Statistische Testverfahren in der Schule

Christoph Richard, FAU Erlangen-Nürnberg
Tag der Mathematik 14. März 2026
Version 16. März 2026

Zusammenfassung

- Mit der Rückkehr zu G9 hat sich der Lehrplan für die Oberstufe geändert. Insbesondere kann im Vertiefungsbereich in der 12. Jahrgangsstufe Statistik unterrichtet werden.
- In diesem Kurs besprechen wir Themen des Vertiefungsbereichs. Dies umfasst Tests auf Erwartungswert kontinuierlicher Zufallsgrößen, Unabhängigkeitstests für diskrete Zufallsgrößen sowie lineare Regression.
- Konzepte des statistischen Testens wurden bereits in 2024 anhand des Binomialtests besprochen.
- Materialien zu beiden Kursen finden sich unter

  `https://www.math.fau.de/stochastik/christoph-richard/statistik/`

Fachlehrplan des ISB (Auszug)

- M12 Zufallsgrößen und Binomialverteilung (22 Std)

  Zufallsvariablen auf endlichen Ergebnisräumen, Verteilungsfunktion, Erwartungswert und Varianz, Urnenmodelle, Bernoulli-Kette, Binomialverteilung

- M12 einseitiger Signifikanztest (8 Std)

  Prinzip des Testens, einseitiger Binomialtest, Fehler 1.Art und Fehler 2.Art, Visualisierung

- M13 Normalverteilung (14 Std)

  kontinuierliche Zufallsgrößen, Dichtefunktion, Intervallwahrscheinlichkeiten, Normalverteilung, σ-Regeln

- M12 Vertiefung: Modul 5 Statistik

  beschreibende Statistik: lineare Regression, schließende Statistik: t-Test oder χ^2-Unabhängigkeitstest, p-Wert, Statistik-Software

Materialien zum Vertiefungsbereich

- Fundamente der Mathematik Bayern Jgst 12, Vertiefungskurs (Cornelsen)
- Lambacher-Schweizer, Mathematik für Gymnasien, Bayern Jgst 12, Vertiefungskurs (Klett Verlag), in Vorbereitung
- Lehrer-Fortbildung Universität Würzburg 2024/2025, Materialien abrufbar über ISB-Webseite

Zu dieser Schulung

Ansatz

- Anhand des Binomialtests wird ein Grundverständnis für statistische Tests erworben (Schulung 2024).
- Für andere Testprobleme können Tests aus Statistik-Programmen als "black box" verwendet werden, nachdem die Passgenauigkeit der zugrundeliegenden Modelle diskutiert worden ist.
- Dieser Zugang wird im Buch *Mathematik in den Life Sciences* von Gerhard Keller [K] auf universitärem Anfänger-Niveau verfolgt.

Gliederung

- Wir besprechen den exakten Test von Fisher auf Unabhängigkeit. [5]
- Dann besprechen wir Tests unter Normalverteilungsannahmen:
  - Gauß-Test [7], t-Test [2]
  - Tests auf Differenz von Mittelwerten [4]
  - χ^2-Anpassungstests, χ^2-Unabhängigkeitstest [3]
- Wir besprechen lineare Regression samt schließender Statistik. [14]

exakter Test von Fisher auf Unabhängigkeit [LW], [G, A11.11]

Hat Smartphone-Nutzung einen Einfluss auf schulische Leistungen?

- 87 Schüler:innen der 12. Jahrgangsstufe schreiben eine Mathematik-Klausur.
- Die Schüler:innen haben über einen Zeitraum von einem Monat vor der Klausur ihre Smartphone-Nutzungsdauer gemessen.
- Ihre Smartphone-Nutzungsdauer war entweder niedrig (N, weniger als 2h/Tag) oder hoch (H, mindestens 2h/Tag).
- Ihr Klausurergebnis war entweder besser (B) oder nicht besser (S) als das durchschnittliche Ergebnis.
- Vierfeldertafel

	B	S	
N	19	19	38
H	16	33	49
	35	52	87

exakter Test von Fisher auf Unabhängigkeit

- Fixiere die Randhäufigkeiten in der Vierfeldertafel.
- Wie wahrscheinlich ist obiges Ergebnis, wenn Nutzungsdauer und Klausurergebnis unabhängig sind?

Urnenmodell

- Betrachte den linken oberen Eintrag (N,B) in der Vierfeldertafel.
- Von 87 Schüler:innen haben 35 ein überdurchschnittliches Ergebnis.
- Ziehe von den 87 Schüler:innen 38 ohne Zurücklegen (nämlich die mit niedriger Nutzungsdauer).
- Anzahl Schüler:innen X mit niedriger Nutzungsdauer und überdurchschnittlichem Ergebnis (N,B) ist hypergeometrisch verteilt

$$\mathbb{P}(X = k) = \frac{\binom{35}{k}\binom{52}{38-k}}{\binom{87}{38}}$$

exakter Test von Fisher auf Unabhängigkeit

- Unabhängigkeitshypothese H_0 kann verworfen werden, wenn linker oberer Eintrag in Vierfeldertafel untypisch klein oder groß ist.
- Wähle $\alpha = 0.05$ als Schranke für Wahrscheinlichkeit eines Fehlers 1.Art (H_0 wird abgelehnt, obwohl H_0 wahr ist).
- Dann ist $[11, 19]$ kleinstes Intervall $[a, b]$ mit

$$\mathbb{P}(a \leq X \leq b) \geq 1 - \alpha \ .$$

- Also kann Unabhängigkeitshypothese zum Signifikanzniveau $\alpha = 0.05$ nicht abgelehnt werden.
- Zu diesem Test existiert eine Optimalitätstheorie [LR 4.6-4.7], welche derjenigen der Binomialtests ähnlich ist, siehe auch [W 6.6].

Konzeptionelles zu Tests auf Abhängigkeit

(I) Es wird auf Abhängigkeit getestet. Falls Unabhängigkeitshypothese H_0 akzeptiert werden muss, kann man aber nicht erwarten, dass die Merkmale unabhängig sind.

