Power: Gliederung

1. Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson
2. Power und ihre Determinanten
3. Vorgehen bei der Poweranalyse
4. Power-Berechnungen mit dem Programm G*Power
5. Power in der Empirie
6. Effektstärken

Power: Literatur

Einführende Literatur

📖 Bortz, J. & Schuster, Ch. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Auflage). Berlin: Springer. [Kap. 7.5]

📖 Diehl, J. M. & Arbinger, R. (2001). *Einführung in die Inferenzstatistik* (3. Auflage). Eschborn bei Frankfurt: Klotz Verlag. [Kap. 3.2 & 29]

📖 Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz. [Kap. 8.2 & 8.7]

Weiterführende Literatur

📖 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Logik der Inferenzstatistik

- Das bisher praktizierte Vorgehen beim Hypothesentesten, auch als NHST (Null Hypothesis Significance Testing) bezeichnet, ist tatsächlich eine Kombination (ein „Hybrid“; Gigerenzer, 1993) aus zwei theoretischen Vorstellungen darüber, wie man beim Signifikanztesten vorgehen sollte: Der von Fisher sowie der von Neyman und Pearson.

Ronald A. Fisher (1890-1962)

- Sir Ronald Aylmer Fisher (engl. Mathematiker & Genetiker, 1890-1962) hat folgendes Vorgehen vorgeschlagen, das über die Zeit (1925, 1935, 1956) leichte Modifikationen erfahren hat:
  - Formuliere die statistische Nullhypothese H_0.
  - Errichte die Stichprobenkennwerteverteilung unter der H_0 und bestimme den p-Wert (Überschreitungswahrscheinlichkeit) der berechneten Prüfgröße.
  - 1935 hat Fisher empfohlen, den p-Wert mit einem Signifikanzniveau α, das per Konvention z.B. auf 0.05 oder 0.01 festgelegt werden kann, zu vergleichen. Ist $p < \alpha$, so wird die H_0 zurückgewiesen und es resultiert ein statistisch signifikanter Effekt. Andernfalls kann man keine Schlüsse ziehen (auch nicht, dass die H_0 richtig ist).
  - 1956 hat er diese Position revidiert und empfohlen, keine Entscheidung zu treffen, sondern im Anschluss an die Auswertung nur den exakten p-Wert zu berichten.

Logik der Inferenzstatistik

Jerzy Neyman (1894-1981)

Egon S. Pearson (1895-1980)

- Jerzy Neyman (polnischer Mathematiker, 1894-1981) and Egon S. Pearson (britischer Mathematiker, 1895-1980, Sohn von Karl Pearson) entwickelten folgendes Vorgehen beim Hypothesentesten:
  - Übertrage die wissenschaftliche Hypothese in zwei sich ausschließende Hypothesen: Die Null- und Alternativhypothese.
  - Lege vor der Studie aufgrund von Kosten-Nutzen Überlegungen den α- und β-Fehler und daraufhin den optimalen Stichprobenumfang *n* fest.
  - Entscheide dich auf der Basis des Tests für die H_0 oder die H_1: Wenn die Prüfgröße in die Zurückweisungsregion der H_0 fällt, akzeptiere die H_1 (d.h. verhalte dich so, als wenn die H_1 richtig wäre); andernfalls akzeptiere die H_0.
  - Durch die vorherige Festlegung kann das Signifikanzniveau α als Fehlerhäufigkeit „in the long run“ betrachtet werden. Wenn die H_0 korrekt ist und man zieht immer wieder Zufallsstichproben aus der Population, so wird man die H_0 in 5% der Fälle fehlerhafter Weise zurückweisen und sich für die H_1 entscheiden.

Entscheidung für	Realität	
	H_0	H_1
H_0	$1-\alpha$	β-Fehler
H_1	α-Fehler	$1-\beta$=Power

Logik der Inferenzstatistik

- Aus der folgenden tabellarischen Darstellung wird nochmals der Hybrid-Charakter der NHST-Logik deutlich, die insgesamt stärker durch die Fisher-Logik geprägt ist:

Vorgehen beim NHST	Quelle
Formulierung von H_0 und H_1	Neyman - Pearson
Entweder: Signifikanzniveau α wird vor der Auswertung festgelegt. Ist $p < \alpha$, so ist das Ergebnis statistisch signifikant.	früher Fisher
Oder: Es wird geschaut, ob der p-Wert jeweils kleiner als 0.05, 0.01 oder 0.001 ist und dies dann jeweils berichtet, $p < .05$ (bzw. *), $p < .01$ (**), $p < .001$ (***) oder andernfalls *ns* (nicht signifikant)	weder Neyman - Pearson noch Fisher
Resultiert kein statistisch signifikantes Ergebnis, so kann die Null-hypothese nicht bestätigt werden, sondern lediglich „beibehalten“ werden.	Fisher
β-Fehler und Power werden keine Beachtung geschenkt.	Fisher

- Im Folgenden werden wir die Neyman-Pearson Logik näher betrachten, die nicht nur den α-Fehler, sondern auch den β-Fehler berücksichtigt.

Logik der Inferenzstatistik

- 90 Korrelationen zwischen 18 Risikoeinschätzungen (Zeilen) und den Big-5 Persönlichkeitsmerkmalen (Spalten) aus Soane, Dewberry & Narendran (2010).

Man beachte: Wären in der Population alle Korrelationen $\rho = 0$, so wären hier $90 \cdot \alpha$ statistisch signifikante Korrelationen zu erwarten (also z.B. $0.05 \cdot 90 = 4.5$)! Hier sind es 23.

Table 1. Descriptive statistics and correlations between personality and risk scales.

	Mean (SD)	Alpha	Emotionality	Extraversion	Openness	Agreeableness	Conscientiousness
Likelihood of ethical risk-taking	1.94 (0.62)	0.74	0.15	0.096	0.13	−0.32**	−0.43**
Benefits of ethical risk-taking	2.10 (0.52)	0.79	0.16	0.20*	0.18*	−0.20*	−0.37**
Costs of ethical risk-taking	3.68 (0.62)	0.82	0.07	−0.12	0.02	0.09	0.07
Likelihood of investment risk	2.68 (0.69)	0.78	0.09	0.04	0.03	−0.15	0.08
Benefits of investment risk	3.09 (0.72)	0.73	0.02	0.16*	0.03	−0.10	0.09
Costs of investment risk	2.91 (0.89)	0.74	0.09	−0.04	0.03	0.08	−0.02
Likelihood of gambling risk	1.55 (0.78)	0.84	0.08	−0.02	0.17*	−0.29**	−0.35**
Benefits of gambling risk	1.86 (0.89)	0.88	0.16*	0.13	0.16	−0.15	−0.27**
Costs of gambling risk	4.07 (0.91)	0.87	−0.10	0.02	0.07	0.12	0.05
Likelihood of health and safety risk	2.41 (0.70)	0.68	−0.05	0.25**	0.24**	−0.12	−0.26**
Benefits of health and safety risk	1.57 (0.47)	0.65	−0.05	0.16	0.14	−0.01	−0.07
Costs of health and safety risk	3.83 (0.64)	0.78	0.26**	−0.06	0.03	0.01	−0.09
Likelihood of recreational risk	2.70 (0.95)	0.84	−0.14	0.16	0.18*	0.01	−0.13
Benefits of recreational risk	2.78 (0.84)	0.82	−0.15	0.10	0.11	0.09	−0.05
Costs of recreational risk	3.26 (0.68)	0.75	0.21*	−0.06	−0.11	−0.03	−0.02
Likelihood of social risk-taking	3.28 (0.55)	0.61	−0.09	0.24**	0.40**	−0.15	−0.01
Benefits of social risk-taking	3.07 (0.47)	0.50	−0.05	0.07	0.18*	−0.06	0.08
Costs of social risk-taking	2.30 (0.52)	0.67	0.24**	−0.01	0.05	−0.26**	−0.11

Notes: n=154; *correlation is significant at the 0.05 level (two–tailed); **correlation is significant at the 0.01 level (two–tailed).

Power: Gliederung

1 Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson

2 Power und ihre Determinanten

3 Vorgehen bei der Poweranalyse

4 Power-Berechnungen mit dem Programm G*Power

5 Power in der Empirie

6 Effektstärken

Power

- Die Power (Teststärke) eines Tests ...
  - entspricht der Komplementärwahrscheinlichkeit des β-Fehlers $1-\beta$
  - bezeichnet die Wahrscheinlichkeit, sich für die H_1 zu entscheiden, wenn sie gilt (= die Wahrscheinlichkeit einen Effekt zu finden, wenn er existiert = die Wahrscheinlichkeit, mit der ein Test korrekter Weise ein statistisch signifikantes Ergebnis produziert)
  - kann nicht so direkt festgesetzt werden wie das α-Risiko.
  - kann nur bestimmt werden, wenn man vorab in der H_1 festlegt, wie stark der Effekt mindestens sein soll (also die Abweichung von der Aussage in der H_0 quantifiziert). Wenn der Effekt dann (mindestens) so groß ist, wird er mit der Wahrscheinlichkeit $1-\beta$ auch gefunden (d.h. es resultiert ein statistisch signifikantes Ergebnis).
  - hängt nicht nur von der Stärke des Effektes sondern von weiteren Einflussgrößen ab, darunter der Stichprobengröße n und dem gewählten α.

Entscheidung für	Realität	
	H_0	H_1
H_0	$1-\alpha$	β-Fehler
H_1	α-Fehler	$1-\beta$=Power

Power

- Nehmen wir an, wir wollen testen, ob Jurastudierende intelligenter sind als andere Studierende. Nehmen wir ferner an, es sei bekannt, dass die mittlere Intelligenz von Studierenden bei 115 mit $\sigma = 15$ liegt.

- Mittels des Ein-Stichproben z-Tests können wir die gerichtete Hypothese z.B. einseitig testen: H_0: $\mu \leq \mu_0 = 115$, H_1: $\mu > 115$. In einer Stichprobe von 100 Jurastudierenden habe sich ein mittlerer Intelligenzwert von 117 ergeben. Als Prüfgröße im Ein-Stichproben z-Test resultiert

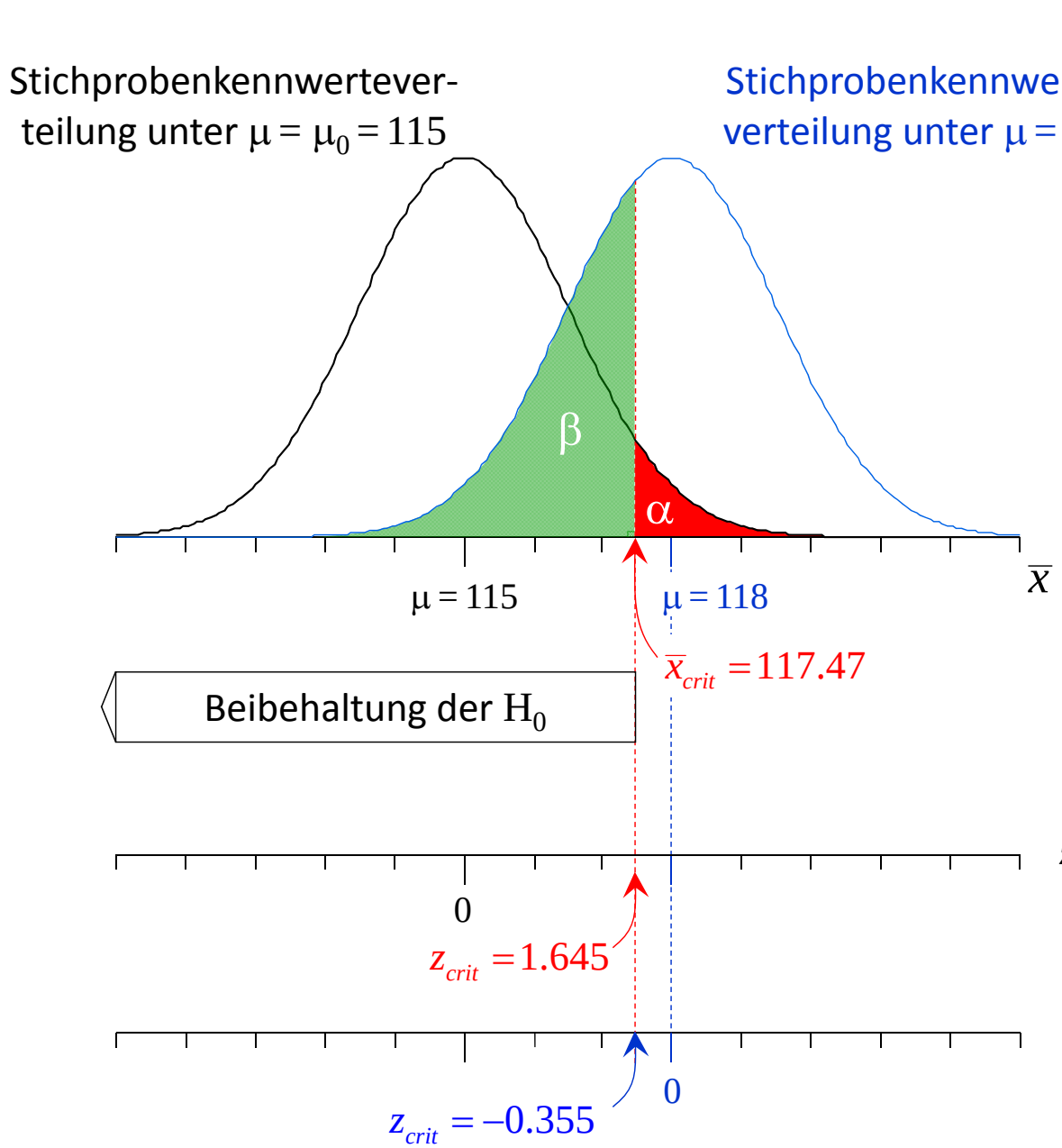
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{117 - 115}{15 / \sqrt{100}} = 1.33$$

