Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun (Übungen)

Übersicht

- Einführung, Klassifikation vs. Regression, parametrisches und nicht-parametrisches überwachtes Lernen
- Neuronale Netze und Support-Vektor-Maschinen
- Häufungsanalysen, Warenkorbanalyse, Empfehlungen
- Statistische Grundlagen: Stichproben, Schätzer, Verteilung, Dichte, kumulative Verteilung, Skalen: Nominal-, Ordinal-, Intervall- und Verhältnisskala, Hypothesentests, Konfidenzintervalle, Reliabilität, Interne Konsistenz, Cronbach Alpha, Trennschärfe
- Bayessche Statistik, Bayessche Netze zur Spezifikation von diskreten Verteilungen, Anfragen, Anfragebeantwortung, Lernverfahren für Bayessche Netze
- Induktives Lernen: Versionsraum, Informationstheorie, Entscheidungsbäume, Lernen von Regeln
- Ensemble-Methoden, Bagging, Boosting, Random Forests
- Clusterbildung, K-Means, Analyse der Variation (Analysis of Variation, ANOVA), Inter-Cluster-Variation, Intra-Cluster-Variation, F-Statistik, Bonferroni-Korrektur, MANOVA, Discriminant Function Analysis
- Analyse Sozialer Strukturen

Warenkorbanalyse: Kombinatorische Explosion

- Operationen über Potenzmengenverband
  - Verbesserung durch:
    - Berechnung von häufigen Artikelmengen mit Beschneidung des Suchraums (Pruning) …
    - … und anschließender Bestimmung von Assoziationsregeln durch Betrachtungen aller binären Partitionierungen der häufigen Artikelmengen
  - Aber: Join für jede Ebene → Jeder mit jedem
    - Quadratischer Aufwand!
    - Ist das wirklich handhabbar?
    - Nicht wenn viele verschiedene Artikel vorkommen!
- Können wir eventuell nur eine Teilmenge der Daten analysieren?

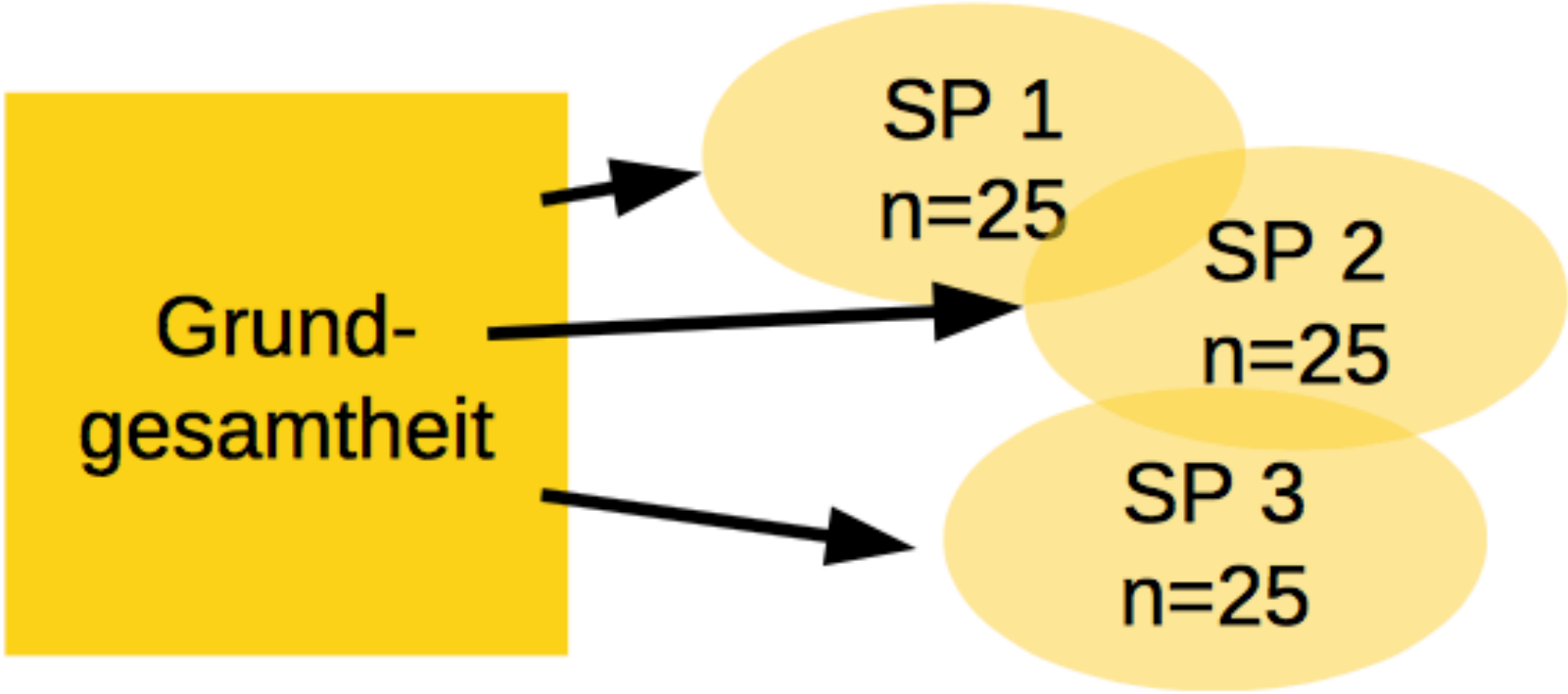
Welche Probleme können vereinfacht werden?

- Bei allgemeiner Warenkorbanalyse erstmal keine
  - Anzahl der Warenkorbtypen zählt
    (verschiedene Artikelmengen)
- Nehmen wir an, die Anzahl der interessierenden Warenkorbtypen sei vorgegeben (und gar nicht so groß)
  - Support- und Konfidenzberechnung immer noch aufwendig:
    Anzahl der Warenkörbe zählt
- Teilmenge der Warenkörbe betrachten?
  - Support- und Konfidenzwerte wie bei Gesamtmenge?
  - Wie groß muss die Teilmenge sein,
    um Aussagen treffen zu können?
  - Welche Teilmenge(n) auswählen?

Betrachtung einer Teilmenge der Daten

Daten, auch Grundgesamtheit oder Population genannt

Erhebung einer Teilmenge der Daten auch Stichprobe (SP) genannt



Betrachtung einer Teilmenge der Daten

Daten, auch Grundgesamtheit oder Population genannt

Teilmenge der Daten, auch Stichprobe (SP) genannt

Definition	Population	Stichprobe
	Grundgesamtheit	Teilmenge einer Grundgesamtheit
Symbole	griechisch	latein
Mittel	μ	$\bar{x}$
Standardab-weichung	σ	s $\hat{\sigma}$

Begriff der statistischen Variable

- Statistische Variable ordnet einem Attribut (Merkmal) einer Erhebungseinheit (Merkmalsträger, Objekt) einen Wert zu (Merkmalsausprägung)
- Beispiele
  - Grundgesamtheit: *Einwohner der Stadt Lübeck*
    - Merkmalsträger: *ein Einwohner*
    - Merkmal: *Geschlecht*
    - Merkmalsausprägung: *männlich*
  - Grundgesamtheit: *Tage eines Untersuchungszeitraums*
    - Merkmalsträger: *ein Tag*
    - Merkmal: *Niederschlagsmenge in Deutschland*
    - Merkmalsausprägung: *1,5 Kubikkilometer*

Statistik

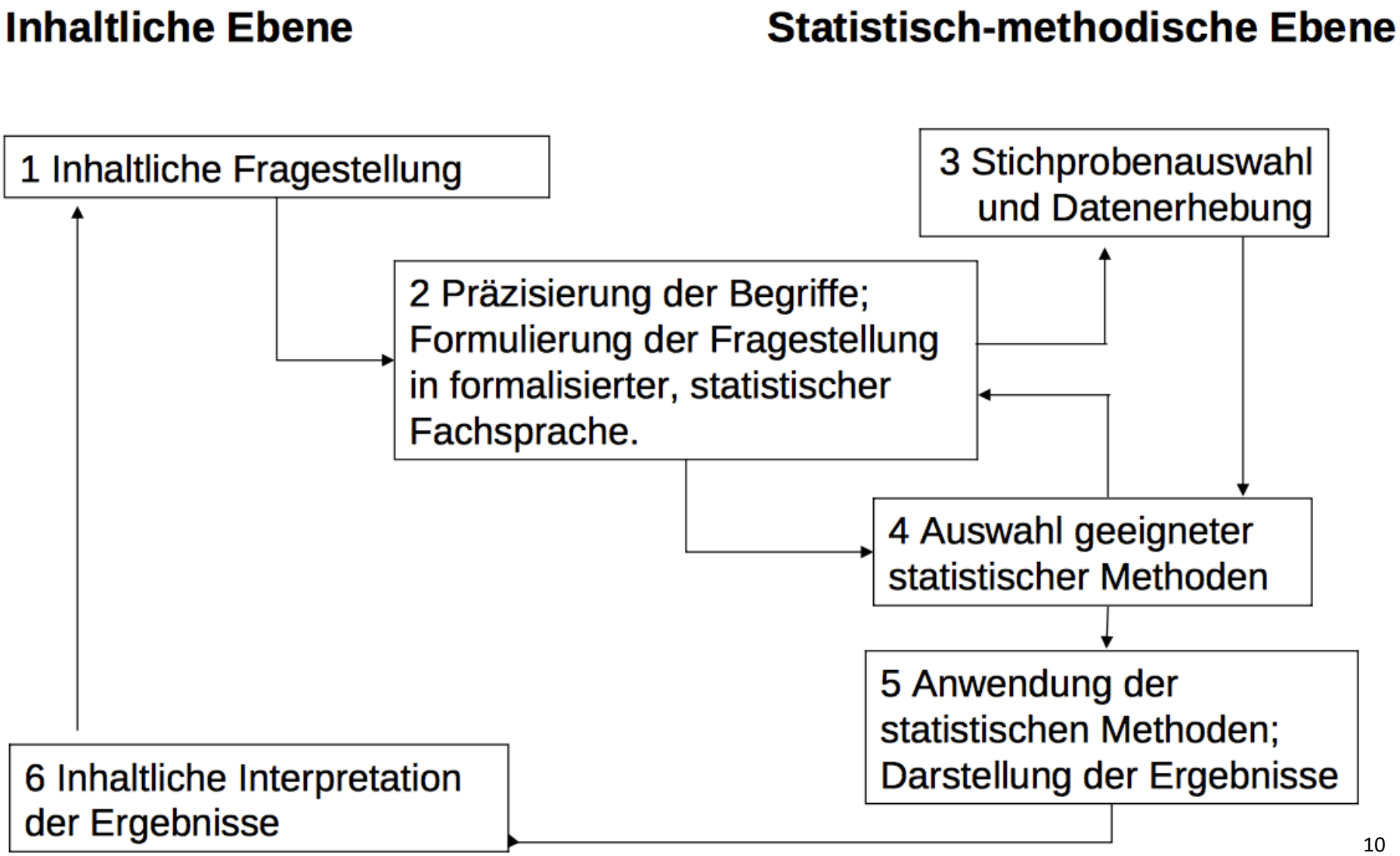
- Deskriptive Statistik
  - Beschreiben von Daten (auch: Teilmenge von Daten)
    - Beispiele: Mittelwert, Varianz, Regression, Korrelation, ...
  - Suchen nach Trends / Mustern
    - Beispiele: Häufige Artikelmengen, Assoziationsregeln
- Induktive Statistik
  - Ziel: Verallgemeinerung der Beschreibung einer Teilmenge der Daten auf Grundgesamtheit
  - Rückschlüsse auf Grundgesamtheit/Population durch Erhebung einer "repräsentativen" Stichprobe

"Repräsentativ"

- Durch Aussagen über Stichprobe kann auf Eigenschaften der Grundgesamtheit geschlossen werden
- Elemente zufällig aus Grundgesamtheit nehmen?
- Größe der Stichprobe sollte ausreichend sein
  - Wir kommen später darauf zurück

- Zunächst: Kein formal definiertes Konzept, basiert je nach Anwendung in vielen Fällen eher auf plausiblen Argumenten

Ablauf systematischer Untersuchungen



Planung von Auswerte-Untersuchungen

- Welche Stichprobeneinheit soll verwendet werden?
  - Welche Skalierung/Normalisierung der Daten?
- Welches räumliche Probennahmemuster verwenden?
  - Welche Aufteilung einer Fläche zur Beprobung?
- Welches zeitliche Probennahmemuster verwenden?
  - Was sind angemessene Intervalle?