Eine Interpretation des Testergebnisses beruht auf dem "fränkischen Prinzip": Die Modellierung des Experiments mit unabhängigen Merkmalen widerspricht nicht den Beobachtungen, selbst wenn die Merkmale abhängig wären.

(II) Es wird auf Abhängigkeit getestet. Ablehnen von H_0 liefert aber keine Information darüber, wie stark die Merkmale voneinander abhängen.

(III) Für das umgekehrte Hypothesenpaar H_0 : *Abhängigkeit* gegen H_1 : *Unabhängigkeit* gibt es keinen brauchbaren statistischen Test, weil die Nullhypothese zu groß ist.

Konzeptionelles zu Tests auf Abhängigkeit

- Tests, welche die Probleme in (II) und (III) angehen, wurden für den zweiseitigen Binomialtest

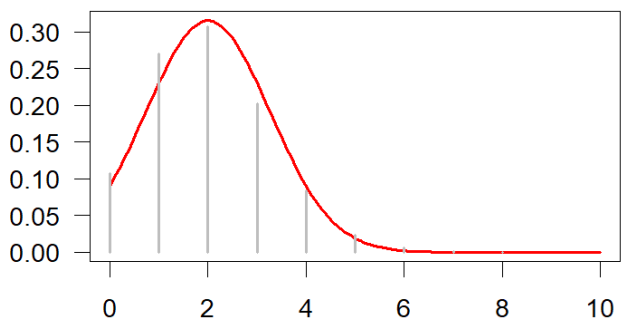
$$H_0 : p = p_0 \ , \qquad H_1 : p \neq p_0$$

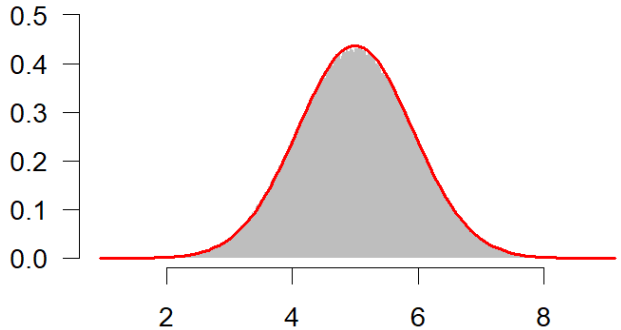
  in der Schulung aus 2024 entwickelt, samt Optimalitätsfragen.
- Analog geht man beim exakten Test von Fisher vorgehen, und auch beim t-Test [W].
- Die entsprechenden Inäquivalenztests oder Äquivalenztests werden in der Medizin, Pharmazie und Psychologie verwendet.

Welcome to Normality!

Beobachtung

- Die Verteilung einer Summe vieler Zufallsvariablen ist oft näherungsweise normal.
- Dies kann man überprüfen, wenn man viele Zufallszahlen hat.
- 10 Summanden aus $Ber(0.2)$ und $U(0,1)$ (Computersimulation)





- Dies erklärt die Bedeutung der Normalverteilungen.

mathematische Präzisierungen

- lokaler Grenzwertsatz für Zähldichte oder Dichte
- zentraler Grenzwertsatz für Verteilungsfunktion

Normalapproximation beim Testen

Tests auf eine Wahrscheinlichkeit

- Beim Binomial-Test beruht die Testentscheidung auf der Zahl der Erfolge bei einem mehrfachen unabhängigen Münzwurf.
- Beim Gauß-Test wird auf den Erwartungswert einer Normalverteilung getestet.
- Bei hoher Stichprobengröße kann man anstelle des Binomial-Tests näherungsweise den Gauß-Test verwenden ($np_0(1-p_0) \geq 10$).
- Der Gauß-Test ist das didaktische Pendant zum Binomialtest.

Unabhängigkeitstests [G A11.10]

- Bei hoher Stichprobengröße kann die hypergeometrische Verteilung durch eine Normalverteilung approximiert werden (alle Zellen ≥ 5).
- Dies führt auf den χ^2-Test auf Unabhängigkeit.

zufällige Experimente und statistische Modelle

- Ein Experiment liefere als Ergebnis reelle Zahlen. Bei unabhängiger Wiederholung entstehen in der Regel unterschiedliche Zahlen. (Temperaturmessung, Münzwurf, ...)
- Beschreibe das Experiment mit einer Zufallsvariable. Die Verteilung dieser Zufallsvariable ist unbekannt.
- Das wiederholte Experiment liefert eine Stichprobe reeller Zahlen. Mit dieser Stichprobe überprüft man Vermutungen über die dem Experiment zugrundeliegende Verteilung.
- Hierzu legt man ein statistisches Modell fest. Dies ist eine Menge von Verteilungen, von denen man annimmt, dass sie das Experiment gut beschreiben.
- Man vermutet, dass die dem Experiment zugrundeliegende Verteilung in einer bestimmten Teilmenge des Modells liegt.
- Mit der Stichprobe überprüft man, ob die Vermutung plausibel ist.

Tests und Fehlentscheidungen [H 29]

Normalverteilungen $N(\mu, \sigma^2)$, Erwartungswert μ unbekannt, Varianz σ^2 bekannt.

- statistisches Modell $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$, mit $\Theta = \mathbb{R}$ und $\mathbb{P}_\vartheta = N(\vartheta, \sigma^2)$.
- Partitioniere Θ in Hypothese H_0 und disjunkte Alternative H_1.
- Stichprobe liefert Entscheidung für H_0 oder für H_1.

Entscheidung	H_0 wahr	H_1 wahr
H_0 ablehnen	Fehler 1.Art	☺
H_0 akzeptieren	☺	Fehler 2.Art

Gütefunktion $G : \Theta \to [0, 1]$ mit $G(\vartheta) = \mathbb{P}_\vartheta(H_0 \text{ ablehnen})$.

- für $\vartheta \in H_0$ ist $G(\vartheta)$ Wkeit für Fehler 1.Art
- für $\vartheta \in H_1$ ist $1 - G(\vartheta)$ Wkeit für Fehler 2.Art

In vielen Beispielen approximiert G mit wachsender Stichprobengröße eine 01-Sprungfunktion (perfekter Test).

