

Deskriptive Statistik

Teilbereiche der Statistik

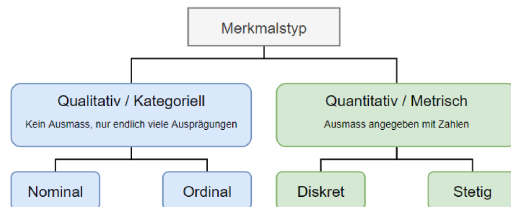
- **Deskriptive Statistik:** Beschreibung und übersichtliche Darstellung von Daten, Ermittlung von Kenngrößen und Datenvalidierung
- **Explorative Statistik:** Weiterführung und Verfeinerung der beschreibenden Statistik, insbesondere die Suche nach Strukturen und Besonderheiten
- **Induktive Statistik:**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



Merkmalstypen

- **Würfelfall (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 ($i = 1, \dots, m$)

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

Würfelfall 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): $Q_3 = x_6 = 8$

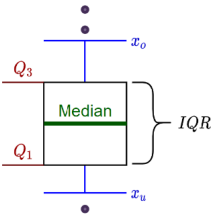
$i = [8 \cdot 0.75] = [6] = 6$

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
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? \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)

Diskrete Wahrscheinlichkeitsräume

Symbole

- Ω : Ergebnisraum (Menge aller möglichen Ergebnisse)
- ω : Ergebnis eines Zufallsexperiments
- $|\Omega|$: Anzahl der Elemente im Ergebnisraum
- $P: \Omega \rightarrow [0, 1]$: 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} : Komplementäreignis 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** $P: 2^\Omega \rightarrow [0, 1]$ ist definiert durch:

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

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: $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$

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

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 $X: \Omega \rightarrow \mathbb{R}$, 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: $x \leq y \Rightarrow F(x) \leq F(y)$
- $P(a < X \leq b) = F(b) - F(a)$

Erwartungswert und Varianz Für eine diskrete Zufallsvariable X :

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

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

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

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:

- $P(X = 5\text{€}) = 1/6$, $P(X = 2\text{€}) = 1/6$, $P(X = -1\text{€}) = 4/6$

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:

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

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

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 ∩ B) = P(A) · P(B)

Eigenschaften der stochastischen Unabhängigkeit

Für unabhängige Ereignisse A und B gilt:

- A und Ω \ B sind unabhängig
- Ω \ A und Ω \ 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 ∈ ℝ:

P(X = x, Y = y) = P(X = x) · P(Y = y)

Prüfung auf stochastische Unabhängigkeit

- 1. **Für Ereignisse**
 - Berechne P(A ∩ B) und P(A) · P(B)
 - Vergleiche die Werte
- 2. **Für Zufallsvariablen**
 - Stelle Verbundverteilung auf und prüfe für alle Wertepaare: P(X = x, Y = y) = P(X = x) · 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) = 1/6
 - P(B) = 1/2
- 2. **Schnittwahrscheinlichkeit:** P(A ∩ B) = 1/12 = 1/6 · 1/2 = P(A) · P(B)
- 3. **Schlussfolgerung:** Die Ereignisse sind stochastisch unabhängig

Kartenziehen ohne Zurücklegen

Aufgabe: Aus einem Kartenspiel werden nacheinander zwei Karten gezogen.

Ereignisse:

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

Lösung:

- 1. **Wahrscheinlichkeiten:**
 - P(A) = 13/52 = 1/4
 - P(B|A) = 12/51
 - P(B|Ā) = 13/51
- 2. **Prüfung:**

P(B) = 13/52 ≠ P(B|A)
- 3. **Schlussfolgerung:** Die Ereignisse sind stochastisch abhängig

Bedingte Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit von B unter der Bedingung A ist:

P(B|A) = P(A ∩ B) / P(A) für P(A) > 0

Multiplikationssatz P(A ∩ B) = P(A) · P(B|A) = P(B) · 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) = P(A ∩ B) / P(A) und P(A|B) = P(A ∩ 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(krank|positiv) = 0.0098 / 0.0593 ≈ 0.165 = 16.5%

