Statistische Methoden der Datenanalyse

W. Waltenberger
wolfgang.waltenberger@gmail.com

Basierend auf den Folien der Vorlesung des Jahres 2018 von R. Frühwirth

VU 260014 (Uni) / SV 142.340 (TU)

23. November 2020

Übersicht über die Vorlesung

Übersicht über die Vorlesung

Teil 5: Konfidenzintervalle

Teil 6: Testen von Hypothesen

Teil 7: Regression und lineare Modelle

Teil 8: Bayesstatistik

Teil I

Deskriptive Statistik

Übersicht Teil 1

Statistische Methoden der Datenanalyse

W. Waltenberger

Abschnitt 1: Einleitung

Unterabschnitt: Grundbegriffe

Grundbegriffe

Definition von Statistik

- Die Erhebung und Speicherung von Daten, z.B. durch statistische Ämter
- Die mathematische Auswertung von Daten, z.B. die Berechnung von Maß- und Kennzahlen, die Schätzung von unbekannten Parametern, das Testen von Hypothesen

Deskriptive Statistik

- Beschreibung von vorhandenen Daten durch Maßzahlen, Tabellen, Graphiken

Grundbegriffe

Induktive Statistik

- Untersuchung von Gesetzmäßigkeiten und Ursachen, die hinter den Daten stehen und die Daten (teilweise) erklären.
- Explorative Datenanalyse: Ziel ist, Hypothesen für die Theoriebildung zu gewinnen
- Konfirmative Datenanalyse: Ziel ist, vorhandene Theorien zu prüfen, z.B. durch Schätzen von Parametern oder Testen von Hypothesen

Unterabschnitt: Merkmal- und Skalentypen

Merkmal- und Skalentypen

Qualitative Merkmale

- **binär** (ja/nein). Beispiel: EU-Bürgerschaft.
- **kategorial** (Klassifizierung).
  Beispiel: ledig/geschieden/verheiratet/verwitwet.
- **ordinal** (Rang). Beispiel: Noten 1–5.

Quantitative Merkmale

- **diskret** (ganzzahlig). Beispiel: Zählvorgang.
- **kontinuierlich** (reellwertig). Beispiel: Messvorgang.

Merkmal- und Skalentypen

Skalentypen

- **Nominalskala**: Zahlenwerte sind nur Bezeichnung für sich ausschließende Kategorien.
- **Ordinalskala**: Ordnung der Zahlen ist wesentlich.
- **Intervallskala**: Ordnung und Differenzen zwischen den Werten sind sinnvoll interpretierbar, der Nullpunkt ist willkürlich festgelegt.
- **Verhältnisskala**: Ordnung, Differenzen und Größenverhältnisse sind sinnvoll interpretierbar, es gibt einen absoluten Nullpunkt.

Merkmal- und Skalentypen

Beispiel

- Der Familienstand einer Person wird durch Zahlen kodiert (1=ledig, 2=verheiratet, 3=geschieden, 4=verwitwet). Nominalskala.
- Der Stand einer Mannschaft in der Meisterschaft wird durch den Rang in der Liga angegeben. Ordinalskala.
- Die Jahreszahlen (2007, 2008, ...) bilden eine Intervallskala, da der Nullpunkt willkürlich festgelegt ist.
- Die Celsius-Skala der Temperatur ist eine Intervallskala, da der Nullpunkt willkürlich festgelegt ist.
- Die Kelvin-Skala der Temperatur ist eine Verhältnisskala, da der Nullpunkt physikalisch festgelegt ist.
- Die Größe einer Person wird in cm angegeben. Es liegt eine Verhältnisskala vor, da ein natürlicher Nullpunkt existiert.

Merkmal- und Skalentypen

Beispiel

In der folgenden Datenmatrix D sind Merkmale von acht Personen zusammengestellt.

Nummer	Geschlecht	Alter	Ausbildung
1	1	34	2
2	2	54	1
3	2	46	3
4	1	27	4
5	1	38	2
6	1	31	3
7	2	48	4
8	2	51	2

Geschlecht: 1=W, 2=M, Alter: in Jahren
Ausbildung: 1=Pflichtschule, 2=Höhere Schule, 3=Bachelor, 4=Master

Unterabschnitt: Aussagen und Häufigkeiten

Aussagen und Häufigkeiten

Der Begriff der Aussage

- Wir nennen die Menge der Untersuchungsobjekte die **Grundgesamtheit** Ω.
- Eine **Aussage** $A(x)$ ist eine Feststellung über Eigenschaften der Elemente $x \in \Omega$.
- Für jedes $x \in \Omega$ muss $A(x)$ entweder **wahr** oder **falsch** sein.

Beispiel

Es sei $A(x)$ die Aussage "x ist weiblich". Dann ist $A(1)$ wahr und $A(2)$ falsch.

Beispiel

Es sei $B(x)$ die Aussage "x ist über 50 Jahre alt". Dann ist $B(2)$ wahr und $B(6)$ falsch.

Aussagen und Häufigkeiten

Verknüpfung von Aussagen

Es seien A und B zwei Aussagen.

Symbol	Name	Bedeutung
$A \cup B$	Disjunktion	A oder B (oder beide)
$A \cap B$	Konjunktion	A und B (sowohl A als auch B)
A'	Negation	nicht A (das Gegenteil von A)
$A \subseteq B$	Implikation	aus A folgt B ($A' \cup B$)

Aussagen und Häufigkeiten

Definition (Absolute Häufigkeit)

Es sei $A(x)$ eine Aussage über $x \in \Omega$. Die **absolute Häufigkeit** $h(A)$ von A ist die Anzahl der Elemente von Ω, für die $A(x)$ wahr ist.

Beispiel

Sei $A(x)$ ist die Aussage “Die Person $x \in D$ hat zumindest Bakkalaureat”. Dann ist $h(A) = 4$.

Aussagen und Häufigkeiten

Definition (Relative Häufigkeit)

Es sei $A(x)$ eine Aussage über $x \in \Omega$. Die **relative Häufigkeit** $f(A) = h(A)/n$ von A ist die absolute Häufigkeit $h(A)$ dividiert durch die Gesamtanzahl $n = |\Omega| \in \mathbb{N}$ der Elemente.

Beispiel

Sei $B(x)$ die Aussage "Die Person $x \in D$ ist älter als dreißig Jahre". Dann ist $f(B) = 7/8$.

Aussagen und Häufigkeiten

Spezielle Aussagen

- $A = \emptyset$: A trifft niemals zu, $h(A) = f(A) = 0$.
- $A = \Omega$: A trifft immer zu, $h(A) = n, f(A) = 1$.

Rechengesetze für Häufigkeiten

Additionsgesetz

$$A \cap B = \emptyset \implies \begin{cases} h(A \cup B) = h(A) + h(B) \\ f(A \cup B) = f(A) + f(B) \end{cases}$$

Aussagen und Häufigkeiten

Siebformel

$$h(A \cup B) = h(A) + h(B) - h(A \cap B)$$
$$f(A \cup B) = f(A) + f(B) - f(A \cap B)$$

Beispiel

33% der Kunden einer Bank haben einen Wohnungskredit, 24% haben einen Kredit zur Finanzierung von Konsumgütern, 11% haben beides. Wie groß ist der Anteil der Kunden, die weder Wohnungs- noch Konsumgüterkredit haben?

Abschnitt 2: Eindimensionale Merkmale

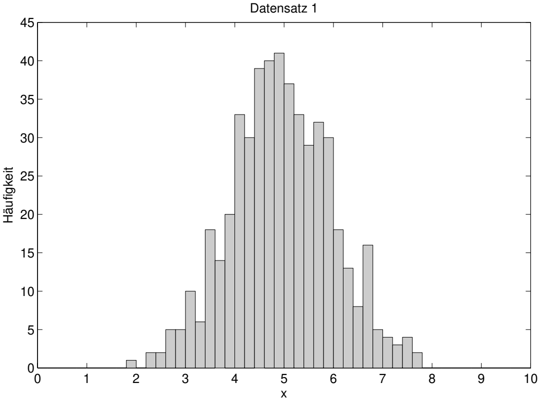
Unterabschnitt: Graphische Darstellung

Graphische Darstellung

- Ein Bild sagt mehr als tausend Worte!
- Graphische Darstellungen von Datensätzen sind daher äußerst beliebt und nützlich.
- Qualitative Variable: Häufigkeitstabelle, Tortendiagramm, Stabdiagramm
- Quantitative Variable: gruppierte Häufigkeitstabelle, Histogramm, Boxplot, empirische Verteilungsfunktion

Graphische Darstellung

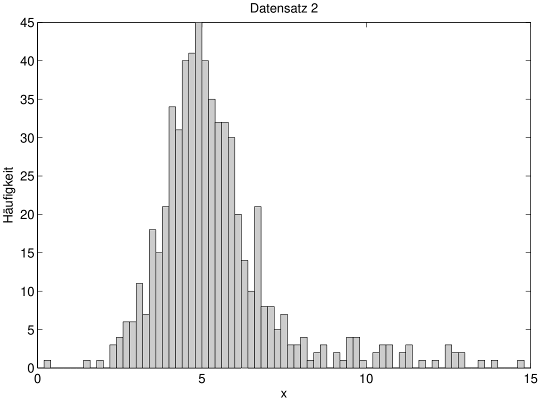
- Datensatz 1 (500 normalverteilte Werte):



Histogramm

Graphische Darstellung

- Datensatz 2 = Datensatz 1 + Kontamination (100 Werte):



Histogramm

Graphische Darstellung

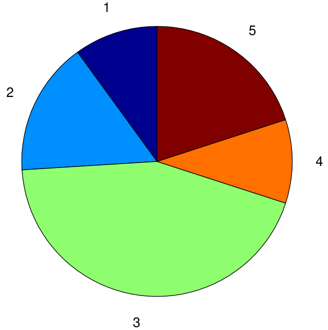
- Datensatz 3 (50 Prüfungsnoten):

Note k	$h(k)$	$f(k)$
1	5	0.10
2	8	0.16
3	22	0.44
4	5	0.10
5	10	0.20
	50	1.00

Häufigkeitstabelle

Graphische Darstellung

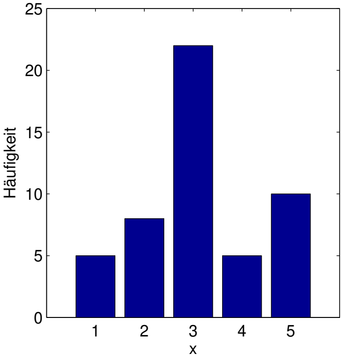
- Datensatz 3 (50 Prüfungsnoten):



Tortendiagramm

Graphische Darstellung

- Datensatz 3 (50 Prüfungsnoten):



Stabdiagramm

Unterabschnitt: Kernschätzer

Kernschätzer

- Die Häufigkeitsverteilung (Histogramm) kann mit einem Kern- oder Dichteschätzer geglättet werden.
- Die Dichte des beobachteten Merkmals wird dabei durch eine Summe von Kernen $K(\cdot)$ approximiert:

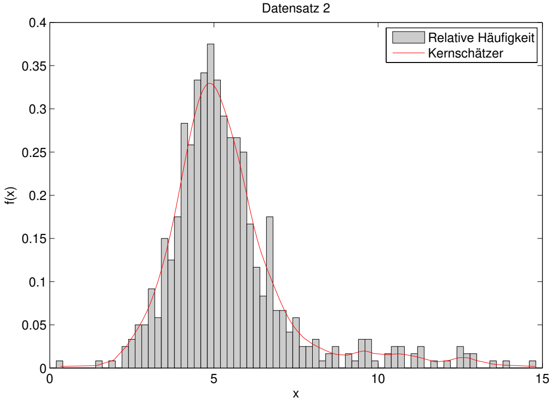
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- h ist die Bandbreite des Kernschätzers.
- Ein beliebter Kern ist der Gaußkern:

$$K(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Kernschätzer

- Datensatz 2:



Glättung des Histogramms durch Kernschätzer

Python: `sklearn.neighbors.KernelDensity`

Unterabschnitt: Maßzahlen

Maßzahlen

- Datenlisten sind oft so umfangreich, dass ihr Inhalt in einigen wenigen Maßzahlen zusammgefasst wird oder werden muss. Welche Maßzahlen dabei sinnvoll sind, hängt vom Skalentyp ab.
- Manche Maßzahlen gehen von der geordneten Datenliste $x_{(1)}, \ldots, x_{(n)}$ aus.
- Wir unterscheiden Lage-, Streuungs-, und Schiefemaße.
- Ein Lagemaß gibt an, um welchen Wert die Daten konzentriert sind.
- Ein Streuungsmaß gibt an, wie groß die Schwankungen der Daten um ihren zentralen Wert sind.
- Ein Schiefemaß gibt an, wie symmetrisch die Daten um ihren zentralen Wert liegen.

Maßzahlen

Lagemaße

Definition (Lagemaß)

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ eine Datenliste. Die Funktion $\ell(\boldsymbol{x})$ heißt ein Lagemaß für $\boldsymbol{x}$, wenn gilt:

- $\ell(a\boldsymbol{x} + b) = a\ell(\boldsymbol{x}) + b$
- $\min(\boldsymbol{x}) \leq \ell(\boldsymbol{x}) \leq \max(\boldsymbol{x})$

- Sinnvolle Lagemaße geben den “typischen” oder “zentralen” Wert der Datenliste an.
- Je nach Skala sind verschiedene Lagemaße sinnvoll.

Maßzahlen

Mittelwert der Daten

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.
- Der Mittelwert minimiert die folgende Funktion:

$$\bar{x} = \operatorname*{argmin}_x \sum_{i=1}^{n} (x_i - x)^2$$

PYTHON: `xbar=numpy.mean(x)`

Maßzahlen

Median der Daten

$$\tilde{x} = x_{(n/2)}$$

- Der Median teilt die geordnete Datenliste in zwei gleich große Teile.
- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.
- Der Median minimiert die folgende Funktion:

$$\tilde{x} = \operatorname*{argmin}_x \sum_{i=1}^{n} |x_i - x|$$

PYTHON: `xmed=numpy.median(x)`

Maßzahlen

- Der Median ist ein Spezialfall eines allgemeineren Begriffs, des Quantils.

α-Quantil

$$Q_\alpha = x_{(\alpha n)}$$

- Das α-Quantil teilt die geordnete Datenliste im Verhältnis $\alpha : 1 - \alpha$.
- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.
- PYTHON: `qa=numpy.quantile(x,alpha)`
- Q_0 ist der kleinste Wert, Q_1 ist der größte Wert der Datenliste. $Q_{0.5}$ ist der Median.
- Die fünf **Quartile** $Q_0, Q_{0.25}, Q_{0.5}, Q_{0.75}, Q_1$ bilden das **five point summary** der Datenliste.
- PYTHON: `fps=numpy.quantile(x,[0,0.25,0.5,0.75,1])`

Maßzahlen

LMS (Least Median of Squares)

Der LMS-Wert ist der Mittelpunkt des kürzesten Intervalls, das $h = \lfloor n/2 \rfloor + 1$ Datenpunkte enthält.

- Der LMS-Wert ist extrem unempfindlich gegen fehlerhafte oder untypische Daten.
- Der LMS-Wert minimiert die folgende Funktion:

$$\tilde{x} = \operatorname*{argmin}_x \operatorname{med}_{i=1}^n (x_i - x)^2$$

- Ein verwandtes Lagemaß ist der "shorth", der Mittelwert aller Daten im kürzesten Intervall, das h Datenpunkte enthält.

Maßzahlen

Modus

Der Modus ist der häufigste Wert einer Datenliste

- Sinnvoll vor allem für qualitative Merkmale.
- Für quantitative Merkmale kann der Modus aus dem Kernschätzer der Dichte bestimmt werden.
- PYTHON: `xmode=scipy.stats.mode(x)`

HSM (Half-sample mode)

- Bestimme das kürzeste Intervall, das $h = \lfloor n/2 \rfloor + 1$ Datenpunkte enthält.
- Wiederhole den Vorgang auf den Daten in diesem Intervall, bis zwei Datenpunkte übrig sind.
- Der HSM-Wert ist das Mittel der beiden letzten Daten.

Maßzahlen

Streuungsmaße

Definition (Streuungsmaß)

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ eine Datenliste. Die Funktion $\sigma(\boldsymbol{x})$ heißt ein Streuungsmaß für $\boldsymbol{x}$, wenn gilt:

- $\sigma(\boldsymbol{x}) \geq 0$
- $\sigma(a\boldsymbol{x} + b) = |a|\,\sigma(\boldsymbol{x})$

- Sinnvolle Streuungsmaße messen die Abweichung der Daten von ihrem zentralen Wert.
- Streuungsmaße sind invariant unter Verschiebung der Daten.
- Je nach Skala sind verschiedene Streuungsmaße sinnvoll.

Maßzahlen

Standardabweichung der Daten

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

- Sinnvoll für Intervall- und Verhältnisskala.
- Die Standardabweichung hat die gleiche Dimension wie die Daten.
- Das Quadrat der Standardabweichung heißt **Varianz der Daten**.
- PYTHON: `xstd=numpy.std(x)`
- PYTHON: `xvar=numpy.var(x)`

Maßzahlen

Interquartilsdistanz

$$IQR = Q_{0.75} - Q_{0.25}$$

- Die Interquartilsdistanz ist die Länge des Intervalls, das die zentralen 50% der Daten enthält.
- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.

PYTHON: `xiqr=scipy.stats.iqr(x)`

Maßzahlen

LoS (Length of the Shorth)

LoS ist die Länge des kürzesten Intervalls, das $h = \lfloor n/2 \rfloor + 1$ Datenpunkte enthält.

- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.

Maßzahlen

Schiefemaße

Definition (Schiefemaß)

Es sei $\boldsymbol{x} = (x_1, \ldots, x_n)$ eine Datenliste. Die Funktion $\mathfrak{s}(\boldsymbol{x})$ heißt ein Schiefemaß für $\boldsymbol{x}$, wenn gilt:

- $\mathfrak{s}(a\boldsymbol{x} + b) = \operatorname{sgn}(a)\,\mathfrak{s}(\boldsymbol{x})$
- $\mathfrak{s}(\boldsymbol{x}) = 0$, wenn $\exists b : \boldsymbol{x} - b = b - \boldsymbol{x}$

- Sinnvolle Schiefemaße messen die Asymmetrie der Daten.
- Schiefemaße sind invariant unter Verschiebung der Daten.
- Je nach Skala sind verschiedene Schiefemaße sinnvoll.

Maßzahlen

Schiefe der Daten

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

- Die Schiefe γ ist gleich 0 für symmetrische Daten.
- Ist $\gamma < 0$, heißen die Daten **linksschief**.
- Ist $\gamma > 0$, heißen die Daten **rechtsschief**.
- Sinnvoll für Intervall- und Verhältnisskala.

PYTHON: `xgamma=scipy.stats.skew(x)`

Maßzahlen

Schiefekoeffizient

$$SK = \frac{R - L}{R + L}$$

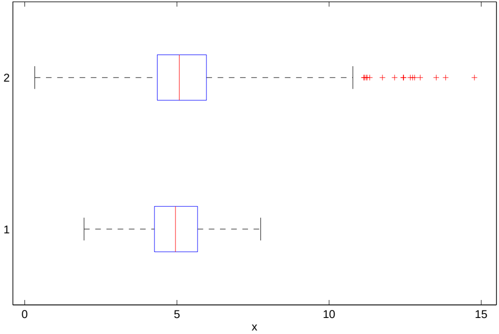
mit $R = Q_{0.75} - Q_{0.5}, L = Q_{0.5} - Q_{0.25}$.

- SK liegt zwischen -1 $(R = 0)$ und $+1$ $(L = 0)$.
- Der Schiefekoeffizient ist gleich 0 für symmetrische Daten.
- Ist $SK < 0$, heißen die Daten **linksschief**.
- Ist $SK > 0$, heißen die Daten **rechtsschief**.
- Sinnvoll für Ordinal-, Intervall- und Verhältnisskala.

Unterabschnitt: Boxplot und Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

- Der **Boxplot** ist die graphische Darstellung des five point summary.
- Vergleich von Datensatz 1 und Datensatz 2:



PYTHON: `matplotlib.pyplot.boxplot(x)`

Boxplot und Empirische Verteilungsfunktion

- Ab Ordinalskala ist es sinnvoll, die Daten zu ordnen.
- Die Häufigkeitstabelle kann durch Summenhäufigkeiten ergänzt werden.
- Datensatz 3 (50 Prüfungsnoten):

Note k	$h(k)$	$H(k)$	$f(k)$	$F(k)$
1	5	5	0.10	0.10
2	8	13	0.16	0.26
3	22	35	0.44	0.70
4	5	40	0.10	0.80
5	10	50	0.20	1.00

Häufigkeitstabelle mit Summenhäufigkeiten

Boxplot und Empirische Verteilungsfunktion

- Die graphische Darstellung der Summenhäufigkeiten wird die **empirische Verteilungsfunktion** der Datenliste genannt.

Definition (Empirische Verteilungsfunktion)

Die empirische Verteilungsfunktion $F_n(x)$ der Datenliste $\boldsymbol{x} = (x_1, \ldots, x_n)$ ist der Anteil der Daten, die kleiner oder gleich x sind:

$$F_n(x) = f(\vec{x} \leq x).$$

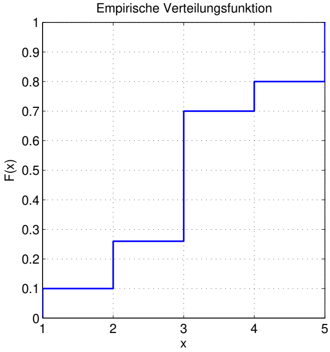
- Ist $x_i \leq x < x_{i+1}$, gilt

$$F_n(x) = f(x_1) + \cdots + f(x_i).$$

- F_n ist eine Sprungfunktion. Die Sprungstellen sind die Datenpunkte, die Sprunghöhen sind die relativen Häufigkeiten der Datenpunkte.

Boxplot und Empirische Verteilungsfunktion

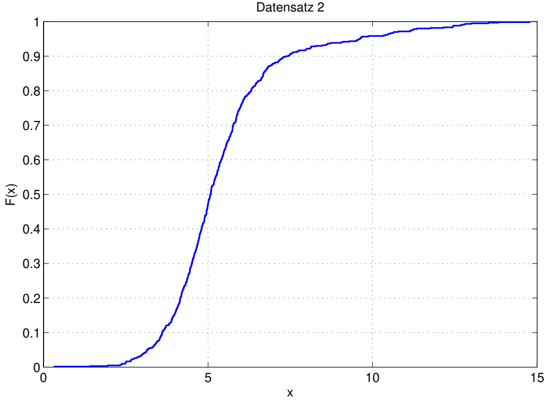
- Datensatz 3: (50 Prüfungsnoten):



Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

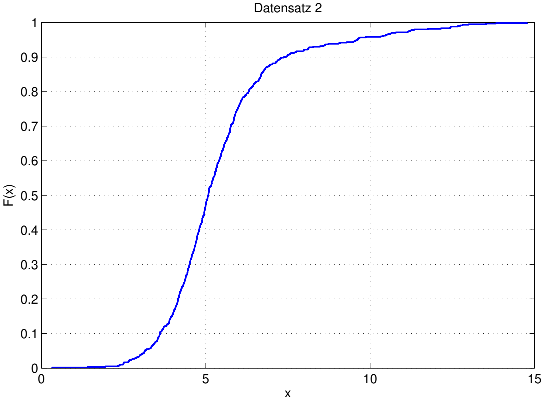
- Datensatz 2 (500 Werte + Kontamination):



Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

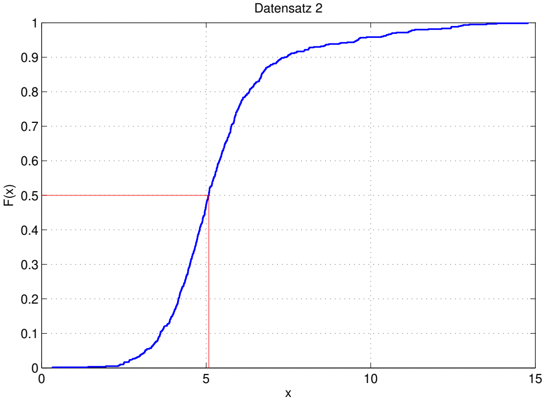
- Aus der empirischen Verteilungsfunktion können Quantile einfach abgelesen werden.
- Median von Datensatz 2:



Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

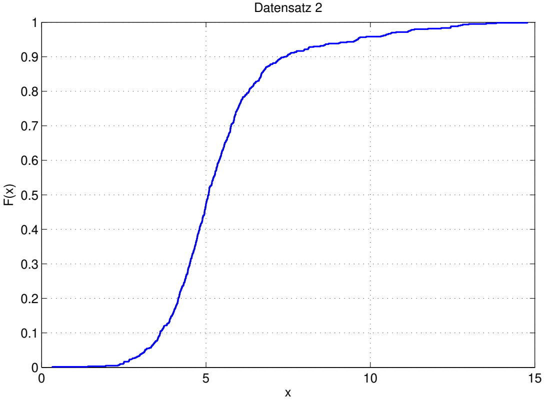
- Aus der empirischen Verteilungsfunktion können Quantile einfach abgelesen werden.
- Median von Datensatz 2:



Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

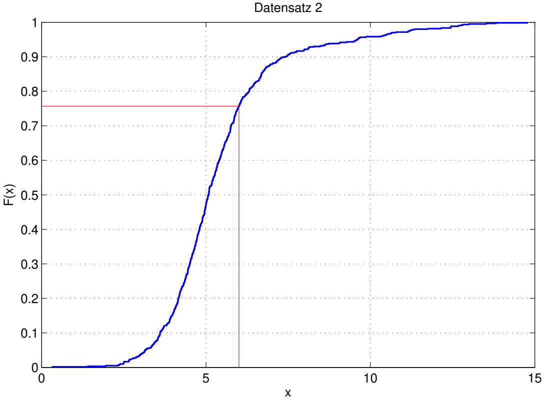
- Es können auch Unter- und Überschreitungshäufigkeiten abgelesen werden.
- Welcher Anteil der Daten ist kleiner oder gleich 6?



Empirische Verteilungsfunktion

Boxplot und Empirische Verteilungsfunktion

- Es können auch Unter- und Überschreitungshäufigkeiten abgelesen werden.
- Welcher Anteil der Daten ist kleiner oder gleich 6?



Empirische Verteilungsfunktion

Unterabschnitt: Beispiele

Beispiele

- Datensatz 1: Symmetrisch, 500 Werte

- Lagemaße:

Mittelwert:	4.9532
Median:	4.9518
LMS:	4.8080
Shorth:	4.8002
HSM:	5.0830

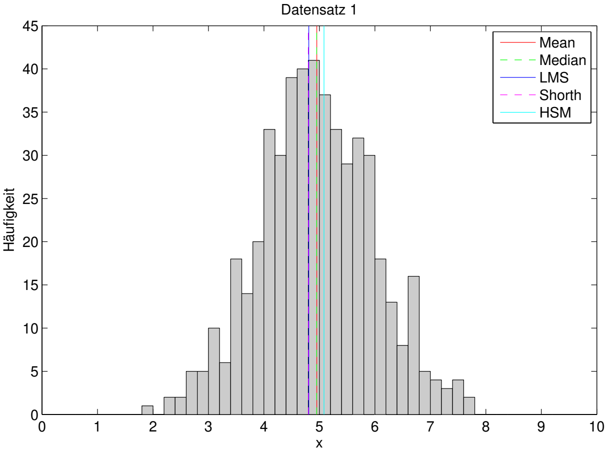
- Streuungsmaße:

Standardabweichung:	1.0255
Interquartilsdistanz:	1.4168
Length of the Shorth:	1.3520

- Schiefemaße:

Schiefe:	0.0375
Schiefekoeffizient:	0.0258

Beispiele



Datensatz 1: Mittelwert, Median, LMS, Shorth, HSM

Beispiele

- Datensatz 2: Datensatz 1 + Kontamination (100 Werte)

- Lagemaße:

Mittelwert:	5.4343
Median:	5.0777
LMS:	5.1100
Shorth:	5.0740
HSM:	4.9985

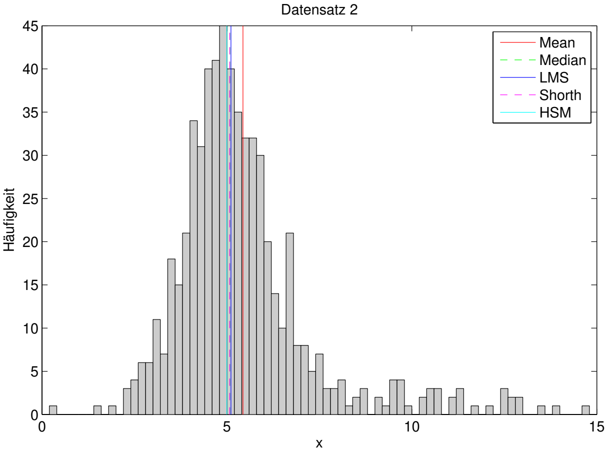
- Schiefemaße:

Schiefe:	1.7696
Schiefekoeffizient:	0.1046

- Streuungsmaße:

Standardabweichung:	1.8959
Interquartilsdistanz:	1.6152
Length of the Shorth:	1.5918

Beispiele



Datensatz 2: Mittelwert, Median, LMS, Shorth, HSM

Beispiele

- Datensatz 3: 50 Prüfungsnoten

- Lagemaße:

Mittelwert:	3.14
Median:	3.0
Modus:	3.0

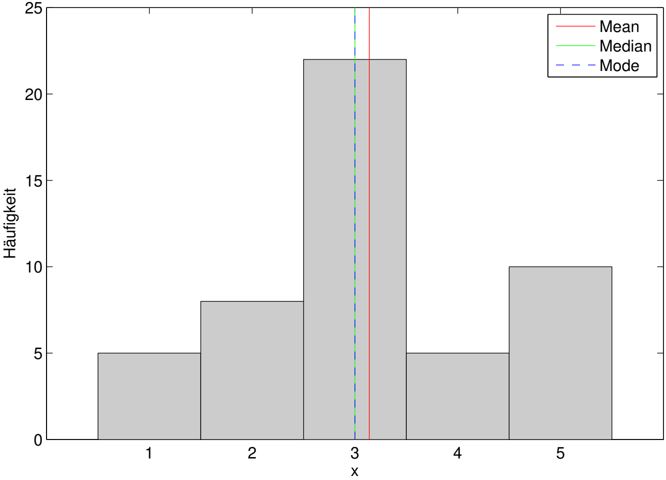
- Streuungsmaße:

Standardabweichung:	1.20
Interquartilsdistanz:	1.75

- Schiefemaße:

Schiefe:	0.0765
Schiefekoeffizient:	0.14

Beispiele



Datensatz 3: Mittelwert, Median, Modus

Abschnitt 3: Zweidimensionale Merkmale

Zweidimensionale Merkmale

- Oft werden zwei oder mehr Merkmale eines Objekts **gleichzeitig** beobachtet.
- Beispiele:
  - Körpergröße und Gewicht einer Person
  - Alter und Einkommen einer Person
  - Schulbildung und Geschlecht einer Person
- Der Zusammenhang zwischen den beiden Merkmalen gibt zusätzliche Information.

Unterabschnitt: Qualitative Merkmale

Qualitative Merkmale

- Wir betrachten zunächst zwei binäre Merkmale A und B.
- Die Häufigkeit des Eintretens von A und B kann in einer **Vierfeldertafel** oder **Kontingenztafel** zusammengefasst werden.
- **Beispiel**:

  A=*"Die Person ist weiblich"*
  B=*"Die Person ist Raucher/in"*

- Vierfeldertafel für 1000 Personen:

	B	B'	
A	228	372	600
A'	136	264	400
	364	636	1000

Qualitative Merkmale

- Allgemeiner Aufbau einer Vierfeldertafel:

	B	B'	
A	$h(A \cap B)$	$h(A \cap B')$	$h(A)$
A'	$h(A' \cap B)$	$h(A' \cap B')$	$h(A')$
	$h(B)$	$h(B')$	n

- Zeilen- und Spaltensummen sind die Häufigkeiten der Ausprägungen A, A' und B, B'.

Qualitative Merkmale

- Die Vierfeldertafel kann mittels Division durch n auf **relative Häufigkeiten** umgerechnet werden:

	B	B'	
A	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
A'	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	1

- Zeilen- und Spaltensummen sind die relativen Häufigkeiten der Ausprägungen A, A' und B, B'.

Qualitative Merkmale

- Der Zusammenhang der beiden Merkmale kann durch die **Vierfelderkorrelation** gemessen werden:

Vierfelderkorrelation

$$\rho(A, B) = \frac{f(A \cap B) - f(A)f(B)}{\sqrt{f(A)f(A')f(B)f(B')}}$$

- Es gilt stets: $-1 \leq \rho(A, B) \leq 1$
- Ist $\rho(A, B) > 0$, heißen A und B **positiv gekoppelt**.
- Ist $\rho(A, B) < 0$, heißen A und B **negativ gekoppelt**.

Qualitative Merkmale

- Das Vorzeichen von $\rho(A, B)$ gibt die **Richtung** der Koppelung an.
- Der Betrag von $\rho(A, B)$ gibt die **Stärke** der Koppelung an.
- Speziell gilt:

$$A = B \implies \rho(A, B) = 1$$
$$A = B' \implies \rho(A, B) = -1$$

- Eine bestehende Koppelung ist kein Beweis für einen kausalen Zusammenhang!
- Die Koppelung kann auch durch eine gemeinsame Ursache für beide Merkmale entstehen.

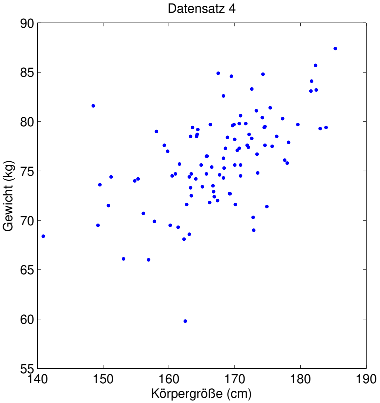
Unterabschnitt: Quantitative Merkmale

Quantitative Merkmale

- Bevorzugte Darstellung von zweidimensionalen Merkmalen: Streudiagramm (Scatter Plot)
- Jeder Punkt entspricht einem Objekt.
- Die beobachteten Merkmale bestimmen die Position des Punktes in der x-y-Ebene.
- Höherdimensionale Merkmale können durch Histogramme und Streudiagramme dargestellt werden. Dabei geht natürlich ein Teil der Information verloren.

Quantitative Merkmale

- Datensatz 4: Körpergröße und Gewicht von 100 Personen



Streudiagramm

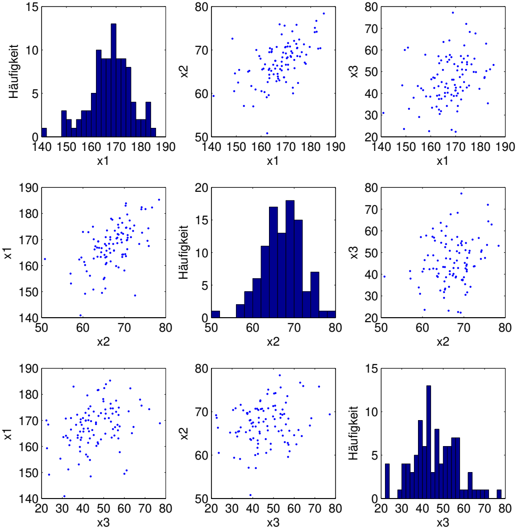
PYTHON: `matplotlib.pyplot.scatter(x,y)`

Quantitative Merkmale

- Datensatz 5:
  Körpergröße, Gewicht und Alter von 100 Personen

Merkmal x_1:	Körpergröße (in cm)
Merkmal x_2:	Gewicht (in kg)
Merkmal x_3:	Alter (in Jahren)

Quantitative Merkmale



Unterabschnitt: Korrelation

Korrelation

Eigenschaften des Streudiagramms

- $(\bar{x}, \bar{y})$ ist der Mittelpunkt der Punktwolke.
- Die Projektion der Punktwolke auf die x-Achse ergibt das Punktediagramm der Datenliste $x_1, \ldots, x_n$.
- Die Projektion der Punktwolke auf die y-Achse ergibt das Punktediagramm der Datenliste $y_1, \ldots, y_n$.

- Aus dem Streudiagramm von Datensatz 4 ist ersichtlich, dass **tendenziell** größere Körpergröße mit größerem Gewicht einhergeht.
- Zwischen den beiden Merkmalen x und y besteht offensichtlich ein Zusammenhang, der auch intuitiv völlig klar ist.

Korrelation

- Wir brauchen eine **Maßzahl** für diesen Zusammenhang.
- Eine nützliche Maßzahl ist der **empirische Korrelationskoeffizient**.
- Sei $(x_1, y_1), \dots, (x_n, y_n)$ eine bivariate Stichprobe.
- Wir berechnen die **Standardscores**:

$$z_{x,i} = \frac{x_i - \bar{x}}{s_x}, \quad z_{y,i} = \frac{y_i - \bar{y}}{s_y}$$

- Wir erinnern uns, dass

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{und} \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Der empirische Korrelationskoeffizient ist der **Mittelwert der Produkte** der Standardscores.

Korrelation

Definition (Empirischer Korrelationskoeffizient)

Der **empirische Korrelationskoeffizient** r_{xy} ist definiert als

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n} z_{x,i} z_{y,i} = \frac{1}{n}\left(z_{x,1} z_{y,1} + \cdots + z_{x,n} z_{y,n}\right)$$

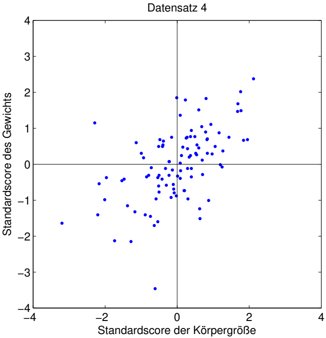
- Es gilt immer:

$$-1 \leq r_{xy} \leq 1$$

Korrelation

- r_{xy} ist positiv, wenn viele Produkte positiv sind, d.h. viele Paare von Standardscores das gleiche Vorzeichen haben.
- Das ist der Fall, wenn die Paare der Standardscores vorwiegend im 1. oder 3. Quadranten liegen.
- x und y heißen dann **positiv korreliert**.
- r_{xy} ist negativ, wenn viele Produkte negativ sind, d.h. viele Paare von Standardscores verschiedenes Vorzeichen haben.
- Das ist der Fall, wenn die Paare der Standardscores vorwiegend im 2. oder 4. Quadranten liegen.
- x und y heißen dann **negativ korreliert**.

- Streudiagramm der Standardscores von Datensatz 4:



- Offensichtlich sind x und y positiv korreliert, da die meisten Punkte im 1. und 3. Quadranten liegen.
- $r_{xy} = 0.5562$

Korrelation

- Eine positive Korrelation muss nicht unbedingt einen kausalen Zusammenhang bedeuten.
- Die positive Korrelation kann auch durch eine gemeinsame Ursache oder einen parallel laufenden Trend verursacht sein.

Beispiel

Zwischen der Kinderzahl und der Anzahl der Störche in Österreich in den letzten 30 Jahren besteht eine positive Korrelation. Warum?

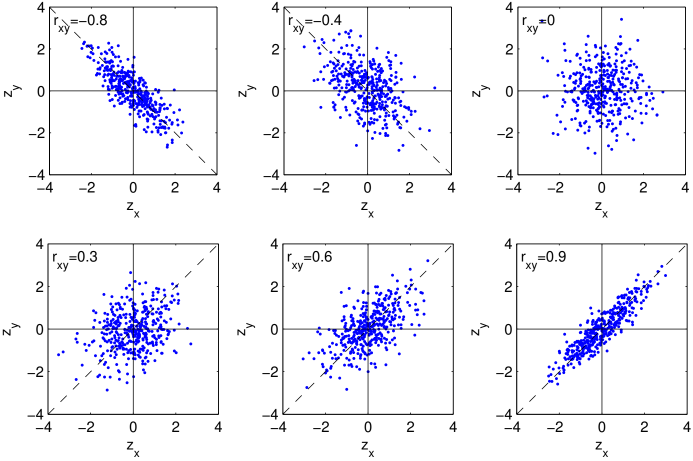
Beispiel

Zwischen dem Butterpreis und dem Brotpreis der letzten 20 Jahre besteht eine positive Korrelation. Warum?

Beispiel

Zwischen der Anzahl an Piraten auf den Weltmeeren und dem globalen CO_2 Ausstoß besteht eine Antikorrelation. Warum?

Korrelation

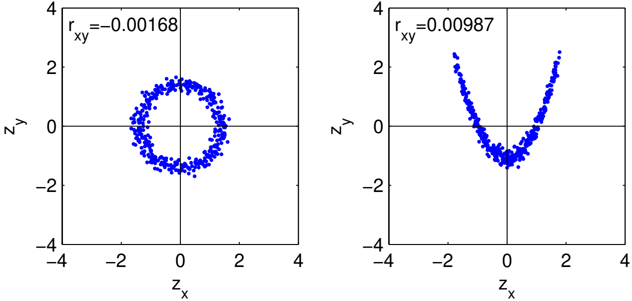


Standardscores mit verschiedenen Korrelationskoeffizienten

Korrelation

- Der Korrelationskoeffizient misst die **Korrelation** der Daten.
- Die Korrelation gibt die Bindung der Punktwolke an eine steigende oder fallende **Gerade**, die **Hauptachse** an.
- Die Korrelation gibt also das Ausmaß der **linearen** Koppelung an.
- Besteht zwischen x und y ein starker, aber **nichtlinearer** Zusammenhang, kann die Korrelation trotzdem sehr klein sein.

Korrelation



Nichtlinearer Zusammenhang zwischen x und y

Korrelation

- Der Korrelationskoeffizient kann auch direkt aus der Stichprobe berechnet werden:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Definition (Kovarianz der Daten)

Die Größe

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

heißt die **Kovarianz der Daten**.

Teil II

Wahrscheinlichkeitsrechnung

Übersicht Teil 2

Abschnitt 4: Einleitung

Einleitung

- Das konkrete Ergebnis eines Experiments kann im Allgemeinen nicht genau vorausgesagt werden; die möglichen Ergebnisse sind jedoch bekannt.
- Mehrere Gründe:
  - Die beobachteten Objekte sind eine zufällige Auswahl (Stichprobe) aus einer größeren Grundgesamtheit.
  - Der beobachtete Prozess ist prinzipiell indeterministisch (Quantenmechanik).
  - Der beobachtete Prozess ist praktisch indeterministisch: mangelnde Kenntnis des Anfangszustandes (Roulette), chaotisches System (Hydrodynamik, Psychologie)
- Zusätzlich können Messfehler dem Ergebnis einen stochastischen (zufälligen) Charakter geben.

Einleitung

- Ziel der Wahrscheinlichkeitsrechnung ist es, Mengen von Ergebnissen, sogenannten Ereignissen, **Wahrscheinlichkeiten** zuzuweisen.
- Zwei Interpretationen der Wahrscheinlichkeit möglich.

Einleitung

Häufigkeitsinterpretation

- Die Wahrscheinlichkeit eines Ereignisses ist seine Häufigkeit, wenn das Experiment sehr oft unter den gleichen Bedingungen wiederholt wird.
- Die darauf basierende Statistik wird „frequentistisch“ genannt.

Beispiel

Die Wahrscheinlichkeit des Ausgangs „1“ beim Würfeln ist der Grenzwert der Häufigkeit für eine große Zahl von Würfen.

Einleitung

Subjektive Interpretation

- Die Wahrscheinlichkeit eines Ausgangs ist eine Aussage über den Glauben der Person, die die Wahrscheinlichkeit angibt.
- Die darauf basierende Statistik wird „bayesianisch“ genannt.

Beispiel

„Die Wahrscheinlichkeit, dass es morgen regnet, ist 40 Prozent“ ist ein Aussage über den Glauben der Person, die diese Aussage tätigt.

Einleitung

- In der Praxis ist der Übergang zwischen den beiden Ansätzen oft fließend.
- In vielen Fällen sind die Resultate identisch, nur die Interpretation ist verschieden.
- Der bayesianische Ansatz ist umfassender und flexibler.
- Der frequentistische Ansatz ist oft einfacher, aber beschränkter.

Abschnitt 5: Ereignisse

Unterabschnitt: Der Ergebnisraum

Der Ergebnisraum

- Grundlegend für die Statistik ist der Begriff des (zufälligen) **Ereignisses**.
- Anschaulich: der **Ausgang eines Experiments**, dessen Ergebnis **nicht genau vorausgesagt** werden kann.
- Die Menge Ω aller möglichen Ergebnisse heißt **Ergebnisraum** oder **Stichprobenraum**.
- Der Ergebnisraum Ω kann diskret (endlich oder abzählbar unendlich) oder stetig (überabzählbar unendlich) sein.

Der Ergebnisraum

Beispiel

- Beim Roulette gibt es 37 mögliche Ergebnisse. Der Ergebnisraum ist diskret und endlich.
- Wird eine radioaktive Quelle beobachtet, ist die Anzahl der Zerfälle pro Sekunde im Prinzip unbeschränkt. Der Ergebnisraum ist diskret und abzählbar unendlich.
- Die Wartezeit zwischen zwei Zerfällen kann jeden beliebigen Wert annehmen. Der Ergebnisraum ist stetig und überabzählbar unendlich.

Unterabschnitt: Die Ereignisalgebra

Die Ereignisalgebra

Definition (Ereignis)

Ein **Ereignis** E ist eine Teilmenge des Ergebnisraums Ω. Ein Ereignis E **tritt ein**, wenn E das Ergebnis $\omega \in \Omega$ des Experiments enthält.

Beispiel

Der Wurf mit einem Würfel hat den Ergebnisraum $\Omega = \{1, 2, 3, 4, 5, 6\}$. Das Ereignis G (gerade Zahl) ist die Teilmenge

$$G = \{2, 4, 6\}$$

G tritt ein, wenn eine gerade Zahl geworfen wird.

Die Ereignisalgebra

Definition (Ereignisalgebra)

Die Menge aller Ereignisse des Ergebnisraums Ω heißt die Ereignisalgebra $\Sigma(\Omega)$.

- Im endlichen oder abzählbar unendlichen Fall kann **jede** Teilmenge als Ereignis betrachtet werden. Die Ereignisalgebra heißt **diskret**.
- Im überabzählbar unendlichen Fall müssen gewisse pathologische (nicht messbare) Teilmengen ausgeschlossen werden. Die Ereignisalgebra heißt **kontinuierlich** oder **stetig**.
- Zwei Ereignisse $A \in \Sigma$ und $B \in \Sigma$ können so wie Aussagen logisch **verknüpft werden**.