Erhebung von Stichproben

- Verbundene Stichproben
  - z.B. wiederholte Messungen am gleichen Versuchsobjekt
  - Stichprobe zu einem Zeitpunkt kann Einfluss auf Stichprobe eines anderen Zeitpunkts haben
- Unverbundene Stichproben
  - Stichproben haben keinen Einfluss aufeinander
  - z.B. unterschiedliche Populationen, Vergleich unterschiedlicher Objekte

Merkmale / Variablen

Objekt/Merkmalsträger	Individuum, an dem interessierende Größen erhoben werden.	Person, die einen Fragebogen ausfüllt
Merkmal	Die interessierende Eigenschaft des Objektes	Qualitativ: Geschlecht, Religionszugehörigkeit Quantitativ: Alter, Einkommen
Merkmalsausprägung	Mögliche Werte eines Merkmals	Kategorien; Zahlenwerte

Experimente werden normalerweise gestaltet, um den Einfluss eines oder mehrerer Faktoren auf eine Variable zu untersuchen

Systematischer Fehler/Trend (Bias)

- Auftretender, meist störender systematischer Effekt mit einer Grundtendenz, so dass Werte von den wahren Ergebnissen abweichen
- Beispiele
  - Schätzung von Fischpopulationen mit Netzen einer bestimmten Maschenweite: kleine Fische können immer entkommen
  - Fangen von Säugetieren: manche Individuen sind "trap happy", manche sind "trap shy"

Lagemaße - Mittelwerte

- Arithmetisches Mittel

$$\bar{x}_{\text{arithm}} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Geometrisches Mittel

$$\bar{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Das geometrische Mittel zweier Zahlen a und b liefert die Seitenlänge eines Quadrates, das den gleichen Flächeninhalt hat wie das Rechteck mit den Seitenlängen a und b

Relevant u.a. für logarithmierte Daten, z.B. Populationswachstum

$$\log_a \bar{x}_{\text{geom}} = \frac{1}{n}\sum_{i=1}^{n} \log_a x_i,$$

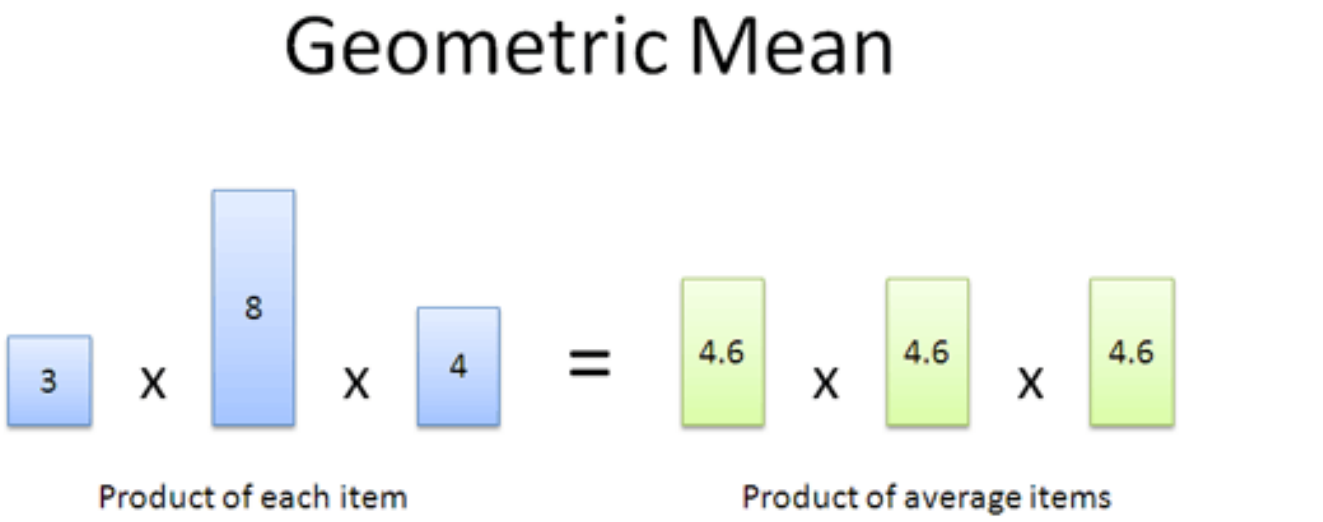
- Harmonisches Mittel

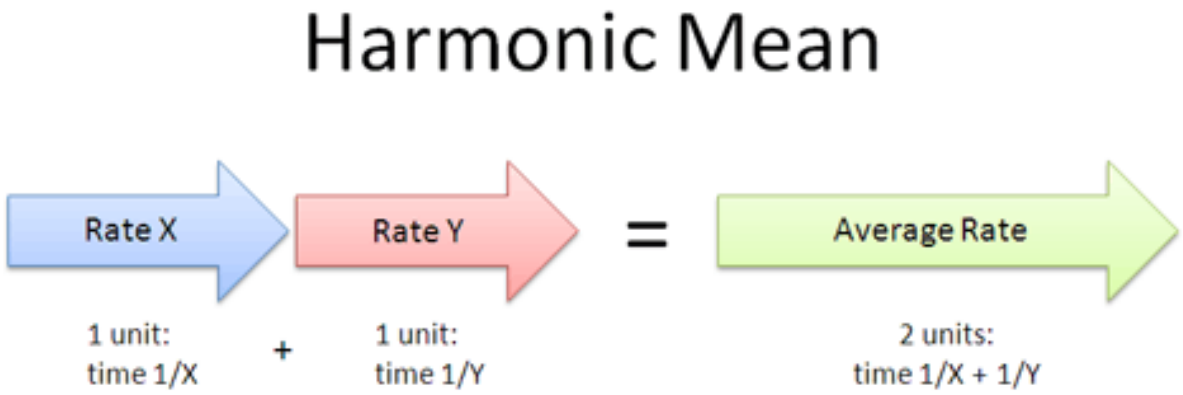
$$\bar{x}_{\text{harm}} = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

$$\frac{1}{\bar{x}_{\text{harm}}} = \frac{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}{n}$$

- $\min(x_1, \ldots, x_n) \leq \bar{x}_{\text{harm}} \leq \bar{x}_{\text{geom}} \leq \bar{x}_{\text{arithm}} \leq \max(x_1, \ldots, x_n)$

Visualisierung





Weitere Lagemaße

- Median (der Wert, der bei einer Auflistung von Zahlenwerten in der Mitte steht)

  4, 1, 37, 2, 1 → Median = 2 (1, 1, 2, 4, 37)

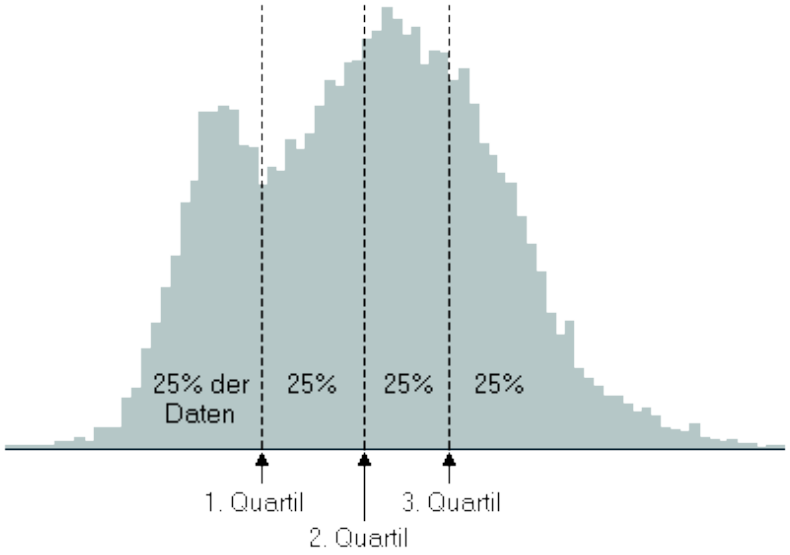
- Modalwert

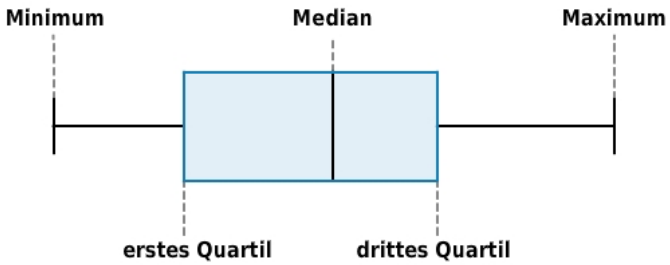
  2, 2, 3, 5, 5, 5, 9, 9, 15

- Quantil, Quartil

  Geordnete Reihe der Merkmalsaus-prägungen wird in gleichgroße Teile zerlegt

  Kumulierte Häufigkeiten





Anwendungen

Name & Meaning	Formula / Example	Used for
Arithmetic Mean [average]	$\frac{sum}{size} = \frac{a+b+c}{3}$	Most situations ("average item")
Median [middle value]	Middle of sorted list (2 middles? Average 'em)	Wildly varying samples (houses, incomes)
Mode [most popular]	Most popular value	No compromises (winner takes all)
Geometric Mean [average factor]	$\sqrt[3]{abc}$	Investments, growth, area, volume
Harmonic Mean [average rate]	$\frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$	Speed, production, cost

Datentypen

Skalenniveaus	*mögliche Aussagen*	*mögliche Methoden (Beispiel Lage)*	*Beispiele*
Nominal (keine Ordnung der Daten möglich)	**1. Gleichheit und Ungleichheit** (=, #) können festgestellt werden)	Häufigkeiten, relative Häufigkeiten, Modalwert	Lieblingszeitungen Geschlecht Studienrichtungen
Ordinal (größenmäßige Ordnung möglich, aber Abstände ohne Aussagekraft)	1 Gleichheit und Ungleichheit **2. Rangreihung (<, >, =)**	dazu: z.B. kumulierte Häufigkeiten, Median	Beliebtheitsrangliste soziale Schichten
Intervall (Abstände können interpretiert werden, nicht aber das Verhältnis von Größen)	1. Gleichheit und Ungleichheit 2. Rangreihung **3. Gleichheit der Unterschiede**	dazu: u.a. arithmetisches Mittel	Intelligenzquotient Temperatur
Verhältnis (die Ausprägungen haben einen absoluten Nullpunkt; das Verhältnis kann interpretiert werden)	1. Gleichheit und Ungleichheit 2. Rangreihung 3. Gleichheit der Unterschiede **4. Proportionalität** $x_{11}=3*x_{12}$	dazu: u.a. geometrisches Mittel	Alter Preis Größe Nahrungswert in Kalorien Inflation

Informationsgehalt

Metrische Variablen

- Intervall- und Verhältnisskala warden oft zur sog. Kardinalskala zusammengefasst.
- Merkmale auf dieser Skala werden dann als metrisch bezeichnet

Kategoriale Variablen

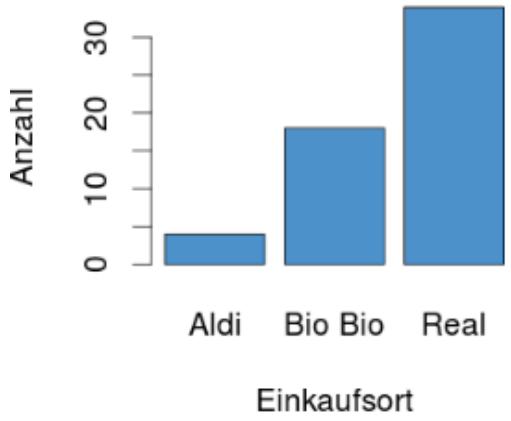
- Nominalskalierte Variablen
- Ordinalskalierte Variablen
- Metrische Variablen, die nur wenige Ausprägungen haben (nicht von allen Autoren unterstützt)
- Variablen, die durch Kategorisierung aus ordinalskalierten oder metrischen Variablen entstanden sind (Beispiel: Variable „Einkommen“ mit den Kategorien „500-999 €“, „1000-1499 €“ usw.)

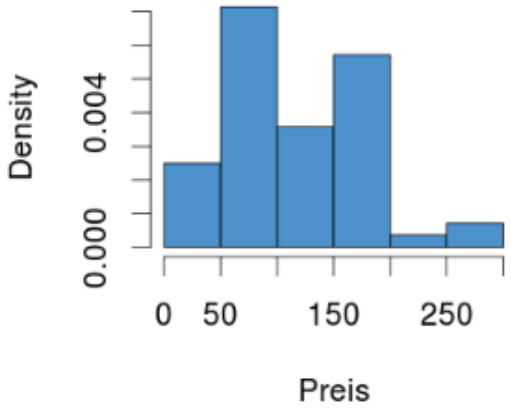
Streuungsmaße

- Spannweite
  - Maximale Differenz zwischen zugrunde liegenden Daten
  - Mindestens Ordinaldaten notwendig
- Varianz
  - Mittlere quadratische Abweichung der einzelnen Datenwerte vom arithmetischen Mittelwert
  - Einheiten quadriert
- Standardabweichung
  - Als Standardabweichung bezeichnet man die Wurzel aus der Varianz
  - Streuungsmaß besitzt dieselbe Einheit wie die Daten und der Mittelwert

Darstellung von Dateneigenschaften

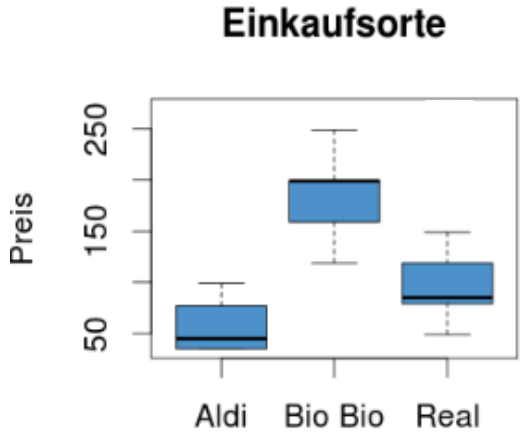
Säulendiagramm

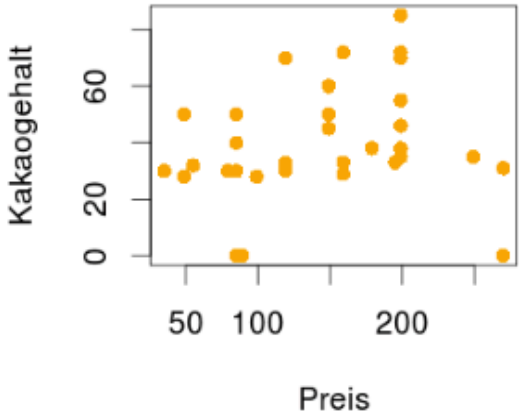




Histogramm

Boxplot





Scatterplot

Darstellung von Daten

Barplot/Säulendiagramm/Balkendiagramm
- Nominale und ordinalskalierte Variablen: Anzahl

Histogramm
- Ordinalskalierte oder metrische Variablen

Scatterplot
- Für 2 Variablen
- Normalerweise metrische Variablen

Boxplot
- Metrische Variablen, die verschiedenen Kategorien angehören können.