- Bei $\alpha = 0.05$ ergibt sich $z_{crit} = z_{1-\alpha} = z_{0.95} = 1.645 > 1.33$ und damit kein statistisch signifikanter Unterschied.

- Frage: Wie groß ist die Wahrscheinlichkeit, einen β-Fehler begangen zu haben? Oder anders gefragt: Wie groß war die Power des Tests, d.h. die Wahrscheinlichkeit, einen bestehenden Intelligenzunterschied auch aufzudecken?

- Diese Frage könnten wir beantworten, wenn wir die Intelligenz der Population der Jurastudierenden kennen würden. Nehmen wir an, wir wüssten, sie wäre $\mu = 118$.

Power

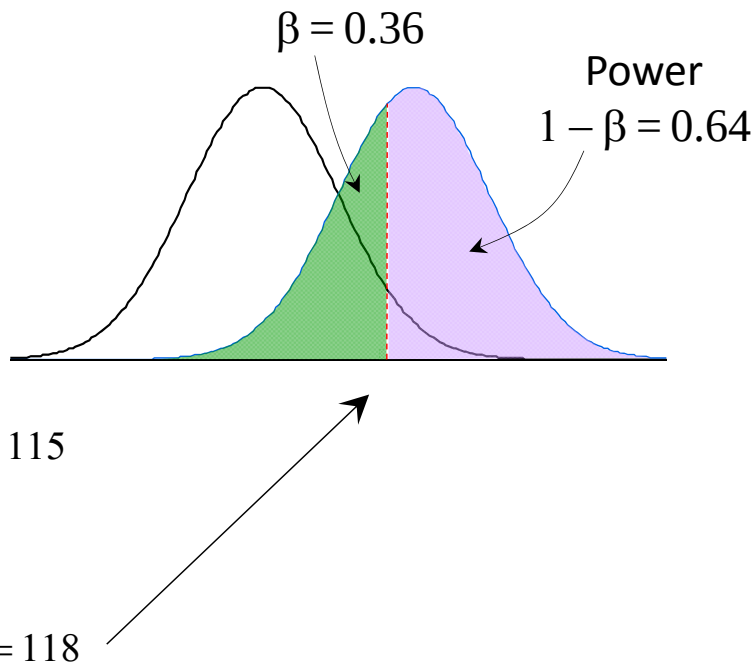


$$z_{crit} = \frac{\bar{x}_{crit} - \mu_0}{\sigma / \sqrt{n}} \Leftrightarrow \bar{x}_{crit} = \mu_0 + z_{crit} \cdot \sigma / \sqrt{n}$$

$$\Rightarrow \bar{x}_{crit} = 115 + 1.645 \cdot 15 / \sqrt{100} = 117.47$$

$$z_{crit} = \frac{\bar{x}_{crit} - \mu_0}{\sigma / \sqrt{n}} = \frac{117.47 - 118}{15 / \sqrt{100}} = -0.355$$

$$\Rightarrow p = 0.361$$



Power

- Wenn wir wüssten, dass die mittlere Intelligenz der Jurastudierenden $\mu = 118$ betragen würde, könnten wir nicht nur die Stichprobenkennwerteverteilung betrachten, die unter der Gültigkeit der H_0 mit dem Mittelwert 115 resultiert, sondern auch die unter $\mu = 118$.
- Wir erkennen dann, dass sich beide Verteilungen beträchtlich überlappen. Ziehen wir Stichproben aus der Population mit $\mu = 118$, so fallen auch einige davon in den Bereich der Beibehaltung der H_0. Wird eine solche Stichprobe gezogen, so erfolgt eine falsche Entscheidung, nämlich die Beibehaltung der H_0. Wir würden einen β-Fehler begehen, und den in der Population bestehenden Unterschied durch unseren Test nicht aufdecken.
- Wie groß ist die Wahrscheinlichkeit für diesen β-Fehler? Um dies zu bestimmen, müssen wir den Flächenanteil der Verteilung mit $\mu = 118$ links von dem kritischen Wert bestimmen. Der kritische z-Wert beträgt in der H_0-Verteilung 1.645, der umgerechnet dem kritischen Intelligenzwert von $\bar{x}_{crit} = \mu_0 + z_{crit} \cdot \sigma / \sqrt{n} = 115 + 1.645 \cdot 15 / \sqrt{100} = 117.47$ entspricht.
- In der Verteilung mit $\mu = 118$ (und gleichem Standardfehler) entspricht diesem kritischen x-Wert der z-Wert

$$z_{crit} = \frac{117.47 - 118}{15 / \sqrt{100}} = -0.355$$

- In der Standardnormalverteilung entspricht $z = -0.355$ eine Wahrscheinlichkeit von 0.361. Die Wahrscheinlichkeit für einen β-Fehler (grüne Fläche) beträgt also 0.361 und die Power ist entsprechend $1 - 0.361 = 0.639$.

Power

- Tatsächlich kennen wir den Populationsmittelwert der Jurastudierenden nicht. Auf der obigen Grundlage können wir aber die Wahrscheinlichkeit eines β-Fehlers bzw. die Power $1 - \beta$ bestimmen, wenn wir die beiden folgenden spezifischen Hypothesen testen:

  $H_0: \mu = \mu_0 = 115$ und $H_1: \mu = \mu_1 = 118$.

- Bei den üblicherweise getesteten zusammengesetzten Hypothesen der Form

  $H_0: \mu \leq 115$ und $H_1: \mu > 115$

  ist eine Powerbestimmung hingegen nicht möglich, da die Power für jeden Wert größer 115 anders ist.

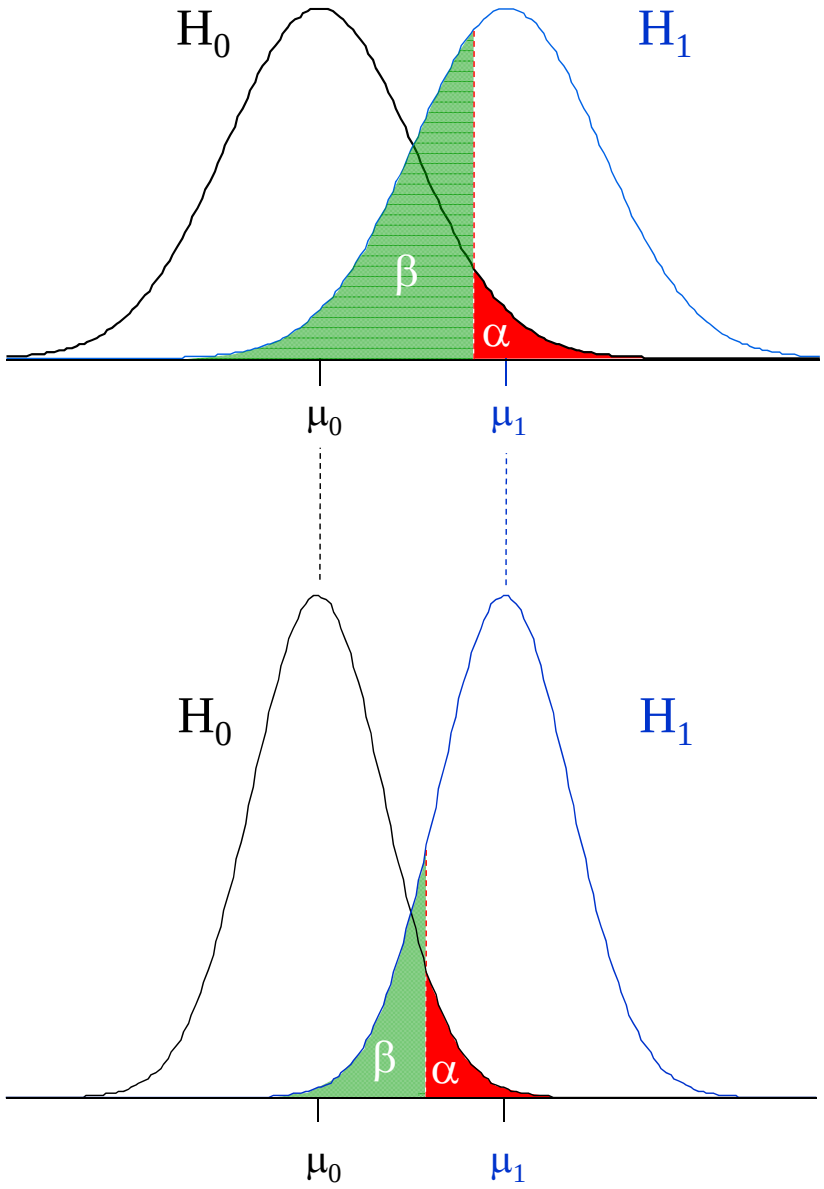
- Für die Powerbestimmung ist dabei nicht der absolute Wert in der H_1 bedeutsam, sondern nur die Differenz (der Abstand) zwischen den Populationswerten unter der H_0 und unter der H_1, die wir bezeichnen als $\varepsilon_1 = \mu_1 - \mu_0$ (z.B. $\varepsilon_1 = 3$).

- Ist die spezifische Alternativhypothese wahr, so ist ε_1 gleich dem wahren Populationseffekt $\varepsilon = \mu - \mu_0$, also der Differenz zwischen dem tatsächlichen (unbekannten) Populations-mittelwert μ und dem unter der H_0 behaupteten Mittelwert μ_0. Sollte gelten, dass $\varepsilon > \varepsilon_1$, so wäre die Power sogar noch höher (wie wir gleich sehen werden).