Die Ereignisalgebra

Verknüpfung von Ereignissen

Disjunktion

Symbol	Name	Bedeutung
$A \cup B$	Disjunktion	A oder B (oder beide)

Konjunktion

Symbol	Name	Bedeutung
$A \cap B$	Konjunktion	A und B (sowohl A als auch B)

Negation

Symbol	Name	Bedeutung
A'	Negation	nicht A (das Gegenteil von A)

Die Ereignisalgebra

- Mit diesen Verknüpfungen ist Σ ist eine **Boole'sche Algebra**: distributiver komplementärer Verband mit Null- und Einselement.
- Das Nullelement $\mathbf{0} = \emptyset$ ist das **unmögliche Ereignis**.
- Das Einselement $\mathbf{1} = \Omega$ ist das **sichere Ereignis**.
- Ein Ereignis, das nur aus einem möglichen Ergebnis besteht, heißt ein **Elementarereignis**.

Die Ereignisalgebra

- Ist Ω (abzählbar oder überabzählbar) unendlich, verlangt man, dass auch abzählbar viele Vereinigungen und Durchschnitte von Ereignissen gebildet werden können.
- Die Ergeignisalgebra ist dann eine sogenannte σ-Algebra.
- Ist im stetigen Fall $\Omega = \mathbb{R}$, so ist die Ereignisalgebra $\Sigma(\Omega)$ die kleinste σ-Algebra, die alle Intervalle enthält.

Unterabschnitt: Wiederholte Experimente

Wiederholte Experimente

- Der Wurf mit einem Würfel hat den Ergebnisraum

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Die Ereignisalgebra $\Sigma(\Omega)$ hat folglich sechs Elementarereignisse:

$$e_1 = \{1\}, e_2 = \{2\}, e_3 = \{3\}, e_4 = \{4\}, e_5 = \{5\}, e_6 = \{6\}$$

und insgesamt $2^6 = 64$ Ereignisse (Teilmengen von Ω).

- Der Ergebnisraum des zweimaligen Würfelns ist das kartesische Produkt $\Omega \times \Omega$:

$$\Omega \times \Omega = \{(i, j) | i, j = 1, \ldots, 6\}$$

Das geordnete Paar (i, j) bedeutet: i beim ersten Wurf, j beim zweiten Wurf. Die Ereignisalgebra $\Sigma(\Omega \times \Omega)$ hat folglich 36 Elementarereignisse e_{ij}:

$$e_{11} = \{(1, 1)\}, \ldots, e_{36} = \{(6, 6)\}$$

Wiederholte Experimente

- Analog ist beim n-maligen Würfeln der Ergebnisraum das n-fache kartesische Produkt $\Omega \times \Omega \times \ldots \times \Omega$.

Beispiel (Ereignisalgebra des Doppelwurfs)

Beispiele von Ereignissen in der Ereignisalgebra des Doppelwurfs sind:

6 beim ersten Wurf:	$\{(6,1),(6,2),\ldots,(6,6)\}$
6 beim zweiten Wurf:	$\{(1,6),(2,6),\ldots,(6,6)\}$
Beide Würfe gleich:	$\{(1,1),(2,2),\ldots,(6,6)\}$
Summe der Würfe gleich 7:	$\{(1,6),(2,5),\ldots,(6,1)\}$

Wiederholte Experimente

Beispiel (Wiederholter Alternativversuch)

Ein Experiment, das nur zwei mögliche Ergebnisse hat, heißt ein **Bernoulli-Experiment** oder **Alternativversuch**. Es gibt zwei Ausgänge, 1 ("Erfolg") und 0 ("Misserfolg").

Wird ein Alternativversuch n-mal durchgeführt, ergibt sich ein Ergebnisraum mit 2^n möglichen Ergebnissen, nämlich allen Folgen der Form $(i_1, \ldots, i_n)$ mit $i_j = 0$ oder 1.

In der Regel interessiert aber nur die **Häufigkeit** des Eintretens von 1 (oder 0). Dann gibt es nur mehr $n + 1$ Ergebnisse: 1 tritt $0, 1, 2, \ldots$ oder n-mal ein. Bezeichnet das Ereignis E_1 das einmalige Eintreten von 1, so ist E_1 die Vereinigung mehrerer Elementarereignisse der ursprünglichen Ereignisalgebra:

$$E_1 = \{(e_1, e_0, \ldots, e_0), (e_0, e_1, e_0, \ldots, e_0), \ldots, (e_0, \ldots, e_0, e_1)\}$$

Ein Beispiel ist das n-malige Werfen einer Münze.

Abschnitt 6: Wahrscheinlichkeit

Unterabschnitt: Wahrscheinlichkeitsmaße

Wahrscheinlichkeitsmaße

Definition (Wahrscheinlichkeitsmaß)

Es sei Σ eine Ereignisalgebra, A und B Ereignisse in Σ, und W eine Abbildung von Σ in $\mathbb{R}$. W heißt ein **Wahrscheinlichkeitsmaß**, wenn gilt:

1. Positivität: $W(A) \geq 0, \ \forall A \in \Sigma$
2. Additivität: $A \cap B = \mathbf{0} \implies W(A \cup B) = W(A) + W(B)$
3. Normierung: $W(\mathbf{1}) = 1$

Wahrscheinlichkeitsmaße

Definition (Wahrscheinlichkeitsraum)

Ist Σ eine σ-Algebra, was für unendliche Ereignisräume vorausgesetzt werden muss, verlangt man für abzählbares J:

4. σ-Additivität: $A_i \in \Sigma,\ i \in J;\ A_i \cap A_j = \mathbf{0},\ i \neq j \implies W(\bigcup_{i \in J} A_i) = \sum_{i \in J} W(A_i)$

Σ heißt dann normiert, und (Σ, W) ein **Wahrscheinlichkeitsraum**. W wird auch als **Wahrscheinlichkeitsverteilung** bezeichnet.

Wahrscheinlichkeitsmaße

Rechengesetze für Wahrscheinlichkeit

Ist (Σ, W) ein Wahrscheinlichkeitsraum, so gilt:

- $W(\mathbf{0}) = 0$
- $W(\mathbf{1}) = 1$
- $0 \leq W(A) \leq 1, \ \forall A \in \Sigma$
- $W(A') = 1 - W(A), \ \forall A \in \Sigma$
- $A \subseteq B \implies W(A) \leq W(B), \ \forall A, B \in \Sigma$
- $W(A \cup B) = W(A) + W(B) - W(A \cap B), \ \forall A, B \in \Sigma$
- Hat Σ höchstens abzählbar viele Elementarereignisse $\{e_i \mid i \in I\}$, so ist $\sum_{i \in I} W(e_i) = 1$.

Wahrscheinlichkeitsmaße

- In einer diskreten Ereignisalgebra ist die Wahrscheinlichkeit eines Ereignisses gleich der Summe der Wahrscheinlichkeiten der Elementarereignisse, deren Vereinigung es ist.
- Daher ist in diesem Fall ein Wahrscheinlichkeitsmaß durch die Werte, die es den Elementarereignissen zuordnet, **eindeutig bestimmt**.
- Andererseits kann jede positive Funktion, die auf der Menge der Elementarereignisse definiert ist und die Normierungsbedingung erfüllt, eindeutig zu einem Wahrscheinlichkeitsmaß fortgesetzt werden.
- Man kann also auf einer diskreten Ereignisalgebra Σ unendlich viele Verteilungen definieren.

Wahrscheinlichkeitsmaße

- In einer kontinuierlichen Ereignisalgebra ist die Wahrscheinlichkeit jedes Elementarereignisses gleich 0.
- Die Wahrscheinlichkeit eines Ereignisses kann daher nicht mehr durch Summation ermittelt werden.
- Statt dessen wird eine **Dichtefunktion** $f(x)$ angegeben, die jedem Elementarereignis x einen nichtnegativen Wert $f(x)$ zuordnet.
- Die Wahrscheinlichkeit eines Ereignisses A wird durch **Integration** über die Dichte ermittelt:

$$W(A) = \int_A f(x)\,\mathrm{d}x$$

- Die Dichtefunktion ist auf 1 **normiert**:

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$$

Unterabschnitt: Gesetz der großen Zahlen

Gesetz der großen Zahlen

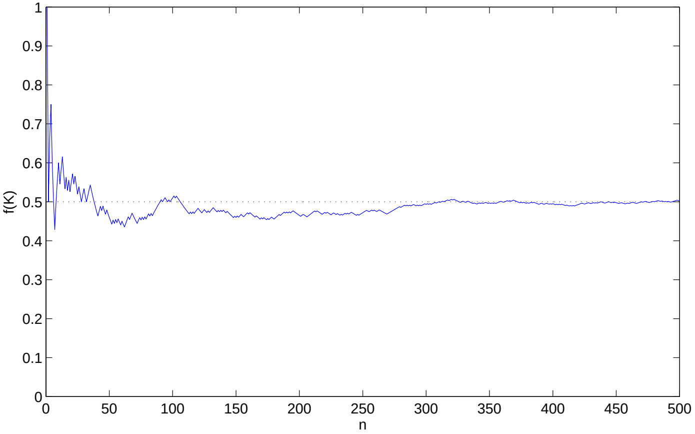
- Wir betrachten ein einfaches Zufallsexperiment: Münzwurf
- Zwei mögliche Ergebnisse: Kopf (K), Zahl (Z)
- Annahme: Münze symmetrisch, K und Z gleichwahrscheinlich
- Experiment wird n-mal wiederholt

n	$h_n(K)$	$f_n(K)$	$\lvert f_n(K) - 0.5\rvert$
10	6	0.6	0.1
100	51	0.51	0.01
500	252	0.504	0.004
1000	488	0.488	0.012
5000	2533	0.5066	0.0066

Häufigkeitstabelle

MATLAB: `make_coin`

Gesetz der großen Zahlen



Entwicklung der relativen Häufigkeit von K

Gesetz der großen Zahlen

- Die relative Häufigkeit des Ereignisses K scheint gegen den Grenzwert 0.5 zu streben.
- Dieser Grenzwert wird als die **Wahrscheinlichkeit** $W(K)$ bezeichnet.

Empirisches Gesetz der großen Zahlen

$$\lim_{n \to \infty} f_n(K) = W(K)$$

- Das mathematische Problem dieser Definition liegt darin, dass die Existenz des Grenzwerts von vornherein nicht einzusehen ist und im klassisch analytischen Sinn tatsächlich nicht gegeben sein muss, sondern nur in einem weiteren, statistischen Sinn.

Abschnitt 7: Bedingte Wahrscheinlichkeit

Unterabschnitt: Kopplung und bedingte Wahrscheinlichkeit

Kopplung und bedingte Wahrscheinlichkeit

- Wir betrachten jetzt zwei Ereignisse A und B, die bei einem Experiment eintreten können.
- **Frage**: Besteht ein Zusammenhang zwischen den Ereignissen?
- Ein solcher Zusammenhang wird **Koppelung** genannt.
- **Positive Koppelung**: Je öfter A eintritt, desto öfter tritt tendenziell auch B ein.
- **Negative Koppelung**: Je öfter A eintritt, desto seltener tritt tendenziell auch B ein.
- Quantifizierung von „oft“ und „selten“ erfolgt durch Häufigkeitstabelle.

Kopplung und bedingte Wahrscheinlichkeit

- Die Häufigkeit des Eintretens von A und B kann in einer **Vierfeldertafel** oder **Kontingenztafel** zusammengefasst werden.
- **Beispiel**:

  A=*"Eine untersuchte Person ist weiblich"*
  B=*"Eine untersuchte Person hat Diabetes"*

- Vierfeldertafel für 1000 Personen:

	B	B'	
A	19	526	545
A'	26	429	455
	45	955	1000

Kopplung und bedingte Wahrscheinlichkeit

- **Gewöhnliche relative Häufigkeiten** werden auf den Umfang n des gesamten Datensatzes bezogen:

$$f(A \cap B) = \frac{h(A \cap B)}{n}$$

- **Bedingte relative Häufigkeiten** werden auf das Eintreten des anderen Merkmals bezogen:

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

- $f(A|B)$ heißt die bedingte relative Häufigkeit von A unter der Bedingung B.

Kopplung und bedingte Wahrscheinlichkeit

- Die Vierfeldertafel U gibt folgende bedingte relative Häufigkeiten:

$$f(A|B) = \frac{19}{45} = 0.422, \quad f(A|B') = \frac{526}{955} = 0.551$$

- Es ist somit zu vermuten, dass die beiden Merkmale gekoppelt sind.
- $f(A|B) > f(A)$ deutet auf eine positive Koppelung, $f(A|B) < f(A)$ auf eine negative Koppelung.

Kopplung und bedingte Wahrscheinlichkeit

- Stammen die Daten aus einem Zufallsexperiment, dann besitzen die Ereigniskombinationen auch Wahrscheinlichkeiten.
- Wahrscheinlichkeitstabelle:

	B	B'	
A	$W(A \cap B)$	$W(A \cap B')$	$W(A)$
A'	$W(A' \cap B)$	$W(A' \cap B')$	$W(A')$
	$W(B)$	$W(B')$	1

- Nach dem empirischen Gesetz der großen Zahl sind diese Wahrscheinlichkeiten die Grenzwerte der entsprechenden relativen Häufigkeiten.

Kopplung und bedingte Wahrscheinlichkeit

- Die bedingten relativen Häufigkeiten konvergieren für $n \to \infty$ gegen einen Grenzwert:

$$f_n(A|B) = \frac{f_n(A \cap B)}{f_n(B)} \to W(A|B) = \frac{W(A \cap B)}{W(B)}$$

Definition (Bedingte Wahrscheinlichkeit)

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

heißt die **bedingte Wahrscheinlichkeit von** A **unter der Bedingung** B, sofern $W(B) \neq 0$.

Kopplung und bedingte Wahrscheinlichkeit

Beispiel (Der symmetrische Würfel)

Ist der Würfel völlig symmetrisch, werden den Elementarereignissen $e_i = \{i\}$ gleiche Wahrscheinlichkeiten zugeordnet:

$$W(e_i) = \frac{1}{6},\ 1 \leq i \leq 6$$

Wir definieren die folgenden Ereignisse:

$$U = \{1, 3, 5\},\ G = \{2, 4, 6\}$$

Dann gilt zum Beispiel

$$W(e_1|U) = \frac{W(e_1 \cap U)}{W(U)} = \frac{W(e_1)}{W(U)} = \frac{1}{3}$$

$$W(e_1|G) = \frac{W(e_1 \cap G)}{W(G)} = \frac{W(\mathbf{0})}{W(G)} = 0$$

Kopplung und bedingte Wahrscheinlichkeit

Beispiel (Fortsetzung)

$$W(U|e_1) = \frac{W(e_1 \cap U)}{W(e_1)} = \frac{W(e_1)}{W(e_1)} = 1$$

$$W(e_1 \cup e_3|U) = \frac{W((e_1 \cup e_3) \cap U)}{W(U)} = \frac{W(e_1 \cup e_3)}{W(U)} = \frac{2}{3}$$

$$W(e_1 \cup e_2|U) = \frac{W((e_1 \cup e_2) \cap U)}{W(U)} = \frac{W(e_1)}{W(U)} = \frac{1}{3}$$

Kopplung und bedingte Wahrscheinlichkeit

- Aus der Definition der bedingten Wahrscheinlichkeit folgt sofort die

Produktformel

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

und die Formel für die

Inverse Wahrscheinlichkeit

$$W(B|A) = \frac{W(A|B)W(B)}{W(A)}$$

- Beide Formeln gelten auch für relative Häufigkeiten!

Unterabschnitt: Satz von Bayes

Satz von Bayes

Definition (Zerlegung)

Die Ereignisse $B_1, B_2, \ldots, B_m$ bilden eine **Zerlegung** des Ergebnisraums Ω, wenn gilt:

1. Unvereinbarkeit: $B_i \cap B_j = \emptyset, i \neq j$
2. Vollständigkeit: $B_1 \cup B_2 \cup \ldots \cup B_m = \Omega$

Satz

Bilden die Ereignisse $B_1, B_2, \ldots, B_m$ eine Zerlegung des Ergebnisraums Ω, dann gilt:

$$W(B_1) + W(B_2) + \ldots + W(B_m) = W(\Omega) = \mathbf{1}$$

Satz von Bayes

- Es sei $B_1, \ldots, B_m$ eine Zerlegung. Dann gilt:

Totale Wahrscheinlichkeit

$$W(A) = W(A|B_1)W(B_1) + \ldots + W(A|B_m)W(B_m)$$

Beispiel

Ein Betrieb erzeugt Widerstände mit 10 kΩ (35% der Produktion), mit 22 kΩ (45%) und mit 47 kΩ (20%). Nach zwei Jahren im Dauerbetrieb sind noch 98% der 10 kΩ-Widerstände funktionsfähig, 96% der 22 kΩ-Widerstände, und 92% der 47 kΩ-Widerstände. Welcher Anteil an allen Widerständen ist nach zwei Jahren noch funktionsfähig?

Satz von Bayes

- Es sei $B_1, \ldots, B_m$ eine Zerlegung. Dann gilt:

Satz von Bayes

$$W(B_i|A) = \frac{W(A|B_i)W(B_i)}{W(A)}$$
$$= \frac{W(A|B_i)W(B_i)}{W(A|B_1)W(B_1) + \ldots + W(A|B_m)W(B_m)}$$

- $W(B_i)$ wird die **a-priori** Wahrscheinlichkeit von B_i genannt, $W(B_i|A)$ die **a-posteriori** Wahrscheinlichkeit.

Satz von Bayes

Beispiel

Ein Betrieb kauft Bauteile von zwei Anbietern, wobei der Anteil des ersten 65% beträgt. Erfahrungsgemäß ist der Ausschussanteil bei Anbieter 1 gleich 3% und bei Anbieter 2 gleich 4%.

1. Wie groß ist der totale Ausschussanteil?
2. Wie groß ist die Wahrscheinlichkeit, dass ein einwandfreier Bauteil von Anbieter 2 kommt?
3. Wie groß ist die Wahrscheinlichkeit, dass ein mangelhafter Bauteil von Anbieter 1 kommt?

Satz von Bayes

Beispiel

Ein Bauteil wird von vier Firmen geliefert, und zwar kommen 20% von Firma 1, 30% von Firma 2, 35% von Firma 3, und 15% von Firma 4. Die Wahrscheinlichkeit, dass der Bauteil im Testbetreib innerhalb von 24 Stunden ausfällt, ist 0.02 für Firma 1, 0.015 für Firma 2, 0.025 für Firma 3, und 0.02 für Firma 4. Ein Bauteil fällt im Testbetrieb nach 16 Stunden aus. Die Wahrscheinlichkeit, dass er von Firma i kommt, ist mittel des Satzes von Bayes zu berechnen.

Unterabschnitt: Unabhängigkeit

Unabhängigkeit

- Zwei Ereignisse sind **positiv gekoppelt**, wenn

$$W(A|B) > W(A) \quad \text{oder} \quad W(A \cap B) > W(A)W(B)$$

- Zwei Ereignisse sind **negativ gekoppelt**, wenn

$$W(A|B) < W(A) \quad \text{oder} \quad W(A \cap B) < W(A)W(B)$$

- Liegt weder positive noch negative Kopppelung vor, sind A und B **unabhängig**.

Unabhängigkeit

Definition (Unabhängigkeit)

Zwei Ereignisse A und B heißen **stochastisch unabhängig**, wenn

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

Die Ereignisse $A_1, A_2, \ldots, A_n$ heißen unabhängig, wenn gilt:

$$W(A_1 \cap \ldots \cap A_n) = W(A_1) \cdot \ldots \cdot W(A_n)$$

Dazu genügt nicht, dass je zwei Ereignisse A_i und A_j paarweise unabhängig sind!

Unabhängigkeit

Beispiel

Wir betrachten den zweimaligen Wurf einer Münze (Kopf/Zahl). Die möglichen Ausgänge sind $\Omega = \{KK, KZ, ZK, ZZ\}$. Ferner definieren wir die Ereignisse:

$$
\begin{aligned}
E_1 &= \{KK, KZ\} \ldots \text{Kopf beim ersten Wurf} \\
E_2 &= \{KK, ZK\} \ldots \text{Kopf beim zweiten Wurf} \\
E_3 &= \{KK, ZZ\} \ldots \text{Gerade Zahl von Köpfen}
\end{aligned}
$$

Dann gilt für alle $i \neq j$

$$W(E_i \cap E_j) = \frac{1}{4} = W(E_i) \cdot W(E_j)$$

aber

$$W(E_1 \cap E_2 \cap E_3) = \frac{1}{4} \neq \frac{1}{8} = W(E_1) \cdot W(E_2) \cdot W(E_3)$$

Unabhängigkeit

- Sind A und B unabhängig, gilt $W(A|B) = W(A)$ und $W(B|A) = W(B)$.
- Die Vierfeldertafel für zwei unabhängige Ereignisse:

	B	B'	
A	$W(A)W(B)$	$W(A)W(B')$	$W(A)$
A'	$W(A')W(B)$	$W(A')W(B')$	$W(A')$
	$W(B)$	$W(B')$	1

Unabhängigkeit

- Die Koppelung kann durch die **Vierfelderkorrelation** gemessen werden:

Vierfelderkorrelation

$$\rho(A,B) = \frac{W(A \cap B) - W(A)W(B)}{\sqrt{W(A)W(A')W(B)W(B')}}$$

Eigenschaften der Vierfelderkorrelation

- $-1 \leq \rho(A,B) \leq 1$
- $\rho(A,B) = 0 \iff A$ und B unabhängig
- $\rho(A,B) > 0 \iff A$ und B positiv gekoppelt
- $\rho(A,B) < 0 \iff A$ und B negativ gekoppelt

Unabhängigkeit

- Das Vorzeichen von $\rho(A, B)$ gibt die **Richtung** der Koppelung an.
- Der Betrag von $\rho(A, B)$ gibt die **Stärke** der Koppelung an.
- Speziell gilt:

$$A = B \implies \rho(A, B) = 1$$
$$A = B' \implies \rho(A, B) = -1$$

- Eine bestehende Koppelung ist kein Beweis für einen kausalen Zusammenhang!
- Die Koppelung kann auch durch eine gemeinsame Ursache für beide Ereignisse entstehen.

Unabhängigkeit

- Zwei Ereignisse können als unabhängig postuliert werden, wenn zwischen ihnen keine wie immer geartete Verbindung besteht, da dann das Eintreten des einen Ereignisses die Wahrscheinlichkeit des anderen nicht beeinflussen kann.
- Zwei Elementarereignisse sind niemals unabhängig, da ihre $\cap$-Verbindung stets das unmögliche Ereignis ist.
- Zwei Elementarereignisse sind sogar höchst „abhängig", weil das Eintreten des einen das Eintreten des anderen mit Sicherheit ausschließt.
- Sind E_1 und E_2 zwei unabhängige Ereignisse eines Wahrscheinlichkeitsraumes (Σ, W), so sind auch E_1 und E_2', E_1' und E_2, sowie E_1' und E_2' unabhängig.

Unabhängigkeit

Beispiel (Wurf mit zwei unterscheidbaren Würfeln)

Es gibt 36 Elementarereignisse $e_{ij} = \{(i,j)\},\ 1 \leq i,j \leq 6$. Das Ereignis E_i^1, beim ersten Wurf eine i zu würfeln, setzt sich so zusammen:

$$E_i^1 = e_{i1} \cup e_{i2} \cup \ldots \cup e_{i6} \text{ und analog}$$
$$E_j^2 = e_{1j} \cup e_{2j} \cup \ldots \cup e_{6j}$$

Klarerweise gilt $E_i^1 \cap E_j^2 = e_{ij}$.
Kann man annehmen, dass alle Elementarereignisse **gleichwahrscheinlich** sind, so gilt:

$$W(E_i^1) = \frac{1}{6},\ W(E_j^2) = \frac{1}{6}$$
$$W(E_i^1 \cap E_j^2) = W(e_{ij}) = \frac{1}{36} = W(E_i^1) \cdot W(E_j^2)$$

Unabhängigkeit

Beispiel (Fortsetzung)

In diesem Fall sind also auch die Elementarereignisse des einfachen Wurfes **gleichwahrscheinlich** und die beiden Teilwürfe sind **unabhängig**. Setzt man umgekehrt voraus, dass für beide Teilwürfe die Elementarereignisse gleichwahrscheinlich sind, und dass E_i^1 und E_j^2 für alle i und j unabhängig sind, so sind die e_{ij} gleichwahrscheinlich. Sind die Teilwürfe nicht unabhängig, so sind die e_{ij} **trotz** der Gleichwahrscheinlichkeit der e_i und e_j **nicht** mehr gleichwahrscheinlich. Ein Beispiel dafür ist der „Wurf" mit einem sehr großen Würfel, der jedesmal bloß um 90^o gedreht werden kann. Das Elementarereignis e_{34} ist hier unmöglich und muss daher die Wahrscheinlichkeit 0 zugewiesen bekommen.

Unabhängigkeit

Beispiel (Wiederholung eines Alternativversuchs)

Die Ereignisalgebra hat 2^n Elementarereignisse, nämlich die Folgen der Form $(i_1, \ldots, i_n)$, $i_j = 0$ oder 1. Sind die Wiederholungen unabhängig, und bezeichnet p die Wahrscheinlichkeit des Eintretens von 1, ist die Wahrscheinlichkeit einer Folge

$$W(\{(i_1, \ldots, i_n)\}) = p^{n_1}(1-p)^{n_0}$$

wo n_0 bzw. n_1 die Anzahl des Eintretens von 0 bzw. 1 angibt. Klarerweise gilt $n_0 + n_1 = n$.

Teil III

Zufallsvariable und Verteilungen

Übersicht Teil 3

Abschnitt 8: Zufallsvariable

Unterabschnitt: Grundbegriffe

Grundbegriffe

Definition (Zufallsvariable)

Eine Abbildung X:

$$\omega \in \Omega \mapsto x = X(\omega) \in \mathbb{R}$$

die jedem Element ω des Ergebnisraums Ω eine reelle Zahl zuordnet, heißt eine (eindimensionale oder univariate) **Zufallsvariable**.

- Ist Ω endlich oder abzählbar unendlich, ist jede beliebige Abbildung X zugelassen.
- Ist Ω überabzählbar unendlich, muss X eine **messbare** Abbildung sein.
- Da der Wert einer Zufallsvariablen vom Ausgang des Experiments abhängt, kann man den möglichen Werten Wahrscheinlichkeiten zuschreiben.

Grundbegriffe

- Nimmt die Zufallsvariable X nur endlich oder abzählbar unendlich viele Werte an, heißt sie **diskret**.
- Nimmt die Zufallsvariable X ein Kontinuum von Werten an, heißt sie **kontinuierlich** oder **stetig**.

Beispiel

Die Abbildung, die beim Würfeln dem Elementarereignis e_i die Augenzahl i zuordnet, ist eine diskrete Zufallsvariable. Natürlich wäre auch die Abbildung $e_i :\longrightarrow 7 - i$ eine diskrete Zufallsvariable.

Beispiel

Die Abbildung, die dem Zerfall eines Teilchens die Lebensdauer x zuordnet, ist eine kontinuierliche Zufallsvariable.

Unterabschnitt: Diskrete Zufallsvariable

Diskrete Zufallsvariable

- Diskrete Zufallsvariable sind oft das Resultat von **Zählvorgängen**.
- Diskrete Zufallsvariable kommen in vielen Bereichen vor: z.B. das Zählen von Ereignissen in einem festen Zeitintervall (Poisson-Verteilung) oder das Zählen von Erfolgen in wiederholten Alternativversuchen (Binomialverteilung),
- Im folgenden nehmen wir an, dass die Werte einer diskreten Zufallsvariablen nichtnegative ganze Zahlen sind. Dies ist keine Einschränkung, weil jede abzählbare Menge von reellen Zahlen bijektiv auf (eine Teilmenge von) $\mathbb{N}_0$ abgebildet werden kann.
- Die Ereignisalgebra ist die Potenzmenge (Menge aller Teilmengen) P von $\mathbb{N}_0$.

Diskrete Zufallsvariable

- Ist auf der Ereignisalgebra $\Sigma(\Omega)$ ein Wahrscheinlichkeitsmaß W definiert, so kann man mit Hilfe der Zufallsvariablen X auf der Potenzmenge P von $\mathbb{N}_0$ ebenfalls ein Wahrscheinlichkeitsmaß definieren.

Definition (Verteilung einer diskreten Zufallsvariablen)

Es sei $\Sigma(\Omega)$ eine diskrete Ereignisalgebra. Die diskrete Zufallsvariable $X : \Omega \mapsto \mathbb{N}_0$ induziert ein Wahrscheinlichkeitsmaß auf $\mathbb{N}_0$ mittels

$$W_X(\{k\}) = W(X^{-1}(k)) = W(\{\omega | \boldsymbol{X}(\omega) = k\})$$

W_X wird als die **Verteilung** von X bezeichnet, und zwar als diskrete oder Spektralverteilung.

Diskrete Zufallsvariable

Beispiel

Wir ordnen den geraden Augenzahlen des Würfels die Zahl 0 zu, den ungeraden die Zahl 1:

$$X : \omega \mapsto \quad \mod(\omega, 2)$$

Die Verteilung von X ist dann gegeben durch

$$W_X(0) = W(X^{-1}(0)) = W(\{2,4,6\}) = \frac{1}{2}$$
$$W_X(1) = W(X^{-1}(1)) = W(\{1,3,5\}) = \frac{1}{2}$$

Diskrete Zufallsvariable

Beispiel

Wir ordnen dem Ausgang eines Doppelwurfs die Summe der Augenzahlen zu:

$$X : (i,j) \mapsto i + j$$

Die Werte von X sind die natürlichen Zahlen von 2 bis 12. Die Verteilung von X ist dann gegeben durch

$$W_X(k) = W(X^{-1}(k)) = \sum_{i+j=k} W(\{(i,j)\}) = \begin{cases} \dfrac{k-1}{36}, & k \leq 7 \\ \dfrac{13-k}{36}, & k \geq 7 \end{cases}$$

Diskrete Zufallsvariable

- Die Zahlen $W_X(k)$ können als Funktionswerte einer Spektralfunktion f_X angesehen werden:

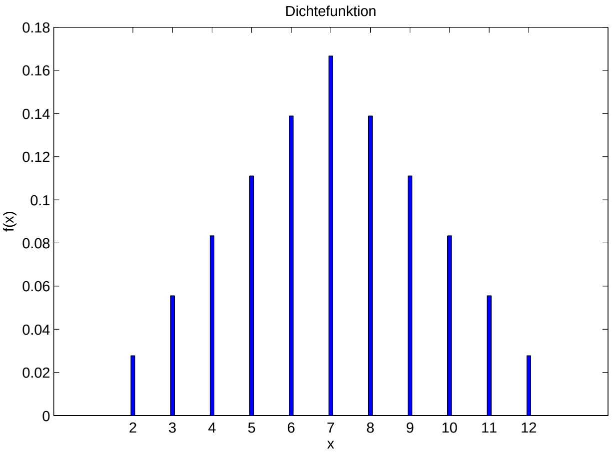
$$f_X(x) = \begin{cases} W_X(k), & \text{wenn } x = k \\ 0, & \text{sonst} \end{cases}$$

Definition (Diskrete Dichtefunktion)

Die Funktion $f_X(x)$ wird als **Wahrscheinlichkeitsdichtefunktion** oder kurz Dichte der Zufallsvariablen X bezeichnet.

Diskrete Zufallsvariable

- Die Dichte der Zufallsvariablen $X = i + j$:



Diskrete Zufallsvariable

- Die Wahrscheinlichkeit $W_X(E)$ eines Ereignisses E lässt sich bequem mit Hilfe der Dichte von X berechnen:

$$W_X(E) = \sum_{k \in E} f_X(k)$$

Definition (Diskrete Verteilungsfunktion)

Ist X eine diskrete Zufallsvariable, so ist die **Verteilungsfunktion** F_X von X definiert durch:

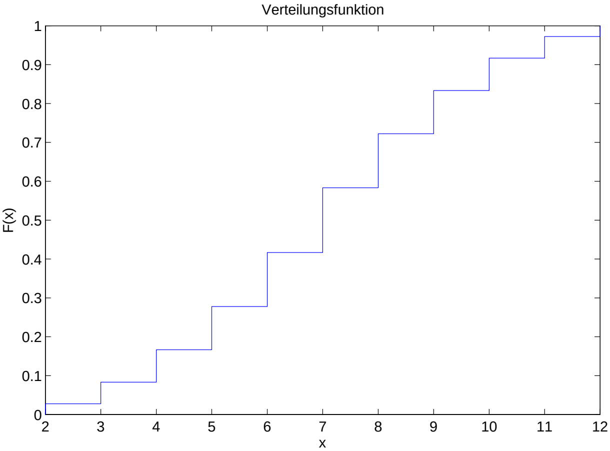
$$F_X(x) = W(X \leq x)$$

Es gilt offenbar:

$$F_X(x) = \sum_{k \leq x} f_X(k) = \sum_{k \leq x} W_X(\{k\})$$

Diskrete Zufallsvariable

- Die Verteilungsfunktion der Zufallsvariablen $X = i + j$:



Diskrete Zufallsvariable

Eigenschaften einer diskreten Verteilungsfunktion F

- F hat eine Sprungstelle in allen Punkten des Wertebereichs
- Die Sprunghöhe im Punkt k ist gleich $f_X(k)$
- $0 \leq F(x) \leq 1, \; \forall x \in \mathbb{R}$
- $x \leq y \implies F(x) \leq F(y), \; \forall x, y \in \mathbb{R}$
- $\lim_{x \to -\infty} F(x) = 0; \; \lim_{x \to \infty} F(x) = 1$
- Die Wahrscheinlichkeit, dass r in das Intervall $(a, b]$ fällt, ist $F(b) - F(a)$:

$$W(r \leq a) + W(a < r \leq b) = W(r \leq b) \implies$$
$$W(a < r \leq b) = F(b) - F(a)$$

Diskrete Zufallsvariable

Wichtige diskrete Verteilungsfamilien

- Poisson-Verteilung $\mathrm{Po}(\lambda)$
- Bernoulli- oder Alternativverteilung $\mathrm{Al}(p)$
- Binomialverteilung $\mathrm{Bi}(n, p)$
- Hypergeometrische Verteilung $\mathrm{Hy}(N, M, n)$

Unterabschnitt: Stetige Zufallsvariable

Stetige Zufallsvariable

- Bisher wurden nur solche Zufallsvariable behandelt, die auf diskreten Ereignisalgebren definiert waren.
- Diese Beschränkung soll nun fallengelassen werden, d.h. es werden jetzt überabzählbar viele Elementarereignisse zugelassen. Das ist notwendig, wenn nicht nur Zählvorgänge, sondern beliebige Messvorgänge betrachtet werden.
- Eine Funktion X, die auf einer solchen überabzählbaren Menge von Elementarereignissen definiert ist, kann beliebige reelle Werte annehmen.

Stetige Zufallsvariable

Definition (Stetige Verteilungsfunktion)

Es sei (Σ, W) ein Wahrscheinlichkeitsraum über einem überabzählbaren Ergebnisraum Ω. X sei eine Zufallsvariable, also eine (messbare) Funktion von Ω in $\mathbb{R}$. Die Funktion F_X, definiert durch:

$$F_X(x) = W(X \leq x)$$

heißt die **Verteilungsfunktion** von X. Die Wahrscheinlichkeit, dass X in ein Intervall $(x, x + \Delta x]$ fällt, ist dann:

$$W(x < X \leq x + \Delta x) = F_X(x + \Delta x) - F_X(x) = \Delta F_X.$$

Stetige Zufallsvariable

Eigenschaften einer stetigen Verteilungsfunktion

- $0 \leq F(x) \leq 1, \ \forall x \in \mathbb{R}$
- $x_1 \leq x_2 \implies F(x_1) \leq F(x_2), \ \forall x_1, x_2 \in \mathbb{R}$
- $\lim_{x \to -\infty} F(x) = 0; \ \lim_{x \to \infty} F(x) = 1$

Definition (Quantil)

Es sei $F_X(x)$ eine stetige Verteilungsfunktion. Der Wert x_α, für den

$$F_X(x_\alpha) = \alpha, \quad 0 < \alpha < 1$$

gilt, heißt das α-**Quantil** der Verteilung von X. Die Funktion

$$x = Q_X(\alpha) = F_X^{-1}(\alpha), \quad 0 < \alpha < 1$$

heißt die **Quantilsfunktion** der Verteilung von X.

Stetige Zufallsvariable

Definition (Quartil)

Die Quantile zu den Werten $\alpha = 0.25, 0.5, 0.75$ heißen **Quartile**. Das Quantil zum Wert $\alpha = 0.5$ heißt **Median** der Verteilung.

- Quantile können auch für diskrete Verteilungen definiert werden, jedoch sind sie dann nicht immer eindeutig.

Stetige Zufallsvariable

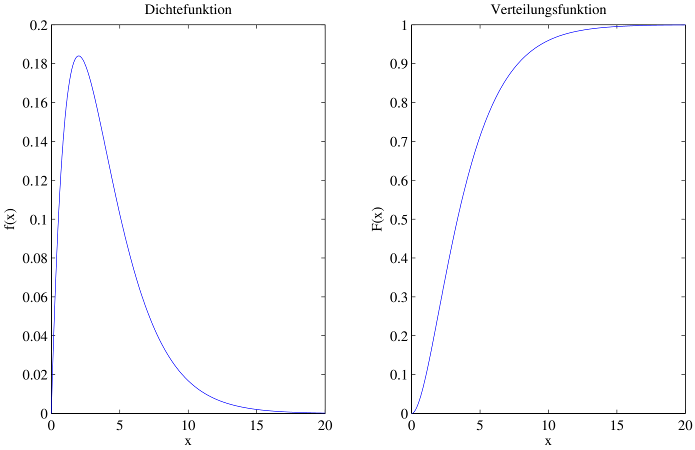
Definition (Stetige Dichtefunktion)

Ist F_X differenzierbar, heißt X eine **stetige Zufallsvariable**. Für die Verteilung von X gilt nach dem Hauptsatz der Integralrechnung:

$$W_X(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\,\mathrm{d}x$$

wobei $f_X(x) = F'_X(x)$ ist. Die Ableitung der Verteilungsfunktion, die Funktion f_X, wird als **Wahrscheinlichkeitsdichtefunktion** oder wieder kurz Dichte von X bezeichnet.

Stetige Zufallsvariable



Stetige Zufallsvariable

- Das Wahrscheinlichkeitsmaß W_X heißt die Verteilung von X. Es ist auf einer Ereignisalgebra Σ definiert, die aus Mengen reeller Zahlen besteht und zumindest alle Intervalle und deren Vereinigungen als Elemente enthält.
- Ähnlich wie bei diskreten Zufallsvariablen lässt sich die Wahrscheinlichkeit W_X einer Menge $M \in \Sigma$ leicht mit Hilfe der Dichte angeben:

$$W_X(M) = \int_M f_X(x)\,\mathrm{d}x$$

- Die Wahrscheinlichkeit eines einzelnen Punktes ist immer gleich 0:

$$W_X(\{x\}) = \int_x^x f_X(x)\,\mathrm{d}x = 0$$

Stetige Zufallsvariable

- Daher ist auch

$$W_X((x_1, x_2]) = W_X((x_1, x_2)) = W_X([x_1, x_2]).$$

- Ganz allgemein erhält man eine Aussage über stetige Zufallsvariable dadurch, dass man in einer Aussage über diskrete Zufallsvariable die Summation durch eine Integration ersetzt.
- Gilt zum Beispiel für eine diskrete Dichte f:

$$\sum_{k \in \mathbb{N}_0} f(k) = 1$$

so gilt für eine stetige Dichte f:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

Stetige Zufallsvariable

Wichtige stetige Verteilungsfamilien

- Normalverteilung $\mathrm{No}(\mu, \sigma^2)$
- Gammaverteilung $\mathrm{Ga}(a, b)$, mit den Spezialfällen
  - Exponentialverteilung $\mathrm{Ex}(\tau)$
  - Chiquadratverteilung $\chi^2(n)$
- Student- oder t-Verteilung $\mathrm{t}(n)$
- Betaverteilung $\mathrm{Be}(a, b)$
- Gleichverteilung $\mathrm{Un}(a, b)$

Unterabschnitt: Unabhängigkeit

Unabhängigkeit

Definition (Unabhängige Zufallsvariable)

Zwei Zufallsvariable X und Y heißen unabhängig, wenn für alle Paare (A, B) von Ereignissen gilt:

$$W(X \in A \cap Y \in B) = W_X(A) \cdot W_Y(B)$$

Definition (Gemeinsame Verteilungsfunktion)

Die gemeinsame Verteilungsfunktion F_{XY} von X und Y ist definiert durch

$$F_{XY}(x, y) = W(X \leq x \cap Y \leq y)$$

Unabhängigkeit

Definition (Gemeinsame Dichte)

Sind X und Y diskrete Zufallsvariable, ist ihre gemeinsame Dichte definiert durch

$$f_{XY}(x,y) = W(X = x \cap Y = y)$$

Sind X und Y stetige Zufallsvariable, ist ihre gemeinsame Dichte definiert durch

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}}{\partial x \, \partial y}$$

Unabhängigkeit

Satz

Zwei Zufallsvariable X und Y sind genau dann unabhängig, wenn ihre gemeinsame Verteilungsfunktion das Produkt der individuellen Verteilungsfunktionen ist:

$$F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$

Satz

Zwei Zufallsvariable X und Y sind genau dann unabhängig, wenn ihre gemeinsame Dichte das Produkt der individuellen Dichtefunktionen ist:

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

Unterabschnitt: Faltung

Faltung

Definition (Faltung)

Es seien X und Y zwei **unabhängige** Zufallsvariable und $Z = X + Y$ ihre Summe. Die Verteilung von Z wird als die **Faltung** der Verteilungen von X und Y bezeichnet.

Beispiel

Ein Experiment misst die Lebensdauer X eines instabilen Teilchens, mit einem Messfehler Y. Sind X und Y unabhängig, ist die Verteilung der Beobachtung $Z = X + Y$ die Faltung der Verteilungen von X bzw. Y.

Faltung

Satz

Es seien X und Y zwei **unabhängige** Zufallsvariable und $Z = X + Y$ ihre Summe. Dann ist die Dichte von Z das **Faltungsprodukt** der Dichten von X und Y.

- Sind X und Y diskret, gilt:

$$f_Z(n) = \sum_k f_X(n-k) f_Y(k)$$

- Sind X und Y stetig, gilt:

$$f_Z(z) = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\,\mathrm{d}y$$

  Es ist zu beachten, dass der effektive Integrationsbereich von z abhängen kann.

Abschnitt 9: Momente

Unterabschnitt: Erwartung

Erwartung

Definition (Erwartung)

Es sei X eine (diskrete oder stetige) Zufallsvariable mit der Dichte $f(x)$. Ferner sei g eine beliebige stetige reelle oder komplexe Funktion. Man definiert $\mathsf{E}_X[g] = \mathsf{E}[g(X)]$ durch:

$$\mathsf{E}[g(X)] = \sum_{k \in \mathbb{N}_0} g(k) f(k) \quad \text{bzw.} \quad \mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) \, \mathrm{d}x$$

$\mathsf{E}_X[g] = \mathsf{E}[g(X)]$ heißt die **Erwartung** von $g(X)$.

Erwartung

Definition (Erwartung einer Zufallsvariablen)

Ist $g(x) = x$, so heißt $\mathsf{E}[g(X)] = \mathsf{E}[X]$ die **Erwartung** oder der **Mittelwert** von X.

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \text{ bzw. } \mathsf{E}[X] = \sum_{k\in\mathbb{N}_0} k\, f(k)$$

Eigenschaften der Erwartung

- $\mathsf{E}[c] = c,\ \ c \in \mathbb{R}$
- $\mathsf{E}[aX + b] = a\mathsf{E}[X] + b$
- $\mathsf{E}[X_1 + X_2] = \mathsf{E}[X_1] + \mathsf{E}[X_2]$
- X_1 und X_2 unabhängig $\implies \mathsf{E}[X_1 X_2] = \mathsf{E}[X_1] \cdot \mathsf{E}[X_2]$

Erwartung

- Die Erwartung ist ein **Lageparameter**.
- Die Erwartung braucht nicht zu existieren. Ein Beispiel ist die Cauchy-Verteilung mit der Dichte

  $$f(x) = \frac{1}{\pi(1+x^2)}, \ x \in \mathbb{R}$$

  In diesem Fall ist der Median ein geeigneter Lageparameter.

Erwartung

Definition (Momente)

Sei X eine Zufallsvariable. Die Erwartung von $g(x) = (x-a)^k$, sofern sie existiert, heißt k**-tes Moment von** X **um** a. Das k-te Moment um 0 wird mit μ'_k bezeichnet. Das k-te Moment um den Erwartungswert $\mathrm{E}[X]$ wird als **zentrales Moment** μ_k bezeichnet.

- Die Erwartung ist das erste Moment um 0.
- Die zentralen Momente $\mu_1, \ldots, \mu_k$ können aus den Momenten um 0 $\mu'_1, \ldots, \mu'_k$ berechnet werden, und umgekehrt.
- Selbst wenn alle Momente einer Verteilung existieren, ist sie dadurch **nicht** eindeutig bestimmt.

Unterabschnitt: Varianz

Varianz

Definition (Varianz)

Das zweite zentrale Moment μ_2 heißt die **Varianz** von X, bezeichnet mit $\text{var}[X]$. Die Wurzel aus der Varianz heißt die **Standardabweichung** von X, bezeichnet mit $\sigma[X]$.

- Die Standardabweichung ist ein Skalenparameter, der die Breite der Verteilung beschreibt.
- Die Standardabweichung hat die gleiche Dimension wie die Werte der Zufallsvariablen.
- Varianz und Standardabweichung sind (wie alle zentralen Momente) invariant gegen Translationen.
- Die Varianz braucht nicht zu existieren. Ein Beispiel ist die Verteilung mit der Dichte

$$f(x) = \frac{1}{(2+x^2)^{3/2}}, \ x \in \mathbb{R}$$

Varianz

Eigenschaften der Varianz

- $\mathsf{var}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$
- $\mathsf{var}[aX + b] = a^2 \mathsf{var}[X]$
- $X_1, \ldots, X_n$ unabhängig:

$$\mathsf{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathsf{var}[X_i]$$

Unterabschnitt: Schiefe

Schiefe

Definition (Schiefe)

Das reduzierte dritte zentrale Moment $\gamma = \mu_3/\sigma^3$ heißt die **Schiefe**.

- Die Schiefe misst die Asymmetrie einer Verteilung. Ist die Schiefe positiv (negativ), heißt die Verteilung rechtsschief (linksschief). Für symmetrische Verteilungen ist sie 0.

Schiefe

Definition (Kurtosis)

Das reduzierte vierte zentrale Moment $\kappa = \mu_4/\sigma^4$ heißt die **Wölbung** oder **Kurtosis**.

