

komplexe Analysemethoden

- Vorlesung + Tutorium
 - das **Induktionsproblem** (Probleme des Signifikanztests, Bayes-Statistik, Metaanalyse...), Regression, allgemeines Lineares Modell
 - Wiederholung deskriptive Statistik und Inferenzstatistik
- Seminar
 - multivariate Verfahren - Regression, varianzanalytische Verfahren, logistische Regression, Mehrebenenanalyse,
- Modulprüfung: Klausur zu Inhalten aus Vorlesung und Seminar,

Inferenzstatistik - Schluss von Stichproben auf die Population
keine Lösung - viele Lösungsvorschläge

was sollte bereits bekannt sein?

- **deskriptive Statistik** → Ziel, empirische Daten durch Tabellen, Kennzahlen, Parameter und Grafiken übersichtlich darzustellen und zu ordnen - vor allem bei umfangreichem Datenmaterial → <https://studyflix.de/statistik/deskriptive-statistik-1052>
 - Merkmale
 - * diskret → endlich/abzählbar unendlich viele Ausprägungen, bspw. Anzahl der Kinder einer Person,
 - * stetig → theoretisch unendlich verschiedene Werte innerhalb eines Intervalls - jede denkbare Zahl mit beliebig vielen Nachkommastellen, bspw. Körpergröße,
 - * quasi-stetig → stetige Daten, die nur gerundet gemessen werden, bspw. Nettoeinkommen,
 - Messung von Merkmalen,
 - Skalentypen
 - * nominalalskaliert → endliche Menge ohne Rangfolge, bspw. Farben von Autos,
 - * ordinalskaliert → endliche Menge mit Rangfolge → keine Interpretation der Abstände möglich, bspw. *trifft zu, trifft weniger zu, trifft überhaupt nicht zu*
 - * intervallskaliert → unendlich viele Ausprägungen → ohne Nullpunkt → keine Verhältnisaussagen möglich, bspw. Temperaturskala
 - * verhältnisskaliert → mit einem absoluten Nullpunkt → Verhältnisaussagen wie *doppelt so viel..., halb so viel...* sind nun möglich, bspw. Einkommen, Alter → <https://www.crashkurs-statistik.de/merkmals-und-skalentypen/>
 - Beschreibung von Daten:
 - * Häufigkeitsverteilungen → tabellarische Aufstellung, wie häufig die Ausprägungen eines oder mehrerer Merkmale beobachtet werden
 - * Kennwerte der Verteilungen
 - ! Mittelwert,
 - ! Median → Zentralwert des Datensatzes → genau in der Mitte,
 - ! Varianz → Maß für die Streuung der Wahrscheinlichkeitsdichte um ihren Schwerpunkt,
 - ! Standardabweichung → Maß für die Streubreite der Werte eines Merkmals rund um dessen Mittelwert
 - ! z- Standardisierung überführt Werte, die mit unterschiedlichen Messinstrumenten erhoben wurden, in eine neue gemeinsame

Einheit: in Standardabweichungs-Einheiten →
<https://www.statistikpsychologie.de/z-transformation/>

- Zusammenhänge zwischen Merkmalen:
 - ! Korrelation → misst die Stärke einer statistischen Beziehung von 2 Variablen zueinander - bei einer positiven Korrelation gilt *je mehr Variable A... desto mehr Variable B*
 - ! einfache Regression → nützliches Verfahren für Prognosen, z.B. Vorhersage von Besucherzahlen - und für die Untersuchung von Zusammenhängen, z.B. Einfluss von Werbeausgaben auf die Verkaufsmenge,
 - ! Effektstärke = Effektgröße → ein Wert kleiner als 0.5 gilt als kleiner Effekt, zwischen 0.5 und 0.8 zählt als mittlerer Effekt und Werte darüber als großer Effekt,

→ Wahrscheinlichkeitstheorie

- Wahrscheinlichkeitsbegriff,
- bedingte Wahrscheinlichkeit fragt nach, wie wahrscheinlich ein Ereignis ist, wenn man ein anderes bereits kennt,
- **Stichprobenverteilungen** → Verteilung der Wahrscheinlichkeit, mit der jeder mögliche Wert aus einer Statistik zufällig aus einer Grundgesamtheit gezogen werden kann,

→ Inferenzstatistik

- Stichprobe und Population:
 - ! Parameterschätzung → Ermittlung eines Schätzwertes für einen unbekannten Populationsparameter der Grundgesamtheit auf der Basis von Stichprobenkennwerten → Punktschätzung, Intervallschätzung
 - ! Konfidenzintervalle → lokalisiert die Lage eines wahren Parameters einer Grundgesamtheit mit einer gewissen Wahrscheinlichkeit,
- Grundidee des Signifikanztests:
 - ! statistische Hypothesen → Annahme, die mit Methoden der mathematischen Statistik auf Basis empirischer Daten geprüft wird,
 - ! Hypothesentest → Überprüfung von Behauptungen,
- Testverfahren:
 - ! t-Test → wenn die Mittelwerte von maximal 2 Gruppen miteinander verglichen werden sollen, bspw. kann man analysieren, ob Männer im Durchschnitt größer als Frauen sind,
 - ! Chi-Quadrat-Test → Analyse zweier nominal oder ordinal skalierten Variablen anhand der beobachteten Häufigkeiten ihrer Merkmalsausprägungen – sind 2 Variablen voneinander unabhängig?
 - ! Varianzanalyse → Untersuchung des Einflusses von uV auf aV,

worum geht's?

- Methodenkritik - genauer: Kritik an der *traditionellen* Anwendung statistischer Verfahren in der Psychologie und den Sozialwissenschaften und **Evidenzbewertung**

es liegen ein nachgewiesener
Zusammenhang/ eine nachgewiesene
Wirksamkeit vor

→ allgemeine statistische Prinzipien - hauptsächlich:

- das Allgemeine Lineare Modell

alle Variablen stehen in Beziehung zueinander → linearer Zusammenhang

- das Induktionsproblem (und Lösungsvarianten)

Frage, ob und wann ein Schluss durch Induktion von Einzelfällen auf ein allgemeingültiges Gesetz zulässig ist

→ Multivariate Verfahren

Methodenkritik: die Schule der Skepsis

- die traditionelle Anwendung inferenzstatistischer Verfahren in der Psychologie ...
 - liefert oftmals missverstandene Ergebnisse,
 - ist unzureichend
 - oder schlicht falsch.
- dies richtet sich vor allem - aber nicht nur - gegen die *ritualisierte* Anwendung von Signifikanztests,
- **dies ist keine nerdy-Außenseiterposition abgedrehter Methodiker, sondern der Stand der Diskussion!**
 - *psychology will never be a real science unless we change the way we analyze data* - Loftus, 1994
 - die APA-Guidelines empfehlen dringlich den Einsatz alternativer Verfahren,
 - renommierte Journals verbieten den Einsatz traditioneller Signifikanztests (s. *Psychological Science*),
- warum? ...und was sollten Sie stattdessen tun?

Methodenkritik: Revolution

- sie leben in Zeiten der wissenschaftlichen Revolution - des Paradigmenwechsels:
 - Replikations-Krise,
 - Probleme des Publikationssystems,
 - naive oder fehlerhafte wissenschaftstheoretische Überzeugungen/Vorgehensweisen,
- *everything is fucked* - Sanjay Srivastava
- welche Probleme gibt es außerhalb der Statistik? was ist der aktuelle Stand der Diskussion? wie sollte sich unsere wissenschaftliche Praxis verändern? wie kann uns Wissenschaftstheorie helfen?

Was ist multivariate Statistik?

ermitteln von statistischen Zusammenhängen zwischen... einem Prädiktor und einer abhängigen Variablen



mehreren Prädiktoren und einer abhängigen Variablen - multivariate Statistik im weiteren Sinne



einem oder mehreren Prädiktoren und mehreren abhängigen Variablen - multivariate Statistik im engeren Sinne



→ Wirksamkeit von Therapien

- was sind gemeinsame Faktoren eines Störungsbildes? → Faktorenanalyse

es dient dazu, aus empirischen Beobachtungen vieler verschiedener manifeste Variablen auf wenige zugrunde liegende latente Variablen zu schließen

- geht es Therapierten besser als Nicht-Therapierten? → ANOVA (Mittelwerte) oder Log-lineare Modelle - Häufigkeiten,

↓

Vergleich der Mittelwerte von mehr als 2 Gruppen

- hinsichtlich zahlreicher Symptome? → MANOVA

ähnlich der ANOVA mit mehreren aV

- auch unter Berücksichtigung von Anfangsunterschieden? → ANCOVA

Verbindung von Varianzanalyse und linearer Regressionsanalyse

- ab der wievielen Therapiesitzung verbessert sich der Zustand nicht mehr? → Regression

die Regression basiert auf der Korrelation und ermöglicht die bestmögliche Vorhersage für eine Variable - im Gegensatz zur Korrelation muss hierbei festgelegt werden, welche Variable durch eine andere Variable vorhergesagt werden soll - die Variable, die vorhergesagt werden soll, nennt man bei der Regression Kriterium.

- was unterscheidet erfolgreiche von nicht-erfolgreichen Therapien? → Diskriminanzanalyse

ermöglicht es, zwei oder mehr Gruppen simultan hinsichtlich einer Mehrzahl von Merkmalsvariablen zu untersuchen,

- was unterscheidet Patienten, bei denen die Therapie anschlägt von therapieresistenten? → logistische Regression

wird verwendet, um ein nominalskaliertes, kategoriales Kriterium vorherzusagen

→ 12.10. Einführung

→ 19.10. was ist Wissenschaft?

- Dienes, Z. (2008). *Understanding psychology as a science: an introduction to scientific and statistical inference*. Basingstoke: Palgrave Macmillan (Kapitel 1)

→ 26.10. was ist Wissenschaft?/Replikationskrise

→ 02.11. die Replikationskrise in der Psychologie

- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). DOI: 10.1126/science.aac4716

- Camerer et al. (2018). Estimating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*. DOI: 10.1038/s41562-018-0399-z
- 09.11. Signifikanztests in der Psychologie, Probleme und Alternativen
 - Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale: Erlbaum. (S. 311-339).
- 16.11. klassische Kritik am Signifikanztest
 - Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- 23.11. Effektstärken
 - Rosenthal, R., Rosnow, R.L & Rubin, D.B. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge: Cambridge University Press (Kapitel 2).
 - Borenstein, M. et al. (2010). *Introduction to meta-analysis*. Chichester: Wiley (S. 17 – 55)
- 30.11. mehr Probleme des Signifikanztests: QRPs, p -Hacking und ein Bayesianischer (Aus-)Blick
 - Sedlmeier und Renkewitz (2018), Kapitel 20. Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med*, 2 (8), 696- 701.
- 07.12. mehr Probleme des Signifikanztests: QRPs, p -Hacking und ein Bayesianischer (Aus-)Blick
- 14.12. Konfidenzintervalle
 - Cumming, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170-180.
 - Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532-574.
- 04.01. Metaanalyse
 - Borenstein, M. et al. (2010). *Introduction to meta-analysis*. Chichester: Wiley
- 11.01. Metaanalyse
- 18.01. Grundlagen der Bayes-Statistik
 - Sedlmeier und Renkewitz (2018), Kapitel 22.
- 25.01. Wiederholung

allgemeine Literaturempfehlungen

- Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
- Cumming, G. (2012). *Understanding the new statistics : effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Sedlmeier P. & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium
- Dienes, Z. (2008). *Understanding psychology as a science: an introduction to scientific and statistical inference*. Basingstoke: Palgrave Macmillan
- Cohen, J., Cohen, P & West, S.G. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences*. Mahwah: Lawrence Erlbaum
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah: Lawrence Erlbaum
- Tabachnick, B.G. & Fidell, L.S. (2013). *Using multivariate statistics*. Boston: Pearson.

- Grimm, L.G. & Yarnold, P.R. (2000). Reading and understanding multivariate statistics. Washington, DC: APA

Eid & Cohen - Standardwerke

19. Oktober - was ist Wissenschaft? - Karl Popper und das Demarkationsproblem

Facial Feedback Hypothese

Gesichtsmuskelbewegungen beeinflussen das eigene emotionale Erleben - Personen, die bspw. angehalten werden, während einer Veranstaltung zu lächeln, werden diese Veranstaltung im Nachhinein wahrscheinlich als positiver und vergnüglicher empfinden als Personen, die ihre Augenbrauen zusammengezogen haben

→ Reaktionen

- Moderatoren: Vorwissen, Videoaufzeichnung, *was ist eigentlich lustig?...*
- *I don't see what we've learned. [...] Given these eight nonreplications, I'm not changing my mind. I have no reason to change my mind.* - Fritz Strack
- natürlich kann die Facial Feedback Hypothese nach wie vor korrekt sein,
- aber warum haben wir dieses Experiment je durchgeführt, rezipiert, zitiert, gelehrt, wenn ein völlig veränderter Befund nichts an unseren Überzeugungen ändert?
- im Sinne der immer wichtiger werdenden Reproduzierbarkeit von Studienergebnissen in der Psychologie wurde diese Hypothese überprüft: dazu wurde das Experiment einem einheitlichen Versuchsprotokoll folgend, wiederholt - der ursprünglich berichtete Effekt in konnte dabei nicht repliziert werden - als Folge wird die Gültigkeit kontrovers diskutiert,
- es konnte jedoch geklärt werden, warum die Befunde nicht repliziert werden konnten - dies ist vor allem darauf zurückzuführen, dass andere eine Kamera auf die Versuchspersonen richteten, die einen direkten Einfluss auf die Urteilsprozesse hatte - dies bestätigten Noah et. al in 2018, indem sie die Kameraverwendung experimentell variierten und fanden, dass der ursprüngliche Effekt ohne Kamera repliziert, während er unter der Kamerabedingung verschwindet,
- in 2019 wurde eine Metaanalyse veröffentlicht, welche die Gültigkeit der Facial-Feedback-Hypothese bestätigt,

logischer Positivismus ≡ logischer Empirismus

wissenschaftstheoretische Strömung, die sich zum Ziel setzte, die Philosophie nach wissenschaftlichen und objektiven Kriterien zu erneuern

- Wiener Kreis: Schlick, Carnap, von Mises, Gödel, Hempel, Quine...
- 2 Typen bedeutungsvoller Aussagen:
 - Definitionen - notwendig wahr
 - verifizierbare (empirische) Aussagen - Verfizierbarkeitskriterium
- theoretische Begriffe (Konzepte) müssen also mit beobachtbaren/messbaren Sachverhalten in Verbindung gesetzt werden,
 - operationale Definition - spezifiziert die zur Messung notwendigen Operationen - Elektron? Intelligenz?
- Verifikation von Aussagen über spezifische Objekte
 - Beobachtungssatz - nach Schlick/Carnap/Quine,

Aussagen, über deren Gültigkeit durch sinnliche Beobachtung eine intersubjektive Übereinkunft erzielt werden kann

- Protokollsatz - Neurath,

Person X hat zur Zeit t am Ort O einen weißen Tisch wahrgenommen
- Basissatz - Popper

statt die Person X hat zur Zeit t am Ort O wahrgenommen: der Tisch ist weiß die Form zur Zeit t am Ort O ist der Tisch weiß, weil dadurch eine objektive Aussage über den Tisch selbst gemacht wird, nicht lediglich eine Aussage über den subjektiven Eindruck einer bestimmten Person
- widerlegbar? intersubjektiv? Konvention? Theoriefrei? reduktionistisch?

das Induktionsproblem

es bezieht sich auf die Frage, ob und wann ein Schluss durch Induktion von Einzelfällen auf ein allgemeingültiges Gesetz zulässig ist

→ Verifikation von All-Aussagen?

**Sam the swan is white;
Georgina the swan is white;
Fred the swan is white;**

...

Emma the swan is white

Conclusion: All swans are white (?)

- klappt nicht...
 - die All-Aussage ist nie *sicher wahr*
 - sie wird durch beliebig viele positive Instanzen auch nicht *wahrscheinlicher*,
 - * der logische Empirismus wurde wegen der Probleme des Reduktionismus und Verifikationismus aufgegeben,
- wir können also nicht zu *sicherer Erkenntnis* gelangen - was tut Wissenschaft dann??
 - Fallibilismus und Falsifikationismus - Popper

kritischer Rationalismus - Popper

- zentrales Problem: Wahl zwischen Theorien/Erklärungen
- zentrale These: auch ohne Induktion möglich
- logische Grundlage: Theorien sind nicht verifizierbar, aber **falsifizierbar**
 - ein schwarzer Schwan...
 - typische Form von Theorien: *wenn A, dann B*
 - * die Beobachtung *A → nicht B* ist eine Falsifikation,
- Widerlegung einer wissenschaftlichen Aussage durch ein Gegenbeispiel
- wir können also kein sicheres Wissen erlangen - aber wir können unsere Vermutungen verbessern,
- zentrales Mittel: Kritik/kritische Diskussion!
- Ihre Aufgabe hier: Sagen Sie mir, wo und warum ich irre!
- Ihre Aufgabe als Wissenschaftler: kritisieren Sie empirisch **und theoretisch** die aktuellen Vermutungen! Finden Sie bessere Vermutungen!

psychologische Forschungspraxis...

- der Signifikanztest prüft Null- und Alternativhypotesen,
 - Alternativhypothese entspricht der inhaltlichen Forschungshypothese,

- Nullhypothese als Gegenteil der Alternativhypothese - ein *Strohmann*, der nach Möglichkeit widerlegt werden soll,
 - ein signifikantes Ergebnis führt zur Zurückweisung der Nullhypothese und Annahme der Alternativhypothese,
 - nicht signifikantes Ergebnis?? führt in verschiedenen Formen des Signifikanztests (und unter bestimmten Bedingungen) zu *keiner Schlussfolgerung* (und nicht zur Annahme der Nullhypothese),
- der Signifikanztest zielt in die falsche Richtung: auf Bestätigung, nicht Widerlegung einer Vermutung (Meehl, 1978),
- sammelt psychologische Empirie *positive Instanzen*, statt kritisch zu prüfen? - soweit sie dies tut, ist sie epistemologisch bedeutungslos - laut Popper

Was ist Wissenschaft?

- woher kommen Theorien und Vermutungen?
- sie entspringen Ihrer Kreativität...
 - irrelevant für die Wissenschaftstheorie und die Frage nach der Unterscheidung von Wissenschaft und Nicht-Wissenschaft,
 - eher Gegenstand der Psychologie → Entdeckungszusammenhang – Reichenbach,

Kontext, in dem i. R. des Forschungsprozesses neue Ideen gewonnen, Vermutungen oder Hypothesen gebildet und Forschungsziele definiert werden

- wie werden Theorien und Vermutungen geprüft?
- Gegenstand der Wissenschaftstheorie → Begründungszusammenhang - Reichenbach
- alle Forschungsoperationen, die zur Bestätigung (Verifikation) oder Widerlegung (Falsifikation) der zu überprüfenden bzw. empirisch zu begründenden Theorien und Hypothesen erforderlich sind
- hier liegt die Trennung von Wissenschaft und Nicht-Wissenschaft
 - logische Beziehung zwischen Empirie und Theorie?
 - * wie kommen empirische Beobachtung und Theorie *in Kontakt*?
 - * wie können empirische Beobachtungen Ihre Überzeugungen ändern, wenn Induktion Theorien weder beweisen noch ihre Wahrscheinlichkeit erhöhen kann?
 - * auf Grundlage von Deduktion...
- Ableitung des Besonderen und Einzelnen vom Allgemeinen; Erkenntnis des Einzelfalles durch ein allgemeines Gesetz

Demarkation

- Deduktion: aus Theorien werden Vorhersagen über beobachtbare Ereignisse abgeleitet - diese werden möglichst kritisch geprüft,
- Voraussetzung: Theorien müssen mögliche Ereignisse ausschließen - sie müssen also Vorhersagen treffen, die falsch sein können,
- Demarkationskriterium: Falsifizierbarkeit
- wir können aus unseren Beobachtungen nur dann ein Feedback erhalten, wenn wir irren können,
 - kein Lernen ohne Feedback - Erkenntnisfortschritt durch trial-and-error,
- nicht-falsifizierbare Aussagen fallen in den Bereich der Metaphysik (oder Pseudo-Wissenschaft) - sie sind damit nicht notwendigerweise sinnlos,
- auch im Bereich der Metaphysik kann Wissen entstehen - durch kritische Diskussion,
- Wissenschaft entsteht, wenn
- falsifizierbare Theorien aufgestellt **und** ernsthafte Falsifikationsversuche

- unternommen werden,
- Wissenschaft ist daher auf *Skepsis* angewiesen,
- die *scientific community* benötigt eine falsifikationistische Haltung,

Wissen?

- Wissen ist stets vorläufig - es gibt kein definitives empirisches Wissen,
- Theorien werden nie beweisen,
 - Theorien, die zahlreiche Falsifikationsversuche überstehen, gelten als bewährt - *corroborated*,
- Poppers Lieblingsbeispiele für Nicht-Wissenschaften: Marxismus und Psychoanalyse,
 - beide Theorien erlauben durchaus die Ableitung falsifizierbarer Vorhersagen,
 - zentral: werden die Ergebnisse empirischer Prüfungen genutzt, um die Theorie zu verändern/zu verbessern?
 - was fehlt, ist eher die falsifikationistische Haltung...
- *welche Beobachtung würde meine Überzeugung über die Richtigkeit der Theorie X verändern?*
 - wenn Sie diese Frage nicht beantworten können, ist Ihre Überzeugung kein empirisches Wissen,
 - ohne eine Antwort kann Ihre Überzeugung auch kein empirisches Wissen werden,

Warum bevorzugen Menschen oft das Essen, mit dem sie aufgewachsen sind, aber nicht immer? - betrachten Sie die folgende Zwei-Faktoren-Theorie des Mögens:

- Faktor 1: wir sind darauf programmiert, vertraute Dinge (z.B. Lebensmittel, Menschen, Tiere, Werkzeuge usw.) zu mögen, weil unser Wissen und unsere Fähigkeiten wahrscheinlich darauf anwendbar sind - sie sind nicht gefährlich, wir können mit ihnen umgehen - es gibt also einen Mechanismus, der uns automatisch dazu bringt, Dinge zu mögen, wenn wir ihnen häufiger begegnen,
- Faktor 2: Bekanntes langweilt uns aber auch, weil es wenig zu lernen gibt und wir einen Lerndrang haben,

diese beiden Faktoren wirken gegensätzlich - es kann also möglich sein, die Exposition der Menschen für eine neue Sache zu erhöhen:

- steigern Sie die Zuneigung der Leute, weil die Vertrautheit bedeutet, dass es sicher ist - Faktor 1 in Betrieb,
- verringern Sie die Zuneigung der Leute, weil sie sich langweilen - Faktor 2 in Betrieb,
- zuerst die Zuneigung steigern, dann abnehmen, weil der erste Faktor anfangs wirkt, bevor die Langeweile stärker eintritt
- zuerst abnehmen und dann die Zuneigung steigern, da zunächst Langeweile auftritt, bevor der erste Faktor stärker wirkt,

Die Theorie ist gut, weil sie all diese Ergebnisse erklärt.

Diskutieren!

Theorie

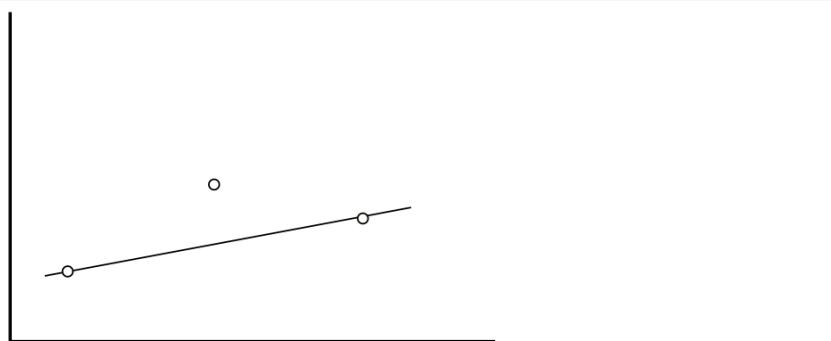
- Menge von *wenn-dann-Aussagen* - arg vereinfacht
- benötigt eine logische nicht empirische Evaluation, sie muss ja falsifizierbar sein,
- wie?
- logische Widerspruchsfreiheit

- empirischer Gehalt:
 - die Menge an Beobachtungen, die durch eine Theorie ausgeschlossen wird,
 - *bessere* Theorien haben einen höheren empirischen Gehalt,
 - Theorien sind also umso besser, je leichter sie falsifiziert werden können,
 - * je **informativer** sie sind!

empirischer Gehalt

- Allgemeinheit: *wenn*-Teil
 - der empirische Gehalt wächst mit der Allgemeinheit - also mit der Größe des Anwendungsbereichs,
- Bestimmtheit: *dann*-Teil
 - Präzision der Vorhersage,
 - der empirische Gehalt wächst mit der Bestimmtheit - also bei präziseren Vorhersagen,
- Beispiele:
 - die Gruppen A und B unterscheiden sich hinsichtlich der Variable X,
 - die Variablen A und B korrelieren,
 - ! schließt nahezu nichts aus → extrem geringe Bestimmtheit → nahezu nicht falsifizierbar → rangiert bestenfalls im *Grenzbereich* von Wissenschaft (Meehl, 1967),
 - die Gruppe A erzielt bessere Ergebnisse als die Gruppe B,
 - die Variablen A und B korrelieren positiv,
 - ! besser, aber immer noch extrem schwach,
 - ! die schwächste mögliche Vorhersage im Bereich der Wissenschaft... (Meehl, 1967)

Figure 1.1

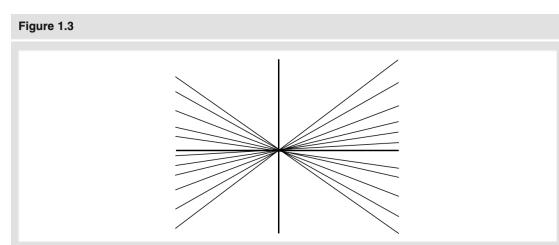
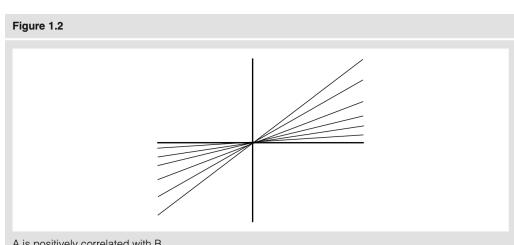


Linear versus quadratic

3Datenpunkte...

...können die Vermutung eines linearen Zusammenhangs falsifizieren,

...können die Vermutung eines quadratischen Zusammenhangs nicht falsifizieren,



schließt wenig aus

schließt nahezu nichts aus

Theorierevision

- was passiert bei Falsifikationen?
- es ist stets möglich, ad hoc (post hoc) Erklärungen für Falsifikationen zu finden - siehe Stracks Reaktion,
- Popper: Revisionen einer Theorie sollten stets einen höheren empirischen Gehalt aufweisen! dasselbe gilt für neue, alternative Theorien,
- diese Anforderung ist vermutlich sehr strikt,
 - der empirische Gehalt sollte zumindest nicht sinken,
 - die Revision muss in jedem Fall falsifizierbar bleiben,
 - es müssen Falsifikationsversuche der Revision unternommen werden - falsifikationistische Haltung,
 - * andernfalls haben wir schlicht das Erkenntnisinteresse verloren und lediglich eine Überzeugung gerechtfertigt,

Falsifizierbarkeit

- Falsifikationen können immer als ungültig zurückgewiesen werden, indem die entsprechende Beobachtung bezweifelt wird,
- möglich, weil die Beobachtung selbst nie theoriefrei ist,
 - wer ist tatsächlich extrovertiert?
 - sind Kreuzchen auf Likert-Skalen ein angemessenes Maß für *Lustigkeit*?
 - *hier ist ein Glas Wasser*
- wir müssen *entscheiden*, welche Beobachtungen wir akzeptieren - dies geschieht auch auf Basis von Übereinkunft und Konvention,
- wenn Ihre Überzeugungen falsifizierbar bleiben sollen, müssen Sie vorab spezifizieren, welche Beobachtungen sie akzeptieren!
- sie sollten diese Entscheidung nicht rückgängig machen - sie sollten generell bereit sein, irgendwelche falsifizierenden Beobachtungen zu akzeptieren,

das Duhem-Quine-Problem

Behauptung der Unterbestimmtheit einer Theorie durch Beobachtungsdaten - demnach besteht eine Theorie aus vielen miteinander verknüpften Aussagen, die zusammen ein möglichst zusammenhängendes Ganzes bilden

- um Theorien (T) zu testen, werden stets Zusatzannahmen (Hilfshypothesen) benötigt,
- nach einer falsifizierenden Beobachtung könnte also die Theorie falsch sein - oder irgendeine der benötigten Zusatzannahmen (oder mehrere),
- wie entscheiden wir, was wir als falsifiziert betrachten?
- Poppers Antwort: wir benötigen Hintergrundwissen,
 - irgendein Teil unseres Wissens muss zumindest vorläufig und für den aktuellen Zweck als unproblematisch akzeptiert werden,
 - nur so wird Kritik an der Theorie möglich...
 - zentrale Idee: führt die Verwendung unterschiedlichen Hintergrundwissens zu derselben
 - Schlussfolgerung im Hinblick auf die Theorie?
 - * konzeptuelle Replikation
- in jedem Fall sollte die falsifizierende Beobachtung dazu führen, dass irgendein Element als falsifiziert betrachtet wird!
- zentrale Frage der deskriptiven Wissenschaftstheorie:
 - was (welche Annahme oder Theorie) wird in der Wissenschaft unter welchen Bedingungen tatsächlich aufgegeben?

- unterschiedliche Antworten bei Kuhn, Lakatos...

Wann sollten Theorien aufgegeben werden?

- sollte eine Theorie nach einer Falsifikation aufgegeben werden? ...nach mehreren?
...dann in jedem Fall?
- vermutlich das umstrittenste Problem...
- auch gegen überaus erfolgreiche Theorien sprachen stets zumindest einzelne
- falsifizierende Beobachtungen,
- war unser Hintergrundwissen (die Zusatzannahmen) doch falsch? - es gilt das Prinzip des Fallibilismus,

es kann keine absolute Gewissheit geben und Irrtümer lassen sich niemals ausschließen
- Popper nahm in seinem späteren Werk an, dass auch *falsche Theorien* als gegenwärtig beste Theorien akzeptiert werden könnten,

Wahrheitsähnlichkeit - *Verisimilitude*

- alle Theorien und Modelle sind Vereinfachungen der Welt und damit notwendig falsch,
- was wir anstreben ist Wahrheitsähnlichkeit!
- eine Theorie T2 ist wahrheitsähnlicher als eine Theorie T1 wenn sie...
 - ...präzisere Vorhersagen macht - höhere Bestimmtheit aufweist **oder**
 - ...mehr Beobachtungen erklären kann - größere Allgemeinheit aufweist **oder**
 - ...mehr kritische Tests überstanden hat,
- wir können T2 also selbst dann bevorzugen, wenn sie auch falsche Vorhersagen macht,

Was ist mit probabilistischen Theorien?

- *Frustration erzeugt Aggression*
 - diese Hypothese meint nicht, dass jede frustrierte Person sich allzeit aggressiver verhalten wird als jede nicht-frustrierte Person,
 - sie meint eher, dass die Aggressivität in der Population frustrierter Person im Mittel größer ausgeprägt ist als in der Population nicht-frustrierter Personen,
 - dies schließt nicht aus, dass wir in einer Stichprobe beliebiger Größe keinen Unterschied beobachten,
 - wie falsifizieren?
- wie immer: kritischer Test!
 - hier: wir benötigen ein Ergebnis, dass bei Gültigkeit der Hypothese deutlich wahrscheinlicher ist als bei Ungültigkeit der Hypothese,
 - nahe an zentralen Ideen des Signifikanztests - aber mit ein paar schwerwiegenden Problemen...

Wissenschaftssoziologie

- wovon hängen Fortschritt und Objektivität in der Wissenschaft ab?
 - nicht so sehr von Kreativität, besonderen Ideen, genialen Individuen...
 - viel mehr von freier Kritik, Diskussion, von frei geäußertem Zweifel,
 - dies ist ein soziales - und offenkundig schwieriges - Unterfangen,
 - von ihrer Kritikfähigkeit und -bereitschaft,
 - von ihrer Bereitschaft, Kritik an der eigenen Arbeit zu ermöglichen und zu suchen,
 - von ihrer Bereitschaft, Zweifel auszuhalten, soziale Unannehmlichkeiten auf sich zu nehmen, Frustration zu tolerieren...

- wissenschaftliche Institutionen sollten Kritik und kritische Diskussion fördern und schützen,
 - sie tun das ganz sicher keineswegs immer!

Platt'scher Test - Platt, 1964

- *strong inference*
- wie wird empirische Kritik an Theorien gefördert?
- indem alternative Theorien gesucht und gegebenenfalls generiert werden!
- sobald mehrere Theorien vorhanden sind, *experimentum crucis* durchführen,

Experiment, dessen Scheitern die dem Experiment zugrunde liegende Theorie falsifiziert oder überwindet

- Theorien gegeneinander testen!

vernünftige Fragen an hypothesenprüfende Paper

- gibt es eine inhaltliche Theorie, die klare Vorhersagen macht?
- wie unterscheidet sich die Theorie von den statistischen Hypothesen?
- wie eindeutig sind die Vorhersagen? ...alternativlos? ...völlig beliebig?
- wie präzise sind die Vorhersagen?
- werden Hilfshypothesen benannt? ...welche Hilfshypothesen sind relevant?
...akzeptieren sie diese als *Hintergrundwissen*? ...wären andere Hilfshypothesen denkbar?
...würden diese die Vorhersagen ändern?
- kennen Sie eine andere Theorie, die dieselben Vorhersagen treffen würde? ...können Sie sich eine solche Theorie auf dem Weg in die Küche ausdenken? gibt es eine Theorie, die durch die vorliegenden Daten falsifiziert wird?
- was würden Sie über die Theorie denken, wenn die Ergebnisse umgekehrt oder irgendwie anders wären?
- hat der Test eine ausreichende statistische Power?
- andernfalls sind die Daten wenig aussagekräftig...

Aufgabenblatt - Sitzung 2: Popper - 27.Oktober

1. Was ist Epistemologie?

- was macht *Wissen* zu *wissenschaftlichem Wissen*?
- Frage nach den Bedingungen von begründetem Wissen

2. aus Perspektive des logischen Positivismus: was sind bedeutungsvolle Aussagen?

- bedeutungsvolle Aussagen sind entweder Definitionen und daher notwendigerweise wahr, wie ein *Dreieck hat drei Seiten*; oder auch überprüfbare empirische Aussagen,
- empirische Aussagen waren nur dann sinnvoll, wenn sie das Verifikationskriterium erfüllten, d.h. wenn man die Schritte angeben könnte, die die Richtigkeit der Aussage überprüfen würden,

3. was ist ein induktiver Schluss? ...erkläre an einem eigenen Beispiel!

- Art des Schlussfolgerns, die vom Besonderen auf das Allgemeine schließt,
- es werden allgemeine Erkenntnisse bzw. Theorien aus der Verallgemeinerung bzw. Abstraktion von Einzelphänomenen gewonnen,
- z.B. Amseln, Rotkehlchen, Adler und Enten können fliegen - daraus ließe sich der induktive Schluss ziehen, dass Vögel fliegen können - dies wäre eine falsifizierbare These, da bei näherer Überprüfung deutlich wird, dass auch Vögel existieren, die nicht fliegen können → Strauss, Pinguin, etc.
- Induktion ist der Prozess des Ableitens universeller Regeln nur aus bestimmten Beobachtungen,

4. was ist ein deduktiver Schluss? ...erkläre an einem eigenen Beispiel!

- Art der Schlussfolgerung, die vom Allgemeinen auf das Besondere schließt,
- es werden Einzelerkenntnisse bzw. Hypothesen aus allgemeinen Theorien abgeleitet,
- z.B. alle Säugetiere sind Warmblüter - Wale sind Säugetiere - daraus folgt: Wale sind Warmblüter,
- Deduktion ist der Prozess, bei dem Schlussfolgerungen gezogen werden, sodass die Schlussfolgerung garantiert wahr ist, wenn die Prämissen wahr sind,

5. was ist Humes Kritik an induktiven Schlüssen?

- Annahme, dass aus bestimmten Beobachtungen keine Verallgemeinerung mit Sicherheit folgt – aber sicherlich wird die Wahrscheinlichkeit der Verallgemeinerung mit jeder erhöht,
- ist es nicht wahrscheinlicher, dass jedes Mal, wenn wir einen weißen Schwan sehen, alle Schwäne weiß sind? - er wies darauf hin, dass dies nicht folgt,
- egal wie oft ein Auto am frühen Morgen gestartet ist, eines Tages wird es nicht mehr starten und eines Tages sind wir an dem Punkt, an dem das Auto nie wieder anspringt - denn dies ist der Einfluss des Alters auf Autos,

6. Was ist das Demarkationskriterium?

- jede wissenschaftliche Aussage muss falsifizierbar sein

Falsifizierbarkeit ist ein Kriterium, das empirische von nichtempirischen Aussagen abgrenzen soll

7. Was ist das Duhem-Quine-Problem und wie sieht Poppers Lösung dazu aus?

- Behauptung der Unterbestimmtheit einer Theorie durch Beobachtungsdaten,
- eine Theorie besteht demnach aus vielen miteinander verknüpften Aussagen, die zusammen ein möglichst kohärentes Ganzes bilden,
- man kann nur das Argumentationssystem als Ganzes falsifizieren, einschließlich Hilfshypothesen, Messtheorien und auch alle vorherigen Beobachtungen, die zur Generierung von Vorhersagen im aktuellen Fall verwendet wurden,
- wie können wir bei einer Fälschung wissen, welche Komponente des Systems abzulehnen ist?
- Popper: damit Kritik überhaupt auftritt, muss ein Teil unseres Wissens für aktuelle Zwecke als unproblematisch hingenommen werden - er nennt solches Wissen *Hintergrundwissen* - z.B. die Behauptung, dass das Gedächtnis in gewisser Weise von der kortikalen Erregung abhängt - d.h., wir müssen eine methodische Entscheidung treffen, wir müssen zumindest einen Teil der bisherigen Forschung akzeptieren, um Fortschritte zu erzielen,

02.November - Replizierbarkeit psychologischer Forschungsbefunde

Zweifel an Replizierbarkeit?

Anreizstruktur in der Wissenschaft - Nosek, Spies, & Motyl (2012)

- neuartige, statistische signifikante, hypothesenbestätigende, konsistente Befunde
- Schwierigkeiten, Null-Ergebnisse zu publizieren,
- Schwierigkeiten, Replikationen zu publizieren,
- Quantität, Geschwindigkeit

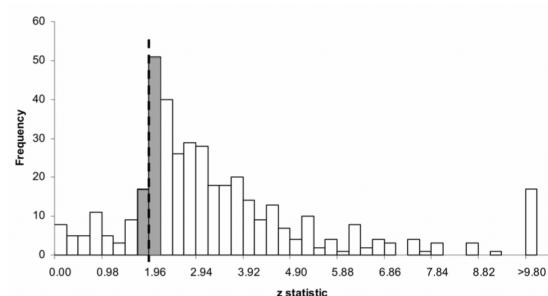
questionable research practices - Simmons, Nelson, & Simonsohn (2011)

- *p*-Hacking
- HARKing
- selektive Veröffentlichung von Experimenten/Studien

kleine Stichproben, kleine Effekte → niedrige Power

- geschätzte mittlere Power: .35 - Bakker, van Dijk, & Wicherts (2012)

überzeugende Evidenz?



Wie viele direkte Replikationen werden veröffentlicht?

Makel, Plucker, & Hegarty (2012)

- Schätzung: 1% Replikationen, 0.15% direkte Replikationen
- Erfolgsquote: 92% bei Replikationen durch denselben Autor, andernfalls 65%

Neuliep & Crandall (1993)

- drei Ausgaben von JPSP im Jahr 1993, 42 Artikel
- 79% Replikationen; keine direkten Replikationen

Sterling (1959)

- vier Zeitschriften im Jahr 1955, 362 Artikel
- keine direkten Replikationen

Reproducibility Project: Psychology (RP:P)

→ Ziel:

- empirische Untersuchung der Reproduzierbarkeit psychologischer Forschungsbefunde
- Identifikation von Moderatoren

- Crowdsourcing-Projekt mit 270 Teilnehmern aus (etwa) 70 Institutionen und 11 Ländern
- koordiniert und finanziell unterstützt durch das *Center for Open Science* (COS; Brian Nosek)
- Project Management Site: Open Science Framework – <http://osf.io>
- Start im November 2011

Design

zentrale Designprobleme

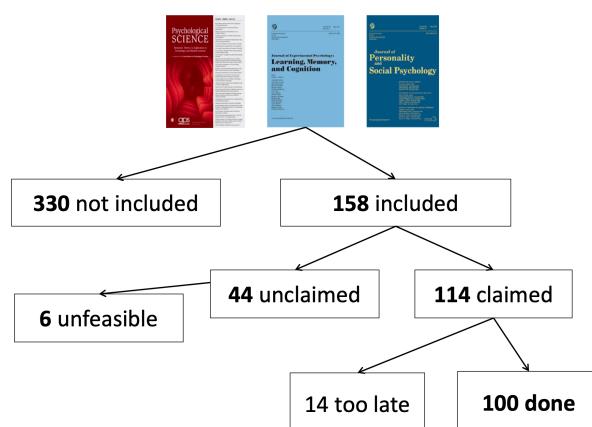
1. generalisierbare Ergebnisse - Stichprobenauswahl
2. Replikationen von hoher Qualität - strukturierter Replikationsprozess

1. generalisierbare Ergebnisse

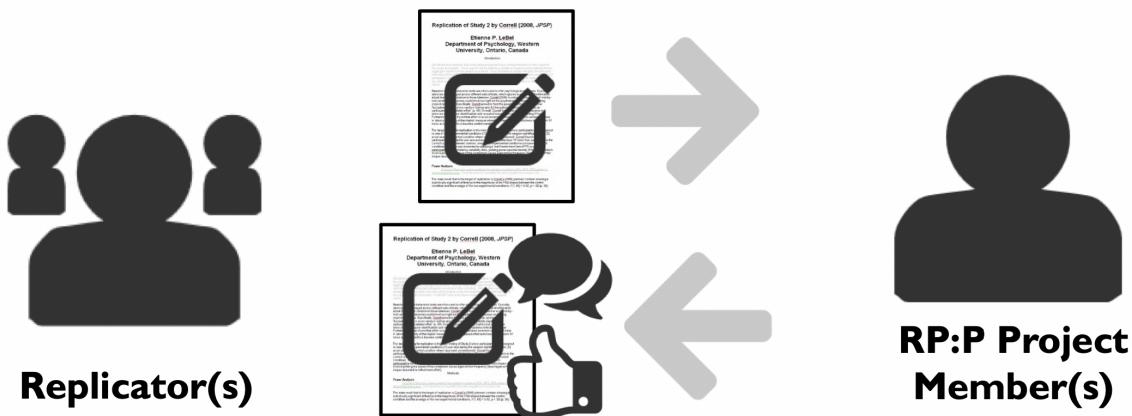
→ Stichprobenauswahl:

- maximal repräsentativ - also vollständige Zufallsauswahl aus der gesamten Psychologie
 - * potentiell zahlreiche sehr aufwendige, schwierige Replikationen → geringe Teilnahme,
 - * zu kleine Stichproben aus Teilgebieten der Psychologie
- maximale Teilnahme → Replikations-Teams können eine Studie frei auswählen,
 - * Selektions-Bias: einfache oder *verdächtige* Studien

→ Festlegung von 3 Journals (Jahrgang 2008): PSCI, JEP, LMC, JPSP



2. Qualität der Replikationen
 - hohe statistische Power
 - ! 93% der Replikationen > 80% Power
 - ! Median der Power = 95%!
 - Originalmaterialien
 - ! 89% der Replikationen konnten mit den Originalmaterialien arbeiten
 - standardisiertes Replikationsprotokoll
 - geplante Stichprobengröße und Sampling-Prozedur
 - exakte Prozedur & Instruktionen
 - Materialien & Maße
 - exakter Plan für Datenaufbereitung & -analyse
 - ! 100% der Replikationsstudien erstellten ein Protokoll vor der Datenerhebung
 - Prüfung des Replikationsprozesses



- ! 90% der Protokolle wurden durch Projektmitglieder **und** Originalautoren geprüft
- öffentlicher Bericht
 - ! 100% der Replikationen veröffentlichten vor der Datenerhebung ein Protokoll
- Design des RP:P
 - solides Sampling kombiniert mit qualitativ hochwertigem Replikationsprozess
 - ! extrem solide!