Power

- Mit ε_1 muss angegeben werden, wie groß der Effekt mindestens sein muss (und in welcher Richtung er liegen muss), damit er praktisch bedeutsam ist.
- Z.B. könnte man im obigen Beispiel festlegen, dass er mindestens 3 IQ-Punkte betragen müsste. In einem Gewichtsreduktionstraining könnte man eine Abnahme um mindestens 5 kg fordern. Oder in einem Gedächtnistraining eine um mindestens 10 Items bessere Leistung in einem Gedächtnistest. Liegt der Effekt darunter, so wird er nicht als „praktisch“ bedeutsam erachtet.
- Die begründete Festlegung dieses Mindesteffektes stellt eine besondere Herausforderung bei der Power-Bestimmung dar. Wir kommen darauf zurück.
- Im Folgenden wird nun zunächst gezeigt, dass die Power mit den folgenden Größen in einem systematischen Zusammenhang steht:
  - der Stichprobengröße n
  - der Streuung des Merkmals in der Population σ
  - der Stärke des (mindestens bedeutsamen) Effekts
  - dem gewählten Signifikanzniveau α

Power: Determinanten Stichprobengröße und Streuung



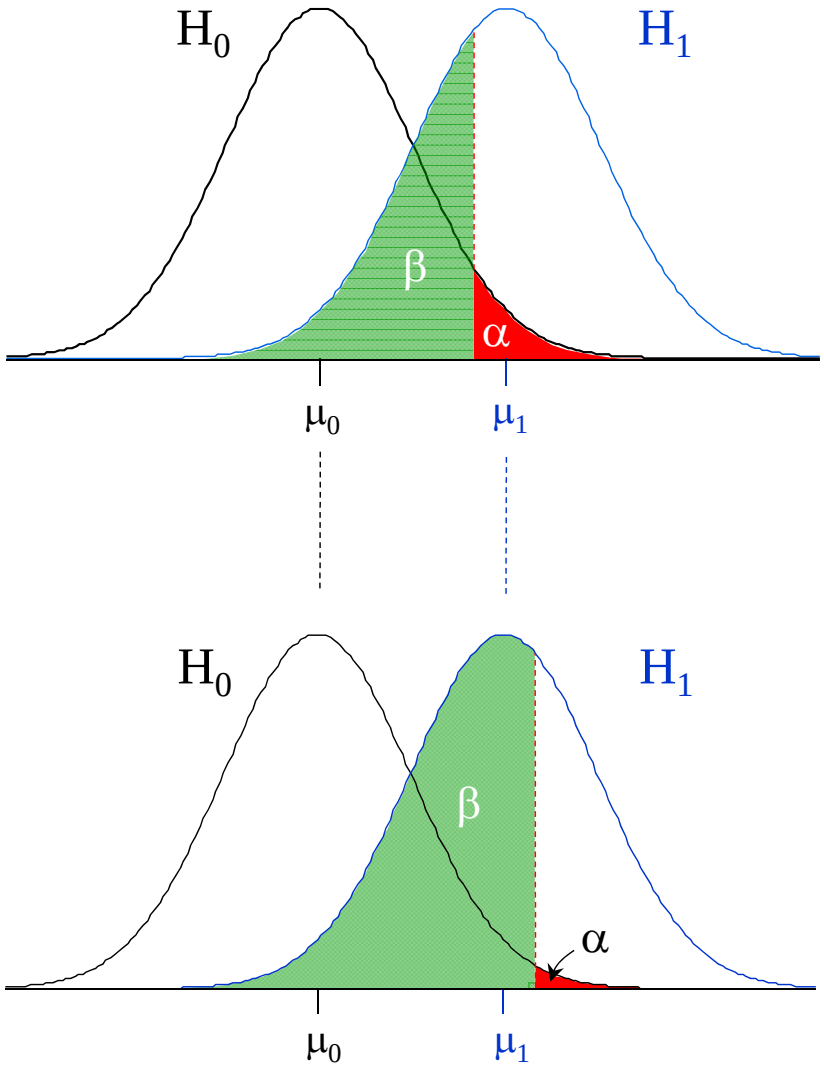
Konstant: μ_0, μ_1, σ, α, einseitige Testung
Variant: $n = 100$ (oben) vs. $n = 200$ (unten)

- Je größer n, desto größer die Power.
- Wenn n größer wird, dann verringert sich der Standardfehler, d.h. die Streuung der Stichprobenkennwerteverteilungen:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Daraus folgt auch: Je kleiner die Streuung des Merkmals σ, desto größer die Power.

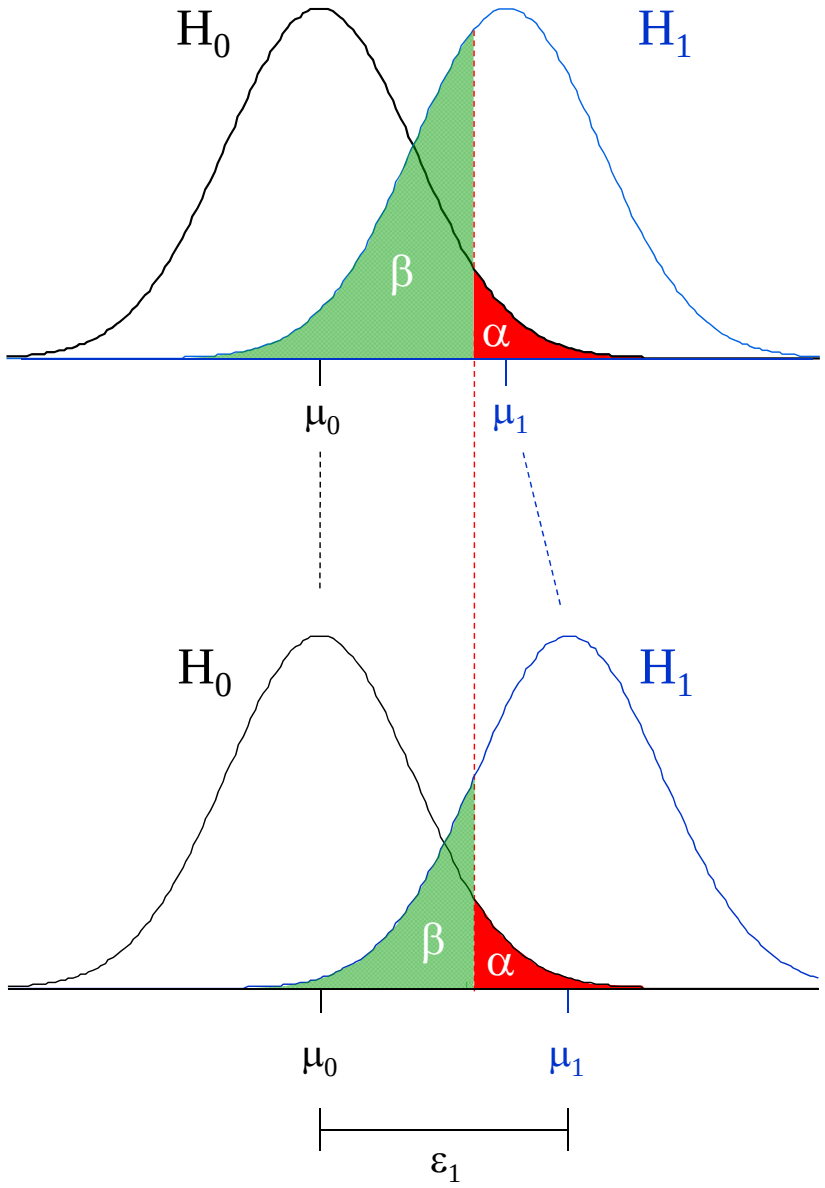
Power: Determinante Signifikanzniveau



Konstant: μ_0, μ_1, σ, n, einseitige Testung
Variant: $\alpha = 0.05$ (oben) vs. $\alpha = 0.01$ (unten)

- Je größer α, desto größer die Power.
- Beide Fehlerarten verhalten sich gegenläufig. Wählt man die Irrtumswahrscheinlichkeit α immer kleiner, so wird das Risiko für einen β-Fehler immer größer.

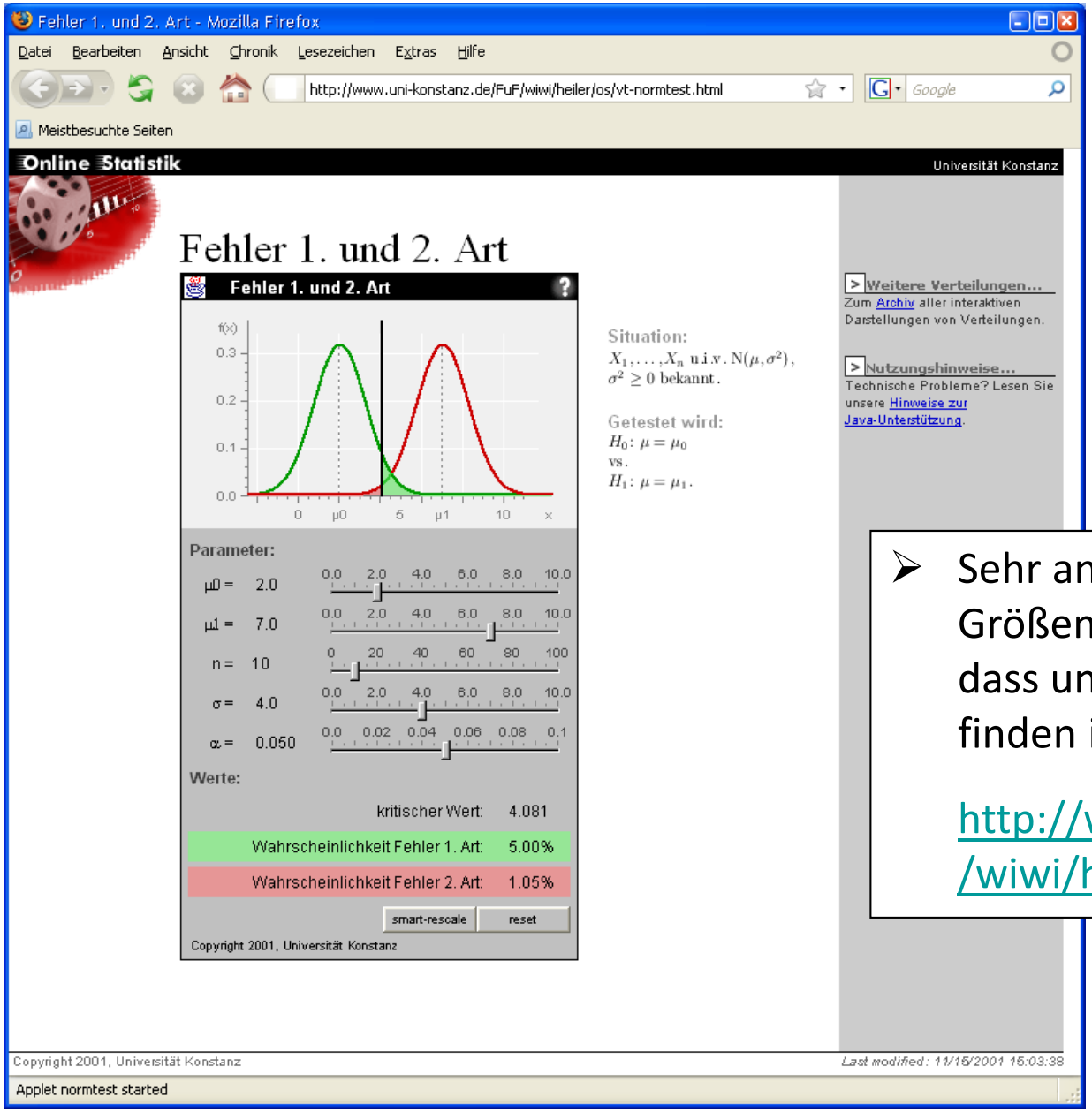
Power: Determinante Effektstärke



Konstant: σ, n, α, einseitige Testung
Variant: Effekt $\varepsilon_1 = \mu_1 - \mu_0$ kleiner (oben) vs. größer (unten)

- Je stärker der relevante Mindesteffekt, desto größer die Power.
- Wenn der als relevant erachtete Effekt größer ist, so wird er in dem Fall, dass er vorliegt (also die H_1 richtig ist), auch leichter gefunden.

Power: Determinanten



- Sehr anschaulich wird der Einfluss dieser Größen mit einem Programm visualisiert, dass unter folgender Internet-Adresse zu finden ist:

  http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-normtest.html

Power: Determinanten

- Die Power hängt zudem noch von weiteren Größen ab. Sie ist größer, ...
  - wenn die Stichprobengröße n zunimmt
  - wenn die Varianz des Merkmals in der Population σ kleiner ist
  - wenn ein größerer Effekt in der H_1 angenommen wird
  - wenn das Signifikanzniveau α größer gewählt wird
  - bei ein- statt zweiseitiger Testung und entsprechender Richtung des Effektes
  - wenn der Versuchsplan bestimmte Eigenschaften aufweist
    - z.B. weist der Vergleich von Mittelwertsunterschieden bei zwei (oder mehr) abhängigen Stichproben dann eine größere Power auf als bei unabhängigen Stichproben, wenn die abhängigen Messungen hoch korreliert sind
    - z.B. wirken sich gleiche Stichprobenumfänge beim Vergleich von Mittelwertsunterschieden zwischen zwei (oder mehr Gruppen) positiv auf die Power aus
  - wenn die Messungen reliabler sind

- Die Power wird auch davon beeinflusst, welcher Test verwendet wird und ob die Voraussetzungen des Tests erfüllt sind (z.B. Normalverteilung, Varianzhomogenität ...).

