Die t-Verteilung, der t-Test - Modell und Realität

Reimund Vehling

15. Januar 2026

Modell und Realität (messerscharf trennen)

Realität: Wir messen endlich viele Patient:innen. Es gibt biologische Variabilität, Messfehler, Bias und Confounding (Verzerrung, Scheinkorrelation). Der wahre Parameter (z. B. der wahre Mittelwert) ist *fix*, aber *unbekannt*.

Modell: Für statistische Aussagen tun wir so, als entstammen die Daten einem vereinfachten Zufallsmodell (z. B. Normal-/t-Modell, unabhängige Stichprobe). *Wahrscheinlichkeiten existieren nur im Modell.*

Was ein t-Test ist (ohne Rezeptdenken)

Ein (zweiseitiger) t-Test prüft, ob die beobachteten Daten *unter Annahme von* H_0 (*im Modell*) so ungewöhnlich sind, dass sie schlecht zu H_0 passen.

Der t-Wert als Signal-zu-Rauschen

$$t \;=\; \frac{\text{Unterschied}}{\text{typische Zufallsschwankung}} \;=\; \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

Der Zähler misst die Abweichung vom Nullwert, der Nenner (Standardfehler) die typische Streuung des Mittelwerts durch Zufall. Der Test bewertet *Daten*, nicht *Wahrheit*.

p-Wert (was er ist, und was nicht)

Der p-Wert ist die Wahrscheinlichkeit, *mindestens so extreme Daten* zu beobachten, *wenn* H_0 *wahr wäre* (im Modell). Er ist *nicht* die Wahrscheinlichkeit, dass H_0 wahr ist.

Konfidenzintervall

Ein 95%-Konfidenzintervall ist die Menge der Parameterwerte, die mit den Daten *im Modell* auf diesem Sicherheitsniveau vereinbar sind. Beim zweiseitigen Test gilt:

$$H_0 \text{ wird auf Niveau } \alpha = 0{,}05 \text{ verworfen} \iff \text{Nullwert liegt nicht im 95\%-KI.}$$

p liegt im Intervall ist falsch: Das Intervall enthält Parameterwerte, nicht p-Werte.

Mini-Beispiel (Therapie vs. Kontrolle)

Eine Studie vergleicht systolischen Blutdruck nach 4 Wochen zwischen Therapie (T) und Kontrolle (K). Wir betrachten den Mittelwertunterschied $\Delta = \mu_T - \mu_K$.

Beobachtung (Beispielzahlen):

$$\bar{x}_T = 128,\ s_T = 12,\ n_T = 25 \qquad \bar{x}_K = 134,\ s_K = 14,\ n_K = 25$$

Damit ist der beobachtete Unterschied:

$$\hat{\Delta} = \bar{x}_T - \bar{x}_K = -6 \text{ mmHg}.$$

Interpretation ohne Rechner:

- Der Test fragt: *Wäre ein Unterschied von -6 mmHg unter $H_0 : \Delta = 0$ im Modell ungewöhnlich?*
- Das 95%-KI fragt: *Welche Δ sind mit den Daten noch vereinbar?*

Typische Ergebnis-Formulierung (wenn z. B. berechnet):

$$95\%\text{-KI für } \Delta : (-11,\ -1) \text{ mmHg}$$

Dann gilt:

- 0 liegt nicht im Intervall $\Rightarrow$ statistisch signifikant (zweiseitig, $\alpha = 0{,}05$).
- Medizinische Frage bleibt: *Ist -6 mmHg klinisch relevant?*

t-Test, p-Wert, Konfidenzintervall - kurz und kompakt

1. Modell vs. Realität

- Realität: endliche Daten, Bias/Confounding/Messfehler. Parameter sind fix, unbekannt.
- Modell: vereinfachtes Zufallsmodell. Wahrscheinlichkeiten gelten *nur im Modell.*

2. t-Test (zweiseitig als Standard)

Frage: Sind die Daten unter H_0 ungewöhnlich?

$$t = \frac{\text{Unterschied}}{\text{Standardfehler}}$$

Merke: Der Test bewertet Daten im Modell, nicht die Wahrheit von H_0.

3. p-Wert

Definition: Wahrscheinlichkeit (im Modell), mindestens so extreme Daten zu sehen, *wenn H_0 gilt.*
Nicht: Wahrscheinlichkeit, dass H_0 wahr ist.
Nicht: klinische Relevanz.

4. Konfidenzintervall (95%)

Idee: Bereich plausibler Parameterwerte (im Modell).
Zweiseitig:

$$p < 0{,}05 \Longleftrightarrow 0 \notin 95\%\text{-KI}$$

Achtung: "p liegt im Intervall" ist falsch.

5. Drei typische Denkfehler

- "Nicht signifikant" $\neq$ "kein Effekt"
- kleiner p-Wert $\neq$ großer Effekt
- statistisch signifikant $\neq$ klinisch relevant

MC-Training (t-Test, p-Wert, KI)

Fragen

1. Ein zweiseitiger t-Test liefert $p = 0{,}12$ bei $\alpha = 0{,}05$. Welche Aussage ist korrekt?
   (a) H_0 ist mit 88% Wahrscheinlichkeit wahr.
   (b) Es gibt sicher keinen Unterschied.
   (c) Die Daten sind unter H_0 nicht ungewöhnlich.
   (d) Das Ergebnis ist klinisch irrelevant.
   (e) Der wahre Parameter liegt mit 95% Wahrscheinlichkeit im 95%-KI.

