Einführung in die Angewandte Stochastik

Prof. Dr. Ansgar Steland
Institut für Statistik und Wirtschaftsmathematik
RWTH Aachen

Grundbegriffe

Zufallsvorgänge

- Ausgang/Ergebnis steht nicht (deterministisch) fest.
- Ausgang hängt vom Zufall ab.

Wir brauchen:

- Mathematisches Modell
- Geeigneten Wahrscheinlichkeitskalkül.

Grundbegriffe: Beispiele/Motivation

Beispiel 1

- 50 Ladekabel, davon 1 defekt.
- Wkeit, dass der nächste Käufer das defekte Kabel greift?

Gesunder Menschenverstand diktiert:

$$p_k = P(\text{,,}k\text{-tes Ladekabel wird gegriffen''}) = \frac{1}{50}$$

für $k = 1, \ldots, 50$. Gleichwahrscheinlichkeit

Grundbegriffe: Beispiele/Motivation

Modellbildung

- Ergebnismenge:

$$\Omega = \{1, \ldots, 50\} \qquad \text{Menge}$$

- Wkeiten p_k, $k = 1, \ldots, 50$. Hier:

$$p_k = \frac{1}{50}, \qquad k = 1, \ldots, 50.$$

- Ereignis Teilmenge der Ergebnismenge

$$A = \text{,,Ladekabel mit gerader Serien-Nummer"} = \{2, 4, \ldots, 50\}.$$

- Wahrscheinlichkeit von A: $P(A) = \frac{25}{50} = \frac{1}{2}$

Grundbegriffe: Beispiele/Motivation

Allgemeiner Ansatz: Modelliere Zufallsvorgang durch

- Ergebnismenge/Grundmenge $\Omega = \{\omega_1, \ldots, \omega_N\}$
- $\omega \in \Omega$: Versuchsausgang/Ergebnis/Ausgang
- Ereignis: $A \subset \Omega$
- Ordne ω_i die Wkeit $p_i \in [0, 1]$ zu.
- $\sum_{i=1}^{N} p_i = 1$.
- Ereignis A tritt ein: $\omega \in A$.

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) = P(\{\omega \in \Omega : \omega \in A\})$$

Grundbegriffe: Beispiele/Motivation

Grundbegriffe: Beispiele/Motivation

Grundbegriffe: Beispiele/Motivation

Beispiel: Max, Niklas, Laura und Sahra wohnen in einer WG. Der Putzplan für die nächsten 2 Wochen wird ausgelost. Sie legen vier Zettel mit ihren Namen in eine Dose. Emma von nebenan spielt die Glücksfee und zieht zwei Zettel. Der zuerst Gezogene putzt in dieser Woche, der andere in der zweiten Woche.

Ergebnismenge Ω?

Formale Darstellung des Ereignisses A, dass die Männer in beiden Wochen putzen müssen?

Grundbegriffe: Beispiele/Motivation

Sprechweisen

Wahrscheinlichkeiten sind Zahlen zwischen 0 und 1 (also Prozentzahlen).

- **Sicheres Ereignis**: $P(A) = 1$, z.B. $A = \Omega$.
- **Unmögliches Ereignis**: $P(A) = 0$, z.B. $A = \emptyset$.
- Zwei Ereignisse A und B heißen **disjunkt**, wenn $A \cap B = \emptyset$.
- Ereignisse $A_1, A_2, \ldots$ heißen **paarweise disjunkt**, wenn

$$A_i \cap A_j = \emptyset \qquad \text{für alle } i, j \geq 1 \text{ mit } i \neq j$$

(d.h., wenn alle Paare von verschiedenen Mengen disjunkt sind).

Grundbegriffe: Beispiele/Motivation

Wartezeit T auf ersten Ausfall eines Cloud-Servers

- $\Omega = \mathbb{N}$ (Es gibt keine obere Schranke für T.)
- Wkeit für Wartezeit $k \in \mathbb{N}$: p_k

Die $\{p_i\} \subset [0, 1]$ erfüllen:

$$\sum_{i=1}^{\infty} p_i = 1$$

(da der Server auf jeden Fall irgendwann ausfällt).

Grundbegriffe: Beispiele/Motivation

Wichtige Beobachtungen:

- Gleichwahrscheinlichkeitsmodell ausgeschlossen!
- Es gibt eine unendliche Folge von paarweise disjunkten Ereignissen:

$$A_k = \{k\}, \qquad k = 1, 2, \ldots$$

  (für die wir gerne Additivität von $P(\cdot)$ hätten).
- Rechnen mit endlichen Summen, $\sum_{k=1}^{N} P(A_k)$, genügt nicht, es treten auch Reihen auf:

$$\sum_{k=1}^{\infty} P(A_k) = P(A_1) + P(A_2) + \ldots$$

Grundbegriffe: Beispiele/Motivation

Harmonische Reihe (divergent)

$$\sum_{i=1}^{\infty} \frac{1}{i} = +\infty$$

Konvergente Reihen sind die geometrische Reihe

$$\sum_{i=0}^{\infty} q^i = \frac{1}{1-q}, \qquad |q| < 1,$$

die Exponentialreihe

$$e^x = \exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

sowie

$$\sum_{k=0}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

Grundbegriffe: Beispiele/Motivation

Im Fall $\Omega = \mathbb{N}$ sind nicht alle Folgen $(p_i : i \in \mathbb{N})$ möglich.
Beispiel: $p_i = c/i$ für $i \in \mathbb{N}$ und eine positive Konstante. Dann gilt:

Beispiel: $p_i = \frac{6}{\pi^2}\frac{1}{i^2}$ für $i \in \mathbb{N}$. Dann gilt:

Beispiel: $p_i = 0.8 \cdot (0.2)^i$ für $i \in \mathbb{N}_0$:

Definition (Ereignisalgebra)

Ist Ω höchstens abzählbar unendliche Menge, dann heißt jede Teilmenge $A \subset \Omega$ **Ereignis** und die Potenzmenge

$$\text{Pot}(\Omega) = \{A : A \subset \Omega\}$$

heißt **Ereignisalgebra**. Die Mengen $A = \{\omega\}$, $\omega \in \Omega$, heißen **Elementarereignisse**.

Ereignisse sind also Mengen. Alle Operationen und Regeln für Mengen übertragen sich daher auf Ereignisse.

Operationen mit Ereignissen

Definition

UND-, ODER- und komplementäres Ereignis $A \subset \Omega$ und $B \subset \Omega$ Ereignisse.

UND-Ereignis

$$A \cap B = \{\omega \mid \omega \in A \text{ und } \omega \in B\}$$

ODER-Ereignis

$$A \cup B = \{\omega \mid \omega \in A \text{ oder } \omega \in B\}$$

NICHT-Ereignis (komplementäres Ereignis):

$$\overline{A} = A^c = \{\omega \mid \omega \in \Omega \text{ und } \omega \notin A\} = \Omega \setminus A$$

Hinweis: Diese Mengenoperationen entsprechen also den logischen Grundoperationen.

Operationen mit Ereignissen

Die wichtigsten Regeln:

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
3. $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$
4. $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

1) und 2) sind die Distributivgesetze, 3) und 4) die Regeln von DeMorgan.
Allgemeiner:

1. $A \cap (B_1 \cup \cdots \cup B_n) = (A \cap B_1) \cup \cdots \cup (A \cap B_n)$
2. $A \cup (B_1 \cap \cdots \cap B_n) = (A \cup B_1) \cap \cdots \cap (A \cup B_n)$
3. $\overline{(A_1 \cup \cdots \cup A_n)} = \overline{A}_1 \cap \cdots \cap \overline{A}_n$
4. $\overline{(A_1 \cap \cdots \cap A_n)} = \overline{A}_1 \cup \cdots \cup \overline{A}_n$

Wahrscheinlichkeitsmaß

Wahrscheinlichkeitsmaß (W-Maß), Wahrscheinlichkeitsverteilung (W-Verteilung)

Ein **W-Maß** (**W-Verteilung**) ist eine Abbildung, die jedem Ereignis $A \subset \Omega$ eine Zahl $P(A) \in \mathbb{R}$ zuordnet, so dass die Kolmogorov-Axiome gelten:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. Sind $A_1, A_2, \ldots$ paarweise disjunkt, dann gilt

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots = \sum_{k=1}^{\infty} P(A_k)$$

Ein **Zufallsexperiment** ist durch Angabe von Ω und P eindeutig beschrieben, also gegeben durch das Tupel (Ω, P).

Endliches Ω

Beispiel

Ist $\Omega = \{\omega_1, \ldots, \omega_N\}$ eine endliche Menge, dann kann P durch eine Tabelle angegeben werden:

ω	ω_1	ω_2	$\ldots$	ω_N
$P(\{\omega\})$	p_1	p_2	$\ldots$	p_N

$P(A)$: Summe der p_i, die zu Elementen ω_i aus A gehören.
Welche Schreibweisen sind korrekt?

$$\square\ P(A) = \sum_{i=1}^{N} P(\{\omega_i\}),\quad \square\ P(A) = \sum_{i:i\in A} P(\{\omega_i\}),\quad \square\ P(A) = \sum_{i:\omega_i\in A} P(\{\omega_i\})$$

Zusammenhang zur deskriptiven Statistik

In der Praxis werden aus Datenbeständen oft relative Häufigkeiten f_j berechnet. Man darf mit den f_j so rechnen wie mit Wahrscheinlichkeiten.

- Merkmal mit Ausprägungen $a_1, \ldots, a_k$ wird erhoben, Stichprobe vom Umfang n
- Jede relative Häufigkeitsverteilung $f_1, \ldots, f_k$ definiert eine Wahrscheinlichkeitsverteilung.

Kategorie	a_1	a_2	$\ldots$	a_k
rel. Hf.	f_1	f_2	$\ldots$	f_k

  Zufallsexperiment: Wähle zufällig mit Wkeit f_j die Kategorie j aus (bzw. ziehe ein Objekt aus dieser Kategorie).

- Bei **Rohdaten** $x_1, \ldots, x_n$:
  Empirische Verteilung mit Masse $\frac{1}{n}$ auf jedem x_i, $i = 1, \ldots, n$.

x	x_1	x_2	$\ldots$	x_n
rel. Hf.	$1/n$	$1/n$	$\ldots$	$1/n$

  Zufallsexperiment: Wähle zufällig ein x_i aus.

Rechenregeln

Elementare Rechenregeln

1. $P(\overline{A}) = 1 - P(A)$.
2. Für $A \subset B$ gilt: $P(B \backslash A) = P(B) - P(A)$.
3. Für *beliebige* Ereignisse A, B gilt:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

4. Für *beliebige* Ereignisse A, B gilt:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B).$$

Rechenregeln

Regel (3) liefert die Ungleichung

$$P(A \cup B) \leq P(A) + P(B)$$

und durch *vollständige Induktion* folgt für Ereignisse $A_1, A_2, \ldots, A_n$ die **Boolesche Ungleichung** ('Union Bound')

$$P\left(\cup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Beispiel

Beispiel

Studie: Unternehmensbefragung von IT-Abteilungen großer Unternehmen:

- 15% *aller IT-Abteilungen haben steigende Kosten.*
- 10% *aller IT-Abteilungen haben fallende Umsätze.*
- 80% *haben weder steigende Kosten noch fallende Umsätze.*

Frage: *Wie hoch ist der Anteil der IT-Abteilungen, die sowohl steigende Kosten als auch fallende Umsätze haben?*

Beispiel

Lösung:
Definiere die Ereignisse:

$$K = \text{,,IT hat steigende Kosten''}$$
$$U = \text{,,IT hat fallende Umsätze''}$$

W-Maß: Empirische Verteilung
Bek: $P(K) = 0.15,\ P(U) = 0.1,\ P(\overline{K \cup U}) = 0.8.$
Ges: $P(K \cap U)$.
Es gilt: $P(K \cup U) = P(K) + P(L) - P(K \cap U)$
Umstellen:

$$\begin{aligned} P(K \cap U) &= P(K) + P(U) - P(K \cup U) \\ &= 0.15 + 0.1 - 0.2 = 0.05. \end{aligned}$$

Beispiel

Gelte nun:

- 60% haben steigende Kosten.
- 50% haben fallende Umsätze.
- 30% haben steigende Kosten und fallende Umsätze.

Frage: Wie hoch ist der Anteil der IT-Abteilungen, die genau eines der beiden Probleme haben?

Beispiel

Bek: $P(K) = 0.6,\ P(U) = 0.5,\ P(K \cap U) = 0.3.$

Ges: $P($ $\underbrace{(K \cup U)}_{\text{Kosten- o. Umsatzproblem}}$ $\setminus$ $\underbrace{(K \cap U)}_{\text{Kosten- u. Umsatzproblem}}$ $) = ?$

Es gilt: $K \cap U \subset K \cup U$ und somit

$$p = P((K \cup U) \setminus (K \cap U)) = P(K \cup U) - P(K \cap U)$$

wobei

$$P(K \cup U) = P(K) + P(U) - P(K \cap U)$$

Einsetzen:

$$\begin{aligned} p &= P(K) + P(U) - P(K \cap U) - P(K \cap U) \\ &= 0.6 + 0.5 - 2 \cdot 0.3 = 0.5 \end{aligned}$$

Beispiel: Probabilistische Garantien

Ein ML-Algorithmus prognostiziere Variablen x_i durch $\hat{x}_i$, $1 \leq i \leq n$.
Die Prognose $\hat{x}_i$ ist ε-genau, wenn $|\hat{x}_i - x_i| \leq \varepsilon$ gilt, $\varepsilon > 0$.
ε: gewünschte Fehlerschranke.
Sei $A_i = \{|\hat{x}_i - x_i| \leq \varepsilon\}$ und $B = \cap_{i=1}^n A_i$ (alle Prognosen sind ε-genau).
Für zufällig gewählte Inputdaten gelte die Garantie:

$$P(A_i^c) = P(|\hat{x}_i - x_i| > \varepsilon) \leq 2\exp(-\frac{\varepsilon^2}{2}), \qquad 1 \leq i \leq n.$$

Dann gilt

$$P(B) = 1 - P(\cup_{i=1}^n A_i^c) \geq 1 - \sum_{i=1}^n P(A_i^c) \geq 1 - 2n\exp\left(-\frac{\varepsilon^2}{2}\right)$$

Beispiel: Probabilistische Garantien

Ziel: Bestimme für gegeben max. Fehlerwahrscheinlichkeit α ein ε derart, dass mit Wahrscheinlichkeit $1-\alpha$ alle Prognosen ε-genau sind.
Ansatz: Um $P(B) \geq 1-\alpha$ zu gewährleisten, bestimmen wir ε so, dass die untere Schranke den Wert $1-\alpha$ annimmt:

$$
\begin{aligned}
& 1 - 2ne^{-\varepsilon^2/2} = 1-\alpha \Leftrightarrow 2ne^{-\varepsilon^2/2} = \alpha \\
& \Leftrightarrow e^{-\varepsilon^2/2} = \frac{\alpha}{2n} \\
& \Leftrightarrow \varepsilon^2/2 = -\ln\left(\frac{\alpha}{2n}\right) \\
& \Leftrightarrow \varepsilon = \sqrt{2ln\left(\frac{2n}{\alpha}\right)}
\end{aligned}
$$

Mit Wahrscheinlichkeit $1-\alpha$ gilt, dass alle Prognosen $\sqrt{2ln\left(\frac{2n}{\alpha}\right)}$-genau sind, d.h.

$$
\max_{1 \leq i \leq n} |\hat{x}_i - x_i| \leq \sqrt{2ln\left(\frac{2n}{\alpha}\right)}.
$$

Laplace-Raum

(Ω, P) heißt **Laplace-Raum**, wenn

$$\Omega = \{\omega_1, \ldots, \omega_K\}$$

endlich ist und das W-Maß durch

$$p(\omega) = P(\{\omega\}) = \frac{1}{K}, \qquad \omega \in \Omega,$$

gegeben ist. P heißt auch **(diskrete) Gleichverteilung auf** Ω. Dann berechnen sich Wahrscheinlichkeiten durch

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}}$$

Laplace-Räume und Kombinatorik

In Laplace-Räumen reduziert sich die Berechnung von $P(A)$ auf das Abzählen von A. Dies erfordert oft kombinatorische Fähigkeiten.

Kombinatorisches Abzählprinzip: Ist $\Omega = \Omega_1 \times \cdots \times \Omega_k$ für ein $k \in \mathbb{N}$ und $A = A_1 \times \cdots \times A_k \subset \Omega$, dann ist

$$|A| = |A_1| \cdot |A_2| \cdot \ldots \cdot |A_k|$$

Wieviele $\omega = (\omega_1, \ldots, \omega_k) \in A$ gibt es?
Zähle aus, wieviel Möglichkeiten es für ω_1 gibt. $\rightarrow$ Genau $|A_1|$.
Zähle aus, wieviel Möglichkeiten es für ω_2 gibt. $\rightarrow$ Genau $|A_2|$.
usw.
$\rightarrow |A|$ ist das Produkt dieser Anzahlen.

Allgemeines Abzählprinzip:

Das Abzählprinzip greift auch, wenn die Mengen der Möglichkeiten für ω_i von $\omega_{i-1}, \ldots, \omega_1$ abhängen: $\omega_i \in A_{\omega_1,\ldots,\omega_{i-1}}$. Dann bestimmt man für jede Kombination $(\omega_1, \ldots, \omega_{i-1})$ die Anzahl der nun möglichen Festlegungen von ω_i und hat

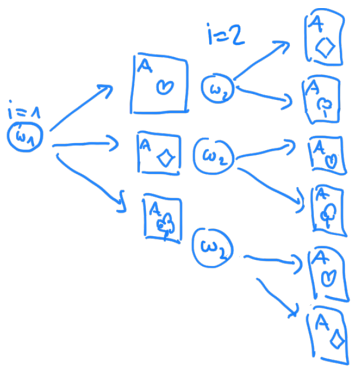
$$\sum_{(\omega_1,\ldots,\omega_{i-1})} |A_{\omega_1,\ldots,\omega_{i-1}}|$$

Möglichkeiten. An einem Beispiel wird klar, was gemeint ist und wie man vorgeht:

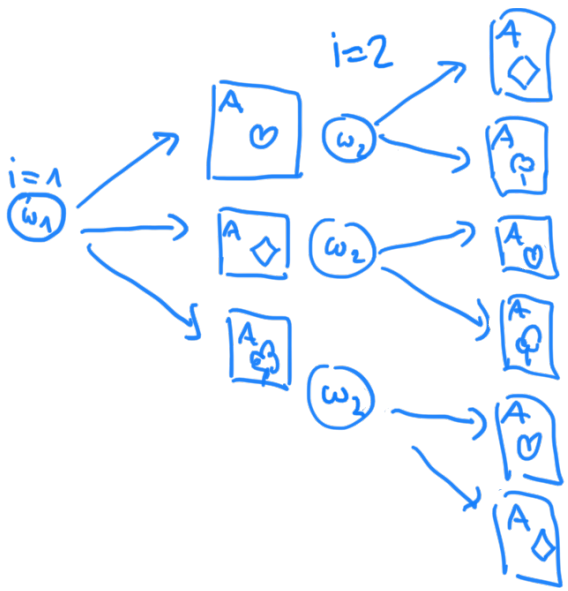
Beispiel: Wähle aus 3 Assen zufällig zwei ohne Zurücklegen aus:
1. Zug: Ziehe aus $A_1 = \Omega = \{Ass_1, Ass_2, Ass_3\}$, also 3 Möglichkeiten.
2. Zug: Es gibt drei Fälle:
Fall 1: $\omega_1 = Ass_1$: Man zieht aus $B_{\omega_1} = B_{Ass_1} = \{Ass_2, Ass_3\}$.
Fall 2: $\omega_1 = Ass_2$: Man zieht aus $B_{\omega_1} = B_{Ass_2} = \{Ass_1, Ass_3\}$.
Fall 3: $\omega_1 = Ass_3$: Man zieht aus $B_{\omega_1} = B_{Ass_3} = \{Ass_1, Ass_2\}$.
In jedem Fall gibt es 2 Möglichkeiten. Insgesamt also $3 \cdot 2 = 6$. Formel:

$$(|B_{Ass_1}| + |B_{Ass_2}| + |B_{Ass_3}|) = \sum_{\omega_1 \in A_1} |B_{\omega_1}|$$

Möglichkeiten. Übersichtliches Abzählen durch Darstellung als Baum:



Laplace-Räume und Kombinatorik



Zwei Urnenmodelle

Urnenmodelle

- Ziehen in Reihenfolge mit Zurücklegen. Sei $A = \{1, \ldots, N\}$.

$$\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_1, \ldots, \omega_n \in A\}$$

$|\Omega| = N^n$.
Als Laplace-Raum: $P(A) = \frac{|A|}{|\Omega|}$ für $A \subset \Omega$

- Ziehen in Reihenfolge ohne Zurücklegen.

$$\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_1, \ldots, \omega_n \in A,\ \omega_i \neq \omega_j,\ i \neq j\}$$

Es gilt: $|\Omega| = N \cdot (N-1) \cdot \ldots \cdot (N-n+1)$.
Als Laplace-Raum: $P(A) = \frac{|A|}{|\Omega|}$ für $A \subset \Omega$.

Zwei Urnenmodelle

Hinweis: Man muss klären, ob das Gleichwahrscheinlichkeitsmodell zutrifft! Das folgt nicht allein schon daraus, dass eine Ergebnismenge mit der eines Urnenmodells übereinstimmt.

$\rightarrow$ Die Urnenmodelle ohne Reihenfolge besprechen wir später.

Beispiele:

a) Max hat 5 Hemden, 2 Krawatten und 1 Anzugshose. Wieviele verschiedene Möglichkeiten hat er, sich für sein Bewerbungsgespräch anzuziehen? $\rightarrow 5 \cdot 2 \cdot 1 = 10$.
(Bei der formalen Modellierung wählt man das Urnenmodell *in Reihenfolge*, da man Hals und Beine nicht vertauschen sollte.)

b) Skat: 32 Karten, 3 Spieler, je 10 Karten, 2 im Skat. Wieviele verschiedene Skatblätter ('unsortierte') gibt es für einen Spieler?
32 Möglichkeiten für 1. Karte, 31 Möglichkeiten für 2. Karte, usw., 23 Möglichkeiten für 10. Karte.
$\rightarrow 32 \cdot 31 \cdot \ldots \cdot (32 - 10 + 1) = 2.34102 \cdot 10^{14}$
Da ein Blatt nicht von den 10! möglichen Anordnungen der Karten auf der Hand abhängt, erhält man die Anzahl verschiedener Blätter durch Division durch 10!.

Beispiele:

c) Das Passwort eines IT-Systems besteht aus einem Buchstaben (A-Z) gefolgt von einer zweistelligen Zahl und einem X oder Z, z.B. $G74X$.
Wie wahrscheinlich ist es, den Geheimcode zu erraten?
Wäre eine dreistellige Zahl sicherer? Wie sicher ist ein 4-stelliger PIN?

Modell:

$$\Omega = \{(\omega_1, \omega_2, \omega_3) | \omega_1 \in \{A, \ldots, Z\}, \omega_2 \in \{0, \ldots, 99\}, \omega_3 \in \{X, Z\}\}$$

Abzählen von $\Omega = \{A, \ldots, Z\} \times \{0, \ldots, 99\} \times \{X, Z\}$:

$$|\Omega| = 26 \cdot 100 \cdot 2 = 5200$$

Die Wahrscheinlichkeit, den Code zu erraten ist $\frac{1}{5200} = 0.0001923077$.
Dreistellige Zahl: $1/1000 = 0.001$. Vierstellige Zahl: $1/10000 = 0.0001$.

Beispiele:

d) Eine faire Münze (Kopf/Zahl gleichwahrscheinlich) wird zweimal geworfen. Ist es wahrscheinlicher, dass beide Male dasselbe erscheint oder einmal Kopf und einmal Zahl?

$\omega_1 \in \{K, Z\}$ Ergebnis des 1. Wurfs, $\omega_2 \in \{K, Z\}$ Ergebnis des 2. Wurfs. $\omega = (\omega_1, \omega_2)$ Ergebnis des Experiments.

Ergebnismenge: $\Omega = \{(\omega_1, \omega_2) | \omega_1, \omega_2 \in \{K, Z\}\}$.

Alle Elementarereignisse $\{(\omega_1, \omega_2)\}$ sind gleichwahrscheinlich.
$\rightarrow |\Omega| = 2 \cdot 2 = 2^2$.

$A = \{(K, K), (Z, Z)\}, B = \{(K, Z), (Z, K)\}$. Es gilt $P(A) = 2/4 = 1/2$ und $P(B) = 2/4 = 1/2$. Also ist keines der Ereignisse wahrscheinlicher als das andere, sondern beide sind gleichwahrscheinlich!

Beispiele:

e) Die Münze aus c) wird nun 6-mal geworfen. Ist das Ergebnis (Z, K, Z, K, Z, K) oder (Z, Z, Z, Z, Z, K) wahrscheinlicher?

Jedes Ergebnis ist beschrieben durch $\omega = (\omega_1, \ldots, \omega_6)$ mit $\omega_1, \ldots, \omega_6 \in \{K, Z\}$.
Hier ist $\Omega = \{(\omega_1, \ldots, \omega_6) | \omega_1, \ldots, \omega_6 \in \{K, Z\}\}$

$\rightarrow |\Omega| = 2^6 = 64$. Es liegt ein Laplace-Experiment vor, da alle Ausgänge gleichwahrscheinlich sind. Daher gilt:

$$P(\{(Z, K, Z, K, Z, K)\}) = \frac{1}{64}, \quad P(\{(Z, Z, Z, Z, Z, K)\}) = \frac{1}{64}$$

Achtung: Beim Ausgang (Z, Z, Z, Z, Z, K) kommt zunächst 5-mal Zahl. Viele Menschen glauben, dass es nun sehr wahrscheinlich ist, dass beim nächsten Mal Kopf kommt. Dies ist allerdings ein Trugschluss: Das Ergebnis (Z, Z, Z, Z, Z, Z) hat aber ebenfalls die Wahrscheinlichkeit $1/64$. Die geworfene Münze landet (aufgrund der Gesetze der Physik) jedesmal zufällig auf der einen oder der anderen Seite, und zwar mit Wkeit $1/2$. Dies gilt ebenso für andere Glückspiele,

Zufällige Speicherbelegung:

k Objekte sollen in einem Speicher mit n Speicherplätzen abgelegt werden. Als Speicheplatz wird jeweils eine zufällige Zahl zwischen 1 und n gewählt. A_{nk} bezeichne das Ereignis, dass eine Kollision auftritt.

Bestimme $P(A_{nk})$ und gebe eine untere Schranke an!

Laplace-Räume und Kombinatorik

Laplace-Räume und Kombinatorik

Das Geburtstagsproblem:
Sie sind auf einer Party mit n Gästen. Wie wahrscheinlich ist es, dass mindestens zwei Gäste am selben Tag Geburtstag haben? Ist man im Vorteil, wenn man darauf wettet?

Die Lösung ist:

$$p_n = 1 - \frac{365 \cdot \ldots \cdot (365 - n + 1)}{365^n}$$

Zahlenbeispiele (gerundet):
$p_5 = 0.027, p_{10} = 0.117, p_{20} = 0.411, p_{23} = 0.507.$

Zwei Urnenmodelle

In Beispielen wie diesen tritt der Ausdruck $\frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}$ auf (setze: $n = N, k = n$).
Es gilt die Abschätzung:

$$\frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{n^k} \leq \exp\left(-\frac{(k-1)k}{2n}\right)$$

Chancen

Alltag: Die Chancen stehen $75 : 25 = 0.75 : 0.25$

Chancen

Die **Chance** (odds) eines Ereignisses A mit Eintrittswahrscheinlichkeit $p = P(A)$ ist definiert durch

$$o = o(A) = \frac{p}{1-p}.$$

Es ist $o(A)$-mal wahrscheinlicher, dass A eintritt, als dass A nicht eintritt. Durch Logarithmieren erhält man die **logarithmierten Chancen** (engl.: *log-odds*):

$$\log(o) = \log(p/(1-p)) = \log(p) - \log(1-p).$$

Symmetrieeigenschaft:

$$\log o(\overline{A}) = \log((1-p)/p) = -\log o(A)$$

Nullpunkt:$\log o(A) = 0$, falls P(A) =1/2.

Chancen-Verhältnis

Odds-Ratio

$$r = \frac{o(A)}{o(B)} = \frac{P(A)/(1-P(A))}{P(B)/(1-P(B))}.$$

Beispiel:

Gruppe 1 mit Risikofaktor, $A =$ "Person aus 1 erkrankt": $o(A) = \frac{0.3}{0.7}$.
Gruppe 2 ohne Risikofaktor: $B =$ "Person aus 2 erkrankt": $o(B) = \frac{0.1}{0.9}$.
Odds-Ratio: $r = \frac{0.3 \cdot 0.9}{0.7 \cdot 0.1} = \frac{0.27}{0.07} \approx 3.856$.

Siebformel (I)

Wir hatten schon die Formel:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Gibt es solch eine Formel auch für 3 Ereignisse, also für $P(A \cup B \cup C)$?

Siebformel (I)

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ - P(B \cap C) + P(A \cap B \cap C).$$

Siebformel (I)

Siebformel (II)

Ganz allgemein: $A_1, A_2, \ldots$

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \mp \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

Hinweis: Dies ist ein klassischer Induktionsbeweis.

Ereignis-Algebra

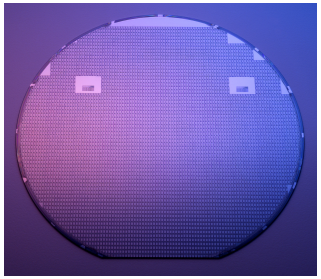
Beispiel

Ergebnismenge für...

1. *zufälligen Gewinn eines Unternehmens:* $\Omega = \mathbb{R}$.
2. *zufälligen Zeitpunkt, an dem die Serverlast zum ersten Mal die Schranke c übersteigt:* $\Omega = [0, \infty)$.

Grundlegender Unterschied:
$\rightarrow$ Ω 'sehr groß' (es gibt überabzählbar viele reelle Zahlen in jedem nichtleeren Intervall).
$\rightarrow$ Es gibt zu viele Teilmengen, so dass man nicht alle als Ereignisse zulassen kann. Ansonsten kein vernünftiger Wahrscheinlichkeitskalkül.
$\rightarrow$ Andere Beschreibung von Wahrscheinlichkeiten (Dichten, Verteilungsfunktionen, später...).

Wafer-Herstellung



- Modelliere Wafer durch Oberfläche Ω
- Staubpartikel landet zufällig an der Stelle $\omega \in \Omega$: CPU defekt.
- (Kleine) Teilfläche $A \subset \Omega$ nutzlos, wenn $\omega \in A$.
- Staubpartikel trifft an einer zufälligen Stelle auf den Wafer. Plausibles Modell:

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

mit $|A| =$ Fläche von A.

Ereignisalgebra

Ereignisalgebra, σ-Algebra

Ein Mengensystem $\mathcal{A} \subset \text{Pot}(\Omega)$ von Teilmengen von Ω heißt **Ereignisalgebra (σ-Algebra)**, wenn die folgenden Eigenschaften gelten:

1. Die Ergebnismenge Ω und die leere Menge $\emptyset$ gehören zu $\mathcal{A}$.
2. Mit A ist auch $\overline{A}$ Element von $\mathcal{A}$.
3. Sind $A_1, A_2, \ldots$ Mengen aus $\mathcal{A}$, dann ist auch $\cup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \ldots$ ein Element von $\mathcal{A}$.

Die Elemente von $\mathcal{A}$ heißen **Ereignisse**.

Bedingte Wahrscheinlichkeiten

Motivation

Wie wahrscheinlich ist es, dass die Übertragung der Buchungsdaten aus Hongkong nicht länger als 20 [s] dauert,...

1. wenn die Übertragung zufällig startet.
2. wenn wir wissen: Die Übertragung erfolgt vormittags.

Frage:

Wie wahrscheinlich ist A gegeben die Information B.

Bedingte Wahrscheinlichkeiten

Beispiel: Expertensystem zur Fehleridentifikation.

Fehlerursachen A führen zu Symptomen B: $A \to B$

Zufallsbehaftete Fehlerursachen $A_1, \ldots, A_K$ (unbeobachtet)

Symptom B (beobachtet)

Beispiel:

Ursachen A_i: Ausfall Bauteil, Kabelbruch, Überspannung, korrodierter Stecker, Überhitzung

Symptome B: Systemausfall, Displayfehler, langsame Datenübertragung, kein WLAN

Relevant: Wie wahrscheinlich ist A_i, wenn B beobachtet wurde?

Bedingte Wahrscheinlichkeiten

Beispiel: Aktienindizes steigen oder fallen zufallsbehaftet. Sie sind beeinflusst von Zinsen.

Daten: 2000 Beobachtungen von Zins und Aktienindex

	Zins fällt (B)	Zins steigt	Summe
Aktienindex fällt (A)	250	950	1200
Aktienindex steigt	750	50	800
Summe	1000	1000	2000

Ablesebeispiel: 250 mal sind Aktienindex und Zins gefallen.

In $\frac{1200}{2000} = 60\%$ der Fälle ist der Aktienindex gesunken.

Relevant: Vorinformation Zins fällt/steigt bekannt.

Wie oft ist der Aktienindex gesunken, wenn der Zins steigt?

Bedingte Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeit

Es seien $A, B \subset \Omega$ Ereignisse mit $P(B) > 0$. Dann heißt

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

bedingte Wahrscheinlichkeit von A **gegeben** B. Liegt speziell ein Laplace-Raum vor, dann ist $P(A|B)$ der Anteil der für das Ereignis $A \cap B$ günstigen Fälle, bezogen auf die möglichen Fälle, welche die Menge B bilden:

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

Bedingte Wahrscheinlichkeiten

Beispiel: Aktienindizes steigen oder fallen zufallsbehaftet. Sie sind beeinflusst von Zinsen.

Daten: 2000 Beobachtungen von Zins und Aktienindex

	Zins fällt	Zins steigt (B)	Summe
Aktienindex fällt (A)	250	950	1200
Aktienindex steigt	750	50	800
Summe	1000	1000	2000

Wahrscheinlichkeit, dass A und B gemeinsam eintreten:

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

Wahrscheinlichkeit, dass B eintritt:

$$P(B) =$$

Wahrscheinlichkeit von A gegeben B (ist eingetreten):

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

Bedingte Wahrscheinlichkeiten

Rechenregel:

A, B seien Ereignisse mit $P(B) > 0$. Dann gilt:

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

Drei Ereignisse A, B, C. Gesucht: Wkeit von C gegeben A und B.
Bedinge auf das Ereignis $A \cap B$ (sofern $P(A \cap B) > 0$):

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

Umstellen:

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

Einsetzen von $P(A \cap B) = P(B|A)P(A)$ (sofern $P(A) > 0$):

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

Beispiel

Beispiel

Ereignisse:

$$A = \text{„Server nicht überlastet“},$$
$$B = \text{„Server antwortet spätestens nach 5 [s]“},$$
$$C = \text{„Download dauert nicht länger als 20 [s]“}.$$

Gesucht: $P(A \cap B \cap C)$.
Gegeben:

1. *Server nicht überlastet mit Wkeit* 0.1.
2. *Wenn Server nicht überlastet, dann Antwort nach spätestens 5 [s] mit Wkeit 0.95.*
3. *In diesem Fall dauert der Download in 8 von 10 Fällen nicht länger als 20 [s].*

Lösung: $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = 0.1 \cdot 0.95 \cdot 0.8 = 0.076$

Rechenregel

Regel

Sind $A, B, C \subset \Omega$ Ereignisse mit $P(A \cap B \cap C) > 0$, dann ist

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

Sind allgemeiner $A_1, \ldots, A_n$ Ereignisse mit $P(A_1 \cap \cdots \cap A_n) > 0$, dann gilt:

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Satz von der totalen Wahrscheinlichkeit

Motivation

Ein Spam-Filter verschiebt E-Mails in den junk-Ordner, wenn gewisse Worte in der E-Mail vorhanden sind, z.B. *win*.
Durch Analysieren von alten E-Mails kann man die bedingten Wahrscheinlichkeiten der Form

$$P(\text{,,E-Mail enthält } \textit{Uni}\text{''} \,|\, \text{,,Email ist Spam''})$$

etc. gut schätzen.
Fragen:

1. Wie wahrscheinlich ist es, dass eine E-Mail Spam ist?
2. Wie groß ist die Wahrscheinlichkeit, dass eine E-Mail tatsächlich Spam ist, wenn das Wort *win* vorkommt?

Spam-Filter

Systematisch: Ereignisse definieren:

$$A = \text{„E-Mail ist Spam“},$$
$$B_1 = \text{„E-Mail enthält das Wort } \textit{Uni}\text{“},$$
$$B_2 = \text{„E-Mail enthält das Wort } \textit{win}\text{“}.$$

Bekannt seien: $P(A), P(B_1|A), P(B_1|\overline{A}), P(B_2|A)$ und $P(B_2|\overline{A})$.

1. Kann man hieraus
$$P(B_i), \qquad i = 1, 2$$
berechnen?
2. Kann man hieraus
$$P(A|B_i)$$
berechnen?

Satz von der totalen Wahrscheinlichkeit

Satz von der totalen Wahrscheinlichkeit

Es sei $A_1, \ldots, A_K$ eine disjunkte Zerlegung von Ω:

$$\Omega = A_1 \cup \cdots \cup A_K, \quad A_i \cap A_j = \emptyset, \ i \neq j.$$

Dann gilt:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_K)P(A_K).$$

In Summenschreibweise:

$$P(B) = \sum_{i=1}^{K} P(B|A_i)P(A_i).$$

Diese Formel gilt auch sinngemäß für $K = \infty$.

Satz von Bayes

Satz von Bayes

Satz von Bayes

$A_1, \ldots, A_K$ sei eine disjunkte Zerlegung von Ω mit $P(A_i) > 0$ für alle $i = 1, \ldots, K$. Dann gilt für jedes Ereignis B mit $P(B) > 0$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{K} P(B|A_k)P(A_k)}.$$

Diese Formel gilt sinngemäß auch für den Fall $K = \infty$.

Satz von Bayes

Satz von Bayes

$A_1, \ldots, A_K$ sei eine disjunkte Zerlegung von Ω mit $P(A_i) > 0$ für alle $i = 1, \ldots, K$. Dann gilt für jedes Ereignis B mit $P(B) > 0$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{K} P(B|A_k)P(A_k)}.$$

Diese Formel gilt sinngemäß auch für den Fall $K = \infty$.

Beispiel: Spam-Filter

Motivation

Ein Spam-Filter verschiebt E-Mails in den junk-Ordner, wenn gewisse Worte in der E-Mail vorhanden sind, z.B. *win*.
Durch Analysieren von alten E-Mails kann man die bedingten Wahrscheinlichkeiten der Form

$$P(\text{,,E-Mail enthält } \textit{Uni}\text{''} \,|\, \text{,,Email ist Spam''})$$

etc. gut schätzen.
Fragen:

1. Wie wahrscheinlich ist es, dass eine E-Mail Spam ist?
2. Wie groß ist die Wahrscheinlichkeit, dass eine E-Mail tatsächlich Spam ist, wenn das Wort *win* vorkommt?

Beispiel: Spam-Filter

Systematisch: Ereignisse definieren:

$$A = \text{„E-Mail ist Spam“},$$
$$B_1 = \text{„E-Mail enthält das Wort } \textit{Uni}\text{“},$$
$$B_2 = \text{„E-Mail enthält das Wort } \textit{win}\text{“}.$$

Bekannt seien: $P(A), P(B_1|A), P(B_1|\overline{A}), P(B_2|A)$ und $P(B_2|\overline{A})$.

1. Kann man hieraus
$$P(B_i), \qquad i = 1, 2$$
berechnen?
2. Kann man hieraus
$$P(A|B_i)$$
berechnen?

Beispiel: Spam-Filter

Mehrstufige Wahrscheinlichkeitsmodelle

Mehrstufige Zufallsexperimente, insbesondere n-malige Wiederholung

$$\Omega = \Omega_1 \times \cdots \times \Omega_n$$

Festlegung der Wahrscheinlichkeiten

$$p(\omega) = P(\{\omega\}) = ?$$

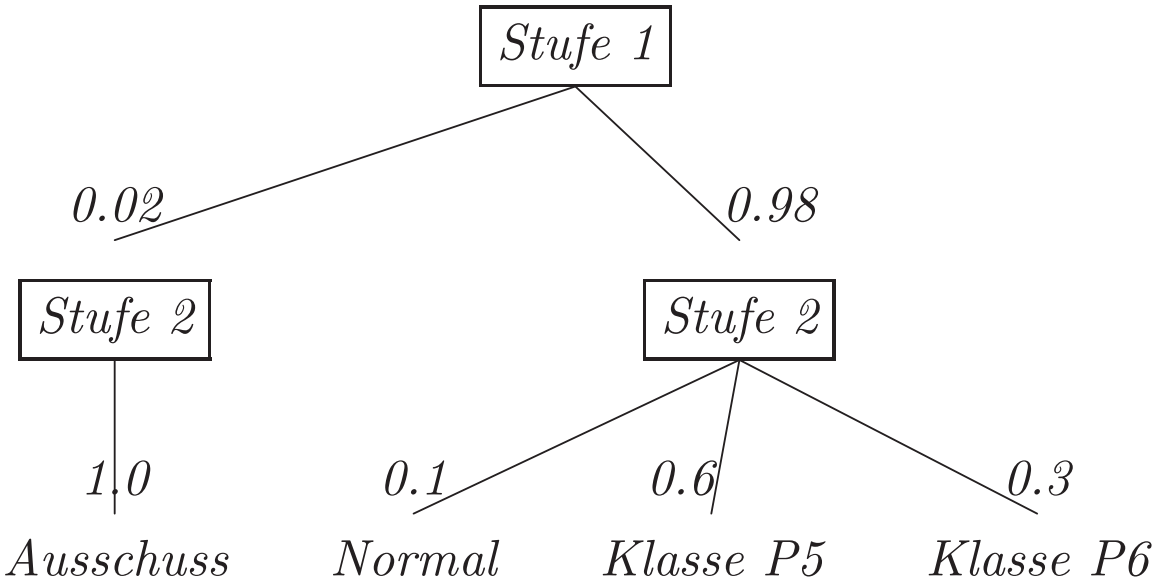
Mehrstufige Wahrscheinlichkeitsmodelle

Fallgestaltung: Produktion von Nadellagern bestehe aus zwei Stufen:

- Stufe 1: Vorbereitende Bearbeitung eines Rohling.
  Mit Wkeit 0.02 genügt ein Rohling nach Stufe 1 nicht den Qualitätsanforderungen.
- Stufe 2: Nachbearbeitung
  Entsprechend der Toleranzen Sortierung in drei Klassen (Normal/P5/P6).

Mit welcher Wkeit erhält man ein Nadellager der Klasse P5?