Relative Häufigkeiten

- Histogramm: Zähler für Anzahl von Ausprägungen
  - Häufigkeitsverteilung
- Normierung der Anzahlen auf [0, 1] (Skalierung) ergibt relative Häufigkeiten
- Verteilung meist in Bezug auf relative Häufigkeiten betrachtet

Verteilungen

- Einige Verteilungen, die natürlich vorkommen
  - Exponentialverteiltung (hatten wir schon)
    - Städte (nominal) Anzahl Einwohner (metrisch)
    - Über Einwohner wird Städte sortierbar
  - Binomialverteilung
  - Normalverteilung

- Beschreibung durch Funktion

$$f: \text{Grundmenge} \longrightarrow [0, 1]$$

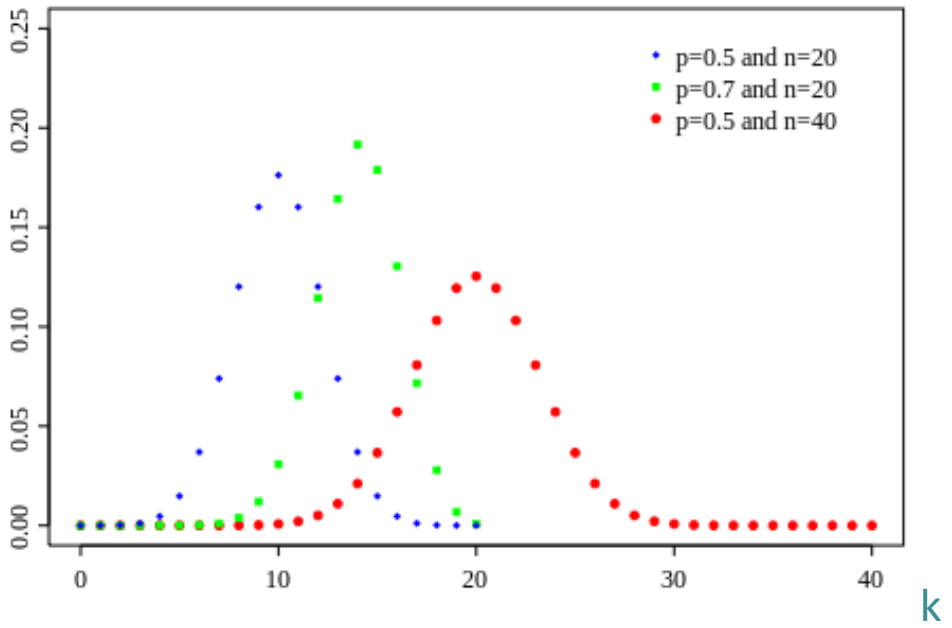
Binomialverteilung

- Beschreibt Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben:
  „Erfolg“ oder „Misserfolg“

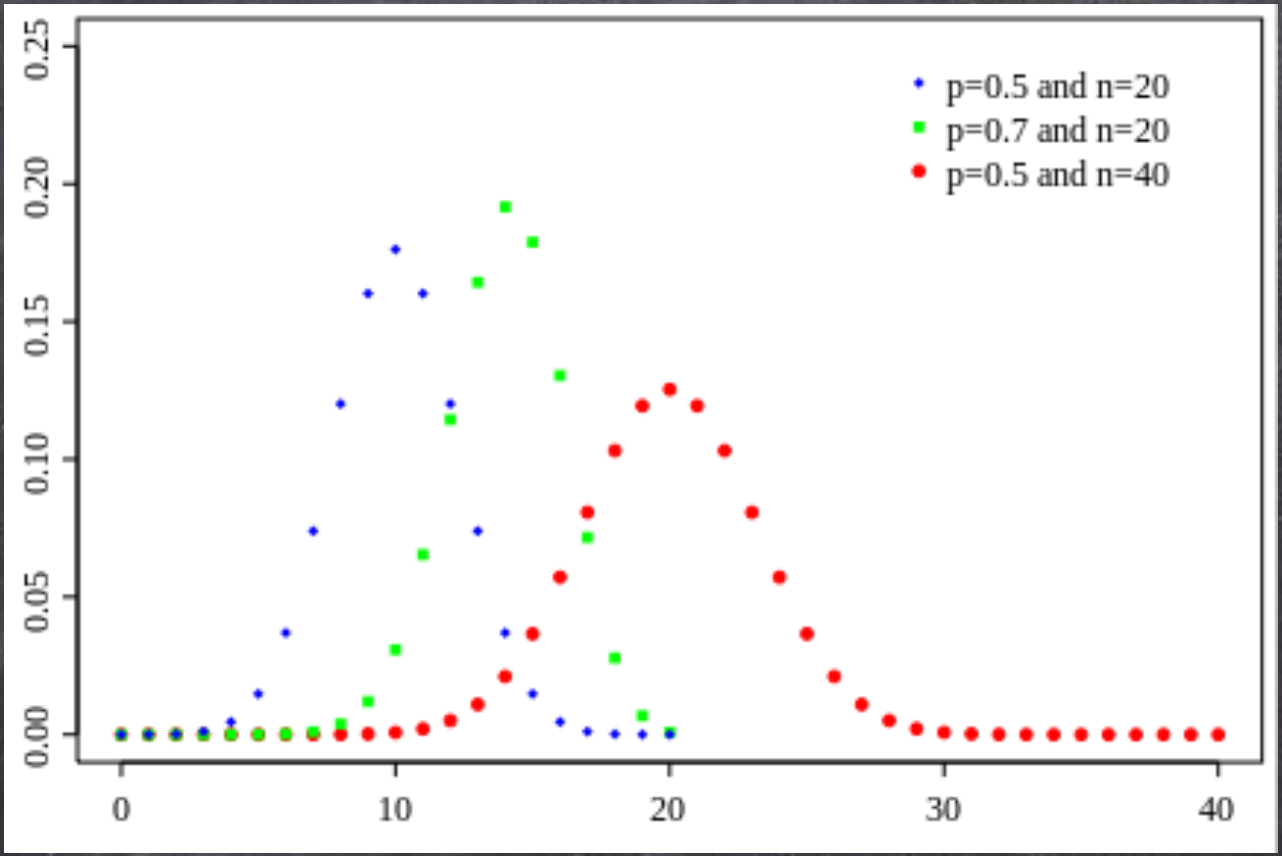
- n = #Versuche
  p = #erfolgr. Vers. / #Versuche

- Beschreibung
  der relativen Häufigkeit, genau k Erfolge zu erzielen, als Funktion $B_{p,n}(k)$

$B_{p,n}(k)$



Aufgabe



Wie bestimmen wir die relative Häufigkeit, bis zu k Erfolge zu erzielen?

$\Sigma_{i=0}^{k} B_{p,n}(i)$

Kumulierte Häufigeit

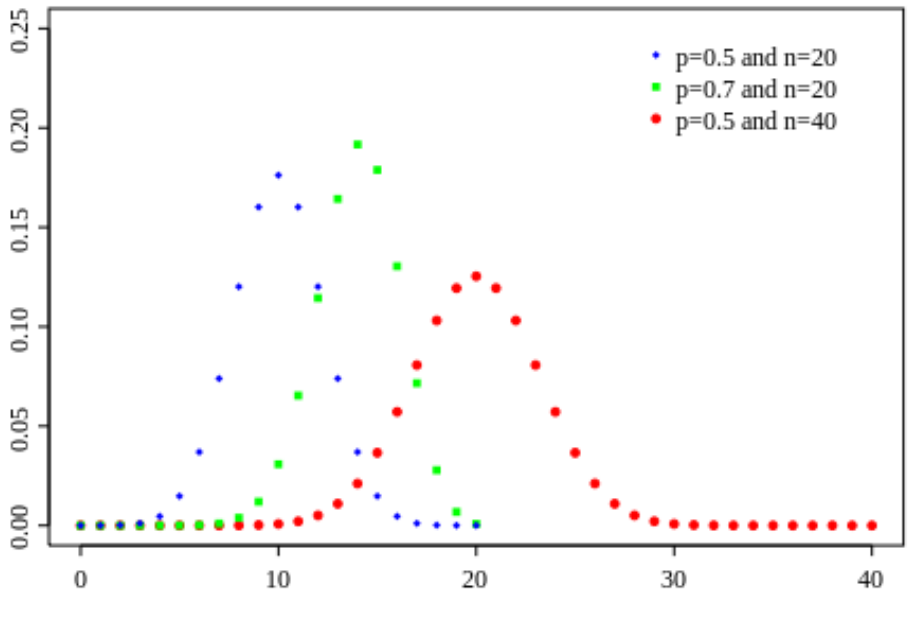
Binomialverteilung

- Beschreibt Anzahl der Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, die jeweils genau zwei mögliche Ergebnisse haben:
  „Erfolg“ oder „Misserfolg“

- n = #Versuche
  p = #erfolgr. Vers. / #Versuche
- Beschreibung
  der relativen Häufigkeit, genau k Erfolge zu erzielen, als Funktion $B_{p,n}(k)$