Ergebnisse

was zählt als erfolgreiche Replikation?

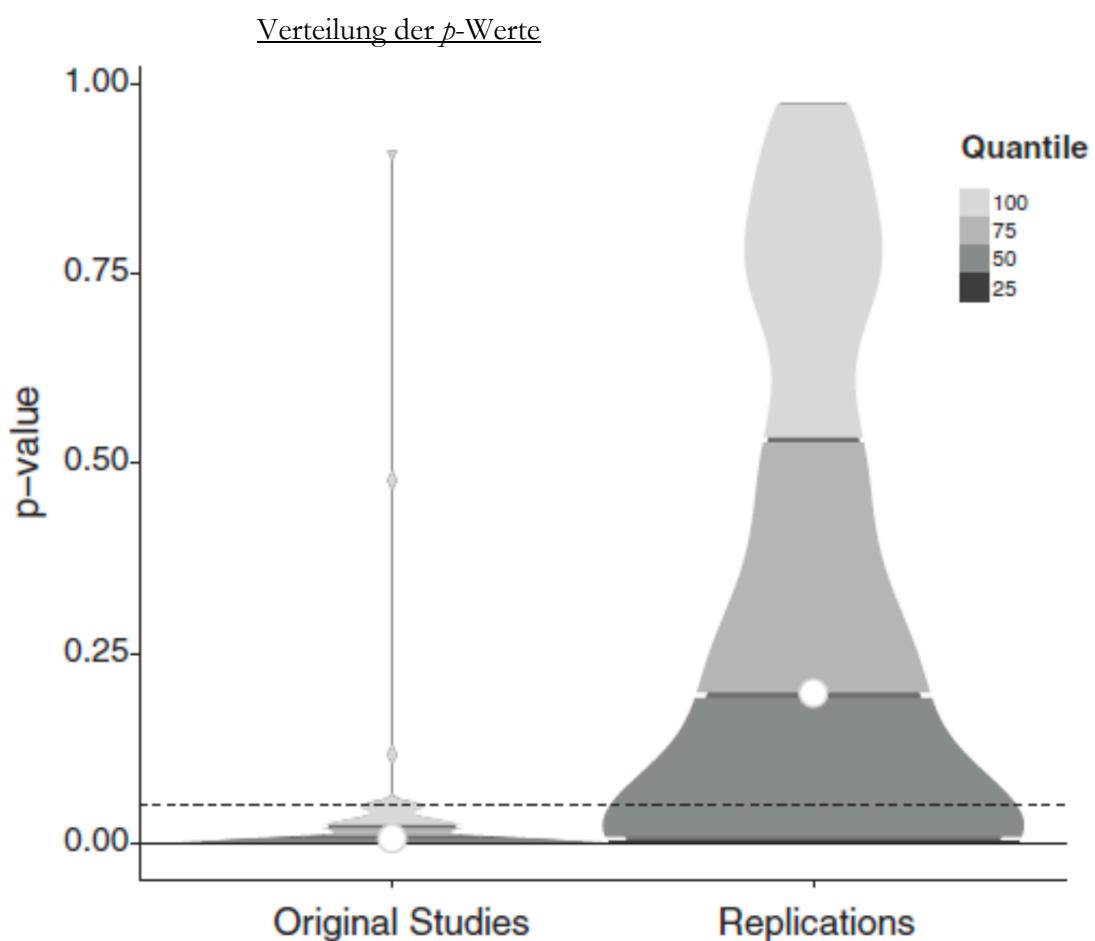
- Fokus auf drei Kriterien:
 - Anzahl der signifikanten Ergebnisse ($p < .05$)
 - Vergleich der Effektstärken
 - meta-analytische Effektstärkenschätzungen

wie viele der Replikationen erzielen signifikante Ergebnisse?

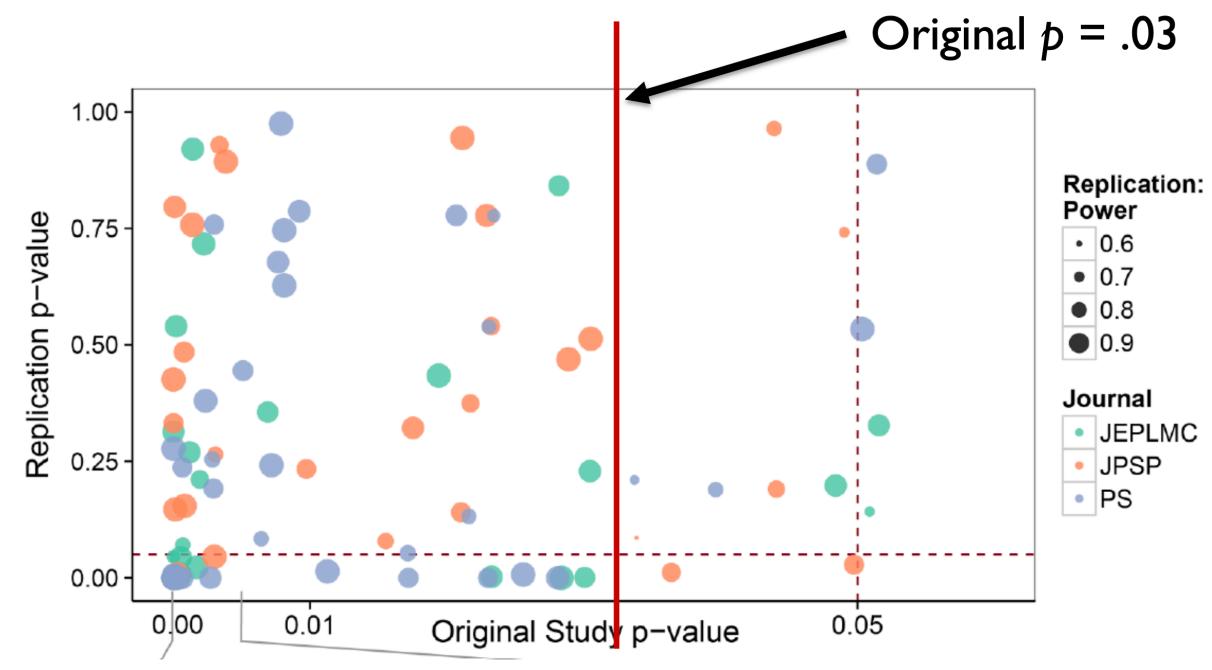
Journal	N Studien	N signifikante Ergebnisse	% signifikante Ergebnisse
JEP: LMC			
JPSP			
PsychScience			
Gesamt	97	35	36

Journal	N Studien	N signifikante Ergebnisse	% signifikante Ergebnisse
JEP: LMC	27	13	48
JPSP	31	7	23
PsychScience	39	15	39
Gesamt	97	35	36

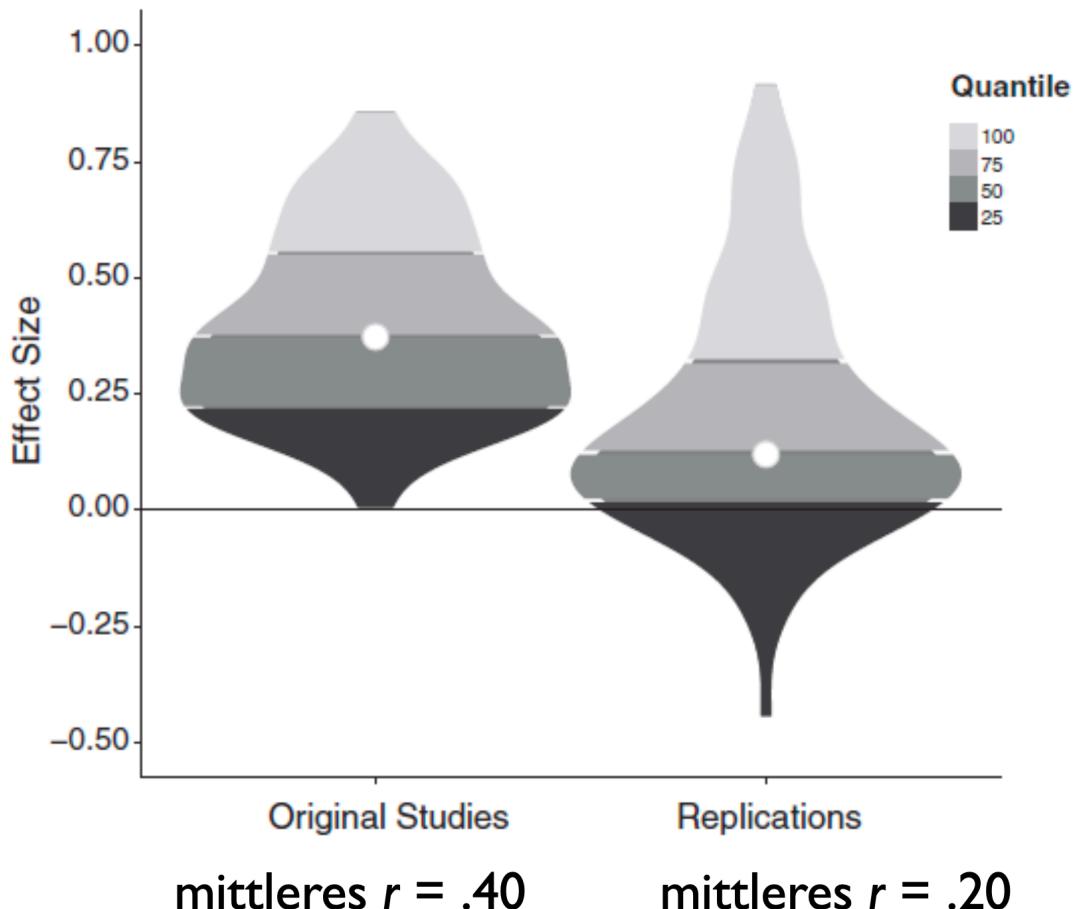
Sub-Disziplin	N Studien	N signifikante Ergebnisse	% signifikante Ergebnisse
Kognitiv	38	20	53
Sozial	53	14	26
Andere	6	1	17
Gesamt	97	35	36



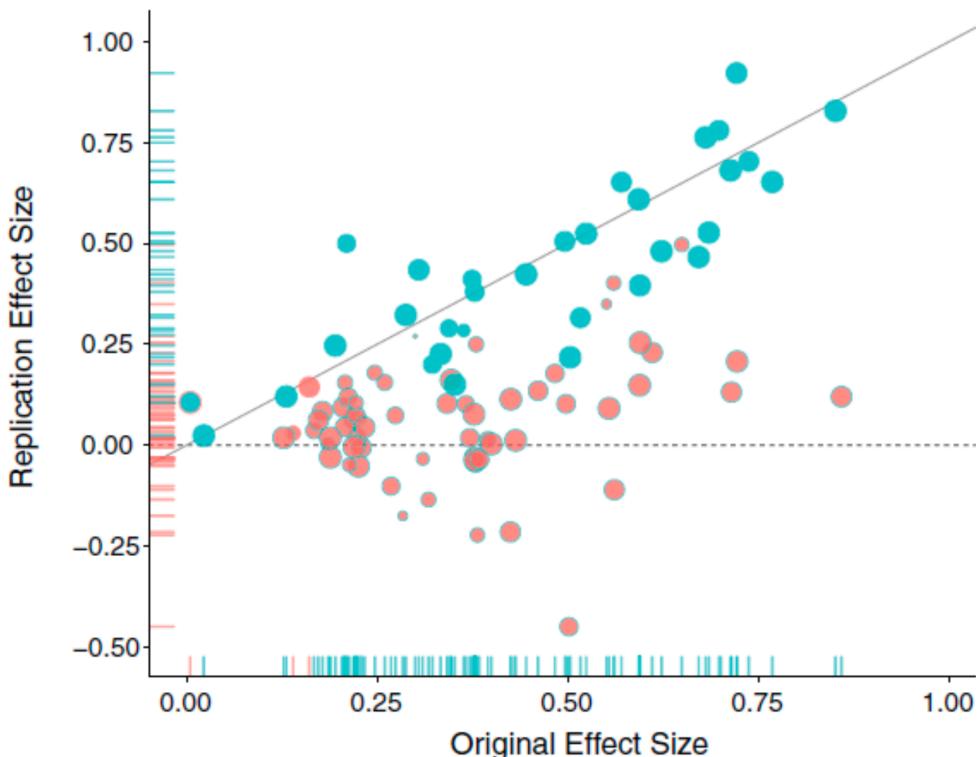
Streudiagramm der p -Werte



Effektstärken



Streudiagramm der Effektstärken



Effektstärken

Journal	Mittlere ES d. Originalstudien	Mittlere ES d. Replikationen	% original ES in Replikations-CI
JEP: LMC	.46	.27	62
JPSP	.29	.07	34
PsychScience	.43	.26	48
<i>Overall</i>	<i>.40 (N = 97)</i>	<i>.20 (N = 97)</i>	<i>47 (N = 95)</i>

Moderatoren

- Rangkorrelationen von Merkmalen der **Originalstudien** mit dem Signifikanzkriterium:

Merkmal	r
p-Wert	-.33
Effektstärke	.30
N	-.15
Relevanz des Effekts	-.12
Überraschungswert des Effekts	-.26
Erfahrung und Expertise des Teams	-.07

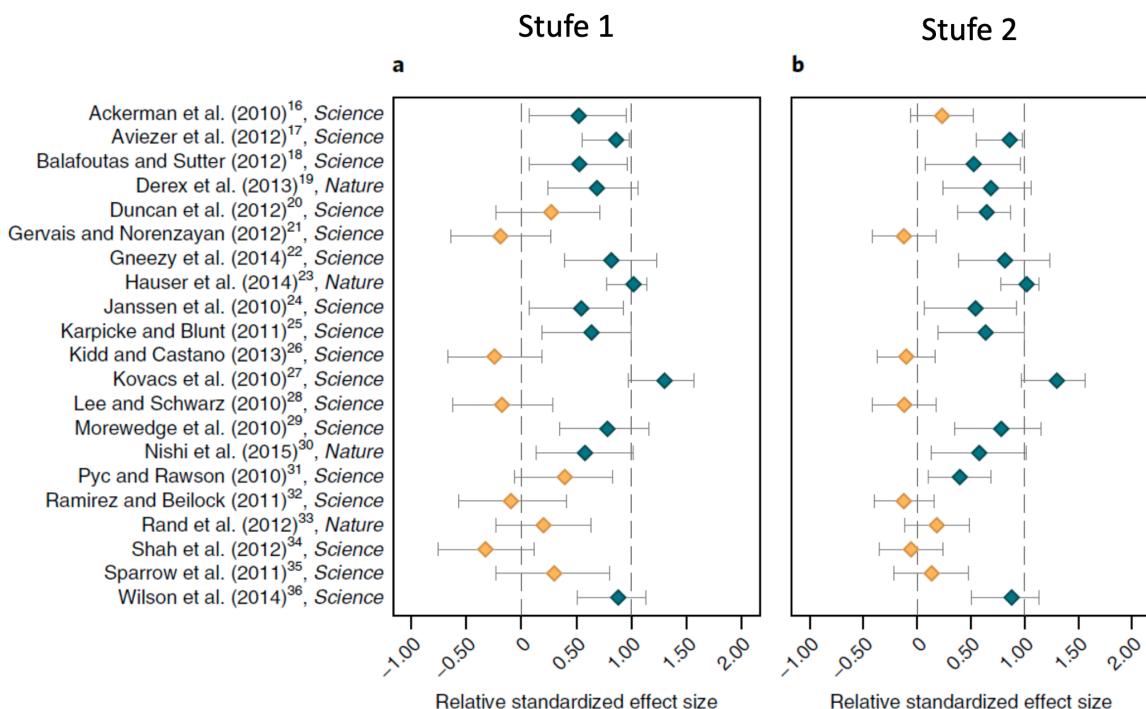
Moderatoren

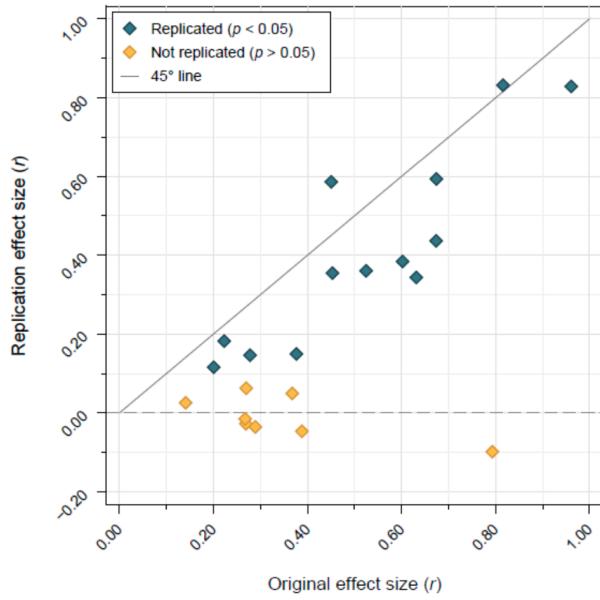
- Rangkorrelationen von Merkmalen der **Replikationsstudien** mit dem Signifikanzkriterium:

Merkmal	<i>r</i>
Power	.37
<i>N</i>	-.09
Schwierigkeit der Replikation	-.22
Erfahrung und Expertise des Teams	-.10
Qualität (Selbstbeurteilung)	-.07

Social Science Replication Project (Camerer et al., 2018)

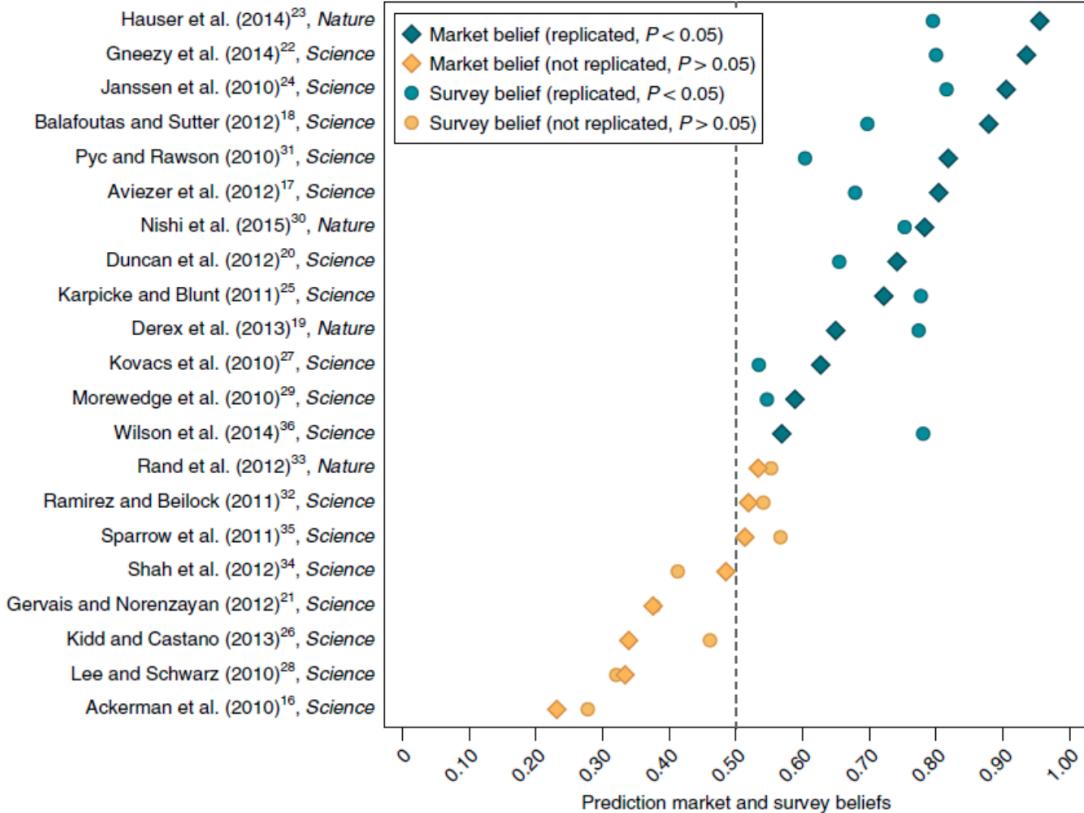
- Replikationen von 21 Studien aus *Nature* und *Science* - veröffentlicht zwischen 2010 und 2015
- zweistufiges Vorgehen
 - Replikation mit 90% Power für 75% der ursprünglichen Effektstärke
 - falls nicht signifikant: Erhöhung der Power auf 90% für 50% der Originalstudie
 - * die Stichprobengrößen der Replikationen sind in Stufe 1 dreimal größer als im Original - 6x in Stufe 2





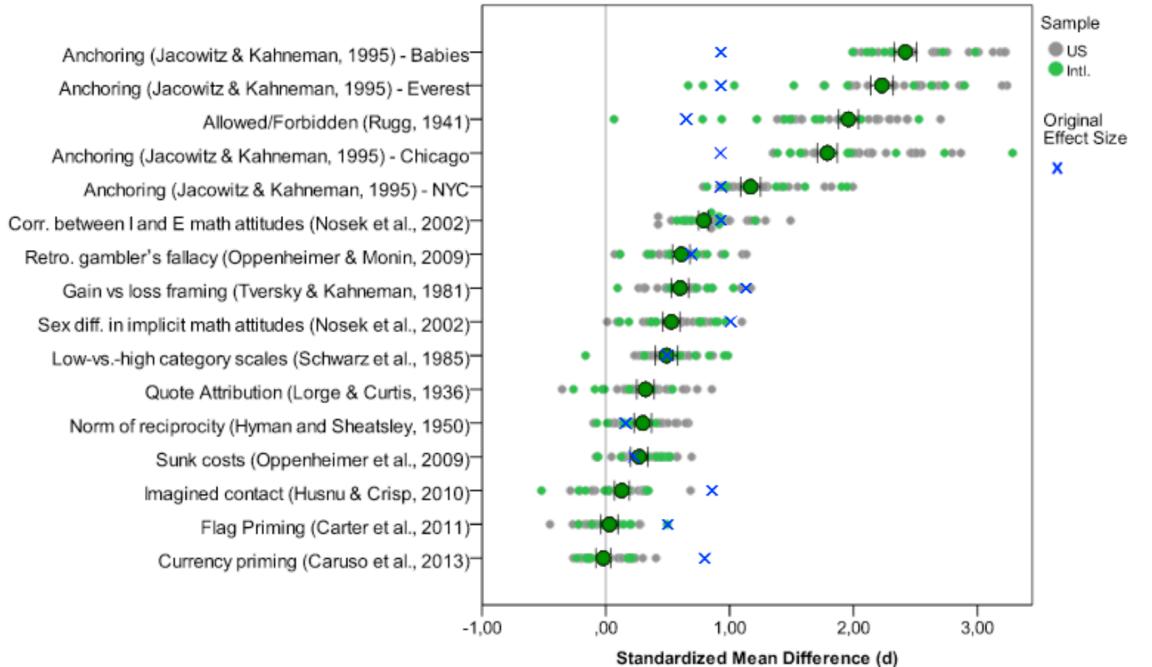
and the dotted line represents a replication effect size equal to zero. The mean standardized effect size (correlation coefficient, r) of the replications is 0.249 ($SD = 0.283$), compared to 0.459 ($SD = 0.229$) in the original studies. This difference is significant (Wilcoxon signed-ranks test, $n = 21$, $z = 3.667$, $p < 0.001$). The mean relative effect size of all the replications is 46.2% [95% CI = $(27.0\%, 65.5\%)$]; the mean relative effect size of the replications that replicated is 74.5% [95% CI = $(60.1\%, 88.9\%)$]; and the mean relative effect size of the replications that did not replicate is 0.3% [95% CI = $(-12.4\%, 13.1\%)$]. The Spearman correlation between the original effect size and the replication effect size is 0.574 [$p = 0.007$; 95% CI = $(18.9\%, 80.6\%)$].

Replikationserfolg ist vorhersagbar!



Diskussion

- etwa 40% - 60% der Befunde in psychologischen Top-Journals sind nicht replizierbar,
...aber nicht notwendigerweise *falsch*
- potentielle Gründe für scheiternde Replikationen?
 - falsch-negative Replikation
 - Kontextabhängigkeit - *hidden moderators*
 - falsch-positiver Originalbefund
- allerdings ist die ursprüngliche Evidenz für die geprüften Hypothesen schwach - und damit wenig aussagekräftig,



- wissenschaftlicher Fortschritt in der Psychologie?
 - zumindest nicht in sonderlich *systematischer* Weise...
 - wir verfehlten (in der Regel) bereits das *Eingangskriterium* für wissenschaftlichen Fortschritt – Replizierbarkeit,
- die Methodenlehre scheint nicht sonderlich erfolgreich zu sein
 - vermitteln wir die falschen Inhalte?
 - bleiben auch die richtigen Inhalte in der Anwendung wirkungslos?
- Gründe?
 - allgemein: Anreizstruktur im Wissenschaftsbetrieb
 - im Forschungs- und Publikationsprozess: fehlerhafter Einsatz ungeeigneter Methoden
 - wir werden unsere Forschungspraxis verändern müssen
 - sieht das in anderen Wissenschaften besser aus?
 - * Replikationsquote in der vorklinischen Krebsforschung:
 - 11% (Begley & Ellis, 2012),
 - 25% (Prinz et al., 2012)
 - eher nicht!

Richard Horton in *The Lancet* (4/2015):

Das Fazit des Symposiums war, dass etwas getan werden muss. Tatsächlich schienen sich alle darin einig zu sein, dass es in unserer Macht stand, dieses Etwas zu tun. Aber was genau zu tun ist oder wie es zu tun ist - dafür gab es keine festen Antworten. Diejenigen, die die Macht haben, zu handeln, scheinen zu denken, dass jemand anderes zuerst handeln sollte. Und jede positive Aktion (z.B. die Finanzierung leistungsstarker Replikationen) hat ein Gegenargument (die Wissenschaft wird weniger kreativ). Die gute Nachricht ist, dass die Wissenschaft einige ihrer schlimmsten Fehler sehr ernst nimmt. Die schlechte Nachricht ist, dass niemand bereit ist, den ersten Schritt zur Bereinigung des Systems zu tun.

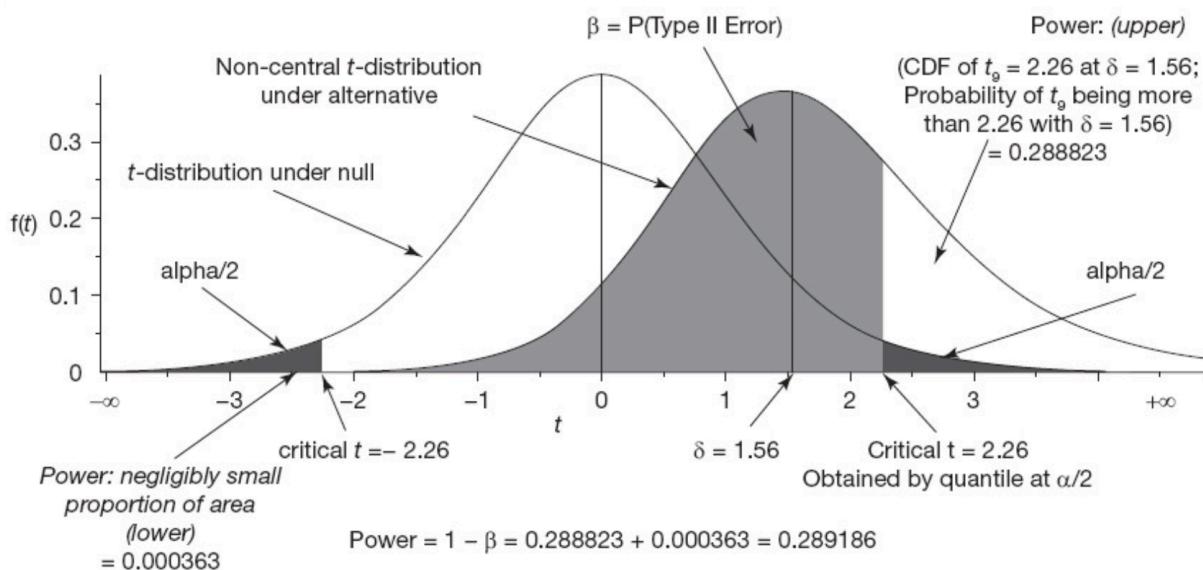
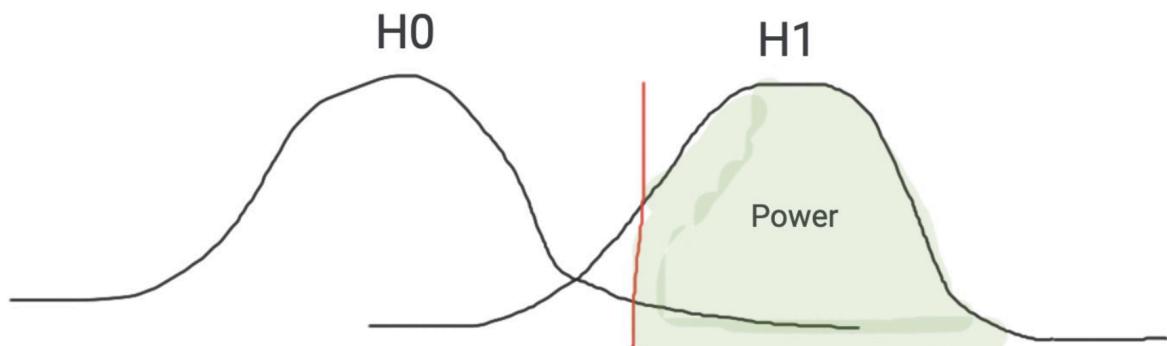
Im Zweifel für den Zweifel! → tocotronic

Tutorium - 03.November - 16:15Uhr

1. wann ist eine Theorie nach Popper bewiesen und etabliert?
 - nie - es gibt keine bewiesenen Theorien nach Popper,
 - die Annahme, dass immer eine Irrtumswahrscheinlichkeit übrig bleibt, nennt man Fallibilismus,
2. Was ist der Kern von Poppers Konzept für Wissenschaft und Gesellschaft?
 - Kritik ausüben,
 - Nur wenn wir uns kritisch mit allem auseinandersetzen, können wir den Anspruch haben, der Welt möglichst gerecht zu werden,
3. Was ist mit Fallibilismus gemeint?
 - erkenntnistheoretische Position, die besagt: es kann nie Gewissheit geben.
4. Warum sind Reduktionismus und Verifikationismus für die Generierung von Wissen problematisch?
 - Reduktionismus - alles ist erklärbar, wenn nur genügend Daten vorhanden sind, z.B. radikaler Behaviorismus - wir haben aber nie alle Daten - es reicht nicht, mit ein paar Daten zu arbeiten,
 - Verifikationismus - Streben nach Bestätigung - durch Sammeln von Positivbeispielen wird die Theorie nicht wahrscheinlicher - wozu forschen, wenn man sowieso nicht feilen kann, sondern immer nur weiter nach Beweisen sucht?
5. was ist laut Text ein Unterschied zwischen Marx und Einstein?
 - Einsteins Ideen waren und sind falsifizierbar,
 - Marx erklärt alle Möglichkeiten, daher sind seine Theorien nicht falsifizierbar,
 - Marx als Philosoph stellte Theorien auf, die nur verifizierbar sind und immer Bestätigung finden - sie können aber nicht falsifiziert werden,
 - erklärende Theorien können nicht durch Beobachtung falsifiziert werden und haben keinen empirischen Wert,
 - für Popper ein typisches Beispiel für Pseudo-Wissenschaft, die den wissenschaftlichen Erkenntnisgewinn verhindert,
 - Einstein als Physiker stellte die Relativitätstheorie auf - durch Beobachtung falsifizierbar

- er lässt Bedingungen zu, unter denen die Theorie nicht mehr haltbar ist, und nennt diese sogar in seinen eigenen Arbeiten,

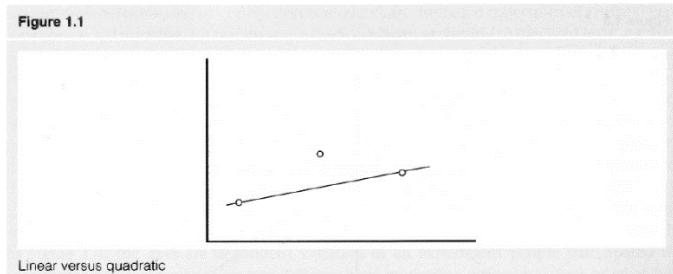
6. *jamovi* ✓
7. Import Datensatz aus erster Sitzung in das *Data Interface* bei *jamovi* ✓
8. Benennung der Variablen wie im google Dokument - size, room, R ✓
9. Wahl einer angemessenen Skala für jede Variable - *measure type* ✓
 - size - continuous
 - room - ordinal
 - R - nominal
10. Zeichnung einer Skizze von einem t-Test: eine Verteilung für die H₀, eine Verteilung für die H₁ und eine Linie, die das Signifikanzniveau markiert, Kennzeichnung der Power in der Zeichnung als Fläche



11. was ist ein Perzentil?
 - Lagemaß aus der Statistik,

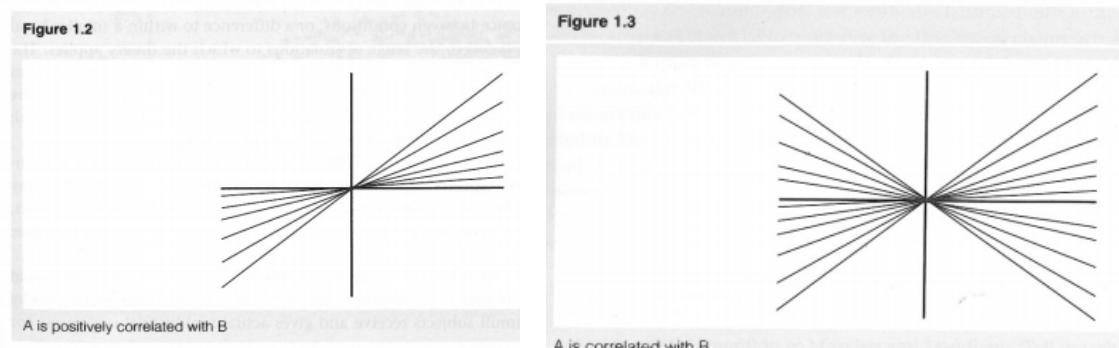
- durch die Perzentile wird ein der Größe nach geordneter Datensatz in 100 umfangsgleiche Teile zerlegt,
- für das 20% Perzentil bedeutet das z.B., dass 20% der Werte unterhalb oder gleich dieses Perzentils liegen,
- Perzentilenkurven zeigen die Streuung einer statistischen Verteilung an,
- die 50er Perzentile markiert den Durchschnittswert,

12.



- die Vorhersage eines linearen Zusammenhangs lässt sich bereits durch drei Datenpunkte falsifizieren, bspw. eine quadratische Beziehung wäre hier noch nicht ausgeschlossen,
- lineare Funktion hat mehr Möglichkeiten zur Falsifikation,
- quadratische Funktion passt für mehr Datenmuster,

13. welche der beiden Hypothesen wären hinsichtlich Popper vorzuziehen?



- in Figure 1.2 wird immerhin die Hälfte aller möglichen Korrelationen ausgeschlossen - somit ist der empirische Gehalt der Theorie höher als in Figure 1.3 und eine Falsifikation wahrscheinlicher,
- Figure 1.2 vorziehen, weil diese den Anspruch an gute Theorien, falsifizierbar zu sein, erfüllt, Figure 1.2 schließt alle negativen Möglichkeiten aus und bietet damit mehr Falsifikationsmöglichkeiten - sie schließt dennoch trotzdem wenig aus,
- Figure 1.3 sagt nichts über die Richtung der Korrelation aus, sodass ihr empirischer Gehalt gering ist, weil keine Möglichkeit ausgeschlossen wird - sie wäre eine schwache Theorie,

14. stelle eine Theorie auf und benenne, wie sie falsifiziert werden könnte!

- *alle Schwäne sind weiß*
 - falsifiziert, wenn mir ein Schwan gezeigt wird, der nicht weiß ist,
- *die meisten Schwäne sind weiß*

- dieser Theorie fehlt es an empirischem Gehalt, weil sie nicht definitiv ist - was heißt **die meisten**?

15. Warum ist falsifizieren besser als verifizieren?

- würde man einer Theorie mehr Glauben schenken, wenn jemand ein Beispiel dazu nennen kann oder wenn sie schon auf hundert verschiedene Weisen versucht wurde, zu widerlegen, aber bisher immer standgehalten hat?

16. stell Dir vor, Du bist EditorIn eines psychologischen Journals. ...was könnte Dich dazu bewegen, eher eine Studie mit signifikantem Effekt als eine mit nicht-signifikantem Effekt abzudrucken?

- am Kriterium der Signifikanz lässt sich entscheiden, ob das Ergebnis einer Stichprobe nur für die Stichprobe gilt oder ob es auf die Population verallgemeinert werden kann,

17. Was ist eine Replikation?

- Wiederholung einer Studie,
Ziel ist dabei die Kontrolle und Überprüfung der berichteten Forschungsergebnisse,
- man erreicht damit letztlich zwei Dinge:
 - die Hypothesen der Originalstudie haben die Chance sich im Popper'schen Sinne zu bewähren → so kann ihre Akzeptanz steigen,
 - außerdem diszipliniert eine Wissenschaft, in der Studien oft repliziert werden, Forschende dazu, sorgfältig zu arbeiten → Betrug und Schummelei können aufgedeckt werden,

18. Was bedeuten die Begriffe *true positive* und *false positive* bei der Interpretation von Signifikanztestergebnissen?

- *true positive*: ein Ergebnis, das auf die Existenz eines Effekts hinweist, wo tatsächlich ein Effekt existiert,
- *false positive*: ein Ergebnis, das auf die Existenz eines Effekts hinweist, obwohl kein Effekt existiert,

		Wirklichkeit	
		H_0 ist wahr	H_1 ist wahr
Entscheidung des Tests für H_0	Richtige Entscheidung (Spezifität) Wahrscheinlichkeit: $1 - \alpha$	Fehler 2. Art Wahrscheinlichkeit: β
	... für H_1	Fehler 1. Art Wahrscheinlichkeit: α	richtige Entscheidung Wahrscheinlichkeit: $1 - \beta$ (Trennschärfe des Tests)

19. Warum werden in RPP und SSRP viele verschiedene Studien repliziert und die Ergebnisse der Replikationen gemeinsam interpretiert?

- um einen Eindruck von der Replizierbarkeit der innerhalb von Forschungsgebieten publizierten Studien zu erhalten,
- von Interesse sind also Aussagen über das Forschungsgebiet oder den Publikationsprozess und nicht über die einzelne Theorie,

20. Was bedeuten die Ergebnisse des RPPs und des SSRPs für die psychologische Forschung?

- zu wenige Replikationen,
- falsche Publikationsanreize - neuartig, signifikant, hoher Effekt
- fragwürdige Wissenschaftskultur

21. Was zeichnet einen Replikationserfolg aus?

- hypothesenkonforme Ergebnisse,
 - Schätzungen von Effektstärken, die nah bei denen der Originalstudie liegen,
 - Schätzungen innerhalb eines Intervalls, siehe prediction intervals

22. Was kannst Du tun, um in eigenen Studien replizierbare Ergebnisse zu erzielen?

- Orientierung an den Standards der *open science* - preregistered, open data, open materials
 - Fallibilismus und Ansporn zur Falsifikation
 - Originalmaterialien zur Verfügung stellen,
 - Präregistrierung
<https://www.uni-erfurt.de/erfurtlab/forschen/informationen-fuer-forschende/prae-registrierung>

23. Was kann man aus diesen Tabellen ablesen?

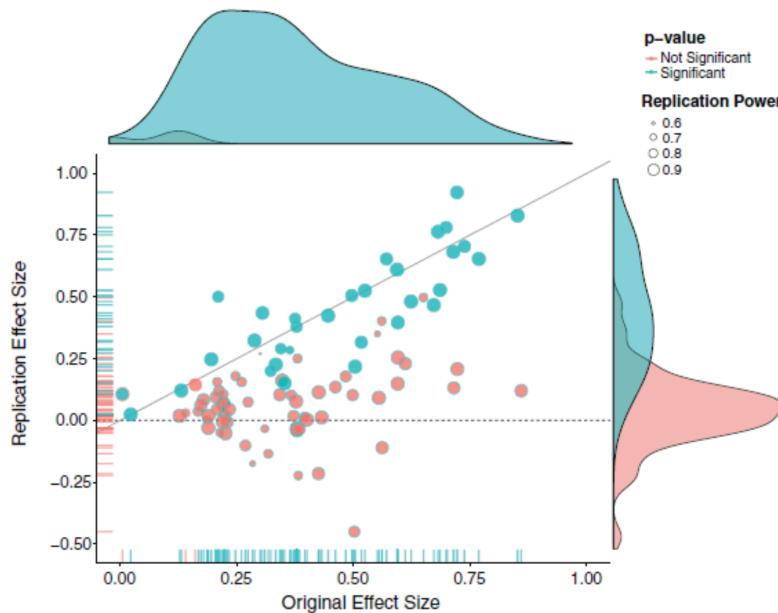
Rangkorrelationen von Merkmalen der
Originalstudien mit dem Signifikanzkriterium:

Merkmal	r
P-Wert	-.33
Effektstärke	.30
N	-.15
Relevanz des Effekts	-.11
Überraschungswert des Effekts	-.26
Erfahrung und Expertise des Teams	-.07

Rangkorrelationen von Merkmalen der
Replikationsstudien mit dem Signifikanzkriterium:

Merkmal	r
Power	.37
N	-.09
Schwierigkeit der Replikation	-.22
Erfahrung und Expertise des Teams	-.10
Qualität (Selbstbeurteilung)	-.07

24. interpretiere!

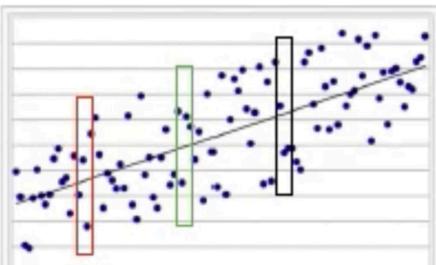


- wir sehen die Größe des ursprünglichen Studieneffekts im Vergleich zur Größe des Replikationseffekts,
- die diagonale Linie repräsentiert einen perfekten Original-Replikations-Ergebnis-Fit,
- die gepunktete Linie repräsentiert die Größe des Replikationseffekts von 0,
- Punkte unterhalb der gepunkteten Linie waren Effekte in der entgegengesetzten Richtung des Originals,
- Interpretation:
 - Ergebnisse verteilen sich nicht gleichmäßig um die Diagonale,
 - Replikationen finden systematisch deutlich kleineren Effekt als Originale und sind häufiger nicht signifikant,
 - die sich ergebenen Verteilungen sind so systematisch, dass man nicht mehr von Zufall ausgehen kann,
→ Publikationsverzerrung!

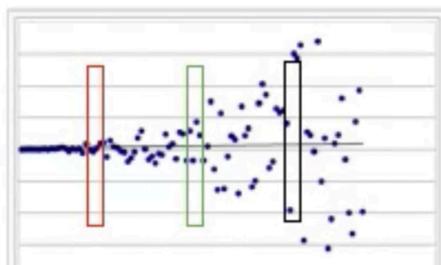
25. Was ist das Ziel einer Replikation? Worüber trifft eine Metaanalyse eine Aussage, wenn sie Studien zu einer spezifischen Theorie zusammenfassen? Worüber soll in Projekten wie RPP eine Aussage getroffen werden?

- direkte Replikation: können wir ähnliche Ergebnisse nochmal finden?
- konzeptuelle Replikation: lassen sich Studienergebnisse ausweiten?
- Metaanalyse: konnten wir die Ergebnisse in vielen Fällen finden?
- RPP: gibt es systematische Probleme im Prozess unseres Erkenntnisgewinns, die nicht am Inhalt der einzelnen Studie hängen?

26. Homoskedastizität vs. Heteroskedastizität



Homoskedastizität: Die Streuung der Punkte um die Gerade in vertikaler Richtung ist konstant.



