

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

QSB I: Epidemiologie, Medizinische Biometrie und Medizinische Informatik

Statistisches Testen I

Dr. rer. nat. Meinhard Mende
IMISE, Universität Leipzig

imise.

Lernziele Statistisches Testen I

Am Ende dieser Lehrveranstaltung

- kennen Sie die Logik des statistischen Testens und
- wenden das Ablaufschema für statistische Tests richtig an,
- treffen Sie anhand von Prüfgröße oder p-Wert Testentscheidungen,
- können Sie einen 1-Stichproben-t-Test sowie den t-Test für 2 unabhängige bzw. 2 gepaarte Stichproben rechnen und das Testergebnis interpretieren sowie
- entscheiden, in welcher Situation man den jeweiligen t-Test anwendet.

V 9

9.1 Grundprinzipien statistischen Testens

Beispiel 9.1 – Ex-DHF-Studie (→ B 1.1)

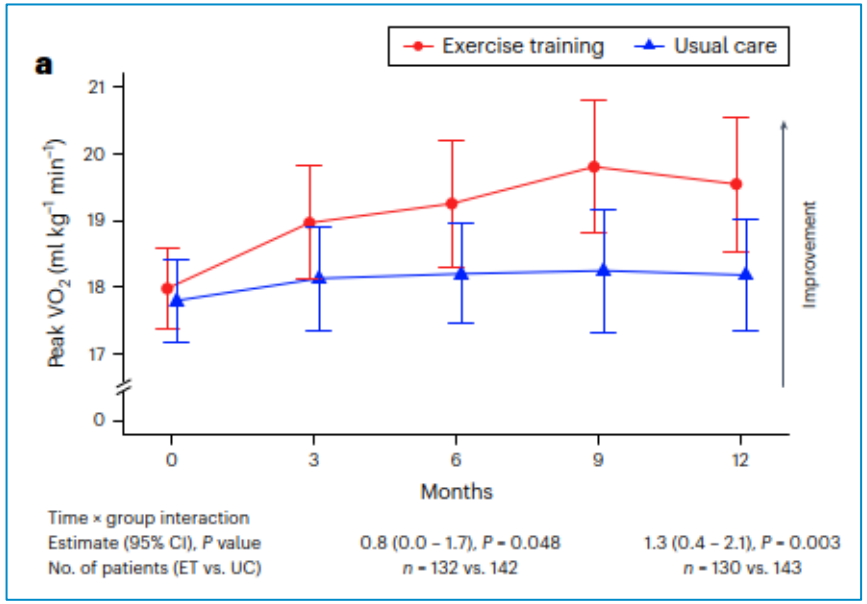
- Patients
  - Probanden mit HFpEF, NYHA-Klasse II / III
- Intervention
  - Training + Usual Care (ET)
- Comparison
  - Usual Care (UC)
- Outcome
  - Gesundheitszustand des Patienten



Edelmann F et al. Nature Medicine 2025;31:306–314

Klinische Fragestellung

Edelmann F et al. Nature Medicine 2025;31:306–314



- Wir beobachten eine Verbesserung der peak VO_2 in der Interventionsgruppe.

Frage: Ist der Unterschied in der Veränderung der peak VO_2 in 12 Monaten allein mit dem Zufall erklärbar?

Idee:

- Differenzbildung zum baseline-Wert
- Test der Differenzen auf Unterschied nach 12 Monaten

Grundidee der Hypothesentests

Methode der Falsifizierung

- Man postuliert das genaue Gegenteil dessen, was man beweisen will:
  Nullhypothese H_0,
- berechnet unter dieser Voraussetzung eine Prüfgröße und
- eine Wahrscheinlichkeit p.

Dann folgert man:

- Die Wahrscheinlichkeit p wurde unter der Annahme von H_0 gewonnen,
- Ist diese Wahrscheinlichkeit sehr klein, war wahrscheinlich die Anfangsannahme falsch.

Methode der Falsifizierung

Studienfrage:

Sind Mäuse auf dem Speicher?

Beobachtung (Daten):
Angeknabberte Kornsäcke

Nullhypothese H_0:
Es ist keine Maus
auf dem Speicher

Die Wahrscheinlichkeit angeknabberter Kornsäcke unter H_0 ist sehr klein

⇨ Folgerung:

Es muss doch eine Maus auf dem Speicher sein;
H_0 wird abgelehnt.

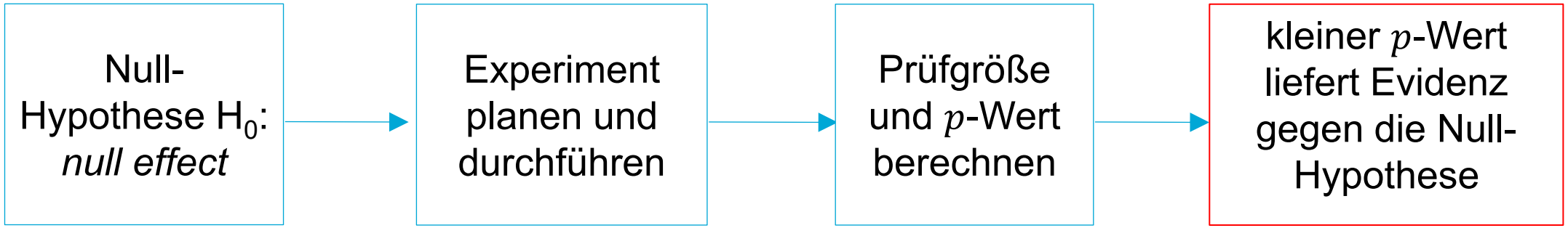


Bild von brgfx auf Freepik

Falsifizierung ist oft einfacher als Verifizierung (Popper).

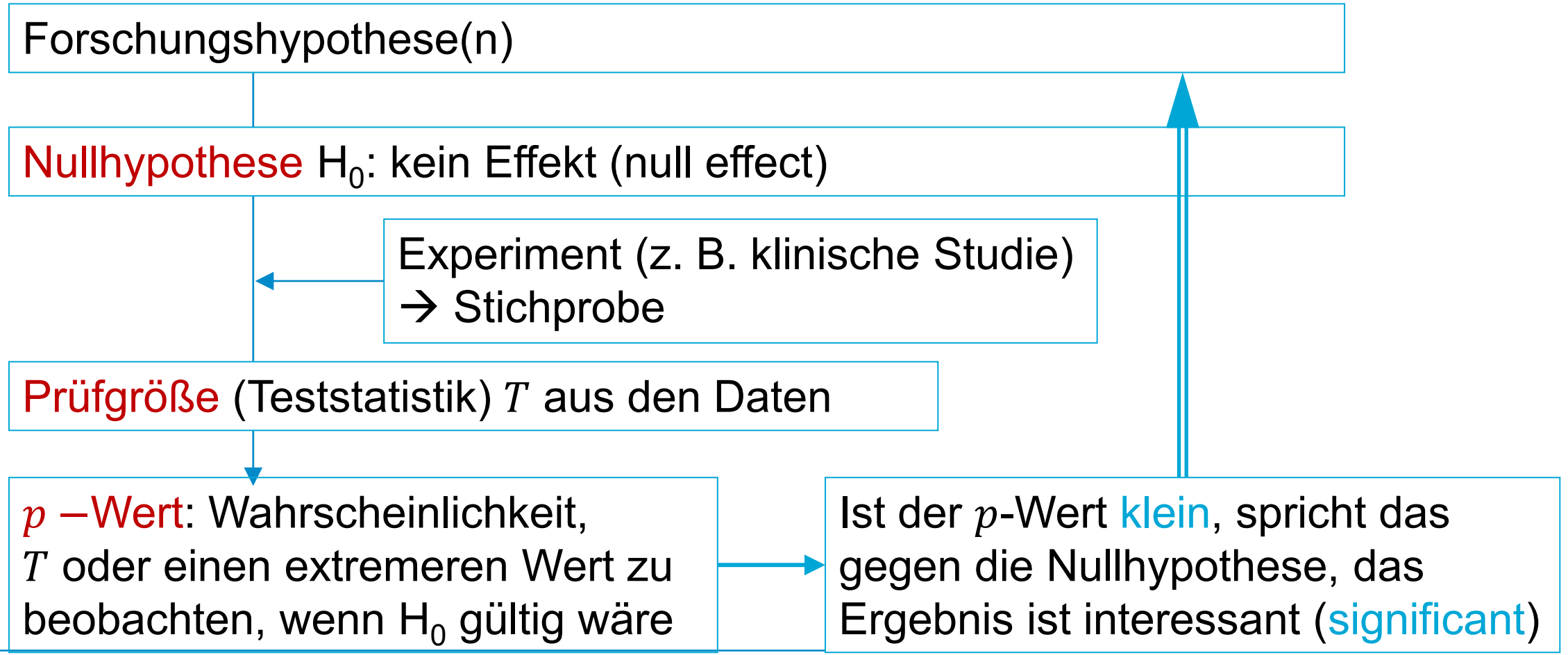
Historische Einordnung 1

- W. S. Gosset (1908), R. A. Fisher (1925): Signifikanztests



- kleiner p-Wert: Ergebnis ist interessant (significant),
- es lohnt sich, an dieser Stelle weiter zu forschen
- p value: evidence measure against the null hypothesis
- inductive inference (induktives Schließen)

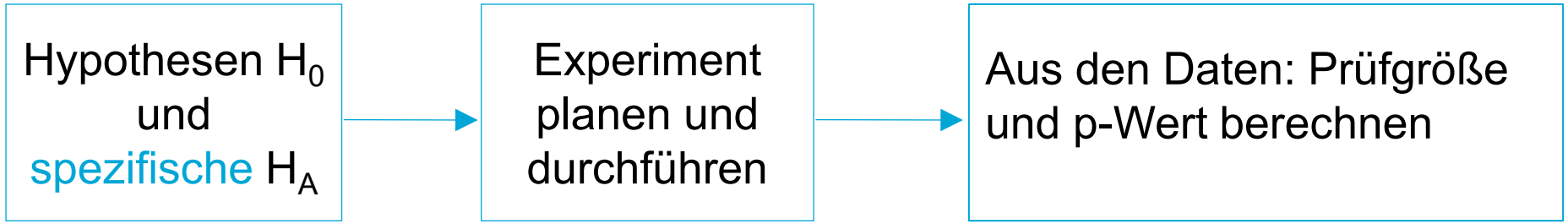
Ablauf eines Signifikanztests (Fisher)



Kritik des Signifikanztests

Das Konzept war nützlich z. B. für die Feldforschung, aber es enthielt ein paar Schwächen:

– Kritik von J. Neyman / E. Pearson (N-P, 1933)



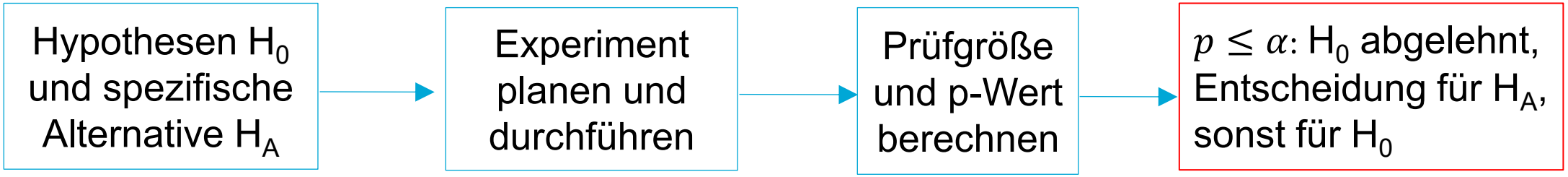
– N-P: zwei Schwächen des Signifikanztests

1. Ist der p-Wert klein, gibt es Evidenz *gegen* H_0, aber was ist, wenn p größer ist?
2. Es fehlt eine Abschätzung, wie oft Fehler passieren.