2. Ein 95%-KI für den Mittelwertunterschied ist $(-11, -1)$ mmHg. Welche Aussage ist korrekt?
   (a) Der p-Wert liegt im Intervall.
   (b) Der Nullwert 0 ist mit den Daten auf 5%-Niveau nicht vereinbar (zweiseitig).
   (c) Der wahre Unterschied liegt mit 95% Wahrscheinlichkeit im Intervall.
   (d) Das Ergebnis ist klinisch relevant.
   (e) H_0 ist bewiesen falsch.
3. Was bedeutet "nicht signifikant" am ehesten?
   (a) Es gibt keinen Effekt.
   (b) Es fehlt Evidenz gegen H_0 im gewählten Modell.
   (c) Der Effekt ist klinisch klein.
   (d) Messfehler sind ausgeschlossen.
   (e) Die Stichprobe ist repräsentativ.
4. Welche Aussage trifft auf den p-Wert zu?
   (a) Er ist die Wahrscheinlichkeit, dass H_0 wahr ist.
   (b) Er misst die Effektstärke.
   (c) Er ist eine Aussage über Daten unter Annahmen.
   (d) Er ist unabhängig von der Stichprobengröße.
   (e) Er ist eine Aussage über einzelne Patient:innen.
5. Zwei Studien finden denselben Mittelwertunterschied, aber Studie A hat $n = 30$, Studie B $n = 300$. Welche Aussage ist am ehesten richtig?
   (a) B hat meist ein engeres KI.
   (b) A hat immer den kleineren p-Wert.
   (c) A ist klinisch relevanter.
   (d) B verletzt das Modell stärker.
   (e) Das KI hat nichts mit n zu tun.
6. Welche Formulierung ist *korrekt*?
   (a) "Mit 95% Wahrscheinlichkeit liegt der wahre Wert im KI."
   (b) "Das KI enthält alle plausiblen Parameterwerte im Modell (95%-Niveau)."
   (c) "p=0,03 beweist, dass H_0 falsch ist."
   (d) "Nicht signifikant bedeutet kein Unterschied."
   (e) "Signifikant bedeutet klinisch wichtig."

Lösungen (mit Kurzbegründung)

1. **c)** $p = 0{,}12$ bedeutet: Daten sind unter H_0 nicht besonders ungewöhnlich (im Modell).
2. **b)** $0 \notin$ 95%-KI $\Leftrightarrow p < 0{,}05$ (zweiseitig, gleiches Niveau).
3. **b)** Nicht signifikant = keine ausreichende Evidenz gegen H_0 im Modell.
4. **c)** p-Wert ist eine Aussage über Daten unter Annahmen, nicht über Wahrheit/Effektstärke.
5. **a)** Größeres $n \Rightarrow$ kleinerer Standardfehler $\Rightarrow$ typischerweise engeres KI.
6. **b)** Das ist die saubere Modell-Interpretation.

Einordnung: p-Werte, Tests und das ASA/Wilkinson-Narrativ

Moderne statistische Leitlinien (z. B. ASA Statement zu p-Werten) betonen, dass Hypothesentests und p-Werte

- keine Wahrheitswahrscheinlichkeiten liefern,
- keine Effektstärken messen und
- nicht ohne Kontext interpretiert werden dürfen.

Ein Hypothesentest bewertet ausschließlich, wie gut beobachtete Daten zu einem idealisierten Modell passen. Die Entscheidung *signifikant / nicht signifikant* ist eine konventionelle Schwelle, keine natürliche Grenze.

Konfidenzintervalle ergänzen Hypothesentests, indem sie die Größenordnung und Unsicherheit des Effekts sichtbar machen. Beide beruhen auf demselben statistischen Modell.

SPSS-Outputs lesen: vom Output zum Modell

Statistiksoftware (z. B. SPSS) führt keine statistischen Entscheidungen aus. Sie berechnet Kennzahlen zu einem impliziten Modell. Die Interpretation liegt vollständig bei der Anwenderin.

Grundprinzip

Jeder SPSS-Output beantwortet drei Fragen:

1. Welcher Parameter wird geschätzt?
2. Wie groß ist die Unsicherheit dieser Schätzung?
3. Sind die beobachteten Daten mit H_0 vereinbar?

Beispiel: t-Test (kontinuierlicher Endpunkt)

Typische SPSS-Ausgaben:

- Mean
- Std. Deviation
- Std. Error Mean
- t, df
- Sig. (2-tailed)
- 95% Confidence Interval

Interpretation:

- Mean: Punktschätzer des Parameters
- Std. Error: typische Zufallsschwankung
- t: Signal-zu-Rauschen-Verhältnis
- Sig. (2-tailed): p-Wert (Kompatibilität der Daten mit H_0)
- 95%-KI: Bereich plausibler Parameterwerte (im Modell)

Alle Größen enthalten dieselbe Information in unterschiedlicher Form.