- Es gilt: $\sum_{i=0}^{n} B_{p,n}(i) = 1$

$B_{p,n}(k)$

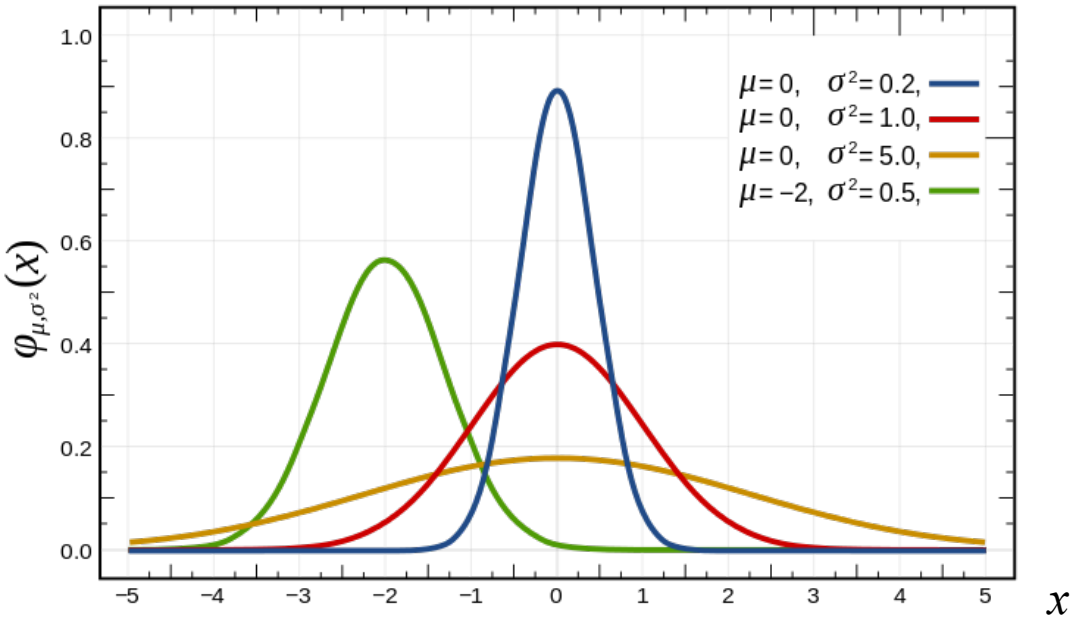


Normalverteilung

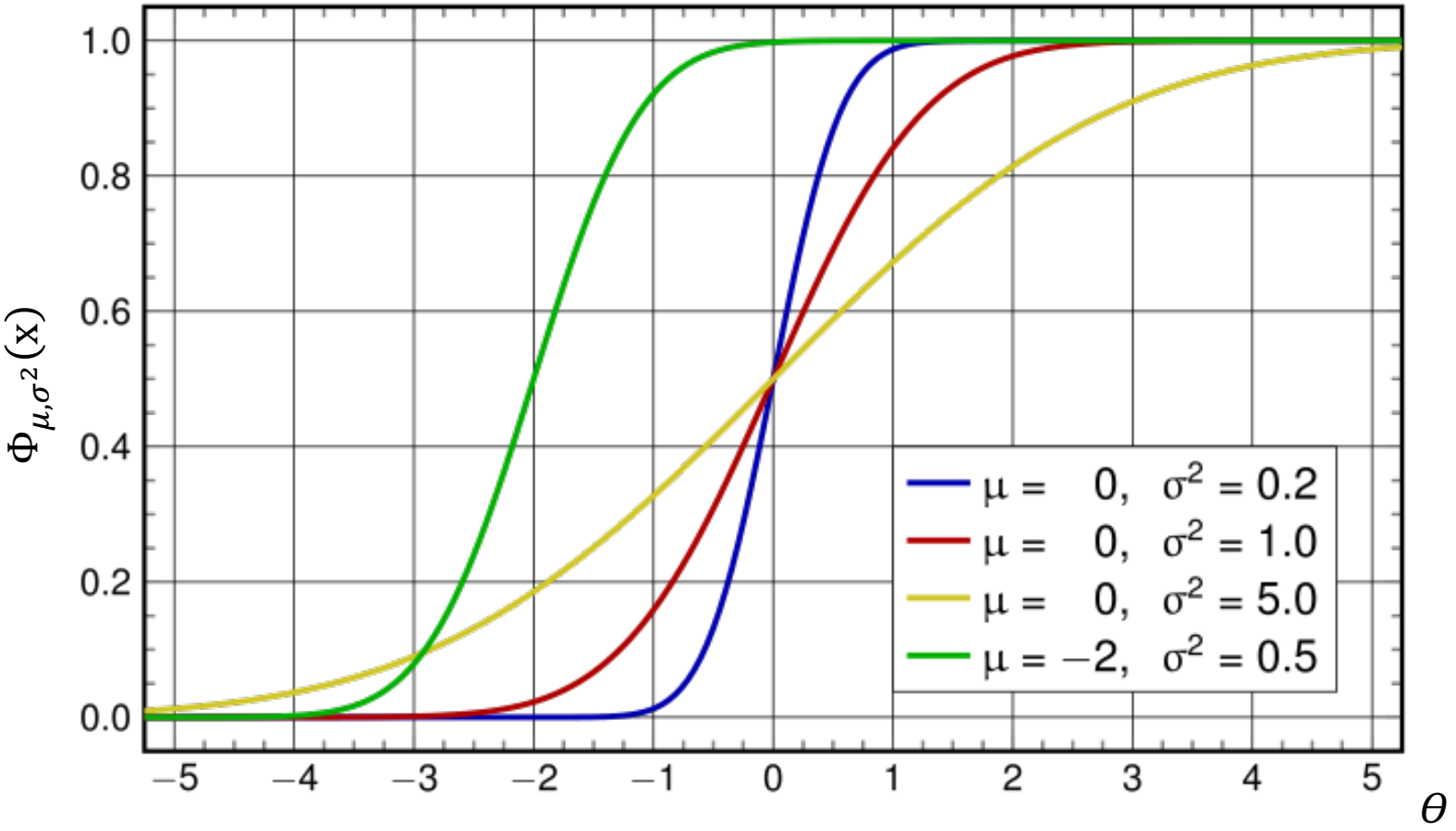
- Grundmenge: $\mathbb{R}$
- Lagemaß: Mittelwert μ
- Streuungsmaß: Varianz σ^2
- Funktion für Häufigkeitsverteilung wird im kontinuierlichen Fall Dichtefunktion genannt:

Bei einer Normalverteilung sind Mittelwert und Median gleich

$$\int_{-\infty}^{\infty} \varphi_{\mu,\sigma^2}(x)dx = 1$$



Verteilung für relative Häufigkeit von $\varphi_{\mu,\sigma^2}(x) \leq \theta$



$\Phi_{\mu,\sigma^2}(x)$ = Fläche unter der Häufigkeitsverteilung $\varphi_{\mu,\sigma^2}(x)$ von $-\infty$ bis θ

→ sog. Verteilungsfunktion

Von relativen Häufigkeiten zu Wahrscheinlichkeiten

- Übergang von relativen Häufigkeiten auf sog. Wahrscheinlichkeiten als Eigenschaften des Daten erzeugenden Prozesses
  - Johann Bernoulli (1667-1748) und Pierre Laplace (1749-1822)
  - Beispiel: Wahrscheinlichkeit, männlich zu sein, wenn man ≥400,000 Euro verdient
  - Aber: Auch bei großen Datenmengen wird die Wahrscheinlichkeit für eine Eigenschaft des die Daten generierenden Prozesses offensichtlich durch #günstige Fälle / #mögliche Fälle nur sehr grob geschätzt
- Betrachtung des Grenzfalles: #mögliche Fälle $\longrightarrow \infty$
  - Richard von Mises (ca. 1883-1953)
- Weitere Entwicklung ab 1930 durch Andrei Kolmogorov

Wahrscheinlichkeits- vs. Dichtefunktion

- Wahrscheinlichkeitsfunktion
  - Wahrscheinlichkeit für jede Merkmalsausprägung
- Geht nicht bei dichter Grundmenge
  - Wahrscheinlichkeit für jeden einzelnen Wert: 0
- Daher in diesem Fall: Dichtefunktion
- Verwendung der Dichte in Verteilungsfunktion
  - Bestimmung der Wahrscheinlichkeit, dass ein gewisses Ereignis höchstens x mal auftritt
  - Verteilungsfunktionen für die Normalverteilung

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

  - Geht für x⟶∞ gegen 1

Normalverteilung

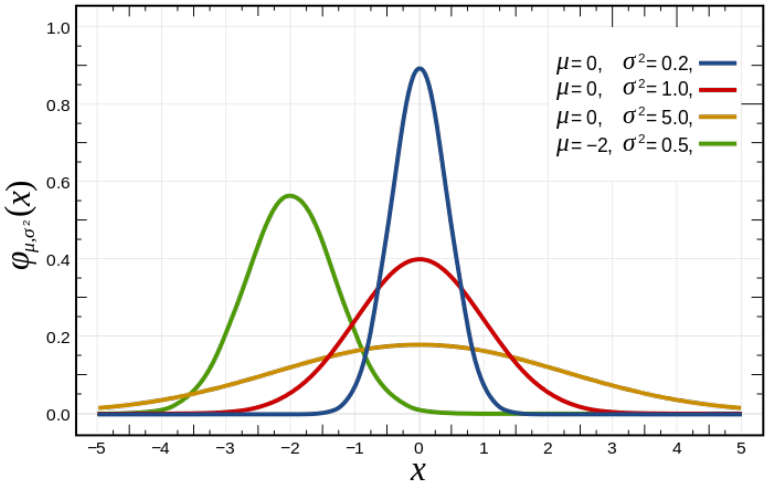
- Dichtefunktion

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

- Standardnormalverteilung:
  $\mu = 0$ und $\sigma = 1$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,\mathrm{d}t.$$



$\phi(x_0)$ Likelihood von x_0

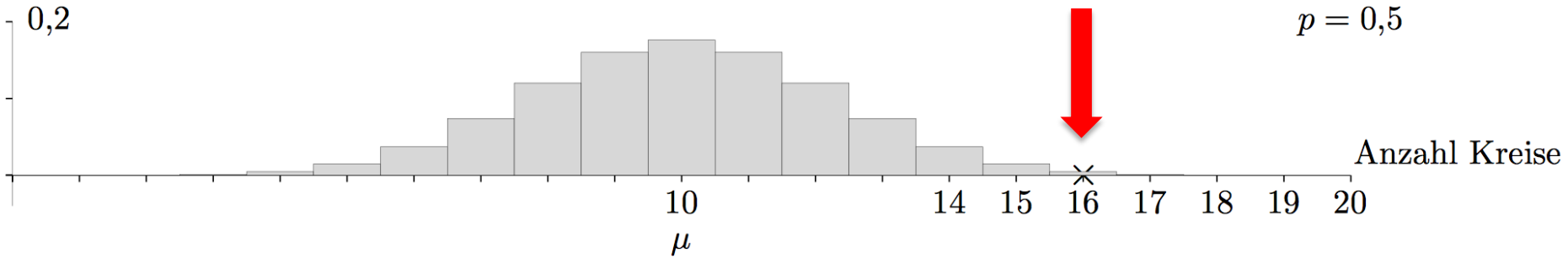
Wahrscheinlichkeit, dass beim Ziehen aus der Grundgesamtheit ein Wert ≤ x auftritt

Hypothesentest

- Vermutung: Küken können Körner schon erkennen und müssen die Form des Futters nicht erst lernen
- Experiment:
  - Kreise und Dreiecke je zur Hälfte zum Picken vorgegeben (sagen wir 20 Objekte insgesamt)
  - Wenn Vermutung wahr, sollte $p_{Kreis} >> 0.5$ gelten
- Hypothese H_0:
  - Küken unterscheiden nicht zwischen Kreisen und Dreiecken, $p_{Kreis} = 0.5$, Mittelwert des Experiments sollte 10 sein, Varianz sei 2
- Hypothese H_1:
  - Küken unterscheiden zwischen Kreis und Dreieck, sie picken häufiger in einen Kreis

Experiment unter Normalverteilungsannahme

- Wenn Vermutung falsch, (also H_0 wahr),
  dann p_{Kreis}=0.5, Mittelwert von μ=10, σ^2=2 (empirisch)

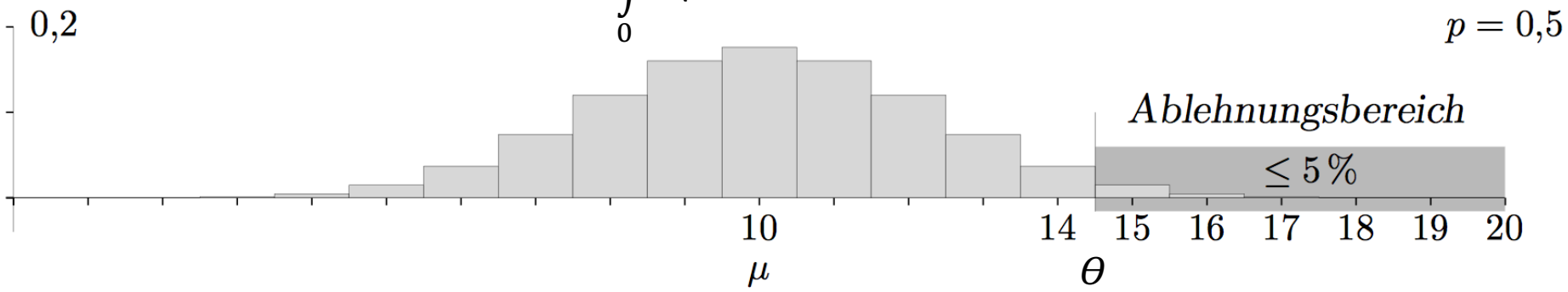


- Ausgang des Experiments:
  - Ausgang: Küken pickt im Mittel 16 mal auf Kreis
- Annahme: Wir wollen die Wahrscheinlichkeit minimieren, H_0 abzulehnen, obwohl sie wahr ist.

Ablehnungsbereich

- Ziel: Wahrscheinlichkeit für Fehler (Ablehnung von H_0, obwohl wahr) klein halten
- Setze Irrtumswahrscheinlichkeit α auf 0.05
- Bestimme θ , so dass

$$\int_0^{\theta} \varphi_{\mu,\sigma^2}(x)dx = 0.95$$



- Fällt Test in Ablehnungsbereich, liegt signifikante Abweichung vor
- Wir sprechen von einem Test mit Signifikanzniveau α

Auswertung des Experiments: Fehleranalyse

- Das Experiment fällt in den Ablehnungsbereich für H_0
- Also: Annahme der Vermutung H_1 als wahr
- Anzahl der Ausgänge mit Kreis sogar 16
- Bestimme

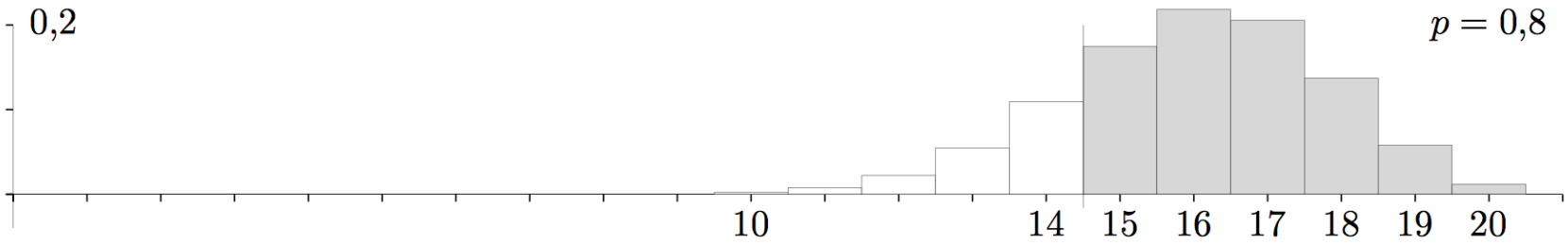
$$\int_0^{16} \varphi_{\mu,\sigma^2}(x)dx = 0.979$$

- Irrtumswahrscheinlichkeit sogar nur 0.021
- Wir sagen α = 0.021 (oder 2.1%)
  und nennen das Fehler 1. Art