Satz der Totalen Wahrscheinlichkeit

P(B) = P(A) · P(B|A) + P(Ā) · P(B|Ā)

- 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) = P(A) · 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) · P(B|A) + P(Ā) · P(B|Ā)
- 3. **Berechne P(A|B)**
 - Setze in Bayes-Formel ein und interpretiere 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|Ā) = 0.03 (Test positiv wenn gut)
- P(T) = 0.05 · 0.98 + 0.95 · 0.03 = 0.0775
- P(F|T) = (0.05 · 0.98) / 0.0775 ≈ 0.632 = 63.2%

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 ≤ x) = ∫_{-∞}^x f(t)dt (CDF)

Gegenüberstellung von diskreten und stetigen Zufallsvariablen

Diskrete Zufallsvariable:

- Dichtefunktion: f(x) = P(X = x)
- Verteilungsfunktion: F(x) = ∑_{x ≤ X} f(x)
- Wahrscheinlichkeiten: P(a ≤ X ≤ b) = ∑_{a ≤ x ≤ b} f(x)
- Erwartungswert: E(X) = ∑_{x ∈ ℝ} x · f(x)
- Varianz: V(X) = ∑_{x ∈ ℝ} (x - E(X))² · f(x)

Stetige Zufallsvariable:

- Dichtefunktion: f(x) = F'(x) ≠ P(X = x)
- Verteilungsfunktion: F(x) = ∫_{-∞}^x f(t)dt
- Wahrscheinlichkeiten: P(a ≤ X ≤ b) = ∫_a^b f(x)dx
- Erwartungswert: E(X) = ∫_{-∞}[∞] x · f(x)dx
- Varianz: V(X) = ∫_{-∞}[∞] (x - E(X))² · f(x)dx

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 (fixe Anzahl)
 - 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 \rightarrow Binomialverteilung
 - Ohne Zurücklegen \rightarrow Hypergeometrische Verteilung
 - Seltene Ereignisse \rightarrow Poisson-Verteilung
- 2. **Prüfe Grundgesamtheit**
 - Endlich, klein \rightarrow Hypergeometrische Verteilung
 - Sehr groß/unendlich \rightarrow Binomialverteilung
 - Zeitlich/räumlich kontinuierlich \rightarrow Poisson-Verteilung
- 3. **Beachte Approximationen**
 - Binomial \rightarrow Poisson für $n \rightarrow \infty, p \rightarrow 0, np = \lambda$
 - Hypergeometrisch \rightarrow 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
- Kenngößen:
- $E(X) = n \cdot \frac{M}{N}$
 - $V(X) = n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$

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

Voraussetzungen für die Bernoulli-Verteilung: Genau zwei mögliche Ausgänge, unabhängige Wiederholungen, konstante Erfolgswahrscheinlichkeit.

Binomialverteilung n -malige **unabhängige Wiederholung** von Bernoulli

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

Poissonverteilung Modelliert **seltene Ereignisse** in festem 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
- Kenngößen:
- $E(X) = \lambda$
 - $V(X) = \lambda$

Stetige Verteilungen

Erwartungswert und Varianz der Normalverteilung

Für eine Zufallsvariable $X \sim N(\mu; \sigma)$ gilt:

- Parameter:
- $E(X) = \mu, \quad V(X) = \sigma^2$
- 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: $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$

Notation: $X \sim N(0, 1)$

Eigenschaften der Normalverteilung

- Symmetrisch bzgl. der Geraden $x = \mu$, Wendepunkte bei $\mu \pm \sigma$
- Änderung μ schiebt in x -Richtung, je grösser σ , desto breiter/flacher
- normiert: $\int_{-\infty}^{\infty} \varphi_{\mu, \sigma}(x) dx = 1$.