Historische Einordnung 2

J. Neyman / E. Pearson (1933): Hypothesentest

- Vorgabe zweier Hypothesen H_0 und H_A, die einander ausschließen
- Definition des Signifikanzniveaus α
  (Mit welcher Wahrscheinlichkeit will ich zulassen, dass H_0 abgelehnt wird, obwohl sie wahr ist?)



- Testentscheidung:

1. $p \leq \alpha$ führt zur Entscheidung für H_A,
2. sonst wird H_0 beibehalten

inductive behaviour (induktives Verhalten)

Historische Einordnung 3

J. Neyman / E. Pearson (1933)

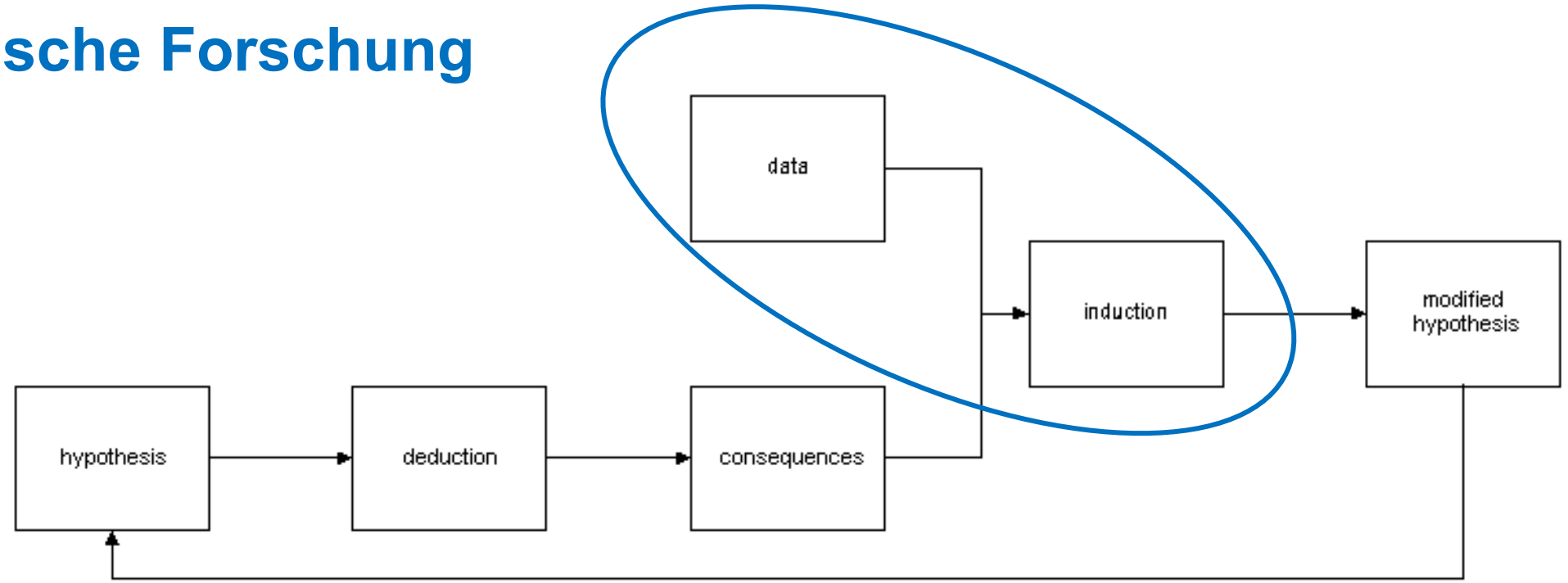


Gewinn:

- Es gibt immer eine Entscheidung für H_A oder H_0
- Man kann den Fehler I. / II. Art begrenzen bzw. abschätzen

Seit der Etablierung der Hypothesentests kritisieren Biometriker die Schwächen im Konzept und in der Anwendung statistischer Tests (F 19 ff).

Empirische Forschung



– Hypothesen haben so lange Bestand, bis es neue Beobachtungen gibt, welche eine Modifizierung erfordern bzw. diese gar widerlegen.

Grafik aus Chance BL, Journal of Statistics Education 2002;10(3).
jse.amstat.org/v10n3/chance.html

Historische Einordnung 4

Bayessches Paradigma:

Thomas Bayes (posthum 1763)

Vorwissen:

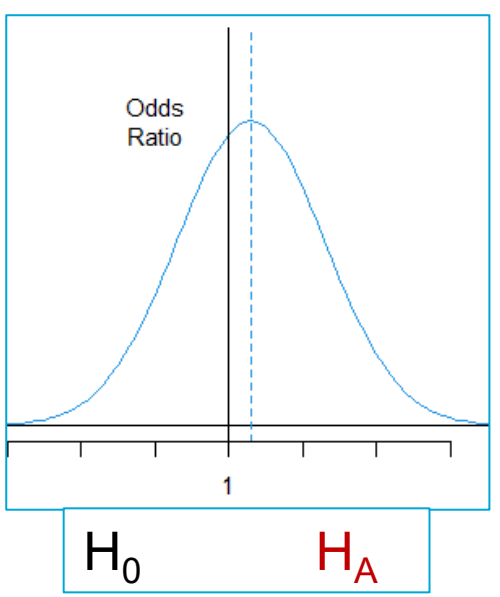
a-priori Wahrscheinlichkeitsverteilung für H_0, subjektive Wahrscheinlichkeit

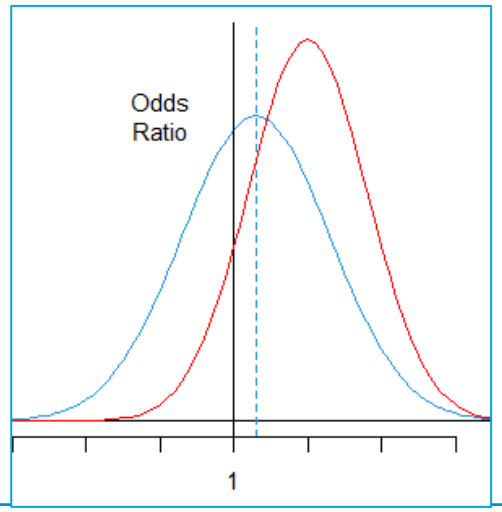
Experiment:

liefert Bayes-Faktor
(Evidenz gegen H_0 und für H_A)

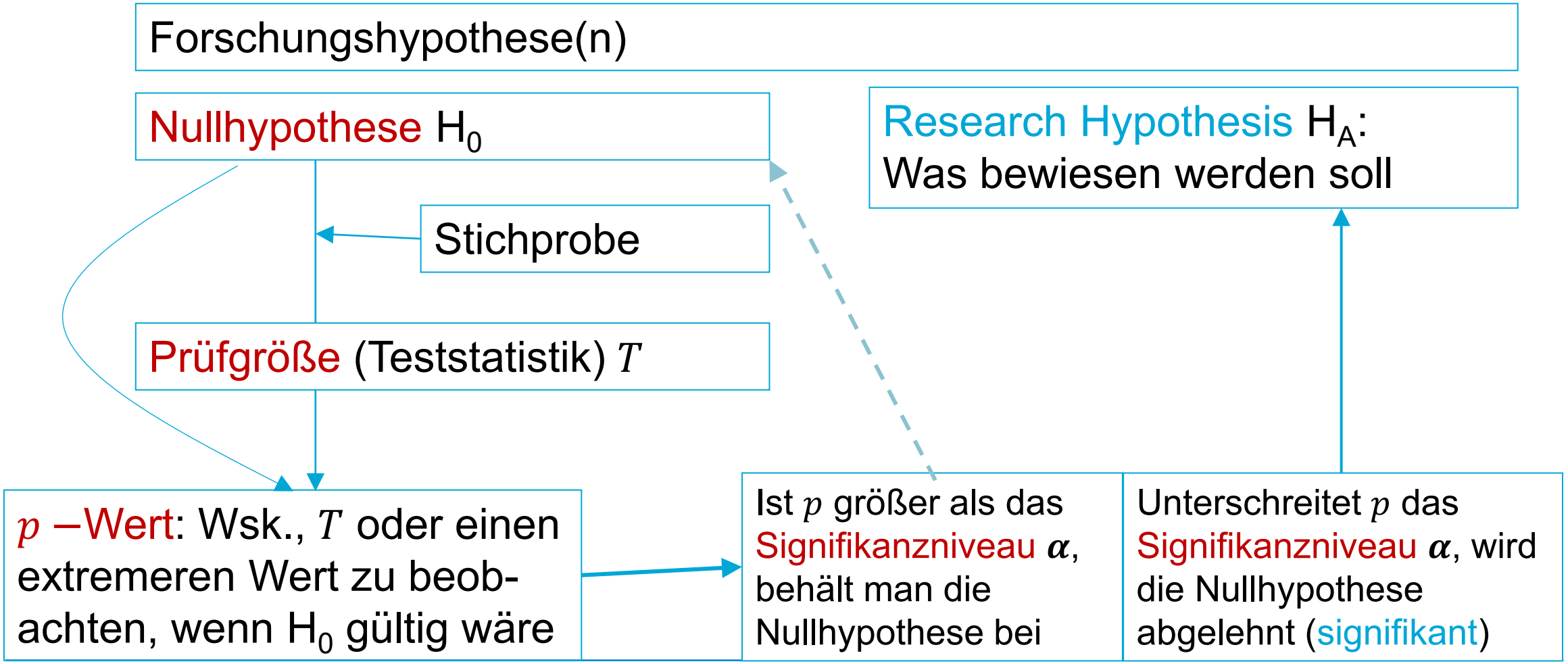
Verbessertes Wissen:

a-posteriori-Verteilung





Ablauf der Hypothesentests (N - P)



Zusammengefasst: Idee der Hypothesentests

- Die Nullhypothese (null effect) bestimmt (unter weiteren Annahmen) eine Verteilung für die Prüfgröße.
- Damit kann man die Wahrscheinlichkeit berechnen, dass die Prüfgröße (Teststatistik) T den beobachteten Wert t oder einen größeren erreicht – den p-Wert:

$$p = P(T \geq t | H_0)$$

- Ist p sehr klein, wäre es unter H_0 sehr unwahrscheinlich einen Wert $\geq t$ zu beobachten
- Entscheidung: H_0 ablehnen

Festlegung des Signifikanzniveaus α I

Motivation 1:

- Um zu entscheiden, ob man die Nullhypothese (d.h. Zufallseffekt als Erklärungsmodell) akzeptieren soll oder ob man seine Skepsis zugunsten eines systematischen Effektes aufgeben soll, muss eine Grenze festgelegt werden: Ablehnbereich (kritischer Bereich, rejection area)
- Je größer die Skepsis gegenüber einem Effekt (H_A) ist, desto kleiner muss ein p-Wert sein, um den Skeptiker zu überzeugen.
- Gedankenlose Konvention: Signifikanzniveau α = 5%.
- Signifikanzniveau ist eindeutiger als der synonym gebrauchte, unscharfe Begriff Irrtumswahrscheinlichkeit.