Weitere Fragestellung

- Nehmen wir an, wir kennen die Verteilung $\mathcal{N}(\mu, \sigma^2)$ für den Falls, dass Küken eine angeborende Körnererkennungsbegabung haben.
- Mit welcher Wahrscheinlichkeit würde die Begabung der Küken nicht erkannt?
- Würden Küken Kreise mit Wahrscheinlichkeit p=0.8 bevorzugen, ergäbe sich:

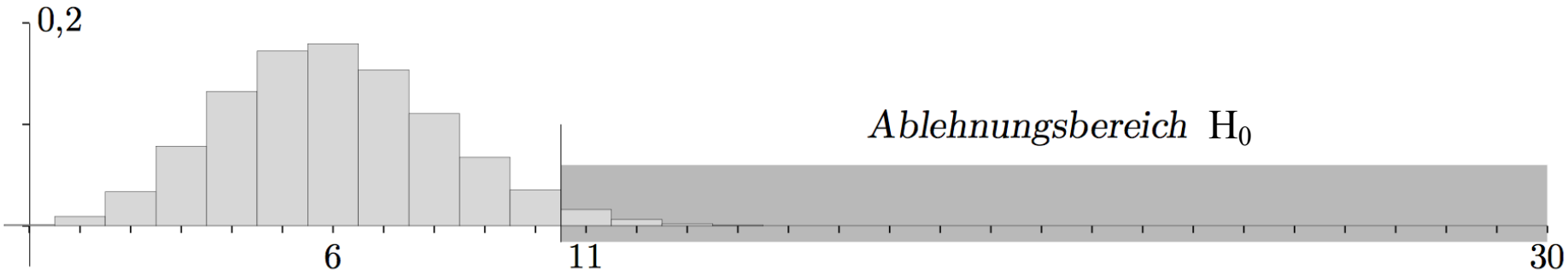
$$\beta = \Phi_{\mu,\sigma^2}(14) = 0{,}196$$



- Je mehr sich p dem Wert 0.5 nähert, umso größer wird der Fehler 2. Art

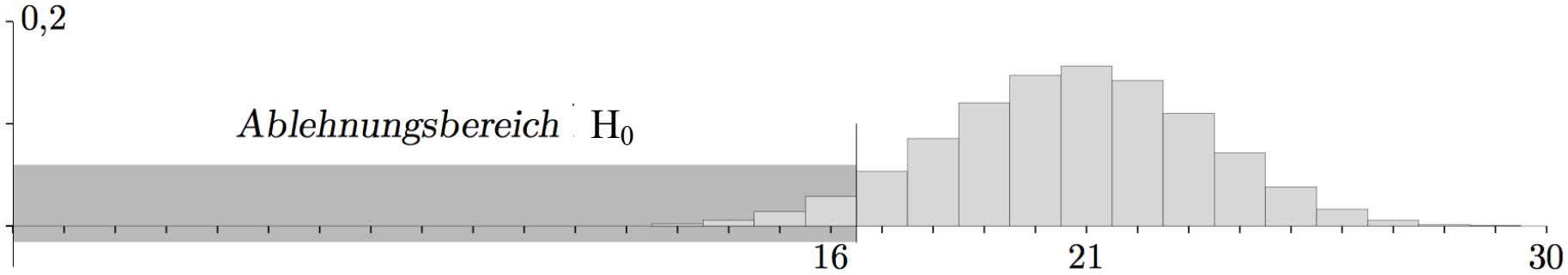
Ablehnungsbereichs rechts

- Behauptung: Ein bestimmtes Medikament verursacht höchstens bei 20 % der Patienten Nebenwirkungen. Wir bezweifeln dies und testen die Nullhypothese auf dem 5%-Niveau. Die Stichprobenlänge sei n = 30
- H_0: $p \leq \theta(\alpha)$ H_1: $p > \theta(\alpha)$



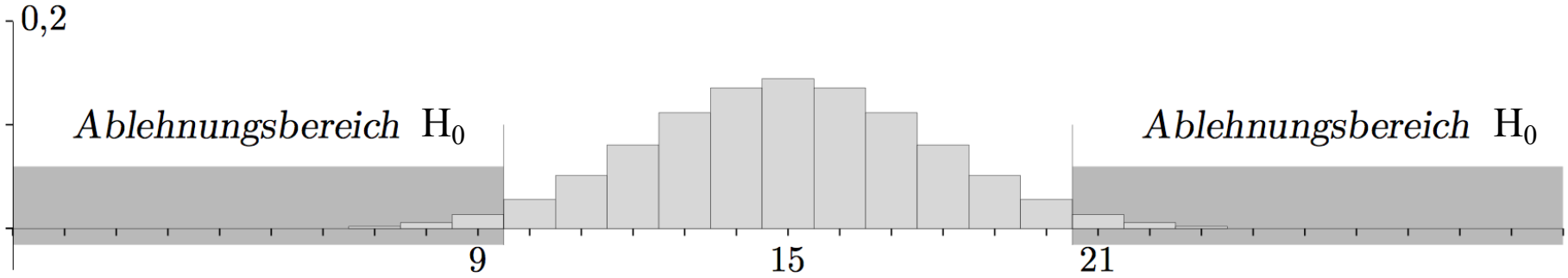
Ablehnungsbereichs links

- Behauptung: Mindestens 70 % der gelieferten Gurken erfüllen die europäsche Krümmungsnorm. Wir vermuten das Gegenteil und testen auf dem 5%-Niveau.
- H_0: $p \geq \theta(\alpha)$ H_1: $p < \theta(\alpha)$



Ablehnungsbereichs beidseitig

- Bei der zufälligen Farbgebung sollen 50 % der Serienprodukte eine helle Tönung besitzen. Wir wollen Abweichungen aufdecken.
- H_0: $p < \theta(\alpha/2)$, $p > \max - \theta(\alpha/2)$ H_1: sonst



Typ-1- und Typ-2-Fehler

- Typ 1: Wir lehnen H_0 ab, obwohl H_0 wahr ist
  - Wenn α=0,05 , dann lehnen wir H_0 in 5% der Fälle ab
  - Wahrscheinlichkeit α, mit der wir H_0 ablehnen, also einen Typ-1-Fehler zu machen

- Typ 2: Wir akzeptieren H_0 obwohl H_0 falsch ist
  - Die Wahrscheinlichkeit einen Typ-2-Fehler zu machen, ist β
  - 1-β ist dann die Wahrscheinlichkeit H_0 (richtigerweise) NICHT zu akzeptieren

- Es werden aber unterschiedliche Verteilungen zugrunde gelegt: $\alpha \neq 1-\beta$

HYPOTHESIS TESTING OUTCOMES

Research \ Reality	The Null Hypothesis Is True	The Alternative Hypothesis is True
The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error β 
The Alternative Hypothesis is True	Type I Error α 	Accurate $1 - \beta$ 

Zusammenfassung: Hypothesentest

- Um eine Hypothese zu beweisen, zeigt man,
  dass die Gegenhypothese wegen eines Testergebnisses
  äußerst unwahrscheinlich ist.
- Welche Hypothese als Nullhypothese getestet wird, hängt von Zielsetzung ab
- Wichtig: Verteilungsannahme der Nullhypothese muss gerechtfertigt sein
- Parameter der jeweils angenommenen Verteilung
  müssen sinnvoll bestimmt werden
- Wie groß sollte die Stichprobe sein?
- Wieviel Datenelemente benötigen wir, um gewisse Aussagen machen zu können?

Schätzung von Parametern

- Auswertung der Daten einer Stichprobe
- Rückschlüsse auf Eigenschaften der Grundgesamtheit
- Wir betrachten zunächst einmal die Normalverteilung
  - Aus Stichprobe Parameter bestimmen

Experimente, Zufallsvariablen, Verteilungen

- Durchführung von Experimenten / Auswertung von Daten
  - Merkmalsausprägungen bestimmen
  - Werte von statistischen Variablen
  - Im Sinne des Ziehens aus Grundgesamtheit: Zufallsvariable
- Beispiel: Zufallsvariable X normalverteilt
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
  - Standardnormalverteilung: $\mu = 0$ und $\sigma = 1$

Erwartungen formal

- Erwartungswert von Zufallsvariable X:
  - Wert, den X im Mittel einnimmt
  - Diskret:

$$\mathrm{E}(X) = \sum_{i \in I} x_i p_i$$

wobei p_i die relative Häufigkeit des Auftretens des Wertes x_i ist

  - Kontinuierlich:

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

f ist die Wahrscheinlichkeits-verteilung von X

  - Notation manchmal auch: $\mathrm{E}[X]$

- $\mathrm{E}[X]$, wenn $\mathrm{X} \sim \mathcal{N}(\mu, \sigma^2)$?

Varianz formal

- Varianz von Zufallsvariable X:
  - Erwartete (quadrierte) Abweichung vom Wert, den X im Mittel einnimmt
  - Definition:

$$\mathrm{Var}(X) \coloneqq \mathrm{E}\big((X-\mu)^2\big)$$

  - Notation manchmal auch: $\mathrm{Var}[X]$

- $\mathrm{Var}[X]$, wenn $X \sim \mathcal{N}(\mu, \sigma^2)$?
  - wobei

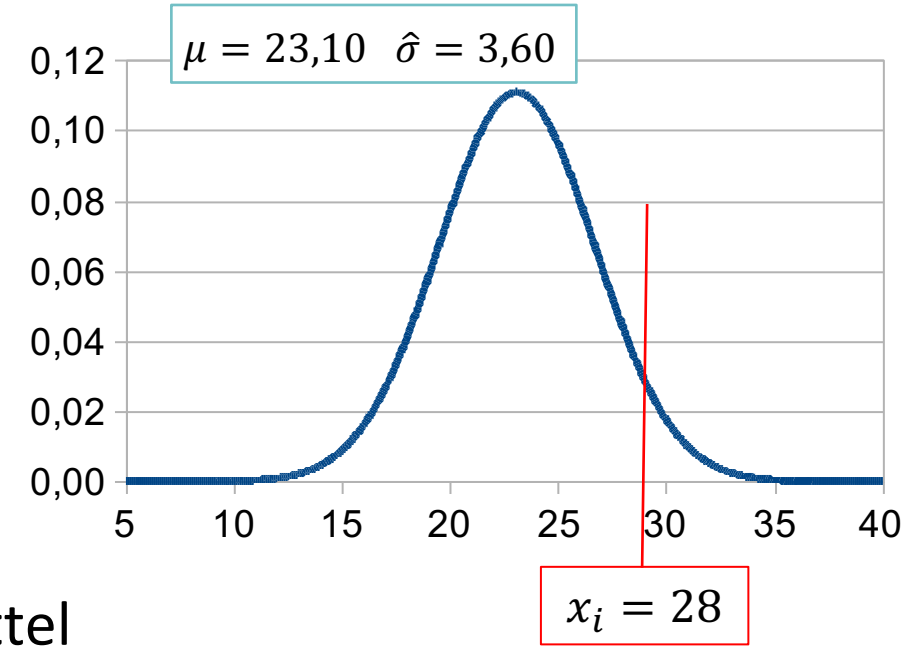
$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) dx$$

$$\mu = \int_{-\infty}^{\infty} x \, f(x) dx$$

Interpretation eines Messwertes

Beispiel $x_i = 28$

- Interpretierbar nur bei gegebener Verteilung
- x_i liegt über dem arithmetischen Mittel
- Genauer: x_i liegt mehr als eine Standardabweichung über dem arithmetischen Mittel
- Genauer: Wie viel Prozent der Gesamtheit haben Werte unter / über 28?
- Um diese Frage zu beantworten, hilft die z-Standardisierung

