TECHNISCHE UNIVERSITÄT
IN DER KULTURHAUPTSTADT EUROPAS
CHEMNITZ

Professur Psychologie digitaler Lernmedien
Institut für Medienforschung
Philosophische Fakultät

Statistik I

t-Test

Oppenheimer (2023). Universal Pictures.

Prof. Dr. Günter Daniel Rey

Überblick

- Einführung
- Beispiel zur Berechnung eines *t*-Tests
- Empirischer *t*-Wert
- Freiheitsgrade beim t-Test
- Geschätzter Standardfehlers der Mittelwertsdifferenz
- Bestimmung des kritischen *t*-Wertes
- Inferenzstatistische Entscheidung und Ergebnisdarstellung
- Inferenzstatistische Voraussetzungen
- Arten von *t*-Tests
- Schritte bei der Durchführung eines *t*-Tests

Einführung
(z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- *t*-Test: Statistisches Verfahren zum Vergleich zweier Mittelwerte
- Differenz der beiden (Gruppen-)Mittelwerte: Wichtigster Wert eines *t*-Tests
- Inferenzstatistischer Vergleich zwischen zwei Mittelwerten: Mittels *t*-Test möglich
- Entscheidungshilfe: Ob gefundener Mittelwertsunterschied rein zufällig entstanden ist oder systematische Unterschiede zwischen den (Gruppen-)Mittelwerten existieren

Beispiel zur Berechnung eines *t*-Tests

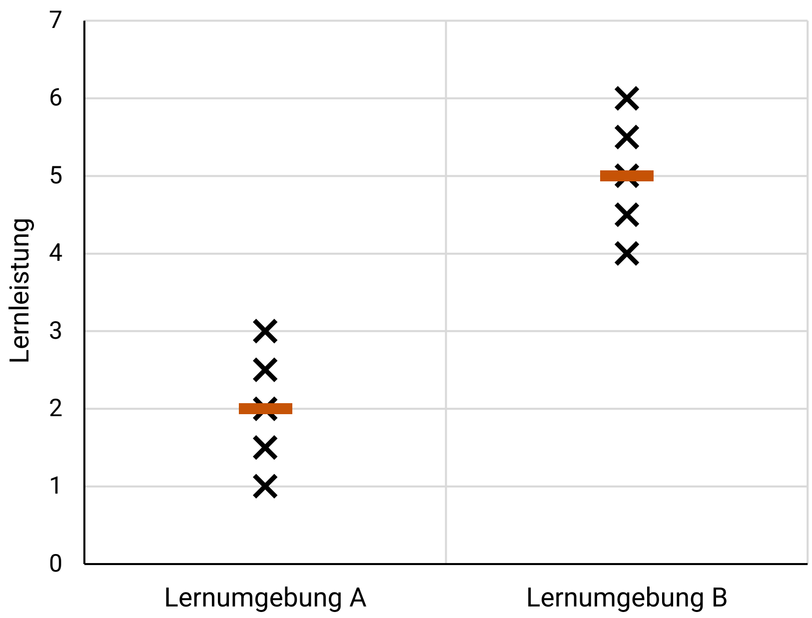
- Beispiel: Fiktive Ergebnisse einer Studie zum Lernen mit Medien
- Rohdaten als schwarze Kreuze, Mittelwerte als orangene Linien

Lernumgebung A

VPN	Transfer
1	2.0
2	1.5
3	1.0
4	2.5
5	3.0
M	2.0
SD	≈ 0.79

Lernumgebung B

VPN	Transfer
6	5.0
7	5.5
8	4.5
9	6.0
10	4.0
M	5.0
SD	≈ 0.79



Berechnung des empirischen *t*-Wertes

- Formel zur Berechnung des empirischen *t*-Wertes (unter der Nullhypothese):

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

t_{df}	= Empirischer *t*-Wert
df	= Freiheitsgrade
$\bar{x}_1$	= Mittelwert der Gruppe 1
$\bar{x}_2$	= Mittelwert der Gruppe 2
$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$	= Geschätzter Standardfehler der Mittelwertsdifferenz

