2. Vorlesung Statistik II
Nicht zusammengesetzte Hypothesentests und Metaanalyse

Jetzt

Thema der ersten Vorlesungen: Interpretation von Gruppen von Hypothesentests:

- ✓ Zusammengesetzte Hypothesentests
- **Mehrere nicht zusammengesetzte Hypothesentests aus einer oder mehreren Stichproben mit unterschiedlichen Hypothesen**
- Mehrere Hypothesentests aus verschiedenen Stichproben mit identischen Hypothesen

Ausgangssituation I

- Ausgangslage: Wir haben eine Gruppe von N Hypothesentests:
  Test 1, Test 2, …, Test j, …, Test N über beliebig viele Stichproben verteilt.

- Mithilfe jedes einzelnen Hypothesentests j, können wir jeweils die Hypothesen H_{0j} und H_{1j} bei einem Signifikanzniveau von α überprüfen. Hierbei ist H_{0j} jeweils die logische Negation von H_{1j} (also das „Gegenteil“).

- Diesmal interessieren wir uns aber direkt für die einzelnen Hypothesen.

- Dies ist der „Normalfall“: Wir haben verschiedene Hypothesen, die wir alle einzeln überprüfen wollen, da uns jeweils interessiert, ob diese wahr oder falsch sind.

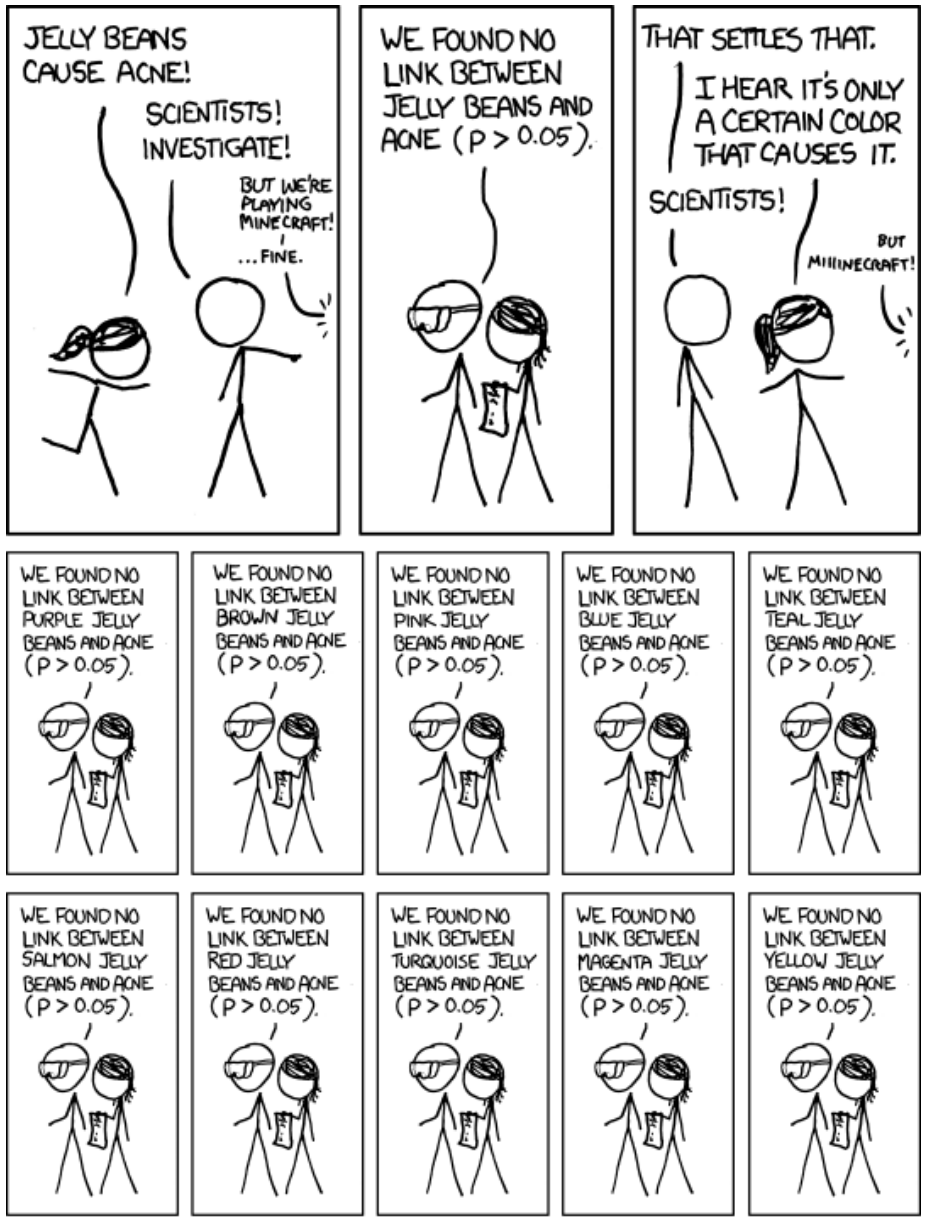
Ausgangssituation II

- Da wir uns nicht für zusammengesetzte Hypothesen interessieren, gibt es auch kein Signifikanzniveau α^* für einen zusammengesetzten Hypothesentest.
- Es gibt also zunächst keinen direkten Grund, warum wir uns über eine mögliche α-Fehler-Kumulierung Gedanken machen sollten.

