Stochastik und Statistik

Jil Zerndt, Lucien Perret,
Moritz Feuchter
January 2025

Deskriptive Statistik

Teilbereiche der Statistik

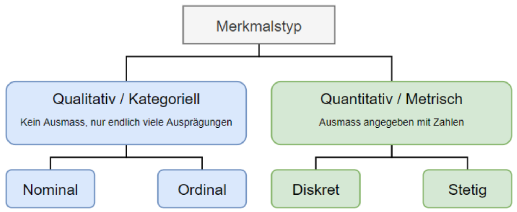
- **Deskriptive Statistik:** Beschreibung und übersichtliche Darstellung von Daten, Ermittlung von Kenngrössen und Datenvalidierung
- **Explorative Statistik:** Weiterführung und Verfeinerung der beschreibenden Statistik, insbesondere die Suche nach Strukturen und Besonderheiten
- **Induktive Statistik:** Versucht mithilfe der Wahrscheinlichkeitsrechnung über die erhobenen Daten hinaus allgemeinere Schlussfolgerungen zu ziehen

Statistische Grundbegriffe

- **Merkmalsträger/Statistische Einheiten:** Objekte, an denen interessierende Grössen beobachtet und erfasst werden (z.B. Wohnungen, Menschen, Unternehmen)
- **Grundgesamtheit:** Alle statistischen Einheiten, über die man Aussagen gewinnen möchte. Kann endlich oder unendlich, real oder hypothetisch sein
- **Vollerhebung:** Eigenschaften werden bei jedem Individuum in der Grundgesamtheit erhoben
- **Stichprobe:** Untersuchte Teilmenge der Grundgesamtheit, soll diese möglichst genau repräsentieren
- **Stichprobenumfang:** Anzahl der Einheiten in der Stichprobe
- **Urliste:** Liste der beobachteten Stichprobenwerte
- **Merkmal:** Interessierende Grösse, die an den statistischen Einheiten beobachtet wird
- **Merkmalsausprägungen:** Verschiedene Werte, die jedes Merkmal annehmen kann

Merkmalstypen

- **Qualitativ/Kategoriell:** eine Ausprägung, kein Ausmass angegeben
  - **Nominal:** Reine Kategorisierung, keine Ordnung
  - **Ordinal:** Ordnung vorhanden, Rangierung möglich
- **Quantitativ/Metrisch:** Es wird ein Ausmass mit Zahlen angegeben
  - **Diskret:** Endlich viele / abzählbar unendlich viele Ausprägungen
  - **Stetig:** Alle Ausprägungen in einem reellen Intervall



Merkmalstypen

- **Würfelwurf (4-mal)** Messniveau: Metrisch diskret
  - Merkmalsausprägungen: Zahlen 1 bis 6
- **Parteiwahl (100 Menschen)** Messniveau: Nominal
  - Merkmalsausprägungen: BDP, CVP, FDP, GLP, etc.
- **Programmrobustheit (100 Tests)** Messniveau: Ordinal
  - Merkmalsausprägungen: schlecht, mittel, sehr gut
- **Programmlaufzeit (100 Tests)** Messniveau: Metrisch stetig
  - Merkmalsausprägungen: Laufzeiten

Häufigkeiten und Verteilungsfunktion

Grundlegende Begriffe

Symbole und Bezeichnungen

- Ω = Grundgesamtheit
- n = Anzahl Objekte (Stichprobenumfang)
- a = Ausprägungen
- a_i = i-te Ausprägung
- m = Anzahl verschiedener Merkmalsausprägungen
- d = Klassenbreite
- X = Stichprobenwerte
- x = Einzelner Stichprobenwert
- h = Absolute Häufigkeit
- f = Relative Häufigkeit
- H = Kumulative Absolute Häufigkeit
- F = Kumulative Relative Häufigkeit

Grundlegende Unterscheidungen

- **Diskrete vs. Stetige Merkmale:**
  - Diskret: PMF, Höhe = Wahrscheinlichkeit
  - Stetig: PDF, Fläche = Wahrscheinlichkeit
- **Nicht-klassiert vs. Klassiert:**
  - Nicht-klassiert: Einzelwerte
  - Klassiert: Intervalle mit Häufigkeitsdichten
- **Absolut vs. Relativ:**
  - Absolut: Konkrete Anzahlen
  - Relativ: Anteile (durch n geteilt)
- **Punktuell vs. Kumulativ:**
  - Punktuell: Häufigkeit an einem Punkt/in einer Klasse
  - Kumulativ: Aufsummierte Häufigkeiten bis zu einem Punkt

Absolute Häufigkeit $h_i = h(x)$

$$\sum_{i=1}^{m} h_i = n$$

h_i: Anzahl des Auftretens eines Wertes/einer Klasse a_i $(i = 1, ..., m)$

Kumulative absolute Häufigkeit:

$$H(x) = \sum_{i:a_i \leq x} h_i$$

Relative Häufigkeit $f_i = \frac{h_i}{n}$

$$\sum_{i=1}^{m} f_i = 1$$

f_i = Anteil der absoluten Häufigkeit h_i am Stichprobenumfang n
Wertebereich: $0 \leq f_i \leq 1$

Kumulative relative Häufigkeit:

$$F(x) = \frac{H(x)}{n} = \sum_{i:a_i \leq x} f_i$$

Übersicht Häufigkeits- und Verteilungsfunktionen

Diskrete Merkmale:

- **PMF**: $f(x) = \frac{h(x)}{n}$, Höhe = rel. Häufigkeit
- **CDF**: $F(x) = \sum_{r \leq x} f(r)$, Treppenfunktion

Stetige/Klassierte Merkmale:

- **Absolute Häufigkeitsdichte**: $h = \frac{h_i}{d_i}$, Höhe im Histogramm
- **PDF**: $f = \frac{h_i}{n \cdot d_i} = \frac{f_i}{d_i}$, Fläche = rel. Häufigkeit
- **CDF**: $F(x) = \int_{-\infty}^{x} f(t)dt$, stetige Funktion

Zusammenhänge:

- $f(x) = F'(x)$ (für stetige Merkmale)
- $F(b) - F(a) = P(a < X \leq b)$ (Wahrscheinlichkeit im Intervall)
- Stets gilt: $0 \leq F(x) \leq 1$ und F monoton steigend

Häufigkeiten und Verteilungsfunktionen für stetige Merkmale

PMF (Probability Mass Function) relative Häufigkeitsfunktion

$$f(x) = P(X = x) = \frac{h(x)}{n}$$

- $f(x)$ ist die Wahrscheinlichkeit, dass X den Wert x annimmt
- Darstellung: Höhe der Säule/des Balkens entspricht $f(x)$
- Eigenschaften:
  - Summe = 1
  - $0 \leq f(x) \leq 1$
  - Keine Interpolation zwischen Werten

CDF (Cumulative Distribution Function)

$$F(x) = P(X \leq x) = \sum_{r \leq x} f(r)$$

- $F(x)$ ist die Wahrscheinlichkeit, dass X kleiner oder gleich x ist
- Darstellung: Treppenfunktion
- Eigenschaften:
  - Monoton steigend
  - Rechtsseitig stetig
  - Sprünge an den Ausprägungen
  - $0 \leq F(x) \leq 1$

Erstellen einer Häufigkeitsverteilung

1. Sammle alle verschiedenen Werte
2. Zähle absolute Häufigkeiten:
   - Wie oft kommt jeder Wert vor?
3. Berechne relative Häufigkeiten:
   - Teile jede absolute Häufigkeit durch n
4. Berechne kumulative Häufigkeiten:
   - Absolute: Summiere h_i von links nach rechts
   - Relative: Summiere f_i von links nach rechts

Würfelwurf Ein Würfel wird 20 Mal geworfen:

a_i	1	2	3	4	5	6	Total
h_i	4	3	4	0	6	3	20
f_i	4/20	3/20	4/20	0	6/20	3/20	1

Anwendung der Verteilungsfunktionen

1. Für kleine diskrete Datensätze: PMF und diskrete CDF verwenden
2. Für große stetige Datensätze:
   - Klassierung durchführen
   - PDF und stetige CDF berechnen
3. Bei klassierten Daten:
   - Klassenbreite beachten
   - Häufigkeitsdichten berechnen
4. Bei der Visualisierung:
   - Säulendiagramm für PMF
   - Histogramm für PDF
   - Treppenfunktion für diskrete CDF
   - Stetige Funktion für stetige CDF

Häufigkeiten und Verteilungsfunktionen für stetige/klassierte Merkmale

Klassierung von Daten Bei grossen Stichproben metrisch stetiger Merkmale teilt man die Stichprobenwerte in Klassen ein:

- Die Klassen sind aneinandergrenzende Intervalle
- Obere Intervallgrenzen zählen immer zum darauffolgenden Intervall
- Relative Häufigkeit eines Intervalls = Anzahl enthaltener Stichprobenwerte / Stichprobengrösse
- Die relative Häufigkeit eines Intervalls entspricht der Fläche des darüber liegenden Rechtecks im Histogramm

Klassenbildung (Faustregeln)

- Die Klassen sollten gleich breit gewählt werden
- Die Anzahl der Klassen sollte etwa zwischen 5 und 20 liegen
- Die Anzahl der Klassen sollte $\sqrt{n}$ nicht wesentlich überschreiten
- Klassengrenzen sollten 'runde' Zahlen sein
- Werte auf Klassengrenzen kommen in die obere Klasse

Absolute Häufigkeitsdichte: $h = \frac{h_i}{d_i}$

Bei klassierten Daten wird die Häufigkeit als Rechtecksfläche über der Klassenbreite d_i dargestellt. Höhe des Rechtecks entspricht der absoluten Häufigkeitsdichte.

