13. Vorlesung Statistik I
False Discovery Rate und Annahmen statistischer Verfahren

Überblick

- Wir werden uns in der heuten Vorlesung mit den folgenden Fragen beschäftigen:
  - Warum ist es wichtig, dass ein statistischer Hypothesentest sowohl ein geringes Signifikanzniveau als auch eine hohe Power aufweist?
  - Wie hoch sind Signifikanzniveau und Power einer typischen psychologischen Studie?
  - Welche Konsequenzen hat dies?

- Außerdem betrachten wir zum Schluss noch, welche Rolle die Annahmen unserer statistischen Verfahren spielen.

Ausgangssituation

Signifikanzniveau und Power in psychologischen Studien I

- In psychologischen Studien werden praktisch ausschließlich Hypothesentests verwendet. Intervallschätzung spielt (zu unrecht) eine sehr viel geringere Rolle.
- Allgemeine Konvention ist hierbei ein Signifikanzniveau von $\alpha = 0.05$.
- Sehr häufig wird keine Stichprobenplanung durchgeführt und die Power der Hypothesentests ignoriert.
- Da eine Entscheidung für die H_0 in der Regel eine Entscheidung für die Falsifikation der überprüften psychologischen Theorie impliziert, sind Studien mit signifikantem Ergebnis - also mit Entscheidung für die H_1 - aus inhaltlicher Sicht interessanter.
- Aus diesem Grund werden in psychologischen Fachzeitschriften praktisch ausschließlich Studien mit signifikanten Ergebnissen veröffentlicht (= **Publikationsbias**).

