12. Vorlesung Statistik I
Effektstärken, Power und Stichprobenplanung

Effektstärken

Motivation

- Wir betrachten eine Parameterdifferenz $\mu_1 - \mu_2$, die der Mittelwertsdifferenz einer stetigen Variable in zwei Populationen entspricht.

- Frage: Ab wann können wir sagen, dass es sich bei dieser Parameterdifferenz um eine große bzw. „bedeutende“ Differenz handelt?

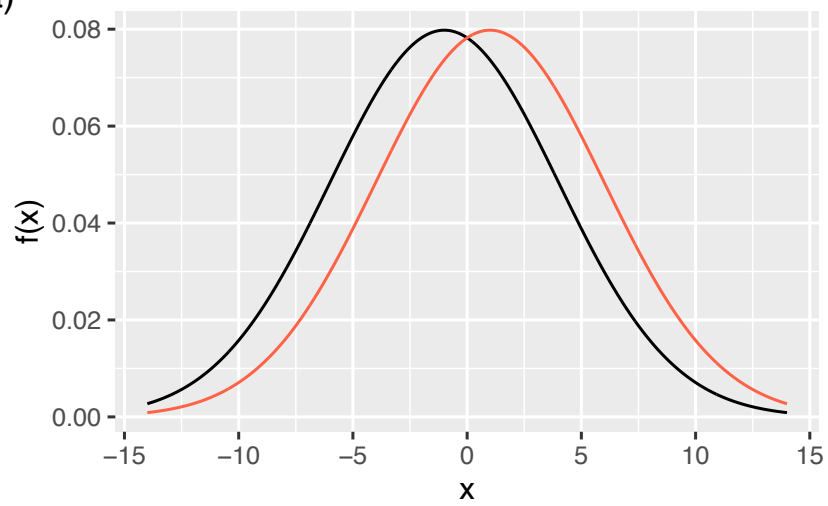
- Mehrere Probleme bei der Beantwortung dieser Frage:
  - Sie hängt von der Einheit der interessierenden Variable ab: Eine Differenz von 10 cm wäre geringer als eine Differenz von 10 km. Problem: Viele psychologische Variablen haben keine direkt interpretierbare Einheit: Was bedeutet z.B. eine Mittelwertsdifferenz von 10 in einem Konzentrationstest?
  - Auch falls die Einheit interpretierbar ist, hängt die Beurteilung der Größe der Differenz vom Kontext ab: Eine Mittelwertsdifferenz von einer Sekunde zwischen zwei Gruppen im 100m Lauf würde einem größeren Unterschied entsprechen als eine Mittelwertsdifferenz von einer Sekunde zwischen zwei Gruppen im 20000m Lauf.

Effektstärken

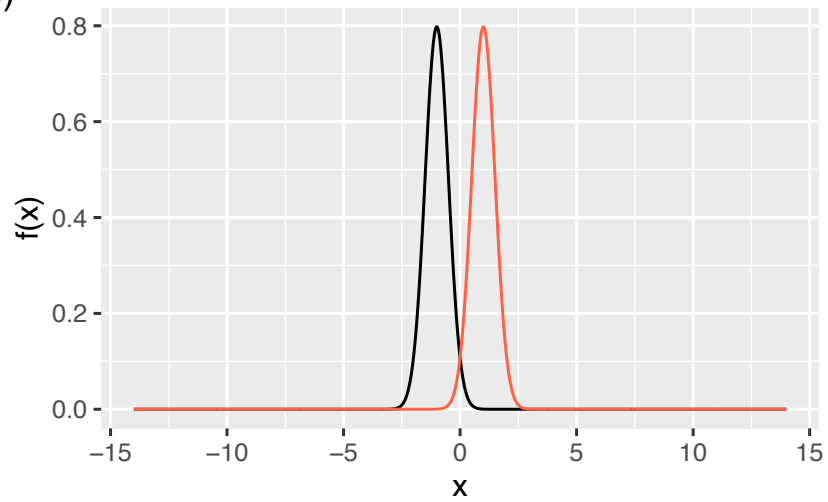
- Wir benötigen also eine Maßzahl für die Größe einer Parameterdifferenz $\mu_1 - \mu_2$.
- Hierfür verwenden wir sogenannte Effektstärken.
- (grobe) Allgemeine Definition: Eine **Effektstärke** ist ein Parameter, der einheitsunabhängig ist und als Maß für die Größe eines Unterschieds oder die Stärke eines Zusammenhangs interpretiert werden kann.
- Synonyme: Effektgröße, Effekt.

Vorüberlegungen I

(a)



(b)



- Wie könnte eine Effektgröße im Fall einer Parameterdifferenz $\mu_1 - \mu_2$ aussehen?
- Zunächst sollte eine Effektgröße in diesem Fall die zwei nebenstehenden Situationen unterscheiden können:
- In den Situationen ist die Parameterdifferenz $\mu_1 - \mu_2$ identisch. In Bild (b) haben aber fast alle Personen aus der Population 2 (rote Dichte) einen höheren Variablenwert als fast alle Personen aus der Population 1 (schwarze Dichte), während es in Bild (a) mehr Überschneidung gibt. In der Situation (b) sollte die Effektstärke also sinnvollerweise einen größeren Wert annehmen.