Mehrstufige Wahrscheinlichkeitsmodelle



Mehrstufige Wahrscheinlichkeitsmodelle

Oft: An verschiedenen Zeitpunkten bestimmen zufällige Ereignisse den Folgezustand.
Darstellung durch **Wahrscheinlichkeitsbaum**.
Modell:

$$\Omega = \Omega_1 \times \cdots \times \Omega_n$$

Startverteilung:

$$p(\omega_1), \qquad \omega_1 \in \Omega_1.$$

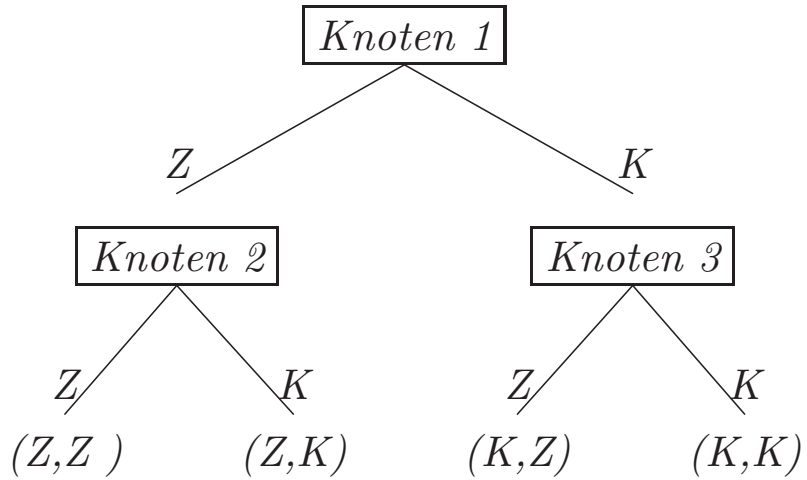
Bedingte Wahrscheinlichkeiten:

$$p(\omega_j | \omega_1, \ldots, \omega_{j-1})$$

Pfadregel für $\omega = (\omega_1, \ldots, \omega_n)$:

$$P(\{\omega\}) = p(\omega_1) p(\omega_2|\omega_1) \cdots p(\omega_n|\omega_1, \ldots, \omega_{n-1})$$

Beispiel 2.3.3 *Eine faire Münze mit Kopf (K) und Zahl (Z) wird zweimal geworfen. Wir können auch dieses Zufallsexperiment als Wahrscheinlichkeitsbaum repräsentieren:*



Unabhängige Ereignisse

Heuristik: B nicht informativ für A, wenn $P(A|B) = P(A)$.

Unabhängige Ereignisse

Zwei Ereignisse A und B heißen **stochastisch unabhängig** (kurz: unabhängig), wenn

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

gilt. Diese Identität wird als **Produktsatz** bezeichnet.

Produktsatz

Produktsatz

k Ereignisse $A_1, \ldots, A_k \subset \Omega$ erfüllen den **Produktsatz**, wenn gilt:

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1) \cdot \ldots \cdot P(A_k).$$

Totale und paarweise Unabhängigkeit

$A_1, \ldots, A_n \subset \Omega$ heißen **(total) stochastisch unabhängig**, wenn für jede Teilauswahl $A_{i_1}, \ldots, A_{i_k}$ von $k \in \mathbb{N}$ Ereignissen der Produktsatz gilt.
$A_1, \ldots, A_n$ heißen **paarweise stochastisch unabhängig**, wenn alle Paare A_i, A_j $(i \neq j)$ stochastisch unabhängig sind.

Totale und paarweise Unabhängigkeit

Totale und paarweise Unabhängigkeit

- $A_1, \ldots, A_n \subset \Omega$ heißen **(total) stochastisch unabhängig**, wenn für jede Teilauswahl $A_{i_1}, \ldots, A_{i_k}$ von $k \in \mathbb{N}$ Ereignissen der Produktsatz gilt.
- $A_1, \ldots, A_n$ heißen **paarweise stochastisch unabhängig**, wenn alle Paare A_i, A_j $(i \neq j)$ stochastisch unabhängig sind.

Ersetzungsregel

Ersetzungsregel

Sind

$$A_1, \ldots, A_n \subset \Omega$$

unabhängig, dann auch

$$B_1, \ldots, B_k, \qquad k \leq n,$$

wobei jedes B_i entweder A_i oder $\overline{A}_i$ ist, für $i = 1, \ldots, k$.

Anwendung zur Unabhängigkeit

Beispiel: Parallelschaltung

- n Datenkabel sind parallel geschaltet. Sie können einzeln genutzt werden und fallen unabhängig voneinander aus.
- Die Übertragung fällt aus, wenn alle Kanäle versagen.
- Die Datenkabel sind durch Stecker verbunden, die unabhängig voneinander ausfallen.

Anwendung zur Unabhängigkeit

• Sei p die Wahrscheinlichkeit, dass ein Datenkabel ausfällt, und A_i das Ereignis, dass das i-te Kabel ausfällt.
Dann sind $A_1, \ldots, A_n$ unabhängig mit $P(A_i) = p$, $i = 1, \ldots, n$.
• Sei B das Ereignis $B =$ „Übertragung fällt aus“. Dann ist

$$B = \bigcap_{i=1}^{n} A_i.$$

• Da $A_1, \ldots, A_n$ unabhängig sind, ergibt sich die Ausfallwahrscheinlichkeit einer Übertragung zu

$$P(B) = P(A_1) \ldots P(A_n) = p^n.$$

• Setzt man beispielsweise vier Datenkabel mit $p = 0.01$ ein, dann erhält man $P(B) = 0.01^4 = 10^{-8}$.

Anwendung zur Unabhängigkeit

Reihenschaltung: Das Datenkabel bestehe aus n Teilkabeln, die mit Steckern verbunden sind. Die Stecker versagen unabhängig voneinander mit Wahrscheinlichkeit q.

• Es bezeichne C_i das Ereignis, dass der i-te Stecker kaputt ist, und D das Ereignis $D =$ „Übertragung fällt aus“. Dann ist

$$D = \bigcup_{i=1}^{n} C_i, \qquad \overline{D} = \bigcap_{i=1}^{n} \overline{C}_i.$$

• Wir erhalten:

$$P(D) = 1 - P(\overline{D}) = 1 - P(\overline{C}_1 \cap \cdots \cap \overline{C}_n).$$

• Da $C_1, \ldots, C_n$ unabhängig sind, sind auch die komplementären Ereignisse $\overline{C}_1, \ldots, \overline{C}_n$ unabhängig. Somit ist:

$$P(\overline{C}_1 \cap \cdots \cap \overline{C}) = (1-q)^n.$$

• Die Übertragung fällt daher mit einer Wkeit von $P(D) = 1 - (1-q)^n$ aus. Für $q = 0.01$ und $n = 10$ ist: $P(D) = 0.0956$.

Zufallsvariablen und ihre Verteilung

Oftmals interessiert nicht $\omega \in \Omega$ (zu fein), sondern

$$x = X(\omega),$$

wobei X eine Abbildung (=Algorithmus) ist: Informationsverdichtung.

Zufallsvariable

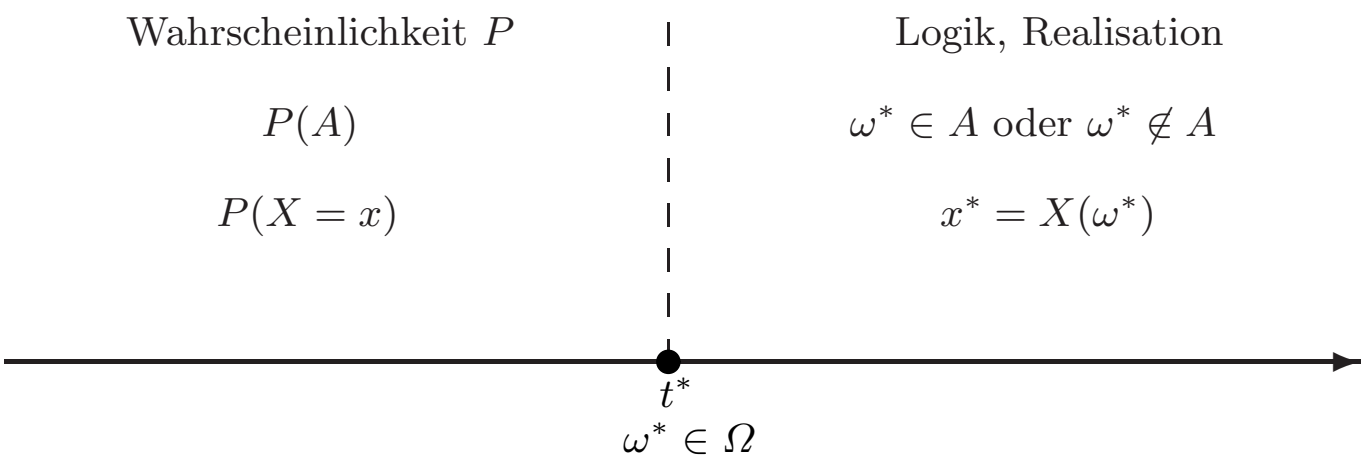
Eine Abbildung

$$X : \Omega \to \mathcal{X} \subset \mathbb{R}, \qquad \omega \mapsto X(\omega),$$

Ω abzählbar, in die reellen Zahlen heißt **Zufallsvariable (mit Werten in $\mathcal{X}$)**.
$x = X(\omega)$: **Realisation**.
Zusatz: Allgemeines Ω: X muss **messbar** sein:

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A} \qquad \text{für alle Ereignisse } B \text{ von } \mathcal{X}.$$

Zufallsvariablen und ihre Verteilung



Diskrete Zufallsvariablen

Ein wichtiger Spezialfall:

Diskrete Zufallsvariable

X heißt **diskrete Zufallsvariable**, wenn

$$\mathcal{X} = \{X(\omega) : \omega \in \Omega\}$$

eine diskrete Menge (endlich oder abzählbar) ist.

Notiz: Ω diskret $\Rightarrow$ Alle ZV sind diskret.

Verteilung einer Zufallsvariablen

Motivation: Glücksspiel-Programm

- Wähle zufällig eine Zahl zwischen 1 und 100.
- Auszahlung:
  - 0 EUR, falls die Zahl kleiner oder gleich 90 ist.
  - 1 EUR, falls die Zahl größer als 90 und kleiner als 100 ist.
  - 2 EUR bei einer 100.

Bestimme die ,,Gewinnverteilung".

Lösung: Der zugrunde liegende Wahrscheinlichkeitsraum ist gegeben durch (Ω, P), mit der Ergebnismenge

$$\Omega = \{\omega : \omega \in \{1, \ldots, 100\}\} = \{1, \ldots, 100\}$$

und dem Wahrscheinlichkeitsmaß $P : Pot(\Omega) \to [0, 1]$,

$$P(\{\omega\}) = \frac{1}{100}, \qquad \text{für alle } \omega \in \Omega$$

Der Gewinn $X : \Omega \to \mathcal{X} \subset \mathbb{R}$, ist eine Abbildung (genauer: Funktion) mit Werten in $\mathcal{X} = \{0, 1, 2\}$ gegeben durch die folgende Tabelle:

ω	$1, \ldots, 90$	$91, \ldots, 99$	100
$x = X(\omega)$	0	1	2
$P(X = x)$	$\frac{90}{100}$	$\frac{9}{100}$	$\frac{1}{100}$

Es gilt:

$$P(X = 1) = P(\{\omega \in \Omega \mid X(\omega) = 1\}) = P(\{91, \ldots, 99\}) = \frac{9}{100}$$

$$P(X = 0) =$$

$$P(X = 2) =$$

Hierdurch ist eine Wahrscheinlichkeitsverteilung auf $\mathcal{X} = \{0, 1, 2\}$ definiert.

Verteilung einer Zufallsvariablen

Sei $A \subset \mathcal{X}$ ein Ereignis.
Das Ereignis, dass sich X in A realisiert, also

$$\{X \in A\} := \{\omega \in \Omega : X(\omega) \in A\},$$

tritt mit Wkeit

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

ein. Betrachte dies als Funktion von A.

Verteilung einer Zufallsvariablen

Verteilung von X

Die Zuordnung, die jedem Ereignis A die Wkeit $P(X \in A)$ zuordnet, heißt **Verteilung von** X.Formal:

$$P_X : A \mapsto P_X(A) = P(X \in A),$$

für Ereignisse $A \subset \mathcal{X}$.

Hinweis: Unterscheide P, das W-Maß auf Ω, und P_X, das W-Maß auf $\mathcal{X}$.

Verteilung einer Zufallsvariablen

Nach Einführung von X interessiert primär die Verteilung von X
Relevant:

- Punktförmige Ereignisse $\{x\}$, $x \in \mathcal{X}$.

$$P_X(\{x\}) = P(X = x)$$

- intervallförmige Ereignisse: $(a, b]$, $a \leq b$

$$P_X((a, b]) = P(X \in (a, b]) = P(a < X \leq b).$$

Berechnung von $P(a < X \leq b)$:
Da $(-\infty, b]$ disjunkt in die Intervalle $(-\infty, a]$ und $(a, b]$ zerlegt werden kann, gilt:

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

Umstellen liefert:

Für die Berechnung von Intervallwahrscheinlichkeiten gilt:

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

Verteilung einer Zufallsvariablen

Beispiel

Für den zufallsbehafteten Gewinn X des nächsten Quartals gelte

- $P(X < 20000) = 0.2$
- $P(X = 20000) = 0.01$
- $P(X \leq 80000) = 0.9$
- $P(X > 100000) = 0$

Berechnen Sie $P(20000 < X \leq 80000)$.

Hinweis: $P(20000 < X \leq 80000) = P(X \leq 80000) - P(X \leq 20000)$
mit

$$P(X \leq 20000) = P(X < 20000) + P(X = 20000) = ...$$

Nun Einsetzen...

Diskrete Zufallsvariablen

X sei diskrete Zufallsvariable mit Werten in $\mathcal{X} = \{x_1, x_2, \dots\} \subset \mathbb{R}$. Dann heißt die Funktion

$$p_X(x) = P(X = x), \qquad x \in \mathbb{R},$$

Wahrscheinlichkeitsfunktion oder **Zähldichte** von X. Es gilt:

$$\sum_{x \in \mathcal{X}} p_X(x) = \sum_{i=1}^{\infty} p_X(x_i) = 1.$$

Sie bestimmt eindeutig die Verteilung von X.

Diskrete Zufallsvariablen (Fs.)

Die Zähldichte kann durch die Punktwahrscheinlichkeiten

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

festgelegt werden: Es gilt $p_X(x_i) = p_i$ und $p_X(x) = 0$, wenn $x \notin \mathcal{X}$. Kann X nur endlich viele Werte $x_1, \ldots, x_k$ annehmen, dann heißt $(p_1, \ldots, p_k)$ auch **Wahrscheinlichkeitsvektor**.

Beispiel

Erinnerung: Glücksspiel-Programm

- Wähle zufällig eine Zahl zwischen 1 und 100.
- Auszahlung:
  - 0 EUR, falls die Zahl kleiner oder gleich 90 ist.
  - 1 EUR, falls die Zahl größer als 90 und kleiner als 100 ist.
  - 2 EUR bei einer 100.

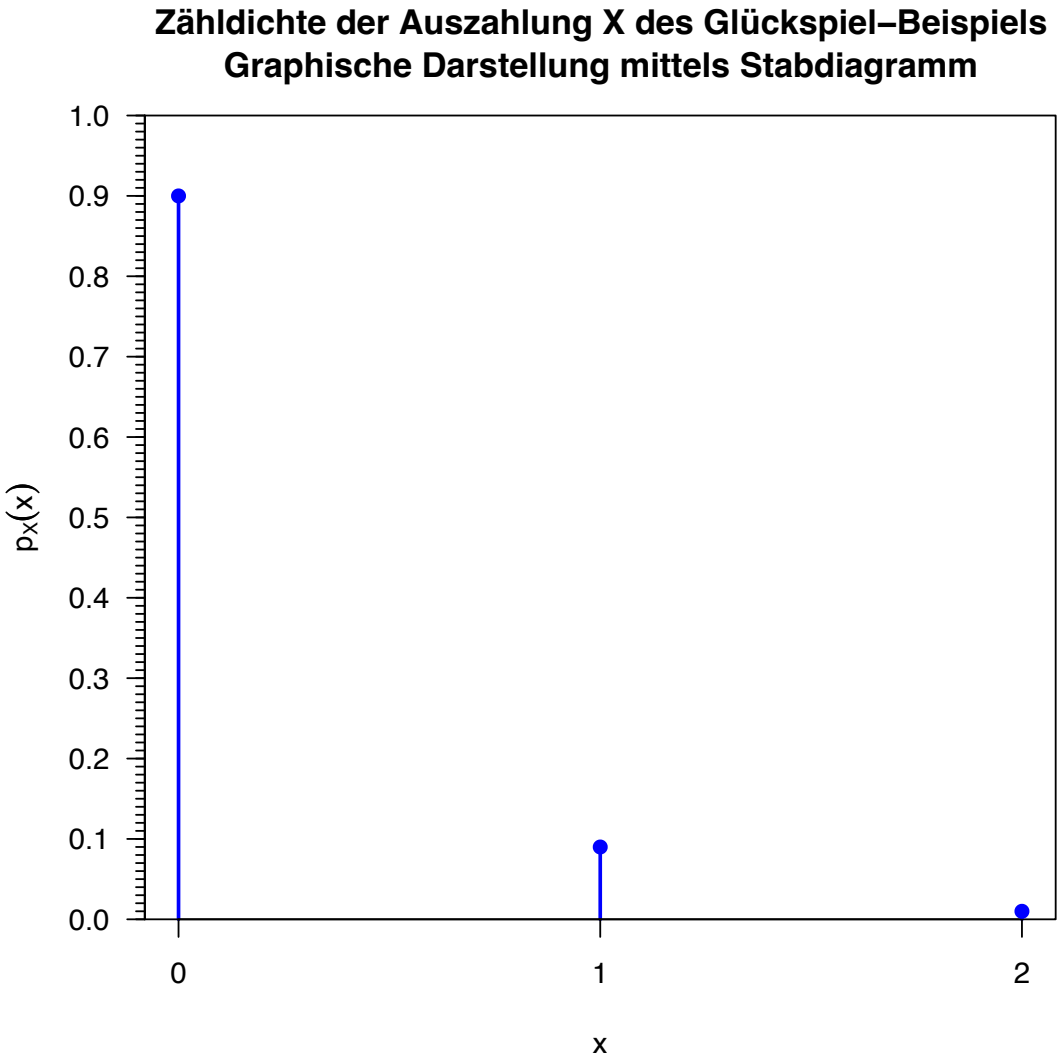
Verteilung des Gewinns X (Tabelle):

x	0	1	2
$P(X = x)$	$\frac{90}{100}$	$\frac{9}{100}$	$\frac{1}{100}$

Zähldichte:

$$p(x) = \begin{cases} 0.9, & x = 0, \\ 0.09, & x = 1, \\ 0.01 & x = 2, \\ 0, & \text{sonst} \end{cases}$$

Beispiele: Diskrete Zufallsvariablen



Beispiele: Diskrete Zufallsvariablen

Beispiele: 1) Überprüfen Sie, ob durch

$$p(x) = \begin{cases} 0.1, & x = -1, \\ 0.8, & x = 0, \\ 0.1, & x = 2, \\ 0, & \text{sonst} \end{cases}$$

eine Zähldichte gegeben ist.

2) Die Zähldichte von X sei gegeben durch

$$p_X(x) = P(X = x) = (1-p)^{x-1}p, \qquad x \in \mathbb{N},$$

für ein $p \in [0,1]$. Für $x \notin \mathbb{N}$ ist $p_X(x) = 0$. Verifiziere, dass hierdurch tatsächlich eine Zähldichte auf $\mathbb{N}$ gegeben ist und leite die Verteilungsfunktion her.

Verteilungsfunktion

Die Funktion $F_X : \mathbb{R} \to [0,1]$,

$$F_X(x) = P(X \leq x), \qquad x \in \mathbb{R},$$

heißt **Verteilungsfunktion von** X. $F_X(x)$ ist monoton wachsend, rechtsstetig und es gilt:

$$F(-\infty) := \lim_{x \to -\infty} F_X(x) = 0, \qquad F(\infty) := \lim_{x \to \infty} F_X(x) = 1.$$

Ferner gilt: $P(X < x) = F(x-) = \lim_{z \uparrow x} F(z)$ und

$$P(X = x) = F(x) - F(x-).$$

Allgemein heißt jede monoton wachsende und rechtsstetige Funktion $F : \mathbb{R} \to [0,1]$ mit $F(-\infty) = 0$ und $F(\infty) = 1$ **Verteilungsfunktion (auf $\mathbb{R}$)** und besitzt obige Eigenschaften.

Beispiel: Die Funktion

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-x}, & x \geq 0, \end{cases}$$

ist eine Verteilungsfunktion, da sie die folgenden Eigenschaften hat:
(1) $0 \leq F(x) \leq 1$ für alle $x \in \mathbb{R}$.

Denn: $0 \leq e^{-x} \leq 1$, $x \geq 0$. *Daher gilt* $F(x) = 1 - e^{-x} \in [0, 1]$ *und somit* $F(x) \geq 0$ *für alle* $x \geq 0$.

Ferner gilt $F(x) = 0 \geq 0$ *für alle* $x < 0$ *nach Definition von* $F(x)$.

(2) $F(-\infty) = \lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} 0 = 0$
(3) $F(\infty) = 1$:

$$\lim_{x \to \infty} F(x) = \lim_{0 \leq x \to \infty} (1 - e^{-x}) = 1 - \lim_{x \to \infty} e^{-x} = 1 - 0 = 1$$

(4) $F(x)$ ist konstant für $x \leq 0$ und streng monton wachsend für $x > 0$:

$$F'(x) = (1 - e^{-x})' = 0 - \frac{d}{dx} e^{-x} = -(-e^{-x}) = e^{-x} > 0$$

Beispiel: Sind $x_1, \ldots, x_n$ reelle Daten (Zahlen, genannt: Stichprobe), dann heißt die Funktion

$$F_n(x) = \frac{\#(x_i \leq x)}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x) = \text{,,Anteil der Daten} \leq x\text{''}$$

empirische Verteilungsfunktion der Stichprobe $x_1, \ldots, x_n$.
(Machen Sie sich eine Zeichnung für $x_1 = -2, x_2 = 1, x_3 = 4$!)

- $F_n(x)$ ist eine Verteilungsfunktion im Sinne der obigen Definition.
- Sind $y_1, \ldots, y_n$ die sortierten x-Werte, dann ist $F_n(x)$ konstant auf den Intervallen $(-\infty, y_1), [y_1, y_2), [y_2, y_3), \ldots, [y_{n-1}, y_n), [y_n, \infty)$ mit Funktionswerten $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$.
- $F_n(x)$ springt an den beobachteten Werten. Die Sprunghöhe ist jeweils der Anteil der jeweiligen Beobachtung in der Stichprobe.

Beispiele: Diskrete Zufallsvariablen

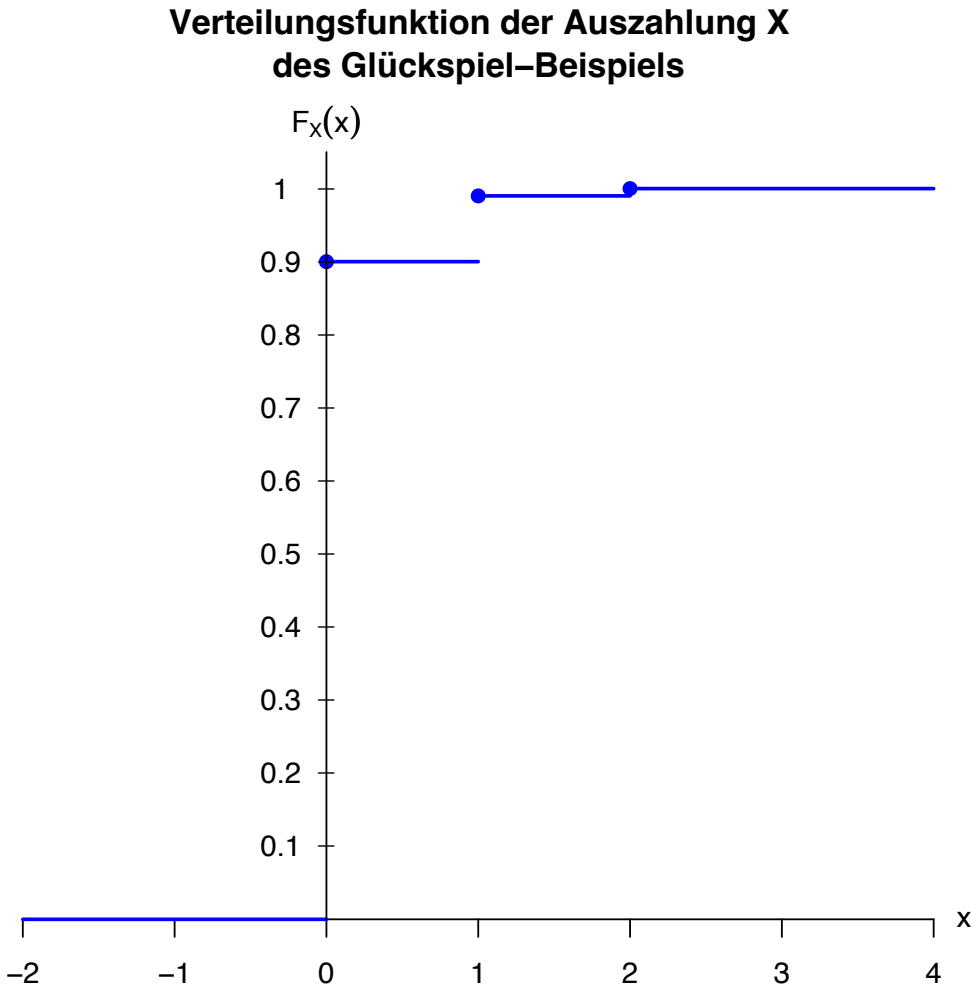
Zähldichte $p(x)$ für $x \in \mathbb{R}$ gegeben durch

$$p(x) = \begin{cases} 0.9, & x = 0, \\ 0.09, & x = 1, \\ 0.01 & x = 2, \\ 0, & \text{sonst} \end{cases}$$

Verteilungsfunktion: (addiere die Zähldichte sukzessive auf...)

$$F(x) = \begin{cases} 0, & x < 0, \\ 0.9, & 0 \leq x < 1, \\ 0.99, & 1 \leq x < 2, \\ 1, & 2 \leq x \end{cases}$$

Beispiele: Diskrete Zufallsvariablen



Beispiel:
Für die Zufallsvariable $X : \{1, 2, 3\} \to \mathbb{R}$ gelte

$$P(X = 1) = 0.1, \quad P(X = 2) = 0.5, \quad P(X = 3) = 0.4.$$

Angabe der Verteilung durch die *Verteilungsfunktion*:

$$F_X(x) = \begin{cases} 0, & x < 1, \\ 0.1, & 1 \leq x < 2, \\ 0.6, & 2 < x \leq 3, \\ 1, & x \geq 3. \end{cases}$$

Sprunghöhen: $0.1, 0.5, 0.4$. Dies sind gerade die Werte der Zähldichte!
Die Sprungstellen sind die x-Werte, an denen die Zähldichte positiv ist.
Also ist die *Zähldichte*:

$$p_X(x) = \begin{cases} 0.1, & x = 1, \\ 0.5, & x = 2, \\ 0.4, & x = 3, \\ 0, & \text{sonst.} \end{cases}$$

Quantilfunktion

Quantilfunktion

$F(x)$ sei eine Verteilungsfunktion.
Die Funktion $F^{-1} : [0, 1] \to \mathbb{R}$,

$$F^{-1}(p) = min\{x \in \mathbb{R} : F(x) \geq p\}, \qquad p \in (0, 1),$$

heißt **Quantilfunktion von** F.

Ist $F(x)$ stetig und streng monoton wachsend, dann ist $F^{-1}(p)$ die Umkehrfunktion von $F(x)$.

Für ein festes p heißt $F^{-1}(p)$ **(theoretisches)** p**-Quantil**.

Anschauliche Beschreibung von Wahrscheinlichkeiten

(Idee: Funktion bestimmt (Intervall-) Wahrscheinlichkeit)

Stetige Zufallsvariable, Dichtefunktion

Eine ZV X heißt **stetig (verteilt)**, wenn es eine integrierbare, nicht-negative Funktion $f(x)$ gibt, so dass für alle Intervalle $(a, b] \subset \mathbb{R}$ gilt:

$$P_X((a, b]) = P(a < X \leq b) = \int_a^b f(x)\, dx.$$

$f_X(x) = f(x)$ heißt dann **Dichtefunktion von** X (kurz: Dichte).
Allgemein heißt jede Funktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) \geq 0, x \in \mathbb{R}, \quad \text{und} \quad \int_{-\infty}^{\infty} f(x)\, dx = 1$$

Dichtefunktion.

Stetige Zufallsvariablen

Punktwahrscheinlichkeiten sind immer 0:

$$P(X = x) = \lim_{\Delta x \downarrow 0} \int_{x}^{x+\Delta x} f(t)\, dt = 0$$

Aufweichen: $P(,,X \approx x'') = P(X \in [x - \Delta x, x + \Delta x])$

$$\int_{x-\Delta x}^{x+\Delta x} f(x)\, dx \approx 2\Delta x \cdot f(x)$$

(Das Integral über kleine Intervalle ist näherungsweise die Recheckfläche).
Also: $P(,,X \approx x'')$ ist proportional zu $f(x)$.
($f(x)$: infinitesimale Wkeit bei x pro x-Einheit).

Stetige Zufallsvariablen

Notation: X hat Dichte $f_X(x)$:

$$X \sim f_X$$

Verteilungsfunktion aus Dichte:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \qquad x \in \mathbb{R}.$$

Dichte aus Verteilungsfunktion:

$$f_X(x) = F_X'(x), \qquad x \in \mathbb{R}.$$

Erster Hauptsatz der Differential- und Integralrechung:
Besitzt die Funktion F auf dem Intervall $[a, b]$ eine stetige (es reicht: Riemann-integrierbare) Ableitung, so ist

$$F(b) = F(a) + \int_a^b F'(x)\,dx$$

und somit

$$\int_a^b F'(x)\,dx = F(b) - F(a)$$

(F ist Stammfunktion des Integranden $f(x) := F'(x)$.)

Zweiter Hauptsatz der Differential- und Integralrechung:
Jede auf $[a, b]$ stetige Funktion besitzt eine Stammfunktion auf $[a, b]$, z.B.:

$$F(x) = \int_a^x f(t)\,dt, \qquad a \leq x \leq b.$$

Man kann diese Ergebnisse auch anwenden, wenn die Voraussetzungen auf einzelnen Intervallen gegeben sind.

Beispiel: Stetige Zufallsvariablen

Beispiel

- *Dichtefunktionen*
- *Dichte - Vf. - Quantilfunktion*

Beispiel: Exponentialverteilung

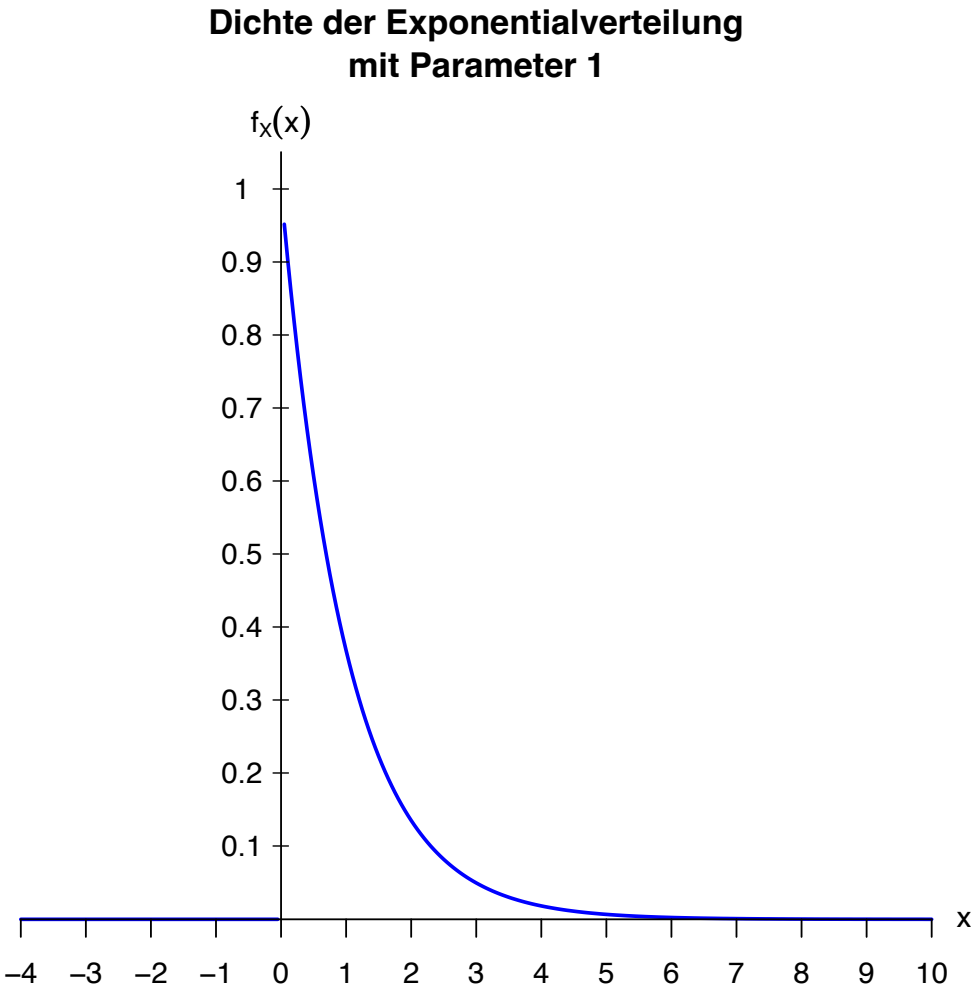
Beispiel: Wir hatten geprüft, dass

$$F(x) = \begin{cases} 1 - e^{-x}, & x \geq 0, \\ 0, & \text{sonst} \end{cases}$$

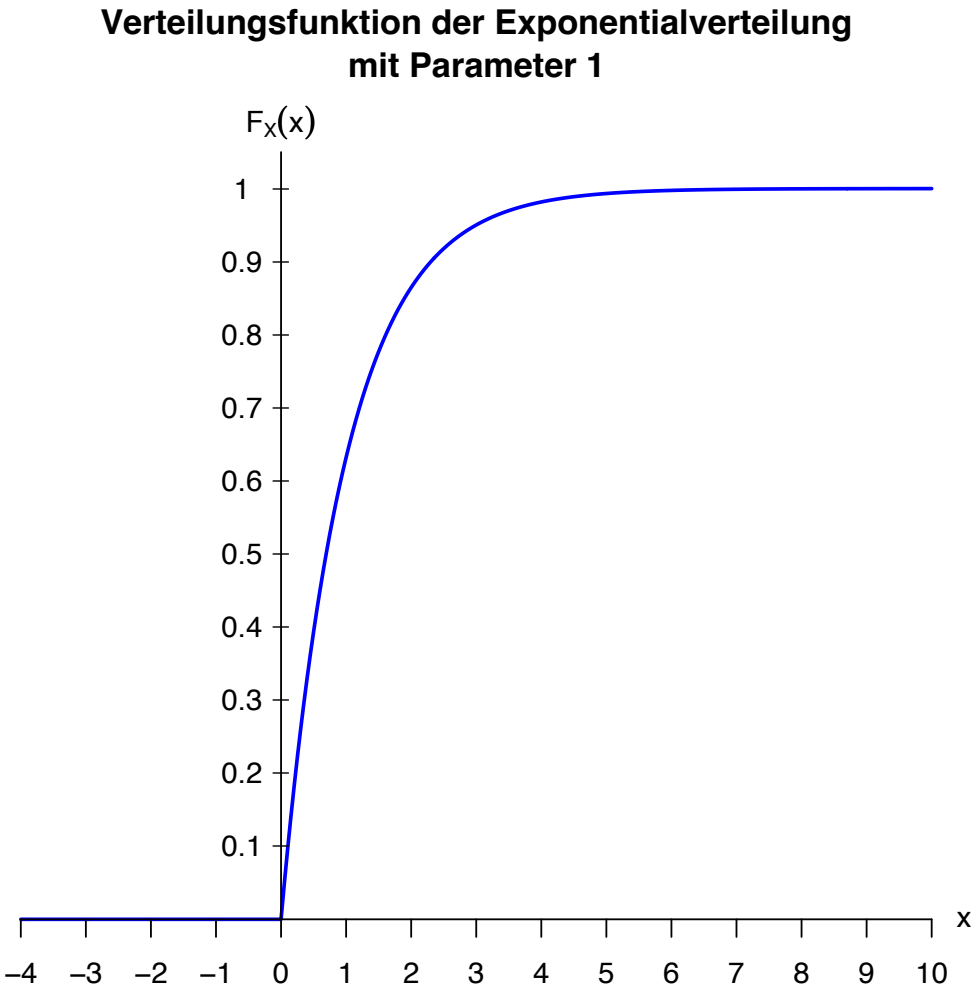
eine auf $\mathbb{R}\backslash\{0\}$ differenzierbare Verteilungsfunktion ist. Die zugehörige Dichte ist gegeben durch 0 auf $(-\infty, 0]$ und durch e^{-x} auf $(0, \infty)$

$$f(x) = F'(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & x \leq 0 \end{cases}$$

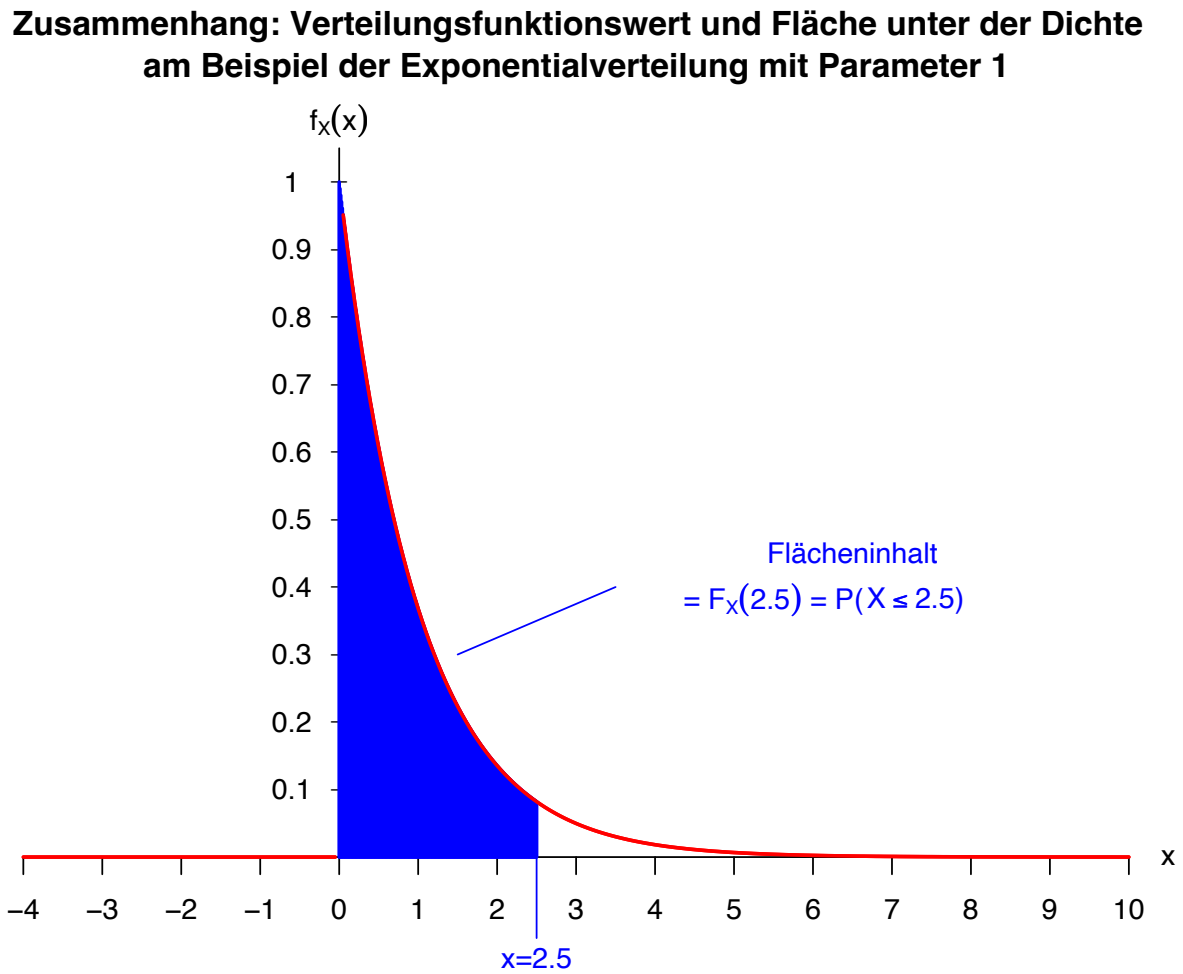
Beispiel: Stetige Zufallsvariablen



Beispiel: Stetige Zufallsvariablen



Beispiel: Stetige Zufallsvariablen



Dichtetransformationssatz

Problem

- Gelte $X \sim f_X$ mit Dichte f_X.
- Relevant: $Y = g(X)$.
- Gesucht: Dichte f_Y von Y.

Dichtetransformationssatz

X sei eine stetige Zufallsvariable mit Werten in $\mathcal{X} = (a, b)$, $a < b$, und mit Dichtefunktion $f_X(x)$.

Weiter sei $y = g(x)$ eine stetig differenzierbare Funktion mit Umkehrfunktion $x = g^{-1}(y)$, so dass $(g^{-1})' \neq 0$ gilt.

Dann hat die Zufallsvariable $Y = g(X)$ die Dichtefunktion

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| .$$

Beispiele: Dichtetransformation

Beispiel

Beispiele zur Dichtetransformation

Unabhängige Zufallsvariablen

Zwei Zufallsvariablen X und Y heißen **stochastisch unabhängig**, wenn die Ereignisse $\{X \in A\}$ und $\{Y \in B\}$ stochastisch unabhängig sind, für alle Ereignisse $A \subset \mathbb{R}$ und $B \subset \mathbb{R}$, d.h.

$$P(X \in A, Y \in B) = P(\{X \in A\} \cap \{Y \in B\}) = P(X \in A) \cdot P(Y \in B)$$

Geg: n Zufallsvariablen $X_1, \ldots, X_n$ mit Werten in Mengen $\mathcal{X}_1, \ldots, \mathcal{X}_n$. $X_1, \ldots, X_n$ heißen (total) **stochastisch unabhängig**, wenn für alle Ereignisse $A_1 \subset \mathcal{X}_1, \ldots, A_n \subset \mathcal{X}_n$ die Ereignisse

$$\{X_1 \in A_1\}, \ldots, \{X_n \in A_n\}$$

stochastisch unabhängig sind. D.h.: Für alle $i_1, \ldots, i_k \in \{1, \ldots, n\}$ gilt:

$$P(X_{i_1} \in A_{i_1}, \ldots, X_{i_k} \in A_{i_k}) = P(X_{i_1} \in A_{i_1}) \cdots P(X_{i_k} \in A_{i_k})$$

Kurz: Stets gilt der Produktsatz für gemeinsame Wahrscheinlichkeiten (d.h. von Schnitten)

Unabhängige Zufallsvariablen

Kriterium für diskrete Zufallsvariablen:
Zwei diskrete Zufallsvariablen X und Y sind stochastisch unabhängig, wenn für alle Realisationen x_i von X und y_j von Y die Ereignisse $\{X = x_i\}$ und $\{Y = y_j\}$ stochastisch unabhängig sind, d.h.