Berechnung der Freiheitsgrade beim *t*-Test (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Freiheitsgrade: Legen die Genauigkeit von Populationsschätzern und damit die Form von Verteilungen fest, die auf Schätzern basieren wie z. B. die *t*-Verteilung
- Zahl der Freiheitsgrade: Gibt an, wie viele Werte theoretisch frei variieren können, wenn das Ergebnis bereits feststeht
- Berechnung der Freiheitsgrade beim *t*-Test:

$$df = n_1 + n_2 - 2$$

n_1 = Anzahl an Versuchspersonen in Gruppe 1
n_2 = Anzahl an Versuchspersonen in Gruppe 2

- Beispiel: Bei 2 x 5 Versuchspersonen ist $df = 5 + 5 - 2 = 8$

Berechnung des geschätzten Standardfehlers der Mittelwertsdifferenz beim *t*-Test (Rasch et al., 2021)

- Formel zur Berechnung des geschätzten Standardfehlers der Mittelwertsdifferenz beim *t*-Test:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$\hat{\sigma}_1^2$	= Geschätzte Varianz der Gruppe 1
$\hat{\sigma}_2^2$	= Geschätzte Varianz der Gruppe 2
n_1	= Anzahl an Versuchspersonen in Gruppe 1
n_2	= Anzahl an Versuchspersonen in Gruppe 2

- Varianz = Quadrierte Standardabweichung (vgl. Sitzung 3, Folie 15)
- Beispiel: Einsetzen ergibt:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} \approx \sqrt{\frac{0.79^2}{5} + \frac{0.79^2}{5}} \approx \sqrt{0.125 + 0.125} \approx 0.5$$

Beispiel zur Ermittlung des empirischen *t*-Wertes

- Berechnung des empirischen *t*-Wertes:

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

- Beispiel: Einsetzen ergibt:

$$t_8 \approx \frac{2 - 5}{0.5} \approx -6.00$$

Bestimmung des kritischen *t*-Wertes (Rasch, Friese, Hofmann & Naumann, 2021)

- Beispiel: Studie mit 10 Versuchspersonen (N = 10), einem Signifikanzniveau von 5% (α = .05) und zweiseitiger Testung (d. h. 2.5% auf jeder Seite der *t*-Verteilung)
- Ergebnis laut Tabelle: $t_8 \approx 2.306$