Signifikanzniveau und Power in psychologischen Studien II

- Für Studien mit ungerichteten Zweistichproben t-Tests für unabhängige Stichproben ist der Median der Stichprobengröße pro Stichprobe in psychologischen Studien gleich 20 (Bakker et. al. 2012).
- Der durchschnittliche Schätzwert für Cohen‘s δ in psychologischen Studien ist 0.5 (Bakker et. al. 2012). Der wahre Wert ist vermutlich niedriger, weil natürlich nur publizierte Studien berücksichtigt werden können, also Studien mit größtenteils signifikanten Ergebnissen.
- Die durchschnittliche Power für t-Tests unter diesen Voraussetzungen ist:

```
Two-sample t test power calculation

              n = 20
              d = 0.5
      sig.level = 0.05
          power = 0.337939
    alternative = two.sided

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

- Da δ vermutlich kleiner als 0.5 ist, ist die tatsächliche Power vermutlich noch geringer.

Signifikanzniveau und Power in psychologischen Studien III

- Auch allgemein (also über die t-Tests hinaus) wird die durchschnittliche Power der in psychologischen Studien verwendeten Hypothesentests auf ungefähr 0.35 geschätzt (Bakker et. al. 2012).

- Wir werden uns überlegen, welche Konsequenzen ein Signifikanzniveau von $\alpha = 0.05$ und eine durchschnittliche Power von $1 - \beta \approx 0.35$ für die Beurteilung signifikanter Ergebnisse haben.

False Discovery Rate

Ausgangssituation

Ausgangssituation:

- Wir betrachten N psychologische Studien, z.B. könnte N die Anzahl aller jemals durchgeführten psychologischen Studien sein.
- In jeder dieser N Studien wird ein statistischer Hypothesentest durchgeführt.
- Der Anteil der Studien, in denen die H_0 wahr ist, ist ρ (griechisch: rho). Absolut betrachtet ist die H_0 also in $\rho \cdot N$ Studien wahr, ρ wird auch **Basisrate** genannt.
- Der Anteil der Studien, in denen die H_1 wahr ist, ist somit $1 - \rho$. Absolut betrachtet ist die H_1 also in $(1 - \rho) \cdot N$ Studien wahr.

Beispiel:

- Wir betrachten 1000 psychologische Studien. Damit ist $N = 1000$.
- Falls in 40% dieser Studien die H_0 wahr ist, ist $\rho = 0.4$
- Es ist also in
  - $\rho \cdot N = 0.4 \cdot 1000 = 400$ Studien die H_0 wahr und in
  - $(1 - \rho) \cdot N = (1 - 0.4) \cdot 1000 = 0.6 \cdot 1000 = 600$ Studien die H_1 wahr.

Annahmen I

Wir treffen die folgenden Annahmen um einfacher weiterrechnen zu können:

- Die Hypothesentests aus den verschiedenen Studien sind unabhängig voneinander.
- Das Signifikanzniveau α des Hypothesentests ist in allen Studien gleich und entspricht in allen Studien der Wahrscheinlichkeit für einen Fehler 1. Art.
- Die Power $1 - \beta$ des statistischen Hypothesentests ist in allen Studien gleich und entspricht in allen Studien der Wahrscheinlichkeit dafür, keinen Fehler 2. Art zu machen.
- Keine Verzerrung, d.h. Annahmen der statistischen Hypothesentests sind erfüllt, einfache Zufallsstichproben liegen vor (d.h. Repräsentativität), keine Fehler bei Datenerhebung und Auswertung, keine Datenfälschung, keine statistischen Tricks, um signifikante Ergebnisse zu bekommen, etc.

Annahmen II

- Bemerkung I: Die Annahme gleicher Power $1 - \beta$ für alle Hypothesentests ist natürlich unrealistisch. Sie führt jedoch zu den gleichen Ergebnissen wie die Annahme, dass die Power in den Studien unterschiedlich und im Durchschnitt $1 - \beta$ ist. Diese Annahme können wir also ohne Probleme treffen.

- Bemerkung II: Auch die vierte Annahme ist unrealistisch. Aber: Wir können uns so zumindest überlegen, wie die Situation im besten Fall aussehen könnte. Falls sich schon hier herausstellt, dass es ernstzunehmende Probleme gibt, ist die Situation bei Vorliegen von Verzerrungen noch ungünstiger.

Anzahl falsch positiver Entscheidungen I

- Wir betrachten nur signifikante Ergebnisse, also Entscheidungen für die H_1.
- Das Signifikanzniveau α entspricht der Wahrscheinlichkeit dafür, dass wir uns fälschlicherweise für die H_1 entscheiden, obwohl die H_0 wahr ist.
- Falsche Entscheidungen für die H_1 nennt man auch **falsch positive Entscheidungen**.
- Falsch positive Entscheidungen können nur in denjenigen Studien getroffen werden, in denen die H_0 tatsächlich wahr ist.
- Dies ist in $\rho \cdot N$ der N Studien der Fall.
- Daraus folgt, dass wir uns im Durchschnitt in $\alpha \cdot \rho \cdot N$ Studien fälschlicherweise für die H_1 entscheiden (frequentistisch gesehen).
- Diese Größe – die durchschnittliche Anzahl falsch positiver Entscheidungen – nennen wir fp:

$$fp = \alpha \cdot \rho \cdot N$$

Anzahl falsch positiver Entscheidungen II

Beispiel:

- $N = 1000$ Studien
- $\rho = 0.4$, d.h. in 40% dieser Studien ist die H_0 wahr.
- Absolut gesehen ist dann in 400 Studien die H_0 wahr. In diesen Studien können wir falsch positive Entscheidungen treffen.
- Falls das Signifikanzniveau in allen Hypothesentests gleich $\alpha = 0.05$ ist, haben wir die Garantie, dass wir uns im Durchschnitt nur in 5% dieser 400 Studien, also in 20 Studien, für die H_1 entscheiden (obwohl ja die H_0 gilt):

$$fp = \alpha \cdot \rho \cdot N = 0.05 \cdot 0.4 \cdot 1000 = 20$$

Anzahl richtig positiver Entscheidungen I

- Die Power $1 - \beta$ entspricht der Wahrscheinlichkeit dafür, dass wir uns für die H_1 entscheiden, falls die H_1 wahr ist.
- Richtige Entscheidungen für die H_1 nennt man auch **richtig positive Entscheidungen**.
- Richtig positive Entscheidungen können nur in denjenigen Studien getroffen werden, in denen die H_1 tatsächlich wahr ist.
- Dies ist in $(1 - \rho) \cdot N$ der N Studien der Fall.
- Daraus folgt, dass wir uns im Durchschnitt in $(1 - \beta) \cdot (1 - \rho) \cdot N$ Studien richtigerweise für die H_1 entscheiden (frequentistisch gesehen).
- Diese Größe – die durchschnittliche Anzahl richtig positiver Entscheidungen – nennen wir rp:

$$rp = (1 - \beta) \cdot (1 - \rho) \cdot N$$

Anzahl richtig positiver Entscheidungen II

Beispiel:

- $N = 1000$ Studien
- $\rho = 0.4$, d.h. in 40% dieser Studien ist die H_0 wahr und in 60% der Studien die H_1
- Absolut gesehen ist dann in 600 Studien die H_1 wahr. In diesen Studien können wir richtig positive Entscheidungen treffen.