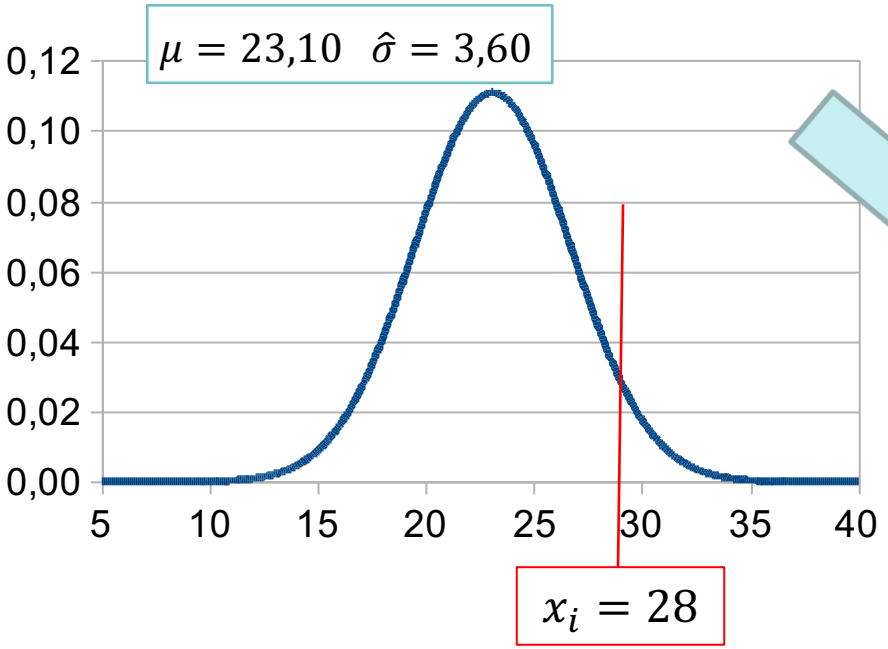


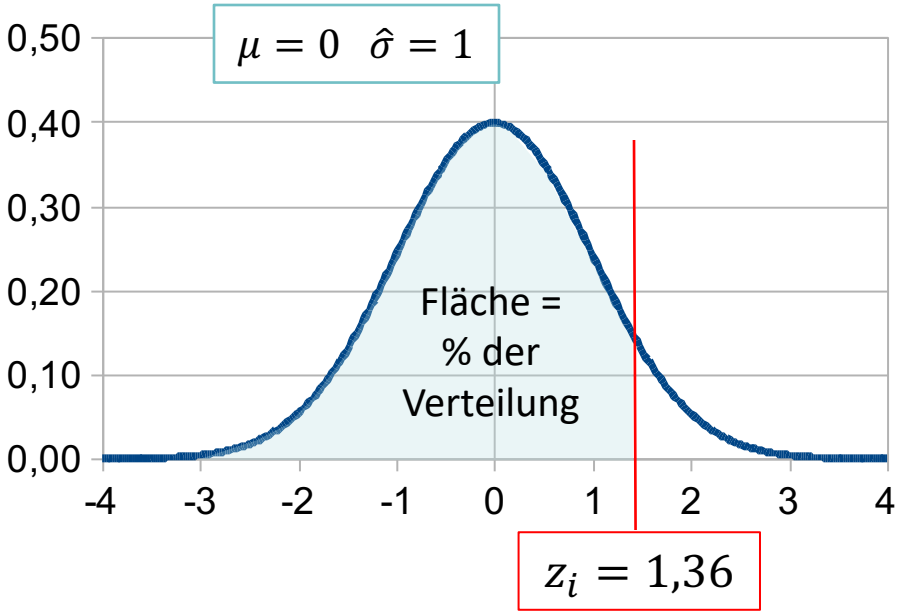
z-Standardisierung

- Mit der ***z-Standardisierung*** wird eine Normalverteilung in eine Standardnormalverteilung umgewandelt.
- Die z-Standardisierung erfolgt in zwei Schritten:
  (1) Zunächst wird von jedem Messwert der *Mittelwert* subtrahiert.
  (2) Dann wird das Ergebnis durch die *Standardabweichung* geteilt.

$$z_i = \frac{x_i - \bar{x}}{\hat{\sigma}}$$

z-Standardisierung





z-Standardisierung

- ***z Werte*** können mit Hilfe einer *z*-Tabelle einfach interpretiert werden.
- In Tabellen zur Standardnormalverteilung ist immer angegeben, wie groß die Fläche unter der Kurve links von einem *z*-Wert ist.
- Die Fläche gibt den Anteil der Verteilung an, deren Werte kleiner oder gleich des „kritischen“ z-Werts ist.
- Beispiel:
  - $x_i = 28$
  - $z_i = 1{,}36$
  - $\text{Fläche}(z_i) = \Phi(z_i) = 0{,}91$
  - Anteil der z-Werte $\leq 1{,}36 \rightarrow 0{,}91$
  - 91% der Population haben z-Werte kleiner oder gleich $1{,}36$
  - 91% der Population haben x-Werte von 28 oder darunter
  - Nur 9% der Population haben x-Werte größer als x_i

z-Standardisierung

Die z-Tabelle (Standardnormalverteilung)

z	*Fläche*	*z*	*Fläche*	*z*	*Fläche*	*z*	*Fläche*
-3.00	0.00	-1.50	0.07	0.00	0.50	1.50	0.93
-2.90	0.00	-1.40	0.08	0.10	0.54	1.60	0.95
-2.80	0.00	-1.30	0.10	0.20	0.58	1.70	0.96
-2.70	0.00	-1.20	0.12	0.30	0.62	1.80	0.96
-2.60	0.00	-1.10	0.14	0.40	0.66	1.90	0.97
-2.50	0.01	-1.00	0.16	0.50	0.69	2.00	0.98
-2.40	0.01	-0.90	0.18	0.60	0.73	2.10	0.98
-2.30	0.01	-0.80	0.21	0.70	0.76	2.20	0.99
-2.20	0.01	-0.70	0.24	0.80	0.79	2.30	0.99
-2.10	0.02	-0.60	0.27	0.90	0.82	2.40	0.99
-2.00	0.02	-0.50	0.31	1.00	0.84	2.50	0.99
-1.90	0.03	-0.40	0.34	1.10	0.86	2.60	1.00
-1.80	0.04	-0.30	0.38	1.20	0.88	2.70	1.00
-1.70	0.04	-0.20	0.42	1.30	0.90	2.80	1.00
-1.60	0.05	-0.10	0.46	1.40	0.92	2.90	1.00

z-Standardisierung

Interpretation der Ausprägung eines normalverteilten Merkmals

- Erhebung einer Stichprobe
  - Berechnung von Mittelwert und Standardabweichung
- Erhebung des Merkmals bei der Person i
- Berechnung des *z*-Werts
- Nachschlagen der Größe der Fläche unterhalb der *z*-Verteilung, die links von z_i liegt
- Die Fläche $f(z_i)$ gibt an, wie viel Prozent der Population Werte kleiner oder gleich z_i bzw. x_i haben.
- $1 - f(z_i)$ gibt an, wie viel Prozent der Population Werte größer z_i bzw. x_i haben.

Prozentränge

- Ein ***Prozentrang*** (PR) gibt an, wie viel Prozent der Population Werte *kleiner oder gleich* einem kritischen Wert haben.

Aufgabe: IQ-Wert-Analyse

Annahme: Normalverteilung
mit $\mu = 100; \sigma = 15$
Welchem Prozentrang entspricht
ein IQ-Wert von
(a) 130; (b) 92.5; (c) 85; (d) 100; (e) 115?

$$z_i = \frac{x_i - \mu}{\sigma}$$

z	*Fläche*	*z*	*Fläche*	*z*	*Fläche*	*z*	*Fläche*
-3.00	0.00	-1.50	0.07	0.00	0.50	1.50	0.93
-2.90	0.00	-1.40	0.08	0.10	0.54	1.60	0.95
-2.80	0.00	-1.30	0.10	0.20	0.58	1.70	0.96
-2.70	0.00	-1.20	0.12	0.30	0.62	1.80	0.96
-2.60	0.00	-1.10	0.14	0.40	0.66	1.90	0.97
-2.50	0.01	-1.00	0.16	0.50	0.69	2.00	0.98
-2.40	0.01	-0.90	0.18	0.60	0.73	2.10	0.98
-2.30	0.01	-0.80	0.21	0.70	0.76	2.20	0.99
-2.20	0.01	-0.70	0.24	0.80	0.79	2.30	0.99
-2.10	0.02	-0.60	0.27	0.90	0.82	2.40	0.99
-2.00	0.02	-0.50	0.31	1.00	0.84	2.50	0.99
-1.90	0.03	-0.40	0.34	1.10	0.86	2.60	1.00
-1.80	0.04	-0.30	0.38	1.20	0.88	2.70	1.00
-1.70	0.04	-0.20	0.42	1.30	0.90	2.80	1.00
-1.60	0.05	-0.10	0.46	1.40	0.92	2.90	1.00

IQ	z(IQ)	PR
130	2.0	98
92.5	-0.5	31
85	-1.0	16
100	0.0	50
115	1.0	84

Prozenträngе

- Ein ***Prozentrang*** (PR) gibt an, wie viel Prozent der Population Werte *kleiner oder gleich* einem kritischen Wert haben.
- Damit entspricht der Prozentrang der Wahrscheinlichkeit des *z*-Werts

Wahrscheinlichkeiten

- Die z-Tabelle ermöglicht es auch, ***Wahrscheinlichkeitsaussagen*** für bestimmte Intervalle zu machen.
- Wie groß ist die Wahrscheinlichkeit für einen IQ-Wert (a) von 85 bis 115; (b) von 70 bis 130; (c) von 0 bis 70; (d) von über 100

IQ	$z(IQ_1)$	$z(IQ_2)$	$p(z_1)$	$p(z_2)$	Δp
85 bis 115	-1.0	1.0	.16	.84	.68
70 bis 130	-2.0	2.0	.02	.98	.96
0 bis 70	-6.7	-2.0	.00	.02	.02
> 100	0	∞	.50	1.00	.50

Wahrscheinlichkeiten

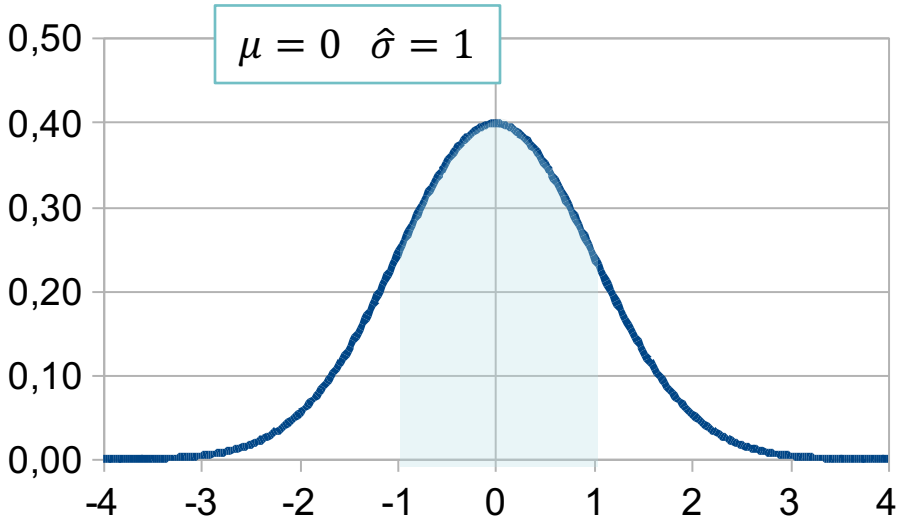
Generell gilt für normalverteilte Merkmale:

- ***68.26%*** der Werte liegen im Bereich:

  $\mu - 1{,}0 \cdot \sigma < x_i < \mu + 1{,}0 \cdot \sigma$ bzw. $-1{,}0 < z_i < 1{,}0$

- ***95.44%*** der Werte liegen im Bereich:

  $\mu - 2{,}0 \cdot \sigma < x_i < \mu + 2{,}0 \cdot \sigma$ bzw. $-2{,}0 < z_i < 2{,}0$

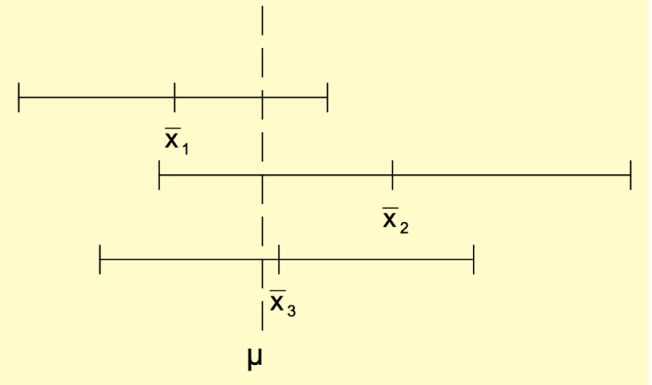


Stichprobenkennwerteverteilungen

- Wir haben verschiedene Stichprobenkennwerte kennengelernt: z.B. Mittelwert, Median, Varianz („Punktschätzer“)
- Meist interessieren nicht die Werte für die konkrete Stichprobe, sondern für die zugrundeliegenden Population
- Die Kennwerte aus einer Stichprobe werden daher als Schätzer für die entsprechenden Populationskennwerte verwendet
- Wir erwarten: Je größer eine (repräsentative) Stichprobe, desto genauer ist die Schätzung

Stichprobenkennwerteverteilungen

- Wenn man aus der gleichen Population immer wieder Stichproben zieht, ergibt sich für jede Stichprobe ein neuer Mittelwert
- Wenn man sehr viele Stichproben erhebt, erhält man auch viele Mittelwerte
- Nun kann man die Verteilung der resultierenden Mittelwerte betrachten
- Diese Verteilung heißt Stichprobenkennwerteverteilung des Mittelwerts



Standardfehler

- Diese „Verteilung der der Mittelwerte“ ist selbst wieder normalverteilt (wenn das Merkmal normalverteilt ist)
- Der Mittelwert der Stichprobenkennwerteverteilung entspricht dem Mittelwert in der Population
- Die Streuung der Stichprobenkennwerteverteilung wird als Standardfehler (des Mittelwerts) bezeichnet
  - Der Standardfehler gibt an, wie nah ein empirischer Stichprobenmittelwert am wahren Populationsmittelwert liegt
  - Dieser Standardfehler des Mittelwertes kann auch aus einer einzigen Stichprobe geschätzt werden:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$$

Begründung

Standardfehler des arithmetischen Mittels [Bearbeiten | Quelltext bearbeiten]

Der Standardfehler des arithmetischen Mittels ist gleich

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}},$$

wobei σ die Standardabweichung einer einzelnen Messung bezeichnet.

