \Omega$ *wird durch* X *und* Y *jeweils genau eine Zahl* $X(\omega)$ *bzw.* $Y(\omega)$ *zugeordnet.*
Formal sind X *und* Y *jeweils Abbildungen des Ergebnisraums auf die reellen*

Zahlen:

$$X : \Omega \to \mathbb{R}, \quad Y : \Omega \to \mathbb{R}$$

Solche Abbildungen nennt man **Zufallsvariablen**. *Dabei ist die Abbildung selbst nicht zufällig, aber da man das Ergebnis des Zufallsexperiments nicht im Vorhinein kennt, kennt man auch vorab nicht den Wert der Zufallsvariablen, worin sich die Zufälligkeit widerspiegelt.*

Definition 7.2 *Diskrete und stetige Zufallsvariable*
Sei Ω die Ergebnismenge eines Zufallsexperiments. Eine Abbildung
$X : \Omega \to \mathbb{R}$ heißt **Zufallsvariable**.
Eine Zufallsvariable ist demnach ein Merkmal, dessen Ausprägungen als Ergebnisse eines Zufallsexperiments resultieren.
Die Menge $X(\Omega) = \{x \in \mathbb{R} : x = X(\omega) \text{ mit } \omega \in \Omega\}$ heißt **Wertebereich** *von X.*
Enthält $X(\Omega)$ nur endlich oder abzählbar viele Werte (d.h. die Elemente von $X(\Omega)$ sind nummerierbar), so heißt X eine **diskrete Zufallsvariable**.
Enthält $X(\Omega)$ alle möglichen Werte eines Intervalls, so heißt X eine **stetige Zufallsvariable**.

Beispiel 7.3 *Zufallsvariablen*

Variable	*Beschreibung*	*Typ*	*Wertebereich*
X_1	*Augensumme beim zweimaligen Würfeln*	*diskret*	$\{2, 3, \ldots, 12\}$
X_2	*Anzahl von "Kopf" beim dreimaligen Münzwurf*	*diskret*	$\{0, 1, 2, 3\}$
X_3	*Lebensdauer einer ausgewählten Glühbirne*	*stetig*	$[0, \infty)$
X_4	*logarithmierte Kursänderung einer Aktie an einem zufällig ausgewählten Börsentag*	*stetig*	$(-\infty, \infty)$

Beachte: auch Funktionen von Zufallsvariablen sind selbst wieder Zufallsvariablen!
Etwa: Differenz zwischen Lebensdauer einer ausgewählten Glühbirne und der laut Hersteller garantierten Mindestlebensdauer

7.2 Verteilungsfunktion und Dichte

Die in der deskriptiven Statistik benutzten Kenngrößen für die Häufigkeitsverteilung von Merkmalen finden ihre Gegenstücke in den entsprechenden Größen für Zufallsvariablen.

Definition 7.4 *Verteilungsfunktion*
Gegeben sei eine Zufallsvariable X. Zu einer vorgegebenen Zahl x betrachte die Wahrscheinlichkeit, dass X einen Wert kleiner oder gleich x annimmt. Die Funktion F, die diese Wahrscheinlichkeit in Abhängigkeit von x beschreibt, heißt die **Verteilungsfunktion** *von X:*

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

Dabei steht die Schreibweise $X \leq x$ abkürzend für das Ereignis $\{\omega \in \Omega : X(\omega) \leq x\}$.

In engem Zusammenhang mit der Verteilungsfunktion steht die **Dichtefunktion**, die das Pendant zur relativen Häufigkeitsverteilung darstellt. Man unterscheidet bei der Definition der Dichte den Fall der diskreten und der stetigen Zufallsvariablen.

Definition 7.5 *Dichte*

1. *Sei X diskrete Zufallsvariable mit Wertebereich $X(\Omega) = \{x_1, x_2, x_3, \ldots\}$ (endlich oder abzählbar unendlich). Die* **diskrete Dichte** *von X ist die Funktion f, so dass für die Verteilungsfunktion F von X gilt:*

$$F(x) = \sum_{x_i \leq x} f(x_i).$$

Dabei kann man die Funktionswerte der diskreten Dichtefunktion angeben als

$$f(x_i) = P(X = x_i), i = 1, 2, \ldots,$$

und es gilt: $f(x_i) \geq 0$ für alle i, $\sum_{i=1}^{\infty} f(x_i) = 1$.
Daraus folgt automatisch, dass $f(x_i) \leq 1$ für alle i.

2. *Sei X stetige Zufallsvariable mit Wertebereich $X(\Omega) = \mathbb{R}$. Die* **stetige Dichte** *von X ist die Funktion f, so dass für die Verteilungsfunktion F von X gilt:*

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Dabei gilt: $f(x) \geq 0$ für alle x, $\int_{-\infty}^{\infty} f(t)dt = 1$.
Daraus folgt **nicht**, *dass immer $f(x) \leq 1$ sein muss!*

Ähnlich wie die relative Häufigkeitsverteilung lässt sich auch die diskrete Dichte mittels eines Stabdiagramms darstellen.

Beispiel 7.6 *Diskrete Dichte und Verteilungsfunktion*
Einmaliges Würfeln, X = gewürfelte Augenzahl;
es ist $X(\Omega) = \{x_1, \ldots, x_6\} = \{1, \ldots, 6\}$ und $P(X = x_i) = 1/6$, $i = 1, \ldots, 6$.

Damit ist die diskrete Dichte von X gegeben als

$$f(x_i) = 1/6, \quad i = 1, \ldots, 6.$$

Weiterhin lassen sich die Werte der Verteilungsfunktion bestimmen als

x_i	*1*	*2*	*3*	*4*	*5*	*6*
$f(x_i)$	*1/6*	*1/6*	*1/6*	*1/6*	*1/6*	*1/6*
$F(x_i)$	*1/6*	*2/6*	*3/6*	*4/6*	*5/6*	*6/6 = 1*

Damit kann man zum Beispiel die Wahrscheinlichkeit bestimmen, bei einem Wurf eine Zahl größer als 1 aber kleiner oder gleich 3 zu werfen:

$$P(1 < X \leq 3) = P(X = 2) + P(X = 3) = f(2) + f(3) = 1/6 + 1/6 = 2/6$$

oder:

$$P(1 < X \leq 3) = P(X \leq 3) - P(X \leq 1) = F(3) - F(1) = 3/6 - 1/6 = 2/6$$

Bemerkung 7.7 *Dichte und Verteilungsfunktion*

1. *Die Verteilungsfunktion ist das Gegenstück zur empirischen Verteilungsfunktion. Für eine diskrete Zufallsvariable sieht die Verteilungsfunktion auch ähnlich aus: es ist eine Treppenfunktion mit Sprüngen an den Stellen x_i und Sprunghöhen $f(x_i) = P(X = x_i)$.*

2. *Für eine* **diskrete** *Zufallsvariable X gilt:*

   - *Die Wahrscheinlichkeit dafür, dass X Werte in einem bestimmten Bereich annimmt, erhält man durch Aufsummieren der Elementarwahrscheinlichkeiten $P(X = x_i)$ für alle x_i aus dem interessierenden Bereich. Etwa:*

$$P(a < X \leq b) = \sum_{x_i : a < x_i \leq b} P(X = x_i).$$

- *Eine solche Wahrscheinlichkeit kann auch mit Hilfe der Verteilungsfunktion bestimmt werden:*

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

3. *Für eine* **stetige** *Zufallsvariable X gilt:*

   - *Der Wert $F(x)$ entspricht der Fläche unter der Kurve der stetigen Dichtefunktion f bis zur Stelle x. Weiterhin lässt sich die stetige Dichte als Ableitung der Verteilungsfunktion schreiben (falls die Ableitung existiert):*

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

   - *$P(X = x) = 0$ für $x \in \mathbb{R}$ fest, d.h. für eine stetige Zufallsvariable ist die Wahrscheinlichkeit, einen bestimmten Wert anzunehmen (sog. Punktwahrscheinlichkeit), gleich Null.*

   - *Die Wahrscheinlichkeit dafür, dass X Werte in einem bestimmten Bereich annimmt, erhält man durch Integration über die Dichte im interessierenden Bereich. Etwa:*

$$P(a < X \leq b) = \int_a^b f(t)dt.$$

     Dies entspricht der Fläche unter der Dichtekurve zwischen den Stellen a und b.

   - *Eine solche Wahrscheinlichkeit kann auch mit Hilfe der Verteilungsfunktion bestimmt werden:*

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

$$= P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Beispiel 7.8 *Stetige Zufallsvariable*

Gegeben sei eine stetige Zufallsvariable mit folgender Dichtefunktion:

$$f(x) = \begin{cases} 0.5, & 0 \leq x \leq 1 \\ 0.25, & 1 < x \leq 3 \\ 0, & \textit{sonst} \end{cases}$$

Überprüfen, ob es sich bei f tatsächlich um eine Dichtefunktion handelt: dazu feststellen, ob $f(x) \geq 0$ und ob $\int_{-\infty}^{\infty} f(t)dt = 1$ gilt.

Offensichtlich ist $f(x) \geq 0$; außerdem

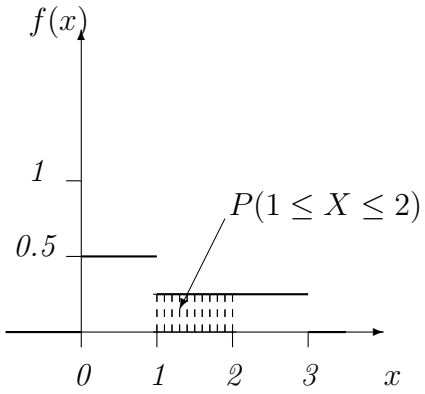
$$\begin{aligned} \int_{-\infty}^{\infty} f(t)dt &= \int_0^3 f(t)dt = \int_0^1 0.5dt + \int_1^3 0.25dt \\ &= 0.5 \cdot t\Big|_0^1 + 0.25 \cdot t\Big|_1^3 \\ &= (0.5 \cdot 1 - 0) + (0.25 \cdot 3 - 0.25 \cdot 1) = 0.5 + 0.5 = 1 \end{aligned}$$

Damit handelt es sich um eine Dichtefunktion.

Weiterhin ist z.B.

$$P(1 \leq X \leq 2) = \int_1^2 f(t)dt = \int_1^2 0.25dt = 0.25 \cdot t\Big|_1^2 = 0.5 - 0.25 = 0.25$$

Skizze:



7.3 Lage- und Streuungsparameter

Analog zu den in der deskriptiven Statistik definierten Lage- und Streuungsmaßen definiert man für Wahrscheinlichkeitsverteilungen bzw. Zufallsvariablen sogenannte Lage- und Streuungsparameter.

Definition 7.9 *Modus*
Für eine Zufallsvariable X ist der **Modus** *der Verteilung von X derjenige x-Wert x_{mod}, für den die Dichte $f(x)$ von X maximal wird. Gibt es keinen eindeutigen x-Wert, der dies erfüllt, so ist der Modus nicht definiert.*

Definition 7.10 *Erwartungswert*
Betrachtet wird eine Zufallsvariable X mit Dichtefunktion f.