PDF (Probability Density Function) $f = \frac{f_i}{d_i}$

- $f(x)$ ist die Dichte der Verteilungsfunktion $F(x)$ (relative Häufigkeitsdichte)
- Darstellung: Fläche unter der Kurve entspricht $F(x)$
- Bei Histogramm: Rechteckfläche = relative Häufigkeit der Klasse

CDF Kumulative Verteilungsfunktion für klassierte Daten

Durch Integration der relativen Häufigkeitsfunktion (PDF) $f(x)$ erhält man die kumulative Verteilungsfunktion (CDF):

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

Eigenschaften der CDF

- $F(x)$ ist stetig, monoton steigend und stückweise differenzierbar
- Die Werte von $F(x)$ an den rechten Klassengrenzen erhält man durch Kumulieren der relativen Häufigkeiten f_i im kompletten Intervall
- $F(x) = \sum_{r \leq x} f(r)$ mit der relativen Häufigkeitsfunktion (PMF)
- $0 \leq F(x) \leq 1$ für alle reellen Zahlen x
- Der Graph von $F(x)$ ist eine rechtsseitig stetige Treppenfunktion
- Es gibt eine reelle Zahl x mit $F(x) = 0$ und y mit $F(y) = 1$
- Der Anteil aller Stichprobenwerte x_i im Bereich $a < x_i \leq b$ berechnet sich als $F(b) - F(a)$

Berechnung der CDF für klassierte Daten

1. Bestimme für jede Klasse: d_i, h_i, f_i
2. Bestimme kumulative Häufigkeiten H_i
3. CDF Berechnung:
   3.1 Bestimme kumulative Häufigkeiten H_i
   3.2 Teile durch Stichprobengröße: $F(x) = \frac{H(x)}{n}$
4. Werte der CDF:
   - An linker Klassengrenze: F(x) entspricht kumulierter Häufigkeit bis vorherige Klasse
   - An rechter Klassengrenze: F(x) entspricht kumulierter Häufigkeit bis aktuelle Klasse

Programmlaufzeiten Ein Programm wird auf 20 Rechnern ausgeführt. Folgende Laufzeiten (in ms) werden gemessen: 400, 399, 398, 400, 398, 399, 397, 400, 402, 399, 401, 399, 400, 402, 398, 400, 399, 401, 399, 399

a_i	397	398	399	400	401	402	Total
h_i	1	3	7	5	2	2	20
f_i	1/20	3/20	7/20	5/20	2/20	2/20	1
H_i	1	4	11	16	18	20	
F_i	1/20	4/20	11/20	16/20	18/20	1	

Kenngrössen

Arten von Kenngrössen

- **Lagemasse:** Beschreiben das Zentrum der Verteilung
- **Streuungsmasse:** Charakterisieren die Abweichung vom Zentrum
- **Schiefemasse:** Beschreiben die Form der Verteilung

Quantile

Quantile Für eine reelle Zahl $0 \leq q \leq 1$ heisst eine Zahl R ein q-Quantil der Stichprobe $x_1, x_2, ..., x_n$, falls:

- Der Anteil der Stichprobenwerte $x_i \leq R$ mindestens q ist
- Der Anteil der Stichprobenwerte $x_i \geq R$ mindestens $1 - q$ ist

Die 0.25, 0.5 und 0.75-Quantile werden auch 1., 2. und 3. Quartil genannt.

Quantil $Q = x_i = x_{\lceil n \cdot q \rceil}$

Position des Quantils: $i = \lceil n \cdot q \rceil$
n: Anzahl der Beobachtungen
q: Quantilswert (zB. 0.25 für Q1)
x_i: Beobachtung an Position i.

Interquartilsabstand

$$IQR = Q_3 - Q_1$$

Q_3: Oberes Quartil (75%)
Q_1: Unteres Quartil (25%)

Berechnung von Quantilen

Für eine geordnete Stichprobe $x_{[1]} \leq x_{[2]} \leq ... \leq x_{[n]}$:

1. Berechne $n \cdot q$
2. Falls $n \cdot q$ eine ganze Zahl ist: $R_q = \frac{1}{2}(x_{n \cdot q} + x_{n \cdot q + 1})$
3. Falls $n \cdot q$ keine ganze Zahl ist: $R_q = x_{\lceil n \cdot q \rceil}$
4. Wobei $\lceil n \cdot q \rceil$ die nächstgrössere ganze Zahl ist

Berechnung von Lageparametern

1. Sortiere die Daten aufsteigend
2. Berechne den Mittelwert: Summe aller Werte / Anzahl Werte
3. Bestimme den Median:
   - Bei ungerader Anzahl: mittlerer Wert
   - Bei gerader Anzahl: Mittelwert der beiden mittleren Werte
4. Finde den Modus (häufigster Wert)
5. Berechne die Quartile:
   - Q1: 25%-Quantil, Q2: Median, Q3: 75%-Quantil

Berechnung von Quantilen Datenreihe: 2, 4, 4, 5, 7, 8, 9, 10 ($n = 8$)

Berechnung Q1 (25%-Quantil): Q1 = $x_2 = 4$
- $i = \lceil 8 \cdot 0.25 \rceil = \lceil 2 \rceil = 2$

Berechnung Q2 (Median): Q2 = $(5 + 7)/2 = 6$
- n gerade → Mittelwert von Position 4 und 5

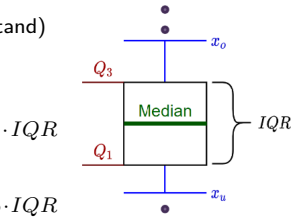
Berechnung Q3 (75%-Quantil): Q3 = $x_6 = 8$
- $i = \lceil 8 \cdot 0.75 \rceil = \lceil 6 \rceil = 6$

Interquartilsabstand: IQR = Q3 - Q1 = 8 - 4 = 4

Boxplot

Boxplot besteht aus:

- Box: Begrenzt durch Q_1 und Q_3
- Mittellinie: Median = $Q_2 = x_{\text{med}}$
- $IQR = Q_3 - Q_1$ (Interquartilsabstand)
- Antennen (Whisker):
  Untere Antenne: x_u :
  $u = \min[Q_1 - 1.5 \cdot IQR, Q_1]$
  → Minimum der Werte $\geq Q_1 - 1.5 \cdot IQR$
  Obere Antenne: x_0 :
  $o = \max[Q_3 + 1.5 \cdot IQR, Q_3]$
  → Maximum der Werte $\leq Q_3 + 1.5 \cdot IQR$
- Ausreisser: alle Werte ausserhalb der Antennen: $x_i < x_u \vee x_i > x_0$



Erstellen eines Boxplots

1. Berechne die Quartile Q_1, Q_2 (Median) und Q_3
2. Bestimme den Interquartilsabstand IQR = $Q_3 - Q_1$
3. Berechne die Grenzen für Ausreisser:
   - Untere Grenze: $Q_1 - 1.5 \cdot IQR$ und Obere Grenze: $Q_3 + 1.5 \cdot IQR$
4. Zeichne Box mit:
   - Unterer Rand bei Q_1, Mittelline bei Q_2, Oberer Rand bei Q_3
5. Zeichne Antennen bis zum:
   - Kleinsten Wert $\geqslant$ untere Grenze
   - Grössten Wert $\leqslant$ obere Grenze
6. Markiere alle Werte ausserhalb als Ausreisser

Boxplot - Praktisches Beispiel Messwerte: 2, 3, 5, 6, 7, 8, 9, 15, 50

1. Sortiere Werte: 2, 3, 5, 6, 7, 8, 9, 15, 50
2. Bestimme Quartile:
   - $Q_1 = 4$ (25%-Quantil), $Q_2 = 7$ (Median), $Q_3 = 12$ (75%-Quantil)
3. IQR = 12 - 4 = 8
4. Ausreisser-Grenzen:
   - Untere: 4 - 1.5 · 8 = -8
   - Obere: 12 + 1.5 · 8 = 24
5. 50 ist ein Ausreisser (> 24)

Lagekennwerte/Lageparameter

Arithmetisches Mittel $\bar{x}$ ist der Durchschnitt der Stichprobenwerte:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} a_i \cdot f_i$$

- a_i: Klassenmitte
- x_i: Einzelner Stichprobenwert
- f_i: Relative Häufigkeit

Median Das 2. Quartil wird auch Median oder Zentralwert genannt:

$$\text{Median}(x_1, ..., x_n) = x_{\text{med}} = \begin{cases} x_{[\frac{n+1}{2}]} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}) & \text{falls } n \text{ gerade} \end{cases}$$

teilt Datensatz in zwei gleich grosse Hälften

Modus x_{mod} = Häufigster Wert in der Stichprobe

- Mittelwert reagiert empfindlich auf Ausreißer (A)
- Median ist robuster gegen Ausreißer
- Modus zeigt Häufungen, kann mehrfach auftreten

Streuungskennwerte

- s: Stichprobenstandardabweichung
- s_{kor}: Korrigierte Stichprobenstandardabweichung
- s^2: Stichprobenvarianz
- s^2_{kor}: Korrigierte Stichprobenvarianz
- $\bar{x}$: Arithmetisches Mittel
- x_i: Einzelner Stichprobenwert

Streuungsmasse

Stichprobenvarianz:

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

Korrigierte Stichprobenvarianz:

$$s^2_{\text{kor}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}s^2$$

Standardabweichung:

$$s = \sqrt{s^2} = \sqrt{\overline{x^2} - \bar{x}^2} \quad \text{bzw.} \quad s_{\text{kor}} = \sqrt{s^2_{\text{kor}}}$$

Berechnung der Stichprobenvarianz

1. Berechne den Mittelwert $\bar{x}$
2. Für jeden Wert x_i:
   2.1 Berechne Abweichung vom Mittelwert $(x_i - \bar{x})$
   2.2 Quadriere die Abweichung $(x_i - \bar{x})^2$
3. Summiere alle quadrierten Abweichungen
4. Teile durch $(n-1)$ für korrigierte Varianz
5. Alternative Berechnung:
   5.1 Berechne $\overline{x^2}$ (Mittelwert der quadrierten Werte)
   5.2 Berechne $(\bar{x})^2$ (Quadrat des Mittelwerts)
   5.3 Varianz $= \overline{x^2} - (\bar{x})^2$

Berechnung von Varianz und Standardabweichung

Gegeben sei die Datenreihe: 2, 4, 4, 6, 9

Schritt 1: Mittelwert berechnen: $\bar{x} = \frac{2+4+4+6+9}{5} = 5$
Schritt 2: Abweichungen quadrieren:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	-3	9
4	-1	1
4	-1	1
6	1	1
9	4	16

Schritt 3: Varianz berechnen: $s^2_{\text{kor}} = \frac{9+1+1+1+16}{5-1} = \frac{28}{4} = 7$

Schritt 4: Standardabweichung berechnen: $s_{\text{kor}} = \sqrt{7} \approx 2.65$

Alternative Berechnung:

- $\overline{x^2} = \frac{4+16+16+36+81}{5} = 30.6$
- $(\bar{x})^2 = 5^2 = 25$
- $s^2 = 30.6 - 25 = 5.6$
- $s^2_{\text{kor}} = \frac{5}{4} \cdot 5.6 = 7$

Form der Verteilung

Verteilungsformen

- **Symmetrisch:** Rechte und linke Hälfte spiegelbildlich
- **Linkssteil (rechtsschief):**
  - Daten links konzentriert
  - $x_{\text{mod}} < x_{\text{med}} < \bar{x}$
- **Rechtssteil (linksschief):**
  - Daten rechts konzentriert
  - $x_{\text{mod}} > x_{\text{med}} > \bar{x}$
- **Modalität:**
  - Unimodal: Ein Maximum
  - Bimodal/Multimodal: Mehrere Maxima

Deskriptive Statistik (mehrere Merkmale)

Multivariate Daten

Multivariate Daten

- **Bivariate Daten:** Zwei Merkmale pro Merkmalsträger
- **Multivariate Daten:** Mehrere Merkmale pro Merkmalsträger

Grafische Darstellung

Darstellungsformen nach Merkmalstypen (Bivariate Daten)

- **Zwei kategorielle Merkmale:** Kontingenztabelle + Mosaikplot
- **Ein kategorielles + ein metrisches Merkmal:** Boxplot oder Stripchart
  - Kennwerte pro Kategorie
- **Zwei metrische Merkmale:** Streudiagramm (Scatterplot)
  - Punktwolke in der (x,y)-Ebene

Kontingenztabelle Studierende nach Studiengang und Geschlecht:

	Männlich	Weiblich	Total
Informatik	120	30	150
Wirtschaft	80	70	150
Total	200	100	300

