

## Deskriptive Statistik

### Teilbereiche der Statistik

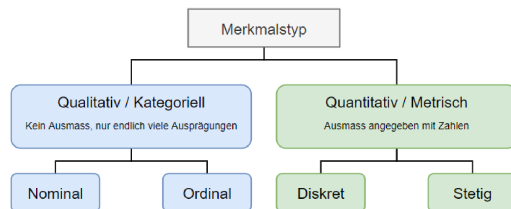
- **Deskriptive Statistik:** Beschreibung und übersichtliche Darstellung von Daten, Ermittlung von Kenngrößen und Datenvalidierung
- **Explorative Statistik:** Weiterführung und Verfeinerung der beschreibenden Statistik, insbesondere die Suche nach Strukturen und Besonderheiten
- **Induktive Statistik:** 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

### Merkmaltypen

- **Qualitativ/Kategorial:** 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



### Merkmaltypen

- **Würfelfwurf (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

- $\Omega$  = 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  $a_i$  ( $i = 1, \dots, m$ )

#### 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

**Würfelfwurf** 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**

- Bestimme für jede Klasse:  $d_i, h_i, f_i$
- Bestimme kumulative Häufigkeiten  $H_i$
- CDF Berechnung:
  - Bestimme kumulative Häufigkeiten  $H_i$
  - Teile durch Stichprobengrösse:  $F(x) = \frac{H(x)}{n}$
- 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	

**Kenngrossen**

**Arten von Kenngrossen**

- 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  $x_1, x_2, \dots, x_n$ , 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**  $Q = x_i = x_{[n \cdot q]}$

Position des Quantils:  $i = [n \cdot q]$   
 $n$ : Anzahl der Beobachtungen  
 $q$ : Quantilswert (zB. 0.25 für  $Q_1$ )  
 $x_i$ : 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  $x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[n]}$ :
- Berechne  $n \cdot q$
  - Falls  $n \cdot q$  eine ganze Zahl ist:  $R_q = \frac{1}{2}(x_{n \cdot q} + x_{n \cdot q + 1})$
  - Falls  $n \cdot q$  keine ganze Zahl ist:  $R_q = x_{[n \cdot q]}$
  - Wobei  $[n \cdot q]$  die nächstgrössere ganze Zahl ist

**Berechnung von Lageparametern**

- Sortiere die Daten aufsteigend
- Berechne den Mittelwert: Summe aller Werte / Anzahl Werte
- Bestimme den Median:
  - Bei ungerader Anzahl: mittlerer Wert
  - Bei gerader Anzahl: Mittelwert der beiden mittleren Werte
- Finde den Modus (häufigster Wert)
- Berechne die Quartile:
  - $Q_1$ : 25%-Quantil,  $Q_2$ : Median,  $Q_3$ : 75%-Quantil

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

**Berechnung  $Q_1$  (25%-Quantil):**  $Q_1 = x_2 = 4$

$i = [8 \cdot 0.25] = [2] = 2$

**Berechnung  $Q_2$  (Median):**  $Q_2 = (5 + 7)/2 = 6$

$n$  gerade  $\rightarrow$  Mittelwert von Position 4 und 5

**Berechnung  $Q_3$  (75%-Quantil):**  $Q_3 = x_6 = 8$

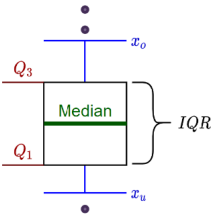
$i = [8 \cdot 0.75] = [6] = 6$

**Interquartilsabstand:**  $IQR = Q_3 - Q_1 = 8 - 4 = 4$

**Boxplot**

**Boxplot** besteht aus:

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



**Erstellen eines Boxplots**

- Berechne die Quartile  $Q_1, Q_2$  (Median) und  $Q_3$
- Bestimme den Interquartilsabstand  $IQR = Q_3 - Q_1$
- Berechne die Grenzen für Ausreisser:
  - Untere Grenze:  $Q_1 - 1.5 \cdot IQR$  und Obere Grenze:  $Q_3 + 1.5 \cdot IQR$
- Zeichne Box mit:
  - Unterer Rand bei  $Q_1$ , Mittellinie bei  $Q_2$ , Oberer Rand bei  $Q_3$
- Zeichne Antennen bis zum:
  - Kleinsten Wert  $\geq$  untere Grenze
  - Grössten Wert  $\leq$  obere Grenze
- Markiere alle Werte ausserhalb als Ausreisser

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

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

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

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

- $s$ : Stichprobenstandardabweichung
- $s_{kor}$ : Korrigierte Stichprobenstandardabweichung
- $s^2$ : Stichprobenvarianz
- $s_{kor}^2$ : 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:

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

Standardabweichung:

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

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:  $s_{kor}^2 = \frac{9+1+1+1+16}{5-1} = \frac{28}{4} = 7$

