Methodenlehre

Vorlesung 12

Prof. Dr. Björn Rasch, Cognitive Biopsychology and Methods
University of Fribourg

Methodenlehre I

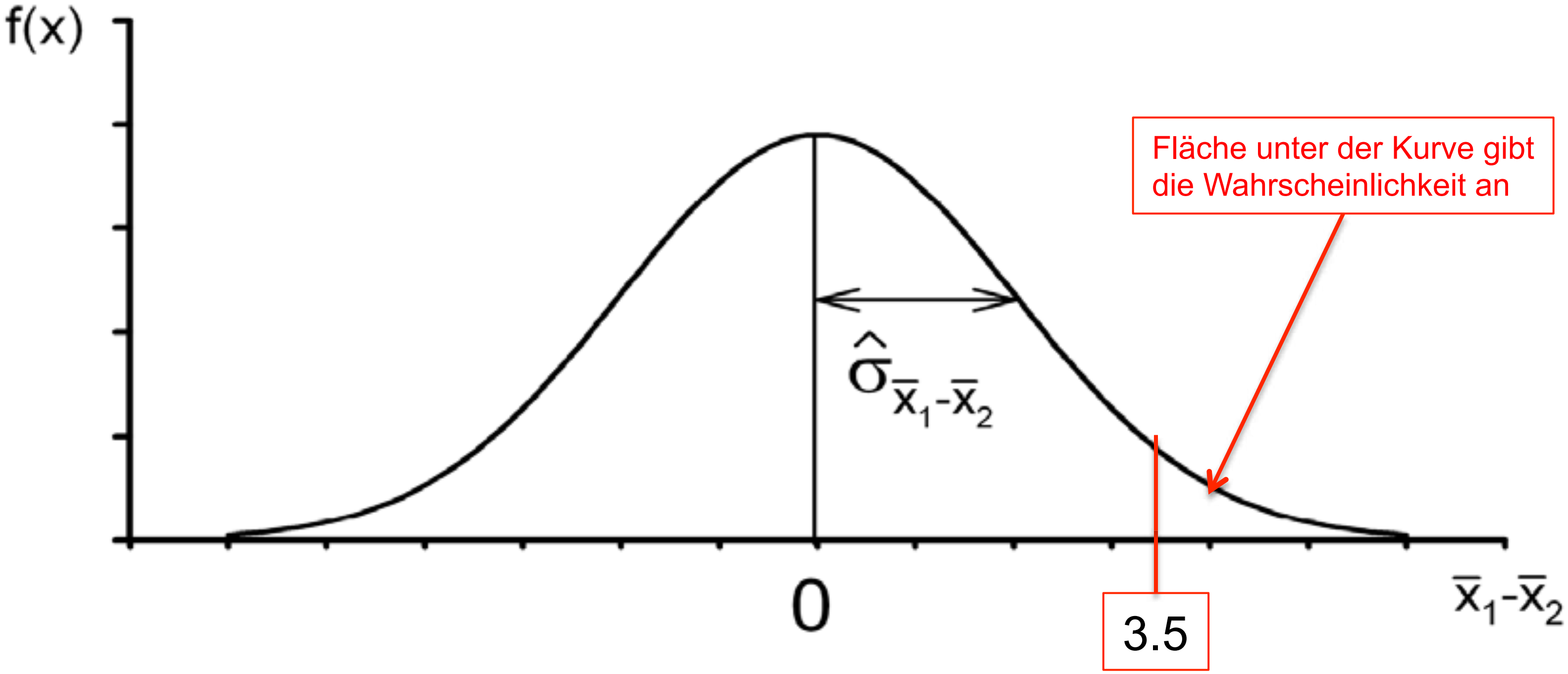
	Woche	Datum	Thema I
	FQ		Einführung, Verteilung der Termine
	1	25.9.13	Psychologie als Wissenschaft
	2	2.10.13	Hypothesen und Variablen
	3	9.10.13	Operationalisieren und Messen
	4	16.10.13	Objektivität, Reliabilität, Validität
	5	23.10.13	Das Experiment
	6	30.10.13	Störvariablen und ihre Kontrolle
	7	6.11.13	Forschungsethik
	8	13.11.13	Durchführung und Berichten eines Experiments
	9	20.11.13	Stichprobe und Population
	10	27.11.13	Statistische Bedeutsamkeit (Signifikanz)
	11	4.12.13	Inhaltliche Bedeutsamkeit
	12	11.12.13	Teststärke und Stichprobenumfangsplanung
	13	18.12.13	Wiederholung und Fragen

Take Home Messages

- Statistische Signifikanz
  - Wie wahrscheinlich ist eine beobachtete (empirische) Mittelwertsdifferenz unter der Annahme der Nullhypothese?
  - Entscheidungsregel: Ist diese Wahrscheinlichkeit kleiner als die (vordefinierte) Signifikanzschwelle, dann lehnen wir die Nullhypothese ab
    - Signifikanzschwelle ist (meist) $P < 0.05$ (5% Fehlerwahrscheinlichkeit)
- Statistische Signifikanz ist abhängig von der Stichprobengrösse
  - Je grösser die Stichprobe, desto eher wird ein Ergebnis signifikant
    - Jeder noch so kleiner Unterscheid kann signifikant „gemacht“ werden
- Effektstärke
  - Angabe der Grösse des Effekts unabhängig von der Stichprobengrösse
    - Distanzmass d, Mass für die Varianzaufklärung η^2
    - Bei signifikantem Ergebnis Effektstärke mit angeben
  - Erlaubt den Vergleich von Ergebnissen mit unterschiedlichen Stichprobengrössen
  - Erlaubt eine Abschätzung der Inhaltlichen Bedeutsamkeit des Ergebnisses
    - Z.B. anhand der Konventionen von Cohen (1988)

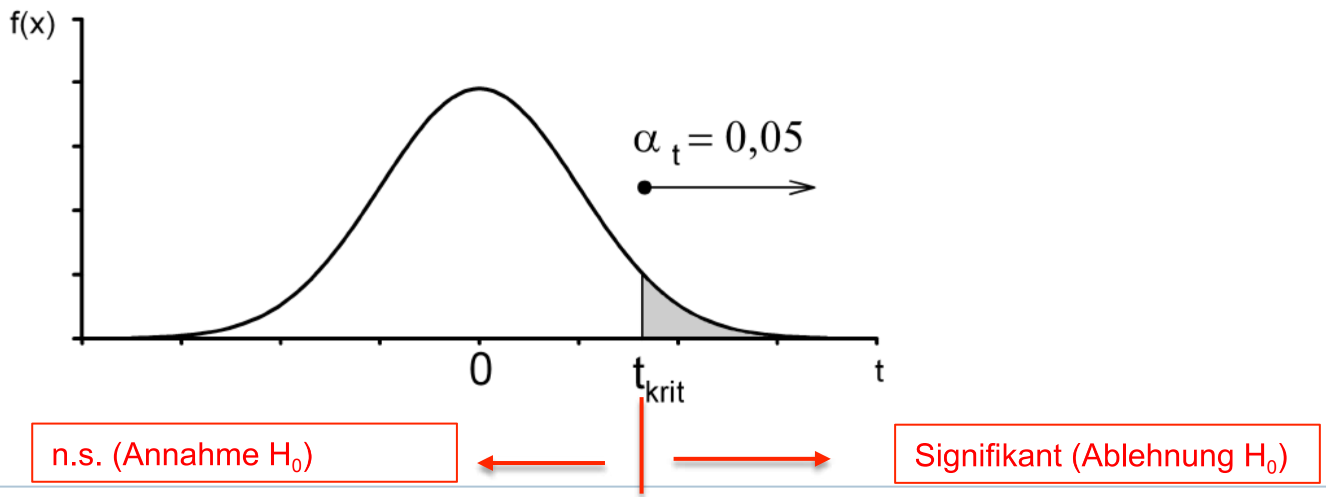
Stichprobe und Population

- Signifikanztest
  - Frage: Wie wahrscheinlich ist das Auftreten der beobachteten Differenz der Stichprobenmittelwerte unter der Annahme der Nullhypothese?
    - Beispiel: Beobachtete Differenz: 42 – 38.5 = 3.5