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j).$$

Dann gilt ferner

$$P(X = x_i|Y = y_j) = P(X = x_i), \quad \text{und} \quad P(Y = y_j|X = x_i) = P(Y = y_j).$$

Unabhängige Zufallsvariablen

Beispiel: X, Y seien unabhängig und es gelte:

$$P(X = 1) = 0.1, \quad P(X = 2) = 0.5, \quad P(X = 3) = 0.4$$

$$P(Y = 0) = 0.2, \quad P(Y = 1) = 0.8$$

Gemeinsame Verteilung:

$Y \setminus X$	1	2	3	
0				0.2
1				0.8
	0.1	0.5	0.4	1

Jedes Kästchen ist das Produkt der Randeinträge! Beispielsweise:

$$P(Y = 0, X = 1) = 0.2 \cdot 0.1 = 0.02$$

Unabhängige Zufallsvariablen

Beispiel: X, Y seien unabhängig und es gelte:

$$P(X = 1) = 0.1, \quad P(X = 2) = 0.5, \quad P(X = 3) = 0.4$$

$$P(Y = 0) = 0.2, \quad P(Y = 1) = 0.8$$

Gemeinsame Verteilung:

$Y \setminus X$	1	2	3	
0	0.02	0.1	0.08	0.2
1	0.08	0.4	0.32	0.8
	0.1	0.5	0.4	1

Jedes Kästchen ist das Produkt der Randeinträge!

Unabhängige Zufallsvariablen

Es gelte

$$P(X = k) = \binom{10}{k} 0.2^k \cdot 0.8^{10-k}, \qquad k = 0, \ldots, 10,$$

und

$$P(Y = j) = \frac{e^{-0.2} 0.2^j}{j!}, \qquad j = 0, 1, \ldots$$

Sind X und Y unabhängig, dann folgt für alle $k = 0, \ldots, 10$ und $j \geq 0$

$$P(X = k, Y = j) = \binom{10}{k} 0.2^k \cdot 0.8^{10-k} \frac{e^{-0.2} 0.2^j}{j!}$$

(Formeln multiplizieren!)

Unabhängige Zufallsvariablen

Kriterium für stetige Zufallsvariablen:
Zwei stetige Zufallsvariablen X und Y sind stochastisch unabhängig, wenn für alle Intervalle $(a, b]$ und $(c, d]$ die Ereignisse

$$\{a < X \leq b\} \qquad \text{und} \qquad \{c < Y \leq d\}$$

unabhängig sind, d.h.

$$\begin{aligned} P(a < X \leq b, c < Y \leq d) &= \int_a^b f_X(x)\,dx \int_c^d f_Y(y)\,dy \\ &= \int_a^b \int_c^d f_X(x) f_Y(y)\,dy dx. \end{aligned}$$

Unabhängige Zufallsvariablen

- X, Y seinen unabhängig und **standardnormalverteilt** nach der Dichte $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Es gelte also für $-\infty < a \leq b < \infty$

$$P(a \leq X \leq b) = P(a \leq Y \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

- Dann gilt für $-\infty < a \leq b < \infty$ und $-\infty < a \leq b < \infty$:

$$\begin{aligned} P(a \leq X \leq b, c \leq Y \leq d) &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \cdot \int_c^d \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ &= \int_a^b \int_c^d \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dy \end{aligned}$$

da $e^{-x^2/2} e^{-y^2/2} = e^{-x^2/2-y^2/2} = e^{-(x^2+y^2)/2}$.
Also: $(X, Y) \sim f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$.
(Das Doppelintegral ist das Volumen unter der Funktion $\frac{1}{2\pi} e^{-(x^2+y^2)/2}$ über dem Rechteck $[a, b] \times [c, d]$.)

Zufallsstichprobe

Das Gesamtexperiment sei wie folgt beschrieben:

- n-fache Wiederholung eines Zufallsexperiments beschrieben durch $X : \Omega \to \mathcal{X}$.
- Die Wiederholungen erfolgen unter identischen Bedingungen.
- Die Ergebnisse hängen nicht voneinander ab.

Stochastisches Modell:

- n Zufallsvariablen $X_1, \ldots, X_n : \Omega \to \mathcal{X}$.
- X_i repräsentiert das Ergebnis der i-ten Wiederholung.

Zufallsstichprobe

$X_1, \ldots, X_n$ bilden eine **(einfache) Zufallsstichprobe**, wenn gilt:

1. $X_1, \ldots, X_n$ sind stochastisch unabhängig und
2. $X_1, \ldots, X_n$ sind identisch verteilt, d.h. alle X_i besitzen dieselbe Verteilung:

$$P(X_i \in A) = P(X_1 \in A), \ i = 1, \ldots, n, \quad \text{für alle Ereignisse } A.$$

Sei $F(x) = F_X(x)$ die Verteilungsfunktion der X_i, so schreibt man kurz:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x).$$

i.i.d. (engl.: *independent and identically distributed*) steht hierbei für **unabhängig und identisch verteilt**.

Zufallsstichprobe

Sprechweisen:

- $X_1, \ldots, X_n$ unabhängige Kopien von X.
- $X_1, \ldots, X_n$ i.i.d. oder i.i.d.(F).
- $X_1, \ldots, X_n$ (random) sample.

Erinnerung: Realisation $\mathbf{x} = (x_1, \ldots, x_n)$, wobei

$$x_i = X_i(\omega), \qquad i = 1, \ldots, n.$$

von den Zufallsvariablen $X_1, \ldots, X_n$.
Literatur: Auch $x_1, \ldots, x_n$ wird oftmals als Stichprobe bezeichnet. Achte also auf den Kontext!

Erwartungswert

Naiv: Arithmetisches Mittel $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Frage

Was ist das theoretische Pendant?

Heuristik:
P_n empirisches W-Maß, Masse $\frac{1}{n}$ auf den Datenpunkten $x_1, \ldots, x_n$:

$$P_n(\{x_i\}) = \frac{1}{n}, \qquad i = 1, \ldots, n.$$

Arithmetisches Mittel:

$$\overline{x} = x_1 \cdot \frac{1}{n} + \cdots + x_n \cdot \frac{1}{n}.$$

Die Trägerpunkte x_i werden mit den zugehörigen Wkeiten gewichtet.

Erwartungswert

$X \sim p_X$ diskrete ZV mit Werten in $\mathcal{X}$, verteilt nach der Zähldicht p_X.
Dann heißt die reelle Zahl

$$E(X) = \sum_{x \in \mathcal{X}} x \cdot p_X(x)$$

Erwartungswert von X, sofern $\sum_{x \in \mathcal{X}} |x| p_X(x) < \infty$.
Wichtiger Spezialfall: $\mathcal{X} = \{x_1, \ldots, x_k\}$ endlich. Dann ist

$$E(X) = x_1 \cdot p_X(x_1) + x_2 \cdot p_X(x_2) + \cdots + x_k \cdot p_X(x_k).$$

Erwartungswert

$X \sim f_X$ stetige ZV, verteilt nach der Dichtefunktion $f_X(x)$.
Die reelle Zahl

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

Erwartungswert von X (sofern $\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$).

Bernoulli-Verteilung

Bernoulli-Experiment

A ein Ereignis. Beobachte, ob A eintritt oder nicht:

$$X = \mathbf{1}_A = \begin{cases} 1, & A \text{ tritt ein} \\ 0, & A \text{ tritt nicht ein.} \end{cases}$$

Träger: $\mathcal{X} = \{0, 1\}$ (binär). Verteilung gegeben durch

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

p: Erfolgswahrscheinlichkeit.

$$X \sim \text{Ber}(p), \qquad X \sim \text{Bin}(1, p)$$

Erwartungswert: $E(X) = p$,

Varianz: $\text{Var}(X) = p(1 - p)$,

Erwartungswert: Rechenregeln

Rechenregeln

Seien X, Y ZVen (mit $E|X|, E|Y| < \infty$) und $a, b \in \mathbb{R}$.

1. $E(X + Y) = E(X) + E(Y)$,
2. $E(aX + b) = aE(X) + b$,
3. $E|X + Y| \leq E|X| + E|Y|$.
4. **Jensen-Ungleichung**: Ist $g(x)$ konvex, dann gilt: $E(g(X)) \geq g(E(X))$ und $E(g(X)) > g(E(X))$, falls $g(x)$ strikt konvex ist. Ist $g(x)$ konkav bzw. strikt konkav, dann kehren sich die Ungleichheitszeichen um.

Erwartungswert: Rechenregeln

Produkteigenschaft

Seien X, Y stochastisch unabhängige ZVen.
Für alle Funktionen $f(x)$ und $g(y)$ (mit $E|f(X)| < \infty$ und $E|g(Y)| < \infty$) gilt:

$$E(f(X)g(Y)) = E(f(X)) \cdot E(g(Y)).$$

Insbesondere: $E(XY) = E(X) \cdot E(Y)$.

Notiz

X, Y unabhängig $\Rightarrow E(XY) - E(X)E(Y) = 0$.

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

ist ein gängiges Maß für Abhängigkeit (später mehr dazu...)

Erwartungswert: Beispiele

Beispiele zu den Rechenregeln.

- X sei ZV mit $P(X = 1) = p$ und $P(X = 0) = 1 - p$, $p \in [0, 1]$. X_1, X_2 unabhängig mit derselben Verteilung wie X.
  (a) $E(X_1 X_2) = ?$
  (b) $E\left((X_1 - p)X_2\right) = ?$
  (c) $E(3X_1 + X_2^2) = ?$

Erwartungswert bzgl. des empirischen Maßes

Erwartungswert bzgl. des empirischen Maßes P_n (Deskriptive Statistik):
Sei $X \sim P_n$, d.h. $P(X = x_i) = \frac{1}{n}, \qquad i = 1, \ldots, n.$
Dann gilt:

$$E_{P_n} f(X) = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

Für $f(x) = (x - \mu)^2$ erhält man $\mu = E_{P_n}(X) = \overline{x}$:

1. $E_{P_n} f(X)$
2. $= E_{P_n}(X - E_{P_n}(X))^2$
3. $= E_{P_n}(X - \overline{x})^2$
4. $= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$

Dies ist die Stichprobenvarianz aus der deskriptiven Statistik!
Stichprobenvarianz: Unter P_n erwartete quadratische Abweichung von $\overline{x}$.

Varianz, Standardabweichung

Varianz

Sei X eine Zufallsvariable. Dann heißt

$$\sigma_X^2 = \mathrm{Var}(X) = E\big((X - E(X))^2\big)$$

Varianz von X, sofern $E(X^2) < \infty$. Die Wurzel aus der Varianz,

$$\sigma_X = \sqrt{\mathrm{Var}(X)},$$

heißt **Standardabweichung von** X.

Erlaubte Schreibweisen: Mit $\mu = E(X) = EX$.
$\mathrm{Var}(X) = E\big((X - \mu)^2\big) = E(X - \mu)^2$
(Tipp: Lieber mehr Klammern und $E(X)$ sowie $E((X - \mu)^2)$ schreiben!)

Verschiebungssatz

Verschiebungssatz

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

Varianz: Rechenregeln

Rechenregeln

X, Y Zufallsvariablen mit existierenden Varianzen und $a \in \mathbb{R}$.

1. $\text{Var}(aX) = a^2\text{Var}(X)$.
2. Falls $E(X) = 0$, dann gilt: $\text{Var}(X) = E(X^2)$.
3. Sind X und Y stochastisch unabhängig, dann gilt:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Varianz, Standardabweichung

Varianz

Sei X eine Zufallsvariable. Dann heißt

$$\sigma_X^2 = \mathrm{Var}(X) = E\big((X - E(X))^2\big)$$

Varianz von X, sofern $E(X^2) < \infty$. Die Wurzel aus der Varianz,

$$\sigma_X = \sqrt{\mathrm{Var}(X)},$$

heißt **Standardabweichung von** X.

Erlaubte Schreibweisen: Mit $\mu = E(X) = EX$.
$\mathrm{Var}(X) = E\big((X - \mu)^2\big) = E(X - \mu)^2$
(Tipp: Lieber mehr Klammern und $E(X)$ sowie $E((X - \mu)^2)$ schreiben!)

Verschiebungssatz

Verschiebungssatz

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

Varianz: Rechenregeln

Rechenregeln

X, Y Zufallsvariablen mit existierenden Varianzen und $a \in \mathbb{R}$.

1. $\text{Var}(aX) = a^2\text{Var}(X)$.
2. Falls $E(X) = 0$, dann gilt: $\text{Var}(X) = E(X^2)$.
3. Sind X und Y stochastisch unabhängig, dann gilt:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Erwartungswert und Varianz: Beispiele

Erwartungswert und Varianz: Beispiele

Erwartungswert: Transformationsformel

Transformationssatz

X ZV und $g : \mathcal{X} \to \mathcal{Y}$ eine Funktion mit $E|g(X)| < \infty$.
Sei $Y = g(X)$.
Zusammenhang zwischen $E(g(X))$ und $E(Y)$:

1. Sind X und $Y = g(X)$ diskrete Zufallsvariablen mit Wahrscheinlichkeitsfunktionen $p_X(x)$ bzw. $p_Y(y)$, dann gilt:

$$E(Y) = \sum_{x \in \mathcal{X}} g(x) p_X(x) = \sum_{y \in \mathcal{Y}} y \cdot p_Y(y).$$

2. Sind X und $Y = g(X)$ stetig, mit den Dichtefunktionen $f_X(x)$ bzw. $f_Y(y)$, dann gilt:

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx = \int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy.$$

Momente, Kurtosis, Exzess

Sei X Zufallsvariable und $k \in \mathbb{N}$. Gelte $E(|X^k|) < \infty$.

- $m_k = E(X^k)$ ist das k-**te Moment** von X.
- $m_k^* = E(|X|^k)$ ist das k-**te absolute Moment** von X.
- $m_k(a) = E(X-a)^k$ ist das k-**te Moment um** a.
- $m_k^*(a) = E(|X-a|^k)$ ist das k-**te absolute Moment um** a. Für $a = E(X)$ spricht man vom k-**ten zentralen absoluten Moment**.

Beachte:

a) $E(X) = m_1$ ist das erste Moment von X.

b) $\mathrm{Var}(X) = m_2^*(E(X))$ ist das zweite zentrale absolute Moment.

c) $X^* = \frac{X - E(X)}{\sqrt{\mathrm{Var}(X)}}$ heißt Standardisierung von X: $E(X^*) = 0$, $\mathrm{Var}(X^*) = 1$.

d) $\beta_2 = E((X^*)^4)$ heißt Kurtosis von X (misst Wölbung).

e) $X \sim N(\mu, \sigma^2) \Rightarrow \beta_2 = 3$.

f) $\gamma_2 = \beta_2 - 3$ heißt Exzess. $\gamma_2 > 0$: Verteilung 'spitzer' als Gaussverteilung, $\gamma_2 < 0$: 'flacher'.

Standard-Verteilungsmodelle

- Anzahl von Erfolgen (ja/nein, gut/schlecht,...) bei n Wiederholungen (Versuchen).
- Wartezeit auf den ersten Erfolg.
- Anzahl von Ereignissen in einem (Zeit-) Intervall.
- Wartezeit auf das erste Eintreten.
- Stetige Gleichverteilung.
- Normalverteilung

Bernoulli-Verteilung

Bernoulli-Experiment

A ein Ereignis. Beobachte, ob A eintritt oder nicht:

$$X = \mathbf{1}_A = \begin{cases} 1, & A \text{ tritt ein} \\ 0, & A \text{ tritt nicht ein.} \end{cases}$$

Träger: $\mathcal{X} = \{0, 1\}$ (binär). Verteilung gegeben durch

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

p: Erfolgswahrscheinlichkeit.

$$X \sim \text{Ber}(p), \qquad X \sim \text{Bin}(1, p)$$

Erwartungswert: $E(X) = p$,

Varianz: $\text{Var}(X) = p(1 - p)$,

Binomialverteilung (I)

Beispiele

- Anzahl der gesetzten Bits in einer zufälligen Bitfolge der Länge n.

  ```
  0 0 1 1 1 0 1 1 0 1    ->  6
  ```

- Umfrage unter n Studierenden: Zähle aus, wieviele mit der Mensa zufrieden sind.
- Anzahl der blauen Autos auf einem Parkplatz.
- Anzahl der erfolgreichen Geschäftsabschlüsse eines Vertreters.
- Anzahl von Überschreitungen einer Benchmark durch ein Aktienkurs
- ...

Binomialverteilung (II)

Binomialverteilung

- Modell: $X_1, \ldots, X_n$ i.i.d. $\sim \text{Ber}(p)$.
- Anzahl der Erfolge gegeben durch:

$$Y = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

- Wie ist Y verteilt?

Binomialverteilung (III)

Gesucht: $P(Y = k), \quad k = 0, 1, \ldots, n.$

- Das Ereignis $\{Y = k\}$ tritt genau dann ein, wenn exakt k Einsen beobachtet werden.
- Beispiele: a) $k = 2, n = 3$; b) $k = 2, n = 4$; c) $k = 4, n = 8$:

a)	b)	c)
1 1 0	1 1 0 0	1 1 1 1 0 0 0 0
1 0 1	1 0 1 0	1 1 1 0 1 0 0 0
0 1 1	1 0 0 1	1 1 1 0 0 1 0 0
	0 1 1 0	...
	0 1 0 1	
	0 0 1 1	0 0 0 0 1 1 1 1
3	6	28

Binomialverteilung (III)

Gesucht: $P(Y = k), \quad k = 0, 1, \ldots, n.$

- Das Ereignis $\{Y = k\}$ tritt genau dann ein, wenn exakt k Einsen beobachtet werden.
- Beispiele: $k = 2$ und $n = 3, n = 4$ sowie $k = 4, n = 8$:

  1 1 0 1 1 0 0 1 1 1 1 0 0 0 0

- Jede Kombination mit k Einsen hat Wkeit $p^k(1-p)^{n-k}$, denn

$$\begin{aligned} P('11110000') &= P(X_1 = 1, \ldots, X_4 = 1, X_5 = 0, \ldots, X_8 = 0) \\ &= p^4(1-p)^4, \ldots \end{aligned}$$

- Wieviele verschiedene Muster gibt es?
  Muster: k Positionen auswählen und eine '1' hinschreiben.

Binomialkoeffizient

Binomialkoeffizient

Für $n \in \mathbb{N}$ und $k \in \{0, \ldots, n\}$ gibt der **Binomialkoeffizient**

$$\binom{n}{k} = \frac{n \cdot (n-1) \ldots (n-k+1)}{k \cdot (k-1) \ldots 2 \cdot 1} = \frac{n!}{k!(n-k)!}$$

die Anzahl der Möglichkeiten an, aus einer n-elementigen Obermenge (aus n Objekten) eine k-elementige Teilmenge (k Objekte ohne Zurücklegen und ohne Berücksichtigung der Reihenfolge) auszuwählen.

Binomialkoeffizienten

Beispiel:

$$\binom{6}{3} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1) \cdot (3 \cdot 2 \cdot 1)} = \frac{120}{6} = 20.$$

Regel von Pascal:

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

für $k = 1, \ldots, n$ und $n \geq 1$, wobei $\binom{m}{m} = 1 = \binom{m}{0}$ für $m \geq 0$.

$$\binom{0}{0} = 1$$

$$\binom{1}{0} = 1 \quad \binom{1}{1} = 1$$

$$\binom{2}{0} = 1 \quad \binom{2}{1} = 2 \quad \binom{2}{2} = 1$$

$$\binom{3}{0} = 1 \quad \binom{3}{1} = 3 = \quad \binom{3}{2} = 3 \quad \binom{3}{3} = 1$$

Binomialkoeffizient

Binomialverteilung

Y heißt binomialverteilt, $Y \sim \text{Bin}(n, p)$, wenn

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n.$$

Erwartungswert: $E(Y) = np$,

Varianz: $\text{Var}(Y) = np(1-p)$,

Zähldichte: $p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, \ldots, n\}.$

Binomialverteilung: Faltung

Eigenschaft

$X \sim \text{Bin}(n_1, p)$ und $Y \sim \text{Bin}(n_2, p)$ unabhängig, dann folgt:
$X + Y \sim \text{Bin}(n_1 + n_2, p)$.

Urnenmodell III

Urnenmodell III: Ziehen ohne Reihenfolge und ohne Zurücklegen

- Urne mit n Kugeln mit Nummern 1 bis n
- Ziehe k Kugeln ohne Zurücklegen.

In Reihenfolge: $\rightarrow$ k-Tupel $(\omega_1, \ldots, \omega_k)$ mit Einträgen $\omega_i \in \{1, \ldots, n\}$, wobei zusätzlich gilt:

$$\omega_i \neq \omega_j, \qquad i \neq j,$$

Anzahl: $n_k = n(n-1) \cdot \cdots \cdot (n-k+1) = \frac{n!}{(n-k)!}$

Jetzt: Ohne Reihenfolge.

Fasse $\omega_1, \ldots, \omega_k$ so zusammen, dass die Anordnung keine Rolle spielt:

$\rightarrow$ Mengen

Urnenmodell III

Ergebnismenge:

$$\Omega_{III} = \{\{\omega_1, \ldots, \omega_k\} : \omega_1, \ldots, \omega_k \in \{1, \ldots, n\}, \omega_i \neq \omega_k, (i \neq j)\}.$$

Wieviele k-Tupel werden auf diese Weise derselben Menge zugeordnet?
$\rightarrow$ Genau $k!$ Tupel, da jede Permutation der k Elemente $\omega_1, \ldots, \omega_k$ zu derselben Menge führt.
Also hat Ω nicht $\frac{n!}{(n-k)!}$ Elemente, sondern nur

$$|\Omega_{III}| = \frac{1}{k!} \frac{n!}{(n-k)!} = \binom{n}{k}$$

Urnenmodell IV

Urnenmodell IV: Ziehen ohne Reihenfolge und mit Zurücklegen

- Urne mit N Kugeln mit Nummern 1 bis N
- Ziehe n Kugeln ohne Reihenfolge und mit Zurücklegen $\rightarrow$ Mehrfachziehungen möglich

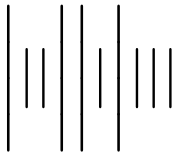
Ergebnismenge: Stelle die Ziehungen als sortierte Tupel dar.

$$\Omega_{IV} = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{1, \ldots, N\}, i = 1, \ldots, n, \omega_1 \leq \cdots \leq \omega_n\}.$$

Urnenmodell IV

Alltag: Strichliste

- Grenze N Felder für die Zahlen $1, \ldots, N$ durch $N-1$ große Striche ab.
- Markiere die gezogenen Kugeln durch kleine Striche in den Feldern.



Jede Stichprobe ist durch $N-1$ große und n kleine Striche charakterisiert. $\rightarrow N-1+n$ Objekte. Es gibt genau $\binom{N-1+n}{n}$ Möglichkeiten, von diesen n auszuwählen und als kleine Striche festzulegen. Die anderen $N-1$ werden die großen Striche. Daher folgt:

$$|\Omega_{IV}| = \binom{N-1+n}{n}$$

Hypergeometrische Verteilung

Problem: Lieferung von N CPUs für Computerproduktion. Ziehe zufällig n aus und teste auf gut/schlecht.

Frage: Wie ist der Anteil X der schlechten Teile in der Stichprobe verteilt?

$\rightarrow$ Urnenmodell III mit $N = R + B$, R: rote Kugeln (schlechte CPUs), B: blaue Kugeln (gute CPUs). (Rot: $K_R = \{1, ..., R\}$. Blau: $\{R + 1, \ldots, N\}$).

Anteil der roten Kugeln in der Urne: $p = \frac{R}{N}$.

Stichprobe vom Umfang $n = r + b$, r: gezogene rote, b: gezogene blaue Kugeln. Ereignis, genau r rote Kugeln zu ziehen: Mit $K_R = \{1, \ldots, R\}$:
$A_r = \{\omega \in \Omega : \exists I = \{i_1, \ldots, i_r\} : \omega_{i_j} \in K_R, 1 \leq j \leq r, \omega_k \notin K_R, \forall k \in I^c\}$

$$P(X = r) = P(A_r) = \frac{\binom{R}{r}\binom{B}{n-r}}{\binom{N}{n}}, \qquad \max(0, n - B) \leq r \leq \min(R, n)$$

X heiß **hypergeometrisch** verteilt mit Par. N, R, B, n.

Geometrische Verteilung

Fragestellungen aus der Praxis:
Eine Software wird täglich mit neuen zufälligen Input-Daten eingesetzt.
Bei einer Fließbandfertigung (z.B. Autoproduktion) wird jedes Produkt einem Endtest unterworfen.

1. Wie lange dauert es im Mittel, bis ein Fehler auftritt?
2. Mit welcher Wahrscheinlichkeit tritt frühestens nach 14 Tagen ein Fehler auf?

Bei Glücksspielen:

1. Wie oft muss man (im Mittel) spielen, bis man gewinnt?
2. Wie wahrscheinlich ist es, dass man mindestens 20 mal spielen muss, um 4 mal zu gewinnen?

Geometrische Verteilung

Beobachte unendliche Bernoulli-Folge (Bitfolge)

0 0 0 0 0 0 0 1 0 0 1 1 0 0 1 . . .

Frage: Verteilung des Auftretens der ersten '1'?
Modell:

$$X_1 X_2 X_3 \cdots$$

wobei

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} \text{Ber}(p)$$

mit $p = P(X_i = 1)$, $i = 1, 2, \ldots$
Zufallsvariable:

$$T = \min\{k \in \mathbb{N} : X_k = 1\}$$

zufälliger Index (Zeitpunkt) der ersten '1'.
zugehörige Wartezeit ist: $W = T - 1$.

Geometrische Verteilung

Beobachte unendliche Bernoulli-Folge (Bitfolge)

0 0 0 0 0 0 0 1 0 0 1 1 0 0 1 . . .

Frage: Verteilung des Auftretens der ersten '1'?
Ereignis

$$\{T = 2\} = \{X_1 = 0, X_2 = 1\}$$
$$\{T = 3\} = \{X_1 = 0, X_2 = 0, X_3 = 1\}$$
$$\vdots$$
$$\{T = n\} = \{X_1 = 0, \ldots, X_{n-1} = 0, X_n = 1\}$$

$$P(T = n) = P(X_1 = 0) \cdot \cdots \cdot P(X_{n-1} = 0) \cdot P(X_n = 1) = (1 - p)^{n-1} p$$

Geometrische Verteilung

T heißt **geometrisch verteilt** mit Parameter $p \in (0, 1]$. Notation: $T \sim \text{Geo}(p)$.

$$P(W = n) = p(1-p)^n, \quad n = 0, 1, \ldots$$
$$P(T = n) = p(1-p)^{n-1}, \quad n = 1, 2, \ldots$$

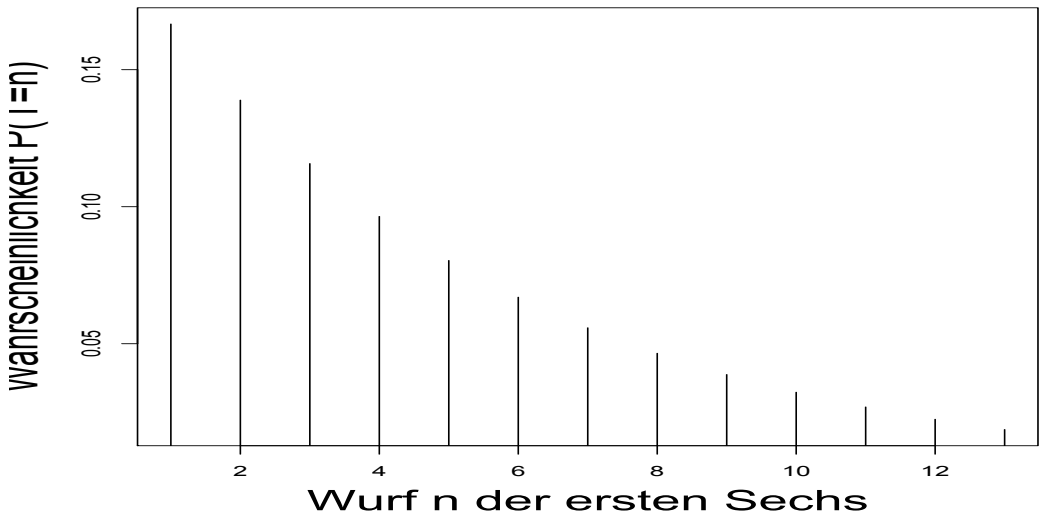
Erwartungswerte: $E(T) = \frac{1}{p}$, $E(W) = \frac{1}{p} - 1$,

Varianzen: $\text{Var}(T) = \frac{1-p}{p^2}$, $\text{Var}(W) = \frac{1-p}{p^2}$.

Geometrische Verteilung

Beispiel: Nach wieviel Würfen kommt im Mittel eine 6?
T: Nummer des ersten Wurfes einer 6. $T \sim \text{Geo}(p)$.

$$E(T) = \frac{1}{p} = \frac{1}{1/6} = 6, \quad \sigma_T = \sqrt{\text{Var}(T)} = \sqrt{\frac{1-p}{p^2}} = \sqrt{30} \approx 5.48$$



$$P(T=1) = p(1-p)^0 = \frac{1}{6} \approx 0.167, \quad P(T=2) = (1-p)p = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36} \approx 0.139$$

Anwendung: Negative Binomialverteilung

Negative Binomialverteilung

Die Verteilung der Summe

$$S = T_1 + \cdots + T_k$$

von k i.i.d. Geo(p)-verteilten ZVen $T_1, \ldots, T_k$ heißt **negativ binomialverteilt**. S_k ist die Anzahl der erforderlichen Versuche, um k Erfolge zu beobachten.

Poisson-Verteilung

- Eingehende Anrufe in einer Notrufzentrale.
- Anzahl der Zeitpunkte, an denen ein Aktienkurs eine Schranke passiert.
- Anzahl von Einbrüchen in einer Zeitperiode.
- Schadstellen in einer elektrischen Leitung.
- Störungen der Internetverbindung.
- Versicherungsfälle.
- Messung radioaktiver (Partikel-) Strahlung.
- ...

Benötigt: Verteilung einer Anzahl von Ereignissen,

- die unabhängig voneinander eintreten,
- deren Eintretenswahrscheinlichkeit nicht von der Zeit abhängt,
- die punktförmig/selten sind.

Poisson-Verteilung

Poisson-Verteilung

Zähle punktförmige Ereignisse in einem Zeitintervall $[0, T]$.Indikatoren:

$$X_t = \begin{cases} 1, & \text{Ereignis zur Zeit } t, \\ 0, & \text{kein Ereignis zur Zeit } t. \end{cases}$$

Annahme: Die X_t sind unabhängig und identisch verteilt.

Anschauung: Ist $I \subset [0, T]$ ein (infinitesimal) kleines Intervall, dann hängt

$$P(\text{"Ereignis in } I\text{"}) = P(X_t = 1, \text{für ein } t \in I)$$

nur von der *Länge*, nicht jedoch von der *Lage* des Intervalls I ab.
$\rightarrow$ Wkeit p, dass ein Ereignis in $[0, T]$ eintritt proportional zu T: $p = \lambda T$

Poisson-Verteilung

Zerlege $[0, T]$ in n gleichbreite Teilintervalle.

$$X_{ni} = \begin{cases} 1, & \text{Ereignis im } i\text{-ten Teilintervall,} \\ 0, & \text{kein Ereignis im } i\text{-ten Teilintervall,} \end{cases}$$

Dann gilt: $X_{n1}, \ldots, X_{nn}$ i.i.d. $\text{Bin}(1, p_n)$ mit

$$p_n = \lambda \cdot \frac{T}{n}$$

λ: Proportionalitätskonstante.
Anzahl:

$$Y = X_{n1} + \cdots + X_{nn} \sim \text{Bin}(n, p_n)$$

Poisson-Grenzwertsatz

Poisson-Grenzwertsatz

Sind $Y_n \sim \text{Bin}(n, p_n)$, $n = 1, 2, \ldots$, binomialverteilte Zufallsvariablen mit $np_n \to \lambda$, $n \to \infty$, dann gilt für festes k:

$$\lim_{n\to\infty} P(Y_n = k) = p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Die Zahlen $p_\lambda(k)$, $k \in \mathbb{N}_0$, definieren eine Verteilung auf $\mathbb{N}_0$.

Y heißt **poissonverteilt** mit Parameter λ. Notation: $Y \sim \text{Poi}(\lambda)$, wenn

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots$$

Anwenden mit λT statt λ:

$$P(Y = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

(Normierung der Zeit: $T = 1$.)

Poisson-Grenzwertsatz

Verwende: $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$.

$$P(Y_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k}$$
$$= \frac{n}{n}\frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{1}{k!} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k}$$

$(np_n)^k \to \lambda^k$, letzter Faktor: $\to e^{-\lambda}$. Also

$$P(Y_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, \cdots$$

Die Zahlen rechts definieren eine Zähldichte (auf $\mathbb{N}_0$), da

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Poisson-Verteilung

Eigenschaften

Es gilt:

$$\begin{aligned} \text{Erwartungswert:} \quad & E(Y) = \lambda, \\ \text{Varianz:} \quad & \mathrm{Var}(Y) = \lambda, \\ \text{Zähldichte:} \quad & p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \ k \in \mathbb{N}_0. \end{aligned}$$

Rechnen mit poissonverteilten Zufallsvariablen:

- $X \sim \mathrm{Poi}(\lambda)$ und $Y \sim \mathrm{Poi}(\mu)$ unabhängig, dann $X + Y \sim \mathrm{Poi}(\lambda + \mu)$.
- $X \sim \mathrm{Poi}(\lambda)$ die Anzahl in $[0, T]$ und Y die Anzahl im Teilintervall $[0, r \cdot T]$, so ist $Y \sim \mathrm{Poi}(r \cdot \lambda)$.

Poisson-Verteilung

Beispiel: Ein Investor beobachtet die minütlichen Aktienkursnotierungen. Y_{12} sei die Anzahl der Überschreitungen des Kursziels 100 während des 12-stündigen Handels, Y_1 während der ersten Stunde.
Y_{12} und Y_1 seien poissonverteilt. Pro Stunde wird eine Überschreitung erwartet.
Bestimmung von λ_1, λ_{12}: $\lambda_1 = E(Y_1) = 1$, $\lambda_{12} = 12 \cdot \lambda_1 = 12$.
Wkeit, dass der Aktienkurs unter 100 bleibt:

$$P(Y_{12} = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-12} = 6.144 \cdot 10^{-6}$$

Wkeit, dass der Aktienkurs während der ersten Stunde 100 übersteigt:

$$\begin{aligned} P(Y_1 > 0) &= 1 - P(Y_1 = 0) = 1 - \frac{\lambda_1^0 e^{-\lambda_1}}{0!} \\ &= 1 - e^{-1} = 1 - 0.3678794 = 0.6321206 \end{aligned}$$

Poisson-Verteilung

Approximation der Binomialverteilung für (sehr) kleine p: $Y \sim Bin(n, p)$

$$P(Y = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

mit $\lambda = np$.

Stetige Verteilungsmodelle

Heuristik: Gleichverteilung
Sei $[a, b]$ ein Intervall und $I \subset [a, b]$ ein (sehr) kleines Teilintervall.
$|I|$ bezeichnet die Länge des Intervalls.
Anschauung: X ist gleichverteilt in $[a, b]$, wenn...

- $P(X \in I)$ ist proportional zu $|I|$
- $P(X \in I)$ invariant (ändert sich nicht) unter Verschiebungen von I:

$$P(X \in I + a) = P(X \in I), \qquad \forall a$$

Stetige Gleichverteilung (uniforme Verteilung)

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \\ 0, & x \notin [a,b]. \end{cases}$$

X heißt dann **(stetig) gleichverteilt auf dem Intervall** $[a,b]$. Notation: $X \sim U[a,b]$. Für die Verteilungsfunktion ergibt sich:

$$F(x) = \frac{x-a}{b-a}, \qquad x \in [a,b],$$

sowie $F(x) = 0$, wenn $x < a$, und $F(x) = 1$, für $x > b$. Es gilt:

$$\text{Erwartungswert:} \quad E(X) = \frac{a+b}{2},$$

$$\text{Varianz:} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Exponentialverteilung

Wartezeit auf poissonverteiltes Ereignis:

1. Die Anzahl von Ereignissen in einem Intervall sei Poi(λ)-verteilt.
2. Anzahl Y_t der Ereignisse in $[0, t]$ ist dann Poi(λt)-verteilt.
3. Sei X die Wartezeit auf das erste Ereignis.
4. Es gilt:

$$X > t \quad \Leftrightarrow \quad Y_t = 0$$

also

$$P(X > t) = P(Y_t = 0) = e^{-\lambda t}$$

Linke Seite ist $1 - F_X(t)$. Also gilt für $t > 0$:

$$F_X(t) = 1 - e^{-\lambda t} \quad \Rightarrow \quad f_X(t) = \lambda e^{-\lambda t}$$

X heißt **exponentialverteilt**, $X \sim \text{Exp}(\lambda)$ mit Parameter $\lambda > 0$.

Exponentialverteilung

X heißt **exponentialverteilt**, $X \sim \text{Exp}(\lambda)$ mit Parameter $\lambda > 0$, wenn X die Dichtefunktion

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0,$$

und $f(x) = 0$ für $x \leq 0$ besitzt.

Erwartungswert:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x) \mathbf{1}_{(0,\infty)}(x \, dx \\ &= \int_{0}^{\infty} x \cdot \lambda \cdot e^{-\lambda x} \, dx \\ &= \dots \\ &= \frac{1}{\lambda}. \end{aligned}$$

Varianz:

$$E(X - \lambda)^2 = E(X^2) - \left(\frac{1}{\lambda}\right)^2 = \dots = \frac{1}{\lambda^2}$$

Normalverteilung

X heißt normalverteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in (0, \infty)$, falls X die Dichte

$$\varphi_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

hat $\rightarrow$ **Gauß'sche Glockenkurve**.
Carl Friedrich Gauß (1777-1885): Berühmter Göttinger Mathematiker

Gauss


3D9
DEUTSCHE BUNDESBANK
Banknote
10
1777 1855 Carl Friedr. Gauß
ZEHN DEUTSCHE MARK
DN6833178D9

Lupe...

Eigenschaften

Verteilungsfunktion:

$$\Phi_{(\mu,\sigma^2)}(x) = \int_{-\infty}^{x} \varphi_{(\mu,\sigma^2)}(t)\, dt = p$$

Quantilfunktion (Umkehrfunktion):

$$x = \Phi^{-1}_{(\mu,\sigma^2)}(p)$$

Es gibt keine expliziten Formeln! $\rightarrow$ Computer / Tabellen
R: `p = pnorm(x), x = qnorm(p).`

Eigenschaften

Eigenschafen der Normalverteilung

1. $E(X) = \int_{-\infty}^{\infty} x\varphi_{(\mu,\sigma^2)}(x)dx = \mu.$
2. $\mathrm{Var}(X) = \int_{-\infty}^{\infty}(x-\mu)^2\varphi_{(\mu,\sigma^2)}(x)dx = \sigma^2.$
3. Dichte hat Symmetriezentrum: μ
4. Dichte hat Wendepunkte bei $\mu \pm \sigma$

Rechnen mit normalverteilten ZVen

Rechenregeln

1. $X \sim N(\mu_1, \sigma_1^2)$ und $Y \sim N(\mu_2, \sigma_2^2)$ unabhängig, dann gilt: $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
2. Ist $X \sim N(\mu, \sigma^2)$ und sind $a, b \in \mathbb{R}$, dann gilt: $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
3. Ist $X \sim N(\mu, \sigma^2)$, dann gilt:

$$X^* = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

   X^*: Standardisierte Version.
4. Ist $X^* \sim N(0, 1)$, dann gilt

$$\mu + \sigma \cdot X^* \sim N(\mu, \sigma^2).$$

Normalverteilung: Rechenregeln

Regel

Sind $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ unabhängig, dann ist das arithmetische Mittel normalverteilt mit Erwartungswert μ und Varianz σ^2/n:

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

und

$$\overline{X}^* = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim N(0, 1).$$

Zufallsvektoren

Ω abzählbar, dann heißt jede Abbildung

$$\mathbf{X} : \Omega \to \mathbb{R}^n, \qquad \omega \mapsto \mathbf{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$$

in den n-dimensionalen Raum $\mathbb{R}^n$ **Zufallsvektor**.
Zusatz: Ist Ω überabzählbar, dann müssen alle X_i, $i = 1, \ldots, n$, messbar sein.

Realisationen von $\mathbf{X} = (X_1, \ldots, X_n)$ sind Vektoren $\mathbf{x}$ im $\mathbb{R}^n$:
$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Verteilung eines Zufallsvektors (X, Y)

Verteilungsfunktion:

$$F(x, y) = P(X \leq x, Y \leq y), \qquad x, y \in \mathbb{R}$$

Es gilt: $F(-\infty, y) = P(\{X \leq -\infty\} \cap \{Y \leq y\}) = P(\emptyset \cap \{Y \leq y\}) = 0$.
Rand-Verteilungsfunktionen:

$$F_X(x) = P(X \leq x) = F(x, \infty) = \lim_{y \to \infty} F(x, y)$$

$$F_Y(y) = P(Y \leq y) = F(\infty, y) = \lim_{x \to \infty} F(x, y)$$

Für $(X_1, \ldots, X_n)$ entsprechend, z. B.

$$F_{(X,Y,Z)}(\infty, y, \infty) = P(Y \leq y), \qquad y \in \mathbb{R}.$$

Verteilung eines Zufallsvektors (X, Y)

Diskrete Zufallsvektoren: Verteilung durch Zähldichten gegeben

$$p(x, y) = P(X = x, Y = y), \qquad (x, y) \in \mathbb{R}^2$$

- Stäbe über der (x, y)-Ebene an denjenigen Stellen (x, y) mit $P(X = x, Y = x) > 0$, sonst 0.
- Entspricht der Kontingenztabelle der gemeinsamen Wahrscheinlichkeiten.
- $P(X \in A, Y \in B) = \sum_{(x,y) \in A \times B} p(x, y)$

(Summe der Stäbe, die in $A \times B$ stehen.)

Stetige Zufallsvektoren: Verteilung gegeben durch Dichte $f(x, y)$

$$f(x, y) \geq 0, (x, y) \in \mathbb{R}^2, \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$$

- 'Gebirge' über der (x, y)-Ebene.
- $P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f(x, y)\, dydx$

Produktverteilung

Produktmodell / Produktverteilung

Produkt-Verteilungsfunktion

$$F(x,y) = F_X(x) \cdot F_Y(y), \qquad (x,y) \in \mathbb{R}^2$$

Produkt-Zähldichte

$$p(x,y) = p_X(x) \cdot p_Y(y), \qquad (x,y) \in \mathbb{R}^2$$

Produkt-Dichtefunktion

$$f(x,y) = f_X(x) \cdot f_Y(y), \qquad (x,y) \in \mathbb{R}^2$$

Dies entspricht der stochastischen Unabhängigkeit von X und Y.

Verteilung eines Zufallsvektors (X, Y)

Beispiel: $X \sim \text{Bin}(10, 0.4)$ und $Y \sim \text{Poi}(2)$ seien unabhängig. Dann gilt:

$$P(X = k, Y = l) = \binom{10}{k} 0.4^k 0.6^{10-k} \cdot \frac{2^l e^{-l}}{l!}, k = 0, \ldots, n, k = 0, 1, 2, \ldots$$

Beispiel: Sind $X, Y \sim N(0, 1)$ unabhängig, dann hat (X, Y) die Dichte

$$\begin{aligned} f(x, y) &= \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \cdot \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) \\ &= \frac{1}{2\pi} \exp(-(x^2 + y^2)/2) \end{aligned}$$

für $(x, y) \in \mathbb{R}^2$.

Beispiel

Beispiel: Gelte $(X, Y) \sim f(x, y)$ wobei

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)}, & x, y > 0 \\ 0, & \text{sonst} \end{cases}$$

Für alle $x, y > 0$ gilt:

$$f(x,y) = \lambda^2 e^{-\lambda(x+y)} = \lambda^2 e^{-\lambda x} e^{-\lambda y} = (\lambda e^{-\lambda x})(\lambda e^{-\lambda y})$$

$f(x, y)$ ist das Produkt von zwei Exp(λ)-Dichten (die 0 sind auf der nichtpositiven Halbachse). Also sind X und Y unabhängig und identisch Exp(λ)-verteilt.