Mini-Beispiel: kontinuierlicher Endpunkt

Unterschied im mittleren systolischen Blutdruck zwischen Therapie (T) und Kontrolle (K).

$$\hat{\Delta} = \bar{x}_T - \bar{x}_K = -6 \text{ mmHg}, \qquad 95\%\text{-KI} = (-11, -1)$$

Interpretation:

- Der Test fragt: Wäre $\hat{\Delta} = -6$ unter $H_0 : \Delta = 0$ ungewöhnlich?
- Das KI zeigt: Welche Effekte sind mit den Daten vereinbar?
- $0 \notin KI \Rightarrow$ statistisch signifikant (zweiseitig).

Mini-Beispiel: binärer Endpunkt

Anteil geheilter Patient:innen:

$$\hat{p}_T = 0.62, \qquad \hat{p}_K = 0.48$$

Ein 95%-KI für den Unterschied der Anteile enthält den Wert 0.
Interpretation:

- Die Daten sind mit $H_0 : p_T = p_K$ vereinbar.
- Nicht signifikant bedeutet: keine ausreichende Evidenz gegen H_0.
- Es bedeutet nicht: kein Effekt.

Mini-Beispiel: Überlebenszeit

Zwei Kaplan–Meier-Kurven werden verglichen. Der Log-Rank-Test ergibt $p = 0.08$.
Interpretation:

- Hypothesentest: Sind diese Kurven unter H_0 ungewöhnlich?
- Ergebnis: Daten sind mit gleicher Überlebensverteilung vereinbar.
- Die Kurven selbst liefern die inhaltliche Information.

Zehn Aussagen, die man bei SPSS-Outputs nicht machen sollte

1. Der p-Wert ist die Wahrscheinlichkeit, dass H_0 wahr ist.
2. Nicht signifikant bedeutet: kein Effekt.
3. Signifikant bedeutet: klinisch relevant.
4. Der Test hat H_0 bewiesen.
5. Das Ergebnis ist zufällig.
6. Der wahre Wert liegt mit 95% Wahrscheinlichkeit im KI.
7. p liegt im Konfidenzintervall.
8. Ein kleiner p-Wert bedeutet einen großen Effekt.
9. Statistik entscheidet über Wahrheit.
10. Das Programm hat getestet.

MC-Training: richtig denken

Frage

Ein zweiseitiger Test ergibt $p = 0.12$.
Welche Aussage ist korrekt?

1. H_0 ist wahrscheinlich wahr.
2. Es gibt keinen Effekt.
3. Die Daten sind unter H_0 nicht ungewöhnlich.
4. Der Effekt ist klinisch irrelevant.
5. Das Ergebnis ist zufällig.

Lösung

Antwort: c)
Begründung: Der p-Wert beschreibt die Kompatibilität der Daten mit H_0 im Modell. Er ist keine Aussage über Wahrheit, Effektgröße oder Relevanz.

Ein Satz, der immer korrekt ist

Statistische Tests bewerten Daten unter Annahmen. Sie entscheiden nicht über Wahrheit.

Rückübersetzung: Vom Output zum Modell

Ziel: Statistische Software (z. B. SPSS) zeigt numerische Ergebnisse. Die statistische Bedeutung entsteht erst durch die Rückübersetzung in das zugrunde liegende Modell.

Grundregel

Statistikprogramme rechnen, sie interpretieren nicht. Der Output enthält Ergebnisse eines Modells, das im Screenshot selbst nicht sichtbar ist. Dieses Modell muss gedanklich rekonstruiert werden.

Die fünf Schritte der Rückübersetzung

Die folgenden fünf Fragen sind *immer* zu beantworten, unabhängig vom Testtyp (t-Test, Chi^2, Fisher, Log-Rank).

1. **Was ist der Parameter?**
   Welche Größe wird geschätzt? (z. B. Mittelwert, Mittelwertdifferenz, Anteil, Hazard Ratio)

2. **Was ist die Nullhypothese H_0?**
   Welcher Parameterwert wird angenommen? (Meist: kein Effekt, Differenz $= 0$, Verhältnis $= 1$)

3. **Wo steht die Schätzung?**
   Was ist der durch die Daten geschätzte Wert? (z. B. Mean, Mean Difference, Estimate)

4. **Wo ist die Unsicherheit?**
   Wie stark kann diese Schätzung aufgrund von Zufall schwanken? (z. B. Standardfehler, Konfidenzintervall)

5. **Ist der Nullwert noch plausibel?**
   Liegt der Nullwert innerhalb des Konfidenzintervalls?

   - ja $\Rightarrow$ nicht signifikant
   - nein $\Rightarrow$ signifikant (zweiseitig)

Wichtige Reihenfolge

$$\textit{Parameter} \rightarrow \textit{Schätzung} \rightarrow \textit{Unsicherheit} \rightarrow \textit{Nullwert} \rightarrow \textit{Entscheidung}$$

Nicht umgekehrt.