- Die Wölbung der Normalverteilung ist 3.
- Verteilungen mit höherer Wölbung haben relativ stärkere Ränder als die Normalverteilung.

Unterabschnitt: Fehlerfortpflanzung

Fehlerfortpflanzung

Affine Transformationen

- Es sei X eine Zufallsvariable mit der Dichte $f(x)$ und $Y = aX + b$.
- Die Dichte $g(y)$ von Y ist dann gleich

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

- Ferner gilt

$$\begin{aligned} \mathsf{E}[Y] &= a\mathsf{E}[X] + b \\ \mathsf{var}[Y] &= a^2 \mathsf{var}[X] \\ \gamma[Y] &= \operatorname{sgn}(a)\gamma[X] \\ \kappa[Y] &= \kappa[X] \end{aligned}$$

Fehlerfortpflanzung

Nichtlineare Transformationen

- Es sei X eine Zufallsvariable mit der Dichte $f(x)$ und $Y = h(X)$.
- Ist $h(x)$ bijektiv, ist die Dichte $g(y)$ von Y gleich

$$g(y) = \left| \frac{\mathrm{d}h^{-1}}{\mathrm{d}y} \right| f\left(h^{-1}(y)\right)$$

- Die Erwartung und die Varianz von $Y = h(X)$ können **näherungsweise** mit Hilfe der Taylorentwicklung von $h(x)$ berechnet werden.

Fehlerfortpflanzung

- Mit der Entwicklungsstelle x_0 gilt in linearer Näherung

$$h(x) \approx h(x_0) + h'(x_0)(x - x_0)$$

- Mit der Wahl $x_0 = \mathsf{E}[X]$ folgt

Satz

$$\mathsf{E}[h(X)] \approx h(\mathsf{E}[X])$$
$$\mathsf{var}[h(X)] \approx h'(\mathsf{E}[X])^2 \cdot \mathsf{var}[X] \quad \text{(Lineare Fehlerfortpflanzung)}$$

- Wird die Entwicklung bis zur 2. Ordnung fortgesetzt, erhält man die verbesserte Näherung

Satz

$$\mathsf{E}[h(X)] \approx h(\mathsf{E}[X]) + \frac{1}{2} h''(\mathsf{E}[X]) \cdot \mathsf{var}[X]$$

Abschnitt 10: Stichprobenfunktionen

Unterabschnitt: Grundbegriffe

Grundbegriffe

- $X_1, \ldots, X_n$ seien unabhängige Zufallsvariable, die alle die gleiche Verteilung F haben.
- Sie bilden dann eine **zufällige Stichprobe** der Verteilung F.
- Eine Zufallsvariable

$$Y = h(X_1, \ldots, X_n)$$

heißt eine **Stichprobenfunktion**.
- In vielen Fällen sind Momente oder die Verteilung von Y zu bestimmen.

Unterabschnitt: Stichprobenmittel

Stichprobenmittel

Definition (Stichprobenmittel)

Das **Stichprobenmittel** $\overline{X}$ der Stichprobe $X_1, \ldots, X_n$ ist definiert durch

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

Momente des Stichprobenmittels

Hat F das Mittel μ und die Varianz σ^2, gilt

- $\mathrm{E}[\overline{X}] = \mu$
- $\mathrm{var}[\overline{X}] = \dfrac{\sigma^2}{n}$

Unterabschnitt: Stichprobenvarianz

Stichprobenvarianz

Definition (Stichprobenvarianz)

Die **Stichprobenvarianz** S^2 der Stichprobe $X_1, \ldots, X_n$ ist definiert durch

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Momente der Stichprobenvarianz

Hat F die Varianz σ^2, gilt

$$\mathsf{E}[S^2] = \sigma^2$$

Existiert das vierte zentrale Moment μ_4 von F, gilt

$$\mathsf{var}[S^2] = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$

Unterabschnitt: Stichprobenmedian

Stichprobenmedian

Definition (Stichprobenmedian)

Der **Stichprobenmedian** $\tilde{X}$ der Stichprobe $X_1, \ldots, X_n$ ist definiert durch

$$\tilde{X} = \begin{cases} X_{((n+1)/2)}, & n \text{ ungerade} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right), & n \text{ gerade} \end{cases}$$

Momente des Stichprobenmedians

Hat F den Median m und die Dichte f, gilt:

- $\lim\limits_{n\to\infty} \mathsf{E}[\tilde{X}] = m$
- $\lim\limits_{n\to\infty} \mathsf{var}[\tilde{X}] = \dfrac{1}{4\,nf^2(m)}, \quad \text{wenn } f(m) > 0$

Teil IV

Punktschätzer

Übersicht Teil 4

Abschnitt 11: Eigenschaften von Punktschätzern

Unterabschnitt: Grundbegriffe

Grundbegriffe

- Ein Punktschätzer ist eine Stichprobenfunktion, die einen möglichst genauen Näherungswert für einen unbekannten Verteilungsparameter ϑ liefern soll:

$$T = g(X_1, \ldots, X_n)$$

- Die Funktion $g(x_1, \ldots, x_n)$ wird die Schätzfunktion genannt.
- Die Konstruktion von sinnvollen Punktschätzern für einen Parameter ϑ ist Aufgabe der Schätztheorie.
- Für einen Parameter ϑ sind viele Punktschätzer möglich. Ein „guter" Punktschätzer sollte jedoch gewisse Anforderungen erfüllen.

Grundbegriffe

Definition (Erwartungstreue)

Ein Punktschätzer T für den Parameter ϑ heißt **erwartungstreu** oder **unverzerrt**, wenn für alle zulässigen Werte von ϑ gilt:

$$\mathsf{E}_\vartheta[T] = \vartheta$$

T heißt **asymptotisch erwartungstreu**, wenn gilt:

$$\lim_{n\to\infty} \mathsf{E}_\vartheta[T] = \vartheta$$

- Ist der unbekannte Parameter gleich ϑ, dann ist die Erwartung des Punktschätzers gleich ϑ.
- Ein erwartungstreuer Punktschätzer hat zwar zufällige Abweichungen vom wahren Wert ϑ, aber keine systematische Verzerrung.

Grundbegriffe

Definition (MSE)

Die **mittlere quadratische Abweichung (mean squared error, MSE)** eines Punktschätzers T für den Parameter ϑ ist definiert durch:

$$\mathsf{MSE}[T] = \mathsf{E}_\vartheta[(T-\vartheta)^2]$$

Definition (MSE-Konsistenz)

Ein Punktschätzer T für den Parameter ϑ heißt **konsistent im quadratischen Mittel (MSE-konsistent)**, wenn gilt:

$$\lim_{n\to\infty} \mathsf{MSE}[T] = 0$$

Grundbegriffe

Definition (MSE-Effizienz)

Ein Punktschätzer T_1 heißt **MSE-effizienter** als der Punktschätzer T_2, wenn für alle zulässigen ϑ gilt:

$$\mathsf{MSE}[T_1] \leq \mathsf{MSE}[T_2]$$

Definition (Effizienz)

Ein erwartungstreuer Punktschätzer T_1 heißt **effizienter** als der erwartungstreue Punktschätzer T_2, wenn für alle zulässigen ϑ gilt:

$$\mathsf{var}[T_1] \leq \mathsf{var}[T_2]$$

Ein erwartungstreuer Punktschätzer T heißt **effizient**, wenn seine Varianz den kleinsten möglichen Wert annimmt.

Grundbegriffe

Definition (Fisher-Information)

Es sei $X_1, \ldots, X_n$ eine Stichprobe mit der gemeinsamen Dichte $g(x_1, \ldots, x_n|\vartheta)$. Die Erwartung

$$I_\vartheta = \mathsf{E}\left[-\frac{\partial^2 \ln g(X_1, \ldots, X_n|\vartheta)}{\partial \vartheta^2}\right]$$

heißt die **Fisher-Information** der Stichprobe.

Satz von Rao und Cramèr

Es sei $X_1, \ldots, X_n$ eine Stichprobe mit der gemeinsamen Dichte $g(x_1, \ldots, x_n|\vartheta)$. Die Varianz eines erwartungstreuen Schätzers T für den Parameter ϑ ist nach unten beschränkt durch:

$$\mathsf{var}[T] \geq 1/I_\vartheta$$

Unterabschnitt: Stichprobenmomente und Stichprobenmedian

Stichprobenmomente und Stichprobenmedian

Satz

- Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Verteilung F mit Erwartung μ. Dann ist das Stichprobenmittel $\overline{X}$ ein erwartungstreuer Punktschätzer von μ.
- Hat F die endliche Varianz σ^2, so ist

$$\mathsf{var}(\overline{X}) = \frac{\sigma^2}{n}$$

- In diesem Fall ist $\overline{X}$ MSE-konsistent.

Stichprobenmomente und Stichprobenmedian

Satz

- Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Verteilung F mit Erwartung μ und Varianz σ^2. Dann ist die Stichprobenvarianz S^2 ein erwartungstreuer Punktschätzer von σ^2.
- Hat F das endliche vierte zentrale Moment μ_4, so ist

$$\mathrm{var}(S^2) = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}$$

- In diesem Fall ist S^2 MSE-konsistent.

Stichprobenmomente und Stichprobenmedian

Satz

- Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der stetigen Verteilung F mit Median m und Dichte f. Dann ist der Stichprobenmedian $\tilde{X}$ ein asymptotisch erwartungstreuer Punktschätzer von m.
- Für symmetrisches F ist $\tilde{X}$ erwartungstreu.
- Der Stichprobenmedian $\tilde{X}$ hat asymptotisch die Varianz

$$\mathrm{var}(\tilde{X}) \approx \frac{1}{4nf(m)^2}$$

- Der Stichprobenmedian ist MSE-konsistent, sofern $f(m) > 0$.

Abschnitt 12: Normalverteilte Daten

Unterabschnitt: Die Gauß- oder Normalverteilung

Die Gauß- oder Normalverteilung

Die Normalverteilung $\mathrm{No}(\mu, \sigma^2)$

- Die **Normalverteilung** ist eine der wichtigsten Verteilungsfamilien in Wissenschaft und Technik. Wir bezeichnen sie mit $\mathrm{No}(\mu, \sigma^2)$.
- Ihre Dichte lautet:

$$f_{\mathrm{No}}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Die Normalverteilung mit $\mu = 0$, $\sigma = 1$ heißt **Standardnormalverteilung**. Deren Dichte wird oft mit $\varphi(x)$ notiert.
- Die Verteilungsfunktion $\Phi(x)$ ist nicht durch elementare Funktionen darstellbar.
- Der Modus M (das Maximum der Dichtefunktion) und der Median m sind bei $x = \mu$.

Die Gauß- oder Normalverteilung

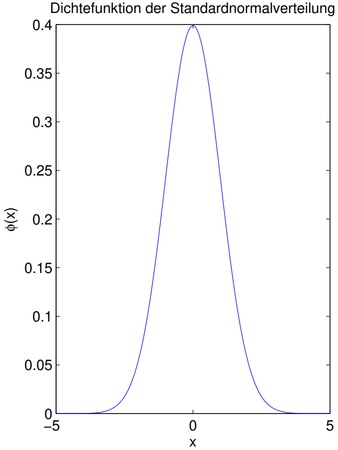
Die Momente und die CF der Normalverteilung

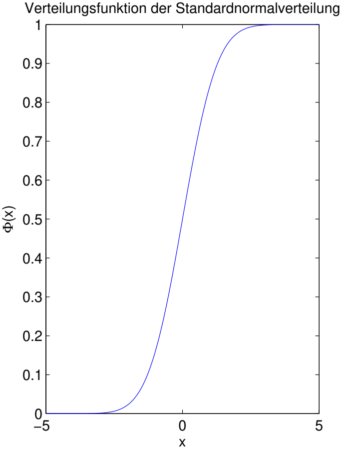
Es sei $X \sim \mathrm{No}(\mu, \sigma^2)$. Dann gilt:

- $\mathsf{E}[X] = \mu$
- $\mathsf{var}[X] = \sigma^2$
- $\gamma[X] = 0$
- $\kappa[X] = 3$
- $\varphi_X(t) = \exp\left(\mathrm{i}\mu t - \sigma^2 t^2/2\right)$

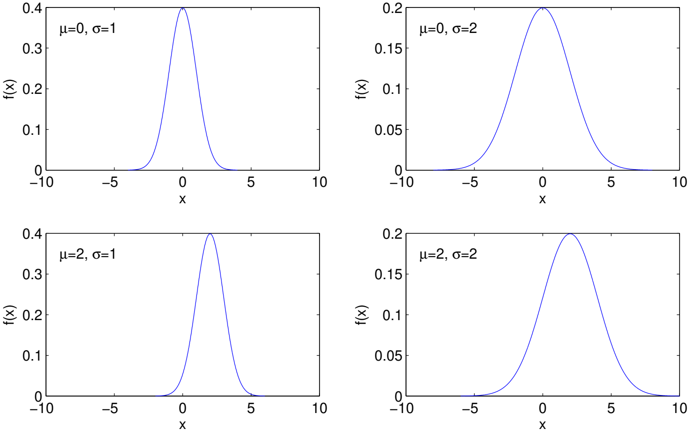
- Das α-Quantil der Standardnormalverteilung wird mit z_α bezeichnet.
- Das α-Quantil von $\mathrm{No}(\mu, \sigma^2)$ ist gleich $\mu + \sigma z_\alpha$.

Die Gauß- oder Normalverteilung



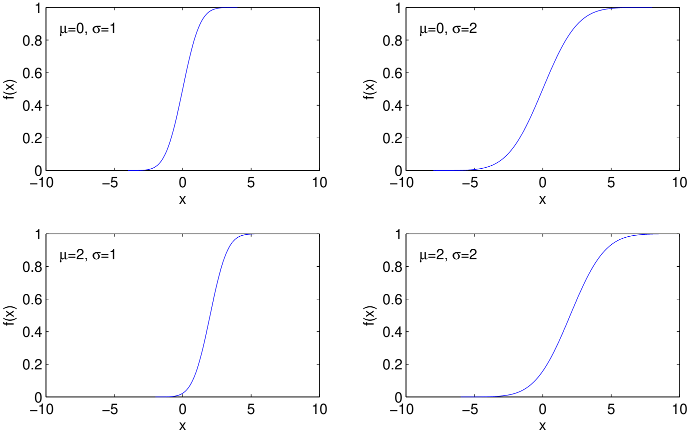


Die Gauß- oder Normalverteilung



Vier Normalverteilungen (Dichtefunktion)

Die Gauß- oder Normalverteilung



Vier Normalverteilungen (Verteilungsfunktion)

Die Gauß- oder Normalverteilung

- Einer der Gründe für die große Bedeutung der Normalverteilung ist der **zentrale Grenzwertsatz**. In seiner einfachsten Formulierung lautet er:

Zentraler Grenzwertsatz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Verteilung F mit Mittel μ und Varianz σ^2. Dann konvergiert die Verteilung von

$$U_n = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

für $n \to \infty$ gegen die Standardnormalverteilung.

- Ist F eine Normalverteilung, so ist U_n klarerweise für alle n exakt standardnormalverteilt.

Die Gauß- oder Normalverteilung

Definition (Standardscore)

Es sei X eine normalverteilte Zufallsvariable mit Mittelwert μ und Varianz σ^2. Dann heißt

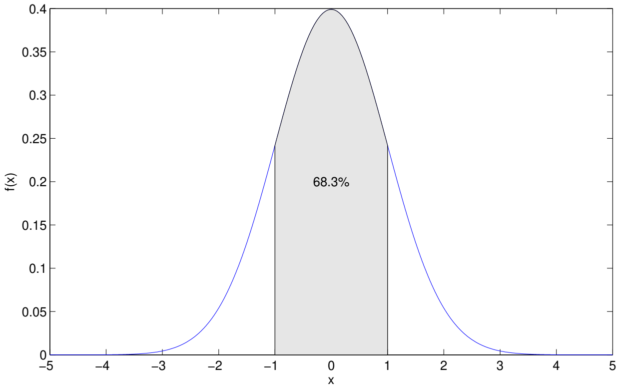
$$Z = \frac{X - \mu}{\sigma}$$

das **Standardscore** von X.

Satz

Ist X normalverteilt, dann ist das Standardscore Z standardnormalverteilt.

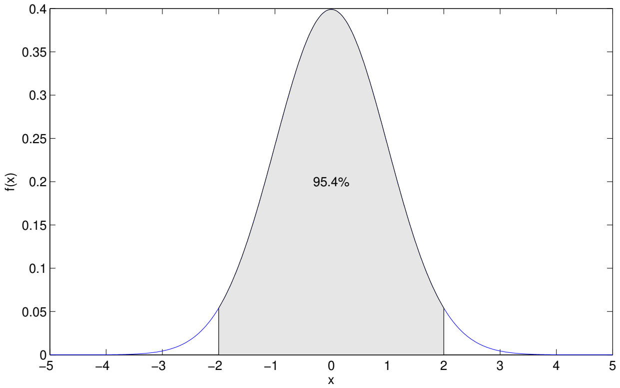
- Wahrscheinlichkeitsinhalte der Standardnormalverteilung



$$W(-1 \leq X \leq 1)$$

Die Gauß- oder Normalverteilung

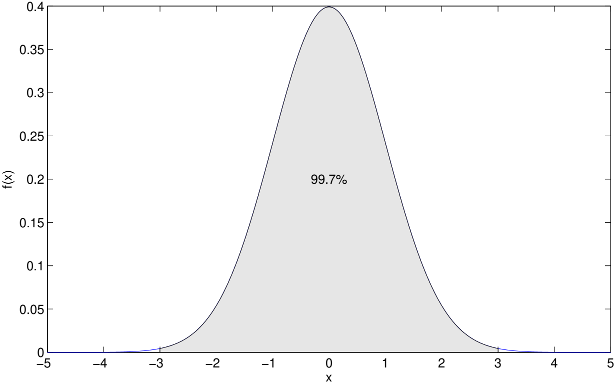
- Wahrscheinlichkeitsinhalte der Standardnormalverteilung



$$W(-2 \leq X \leq 2)$$

Die Gauß- oder Normalverteilung

- Wahrscheinlichkeitsinhalte der Standardnormalverteilung



$$W(-3 \leq X \leq 3)$$

Die Gauß- oder Normalverteilung

Fisher-Information einer normalverteilten Stichprobe

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$. Dann gilt:

- $I_\mu = \dfrac{n}{\sigma^2}$
- $I_{\sigma^2} = \dfrac{n}{2\sigma^4}$

Unterabschnitt: Schätzung des Mittelwerts

Schätzung des Mittelwerts

Satz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$ und $\overline{X}$ das Stichprobenmittel. Dann gilt:

- $\overline{X}$ ist normalverteilt gemäß $\mathrm{No}(\mu, \sigma^2/n)$.
- $\overline{X}$ ist ein erwartungstreuer und konsistenter Schätzer von μ.
- $\overline{X}$ ist effizient.

Unterabschnitt: Schätzung der Varianz

Schätzung der Varianz

Satz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$ und S^2 die Stichprobenvarianz. Dann gilt:

- $(n-1)S^2/\sigma^2$ ist χ^2-verteilt mit $n-1$ Freiheitsgraden.
- $\mathsf{E}[S^2] = \sigma^2$
- $\mathsf{var}[S^2] = \dfrac{2\sigma^4}{n-1}$
- S^2 ist ein erwartungstreuer und konsistenter Schätzer von σ^2.
- S^2 ist asymptotisch effizient.

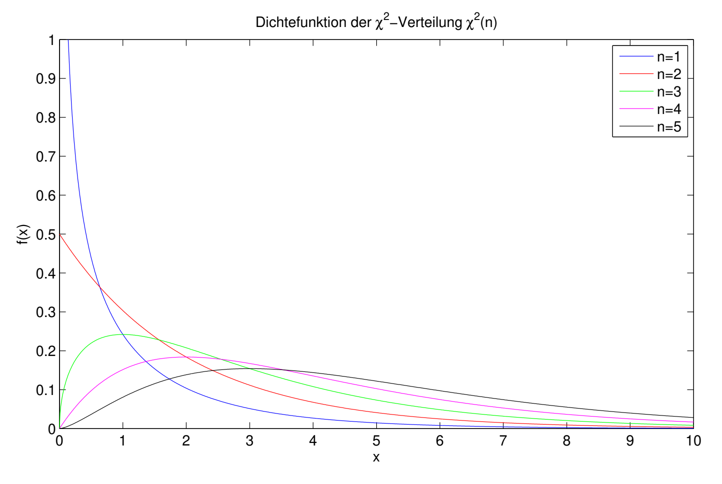
Exkurs: Die χ^2-Verteilung $\chi^2(n)$

- Die Dichte der $\chi^2(n)$-Verteilung mit n Freiheitsgraden lautet:

$$f_{\chi^2}(x;n) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})}\, x^{n/2-1} e^{-x/2} \cdot I_{[0,\infty)}(x)$$

- Sie ist die Gammaverteilung $\mathrm{Ga}(n/2, 2)$.
- Die $\chi^2(2)$-Verteilung ist die Exponentialverteilung $\mathrm{Ex}(2)$.
- Der Modus der Dichte ist bei $x = \max(n-2, 0)$.
- Die Dichte der $\chi^2(1)$-Verteilung hat einen Pol bei $x = 0$.
- Für $n \to \infty$ konvergiert die $\chi^2(n)$-Verteilung gegen die Normalverteilung $\mathrm{No}(n, 2n)$.
- Sind $X_1, \ldots, X_n$ unabhängig und standardnormalverteilt, so ist $Y = \sum_{i=1}^n X_i^2$ $\chi^2(n)$-verteilt mit n Freiheitsgraden.

Schätzung der Varianz



Schätzung der Varianz

Die Momente und die CF der $\chi^2(n)$-Verteilung

Es sei $X \sim \chi^2(n)$. Dann gilt:

- $\mathsf{E}[X] = n$
- $\mathsf{var}[X] = 2n$
- $\gamma[X] = \sqrt{8/n}$
- $\kappa[X] = 3 + 12/n$
- $\varphi_X(t) = (1 - 2\mathrm{i}t)^{-n/2}$

- Das α-Quantil der $\chi^2(n)$-Verteilung wird mit $\chi^2_{\alpha;n}$ bezeichnet.

Schätzung der Varianz

Exkurs: Die Gammaverteilung $\mathrm{Ga}(a, b)$

- Die Dichte der Gammaverteilung lautet:

$$f_{\mathrm{Ga}}(x; a, b) = \frac{x^{a-1} e^{-x/b}}{b^a \Gamma(a)} \cdot I_{[0,\infty)}(x)$$

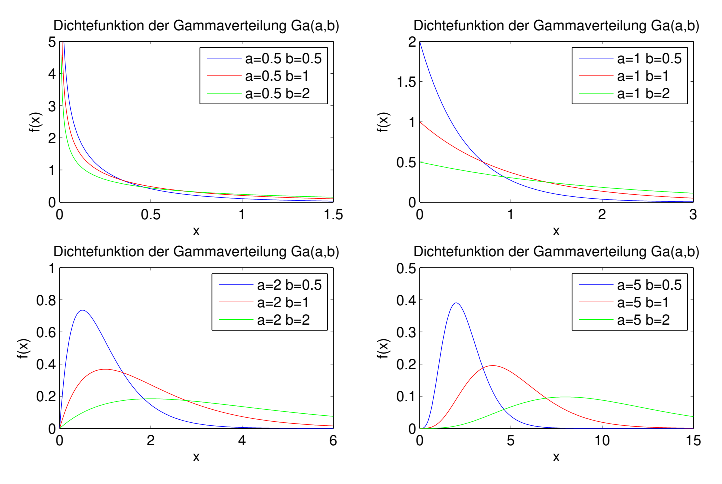
- Ihre Verteilungsfunktion ist die regularisierte unvollständige Gammafunktion:

$$F_{\mathrm{Ga}}(x; a, b) = \int_0^x \frac{t^{a-1} e^{-t/b}}{b^a \Gamma(a)} \,\mathrm{d}t = \frac{\gamma(a, x/b)}{\Gamma(a)}$$

- Der Modus M ist bei $m = (a-1)b$, wenn $a \geq 1$.
- Das α-Quantil der $\mathrm{Ga}(a, b)$-Verteilung wird mit $\gamma_{\alpha;a,b}$ bezeichnet.

PYTHON: `scipy.stats.gamma.pdf(x,a=`a`,scale=`b`)`

Schätzung der Varianz



Schätzung der Varianz

Die Momente und die CF der $\mathrm{Ga}(a, b)$-Verteilung

Es sei $X \sim \mathrm{Ga}(a, b)$. Dann gilt:

- $\mathsf{E}[X] = ab$
- $\mathsf{var}[X] = ab^2$
- $\gamma[X] = 2/\sqrt{a}$
- $\kappa[X] = 3 + 6/a$
- $\varphi_X(t) = (1 - \mathrm{i}bt)^{-a}$

Weitere Eigenschaften der $\mathrm{Ga}(a, b)$-Verteilung

- Ist $X \sim \mathrm{Ga}(a, b)$, dann ist $cX \sim \mathrm{Ga}(a, cb)$
- $\gamma_{\alpha;a,b} = b\,\gamma_{\alpha;a,1}$
- Sind $X_1 \sim \mathrm{Ga}(a_1, b)$ und $X_2 \sim \mathrm{Ga}(a_2, b)$ unabhängig, ist $Y = X_1 + X_2 \sim \mathrm{Ga}(a_1 + a_2, b)$

Unterabschnitt: Schätzung des Medians

Schätzung des Medians

Satz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$ und $\tilde{X}$ der Stichprobenmedian. Dann gilt:

- $\mathsf{E}[\tilde{X}] = \mu$
- $\mathsf{var}[\tilde{X}] \approx \dfrac{2\,\pi\sigma^2}{4\,n} \approx 1.57 \cdot \dfrac{\sigma^2}{n}$

Für großes n ist die Varianz von $\tilde{X}$ also um mehr als 50 Prozent größer als die Varianz von $\overline{X}$.

- Es gibt jedoch Verteilungen, für die der Stichprobenmedian eine kleinere Varianz hat als das Stichprobenmittel, wie das folgende Beispiel zeigt.

Schätzung des Medians

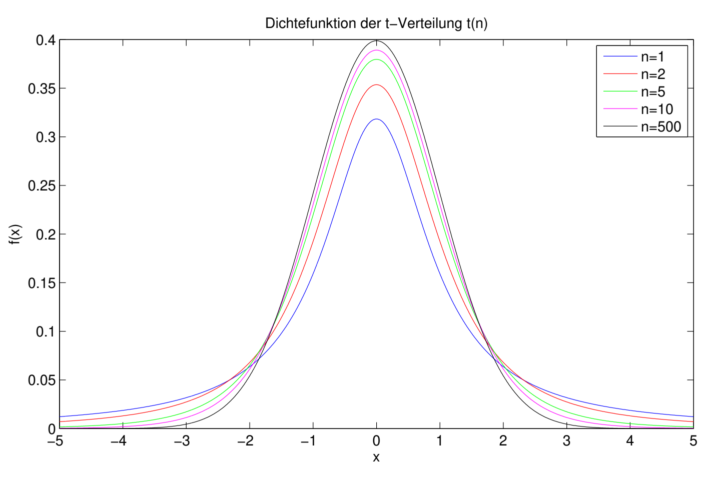
Exkurs: Die t-Verteilung $\mathrm{t}(n)$

- Die Dichte der $\mathrm{t}(n)$-Verteilung mit n Freiheitsgraden lautet:

$$f_{\mathrm{t}}(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

- Die $\mathrm{t}(1)$-Verteilung wird auch Cauchy- oder (in der Teilchenphysik) Breit–Wigner-Verteilung genannt.
- Das k-te Moment μ_k existiert nur für $k < n$.
- Für $n \to \infty$ konvergiert die $\mathrm{t}(n)$-Verteilung gegen die Standardnormalverteilung.

Schätzung des Medians



Schätzung des Medians

Die Momente der $\mathrm{t}(n)$-Verteilung

Es sei $X \sim \mathrm{t}(n)$. Dann gilt:

- $\mathsf{E}[X] = 0, \quad n > 1$
- $\mathsf{var}[X] = \dfrac{n}{n-2}, \quad n > 2$
- $\gamma[X] = 0, \quad n > 3$
- $\kappa[X] = \dfrac{3n-6}{n-4}, \quad n > 4$

- Das α-Quantil der $\mathrm{t}(n)$-Verteilung wird mit $\mathrm{t}_{\alpha;n}$ bezeichnet.

Schätzung des Medians

Beispiel

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der t-Verteilung $\mathrm{t}(3)$. Die Varianz von $\overline{X}$ ist gleich

$$\mathsf{var}(\overline{X}) = \frac{3}{n}$$

Die Varianz von $\tilde{X}$ ist für großes n gleich

$$\mathsf{var}(\tilde{X}) = \frac{1}{4\,n f(0)^2} = \frac{1.8506}{n} \approx 0.62 \cdot \frac{3}{n}$$

Sie ist also fast um 40 Prozent kleiner als die Varianz von $\overline{X}$.

Abschnitt 13: Exponentialverteilte Daten

Unterabschnitt: Die Exponentialverteilung

Die Exponentialverteilung

Die Exponentialverteilung $\mathrm{Ex}(\tau)$

- Die **Eponential**- oder **Wartezeitverteilung** ist eine wichtige Verteilungsfamilie in Naturwissenschaft und Technik. Wir bezeichnen sie mit $\mathrm{Ex}(\tau)$.
- Ihre Dichte lautet:

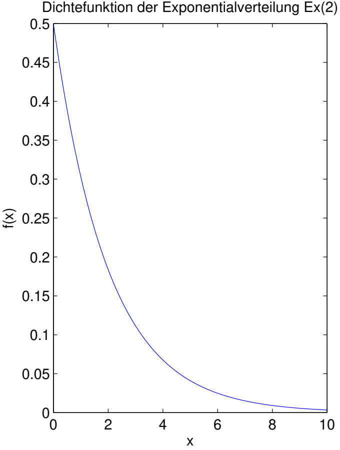
$$f_{\mathrm{Ex}}(x;\tau) = \frac{1}{\tau}\, e^{-x/\tau}$$

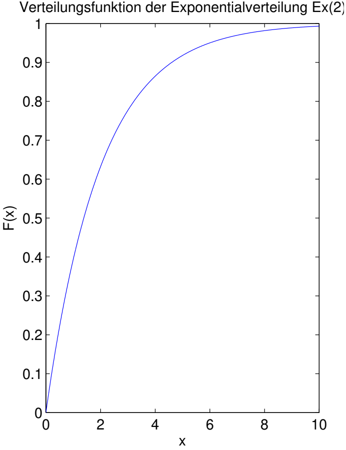
- Ihre Verteilungsfunktion lautet:

$$F_{\mathrm{Ex}}(x;\tau) = 1 - e^{-x/\tau}$$

- Die Exponentialverteilung ist ein Spezialfall der **Gammaverteilung**: $\mathrm{Ex}(\tau)$ ist gleich $\mathrm{Ga}(1,\tau)$.
- Der Modus M ist bei $x = 0$.

Die Exponentialverteilung





Die Exponentialverteilung

- Das α-Quantil der Exponentialverteilung $\mathrm{Ex}(\tau)$ lautet:

$$\gamma_{\alpha;1,\tau} = -\tau \ln(1-\alpha)$$

- Der Median der Verteilung ist folglich bei $x = \tau \ln 2$.
- Die Exponentialverteilung ist die Verteilung einer Wartezeit **ohne Gedächtnis**: Die Verteilung ist unabhängig vom Zeitpunkt, an dem die Zeitmessung startet.
- Dieses Verhalten ist typisch für physikalische Prozesse wie der radioaktive Zerfall eines Atoms oder der Zerfall eines instabilen Elementarteilchens.

Die Exponentialverteilung

Die Momente und die CF der Exponentialverteilung

Es sei $X \sim \mathrm{Ex}(\tau)$. Dann gilt:

- $\mathsf{E}[X] = \tau$
- $\mathsf{var}[X] = \tau^2$
- $\gamma[X] = 2$
- $\kappa[X] = 9$
- $\varphi_X(t) = (1 - \mathrm{i}\tau t)^{-1}$

Fisher-Information einer exponentialverteilten Stichprobe

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Exponentialverteilung $\mathrm{Ex}(\tau)$. Dann gilt:

$$I_\tau = \frac{n}{\tau^2}$$

Unterabschnitt: Schätzung des Mittelwerts

Schätzung des Mittelwerts

Satz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Exponentialverteilung $\mathrm{Ex}(\tau)$ und $\overline{X}$ das Stichprobenmittel. Dann gilt:

- $\overline{X}$ ist gammaverteilt gemäß $\mathrm{Ga}(n, \tau/n)$.
- $\overline{X}$ ist ein erwartungstreuer und konsistenter Schätzer von τ.
- $\overline{X}$ ist effizient.
- $\overline{X}$ ist asymptotisch normalverteilt mit Mittel $\mu = \tau$ und Varianz $\sigma^2 = \tau^2/n$.

Abschnitt 14: Poisson-verteilte Daten

Unterabschnitt: Der Poisson-Prozess

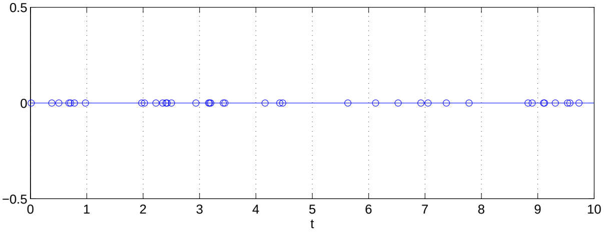
Der Poisson-Prozess

- Wir beobachten einen Prozess, bei dem gewisse Ereignisse zu zufälligen Zeitpunkten eintreten.
- Sind die Wartezeiten zwischen den Ereignissen unabhängig und exponentialverteilt mit Mittel τ, sprechen wir von einem **Poisson-Prozess** mit Intensität $\lambda = 1/\tau$.
- Die Anzahl X der Ereignisse pro Zeiteinheit ist dann unabhängig und **Poisson-verteilt** gemäß $\mathrm{Po}(\lambda)$.
- Ist umgekehrt die Anzahl X der Ereignisse pro Zeiteinheit unabhängig und Poisson-verteilt gemäß $\mathrm{Po}(\lambda)$, so sind die Wartezeiten exponentialverteilt mit Mittel $\tau = 1/\lambda$.
- In einem Poisson-Prozess der Intensität λ ist die Anzahl der Ereignisse pro Zeitintervall der Länge T wieder Poisson-verteilt gemäß $\mathrm{Po}(\lambda T)$.
- Die Summe von zwei Poisson-Prozessen ist wieder ein Poisson-Prozess. Seine Intensität ist die Summe der Intensitäten der Summanden.

Der Poisson-Prozess

Beispiele von Poisson-Prozessen

- Die Anzahl von Zerfällen pro Zeiteinheit in einer (großen) radioaktiven Quelle.
- Die Anzahl von Teilchen pro Zeiteinheit in der Höhenstrahlung.
- Die Anzahl der Pixelfehler auf einem TFT-Display.
- Die Anzahl seltener Ereignisse pro Zeiteinheit (Versicherungsfälle, Selbstmorde, Unfälle)



Der Poisson-Prozess

Die Poisson-Verteilung $\mathrm{Po}(\lambda)$

- Die Dichte der Poisson-Verteilung lautet:

$$f_{\mathrm{Po}}(k;\lambda) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad k \in \mathbb{N}_0$$

- Der Modus M ist gleich

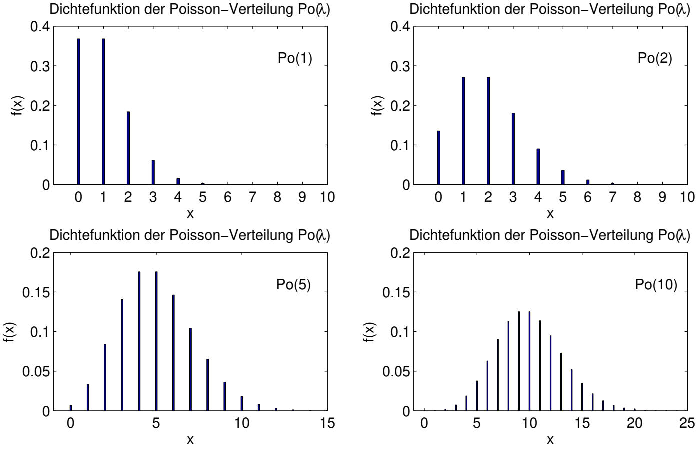
$$M = \begin{cases} \lfloor \lambda \rfloor, & \text{wenn } \lambda \notin \mathbb{N} \\ \lambda \text{ und } \lambda - 1, & \text{wenn } \lambda \in \mathbb{N} \\ 0, & \text{wenn } \lambda = 0 \end{cases}$$

- Die Verteilungsfunktion kann durch die Verteilungsfunktion der Gammaverteilung $\mathrm{Ga}(k+1,1)$ ausgedrückt werden:

$$F_{\mathrm{Po}}(k;\lambda) = 1 - F_{\mathrm{Ga}}(\lambda; k+1, 1)$$

Der Poisson-Prozess

Statistische Methoden der Datenanalyse
W. Waltenberger
Eigenschaften von Punktschätzern
Normalverteilte Daten
Exponentialverteilte Daten
Poisson-verteilte Daten
Der Poisson-Prozess
Schätzung des Mittelwerts
Daten aus Bernoulli- und Ziehungsexperimenten
Maximum-Likelihood-Schätzer
Mischverteilungen und EM-Algorithmus



Die Momente und die CF der Poisson-Verteilung

Es sei $X \sim \mathrm{Po}(\lambda)$. Dann gilt:

- $\mathsf{E}[X] = \lambda$
- $\mathsf{var}[X] = \lambda$
- $\gamma[X] = 1/\sqrt{\lambda}$
- $\kappa[X] = 3 + 1/\lambda$
- $\varphi_X(t) = \exp\left(\lambda\left(e^{\mathrm{i}t} - 1\right)\right)$

Fisher-Information einer Poisson-verteilten Stichprobe

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Poisson-Verteilung $\mathrm{Po}(\lambda)$. Dann gilt:

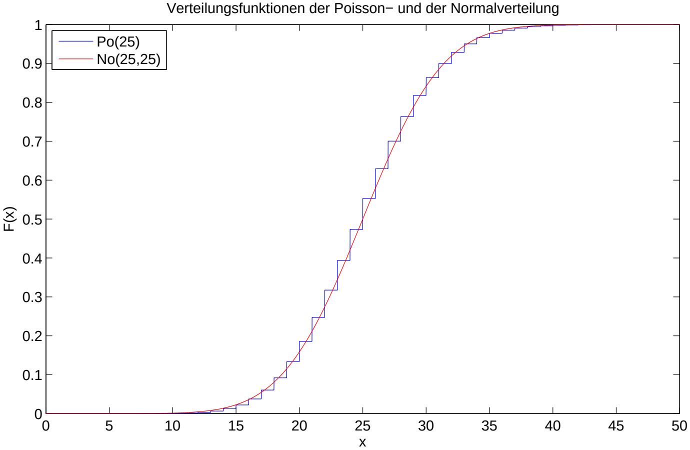
$$I_\lambda = \frac{n}{\lambda}$$

Der Poisson-Prozess

Die Poisson-Verteilung für großes λ

- Eine gemäß $\mathrm{Po}(\lambda)$ verteilte Zufallsvariable kann als Summe von λ $P(1)$-verteilten Zufallsvariablen dargestellt werden.
- Nach dem zentralen Grenzwertsatz muss daher die Poisson-Verteilung für $\lambda \to \infty$ gegen eine Normalverteilung streben.
- Die folgende Abbildung zeigt die Verteilungsfunktion der Poisson-Verteilung $\mathrm{Po}(\lambda)$ mit $\lambda = 25$, sowie die Verteilungsfunktion der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$ mit $\mu = \lambda = 25$ und $\sigma^2 = \lambda = 25$.

Der Poisson-Prozess



Unterabschnitt: Schätzung des Mittelwerts

Schätzung des Mittelwerts

Satz

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Poisson-Verteilung $\mathrm{Po}(\lambda)$ und $\overline{X}$ das Stichprobenmittel. Dann gilt:

- $\overline{X}$ ist ein erwartungstreuer und konsistenter Schätzer von λ.
- $\overline{X}$ ist effizient.
- $\overline{X}$ ist asymptotisch normalverteilt mit Mittel $\mu = \lambda$ und Varianz $\sigma^2 = \lambda/n$.

Abschnitt 15: Daten aus Bernoulli- und Ziehungsexperimenten

Unterabschnitt: Wiederholte Bernoulli-Experimente

Wiederholte Bernoulli-Experimente

Die Alternativ- oder Bernoulliverteilung $\mathrm{Al}(p)$

- Ein Bernoulli-Experiment oder Alternativversuch hat zwei mögliche Ergebnisse, "Erfolg" bzw. "Misserfolg".
- Die Zufallsvariable X ordnet dem Erfolg den Wert 1 und dem Misserfolg den Wert 0 zu.
- Ist p die Wahrscheinlichkeit eines Erfolgs, lautet die Dichte $f_{\mathrm{Al}}(x;p)$ folgendermaßen:

$$f_{\mathrm{Al}}(0;p) = 1 - p,\ \ f_{\mathrm{Al}}(1;p) = p$$

oder

$$f_{\mathrm{Al}}(x;p) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}$$

Wiederholte Bernoulli-Experimente

- Wird der Alternativversuch n mal unabhängig durchgeführt, so gibt es 2^n mögliche Ergebnisse, nämlich alle Folgen der Form $\boldsymbol{e} = (i_1, \ldots, i_n), \quad i_j \in \{0, 1\}$.
- Die diskrete Zufallsvariable Y bildet die Folge $\boldsymbol{e}$ auf die Häufigkeit von 1 ab:

$$Y(\boldsymbol{e}) = \sum_{j=1}^{n} i_j$$

- Der Wertebereich von Y ist die Menge $\{0, 1, \ldots, n\}$. Auf die Zahl k $(0 \leq k \leq n)$ werden alle Folgen abgebildet, bei denen 1 genau k-mal auftritt. Es gibt C_k^n solche Folgen, und jede hat die Wahrscheinlichkeit $p^k(1-p)^{n-k}$.

Wiederholte Bernoulli-Experimente

Die Binomialverteilung $\mathrm{Bi}(n,p)$

- Die Dichte f von Y ist daher:

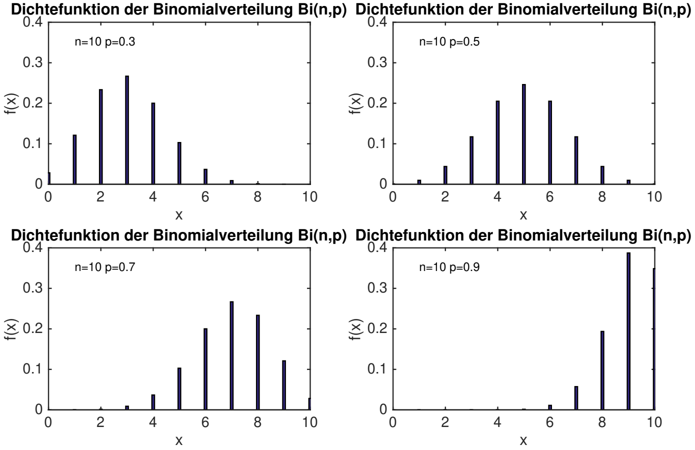
$$f_{\mathrm{Bi}}(k;p) = \binom{n}{k} p^k (1-p)^{n-k},\ 0 \le k \le n$$

- Die Verteilung von Y wird als **Binomialverteilung** $\mathrm{Bi}(n,p)$ mit den Parametern n und p bezeichnet.
- Es gilt:

$$\sum_{k=0}^{n} f_{\mathrm{Bi}}(k;p) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

Das ist gerade der binomische Lehrsatz.

Wiederholte Bernoulli-Experimente



Wiederholte Bernoulli-Experimente

- Der Modus M ist gleich

$$M = \begin{cases} \lfloor (n+1)p \rfloor, & \text{wenn } p = 0 \text{ oder } (n+1)p \notin \mathbb{N} \\ (n+1)p \text{ und} & \\ (n+1)p - 1, & \text{wenn } (n+1)p \in \{1, \dots, n\} \\ 1, & \text{wenn } p = 1 \end{cases}$$

- Die Verteilungsfunktion kann durch die Verteilungsfunktion der Betaverteilung $\mathrm{Be}(x; n-k, k+1)$ ausgedrückt werden:

$$F_{\mathrm{Bi}}(k; n, p) = F_{\mathrm{Be}}(1-p; n-k, k+1)$$

Exkurs: Die Betaverteilung $\mathrm{Be}(a,b)$

- Die Dichte der Betaverteilung lautet:

$$f_{\mathrm{Be}}(x;a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \cdot I_{[0,1]}(x)$$

- Ihre Verteilungsfunktion ist die regularisierte unvollständige Betafunktion:

$$F_{\mathrm{Be}}(x;a,b) = \int_0^x \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)}\,\mathrm{d}t$$

- Der Modus M ist bei $x=(a-1)/(a+b-2)$, wenn $a,b>1$.

Wiederholte Bernoulli-Experimente

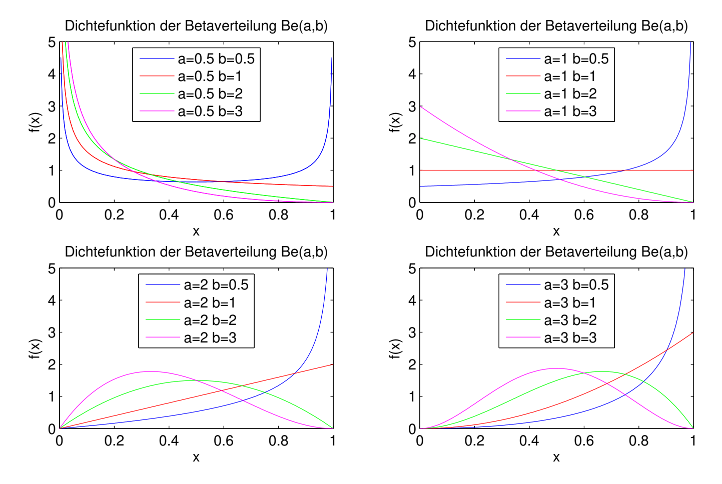
Die Momente der $\mathrm{Be}(a,b)$-Verteilung

Es sei $X \sim \mathrm{Be}(a,b)$. Dann gilt:

- $\mathsf{E}[X] = \dfrac{a}{a+b}$
- $\mathsf{var}[X] = \dfrac{ab}{(a+b)^2(a+b+1)}$
- $\gamma[X] = \dfrac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$
- $\kappa[X] = \dfrac{6(a-b)^2+3ab(a+b+2)}{ab(a+b+2)(a+b+3)}$

- Das α-Quantil der $\mathrm{Be}(a,b)$-Verteilung wird mit $\beta_{\alpha;a,b}$ bezeichnet.

