Einführung in Web- und Data-Science

Statistische Grundlagen

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

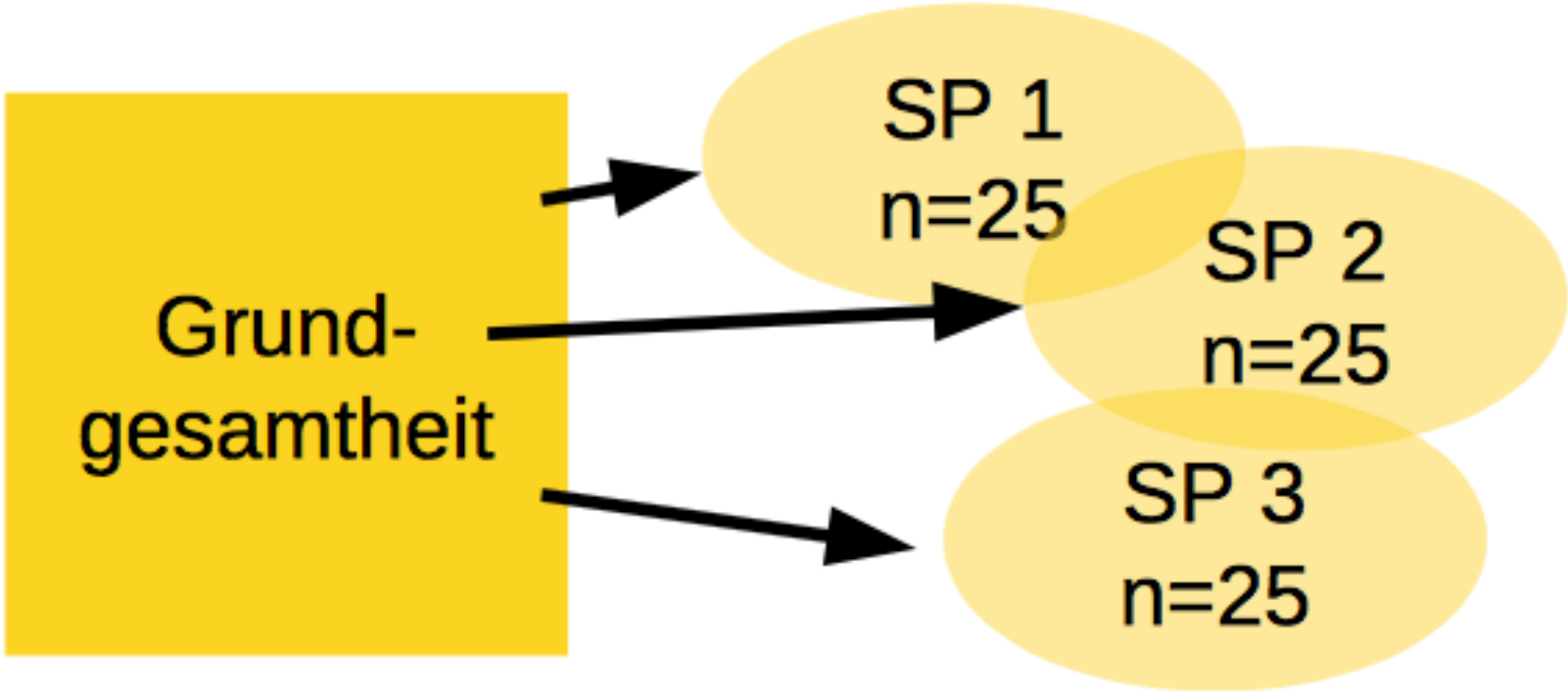
Statistische Grundlagen

BEGRIFFSBESTIMMUNGEN

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

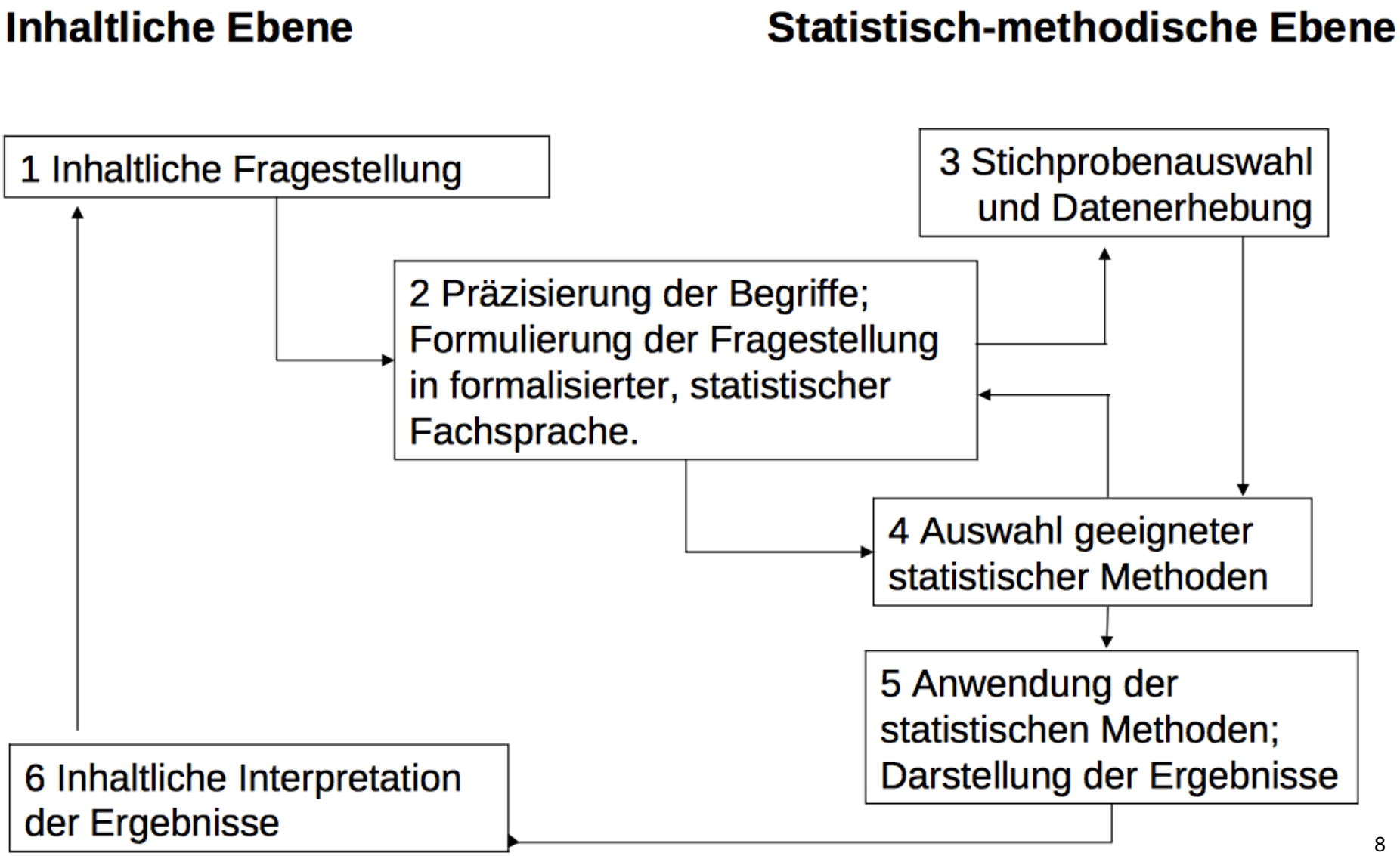
- Deskriptive Statistik
  - Beschreiben von Daten (auch: Teilmenge von Daten)
    - Beispiele: Mittelwert, Varianz, ...
    - … jeweils auch „Statistik“ genannt
  - Suchen nach Trends / Mustern
    - Beispiele: Häufige Artikelmengen, Assoziationsregeln
- Induktive Statistik
  - Ziel: Verallgemeinerung 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 erst einmal 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?

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

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

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

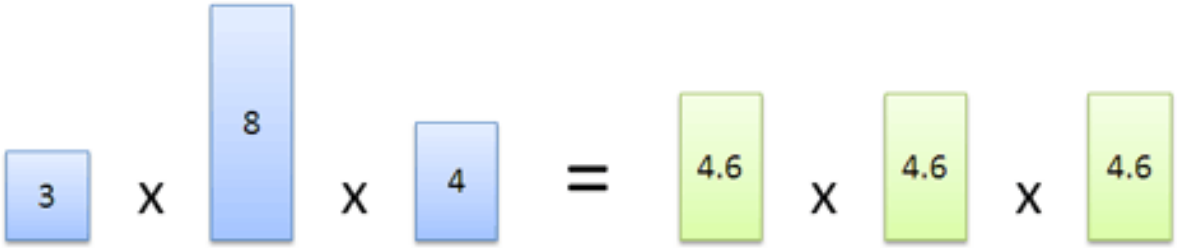
$$\bar{x}_{\text{harm}} = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

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

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

Visualisierung

Geometrisches Mittel



Produkt der Werte

Produkt der gemittelten Werte

Harmonisches Mittel

Ich mache 1 Bild pro Stunde
Sie machen 2 Bilder pro Stunde

Zusammen machen wir im Mittel 1.5 Bilder pro Stunde

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

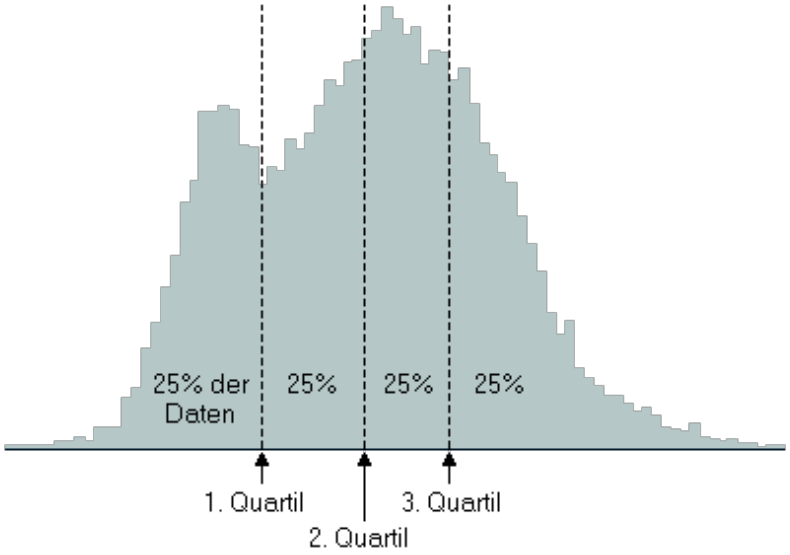
  Most frequent value

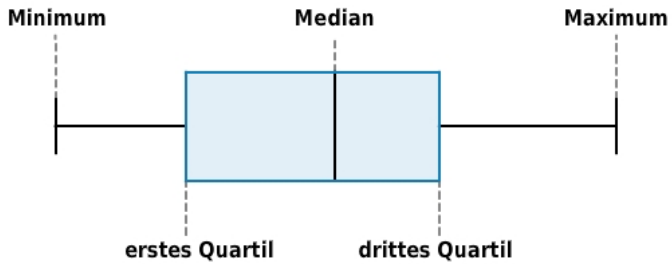
  2, 2, 3, 5, 5, 5, 9, 9, 15

- Quantil, Quartil

  Geordnete Reihe der Merkmalsausprägungen wird in gleichgroße Teile zerlegt

  Kumulierte Häufigkeiten





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 Reihenfolgen
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	Zeitpunkte 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 werden oft zur sog. Kardinalskala zusammengefasst.
- Merkmale auf dieser Skala werden dann als metrisch bezeichnet

Kategoriale Variablen

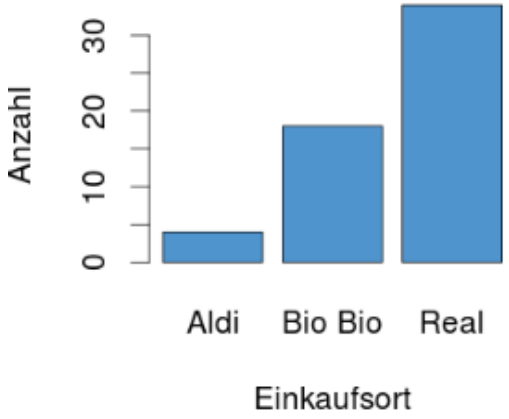
- Nominalskalierte Variablen
- Ordinalskalierte Variablen
- Variablen, die durch Kategorisierung aus ordinalskalierten oder metrischen Variablen entstanden sind (Beispiel: Variable „Einkommen“ mit den Kategorien „500-999 €“, „1000-1499 €“ usw.)