Festlegung des Signifikanzniveaus α II

Motivation 2:

- Das Ziel von N & P war es, die Häufigkeit des Fehlers I. Art (= falsch positive Befunde) nach oben zu begrenzen.
- Der Anwender legt das Signifikanzniveau α fest, die maximale Wahrscheinlichkeit für einen signifikanten Test, wenn kein Effekt vorliegt (also H_0 gilt).

		„Die Wahrheit“	
		H_0 wahr	H_0 falsch
Testentscheidung	H_0 beibehalten	☺	
	H_0 abgelehnt	Fehler I. Art: falsch positiv	

Power des Tests

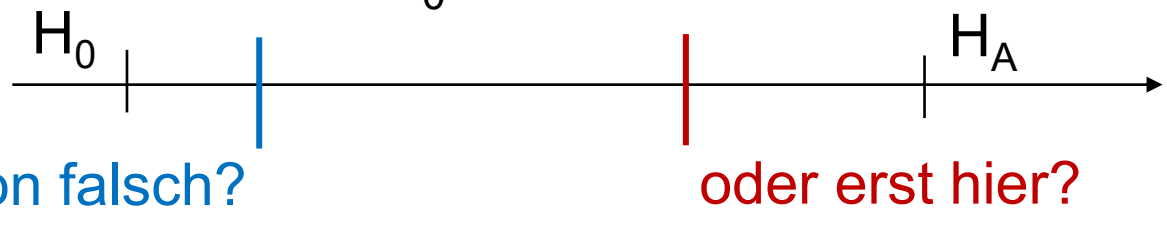
- Der Fehler II. Art geschieht, wenn ein vorhandener Effekt nicht durch eine Signifikanz bestätigt wird. Er ergibt sich aus
  - dem unbekannten Effekt (z. B. Behandlungseffekt),
  - der unbekannten Streuung und
  - der Fallzahl der Stichprobe.

		„Die Wahrheit“	
		H_0 wahr	H_0 falsch
Testent-scheidung	H_0 beibehalten	☺	Fehler II. Art: falsch negativ
	H_0 abgelehnt	Fehler I. Art	☺ Power des Tests

Konzeptionelle Schwächen I

Der Hypothesentest (N-P) ist eines der wichtigsten Werkzeuge der empirischen Wissenschaft – ungeachtet etlicher Schwächen:

1. H_0 ist in der Regel immer falsch! Kleine Differenzen sind der Normalfall. Aber ab welcher Differenz lehnt man H_0 ab?



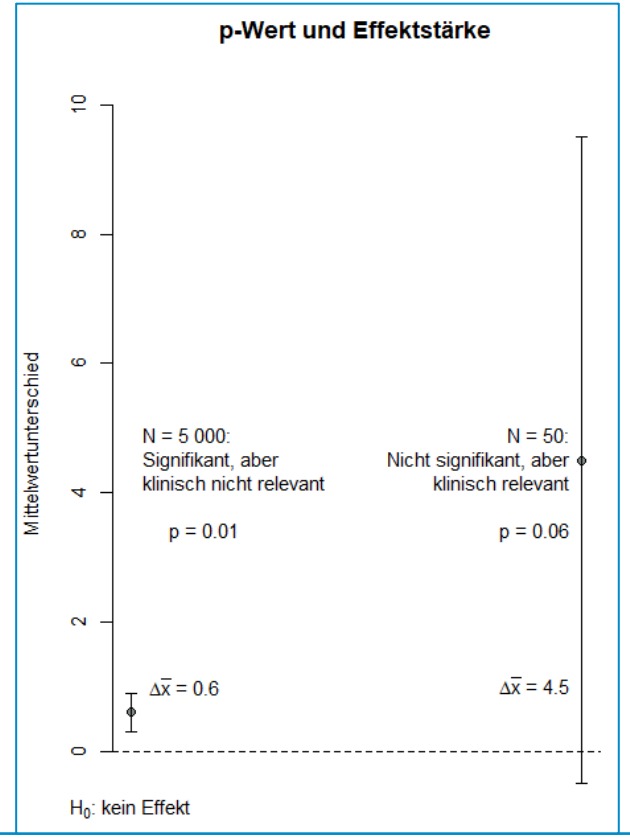
2. Die eigentlich interessierende Wahrscheinlichkeit, dass H_0 wahr ist, ist nicht definiert! Der p-Wert ist ein schwer zu interpretierendes Konzept.
3. (Sehr) kleine p-Werte verführen oft dazu, kausale Schlussfolgerungen zu ziehen.

Konzeptionelle Schwächen II

4. p-Werte bewegen sich im Kontinuum im Intervall [0, 1]. Das spiegelt die dichotome Testentscheidung nicht wider.

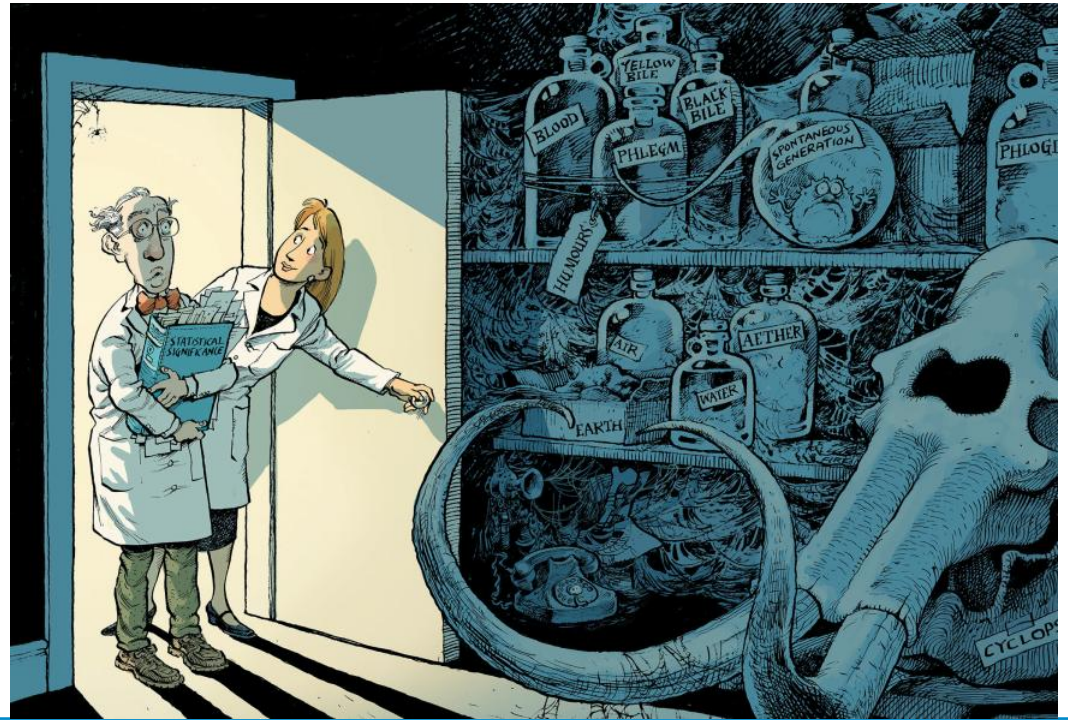


5. p-Werte sagen nichts über die Größe des Effekts aus. Um Gruppenunterschiede eindeutig zu charakterisieren, sollte ein p-Wert immer durch ein Effektmaß (inkl. 95%-C.I.) ergänzt werden.



Konzeptionelle Schwächen III

6. Ist $p > \alpha$, gibt das keine Evidenz, weder für H_0 noch für H_A.
7. Der Fehler I. Art α wird nur im long run eingehalten, im konkreten Testfall ist *die Wahrheit* unklar.
8. Das Konzept des Hypothesentests beinhaltet Inkonsistenzen.



Retire statistical significance

Amrhein V et al. Nature 2019;567:305-7

Fehlerhafte Schlussweise!



Jotquadrat: Grafitto in Bonn. commons. wikipedia.org

Typische Fehl-Schlussweise in älteren Publikationen:

- Szenario: RCT, Prüfung zweier Therapien
- Der statistische Test ergab, dass die Nullhypothese (gleiche Wirksamkeit) nicht verworfen werden konnte.
- Behauptung: ~~*Also sind die beiden Therapien äquivalent*~~.

⇨ Diese Schlussfolgerung ist ***falsch!***

Richtige Formulierung:

- *Aufgrund der in der Stichprobe vorliegenden Information lässt sich kein ausreichend großer Unterschied nachweisen, der erlauben würde, die Null-Hypothese abzulehnen.*

Hypothesentests: Fragen zur Wiederholung

Kreuzen Sie an bzw. beantworten Sie die Fragen!

1. Wählt man als Signifikanzniveau $\alpha = 1\%$ statt $\alpha = 5\%$, wird man im Mittel
   ○ *häufiger* ○ *seltener* H_0 ablehnen, wenn kein Effekt vorliegt.

2. Überlegen Sie, warum beim Hypothesentest Formulierungen wie
   + *hoch signifikant* oder gar
   + *borderline signifikant* unsinnig sind!

3. Setzen Sie den Satz fort: *Die Power ist die Wahrscheinlichkeit,*
   _____, *wenn*
   _____.

4. Ist $p > \alpha$ ein Beweis für die Richtigkeit der Nullhypothese?

Hypothesentests: Antworten

Hypothesentests: Antworten

V 9

9.2 Testalgorithmus am konkreten Beispiel

Statistisches Testen – Beispiel 9.1

Forschungshypothese:

- Der Zuwachs der peak VO_2 in 12 Monaten unterscheidet sich zwischen Probanden der Trainingsgruppe (ET) und der Kontrollgruppe (UC).