Wiederholte Bernoulli-Experimente



Wiederholte Bernoulli-Experimente

Die Momente und die CF der Binomialverteilung

Es sei $X \sim \mathrm{Bi}(n, p)$. Dann gilt:

- $\mathsf{E}[X] = np$
- $\mathsf{var}[X] = np(1-p)$
- $\gamma[X] = \dfrac{1-2p}{\sqrt{np(1-p)}}$
- $\kappa[X] = 3 - \dfrac{6}{n} + \dfrac{1}{np(1-p)}$
- $\varphi_X(t) = \left(1 - p + p e^{\mathrm{i}t}\right)^n$

Fisher-Information einer binomialverteilten Beobachtung

Es sei X eine Beobachtung aus der Binomialverteilung $\mathrm{Bi}(n, p)$. Dann gilt:

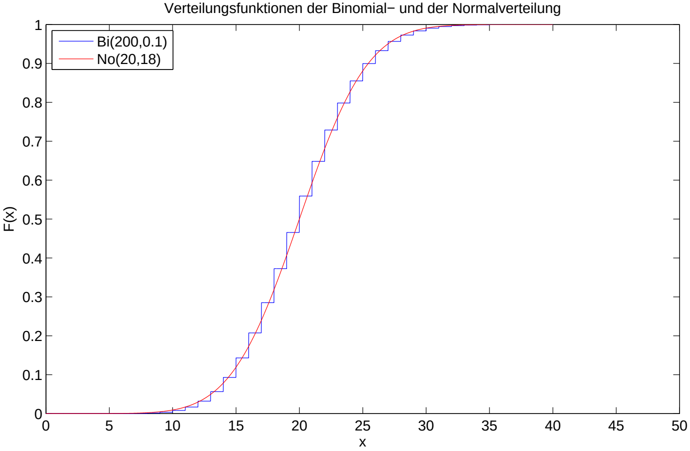
$$I_p = \frac{n}{p(1-p)}$$

Wiederholte Bernoulli-Experimente

Die Binomialverteilung für großes n

- Eine gemäß $\mathrm{Bi}(n, p)$ verteilte Zufallsvariable kann als Summe von n Bernoulli-verteilten Zufallsvariablen dargestellt werden.
- Nach dem zentralen Grenzwertsatz muss daher die Binomial-Verteilung für $n \to \infty$ und festes p gegen eine Normalverteilung streben.
- Die folgende Abbildung zeigt die Verteilungsfunktion der Binomial-Verteilung $\mathrm{Bi}(n, p)$ mit $n = 200$ und $p = 0.1$, sowie die Verteilungsfunktion der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$ mit $\mu = np = 20$ und $\sigma^2 = np(1 - p) = 18$.

Wiederholte Bernoulli-Experimente



Wiederholte Bernoulli-Experimente

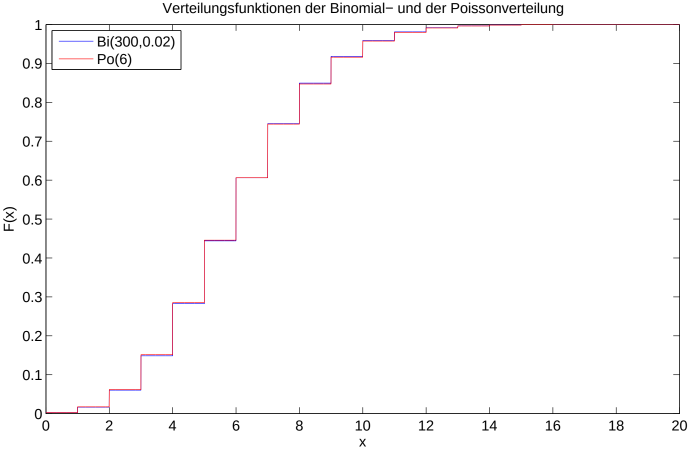
Die Binomialverteilung für großes n und $p = \lambda/n$

- Im Grenzwert $n \to \infty$ und $p = \lambda/n \to 0$ strebt die Binomialverteilung $\mathrm{Bi}(n, p)$ gegen die Poisson-Verteilung $\mathrm{Po}(\lambda)$:

$$\lim_{n\to\infty} \binom{n}{k} (\lambda/n)^k (1-\lambda)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

- Wird also die Versuchsanzahl n vergrößert und die Erfolgswahrscheinlichkeit verkehrt proportional dazu verkleinert, ist die Erfolgsanzahl asymptotisch Poisson-verteilt.

Wiederholte Bernoulli-Experimente



Schätzung der Erfolgswahrscheinlichkeit

Satz

Es sei k eine Beobachtung aus der Binomialverteilung $\mathrm{Bi}(n, p)$. Dann gilt:

- $\hat{p} = k/n$ ist ein erwartungstreuer Schätzer von p.
- $\hat{p}$ ist effizient.
- $\hat{p}$ ist asymptotisch normalverteilt mit Mittel $\mu = p$ und Varianz $\sigma^2 = p(1-p)/n$.

Unterabschnitt: Ziehungsexperimente

Ziehungsexperimente

Ziehen mit Zurücklegen

- Grundgesamtheit von N Objekten, davon sind M Objekte **Merkmalsträger**, haben also eine bestimmte Eigenschaft E.
- Es werden n Objekte gezogen, wobei jedes Objekt die gleiche Wahrscheinlichkeit hat, gezogen zu werden.
- Jedes gezogene Objekte wird sofort wieder zurückgelegt.
- Die Anzahl der gezogenen Merkmalsträger ist eine Zufallsvariable X.
- X ist dann **binomialverteilt** gemäß $\mathrm{Bi}(n, M/N)$.
- Sind N und M viel größer als n, ist X auch dann in guter Näherung binomialverteilt, wenn gezogene Objekte **nicht** zurückgelegt werden.
- Ein Beispiel dafür ist eine Umfrage, in der die Stichprobengröße n wesentlich kleiner als N und M ist.

Ziehungsexperimente

Ziehen ohne Zurücklegen

- Grundgesamtheit von N Objekten, davon sind M Objekte **Merkmalsträger**, haben also eine bestimmte Eigenschaft E.
- Es werden n Objekte gezogen, wobei jedes Objekt die gleiche Wahrscheinlichkeit hat, gezogen zu werden.
- Einmal gezogene Objekte werden **nicht** wieder zurückgelegt.
- Die Anzahl der gezogenen Merkmalsträger ist eine Zufallsvariable X.
- Die Verteilung von X wird **hypergeometrische Verteilung** $\mathrm{Hy}(N, M, n)$ genannt.

Ziehungsexperimente

Die hypergeometrische Verteilung $\mathrm{Hy}(N, M, n)$

- Ihre Dichte lautet:

$$f_{\mathrm{Hy}}(m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}, \quad 0 \leq m \leq \min(n, M)$$

- Der Modus ("mode") ist gleich

$$\mathrm{mode} = \left\lfloor \frac{(n+1)(M+1)}{N+2} \right\rfloor$$

Ziehungsexperimente

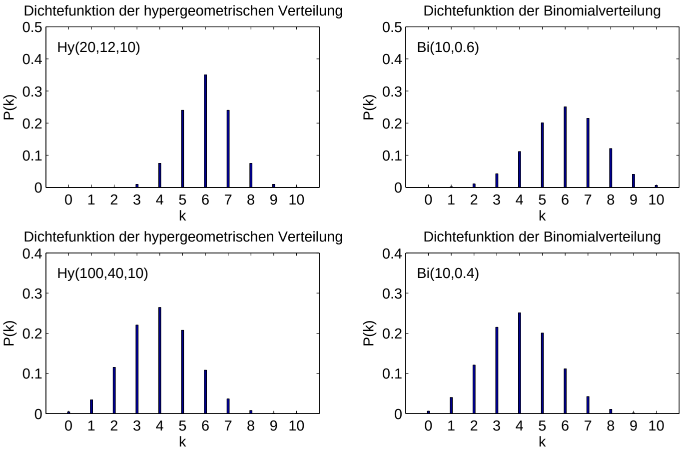
Die Momente der hypergeometrischen Verteilung

Es sei $X \sim \mathrm{Hy}(N, M, n)$ und $p = M/N$. Dann gilt:

- $\mathsf{E}[X] = \dfrac{nM}{N} = np$
- $\mathsf{var}[X] = \dfrac{nM}{N} \dfrac{N-M}{N} \dfrac{N-n}{N-1} = np(1-p)\dfrac{N-n}{N-1}$

- Der Term $(N-n)(N-1)$ in der Varianz wird **Endlichkeitskorrektur** gennant. Er ist nahe bei 1, wenn $N \gg n$.
- Ist $N \gg n$ und $M \gg n$, kann die hypergeometrische Verteilung durch die Binomialverteilung angenähert werden, weil dann die Ziehung der Stichprobe die Zusammensetzung der Grundgesamtheit nur unwesentlich verändert.

Ziehungsexperimente



Ziehungsexperimente

Schätzung der Anzahl der Merkmalsträger

Satz

Es sei k eine Beobachtung aus der hypergeometrischen Verteilung $\mathrm{Hy}(N, M, n)$ mit bekanntem N, aber unbekanntem M. Dann gilt:

- $\hat{M} = \dfrac{kN}{n}$ ist ein erwartungstreuer Schätzer von M.
- $\mathrm{var}[\hat{M}] = M\,\dfrac{N-M}{n}\,\dfrac{N-n}{N-1}$

Ziehungsexperimente

Schätzung der Größe der Grundgesamtheit

Satz

Es sei k eine Beobachtung aus der hypergeometrischen Verteilung $\mathrm{Hy}(N, M, n)$ mit bekanntem M, aber unbekanntem N. Dann gilt:

- $\hat{N} = \dfrac{nM}{k}$ ist ein nicht erwartungstreuer Schätzer von N.
- $1/\hat{N} = \dfrac{k}{nM}$ ist ein erwartungstreuer Schätzer von $1/N$.
- Mit Fehlerfortpflanzung erhalten wir:

$$\mathsf{E}[\hat{N}] \approx N + \frac{N-M}{M}\,\frac{N-n}{n}$$

$$\mathsf{var}[\hat{N}] \approx N\,\frac{N-M}{M}\,\frac{N-n}{n}$$

Abschnitt 16: Maximum-Likelihood-Schätzer

Maximum-Likelihood-Schätzer

Definition (ML-Schätzer)

- Es sei $X_1, \ldots, X_n$ eine Stichprobe mit der gemeinsamen Dichte $g(x_1, \ldots, x_n|\vartheta)$. Die Funktion

$$L(\vartheta|X_1, \ldots, X_n) = g(X_1, \ldots, X_n|\vartheta)$$

heißt die **Likelihoodfunktion** der Stichprobe.
- Der **plausible** oder **Maximum-Likelihood-Schätzer** $\hat{\vartheta}$ ist jener Wert von ϑ, der die Likelihoodfunktion der Stichprobe maximiert.

- Oft wird statt der Likelihoodfunktion ihr Logarithmus, die Log-Likelihoodfunktion $\ell(\vartheta) = \ln L(\vartheta)$ maximiert.
- In einfachen Fällen gelingt die Maximierung analytisch, sonst muss numerisch maximiert werden.

Maximum-Likelihood-Schätzer

- Der ML-Schätzer ist invariant unter (differenzierbaren) Transformationen des Parameters: Ist $\hat{\vartheta}$ der ML-Schätzer von ϑ, so ist $h(\hat{\vartheta})$ der ML-Schätzer von $h(\vartheta)$.

Beispiel (ML-Schätzung eines Bernoulli-Parameters)

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Bernoulli-Verteilung $\mathrm{Al}(p)$. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n - \sum x_i}$$

Die Log-Likelihoodfunktion ist daher:

$$\ell(p) = \sum_{i=1}^{n} X_i \ln p + \left(n - \sum_{i=1}^{n} X_i \right) \ln(1-p)$$

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Ableiten nach p ergibt:

$$\frac{\partial \ell(p)}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} X_i - \frac{1}{1-p} \left(n - \sum_{i=1}^{n} X_i \right)$$

Nullsetzen der Ableitung und Auflösen nach p ergibt:

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

Der ML-Schätzer ist unverzerrt und effizient.

Maximum-Likelihood-Schätzer

Beispiel (ML-Schätzung eines Poisson-Parameters)

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Poisson-Verteilung $\mathrm{Po}(\lambda)$. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Die Log-Likelihoodfunktion ist daher:

$$\ell(\lambda) = \sum_{i=1}^{n} [X_i \ln \lambda - \lambda - \ln(x_i!)]$$

Ableiten nach λ ergibt:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n$$

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Nullsetzen der Ableitung und Auflösen nach λ ergibt:

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

Der ML-Schätzer ist unverzerrt und effizient.

Maximum-Likelihood-Schätzer

Beispiel (ML-Schätzung einer mittleren Lebensdauer)

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Exponentialverteilung $\mathrm{Ex}(\tau)$. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | \tau) = \prod_{i=1}^{n} \frac{e^{-x_i/\tau}}{\tau}$$

Die Log-Likelihoodfunktion ist daher:

$$\ell(\tau) = \sum_{i=1}^{n} [-\ln \tau - \frac{1}{\tau} X_i]$$

Ableiten nach τ ergibt:

$$\frac{\partial \ell(\tau)}{\partial \tau} = -\frac{n}{\tau} + \frac{1}{\tau^2} \sum_{i=1}^{n} X_i$$

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Nullsetzen der Ableitung und Auflösen nach τ ergibt:

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

Der ML-Schätzer ist unverzerrt und effizient.

Maximum-Likelihood-Schätzer

Beispiel (ML-Schätzung der Parameter einer Normalverteilung)

Es sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\,\sigma^2}\right]$$

Die Log-Likelihoodfunktion ist daher:

$$\ell(\mu, \sigma^2 | \boldsymbol{X}) = \sum_{i=1}^{n} \left[-\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma^2 - \frac{(X_i - \mu)^2}{2\,\sigma^2}\right]$$

Ableiten nach μ und σ^2 ergibt:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma^2}, \quad \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \left[-\frac{1}{2\,\sigma^2} + \frac{(X_i - \mu)^2}{2\,\sigma^4}\right]$$

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Nullsetzen der Ableitungen und Auflösen nach μ und σ^2 ergibt:

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

Der ML-Schätzer von μ ist unverzerrt und effizient. Der ML-Schätzer von σ^2 ist asymptotisch unverzerrt und asymptotisch effizient.
Der ML-Schätzer von σ ist gleich

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}}S$$

Er ist ebenfalls asymptotisch unverzerrt und asymptotisch effizient.

Maximum-Likelihood-Schätzer

- Die normierte Likelihoodfunktion kann als a-posteriori Verteilung des geschätzten Parameters interpretiert werden.
- Für großes n kann man die Varianz der Likelihoodschätzung $\hat{\vartheta}$ daher aus dem zweiten zentralen Moment der normierten Likelihoodfunktion ablesen.
- Ist der geschätzte Parameter ϑ das Mittel einer Normalverteilung, so ist diese Vorgangsweise für beliebiges n exakt:

$$L(\vartheta) = \frac{1}{\sigma^n \sqrt{2\,\pi}^n}\, \exp\left[-\frac{n}{2\,\sigma^2}\left((\hat{\vartheta}-\vartheta)^2 + \frac{1}{n}\sum (x_i - \hat{\vartheta})^2\right)\right]$$

- Wird $L(\vartheta)$ normiert, so entsteht die „Dichte“ einer Normalverteilung mit Mittel $\hat{\vartheta}$ und Varianz $\frac{\sigma^2}{n}$, also gerade die Varianz der Schätzung $\hat{\vartheta} = \frac{1}{n}\sum x_i$.

Beispiel (Schätzung des Parameters a einer Gammaverteilung)

Die Stichprobe $\boldsymbol{X} = (X_1, \ldots, X_n)$ besteht aus $n = 200$ Werten, die unabhängig aus einer $\mathrm{Ga}(a, 1)$-Verteilung gezogen werden:

$$f(x_i|a) = \frac{x_i^{a-1} e^{-x_i}}{\Gamma(a)}, \quad i = 1, \ldots, n$$

Der (unbekannte) wahre Wert von a ist $a_{\mathrm{w}} = 2$. Die Log-Likelihoodfunktion lautet

$$\ln L(a|\boldsymbol{X}) = \sum_{i=1}^{n} \ln f(X_i|a) = (a-1)\sum_{i=1}^{n} \ln X_i - \sum_{i=1}^{n} X_i - n \ln \Gamma(a)$$

Maximum-Likelihood-Schätzer

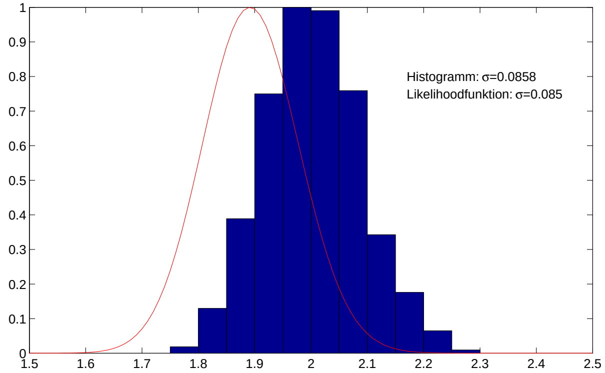
Beispiel (Fortsetzung)

Numerische Maximierung von $\ln L(a)$ gibt die Maximum-Likelihood-Schätzung $\hat{a}$. Das Experiment wird N-mal wiederholt und die Schätzungen der einzelnen Experimente $(\hat{a}^{(k)}, k = 1, \ldots, N)$ werden histogrammiert. Der Vergleich der individuellen (normierten) Likelihoodfunktion mit dem Histogramm $(N = 500)$ zeigt gute Übereinstimmung der Standardabweichungen.

MATLAB: `make_ML_gamma`

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)



Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Bei bekanntem b ist die Fisher-Information der Stichprobe bezüglich a gleich

$$I_a = n \frac{\mathrm{d}^2 \ln \Gamma(a)}{\mathrm{d}a^2} \approx 128.99$$

Dies entspricht einer Standardabweichung $\sigma = 0.088$. Der ML-Schätzer ist also praktisch effizient.

Wegen $\mathsf{E}[\overline{X}] = a$ ist $\overline{X}$ ein für alle n erwartungstreuer Schätzer von a. Er hat jedoch eine größere Standardabweichung als der ML-Schätzer:

$$\sigma[\overline{X}] = \sqrt{\frac{a}{n}} = 0.1$$

Andererseits ist er wesentlich einfacher zu berechnen.

Maximum-Likelihood-Schätzer

- Der ML-Schätzer hat die folgende wichtige Eigenschaft:

Satz

Existieren die ersten beiden Ableitungen von $L(\vartheta)$, existiert die Information $I_g(\vartheta)$ für alle ϑ und ist $E\left[(\ln L)'\right] = 0$, so ist die Likelihoodschätzung $\hat{\vartheta}$ asymptotisch normalverteilt mit Mittel ϑ und Varianz $1/I_g(\vartheta)$. $\hat{\vartheta}$ ist daher asymptotisch erwartungstreu und asymptotisch effizient.

- Daraus folgt sofort die nächste Eigenschaft:

Satz

Der Likelihoodschätzer $\hat{\vartheta}$ ist (unter den selben Voraussetzungen) konsistent.

Maximum-Likelihood-Schätzer

Beispiel (ML-Schätzung des Lageparameters einer Cauchyverteilung)

Es sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus der Cauchyverteilung $\mathrm{t}(1)$ mit Lageparameter μ. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\pi[1 + (x_i - \mu)^2]}$$

Die Log-Likelihoodfunktion ist daher:

$$\ell(\mu|\boldsymbol{X}) = -n \ln \pi - \sum_{i=1}^{n} \ln[1 + (X_i - \mu)^2]$$

Das Maximum $\hat{\mu}$ von $\ell(\mu|\boldsymbol{X})$ muss numerisch gefunden werden.

MATLAB: `make_ML_cauchy`

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Man kann zeigen, dass die Fisher-Information der Stichprobe gleich

$$I_\mu = \frac{n}{2}$$

ist. Für große Stichproben muss daher die Varianz des ML-Schätzers $\hat{\mu}$ ungefähr gleich $2/n$ sein.

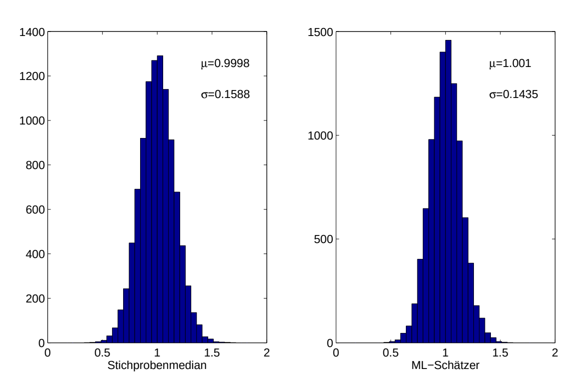
Der Stichprobenmedian $\tilde{X}$ ist ebenfalls ein konsistenter Schätzer für μ. Seine Varianz ist asymptotisch gleich $\pi^2/(4n) \approx 2.47/n$. Sie ist also um etwa 23 Prozent größer als die Varianz des ML-Schätzers.

Das Stichprobenmittel $\overline{X}$ ist dagegen **kein** konsistenter Schätzer für μ. Man kann nämlich zeigen, dass $\overline{X}$ die gleiche Verteilung wie eine einzelne Beobachtung hat.

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Simulation von 10000 Stichproben der Größe $n = 100$:

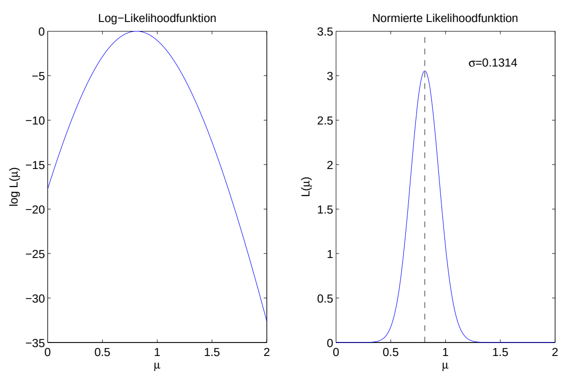


Die Korrelation zwischen $\tilde{x}$ und $\hat{\mu}$ ist etwa 90%.

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Die Standardabweichung des ML-Schätzers kann wieder näherungsweise aus der normierten Likelihoodfunktion einer Stichprobe abgelesen werden:



Maximum-Likelihood-Schätzer

Beispiel (ML-Schätzung des Obergrenze einer Gleichverteilung)

Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Gleichverteilung $\mathrm{Un}(0, b)$ mit Obergrenze b. Die gemeinsame Dichte lautet:

$$g(x_1, \ldots, x_n | b) = \frac{1}{b^n}, 0 \leq x_1, \ldots, x_n \leq b$$

Der größte Wert der Likelihoodfunktion ist daher bei

$$\hat{b} = \max_i X_i$$

Da ein Randmaximum vorliegt, gelten die üblichen asymptotischen Eigenschaften nicht.

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Die Dichte von $\hat{b} = \max_i X_i$ lautet:

$$f(x) = \frac{nx^{n-1}}{b^n}$$

Daraus können Erwartung und Varianz berechnet werden:

$$\mathsf{E}[\hat{b}] = \frac{bn}{n+1}, \quad \mathsf{var}[\hat{b}] = \frac{b^2 n}{(n+2)(n+1)^2}$$

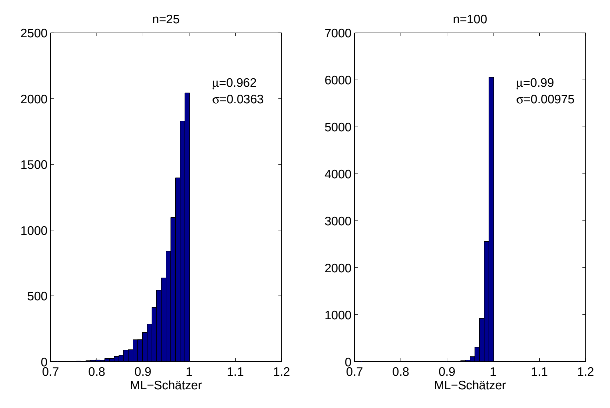
Der Schätzer ist asymptotisch erwartungstreu, die Varianz geht aber wie $1/n^2$ gegen Null! Der Schätzer ist auch nicht asymptotisch normalverteilt.

MATLAB: `make_ML_uniform`

Maximum-Likelihood-Schätzer

Beispiel (Fortsetzung)

Simulation von 10000 Stichproben ($b = 1$) der Größe $n = 25$ bzw. $n = 100$:



Maximum-Likelihood-Schätzer

- Ist die Likelihoodfunktion (annähernd) normal, ist die Log-Likelihoodfunktion (annähernd) parabolisch.
- In diesem Fall können aus der Log-Likelihoodfunktion Fehlerintervalle des ML-Schätzers abgelesen werden.

Breite der Log-Likelihoodfunktion

Es sei $\ell(\vartheta)$ die Log-Likelihoodfunktion des Parameters ϑ, ferner $\hat{\vartheta}$ der ML-Schätzwert und $\sigma[\hat{\vartheta}]$ seine Standardabweichung. Dann gilt näherungsweise:

$$\ell(\hat{\vartheta} - k\sigma[\hat{\vartheta}]) = \ell(\hat{\vartheta} + k\sigma[\hat{\vartheta}]) = \ell(\hat{\vartheta}) - k^2/2$$

Maximum-Likelihood-Schätzer

- Werden zwei Parameter $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$ gleichzeitig geschätzt, ist die Log-Likelihoodfunktion $\ell(\boldsymbol{\vartheta})$ asymptotisch in der Regel ein elliptisches Paraboloid.
- Die Höhenschichtlinien von $\ell(\boldsymbol{\vartheta})$ sind dann Ellipsen.
- Soll eine Höhenschichtlinie den Wahrscheinlichkeitsinhalt $1 - \alpha$ haben, so ist ihre Höhe z gleich

$$z = \ell(\hat{\boldsymbol{\vartheta}}) - \frac{1}{2}\chi^2_{1-\alpha;2} = \ell(\hat{\boldsymbol{\vartheta}}) + \ln(\alpha)$$

- Zum Beispiel ist die Ellipse mit $1 - \alpha = 0.95$ die Höhenschichtlinie zur Höhe $z = \ell(\hat{\boldsymbol{\vartheta}}) - 2.996$.
- Die Kovarianzmatrix des ML-Schätzers $\hat{\boldsymbol{\vartheta}}$ kann durch die inverse negative Hesse-Matrix im Maximum angenähert werden.

Maximum-Likelihood-Schätzer

Beispiel

Mittelwert μ und Standardabweichung σ einer Normalverteilung werden aus einer Stichprobe vom Umfang $n = 500$ geschätzt. Die wahren Werte sind $\mu = 5, \sigma = 1.5$, die Schätzwerte sind $\hat{\mu} = 4.980, \hat{\sigma} = 1.529$. Die inverse negative Hesse-Matrix im Maximum der Log-Likelihoodfunktion ist gleich

$$\mathbf{V} = \begin{pmatrix} 0.0047 & 0 \\ 0 & 0.0023 \end{pmatrix}$$

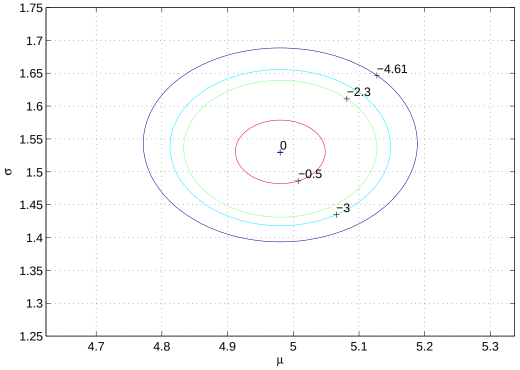
Das entspricht einem Standardfehler von $\sigma[\hat{\mu}] = 0.0684$ und $\sigma[\hat{\sigma}] = 0.0483$. Die inverse Fisher-Informationsmatrix ist gleich

$$\mathbf{I}_{\mu,\sigma} = \begin{pmatrix} 0.0045 & 0 \\ 0 & 0.0022 \end{pmatrix}$$

MATLAB: `make_ML_normal_2D`

Beispiel (Fortsetzung)

Höhenschichtlinien der Log-Likelihoodfunktion mit den Wahrscheinlichkeitsinhalten $0.3935, 0.9, 0.95, 0.99$.



Numerische Berechnung der Hesse-Matrix

- Es sei $\ell(\vartheta_1, \vartheta_2)$ die Log-Likelihoodfunktion, mit Maximum in $(\hat{\vartheta}_1, \hat{\vartheta}_2)$.
- Ferner sei $\varepsilon_1, \varepsilon_2 > 0$ und

$$\mathbf{L} = \begin{pmatrix} \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2 - \varepsilon_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2 - \varepsilon_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2 - \varepsilon_2) \\ \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2) \\ \ell(\hat{\vartheta}_1 - \varepsilon_1, \hat{\vartheta}_2 + \varepsilon_2) & \ell(\hat{\vartheta}_1, \hat{\vartheta}_2 + \varepsilon_2) & \ell(\hat{\vartheta}_1 + \varepsilon_1, \hat{\vartheta}_2 + \varepsilon_2) \end{pmatrix}$$

- Die Hesse-Matrix ist dann gleich

$$\mathbf{H} = \begin{pmatrix} \dfrac{\mathbf{L}_{21} + \mathbf{L}_{23} - 2\,\mathbf{L}_{22}}{\varepsilon_1^2} & \dfrac{\mathbf{L}_{11} + \mathbf{L}_{33} - \mathbf{L}_{13} - \mathbf{L}_{31}}{4\,\varepsilon_1 \varepsilon_2} \\ \dfrac{\mathbf{L}_{11} + \mathbf{L}_{33} - \mathbf{L}_{13} - \mathbf{L}_{31}}{4\,\varepsilon_1 \varepsilon_2} & \dfrac{\mathbf{L}_{12} + \mathbf{L}_{32} - 2\,\mathbf{L}_{22}}{\varepsilon_2^2} \end{pmatrix}$$

Abschnitt 17: Mischverteilungen und EM-Algorithmus

Mischverteilungen und EM-Algorithmus

- Experimentelle Beobachtungen stammen oft aus verschiedenen Verteilungen, z.B. Signal und Untergrund.
- Solche Daten können durch eine **Mischverteilung** beschrieben werden.

Definition (Mischverteilung)

Eine **Mischverteilung** mit k Komponenten ist eine Verteilung, deren Dichte die folgende Form hat:

$$f(x) = \sum_{j=1}^{k} w_j f_j(x), \quad w_j \geq 0, \quad \sum_{j=1}^{k} w_j = 1$$

Die w_j sind nichtnegativ und werden als die **Gewichte** der Komponenten bezeichnet.

- Die Komponenten f_j sind typischerweise Normalverteilungen oder andere einfache Verteilungen.

Mischverteilungen und EM-Algorithmus

- Das l-te Moment μ'_l um 0 der Mischverteilung ist die Mischung der entsprechenden Momente μ'_{lj} um 0 der Komponenten:

$$\mu'_l = \sum_{j=1}^{k} w_j \mu'_{lj}$$

Erwartung und Varianz einer Mischverteilung

Eine Mischverteilung bestehe aus k Komponenten mit den Mittelwerten μ_j und den Varianzen σ_j^2. Dann sind der Mittelwert μ und die Varianz σ^2 der Mischverteilung gegeben durch:

$$\mu = \sum_{j=1}^{k} w_j \mu_j, \qquad \sigma^2 = \sum_{j=1}^{k} w_j \left(\mu_j^2 + \sigma_j^2\right) - \mu^2$$

Mischverteilungen und EM-Algorithmus

Beispiel

- Eine Mischverteilung aus k normalverteilten Komponenten hat $3k-1$ Parameter, nämlich

$$\boldsymbol{\vartheta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, w_1, \ldots, w_{k-1})$$

- Eine Mischverteilung aus k exponentialverteilten Komponenten hat $2k-1$ Parameter, nämlich

$$\boldsymbol{\vartheta} = (\tau_1, \ldots, \tau_k, w_1, \ldots, w_{k-1})$$

Mischverteilungen und EM-Algorithmus

- Die Schätzung der Parameter kann mit der Maximum-Likelihood-Methode erfolgen.
- Im Fall eines Gemischs aus Normalverteilungen lautet die Log-Likelihoodfunktion einer Stichprobe $\boldsymbol{X}$ vom Umfang n:

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} w_j \cdot \varphi(x_i; \mu_j, \sigma_j^2) \right)$$

- Die Log-Likelihoodfunktion muss numerisch maximiert werden, unter Berücksichtigung der Nebenbedingung $\sum_{j=1}^{k} w_j = 1$.
- Die Log-Likelihoodfunktion kann mehrere lokale Maxima haben. In diesem Fall empfiehlt sich wiederholte Maximierung mit verschiedenen Startwerten.

Mischverteilungen und EM-Algorithmus

- Eine Methode zur Maximierung der Log-Likelihoodfunktion ist der **EM-Algorithmus**.
- Der EM-Algorithmus ist iterativ. Die Wahl der Startwerte kann beeinflussen, welches lokale Maximum erreicht wird.
- In jeder Iteration werden mit dem Satz von Bayes die a-posteriori-Wahrscheinlichkeiten p_{ij} berechnet, dass Beobachtung i aus Komponente j stammt.
- Die Mittelwerte und Varianzen der Komponenten werden dann durch gewichtete Stichprobenmittel bzw. gewichtete Stichprobenvarianzen geschätzt.
- Diese beiden Schritte werden iteriert, bis die Schätzung sich stabilisiert.
- Die Log-Likelihoodfunktion kann in keiner Iteration kleiner werden; das Erreichen des globalen Maximums ist jedoch **nicht** garantiert.

Mischverteilungen und EM-Algorithmus

EM-Algorithmus für ein Gemisch von Normalverteilungen

1. Wahl von Startwerten für die Mischungsparameter $\boldsymbol{\vartheta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, w_1, \ldots, w_k)$
2. Berechnung der p_{ij} und p_j:

$$p_{ij} = \frac{w_j \, \varphi(X_i; \mu_j, \sigma_j^2)}{\sum_{l=1}^{k} w_l \; \varphi(X_i; \mu_l, \sigma_l^2)}, \quad p_j = \sum_{i=1}^{n} p_{ij}$$

3. Schätzung der Gewichte und Momente:

$$w_j = \frac{p_j}{n}, \;\; \mu_j = \frac{\sum_{i=1}^{n} p_{ij} X_i}{p_j}, \;\; \sigma_j^2 = \frac{\sum_{i=1}^{n} p_{ij} (X_i - \mu_j)^2}{p_j}$$

4. Wiederholung der Schritte 2 und 3 bis zur Konvergenz.

Mischverteilungen und EM-Algorithmus

Beispiel

Datensatz 2 ist ein Gemisch von 500 Werten aus $\mathrm{No}(5,1)$ und 100 Werten aus $\mathrm{No}(8,9)$. Der EM-Algorithmus ergibt folgende Schätzwerte der Mischungsparameter:

$$\mu_1 = 4.946, \sigma_1 = 1.053, w_1 = 0.850$$
$$\mu_2 = 8.206, \sigma_2 = 2.944, w_2 = 0.150$$

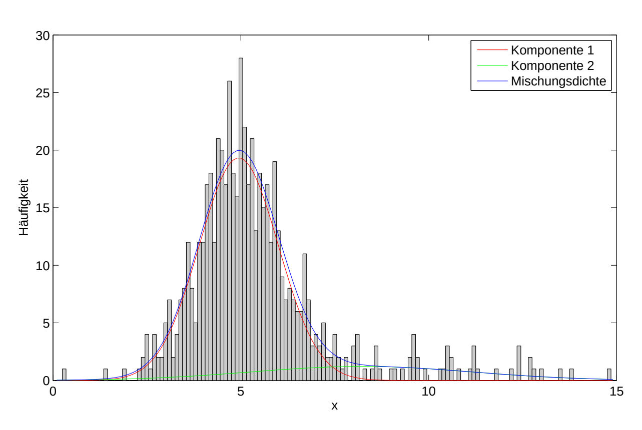
Maximierung der Log-Likelihoodfunktion mit dem Simplexalgorithmus nach Nelder und Mead ergibt fast identische Werte:

$$\mu_1 = 4.946, \sigma_1 = 1.052, w_1 = 0.850$$
$$\mu_2 = 8.195, \sigma_2 = 2.946, w_2 = 0.150$$

MATLAB: `make_EM_normal`

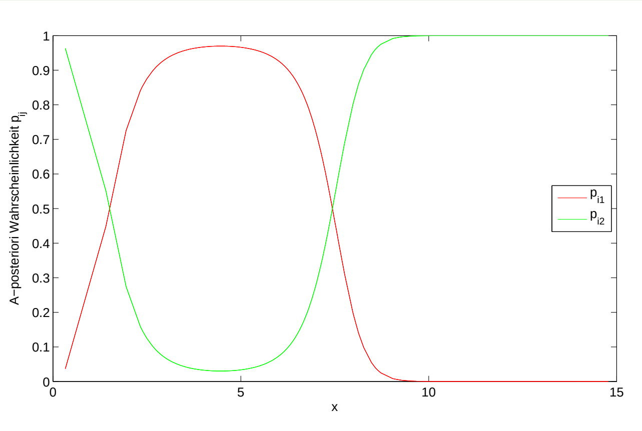
Beispiel (Fortsetzung)

Dichten der Komponenten und Mischungsdichte:



Beispiel (Fortsetzung)

A-posteriori Wahrscheinlichkeiten p_{ij}:



Teil V

Konfidenzintervalle

Übersicht Teil 5

Abschnitt 18: Grundbegriffe

Grundbegriffe

- Neben dem Schätzwert selbst ist auch seine Streuung um den wahren Wert von Interesse.
- Wir wollen aus einer Stichprobe ein Intervall bestimmen, das den wahren Wert mit einer gewissen Wahrscheinlichkeit enthält.

Definition (Konfidenzintervall)

Es sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus der Verteilung F mit dem unbekannten Parameter ϑ und $0 < \alpha < 1$. Ein Intervall mit den Grenzen $G_1 = g_1(\boldsymbol{X})$ und $G_2 = g_2(\boldsymbol{X})$ heißt ein **Konfidenzintervall** mit Sicherheit $1 - \alpha$, wenn gilt:

$$W(G1 \leq G2) = 1$$
$$W(G1 \leq \vartheta \leq G2) \geq 1 - \alpha$$

Ein solches Intervall wird kurz als $(1 - \alpha)$-Konfidenzintervall bezeichnet.

Grundbegriffe

- Zu jedem Wert der Sicherheit $1-\alpha$ gibt es viele verschiedene Konfidenzintervalle.
- Ist F stetig, gibt es unendlich viele Konfidenzintervalle mit Sicherheit $1-\alpha$.
- Ist F diskret, ist die Sicherheit in der Regel größer als $1-\alpha$.
- Ein **symmetrisches** Konfidenzintervall liegt vor, wenn gilt:

$$W(\vartheta \le G_1) = W(\vartheta \ge G_2)$$

- Ein **einseitiges** Konfidenzintervall liegt vor, wenn gilt:

linksseitig: $W(\vartheta \le G_2) \ge 1-\alpha$ oder

rechtsseitig: $W(G_1 \le \vartheta) \ge 1-\alpha$

Abschnitt 19: Allgemeine Konstruktion nach Neyman

Allgemeine Konstruktion nach Neyman

- Es sei $Y = h(\boldsymbol{X})$ eine Stichprobenfunktion. Die Verteilung G von Y hängt dann ebenfalls vom unbekannten Parameter ϑ ab.
- Für jeden Wert von ϑ bestimmen wir ein **Prognoseintervall** $[y_1(\vartheta), y_2(\vartheta)]$ vom Niveau $1-\alpha$:

$$W(y_1(\vartheta) \leq Y \leq y_2(\vartheta)) \geq 1 - \alpha$$

- Ist die Beobachtung gleich $Y = y_0$, so ist das Konfidenzintervall $[G_1(Y), G_2(Y)]$ gegeben durch:

$$G1 = \min_{\vartheta}\{\vartheta | y_1(\vartheta) \leq y_0 \leq y_2(\vartheta)\}$$
$$G2 = \max_{\vartheta}\{\vartheta | y_1(\vartheta) \leq y_0 \leq y_2(\vartheta)\}$$

Allgemeine Konstruktion nach Neyman

Beispiel

Es sei $\boldsymbol{X}$ eine Stichprobe aus $\mathrm{No}(\mu, \sigma^2)$ mit unbekannter Varianz σ^2. Dann ist $(n-1)S^2/\sigma^2$ χ^2-verteilt mit $n-1$ Freiheitsgraden. Es gilt daher:

$$W\left(\frac{\sigma^2\chi^2_{\alpha/2;n-1}}{n-1} \leq S^2 \leq \frac{\sigma^2\chi^2_{1-\alpha/2;n-1}}{n-1}\right) = 1-\alpha$$

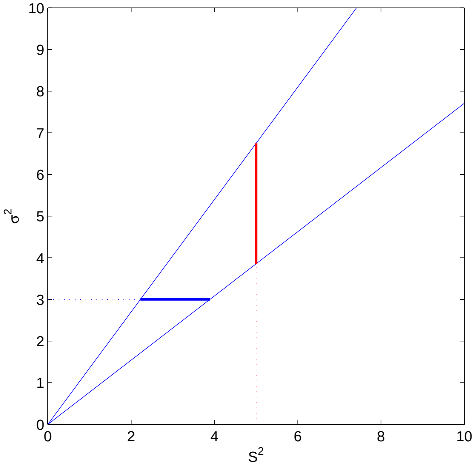
Der Ausdruck in der Klammer kann umgeformt werden zu:

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

Daraus folgt

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

Allgemeine Konstruktion nach Neyman



Blau: Prognoseintervall für $\sigma^2 = 3$; rot: Konfidenzintervall für $S^2 = 5$

Abschnitt 20: Normalverteilte Daten

Unterabschnitt: Mittelwert

Mittelwert

- Es sei $\boldsymbol{X}$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$.
- $\overline{X}$ ist normalverteilt gemäß $\mathrm{No}(\mu, \sigma^2/n)$.
- Ist σ^2 bekannt, ist das Standardscore

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

standardnormalverteilt. Aus

$$W(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

folgt dann

$$W(\overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

Mittelwert

Symmetrisches KI für den Mittelwert, bekannte Varianz

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}$$

- Analog dazu werden die links- und rechtsseitigen Konfidenzintervalle gebildet.

Linksseitiges KI für den Mittelwert, bekannte Varianz

$$G_1(\boldsymbol{X}) = -\infty, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha}\sigma/\sqrt{n}$$

Rechtsseitiges KI für den Mittelwert, bekannte Varianz

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha}\sigma/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \infty$$

Mittelwert

Quantile der Standardnormalverteilung

α	$z_{1-\alpha/2}$	$z_{1-\alpha}$
0.001	3.29	3.09
0.002	3.09	2.88
0.005	2.81	2.58
0.01	2.58	2.33
0.02	2.33	2.05
0.05	1.96	1.64
0.1	1.64	1.28

Mittelwert

- Ist σ^2 unbekannt, wird σ^2 durch die Stichprobenvarianz S^2 geschätzt. Das Standardscore

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ist dann t-verteilt mit $n-1$ Freiheitsgraden. Aus

$$W(-t_{1-\alpha/2;n-1} \leq T \leq t_{1-\alpha/2;n-1}) = 1 - \alpha$$

folgt

Symmetrisches KI für den Mittelwert, unbekannte Varianz

$$G_1(\boldsymbol{x}) = \overline{X} - t_{1-\alpha/2;n-1} S/\sqrt{n},\ G_2(\boldsymbol{X}) = \overline{X} + t_{1-\alpha/2;n-1} S/\sqrt{n}$$

Mittelwert

Beispiel

Eine Stichprobe vom Umfang $n = 50$ aus der Standardnormalverteilung hat das Stichprobenmittel $\overline{X} = 0.0540$ und die Stichprobenvarianz $S^2 = 1.0987$. Wird die Varianz als bekannt vorausgesetzt, lautet das symmetrische 95%-Konfidenzintervall für μ:

$$G_1 = 0.0540 - 1.96/\sqrt{50} = -0.2232$$
$$G_2 = 0.0540 + 1.96/\sqrt{50} = 0.3312$$

Wird die Varianz als unbekannt angenommen, lautet das symmetrische 95%-Konfidenzintervall für μ:

$$G_1 = 0.0540 - 2.01 \cdot 1.0482/\sqrt{50} = -0.2439$$
$$G_2 = 0.0540 + 2.01 \cdot 1.0482/\sqrt{50} = 0.3519$$

Unterabschnitt: Varianz

Varianz

- Es sei $X_1, \ldots, X_n$ eine Stichprobe aus der Normalverteilung $\mathrm{No}(\mu, \sigma^2)$.
- $(n-1)S^2/\sigma^2$ ist χ^2-verteilt mit $n-1$ Freiheitsgraden. Aus

$$W\left(\chi^2_{\alpha/2;n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2;n-1}\right) = 1-\alpha$$

folgt

Symmetrisches KI für die Varianz

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}$$

Varianz

Linksseitiges KI für die Varianz

$$G_1(\boldsymbol{X}) = 0, \quad G_2(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{\alpha;n-1}}$$

Rechtsseitiges KI für die Varianz

$$G_1(\boldsymbol{X}) = \frac{(n-1)S^2}{\chi^2_{1-\alpha;n-1}}, \quad G_2(\boldsymbol{X}) = \infty$$

Varianz

Beispiel

Eine Stichprobe vom Umfang $n = 50$ aus der Normalverteilung $\mathrm{No}(0, 4)$ hat die Stichprobenvarianz $S^2 = 4.3949$. Das symmetrische 95%-Konfidenzintervall für σ^2 lautet:

$$\begin{aligned} G_1 =& 49 \cdot 4.3949/70.2224 = 3.0667 \\ G_2 =& 49 \cdot 4.3949/31.5549 = 6.8246 \end{aligned}$$

Werden die Quantile der χ^2-Verteilung $\chi^2(n-1)$ durch die Quantile der Normalverteilung $\mathrm{No}(n-1, 2(n-1))$ ersetzt, lautet das Konfidenzintervall:

$$\begin{aligned} G_1 =& 49 \cdot 4.3949/68.4027 = 3.1483 \\ G_2 =& 49 \cdot 4.3949/29.5973 = 7.2760 \end{aligned}$$

Unterabschnitt: Differenz von zwei Mittelwerten

Differenz von zwei Mittelwerten

- Es seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ zwei unabhängige Stichproben aus den Normalverteilungen $\mathrm{No}(\mu_x, \sigma_x^2)$ bzw. $\mathrm{No}(\mu_y, \sigma_y^2)$.
- Wir suchen ein Konfidenzintervall für $\mu_x - \mu_y$. Die Differenz $D = \overline{X} - \overline{Y}$ ist normalverteilt gemäß $\mathrm{No}(\mu_x - \mu_y, \sigma^2)$, mit $\sigma_D^2 = \sigma_x^2/n + \sigma_y^2/m$.
- Sind die Varianzen bekannt, ist das Standardscore von D standardnormalverteilt. Aus

$$W\left(-z_{1-\alpha/2} \leq \frac{D - (\mu_x - \mu_y)}{\sigma_D} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

folgt

Symmetrisches KI für die Differenz von zwei Mittelwerten

$$G_1(\boldsymbol{X}) = D - z_{1-\alpha/2}\sigma_D, \quad G_2(\boldsymbol{X}) = D + z_{1-\alpha/2}\sigma_D$$

Differenz von zwei Mittelwerten

- Sind die Varianzen unbekannt und gleich, ist

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

χ^2-verteilt mit $k = n + m - 2$ Freiheitsgraden.