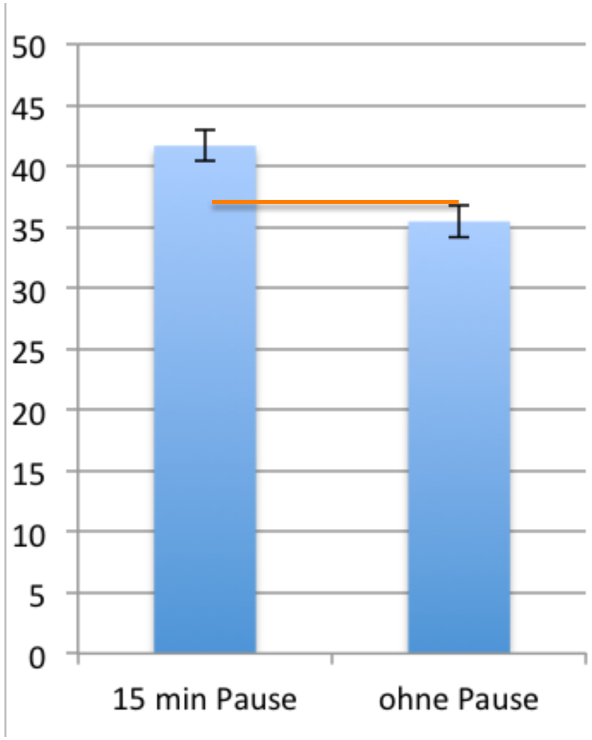
Signifikanzprüfung

- Entscheidungsregel
  - Ist die Wahrscheinlichkeit eines empirischen Ergebnisse unter der Nullhypothese kleiner als das festgelegte Signifikanzniveau, so lehne ich die Nullhypothese ab
    - Signifikant: Ablehnung der H_0
    - Es gibt einen Unterschied zwischen den Gruppen.
  - Ist die Wahrscheinlichkeit grösser, so nehme ich die Nullhypothese an.
    - Nicht signifikant: Annahme der H_0
    - Es gibt keinen Unterschied zwischen den Gruppen.

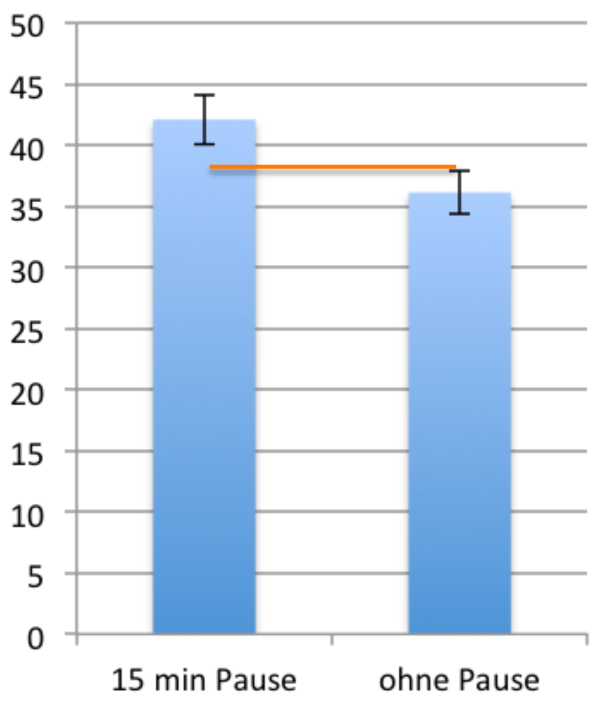


Beispiel

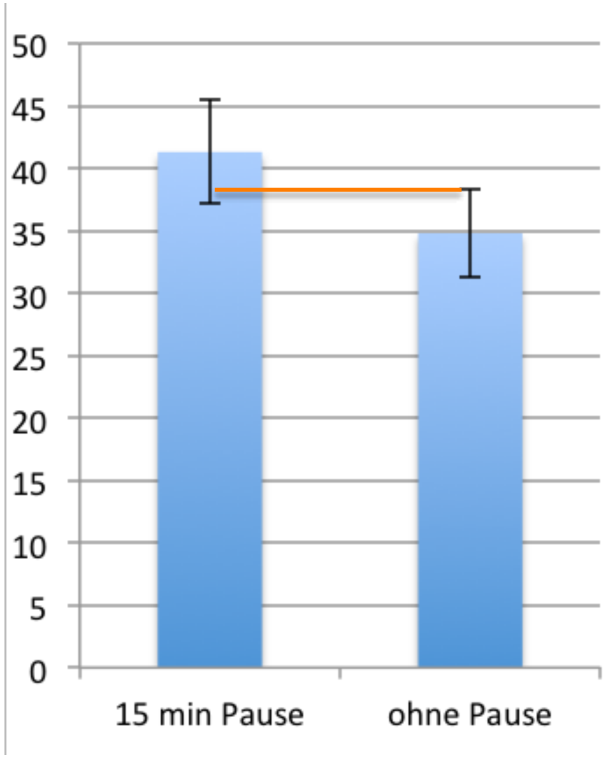
- Ergebnisvergleich



N = 100
P = 0.001
signifikant



N = 50
P = 0.033
signifikant



N = 20
P = 0.25
n.s.

Inhaltliche Bedeutsamkeit

- Problem:
  - Statistische Signifikanzen stark von der Stichprobengrösse abhängig
  - Ergebnisse zwischen Studien mit unterschiedlichen Stichprobengrössen nicht vergleichbar
- Lösung
  - Effektstärken
    - Effektstärken geben die Grösse eines Effekts unabhängig von der Stichprobengrösse an
    - Erlaubt den Vergleich zwischen Studien
    - Erlaubt die Einschätzung der inhaltlichen Bedeutsamkeit eines Effekts.
- Handlungsanweisung
  - Bei einem statistisch signifikanten Ergebnis immer auch die Effektstärke mit angeben

Effektstärken

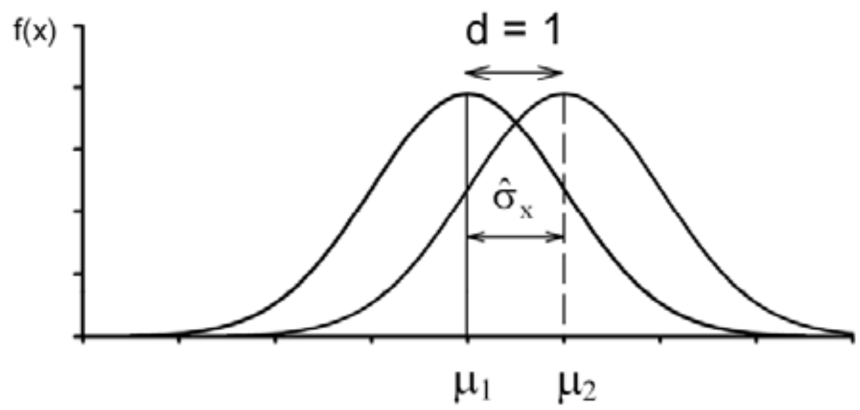
- **Effektstärke als Distanzmass**
  - Distanz zwischen Populationsmittelwerten
  - Effektstärkenmass d

$$d = \frac{\mu_1 - \mu_2}{\sigma_x}$$

$\mu_1; \mu_2$: Mittelwerte der Populationen, aus denen die Stichproben gezogen werden

σ_x : Streuung der Population innerhalb einer Bedingung

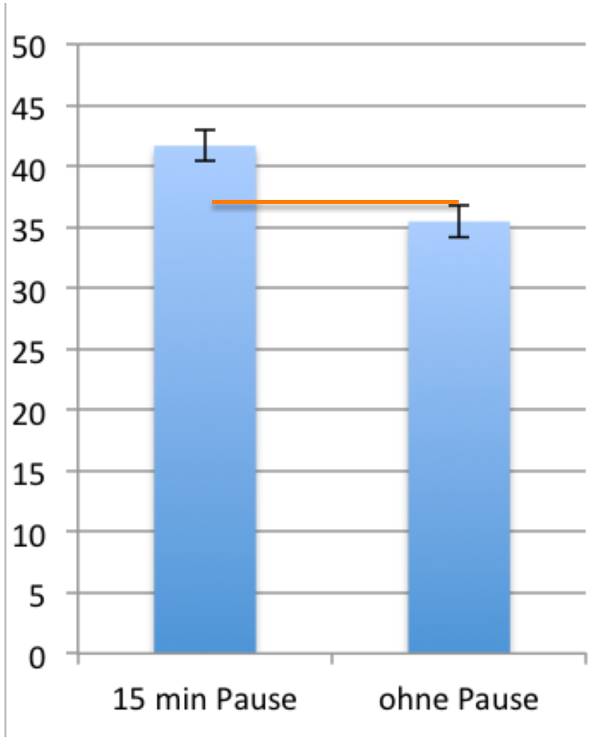
- **Interpretation**
  - d = 1: Populationen unterschieden sich um eine Streuungseinheit