- Falls die Power in allen Hypothesentests gleich $1 - \beta = 0.95$ ist, haben wir die Garantie, dass wir uns im Durchschnitt in 95% dieser 600 Studien, also in 570 Studien richtigerweise für die H_1 entscheiden:

$$rp = (1 - \beta) \cdot (1 - \rho) \cdot N = 0.95 \cdot 0.6 \cdot 1000 = 570$$

Gesamtanzahl positiver Entscheidungen

- Die durchschnittliche Anzahl **aller Entscheidungen für die $\boldsymbol{H_1}$** entspricht der Summe der durchschnittlichen **Anzahl der falschen** Entscheidungen für die H_1 $(= fp)$ und der durchschnittlichen **Anzahl der richtigen** Entscheidungen für die H_1 $(= rp)$.

- Diese Größe – die durchschnittliche Gesamtanzahl positiver Entscheidungen – nennen wir gp:

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

False Discovery Rate

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

$$FDR = \frac{fp}{gp}$$

- Die FDR ist die zentrale Größe bei der Beurteilung signifikanter Ergebnisse:
  - falls die FDR klein ist, bedeutet dies, dass im Durchschnitt ein großer Anteil der signifikanten Ergebnisse richtig ist.
  - falls die FDR groß ist, bedeutet dies, dass im Durchschnitt ein großer Anteil der signifikanten Ergebnisse falsch ist.
- Die FDR ist unabhängig von der Anzahl N der Studien:

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

Zähler und Nenner durch N teilen

Einflussgrößen auf die False Discovery Rate

Einflussgrößen auf die FDR

- Aus

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

ergibt sich, dass die FDR von den folgenden Größen abhängt:

  - dem Signifikanzniveau α,
  - der Power $1 - \beta$,
  - der Basisrate ρ.

- Wie genau diese Abhängigkeit aussieht, werden wir uns auf den folgenden Folien ansehen.

FDR und Signifikanzniveau

- Allgemein ergibt sich bei gegebenem ρ und $1 - \beta$ für $\alpha \neq 0$ aus

Zähler und Nenner durch α teilen

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

dass die **FDR umso höher ist, je größer das Signifikanzniveau α ist**.

- Extremfälle: $FDR = \frac{\rho}{\rho + (1-\beta) \cdot (1-\rho)}$ bei $\alpha = 1$ und $FDR = 0$ bei $\alpha = 0$.
- Wichtige Erkenntnisse:
  - Unsere Hypothesentests sollten ein niedriges Signifikanzniveau aufweisen.
  - Aber: Ein niedriges Signifikanzniveau alleine ist keine Garantie dafür, dass die FDR niedrig ist!

FDR und Power

- Allgemein ergibt sich bei gegebenem ρ und α aus

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

dass die **FDR umso höher ist, je niedriger die Power $1 - \beta$ ist**.

- Extremfälle: $FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + 1 - \rho}$ bei $1 - \beta = 1$ und $FDR = 1$ bei $1 - \beta = 0$.
- Wichtige Erkenntnisse:
  - Unsere Hypothesentests sollten eine hohe Power aufweisen.
  - Auch bei einem sehr geringen Signifikanzniveau kann eine geringe Power zu einer sehr hohen FDR führen, im schlimmsten Fall sogar zu einer FDR nahe 1.
  - Das Signifikanzniveau bestimmt, wie klein die FDR für eine gegebene Basisrate im besten Fall werden kann (d.h. im Extremfall $1 - \beta = 1$, siehe oben)

FDR und Basisrate I

- Allgemein ergibt sich bei gegebenem α und $1-\beta$ für $\rho \neq 0$ aus

Zähler und Nenner durch ρ teilen

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1-\beta) \cdot (1-\rho)} = \frac{\alpha}{\alpha + (1-\beta) \cdot \frac{(1-\rho)}{\rho}} = \frac{\alpha}{\alpha + (1-\beta) \cdot \left(\frac{1}{\rho} - 1\right)}$$

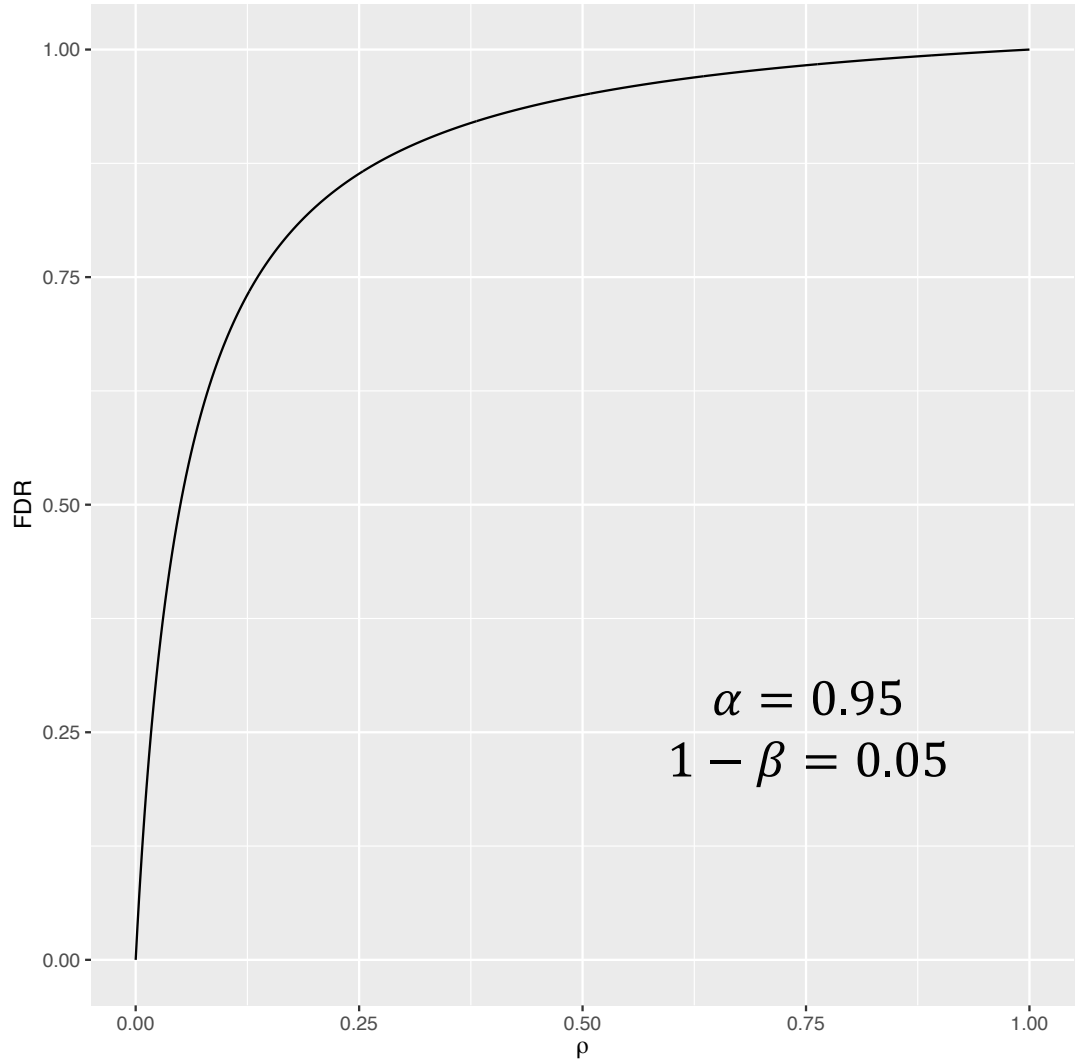
dass die **FDR umso höher ist, je höher die Basisrate ρ ist**.

- Extremfälle: $FDR = 1$ bei $\rho = 1$ und $FDR = 0$ bei $\rho = 0$.

- Sehr wichtig: Wie stark die FDR mit der Basisrate steigt, hängt von α und $1-\beta$ ab!

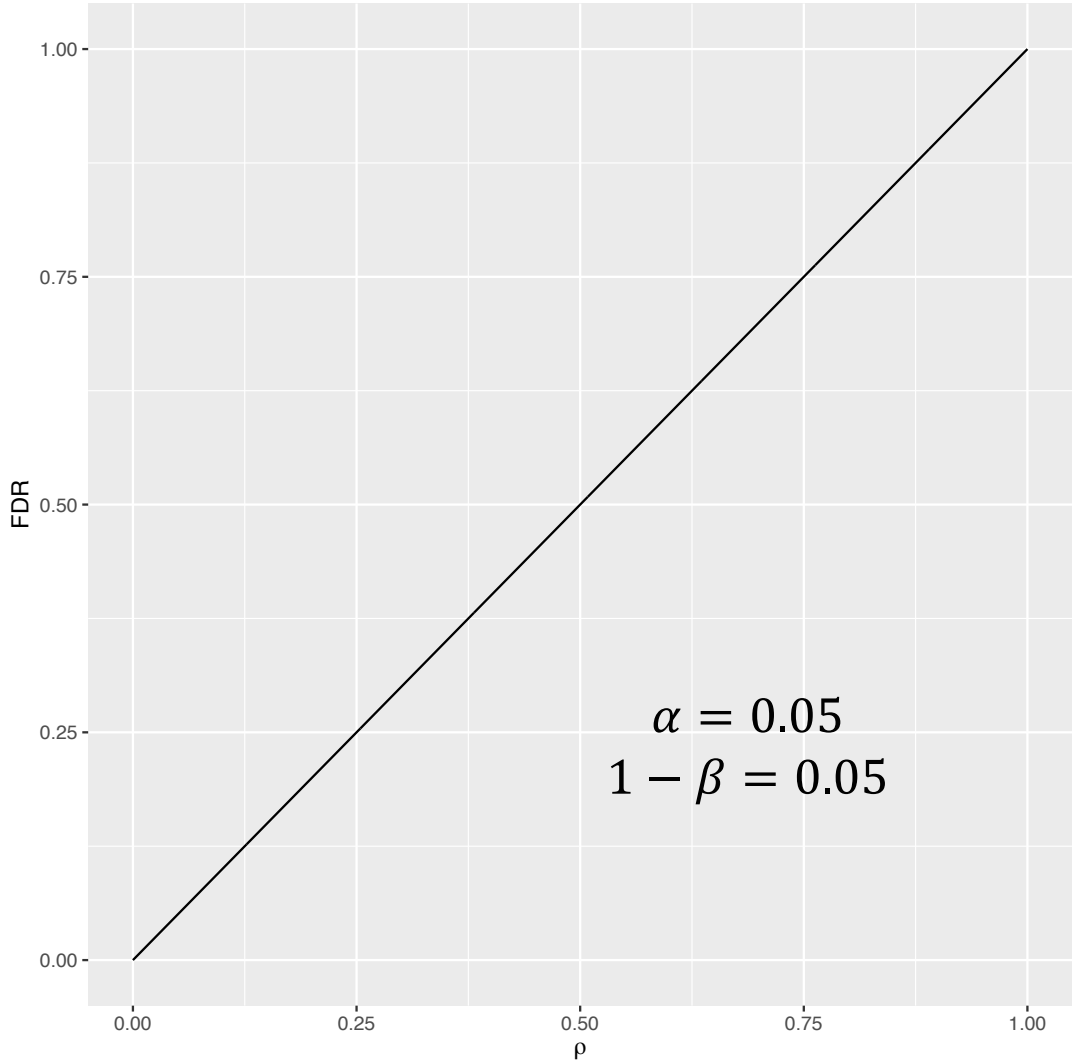
FDR und Basisrate II

- Bei einem hohen Signifikanzniveau und einer geringen Power wächst die FDR sehr schnell mit der Basisrate ρ. Schon bei einer geringen Basisrate von $\rho = 0.1$ wären hier im Durchschnitt weit über die Hälfte der signifikanten Ergebnisse falsch:



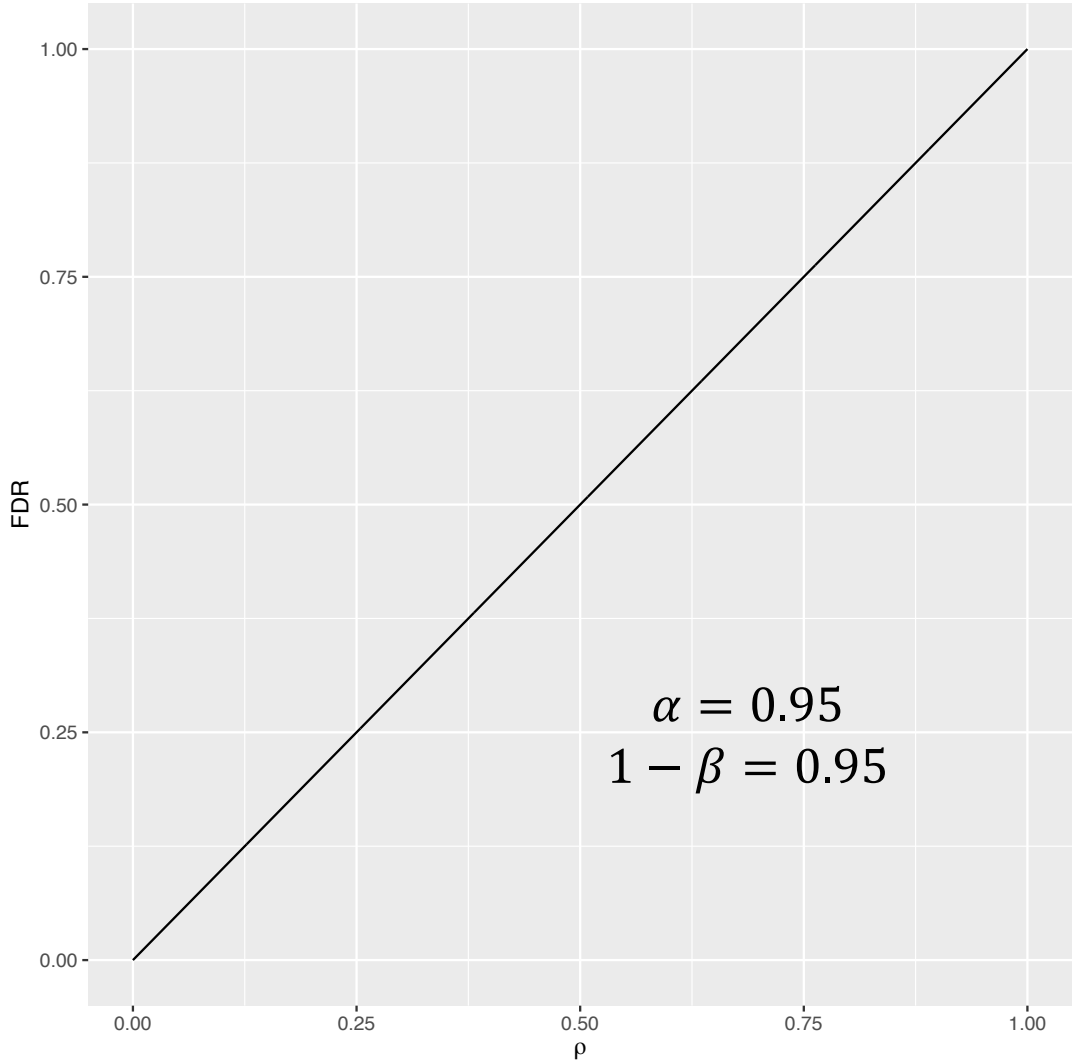
FDR und Basisrate III

- Bei einem geringen Signifikanzniveau α und einer geringen Power $1 - \beta$ wächst die FDR immer noch schnell mit der Basisrate ρ:



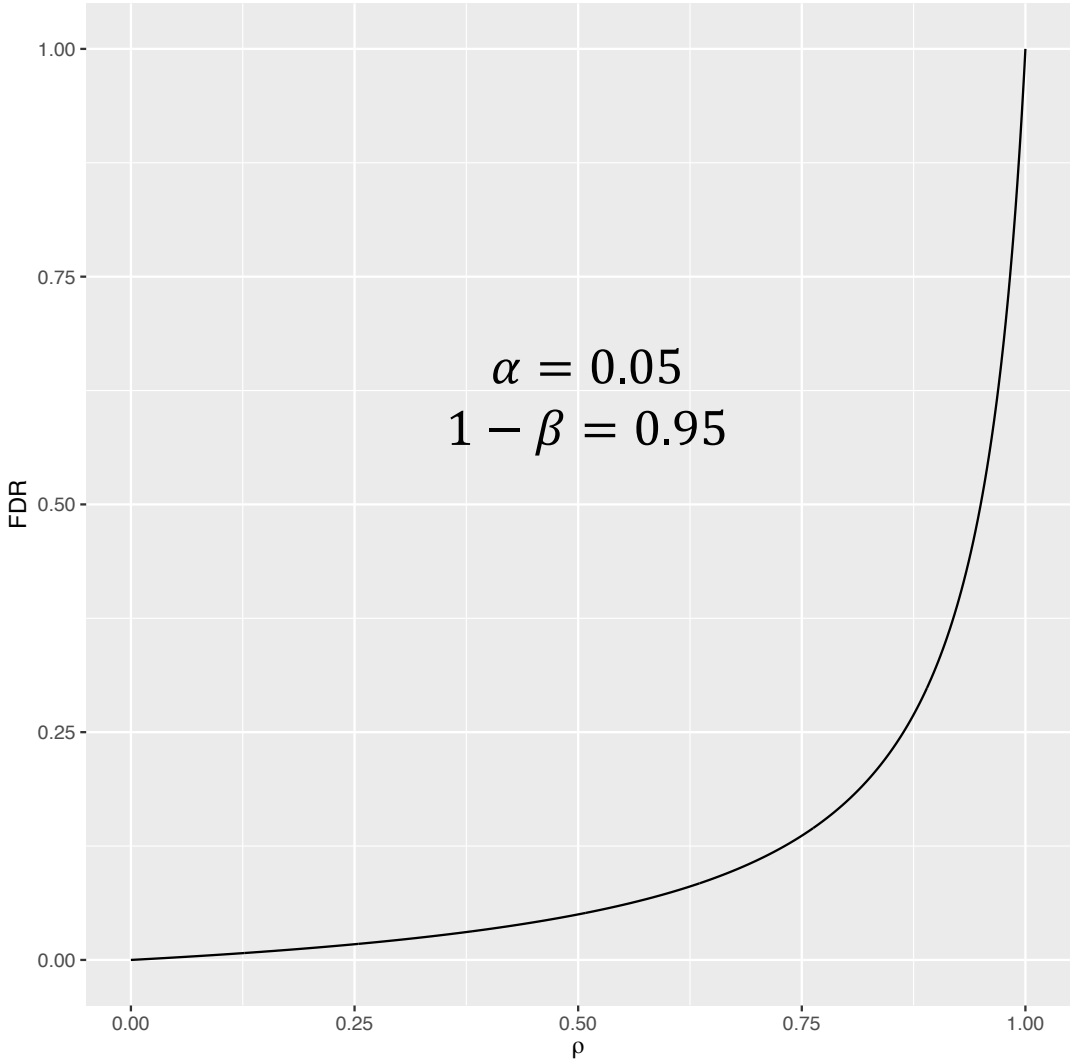
FDR und Basisrate IV

- Gleiches gilt für hohes Signifikanzniveau α und eine hohe Power $1 - \beta$:



FDR und Basisrate V

- Bei einem geringen Signifikanzniveau und einer hohen Power wächst die FDR sehr langsam mit der Basisrate ρ. Auch bei einer ungünstigen Basisrate von $\rho = 0.9$ wären hier im Durchschnitt nur ca. 30% der signifikanten Ergebnisse falsch:



FDR und Basisrate VI

- Wichtig: **Die Basisrate können wir nicht kontrollieren.** Wir können nie wissen, wie viele unserer Alternativhypothesen wahr sind. Ob die Alternativhypothese stimmt wollen wir mithilfe unseres Hypothesentests ja erst herausfinden.

- Wir können uns gegen einen zu starken Einfluss der Basisrate auf die FDR aber durch ein niedriges Signifikanzniveau und eine hohe Power unserer Hypothesentests absichern.

- Dies ist der Grund dafür, warum statistische Hypothesentests nur dann zu verlässlichen Ergebnissen führen, falls sie **sowohl ein niedriges Signifikanzniveau als auch eine hohe Power** aufweisen.

False Discovery Rate in der Psychologie

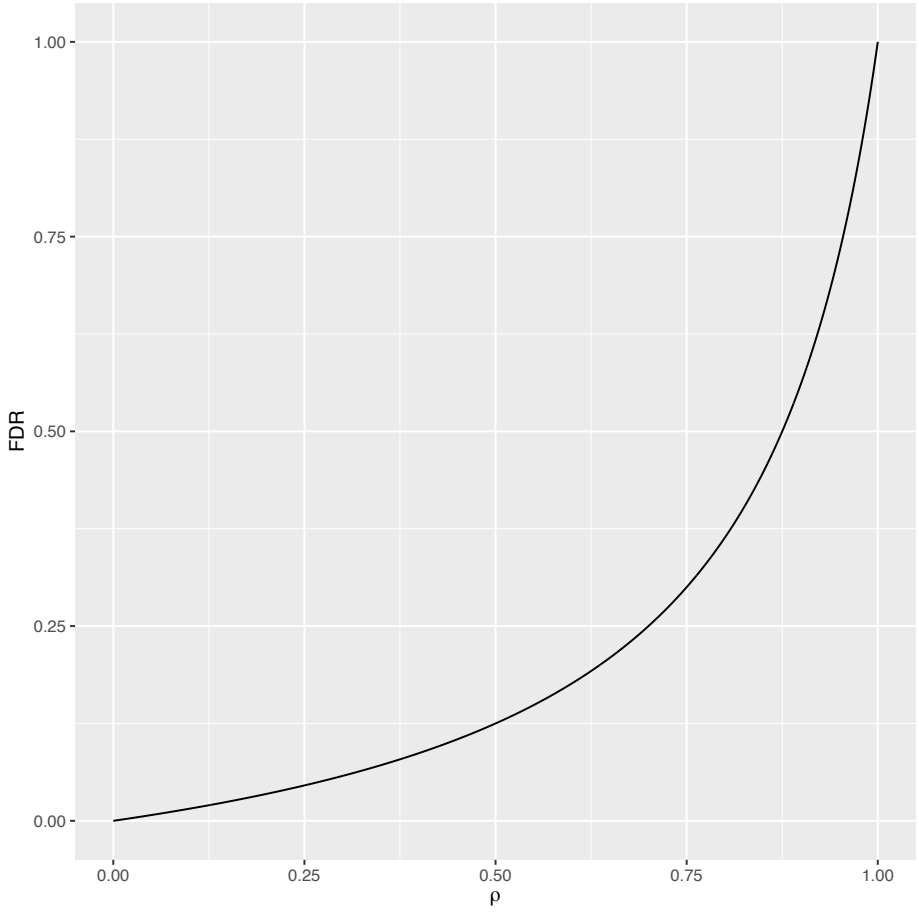
FDR in der Psychologie

- Zur Erinnerung - Ausgangssituation in der Psychologie:
  - $\alpha = 0.05$
  - $1 - \beta = 0.35$ (wohlwollende Schätzung, wahrscheinlich niedriger)
- Damit folgt für die FDR, also den durchschnittlichen Anteil falsch positiver Ergebnisse:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.05 \cdot \rho}{0.05 \cdot \rho + 0.35 \cdot (1 - \rho)}$$

- Problem: Wir können die genaue FDR nur bestimmen, falls wir die Basisrate ρ, d.h. den Anteil an wahren Nullhypothesen, kennen. Diese kennen wir natürlich nicht.
- Zwei Möglichkeiten, um die Situation zu beurteilen:
  - Theoretisch: Bestimmen wie stark die FDR in diesem Fall von der Basisrate abhängt.
  - Empirisch: Schätzung der FDR.

Verschiedene Basisraten I



- Sehr wohlwollende Abschätzung mit $\rho = 0.1$:

$$FDR = \frac{0.05 \cdot \rho}{0.05 \cdot \rho + 0.35 \cdot (1 - \rho)}$$
$$= \frac{0.05 \cdot 0.1}{0.05 \cdot 0.1 + 0.35 \cdot 0.9} \approx 0.02$$

- Eher konservative Abschätzung mit $\rho = 0.9$:

$$FDR = \frac{0.05 \cdot \rho}{0.05 \cdot \rho + 0.35 \cdot (1 - \rho)}$$
$$= \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.35 \cdot 0.1} \approx 0.56$$

Verschiedene Basisraten II

- Wir sehen also: Starke Schwankungen der FDR je nachdem wie hoch die Basisrate ist. Ist die wahre Basisrate gleich 0.1, ist die Situation ok. Ist die wahre Basisrate aber gleich 0.9, sind im Durchschnitt mehr als die Hälfte der Ergebnisse in psychologischen Fachzeitschriften falsch.

- Aus theoretischer Sicht sollten statistische Methoden sicherstellen, dass nicht kontrollierbare Größen wie die Basisrate keinen so großen Einfluss auf die Qualität der Ergebnisse haben.

- Dies ist bei einem Signifikanzniveau von $\alpha = 0.05$ und einer Power von $1 - \beta = 0.35$ nicht der Fall.

- Außerdem: Dies alles gilt unter der Annahme, dass keine Verzerrungen vorliegen. Diese würden die FDR für gegebene Basisraten noch weiter erhöhen. Dies ist also ein best-case Szenario.

- Fazit aus theoretischer Sicht also:
  - Starke Abhängigkeit der in der Psychologie verwendeten statistischen Methoden von der Basisrate, d.h. von einer unkontrollierbaren Größe.

Empirische Schätzung der FDR I

- Aber: Es könnte natürlich sein, dass die psychologischen Theorien so gut sind, dass der Anteil der wahren Nullhypothesen so niedrig ist, dass die FDR trotzdem in einem annehmbaren Bereich liegt.
- Wie könnte man die FDR in der Psychologie empirisch schätzen?
- Idee von Nosek et. al. (2015):
  - Zufällige Auswahl von 97 Studien mit signifikantem Ergebnis aus hochrangigen psychologischen Fachzeitschriften aus dem Jahr 2008.
  - Replikation (d.h. nochmaliges Durchführen) dieser Studien in größeren Stichproben mit höherer Power.
  - Schätzwert für die FDR = Anteil der Studien, bei denen die Replikation nicht wieder zu einem signifikanten Ergebnis führt.

Empirische Schätzung der FDR II

- (Etwas vereinfachtes) Ergebnis: Nur in 37% der Studien ergab sich in der Replikation wieder ein signifikantes Ergebnis.
- Die FDR kann somit in der Psychologie auf 1 – 0.37 = 0.63 geschätzt werden.
- Auf der Basis dieser Ergebnisse müssen wir also davon ausgehen, dass **im Durchschnitt 63% aller veröffentlichten signifikanten Ergebnisse falsch** sind.
- Dies hat die sogenannte **Replikationskrise** in der Psychologie mit ausgelöst.
- Selbstverständlich ist der Wert 0.63 nur ein Punktschätzwert und natürlich kann die Methodik von Nosek et. al. (2015) kritisiert werden. Aber die Daten sprechen durchaus dafür, dass mehr als die Hälfte der veröffentlichten signifikanten Ergebnisse falsch sein könnten.
- Zudem gibt es mittlerweile weitere Studien, die zu ähnlichen Schätzwerten kommen (Klein et. al. 2018).

Mögliche Lösungen

Mögliche Lösungsansätze

- Problem: Wir können

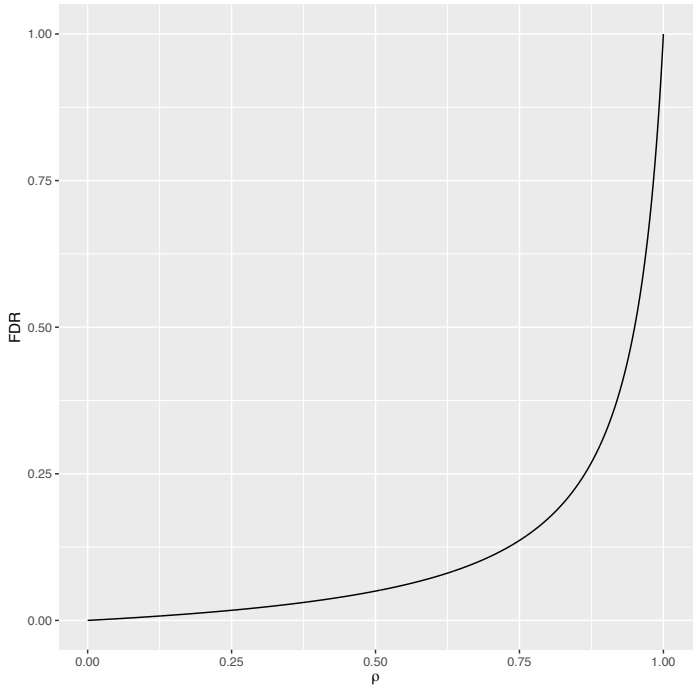
$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)}$$

  nicht direkt kontrollieren, da wir die Basisrate ρ nicht kontrollieren können.

- Die verschiedenen Lösungsansätze müssen also an den von uns kontrollierbaren Einflussgrößen auf die FDR ansetzen:
  - dem Signifikanzniveau α,
  - der Power $1 - \beta$.
- Wir werden einige Lösungsansätze diskutieren.

Erster Ansatz I

- Erster Lösungsansatz:
  - $\alpha = 0.05$ beibehalten.
  - Größere Stichproben und somit höhere Power $1 - \beta$.
- Aus statistischer Sicht ist dies eine zunächst naheliegende Lösung, da wir gesehen haben, dass der Einfluss der Basisrate auf die FDR für geringes α und hohes $1 - \beta$ weniger stark ist:



Erster Ansatz II

- In dieser Hinsicht ist in den letzten Jahren auch ein leichter positiver Trend erkennbar. Immer mehr Fachzeitschriften verlangen größere Stichproben und Poweranalysen.

- Aber: Eine Power von $1 - \beta = 0.95$ kann bei einem Signifikanzniveau von $\alpha = 0.05$ immer noch zu niedrig sein.

- Bei einer Basisrate von $\rho = 0.9$ ergibt sich hier:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.95 \cdot 0.1} \approx 0.32$$

- Die FDR ist bei $\alpha = 0.05$ sogar im positiven Extremfall einer Power von $1 - \beta = 1$ bei einer Basisrate von $\rho = 0.9$ noch sehr hoch:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 1 \cdot 0.1} \approx 0.31$$

Zweiter Ansatz

- In vielen Fällen ist es sehr schwierig bis unmöglich, große Stichproben zu erheben. Beispiele:
  - Babystudien in der Entwicklungspsychologie.
  - Studien mit aufwendigen Bildgebungsverfahren in der Neuropsychologie.
- Daher zweiter Lösungsansatz:
  - Viele einzelne Studien mit identischem Versuchsaufbau mit jeweils kleinen Stichproben durchführen, aber **nicht interpretieren**.
  - Kombinierte Auswertung der Daten dieser Studien mithilfe metaanalytischer statistischer Methoden, die eine höhere Power aufweisen (siehe Statistik 2).
- Wichtige Voraussetzung: Alle Einzelstudien werden veröffentlicht.
- Auch in diese Richtung ist ein leichter positiver Trend beobachtbar:
  Many Labs Studien, Open Data Repositories, Präregistrierung.

Dritter Ansatz I

- Dritter Lösungsansatz (Benjamin et. al. 2017):
  - $\alpha = 0.005$ statt $\alpha = 0.05$ als neue Konvention für das Signifikanzniveau.
- Warum könnte dies sinnvoll sein?
- Zunächst: Je niedriger das Signifikanzniveau, desto niedriger die FDR bei gleichbleibender Power und gleichbleibender Basisrate.
- Aber: Je niedriger das Signifikanzniveau, desto geringer die Power. Eine geringere Power erhöht wiederum die FDR.
- Es zeigt sich aber, dass der Effekt der Senkung des Signifikanzniveaus auf die FDR in der Regel größer ist als der Effekt der damit einhergehenden Senkung der Power.
- Zudem verringert sich die untere Grenze der FDR für gegebene Basisraten (siehe Folie 20) bei einer Power von 1.

Dritter Ansatz II

- Powerberechnung für einen ungerichteten Zweistichproben t-Test für unabhängige Stichproben bei einem Signifikanzniveau von $\boldsymbol{\alpha = 0.05}$ und einem Effekt unter der H_1 von $\delta_{H_1} = 0.5$ bei einer Stichprobengröße von $n = 20$ pro Stichprobe:

```
Two-sample t test power calculation

              n = 20
              d = 0.5
      sig.level = 0.05
          power = 0.337939
    alternative = two.sided

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

- Als FDR für eine ungünstige Basisrate von $\rho = 0.9$ ergibt sich in diesem Fall:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.34 \cdot 0.1} \approx 0.57$$

Dritter Ansatz III

- Powerberechnung für einen ungerichteten Zweistichproben t-Test für unabhängige Stichproben bei einem Signifikanzniveau von $\boldsymbol{\alpha = 0.005}$ und einem Effekt unter der H_1 von $\delta_{H_1} = 0.5$ bei einer Stichprobengröße von $n = 20$ pro Stichprobe:

```
Two-sample t test power calculation

              n = 20
              d = 0.5
      sig.level = 0.005
          power = 0.09572436
    alternative = two.sided

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

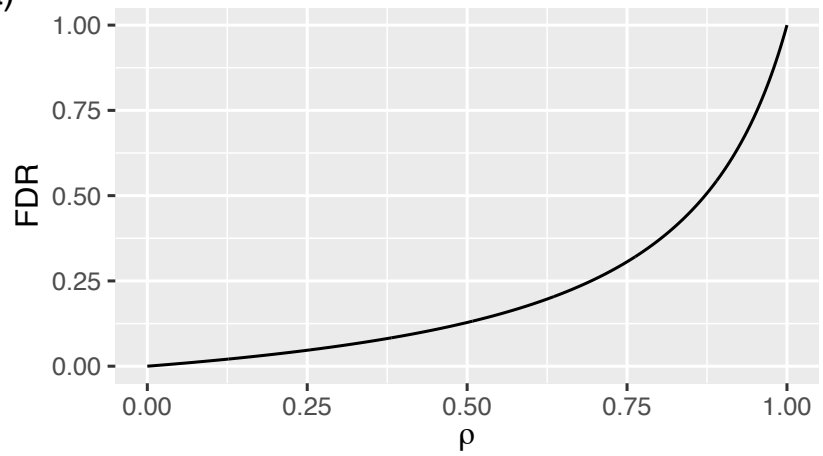
- Als FDR für eine ungünstige Basisrate von $\rho = 0.9$ ergibt sich in diesem Fall:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.005 \cdot 0.9}{0.005 \cdot 0.9 + 0.1 \cdot 0.1} \approx 0.31$$

- Also: In kleinen Stichproben etwas besser, aber immer noch hoch.

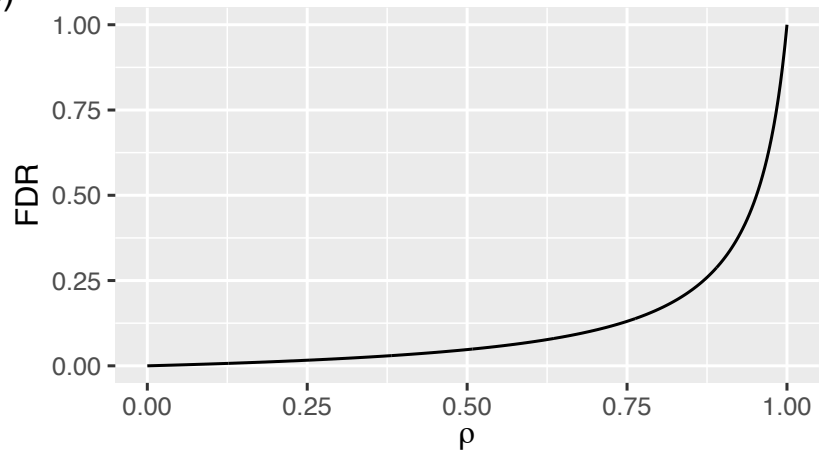
Dritter Ansatz IV

(a)



$\alpha = 0.05$ und $1 - \beta = 0.34$

(b)



$\alpha = 0.005$ und $1 - \beta = 0.1$

Auch wenn wir uns offen lassen wollen, welche Basisrate ρ tatsächlich vorliegt, zeigt sich im unteren Fall für das geringere Signifikanzniveau und die niedrigere Power ein günstigerer Verlauf der FDR.

Dritter Ansatz V

- Der Vorteil von $\alpha = 0.005$ wird noch deutlicher, falls man zusätzlich größere Stichproben und somit eine höhere Power fordert:

- Bei einer Power von $1 - \beta = 0.8$, einem Signifikanzniveau von $\alpha = 0.005$ und einer Basisrate von $\rho = 0.9$ ergibt sich:

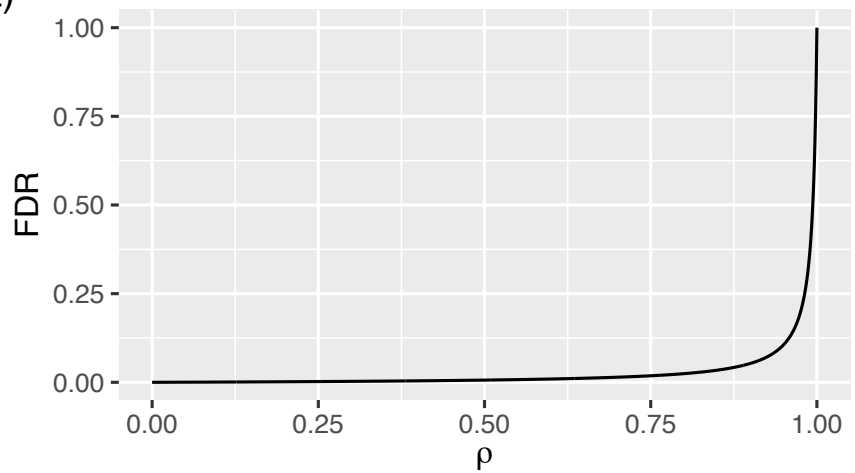
$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.005 \cdot 0.9}{0.005 \cdot 0.9 + 0.8 \cdot 0.1} \approx 0.05$$

- Im Vergleich dazu bei einer Power von $1 - \beta = 0.95$, einem Signifikanzniveau von $\alpha = 0.05$ und einer Basisrate von $\rho = 0.9$:

$$FDR = \frac{\alpha \cdot \rho}{\alpha \cdot \rho + (1 - \beta) \cdot (1 - \rho)} = \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.95 \cdot 0.1} \approx 0.32$$

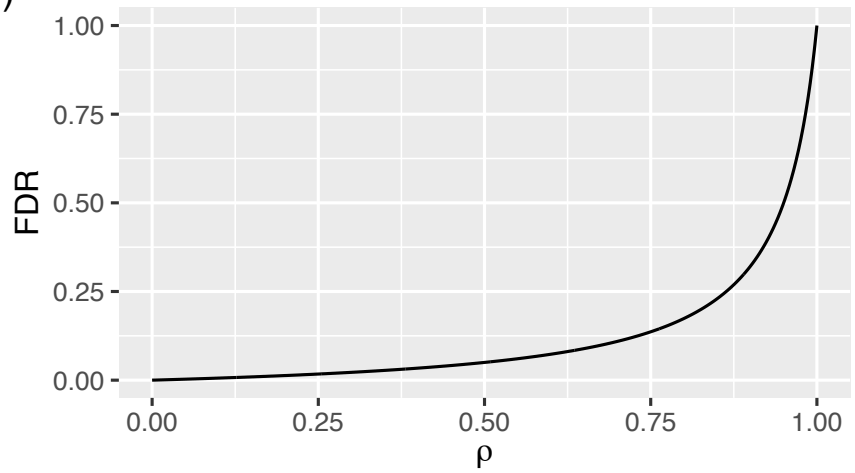
Dritter Ansatz VI

(a)



α = 0.005 und 1 – β = 0.8

(b)



α = 0.05 und 1 – β = 0.95

- Also: Deutlich geringere Abhängigkeit der FDR von der Basisrate bei $\alpha = 0.005$ und $1 - \beta = 0.8$ als bei $\alpha = 0.05$ und $1 - \beta = 0.95$.
- Bei $\alpha = 0.005$ und $1 - \beta = 0.8$ müsste die Basisrate sehr hoch sein, damit eine problematische FDR resultieren würde.
- Fall (a) und (b) unterscheiden sich in der benötigten Stichprobengröße außerdem kaum.

Dritter Ansatz VII

Gegenüberstellung der t-Test Stichprobenplanung für $\alpha = 0.005$ und $1 - \beta = 0.8$ und für $\alpha = 0.05$ und $1 - \beta = 0.95$, falls wir zudem berücksichtigen, dass $|\delta|_{H_1} = 0.5$ den wahren Effekt vermutlich überschätzt, und stattdessen $|\delta|_{H_1} = 0.3$ voraussetzen:

```
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
```

vs.

```
Two-sample t test power calculation

              n = 289.7353
              d = 0.3
      sig.level = 0.05
          power = 0.95
    alternative = two.sided

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

Fazit

- **Eine neue Konvention von $\alpha = 0.005$ und $1 - \beta = 0.8$ könnte zu einer deutlichen Verbesserung der Situation führen.**
- Bemerkung 1: Ein $\alpha = 0.005$ könnte auch als Daumenregel für das rückwirkende Filtern von schon veröffentlichten Studien verwendet werden (ausgenommen natürlich Studien, die unter Ansatz 2 fallen).
- Bemerkung 2: Vorsicht bei Hypothesentests in denen die H_0 die Wunschhypothese darstellt. Hier ist statt der FDR der durchschnittliche Anteil der falschen H_0 Entscheidungen an allen H_0 Entscheidungen relevant. Es kann daher sinnvoll sein, in solchen Fällen für analoge Überlegungen α und β zu vertauschen und z.B.: $\alpha = 0.2$ und $1 - \beta = 0.995$ zu wählen.
- Aber zur Erinnerung: Wir müssen zusätzlich voraussetzen, dass keine Verzerrung vorliegt, d.h. Annahmen der statistischen Hypothesentests sind erfüllt, einfache Zufallsstichproben liegen vor (d.h. Repräsentativität), keine Fehler bei Datenerhebung und Auswertung, keine Datenfälschung, keine statistischen Tricks, um signifikante Ergebnisse zu bekommen, etc.

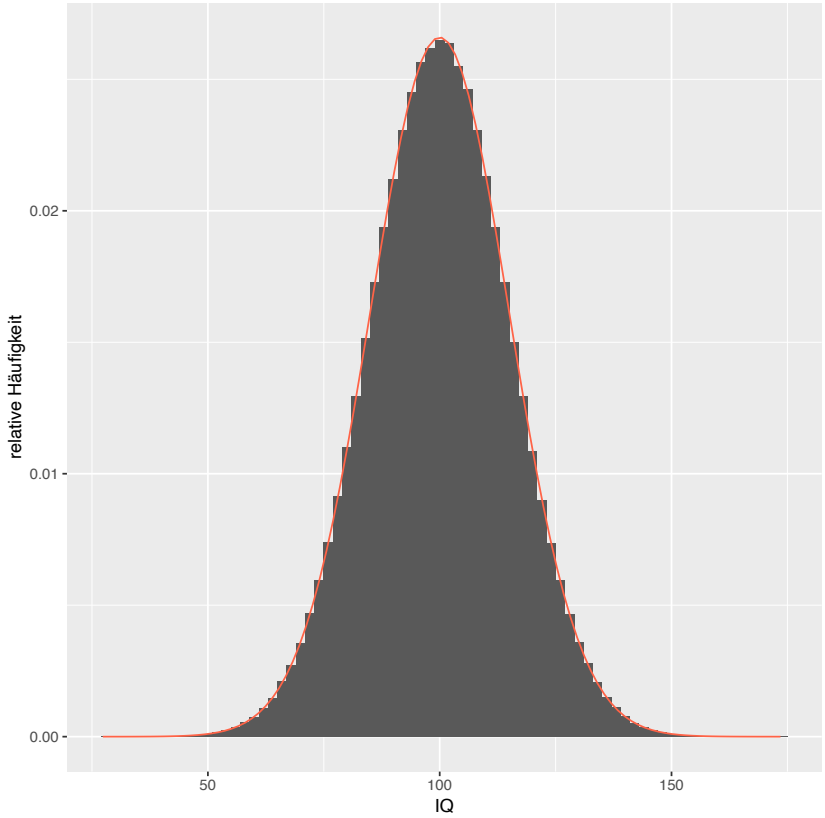
Annahmen inferenzstatistischer Verfahren

Wiederholung I

- Ausgangssituation:
  - Wir interessieren uns für die relative Häufigkeit einer Messwertausprägung einer diskreten Variable in einer Population.
  - Wir ziehen eine einfache Zufallsstichprobe aus dieser Population.
- Annahmen:
  - **keine**
- Inferenzstatistische Verfahren:
  - Intervallschätzung: Konfidenzintervall für π
  - Hypothesentests: Binomialtests.

Wiederholung II

- Ausgangssituation:
  - Wir interessieren uns für den Mittelwert einer metrischen Variable in einer Population.
  - Wir ziehen eine einfache Zufallsstichprobe aus dieser Population.
- Annahmen:
  - **Das Histogramm der interessierenden Variable in der Population kann durch die Dichtefunktion einer Normalverteilung approximiert werden.**
- Inferenzstatistische Verfahren:
  - Intervallschätzung: Konfidenzintervall für μ.
  - Hypothesentests: Einstichproben t-Tests.



Wiederholung III

- Ausgangssituation:
  - Wir interessieren uns für die Differenz der Mittelwerte einer metrischen Variable in zwei Populationen.
  - Wir ziehen zwei unabhängige einfache Zufallsstichproben aus den beiden Populationen.
- Annahmen:
  - **Das Histogramm der interessierenden Variable kann in beiden Populationen durch die Dichtefunktion einer Normalverteilung approximiert werden.**
  - **Die empirische Varianz der interessierenden Variable ist in beiden Populationen gleich groß.**
- Inferenzstatistische Verfahren:
  - Intervallschätzung: Konfidenzintervall für $\mu_1 - \mu_2$ bei unabhängigen Stichproben.
  - Hypothesentests: Zweistichproben t-Tests für unabhängige Stichproben.

Wiederholung IV

- Ausgangssituation:
  - Wir interessieren uns für die Differenz der Mittelwerte einer metrischen Variable in zwei Populationen.
  - Wir ziehen zwei abhängige einfache Zufallsstichproben aus den beiden Populationen.
- Annahmen:
  - **Das Histogramm der interessierenden Variable kann in beiden Populationen durch die Dichtefunktion einer Normalverteilung approximiert werden.**
- Statistische Verfahren:
  - Intervallschätzung: Konfidenzintervall für $\mu_1 - \mu_2$ bei abhängigen Stichproben.
  - Hypothesentests: Zweistichproben t-Tests für abhängige Stichproben.