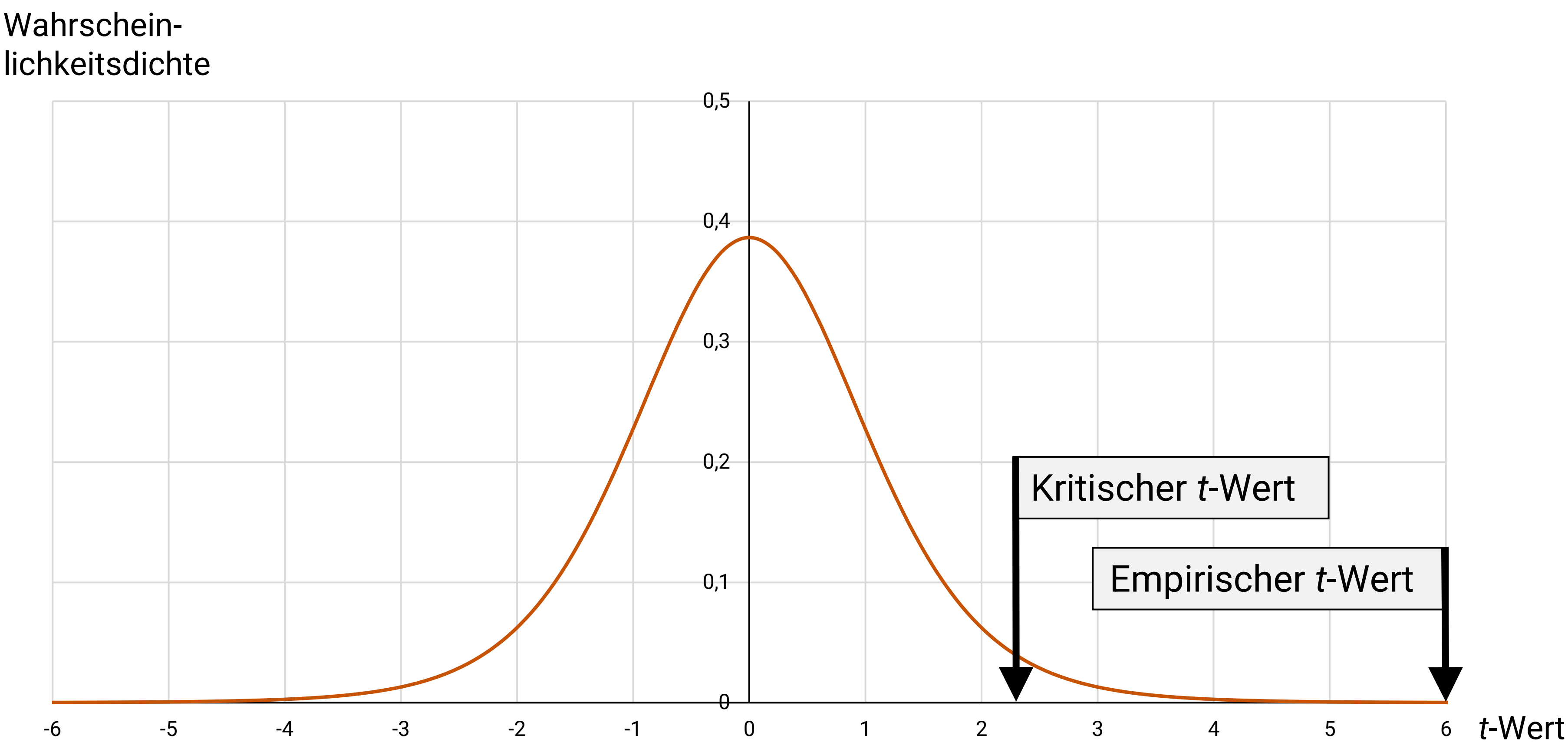
Tab. B: *t*-Verteilung

(Quelle: Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology* (p. 521). Englewood Cliffs: Prentice-Hall.)

df	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,975	0,990	0,995	0,9995
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,660	636,620
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587

Quelle: Rasch et al.(2021; S. 159)

Inferenzstatistische Entscheidung



Inferenzstatistische Entscheidung und Ergebnisdarstellung

- Zwei Möglichkeiten bei der inferenzstatistischen Entscheidung
  - Nicht signifikant ($|t_{emp}| < t_{krit}$): H_0 wird vorläufig beibehalten
  - Signifikant ($|t_{emp}| \geq t_{krit}$): H_0 wird zugunsten der H_1 verworfen
- Beispiel: Da $|t_{emp}| \approx 6.00 \geq t_{krit} \approx 2.306$ wird H_0 zugunsten der H_1 verworfen, d. h. das Ergebnis ist signifikant
- Inferenzstatistische Ergebnisdarstellung
  - Signifikantes Ergebnis: Angabe des empirischen *t*-Wertes (inklusive df) und *p*-Wertes sowie der Effektstärke
  - Nicht signifikantes Ergebnis: Ggf. zusätzlich Angabe der Teststärke
- Beispiel: $t(8) = 6.00$, $p < .001$, $d = 3.8$

Prüfung der inferenzstatistischen Voraussetzungen (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Intervallskalenniveau der abhängigen Variable: Überprüfung aufwändig und schwierig
- Normalverteilung der abhängigen Variable in der Population: Überprüfung z. B. mittels Shapiro-Wilk-Test
- Varianzhomogenität als Gleichheit der Populationsvarianzen, aus denen die beiden Stichproben stammen: Überprüfung mittels Levene-Test