Aber:

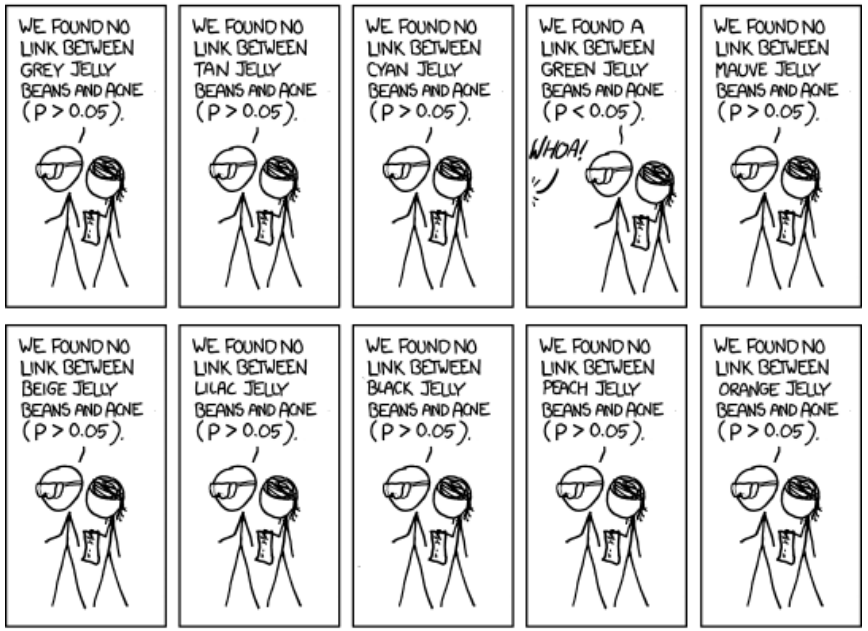
- Könnte es nicht trotzdem sinnvoll sein, die Wahrscheinlichkeit dafür, mindestens eine falsche Entscheidung für eine H_{1j} zu treffen, also die FWER, zu kontrollieren (z.B. mithilfe der Bonferroni-Methode)?

Beispiel I



https://xkcd.com/882/

Beispiel II



https://xkcd.com/882/

Beispiel III

Was passiert in diesem Beispiel aus rein statistischer Sicht?

- Es wird eine Reihe von Hypothesentests jeweils mit einem bestimmten Signifikanzniveau α durchgeführt.
- Bei genau einem der Hypothesentests entscheiden wir uns für die H_{1j}. Bei allen anderen für die H_{0j}.
- Es wird keine Aussage über eine zusammengesetzte Hypothese getroffen, sondern nur die einzelne Entscheidung für die H_{1j} berichtet.
- Die Wahrscheinlichkeit dafür, uns in diesem Fall **mindestens einmal fälschlicherweise** für eine H_{1j} zu entscheiden, ist aufgrund der α-Fehler-Kumulierung relativ hoch.
- Sollten wir also in diesem Fall die FWER kontrollieren, auch wenn wir uns nicht für zusammengesetzte Hypothesen interessieren?

Probleme

- Für welche Gruppe von Tests soll die FWER genau kontrolliert werden? Im Gegensatz zum Fall mit zusammengesetzten Hypothesen ist dies hier nicht klar:
  - Was passiert, wenn in einer anderen Studie Jelly Beans mit weiteren Farben untersucht werden? Müssen diese Tests dann zur Testgruppe hinzugefügt werden und alle p-Werte aus der Originalstudie rückwirkend nachkorrigiert werden?
  - Was passiert, wenn eine weitere Forschungsgruppe die gleiche Untersuchung mit Smarties statt mit Jelly Beans durchführt? Werden diese Tests dann in die Gruppe der Tests aufgenommen, für welche die FWER kontrolliert werden soll?
  - Warum nicht auch noch alle Hypothesentests zu anderen Einflussfaktoren auf Akne mit in die Gruppe der Tests aufnehmen?
  - Was ist mit beliebigen Gruppen von Hypothesentests mit inhaltlich völlig unabhängigen einzelnen Hypothesen? Rein statistisch gesehen ist die Situation hier die gleiche wie im Jelly Beans Beispiel.
- Problem hierbei: Die FWER hängt von der Größe der jeweils gewählten Testgruppe ab.

Alternative

- Außerdem: Auch wenn es möglich wäre, hier eine eindeutige Entscheidung zu treffen, ist die Forderung nach einer niedrigen FWER immer noch ein sehr strenges Kriterium.
- Die FWER entspricht schließlich der Wahrscheinlichkeit, bei einer Gruppe von Hypothesentests **mindestens einen Fehler erster Art** zu begehen. Dieses Kriterium bleibt also gleich, egal wie viele Hypothesentests betrachtet werden.
- Falls wir uns für einzelne H_{1j} Entscheidungen interessieren, kann man sehr gut argumentieren, dass es ausreicht, den **durchschnittlichen Anteil der falschen H_{1j} Entscheidungen unter allen Entscheidungen für die einzelnen Alternativhypothesen** zu kontrollieren (siehe Statistik I).
- Das heißt: Es würde ausreichen, die **False Discovery Rate** zu kontrollieren.