Die Verteilungsfunktion der Normalverteilung

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

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

Standardisierung der Normalverteilung

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

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

Diese Zufallsvariable U ist standardnormalverteilt $N(0, 1)$.

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

Erwartungswert und Varianz für Zufallsvariablen

Für n unabhängige Zufallsvariablen X_1, X_2, \dots, X_n definieren wir:

n -te Summe $S_n = X_1 + \dots + X_n = \sum_{i=1}^n X_i$

arithmetische Mittel der Zufallsvariablen: $\bar{X}_n = \frac{S_n}{n}$

Für diese beiden neuen Zufallsvariablen gilt:

- $E(S_n) = E(X_1) + \dots + E(X_n) = E(X_1 + \dots + X_n)$
- $E(\bar{X}_n) = \frac{1}{n}(E(X_1) + \dots + E(X_n)) = E(\frac{1}{n}(X_1 + \dots + X_n))$

Sind die Zufallsvariablen paarweise stochastisch unabhängig gilt:

- $V(S_n) = V(X_1) + \dots + V(X_n) = V(X_1 + \dots + X_n)$
- $V(\bar{X}_n) = \frac{1}{n^2}(V(X_1) + \dots + V(X_n)) = V(\frac{1}{n}(X_1 + \dots + X_n))$

Zentraler Grenzwertsatz

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

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

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

Die standardisierte Zufallsvariable:

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

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 $n \rightarrow \infty$ gegen die Verteilungsfunktion $\phi(u)$ der Standardnormalverteilung:

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

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

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 → Normalverteilung:

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

Binomialverteilung → Poissonverteilung:

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

Hypergeometrisch → 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:

$$B(n, p) \approx Poi(n \cdot p)$$

- 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
 $Poi(\lambda)$ = 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 → Normal: Große Stichproben, mittleres p
- Binomial → Poisson: Große n , kleines p
- Hypergeometrisch → Binomial: Kleine Stichprobe relativ zur Grundgesamtheit

4. Beachte

- Stetigkeitskorrektur bei Normal
- Rundungsregeln bei Grenzen
- Vergleich mit exakter Lösung wenn möglich

Approximation der Binomialverteilung

Aufgabe: Eine Produktionsanlage produziert mit Ausschusswahrscheinlichkeit 5%. In einer Charge von 200 Teilen:

- Wie groß ist die Wahrscheinlichkeit für 15 oder mehr defekte Teile?

Lösung:

1. Prüfung Approximationsbedingung:

- $npq = 200 \cdot 0.05 \cdot 0.95 = 9.5 > 9$
- Normalapproximation ist zulässig

2. Parameter der Normalverteilung:

- $\mu = np = 200 \cdot 0.05 = 10$
- $\sigma = \sqrt{npq} = \sqrt{9.5} \approx 3.08$

3. Berechnung mit Stetigkeitskorrektur:

$$\begin{aligned} P(X \geq 15) &= 1 - P(X \leq 14) \\ &= 1 - P(X \leq 14.5) \\ &= 1 - \Phi(\frac{14.5 - 10}{3.08}) \\ &= 1 - \Phi(1.46) \\ &\approx 0.0721 \end{aligned}$$

Approximation durch Poissonverteilung

Aufgabe: Ein seltener Gendefekt tritt mit Wahrscheinlichkeit $p = 0.001$ auf. In einer Gruppe von 2000 Menschen:

- Wie groß ist die Wahrscheinlichkeit für genau 3 Betroffene?

Lösung:

1. Prüfung Approximationsbedingung:

- $n = 2000 \geq 50$ und $p = 0.001 \leq 0.1$
- Poissonapproximation ist zulässig

2. Parameter:

- $\lambda = np = 2000 \cdot 0.001 = 2$

3. Berechnung:

$$P(X = 3) = \frac{2^3}{3!} \cdot e^{-2} \approx 0.180$$

4. Vergleich mit Binomialverteilung:

$$P_{Bin}(X = 3) = \binom{2000}{3} \cdot 0.001^3 \cdot 0.999^{1997} \approx 0.180$$

Die Methode der kleinsten Quadrate

Einführung

Einführung

Weit verbreitete Optimierungsmethode zur Modellierung mathematischer Zusammenhänge in großen Datenmengen. Ziel: optimale Parameter zu finden, die funktionalen Zusammenhang zwischen Messdaten am besten beschreiben.