Bedingte Verteilung, Unabhängigkeit

X, Y diskret mit Werten in $\mathcal{X} = \{x_1, x_2, \ldots, \}$ bzw. $\mathcal{Y} = \{y_1, y_2, \ldots, \}$.

1. Bedingte Wahrscheinlichkeit von $X = x_i$ gegeben $Y = y_j$:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p(x_i, y_j)}{p_Y(y_j)}$$

   $p(x_i, y_j)$: gem. Zähldichte, $p_Y(y_j)$: Zähldichte von Y.

2. Definiert die **bedingte Zähldichte**

$$p(x|y) = p_{X|Y}(x|y) = P(X = x|Y = y) = \begin{cases} \frac{p(x,y)}{p_Y(y)}, & y \in \{y_1, y_2, \ldots\} \\ p_X(x), & y \notin \{y_1, y_2, \ldots, \} \end{cases}$$

3. Endlicher Fall: $p(x_i, y_j)$: Tabelle (Kontingenzafel), $p_Y(y_j)$: Rand
4. Entsprechend definiert man die bedingte Wahrscheinlichkeit von $Y = y_j$ gegeben $X = x_i$.

Bedingte Zähldichten

Beispiel: X, Y seien Zufallsvariable und es gelte:

$$P(X = 1) = 0.1, \quad P(X = 2) = 0.5, \quad P(X = 3) = 0.4$$

$$P(Y = 0) = 0.2, \quad P(Y = 1) = 0.8$$

Stochastische Abhängigkeit bei Verletzung der Produktregel:

$Y \setminus X$	1	2	3	
0	0.02	0.08	0.1	0.2
1	0.08	0.42	0.3	0.8
	0.1	0.5	0.4	1

Bedingte Wahrscheinlichkeit für $X = 2$ gegeben $Y = 1$:

$$P(X = 2|Y = 1) = \frac{p_{XY}(2, 1)}{p_Y(1)} = \frac{0.42}{0.8} = 0.525$$

Beispiel: Klassifikation im Machine Learning

Problem: Klassifiziere Objekte in 2 Klassen, 0: ok, 1: defekt
Klassifizierer (Algorithmus) berechnet $Y = g(X) \in \{0, 1\}$.
X: Zufallsvariable(n) (Input-Feature(s))
Z: wahres Label mit Werten in $\{0, 1\}$ (unbeobachtet).

Beispiele: Virus (defekt=infiziert), Produktion (Objekt=Produkt), ...

Entscheidung für Klasse 0: „0" $= \{Y = 0\}$,
Entscheidung für Klasse 1: „1" $= \{Y = 1\}$

Wahre Wahrscheinlichkeiten in der Population:

	wahres Label Z		
Klassifizierer Y	0	1	
„0"	p_{00}	p_{01}	$p_{00} + p_{01}$
„1"	p_{10}	p_{11}	$p_{10} + p_{11}$

Fehlklassifikationswahrscheinlichkeit: $p_{err} = p_{01} + p_{10}$.
Falsch-Positiv-Rate: $P(Z = 0|Y = 1) = \frac{p_{10}}{p_{10}+p_{11}}$

Beispiel: Klassifikation im Machine Learning

Falsch-Positiv-Rate: $P(Z = 0|Y = 1) = \frac{p_{10}}{p_{10}+p_{11}}$
Analyse des Zählers:

$$\begin{aligned} p_{10} &= P(Y = 1|Z = 0)P(Z = 0) \\ &= (1 - P(Y = 0|Z = 0)) \cdot (1 - P(Z = 1)) \end{aligned}$$

Hier ist $P(Y = 0|Z = 0)$ die **Spezifität** des Klassifiziers (richtige Entscheidung bei Nicht-Defekten) und $P(Z = 1)$ der Defektanteil.
Falsch-Positiv-Rate kann sehr groß sein, wenn Spezifität nicht nahe 1 und/oder Defektanteil klein!
Sensitivität: $P(Y = 1|Z = 1)$ (Erfolgsrate unter Defekten)

- Algorithmen optimieren in der Regel $p_{err} = p_{01} + p_{11}$
- Sensitivität und Spezifität schätzbar aus Testfällen mit Z bekannt.
- Anwendung problematisch, wenn Falsch-Positiv-Rate hoch!

Simpson's Paradoxon

Situation: Verteilung von Y gegeben $X = x$ interessiert. Zusätzliche Variable Z

Beispiel: Fahrprüfungen: Y: best. / n. best., X: M/W, Z: Tag 1, 2

Tabellen nach Tagen (Werten von Z):

Tag $Z = 1$:

$Y \setminus X$	M	F
best.	1	9
n. best.	0	1
	1	10

Durchfallquoten: 0 %, 10 %

Tag $Z = 2$:

$Y \setminus X$	M	F
best.	3	1
n. best.	1	1
	4	2

Durchfallquoten: 25%, 50 %

Aggregiert über Z:

$Y \setminus X$	M	F
best.	4	10
n. best.	1	2
	5	12

Durchfallquoten: 20%, 16.67 %

Simpson's Paradoxon

Paradoxon: Das Ergebnis kehrt sich um, d.h. aggregriert anderes Ergebnis als aufgeschlüsselt!

Ursache: Verteilung von Variable Z unterschiedlich für Werte von X:

Verteilung von Z unter Männern: $1/5, 4/5$

Verteilung von Z unter Frauen: $10/12, 2/12$

(Männer wählen meist Tag 2, Frauen Tag 1)

Rechnerisch gilt: ($Y = 1$ für n. best., rel. Hf. als Wkeiten)

$$P(Y=1|X=M) = \underbrace{P(Y=1|X=M,Z=1)}_{\text{Quote aus Teilpopulation } Z=1} \cdot \underbrace{P(Z=1|X=M)}_{\text{Anteil } Z=1 \text{ unter } X=M}$$

$$+ \underbrace{P(Y=1|X=M,Z=2)}_{\text{Quote aus Teilpopulation } Z=2} \cdot \underbrace{P(Z=2|X=M)}_{\text{Anteil } Z=2 \text{ unter } X=M}$$

$$= \frac{0}{1} \cdot \frac{1}{5} + \frac{1}{4} \cdot \frac{4}{5} = \frac{1}{5}$$

$$P(Y=1|X=F) = \frac{1}{10}\frac{10}{12} + \frac{1}{2}\frac{2}{12} = \frac{2}{12} = \frac{1}{6} = 0.1\bar{6}$$

Simpson's Paradoxon: Klassisches Beispiel

Klassisches Beispiel: Y: Zulassung zum Studium, X: Geschlecht, Z: Studienfach.

Berkeley 1973: 35% der Frauen zugelassen, aber 44% der Männer.

Frage: Liegt eine Diskriminierung vor?

Analyse aufgeschlüsselt nach Studienfächern (Fakultäten) bestätigte den Verdacht nicht.

Ursache: Verteilung des Studienfachs Z bei Frauen anders als bei Männern. Frauen bewarben sich häufiger als Männer für Fächer mit niedrigen Zulassungsquoten für beide Geschlechter.

Simpson's Paradoxon und Big Data/ML/AI

• In der Regel ist die Detailanalyse nach den Werten Z informiver und liefert eine genauere Interpretation.
• Generell Vorsicht bei aggregierten Daten/Tabellen! Besser aufschlüsseln nach weiteren (potentiellen) Einflussvariablen.

Allgegenwärtiges, reales Problem bei Big Data, ML, AI:

Diskriminierung, Bias, Fairness-Begriff

Datenbestände können ungewollten gesellschaftlichen Bias und Diskriminierung abbilden. Schlechterstellung/Nachteil (sagen wir $Y = 1$) bei Vorliegen von bestimmten Ausprägungen ($X = x^*$) häufiger als bei anderen Ausprägungen von X. KI-Verfahren lernen dies aus den Daten. Was soll Fairness bedeuten? Gleiche Quoten marginal oder in Teilpopulationen ($Z = z$)?

Bedingte Dichten

Bedingte Dichte

X und Y stetig verteilt: $(X, Y) \sim f(x, y)$.
Bedingte Dichte von X gegeben $Y = y$ (y fest):

$$f(x|y) = f_{X|Y}(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)}, & f_Y(y) > 0, \\ f_X(x), & f_Y(y) = 0, \end{cases}$$

Dies ist als Funktion von x eine Dichte (für jedes $y \in \mathbb{R}$).
Notation: $X|Y = y \sim f_{X|Y}(x|y)$.

Bedingte Dichten

Beispiel: Bedingte Dichte für $y > 0$:

$$f(x|y) = y \cdot e^{-y \cdot x} \mathbf{1}_{(0,\infty)}(x), \qquad x \in \mathbb{R}.$$

Ist $Y \sim U(1,3)$, dann ist die gemeinsame Dichte

$$f(x,y) = f(x|y) f_Y(y) = \begin{cases} \frac{1}{2} y \cdot e^{-y \cdot x}, & y \in [1,3], x > 0, \\ 0, & x \leq 0 \text{ oder } y < 1 \text{ oder } y > 3. \end{cases}$$

Kriterien für Unabhängigkeit

1. Diskreter Fall: X und Y sind genau dann stochastisch unabhängig, wenn für alle x und y gilt:

$$p_{X|Y}(x|y) = p_X(x) \qquad \text{bzw.} \qquad p_{Y|X}(y|x) = p_Y(y).$$

   bzw. $p(x,y) = p_X(x)p_Y(y)$
   (gem. Zähldichte = Produkt-Zähldichte)

2. Stetiger Fall: X und Y genau dann stochastisch unabhängig, wenn für alle x und y gilt:

$$f_{X|Y}(x) = f_X(x) \qquad \text{bzw.} \qquad f_{Y|X}(y) = f_Y(y).$$

   bzw. $f(x,y) = f_X(x)f_Y(y)$ (gem. Dichte = Produktdichte).

3. X, Y ist genau dann stochastisch unabhängig, wenn für die gemeinsame Verteilungsfunktion $F_{(X,Y)}(x,y)$ für alle $x, y \in \mathbb{R}$ gilt: $F_{(X,Y)}(x,y) = F_X(x) \cdot F_Y(y)$.

Bedingter Erwartungswert

Berechne EW mit bedingter Verteilung.

1. Sei (X, Y) nach der Zähldichte $p(x, y)$ verteilt.
   Bedingte Erwartungswert von X gegeben $Y = y$ gegeben durch

$$E(X|Y = y) = \sum_{x \in \mathcal{X}} x \cdot p_{X|Y}(x|y)$$

2. Sei $(X, Y) \sim f_{(X,Y)}(x, y)$ stetig. Bedingter Erwartungswert:

$$E(X|Y = y) = \int x \cdot f_{X|Y}(x|y)\, dx$$

3. $g(y) = E(X|Y = y)$ ist eine Funktion von y.
4. Einsetzen der Zufallsvariable Y liefert **bedingte Erwartung von X gegeben** Y. Notation: $E(X|Y) := g(Y)$.

Bedingte Dichten

Beispiel: Bedingte Dichte für $y > 0$:

$$f(x|y) = y \cdot e^{-y \cdot x} \mathbf{1}_{(0,\infty)}(x), \qquad x \in \mathbb{R}.$$

Bedingter Erwartungswert

$$E(X|Y = y) = \int_{-\infty}^{\infty} x \cdot f(x|y)\, dx = \int_{0}^{\infty} xy \cdot e^{-y \cdot x}\, dx = \frac{1}{y}$$

(Wie beim partiellen Differenzieren behandle y als Konstante und integriere nach x).
Also zum Beispiel: Für $Y = 2$ erwarten wir

$$E(X|Y = 2) = \frac{1}{2}$$

Quantil-Transformation:

Ist $U \sim U[0,1]$, dann ist die Zufallsvariable $F^{-1}(U)$ nach der Verteilungsfunktion F verteilt.

Beispiel: $-\ln(U)/\lambda$ ist $\mathrm{Exp}(\lambda)$-verteilt.

Normalverteilung: Box-Muller-Methode

Sind U_1, U_2 unabhängig und identisch $U[0,1]$-verteilt, dann sind

$$Z_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2),$$
$$Z_2 = \sqrt{-2 \ln U_1} \sin(2\pi U_2),$$

unabhängig und identisch $N(0,1)$-verteilt und

$$X_i = \mu + \sigma Z_i, \qquad i = 1, 2,$$

unabhängig und identisch $N(\mu, \sigma^2)$-verteilt.

Erwartungswertvektor

Erwartungswertvektor

$\mathbf{X} = (X_1, \ldots, X_n)'$ Zufallsvektor.
Gelte: $\mu_i = E(X_i)$, $i = 1, \ldots, n$, existieren.
Der (Spalten-) Vektor $\boldsymbol{\mu} = (E(X_1), \ldots, E(X_n))'$ heißt
Erwartungswertvektor von X.

Rechenregeln übertragen sich. Insbesondere gilt für zwei Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$ sowie Skalare $a, b \in \mathbb{R}$:

$$E(a \cdot \mathbf{X} + b \cdot \mathbf{Y}) = a \cdot E(\mathbf{X}) + b \cdot E(\mathbf{Y}).$$

Rechenbeispiel

Beispiel: Der Zufallsvektor $\mathbf{X} = (X_1, X_2)^\top$ sei gegeben durch

X_1 : Anzahl der Sechsen bei 12 Würfen mit einem fairen Würfel,

$X_2 = 10 + 2 \cdot Z$

wobei $Z \sim N(0, 1)$. Es gilt $X_1 \sim \text{bin}(12, 1/6)$ und somit

$$E(X_1) = 12 \cdot \frac{1}{6} = 2.$$

Nach den Rechenregeln für normalverteilte Zufallsvariablen ist $X_2 \sim N(10, 2^2)$ und somit $E(X_2) = 10$. Daher ist

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} = \begin{pmatrix} 2 \\ 10 \end{pmatrix}.$$

Transformationsformel für Erwartungswerte

$E(Y) =?$ **für** $Y = g(\mathbf{X})$**:**
Ist $\mathbf{X}$ nach der diskreten Zähldichte $p_{\mathbf{X}}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n$, verteilt, dann gilt

$$E(Y) = E(g(\mathbf{X})) = \sum_{\mathbf{x}} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}).$$

Ist $\mathbf{X}$ nach der stetigen Dichte $f_{\mathbf{X}}(\mathbf{x})$ verteilt, dann gilt

$$E(Y) = E(g(\mathbf{X})) = \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}.$$

Beispiel:
Es gelte $\mathbf{X} = (X_1, X_2)' \sim f_{\mathbf{X}}$ mit

$$f_{(X_1,X_2)}(x_1, x_2) = \begin{cases} x_2^3 & \text{falls } x_1 \in [0,4] \text{ und } x_2 \in [0,1], \\ 0, & \text{sonst.} \end{cases}$$

Zu bestimmen sei $Eg(X_1, X_1)$ für die Funktion $g(x_1, x_2) = x_1 \cdot x_2,\ x_1, x_2 \in \mathbb{R}$. Wir erhalten

$$\begin{aligned} E(X_1 \cdot X_2) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1 x_2 f_{(X_1,X_2)}(x_1, x_2) dx_1 dx_2 \\ &= \int_0^1 \int_0^4 x_1 x_2 x_2^3 dx_1 dx_2 \\ &= \int_0^1 x_2^4 \left(\frac{x_1^2}{2} \Big|_{x_1=0}^{x_1=4} \right) dx_2 = \ldots = \frac{8}{5}. \end{aligned}$$

Kovarianz

Bekannt: X, Y unabhängig, dann gilt: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
Frage: Formel für den allgemeinen Fall?
Sei $\mu_X = E(X)$ und $\mu_Y = E(Y)$.
Ausquadrieren und Linearität des Erwartungswertes ausnutzen:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E\left(\left(X + Y - (\mu_X + \mu_Y)\right)^2 \right) \\
&= E\left(\left((X-\mu_X) + (Y-\mu_Y)\right)^2 \right) \\
&= E\left((X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2 \right) \\
&= E\left((X-\mu_X)^2\right) + 2E\left((X-\mu_X)(Y-\mu_Y)\right) + E\left((Y-\mu_Y)^2\right) \\
&= \mathrm{Var}(X) + 2E(X-\mu_X)(Y-\mu_Y) + \mathrm{Var}(Y).
\end{aligned}$$

Sind X und Y stochastisch unabhängig, dann gilt für den mittleren Term

$$E((X-\mu_X)(Y-\mu_Y)) = E(X-\mu_X) \cdot E(Y-\mu_Y) = 0.$$

Kovarianz

Kovarianz, Kovarianzmatrix

1. Sind X und Y Zufallsvariablen mit existierenden Varianzen, dann heißt
$$\mathrm{Cov}(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big)$$
Kovarianz von X und Y.
2. X, Y heißen **unkorreliert**, falls $\mathrm{Cov}(X,Y) = 0$.
3. Ist $\mathbf{X} = (X_1, \ldots, X_n)$ ein Zufallsvektor, dann heißt die symmetrische $(n \times n)$-Matrix $\mathrm{Var}(\mathbf{X}) = (\mathrm{Cov}(X_i, X_j))_{i,j}$ der n^2 Kovarianzen **Kovarianzmatrix von X.**
4. Alle X_i paarweise unkorreliert: $\mathrm{Var}(\mathbf{X})$ Diagonalmatrix.
5. Korrelation:
$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Kovarianz: Rechenregeln

Rechenregeln

X, Y und Z Zufallsvariablen mit endlichen Varianzen. Dann gelten für alle $a, b \in \mathbb{R}$ die folgenden Rechenregeln:

1. $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$.
2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
3. $\mathrm{Cov}(X, Y) = 0$, wenn X und Y unabhängig sind.
4. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
5. Cauchy-Schwarz-Ungleichung: $\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$.

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)} = \sigma_X \sigma_Y.$$

(liefert: $\mathrm{Cor}(X, Y) \in [-1, 1]$.)

Kovarianz: Rechenbeispiel

Sei $Z \sim N(0,1)$ und $\mathbf{X} = (X_1, X_2)'$ gegeben durch

$$X_1 = 1 + 2Z,$$
$$X_2 = 3Z.$$

Dann gilt

$$\mathrm{Var}(X_1) = 4, \qquad \mathrm{Var}(X_2) = 9$$

und die Kovarianz zwischen X_1 und X_2 berechnet sich zu

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \mathrm{Cov}(1 + 2Z, 3Z) \\ &= \mathrm{Cov}(2Z, 3Z) \\ &= 2 \cdot 3 \cdot \mathrm{Cov}(Z, Z) \\ &= 6\mathrm{Var}(Z) = 6.\end{aligned}$$

Somit erhalten wir für die Kovarianzmatrix

$$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 9 \end{pmatrix}.$$

Anwendung: Multivariate Normalverteilung

Definition: Sind $X_1, \ldots, X_n$ unabhängig und identisch $N(0,1)$-verteilte Zufallsvariablen, dann ist die gemeinsame Dichtefunktion des Zufallsvektors $\mathbf{X} = (X_1, \ldots, X_n)'$ gegeben durch

$$\varphi(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right), \qquad x_1, \ldots, x_n \in \mathbb{R}.$$

$\mathbf{X}$ heißt **multivariat** oder n-dimensional **standardnormalverteilt**.
Notation: $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$

Erwartungswertvektor und Kovarianzmatrix von $\mathbf{X}$ sind gegeben durch

$$\boldsymbol{\mu} = E(\mathbf{X}) = \mathbf{0} = (0, \ldots, 0)' \in \mathbb{R}^n, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Multivariate Normalverteilung

Beispiel: Produktverteilung (Erinn: $\prod_{i=1}^{n} \varphi_i(x_i)$) aus Normalverteilungen

Sind $X_i \sim N(\mu_i, \sigma_i^2)$ unabhängig, $1 \leq i \leq n$, dann hat $\mathbf{X} = (X_1, \ldots, X_n)'$ die gemeinsame Produktdichte

$$\varphi(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right), \qquad x_1, \ldots, x_n \in \mathbb{R}.$$

Erwartungswert: $E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$
Kovarianzmatrix: $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (Diagonalmatrix).

Notation: $\mathbf{X} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$

Multivariate Normalverteilung

Satz: Ist $\mathbf{a} = (a_1, \ldots, a_n)' \in \mathbb{R}^n$ ein Spaltenvektor und gilt $\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\boldsymbol{\mu}, \mathbf{I})$, dann ist die Linearkombination $\mathbf{a}'\mathbf{X} = a_1X_1 + \cdots + a_nX_n$ ebenfalls normalverteilt mit Erwartungswert

$$E(a_1X_1 + \cdots + a_nX_n) = a_1\mu_1 + \cdots a_n\mu_n = \mathbf{a}'\boldsymbol{\mu}$$

und Varianz

$$\mathrm{Var}(a_1X_1 + \cdots + a_nX_n) = \mathrm{Var}(a_1X_1) + \cdots + \mathrm{Var}(a_nX_n) = a_1^2 + \cdots + a_n^2 = \mathbf{a}'\mathbf{a}.$$

Kurz: Wenn $\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\boldsymbol{\mu}, \mathbf{I})$ und $\mathbf{a} = (a_1, \ldots, a_n)' \in \mathbb{R}^n$, dann folgt

$$\mathbf{a}'\mathbf{X} \sim N_n(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\mathbf{a}).$$

Multivariate Normalverteilung

Anwendungsbeispiel: Künstliche Neuronale Netze (z.B. CNNs) berechnen für einen (zufälligen) d-dim. Input $\mathbf{X}$ lineare Transformationen

$$Z_j = b_j + \mathbf{a}_j'\mathbf{X}, \qquad 1 \leq j \leq m,$$

mit Gewichtungsvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^d$ und Intercept-Termen $b_1, \ldots, b_m$, und dann Ausgaben

$$Y_j = \sigma(Z_j), \qquad 1 \leq j \leq m,$$

mit einer Aktivierungsfunktion $\sigma : \mathbb{R} \to \mathbb{R}$ (als Modell für ein Neuron mit m Dendriten).

Für normalverteilten Input $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ sind die Zwischenergebnisse (Projektionen! - vgl. LA) Z_j multivariat normalverteilt:

$$\mathbf{Z} = (Z_1, \ldots, Z_m)' = \mathbf{b} + \mathbf{A}\mathbf{X} \sim N(\mathbf{b} + \mathbf{a}'\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

mit $\boldsymbol{\Sigma} = (\mathrm{Cov}(Z_i, Z_j))_{\substack{1 \leq i \leq m \\ 1 \leq j \leq m}} = (\mathbf{a}_i'\mathbf{a}_j)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq m}}$, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)'$, $\mathbf{b} = (b_1, \ldots, b_m)'$.

Multivariate Normalverteilung

Seien allg. $\mathbf{a} = (a_1, \ldots, a_n)'$ und $\mathbf{b} = (b_1, \ldots, b_n)'$ Spaltenvektoren sowie

$$U = \mathbf{a}'\mathbf{X} = a_1 X_1 + \cdots + a_n X_n,$$
$$V = \mathbf{b}'\mathbf{X} = b_1 X_1 + \cdots + b_n X_n,$$

zwei Linearkombinationen der Zufallsvariablen $X_1, \ldots, X_n$.

Ist der Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_n)'$ nun $N_n(\mathbf{0}, \mathbf{I})$-verteilt, dann ist aufgrund der Unabhängigkeit der X_i

$$\begin{aligned}\mathrm{Cov}(U, V) &= \mathrm{Cov}(a_1 X_1 + \cdots + a_n X_n, b_1 X_1 + \cdots b_n X_n) \\ &= \mathrm{Cov}(a_1 X_1, b_1 X_1) + \cdots + \mathrm{Cov}(a_n X_n, b_n X_n) \\ &= a_1 b_1 + \cdots + a_n b_n = \mathbf{a}'\mathbf{b}.\end{aligned}$$

Somit sind die Zufallsvariablen U und V genau dann unkorreliert (also unabhängig), wenn $\mathbf{a}'\mathbf{b} = 0$.

Multivariate Normalverteilung

Ist $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$, $\mathbf{b} \in \mathbb{R}^n$ und $\mathbf{A}$ eine $m \times n$-Matrix, dann ist

$$\mathbf{Y} = \mathbf{b} + \mathbf{AX} \sim N(\mathbf{b}, \mathbf{\Sigma})$$

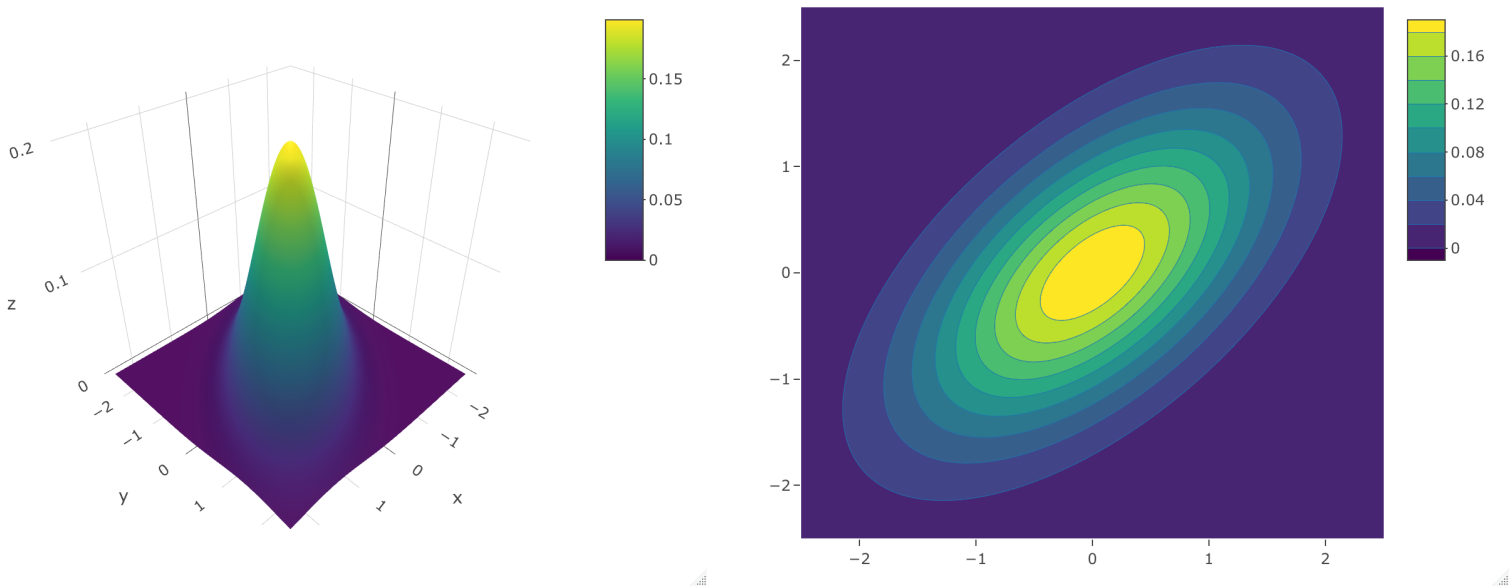
mit $\mathbf{\Sigma} = \mathbf{AA}'$.

Erzeugung von $N(\boldsymbol{\mu}, \mathbf{\Sigma})$-verteilten Vektoren:

- Bestimme Matrix $\mathbf{A}$ mit $\mathbf{\Sigma} = \mathbf{AA}'$ (geht über Singulärwertzerlegung)
- Erzeuge $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ (Box-Muller)
- Berechne $\mathbf{Y} = \mathbf{b} + \mathbf{AX}$

Multivariate Normalverteilung

Dichte einer bivariaten Normalverteilung (Korrelation 0.6, $\boldsymbol{\mu} = \mathbf{0}$).



Höhenlinien sind Ellipsen (Achsen d. Ellipsen gegeben durch Eigenvektoren $\mathbf{v}_i$ von $\boldsymbol{\Sigma}$, Länge der Halbachsen = Wurzel der Eigenwerte).
(Plots von Stichproben sind aufsteigende Punktewolken, die ellipsenförmig aussehen.)

Grenzwertsätze und Konvergenzbegriffe

Gegeben: Eine Folge von Zufallsvariablen:

$$X_1, X_2, X_3, \ldots$$

Formaler:

$$X_n, \qquad n = 1, 2, \ldots$$

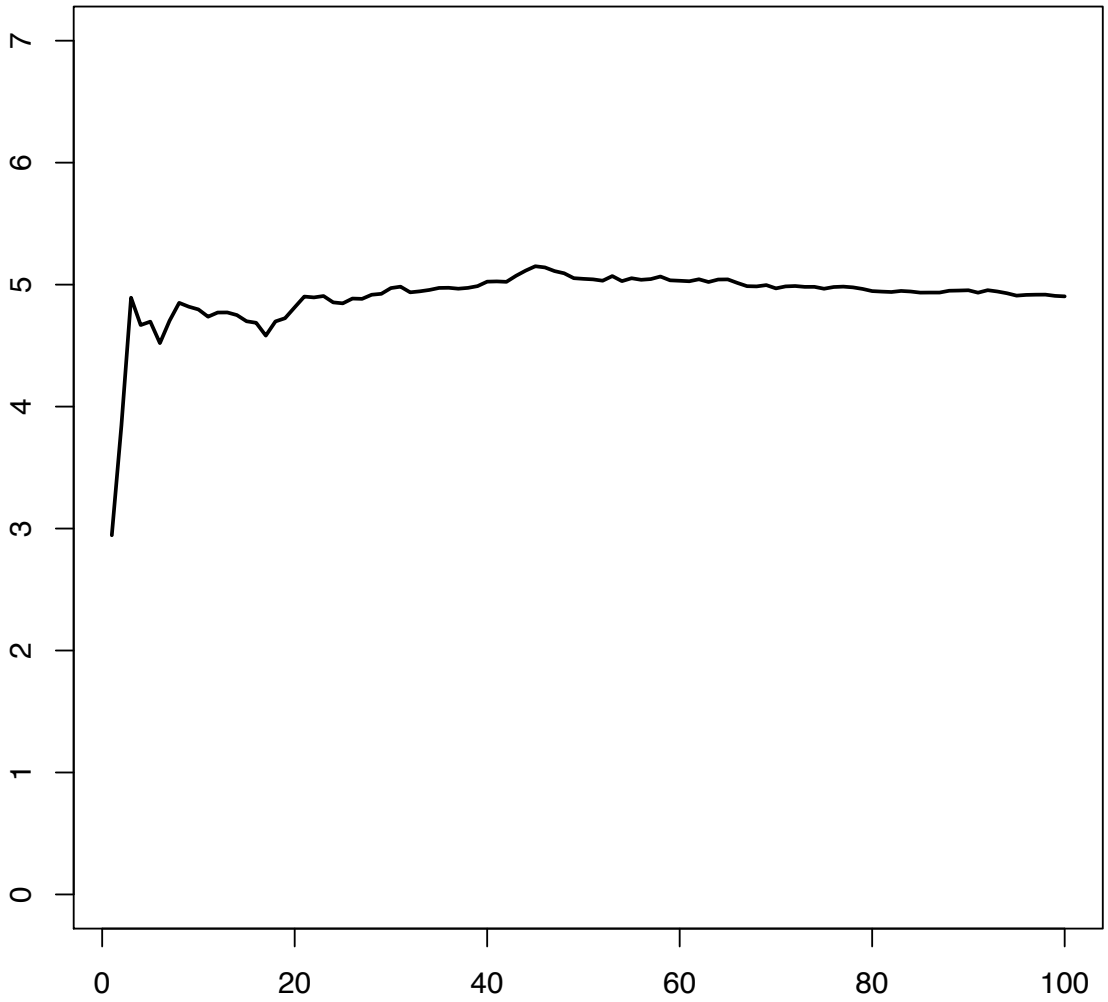
Betrachte die Folge der arithmetischen Mittelwerte:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad n = 1, 2, \ldots$$

$$X_1, \quad \frac{X_1 + X_2}{2}, \quad \frac{X_1 + X_2 + X_3}{3}, \ldots$$

Computerexperiment I: Folge der arithmetischen Mittel

Computerexperiment I:



Gesetz der großen Zahl

Beobachtung:

- Die Folge der $\overline{X}_n$ nähert sich einem festen Wert an.
- Welcher Wert ist das?
- Wie kann man diese 'Konvergenz' beschreiben?

Fundamentales Resultat 1: Gesetz der großen Zahl

Grenzwertsätze und Konvergenzbegriffe

Gegeben: Eine Folge von Zufallsvariablen:

$$X_1, X_2, X_3, \ldots$$

Formaler:

$$X_n, \qquad n = 1, 2, \ldots$$

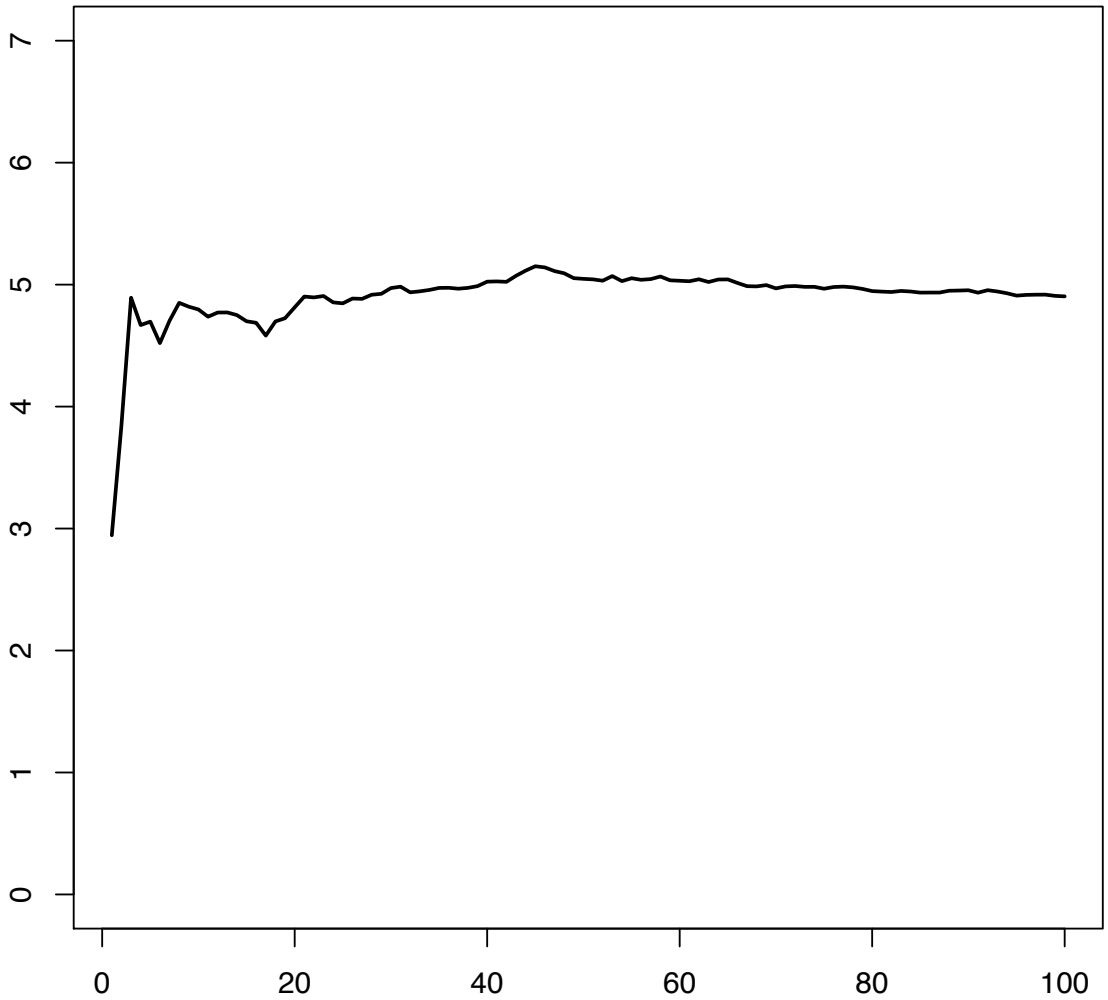
Betrachte speziell die Folge der arithmetischen Mittelwerte:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad n = 1, 2, \ldots$$

$$X_1, \quad \frac{X_1 + X_2}{2}, \quad \frac{X_1 + X_2 + X_3}{3}, \ldots$$

Computerexperiment I: Folge der arithmetischen Mittel

Computerexperiment I:



Gesetz der großen Zahl

Beobachtung:

- Die Folge der $\overline{X}_n$ nähert sich einem festen Wert an.
- Welcher Wert ist das?
- Wie kann man diese 'Konvergenz' beschreiben?

Fundamentales Resultat 1: Gesetz der großen Zahl

Computerexperiment II: Verteilung von $\overline{X}_n^*$

Simuliere auf dem Computer eine Zufallsstichprobe vom Umfang n

$$X_1, \ldots, X_n \sim F,$$

F: vorgegeben, und berechne $\overline{X}_n$. Beispiel in R:

```
x = rexp(100)
res = mean(x)
```

Wiederhole dies $S = 10000$-mal, um eine Stichprobe von $\overline{X}_n$-Werten zu erhalten. Führe dies für verschiedene Verteilungen durch.

Um Histogramme von verschiedenen Simulationen besser vergleichen zu können, standardisieren wir $\overline{X}_n$. D.h.: Berechne statt $\overline{X}_n$

$$\overline{X}_n^* = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

wobei μ der Erwartungswert und σ die Standardabweichung der gewählten Verteilung ist.

Beispielprogramme

R/Splus: Simuliere 10000 $\overline{X}_{100}$-Werte für $X_i \sim U[0,1]$.

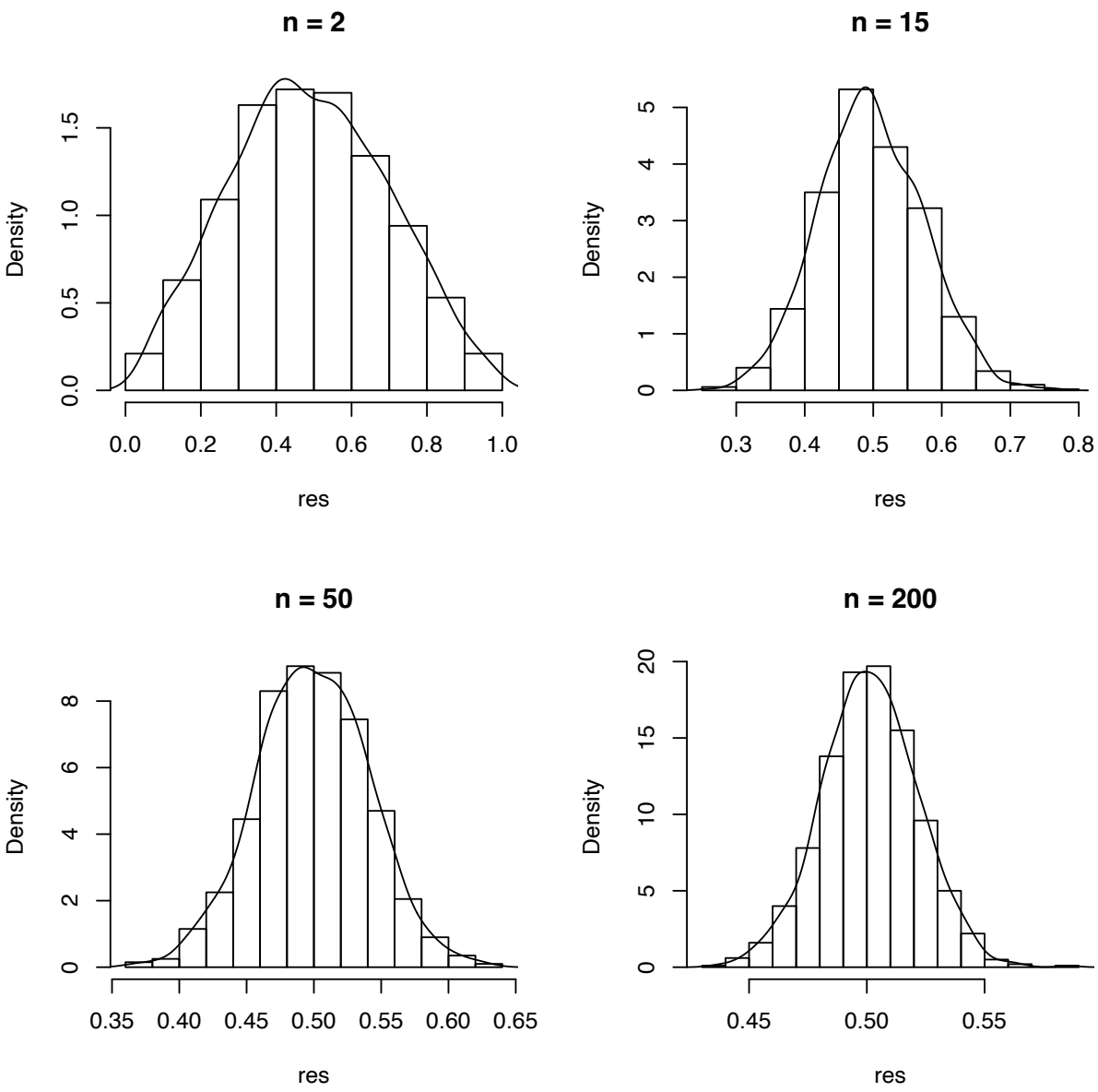
```
res = numeric(10000)
for ( i in (1:10000) ) {
 x = runif(100)
 res[i] = mean(x)
}
hist(res)
```

Matlab/Scilab:

```
res = zeros(10000,1);
for i = 1:10000, x = rand(10,1,'exp');
 res(i)= mean(x);
end;
histplot( 8, res );
```

Frage: Wie muss das Programm modifiziert werden, um Histogramme für die standardisierten Mittelwerte zu erhalten?

Computerexperiment 1: Beobachtungen normalverteilt



Computerexperiment 1: Beobachtungen normalverteilt

Erklärung:

Wir wissen, dass Linearkombinationen

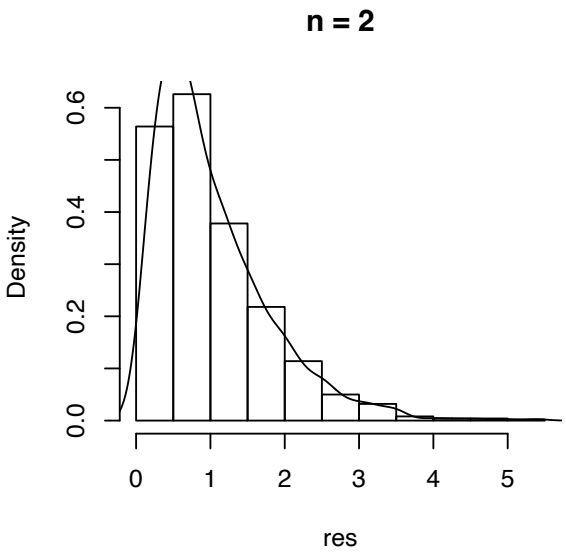
$$a_1 X_1 + \cdots + a_k X_k$$

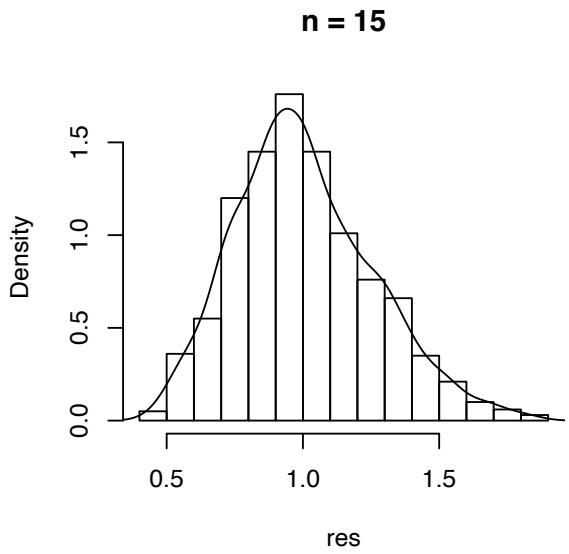
von normalverteilten Zufallsvariablen wieder normalverteilt sind. Daher ist auch $\overline{X}_n$ normalverteilt.