Heteroskedastizität: Die Streuung der Punkte um die Gerade wächst nach rechts hin stärker als linear an.

die Varianz der Residuen ist in einer Regressionsanalyse für alle Werte des Prädiktors konstant, d.h., die Abweichungen der vorhergesagten Werte von den wahren Werten sind in etwa immer gleich groß - unabhängig wie hoch oder niedrig der Wert des Prädiktors ist, um eine Regressionsanalyse sinnvoll interpretieren zu können, ist es wichtig, dass Homoskedastizität vorliegt,

die Varianz der Residuen verändert sich mit ansteigenden oder abfallenden Werten des Prädiktors, bspw. kann man bei Heteroskedastizität mit einer Vorhersage systematisch umso weiter daneben liegen, je größer der Prädiktorwert ist, für den man ein Kriterium schätzen möchte,

der Signifikanztest in der Psychologie

→ das Problem:

- es liegen Daten aus einer Stichprobe vor, gesucht sind aber Aussagen über eine Population,

- aus den Daten in einer Stichprobe soll also auf die Population geschlossen werden,
- induktiver Schluss - ein solcher Schluss kann niemals sicher sein,
 - das Resultat ist eine Wahrscheinlichkeitsaussage,
 - es gibt keine unumstrittene, optimale Methode der Induktion, sondern zahlreiche Lösungsvorschläge,

→ populäre Lösungsvorschläge:

- Signifikanztest nach R. A. Fisher
- Signifikanztest nach J. Neyman und E. S. Pearson
- Bayes-Statistik

→ 2 essenzielle Zutaten in allen Lösungsvorschlägen

- Wahrscheinlichkeitskonzept
 - frequentistischer Wahrscheinlichkeitsbegriff
 - aleatorische/ontische/objektive/statistische Wahrscheinlichkeit
 - hierhin gehören Neyman, Pearson, Fisher (mit einigen Widersprüchen) und ein Großteil der modernen Statistik
 - subjektivistischer Wahrscheinlichkeitsbegriff
 - epistemische/personelle/subjektive/Bayessche Wahrscheinlichkeit,
 - hierhin gehören Bayes, deFinetti, Savage
- Stichproben(kennwerte)verteilungen

frequentistischer Wahrscheinlichkeitsbegriff

- Wahrscheinlichkeit → relative Häufigkeit eines Ereignisses in einer Referenzklasse,
- Beispiel: relative Häufigkeit von *Kopf* bei (unendlich) vielen Münzwürfen? → die Wahrscheinlichkeit von *Kopf* beträgt 0,5
- singulären Ereignissen kann keine Wahrscheinlichkeit zugeordnet werden!
 - Margarete liebt mich,
 - Frankreich gewinnt die WM 2022,
 - Linda ist eine Bankkauffrau,
 - die mittlere Aggressivität von Frustrierten ist größer als die Aggressivität von Nicht-Frustrierten,
 - die Hypothese ist korrekt,

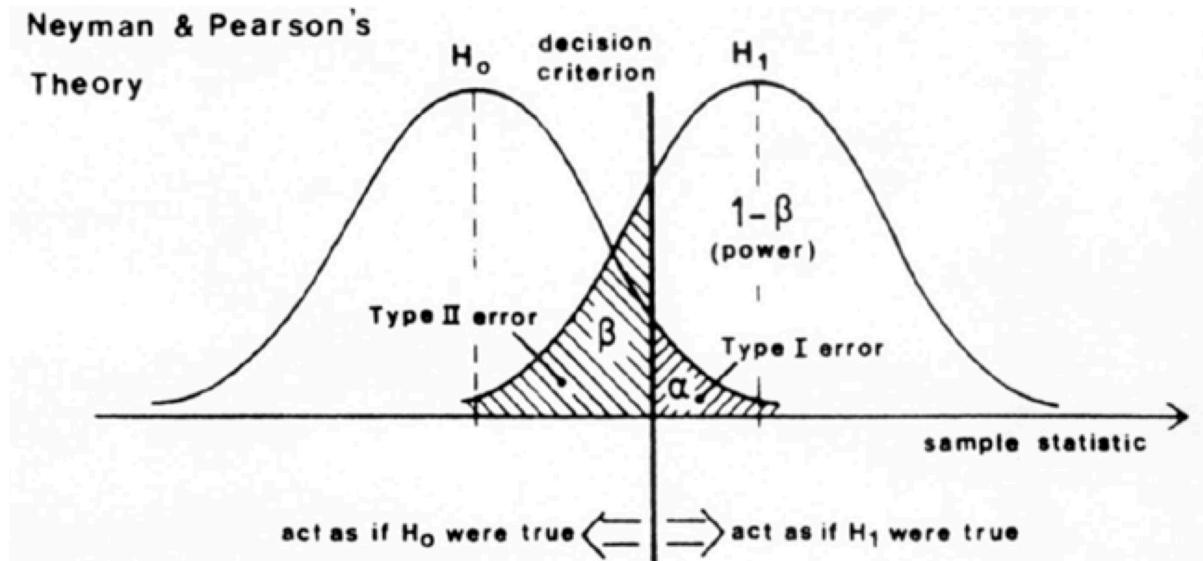
→ derartige Aussagen sind wahr oder falsch- der Begriff der Wahrscheinlichkeit kann auf sie nicht sinnvoll angewendet werden,

subjektivistischer Wahrscheinlichkeitsbegriff

- Wahrscheinlichkeit → Grad der Überzeugung (*degree of belief*)
- Messung durch die Akzeptanz von Wetten
 - sie können 100€ gewinnen,
 - in diesem Jahr liegt Weihnachten Schnee
 - wie viel sind Sie bereit einzusetzen? → 20€
 - die subjektive Wahrscheinlichkeit für Schnee zu Weihnachten beträgt $20/100 = 0,2!$
- sie sind in der Wahl Ihrer Überzeugungen frei - eine subjektive Wahrscheinlichkeit kann also nicht richtig oder falsch sein,
- allerdings können subjektive Wahrscheinlichkeiten inkohärent und damit irrational sein,
 - im obigen Beispiel müssten Sie bereit sein, 80€ darauf zu setzen, dass Weihnachten *kein* Schnee liegt - andernfalls sind Ihre subjektiven Wahrscheinlichkeiten inkohärent,

- die Wahrscheinlichkeitstheorie beschreibt, wie Wahrscheinlichkeiten kohärent verrechnet werden,

Stichprobenkennwerteverteilungen

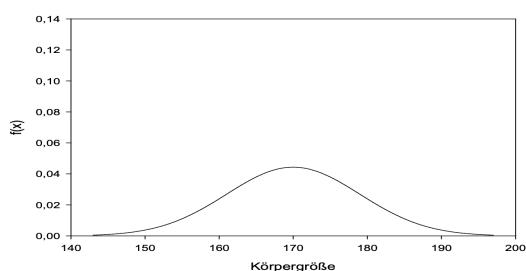


die Hauptschwierigkeit beim Verständnis des Signifikanztests: was sind das für Kurven?

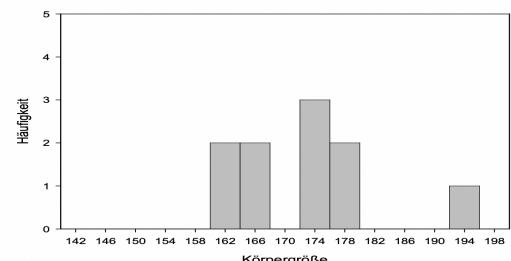
- Verteilung von Kennwerten aus (unendlich) vielen Stichproben,
- Kennwert: Maßzahl zur Beschreibung einer Häufigkeitsverteilung
 - Mittelwert, Median, Varianz, Anteil, Mittelwertsdifferenz...
- Wie kommt eine Stichprobenkennwerteverteilung zustande?
 - empirisch:
 - aus einer Population werden (unendlich) viele gleich große Stichproben gezogen,
 - in jeder dieser Stichproben wird ein Kennwert (z.B. der Mittelwert) berechnet,
 - diese (unendlich) vielen Kennwerte bilden eine neue Verteilung
⇒ die Stichprobenkennwerteverteilung!
 - Stichprobenkennwerteverteilungen können theoretisch hergeleitet werden!

Verteilungsarten in der Inferenzstatistik

Populationsverteilung
(oftmals eine Normalverteilung)



Häufigkeitsverteilung (der Daten in *einer* Stichprobe) - weicht zufällig von der Populationsverteilung ab!

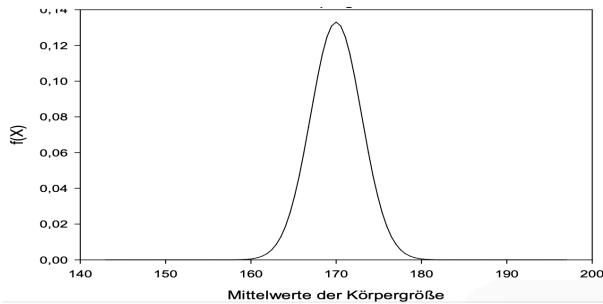


Stichprobenkennwerteverteilung (von Mittelwerten)

- Form: bei großem n stets normal
- Mittelwert \rightarrow Mittelwert der Populationsverteilung
- Varianz: kleiner als Varianz der Populationsverteilung

\rightarrow

$$\sigma_x^2 = \sigma^2 / n$$



wichtige Eigenschaften vom Stichprobenverteilungen

- die Form der Stichprobenverteilung hängt u.a. vom untersuchten Kennwert ab,
- zentraler Grenzwertsatz
 - mit steigender Stichprobengröße nähert sich *jede* gebräuchliche Stichprobenkennwerteveerteilung der Normalverteilung an
 - bei *kleinen* Stichproben ergeben sich andere Verteilungsformen (t -Verteilung, F -Verteilung, Chi-Quadrat-Verteilung...)
- empirisches Gesetz der großen Zahlen:
 - mit steigender Stichprobengröße nimmt die Varianz der Stichprobenkennwerteveerteilung ab, d.h. Schätzungen von Populationswerten werden tendenziell genauer,

Illustration des zentralen Grenzwertsatzes
 $P(\text{Ereignis}) = 0,1$

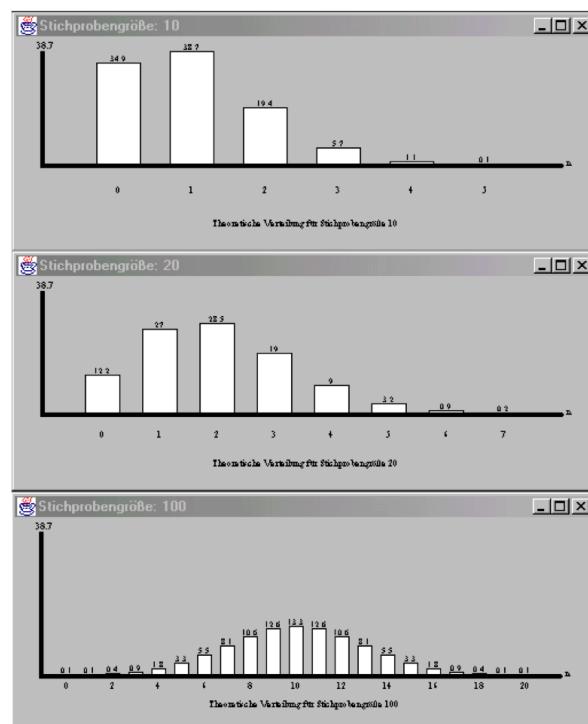
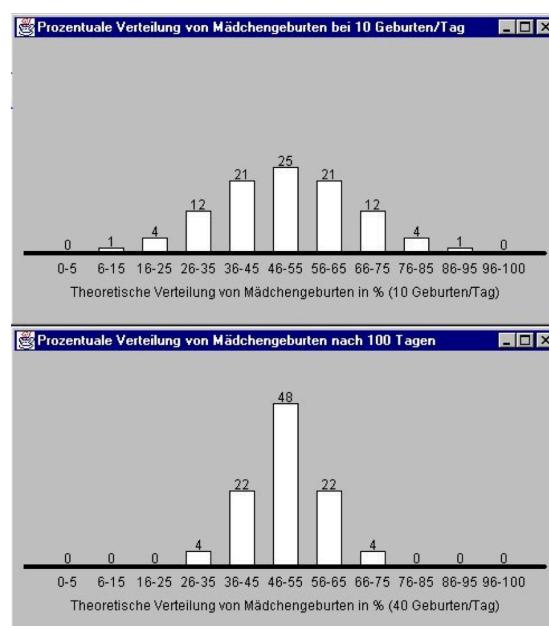


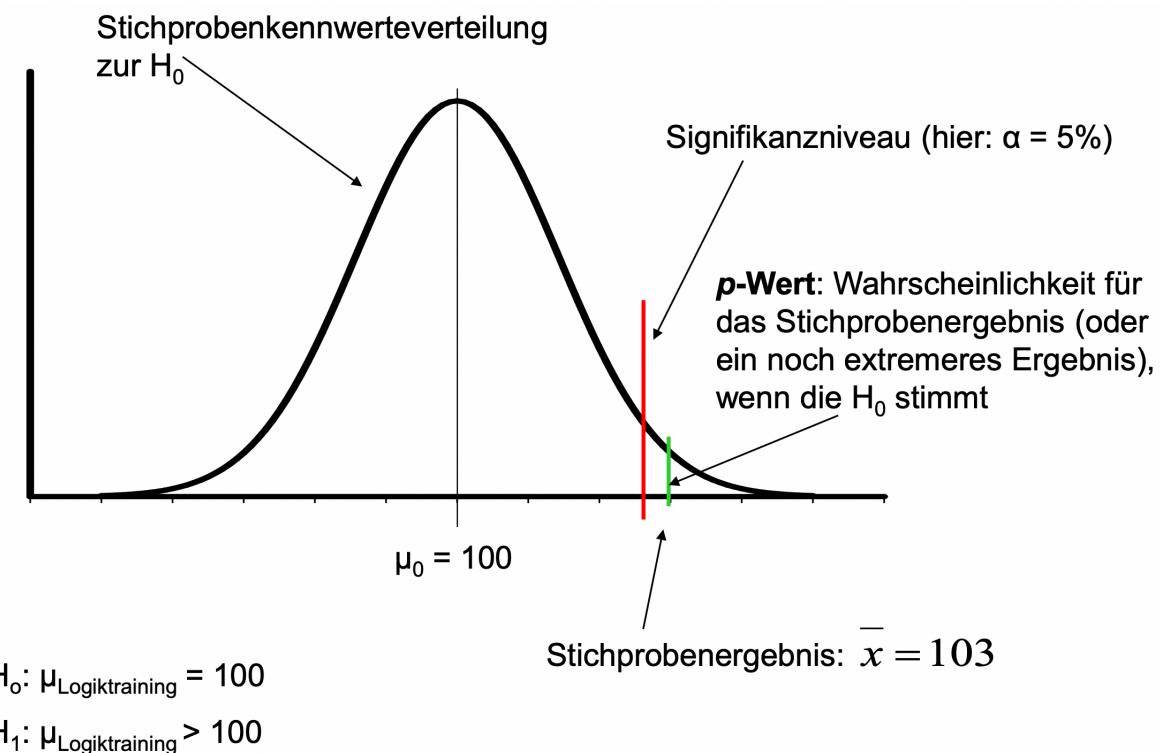
Illustration des empirischen Gesetzes der großen Zahlen
 $P(\text{Ereignis}) = 0,5$



der Signifikanztest nach R.A. Fisher

1. eine Nullhypothese festlegen, die nach Möglichkeit widerlegt werden soll
 - die H_0 besagt das Gegenteil der Alternativhypothese

- in der Psychologie lautet die H_0 nahezu immer, dass *kein* Effekt besteht - dies müsste nicht so sein!
 - Beispiel: die Populationsmittelwerte von Kontrollgruppe und Experimentalgruppe unterscheiden sich nicht,
2. Stichprobenverteilung zur H_0 bestimmen,
 3. Signifikanzniveau α festlegen,
 - Fisher schlägt $\alpha = 5\%$ und $\alpha = 1\%$ vor,
 4. p-Wert zu den Stichprobendaten berechnen,
 - $p\text{-Wert} = p(D \mid H_0)$ = Wahrscheinlichkeit des Stichprobenkennwerts (oder eines noch extremeren Werts), gegeben, dass die H_0 richtig ist,
 5. wenn p kleiner oder gleich α , ist das Ergebnis signifikant \rightarrow die H_0 wird zurückgewiesen
 \rightarrow andernfalls ist keine Schlussfolgerung möglich,



Was bedeutet ein signifikantes Ergebnis?

- unstrittig:
 - wenn die H_0 korrekt ist und α vor der Studie festgesetzt wurde, treten bei wiederholten Replikationen in α -Prozent aller Untersuchungen signifikante Ergebnisse auf,
 - α gibt damit die Wahrscheinlichkeit von Fehlentscheidungen zugunsten der H_1 an,
- Fisher interpretierte signifikante Ergebnisse epistemisch - das Wissen betreffend, erkenntnistheoretisch bedeutsam,
- ein signifikantes Ergebnis erlaubt demnach Aussagen über die Richtigkeit der Hypothese,
- es beeinflusst die Konfidenz, die wir in die H_0 haben sollten,
- dies ist (relativ) nah an einem subjektivistischen Wahrscheinlichkeitsbegriff und damit inkonsistent mit dem sonstigen Fisherschen Werk,
- aber:
- errechnet wurde $p(D \mid H_0)$ – die Wahrscheinlichkeit der Daten ergeben die H_0

- eine epistemische Interpretation impliziert eine Aussage über $p(H_0 | D)$ – die Wahrscheinlichkeit der Hypothese ergeben die Daten,
- zumeist: $p(A | B) \neq p(B | A)$
- Beispiel:
 - Wahrscheinlichkeit, dass eine Person schwanger ist, gegeben dass es sich um eine Frau handelt? → vielleicht (knapp) 2%
 - Wahrscheinlichkeit, dass eine Person eine Frau ist, gegeben dass die Person schwanger ist? → offensichtlich 100%
- es gibt keinen Weg (ohne Zusatzannahmen) aus $p(D | H_0)$ auf $p(H_0 | D)$ zu schließen,

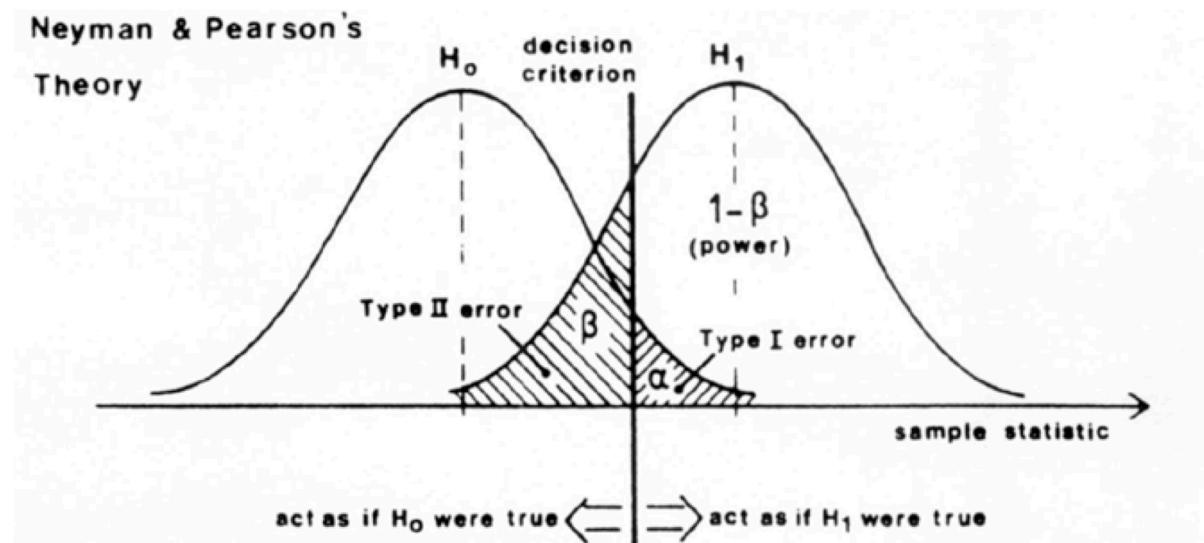
Was bedeutet ein nicht-signifikantes Ergebnis?

- früher Fisher:
 - nichts!
 - $p(D | H_0) = 0,06$ - offensichtlich kein starkes Argument für die H_0
- später Fisher:
 - ein nicht-signifikantes Ergebnis stärkt die Nullhypothese,
 - diese Aussage ist zumindest sehr erstaunlich - man könnte sie schlicht für falsch halten!

Wie und wann wird das Signifikanzniveau α festgelegt?

- früher Fisher:
 - das Signifikanzniveau wird vor der Studie per Konvention festgesetzt ($\alpha = 5\%$ und $\alpha = 1\%$) - das Signifikanzniveau ist damit eine Eigenschaft des Tests,
 - diese Auffassung gerät unter dem Einfluss von Neyman & Pearson unter Druck,
 - kritisch ist insbesondere, dass das Signifikanzniveau damit ohne Berücksichtigung der Fragestellung und der technischen Randbedingungen (vor allem der Stichprobengröße) festgesetzt wird,
- später Fisher:
 - das Signifikanzniveau ist identisch mit dem p-Wert - dieser wird kommuniziert,
 - die Bewertung dieser Information bleibt dem Leser/Rezipienten überlassen - das Signifikanzniveau ist damit eine Eigenschaft der Daten,
- 2 Bedeutungen des Begriffs Signifikanzniveau:
 - *standard level of significance*: eine Konvention
 - *exact level of significance*: der p-Wert

Der Signifikanztest nach Neyman & Pearson



1. lege eine Nullhypothese fest - wie gehabt
2. lege eine spezifische Alternativhypothese fest - formuliert als standardisierte Effektgröße oder in den ursprünglichen Kennwerten, z.B. Mittelwertsdifferenz,
3. lege akzeptable Größen für α und β und fest, ermitte eine entsprechende Stichprobengröße und bestimme die Stichprobenverteilungen → A priori Poweranalyse oder:
lege eine akzeptable Stichprobengröße fest, wäge die Wichtigkeit von α und β ab und bestimme die Stichprobenverteilungen,
4. berechne den p -Wert zu den Stichprobendaten,
5. wenn der p -Wert kleiner oder gleich α ist, ist das Ergebnis signifikant - die H_1 wird angenommen - andernfalls ist das Ergebnis nicht signifikant - die H_0 wird angenommen,

Was bedeutet ein signifikantes/nicht-signifikantes Ergebnis?

- der Signifikanztest ist nun symmetrisch → signifikante und nicht-signifikante Ergebnisse sind interpretierbar,
- α gibt die relative Häufigkeit von Fehlentscheidungen zugunsten der H_1 (bei wiederholten Replikationen) an,
- β gibt die relative Häufigkeit von Fehlentscheidungen zugunsten der H_0 an,
- keine epistemische Interpretation des Ergebnisses,
- das Ergebnis liefert also keinerlei Aussage über die Richtigkeit der Hypothese(n),
- stattdessen: Verhaltensinterpretation
 - bei signifikantem Ergebnis: verhalte Dich, als ob die H_1 wahr wäre,
 - bei nicht-signifikantem Ergebnis: verhalte Dich, als ob die H_0 wahr wäre,
- Verhaltensinterpretation
ein Beispiel aus der Qualitätskontrolle:
 - in der Produktion einer Firma befinden sich regulär 10% Ausschuss → Nullhypothese,
 - wenn der Ausschuss über 15% liegt, wird die Produktion gestoppt → Alternativhypothese,
 - täglich wird eine Stichprobe gezogen und per Signifikanztest über die weitere Produktion entschieden,
 - da ein zu großer Ausschuss deutlich kostspieliger ist als ein Produktionsstopp wird $\beta = 1\%$ und $\alpha = 10\%$ festgesetzt,
 - was tun Sie bei einem signifikanten Ergebnis?
 - sie stoppen die Produktion - dies ist vernünftig, da Sie so im *long run* Geld sparen,
 - sie werden dennoch nicht unbedingt glauben, dass die Alternativhypothese korrekt ist - dies ist (mehr oder weniger) häufig **nicht** der Fall!

Wie und wann wird das Signifikanzniveau α festgelegt?

- **vor dem Test!** - nur so können α und β als relative Häufigkeiten von Fehlentscheidungen interpretiert werden,
- die Festlegung erfolgt aufgrund einer Abwägung der relativen Wichtigkeit von α und β ,
- Beispiel:
 - Signifikanztest zum Effekt eines neuen Medikaments,
 - das Medikament soll eine bisher nicht behandelbare Krankheit heilen - Nebenwirkungen sind nicht zu erwarten,
 - kleines β (und relativ großes α)
 - das Medikament soll bei Schnupfen helfen - es sind starke Nebenwirkungen zu erwarten,

- kleines α (und relativ großes β)

Kritik am Neyman-Pearson-Ansatz

- Fisher hat die Vorschläge von Neyman und Pearson mit verständlichen und unverständlichen Argumenten massiv kritisiert,
- ein aus meiner Sicht berechtigter Kritikpunkt bei grundlagen- wissenschaftlichen Anwendungen:
 - das Erkenntnisinteresse ist hier tatsächlich epistemischer Natur,
 - eine Verhaltensinterpretation von Testergebnissen ist zuweilen wenig sinnvoll, da mit dem Test keine Entscheidung über ein Verhalten intendiert ist,
 - dies macht eine Abwägung von α und β schwierig,

welche Faktoren beeinflussen das Ergebnis eines Signifikanztests?

- Effekt in der Population → je größer, desto eher signifikant
- Abwägung von α und β
 - α : je größer, desto eher signifikant
 - β : je kleiner, desto eher signifikant
 - (α und β sind komplementär)
- Stichprobengröße → je größer, desto eher signifikant

<https://rpsychologist.com/d3/nhst/>

was in der Psychologie daraus wurde - der Hybrid

- das Über-Ich - oder wie der Signifikanztest durchgeführt werden sollte
 - als *normatives Ideal* gilt der Neyman-Pearson-Ansatz,
- das Ich - oder wie der Signifikanztest tatsächlich eingesetzt wird
 - die tatsächliche Nutzung folgt hauptsächlich dem Fisher-Ansatz,
 - keine (spezifische) Alternativhypothese,
 - keine Poweranalyse - was sich anscheinend in jüngster Zeit ändert,
 - das Signifikanzniveau wird post hoc durch Aufrunden des p-Werts bestimmt - dies ist in jedem Ansatz Unsinn!
 - epistemische Aussagen über die Bedeutung des Resultats,
- das Es - oder wie der Signifikanztest oftmals interpretiert wird
 - häufig bayesianische Interpretation
 - das Ergebnis des Signifikanztests wird als Wahrscheinlichkeit der Nullhypothese aufgefasst - dies ist eindeutig falsch!

wie es vielleicht gehen könnte - der optimierte Hybrid

1. lege eine Nullhypothese fest
2. lege eine spezifische Alternativhypothese fest
3. lege akzeptable Größen für α und β und fest, ermitte eine entsprechende Stichprobengröße und bestimme die Stichprobenverteilungen → a priori Poweranalyse dabei wird man sich zumeist auf Konventionen stützen müssen!
4. berechne den p -Wert zu den Stichprobendaten,
5. wenn der p -Wert kleiner oder gleich α ist, ist das Ergebnis signifikant - die H_1 wird angenommen, andernfalls ist das Ergebnis nicht signifikant - die H_0 wird angenommen,

verschiedene inferenzstatistische Methoden

- Signifikanztest nach R. A. Fisher
 - o kein β -Fehler und keine Power,
 - o Stichprobenverteilung bezieht sich auf *hypothetische Population*,
 - o Signifikanzniveau aufgrund von Konventionen,

- epistemische Interpretation des Ergebnisses,
 - keine Alternativhypothese,
- Signifikanztest nach J. Neyman und E. S. Pearson
- spezifische Alternativhypothese,
 - β -Fehler und Power,
 - Signifikanzniveau aufgrund einer Abwägung von α - und β -Fehler,
 - Verhaltensinterpretation des Ergebnisses,
- Bayes-Statistik
- Berechnung der Wahrscheinlichkeit von Hypothesen

17.November - Tutorium

1. zeichne schematisch einen t-Test, der eine Power von 50% hat!



2. wie lautet der empirische t-Wert? wie lautet der kritische t-Wert? war das Ergebnis signifikant? welche Aussage kannst du aufgrund des Ergebnisses treffen?
 - du wertest eine Medikamentenstudie an Betroffenen zu krankheitserregenden Bakterien aus - mit den TeilnehmerInnen der Gruppe A wird eine Phagentherapie durchgeführt - die TeilnehmerInnen der Gruppe B bekommen ein Placebo verabreicht - anschließend werden Proben genommen und die Menge der Bakterien bestimmt - jeder TN erhält einen Wert zwischen 0 (keine Bakterien) und 7 (viele Bakterien) - führe einen t-Test für unabhängige Gruppen durch ($\alpha = 0,05$),
 - die Werte der Personen:
 - Gruppe A: 2, 4, 1, 1, 3, 1, 2
 - Gruppe B: 6, 3, 6, 4, 6

$$df = n_A + n_B - 2$$

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{\hat{\sigma}_{\bar{x}_A - \bar{x}_B}}$$

$$\hat{\sigma}_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{(n_A - 1) \cdot \hat{\sigma}_A^2 + (n_B - 1) \cdot \hat{\sigma}_B^2}{(n_A - 1) + (n_B - 1)} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}$$

Varianz in der Stichprobe

$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Schätzung der Populationsvarianz

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$

- $t_{krit} = 1,813$ ($df = 10$, alpha = 5%)
 - $t_{emp} = 4,050463$
 - $t_{krit} < t_{emp} \rightarrow$ signifikant
 - aufgrund des signifikanten Ergebnisses könnte man sich so verhalten, als wäre die in der Stichprobe beobachtete Wirkung des Medikaments auf die Population verallgemeinerbar,
 - da es sich hierbei um eine extrem kleine Stichprobe einer einzelnen Studie handelt, wäre ein induktiver Schluss verfrüht,
3. was besagt das Gesetz der großen Zahlen?
- ! die Streuung (Varianz) einer Stichprobenkennwerteverteilung nimmt mit steigender Stichprobengröße ab, d.h. Schätzungen von Populationswerten werden tendenziell genauer,



4. was besagt der zentrale Grenzwertsatz?

- ! die SKV nähert sich mit steigender Stichprobengröße der Form einer Standardnormalverteilung an,
- ! haben bei kleinen Stichproben einzelne, vom wahren Wert abweichende, Werte noch eine große Bedeutung und Einfluss auf z.B. den Mittelwert, so verlieren sie in größeren Stichproben an Relevanz,

5. fülle die Tabelle!

	früher Fisher	später Fisher	Neyman&Pearson
welcher Wahrscheinlichkeitsbegriff wird verwendet?	frequentistische gemischt mit subjektivistischer Interpretation	subjektivistische Interpretation bekommt mehr Gewicht	frequentistische Interpretation
wie wird das Signifikanzniveau bestimmt und berichtet?	per Konvention - es ist eine Eigenschaft des Tests	identisch mit dem p-Wert - eine Eigenschaft der Daten und wird exakt berichtet	Festlegung findet vor dem Test statt - wichtig ist die Abwägung der relativen Wichtigkeit von alpha und beta
wie wird ein signifikantes Ergebnis interpretiert?	epistemische Interpretation: wir lernen etwas über die Hypothese im spezifischen Experiment - das Signifikanzniveau berichtet, wie stark unsere Überzeugung durch das Ergebnis geändert wird	epistemische Interpretation: wir lernen etwas über die Hypothese im spezifischen Experiment - das Signifikanzniveau berichtet, wie stark unsere Überzeugung durch das Ergebnis geändert wird	verhalte Dich, als ob die H_1 wahr wäre, das Ergebnis lässt keine Aussage über die Richtigkeit der Hypothese zu, in Forschung mit oft wiederholten Experimenten wird eine festgelegte Fehlerrate akzeptiert - alpha
wie wird ein nicht-signifikantes Ergebnis interpretiert?	nicht interpretierbar	bestärkt die H_0 , bestätigt sie aber nicht	verhalte Dich, als ob die H_0 wahr wäre, das Ergebnis lässt keine Aussage über die Richtigkeit der Hypothese zu,

6. Warum bezeichnet Gigerenzer das Signifikanztestkonzept von Neyman & Pearson als *Über-ich*?

- ! wegen der konsequenten Beschränkung auf ein einzelnes frequentistisches Wahrscheinlichkeitskonzept, die sich auch in der nicht epistemischen Interpretation des Testergebnisses wiederfindet,

7. Warum ist die Varianz der SKV, bspw. vom Mittelwert kleiner als die Varianz der Populationsverteilung?

- ! in die Populationsverteilung gehen alle Werte ein, d.h. auch Ausreißer- und Extremwerte,
- ! die SKV umfasst hingegen keine einzelnen Werte, sondern Kennwerte, die über die Stichproben gemittelt werden - aus diesem Grund fällt ein Großteil der Varianz weg,
- ! Standardabweichung des Mittelwertes:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} = \frac{s}{\sqrt{n-1}}$$

8. begründe folgende Aussage einmal mit einem frequentistischen und einmal mit einem subjektivistischen Wahrscheinlichkeitskonzept: *morgen scheint zu 75% die Sonne!*

- ! frequentistisch:
 - Wahrscheinlichkeit nach dem frequentistischen Ansatz beschreibt die Häufigkeit eines Ereignisses in einer Referenzklasse - die Entscheidung,

ob morgen die Sonne scheint, ist jedoch ein singuläres Ereignis (dichotom mögliches Ereignis, singulären Ereignissen kann keine WK zugeordnet werden),

- aus diesem Grund können wir keine Wahrscheinlichkeitsaussage machen,
- eine solche wäre nur in Bezug zu einer Referenzklasse möglich,

! subjektivistisch:

- mit allen mir bisher vorliegenden Informationen komme ich zu dem Schluss, dass es 3x so wahrscheinlich ist, dass die Sonne scheint, gegenüber der Möglichkeit, dass sie nicht scheint,
- die 75% drückt meine Überzeugung aus, dass dieses Ereignis morgen eintritt,
- die 25% ist meine verbleibende Unsicherheit, dass dieses Ereignis doch nicht eintreten wird,

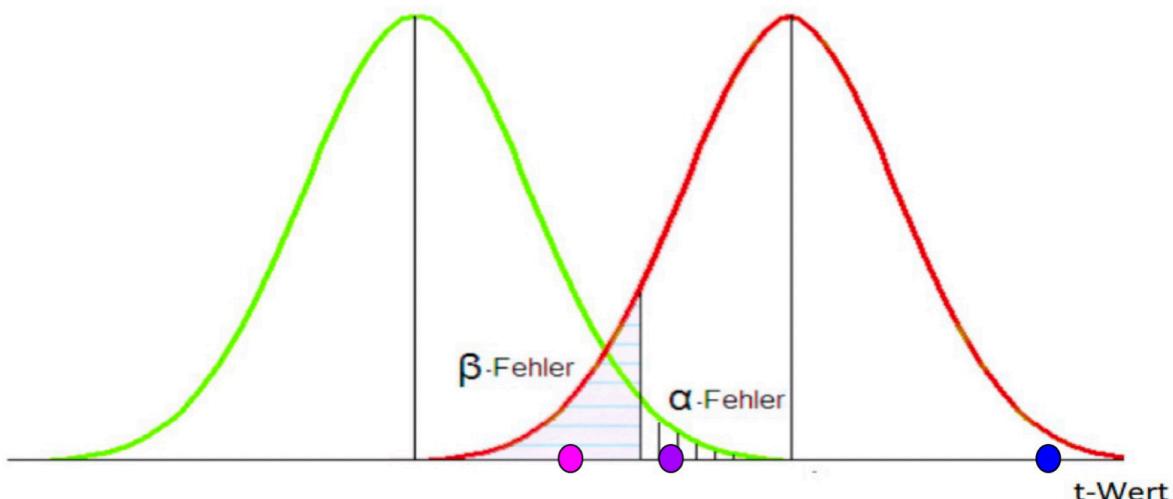
9. Was ist der Bayesian Cake und was ist laut Gigerenzer Fishers Beziehung zu diesem?

- die Interpretation des Signifikanztestergebnisses als $p(H_0 | D)$ anstelle von $p(D | H_0)$
- *he wanted to eat and reject it at the same time* - er wollte essen und es gleichzeitig ablehnen

10. was kann über die 3 Punkte aus Perspektive der Verteilungen gesagt werden?

Signifikanzniveau bei 5%

was sind deine Möglichkeiten, um möglichst sicher zu gehen, dass man einen besonders kleinen Effekt in einer empirischen Untersuchung nachweisen kann?



11. gegeben ist $p(D | H_0) = 70\%$ - formuliere diese Aussage aus -

wie wahrscheinlich ist die H_0 ?

- die Wahrscheinlichkeit für die Daten beträgt 70% unter der Annahme, dass die H_0 wahr ist,
- die H_0 setzen wir hier als gegeben voraus und deshalb hat sie streng genommen keine Wahrscheinlichkeit,

12. welche statistische Größe sollte laut Cohen im Bericht über eine Studie anstelle des Signifikanzkriteriums stärker in den Vordergrund gestellt werden?

- die Effektgröße - im besten Fall mit dem zugehörigen Konfidenzintervall,

13. warum kritisiert Cohen den interpretatorischen Fokus auf die Korrelation und fordert ein, Regressionskoeffizienten stärker zu fokussieren?

- Korrelationen sind standardisierte Effektgrößen und damit unabhängig von der gemessenen Einheit,
- durch die Überbetonung von standardisierten Effektgrößen fehlt aus Perspektive Cohens eine stärkere Auseinandersetzung mit den gemessenen Einheiten,

14. theoretisch kann man Signifikanztests auch zur Falsifikation von Forschungshypothesen verwenden - inwiefern unterscheidet sich dieses Vorgehen von der (bspw. in der Psychologie) typischen Vorgehensweise?

- das Forschungsinteresse läge auf der H_0 , wodurch man versuchen würde, die eigene Hypothese zu widerlegen,
- ein nicht signifikantes Ergebnis wäre in diesem Fall für die Bewährung der Theorie erstrebenswert,

15. welchen Vorteil bringen Präregistrierungen oder registered reports für das Signifikanztesten?

- es ist möglich das Signifikanzniveau vorher festzulegen, so dass der p- Wert eine eindeutige Interpretation bekommt und der Prozess transparenter wird,

16. fasste in eigenen Worten mindestens 3 Probleme des bisher behandelten Signifikanztestens (Fisher, Neyman & Pearson) zusammen.

- die Untersuchung entspricht nicht unserem Forschungsinteresse, wir betrachten $p(D | H_0)$, nicht $p(H_0 | D)$,
- die H_0 ist häufig unmöglich: sie wird aufgrund marginaler Unterschiede (bspw. durch Messfehler) falsch sein - vor allem mit steigender Power,
- die verschiedenen Konzeptionen werden häufig durcheinandergeworfen und fehlinterpretiert,

17. das *Journal of Experimental Psychology* machte 1962 die Publikationswahrscheinlichkeit eines Papers vom gefundenen p- Wert abhängig - somit wurden Ergebnisse mit $p < 0.05$ seltener publiziert als Artikel mit $p < 0.01$ - wie kann man dieses Vorgehen bewerten?

- es gibt viele Gründe dieses Vorgehen zu kritisieren - grundsätzlich wäre es sinnvoll, alle methodisch akzeptablen Studien zu veröffentlichen, um ein adäquates empirisches Bild zu erhalten,
- außerdem ist der p-Wert der einzelnen Studie stark vom Zufall abhängig,
- gleichzeitig bietet das Vorgehen Gründe für p-Hacking, HARKing, etc.,...

18. was ist der Unterschied zwischen essentieller und nicht-essentieller Multikolinearität?

- essentiell: nahezu lineare Beziehung zw mind. 2UV ohne den Achsenabschnitt,
- nicht-essentiell: nahezu lineare Beziehung zw. dem Achsenabschnitt und mind. einer der verbliebenen UV,

Effektgrößen

wozu?

- der typische Signifikanztest stellt (und beantwortet) die Frage *gibt es einen Effekt?*
 - diese Frage ist nahezu immer unzureichend und impliziert, dass kein Vorwissen existiert,
 - die relevante Frage lautet nahezu immer *wie groß ist der Effekt?*
 - wirkt die Heizung?
 - wie groß ist der Effekt von Psychotherapie?
 - wie groß ist der Effekt einer Kopplung von UCS und CS auf den Speichelfluss bzw. die Assoziationsstärke?
 - eine Möglichkeit Effektgrößen anzugeben, besteht in der Verwendung des Mittelwertsunterschieds,
 - *der durchschnittliche Jahresverdienst von Akademikern liegt 10000 Euro über dem Jahresverdienst von Nicht-Akademikern*
 - diese Möglichkeit ist in der Psychologie zumeist unzureichend!
 - Maßeinheiten sind in psychologischen Untersuchungen häufig willkürlich gewählt und zunächst bedeutungslos,
 - *das durchschnittliche Befinden von therapierten Depressiven war um 3 Punkte besser als das durchschnittliche Befinden von unbehandelten Depressiven*
 - Mittelwertsunterschiede aus verschiedenen Untersuchungen sind idR nicht vergleichbar,
 - wie kann der Effekt einer Pause auf das Vokabellernen mit dem Effekt einer Pause auf das Erlernen eines Gedichts verglichen werden?
- Lösung: standardisierte Effektgrößen**

Klassen von standardisierten Effektgrößen

- Abstandsmaße (standardisierter Mittelwertsunterschied) → d, g
- Zusammenhangsmaße → r, φ
- Maße der Varianzaufklärung → r^2, η^2

Beispieleperiment:

Fragestellung:

wie wirkt das Einlegen von Pausen während des Lernens auf den Lernerfolg?

UV: lernen mit/ ohne Pause

Kontrollgruppe:

eine Stunde Vokabeln lernen ohne Pause

Experimentalgruppe:

eine Stunde Vokabeln lernen mit 10minütiger Pause

AV: Anzahl gelöster Aufgaben im Vokabeltest

- wie groß ist der Effekt der Pause?