Lineare Regression: linearer Zusammenhang zwischen Daten vermutet und versucht, optimale Gerade in 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:

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

Die Residuen ϵ_i ergeben sich als: $\epsilon_i = y_i - \hat{y}_i = y_i - (mx_i + q)$

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 \hat{s}_ϵ^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: $\hat{s}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = (\frac{1}{n} \sum_{i=1}^n x_i^2) - \bar{x}^2$

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

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

Lineare Regression berechnen

1. Berechne arithmetische Mittel \bar{x} und \bar{y}

2. Berechne Kovarianzen und Varianzen

3. Berechne Steigung m und y -Achsenabschnitt d :

$$m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}, d = \bar{y} - m\bar{x}$$

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

- $\bar{x} = 3, \bar{y} = 6.02$
- Kovarianzen und Varianzen:
 - $s_{xy} = 3.945, s_x^2 = 2, s_y^2 = 8.4916$
- Parameter:
 - $m = \frac{3.945}{2} = 1.9725, d = 6.02 - 1.9725 \cdot 3 = 0.1025$
- 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:

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

\hat{s}_y^2 : prognostizierte (erklärte) Varianz, \hat{s}_ϵ^2 : Residualvarianz

Bestimmtheitsmass R^2 (zwischen 0 und 1)

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

$$R^2 = \frac{\hat{s}_y^2}{s_y^2}$$

s_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_x^2 = Varianz der x -Werte, s_y^2 = Varianz der y -Werte

s_{xy} = Kovarianz von x und y

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

- Berechne die totale Varianz s_y^2
- Berechne die Residualvarianz s_ϵ^2
- Berechne die erklärte Varianz \hat{s}_y^2
- Berechne das Bestimmtheitsmass:

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

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

- Berechne die Residuen für jeden Datenpunkt:

- $\epsilon_i = y_i - (mx_i + d)$

- Erstelle Residuenplot:

- x-Achse: Prognostizierte Werte $\hat{y}_i = mx_i + d$
- y-Achse: Residuen ϵ_i

- Prüfe Eigenschaften:

- Residuen sollten zufällig um Null streuen
- Keine systematischen Muster erkennbar
- Gleiche Streubreite über alle \hat{y}_i

Gütekriterien für Regression

- 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

- Residuenanalyse:

- Residuen sollten zufällig um 0 schwanken
- Keine systematischen Muster erkennbar
- Residuen sollten normalverteilt sein

- 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.

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

- Bestimme passende Transformation aus Tabelle
- Führe Transformation durch
- Wende lineare Regression auf transformierte Daten an
- 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

- 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

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

Allgemeines Vorgehen bei der Regression

Matrix-Darstellung m, q der Regressionsgeraden mit A berechnen:

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

Matrix-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-Methode für lineare Regression

- Erstelle Design-Matrix A : $A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$
- Berechne $A^T \cdot A$
- Berechne $(A^T \cdot A)^{-1}$
- Berechne Parameter: $\begin{pmatrix} m \\ q \end{pmatrix} = (A^T \cdot A)^{-1} \cdot A^T \cdot \bar{y}$

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: $A = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1(k-1)} & 1 \\ x_{21} & x_{22} & \dots & x_{2(k-1)} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{n(k-1)} & 1 \end{pmatrix}$
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 Regression mit Polynomen höheren Grades:

1. Erweitere Designmatrix: $A = \begin{pmatrix} x_1^n & x_1^{n-1} & \dots & x_1 & 1 \\ x_2^n & x_2^{n-1} & \dots & x_2 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_m^n & x_m^{n-1} & \dots & x_m & 1 \end{pmatrix}$
2. Löse wie bei linearer Regression: $\vec{p} = (A^T A)^{-1} A^T \vec{y}$
3. Polynom aufstellen: $y = p_1 x^n + p_2 x^{n-1} + \dots + p_n x + p_{n+1}$

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

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, \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:

- Alle Stichprobenvariablen sind stochastisch unabhängig
- Alle X_i folgen derselben Verteilung $F(x)$ der Grundgesamtheit

Parameterschätzungen

Schätzfunktionen

Schätzfunktion

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

Grundlagen der Schätztheorie

Die Schätztheorie befasst sich mit zwei Hauptproblemen:

- Punktschätzung: Bestimmung eines einzelnen Schätzwerts
- Intervallschätzung: Bestimmung eines Vertrauensbereichs

Wichtige Begriffe:

- θ : Unbekannter Parameter der Grundgesamtheit
- Θ : Schätzfunktion (Zufallsvariable)
- $\hat{\theta}$: Schätzwert (konkreter Wert)
- n : Stichprobenumfang

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: $E(\Theta) \rightarrow \theta$ und $V(\Theta) \rightarrow 0$ für $n \rightarrow \infty$
→ Grenzwert für $n \rightarrow \infty$ betrachten

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)?

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

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

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:

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

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

Wichtige Schätzfunktionen

Schätzfunktionen für wichtige Parameter

Erwartungswert:

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

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)

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

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

Konstruktion eines Vertrauensintervalls

- 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: Konstruktion eines Vertrauensintervalls (Normalverteilung)
Gegeben: **normalverteilte** Zufallsvariable X mit unbekanntem Parameter μ und bekannter Varianz σ^2 . $\gamma = 0.95$.

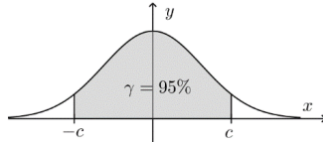
Aufgabe: Konstruiere ein Vertrauensintervall für den Mittelwert μ .

- 1. Verteilungstyp bestimmen siehe Übersicht Vertrauensintervalle
- 2. Gleichung aufstellen Die Schätzfunktion für μ liefert Ausgangspunkt für Vertrauensintervall, ausgehend von diesem Punkt berechnen wir die Grenzen.
Die Grenzen sollen symmetrisch um \bar{X} liegen. Wir suchen also eine Schranke e so dass gilt:

$$P(\bar{X} - e \leq \mu \leq \bar{X} + e) = \gamma \text{ bzw. } P(|\bar{X} - \mu| \leq e) = \gamma(*)$$

- 3. Standardisierung Nach dem Zentralen Grenzwertsatz ist \bar{X} normalverteilt mit $E(\bar{X}) = \mu$ und $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Damit wir Wahrscheinlichkeiten bestimmen können, müssen wir statt \bar{X} die standardisierte Zufallsvariable $U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ verwenden; U ist standardnormalverteilt. Die Gleichung (*) lässt sich umformen zu: $P(|U| \leq \frac{e}{\sigma/\sqrt{n}}) = \gamma$

4. c bestimmen



Die Illustration zeigt die Situation, dabei ist $c = \frac{e}{\sigma/\sqrt{n}}$.
Wir suchen also:
das c mit $\phi(c) = \frac{1+\gamma}{2} = 0.975$.

Aus der Standardnormalverteilungstabelle erhalten wir $c = 1.96$.