1. *Ist X diskrete Zufallsvariable, so ist der* **Erwartungswert** *$E(X)$ von X das mit den Werten $f(x_i)$ der Dichte gewichtete Mittel der möglichen Werte x_i von X:*

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot f(x_i) = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \dots$$

2. *Ist X stetige Zufallsvariable, so ist der* **Erwartungswert** *$E(X)$ von X definiert als*

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

Beispiel 7.11 *Erwartungswert*

1. *X = Augenzahl beim einmaligen Würfeln*

   es war $f(x_i) = 1/6$, $x_i = 1, \dots, 6$

   damit ist

   $E(X) = \sum_{i=1}^{\infty} x_i \cdot f(x_i) = \sum_{i=1}^{6} x_i \cdot f(x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$

   Das heißt: Wenn man das Experiment "Würfeln mit einem Würfel"

häufig wiederholt, erwartet man, dass die mittlere geworfene Augenzahl bei 3.5 liegt.

2. *X = Wartezeit auf die Straßenbahn bei zufälliger Ankunft an der Haltestelle, wenn regelmäßig alle 5 Minuten eine Bahn kommt*

   man kann annehmen, dass die Dichte von X folgende Gestalt hat: $f(x) = \frac{1}{5}$ *für* $0 \leq x \leq 5$ *(und* $f(x) = 0$ *sonst)*
   damit:
   $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_0^5 x \cdot f(x) dx = \int_0^5 x \cdot \frac{1}{5} dx = \frac{1}{10} x^2 \Big|_0^5 = \frac{25}{10} = 2.5$
   Das heißt: Wenn man häufig zur Haltestelle geht, ohne sich mit dem Fahrplan auseinander zu setzen, wartet man im Schnitt 2.5 Minuten, bis die nächste Bahn kommt.

Bemerkung 7.12 *Eigenschaften und Rechenregeln*

- *Der Erwartungswert existiert nicht immer. Es kann Dichten geben, so dass Summe bzw. Integral von $x \cdot f(x)$ nicht endlich ist. In diesem Fall sagt man, dass $E(X)$ nicht existiert.*

- *Der Erwartungswert ist das theoretische Gegenstück zum arithmetischen Mittel. Man kann $E(X)$ interpretieren als den "Schwerpunkt" der Dichte, d.h. als die Stelle, an der man die Dichtefunktion unterstützen müsste, um sie im Gleichgewicht zu halten.*

- *Ist die Dichtefunktion f von X symmetrisch um eine Stelle a, d.h. $f(a + x) = f(a - x)$ für alle x, dann ist $E(X) = a$.*

- *Transformiert man die Zufallsvariable X linear, d.h. man betrachtet $Y = a \cdot X + b$ für Konstanten a, b, so gilt:*
  $$E(Y) = E(a \cdot X + b) = a \cdot E(X) + b$$
  (sog. **Linearität des Erwartungswerts***).*

- *Transformiert man die Zufallsvariable X mit einer beliebigen Funktion g, d.h. man betrachtet $Y = g(X)$, so gilt:*

$$E(Y) = E(g(X)) = \sum_{i=1}^{\infty} g(x_i) \cdot f(x_i),$$

falls X diskrete Zufallsvariable, bzw.

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

falls X stetige Zufallsvariable ist.

Definition 7.13 *Quantile*
Sei $p \in (0, 1)$. Analog zur Definition aus der deskriptiven Statistik bestimmt man auch für Zufallsvariablen bzw. deren Verteilungen **p-Quantile**. *Dabei spricht man auch hier für $p = 0.5$ vom* **Median** *und für $p = 0.25$ bzw. $p = 0.75$ vom* **unteren** *bzw.* **oberen Quartil**.

1. *Für eine diskrete Zufallsvariable X heißt eine Zahl x_p* **(theoretisches) p-Quantil**, *wenn gilt:*

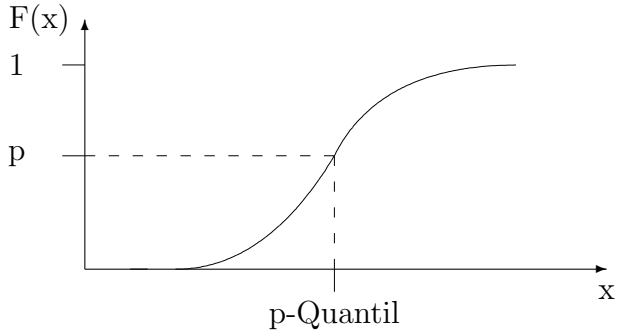
$$P(X < x_p) \leq p \quad \textit{und} \quad P(X > x_p) \leq 1 - p.$$

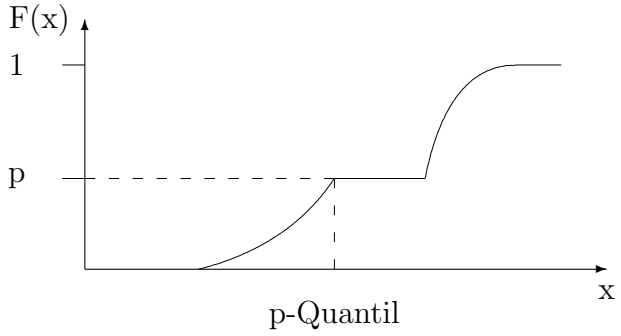
Falls x_p aus dieser Beziehung nicht eindeutig bestimmbar ist, wählt man den kleinsten Wert, der diese Bedingung erfüllt.

2. *Für eine stetige Zufallsvariable X heißt eine Zahl x_p* **(theoretisches) p-Quantil**, *wenn gilt:*

$$F(x_p) = p.$$

Auch hier wählt man bei Nicht-Eindeutigkeit den kleinsten Wert x_p, der dies erfüllt.





Zur Beurteilung der Streuung einer Zufallsvariablen benutzt man wie im deskriptiven Fall Varianz und Standardabweichung.

Definition 7.14 *Varianz, Standardabweichung*
Sei X eine Zufallsvariable mit Dichtefunktion f, und der Erwartungswert $E(X)$ existiere.

1. *Ist X diskret, so ist die* **Varianz** *von X definiert durch*

$$Var(X) = E\left((X - E(X))^2\right) = \sum_{i=1}^{\infty}(x_i - E(X))^2 \cdot f(x_i).$$

Die Varianz wird oft mit σ^2 ("sigma-Quadrat") benannt.
Die Größe $\sigma_X = \sqrt{Var(X)}$ heißt **Standardabweichung** *von X.*

2. *Ist X stetig, so ist die* **Varianz** *von X definiert durch*

$$Var(X) = E\left((X - E(X))^2\right) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\, dx,$$

und $\sigma_X = \sqrt{Var(X)}$ heißt **Standardabweichung** *von X.*

Bemerkung 7.15 *Eigenschaften und Rechenregeln*

- *In der Formel für die Varianz einer diskreten Zufallsvariable,*

$$Var(X) = \sum_{i=1}^{\infty} (x_i - E(X))^2 \cdot f(x_i),$$

sieht man die Analogie zur empirischen Varianz:

$$\widetilde{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \frac{1}{n}.$$

- *Wie bei der empirischen Varianz ist es auch bei der Varianz oft günstiger, sie über den* **Verschiebungssatz** *zu berechnen:*

$$Var(X) = E(X^2) - (E(X))^2$$

- *Transformiert man die Zufallsvariable X linear, d.h. man betrachtet $Y = a \cdot X + b$ für Konstanten a, b, so gilt:*

$$Var(Y) = Var(a \cdot X + b) = a^2 \cdot Var(X)$$

und

$$\sigma_Y = |a| \cdot \sigma_X.$$

Beispiel 7.16 *Varianz*

1. *X = Augenzahl beim einmaligen Würfeln*

   mit $f(x_i) = 1/6$, $x_i = 1, \ldots, 6$
   ist $Var(X) = E(X^2) - (E(X))^2 = E(X^2) - 3.5^2$ *(vgl. Bsp. 7.11).*
   $E(X^2) = E(g(X))$ *mit* $g(X) = X^2$, *daher* $E(X^2) = \sum_{i=1}^{\infty} g(x_i) \cdot f(x_i) = \sum_{i=1}^{6} x_i^2 \cdot \frac{1}{6} = \frac{1}{6} \sum_{i=1}^{6} x_i^2 = \frac{1}{6} \cdot (1^2 + 2^2 + \ldots + 6^2) = \frac{91}{6} = 15.17$
   und $Var(X) = 15.17 - 3.5^2 = 2.92$.

2. *X = Wartezeit auf die Straßenbahn (vgl. Bsp. 7.11)*

   mit $f(x) = \frac{1}{5}$ *für* $0 \leq x \leq 5$ *(und* $f(x) = 0$ *sonst)*
   ist $Var(X) = E(X^2) - (E(X))^2 = E(X^2) - 2.5^2$
   $E(X^2) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx = \int_0^5 x^2 \cdot f(x) dx = \int_0^5 x^2 \cdot \frac{1}{5} dx = \frac{1}{5} \cdot \int_0^5 x^2 dx = \frac{1}{5} \cdot \frac{1}{3} \cdot x^3 \Big|_0^5 = \frac{125}{15} = 8.33$
   und $Var(X) = 8.33 - 2.5^2 = 2.08$.

7.4 Mehrdimensionale Zufallsvariablen

Betrachtet man mehrere Zufallsvariablen gemeinsam, so spricht man auch von einer mehrdimensionalen Zufallsvariablen. Beim Rechnen mit mehreren Zufallsvariablen benötigt man ähnlich wie bei der Behandlung mehrerer Merkmale die **gemeinsame Dichte**, die **Randdichte** und die **bedingte Dichte**. Wieder kann man für diskrete Zufallsvariablen die gemeinsame Dichte durch Wahrscheinlichkeiten darstellen.

Definition 7.17 *Gemeinsame Dichte, Randdichte, bedingte Dichte*

1. *Seien X und Y diskrete Zufallsvariablen mit Wertebereichen* $X(\Omega) = \{x_1, x_2, \ldots\}$, $Y(\Omega) = \{y_1, y_2, \ldots\}$.

- *Dann heißt* $f_{X,Y}$ *mit*

$$f_{X,Y}(x_i, y_j) = P(X = x_i \text{ und } Y = y_j) = P(X = x_i, Y = y_j)$$

die **gemeinsame diskrete Dichte** *von* X *und* Y*.*

- *In diesem Fall nennt man* f_X *und* f_Y *mit* $f_X(x_i) = P(X = x_i)$ *und* $f_Y(y_j) = P(Y = y_j)$ *die* **Randdichten** *von* X *und* Y*.*

- *Die beiden Funktionen* $f_{X|Y}$ *und* $f_{Y|X}$ *mit*

$$f_{X|Y}(x_i|y_j) = \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)} = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

für alle y_j *mit* $f_Y(y_j) \neq 0$ *und*

$$f_{Y|X}(y_j|x_i) = \frac{f_{X,Y}(x_i, y_j)}{f_X(x_i)} = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

für alle x_i *mit* $f_X(x_i) \neq 0$

die **bedingte Dichte von** X **gegeben** Y *bzw. die* **bedingte Dichte von** Y **gegeben** X*.*

2. *Sind* X *und* Y *stetige Zufallsvariablen, definiert man die Randdichte und die bedingte Dichte in entsprechender Weise über die* **gemeinsame stetige Dichte** $f_{X,Y}(x, y)$*.*

Analog zum Fall der relativen Häufigkeiten in der deskriptiven Statistik kann die Randdichte einer diskreten Zufallsvariablen X durch das Aufsummieren der gemeinsamen Dichte über alle Werte von Y bestimmt werden:

$$f_X(x_i) = \sum_{j=1}^{\infty} f_{X,Y}(x_i, y_j).$$

Im Fall stetiger Zufallsvariablen wird die Summation durch eine Integration ersetzt:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy.$$

Wie in der deskriptiven Statistik interessiert man sich für Abhängigkeiten zwischen Zufallsgrößen. Zwei Zufallsvariablen werden stochastisch unabhängig genannt, wenn sich ihre gemeinsame Dichte schreiben lässt als das Produkt der beiden Randdichten.