Analyse von Streudiagrammen

1. Untersuche die **Form** des Zusammenhangs:
   - Linear: Punkte streuen um Gerade
   - Gekrümmt: Punkte folgen einer Kurve
   - Mehrere Punktwolken vorhanden?
2. Bestimme die **Richtung**:
   - Positiv: y-Werte steigen mit x-Werten
   - Negativ: y-Werte fallen mit x-Werten
   - Kein Trend erkennbar
3. Beurteile die **Stärke**:
   - Wenig Streuung: starker Zusammenhang (Punkte nahe an Gerade)
   - Große Streuung: schwacher Zusammenhang
   - Auf Ausreisser achten

Darstellung multivariater Daten

- **Kategorielle Merkmale:**
  - Mehrdimensionale Kontingenztabellen
  - Farbliche Codierung zusätzlicher Dimensionen
- **Metrische Merkmale:**
  - Matrix von Streudiagrammen
  - Korrelationsmatrix

Varianz und Kovarianz

Abkürzungen

Mittelwert x-Werte: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Mittelwert y-Werte: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ Mittelwert Produkte: $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$

Varianz und Kovarianz

Die **Varianz** ist ein Maß für die Streuung eines Merkmals:

$$(s_x)^2 = \overline{x^2} - \bar{x}^2, \quad (s_y)^2 = \overline{y^2} - \bar{y}^2$$

Die **Kovarianz** ist ein Maß für den linearen Zusammenhang:

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

Berechnung der Kovarianz

1. Methode (direkte Formel):
   - Berechne Mittelwerte $\bar{x}$ und $\bar{y}$
   - Für jedes Paar (x_i, y_i): Berechne $(x_i - \bar{x})(y_i - \bar{y})$
   - Summiere alle Produkte und teile durch n
2. Methode (schnellere Berechnung):
   - Berechne $\overline{xy}$ (Mittelwert der Produkte) und $\bar{x} \cdot \bar{y}$
   - Kovarianz $= \overline{xy} - \bar{x} \cdot \bar{y}$

Rang $rg(x_i)$ des Stichprobenwertes x_i ist definiert als der Index von x_i in der nach der Grösse geordneten Stichprobe.

i	1	2	3	4	5	6
x_i	23	27	35	35	42	59
$rg(x_i)$	1	2	3.5	3.5	5	6

Rang-Varianz und Kovarianz

Varianz (Ränge) $(s_{rg(x)})^2, (s_{rg(y)})^2$:

$$(s_{rg(x)})^2 = \overline{rg(x)^2} - (\overline{rg(x)})^2, \quad (s_{rg(y)})^2 = \overline{rg(y)^2} - (\overline{rg(y)})^2$$

Kovarianz (Ränge) $s_{rg(xy)}$:

$$s_{rg(xy)} = \overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)} = \overline{rg(xy)} - \frac{(n+1)^2}{4}$$

Rangberechnung und Bindungen

1. Sortiere die Werte aufsteigend
2. Ränge zuweisen: Kleinster Wert: Rang 1, Zweitkleinster: Rang 2, ...
3. Bei Bindungen (gleiche Werte):
   - Identifiziere gleiche Werte
   - Berechne Durchschnittsrang: $\frac{\text{Summe der Rangplätze}}{\text{Anzahl gebundener Werte}}$
   - Weise allen gleichen Werten diesen Rang zu

Rangberechnung mit Bindungen Datenreihe: 3, 7, 7, 4, 9, 7, 2
Schritt 1: Sortieren: $2, 3, 4, 7, 7, 7, 9$
Schritt 2: Ränge zuweisen:

- 2: Rang 1
- 3: Rang 2
- 4: Rang 3
- 7: Durchschnittsrang $\frac{4+5+6}{3} = 5$
- 9: Rang 7

Schritt 3: Finale Rangzuordnung:

Wert	3	7	7	4	9	7	2
Rang	2	5	5	3	7	5	1

Korrelation

Korrelationskoeffizient nach Pearson normiert die Kovarianz:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

Eigenschaften:
- $-1 \leq r_{xy} \leq 1$
- $r_{xy} \approx 1$: starker positiver linearer Zusammenhang
- $r_{xy} \approx -1$: starker negativer linearer Zusammenhang
- $r_{xy} \approx 0$: kein linearer Zusammenhang

Interpretation des Korrelationskoeffizienten

Verschiedene Datensätze mit jeweils 20 (x, y)-Paaren:
Fall A: $r_{xy} = 0.95 \rightarrow$ Starker positiver linearer Zusammenhang
- y steigt fast proportional mit x
- Nur geringe Streuung um die Regressionsgerade

Fall B: $r_{xy} = -0.82 \rightarrow$ Starker negativer linearer Zusammenhang
- y sinkt mit steigendem x
- Moderate Streuung vorhanden

Fall C: $r_{xy} = 0.12 \rightarrow$ Kaum linearer Zusammenhang
- Starke Streuung der Punkte
- Möglicherweise nichtlinearer Zusammenhang

Rangkorrelationskoeffizient nach Spearman
Für monotone Zusammenhänge:

$$r_{sp} = \frac{s_{rg(xy)}}{s_{rg(x)} \cdot s_{rg(y)}} = \frac{\overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)}}{\sqrt{\overline{rg(x)^2} - (\overline{rg(x)})^2} \cdot \sqrt{\overline{rg(y)^2} - (\overline{rg(y)})^2}}$$

Vereinfachte Formel, sofern **alle Ränge unterschiedlich** sind:

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

mit $d_i = rg(x_i) - rg(y_i)$ (Rangdifferenzen)

Berechnung des Spearman-Korrelationskoeffizienten
1. Weise beiden Merkmalen Ränge zu:
   - Sortiere x-Werte, vergebe Ränge (ebenfalls für y-Werte)
   - Bei Bindungen: Durchschnittsränge
2. Falls keine Bindungen vorhanden:
   - Berechne Rangdifferenzen d_i
   - Quadriere Differenzen d_i^2 und summiere sie
   - Verwende vereinfachte Formel für r_{sp}
3. Bei Bindungen:
   - Berechne Rangmittelwerte
   - Berechne Rangvarianzen und -kovarianz
   - Verwende allgemeine Formel

Unterschied Pearson und Spearman
- **Pearson:**
  - Misst linearen Zusammenhang
  - Empfindlich gegen Ausreißer
  - Für metrische Daten
- **Spearman:**
  - Misst (nichtlinearen) monotonen Zusammenhang
  - Robust gegen Ausreißer
  - Auch für ordinale Daten

Vergleich Pearson und Spearman

Gegeben seien die Wertepaare: $(1, 1), (2, 4), (3, 9), (4, 16), (5, 25)$

Pearson-Korrelation: $r_{xy} = 0.975$
- Zeigt starken linearen Zusammenhang

Spearman-Korrelation: $r_{sp} = 1.000$
- Perfekter monotoner Zusammenhang

Vergleich:
- Pearson erfasst nur linearen Zusammenhang
- Spearman erfasst jeden monotonen Zusammenhang
- Hier: Quadratischer Zusammenhang
- Spearman robuster gegen Ausreißer

Berechnung von Kovarianz und Korrelation

Gegeben seien die Wertepaare: $(1, 2), (2, 4), (3, 5), (4, 8)$

Schritt 1: Mittelwerte berechnen:

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5, \quad \bar{y} = \frac{2+4+5+8}{4} = 4.75$$

Schritt 2: Kovarianz berechnen: $s_{xy} = 14.25 - 11.875 = 2.375$

$$\overline{xy} = \frac{2+8+15+32}{4} = 14.25, \quad \bar{x} \cdot \bar{y} = 2.5 \cdot 4.75 = 11.875$$

Schritt 3: Korrelationskoeffizient berechnen

$s_x^2 = \frac{1+4+9+16}{4} - 2.5^2 = 1.25, \quad s_y^2 = \frac{4+16+25+64}{4} - 4.75^2 = 5.6875$

$$r_{xy} = \frac{2.375}{\sqrt{1.25} \cdot \sqrt{5.6875}} = 0.894$$

Grenzen der Korrelation

Scheinkorrelation Eine Korrelation zwischen zwei Merkmalen bedeutet nicht automatisch einen kausalen Zusammenhang:
- Ein drittes Merkmal könnte beide beeinflussen
- Der Zusammenhang könnte zufällig sein
- Ausreißer können das Ergebnis verzerren
- Nichtlinearer Zusammenhang möglich

Prüfung auf Scheinkorrelation
1. Betrachte die Datenpunkte im Streudiagramm:
   - Gibt es Ausreißer?
   - Ist der Zusammenhang wirklich linear?
2. Überlege fachlich:
   - Gibt es plausible Kausalität?
   - Könnte ein drittes Merkmal beide beeinflussen?
3. Prüfe Teilstichproben:
   - Bleibt Korrelation in Untergruppen bestehen?
   - Ändert sich die Stärke deutlich?
4. Bei Zweifeln:
   - Spearman-Korrelation prüfen und weitere Merkmale einbeziehen
   - Fachexperten konsultieren (sure, eifach Dozent frage wäred de Prüefig)

Kombinatorik

Grundbegriffe

Fakultät $n!$ einer natürlichen Zahl n ist definiert als das Produkt aller positiven ganzen Zahlen bis zu dieser Zahl:

$$n! = 1 \cdot 2 \cdot \ldots \cdot n = \prod_{a=1}^{n} a \text{ mit } 0! = 1 \text{ als Definitionsvereinbarung}$$

Parameter:
- n = Die positive ganze Zahl, für die die Fakultät berechnet wird
- a = Laufvariable in der Produktnotation
- $\prod$ = Produkt aller Terme von $a = 1$ bis n

Binomialkoeffizient $\binom{n}{k}$ für natürliche Zahlen $0 \leq k \leq n$:

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$ → Anzahl Möglichkeiten, aus n Objekten k Objekte auszuwählen.

Eigenschaften Für den Binomialkoeffizienten gelten:

Leere Menge: $\binom{n}{0} = \binom{n}{n} = 1$

Symmetrie: $\binom{n}{k} = \binom{n}{n-k}$

Pascal'sche Rekursion: $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$

Summe: $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

Berechnung von Binomialkoeffizienten
1. **Prüfe Spezialfälle:** $\binom{n}{0} = \binom{n}{n} = 1$ und $\binom{n}{1} = n$
2. **Nutze Symmetrie:** $\binom{n}{k} = \binom{n}{n-k}$
3. **Pascal'sches Dreieck:** Baue schrittweise auf, nutze Rekursionsformel
4. **Direkte Berechnung:** Nur wenn nötig, kürze vor dem Ausrechnen

Grundlegende Abzählmethoden

Systematik der Kombinatorik

	Mit Wiederholung	Ohne Wiederholung
Variation (Reihenfolge wichtig)	n^k	$\frac{n!}{(n-k)!}$
Kombination (Reihenfolge unwichtig)	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Bestimmung der Abzählmethode
1. **Analysiere das Problem:**
   - n: Anzahl verfügbarer Objekte, k: Anzahl auszuwählender Objekte
2. **Prüfe die Reihenfolge:**
   - Ist die Reihenfolge wichtig? → Variation
   - Ist nur die Auswahl wichtig? → Kombination
3. **Prüfe Wiederholungen:**
   - Dürfen Objekte mehrfach vorkommen? → Mit Wiederholung
   - Darf jedes Objekt nur einmal? → Ohne Wiederholung