Beispielhafte Rückübersetzung

Angenommen, ein Output zeigt:

Mean Difference $= -6$
95% Confidence Interval $= (-11, -1)$

Dann gilt:

- Parameter: Mittelwertdifferenz
- H_0: Differenz $= 0$
- Schätzung: -6
- Unsicherheit: plausible Werte zwischen -11 und -1
- Nullwert nicht im Intervall $\Rightarrow$ signifikant

Der p-Wert enthält keine zusätzliche Information, sondern bestätigt diese Entscheidung lediglich.

Merksätze

- Kein Test ohne Parameter.
- Eine Zahl ohne Unsicherheit ist unvollständig.
- Der Test bewertet Daten im Modell, nicht Wahrheit.
- Das Konfidenzintervall zeigt, was plausibel ist.

Checkliste: Vom Output zum Modell

Ziel: Jeden statistischen Output (z. B. aus SPSS) gedanklich in sein statistisches Modell zurückübersetzen.

1. **Parameter identifizieren**
   Was wird geschätzt? *(Mittelwert, Mittelwertdifferenz, Anteil, Hazard Ratio, ...)*
2. **Nullhypothese formulieren**
   Welcher Parameterwert wird angenommen? *(meist: kein Effekt, Differenz = 0, Verhältnis = 1)*
3. **Schätzung finden**
   Was ist der beobachtete Wert aus den Daten? *(Mean, Mean Difference, Estimate)*
4. **Unsicherheit bestimmen**
   Wie stark kann diese Schätzung zufällig schwanken? *(Standardfehler, Konfidenzintervall)*
5. **Nullwert prüfen**
   Liegt der Nullwert im Konfidenzintervall?
   - ja ⇒ nicht signifikant
   - nein ⇒ signifikant (zweiseitig)

Wichtige Reihenfolge:

$$\text{Parameter} \rightarrow \text{Schätzung} \rightarrow \text{Unsicherheit} \rightarrow \text{Nullwert} \rightarrow \text{Entscheidung}$$

Merksätze:

- Kein Test ohne Parameter.
- Eine Zahl ohne Unsicherheit ist unvollständig.
- Der p-Wert bestätigt nur, was das Intervall zeigt.

Anhang: Die t-Verteilung

Die **t-Verteilung** (Student-t-Verteilung) ist eine stetige Wahrscheinlichkeitsverteilung, die immer dann verwendet wird, wenn Aussagen über einen Mittelwert gemacht werden sollen, **aber die Populationsstandardabweichung unbekannt ist**.

In der Praxis ist dies fast immer der Fall, da man meist nur eine Stichprobe und nicht die gesamte Grundgesamtheit beobachtet.

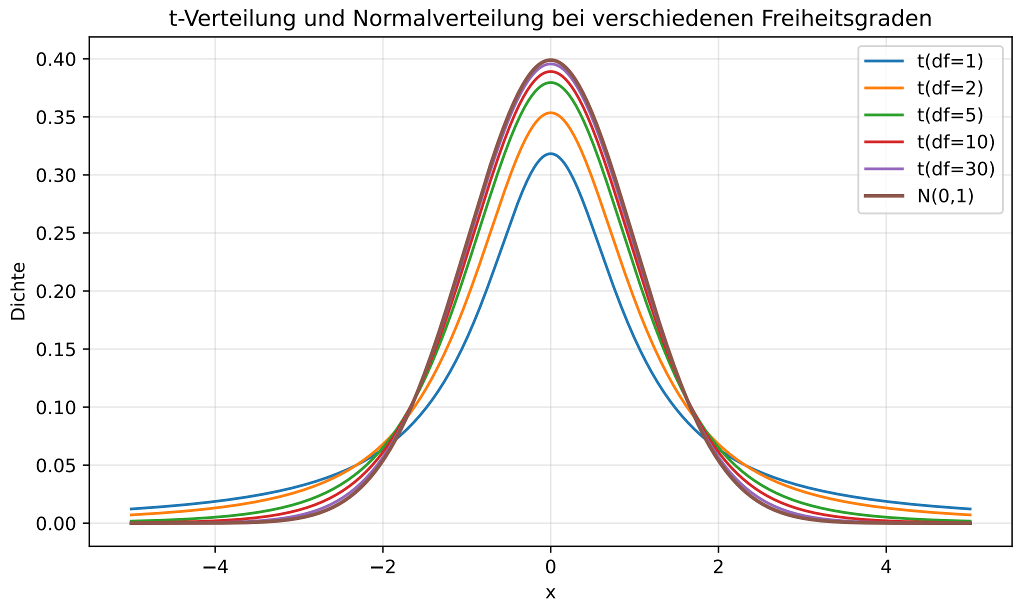

Abbildung 1: Vergleich der t-Verteilungen für verschiedene Freiheitsgrade mit der Standardnormalverteilung. Mit wachsender Stichprobengröße nähert sich die t-Verteilung der Normalverteilung an.

Der t-Wert: Definition und Intuition

Der t-Wert misst, **wie viele Standardfehler** der beobachtete Stichprobenmittelwert vom unter der Nullhypothese angenommenen Mittelwert entfernt ist.