Definition 7.18 *Unabhängigkeit von Zufallsvariablen*
Es seien X und Y zwei Zufallsvariablen mit Dichten f_X, f_Y und gemeinsamer Dichtefunktion $f_{X,Y}$. Dann sind X und Y **(stochastisch) unabhängig**, *wenn*

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

für alle $x \in X(\Omega)$ und für alle $y \in Y(\Omega)$. (Beachte: hier sind die beiden Fälle diskreter und stetiger Zufallsvariablen beide abgedeckt.)

Zur Beschreibung der Abhängigkeit von zwei Zufallsvariablen X und Y dienen **Kovarianz** und **Korrelation**.

Definition 7.19 *Kovarianz, Korrelation*
Für zwei Zufallsvariablen X und Y ist die **Kovarianz** *zwischen X und Y definiert als*

$$Cov(X,Y) = E\Big((X - E(X)) \cdot (Y - E(Y))\Big).$$

Der **Korrelationskoeffizient** *(kurz: die* **Korrelation***) zwischen X und Y ist gegeben als*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}.$$

- *Sind X und Y diskret, so lässt sich die Formel für die Kovarianz darstellen durch:*

$$Cov(X,Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (x_i - E(X)) \cdot (y_j - E(Y)) \cdot f_{X,Y}(x_i, y_j).$$

- *Für zwei stetige Zufallsvariablen X, Y ergibt sich*

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - E(X)) \cdot (y - E(Y)) \cdot f_{X,Y}(x,y)\, dx\, dy.$$

Bemerkung 7.20 *Rechenregeln und Eigenschaften*

- *Zur vereinfachten Berechnung der Kovarianz verwendet man:*

$$Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y),$$

  wobei im diskreten Fall $E(X \cdot Y) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i \cdot y_j \cdot f_{X,Y}(x_i, y_j)$, *im stetigen Fall* $E(X \cdot Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x,y)\, dx\, dy$.

- *Transformiert man X und Y linear in $a \cdot X + b$ und $c \cdot Y + d$ für konstante Werte a, b, c, d, so gilt:*

$$Cov(a \cdot X + b, c \cdot Y + d) = a \cdot c \cdot Cov(X,Y).$$

- *Für zwei Zufallsvariablen X und Y gilt außerdem:*

$$\begin{aligned} Var(X+Y) &= Var(X) + Var(Y) + 2 \cdot Cov(X,Y), \\ Var(X-Y) &= Var(X) + Var(Y) - 2 \cdot Cov(X,Y). \end{aligned}$$

- *Wenn X und Y stochastisch unabhängig sind, so gilt $Cov(X,Y) = 0$. Der Umkehrschluss ist* **nicht** *zulässig! D.h. aus $Cov(X,Y) = 0$ folgt nicht die Unabhängigkeit der beiden Zufallsvariablen.*

Betrachtet man nicht nur zwei, sondern eventuell auch mehr als zwei Zufallsvariablen $X_1, \ldots, X_n$ gemeinsam, so gelten außerdem noch folgende Rechenregeln:

- $X_1, \ldots, X_n$ sind stochastisch unabhängig, falls

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n).$$

Dabei bezeichnet $f_{X_1,\ldots,X_n}$ die gemeinsame Dichte von $X_1, \ldots, X_n$ und f_{X_i} die Randdichte von X_i, $i = 1, \ldots, n$.

- Für den Erwartungswert gilt immer:

$$E(\sum_{i=1}^{n} X_i) = E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n) = \sum_{i=1}^{n} E(X_i),$$

$$E(X_1 - \ldots - X_n) = E(X_1) - \ldots - E(X_n).$$

- Falls $X_1, \ldots, X_n$ stochastisch unabhängig, gilt für die Varianz:

$$Var(\sum_{i=1}^{n} X_i) = Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n) = \sum_{i=1}^{n} Var(X_i).$$

8 Verteilungen

In der Praxis tauchen häufig bestimmte Typen von Zufallsvariablen auf. Diese unterscheidet man nach ihrer Verteilungsfunktion und spricht von sog. **Verteilungen**. Man unterscheidet wieder nach diskreten und stetigen Verteilungen.

8.1 Diskrete Verteilungen

Die Bernoulli-Verteilung

Beispiel 8.1 *Bernoulli-Experiment*
Ein Student schreibt eine Klausur. Das Ergebnis ist die erreichte Punktzahl, d.h. $\Omega = \{$*mögliche Punktzahlen in der Klausur*$\}$.
Es interessiert ihn aber nur, ob er die Klausur bestanden hat. Dies drückt er aus durch die Zufallsvariable X *mit*

$$X(\omega) = \begin{cases} 1, & \textit{falls Punktzahl zum Bestehen führt} \\ 0, & \textit{sonst} \end{cases}$$

Mit Wahrscheinlichkeit p besteht er die Klausur. Damit ist $P(X = 1) = p$ und $P(X = 0) = 1 - p$.

Ein solches Zufallsexperiment mit den zwei interessierenden Ereignissen "Erfolg" und "Misserfolg" sowie Erfolgswahrscheinlichkeit p nennt man auch **Bernoulli-Experiment**.
Die Dichtefunktion von X lässt sich nach obiger Überlegung beschreiben durch $f(0) = 1 - p$, $f(1) = p$ ($f = 0$ an allen anderen Stellen). Dies kann man zusammenfassend schreiben als $f(x_i) = p^{x_i} \cdot (1 - p)^{1-x_i}$ für $x_i = 0, 1$. Eine Zufallsvariable, deren Dichte in dieser Form dargestellt werden kann, besitzt eine sogenannte **Bernoulli-Verteilung**. *Die Erfolgswahrscheinlichkeit p heißt der* **Parameter** *dieser Verteilung.*

Definition 8.2 *Bernoulli-Verteilung*
Betrachtet wird ein interessierendes Ereignis A. Eine Zufallsvariable X, die den Wert 1 annimmt, falls A eintritt, und den Wert 0, falls A nicht eintritt, heißt **Bernoulli-verteilt mit Parameter** *p, wenn ihre Dichtefunktion f die folgende Form hat:*

$$f(x_i) = p^{x_i} \cdot (1 - p)^{1-x_i}$$

für $x_i = 0, 1$.
Schreibweise: $X \sim Bin(1, p)$.

Bemerkung 8.3 *Eigenschaften*

- *Der Erwartungswert einer $Bin(1, p)$-verteilten Zufallsvariablen ist*

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot f(x_i) = 0 \cdot f(0) + 1 \cdot f(1) = 0 \cdot (1 - p) + 1 \cdot p = p$$

- *Die Varianz ergibt sich als*

$$\begin{aligned} Var(X) &= \sum_{i=1}^{\infty}(x_i - E(X))^2 \cdot f(x_i) = (0-p)^2 \cdot f(0) + (1-p)^2 \cdot f(1) \\ &= p^2 \cdot (1-p) + (1-p)^2 \cdot p = p \cdot (1-p) \end{aligned}$$

Die Binomialverteilung

Beispiel 8.4 *n Bernoulli-Experimente*

Es schreiben n Studierende (hoffentlich) unabhängig die Klausur. Die Dozentin ist interessiert an der Anzahl der bestandenen Klausuren. Modellhaft geht man davon aus, dass die Wahrscheinlichkeit des Bestehens für jeden Studierenden die gleiche ist, nämlich p. Für den i-ten Studierenden wird das Ergebnis durch die Zufallsvariable X_i beschrieben mit

$$X_i = \begin{cases} 1, & \textit{falls Punktzahl zum Bestehen führt} \\ 0, & \textit{sonst} \end{cases}$$

Die Anzahl der bestandenen Klausuren ergibt sich dann als Summe der X_i: $X = \sum_{i=1}^{n} X_i$. Anders ausgedrückt, beschreibt X die Anzahl der Erfolge in n unabhängigen Bernoulli-Experimenten mit gleicher Erfolgswahrscheinlichkeit. Diese Zufallsvariable X besitzt eine sogenannte **Binomialverteilung**.

Definition 8.5 *Binomialverteilung*

Eine diskrete Zufallsvariable X, die Werte $0, 1, \ldots, n$ annehmen kann, mit Dichtefunktion

$$f(x_i) = \binom{n}{x_i} \cdot p^{x_i} \cdot (1-p)^{n-x_i}$$

für $x_i = 0, \ldots, n$

heißt **binomialverteilt mit Parametern n und p.**

Schreibweise: $X \sim Bin(n, p)$.

Bemerkung 8.6 *Eigenschaften*

- *Die Bernoulliverteilung ist ein Spezialfall der Binomialverteilung mit* $n = 1$.
- *Sind* $X_1, \ldots, X_n$ *stochastisch unabhängig mit* $X_i \sim Bin(1, p)$, $i = 1, \ldots, n$, *dann ist* $X = \sum_{i=1}^{n} X_i \sim Bin(n, p)$.
- *Erwartungswert und Varianz einer* $Bin(n, p)$*-verteilten Zufallsvariablen sind* $E(X) = n \cdot p$, $Var(X) = n \cdot p \cdot (1 - p)$.
- *Ist* $X \sim Bin(n, p)$, *dann ist* $n - X \sim Bin(n, 1 - p)$.

Die Binomialverteilung kommt auch immer dann zum Tragen, wenn man ein Urnenmodell unterstellen kann und sich in der Situation des Ziehens mit Zurücklegen befindet. Hat man beispielsweise eine Urne mit 60 Kugeln, von denen 24 rot sind, und zieht 20 Kugeln mit Zurücklegen, dann ist die Anzahl X der gezogenen roten Kugeln eine binomialverteilte Zufallsvariable:

$$X \sim Bin(20, \frac{24}{60}).$$

Die hypergeometrische Verteilung

Die hypergeometrische Verteilung spiegelt das Modell des Ziehens ohne Zurücklegen wider: sind in einer Urne N Kugeln, von denen M eine interessierende Eigenschaft besitzen, und zieht man n Kugeln ohne Zurücklegen, so ist die Zufallsvariable X, die die Anzahl der gezogenen Kugeln mit der interessierenden Eigenschaft beschreibt, **hypergeometrisch verteilt**. (Trifft zum Beispiel auf die Situation des Lottospiels zu.)