Lösen komplexer kombinatorischer Probleme
1. **Problem zerlegen**
   - Teile das Problem in unabhängige Teilprobleme
   - Identifiziere abhängige Entscheidungen
2. **Für jedes Teilproblem:** Wähle passende Abzählmethode
3. **Kombiniere Teillösungen**
   - Unabhängige Ereignisse: Multipliziere
   - Sich ausschließende Ereignisse: Addiere
   - Prüfe Überlappungen (Inklusions-Exclusions)

Elementare Wahrscheinlichkeitsrechnung

Diskrete Wahrscheinlichkeitsräume

Symbole

- Ω: Ergebnisraum (Menge aller möglichen Ergebnisse)
- ω: Ergebnis eines Zufallsexperiments
- $|\Omega|$: Anzahl der Elemente im Ergebnisraum
- $P : \Omega \to [0, 1]$: Wahrscheinlichkeitsmaß (Zähldichte) ordnet jedem Ergebnis $\omega \in \Omega$ seine Wahrscheinlichkeit zu, wobei $\sum_{\omega \in \Omega} P(\omega) = 1$ gilt
- 2^Ω: Ereignisraum (Menge aller möglichen Ereignisse)
- $P(A)$: Wahrscheinlichkeit des Ereignisses A
- $|A|$: Anzahl der Elemente im Ereignis A
- A, B, C: Ereignisse (Teilmengen von Ω)
- $\bar{A}$: Komplementärereignis von A

Zufallsexperiment

folgende Bedingungen müssen erfüllt sein:

- Der Vorgang lässt sich unter den gleichen äußeren Bedingungen beliebig oft wiederholen
- Es sind mehrere sich gegenseitig ausschließende Ergebnisse möglich
- Das Ergebnis lässt sich nicht mit Sicherheit voraussagen, sondern ist zufallsbedingt

Ereignisse und Wahrscheinlichkeitsraum

Das **Wahrscheinlichkeitsmaß** $P : 2^\Omega \to [0, 1]$ ist definiert durch:

$$P(A) = \sum_{\omega \in A} \rho(\omega) \text{ für } A \subseteq \Omega$$

Ein **Laplace-Raum** liegt vor, wenn alle Elementarereignisse gleich wahrscheinlich sind:

$$P(A) = \frac{|A|}{|\Omega|}$$

Eigenschaften von Wahrscheinlichkeitsräumen

Für einen diskreten Wahrscheinlichkeitsraum (Ω, P) gelten:

- Unmögliches Ereignis: $P(\emptyset) = 0$
- Sicheres Ereignis: $P(\Omega) = 1$
- Komplementäres Ereignis: $P(\Omega \setminus A) = 1 - P(A)$
- Vereinigung: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Sigma-Additivität: Für paarweise disjunkte Ereignisse gilt: $P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2) + P(A_3) + ...$

Wahrscheinlichkeits-Ausdrücke und Regeln

- $P(A)$ = Wahrscheinlichkeit von Ereignis A
- $P(B)$ = Wahrscheinlichkeit von Ereignis B
- $P(\bar{A})$ = Wahrscheinlichkeit des Gegenereignisses von A
- $P(B|A)$ = Wahrscheinlichkeit von B unter der Bedingung dass A eingetreten ist
- $P(B|\bar{A})$ = Wahrscheinlichkeit von B unter der Bedingung dass A nicht eingetreten ist
- $P(A \cap B)$ = Wahrscheinlichkeit dass beide Ereignisse eintreten
- $P(A \cup B)$ = Wahrscheinlichkeit dass mindestens eines der Ereignisse eintritt
- **Additionssatz:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Komplementärregel:** $P(\bar{A}) = 1 - P(A)$
- **Multiplikationssatz:** $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

Strategien zur Berechnung von Wahrscheinlichkeiten

Grundschritte der Wahrscheinlichkeitsberechnung

1. **Ergebnisraum identifizieren**
   - Alle möglichen Ergebnisse auflisten
   - Prüfen, ob es sich um einen Laplace-Raum handelt
2. **Ereignis präzisieren**
   - Exakte mathematische Beschreibung des gesuchten Ereignisses
   - Zerlegung in Teilmengen falls nötig
3. **Berechnungsstrategie wählen**
   - Direkte Berechnung: $P(A) = \frac{|A|}{|\Omega|}$
   - Über Gegenereignis: $P(A) = 1 - P(\bar{A})$
   - Über bedingte Wahrscheinlichkeit falls abhängig
4. **Berechnung durchführen** Kombinatorische Formeln anwenden
   - Zwischenergebnisse notieren
   - Probe durch Plausibilitätskontrolle

Problemlösung mit Gegenereignis

Oft einfacher

1. **Prüfe, ob Gegenereignis einfacher ist**
   - Original: "Mindestens eine..." oder "Mehr als..."
   - Gegenereignis: "Keine..." oder "Höchstens..."
2. **Berechne Wahrscheinlichkeit des Gegenereignis**
   - Oft einfacher zu zählen
   - Weniger Fälle zu berücksichtigen
3. **Wende Komplementärregel an:** $P(A) = 1 - P(\bar{A})$
   - Überprüfe Plausibilität des Ergebnisses

Zufallsvariablen

Symbole

- X, Y, Z: Zufallsvariablen (Funktionen von Ω nach $\mathbb{R}$)
- x, y, z: Mögliche Werte der Zufallsvariablen
- $P(X = x)$: Wahrscheinlichkeit, dass X den Wert x annimmt
- $P(X \leq x)$: Wahrscheinlichkeit, dass X kleiner oder gleich x ist
- $P(X = x, Y = y)$: Wahrscheinlichkeit, dass $X = x$ und $Y = y$ sind
- $f(x)$: Wahrscheinlichkeitsfunktion (PMF) von X
- $F(x)$: Verteilungsfunktion (CDF) von X
- $E(X)$: Erwartungswert von X
- $V(X)$: Varianz von X
- $S(X)$: Standardabweichung von X
- α, β, γ: Konstanten
- $\sum_{x \in \mathbb{R}}$ = Summe über alle möglichen Werte von $x \in \mathbb{R}$

Zufallsvariablen

Eine **Zufallsvariable** X ist eine Funktion $X : \Omega \to \mathbb{R}$, die jedem Ergebnis eine reelle Zahl zuordnet.
Die **Wahrscheinlichkeitsfunktion** (PMF) ist definiert durch:

$$f(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

Die **Verteilungsfunktion** (CDF) ist definiert durch:

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

Eigenschaften von PMF und CDF

- $\sum_{x \in \mathbb{R}} f(x) = 1$ und $F(x) = \sum_{t \leq x} f(t)$
- $\lim_{x \to \infty} F(x) = 1$ und $\lim_{x \to -\infty} F(x) = 0$
- Monotonie: $x \leq y \Rightarrow F(x) \leq F(y)$
- $P(a < X \leq b) = F(b) - F(a)$

Kenngrössen

Erwartungswert und Varianz

Für eine diskrete Zufallsvariable X:

Erwartungswert: $E(X) = \sum_{x \in \mathbb{R}} x \cdot f(x)$

Varianz: $V(X) = E((X - E(X))^2) = \sum_{x \in \mathbb{R}} (x - E(X))^2 \cdot f(x)$

Standardabweichung: $S(X) = \sqrt{V(X)}$

Rechenregeln

für **stochastisch unabhängige** Zufallsvariablen X, Y:

- **Addition:** $E(X + Y) = E(X) + E(Y)$, $V(X \pm Y) = V(X) + V(Y)$
- **Multiplikation:** $E(X \cdot Y) = E(X) \cdot E(Y)$
- **Linearität:** $E(aX + b) = aE(X) + b$
- **Verschiebungssatz:** $V(X) = E(X^2) - (E(X))^2$ wobei $E(X^2) = \sum_{x \in \mathbb{R}} P(X = x) \cdot x^2$
- **Lineare Transformation:** $V(aX + b) = a^2 V(X)$

Erwartungswert und Varianz

Für eine stetige Zufallsvariable X:

$$E(X) = \int_{-\infty}^{\infty} f(x) \cdot x dx \quad V(X) = \int_{-\infty}^{\infty} f(x) \cdot (x - E(X))^2 dx$$

Berechnung von Erwartungswert und Varianz

1. **Erwartungswert bestimmen:**
Formel je nach Art der Zufallsvariable (diskret/stetig)
2. **Varianz berechnen:** direkt (Formel) oder über Verschiebungssatz
3. **Bei Standardabweichung:** Wurzel aus Varianz ziehen
   - Einheit beachten (gleich wie Ursprungsdaten)

Erwartungswert bei Würfelspiel

Bei einem Würfelspiel gewinnt man:

- Bei 6: 5€, bei 5: 2€, bei 1-4: verliert man 1€

1. **Wahrscheinlichkeiten und Werte aufstellen:**
   - P(X = 5€) = 1/6, P(X = 2€) = 1/6, P(X = -1€) = 4/6
2. **Erwartungswert berechnen:**

$$E(X) = 5 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + (-1) \cdot \frac{4}{6} = \frac{5 + 2 - 4}{6} = \frac{3}{6} = 0.5$$

3. **Varianz berechnen:**

$$E(X^2) = 25 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 1 \cdot \frac{4}{6} = \frac{25 + 4 + 4}{6} = \frac{33}{6}$$

$$V(X) = E(X^2) - (E(X))^2 = \frac{33}{6} - (\frac{1}{2})^2 = \frac{33}{6} - \frac{1}{4} \approx 5.25$$

Interpretation:

- Positiver Erwartungswert: Spiel ist langfristig profitabel
- Hohe Varianz: Große Schwankungen möglich

Interpretation von Erwartungswert und Varianz

Erwartungswert: Nicht unbedingt ein möglicher Wert
- Langfristiger Durchschnitt, Schwerpunkt der Verteilung

Varianz: Mass für die Streuung (Quadratische Einheit beachten)
- Je größer, desto unsicherer die Vorhersage

Standardabweichung: Typische Abweichung vom Mittelwert
- Oft für Konfidenzintervalle verwendet (Gleiche Einheit wie Daten)

Stochastische Unabhängigkeit

Stochastische Unabhängigkeit Ereignisse

Zwei Ereignisse A und B heißen **stochastisch unabhängig**, falls:

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

Eigenschaften der stochastischen Unabhängigkeit

Für unabhängige Ereignisse A und B gilt:

- A und $\Omega \setminus B$ sind unabhängig
- $\Omega \setminus A$ und $\Omega \setminus B$ sind unabhängig
- $P(A|B) = P(A)$ falls $P(B) > 0$

Stochastische Unabhängigkeit Zufallsvariablen

Zwei Zufallsvariablen X und Y heißen **stochastisch unabhängig**, falls für alle $x, y \in \mathbb{R}$:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Prüfung auf stochastische Unabhängigkeit

1. **Für Ereignisse**
   - Berechne $P(A \cap B)$ und $P(A) \cdot P(B)$
   - Vergleiche die Werte
2. **Für Zufallsvariablen**
   - Stelle Verbundverteilung auf und prüfe für alle Wertepaare: $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
   - Alternative: Prüfe Kovarianz $= 0$
3. **Praktische Überlegungen**
   - Physikalische/logische Abhängigkeit?
   - Kausaler Zusammenhang?
   - Gemeinsame Einflussfaktoren?