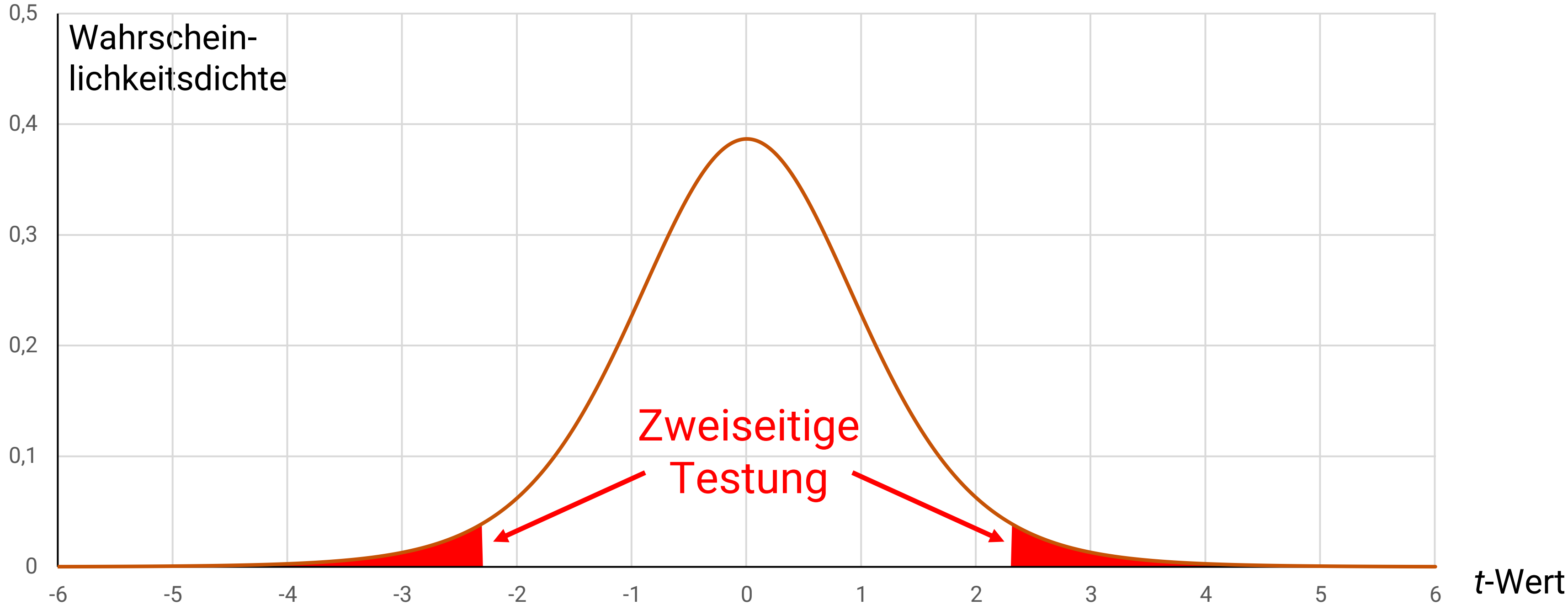
- Überprüfung der drei Voraussetzungen in der Forschungspraxis (leider) eher unüblich

Prüfung der inferenzstatistischen Voraussetzungen (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Verletzungen der Voraussetzungen: Bei Verletzungen einer der drei genannten Voraussetzungen kann statt des *t*-Tests (parametrisches Verfahren) u. a. ein nonparametrisches Verfahren (wird in Vorlesung „Statistik II“ behandelt) verwendet werden
- Robustheit: Allerdings reagiert der *t*-Test unter folgenden Bedingungen relativ robust gegenüber Verletzungen der Voraussetzungen
    - Ungefähr gleichgroße Stichproben der beiden Gruppen
    - Hinreichend große Stichproben ($n_1 = n_2 > 30$)
- In der Praxis: Nonparametrische Verfahren selten genutzt

Arten von *t*-Tests (z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Einseitige vs. zweiseitige Testung
  - Einseitige Testung: Bei gerichteten Hypothesen
  - Zweiseitige Testung: Bei ungerichteten Hypothesen
  - Teststärke: Höhere Teststärke bei einseitiger Testung



Arten von *t*-Tests
(z. B. Rasch, Friese, Hofmann & Naumann, 2021)

- Unabhängige vs. abhängige Stichproben
  - U. a. bei Messwiederholungen liegen abhängige Stichproben vor
  - In diesem Fall Berechnung des *t*-Wertes über die Differenzen der Messwerte
- *t*-Test für eine Stichprobe (Einstichprobentest; engl. one-sample *t*-test): Überprüfung, ob Stichprobe aus einer bestimmten Population stammt

Schritte bei der Durchführung eines *t*-Tests (Rasch, Friese, Hofmann & Naumann, 2021)