Wiederholung False Discovery Rate

- Die **False Discovery Rate (FDR)** ist definiert als durchschnittlicher Anteil der falsch positiven Entscheidungen an allen positiven Entscheidungen:

$$FDR = \frac{fp}{gp} = \frac{fp}{fp + rp} = \frac{\alpha \cdot \rho \cdot N}{\alpha \cdot \rho \cdot N + (1 - \beta) \cdot (1 - \rho) \cdot N} = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)}$$

N = Anzahl der durchgeführten Hypothesentests

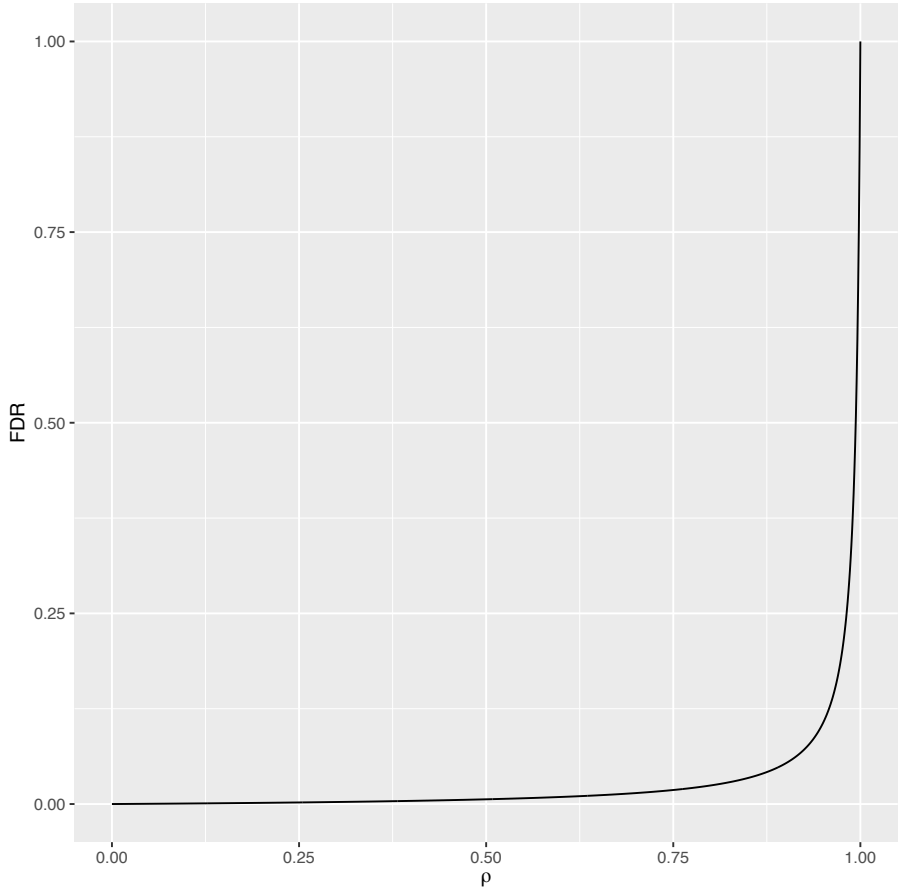
α = Signifikanzniveau der einzelnen Hypothesentests

ρ = Basisrate, d.h. der Anteil der Tests, in denen die Nullhypothese H_{0j} gilt

$1 - \beta$ = durchschnittliche Power der betrachteten Hypothesentests

- **Problem:** FDR kann nicht direkt kontrolliert werden, da deren Höhe von der unbekannten Basisrate abhängt.
- **Lösung:** Absichern gegen eine ungünstige Basisrate durch ein geringes α und ein hohes $1 - \beta$.

