Statistik für Data Scientists

Vorlesung 7: Hypothesentests

Prof. Dr. Siegfried Handschuh
DS-NLP | Universität St. Gallen

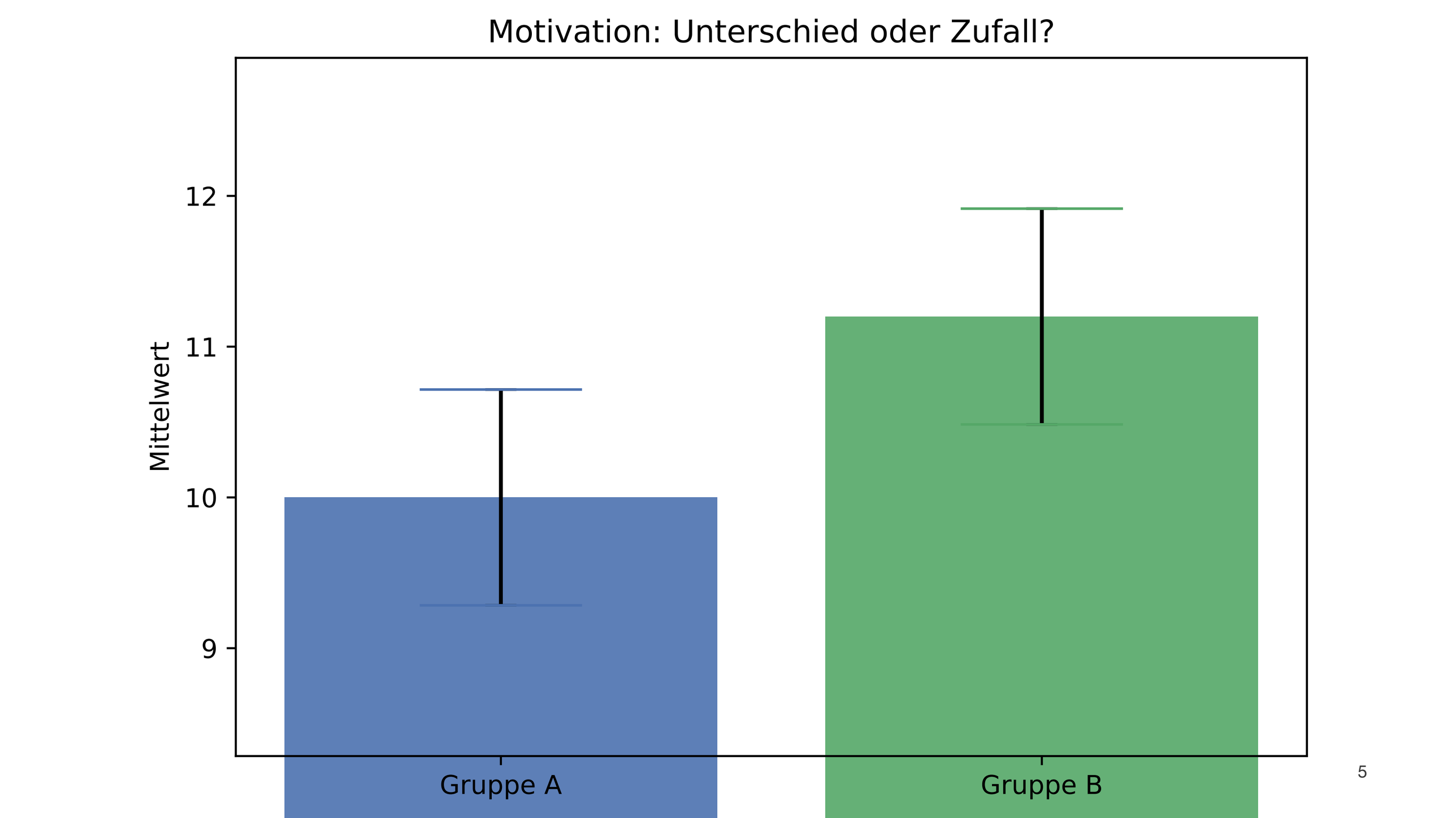
Recap & Ziele heute

- Rückblick VL6: Schätzen & Konfidenzintervalle.
- Heute: Grundlagen der Hypothesentests.
- Null- und Alternativhypothese (H_0, H_1).
- Fehlerarten & Teststärke (Power).
- p-Wert richtig interpretieren.
- t-Test an Beispielen.
- Kritik & gute Praxis.

Motivation & Grundidee

Warum Hypothesentests?

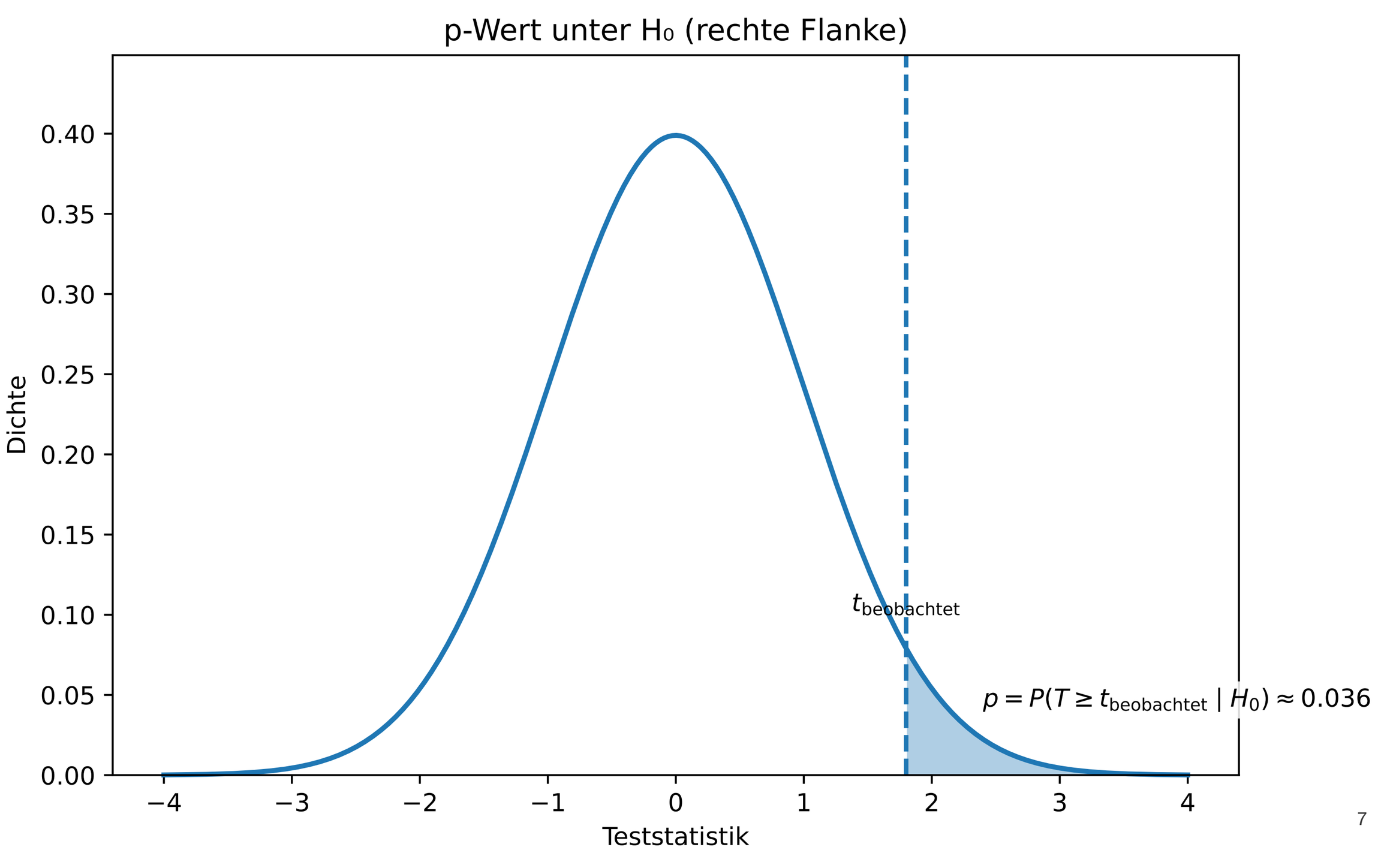
- Frage: Ist ein Unterschied zufällig oder echt?
- Beispiele:
  - Wirkt ein Medikament tatsächlich?
  - Funktioniert A/B-Variante besser?
  - Verdienen Männer und Frauen gleich viel?
- Statistik liefert ein objektives Verfahren, um Unsicherheit zu quantifizieren.