- Aufstellen einer Hypothese
- Prüfung der inferenzstatistischen Voraussetzungen
- Festlegung eines Populationseffekts
- Festlegung des Signifikanzniveaus
- Stichprobenumfangsplanung
- Bestimmung des kritischen *t*-Wertes (t_{krit})
- Prüfung des empirischen *t*-Wertes (t_{emp}) auf Signifikanz
- Interpretation des Ergebnisses

Aufstellen einer Hypothese
(Rasch, Friese, Hofmann & Naumann, 2021)

- Hypothesenformulierung: Aufstellung einer möglichst spezifischen Hypothese
- H_0 und H_1: Definition der (statistischen) Nullhypothese und (statistischen) Alternativhypothese ableiten
- Beispiel:
  - H_1: $\mu_1 - \mu_2 > 0$
  - H_0: $\mu_1 - \mu_2 \leq 0$

Populationsmittelwerte:
Experimentalgruppe: μ_1
Kontrollgruppe: μ_2

Prüfung der inferenzstatistischen Voraussetzungen (Rasch, Friese, Hofmann & Naumann, 2021)

- Intervallskalenniveau der abhängigen Variable: In der Regel keine Überprüfung
- Normalverteilung der abhängigen Variable in der Population: Überprüfung z. B. mittels Shapiro-Wilk-Test
  - Shapiro-Wilk-Test nicht signifikant: Annahme der Normalverteilung wird (vorerst) beibehalten
  - Shapiro-Wilk-Test signifikant: Unter Umständen Einsatz eines nonparametrischen Verfahrens
- Varianzhomogenität: Überprüfung mittels Levene-Test
  - Levene-Test nicht signifikant: Annahme der Varianzhomogenität wird (vorerst) beibehalten
  - Levene-Test signifikant: Korrektur der Freiheitsgrade (erfolgt in vielen Statistikprogrammen automatisch)

Festlegung eines Populationseffekts (Rasch, Friese, Hofmann & Naumann, 2021)

- Inhaltliche Überlegungen: Festlegung eines Populationseffektes abhängig von inhaltlichen (≠ statistischen) Überlegungen
- Orientierungsmöglichkeiten: Bereits vorliegende Studien oder Metaanalysen zu vergleichbaren Fragestellungen
- Koventionen: Falls keine Studien oder Metaanalysen vorliegen sollten, dann Rückgriff auf Konventionen von Cohen (1988)

Effektgröße	Kleiner Effekt	Mittlerer Effekt	Großer Effekt
d	0.20	0.50	0.80
r	0.10	0.30	0.50

Festlegung des Signifikanzniveaus (Rasch, Friese, Hofmann & Naumann, 2021)

- 5%-Niveau: Gängiges Signifikanzniveau per Konvention
- Andere Signifikanzniveaus: Abhängig von der Fragestellung auch andere Signifikanzniveaus denkbar
- Statistische Signifikanz: Für die Entscheidungsfindung auf Basis der berechneten Wahrscheinlichkeit (*p*-Wert)

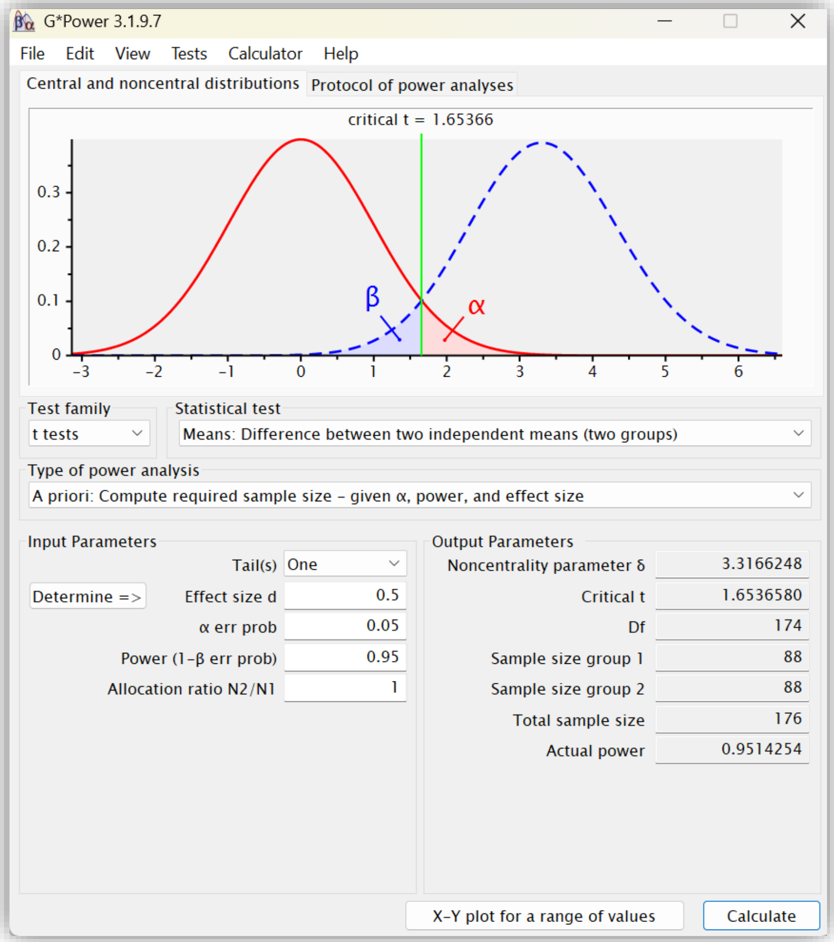
	Nicht signifikant	Signifikant	Sehr signifikant	Hoch signifikant
p-Wert	> 5%	≤ 5%	≤ 1%	≤ 0.1%
Abkürzung	n.s.	*	**	***