Wiederholung False Discovery Rate II



- Vorschlag: $\alpha = 0.005$ (Benjamin et. al. 2017) und $1 - \beta = 0.8$
- In diesem Fall müsste die Basisrate sehr hoch sein, damit eine problematische FDR resultieren würde:
- Die benötigten Stichprobengrößen sind jedoch dementsprechend groß!

```
Two-sample t test power calculation

              n = 297.8117
              d = 0.3
      sig.level = 0.005
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Vorteile der FDR

- Unter den genannten Bedingungen können wir davon ausgehen, dass zwar mit hoher Wahrscheinlichkeit mindestens eine falsche Entscheidung für eine H_{1j} unter allen Entscheidungen für die einzelnen Alternativhypothesen ist (da keine FWER-Kontrolle), aber der durchschnittliche Anteil dieser falschen Entscheidungen für die H_{1j} an allen Entscheidungen für die einzelnen Alternativhypothesen gering ist.
- Ein weiterer Vorteil der FDR gegenüber der FWER ist, dass sie nicht von der Anzahl der Tests und der jeweils gewählten Testgruppe abhängt: Falls wir eine Gruppe von N_1 Tests mit einer FDR von 0.05 haben und eine zweite Gruppe von N_2 Tests mit derselben FDR, dann ist die FDR für die Gesamtgruppe der $N_1 + N_2$ Tests ebenfalls 0.05 .

Bemerkungen I

- Bemerkung I: Dass bei geringer FDR die FWER nicht unbedingt korrigiert werden muss, bedeutet nicht, dass das Vorgehen, sehr viele verschiedene einzelne Hypothesen zu testen und nur die signifikanten Ergebnisse zu berichten (wie im Jelly Beans Beispiel), unproblematisch ist. Es sollten trotzdem immer alle Entscheidungen, also auch die H_{0j} Entscheidungen, berichtet werden (mehr dazu in der nächsten Vorlesung).
- Bemerkung II: Ein Fall, in dem es sinnvoll sein kann, die FWER zu kontrollieren, obwohl man sich für die einzelnen Hypothesen interessiert, ist, wenn man gleichzeitig Aussagen über einzelne Hypothesen und über aus diesen zusammengesetzte Hypothesen treffen will.

  Beispiel: Wir wollen überprüfen, ob Persönlichkeit einen Einfluss auf Depression hat (zusammengesetzte Hypothese) und gleichzeitig Aussagen darüber treffen, welche Persönlichkeitsdimensionen einen Einfluss haben (einzelne Hypothesen).

Bemerkungen II

- Bemerkung III: Für große Gruppen von Tests gibt es Methoden, um die FDR approximativ zu kontrollieren:
  - Benjamini & Hochberg (1995): Sehr verbreitete Methode zur Korrektur der p-Werte, die genau genommen nicht direkt die FDR, sondern $FDR \cdot P(S > 0)$ kontrolliert, wobei S die Anzahl der Entscheidungen für die Alternativhypothese ist. Daher ist sie vor allem für große Gruppen von Tests geeignet, weil man hier davon ausgehen kann, dass auch nach der Korrektur die Wahrscheinlichkeit $P(S > 0)$ dafür, dass es mindestens ein signifikantes Ergebnis gibt, nahe 1 ist und deshalb $FDR \cdot P(S > 0) \approx FDR$ ist.
  - Komplexere Methoden (z.B. Storey, 2003 oder Efron, 2004), die aus einer großen Gruppe von Tests auf der Basis der Realisationen der einzelnen Teststatistiken die Basisrate schätzen und zur Kontrolle der FDR verwenden.

Thema der ersten Vorlesungen: Interpretation von **Gruppen von Hypothesentests**:

- ✓ Zusammengesetzte Hypothesentests
- ✓ Mehrere nicht zusammengesetzte Hypothesentests aus einer oder mehreren Stichproben mit unterschiedlichen Hypothesen
- **Mehrere Hypothesentests aus verschiedenen Stichproben mit identischen Hypothesen**

Ausgangssituation I

- Ausgangslage: Wir haben eine Gruppe von N Hypothesentests:
  Test 1, Test 2, …, Test j, … , Test N alle aus unterschiedlichen Stichproben.
- Mithilfe jedes einzelnen Hypothesentests j, können wir jeweils die Hypothesen H_{0j} und H_{1j} bei einem Signifikanzniveau von α überprüfen. Hierbei ist H_{0j} jeweils die logische Negation von H_{1j} (also das „Gegenteil“).
- Außerdem: Alle einzelnen Hypothesentests überprüfen **die gleichen Hypothesen**.
- Warum sollte man mehrfach die gleichen Hypothesen überprüfen?

Wdh.: Probleme kleiner Stichproben

- Probleme von Studien mit kleinen Stichproben (u.a.):
  - falls wir Parameter schätzen wollen: zu große Konfidenzintervalle
  - falls wir Tests durchführen wollen: zu geringe Power und deshalb hohe FDR
- Fazit: sehr geringer Erkenntnisgewinn.
- Eine aus statistischer Sicht „einfache“ Lösung dieses Problems besteht darin, große Stichproben zu erheben.
- Dies ist in der Praxis aus folgenden Gründen leider oft nicht so leicht:
  - zu wenige verfügbare Versuchspersonen (z.B. Patienten in Psychotherapiestudien)
  - aufwendige Erhebungsmethoden (z.B. fMRI)
  - begrenzte Zeit zur Erhebung (z.B. Bachelorarbeit)

Beispiel

- Beispiel Psychotherapiestudie:
  - Wir wollen zwei Therapieformen (Verhaltenstherapie und Gestalttherapie) in ihrer Wirksamkeit bei Zwangsstörungen vergleichen und führen zu diesem Zweck eine Stichprobenplanung durch.
  - Bei einem Signifikanzniveau von 0.005 und einer Power von mindestens 0.8 wollen wir einen kleinen Effekt (δ = 0.2) erkennen.
- Wie groß ist der Mindeststichprobenumfang?

```
Two-sample t test power calculation

              n = 667.606
              d = 0.2
      sig.level = 0.005
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Kombination mehrerer Studien

- Angenommen, wir können aus praktischen Gründen nur 80 Versuchspersonen erheben (n_1 = 40 und n_2 = 40).
- Sollten wir die Studie dann gar nicht durchführen?
- Doch: Auch wenn unsere Studie für sich genommen wenig Erkenntnisgewinn liefert, lassen sich die **Ergebnisse mehrerer kleiner Studien kombinieren**.
- Wenn wir Ergebnisse aus einzelnen Studien mit großen Stichproben vorliegen haben, ist es natürlich ebenfalls sinnvoll, diese zu kombinieren.
- Hierfür gibt es grundsätzlich zwei Möglichkeiten:
  - **Qualitatives Review**
  - **Metaanalyse**

Erster Ansatz: qualitatives Review

- Unter einem **qualitativen Review** versteht man eine **sprachliche Zusammenfassung der Ergebnisse einzelner Studien**.
- typischer Auszug aus einem Review-Artikel:

  „Zum Vergleich zwischen Verhaltenstherapie und Gestalttherapie bei Zwangsstörungen liegen inkonsistente Ergebnisse vor. Nur in zwei von zehn Studien zeigte sich eine signifikante Überlegenheit der Verhaltenstherapie."
- Die Autorinnen scheinen hier davon auszugehen, dass für ein konsistentes Bild zumindest in der Mehrzahl der 10 Einzelstudien (wenn nicht sogar in allen) eine Entscheidung für die H_{1j} getroffen werden sollte.
- Aber können wir überhaupt ein solches „konsistentes" Ergebnis erwarten, selbst wenn der gesuchte Effekt auch tatsächlich vorliegt?