Niveau α-Tests [H 29]

Motivation

- Mit wenigen Daten kann man nur einen Fehler gut kontrollieren.
- Kontrolliere den Fehler 1.Art: Wähle $\alpha > 0$ mit

$$\mathbb{P}_\vartheta(H_0 \text{ ablehnen}) \leq \alpha \qquad \text{für alle } \vartheta \in H_0$$

- Dann liegt Wkeit für Fehler 2.Art fest. Sie kann sehr hoch sein, wenn $\vartheta \in H_1$ nahe bei H_0 liegt.

Testentscheidung:

- Falls man H_0 ablehnen kann, kontrolliert man den Fehler für eine falsche Entscheidung.
- Falls H_0 akzeptiert werden muss, gilt das Prinzip *in dubio pro reo*, obwohl eine falsche Entscheidung sehr wahrscheinlich sein könnte.

Wähle also für H_0 das *Gegenteil* der zu überprüfenden Hypothese!

Ist H_1 wahr, wenn ich H_0 ablehnen kann?

Niveau α-Tests

- Wie wahrscheinlich ist es, dass H_1 wahr ist, wenn wir H_0 abgelehnt haben?
- Hierauf kann der obige Niveau α-Test keine Antwort geben!
- Dies liegt daran, dass der obige Niveau α-Test kein Vorwissen darüber verwendet, wie plausibel bestimmte Werte in Θ sind.
- Dies ist ein unauflösbares Problem der statistischen Inferenz ohne Vorwissen.

Analogie zu medizinischen Massentests

- Wie wahrscheinlich ist es, dass ich Corona habe, wenn mein Test positiv ist?
- Man kann diese bedingte Wahrscheinlichkeit berechnen, wenn man weiss, mit welcher Wahrscheinlichkeit eine zufällig ausgewählte Person krank ist.

Konstruktion des Gauß-Tests

hier linksseitiges Testproblem

$$H_0 : \mu \geq \mu_0 \ , \qquad H_1 : \mu < \mu_0$$

- Testentscheidung mit arithmetischem Mittel der Stichprobe

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

- guter Schätzer für μ wegen

$$\mathbb{E}_\mu[\overline{X}_n] = \mu \ , \qquad \mathbb{V}_\mu[\overline{X}_n] = \sigma^2/n \ .$$

- akzeptiere H_0, falls $\overline{X}_n$ auf Stichprobe großen Wert annimmt

Forderungen:

- Ablehnungsbereich muss klein genug sein, so dass Fehler 1.Art mit Wkeit höchstens α auftritt.
- Akzeptanzbereich sollte möglichst klein sein, denn dann tritt Fehler 2.Art mit möglichst geringer Wkeit auf.

Konstruktion des linksseitigen Gauß-Tests

- verwende anstelle von $\overline{X}_n$ die standardisierte Teststatistik

$$T_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

- Entscheidungsregel:

$$\text{akzeptiere } H_0 \iff T_n \in [c, \infty)$$

für einen zu bestimmenden kritischen Wert c.

Verteilung von T_n

- für $\mu = \mu_0$ hat T_n Verteilung $N(0,1)$ mit Verteilungsfunktion Φ.
- für allgemeines μ hat T_n Verteilung $N(\lambda, 1)$ mit Nichtzentralitätsparameter $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$ wegen

$$\mathbb{P}_\mu(T \leq z) = \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z - \lambda\right) = \Phi(z - \lambda)$$

Hiermit lassen sich alle Fehlerwahrscheinlichkeiten berechnen.

Konstruktion des linksseitigen Gauß-Tests

- Fazit der bisherigen Überlegungen:

$$\mathbb{P}_\mu(H_0 \text{ ablehnen}) = \mathbb{P}_\mu(T_n < c) = \Phi(c - \lambda)$$

  mit $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$ und $\lambda \geq 0$ auf H_0

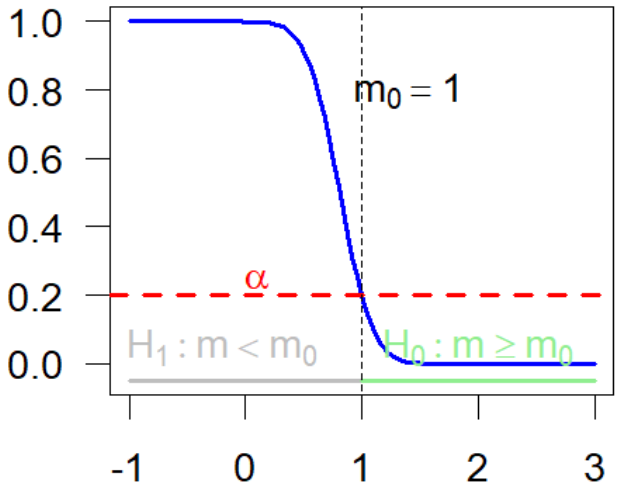
- Ablehnungsbereich muss klein genug sein, so dass Fehler 1.Art mit Wkeit höchstens α auftritt.
- Ablehnungsbereich sollte möglichst groß sein, denn dann tritt Fehler 2.Art mit möglichst geringer Wkeit auf.
- Also sollten wir den kritischen Wert c wählen als

$$\begin{aligned} c &= \max\{t \in \mathbb{R} : \sup_{\mu \in H_0} \Phi(t - \lambda) \leq \alpha\} \\ &= \max\{t \in \mathbb{R} : \Phi(t) \leq \alpha\} = \Phi^{-1}(\alpha) =: z_\alpha \end{aligned}$$

- Die Gütefunktion des linksseitigen Gauß-Tests ist also

$$G(\mu) = \mathbb{P}_\mu(H_0 \text{ ablehnen}) = \Phi(z_\alpha - \lambda)$$

Gütefunktion linksseitiger Gauß-Test



$G(m) = \mathbb{P}_m(H_0 \text{ ablehnen})$

- Auf H_0 Wahrscheinlichkeit für Fehler 1.Art
- Auf H_1 Gegenwahrscheinlichkeit zum Fehler 2.Art

rechtsseitiger Test $H_0 : m \leq m_0$ gegen $H_1 : m > m_0$

- kann analog behandelt werden, Akzeptanzbereich $A = (-\infty, z_{1-\alpha}]$
- Gütefunktion gleicht der vom linksseitigen Test, bis auf Spiegelung an der Vertikale $m = m_0$.