- Das Standardscore

$$T = \frac{D - (\mu_x - \mu_y)}{S_D}$$

mit $S_D = S\sqrt{1/n + 1/m}$ ist daher t-verteilt mit $k = n + m - 2$ Freiheitsgraden.

Differenz von zwei Mittelwerten

- Aus

$$W\left(-t_{1-\alpha/2;k} \leq T \leq t_{1-\alpha/2;k}\right) = 1-\alpha$$

folgt

Symmetrisches KI für die Differenz von zwei Mittelwerten, unbekannte Varianz

$$G_1(\boldsymbol{X}) = D - t_{1-\alpha/2;k} S_D, \quad G_2(\boldsymbol{X}) = D + t_{1-\alpha/2;k} S_D$$

mit $k = n + m - 2$.

Differenz von zwei Mittelwerten

Beispiel

Eine Stichprobe aus $\mathrm{No}(2,4)$ vom Umfang $n = 50$ hat Stichprobenmittel $\overline{X} = 2.1080$ und Stichprobenvarianz $S_x^2 = 4.3949$; eine zweite Stichprobe aus $\mathrm{No}(1,4)$ vom Umfang $m = 25$ hat Stichprobenmittel $\overline{Y} = 1.6692$ und Stichprobenvarianz $S_y^2 = 5.2220$. Werden die Varianzen als bekannt vorausgesetzt, lautet das 95%-Konfidenzintervall für $\mu_x - \mu_y$:

$$G_1 = 0.4388 - 1.96 \cdot 0.4899 = -0.5213$$

$$G_2 = 0.4388 + 1.96 \cdot 0.4899 = 1.3990$$

Differenz von zwei Mittelwerten

Beispiel (Fortsetzung)

Werden die Varianzen als unbekannt angenommen, ist $S^2 = 4.6668$ und $S_D = 0.5292$. Das 95%=Konfidenzintervall für $\mu_x - \mu_y$ lautet dann:

$$G_1 = 0.4388 - 1.993 \cdot 0.5292 = -0.6158$$
$$G_2 = 0.4388 + 1.993 \cdot 0.5292 = 1.4935$$

Abschnitt 21: Exponentialverteilte Daten

Exponentialverteilte Daten

- Es sei $\boldsymbol{X}$ eine Stichprobe aus der Exponentialverteilung $\mathrm{Ex}(\tau)$.
- Das Stichprobenmittel $\overline{X}$ ist dann Gamma-verteilt gemäß $\mathrm{Ga}(n, \tau/n)$ und hat die folgende Dichte:

$$f(x) = \frac{x^{n-1}}{(\tau/n)^n \Gamma(n)} \exp\left(-\frac{x}{\tau/n}\right)$$

- Für jedes τ gilt daher:

$$W\left(\gamma_{\alpha/2;n,\tau/n} \leq \overline{X} \leq \gamma_{1-\alpha/2;n,\tau/n}\right) = 1 - \alpha$$

Exponentialverteilte Daten

- Daraus folgt

$$W\left(\gamma_{\alpha/2;n,1/n} \leq \frac{\overline{X}}{\tau} \leq \gamma_{1-\alpha/2;n,1/n}\right) = 1-\alpha$$

und

$$W\left(\frac{\overline{X}}{\gamma_{1-\alpha/2;n,1/n}} \leq \tau \leq \frac{\overline{X}}{\gamma_{\alpha/2;n,1/n}}\right) = 1-\alpha$$

- Damit gilt:

Symmetrisches KI für den Mittelwert

$$G_1(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{1-\alpha/2;n,1/n}}, \quad G_2(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{\alpha/2;n,1/n}}$$

Exponentialverteilte Daten

Linksseitiges KI für den Mittelwert

$$G_1(\boldsymbol{X}) = 0, \quad G_2(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{\alpha;n,1/n}}$$

Rechtsseitiges KI für den Mittelwert

$$G_1(\boldsymbol{X}) = \frac{\overline{X}}{\gamma_{1-\alpha;n,1/n}}, \quad G_2(\boldsymbol{X}) = \infty$$

Abschnitt 22: Poisson-verteilte Daten

Poisson-verteilte Daten

- Es sei K eine Beobachtung aus der Poisson-Verteilung $\mathrm{Po}(\lambda)$.
- Die Quantile der Poisson-Verteilung können mit Hilfe der Verteilungsfunktion der Gammaverteilung berechnet werden.
- Für die praktische Berechnung des Konfidenzintervalls können daher die **Quantile der Gammaverteilung** benützt werden.

Symmetrisches KI für den Mittelwert

$$G_1(K) = \gamma_{\alpha/2;K,1}, \quad G_2(K) = \gamma_{1-\alpha/2;K+1,1}$$

Poisson-verteilte Daten

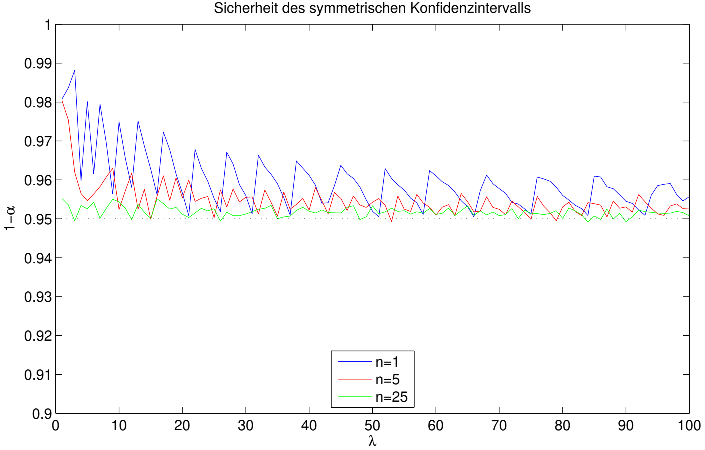
- Liegen n Beobachtungen $K_1, \ldots, K_n$ vor, so ist $L = \sum K_i$ Poisson-verteilt mit Mittel $n\lambda$. Das symmetrische Konfidenzintervall für λ ist dann:

Symmetrisches KI für den Mittelwert

$$G_1(L) = \gamma_{\alpha/2;L,1/n}, \quad G_2(L) = \gamma_{1-\alpha/2;L+1,1/n}$$

- Dieses Intervall ist **konservativ** in dem Sinn, dass die Sicherheit praktisch immer größer als $1 - \alpha$ ist.

Poisson-verteilte Daten



Poisson-verteilte Daten

- Auf analoge Weise werden die einseitigen Konfidenzintervalle bestimmt.

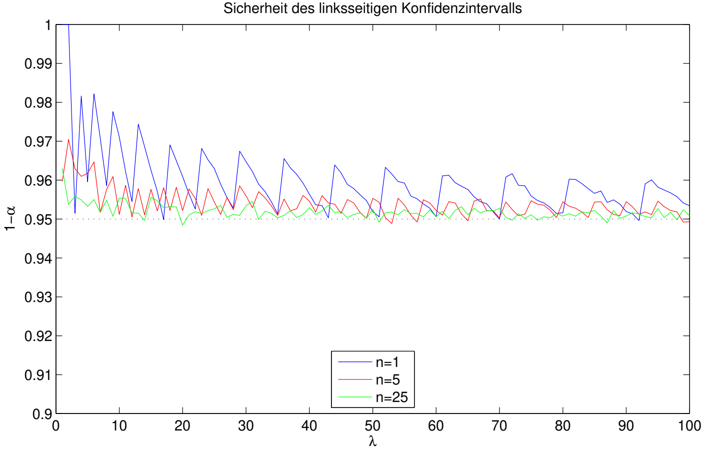
Linksseitiges KI für den Mittelwert

$$G_1(L) = 0, \quad G_2(L) = \gamma_{1-\alpha;L+1,1/n}$$

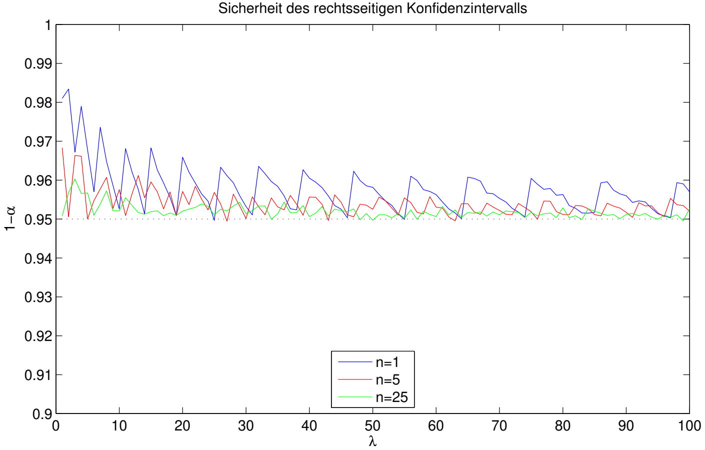
Rechtsseitiges KI für den Mittelwert

$$G_1(L) = \gamma_{\alpha;L,1/n}, \quad G_2(L) = \infty$$

Poisson-verteilte Daten



Poisson-verteilte Daten



Abschnitt 23: Binomialverteilte Daten

Binomialverteilte Daten

- Es sei K eine Beobachtung aus der Binomialverteilung $\mathrm{Bi}(n, p)$.
- Die Quantile der Binomialverteilung können mit Hilfe der Verteilungsfunktion der Betaverteilung berechnet werden.
- Für die praktische Berechnung des Konfidenzintervalls können daher die **Quantile der Betaverteilung** benützt werden.

Symmetrisches KI nach Clopper und Pearson

$$G_1(K) = \beta_{\alpha/2;K,n-K+1}, \quad G_2(K) = \beta_{1-\alpha/2;K+1,n-K}$$

- Sonderfälle: für $K = 0$ ist $G_1(0) = 0$, für $K = n$ ist $G_2(n) = 1$.
- Dieses Intervall ist **konservativ** in dem Sinn, dass die Sicherheit praktisch immer größer als $1 - \alpha$ ist.

Binomialverteilte Daten

- Auf analoge Weise werden die einseitigen Konfidenzintervalle bestimmt.

Linksseitiges KI nach Clopper und Pearson

$$G_1(K) = 0, \quad G_2(K) = \beta_{1-\alpha;K+1,n-K}$$

- Ist $K = 0$, erstreckt sich das linksseitige Konfidenzintervall von 0 bis $\beta_{1-\alpha;1,n} = 1 - \sqrt[n]{\alpha}$.

Rechtsseitiges KI nach Clopper und Pearson

$$G_1(K) = \beta_{\alpha;K,n-K+1}, \quad G_2(K) = 1$$

- Ist $K = n$, erstreckt sich das rechtsseitige Konfidenzintervall von $\beta_{\alpha;n,1} = \sqrt[n]{\alpha}$ bis 1.

Binomialverteilte Daten

- Für genügend großes n ist $\hat{p} = K/n$ annähernd normalverteilt gemäß $\mathrm{No}(p, p(1-p)/n)$.
- Das Standardscore
$$Z = \frac{\hat{p} - p}{\sigma[p]}$$
ist dann annähernd standardnormalverteilt.
- Aus
$$W(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$
folgt

$$W(\hat{p} - z_{1-\alpha/2}\sigma[p] \leq p \leq \hat{p} + z_{1-\alpha/2}\sigma[p]) = 1 - \alpha$$

Binomialverteilte Daten

- Da p nicht bekannt ist, muss $\sigma[p]$ näherungsweise bestimmt werden.
- **Bootstrap-Verfahren**: p wird durch $\hat{p}$ angenähert:

$$\sigma[p] \approx \sigma[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Symmetrisches KI nach dem Bootstrap-Verfahren

$$G_1(K) = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$G_2(K) = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Binomialverteilte Daten

- **Robustes Verfahren**: p wird so gewählt, dass $\sigma[p]$ maximal ist, also $p = 0.5$ und

$$\sigma[p] \approx \frac{1}{2\sqrt{n}}$$

Symmetrisches KI nach dem robusten Verfahren

$$G_1(K) = \hat{p} - z_{1-\alpha/2}\frac{1}{2\sqrt{n}}$$

$$G_2(K) = \hat{p} + z_{1-\alpha/2}\frac{1}{2\sqrt{n}}$$

Binomialverteilte Daten

- Verfahren nach **Agresti–Coull**: Wir setzen

$$c = z_{1-\alpha/2}, K' = K + c^2/2, n' = n + c^2, p' = K'/n'$$

und

$$\sigma[p'] = \sqrt{\frac{p'(1-p')}{n'}}$$

Symmetrisches KI nach Agresti-Coull

$$G_1(K) = p' - z_{1-\alpha/2}\sqrt{\frac{p'(1-p')}{n}}$$

$$G_2(K) = p' + z_{1-\alpha/2}\sqrt{\frac{p'(1-p')}{n}}$$

Binomialverteilte Daten

Beispiel

Angabe: Bei einer Umfrage unter $n = 400$ Personen geben $k = 157$ Personen an, Produkt X zu kennen. Wir suchen ein 95%-Konfidenzintervall für den Bekanntheitsgrad p.

Clopper-Pearson:

$$G_1(k) = \beta_{0.025;157,244} = 0.3443$$
$$G_2(k) = \beta_{0.975;158,243} = 0.4423$$

Approximation durch Normalverteilung:

Es gilt $\hat{p} = 0.3925$ und $z_{0.975} = 1.96$. Mit dem Bootstrap-Verfahren ergibt sich $\sigma[\hat{p}] = 0.0244$. Die Grenzen des Konfidenzintervalls sind daher

$$G_1 = 0.3925 - 1.96 \cdot 0.0244 = 0.3446$$
$$G_2 = 0.3925 + 1.96 \cdot 0.0244 = 0.4404$$

Binomialverteilte Daten

Beispiel (Fortsetzung)

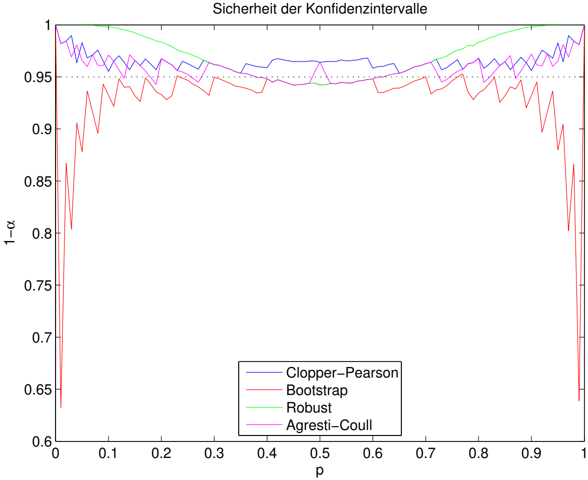
Mit dem robusten Verfahren ergibt sich $\sigma[\hat{p}] = 0.025$ und die Grenzen

$$G_1 = 0.3925 - 1.96 \cdot 0.025 = 0.3435$$
$$G_2 = 0.3925 + 1.96 \cdot 0.025 = 0.4415$$

Das robuste Intervall ist nur unwesentlich länger als das Bootstrap-Intervall.
Mit der Korrektur von Agresti-Coull ergibt sich $\hat{p} = 0.3936$. Die Grenzen des Konfidenzintervalls sind dann

$$G_1 = 0.3936 - 1.96 \cdot 0.0244 = 0.3457$$
$$G_2 = 0.3936 + 1.96 \cdot 0.0244 = 0.4414$$

Binomialverteilte Daten



Abschnitt 24: Daten aus anderen Verteilungen

Daten aus anderen Verteilungen

- Es sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe aus der Verteilung F mit Mittel μ und Varianz σ^2.
- Aufgrund des zentralen Grenzwertsatzes ist das Standardscore Z des Stichprobenmittels:

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

für große Stichproben annähernd normalverteilt. Es gilt also **näherungsweise**:

$$W(\overline{X} - z_{1-\alpha/2} S/\sqrt{n} \leq \mu \leq \overline{X} + z_{1-\alpha/2} S/\sqrt{n}) \approx 1 - \alpha$$

Näherungsweises KI für den Mittelwert

$$G_1(\boldsymbol{X}) = \overline{X} - z_{1-\alpha/2} S/\sqrt{n}, \quad G_2(\boldsymbol{X}) = \overline{X} + z_{1-\alpha/2} S/\sqrt{n}$$

Daten aus anderen Verteilungen

Beispiel

Für exponentialverteilte Stichproben vom Umfang n gibt die folgende Tabelle die Sicherheit des 95%-Konfidenzintervalls in Näherung durch Normalverteilung, geschätzt aus $N = 20000$ Stichproben:

n	25	50	100	200	400
$1-\alpha$	0.9112	0.9289	0.9408	0.9473	0.9476

Teil VI

Testen von Hypothesen

Übersicht Teil 6

Abschnitt 25: Einleitung

Einleitung

- Wir beobachten eine Stichprobe $X_1, \ldots, X_n$ aus einer Verteilung F.
- Ein Test soll feststellen, ob die Beobachtungen mit einer gewissen Annahme über F verträglich sind.
- Die Annahme wird als **Nullhypothese** H_0 bezeichnet.
- Ist die Form von F bis auf einen oder mehrere Parameter spezifiziert, heißt der Test **parametrisch**.
- Ist die Form von F nicht spezifiziert, heißt der Test **nichtparametrisch** oder **parameterfrei**.
- Der Test entscheidet, ob die Stichprobe mit der Hypothese vereinbar ist, nicht ob die Hypothese richtig ist!
- Ist die Stichprobe mit der Hypothese vereinbar, kann die Hypothese dennoch von der Wahrheit abweichen, wenn auch nur in eingeschränktem Ausmaß.

Einleitung

Allgemeine Vorgangsweise

- Aus der Stichprobe wird eine Testgröße (Teststatistik) T berechnet.
- Der Wertebereich von T wird, in Abhängigkeit von H_0, in einen **Ablehnungsbereich** (kritischen Bereich) C und einen **Annahmebereich** C' unterteilt.
- Der Annahmebereich ist meist ein **Prognoseintervall** für T.
- Fällt der Wert von T in den Ablehnungsbereich, wird H_0 verworfen.
- Andernfalls wird H_0 vorläufig beibehalten.
- Das ist jedoch keine Bestätigung von H_0. Es heißt lediglich, dass die Daten mit der Hypothese vereinbar sind.

Einleitung

Einseitige und zweiseitige Tests

- Ist der Annahmebereich das symmetrische Prognoseintervall für T, wird der Test **zweiseitig** genannt. Der kritische Bereich zerfällt dann in zwei Teilintervalle.
- Ist der Annahmebereich ein Intervall der Form $T \leq c$ oder $T \geq c$, wird der Test **einseitig** genannt. Der kritische Bereich ist dann ein Intervall der Form $T > c$ bzw. $T < c$.

Tests und Konfidenzintervalle

- In vielen Fällen ist der Annahmebereich ein Konfidenzintervall für den zu testenden Parameter.
- Die Nullhypothese wird abgelehnt, wenn der hypothetische Wert außerhalb des Konfidenzintervalls liegt, da er dann nicht genügend plausibel ist.

Einleitung

Der p-Wert

- Der Test kann alternativ auch unter Benützung des p-Werts $P(T)$ durchgeführt werden.
- Der p-Wert gibt an, wie wahrscheinlich es ist, unter Annahme der Nullhypothese mindestens den Wert T bzw. höchstens den Wert T zu beobachten.
- Zweiseitiger Test: Ist $F_0(x)$ die Verteilungsfunktion von T unter der Nullhypothese, so ist der p-Wert gleich

$$P(T) = 2\min(F_0(T), 1 - F_0(T))$$

- Einseitiger Test: Ist $F_0(x)$ die Verteilungsfunktion von T unter der Nullhypothese, so ist der p-Wert gleich

$$P(T) = F_0(T) \text{ bzw. } P(T) = 1 - F_0(T)$$

- Die Nullhypothese wird verworfen, wenn $P(T) < \alpha$.

Einleitung

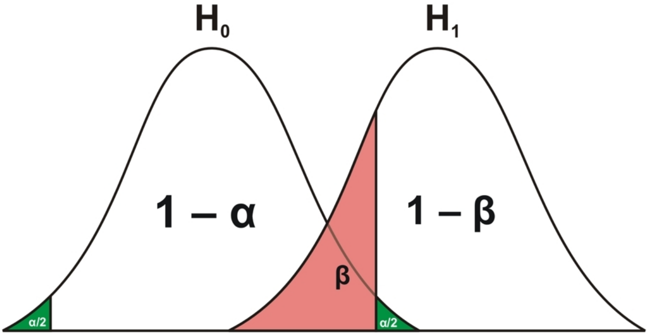
Signifikanz und Güte

- Bei jedem Testverfahren sind zwei Arten von Fehlern möglich.
  - **Fehler 1. Art**: Die Hypothese H_0 wird abgelehnt, obwohl sie zutrifft.
  - **Fehler 2. Art**: Die Hypothese H_0 wird beibehalten, obwohl sie nicht zutrifft.
- Die Verteilung von T unter Annahme von H_0 wird bestimmt.
- Der Ablehnungsbereich wird so festgelegt, dass die Wahrscheinlichkeit eines Fehlers 1. Art maximal gleich einem Wert α ist.
- α heißt das **Signifikanzniveau** des Tests. Gängige Werte sind $\alpha = 0.05, 0.01, 0.005$.

Einleitung

- Ist der Ablehnungsbereich festgelegt, kann für eine Gegenhypothese („alternative hypothesis") H_1 die Wahrscheinlichkeit $\beta(H_1)$ eines Fehlers 2. Art berechnet werden.
- $1 - \beta(H_1)$ heißt die **Güte** des Tests für H_1.
- Die Güte sollte nie kleiner als α sein.
- Ist die Güte nie kleiner als α, heißt der Test **unverzerrt**.
- Ein Ziel der Testtheorie ist es, unverzerrte Tests mit maximaler Güte (UMPU) zu konstruieren.

Einleitung



Abschnitt 26: Parametrische Tests

Unterabschnitt: Grundlagen

Grundlagen

- Wir betrachten eine Stichprobe $X_1, \ldots, X_n$ aus einer Verteilung F, die bis auf einen oder mehrere Parameter spezifiziert ist.
- Tests von Hypothesen über F heißen **parametrisch**.
- Eine Nullhypothese H_0 kann als eine Teilmenge des Parameterraums Θ aufgefasst werden.
- Der Test entscheidet, ob die Stichprobe mit der Hypothese vereinbar ist.
- Vor der Anwendung ist zu klären, ob die angenommene parametrische Form plausibel ist.

Grundlagen

- Zunächst wird die Teststatistik T und das Signifikanzniveau α gewählt.
- Dann wird der kritische Bereich C so festgelegt, dass

$$W(T \in C | \vartheta \in H_0) \leq \alpha$$

- Zu einer Nullhypothese H_0 kann eine **Gegenhypothese** H_1 formuliert werden.
- H_1 kann ebenfalls als Teilmenge des Parameterraums Θ aufgefasst werden.
- Ist das Signifikanzniveau α festgelegt, kann für jedes $\vartheta \in H_1$ die Güte berechnet werden:

$$1 - \beta(\vartheta) = W(T \in C | \vartheta \in H_1)$$

- $1 - \beta(\vartheta)$ heißt die **Gütefunktion** des Tests.

Grundlagen

Beispiel mit Exponentialverteilung

- $X_1, \ldots, X_n$ ist eine exponentialverteilte Stichprobe aus $\mathrm{Ex}(\tau)$.
- Die Hypothese $H_0 : \tau = \tau_0$ soll anhand der Stichprobe getestet werden.
- Als Teststatistik T wählen wir das Stichprobenmittel: $T = \overline{X}$.
- Unter Annahme von H_0 hat T die folgende Dichte:

$$f(t) = \frac{t^{n-1}}{(\tau_0/n)^n \Gamma(n)} \exp\left(-\frac{t}{\tau_0/n}\right)$$

- T ist also verteilt gemäß $\mathrm{Ga}(n, \tau_0/n)$.
- Das symmetrische Prognoseintervall $[y_1(\tau_0), y_2(\tau_0)]$ für T zum Niveau $1-\alpha$ erhält man mit:

$$y_1(\tau_0) = \gamma_{\alpha/2;n,\tau_0/n}, \quad y_2(\tau_0) = \gamma_{1-\alpha/2;n,\tau_0/n}$$

Grundlagen

- Der Verwerfungsbereich mit Signifikanzniveau α ist daher die Menge

$$C = [0, y_1(\tau_0)] \cup [y_2(\tau_0), \infty[$$

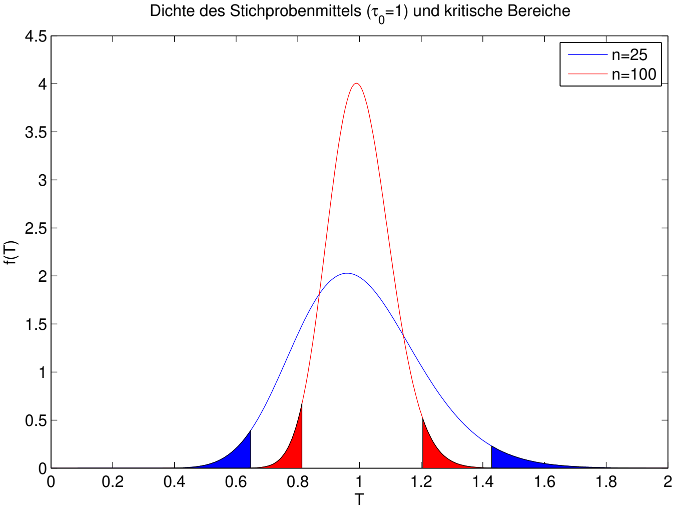
- H_0 wird also abgelehnt, wenn T "weit entfernt" vom hypothetischen Wert τ_0 ist.
- Die Gütefunktion für einen Wert τ ergibt sich durch:

$$1 - \beta(\tau) = W(T \in C) = G(y_1(\tau)) + 1 - G(y_2(\tau))$$

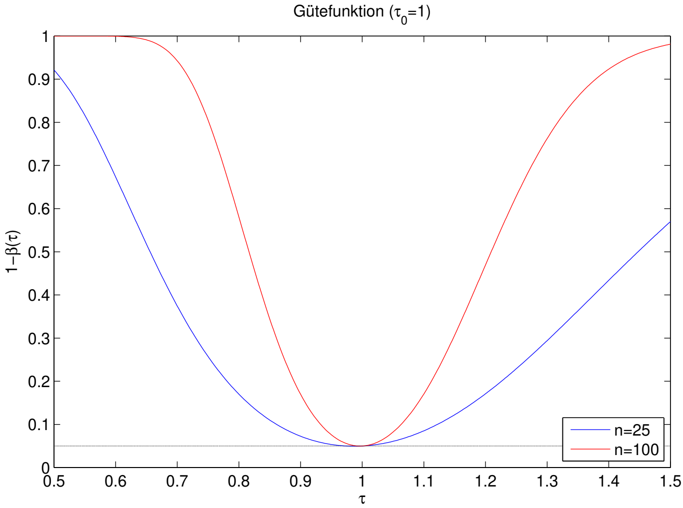
  wo G die Verteilungsfunktion der $\mathrm{Ga}(n, \tau/n)$-Verteilung ist.
- Der Test ist nicht unverzerrt, da z.B. für $\tau_0 = 1$ und $n = 25$

$$1 - \beta(0.986) = 0.0495 < \alpha$$

Grundlagen


Dichte des Stichprobenmittels (τ_0=1) und kritische Bereiche
n=25
n=100
f(T)
T
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0.2
0.4
0.6
0.8
1.2
1.4
1.6
1.8
2

Grundlagen


Gütefunktion (τ0=1)
1−β(τ)
τ
n=25
n=100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5

Unterabschnitt: Tests für normalverteilte Daten

Erwartungswert bei bekannter Varianz

- $X_1, \ldots, X_n$ ist eine normalverteilte Stichprobe aus $\mathrm{No}(\mu, \sigma^2)$ mit bekanntem σ^2.
- Die Hypothese $H_0 : \mu = \mu_0$ soll anhand der Stichprobe gegen die Gegenhypothese $H_1 : \mu \neq \mu_0$ getestet werden.
- Als Teststatistik T wählen wir das Standardscore des Stichprobenmittels:

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$$

- Unter Annahme von H_0 ist $T \sim \mathrm{No}(0, 1)$.
- H_0 wird abgelehnt, wenn T nicht in einem Prognoseintervall vom Niveau $1 - \alpha$ der Standardnormalverteilung liegt.

Tests für normalverteilte Daten

Zweiseitiger Test

- Die Hypothese H_0 wird abgelehnt, wenn

$$|T| = \frac{\sqrt{n}\left|\overline{X} - \mu_0\right|}{\sigma} > z_{1-\alpha/2}$$

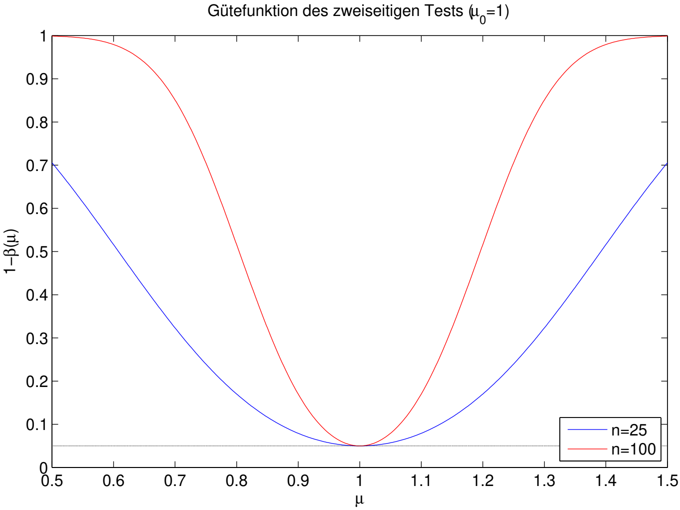
- Die Gütefunktion für einen Wert μ ergibt sich durch:

$$1 - \beta(\mu) = W(T \in C) = G(z_{\alpha/2}) + 1 - G(z_{(1-\alpha)/2})$$

  wo G die Verteilungsfunktion der $\mathrm{No}(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$-Verteilung ist.

- Der Test ist unverzerrt.

Tests für normalverteilte Daten



Tests für normalverteilte Daten

Einseitiger Test

- Die Hypothese $H_0 : \mu \leq \mu_0$ soll mit der Teststatistik T gegen die Gegenhypothese $H_1 : \mu > \mu_0$ getestet werden.
- H_0 wird abgelehnt, wenn T „zu groß“ ist.
- Ein Verwerfungsbereich mit Signifikanzniveau α ist die Menge

$$C = [z_{1-\alpha}, \infty[$$

- Die Hypothese H_0 wird also abgelehnt, wenn

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{\sigma} > z_{1-\alpha}$$

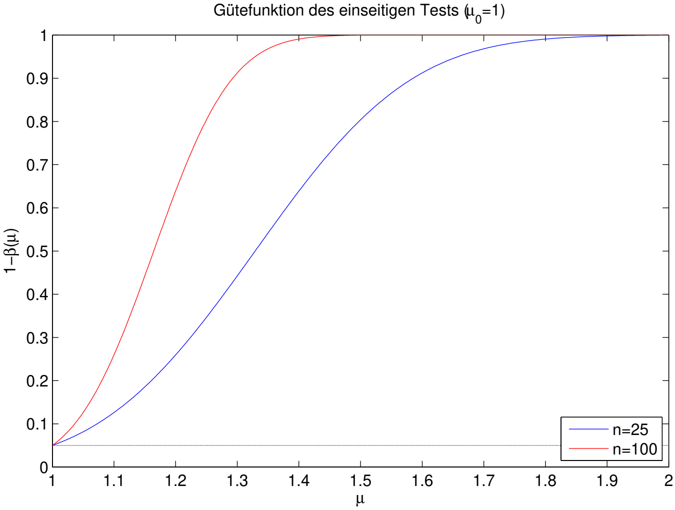
Tests für normalverteilte Daten

- Die Gütefunktion für einen Wert $\mu > \mu_0$ ergibt sich durch:

$$1 - \beta(\mu) = W(T \in C) = 1 - G(z_{1-\alpha})$$

wo G die Verteilungsfunktion der $\mathrm{No}(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$-Verteilung ist.

- Analog verläuft der Test mit $H_0 : \mu \geq \mu_0$ und $H_1 : \mu < \mu_0$.


Gütefunktion des einseitigen Tests (μ0=1)
1−β(μ)
μ
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
n=25
n=100

Erwartungswert bei unbekannter Varianz: t-Test

- $X_1, \ldots, X_n$ ist eine normalverteilte Stichprobe aus $\mathrm{No}(\mu, \sigma^2)$ mit unbekanntem σ^2.
- Die Hypothese $H_0 : \mu = \mu_0$ soll anhand der Stichprobe gegen die Gegenhypothese $H_1 : \mu \neq \mu_0$ getestet werden.
- Als Teststatistik T wählen wir das Standardscore des Stichprobenmittels, unter Benützung der Stichprobenvarianz S^2:

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$$

- Unter Annahme von H_0 ist $T \sim \mathrm{t}(n-1)$.

Tests für normalverteilte Daten

- H_0 wird abgelehnt, wenn T nicht in einem Prognoseintervall vom Niveau $1-\alpha$ der t-Verteilung mit $n-1$ Freiheitsgraden liegt.
- Ein Verwerfungsbereich mit Signifikanzniveau α ist die Menge

$$C =]-\infty, t_{\alpha/2;n-1}] \cup [t_{1-\alpha/2;n-1}, \infty[$$

  wo $t_{p;n}$ das Quantil der t-Verteilung mit $n-1$ Freiheitsgraden zum Niveau p ist.
- Die Hypothese H_0 wird also abgelehnt, wenn

$$|T| = \frac{\sqrt{n}\,\left|\overline{X}-\mu_0\right|}{S} > t_{1-\alpha/2;n-1}$$

Tests für normalverteilte Daten

- Die Gütefunktion für einen Wert μ ergibt sich durch:

$$1 - \beta(\mu) = W(T \in C) = G(z_{\alpha/2}) + 1 - G(z_{(1-\alpha)/2})$$

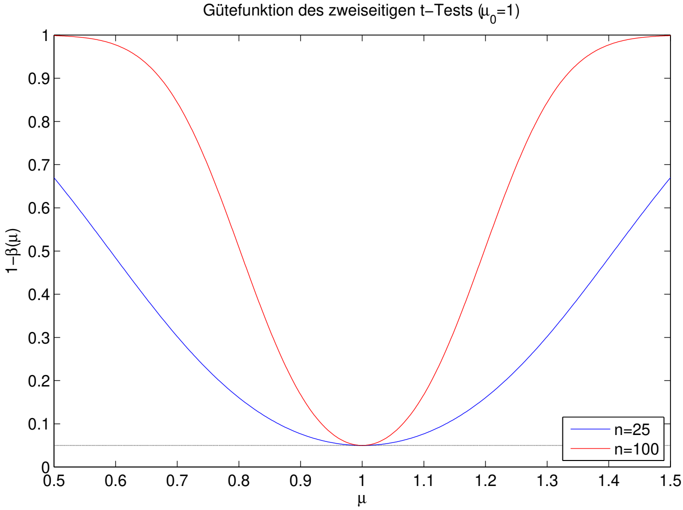
wo G die Verteilungsfunktion der nichtzentralen $\mathrm{t}(n-1, \delta)$-Verteilung mit

$$\delta = \sqrt{n}(\mu - \mu_0)/\sigma$$

ist.

- Der Test ist unverzerrt.

Tests für normalverteilte Daten



Tests für normalverteilte Daten

Gleichheit von zwei Erwartungswerten

- $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ sind zwei unabhängige normalverteilte Stichproben aus $\mathrm{No}(\mu_x, \sigma_x^2)$ bzw. $\mathrm{No}(\mu_y, \sigma_y^2)$.
- Die Hypothese $H_0 : \mu_x = \mu_y$ soll anhand der Stichproben gegen die Gegenhypothese $H_1 : \mu_x \neq \mu_y$ getestet werden.
- Sind die **Varianzen bekannt**, wählen wir als Teststatistik T die Differenz der Stichprobenmittel:

$$T = \overline{X} - \overline{Y}$$

- Unter Annahme von H_0 ist $T \sim \mathrm{No}(0, \sigma_x^2/n + \sigma_y^2/m)$.

Tests für normalverteilte Daten

- Das Standardscore

$$Z = \frac{T}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

ist dann standardnormalverteilt.

- Die Hypothese H_0 wird also abgelehnt, wenn

$$|Z| > z_{1-\alpha/2}$$

oder

$$\frac{|\overline{X} - \overline{Y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} > z_{1-\alpha/2}$$

Tests für normalverteilte Daten

- Sind die **Varianzen unbekannt und gleich**, kann die Varianz aus der kombinierten („gepoolten") Stichprobe geschätzt werden:

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

- Unter Annahme von H_0 ist

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2(1/n + 1/m)}}$$

  t-verteilt mit $n + m - 2$ Freiheitsgraden.

- Die Hypothese H_0 wird also abgelehnt, wenn

$$|T| > t_{1-\alpha/2;n+m-2}$$

  wo $t_{1-\alpha/2;n+m-2}$ das Quantil der t-Verteilung mit $n + m - 2$ Freiheitsgraden ist.

t-Test für gepaarte Stichproben

- Gepaarte Stichproben $(X_1, Y_1), \ldots, (X_n, Y_n)$ entstehen, wenn für jedes beobachtete Objekt die selbe Größe zweimal gemessen wird, vor und nach einer bestimmten Intervention.
- Die Wirkung der Intervention wird durch die Differenzen $W_i = Y_i - X_i, i = 1, \ldots, n$ beschrieben.
- Wir nehmen an, dass $W_1, \ldots, W_n$ normalverteilt mit Mittel μ_w und unbekannter Varianz σ_w^2 ist.
- Die Hypothese $H_0 : \mu_w = 0$ (keine Wirkung der Intervention) soll anhand der Stichprobe gegen die Gegenhypothese $H_1 : \mu_w \neq 0$ getestet werden.
- Dies erfolgt mit dem t-Test für einzelne Stichproben.

Tests für normalverteilte Daten

Test der Varianz

- $X_1, \ldots, X_n$ ist eine normalverteilte Stichprobe mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2.
- Die Hypothese $H_0 : \sigma^2 = \sigma_0^2$ soll anhand der Stichprobe gegen die Gegenhypothese $H_1 : \sigma^2 \neq \sigma_0^2$ getestet werden.
- Als Teststatistik T wählen wir:

$$T = \frac{(n-1)S^2}{\sigma_0^2}$$

- Unter Annahme von H_0 ist T χ^2-verteilt mit $n-1$ Freiheitsgraden.

Tests für normalverteilte Daten

- Die Hypothese H_0 wird also abgelehnt, wenn

$$T < \chi^2_{\alpha/2;n-1} \quad \text{oder} \quad T > \chi^2_{1-\alpha/2;n-1}$$

  wo $\chi^2_{p;k}$ das p-Quantil der χ^2-Verteilung mit k Freiheitsgraden ist.

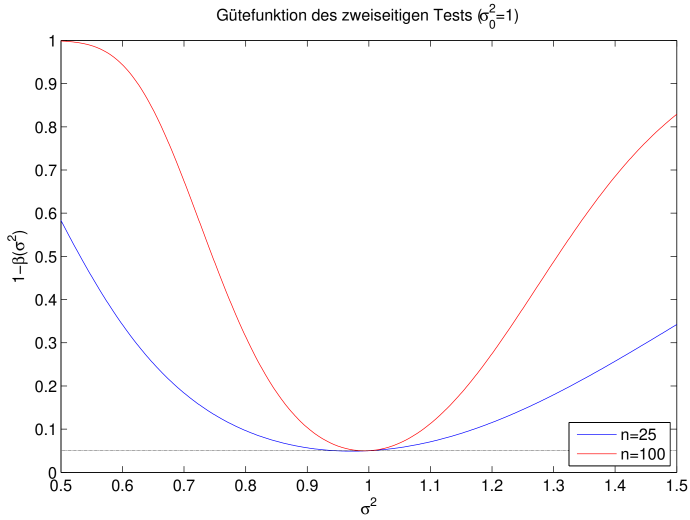
- Die Gütefunktion für einen Wert σ^2 ergibt sich durch:

$$1 - \beta(\sigma^2) = G(\sigma_0^2/\sigma^2 \cdot \chi^2_{\alpha/2;n-1}) + 1 - G(\sigma_0^2/\sigma^2 \cdot \chi^2_{1-\alpha/2;n-1})$$

  wo G die Verteilungsfunktion der $\chi^2(n-1)$-Verteilung ist.

- Der Test ist nicht unverzerrt.

Tests für normalverteilte Daten



Unterabschnitt: Tests für Poisson-verteilte Daten

Tests für Poisson-verteilte Daten

Zweiseitiger Test auf den Erwartungswert

- Es sei $X_1, \ldots, X_n$ eine Poisson-verteilte Stichprobe aus $\mathrm{Po}(\lambda)$.
- Die Hypothese $H_0 : \lambda = \lambda_0$ soll anhand der Stichprobe gegen die Gegenhypothese $H_1 : \lambda \neq \lambda_0$ getestet werden.
- Als Teststatistik T wählen wir die Stichprobensumme:

$$T = \sum_{i=1}^{n} X_i$$

- T ist Poisson-verteilt gemäß $\mathrm{Po}(n\lambda)$.
- H_0 wird abgelehnt, wenn T „zu klein“ oder „zu groß“ ist, also wenn

$$\sum_{k=0}^{T} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha/2 \text{ oder } \sum_{k=T}^{\infty} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha/2$$

Tests für Poisson-verteilte Daten

- Wird die Verteilungsfunktion der Poisson-Verteilung durch die Verteilungsfunktion der Gammaverteilung ausgedrückt, ergibt sich der folgende kritische Bereich:

$$n\lambda_0 < \gamma_{\alpha/2;T,1} \quad \text{oder} \quad n\lambda_0 > \gamma_{1-\alpha/2;T+1,1}$$

- H_0 wird also abgelehnt, wenn $n\lambda_0$ nicht im Konfidenzintervall liegt, das auf der Beobachtung T basiert.

Tests für Poisson-verteilte Daten

Einseitiger Test auf den Erwartungswert

- Die Hypothese $H_0 : \lambda \leq \lambda_0$ wird abgelehnt, wenn T „zu groß" ist und damit der p-Wert zu klein:

$$P(T) = \sum_{k=T}^{\infty} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha \quad \text{oder} \quad n\lambda_0 < \gamma_{\alpha;T,1}$$

- Die Hypothese $H_0 : \lambda \geq \lambda_0$ wird abgelehnt, wenn T „zu klein" ist und damit auch der p-Wert zu klein:

$$P(T) = \sum_{k=0}^{T} \frac{(n\lambda_0)^k e^{-n\lambda_0}}{k!} < \alpha \quad \text{oder} \quad n\lambda_0 > \gamma_{1-\alpha;T+1,1}$$

Tests für Poisson-verteilte Daten

Beispiel

Ein Hersteller strebt an, dass in einer Fabrik täglich im Mittel nicht mehr als 25 defekte Bauteile hergestellt werden. Eine Stichprobe von 5 Tagen ergibt 28,34,32,38 und 22 defekte Bauteile. Hat der Hersteller sein Ziel erreicht?

Die Testgröße ist gleich $T = 154$. Es gilt also:

$$P(T) = \sum_{k=T}^{\infty} \frac{(125)^k e^{-125}}{k!} = 0.0067 < 0.01$$

$$\gamma_{0.01;154,1} = 126.61 < 154$$

Die Hypothese lässt sich also auf einem Signifikanzniveau von 1 Prozent widerlegen.

Tests für Poisson-verteilte Daten

Näherung durch Normalverteilung

- Ist n genügend groß, kann die Verteilung von T durch eine Normalverteilung $\mathrm{No}(n\lambda, n\lambda)$ angenähert werden.
- H_0 wird abgelehnt, wenn das Standardscore

$$Z = \frac{T - n\lambda_0}{\sqrt{n\lambda_0}}$$

nicht in einem Prognoseintervall vom Niveau $1 - \alpha$ der Standardnormalverteilung liegt.

Beispiel

Mit der Angabe des letzten Beispiels ergibt die Näherung:

$$Z = 2.5938 > z_{0.99} = 2.3263$$

Die Hypothese kann also auf einem Signifikanzniveau von 1 Prozent abgelehnt werden.

Unterabschnitt: Tests für binomialverteilte Daten

Tests für binomialverteilte Daten

Zweiseitiger Test für den Parameter p

- k ist eine Beobachtung aus der Binomialverteilung $\mathrm{Bi}(n, p)$.
- Die Hypothese $H_0 : p = p_0$ soll anhand der Beobachtung gegen die Gegenhypothese $H_1 : p \neq p_0$ getestet werden.
- H_0 wird abgelehnt, wenn k unter Annahme von H_0 nicht im symmetrischen Prognoseintervall $[y_1(p_0), y_2(p_0)]$ liegt, also „zu klein“ oder „zu groß“ ist.

Tests für binomialverteilte Daten

- Das ist der Fall, wenn entweder

$$\sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i} = F_{\mathrm{Be}}(p_0; k, n-k+1) < \alpha/2$$

oder

$$\sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} = F_{\mathrm{Be}}(1-p_0; n-k, k+1) < \alpha/2$$

gilt, wo $F_{\mathrm{Be}}(x; a, b)$ die Verteilungsfunktion von $\mathrm{Be}(a, b)$ ist.