Nullhypothese H_0:

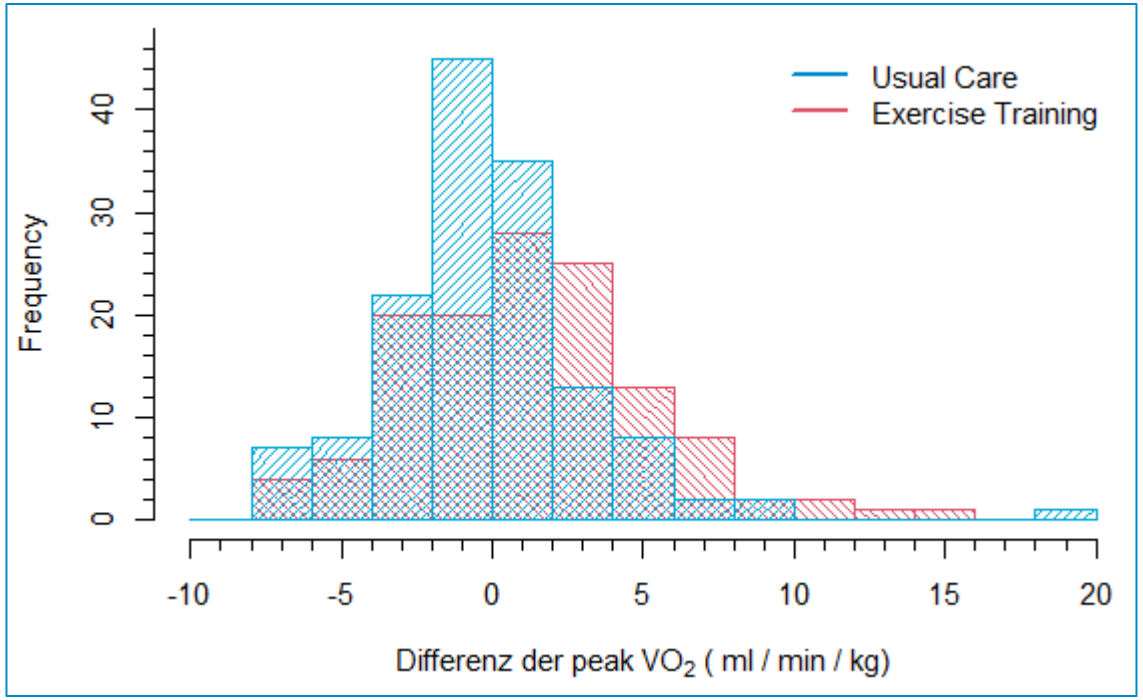
- mittlerer Zuwachs der peak VO_2 in der Trainingsgruppe = mittlerer Zuwachs der peak VO_2 in der Kontrollgruppe

$$H_0: \ \Delta\mu_{ET} = \ \Delta\mu_{UC}$$

- d. h. Hypothese der Gleichheit zweier Mittelwerte

Beobachtung

- Ein Anteil von über 25 Probanden des ET-Arms verbessert sich mehr als Probanden des UC-Arms.



- Ist das zufällig oder gibt es Evidenz für einen Interventionseffekt?

Evidenz:

- Gesamtheit des Wissensstands bezüglich einer spezifischen Frage (IQWiG-Glossar)
- im engeren Sinne: Überzeugtsein durch einen statistisch signifikanten Test.

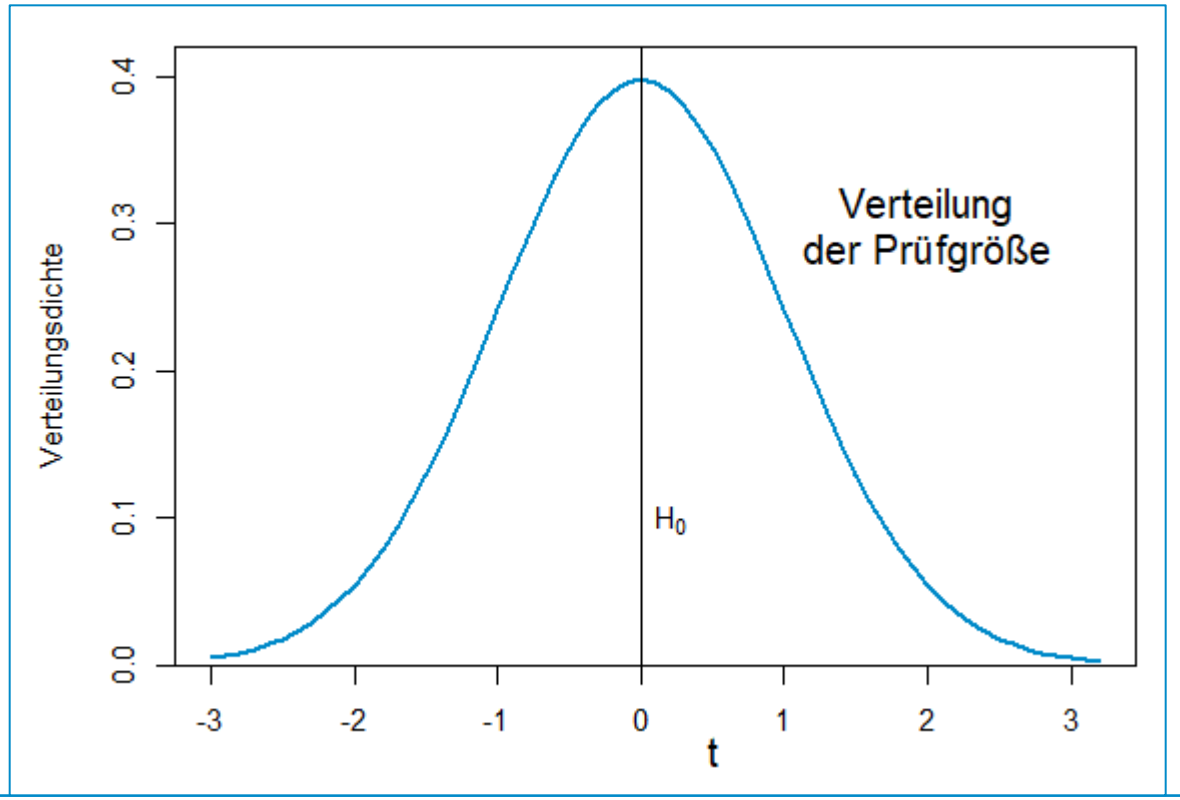
Prüfgröße

- Quantifizierung der Mittelwertdifferenz durch Standardisierung

- Prüfgröße: $t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$

noch zu klären:

- Wie berechnet man $SE_{\bar{x}_1 - \bar{x}_2}$
- Wie ist die Verteilung von t?



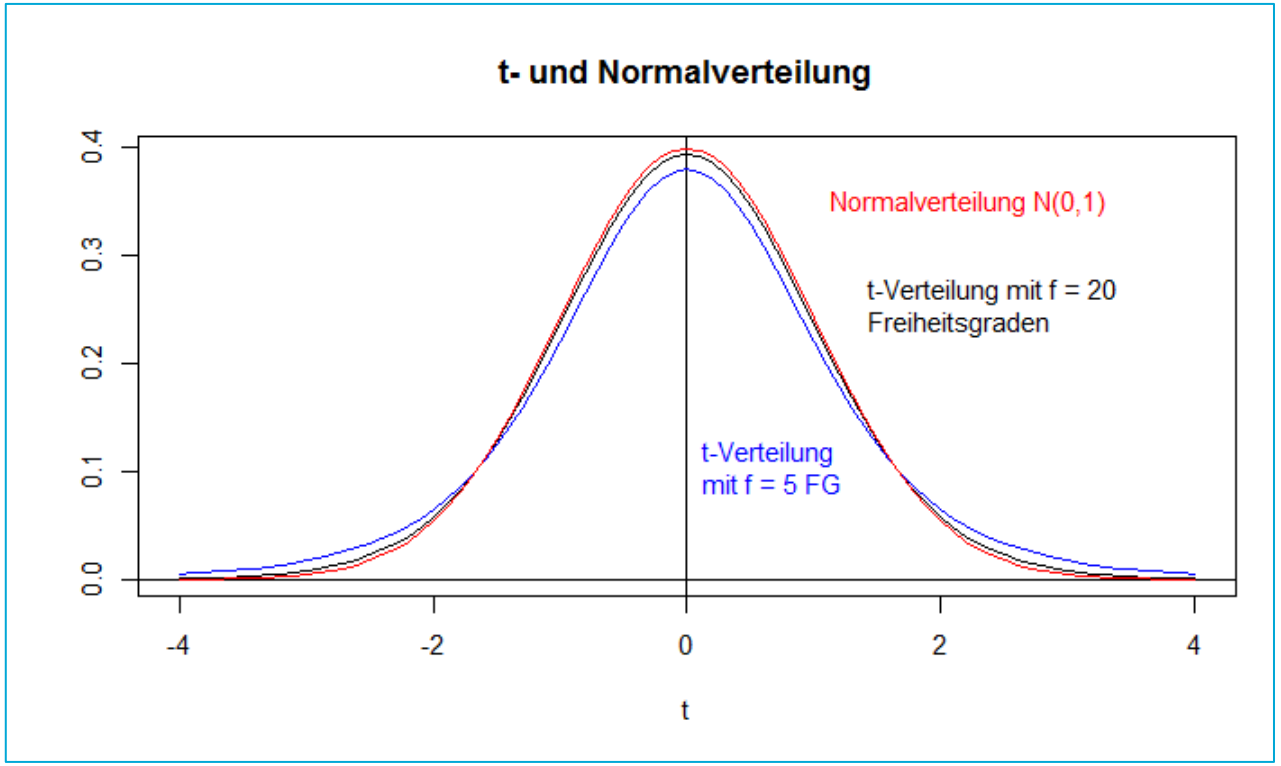
Prüfverteilung

- Die Prüfgröße ist so konstruiert, dass sie bei hypothetischer Annahme der Nullhypothese (H_0) ermöglicht,
- die Prüfverteilung zu berechnen.
- Diese zeigt alle möglichen Mittelwertdifferenzen zweier Stichproben, wenn diese aus Populationen mit identischem Mittelwert μ stammen würden: t-Verteilung
- Vgl. Skriptenheft 5.7.4, S. 64!
- R. A. Fisher (1925) Formel für t-Statistik

t-Verteilung mit f FG

Die T-Statistik ist unter H_0
t-verteilt mit f Freiheitsgraden
(vgl. V 6).