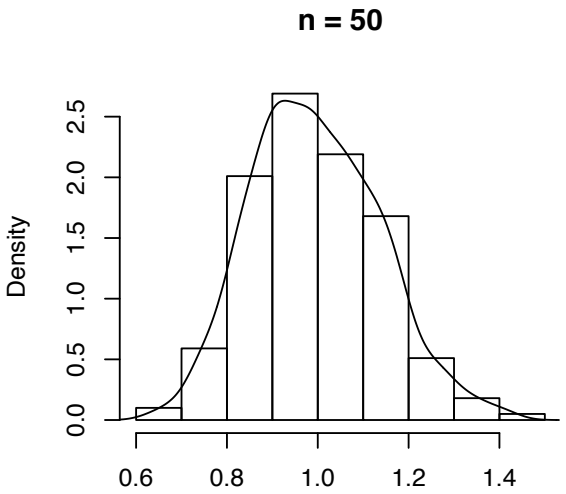
Das Computerexperiment ist also im Einklang mit der Theorie, wenn die X_i normalverteilt sind.

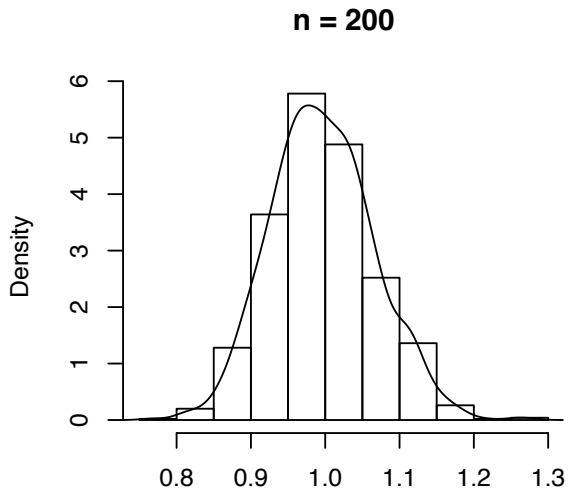
Frage: Was passiert, wenn die X_i anderen Verteilungen folgen?

Computerexperiment 2: Beobachtungen nicht normalverteilt









Asymptotische Verteilung

Beobachtung:

- Für kleine n ist die Verteilung sehr schief und nicht durch eine Normalverteilungsdichte approximierbar.
- Für großes n scheint die Verteilung von $\overline{X}_n$ der Normalverteilung sehr ähnlich zu sein,
- und zwar auch dann, wenn die einzelnen X_i nicht normalverteilt sind!
- Gilt das tatsächlich?
- Wie kann man diese Form der 'Konvergenz' beschreiben?

Fundamentales Resultat 2: Zentraler Grenzwertsatz

Aber zunächst zum ersten fundamentalen Resultat...

Das Gesetz der großen Zahlen

$X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit

$$\mu = E(X_1), \qquad \sigma^2 = \mathrm{Var}(X_1)$$

Sei

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

Frage

Wie groß ist der Fehler, wenn man $\overline{X}_n$ statt μ verwendet?

Das Gesetz der großen Zahlen

Fehler:

$$F_n = |\overline{X}_n - \mu|$$

Toleranz

$$\varepsilon > 0$$

Ereignis:

$$\{F_n > \varepsilon\} = \{|\overline{X}_n - \mu| > \varepsilon\}$$

Wahrscheinlichkeit

$$P(|\overline{X}_n - \mu| > \varepsilon) = ?$$

Tschebyschow-Ungleichung

Tschebyschow (Chebychev, Tschebyscheff, Čebyšëv)-Ungleichung

$X_1, \ldots, X_n$ i.i.d. mit Varianz $\sigma^2 \in (0, \infty)$ und Erwartungswert μ.
Dann gilt:

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Diskussion der oberen Schranke:

1. Schranke umso besser, je kleiner σ^2.
2. Schranke umso besser, je größer n.
3. Die Verteilung der X_i geht nur über σ^2 ein!

Das Gesetz der großen Zahlen

Schwaches Gesetz der großen Zahlen

$X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ und Varianz σ^2, $\sigma^2 \in (0, \infty)$.

Dann konvergiert das arithmetische Mittel $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ im stochastischen Sinne gegen den Erwartungswert μ, d.h. für jede Toleranzabweichung $\varepsilon > 0$ gilt:

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0,$$

wenn n gegen ∞ strebt.

Konvergenzbegriff: Stochastische Konvergenz

Allgemein definiert man:
Definition: (i) Eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen **konvergiert stochastisch** oder **in Wahrscheinlichkeit** gegen die Zufallsvariable X, wenn für alle $\varepsilon > 0$ gilt:

$$P(|X_n - X| > \varepsilon) \to 0, \qquad n \to \infty.$$

Notation: $X_n \xrightarrow{P} X$, $n \to \infty$.
(ii) Eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen **konvergiert stochastisch** oder **in Wahrscheinlichkeit** gegen die Konstante a, wenn für alle $\varepsilon > 0$ gilt:

$$P(|X_n - a| > \varepsilon) \to 0, \qquad n \to \infty.$$

Konvergenzbegriff: Stochastische Konvergenz

Rechenregeln: $X_1, X_2, \cdots$ und $Y_1, Y_2, \cdots$ seien Folgen von ZVen.
(i) Aus $X_n \xrightarrow{P} a$, $n \to \infty$ und $Y_n \xrightarrow{P} b$, $n \to \infty$ folgt für $\lambda, \mu \in \mathbb{R}$:

$$\lambda \cdot X_n \pm \mu \cdot Y_n \xrightarrow{P} \lambda \cdot a \pm \mu \cdot b, \qquad n \to \infty.$$

(ii) Aus $X_n \xrightarrow{P} X$, $n \to \infty$, und $Y_n \xrightarrow{P} b$, $n \to \infty$ folgt

$$X_n \cdot Y_n \xrightarrow{P} X \cdot b, \qquad n \to \infty.$$

und, falls $b \neq 0$ und ein $n_0 \in \mathbb{N}$ existiert mit $P(Y_n \neq 0)$ für $n > n_0$,

$$\frac{X_n}{Y_n} \xrightarrow{P} \frac{X}{b}, \qquad n \to \infty$$

(iii) Aus $X_n \xrightarrow{P} X$, $n \to \infty$, und $Y_n \xrightarrow{P} Y$, $n \to \infty$, folgt für jede stetige Funktion f: $f(X_n) \xrightarrow{P} f(X)$, $n \to \infty$.
(iv) Aus $X_n \xrightarrow{P} X$, $n \to \infty$, und $Y_n \xrightarrow{P} Y$, $n \to \infty$, folgt für jede stetige Funktion $f : \mathbb{R}^2 \to \mathbb{R}$: Falls $f(X, Y)$ und $f(X_n, Y_n)$ definiert sind für alle $n \in \mathbb{N}$, dann gilt: $f(X_n, Y_n) \xrightarrow{P} f(X, Y), n \to \infty$.

Beispiele: Gesetz der großen Zahlen/stochastische Konvergenz

Beispiele:

(i) $\overline{X}_n$ und $\overline{Y}_n$ seien die arithmetischen Mittelwerte aus zwei i.i.d-Stichproben $X_1, X_2, \cdots$ und $Y_1, Y_2, \cdots$ mit Erwartungswerten μ_X und $\mu_Y > 0$. Dann gilt $\overline{X}_n / \overline{Y}_n \xrightarrow{P} \mu_X / \mu_Y, \qquad n \to \infty.$

(ii) Um die Laufzeiten ℓ_x und ℓ_y von zwei nacheinander geschalteten Algorithmen abzuschätzen, werden die Algorithmen jeweils n mal unabhängig voneinander unter identischen Bedingungen gestartet, so dass die einzelnen Laufzeiten $X_1, \cdots, X_n$ und $Y_1, \cdots, Y_n$ als einfache Stichproben mit Erwartungswerten ℓ_x und ℓ_y angesehen werden können. Man approximiert nun ℓ_x durch $\overline{X}_n$ und ℓ_y durch $\overline{Y}_n$. Da nach dem schwachen Gesetz großer Zahlen $\overline{X}_n \xrightarrow{P} \ell_x$ und $\overline{Y}_n \xrightarrow{P} \ell_y$, wenn $n \to \infty$, folgt nach Rechenregel (i):

$$\overline{X}_n + \overline{Y}_n \xrightarrow{P} \ell_x + \ell_y, \qquad n \to \infty.$$

Das Gesetz der großen Zahlen

Fixiere einen Ausgang $\omega \in \Omega$.

- Die Realisationen

$$\overline{x}_1 = \overline{X}_1(\omega),\ \overline{x}_2 = \overline{X}_2(\omega), \ldots$$

  sind eine reelle Zahlenfolge.
- In Abhängigkeit von ω gilt

$$\lim_{n\to\infty} \overline{X}_n(\omega) = \mu$$

  oder nicht.
- Das starke Gesetz macht eine Aussage über die Menge der ω, für die Konvergenz vorliegt:

$$\left\{ \lim_{n\to\infty} \overline{X}_n(\omega) = \mu \right\}$$

Das Gesetz der großen Zahlen

Starkes Gesetz der großen Zahlen

$X_1, \ldots, X_n$ i.i.d. mit $E|X_1| < \infty$ und Erwartungswert μ.

Dann konvergiert das arithmetische Mittel mit Wahrscheinlichkeit 1 gegen μ, d.h.

$$P(\overline{X}_n \to \mu) = P(\{\omega | \overline{X}_n(\omega) \text{ konvergiert gegen } \mu\}) = 1.$$

Diskussion:

1. Schwächere Voraussetzungen: Varianz muss nicht existieren.
2. Aussage stärker: Konvergenz mit Wkeit 1.

Hauptsatz der Statistik

1. Der *Hauptsatz der Statistik* macht eine Aussage über die Konvergenz der empirischen Verteilungsfunktion.
2. Dieses Ergebnis ist fundamental, da sehr viele Funktion von Stichprobenvariablen $X_1, \ldots, X_n$ hierüber ausgedrückt werden können.
3. Also: Der Hauptsatz der Statistik liefert die rigorose Begründung, warum Statistik 'funktioniert'.

Hauptsatz der Statistik

Die Zufallsvariablen $X_1, \ldots, X_n, \ldots$ seien unabhängig und identisch (i.i.d.) nach der Verteilungsfunktion F verteilt.

Dann konvergiert der (maximale) Abstand zwischen der **empirischen Verteilungsfunktion** $F_n(x)$ und der **wahren Verteilungsfunktion** $F(x)$ mit Wahrscheinlichkeit 1 gegen 0:

$$P\left(\lim_{n\to\infty} \max_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1.$$

Konvergenzbegriff: Fast sichere Konvergenz

Allgemein definiert man:
Definition: (i) Eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen **konvergiert fast sicher** oder **mit Wahrscheinlichkeit 1** gegen die Zufallsvariable X, wenn gilt:

$$P(\lim_{n\to\infty} |X_n - X| = 0) = 1$$

Notation: $X_n \overset{f.s.}{\to} X$, $n \to \infty$.
(ii) Eine Folge $X_1, X_2, \ldots$ von Zufallsvariablen **konvergiert fast sicher** oder **mit Wahrscheinlichkeit 1** gegen die Konstante a, wenn gilt:

$$P(\lim_{n\to\infty} |X_n - a| = 0) = 1$$

Konvergenzbegriff: Fast sichere Konvergenz

Rechenregeln: $X_1, X_2, \cdots$ und $Y_1, Y_2, \cdots$ seien Folgen von ZVen.
(i) Aus $X_n \overset{f.s.}{\to} a$, $n \to \infty$ und $Y_n \overset{f.s.}{\to} b$, $n \to \infty$ folgt für $\lambda, \mu \in \mathbb{R}$:

$$\lambda \cdot X_n \pm \mu \cdot Y_n \overset{f.s.}{\to} \lambda \cdot a \pm \mu \cdot b, \qquad n \to \infty.$$

(ii) Aus $X_n \overset{f.s.}{\to} X$, $n \to \infty$, und $Y_n \overset{f.s.}{\to} b$, $n \to \infty$ folgt

$$X_n \cdot Y_n \overset{f.s.}{\to} X \cdot b, \qquad n \to \infty.$$

und, falls $b \neq 0$ und ein $n_0 \in \mathbb{N}$ existiert mit $P(Y_n \neq 0) = 1$ für $n > n_0$,

$$\frac{X_n}{Y_n} \overset{f.s.}{\to} \frac{X}{b}, \qquad n \to \infty$$

(iii) Aus $X_n \overset{f.s.}{\to} X$, $n \to \infty$, und $Y_n \overset{f.s.}{\to} Y$, $n \to \infty$, folgt für jede stetige Funktion f: $f(X_n) \overset{f.s.}{\to} f(X)$, $\qquad n \to \infty$.
(iv) Aus $X_n \overset{f.s.}{\to} X$, $n \to \infty$, und $Y_n \overset{f.s.}{\to} Y$, $n \to \infty$, folgt für jede stetige Funktion $f : \mathbb{R}^2 \to \mathbb{R}$: Falls $f(X, Y)$ und $f(X_n, Y_n)$ definiert sind für alle $n \in \mathbb{N}$, dann gilt: $f(X_n, Y_n) \overset{f.s.}{\to} f(X, Y), n \to \infty$.

Anwendung: Konvergenz der Stichprobenvarianz

Sind $X_1, \cdots, X_n$ i.i.d. mit Erwartungswert μ und Varianz $\sigma^2 \in (0, \infty)$, so heißt

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\overline{X}_n\right)^2$$

Stichprobenvarianz.[1] Es gilt
$\mathrm{Var}(X_1) = \sigma^2 = E(X_1^2) - \mu^2 \Leftrightarrow E(X_1^2) = \sigma^2 + \mu^2 \in \mathbb{R}$. $X_1^2, \cdots, X_n^2$ erfüllen die Voraussetzungen des starken Gesetzes großer Zahlen:

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \overset{f.s.}{\to} E(X_1^2) = \sigma^2 + \mu^2, \quad n \to \infty.$$

Ferner: $\overline{X}_n \overset{f.s.}{\to} \mu$ und$f(x) = x^2$ stetig $\overset{(iii)}{\Rightarrow} \left(\overline{X}_n\right)^2 = f(\overline{X}_n) \overset{f.s.}{\to} f(\mu) = \mu^2, n \to \infty$.
Rechenregel (i) liefert nun die f.s. Konvergenz der Stichprobenvarianz:

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\overline{X}_n\right)^2 \overset{f.s.}{\to} (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$$

[1] Die zweite Formel folgt durch Ausmultiplizieren von $(X_i - \overline{X}_n)^2$ und zusammenfassen.

Der zentrale Grenzwertsatz: Vorbereitung

$X_1, X_2, \cdots$ i.i.d.-Zufallsvariablen mit Erwartungswert μ und Varianz $\sigma^2 \in (0, \infty)$.

Gesetz der großen Zahlen:

$$\overline{X}_n \overset{f.s.}{\to} \mu, \qquad n \to \infty$$

Äquivalent:

$$\overline{X}_n - \mu \overset{f.s.}{\to} 0, \qquad n \to \infty$$

Skaliert man die Abweichung $\overline{X}_n - \mu$ mit $\sqrt{n}$, so 'kollabiert' der Ausdruck nicht mehr gegen 0.

Die mit $\sqrt{n}$ skalierte Abweichung folgt einem universellem Fehlergesetz.

Der zentrale Grenzwertsatz

Die Simulation der Verteilung von $\overline{X}_n^*$ legt folgende Vermutung für Ereignisse A nahe.

Vermutung: In großen Stichproben gilt für Intervalle $[A, B]$:

$$P(\overline{X}_n^* \in [A, B]) \approx P(Z \in [a, b]), \qquad (Z \sim N(0, 1)),$$

und somit:

$$P(\overline{X}_n \in [a, b]) \approx P(Z_n \in [a, b]), \qquad Z_n \sim N(\mu, \sigma^2/n).$$

Zusammenhang: $P\left(\overline{X}_n \in [a, b]\right) = P\left(\overline{X}_n^* \in [\frac{a-\mu}{\sigma/\sqrt{n}}, \frac{b-\mu}{\sigma/\sqrt{n}}]\right)$.

Der zentrale Grenzwertsatz

Zentrale Grenzwertsatz (ZGWS)

$X_1, \ldots, X_n$ i.i.d. mit

$$\mu = E(X_1), \qquad \sigma^2 = \mathrm{Var}(X_1) \in (0, \infty).$$

Dann gilt: $\overline{X}_n$ ist asymptotisch $N(\mu, \sigma^2/n)$-verteilt,

$$\overline{X}_n \sim_{approx} N(\mu, \sigma^2/n),$$

in dem Sinne, dass die Verteilungsfunktion der standardisierten Version gegen die Verteilungsfunktion der $N(0,1)$-Verteilung konvergiert:

$$P\left(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq x\right) \to \Phi(x), \qquad n \to \infty.$$

Der zentrale Grenzwertsatz

Die Aussage des zentralen Grenzwertsatzes bleibt gültig, wenn die in der Praxis meist unbekannte Streuung σ durch die empirische Standardabweichung

$$\widehat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

ersetzt wird.

Szenarioanalyse

Fallgestaltung und Aufgabe:

Ein Autohersteller betrachtet im Rahmen einer Szenarioanalyse die Wahrscheinlichkeit, dass der Gesamtgewinn von n Autohäusern die Benchmark b erreicht.

Hierzu wurden n Autohäuser ausgewählt, die in verschiedenen - aber vergleichbaren - Großstädten unabhängig voneinander operieren und von vergleichbarer Größe sind.

Es ist zu klären, wie die gesuchte Wahrscheinlichkeit - zumindest näherungsweise - ermittelt werden kann.

ZGWS: Fallgestaltung

Modellbildung:

Die zukünftigen zufallsbehafteten Quartalsgewinne $X_1, \ldots, X_n$ werden als einfache (i.i.d.-) Zufallsstichprobe aufgefasst mit Erwartungswert μ und Varianz $\sigma^2 \in (0, \infty)$.

Die genaue Verteilung der Quartalsgewinne ist unbekannt. Lediglich der erwartete Quartalsgewinn wird im Rahmen der Szenarien spezifiert. Als relevanter Bereich wird festgelegt:

$$1.2 \leq \mu \leq 1.6$$

ZGWS: Fallgestaltung

Problemformulierung

Bestimme (approximativ) die Wahrscheinlichkeit, dass der Gesamtgewinn

$$G = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

die Benchmark b übersteigt, also

$$P(G > b) = P\left(\sum_{i=1}^{n} X_i > b\right).$$

Wir werden sehen, dass der ZGWS eine praktikable Lösung ermöglicht, sofern n nicht zu klein ist.

Der zentrale Grenzwertsatz

1. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x)$ mit

$$\mu = E(X_1), \qquad \sigma^2 = \mathrm{Var}(X_1) \in (0, \infty).$$

2. Arithmetisches Mittel:

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

erfüllt: $E(\overline{X}_n) = \mu$ und $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.

3. Dann erfüllt die standardisierte Version

$$\overline{X}_n^* = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

$E(\overline{X}_n^*) = 0$ und $\mathrm{Var}(\overline{X}_n^*) = 1$ für alle $n \in \mathbb{N}$!

ZGWS: Anwendung auf die Fallgestaltung

Der Zentrale Grenzwertsatz liefert die Näherung

$$
\begin{aligned}
P\left(\sum_{i=1}^{n} X_i > b\right) &= P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} > \frac{b - n\mu}{\sqrt{n\sigma^2}}\right) \\
&\approx 1 - \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) \\
&\approx 1 - \Phi\left(\frac{b - n\mu}{\sqrt{n}\widehat{\sigma}_n}\right)
\end{aligned}
$$

Schritt 1: Standardisieren.
Schritt 2: Rückführung auf die (tabellierte) Verteilungsfunktion der $N(0,1)$-Verteilung.
Schritt 3: Ersetzen von σ durch $\widehat{\sigma}_n$.

Hierbei ist

b	:	Benchmark
μ	:	erwarteter Quartalsgewinn im Szenario
$\widehat{\sigma}_n$	:	Schätzung für σ

ZGWS: Anwendung auf die Fallgestaltung

Zahlenbeispiel:
Sei $n = 36$. Für die Standardabweichung liege eine Schätzung aus historischen Daten vor:

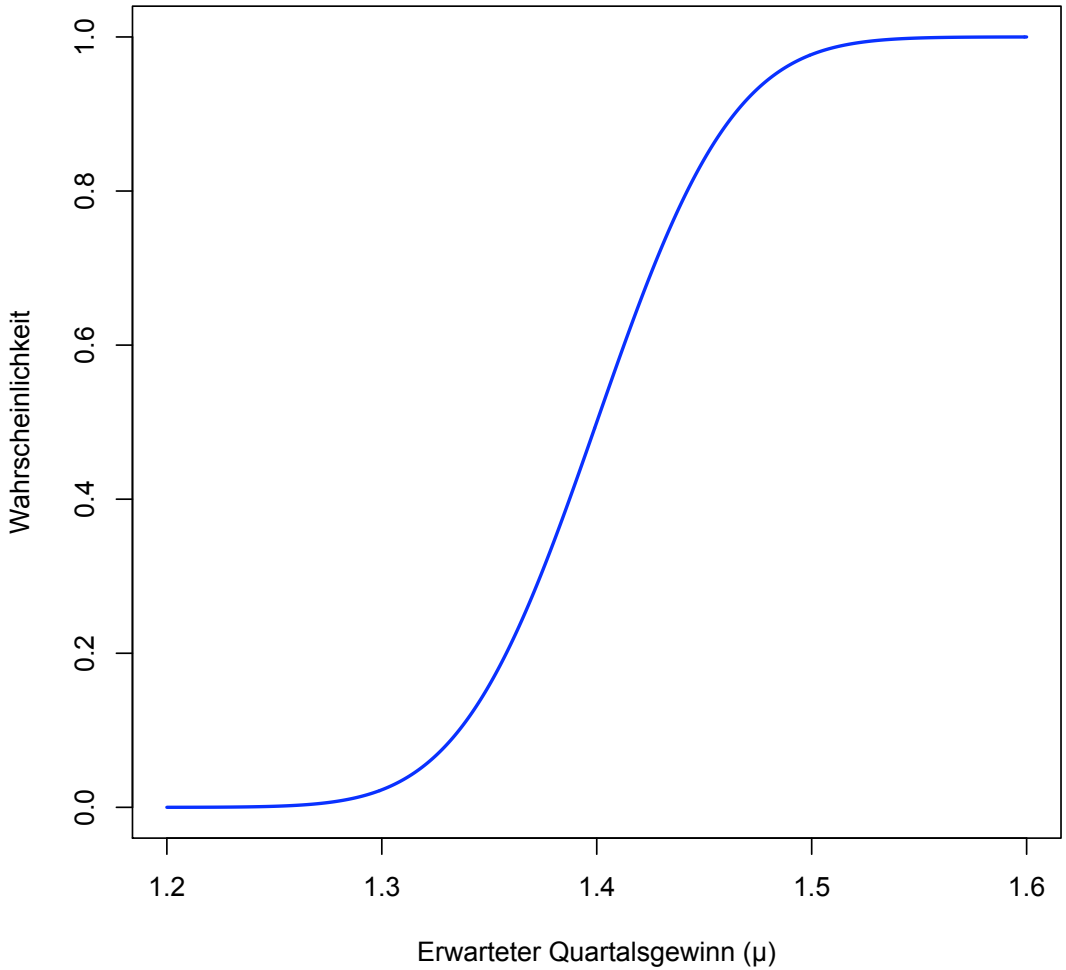
$$\widehat{\sigma}_n = 0.2 \text{ Mio Euro.}$$

Benchmark: $b = 50.4$, d.h. 1.4 Mio pro Händler.

Als relevante Szenarien werden erwartete Quartalsgewinne zwischen 1.2 und 1.6 Mio Euro betrachtet.

Die folgende Grafik zeigt die zugehörige approximative Wahrscheinlichkeit, die Benchmark zu erreichen oder sogar zu schlagen.

ZGWS: Anwendung auf die Fallgestaltung



Statistik

Prof. Dr. Ansgar Steland

EAS 2023

Ziele der Deskriptiven Statistik:

- Empirische Daten durch Tabellen, Grafiken und Kennzahlen übersichtlich darstellen und ordnen.
- Daten durch aussagekräftige Kennzahlen zahlenmäßig zu beschreiben und verdichten.
- Interpretation der aufbereiteten Daten.
- Generierung von Hypothesen.

Hierbei werden keine stochastischen Modelle verwendet, so dass getroffene Ausagen nicht durch Fehlerwahrscheinlichkeiten abgesichert sind. Dies ist Aufgabe der **Schließenden Statistik (Inferenzstatistik)**.

Vorgeschaltet ist die **Planung der statistischen Studie**:

- Was erheben? Wie erheben?
- Worüber sollen Aussagen getroffen? Welche Fragen sind zu beantworten?
- Definition der zu erhebenen Variablen
- Ein- und Ausschlusskriterien
- Sicherstellung der Datenqualität.
- Umgang mit fehlenden Daten.
- Festlegung von Verantwortlichkeiten, Zugriffsrechten.
- Datenspeicherung, Datenschutz.
- Planung der eigentlichen statistischen Analyse: Welche Methoden?

Vollständige Dokumentation.

Grundbegriffe:

Statistische Analyse von Daten:

1. Definition der relevanten **statistischen Einheiten (Untersuchungseinheiten, Merkmalsträger)**
2. Die **Grundgesamtheit** G ist die Menge aller statistischen Einheiten.
3. Erhebe Daten (**Merkmale**, Variablen) an allen (Totalerhebung) oder ausgewählten Einheiten.
4. Werden die Daten durch Experimente gewonnen, dann heißen die $g \in G$ auch **Versuchseinheiten** (experimental units). Werden die Daten durch Beobachtunge gewonnen, so spricht man von **Beobachtungseinheiten** (observational units).
5. Merkmale X nehmen gewissen **Merkmalsausprägungen** M. Formal:

$$X : G \to M, \qquad g \mapsto X(g)$$

(o.E. (durch Kodieren) $M \subset \mathbb{R}$)

6. Zufallsauswahl: Ziehe n Mal aus der 'Urne' G mit Zurücklegen:

$$\Omega = G \times \cdots \times G$$

7. Zufallsstichprobe: $X_1, \ldots, X_n : \Omega \to \mathbb{R}$, unabhängig und identisch verteilte Zufallsvariablen (Zufallsvektoren, wenn mehrere Variablen erhoben werden).

8. Bei Experimenten werden den gezogenen $g \in G$ gewisse Ausprägungen zugeordnet (z.B. Kontrollgruppe/Behandlungsgruppe). Diese haben i.d.R. nur wenige mögliche Ausprägungen (z.B. binär 0/1).

9. Deskriptive Statistik betrachtet Realisation $(x_1, \ldots, x_n)'$ als Input, die **Datenmatrix**.

statistische Einheit	Merkmal	Merkmalsausprägungen
Studierender	Studienfach	BWL/Informatik/WiIng/Biologie/...
	Geschlecht	M/W/D
	Alter	$\mathbb{R}^+$
IT-Unternehmen	Mitarbeiterzahl	$\mathbb{N}$
	Umsatz	$\mathbb{R}_0^+$
	Gewinn/Verlust	$\mathbb{R}$
Arbeitnehmer	Einkommen	$\mathbb{R}^+$
	Bildungsniveau	Abitur/Bachelor/Master/...
	Arbeitszeit	R_0^+
Regionen	Arbeitslosenquote	$[0, 1]$
	Wirtschaftskraft	$\mathbb{R}^+$
Ballungsräume	Populationsdichte	$\mathbb{N}$ oder $\mathbb{R}$
	politische Funktion	Mittelzentrum / Landeshauptstadt / Hauptstadt
Staaten	Bruttoinlandsprodukt	$\mathbb{R}^+$
	Verschuldung (in %)	$[0, 100]$

Skalenniveaus: $X : G \to M$

Diskrete Merkmale: M endlich oder abzählbar unendlich.

Stetige Merkmalee: $M \subset \mathbb{R}$ Intervall (oder ganz $\mathbb{R}$).

In der Praxis werden stetige Merkmale oft vergröbert (komprimiert) durch **Gruppierung**.
Bsp: Einkommensklassen $[0, 500], (500, 1000], (1000, 5000], (5000, \infty)$.

Klassifikation von Merkmalen aufgrund des Skalenniveaus:
• **Nominalskala**: Ausprägungen nur unterscheidbar (Labels)
• **Ordinalskala**: Ausprägungen können verglichen werden (Schulnoten, Grad der Zustimmung 1-5, ...).
• **Metrische Skala (Kardinalskala, Intervallskala, Ratioskala)**:
Kardinalskala: Messe Vielfache einer Grundeinheit (analog Messtab).
Intervallskala: Nullpunkt willkürlich. Dann können Quotienten nicht interpretiert werden (Temperatur).
Verhältnis-, Quotienten- o. Ratioskala: Nullpunkt physikalisch zwingend (Längen, Gewichte, Geld, Anzahlen)

ACHTUNG:

• Daten sind oft durch Zahlen kodiert. Dies heißt noch lange nicht, dass Rechenoperationen sinnvoll sind.

• Welche Rechenoperationen und statistischen Verfahren sinnvoll angewendet werden können, hängt oft vom Skalenniveau der Daten ab.

Ziele: Tabellarische und grafische Aufbereitung von Zahlenmaterial.
Ausgangspunkt: **Rohdaten** (**Primärdaten**, **Urliste**) nach der Erhebung.
Allgemeine Situation: Erhebe p Merkmale an n statistischen Einheiten.

Darstellung der Daten in der **Datenmatrix** (Tabelle):

stat. Einheit Nr.	Geschlecht	Alter	Größe	Messwert
1	M	18	72.6	10.2
2	W	21	18.7	9.5
⋮				⋮
n	W	19	15.6	5.6

i-te Zeile: Werte der p Variablen für die i-ten statistischen Einheit (Beob.)
j-te Spalte: Stichprobe der n beobachteten Werte des j-ten Merkmals.

Zeilen = Beobachtungen, Spalten = Variablen

Selektiere Spalte:
$\rightarrow$ Stichprobe $x_1, \ldots, x_n$,
$\rightarrow$ Datenvektor $\mathbf{x} = (x_1, \ldots, x_n)'$.

Aufgabe: Visualisierung von Zahlenmaterial:

Prinzip der Flächentreue: Sollen Zahlen grafisch durch Flächenelemente visualisiert werden, so müssen die Flächen proportional zu den Zahlen gewählt werden.
Grund: Gehirn spricht auf Fläche an, nicht auf Höhe oder Breite eines grafischen Elements.

Beispiel: Visualisierung durch Kreisflächen.

$$F = \pi r^2$$

r : Radius, F : Fläche.
Man muss die Radii proportional zur Wurzel der Zahlen wählen.

Nominale/ordinale Daten:
Zähle aus, wie oft die Ausprägungen im Datensatz vorkommen.
Nominales Merkmal mit den Ausprägungen $a_1, \ldots, a_k$
Die **absoluten Häufigkeiten** (engl.: *frequencies, counts*) $h_1, \ldots, h_k$, sind durch

$$\begin{aligned} h_j &= \text{Anzahl der } x_i \text{ mit } x_i = a_j \\ &= \sum_{i=1}^{n} \mathbf{1}(x_i = a_j), \end{aligned}$$

$j = 1, \ldots, k$ gegeben. Die (tabellarische) Zusammenstellung der absoluten Häufigkeiten $h_1, \ldots, h_k$ heißt **absolute Häufigkeitsverteilung.**
Es gilt:

$$n = h_1 + \cdots + h_k.$$

Dividiert man die absoluten Häufigkeiten durch den Stichprobenumfang n, so erhält man die **relativen Häufigkeiten** $f_1, \ldots, f_k$. Für $j = 1, \ldots, k$ berechnet sich f_j durch

$$f_j = \frac{h_j}{n}.$$

f_j ist der Anteil der Beobachtungen, die den Wert a_j haben.

Die (tabellarische) Zusammenstellung der $f_1, \ldots, f_k$ heißt **relative Häufigkeitsverteilung.**

Die relativen Häufigkeiten summieren sich zu 1 auf: $f_1 + \cdots + f_k = 1$.

Darstellung durch Stab-, Balken- oder Kreisdiagramme.

Kreisdiagramm (Kuchendiagramm): Die Winkelsumme von 360° (Gradmaß) bzw. 2π (Bogenmaß) wird entsprechend den absoluten oder relativen Häufigkeiten aufgeteilt.

Zu einer relativen Häufigkeit f_i gehört also der Winkel

$$\varphi_i = \frac{h_i}{n} \cdot 360^\circ = 2\pi f_i [\text{rad}].$$

$\rightarrow$ Ordinales Merkmal: Ordne die Stäbe, Balken oder Kreissegmente entsprechend der Anordnung der Ausprägungen an.

Tipp: Zum Erkennen von Zusammenhängen mit einem anderen Merkmal Y ordne die Stäbe, Balken oder Kreissegmente nach dem anderen Merkmal Y an! (s. Beispiel zu Öleinnahmen und BIP im Buch).

Die sortierten Beobachtungen werden mit $x_{(1)}, \ldots, x_{(n)}$ bezeichnet. Die Klammer um den Index deutet somit den Sortiervorgang an. Es gilt:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

$x_{(i)}$ heißt *i*-**te Ordnungsstatistik**,
$(x_{(1)}, \ldots, x_{(n)})$ heißt **Ordnungsstatistik** der Stichprobe $x_1, \ldots, x_n$.
Das **Minimum** $x_{(1)}$ wird auch mit $x_{\min}$ bezeichnet, das **Maximum** $x_{(n)}$ entsprechend mit $x_{\max}$.

Tipp: Für wenig Daten: Markiere die Beobachtungen x_i auf der reellen Zahlenachse und schreibe jeweils x_i drüber. Dann hat man von links nach rechts die Ordnungsstatistik und zugleich die Zuordnung zu den Ausgangsdaten $x_1, \ldots, x_n$. Zudem erkennt man, in welchen Bereichen sich die Daten häufen.

Messbereich (range): $[x_{\min}, x_{\max}]$ (kleinste Intervall, das alle Daten enthält).

Gruppierung (Klassierung) von Daten:
Lege k Intervalle

$$I_1 = [g_1, g_2],\ I_2 = (g_2, g_3],\ \ldots,\ I_k = (g_k, g_{k+1}],$$

fest, welche den Messbereich überdecken.
I_j heißt j-te **Gruppe** oder **Klasse** und ist für $j = 2, \ldots, k$ gegeben durch $I_j = (g_j, g_{j+1}]$. Die Zahlen $g_1, \ldots, g_{k+1}$ heißen **Gruppengrenzen**. Des Weiteren führen wir noch die k **Gruppenbreiten**

$$b_j = g_{j+1} - g_j, \qquad j = 1, \ldots, k,$$

und die k **Gruppenmitten**

$$m_j = \frac{g_{j+1} + g_j}{2}, \qquad j = 1, \ldots, k,$$

ein.

Histogramm:
Das Histogramm ist eine grafische Darstellung der relativen Häufigkeitsverteilung, die dem Prinzip der Flächentreue folgt.

1. Gruppiere in k Klassen mit Gruppengrenze $g_1 < \cdots < g_{k+1}$.
2. Berechne zugehörige relative Häufigkeiten $f_1, \ldots, f_k$.
3. Zeichne über Gruppe j ein Rechteck der Fläche f_j

Hierzu bestimmen wir die Höhe l_j des j-ten Rechtecks so, dass die Fläche $F_j = b_j l_j$ des Rechtecks der relativen Häufigkeit f_j entspricht:

$$F_j = b_j l_j \stackrel{!}{=} f_j \qquad \Rightarrow \qquad l_j = \frac{f_j}{b_j}, \qquad j = 1, \ldots, k.$$

Beispiel: Histogramm von $n = 30$ Leistungsdaten der Solarmodule.

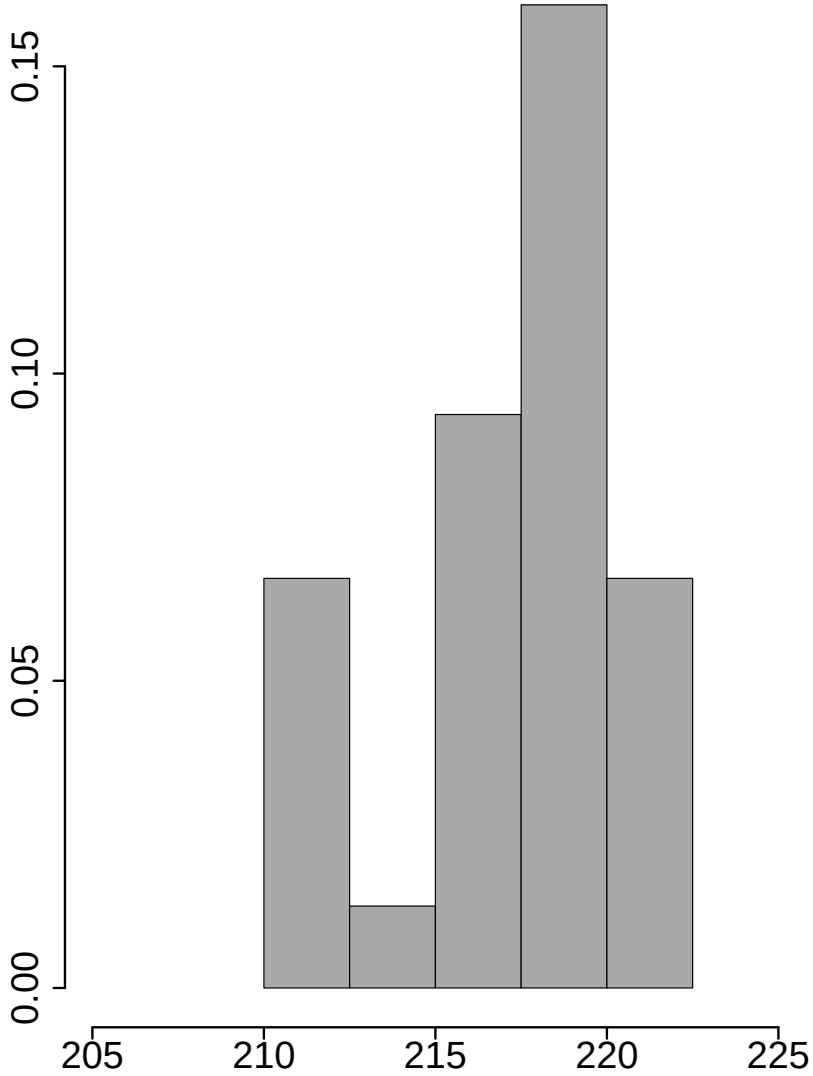
214.50	210.07	219.75	210.48	217.93	217.97	217.07	219.05
218.43	217.69	217.19	220.42	217.60	222.01	219.58	217.87
212.38	222.44	219.72	217.99	217.87	221.96	210.42	217.48
211.61	217.40	216.78	216.11	217.03	222.08		

Wir wählen 5 äquidistante Gruppen der Breite 2.5.
Mit den $k = 6$ Gruppengrenzen

$$g_1 = 210,\ g_2 = 212.5, \ldots,\ g_6 = 222.5$$

erhält man folgende Arbeitstabelle:

j	I_j	h_j	f_j	l_j
1	[210.0,212.5]	5	0.167	0.067
2	(212.5,215.0]	1	0.033	0.013
3	(215.0,217.5]	7	0.233	0.093
4	(217.5,220.0]	12	0.400	0.160
5	(220.0,222.5]	5	0.167	0.067


0.00
0.05
0.10
0.15
205
210
215
220
225

Der obere Rand des Histogramms definiert eine Treppenfunktion $\widehat{f}(x)$, die über dem j-ten Intervall I_j der Gruppeneinteilung den konstanten Funktionswert l_j annimmt. Außerhalb der Gruppeneinteilung setzt man $\widehat{f}(x)$ auf 0.

$$\widehat{f}(x) = \begin{cases} 0, & x < g_1, \\ l_1, & x \in [g_1, g_2], \\ l_j, & x \in (g_j, g_{j+1}],\ j = 2, \ldots, k, \\ 0, & x > g_{k+1}. \end{cases}$$

$\widehat{f}(x)$ heißt **Häufigkeitsdiche** oder auch **Dichteschätzer**.
$\rightarrow$ Die aus dem Histogramm abgeleitete Häufigkeitsdichte ist ein Schätzer für die Wahrscheinlichkeitsdichte $f(x)$ des Merkmals.

Die Häufigkeitsdichte ist selbst eine Wahrscheinlichkeitsdichte:
a) $\widehat{f}(x) \geq 0$ für alle $x \in \mathbb{R}$.
b) Für $x \in (g_j, g_{j+1}]$ ist sie konstant mit Wert

$$\widehat{f}(x) = l_j = \frac{f_j}{g_{j+1} - g_j}$$

so dass

$$\int_{g_j}^{g_{j+1}} \widehat{f}(x)dx = (g_{j+1} - g_j)\widehat{f}(x) = f_j.$$

Summation über j liefert daher den Wert 1 und somit

$$\begin{aligned}\int_{-\infty}^{\infty} \widehat{f}(x)\,dx &= \int_{g_1}^{g_{k+1}} \widehat{f}(x)\,dx \\ &= \int_{g_1}^{g_2} \widehat{f}(x)\,dx + \cdots + \int_{g_k}^{g_{k+1}} \widehat{f}(x)\,dx \\ &= \sum_{j=1}^{k} f_j = 1.\end{aligned}$$

Quantifizierung der Gestalt empirischer Verteilungen

Ziel

Beschreibe Zentrum der Daten, um das die Zahlen streuen.

Beispiel-Datensatz: Ozonkonzentration in 1000 [ppm]

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	66	52	49	64	68	26	86	52	43	75	87	188	118

Ordinal skalierte Daten

Definition

x_{med} heißt **Median** von $x_1, \ldots, x_n$, wenn

- mind. 50 % der Daten kleiner oder gleich x_{med} sind *und*
- mind. 50 % der Daten großer oder gleich x_{med} sind.

Median

Berechnung

- n ungerade: $x_{\text{med}} = x_{(k)}$, $k = \frac{n+1}{2}$.
- n gerade: Jede Zahl des Intervalls $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$.

Median

Konvention (metrisch skalierte Daten)

$$x_{\text{med}} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ ungerade}, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & n \text{ gerade}. \end{cases}$$

Beispiel: Median

Beispiel

Sortiere die Daten...

$$26 \quad 43 \quad 49 \quad 52 \quad 52 \quad 64 \quad 66 \quad 68 \quad 75 \quad 86 \quad 87 \quad 118 \quad 188$$

Der Median dieser 13 *Messungen ist der* 7*-te Wert,* $x_{(7)} = 66$*, der sortierten Messungen.*

Median: Eigenschaften

Eigenschaften

- Vollzieht affin-lineare Transf. nach (Umrechnung von Einheiten!)

$$y_i = a + b \cdot x_i, \qquad i = 1, \ldots, n.$$

  Dann: $y_{\text{med}} = a + b \cdot x_{\text{med}}$.

- Vollzieht monotone Transformationen $f(x)$ nach:

$$y_i = f(x_i), \qquad i = 1, \ldots, n.$$

  Dann gilt: $y_{\text{med}} = f(x_{\text{med}})$.