Ergebnisse:

ohne Pause	mit Pause
18	22
16	24
20	17
14	25
23	23
	26

Mittelwerte:

$$x_{op} = 18,2 \quad x_{mp} = 22,83$$

die Effektgrößen d und g

- der Mittelwertsunterschied zw. zwei Gruppen wird an der Standardabweichung relativiert,
- d gibt die Effektgröße in einer Stichprobe oder Population an,
- g entspricht der bestmöglichen Schätzung des Populationseffekts aufgrund von Stichprobendaten,

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Standardabweichung in der Stichprobe

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Standardabweichung in der Population

$$\hat{\sigma} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Schätzung der Standardabweichung in der Population

die Effektgröße d

$$d = \frac{\mu_A - \mu_B}{\sigma_{AB}}$$

σ_{AB} : gepoolte Standardabweichung

$$\text{Falls } n_A = n_B: \sigma_{AB} = \sqrt{\frac{s_A^2 + s_B^2}{2}} \quad \text{Falls } n_A \neq n_B: \sigma_{AB} = \sqrt{\frac{n_A \cdot s_A^2 + n_B \cdot s_B^2}{n_A + n_B}}$$

Im Beispiel: $s_{mP}^2 = 8,47$ $s_{oP}^2 = 9,76$

$$d = \frac{22,83 - 18,2}{\sqrt{\frac{6 \cdot 8,47 + 5 \cdot 9,76}{5 + 6}}} = \frac{22,83 - 18,2}{3,01} = 1,54$$

die Effektgröße g

$$g = \frac{\mu_A - \mu_B}{\hat{\sigma}_{AB}}$$

Falls $n_A = n_B$:

$$\sigma_{AB} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2}}$$

Falls $n_A \neq n_B$:

$$\sigma_{AB} = \sqrt{\frac{(n_A - 1) \cdot \hat{\sigma}_A^2 + (n_B - 1) \cdot \hat{\sigma}_B^2}{N - 2}}$$

Im Beispiel: $\hat{\sigma}_{mP}^2 = 10,164$ $\hat{\sigma}_{oP}^2 = 12,2$

$$g = \frac{22,83 - 18,2}{\sqrt{\frac{5 \cdot 10,164 + 4 \cdot 12,2}{11 - 2}}} = \frac{22,83 - 18,2}{3,33} = 1,39$$

Definitionsprobleme

- Rosenthals g ist identisch mit Borensteins d ,
- Rosenthals d kommt bei Borenstein nicht vor,
- auch Rosenthals g ist KEIN unverzerrter Schätzer des Populationseffekts,
- ein unverzerrter Schätzer ergibt sich durch eine Korrektur,

Korrekturfaktor:

$$J = 1 - \frac{3}{4df - 1}$$

→ **Borensteins $g = Rosenthals \ g * J$**

Varianz der Effektstärken

- würden wir eine Studie beliebig oft wiederholen und jeweils eine Effektstärke berechnen, so ergäbe sich die Stichprobenverteilung der Effektstärke,
- Varianz von Borensteins d :

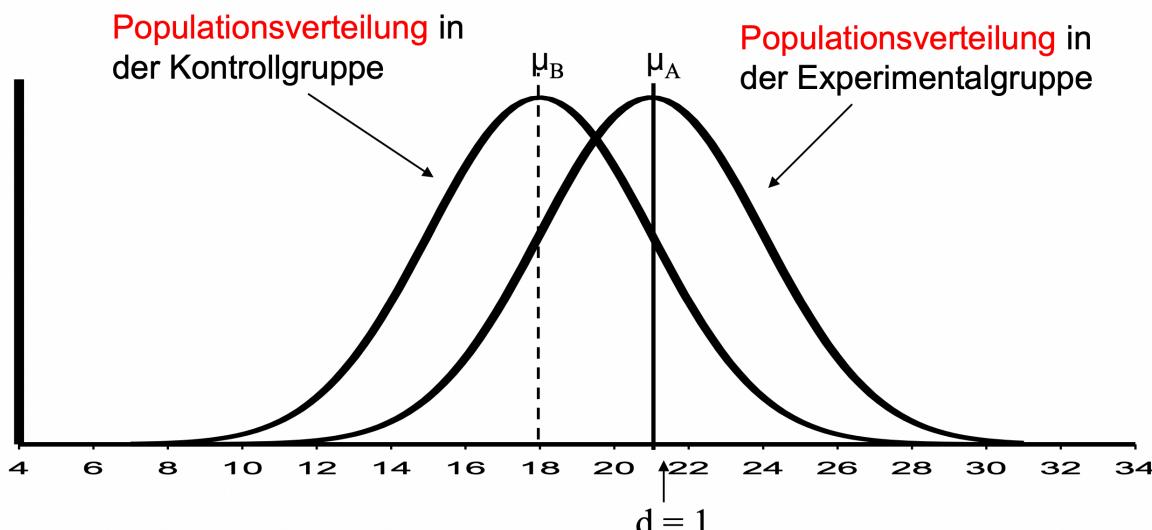
$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

- Varianz von Borensteins g :

$$V_g = J^2 \times V_d$$

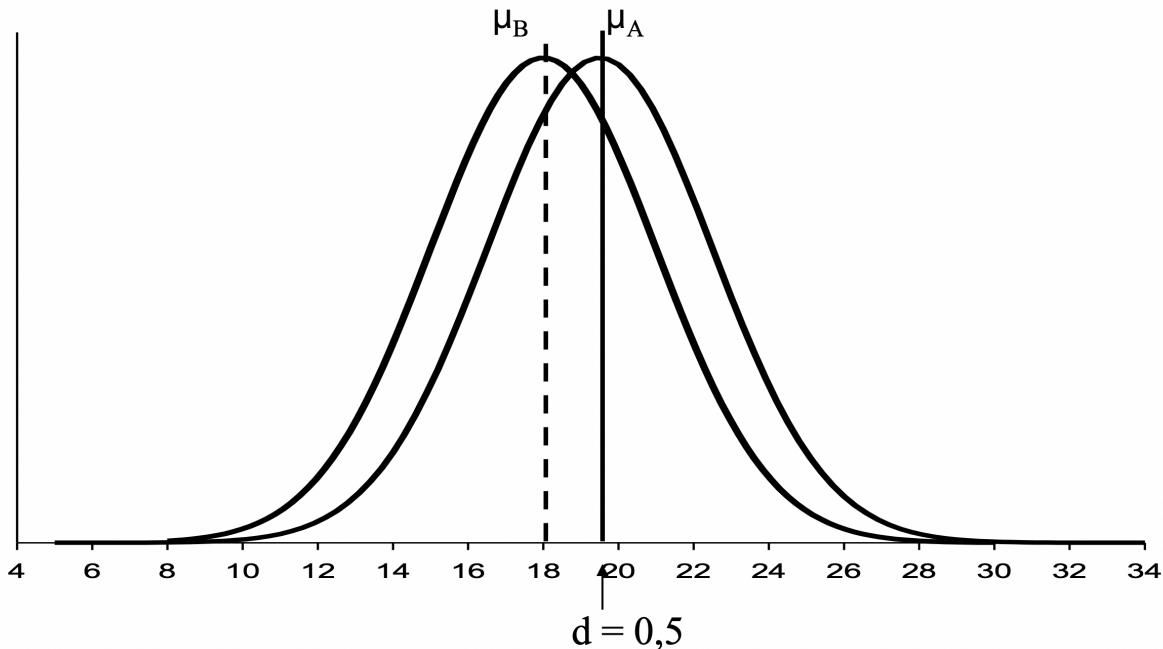
Was bedeutet die Effektgröße d inhaltlich?

- Situation bei $d = 1$ ($\mu_A = 21$; $\mu_B = 18$; $\sigma_{AB} = 3$):



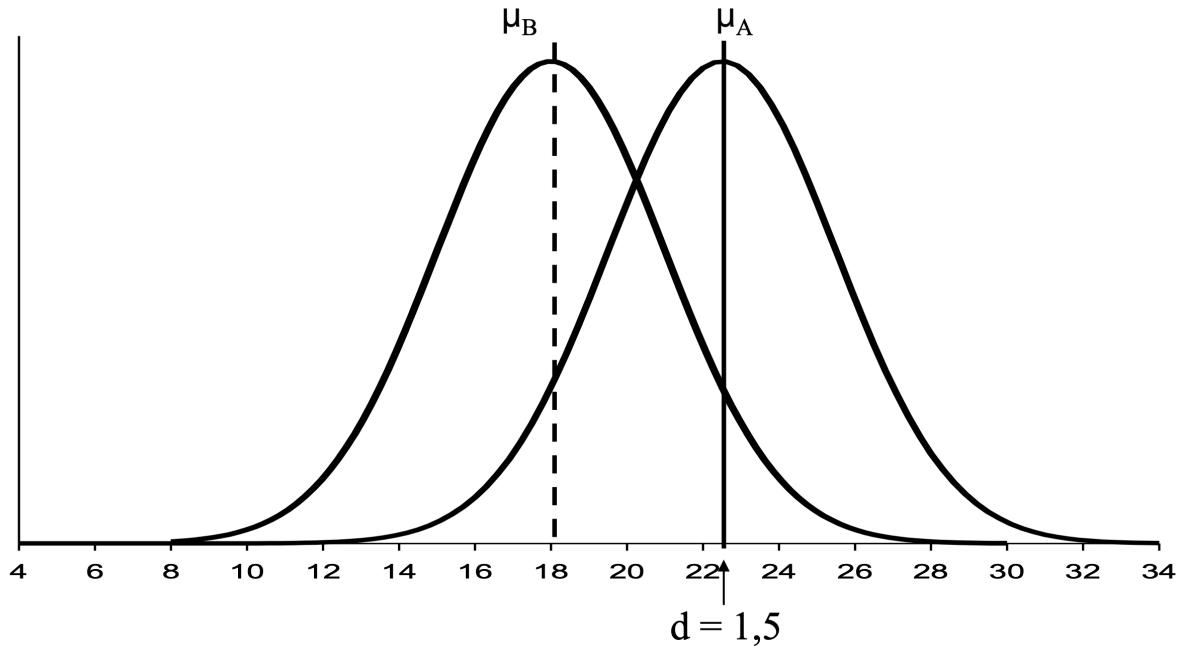
- ! bei $d = 1$ entspricht ein durchschnittlicher Wert in der Experimentalgruppe einem Prozentrang von 84 in der Kontrollgruppe - wenn das Merkmal normalverteilt ist!
- ! generell gibt d an, wie stark die Populationsverteilungen aus Experimental- und Kontrollgruppe überlappen!

→ Situation bei $d = 0,5$ ($\mu_A = 19,5$; $\mu_B = 18$; $\sigma_{AB} = 3$):



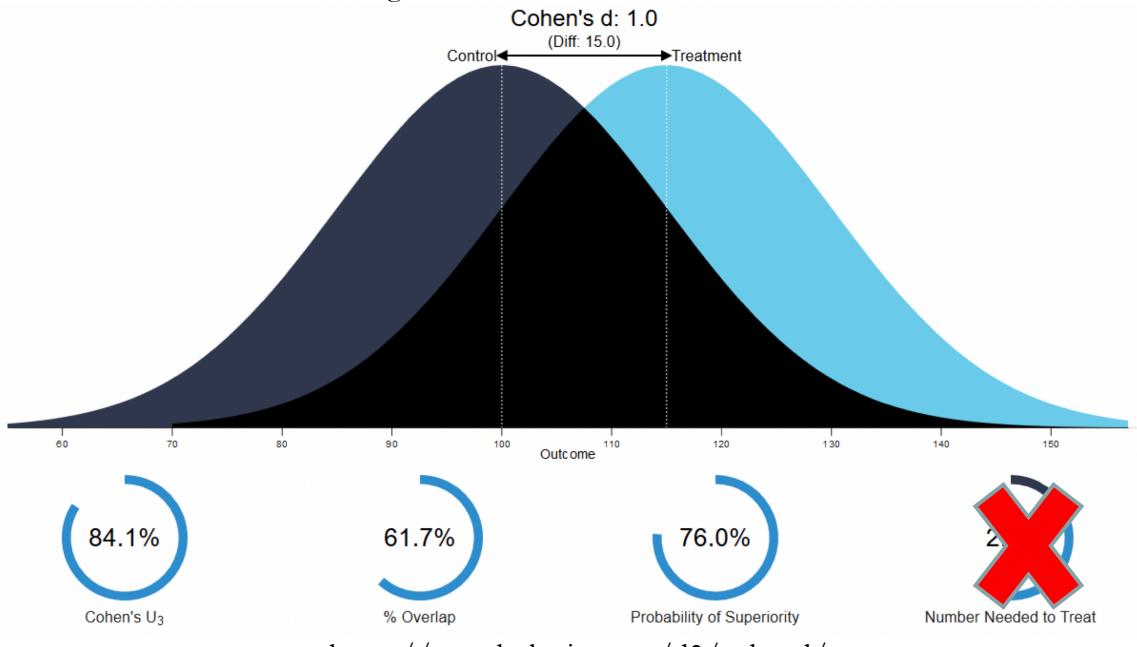
- ! bei $d = 0,5$ entspricht ein durchschnittlicher Wert in der Experimentalgruppe einem Prozentrang von 69 in der Kontrollgruppe - wenn das Merkmal normalverteilt ist!

→ Situation bei $d = 1,5$ ($\mu_A = 22,5$; $\mu_B = 18$; $\sigma_{AB} = 3$):



- ! bei $d = 1,5$ entspricht ein durchschnittlicher Wert in der Experimentalgruppe einem Prozentrang von 93 in der Kontrollgruppe - wenn das Merkmal normalverteilt ist!

mehr Illustrationen der Bedeutung von d



- wieviel Prozent der Experimentalgruppe (eigentlich: der entsprechenden Population) liegen oberhalb des Mittelwerts in der Kontrollgruppe?
 - Cohen's U₃ = $\Phi(d)$

- Überlappung → 2Definitionen
 - wieviel Prozent der Mitglieder beider Populationen liegen innerhalb der Schnittmenge?

$$OVL = 2\Phi\left(\frac{-|d|}{2}\right)$$

- Also, für $d = 1$:

$$\bullet \quad OVL = 2\Phi\left(\frac{-1}{2}\right) = 2\Phi(-0,5) = 2 \cdot 0,3085 = 0,617$$

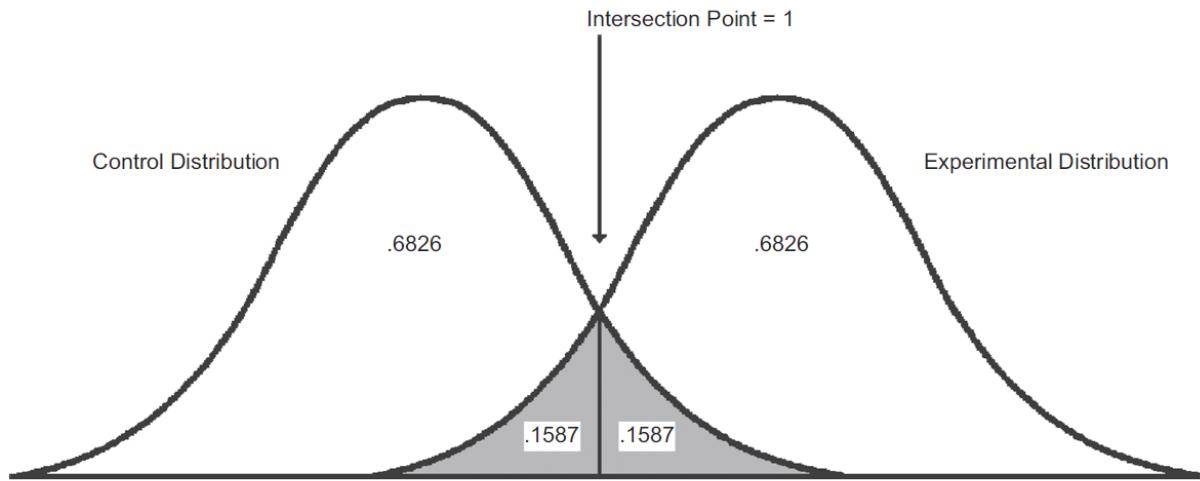
- die Schnittmenge überdeckt wieviel Prozent der Fläche beider Verteilungen?

$$OVL_C = \Phi\left(\frac{-|d|}{2}\right) / \Phi\left(\frac{|d|}{2}\right)$$

- Also, für $d = 1$:

$$\bullet \quad U_1 = 0,3085 / \Phi(0,5) = \frac{0,3085}{0,6915} = 0,446$$

- $d = 2$
- der Schnittpunkt liegt stets bei $d/2$ (hier also bei $d = 1$)
- $\Phi(d=1) = 0,8413 \quad \Phi - d = -1 = 0,1587$



- Fläche einer Verteilung: $0,6826 + 0,1587 + 0,1587 = 1$
- OVL: $2 * 15,87\% = 31,74\%$ dieser Fläche werden von der anderen Verteilung überdeckt,
- Gesamtfläche beider Verteilungen: $0,6826 + 0,6826 + 0,1587 + 0,1587 = 1,6826$
- die Fläche der Schnittmenge beträgt 0,3174
- also: $OVL_C = 31,74\% / 168,26\% = 18,86\%$

wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person aus der Experimentalgruppe einen höheren Wert hat als eine zufällig ausgewählte Person aus der Kontrollgruppe?

$$AUC = \Phi\left(\frac{d}{\sqrt{2}}\right)$$

- Also, für $d = 1$:

$$\bullet \quad AUC = \Phi\left(\frac{1}{\sqrt{2}}\right) = \Phi(0,71) = 0,761$$

einige exemplarische Werte:

d	U_3	OVL	OVL_C	AUC
0,2	0,58	0,92	0,85	0,56
0,5	0,69	0,80	0,67	0,64
0,8	0,79	0,69	0,53	0,71
1	0,84	0,61	0,45	0,76
1,2	0,93	0,45	0,29	0,86
2	0,98	0,32	0,19	0,92

Berechnung von d und g aus Signifikanztestergebnissen

!!!! Effektgröße = Signifikanztest/Größe der Studie (N) !!!!

falls $n_A = n_B \rightarrow$

$$d = \frac{2t}{\sqrt{df}}$$

$$g = \frac{2t}{\sqrt{N}}$$

falls $n_A \neq n_B \rightarrow$

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

$$g = \frac{t \cdot (n_1 + n_2)}{\sqrt{N} \cdot \sqrt{n_1 \cdot n_2}}$$

im Beispiel: $t = 2,298$; $n_A = 6$; $n_B = 5$; $df = 9$

$$d = \frac{t \cdot (6 + 5)}{\sqrt{9} \cdot \sqrt{6 \cdot 5}} = 1,54$$

$$g = \frac{t \cdot (6 + 5)}{\sqrt{11} \cdot \sqrt{6 \cdot 5}} = 1,39$$

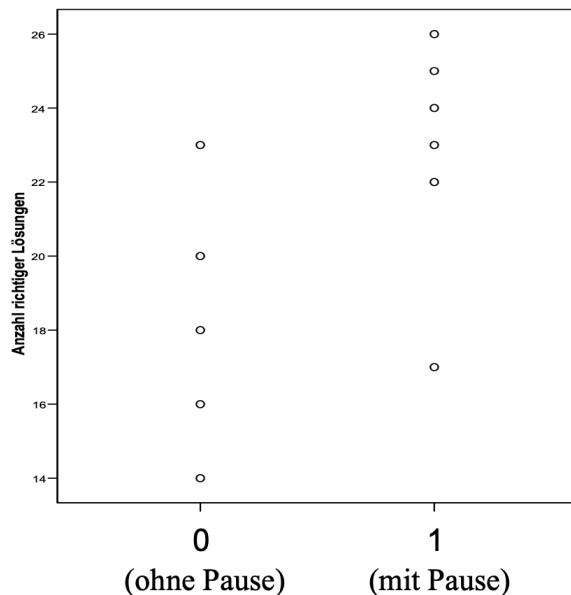
Äquivalenz von Abstandsmaßen (d) und Zusammenhangsmaßen (r)

Ohne Pause	Mit Pause
18	22
16	24
20	17
14	25
23	23
	26

Wie groß ist der standardisierte
Mittelwertsunterschied im
Lernerfolg? $\Rightarrow d!$



Gleichbedeutend!
(Standardisierter Mittelwertsunterschied \approx
Korrelation)



Wie groß ist der Zusammenhang
zwischen dem Einlegen einer Pause
und dem Lernerfolg? $\Rightarrow r!$

die Effektgröße r

Pause? UV	Anzahl richtiger Lösungen AV
0	18
0	16
0	20
0	14
0	23
1	22
1	24
1	17
1	25
1	23
1	26

0 - ohne Pause
1 - mit Pause

$$r_{XY} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y} \quad \text{oder} \quad r_{pbis} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \sqrt{\frac{n_1 n_0}{N^2}}$$

→ x: AV; y: UV

dabei sind:

n_0 : Anzahl der Personen mit $y = 0$

n_1 : Anzahl der Personen mit $y = 1$

\bar{x}_0 : Mittelwert von x über alle Personen mit $y = 0$

\bar{x}_1 : Mittelwert von x über alle Personen mit $y = 1$

s_x : Standardabweichung von x in der Gesamtstichprobe

im Beispiel:

$$s_x = 3,79$$

$$r = \frac{22,83 - 18,2}{3,79} \cdot \sqrt{\frac{5 \cdot 6}{11^2}} = 0,61$$

Berechnung von r aus Signifikanztestergebnissen

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

im Beispiel: $t = 2,298$; $df = 9$

$$r = \sqrt{\frac{2,298^2}{2,298^2 + 9}} = 0,61$$

Varianz von r

$$V_r = \frac{(1 - r^2)^2}{n - 1}$$

- die Varianz der Korrelation ist damit stark vom beobachteten Effekt abhängig - großer Effekt > kleine Varianz → dies soll vermieden werden,
- häufig werden daher Fisher-Z-transformierte Korrelationen verwendet,
- Fisher-Z-Transformation:

$$z = 0,5 \times \ln\left(\frac{1 + r}{1 - r}\right)$$

- Varianz Fisher-Z-Werte:

$$V_z = \frac{1}{n - 3}$$

- Rücktransformation:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Was bedeutet die Effektgröße r inhaltlich? - BESD

- fiktives Setting
 - 100 Vpn in Experimentalgruppe,
 - 100 Vpn in Kontrollgruppe,
 - anhand der AV wird ein Mediansplit mit allen 200 Vpn durchgeführt,
 - wir erhalten eine Gruppe von Vpn mit den 100 schlechtesten Ergebnissen und eine zweite Gruppe von Vpn mit den 100 besten Ergebnissen,

	Ergebnis		100
	gut	schlecht	
Bedingung	EG		100
	KG		100
	100	100	

- wie viele der Vpn in der Experimentalgruppe erzielen gute Ergebnisse? - Erfolgsrate in der EG
→ diese Frage lässt sich anhand der Effektgröße r beantworten!

- Erfolgsrate in der EG: $.50 + r/2$
- Erfolgsrate in der KG: $.50 - r/2$
- für $r = 0,6$ ergibt sich also:

		Ergebnis		
		gut	schlecht	
Bedingung	EG	80	20	100
	KG	20	80	100
		100	100	

- Rationale:
 - anhand der Werte in einer Vierfeldertafel lässt sich eine korrelative Effektgröße r berechnen,
 - die Werte im BESD führen wiederum zur ursprünglichen Effektgröße r

BESD mit dichotomer AV

tatsächliches Studienergebnis:

		Herzinfarkt		
		ja	nein	
Bedingung	Aspirin	104	10933	11037
	Placebo	189	10845	11034
		293	21778	

$$\rightarrow r = 0,034$$

BESD:

		Herzinfarkt		
		ja	nein	
Bedingung	Aspirin	48,3	51,7	100
	Placebo	51,7	48,3	100
		100	100	

Transformation von Effektstärken

$$d = \frac{2r}{\sqrt{1-r^2}}$$

$$d = g \sqrt{\frac{N}{N-2}}$$

$$g = d \sqrt{\frac{N-2}{N}}$$

$$g = \frac{2r}{\sqrt{1-r^2}} \sqrt{\frac{N-2}{N}}$$

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

$$r = \frac{g}{\sqrt{g^2 + 4 \frac{N-2}{N}}}$$

Was ist ein großer Effekt?

- die Frage ist äquivalent zu *Was ist eine große Länge?* - sie ist in dieser Form also – streng genommen – sinnlos!
- die Antwort richtet sich grundsätzlich nach dem jeweiligen Kontext - vor allem Effektgrößen aus vergleichbaren Studien,
- fehlen relevante Vergleichsgrößen, können Konventionen zur Beurteilung von Effektgrößen helfen (Cohen, 1988):

	<i>d</i>	<i>r</i>
klein	0,2	0,1
mittel	0,5	0,3
groß	0,8	0,5

- diese Konventionen haben sich in der sozialwissenschaftlichen Forschung bewährt,
 - d.h., sie entsprechen recht gut den üblicherweise gefundenen Effekten

Die Effektstärke η^2

- als Effektstärke wird in der Varianzanalyse η^2 (griechisch: *eta*) verwendet.,
- sie gibt den Anteil der (durch die UV) erklärten Variation an der gesamten Variation an,

$$\eta^2 = \frac{QS_{zw}}{QS_{Ges}}$$

- falls die Quadratsummen nicht verfügbar sind:

$$\eta^2 = \frac{F \cdot df_{zw}}{F \cdot df_{zw} + df_{inn}}$$

- falls exakt 2 Gruppen verglichen werden, gilt:

$$\eta^2 = r^2$$

- Konventionen zur Beurteilung der Effektgröße:

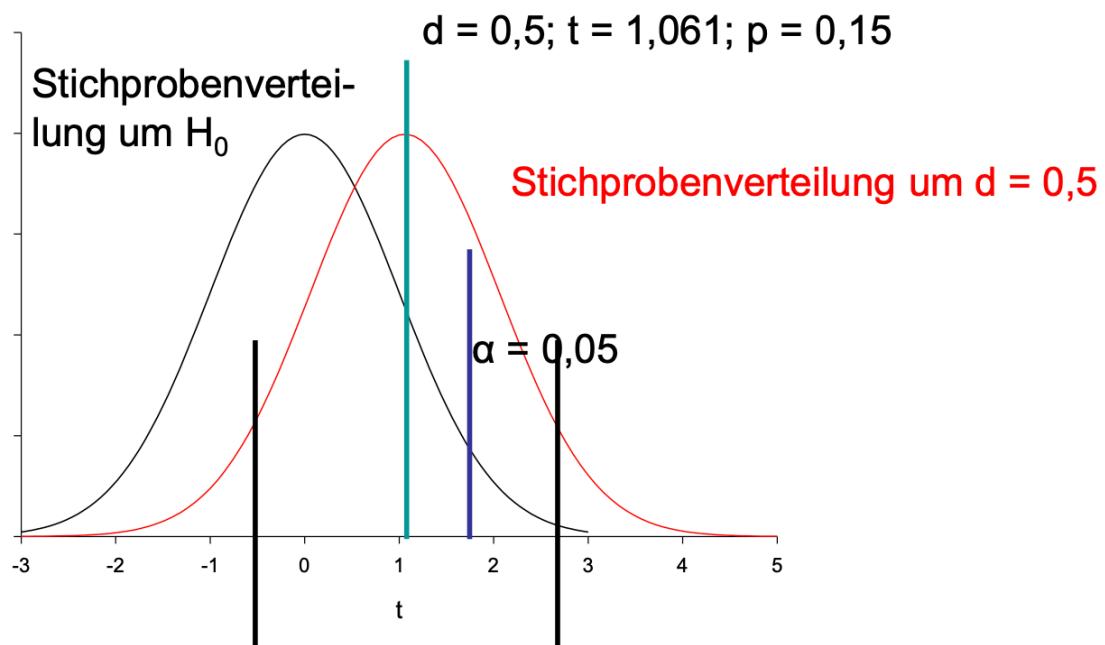
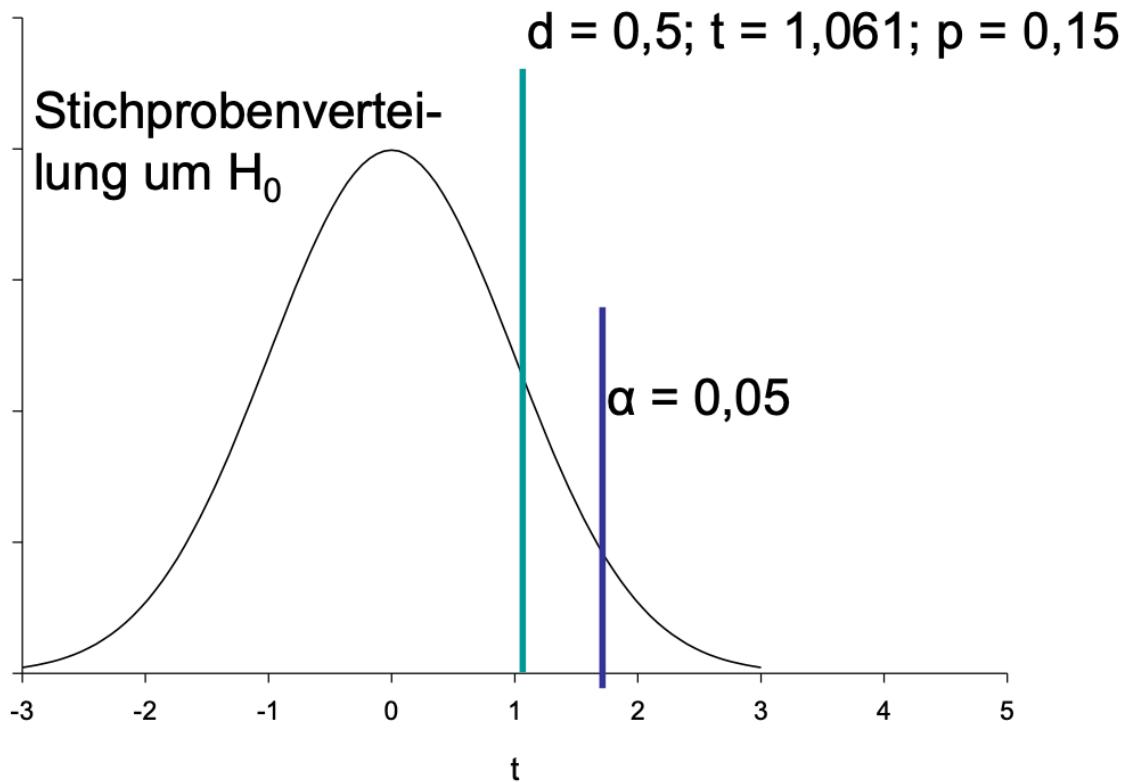
- kleiner Effekt: $\eta^2 = 0,01$
- mittlerer Effekt: $\eta^2 = 0,06$
- großer Effekt: $\eta^2 = 0,14$

Counternull-Werte

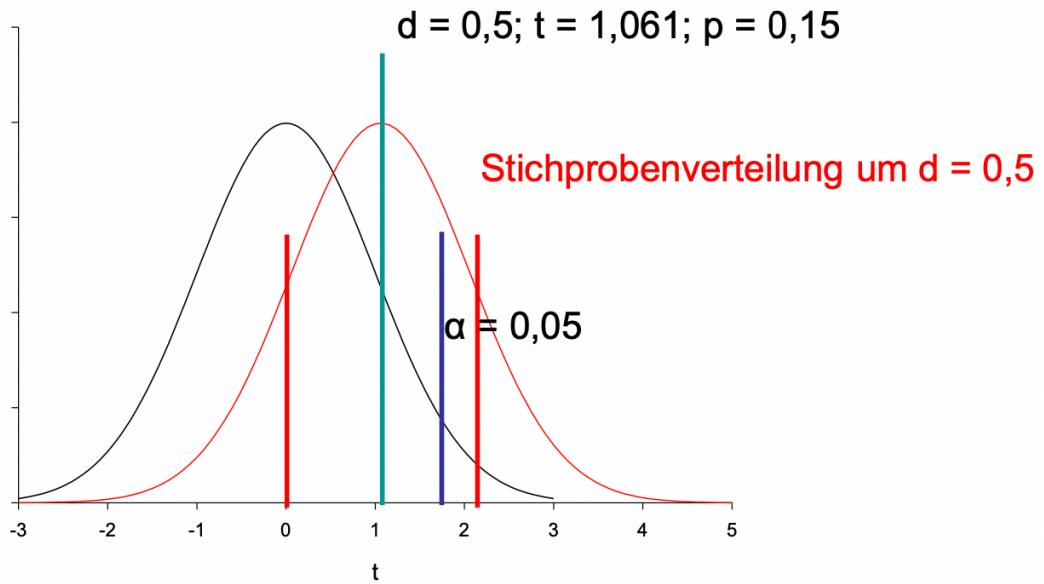
$n_1 = n_2 = 15$

$H_0: d = 0$

beobachtete Effektstärke: $d = 0,5$



→ 90%-Konfidenzintervall: der wahre Effekt liegt mit 90%iger Sicherheit zwischen $d = -0,32$ und $d = 1,32$



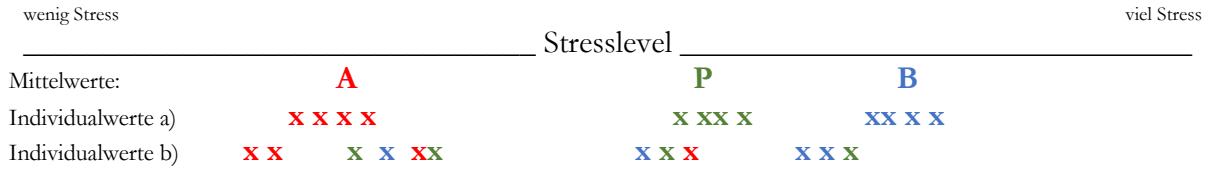
- Counternull-Wert = $2 \cdot ES_{\text{beob}} - ES_{\text{null}} = 2 * 0,5 = 1$ bzw. $t = 2,122$
- Counternull-Intervall: der wahre Effekt liegt mit **70%iger** Sicherheit zwischen $d = 0$ und $d = 1$
- die Intervallgrenzen ergeben sich aus dem ursprünglichen p -Wert!

1. Neyman-Pearson-t-Test: man vergleicht einen CoronaSchnelltest (H_0 = infiziert) und den Hypothesentest in einer psychologischen Fragestellung - bei welchem der beiden Tests würde man eher einen Fehler 1.Art als 2.Art in Kauf nehmen?
 - bei einem CoronaSchnelltest sollte der Fehler 2.Art so gering wie möglich sein, damit das Risiko, dass eine fälschlicherweise negativ getestete Person Andere ansteckt, so gering wie möglich ist → Fehler 1.Art bei CoronaSchnelltest sollte in Kauf genommen werden
2. was sind die Möglichkeiten, um möglichst sicher zu gehen, dass man einen besonders kleinen Effekt in einer empirischen Untersuchung nachweisen kann?
 - besonders große Stichprobe,
 - kleines alpha
3. es liegen die Medikamente A und B vor - beide wurden entwickelt, um Stress zu reduzieren, den man mit einem hormonellen Marker im Speichel an einem einzelnen Untersuchungszeitpunkt aufnimmt - in der Studie zu diesen Medikamenten muss man außerdem eine Placebo-Gruppe einführen!

wie viele Gruppen würde man in einem between-subjects Querschnitt-Design untersuchen?

was ist die AV?

 - 3Gruppen: Medikament A, Medikament B, PlaceboGruppe
 - AV - Stresslevel



bei welchem Ergebnis kann man sich sicherer sein, dass es sich überhaupt um verschiedene Medikamente handelt?

- bei a), weil hier Medikament A stärker abgegrenzt ist,
- die Varianz drückt die Abstände zwischen den einzelnen Werten als durchschnittliche quadrierte Abweichung aus,
- ebenso auch die Abstände zwischen den einzelnen Gruppenmittelwerten,
- Bildung eines GesamtMittelwertes möglich,
- ein Test, der die Unterschiede in der Interpretation zwischen a) und b) deutlich macht, ist die Varianzanalyse

$$F = \frac{\hat{\sigma}_{zw}^2}{\hat{\sigma}_{inn}^2} = \frac{\text{[Diagramm: zwei Dichtekurven, eine breiter unter dem Bruchstrich, eine schmäler darüber]}}{=} = \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$$

- entspricht die Zeichnung eher Fall a) oder Fall b)?
 - Fall a), weil bei b) die Varianz zwischen den Werten deutlich größer ist und somit die Kurve unter dem Bruchstrich breiter sein müsste
- zeichne einen Fall für $F = 1$

4. Wann solltest du Hypothesen aufstellen: vor oder nach der Datenerhebung?
 - konfirmatorisch - ausschließlich vor der Erhebung
 - explorativ - nach der Erhebung ist möglich

5. versuche die Idee des HARKing kurz zusammenzufassen...
 - dabei werden große Datenmengen gesammelt und zahlreiche Signifikanztests durchgeführt - für ein beobachtetes signifikantes Resultat wird im Nachhinein nach einer passenden Hypothese gesucht und das gesamte Vorgehen als Test dieser Hypothese ausgegeben...

6. der Effekt einer Studie konnte nicht repliziert werden - woran kann das liegen?
 - Originalstudie falsch positiv,
 - Replikation falsch negativ,
 - zu große Abweichung zwischen den Studien, sodass sie verschiedene Effekte testen,

7. zu welchem Zweck werden Effektgrößen standardisiert?
 - um Aussagen über die Größe und Richtung eines Effektes unabhängig von der Maßeinheit der Skala treffen zu können - bei Standardisierung fällt Maßeinheit weg,
 - Maßeinheiten werden häufig willkürlich gewählt,
 - z.B. *das durchschnittliche Befinden ist bei Behandelten um 3 Punkte besser oder:
der Effekt einer Pause auf Diktat im Vergleich zu Vokabellernen*
! erst durch Standardisierung wird ein Effekt interpretierbar

8. wie passiert Standardisierung?
 - den Effekt (z.B. Abstand von Mittelwerten) ins Verhältnis mit der (gepoolten) Standardabweichung setzen,

9. nenne die drei verschiedenen Klassen standardisierter Effektgrößen und erläutere kurz den Kerngedanken je eines Vertreters!
 - Abstandsmaße d und g geben an, wie groß standardisierte Mittelwertsunterschiede sind,
 - Zusammenhangsmaße r und φ (Phi) geben an, wie groß der Zusammenhang zwischen zwei Variablen ist,
 - Maße der Varianzaufklärung η² oder R² geben an, wie viel Varianz an der Gesamtvarianz aufgeklärt wird bzw. wie viel die uV an der aV erklärt

10. erkläre in eigenen Worten den Unterschied zwischen Cohens d und Rosenthals g - wann sollte ich welches Maß verwenden?
 - Cohens d ist ein Maß für Stichprobenunterschiede,
 - es berechnet sich mithilfe der Stichprobenvarianz,
 - sie bezeichnet den standardisierten Mittelwertsunterschied in der Stichprobe,

$$d = \frac{\mu_A - \mu_B}{\sigma_{AB}}$$

- Rosenthals g berechnet sich mithilfe der geschätzten Populationsvarianz,
- bestmögliche Schätzung des Populationseffekts aus den Stichprobendaten,
- eher bei kleinen Stichproben, weil dort mehr korrigiert wird,

$$g = \frac{\mu_A - \mu_B}{\hat{\sigma}_{AB}}$$

11. was bedeutet es für die zu korrigierende Effektgröße, wenn J sich an 1 bzw. 0 annähert?

- je näher J an 1 ist, desto besser hat die zu korrigierende Effektgröße bereits den Effekt für die Population geschätzt,
- je näher J an 0 kommt, desto mehr hat die zu korrigierende Effektgröße den Effekt überschätzt - de facto ist J aber nur selten kleiner als 0,9,
- Rosenthal $g * J =$ Borensteins g

12. du möchtest die folgenden Effektstärken und Prüfgrößen aus den verschiedenen Studien zusammen fassen - der erste Schritt besteht darin, Borensteins g aus den jeweiligen Daten zu berechnen - außerdem brauchst du gegebenenfalls die jeweiligen Varianzen! Hinweise: es handelt sich immer um *between subjects*- Designs mit 2 Gruppen ($df = N - 2$) - im Zweifel sind die Gruppengrößen gleich - bei 2 Gruppen gilt $F = t^2!$ - vergiss nicht, dass Rosenthals g Borensteins d entspricht!

Studie	Daten
1	$r=0.3, N=50$
2	$d=0.5, N=90$
3	$M1=7, M2=9, s1=s2=4, n1=30$
4	$t(54)=1.83, n1=27, n2=29$
6	$M1=120, M2=150, s1=s2=80, N=50$
8	$F(1,48)=4.01$

- Borenstein's g bietet die bestmögliche Schätzung durch die erneute Korrektur des Effekts!

Merkregel zur Umrechnung von Cohen/Rosenthal/Borenstein:

Cohens...

Rosenthals...

Borenstein...

$$d \xrightarrow{\sqrt{\frac{(N-2)}{N}}} g = d \xrightarrow{J = 1 - \frac{3}{4*df-1}} g$$

*bestmögliche Schätzung durch erneute Korrektur des Effekts!

$$J = 1 - \frac{3}{4*df-1} \rightarrow \text{Korrekturfaktor!}$$

Studie	Daten	Borensteins g	Varianz
1	r=0.3, N=50		
2	d=0.5, N=90		
3	M1=7, M2=9, s1=s2=4, n1=30		
4	t(54)=1.83, n1=27, n2=29		
6	M1=120, M2=150, s1=s2=80, N=50		
8	F(1,48)=4.01		

13. du liest in einem Artikel von einer Studie mit zwei Bedingungen mit jeweils 11 Personen:
Schlafentzug (4h Schlaf vs. 8h Schlaf) verringert das Wohlbefinden am Morgen signifikant (alpha = 5%) - es handelt sich um ein *between subject*- Design, das mit einem t-Test berichtet wurde
- die Studie berichtet aber leider keinen Effekt, sondern nur den empirischen p-Wert
→ .05 - berechne Cohen's d!
- Df = 20,
- einseitig t = (ca.) 1.725
- $d = 2 * 1.725 / \sqrt{20} = .771$

mehr Probleme des Signifikanz-Tests und ein Bayesianischer (Aus-)Blick

Thesen

1. wir sollten über die Wahrscheinlichkeit von Hypothesen nachdenken!
- vor und nach der Sammlung von Evidenzen/Daten
2. wir tun das zumindest *offiziell* nie
3. dies führt dazu, dass wir vorliegende Evidenz nicht in einer kohärenten Weise zur Bewertung von Hypothesen nutzen
4. es führt zudem dazu, dass wir die Unsicherheit in unseren Schlussfolgerungen nicht angemessen kommunizieren
5. die Konsequenzen sind zahlreiche falsche Forschungsergebnisse und mangelnder wissenschaftlicher Fortschritt
6. all dies basiert auch auf einer fehlerhaften Interpretation des Ergebnisses des Signifikanztests - also $p(D | H_0)$,

Bayes-Theorem

→ Beschreibung der Beziehung zwischen zwei bedingten Wahrscheinlichkeiten,

$$p(H | D) = \frac{p(D | H) \cdot p(H)}{p(D)}$$

- Ziel: Wahrscheinlichkeitsrevision!
- es beschreibt, wie wir unsere Überzeugungen ($p(H)$) ändern sollten (in $p(H | D)$), nachdem ein neues Datum (D) beobachtet wurde,
- es besteht ein Verhältnis zwischen der bedingten Wahrscheinlichkeit zweier Ereignisse $P(A | B)$ und der umgekehrten Form $P(B | A)$,

$$p(H | D) = \frac{p(D | H) \cdot p(H)}{p(D)}$$

↑ ↑

mit Info
über Daten **ohne Info**
über Daten

oder

$$p(H_1 | D) = \frac{p(D | H_1) \cdot p(H_1)}{p(D | H_1) \cdot p(H_1) + p(D | H_0) \cdot p(H_0)}$$

allgemein:

$$p(H | D) = \frac{\sum_H p(D | H) \cdot p(H)}{\sum_H p(D | H) \cdot p(H)}$$

klassische Kritik am Signifikanztest

das Ergebnis des Tests entspricht nicht dem eigentlichen Interesse - z.B. Cohen, 1994,
der Test ist als Mittel der Theorieprüfung belanglos - z.B. Meehl, 1967, 1997,
die Fehlerrate des Tests ist extrem hoch - z.B. Hunter, 1997,
...[der Signifikanztest] ist ein potenter, aber steriler intellektueller Wüstling, der auf seinem vergnügten Weg einen langen Zug geschändeter Mädchen zurücklässt, aber keine lebensfähigen wissenschaftlichen Nachkommen.
- Meehl, 1967

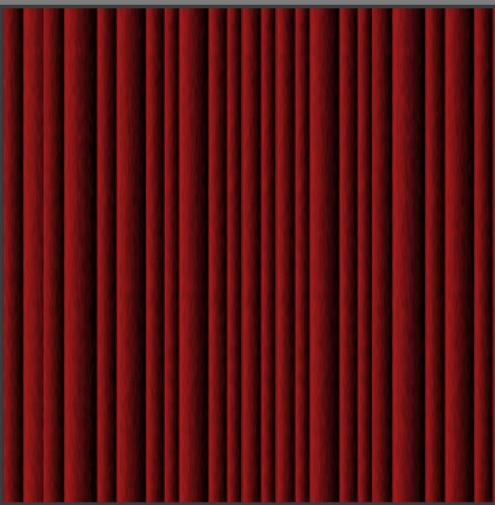
Probleme des Signifikanztests

- ein nicht-signifikantes Ergebnis ist bestenfalls schwache Evidenz für die H_0
 - das *Power-Problem*,
- ein signifikantes Ergebnis ist oftmals schwache und unzureichende Evidenz für die H_1

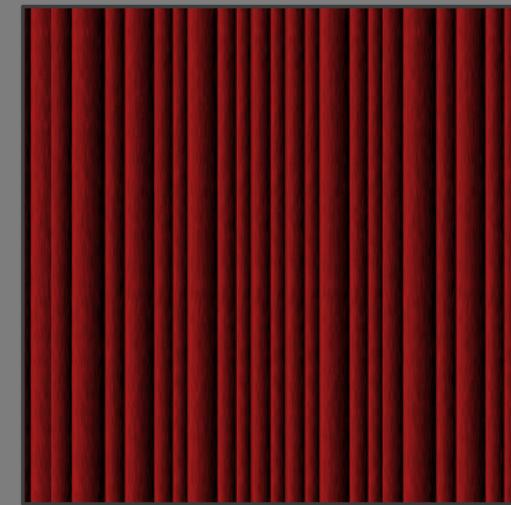


- Bem, 2011
 - untersucht das Phänomen der precognition - in etwa *Hellseherei*
 - die übergeordnete Idee ist, etablierte Effekte *zeitlich umzukehren*
 - Reaktionen werden erfasst, bevor die Ursachen auftreten!
 - 10 Studien dieser Art - 9 Studien finden signifikante Effekte und demonstrieren so die Existenz von precognition,
 - ein Beispiel...

Wo ist der Zielreiz?



A



B

- Bem beachtete die üblichen methodischen und statistischen Standards in der Psychologie äußerst genau,
- *Bem played by the implicit rules that guide academic publishing - in fact, Bem presented many more studies than would usually be required*
- irgendwas an den Standards der Evidenzbewertung in der Psychologie muss also falsch sein!

→ 3 Haupt-Probleme:

- wie hoch ist die Wahrscheinlichkeit für die H_1 nach einem signifikanten Ergebnis?
 - und wovon hängt das eigentlich ab?
- was passiert, wenn wir mit dem Signifikanztest *unsicher* umgehen? - der Zufall und seine Gefahren...
- wie wir es der H_1 zu leicht machen: Evidenz ist ein relatives Konzept!

Wie hoch ist $p(H_1 | \text{signifikantes } D)$?