Würfelwurf und Münzwurf

Aufgabe: Ein Würfel wird geworfen und eine Münze geworfen.
Ereignisse:

- A: "Würfel zeigt eine 6"
- B: "Münze zeigt Kopf"

Lösung:

1. **Einzelwahrscheinlichkeiten:**
   - $P(A) = \frac{1}{6}$
   - $P(B) = \frac{1}{2}$
2. **Schnittwahrscheinlichkeit:** $P(A \cap B) = \frac{1}{12} = \frac{1}{6} \cdot \frac{1}{2} = P(A) \cdot P(B)$
3. **Schlussfolgerung:** Die Ereignisse sind stochastisch unabhängig

Kartenziehen ohne Zurücklegen

Aufgabe: Aus einem Kartenspiel werden nacheinander zwei Karten gezogen.
Ereignisse:

- A: Ërste Karte ist Herz"
- B: SZweite Karte ist Herz"

Lösung:

1. **Wahrscheinlichkeiten:**
   - $P(A) = \frac{13}{52} = \frac{1}{4}$
   - $P(B|A) = \frac{12}{51}$
   - $P(B|\bar{A}) = \frac{13}{51}$
2. **Prüfung:**

$$P(B) = \frac{13}{52} \neq P(B|A)$$

3. **Schlussfolgerung:** Die Ereignisse sind stochastisch abhängig

Bedingte Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit

von B unter der Bedingung A ist:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{für } P(A) > 0$$

Multiplikationssatz

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

Anwendung:

- Berechnung von Schnittwahrscheinlichkeiten
- Prüfung auf stochastische Unabhängigkeit
- Zerlegung von mehrstufigen Experimenten

Erstellen einer Vierfeldertafel

1. **Aufbau der Tabelle**
   - Zeilen: Erstes Merkmal (A und nicht A)
   - Spalten: Zweites Merkmal (B und nicht B)
   - Randwahrscheinlichkeiten notieren
2. **Eintragen der Wahrscheinlichkeiten**
   - Schnittwahrscheinlichkeiten in die Felder
   - Zeilensummen = P(A) bzw. P(nicht A)
   - Spaltensummen = P(B) bzw. P(nicht B)
3. **Berechnung bedingter Wahrscheinlichkeiten**
   - $P(B|A) = \frac{P(A \cap B)}{P(A)}$ und $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Medizinischer Test

Aufgabe: Ein Test auf eine Krankheit hat folgende Eigenschaften:

- 1% der Bevölkerung hat die Krankheit
- Test ist bei Kranken zu 98% positiv
- Test ist bei Gesunden zu 95% negativ

Lösung mit Vierfeldertafel:

	Test +	Test -	Summe
Krank	0.0098	0.0002	0.01
Gesund	0.0495	0.9405	0.99
Summe	0.0593	0.9407	1

Berechnung: Wahrscheinlichkeit krank bei positivem Test:

$$P(\text{krank}|\text{positiv}) = \frac{0.0098}{0.0593} \approx 0.165 = 16.5\%$$

Satz der Totalen Wahrscheinlichkeit

$$P(B) = P(A) \cdot P(B|A) + P(\bar{A}) \cdot P(B|\bar{A})$$

Anwendung:

- Berechnung von P(B) durch Fallunterscheidung
- Basis für den Satz von Bayes
- Wichtig bei Entscheidungsbäumen

Ereignisbäume

1. **Aufbau**
   - Von links nach rechts zeichnen
   - Alle Verzweigungen vollständig angeben
   - Übergangswahrscheinlichkeiten an Äste schreiben
2. **Pfadwahrscheinlichkeiten**
   - Multiplikation entlang des Pfades
   - Für jedes Endereignis alle Pfade addieren
   - Summe aller Pfadwahrscheinlichkeiten = 1

Satz von Bayes

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

Anwendung:

- Umkehrung bedingter Wahrscheinlichkeiten
- Aktualisierung von Wahrscheinlichkeiten
- Diagnostische Tests

Anwendung des Satzes von Bayes

1. **Identifiziere die bekannten Größen**
   - A priori Wahrscheinlichkeit P(A)
   - Bedingte Wahrscheinlichkeit P(B|A)
   - Totale Wahrscheinlichkeit P(B)
2. **Berechne P(B) falls nötig**
   - Nutze Satz der totalen Wahrscheinlichkeit
   - $P(B) = P(A) \cdot P(B|A) + P(\bar{A}) \cdot P(B|\bar{A})$
3. **Berechne P(A|B)**
   - Setze in Bayes-Formel ein
   - Interpretiere das Ergebnis

Qualitätskontrolle

Aufgabe: Eine Maschine produziert Teile.

- 95% der Teile sind fehlerfrei
- Ein Test erkennt fehlerhafte Teile zu 98%
- Der Test klassifiziert 3% der guten Teile falsch

Gesucht: Wahrscheinlichkeit für tatsächlich fehlerhaftes Teil bei positivem Test
Lösung:

- $P(F) = 0.05$ (fehlerhaft)
- $P(T|F) = 0.98$ (Test positiv wenn fehlerhaft)
- $P(T|\bar{F}) = 0.03$ (Test positiv wenn gut)
- $P(T) = 0.05 \cdot 0.98 + 0.95 \cdot 0.03 = 0.0775$
- $P(F|T) = \frac{0.05 \cdot 0.98}{0.0775} \approx 0.632 = 63.2\%$

Kovarianz und Korrelation

Kovarianz und Korrelation

Die **Kovarianz** zweier Zufallsvariablen ist:

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Der **Korrelationskoeffizient** ist:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

Eigenschaften:

- $-1 \leq \rho_{XY} \leq 1$
- $\rho_{XY} = \pm 1$: perfekter linearer Zusammenhang
- $\rho_{XY} = 0$: unkorreliert

Anwendung von Kovarianz und Korrelation

1. **Kovarianz berechnen**
   - Direkter Weg: $Cov(X, Y) = E(XY) - E(X)E(Y)$
   - Alternativ: $\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$
2. **Korrelation bestimmen**
   - Kovarianz durch Produkt der Standardabweichungen
   - Normierung auf [-1,1]
3. **Interpretation**
   - Vorzeichen: Richtung des Zusammenhangs
   - Betrag: Stärke des Zusammenhangs
   - Unabhängig von Maßeinheiten

Spezielle Verteilungen

Diskrete und Stetige Zufallsvariablen

Diskrete und Stetige Zufallsvariablen Bei einer **diskreten Zufallsvariable** gibt es immer Lücken zwischen den Werten; sie kann nur bestimmte Werte annehmen.
Eine **stetige Zufallsvariable** hat ein kontinuierliches Spektrum von möglichen Werten.
Berechnung von Wahrscheinlichkeiten:
- Diskret: $P(X = x) = f(x)$ (PMF)
- Stetig: $P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ (CDF)

Gegenüberstellung von diskreten und stetigen Zufallsvariablen
Diskrete Zufallsvariable:
- Dichtefunktion: $f(x) = P(X = x)$
- Verteilungsfunktion: $F(x) = \sum_{x \leq X} f(x)$
- Wahrscheinlichkeiten: $P(a \leq X \leq b) = \sum_{a \leq x \leq b} f(x)$
- Erwartungswert: $E(X) = \sum_{x \in \mathbb{R}} x \cdot f(x)$
- Varianz: $V(X) = \sum_{x \in \mathbb{R}} (x - E(X))^2 \cdot f(x)$

Stetige Zufallsvariable:
- Dichtefunktion: $f(x) = F'(x) \neq P(X = x)$
- Verteilungsfunktion: $F(x) = \int_{-\infty}^{x} f(t)dt$
- Wahrscheinlichkeiten: $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$
- Erwartungswert: $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$
- Varianz: $V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)dx$

Diskrete Verteilungen

Übersicht der diskreten Verteilungen
1. **Hypergeometrische Verteilung:** Ziehen ohne Zurücklegen
   - Endliche Grundgesamtheit
   - Veränderliche Wahrscheinlichkeiten
2. **Bernoulli-Verteilung:** Genau zwei mögliche Ausgänge
   - Ein einzelner Versuch
   - Konstante Erfolgswahrscheinlichkeit
3. **Binomial-Verteilung:** Mehrere unabhängige Versuche
   - Feste Anzahl unabhängiger Versuche
   - Mit Zurücklegen/große Grundgesamtheit
   - Konstante Erfolgswahrscheinlichkeit
4. **Poisson-Verteilung:** Seltene Ereignisse
   - Festes Zeitintervall/Raumbereich
   - Rate λ bekannt

Wahl der richtigen Verteilung
1. **Prüfe Ziehungsart**
   - Mit Zurücklegen → Binomialverteilung
   - Ohne Zurücklegen → Hypergeometrische Verteilung
   - Seltene Ereignisse → Poisson-Verteilung
2. **Prüfe Grundgesamtheit**
   - Endlich, klein → Hypergeometrische Verteilung
   - Sehr groß/unendlich → Binomialverteilung
   - Zeitlich/räumlich kontinuierlich → Poisson-Verteilung
3. **Beachte Approximationen**
   - Binomial → Poisson für $n \to \infty$, $p \to 0$, $np = \lambda$
   - Hypergeometrisch → Binomial für $\frac{n}{N} \leq 0.05$

Hypergeometrische Verteilung
Ziehen **ohne Zurücklegen** aus einer endlichen Grundgesamtheit.

Wahrscheinlichkeitsfunktion: $P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$

Notation: $X \sim H(N, M, n)$
Parameter:
- N: Grundgesamtheit
- M: Anzahl Merkmalsträger
- n: Stichprobenumfang

Kenngrößen:
- $E(X) = n \cdot \frac{M}{N}$
- $V(X) = n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$

Bernoulli-Verteilung Experiment mit genau zwei möglichen Ausgängen (Erfolg/Misserfolg bzw 1/0).

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

Notation: $X \sim B(1, p)$
Parameter:
- p = Erfolgswahrscheinlichkeit
- $q = 1 - p$ = Gegenwahrscheinlichkeit

Kenngrößen:
- $E(X) = E(X^2) = p$
- $V(X) = p \cdot (1 - p) = pq$

Voraussetzungen für die Bernoulli-Verteilung:
- Genau zwei mögliche Ausgänge
- Unabhängige Wiederholungen
- Konstante Erfolgswahrscheinlichkeit

Binomialverteilung
n-malige **unabhängige Wiederholung** eines Bernoulli-Experiments

Wahrscheinlichkeitsfunktion: $P(X = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$

Notation: $X \sim B(n, p)$
Parameter:
- n: Anzahl Versuche
- p: Erfolgswahrscheinlichkeit
- $q = 1 - p$: Gegenwahrscheinlichkeit

Kenngrößen:
- $E(X) = n \cdot p$
- $V(X) = n \cdot p \cdot q$

Poissonverteilung
Modelliert **seltene Ereignisse** in einem festen Intervall.