Definition 8.7 *Hypergeometrische Verteilung*

Eine diskrete Zufallsvariable X, die die Werte $0, 1, \ldots, n$ annehmen kann, mit Dichte

$$f(x_i) = \frac{\binom{M}{x_i} \cdot \binom{N-M}{n-x_i}}{\binom{N}{n}}$$

für $x_i = 0, \ldots, n$

heißt **hypergeometrisch verteilt mit Parametern n, M, N.**

Schreibweise: $X \sim Hyp(n, M, N)$.

Bemerkung 8.8 *Eigenschaften*

- *Ist $X \sim Hyp(n, M, N)$, so gilt:*
  $E(X) = n \cdot \frac{M}{N}$, $Var(X) = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}$.

- *Obwohl in der Situation des Ziehens ohne Zurücklegen die hypergeometrische Verteilung angebracht ist, kann man unter gewissen Bedingungen das Ziehen ohne Zurücklegen behandeln wie das Ziehen mit Zurücklegen. Falls nämlich der Umfang N der Grundgesamtheit erheblich größer ist als der Umfang n der Stichprobe, d.h. falls das Verhältnis $\frac{n}{N}$ klein ist, ist der Unterschied zwischen beiden Ziehungstechniken vernachlässigbar. Die hypergeometrische Verteilung wird, falls $\frac{n}{N}$ klein, durch die Binomialverteilung approximiert:*

$$\begin{array}{ccc} Hyp(n, M, N) & \approx & Bin(n, \frac{M}{N}). \\ & \uparrow & \\ & \frac{n}{N} \text{ klein} & \end{array}$$

  Faustregel: dieser Zusammenhang gilt für $\frac{n}{N} < 0.05$.

Die Poisson-Verteilung

Interessiert man sich für Zufallsvariablen wie die Anzahl der Verkehrsunfälle in Halle an einem Stichtag oder die Anzahl neu ankommender Kunden, die sich innerhalb der nächsten Minute in eine Warteschlange einreihen (z.B. am Postschalter), so ist es in beiden Fällen nicht möglich, zur Modellierung ein Urnenmodell heran zu ziehen. Es gibt außerdem keine feste obere Grenze für die möglichen Ausprägungen.

Tritt das interessierende Ereignis außerdem nur mit geringer Wahrscheinlichkeit ein, sind solche Anzahlen häufig **Poisson-verteilt**.

Definition 8.9 *Poisson-Verteilung*

Eine diskrete Zufallsvariable X, die Werte $0, 1, 2, \ldots$ annehmen kann, mit Dichtefunktion

$$f(x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

für $x_i = 0, 1, 2, \ldots$ und $\lambda > 0$
heißt **Poisson-verteilt mit Parameter** λ *("lambda").*
Schreibweise: $X \sim Poi(\lambda)$.

Bemerkung 8.10 *Eigenschaften*

- *Ist $X \sim Poi(\lambda)$, so gilt:*
  $E(X) = \lambda$, $Var(X) = \lambda$.
- *Man nennt die Poisson-Verteilung auch die Verteilung der seltenen Ereignisse.*
- *Für die Exponentialfunktion e^x gibt es auch die Schreibweise* $\exp(x)$, *und es ist $e^x = \exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$.*

8.2 Stetige Verteilungen

Die stetige Gleichverteilung

Im Fall einer diskreten Zufallsvariablen spricht man von einer Gleichverteilung, wenn alle möglichen Werte aus dem Wertebereich mit derselben Wahrscheinlichkeit realisiert werden können (vgl. die Augenzahl beim einfachen Würfeln). Im stetigen Fall heißt eine Zufallsvariable **gleichverteilt** auf einem Intervall, wenn ihre Dichte über diesem Intervall konstant ist und außerhalb den Wert Null annimmt.

Definition 8.11 *Stetige Gleichverteilung*
Eine stetige Zufallsvariable X mit Werten in $\mathbb{R}$ und Dichte

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & sonst \end{cases}$$

heißt **gleichverteilt (rechteckverteilt) auf dem Intervall** $[a, b]$.
Schreibweise: $X \sim G[a, b]$.

Bemerkung 8.12 *Eigenschaften*

- *Ist* $X \sim G[a, b]$, *dann gilt:*
  $E(X) = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$.

Die Normalverteilung

Merkmale, die durch das Zusammenwirken vieler zufälliger Einflüsse entstehen, wie beispielsweise Größe, Gewicht, physikalische Messgrößen, folgen häufig einer sogenannten **Normalverteilung**. Die Normalverteilung ist die wichtigste stetige Verteilung, zumal sich viele andere Verteilungen für große Stichprobenumfänge n annähernd wie eine Normalverteilung verhalten.

Definition 8.13 *Normalverteilung*

Eine stetige Zufallsvariable X mit Werten in $\mathbb{R}$ und Dichte

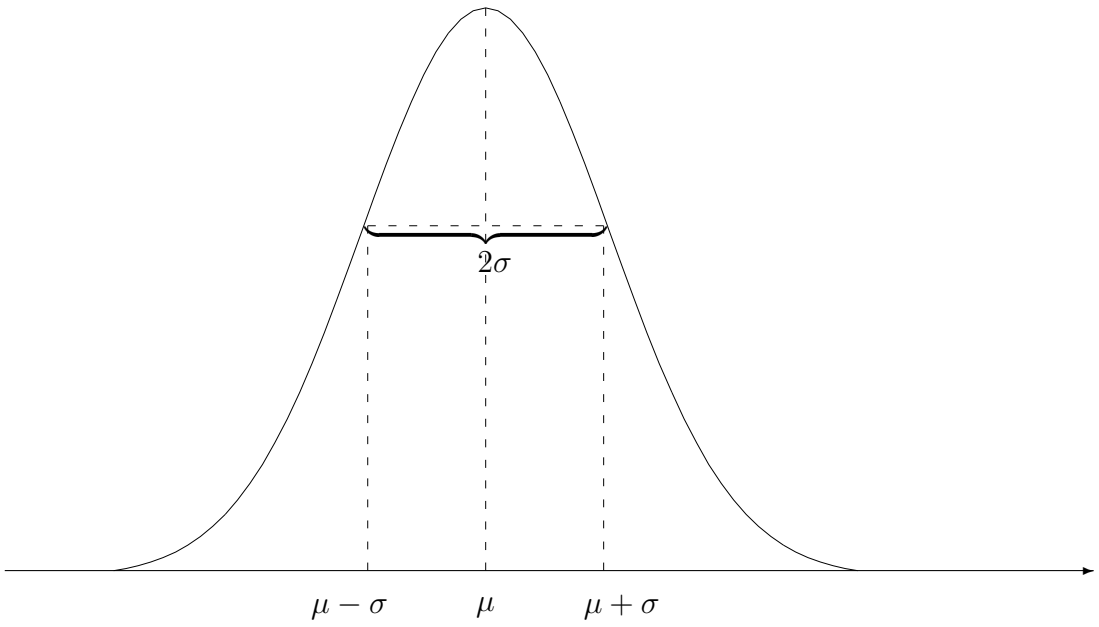
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left(-\frac{(x-\mu)^2}{2 \cdot \sigma^2}\right)$$

heißt **normalverteilt mit Parametern μ und σ^2** *("my" und "sigma-Quadrat").*

Schreibweise: $X \sim N(\mu, \sigma^2)$.

Die spezielle Normalverteilung $N(0,1)$ mit Parametern $\mu = 0$ und $\sigma^2 = 1$ heißt **Standardnormalverteilung**. *Ihre Verteilungsfunktion wird mit Φ ("Phi") bezeichnet.*

Skizze: Dichte der Normalverteilung



Bemerkung 8.14 *Eigenschaften*

- *Die Normalverteilung wird auch Gaußverteilung genannt, ihre Dichte ist auch bekannt als die Gauß'sche Glockenkurve.*

- *Ist* $X \sim N(\mu, \sigma^2)$*, dann gilt:*
  $E(X) = \mu$*,* $Var(X) = \sigma^2$*.*

- *Die Dichtekurve ist symmetrisch um* μ*, die Varianz* σ^2 *entscheidet über die Form der "Glocke", vgl. Skizze oben. Es ist* $f(x) > 0$ *für alle* $x \in \mathbb{R}$*.*

- *Eine* $N(\mu, \sigma^2)$*-verteilte Zufallsvariable* X *kann immer so* **standardisiert** *werden, dass ihre Transformation* Z $N(0,1)$*-verteilt ist:*
  ist $X \sim N(\mu, \sigma^2)$*, dann ist*

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

  d.h. $P(Z \leq z) = \Phi(z)$*.*

- *Ist* $X \sim N(\mu, \sigma^2)$*, dann ist eine lineare Transformation* Y *von* X *wieder normalverteilt, und zwar:*

$$Y = a \cdot X + b \sim N(a \cdot \mu + b, a^2 \cdot \sigma^2).$$

- *Ist* $X \sim N(\mu, \sigma^2)$*, dann ist*
  $P(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.68$
  $P(X \in [\mu - 2\sigma, \mu + 2\sigma]) \approx 0.95$
  $P(X \in [\mu - 3\sigma, \mu + 3\sigma]) \approx 0.99$
  (vgl. die Schwankungsbereiche auf Basis von $\overline{x}$ *und* $\widetilde{s}$ *in der deskriptiven Statistik).*

- *Speziell für die Verteilungsfunktion der Standardnormalverteilung gilt:*

$$\Phi(-z) = P(Z \leq -z) = P(Z \geq z) = 1 - \Phi(z)$$

  (mit $Z \sim N(0,1)$*).*

- *Sind $X_1, \ldots, X_n$ stochastisch unabhängig und ist jeweils $X_i \sim N(\mu_i, \sigma_i^2)$, dann ist*

$$\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2).$$

Im Spezialfall $X_i \sim N(\mu, \sigma^2)$ für alle i ist dann

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}).$$

Bemerkung 8.15 *Ermittlung von $P(a \leq X \leq b)$*

Die Verteilungsfunktion Φ der Standardnormalverteilung ist nicht in geschlossener Form darstellbar. Die Funktionswerte von Φ sind aber tabelliert und in Standard-Statistikprogrammen abrufbar.