Stichprobenumfangsplanung (Rasch, Friese, Hofmann & Naumann, 2021)

- Wichtigkeit: Stichprobenumfangsplanung wichtiger Schritt bei allen empirischen Studien mit inferenzstatistischer Datenauswertung
- Voraussetzung (u. a.): Festlegung des Populationseffekts, des Signifikanzniveaus und der gewünschten Teststärke
- Berechnung: Erfolgt mithilfe des Nonzentralitätsparameters λ und wird in der Vorlesung „Statistik II“ erläutert
- G*Power: Programm zur Berechnung des Stichprobenumfanges (oder anderer statistischer Kenngrößen)

Exkurs: G*Power

- Beispiel: Berechnung des Stichprobenumfangs mit G*Power für ein Zweigruppendesign ohne Messwiederholung
  - *t*-Test
  - Einseitige Testung
  - Angenommene Effektgröße: *d* = 0.5 (mittlerer Effekt)
  - Signifikanzniveau: 0.05
  - Gewünschte Teststärke: 0.95
- Benötigter Stichprobenumfang: 176



Bestimmung des kritischen *t*-Wertes (Rasch, Friese, Hofmann & Naumann, 2021)

- Bestimmung des kritischen *t*-Wertes (t_{krit}): Erst nach Durchführung der Datenerhebung sinnvoll, weil erst dann sicher ist, wie viele Daten in die Auswertung eingehen
- Beispiel: Studie mit 113 Versuchspersonen (N = 113), einem Signifikanzniveau von 5% (α = .05) und einseitiger Testung
- $t_{krit(df=111)} \approx 1.66$

Tab. B: *t*-Verteilung

(Quelle: Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology* (p. 521). Englewood Cliffs: Prentice-Hall.)

df	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,975	0,990	0,995	0,9995
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,660	636,620
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
...													
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
z	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Quelle: Rasch et al.(2021; S. 159)

Prüfung des empirischen *t*-Wertes auf Signifikanz (Rasch, Friese, Hofmann & Naumann, 2021)