Wie würden „konsistente“ Ergebnisse aussehen?

- Nehmen wir an, dass in jeder der zehn Studien 80 Personen (n_1 = 40 und n_2 = 40) untersucht wurden und die Verhaltenstherapie tatsächlich etwas besser wirkt (δ = 0.2).
- Mithilfe von R lässt sich leicht berechnen, dass jede einzelne Studie in diesem Fall eine Power von 0.03 hat.
- Die Wahrscheinlichkeit, dass alle Studien signifikant werden, beträgt somit

$$0.03^{10} \approx 0.00000000000000059049.$$

- „Inkonsistente“ Ergebnisse sind also bei Studien mit geringem Stichprobenumfang zu erwarten. Es wäre **sehr unwahrscheinlich**, dass **alle Einzeltests signifikant** ausfallen.

Zweiter Ansatz: Metaanalyse

- Unter einer **Metaanalyse** versteht man eine **quantitative Zusammenfassung der Ergebnisse der einzelnen Studien**.
- Quantitativ bedeutet, dass aus den einzelnen Studien ein **kombiniertes Konfidenz-intervall** bzw. ein **kombinierter Hypothesentest** berechnet wird.
- In unserem Beispiel würden wir also alle Stichproben der einzelnen Studien zu einer einzigen großen Stichprobe zusammenfassen und ein einziges *KI* bzw. einen einzigen Test für die Differenz in der Wirksamkeit der beiden Therapieformen berechnen.
- Mögliches Problem bei diesem Vorgehen ist, dass es in vielen Fällen nicht möglich ist, an die Rohdaten der einzelnen Studien zu kommen.
- Weiteres Problem: Was machen wir, wenn in den einzelnen Studien unterschiedliche Messinstrumente zur Erfassung der Schwere der Zwangsstörung verwendet wurden?
    - Zum Beispiel könnte in einigen Studien ein Fragebogen eingesetzt worden sein, der die Schwere der Zwangsstörung auf einer Skala von 0 bis 27 erfasst, während in den anderen Studien eine Skala von -50 bis +50 verwendet wurde.
- Wie können wir hiermit umgehen?

Verwendung von Effektgrößen bei Metaanalysen

- Lösung des Problems unterschiedlicher Messinstrumente:
  Übergang zu **Effektstärken**
- In unserem Beispiel würden wir Cohens δ verwenden, da wir uns für den Unterschied zwischen zwei Gruppen interessieren (Verhaltenstherapie und Gestalttherapie).
- Wir würden also das kombinierte KI und den kombinierten Test **auf der Basis der jeweiligen Schätzwerte für Cohens** δ aus den einzelnen Studien berechnen.

Metaanalyse mit festen Effekten

- Man unterscheidet je nach Art der statistischen Annahmen zwei Arten der Metaanalyse:
  - Metaanalyse mit festen Effekten
  - Metaanalyse mit zufälligen Effekten

- Wir werden nur die Metaanalyse mit **festen Effekten** besprechen, da sich an dieser das allgemeine Prinzip der Metaanalyse leichter veranschaulichen lässt.

Ausgangslage Metaanalyse

- Ausgangslage allgemein:
  - Insgesamt N Studien wurden unabhängig voneinander durchgeführt.
  - In jeder Studie *j* = 1, 2, ... , N wird eine uns interessierende (unbekannte) Effektstärke θ durch eine Schätzfunktion $\hat{\theta}_j$ geschätzt.
  - Wir interessieren uns für θ bzw. für Hypothesen über θ.

- Unser Beispiel:
  - Wir interessieren uns für den Unterschied zwischen Verhaltenstherapie und Gestalttherapie bei Zwangsstörungen.
  - Insgesamt 10 Studien wurden unabhängig voneinander durchgeführt.
  - In diesem Fall ist die interessierende Effektstärke Cohens δ, d.h. $\theta = \delta$.
  - In jeder der 10 Studien wurde dieses δ durch die Schätzfunktion D_j geschätzt, d.h. $\hat{\theta}_j = D_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{S_{pool,j}}$.

Annahmen der Metaanalyse (I)

- Falls die Schätzfunktionen $\hat{\theta}_j$ aus den einzelnen Studien jeweils erwartungstreu sind, gilt:

$$E(\hat{\theta}_j) = \theta \text{ für alle } j$$

- Die Varianz der einzelnen Schätzfunktionen $\hat{\theta}_j$ bezeichnen wir mit σ_j^2:

$$\sigma_j^2 = Var(\hat{\theta}_j) = SE(\hat{\theta}_j)^2$$

- Zudem wird angenommen, dass die einzelnen $\hat{\theta}_j$ normalverteilt sind. Dies ist für viele Schätzfunktionen zumindest approximativ erfüllt, falls die statistischen Annahmen in den einzelnen Studien erfüllt sind.

Annahmen der Metaanalyse (II)

- Fassen wir diese Annahmen kompakt zusammen, ergibt sich für die Metaanalyse mit festen Effekten:

$$\hat{\theta}_j \overset{ind}{\sim} N(\theta, \sigma_j^2)$$

- Die einzelnen Schätzfunktionen $\hat{\theta}_j$ seien also normalverteilt mit dem gemeinsamen Erwartungswert θ, wobei die Standardfehler bzw. die Varianzen σ_j^2 der einzelnen $\hat{\theta}_j$ von der Stichprobengröße der jeweiligen Studie *j* abhängen.