W. Gosset (1908), genannt student



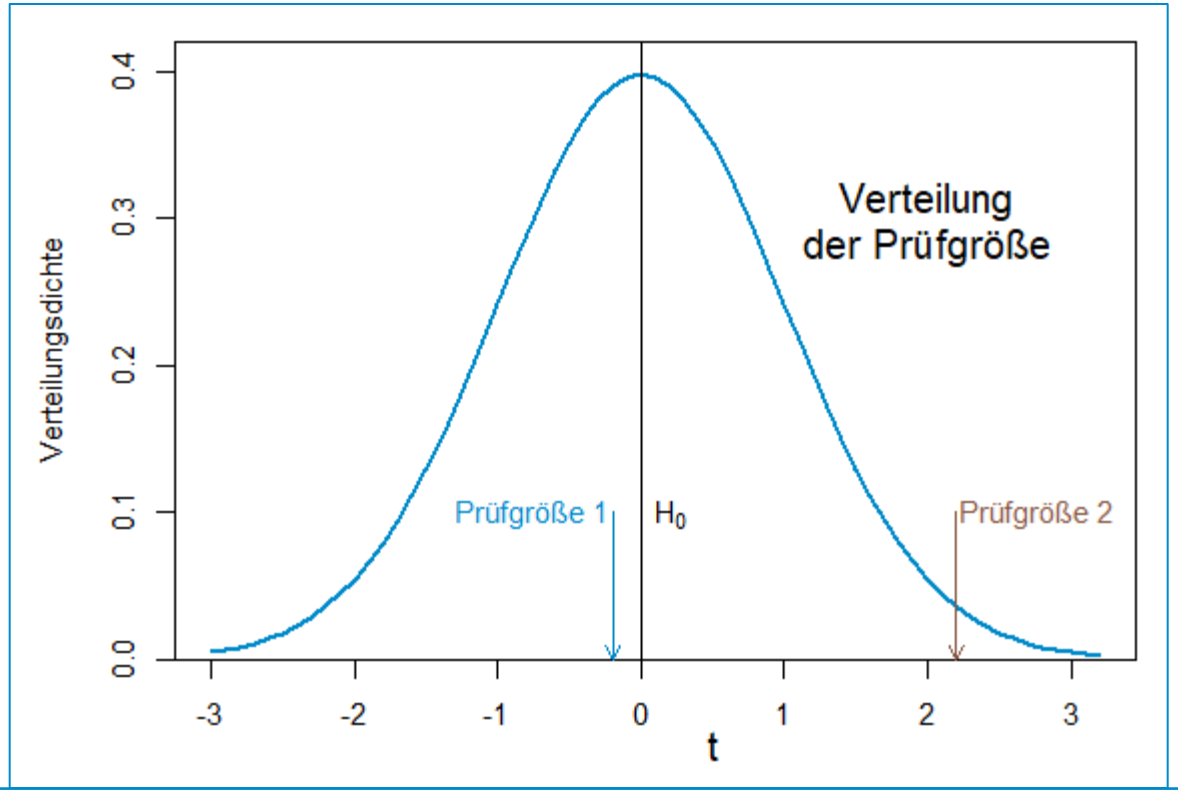
Die t-Verteilungsdichte

- ähnelt sehr der NV-Dichte,
- ist symmetrisch um 0,
- mit wachsender Zahl der Freiheitsgrade → sinkt die Streuung und
- nähert sich die Verteilung der Normalverteilung,
  - für $n = 20$ Freiheitsgrade kaum noch von der NV zu unterscheiden,
  - für $n = 100$ Freiheitsgrade fast identisch (→ ZGWS)

Testidee I

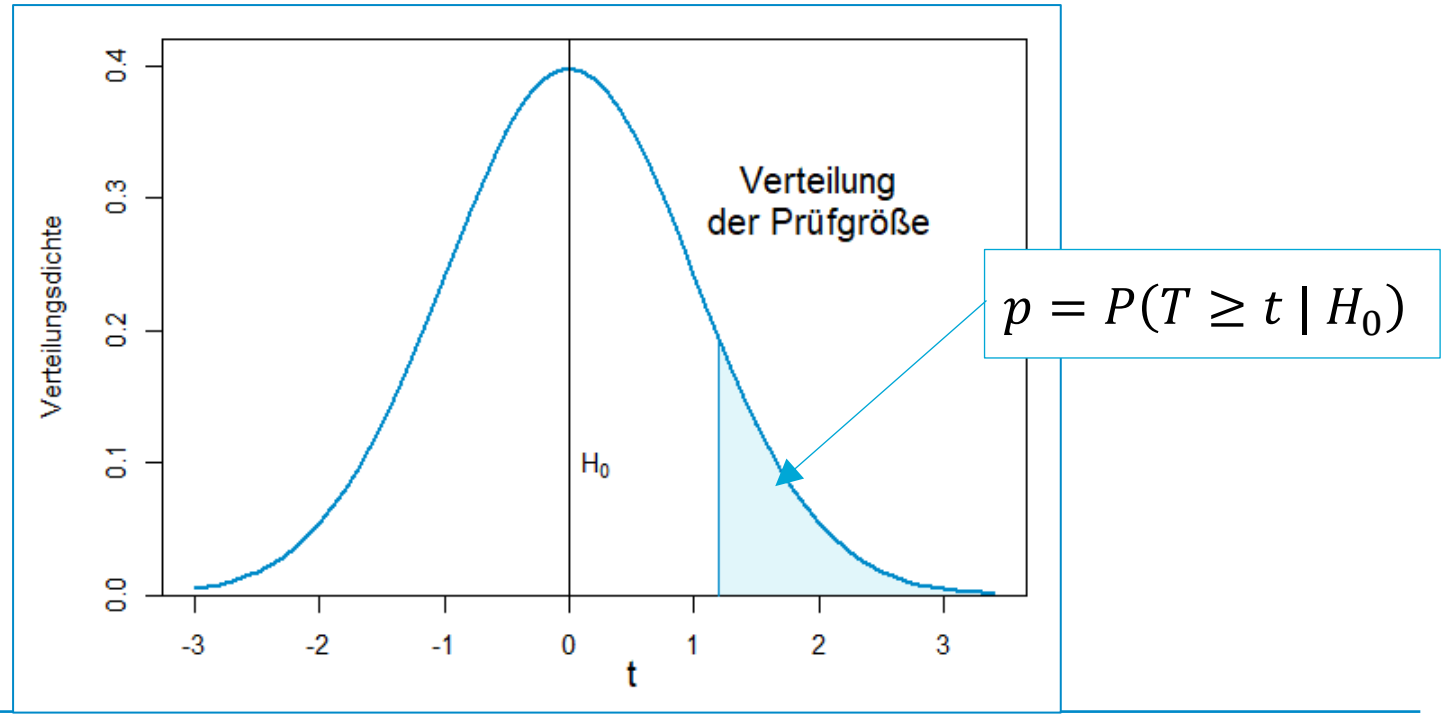
Je nach Stichprobe fällt der Wert der Prüfgröße mehr oder weniger ins Zentrum der Prüfverteilung:

- Stichprobe 1: Prüfgröße 1 gut mit H_0 vereinbar.
- Stichprobe 2: Prüfgröße 2 kaum noch mit H_0 vereinbar.
- Ab wann lehnen wir H_0 ab?



R. A. Fisher‘s Signifikanztest

- Wie wahrscheinlich ist es, unter der vorgegebenen Nullhypothese H_0 Stichproben zu erhalten, die mindestens so extrem sind wie diejenige, die man beobachtet hat?
- Sir R. A. Fishers Antwort: der p-Wert
- Fisher wertete kleine p-Werte als evidence against the null (significant = interessant)



p-Wert

- Der p-Wert (engl.: p-value) ist die Wahrscheinlichkeit
  - bei einer Wiederholung des Experimentes unter identischen Bedingungen (d. h. frequentistisches Paradigma)
  - unter Zugrundelegung der Nullhypothese (und der aus ihr resultierenden Wahrscheinlichkeitsverteilung)
  - und weiterer Modellannahmen
- das beobachtete oder ein noch extremeres Ergebnis zu erzielen.

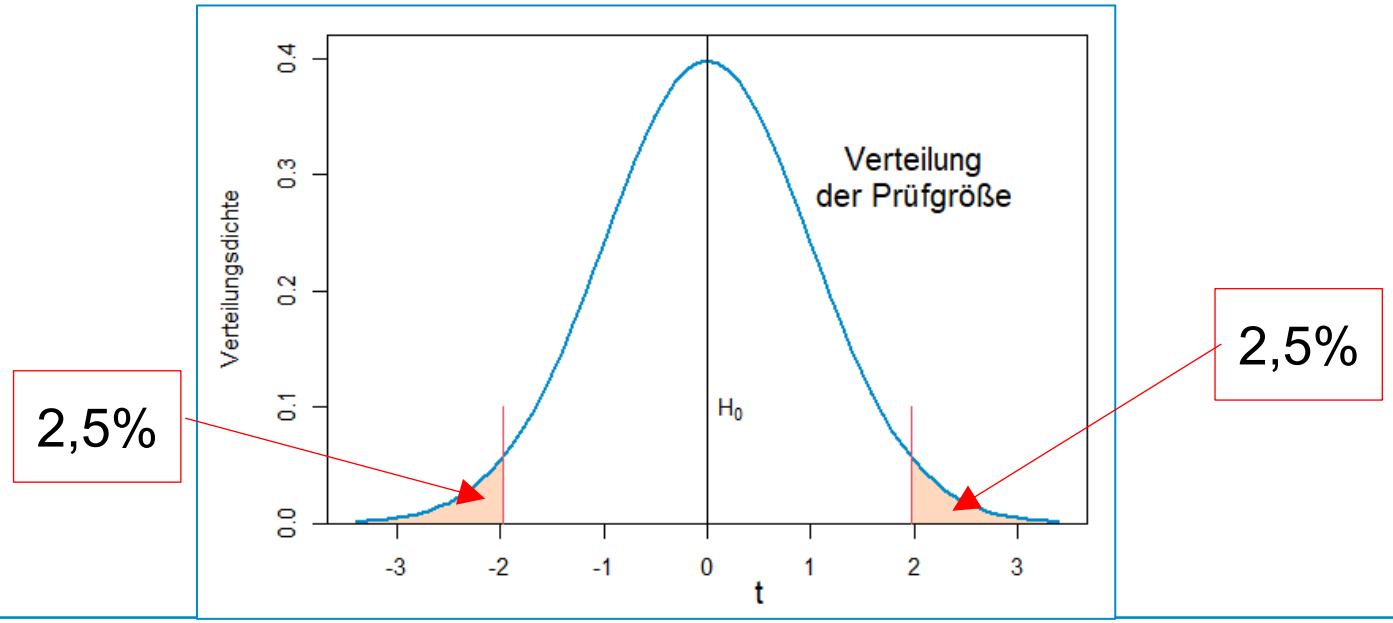


Gonick & Smith: Statistik in Cartoons

Neyman & Pearson

Ablehnbereich (rejection area, kritischer Bereich):

- links und rechts von H_0 je $\frac{\alpha}{2}$ Wahrscheinlichkeit
- bei symmetrischen Verteilungen: außerhalb des $\frac{\alpha}{2}$- und $(1-\frac{\alpha}{2})$-Quantils der Prüfverteilung.