zweiseitiger Gauß-Test

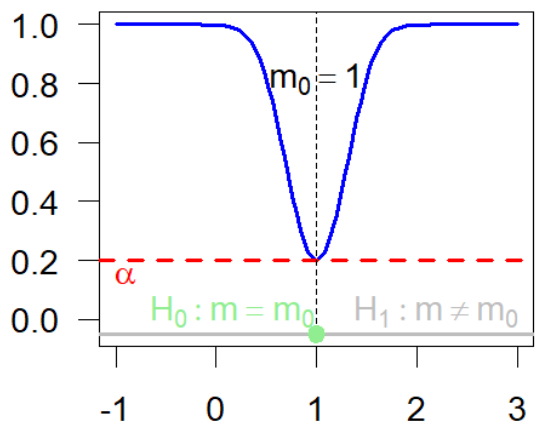
- Hypothesenpaar

$$H_0 : m = m_0 \ , \qquad H_1 : m \neq m_0$$

- Kombination zweier einseitiger Tests

$$H_0 : m \geq m_0 \text{ und } m \leq m_0 \ , \qquad H_1 : m < m_0 \text{ oder } m > m_0$$

- also Akzeptanzbereich $A = [z_{\alpha'}, z_{1-\alpha-\alpha'}]$
- kürzeste Länge bei symmetrischer Wahl $\alpha' = \alpha/2$
- Gütefunktion des zweiseitigen Gauß-Tests



Gauß-Test und p-Werte

Umformulierung der Testentscheidung:

- akzeptiere $H_0 \iff \alpha \leq p(t)$, mit t der Wert von T_n

Für die drei Arten des Gauß-Tests gilt

- linksseitiger Test:

$$t \in [z_\alpha, \infty) \iff \alpha \leq \Phi(t)$$

- rechtsseitiger Test:

$$t \in (-\infty, z_{1-\alpha}] \iff \alpha \leq 1 - \Phi(t)$$

- zweiseitiger Test:

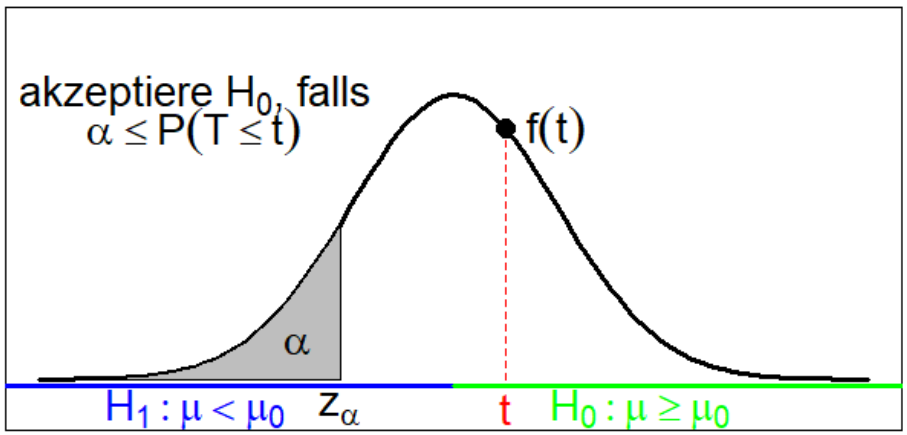
$$t \in [z_{\alpha/2}, z_{1-\alpha/2}] \iff \alpha \leq 2 \min\{\Phi(t), 1 - \Phi(t)\} = 2\Phi(-|t|)$$

Verteilung des p-Werts

- $\mathbb{P}_\mu(p \circ T_n < \alpha) = \mathbb{P}_\mu(H_0 \text{ ablehnen}) = G(\mu)$
- Also $U(0,1)$ bei $\mu = \mu_0$ wegen $G(\mu_0) = \alpha$!

graphische Interpretation des p-Werts

- linksseitiger Test: $A = [z_\alpha, \infty)$ und $p(t) = \mathbb{P}_{\mu_0}(T \leq t)$
- H_0 ablehnen $\iff p(t) < \alpha$, H_0 akzeptieren $\iff p(t) \geq \alpha$



Tests und Konfidenzintervalle

mit $t = \sqrt{n}(\overline{x} - \mu_0)/\sigma$ ergibt sich

- linksseitiger Test:

$$t \in [z_\alpha, \infty) \iff \mu_0 \leq \overline{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

- rechtsseitiger Test:

$$t \in (-\infty, z_{1-\alpha}] \iff \mu_0 \geq \overline{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

- zweiseitiger Test:

$$t \in [z_{\alpha/2}, z_{1-\alpha/2}] \iff \mu_0 \in \left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

H_0 akzeptieren $\iff$ μ_0 liegt im (1-α)-Konfidenzintervall für μ

rechtsseitiger t-Test

- statistisches Modell $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$
- hier $\Theta = \mathbb{R}\times\mathbb{R}_{>0}$ und $\mathbb{P}_\vartheta = N(\vartheta)$ mit $\vartheta = (\mu,\sigma^2)$.
- rechtsseitiger Test $H_0 : \mu \le \mu_0$ gegen $H_1 : \mu > \mu_0$.
- schätze unbekannte Varianz durch empirische Varianz

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

- Teststatistik

$$T_n = \frac{\overline{X}_n - \mu_0}{\sqrt{S_n^2/n}}$$

  Verteilung unter $\mu = \mu_0$ unabhängig von μ_0 und σ^2!
- Entscheidungsregel:

$$\text{akzeptiere } H_0 \iff T_n \in [c,\infty)$$

  für einen zu bestimmenden kritischen Wert c.

rechtsseitiger t-Test

Konstruktion des t-Tests genau wie beim Gauß-Test

- für $\mu = \mu_0$ ist die Verteilung T_n unabhängig von μ_0 und σ^2. Die Verteilung ist symmetrisch um 0 und verallgemeinert die Standardnormalverteilung. Sie wird mit t_{n-1} bezeichnet, die t-Verteilung mit $n-1$ Freiheitsgraden.
- für allgemeines μ hängt die Verteilung von T_n nur vom Nichtzentralitätsparameter $\lambda = \sqrt{n}(\mu - \mu_0)/\sigma$ ab. Sie wird mit $t_{n-1,\lambda}$ bezeichnet, die t-Verteilung mit $n-1$ Freiheitsgraden und Nichtzentralitätsparameter λ.
- Es ergeben sich dieselben Formeln wie beim Gauß-Test, mit den Ersetzungen $\sigma^2 || S_n^{(2)}$ und $\Phi || t_{n-1}$.