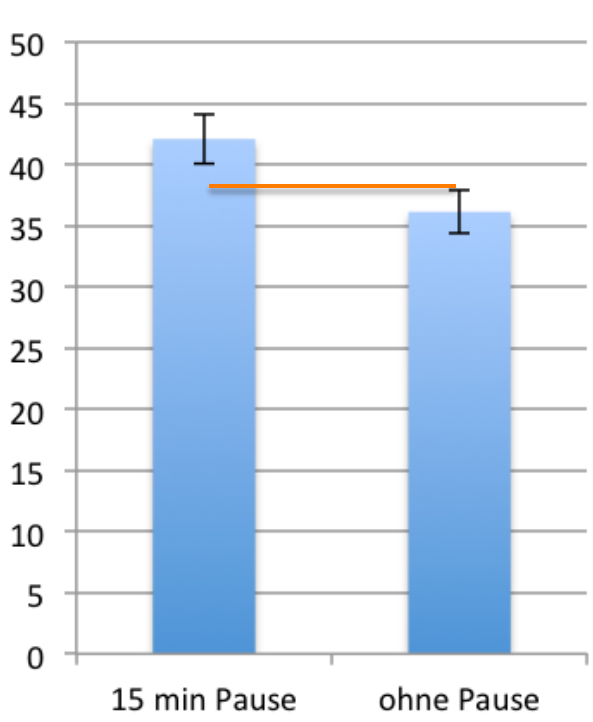
- **Konventionen**
  - d = 0.2: kleiner Effekt
  - d = 0.5: mittlerer Effekt
  - d = 0.8: grosser Effekt
    - Konventionen gelten nur für nicht messwiederholte Mittelwertsvergleiche

Beispiel

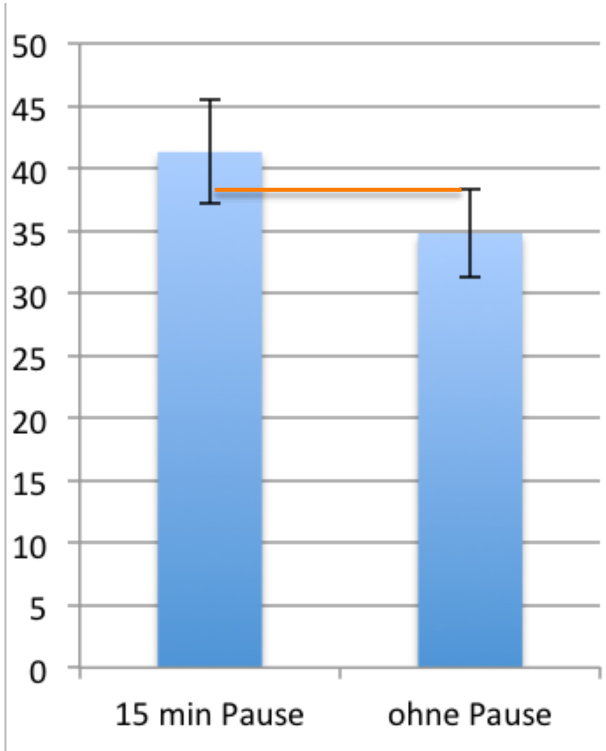
- Ergebnisvergleich



N = 100
P = 0.001
Signifikant
d = 0.69



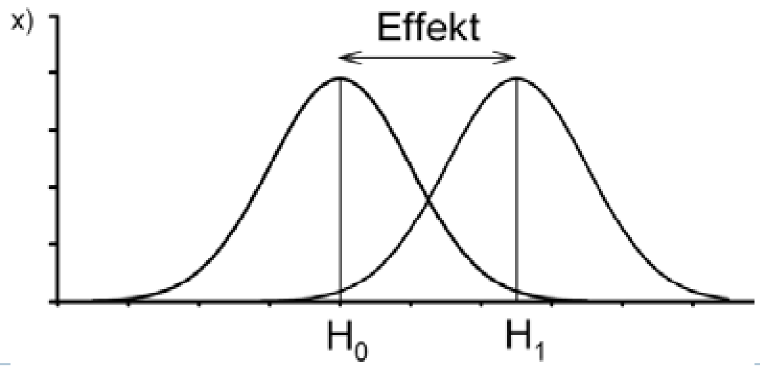
N = 50
P = 0.033
Signifikant
d = 0.62



N = 20
P = 0.25
n.s.
d = 0.59

Effektstärke

- Alternativhypothese (H_1)
  - Es liegt ein Effekt der experimentellen Manipulation in der Population vor.
- Spezifische Alternativhypothese
  - Es liegt ein Effekt einer bestimmte Grösse vor
- Effektstärke
  - Wie stark unterscheidet sich die Alternativhypothese von der Nullhypothese?
    - Wenn H_0 zutrifft: Empirische Ergebnisse verteilen sich um Null
    - Wenn H_1 zutrifft: Empirische Ergebnisse verteilen sich um die wahre Differenz zwischen den Populationsmittelwerten



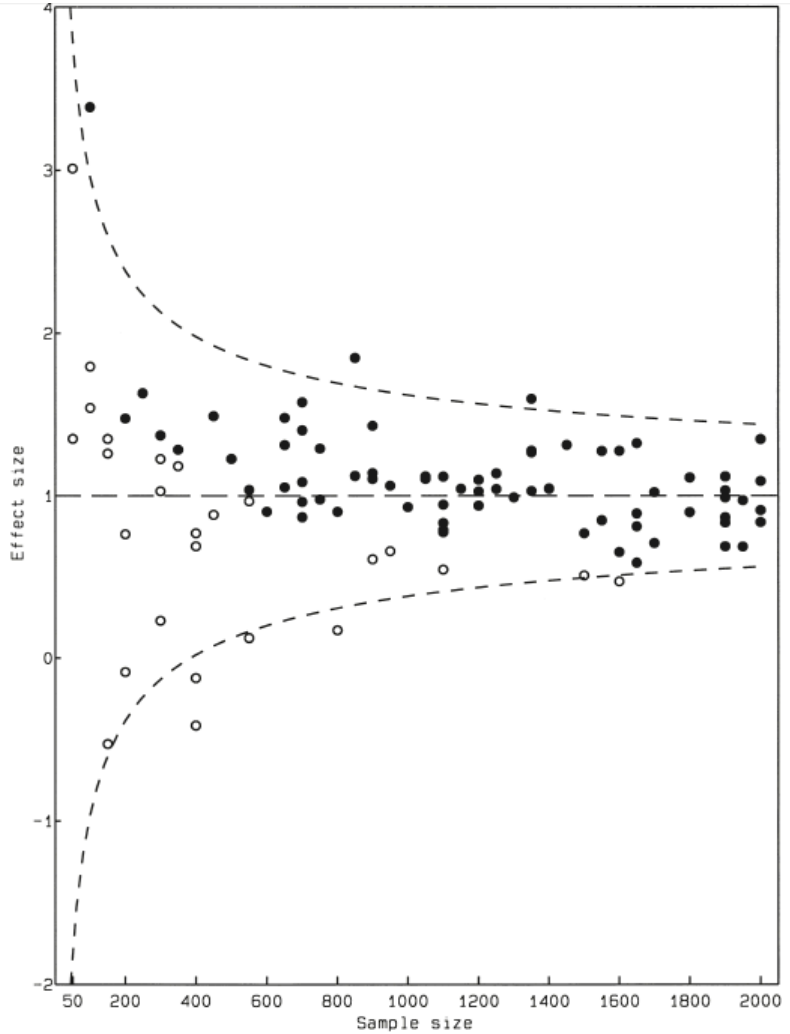
Effektstärken

- Effektstärken geben die Stärke des experimentellen Effekts an
  - Effektstärken sind unabhängig von der Stichprobengrösse
  - Erlauben der Vergleich zwischen verschiedenen Studien
    - Meta-Analyse von mehreren Ergebnissen
    - Berechnung einer mittleren Effektstärke möglich

- Schätzung des „wahren" Effekts um so besser, je grösser die Stichprobe ist
  - Vertrauensintervall (Konfidenzintervall) von Effektstärken
    - Effektstärken verteilen sich um den „wahren" Effekt
    - Normalverteilung von den aus Stichproben geschätzten Effektstärken
  - Bei kleinen Stichproben kann der empirische Effekt zufällig sehr viel grösser oder kleiner als der „wahre" Effekt sein.