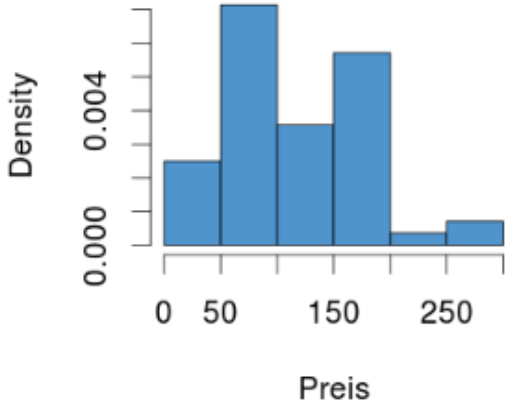
Streuungsmaße

- 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

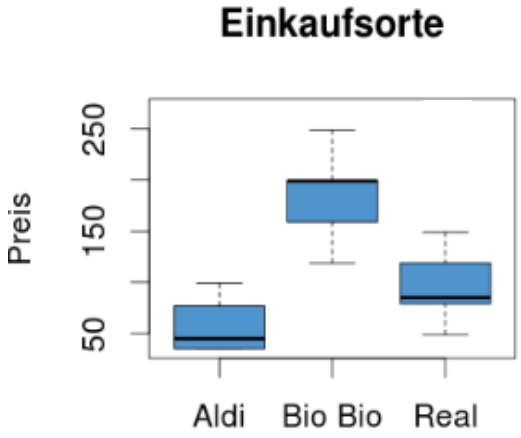
Säulendiagramm

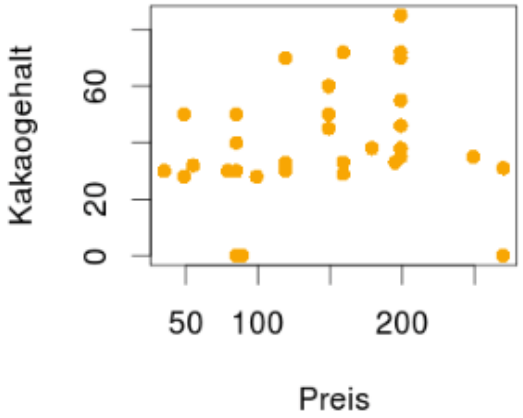




Histogramm

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

Statistische Grundlagen

VERTEILUNGEN

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

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

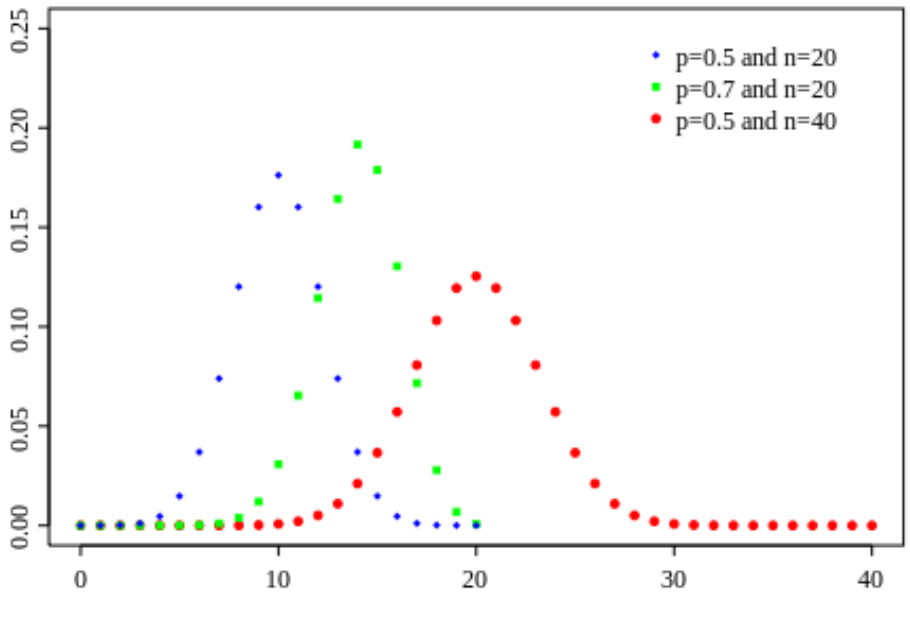
Binomialverteilung

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

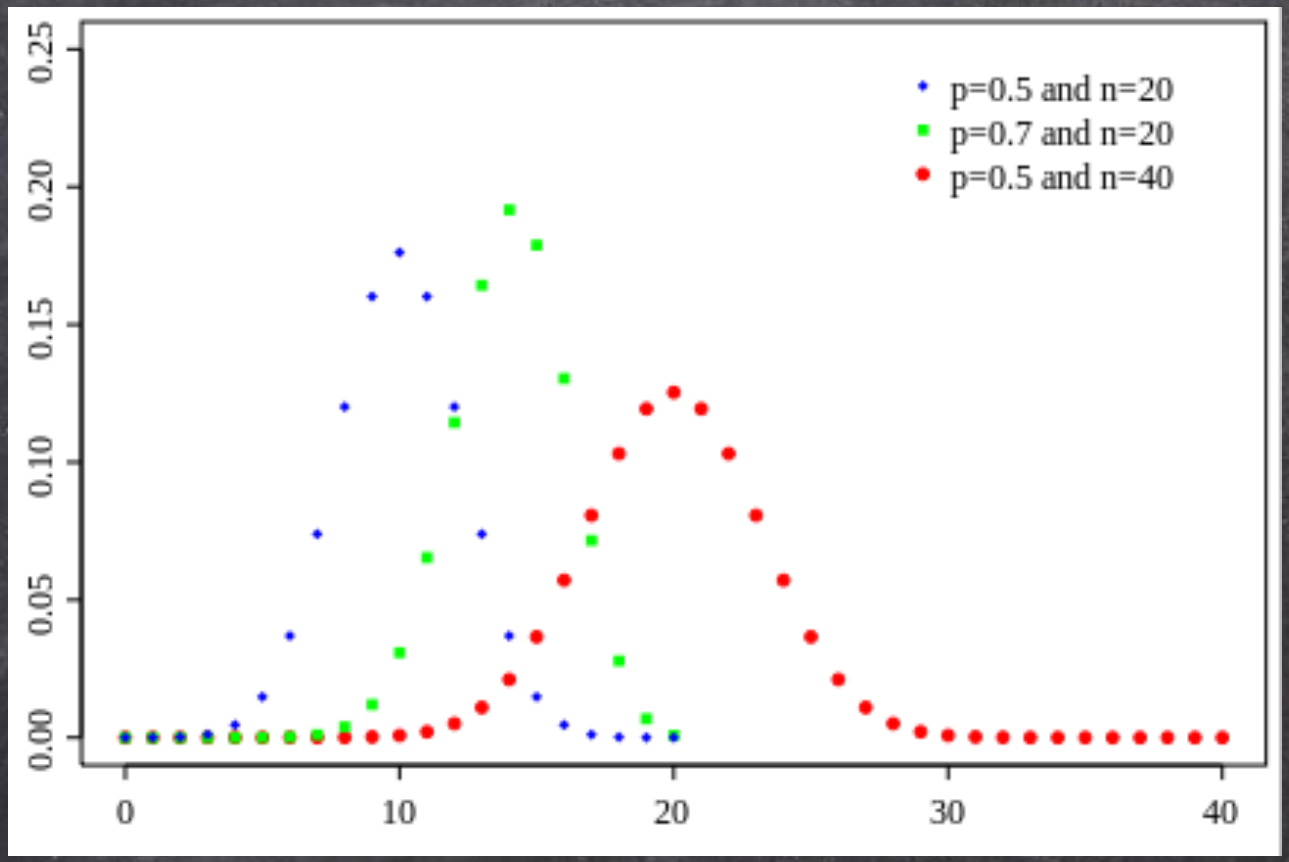
- n = #Versuche
  p = #erfolgr. Vers. / #Versuche

- 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?