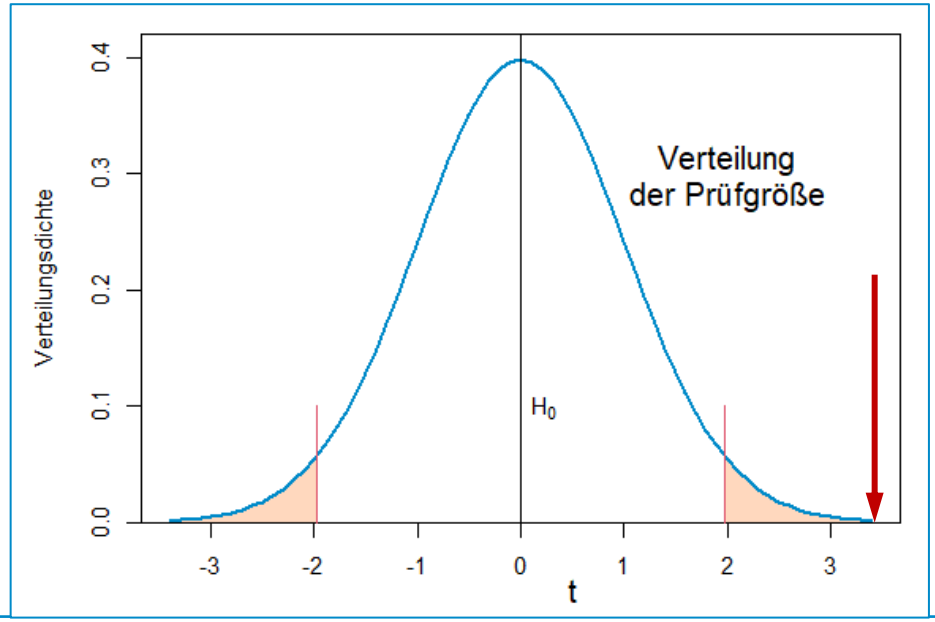


Testentscheidung Hypothesentest I

a) Handrechnung: Liegt der Wert t der Prüfgröße im Ablehnbereich:
$|t| \geq t_{f;\,1-\frac{\alpha}{2}}$, wird H_0 abgelehnt.

Ablauf:

- $t_{1-\frac{\alpha}{2}}$-Quantil aus der Tafel ablesen (Skriptenheft)
- Ist $|t| \geq t_{f;\,1-\frac{\alpha}{2}}$, dann wird H_0 abgelehnt,
- ansonsten beibehalten.



Testentscheidung Hypothesentest II

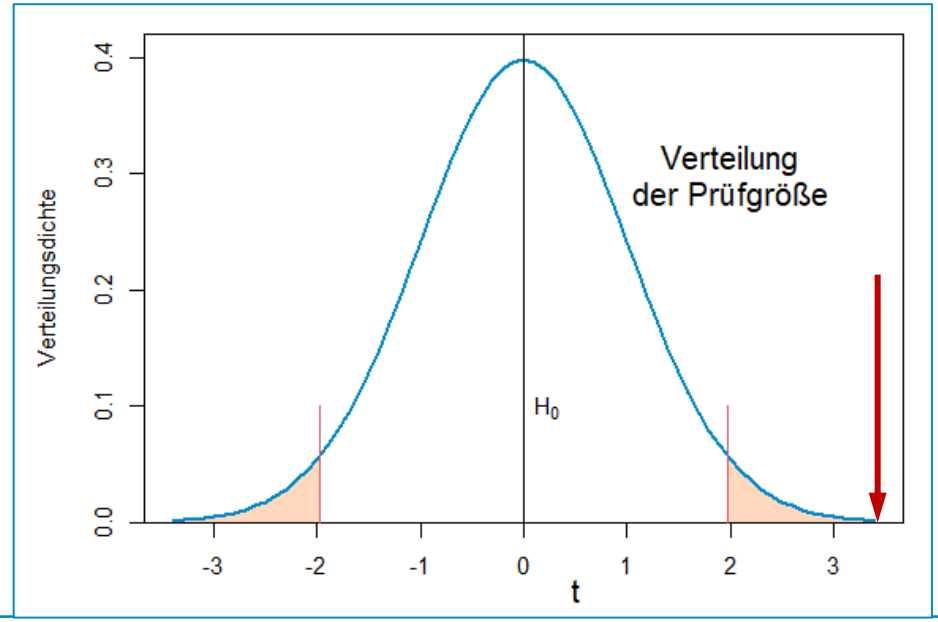
b) Software (z. B. R): Ist wird H_0 abgelehnt.

$$p \leq \alpha,$$

```
> t.test(delta$dvo[delta$arm == 1], delta$dvo[delta$arm == 2])

        Welch Two Sample t-test

data:  delta$dvo[delta$arm == 1] and delta$dvo[delta$arm == 2]
t = 3.3605, df = 255.6, p-value = 0.0008968
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6391192 2.4484333
sample estimates:
  mean of x   mean of y
 1.48153846 -0.06223776
```



Einseitige Alternative

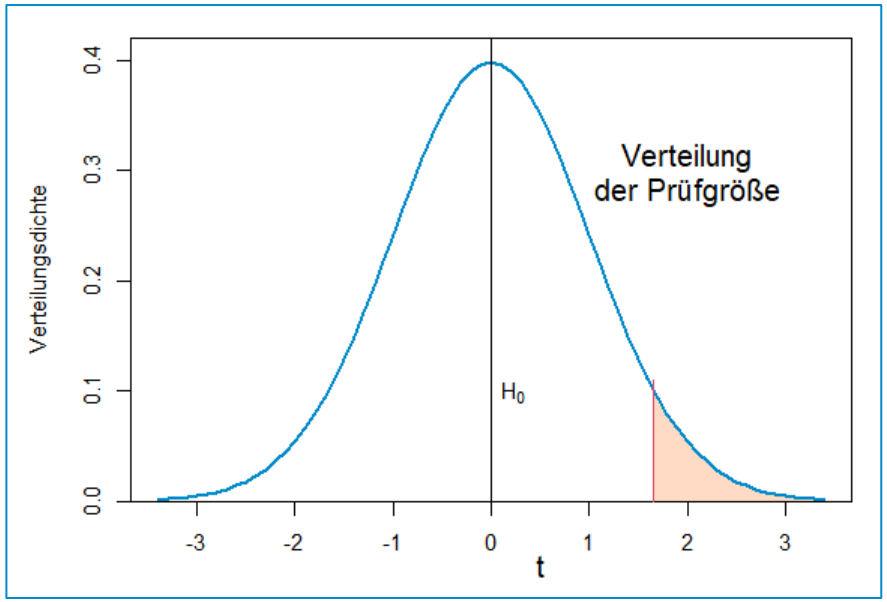
Kann man von vornherein eine Richtung des Unterschieds ausschließen, z. B. weil

- das medizinisch unmöglich ist oder nur
- eine Richtung interessant ist,

formuliert man die Alternativhypothese einseitig:

- H_0: $\mu_1 = \mu_2$
- H_A: $\mu_1 > \mu_2$

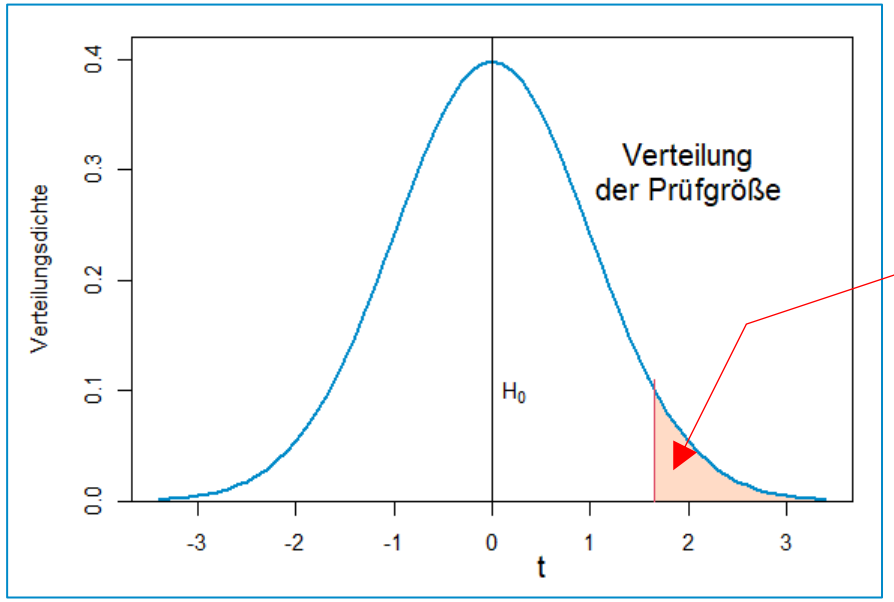
Dann ist der Ablehnbereich einseitig.



Einseitige Tests sind im Umfeld klinischer Studien die Ausnahme, welche stichhaltig begründet werden muss.

Testentscheidung (einseitiger Test)

- Der Ablehnbereich wird dann z. B. durch die Bedingung $t \geq t_{f;\,1-\alpha}$ beschrieben.



- Ist diese Bedingung erfüllt, ‘ wird H_0 abgelehnt.

- Mit Software gilt nach wie vor:
  H_0 wird abgelehnt, wenn $p \leq \alpha$ ist, sonst beibehalten.
  (Achtung: Einseitigen Test beachten!)

Signifikanzniveau und p-Wert

Konstante,

- die vor Studienbeginn
- unabhängig von den Daten basierend auf klinischen und ethischen Abwägungen festgelegt wird,
- die den Ablehnbereich für die Testentscheidung bestimmt und
- im long run den Fehler I. Art nach oben begrenzt.

- Zufallsgröße,
- die aus der konkreten Stichprobe berechnet wird,
- die als Wahrscheinlichkeit definiert ist, diese oder eine noch extremere Stichprobe aus der GG zu ziehen, wenn H_0 gelten würde.
- Der p-Wert ist ein (mangelhaftes) Evidenzmaß gegen H_0.