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

Alternative Berechnung:

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

Verteilungsformen

- **Symmetrisch:** Rechte und linke Hälfte spiegelbildlich
- **Linkssteil (rechtsschief):**
  - Daten links konzentriert
  - $x_{mod} < x_{med} < \bar{x}$
- **Rechtssteil (linksschief):**
  - Daten rechts konzentriert
  - $x_{mod} > x_{med} > \bar{x}$
- **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:** Kontingenztafel + Mosaikplot
- **Ein kategorielles + ein metrisches Merkmal:** Boxplot oder Strip-chart
  - Kennwerte pro Kategorie
- **Zwei metrische Merkmale:** Streudiagramm (Scatterplot)
  - Punktwolke in der (x,y)-Ebene

Kontingenztafel 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 Kontingenztafeln
  - Farbliche Codierung zusätzlicher Dimensionen
- **Metrische Merkmale:**
  - Matrix von Streudiagrammen
  - Korrelationsmatrix

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:

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

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

**Korrelationskoeffizient nach Pearson** normiert die Kovarianz:

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

- 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:

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

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 wäre de Prüfig)

**Kombinatorik**

**Grundbegriffe**

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

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

**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  $0 \leq k \leq n$ :

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

**Eigenschaften** Für den Binomialkoeffizienten gelten:

**Leere Menge:**  $\binom{n}{0} = \binom{n}{n} = 1$       **Summe:**  $\sum_{k=0}^n \binom{n}{k} = 2^n$

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

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

**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
<b>Variation</b> (Reihenfolge wichtig)	$n^k$	$\frac{n!}{(n-k)!}$
<b>Kombination</b> (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?  $\rightarrow$  Variation
  - Ist nur die Auswahl wichtig?  $\rightarrow$  Kombination
3. **Prüfe Wiederholungen:**
  - Dürfen Objekte mehrfach vorkommen?  $\rightarrow$  Mit Wiederholung
  - Darf jedes Objekt nur einmal?  $\rightarrow$  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

**Ergebnisraum und Zähl-dichte** Ein Zufallsexperiment ist ein Vorgang, bei dem folgende Bedingungen erfüllt sind:

- 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

Der **Ergebnisraum**  $\Omega$  ist die Menge aller möglichen Ergebnisse des Zufallsexperiments.

Die **Zähl-dichte**  $\rho : \Omega \rightarrow [0,1]$  ordnet jedem Ergebnis  $\omega \in \Omega$  seine Wahrscheinlichkeit zu, wobei  $\sum_{\omega \in \Omega} \rho(\omega) = 1$  gilt.

**Ereignisse und Wahrscheinlichkeitsraum** Ein **Ereignis** ist eine Teilmenge des Ergebnisraums  $\Omega$ . Der **Ereignisraum**  $2^\Omega$  ist die Menge aller möglichen Ereignisse (Potenzmenge von  $\Omega$ ).

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

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

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

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

**Eigenschaften von Wahrscheinlichkeitsräumen** Für einen diskreten Wahrscheinlichkeitsraum  $(\Omega, P)$  gelten:

- (A1) Unmögliches Ereignis:  $P(\emptyset) = 0$
- (A2) Sicheres Ereignis:  $P(\Omega) = 1$
- (A3) Komplementäres Ereignis:  $P(\Omega \setminus A) = 1 - P(A)$
- (A4) Vereinigung:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (A5) Sigma-Additivität: Für paarweise disjunkte Ereignisse gilt:  $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$

Zufallsvariablen

**Zufallsvariablen** Eine **Zufallsvariable**  $X$  ist eine Funktion  $X : \Omega \rightarrow \mathbb{R}$ , 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  $F(x) = \sum_{t \leq x} f(t)$
- $\lim_{x \rightarrow \infty} F(x) = 1$  und  $\lim_{x \rightarrow -\infty} F(x) = 0$
- 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  $X$  sind definiert:

**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 Erwartungswert und Varianz**

- Linearität:  $E(aX + b) = aE(X) + b$
- Verschiebungssatz:  $V(X) = E(X^2) - (E(X))^2$
- Lineare Transformation:  $V(aX + b) = a^2 V(X)$

Bedingte Wahrscheinlichkeit

**Bedingte Wahrscheinlichkeit** Die **bedingte Wahrscheinlichkeit** von  $B$  unter der Bedingung  $A$  ist:

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

**Multiplikationssatz**

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

**Satz von der totalen Wahrscheinlichkeit**

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

**Satz von Bayes**

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

**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

Stochastische Unabhängigkeit

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

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

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

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

**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  $P(B) > 0$

Für unabhängige Zufallsvariablen  $X$  und  $Y$  gilt:

- $E(X \cdot Y) = E(X) \cdot E(Y)$
- $V(X + Y) = V(X) + V(Y)$

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 ZV	Stetige ZV
Dichtefunktion	$f(x) = P(X = x)$	$f(x) = f'(x)$
Verteilungsfunktion	$F(x) = \sum_{x \leq X} f(x)$	$F(x) = \int_{-\infty}^x f(t)dt$
Wahrscheinlichkeiten	$P(a \leq X \leq b) = \sum_{a \leq x \leq b} f(x)$	$P(a \leq X \leq b) = \int_a^b f(t)dt$
Erwartungswert	$E(X) = \sum_{x \in \mathbb{R}} x \cdot f(x)$	$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$
Varianz	$V(X) = \sum_{x \in \mathbb{R}} (x - E(X))^2 \cdot f(x)$	$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)dx$

Diskrete Verteilungen

**Übersicht der diskreten Verteilungen**

Verteilung	Notation	$E(X)$	$V(X)$
Hypergeometrisch	$H(N, M, n)$	$n \cdot \frac{M}{N}$	$n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$
Binomial	$B(n, p)$	$n \cdot p$	$n \cdot p \cdot q$
Poisson	$Poi(\lambda)$	$\lambda$	$\lambda$

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

**Parameter:**

- $N$ : Grundgesamtheit
- $M$ : Anzahl Merkmalsträger
- $n$ : Stichprobenumfang

**Wahrscheinlichkeitsfunktion:**

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

**Kenngößen:**

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

**Notation:**  $X \sim H(N, M, n)$

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

**Parameter:**

- $n$ : Anzahl Versuche
- $p$ : Erfolgswahrscheinlichkeit
- $q = 1 - p$ : Gegenwahrscheinlichkeit

**Wahrscheinlichkeitsfunktion:**

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

**Kenngößen:**

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

**Notation:**  $X \sim B(n, p)$

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

**Parameter:**

- $\lambda$ : Erwartungswert/Rate pro Intervall

**Wahrscheinlichkeitsfunktion:**

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

**Kenngößen:**

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

**Notation:**  $X \sim Poi(\lambda)$

## Stetige Verteilungen

**Normalverteilung** Die Dichtefunktion der Normalverteilung ist:

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

**Parameter:**

- $\mu$ : Erwartungswert (Lage)
- $\sigma$ : Standardabweichung (Streuung)

**Eigenschaften:**

- Symmetrisch um  $\mu$
- 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  $[\mu - 3\sigma, \mu + 3\sigma]$

**Notation:**  $X \sim N(\mu, \sigma)$

## Zentraler Grenzwertsatz und Approximationen

**Zentraler Grenzwertsatz** Für die Summe  $S_n = X_1 + \dots + X_n$  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
- $E(\bar{X}_n) = \mu$
- $V(\bar{X}_n) = \frac{\sigma^2}{n}$

**Approximationsregeln** Binomialverteilung  $\rightarrow$  Normalverteilung:

- Bedingung:  $npq > 9$
- $B(n, p) \approx N(np, \sqrt{npq})$
- Stetigkeitskorrektur beachten!

**Binomialverteilung  $\rightarrow$  Poissonverteilung:**

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

**Hypergeometrisch  $\rightarrow$  Binomialverteilung:**

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

**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:  $\pm 0.5$  an den Grenzen
- $P(X \leq k) \approx P(X \leq k + 0.5)$
- $P(X = k) \approx P(k - 0.5 \leq X \leq k + 0.5)$

## 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

$\epsilon_i$ : Residuen (oder auch Fehler)

#### 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 Kenngößen:**

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:

$$\tilde{s}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left( \frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

Varianz der  $y_i$ -Werte:

$$\tilde{s}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \left( \frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2$$

Kovarianz:

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

Residualvarianz:

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

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  $\tilde{s}_{\hat{y}}^2$  an der totalen Varianz  $\tilde{s}_y^2$ :

$$R^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2} = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2 \tilde{s}_y^2} = r_{xy}^2$$

Das Bestimmtheitsmass  $R^2$  stimmt überein mit dem Quadrat des Korrelationskoeffizienten (nach Bravais-Pearson).  
Interpretation:  
-  $R^2 = 0.75$  bedeutet, dass 75% der gesamten Varianz durch die Regression erklärt sind  
- Die restlichen 25% sind Zufallsstreuung

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

Nichtlineares Verhalten

Linearisierung

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

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

Allgemeines Vorgehen

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  
 $\epsilon$ : Vektor der Residuen  
 $X$ : Matrix der Eingangswerte

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

Vorgehen bei Mehrfachregression

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

Schmessende Statistik – Parameter- und 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 weil:  
• Der Zeit- und Kostenaufwand für eine Vollerhebung zu hoch ist  
• Die Anzahl Objekte zu groß ist  
• Die Objekte bei der Untersuchung zerstört werden