Einseitiger Test für den Parameter p

- Die Hypothese $H_0 : p \leq p_0$ soll anhand der Beobachtung k gegen die Gegenhypothese $H_1 : p > p_0$ getestet werden.
- H_0 wird abgelehnt, wenn k „zu groß" ist und damit der p-Wert zu klein:

$$P(k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i} = B(p_0; k, n-k+1) < \alpha$$

- Die Hypothese $H_0 : p \geq p_0$ wird abgelehnt, wenn k „zu klein" ist und damit auch der p-Wert zu klein:

$$P(k) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1 - p_0)^{n-i} = B(1 - p_0; n-k, k+1) < \alpha$$

Tests für binomialverteilte Daten

Beispiel

Ein Hersteller behauptet, dass nicht mehr als 2 Prozent eines gewissen Bauteils fehlerhaft sind. In einer Stichprobe vom Umfang 300 sind 9 Stück defekt. Kann die Behauptung des Herstellers widerlegt werden?

Es gilt:

$$P(k) = \sum_{i=9}^{300} \binom{300}{i} 0.02^i 0.98^{300-i} = 0.1507$$

Die Behauptung des Herstellers lässt sich also auf einem Signifikanzniveau von 5 Prozent nicht widerlegen.

Tests für binomialverteilte Daten

Näherung durch Normalverteilung

- Ist n genügend groß, kann die Verteilung von k durch eine Normalverteilung $\mathrm{No}(np, np(1-p))$ angenähert werden.
- H_0 wird abgelehnt, wenn das Standardscore

$$Z = \frac{k - np_0}{\sqrt{np_0(1-p_0)}}$$

nicht in einem Prognoseintervall vom Niveau $1-\alpha$ der Standardnormalverteilung liegt.
- Zweiseitiger Test: H_0 wird abgelehnt wenn

$$Z < z_{\alpha/2} \text{ oder } Z > z_{1-\alpha/2}$$

- Einseitiger Test: H_0 wird abgelehnt wenn

$$Z < z_{\alpha} \text{ bzw. } Z > z_{1-\alpha}$$

Tests für binomialverteilte Daten

Beispiel

Mit der Angabe des letzten Beispiels ergibt die Näherung:

$$Z = 1.2372 < z_{0.95} = 1.6449$$

Die Hypothese kann also nicht abgelehnt werden.

Unterabschnitt: Likelihood-Ratio Tests

Likelihood-Ratio Tests

- Eine Hypothese heißt **einfach**, wenn sie keine freien Parameter hat, d.h. wenn das statistische Modell vollständig gegeben ist.
- Wenn zwei *einfache* Hypothesen $H_0 : \theta = \theta_0$, und $H_1 : \theta = \theta_1$, gegeben sind, kann ein Test aus dem Verhältnis ("ratio") der Likelihoods konstruiert werden.

$$\Lambda(x) := \frac{L(\theta_0|x)}{L(\theta_1|x)}$$

- Die Nullhypothese H_0 wird abgelehnt wenn die Teststatistik unterhalb eines bestimmten Wertes $\Lambda(X) \leq \eta$ fällt.
- Hierbei wird η so gewählt dass $P(\Lambda(X) \leq \eta | H_0) \leq \alpha$. Wie üblich wird das Signifikanzniveau α wiederum frei gewählt. Übliche Werte sind $\alpha = 0.05, 0.1$ oder 0.01.

Likelihood-Ratio Tests

Beispiel (Likelihood-Ratio Test)

Sei $(X_1, \ldots, X_n)$ eine Stichprobe einer Normalverteilung $\mathrm{N}(\mu, \sigma^2)$ mit festem μ. Wir testen auf die *Varianz* der Verteilung, mit $H_0 : \sigma^2 = \sigma_0^2$ und $H_1 : \sigma^2 = \sigma_1^2, \sigma_1^2 > \sigma_0^2$.

- Die Likelihood-Ratio ist

$$\Lambda(X) = \frac{L(\sigma_0^2|X)}{L(\sigma_1^2|X)} = \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{-\frac{n}{2}} \exp\left\{-\frac{\sigma_0^{-2} - \sigma_1^{-2}}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right\}$$

- $\sum_{i=1}^{n} (x_i - \mu)^2$ ist der einzige Term der Likelihood-Ratio der von den Daten abhängig ist, deswegen wird H_0 abgelehnt wenn die Summe extrem genug unter der Nullhypothese ist. Da wir verlangten dass $\sigma_1^2 > \sigma_0^2$ ist, können wir uns darauf beschränken zu testen, ob die Summe zu groß ist (einseitiger Test).

Likelihood-Ratio Tests

Neyman-Pearson Lemma

Der Likelihood-Ratio Test für zwei einfache Hypothesen ist ein Test mit maximaler Güte bei gegebenem Signifikanzniveau α.

- Eine Hypothese die nicht *einfach* ist, heißt **zusammengesetzt**.
- Die Gegenhypothese H_1 wird zumeist wieder als die Komplementärmenge der Nullhypothese konstruiert, d.h. $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_0'$.
- Man spricht man von einem **verschachtelten Modell**. In diesem Fall gilt mit $\Theta = \Theta_0 \cup \Theta_0'$ das Neyman-Pearson Lemma für:

$$\Lambda(x) = \frac{\sup\{L(\theta|x) : \theta \in \Theta_0\}}{\sup\{L(\theta|x) : \theta \in \Theta\}}$$

Likelihood-Ratio Tests

Es gilt zudem auch das:

Wilks' Theorem

Falls kein Randextremum vorliegt, konvergiert die Teststatistik $T = -2 \ln \Lambda(X)$ für verschachtelte Modelle unter H_0 asymptotisch gegen eine $\chi^2(n)$ Verteilung mit n Freiheitsgraden, wobei n gegeben ist als die Differenz der Dimensionalitäten von Θ and Θ_0: $n = \dim(\Theta) - \dim(\Theta_0)$.

- Wilks' Theorem kann benutzt werden um η festzulegen: $\eta = \chi^2_{\alpha;n}$, wobei α wiederum das Signifikanzniveau des Testes ist und n die Differenz der Anzahl der freien Parameter der Hypothesen ist (siehe oben).

Likelihood-Ratio Tests

Beispiel (Likelihood-Ratio Test für $\mathrm{Po}(\lambda)$)

Wir konstruieren einen Test für den Intensitätsparameter λ einer Poissonverteilung, $H_0 : \lambda = \lambda_0, H_1 : \lambda \neq \lambda_0$. Sei $(X_1, \ldots, X_n)$ eine Stichprobe einer poisson-verteilten Zufallsvariablen.

- Das Supremum der Gegenhypothese H_1 (die ,,Maximum Likelihood") wird erreicht bei $\lambda = \overline{X}$.
- Daraus folgt für die Likelihood-Ratio $\Lambda(X)$:

$$\Lambda(X) = e^{n(\overline{X}-\lambda_0)} \prod_{i=1}^{n} \left(\frac{\lambda_0}{\overline{X}}\right)^{X_i}$$

- Die Teststatistik $T = -2 \ln \Lambda(X)$:

$$T = -2n \left[\overline{X} - \lambda_0 + \overline{X} \ln\left(\frac{\lambda_0}{\overline{X}}\right)\right]$$

Likelihood-Ratio Tests

Beispiel (Likelihood-Ratio Test für $\mathrm{Po}(\lambda)$, Fortsetzung)

- Die Teststatistik $T = -2\ln\Lambda(X)$:

$$T = 2n\left[\lambda_0 - \overline{X} + \overline{X}\ln\left(\frac{\overline{X}}{\lambda_0}\right)\right]$$

- Nach Wilks' Theorem ist T asymptotisch χ^2-verteilt mit einem Freiheitsgrad. Das 95% Quantil von $\chi^2(1)$ liegt bei 3.84. Die kritische Region des Tests bei einem Signifikanzniveau von 5% ist daher

$$C = [3.84, \infty)$$

.

Abschnitt 27: Anpassungstests

Anpassungstests

- Ein Test, der die Hypothese überprüft, ob die Daten einer gewissen Verteilung entstammen können, heißt ein **Anpassungstest**.
- Die Verteilung kann völlig oder bis auf unbekannte Parameter bestimmt sein.
- Ein Anpassungstest kann einem parametrischen Test vorausgehen, um dessen Anwendbarkeit zu überprüfen.

Unterabschnitt: Der Chiquadrat-Test

Der Chiquadrat-Test für diskrete Daten

- Die Stichprobe $X_1, \ldots, X_n$ entstammt einer diskreten Verteilung mit Wertebereich $\{1, \ldots, k\}$.
- Wir testen die Hypothese H_0, dass die Dichte f die Werte $f(j) = p_j, j = 1, \ldots, k$ hat:

  $$H_0 : W(X_i = j) = p_j, j = 1, \ldots, k$$

  gegen

  $$H_1 : W(X_i = j) \neq p_j, \text{für ein } j$$

- Es sei Y_j die Zahl der Beobachtungen, die gleich j sind.
- Unter der Nullhypothese ist $Y_1, \ldots, Y_k$ multinomial verteilt gemäß $\mathrm{Mu}(n, p_1, \ldots, p_k)$ und $\mathsf{E}[Y_j] = np_j$.

Der Chiquadrat-Test

Exkurs: Die Multinomialverteilung $\mathrm{Mu}(n, p_1, \ldots, p_d)$

- Der Alternativversuch kann dahingehend verallgemeinert werden, dass man nicht nur zwei, sondern d Elementarereignisse $e_1, \ldots, e_d$ zulässt, denen die Wahrscheinlichkeiten $p_1, \ldots, p_d$ zugeordnet werden, die nur

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

erfüllen müssen.

- Führt man den **verallgemeinerten Alternativversuch** n-mal durch, so sind die Elementarereignisse die Folgen der Form:

$$(e_{i_1}, \ldots, e_{i_n}),\ 1 \leq i_j \leq d$$

Der Chiquadrat-Test

- Sind die n Teilversuche unabhängig, gilt:

$$W(e_{i_1}, \ldots, e_{i_n}) = \prod_{j=1}^{n} W(e_{i_j}) = \prod_{j=1}^{n} p_{i_j} = \prod_{i=1}^{d} p_i^{n_i}$$

  wobei n_i die Anzahl des Eintretens von e_i ist. Die Summe der n_i ist daher n.

- Die d-dimensionale Zufallsvariable $\boldsymbol{X} = (X_1, \ldots, X_d)$ bildet die Folge $(e_{i_1}, \ldots, e_{i_n})$ auf den Vektor $(n_1, \ldots, n_d)$ ab. Dabei werden $n!/(n_1! \cdots n_d!)$ Folgen auf den gleichen Vektor abgebildet.
- Die Dichte von $\boldsymbol{X}$ lautet daher:

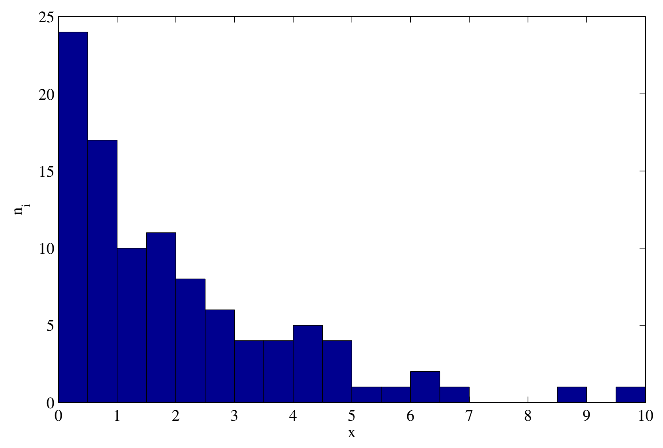
$$f_{\boldsymbol{X}}(n_1, \ldots, n_d) = \frac{n!}{n_1! \ldots n_d!} \prod_{i=1}^{d} p_i^{n_i}, \; \sum_{i=1}^{d} n_i = n, \; \sum_{i=1}^{d} p_i = 1$$

Der Chiquadrat-Test

- Die Verteilung von $\boldsymbol{X}$ wird als **Multinomialverteilung** mit den Parametern n und $p_1, \ldots, p_d$ bezeichnet: $W_{\boldsymbol{X}} = \mathrm{Mu}(n, p_1, \ldots, p_d)$
- Das klassische Beispiel eines multinomialverteilten Zufallsvektors ist das Histogramm (gruppierte Häufigkeitsverteilung), das zur graphischen Darstellung der (absoluten) **experimentellen Häufigkeit** verwendet wird.
- X_i ist die Anzahl der Fälle, in denen die Zufallsvariable R, das experimentelle Ergebnis, in Gruppe i fällt.
- Die Wahrscheinlichkeit, dass R in Gruppe i fällt, sei gleich p_i.
- Werden in das Histogramm n Ergebnisse eingefüllt, so sind die Gruppeninhalte $(X_1, \ldots, X_d)$ multinomial nach $\mathrm{Mu}(n, p_1, \ldots, p_d)$ verteilt.

Der Chiquadrat-Test

- Ein Histogramm



Der Chiquadrat-Test

Die Momente der $\mathrm{Mu}(n, p_1, \ldots, p_d)$-Verteilung

Es sei $\boldsymbol{X} = (X_1, \ldots, X_d) \sim \mathrm{Mu}(n, p_1, \ldots, p_d)$. Dann gilt:

- $\mathsf{E}[X_i] = np_i$
- $\mathsf{var}[X_i] = np_i(1 - p_i)$
- $\mathsf{cov}[X_i, X_j] = -np_ip_j$

Definition (Kovarianz)

Die **Kovarianz** von zwei Zufallsvariablen X und Y ist definiert durch:

$$\mathsf{cov}[X, Y] = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$$

Ist $\mathsf{cov}[X, Y] = 0$, heißen X und Y **unkorreliert**.

Der Chiquadrat-Test

Eigenschaften der Kovarianz

- $\mathrm{cov}[aX + b, cY + d] = ac \cdot \mathrm{cov}[X, Y]$
- $\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\mathrm{cov}[X, Y]$
- $\mathrm{cov}[X, X] = \mathrm{var}[X]$

Definition (Korrelation)

Der **Korrelationskoeffizient** $\rho[X, Y]$ ist definiert durch:

$$\rho[X, Y] = \frac{\mathrm{cov}[X, Y]}{\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}}$$

Der Chiquadrat-Test

Eigenschaften des Korrelationskoeffizienten

- $-1 \leq \rho[X, Y] \leq 1$
- $\rho[aX + b, cY + d] = \rho[X, Y]$
- $\rho[X, X] = 1$

Der Chiquadrat-Test

Definition (Kovarianzmatrix)

Es sei $\boldsymbol{X} = (X_1, \ldots, X_d)$ ein Zufallsvektor. Existieren alle Varianzen und Kovarianzen, werden sie in der **Kovarianzmatrix** $\mathbf{C} = \mathsf{Cov}[\boldsymbol{X}]$ zusammengefasst:

$$\mathbf{C}_{ij} = \mathsf{cov}[X_i, X_j]$$

Analog dazu werden die Korrelationen in der **Korrelationsmatrix** $\mathbf{R} = \mathsf{Cor}[X, Y]$ zusammengefasst:

$$\mathbf{R}_{ij} = \rho[X_i, X_j]$$

- Die Kovarianzmatrix ist stets **symmetrisch** und **positiv definit**. Alle Eigenwerte sind reell und positiv.
- Die Korrelationsmatrix ist ebenfalls **symmetrisch** und **positiv definit**. Alle Diagonalelemente sind gleich 1.

Der Chiquadrat-Test

- Die Testgröße vergleicht die beobachteten Häufigkeiten Y_j mit ihren Erwartungswerten:

$$T = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$$

- Die Nullhypothese wird verworfen, wenn T groß ist.
- Der kritische Bereich kann nach dem folgenden Ergebnis bestimmt werden.

Satz

Unter Annahme der Nullhypothese ist die Zufallsvariable T asymptotisch, d.h. für $n \to \infty$, χ^2-verteilt mit $k-1$ Freiheitsgraden.

Der Chiquadrat-Test

- Soll der Test Signifikanzniveau α haben, wird H_0 abgelehnt, wenn

$$T \geq \chi^2_{1-\alpha;k-1}$$

  wo $\chi^2_{1-\alpha;k}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $k-1$ Freiheitsgraden ist.
- Der Grund dafür, dass T nur $k-1$ Freiheitsgrade hat, ist der lineare Zusammenhang zwischen den Y_j:

$$\sum_{j=1}^{k} Y_j = n$$

- Als Faustregel gilt: n sollte so groß sein, dass $np_j > 5, j = 1, \ldots, k$.
- Ist das nicht erfüllt, sollte der Ablehnungsbereich durch Simulation bestimmt werden.

Der Chiquadrat-Test

Beispiel

Wir testen anhand einer Stichprobe vom Umfang 50, ob ein Würfel symmetrisch ist, d.h. ob die Augenzahl X folgende Verteilung hat:

$$W(X=1)=\ldots=W(X=6)=\frac{1}{6}$$

Eine Simulation von $N=100000$ Stichproben ergibt:

$$\overline{T}=5.000, \quad S_T^2=9.789$$

Das 0.95-Quantil der χ^2-Verteilung mit fünf Freiheitsgraden ist $\chi^2_{0.95;5}=11.07$, und

$$W(T\geq 11.07)=0.048$$

Der Chiquadrat-Test

Der Chiquadrat-Test für stetige Daten

- Die Stichprobe $X_1, \ldots, X_n$ entstammt einer stetigen Verteilung F.
- Wir testen die Hypothese $H_0 : F(x) = F_0(x)$.
- Dazu wird der Wertebereich von X in k Gruppen $G_1, \ldots, G_k$ eingeteilt.
- Es sei Y_j die Zahl der Beobachtungen in Gruppe G_j.
- Unter der Nullhypothese ist $Y_1, \ldots, Y_k$ multinomial verteilt gemäß $\mathrm{Mu}(n, p_1, \ldots, p_k)$ und $\mathsf{E}[Y_j] = np_j$, mit

$$p_j = W(X \in G_j | H_0)$$

- Der Test verläuft weiter wie im diskreten Fall.

Der Chiquadrat-Test

Unbekannte Parameter

- Die Nullhypothese muss nicht vollständig spezifiziert sein. Wir betrachten den Fall, dass die p_j noch von unbekannten Parametern $\boldsymbol{\vartheta}$ abhängen:

$$W(X \in G_j) = p_j(\boldsymbol{\vartheta})$$

- Die Statistik T ist nun eine Funktion der unbekannten Parameter:

$$T(\boldsymbol{\vartheta}) = \sum_{j=1}^{k} \frac{(Y_j - np_j(\boldsymbol{\vartheta}))^2}{np_j(\boldsymbol{\vartheta})}$$

- Zunächst werden die Parameter geschätzt, durch ML-Schätzung oder Minimierung von T:

$$\tilde{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} T(\boldsymbol{\vartheta})$$

Der Chiquadrat-Test

- Der kritische Bereich kann nach dem folgenden Ergebnis bestimmt werden.

Satz

Werden m Parameter aus der Stichprobe geschätzt, so ist $T(\tilde{\boldsymbol{\vartheta}})$ asymptotisch χ^2-verteilt mit $k-1-m$ Freiheitsgraden.

- Soll der Test Signifikanzniveau α haben, wird H_0 abgelehnt, wenn

$$T \geq \chi^2_{1-\alpha;k-1-m}$$

wo $\chi^2_{1-\alpha;k-1-m}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $k-1-m$ Freiheitsgraden ist.

Der Chiquadrat-Test

Beispiel

Angabe: Die Zahl der Arbeitsunfälle wurde in einem großen Betrieb über 30 Wochen erhoben. Es ergaben sich folgende Werte:

$$\boldsymbol{X} = \{8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0, 1, 9, 3, 4, 5, 3, 3, 4, 7, 4, 0, 1, 2, 1, 2\}$$

Es soll die Hypothese überprüft werden, dass die Beobachtungen Poisson-verteilt gemäß $\mathrm{Po}(\lambda)$ sind.

Lösung: Die Beobachtungen werden in fünf Gruppen eingeteilt:

Gruppe	1	2	3	4	5
X	0	1	2–3	4–5	> 5

Die Häufigkeiten der Gruppen sind:

$$Y_1 = 6, Y_2 = 5, Y_3 = 8, Y_4 = 6, Y_5 = 5$$

Der Chiquadrat-Test

Beispiel (Fortsetzung)

Der Schätzwert für λ ist das Stichprobenmittel:

$$\tilde{\lambda} = 3.1667$$

Die Erwartungswerte der Y_j unter Annahme von $H_0 = \mathrm{Po}(\tilde{\lambda})$ sind:

j	1	2	3	4	5
$\mathsf{E}[Y_1]$	1.2643	4.0037	13.0304	8.6522	3.0493

Die Testgröße T ist gleich $T = 21.99$. Das 99%-Quantil der χ^2-Verteilung mit drei Freiheitsgraden ist gleich $\chi^2_{0.99;3} = 11.35$. Die Hypothese, dass die Beobachtungen Poisson-verteilt sind, ist also abzulehnen.

Unterabschnitt: Der Kolmogorov-Smirnov-Test

Der Kolmogorov-Smirnov-Test

Eine Stichprobe

- Die Stichprobe $X_1, \ldots, X_n$ ist aus der stetigen Verteilung mit Verteilungsfunktion F.
- Wir testen die Hypothese $H_0 : F(x) = F_0(x)$.
- Die Testgröße D_n ist die maximale absolute Abweichung der empirischen Verteilungsfunktion $F_n(x)$ der Stichprobe von der hypothetischen Verteilungsfunktion $F_0(x)$:

$$D_n = \max_x |F_n(x) - F_0(x)|$$

- Für Stichproben aus F_0 strebt die Verteilungsfunktion von $\sqrt{n}D$ für $n \to \infty$ gegen:

$$K(x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$$

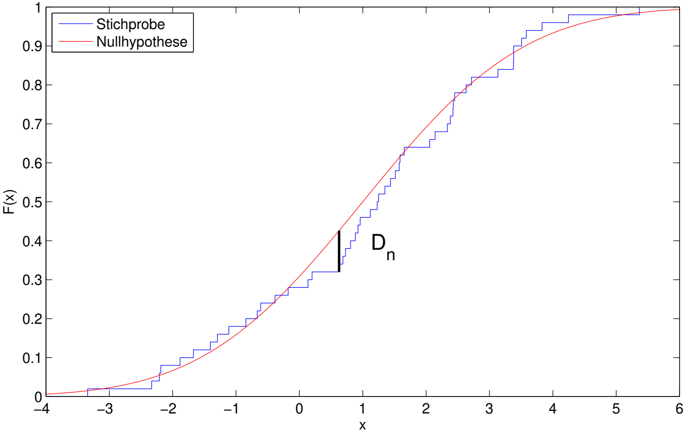
Der Kolmogorov-Smirnov-Test

- Aus der asymptotischen Verteilungsfunktion können Quantile $K_{1-\alpha}$ berechnet werden.
- Die Nullhypothese wird abgelehnt, wenn

$$\sqrt{n}D_n > K_{1-\alpha}$$

- Werden vor dem Test Parameter von F_0 geschätzt, sind die Quantile nicht mehr gültig.
- In diesem Fall muss der Ablehnungsbereich durch Simulation ermittelt werden.

Der Kolmogorov-Smirnov-Test



D_n, die Teststatistik des Kolmogorov-Smirnov-Test

Der Kolmogorov-Smirnov-Test

Zwei Stichproben

- Wir testen, ob zwei Stichproben vom Umfang n bzw. m aus der gleichen Verteilung F stammen.
- Die Testgröße ist die maximale absolute Differenz der empirischen Verteilungsfunktionen:

$$D_{n,m} = \max_x \left| F_n^1(x) - F_m^2(x) \right|$$

- Die Nullhypothese wird abgelehnt, wenn

$$\sqrt{\frac{nm}{n+m}} D_{n,m} > K_{1-\alpha}$$

Teil VII

Regression und lineare Modelle

Übersicht Teil 7

Abschnitt 28: Einleitung

Einleitung

- Regressionsanalyse untersucht die Abhängigkeit der Beobachtungen von diversen Variablen.
- **Einflussvariable** (unabhängige Variable) $\boldsymbol{x} = (x_1, \ldots, x_r)$.
- **Ergebnisvariable** (abhängige Variable) Y.
- **Regressionsmodell**:

$$Y = f(\boldsymbol{\beta}, \boldsymbol{x}) + \varepsilon$$

  mit **Regressionskoeffizienten** $\boldsymbol{\beta}$ und **Fehlerterm** ε.
- Ziel ist die **Schätzung von** $\boldsymbol{\beta}$ anhand von Beobachtungen $Y_1, \ldots, Y_n,\ n \geq r$.
- Eine Einflussvariable: einfache Regression;
  Mehrere Einflussvariable: mehrfache (multiple) Regression.

Einleitung

- Jede Beobachtung Y_i hat einen Fehlerterm ε_i.
- Die Fehlerterme brauchen nicht alle die gleiche Verteilung zu haben und auch nicht unabhängig zu sein.
- Häufig wird angenommen, dass der Zufallsvektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ eine **multivariate Normalverteilung** hat.
- Es sind aber auch andere Verteilungen von $\boldsymbol{\varepsilon}$ denkbar.

Abschnitt 29: Mehrdimensionale Zufallsvariable

Unterabschnitt: Grundbegriffe

Grundbegriffe

Definition (Mehrdimensionale Zufallsvariable)

Eine Abbildung $\boldsymbol{X}$:

$$\omega \in \Omega \mapsto \boldsymbol{x} = \boldsymbol{X}(\omega) \in \mathbb{R}^d$$

die jedem Element ω des Ergebnisraums Ω einen reellen Vektor $\boldsymbol{x} \in \mathbb{R}^d$ zuordnet, heißt eine d-dimensionale Zufallsvariable.

- Jede Komponente einer d-dimensionalen Zufallsvariablen ist selbst eine Zufallsvariable.
- Jede Komponente kann diskret oder stetig sein.

Grundbegriffe

Definition (Verteilungsfunktion)

Ist $\boldsymbol{X} = (X_1, \ldots, X_d)$ eine d-dimensionale Zufallsvariable, so ist ihre Verteilungsfunktion $F_{\boldsymbol{X}}$ durch

$$F_{\boldsymbol{X}}(x_1, \ldots, x_d) = W(X_1 \leq x_1 \cap \ldots \cap X_d \leq x_d)$$

definiert.

Definition (Dichtefunktion)

Ist $\boldsymbol{X} = (X_1, \ldots, X_d)$ eine d-dimensionale diskrete Zufallsvariable, so ist ihre Dichtefunktion $f_{\boldsymbol{X}}$ durch

$$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = W(X_1 = x_1 \cap \ldots \cap X_d = x_d)$$

definiert.

Grundbegriffe

Beispiel

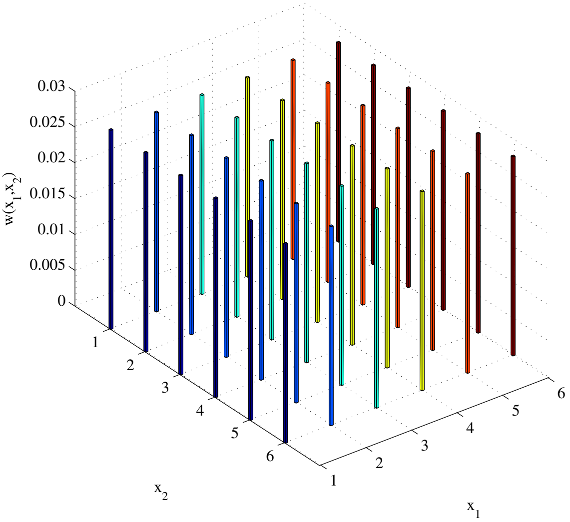
Die bivariate (zweidimensionale) Zufallsvariable $\boldsymbol{X} = (X_1, X_2)$ ordnet dem Ergebnis (e_i, e_j) des Wurfs mit zwei Würfeln die Augenzahlen (i, j) zu. Sind alle Ausgänge gleichwahrscheinlich, so ist $W_{\boldsymbol{X}}$ gegeben durch:

$$W_{\boldsymbol{X}}\left(\{(i,j)\}\right) = \frac{1}{36}$$

Die Dichte $f_{\boldsymbol{X}}$ lautet:

$$f_{\boldsymbol{X}}(x_1, x_2) = \begin{cases} \frac{1}{36}, & x_1 \in \{1, \ldots, 6\} \cap x_2 \in \{1, \ldots, 6\} \\ 0, & \text{sonst} \end{cases}$$

Grundbegriffe



Grundbegriffe

Beispiel (Fortsetzung)

Die Verteilungsfunktion $F_{\boldsymbol{X}}$ ist daher:

$$F_{\boldsymbol{X}}(x_1, x_2) = W(X_1 \leq x_1 \cap X_2 \leq x_2) = \sum_{i \leq x_1 \cap j \leq x_2} f(i, j)$$

Beispielsweise ist $F_{\boldsymbol{X}}(3, 4) = \sum_{i \leq 3 \cap j \leq 4} \frac{1}{36} = \frac{12}{36} = \frac{1}{3}$.

- Wegen der Abzählbarkeit der Elementarereignisse können diese auch durch eine univariate Zufallsvariable Y eindeutig in $\mathbb{R}$ abgebildet werden, z. B.:

  $$Y : (e_i, e_j) \longrightarrow 6i + j - 6$$

  Der Wertevorrat von Y sind die natürlichen Zahlen zwischen 1 und 36, und W_Y ist gegeben durch:

  $$W_Y(\{k\}) = \frac{1}{36}, \; 1 \leq k \leq 36$$

Grundbegriffe

Definition (Dichtefunktion)

Ist $\boldsymbol{X} = (X_1, \ldots, X_d)$ eine d-dimensionale stetige Zufallsvariable, so ist ihre Dichtefunktion $f_{\boldsymbol{X}}$ durch

$$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = \frac{\partial^d F_{\boldsymbol{X}}}{\partial x_1 \ldots \partial x_d}$$

definiert.

Definition (Momente)

- Erwartung: $\mathsf{E}[\boldsymbol{X}] = (\mathsf{E}[X_1] \cdots \mathsf{E}[X_d])$
- Kovarianzmatrix: $\mathsf{Cov}[\boldsymbol{X}]_{ij} = \mathsf{cov}[X_i, X_j]$

Unterabschnitt: Randverteilungen und bedingte Verteilungen

Randverteilungen und bedingte Verteilungen

- Sind X_1 und X_2 zwei (diskrete oder stetige) univariate Zufallsvariable, so ist $\boldsymbol{X} = (X_1, X_2)$ eine bivariate Zufallsvariable. Die Verteilung (Verteilungsfunktion, Dichte) von $\boldsymbol{X}$ heißt auch die **gemeinsame Verteilung** (Verteilungsfunktion, Dichte) von X_1 und X_2.
- Es stellt sich nun das folgende Problem: Kann man die Verteilung von X_1 bzw. X_2 aus der gemeinsamen Verteilung berechnen?

Randverteilungen und bedingte Verteilungen

- Es sei F die Verteilungsfunktion und f die Dichte der stetigen Zufallsvariablen $\boldsymbol{X} = (X_1, X_2)$. Dann ist die Verteilungsfunktion F_1 von X_1 gegeben durch:

$$F_1(x_1) = W(X_1 \leq x_1) = W(X_1 \leq x_1 \cap -\infty < X_2 < \infty) =$$
$$= \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1$$

- Daraus folgt:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2$$

ist die Dichte von X_1.

Randverteilungen und bedingte Verteilungen

Definition (Randverteilung)

Es sei $\boldsymbol{X} = (X_1, X_2)$ eine bivariate stetige Zufallsvariable mit der Verteilungsfunktion F und der Dichte f. Die Verteilung von X_1 heißt die **Randverteilung** von X_1 bezüglich $\boldsymbol{X}$. Ihre Dichte f_1 lautet:

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2.$$

Ist $\boldsymbol{X} = (X_1, X_2)$ diskret mit der Dichte f, so ist analog die Dichte f_1 der Randverteilung von X_1 bezüglich $\boldsymbol{X}$ gegeben durch:

$$f_1(k_1) = \sum_{k_2} f(k_1, k_2)$$

Randverteilungen und bedingte Verteilungen

- Die Verteilungen von X_1 und X_2 lassen sich also aus der gemeinsamen Verteilung von X_1 und X_2 berechnen.
- Der umgekehrte Vorgang ist im allgemeinen nicht möglich, da die gemeinsame Verteilung auch Information über Zusammenhänge (Kopplung) zwischen X_1 und X_2 enthält.
- Es seien X_1 und X_2 zwei diskrete Zufallsvariable mit der gemeinsamen Dichte $f(k_1, k_2)$ und den Randverteilungsdichten $f_1(k_1)$ und $f_2(k_2)$. Dann ist die bedingte Wahrscheinlichkeit des Ereignisses $X_1 = k_1$ unter der Bedingung $X_2 = k_2$ gegeben durch:

$$W(X_1 = k_1 | X_2 = k_2) = \frac{W(X_1 = k_1 \cap X_2 = k_2)}{W(X_2 = k_2)} = \frac{f(k_1, k_2)}{f_2(k_2)}$$

Randverteilungen und bedingte Verteilungen

Definition (Bedingte Dichte)

Es sei $\boldsymbol{X} = (X_1, X_2)$ eine bivariate diskrete Zufallsvariable mit der Dichte $f(k_1, k_2)$ und den Randverteilungsdichten $f_1(k_1)$ bzw. $f_2(k_2)$. Die Funktion $f(k_1|k_2)$, definiert durch:

$$f(k_1|k_2) = \frac{f(k_1, k_2)}{f_2(k_2)}$$

heißt die durch X_2 **bedingte Dichte** von X_1.

- Die bedingte Dichte ist für festes k_2 die Dichte einer Verteilung, der durch $X_2 = k_2$ **bedingten Verteilung** von X_1.

Randverteilungen und bedingte Verteilungen

- Ist $\boldsymbol{X} = (X_1, X_2)$ stetig, so ist analog $f(x_1|x_2)$ definiert durch:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \ (f_2(x_2) \neq 0)$$

- $f(x_1|x_2)$ ist für festes x_2 die Dichte einer Verteilung, der durch $X_2 = x_2$ bedingten Verteilung von X_1.
- Dass $f(x_1|x_2)$ tatsächlich eine Dichte ist, läßt sich leicht nachprüfen:

$$\int_{-\infty}^{\infty} f(x_1|x_2)\,\mathrm{d}x_1 = \int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_2(x_2)}\,\mathrm{d}x_1 = \frac{f_2(x_2)}{f_2(x_2)} = 1$$

und analog für diskretes $\boldsymbol{X}$.

Randverteilungen und bedingte Verteilungen

Eigenschaften der bedingen Dichte

Es gilt:

$$f(x_1, x_2) = f(x_1|x_2) \cdot f_2(x_2)$$
$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1|x_2) \cdot f_2(x_2)\,\mathrm{d}x_2$$

und analog für diskrete Dichten.

Definition (Unabhängigkeit von Zufallsvariablen)

Ist die (unbedingte) Dichte der Randverteilung von X_1 gleich der durch X_2 bedingten Dichte, so heißen X_1 und X_2 **unabhängig**.

$$X_1 \text{ und } X_2 \text{ unabhängig} \iff f(x_1|x_2) = f_1(x_1)$$

Randverteilungen und bedingte Verteilungen

- Für unabhängige Zufallsvariable X_1 und X_2 gilt:

$$\begin{aligned} f(x_1|x_2) = f_1(x_1) &\iff f(x_2|x_1) = f_2(x_1) \\ &\iff f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2) \end{aligned}$$

und analog für diskretes $\boldsymbol{X}$.

Eigenschaften von unabhängigen Zufallsvariablen

- $f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$
- $\mathsf{E}[X_1 X_2] = \mathsf{E}[X_1] \cdot \mathsf{E}[X_2]$
- $\mathsf{cov}[X_1, X_2] = \rho[X_1, X_2] = 0$

- Unkorrelierte Zufallsvariable sind nicht notwendigerweise auch unabhängig!

Randverteilungen und bedingte Verteilungen

- Ist $\boldsymbol{X} = (X_1, \ldots, X_d),\ d > 2$, so müssen die Definitionen der Randverteilung, der bedingten Dichten und der Unabhängigkeit entsprechend verallgemeinert werden.
- Die Dichte $f_{i_1,\ldots,i_m}$ der Randverteilung von $X_{i_1}, \ldots, X_{i_m}$ erhält man, indem über alle anderen Variablen integriert bzw. aufsummiert wird.
- Die durch X_j bedingte Dichte von X_i ist gegeben durch:

$$f(x_i|x_j) = \frac{f_{i,j}(x_i, x_j)}{f_j(x_j)}$$

  wobei $f_{i,j}(x_i, x_j)$ die Randverteilungsdichte von X_i, X_j ist.
- $X_{i_1}, \ldots, X_{i_k}$ heißen unabhängig, wenn die Dichte ihrer Randverteilung das Produkt der Dichten der Randverteilungen der einzelnen X_{i_j} ist.

Randverteilungen und bedingte Verteilungen

Beispiel (Die Akzeptanz oder Nachweiswahrscheinlichkeit)

X sei eine Zufallsvariable mit der Dichte $f(x)$. Nimmt X den Wert x an, so gibt es eine Wahrscheinlichkeit $a(x)$ dafür, dass x auch tatsächlich beobachtet wird. Man definiert nun eine Zufallsvariable I, die 1 ist, wenn x beobachtet wird, und 0 sonst. Dann ist I unter der Bedingung $X = x$ Bernoulli-verteilt gemäß $\mathrm{Al}(a(x))$:

$$W(I = 1|X = x) = a(x)$$
$$W(I = 0|X = x) = 1 - a(x)$$

Die gemeinsame Dichte von X und I ist daher:

$$f(x, 1) = a(x)f(x)$$
$$f(x, 0) = [1 - a(x)]f(x)$$

Randverteilungen und bedingte Verteilungen

Beispiel (Fortsetzung)

Da der Experimentator nur mit beobachteten Größen arbeiten kann, schränkt er seine Grundgesamtheit auf die nachgewiesenen Ereignisse ein, d.h. er braucht die Dichte von X unter der Bedingung, dass X beobachtet wird:

$$f_A(x) = f(x|I=1) = \frac{f(x,1)}{f_2(1)} = \frac{a(x)f(x)}{\int a(x)f(x)\,\mathrm{d}x}$$

Als konkretes Beispiel diene die Messung einer Lebensdauer. Die Messung möge bei $t_{\min}$ beginnen und bei $t_{\max}$ enden. Dann hat $a(t)$ die folgende Gestalt:

$$a(t) = \begin{cases} 0, & \text{für } t \le t_{\min} \\ 1, & \text{für } t_{\min} < t \le t_{\max} \\ 0, & \text{für } t > t_{\max} \end{cases}$$

Randverteilungen und bedingte Verteilungen

Beispiel (Fortsetzung)

Für die gemessene Wahrscheinlichkeitsdichte gilt:

$$f_A(t) = \begin{cases} 0, \ t \le t_{\min} \\ \dfrac{\frac{1}{\tau}\exp(-t/\tau)}{\exp(-t_{\min}/\tau) - \exp(-t_{\max}/\tau)}, \ t_{\min} \le t < t_{\max} \\ 0, \ t > t_{\max} \end{cases}$$

Der Nenner $[\exp(-t_{\min}/\tau) - \exp(-t_{\max}/\tau)]$ korrigiert für jene Teilchen, die vor $t_{\min}$ oder nach $t_{\max}$ zerfallen.

Die Nachweiswahrscheinlichkeit $a(t)$ kann auch eine wesentlich kompliziertere Abhängigkeit von t haben. So kann es etwa von der Art und geometrischen Konfiguration der Zerfallsprodukte abhängen, ob ein Zerfall bei t beobachtet werden kann oder nicht.

Unterabschnitt: Die multivariate Normalverteilung

Die multivariate Normalverteilung

Die multivariate Normalverteilung $\mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$

- Ihre Dichte lautet:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\mathbf{V}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Momente

Es sei $\boldsymbol{X} = (X_1, \ldots, X_d) \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$. Dann gilt:

- $\mathsf{E}[\boldsymbol{X}] = \boldsymbol{\mu}$
- $\mathsf{Cov}[\boldsymbol{X}] = \mathbf{V}$
- $\mathbf{V}^{-1}$ ist ebenfalls symmetrisch und positiv definit, und wird als Gewichts- oder Informationsmatrix bezeichnet.

Die multivariate Normalverteilung

Lineare Transformationen

Es sei $\boldsymbol{X} \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$ und $\mathbf{H}$ eine $m \times d$ Matrix. Dann ist $\boldsymbol{Y} = \mathbf{H}\boldsymbol{X} \sim \mathrm{No}(\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\mathbf{V}\mathbf{H}^{\mathrm{T}})$.

Randverteilungen

Jede Randverteilung einer Normalverteilung ist wieder eine Normalverteilung. Mittelwert und Kovarianzmatrix der Randverteilung entstehen durch Streichen der Spalten und Zeilen der restlichen Variablen.

Bedingte Verteilungen

Jede bedingte Verteilung einer Normalverteilung ist wieder eine Normalverteilung.

Die multivariate Normalverteilung

- Ist $\boldsymbol{X} \sim \mathrm{No}(\boldsymbol{\mu}, \mathbf{V})$, so kann $\mathbf{V}$ als positiv definite symmetrische Matrix mittels einer orthogonalen Transformation (Rotation) auf Diagonalform gebracht werden:

$$\mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}} = \mathbf{D}^2$$

- Alle Diagonalelemente von $\mathbf{D}^2$ sind positiv. Die Zufallsvariable $\boldsymbol{Z} = \mathbf{D}\mathbf{U}(\boldsymbol{X} - \boldsymbol{\mu})$ ist dann multivariat standardnormalverteilt, d.h.:

$$\mathsf{E}[\boldsymbol{Z}] = \mathbf{0}, \quad \mathsf{Cov}[\boldsymbol{Z}] = \mathbf{I}$$

Die Drehung $\mathbf{U}$ heißt **Hauptachsentransformation**.

Die multivariate Normalverteilung

Die bivariate Normalverteilung

- Für $d = 2$ und $\boldsymbol{\mu} = \mathbf{0}$ kann die Dichte folgendermaßen angeschrieben werden:

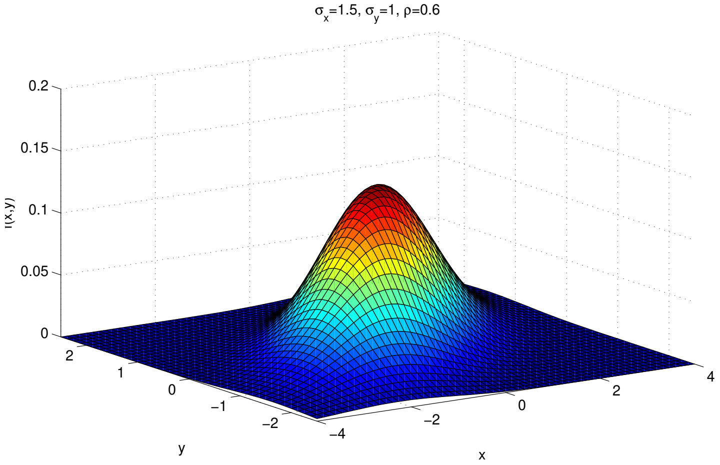
$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2\,\rho\,x\,y}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)\right]$$

- $\rho = \sigma_{xy}/(\sigma_x\sigma_y)$ ist der Korrelationskoeffizient. Sind X und Y unkorreliert, also $\rho = 0$, folgt:

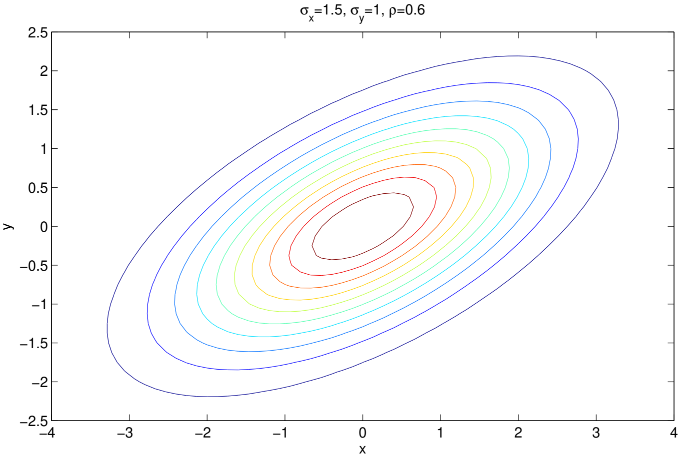
$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] = f_1(x) \cdot f_2(y)$$

- Zwei unkorrelierte normalverteilte Zufallsvariable mit gemeinsamer Normalverteilung sind daher **unabhängig**.

Die multivariate Normalverteilung



Dichtefunktion einer bivariaten Normalverteilung



Höhenschichtlinien einer bivariaten Normalverteilung

Die multivariate Normalverteilung

- Die **bedingte Dichte** $f(y|x)$ ist gegeben durch

$$f(y|x) = \frac{f(x,y)}{f(x)} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,\sigma_y^2\,(1-\rho^2)}\left(y - \frac{\rho\,y\,\sigma_y}{\sigma_x}\right)^2\right]$$

- $Y|X = x$ ist also eine normalverteilte Zufallsvariable mit der Erwartung

$$\mathrm{E}[Y|X] = \rho\, x\, \sigma_y / \sigma_x$$

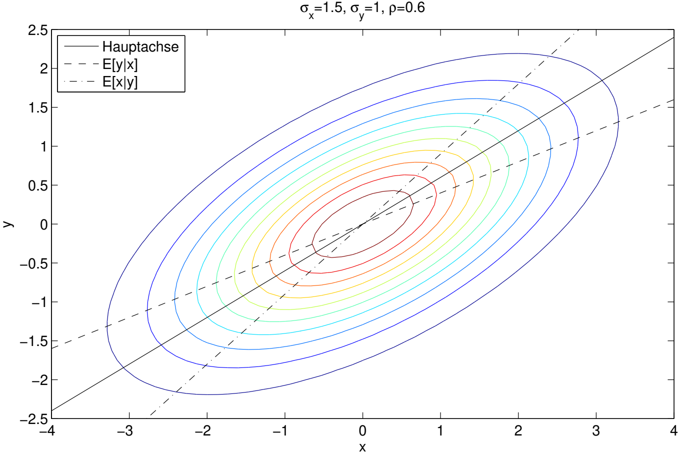
- $\mathrm{E}[Y|X]$ heißt die **bedingte Erwartung** oder **Regression von** y **auf** x.