- es ist unklar, wie precognition funktionieren sollte
- keine Evidenz für precognition im Alltag
- wir sollten der H_1 also eine geringe Prior-Wahrscheinlichkeit zuordnen,
- Annahme: $p(H_0) = .999$ und damit $p(H_1) = .001$

Bayes-Theorem

$$p(H_1 | D) = \frac{p(D | H_1) \cdot p(H_1)}{p(D | H_0) \cdot p(H_0) + p(D | H_1) \cdot p(H_1)}$$

$$p(H_1 | D) = \frac{0.95 \cdot 0.001}{0.05 \cdot 0.999 + 0.95 \cdot 0.001} = 0.019$$

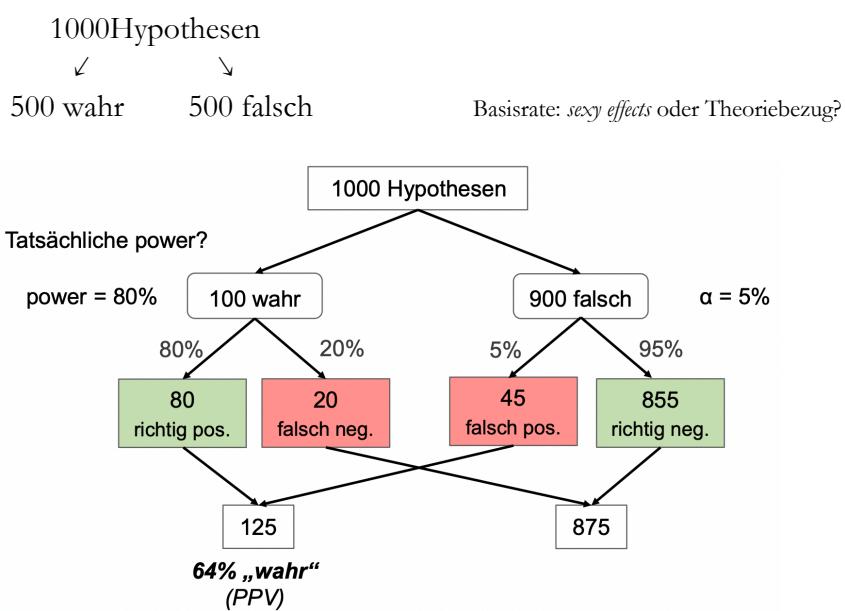
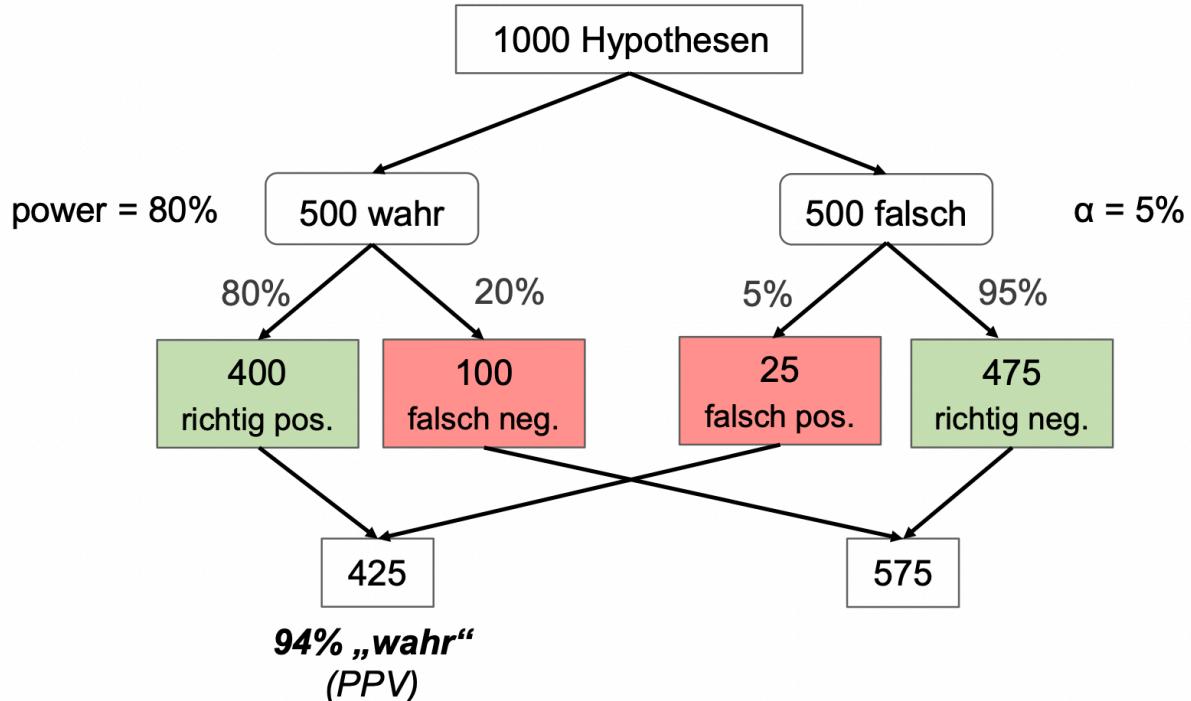
→ außergewöhnliche Behauptungen verlangen außergewöhnliche Evidenz!

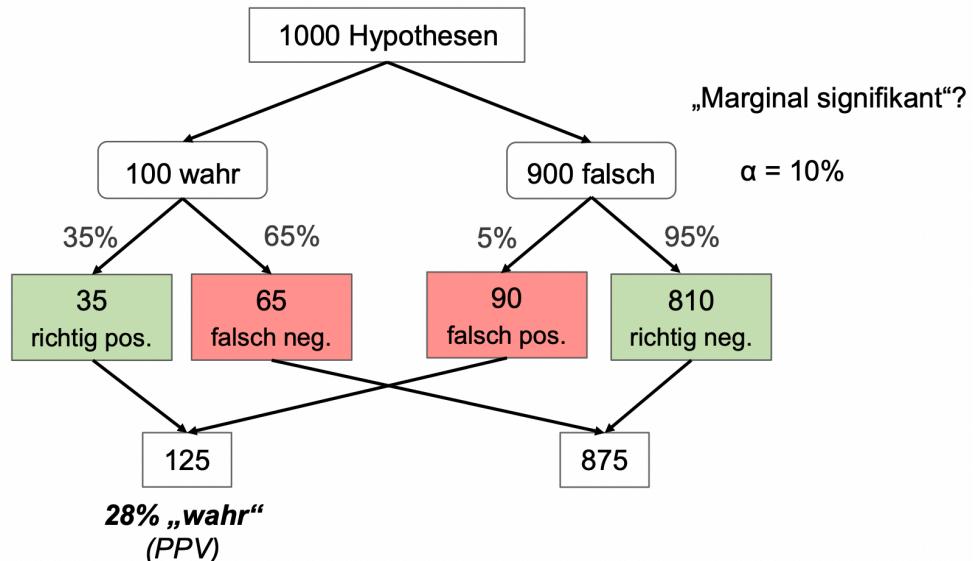
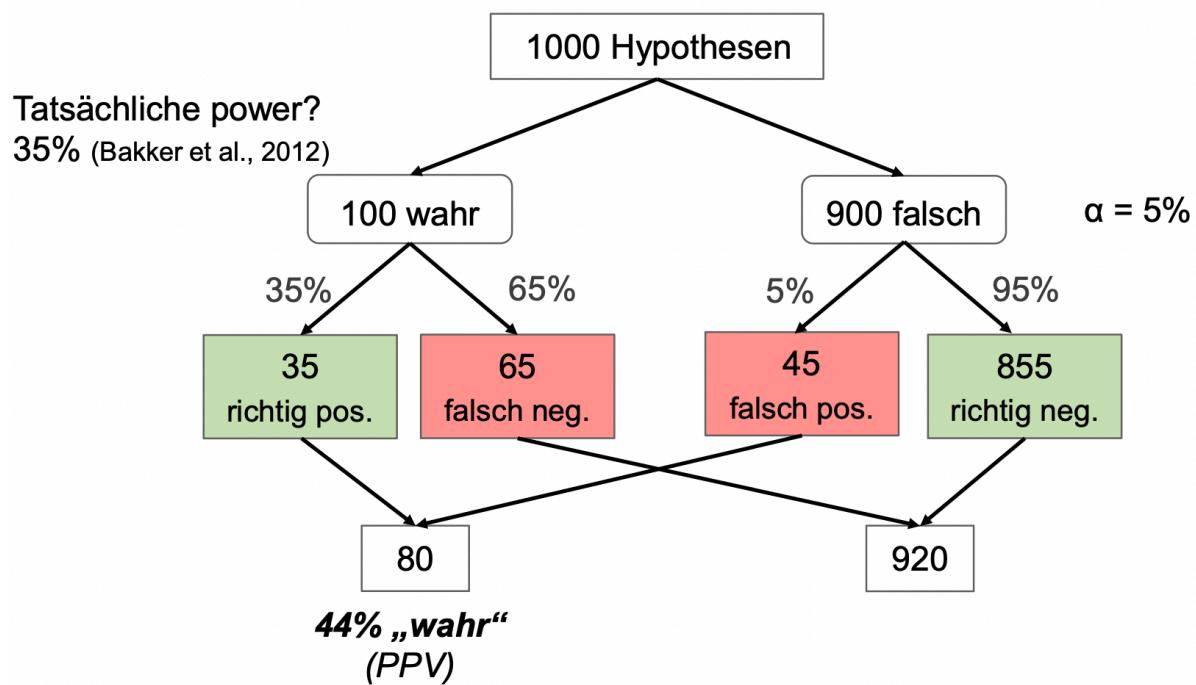
→ Standard Gegenargument: Prior-Wahrscheinlichkeit ist subjektiv,

- ich halte es für völlig irrational, keine Prior-Wahrscheinlichkeit zu vergeben!

- Subjektiv meint nicht arbiträr - nach Ermessen/willkürlich,
- Alltags-Kognition würde ohne Prior-Wahrscheinlichkeit nicht funktionieren,
- Wissenschaft zielt auf Erkenntnisgewinn - sie SOLLTE zumindest zu einer Veränderung der Wahrscheinlichkeiten von Hypothesen führen,
- eine prinzipielle Ablehnung von Prior-Wahrscheinlichkeit impliziert, dass nichts gelernt werden kann,
- die Publikation von Bems Experimenten in JPSP impliziert eine niedrige Prior-Wahrscheinlichkeit für precognition,
 - * können Sie sich vorstellen, dass JPSP einen Artikel mit dem Titel *there is no precognition – conclusive evidence from nine studies* publiziert hätte??

Signifikanztests





		Tatsächlicher Effekt?		Total
		Ja	Nein	
Befund	$p \leq .05$			
	$p > .05$			
Total			1000	

power $\approx 50\%$
Prior Wahrscheinlichkeit, dass eine (in der Psychologie getestete) H_1 wahr ist
- Annahme: 50%

		Tatsächlicher Effekt?		Total
		Ja	Nein	
Befund	$p \leq .05$	250	25	275
	$p > .05$	250	475	725
Total		500	500	1000

$$p(H_1 | D) = \frac{250}{250+25} = 90.9\%$$

Beispiel von Ioannidis:
es wird geprüft, welche von 100000 Genvarianten mit Schizophrenie assoziiert sind - aus theoretischen Gründen kann erwartet werden, dass 10Genvarianten tatsächlich mit Schizophrenie assoziiert sind, also:

		Tatsächlicher Effekt?		Total
		Ja	Nein	
Befund	$p \leq .05$	6	4,999.5	5005.5
	$p > .05$	4	94,990.5	94,994.5
Total		10	99,990	100,000

$$p(H_1 | D) = \frac{6}{6+4999.5} \approx 0.12\%$$

Wie hoch ist der Anteil wahrer Alternativhypthesen in der Psychologie?
natürlich unbekannt - es gibt aber gute Gründe anzunehmen, dass der Anteil deutlich kleiner ist als 50% - wir akzeptieren sogar Hypothesen über precognition - in einer typischen Studie wird eine Vielzahl von Tests durchgeführt, z.B. diverse AVs, Geschlechtsunterschiede, sonstige Kovariaten...

		Tatsächlicher Effekt?		Total
		Ja	Nein	
Befund	$p \leq .05$	50	45	95
	$p > .05$	50	855	945
Total		100	900	1000

$$p(H_1 | D) = \frac{50}{50+45} = 52.6\%$$

generell ist $p(H_1 | \text{signifikantes } D)$ größer als 50%, wenn:

$$\text{power} \cdot \frac{p(H_1)}{1-p(H_1)} > \alpha$$

mit power = .5 und $\alpha = .05$:

$$.5 \cdot \frac{p(H_1)}{1-p(H_1)} > .05$$

$$\frac{p(H_1)}{1-p(H_1)} > .1 \Leftrightarrow p(H_1) > .091$$

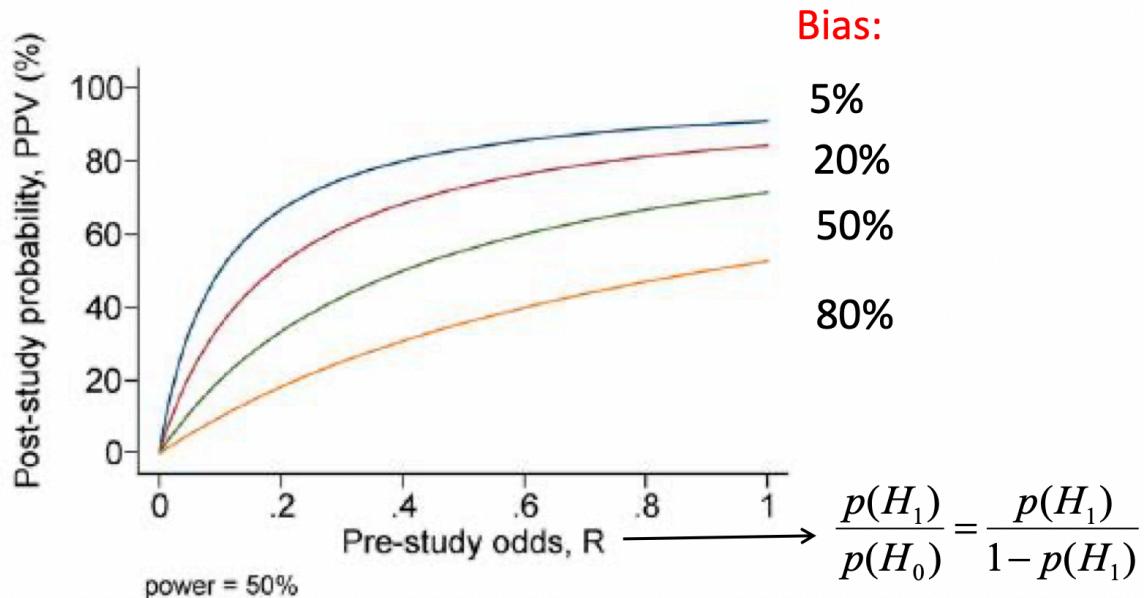
→ veröffentlichte Resultate unterliegen zudem einem Bias...

- Bias: Anteil *künstlich generierter* signifikanter Ergebnisse
- Manipulation, Betrug

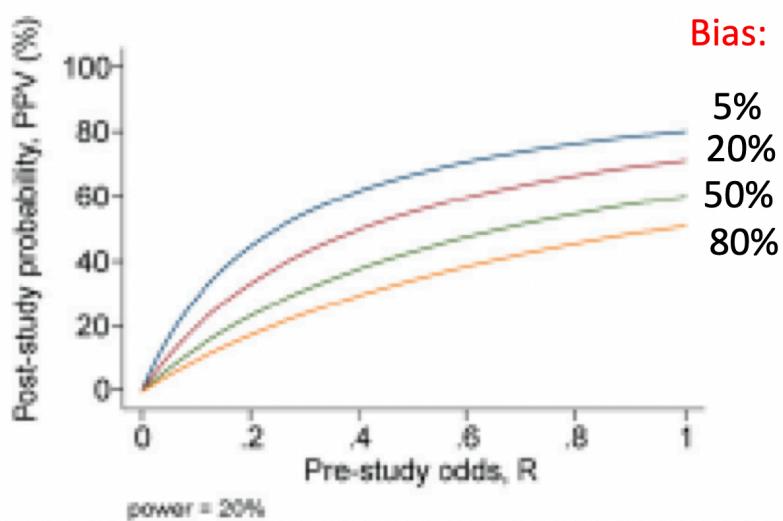
- reporting Bias: eine korrekte H_0 wird wiederholt getestet bis ein signifikantes Ergebnis erzielt wird - dieses wird veröffentlicht,
- wir können sicher sein, dass dies vorkommt... aber wie oft???
 - Ioannidis nimmt Bias-Raten zwischen 5% und 80% an!

Bias: 20%		Tatsächlicher Effekt?			
		Ja	Nein	Total	
Befund	$p \leq .05$	260	1250	1510	
	$p > .05$	260	480	740	
Total		500	500	1000	

$p(H_1 | D) = 71.4\%$



Ioannidis, 2005



→ die Antwort hängt ab von:

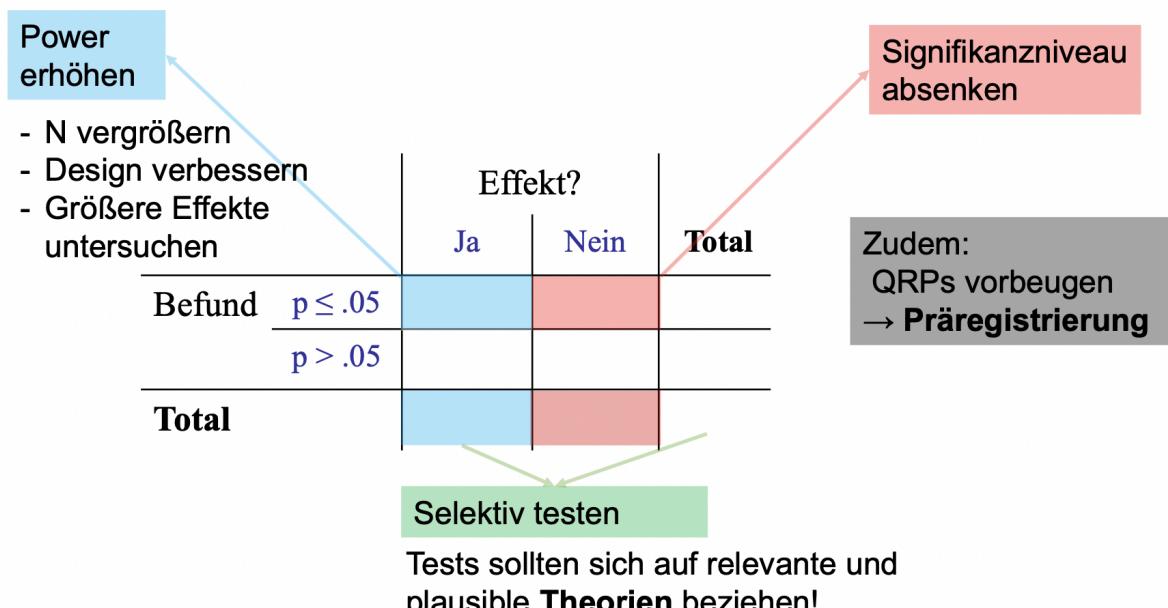
- der Prior-Wahrscheinlichkeit der Forschungshypothese - dem Anteil wahrer Forschungshypothesen in einem Feld,
- der Power, mit der die Hypothese getestet wurde - der *typischen* Power in einem Feld,
- Auftreten und Ausmaß von Publication Biases und Questionable Research Practices,
- dem verwendeten Signifikanzkriterium α

→ einige Implikationen - Ioannidis, 2005:

- ! je kleiner die Stichprobengrößen in einem Feld, desto kleiner die Wahrscheinlichkeit, dass Forschungsbefunde *wahr* sind,
- ! je kleiner die Effektgrößen in einem Feld, desto kleiner die Wahrscheinlichkeit, dass Forschungsbefunde *wahr* sind,
- ! je größer die Anzahl und je schwächer die Auswahl der überprüften Hypothesen in einem Feld, desto kleiner die Wahrscheinlichkeit, dass Forschungsbefunde *wahr* sind,
- ! je größer die Flexibilität in Designs, Definitionen, aVs und statistischen Prozeduren in einem Feld, desto kleiner die Wahrscheinlichkeit, dass Forschungsbefunde *wahr* sind,

→ **ich fürchte, all dies heißt nichts Gutes für die Psychologie...**

was können wir statistisch tun?

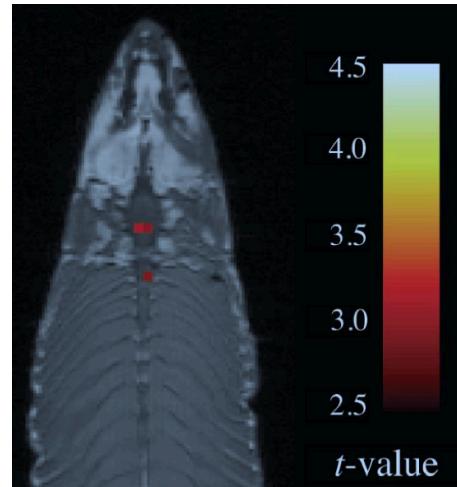


→ gilt für jedes hypothesenprüfende Verfahren - nicht nur für Signifikanztests!

der Zufall und seine Gefahren...

- psychologische Daten hängen vom Zufall ab,
- zuweilen wird der Zufall selbst bei der absurdesten Forschungshypothese zu einem signifikanten Ergebnis führen - in 5% aller Studien, wenn $\alpha = 5\%$,
- ein extremes Beispiel...

ein toter Lachs im fMRI-Scanner - Bennett et al., 2010:
der Lachs identifiziert menschliche Emotionen auf Fotos - Veränderung im BOLD-Signal zwischen Präsentation und Pausen - wird in mehreren Tausend Voxeln getestet,
- $\alpha = .001$,
Schwelle: Ausdehnung von 3 Voxeln,
dies könnte durch eine angemessene α -Korrektur behoben werden - aber:
...a sizable percentage of results still utilize uncorrected statistics - about 30%



- wir können also alles mit dem Signifikanztest beweisen, wenn wir es nur oft genug versuchen,
- was ist *oft genug*?
- bei einem $\alpha = 5\%$ erwarten wir ein signifikantes Ergebnis in 20 Studien,
- Ihnen fehlen Zeit und Geld für 20 Studien? - nun, da gibt's noch zahlreiche Möglichkeiten...
- um ein signifikantes Ergebnis zu erzielen, muss grundsätzlich lediglich die Zahl der Tests erhöht werden:
 - berücksichtigen Sie zusätzliche und austauschbare aVs,
 - erfassen Sie so viele Kovariaten, wie Ihnen einfallen - einige werden die Fehlervarianz reduzieren,
 - probieren Sie verschiedene Transformationen Ihrer Daten,
 - benutzen Sie andere statistische Tests,
 - schließen Sie einige Teilnehmer aus - es gibt bestimmt einen guten Grund!
 - testen Sie nach 20, 30, 40, 50... Teilnehmern - beenden Sie den Test, sobald Sie ein signifikantes Ergebnis gefunden haben - und nur dann → *data peeking*
- Simulationsstudie - Simmons, Nelson & Simonsohn, 2011:

Table I. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

→ wiederholte t-Tests nach je einer zusätzlichen Vp in jeder Bedingung - Simmons, Nelson & Simonsohn, 2011:

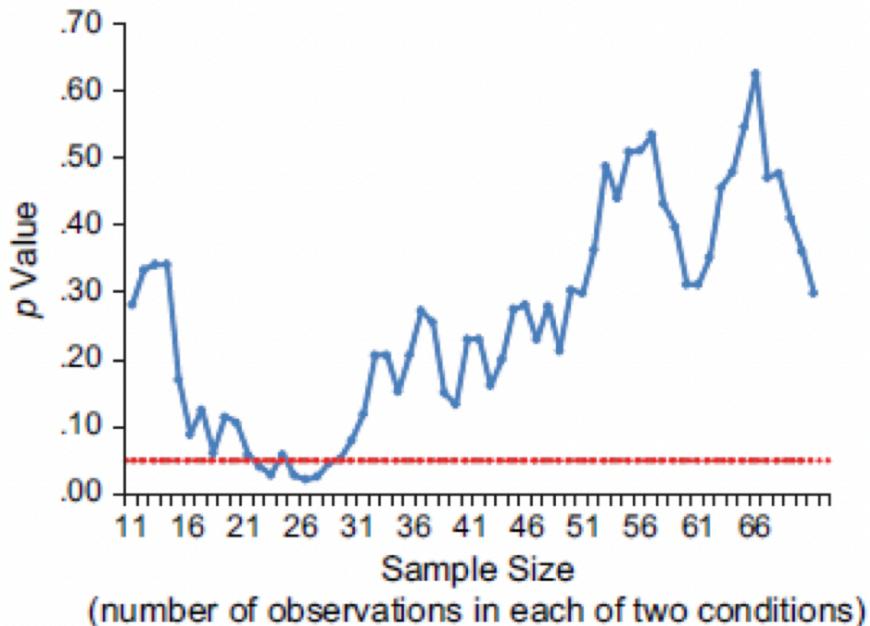


Fig. 2. Illustrative simulation of p values obtained by a researcher who continuously adds an observation to each of two conditions, conducting a t test after each addition. The dotted line highlights the conventional significance criterion of $p \leq .05$.

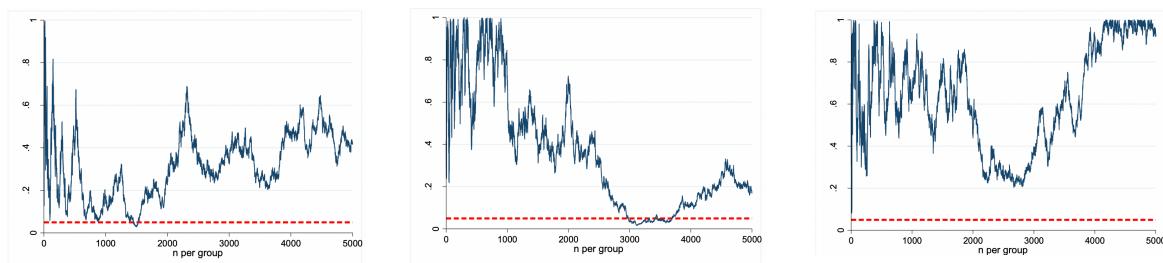
→ p -Werte fluktuieren endlos wenn die H_0 wahr ist,

path 2: significant in 1.4% of all sample sizes

path 3: significant in 13.4% of all sample sizes

path 4: significant in 0% of all sample sizes

p -values are highly volatile if N is small → p -hacking is easier with small N



→ gibt es Evidenz für solche *explorativen Verfahrensweisen* in Bems Studien? - aber ja...

- Experiment 1 prüfte nicht nur erotische Fotos, sondern auch neutrale, positive, negative und romantische aber nicht-erotische Fotos,
- es werden vermutlich unnötige Daten-Transformationen verwendet,
- es wird grundsätzlich auf Geschlechtsunterschiede getestet - explizit ohne theoretischen Grund,
- es gibt Hinweise darauf, dass die Zahl der Teilnehmer post hoc bestimmt wurde,

→ einige Empfehlungen hat Bem in *writing the empirical article* zusammengefasst:

Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, place them aside temporarily and see if any coherent patterns emerge. Go on a fishing expedition for something—anything—interesting. (Bem, 2000, pp. 4–5)

untersuchen Sie aus jedem Blickwinkel - analysieren Sie die Geschlechter getrennt - wenn Daten eine neue Hypothese nahelegen, versuchen Sie, Beweise an anderer Stelle zu finden - wenn Sie schwache Spuren interessanter Muster sehen, versuchen Sie, die Daten neu zu ordnen, um sie deutlicher hervorzuheben - wenn es Teilnehmer gibt, die Sie nicht mögen, oder Studien, Beobachter oder Interviewer, die Ihnen anomale Ergebnisse lieferten, legen Sie sie vorübergehend beiseite und sehen Sie, ob kohärente Muster auftauchen - fischen Sie nach etwas interessantem!

→ anscheinend gilt das *explorative Testen* in der Psychologie als akzeptierte, ja sogar empfohlene Praxis...

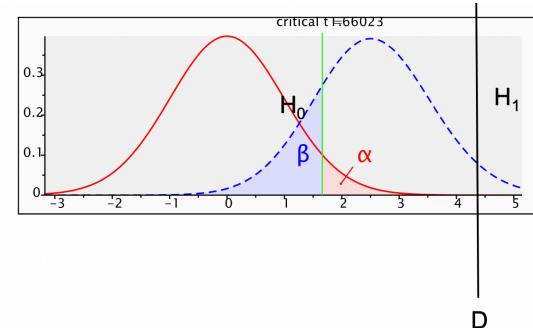
Evidenz ist ein relatives Konzept!

- Sie haben im Lotto gewonnen - kann das Zufall sein?
- eher nicht: die Wahrscheinlichkeit betrug $p = 6.4 \cdot 10^{-8}$
- sollten wir also schließen, dass Sie betrogen haben?
- nein, denn unter jeder Alternativerklärung ist das Ergebnis noch unwahrscheinlicher!
- die Stärke der Evidenz für jede spezifische Hypothese hängt davon ab, wie gut die Daten zu alternativen Hypothesen passen!
- lediglich die H_0 zu betrachten, ist ein logischer Fehler!

$p(D | H_0)$ ist klein - die Daten passen schlecht zur H_0

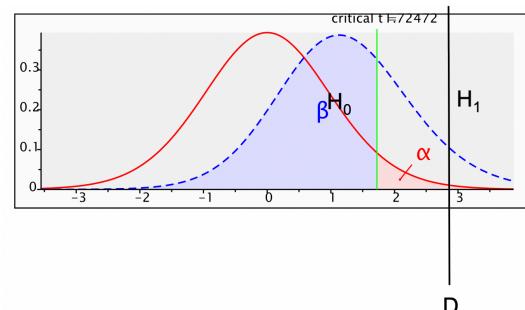
$p(D | H_1)$ ist groß - die Daten passen gut zur H_1

→ starke Evidenz für die H_1



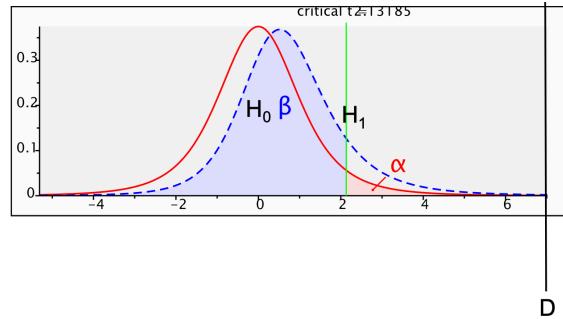
$p(D | H_0)$ ist relativ groß und nicht signifikant
 $p(D | H_1)$ ist groß - bestmögliche Evidenz für die H_1

→ die Daten stärken die H_1 !



die Situation in Bems Studien:

$p(D | H_0)$ ist klein und signifikant,
 $p(D | H_1)$ ist ebenfalls klein,
 → die Daten liefern schwache Evidenz
 für beide Hypothesen!



→ die gleiche Idee, etwas andere Umsetzung → Bayes Factor!

$$p(H | D) = \frac{p(D | H) \cdot p(H)}{p(D)}$$

Posterior

Evidence

Likelihood

Prior

$$\frac{p(H_0 | D)}{p(H_1 | D)} = \frac{p(D | H_0) \cdot p(H_0) / p(D)}{p(D | H_1) \cdot p(H_1) / p(D)} = \frac{p(D | H_0)}{p(D | H_1)} \cdot \frac{p(H_0)}{p(H_1)}$$

$$BF_{01} = \frac{p(D | H_0)}{p(D | H_1)}$$

- der BayesFactor gibt also an, wie stark die Daten unsere Überzeugung verändern sollten,
- der BayesFactor im Bereich von 1 besagt, dass die Daten keinen Informationswert haben,

Bayesian t-Test - Rouder et al., 2009

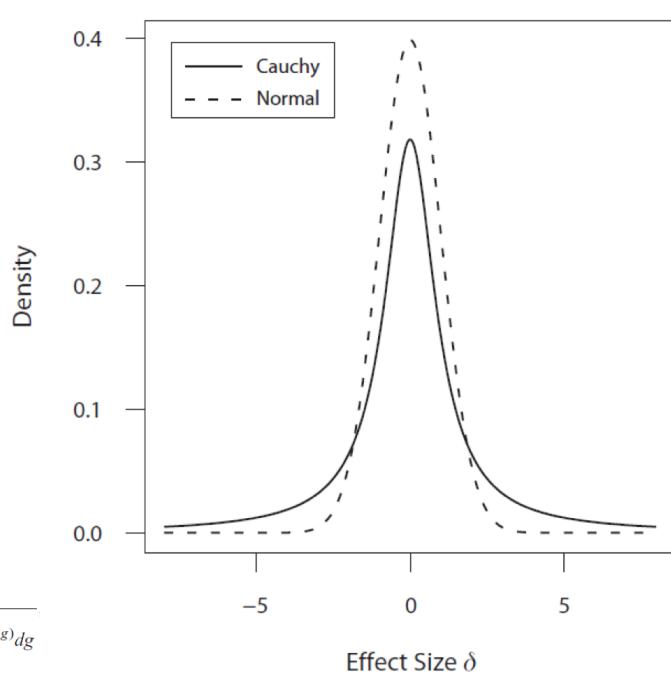
- basiert ausschließlich auf dem BayesFactor
- unabhängig von Prior-Wahrscheinlichkeit für H_0 bzw. M_0 und H_1 bzw. M_1
- betrachtet anstelle einer spezifischen Alternativhypothese eine Verteilung möglicher Effekte,
- wählt diese Verteilung so uninformativ wie möglich,
- <http://pcl.missouri.edu/bayesfactor>

Interpretation:

1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

- **Modell 0:** $\delta = 0$
- **Modell 1:** $\delta \sim \text{Cauchy}$

Prior für Effektstärken in M1:



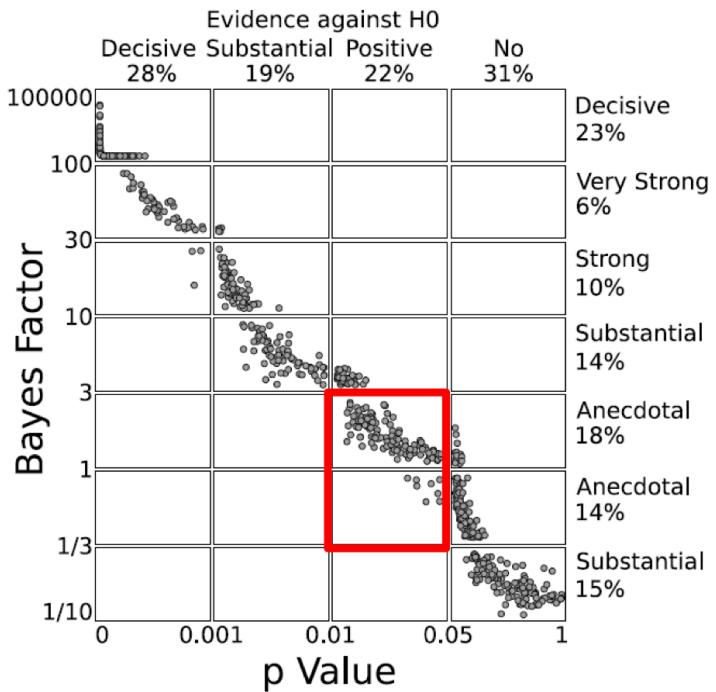
$$B_{01} = \frac{\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}}{\int_0^\infty (1 + Ng)^{-1/2} \left(1 + \frac{t^2}{(1 + Ng)v}\right)^{-(v+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)} dg}$$

→ Ergebnisse für Bems Experimente:

<u>Study</u>	<u>p</u>	<u>BF₀₁</u>	<u>in favor of</u>	<u>strength</u>
1	.01	.61	H1	anecdotal
2	.01	.95	H1	anecdotal
3	.006	.55	H1	anecdotal
4	.02	1.71	H0	anecdotal
5	.01	1.14	H0	anecdotal
6a	.04	3.14	H0	substantial
6b	.04	3.49	H0	substantial
7	.10	7.61	H0	substantial
8	.03	2.11	H0	anecdotal
9	.002	0.17	H1	substantial

→ andere Beispiele:

- Mussweiler (PsychScience, 2006)
 - do stereotypic movements activate the stereotype?
 - Results: $t(18) = 2.09$; $p < .05$, $\text{BF}_{01} = 0.64$
- Dijksterhuis (Science, 2006)
 - is there an unconscious thought effect?
 - Study 1: $t(38) = 2.48$; $p < .05$, $\text{BF}_{01} = 0.34$
- Topolinski (PsychScience, 2011)
 - does dialing a phone number activate the affective valence of emotional words (5683 = love)?
 - Study 2: $t(36) = 2.24$; $p = .03$, $\text{BF}_{01} = 0.79$
- Grider & Malmberg (Memory & Cognition, 2008)
 - are emotional words remembered better than neutral words?
 - Study 2: $t(79) = 2.03$; $p = .03$, $\text{BF}_{01} = 1.56$
- Wetzels et al. (2011)



Zusammenfassung:

- liefert ein signifikantes Ergebnis (z.B. $p = .02$) starke Evidenz gegen die H_0 ?
- die Antwort hängt von den betrachteten Alternativhypotesen ab!
- keine Alternativhypotesen zu betrachten (wie im Signifikanztest üblich), hilft nicht,
- dies führt lediglich dazu, dass wir nicht wissen, ob Evidenz gegen die H_0 vorliegt,
- die Evidenz eines signifikanten Ergebnisses kann schwach und uneindeutig sein - sie kann sogar die H_0 favorisieren,

...aber:

- auch die Verwendung von Bayes-Faktoren ist umstritten,
- zentrales Gegenargument: der Schwerpunkt der Datenanalyse sollte generell auf Parameterschätzungen liegen und nicht auf Hypothesentests,

eine Zusammenfassung zum Signifikanztest

- ein nicht-signifikantes Ergebnis bedeutet idR nicht, dass die H_0 wahrscheinlich korrekt ist,
- ein signifikantes Ergebnis bedeutet nicht notwendigerweise, dass die Forschungshypothese wahrscheinlich korrekt ist,
- ein signifikantes Ergebnis bedeutet nicht notwendigerweise, dass Evidenz für die Forschungshypothese vorliegt,
- mit dem Signifikanztest kann alles belegt werden,
- selbst eine korrekte H_1 ist u.U. nur schwache Evidenz für eine Theorie,

eine Zusammenfassung zur Bayesianischen Statistik

- Bayes-Statistik sagt uns, wie überzeugt wir von Hypothesen sein sollten,
- Bayes-Statistik sagt uns, welchen Informationswert Daten haben,
- Bayes-Statistik erlaubt uns, Evidenz symmetrisch zu interpretieren,
- Bayes-Statistik ist robuster gegenüber Publication Biases, Data Peeking und geistig armem Betrug,
- kurz: **Bayes-Statistik ist, was wir wollen!**

15.Dezember - Tutorium

1. eine Studie hat ein Medikament getestet und dafür ein $r = 0.15$ gefunden - der Effekt ist nicht signifikant geworden ($H_0 \rightarrow$ kein Effekt) mit einem empirischen p-Wert von $p = 0.12$ -

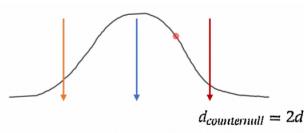
erstelle drei BESD - binomial effect size display: eins für die gefundene Effektstärke, eins für die H_0 und eins für den Counter-Null-Wert!

H_0		
gut	schlecht	
50	50	100
50	50	100
100	100	200

Effektstärke		
gut	schlecht	
.50 + r/2	42,5	100
42,5	57,5	100
100	100	200

$$r_{counternull} = \frac{2d}{\sqrt{4d^2 + 4}}$$

$$r_{counternull} = \sqrt{\frac{4r^2}{1 + 3r^2}}$$

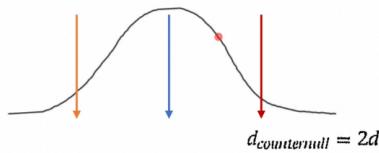


Counter-Null ($r = 0,293$)

Counter-Null ($r = 0,293$)		
gut	schlecht	
50	50	100
50	50	100
100	100	200

$$r_{counternull} = \frac{2d}{\sqrt{4d^2 + 4}}$$

$$r_{counternull} = \sqrt{\frac{4r^2}{1 + 3r^2}}$$



Counter-Null ($r = 0,293$)

Counter-Null ($r = 0,293$)		
gut	schlecht	
		100
		100
100	100	200

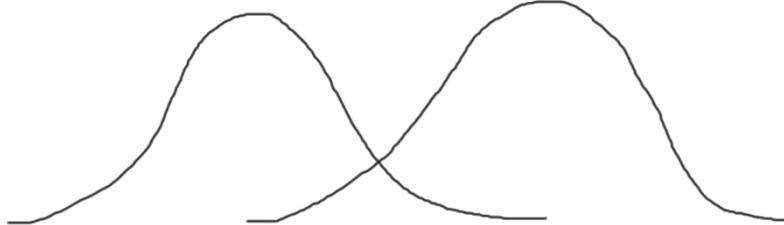
2. zeichne den Bereich in der Abbildung ein:

d

d/2

$$\Phi\left(\frac{-|d|}{2}\right)$$

Aufgabe c) lässt sich eigentlich nicht 1:1 auf die Abbildung übertragen - woran liegt das?



3. beschreibe die Beziehung zwischen p-Wert und Cohen's U₃

- beide Werte beschreiben den Flächeninhalt einer Stichprobenkennwerteverteilung,

- U_3 ist ein Anteil an der Standardnormalverteilung und der p-Wert (im Fall des t-Tests) ein Anteil an der t-Verteilung,
- mit steigender Stichprobengröße nähert sich die Form der t-Verteilung an die Standardnormalverteilung an,
- im Fall einer unendlich großen Stichprobe sind der empirische p-Wert und der U_3 -Wert identisch - d.h. der p-Wert ist der Flächenanteil der Dichtefunktion der Standardnormalverteilung und U_3 der Wert der y-Achse in der kumulierten Verteilungsfunktion (deren y-Achse den kumulierten Flächeninhalt der Dichtefunktion abbildet),

Konfidenzintervalle

APA Publication Manual (2001) - *because confidence intervals combine information on location and precision ... they are, in general, the best reporting strategy - the use of confidence intervals is therefore strongly recommended,*

→ weil Konfidenzintervalle Informationen über Standort und Genauigkeit kombinieren ... sie im Allgemeinen die beste Berichtsstrategie sind, wird die Verwendung von Konfidenzintervallen daher dringend empfohlen,

APA Task Force on Statistical Inference (1999) - *in all figures, include graphical representations of interval estimates whenever possible,*

→ alle Abbildungen sollten nach Möglichkeit grafische Darstellungen von Intervallschätzungen einschließen,

Punktschätzung von Populationsparametern

→ was ist die beste Schätzung eines Populationskennwerts aufgrund eines Stichprobenergebnisses?

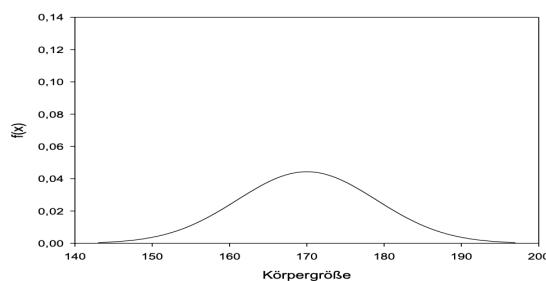
- wichtiges Kriterium: Erwartungstreue
- ein Stichprobenkennwert ist ein erwartungstreuer Schätzer, wenn der Erwartungswert der Stichprobenkennwerteveerteilung dem gesuchten Populationskennwert entspricht,
- Beispiel: der Stichprobenmittelwert!

$$E(x_{\text{quer}}) = \mu$$

! der Stichprobenmittelwert ist also die beste Schätzung für den Populationsmittelwert!