5. Intervallgrenzen berechnen $e = c \cdot \frac{\sigma}{\sqrt{n}}$

Formeln für die Stichprobenfunktionen und Grenzen:

$$\Theta_u = \bar{X} - \underbrace{c \cdot \frac{\sigma}{\sqrt{n}}}_e, \quad \Theta_o = \bar{X} + \underbrace{c \cdot \frac{\sigma}{\sqrt{n}}}_e$$

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)

- c : Quantil der Verteilung
- p : Wahrscheinlichkeit für Quantil
- f : Freiheitsgrade
- s : Schätzwert für σ
- S^2 : Schätzvarianz
- n : Stichprobenumfang
- \bar{X} : Stichprobenmittelwert
- γ : Vertrauensniveau
- α : Irrtumswahrscheinlichkeit
- μ : Wahre Parameter der Grundgesamtheit
- σ : Wahre Varianz der Grundgesamtheit
- Θ_u, Θ_o : Unter- und Obergrenze des Intervalls
- c_1, c_2 : Quantile der Verteilung
- p_1, p_2 : Wahrscheinlichkeiten für Quantile

γ gibt Wahrscheinlichkeit an, dass das Intervall den wahren Parameter θ enthält. Irrtumswahrscheinlichkeit $\alpha = 1 - \gamma$.
Beispiel: $\gamma = 0.95$ bedeutet, dass in 95% der Fälle das Intervall den wahren Parameter enthält $\rightarrow \alpha = 0.05$
Meist kann das Vertrauensniveau γ frei gewählt werden (α möglichst klein). Häufig wird $\gamma = 0.95$ oder $\gamma = 0.99$ gewählt.

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. $e = c \cdot \frac{S}{\sqrt{n}} = 4.111, \Theta_u = \bar{X} - e = 97.89, \Theta_o = \bar{X} + e = 106.11$

Übersicht Vertrauensintervalle zum Niveau γ

Verteilung der Grundgesamtheit	Param.	Schätzfunktionen	standardisierte Zufallsvariable	Verteilung und benötigte Quantile	Intervallgrenzen
1 Normalverteilung (σ^2 bekannt)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$	$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	Standardnormalverteilung (Tab.2) $c = u_p$ mit $p = \frac{1+\gamma}{2}$	$\bar{x} \pm c \cdot \frac{\sigma}{\sqrt{n}}$
2 Normalverteilung (σ^2 unbek.)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	t-Verteilung (Tab.4) $f = n - 1$ $c = t_{p,f}$ mit $p = \frac{1+\gamma}{2}$	$\bar{x} \pm c \cdot \frac{s}{\sqrt{n}}$
3 Normalverteilung	σ^2	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$	$Z = (n-1) \frac{S^2}{\sigma^2}$	χ^2 -Verteilung (Tab.3) $f = n - 1$ $c_1 = z_{p_1, f}, p_1 = \frac{1+\gamma}{2}$ $c_2 = z_{p_2, f}, p_2 = \frac{1-\gamma}{2}$	$\Theta_u = \frac{(n-1)s^2}{c_2}$ $\Theta_o = \frac{(n-1)s^2}{c_1}$
4 Bernoulli-Verteilung Anteilsschätzung	p	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ X_i 0/1-wertig $P(X_i = 1) = p$	$U = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$	Standardnormalverteilung (Tab.2) $c = u_p$ mit $q = \frac{1+\gamma}{2}$	$\bar{x} \pm c \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$
5 beliebig mit $n \geq 30$	μ, σ^2				

Wie im Fall 1 (gegebenenfalls mit s als Schätzwert), bzw. im Fall 3

Beispiele

Kombinatorik

Komplexeres Beispiel: Passwörter

Aufgabe: Ein Passwort muss bestehen aus:

- Genau 8 Zeichen
- Mindestens ein Großbuchstabe (26 mögliche)
- Mindestens eine Ziffer (10 mögliche)
- Kleine Buchstaben erlaubt (26 mögliche)

Lösung: 1. Gesamtzahl aller möglichen 8-stelligen Passwörter mit den Zeichen:

- $n = 26 + 26 + 10 = 62$ Zeichen
- Variation mit Wiederholung: 62^8

2. Abziehen der ungültigen Kombinationen:

- Ohne Großbuchstaben: $(36)^8$
- Ohne Ziffern: $(52)^8$
- Ohne beide: $(26)^8$

3. Nach dem Inklusions-Exclusions-Prinzip: Gültige Passwörter $= 62^8 - 36^8 - 52^8 + 26^8$

Wahrscheinlichkeitsrechnung Beispiele

Lotterie mit bedingten Gewinnen

Aufgabe: Bei einer Lotterie gewinnt man zunächst mit $p = 0.1$ einen Bonus-Los. Mit diesem Los kann man dann mit $p = 0.2$ den Hauptpreis von 1000€ gewinnen. Berechne den Erwartungswert.