Motivation: Unterschied oder Zufall?
Mittelwert
12
11
10
9
Gruppe A
Gruppe B

Grundidee des Tests

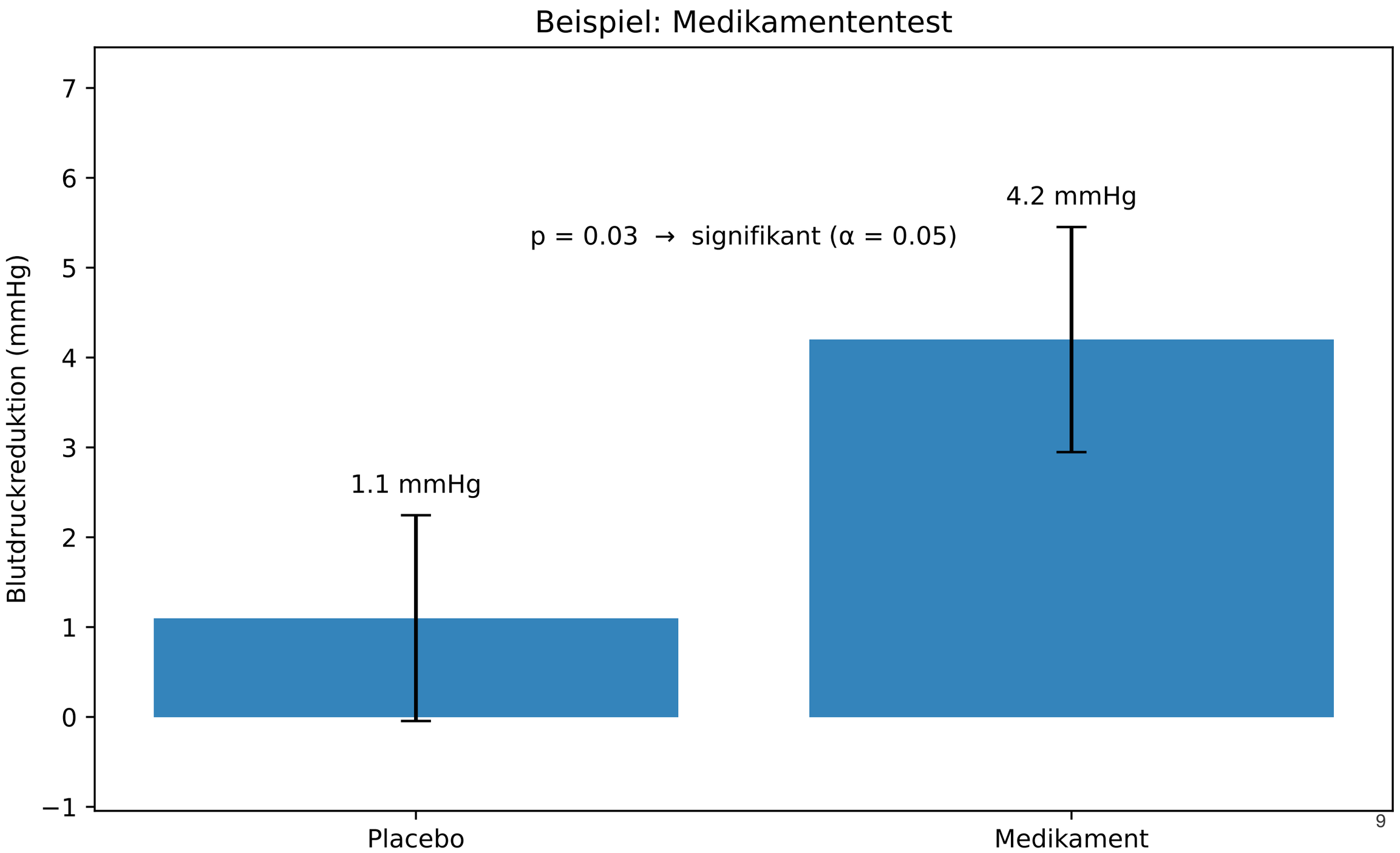
- Formuliere H_0: kein Effekt.
- Formuliere H_1: Effekt vorhanden.
- Wähle eine Teststatistik T.
- Vergleiche den beobachteten Wert $t_{\text{beobachtet}}$ mit der Referenzverteilung unter H_0.
- Kleine p-Werte → starke Evidenz gegen H_0.

$$p = P(T \geq t_{\text{beobachtet}} \mid H_0)$$


p-Wert unter H₀ (rechte Flanke)
Dichte
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−4
−3
−2
−1
0
1
2
3
4
Teststatistik
$t_{\text{beobachtet}}$
$p = P(T \geq t_{\text{beobachtet}} \mid H_0) \approx 0.036$

Beispiel: Medikamententest

- Zwei Gruppen: Medikament vs. Placebo.
- Mittelwert der Blutdruckreduktion: 4.2 mmHg vs. 1.1 mmHg.
- t-Test → $p = 0.03$ bei $\alpha = 0.05$ → H_0 verwerfen.
- Signifikant? Ja; praktisch relevant? Fraglich.


Beispiel: Medikamententest
Blutdruckreduktion (mmHg)
7
6
5
4
3
2
1
0
−1
4.2 mmHg
p = 0.03 → signifikant (α = 0.05)
1.1 mmHg
Placebo
Medikament

Take-away: Motivation & Grundidee

- Hypothesentests = Werkzeug gegen Zufallsurteile.
- Formale Bausteine: H_0, H_1, Teststatistik T, p-Wert.
- p-Wert misst Seltenheit der Daten unter H_0, nicht deren Wahrheit.
- Beispiel: Medikament – signifikant, aber nicht immer relevant.

Statistik liefert Entscheidungen unter Unsicherheit: nicht absolute Wahrheiten.

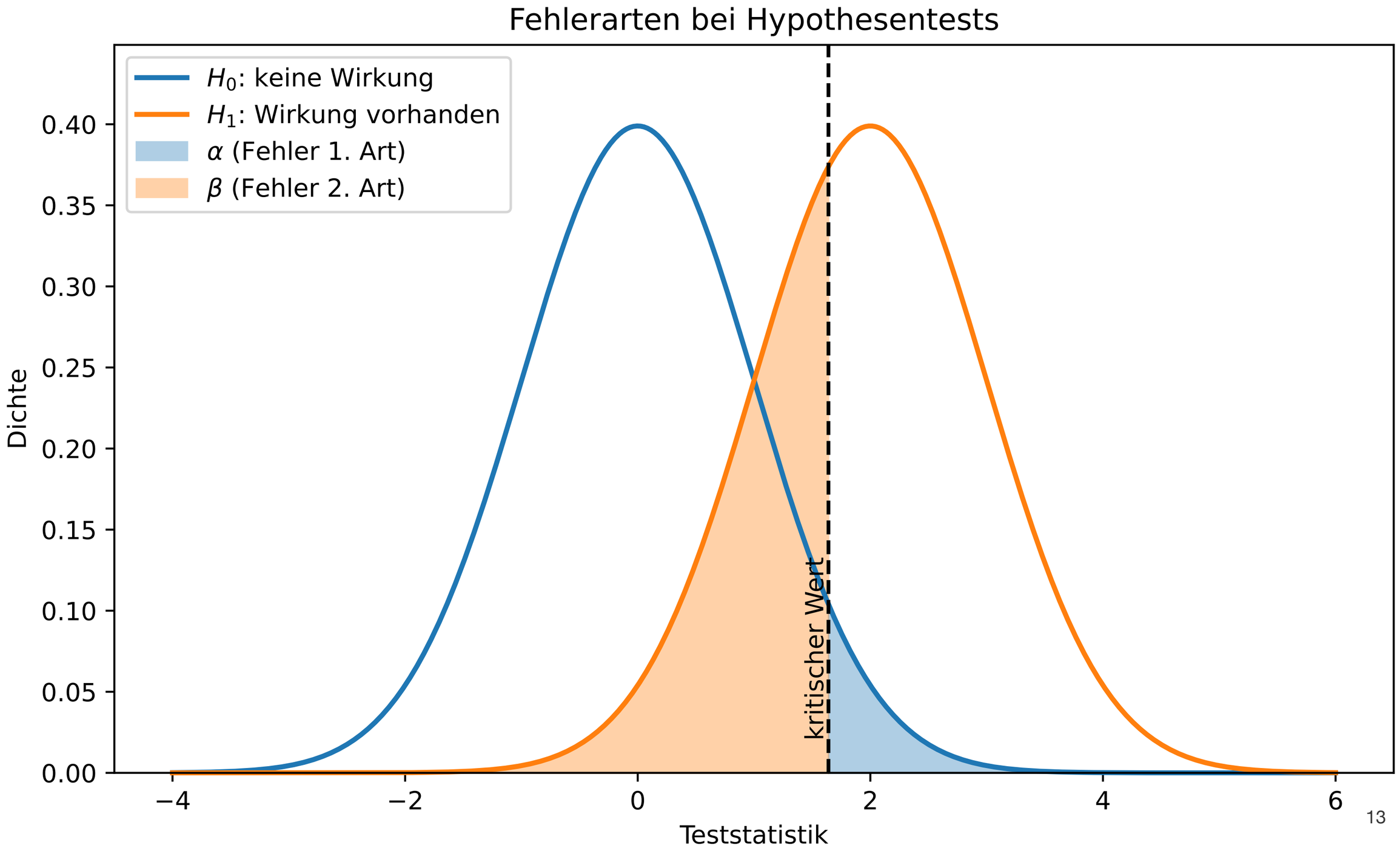
Fehlerarten & Power

Fehlerarten

Typ	Beschreibung	Symbol
Fehler 1. Art	H_0 ist wahr, wird aber verworfen. *‹False Positive› → Fehlalarm.*	α
Fehler 2. Art	H_0 ist falsch, wird aber nicht verworfen. *‹False Negative› → übersehener Effekt.*	β

Teststärke (Power): $1 - \beta$

Ziel: α klein halten, Power gross halten.