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

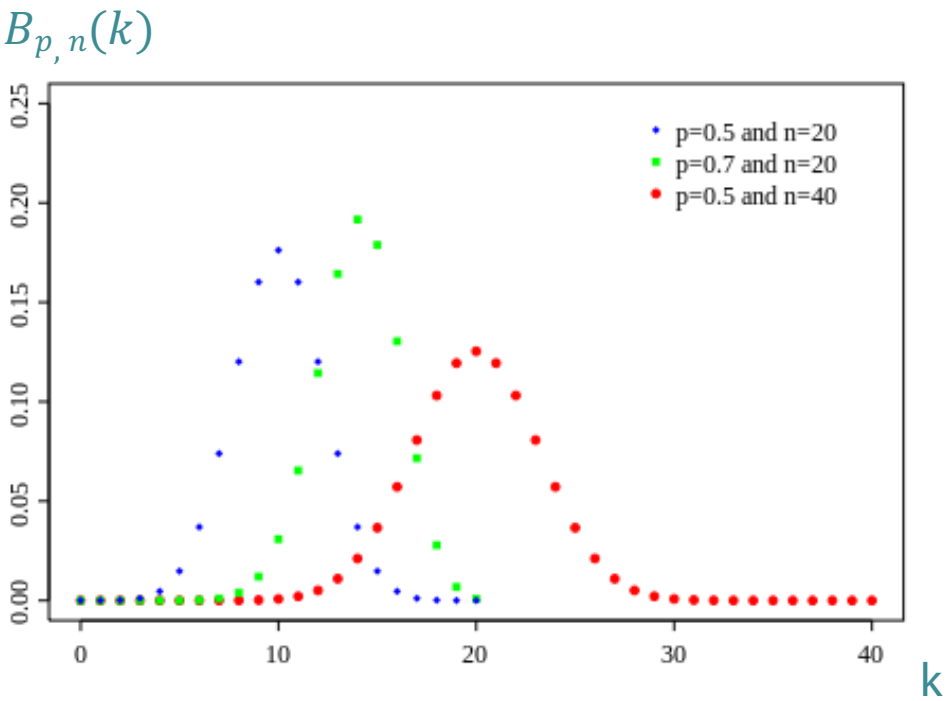
Kumulierte Häufigeit

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

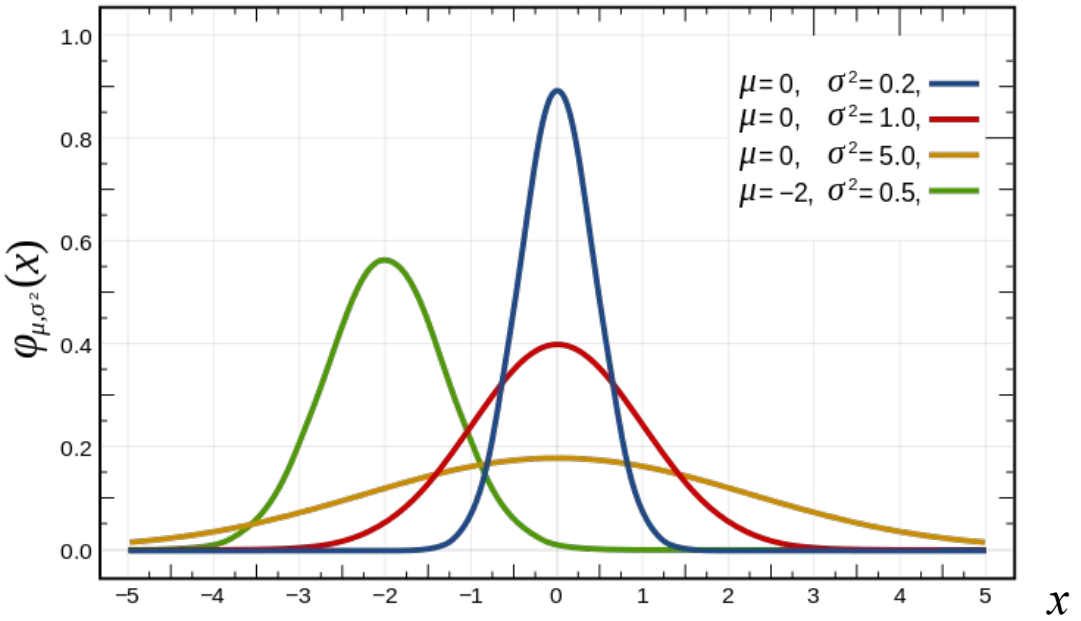


Normalverteilung

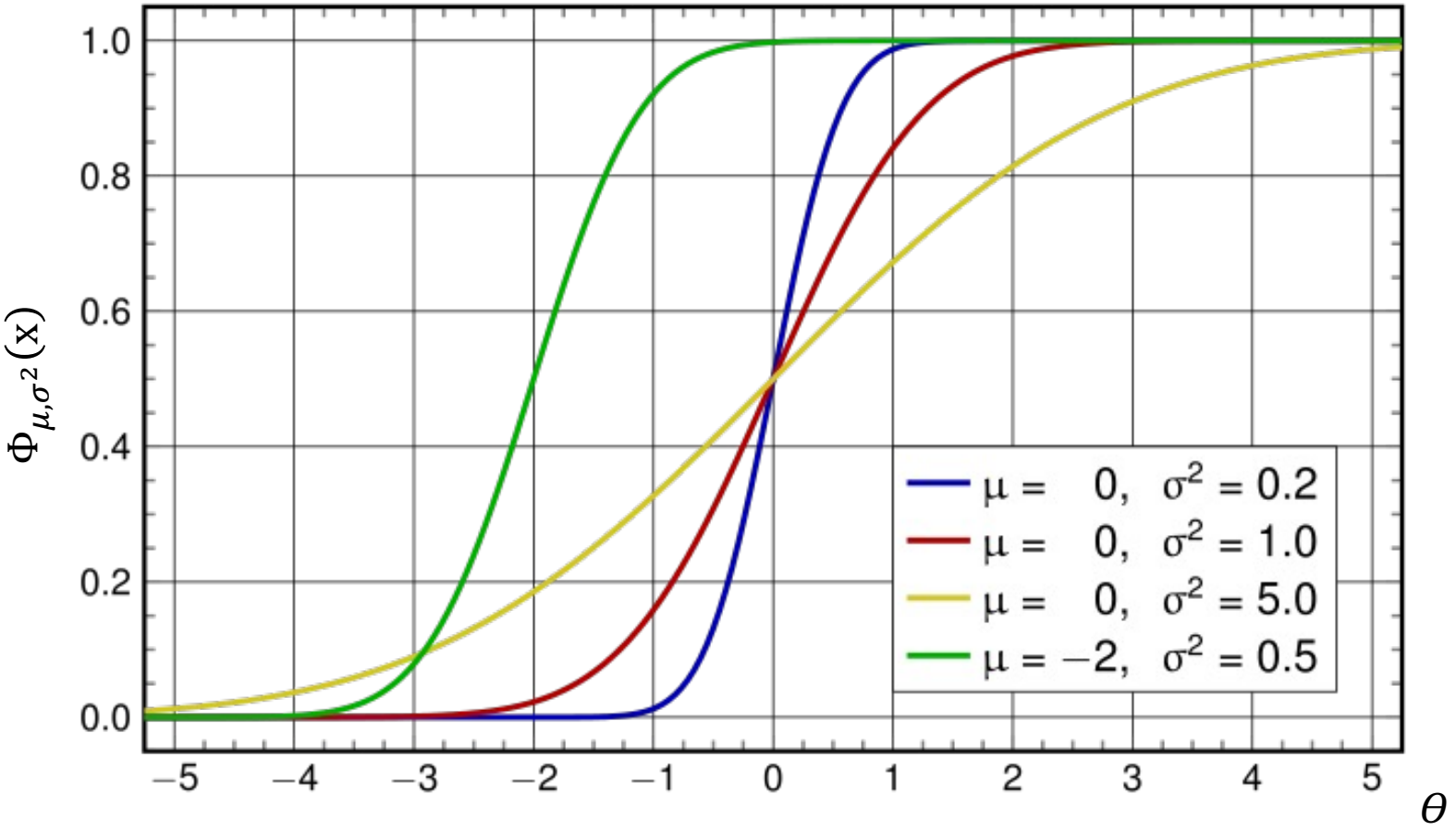
- Grundmenge: $\mathbb{R}$
- Lagemaß: Mittelwert μ
- Streuungsmaß: Varianz σ^2
- Funktion für Häufigkeitsverteilung wird im kontinuierlichen Fall Dichtefunktion genannt
- Es gilt:

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

Bei einer Normalverteilung sind Mittelwert und Median gleich



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 für ZV X?
  - Wahrscheinlichkeit für jede Merkmalsausprägung von X
- Geht nicht bei dichter Grundmenge für X
  - Wahrscheinlichkeit für jeden einzelnen Wert: 0
- Daher in diesem Fall: Dichtefunktion
- Verwendung der Dichte in Verteilungsfunktion
  - Bestimmung der Wahrscheinlichkeit, dass ein gewisses Ereignis $\mathrm{X} \leq x$ auftritt
  - Verteilungsfunktion für die Normalverteilung

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

  - Geht für $x \longrightarrow \infty$ gegen 1

Normalverteilung

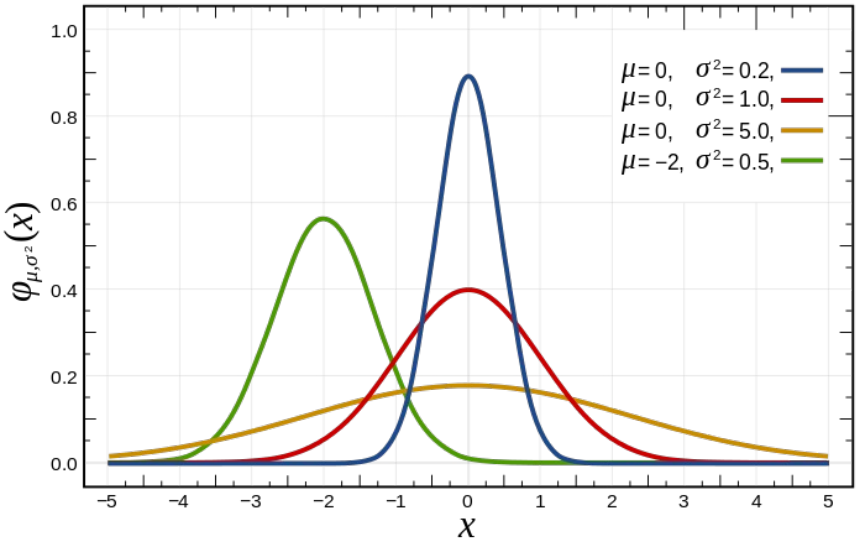
- Dichtefunktion

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

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

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

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



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

$\phi(x_0)$ Likelihood von x_0

Statistische Grundlagen

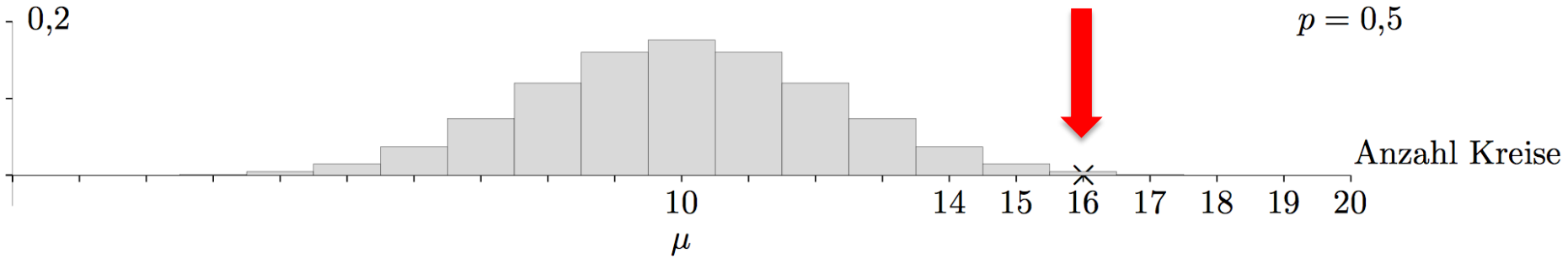
HYPOTHESENTEST

Hypothesentest

- Vermutung: Küken können Körner (rund) schon von Geburt an 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)

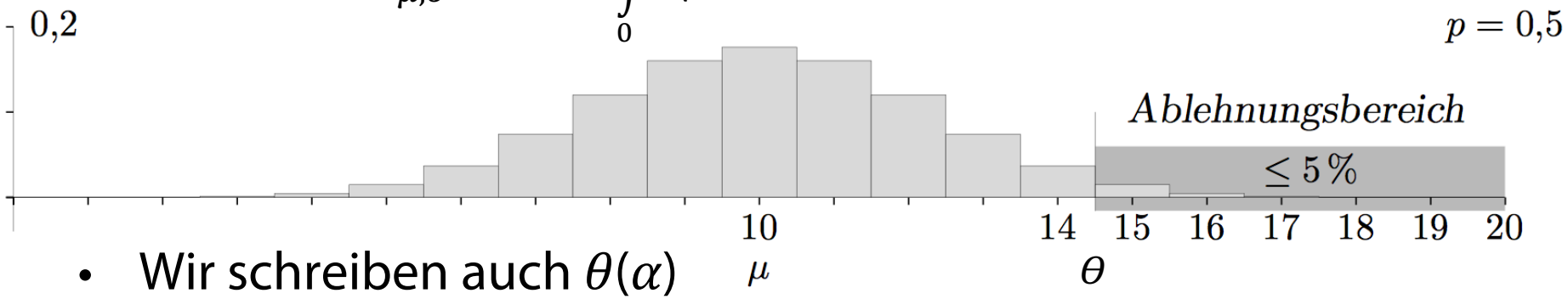


- Ausgang des Experiments:
  - Ausgang: Küken pickt im Mittel 16 mal auf Kreis
- Annahme: Wir wollen die Wahrscheinlichkeit minimieren, H_0 abzulehnen, obwohl sie wahr ist.

Ablehnungsbereich

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

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



- Wir schreiben auch $\theta(\alpha)$
- Fällt Test in Ablehnungsbereich, liegt signifikante Abweichung vor und 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

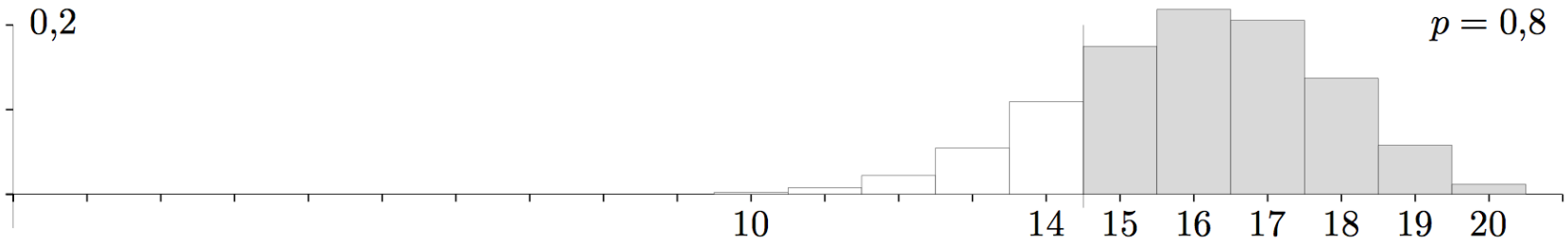
$$\int_0^{16} \varphi_{\mu,\sigma^2}(x)dx = 0.979$$

- Irrtumswahrscheinlichkeit sogar nur 0.021
- Wir sagen, der p-Wert beträgt p=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:

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

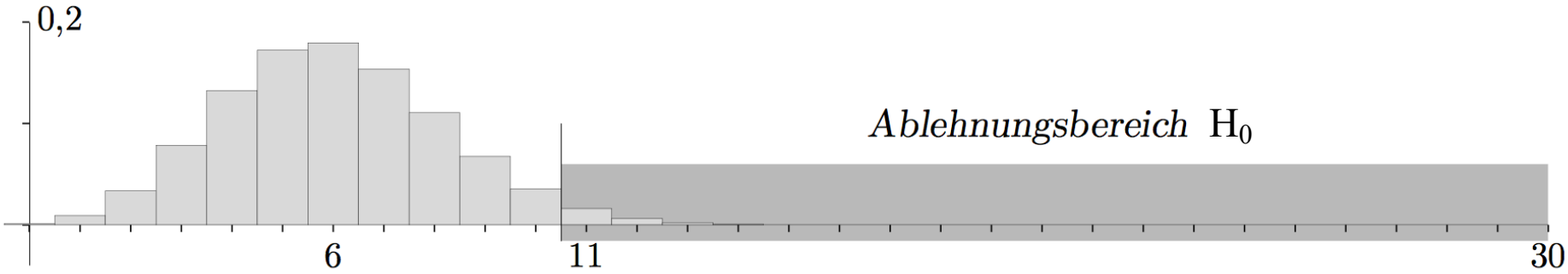


- Je mehr sich p dem Wert 0.5 nähert, umso größer wird der Fehler 2. Art

Ablehnungsbereichs rechts

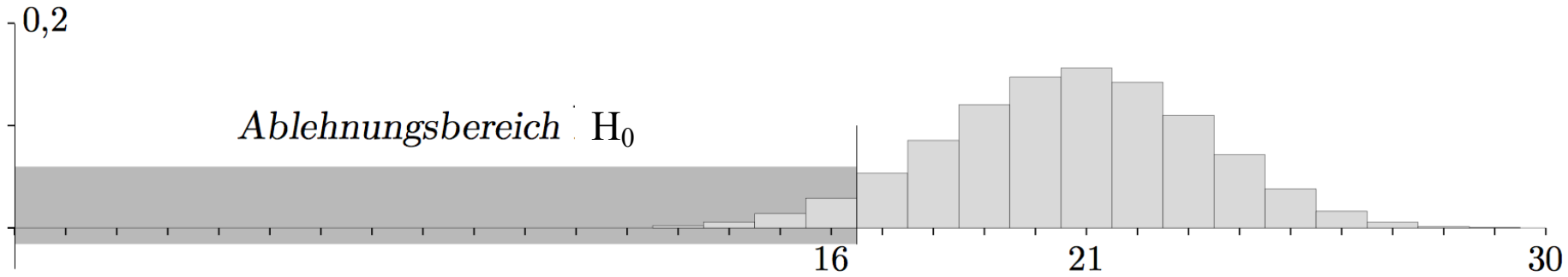
- Behauptung: Ein bestimmtes Medikament verursacht höchstens bei 20 % der Patienten Nebenwirkungen. Wir bezweifeln dies und testen die Nullhypothese auf dem 5%-Niveau.
- Die Stichprobenlänge sei n = 30
- Wähle H_0, wenn $p \leq \theta(\alpha)$ und H_1, wenn $p > \theta(\alpha)$

20% von 30 = 6



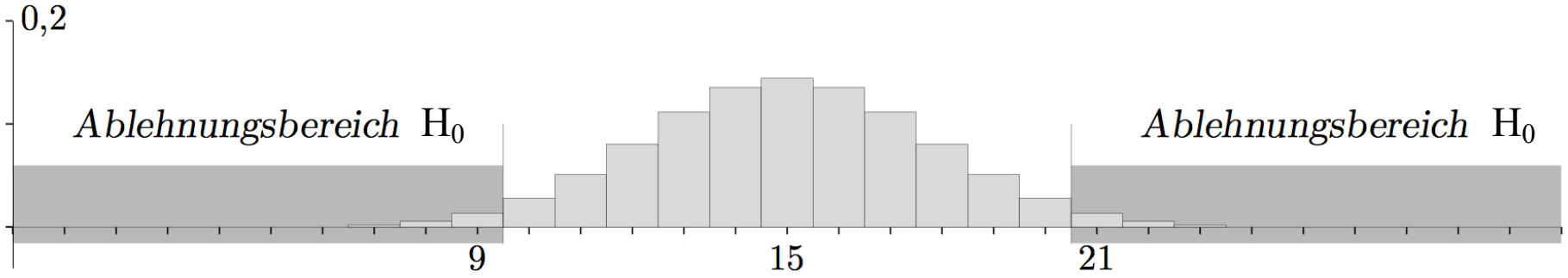
Ablehnungsbereichs links

- Behauptung: Mindestens 70 % der gelieferten Gurken erfüllen die europäische Krümmungsnorm. Wir vermuten das Gegenteil und testen auf dem 5%-Niveau.
- Die Stichprobenlänge sei n = 30
- Wähle H_0, wenn $p \geq \theta(1-\alpha)$ und H_1, wenn $p < \theta(1-\alpha)$



Ablehnungsbereichs beidseitig

- Bei der zufälligen Farbgebung sollen 50 % der Serienprodukte eine helle Tönung besitzen. Wir wollen Abweichungen aufdecken.
- Die Stichprobenlänge sei n = 30
- Wähle H_0, wenn $p \geq \theta(1-\alpha/2)$ und $p \leq \theta(\alpha/2)$; 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, i.A. gilt: $\alpha \neq 1-\beta$

Mögliche Resultate vom Hypothesentest

Wirklichkeit

Ergebnis Hypothesentest	H_0 **wahr**	H_1 **wahr**
H_0 **wahr**	Richtig $1 - \alpha$ 	Typ II Fehler β 
H_1 **wahr**	Typ I Fehler α 	Richtig $1 - \beta$ 

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 der 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?

Statistische Grundlagen

SCHÄTZER

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 Dichte von X

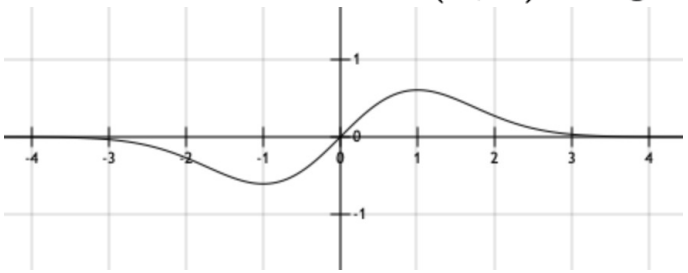
  - Notation manchmal auch: $\mathrm{E}[X]$

Erwartungswert der Standardnormalverteilung

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

Der Erwartungswert der Standardnormalverteilung ist 0. Es sei $X \sim \mathcal{N}(0,1)$, so gilt

$$\mathrm{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x\, e^{-\frac{1}{2}x^2}\, \mathrm{d}x = 0,$$

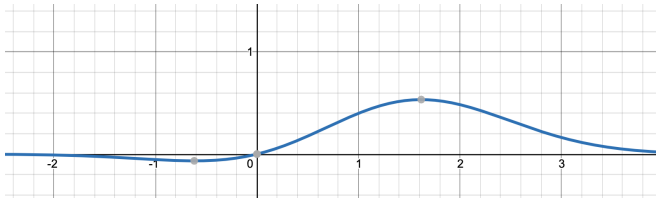


da der Integrand integrierbar und punktsymmetrisch ist.

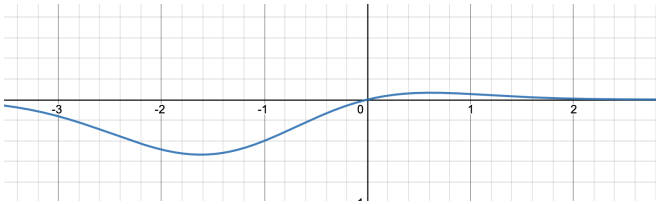
Erwartungswert

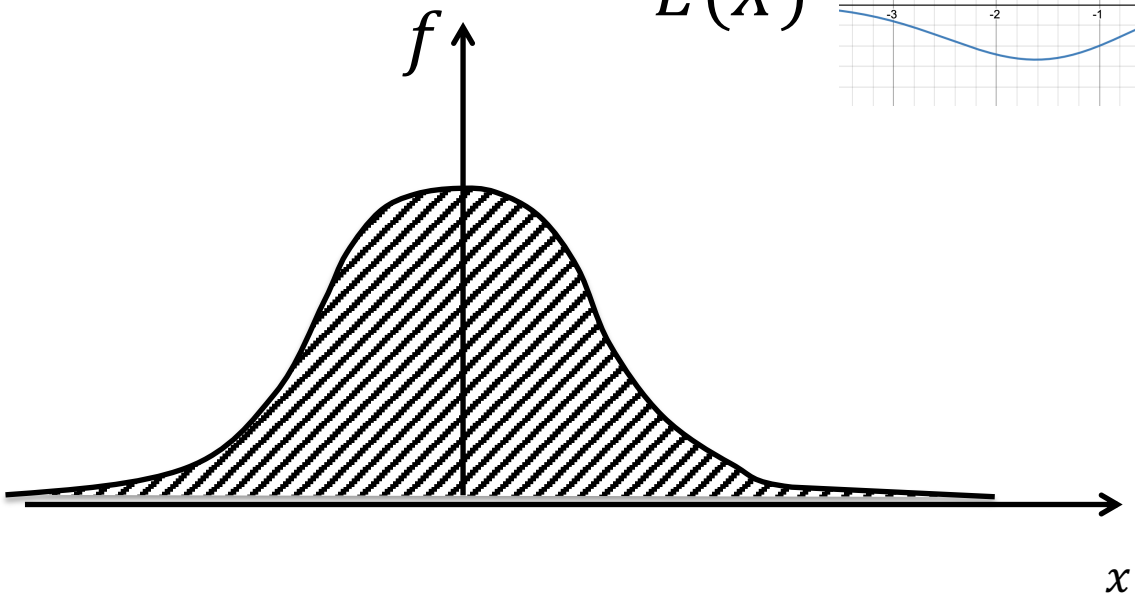
$$E[X] = \int x\, f(x) dx$$

$E(X)$



$E(X)$





Varianz formal

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

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

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

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

Mehrdimensionale Verteilungen

Wir betrachten eine zweidimensionale Verteilung mit Zufallsvariablen X und Y

Definition

Seien X und Y zwei Zufallsvariablen. Dann heißt

$$\sigma_{X,Y} := \mathrm{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right)$$

Kovarianz von X und Y.

Korrelation

Definition

Zwei Zufallsvariablen X und Y mit $\mathrm{Cov}(X, Y) = 0$ heißen unkorreliert.

Korrelation

Definition

Gegeben seien zwei Zufallsvariablen X und Y. Dann heißt

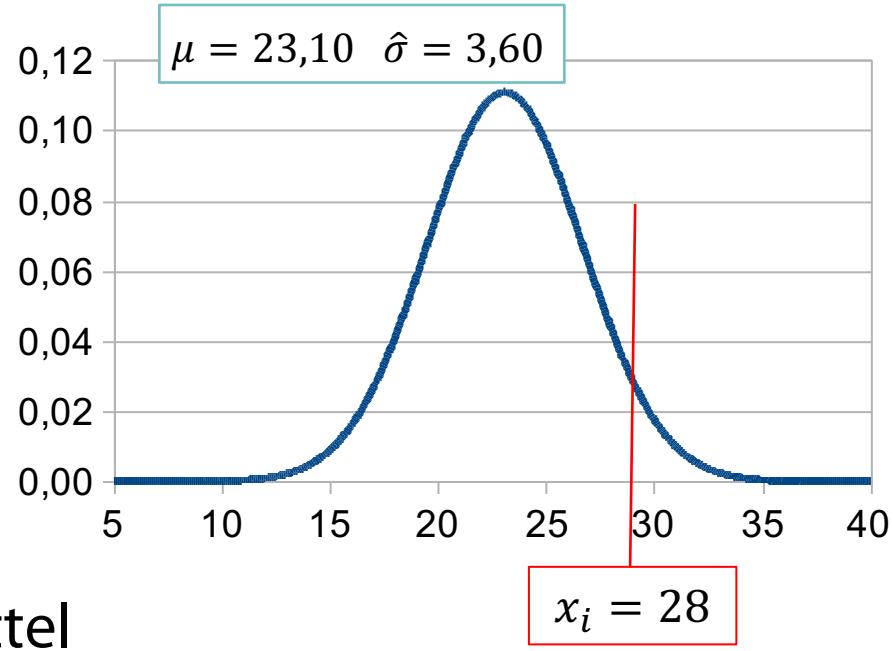
$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Korrelationskoeffizient von X und Y.