Wahrscheinlichkeitsfunktion: $P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad \lambda > 0$

Notation: $X \sim Poi(\lambda)$
Parameter:
- λ: Rate/Erwartungswert pro Intervall

Kenngrößen:
- $E(X) = \lambda$
- $V(X) = \lambda$

Stetige Verteilungen

Erwartungswert und Varianz der Normalverteilung
Für eine Zufallsvariable $X \sim N(\mu; \sigma)$ gilt:

$$E(X) = \mu, \quad V(X) = \sigma^2$$

Parameter:
μ = Erwartungswert (Lage)
σ^2 = Varianz
σ = Standardabweichung (Streuung)

Gauss-Verteilung/Normalverteilung Die stetige Zufallsvariable X folgt der Normalverteilung mit den Parametern $\mu, \sigma \in \mathbb{R}, \sigma > 0$:

Dichtefunktion der Normalverteilung: $\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
Notation: $X \sim N(\mu, \sigma)$

Standardnormalverteilung ($\mu = 0$ und $\sigma = 1$):
Dichtefunktion der Standardnormalverteilung: $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$

Eigenschaften der Normalverteilung
- Symmetrisch um μ
- Wendepunkte bei $\mu \pm \sigma$
- Ca. 68% der Werte in $[\mu - \sigma, \mu + \sigma]$
- Ca. 95% der Werte in $[\mu - 2\sigma, \mu + 2\sigma]$
- Ca. 99,7% der Werte in $[\mu - 3\sigma, \mu + 3\sigma]$

Die Verteilungsfunktion der Normalverteilung
Die kumulative Verteilungsfunktion (CDF) von $\varphi_{\mu,\sigma}(x)$ wird mit $\phi_{\mu,\sigma}(x)$ bezeichnet. Sie ist definiert durch:

$$\phi_{\mu,\sigma}(x) = P(X \leq x) = \int_{-\infty}^{x} \varphi_{\mu,\sigma}(t)dt = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt$$

$\varphi_{\mu,\sigma}(x)$ = Dichtefunktion der Normalverteilung
$P(X \leq x)$ = Wahrscheinlichkeit dass X kleiner oder gleich x ist

Standardisierung der Normalverteilung
Bei einer stetigen Zufallsvariable X lässt sich die Verteilungsfunktion als Integral einer Funktion f darstellen:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u) \cdot du$$

Liegt eine beliebige Normalverteilung $N(\mu, \sigma)$ vor, muss standardisiert werden. Statt ursprünglichen Zufallsvariablen X betrachtet man die Zufallsvariable:

$$U = \frac{X - \mu}{\sigma}$$

$F(x)$ = Verteilungsfunktion
$P(X \leq x)$ = Wahrscheinlichkeit dass X kleiner oder gleich x ist
$f(u)$ = Dichtefunktion
U = Standardisierte Zufallsvariable
X = Ursprüngliche Zufallsvariable

Arbeiten mit der Normalverteilung
1. **Standardisierung**
   - $Z = \frac{X-\mu}{\sigma}$ transformiert zu N(0,1)
   - Benutze Tabelle der Standardnormalverteilung
   - Beachte: $\phi(z) = 1 - \phi(-z)$
2. **Stetigkeitskorrektur**
   - Bei Approximation diskreter Verteilungen
   - Untere Grenze: $a - 0.5$
   - Obere Grenze: $b + 0.5$
3. **Faustregel für Intervalle**
   - $\mu \pm \sigma$: ca. 68% der Werte
   - $\mu \pm 2\sigma$: ca. 95% der Werte
   - $\mu \pm 3\sigma$: ca. 99.7% der Werte

Zentraler Grenzwertsatz und Approximationen

Zentraler Grenzwertsatz

Für eine Folge von Zufallsvariablen $X_1, X_2, \ldots, X_n$ mit gleichem Erwartungswert μ und gleicher Varianz σ^2 gilt:

$$E(S_n) = n \cdot \mu, \quad V(S_n) = n \cdot \sigma^2$$

$$E(\bar{X}_n) = \mu, \quad V(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{1}{n^2} \cdot V(S_n)$$

S_n = Summe der Zufallsvariablen
$\bar{X}_n$ = Arithmetisches Mittel der Zufallsvariablen
n = Anzahl der Zufallsvariablen
μ = Erwartungswert der einzelnen Zufallsvariablen
σ^2 = Varianz der einzelnen Zufallsvariablen

Die standardisierte Zufallsvariable:

$$U_n = \frac{((X_1 + X_2 + \cdots + X_n) - n\mu)}{\sqrt{n} \cdot \sigma} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Sind die Zufallsvariablen alle identisch $N(\mu, \sigma)$ verteilt, so sind die Summe S_n und das arithmetische Mittel $\bar{X}_n$ wieder normalverteilt mit:

- S_n : $N(n \cdot \mu, \sqrt{n} \cdot \sigma)$
- $\bar{X}_n$: $N(\mu, \frac{\sigma}{\sqrt{n}})$

Verteilungsfunktion $F_n(u)$ konvergiert für $n \to \infty$ gegen die Verteilungsfunktion $\phi(u)$ der Standardnormalverteilung:

$$\lim_{n\to\infty} F_n(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{u} e^{-\frac{1}{2}t^2} dt$$

Weitere Eigenschaften

Für die Summe $S_n = X_1 + \ldots + X_n$ von n unabhängigen, identisch verteilten Zufallsvariablen mit $E(X_i) = \mu$ und $V(X_i) = \sigma^2$ gilt:

- S_n ist approximativ normalverteilt
- $E(S_n) = n\mu$
- $V(S_n) = n\sigma^2$

Für das arithmetische Mittel $\bar{X}_n = \frac{S_n}{n}$ gilt:

- $\bar{X}_n$ ist approximativ normalverteilt
- $E(\bar{X}_n) = \mu$
- $V(\bar{X}_n) = \frac{\sigma^2}{n}$

Anwendung des Zentralen Grenzwertsatzes

1. **Prüfe Voraussetzungen**
   - Unabhängige Zufallsvariablen
   - Identische Verteilung
   - Endliche Varianz
   - Genügend große Stichprobe (n $\geq$ 30)
2. **Berechne Parameter**
   - $\mu_{S_n} = n\mu$
   - $\sigma_{S_n} = \sqrt{n}\sigma$
   - $\mu_{\bar{X}} = \mu$
   - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
3. **Standardisiere**
   - Transformiere zu $Z = \frac{X-\mu}{\sigma}$
   - Verwende Tabelle der Standardnormalverteilung

Approximationen

Approximation durch die Normalverteilung

- Binomialverteilung: $\mu = np, \sigma^2 = npq$
- Poissonverteilung: $\mu = \lambda, \sigma^2 = \lambda$

$$P(a \leq X \leq b) = \sum_{x=a}^{b} P(X = x) \approx \phi_{\mu,\sigma}(b + \frac{1}{2}) - \phi_{\mu,\sigma}(a - \frac{1}{2})$$

$P(a \leq X \leq b)$ = Wahrscheinlichkeit dass X zwischen a und b liegt
$\phi_{\mu,\sigma}$ = Verteilungsfunktion der Normalverteilung
a, b = Untere und obere Grenze

Approximationsregeln

Binomialverteilung $\to$ Normalverteilung:
- Bedingung: $npq > 9$
- Parameter: $\mu = np$, $\sigma^2 = npq$
- $B(n,p) \approx N(np, \sqrt{npq})$
- Stetigkeitskorrektur beachten!

Binomialverteilung $\to$ Poissonverteilung:
- Bedingung: $n \geq 50$ und $p \leq 0.1$
- $B(n,p) \approx Poi(np)$

Hypergeometrisch $\to$ Binomialverteilung:
- Bedingung: $n \leq \frac{N}{20}$
- $H(N,M,n) \approx B(n, \frac{M}{N})$

Faustregeln für Approximationen

- Die Approximation (Binomialverteilung) kann verwendet werden, wenn $npq > 9$
- Für grosses $n(n \geq 50)$ und kleines $p(p \leq 0.1)$ kann die Binomial- durch die Poisson-Verteilung approximiert werden:

$$B(n,p) \approx \text{Poi}(n \cdot p)$$

- Eine Hypergeometrische Verteilung kann durch eine Binomialverteilung angenähert werden, wenn $n \leq \frac{N}{20}$:

$$H(N,M,n) \approx B(n, \frac{M}{N})$$

$H(N,M,n)$ = Hypergeometrische Verteilung
$B(n,p)$ = Binomialverteilung
$\text{Poi}(\lambda)$ = Poissonverteilung mit Parameter $\lambda = n \cdot p$
N = Grundgesamtheit
M = Anzahl der Erfolge in der Grundgesamtheit
n = Stichprobengröße

Wahl der richtigen Verteilung

1. **Diskrete Verteilungen:**
   - Ziehen ohne Zurücklegen: Hypergeometrisch
   - Unabhängige Versuche: Binomial
   - Seltene Ereignisse: Poisson
2. **Approximationen prüfen:**
   - $npq > 9$: Normal-Approximation möglich
   - $n \geq 50, p \leq 0.1$: Poisson-Approximation möglich
   - $n \leq \frac{N}{20}$: Binomial-Approximation möglich
3. **Stetigkeitskorrektur:**
   - Bei Normal-Approximation: ± 0.5 an den Grenzen
   - $P(X \leq k) \approx P(X \leq k + 0.5)$
   - $P(X = k) \approx P(k - 0.5 \leq X \leq k + 0.5)$

Entscheidung über Approximationen

1. **Prüfe Stichprobenumfang**
   - Klein (n $<$ 30): Exakte Verteilung
   - Mittel (30 $\leq$ n $<$ 50): Je nach p
   - Groß (n $\geq$ 50): Approximation möglich
2. **Prüfe Wahrscheinlichkeit**
   - p $\leq$ 0.1: Poisson möglich
   - 0.1 $<$ p $<$ 0.9: Normal möglich
   - npq $>$ 9: Normal empfohlen
3. **Wähle Approximation**
   - Binomial $\to$ Normal: Große Stichproben, mittleres p
   - Binomial $\to$ Poisson: Große n, kleines p
   - Hypergeometrisch $\to$ Binomial: Kleine Stichprobe relativ zur Grundgesamtheit
4. **Beachte**
   - Stetigkeitskorrektur bei Normal
   - Rundungsregeln bei Grenzen
   - Vergleich mit exakter Lösung wenn möglich

Die Methode der kleinsten Quadrate

Einführung

Einführung

Die Methode der kleinsten Quadrate ist eine weit verbreitete Optimierungsmethode zur Modellierung mathematischer Zusammenhänge in großen Datenmengen. Das Ziel ist es, optimale Parameter zu finden, die den funktionalen Zusammenhang zwischen Messdaten am besten beschreiben. Bei der linearen Regression wird beispielsweise ein linearer Zusammenhang zwischen den Daten vermutet und versucht, eine optimale Gerade in die Datenmenge einzupassen.

Lineare Regression

Lineare Regression

Gegeben sind Datenpunkte $(x_i; y_i)$ mit $1 \leq i \leq n$, die näherungsweise auf einer Geraden liegen.
Die Residuen oder Fehler $\epsilon_i = y_i - g(x_i)$ dieser Datenpunkte sind die Abstände in y-Richtung zwischen y_i und der Geraden g.

Die"bestmögliche" Gerade, die Ausgleichs- oder Regressionsgerade, ist diejenige Gerade, für die die Summe der quadrierten Residuen $\sum_{i=1}^{n} \epsilon_i^2$ am kleinsten ist:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

mit:
y_i: beobachtete y-Werte
$\hat{y}_i$: prognostizierte bzw. erklärte y-Werte
ϵ_i: Residuum (oder auch Fehler/Abweichung) des i-ten Datenpunktes
$g(x_i)$ = Wert der Regressionsgerade an der Stelle x_i
n = Anzahl der Datenpunkte
(x_i, y_i) = Datenpunkte

Parameter der Regressionsgerade

Die Regressionsgerade $g(x) = mx + d$ mit den Parametern m und d ist die Gerade, für die die Residualvarianz $\tilde{s}_\epsilon^2$ minimal ist.