Angabe von p-Werten

- Die Angabe von p-Werten sollte erfolgen unter der Angabe
  - der Hypothesen H_0 und H_A, für die er errechnet wurde und
  - der Fallzahl, auf der er basiert.
- mit dem genauen Wert (gerundet auf zwei nichttriviale Dezimalen),
  - z. B. sinnvolle Angaben: $p = 0{,}017$ oder $p = 0{,}34$;
  - zu vermeidende Angaben: $p < 0{,}05$ oder $p = 0{,}7823$

Ein kleiner p-Wert sagt nichts darüber aus,

- ob ein beobachteter Effekt klinisch relevant ist und
- ob ein kausaler Zusammenhang besteht

Ablaufschema für statistische Tests I

1. Festlegung der Zielsetzung (Was will man wissen?):
   Fragestellung / Zielgröße
2. Stichproben (Versuchsplanung)
   Anzahl / Paarung
3. Formulierung der Hypothesen H_0 und H_A
4. Wahl des Signifikanzniveaus α
5. Auswahl der Prüfgröße und der dazugehörigen Testverteilung
6. Prüfung von Testvoraussetzungen
   - Modellannahmen (z. B. Verteilung, Varianzhomogenität)
   - ggf. Modifikation von 5.

Ablaufschema für statistische Tests II

7. Ermittlung des Ablehnbereiches (bzw. Annahmebereichs)
8. Berechnung der Prüfgröße aus der Stichprobe
9. Testentscheidung
   - Handrechnung: Vergleich der berechneten Prüfgröße mit dem Quantil der Prüfverteilung oder
   - Software: Ausgabe des *p*-Wertes: Vergleich mit α

   → H_0 ablehnen oder nicht
10. Interpretation mit Angabe eines Effektes (z. B. Mittelwertdifferenz)

Parallelen

	Hypothesen (*die Wahrheit*)	
Test	H_0	H_A
H_0 beibehalten	☺	Fehler II. Art
H_0 _ ablehnen	Fehler I. Art	☺

	Krankheitsstatus	
Test		
Test negativ	☺	
Test positiv		☺

Basiswahrscheinlichkeit für einen Effekt ← Prävalenz

Parametrische / nichtparametrische Tests

- Parametrische Tests
  - setzen eine bestimmte Verteilung (z. B. Normalverteilung) voraus.
  - Die Testgröße wird aus dem geschätzten Parameter (z. B. Mittelwert) berechnet
  - und daraus der p-Wert bestimmt.

- Nichtparametrische Tests (syn. verteilungsfreie Tests)
  - nur sehr allgemeine Annahmen zur Verteilung
  - Ermittlung der Prüfgröße z. B. aus Rängen der Stichprobe
  - V 10

V 9

9.3 Mittelwertvergleiche

t-Tests zum Vergleich von Mittelwerten

1. Vergleich eines Mittelwertes mit einem vorgegebenen Referenzwert μ_0

$$H_0: \ \mu = \mu_0$$

1 Stichprobe

2. Vergleich der Mittelwerte zweier Gruppen (z. B. Studienarme)

$$H_0: \ \mu_1 = \mu_2$$

2 Stichproben (unabhängig)

3. Vergleich der Mittelwerte vor und nach einer Intervention an einer Versuchsgruppe

$$H_0: \ \mu_{vor} = \mu_{nach}$$

2 Stichproben (gepaart / verbunden)

Beispiel 9.2 Normwertvergleich

- Ein bestimmtes Arzneimittel wird in Flaschen zu je 50 ml angeboten (Standardabweichung: 1 ml).
- Bei der Qualitätskontrolle werden regelmäßig Stichproben von je 20 Flaschen geprüft.
- Ist der erhaltene Mittelwert von 51,09 ml mit der Zielmenge 50 ml vereinbar?
- Dabei sollen Abweichungen sowohl nach unten als auch nach oben berücksichtigt werden!
- Signifikanzniveau 5 %
- Normalverteilung der erhobenen Werte wird vorausgesetzt.

→ 1-Stichproben-t-Test

1-Stichproben t-Test

Hypothesen

- H_0: $\mu = \mu_0$ vs. H_A: $\mu \neq \mu_0$ (zweiseitige Fragestellung)

Prüfgröße:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

ist *t*-verteilt mit *n-1* Freiheitsgraden

1-Stichproben t-Test – Beispiel 9.2

1. Zielgröße: mittleres Flaschenvolumen
2. Stichprobe: $\bar{x} = 51.09\,, s = 1,\ n = 20,\ \mu_0 = 50$!
3. Hypothesen: $H_0\colon\ \mu = \mu_0 = 50\ \text{ml} \quad \Leftrightarrow \quad \mu - \mu_0 = 0$

   $H_A\colon\ \mu \neq \mu_0$
4. Signifikanzniveau: $\alpha = 0{,}05$; zweiseitig
5. Testvoraussetzungen: keine markanten Abweichungen von der Normalverteilung
6. Prüfgröße: $t = \frac{Mittelwert - \mu_0}{S.E.M.} = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$

Beispiel 9.2 II

8. Berechnung der Prüfgröße: $t = \frac{51{,}09-50}{1} \cdot \sqrt{20} = 4{,}875$

9. Testentscheidung: $|t| = 4{,}875 > t_{19;1-0{,}05/2} = 2{,}09$, d. h.

   H_0 wird abgelehnt

10. Interpretation:

Die mittlere Flaschenfüllung dieser Stichprobe ist gegenüber dem Normwert μ_0 = 50 ml zum Signifikanzniveau von 5% im Mittel um 1,09 ml signifikant höher.

t-Tests zum Vergleich von Mittelwerten

1. Vergleich eines Mittelwertes mit einem vorgegebenen Referenzwert μ_0

✓

$$H_0: \ \mu = \mu_0$$

1 Stichprobe

2. Vergleich der Mittelwerte zweier Gruppen (z. B. Studienarme)

$$H_0: \ \mu_1 = \mu_2$$

2 Stichproben (unabhängig)

3. Vergleich der Mittelwerte vor und nach einer Intervention an einer Versuchsgruppe

$$H_0: \ \mu_{vor} = \mu_{nach}$$

2 Stichproben (gepaart / verbunden)

B 9.1 Ex-DHF-Studie

Studienfrage:

- Gibt es im Trainingsarm unter der Intervention einen Unterschied in der Veränderung der peak VO_2 in 12 Monaten zum Kontrollarm.

Szenario:

- Werte einer metrischen Zufallsvariable, erfasst von Probanden zweier Gruppen: $x_{11}, x_{12}, \ldots, x_{1n_1}; x_{21}, x_{22}, \ldots, x_{2n_2};$
- Normalverteilung wird angenommen: $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2),$
- $H_0: \mu_1 = \mu_2$ vs. $H_A: \mu_1 \neq \mu_2$
- Signifikanzniveau: 1%

t-Test (Welch)

Varianz: $\sigma_1^2 \neq \sigma_2^2$; Freiheitsgrade (FG) aufwändig zu berechnen

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_f$$

t – verteilt mit f Freiheitsgraden

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

(Welch BL, Biometrika 1947;34:28-35)

t-Test (Sonderfall: gleiche Varianzen)

- Varianz: $\boldsymbol{\sigma_1^2 = \sigma_2^2}$, $n_1 = n_2 = n$
- Prüfgröße: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 + s_2^2}} \cdot \sqrt{n} \sim t_{2n-2}$

Empfehlung:

- Nach Möglichkeit sollte immer Welchs Test (vorige Folie) angewandt werden.
  - Er ist sehr robust gegenüber Abweichungen von der Normalverteilung
  - und hat insbesondere für ungleiche Varianzen höhere Power.

  (z. B. Rasch D et al. Stat Papers 2011;52:219-231)

t-Test (2 unabhängige Stichproben) – Beispiel I

1. Zielgröße: peak VO_2
2. Stichproben: 2 unabhängige
3. Hypothesen: $H_0: \mu_{ET} = \mu_{UC} \quad vs. \quad H_A: \mu_{ET} \neq \mu_{UC}$

– H_0: Die maximale Sauerstoffaufnahme hat sich in beiden Studienarmen im Mittel in gleicher Weise verändert.

– H_A: Die Differenz der maximale Sauerstoffaufnahme (12 Monate – baseline) unterscheidet sich in beiden Studienarmen.

4. Signifikanzniveau 1%, zweiseitiger Test

t-Test (2 unabhängige Stichproben) – Beispiel I

5. Prüfgröße:

Stichprobe: $\bar{x}_1 = 1,5;\ \bar{x}_2 = -0,1; s_1 = 3,5 \approx s_2 = 3,6$

$n_1 = n_2 = 144;$

$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{{s_1}^2 + {s_2}^2}} \cdot \sqrt{n}$, also $t = \frac{1,5-(-0,1)}{\sqrt{3,6^2+3,6^2}} \cdot \sqrt{144} = 3,77$

$f = 2 \cdot 144 - 2 = 286$

6. Testannahmen:

Normalverteilung, keine extremen Ausreißer

Varianz in den Gruppen ist gleich, $n_1 = n_2$

t-Test (2 unabhängige Stichproben) – Beispiel II

7. Kritischer Bereich: $|t| \geq t_{286,0.995} = 2{,}58$ (wegen $f > 100$ Normalquantil)
8. Testentscheidung: $t = 3{,}77 > 2{,}58$; d. h.,
   H_0 wird abgelehnt.

9. Interpretation:

– Patienten im Interventionsarm verbesserten ihre peak VO_2 im Mittel um 1,5 ml / min / kg, währenddessen diese im Kontrollarm fast gleich blieb.
– Das ist signifikant zum 1%-Niveau.
– Der Unterschied in den Differenzen der peak VO_2 beträgt 1,6 ml / min / kg.

t-Test (2 unabhängige Stichproben) – Beispiel III

7. Kritischer Bereich: $|t| \geq t_{286,0.995} = 2{,}58$ (wegen $f > 100$ Normalquantil)

8. Die Software R liefert: $p = 0{,}00016 < 0{,}01$
   D. h., H_0 wird abgelehnt.

9. Interpretation wie vorige Folie.

t-Tests zum Vergleich von Mittelwerten

1. Vergleich eines Mittelwertes mit einem vorgegebenen Referenzwert μ_0

$$H_0: \ \mu = \mu_0$$

1 Stichprobe

2. Vergleich der Mittelwerte zweier Gruppen (z. B. Studienarme)

$$H_0: \ \mu_1 = \mu_2$$

2 Stichproben (unabhängig)

3. Vergleich der Mittelwerte vor und nach einer Intervention an einer Versuchsgruppe

$$H_0: \ \mu_{vor} = \mu_{nach}$$

2 Stichproben (gepaart / verbunden)

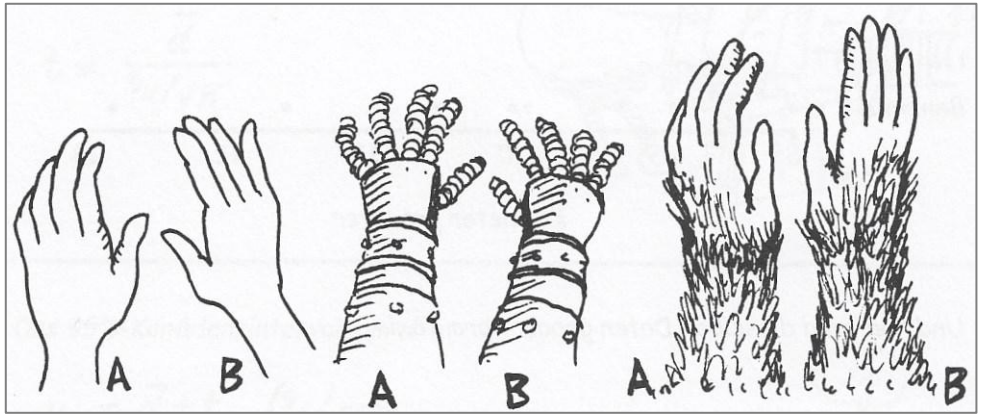
Abhängige Stichproben I

- Abhängige (syn.: gepaarte oder verbundene) Stichproben liegen vor, wenn:
  - es zu denselben Beobachtungseinheiten mehrere Messreihen gibt,
  - diese also voneinander abhängig sind.
- Typischer Anwendungsfall: Cross-over-Studien (→ V 1)
  - Da beide Gruppen dieselben Behandlungen (in unterschiedlicher Reihenfolge) erhalten,
  - sind die Probanden ihre eigenen Kontrollen und
  - wird ein Großteil von Störgrößen ausgeschaltet und die Aussagekraft der Analysen erheblich gesteigert!