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

Interpretation der Bestandteile

- $\bar{x} - \mu_0$: beobachtete Abweichung vom Nullhypothesenwert
- s: Streuung der Daten
- $\sqrt{n}$: Mittelwerte streuen weniger als Einzelwerte

Der Nenner $s/\sqrt{n}$ ist der sogenannte **Standardfehler des Mittelwerts**. Er beschreibt die typische Schwankung des Stichprobenmittelwerts.

Warum genau diese Form?

- große Abweichung ⇒ großer t-Wert
- große Streuung ⇒ kleinerer t-Wert
- große Stichprobe ⇒ kleinerer Standardfehler

Der t-Wert ist damit eine **standardisierte Effektgröße**.

Wichtige Eigenschaften

- symmetrisch um 0
- glockenförmig ähnlich zur Normalverteilung
- besitzt **schwerere Ränder (heavy tails)**
- hängt von der Anzahl der **Freiheitsgrade** ab

Die Freiheitsgrade ergeben sich meist aus der Stichprobengröße, z. B.

$$df = n - 1$$

Unterschied zwischen t-Verteilung und Normalverteilung

Normalverteilung

Die Normalverteilung wird verwendet, wenn

- der Populationsmittelwert μ bekannt oder zu prüfen ist und
- die Populationsstandardabweichung σ **bekannt** ist.

Die standardisierte Teststatistik folgt dann einer Standardnormalverteilung:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

t-Verteilung

In der Realität ist σ fast nie bekannt und wird durch die Stichprobenstandardabweichung s geschätzt. Dadurch entsteht zusätzliche Unsicherheit.

Die Teststatistik lautet:

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

Da s zufällig ist, folgt diese Statistik **nicht mehr exakt der Normalverteilung**, sondern einer t-Verteilung.

Vergleich

	Normalverteilung	t-Verteilung
Varianz	bekannt (σ)	geschätzt (s)
Ränder	schmal	breiter
Abhängigkeit von n	nein	ja (Freiheitsgrade)
Grenzfall $n \to \infty$	—	Normalverteilung

Mit wachsender Stichprobengröße nähert sich die t-Verteilung der Normalverteilung an.

Der Welch-t-Test

Warum der Welch-Test?

Beim Vergleich zweier Mittelwerte ist die Annahme gleicher Varianzen oft nicht realistisch, insbesondere bei medizinischen Daten. Der Welch-t-Test verzichtet auf diese Annahme und ist daher robuster als der klassische Zwei-Stichproben-t-Test.

Wichtig: Der Welch-Test testet dieselbe Nullhypothese wie der klassische t-Test, verwendet aber ein flexibleres Modell für die Streuung.

Nullhypothese

$$H_0: \ \mu_1 - \mu_2 = 0$$

Es wird geprüft, ob der beobachtete Unterschied der Mittelwerte unter dieser Annahme im Modell ungewöhnlich ist.

Teststatistik (Welch)

Die Teststatistik hat die bekannte Form

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

Interpretation:

- Zähler: beobachteter Unterschied der Mittelwerte
- Nenner: typische Zufallsschwankung dieses Unterschieds
- t misst ein *Signal-zu-Rauschen-Verhältnis*

Die Struktur ist identisch zum klassischen t-Test; lediglich die Modellierung der Streuung ist allgemeiner.

Freiheitsgrade (Welch–Satterthwaite)

Da die Varianzen getrennt geschätzt werden, sind die Freiheitsgrade nicht mehr ganzzahlig. Sie werden durch die Welch–Satterthwaite-Approximation gegeben:

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

Bemerkung: Diese Formel muss nicht hergeleitet werden. Sie spiegelt wider, dass die Unsicherheit der Varianzschätzungen in die Testverteilung eingeht.

Was bleibt gleich?

- Die Teststatistik folgt (approximativ) einer t-Verteilung.
- Es gibt einen beobachteten t_{obs}.
- Es gibt kritische Werte $\pm t_{\text{krit}}$.
- Die Entscheidungsstruktur ist unverändert:

$$|t_{\text{obs}}| > t_{\text{krit}} \iff \text{Ablehnung von } H_0$$

Beispiel: Der Welch-t-Test (zwei unabhängige Gruppen) - mit SPSS-Ausgabe

Diese Simulation und Ausgabe wurde mit dem Programm Python erzeugt.
Wir simulieren systolische Blutdruckdaten (mmHg) für zwei unabhängige Gruppen (Therapie vs. Kontrolle) und testen die Nullhypothese

$$H_0 : \ \mu_{\text{Therapie}} - \mu_{\text{Kontrolle}} = 0$$

mit dem **Welch-t-Test**, der keine Varianzgleichheit voraussetzt. Die Teststatistik

$$t = \frac{\bar{x}_T - \bar{x}_K}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_K^2}{n_K}}}$$

misst den beobachteten Mittelwertunterschied relativ zu seiner zufälligen Streuung. Die (nicht-ganzzahligen) Freiheitsgrade werden über die Welch–Satterthwaite-Approximation bestimmt; Entscheidungsstruktur und Interpretation bleiben identisch zum klassischen t-Test.