Parameter:

Steigung: $m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$

y-Achsenabschnitt: $d = \bar{y} - m\bar{x}$

Wichtige Kenngrößen:

Arithmetische Mittel: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Varianz der x_i-Werte:
$\tilde{s}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2$

Varianz der y_i-Werte:
$\tilde{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = (\frac{1}{n}\sum_{i=1}^{n} y_i^2) - \bar{y}^2$

Kovarianz:
$\tilde{s}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i y_i) - \bar{x}\bar{y}$

Residualvarianz:
$\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$

Lineare Regression berechnen

1. Berechne arithmetische Mittel $\bar{x}$ und $\bar{y}$
2. Berechne Kovarianzen und Varianzen:
   - $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
   - $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
   - $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
3. Berechne Steigung m und y-Achsenabschnitt d:
   - $m = \frac{s_{xy}}{s_x^2}$
   - $d = \bar{y} - m\bar{x}$
4. Regressionsgerade: $g(x) = mx + d$

Lineare Regression

Gegeben sind die Datenpunkte:

x_i	1	2	3	4	5
y_i	2.1	4.0	6.3	7.8	9.9

1. $\bar{x} = 3$, $\bar{y} = 6.02$
2. Kovarianzen und Varianzen:
   - $s_{xy} = 3.945$
   - $s_x^2 = 2$
   - $s_y^2 = 8.4916$
3. Parameter:
   - $m = \frac{3.945}{2} = 1.9725$
   - $d = 6.02 - 1.9725 \cdot 3 = 0.1025$
4. Regressionsgerade: $g(x) = 1.9725x + 0.1025$

Varianzzerlegung und Bestimmtheitsmass

Varianzzerlegung

Die Totale Varianz setzt sich zusammen aus der Residualvarianz und der Varianz der prognostizierten Werte:

$$\tilde{s}_y^2 = \tilde{s}_\epsilon^2 + \tilde{s}_{\hat{y}}^2$$

mit:
$\tilde{s}_y^2$: Totale Varianz
$\tilde{s}_{\hat{y}}^2$: prognostizierte (erklärte) Varianz
$\tilde{s}_\epsilon^2$: Residualvarianz

Bestimmtheitsmass

Das Bestimmtheitsmass R^2 beurteilt die globale Anpassungsgüte einer Regression über den Anteil der prognostizierten Varianz $s_{\hat{y}}^2$ an der totalen Varianz s_y^2:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

R^2 = Bestimmtheitsmass (zwischen 0 und 1)
$s_{\hat{y}}^2$ = Varianz der prognostizierten Werte
s_y^2 = Totale Varianz

Das Bestimmtheitsmass R^2 entspricht dem Quadrat des Korrelationskoeffizienten (nach Bravais-Pearson):

$$R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = (r_{xy})^2$$

s_{xy} = Kovarianz von x und y
s_x^2 = Varianz der x-Werte
s_y^2 = Varianz der y-Werte
r_{xy} = Korrelationskoeffizient

Interpretation des Bestimmtheitsmasses

- $R^2 = 0.75$ bedeutet, dass 75% der gesamten Varianz durch die Regression erklärt sind
- Die restlichen 25% sind Zufallsstreuung

Bestimmtheitsmass berechnen

1. Berechne die totale Varianz s_y^2 2. Berechne die Residualvarianz s_ϵ^2 3. Berechne die erklärte Varianz $s_{\hat{y}}^2$ 4. Berechne das Bestimmtheitsmass:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_\epsilon^2}{s_y^2}$$

5. Interpretation:
   - $R^2 \approx 1$: Sehr gute Anpassung
   - $R^2 \approx 0$: Schlechte Anpassung

Residuenbetrachtung

Residuenplot

Die Residuen werden bezogen auf die prognostizierten y-Werte $\hat{y}$ dargestellt. Auf der horizontalen Achse werden die prognostizierten y-Werte $\hat{y}$ und auf der vertikalen Achse die Residuen angetragen.

Beurteilungskriterien:
- Residuen sollten unsystematisch (d.h. zufällig) streuen
- Überall etwa gleich um die horizontale Achse streuen
- Betragsmäßig kleine Residuen sollten häufiger sein als große

Residuen und Residuenplot analysieren

1. Berechne die Residuen für jeden Datenpunkt:
   - $\epsilon_i = y_i - (mx_i + d)$
2. Erstelle Residuenplot:
   - x-Achse: Prognostizierte Werte $\hat{y}_i = mx_i + d$
   - y-Achse: Residuen ϵ_i
3. Prüfe Eigenschaften:
   - Residuen sollten zufällig um Null streuen
   - Keine systematischen Muster erkennbar
   - Gleiche Streubreite über alle $\hat{y}_i$

Nichtlineares Verhalten

Linearisierung

Wichtige Transformationen:

Oft können nichtlineare Regressionsmodelle durch geeignete Transformation auf ein lineares Modell zurückgeführt werden.

Ausgangsfunktion	Transformation
$y = q \cdot x^m$	$\log(y) = \log(q) + m \cdot \log(x)$
$y = q \cdot m^x$	$\log(y) = \log(q) + \log(m) \cdot x$
$y = q \cdot e^{m \cdot x}$	$\ln(y) = \ln(q) + m \cdot x$
$y = \frac{1}{q + m \cdot x}$	$V = q + m \cdot x; V = \frac{1}{y}$
$y = q + m \cdot \ln(x)$	$y = q + m \cdot U; U = \ln(x)$
$y = \frac{1}{q \cdot m^x}$	$\log(\frac{1}{y}) = \log(q) + \log(m) \cdot x$

y = Abhängige Variable
x = Unabhängige Variable
q, m = Parameter der Funktion

Nichtlineare Regression durch Linearisierung

1. Bestimme passende Transformation aus Tabelle
2. Führe Transformation durch
3. Wende lineare Regression auf transformierte Daten an
4. Transformiere Parameter zurück

Exponentielles Wachstum

$y = q \cdot e^{mx}$ mit gegebenen Messwerten:

x	1	2	3	4
y	2.1	4.2	8.1	15.9

1. Transformation $\ln(y) = \ln(q) + mx \rightarrow Y = \ln(y)$, $b = \ln(q)$:

x	1	2	3	4
Y	0.742	1.435	2.092	2.766

2. Lineare Regression: $Y = mx + b \; rightarrow \; Y = 0.674x + 0.071$
3. Rücktransformation: $q = e^b$
   - $m = 0.674$
   - $q = e^{0.071} = 1.074$
4. Ergebnis: $y = 1.074 \cdot e^{0.674x}$

Allgemeines Vorgehen bei der Regression

Matrix-Darstellung

Für die Methode der kleinsten Quadrate mit mehreren Variablen wird ein lineares Gleichungssystem aufgestellt:
$y = Xp + \epsilon$
mit:
p: Vektor der Parameter
y: Vektor der Messwerte
ϵ: Vektor der Residuen
X: Matrix der Eingangswerte

Die Lösung ist:
$p = (X^T X)^{-1} X^T y$
falls $(X^T X)$ invertierbar

Matrix-Darstellung

Die Parameter m und q der Regressionsgeraden werden mit der Matrix A berechnet:

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \quad A^T \cdot A \cdot \begin{pmatrix} m \\ q \end{pmatrix} = A^T \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Matrix-Methode für lineare Regression

1. Erstelle Design-Matrix A:

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

2. Berechne $A^T \cdot A$ 3. Berechne $(A^T \cdot A)^{-1}$ 4. Berechne Parameter:

$$\begin{pmatrix} m \\ q \end{pmatrix} = (A^T \cdot A)^{-1} \cdot A^T \cdot \vec{y}$$

Residuenberechnung

Die Residuen ϵ_i ergeben sich als:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (mx_i + q)$$

Die Summe der quadrierten Residuen wird minimiert:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (mx_i + q))^2 \to \min$$

Vorgehen bei Mehrfachregression

1. Aufstellen der Matrix X mit den Eingangswerten 2. Berechnung der Parameter $p = (X^T X)^{-1} X^T y$ 3. Berechnung der Residuen $\epsilon = y - Xp$
4. Überprüfung der Modellgüte durch:

- Bestimmtheitsmass R^2
- Residuenanalyse
- Plausibilität der Parameter

Mehrfachregression

1. Aufstellen der Designmatrix:

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1(k-1)} & 1 \\ x_{21} & x_{22} & \cdots & x_{2(k-1)} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n(k-1)} & 1 \end{pmatrix}$$

2. Berechnung der Parameter:

$$\vec{p} = (A^T A)^{-1} A^T \vec{y}$$

3. Residuen berechnen:

$$\vec{\epsilon} = \vec{y} - A\vec{p}$$

4. Bestimmtheitsmass:

$$R^2 = 1 - \frac{\sum \epsilon_i^2}{\sum (y_i - \bar{y})^2}$$

Mehrfachregression

Ein Gebrauchtwagenhändler möchte den Preis (P) seiner Autos basierend auf Alter (A) und Kilometerstand (K) berechnen. Gegeben sind folgende Daten:

Auto	Alter (Jahre)	km (10000)	Preis (1000 CHF)
1	2	3	25
2	3	4	20
3	4	6	15
4	5	7	12

1. Designmatrix aufstellen:

$$A = \begin{pmatrix} 2 & 3 & 1 \\ 3 & 4 & 1 \\ 4 & 6 & 1 \\ 5 & 7 & 1 \end{pmatrix}$$

2. Parameter berechnen:

$$\vec{p} = \begin{pmatrix} -3 \\ -1.5 \\ 35 \end{pmatrix}$$

3. Resultierende Funktion:

$$P = -3A - 1.5K + 35$$

Polynomiale Regression

Für Regression mit Polynomen höheren Grades:
1. Erweitere Designmatrix:

$$A = \begin{pmatrix} x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ x_2^n & x_2^{n-1} & \cdots & x_2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_m^n & x_m^{n-1} & \cdots & x_m & 1 \end{pmatrix}$$

2. Löse wie bei linearer Regression:

$$\vec{p} = (A^T A)^{-1} A^T \vec{y}$$

3. Polynom aufstellen:

$$y = p_1 x^n + p_2 x^{n-1} + \ldots + p_n x + p_{n+1}$$

Quadratische Regression

Gegeben sind Messwerte:

x	0	1	2	3	4
y	1	2.1	5.2	10.1	17.2

1. Designmatrix für quadratisches Polynom:

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{pmatrix}$$

2. Parameter berechnen:

$$\vec{p} = \begin{pmatrix} 1 \\ 0.1 \\ 1 \end{pmatrix}$$

3. Quadratische Funktion:

$$y = x^2 + 0.1x + 1$$

Gütekriterien für Regression

1. Bestimmtheitsmass R^2:
   - $R^2 > 0.9$: Sehr gute Anpassung
   - $0.7 < R^2 < 0.9$: Gute Anpassung
   - $0.5 < R^2 < 0.7$: Mittelmässige Anpassung
   - $R^2 < 0.5$: Schlechte Anpassung
2. Residuenanalyse:
   - Residuen sollten zufällig um 0 schwanken
   - Keine systematischen Muster erkennbar
   - Residuen sollten normalverteilt sein
3. Prognosegüte:
   - Mittlerer quadratischer Fehler (MSE)
   - Wurzel des mittleren quadratischen Fehlers (RMSE)
   - Mittlerer absoluter Fehler (MAE)

Modellwahl durch Residuenanalyse

Für einen Datensatz wurden drei Modelle getestet:
- Linear: $y = 2x + 1$
- Quadratisch: $y = x^2 + x + 1$
- Exponentiell: $y = 2e^{0.5x}$

Bestimmtheitsmasse:
- Linear: $R^2 = 0.85$
- Quadratisch: $R^2 = 0.98$
- Exponentiell: $R^2 = 0.92$

Residuenanalyse zeigt:
- Linear: Systematische Krümmung in Residuen
- Quadratisch: Zufällige Verteilung der Residuen
- Exponentiell: Leichte Systematik in Residuen

Schlussfolgerung: Das quadratische Modell ist am besten geeignet.