Fazit

- Falls wir uns für die relative Häufigkeit einer Messwertausprägung einer diskreten Variable in einer Population interessieren, sind keine Annahmen nötig.

- In allen anderen Fällen müssen wir Annahmen treffen. Diese Annahmen können falsch sein:

  - Es könnte sein, dass die uns interessierende Variable zwar stetig ist, aber ihr Histogramm in der Population bzw. den Populationen nicht durch die Dichtefunktion einer Normalverteilung, sondern nur durch die Dichtefunktion irgendeiner anderen stetigen Wahrscheinlichkeitsverteilung approximiert werden kann.

  - Es könnte sein, dass die uns interessierende Variable überhaupt nicht stetig ist (sehr häufig in der Psychologie) und schon allein deshalb ihr Histogramm nicht durch die Dichtefunktion einer Normalverteilung approximiert werden kann.

  - Im unabhängigen Zweistichprobenfall könnte es sein, dass die empirischen Varianzen nicht gleich sind, dass die Normalverteilungsannahme in mindestens einer der Gruppen nicht gegeben ist, oder dass beide Annahmen nicht erfüllt sind.

Konsequenzen

- Wir haben alle bisher besprochenen inferenzstatistischen Verfahren unter der Bedingung hergeleitet, dass die Annahmen erfüllt sind.
- Falls die Annahmen falsch sind, haben diese Verfahren nicht die von uns gewünschten Eigenschaften:
  - Die Konfidenzintervalle haben nicht das von uns festgesetzte Konfidenzniveau.
  - Die Hypothesentests haben nicht das von uns festgesetzte Signifikanzniveau und nicht die von uns im Rahmen der Stichprobenplanung festgesetzte Power.

Überprüfung der Annahmen

- Es gibt verschiedene Methoden, um die Annahmen auf der Basis der Stichprobendaten zu überprüfen.
- Auf der einen Seite gibt es graphische Methoden zur Überprüfung der Plausibilität dieser Annahmen mithilfe von Histogrammen, Boxplots etc.
- Auf der anderen Seite gibt es statistische Hypothesentests, um die Annahmen zu überprüfen z.B.
  - den Levene-Test zur Überprüfung der Gleichheit der Varianzen im unabhängigen Zweistichprobenfall.
  - den Kolmogorov-Smirnov-Test und den Shapiro-Wilk-Test zur Überprüfung der Normalverteilungsannahmen.
- Probleme:
  - Die graphischen Methoden sind sehr subjektiv.
  - Die Hypothesentests treffen teilweise selbst wieder Annahmen (z.B. setzt der Levene-Test die Normalverteilungsannahme voraus)

Vorgehen

- Außerdem: Wir müssen auch ohne Überprüfung realistischerweise davon ausgehen, dass unsere Annahmen falsch sind. Warum sollte z.B. die empirische Varianz einer Variable in zwei Populationen exakt gleich sein?

- Wir beschäftigen uns daher nicht weiter damit, wie wir die Annahmen überprüfen könnten, sondern damit, wie wir inferenzstatistische Verfahren konstruieren könnten, falls wir sicher wüssten, dass die Annahmen falsch sind.

- Die in diesem Fall resultierenden Konfidenzintervalle und Hypothesentests werden wir dann mit den Konfidenzintervallen und Hypothesentests vergleichen, die wir mithilfe der falschen Annahmen hergeleitet haben.

- Wir werden das Vorgehen beispielhaft an dem Fall illustrieren, in dem wir uns für den Mittelwert einer metrischen Variable in einer einzelnen Population interessieren.

Inferenzstatistik ohne Annahmen

Ausgangssituation ohne Annahmen

- Wir interessieren uns für der Mittelwert $\bar{x}_{Pop}$ einer beliebigen metrischen (nicht unbedingt stetigen) Variable in der Population. Die empirische Varianz dieser Variable in der Population sei $s^2_{emp\ Pop}$.

- Wir setzen lediglich voraus, dass eine einfache Zufallsstichprobe mit Umfang n aus der Population gezogen wird.

- Wir nehmen also **nicht** an, dass das Histogramm der interessierenden Variable in der Population durch die Dichtefunktion einer Normalverteilung approximiert werden kann.

- Seien $X_1, X_2, \ldots, X_n$ diskrete oder stetige Zufallsvariablen, deren Realisationen jeweils für den Messwert der gezogenen Personen auf der interessierenden Variable stehen.

Erste Folgerungen

- Da alle Personen unabhängig aus der gleichen Population gezogen wurden, sind die Zufallsvariablen $X_1, X_2, \ldots, X_n$ iid (Beweis etwas schwierig). Sie folgen alle einer unbekannten (diskreten oder stetigen) Wahrscheinlichkeitsverteilung P:

$$X_i \overset{\text{iid}}{\sim} P$$

- Zudem kann man in dieser Situation zeigen, dass
  - der Erwartungswert $E(X_i)$ für alle i gleich ist und dem Mittelwert $\bar{x}_{Pop}$ in der Population entspricht,
  - die Varianz $Var(X_i)$ für alle i gleich ist und der empirischen Varianz $s^2_{emp\,Pop}$ in der Population entspricht.
- Das heißt also:
  - Aussagen über $E(X_i)$ sind äquivalent zu Aussagen über $\bar{x}_{Pop}$,
  - Aussagen über $Var(X_i)$ sind äquivalent zu Aussagen über $s^2_{emp\,Pop}$.

Punktschätzung

- Als Schätzfunktion für $E(X_i)$ (und somit für $\bar{x}_{Pop}$) können wir wegen

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n \cdot E(X_i) = E(X_i)$$

  einfach $\bar{X}$ verwenden.

- Als Schätzfunktion für $Var(X_i)$ (und somit für $s^2_{emp\ Pop}$) können wir wegen

$$E(S^2) = \cdots = Var(X_i)$$

  einfach S^2 verwenden (der Beweis hierfür ist etwas aufwendiger).

- **Erste Erkenntnis also: Ob wir die Normalverteilungsannahme treffen oder nicht, wirkt sich nicht auf die Wahl der Schätzfunktionen aus.**

z-Standardisierung

- Die Varianz von $\bar{X}$ ist in dieser Situation

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{Var(X_i)}{n}$$

- z-Standardisierung von $\bar{X}$ führt damit zu

$$Z = \frac{\bar{X} - E(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{Var(X_i)}{n}}}$$

- Falls wir nun noch die unbekannte Varianz $Var(X_i)$ im Nenner durch die Schätzfunktion S^2 ersetzen, gelangen wir zu

$$Z^* = \frac{\bar{X} - E(\bar{X})}{\widehat{SD}(\bar{X})} = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{S^2}{n}}}$$

- Falls wir nun die Wahrscheinlichkeitsverteilung der Zufallsvariable

$$Z^* = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{S^2}{n}}}$$

kennen würden, könnten wir sofort die entsprechenden Quantile bestimmen und ein Konfidenzintervall für $E(X_i)$ konstruieren.

- Außerdem würden wir dann auch die Wahrscheinlichkeitsverteilung von Z^* für bestimmte Werte $E(X_i) = \mu_0$ kennen und könnten leicht kritische Bereiche und p-Werte für die Teststatistik

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

bestimmen.

Problem und Lösung: Approximative Wahrscheinlichkeitsverteilung

- Problem: Wir können die Wahrscheinlichkeitsverteilung von Z^* nicht bestimmen, da wir nicht einmal die Wahrscheinlichkeitsverteilung P der einzelnen Zufallsvariablen X_i kennen.
- Genau hierfür hatten wir die Normalverteilungsannahme verwendet: Falls die X_i alle iid normalverteilt sind, folgt Z^* einer t-Verteilung mit $\nu = n - 1$.
- Was können wir ohne Normalverteilungsannahme tun?
- Genau dieser Fall ist bereits bei der Konstruktion eines Konfidenzintervalls für π aufgetaucht (VL 8 ab Folie 13).
- Bereits hier hatten wir uns zu Nutze gemacht, dass $Z^* \overset{a}{\sim} N(0,1)$ (der Beweis dafür ist sehr schwierig). Diese Lösung können wir auch in dem Fall verwenden, falls die Wahrscheinlichkeitsverteilung P der einzelnen Zufallsvariablen X_i gänzlich unbekannt ist.
- Es gilt also auch hier für große Stichproben, d.h. für großes n:

$$Z^* = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{S^2}{n}}} \overset{a}{\sim} N(0,1)$$

Konstruktion des Konfidenzintervalls I

- Das heißt: Auch wenn wir wissen, dass die Normalverteilungsannahme falsch ist, können wir das Konfidenzintervall trotzdem so konstruieren, als ob sie wahr wäre.

- Wir haben dann in großen Stichproben die Garantie, dass das so konstruierte Konfidenzintervall auch bei Verletzung der Normalverteilungsannahme ein approximatives Konfidenzniveau von $1 - \alpha$ aufweist.

- Wichtig: In kleinen Stichproben haben wir diese Garantie nicht.