- x_{med} minimiert $Q(m) = \sum_{i=1}^{n} |x_i - m|$.

Metrische skalierte Daten

Definition

Die Kennzahl

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

heißt **arithmetisches Mittel** oder **arithmetischer Mittelwert**.

Gruppierte Daten:

- $f_1, \ldots, f_k$: rel. Hf.
- $m_1, \ldots, m_k$: Gruppenmitten

Dann verwendet man:

$$\overline{x}_g = f_1 m_1 + \cdots + f_k m_k$$

Beispiel: Ozondaten

Beispiel

Für die Ozondaten erhält man:

$$\sum_{i=1}^{n} x_i = 66 + 52 + 49 + 64 + 68 + 26 + 86 + 52 + 43 + 75 + 87 + 188 + 118$$
$$= 974$$

und hieraus $\overline{x} = \frac{974}{13} = 74.923$.

Eigenschaften

Eigenschaften

- Schwerpunkteigenschaft
- Hochrechnung
- Verhalten unter affin-linearen Transformationen
- $\overline{x}$ minimiert $Q(m) = \sum_{i=1}^{n}(x_i - m)^2$.

Minimierungseigenschaft

$\overline{x}$ minimiert $Q(m) = \sum_{i=1}^{n}(x_i - m)^2$, $m \in \mathbb{R}$:

Ableitung von $(x_i - m)^2$ nach m: $2(x_i - m) \cdot (-1) = -2(x_i - m)$.

Ableitungen von $Q(m)$: Für alle $m \in \mathbb{R}$ gilt:

$$\begin{aligned} Q'(m) &= -2\sum_{i=1}^{n}(x_i - m) \\ &= -2\sum_{i=1}^{n} x_i + 2 \cdot n \cdot m, \\ Q''(m) &= 2n > 0 \end{aligned}$$

Nullsetzen der 1. Ableitung:

$$Q'(\widehat{m}) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad 2n\widehat{m} = 2\sum_{i=1}^{n} x_i \quad \Leftrightarrow \quad \underline{\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Robustheit

Median oder arithmetisches Mittel?

- 9 arme Bauern: (Einkommen (in Euro): 1000) und 1 Reicher: (20000)
- $\overline{x} = (9 \cdot 1000 + 20000)/10 = 2900.$

Der Reiche ist ein **Ausreißer**. $\overline{x}$ reagiert sehr empfindlich auf solche Ausreißer!

- $x_{\text{med}} = 1000$ (Median-Einkommen)

Nominale/ordinale Daten

Streuung kategorialer Daten

Ausgangspunkt: relative Häufigkeitsverteilung

1	2	$\cdots$	k
f_1	f_2	$\cdots$	f_k

Nominale/ordinale Daten

Streuung kategorialer Daten

Keine Streuung:

1	2	$\cdots$	k
?	?	$\cdots$	?

Streuung kategorialer Daten

Keine Streuung: Nur eine Kategorie besetzt, z.B.:

1	2	$\cdots$	k
1	0	$\cdots$	0

Nominale/ordinale Daten

Streuung kategorialer Daten

Maximale Streuung:

1	2	$\cdots$	k
?	?	$\ldots$	?

Streuung kategorialer Daten

Maximale Streuung: Alle Kategorien gleich stark besetzt, d.h.:

1	2	$\cdots$	k
$\frac{1}{k}$	$\frac{1}{k}$	$\ldots$	$\frac{1}{k}$

Entropie

Betrachte: Gleichverteilung auf $r \leq k$ Kategorien $\rightarrow f_j = 1/r$

Anzahl r misst Streuung. In Binärdarstellung $001, 010, \ldots$ benötigte Bits:

$$b = \log_2(r) = -\log_2(1/r) = -\log_2(f_j)$$

Umlegen auf r Kategorien:

$$-\frac{1}{r}\log_2\left(\frac{1}{r}\right) = -f_j \log_2(f_j)$$

Erinnerung: Umrechnung Logarithmen: $\log_a(x) = \log_b(x) \cdot \log_a(b)$

Entropie

Definition

Die Kennzahl

$$H = -\sum_{j=1}^{k} f_j \log(f_j)$$

heißt **Shannon-Wieder-Index** oder **Shannon-Entropie**.

$$J = \frac{H}{\log(k)}$$

heißt **relative Entropie**

Eigenschaften

Eigenschaften

- $0 \leq H \leq \log(k)$
- $0 \leq J \leq 1$
- Minimalwert: 1-Punkt-Verteilung
- Maximalwert: Gleichverteilung auf k Kategorien

Metrisch skalierte Daten

Datenvektor: $\mathbf{x} = (x_1, \ldots, x_n)$

Definition

Stichprobenvarianz (empirische Varianz):

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

Bei gruppierten Daten:

$$s_g^2 = \sum_{j=1}^{k} f_j (m_j - \overline{x}_g)^2$$

Standardabweichung $s = \sqrt{s^2}$

Eigenschaften

Maßstabsänderung von Datenvektoren $\mathbf{x} = (x_1, \ldots, x_n)$

$$b \cdot \mathbf{x} = (b \cdot x_1, \ldots, b \cdot x_n)$$

Lageänderung

$$\mathbf{x} + a = (x_1 + a, \ldots, x_n + a)$$

Rechenregeln

- Invarianz unter Lageänderung

$$\text{var}(a + \mathbf{x}) = \text{var}(\mathbf{x})$$

- Quadratische Reaktion auf Maßstabsänderung

$$\text{var}(b \cdot \mathbf{x}) = b^2 \cdot \text{var}(\mathbf{x})$$

Verschiebungssatz

Verschiebungssatz

Es gilt:

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

sowie

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

Was macht die Praxis?

Praxis:

In der Praxis wird die Formel

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

verwendet. (Begründung in der LV *Statistik*).

Quantile

Beispiel

PC-Händler bestellt monatlich TFT-Monitore. In 9 von 10 Fällen soll die Lieferung bis zum Monatsende reichen.

Ansatz: Daten $x_{(1)} \leq \cdots \leq x_{(9)} \leq x_{(10)}$.
Für jede Zahl $x \in [x_{(9)}, x_{(10)}]$ gilt:

- Mindestens $9/10$ der x_i sind $\leq x$ und
- mindestens $1/10$ der x_i sind $\geq x$

(Für jedes $x \in (x_{(9)}, x_{(10)})$ gilt: Genau $9/10$ sind $\leq x$ und genau $1/10$ sind $\geq x$).

Quantile

Definition

Ein **empirisches** p-**Quantil**, $p \in (0,1)$, von $x_1, \ldots, x_n$ ist jede Zahl $\widetilde{x}_p$, so dass

- mindestens $100 \cdot p\%$ der Datenpunkte sind $\leq \widetilde{x}_p$ *und*
- mindestens $100 \cdot (1-p)\%$ der Datenpunkte sind $\geq \widetilde{x}_p$ *und*

Quantile

Berechnung

- np ganzzahlig: Jede Zahl aus $[x_{(np)}, x_{(np+1)}]$.
  Nicht immer ist das Merkmal metrisch skaliert. Dann sind mitunter nur bestimmte x-Werte interpretierbar, nicht jedoch 'Zwischenwerte'. Dann sind (nur) $x_{(np)}$ und $x_{(np+1)}$ Quantile.
- sonst: $\widetilde{x}_p = x_{(\lfloor np \rfloor + 1)}$.

Hierbei ist

$$\lfloor x \rfloor$$

die **Abrundung** einer Zahl $x \in \mathbb{R}$.

Quantile

Für metrische skalierte Daten gibt es verschiedene Konventionen, um die Defintion des Quantils eindeutig zu machen. Zum Beispiel:

Konvention: Intervallmitte: $\frac{1}{2}(x_{(np)} + x_{(np+1)})$

Quartile

Quartile:

$Q_1 = \widetilde{x}_{0,25}$: unteres Quartile (grenzt das untere Viertel ab)
$Q_2 = \widetilde{x}_{0,5}$: Median (grenzt die untere Hälfte ab, teilt die Verteilung)
$Q_3 = \widetilde{x}_{0,75}$: oberes Quartil (grenzt das obere Viertel ab).

Zwischen Q_1 und Q_3 liegen die zentralen 50% der Datenpunkte (die Mitte)!

$Q_3 - Q_1$ heißt **Interquartilabstand (IQR)** und ist ein robustes Streuungsmaß.

Beispiel

Fünfpunkte-Zusammenfassung und Boxplot
Die 5 Statistiken (Kennzahlen) $x_{\min}, Q_1, \widetilde{x}_{0.5} = x_{med}, Q_3, x_{\max}$ heißt **Fünfpunkte-Zusammenfassung**.

Boxplot: Grafische Darstellung der 5-Punkte-Zusammenfassung:

EAS 2023 - Inferenzstatistik

Prof. Dr. Ansgar Steland

2021

Agenda

- Einstieg in die Statistische Inferenz
- Parametrische Modelle
- Schätzverfahren

Wozu schließende Statistik?

Beispiel

Sind User der Spielekonsole __ zufrieden?

- *Umfrage unter* $n = 500$ *zufällig ausgewählten registrierten Usern.*
- $k = 400$ *sind mit ihrer Konsole zufrieden.*

Sind diese Zahlen belastbar*?*

1. *Ist der Anteil von* $k/n = 80\%$ *zufriedenen Nutzern eine gute Schätzung des wahren Anteils in der Grundgesamtheit?*
2. *Wie stark streut das Stichprobenergebnis? Wie sicher ist die Schätzung?*
3. *Wie kann objektiv nachgewiesen werden, dass der wahre Anteil zufriedener User zumindest höher als (z. B.) 75% ist?*

Wege zur Lösung...

1. Finde geeignetes Verteilungsmodell für die Daten.
   Hier: Bin(n, p), p unbekannt.
2. Wie kann man p (optimal?) aus den Daten schätzen?
3. Wie kann man die Hypothese $p > 0.75$ nachweisen?

Wahrscheinlichkeitsrechnung:

- Liefert Regeln, wie man mit Wahrscheinlichkeiten und Verteilungen rechnet.
- Gegeben: Stochastisches Modell $X \sim F$.
  Oft: $F = F_\vartheta$ (parametrisiert durch ϑ).
- F wird (gedanklich) als bekannt/gegeben angenommen.
- Bsp: $X \sim N(\mu, \sigma^2) \Rightarrow P(X \leq 2) = \Phi((2-\mu)/\sigma)$.
  Liefert eine Formel, die von $\vartheta = (\mu, \sigma^2)$ abhängt.
  Einsetzen spezieller Werte, z.B. $\vartheta = (4, 2)$, liefert eine konkrete Zahl.

Schließende Statistik:

- Gegeben: Verrauschte (zufallsbehaftete) Daten $X_1, \ldots, X_n \sim F_\vartheta$.
- Gesucht: Das Modell F_ϑ, also ϑ.
- Ziel: Schließe aus den Daten auf das zugrunde liegende Modell.
- Relevant Schritte:
  1. Gute Modellklasse für die Daten finden. — Modellierung
  2. Schätzen des Modells aus den Daten. — Schätzen
  3. Testen: Gilt $\vartheta \in \Theta_0$ oder $\vartheta \in \Theta_1$? — Testen
  4. Untersuche, ob das Modell die Daten gut erklärt. — Modellvalidierung

Grundbegriffe: Stichprobe

Stichprobe

$X_1, \ldots, X_n$ heißt **Stichprobe** vom **Stichprobenumfang** n, wenn

$$X_1, \ldots, X_n : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$$

Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ sind.
Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_n)$ nimmt Werte im **Stichprobenraum**

$$\mathcal{X} = \{\mathbf{X}(\omega) : \omega \in \Omega\} \subset \mathbb{R}^n$$

an. Realisierungen: Vektoren $(x_1, \ldots, x_n) \in \mathcal{X}$.

Hinweis

In der Statistik interessiert i.d.R. der zugrunde liegende W-Raum $(\Omega, \mathcal{A}, P)$ nicht, sondern lediglich der Stichprobenraum $\mathcal{X}$ und die Verteilung $P_{\mathbf{X}}$ von $\mathbf{X} = (X_1, \ldots, X_n)'$ hierauf!

Grundbegriffe: Verteilungsmodell

Verteilungsmodell

Eine Menge $\mathcal{P}$ von (möglichen) Verteilungen auf $\mathbb{R}^n$ (für die Stichprobe $(X_1, \ldots, X_n)$) heißt **Verteilungsmodell**.

$\mathcal{P}$ heißt **parametrisches Verteilungsmodell**, falls

$$\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$$

für eine Menge $\Theta \subset \mathbb{R}^k$ von Parametervektoren.

Θ: **Parameterraum**.
D.h.: Es gibt eine Bijektion $\mathcal{P} \leftrightarrow \Theta$.

Ein Verteilungsmodell, das nicht durch einen endlichdimensionalen Parameter parametrisiert werden kann, heißt **nichtparametrisches Verteilungsmodell**.

Beispiele:

Beispiel

Parametrische Verteilungsmodelle:

1). $\mathcal{P} = \{Bin(n,p) : p \in [0,1]\}$ *für ein festes* n.
Parameter: $\vartheta = p \in \Theta = [0,1]$.

2). $\mathcal{P} = \{N(\mu,\sigma^2) : \mu \in \mathbb{R}, 0 < \sigma^2 < \infty\}$. *Parameter:*
$\vartheta = (\mu,\sigma^2) \in \Theta = \mathbb{R} \times (0,\infty)$.

3). *Sei* $Y = g_{net}(\mathbf{X})$ *mit* $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_p)$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $p \in \mathbb{N}$
$y = g_{net}(\mathbf{x})$ *Deep Learner mit Netzparametern* $\mathbf{w} \in \mathbf{W} \subset \mathbb{R}^q$, $q \in \mathbb{N}$.
Bezeichne $G_{(\boldsymbol{\mu},\mathbf{w})}(y)$ *die Verteilungsfunktion von* Y *bei Input* $\mathbf{X}$.
$\mathcal{P} = \{G_{(\boldsymbol{\mu},\mathbf{w})} : \boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{w} \in \mathbf{W}\}$ *Menge möglicher Verteilungen für* Y.
Parameter: $\boldsymbol{\vartheta} = (\boldsymbol{\mu},\boldsymbol{\theta}) \in \Theta = \mathbb{R} \times (0,\infty)$.

Nichtparametrische Verteilungsmodelle:

4). $\mathcal{P} = \{F : \mathbb{R} \to [0,1] : F$ *ist Verteilungsfunktion*$\}$

5). $\mathcal{P} = \{f : \mathbb{R} \to \mathbb{R}^+ : f$ *ist stetige Dichtefunktion*$\}$

Statistik, Schätzfunktion, Schätzer

Statistik,...

Sei $X_1, \ldots, X_n$ eine Stichprobe (und o.E. $\mathcal{X} = \mathbb{R}^n$).
Eine Abbildung

$$T : \mathbb{R}^n \to \mathbb{R}^d$$

mit $d \in \mathbb{N}$ (oft: $d = 1$) heißt **Statistik**.

Bildet T in den Parameterraum ab, d.h.

$$T : \mathbb{R}^n \to \Theta,$$

dann heißt T **Schätzfunktion** oder kürzer **Schätzer** (für ϑ).

Allgemein: Schätzung von Funktionen $g(\vartheta)$ von ϑ durch Statistiken $T : \mathbb{R}^n \to \Gamma$ mit $\Gamma = g(\Theta) = \{g(\vartheta) | \vartheta \in \Theta\}$.

Beispiele

Beispiel: Seien $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, und

$$T_1(X_1, \ldots, X_n) = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ ,$$

$$T_2(X_1, \ldots, X_n) = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 .$$

$T_1(X_1, \ldots, X_n)$ bildet in den Parameterraum $\Theta_1 = \mathbb{R}$ für μ ab und ist daher eine Schätzfunktion für μ.
$T_1(X_1, \ldots, X_n)$ bildet in den Parameterraum $\Theta_2 = (0, \infty)$ von σ^2 ab und ist daher eine Schätzfunktion für σ^2.
Standard-Notation: Ist $T : \mathbb{R}^n \to \Theta$ ein Schätzer für ϑ, dann schreibt man

$$\widehat{\vartheta} = T(X_1, \ldots, X_n)$$

zu schreiben. Analog: $\widehat{F}_n(x)$ ist Schätzer für $F(x)$, etc.

Empirische Verteilungsfunktion

Allgemeinstes nichtparametrisches Modell:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x)$$

mit beliebiger Verteilungsfunktion

$$F(x) = P(X_1 \leq x), \qquad x \in \mathbb{R}.$$

Frage:

1. Wie kann man $F(x)$ ohne zusätzliche Annahmen schätzen?
2. Wie kann man einen solchen Schätzer $\widehat{F}(x)$ rechtfertigen?

Empirische Verteilungsfunktion

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i), \qquad x \in \mathbb{R}.$$

Hierbei: $\mathbf{1}_{(-\infty,x]}(X_i) = \mathbf{1}(X_i \leq x)$.

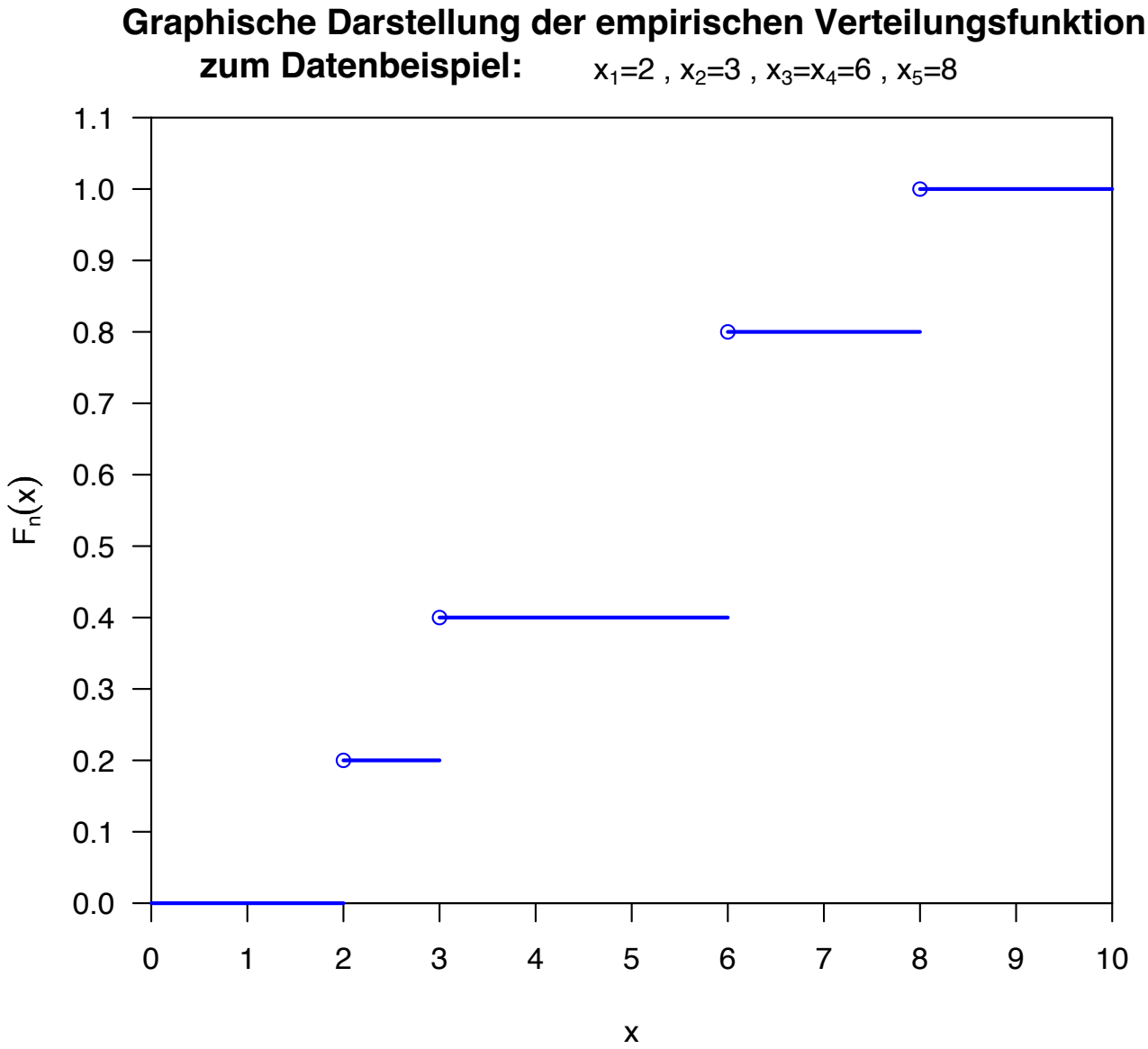
$\widehat{F}_n(x)$: Anteil der Beobachtungen, die kleiner oder gleich x sind.

1. Die Anzahl $n\widehat{F}_n(x)$ der Beobachtungen $\leq x$ ist binomialverteilt mit Parametern n und $p(x) = E(\mathbf{1}(X_i \leq x)) = F(x)$.
2. Daher folgt:

$$E(\widehat{F}_n(x)) = P(X_i \leq x) = F(x), \quad \text{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

3. Nach dem Hauptsatz der Statistik konvergiert $\widehat{F}_n(x)$ mit Wahrscheinlichkeit 1 gegen $F(x)$ (gleichmäßig in x).

Empirische Verteilungsfunktion



Empirische Verteilungsfunktion

Sehr viele Statistiken leiten sich von der empirischen Verteilungsfunktion ab, z.B.:

- Arithmetisches Mittel $\overline{X}_n$.
- Stichprobenvarianz S^2.
- Empirisches Quantil.

(da die Funktion $\widehat{F}(x)$ die geordnete Stichprobe kodiert).

Dichteschätzung

Nichtparametrisches Verteilungsmodell:

$$X_1, \ldots, X_n \sim f(x)$$

mit einer Dichtefunktion $f(x)$.
Mögliche Schätzer:

- Histogramm-Schätzer (schätzt eine Vergröberung der Dichte).
- Kerndichteschätzer $\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x-X_i}{h})$,
  $K : \mathbb{R} \to [0,1]$, $h > 0$ Bandbreite. $K(z) = \frac{1}{2}\mathbf{1}_{[-1,1]}(z)$ liefert das gleitende Histogramm ($\widehat{f}_n(x)$: Anteil der Beob. in $[x-h, x+h]$).

Das Likelihood-Prinzip

Wichtiges Schätzprinzip der parametrischen Statistik.

Motivation:

Information:

1. Ein Restaurant hat zwei Köche A und B.
2. Koch A versalzt die Suppe mit Wkeit 0.1.
3. Koch B versalzt die Suppe mit Wkeit 0.3.

Sie gehen ins Restaurant und essen eine Suppe. Die Suppe ist versalzen.
Wer war der Koch?

Formalisierung

Wir beobachten $x \in \{0, 1\}$. (1: Suppe versalzen, 0: nicht versalzen).
Parameter: $\vartheta \in \Theta = \{A, B\}$ (der wahre Koch).
Statistisches Problem: Schätze ϑ bei Vorliegen der Beobachtung x.
Jeder Koch erzeugt eine W-Verteilung auf $\mathcal{X} = \{0, 1\}$.

$\diagdown\, p_\vartheta(x)$	Beobachtung		
$\vartheta \diagdown$	0	1	Summe
A	0.9	0.1	1.0
B	0.7	0.3	1.0

Lösungsheuristik: ϑ umso plausibler, größer $p_\vartheta(x)$ ist.

Likelihood-Funktion

Likelihood-Funktion

Sei $p_\vartheta(x)$ eine Zähldichte (in $x \in \mathcal{X}$) und $\vartheta \in \Theta$ ein Parameter.
Für eine gegebene (feste) Beobachtung $x \in \mathcal{X}$ heißt die Funktion

$$L(\vartheta|x) = p_\vartheta(x), \qquad \vartheta \in \Theta,$$

Likelihood-Funktion.

Likelihood-Prinzip

Ein Verteilungsmodell ist bei gegebenen Daten plausibel, wenn es die Daten mit hoher Wahrscheinlichkeit erzeugt. Entscheide Dich für das plausibelste Verteilungsmodell!

Verallgemeinerung

Situation 1:
Diskreter Parameterraum $\Theta = \{\vartheta_1, \ldots, \vartheta_L\}$.
Diskreter Stichprobenraum $\mathcal{X} = \{x_1, \ldots, x_K\}$.

	x_1	$\ldots$	x_K	Summe
ϑ_1	$p_{\vartheta_1}(x_1)$	$\ldots$	$p_{\vartheta_1}(x_K)$	1
ϑ_2	$p_{\vartheta_2}(x_1)$	$\ldots$	$p_{\vartheta_2}(x_K)$	1
$\vdots$	$\vdots$		$\vdots$	
ϑ_L	$p_{\vartheta_L}(x_1)$	$\ldots$	$p_{\vartheta_L}(x_K)$	1

Algorithmus: Bestimme Spaltenmaximum für gegebene Beobachtung $x \in \{x_1, \ldots, x_K\}$.

Verallgemeinerung

Situation 2: (Standardfall bei diskreten Beobachtungen)
Parameterraum $\Theta \subset \mathbb{R}$ Intervall oder ganz $\mathbb{R}$
Diskreter Stichprobenraum: $\mathcal{X} = \{x_1, x_2, \dots\}$.
Keine Tabellendarstellungen mehr. Zeit für eine formale Definition:

Maximum-Likelihood-Schätzer

$p_\vartheta(x)$ sei Zähldichte (in $x \in \mathcal{X}$).
$\vartheta \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$ wie oben.
Dann heißt $\widehat{\vartheta} = \widehat{\vartheta}(x) \in \Theta$ **Maximum-Likelihood-Schätzer (ML-Schätzer)**, wenn für festes x gilt:

$$p_{\widehat{\vartheta}}(x) \geq p_\vartheta(x) \qquad \text{für alle } \vartheta \in \Theta.$$

(Falls Maximum nicht eindeutig, so wähle eines aus).
Hierdurch ist eine Funktion $\widehat{\vartheta} : \mathcal{X} \to \Theta$ definiert.

Maximum-Likelihood

- Also: Maximiere $(\vartheta, x) \mapsto p_\vartheta(x)$ für festes x in der Variablen $\vartheta \in \Theta$.
- Typischerweise ist $p_\vartheta(x)$ differenzierbar in ϑ.
- Wende bekannte Methoden zur Maximierung an.

Likelihood für Dichten

Problem: Was tun bei stetigen Variablen: $X \sim f_X(x)$?
Für alle $x \in \mathbb{R}$ gilt:

$$P(X = x) = 0$$

Wie kann man jetzt eine Likelihood-Funktion definieren?

Likelihood für Dichten

Idee:

1. Beobachtung x sei fest.
2. Vergröbere die Information 'x beobachtet' zu: 'ungefähr x beobachtet':
$$\{x\} \mapsto [x - dx, x + dx].$$
dx 'infinitesimal' klein.
3. Jetzt ist die Likelihood wie oben definiert:
$$L(\vartheta|[x - dx, x + dx]) = \int_{x-dx}^{x+dx} f_\vartheta(s)\, ds \approx f_\vartheta(x) \cdot (2dx).$$
4. Die rechte Seite wird maximiert, wenn $\vartheta \mapsto f_\vartheta(x)$ maximiert wird.

Likelihood für Dichten

$f_\vartheta(x)$ eine Dichtefunktion (in x), $\vartheta \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$.
Für festes x heißt die Funktion

$$L(\vartheta|x) = f_\vartheta(x), \qquad \vartheta \in \Theta,$$

Likelihood-Funktion. $\widehat{\vartheta} \in \Theta$ heißt **Maximum-Likelihood-Schätzer**, wenn bei festem x gilt:

$$f_{\widehat{\vartheta}}(x) \geq f_\vartheta(x), \quad \text{für alle } \vartheta \in \Theta.$$

Likelihood einer Stichprobe

Kompakt: $X \sim f_\vartheta(x)$, f_ϑ eine Zähldichte oder Dichtefunktion. Dann ist

$$L_\vartheta(x) = f_\vartheta(x)$$

Sei nun speziell $\mathbf{X} = (X_1, \ldots, X_n)'$ mit

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$$

Dann ist die gemeinsame (Zähl-) Dichte die Produkt-(Zähl-) Dichte.
Also:

$$L_\vartheta(\mathbf{x}) = f_\vartheta(x_1) \cdots f_\vartheta(x_n)$$

(Gilt für Zähldichten und Dichtefunktionen).

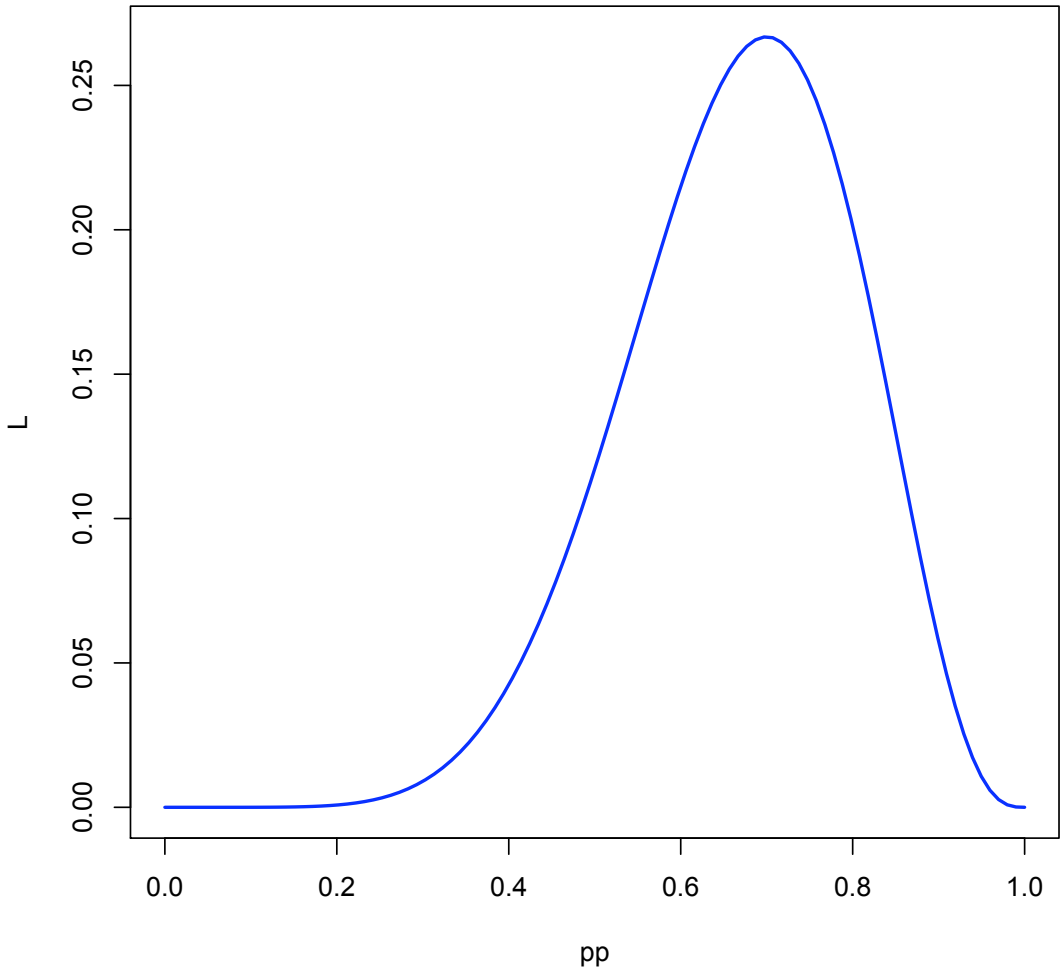
Beispiele

Beispiele...

Computerexperiment: Likelihood

```
# Likelihood der B(n,p)-Verteilung
# p: (Vektor der) Erfolgswahrscheinlichkeit(en)
# y: beobachtete Anzahl der Erfolge
# n: Stichprobenumfang

likeli = function( p, n, y ) {
  choose(n,y) * p^y * (1-p)^(n-y)
}

# Bsp: n = 10, y = 7 Erfolge

n = 10; y = 7
pp = seq( 0, 1, len=100 )
L = likeli( pp, n, y )
plot( pp, L, type="l", lwd=2, col="blue" )

# ML-Schä‰zer (numerisch im Intervall [0,1] mit max. Fehler 1e-10)

optimize( likeli, c(0,1), maximum=TRUE, tol=1e-10, n=10, y=7 )
```

Computerexperiment: Likelihood



Schätzer

Unterscheide (konzeptionell):

- Die Abbildung $\widehat{\vartheta}_n = \widehat{\vartheta}_n(x)$,

$$x = (x_1 \ldots, x_n) \mapsto \widehat{\vartheta}_n(x),$$

  die jeder Realisation x des Stichprobenraums $\mathcal{X}$ einen Schätzwert zuordnet; gedanklich nach Durchführung des Zufallsexperiments.
- Die Abbildung $\widehat{\Theta}_n = \widehat{\vartheta}_n(X)$,

$$X = (X_1, \ldots, X_n) \mapsto \widehat{\vartheta}_n(X),$$

  die jedem (zufälligen) Vektor X die Zufallsgröße $\widehat{\vartheta}_n(X)$ zuordnet; gedanklich vor Durchführung des Zufallsexperiments).

Es ist üblich, in beiden Fällen $\widehat{\vartheta}_n$ zu schreiben und von einem 'Schätzer' zu sprechen. Ob die Statistik (als Zufallsvariable bzw. Zufallsvektor) oder eine Realisation derselben gemeint ist, muss aus dem Kontext erschlossen werden.

Gütekriterien: Erwartungstreue

- Sei $\widehat{\vartheta}_n = \widehat{\vartheta}_n(X)$ ein Schätzer für ϑ.
- Da $\widehat{\vartheta}_n = T_n(X_1, \ldots, X_n)$ von den Zufallsvariablen $X_1, \ldots, X_n$ abhängt, ist $\widehat{\vartheta}_n$ zufällig, streut also.
- Frage: Um welchen Wert streut der Schätzer?
- Berechne den Erwartungswert:

$$E(\widehat{\vartheta}_n) = E(T_n(X_1, \ldots, X_n)) = ...?$$

- Das Ergebnis der Berechnung hängt von der Verteilung der $X_i \sim F_\vartheta$ ab! Um diese Abhängigkeit zum Ausdruck zu bringen schreibt man mitunter $E_\vartheta(\cdots)$ statt $E(\cdots)$.
- Im Allgemeinen ist $E(\widehat{\vartheta}_n)$ daher eine Funktion des Parameters ϑ!

Erwartungstreue

Ein Schätzer $\widehat{\vartheta}_n$ heißt **erwartungstreu für** ϑ, wenn für alle $\vartheta \in \Theta$ gilt:

$$E(\widehat{\vartheta}_n) = \vartheta$$

$g(\widehat{\vartheta}_n)$ heißt **erwartungstreu für** $g(\vartheta)$, wenn für alle $\vartheta \in \Theta$ gilt:

$$E(g(\widehat{\vartheta}_n)) = g(\vartheta)$$

Sinngemäß gelten diese Definitionen auch für nichtparametrische Modelle: T_n heißt erwartungstreu für eine Kenngröße $g(F)$, wenn $E(T_n) = E_F(T_n) = g(F)$ für alle Verteilungsfunktionen F der betrachteten Verteilungsklasse. Hierbei deutet $E_F(\cdot)$ an, dass der EW unter der Annahme $X_i \sim F$ berechnet wird.

Erwartungstreue

Beispiele: a) $X_1, \ldots, X_n$ seinen unabhängig und identisch verteilt mit Erwartungswert $\mu \in \mathbb{R}$. $\overline{X}_n$ ist erwartungstreu für μ.

b) Parameter: $\vartheta = g(\mu) = \mu^2$.
$g(\overline{X}_n) = (\overline{X}_n)^2$ ist nicht erwartungstreu für $\vartheta = g(\mu) = \mu^2$.

c) $X_1, \ldots, X_n \sim U(0, \vartheta)$ mit $\vartheta > 0$ unbekannt.

Der ML-Schätzer $\widehat{\vartheta}_n = \max_{i=1,\ldots,n} X_i$ für ϑ ist nicht erwartungstreu, aber der Schätzer

$$\widehat{\vartheta}_n^* = \frac{n+1}{n} \widehat{\vartheta}_n$$

Erwartungstreue: Anschauung

Anschauung:

- Wende erwartungstreuen Schätzer N Mal auf Stichproben vom Umfang n an.
- N Schätzungen: $\widehat{\vartheta}_n(1), \ldots, \widehat{\vartheta}_n(N)$.
- Wende Gesetz der großen Zahlen an!

$$\frac{1}{N} \sum_{i=1}^{N} \widehat{\vartheta}(i) \to E(\widehat{\vartheta}_n(i)) = E(\widehat{\vartheta}_n)$$

- $\widehat{\vartheta}_n$ erwartungstreu: rechte Seite ist ϑ unabhängig von $\vartheta \in \Theta$.
- sonst: rechte Seite $\neq \vartheta$.

Werden Schätzungen aus einer täglichen Stichprobe vom Umfang n über einen langen Zeitraum gemittelt, so schwankt dieses Mittel um $E(\widehat{\vartheta}_n)$. Bei einer erwartungstreuen Schätzfunktion also um den wahren Wert ϑ.

Bias (Verzerrung)

Verzerrung (Bias)

Die **Verzerrung** (engl.: *bias*) wird gemessen durch

$$\text{Bias}(\widehat{\vartheta}_n; \vartheta) = E_\vartheta(\widehat{\vartheta}) - \vartheta.$$

Beispiele

Beispiele:

$X_1, \ldots, X_n$ seinen unabhängig und identisch verteilt mit EW μ und Varianz $\sigma^2 > 0$.

Der Bias von $(\overline{X}_n)^2$ bzgl. des Parameters μ^2 ist:

$$\text{Bias}((\overline{X}_n)^2; \mu^2) = \frac{\sigma^2}{n}$$

Gütekriterien: Asymptotische Erwartungstreue

(Asymptotische) Erwartungstreue, Unverfälschtheit

Ein Schätzer $\widehat{\vartheta}_n$ für einen Parameter ϑ heißt **asymptotisch erwartungstreu für** ϑ, wenn für alle ϑ

$$E_\vartheta(\widehat{\vartheta}_n) \to \vartheta,$$

gilt.

Gütekriterien: Konsistenz

Das Gütekriterium der **Konsistenz** fragt danach, ob bei wachsendem Stichprobenumfang n die Wahrscheinlichkeit gegen 1 strebt, dass der Unterschied zwischen Schätzer $\widehat{\vartheta}_n$ und wahrem Wert ϑ kleiner als eine beliebig vorgegebene Toleranz $\delta > 0$ ist:
Für beliebiges $\delta > 0$ gilt:

$$P(|\widehat{\vartheta}_n - \vartheta| \leq \delta) \to 1, \qquad n \to \infty$$

oder gleichbedeutend:

$$P(|\widehat{\vartheta}_n - \vartheta| > \delta) \to 0, \qquad n \to \infty$$

Diese Eigenschaft entspricht der **stochastischen Konvergenz**:

$$\widehat{\vartheta}_n \xrightarrow{P} \vartheta, \qquad n \to \infty.$$

Gütekriterien: Konsistenz

Konsistenz

Ein Schätzer $\widehat{\vartheta}_n = T(X_1, \ldots, X_n)$ basierend auf einer Stichprobe vom Umfang n heißt **(schwach) konsistent für** ϑ, falls

$$\widehat{\vartheta}_n \xrightarrow{P} \vartheta, \qquad n \to \infty,$$

Gilt sogar fast sichere Konvergenz, dann heißt $\widehat{\vartheta}_n$ **stark konsistent für** ϑ.

1. Ist $\widehat{\vartheta}_n$ konsistent für ϑ und ist g stetig, dann ist $g(\widehat{\vartheta}_n)$ konsistent für den abgeleiteten Parameter $g(\vartheta)$.
2. Die obige Aussage gilt auch für vektorwertige Parameter und ihre Schätzer. Insbesondere folgt aus der Konsistenz von $\widehat{\vartheta}_n$ für ϑ und $\widehat{\xi}_n$ für ξ die Konsistenz von $\widehat{\vartheta}_n \pm \widehat{\xi}_n$ für $\vartheta \pm \xi$.

Konsistenz: Beispiele

1. $X_1, \ldots, X_n$ i.i.d. mit $\mu = E(X_1)$. Dann ist $\widehat{\mu}_n = \overline{X}_n$ konsistent für μ.
2. $g(\overline{X}_n) = (\overline{X}_n)^2$ ist konsistent für den abgeleiteten Parameter $g(\mu) = \mu^2$.
3. Gilt $E(X_1^2) < \infty$, dann folgt (starkes Gesetz der großen Zahlen):

$$\widehat{m}_{2,n} = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

ist (stark) konsistent für das zweite Moment $m_2 = E(X_1^2)$.
Dann ist auch die Stichprobenvarianz

$$\widehat{\sigma}_n^2 = \widehat{m}_{2,n} - \widehat{\mu}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2$$

konsistent für $\sigma^2 = E(X_1^2) - (E(X_1))^2 = \mathrm{Var}(X_1)$.

Konsistenz: Beispiele

Schätzung der Varianz σ^2: (s. Basiswissen, S. 192)
$X_1, \ldots, X_n$ einfache Zufallsstichprobe mit $\mu = E(X_1)$, $\sigma^2 = \mathrm{Var}(X_1) < \infty$.
Stichprobenvarianz:

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Dieser Schätzer ist konsistent aber nicht erwartungstreu:

$$E(\widehat{\sigma}_n^2) = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{\sigma^2}{n}$$

Im Mittel wird σ^2 unterschätzt. Man verwendet daher

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Dieser Schätzer ist konsistent und erwartungstreu für σ^2.

Effizienz und mittlerer quadratischer Fehler

- Mitunter stehen mehrere Schätzfunktionen zur Auswahl.
- Angenommen, alle sind erwartungstreu. Welche sollte man nehmen?

Effizienz

1. Sind T_1 und T_2 zwei erwartungstreue Schätzer für ϑ und gilt $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$, so heißt T_1 **effizienter** als T_2.
2. T_1 ist **effizient**, wenn T_1 effizienter als jede andere erwartungstreue Schätzfunktion ist.

Beispiele

Beispiel: $X_1, \ldots, X_n$ sei eine einfache Stichprobe. Betrachte

$$T_1 = \frac{X_1 + X_n}{2}, \qquad T_2 = \overline{X}_n.$$

Welche Schätzfunktion ist effizienter für die Schätzung von μ?

Beide Schätzfunktionen sind erwartungstreu: $E(T_2) = E(\overline{X}_n) = \mu$ und

$$E(T_1) = \frac{1}{2}E(X_1 + X_n) = \frac{1}{2}(E(X_1) + E(X_n)) = \frac{2\mu}{2} = \mu$$

Vergleich der Varianzen:

$$\mathrm{Var}(T_1) = \frac{\sigma^2}{2}, \qquad \mathrm{Var}(T_2) = \frac{\sigma^2}{n}$$

Für $n > 2$ ist T_2 effizienter als T_1.

Beispiele

Beispiel

Gelte $X_1, \ldots, X_n \sim G[0, \vartheta]$, $\vartheta > 0$ *unbekannt.*
Zwei erwartungstreue Schätzer für ϑ*:*

$$T_1 = 2\overline{X} \quad \textit{und} \quad T_2 = \frac{n+1}{n} \max(X_1, \ldots, X_n).$$

Welche Schätzfunktion ist effizienter?