Fehlerarten bei Hypothesentests
H_0: keine Wirkung
H_1: Wirkung vorhanden
α (Fehler 1. Art)
β (Fehler 2. Art)
kritischer Wert
Dichte
Teststatistik
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−4
−2
0
2
4
6

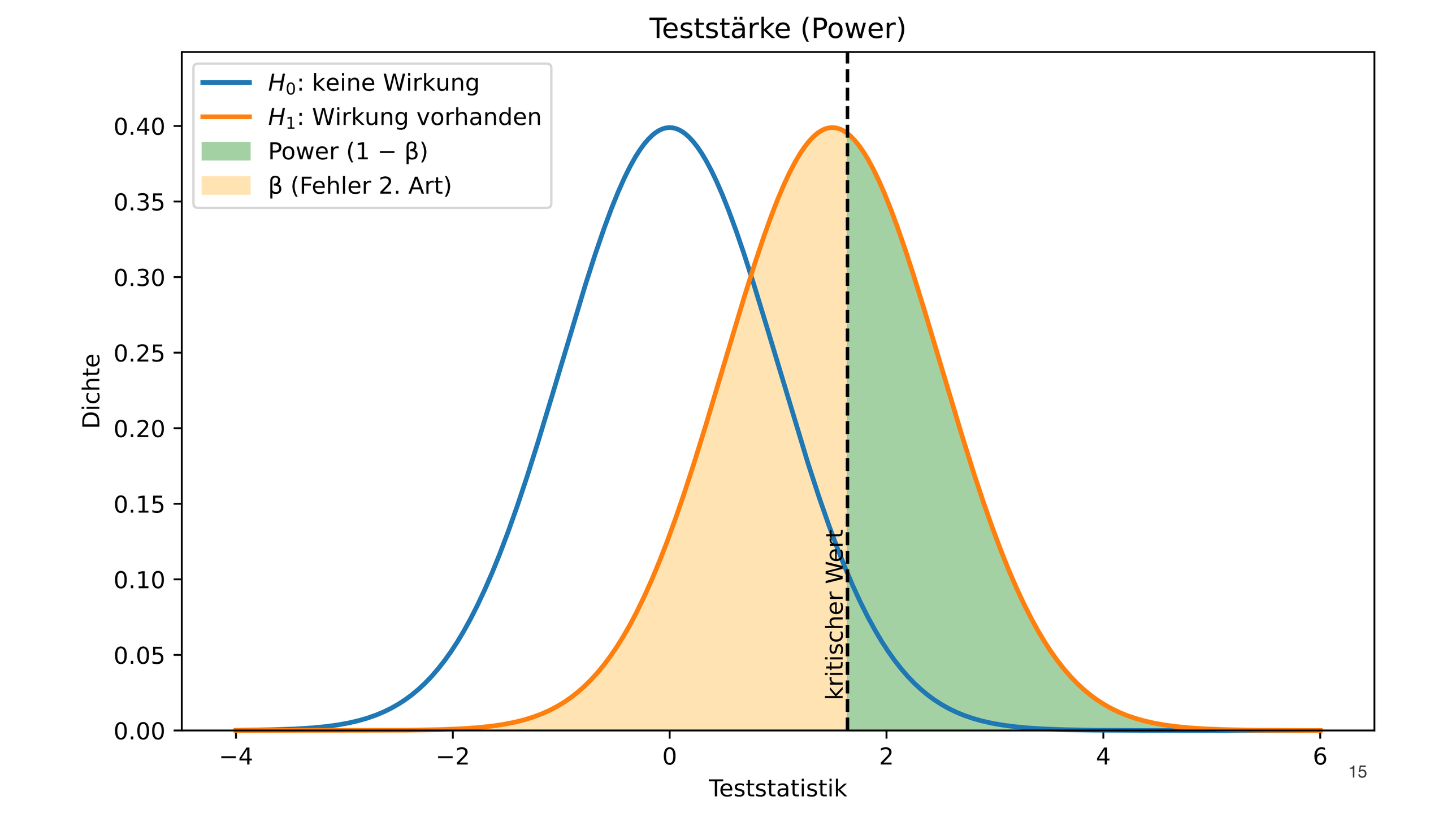
Teststärke (Power)

Definition: Wahrscheinlichkeit, einen echten Effekt zu entdecken

$$\text{Power} = 1 - \beta$$

Beispiel:
Ein Medikament wirkt tatsächlich.
- Power 80 % → In 8 von 10 Studien wird die Wirkung erkannt.
- Power 20 % → In 8 von 10 Studien wird sie übersehen.

Faustregel: Power ≥ 0.80 anstreben

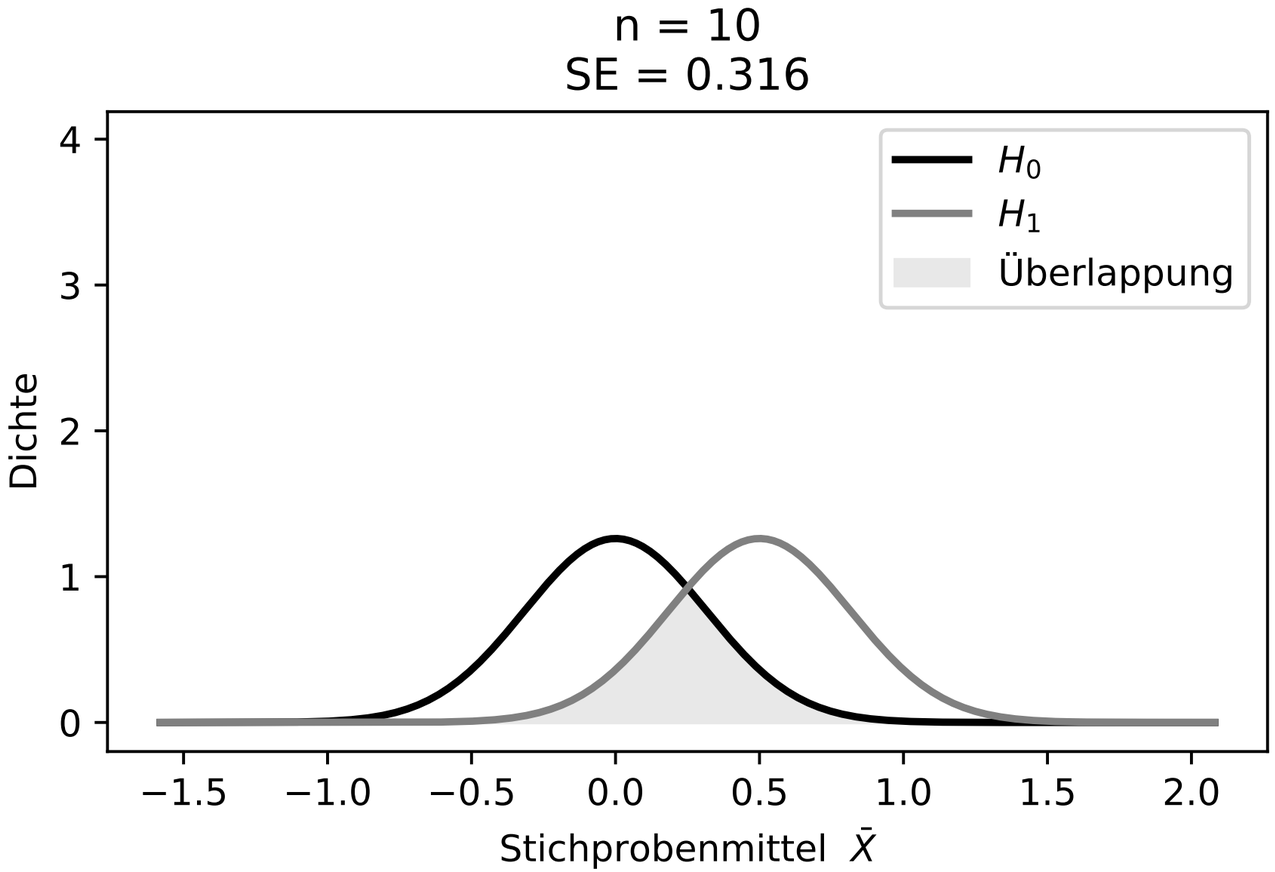

Teststärke (Power)
H_0: keine Wirkung
H_1: Wirkung vorhanden
Power (1 − β)
β (Fehler 2. Art)
kritischer Wert
Dichte
Teststatistik
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−4
−2
0
2
4
6