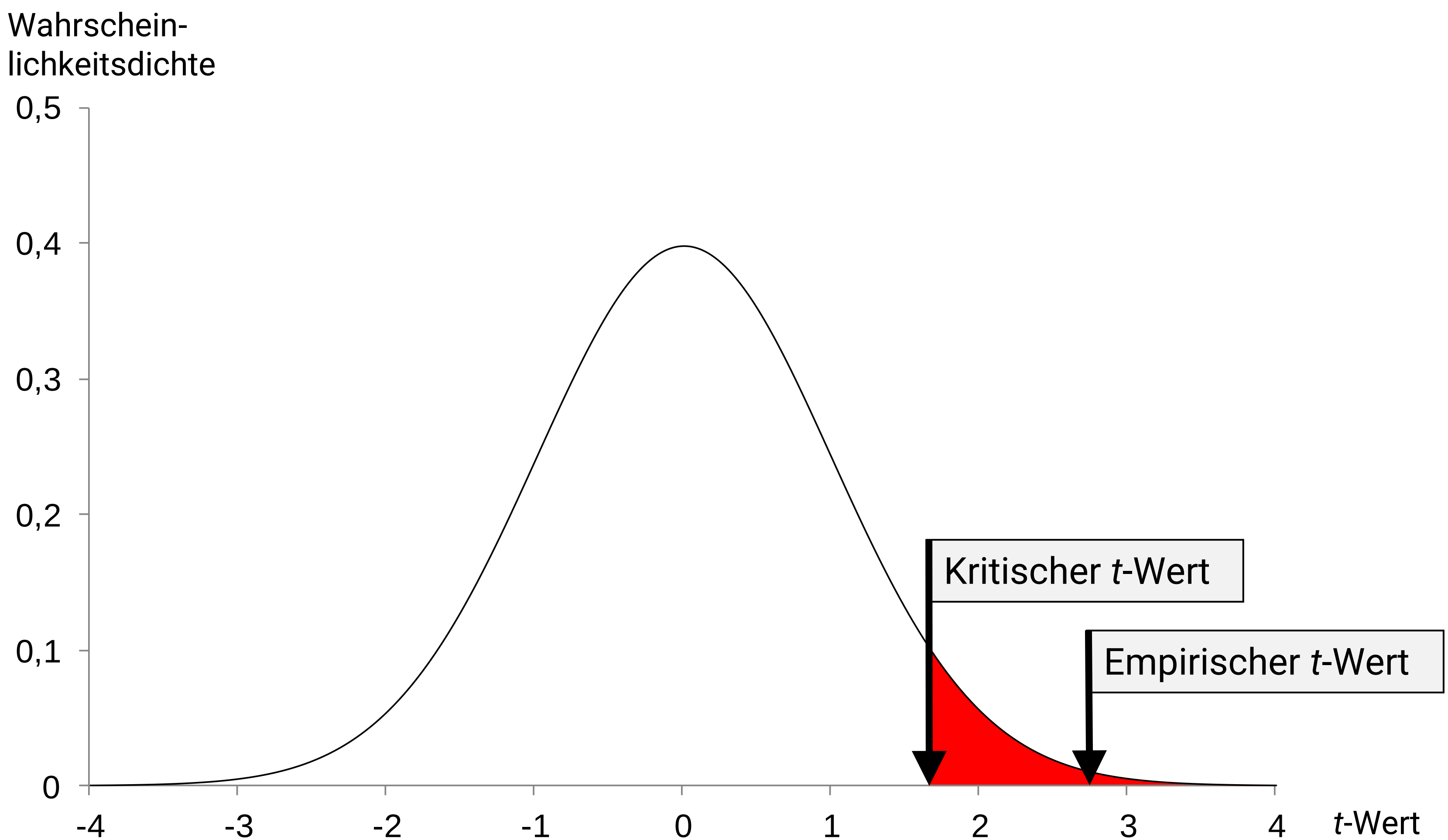
- Berechnung des empirischen *t*-Wertes (t_{emp}): Berechnung der empirischen Mittelwertsdifferenz und Schätzung der Populationsstreuung
- Formel zur Berechnung des empirischen *t*-Wertes (unter der Nullhypothese):

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

- Beispiel:

$$t_{111} \approx \frac{5.71 - 4.75}{0.35} \approx 2.77$$

Inferenzstatistische Entscheidung treffen (Rey, 2020)



Interpretation des Ergebnisses (Rasch, Friese, Hofmann & Naumann, 2021)

- Zwei Möglichkeiten bei der inferenzstatistischen Entscheidung
  - Nicht signifikant ($t_{emp} < t_{krit}$): H_0 wird vorläufig beibehalten
  - Signifikant ($t_{emp} \geq t_{krit}$): H_0 wird zugunsten der H_1 verworfen
- Beispiel: Da $t_{emp} = 2.77 \geq t_{krit} \approx 1.66$ wird H_0 zugunsten der H_1 verworfen, d. h. das Ergebnis ist signifikant
- Inferenzstatistische Ergebnisdarstellung
  - Signifikantes Ergebnis: Angabe des empirischen *t*-Wertes (inklusive df) und *p*-Wertes sowie der Effektstärke
  - Nicht signifikantes Ergebnis: Ggf. zusätzlich Angabe der Teststärke
- Beispiel: $t(111) = 2.77$, $p < .01$, $d = 0.52$