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

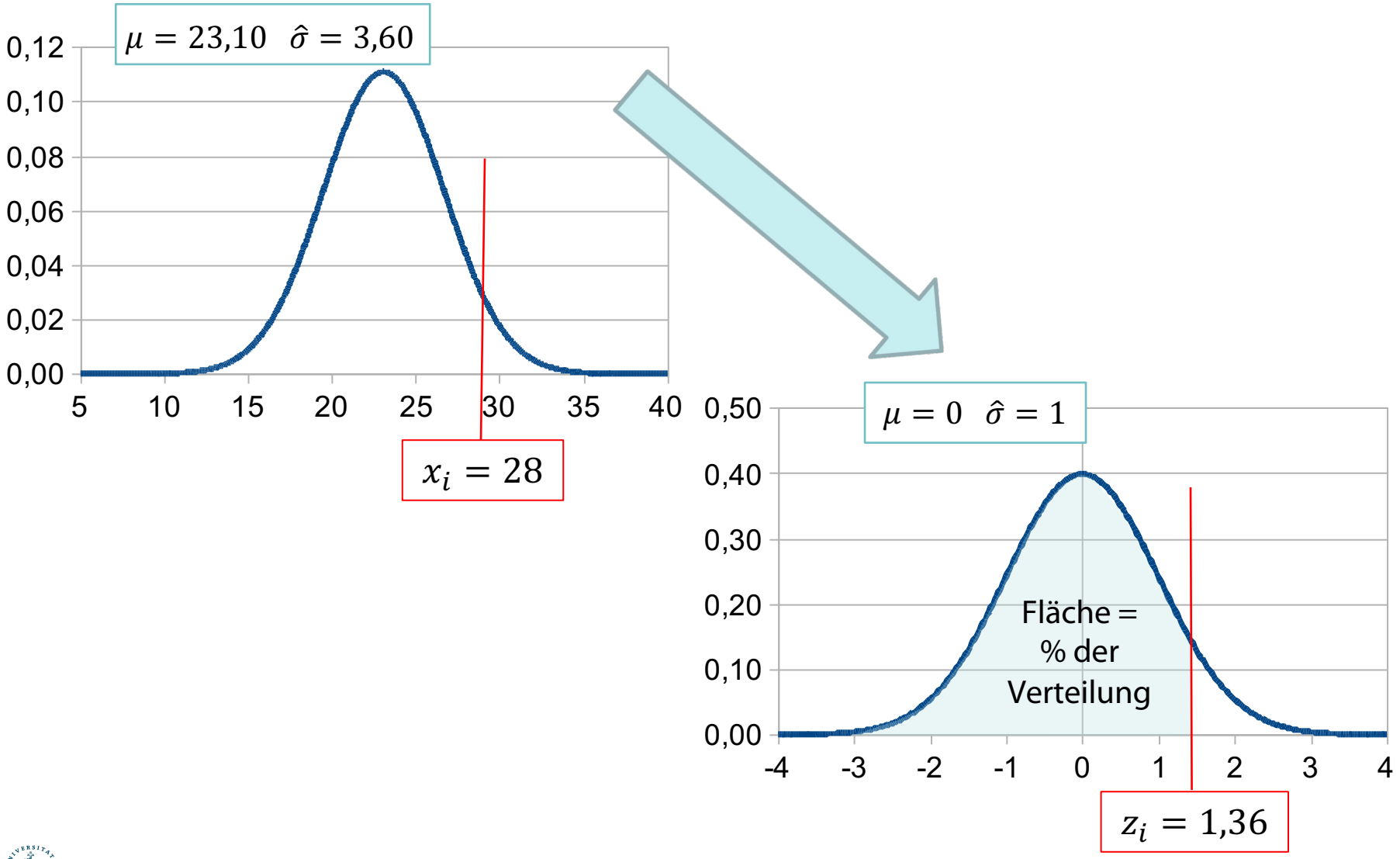


z-Standardisierung

- 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

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

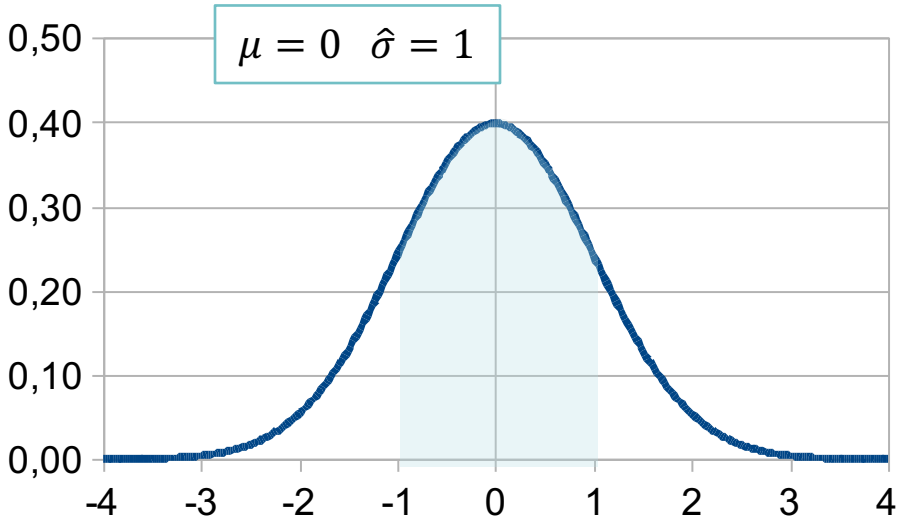
Generell gilt für normalverteilte Merkmale:

- ***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. $-2{,}0 < z_i < 2{,}0$

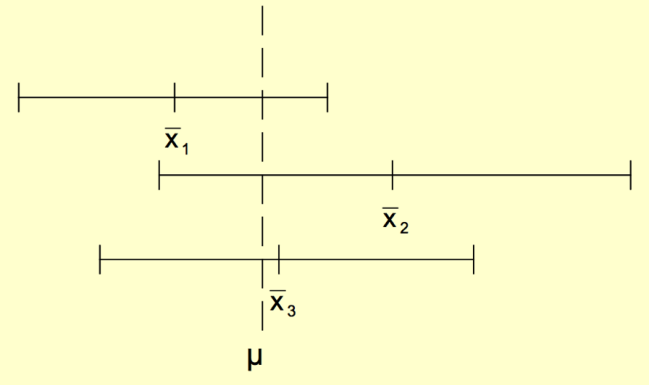


Stichprobenkennwerte

- 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

Stichprobenkennwertverteilungen

- 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 Stichprobenkennwertverteilung des Mittelwerts



Standardfehler

- Diese „Verteilung der Mittelwerte“ ist selbst wieder normalverteilt (wenn das Merkmal normalverteilt ist)
- Der Mittelwert der Stichprobenkennwertverteilung für die Mittelwerte der Stichproben 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:

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

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: $\mathrm{P}(57 < \mu < 63\ ) = p$

Statistische Grundlagen

KONFIDENZINTERVALLE

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

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{mit } \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

ZV= 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 $\bar{X}-\mu \geq 0$: $\bar{X}-\mu \leq d \rightarrow \bar{X}-d \leq \mu$
  - Fall $\bar{X}-\mu \leq 0 = -(\bar{X}-\mu) \geq 0$: $-\bar{X}+\mu \leq d \rightarrow \bar{X}+d \geq \mu$
- 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: $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ $Z \sim \mathcal{N}(0,1)$

Plan: In $\mathrm{P}(|\bar{X}-\mu| \leq d)$ einsetzen

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

Konfidenzintervall: Herleitung

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

  Abkürzung: $\mathrm{z}_{1-\frac{\alpha}{2}} = \frac{d}{\sigma}\cdot\sqrt{n}$

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

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

$$= \mathrm{P}\left(|Z| \leq \mathrm{z}_{1-\frac{\alpha}{2}}\right)$$

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

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

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

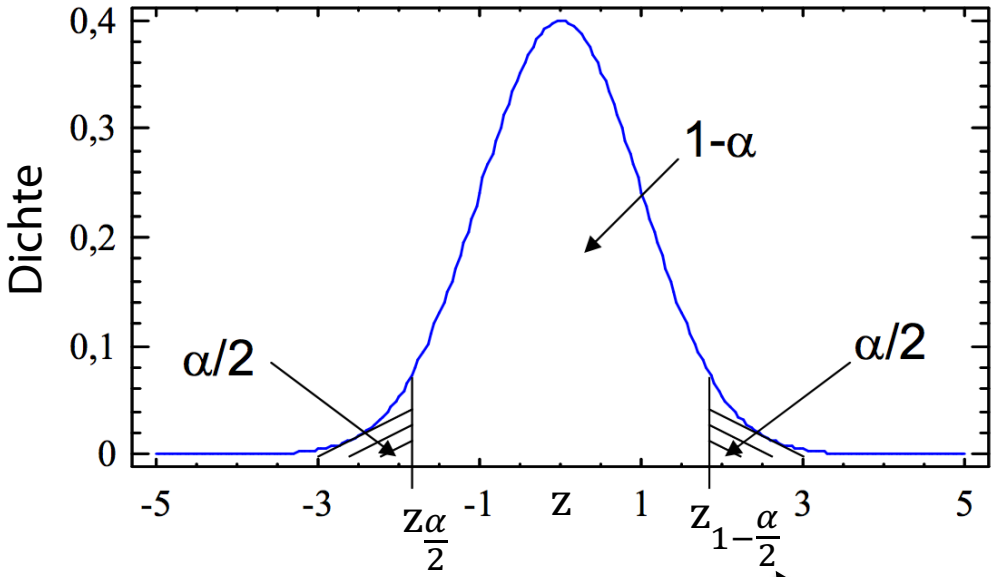
$$\mathrm{z}_{\frac{\alpha}{2}} = -\mathrm{z}_{1-\frac{\alpha}{2}}$$

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:

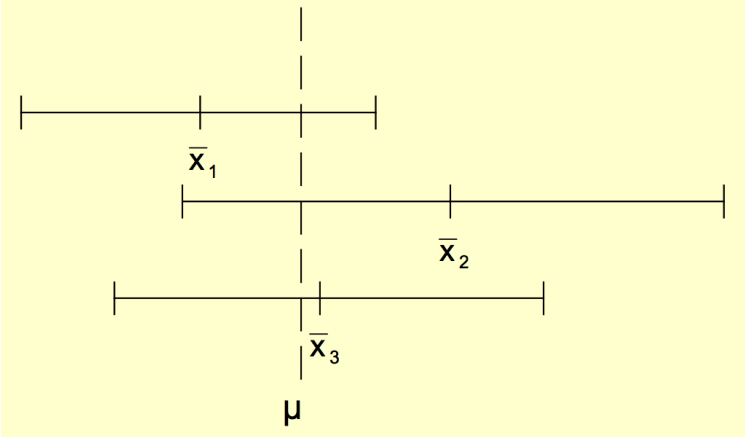
$1-\alpha$	α	$z_{1-\frac{\alpha}{2}}$	$z_{1-\alpha}$
0,95	0,05	1,96	1,64
0,99	0,01	2,58	2,33
0,999	0,001	3,29	3,09

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

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

Konfidenzintervall: Bemerkungen

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

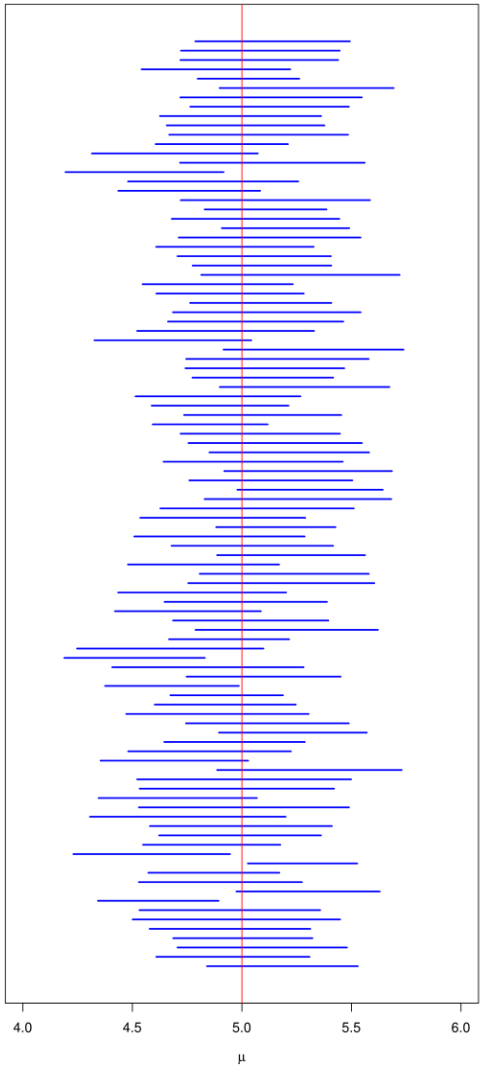


- Bei einem Konfidenzniveau von $(1 - \alpha) = 0{,}95$ heißt das:
  - In 95% aller Fälle enthält das Konfidenzintervall den unbekannten Parameter der GG und in 5% der Fälle nicht.
  - D.h.: Behauptet man k mal, der unbekannte Parameter liege im Vertrauensbereich, so hat man im Mittel $\alpha \cdot k$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