Abhängige Stichproben II

weitere Beispiele für abhängige Stichproben:

- vorher - nachher Untersuchungen
- paarige Organe
- Zellkulturen des gleichen Klons
- Zwillingsexperimente
- Evaluation durch mehrere Gutachter



Gonick & Smith: Statistik in Cartoons

t-Test (2 abhängige Stichproben) I

Grundidee: t-Test für 2 abhängige Gruppen

- Bilde die Differenzen bei jedem Wertepaar
- und teste die Verteilung der Differenzen
- auf Abweichungen von 0!

Voraussetzungen:

- $x_{11}, x_{12}, \ldots, x_{1n}; \quad x_{21}, x_{22}, \ldots, x_{2n}$
  2 Messreihen zu denselben Beobachtungseinheiten
- normalverteilt

t-Test (2 abhängige Stichproben) II

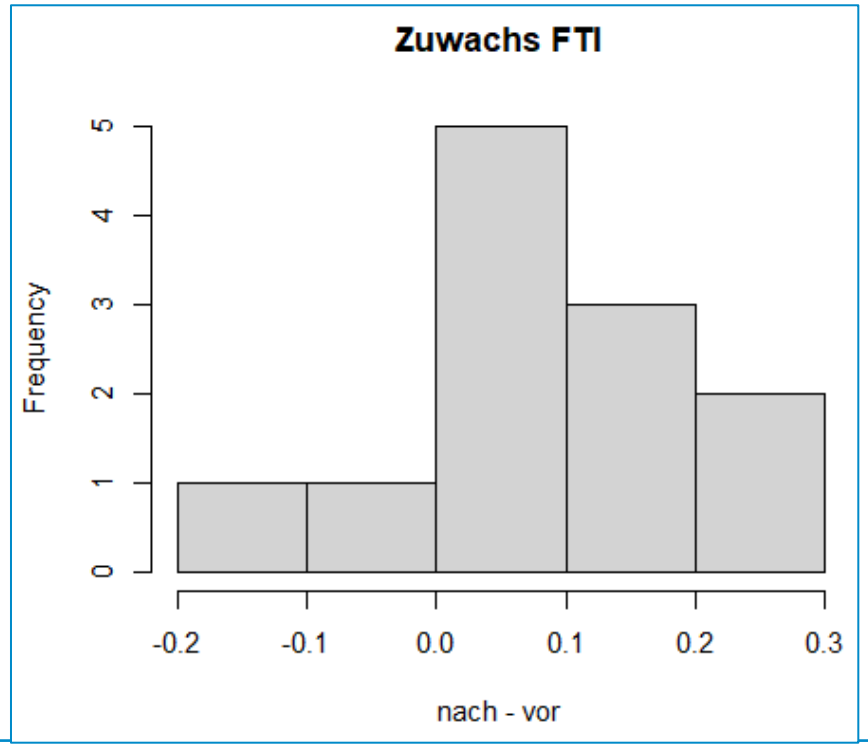
- Differenzen pro Beobachtungseinheit: $d_i = x_{1i} - x_{2i} \quad i = (1, \ldots, n)$
- Hypothesen H_0: $\mu_d = 0$ $(\mu_1 = \mu_2)$ H_A: $\mu_d \neq 0$ $(\mu_1 \neq \mu_2)$
- Anwendung des 1-Stichproben-t-Tests mit Referenzwert 0
- Prüfgröße: $$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} = \frac{\bar{d}}{s_d} \cdot \sqrt{n} \sim t_{n-1}$$ t-verteilt mit (n-1) Freiheitsgraden

t-Test (2 abhängige Stichproben) – Beispiel I

In einer Stichprobe von 12 Patienten wurde der freie Thyroxin-Index (FTI) vor und nach Injektion des Hypophysenhormons TSH bestimmt.

- Unterscheidet sich der freie Thyroxin-Index im Mittel vor und nach Injektion von TSH?

- Normalverteilung der Differenzen (vor-nach) wird vorausgesetzt.
- Signifikanzniveau 2 %



Daten

Thyroxin-Index			
Pat.-Nr.	vor Injektion	nach Injektion	Differenz
1	1,04	1,06	-0,02
2	0,88	0,91	-0,03
3	1,02	1,13	-0,11
4	1,09	1,17	-0,08
5	0,85	1,00	-0,15
6	0,97	1,26	-0,29
7	0,92	1,11	-0,19
8	1,17	1,05	0,12
9	1,07	0,98	0,09
10	0,85	0,95	-0,1
11	0,96	1,04	-0,08
12	0,84	1,05	-0,21
Arith. Mittel	0,97	1,06	$\bar{d} = -0{,}0875$
Std.-Abw.	0,11	0,10	$s_d = 0{,}118$

t-Test (2 abhängige Stichproben) – Beispiel III

1. Zielgröße: freie Thyroxin-Index - Mittelwert
2. Stichproben: 2 gepaarte
3. Hypothesen: H_0: $\mu_d = 0$ H_A: $\mu_d \neq 0$ zweiseitige Fragestellung

oder:

- H_0: Der freie Thyroxin-Index vor und nach Injektion des Hypophysenhormons TSH unterscheidet sich im Mittel *nicht*.
- H_A: Der freie Thyroxin-Index vor und nach Injektion des Hypophysenhormons TSH unterscheidet sich im Mittel.

t-Test (2 abhängige Stichproben) – Beispiel IV

4. Signifikanzniveau: $\alpha = 0{,}02$; zweiseitig

5. Prüfgröße: Stichprobe: $\bar{d}, s_d, n$ $\quad t = \frac{\bar{d}}{s_d} \cdot \sqrt{n}$

   t-verteilt mit $n - 1$ Freiheitsgraden

6. Testvoraussetzungen: Die Differenzen (vor-nach) des freien Thyroxin-Index sind normalverteilt.

t-Test (2 abhängige Stichproben) – Beispiel V

7. Kritischer Bereich: $\mathrm{f} = n - 1 = 12 - 1 = 11$ und $t_{11;1-0{,}02/2} = 2{,}72$;

   $|t| \geq 2{,}72$

8. Berechnung der Prüfgröße: $\bar{d} = -0{,}0875$, $s_d = 0{,}118$, $n = 12$

$$t = \frac{\bar{d}}{s_d} \cdot \sqrt{n} = \frac{-0{,}0875}{0{,}118} \cdot \sqrt{12} = -2{,}56$$

9. Testentscheidung: $|t| = 2{,}56 < t_{11;\,1-0{,}02/2} = 2{,}72$

   D. h., H_0 kann nicht verworfen werden.

SPSS liefert den p-Wert: $p = 0{,}026 > \alpha = 0{,}02$

→ H_0 wird beibehalten.

t-Test (2 abhängige Stichproben) – Beispiel VI

- Interpretation: *Der freie Thyroxin-Index (FTI) ist nach Injektion von TSH im Mittel um 0,088 Einheiten höher als vor Injektion. Das ist nicht signifikant zum Niveau 2%* $(p = 0{,}026)$.

- Wir dürfen nicht argumentieren:
  ~~*Der freie Thyroxin-Index hat sich im Mittel nicht verändert.*~~
- Immerhin ist der Mittelwert um 0,088 Einheiten angestiegen.
- Es ist zu klären, ob diese Veränderung klinisch relevant ist.



t-Tests zum Vergleich von Mittelwerten

1. Vergleich eines Mittelwertes mit einem vorgegebenen Referenzwert μ_0

$$H_0: \ \mu = \mu_0$$

1 Stichprobe

2. Vergleich der Mittelwerte zweier Gruppen (z. B. Studienarme)

$$H_0: \ \mu_1 = \mu_2$$

2 Stichproben (unabhängig)

3. Vergleich der Mittelwerte vor und nach einer Intervention an einer Versuchsgruppe

$$H_0: \ \mu_{vor} = \mu_{nach}$$

2 Stichproben (gepaart / verbunden)

Zusammenfassung I

- t-Tests prüfen die Lage von Verteilungen stetiger Merkmale.
- Die Prüfgröße der t-Tests bei Mittelwertvergleichen folgt in allen 3 Szenarien derselben Logik.
  Folglich ähneln sich die Formeln der Prüfgrößen:

$$\frac{Mittelwert}{Standardfehler} \sim t - \text{verteilt}$$

mit einer geeigneten Zahl von Freiheitsgraden

Zusammenfassung II

- Außer dem Testergebnis sollte immer auch ein Maß für den Effekt (z. B. Mittelwertunterschied), am besten mit C.I. angegeben werden.
- Der Test nach Welch ist sehr robust gegenüber
  - Abweichungen gegenüber der Normalverteilung und
  - unterschiedlichen Varianzen (vgl. Rasch et al. Stat Papers 2011;52:219–231).

- Er kann also ohne Vortests auf Normalverteilung oder Varianzhomogenität angewandt werden. Ausnahme sind ordinale Größen oder kleine Stichproben, die schief verteilt sind (Ausreißer!).
- Das kann man mit Hilfe von Histogrammen prüfen.



UNIVERSITÄT LEIPZIG

Medizinische Fakultät

VIELEN DANK!

Dr. rer. nat. Meinhard Mende

Senior-Wissenschaftler, Institut für Medizinische Informatik, Statistik und Epidemiologie, Härtelstraße 16-18, 04107 Leipzig

Meinhard.mende@zks.uni-leipzig.de