Die multivariate Normalverteilung

- Je nach Vorzeichen von ρ fällt oder wächst die bedingte Erwartung von Y, wenn X wächst.
- Ist $\rho = 1$, sind X und Y proportional: $Y = X\,\sigma_y/\sigma_x$.
- Die Höhenschichtlinien der Dichtefunktion sind **Ellipsen**.
- Die **Hauptachsentransformation** ist jene Drehung, die die Ellipsen in achsenparallele Lage bringt.
- Sie hängt im Fall $d = 2$ nur von ρ ab. Ist $\rho = 0$, sind X und Y bereits unabhängig, und der Drehwinkel ist gleich 0. Ist $\rho \neq 0$, ist die Drehmatrix U gleich

$$U = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \text{ mit } \varphi = -\frac{1}{2}\,\mathrm{arccot}\,\frac{\sigma_y^2 - \sigma_x^2}{2\rho\sigma_x\sigma_y}$$

Die multivariate Normalverteilung



Hauptachse und Regressionsgeraden einer bivariaten Normalverteilung

Unterabschnitt: Multivariate Fehlerfortpflanzung

Multivariate Fehlerfortpflanzung

Affine Transformationen

- Es sei $\boldsymbol{X}$ eine Zufallsvariable mit der Dichte $f(\boldsymbol{x})$ und $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{b}$ mit regulärem $\mathbf{A}$.
- Die Dichte $g(\boldsymbol{y})$ von $\boldsymbol{Y}$ ist dann gleich

$$g(\boldsymbol{y}) = \frac{1}{|\mathbf{A}|} f\left(\mathbf{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})\right)$$

- Ferner gilt für beliebiges $\mathbf{A}$:

$$\mathsf{E}[\boldsymbol{Y}] = \mathbf{A} \cdot \mathsf{E}[\boldsymbol{X}] + \boldsymbol{b}$$
$$\mathsf{Cov}[\boldsymbol{Y}] = \mathbf{A} \cdot \mathsf{Cov}[\boldsymbol{X}] \cdot \mathbf{A}^{\mathrm{T}}$$

Multivariate Fehlerfortpflanzung

Nichtlineare Transformationen

- Es sei $\boldsymbol{X}$ eine Zufallsvariable mit der Dichte $f(\boldsymbol{x})$ und $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$.
- Ist $\boldsymbol{h}(\boldsymbol{x})$ bijektiv, ist die Dichte $g(\boldsymbol{y})$ von $\boldsymbol{Y}$ gleich

$$g(\boldsymbol{y}) = \left| \frac{\partial \boldsymbol{h}^{-1}}{\partial \boldsymbol{y}} \right| f\left(\boldsymbol{h}^{-1}(\boldsymbol{y})\right)$$

- Die Erwartung und die Varianz von $\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{X})$ können **näherungsweise** mit Hilfe der Taylorentwicklung von $\boldsymbol{h}(\boldsymbol{x})$ berechnet werden, auch wenn $\boldsymbol{h}(\boldsymbol{x})$ nicht bijektiv ist.

Multivariate Fehlerfortpflanzung

- Mit der Entwicklungsstelle $\boldsymbol{x}_0$ gilt in linearer Näherung

$$\boldsymbol{h}(\boldsymbol{x}) \approx \boldsymbol{h}(\boldsymbol{x}_0) + \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0), \quad \mathbf{H} = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

- Mit der Wahl $\boldsymbol{x}_0 = \mathsf{E}[\boldsymbol{X}]$ folgt

Satz

$$\mathsf{E}[\boldsymbol{h}(\boldsymbol{X})] \approx \boldsymbol{h}(\mathsf{E}[\boldsymbol{X}])$$

$$\mathsf{Cov}[\boldsymbol{h}(\boldsymbol{X})] \approx \mathbf{H} \cdot \mathsf{Cov}[X] \cdot \mathbf{H}^{\mathrm{T}} \quad \text{(Lineare Fehlerfortpflanzung)}$$

Abschnitt 30: Einfache Regression

Unterabschnitt: Lineare Regression

Lineare Regression

- Das einfachste Regressionsmodell ist eine Gerade:

$$Y = \alpha + \beta x + \varepsilon, \quad \mathsf{E}[\varepsilon] = 0,\ \mathsf{var}[\varepsilon] = \sigma^2$$

- Es seien nun $Y_1, \ldots, Y_n$ die Ergebnisse für die Werte $x_1, \ldots, x_n$ der Einflussvariablen x.
- Die Schätzung von α und β kann nach dem Prinzip der kleinsten Fehlerquadrate erfolgen.
- Die folgende Zielfunktion wird minimiert:

$$SS = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$$

- Gradient von SS:

$$\frac{\partial SS}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i),\ \frac{\partial SS}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (Y_i - \alpha - \beta x_i)$$

Lineare Regression

- Nullsetzen des Gradienten gibt die **Normalgleichungen**:

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i Y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

- Die geschätzten Regressionskoeffizienten lauten:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{x} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\bar{x}$$

Lineare Regression

- Es gilt

$$\mathsf{E}[\hat{\alpha}] = \alpha, \quad \mathsf{E}[\hat{\beta}] = \beta$$

- Die Varianz des Fehlerterms wird erwartungstreu geschätzt durch:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2$$

mit

$$r_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- Die r_i werden als **Residuen** der Regression bezeichnet.

Lineare Regression

- Die Kovarianzmatrix der geschätzten Regressionkoeffizienten erhält man durch lineare Fehlerfortpflanzung:

$$\mathsf{Cov}[\hat{\alpha}, \hat{\beta}] = \sigma^2 \begin{pmatrix} \dfrac{\sum x_i^2}{n\left(\sum x_i^2 - n\bar{x}^2\right)} & -\dfrac{\sum x_i}{n\left(\sum x_i^2 - n\bar{x}^2\right)} \\ -\dfrac{\sum x_i}{n\left(\sum x_i^2 - n\bar{x}^2\right)} & \dfrac{1}{\sum x_i^2 - n\bar{x}^2} \end{pmatrix}$$

Lineare Regression

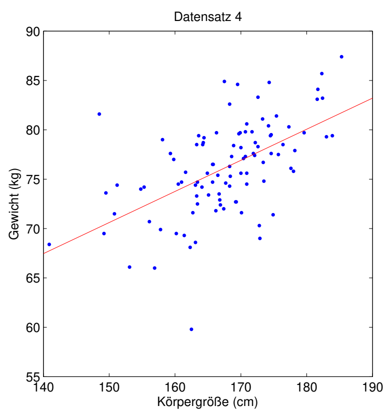
Beispiel

Datensatz 4:

$$\begin{aligned} \bar{x} &= 167.60 & r_{xy} &= 0.5562 \\ \bar{y} &= 76.16 & \hat{a} &= 23.37 \\ s_x &= 8.348 & \hat{b} &= 0.3150 \\ s_y &= 4.727 \end{aligned}$$

Lineare Regression

Beispiel (Fortsetzung)



Streudiagramm mit Regressionsgerade

Lineare Regression

- Die Streuung der Werte Y_i hat im Regressionsmodell unterschiedliche Ursachen.
- Einerseits gibt es systematische Unterschiede durch unterschiedliche Werte von x.
- Dazu kommt noch die zufällige Streuung der Daten.

Erklärbare Streuung	$SS^* = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = r_{xy}^2 n s_Y^2$
Reststreuung	$SS_R = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (1 - r_{xy}^2) n s_Y^2$
Totale Streuung	$SS_T = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = n s_Y^2$

Lineare Regression

Streuungszerlegung

$$SS_T = SS^* + SS_R$$

- Die Güte der Regressionsgeraden kann durch das **Bestimmtheitsmaß** angegeben werden:

Bestimmheitsmaß der Regression

$$B = \frac{SS^*}{SS_T} = r_{xy}^2$$

- Es gibt an, welcher Anteil an der Gesamtstreuung durch die Korrelation von x und Y erklärt werden kann.

Unterabschnitt: Tests, Konfidenz- und Prognoseintervalle

Tests, Konfidenz- und Prognoseintervalle

- Ist $\beta = 0$, hängt das Ergebnis überhaupt nicht von den Einflussvariablen ab.
- Ein Test der Nullhypothese $H_0 : \beta = 0$ gegen $H_1 : \beta \neq 0$ beruht auf dem folgenden Satz.

Satz

Ist ε normalverteilt, so sind

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}}, \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}}$$

t-verteilt mit $n - 2$ Freiheitsgraden, wobei

$$\hat{\sigma}^2_{\hat{\alpha}} = \frac{\hat{\sigma}^2 \sum x_i^2}{n \left(\sum x_i^2 - n\bar{x}^2\right)}, \quad \hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum x_i^2 - n\bar{x}^2}$$

Tests, Konfidenz- und Prognoseintervalle

- Die Nullhypothese $H_0 : \beta = 0$ wird abgelehnt, wenn die Testgröße

$$T = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

relativ klein oder relativ groß ist, also wenn

$$\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}} > t_{1-\alpha/2;n-2}$$

wo $t_{p;n-2}$ das p-Quantil der t-Verteilung mit $n-2$ Freiheitsgraden ist.

- Ein analoger Test kann für die Nullhypothese $H_0 : \alpha = 0$ durchgeführt werden.

Tests, Konfidenz- und Prognoseintervalle

Symmetrische Konfidenzintervalle

$$\hat{\alpha} \pm \hat{\sigma}_{\hat{\alpha}} \cdot t_{1-\alpha/2;n-2}, \quad \hat{\beta} \pm \hat{\sigma}_{\hat{\beta}} \cdot t_{1-\alpha/2;n-2}$$

- Für $n \gtrsim 50$ können die Quantile der t-Verteilung durch Quantile der Standardnormalverteilung ersetzt werden.
- Es soll nun das Ergebnis $Y_0 = Y(x_0)$ für einen bestimmten Wert x_0 der Einflussvariablen x prognostiziert werden.
- Der Erwartungswert von Y_0 ist

$$\mathsf{E}[Y_0] = \hat{\alpha} + \hat{\beta} x_0$$

- Die Varianz von $\mathsf{E}[Y_0]$ ergibt sich mittels Fehlerfortpflanzung:

$$\mathsf{var}[\mathsf{E}[Y_0]] = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} \right]$$

Tests, Konfidenz- und Prognoseintervalle

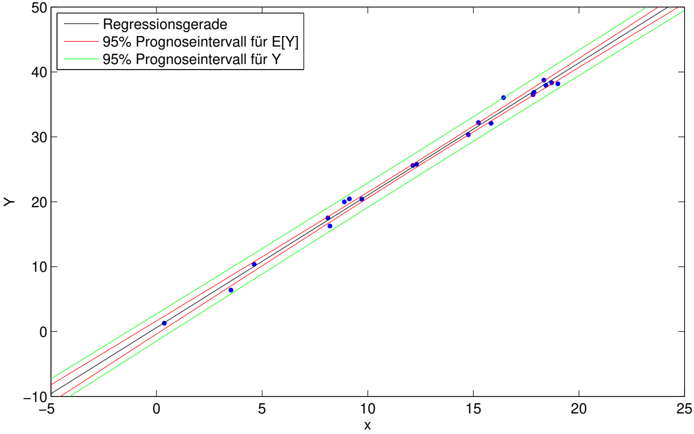
- Da Y_0 um seinen Erwartungswert mit Varianz σ^2 streut, ergibt sich:

$$\text{var}[Y_0] = \sigma^2 \left[\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2} \right]$$

- Das symmetrische Prognoseintervall für Y_0 mit Sicherheit α ist daher gleich:

$$\hat{\alpha} + \hat{\beta} x_0 \pm t_{1-\alpha/2;n-2} \hat{\sigma} \sqrt{\frac{n+1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n\bar{x}^2}}$$

Tests, Konfidenz- und Prognoseintervalle



Prognosebänder für $\mathsf{E}[Y]$ und Y

Tests, Konfidenz- und Prognoseintervalle

- Die Angemessenheit des Modells kann durch Untersuchung der **studentisierten Residuen** (Restfehler) überprüft werden.
- Das Residuum r_k hat die Varianz

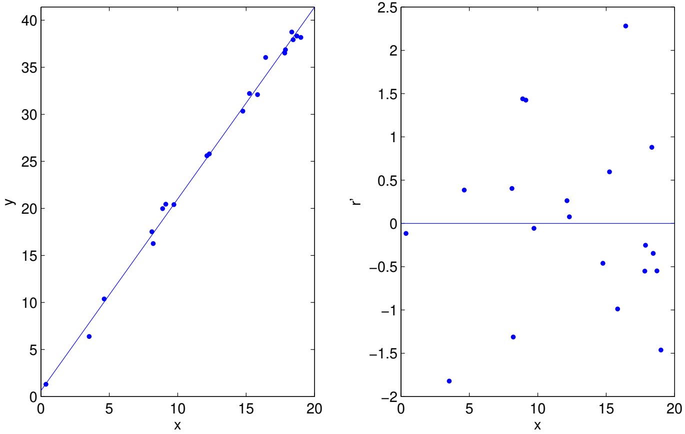
$$\mathsf{var}[r_k] = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_k - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}\right]$$

- Das studentisierte Residuum ist dann

$$r'_k = \frac{r_k}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_k - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}}$$

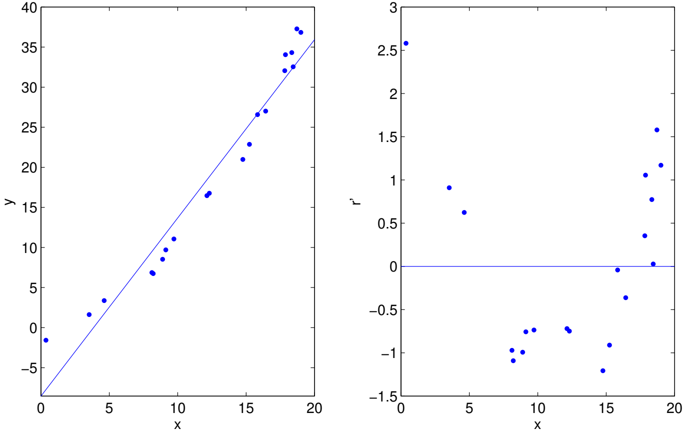
- Es hat Erwartung 0 und Varianz 1.

Tests, Konfidenz- und Prognoseintervalle



Regressionsgerade und studentisierte Residuen

Tests, Konfidenz- und Prognoseintervalle



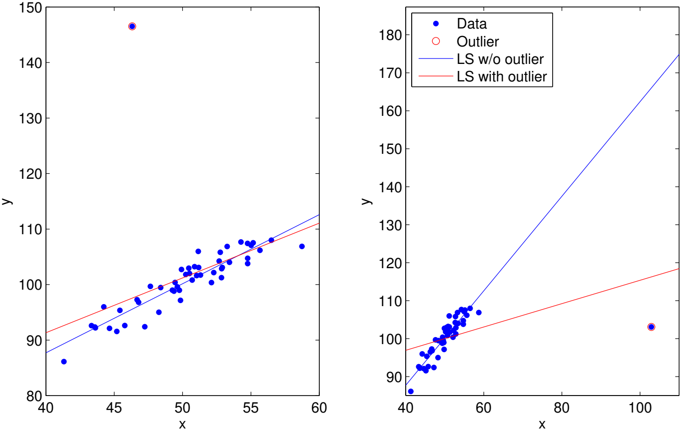
Regressionsgerade und studentisierte Residuen

Unterabschnitt: Robuste Regression

Robuste Regression

- Als LS-Schätzer ist die Regressionsgerade nicht robust, d.h. empfindlich gegen Ausreißer.
- Einzelne Ausreißer in der Ergebnisvariablen y haben in der Regel eine leichte Verzerrung der Regressiongeraden zur Folge.
- Einzelne Ausreißer in der Einflussvariablen x, sogenannte Hebelpunkte (leverage points), können katastrophale Verzerrungen zur Folge haben.

Robuste Regression

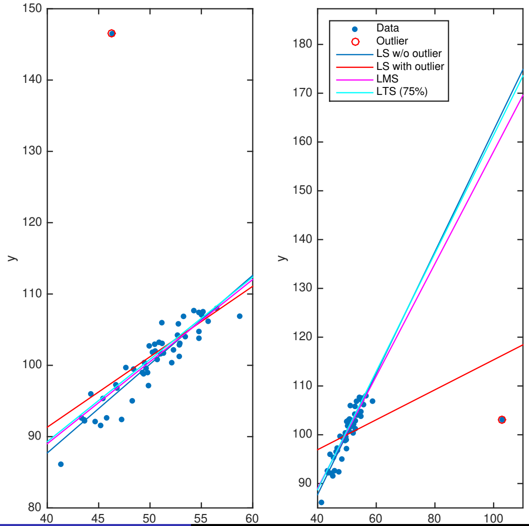


Lineare Regression mit Ausreißern

Robuste Regression

- **LMS (Least Median of Squares)**: Anstatt der Summe der Fehlerquadrate wird der **Median** der Fehlerquadrate minimiert.
- "Exact fit property": Die LMS-Gerade geht durch zwei Datenpunkte.
- Berechnung kombinatorisch.
- **LTS (Least Trimmed Squares)**: Es wird die Summe einer festen Anzahl $h \leq n$ von Fehlerquadraten minimiert.
- Berechnung iterativ (FAST-LTS).
- Beide Methoden gehen auf P. Rousseeuw zurück.

Robuste Regression



Unterabschnitt: Polynomiale Regression

Polynomiale Regression

- Ist der Zusammenhang zwischen x und Y nicht annähernd linear, kann man versuchen, ein Polynom anzupassen.
- Das Modell lautet dann:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \varepsilon, \quad \mathsf{E}[\varepsilon] = 0,\ \mathsf{var}[\varepsilon] = \sigma^2$$

- Es seien wieder $Y_1, \ldots, Y_n$ die Ergebnisse für die Werte $x_1, \ldots, x_n$ der Einflussvariablen x.
- In Matrix-Vektor-Schreibweise:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathsf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$$

mit

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^r \\ 1 & x_2 & x_2^2 & \cdots & x_2^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^r \end{pmatrix}$$

Polynomiale Regression

- Die folgende Zielfunktion wird minimiert:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Gradient von SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Nullsetzen des Gradienten gibt die **Normalgleichungen**:

$$\mathbf{X}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- Die Lösung lautet:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$$

Polynomiale Regression

- $\hat{\boldsymbol{\beta}}$ ist ein erwartungstreuer Schätzer von $\boldsymbol{\beta}$.
- Die Varianz des Fehlerterms wird erwartungstreu geschätzt durch:

$$\hat{\sigma}^2 = \frac{1}{n-r-1} \sum_{i=1}^{n} r_i^2$$

mit dem Vektor der Residuen

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \quad \hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

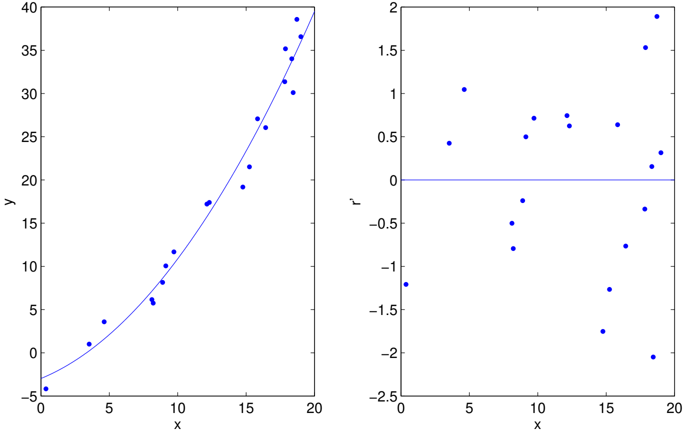
- Kovarianzmatrix der geschätzten Regressionkoeffizienten:

$$\mathsf{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

- Kovarianzmatrix der Residuen $\boldsymbol{r}$:

$$\mathsf{Cov}[\boldsymbol{r}] = \sigma^2 \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]$$

Polynomiale Regression



Regressionsparabel und studentisierte Residuen

Abschnitt 31: Mehrfache Regression

Unterabschnitt: Das lineare Modell

Das lineare Modell

- Hängt das Ergebnis Y von mehreren Einflussvariablen ab, lautet das einfachste lineare Regressionmodell:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \varepsilon, \quad \mathsf{E}[\varepsilon] = 0, \ \mathsf{var}[\varepsilon] = \sigma^2$$

- Es seien wieder $Y_1, \ldots, Y_n$ die Ergebnisse für n Werte $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ der Einflussvariablen $\boldsymbol{x} = (x_1, \ldots, x_r)$.
- In Matrix-Vektor-Schreibweise:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathsf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}$$

mit

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,r} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,r} \end{pmatrix}$$

Unterabschnitt: Schätzung, Tests und Prognoseintervalle

Schätzung, Tests und Prognoseintervalle

- Die folgende Zielfunktion wird minimiert:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Gradient von SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Nullsetzen des Gradienten gibt die **Normalgleichungen**:

$$\mathbf{X}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- Die Lösung lautet:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$$

Schätzung, Tests und Prognoseintervalle

- $\hat{\boldsymbol{\beta}}$ ist ein erwartungstreuer Schätzer von $\boldsymbol{\beta}$.
- Die Varianz des Fehlerterms wird erwartungstreu geschätzt durch:

$$\hat{\sigma}^2 = \frac{1}{n-r-1} \sum_{i=1}^{n} r_i^2$$

mit

$$\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}, \quad \hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Kovarianzmatrix der geschätzten Regressionkoeffizienten:

$$\mathsf{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

- Kovarianzmatrix der Residuen $\boldsymbol{r}$:

$$\mathsf{Cov}[\boldsymbol{r}] = \sigma^2 \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]$$

Schätzung, Tests und Prognoseintervalle

- Ist $\beta_k = 0$, hängt das Ergebnis überhaupt nicht von den Einflussvariablen x_k ab.
- Ein Test der Nullhypothese $H_0 : \beta_k = 0$ gegen $H_1 : \beta_k \neq 0$ beruht auf dem folgenden Satz.

Satz

Ist ε normalverteilt, so ist

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

t-verteilt mit $n - r - 1$ Freiheitsgraden, wobei $\hat{\sigma}^2_{\hat{\beta}_k}$ das k-te Diagonalelement der geschätzten Kovarianzmatrix

$$\hat{\sigma}^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

ist.

Schätzung, Tests und Prognoseintervalle

- Die Nullhypothese $H_0 : \beta_k = 0$ wird abgelehnt, wenn die Testgröße

$$T = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$$

relativ klein oder relativ groß ist, also wenn

$$\frac{|\hat{\beta}_k|}{\hat{\sigma}_{\hat{\beta}_k}} > t_{1-\alpha/2;n-r-1}$$

- Das symmetrische Konfidenzintervall für β_k mit 95% Sicherheit lautet:

$$\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} \cdot t_{1-\alpha/2;n-r-1}$$

Schätzung, Tests und Prognoseintervalle

- Es soll nun das Ergebnis $Y_0 = Y(\boldsymbol{x}_0)$ für einen bestimmten Wert $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0r})$ der Einflussvariablen prognostiziert werden.
- Wir erweitern $\boldsymbol{x}_0$ um den Wert 1: $\boldsymbol{x}_+ = (1, x_{01}, \ldots, x_{0r})$.
- Der Erwartungswert von Y_0 ist dann

$$\mathsf{E}[Y_0] = \boldsymbol{x}_+ \cdot \hat{\boldsymbol{\beta}}$$

- Die Varianz von $\mathsf{E}[Y_0]$ ergibt sich mittels Fehlerfortpflanzung:

$$\mathsf{var}[\mathsf{E}[Y_0]] = \sigma^2 \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} {\boldsymbol{x}_+}^{\mathrm{T}}$$

Schätzung, Tests und Prognoseintervalle

- Da Y_0 um seinen Erwartungswert mit Varianz σ^2 streut, ergibt sich:

$$\mathrm{var}[Y_0] = \sigma^2 \left[1 + \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \boldsymbol{x}_+{}^{\mathrm{T}}\right]$$

- Das symmetrische Prognoseintervall für Y_0 mit Sicherheit α ist daher gleich:

$$\boldsymbol{x}_+ \cdot \hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2;n-k-1} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_+ \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \boldsymbol{x}_+{}^{\mathrm{T}}}$$

Unterabschnitt: Gewichtete Regression

Gewichtete Regression

- Im allgemeinen Fall kann der Fehlerterm eine beliebige Kovarianzmatrix haben:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \mathbf{V}$$

- Ist $\mathbf{V}$ bekannt, lautet die Zielfunktion:

$$SS = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}), \quad \mathbf{G} = \mathbf{V}^{-1}$$

- Gradient von SS:

$$\frac{\partial SS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Nullsetzen des Gradienten gibt die **Normalgleichungen**:

$$\mathbf{X}^{\mathrm{T}}\mathbf{G}\boldsymbol{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\boldsymbol{\beta}$$

- Die Lösung lautet:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{G}\boldsymbol{Y}$$

Gewichtete Regression

- Kovarianzmatrix der geschätzten Regressionkoeffizienten:

$$\mathsf{Cov}[\hat{\boldsymbol{\beta}}] = \left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}$$

- Kovarianzmatrix der Residuen $\boldsymbol{r}$:

$$\mathsf{Cov}[\boldsymbol{r}] = \mathbf{V} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{G}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$

- Tests und Prognoseintervalle können entsprechend modifizert werden.

Unterabschnitt: Nichtlineare Regression

Nichtlineare Regression

- In der Praxis ist die Abhängigkeit der Ergebnisse von den Regressionskoeffizienten oft nichtlinear:

$$\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \quad \mathsf{Cov}[\boldsymbol{\varepsilon}] = \mathbf{V}$$

- Ist $\mathbf{V}$ bekannt, lautet die Zielfunktion:

$$SS = [\boldsymbol{Y} - \boldsymbol{h}(\boldsymbol{\beta})]^{\mathrm{T}}\mathbf{G}[\boldsymbol{Y} - \boldsymbol{h}(\boldsymbol{\beta})], \quad \mathbf{G} = \mathbf{V}^{-1}$$

- SS kann mit dem Gauß-Newton-Verfahren minimiert werden.
- Dazu wird $\boldsymbol{h}$ an einer Stelle $\boldsymbol{\beta}_0$ linearisiert:

$$\boldsymbol{h}(\boldsymbol{\beta}) \approx \boldsymbol{h}(\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \boldsymbol{c} + \mathbf{H}\boldsymbol{\beta}, \quad \mathbf{H} = \left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0}$$

Nichtlineare Regression

- Die Schätzung von $\boldsymbol{\beta}$ lautet:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{G}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \boldsymbol{c})$$

- $\boldsymbol{h}$ wird neuerlich an der Stelle $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}$ linearisiert.
- Das Verfahren wird iteriert, bis die Schätzung sich nicht mehr wesentlich ändert.
- Viele andere Methoden zur Minimierung von SS sind verfügbar.

Abschnitt 32: Ausgleichsrechnung

Lineare Zwangsbedingungen

Es sei $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ ein Vektor von Beobachtungen der n unbekannten Größen $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$:

$$\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}, \mathsf{Cov}[\boldsymbol{\epsilon}] = \mathbf{V},$$

mit bekanntem $\mathbf{V}$. Außerdem sei bekannt, dass die Größen $\boldsymbol{\beta}$ m unabhängige lineare Zwangsbedingungen erfüllen müssen, die r unbekannte Parameter $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_r)^{\mathrm{T}}$ enthalten:

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}$$

Hierbei ist $\mathbf{A}$ eine Matrix der Dimension $m \times n$, $\mathbf{B}$ eine Matrix der Dimension $m \times r$ und $\boldsymbol{c}$ ein $m \times 1$ dimensionaler Spaltenvektor.

Beispiel (Fortsetzung)

$\boldsymbol{\beta}$ und $\boldsymbol{\vartheta}$ werden nach dem Prinzip der kleinsten Fehlerquadrate geschätzt. Die Zielfunktion ist

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{G} (\boldsymbol{Y} - \boldsymbol{\beta}), \mathbf{G} = \mathbf{V}^{-1}$$

unter der Bedingung:

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}.$$

Das Problem ist bestimmt, wenn $r \leq m$ und widerspruchsfrei, wenn $m \leq n + r$. Die Minimierung von S kann mit Hilfe von m Lagrangefaktoren $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$ erfolgen. Die erweiterte Zielfunktion ist

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \boldsymbol{\lambda}) = (\boldsymbol{Y} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{G} (\boldsymbol{Y} - \boldsymbol{\beta}) + 2\boldsymbol{\lambda}^{\mathrm{T}} (\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} - \boldsymbol{c})$$

mit den Gradienten

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{G}(\boldsymbol{Y} - \boldsymbol{\beta}) + 2\mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda}, \frac{\partial L}{\partial \boldsymbol{\vartheta}} = 2\mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda}, \frac{\partial L}{\partial \boldsymbol{\lambda}} = 2(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} - \boldsymbol{c})$$

Beispiel (Fortsetzung)

Nullsetzen des Gradienten führt auf das folgende lineare Gleichungssystem:

$$\mathbf{G}\boldsymbol{\beta} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{G}\boldsymbol{Y}$$

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0}$$

$$\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\vartheta} = \boldsymbol{c}$$

Mit der zusätzlichen Annahme $m \leq n$ und den Bezeichnungen $\mathbf{G}_{\mathrm{A}} = (\mathbf{AVA}^{\mathrm{T}})^{-1}, \mathbf{V}_{\mathrm{B}} = (\mathbf{B}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}\mathbf{B})^{-1}$ erhält man die folgende geschlossene Lösung:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{VA}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}[\mathbf{I} - \mathbf{BV}_{\mathrm{B}}\mathbf{B}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}](\boldsymbol{c} - \mathbf{A}\boldsymbol{Y})$$

$$\hat{\boldsymbol{\vartheta}} = \mathbf{V}_{\mathrm{B}}\mathbf{B}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}(\boldsymbol{c} - \mathbf{A}\boldsymbol{Y})$$

Die gemeinsame Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\vartheta}}$ lautet:

$$\mathsf{Cov}\left[\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\vartheta}}\end{pmatrix}\right] = \begin{pmatrix}\mathbf{V} + \mathbf{VA}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}(\mathbf{BV}_{\mathrm{B}}\mathbf{B}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}} - \mathbf{I})\mathbf{AV} & -\mathbf{VA}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}\mathbf{BV}_{\mathrm{B}}\\ -\mathbf{V}_{\mathrm{B}}\mathbf{V}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}\mathbf{AV} & \mathbf{V}_{\mathrm{B}}\end{pmatrix}$$

Beispiel (Fortsetzung)

Enthalten die Zwangsbedingungen keine unbekannten Parameter, gilt:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{V}\mathbf{A}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}(\boldsymbol{c} - \mathbf{A}\boldsymbol{Y}), \mathsf{Cov}[\hat{\boldsymbol{\beta}}] = \mathbf{V} - \mathbf{V}\mathbf{A}^{\mathrm{T}}\mathbf{G}_{\mathrm{A}}\mathbf{A}\mathbf{V}$$

Die χ^2-Statistik des Ausgleichs ist definiert durch:

$$\chi^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{G}(\boldsymbol{Y} - \hat{\boldsymbol{\beta}})$$

Ist $\boldsymbol{Y}$ normalverteilt mit der Kovarianzmatrix $\mathbf{V} = \mathbf{G}^{-1}$, so ist χ^2 χ^2-verteilt mit $m - r$ Freiheitsgraden.

Ausgleichsrechnung

Beispiel (Ausgleich der Winkel im Dreieck)

In einem Dreieck werden die Winkel $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ gemessen. Die Werte in Grad seien $\boldsymbol{Y} = (34.26, 86.07, 59.52)$, die Messfehler $\sigma_1 = 0.1, \sigma_2 = 0.08, \sigma_3 = 0.12$. Die gemessenen Winkel können mit der Bedingung $\beta_1 + \beta_2 + \beta_3 = 180$ ausgeglichen werden.

- Es ist $\mathbf{A} = (1, 1, 1), c = 180$, und $\mathbf{V} = \mathrm{diag}(.1^2, .08^2, .12^2)$.
- Weiters ist $\mathbf{G_A} = (\mathbf{AVA}^{\mathrm{T}})^{-1} = 32.468$.
- Somit ergibt sich $\hat{\boldsymbol{\beta}} = \boldsymbol{Y} + \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}(c - \mathbf{A}\boldsymbol{Y}) = (34.31, 86.10, 59.59)$
- $\mathrm{Cov}[\hat{\boldsymbol{\beta}}] = \mathbf{V} - \mathbf{VA}^{\mathrm{T}}\mathbf{G_A}\mathbf{AV} = 10^{-2} \cdot \begin{pmatrix} .68 & -.21 & -.47 \\ -.21 & .51 & -.30 \\ -.47 & -.30 & .77 \end{pmatrix}$.

  Alle Schätzwerte sind negativ korreliert.
- $\chi^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\beta}})^T \mathbf{G}(\boldsymbol{Y} - \hat{\boldsymbol{\beta}}) = 0.846, p = 0.36$. Die Fehler scheinen realistisch zu sein.

Statistische Methoden der Datenanalyse

W. Waltenberger

Einleitung

Mehrdimensionale Zufallsvariable

Einfache Regression

Mehrfache Regression

Ausgleichsrechnung

Nichtlineare Zwangsbedingungen

Sind die Zwangsbedingungen nichtlinear, können sie in der folgenden allgemeinen Form angeschrieben werden:

$$\boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = \mathbf{0}$$

Zunächst wird $\boldsymbol{h}$ an geeigneter Stelle $(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0)$ linearisiert:

$$\boldsymbol{h}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) \approx \boldsymbol{h}(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0) + \mathbf{A}_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \mathbf{B}_0(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) = \mathbf{A}_0\boldsymbol{\beta} + \mathbf{B}_0\boldsymbol{\vartheta} - \boldsymbol{c}_0$$

mit

$$\mathbf{A}_0 = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}}, \mathbf{B}_0 = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\vartheta}}, \boldsymbol{c}_0 = \mathbf{A}_0\boldsymbol{\beta}_0 + \mathbf{B}_0\boldsymbol{\vartheta}_0 - \boldsymbol{h}(\boldsymbol{\beta}_0, \boldsymbol{\vartheta}_0).$$

Die Schätzung folgt wie im linearen Fall. Anschließend wird $\boldsymbol{h}$ an der Stelle $\boldsymbol{\beta}_1, \boldsymbol{\vartheta}_1 = \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\vartheta}}$ neu linearisiert. Es wird wiederholt bis die Zwangsbedingungen genau genügend erfüllt sind.

Teil VIII

Bayesstatistik

Übersicht Teil 8

Abschnitt 33: Einleitung und Grundbegriffe

Einleitung und Grundbegriffe

- In der Bayes-Statistik werden Wahrscheinlichkeiten als rationale Einschätzungen von Sachverhalten interpretiert.
- Neben den Daten werden auch unbekannte Parameter von Verteilungen als **Zufallsvariable** betrachtet.
- Die Vorinformation über die Parameter wird in einer **a-priori-Verteilung** zusammengefasst.
- Die Information in den Daten führt durch Anwendung des Satzes von Bayes zu einer verbesserten Information über die Parameter, die durch die **a-posteriori-Verteilung** ausgedrückt wird.
- Die a-posteriori Verteilung kann zum Schätzen von Parametern, zur Berechnung von Vertrauensintervallen, zum Testen, zur Prognose, zur Modellwahl etc. verwendet werden.

Einleitung und Grundbegriffe

- Durch die Wahl der a-priori-Verteilung wird eine subjektive Komponente in die Datenanalyse eingeführt.
- Dies kann besonders bei sehr kleiner Anzahl der Daten einen merkbaren Einfluss auf das Resultat haben.
- Es gibt jedoch Verfahren zur Konstruktion von a-priori-Dichten, die diesen Einfluss minimieren.
- Liegen viele Daten vor, wird der Einfluss der a-priori-Verteilung vernachlässigbar.
- In diesem Fall liefert die Bayes-Analyse annähernd die gleichen Resultate wie klassische "frequentistische" Verfahren.
- In der a-posteriori-Verteilung ist jedoch mehr Information enthalten als in klassischen Punkt- oder Intervallschätzern.

Einleitung und Grundbegriffe

- Der **Stichprobenraum** $\mathcal{Y}$ ist die Menge aller möglichen Stichproben $\boldsymbol{y}$.
- Der **Parameterraum** Θ ist die Menge aller möglichen Werte des Parameters ϑ.
- Die **a-priori-Verteilung** $\pi(\vartheta)$ beschreibt unsere Einschätzung, ob ein bestimmter Wert ϑ die wahre Verteilung der Daten beschreibt.
- Die **Likelihoodfunktion** $p(\boldsymbol{y}|\vartheta)$ beschreibt die Wahrscheinlichkeit, dass die Stichprobe $\boldsymbol{y}$ beobachtet wird, wenn ϑ wahr ist.
- Wird $\boldsymbol{y}$ beobachtet, kann unsere Einschätzung von ϑ aktualisiert werden, indem mit dem Satz von Bayes die **a-posteriori-Verteilung** berechnet wird:

$$p(\vartheta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\vartheta)\pi(\vartheta)}{\int_\Theta p(\boldsymbol{y}|\vartheta)\pi(\vartheta)\,\mathrm{d}\vartheta}$$

Einleitung und Grundbegriffe

- Die a-posteriori-Verteilung beschreibt unsere Einschätzung von ϑ im Lichte der Daten $\boldsymbol{y}$.
- Kommen neue Daten dazu, kann die a-posteriori-Verteilung als die neue a-priori-Verteilung benützt werden.
- Kann das Integral im Nenner nicht analytisch berechnet werden, muss man zu Monte-Carlo-Methoden greifen.
- Als Punktschätzer von ϑ werden häufig der **Erwartungswert (posterior mean)** oder der **Modus (posterior mode)** von $p(\vartheta|\boldsymbol{y})$ verwendet, gelegentlich auch der **Median (posterior median)**.

Abschnitt 34: A-priori-Verteilungen

A-priori-Verteilungen

- Die Wahl der a-priori-Verteilung kann bei kleinen Stichproben das Ergebnis relativ stark beeinflussen.
- Wir unterscheiden **informative** und **nicht informative** a-priori-Verteilungen.
- Eine informative a-priori-Verteilung beschreibt tatsächliche Information über den unbekannten Parameter ϑ. Diese kann subjektiv oder objektiv sein.
- Eine nicht informative a-priori-Verteilung beschreibt die Abwesenheit solcher Information.
- In jedem Fall sollte die Sensitivität des Resultats bezüglich der a-priori-Verteilung untersucht werden.

A-priori-Verteilungen

- Die a-priori-Verteilung braucht nicht normiert zu sein, da der Normierungsfaktor sich wegkürzt.
- Kann die a-priori-Verteilung normiert werden, wird sie als **proper** bezeichnet, sonst als **improper**.
- Auch eine impropere a-priori-Verteilung kann zu einer properen a-posteriori-Verteilung führen.
- Die Wahl der a-priori-Verteilung wird auch oft von rein rechnerischen Überlegungen beeinflusst.
- Für einige Formen der Likelihoodfunktion gibt es a-priori-Verteilungen, die a-posteriori-Verteilungen aus der gleichen Verteilungsfamilie erzeugen.
- Solche a-priori-Verteilungen heißen **konjugierte** a-priori-Verteilungen.
- In einigen Fällen ist eine nicht informative a-priori-Verteilung auch eine konjugierte a-priori-Verteilung.

A-priori-Verteilungen

- Es gibt verschiedene Vorschläge zur Wahl einer nicht informativen a-priori-Verteilung:
  - Prinzip der **maximalen Entropie**
  - Invarianz unter Parametertransformationen: **Jeffrey's prior**
  - Minimaler Einfluss der a-priori-Verteilung auf die a-posteriori-Verteilung: **Reference prior**
- Für eindimensionales ϑ sind Jeffrey's prior und Reference prior meist identisch; im mehrdimensionalen Fall gilt dies nicht mehr.

A-priori-Verteilungen

Stetige Verteilungen mit maximaler Entropie

- E$[X]$ und var$[X]$ gegeben: Normalverteilung
- $X \geq 0$, E$[\ln X]$ und var$[\ln X]$ gegeben: Lognormalverteilung
- $X \geq 0$, E$[X]$ und E$[\ln X]$ gegeben: Gammaverteilung
- $X \geq 0$, E$[X]$ gegeben: Exponentialverteilung
- $X \in [a, b]$: Gleichverteilung auf $[a, b]$

Diskrete Verteilungen mit maximaler Entropie

- $X \in \{1, \ldots, n\}$: Gleichverteilung auf $\{1, \ldots, n\}$
- $X \in \mathbb{N}$, E$[X]$ gegeben: Geometrische Verteilung

A-priori-Verteilungen

- **Jeffrey's prior** wird so konstruiert, dass die a-priori-Verteilung invariant unter einer Transformation des Parameters ϑ ist.
- Es sei $\tau = h(\vartheta)$ der transformierte Parameter und $\pi_{\mathrm{J}}(\vartheta)$ Jeffrey's prior in ϑ.
- Dann ist die transformierte a-priori-Verteilung in τ gleich

$$\pi(\tau) = \pi_{\mathrm{J}}(\vartheta) \left| \frac{\mathrm{d}\vartheta}{\mathrm{d}\tau} \right|$$

- Die transformierte Fisher-Information ist gleich

$$I(\tau) = I(\vartheta) \left(\frac{\mathrm{d}\vartheta}{\mathrm{d}\tau} \right)^2$$

- Jeffrey's prior wird daher so gewählt:

$$\pi_{\mathrm{J}}(\vartheta) \propto \sqrt{I(\vartheta)}$$

Abschnitt 35: Binomialverteilte Daten

Binomialverteilte Daten

- Wir betrachten ein n mal wiederholtes Bernoulli-Experiment mit Erfolgswahrscheinlichkeit ϑ.
- Die Anzahl Y der Erfolge ist dann binomialverteilt gemäß $\mathrm{Bi}(n, \vartheta)$.
- Wir wollen aus einer Beobachtung y eine Aussage über die Erfolgswahrscheinlichkeit ϑ gewinnen.
- Dazu brauchen wir eine a-priori-Verteilung von ϑ.
- Das Prinzip der maximalen Entropie ergibt die Gleichverteilung $\mathrm{Un}(0, 1) = \mathrm{Be}(1, 1)$ als a-priori-Verteilung:

$$\pi(\vartheta) = I_{[0,1]}(\vartheta)$$

- Die Likelihoodfunktion ist gleich

$$p(y|\vartheta) = \binom{n}{y} \vartheta^y (1 - \vartheta)^{n-y}$$

Binomialverteilte Daten

- Der Satz von Bayes liefert die a-posteriori-Dichte:

$$p(\vartheta|y) \propto \vartheta^y (1-\vartheta)^{n-y} I_{[0,1]}(\vartheta)$$

- Da die a-posteriori-Dichte proportional zur Dichte der Betaverteilung $\mathrm{Be}(y+1, n-y+1)$ ist, muss sie mit dieser identisch sein.
- Der Erwartungswert der a-posteriori-Verteilung ist gleich

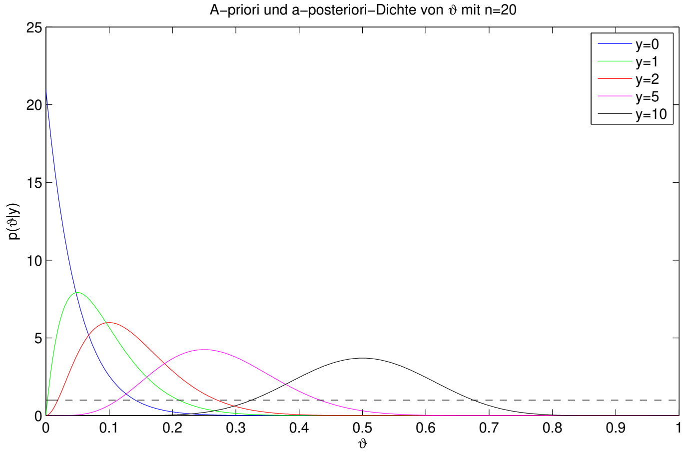
$$\mathsf{E}[\vartheta|y] = \frac{y+1}{n+2}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\vartheta} = \frac{y}{n}$$

und damit der Maximum-Likelihood-Schätzer von ϑ.

Binomialverteilte Daten



Binomialverteilte Daten

- Die Fisher-Information einer Beobachtung ist gleich

$$I(\vartheta) = \frac{1}{\vartheta(1-\vartheta)}$$

- Jeffrey's prior ist also $\mathrm{Be}(0.5, 0.5)$.
- Der Satz von Bayes liefert wieder die a-posteriori-Dichte:

$$p(\vartheta|y) \propto \vartheta^{y-0.5}(1-\vartheta)^{n-y-0.5} I_{[0,1]}(\vartheta)$$

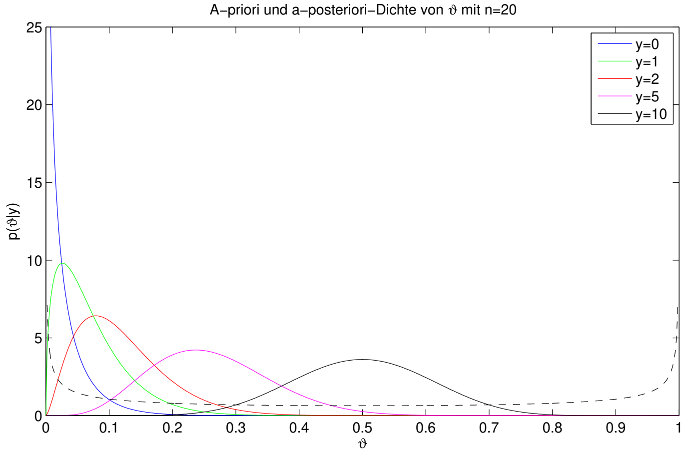
- Da die a-posteriori-Dichte proportional zur Dichte der Betaverteilung $\mathrm{Be}(y+0.5, n-y+0.5)$ ist, muss sie mit dieser identisch sein.
- Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\vartheta|y] = \frac{y+0.5}{n+1}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\vartheta} = \frac{y-0.5}{n-1}$$

Binomialverteilte Daten



Binomialverteilte Daten

- Liegt Vorinformation über ϑ vor, kann diese durch eine geeignete a-priori-Dichte inkludiert werden.
- Die rechnerisch einfachste Behandlung ergibt sich, wenn auch die a-priori-Verteilung eine Betaverteilung ist.
- Sei also

$$\pi(\vartheta) = \frac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a,b)}$$

- Dann ist die a-posteriori-Dichte gleich

$$\begin{aligned} p(\vartheta|y) &\propto \vartheta^{y}(1-\vartheta)^{n-y}\vartheta^{a-1}(1-\vartheta)^{b-1} \\ &= \vartheta^{y+a-1}(1-\vartheta)^{n-y+b-1} \end{aligned}$$

- Die a-posteriori-Verteilung ist wieder eine Betaverteilung, nämlich $\mathrm{Be}(y+a, n-y+b)$.