Einfache Zufallsstichprobe

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

Parameterschätzungen

Schätzfunktionen

Schätzfunktion

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

Kriterien für eine optimale Schätzfunktion

Optimale Schätzfunktionen

Eine Schätzfunktion sollte folgende Eigenschaften haben:  
1. Erwartungstreu:  $E(\Theta) = \theta$   
2. Effizient: Kleinste Varianz unter allen erwartungstreuen Schätzern  
3. Konsistent:  $E(\Theta) \rightarrow \theta$  und  $V(\Theta) \rightarrow 0$  für  $n \rightarrow \infty$   
Interpretation:  
• Erwartungstreue bedeutet, dass im Mittel der richtige Wert geschätzt wird  
• Effizienz bedeutet möglichst geringe Streuung der Schätzung  
• Konsistenz bedeutet, dass die Schätzung mit wachsendem Stichprobenumfang immer genauer wird

Wichtige Schätzfunktionen

Schätzfunktionen für wichtige Parameter

Erwartungswert:

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

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

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

Eigenschaften:  
• Erwartungstreu:  $E(S^2) = \sigma^2$   
• Konsistent:  $V(S^2) \rightarrow 0$  für  $n \rightarrow \infty$   
Anteilswert: (bei Bernoulli-Verteilung)

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

Maximum-Likelihood-Schätzung

Likelihood-Funktion

Für eine Stichprobe vom Umfang  $n$  mit den Werten  $x_1, x_2, \dots, x_n$  ist die Likelihood-Funktion definiert als:  
$$L(\theta) = f_X(x_1|\theta) \cdot f_X(x_2|\theta) \cdot \dots \cdot f_X(x_n|\theta)$$
  
wobei  $f_X(x|\theta)$  die Wahrscheinlichkeitsdichte der Verteilung ist.

Maximum-Likelihood-Schätzung

1. Likelihood-Funktion  $L(\theta)$  aufstellen 2. Log-Likelihood  $\ln(L(\theta))$  bilden (vereinfacht die Rechnung) 3. Ableitung  $\frac{d}{d\theta} \ln(L(\theta)) = 0$  setzen 4. Nach  $\theta$  auflösen für  $\hat{\theta}_{ML}$  5. Maximum überprüfen durch zweite Ableitung
- Beispiel für Normalverteilung:**

1.  $L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$
2.  $\ln(L) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$
3.  $\frac{\partial}{\partial \mu} \ln(L) = 0$  und  $\frac{\partial}{\partial \sigma^2} \ln(L) = 0$
4. Ergibt:  $\hat{\mu}_{ML} = \bar{x}$  und  $\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$

Vertrauensintervalle

Vertrauensintervall

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

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

$\gamma$ : Vertrauensniveau (statistische Sicherheit)  
 $\alpha = 1 - \gamma$ : Irrtumswahrscheinlichkeit  
 $\Theta_u, \Theta_o$ : Unter- und Obergrenze

Vertrauensintervall-Typen

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

mit  $p = \frac{1+\gamma}{2}, p_1 = \frac{1-\gamma}{2}, p_2 = \frac{1+\gamma}{2}$

Vertrauensintervalle berechnen

1. Verteilungstyp bestimmen:
- Parameter ( $\mu$  oder  $\sigma^2$ )
  - $\sigma^2$  bekannt oder unbekannt
2. Quantile bestimmen:
- $\gamma$  und  $\alpha$  beachten
  - Richtige Tabelle wählen
  - Freiheitsgrade  $f = n - 1$  beachten
3. Intervallgrenzen berechnen:
- Standardfehler berechnen
  - Grenzen  $\Theta_u$  und  $\Theta_o$  bestimmen

Stichprobenumfang bestimmen

1. Bei gegebener Genauigkeit  $d$  und Vertrauensniveau  $\gamma$ :
- $\sigma^2$  bekannt:  $n \geq (\frac{2c\sigma}{d})^2$
  - Auf nächste ganze Zahl aufrunden
  - $c$  aus entsprechender Verteilung
2. Bei unbekannter Varianz:
- Vorerhebung durchführen
  - Varianz schätzen
  - t-Verteilung verwenden

Vertrauensintervall für Mittelwert

Gegeben:  $n = 25$  Messungen,  $\bar{x} = 102, s = 4, \gamma = 0.95$

1. Verteilungstyp: t-Verteilung ( $\sigma^2$  unbekannt)
- $f = 24$  Freiheitsgrade
  - $p = 0.975$
  - $c = t_{(0.975; 24)} = 2.064$
2. Grenzen berechnen:
- $e = 2.064 \cdot \frac{4}{\sqrt{25}} = 1.652$
  - $[102 - 1.652; 102 + 1.652]$
  - $[100.348; 103.652]$