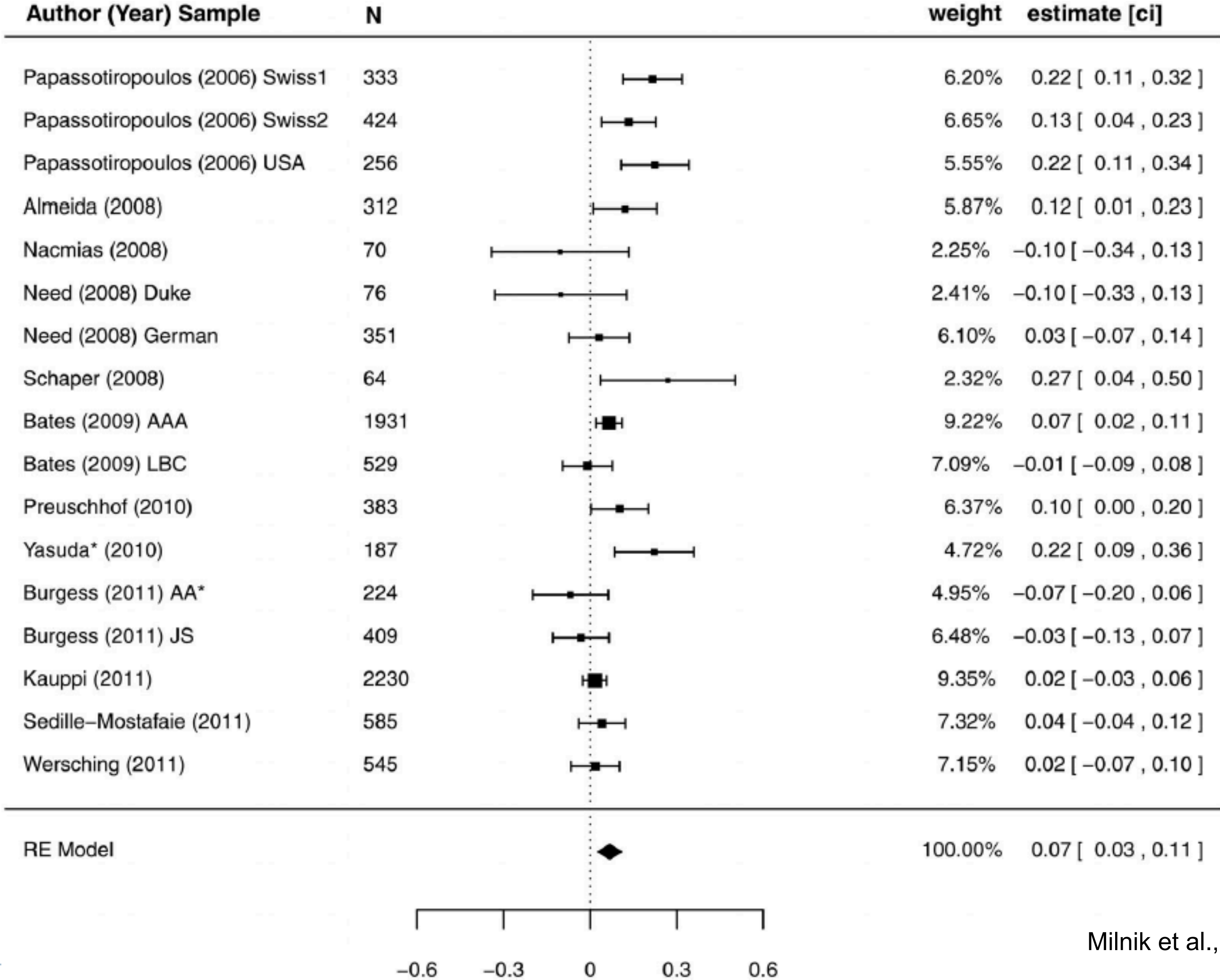
Effektstärken

- Simulierte Verteilung von Effektstärken
  - „wahrer“ Effekt d = 1
    - Schwarz: Effekte sind signifikant
    - Weiss: Effekte sind nicht signifikant
  - „Funnel plot“
- Problem:
  - Signifikante Ergebnisse werde häufiger publiziert als nicht signifikante
    - Publication Bias
  - Bei vorliegen eines Publication Bias ist Funnel Plot nicht vollständig
    - Weissen Punkte fehlen



Thornton & Lee, 2000, *Clin Epidemiology*

Meta-Analyse



Milnik et al., 2012, *Medial Genetics*

Teststärke und Stichprobenumfang

- Problem
  - Statistische Signifikanz abhängig vom Stichprobengrösse
    - Je grösser die Stichprobe, desto eher wird ein Unterschied signifikant
      - „Jeder noch so kleine Effekt kann signifikant gemacht werden“
- Lösung
  - Optimaler Stichprobenumfang, um inhaltlich relevanten Effekt zu finden
    - Falls der Effekt (in der Population) existiert
- Beispiel:
  - Metalldetektor “Goldi_2000“ am Strand auf der Suche nach Goldmünzen
    - Je besser mein Metalldetektor, desto eher finde ich auch kleine Münzen
      - „Ich kann jeden noch so kleinen Metallsplitter entdecken, wenn mein Metalldetektor nur gut genug ist“
      - Aber was soll ich mit kleinen Metallsplittern? Ich will Münzen!!!

Teststärke und Stichprobenumfang

- Goldi_2000
  - Entdeckt Münzen mit Durchmesser von 2 cm mit p = 95%
    - Münzen mit Ø = 3 cm: p = 99%
    - Münzen mit Ø = 1 cm: p = 70%
    - Metallsplitter mit Ø = 0.5 cm: p = 50%
- Goldi_10000
  - Entdeckt Münzen mit Durchmesser von 1 cm mit p = 95%
    - Münzen mit Ø = 3 cm: p = 99.99%
    - Münzen mit Ø = 2 cm: p = 99%
    - Metallsplitter mit Ø = 0.5 cm: p = 70%
- Goldi_20B
  - Entdeckt Münzen mit Durchmesser von 3 cm mit p = 60%
    - Münzen mit Ø = 2 cm: p = 50%
    - Münzen mit Ø = 1 cm: $p < 50\%$
    - Metallsplitter mit Ø = 0.5 cm: $p < 50\%$

Teststärke

- Teststärke (statistical power)
  - Wahrscheinlichkeit einen Effekt einer bestimmten Grösse zu finden, falls er wirklich existiert
    - Verteilung des Stichprobenkennwerts unter der Alternativhypothese
      - Alternativhypothese: Annahme, das ein Effekt einer bestimmten Grösse in der Population existiert
  - Teststärke ist abhängig von
    - Grösse des (gesuchten oder wahren) Effekts
    - Stichprobengrösse
    - Signifikanzniveau
  - Teststärke sollte mindestens > 80% sein
    - besser: Teststärke > 90%
    - Hängt von der Grösse des gesuchten (relevanten) Effekts ab

Statistische Signifikanz und Teststärke

- Statistische Signifikanz
  - Wahrscheinlichkeit eines empirischen Ergebnisses unter der Annahme der Nullhypothese (H_0)
- Teststärke
  - Wahrscheinlichkeit für ein signifikantes Ergebnis unter der Annahme einer (spezifischen) Alternativhypothese (H_1)