Binomialverteilte Daten

- Offenbar gibt eine a-priori Betaverteilung mit einer Binomial-Likelihoodfunktion wieder eine a-posteriori Betaverteilung.
- Die Betaverteilung ist daher die **konjugierte** a-priori-Verteilung zur Binomialverteilung.
- Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\vartheta|y] = \frac{a+y}{a+b+n}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\vartheta} = \frac{a+y-1}{a+b+n-2}$$

Binomialverteilte Daten

- Der Erwartungswert der a-posteriori-Verteilung kann als gewichtetes Mittel aus a-priori-Information und Daten angeschrieben werden:

$$\begin{aligned}\mathsf{E}[\vartheta|y] =& \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ =& \frac{a+b}{a+b+n} \times \text{a-priori-Erwartung} \\ +& \frac{n}{a+b+n} \times \text{Mittel der Daten}\end{aligned}$$

- a und b können als "a-priori-Daten" interpetiert werden:

a	Anzahl der Erfolge a-priori
b	Anzahl der Misserfolge a-priori
$a+b$	Anzahl der Versuche a-priori

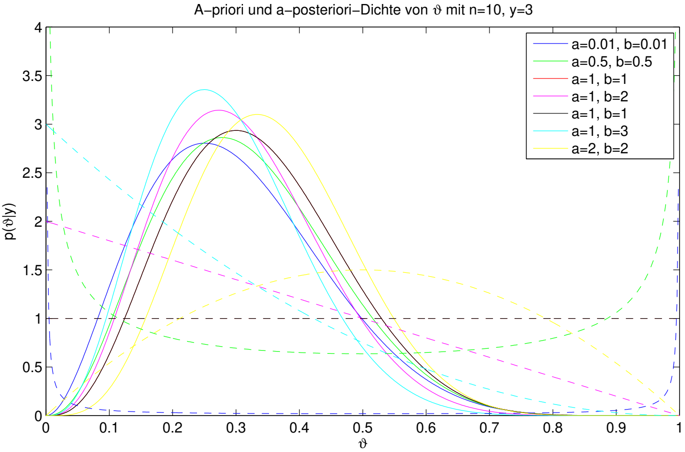
Binomialverteilte Daten

- Wird die Anzahl der Versuche a-priori gleich 0 gesetzt, erhält man **Haldane's prior**:

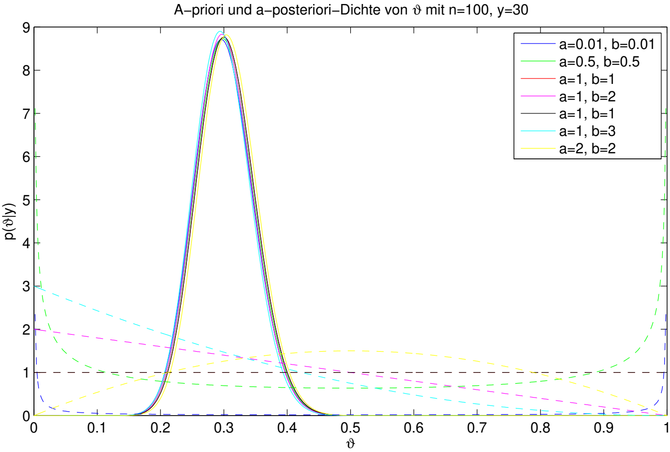
$$\pi_{\mathrm{H}} = \frac{1}{\vartheta(1-\vartheta)}$$

- Haldane's prior kann als $\mathrm{Be}(0,0)$ interpretiert werden, ist allerdings improper.
- Der a-posteriori-Mittelwert ist dann gleich dem ML-Schätzer.
- Paradox: Die a-priori-Verteilung ohne jegliche Vorinformation gibt den Werten $\vartheta = 0$ und $\vartheta = 1$ die größte Wahrscheinlichkeit!

Binomialverteilte Daten



Binomialverteilte Daten



Binomialverteilte Daten

- Wir wollen nun Teilbereiche des Parameterraums Θ konstruieren, die den wahren Wert von ϑ mit großer Sicherheit $1-\alpha$ enthalten.
- Ein solcher Bereich wird ein **Vertrauensbereich** genannt. Er ist meistens ein Intervall (Vertrauensintervall).
- Die einfachste Konstruktion eines Vertrauensintervalls $[\vartheta_1(y), \vartheta_2(y)]$ benützt die Quantile der a-posteriori-Verteilung.
- Das symmetrische Vertrauensintervall lautet:

$$\vartheta_1(y) = q_{\alpha/2}, \quad \vartheta_1(y) = q_{1-\alpha/2}$$

wo q_p das p-Quantil der a-posteriori-Verteilung $p(\vartheta|y)$ ist.

Binomialverteilte Daten

Beispiel

Es sei $y = 4$ die Anzahl der Erfolge in $n = 20$ unabhängigen Alternativversuchen mit Erfolgswahrscheinlichkeit ϑ. Mit der Gleichverteilung als a-priori-Verteilung ist die a-posteriori-Verteilung von ϑ eine $\mathrm{Be}(5, 17)$-Verteilung. Das symmetrische Vertrauensintervall mit $1 - \alpha = 0.95$ hat dann die Grenzen

$$\vartheta_1(y) = \beta_{0.025;5,17} = 0.0822$$
$$\vartheta_2(y) = \beta_{0.975;5,17} = 0.4191$$

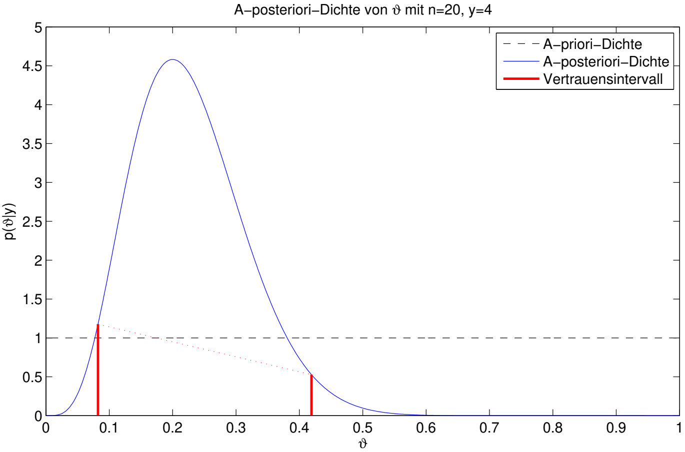
Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\vartheta|y] = \frac{5}{22} = 0.2273$$

Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\vartheta} = \frac{4}{20} = 0.2$$

Binomialverteilte Daten



Binomialverteilte Daten

Beispiel (Fortsetzung)

Mit Jeffrey's prior ist die a-posteriori-Verteilung von ϑ eine $\mathrm{Be}(4.5, 16.5)$-Verteilung. Das symmetrische Vertrauensintervall mit $1-\alpha = 0.95$ hat dann die Grenzen

$$\vartheta_1(y) = \beta_{0.025;4.5,16.5} = 0.0715$$
$$\vartheta_2(y) = \beta_{0.975;4.5,16.5} = 0.4082$$

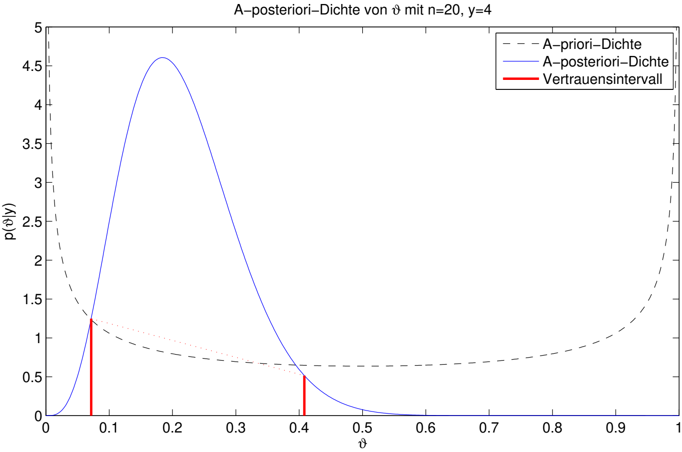
Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\vartheta|y] = \frac{4.5}{21} = 0.2143$$

Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\vartheta} = \frac{3.5}{19.5} = 0.1795$$

Binomialverteilte Daten



Binomialverteilte Daten

- Das symmetrische Vertrauensintervall enthält Werte von ϑ, die geringere a-posteriori-Wahrscheinlichkeit haben als manche Punkte außerhalb des Intervalls.
- Ein Bereich, in dem alle Werte von ϑ höhere a-posteriori-Wahrscheinlichkeit haben als alle Werte außerhalb des Bereichs, heißt ein High-Posterior-Density-Bereich, kurz **HPD**-Bereich.
- Ist die a-posteriori-Dichte unimodal, ist der HPD-Bereich ein HPD-Intervall.
- In diesem Fall erhält man die Grenzen ϑ_1, ϑ_2 des HPD-Intervalls als Lösung des Gleichungssystems:

$$
\begin{aligned}
p(\vartheta_2|y) - p(\vartheta_1|y) &= 0 \\
P(\vartheta_2|y) - P(\vartheta_1|y) &= 1 - \alpha
\end{aligned}
$$

Hier ist $P(\vartheta|y)$ die a-posteriori-Verteilungsfunktion.

Binomialverteilte Daten

- Das Gleichungssystem muss in der Regel numerisch gelöst werden.

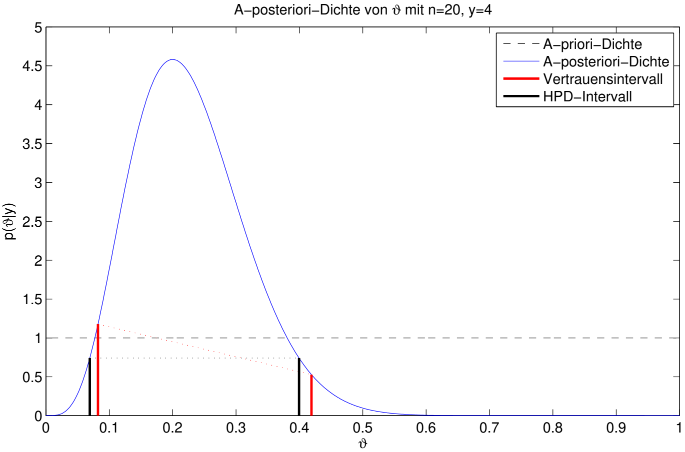
Beispiel (Fortsetzung)

Das HPD-Intervall mit $1-\alpha = 0.95$ hat die Grenzen

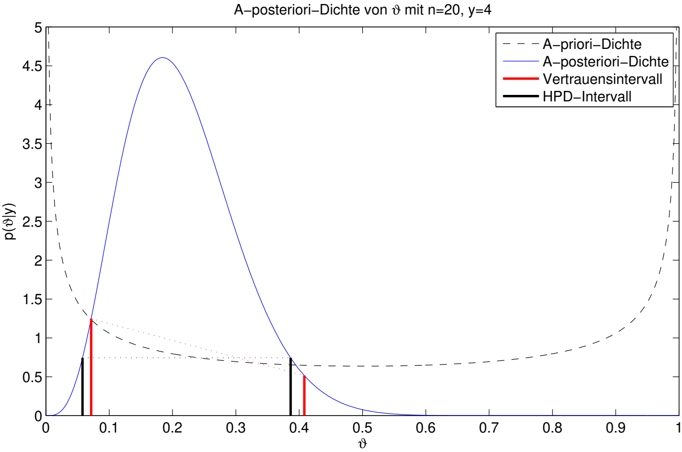
$$\vartheta_1(y) = 0.06921, \quad \vartheta_1(y) = 0.3995$$

Es ist mit einer Länge von 0.3303 kürzer als das symmetrische Vertrauensintervall, das eine Länge von 0.3369 hat.

Binomialverteilte Daten



Binomialverteilte Daten



Binomialverteilte Daten

Beispiel

Ein Alternativversuch wird $n = 20$ mal wiederholt und ergibt keinen einzigen Erfolg ($k = 0$). Was kann man über die Erfolgswahrscheinlichkeit ϑ aussagen?

- Mit der a-priori-Dichte $\pi(\vartheta) = 1$ ist die a-posteriori-Verteilung $\mathrm{Be}(1, 21)$. Der Modus ist gleich 0, der Erwartungswert ist gleich 0.0455. Das HPD-Intervall mit $1 - \alpha = 0.95$ ist gleich $[0, 0.1329]$.
- Mit Jeffrey's prior ist die a-posteriori-Verteilung $\mathrm{Be}(0.5, 20.5)$. Der Modus ist gleich 0, der Erwartungswert ist gleich 0.0238. Das HPD-Intervall mit $1 - \alpha = 0.95$ ist gleich $[0, 0.0905]$.
- Ist bekannt, dass ϑ eher klein ist, kann z.B. $\mathrm{Be}(0.5, 5)$ als a-priori-Verteilung gewählt werden. Die a-posteriori-Verteilung ist dann $\mathrm{Be}(0.5, 25)$. Der Modus ist gleich 0, der Erwartungswert gleich 0.0196, und das HPD-Intervall gleich $[0, 0.0747]$.

Binomialverteilte Daten

Beispiel (Fortsetzung)

- Der Likelihoodschätzer von ϑ ist gleich 0.
- Das einseitige Konfidenzintervall nach Clopper-Pearson ist gleich $[0, 0.1391]$.
- Die Näherung durch Normalverteilung ist nur sinnvoll mit der Korrektur gemäß Agresti-Coull, da sonst das Konfidenzintervall auf den Nullpunkt schrumpft. Die Schätzung ist $\hat{\vartheta} = 0.0833$, das symmetrische Konfidenzintervall ist $[-0.0378, 0.2045]$, das linksseitige Konfidenzintervall ist $[0, 0.1850]$.

Abschnitt 36: Poissonverteilte Daten

Poissonverteilte Daten

- Die Beobachtung eines Poissonprozesses ergibt die Werte $\boldsymbol{y} = y_1, \ldots, y_n$.
- Wir wollen aus den Daten eine Einschätzung der Intensität λ des Prozesses gewinnen.
- Die Likelihoodfunktion lautet:

$$p(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \propto \lambda^{\sum y_i} e^{-n\lambda}$$

- Sie hängt von den Daten nur über $s = \sum y_i$ ab.
- Als nicht informative a-priori-Verteilungen kommen in Frage:
    - Die impropere Dichte $\pi(\lambda) = 1$
    - Jeffrey's prior $\pi_{\mathrm{J}}(\lambda) = \lambda^{-1/2}$, ebenfalls improper

Poissonverteilte Daten

- Mit $\pi(\lambda) = 1$ ist die a-posteriori-Dichte proportional zur Likelihoodfunktion:

$$p(\lambda|s) \propto \lambda^s e^{-n\lambda}$$

- Da die a-posteriori-Dichte proportional zur Dichte der Gammaverteilung $\mathrm{Ga}(s+1, 1/n)$ ist, muss sie mit dieser identisch sein:

$$p(\lambda|s) = \frac{\lambda^s e^{-n\lambda}}{n^{-(s+1)}\Gamma(s+1)}$$

- Der Erwartungswert der a-posteriori-Verteilung ist gleich

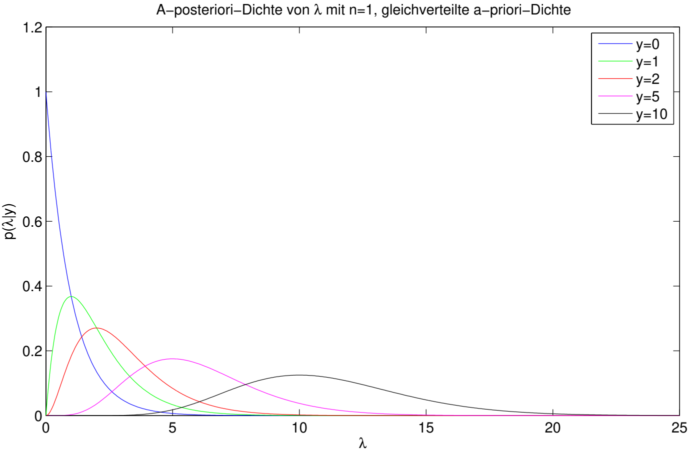
$$\mathsf{E}[\lambda|s] = \frac{s+1}{n}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\lambda} = \frac{s}{n}$$

  und damit der Maximum-Likelihood-Schätzer von λ.

Poissonverteilte Daten



Poissonverteilte Daten

- Mit $\pi(\lambda) = \lambda^{-1/2}$ lautet die a-posteriori-Dichte:

$$p(\lambda|s) \propto \lambda^{s-0.5} e^{-n\lambda}$$

- Da die a-posteriori-Dichte proportional zur Dichte der Gammaverteilung $\mathrm{Ga}(s+0.5, 1/n)$ ist, muss sie mit dieser identisch sein:

$$p(\lambda|s) = \frac{\lambda^{s-0.5} e^{-n\lambda}}{n^{-(s+0.5)} \Gamma(s+0.5)}$$

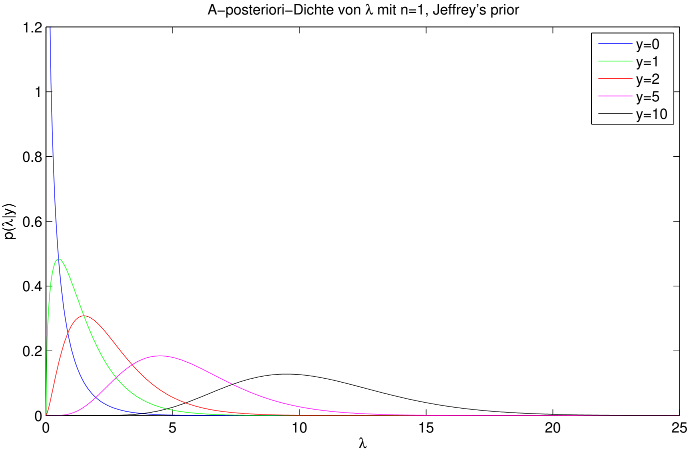
- Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\lambda|s] = \frac{s+0.5}{n}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\lambda} = \frac{s-0.5}{n}$$

Poissonverteilte Daten



Poissonverteilte Daten

- Liegt Vorinformation über λ vor, kann diese durch eine geeignete a-priori-Dichte inkludiert werden.
- Die rechnerisch einfachste Behandlung ergibt sich, wenn die a-priori-Verteilung eine Gammaverteilung ist.
- Sei also

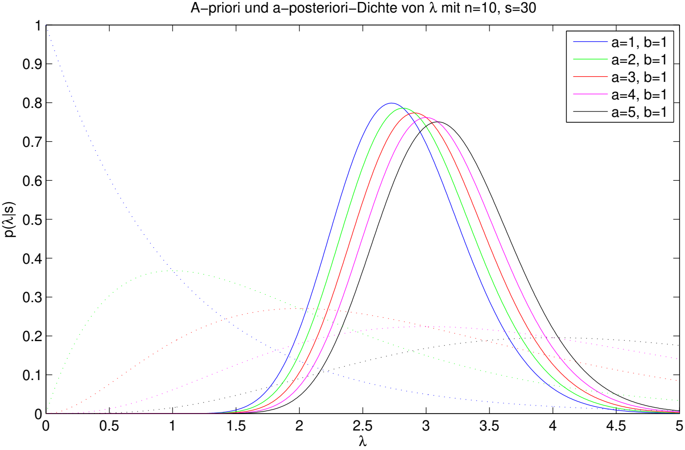
$$\pi(\lambda) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$

  Dies ist die Dichte der Gammaverteilung $\mathrm{Ga}(a, 1/b)$.
- Dann ist die a-posteriori-Dichte gleich

$$\begin{aligned} p(\lambda|s) &\propto \lambda^s e^{-n\lambda} \lambda^{a-1} e^{-b\lambda} \\ &= \lambda^{s+a-1} e^{-(b+n)\lambda} \end{aligned}$$

- Die a-posteriori-Verteilung ist die Gammaverteilung $\mathrm{Ga}(a+s, 1/(b+n))$. Die Gammaverteilung ist also konjugiert zur Poissonverteilung.

Poissonverteilte Daten



Poissonverteilte Daten

- Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\lambda|s] = \frac{a+s}{b+n}$$

- Der Modus der a-posteriori-Verteilung ist gleich

$$\hat{\lambda} = \frac{a+s-1}{b+n}$$

- Der Erwartungswert der a-posteriori-Verteilung kann als gewichtetes Mittel aus a-priori-Information und Daten angeschrieben werden:

$$\begin{aligned}\mathsf{E}[\lambda|s] = \frac{a+s}{b+n} &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{s}{n} \\ &= \frac{b}{b+n} \times \text{a-priori-Erwartung} \\ &+ \frac{n}{b+n} \times \text{Mittel der Daten}\end{aligned}$$

Poissonverteilte Daten

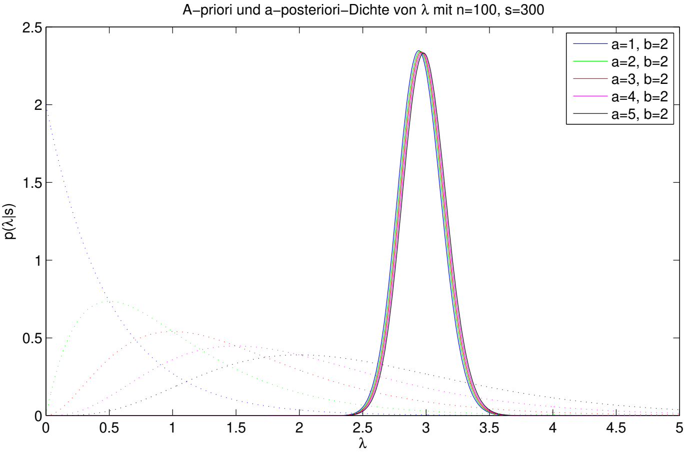
- Die a-priori-Parameter a und b können folgendermaßen interpetiert werden:

a	Summe der Daten a-priori
b	Anzahl der Daten a-priori

- Ist $b \ll n$, dominieren die Daten:

$$\mathsf{E}[\lambda|s] \approx \frac{s}{n}$$

Poissonverteilte Daten



Poissonverteilte Daten

- Vertrauensintervalle $[\lambda_1(s), \lambda_2(s)]$ können leicht mittels der Quantile der a-posteriori-Gammaverteilung konstruiert werden.
- Das symmetrische Vertrauensintervall ist gleich

$$\left[\gamma_{\alpha/2;a+s,1/(b+n)}, \gamma_{1-\alpha/2;a+s,1/(b+n)}\right]$$

- Die einseitigen Vertrauensintervalle sind

$$\left[0, \gamma_{\alpha;a+s,1/(b+n)}\right] \quad \text{bzw.} \quad \left[\gamma_{\alpha;a+s,1/(b+n)}, \infty\right]$$

- HPD-Bereiche können mit numerischen Methoden bestimmt werden.

Poissonverteilte Daten

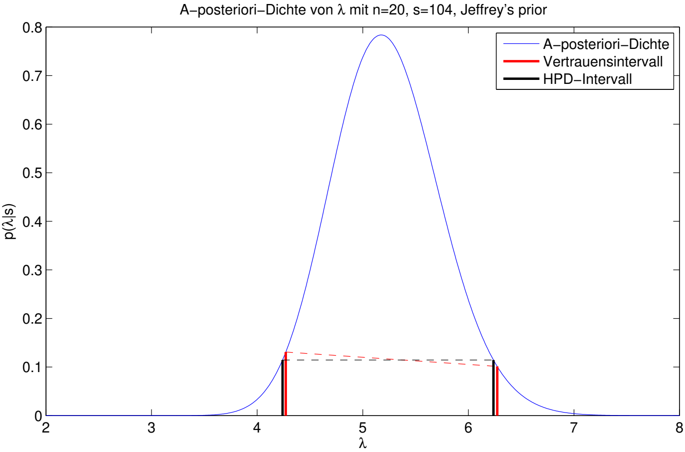
Beispiel

Sie messen die Hintergrundstrahlung in einem Labor über eine Periode von 20 Sekunden. Die Zählwerte sind

$$6, 2, 6, 1, 6, 8, 5, 3, 8, 4, 2, 5, 7, 8, 5, 4, 7, 9, 4, 4$$

Ihre Summe ist gleich $s = 104$. Mit Jeffrey's prior ist die a-posteriori-Verteilung die Gammaverteilung $\mathrm{Ga}(104.5, 0.05)$. Ihre Erwartung ist 5.225, ihr Modus ist 5.1750. Das symmetrische Vertrauensintervall mit $1 - \alpha = 0.95$ ist $[4.2714, 6.2733]$, das HPD-Intervall ist $[4.2403, 6.2379]$. Da die Verteilung fast symmetrisch ist, sind die beiden Intervalle praktisch gleich lang.

Poissonverteilte Daten



Poissonverteilte Daten

Beispiel

Eine Beboachtung aus einer Poissonverteilung hat den Wert $k = 0$. Was kann über den Mittelwert λ ausgesagt werden?

- Mit der a-priori-Dichte $\pi(\lambda) = 1$ ist die a-posteriori-Verteilung $\mathrm{Ga}(1, 1)$. Der Modus ist 0, der Mittelwert ist 1. Das HPD-Intervall mit $1 - \alpha = 0.95$ ist $[0, 2.9957]$.
- Mit Jeffrey's prior ist die a-posteriori-Verteilung $\mathrm{Ga}(0.5, 1)$. Der Modus ist 0, der Mittelwert ist 0.5. Das HPD-Intervall mit $1 - \alpha = 0.95$ ist $[0, 1.9207]$.
- Ist bekannt, dass λ deutlich kleiner als 1 ist, kann z.B. $\mathrm{Ga}(0.5, 0.5)$ als a-priori-Verteilung gewählt werden. Die a-posteriori-Verteilung ist dann $\mathrm{Ga}(0.5, 0.6667)$. Der Modus ist 0, der Mittelwert ist 0.3333, das HPD-Intervall ist $[0, 1.2805]$.
- Der Likelihoodschätzer von λ ist 0.
- Das linksseitige Konfidenzintervall ist $[0, 2.9957]$, ist also identisch mit dem HPD-Intervall mit improperer a-priori-Dichte.

Abschnitt 37: Normalverteilte Daten

Normalverteilte Daten

- Wir betrachten eine normalverteilte Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- Wir wollen eine Einschätzung von Mittelwert und Varianz der Verteilung gewinnen, aus der die Daten stammen.
- Wir nehmen zunächst an, dass die Varianz σ^2 bekannt ist.
- Dann lautet die Likelihoodfunktion:

$$p(\boldsymbol{y}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \mu)^2}{2\,\sigma^2}\right]$$

- In Abwesenheit jeglicher Vorinformation über den tatsächlichen Wert von μ wählen wir die impropere a-priori-Dichte $\pi(\mu) = 1$, die gleichzeitig auch Jeffrey's prior ist.

Normalverteilte Daten

- Dann ist die a-posteriori-Dichte gleich

$$\begin{aligned} p(\mu|\boldsymbol{y},\sigma^2) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i-\mu)^2}{2\,\sigma^2}\right] \\ &\propto \exp\left[-\frac{n(\mu-\bar{y})^2}{2\,\sigma^2}\right] \end{aligned}$$

- Da die a-posteriori-Dichte proportional zur Dichte der Normalverteilung $\mathrm{No}(\bar{y},\sigma^2/n)$ ist, muss sie mit dieser identisch sein.
- Der Erwartungswert der a-posteriori-Verteilung ist gleich

$$\mathsf{E}[\mu|\boldsymbol{y}] = \bar{y}$$

und damit der Maximum-Likelihood-Schätzer von μ.
- Der Modus ist ebenfalls gleich $\bar{y}$.

Normalverteilte Daten

- Liegt Vorinformation über μ vor, kann diese durch eine geeignete a-priori-Dichte inkludiert werden.
- Die rechnerisch einfachste Behandlung ergibt sich, wenn die a-priori-Verteilung ebenfalls eine Normalverteilung ist.
- Sei also

$$\pi(\mu|\sigma^2) = \frac{1}{\sqrt{2\pi}\tau_0} \exp\left[-\frac{(\mu-\mu_0)^2}{2\,\tau_0^2}\right]$$

- Dann ist die a-posteriori-Dichte gleich

$$\begin{aligned} p(\mu|\boldsymbol{y},\sigma^2) &\propto \exp\left[-\frac{n(\mu-\bar{y})^2}{2\,\sigma^2}\right] \exp\left[-\frac{(\mu-\mu_0)^2}{2\,\tau_0^2}\right] \\ &\propto \exp\left[-\frac{a(\mu-b/a)^2}{2}\right] \end{aligned}$$

mit

Normalverteilte Daten

$$a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad \text{und} \quad b = \frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}$$

- Die a-posteriori-Verteilung ist daher die Normalverteilung $\mathrm{No}(\mu_n, \tau_n^2)$ mit

$$\mu_n = \frac{b}{a} = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

- Der Mittelwert μ_n ist das gewichtete Mittel aus der Vorinformation μ_0 und dem Mittelwert der Daten $\bar{y}$, wobei die Gewichte durch die **Präzision**, d.h. die inverse Varianz gegeben sind.

Normalverteilte Daten

- Die Präzision $1/\tau_n^2$ ist die Summe der Präzision der Vorinformation μ_0 und der Präzision des Mittelwerts der Daten $\bar{y}$.
- Je kleiner die Präzision der Vorinformation ist, d.h. je größer τ_0^2 ist, desto kleiner ist der Einfluss der Vorinformation auf die a-posteriori-Verteilung.

Normalverteilte Daten

- Ist die Varianz σ^2 unbekannt, brauchen wir eine gemeinsame a-priori-Dichte für μ und σ^2.
- Die Rechnung ist einfacher, wenn wir statt der Varianz σ^2 die Präzision $\zeta = 1/\sigma^2$ verwenden.
- Eine mögliche Wahl ist die impropere a-priori-Dichte

$$\pi(\mu, \zeta) = \frac{1}{\zeta}$$

- Dann gilt:

$$\begin{aligned} p(\mu, \zeta|\boldsymbol{y}) &\propto \frac{1}{\zeta} \prod_{i=1}^{n} \sqrt{\zeta} \exp\left[-\frac{\zeta}{2}(y_i - \mu)^2\right] \\ &\propto \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \end{aligned}$$

Normalverteilte Daten

- Integration über μ gibt die a-posteriori-Dichte von ζ:

$$p(\zeta|\boldsymbol{y}) \propto \int \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \mathrm{d}\mu$$
$$\propto \zeta^{(n-3)/2} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

- ζ ist daher a-posteriori Gammaverteilt gemäß $\mathrm{Ga}((n-1)/2, 2/\sum(y_i-\bar{y})^2)$. Wegen

$$\sum(y_i-\bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

hängt die Verteilung nur von $s_1 = \sum y_i$ und $s_2 = \sum y_i^2$ bzw. nur vom Stichprobenmittel $\bar{y}$ und der Stichprobenvarianz S^2 ab.

Normalverteilte Daten

- Die a-posteriori-Dichte von σ^2 erhält man durch die Transformation $\sigma^2 = 1/\zeta$:

$$g(\sigma^2) = \frac{p(1/\sigma^2|\boldsymbol{y})}{\sigma^4}$$

- Die Verteilung von σ^2 ist eine inverse Gammaverteilung.

Die Inverse Gammaverteilung $\mathrm{IG}(a, b)$

- Die Dichte der inversen Gammaverteilung lautet:

$$f_{\mathrm{IG}}(x; a, b) = \frac{(1/x)^{a+1} e^{-1/(xb)}}{b^a \Gamma(a)} \cdot I_{[0,\infty)}(x)$$

- Ihre Verteilungsfunktion lautet:

$$F_{\mathrm{IG}}(x; a, b) = 1 - F_{\mathrm{Ga}}(1/x; a, b)$$

- Der Mittelwert ist $1/(b(a-1))$, wenn $a > 1$; der Modus ist $m = 1/(b(a+1))$.

Normalverteilte Daten

- Die durch ζ bedingte a-posteriori-Verteilung von μ erhalten wir aus

$$\begin{aligned} p(\mu|\zeta, \boldsymbol{y}) &= \frac{p(\mu, \zeta|\boldsymbol{y})}{p(\zeta|\boldsymbol{y})} \\ &\propto \zeta^{0.5} \exp\left\{-\frac{\zeta}{2}\left[\sum(y_i - \mu)^2 - \sum(y_i - \bar{y})^2\right]\right\} \\ &\propto \zeta^{0.5} \exp\left[-\frac{n\zeta}{2}(\mu - \bar{y})^2\right] \\ &= \frac{1}{\sigma} \exp\left[-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right] \end{aligned}$$

- Das ist die Normalverteilung $\mathrm{No}(\bar{y}, \sigma^2/n)$.

Normalverteilte Daten

- Schließlich erhalten wir die a-posteriori-Verteilung von μ durch Integration über ζ:

$$p(\mu|\boldsymbol{y}) \propto \int \zeta^{n/2-1} \exp\left[-\frac{\zeta}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \mathrm{d}\zeta$$

$$\propto \frac{1}{\left[\sum(\mu-y_i)^2\right]^{n/2}}$$

- Daraus folgt durch Umformung, dass das Standardscore

$$t = \frac{\mu-\bar{y}}{S/\sqrt{n}}$$

a-posteriori t-verteilt mit $n-1$ Freiheitsgraden ist.

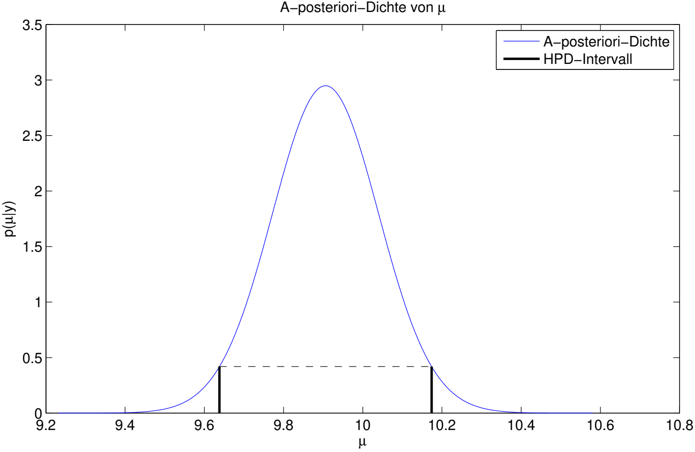
Normalverteilte Daten

- Will man eine eigentliche a-priori-Dichte für ζ, bietet sich die Gammaverteilung an, da diese die konjugierte Verteilung ist.
- Die a-posteriori-Verteilung $p(\zeta|\boldsymbol{y})$ ist dann wieder eine Gammaverteilung.

Beispiel

Von einem normalverteilten Datensatz des Umfangs $n = 100$ aus $\mathrm{No}(10, \sqrt{2})$ sind $\bar{y} = 9.906$ und $S = 1.34981$ gegeben. Die a-posteriori-Dichte von μ ist eine mit $S/\sqrt{n}$ skalierte und um $\bar{y}$ verschobene $\mathrm{t}(n-1)$-Verteilung. Das HPD-Intervalls mit $1 - \alpha = 0.95$ ist gleich $[9.6382, 10.1739]$.

Normalverteilte Daten



Normalverteilte Daten

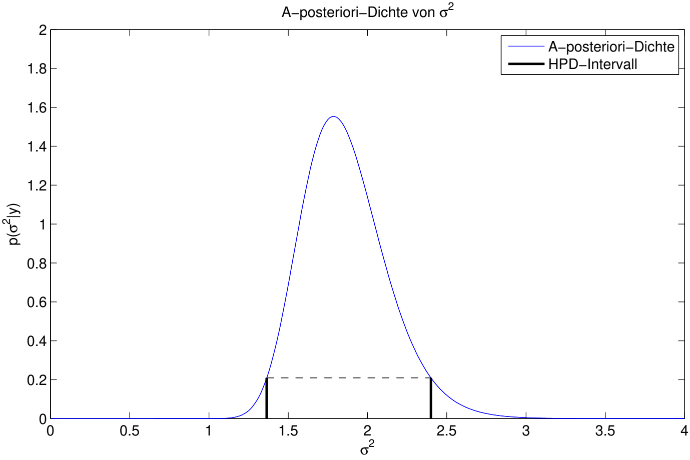
Beispiel (Fortsetzung)

Die a-posteriori-Dichte $p(\sigma^2|\boldsymbol{y})$ der Varianz σ^2 ist die inverse Gammaverteilung

$$\mathrm{IG}((n-1)/2, 2/(S^2(n-1))) = \mathrm{IG}(49.5, 0.0110879)$$

Das HPD-Intervall mit $1-\alpha = 0.95$ ist gleich $[1.3645, 2.4002]$.

Normalverteilte Daten



Abschnitt 38: Exponentialverteilte Daten

Exponentialverteilte Daten

- Wir betrachten eine exponentialverteilte Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n),\ n > 1$.
- Wir wollen eine Einschätzung des Mittelwerts τ der Exponentialverteilung gewinnen, aus der die Daten stammen.
- Die Likelihoodfunktion lautet:

$$p(\boldsymbol{y}|\tau) = \prod_{i=1}^{n} \frac{1}{\tau} \exp\left(-\frac{y_i}{\tau}\right) = \frac{1}{\tau^n} \exp\left(-\frac{\sum y_i}{\tau}\right)$$

- Die Fisher-Information ist gleich $I_\tau = n/\tau^2$; Jeffrey's prior ist daher $\pi_{\mathrm{J}}(\tau) = \tau^{-1}$ und damit improper.
- Die a-posteriori-Verteilung ist dann proportional zu

$$p(\tau|\boldsymbol{y}) \propto \frac{1}{\tau^{n+1}} \exp\left(-\frac{\sum y_i}{\tau}\right) = \frac{1}{\tau^{n+1}} \exp\left(-\frac{s}{\tau}\right)$$

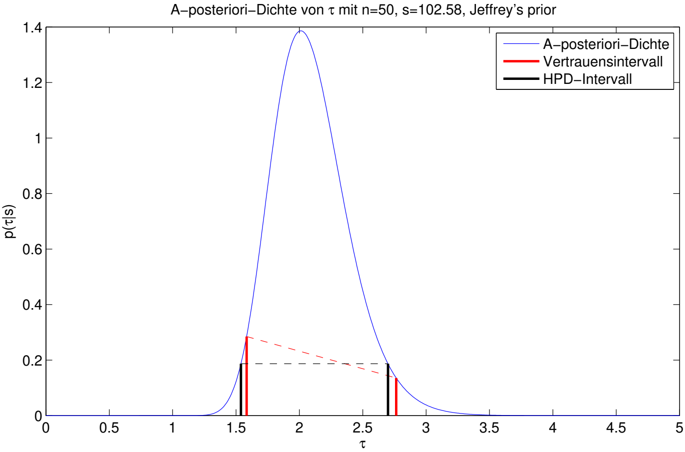
Exponentialverteilte Daten

- Die a-posteriori-Verteilung ist daher die inverse Gammaverteilung $\mathrm{IG}(n, 1/s)$.
- Der Mittelwert ist $s/(n-1)$, der Modus ist $s/(n+1)$.

Beispiel

Es liegt eine exponentialverteilte Stichprobe vom Umfang $n = 50$ vor. Die Summe der Daten ist $s = 102.58$. Der a-posteriori Mittelwert ist 2.0935, der a-posteriori Modus ist 2.0114. Das HPD-Intervall mit $1 - \alpha = 0.95$ ist $[1.5389, 2.6990]$.

Exponentialverteilte Daten



Exponentialverteilte Daten

- Die Likelihoodfunktion kann auch mit der mittleren Rate λ parametrisiert werden:

$$p(\boldsymbol{y}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda y_i) = \lambda^n \exp\left(-\lambda \sum y_i\right) = \lambda^n \exp(-\lambda s)$$

- Jeffrey's prior ist $\pi_{\mathrm{J}}(\lambda) = \lambda^{-1}$.
- Die a-posteriori-Verteilung ist dann proportional zu

$$p(\lambda|\boldsymbol{y}) \propto \lambda^{n-1} \exp(-\lambda s)$$

- Die a-posteriori-Verteilung ist daher die Gammaverteilung $\mathrm{Ga}(n, 1/s)$ mit Mittelwert n/s und Modus $(n-1)/s$. Der Mittelwert ist gleich dem Maximum-Likelihood-Schätzer von λ.
- Die Gammaverteilung ist auch die konjugierte a-priori-Verteilung.

Abschnitt 39: Markov Chain Monte Carlo

Markov Chain Monte Carlo

- Die Folge $X_0, X_1, \ldots$ diskreter Zufallsvariablen heißt **diskrete Markow-Kette**, falls die **Markow-Eigenschaft** zutrifft:

$$W(X_{n+1} = j | X_1 = i_1, \ldots, X_n = i_n) = W(X_{n+1} = j | X_n = i_n)$$

- Die Übergangswahrscheinlichkeiten werden in einer **Übergangsmatrix** $p_{ij}^{(n)}$ zusammengefasst, mit:

$$p_{ij}^{(n)} = W(X_{n+1} = j | X_n = i)$$

$p_{ij}^{(n)}$ ist eine **stochastische Matrix**: die Zeilensummen sind 1.

- In einer **zeitlich homogenen** Markow-Kette gilt $p_{ij}^{(n)} = p_{ij}$ für alle n.
- Sei $\boldsymbol{\pi}^{(0)}$ die Dichte von X_0. Dann ist die Dichte von X_n

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \, p_{ij}^{(n)}$$

Markov Chain Monte Carlo

- Unter bestimmten Umständen (die Kette muss irreduzibel sein, alle Zustände müssen positiv rekurrent sein) hat eine zeitlich homogene Markow-Kette eine **stationäre Verteilung** $\boldsymbol{\pi}$, mit

$$\boldsymbol{\pi} = \boldsymbol{\pi}\, p_{ij}$$

- Offensichtlich ist $\boldsymbol{\pi}$ ein linker Eigenvektor von p_{ij} mit dem Eigenwert $\lambda = 1$.

Beispiel

Einfaches Wettermodell. Zustände: sonnig=1, regnerisch=2. Übergangsmatrix:

$$\mathbf{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}$$

Markov Chain Monte Carlo

Beispiel (Fortsetzung)

Zeitliche Entwicklung, beginnend mit einem sonnigen Tag:

n	$W(1)$	$W(2)$
0	1.0	0.0
1	0.9	0.1
2	0.85	0.15
3	0.825	0.175
4	0.813	0.188
5	0.806	0.194
...	...	...

Markov Chain Monte Carlo

Beispiel (Fortsetzung)

Zeitliche Entwicklung, beginnend mit einem regnerischen Tag:

n	$W(1)$	$W(2)$
0	0.0	1.0
1	0.4	0.6
2	0.6	0.4
3	0.7	0.3
4	0.75	0.25
5	0.775	0.225
...	...	...

Stationäre Verteilung:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix}$$

Markov Chain Monte Carlo

- Eine diskrete Markow-Kette heißt **reversibel** falls die Bedingung der **detailed balance** gilt, d.h. wenn es eine Verteilung p_{ij} gibt für die gilt:

$$\pi_i p_{ij}^{(n)} = \pi_j p_{ji}^{(n)}$$

- In dem Fall ist p_{ij} eine stationäre Verteilung der Kette.
- Für **kontinuierliche** Zufallsvariablen X_i, wird die Übergangsmatrix durch einen **Übergangs-Kernel** $q(x, y)$ ersetzt mit:

$$\int q(x, y) \, \mathrm{d}y = 1$$

- Wenn zur Zeit n der Zustand der Kette gleich x ist, dann ist der Zustand y zur Zeit $n + 1$ gegeben durch $q(x, y)$.

Markov Chain Monte Carlo

- **Markov Chain Monte Carlo (MCMC)** geht umgekehrt vor: gegeben eine Zielverteilung $\pi(x)$, wir suchen nach einem Übergangs-Kernel $q(x, y)$ der $\pi(x)$ als stationäre Verteilung hat.
- Im Allgemeinen ist es nicht möglich, einen derartigen Kernel explizit zu finden. Eine praktische Lösung ist jedoch der **Metropolis–Hastings** Algorithmus.
- Angenommen, dass für gewisse x, y gilt:

$$\pi(x)q(x, y) > \pi(y)q(y, x)$$

  Dann entwickelt sich die Kette von x nach y häufiger als von y nach x.
- Um das zu korrigieren, wird die Häufigkeit von Entwicklungen von x nach y reduziert, indem man eine **Annehmwahrscheinlichkeit** $\alpha(x, y) < 1$ einführt.

Markov Chain Monte Carlo

- Von der detailed balance Bedingung erhalten wir

$$\alpha(x,y) = \min\left[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right]$$

- $\alpha(x,y)$ kann berechnet werden **ohne Kenntnis der Normalisierungskonstante von** $\pi(x)$!

Der Metropolis–Hastings Algorithmus

- Generiere einen Startwert x_0.
- Für i=1,. . . ,N:
  - Generiere einen Vorschlagswert y aus $q(x_{i-1}, y)$ und u aus $\mathrm{Un}(0,1)$.
  - Wenn $u \leq \alpha(x_{i-1}, y)$, setze $x_i = y$, sonst $x_i = x_{i-1}$.
- Verwerfe die ersten n Glieder der Kette („burn-in").

Markov Chain Monte Carlo

- Der M–H Algorithmus kann verwendet werden um aus Dichten mit unbekannter Normalisierung, wie z.B. posteriori-Dichten zu ziehen.
- Wenn $q(x, y) = q(y, x)$, dann ist die Annehmwahrscheinlichkeit

$$\alpha(x, y) = \min \left[1, \frac{\pi(y)}{\pi(x)} \right]$$

- Wenn $\pi(y) > \pi(x)$, dann nimm den Entwicklungsschritt an; ansonsten nimm den Schritt an mit einer Wahrscheinlichkeit von $\pi(y)/\pi(x)$. Dies ist die Basis von **Optimierung durch „simulated annealing"**.
- Wenn $q(x, y) = g(y - x)$, dann ist die Kette ein **random walk**, da $y = x + z$, z beschreibt das stochastiche Rauschen mit einer Verteilung $g(z)$. Wenn $g(x)$ symmetrisch ist, reduziert sich die Annehmwahrscheinlichkeit wiederum wie oben.

Markov Chain Monte Carlo

- Falls $q(x, y) = p(y)$ nicht von x abhängt, wird der Algorithmus auch **Independence Sampler** genannt, und $p(y)$ die **Vorschlagsdichte**. The Annehmwahrscheinlichkeit ist

$$\alpha(x, y) = \min\left[1, \frac{\pi(y)p(x)}{\pi(x)p(y)}\right]$$

Literatur

1. S. Chib and E. Greenberg, Understanding the Metropolis–Hastings Algorithm. The American Statistician 49/4, 1995.
2. L. Tierney, Markov Chains for Exploring Posterior Distributions. The Annals of Statistics 22/4, 1994.
3. W.K. Hastings, Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57/1, 1970.
4. O. Cencic and R. Frühwirth, Data reconciliation with non-normal distributions I: linear constraints. Submitted to Computers & Chemical Engineering.