Power

- Oben wurde der Effekt als Differenz der unter der H_0 und der H_1 postulierten Populationsmittelwerte spezifiziert: $\varepsilon_1 = \mu_1 - \mu_0$ bzw. analog der tatsächliche Effekt als $\varepsilon = \mu - \mu_0$.
- Ein Nachteil dieser Effekt-Definition liegt darin, dass sie abhängig vom Maßstab der Variablen (IQ-Punkte, Gewichtsreduktion in kg, Zahl der Items im Gedächtnistest) und damit nicht über Studien hinweg vergleichbar ist, wenn die Messinstrumente eine andere Metrik aufweisen. Außerdem berücksichtigt er nicht, wie stark die Werte streuen.
- Daher standardisiert man die Differenzen an der Standardabweichung und kommt so zu dem Effektstärkemaß δ („delta“, Cohens δ) bzw. δ_1:

$$\delta = \frac{\varepsilon}{\sigma} = \frac{\mu - \mu_0}{\sigma} \qquad \delta_1 = \frac{\varepsilon_1}{\sigma} = \frac{\mu_1 - \mu_0}{\sigma}$$

- Dieses Effektstärkemaß lässt sich in ähnlicher Weise z.B. auf den Fall des Vergleichs von Mittelwertsunterschieden in zwei Gruppen anwenden:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$ (bei unabhängigen Gruppen) $$\delta = \frac{\mu_d}{\sigma_d}$$ (bei abhängigen Gruppen)

- Für andere statistische Tests resultieren andere Effektstärkemaße (vgl. Cohen, 1988; Eid, Gollwitzer & Schmidt, 2010; Faul et al., 2007). Der Einfachheit halber bezeichnen wir diese im Folgenden auch mit δ_1 bzw. δ (obwohl sie meist andere Bezeichnungen haben, s.u.).

Power

- Eine begründete Annahme über die Stärke des Effektes kann resultieren ...
  - aus der Kenntnis früherer Studien. Man kann Effektstärken in früheren Studien bestimmen oder auf Metaanalysen zurückgreifen, in denen Effektstärken über Studien zu einem bestimmten Effekt aggregiert wurden.

    Beispiel: Eine Metaanalyse zum Einfluss von Diät auf die Gewichtsreduktion zeigt bei übergewichtigen erwachsenen Personen eine mittlere Abnahme von $10.7\ \mathrm{kg}$ in durchschnittlich 16 Wochen (Miller, Koceja & Hamilton, 1997) und eine Effektstärke von $\hat{\delta} = 5.1$.
  - aufgrund inhaltlicher Überlegungen, wie groß ein Effekt mindestens sein müsste, um als bedeutsam erachtet zu werden.

    Beispiel: Ärzte empfehlen eine Abnahme um ca. 1 kg pro Woche (Miller, Koceja & Hamilton, 1997).
  - aufgrund des Vergleichs mit alternativen treatments oder Zielgruppen.

    Beispiel: Eine Metaanalyse zeigt, dass Aerobic-Kurse bei übergewichtigen erwachsenen Personen zu einer durchschnittlichen Gewichtsreduktion von $2.1\ \mathrm{kg}$ pro Woche führen (bei einer mittleren Dauer von 21 Wochen, $\hat{\delta} = 2.1$, Miller, Koceja & Hamilton, 1997). Man kann evtl. bei Kindern einen Effekt in derselben Größenordnung erwarten.

Power

- Eine begründete Annahme über die Stärke des Effektes kann resultieren ... (Fs.)
  - aufgrund allgemeiner heuristischer Empfehlungen von Cohen (1988). Cohen hat für verschiedene Effektstärkemaße/Tests Vorschläge gemacht, welche Effektstärken jeweils „kleinen“, „mittleren“ oder „großen“ Effekten entsprechen sollen. Hier muss man entscheiden, ob man einen kleineren, mittleren oder großen Effekt erwartet und kann dann die entsprechende Effektstärke nachschlagen und einsetzen.

    Beispiel: Im obigen Fall des Ein-Stichproben-Tests lauten die Konventionen: $|\delta| \approx 0.14$ ist „klein“, $|\delta| \approx 0.35$ ist „mittel“, und $|\delta| \approx 0.57$ ist „groß“ . Würden wir also einen mittleren Effekt bzgl. des Intelligenzunterschieds zwischen Jura- und anderen Studierenden erwarten, so würden wir von $\delta_1 = 0.35$ ausgehen. (Dies entspricht $\delta_1 \cdot \sigma = \varepsilon_1 = 0.35 \cdot 15 = 5.25$ Intelligenzpunkten, statt den im Beispiel angenommenen 3 Punkten.)

    Obwohl Cohen dieses Vorgehen nur für den Fall vorgeschlagen hat, wenn keine andere Möglichkeit greift, wird in der Forschungspraxis sehr häufig darauf zurückgegriffen.

Power: Gliederung

1 Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson

2 Power und ihre Determinanten

3 Vorgehen bei der Poweranalyse

4 Power-Berechnungen mit dem Programm G*Power

5 Power in der Empirie

6 Effektstärken

Power

- A priori Poweranalysen werden vor einer Studie durchgeführt, um zu prüfen, ob die Power ausreichend ist bzw. die optimale Stichprobengröße festzulegen. Man legt dann α und β (oder $1 - \beta$) sowie die Mindestgröße des Effektes fest und berechnet dann n.

- Optimal bedeutet, dass die Stichprobengröße n ...
  - nicht zu klein ist. Bei zu kleinem n wäre die Wahrscheinlichkeit, einen bestehenden Effekt der festgelegten Größe auch zu finden, kleiner als die gewünschte Power.
  - nicht zu groß ist. Bei zu großem n würden auch bereits kleinere als der spezifizierte Effekt statistisch signifikant. Dies erscheint aber eben aufgrund der Vorüberlegungen nicht als bedeutsam. Außerdem wäre der Aufwand bei der Datenerhebung größer.
  - so groß ist, dass dann, wenn (mindestens) der in der H_1 festgelegte Effekt existiert, die Wahrscheinlichkeit für eine Entscheidung zugunsten der H_1 (mindestens) $1 - ß$ beträgt, die gewünschte Power also erzielt wird. (Das Risiko einer Fehlentscheidung bei der Annahme von H_1 beträgt weiterhin α.)

Power

- Alternativ können Poweranalysen auch a posteriori durchgeführt werden, wenn ein insignifikantes Ergebnis resultierte. (Andernfalls kann man keinen β-Fehler begehen.)
- In diesem Fall will man wissen, ob die Power in dieser Studie ausreichend war, also eine ausreichend hohe Chance bestand, einen bestimmten Effekt auch zu finden. Dies ist vor allem auch dann von Bedeutung, wenn die wissenschaftliche Hypothese in der H_0 lag. Dabei sind zwei verschiedene Vorgehensweisen anzutreffen, die hier als Post-hoc- und retrospektive Poweranalysen bezeichnet werden (Faul et al., 2007; die Terminologie ist im Allgemeinen sehr uneinheitlich):
  - Post-hoc Poweranalyse: Mit den Angaben α und n aus der Studie kann man für einen theoretisch gewählten minimalen Effekt δ_1 prüfen, wie groß die Power $1 - \beta$ in dieser Studie war. Hier ist die Bezeichnung Post-hoc Poweranalyse insofern irreführend, als der Effekt ja unabhängig vom Ergebnis der Studie festgelegt wird.
  - Retrospektive Poweranalyse: Wie vorher, aber δ_1 wird hier aus den Daten der Studie (also z.B. d) geschätzt. Die Annahme, dass die Effektstärke in der Stichprobe der der Population entspricht, ist dabei problematisch; die Schätzung ist – insbesondere bei kleinen Stichproben – verzerrt. Außerdem widerspricht dieses Vorgehen der Logik der theoretisch begründeten Festlegung des minimalen Effektes. Aus diesen und anderen Gründen raten viele Autoren von dieser Art der Poweranalyse ab.

Power: Gliederung

1 Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson

2 Power und ihre Determinanten

3 Vorgehen bei der Poweranalyse

4 Power-Berechnungen mit dem Programm G*Power

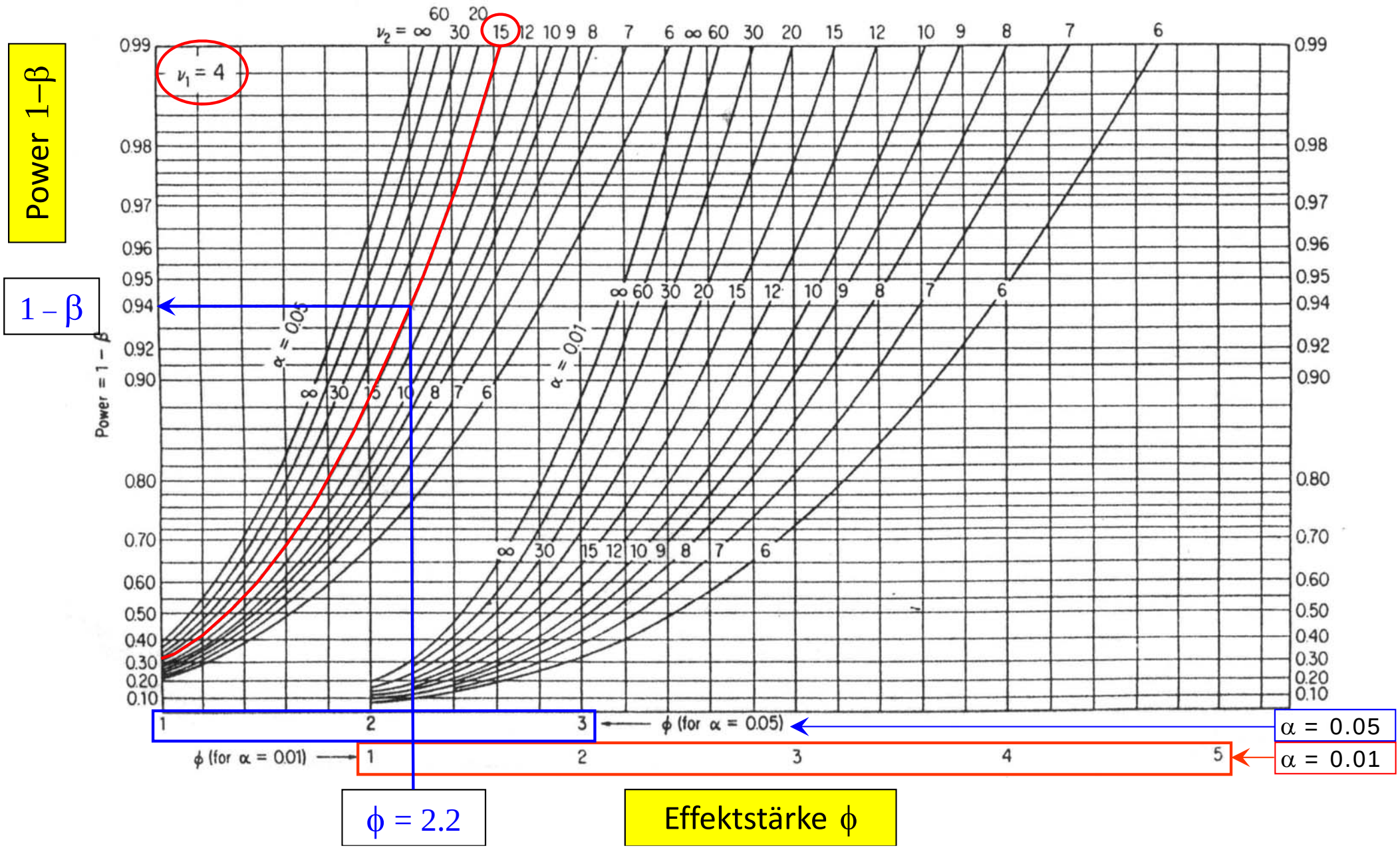
5 Power in der Empirie

6 Effektstärken

Power-Berechnungen

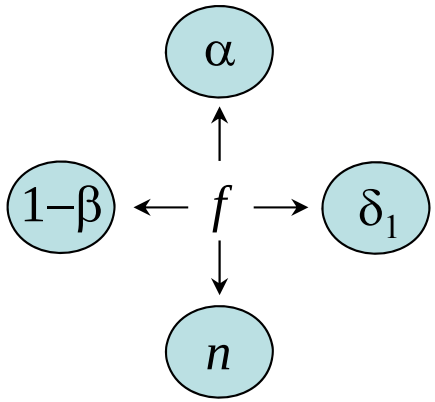
- Die numerische Bestimmung der Power ist im Allgemeinen komplizierter als in unserem Beispiel des Ein-Stichproben z-Tests. Dazu müssen die Stichprobenkennwerteverteilungen unter der H_1 bestimmt werden. Diese bezeichnet man auch als nonzentrale Verteilungen, z.B. im Falle der t-Tests als nonzentrale t-Verteilung (die nicht nur von den *df* sondern zusätzlich noch von einen Nonzentralitätsparameter abhängen und nicht mehr symmetrisch sind).
- Zur praktischen Berechnung wurde früher mit Tabellen (z.B. im Buch von Cohen, 1988) oder grafischen Darstellungen (sog. Nomogramen) gearbeitet, heute verwendet man Computerprogramme.
- SPSS kann nur retrospektive Poweranalysen durchführen und das auch nur in einigen Prozeduren. Andere Statistikprogramme wie Systat bieten mehr Möglichkeiten.
- Ein ausgezeichnetes Freeware-Programm ist G*Power (Faul, Erdfelder, Lang & Buchner, 2007), das auch im CIP-Pool auf allen Rechnern installiert ist. Windows- und MAC OS-Versionen sind frei downloadbar unter http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/.