- *Für eine $N(0,1)$-verteilte Zufallsvariable Z ist*

$$P(Z \leq z) = \Phi(z) \text{ und } P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

- *Für eine $N(\mu, \sigma^2)$-verteilte Zufallsvariable X ist*

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

und

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Beispiel 8.16 *Wahrscheinlichkeitsbestimmung*

- *Für $Z \sim N(0,1)$ gilt:*

z	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
$\Phi(z)$	0.5793	0.6554	0.7257	0.7881	0.8413	0.8849	0.9192	0.9452

Damit ist beispielsweise

$P(Z \leq 1.2) = \Phi(1.2) = 0.8849$ *und*

$P(0.4 \leq Z \leq 1.6) = \Phi(1.6) - \Phi(0.4) = 0.9452 - 0.6554 = 0.2898.$

Tabellierte Werte werden üblicherweise nur für positive Werte von z angegeben. Zur Bestimmung von $P(Z \leq -1)$ nutze die "Symmetrie" von Φ:

$P(Z \leq -1) = \Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587.$

- *Sei $X \sim N(30, 25)$.*

  Dann ist

$$P(X \leq 35) = P\left(\frac{X - 30}{5} \leq \frac{35 - 30}{5}\right) = \Phi(1) = 0.8413$$

  und

$$\begin{aligned} P(31 \leq X \leq 35) &= \Phi\left(\frac{35 - 30}{5}\right) - \Phi\left(\frac{31 - 30}{5}\right) \\ &= \Phi(1) - \Phi(0.2) = 0.8413 - 0.5793 = 0.2620. \end{aligned}$$

Bemerkung 8.17 *Ermittlung von Quantilen*

Auch zur Ermittlung von Quantilen der Normalverteilung benötigt man entsprechende Tabellen bzw. Programme. Entsprechend zu der Verteilungstabelle sind die Quantilswerte nur für die Standardnormalverteilung tabelliert. Eine solche Tabelle hat die folgende Gestalt (dabei bezeichnet z_p das p-Quantil der $N(0, 1)$):

p	0.95	0.96	0.97	0.98	0.99
z_p	1.6449	1.7507	1.8808	2.0537	2.3263

- *Zur Bestimmung des p-Quantils z_p der Standardnormalverteilung: lese z_p direkt aus der Tabelle ab.*

- *Zur Bestimmung des p-Quantils x_p der $N(\mu, \sigma^2)$-Verteilung: beachte, dass für x_p gelten muss $F(x_p) = p$ mit F = Verteilungsfunktion der $N(\mu, \sigma^2)$-Verteilung. Das heißt für $X \sim N(\mu, \sigma^2)$:*

$$p = F(x_p) = P(X \leq x_p) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x_p - \mu}{\sigma}\right) = \Phi(z_p)$$

(wegen $\Phi(z_p) = p$, wobei z_p das p-Quantil der $N(0, 1)$-Verteilung).
Das bedeutet:

$$z_p = \frac{x_p - \mu}{\sigma} \Rightarrow x_p = \sigma \cdot z_p + \mu$$

Das heißt: zur Bestimmung von x_p ermittle zunächst z_p aus der Quantiltabelle und bestimme daraus x_p nach obiger Formel.

Beispiel 8.18 *Quantile der Normalverteilung*

- *Gesucht: 95%-Quantil der Standardnormalverteilung.*
  Aus der Tabelle liest man für $p = 0.95$ ab, dass $z_{0.95} = 1.6449$.

- *Gesucht: 99%-Quantil der $N(22, 9)$-Verteilung.*
  Aus der Tabelle: für $N(0, 1)$ ist $z_{0.99} = 2.3263$, so dass

$$x_{0.99} = 3 \cdot 2.3263 + 22 = 28.9789.$$

Beispiel 8.19 *Normalverteilung*

Eine Firma möchte ein neues Lager einrichten. Bei der Dimensionierung der Fläche geht man davon aus, dass im Schnitt täglich 120 m^2 genutzt werden. Natürlich muss man eine gewisse Schwankung der genutzten Fläche einkalkulieren. Man unterstellt für die täglich genutzte Fläche X eine Normalverteilung mit Erwartungswert 120 und Varianz 400.
Probleme für die Firma treten immer dann auf, wenn mehr untergebracht werden muss, als Lagerkapazität zur Verfügung steht, oder wenn weniger als

die Hälfte der Lagerfläche auch genutzt wird. Die Firma entscheidet sich, das Lager mit einer Fläche von 150 m^2 zu dimensionieren. Damit ist die Wahrscheinlichkeit, dass Schwierigkeiten auftreten

$$\begin{aligned} P(X < 75 \text{ oder } X > 150) &= 1 - P(75 \leq X \leq 150) \\ &= 1 - (\Phi\left(\frac{150-120}{20}\right) - \Phi\left(\frac{75-120}{20}\right)) \\ &= 1 - \Phi(1.5) + \Phi(-2.25) \\ &= 1 - \Phi(1.5) + 1 - \Phi(2.25) \\ &= 1 - 0.9332 + 1 - 0.9878 = 0.079. \end{aligned}$$

Damit ist mit einer Wahrscheinlichkeit von 0.079 oder an rund 8 von 100 Tagen mit Schwierigkeiten zu rechnen.

Möchte man andererseits im Vorhinein abschätzen, wie groß das Lager sein sollte, damit an nicht mehr als 1% der Tage die Lagerkapazität nicht ausreicht, bestimmt man entsprechend das 99%-Quantil von X:
Es ist $z_{0.99} = 2.3263$, so dass

$$x_{0.99} = 20 \cdot 2.3263 + 120 = 166.526.$$

Mit einer Lagergröße von rd. 167 m^2 wird man an höchstens 1% der Tage keine ausreichende Lagerkapazität haben.

Die große Bedeutung der Normalverteilung ergibt sich aus dem sogenannten **zentralen Grenzwertsatz**. Dieser zentrale Grenzwertsatz besagt, dass das arithmetische Mittel aus n Zufallsvariablen (das selbst eine Zufallsvariable ist) asymptotisch normalverteilt ist. Genauer gilt folgendes: sind $X_1, X_2, \ldots$ stochastisch unabhängige Zufallsvariablen, die alle dieselbe Verteilung mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$, $0 < \sigma^2 < \infty$, besitzen, und wird das arithmetische Mittel aus $X_1, \ldots, X_n$ bezeichnet mit $\overline{X}_n$, dann gilt folgender Zusam-

menhang:

$$P\left(\sqrt{n}\cdot\frac{\overline{X}_n-\mu}{\sigma}\leq z\right)\longrightarrow\Phi(z)\quad(n\to\infty),$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung $N(0,1)$ bezeichnet. Dabei spielt es keine Rolle, welche Verteilung die X_i selbst ursprünglich besitzen.

9 Schätzer

Bisher: Charakterisierung von Verteilungen auf theoretischer Ebene durch Dichte, Verteilungsfunktion, Parameter usw.

Problem: tatsächliche Verteilung eines Merkmals in der Grundgesamtheit i.d.R. unbekannt

Lösung: Ziehe "gute" Stichprobe aus der Grundgesamtheit, schließe aus dieser auf die Verteilung des interessierenden Merkmals ("schätze" die Verteilung)
→ Aufgabe der **induktiven Statistik**

Bemerkung 9.1 *Modellvorstellung*
Betrachtet wird ein interessierendes Merkmal X, das als Zufallsvariable aufgefasst wird. Anhand einer Stichprobe $x_1,\ldots,x_n$ von Realisierungen des Merkmals soll eine Aussage über die Verteilung von X getroffen werden. Man stellt sich das i-te Stichprobenelement x_i vor als Realisation einer Zufallsvariablen X_i, die dieselbe Verteilung besitzt wie X.
Damit: betrachte $x_1,\ldots,x_n$ als Realisationen von n (unabhängigen) Zufallsvariablen $X_1,\ldots,X_n$, die alle dieselbe Verteilung besitzen wie X.

Typisches Vorgehen in dieser Situation:

- Annahme eines bestimmten Verteilungstyps (z.B. durch Vorkenntnis bedingt, weil gewisse Verteilungen in gewissen Situationen häufig auftreten), etwa Normalverteilung, Binomialverteilung, ...
- Schätzung der charakterisierenden Parameter der unterstellten Verteilung (Erwartungswert, Varianz, Erfolgswahrscheinlichkeit, etc.) durch Verdichtung der Information in $X_1, \ldots, X_n$

Zwei Arten von Schätzern:

- direkte Angabe eines Wertes für einen Parameter: sog. **Punktschätzer**
- Angabe eines Bereichs, in dem ein Parameter mit gewisser Wahrscheinlichkeit liegt: sog. **Intervallschätzer**

9.1 Punktschätzer

Beispiel 9.2 *Erfolgswahrscheinlichkeit bei Klausur*
Situation aus Beispiel 8.4 (n Studierende schreiben eine Klausur)
Bekannt: in dieser Situation ist die Zufallsvariable X_i, die den Klausurerfolg angibt, Bernoulli-verteilt mit Parameter p.

Die Dozentin interessiert sich für die Erfolgswahrscheinlichkeit p, dass ein Studierender die Klausur besteht.
Die Teilnehmer an der Klausur werden als reine Zufallsstichprobe aus allen potenziellen Klausurteilnehmern angesehen. Die vorliegenden Klausurergebnisse $x_1, \ldots, x_n$ betrachtet die Dozentin als Realisationen von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$, die alle $Bin(1, p)$-verteilt sind
$\rightarrow$ um p zu schätzen, bestimmt sie den Anteil der bestandenen Klausuren, d.h. $\frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$

→ sie zieht den Schluss, dass in der Grundgesamtheit aller potenziellen Klausurteilnehmer die Wahrscheinlichkeit des Bestehens dem in der Stichprobe realisierten Anteil $\overline{x}$ entspricht.

*Der in der Stichprobe realisierte Anteil $\overline{x}$ (***Schätzung***) ist selbst wieder die Realisation der zufälligen Größe $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (***Schätzer** *oder* **Schätzfunktion***).*

Definition 9.3 *Schätzfunktion, Schätzer, Schätzung*
Betrachtet wird eine interessierende Zufallsvariable X. Über die Verteilung von X sei bekannt, dass sie aus einer bestimmten Klasse von Verteilungen stammt und durch einen oder mehrere Parameter charakterisiert wird.
Seien weiter $X_1, \ldots, X_n$ n unabhängige Zufallsvariablen mit derselben Verteilung wie X. Eine Funktion t, die auf Basis von $X_1, \ldots, X_n$ den oder die Parameter durch $t(X_1, \ldots, X_n)$ schätzt, heißt **Schätzfunktion** *oder* **Schätzer** *(***Punktschätzer***).*
Die Realisierung $t(x_1, \ldots, x_n)$ von t an einer konkreten Stichprobe $x_1, \ldots, x_n$ heißt **Schätzung**.