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

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

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

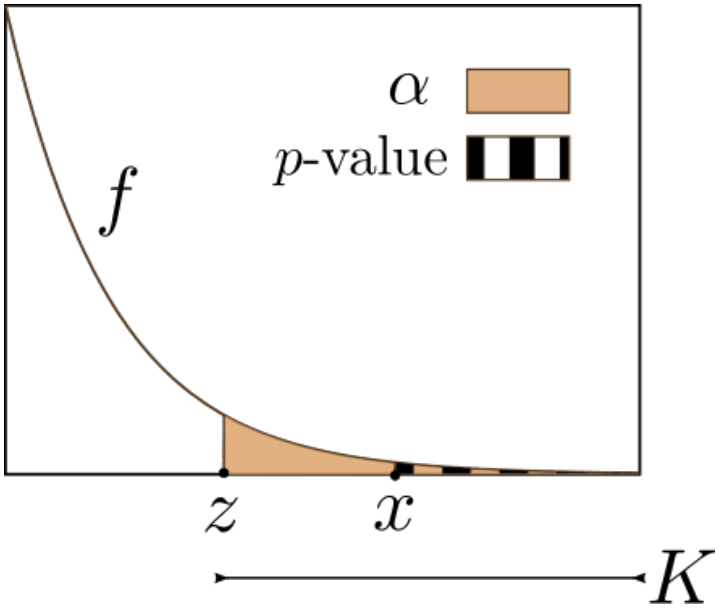
$$n = \frac{\sigma^2}{d^2} \cdot z^2_{1-\frac{\alpha}{2}}$$

Konfidenzintervall für Varianz

- Ähnliche Überlegungen
- Auch hierfür Herleitung der erforderlichen Stichprobengröße möglich

P-Wert (einseitiger Ablehnungsbereich)

- Hypothesentest H_0 vs. H_1
- Wie extrem ist der auf Basis der erhobenen Daten berechnete Wert der Teststatistik?
- P-Wert = Wahrscheinlichkeit, bei Gültigkeit von H_0 den bestimmten oder einen extremeren Wert der Teststatistik zu erhalten



Für diese Realisation x im Ablehnbereich K ist der p-Wert kleiner als α, oder dazu äquivalent ist die Realisation der Teststatistik x größer als der kritische Wert z. Hier ist f die Wahrscheinlichkeitsdichte der Verteilung unter der Nullhypothese

In manchen Veröffentlichungen wird α als p-Wert bezeichnet!

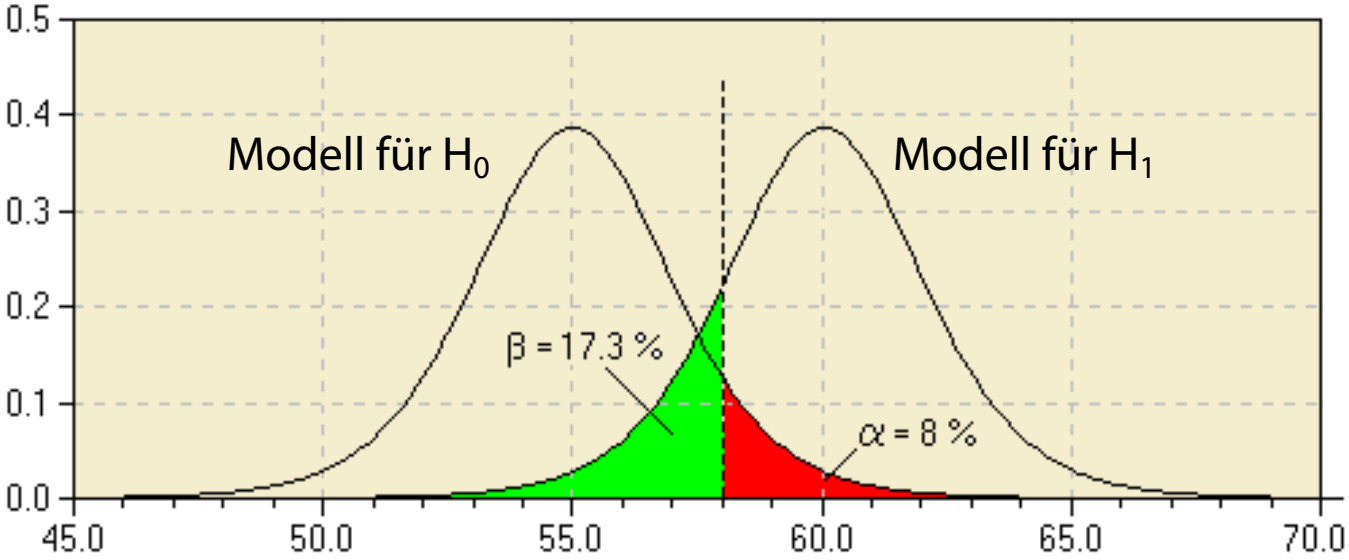
Nicht nur α ist relevant

- Uns interessiert auch:
  - Mit welcher Wahrscheinlichkeit weist ein statistischer Test die abzulehnende Nullhypothese H_0 korrekt zurück, wenn die Alternativhypothese H_1 wahr ist
- Interpretiert als „Trennschärfe“ des Tests
  - Hohe Trennschärfe des Tests spricht gegen, niedrige Trennschärfe für die Nullhypothese H_0
- Ziel:
  - Ablehnungsbereich A so bestimmen, dass die Wahrscheinlichkeit für die Ablehnung einer „falschen Nullhypothese“ H_0, d. h. für Annahme der Alternativhypothese H_1 unter der Bedingung, dass H_1 wahr ist, möglichst groß ist

Trennschärfe (Power) eines Tests

	H_0 ist wahr	H_1 ist wahr
Durch einen statistischen Test fällt eine **Entscheidung für** H_0	Richtige Entscheidung (Spezifität) Wahrscheinlichkeit: 1 - α	**Fehler 2. Art** Wahrscheinlichkeit: β
Durch einen statistischen Test fällt eine **Entscheidung für** H_1	**Fehler 1. Art** Wahrscheinlichkeit: α	richtige Entscheidung Wahrscheinlichkeit: 1-β (Trennschärfe des Tests, Sensitivität)

- Trennschärfe hat den Wert 1-β
- Wahl des β-Niveaus?
- Modell für H_1 benötigt



Determinanten der Trennschärfe

Die Trennschärfe (1-β) wird größer:[6]

- mit wachsender Differenz von $(\mu_0 - \mu_1)$
- mit kleiner werdender Merkmalsstreuung σ
- mit größer werdendem Signifikanzniveau α (sofern β nicht festgelegt ist)
- mit wachsendem Stichprobenumfang, da der Standardfehler dann kleiner wird: $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$
- bei einseitigen Tests im Vergleich zu zweiseitigen Tests: Für den zweiseitigen Test braucht man einen etwa um 25 % größeren Stichprobenumfang, um dieselbe Trennschärfe wie für den einseitigen Test zu erreichen.

[Wikipedia]

Weitere Gütekriterien von Schätzern

- Erwartungstreue
- Konsistenz
- Effizienz
- Reliabilität
- Validität
- Objektivität

Es soll hier nur ein grober Überblick vermittelt werden

Statistische Grundlagen

KORRIGIERTE STICHPROBENVARIANZ

Stichprobenfunktion

- In der Statistik fasst eine Stichprobenfunktion, auch Stichprobenstatistik oder schlicht Statistik, Informationen aus einer Stichprobe in spezifischer Form als Funktion zusammen.
- Beispiel für Stichprobenfunktion: Schätzfunktion
- Notation:

Arithmetisches Mittel	Stichprobenfunktion
$\overline{x} := \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$	$\overline{X} := \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

x_i: konkrete Werte

X_i: noch zu bestimmende Werte (Zufallsvariablen)

Name der Schätzfunktion

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 $x_i \sim \mathcal{N}(\mu, \sigma^2)$ zufällig aus GG gezogen, dann $\mathrm{E}(\bar{x}) = \mu$
  - Erwartungswert von $\bar{x}$:

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

- Stichprobenmittel also erwartungstreuer Schätzer von μ

[Wikipedia]

Korrigierte Stichprobenvarianz

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

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

Warum $n-1$? wenn Stichproben-mittel $\bar{x}$ verwendet?

  - 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

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

μ nicht $\bar{x}$

- 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 $n-1$?

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

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

Wir unterscheiden:
empirische und korrigierte
Varianz

Statistische Grundlagen

GÜTE VON SCHÄTZERN

Konsistenz und Effizienz eines Schätzers

- Ein Schätzer ist konsistent, wenn er für immer größere Stichproben immer genauer wird
  - Man die Schätzung beliebig genau machen, indem man die Stichprobengröße weit genug erhöht
- Die Effizienz (Wirksamkeit) des Schätzwertes kennzeichnet die Präzision, mit der er Parameter schätzt
  - Eine Schätzfunktion ist um so effizienter, je kleiner die Streuung (oder Varianz) der Schätzwerte um den Parameter ist
  - Je größer die Streuung der Stichprobenkennwertverteilung, desto geringer ist die Effizienz des entsprechenden Schätzwertes

Reliabilität / Zuverlässigkeit

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 zusammengefassten 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)

Wiederholung

Gegeben seien zwei Zufallsvariablen X und Y. Dann heißt

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Korrelationskoeffizient von X und Y.

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 r zwischen allen Einzelitems
  - Ausreichende Reliabilität: r: = 0.75
  - Gute Reliabilität: r = 0.90

Weitere Gütekriterien

- Validität: Misst ein Test das, was er messen soll?

- Objektivität: Unabhängigkeit der Versuchsergebnisse von den Rahmenbedingungen (Randbedingungen) und verfälschenden Drittfaktoren

Statistische Grundlagen

UNTERSCHIEDSHYPOTHESEN

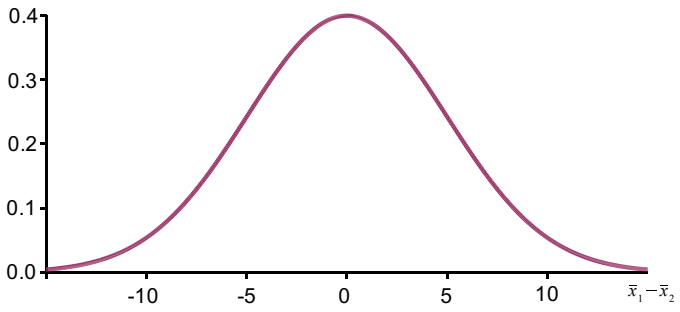
Unterschiedshypothesen

- Bekommen Frauen mehr Gehalt als Männer?
  - Unterscheiden sich die Mittelwerte von zwei Gruppen?
  - Unabhängige Stichproben
- Ist der Mittelwert der Gehälter nach einer Fortbildung größer als vor der Fortbildung?
  - 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 vorgegebenen Wert?
  - Test bzgl. Gruppe

Unterschiedshypothesen: Unabhängige Stichproben

Unterscheiden sich die Mittelwerte von zwei Gruppen?