MSE

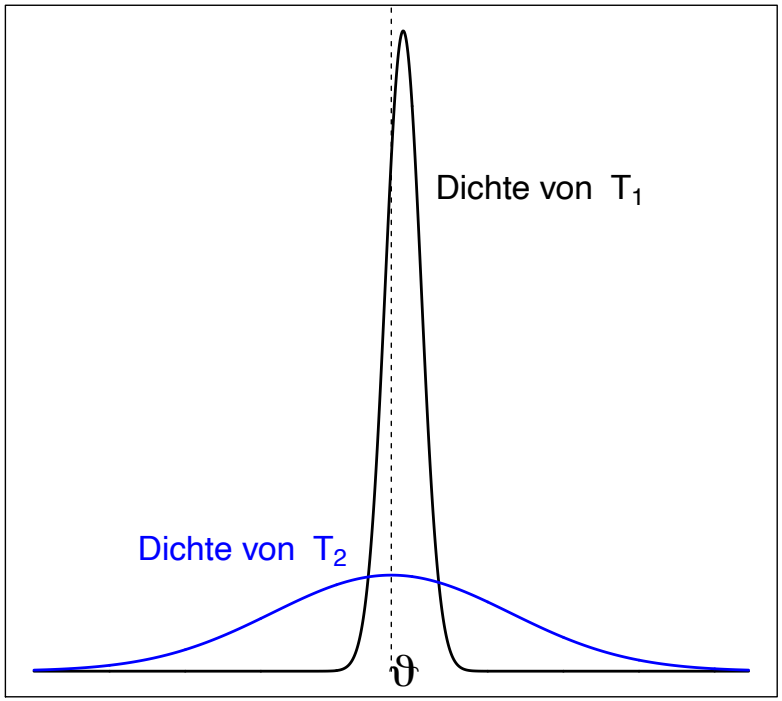


Abbildung: Dargestellt sind Dichten von zwei Schätzern zur Schätzung des Parameters ϑ. T_1 ist zwar verzerrt, hat aber eine viel kleinere Streuung.

MSE

MSE: Mean Squared Error

Der MSE ist das wichtigste Gütemaß für Bewertung und Vergleiche von Schätzern. Er integriert die Varianz (als Streuungsmaß) und den Bias in einer Kennzahl.

MSE

$$\text{MSE}(\widehat{\vartheta}_n; \vartheta) = E_\vartheta(\widehat{\vartheta}_n - \vartheta)^2$$

Additive Zerlegung

Ist $\widehat{\vartheta}_n$ eine Schätzfunktion mit $\text{Var}_\vartheta(\widehat{\vartheta}_n) < \infty$, dann gilt die additive Zerlegung

$$\text{MSE}(\widehat{\vartheta}_n; \vartheta) = \text{Var}_\vartheta(\widehat{\vartheta}_n) + [\text{Bias}(\widehat{\vartheta}_n; \vartheta)]^2.$$

MSE

MS-Effizienz

1. Sind T_1 und T_2 zwei Schätzer für ϑ und gilt $\mathrm{MSE}(T_1;\vartheta) < \mathrm{MSE}(T_2;\vartheta)$, so heißt T_1 **effizienter** als T_2.
2. T_1 ist **effizient**, wenn T_1 effizienter als jede andere erwartungstreue Schätzfunktion ist.

Beispiel: $X_1, \ldots, X_n \sim G(0, \vartheta)$. Effizienzvergleich[1] von

$$T_1 = 2\overline{X}_n, \qquad T_2 = \frac{n+1}{n} \max_{1 \le i \le n} X_i.$$

Schritt 1: Berechne $MSE(T_1; \vartheta)$:
Erwartungswert und Varianz von T_1:

$$E(T_1) = \vartheta, \qquad \text{Var}(T_1) = 4\text{Var}(\overline{X}_n) = 4\frac{\sigma^2}{n}$$

mit $\sigma^2 = \text{Var}(X_i) = \frac{\vartheta^2}{12}$. Also: $\text{Var}(T_1) = \frac{\vartheta^2}{3n}$. Damit ist

$$MSE(T_1; \vartheta) = \frac{\vartheta^2}{3n}$$

[1] Diese Aufgabe ist eine gute Übung für das Zusammenspiel von Erwartungswerten, Varianzen, Termumformungen und Berechnung von Integralen (Arbeiten mit Dichten)!

Schritt 2: Berechne $MSE(T_2;\vartheta)$:
Berechne die Varianz von $Z = \max_{1\le i\le n} X_i$ mit Verschiebungssatz:

$$\mathrm{Var}(Z) = E(Z^2) - (E(Z))^2$$

Oben schon berechnet: $E(Z) = E(X_{(n)}) = \frac{n}{n+1}\vartheta$ und

$$f_Z(x) = \frac{n}{\vartheta^n} x^{n-1} \mathbf{1}_{(0,\vartheta)}(x), \qquad x \in \mathbb{R}.$$

$$\begin{aligned}
\Rightarrow E(Z^2) &= \int_{-\infty}^{\infty} x^2 \cdot f_Z(x)\, dx = \int_0^{\vartheta} x^2 \frac{n}{\vartheta^n} x^{n-1}\, dx \\
&= \frac{n}{\vartheta^n} \int_0^{\vartheta} x^{n+1}\, dx = \frac{n}{\vartheta^n} \frac{\vartheta^{n+2}}{n+2} \\
&= \frac{n}{n+2}\vartheta^2
\end{aligned}$$

Fs. Schritt 2:

$$\begin{aligned}\mathrm{Var}(Z) = E(Z^2) - (E(Z))^2 &= \frac{n}{n+2}\vartheta^2 - \left(\frac{n}{n+1}\vartheta\right)^2 \\ &= \vartheta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) \\ &= \vartheta^2\left(\frac{n(n+1)^2 - (n+2)n^2}{(n+2)(n+1)^2}\right)\end{aligned}$$

Vereinfachen des Ausdrucks im Zähler des Bruchs:

$$\begin{aligned}n(n+1)^2 - (n+2)n^2 &= n(n^2+2n+1) - (n^3+2n^2) \\ &= (n^3+2n^2+n) - n^3 - 2n^2 = n.\end{aligned}$$

Damit folgt:

$$\mathrm{Var}(Z) = \vartheta^2 \frac{n}{(n+2)(n+1)^2}$$

Fs. Schritt 2:

$$\text{Var}(Z) = \vartheta^2 \frac{n}{(n+2)(n+1)^2}$$

Mit $T_2 = \frac{n+1}{n} Z$ ergibt sich

$$\text{Var}(T_2) = \frac{(n+1)^2}{n^2} \cdot \vartheta^2 \frac{n}{(n+2)(n+1)^2} = \frac{\vartheta^2}{n(n+2)}$$

Da T_2 erwartungstreu für ϑ ist, ergeben sich also die folgenden MSEs:

$$MSE(T_1; \vartheta) = \frac{\vartheta^2}{3n} \qquad MSE(T_2; \vartheta) = \frac{\vartheta^2}{n(n+2)}$$

Schritt 4: Vergleich der Ausdrücke:

$$MSE(T_1; \vartheta) > MSE(T_2; \vartheta) \Leftrightarrow \frac{\vartheta^2}{3n} > \frac{\vartheta^2}{n(n+2)} \Leftrightarrow n^2 + 2n > 3n$$

Dies ist für alle $n > 1$ der Fall (für $n = 1$ sind die Ausdrücke gleich).

Testverteilungen

Seien $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ und

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

sowie

$$S^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

t-Verteilung

Die Verteilung von

$$T = \sqrt{n} \frac{\overline{X} - \mu}{S}$$

heißt t-**Verteilung mit** $n-1$ **Freiheitsgraden**.
Notation: $t(n-1)$. p-Quantil: $t(n-1)_p$.

χ^2-Verteilung

Sind $U_1, \ldots, U_k$ i.i.d. $\sim N(0,1)$, dann heißt die Verteilung von

$$Q = \sum_{i=1}^{k} U_i^2$$

χ^2-Verteilung mit k Freiheitsgraden.
Momente: Es gilt: $E(Q) = k$ und $\mathrm{Var}(Q) = 2k$.
Gilt mit einer Konstanten $c > 0$:

$$T/c \sim \chi^2(k),$$

dann heißt T **gestreckt χ^2-verteilt mit k Freiheitsgraden**.
Man schreibt auch: $T \sim c \cdot \chi^2(k)$.

Verteilung der Varianzschätzer

Annahme: Normalverteilungsmodell, d.h.

$$X_1, \ldots, X_n \overset{d}{\sim} N(\mu, \sigma^2)$$

Welchen Varianzschätzer $\widehat{\sigma}_n^2$ wann verwenden?

Fall 1: μ bekannt: Verwende $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Dann gilt (per def.)

$$\frac{n}{\sigma^2}\widehat{\sigma}_n^2 \sim \chi^2(n)$$

Fall 2: μ unbekannt. Verwende $\widehat{\sigma}_n^2 := S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Dann:

$$\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2(n-1).$$

F-Verteilung

Seien $Q_1 \sim \chi^2(n_1)$ und $Q_2 \sim \chi^2(n_2)$ unabhängig. Dann heißt die Verteilung des Quotienten

$$F = \frac{Q_1/n_1}{Q_2/n_2}$$

F-Verteilung mit n_1 und n_2 Freiheitsgraden.

Notation: $F(n_1, n_2)$.

p-Quantil: $F(n_1, n_2)_p$.

Momente: $E(F) = \frac{n_2}{n_2-2}$, $\mathrm{Var}(F) = \frac{2n_2^2(n_2+n_1-2)}{n_1(n_2-2)^2(n_2-4)}$.

F-Verteilung: Vergleich von Varianzschätzungen

$X_{11}, \ldots, X_{1,n_1}$ und $X_{21}, \ldots, X_{2,n_2}$ seien zwei unabhängige normalverteilte Stichproben mit

$$X_{1i} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \qquad i = 1, \ldots, n_1,$$

und

$$X_{2i} \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2), \qquad i = 1, \ldots, n_2,$$

Erwartungstreue und unabhängig Schätzungen der Varianzen σ_1^2 und σ_2^2 sind

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2, \qquad i = 1, 2.$$

Zahlenbeispiel: $s_1^2 = 3.5$ und $s_2^2 = 5.5$. Frage: Besteht tatsächlich ein Unterschied?

F-Verteilung

Man kann prinzipiell $S_2^2 - S_1^2$ mit 0 vergleichen oder S_1^2/S_2^2 mit 1.
In der Statistik betrachtet man den Quotienten, da dieser einer (gestreckten) F-Verteilung folgt:

$$Q_i = \frac{n_i - 1}{\sigma_i^2} S_i^2 \sim \chi^2(n_i - 1), \qquad i = 1, 2.$$

Q_1 und Q_2 sind unabhängig und χ^2-verteilt. Daher gilt:

$$F = \frac{Q_1/(n_1 - 1)}{Q_2/(n_2 - 1)} \sim F(n_1 - 1, n_2 - 1)$$

Ausrechnen:

$$F = \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2}.$$

Im Fall $\sigma_1^2 = \sigma_2^2$ folgt: $F \sim F(n_1 - 1, n_2 - 1)$.

Konfidenzintervalle

Kritik an Punktschätzungen:
Ein Datensatz liefere:

$$\overline{x} = 11.34534, \qquad s/\sqrt{n} = 5.45$$

Hinweis: $s/\sqrt{n}$ ist eine Schätzung der Standardabweichung von $\overline{X}$ und heißt **Standardfehler**.
Die Angabe vieler Nachkommastellen suggeriert eine Genauigkeit, die statistisch nicht unbedingt gerechtfertigt ist!
Besser:
Gebe ein datenbasiertes Intervall $[L, U]$ an, welches mit einer definierten (Mindest-) Wahrscheinlichkeit den Parameter überdeckt.

Anschauung: Sollte die Schätzung mit einem Microliner oder einem mehr oder weniger dicken Edding markiert werden?

Konfidenzintervall

Ein Intervall $[L, U]$ mit datenabhängigen Intervallgrenzen

$$L = L(X_1, \ldots, X_n)$$
$$U = U(X_1, \ldots, X_n)$$

heißt **Konfidenzintervall (Vertrauensbereich) zum Konfidenzniveau** $1 - \alpha$, wenn f¸r alle $\vartheta \in \Theta$ gilt:

$$P([L, U] \ni \vartheta) \geq 1 - \alpha.$$

Im Unterschied hierzu: **Prognoseintervall** für eine ZV X:

$$P(a < X \leq b) \geq 1 - \alpha$$

(Nehme Quantile $a = F_X^{-1}(\alpha/2)$ und $b = F_X^{-1}(1 - \alpha/2)$.)

Konfidenzintervall für μ

Modell:

$$X_1, \ldots, X_n \overset{d}{\sim} N(\mu, \sigma^2)$$

Ausgangspunkt: Prognoseintervall für $T = \sqrt{n}(\overline{X} - \mu)/S \sim t(n-1)$:
Mit Wahrscheinlichkeit $1 - \alpha$ gilt:

$$-t(n-1)_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X} - \mu}{S} \leq t(n-1)_{1-\alpha/2}$$

(Beachte: $t(n-1)_{1-\alpha/2}$ ist das $(1 - \frac{\alpha}{2})$-Quantil der $t(n-1)$-Verteilung!)
Umformen, so dass nur μ in der Mitte stehen bleibt:

$$\overline{X} - t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}.$$

$(1 - \alpha)$-KI für μ ist gegeben durch:

$$[L, U] = \left[\overline{X} - t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Konfidenzintervall für μ

Verbreitet in der Praxis: 'Error Bounds' $\overline{X}_n \pm S_n/\sqrt{n}$ (zu optimistisch!).

Statistiker verwendet $\overline{X}_n \pm t(n-1)_{1-\alpha/2}\frac{S_n}{\sqrt{n}}$ (klare Interpretation).

Mit einer Fehlerwahrscheinlichkeit von α ist die Aussage

$$\overline{X} - t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

über μ korrekt.

Konfidenzintervall für μ

Zweiseitiges KI, σ unbekannt:

$$\left[\overline{X} - t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t(n-1)_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Einseitige KIs:

1. Einseitiges unteres KI: $\left(-\infty, \overline{X} + t(n-1)_{1-\alpha} \cdot S/\sqrt{n}\right]$.
   Mit Wkeit $1-\alpha$ ist die Aussage „$\mu \leq \overline{X} + t(n-1)_{1-\alpha} \cdot S/\sqrt{n}$“ richtig (obere Schranke).
2. Einseitiges oberes KI: $\left[\overline{X} - t(n-1)_{1-\alpha} \cdot S/\sqrt{n}, \infty\right)$ liefert analog eine untere Schranke.

Falls σ bekannt ist: Ersetze in den Formeln:

1. S durch σ.
2. $t(n-1)_{1-\alpha/2}$ durch $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$.
3. $t(n-1)_{1-\alpha}$ durch $z_{1-\alpha}$.

$z_{1-\alpha}$: $(1-\alpha)$-Quantil der $N(0,1)$-Verteilung.

Konfidenzintervall für μ

Computersimulation: Simulation von 10 Stichproben vom Umfang n (=10) aus einer $N(2, 1)$-Verteilung.

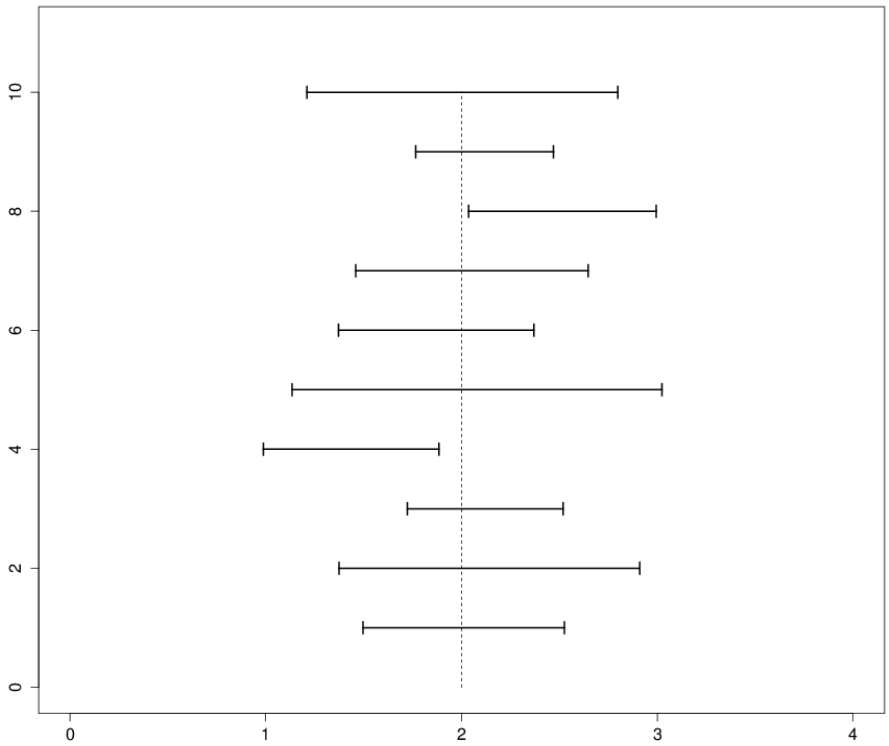


Abbildung: Computersimulation: Dargestellt sind 10 Konfidenzintervalle für μ, die aus 10 unabhängigen Stichproben berechnet wurden. Der im Experiment eingestellte Wert $\mu = 2$ ist gestrichelt eingezeichnet.

Konfidenzintervall für σ^2

Ausgangspunkt: Schätzer $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Mit Wahrscheinlichkeit $1-\alpha$ gilt:

$$\chi^2(n-1)_{\alpha/2} \leq \frac{(n-1)\widehat{\sigma}^2}{\sigma^2} \leq \chi^2(n-1)_{1-\alpha/2}$$

Umformen liefert zweiseitiges Konfidenzintervall für σ^2:

$$\left[\frac{n-1}{\chi^2(n-1)_{1-\alpha/2}}\widehat{\sigma}^2, \frac{n-1}{\chi^2(n-1)_{\alpha/2}}\widehat{\sigma}^2\right]$$

Analog:

- einseitiges oberes Konfidenzintervall: $[0, (n-1)\widehat{\sigma}^2/\chi^2(n-1)_{\alpha}]$
- einseitiges unteres Konfidenzintervall $[(n-1)\widehat{\sigma}^2/\chi^2(n-1)_{1-\alpha}, \infty)$

Konfidenzintervall für p

Modell: $Y \sim \text{Bin}(n, p)$.
Approximatives Konfidenzintervall (aus ZGWS):

$$
\begin{aligned}
L &= \widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \\
U &= \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}
\end{aligned}
$$

Konfidenzintervall für p

ZGWS für Binomialverteilung mit geschätztem $\sigma = \sqrt{p(1-p)}$ im Nenner: $\sqrt{n}\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}} \sim_{approx} N(0,1)$.
Mit Wahrscheinlichkeit $1-\alpha$ gilt näherungsweise (für großes n):

$$-z_{1-\alpha/2} \leq \sqrt{n}\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})}} \leq z_{1-\alpha/2}$$

Dies ist äquivalent zu (Umformen, so das p in der Mitte stehen bleibt):

$$\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p}}{)}n}$$

Somit überdeckt $[L, U]$ die unbekannte Erfolgswahrscheinlichkeit p mit Wahrscheinlichkeit $1-\alpha$.

Konfidenzintervall für p

Besser (bei kleinen Stichprobenumfängen):
Konfidenzintervalle $[p_L, p_U]$ nach Pearson-Clopper:

$$p_L = \frac{y \cdot f_{\alpha/2}}{n - y + 1 + y \cdot f_{\alpha/2}}, \quad p_U = \frac{(y+1) f_{1-\alpha/2}}{n - y + (y+1) f_{1-\alpha/2}}$$

mit den folgenden Quantilen der F-Verteilung:

$$f_{\alpha/2} = F(2y, 2(n-y+1))_{\alpha/2},$$
$$f_{1-\alpha/2} = F(2(y+1), 2(n-y))_{1-\alpha/2}.$$

Konfidenzintervall für p:

Wie genau sind Wahlumfragen?

Forschungsgruppe Wahlen: $n = 2500$ (Politbarometer).

Allensbach: $n = 1000$

Sonntagsfrage Januar 2013:

Partei	Allensbach	Forschungsgruppe Wahlen	Bundestagswahl 2009
CDU/CSU	39.0	41.0	33.8
SPD	28.0	29.0	23.0
GRÜNE	14.0	13	10.7
FDP	5	4	14.6
DIE LINKE	7	6	11.9
PIRATEN	3	3	2.0
Sonstige	4	4	4.0

Konfidenzintervall für p:

Beispiel: CDU/CSU als große Partei.

Schätzungen 39.0% (Allensbach) bzw. 41.0% (FG Wahlen). Wir berechnen die KI zur Konfidenz 95%.

Auswertung Allensbach-Umfrage:

Mit $z_{0.975} \approx 1.96$ und $n = 1000$ ergibt sich das realisierte KI

$$\left[0.39 - 1.96\sqrt{\frac{0.39(1-0.39)}{1000}}, 0.39 + 1.96\sqrt{\frac{0.39(1-0.39)}{1000}}\right] = [0.3598; 0.4202].$$

Auswertung FG-Wahlen-Umfrage mit $n = 2500$:

$$\left[0.41 - 1.96\sqrt{\frac{0.41(1-0.41)}{2500}}, 0.41 + 1.96\sqrt{\frac{0.41(1-0.41)}{2500}}\right] = [0.3907; 0.4293].$$

Konfidenzintervall für p:

Kleine Parteien: Wir nehmen die Daten der FG Wahlen (größeres n): Auswertung PIRATEN, Schätzung 3%.

Es ergibt sich das realisierte KI

$$\left[0.03 - 1.96\sqrt{\frac{0.03(1-0.03)}{2500}}, 0.03 + 1.96\sqrt{\frac{0.03(1-0.03)}{2500}}\right] = [0.0233; 0.0367]$$

Auswertung FDP: Schätzung 4% :

$$\left[0.04 - 1.96\sqrt{\frac{0.04(1-0.04)}{2500}}, 0.04 + 1.96\sqrt{\frac{0.04(1-0.04)}{2500}}\right] = [0.0323; 0.0477]$$

Aufgabe 37

Textaufgabe:

Eine Fluggesellschaft möchte wissen, wie hoch der Anteil p der Passagiere ist, die ihren Flug nicht antreten. Hierzu soll ein Konfidenzintervall für p bestimmt werden.

Die Überprüfung von 1000 zufällig ausgewählten Passagieren ergibt, dass 74 von ihnen den Flug nicht angetreten haben.

Bestimmen Sie anhand dieses Ergebnisses ein approximatives zweiseitiges Konfidenzintervall für p zum Konfidenzniveau 90%.

Aufgabe 37

Beispiel: Beobachte die Anzahl Y der von einer künstlichen Intelligenz richtig erkannten Testbeispiele unter $n = 30$ Beispielen.

Modell: $Y \sim \text{Bin}(n = 30, p)$

p: wahre Wahrscheinlichkeit, dass der Detektor korrekt erkennt. p ist unbekannt.

Entscheidungproblem:

$p = p_0 = 1/2$: nur so gut wie eine Entscheidung per Münzwurf.

$p = p_1 = 0.9$: Wunschrate korrekter Detektionen.

Wir wollen entscheiden zwischen $p = p_0$ und $p = p_1$.

$\rightarrow$ Zwei Verteilungen (Zähldichten) für Y.

$$\text{bin}(30, 1/2) \text{ oder } \text{bin}(30, 0.9).$$

Beispiel: Erhebe n Messungen $X_1, \ldots, X_n$ der Ozonkonzentration X (in $\mu g/m^3$). Aus langjährigen Voruntersuchungen sei die Standardabweichung $\sigma = 5$ bekannt.

Modell: $X_1, \ldots, X_n \sim N(\mu, 5^2)$

μ: wahre Ozonkonzentration ($\mu = E(X)$), μ unbekannt

Entscheidungsproblem:

$\mu = \mu_0 = 240$: Alarmschwellwert lt. Ozon-Gesetz

$\mu = \mu_1 = 200$: Zielwert der Gemeinde

$\rightarrow$ Zwei Verteilungen (Dichten) für die Daten X:

$$\varphi_{(240,25)}(x) \text{ oder } \varphi_{(200,25)}(x).$$

Testproblem, Nullhypothese, Alternative

Sind f_0 und f_1 zwei mögliche Verteilungen für eine Zufallsvariable X, dann wird das **Testproblem**, zwischen $X \sim f_0$ und $X \sim f_1$ zu entscheiden, in der Form

$$H_0 : f = f_0 \qquad \text{gegen} \qquad H_1 : f = f_1$$

notiert, wobei f die wahre Verteilung von X bezeichnet. H_0 heißt **Nullhypothese** und H_1 **Alternative** (**Alternativhypothese**).

Statistische Testtheorie

Datenmaterial $X_1, \ldots, X_n$
Statistik $T = T(X_1, \ldots, X_n)$

Statistischer Test

Ein **(statistischer) Test** ist eine Entscheidungsregel, die basierend auf T entweder zugunsten von H_0 (Notation: „H_0“) oder zugunsten von H_1 („H_1“) entscheidet.

Fehler 1. und 2. Art

Entscheidung für H_1, obwohl H_0 richtig ist, heißt **Fehler 1. Art**. H_0 wird dann fälschlicherweise verworfen. Eine Entscheidung für H_0, obwohl H_1 richtig ist, heißt **Fehler 2. Art**. H_0 wird fälschlicherweise akzeptiert.

Insgesamt sind vier Konstellationen möglich, die in der folgenden Tabelle zusammengefasst sind:

	H_0	H_1
„H_0“	$\surd$	Fehler 2. Art
„H_1“	Fehler 1. Art	$\surd$

Signifikanzniveau, Test zum Niveau α

Bezeichnet „H_1“ eine Annahme der Alternative und „H_0“ eine Annahme der Nullhypothese durch eine Entscheidungsregel, dann ist durch diese Regel ein **statistischer Test zum Signifikanzniveau (Niveau)** α gegeben, wenn

$$P_{H_0}(\text{„}H_1\text{“}) \leq \alpha\,.$$

Genauer ist die linke Seite ist das tatsächliche Signifikanzniveau des Tests und die rechte Seite das vorgegebene **nominale** Signifikanzniveau.

Hinweis: Die Wahrscheinlichkeit eines Fehlers 2. Art wird nicht unbedingt kontrolliert. Dies erfordert eine Planung der Stichprobengröße.

Schärfe (Power)

Die Wahrscheinlichkeit eines Fehlers 2. Art wird üblicherweise mit β bezeichnet. Die Gegenwahrscheinlichkeit,

$$1 - \beta = P_{H_1}(\text{„}H_1\text{“})(= E_{H_1}(1 - \phi)),$$

dass der Test die Alternative H_1 tatsächlich aufdeckt, heißt **Schärfe (Power)** des Testverfahrens.

Entscheidungskonstellationen und die Wahrscheinlichkeiten:

	H_0	H_1
„H_0“	$\surd$ $1 - \alpha$	Fehler 2. Art β
„H_1“	Fehler 1. Art α	$\surd$ $1 - \beta$: Schärfe (Power)

Frage: Wie sollen die Hypothesen H_0 und H_1 zugeordnet werden?

Vorgehen 1: Risikoüberlegung

$\rightarrow$ Ein Signifikanztest kontrolliert stets den Fehler 1. Art, aber nicht unbedingt den Fehler 2. Art.

- Entscheide für das vorliegende Problem, welcher Fehler schlimmer ist und auf jeden Fall kontrolliert werden soll.
- Stelle Hypothesen so auf, dass der Fehler 1. Art der schlimmere ist.

Vorgehen 2: Nachweisformulierung

• Sehr oft stellt eine der Hypothesen das etablierte Wissen (Stand der Technik) da und die andere Hypothese den vermuteten neuen, besonderen Effekt.

• Der Effekt kann z.B. ein Überschreiten eines Grenzwerts, ein Unterschreiten einer Zielvorgabe des Managements, die Wirksamkeit eines neuen Wirkstoffs für ein Medikament oder die Überlegenheit einer künstlichen Intelligenz für Problem X sein.

$\rightarrow$ Wer den Effekt nachweisen will, muss die Gegenseite überzeugen.

$\rightarrow$ Die Anhänger des etablierten Wissens werden sich nur dann von dem Effekt überzeugen lassen und ihre Meinung ändern, wenn die Wahrscheinlichkeit einer Fehlentscheidung zu Gunsten des Effekts (sehr) klein ist!

Vorgehen 2: Nachweisformulierung

Ansatz: Test kontrolliert $P_{H_0}(\text{„}H_1\text{“}) \leq \alpha$.

Lege H_0 und H_1 so fest, dass gilt:

$$P(\text{„Entscheidung für Effekt, obwohl kein Effekt existiert.“}) = P_{H_0}(\text{„}H_1\text{“})$$

Die Hypothese, die den Effekt beschreibt wird die Alternativhypothese!
Die Hypothese, die das etablierte Wissen (kein Effekt) beschreibt, wird die Nullhypothese!

In dieser Formulierung wird die Fehlerwahrscheinlichkeit durch α kontrolliert, fälschlicherweise von dem Vorliegen des Effekts auszugehen.

Verallgemeinerung: s. Buch

$$H_0 : \vartheta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \vartheta \in \Theta_1,$$

wobei $\Theta_0 \cup \Theta_1 = \Theta$ eine disjunkte Zerlegung des Parameterraums Θ ist.

Beispiel: Modell $N(\mu, 25)$, $\mu \in \mathbb{R}$. Gilt $\mu = 200$ oder $\mu = 240$?

$\Theta = \{200, 240\}$.
$\Theta_0 = \{200\} \leftrightarrow H_0 : \mu = 200$
$\Theta_1 = \{240\} \leftrightarrow H_1 : \mu = 240$.

Hypothesen

Meist möchte man aber nicht nur zwei Verteilungen gegeneinander testen, sondern z.B. $\mu \leq 240$ (Grenzwert eingehalten) testen gegen $\mu > 240$ (Alarmwert überschritten).

Hypothesen (über den Erwartungswert μ)

Einseitiges Testproblem:

$$H_0 : \mu \leq \mu_0 \qquad \text{gegen} \qquad H_1 : \mu > \mu_0,$$

bzw.

$$H_0 : \mu \geq \mu_0 \qquad \text{gegen} \qquad H_1 : \mu < \mu_0.$$

Zweiseitiges Testproblem:

$$H_0 : \mu = \mu_0 \qquad \text{gegen} \qquad H_1 : \mu \neq \mu_0.$$

WICHTIG: Der Grenzfall $'=$' wird immer H_0 zugeschlagen.

Der Gauß-Test

Gegeben: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ mit *bekannter* Varianz $\sigma^2 \in (0, \infty)$

Teststatistik: $T = \sqrt{n}\,\dfrac{\overline{X}_n - \mu_0}{\sigma}$ ($\mu_0 \in \mathbb{R}$ vorgegebener Sollwert)

Verteilung der Teststatistik: $T \sim \mathrm{N}(0,1)$ für $\mu = \mu_0$

(In der Teststatistik wird $\overline{X}_n$ mit μ_0 verglichen, dem am schwersten von H_1 zu unterscheidendem Fall.)

Der Gauß-Test

Einseitiger Gauß-Test (1)

Der einseitige Gaußtest verwirft die Nullhypothese $H_0 : \mu \leq \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu > \mu_0$, wenn $T > z_{1-\alpha}$.

Einseitiger Gauß-Test (2)

Der einseitige Gaußtest verwirft $H_0 : \mu \geq \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu < \mu_0$, wenn $T < -z_{1-\alpha} = z_\alpha$.

Zweiseitiger Gauß-Test

Der zweiseitige Gauß-Test verwirft die Nullhypothese $H_0 : \mu = \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu \neq \mu_0$, wenn $|T| > z_{1-\alpha/2}$.

(Hierbei bezeichnet z_p das p-Quantil zu $N(0,1)$ für $p \in (0,1)$.)

Beispiel

Beispiel: Parfum wird in 100 ml Flaschen abgefüllt. Durch eine Zufallsstichprobe vom Umfang $n = 36$ möchte das Management überprüfen, ob nicht etwa die Füllmenge zu hoch ist. Die Messungen können als normalverteilt angesehen werden. Die Fehlerwahrscheinlichkeit, fälschlicherweise aus den Daten auf eine zu hohe Füllmenge zu schlieflen, wird vom Management mit $\alpha = 1\%$ festgelegt.

$$\overline{x} = 101.78, \quad \sigma = 2.34, \quad n = 36$$

Modell: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, μ: wahre Füllmenge, $\sigma^2 = 2.34^2$

1. Formulierung des Testproblems:

$H_0 : \mu \leq 100$ versus $H_1 : \mu > 100$

2. Testverfahren:

Einseitiger 1-SP-Gaufltest (nach oben) mit $\mu_0 = 100$ durchgeführt.

3. Testdurchführung:

3a. Berechne $t = \sqrt{n}\frac{\overline{x}-\mu_0}{\sigma}$.

3b. Berechne kritischen Wert $c_{krit} = q_{0.99}$

3c. Lehne H_0 ab, falls $t > c_{krit}$. Behalte H_0 bei, falls $t \leq c_{krit}$

Beispiel

Datenanalyse und -interpretation:

$$\overline{x} = 101.78, \quad \sigma = 2.34, \quad n = 36$$

Berechnung der Teststatistik:

$$t = T_{obs} = \sqrt{36}\frac{101.78 - 100}{2.34} = 4.564...$$

Berechnung des kritischen Werts (z.B. Raussuchen Tabelle Buch S. 304):

$$c_{krit} = q_{1-0.01} = q_{0.99} \approx 2.33$$

Testdurchführung: Da $t = 4.465 > 2.33 = c_{krit}$, wird H_0 auf dem Niveau $\alpha = 0.01$ abgelehnt.
Antwortsatz: Basierend auf einer normalverteilten Zufallsstichprobe vom Umfang $n = 36$ konnte durch Anwendung eines einseitigen Gaußtests statistisch auf einem Signifikanzniveau von $\alpha = 1\%$ nachgewiesen werden, dass die Füllmenge größer als der Sollwert 100[ml] ist.

Der t-Test (für eine Stichprobe)

Gegeben: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ mit *unbekannter* Varianz σ^2

Idee: Ersetze die unbekannte Varianz σ^2 durch den (erwartungstreuen und konsistenten) Schätzer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \ .$$

Teststatistik: $T = \sqrt{n} \, \dfrac{\overline{X}_n - \mu_0}{S}$ ($\mu_0 \in \mathbb{R}$ vorgegebener Sollwert)

Verteilung der Teststatistik: $T \sim t(n-1)$ für $\mu = \mu_0$

Der t-Test (für eine Stichprobe)

Einseitiger t-Test (1)

Der einseitige t-Test verwirft die Nullhypothese $H_0 : \mu \leq \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu > \mu_0$, wenn $T > t(n-1)_{1-\alpha}$.

Einseitiger t-Test (2)

Der einseitige t-Test verwirft die Nullhypothese $H_0 : \mu \geq \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu < \mu_0$, wenn $T < -t(n-1)_{1-\alpha} = t(n-1)_{\alpha}$.

Zweiseitiger t-Test

Der zweiseitige t-Test verwirft die Nullhypothese $H_0 : \mu = \mu_0$ auf dem Signifikanzniveau α zugunsten von $H_1 : \mu \neq \mu_0$, wenn $|T| > t(n-1)_{1-\alpha/2}$.

(Hierbei bezeichnet $t(n-1)_p$ das p-Quantil zur t-Verteilung mit $n-1$ Freiheitsgraden für $p \in (0,1)$.)

Der t-Test (für eine Stichprobe)

Beispiel

Die Schätzung der mittleren Ozonkonzentration während der Sommermonate ergaben für eine Großstadt anhand von $n = 26$ Messungen den Mittelwert $\overline{x}_n = 244$ und die Stichproben-Standardabweichung $s = 5.1$ (jeweils in $\mu g/m^3$).

Der im Ozongesetz von 1995 festgelegte verbindliche Alarmwert beträgt 240 $\mu g/m^3$. Kann das gemessene Ergebnis als signifikante Überschreitung des Warnwerts gewertet werden zum Signifikanzniveau $\alpha = 0.01$?

Der t-Test (für eine Stichprobe)

Lösung:

- Als beobachtete Teststatistik erhalten wir aus den Daten:

$$t = T_{obs} = \sqrt{26}\frac{244 - 240}{5.1} = 3.999,$$

Die Statistik T ist $t(n-1=25)$-verteilt, wenn $\mu = \mu_0$, also am Rand der Nullhypothese.

- Bestimmung des kritischen Werts: $c_{krit} = t(25)_{0.99} = 2.485$.
- Testentscheidung: Da $t > 2.485$ wird die Nullhypothese $H_0 : \mu \leq 240$ zu Gunsten von $H_1 : \mu > 240$ verworfen.
- Antwortsatz: Durch einen Signifikanztest (1-Stichproben $t-$Test) konnte basierend auf einer Zufallsstichprobe vom Umfang $n = 26$ auf einem Signifikanzniveau von 1% statistisch nachgewiesen werden, dass von einer Überschreitung des Alarmwert 240 ausgegangen werden kann.

Wichtige Angaben: n, α, verwendetes Testverfahren, die Hypothesenformulierung muss ersichtlich sein.

Zusammenhang Test $\leftrightarrow$ Konfidenzintervall

Der **zweiseitige** t-**Test** akzeptiert H_0 auf dem Niveau α, wenn

$$\left| \sqrt{n} \frac{\overline{X} - \mu_0}{S} \right| \leq t(n-1)_{1-\alpha/2}$$

(sonst wird H_0 abgelehnt). Dies ist äquivalent zur Ungleichungskette

$$\mu_0 - t(n-1)_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \overline{X} \leq \mu_0 + t(n-1)_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}.$$

Man kann also auch $\overline{X}$ mit $\mu_0 \pm t(n-1)_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$ vergleichen.
Weiteres Umformen liefert:

$$\overline{X} - t(n-1)_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \mu_0 \leq \overline{X} + t(n-1)_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}.$$

$H_0 : \mu = \mu_0$ wird somit genau dann akzeptiert, wenn der Sollwert μ_0 vom $(1-\alpha)$-Konfidenzintervall für μ überdeckt wird.

Zusammenhang Test $\leftrightarrow$ Konfidenzintervall

Merke: (Erfahrene Data Analysts kennen solche Zusammenhänge...)

Das (zweiseitige) Konfidenzintervall für μ liefert die wertvolle Information, welche Nullhypothesen vom vorliegenden Datenmaterial durch den (zweiseitigen) t-Test abgelehnt werden.

p-Wert

Durchführung eines statistischen Tests:

1. Formuliere H_0 und H_1.
2. Wähle Signifikanzniveau α.
3. Bestimme kritischen Wert c_{krit}.
4. Berechne t_{obs}.
5. Vergleiche t_{obs} mit c_{krit}.

Nachteile:

Bei Änderung von α müssen 3. bis 5. neu durchgeführt werden.

In der Praxis geht man so vor:

Angabe einer Zahl p, so dass folgende Regel gilt:

$$p < \alpha \qquad \Leftrightarrow \qquad H_0 \text{ ablehnen}$$

p-Wert

Frage

Wie wahrscheinlich ist es, bei einer (gedanklichen) Wiederholung des Experiments, einen Teststatistik-Wert zu beobachten, der noch deutlicher gegen H_0 spricht als t_{obs}?

p-Wert

Einseitige Tests

Testproblem: $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$

$$p = P_{\mu_0}(T > t_{obs})$$

Testproblem: $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$

$$p = P_{\mu_0}(T < t_{obs})$$

Lehne H_0 genau dann ab, wenn $p < \alpha$.

p-Wert

Zweiseitiger Test

Testproblem: $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$

$$p_{zweis} = P_{\mu_0}(|T| > |t_{obs}|)$$

Lehne H_0 genau dann ab, wenn $p_{zweis} < \alpha$.

Einseitige Tests

Gegeben: t_{obs} und p_{zweis}.

1. Lehne $H_0 : \mu \leq \mu_0$ zugunsten von $H_1 : \mu > \mu_0$ ab, falls

$$t_{obs} \geq 0 \quad \text{und} \quad \frac{p_{zweis}}{2} \overset{t_{obs} \geq 0}{=} P_{\mu_0}(T > t_{obs}) < \alpha \ .$$

2. Lehne $H_0 : \mu \geq \mu_0$ zugunsten von $H_1 : \mu < \mu_0$ ab, falls

$$t_{obs} \leq 0 \quad \text{und} \quad \frac{p_{zweis}}{2} \overset{t_{obs} \leq 0}{=} P_{\mu_0}(T < t_{obs}) < \alpha \ .$$

Gütefunktion

Frage: Wie wahrscheinlich ist es, dass die Alternative H_1 tatsächlich aufgedeckt wird?

Gesucht: $P_{H_1}(\text{„}H_1\text{“})$.

Gütefunktion

Die Funktion

$$G(\mu) = P_\mu(\text{„}H_1\text{“}) = P(\text{„}H_1\text{“}|\mu, \sigma^2), \qquad \mu \in \mathbb{R},$$

heißt Gütefunktion (an der Stelle μ).

Beispiel

Beispiel

Sei $\mu_0 = 150$ *und* $\sigma = 10$. *Betrachte das Testproblem*

$$H_0 : \mu \leq 150, \qquad \textit{versus} \qquad H_1 : \mu > 150.$$

Wähle $\alpha = 0.01$. *Der einseitige Gauß-Test verwirft* H_0, *falls*

$$T > z_{0.99} = 2.3263$$

Bestimme $G(\mu)$ *für* $\mu \in \{155, 160\}$.

Fallzahlplanung:

Lösung: Berechnung der Gütefunktion

$$G(\mu) = P_\mu(T > 2.3263).$$

Ist μ der wahre Erwartungswert, dann ist T nicht korrekt zentriert.
Korrektur:

$$\frac{\overline{X} - 150}{10/\sqrt{n}} = \underbrace{\frac{\overline{X} - \mu}{10/\sqrt{n}}}_{\sim N(0,1)} + \frac{\mu - 150}{10/\sqrt{n}}.$$

Schreibe $G(\mu)$ um:

$$\begin{aligned}
G(\mu) &= P_\mu\left(\frac{\overline{X} - 150}{10/\sqrt{n}} > 2.3263\right) \\
&= P_\mu\left(\frac{\overline{X} - \mu}{10/\sqrt{n}} + \frac{\mu - 150}{10/\sqrt{n}} > 2.3263\right) \\
&= P_\mu\left(\frac{\overline{X} - \mu}{10/\sqrt{n}} > 2.3263 - \frac{\mu - 150}{10/\sqrt{n}}\right) \\
&= 1 - \Phi\left(2.3263 - \frac{\mu - 150}{10/\sqrt{n}}\right)
\end{aligned}$$

Fallzahlplanung:

Für $n = 25$ und $\mu = 155$ erhalten wir

$$G(155) = \Phi(-2.3263 + 2.5) = \Phi(0.1737) \approx 0.569.$$

Genauso berechnet man $G(160) = \Phi(2.6737) \approx 0.9962$.

Eine Abweichung von 10 Einheiten wird also mit sehr hoher Wahrscheinlichkeit entdeckt, 5 Einheiten jedoch lediglich mit Wahrscheinlichkeit ≈ 0.57.

Ersetzt man in der obigen Herleitung 2.3263 durch $z_{1-\alpha}$, 150 durch μ und 10 durch σ, so erhält man die allgemeine Formel für die Güte des einseitigen Gaußtests:

Gütefunktion

Formel für die Güte des einseitigen Gaußtests:

$$G(\mu) = \Phi\left(-z_{1-\alpha} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

Analog für den zweiseitigen Test:

$$G_{\text{zweis.}}(\mu) = 2\Phi\left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

Hinweis: In der Praxis wird σ aus Trainingsdaten (historischen Daten) durch S geschätzt.