Herleitung [Bearbeiten | Quelltext bearbeiten]

Der Mittelwert einer Stichprobe vom Umfang n ist definiert durch

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

Betrachtet man die Schätzfunktion

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

mit unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit endlicher Varianz σ^2, so ist der Standardfehler definiert als die Wurzel aus der Varianz von $\overline{X}$. Man berechnet unter Verwendung der Rechenregeln für Varianzen und der Gleichung von Bienaymé:

$$\sigma(\overline{X})^2 = \operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Rechenregeln für Varianzen

- $\operatorname{Var}(cx) = c^2 \cdot \operatorname{Var}(x)$
- $\operatorname{Var}(\Sigma x) = \Sigma \operatorname{Var}(x)$

https://de.wikipedia.org/wiki/Standardfehler#Standardfehler_des_arithmetischen_Mittels

Standardfehler

Beispiel: Unter den Mitarbeitern einer großen Firma soll die Leistungsmotivation bestimmt werden. Es werden 10 Mitarbeiter zufällig ausgewählt und getestet

- Es ergibt sich ein Mittelwert von 60 bei einer geschätzten Populationsvarianz von 90

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}}$$

- Wie groß ist der Standardfehler dieses Mittelwerts?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{10}} = \sqrt{9} = 3$$

- Wie groß wäre der Standardfehler bei $\sigma^2=250$ und $n=10$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{250}{10}} = \sqrt{25} = 5$$

- Wie groß wäre der Standardfehler bei $\sigma^2=90$ und $n=90$?

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{90}{90}} = \sqrt{1} = 1$$

Bereich um den Mittelwert

- Der Standardfehler ist die Standardabweichung der Stichprobenkennwerteverteilung
- Da die Stichprobenkennwerteverteilung normalverteilt ist, kann die Wahrscheinlichkeit dafür berechnet werden, dass der Mittelwert in einem bestimmten Intervall liegt
- Mit $p = 0{,}68$ ist der Populationsmittelwert höchstens einen Standardfehler vom Stichprobenmittelwert entfernt
- Beispiel:
  - Wenn $\bar{x} = 60$ und $\hat{\sigma}_{\bar{x}} = 3$, dann gilt mit $p = 0{,}68$ für den Populationsmittelwert: $57 < \mu < 63$

- Notation: $\mathrm{P}(Bedingung) = p$ mit $p \in [0, 1]$
- Beispiel: $\mathrm{P}(57 < \mu < 63\ ) = p$

Konfidenzintervalle

- Ein Konfidenzintervall ist ein symmetrischer Bereich um den Stichprobenmittelwert, in welchem der Populationsmittelwert mit einer bestimmten Wahrscheinlichkeit liegt.

$$\mathrm{P}(\bar{x} - 1{,}00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1{,}00 \cdot \hat{\sigma}_{\bar{x}}) = 0{,}682$$
$$\mathrm{P}(\bar{x} - 2{,}00 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2{,}00 \cdot \hat{\sigma}_{\bar{x}}) = 0{,}954$$
$$\mathrm{P}(\bar{x} - 1{,}96 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 1{,}96 \cdot \hat{\sigma}_{\bar{x}}) = 0{,}95$$
$$\mathrm{P}(\bar{x} - 2{,}57 \cdot \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + 2{,}57 \cdot \hat{\sigma}_{\bar{x}}) = 0{,}99$$

Konfidenzintervall

- Die Lage und Breite des Konfidenzintervalls ist abhängig von den zufälligen Konfidenzgrenzen
- Diese hängen ab von:
  - dem Stichprobenumfang
  - der Schätzfunktion und deren Verteilung und
  - dem sog. Konfidenzniveau α
- Breite des Konfidenzintervalls ist Ausdruck für die Genauigkeit der Parameterschätzung!
  - Ein höheres Konfidenzniveau (kleineres α) führt zu einer Verbreiterung des Konfidenzintervalls und …
  - … ein größerer Stichprobenumfang führt zu einer Verkleinerung des Konfidenzintervalls

Konfidenzintervall: Herleitung

Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ eine normalverteilte ZV und $(X_1, \ldots, X_n)$ eine mathematische Stichprobe aus der GG X.

1. Fall: Die Varianz σ^2 der normalverteilten GG sei bekannt.
Für den unbekannten Parameter μ ist eine Konfidenzschätzung anzugeben.

Als Punktschätzer für μ wählen wir das arithmetische Mittel

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

mit $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

ZG= Zufallsvariable GG=Grundgesamtheit

Konfidenzintervall: Herleitung

- Die Wahrscheinlichkeit, dass der Betrag des Schätzfehlers kleiner als die Schranke d ist, wird mit $(1-\alpha)$ vorgegeben:

$$\mathrm{P}(|\bar{X}-\mu| \leq d) = 1-\alpha \quad \bar{X}-\mu \text{ ist der Schätzfehler}$$

- Betrag auflösen:
  - Fall $\bar{X}-\mu \geq 0$: $\quad \bar{X}-\mu \leq d \rightarrow \bar{X}-d \leq \mu$
  - Fall $\bar{X}-\mu \leq 0 = -(\bar{X}-\mu) \geq 0$: $\quad -\bar{X}+\mu \leq d \rightarrow \bar{X}+d \geq \mu$
- Umformung:

$$\mathrm{P}(|\bar{X}-\mu| \leq d) = \mathrm{P}(\bar{X}-d \leq \mu \leq \bar{X}+d) = 1-\alpha$$

(Symmetrie der NV-Dichtefunktion)

NV = Normalverteilung

Konfidenzintervall: Herleitung

- Zur Bestimmung der Größe d z-standardisieren wir die ZV $\bar{X}$:
  - gegeben $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $z = \frac{x-\mu}{\sigma}$ $\sigma = \sqrt{\frac{\sigma^2}{n}}$

    (Standardabweichung σ = Wurzel der Varianz)
  - Z-standardisiert: $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ $Z \sim \mathcal{N}(0,1)$
  - In $\mathrm{P}(|\bar{X}-\mu| \leq d)$,

    $\rightarrow \mathrm{P}\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right| \leq \frac{d}{\sigma}\cdot\sqrt{n}\right)$

Konfidenzintervall: Herleitung

- Gegeben $\mathrm{P}\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right| \leq \frac{d}{\sigma}\cdot\sqrt{n}\right) = 1-\alpha,$

$$z_{1-\frac{\alpha}{2}} = \frac{d}{\sigma}\cdot\sqrt{n}$$

- Daraus folgt mit $\mathrm{P}(|\bar{X}-\mu| \leq d) = \mathrm{P}(\bar{X}-d \leq \mu \leq \bar{X}+d)$:

$$\mathrm{P}\left(\left|\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right| \leq \frac{d}{\sigma}\cdot\sqrt{n}\right)$$

$$= \mathrm{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right) \qquad \rightarrow \boxed{d = \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}}$$

$$= \mathrm{P}\left(z_{\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \qquad z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$$

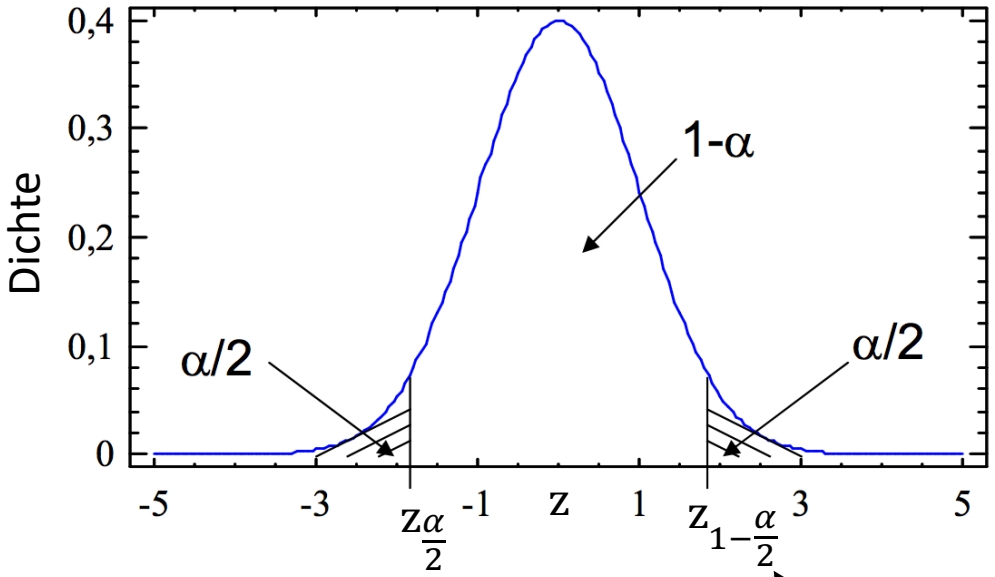
$$= \boxed{\mathrm{P}\left(\bar{X}-\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}} < \mu < \bar{X}+\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right) = 1-\alpha}$$

Konfidenzintervall: Interpretation

- Das Konfidenzintervall

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right]$$

überdeckt also den wahren Parameter μ mit der Wahrscheinlichkeit $(1-\alpha)$.



Dichtefunktion der Standard-Normalverteilung

$1-\frac{\alpha}{2}$-Quantil der Standard-NV

Konfidenzintervall: Interpretation

- Jede konkrete Stichprobe liefert uns eine Realisierung der ZV $\bar{X}$ und damit ein realisiertes Konfidenzintervall:

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right]$$

- Einige typische $z_{1-\frac{\alpha}{2}}$-Werte (2-seitige Fragestellung) und $z_{1-\alpha}$-Werte (1-seitige Fragestellung) enthält die Tabelle:

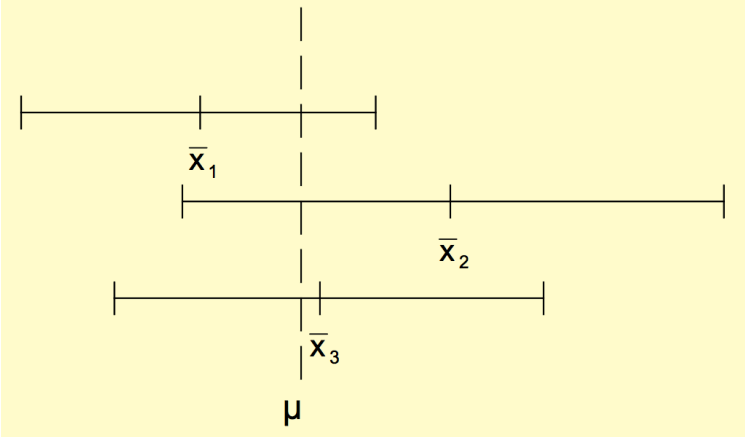
$1-\alpha$	α	$z_{1-\frac{\alpha}{2}}$	$z_{1-\alpha}$
0,95	0,05	1,96	1,64
0,99	0,01	2,58	2,33
0,999	0,001	3,29	3,09

$$\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

$$\Phi(z_{1-\alpha}) = 1 - \alpha$$

Konfidenzintervall: Bemerkungen

- Die Lage des konkreten Konfidenzintervalls wird durch die konkrete Stichprobe bestimmt.

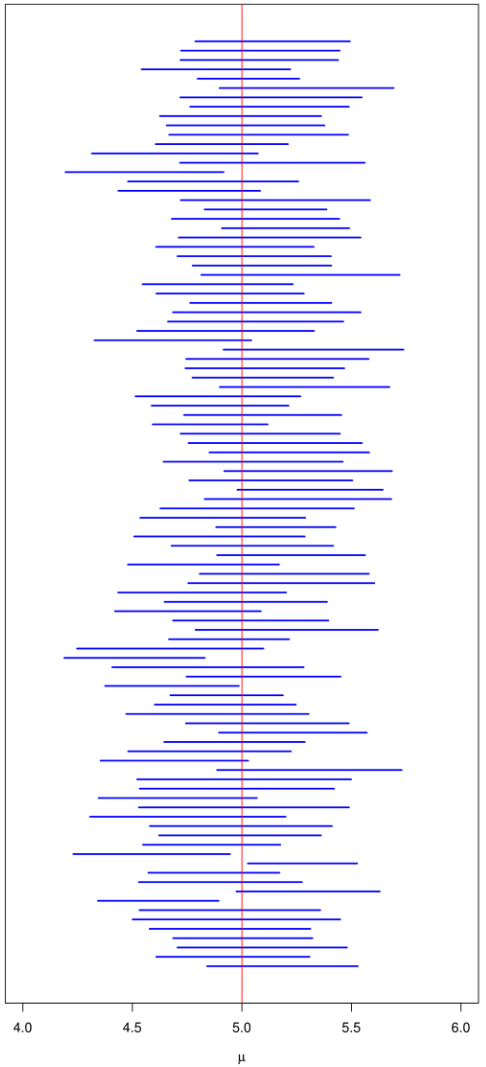


- Bei einem Konfidenzniveau von $(1-\alpha) = 0{,}95$ heißt das:
  - In 95% aller Fälle enthält der Vertrauensbereich den unbekannten Parameter der GG und in 5% der Fälle nicht.
  - D.h.: Behauptet man k mal, der unbekannte Parameter liege im Vertrauensbereich, so hat man im Mittel $\alpha \cdot k$ Fehlschüsse zu erwarten.