Schätzfunktion für $\boldsymbol{\theta}$

- Eine intuitiv naheliegende Schätzfunktion für θ wäre der Mittelwert der Schätzfunktionen aus den einzelnen Studien:

$$\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j$$

- Problem: Diese Schätzfunktion ist zwar erwartungstreu, aber **nicht effizient**. Dies liegt daran, dass die Varianzen σ_j^2 in den einzelnen Studien *j* unterschiedlich sein können.

Gewichtung der Schätzfunktion für $\boldsymbol{\theta}$ (I)

- Idee: Da Schätzungen aus Studien mit kleinerem geschätzten Standardfehler bzw. kleinerer geschätzter Varianz σ_j^2 genauer sind, sollten diese **stärker gewichtet** werden: Statt dem **ungewichteten Mittelwert** verwenden wir als Schätzfunktion einen **gewichteten Mittelwert**:

$$\hat{\theta} = \frac{1}{\sum_{j=1}^{N} W_j} \sum_{j=1}^{N} W_j \cdot \hat{\theta}_j$$

- In dieser Schätzfunktion wird **jede Schätzfunktion** aus den einzelnen Studien **mit dem Gewicht W_j gewichtet**. Dieses Gewicht ist definiert als

$$W_j = \frac{1}{\hat{\sigma}_j^2}$$

Gewichtung der Schätzfunktion für $\boldsymbol{\theta}$ (II)

- Schätzfunktionen aus Studien mit **größerem geschätzten Standardfehler** bzw. größerer Varianz σ_j^2 erhalten somit ein **kleineres Gewicht $\boldsymbol{W_j}$** , während Schätzungen aus Studien mit **kleinerem geschätzten Standardfehler** bzw. kleinerer Varianz σ_j^2 ein **größeres Gewicht $\boldsymbol{W_j}$** erhalten.

- Bemerkung: Die Summe dieser gewichteten Schätzfunktionen wird nicht durch N geteilt, sondern durch die Summe der Gewichte. Dies liegt daran, dass die Schätzfunktion sonst nicht erwartungstreu ist.

Gewichtung der Schätzfunktion für $\boldsymbol{\theta}$: Beispiel

- Intuitiv lässt sich die Notwendigkeit einer solchen Gewichtung an folgendem Beispiel plausibel machen:

  Wir nehmen an, dass ein unbekannter Parameter θ in zwei Studien geschätzt wurde. In der ersten Studie wurden 1000 Versuchspersonen erhoben und ein Schätzwert von 0.5 berechnet, während in der zweiten Studie nur 10 Personen untersucht wurden und ein Schätzwert von 1.5 berechnet wurde.

- Da die Schätzgenauigkeit in der ersten Studie aufgrund der großen Stichprobe viel höher als in der zweiten Studie ist, ist es plausibel anzunehmen, dass der Schätzwert $\hat{\theta}_{1\,Wert}$ aus der ersten Studie näher an dem wahren Wert θ liegt, als der Schätzwert $\hat{\theta}_{2\,Wert}$ aus der zweiten Studie (im Durchschnitt).

- Würden wir nun θ schätzen, indem wir einfach den Mittelwert der beiden Schätzwerte berechnen, würde dies nicht berücksichtigt. Als **kombinierter Schätzwert** ergäbe sich

$$\hat{\theta}_{Wert} = \frac{1}{2}(0.5 + 1.5) = 1$$

  also einfach der Wert, der genau in der Mitte zwischen den beiden einzelnen Schätzwerten liegt.

Gewichtung der Schätzfunktion für $\boldsymbol{\theta}$: Beispiel

- Nehmen wir an, wir haben in der ersten Studie einen Standardfehler der Schätzfunktion von 0.1 geschätzt und in der zweiten Studie einen Standardfehler von 1.
- Damit ergeben sich als realisierte Gewichte:

$$\boldsymbol{w_1} = \frac{1}{\widehat{\sigma}^2_{1\,Wert}} = \frac{1}{0{,}1^2} = \mathbf{100} \text{ und } \boldsymbol{w_2} = \frac{1}{\widehat{\sigma}^2_{2\,Wert}} = \frac{1}{1^2} = \mathbf{1}$$

- Der Schätzwert aus der ersten Studie erhält somit ein deutlich höheres Gewicht als der Schätzwert aus der zweiten Studie.
- Eingesetzt in die Formel für den gewichteten Mittelwert ergibt sich

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{Wert}} = \frac{1}{w_1 + w_2}\left(w_1 \cdot \hat{\theta}_{1\,Wert} + w_2 \cdot \hat{\theta}_{2\,Wert}\right) = \frac{1}{100 + 1}(100 \cdot 0.5 + 1 \cdot 1.5) = \mathbf{0.51}$$

- Der **gewichtete kombinierte Schätzwert** liegt also näher an dem genaueren Schätzwert aus der ersten Studie.

Teststatistik und Konfidenzintervall

- Man kann zeigen, dass die folgende **Teststatistik approximativ standardnormalverteilt** ist:

$$T = \frac{\left(\hat{\theta} - \theta_0\right)}{\frac{1}{\sqrt{\sum_{j=1}^{N} W_j}}} = \left(\hat{\theta} - \theta_0\right)\sqrt{\sum_{j=1}^{N} W_j}$$

wobei θ_0 der unter der Nullhypothese angenommene Effekt ist (typischerweise $\theta_0 = 0$).