2-Stichproben-Tests: gepaarte Stichproben

Fragestellung:

- Stichproben $X = (X_1, \ldots, X_n)$ und $Y = (Y_1, \ldots, Y_n)$ aus i.a. verschiedenen Verteilungen.
- Untersuche, ob deren Mittelwerte sich unterscheiden.
- Falls X, Y in natürlicher Weise "gepaart" sind, ist eine genauere Analyse möglich als im ungepaarten Fall.

Abnutzung von Schuhen Modell A, B

- Jede Person trägt A am linken und B am rechten Fuß.
- Dann werden Unterschiede in den Modellen nicht von unterschiedlichem Gehverhalten der Personen überdeckt.

Modell

- Differenzen $X_1 - Y_1, \ldots, X_n - Y_n$ identisch normalverteilt
- verwende Gauß-Test oder t-Test von oben

2-Stichproben-Tests: ungepaarte Stichproben

Fragestellung:

- $X_1, \ldots, X_m$ iid $N(\mu_1, \sigma_1^2)$, $Y_1, \ldots, Y_n$ iid $N(\mu_2, \sigma_2^2)$, alle unabhängig
- Unterscheiden sich die Erwartungswerte μ_1, μ_2?
- Differenz der arithmetischen Mittel

$$\overline{X}_m - \overline{Y}_n \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

Fall 1: $\sigma_1 = \sigma_2 = \sigma$ bekannt

- Teststatistik

$$T_{mn} = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

- $N(0,1)$ verteilt für $\mu_1 = \mu_2$

2-Stichproben-Tests: ungepaarte Stichproben

Fall 2: $\sigma_1 = \sigma_2 = \sigma$ unbekannt

- schätze σ mit gepoolter empirischer Varianz

$$S_{XY}^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

- Teststatistik

$$T_{mn} = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{S_{XY}^2 \left(\frac{1}{m} + \frac{1}{n}\right)}}$$

- t_{m+n-2} verteilt für $\mu_1 = \mu_2$

2-Stichproben-Tests: ungepaarte Stichproben

Fall 3: σ_1, σ_2 unbekannt

- Behrens-Fisher-Problem: Was ist die Verteilung von

$$T_{mn} = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \qquad ?$$

- Welch 1938: für $\mu_1 = \mu_2$ näherungsweise t_ν mit

$$\nu = \frac{\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^2}{\frac{\sigma_1^4}{m^2(m-1)} + \frac{\sigma_2^4}{n^2(n-1)}}$$

- Dies kann man mit Zufallszahlen aus zwei verschiedenen Normalverteilungen numerisch überprüfen.

vom Binomialtest zum χ^2-Anpassungstest [K 10.2]

zweiseitiger Binomialtest $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$

- Akzeptiere H_0, falls sowohl *Zahl der Erfolge* S_n als auch *Zahl der Misserfolge* nahe an ihren unter H_0 erwarteten Werten sind.
- Wähle als Teststatistik also zum Beispiel

$$T_n = \frac{(S_n - np_0)^2}{np_0} + \frac{((n - S_n) - n(1 - p_0))^2}{n(1 - p_0)}$$

- Mit dieser Wahl gilt $T_n = Y_n^2$, mit Y_n Standardisierung von S_n

$$Y_n = \frac{S_n - np_0}{\sqrt{np_0(1 - p_0)}}$$

- T_n asymptotisch verteilt wie das Quadrat einer $N(0, 1)$-Zufallsgröße
- Letztere Verteilung heißt χ^2-Verteilung mit einem Freiheitsgrad.
- Lehne H_0 ab, falls T_n zu groß ist.
- Dies liefert einen asymptotisch exakten Test.

χ^2-Anpassungstest

- Gesucht ist die Verteilung einer Zufallsgröße mit s Werten, die mit positiven Wkeiten $\pi = (\pi_1, \ldots, \pi_s)$ auftreten. Es wird auf $H_0 : \pi = p$ getestet, mit $p = (p_1, \ldots, p_s)$.
- Für eine zufällige Stichprobe von n Zahlen aus obiger Verteilung seien $N_1, \ldots, N_s$ die Häufigkeiten für die Werte $1, \ldots, s$. Setze

$$T_n = \sum_{i=1}^{s} \frac{(N_i - np_i)^2}{np_i}$$

- T_n ist unter H_0 asymptotisch χ^2_{s-1}-verteilt, d.h. wie Summe von $s-1$ unabhängigen quadrierten $N(0,1)$-Zufallsvariablen. Diese Verteilung ist unabhängig von p!
- Lehne H_0 ab, falls T_n zu groß ist. Dies ergibt einen asymptotisch exakten Test, für den eine komplizierte asymptotische Optimalitätsaussage gilt [LR, Thm 14.3.2].
- Die exakte Verteilung von T_n kann simuliert werden.