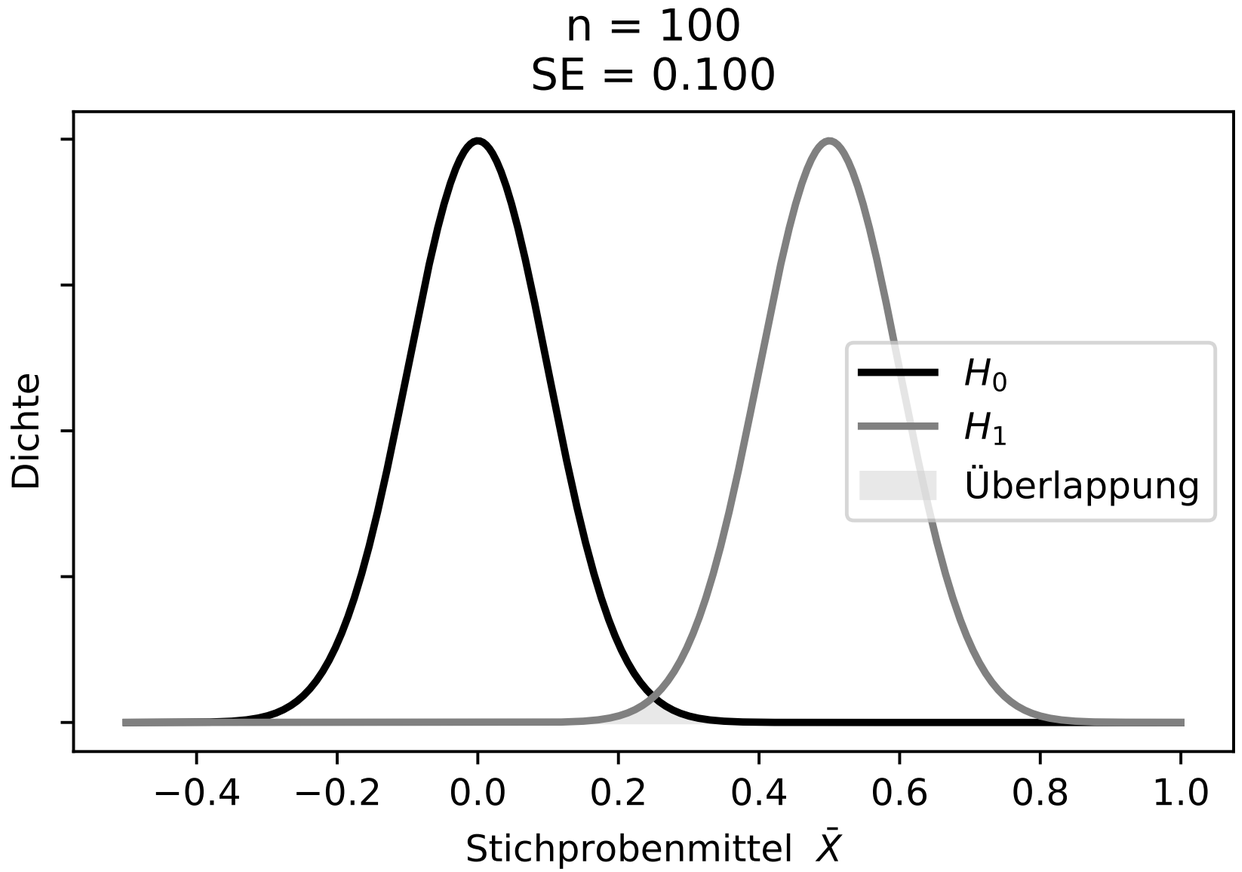
Einfluss auf Power

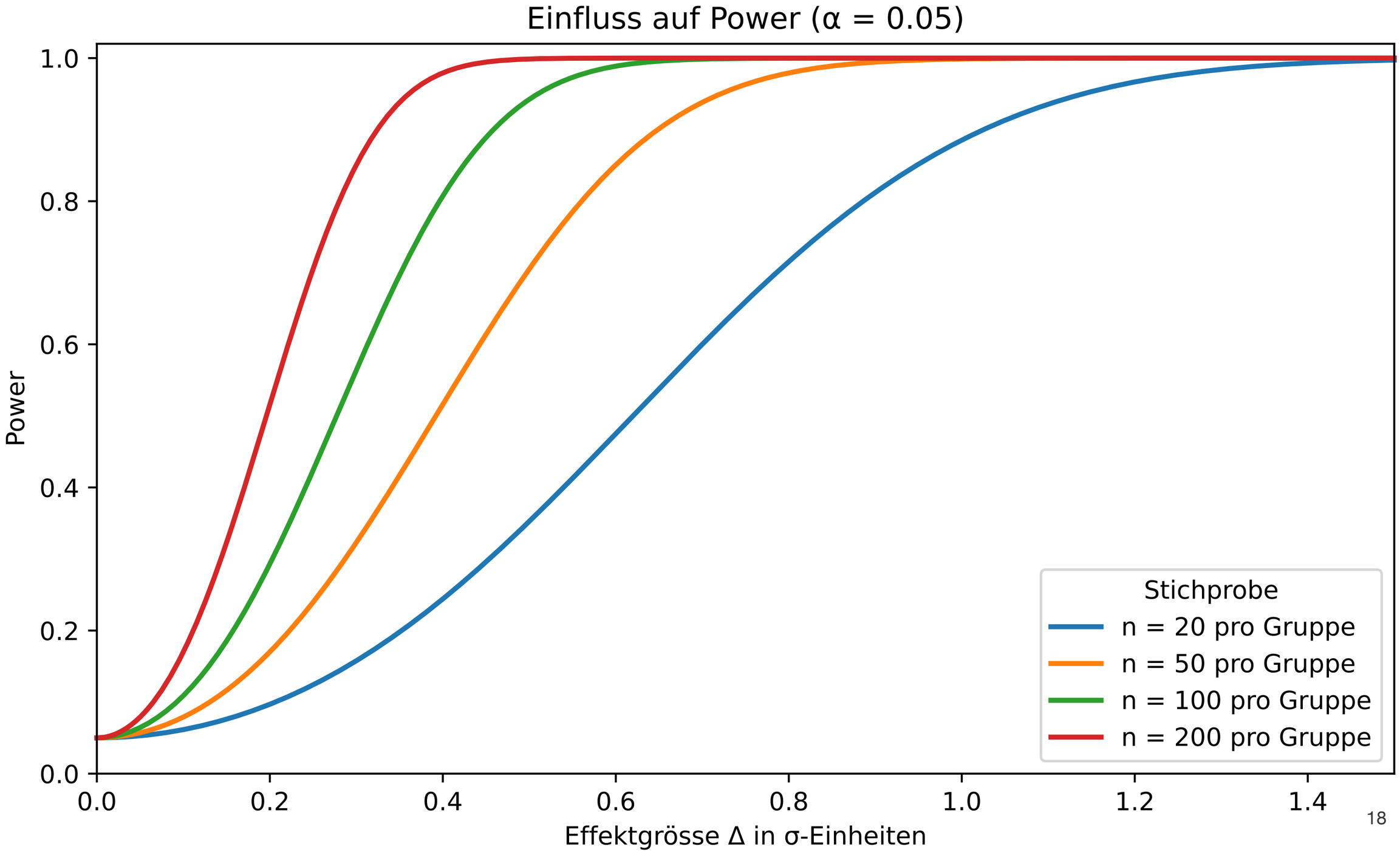
- Stichprobengrösse n: grösser → kleinerer Standardfehler → höhere Power.
- Effektgrösse Δ: grösser → leichter erkennbar.
- Signifikanzniveau α: höher → mehr Power, aber auch mehr Fehlalarme.

$$\text{Power} = f(n, \alpha, \Delta, \sigma)$$

Grössere Stichprobe → schmalere Kurven → geringere Überlappung → höhere Power






Einfluss auf Power (α = 0.05)
Power
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Effektgrösse Δ in σ-Einheiten
Stichprobe
n = 20 pro Gruppe
n = 50 pro Gruppe
n = 100 pro Gruppe
n = 200 pro Gruppe

Take-away: Fehlerarten & Power

- Fehler 1 (Typ I): H_0 wahr, aber verworfen → α (Fehlalarm).
- Fehler 2 (Typ II): H_0 falsch, aber beibehalten → β (übersehen).
- Power = $1-\beta$: Wahrscheinlichkeit, echten Effekt zu entdecken.
- Einflussfaktoren auf Power: $n \uparrow$, Effektgrösse $\Delta \uparrow$, $\alpha \uparrow$
  (Achtung: $\alpha \uparrow \Rightarrow$ mehr Fehlalarme).