Power-Berechnungen: Nomogramm



Power-Berechnungen

- Bei der Poweranalyse besteht eine (komplizierte) funktionale Beziehung zwischen den Größen α, β (bzw. $1-\beta$), n, und δ_1. Je nach Anwendung kann man eine Größe bestimmen, wenn die anderen drei Größen bekannt sind (für weitere Anwendungen vgl. Faul et al., 2007):
  - Bestimmung der optimalen Stichprobengröße: Festgesetzt werden α, $1-\beta$ und δ_1. Gesucht wird n.
  - Post-hoc Poweranalyse: Festgesetzt sind α, δ_1, in der Studie existiert ein bestimmtes n. Bestimmt wird $1-\beta$.
  - Kompromiss Poweranalyse (Faul et al., 2007): Gegeben sind ein maximal zu erreichendes n, δ_1 und ein akzeptables Verhältnis der beiden Fehlerwahrscheinlichkeit β / α (z.B. $\beta / \alpha = 1$ oder $\beta / \alpha = 2$). Bestimmt werden α und $1-\beta$.



Power-Berechnungen mittels G*Power

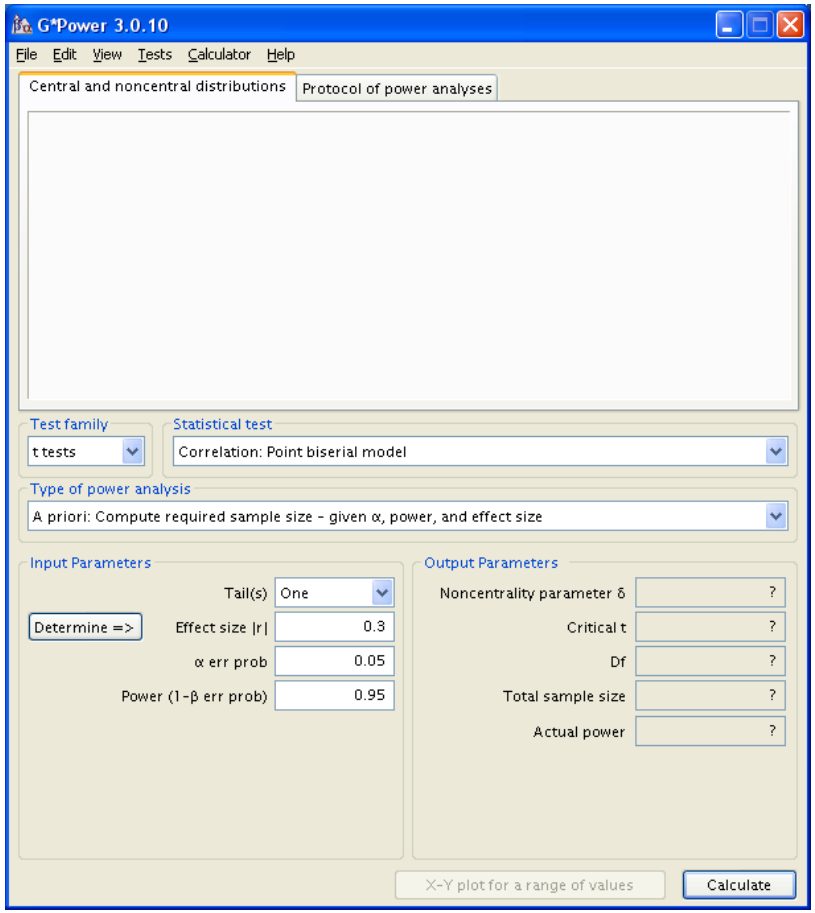
- Die Berechnungen im Programm G*Power 3 (aktuell Version 3, Faul et al., 2007) erfolgen in den folgenden vier Schritten:

  (1) Wähle den angemessenen Test aus. Dabei ist zunächst die Test-Familie aufgrund der Verteilungsform der Prüfgröße zu bestimmen (also z-, t-, F- und χ^2-Verteilung) und dann der Test.

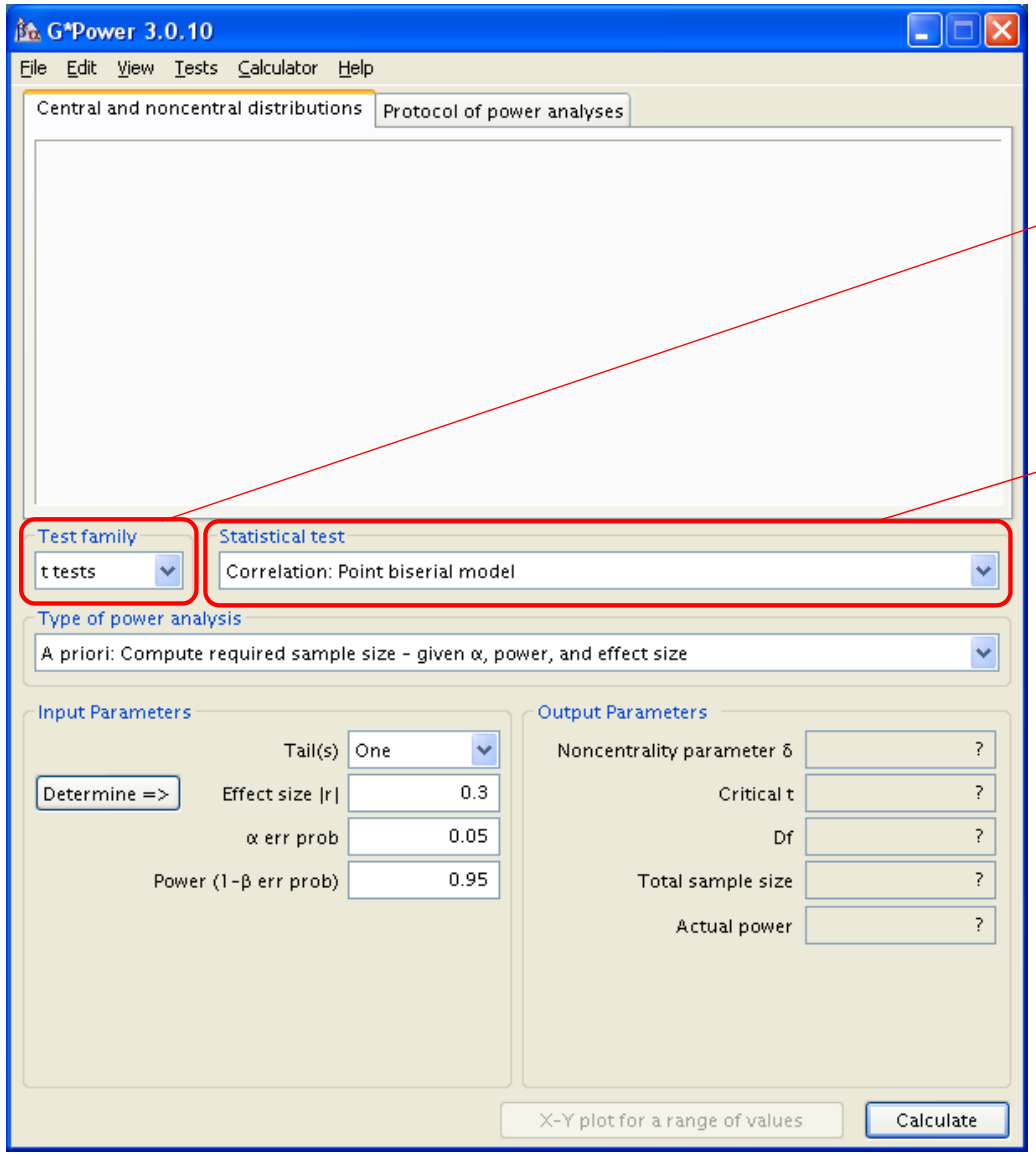
  (2) Wähle die Art der Analyse (folgende drei von fünf Typen wurden oben dargestellt: „A priori“, „Post hoc“ und „Compromise“)

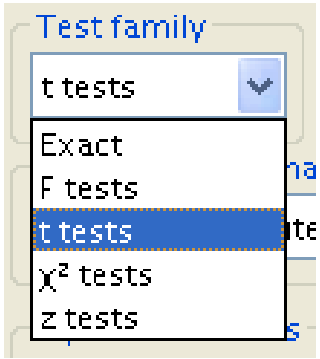
  (3) Gebe die erforderlichen Größen an.

  (4) Führe die Berechnungen aus.



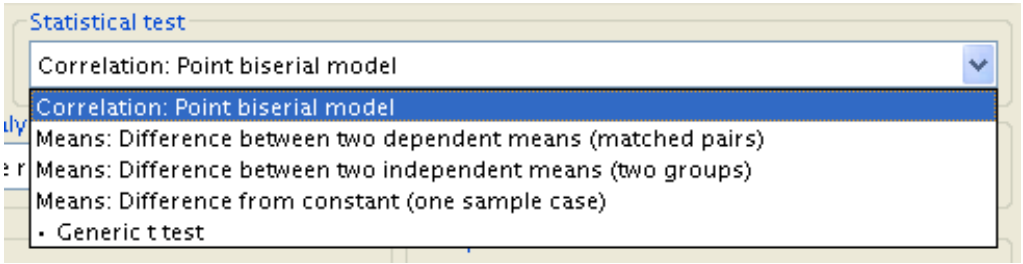
Power-Berechnungen mittels G*Power



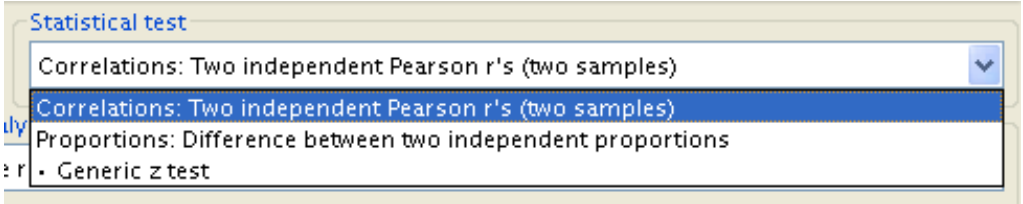


Alternativ kann das Testverfahren auch über das Hauptmenü unter Tests/… ausgewählt werden.