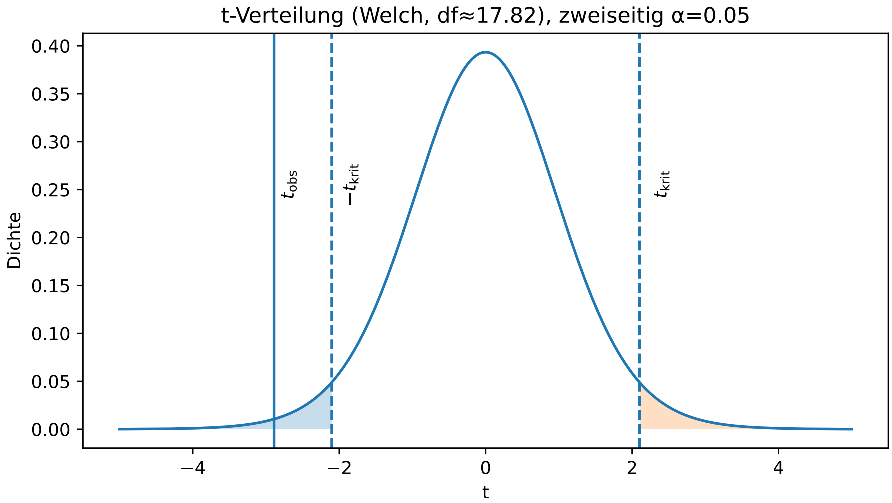

Abbildung 2: Ausgabe der t-Verteilung mit t_{krit} und t_{obs}

T-Test (Unabhängige Stichproben) — SPSS-Style

Endpunkt: Systolischer Blutdruck (mmHg), Therapie vs. Kontrolle

Group Statistics

Group	N	Mean	Std. Deviation	Std. Error Mean
Therapy	10	123.97	11.2	3.54
Control	10	137.78	10.13	3.2

Independent Samples Test (Welch)

t	df	Sig. (2-tailed)	Mean Diff (T-K)	95% CI Lower	95% CI Upper
-2.89	17.82	0.01	-13.8	-23.84	-3.76

Abbildung 3: Ausgabe der Ergebnisse in SPSS-Darstellung

```python
# ----------------------------
# 1) Beispiel: Blutdruckdaten
# ----------------------------
rng = np.random.default_rng(42)

n = 10
therapy = rng.normal(128, 12, size=n)   # Therapiegruppe
control = rng.normal(134, 14, size=n)   # Kontrollgruppe

# Welch t-test
t_stat, p_val = stats.ttest_ind(therapy, control, equal_var=False)

n1, n2 = len(therapy), len(control)
m1, m2 = therapy.mean(), control.mean()
s1, s2 = therapy.std(ddof=1), control.std(ddof=1)
v1, v2 = s1**2, s2**2

# Welch df
df = (v1/n1 + v2/n2)**2 / ((v1/n1)**2/(n1-1) + (v2/n2)**2/(n2-1))

diff = m1 - m2
se = math.sqrt(v1/n1 + v2/n2)
tcrit = stats.t.ppf(0.975, df)
ci = (diff - tcrit*se, diff + tcrit*se)
```

Abbildung 4: Berechnungen mit Python

Das sollte man wissen:

- Der Welch-Test ist der Standard, wenn Varianzen unbekannt oder ungleich sind.
- Nicht-ganzzahlige Freiheitsgrade sind kein Problem, sondern ein Feature.
- Die Modellstruktur des Hypothesentests bleibt unverändert.

Durchgerechnetes Zahlenbeispiel

Fragestellung

Eine Maschine soll im Mittel $\mu_0 = 100$ g abfüllen. Es wird überprüft, ob der wahre Mittelwert davon abweicht (zweiseitiger Test).

Stichprobe

$$98,\ 101,\ 99,\ 102,\ 100,\ 97,\ 103,\ 99,\ 100,\ 101$$

Stichprobengröße: $n = 10$

Mittelwert

$$\bar{x} = \frac{1000}{10} = 100$$

Standardabweichung

Quadratsummen der Abweichungen:

$$\sum (x_i - \bar{x})^2 = 30$$

$$s^2 = \frac{30}{9} = 3.33 \quad \Rightarrow \quad s \approx 1.83$$

Teststatistik

$$t = \frac{100 - 100}{1.83/\sqrt{10}} = 0$$

Freiheitsgrade:

$$df = n - 1 = 9$$

p-Wert und Entscheidung

Für $t = 0$ ergibt sich ein zweiseitiger p-Wert von

$$p = 1.0$$

Bei $\alpha = 0.05$ wird die Nullhypothese nicht verworfen.

95%-Konfidenzintervall

Kritischer t-Wert:

$$t_{0.975,9} \approx 2.262$$

$$KI = 100 \pm 2.262 \cdot \frac{1.83}{\sqrt{10}} = (98.7,\ 101.3)$$

Der Sollwert 100 liegt im Konfidenzintervall.