Take-away:

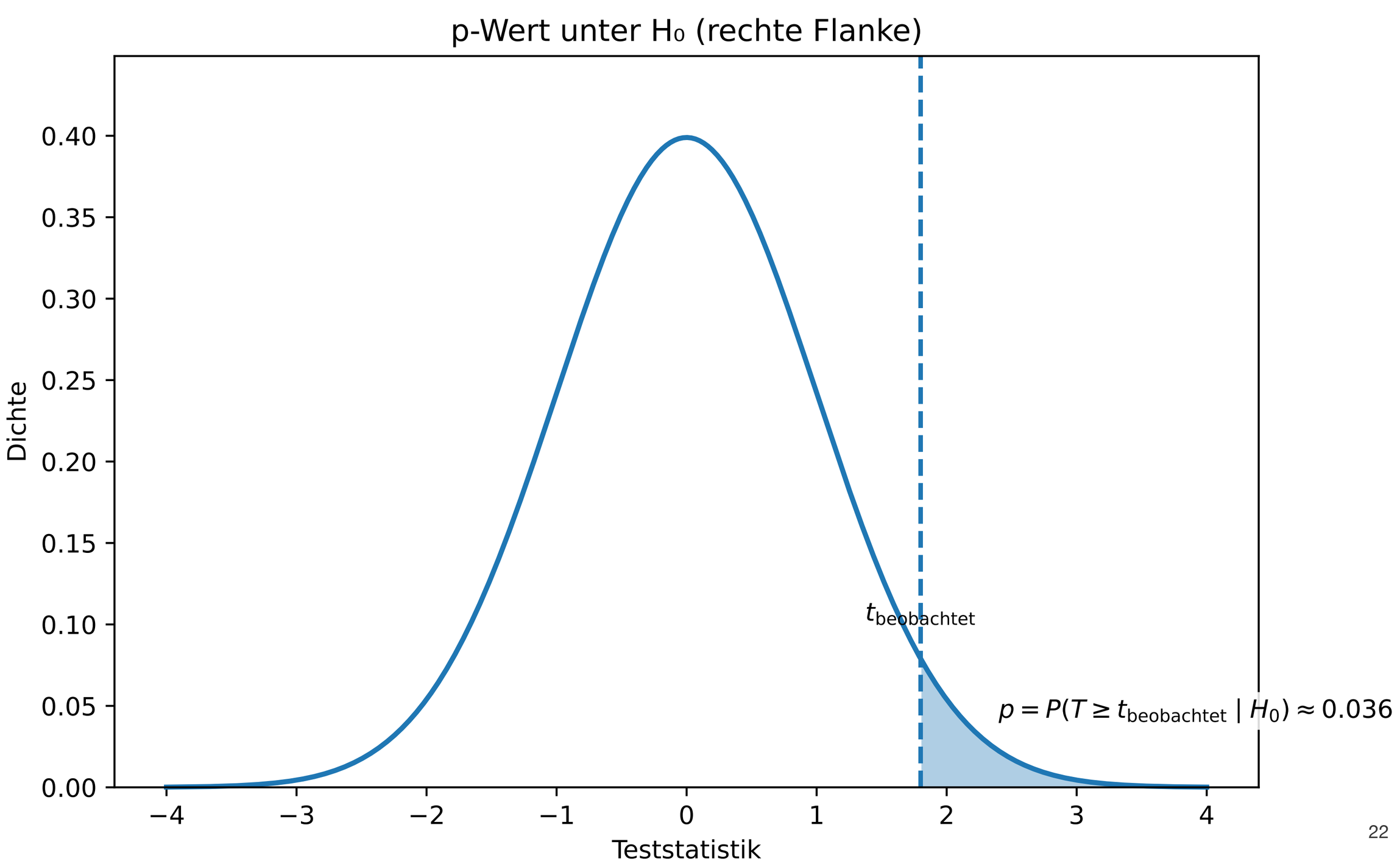
α und β sind keine Fehler, sondern Designentscheidungen $\rightarrow$ gute Statistik balanciert Risiko und Sensitivität.

p-Wert

p-Wert verstehen

- Der p-Wert ist die Wahrscheinlichkeit, unter H_0 ein Ergebnis so extrem oder extremer zu beobachten.
- Er misst die Kompatibilität der Daten mit H_0, nicht die Wahrscheinlichkeit, dass H_0 wahr ist.
- Kleine p → Daten ungewöhnlich unter H_0.
- Grosse p → Daten typisch unter H_0.

$$p = P(T \geq t_{\text{beobachtet}} \mid H_0)$$


p-Wert unter H₀ (rechte Flanke)
Dichte
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−4
−3
−2
−1
0
1
2
3
4
Teststatistik
$t_{\text{beobachtet}}$
$p = P(T \geq t_{\text{beobachtet}} \mid H_0) \approx 0.036$

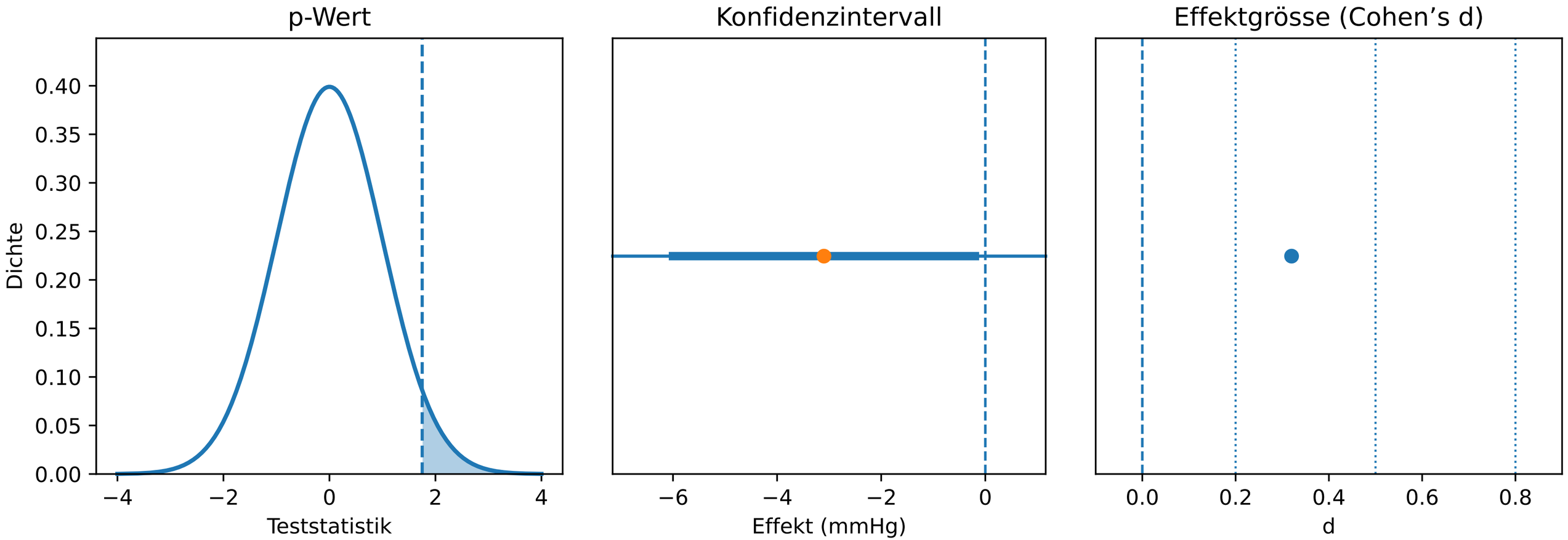
Typische Missverständnisse

- ❌ p = 0.03 → „H_0 ist zu 97 % falsch.“
- ❌ p = 0.20 → „H_0 stimmt.“
- ✅ Korrekt: Der p-Wert beschreibt, wie ungewöhnlich die Daten unter H_0 sind.
- Je kleiner p, desto weniger plausibel sind die Daten, wenn H_0 wahr wäre.

Beispiel: Medikamentenstudie

- Studie: Medikament senkt Blutdruck.
- Ergebnis: $p = 0.04 \rightarrow H_0$ verwerfen (bei $\alpha = 0.05$).
- Aber: Effektgrösse und Konfidenzintervall wichtiger als p allein.
- Immer p, CI und Effektgrösse gemeinsam berichten.