Fallzahlplanung

Ziel: Bestimme den Stichproben-Umfang n so, dass eine vorgegebene Lageänderung d $(= \mu - \mu_0)$ von μ_0 mit einer Mindestwahrscheinlichkeit von $1 - \beta$ aufgedeckt wird.

Hierdurch wird auch der Fehler 2. Art kontrolliert: Die Fehlerwahrscheinlichkeit 2. Art ist ein Abweichung d (oder schlimmer) höchstens β.

Beispiel

Ansatz: Bestimme n so, dass eine Abweichung von 5 mit Wahrscheinlichkeit von mindestens 90% aufgedeckt wird.
Mit $\mu(= 155)$ ist n so zu bestimmen, dass

$$\Phi\left(-2.3263 + \frac{\mu - 150}{10/\sqrt{n}}\right) \geq 0.9.$$

Bezeichne das Argument von Φ mit z. Zu Lösen ist also $\Phi(z) \geq 0.9$. Da $\Phi(z)$ und die Inverse $\Phi^{-1}(p)$ streng monton wachsend sind, ist

$$\Phi(z) \geq 0.9 \Leftrightarrow z \geq z_{0.9}$$

(allg.: $\Phi(z) \geq 1 - \beta \Leftrightarrow z \geq z_{1-\beta}$). Also:

$$z = -2.3263 + \sqrt{n}\frac{\mu - 150}{10} \geq z_{0.9}$$

Formales Auflösen nach n liefert für $\mu = 155$ und $z_{0.9} = 1.12816..$:

$$n \geq \frac{10^2}{5^2}(2.3263 + 1.2816)^2 = 52.068$$

$\rightarrow$ **Die gewünschte Schärfe des Tests erfordert $n \geq 53$.**

Beispiel

Beispiel

Hatten:

$$G(\mu) = \Phi\left(-2.3263 + \frac{\mu - 150}{10/\sqrt{n}}\right), \qquad \mu \in \mathbb{R}.$$

Finde den minimalen Stichprobenumfang $n \in \mathbb{N}$, so dass eine Abweichung von $d = 5$ mit einer Wahrscheinlichkeit von mindestens 90% aufgedeckt wird.

Fallzahlplanung

Mindestfallzahl

$$n \geq \frac{\sigma^2}{|\mu - \mu_0|^2}(z_{1-\alpha} + z_{1-\beta})^2.$$

Für den zweiseitigen Fall ergibt sich die Forderung

$$n \geq \frac{\sigma^2}{|\mu - \mu_0|^2}(z_{1-\alpha/2} + z_{1-\beta})^2,$$

damit Abweichungen größer oder gleich $\Delta = |\mu - \mu_0|$ mit einer Mindestwahrscheinlichkeit von $1 - \beta$ aufgedeckt werden.

2-Stichproben-Tests

Zwei Grundsituationen

1. Verbundenes Design
2. Unverbundenes Design

Verbundenes Design

Typische Anwendungssituation:
'Vorher-Nachher-Design' zur Analyse von zeitlichen Effekten bzw. Effekten nach Änderung der Versuchsbedingungen.

Beispiele:

- Änderung Druck/Temperatur/Spannung/...,
- neue Marketing-Maßnahme
- Schulung
- Umstrukturierung
- Medikament

Verbundenes Design

Für $i = 1, \ldots, n$ erhebe

X_i : Messung an der i-ten Versuchseinheit vorher,

Y_i : Messung an der i-ten Versuchseinheit nachher.

Modell: Bivariate einfache Stichprobe

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

von normalverteilten Zufallsvektoren mit

$$\mu_X = E(X_i) \qquad \text{und} \qquad \mu_Y = E(Y_i)$$

Betrachte die Differenzen (nachher - vorher):

$$D_i = Y_i - X_i, \qquad i = 1, \ldots, n.$$

Erwartungswert der Differenzen:

$$E(D_i) = E(Y_i) - E(X_i) = \mu_Y - \mu_X = \delta.$$

Verbundenes Design

Es sei $\sigma_D^2 = Var(D_1) = \ldots = Var(D_n)$ unbekannt.

Verwerfe dann

$$H_0 : \delta = 0 \Leftrightarrow \mu_X = \mu_Y \qquad \text{(kein Effekt)}$$

zugunsten von

$$H_1 : \delta \neq 0 \Leftrightarrow \mu_X \neq \mu_Y \qquad \text{(Effekt vorhanden)}$$

falls

$$|T| > t(n-1)_{1-\alpha/2},$$

wobei

$$T = \sqrt{n}\,\frac{\overline{D}}{S_D} \quad \text{mit} \quad \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \;,\; S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}\;.$$

Analog konstruiert man einseitige Tests, s. Buch.

Unverbundes 2-Stichprobenproblem

Modell: Zwei unabhängige Stichproben

$$X_{11}, \ldots, X_{1n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$$

$$X_{21}, \ldots, X_{2n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$$

Schritte:

1. Test auf Varianzhomogenität: Gilt $\sigma_1^2 = \sigma_2^2$?
2. Test auf Lageunterschied: Gilt $\mu_1 = \mu_2$?

Test auf Varianzhomogenität

Testproblem

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{versus} \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Varianzschätzungen:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1j} - \overline{X}_1)^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \overline{X}_2)^2$$

Teststatistik: $F = \frac{S_1^2}{S_2^2}$. Unter $H_0 : \sigma_1^2 = \sigma_2^2$ ist F F-verteilt!

Test

H_0 ablehnen, falls

$$F < F(n_1 - 1, n_2 - 1)_{\alpha/2} \qquad \text{oder} \qquad F > F(n_1 - 1, n_2 - 1)_{1-\alpha/2}$$

Äquivalent: Nummeriere so, dass $S_1^2 \leq S_2^2$ und lehne H_0 ab, falls $F < F(n_1 - 1, n_2 - 1)_{\alpha/2}$.

2-Stichproben-t-Test auf Lageunterschied

Annahme: $\sigma_1 = \sigma_2 =: \sigma^2$ (Varianzhomogenität).

Testproblem (zweiseitig):

$$H_0 : \mu_1 = \mu_2 \qquad \text{(kein Lageunterschied)}$$

versus

$$H_1 : \mu_1 \neq \mu_2 \quad \text{(Lageunterschied)}$$

Testprobleme (einseitig):

$$H_0 : \mu_1 \geq \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 < \mu_2.$$

bzw.

$$H_0 : \mu_1 \leq \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 > \mu_2.$$

2-Stichproben-t-Test auf Lageunterschied

Teststatistik: $T = \dfrac{\overline{X}_2 - \overline{X}_1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, S}$ mit

$$\begin{aligned} S^2 &= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 \\ &= \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_{1i} - \overline{X}_1)^2 + \sum_{j=1}^{n_2} (X_{1j} - \overline{X}_2)^2 \right) . \end{aligned}$$

2-Stichproben-t-Test

1. Lehne $H_0 : \mu_1 = \mu_2$ zugunsten von $H_1 : \mu_1 \neq \mu_2$ ab, wenn $|T| > t(n_1 + n_2 - 2)_{1-\alpha/2}$.
2. Lehne $H_0 : \mu_1 \geq \mu_2$ zugunsten von $H_1 : \mu_1 < \mu_2$ ab, wenn $T > t(n_1 + n_2 - 2)_{\alpha}$.
3. Lehne $H_0 : \mu_1 \leq \mu_2$ zugunsten von $H_1 : \mu_1 > \mu_2$, falls $T < t(n_1 + n_2 - 2)_{1-\alpha}$.

Welch-Test auf Lageunterschied

Bei Varianzinhomogenität $\sigma_1^2 \neq \sigma_2^2$ verwende man den Welch-Test.

Teststatistik:

$$T = \frac{\overline{X}_2 - \overline{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Lehne $H_0 : \mu_1 = \mu_2$ auf dem Niveau α ab, wenn $|T| > t(df)_{1-\alpha/2}$, wobei

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1-1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2-1}}$$

Falls $df \notin \mathbb{N}$, dann vorher auf nächste ganze Zahl abrunden.

Fallzahlplanung

Für $n = n_1 = n_2$ kann man folgende Näherungen verwenden:
Zweiseitiger Test: Wähle

$$n \geq \frac{\sigma^2}{\Delta^2}(z_{1-\alpha/2} + z_{1-\beta})^2,$$

um eine Schärfe von $1 - \beta$ bei einer Abweichung von $\Delta = |\mu_A - \mu_B|$ näherungsweise zu erzielen.
Einseitiger Test: Wähle

$$n \geq \frac{\sigma^2}{\Delta^2}(z_{1-\alpha} + z_{1-\beta})^2,$$

um eine Schärfe von $1 - \beta$ bei einer Abweichung von $\Delta = |\mu_A - \mu_B|$ näherungsweise zu erzielen.

Vorzeichentest, Binomialtest

Test für den Median

Modell: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$.

F besitze den eindeutigen Median $m = F^{-1}(0.5)$.

Testproblem

Einseitig

$$H_0 : m \geq m_0 \qquad \text{versus} \qquad H_1 : m < m_0$$

bzw.

$$H_0 : m \leq m_0 \qquad \text{versus} \qquad H_1 : m > m_0$$

Rückführung auf Binomialtest

Zähle Anzahl der Beobachtungen, die größer als m_0 sind $\rightarrow Y$

$$Y \sim \text{Bin}(n, p), \qquad p = P(Y_1 > m_0).$$

Für $m \overset{\leqq}{=} m_0$ gilt $p \overset{\leqq}{=} p_0 = 1/2$.

Exakter Binomialtest

Modell: Sei $Y \sim \text{Bin}(n, p)$.
Testproblem:

$$H_0 : p = p_0 \qquad \text{versus} \qquad H_1 : p \neq p_0.$$

Einseitig, z.B.

$$H_0 : p \leq p_0 \qquad \text{versus} \qquad H_1 : p > p_0.$$

Exakter Binomialtest

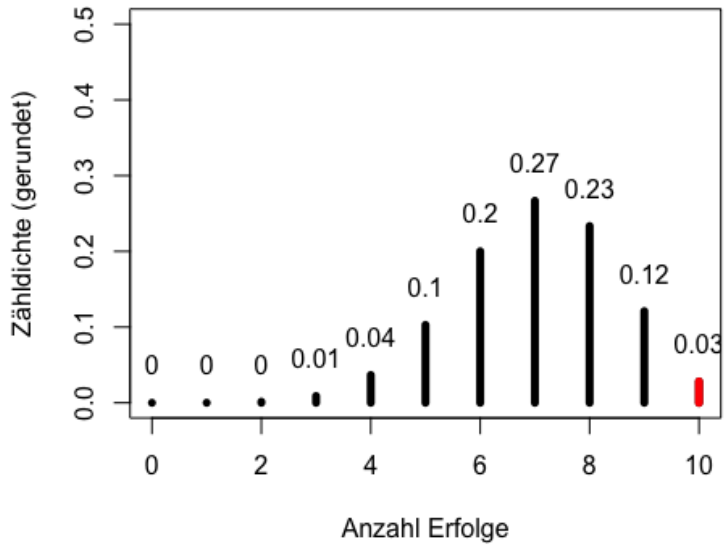
Lehne $H_0 : p \leq p_0$ zugunsten von $H_1 : p > p_0$, wenn $Y > c_{\text{krit}}$ ist. Hierbei ist c_{krit} die kleinste ganze Zahl, so dass

$$\sum_{k=c_{\text{krit}}+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha.$$

Exakter Binomialtest

Beispiel: $Y \sim \text{Bin}(10, p)$, $H_0 : p \leq 0.7$ versus $H_1 : p > 0.7$

Nominales Signifikanzniveau $\alpha = 0.05$



Kritischer Wert: $c_{\text{krit}} = 9$. H_0 wird abgelehnt, falls $Y > 9$.

Reales Signifikanzfikanzniveau: $P_{0.8}(Y > 9) = 0.03$.

$\Rightarrow$ Konservativer Test: Das Signifikanzniveau wird nicht ausgeschöpft.

Asymptotischer Binomialtest

Asymptotischer Test

1. Lehne $H_0 : p \leq p_0$ auf dem Niveau α zugunsten von $H_1 : p > p_0$ ab, wenn
$$T = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} > z_{1-\alpha}.$$
Äquivalent zu: $Y > np_0 + z_{1-\alpha}\sqrt{np_0(1-p_0)}$.
2. Lehne $H_0 : p \geq p_0$ zugunsten $H_1 : p < p_0$ ab, wenn $T < -z_{1-\alpha}$.
3. Lehne $H_0 : p = p_0$ zugunsten von $H_1 : p \neq p_0$ ab, wenn $|T| > z_{1-\alpha/2}$.

$z_{1-\alpha}$: $(1-\alpha)$-Quantil der $N(0,1)$-Verteilung.

Spezialfall $p_0 = 1/2$: Teststatistik vereinfacht sich zu

$$T = \frac{Y - n/2}{\sqrt{n/4}} = 2\frac{Y - n/2}{\sqrt{n}}.$$

2-Stichproben-Binomialtest

Modell: $Y_1 \sim \text{Bin}(n_1, p_1)$, $Y_2 \sim \text{Bin}(n_2, p_2)$ stochastisch unabhängig.
Testproblem:

$$H_0 : p_1 = p_2 \qquad \text{versus} \qquad H_1 : p_1 \neq p_2.$$

Schätzer: (Erfolgsraten in den Stichproben)

$$\widehat{p}_1 = Y_1/n_1, \qquad \widehat{p}_2 = Y_2/n_2$$

Teststatistik:

$$T = \frac{\widehat{p}_2 - \widehat{p}_1}{\sqrt{\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2} + \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}}}$$

Lehne H_0 ab, falls $|T| > z_{1-\alpha/2}$.

Test auf Korrelation

Modell: $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} N(\mu, \mathbf{\Sigma})$ bivariat normalverteilt mit

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

und Kovarianzmatrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \gamma \\ \gamma & \sigma_Y^2 \end{pmatrix}$$

Kovarianz

$$\gamma = \rho_{XY} = \mathrm{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y)$$

Korrelation

$$\rho = \mathrm{Cor}(X, Y) = \frac{\gamma_{XY}}{\sigma_X \sigma_Y}$$

Test auf Korrelation

Testproblem:

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_1 : \rho \neq 0$$

(Empirischer) Korrelationskoeffizient:

$$\widehat{\rho} = r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}},$$

Teststatistik:

$$T = \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}}$$

Unter H_0 gilt:

$$T \sim t(n-2)$$

1. Lehne $H_0 : \rho = 0$ ab, falls $|T| > t(n-2)_{1-\alpha/2}$.
2. Lehne $H_0 : \rho \geq 0$ ab, falls $T < -t(n-2)_{1-\alpha}$.
3. Lehne $H_0 : \rho \leq 0$ ab, falls $T > t(n-2)_{1-\alpha}$.

Beispiel: Die erwarteten Umsatzerlöse (Y) des Online-Shop SuperBuy4U eines Startups seien linear abhängig von der kumulierten Betrachtungszeit X der Werbespots auf Youtube (X = Viewer Retention · Clicks). Basierend auf den Daten der letzten $n = 25$ Wochen erhielt man folgende Statisiken:

$$\sum_{i=1}^{n} x_i y_i = 2603.316, \quad \sum_{i=1}^{n} x_i^2 = 16256.15, \quad \sum_{i=1}^{n} y_i^2 = 420.4859$$

sowie $\overline{x} = 24.96432$ und $\overline{y} = 4.049282$.

1. Formulieren Sie das zugehörige lineare Regressionsmodell unter Normalverteilungsannahme.
2. Berechnen Sie die Ausgleichsgerade.

Regressionsproblem

Gegeben: Punkte $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ **Punktewolke**.
Modell: Daten streuen um Gerade

$$f(x) = a + b \cdot x, \qquad x \in \mathbb{R}.$$

- Finde diejenige Gerade, die den Datensatz optimal approximiert.
- y_i : Zielwert (target, response, output)
- x_i : Regressor (erklärende Variable, input)

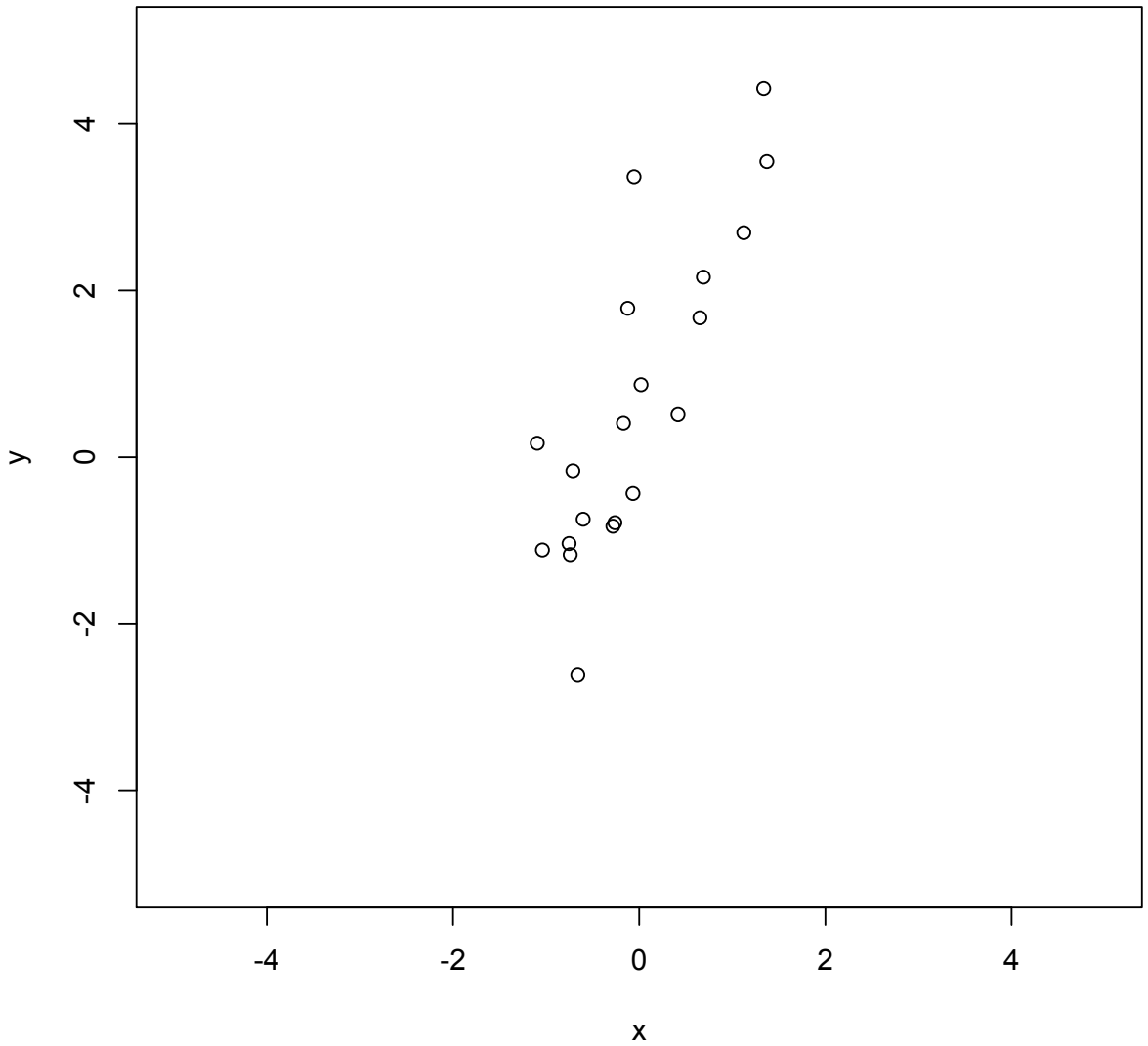
Idee

n Abstände der Punkte (x_i, y_i) zur Gerade in y-Richtung:

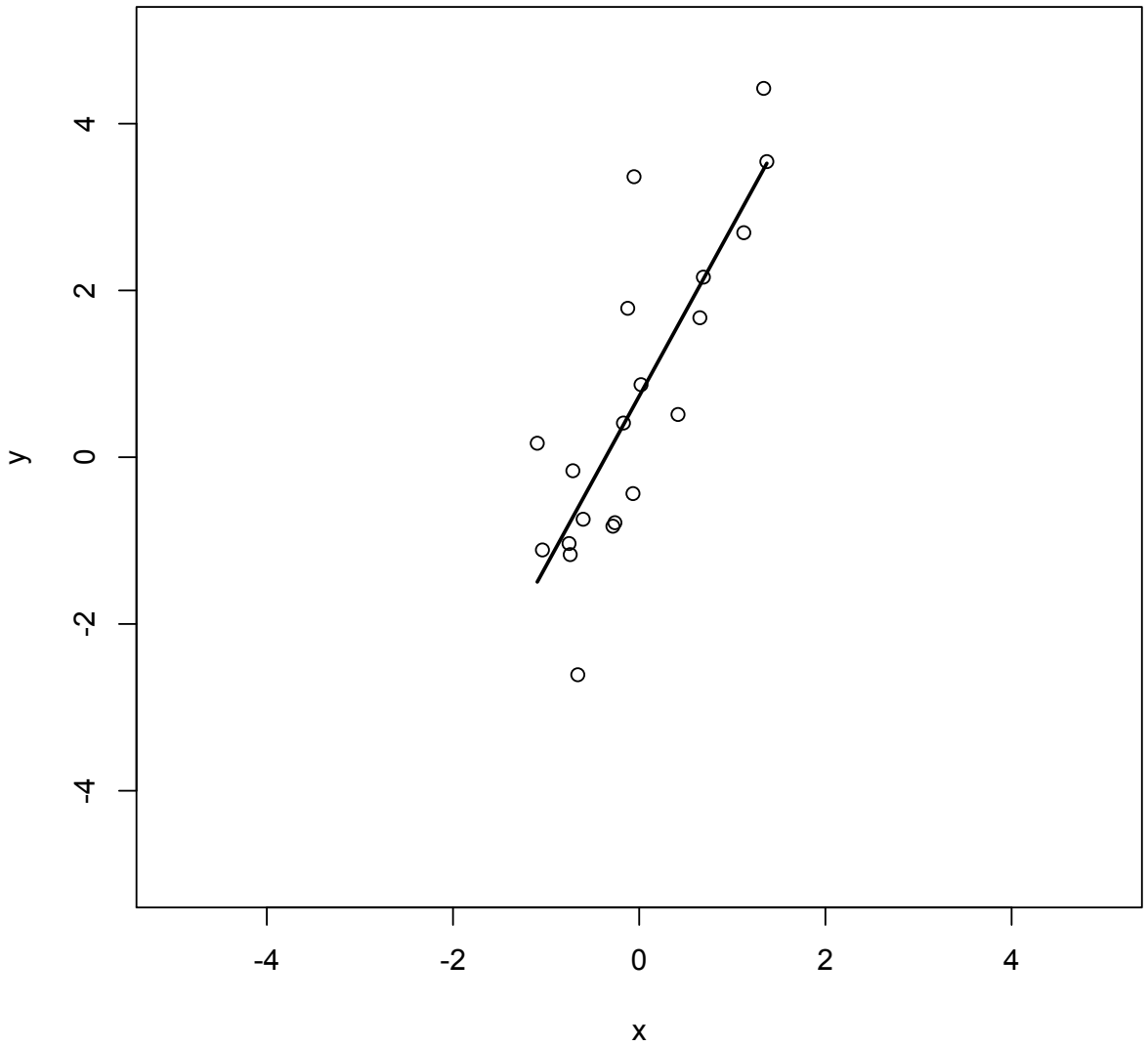
$$|y_i - (a + b \cdot x_i)|, \qquad i = 1, \ldots, n.$$

Alle n quadrierten Abstände $(y_i - (a + b \cdot x_i))^2$ sollen gleichmäßig klein sein.

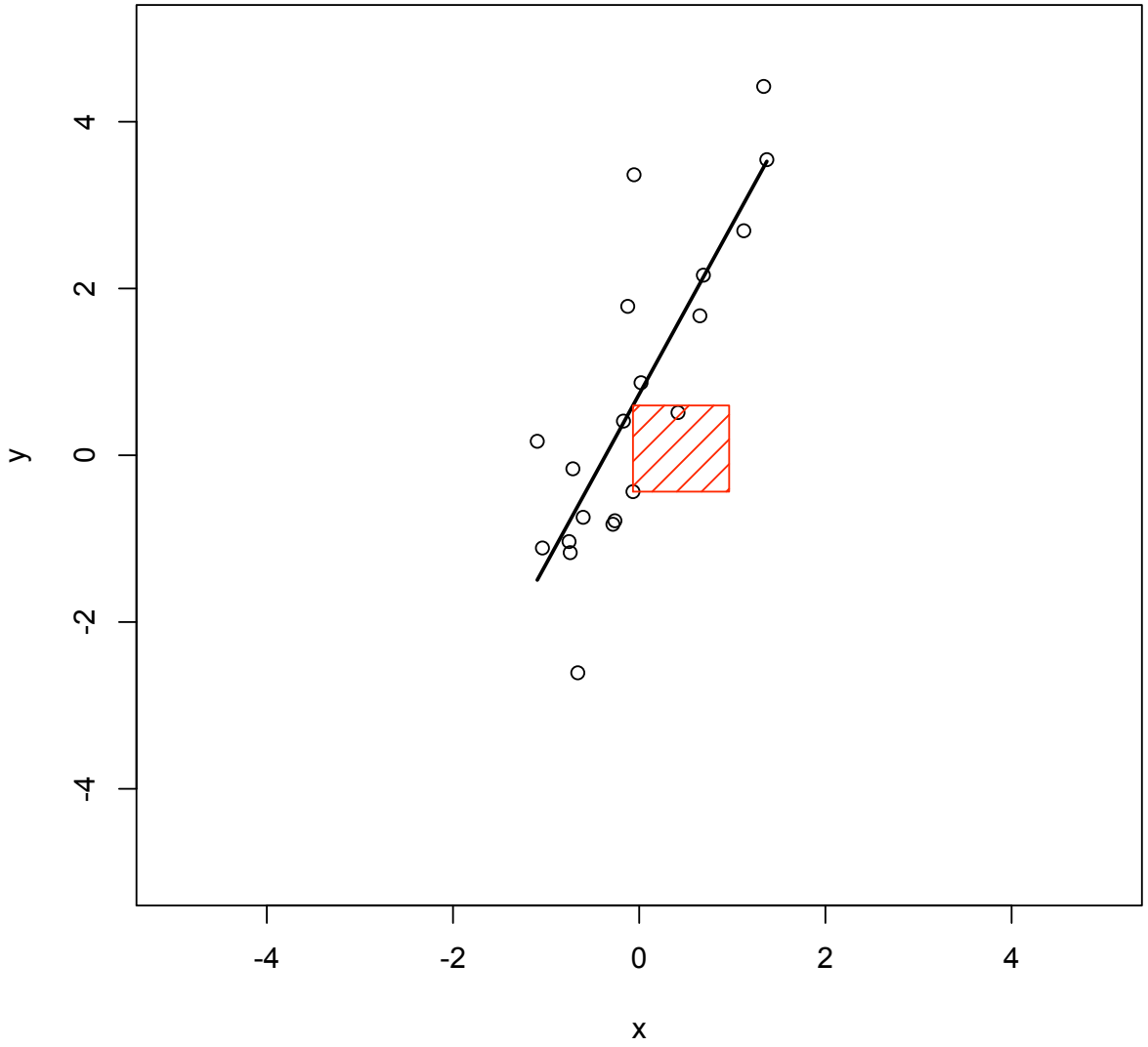
Regression. KQ-Schätzung



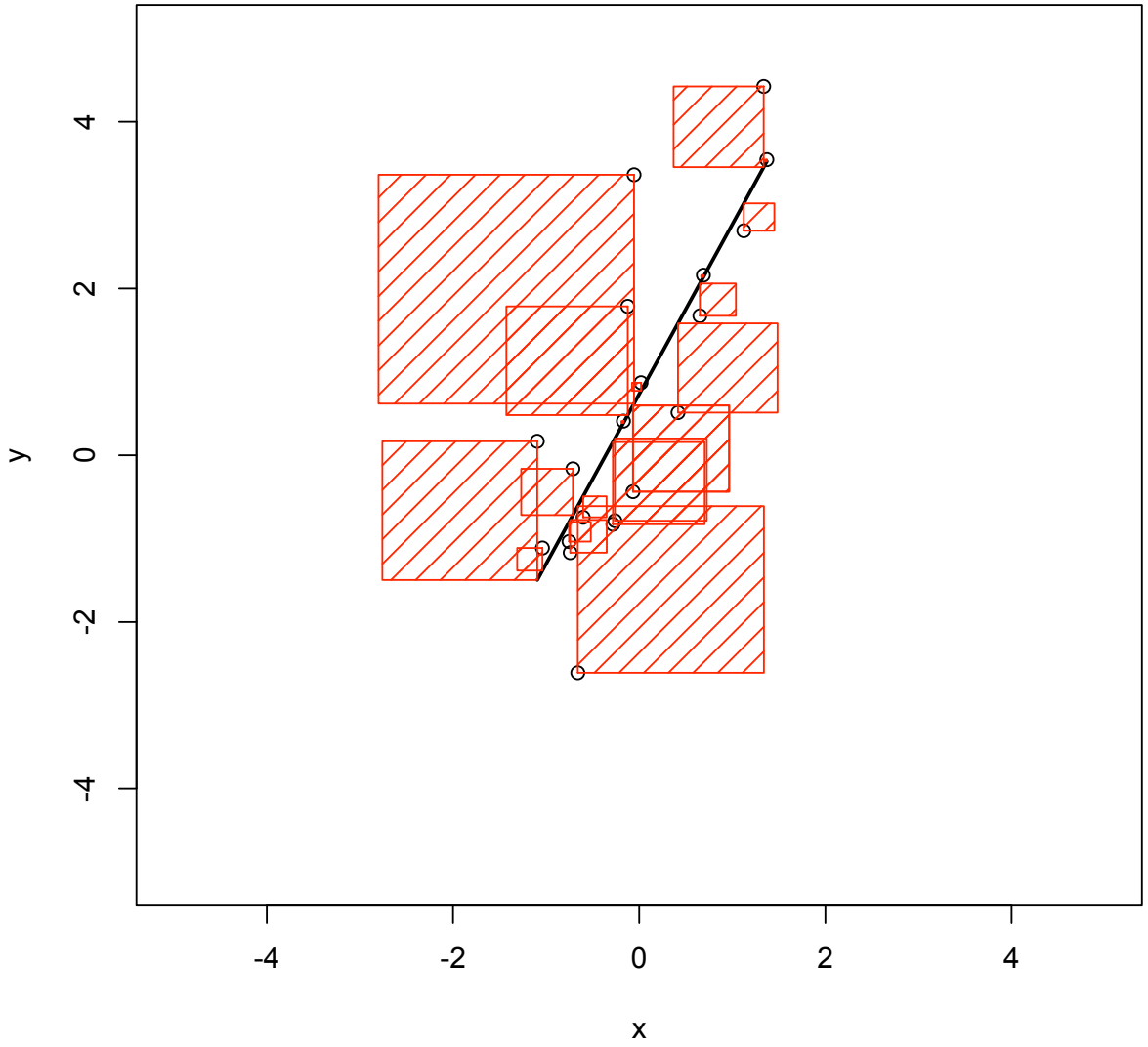
Regression. KQ-Schätzung



Regression. KQ-Schätzung



Regression. KQ-Schätzung



Methode der kleinsten Quadrate

KQ-Methode

Minimiere

$$Q(a,b) = \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2, \qquad (a,b) \in \mathbb{R}^2.$$

Lösungen $(\widehat{a}, \widehat{b})$:

$$\widehat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

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

Methode der kleinsten Quadrate

$Q(a,b)$ stetig partiell differenzierbar mit $\lim_{|a|\to\infty} Q(a,b) = \lim_{|b|\to\infty} Q(a,b) = \infty$.

$$\frac{\partial Q(a,b)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i), \qquad \frac{\partial Q(a,b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i\,.$$

Ist $(\widehat{a},\widehat{b})$ eine Minimalstelle, dann gilt nach dem notwendigen Kriterium 1. Ordnung:

$$0 = -\sum_{i=1}^{n} y_i + n\widehat{a} + \widehat{b}\sum_{i=1}^{n} x_i\,,$$

$$0 = -\sum_{i=1}^{n} y_i x_i + \widehat{a}\sum_{i=1}^{n} x_i + \widehat{b}\sum_{i=1}^{n} x_i^2\,.$$

Dies ist ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten. Division der ersten Gleichung durch $n > 1$ führt auf:

$$0 = -\overline{y} + \widehat{a} + \widehat{b}\cdot\overline{x}.$$

Löst man diese Gleichung nach $\widehat{a}$ auf, so erhält man $\widehat{a} = \overline{y} - \widehat{b}\,\overline{x}$. Einsetzen in die zweite Gleichung und anschließendes Auflösen nach $\widehat{b}$ ergibt

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

Berechnet man die Hesse-Matrix, so stellt sich $(\widehat{a},\widehat{b})$ als Minimalstelle heraus (vgl.

Lineare Regression

Schätzer:

$$\begin{aligned} \widehat{b} &= \frac{\sum_{i=1}^{n} y_i x_i - n \cdot \overline{yx}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2} = \frac{s_{xy}}{s_x^2}, \\ \widehat{a} &= \overline{y} - \widehat{b} \cdot x. \end{aligned}$$

Geschätzte Regressionsgleichung (Ausgleichsgerade):

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad \text{für} \quad x \in [x_{\min}, x_{\max}].$$

Vorhersage- oder Prognosewerte:

$$\widehat{y}_i = \widehat{a} + \widehat{b} \cdot x_i, \qquad i = 1, \ldots, n.$$

Geschätzte Residuen:

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i, \qquad i = 1, \ldots, n,$$

Handrechnungen...

Es gilt:

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n \cdot (\overline{x} \cdot \overline{y}),$$

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n \cdot (\overline{x})^2$$

Daher:

$$\widehat{b} = \frac{\sum_i x_i y_i - n \cdot (\overline{x} \cdot \overline{y})}{\sum_i x_i^2 - n \cdot (\overline{x})^2}$$

Regression

Ausgleichsgerade

Ausgleichs- oder Regressionsgerade:

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad x \in [x_{\min}, x_{\max}]$$

$[x_{\min}, x_{\max}]$: Stützbereich der Regression.

Anpassungsgüte

Streuungszerlegung:

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}) + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = SSR + SSE$$

Bestimmtheitsmaß:

$$R^2 = \frac{SSR}{SST} = r_{XY}^2$$

Residuenplot: Plotte Index $i = 1, \ldots, n$ gegen $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$.

Regression: Zahlenbeispiel

Gegeben seien die folgenden Daten:

x	1	2	3	4	5	6	7
y	1.7	2.6	2.0	2.7	3.2	3.6	4.6

Hieraus berechnet man:

$$\sum_{i=1}^{7} x_i = 28, \qquad \sum_{i=1}^{7} x_i^2 = 140, \qquad \overline{x} = 4\,,$$

$$\sum_{i=1}^{7} y_i = 20.4, \qquad \sum_{i=1}^{7} y_i^2 = 65.3, \qquad \overline{y} = 2.91429\,,$$

sowie $\sum_{i=1}^{7} y_i x_i = 93.5$. Die geschätzten Regressionskoeffizienten lauten somit:

$$\begin{aligned} \widehat{b} &= \frac{\sum_{i=1}^{7} y_i x_i - n \cdot \overline{x}\,\overline{y}}{\sum_{i=1}^{7} x_i^2 - n \cdot \overline{x}^2} \\ &\approx \frac{93.5 - 7 \cdot 4 \cdot 2.91}{140 - 7 \cdot (4)^2} = \frac{12.02}{28} \approx \underline{0.4293}. \end{aligned}$$

$$\widehat{a} = \overline{y} - \widehat{b} \cdot \overline{x} = 2.91 - 0.4293 \cdot 4 = \underline{1.1928}.$$

Die Ausgleichsgerade ist somit gegeben durch:

$$\widehat{f}(x) = 1.1928 + 0.4293 \cdot x, \qquad x \in [1, 7].$$

Lineare Regression

Modell: Unabhängige identisch verteilte Zufallsvektoren

$$(Y, x), (Y_1, x_1), \ldots, (Y_n, x_n)$$

mit

- Y_i: gemessener Wert (zufallsbehaftet) der Zielgröße
- x_i: zugehöriger Wert (fest) der erklärenden Variable

Stochastisches Regressionsmodell:

$$Y_i = a + bx_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

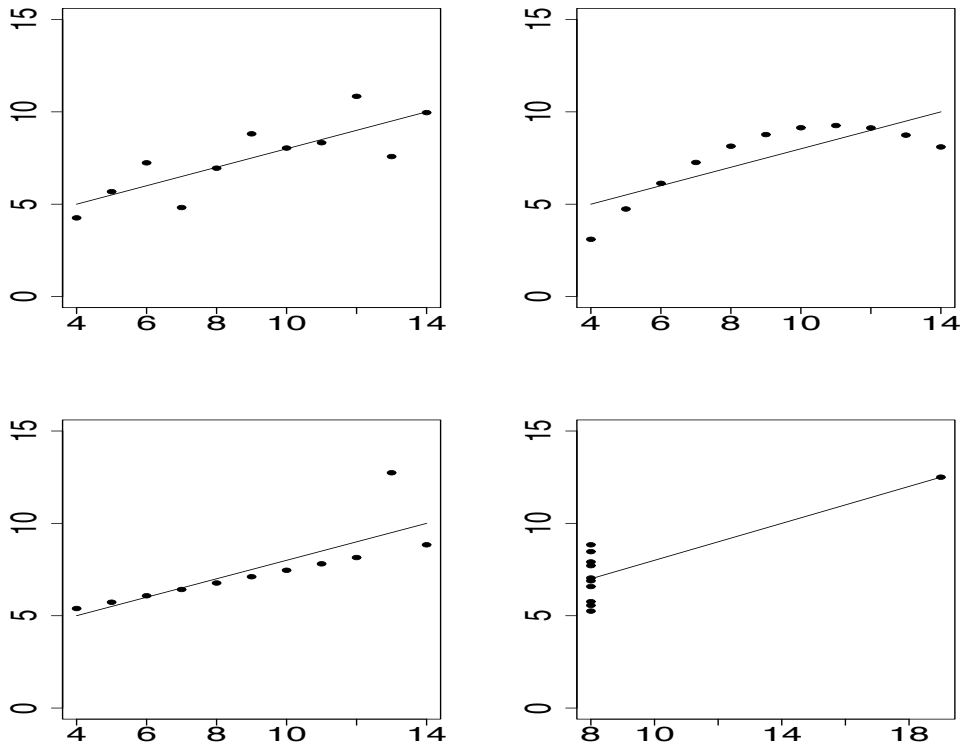
mit Störtermen (Messfehlern, Rauschen (noise)) $\epsilon_1, \ldots, \epsilon_n$,

$$E(\epsilon_i) = 0, \qquad \mathrm{Var}(\epsilon_i) = \sigma^2 \in (0, \infty), \qquad i = 1, \ldots, n.$$

Standardannahmen (klassisch)

1. $\epsilon_1, \ldots, \epsilon_n$ i.i.d $N(0, \sigma^2)$-verteilt.
2. $x_1, \ldots, x_n$ vorgegeben (fest, fixed design).
3. a, b: unbekannte Parameter, **Regressionskoeffizienten**.

Beispiel



4 Datensätze mit identischen Ausgleichsgeraden!

Lineare Regression

Schätzer:

$$\begin{aligned} \widehat{b} &= \frac{\sum_{i=1}^{n} Y_i x_i - n \cdot \overline{Y}\overline{x}}{\sum_{i=1}^{n} x_i^2 - n \cdot (\overline{x})^2} = \frac{s_{xy}}{s_x^2}, \\ \widehat{a} &= \overline{Y} - \widehat{b} \cdot x. \end{aligned}$$

Die Schätzer sind zufällig (Zufallsvariablen/Statistiken).
Geschätzte Regressionsgleichung (Ausgleichsgerade):

$$\widehat{f}(x) = \widehat{a} + \widehat{b} \cdot x, \qquad \text{für} \quad x \in [x_{\min}, x_{\max}].$$

Vorhersage- oder Prognosewerte:

$$\widehat{Y}_i = \widehat{a} + \widehat{b} \cdot x_i, \qquad i = 1, \ldots, n.$$

Geschätzte Residuen:

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \qquad i = 1, \ldots, n,$$

Eigenschaften

Die Schätzer $\widehat{a}$ und $\widehat{b}$ sind erwartungstreu und konsistent für die Regressionskoeffizienten a bzw. b. Ihre Varianzen können durch

$$\widehat{\sigma}_b^2 = \frac{\widehat{\sigma}^2}{n \cdot s_x^2} \qquad \text{sowie} \qquad \widehat{\sigma}_a^2 = \frac{\sum_{i=1}^n x_i^2}{n \cdot s_x^2} \cdot \widehat{\sigma}^2$$

geschätzt werden. Hierbei ist

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

eine erwartungstreue und konsistente Schätzung des Modellfehlers σ^2.

Statistische Eigenschaften

Verteilung

Sind $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $N(0, \sigma^2)$-verteilt, dann gilt:

$$T_b = \frac{\widehat{b} - b}{\widehat{\sigma}_b} \sim t(n-2), \quad T_a = \frac{\widehat{a} - a}{\widehat{\sigma}_a} \sim t(n-2),$$

und

$$Q = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Konfidenzintervall

$(1-\alpha)$-Konfidenzintervall für b:

$$\widehat{b} \pm t(n-2)_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

$(1-\alpha)$-Konfidenzintervall für σ^2:

$$\left[\frac{(n-2)\widehat{\sigma}^2}{\chi^2(n-2)_{1-\alpha/2}}, \frac{(n-2)\widehat{\sigma}^2}{\chi^2(n-2)_{\alpha/2}}\right]$$

Test der Regressionskoeffizienten

Test des Steigungsmaßes b und Intercepts a:

Teststatistiken mit H_0-Wert eingesetzt: $T_a = \frac{\widehat{a}-a_0}{\widehat{\sigma}_a}$, $T_b = \frac{\widehat{b}-b_0}{\widehat{\sigma}_b}$.

1. $H_0 : b = b_0$ gegen $H_1 : b \neq b_0$.
   H_0 ablehnen, wenn $|T_b| > t(n-2)_{1-\alpha/2}$.
2. $H_0 : b \leq b_0$ gegen $H_1 : b > b_0$.
   H_0 ablehnen, falls $T_b > t(n-2)_{1-\alpha}$.
3. $H_0 : b \geq b_0$ gegen $H_1 : b < b_0$.
   H_0 ablehnen, falls $T_b < -t(n-2)_{1-\alpha}$.

Die entsprechenden Tests für den Parameter a erhält man durch Ersetzen von b durch a in den Hypothesen und Ersetzen von T_b durch T_a.

Test des Modellfehlers σ^2:

1. $H_0 : \sigma^2 = \sigma_0^2$ gegen $H_1 : \sigma^2 \neq \sigma_0^2$.
   H_0 ablehnen, wenn $Q < \chi^2(n-2)_{\alpha/2}$ oder $Q > \chi^2(n-2)_{1-\alpha/2}$.
2. $H_0 : \sigma^2 \leq \sigma_0^2$ gegen $H_1 : \sigma^2 > \sigma_0^2$.
   H_0 ablehnen, falls $Q > \chi^2(n-2)_{1-\alpha}$.
3. $H_0 : \sigma^2 \geq \sigma_0^2$ gegen $H_1 : \sigma^2 < \sigma_0^2$.
   H_0 ablehnen, falls $Q < \chi^2(n-2)_{\alpha}$.