Verteilungsarten in der Inferenzstatistik

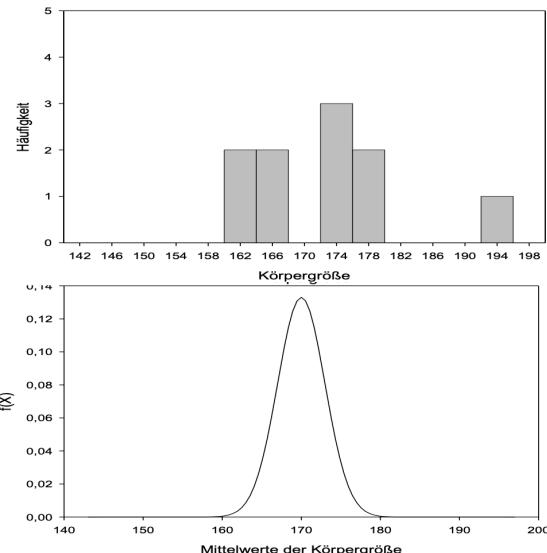
1. Populationsverteilung - oftmals eine Normalverteilung



2. Häufigkeitsverteilung der Daten in einer Stichprobe - weicht zufällig von der Populationsverteilung ab!

3. Stichprobenkennwerteverteilung von Mittelwerten
 - Form: bei großem n stets normal
 - Mittelwert = Mittelwert der Populationsverteilung
 - Varianz: kleiner als Varianz der Populationsverteilung

$$\rightarrow \hat{\sigma}^2 = \sigma^2 / n$$



Punktschätzung von Populationsparametern

- ! die Stichprobenvarianz ist kein erwartungstreuer Schätzer der Populationsvarianz!
 - erwartungstreuer Schätzer der Populationsvarianz:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{n}{n-1} \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

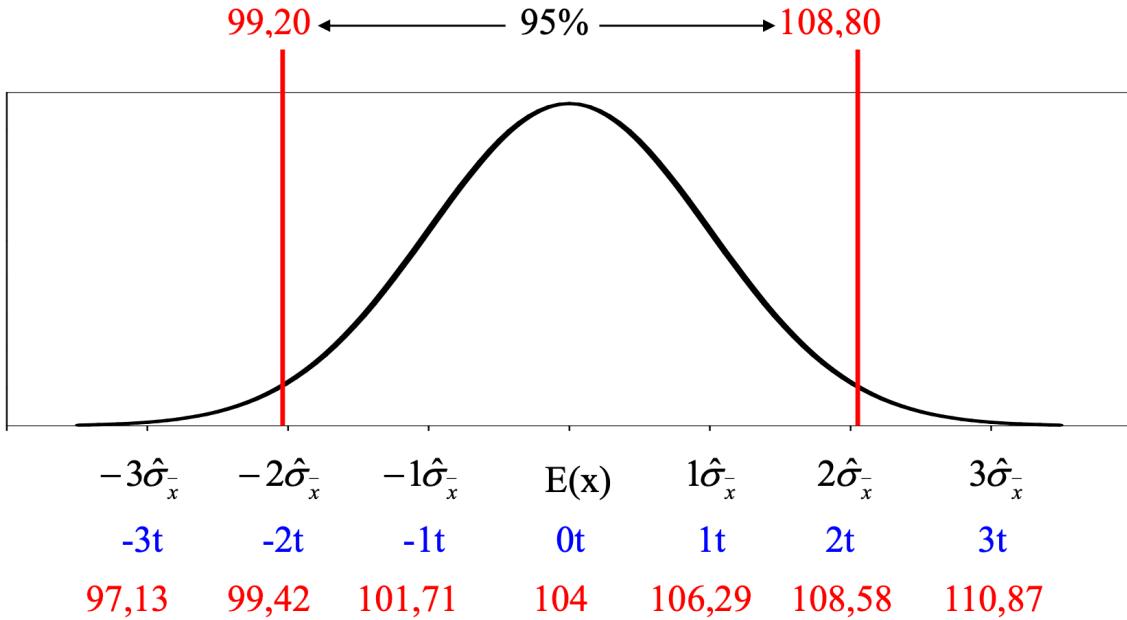
- erwartungstreuer Schätzer des Standardfehlers des Mittelwerts:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} = \frac{s}{\sqrt{n-1}}$$

für eine Herleitung – siehe Sedlmeier & Renkewitz, S.327 f

Intervallschätzung von Populationsparametern - Konfidenzintervalle

- Beispiel:
 ein standardisierter Schulleistungstest zur Messung der Mathematikkenntnisse in der 5.Klasse hat einen Mittelwert von 100 und eine Standardabweichung von 15 - eine Gruppe von Pädagogen ist überzeugt, mit einem neuartigen Förderprogramm die durchschnittlichen Mathematikkenntnisse verbessern zu können - tatsächlich erzielt eine Stichprobe von 20 zufällig ausgewählten Kindern, die an dem Förderprogramm teilgenommen haben, in dem Mathematiktest einen Mittelwert von 104 - die Standardabweichung in der Stichprobe beträgt 10 - in welchem Bereich liegt der Populationsmittelwert der geförderten Kinder mit 95%-iger Sicherheit?
- zur Bestimmung eines Konfidenzintervalls wird zunächst ein (passender) Stichprobenkennwert (Mittelwert, Anteil...) ermittelt,
- ein X%-Konfidenzintervall überdeckt die mittleren x% der entsprechenden Stichprobenkennwerteverteilung,
- Beispiel: 95%-Konfidenzintervall für einen Mittelwert von 104 (s = 10, n = 20),



Berechnung eines x%-Konfidenzintervalls - allgemein

1. Stichprobe ziehen und Kennwert berechnen
 - dieser Stichprobenkennwert wird verwendet, um die beste Punktschätzung für den gesuchten Populationsparameter zu bestimmen - um diese Punktschätzung wird das KI konstruiert!
2. Stichprobenverteilung des Kennwerts bestimmen
 - Form der Stichprobenverteilung? (t, z, binomial...)
 - Standardfehler SE - Standardabweichung der Stichprobenverteilung
3. diejenigen standardisierten Werte in der Stichprobenverteilung bestimmen, die die mittleren x% der Verteilung begrenzen
 - z.B. Normalverteilung, 95% Konfidenz: $z_{2,5\%} = -1,96$, $z_{97,5\%} = 1,96$
4. standardisierte Werte in ursprüngliche Einheiten umrechnen
 - z.B. Normalverteilung: $z = (x - \bar{x}) / \sigma_x \Rightarrow x = \bar{x} + \sigma_x \cdot z$
 - Margin of Error: $w = \text{standardisierter Wert} \cdot SE$
 - Konfidenzintervall:
 - untere Grenze: Kennwert - w
 - obere Grenze: Kennwert + w

Berechnung eines x%-Konfidenzintervalls für einen Mittelwert

1. Stichprobe ziehen und Kennwert - hier: Mittelwert - berechnen
 - $M = 104$
2. Stichprobenverteilung des Mittelwerts bestimmen
 - Form der Stichprobenverteilung?
 - großes N: Normalverteilung
 - kleines N: t-Verteilung - praktisch kann immer t benutzt werden - die t-Verteilung geht bei großem N in die Normalverteilung über,
3. diejenigen t-Werte bestimmen, die die mittleren x% der Verteilung begrenzen
 - df = $n - 1 = 19$
 - $t_{2,5\%} = -2,093$
 - $t_{97,5\%} = 2,093$

4. t-Werte in ursprüngliche Einheiten umrechnen

$$\rightarrow t = (x - \bar{x}) / \hat{\sigma}_x \Rightarrow x = \bar{x} + \hat{\sigma}_x \cdot t$$

$$\rightarrow \text{untere Grenze: } 104 - 2,29 \cdot 2,093 = 99,20$$

$$\rightarrow \text{obere Grenze: } 104 + 2,29 \cdot 2,093 = 108,80$$

! das Intervall zwischen 99,20 und 108,80 überdeckt den Populationsmittelwert der geförderten Kinder mit 95%-iger Sicherheit,

Berechnung eines x%-Konfidenzintervalls für einen Anteil

in einer Wahlumfrage geben von 1200 zufällig ausgewählten Befragten 78 an, dass sie die FDP wählen wollen - in welchem Bereich liegt der Anteil der FDP-Wähler in der Population mit 95%-iger Sicherheit?

1. Stichprobe ziehen und Kennwert berechnen: 78 von 1200 - 6,5%

2. Stichprobenverteilung des Anteils bestimmen

→ Form der Stichprobenverteilung?

- Binomialverteilung

- geht bei großem N in Normalverteilung über

→ Faustregel: Normalverteilung anwendbar, wenn: $np(1-p) > 9$

$$\rightarrow \text{Standardfehler: } SE = \sqrt{np(1-p)} \quad SE = \sqrt{1200 \cdot 0,065 \cdot 0,935} = 8,54$$

3. diejenigen z-Werte bestimmen, die die mittleren x% der Standardnormalverteilung begrenzen:

$$z_{2,5\%} = -1,96$$

$$z_{97,5\%} = 1,96$$

4. z-Werte in ursprüngliche Einheiten umrechnen:

$$\text{untere Grenze: } 78 - 8,54 \cdot 1,96 = 61,26 - 5,1\%$$

$$\text{obere Grenze: } 78 + 8,54 \cdot 1,96 = 94,74 - 7,9\%$$

das Intervall zwischen 5,1% und 7,9% beinhaltet den Anteil der FDP-Wähler in der Population mit 95%-iger Sicherheit,

Berechnung eines x%-Konfidenzintervalls für die Effektgröße d

in einer Untersuchung zur Wirkung von verteiltem Lernen mit $n_1 = n_2 = 10$ wurde eine Effektgröße von $d = 0,5$ beobachtet - in welchem Bereich liegt die Effektgröße in der Population mit 95%-iger Sicherheit?

1. Stichprobe ziehen und Kennwert berechnen: $d = 0,5$

2. Stichprobenverteilung der Effektgröße d (Borenstein) bestimmen

→ Form der Stichprobenverteilung?

- t-Verteilung mit $df = N - 2$

→ Standardfehler:

$$se = \sqrt{\left(\frac{n_1 + n_2}{n_1 \cdot n_2} + \frac{d^2}{2(n_1 + n_2)} \right)}$$

3. diejenigen t-Werte bestimmen, die die mittleren x% der Standardnormalverteilung begrenzen

$$t_{2,5\%} = -2,101$$

$$t_{97,5\%} = 2,101$$

4. t-Werte in ursprüngliche Einheiten umrechnen

$$t = \frac{d - d_{emp}}{se}$$

$$d = d_{emp} + t \cdot se$$

- untere Grenze: $0,5 - 21,01 \cdot 0,45 = -0,45$
- obere Grenze: $0,5 + 21,01 \cdot 0,45 = 1,45$
! das Intervall zwischen -0,45 und 1,45 beinhaltet die Effektgröße in der Population mit 95%-iger Sicherheit

zentrale und non-zentrale t-Verteilungen

- die Berechnung des KIs für d über die zentrale t-Verteilung liefert lediglich eine **approximative** Lösung
- tatsächlich folgt die Stichprobenverteilung von d im Fall von $d \neq 0$ einer non-zentralen t-Verteilung,
- non-zentrale t-Verteilungen sind asymmetrisch,
- nonzentralitäts-Parameter Δ
 - im Fall einer Stichprobe:

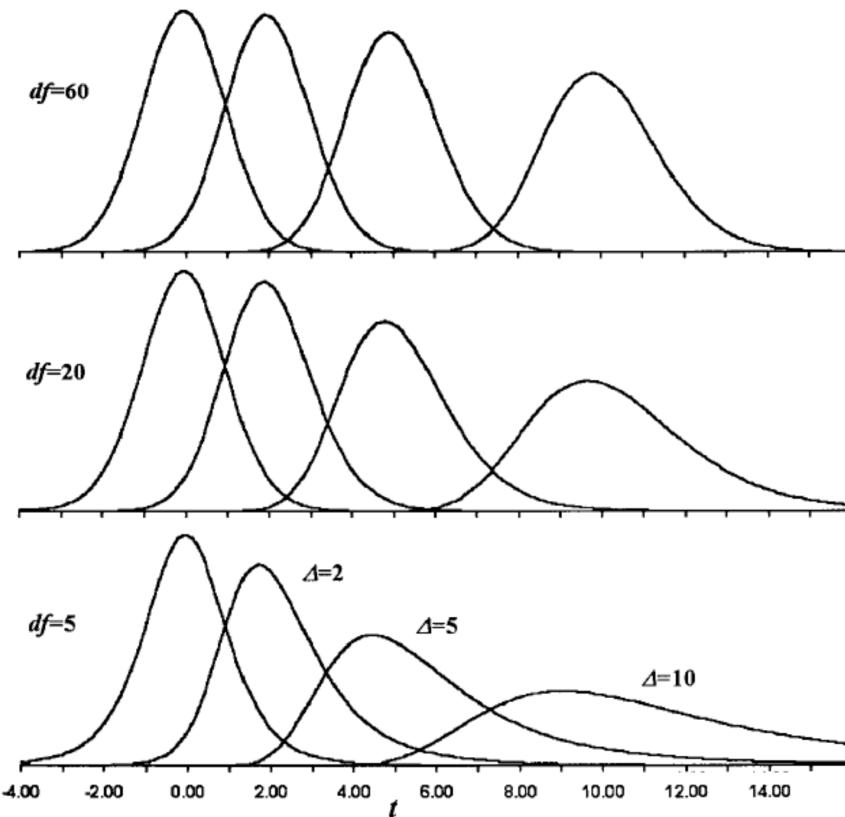
$$\Delta = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}$$

- zudem

$$d = \frac{\mu - \mu_0}{\sigma} \text{ und also } \Delta = d\sqrt{n}$$

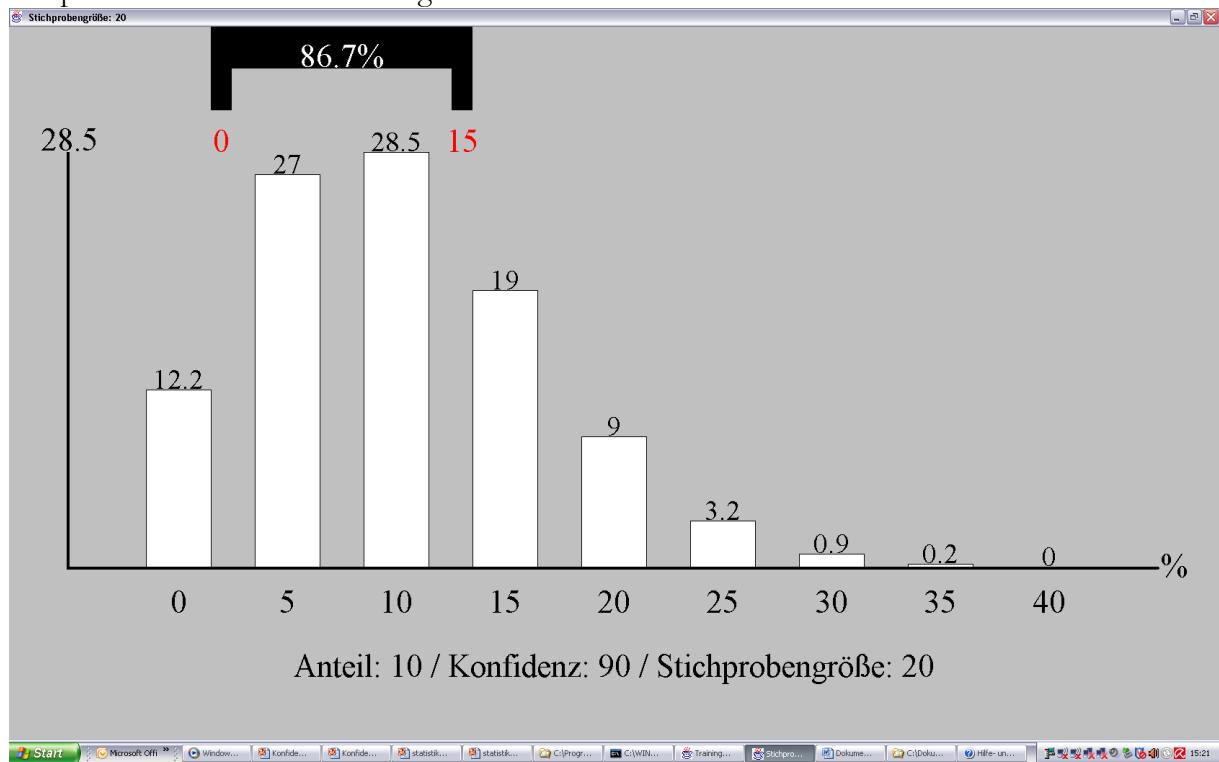
- die Asymmetrie der non-zentralen t-Verteilungen steigt mit Δ ,
- die Asymmetrie der non-zentralen t-Verteilungen sinkt mit den Freiheitsgraden - df

non-zentrale t-Verteilungen



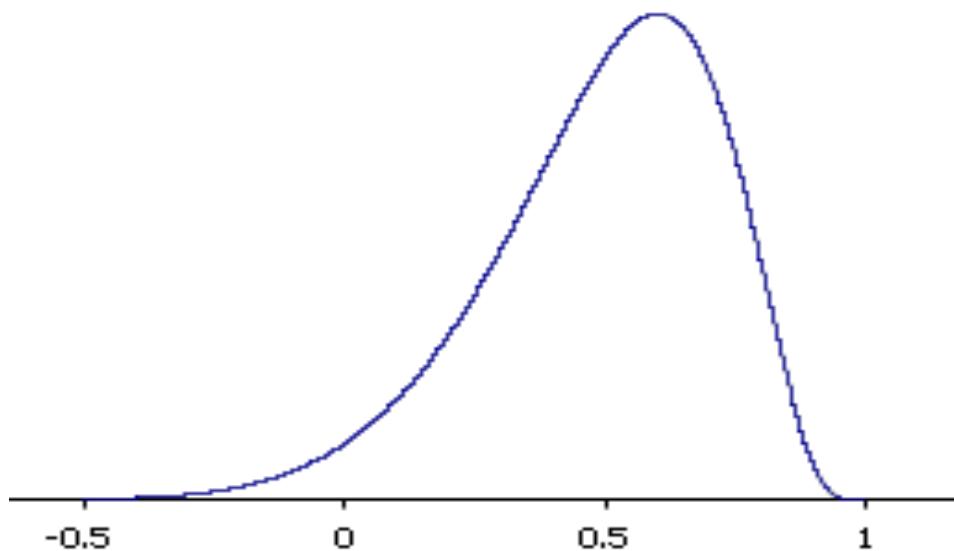
Konfidenzintervall für einen Anteil: N = 20; Stichprobenergebnis P = 10%

entsprechende Binomialverteilung:



- ! bei asymmetrischen Stichprobenverteilungen kann die Berechnung von Konfidenzintervallen nicht auf Basis des Standardfehlers erfolgen!

Stichprobenverteilung des Korrelationskoeffizienten r bei $\rho = 0,6$ und $N = 12$:



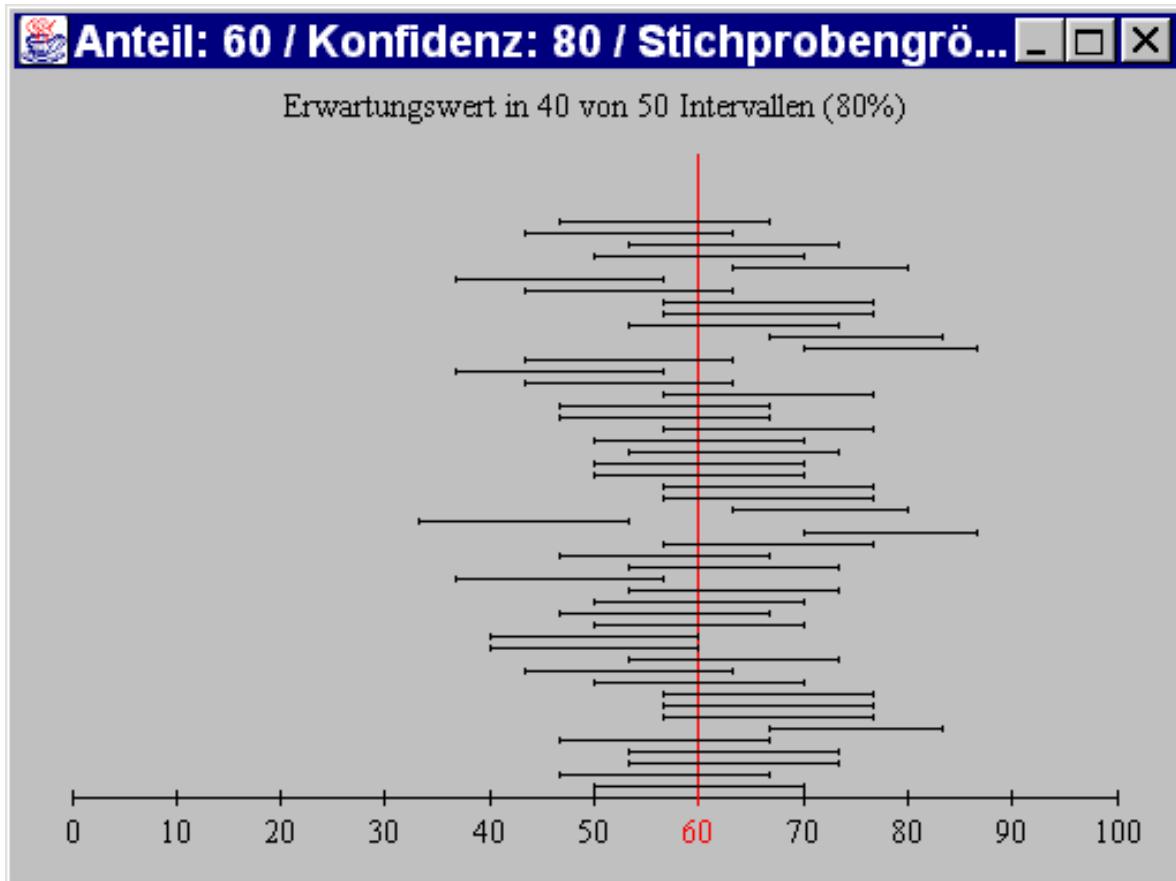
- ! die Folge sind asymmetrische Konfidenzintervalle!

Interpretation von Konfidenzintervallen

- Problem: gemäß der **frequentistischen Wahrscheinlichkeitskonzeption** sind Wahrscheinlichkeiten relative Häufigkeiten
 - 50% aller Münzwürfe ergeben Kopf

- Konfidenzintervalle beziehen sich auf Populationsparameter
- es gibt jeweils nur **einen** Populationsparameter!
 - 95% der Populationsparameter liegen im Konfidenzintervall??? → offensichtlich unsinnig!
- wo ist also die relative Häufigkeit, die der Wahrscheinlichkeitsaussage zugrunde liegt?

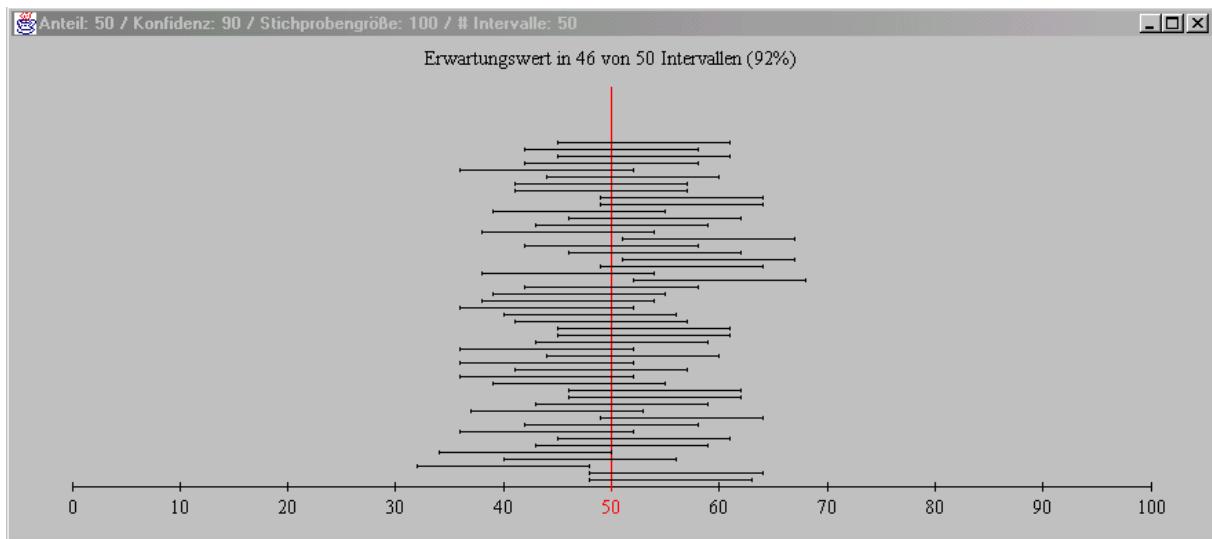
nehmen wir an, dass der Stimmenanteil der CSU in dörflichen Gebieten Bayerns bei 60% liegt - durch insgesamt 50 Umfragen mit je 30 zufällig ausgewählten Teilnehmern soll dieser Anteil ermittelt werden - bei jeder Umfrage wird ein 80%-Konfidenzintervall berechnet - bei wie vielen Umfragen kann man erwarten, dass das Konfidenzintervall den Wert 60% einschließt?



- Illustration der korrekten Interpretation von Konfidenzintervallen - die Population bestand aus Wahlberechtigten, bei denen CSU-Wähler einen Anteil von 60% ausmachten - in 50 Zufallsstichproben wurden jeweils 30 Wahlberechtigte gezogen - aufgrund der so ermittelten Anteile von CSU-Wählern wurden 80%-Konfidenzintervalle berechnet - in dem dargestellten Simulationsergebnis schließen 40 der 50 Konfidenzintervalle den Wert 60% ein,

Konfidenzintervall: Interpretation

- die Wahrscheinlichkeit, dass ein x%-Konfidenzintervall den wahren Wert überdeckt, ist x% - wenn unendlich oft Stichproben gezogen und daraus Konfidenzintervalle berechnet würden, dann würden x% dieser Konfidenzintervalle den wahren Wert überdecken,
- Beispiel:
fünfzig 90%-Konfidenzintervalle, bei einem tatsächlichen Anteil von 50% und jeweils n = 100



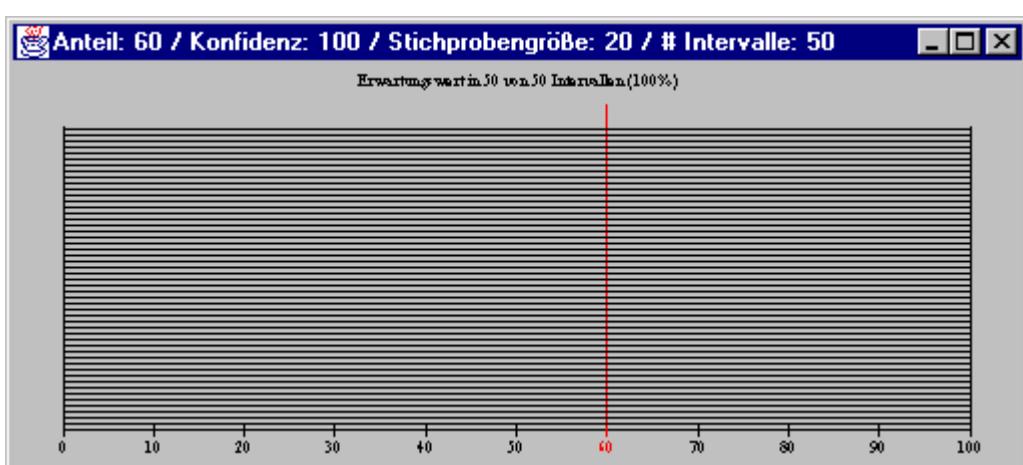
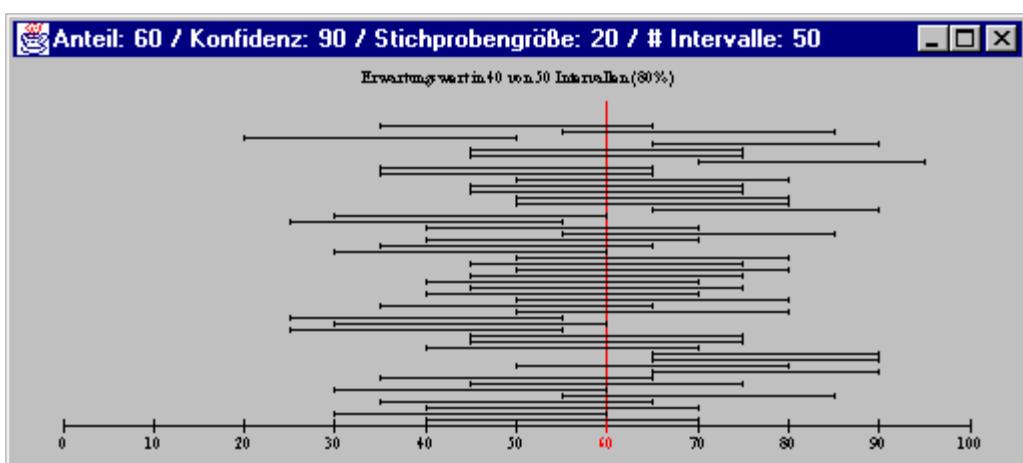
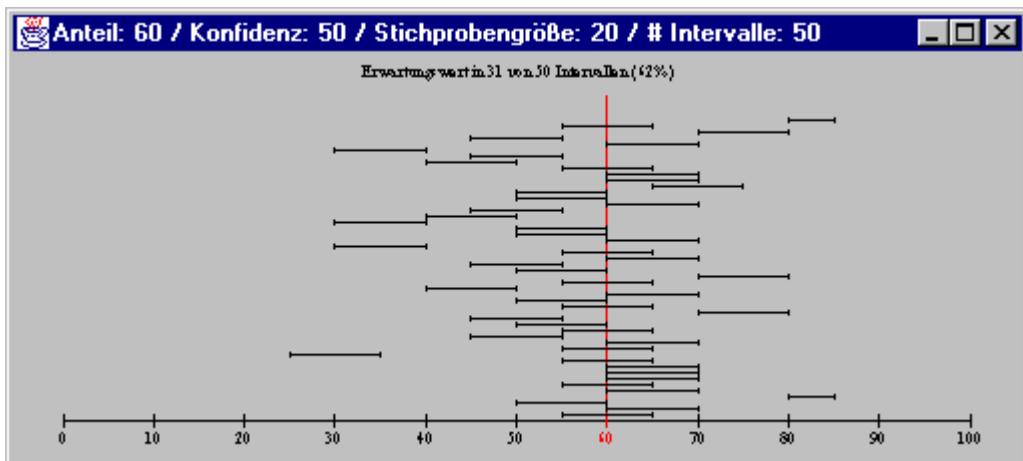
- ! die Wahrscheinlichkeitsaussage bezieht sich also auf die Intervalle, nicht auf den Populationsparameter!
- ! 90% aller 90%-Konfidenzintervalle überdecken den wahren Wert!
- ! allerdings überdecken diese Intervalle jeweils einen anderen Wertebereich
- ! Konfidenzintervalle erlauben daher keine Wahrscheinlichkeitsaussage über den tatsächlichen Wert in der Population!

- ! das Konzept des Konfidenzintervalls wurde von Neyman (1934) entwickelt - er verwendete dabei einen frequentistischen Wahrscheinlichkeitsbegriff und traf folglich ausschließlich Wahrscheinlichkeitsaussagen über die Intervalle,
- ! angezielt war aber eine Aussage über den Populationsparameter
- ! um Aussagen über den Populationsparameter zu treffen, führte Neyman den Begriff *Konfidenz* ein,
- ! die Interpretation eines Konfidenzintervalls folgt damit letztlich eher einem subjektivistischen Wahrscheinlichkeitsbegriff,
 - *ich bin zu 90% sicher, dass der wahre Wert zwischen X und Y liegt*
- ! dies führte zu allen Zeiten zu Konfusion und Unsicherheit bei der Interpretation
- ! der **Konfidenz-Trick**

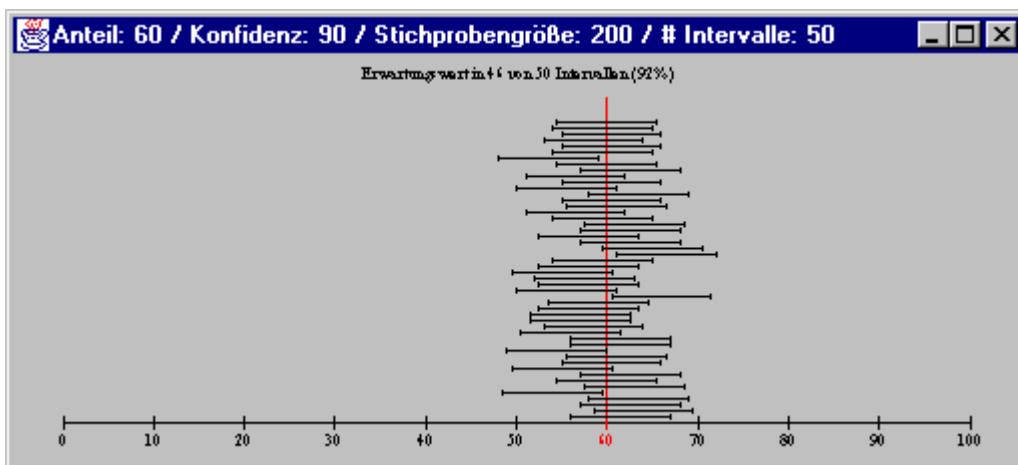
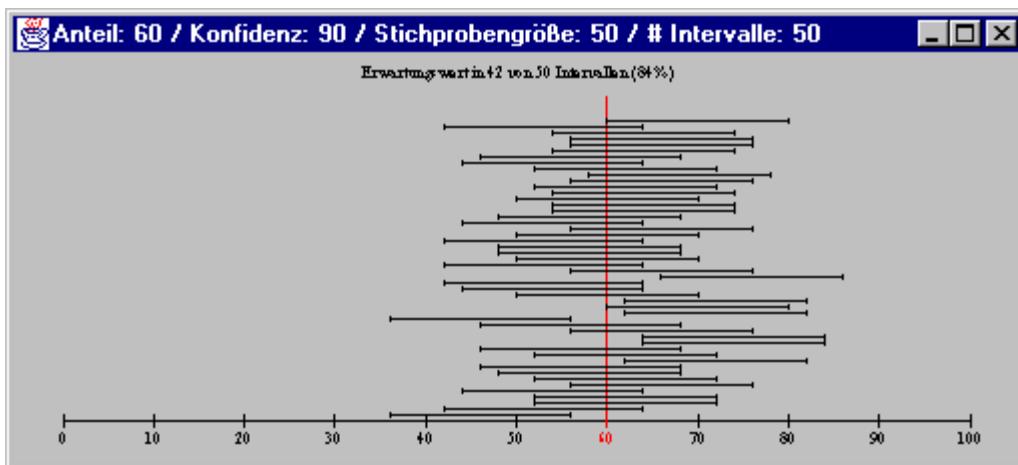
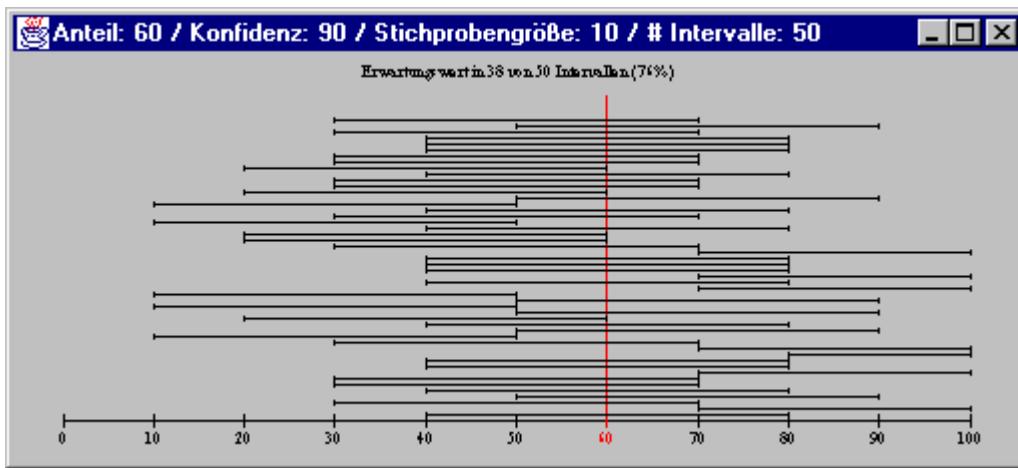
sprachliche Interpretationsalternativen:

- der wahre Wert liegt mit x%-iger Sicherheit zwischen a und b
- wir können zu x% konfident sein, dass das Konfidenzintervall den wahren Wert beinhaltet
- das Konfidenzintervall gibt den Bereich plausibler Werte für den Populationsparameter an - Werte außerhalb des KIs sind unplausibel

- wodurch wird die Größe eines Konfidenzintervalls beeinflusst?
 - ! Höhe der Konfidenz
 - ! Stichprobengröße

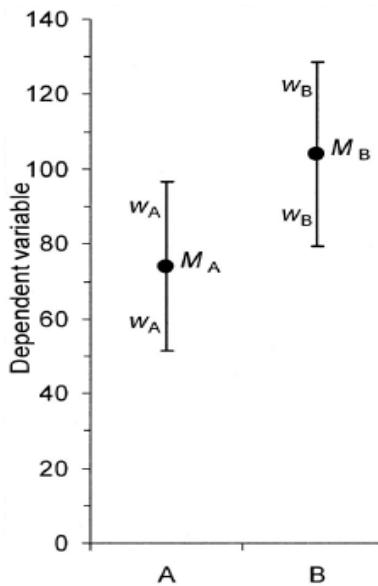


→ der Einfluss der Höhe der Konfidenz (wieviel %) auf die Länge des Konfidenzintervalls für einen Anteil von 60%, n = 20 und Konfidenzen von 50%, 90% und 100%



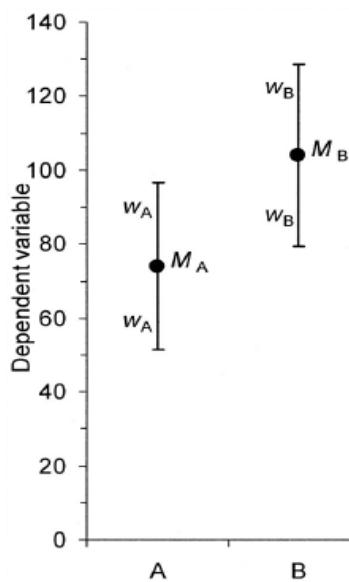
- der Einfluss der Stichprobengröße auf die Länge des Konfidenzintervalls für einen Anteil von 60%, eine Konfidenz von 90% und Stichprobengrößen von 10, 50 und 200
- Grund: Gesetz der Großen Zahlen!

Fehlerplots



Welche Informationen sind zu sehen?

- die deskriptiven Ergebnisse der Studie - hier Mittelwerte - in den ursprünglichen Einheiten!
- Intervallschätzungen der Populationsparameter
 - plausible Annahmen über die Populationsparameter werden erkennbar,
- die Genauigkeit/ Präzision der Studie
 - je kleiner das KI desto präziser/ besser ist die Studie
- das Ergebnis eines Signifikanztests zum Vergleich der Mittelwerte lässt sich ableiten - bei Designs mit unabhängigen Gruppen,
- KIs erlauben die Kombination/ Aggregation der Daten aus mehreren Studien



für eine vollständige und korrekte Interpretation des Plots werden benötigt:

- Angaben über die Höhe der Konfidenz (90%, 95%, SE??)
- Angaben zum Design der Studie

Zusammenhang Konfidenzintervall und Signifikanz im t-Test

$H_0: \mu = 100;$

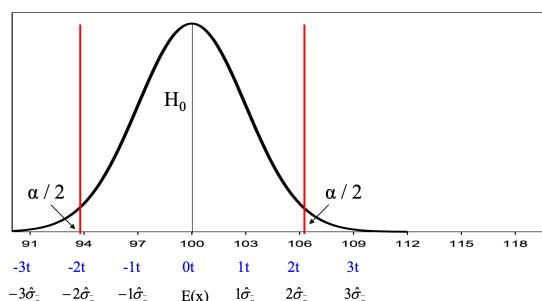
zweiseitiger Ein-Stichproben t-Test mit $\alpha = 5\%$

Studie mit $N = 30$

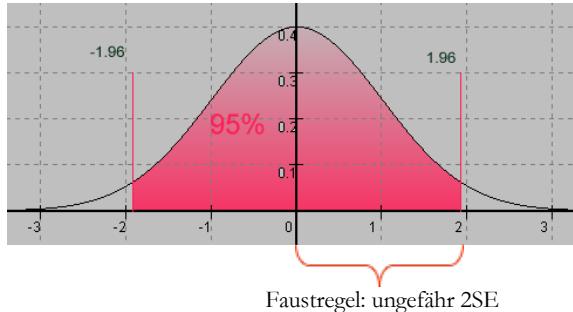
$SE = 3$

$t_{krit} = 2,045$

$M = 106,14$



- wenn ein x% Konfidenzintervall den Wert für H_0 nicht überdeckt, dann ist das Testergebnis signifikant bei einem zweiseitigen $\alpha = 100\% - x\%$
 - z.B: wenn das 95% Konfidenzintervall bei einem Test auf Abweichung von einem vorgegebenen Mittelwert diesen Mittelwert (H_0) nicht überdeckt, dann ist der entsprechende t-Test bei einem zweiseitigen $\alpha = 5\%$ signifikant,



keine Überlappung → Mittelwert im Konfidenzintervall liegt außerhalb des Kriteriums

t-Test für unabhängige Stichproben und Konfidenzintervalle

die H_0 bezieht sich im t-Test für unabhängige Stichproben auf einen Mittelwertsunterschied, das entsprechende Konfidenzintervall bezieht sich also ebenfalls auf einen Mittelwertsunterschied!