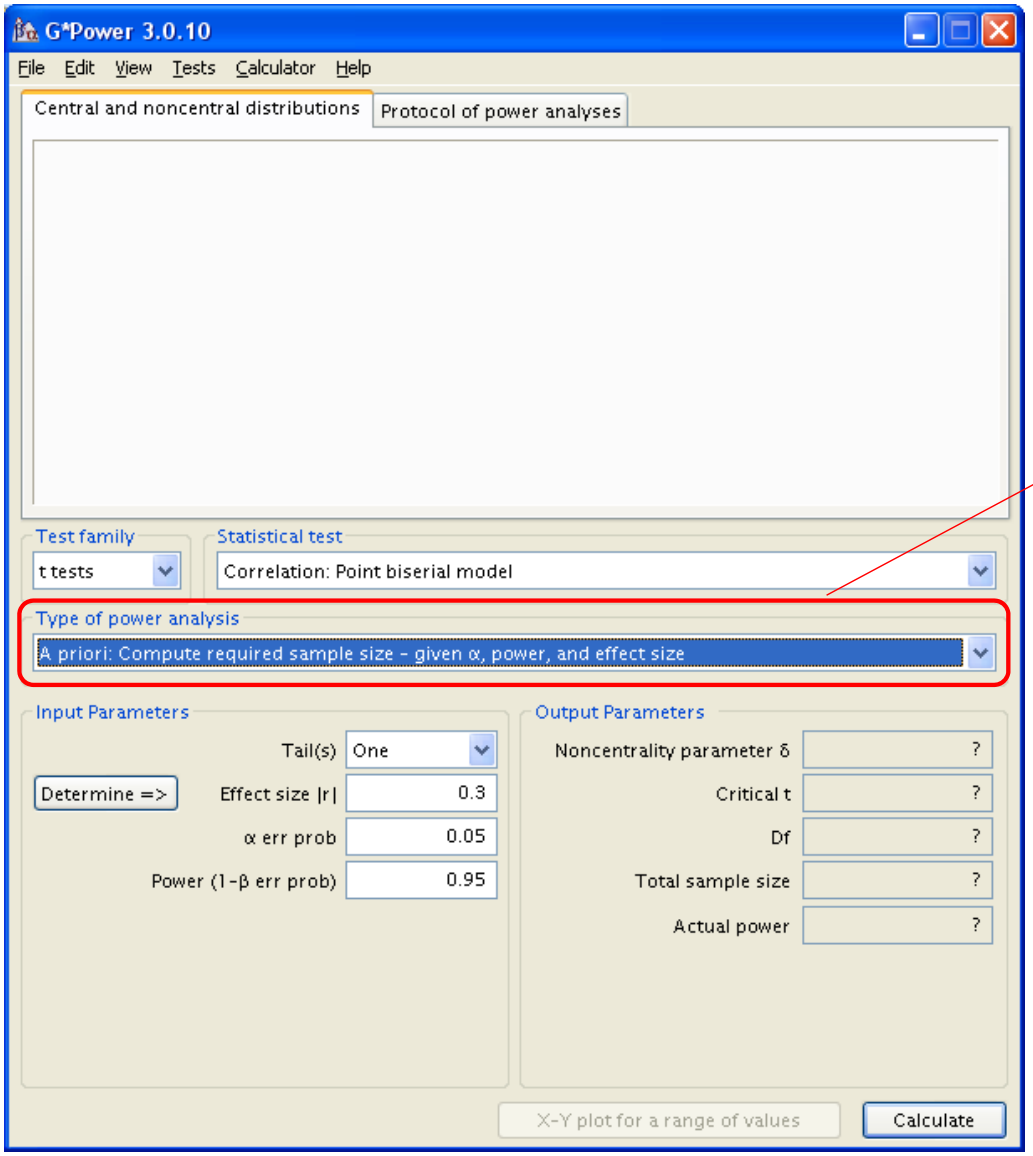
Nach Festsetzen der „Test family“ kann dann der genaue „Statistical test“ spezifiziert werden, z.B. bei der t-Verteilungs-Familie (Version 3.0):



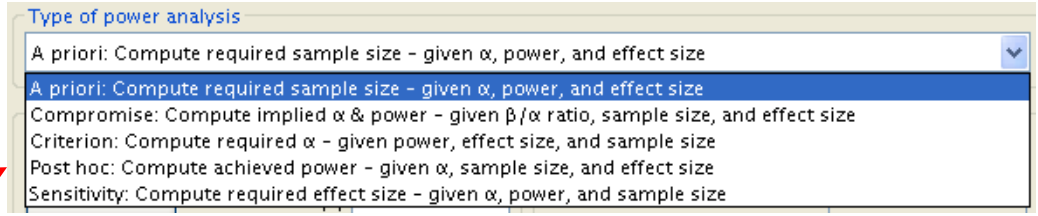
oder bei der z-Familie (Normalverteilung):



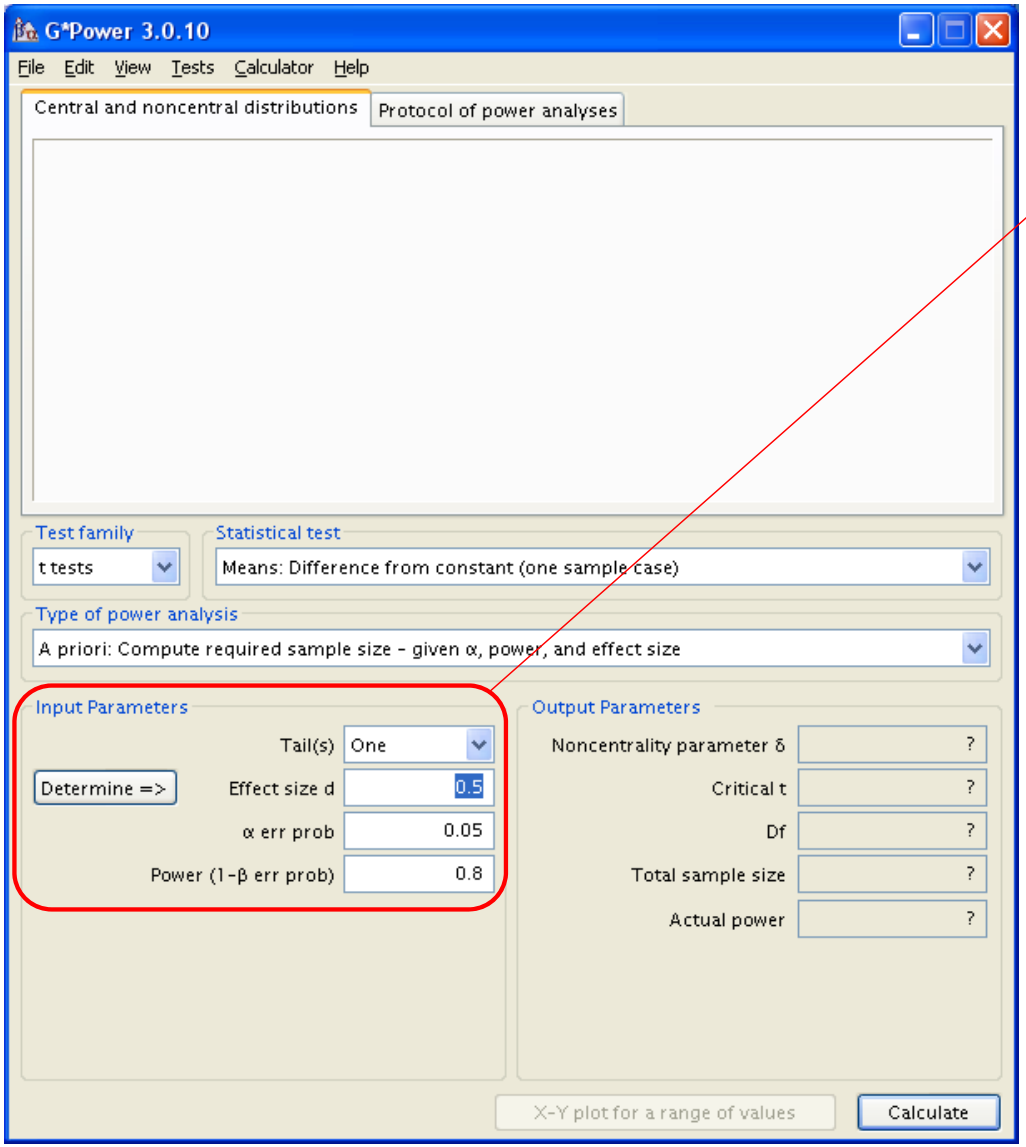
Power-Berechnungen mittels G*Power



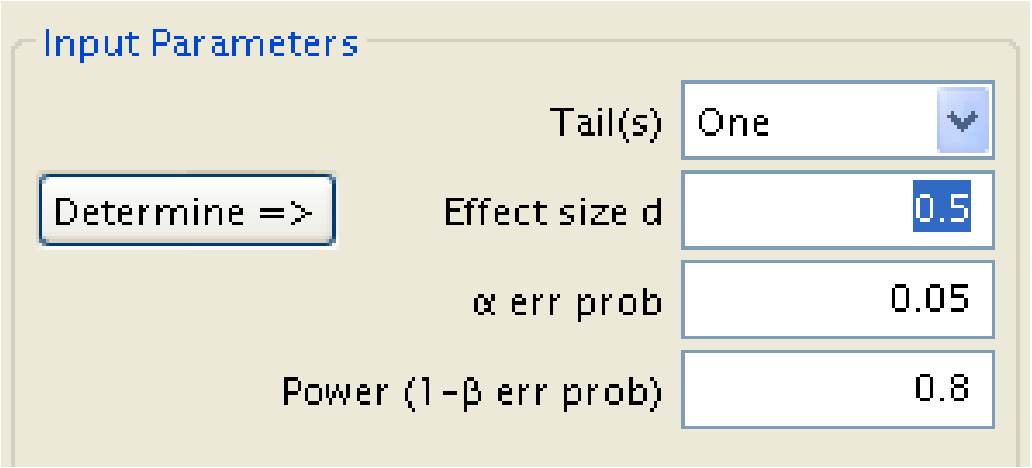
Anschließend ist der Typ der Berechnung („Type of power analysis“) festzulegen. Behandelt wurden „A priori“ (Voreinstellung), „Compromise“ und „Post hoc“:



Power-Berechnungen mittels G*Power



In Abhängigkeit vom gewählten Analyse-Typ müssen dann die bekannten Größen („Input Parameters“) spezifiziert werden.
Bei der a priori Berechnung des optimalen Stichprobenumfangs für den Ein-Stichproben t-Test sind dies α („α err prob“), Power („1-β err prob“), δ_1 („Effect size d“) und die Information, ob der Test einseitig [„Tail(s)“=One] oder zweiseitig [„Tail(s)“=Two] durchgeführt werden soll:



Power-Berechnungen mittels G*Power

- Beispiel 1: Mehl, Vazire, Ramirez-Esparza, Slatcher und Pennebaker (2007) untersuchten die Hypothese, dass Frauen gesprächiger als Männer sind. Dazu führten männliche und weibliche Studierende über mehrere Tage im Alltag ein Sprachaufzeichnungsgerät mit, dass alle 12.5 Minuten 30 Sekunden lang aufzeichnete. Bei der Auswertung ergaben sich hochgerechnet bei einer durchschnittlichen Wachzeit von 17 Stunden die folgenden Wortzahlen:

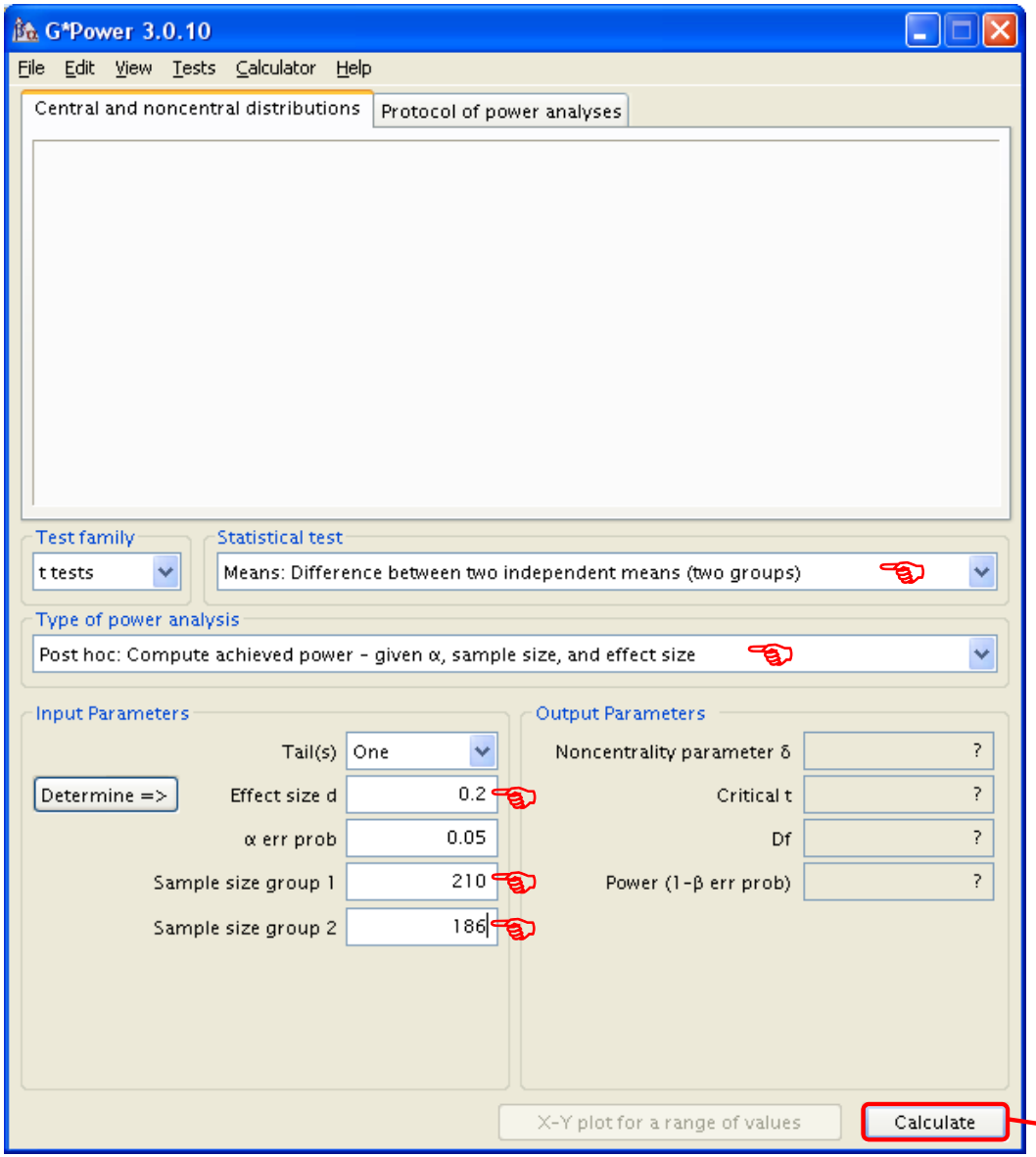
	n	$\bar{x}$	s
Frauen	210	16215	7301
Männer	186	15669	8633

- Der t-Test für unabhängige Gruppen ergab eine Prüfgröße von $t(394) = 0.682$ mit einem p-Wert von 0.248 bei einseitiger Testung. Bei $\alpha = 0.05$ resultiert also kein statistisch signifikanter Effekt.

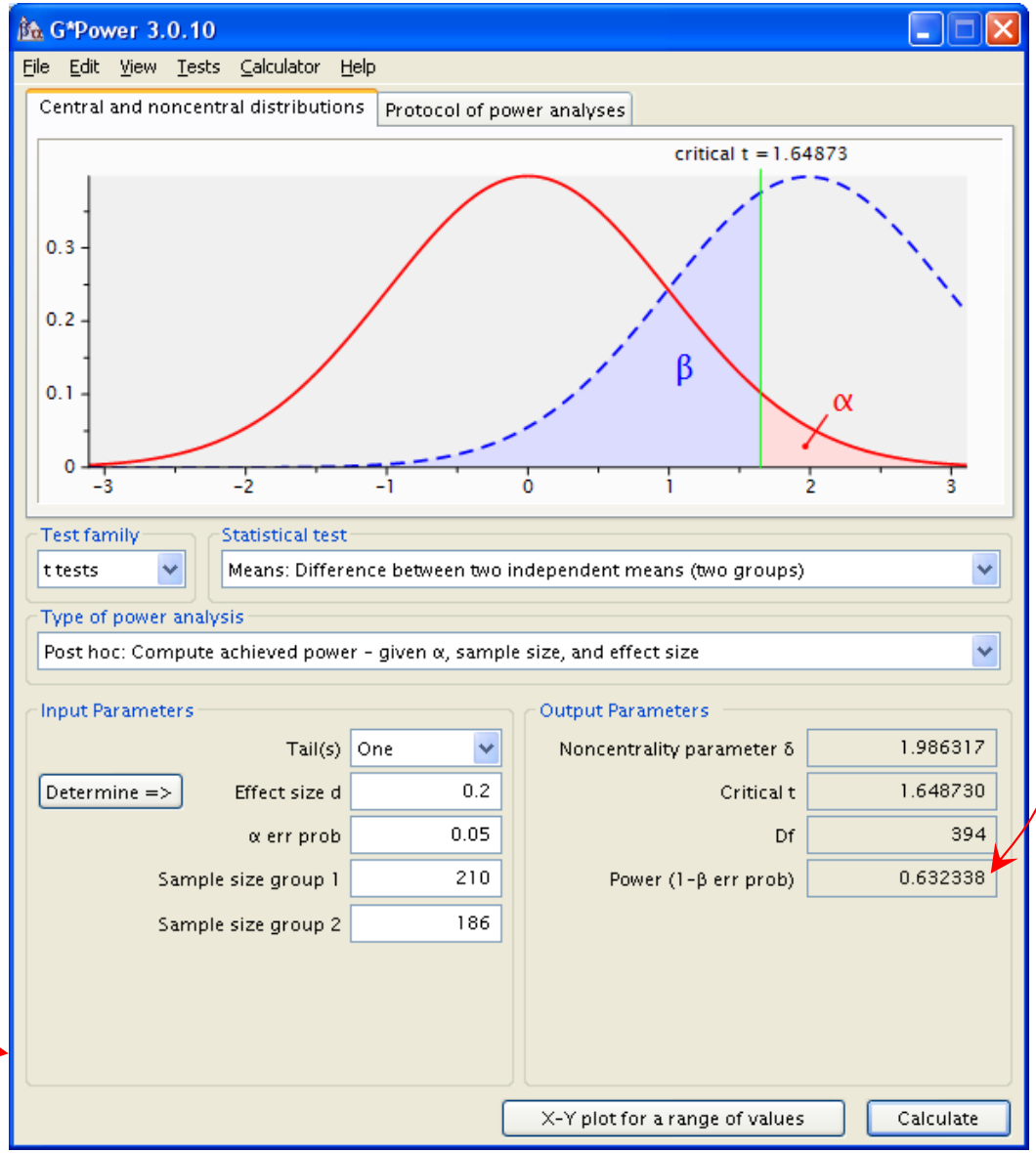
Die Autoren folgern: „We therefore conclude, on the basis of available empirical evidence, that the widespread and highly publicized stereotype about female talkativeness is unfounded.“ (S. 82)

Frage: Wie sicher kann man sein, dass keine Geschlechtsunterschiede existieren? Wie groß war in der Studie die Chance, einen statistisch signifikanten Geschlechtsunterschied zu finden, wenn dieser klein ($\delta_1 = 0.20$) ist?