Konstruktion des Konfidenzintervalls II

- Einsetzen von

$$Z^* = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{S^2}{n}}}$$

in $P\left(z_{\frac{\alpha}{2}} \leq Z^* \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$ und Umstellen (analog zum Fall mit Normalverteilungsannahme VL 7 Folien 66-68) ergibt das folgende Konfidenzintervall für $E(X_i)$:

$$I(X_1, \dots, X_n) = \left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}\right]$$

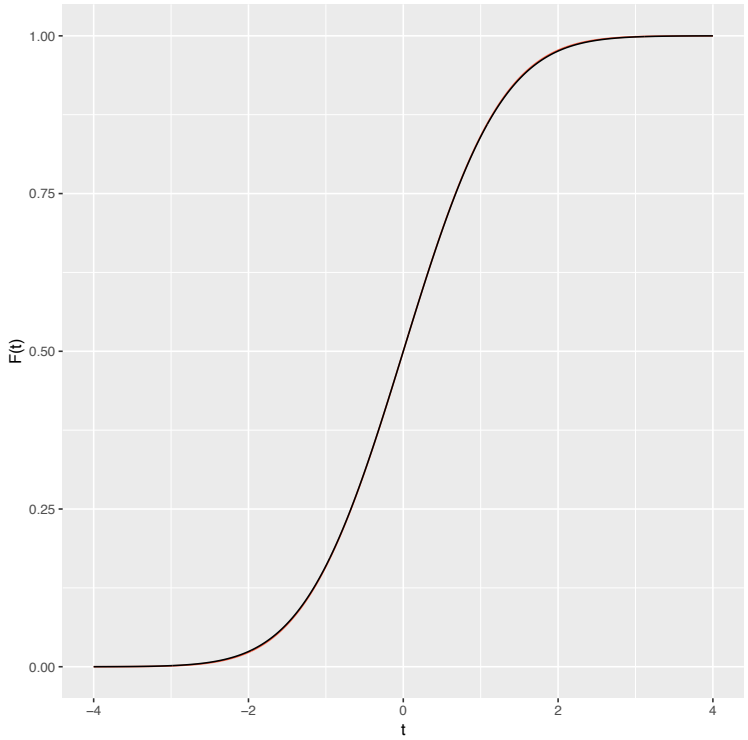
- Diese Konfidenzintervall hat ein approximatives Konfidenzniveau von $1 - \alpha$.

Konstruktion des Konfidenzintervalls III

- Wir wissen zudem, dass die t-Verteilung für $\nu \to \infty$ gegen die Standardnormalverteilung konvergiert:

$$\lim_{\nu \to \infty} t(\nu) = N(0,1)$$

- Falls wir also in großen Stichproben $\nu = n - 1$ wählen, ist ν ebenfalls groß und wir können die Standardnormalverteilung durch eine t-Verteilung mit $\nu = n - 1$ approximieren:



rot: $N(0,1)$
schwarz: $t(100)$

Konstruktion des Konfidenzintervalls IV

- Da sich in diesem Fall die Verteilungsfunktionen nur geringfügig unterscheiden, unterscheiden sich auch die Quantile $z_{1-\frac{\alpha}{2}}$ und $t_{1-\frac{\alpha}{2}}$ nur geringfügig und es gilt

$$z_{1-\frac{\alpha}{2}} \approx t_{1-\frac{\alpha}{2}}$$

- Ersetzen von $z_{1-\frac{\alpha}{2}}$ durch $t_{1-\frac{\alpha}{2}}$ in unserem approximativen Konfidenzintervall auf der vorletzten Folie ergibt als alternatives approximatives Konfidenzintervall

$$I(X_1, \dots, X_n) = \left[\bar{X} - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{n}}\right]$$

wobei $t_{1-\frac{\alpha}{2}}$ das $1-\frac{\alpha}{2}$ - Quantil einer t-Verteilung mit $\nu = n-1$ ist.

- Dieses Konfidenzintervall ist dann identisch mit dem Konfidenzintervall, dass wir unter der Normalverteilungsannahme hergeleitet haben.

Konstruktion des Konfidenzintervalls V

- Das heißt: Auch wenn wir wissen, dass die Normalverteilungsannahme falsch ist, können wir das Konfidenzintervalls trotzdem so konstruieren, als ob sie wahr wäre.

- Wir haben dann in großen Stichproben die Garantie, dass das so konstruierte Konfidenzintervall auch bei Verletzung der Normalverteilungsannahme ein approximatives Konfidenzniveau von $1 - \alpha$ aufweist.

- Wichtig: In kleinen Stichproben haben wir diese Garantie nicht.

Hypothesentests I

- Wir wissen, dass

$$Z^* = \frac{\bar{X} - E(X_i)}{\sqrt{\frac{S^2}{n}}}$$

für große Stichproben approximativ einer Standardnormalverteilung folgt.

- Unter der Voraussetzung, dass $E(X_i) = \mu_0$ ist, ist damit die Teststatistik

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

approximativ standardnormalverteilt. Diese Teststatistik ist identisch mit der Teststatistik der Einstichproben-t-Tests.

Hypothesentests II

- Wir können für diese Teststatistik und ihre Realisation daher approximative kritische Bereiche und approximative p-Werte auf der Basis der Standardnormalverteilung konstruieren.

- Zum Beispiel wäre bei einer linksgerichteten Alternativhypothese

$$H_0: E(X_i) \geq \mu_0$$
$$H_1: E(X_i) < \mu_0$$

der kritische Wert t_{krit} derjenige Wert, für den $P(T \leq t_{krit}) = F(t_{krit}) = \alpha$ ist, wobei P eine Standardnormalverteilung und F deren Verteilungsfunktion ist.

- Auch hier können wir außerdem wieder wie bei den Konfidenzintervallen die Standardnormalverteilung durch eine t-Verteilung mit $\nu = n - 1$ approximieren.

- In diesem Fall wäre der kritische Wert t_{krit} derjenige Wert, für den $P(T \leq t_{krit}) = F(t_{krit}) = \alpha$ ist, wobei P eine t-Verteilung mit $\nu = n - 1$ und F deren Verteilungsfunktion ist. Dies ist der gleiche kritische Wert wie im linksgerichteten Einstichproben-t-Test.

- Das heißt auch hier: Falls wir davon ausgehen müssen, dass die Normalverteilungsannahme falsch ist, können wir den kritischen Bereich bzw. den p-Wert trotzdem so konstruieren, als ob sie wahr wäre.

- Wir haben dann in großen Stichproben die Garantie, dass der so konstruierte Hypothesentest auch bei Verletzung der Normalverteilungsannahme ein approximatives Signifikanzniveau von α aufweist.

- Wichtig: In kleinen Stichproben haben wir diese Garantie nicht.

- Bemerkung: Für die Poweranalyse bzw. die Stichprobenplanung gilt dies ebenfalls.

Allgemeines Fazit

- Fazit: Die Verletzung der Annahme kann in diesem Fall bei großen Stichproben vernachlässigt werden. Wir können einfach die schon bekannten inferenzstatistischen Verfahren verwenden.
- Aber: Das Ziehen einer einfachen Zufallsstichprobe wird immer noch vorausgesetzt.
- Für die beiden Zweistichprobenfälle (abhängig und unabhängig) gilt dies ebenfalls: Auch hier können wir in großen einfachen Zufallsstichproben auch bei Verletzung der Annahmen einfach die entsprechenden schon bekannten Konfidenzintervalle und t-Tests verwenden.
- Die Argumentation ist in diesen Fällen weitgehend analog.

Bemerkungen

- Je nachdem, wie weit die tatsächliche Wahrscheinlichkeitsverteilung P der Zufallsvariablen X_i von der Normalverteilung entfernt ist, kann die für eine einigermaßen genaue Approximation durch die Standardnormalverteilung benötigte Stichprobengröße n sehr groß sein.

- Für bestimmte Stichprobengrößen und spezifische Verletzungen der Annahmen kann es unter Umständen bessere Alternativen zu den auf den Annahmen basierenden inferenzstatistischen Verfahren geben:
  - Welch-Test für zwei unabhängige Stichproben.
  - Klassische nonparametrische Verfahren: z.B: Wilcoxon Tests.
  - Robuste Verfahren (wrs2 package in R).

- Außerdem: Obwohl dies bei allen von uns bis jetzt behandelten Verfahren der Fall ist, entsprechen die unter bestimmten Annahmen hergeleiteten statistischen Verfahren nicht immer den approximativen Verfahren, die man ohne Annahmen herleiten würde.

Zusammenfassung

- Die False Discovery Rate gibt den Anteil der falschen Entscheidungen für die H_1 an allen Entscheidungen für die H_1 an.
- Die FDR hängt dabei neben Signifikanzniveau und Power maßgeblich von der Grundquote ab, also dem Anteil der tatsächlich gültigen H_1 an allen Hypothesen.
- Da die Grundquote nur schwer geschätzt werden kann ist es zweckmäßig, Signifikanzniveau und Power so zu wählen, dass die FDR für einen großen Bereich möglicher Ausprägungen der Grundquote gering ist.
- Eine Wahl von $\alpha = 0.005$ und $1 - \beta = 0.8$ hat hier günstige Auswirkungen.
- Eine nachträgliche Verringerung von α verringert zwar auch die Power, insgesamt ergibt sich dadurch jedoch trotzdem eine bessere FDR.
- Die Verteilungsannahmen sind für die Konstruktion der von uns besprochenen Verfahren essentiell.
- Sollten diese in Wahrheit nicht gelten, sorgen jedoch mathematische Grenzwertsätze dafür, dass mit größerem n (Stichprobengröße) zunehmend die Eigenschaften der Verfahren approximativ eingehalten werden.