- Der kritische Bereich bzw. der p-Wert kann somit mithilfe der Quantile der Standardnormalverteilung berechnet werden.

- Außerdem lässt sich ein approximatives $1 - \alpha$-Konfidenzintervall konstruieren:

$$I\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N\right) = \left[\hat{\theta} - z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{N} W_j}}, \hat{\theta} + z_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{\sum_{j=1}^{N} W_j}}\right]$$

Zwischenfazit

- Wir interessieren uns für einen Parameter θ, der in N Studien jeweils durch eine Schätzfunktion $\widehat{\theta_j}$ geschätzt wurde.
- Wir haben zwar keinen Zugriff auf die Rohdaten, können jedoch aus Daten, die über die einzelnen Studien in der Regel bekannt sind, eine gemeinsame Schätzung von θ vornehmen, bei der die Qualität einzelner Studien berücksichtigt wird.
- Wie gut θ in jeder einzelnen Studie j geschätzt werden konnte, spiegelt sich in der Varianz σ_j^2 der Schätzfunktion $\hat{\theta}_j$ wider. Diese Varianz ist u.a. dann kleiner, je größer die Stichrobe der Studie ist.
- Mit Hilfe von $\hat{\theta}_j$ und $\hat{\sigma}_j^2$ lässt sich nun ein gewichteter Mittelwert als neuer kombinierter Schätzwert für θ berechnen. Das jeweilige Gewicht ergibt sich aus $W_j = \frac{1}{\hat{\sigma}_j^2}$.
- Außerdem lässt sich mit der Standardnormalverteilung ein (bei Geltung der statistischen Annahmen) approximatives Konfidenzintervall für θ konstruieren.
- Um Hypothesen über θ zu testen kann schließlich auch noch eine Teststatistik konstruiert werden, die (bei Geltung der statistischen Annahmen) approximativ standardnormalverteilt ist.

Schätzwerte für die Varianzen σ_j^2

- Um die realisierten Gewichte

$$w_j = \frac{1}{\hat{\sigma}^2_{j\,wert}}$$

berechnen zu können, benötigen wir die Schätzwerte für die Varianzen σ_j^2 der Schätzfunktionen der einzelnen Studien. **Wie diese berechnet werden, unterscheidet sich je nach Effektgröße**.

- Für Cohens δ kann der Schätzwert hier wie folgt bestimmt werden:

$$\hat{\sigma}^2_{j\,wert} = \frac{n_{1j} + n_{2j}}{n_{1j} \cdot n_{2j}} + \frac{d_j^2}{2 \cdot \left(n_{1j} + n_{2j}\right)}$$

- Dabei sind n_{1j} und n_{2j} die Stichprobengrößen der beiden einfachen Zufallsstichproben in Studie *j*. Ist $n_{1j} = n_{2j} = n_j$ vereinfacht sich die Formel entsprechend.

- Für eine Übersicht zur Berechnung der Schätzwerte für die Varianzen σ_j^2 der Schätzfunktionen siehe z.B.: Cooper, H., Hedges, L.V. & Valentine, J.C. (2009). *The Handbook of Research Synthesis and Meta-Analysis.* New York: Sage.

Praktische Durchführung

- Wir erinnern uns an unser Beispiel: Auf der Basis von zehn Studien zum Vergleich der Wirksamkeit von Verhaltenstherapie und Gestalttherapie wollen wir ein kombiniertes *KI* berechnen und einen kombinierten Test durchführen.
- Um unsere Metaanalyse durchführen zu können, benötigen wir folgende Größen:
  - die Schätzwerte d_j für die Effektgröße Cohens δ aus den einzelnen Studien
  - die zugehörigen Schätzwerte $\hat{\sigma}^2_{j\,wert}$ für die unbekannten Varianzen
- Häufig sind in den einzelnen Artikeln keine Schätzwerte für die Effektgrößen, sondern lediglich die Ergebnisse der *t*-Tests (für unabhängige Stichproben) angegeben bzw. komplett andere Analysen.
- In diesem Fall müssen die Effektgrößen per Hand berechnet werden. Die Berechnung von Cohens d in Studie *j* bei aus dem realisierten *t*-Wert lautet:

$$d_j = t_j \cdot \sqrt{\frac{n_{1j} + n_{2j}}{n_{1j} \cdot n_{2j}}}$$

- Formel für die Berechnung von Cohens d aus deskriptiv-statistischen Werten:

$$d_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s_{pool,j}}$$

Praktische Durchführung: Daten

	t	ν	p-Wert
Studie 1	1.034	78	0.304
Studie 2	-0.706	78	0.482
Studie 3	-0.191	78	0.849
Studie 4	2.776	78	0.007
Studie 5	0.587	78	0.558
Studie 6	0.708	78	0.481
Studie 7	1.977	78	0.052
Studie 8	0.729	78	0.468
Studie 9	2.020	78	0.047
Studie 10	1.412	78	0.162

- Auch wenn es in der Regel schwierig ist, an Rohdaten einzelner Studien zu kommen sind doch mindestens Angaben zu den verwendeten statistischen Verfahren in Fachartikeln verfügbar.
- Für unser Beispiel liegen entsprechend die nebenstehenden Ergebnisse aus den einzelnen Studien vor.