Lösung:

1. **Ereignisbaum erstellen:**

- $P(\text{Bonus}) = 0.1$
- $P(\text{Hauptgewinn}|\text{Bonus}) = 0.2$

2. **Mögliche Ausgänge:**

- 1000€: $P = 0.1 \cdot 0.2 = 0.02$
- 0€: $P = 0.98$

3. **Erwartungswert:** $E(X) = 1000 \cdot 0.02 + 0 \cdot 0.98 = 20$

Aktienportfolio **Aufgabe:** Ein Portfolio besteht aus:

- Aktie A: 60% Anteil, $E(A) = 8\%$, $V(A) = 25$
- Aktie B: 40% Anteil, $E(B) = 12\%$, $V(B) = 36$

Lösung:

1. **Erwartungswert des Portfolios:**

$$\begin{aligned} E(P) &= 0.6 \cdot E(A) + 0.4 \cdot E(B) \\ &= 0.6 \cdot 8\% + 0.4 \cdot 12\% \\ &= 4.8\% + 4.8\% = 9.6\% \end{aligned}$$

2. **Varianz des Portfolios** (bei Unabhängigkeit):

$$\begin{aligned} V(P) &= (0.6)^2 \cdot V(A) + (0.4)^2 \cdot V(B) \\ &= 0.36 \cdot 25 + 0.16 \cdot 36 \\ &= 9 + 5.76 = 14.76 \end{aligned}$$

3. **Standardabweichung:** $S(P) = \sqrt{14.76} \approx 3.84\%$

Hypergeometrische Verteilung

Ziehung ohne Zurücklegen **Aufgabe:** In einer Urne sind 20 Kugeln, davon 8 rot. Es werden 5 Kugeln ohne Zurücklegen gezogen.

Lösung:

1. **Parameter:** $N = 20$ (Gesamtanzahl), $M = 8$ (rote Kugeln), $n = 5$ (Ziehungen)
2. **Erwartungswert:**

$$E(X) = 5 \cdot \frac{8}{20} = 2$$

3. **Varianz:**

$$V(X) = 5 \cdot \frac{8}{20} \cdot \frac{12}{20} \cdot \frac{15}{19} \approx 1.184$$

4. **P(genau 2 rote):** $P(X = 2) = \frac{\binom{8}{2} \binom{12}{3}}{\binom{20}{5}} \approx 0.3682$

Bernoulli-Verteilung

Münzwurf

Aufgabe: Faire Münze wird geworfen. $X = 1$ bei Kopf, $X = 0$ bei Zahl.

Lösung:

- $p = 0.5$ (faire Münze)
- $E(X) = 0.5$
- $V(X) = 0.5 \cdot 0.5 = 0.25$
- $P(X = 1) = 0.5$
- $P(X = 0) = 0.5$

Binomialverteilung

Qualitätskontrolle mit Binomialverteilung

Aufgabe: Eine Maschine produziert Teile mit Ausschussquote 5%. In einer Stichprobe von 100 Teilen:

- a) Wie viele defekte Teile sind zu erwarten?
- b) Wie groß ist die Wahrscheinlichkeit für genau 3 defekte Teile?
- c) Wie groß ist die Wahrscheinlichkeit für höchstens 2 defekte Teile?