Cohen's d: Effektgrösse $\rightarrow$ StatCoach fragen


p-Wert
Konfidenzintervall
Effektgrösse (Cohen’s d)
Dichte
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
Teststatistik
−6
−4
−2
0
Effekt (mmHg)
0.0
0.2
0.4
0.6
0.8
d

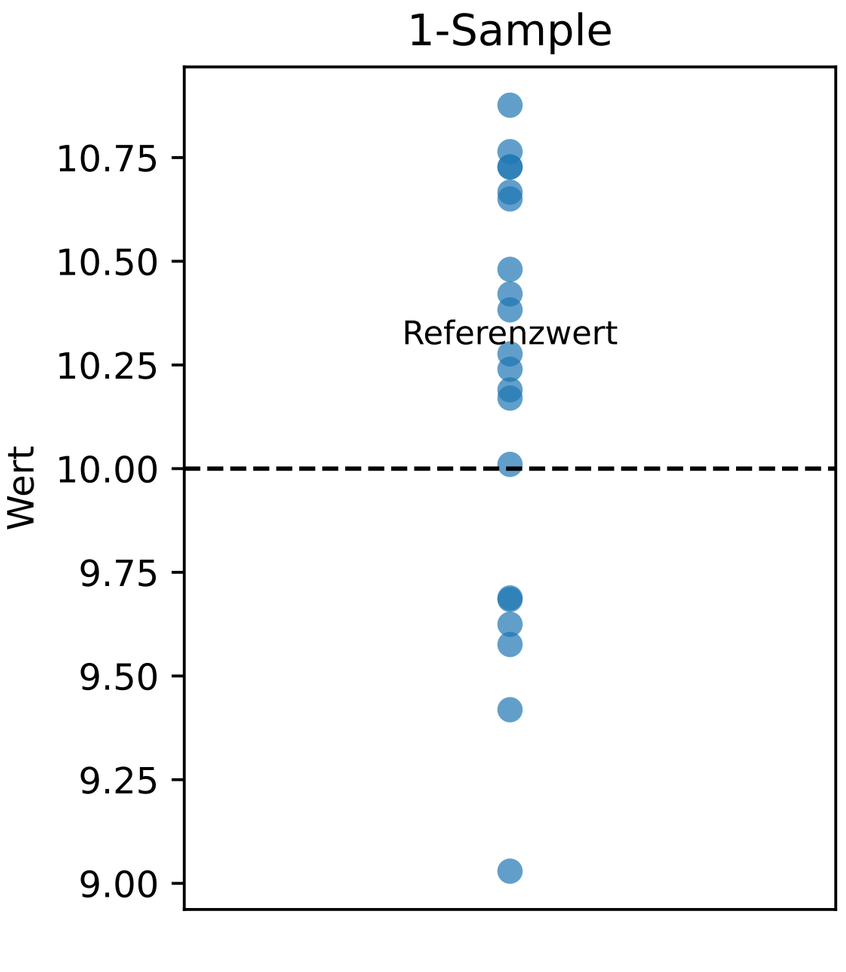
Take-away: p-Wert

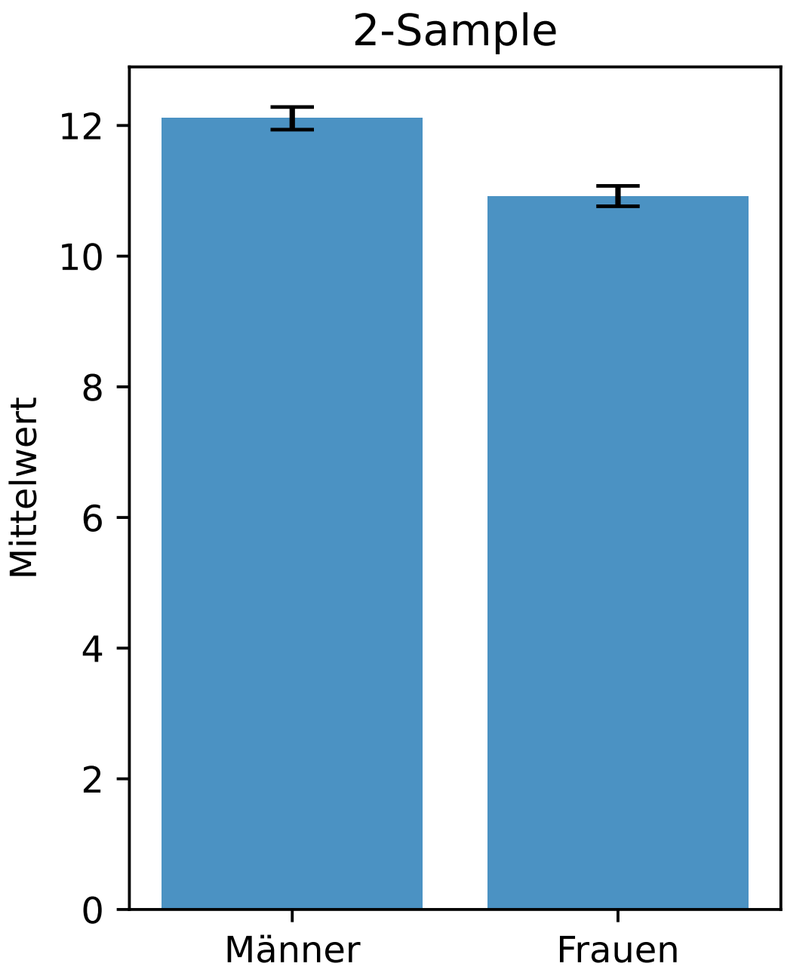
- p-Wert: $P(T \geq t_{\text{obs}} \mid H_0)$: Seltenheit der Daten unter H_0.
- Kein Maß dafür, dass H_0 falsch/wahr ist (kein posterior).
- Kleine p → Daten sind unter H_0 ungewöhnlich;
  grosse p → typisch unter H_0.
- Berichte immer p + Konfidenzintervall + Effektgrösse (z. B. d):
  Kontext entscheidet.

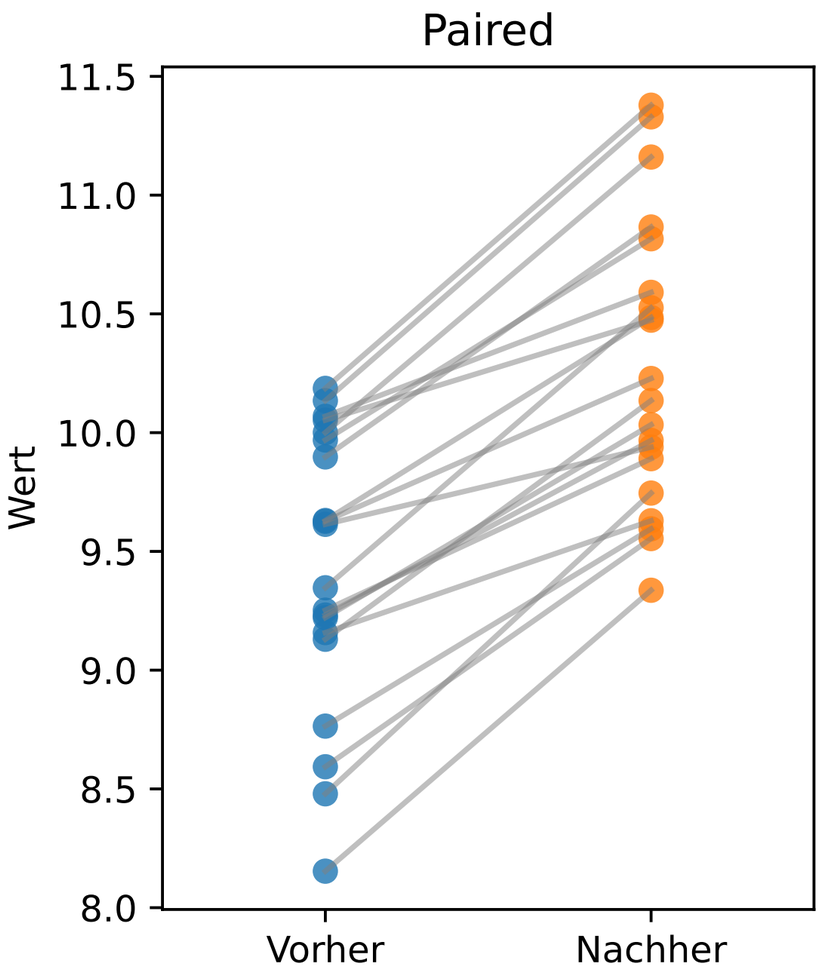
t-Tests

Überblick t-Tests

Testtyp	Anwendung	Beispiel
1-Sample	Mittelwert einer Stichprobe vs. Referenzwert	$\mu = 10$?
2-Sample	Vergleich zweier unabhängiger Gruppen	Männer vs. Frauen
Paired	Vergleich innerhalb von Personen (Vorher–Nachher)	Trainingseffekt