- Differenz der Mittelwerte zweier Stichproben: $\Delta_x = \bar{x}_1 - \bar{x}_2$
- H_0: Differenz irrelevant
- Schätze die Dichtefunktion für Δ_x wenn H_0 war ist
- Stichprobenkennwerteverteilung: Verteilung der Mittelwertsdifferenzen unter ***H_0***
- Wie verteilen sich empirische Mittelwertsdifferenzen, wenn man sehr oft Stichproben zieht?
- Verteilung von Mittelwerts-differenzen bei großen Stichproben normalverteilt



Standardfehler der Kennwerteverteilung

- Für Mittelwert

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

- Für Differenzen hängt er von den Varianzen und den Größen der beiden Teilstichproben ab:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

- Benötigt, um gefundene Mittelwertsdifferenz interpretieren zu können

t-Verteilung

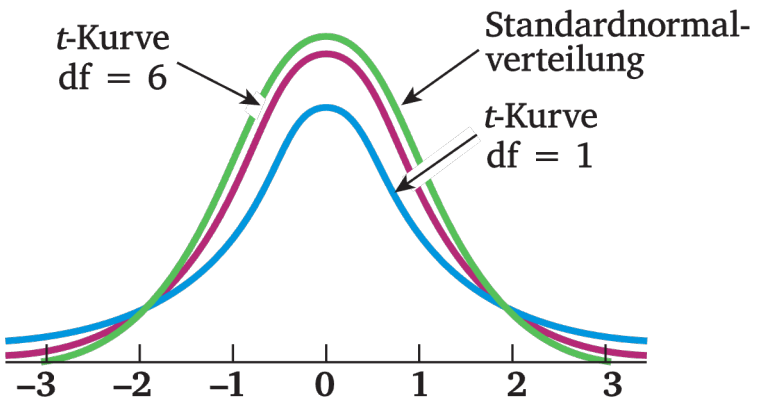
- Empirische (gefundene) Mittelwertsdifferenz durch Standardfehler dividiert ergibt sog. t-Verteilung

$$t_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

- Die genaue Form der t-Verteilung hängt von deren Freiheitsgraden (*df* = *degree of freedom*) ab

$$df = N_1 + N_2 - 2$$

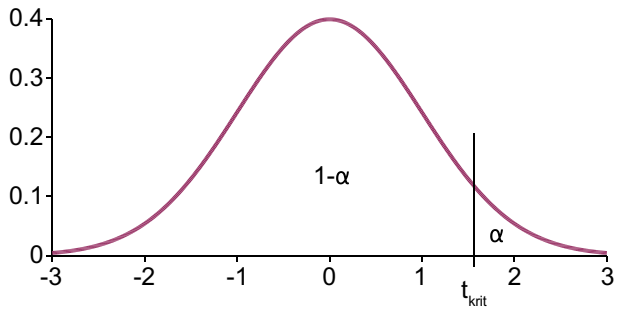
  - Bei *df* >120 nahezu identisch mit *z*-Verteilung (St.Norm.V.)
  - Je kleiner df, desto schmalgipfliger die t-Verteilung



Die *t*-Verteilung

df	p=0,8	p=0,9	p=0,95	p=0,975	p=0,99	p=0,995
1	1,376	3,078	6,314	12,706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
20	0,860	1,325	1,725	2,086	2,528	2,845
30	0,854	1,310	1,697	2,042	2,457	2,750
40	0,851	1,303	1,684	2,021	2,423	2,704
50	0,849	1,299	1,676	2,009	2,403	2,678
60	0,848	1,296	1,671	2,000	2,390	2,660
70	0,847	1,294	1,667	1,994	2,381	2,648
80	0,846	1,292	1,664	1,990	2,374	2,639
90	0,846	1,291	1,662	1,987	2,368	2,632
100	0,845	1,290	1,660	1,984	2,364	2,626
200	0,843	1,286	1,653	1,972	2,345	2,601
1000	0,842	1,282	1,646	1,962	2,330	2,581

Für einen 2-seitigen Test muss t_{krit} so gewählt werden, dass ein Bereich von α/2 „von der Verteilung abgeschnitten wird“



Kritische t-Werte:

α = .05, einseitig, *df*=100:

$t_{krit}(100) = 1.66$

α = .05, zweiseitig, *df*=100:

$t_{krit}(100) = 1.98$

α = .01, einseitig, *df*=100:

$t_{krit}(100) = 2.36$

Voraussetzungen für t-Test-Anwendung

(1) Variable besitzt Intervallskala (arithm. Mittel ist definiert)

(2) Normalverteilung des Merkmals in der Grundgesamtheit
- Kann geprüft werden (Kolmogorov-Smirnov-Test)
  - Hier nicht vertieft

(3) Varianzhomogenität
- „Gleiche“ Varianzen des Merkmals in beiden Populationen
- „Varianz der Varianz“ klein
- Kann geprüft werden (Levene-Test)
  - Hier nicht vertieft

(4) Unabhängigkeit der Stichproben

Unterschiedshypothesen: Abhängige Stichproben

- Ziehung eines Merkmalsträgers in die erste Stichprobe beeinflusst die Zugehörigkeit eines Merkmalsträgers zur zweiten Stichprobe
- Werte zweier Stichproben paarweise zugeordnet.
  - Beide Teilstichproben immer gleich groß!
- Messwiederholung
  - Gleiches Merkmal zweimal (oder mehrmals) bei den gleichen Personen erhoben
- Parallelisierung
  - Jeweils ähnliche 2 Personen einander zugeordnet
- Matching
  - Jeder Person der Stichprobe 1 ist einer Person der Stichprobe 2 zugeordnet

Abhängige Stichproben: Beispielrechnung

- Verändert sich die Einstellung zum Studienfach Informatik innerhalb der ersten 6 Wochen des Studiums?
- Abh. Variable: Einstellung zum Studium Informatik (Wertebereich 5 bis 25)
- Unabh. Variable: Messzeitpunkt (1. Woche vs. 6. Woche)

Versuchs-person	1. Woche	6. Woche
1	16	20
2	18	19
3	23	23
4	14	16
	…	…
mean	19.67	18.98

Beispielrechnung

- Für jede Person kann die Differenz der Messwerte berechnet werden (Einstellungsänderung)

Vp	1. Woche	6. Woche	$D=x_2-x_1$
1	16	20	4
2	18	19	1
3	23	23	0
4	16	14	-2
	…	…	…
mean	19.67	18.98	.68

Hypothesen

- Die statistischen Hypothesen des t-Tests für abhängige Stichproben beziehen sich auf den Mittelwert der Differenzen aller Personen
  - Vorteil: Es ist nun unerheblich, ob innerhalb der Messzeitpunkte große Varianz gegeben ist.
- Ungerichtete Hypothese:
  - H_0: $\mu_d = 0$
  - H_1: $\mu_d \neq 0$
- Gerichtete Hypothese (1):
  - H_0: $\mu_d \leq 0$
  - H_1: $\mu_d > 0$
- Gerichtete Hypothese (2):
  - H_0: $\mu_d \geq 0$
  - H_1: $\mu_d < 0$

Standardfehler und *t*-Wert

- Um die empirisch gefundene Differenz beurteilen zu können, wird der Standardfehler benötigt

$$\hat{\sigma}_{\bar{x}_d} = \frac{\hat{\sigma}_{x_d}}{\sqrt{N}} \qquad mit \quad \hat{\sigma}_{x_d} = \sqrt{\frac{\sum_{i=1}^{N} (x_{di} - \bar{x}_d)^2}{N-1}}$$

Basierend auf korrigierter Stichprobenvarianz

- Mit dem Standardfehler kann nun ein empirischer normalisierter *t*-Wert berechnet werden
  - Normalisierung bzgl. Standardabweichung (vgl. z-Standardisierung)

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

$$t_{df} = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}} \text{ mit } df = N-1$$

Standardfehler und *t*-Wert

Im Beispieldatensatz:

$$\bar{x}_d = 0.68$$

$$\hat{\sigma}_{x_d} = 2.78$$

$$N = 60$$

- Es ergibt sich :

$$\hat{\sigma}_{\bar{x}_d} = \frac{2.78}{\sqrt{60}} = 0.36$$

$$t_{59} = \frac{0.68}{0.36} = 1.89$$

Kritischer *t*-Wert & Interpretation

- $T_{emp,59} = 1.89$
- $T_{krit,59} = ?$
  - Offene Fragestellung
    ⇨ zweiseitiger Test
  - $\alpha = .05$
- Interpretation:
  - $t_{emp} < t_{krit}$
  - Also: Kein bedeutsamer Unterschied!

df	p=0,8	p=0,9	p=0,95	p=0,975	p=0,99	p=0,995
1	1,376	3,078	6,314	12,706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
20	0,860	1,325	1,725	2,086	2,528	2,845
30	0,854	1,310	1,697	2,042	2,457	2,750
40	0,851	1,303	1,684	2,021	2,423	2,704
50	0,849	1,299	1,676	2,009	2,403	2,678
60	0,848	1,296	1,671	2,000	2,390	2,660
70	0,847	1,294	1,667	1,994	2,381	2,648
80	0,846	1,292	1,664	1,990	2,374	2,639
90	0,846	1,291	1,662	1,987	2,368	2,632
100	0,845	1,290	1,660	1,984	2,364	2,626
200	0,843	1,286	1,653	1,972	2,345	2,601
1000	0,842	1,282	1,646	1,962	2,330	2,581

Eingruppen *t*-Test

- Ziel: Vergleich des Mittelwerts einer Stichprobe mit einem vorgegebenen (konstanten) Wert
- Beispiele:
  - Prüfe, ob eine bestimmte Personengruppe sich in ihrer Intelligenz vom Populationsmittelwert (100) unterscheidet
  - Prüfe, ob sich die tatsächliche Studiendauer von der Regelstudienzeit unterscheidet
  - Prüfe, ob sich die Differenz von Reaktionszeiten unter zwei Bedingungen von Null unterscheidet

Eingruppen *t*-Test

Voraussetzungen

- *Normalverteilung* des Merkmals
- *Intervallskalenniveau* des Merkmals
- Es handelt sich um eine *Zufallsstichprobe*

Eingruppen *t*-Test

Statistische Hypothesen

- Ungerichtete Hypothese:
  - H_0: $\mu = c$
  - H_1: $\mu \neq c$
- Gerichtete Hypothese (1):
  - H_0: $\mu \leq c$
  - H_1: $\mu > c$
- Gerichtete Hypothese (2):
  - H_0: $\mu \geq c$
  - H_1: $\mu < c$

Standardfehler und *t*-Wert

- Berechnung des Standardfehlers

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

- Berechnung des t-Werts

$$t\left(df = N - 1\right) = \frac{\bar{x} - c}{\hat{\sigma}_{\bar{x}}}$$

Beispiel

- Liegt der IQ der Kinder, die als hochbegabten klassifiziert werden, wirklich über dem Populationsmittelwert (100)?
- Hypothesen:
  - H_0: $\mu \leq 100$
  - H_1: $\mu > 100$
- Stichprobenkennwerte bei N=10:
  - Mittelwert: 108.50
  - Standardabweichung: 14.35

$$\hat{\sigma}_{\bar{x}} = \frac{14.35}{\sqrt{10}} = 4.54 \qquad t(9) = \frac{108.5 - 100}{4.54} = 1.87$$

Beispiel

- $t_{emp}(9) = 1.87$
- $t_{krit}(9) = ?$
  - Gerichtete Fragestellung
    ⇨ einseitiger Test
  - $\alpha = .05$
- Interpretation:
  - $t_{emp} > t_{krit}$
  - H_0 wird verworfen

df	p=0,8	p=0,9	p=0,95	p=0,975	p=0,99	p=0,995
1	1,376	3,078	6,314	12,706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
20	0,860	1,325	1,725	2,086	2,528	2,845
30	0,854	1,310	1,697	2,042	2,457	2,750
40	0,851	1,303	1,684	2,021	2,423	2,704
50	0,849	1,299	1,676	2,009	2,403	2,678
60	0,848	1,296	1,671	2,000	2,390	2,660
70	0,847	1,294	1,667	1,994	2,381	2,648
80	0,846	1,292	1,664	1,990	2,374	2,639
90	0,846	1,291	1,662	1,987	2,368	2,632
100	0,845	1,290	1,660	1,984	2,364	2,626
200	0,843	1,286	1,653	1,972	2,345	2,601
1000	0,842	1,282	1,646	1,962	2,330	2,581

Zusammenfassung der 3 Arten des *t*-Tests

	unabhängige Stichproben	abhängige Stichproben	Eingruppen *t*-Test
Fragestellung			
Voraus-setzungen			

Testverfahren

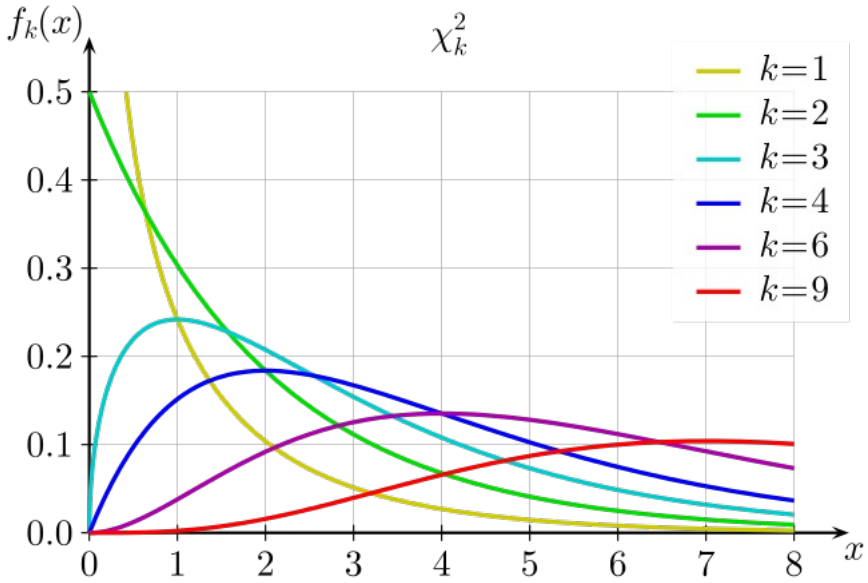
- ***Parametrische Verfahren***
  - Beteiligte Variablen müssen geforderte Verteilungsform aufweisen (z.B. Normalverteilung für den *t*-Test)
  - Intervallskalen erforderlich
  - Dann aber gute "Aussagekraft“ (siehe den Begriff der Trennschärfe)
- ***Nonparametrische Verfahren*** werden eingesetzt…
  - ⇨ Für die Analyse von ordinal- oder nominalskalierten Variablen
  - ⇨ Wenn die Normalverteilungsannahme des Gesamtmerkmals verletzt ist
    - ⇨ Beispiel: Summe der Quadrate von k normalverteilten Zufallsvariablen ist nicht normalverteilt

Statistische Grundlagen

CHI-QUADRAT-TEST

χ^2-Verteilung

χ^2 ist eine der Verteilungen, die aus der Normalverteilung $\mathcal{N}\left(\mu, \sigma^2\right)$ abgeleitet werden kann: Hat man k Zufallsvariablen Z_i, die unabhängig und standardnormalverteilt sind, so ist die Chi-Quadrat-Verteilung mit k Freiheitsgraden definiert als die Verteilung der Summe der quadrierten Zufallsvariablen $Z_1^2 + \cdots + Z_k^2$



[Wikipedia]

Der χ^2 -Test

- Der ***χ^2-Test*** („Chi-Quadrat-Test“) dient dem Vergleich von *beobachteten* und *erwarteten* Häufigkeiten. Er kann eingesetzt werden, wenn 1 oder 2 nominalskalierte unabhängige Variablen vorliegen.