Lösung:

1. **Parameter:**

- $n = 100$ (Stichprobenumfang)
- $p = 0.05$ (Ausschusswahrscheinlichkeit)
- $X \sim B(100, 0.05)$

2. **Erwartungswert:**

$$E(X) = np = 100 \cdot 0.05 = 5$$

3. **Genau 3 defekte:**

$$P(X = 3) = \binom{100}{3} (0.05)^3 (0.95)^{97} \approx 0.1404$$

4. **Höchstens 2 defekte:**

$$P(X \leq 2) = \sum_{k=0}^2 \binom{100}{k} (0.05)^k (0.95)^{100-k} \approx 0.0861$$

Poisson-Verteilung

Poisson-Verteilung in der Praxis

Aufgabe: Ein Callcenter erhält durchschnittlich 3 Anrufe pro 10 Minuten.

- a) Wahrscheinlichkeit für genau 2 Anrufe in 10 Minuten?
- b) Wahrscheinlichkeit für mehr als 4 Anrufe?

Lösung:

1. **Parameter:** $\lambda = 3$ (Erwartungswert), $X \sim \text{Poi}(3)$

2. **Genau 2 Anrufe:**

$$P(X = 2) = \frac{3^2}{2!} e^{-3} \approx 0.2240$$

3. **Mehr als 4 Anrufe:** $P(X > 4) = 1 - \sum_{k=0}^4 \frac{3^k}{k!} e^{-3} \approx 0.1847$

Normalverteilung

Körpergrößen

Aufgabe: Körpergrößen in einer Population sind normalverteilt mit $\mu = 175$ cm und $\sigma = 10$ cm.

Berechnung:

- $P(X \leq 185) = \Phi\left(\frac{185-175}{10}\right) = \Phi(1) \approx 0.8413$
- $P(165 \leq X \leq 185) = \Phi(1) - \Phi(-1) \approx 0.6826$
- $P(X > 195) = 1 - \Phi(2) \approx 0.0228$

Parameter-/Intervallschätzung

Intervallschätzung für die Varianz

Für die Varianz σ^2 einer Normalverteilung mit Stichprobenumfang $n = 10$ und Stichprobenvarianz $s^2 = 16$ soll ein 99%-Vertrauensintervall berechnet werden.

1. Verteilungstyp: Chi-Quadrat-Verteilung

2. Freiheitsgrade: $f = n - 1 = 9$

3. Quantile: $c_1 = \chi^2_{(0.005;9)} = 1.735$, $c_2 = \chi^2_{(0.995;9)} = 23.589$

4. Vertrauensintervall:

$$\frac{(n-1)s^2}{c_2} \leq \sigma^2 \leq \frac{(n-1)s^2}{c_1}$$

n = Stichprobenumfang

s^2 = Stichprobenvarianz

c_1, c_2 = Chi-Quadrat-Quantile

σ^2 = Wahre Varianz der Grundgesamtheit

$$\frac{9 \cdot 16}{23.589} \leq \sigma^2 \leq \frac{9 \cdot 16}{1.735}$$

$$6.10 \leq \sigma^2 \leq 82.99$$

Bernoulli-Anteilsschätzung

Ein Vertrauensintervall für den Parameter p einer Bernoulli-Verteilung soll aus einer Stichprobe mit $n = 100$ und $\bar{x} = 0.42$ bei einem Vertrauensniveau von 95% berechnet werden.

1. Prüfen der Voraussetzung: $n\hat{p}(1 - \hat{p}) = 100 \cdot 0.42 \cdot 0.58 = 24.36 > 9$

2. Quantil: $c = u_{0.975} = 1.96$

3. Standardfehler: $\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} = \sqrt{\frac{0.42 \cdot 0.58}{100}} = 0.0494$

4. Vertrauensintervall:

$$0.42 \pm 1.96 \cdot 0.0494 = [0.323; 0.517]$$

n = Stichprobenumfang

\bar{x} = Stichprobenmittelwert (Anteil der Erfolge)

\hat{p} = Geschätzter Parameter der Bernoulli-Verteilung

$u_{0.975} = 97.5$