Prüfungsaufgaben lösen

1. Aufgabentyp identifizieren:
   - Einfache lineare Regression
   - Mehrfachregression
   - Nichtlineare Regression mit Transformation
   - Polynomiale Regression
2. Vorgehen wählen:
   - Linear: Direkte Berechnung mit Formeln
   - Nichtlinear: Transformation und lineare Regression
   - Polynomial: Erweiterte Designmatrix
   - Mehrfach: Matrix-Methode
3. Berechnungen durchführen:
   - Parameter bestimmen
   - Bestimmtheitsmass berechnen
   - Residuen analysieren
4. Ergebnisse interpretieren:
   - Modellgüte bewerten
   - Residuen beurteilen
   - Prognosen erstellen

Klausuraufgabe - Linearisierung Gegeben sind Messwerte für ein exponentielles Wachstum:

t (h)	0	1	2	3
N	100	150	225	340

Finden Sie eine Funktion der Form $N(t) = N_0 e^{kt}$

1. Transformation:

$$\ln(N) = \ln(N_0) + kt$$

2. Neue Wertetabelle:

t	0	1	2	3
$\ln(N)$	4.61	5.01	5.42	5.83

3. Lineare Regression:

$$\ln(N) = 0.405t + 4.61$$

4. Rücktransformation:

$$N(t) = 100.4e^{0.405t}$$

5. Bestimmtheitsmass: $R^2 = 0.999$

Schliessende Statistik - Parameter- / Intervallschätzung

Zufallsstichproben

Grundlagen der Zufallsstichproben

Die Grundgesamtheit ist eine Menge von gleichartigen Objekten oder Elementen. Sie kann endlich oder unendlich viele Objekte enthalten.
Eine Stichprobe vom Umfang n wird entnommen, um Informationen über die Grundgesamtheit zu gewinnen. Dies ist oft notwendig, da der Zeit- und Kostenaufwand für eine Vollerhebung zu hoch ist

Einfache Zufallsstichprobe

Eine einfache Zufallsstichprobe vom Umfang n ist eine Folge von Zufallsvariablen $X_1, X_2, \ldots, X_n$ (Stichprobenvariablen). Dabei bezeichnet X_i die Merkmalsausprägung des i-ten Elements in der Stichprobe.
Die beobachteten Merkmalswerte $x_1, x_2, \ldots, x_n$ der n Elemente sind Realisierungen der Zufallsvariablen und heißen Stichprobenwerte.
Wichtige Eigenschaften:
- Alle Stichprobenvariablen sind stochastisch unabhängig
- Alle X_i folgen derselben Verteilung $F(x)$ der Grundgesamtheit

Parameterschätzungen

Schätzfunktionen

Schätzfunktion

Eine Schätzfunktion $\Theta = g(X_1, X_2, \ldots, X_n)$ ist eine spezielle Stichprobenfunktion zur Schätzung eines Parameters θ der Grundgesamtheit.
Der Schätzwert $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ ergibt sich durch Einsetzen der konkreten Stichprobenwerte.
θ ist der wahre, unbekannte Parameterwert der Grundgesamtheit.

Kriterien für eine optimale Schätzfunktion

Optimale Schätzfunktionen

Eine Schätzfunktion sollte folgende Eigenschaften haben:
- Erwartungstreu: $E(\Theta) = \theta$
- Effizient: Kleinste Varianz unter allen Schätzern $V(\Theta_1) < V(\Theta_2)$
- Konsistent: $E(\Theta) \to \theta$ und $V(\Theta) \to 0$ für $n \to \infty$

Interpretation:
- Erwartungstreue: im Mittel wird der richtige Wert geschätzt
- Effizienz: möglichst geringe Streuung der Schätzung
- Konsistenz: Schätzung wird mit wachsender Stichprobe genauer

Beispiel Erwartungstreue einer Schätzfunktion

Grundgesamtheit mit Erwartungswert μ, Varianz σ^2 und Zufallsstichprobe X_1, X_2, X_3. Die folgende Schätzfunktion ist gegeben:

$$\Theta_1 = \frac{1}{3} \cdot (2X_1 + X_2)$$

Ist diese Schätzfunktion erwartungstreu (wahrer Parameter: μ)?

$$E(\Theta_1) = E(\frac{1}{3} \cdot (2X_1 + X_2)) = \frac{1}{3} \cdot (2E(X_1) + E(X_2))$$
$$E(\Theta_1) = \frac{1}{3} \cdot (2\mu + \mu) = \frac{3\mu}{3} = \mu$$

$E(X_1), E(X_2) =$ Erwartungswerte der einzelnen Zufallsvariablen

Da $E(\Theta_1) = \mu$ ist die Funktion erwartungstreu.

Effizienz einer Schätzfunktion

Grundgesamtheit mit Erwartungswert μ, Varianz σ^2 und Zufallsstichprobe X_1, X_2, X_3. Gegeben ist die Schätzfunktion:

$$\Theta_1 = \frac{1}{3} \cdot (2X_1 + X_2)$$

Berechnung der Effizienz:

$$\begin{aligned} V(\Theta_1) &= V(\frac{1}{3} \cdot (2X_1 + X_2)) \\ &= \frac{1}{9} \cdot V(2X_1 + X_2) \\ &= \frac{1}{9} \cdot (V(2X_1) + V(X_2)) \\ &= \frac{1}{9} \cdot (4V(X_1) + V(X_2)) \\ &= \frac{1}{9} \cdot (4\sigma^2 + \sigma^2) \\ &= \frac{5\sigma^2}{9} \end{aligned}$$

$V(\Theta_1) =$ Varianz der Schätzfunktion
$V(X_1), V(X_2) =$ Varianzen der einzelnen Zufallsvariablen
$\sigma^2 =$ Varianz der Grundgesamtheit

Die Effizienz der Schätzfunktion ist also $\frac{5\sigma^2}{9}$.

Wichtige Schätzfunktionen

Schätzfunktionen für wichtige Parameter

Erwartungswert:

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

Eigenschaften:
- Erwartungstreu: $E(\bar{X}) = \mu$
- Konsistent: $V(\bar{X}) = \frac{\sigma^2}{n} \to 0$ für $n \to \infty$

Varianz:

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

Eigenschaften:
- Erwartungstreu: $E(S^2) = \sigma^2$
- Konsistent: $V(S^2) \to 0$ für $n \to \infty$

Anteilswert: (bei Bernoulli-Verteilung)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Vertrauensintervalle

Vertrauensintervall

Ein Vertrauensintervall $[\Theta_u, \Theta_o]$ zum Niveau γ ist ein zufälliges Intervall mit:

$$P(\Theta_u \leq \theta \leq \Theta_o) = \gamma$$

γ: Vertrauensniveau (statistische Sicherheit)
$\alpha = 1 - \gamma$: Irrtumswahrscheinlichkeit
Θ_u, Θ_o: Unter- und Obergrenze

Vertrauensintervall-Typen

Fall	Verteilung	Test-Statistik	Grenzen
μ (σ^2 bekannt)	Standard-normalvert.	$U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$\bar{x} \pm c\frac{\sigma}{\sqrt{n}}$ $c = u_p$
μ (σ^2 unbek. *)	t-Verteilung mit $f = n-1$	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$\bar{x} \pm c\frac{s}{\sqrt{n}}$ $c = t_{p,f}$
σ^2	χ^2-Verteilung mit $f = n-1$	$Z = \frac{(n-1)S^2}{\sigma^2}$	$[\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1}]$ $c_1 = \chi^2_{p_1,f}, c_2 = \chi^2_{p_2,f}$

* $n \leq 30$; sonst Fall 1 mit s^2 (Varianz von Stichprobe) als Schätzwert für σ
mit $p = \frac{1+\gamma}{2}$, $p_1 = \frac{1-\gamma}{2}$, $p_2 = \frac{1+\gamma}{2}$

Vertrauensintervalle berechnen

1. Verteilungstyp bestimmen:
   - Parameter (μ oder σ^2)
   - σ^2 bekannt oder unbekannt
2. Quantile bestimmen:
   - γ und α beachten
   - Richtige Tabelle wählen
   - Freiheitsgrade $f = n - 1$ beachten
3. Intervallgrenzen berechnen:
   - Standardfehler berechnen
   - Grenzen Θ_u und Θ_o bestimmen

Beispiel: Berechnung eines Vertrauensintervalls (t-Verteilung)

Geben Sie das Vertrauensintervall für μ an (σ^2 unbekannt). Gegeben sind:

$$n = 10, \quad \bar{x} = 102, \quad s^2 = 16, \quad \gamma = 0.99$$

1. Verteilungstyp mit Param μ und σ^2 unbekannt $\rightarrow$ T-Verteilung
2. $f = n - 1 = 9$, $p = \frac{1+\gamma}{2} = 0.995$, $c = t_{(p;f)} = t_{(0.995;9)} = 3.25$
3. $e = c \cdot \frac{S}{\sqrt{n}} = 4.111$, $\Theta_u = \bar{X} - e = 97.89$, $\Theta_o = \bar{X} + e = 106.11$

Bestimmung des Stichprobenumfangs

1. Gegebene Verteilung und Parameter:
   - Normalverteilung mit σ^2 bekannt
   - Vertrauensniveau γ
   - Maximal zulässige Intervallbreite d
2. Kritischen Wert bestimmen:
   - $p = \frac{1+\gamma}{2}$
   - $c = u_p$ für Normalverteilung
3. Stichprobenumfang berechnen:
   - $n \geq (\frac{2c\sigma}{d})^2$
   - Auf nächste ganze Zahl aufrunden
4. Bei unbekannter Varianz:
   - Vorerhebung durchführen
   - Varianz schätzen
   - t-Verteilung statt Normalverteilung

Beispiel: Stichprobenumfang bestimmen

Ein Prozess produziert Teile mit bekannter Standardabweichung $\sigma = 0.5$ mm. Der Mittelwert soll mit einer Genauigkeit von ± 0.2 mm bei einem Vertrauensniveau von 99% geschätzt werden.

1. Gesucht:
   - Intervallbreite $d = 0.4$ mm
   - $\gamma = 0.99$
2. Kritischer Wert:
   - $p = \frac{1+0.99}{2} = 0.995$
   - $c = u_{0.995} = 2.576$
3. Stichprobenumfang:
   - $n \geq (\frac{2 \cdot 2.576 \cdot 0.5}{0.4})^2 = 41.47$
   - $n = 42$ (aufgerundet)