Nullhypothese im t-Test:

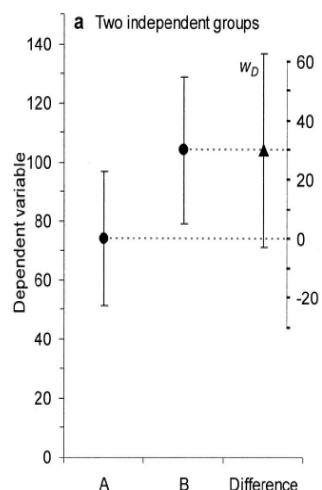
$$\mu_1 - \mu_2 = 0 \text{ bzw. } \mu_1 = \mu_2$$

die relevante Stichprobenkennwerteverteilung ist eine Verteilung von Mittelwertsdifferenzen, Prüfgröße:

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{\hat{\sigma}_{\bar{x}_A - \bar{x}_B}}$$

falls $n_A = n_B$:

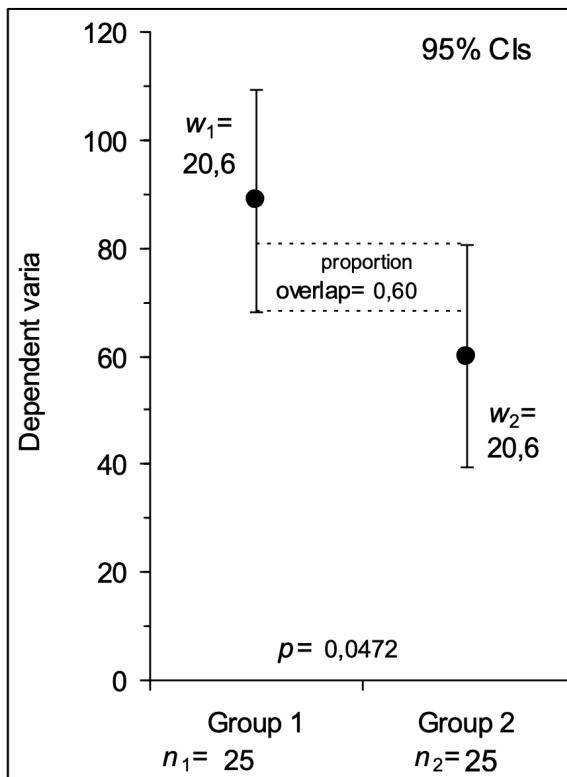
$$\hat{\sigma}_{\bar{X}_A - \bar{X}_B} = \sqrt{\hat{\sigma}_{\bar{X}_A}^2 + \hat{\sigma}_{\bar{X}_B}^2}$$



$$\hat{\sigma}_{\bar{X}_A - \bar{X}_B} = \sqrt{\hat{\sigma}_{\bar{X}_A}^2 + \hat{\sigma}_{\bar{X}_B}^2}$$

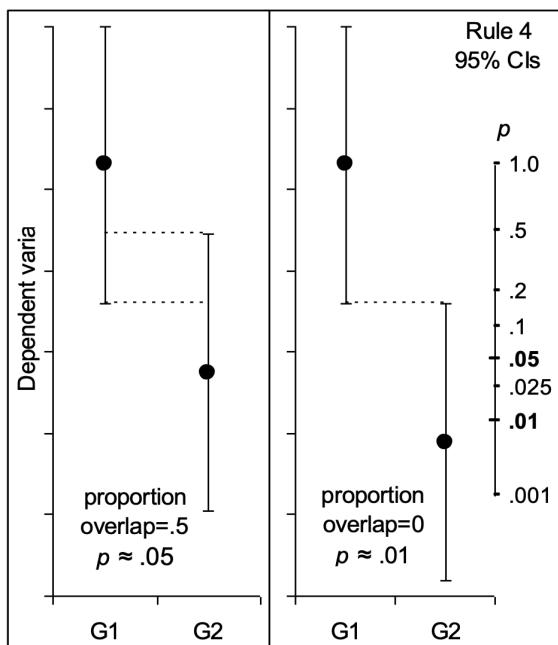
wenn $\hat{\sigma}_{\bar{X}_A}^2 = \hat{\sigma}_{\bar{X}_B}^2$ dann $\hat{\sigma}_{\bar{X}_A - \bar{X}_B} = \sqrt{\hat{\sigma}_{\bar{X}}^2 + \hat{\sigma}_{\bar{X}}^2} = \sqrt{2\hat{\sigma}_{\bar{X}}^2} = 1,41\hat{\sigma}_{\bar{X}}$

der Standardfehler der Mittelwertsdifferenz beträgt also etwa das 1,41fache des Standardfehlers der Mittelwerte!
Breite des Konfidenzintervalls: $2 \cdot SE \cdot t_{krit}$
entsprechend sind also auch der margin of error und das Konfidenzintervall der Mittelwertsdifferenz etwa 1,4mal so breit wie der margin of error und das Konfidenzintervall der Mittelwerte

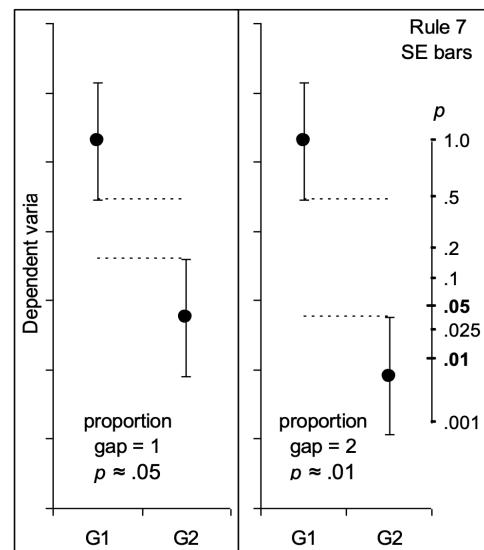


- Mittelwertsdifferenz: 28,84
- margin of error der Mittelwerte: 20,6
- Breite KI der Mittelwerte: 41,2
- Margin of Error
(Mittelwertsdifferenz):
 $20,6 \cdot 1,4 = 28,84$
- $p = 0,047$
- Überlappung:
 $28,84 - 2 \cdot (28,84 - 20,6) = 12,36$
- Anteil Überlappung:
 $12,36 / 20,6 = 0,6$
- wenn $n = n$ und $\hat{\sigma}^2 = \hat{\sigma}^2$, dann ist eine 60%ige Überlappung mit einem signifikanten Ergebnis verbunden,

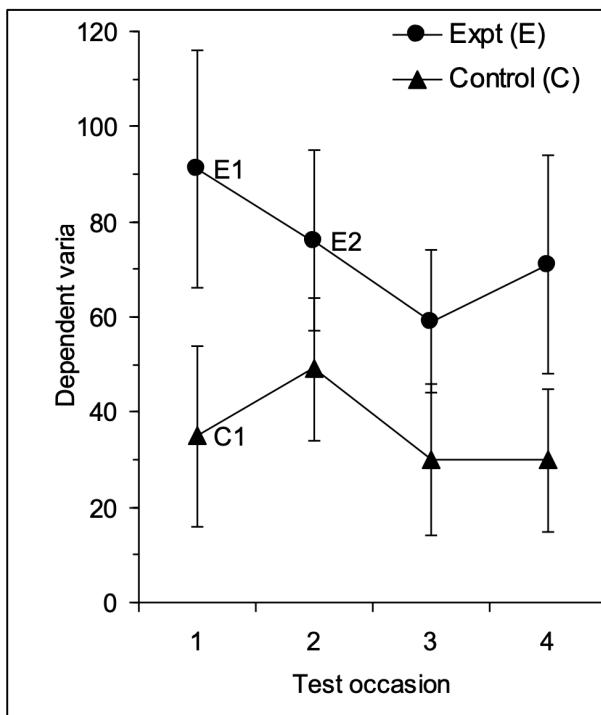
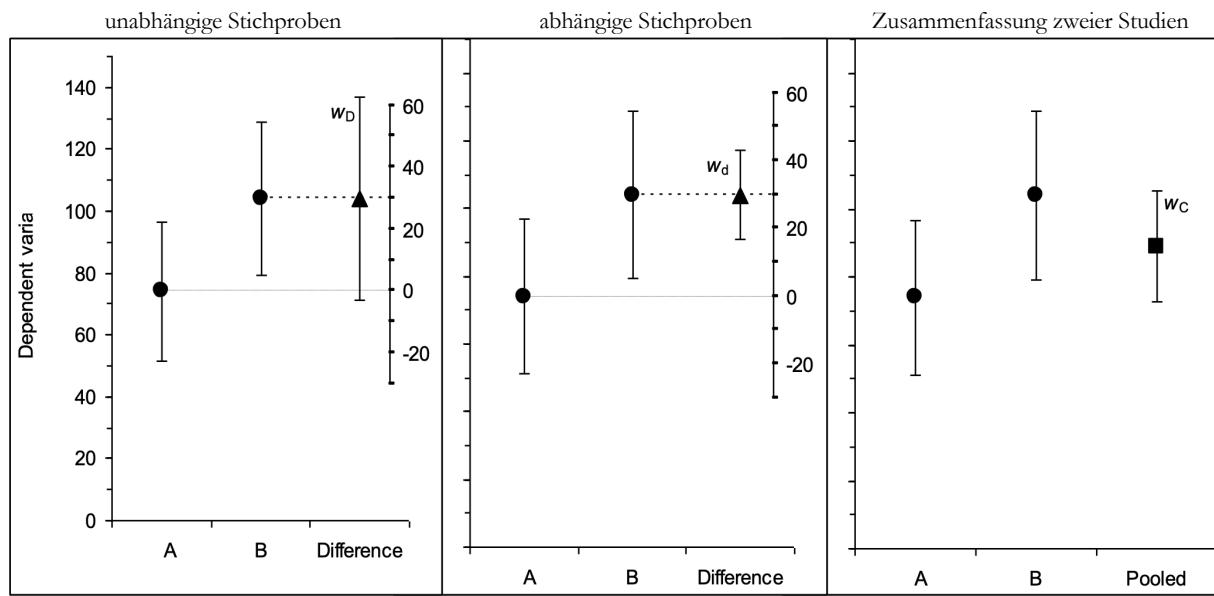
allgemeine Faustregeln - auch bei kleinen und ungleichen n und ungleichen Standardfehlern
- 95% KIs:



allgemeine Faustregeln - auch bei kleinen und ungleichen n und ungleichen Standardfehlern
- SE-Balken:



SE-Balken geben bei $n > 10$ in etwa ein 68% KI an!



ein between-Faktor
ein within-Faktor

für jeden spezifischen Vergleich
(Haupteffekte, Interaktion, simple main effects, Kontraste) könnte und müsste ein eigenes KI angegeben werden!

t-Test für abhängige Stichproben

Nullhypothese:

- $\mu_d = 0$
- der Mittelwert der Differenzen der beiden Messwerte in jedem Wertepaar (μ_d) beträgt in der Population 0
- gleichbedeutend mit: $\mu_1 = \mu_2$

Prüfgröße im t-Test für abhängige Stichproben

$$t = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{x}_d}}$$

t-verteilt mit df = n - 1
n: Anzahl der Wertepaare!

$$\bar{x}_d$$

→ Mittelwert der Differenzen zwischen Wert 1 und Wert 2 in jedem Wertepaar

$$\mu_d$$

→ unter der H_0 erwarteter Mittelwert der Differenzen in der Population

→ zumeist besagt die H_0 , dass der Mittelwert der Differenzen 0 beträgt - also: $\mu_d = 0$

damit vereinfacht sich die Prüfgröße zu:

$$t = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}}$$

$$\hat{\sigma}_{\bar{x}_d}$$

→ geschätzter Standardfehler des Mittelwerts der Differenzen

Standardfehler des Mittelwerts von Differenzen

$$\hat{\sigma}_{\bar{x}_d} = \sqrt{\frac{\hat{\sigma}_d^2}{n}} = \frac{\hat{\sigma}_d}{\sqrt{n}} = \frac{s_d}{\sqrt{n-1}}$$

$$s_d^2 = \frac{1}{n} \sum_{i=1}^n (x_{i,d} - \bar{x}_d)^2 \quad \text{Varianz der Differenzen in der Stichprobe}$$

$$\hat{\sigma}_d^2 = \frac{n}{n-1} s_d^2 \quad \text{beste Schätzung der Varianz der Differenzen in der Population}$$

→ Beispiel: gleiche Werte, unterschiedliche Paarungen

Messung A	Messung B	Diff
27	21	6
25	25	0
30	23	7
29	26	3
30	27	3
33	26	7
31	29	2
35	31	4

$$M_A = 30$$

$$S_A = 2,958$$

$$r_{AB} = .69$$

$$r_{AB} = .00$$

Messung A	Messung B	Diff
27	29	-2
25	25	0
30	21	9
29	27	2
30	23	7
33	26	7
31	31	0
35	26	9

$$M_A = 30$$

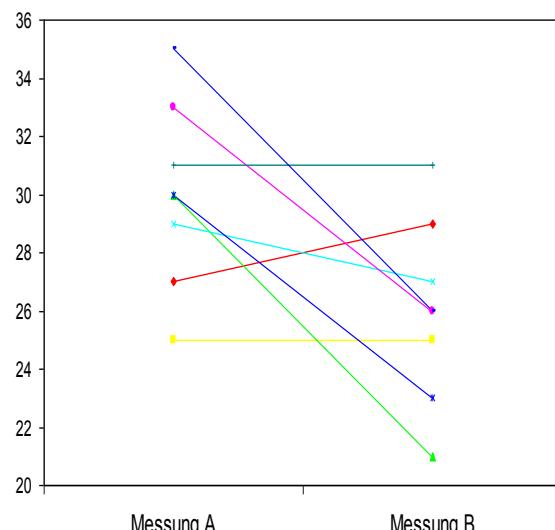
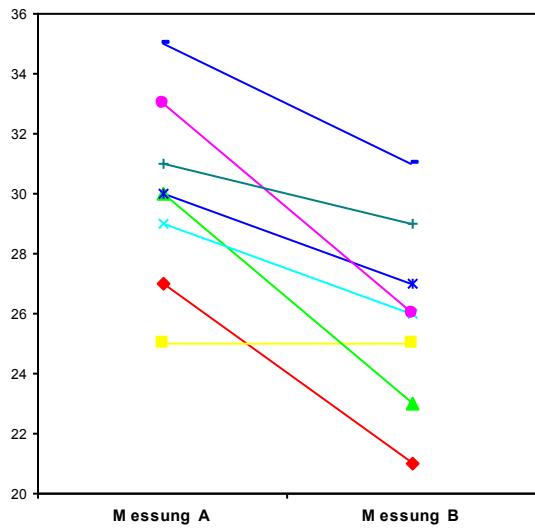
$$S_A = 2,958$$

$$M_B = 26 \\ M_d = 4$$

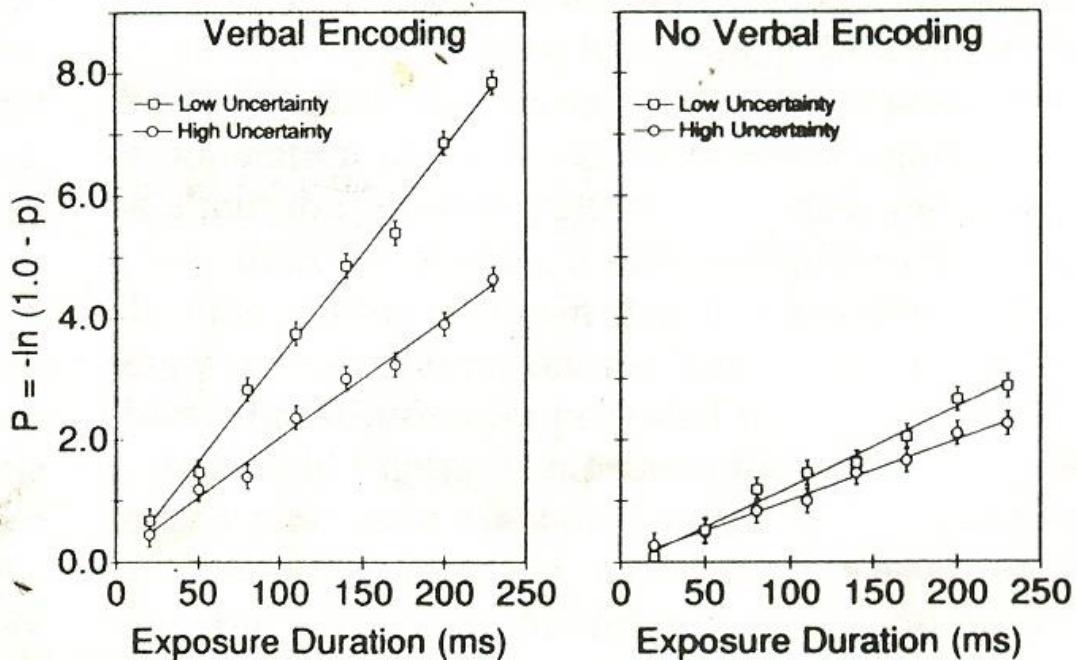
$$S_B = 2,958 \\ S_d = 2,345$$

$$M_B = 26 \\ M_d = 4$$

$$S_B = 2,958 \\ S_d = 4,183$$



- je höher die Messwertreihen korrelieren, desto geringer ist die Standardabweichung der Differenzen,
- bei einer höheren Korrelation ist also auch die Power des Tests größer,
- bei $r_{AB} = 0$ führen der t-Test für abhängige Stichproben und der t-Test für unabhängige Stichproben zum gleichen t-Wert,
- Beispiel: Fehlerplot



Lofus, 1993

- ! alle Effekte sind signifikant!
- ! die Beziehung zwischen Leistung und Dauer entspricht dem vorhergesagten Verlauf!
- ! Enkodierung und Unsicherheit haben die vorhergesagten Effekte
- ! die Schätzung der Populationsmittelwerte ist sehr genau

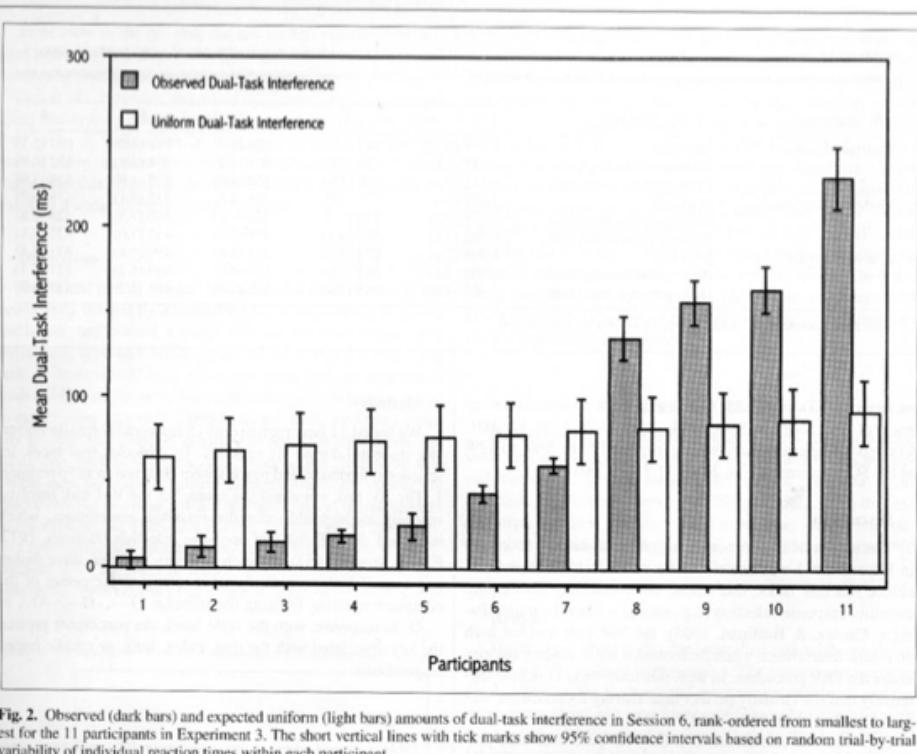


Fig. 2. Observed (dark bars) and expected uniform (light bars) amounts of dual-task interference in Session 6, rank-ordered from smallest to largest for the 11 participants in Experiment 3. The short vertical lines with tick marks show 95% confidence intervals based on random trial-by-trial variability of individual reaction times within each participant.

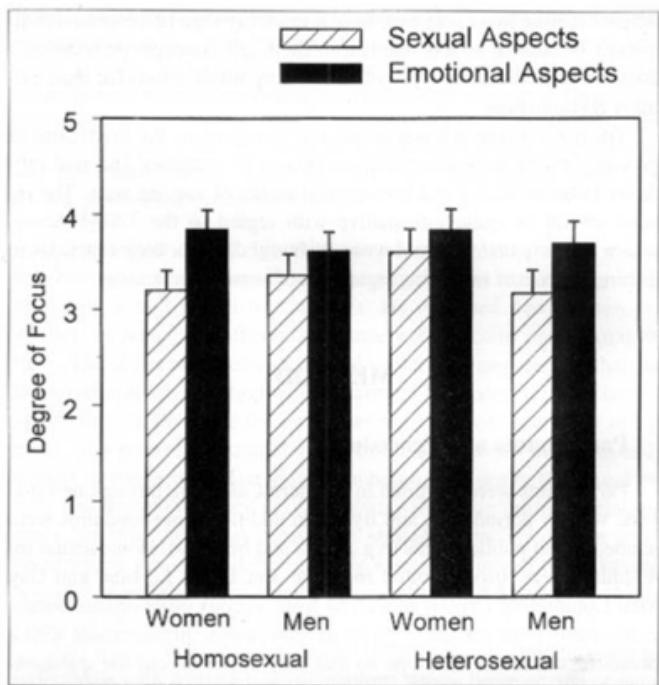


Fig. 2. Reactions to a mate's actual infidelity: Mean degree (1 = *not at all*; 5 = *completely*) to which participants reported focusing on sexual and emotional aspects of the infidelity. Error bars represent standard errors.

Metaanalyse

Fragestellung

- In der Psychologie gibt es zu zahlreichen Fragestellungen hunderte von Studien!
 - Wirkt Psychotherapie?
 - Gibt es Geschlechtsunterschiede bei Persönlichkeitseigenschaften?
 - Sind Menschen sensitiv für Häufigkeiten?
 - Welche Eigenschaften des Modells beeinflussen, ob Modellverhalten nachgeahmt wird?
- Typisches Ergebnis: 70mal war der untersuchte Effekt signifikant, 30 mal nicht!
 - Was sagt uns das bloß?
- Es müsste möglich sein, die Ergebnisse der Studien zusammenzufassen und so eine Wissensakkumulation zu erreichen!
⇒ Dies leistet die Metaanalyse!!!

Was kommt raus?

- Das Hauptergebnis ist eine Schätzung des gesuchten Populationseffekts
⇒ **Gewichtete mittlere Effektstärke!**
 - Die klassische Studie: Smith & Glass (1977) zur Wirkung von Psychotherapie
 - Gewichtete mittlere Effektgröße: $d = 0,68$

Was kommt raus?

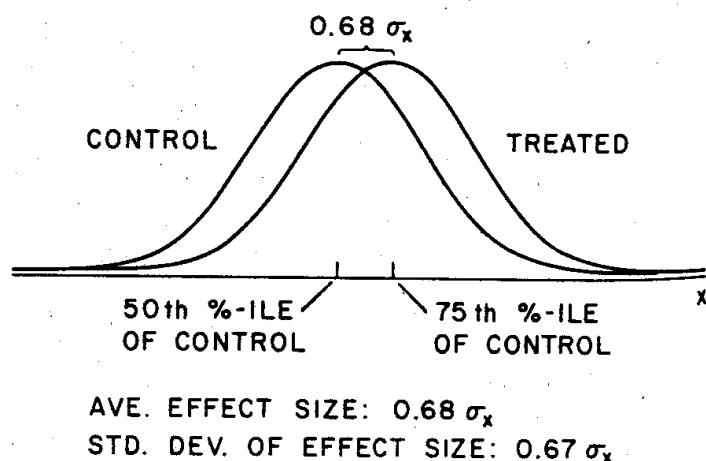


Figure 1. Effect of therapy on any outcome.
(Data based on 375 studies; 833 data points.)

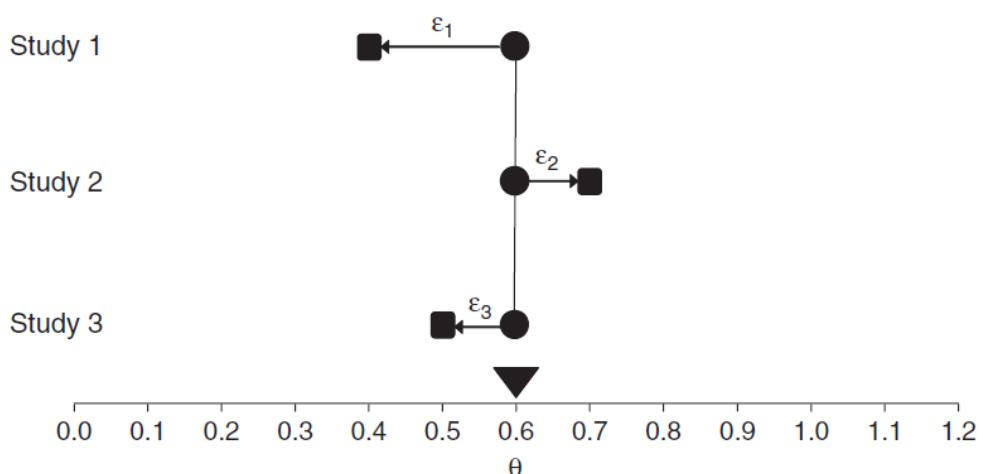
Ausgangsbasis: Empirische Stichprobenverteilungen

- In 60 Studien wurde die gleiche Fragestellung untersucht. In allen Studien $N = 30$
- Verteilung der 60 Effektgrößen:

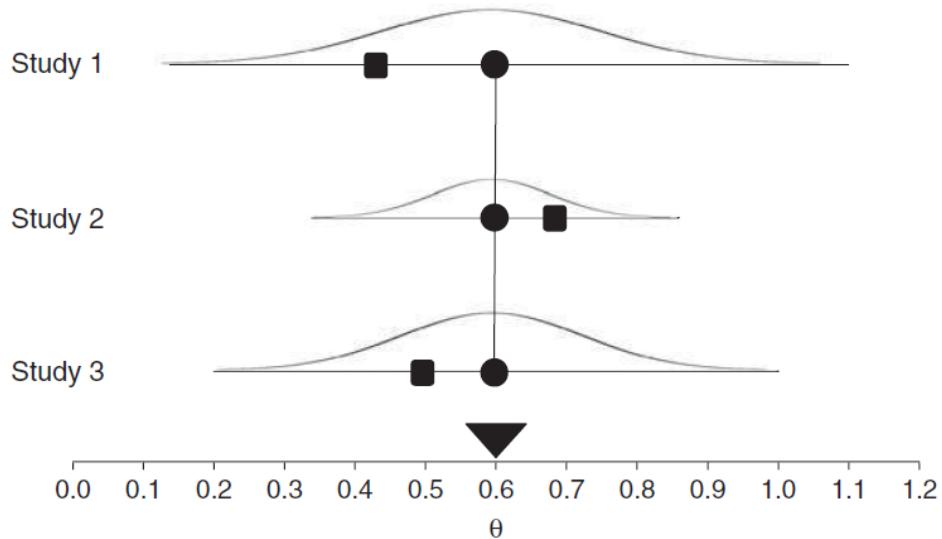
$\alpha = 0,05$, zweiseitig	signifikant	7 0 6 5 5 0,0,1,3,4,5,5,9 4 2,3,3,3,3,4,6,7,8
	nicht signifikant	3 0,0,1,3,3,6,6,7,7,7,8,9 2 0,0,1,3,3,4,7,8,9,9 1 1,1,3,5,6,6,8,9,9 0 4,5,5,8,9 -0 0 -1 0,2,7

- Schätzung des Populationseffekts: $r = 0,33$
- Range der Einzelergebnisse: $r = -,17$ bis $r = ,70$
- Die Variation geht (fast) ausschließlich auf Stichprobenfehler zurück!
- **Die Einzelstudie ist nicht sonderlich aussagekräftig!!!**

Einzelstudie und “wahrer” Effekt



Präzision der Einzelstudie



Was ist zu berücksichtigen?

- **Stichprobengröße**
 - Größere Stichproben liefern präzisere Schätzungen des Populationseffekts \Rightarrow Gesetz der Großen Zahlen!
 - Präzision: SE oder $1/SE$
 - Für den Mittelwert: $SE = \frac{\hat{\sigma}}{\sqrt{N}}$
 - Beispiel: $N = 100$ und $N = 25$: $SE = \frac{10}{\sqrt{100}} = 1$ $SE = \frac{10}{\sqrt{25}} = 2$
 - Die Effektgrößen aus den Einzelstudien werden daher gewichtet mit ihrer Stichprobengröße oder (häufiger und besser) mit ihrer Varianz (SE^2) gemittelt:

$$M = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i}$$

Was ist zu berücksichtigen?

Formeln

- Gewicht der einzelnen Effektgröße:
- Gewichtete mittlere Effektgröße:

$$W_i = \frac{1}{V_{Y_i}}$$

$$M = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i}$$

- Varianz und Standardabweichung der *mittleren* Effektgröße:

$$V_M = \frac{1}{\sum_{i=1}^k W_i} \quad SE_M = \sqrt{V_M}$$

- Test der Nullhypothese $M = 0$:

$$Z = \frac{M}{SE_M}$$

Was ist zu berücksichtigen?

Formeln

- Konfidenzintervall der mittleren Effektgröße:

$$LL_M = M - 1.96 \times SE_M$$

$$UL_M = M + 1.96 \times SE_M$$

Was ist zu berücksichtigen?

Formeln

- Metaanalysen werden (in aller Regel) mit standardisierten Effektgrößen durchgeführt.
- Benötigt werden also die Formeln für die Varianz und Standardfehler der gängigen Effektmaße.
- Varianz und Standardfehler d :

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \quad SE_d = \sqrt{V_d}$$

- Varianz und Standardfehler g :

$$V_g = J^2 \times V_d \quad SE_g = \sqrt{V_g}$$

mit:

$$J = 1 - \frac{3}{4df - 1}$$

Was ist zu berücksichtigen?

Formeln

- Varianz und Standardfehler r :

$$V_r = \frac{(1 - r^2)^2}{n - 1}$$

- Die Varianz der Korrelation ist damit stark vom beobachteten Effekt abhängig (Großer Effekt > Kleine Varianz). Dies soll vermieden werden. Metaanalysen werden daher mit *Fisher-Z-transformierten* Korrelationen durchgeführt.
- Fisher-Z-Transformation:

$$z = 0,5 \times \ln\left(\frac{1+r}{1-r}\right)$$

Was ist zu berücksichtigen?

Formeln

- Varianz und Standardfehler *Fisher-Z-Werte*:

$$V_z = \frac{1}{n - 3} \quad SE_z = \sqrt{V_z}$$

- Rücktransformation:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Was ist zu berücksichtigen?

- Methodische Qualität der Studien
 - Gesonderte Analyse für “hochrangige” Publikationen und “graue Literatur”
 - Inhaltliche Unterschiede
 - Unterschiedliche Operationalisierungen von UV und AV
 - Systematisch unterschiedliche Stichproben
- ⇒ Moderatoranalyse!

Praktisches Vorgehen

- Literaturrecherche
- Auswahl
- Bestimmung einer einheitlichen Effektgröße in *allen* Studien
 - Vergleichbarkeit von Effektgrößen beachten (z.B. Effektgrößen in Studien mit between-subjects und within-subjects Design)
 - Mögliche Abhängigkeiten von Effektgrößen beachten.
 - Korrekturformeln beachten (für Reliabilitätsunterschiede, Varianzeinschränkungen etc.)
- Aggregation der Effektgrößen

Berechnung und Aggregation von Effektstärken

Berechnen Sie den mit n gewichteten Mittelwert von r (gleiche Gruppengrößen) für die folgenden 5 Ergebnisse:

Ergebnisse aus 5 Studien (Mittelwertsunterschiede, 2 Gruppen):

1. Mittelwert_A=1.6, Mittelwert_B=1.4, s_A=s_B=0.4, n=60
2. r=.4, n = 20
3. d=.8, n = 40
4. t(48) = 1.6
5. F(1,36) = 2,28

$$1. d=(1.6-1.4)/0.4=0.5 \Rightarrow r=0.5/\sqrt{4.25}=0.24$$

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

$$3. r=0.8/\sqrt{4.64}=0.37$$

$$4. r=\sqrt{(2.56/50.56)}=0.23$$

$$5. r=\sqrt{(2.28/38.28)}=0.24$$

$$r = \sqrt{\frac{t^2}{t^2 + df}} \text{ und } r = \sqrt{\frac{F}{F + df_{within}}}, \text{ für } F(1, x)$$

$$\begin{aligned} r_{\text{Mittel}} &= (0.24 \times 60 + 0.4 \times 20 + 0.37 \times 40 + 0.23 \times 50 + 0.24 \times 38) / 208 \\ &= (14.4 + 8 + 14.8 + 11.5 + 9.12) / 208 = 0.28 \end{aligned}$$

Berechnung und Aggregation von Effektstärken

Study	Treated			Control		
	Mean	SD	n	Mean	SD	n
Carroll	94	22	60	92	20	60
Grant	98	21	65	92	22	65
Peck	98	28	40	88	26	40
Donat	94	19	200	82	17	200
Stewart	98	21	50	88	22	45
Young	96	21	85	92	22	85

Für die erste Studie:

$$s_{pooled} = \sqrt{\frac{(60-1) \times 22 + (60-1) \times 20}{60+60-2}} = 21,0123 \quad d = \frac{94 - 92}{21,0238} = 0,0951$$

$$V_d = \frac{60+60}{60 \times 60} + \frac{0,0951^2}{2(60+60)} = 0,0334 \quad J = \left(1 - \frac{3}{4 \times 118 - 1}\right) = 0,9936$$

$$g = 0,9936 \times 0,0951 = 0,0945 \quad V_g = 0,9936^2 \times 0,0334 = 0,0329$$

Berechnung und Aggregation von Effektstärken

Study	Effect size	Variance Within	Weight	Calculated quantities		
				Y	V _Y	W
Carroll	0,095	0,033	30,352	2,869	0,271	921,214
Grant	0,277	0,031	32,568	9,033	2,505	1060,682
Peck	0,367	0,050	20,048	7,349	2,694	401,931
Donat	0,664	0,011	95,111	63,190	41,983	9046,013
Stewart	0,462	0,043	23,439	10,824	4,999	549,370
Young	0,185	0,023	42,698	7,906	1,464	1823,115
Sum			244,215	101,171	53,915	13802,325

Gewichtete mittlere Effektstärke:

$$M = \frac{101,171}{244,215} = 0,4143 \quad V_M = \frac{1}{244,215} = 0,0041 \quad SE_M = \sqrt{0,0041} = 0,0640$$

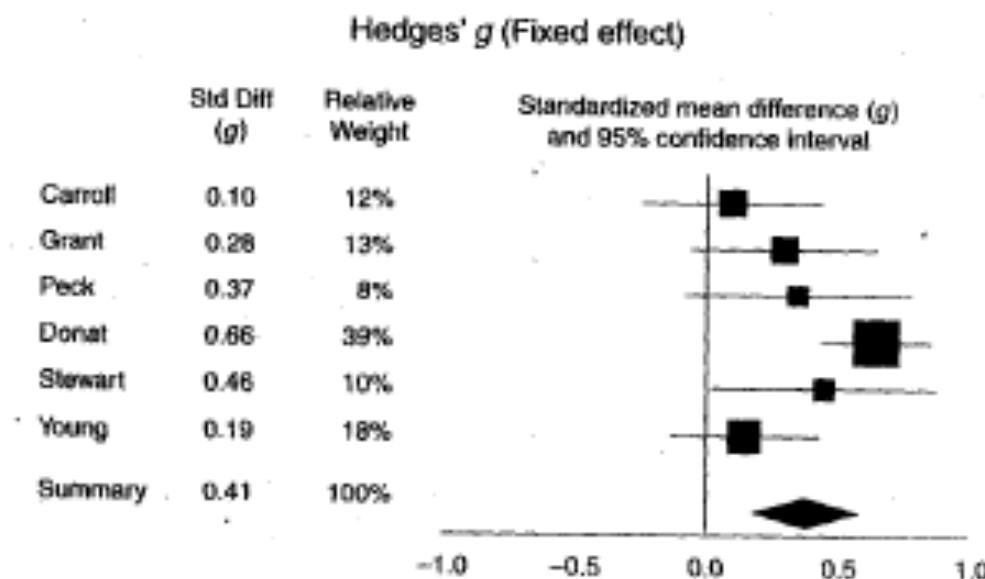
Konfidenzintervall:

$$LL_M = 0,4143 - 1,96 \times 0,0640 = 0,2889$$

$$UL_M = 0,4143 + 1,96 \times 0,0640 = 0,5397$$

$$\text{Signifikanztest: } Z = \frac{0,4143}{0,0640} = 6,4739$$

Forest Plot



Berechnung und Aggregation von Effektstärken

Study	Correlation	N
Fonda	0.50	40
Newman	0.60	90
Grant	0.40	25
Granger	0.20	400
Miland	0.70	60
Finch	0.45	50

Transformation für die erste Studie:

$$z_1 = 0.5 \times \ln\left(\frac{1 + 0.50}{1 - 0.50}\right) = 0.5493, \quad V_1 = \frac{1}{40 - 3} = 0.0270$$

Berechnung und Aggregation von Effektstärken

Study	Effect size \bar{Y}	Variance Within $V_{\bar{Y}}$	Weight W	Calculated quantities		
				$W\bar{Y}$	$W\bar{Y}^2$	W^2
Fonda	0.5493	0.0270	37.000	20.324	11.164	1369.000
Newman	0.6931	0.0115	87.000	60.304	41.799	7569.000
Grant	0.4236	0.0455	22.000	9.320	3.949	484.000
Granger	0.2027	0.0025	397.000	80.485	16.317	157609.000
Milland	0.8673	0.0175	57.000	49.436	42.876	3249.000
Finch	0.4847	0.0213	47.000	22.781	11.042	2209.000
Sum			647.000	242.650	127.147	172489.000

Gewichtete mittlere Effektstärke (Fisher Z):

$$M = \frac{242.650}{647.000} = 0.3750, \quad V_M = \frac{1}{647.000} = 0.0015, \quad SE_M = \sqrt{0.0015} = 0.0393,$$

Konfidenzintervall:

$$LL_M = 0.3750 - 1.96 \times 0.0393 = 0.2980,$$

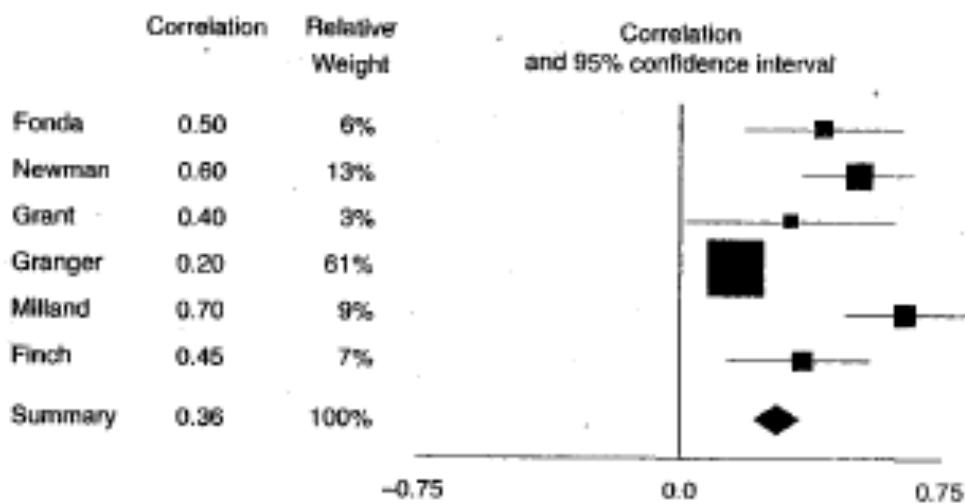
$$UL_M = 0.3750 + 1.96 \times 0.0393 = 0.4521$$

Signifikanztest:

$$Z = \frac{0.3750}{0.0393} = 9.5396$$

$$r = \frac{e^{(2 \times 0.3750)} - 1}{e^{(2 \times 0.3750)} + 1} = 0.3584.$$

Forest Plot



Fixed- und Random-Effects

Bisher: Fixed-Effects-Modell

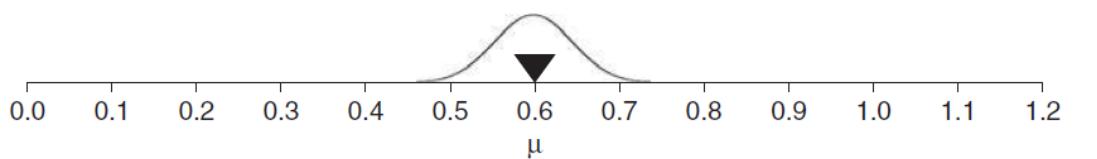
- Alle Studien schätzen denselben wahren Effekt (es gibt also nur einen wahren Effekt).
- Die gewichtete mittlere Effektstärke ist damit die bestmögliche Schätzung des wahren Effekts.

Random-Effects-Modell

- Realistischer ist zumeist die Annahme, dass der wahre Effekt von Studie zu Studie variiert.
- Unterschiedliche Stichprobenzusammensetzungen (z.B. Alter), unterschiedliche Implementierungen der UV, unterschiedliche Messmethoden, unterschiedliche Randbedingungen (z.B. Wetter).
- Mögliche Moderatoren sind oftmals nicht bekannt und wurden nicht erfasst.
- Der Einfluss dieser Moderatoren kann daher als **Zufall** aufgefasst werden.

Random-Effects-Modell

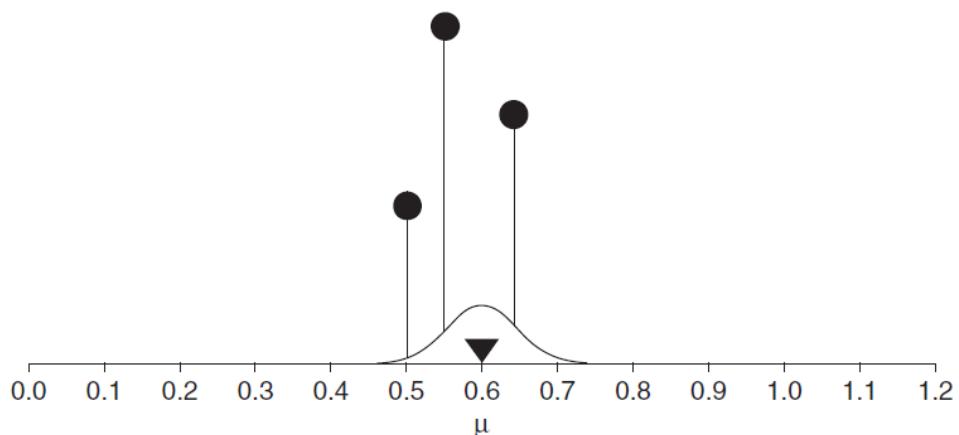
Variation des wahren Effekts:



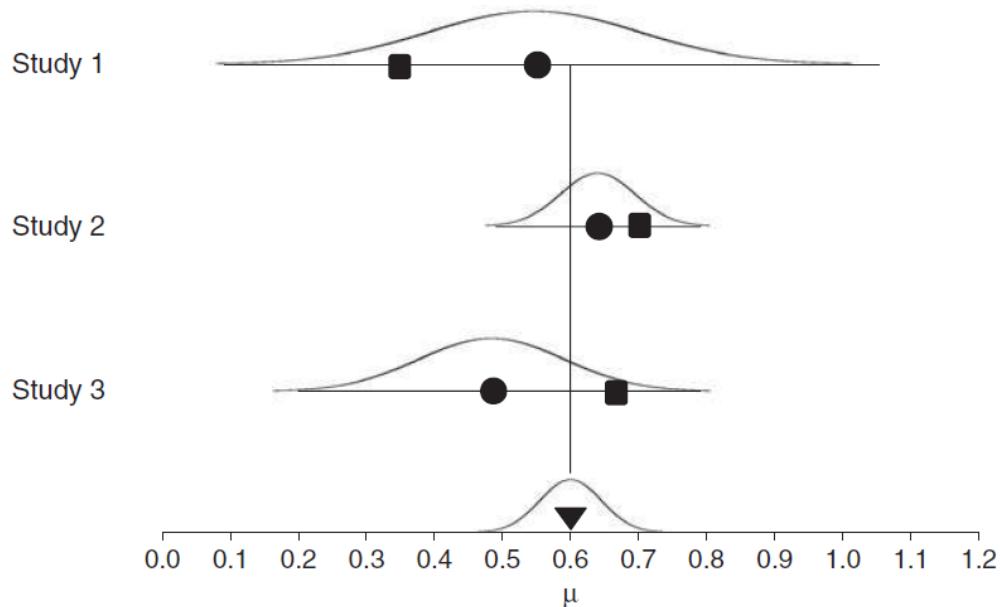
Study 1

Study 2

Study 3

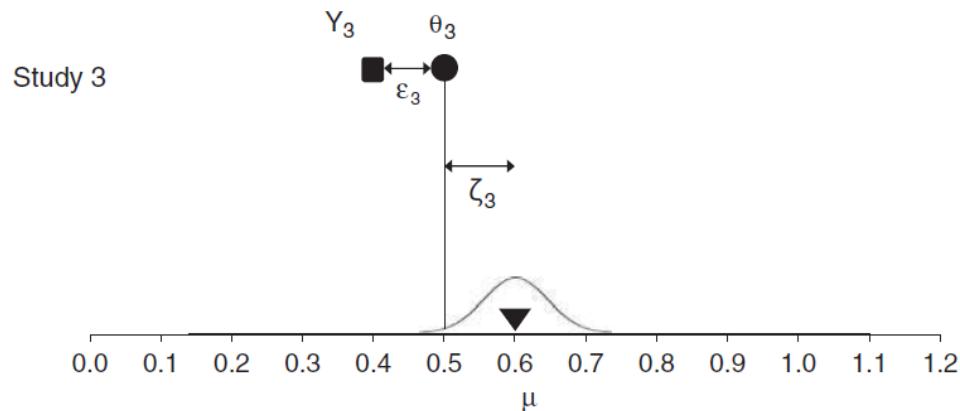


Random-Effects-Modell



Random-Effects-Modell

Warum weicht der beobachtete Effekt in einer Studie vom *Mittelwert der wahren Effekte* ab:



ε : Der beobachtete Effekt ist nicht identisch mit dem wahren Effekt in dieser Studie.