	Merkmal		
	Auspr. 1	...	Auspr. k
Beobachtet	$f_{b,1}$		$f_{b,k}$
Erwartet	$f_{e.1}$		$f_{e,k}$

Beispiele:

- Leiden junge und alte Personen gleich häufig an einer bestimmten Erkrankung?
- Leisten hoch-ängstlich und gering-ängstliche Personen gleich häufig Hilfe in einer Notsituation?

Der χ^2 -Test

Voraussetzung für den χ^2 -Test (Faustregeln)

(1) Weniger als 1/5 aller Zellen hat eine *erwartete Häufigkeit* kleiner als 5.

(2) Keine Zelle weist eine *erwartete Häufigkeit* kleiner als 1 auf.

Wenn Voraussetzungen nicht erfüllt → andere Tests

Der χ^2 -Test

χ^2 -Test – Beispiel 1

- Es soll geprüft werden, ob die Verteilung von Männern und Frauen in einer Gruppe signifikant von einer Gleichverteilung abweicht
- $N = 76$ (Frauen: 56; Männer: 20)
- Statistische Hypothesen
  - $\boldsymbol{H_0}$: π(Frau) = π(Mann)
  - $\boldsymbol{H_1}$: π(Frau) ≠ π(Mann)

π(x) = Relative Häufigkeit, dass Merkmalswert x auftritt

Der χ^2 -Test

Schritt 1:

- Zunächst werden die nach der H_0 zu erwarteten Häufigkeiten berechnet:
- *Beobachtet:* $N_F = 56; N_M = 20$
- *Erwartet*: ???
  - Gesamtzahl: 76
  - Bei einer Gleichverteilung wären also ☐ Männer und ☐ Frauen zu erwarten.

Der χ^2 -Test

Schritt 2:

- Nun wird der (empirische) χ^2-Wert berechnet:

$$\chi^2_{df=k-1} = \sum_{i=1}^{k} \frac{\left(f_{b,i} - f_{e,i}\right)^2}{f_{e,i}}$$

	Merkmal		
	Auspr. 1	...	Auspr. k
Beobachtet	$f_{b,1}$		$f_{b,k}$
Erwartet	$f_{e,1}$		$f_{e,k}$

mit:

- k: Anzahl der Stufen der beiden Variablen
- $f_{b,i}$: Beobachtete Häufigkeit in der Zelle (i)
- $f_{e,i}$: Erwartete Häufigkeit in der Zelle (i)

Der χ^2 -Test

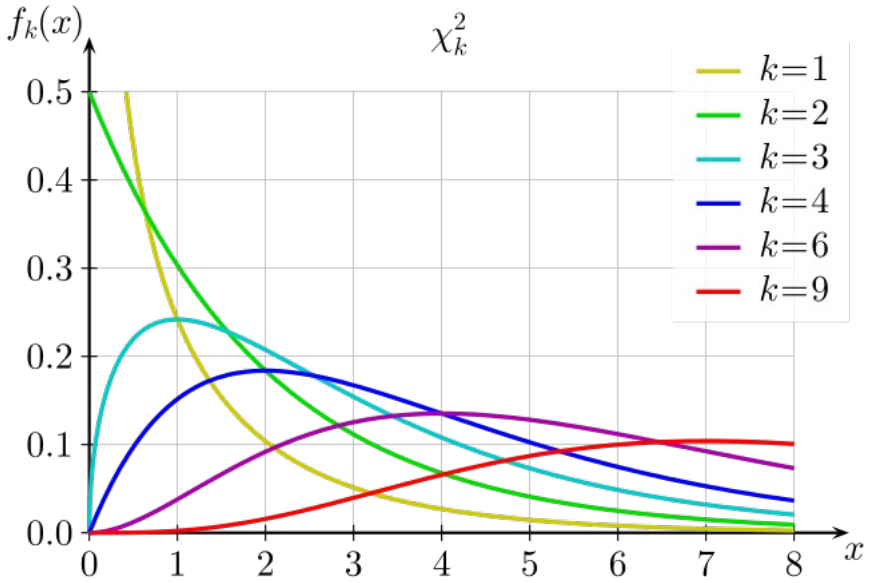
	Geschlecht		
	Frau	Mann	
Beobachtet	56	20	76
Erwartet	38	38	76

$$\chi^2_{df=k-1} = \sum_{i=1}^{k} \frac{\left(f_{b,i} - f_{e,i}\right)^2}{f_{e,i}}$$

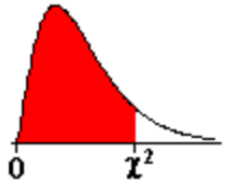
$$\chi^2_{df=1} = \frac{(56-38)^2}{38} + \frac{(20-38)^2}{38} = \frac{18^2}{38} + \frac{(-18)^2}{38} = 8.53 + 8.53 = 17.05$$

Der χ^2 -Test

- ***Schritt 3***: Vergleich des empirischen χ^2-Werts mit dem kritischen χ^2-Wert.
- Der kritische χ^2-Wert wird in Abhängigkeit von den Freiheitsgraden k-1 und dem gewählten *α-Niveau* aus einer Tabelle zur *χ^2-Verteilung* abgelesen



χ^2 -Tabelle



Lesebeispiel: Gesucht sei der χ^2-Wert, unter dem bei df=10 Freiheitsgraden 95% aller möglichen Werte einer χ^2-verteilten Zufallsvariablen X^2 liegen. In der Zeile für df=10 finden Sie in der Spalte $1-\alpha$ =0,95 den gesuchten Wert χ^2= 18,31.

df	(rote/dunkle) Fläche $1-\alpha$								
	0,7	**0,75**	**0,8**	**0,85**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84
4	4,88	5,39	5,99	6,74	7,78	9,49	11,14	13,28	14,86
5	6,06	6,63	7,29	8,12	9,24	11,07	12,83	15,09	16,75
6	7,23	7,84	8,56	9,45	10,64	12,59	14,45	16,81	18,55
7	8,38	9,04	9,80	10,75	12,02	14,07	16,01	18,48	20,28
8	9,52	10,22	11,03	12,03	13,36	15,51	17,53	20,09	21,95
9	10,66	11,39	12,24	13,29	14,68	16,92	19,02	21,67	23,59
10	11,78	12,55	13,44	14,53	15,99	18,31	20,48	23,21	25,19

Der χ^2 -Test

- ***Schritt 3***: Vergleich des empirischen χ^2-Werts mit dem kritischen χ^2-Wert.
- Für α=.05 ergibt sich bei *df=1*:

$$\chi^2_{emp} = 17.05$$

$$\chi^2_{krit} = 3.84$$

- Die $\boldsymbol{H_0}$ muss verworfen werden; folglich kann ein Unterschied nachgewiesen werden.

Der χ^2 -Test

χ^2 -Test – Beispiel 2

	Geschlecht		
Gehalt	Frau	Mann	
gering	23	14	37
hoch	35	6	41
	58	20	78

- Frage: Ist die relative Häufigkeit hoher bzw. geringer Gehälter bei Männern und Frauen gleich?
- Statistische Hypothesen
  - $\boldsymbol{H_0}$: $\pi(\text{Gehalt} \mid \text{Frau}) = \pi(\text{Gehalt} \mid \text{Mann})$
  - $\boldsymbol{H_1}$: $\pi(\text{Gehalt} \mid \text{Frau}) \neq \pi(\text{Gehalt} \mid \text{Mann})$

$\pi(x|y) =$ Relative Häufigkeit, dass Merkmalswert x auftritt, wenn Merkmalswert y auftritt

Der χ^2 -Test

Schritt 1: Zunächst werden aus den Randsummen die nach der H_0 zu erwarteten Häufigkeiten geschätzt:

Beobachtet:

	Geschlecht		
Gehalt	Frau	Mann	
gering	23	14	37
hoch	35	6	41
	58	20	78

$$f_{e(i,j)} = \frac{f_{b(i.)}}{N} \cdot \frac{f_{b(.j)}}{N} \cdot N = \frac{f_{b(i.)} \cdot f_{b(.j)}}{N}$$

Erwartet:

	Geschlecht		
Gehalt	Frau	Mann	
gering	27	10	37
hoch	31	10	41
	58	20	78

b(1 ,) = 37 b(, 1) = 58
N = 78
(37 · 58) / 78 = 27

Der χ^2 -Test

Schritt 2: Nun wird der (empirische) χ^2-Wert berechnet:

$$\chi^2_{df=(k-1)\cdot(l-1)} = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(f_{b(i,j)} - f_{e(i,j)}\right)^2}{f_{e(i,j)}}$$

mit:

- *k*, *l:* Anzahl der Stufen der beiden Variablen
- $f_{b(i,j)}$: Beobachtete Häufigkeit in der Zelle (i,j)
- $f_{e(i,j)}$: Erwartete Häufigkeit in der Zelle (i,j)

Der χ^2 -Test

Beobachtet:

	Geschlecht		
Gehalt	Frau	Mann	
gering	23	14	37
hoch	35	6	41
	58	20	78

Erwartet:

	Geschlecht		
Gehalt	Frau	Mann	
gering	27	10	37
hoch	31	10	41
	58	20	78

$$\chi^2_{df=(k-1)\cdot(l-1)} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(f_{b(i,j)} - f_{e(i,j)}\right)^2}{f_{e(i,j)}}$$

$$\chi^2_{df=1} = \frac{(23-27)^2}{27} + \frac{(35-31)^2}{31} + \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10}$$

$$= 0{,}59 + 0{,}51 + 1{,}60 + 1{,}60 = 4{,}30$$

Der χ^2 -Test

- ***Schritt 3***: Vergleich des empirischen χ^2-Werts mit dem kritischen χ^2-Wert.

df	(rote/dunkle) Fläche 1-α								
	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84

- Für α=.05 ergibt sich bei *df=1*:

$$\chi^2_{emp} = 4.30$$

$$\chi^2_{krit} = 3.84$$

- Die $\boldsymbol{H_0}$ muss verworfen werden; folglich kann ein Unterschied nachgewiesen werden.

Zusammenfassung

- *Nonparametrische Testverfahren* können verwendet werden, wenn
  a) die vorliegenden Daten kein Intervallskalenniveau aufweisen oder
  b) die Normalverteilungsannahme der parametrischen Tests verletzt ist.
- Der ***χ^2-Test*** überprüft, ob *beobachtete* und *erwartete* Häufigkeiten *signifikant* voneinander abweichen

Statistik als bedeutsame Wissenschaft

- Sehr viele statistische Tests wurden entwickelt
- Wir können hier nur die Spitze des Eisbergs beleuchten (Life-long Learning ist notwendig)
- Ohne statistische Kenntnisse ist man in der Informatik verloren
  - Häufige Fehler:
    - Nur Mittelwerte berechnen ohne auch Varianzen anzugeben
    - Verwendung von Schwellwerten anstelle von (modellbasierten) Hypothesentests