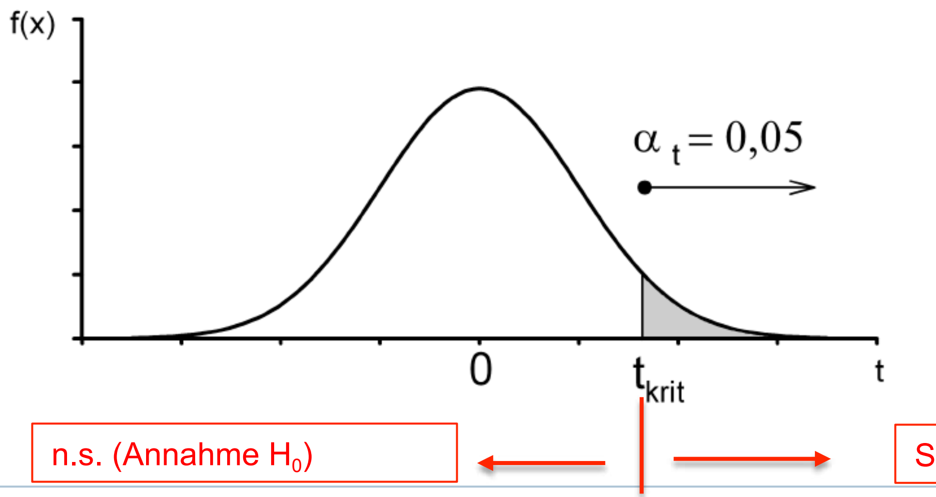
	In Wirklichkeit gilt die H_0	In Wirklichkeit gilt die H_1
Entscheidung zugunsten der H_0	Richtige Entscheidung	**β-Fehler**
Entscheidung zugunsten der H_1	**α-Fehler**	Richtige Entscheidung

Teststärke

Signifikanz

Signifikanzprüfung

- Entscheidungsregel
  - Ist die Wahrscheinlichkeit eines empirischen Ergebnisse unter der Nullhypothese kleiner als das festgelegte Signifikanzniveau, so lehne ich die Nullhypothese ab
    - Signifikant: Ablehnung der H_0
    - Es gibt einen Unterschied zwischen den Gruppen.
  - Ist die Wahrscheinlichkeit grösser, so nehme ich die Nullhypothese an.
    - Nicht signifikant: Annahme der H_0
    - Es gibt keinen Unterschied zwischen den Gruppen.

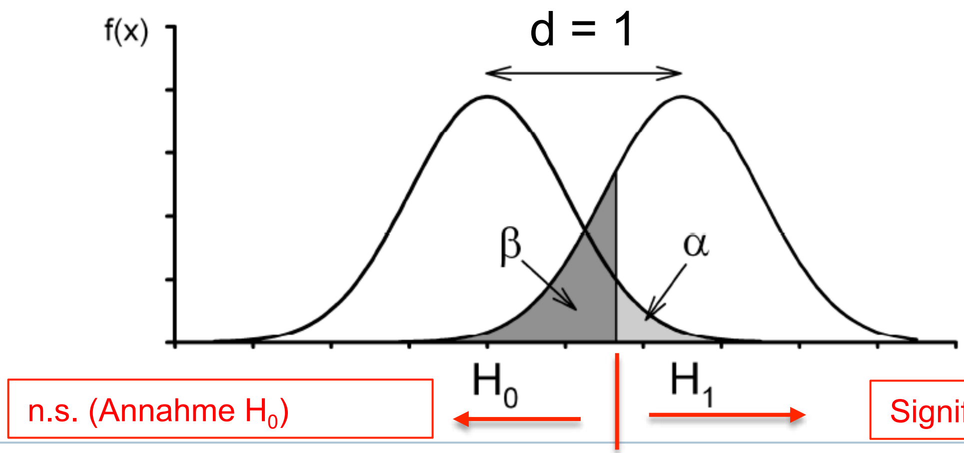


Entscheidungsgrenze wird durch Signifikanzniveau festgelegt (meist p = 0.05)

Statistische Signifikanz und Teststärke

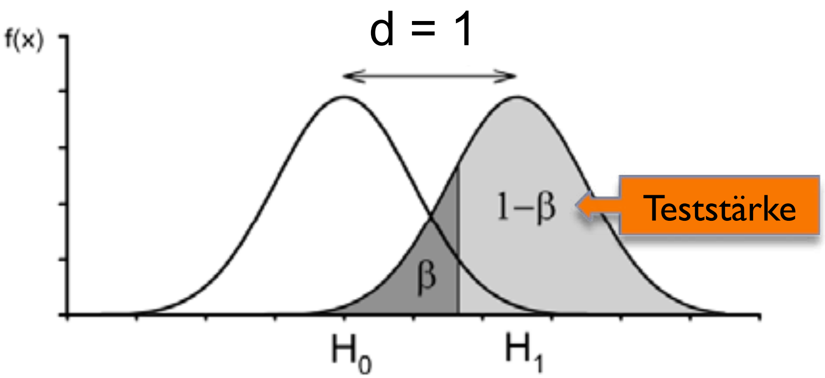
- Ziel einer guten Studie:
  - Wenn die H_0 in Wirklichkeit zutrifft:
    - Wahrscheinlichkeit sollte gering sein, dass ich fälschlicherweise die H_1 annehme
      - Signifikanzniveau $p < 0.05$
      - Fehler 1.Art (false positive)
  - Wenn die H_1 in Wirklichkeit zutrifft
    - Wahrscheinlichkeit sollte gering sein, dass ich fälschlicherweise die H_0 annehme
      - β-Fehler sollte klein sein (Fehler 2. Art, false negative)
      - Teststärke (1-β) > 80%

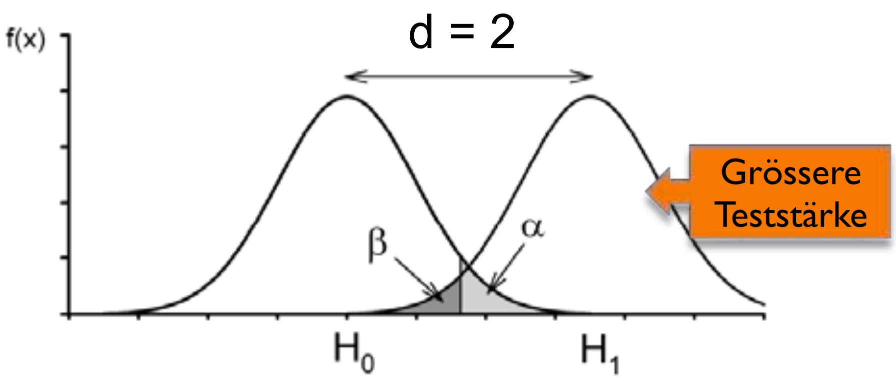
Entscheidungsgrenze wird durch Signifikanzniveau festgelegt (meist p = 0.05)



Teststärke und Effektstärke

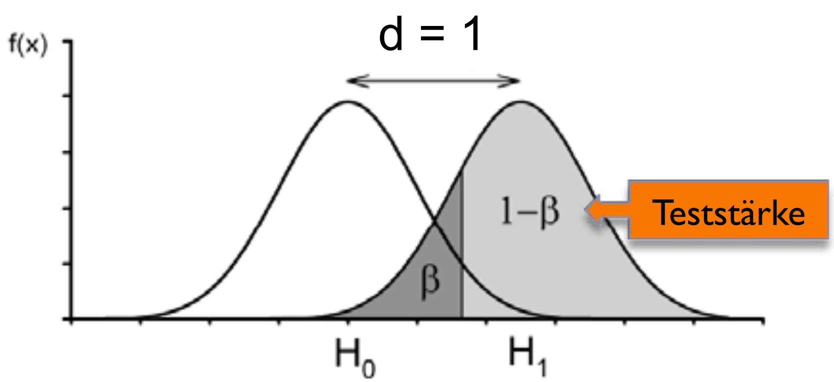
- Teststärke ist immer (!!!) abhängig von der Effektstärke
  - Effektgrösse in der Population
  - Wenn nicht bekannt: inhaltlich relevanter (gesuchter) Effekt
    - Muss vom Wissenschaftler festgelegt werden
    - Meta-Analysen, vorangegangen Studien, Pilotstudien, inhaltliche Überlegungen
  - Je grösser der gesuchte Effekt, desto weniger Probanden sind notwendig, um eine ausreichende Teststärke zu erzielen
  - Je kleiner der gesuchte Effekt, desto mehr Probanden sind notwendig, um eine ausreichende Teststärke zu erzielen