χ^2-Anpassungstest, χ^2-Unabhängigkeitstest

- Der χ^2-Anpassungstest untersucht, ob eine Zufallsgröße mit endlich vielen Werten bestimmte vermutete Wkeiten besitzt [H 29.7-29.9]. Dies verallgemeinert den zweiseitigen Binomialtest [K 10.2].
- Der χ^2-Unabhängigkeitstest untersucht, ob zwei Zufallsgrößen mit endlich vielen Werten unabhängig sind. Dieser Test wird aus dem χ^2-Anpassungstest entwickelt.
- Die diskreten Teststatistiken beider Tests sind asymptotisch χ^2_r-verteilt mit passendem r, d.h. wie die kontinuierliche Summe von r unabhängigen quadrierten $N(0,1)$-Zufallsvariablen.
- Die Theorie zum χ^2-Anpassungstest ist komplex [G 11.2], [LR 14.3.1], die Theorie zum χ^2-Unabhängigkeitstest ist komplexer [G 11.3], [LR 14.3.3].
- Um den Fehler der Normalapproximation zu untersuchen, können die Teststatistiken auf dem Computer simuliert werden. Dies gilt auch für den Gauß-Test und den t-Test jenseits der Normalverteilung.

Regression als Optimierungsproblem [K,G]

- Sei $x = (x_1, \ldots, x_n)$ ein Datensatz reeller Zahlen.
- Welche Zahl beschreibt die Lage des Datensatzes am besten?

arithmetisches Mittel $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$?
Median $\widetilde{x}$?

- n ungerade: Zahl an mittlerer Position im geordneten Datensatz.
- n ungerade: jeder Wert zwischen den beiden Zahlen an mittlerer Position im geordneten Datensatz.

Theorem

$\overline{x}$ minimiert die Summe der Fehlerquadrate $m \mapsto \sum_{i=1}^{n} (x_i - m)^2$.
$\widetilde{x}$ minimiert die Summe der Fehlerbeträge $m \mapsto \sum_{i=1}^{n} |x_i - m|$.

- Summe der Fehlerquadrate sehr einfach zu minimieren (Parabel).
- Aber Ergebnis nicht robust gegenüber Ausreißern!

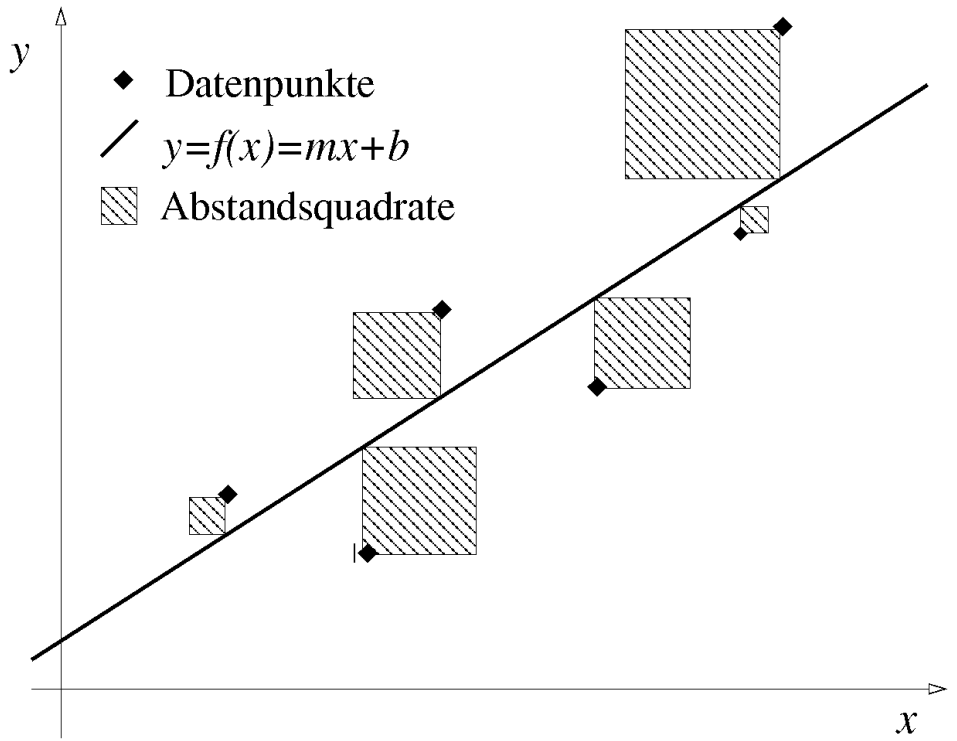
einfache lineare Regression

- Sei $(t_1, x_1), \ldots, (t_n, x_n)$ ein zweidimensionaler Datensatz.
- Welche Gerade $x = a + bt$ beschreibt diesen Datensatz am besten?
- Minimiere die Summe der Fehlerquadrate