Beachte:
Schätzer = selbst zufällige Größe, Zufallsvariable
Schätzung = realisierter Wert, hängt von der konkreten Stichprobe ab

Bemerkung 9.4 *Schätzer für den Erwartungswert*
Sind $X_1, \ldots, X_n$ Zufallsvariablen, die alle denselben Erwartungswert besitzen, d.h. $E(X_i) = \mu$ für alle i, dann ist $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ein Schätzer für μ. Es gilt:

- *$E(\overline{X}) = \mu$, das heißt, im "Mittel" schätzt man den zu bestimmenden Erwartungswert μ durch $\overline{X}$ korrekt.*

Sind $X_1, \ldots, X_n$ darüber hinaus unabhängig und besitzen sogar alle dieselbe Verteilung mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$, dann gilt:

- $Var(\overline{X}) = \frac{\sigma^2}{n}$, *das heißt, je größer die Stichprobe, desto genauer wird die Schätzung, desto näher liegt $\overline{X}$ an μ.*

Bemerkung 9.5 *Schätzer für die Varianz*
Sind $X_1, \ldots, X_n$ Zufallsvariablen, die alle denselben Erwartungswert und dieselbe Varianz besitzen, d.h. $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$ für alle i, dann sind

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

Schätzer für σ^2.
Sind $X_1, \ldots, X_n$ darüber hinaus unabhängig und besitzen alle dieselbe Verteilung, dann gilt:

- $E(\widetilde{S}^2) = \frac{n-1}{n}\sigma^2$, *das heißt, im "Mittel" unterschätzt man die zu bestimmende Varianz σ^2 durch $\widetilde{S}^2$.*
- $E(S^2) = \sigma^2$, *das heißt, im "Mittel" schätzt man σ^2 durch S^2 richtig.*

Beispiel 9.6 *Erwarteter Lohn*
In einer repräsentativen Untersuchung zu Lohnzahlungen in Niedriglohnberufen wurden $n = 39$ Raumpflegerinnen und Raumpfleger nach ihrem Stundenverdienst gefragt. Die Resultate sind in folgender Tabelle zusammengestellt:

7.91	*7.97*	*7.35*	*7.51*	*7.79*	*8.04*	*8.22*	*7.49*	*7.84*	*7.36*
7.92	*7.50*	*7.80*	*7.45*	*8.58*	*7.91*	*7.51*	*7.52*	*6.42*	*8.64*
7.98	*7.57*	*8.26*	*8.23*	*7.28*	*7.30*	*7.91*	*8.02*	*7.79*	*8.01*
8.27	*6.90*	*8.01*	*6.90*	*7.96*	*7.97*	*7.98*	*7.63*	*8.41*	

Als Schätzer für den erwarteten Lohn in dieser Berufsgruppe benutzt man den Mittelwert $\overline{X}$, der sich in dieser Stichprobe realisiert zum Schätzwert von

$$\overline{x} = \frac{1}{39} \sum_{i=1}^{39} x_i = 7.77.$$

In einer zweiten Studie zum selben Thema wurde eine ebenfalls repräsentative Stichprobe von $n = 30$ Personen dieser Berufsgruppe befragt. Hier wurden folgende Stundenlöhne genannt:

7.90	*7.99*	*7.39*	*7.50*	*7.86*	*8.03*	*8.23*	*7.59*	*7.94*	*7.38*
7.97	*7.52*	*7.85*	*7.40*	*8.61*	*7.91*	*7.56*	*7.50*	*6.37*	*8.66*
7.98	*7.53*	*8.29*	*8.20*	*7.30*	*7.27*	*7.96*	*8.05*	*7.78*	*6.40*

Hier ergibt sich für denselben Schätzer $\overline{X}$ nun die Schätzung

$$\overline{x} = \frac{1}{30} \sum_{i=1}^{30} x_i = 7.73.$$

Man sieht, dass zwar in beiden Fällen dieselbe Schätzfunktion für den Erwartungswert des Stundenlohns angewendet wurde, sich als konkrete Schätzungen jedoch zwei verschiedene Werte ergaben. Dies liegt daran, dass der Schätzer selbst eine Zufallsvariable ist, das konkrete Schätzergebnis daher von der jeweiligen Stichprobe abhängt.

Generell möchte man vernünftige Schätzfunktionen konstruieren. Dazu werden Gütekriterien aufgestellt, die von guten Schätzern erfüllt werden sollen. Zwei wichtige solche Kriterien sind die sog. **Erwartungstreue** und eine möglichst **geringe Varianz** des Schätzers: er soll im Mittel richtig schätzen und die Schätzungen sollen nicht zu weit streuen. Die oben angegebenen Schätzer $\overline{X}$ und S^2 erfüllen diese Bedingungen.
Zur Konstruktion geeigneter Schätzfunktionen gibt es verschiedene Prinzipien → Veranstaltung "Schätzen und Testen" im Masterstudium.

9.2 Intervallschätzer

Idee: Statt genau einen Wert als Schätzung für eine interessierende Größe anzugeben, trifft man zusätzliche Aussage über die Genauigkeit der Schätzung
$\rightarrow$ Angabe eines ganzen Bereichs (sog. **Konfidenzintervall**), der den interessierenden Parameter mit gewisser Wahrscheinlichkeit überdeckt.
Die Breite des Bereichs gibt dabei Auskunft über die Genauigkeit der Schätzung.

Definition 9.7 *Konfidenzintervall*
*Betrachtet wird eine interessierende Zufallsvariable X, wobei X eine Verteilung besitzt, die durch einen Parameter θ ("theta") charakterisiert wird. Seien weiter $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit derselben Verteilung wie X. Zu einer vorgegebenen Wahrscheinlichkeit α ("alpha"), $0 < \alpha < 1$, heißt ein Intervall $[K_u(X_1, \ldots, X_n), K_o(X_1, \ldots, X_n)]$ mit $K_u < K_o$ ein $(1-\alpha)$-***Konfidenzintervall (KI) für*** θ, *wenn*

$$P(\theta \in [K_u(X_1, \ldots, X_n), K_o(X_1, \ldots, X_n)]) = 1 - \alpha.$$

Die Wahrscheinlichkeit α heißt **Irrtumswahrscheinlichkeit**, *$1-\alpha$ heißt* **Konfidenzniveau** *oder* **Vertrauenswahrscheinlichkeit**.

Bemerkung 9.8 *Eigenschaften*

- *Typische Werte für α sind $\alpha = 0.1,\ 0.05,\ 0.01$.*
- *Die am häufigsten verwendete Art von KI ist das symmetrische KI auf Basis eines Schätzers $t(X_1, \ldots, X_n)$ für θ:*

  $$[t(X_1, \ldots, X_n) - d, t(X_1, \ldots, X_n) + d]$$

  für einen geeigneten Wert $d \in \mathbb{R}$.

Bemerkung 9.9 *Bestimmung eines symmetrischen KI*

Betrachte ein Normalverteilungsmodell, d.h. $X \sim N(\mu, \sigma^2)$*, wobei* σ^2 *bekannt sei,* μ *unbekannt. Der interessierende Parameter ist in diesem Fall* $\theta = \mu$*. Seien* $X_1, \ldots, X_n$ *unabhängig und identisch verteilt wie* X*.*
Gesucht ist ein symmetrisches KI für μ*.*

Bekannt: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ *ist ein vernünftiger Schätzer für* μ*.*
Damit gesucht: d*, so dass*

$$\begin{aligned}
& P(\theta \in [t(X_1, \ldots, X_n) - d, t(X_1, \ldots, X_n) + d]) = 1 - \alpha \\
\Leftrightarrow \quad & P(\mu \in [\overline{X} - d, \overline{X} + d]) = 1 - \alpha \\
\Leftrightarrow \quad & P(\overline{X} - d \leq \mu \leq \overline{X} + d) = 1 - \alpha \\
\Leftrightarrow \quad & P(-d \leq \mu - \overline{X} \leq d) = 1 - \alpha \\
\Leftrightarrow \quad & P(-d \leq \overline{X} - \mu \leq d) = 1 - \alpha \\
\Leftrightarrow \quad & P(\mu - d \leq \overline{X} \leq \mu + d) = 1 - \alpha
\end{aligned}$$

Nutze, dass

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \sim N(0, 1),$$

wenn die X_i *selbst* $N(\mu, \sigma^2)$*-verteilt sind. Damit gilt weiter:*

$$\begin{aligned}
& P(\theta \in [t(X_1, \ldots, X_n) - d, t(X_1, \ldots, X_n) + d]) = 1 - \alpha \\
\Leftrightarrow \quad & P(\sqrt{n} \cdot \frac{\mu - d - \mu}{\sigma} \leq \sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \leq \sqrt{n} \cdot \frac{\mu + d - \mu}{\sigma}) = 1 - \alpha \\
\Leftrightarrow \quad & P(-\sqrt{n} \cdot \frac{d}{\sigma} \leq Z \leq \sqrt{n} \cdot \frac{d}{\sigma}) = 1 - \alpha \\
\Leftrightarrow \quad & \Phi(\sqrt{n} \cdot \frac{d}{\sigma}) - \Phi(-\sqrt{n} \cdot \frac{d}{\sigma}) = 1 - \alpha \\
\Leftrightarrow \quad & \Phi(\sqrt{n} \cdot \frac{d}{\sigma}) - 1 + \Phi(\sqrt{n} \cdot \frac{d}{\sigma}) = 1 - \alpha
\end{aligned}$$

$$
\begin{aligned}
&\Leftrightarrow \quad 2 \cdot \Phi(\sqrt{n} \cdot \frac{d}{\sigma}) = 2 - \alpha \\
&\Leftrightarrow \quad \Phi(\sqrt{n} \cdot \frac{d}{\sigma}) = 1 - \alpha/2 \\
&\Leftrightarrow \quad \sqrt{n} \cdot \frac{d}{\sigma} = z_{1-\alpha/2} \\
&\Leftrightarrow \quad d = \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}
\end{aligned}
$$

Insgesamt ergibt sich ein symmetrisches $(1-\alpha)$-Konfidenzintervall für μ als

$$[\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}].$$

Bemerkung 9.10 *Konfidenzintervalle für μ im Normalverteilungsmodell*
Betrachte eine Zufallsvariable X mit $X \sim N(\mu, \sigma^2)$; seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt wie X. Gegeben sei weiter eine Irrtumswahrscheinlichkeit α, $0 < \alpha < 1$.

1. *Falls σ^2 bekannt ist, so ist*

   $$[\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}]$$

   ein $(1-\alpha)$-Konfidenzintervall für μ. Dabei bezeichnet $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der $N(0,1)$.

2. *Falls σ^2 unbekannt ist, ist*

   $$[\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}]$$

   ein $(1-\alpha)$-Konfidenzintervall für μ.
   Dabei ist $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$, und $t_{n-1;1-\alpha/2}$ bezeichnet das $(1-\alpha/2)$-Quantil der sogenannten t-Verteilung mit $n-1$ Freiheitsgraden (s.u.).

Bemerkung 9.11 *t-Verteilung*

Die sogenannte ***t*-Verteilung mit n Freiheitsgraden**, *auch* **Student-*t*-Verteilung** *genannt, besitzt die folgenden Eigenschaften:*

- *Sie ist symmetrisch um 0.*
- *Für das p-Quantil gilt:* $t_{n;p} = -t_{n;1-p}$ *(wegen der Symmetrie).*
- *Für* $n \geq 2$ *existiert der Erwartungswert einer* t_n*-verteilten Zufallsvariablen* X*, und es ist* $E(X) = 0$*.*
- *Für* $n \geq 3$ *existiert die Varianz, und es ist* $Var(X) = \frac{n}{n-2}$*.*
- *Für große Werte von* n *nähert sich die* t_n*-Verteilung der* $N(0,1)$*:*

$$t_n \longrightarrow N(0,1) \quad (n \to \infty).$$

  Faustregel: Approximation gut ab $n \geq 30$*.*

- *Sind* $X_1, \ldots, X_n$ *unabhängig und identisch* $N(\mu, \sigma^2)$*-verteilt, so ist*

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{S} \sim t_{n-1}.$$

- *Wie bei der Normalverteilung: Quantile der t-Verteilung sind vertafelt für verschiedene Freiheitsgrade; alternativ: mittels Statistik-Programm bestimmen.*

Beispiel 9.12 *Niedriglöhne*

In der Situation aus Beispiel 9.6 geht man davon aus, dass die Löhne X_i *der Reinigungskräfte normalverteilte Zufallsvariablen sind:* $X_i \sim N(\mu, \sigma^2)$*. Will man in diesem Fall ein KI für den erwarteten Lohn angeben, so greift man auf die Formel aus Bemerkung 9.10, Fall 2, zurück, da die Varianz* σ^2 *unbekannt ist.*