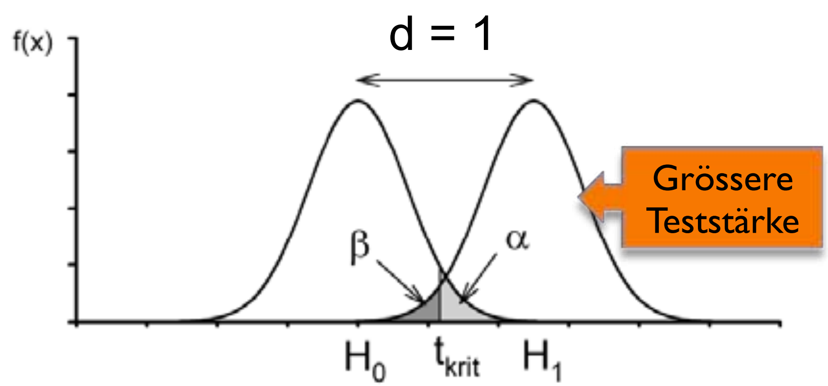




Teststärke und Stichprobengrösse

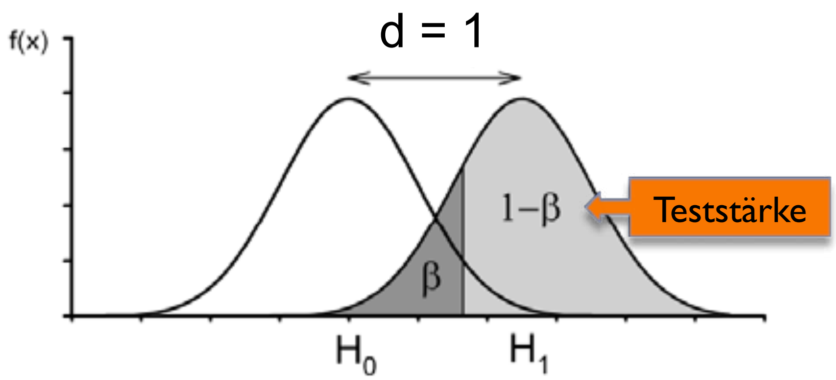
- Teststärke ist abhängig von der Stichprobengrösse
  - Bei gleicher Effektstärke
    - Je grösser die Stichprobe, desto grösser die Teststärke
      - Je grösser die Stichprobe, desto grösser die Wahrscheinlichkeit, einen bestimmten Effekt zu finden (falls er existiert)
      - Stichprobenkennwerteverteilungen werden schmaler
    - Je kleiner die Stichprobe, desto kleiner die Teststärke
      - Je kleiner die Stichprobe, desto kleiner die Wahrscheinlichkeit, einen bestimmten Effekt zu finden (falls er existiert)

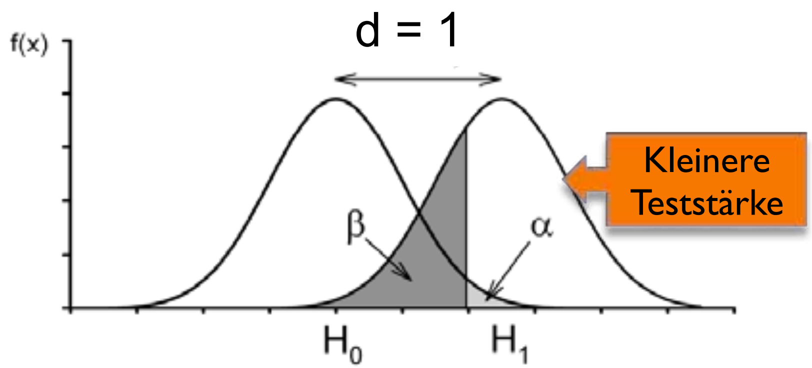




Teststärke und Signifikanzniveau

- Teststärke ist abhängig von dem gewählten Signifikanzniveau
  - Je kleiner das Signifikanzniveau, desto kleiner die Teststärke
    - Je kleiner der Fehler 1. Art (α), desto grösser der Fehler 2. Art (β)
      - Je weniger Fehler ich mache möchte, die H_0 abzulehnen wenn sie eigentlich gilt, desto eher mach ich den Fehler, die H_0 anzunehmen, obwohl in Wirklichkeit die H_1 gilt.
      - Je weniger „false positives", desto mehr „false negatives".
  - Je grösser das Signifikanzniveau, desto grösser die Teststärke



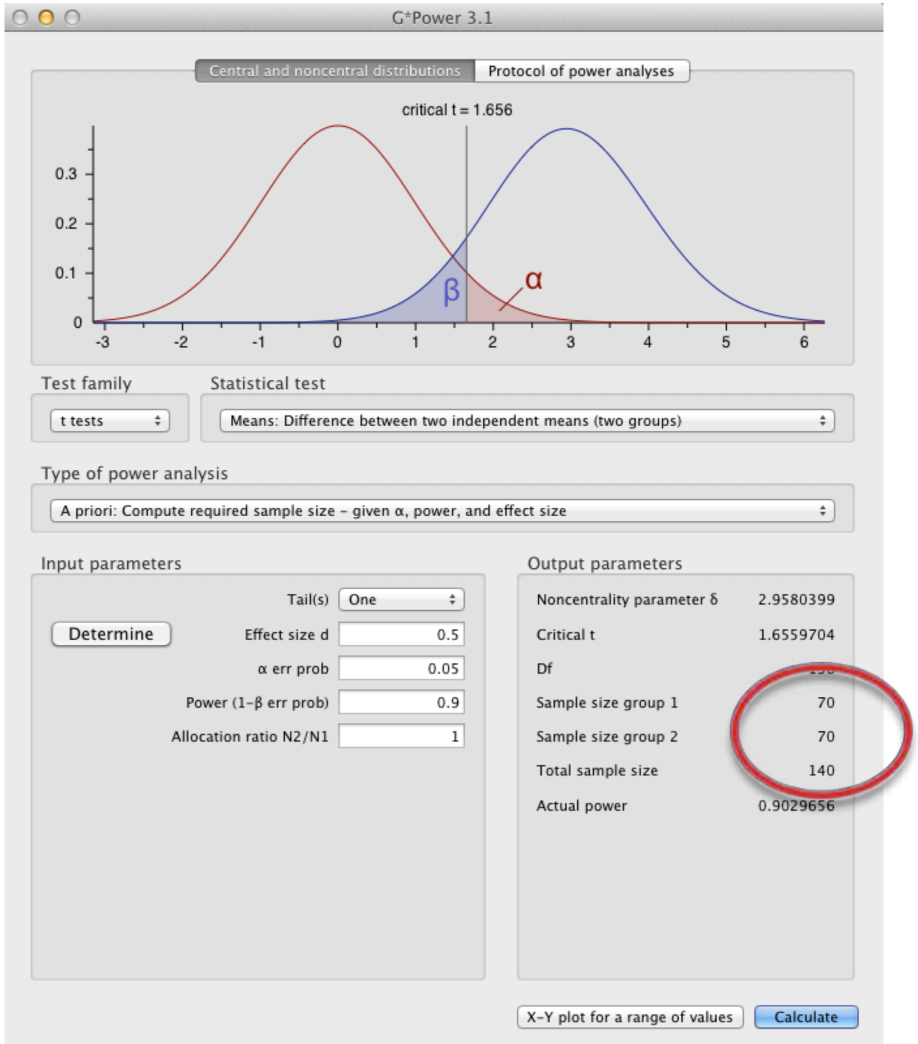


Teststärke

- Wann ist die Teststärke wichtig?
  - Vor der Durchführung eines Experiments
    - A priori Teststärkebestimmung
    - Festlegung der gewünschten Teststärke eines Experiments
      - Z.B. 80% oder 90%
    - Festlegung der relevanten (gesuchten) Effektstärke
      - Meta-Analysen, vorangegangen Studien, Pilotstudien, inhaltliche Überlegungen
    - Festlegung des Signifikanzniveaus (z.B. p = 0.05)
    - Berechnung des erforderlichen Stichprobenumfangs
      - Stichprobenumfangsplanung
  - Nach der Durchführung eines Experiments
    - A posteriori Teststärkenbestimmung
      - Nur wichtig bei nicht signifikanten Ergebnissen !!!!
    - Festlegung der relevanten (gesuchten) Effektstärke erforderlich
      - Meta-Analysen, vorangegangen Studien, Pilotstudien, inhaltliche Überlegungen
      - Auch möglich: empirischer Effekt aus Stichprobe (wenn empirische Effektgrösse inhaltlich relevante Grösse erreicht hat!!!)