$$(a, b) \mapsto \sum_{i=1}^{n} (x_i - (a + bt_i))^2$$

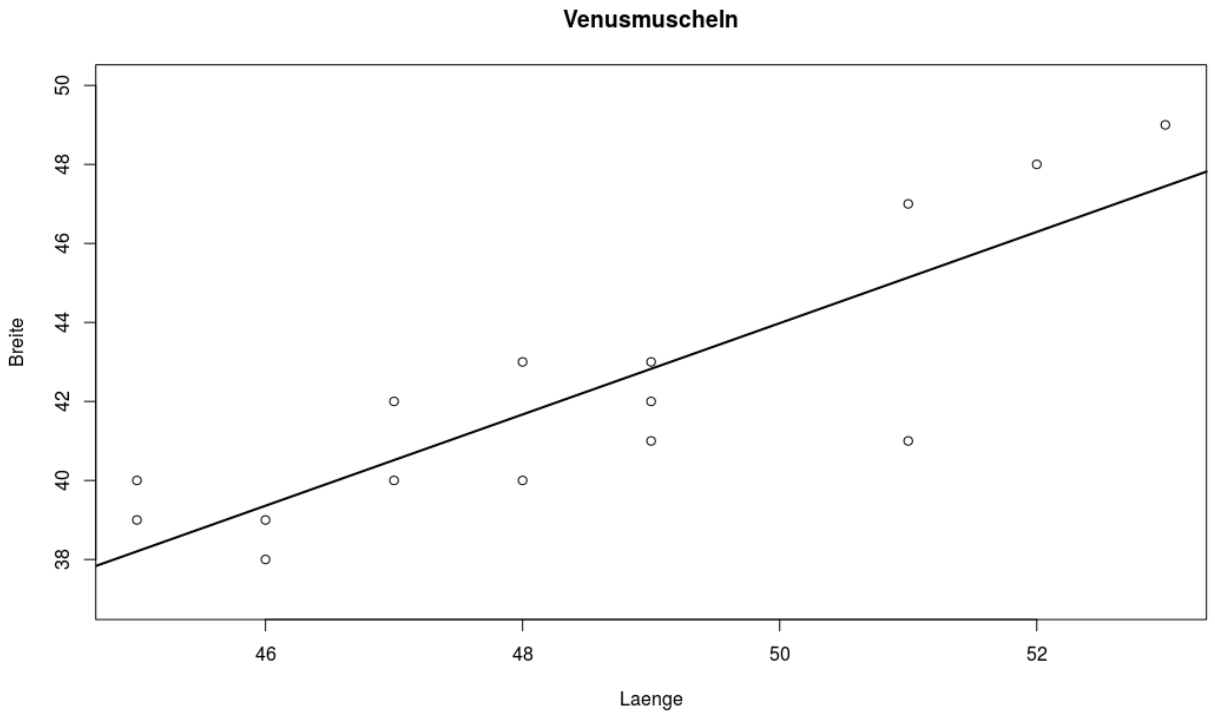
- Minimalstelle eines Paraboloids, kann durch Ableiten nach den Parametern bestimmt werden
- Ergebnis nicht robust gegenüber Ausreißern!
- Alternative: (numerische) robuste Regression mit der 1-Norm

Methode der kleinsten Quadrate



aus [K, S. 45]

Ausgleichsgerade für Venusmuscheln [K, Bsp. 4.4.1]



Wachstumsmodelle und einfache lineare Regression

Linearisierung komplexerer Zusammenhänge

- ungefährer linearer Zusammenhang für das Auge leicht erkennbar
- Linearisierung macht komplexen Zusammenhang leicht erkennbar
- lineare Regression liefert dann Schätzwerte für Parameter

exponentielles Wachstum $x = ae^{bt}$

- Linearisierung durch einfach logarithmische Darstellung

$$\log x = \log a + bt$$

allometrisches Wachstum $x = at^b$

- Linearisierung durch doppelt logarithmische Darstellung

$$\log x = \log a + b \log t$$

Wann ist (lineare) Regression sinnvoll?

Falls ein zweidimensionaler Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$ gut mit einer Gerade beschrieben werden kann, sind die eindimensionalen Datensätze $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ "korreliert".

- empirische Kovarianz und empirische Varianz

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \ , \qquad s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Weichen x_i, y_i von ihren Mittelwerten in dieselbe Richtung ab?

Korrelationskoeffizient

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

- $-1 \leq r_{xy} \leq 1$ mit $r_{xy}^2 = 1$ genau für linearen Zusammenhang
- Bei "schwacher Korrelation" liegt r_{xy} nahe bei 0.
- r_{xy}^2 kann auch für nicht linearen Zusammenhang nahe bei eins liegen!

Ein Test auf Korrelation [K, Kap. 12]

Modellannahme

- Stichprobe $(X_1, Y_1), \ldots, (X_n, Y_n)$ aus 2d Normalverteilung (X, Y)
- äquivalent zu $aX + bY$ normalverteilt für alle Wahlen von a, b
- problematisch, wenn x_i nicht zufällig (zB äquidistante Zeitpunkte)

F-Test für den Korrelationskoeffizienten

- Betrachte die Kovarianz S_{XY} von X und Y.

$$S_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- Nullhypothese H_0: $S_{XY} = 0$
- Für Normalverteilung äquivalent zu Unabhängigkeit von X und Y!
- Teststatistik ist eine bestimmte Funktion von $r_{xy} = s_{xy}/(s_x s_y)$.
- Falls man H_0 ablehnen kann, ist eine Regression sinnvoll.

Statistik zur einfachen linearen Regression

Gaußsches Produktmodell

- Modellannahme: verrauschter linearer Zusammenhang:

$$x_i = a + bt_i + \xi_i$$

  $\xi_1, \ldots, \xi_n$ unabhängige Zufallsvariablen, Verteilungen $N(0, \sigma^2)$
- Normalapproximation bei großer Stichprobe zulässig, da Formeln für a, b Summen enthalten.
- Man kann in diesem Modell die Ungenauigkeit der Schätzung von a und von b quantifizieren.
- Man kann auch testen, ob a, b signifikant von 0 verschieden sind.

Beachte:

- Man kann dieses Modell auf dem Computer implementieren und mit Zufallszahlen alle Aussagen aus der Theorie numerisch überprüfen.
- Man kann auch deutlich kompliziertere Modelle mit dem Computer simulieren.

Venusmuscheln: Kenngrößen mit R

```
> summary(lm(Breite~Laenge,data=Muscheln))

Call:
lm(formula = Breite ~ Laenge, data = Muscheln)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1370 -1.0936  0.1735  1.5183  1.8630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.7808     9.3971  -1.466    0.166
Laenge        1.1553     0.1939   5.958 4.76e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.815 on 13 degrees of freedom
Multiple R-squared:  0.7319,        Adjusted R-squared:  0.7113
F-statistic: 35.49 on 1 and 13 DF,  p-value: 4.765e-05
```

Statistik zu den Venusmuscheln I

- Parameterschätzwerte $\widehat{a}, \widehat{b}$ liefern Vorhersagen $\widehat{x}_i = \widehat{a} + \widehat{b}t_i$
- Man nennt die Vorhersagefehler $\varrho_i = x_i - \widehat{x}_i$ die Residuen.
- Es werden Kenngrößen der Residuen $\varrho_1, \ldots, \varrho_n$ ausgegeben.
- Bei passendem Modell: ϱ_i näherungsweise $N(0, \sigma^2)$-Zufallszahlen
- Man kann die Varianz σ^2 gut schätzen durch den Ausdruck

$$\widehat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^{n} (x_i - \widehat{x}_i)^2$$

- `Residual standard error` ist Wurzel aus diesem Ausdruck.
- Überprüfe Passgenauigkeit des Modells mit einem Residuenplot.