Zusammenfassung

- Die t-Verteilung berücksichtigt Unsicherheit durch geschätzte Varianz
- Der t-Test ist konservativer als der z-Test
- Mit wachsendem n verschwindet der Unterschied zur Normalverteilung
- p-Wert und Konfidenzintervall liefern konsistente Entscheidungen

Zwei-Stichproben-t-Test (unabhängige Stichproben)

Beim Zwei-Stichproben-t-Test wird geprüft, ob sich die Mittelwerte zweier **unabhängiger Gruppen** unterscheiden. Typische Fragestellung: Unterscheidet sich der mittlere Lernerfolg zwischen Methode A und B?

Hypothesen (zweiseitig)

$$H_0: \ \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1: \ \mu_1 \neq \mu_2$$

Seien $\bar{x}_1, s_1, n_1$ Kennwerte der ersten Stichprobe und $\bar{x}_2, s_2, n_2$ der zweiten.

0.1 Welch-t-Test (empfohlen bei ungleichen Varianzen)

Der Welch-t-Test setzt **keine Varianzgleichheit** voraus und ist in der Praxis oft die robuste Standardwahl.

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

Die Freiheitsgrade werden näherungsweise mit der Welch–Satterthwaite-Formel berechnet:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

Klassischer Zwei-Stichproben-t-Test (gepoolte Varianz)

Dieser Test setzt **Varianzgleichheit** voraus ($\sigma_1^2 = \sigma_2^2$). Dann wird eine gemeinsame (gepoolte) Varianz geschätzt:

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

Die Teststatistik lautet:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{mit} \qquad df = n_1 + n_2 - 2$$

Entscheidung und Interpretation

Wie beim Ein-Stichproben-t-Test gilt:

- p-Wert $< \alpha \Rightarrow$ Nullhypothese verwerfen (signifikanter Unterschied)
- p-Wert $\geq \alpha \Rightarrow$ keine ausreichende Evidenz für einen Unterschied

Ein passendes 95%-Konfidenzintervall für $\mu_1 - \mu_2$ ergibt sich durch

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.975,df} \cdot SE$$

wobei $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (Welch) bzw. $SE = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ (gepoolt) ist.

Ein-Stichprobentest

Unter `github.com/RVeh/statistic_medicine` befindet sich das zugehörige Programm. Dort können die Parameter verändert werden.

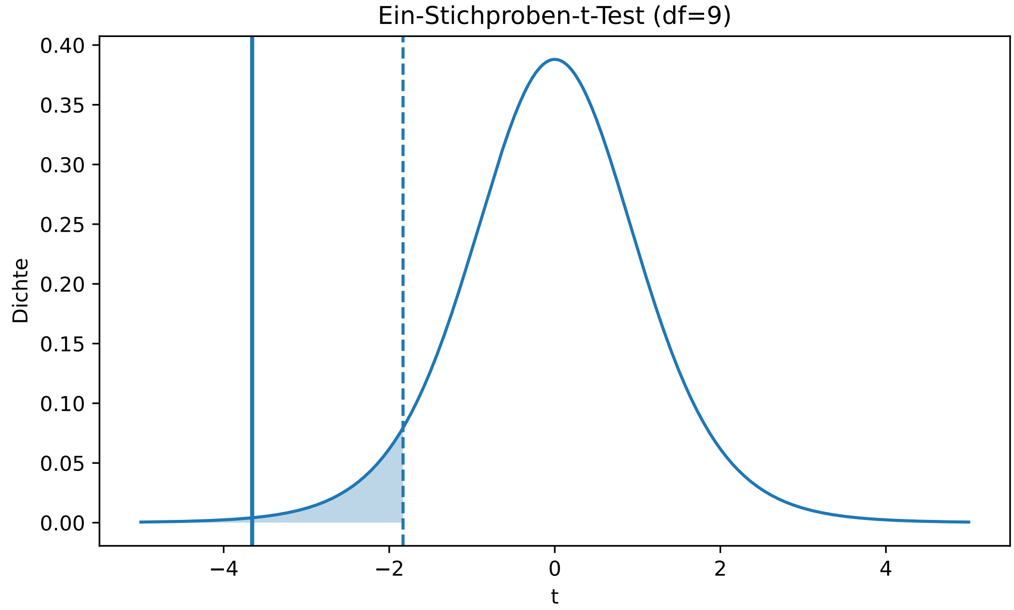

Abbildung 5: Ein-Stichprobentest (linksseitig) mit t_{krit} und t_{obs}

```
Ein-Stichproben-t-Test
----------------------
Beobachtete Daten: [122 125 130 128 127 126 124 129 131 123]
Datenumfang n         = 10
Referenzwert mu0      = 130
Stichprobenmittel     = 126.500
Stichproben-SD        = 3.028

Testseite             = left
Signifikanzniveau α   = 0.05
Freiheitsgrade df     = 9

Beobachteter t-Wert   = -3.656
Kritischer t-Wert     = -1.833
p-Wert                = 0.0026
```