Quelle: https://de.wikipedia.org/wiki/Konfidenzintervall#/media/File:Confidence_intervall_normal_dist.svg

Konfidenzintervall: Beispiel

- Experiment
  - Normalverteilte GG
  - $\mu = 5$
  - $\alpha = 0{,}05$
  - 100 Stichproben
  - $n = 30$
- Mittelwerte, Konf.intervalle für jede Stichprobe ausrechnen
  - 94 der Intervalle überdecken μ
  - 6 Intervalle tun das nicht
- Mittelwerte der 100 Stichproben normalverteilt
  - Stichprobenkennwerteverteilung



Quelle: https://de.wikipedia.org/wiki/Konfidenzintervall#/media/File:Confidence_intervall_normal_dist.svg

Konfidenzintervall: Bemerkungen

- Die Breite des Konvidenzintervalls für den Erwartungswert μ beträgt $2d$ und ist von α, n, σ und der Verteilung der zugehörigen Schätzfunktion abhängig.

$$2d = 2 \cdot \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

  - Je größer α (n konstant), desto größer das Konf.intervall
  - Je größer n, desto kleiner das Konfidenzintervall
  - $2d$: Maß für die Genauigkeit der Schätzung von μ
  - α ein Maß für das Risiko
- Planung des Stichprobenumfangs

Konfidenzintervall: Bemerkungen

- Planung des Stichprobenumfangs
  - Gegeben: halbe Breite des Konfidenzintervalls d, Varianz σ^2 Konfidenzniveau $(1-\alpha)$
  - Gesucht: Stichprobenumfang n

$$2d = 2 \cdot \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

$$d = \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

$$\sqrt{n} = \frac{\sigma}{d} \cdot z_{1-\frac{\alpha}{2}}$$

$$n = \frac{\sigma^2}{d^2} \cdot z^2_{1-\frac{\alpha}{2}}$$

Schätzung der Varianz

- Ähnliche Überlegungen
- Auch hierfür Herleitung der erforderlichen Stichprobengröße möglich

Begriff der Erwartungstreue

- Ein Schätzer heißt erwartungstreu, wenn sein Erwartungswert gleich dem wahren Wert des zu schätzenden Parameters ist
  - Schätzung von μ der GG durch Stichprobenmittel

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

  - Wenn x_i zufällig aus GG gezogen, dann $\mathrm{E}(x_i) = \mu$
  - Erwartungswert $\bar{x}$

$$\mathrm{E}(\bar{x}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$$

- Stichprobenmittel also erwartungstreuer Schätzer von μ

[Wikipedia]

Korrigierte Stichprobenvarianz

- Gegeben Stichprobenwerte $(x_1, \dots, x_n)$
- Korrigierte Stichprobenvarianz

Warum $n-1$?

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

  - Mit Stichprobenmittel $\bar{x}$

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

Korrigierte Stichprobenvarianz: Warum $n-1$?

- Stichprobenwerte $(x_1, \ldots, x_n)$ sind Ausprägungen der unabhängig identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit Varianz σ^2 und Mittelwert μ der GG
- Dann ist S_0^2 eine erwartungstreue Schätzfunktion für σ^2 und s_0^2 eine erwartungstreue Schätzung der Varianz

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Es gilt

$$\mathrm{E}(S_0^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left((X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}n\sigma^2 = \sigma^2$$

Korrigierte Stichprobenvarianz: Warum $n-1$?

- Überlicherweise μ unbekannt, geschätzt durch

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

- Als Schätzfunktion eingesetzt, erhält man für σ^2 als Schätzung

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \rightarrow \quad s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Erwartungstreue testen über Erwartungswert von S_1^2

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Korrigierte Stichprobenvarianz: Warum $n-1$?

$$
\begin{aligned}
\mathrm{E}(S_1^2) &= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu+\mu-\bar{X})^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - 2(X_i-\mu)(\bar{X}-\mu) + (\bar{X}-\mu)^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - 2\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu) + n(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + n(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - 2n(\bar{X}-\mu)(\bar{X}-\mu) + n(\bar{X}-\mu)^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n}\mathrm{E}\big((X_i-\mu)^2\big) - n\cdot\mathrm{E}\big((\bar{X}-\mu)^2\big)\right) \\
&= \frac{1}{n}\big(n\cdot\mathrm{Var}(X) - n\cdot\mathrm{Var}(\bar{X})\big) = \mathrm{Var}(X) - \mathrm{Var}(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Korrigierte Stichprobenvarianz: Warum $n-1$?

- Ergebnis von $\mathrm{E}(S_1^2)$

$$\mathrm{E}(S_1^2) = \frac{n-1}{n}\sigma^2$$

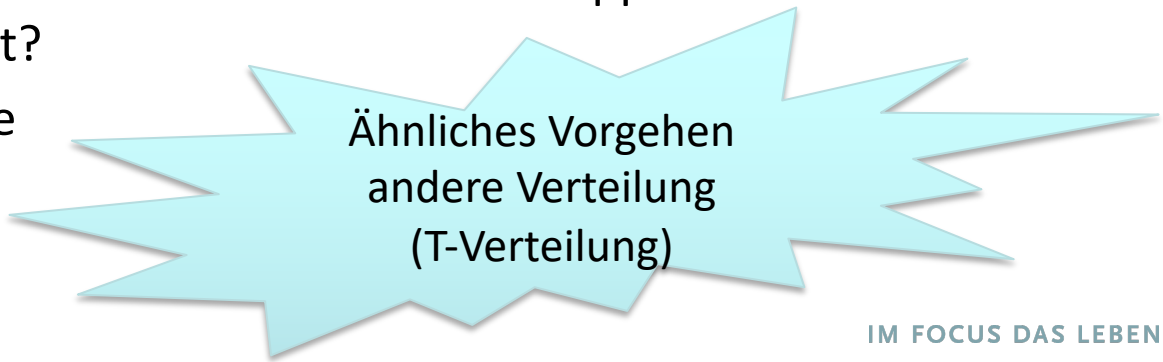
- Schätzfunktion S_1^2 nicht erwartungstreu für σ^2
- Lösung: multiplizieren mit $\frac{n}{n-1}$
- Erwartungstreue Schätzfunktion für σ^2 $\quad S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$S_1^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Damit gilt $\mathrm{E}(S_1^2) = \sigma^2$

Unterschiedshypothesen

- Sind Frauen ängstlicher als Männer?
  - Unterscheiden sich die Mittelwerte von zwei Gruppen?
  - Unabhängige Stichproben
- Ist der Mittelwert der Ängstlichkeit nach einer Therapie größer als vor der Therapie?
  - Unterscheidet sich der Mittelwert einer Stichprobe zu zwei Messzeitpunkten?
  - Abhängige Stichproben
- Liegt der mittlere IQ einer Gruppe über 100?
  - Unterscheidet sich der Mittelwert einer Gruppe von einem vorgegeben Wert?
  - Test bzgl. Gruppe



Reliabilität

- Zuverlässigkeit
  - Angegeben beispielsweise durch Konfidenzintervall

- Messgenauigkeit eines Tests mit mehreren Indikatoren bzw. Merkmalen (Beispiel: Fragebogen) und z.B. Mittelung der ermittelten Werte der Teilmerkmale
  - Interne (innere) Konsistenz
    - Wird von verschiedenen Merkmalen (z.B. an verschiedenen Stellen eines Fragebogens) dasselbe gemessen?
  - Zeitliche Stabilität
    - Wird zu verschiedenen Zeitpunkten (bei Testwiederholung) dasselbe gemessen?

Bestimmung der Reliabilität eines Tests

- Re-Test-Reliablität :
  - Bestimmung des statistischen Zusammenhangs (Korrelation) zwischen zwei aufeinanderfolgenden Messungen

- Ein Test misst dann genau, wenn er zu mehreren Zeitpunkten dasselbe Ergebnis liefert.

- Korrelation desselben Fragebogen-Gesamtwerts zu verschiedenen Zeitpunkten mit denselben Probanden (ungeeignet bei vorübergehenden Merkmalen, z.B. Stimmung)

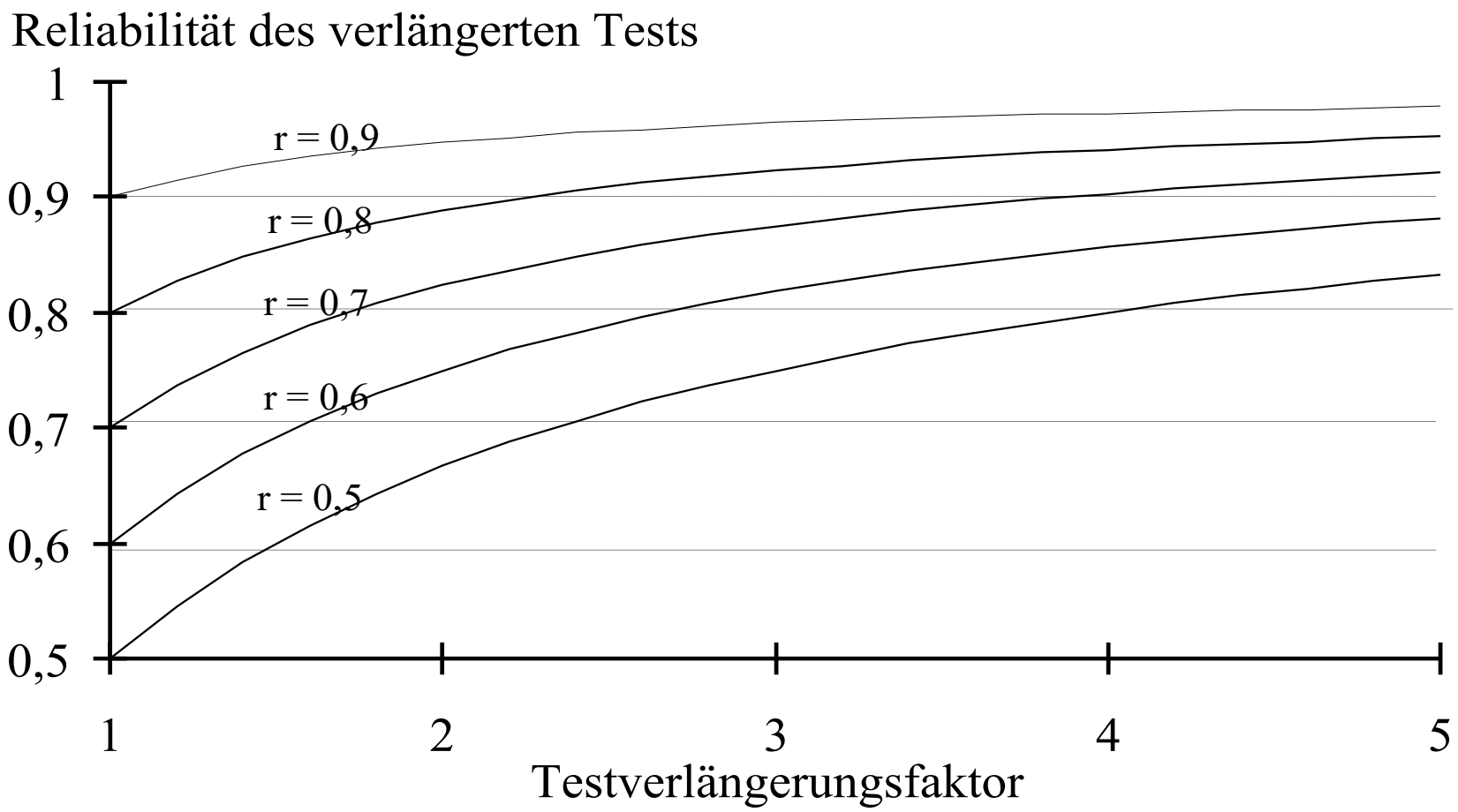
Bestimmung der Reliabilität eines Tests

- Split-Half-Reliabilität:
  - Korrelation zwischen zwei Hälften der Items eines Tests

- Cronbachs Alpha (Maß für sog. Interne Konsistenz):
  - Mittelwert der Korrelationen zwischen allen Einzelitems
  - Ausreichende Reliabilität: r: = 0.75
  - Gute Reliabilität: r = 0.90

- Probleme:
  - Die Messgenauigkeit kann nur für mehrere Items (Skala, Test, Subtest) bestimmt werden, nicht für Einzelitems
  - Daher liefert ein Test, der nicht vollständig durchgeführt wurde, keine zuverlässige Messung
  - Je mehr Items ein Test (Subtest, Skala) enthält, desto „genauer“ wird er

Reliabilitätssteigerung durch Testverlängerung



Validität

- Validität: Gültigkeit
- Misst ein Test das, was er messen soll?
  - Zusammenhang zwischen dem Testergebnis und anderen Kriterien für das Zielverhalten
  - Evaluation durch Bestimmung des Zusammenhangs (Korrelation) zwischen dem Testergebnis und anderen Kriterien für das messende Verhalten

Zusammenfassung

- Konzept der Stichprobe
- Relative Häufigkeiten
- Verteilungen
- Beschreibungsmaße
  - Mittelwert, Varianz (Streuung), …
- Wahrscheinlichkeiten
- Hypothesentest, Signifikanzniveau
- Zufallsvariable,
- Normalverteilung, Standardnormalverteilung
- Standardfehler
- Konfidenzintervall, Stichprobenumfang
- Erwartungstreue