Power-Berechnungen mittels G*Power



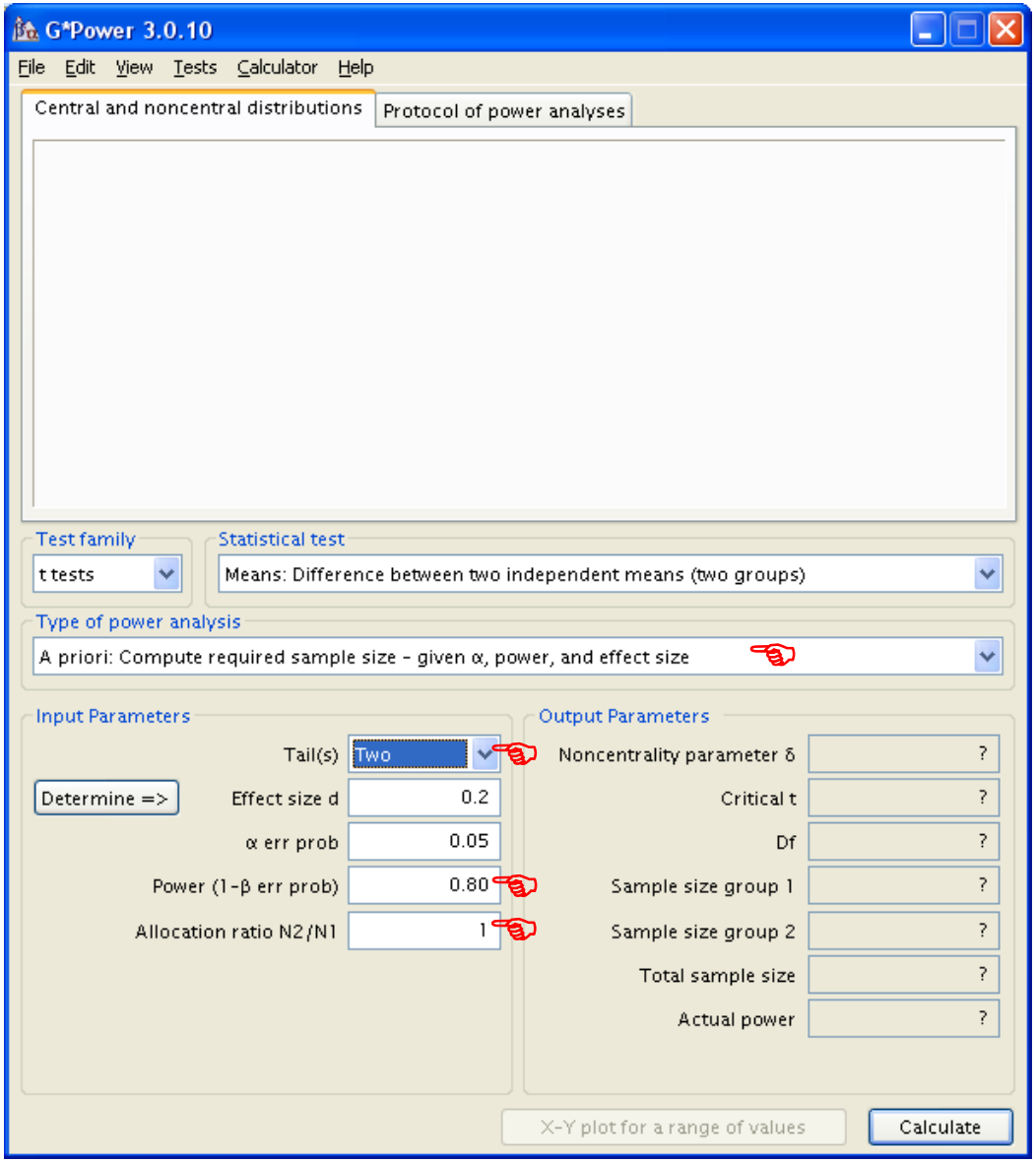
Die Power betrug also nur 0.63

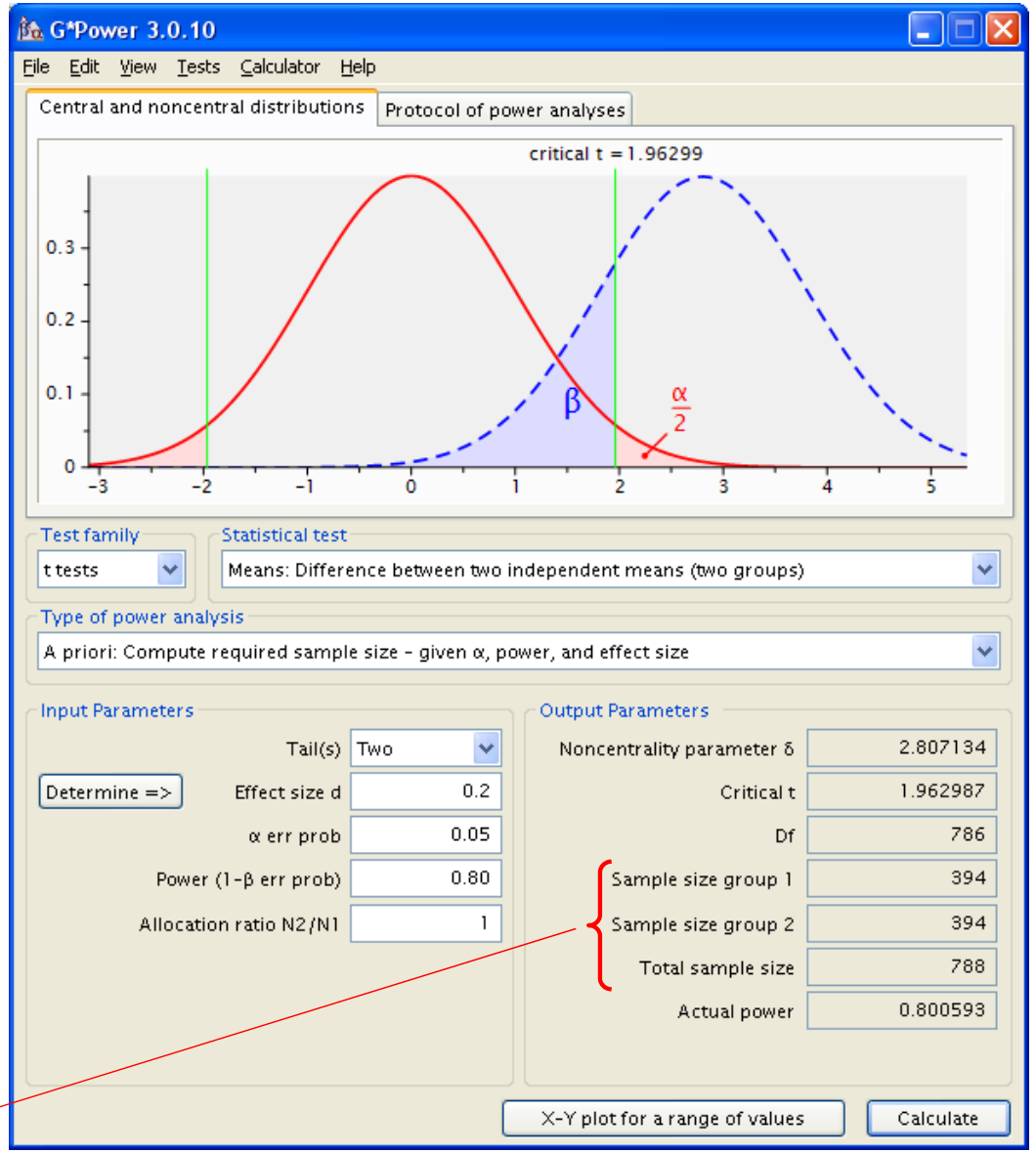


Power-Berechnungen mittels G*Power

- Beispiel 2: Wie groß müssten die Stichproben gewählt werden, um bei obiger Studie von Mehl et al. (2007) einen kleinen Effekt (Geschlechtsunterschied) bei $\alpha = 0.05$ und zweiseitiger Testung mit einer Power von 0.80 aufdecken zu können?

Power-Berechnungen mittels G*Power





Die Stichprobengrößen müssten $n_1 = n_2 = 394$ betragen.

Power: Gliederung

1 Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson

2 Power und ihre Determinanten

3 Vorgehen bei der Poweranalyse

4 Power-Berechnungen mit dem Programm G*Power

5 Power in der Empirie

6 Effektstärken

Power in der Empirie

- Cohen und andere empfehlen im Regelfall eine Power von mindestens 0.80 bei einem mittleren Effekt und $\alpha = 0.05$. (Dies entspricht einem Verhältnis von 1:4 zwischen den Wahrscheinlichkeiten, einen α-Fehler und einen β-Fehler zu begehen.)

- Cohen (1962) untersuchte die Power von 70 publizierten Studien (aus den Bereichen Klinische Psychologie und Sozialpsychologie). Ausgehend von den berichteten Stichprobenumfängen und den durchgeführten Analysen bei einem α von 0.05 und zweiseitiger Testung errechnete er bei den insgesamt durchgeführten 2088 Tests eine durchschnittliche Power von $.18$ für einen kleinen, $.48$ für einen mittleren und $.83$ für einen großen Effekt.

- Bei einer mittleren Effektstärke war die Wahrscheinlichkeit, korrekter Weise ein statistisch signifikantes Ergebnis zu erhalten, im Durchschnitt also weniger als 50%! Oder anders ausgedrückt: Die Chance, einen mittleren Effekt zu entdecken (bzw. zu übersehen) war ca. 50 : 50.

- Neuere Power-Studien zeigen, dass die Power auch in später publizierten Studien in den verschiedensten Forschungsfeldern nicht wesentlich besser geworden ist: Über 15 solche Power-Studien gemittelt berichten Onwuegbuzie und Leech (2004) mittlere Power-Werte von $.24$ (kleiner Effekt), $.63$ (mittlerer Effekt) und $.85$ (großer Effekt).

Power in der Empirie

- Es stellt sich natürlich die Frage, wieso in Zeitschriften so viele statistisch signifikante Ergebnisse berichtet werden können, wenn doch die Power der Studien häufig so niedrig ist. Dafür kann es prinzipiell zwei Ursachen geben:
- Ursache 1: Die Effekte sind tatsächlich häufig groß. Dies ist allerdings durch Metaanalysen widerlegt, die in ihrer überwiegenden Zahl eher kleine bis mittlere Effekte zeigen.
- Ursache 2: Die Effekte existieren häufig nicht und sind das Ergebnis von Stichprobenfehlern. Die große Zahl an signifikanten Ergebnissen ist dann auch zurückzuführen auf ...
  - Publikationsverzerrungen, die zustande kommen, wenn nur die signifikanten Studien veröffentlicht und die anderen in den Schubladen bleiben.
  - das Durchführen einer Vielzahl von Tests in einer Studie, bei denen dann im Nachhinein die signifikanten „herausgefischt“ werden und entsprechend der Ergebnisse die Hypothesen formuliert werden. [Z.B. ergaben sich bei der Power-Studie von Cohen (1962) allein durchschnittlich $2088/70 \approx 30$ statistische Analysen pro Studie.]

Power in der Empirie

- Poweranalysen sind also heute immer noch selten. Mögliche Ursachen sind u.a.
  - Unzureichende methodische Kenntnisse
  - Schwierigkeit bei der Begründung des Mindesteffektes
  - Apriori Poweranalysen fordern hohe Stichprobengrößen, die praktisch kaum zu realisieren sind

- Vorteile von Poweranalysen
  - Wir können prüfen, ob es bei einer geplanten Studie unter den gegebenen Randbedingungen (n, α ...) überhaupt eine realistische Chance gibt, einen vorhanden Effekt auch zu finden bzw. sie entsprechend planen, so dass dies der Fall ist.
  - Die Nullhypothese kann auch angenommen (nicht nur beibehalten) werden, da die Wahrscheinlichkeit β, dass dies eine fehlerhafte Entscheidung ist, kontrolliert wird. Dies ist besonders wichtig, wenn die Hypothese in der H_0 liegt.

Power: Gliederung

1 Logik der Inferenzstatistik: Fisher vs. Neyman-Pearson

2 Power und ihre Determinanten

3 Vorgehen bei der Poweranalyse

4 Power-Berechnungen mit dem Programm G*Power

5 Power in der Empirie

6 Effektstärken

Effektstärken

- Unabhängig davon, dass den Effektstärken (Effektgrößen, manchmal auch als „Maße der praktischen Signifikanz“ bezeichnet) eine besondere Rolle bei der Powerbestimmung zukommt, ist deren Bestimmung und Bericht allgemein (zusätzlich zur Angabe der Signifikanz bzw. der Konfidenzintervalle) sinnvoll!
- Denn: Bei sehr großem n können auch triviale, minimale Effekte signifikant werden. Und umgekehrt: Bei sehr kleinem n können auch starke Effekt zu einem insignifikanten Ergebnis führen.
- Effektstärken sind unabhängig vom Maßstab der betrachteten abhängigen Variablen (dimensionslose Größen) und eignen sich daher besonders zum Vergleich zwischen Studien und zur Aggregierung (Mittelung) von Effekten über Studien hinweg. Solche Aggregierungen werden in sog. Metaanalysen vorgenommen und geben dann (u.a.) die Stärke des Effektes einer bestimmten Maßnahme oder eines bestimmten Zusammenhangs über Studien hinweg an.
- Im Folgenden wird am Beispiel des t-Tests für unabhängige Gruppen die Bestimmung eines geeigneten Effektstärkemaßes exemplarisch dargestellt.

Effektstärken

- Beim t-Test für unabhängige Gruppen ergibt sich als Effektstärke-Maß δ („delta“) in der Population die an der Standardabweichung σ normierte Differenz der Mittelwerte in beiden Gruppen (Entsprechend der Voraussetzung des t-Tests gilt $\sigma = \sigma_1 = \sigma_2$):

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

- Eine Schätzung für δ ergibt sich durch Einsetzen der Stichprobenkennwerte, wobei im Nenner die aus beiden Gruppen gepoolte Standardabweichung verwendet wird. Als Schätzung d (auch als Hedges g bezeichnet) resultiert:

$$d = \hat{\delta} = \frac{\overline{x}_1 - \overline{x}_2}{s_g} \text{ mit } s_g = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

- Die Schätzung ist nicht erwartungstreu und kann entsprechend (vor allem sinnvoll bei kleinen Stichproben) noch mittels des Faktors $1 - 3 / [(4 \cdot (n_1 + n_2) - 9]$ korrigiert werden (dazu und zu alternativen Schätzungen vgl. z.B. Ellis, 2010).

Effektstärken

- Nach Cohen (1988) werden hier Effektstärken um 0.20 als „klein“, um 0.50 als „mittel“ und um 0.80 als „groß“ klassifiziert.
- Für Effektstärken lassen sich ebenfalls Signifikanztests oder – informativer – Konfidenz-intervalle angeben (werden hier nicht berichtet).
- Verschiedene Effektstärke-Maße lassen sich (approximativ) ineinander umrechnen. Z.B. besteht zwischen d und der (punktbiserialen) Korrelation folgender Zusammenhang (vgl. Rosenthal 1994):

$$r = \frac{d}{\sqrt{d^2 + \frac{n^2 - 2 \cdot n}{n_1 \cdot n_2}}}$$

- Es bestehen auch funktionale Zusammenhänge zwischen der Prüfgröße und dem Effektstärkemaß. Z.B. besteht zwischen der t-Prüfgröße und d folgender Zusammenhang:

$$d = t \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} = \frac{t}{\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}}$$

Effektstärken

- Folgende Tabelle gibt exemplarisch Definitionen für Stichproben-Effektstärken mit Empfehlungen für kleine, mittlere und große Effekte an (ausführlicher Cohen, 1988; Cooper & Hedges, 1994; Eid, Gollwitzer & Schmidt, 2010; Faul et al., 2007):

Testverfahren	Effektstärkemaß	klein	mittel	groß
Ein-Stichproben z- oder t-Test	$d = \frac{\bar{x} - \mu_0}{s}$	.14	.35	.57
t-Test für Mittelwertsunterschiede zweier unabhängiger Gruppen	$d = g = \frac{\bar{x}_1 - \bar{x}_2}{s_g}$	.20	.50	.80
Levene-Test für Unterschiede in Varianzen zweier unabhängiger Gruppen	$v = \frac{s_1^2}{s_2^2}$	1.1	1.5	2.0
Vierfelder χ^2-Test zum Vergleich zweier Anteile aus unabhängigen Stichproben	φ	.10	.30	.50
Produkt-Moment Korrelation	r	.10	.30	.50

Power: Zitierte Quellen

Zitierte Quellen:

- Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.
- Cooper, H. & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York, NY: Russell Sage. [Kap. 16 & 17)
- Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge: Cambridge University Press.
- Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedicl sciences. Behavior *Research Methods*, *39*, 175-191.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.
- Gigerenzer, G. (1993). The supergo, the ego, and the id in statistical reasoning. In G. Kerren & Ch. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ. Erlbaum.
- Miller, W. C., Koceja, D. M. & Hamilton, E. J. (1997). A meta-analysis of the past 25 years of weight loss research using diet, exercise or diet plus exercise intervention. *International Journal of Obesity*, *21*, 941-947.
- Onwuegbuzie, A. J. & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*, 201-230.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- Soane, E., Dewberry, C. & Narendam, S. (2010). The role of perceived costs and perceived benefits in the relationship between personality and risk-related choice. *Journal of Risk Research*, *13*, 303-318.