Statistik zu den Venusmuscheln II

Parameterschätzwerte $\widehat{a}, \widehat{b}$

- 2σ-Faustregel: Bei passendem Modell liegt in 95% aller Experimente der unbekannte Parameter innerhalb von 2 Standardabweichungen um seinen Schätzwert.
- Test auf verschwindenden Parameter mit p-Wert (und t-Wert Schätzwert/Standardabweichung)

Überprüfung der Korrelation

- `multiple R-squared` oder Bestimmtheitsmaß

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{x}_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

(Anteil der im linearen Modell erklärten Varianz des Datensatzes x)

- Man kann zeigen, dass $R^2 = r_{xy}^2$ gilt!
- F-Test auf Korrelationskoeffizienten mit p-Wert (und F-Wert)

mehrfache lineare Regression und das lineare Modell [G]

mehrfache lineare Regression:

- Zielgröße x hängt linear von d Einflussgrößen $t_1, \ldots, t_d$ ab:

$$x = \gamma_0 + t_1\gamma_1 + \ldots + t_d\gamma_d$$

- Die am besten passsenden Parameter $\gamma_0, \ldots, \gamma_d$ können mit der Methode der kleinsten Quadrate aus verrauschten Daten $(x_1, t_{1,1}, \ldots, t_{1,d}), \ldots, (x_n, t_{n,1}, \ldots, t_{n,d})$ bestimmt werden.

Theorem (lineares Modell)

Betrachte für $X \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times s}$ und $\gamma \in \mathbb{R}^s$ das Gleichungssystem $X = A\gamma$. Falls $n > s$ gilt und die Matrix A maximalen Rang hat, ist die eindeutige Lösung der Methode der kleinsten Quadrate gegeben durch

$$\widehat{\gamma} = (A^T A)^{-1} A^T X \ .$$

Das lineare Gauß-Modell [GWHT]

Statistik im linearen Gauß-Modell

- Modellannahme: verrauschter linearer Zusammenhang:

$$x_i = \gamma_0 + t_{i,1}\gamma_1 + \ldots + t_{i,s}\gamma_s + \xi_i$$

  $\xi_1, \ldots, \xi_n$ unabhängige Zufallsvariablen, Verteilung $N(0, \sigma^2)$
- Man kann die Ungenauigkeit der Schätzung der Parameter $\gamma_0, \ldots, \gamma_s$ quantifizieren.
- Man kann testen, ob Parameter $\gamma_0, \ldots, \gamma_s$ signifikant von Null verschieden sind.
- So kann man relevante Einflussgrößen auf eine Zielgröße erkennen.

Beachte:

- Schätzen, Testen und Modellwahl sind zB in R implementiert.
- Man kann das Modell auf dem Computer implementieren und mit Zufallszahlen alle Aussagen aus der Theorie numerisch überprüfen.

verallgemeinerte lineare Modelle [AK, EH]

Lineares Modell ist oft eine unpassende Beschreibung für ein Experiment.

- Zielgröße x_i folgt einer vorgegebenen Verteilung.
- Einfluss auf Zielgröße mit linearem Prediktor $\gamma_0 + t_{i,1}\gamma_1 + \ldots + t_{i,s}\gamma_s$
- Zusammenhang zwischen Zielgröße und Prediktor wird durch Link-Funktion g beschrieben: Mit $\mu_i = \mathbb{E}[x_i | t_{i,1}, \ldots, t_{i,n}]$ gilt

$$g(\mu_i) = \gamma_0 + t_{i,1}\gamma_1 + \ldots + t_{i,s}\gamma_s$$

zwei wichtige Modellklassen

- lineares Gauß-Modell: x_i normalverteilt, $g(x) = x$
- logistische Regression: x_i Bernoulli, $g(x) = \log(x/(1-x))$

Modelle und Vorhersage versus Erkenntnis

- Einfache Modelle können komplexe Zufallsexperimente perfekt beschreiben, im Rahmen der gegebenen experimentellen Daten.
- Damit erklären sie den zugrundeliegenden Zufallsmechanismus.
- Beispiel: Tore in einem Fußballspiel und das Münzwurf-Modell.
- Einfache Modelle werden oft auch dann verwendet, wenn sie nicht perfekt passen. Dies ist bei großem Datenmaterial der Fall.
- Hoch komplexe Modelle aus der KI liefern oft gute Vorhersagen zum Ausgang eines Zufallsexperiments.
- Allerdings liefern solche Modelle keine einfachen Erklärungen für den zugrundeliegenden Zufallsmechanismus.

ausgewählte Literatur zu Statistik und Data Science

K G. Keller, Mathematik in den Life Sciences, Ulmer TB (2011)
Modellbildung und Statistik mit R für Erstsemester, mit p-Werten

CCM C. Christensen, B. Christensen, M. Missong, Statistik klipp & klar, Springer (2019) *beispielorientierter Kompaktkurs für WiWis, mit p-Werten*

LW J. Lehn, H. Wegmann, Einführung in die Statistik, Teubner (1999)
moderates mathematisches Niveau, separates Buch mit Aufgaben, 5. Auflage

H N. Henze, Stochastik für Einsteiger, Springer Spektrum (2013)
moderates mathematisches Niveau, 10. Auflage

G H.O. Georgii, Stochastik, de Gruyter (2009)
Lehrbuch für Mathematik-Studierende, 5. Auflage

LR E. Lehmann, J. Romano, Testing Statistical Hypotheses, Springer (2022)

W S. Wellek, Testing Statistical Hypotheses of Equivalence and Noninferiority, CRC Press (2010)

JWHT G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer (2021) *Statistik mit dem Computer*

AK A. Agresti, M. Kateri, Foundations of Statistics for Data Scientists, CRC Press (2022) *Statistik mit dem Computer*

EH B. Efron, T. Hastie, Computer Age Statistical Inference, CUP (2016)
Streifzug durch die Statistik von den Anfängen bis zu Data Science