Es war für die erste Stichprobe $\overline{x} = 7.77$*;*
weiterhin muss noch $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ *bestimmt werden,*
hier: $s = 0.46$*.*

Zum Konfidenzniveau $1 - \alpha = 0.95$ *erhält man das 95%-KI für* μ *damit als*

$$[\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}] = [7.77 - \frac{0.46}{\sqrt{39}} \cdot t_{38;0.975}, 7.77 + \frac{0.46}{\sqrt{39}} \cdot t_{38;0.975}]$$

Aus Tabellen bzw. Rechner: $t_{38;0.975} = 2.024$*, und insgesamt:*

$$[7.62, 7.92]$$

ist ein 95%-Konfidenzintervall für μ*, d.h. der erwartete Stundenlohn liegt mit einer Wahrscheinlichkeit von 0.95 zwischen 7.62 und 7.92 Euro.*

Abschließend: Was tun, wenn die betrachteten Zufallsvariablen X_i nicht normalverteilt sind?
Dann nutze den zentralen Grenzwertsatz, nämlich, dass

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \sim N(0, 1) \text{ approximativ.}$$

Bemerkung 9.13 *Approximative Konfidenzintervalle für* μ
Sind in der Situation von Bemerkung 9.10 die Zufallsvariablen nicht notwendig normalverteilt, besitzen aber alle dieselbe Verteilung mit Erwartungswert μ *und Varianz* σ^2*, so gilt: für* $n \geq 30$ *ist jeweils*

1. *bei bekannter Varianz* σ^2

$$[\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}]$$

2. *bei unbekannter Varianz* σ^2

$$[\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2}]$$

ein approximatives $(1 - \alpha)$*-Konfidenzintervall für* μ*.*

10 Statistische Hypothesentests

10.1 Prinzip des Testens

Empirischen Untersuchungen liegen oft gewisse Forschungshypothesen zu Grunde. Dies sind Vermutungen über Sachverhalte oder Zusammenhänge **in der Grundgesamtheit**. Diese Vermutungen sollen an Hand erhobener Daten beurteilt werden.

Definition 10.1 *Hypothese, Alternative, Testproblem, Test*
Zur Überprüfung einer Vermutung über eine Grundgesamtheit formuliert man eine sogenannte **Nullhypothese** H_0 *(kurz:* **Hypothese***) und eine dazu komplementäre* **Alternativhypothese** H_1 *(kurz:* **Alternative***). Auf Grund der Besonderheiten statistischer Testtheorie wird die zu untersuchende Forschungshypothese als Alternativhypothese formuliert.*
Die Aufgabenstellung, zwischen der Gültigkeit von H_0 *und* H_1 *zu entscheiden, bezeichnet man als* **statistisches Testproblem**. *Man schreibt kurz:* H_0 *vs.* H_1 *("versus").*
Eine formale Entscheidungsregel, die eine Antwort auf das Testproblem gibt, heißt **statistischer Test**.

Beispiel 10.2 *Testproblem*
Bei Wahlen zu Bundestag oder Landtagen ist es insbesondere für kleinere Parteien interessant, ob sie die 5%-Hürde nehmen oder nicht. Nehmen sie sie nicht, gehen ihnen Einfluss und Geld verloren. Sie interessieren sich daher für die Forschungshypothese, dass p, der Anteil der von ihnen erzielten Stimmen, kleiner ist als 5%. Gesucht wird eine Entscheidung zwischen $H_0 : p \geq 0.05$ *und* $H_1 : p < 0.05$. *Als Testproblem geschrieben:*

$$H_0 : p \geq 0.05 \text{ vs. } H_1 : p < 0.05$$

Auf Basis des bereits ausgezählten Teils der abgegebenen Stimmen wird über dieses Testproblem am Wahltag mit jeder Hochrechnung von Neuem entschieden.

Generell sind beim Testen zwei Entscheidungen möglich:

1. H_0 wird verworfen. Das bedeutet: es gibt in der erhobenen Stichprobe starke Hinweise darauf, dass H_0 nicht gelten kann. Diese Hinweise sind so stark, dass man nicht von einem zufälligen Zustandekommen ausgehen kann.

2. H_0 wird nicht verworfen. Das bedeutet: man hat keine Hinweise gefunden, die gegen H_0 sprechen; alle aufgetretenen Effekte, die gegen H_0 sprechen könnten, könnten genauso gut zufallsbedingt sein.

Da beide Entscheidungen von der jeweils gezogenen Stichprobe abhängen und damit selbst zufallsbehaftet sind, können mit gewissen Wahrscheinlichkeiten Fehler auftreten. Trifft man Entscheidung 1, obwohl tatsächlich H_0 korrekt ist, oder Entscheidung 2, obwohl H_1 korrekt ist, so handelt es sich um Fehlentscheidungen.

Definition 10.3 *Fehler beim Testen*
Gegeben sei ein statistisches Testproblem H_0 vs. H_1 und ein für dieses Problem geeigneter statistischer Test.

- *Falls H_0 gilt, die Testentscheidung aber lautet, H_0 zu verwerfen, so begeht man den* **Fehler 1. Art**.

- *Falls H_1 gilt, die Testentscheidung aber lautet, H_0 beizubehalten, so begeht man den* **Fehler 2. Art**.

Beispiel 10.4 *Fehler*

Liegt im Fall von Beispiel 10.2 der wahre Stimmenanteil bei 7%, aber die Hochrechnung geht davon aus, dass die Partei nicht im Landtag vertreten sein wird $\rightarrow$ *Fehler 1. Art.*

Liegt der Anteil bei 3%, laut Hochrechnung gibt man aber die Prognose ab, dass die Partei im Landtag vertreten sein wird $\rightarrow$ *Fehler 2. Art.*

Idealerweise möchte man statistische Tests so konstruieren, dass die Wahrscheinlichkeit für das Begehen beider Fehler möglichst minimal ist. Leider kann man nicht beide Fehlerarten gleichzeitig kontrollieren. Daher sichert man in der Regel die Wahrscheinlichkeit für den Fehler 1. Art durch eine vorgegebene Schranke nach oben ab (sog. **Signifikanzniveau**). Unter allen statistischen Tests für ein Problem, die diese Schranke einhalten, wählt man dann denjenigen mit der kleinsten Wahrscheinlichkeit für den Fehler 2. Art. Diese unsymmetrische Behandlungsweise ist der Grund, dass die interessierende Forschungshypothese in der Regel in die Alternative geschrieben wird (da man die Wahrscheinlichkeit, sich fälschlicherweise für die Alternative und damit für die interessierende Aussage zu entscheiden, durch die vorgegebene Schranke absichert).

Definition 10.5 *Test zum Niveau* α

Ein statistischer Test für das Testproblem H_0 *vs.* H_1 *heißt* **Test zum Niveau** α*, wenn für einen vorgegebenen Wert von* α*,* $0 < \alpha < 1$*, gilt:*

$$P(\textit{Entscheidung für } H_1 | H_0 \textit{ ist wahr}) \leq \alpha.$$

Die Schranke α *heißt* **Signifikanzniveau** *(oder kurz:* **Niveau***) des Tests.*

Übliche Werte für α sind $\alpha = 0.1, 0.05, 0.01$.

Bemerkung 10.6 *Interpretation von Testergebnissen*

- *Beim Testen wird nur die Wahrscheinlichkeit für den Fehler 1. Art kontrolliert, d.h. P(H_0 ablehnen $|H_0$ ist wahr) $\leq \alpha$. Wenn also H_0 tatsächlich gilt, wird man sich nur in $\alpha \cdot 100\%$ der Fälle für H_1 entscheiden.*

  Die Entscheidung für H_1 ist in diesem Sinn statistisch abgesichert. Bei Entscheidung gegen H_0 und damit für H_1 spricht man von einem signifikanten Ergebnis.

- *Die Wahrscheinlichkeit für den Fehler 2. Art wird dagegen nicht kontrolliert. Die Entscheidung, H_0 beizubehalten, ist statistisch nicht abgesichert. Kann man H_0 nicht verwerfen, so bedeutet das daher nicht, dass man sich "aktiv" für H_0 entscheidet (es spricht nur nichts gegen H_0).*

10.2 Spezielle Tests

Beispiel 10.7 *Test über den Erwartungswert*

Eine Firma produziert Brötchen. Bekannt ist: das Brötchengewicht X ist eine normalverteilte Zufallsgröße mit $X \sim N(\mu, \sigma^2)$, wobei $\sigma^2 = 1.44$ gilt.

Behauptung der Firma: die produzierten Brötchen sind im Mittel 50 Gramm schwer, d.h. $\mu = 50$.

Den Verbraucher interessiert: stimmt diese Angabe, bzw. sind die Brötchen (zu Gunsten des Verbrauchers) vielleicht sogar etwas schwerer? In diesem Fall wäre der Verbraucher zufrieden und würde die Brötchen anstandslos akzeptieren.

Falls aber das mittlere Brötchengewicht kleiner wäre als 50 Gramm, würde der Verbraucher protestieren.

Aus Verbrauchersicht ergibt sich also folgendes Testproblem:

$$H_0 : \mu \geq 50 \text{ vs. } H_1 : \mu < 50$$

Um dieses Problem an Hand einer Stichprobe aus n Brötchen zu entscheiden, schätzt man zunächst das erwartete Brötchengewicht mit einem geeigneten Schätzer: betrachte Gewichte $x_1, \ldots, x_n$ als Realisationen von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$, alle mit derselben Verteilung wie X. Dann ist $\overline{X}$ ein vernünftiger Schätzer für μ.

Ist das durch $\overline{X}$ geschätzte erwartete Gewicht deutlich größer als 50
→ spricht nicht gegen H_0.
Ist $\overline{X}$ ungefähr gleich 50 oder knapp darunter
→ spricht auch noch nicht gegen H_0.
Ist $\overline{X}$ aber deutlich kleiner als 50
→ spricht gegen H_0 und damit für H_1.

$\overline{X}$ dient also als sog. **Prüfgröße** *oder* **Teststatistik**.

Frage: Wo setzt man die Grenze?
Dies geschieht durch die Vorgabe des Signifikanzniveaus α. Die Grenze hängt ab von der gewünschten Wahrscheinlichkeit für den Fehler 1. Art.

Dazu betrachtet man die Stelle, an der Hypothese und Alternative sich "treffen", d.h. man betrachtet den Fall $\mu = 50$.
Bekannt: in der oben beschriebenen Modellsituation ist

$$\sqrt{n} \cdot \frac{\overline{X} - 50}{1.2} \sim N(0, 1),$$

falls exakt $\mu = 50$ gilt. Verwende daher statt $\overline{X}$ die standardisierte Größe als Teststatistik.
Bei Gültigkeit der Hypothese soll die Wahrscheinlichkeit für den Fehler 1.

Art höchstens gleich α sein. Man stellt diesen Zusammenhang wieder für den Trennpunkt zwischen Hypothese und Alternative her, d.h.

$$P(\textit{Fehler 1. Art}) = P(\sqrt{n} \cdot \frac{\overline{X} - 50}{1.2} < c) \leq \alpha.$$

Gleichzeitig möchte man die Schranke c bei dem hier untersuchten Testproblem möglichst groß wählen, damit Abweichungen nach unten vom postulierten Gewicht von 50g möglichst schnell erkannt werden.
Beides zusammen liefert: c ist z_α, das α-Quantil der $N(0,1)$.