Abbildung 6: Ein-Stichprobentest (linksseitig) - Ausgabe

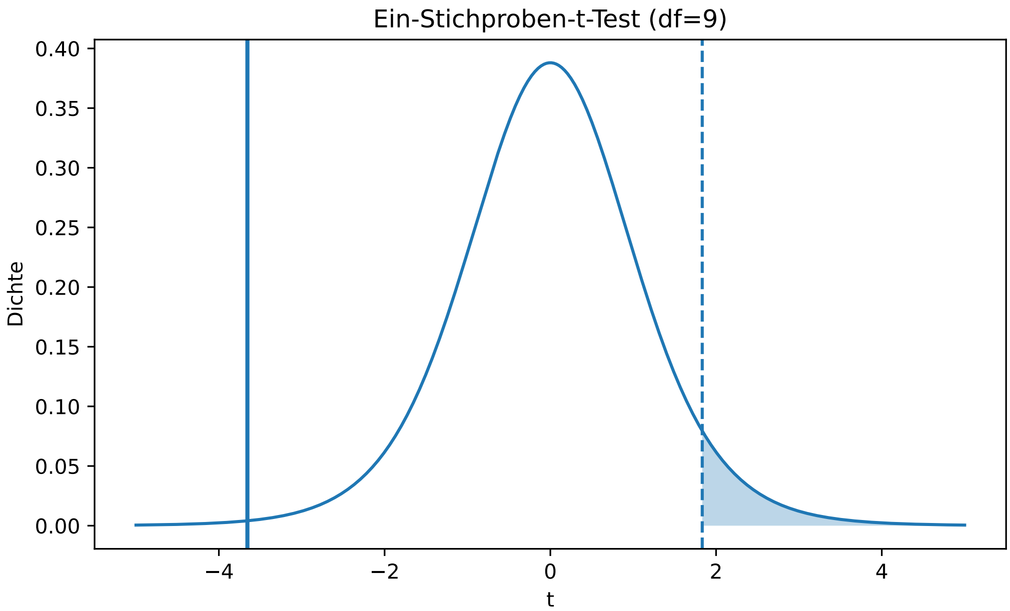

Abbildung 7: Ein-Stichprobentest (rechtsseitig) mit t_{krit} und t_{obs}

```
Ein-Stichproben-t-Test
----------------------
Beobachtete Daten: [122 125 130 128 127 126 124 129 131 123]
Datenumfang n         = 10
Referenzwert mu0      = 130
Stichprobenmittel     = 126.500
Stichproben-SD        = 3.028

Testseite             = right
Signifikanzniveau α   = 0.05
Freiheitsgrade df     = 9

Beobachteter t-Wert   = -3.656
Kritischer t-Wert     = 1.833
p-Wert                = 0.9974
```

Abbildung 8: Ein-Stichprobentest (rechtsseitig) - Ausgabe

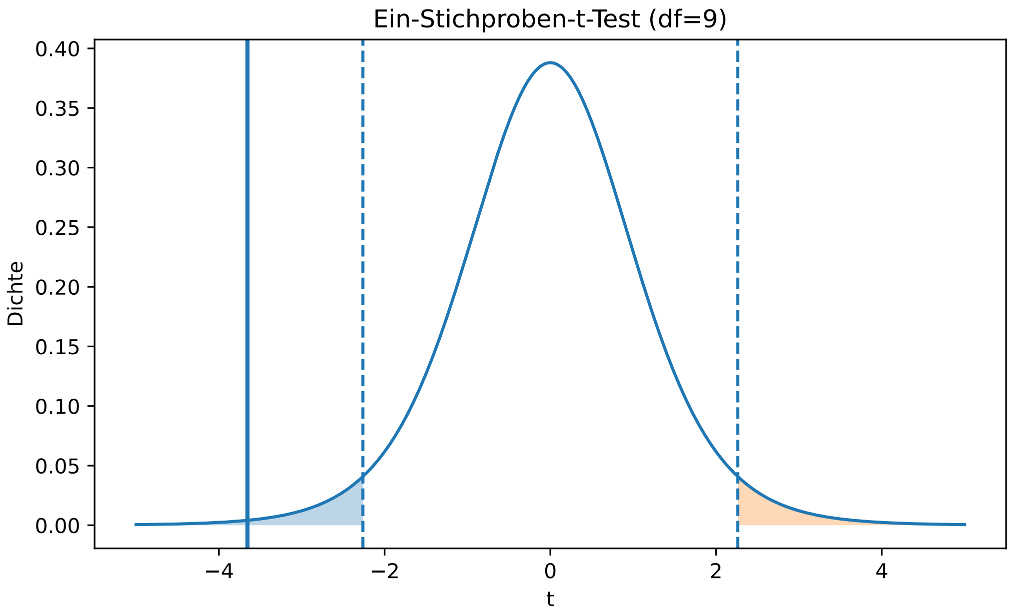

Abbildung 9: Ein-Stichprobentest (rechtsseitig) mit t_{krit} und t_{obs}

```
Ein-Stichproben-t-Test
----------------------
Beobachtete Daten: [122 125 130 128 127 126 124 129 131 123]
Datenumfang n         = 10
Referenzwert mu0      = 130
Stichprobenmittel     = 126.500
Stichproben-SD        = 3.028

Testseite             = two-sided
Signifikanzniveau α   = 0.05
Freiheitsgrade df     = 9

Beobachteter t-Wert   = -3.656
Kritischer t-Wert     = 2.262
p-Wert                = 0.0053
```

Abbildung 10: Ein-Stichprobentest (zweiseitig) - Ausgabe