A priori Teststärkenbestimmung

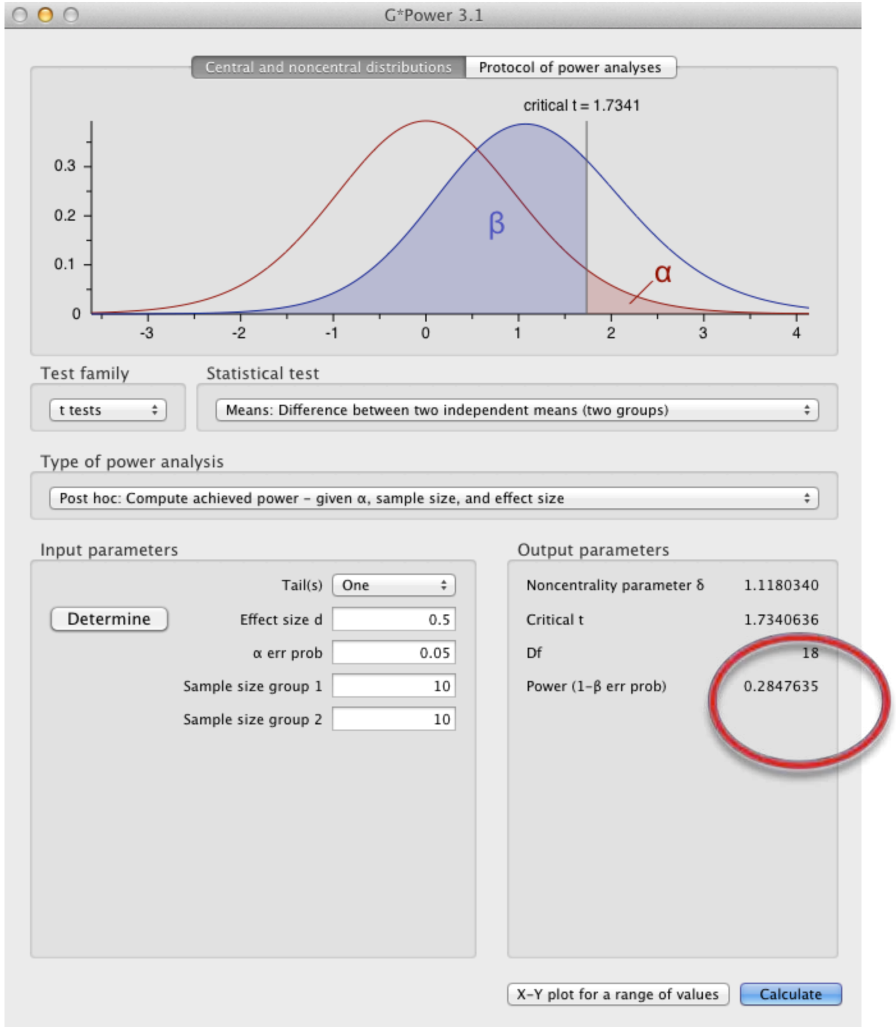
- Beispiel:
  - Relevanter (gesuchter) Effekt:
    - d = 0.5
  - Gewünschte Teststärke:
    - $1-\beta$ = 90%
  - Signifikanzniveau:
    - α = 0.05
- Berechnung über Software (G*Power)
  - Frei verfügbar unter:
    - http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/



Optimaler Stichprobenumfang:
N = 140 (70 Personen pro Gruppe)

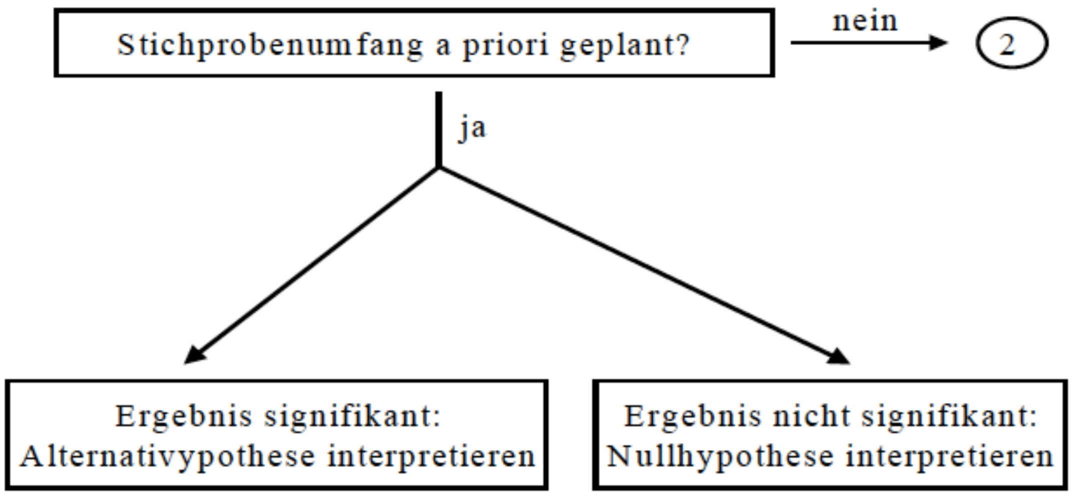
A posteriori Teststärkenbestimmung

- Beispiel
  - Problemlösefähigkeit
    - N = 10 pro Gruppe
    - α = 0.05
    - Ergebnis nicht signifikant (P = 0.25)
  - Inhaltlicher relevanter Effekt: d = 0.05
- Berechnung mit G*Power
  - Post hoc statistical power
- Teststärke ist 28.5%
  - Die Wahrscheinlichkeit, einen Effekt der Grösse d = 0.5 zu finden, war 28%.
- H1 kann <u>nicht</u> abgelehnt werden
  - Wahrscheinlichkeit eines „false negatives“ ist zu gross (72%).
- Keine Entscheidung möglich
  - Nutzloses Experiment



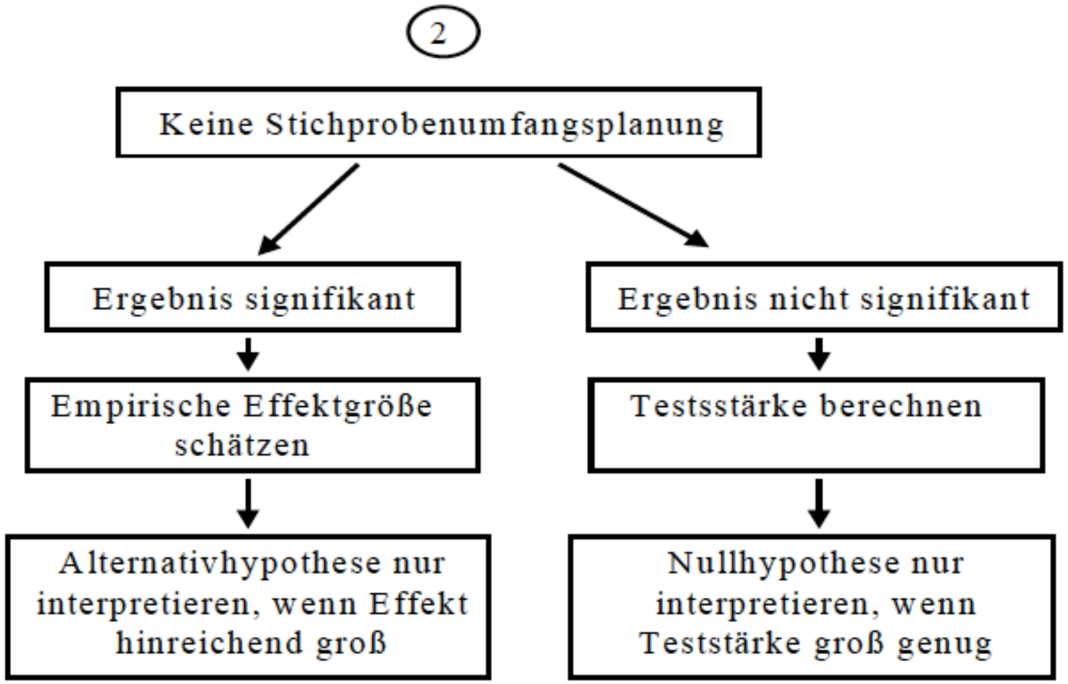
Ablauf

- Nach a priori Teststärkenbestimmung:
  - Ergebnis eines statistischen Tests eindeutig interpretierbar
    - Signifikantes Ergebnis = Annahme eines inhaltlich relevanten Effekts
    - Nicht signifikantes Ergebnis: Ablehnung eines inhaltlichen Effekts mit ausreichender Teststärke
      - Wahrscheinlichkeit für ein false negative (β -Fehler) ist ausreichend klein