Beispielhafte Berechnung eines realisierten Gewichts

- Aus diesen Ergebnissen können wir mit den oben genannten Formeln für jede Studie den Schätzwert d_j für Cohens δ und das realisierte Gewicht w_j berechnen.
- Für die erste Studie ist dies beispielhaft dargestellt:

  Wir erinnern uns, dass für jede Studie $n_1 = n_2 = 40$ gilt. In der ersten Studie wurde ein realisierter *t*-Wert von $t = 1.034$ beobachtet. Damit ergibt sich

$$d_1 = t_1 \cdot \sqrt{\frac{n_{11} + n_{21}}{n_{11} \cdot n_{21}}} = 1.034 \cdot \sqrt{\frac{40 + 40}{40 \cdot 40}} = 0.231$$

  und für den Schätzwert $\hat{\sigma}^2_{1\,wert}$:

$$\hat{\sigma}^2_{1\,wert} = \frac{n_{11} + n_{21}}{n_{11} \cdot n_{21}} + \frac{d_1^2}{2 \cdot (n_{11} + n_{21})} = \frac{40 + 40}{40 \cdot 40} + \frac{0.231^2}{2 \cdot (40 + 40)} = 0.050$$

  Für das realisierte Gewicht der Studie 1 ergibt sich also

$$w_1 = \frac{1}{\hat{\sigma}^2_{1\,wert}} = \frac{1}{0.05} = 20$$

Tabellarische Zusammenfassung der Schätzwerte d_j und w_j

	d_j	w_j
Studie 1	0.231	20
Studie 2	-0.158	20
Studie 3	-0.043	20
Studie 4	0.621	19.231
Studie 5	0.131	20
Studie 6	0.158	20
Studie 7	0.442	19.608
Studie 8	0.163	20
Studie 9	0.452	19.608
Studie 10	0.316	19.608

- Analog kann Cohens d_j und $w_j = \frac{1}{\widehat{\sigma}^2_{j\,wert}}$ für alle Studien berechnet werden.
- Warum sind die Gewichte in diesem Fall so ähnlich?

Berechnung des kombinierten Schätzwerts $\boldsymbol{d}$

	d_j	w_j	$d_j * w_j$
Studie 1	0.231	20	4.62
Studie 2	-0.158	20	-3.16
Studie 3	-0.043	20	-0.86
Studie 4	0.621	19.231	11.94
Studie 5	0.131	20	2.62
Studie 6	0.158	20	3.16
Studie 7	0.442	19.608	8.67
Studie 8	0.163	20	3.26
Studie 9	0.452	19.608	8.86
Studie 10	0.316	19.608	6.20
$\sum_{j=1}^{N}$		198.055	45.308

- Aus diesen Größen können wir den **kombinierten Schätzwert** $\boldsymbol{d}$ für δ berechnen:

$$d = \frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j \cdot d_j = \frac{45.308}{198.055} = 0.229$$

wobei $\sum_{j=1}^{N} w_j = 198.055$ und $\sum_{j=1}^{N} w_j \cdot d_j = 45.308$ ist.

Berechnung der Realisation der Teststatistik und des konkreten 0.95-Konfidenzintervalls

- Wir erinnern uns: Die interessierende Effektstärke θ ist in unserem Fall Cohens δ. Für dieses δ können wir nun einen Test durchführen und ein KI berechnen.

- Berechnung der Realisation der Teststatistik für die H_0: $\theta = \delta = 0$:

$$t = \left(\hat{\theta}_{wert} - \theta_0\right)\sqrt{\sum_{j=1}^{N} w_j} = (0.229 - 0)\sqrt{198.055} = 3.222$$

- Berechnung eines 95%-Konfidenzintervalls für $\theta = \delta$:

$$\left[\hat{\theta}_{wert} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\sum_{j=1}^{N} w_j}}; \hat{\theta}_{wert} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\sum_{j=1}^{N} w_j}}\right] = \left[0.229 - \frac{1.96}{\sqrt{198.055}}; 0.229 + \frac{1.96}{\sqrt{198.055}}\right] = [0.090; 0.368]$$

Interpretation der Ergebnisse

- Ergebnis des statistischen Tests bei einem Signifikanzniveau von $\alpha = 0.005$:

```
> qnorm(0.9975)
[1] 2.807034
```

$$t = 3.222 \in K_T = [-\infty; -2.807] \cup [+2.807; +\infty]$$

- konkretes 0.95-Konfidenzintervall:

$$[0.090; 0.368]$$

- Interpretation:
  - Wir entscheiden uns für die H_1 und damit dafür, dass sich die beiden Therapien in ihrer Wirksamkeit unterscheiden.
  - Wir können im Sinne der Klassifikation von Cohen von einem kleinen Effekt ausgehen, da die plausiblen Werte für Cohens δ zwischen 0.090 und 0.368 liegen.

Zusammenfassung

- Zusammengesetzte Hypothesentests:
  - „oder“-Verknüfung in der H_1: **Korrektur** der Signifikanzniveaus der einzelnen Tests oder entsprechender **Omnibustest** (siehe VL im weiteren Verlauf des Semesters).
  - „und“-Verknüfung in der H_1: **keine Korrektur** der Signifikanzniveaus.
- Mehrere nicht zusammengesetzte Hypothesentests aus einer oder mehreren Stichproben mit unterschiedlichen Hypothesen: **in der Regel keine Korrektur (aber niedrige FDR sicherstellen!)**
- Mehrere Hypothesentests aus verschiedenen Stichproben mit identischen Hypothesen: **Metaanalyse**