Beispiele für *t*-Tests in Fachzeitschriften

two multiple comparisons in this analysis. Whereas there was no difference between the two non-identical clockwork types functioning and non-functioning, $t(69) = 0.50$, $p = .62$, $d = 0.03$, search times for both non-identical functioning and non-identical non-functioning clockworks were faster than search times for identical clockworks, $t(69) = 3.78$, $p < .001$, $d = 0.18$ and $t(69) = 2.61$, $p = .011$, $d = 0.15$, respectively. However, the interaction of clockwork type

Quelle: Huff, Bauhoff und Schwan (2012)

Table 2
Means (and standard deviations) for the tests assessing spatial abilities and working memory span.

Learner-related factors	Presentation mode		*t*	*df*
	Vexing image	Split screen		
Spatial abilities				
Mental rotation (MRT)	22.59 (11.93)	23.39 (14.40)	<1	71.77
Visual patterns (VPT)	9.03 (1.61)	8.55 (1.55)	1.31	74.99
Spatial orientation (SO)	32.87 (21.85)	33.84 (30.80)	<1	66.58
Working memory span				
Zahlen merken (ZM)	34.15 (7.67)	32.34 (7.04)	1.08	74.74

* $p < .01$.
** $p < .001$.

Quelle: Huff, Bauhoff und Schwan (2012)

The bottom two rows of Table 1 show the means (and standard deviations) on the retention and transfer tests for the pre-training and no-training groups. *T* tests (for independent means) were conducted on the retention test data and the transfer test data. The pre-training group scored significantly higher than the no-training group on the retention test, $t(31) = 2.30$, $p < .05$, and on the transfer test, $t(31) = 2.50$, $p < .05$. The effect sizes (based on Cohen's *d*) were 0.64 for retention and 1.54 for transfer. Consistent with the two-stage theory of mental model construction and the results of Experiment 1, students in the pre-training group outperformed the other students in applying what they had learned in new situations, confirming that computer-based pre-training led to a deeper understanding in the same way as paper-based pre-training.

Quelle: Mayer, Mathias und Wetzell (2002)

Zusammenfassung

- *t*-Test: Statistisches Verfahren zum Vergleich zweier Mittelwerte
- Grundprinzip: Inferenzstatistischer Vergleich zwischen zwei Mittelwerten

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

- Inferenzstatistische Überprüfung: Vergleich eines empirischen *t*-Wertes mit einem kritischen *t*-Wert
- Voraussetzung von Varianzanalysen: Intervallskalenniveau, Normalverteilung und Varianzhomogenität
- Arten von *t*-Tests: Einseitige vs. zweiseitige Testung; unabhängige vs. abhängige Stichproben; Einstichprobentest
- Verschiedene Schritte bei der Durchführung eines *t*-Tests

Prüfungsliteratur

- Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2021). *Quantitative Methoden 1: Einführung in die Statistik für Psychologie, Sozial- & Erziehungswissenschaften* (5. Aufl.). Heidelberg: Springer.
  - Der *t*-Test (S. 35–82)

Weiterführende Literatur I

- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.
  - Hypothesentesten (S. 97–-115)
  - Tests zur Überprüfung von Unterschiedshypothesen (S. 117–136)
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5. Aufl.). Weinheim: Beltz.
  - Grundlagen der Inferenzstatistik (S. 217–277)
  - Vergleich zweier Stichprobenmittelwerte (S. 331–343)
- Leonhart, R. (2022). *Lehrbuch Statistik. Einstieg und Vertiefung* (5. Auflage). Bern: Huber.
  - Einführung in die inferenzstatistische Hypothesenprüfung (S. 175–199)
  - Parametrische Testverfahren (S. 203–226)

Weiterführende Literatur II

- Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik: Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (3. Aufl.). München: Pearson.
  - Inferenzstatistik (S. 309–428)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Hillsdale, NJ: Erlbaum.
- Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*, 135–147.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*, 21–33.