1-Sample
Wert
10.75
10.50
10.25
10.00
9.75
9.50
9.25
9.00
Referenzwert


2-Sample
Mittelwert
12
10
8
6
4
2
0
Männer
Frauen


Paired
Wert
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
Vorher
Nachher

Beispiel: Zwei-Stichproben-t-Test

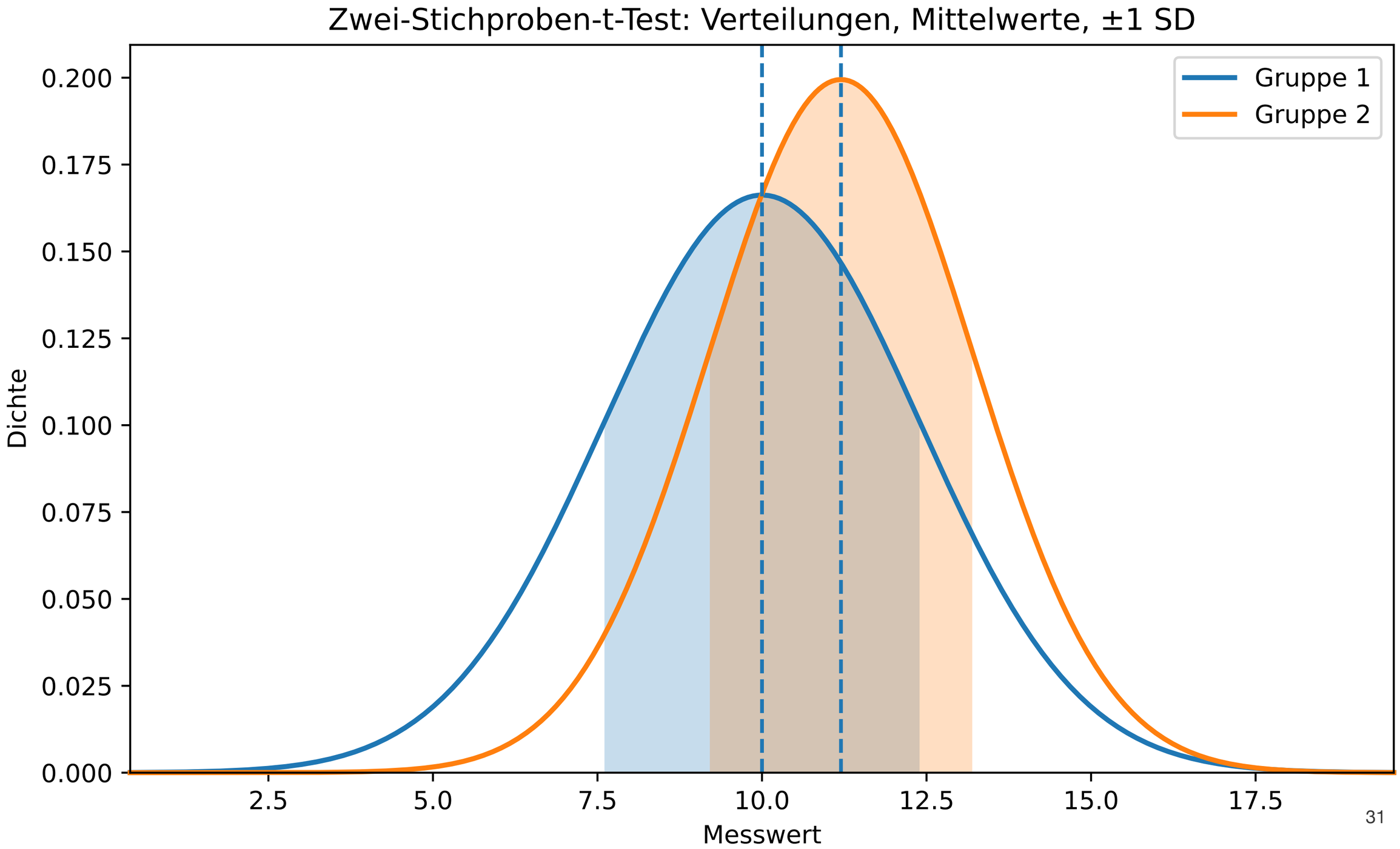
- Prüft, ob sich zwei Gruppen im Mittelwert unterscheiden.
- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

Teststatistik:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

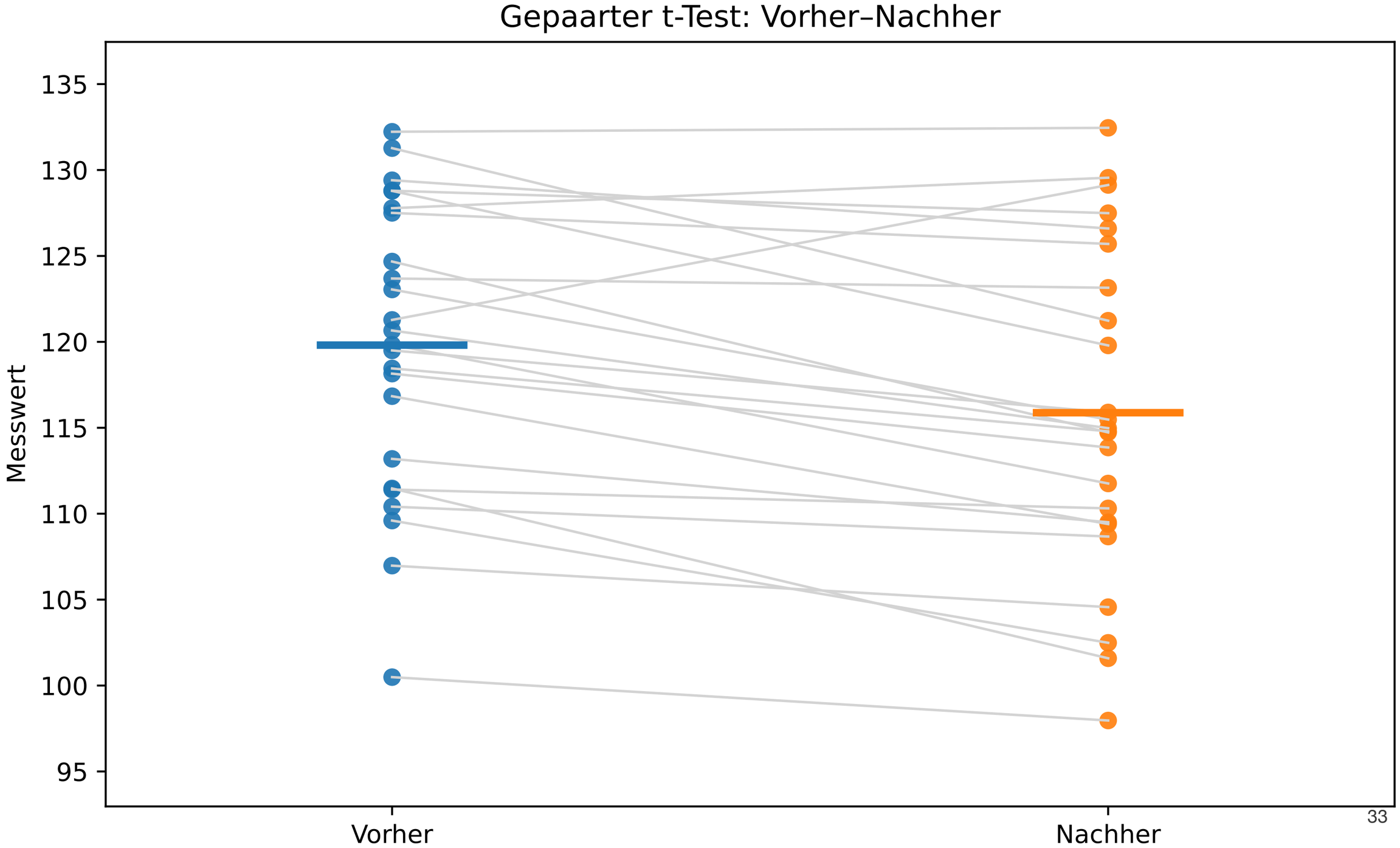
mit

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$


Zwei-Stichproben-t-Test: Verteilungen, Mittelwerte, ±1 SD
Gruppe 1
Gruppe 2
Dichte
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Messwert

Gepaarter t-Test

- Für Vorher–Nachher-Daten oder gepaarte Beobachtungen.
- Testet, ob der Mittelwert der Differenzen $\bar{d} \neq 0$ ist.
- H_0: $\mu_d = 0$
- Teststatistik: $t = \dfrac{\bar{d}}{s_d / \sqrt{n}}$


Gepaarter t-Test: Vorher–Nachher
Messwert
135
130
125
120
115
110
105
100
95
Vorher
Nachher

Interpretation von t-Tests

Berichte immer:

- Mittelwerte und Standardabweichungen.
- Freiheitsgrade, t-Wert, p-Wert.
- Effektgrösse (z. B. Cohen’s d) und Konfidenzintervall.

Beispiel-Bericht:

t(58) = 2.13, p = 0.037, d = 0.55 → mittlerer Effekt.

Ergebnis ≠ Schlussfolgerungw → Kontext zählt.