Insgesamt:
Lehne $H_0 : \mu \geq 50$ zu Gunsten von $H_1 : \mu < 50$ ab, falls

$$\sqrt{n} \cdot \frac{\overline{X} - 50}{1.2} < z_\alpha = -z_{1-\alpha}.$$

Ein Verbraucher kauft $n = 25$ zufällig ausgewählte Brötchen und ermittelt als durchschnittliches Gewicht einen realisierten Wert von $\overline{x} = 49.5g$. Für den Test zum Niveau $\alpha = 0.05$ ermittelt er

$$\sqrt{n} \cdot \frac{\overline{x} - 50}{1.2} = \sqrt{25} \cdot \frac{49.5 - 50}{1.2} = -2.083 < -1.6449 = -z_{0.95}.$$

Die Hypothese kann also zum Niveau $\alpha = 0.05$ verworfen werden. Das erwartete Brötchengewicht liegt unter 50g.

Bemerkung 10.8 *Test*

- *Die Testentscheidung ist vom Signifikanzniveau abhängig. Je kleiner α, desto geringer die Wahrscheinlichkeit für den Fehler 1. Art, desto größer müssen aber auch die realisierten "Abweichungen" sein, um H_0 zu verwerfen.*

  In Beispiel 10.7: wählt man statt $\alpha = 0.05$ einen Wert von $\alpha = 0.01$, so ist

  $$\sqrt{n} \cdot \frac{\overline{x} - 50}{1.2} = -2.083 > -2.3263 = -z_{0.99},$$

  und H_0 kann zum Niveau 0.01 nicht verworfen werden!

- *Auch andere Hypothesen bzw. Alternativen können interessant sein.*

  Für den Produzenten ist beispielsweise eher die Hypothese $H_0 : \mu \leq 50$ *(vs.* $H_1 : \mu > 50$*) von Interesse, da aus seiner Sicht die Brötchen nicht zu viel wiegen sollten.*

  Im Rahmen einer Qualitätskontrolle kann die möglichst genaue Einhaltung des Sollgewichts von 50g interessieren:

  $H_0 : \mu = 50$ *vs.* $H_1 : \mu \neq 50$.

  Beide Testprobleme führen zu anderen Vorschriften, nach denen die jeweilige Hypothese zu verwerfen ist.

Bemerkung 10.9 *Gauß-Test*

Seien $X_1, \ldots, X_n$ *unabhängige und identisch normalverteilte Zufallsvariablen,* $X_i \sim N(\mu, \sigma^2)$*, und sei* σ^2 *bekannt. Zum Testen von Hypothesen über den unbekannten Erwartungswert* μ *dient der sogenannte* **Gauß-Test**. *Dabei werden zum Niveau* α *(*$0 < \alpha < 1$*) die folgenden Entscheidungen für die angegebenen Testprobleme getroffen:*

Testproblem	*Entscheidung*
H_0 *vs.* H_1	*Lehne* H_0 *ab, falls*
$\mu = \mu_0$ *vs.* $\mu \neq \mu_0$	$\sqrt{n} \cdot \frac{\lvert\overline{X}-\mu_0\rvert}{\sigma} > z_{1-\alpha/2}$
$\mu \geq \mu_0$ *vs.* $\mu < \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{\sigma} < -z_{1-\alpha}$
$\mu \leq \mu_0$ *vs.* $\mu > \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{\sigma} > z_{1-\alpha}$

Dabei bezeichnet z_α *das* α*-Quantil der Standardnormalverteilung.*

Bemerkung 10.10 *t-Test*

Die Voraussetzungen seien wie in Bemerkung 10.9, aber sei σ^2 *unbekannt. Zum Testen von Hypothesen über* μ *dient in diesem Fall der* **t-Test**. *Dabei wird in den Teststatistiken des Gauß-Tests die unbekannte Standardabweichung* σ *durch den Schätzer* $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ *ersetzt, und statt der Quantile der* $N(0,1)$ *werden Quantile der* t*-Verteilung benutzt:*

Testproblem H_0 *vs.* H_1	*Entscheidung* *Lehne* H_0 *ab, falls*
$\mu = \mu_0$ *vs.* $\mu \neq \mu_0$	$\sqrt{n} \cdot \frac{\lvert\overline{X}-\mu_0\rvert}{S} > t_{n-1;1-\alpha/2}$
$\mu \geq \mu_0$ *vs.* $\mu < \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{S} < -t_{n-1;1-\alpha}$
$\mu \leq \mu_0$ *vs.* $\mu > \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{S} > t_{n-1;1-\alpha}$

Dabei bezeichnet $t_{n-1;\alpha}$ *das* α*-Quantil der t-Verteilung mit n-1 Freiheitsgraden.*

Bemerkung 10.11 *Approximativer Gauß-Test*

Betrachtet man n unabhängige und identisch verteilte Zufallsvariablen, die aber nicht notwendig normalverteilt sind, mit $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, *wobei* σ^2 *unbekannt ist, so kann man für* $n \geq 30$ *den zentralen Grenzwertsatz ausnutzen. Damit ergibt sich der* **approximative Gauß-Test zum Nivau** $\boldsymbol{\alpha}$ *zu folgenden Testproblemen mit den angegebenen Entscheidungsregeln:*

Testproblem H_0 *vs.* H_1	*Entscheidung* *Lehne* H_0 *ab, falls*
$\mu = \mu_0$ *vs.* $\mu \neq \mu_0$	$\sqrt{n} \cdot \frac{\lvert\overline{X}-\mu_0\rvert}{S} > z_{1-\alpha/2}$
$\mu \geq \mu_0$ *vs.* $\mu < \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{S} < -z_{1-\alpha}$
$\mu \leq \mu_0$ *vs.* $\mu > \mu_0$	$\sqrt{n} \cdot \frac{\overline{X}-\mu_0}{S} > z_{1-\alpha}$

Beispiel 10.12 *Test auf einen Anteil*

Situation aus Beispiel 10.2: die kleine Partei A möchte wissen, ob sie im Landtag vertreten sein wird. Dazu werden $n = 100$ *zufällig ausgewählte Personen über ihr Wahlverhalten befragt. Unter den Befragten bezeichnen sich 4 Personen als A-Wähler. Bezeichne p den Anteil von A-Wählern in der Grundgesamtheit der Wahlteilnehmer. Die Partei möchte eine Entscheidung von*

$$H_0 : p \geq 0.05 \text{ vs. } H_1 : p < 0.05$$

zum Niveau $\alpha = 0.01$ treffen.

Bekannt: $\overline{X}$ (= Anteil der A-Wähler in der Stichprobe) ist ein vernünftiger Schätzer für p, außerdem gilt:

$$\sqrt{n} \cdot \frac{\overline{X} - p}{\sqrt{p \cdot (1-p)}} \sim N(0,1) \quad \textit{asymptotisch.}$$

Damit kann man analog zu der Situation in Beispiel 10.7 die Entscheidung treffen, H_0 zu verwerfen, falls die Abweichung nach unten von $\overline{X}$ zum interessierenden Wert $p_0 = 0.05$ zu groß wird:
lehne H_0 ab, falls

$$\sqrt{n} \cdot \frac{\overline{X} - p_0}{\sqrt{p_0 \cdot (1-p_0)}} < -z_{1-\alpha}.$$

Mit $n = 100$ und $\overline{x} = 4/100 = 0.04$ ist für $\alpha = 0.01$

$$\sqrt{n} \cdot \frac{\overline{X} - p_0}{\sqrt{p_0 \cdot (1-p_0)}} = \sqrt{100} \cdot \frac{0.04 - 0.05}{\sqrt{0.05 \cdot 0.95}} = -0.4588 > -2.3263 = -z_{0.99}.$$

Die Hypothese kann daher zum Niveau 0.01 nicht verworfen werden. Es spricht nichts dagegen, dass Partei A im Landtag vertreten sein wird.

Bemerkung 10.13 *Test auf einen Anteil*

Besitzt ein Anteil p der Grundgesamtheit eine interessierende Eigenschaft, und werden Zufallsvariablen $X_1, \ldots, X_n$ betrachtet mit $X_i = 1$, falls das i-te Element die Eigenschaft besitzt, $X_i = 0$ sonst, so kann man zu den folgenden Testproblemen über p zum Niveau α die folgenden Entscheidungen als approximative Testentscheidungen verwenden:

Testproblem H_0 *vs.* H_1	*Entscheidung* *Lehne* H_0 *ab, falls*
$p = p_0$ *vs.* $p \neq p_0$	$\sqrt{n} \cdot \frac{\lvert\overline{X} - p_0\rvert}{\sqrt{p_0 \cdot (1-p_0)}} > z_{1-\alpha/2}$
$p \geq p_0$ *vs.* $p < p_0$	$\sqrt{n} \cdot \frac{\overline{X} - p_0}{\sqrt{p_0 \cdot (1-p_0)}} < -z_{1-\alpha}$
$p \leq p_0$ *vs.* $p > p_0$	$\sqrt{n} \cdot \frac{\overline{X} - p_0}{\sqrt{p_0 \cdot (1-p_0)}} > z_{1-\alpha}$

Zur Zusammenhangsanalyse in Kontingenztafeln existiert neben der Betrachtung des korrigierten Kontingenzkoeffizienten K^* (vgl. Def. 3.10) ein formaler Test auf Unabhängigkeit zwischen zwei Zufallsvariablen X und Y. Dieser basiert auf der Berechnung des χ^2-Koeffizienten (Def. 3.8).

Bemerkung 10.14 *χ^2-Unabhängigkeitstest*
Betrachtet werden zwei Zufallsvariablen X, Y. Die Beobachtungspaare (x_i, y_i) werden als Realisationen von Paaren (X_i, Y_i) betrachtet, wobei jedes Paar (X_i, Y_i) die gleiche gemeinsame Verteilung besitzt wie (X, Y). Die Beobachtungen (x_i, y_i) seien analog zu Definition 3.8 in einer $(k \times m)$-Kontingenztafel zusammengefasst. Die gemeinsamen absoluten Häufigkeiten in der Tafel seien bezeichnet mit h_{ij}, die Randhäufigkeiten mit $h_{i\bullet}$ bzw. $h_{\bullet j}$, und die unter Unabhängigkeit von X und Y erwarteten Häufigkeiten mit $e_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$, $i = 1, \ldots, k$, $j = 1, \ldots, m$.
Für das Testproblem

$$H_0 : X, Y \text{ unabhängig vs. } H_1 : X, Y \text{ abhängig}$$

ist ein Test zum Niveau α gegeben durch folgende Entscheidungsregel:
H_0 wird verworfen, falls

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}} > \chi^2_{(k-1)\cdot(m-1);1-\alpha}$$

Dabei bezeichnet $\chi^2_{q;\alpha}$ das α-Quantil der χ^2-Verteilung mit q Freiheitsgraden.

Bemerkung 10.15 *χ^2-Verteilung*
Die **χ^2-Verteilung mit q Freiheitsgraden** *besitzt folgende Eigenschaften:*

- *Sie ist nicht symmetrisch.*
- *Ist $X \sim \chi^2_q$, so ist $E(X) = q$, $Var(X) = 2q$.*