Ablauf

- Ohne a priori Teststärkenbestimmung
  - Ergebnis nicht eindeutig interpretierbar
    - Signifkant: Ist der gefundene Effekt auch inhaltlich relevant? (Effektstärke berechnen)
    - Nicht signifikant: War die Teststärke gross genug, einen inhaltlich relevanten Effekt überhaupt zu finden? (A posteriori Teststärkebestimmung)



Take-Home Messages

- Entscheidungsregel
  - Ist die Wahrscheinlichkeit eines empirischen Ergebnisses kleiner als das Signifikanzniveau α, dann lehne ich die H0 ab
    - Fehler 1. Art (α-Fehler): Ablehnung der H0, obwohl H0 gilt (false positive)
    - Fehler 2. Art (β-Fehler): Ablehnung der H1, obwohl H1 gilt (false negative)
- Teststärke
  - Wahrscheinlichkeit, dass ein bestimmter Effekt gefunden (signifikant) wird, falls er wirklich existiert (Teststärke = $1-\beta$)
  - Teststärke abhängig von Grösse des Effekts, Stichprobengrösse und Signifikanzniveaus α
  - Teststärke für inhaltlich relevanten Effekt sollte mindestens 80% (besser 90%) betragen
- A priori Teststärkenbestimmung
  - Festlegung inhaltlich relevanter Effekt, gewünschte Teststärke und Signifikanzniveaus α
  - Berechnung des optimalen Stichprobenumfangs (Stichprobenumfangsplanung)
  - Ergebnisse eindeutig interpretierbar
- A posteriori Teststärkenbestimmung
  - Bei Studien ohne Stichprobenumfangsplanung notwendig, wenn das Ergebnis nicht signifikant wurde
  - Ablehnung der H1 nur bei ausreichender Teststärke möglich.

Vielen Dank für Ihre Aufmerksamkeit

Effektstärke

- Berechnung von d aus Daten

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}^2_{x_1} + \hat{\sigma}^2_{x_2}}{2}}}$$

Gruppenstatistiken

	Gruppe	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Problemlösefähigkeit	15 min Pause	50	41,6600	8,95478	1,26640
	"keine Pause"	50	35,4400	9,04673	1,27940

$$d = \frac{41.66 - 35.44}{\sqrt{\frac{8.95^2 + 9.05^2}{2}}} = 0.69$$

$\Rightarrow$ Mittlerer Effekt

$\Rightarrow$ $P = 0.001$

Effektstärke

- Berechnung von d aus Daten

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}^2_{x_1} + \hat{\sigma}^2_{x_2}}{2}}}$$

Gruppenstatistiken

	Gruppe	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Problemlösefähigkeit	15 min Pause	25	42,0800	10,07439	2,01488
	"keine Pause"	25	36,1200	9,04765	1,80953

$$d = \frac{42.08 - 36.12}{\sqrt{\frac{10.07^2 + 9.05^2}{2}}} = 0.62$$

$\Rightarrow$ Mittlerer Effekt

$\Rightarrow P = 0.033$

Effektstärke

- Berechnung von d aus Daten

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}^2_{x_1} + \hat{\sigma}^2_{x_2}}{2}}}$$

Gruppenstatistiken

	Gruppe	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Problemlösefähigkeit	15 min Pause	10	41,3000	13,18290	4,16880
	"keine Pause"	10	34,8000	11,19325	3,53962

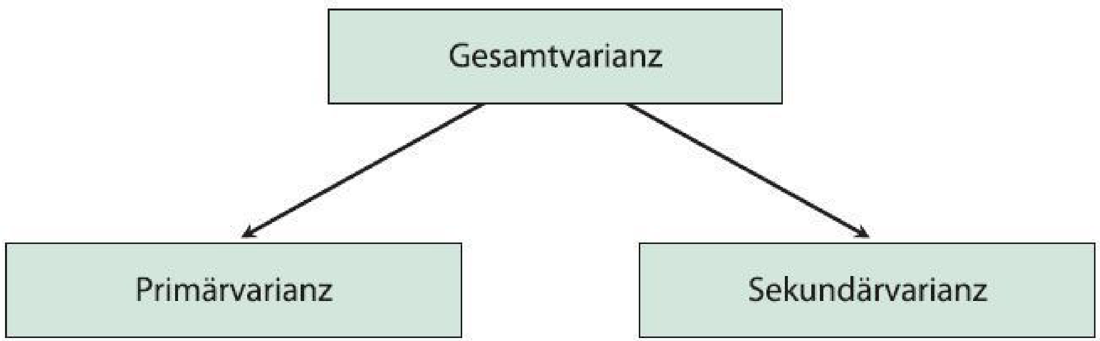
$$d = \frac{41.3 - 34.8}{\sqrt{\frac{13.18^2 + 11.19^2}{2}}} = 0.53$$

$\Rightarrow$ Mittlerer Effekt

$\Rightarrow$ $P = 0.25$ *n.s.*

Effektstärke

- **Effektstärke als Mass der Varianzaufklärung**
  - Wie viel Prozent der gesamten Unterschiede zwischen allen Versuchspersonen wird durch die experimentelle Manipulation (UV) aufgeklärt?
  - Anteil der aufgeklärten Varianz an der Gesamtvarianz
    - Synonyme für aufgeklärte Varianz:
      - Primärvarianz, Effektvarianz, systematische Varianz



Experimentelle Richtlinie: Maximiere die Primärvarianz und minimiere die Sekundärvarianz!

Effektstärke

- Effektstärke aus Mass der Varianzaufklärung
  - Primärvarianz/ Gesamtvarianz
    - Gesamtvarianz besteht aus Primär + Sekundärvarianz
    - Synonym: Gesamtvarianz besteht aus systematischer Varianz und Fehlervarianz
  - Effektmass auf Populationsebene
    - Ω^2 (Omega Quadrat)
      - Wird in der Literatur wenig verwendet

$$\Omega^2 = \frac{\sigma^2_{\text{sys}}}{\sigma^2_{\text{Gesamt}}} = \frac{\sigma^2_{\text{sys}}}{\sigma^2_{\text{sys}} + \sigma^2_{\text{x}}}$$

  - Effektmass auf Stichprobenebene
    - η^2 (eta Quadrat)
      - QS: Quadratsummen
        - (Mass für Varianz auf Stichprobenebene)
      - Bei mehreren Faktoren
        - Partielles η^2

$$\eta^2 = \frac{QS^2_{\text{sys}}}{QS^2_{\text{Gesamt}}} = \frac{QS^2_{\text{sys}}}{QS^2_{\text{sys}} + QS^2_{\text{x}}}$$

Effekstärken

- Konventionen
  - $\Omega^2 = 0.01$: kleiner Effekt
    - 1% aufgeklärte Varianz
  - $\Omega^2 = 0.06$: mittlerer Effekt
    - 6% aufgeklärte Varianz
  - $\Omega^2 = 0.14$: grosser Effekt
    - 14% aufgeklärte Varianz
    - η^2 fällt häufig etwas grösser aus (Überschätzung des Populationseffekts)
      - Konventionen gelten nur für nicht-messwiederholte Mittelwertsvergleiche
      - Basieren auf Cohen (1988)
  - Berechnung über Kennwert des t-Tests (t-Wert)
    - Kann bei Varianzanalyse von SPSS ausgegeben werden

$$f_S^2 = \frac{t^2}{df} \quad \text{->} \quad \eta^2 = \frac{f_S^2}{1 + f_S^2}$$

Beispiel

- Effektstärkeberechnung aus den Daten

Gruppenstatistiken

	Gruppe	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Problemlösefähigkeit	15 min Pause	50	41,6600	8,95478	1,26640
	"keine Pause"	50	35,4400	9,04673	1,27940

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
Problemlösefähigkeit	Varianzen sind gleich	,020	,887	3,455	98	,001	6,22000	1,80017	2,64761	9,79239
	Varianzen sind nicht gleich			3,455	97,990	,001	6,22000	1,80017	2,64761	9,79239

$$f^2 = \frac{3.455^2}{98} = 0.12 \qquad \eta^2 = \frac{0.12}{1+0.12} = 0.11$$

- Interpretation:
  - Es wurden ca. 11% der Gesamtvarianz in der Problemlösefähigkeit durch den Faktor „Pausen“ erklärt. Dies ist ein mittlerer Effekt.