Take-away: t-Tests

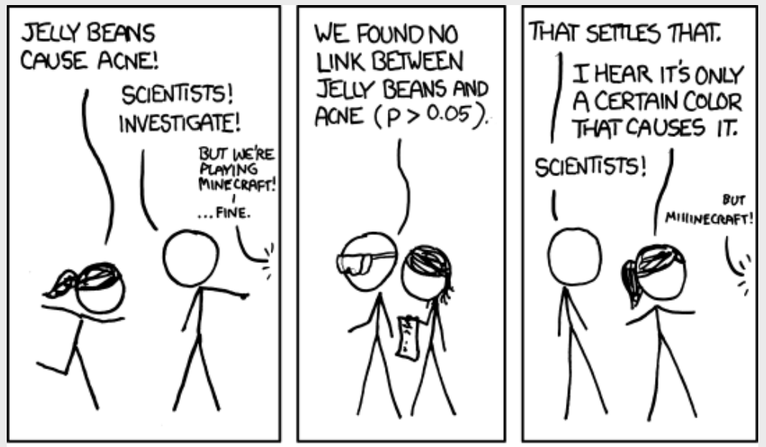
- 1-Sample: $\bar{x}$ vs. μ_0. (eine Gruppe, Referenzwert).
- 2-Sample: zwei unabhängige Gruppen.
- Paired: Vorher–Nachher innerhalb derselben Personen.
- Ziel: prüfen, ob Mittelwertsunterschiede Zufall sind.
- Immer p + CI + Effektgrösse berichten.

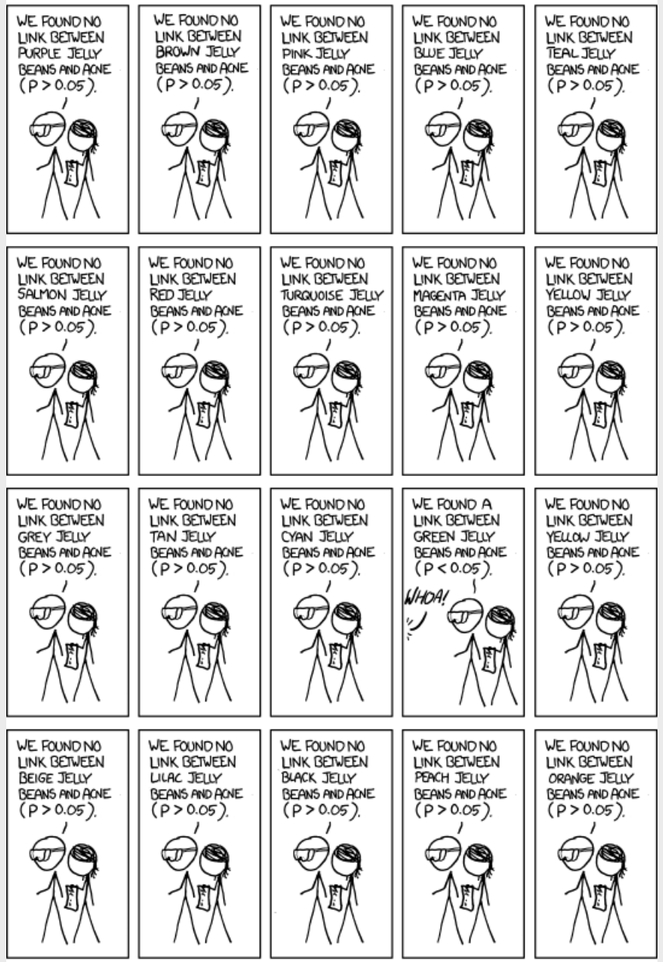
t-Tests zeigen, ob Mittelwertsunterschiede plausibel sind $\rightarrow$ aber erst Kontext und Effektgrösse machen sie bedeutsam.

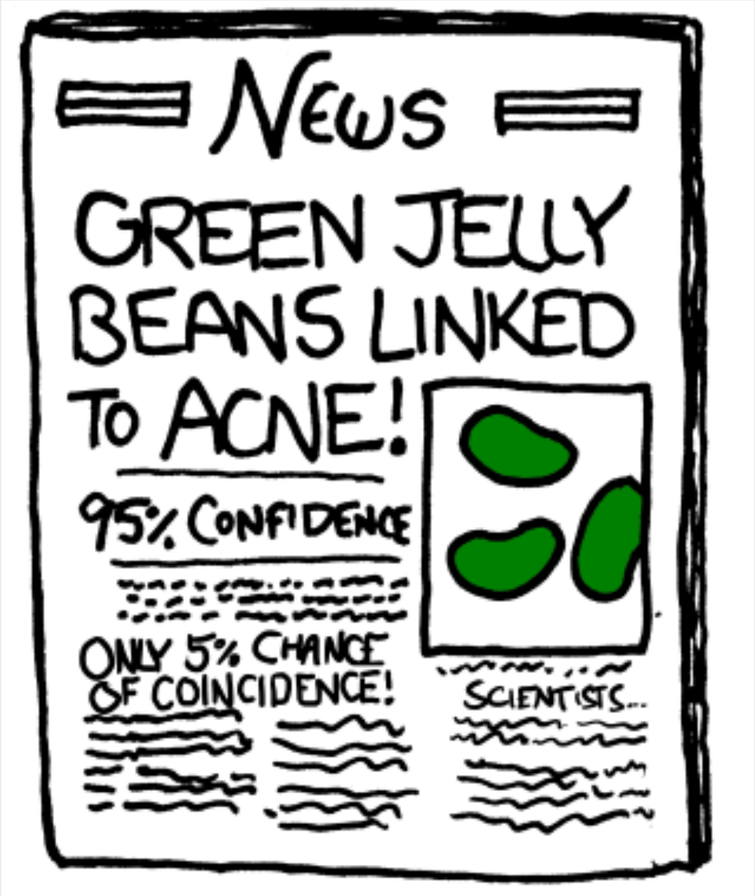
Praxis & Kritik

p-Hacking: Wenn man das «p» jagt

- Definition: Wiederholtes Testen oder Subgruppen-Analysen, bis $p < 0.05$ erreicht ist.
- Ziel: ein „signifikantes“ Ergebnis finden, obwohl H_0 wahr ist.
- Beispiel-Kaskade: «Äpfel & Akne»
  - Test 1: alle Äpfel → p = 0.45
  - Test 2: nur grüne → p = 0.08
  - Test 3: nur Männer → p = 0.20
  - …
  - Test 20: gelbe Äpfel bei Teenagern → p = 0.04 → scheinbar signifikant!
- Problem: Nur der letzte Test wird berichtet → Falsch-Positiver Befund.







xkcd Comic: Significant

p-Hacking & Relevanz

- p-Hacking: wiederholtes Testen, bis $p < 0.05$.
- Problem: erhöht massiv die Wahrscheinlichkeit für Fehlalarme (α-Inflation).
- Immer dokumentieren, wie viele Tests durchgeführt wurden.

Signifikanz vs. Relevanz

- Statistisch signifikant ≠ praktisch bedeutsam.
- Beispiel: Puls sinkt um 0.3 bpm, p = 0.001 → irrelevant im Alltag.
- Deshalb immer:
  - Effektgrösse (z. B. Cohen’s d).
  - Konfidenzintervall.
  - Kontext und Bedeutung.

Signifikanz sagt, dass etwas auffällt → nicht, dass es wichtig ist.

Best Practices

- Hypothesen vorab registrieren (Pre-Registration).
- Effektgrössen & Konfidenzintervalle immer berichten.
- Bei vielen Tests: Bonferroni oder FDR (Statcoach) anwenden.
- Kontext erläutern: Statistik liefert Evidenz, kein Urteil.
- Ziel: Reproduzierbare Forschung, nicht überraschende p-Werte.

Take-away: Praxis & Kritik

- p-Hacking: viele Tests → mehr Zufallstreffer.
- Signifikanz ≠ Relevanz: immer Effektgrösse & Kontext prüfen.
- Best Practice: Pre-Registration, CI, Transparenz.

Klar planen, sauber berichten, ehrlich interpretieren.

Reflexion & Take-aways

Was wir heute gelernt haben

- Zweck von Hypothesentests: Zufall von echtem Effekt trennen.
- Kernkonzepte: H_0, H_1, α, β, Power, p-Wert.
- t-Test: prüft Mittelwertsunterschiede – Beispiel für angewandte Inferenz.
- p-Wert: misst Seltenheit unter H_0, kein Wahrheitsmass.
- Kritik & Praxis: Signifikanz ≠ Relevanz, p-Hacking vermeiden, Transparenz fördern.

Transferfrage: Interpretation üben

Eine Studie findet $p = 0.02$ und eine Effektgrösse $d = 0.1$.

Was bedeutet das?

- Statistisch signifikant (selten unter H_0).
- Aber praktisch irrelevant (winziger Effekt).

→ Beispiel für „signifikant, aber nicht bedeutsam“.

Quiz: Aktive Wiederholung

Kahoot Quiz VL7: Hypothesentests

Nächste Woche: Hypothesentests II

- Erweiterung auf komplexere Designs
  - A/B-Tests
  - Permutationstests
  - ANOVA
- Fokus auf Effektgrössen, Power-Analyse und Multiple Testing.

👉 Vorbereitung: PSDS Kapitel 4