ζ : Der wahre Effekt der Studie ist nicht identisch mit dem mittleren wahren Effekt

Random-Effects-Modell

- Neben dem Standardfehler (SE , *within study variation*) muss nun also auch die Variation der wahren Effekte zwischen Studien (τ^2 , *between study variation*) geschätzt werden.
- **Formeln:**

$$T^2 = \frac{Q - df}{C}$$

$$Q = \sum_{i=1}^k W_i Y_i^2 - \frac{\left(\sum_{i=1}^k W_i Y_i \right)^2}{\sum_{i=1}^k W_i}$$

$$df = k - 1$$

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

Fixed-Effects- Modell

$$V_{Y_i} = \frac{\sigma^2}{n}$$

$$W_i = \frac{1}{\sigma^2/n} = \frac{n}{\sigma^2}$$

$$V_M = \frac{1}{\sum_{i=1}^k W_i} = \frac{1}{k \times n / \sigma^2} = \frac{\sigma^2}{k \times n}$$

$$SE_M = \sqrt{\frac{\sigma^2}{k \times n}}$$

Random-Effects- Modell

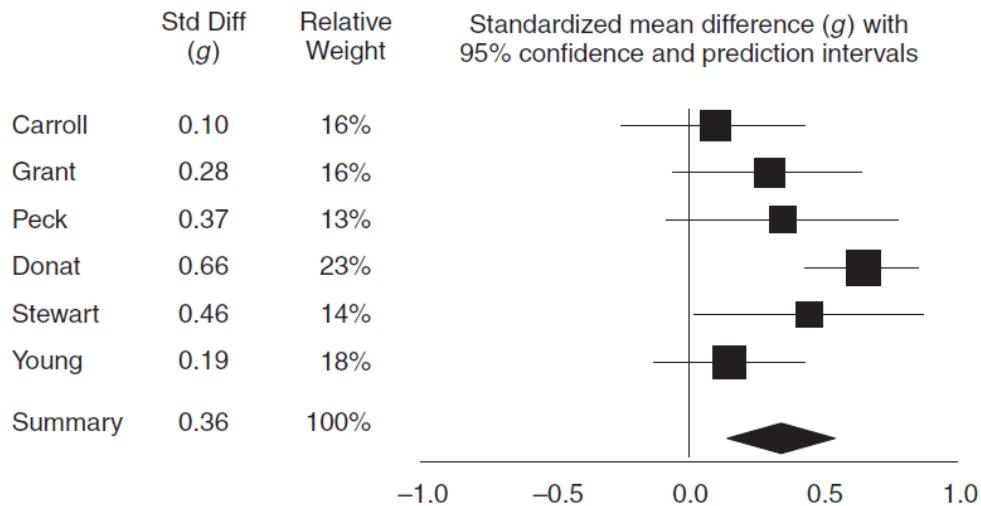
$$V_{Y_i}^* = V_{Y_i} + T^2$$

$$W_i^* = \frac{1}{(\sigma^2/n) + \tau^2}$$

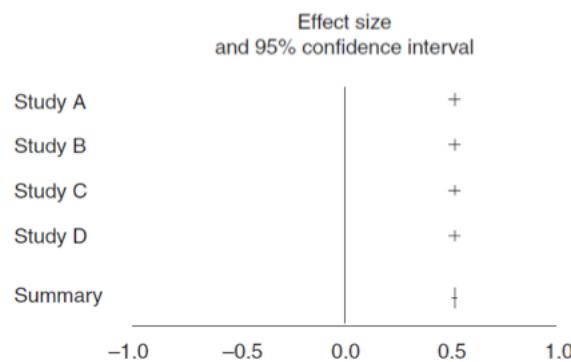
$$SE_{M^*} = \sqrt{\frac{\sigma^2}{k \times n} + \frac{\tau^2}{k}}$$

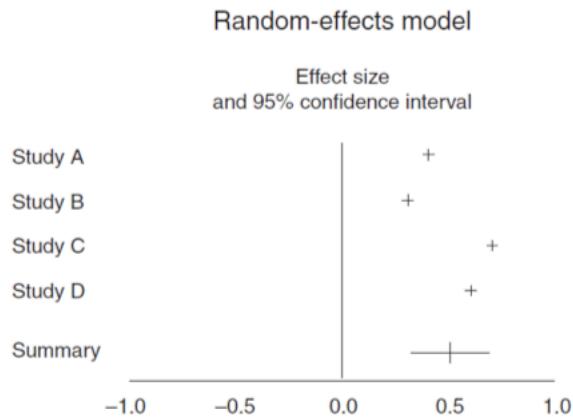
Random-Effects-Modell

Impact of Intervention (Random effects)



Fixed-effect model





Random-Effects-Modell

Formeln:

$$V_{Y_i}^* = V_{Y_i} + T^2 \quad W_i^* = \frac{1}{V_{Y_i}^*} \quad M^* = \frac{\sum_{i=1}^k W_i^* Y_i}{\sum_{i=1}^k W_i^*}$$

$$V_{M^*} = \frac{1}{\sum_{i=1}^k W_i^*} \quad SE_{M^*} = \sqrt{V_{M^*}}$$

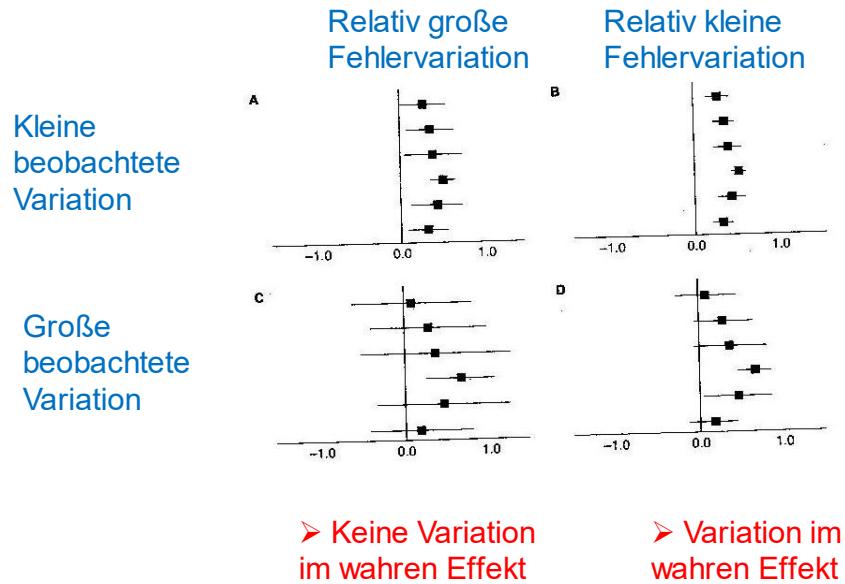
$$LL_{M^*} = M^* - 1.96 \times SE_{M^*}$$

$$UL_{M^*} = M^* + 1.96 \times SE_{M^*}$$

$$Z^* = \frac{M^*}{SE_{M^*}}$$

Heterogenitätsmaße

- Heterogenität: Variation der *wahren* Effektstärken
- Die *beobachtete* Variation setzt sich zusammen aus der Variation der wahren Effektstärken (*between study variation*) und der Fehlervariation (*within study variation*).



Heterogenitätsmaße

- Logik der Berechnung der Variation der *wahren* Effektstärken:
 - Berechnung der beobachteten Variation
 - Berechnung der *erwarteten* Fehlervariation (wenn der wahre Effekt in allen Studien identisch wäre).
 - Die Differenz beider Variationen ist ein Indikator der Variation der wahren Effektstärken.

Heterogenitätsmaße

- Q ist ein standardisiertes Maß der beobachteten Variation

$$Q = \sum_{i=1}^k W_i (Y_i - M)^2,$$

- Gewichtete Quadratsumme (Gewichte: inverse Varianz)

$$Q = \sum_{i=1}^k \left(\frac{Y_i - M}{S_i} \right)^2$$

- Q ist also auch ein standardisiertes Maß

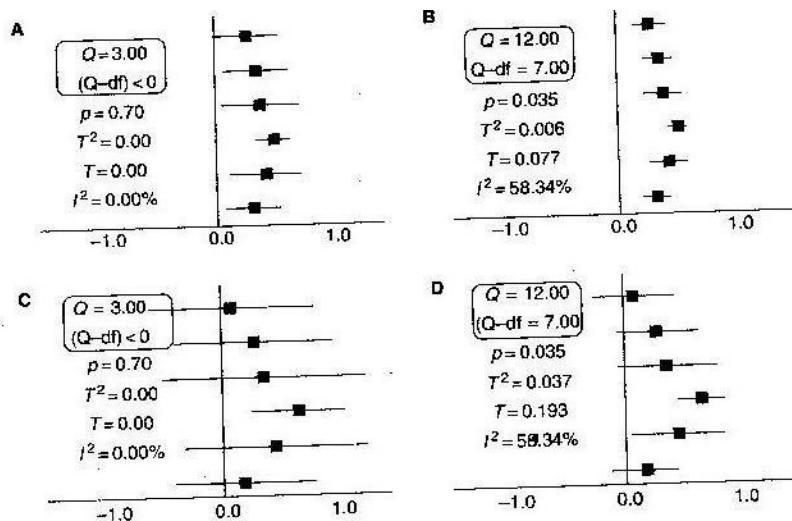
- Erwartete Fehlervariation:

$$df = k - 1$$

- Variation der wahren Effektstärken (standardisiert):

$$Q - df,$$

Heterogenitätsmaße



Heterogenitätsmaße

- Test der Nullhypothese, dass die Variation der wahren Effektstärken 0 beträgt ($Q - df = 0$):

$$-\chi^2 = Q \quad \text{mit } df = k - 1$$

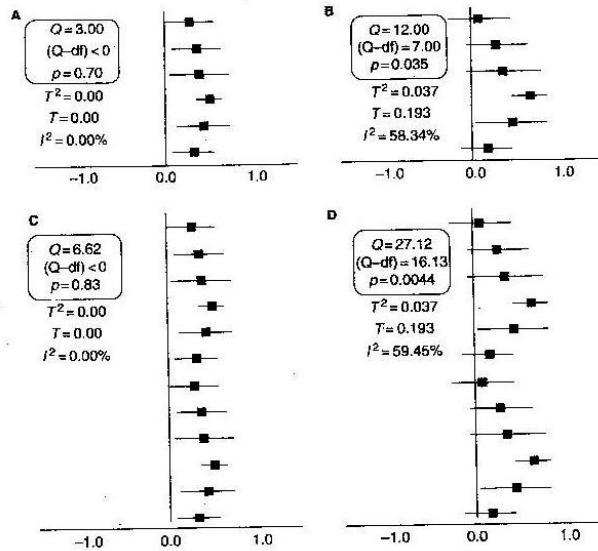
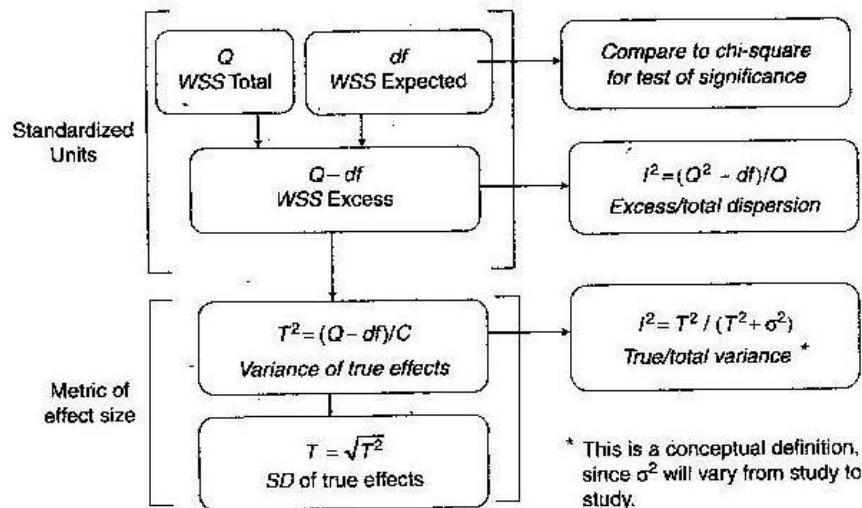


Figure 16.4 Impact of Q and number of studies on the p -value.

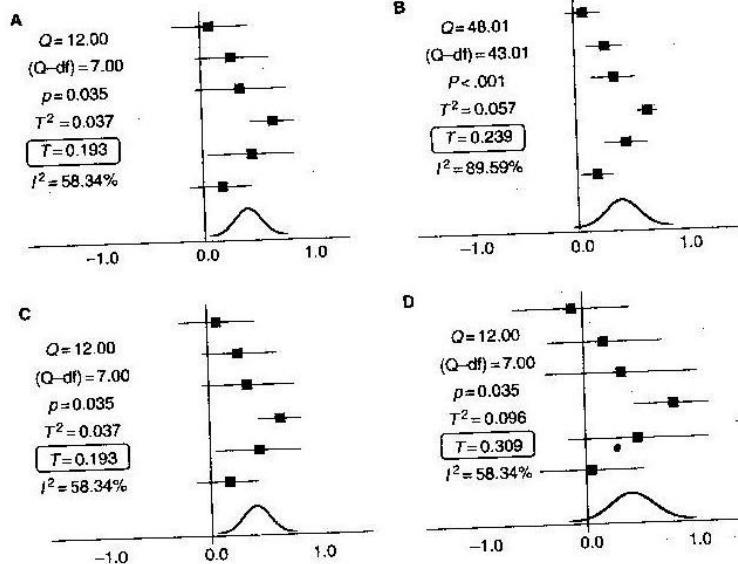
Heterogenitätsmaße



Heterogenitätsmaße

- τ^2 ist Maß der Variation der wahren Effektstärken in den ursprünglichen Maßeinheiten.

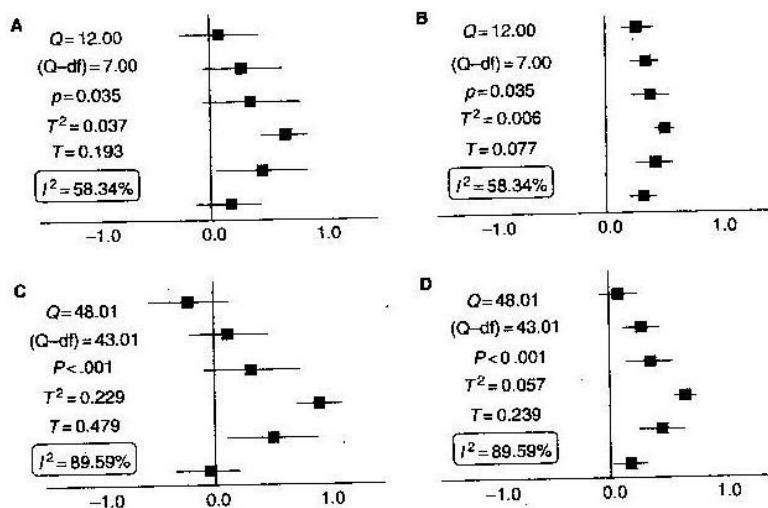
$$T^2 = \frac{Q - df}{C} \quad C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$



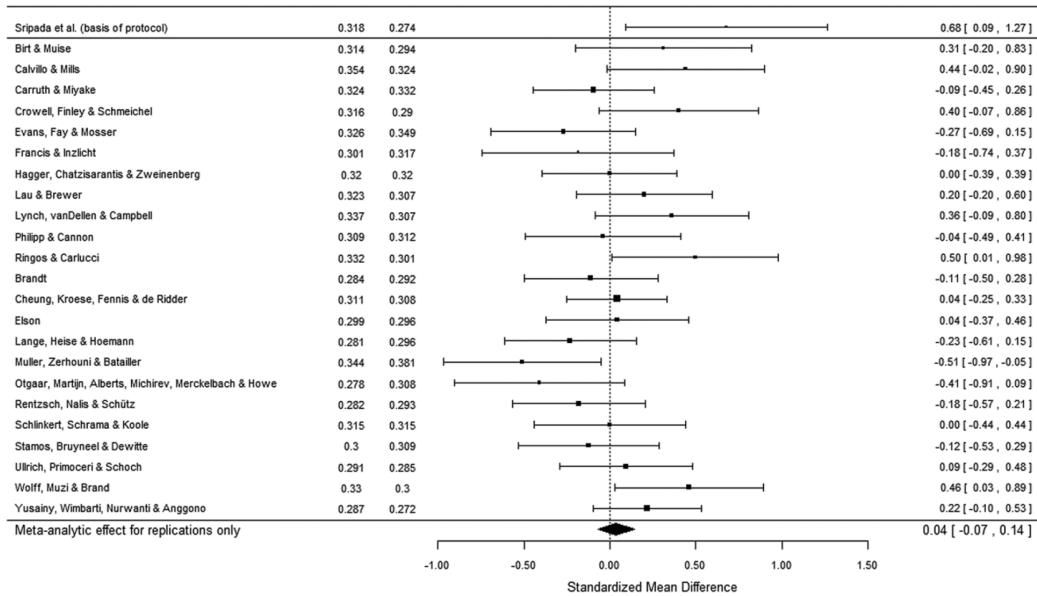
Heterogenitätsmaße

- I^2 bezeichnet den Anteil der Variation der wahren Effektstärken an der beobachteten Variation.

$$I^2 = \left(\frac{Q - df}{Q} \right) \times 100\% \quad I^2 = \left(\frac{\text{Variance}_{bet}}{\text{Variance}_{total}} \right) \times 100\% = \left(\frac{\tau^2}{\tau^2 + V_y} \right) \times 100\%,$$



Beispiel: Ego-Depletion (Sripada et al., 2014)



- $\tau = 0.16$ ($I^2 = 0.36$, Hagger & Chatzirantis, 2016)

Moderatoranalyse

- Gibt es *systematische* Unterschiede in der Effektstärke zwischen Gruppen von Studien?
 - Therapie A vs. Therapie B
 - Intervention A vs. Intervention B
 - Bei Studien mit unterschiedlichen Populationen von Versuchsteilnehmern
 - Studierende vs. Andere
 - Asiaten vs. Europäer
 - Studien mit unterschiedlichen Randbedingungen
 - Leistungsabhängige Bezahlung vs. “Teilnahmeentschädigung”
 - Studien mit unterschiedlichen Messmethoden der AV
 - Verhalten vs. Rating

Moderatoranalyse im Fixed-Effect Modell

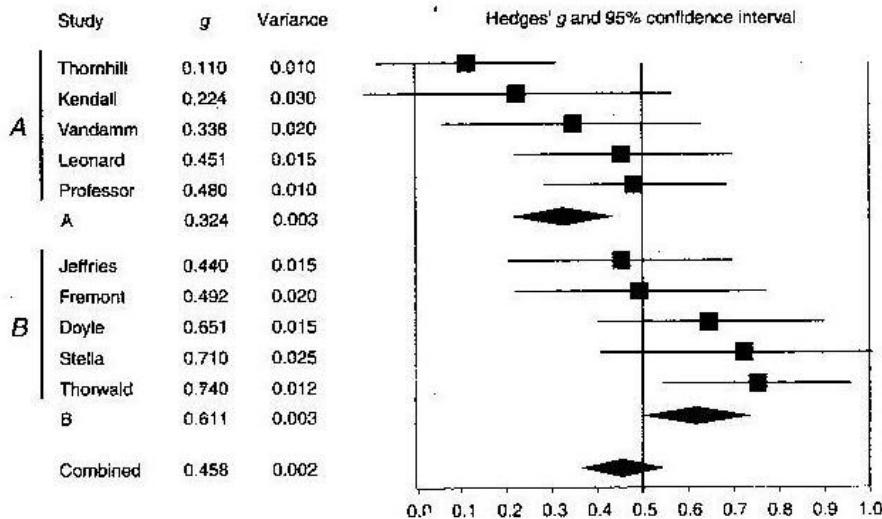
- Separate Analyse der verschiedenen Subgruppen von Studien
- Vergleich der mittleren Effekte in verschiedenen Subgruppen mittels Konfidenzintervallen und Signifikanztests.
- Eine Möglichkeit für den Signifikanztest: Z-Test
 - Alternative: (Q -Test analog zur ANOVA)
 - Test der Nullhypothese, dass die beiden wahren Effektstärken gleich sind.

$$H_0 : \theta_A = \theta_B,$$

– Formel: $Z_{Diff} = \frac{Diff}{SE_{Diff}}$

Mit: $Diff = M_B - M_A$, $SE_{Diff} = \sqrt{V_{M_A} + V_{M_B}}$.

Moderatoranalyse im Fixed-Effect Modell



Moderatoranalyse im Fixed-Effect Modell

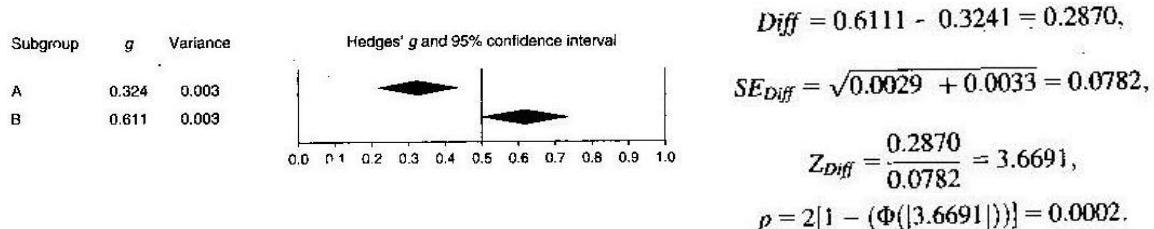
Table 19.1 Fixed effect model – computations.

Study	Effect size γ	Variance Within V_γ	Variance Between τ^2	Variance Total V	Weight W	Calculated quantities		
						WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0000	0.0100	100.000	11.000	1.210 10000.000
	Kendall	0.224	0.0300	0.0000	0.0300	33.333	7.467	1.673 1111.111
	Vandamm	0.338	0.0200	0.0000	0.0200	50.000	16.900	5.712 2500.000
	Leonard	0.451	0.0150	0.0000	0.0150	66.667	30.067	13.560 4444.444
	Professor	0.480	0.0100	0.0000	0.0100	100.000	48.000	23.040 10000.000
	Sum A				350.000	113.433	45.195	28055.556
B	Jeffries	0.440	0.0150	0.0000	0.0150	66.667	29.333	12.907 4444.444
	Fremont	0.492	0.0200	0.0000	0.0200	50.000	24.600	12.103 2500.000
	Doyle	0.651	0.0150	0.0000	0.0150	66.667	43.400	28.253 4444.444
	Stella	0.710	0.0250	0.0000	0.0250	40.000	28.400	20.164 1600.000
	Thorwald	0.740	0.0120	0.0000	0.0120	83.333	61.667	45.633 6944.444
	Sum B				306.667	187.400	119.061	19933.333
	Sum				656.667	300.833	164.255	47988.889

Moderatoranalyse im Fixed-Effect Modell

Table 19.2 Fixed-effect model – summary statistics.

	A	B	Combined
γ	0.3241	0.6111	0.4581
V	0.0029	0.0033	0.0015
SE_γ	0.0535	0.0571	0.0390
LL_γ	0.2193	0.4992	0.3816
UL_γ	0.4289	0.7230	0.5346
Z	6.0633	10.7013	11.7396
$p2$	0.0000	0.0000	0.0000
Q	8.4316	4.5429	26.4371
df	4.0000	4.0000	9.0000
$p\text{-value}$	0.0770	0.3375	0.0017
Numerator	4.4316	0.5429	17.4371
C	269.8413	241.6667	583.5871
τ^2	0.0164	0.0022	0.0299
I^2	52.5594	11.9506	65.9569

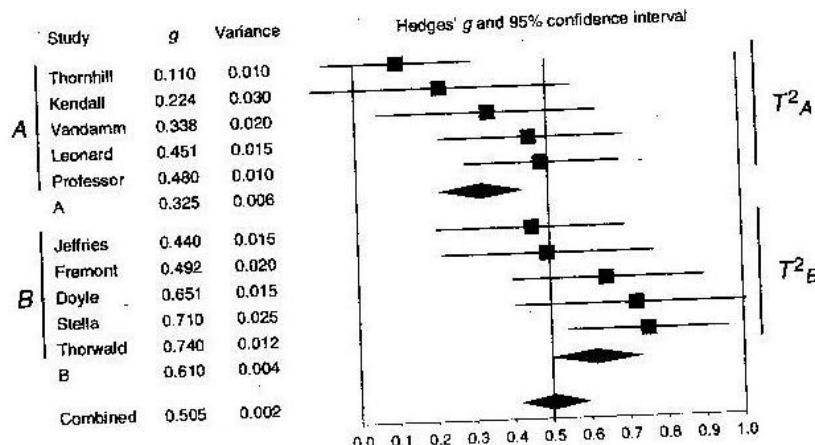


Moderatoranalyse im Random-Effect Modell

- Zwischen Studiengruppen bestehen möglicherweise systematische Unterschiede.
- Innerhalb der Gruppen kann der wahre Effekt dennoch zufällig variieren.
- Separate Analyse der verschiedenen Subgruppen von Studien mit dem Random-Effect Modell
- Wird auch als Mixed-Effect Modell bezeichnet (Random-Effect innerhalb der Gruppen, Fixed Effect zwischen Gruppen).
- In diesem Fall wird die (Zufalls-)Variation der wahren Effektstärke (τ^2) in beiden Gruppen separat berechnet und berichtet.
 - Für die weitere Analyse kann entweder eine *gepoolte* Schätzung von τ^2 verwendet werden oder eine jeweils eigene Schätzung pro Subgruppe
 - Erwarten wir, dass die Zufallsvariation in allen Gruppen gleich ist?

Moderatoranalyse im Random-Effect Modell

- Vergleich der mittleren Effekte in verschiedenen Subgruppen mittels Konfidenzintervallen und Signifikanztests.
- Eine Möglichkeit für den Signifikanztest: Z-Test
 - Alternative: (Q -Test analog zur ANOVA)



Moderatoranalyse im Random-Effect Modell

Table 19.5 Random-effects model (separate estimates of τ^2) – computations.

Study	Effect size Y	Variance Within V_Y	Variance Between τ^2	Variance Total V	Weight W	Calculated quantities		
						WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0164	0.0264	37.846	4.163	0.458 1432.308
	Kendall	0.224	0.0300	0.0164	0.0464	21.541	4.825	1.081 464.017
	Vandamm	0.338	0.0200	0.0164	0.0364	27.455	9.280	3.137 753.788
	Leonard	0.451	0.0150	0.0164	0.0314	31.824	14.353	6.473 1012.757
	Professor	0.480	0.0100	0.0164	0.0264	37.846	18.166	8.720 1432.308
	Sum A					156.512	50.787	19.868 5095.179
B	Jefferies	0.440	0.0150	0.0022	0.0172	57.983	25.512	11.225 3362.002
	Fremont	0.492	0.0200	0.0022	0.0222	44.951	22.116	10.881 2020.582
	Doyle	0.651	0.0150	0.0022	0.0172	57.983	37.747	24.573 3362.002
	Stella	0.710	0.0250	0.0022	0.0272	36.702	26.058	18.501 1347.034
	Thorwald	0.740	0.0120	0.0022	0.0142	70.193	51.943	38.438 4927.012
	Sum B					267.811	163.376	103.619 15018.633
Sum						424.323	214.163	123.487 20113.812

Moderatoranalyse im Random-Effect Modell

Table 19.6 Random-effects model (separate estimates of τ^2) – summary statistics.

	A	B	Combined
Y	0.3245	0.6100	0.5047
V	0.0064	0.0037	0.0024
SE_Y	0.0799	0.0611	0.0485
LL_Y	0.1678	0.4903	0.4096
UL_Y	0.4812	0.7298	0.5999
Z	4.0595	9.9833	10.3967
$p2$	0.0000	0.0000	0.0000
Q	3.3882	3.9523	15.3952

$$Z_{Diff}^* = \frac{Diff^*}{SE_{Diff^*}} \quad SE_{Diff^*} = \sqrt{V_M^A + V_M^B}.$$

$$Diff^* = 0.6100 - 0.3245 = 0.2856,$$

Study	g	Variance	Hedges' G and 95% confidence interval
A	0.325	0.006	
B	0.610	0.004	

$$SE_{Diff^*} = \sqrt{0.0064 + 0.0037} = 0.1006,$$

$$Z_{Diff^*} = \frac{0.2856}{0.1006} = 2.8381.$$

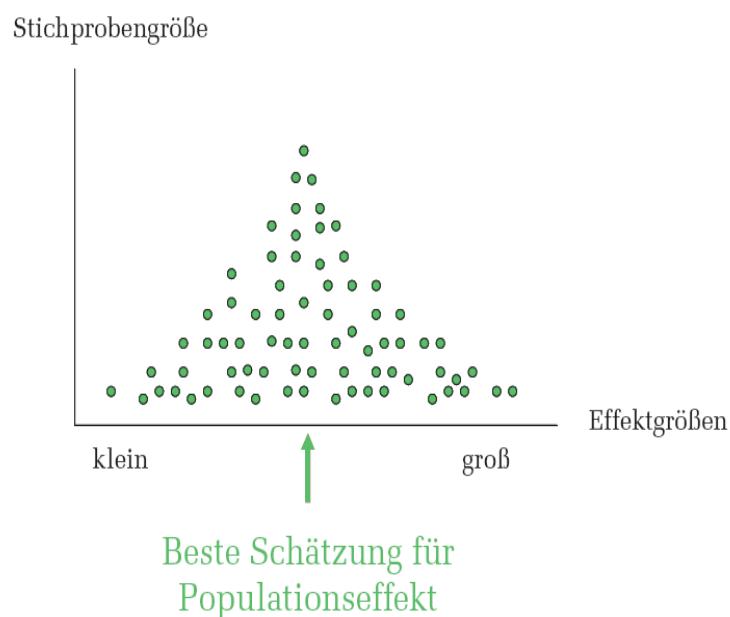
$$p = ,0045$$

Selektionsbias

- In der Psychologie werden oftmals nur signifikante Ergebnisse publiziert.
- Signifikante Ergebnisse sind zu einem Teil solche Ergebnisse, bei denen *zufällig* ein großer Effekt auftrat.
- Liegt ein solcher Publikationsbias vor, führt die Metaanalyse zu einer systematischen Überschätzung des Populationseffekts.
- Wie kann ein Publikationsbias identifiziert werden??

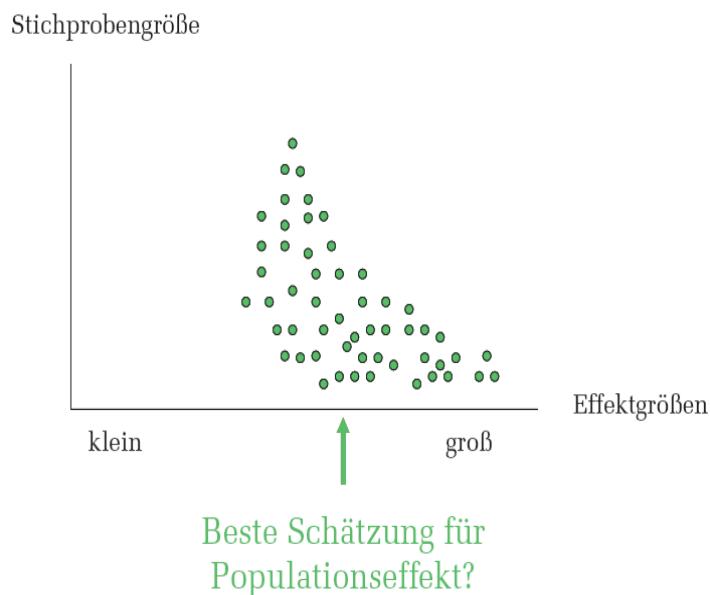
Funnel-Plot

- Kein Selektionsbias:



Funnel-Plot

- Selektionsbias:



Quantitative Metaanalyse - Fazit -

- Aufgrund relativ kleiner Stichproben sind psychologische Untersuchungen häufig mit einem großen *Stichprobenfehler* behaftet.
- Aufgrund unpräziser Messinstrumente sind psychologische Untersuchungen häufig mit einem großen *Messfehler* behaftet.
- Einzelunteruntersuchungen führen daher generell oftmals zu unpräzisen Schätzungen der Effektstärke in der Population.
- Einzeluntersuchungen erlauben also (schon aus statistischen Gründen) keine definitive Aussage über einen Effekt und damit keinen definitiven “Erkenntnisgewinn”.
- Der Signifikanztest “kaschiert” dieses Problem (durch Aussagen wie “Es gibt einen / keinen Effekt”).
- Die Ergebnisse von Signifikanztests erlauben keine sinnvolle Aggregation der Ergebnisse aus mehreren Studien.

Quantitative Metaanalyse

- Fazit -

- Die Metaanalyse löst dieses Problem auf Basis von *Effektstärken*.
- Die Metaanalyse liefert die bestmögliche Schätzung eines Populationseffekts auf der Grundlage *aller* bisherigen Untersuchungen und ermöglicht somit **Wissensakkumulation!**
- Die Metaanalyse erlaubt zudem eine sinnvolle Bestimmung von Moderatorvariablen für einen Effekt!

Quantitative Metaanalyse

- Fazit -

- Aus diesen Sachverhalten leiten manche Autoren (z.B. Schmidt, 1992) Forderungen über einen grundsätzlich veränderten Forschungsprozess ab.
 - Jede methodisch einwandfreie Studie sollte publiziert werden (um einen Selektionsbias zu verhindern).
 - Die “Forschercommunity” könnte sich in zwei Gruppen aufspalten:
 - Eine Gruppe von Forschern führt Einzelstudien durch. Die Ergebnisse dieser Studien werden ausschließlich als “Rohmaterial” für Metaanalysen verwendet und nicht theoretisch interpretiert!
 - Eine zweite Gruppe von Forschern führt die Metaanalysen durch und “gewinnt die Erkenntnisse”.
- Die Anzahl der publizierten Metaanalysen ist in den letzten zwei Jahrzehnten dramatisch gestiegen (s. z.B. Psychological Bulletin)!

26.Januar - Tutorium

1. zur fixed-effect-Metaanalyse
 - a. was ist die Formel für den Haupteffekt dieser Form von Metaanalyse? was stellen Y und W in dieser Formel dar?

$$M = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i}$$

Y: Effektstärke der einzelnen Studie
W: Gewicht der einzelnen Studie - Kehrwert der i. Varianz
 - b. was sind Nutzen eines Konfidenzintervalls (KI) für die Interpretation der gewichteten mittleren Effektstärke einer Metaanalyse?
 - ! neben der Lokation wird auch die Präzision der gewichteten mittleren Effektstärke berichtet und interpretierbar,
 - ! im Vergleich zu den Einzelstudien wird also deutlich, dass der Effekt genauer geschätzt wird,
 - c. gibt es einzelne KIs, die kleiner sind als das der gewichteten mittleren Effektstärke?
 - ! nein
 - d. rechne selbst eine fixed-effect-Metaanalyse (berichte die gewichtete mittlere Effektstärke und ihr KI, einen Signifikanztest gegen eine H_0 brauchst du nicht rechnen) mit den folgenden Studien:
 Studie 1: $g = 0,45$ $V_g = 0,04$
 Studie 2: $g = 0,70$ $V_g = 0,01$
 Studie 3: $d = 0,50$ $n_1 = 25$ $n_2 = 25$
 ! die Bezeichnungen der Effektgrößen erfolgten nach Borenstein!

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

$$J = \text{Korrekturfaktor} = 1 - \frac{3}{4df - 1}$$

$$W_1 = 25, \quad W_2 = 100, \quad W_3 = 12.516, \quad V_d = 0,0825, \quad V_g = 0.0799, \quad g = 0.49$$

$$M = \frac{(25 * 0.45 + 100 * 0.70 + 12.516 * 0.49)}{(25 + 100 + 12.516)} = 0.635$$

$$LL_M = M - 1,96 \times SE_M$$

$$M = 0.635 \quad SE_M = 0.084$$

$$LL_M = 0.635 - 1.96 * 0.084 = 0.470$$

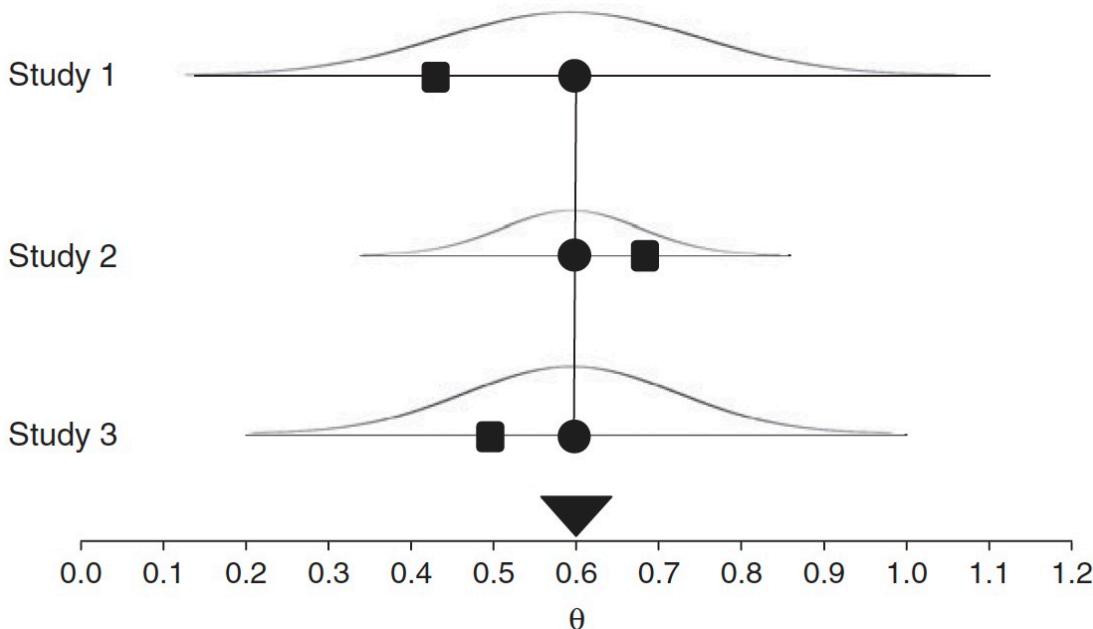
$$UL_M = M + 1,96 \times SE_M$$

$$UL_M = 0.635 + 1.96 * 0.084 = 0.800$$

2. beschreibe den grundlegenden Unterschied oder die grundlegende Idee von *random* vs. *fixed*? wann sollte welche Methodik angewendet werden?

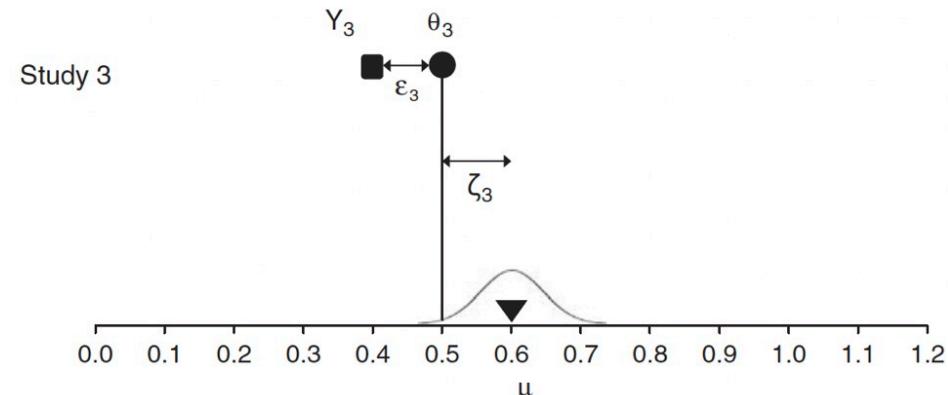
- ! *fixed*: es gibt einen wahren Effekt, der durch alle Studien versucht wird, zu erfassen, *random*: der wahre Effekt ist nicht fest, sondern variiert zufällig in Form einer Normalverteilung,
- ! dadurch schätzt jede Studie (abhängig von ihrem exakten Design, der Stichprobe, dem Zeitpunkt, etc.) einen anderen Punkt aus dieser Normalverteilung,
- ! Fälle für ein fixed-effect-model:
 - wir wissen aufgrund theoretischer Vorüberlegungen mit Sicherheit, dass der wahre Effekt nicht variieren kann - in der Psychologie wohl ausgeschlossen!
 - Tau-Quadrat wird 0,
 - in diesem Fall rechnen wir prinzipiell wieder mit einem fixed-effect-design,

3. was wird in dieser Abbildung dargestellt? wo werden in dieser Abbildung Quellen für Varianz dargestellt? ist hier ein fixed-effect- oder ein random-effects-modell abgebildet?



- ! die Kreise zeigen die wahren Effekte an, die Quadrate zeigen die tatsächlich gemessenen Effekte an,
- ! die wahren Effekte selbst variieren nicht, sondern liegen alle bei 0.6 → starker Hinweis für fixed-effect-modell
- ! drei Studien, ihr gemessener Effekt, die Lage des metaanalytisch bestimmten wahren Effekts und wie die Werte der einzelnen Studien aus Sicht des genutzten metaanalytischen Modells zu verorten sind - also wie wahrscheinlich ihr Ergebnis war, gegeben ihre Stichprobengröße und den hier bestimmten wahren Effekt,
- ! Varianz wird als Abweichung vom wahren Effekt in Form des Messfehlers der einzelnen Untersuchung dargestellt,

4. die Abbildung visualisiert grundlegende Annahmen hinter einem der beiden metaanalytischen Modelle - finde heraus, um welches Modell es sich handelt und beschreibe die abgebildeten Annahmen!



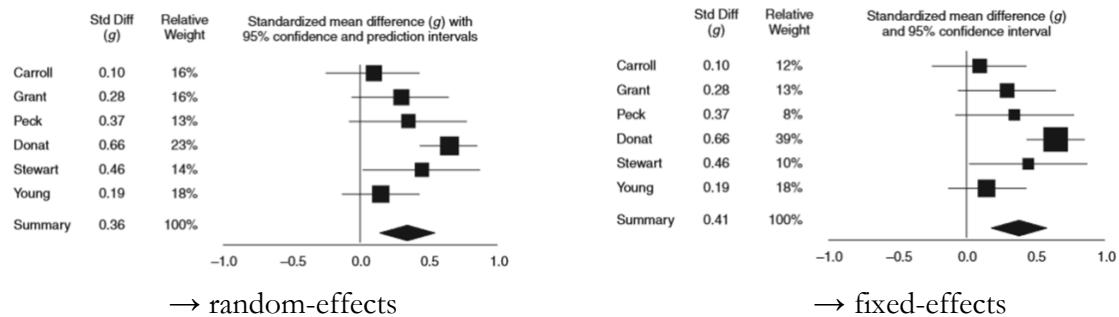
μ Mittelwert der Verteilung der wahren Effekte
 θ wahrer Effekt der Studie
 ζ (Zeta) Abstand zwischen dem wahren Effekt der Studie und μ Y ist der empirisch gefundene Wert der Studie
 ϵ Messfehler der Studie

→ Random-Effects-Modell

5. erkläre die Annahme der Homogenität von Effekten in eigenen Worten!
- ! die Annahme, dass alle in der Metaanalyse genutzten Studien die gleiche wahre Effektstärke besitzen und demnach die Varianz zwischen diesen Studien nur auf Messfehler der einzelnen Studien zurückzuführen ist,
6. mit den gleichen 40Studien wurden eine mittlere gewichtete Effektstärke nach random-effects-modell und eine nach fixed-effect-modell berechnet - die beiden Modelle kommen zu deutlich verschiedenen Ergebnissen - welche Aussage kannst du daraus über die Heterogenität zwischen den Effektstärken der 40Studien treffen?
- ! sie ist sehr groß!
7. formuliere in eigenen Worten was durch Q ausgedrückt wird - erkläre außerdem, warum die hier notierten Formeln zwar für das Verständnis nützlich, aber im rechnerischen Vorgehen manchmal nicht praktikabel sind!
- ! Q sind die gewichteten quadrierten Abweichungen der einzelnen Effektstärken von ihrem Mittelwert,
 - ! diese Formel benötigt den Mittelwert aus dem fixed-effect-modell,
 - ! wenn dieser nicht bereits vorliegt, ist es praktikabler, die Version der Formel zu nutzen, die sich auf die Gewichte und Effektstärken beschränkt,

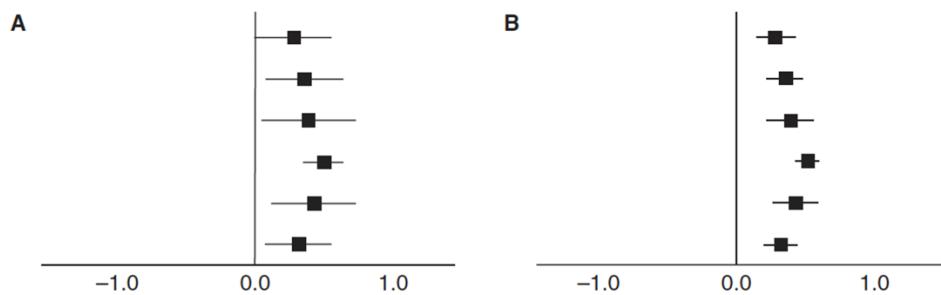
$$\sum_{i=1}^k W_i (Y_i - M)^2 = \sum_{i=1}^k \left(\frac{Y_i - M}{S_i} \right)^2 \quad Q = \sum_{i=1}^k W_i Y_i^2 - \frac{\left(\sum_{i=1}^k W_i Y_i \right)^2}{\sum_{i=1}^k W_i}$$

8. vergleiche die beiden Abbildungen miteinander - bei welcher handelt es sich um ein fixed-effect-modell und welche beschreibt das random-effects-modell? begründe deine Entscheidung!



- ! da in beiden Fällen die gleichen Studien ausgewertet wurden, die Gewichte aber in der linken Abbildung deutlich ausgeglichener sind, liegt nahe, dass hier der Einfluss von τ^2 sichtbar ist,

9. in welcher Abbildung vermutest du das höhere Q: A oder B? ...begründe!



- ! in Abbildung B, da die einzelnen Studien kleinere Konfidenzintervalle aufweisen - diese lassen auf kleinere Standardabweichungen der einzelnen Studien schließen,

$$\sum_{i=1}^k W_i (Y_i - M)^2 = \sum_{i=1}^k \left(\frac{Y_i - M}{S_i} \right)^2$$