Vorüberlegungen II

- Eine Effektstärke für die Größe einer Parameterdifferenz sollte also bei gleicher Differenz $\mu_1 - \mu_2$ umso größer sein, je kleiner die Streuung der interessierenden Variable innerhalb der beiden Populationen ist.

- Welcher Parameter quantifiziert die Streuung innerhalb der Populationen (bei unabhängigen Stichproben)? → $\boldsymbol{\sigma^2}$

- Wir fordern also, dass eine Effektstärke bei gleicher Differenz $\mu_1 - \mu_2$ umso größer sein sollte, je kleiner σ^2 ist.

Cohen‘s δ I

- Diese Überlegungen liegen der Definition der Effektstärke **Cohen‘s δ** zugrunde.
- Diese ist für unabhängige Stichproben als

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2}}$$

und für abhängige Stichproben als

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{Diff}}}$$

definiert.

- Bemerkung: Bei abhängigen Stichproben entspricht σ^2_{Diff} nicht genau der empirischen Varianz innerhalb der beiden Populationen, hängt aber mit dieser zusammen.

Cohen‘s δ II

- Wichtige Eigenschaften von Cohen‘s δ:
  - Cohen‘s δ ist bei gleicher Parameterdifferenz $\mu_1 - \mu_2$ umso größer, je kleiner die Varianz σ^2 bzw. σ^2_{Diff} ist.
  - Cohen‘s δ ist unabhängig von der Einheit der interessierenden Variable: Falls wir die Einheit der Variable ändern, ändert sich Cohen‘s δ nicht.
  - Cohen‘s δ hat selbst zwar keine Einheit, trotzdem kann man für die Interpretation sagen, dass es eine Parameterdifferenz „in Standardabweichungen“ angibt. Ist beispielsweise im unabhängigen Fall die Parameterdifferenz genau so groß wie $\sqrt{\sigma^2}$, dann ist $\delta = 1$. Ist die Differenz dagegen nur halb so groß wie die Standardabweichung $\sqrt{\sigma^2}$, wird $\delta = 0.5$, etc..
  - Cohen‘s δ ist negativ, falls $\mu_1 - \mu_2$ negativ ist, und positiv, falls $\mu_1 - \mu_2$ positiv ist. Für die Beurteilung der Größe der Parameterdifferenz ist daher nur der Betrag $|\delta|$ relevant. Das Vorzeichen liefert aber Informationen über die Richtung der Parameterdifferenz.

Daumenregeln nach Cohen

- Daumenregel für die Interpretation von δ nach Cohen (1988):

$\lvert\delta\rvert$	0.2	0.5	0.8
Interpretation	kleiner Effekt	mittlerer Effekt	großer Effekt

- Effektstärken sollten immer mit Bezug auf die inhaltliche Fragestellung interpretiert werden (was ist ein bedeutsamer Effekt?)

Parameterschätzung für Effektstärken

Punktschätzung

Punktschätzung

- Um erwartungstreue und konsistente Schätzfunktionen für δ in abhängigen und unabhängigen Stichproben zu erhalten, ersetzen wir einfach alle unbekannten Größen in

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2}}$$

bzw.

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2_{Diff}}}$$

durch die jeweiligen Schätzfunktionen (siehe VL 9).

- Bemerkung: Der Schätzwert $\hat{\delta}_{Wert}$ für δ wird in der Literatur oft auch Cohen‘s d genannt.

Punktschätzung in unabhängigen Stichproben

- In **unabhängigen** Stichproben ergibt sich also als Schätzfunktion

$$\hat{\delta} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{pool}}}$$

und somit als Schätzwert

$$\hat{\delta}_{Wert} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2_{pool}}}$$

Punktschätzung in abhängigen Stichproben

- In **abhängigen** Stichproben ergibt sich als Schätzfunktion

$$\hat{\delta} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2_{Diff}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{Diff}}}$$

und somit als Schätzwert

$$\hat{\delta}_{Wert} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{Diff}}}$$

Punktschätzung Beispiel

- Beispiel: Uns interessiert der Unterschied zwischen Depressiven und Nicht-Depressiven in der durchschnittlichen Konzentrationsleistung. Wir setzen voraus, dass das Histogramm der durch einen psychologischen Test erfassten Konzentrationsleistung in beiden Populationen durch die Dichte einer Normalverteilung approximiert werden kann.

- Wir ziehen zwei unabhängige einfache Zufallsstichproben:
  - Stichprobe 1 aus der Population der Depressiven mit Umfang $n_1 = 101$.
  - Stichprobe 2 aus der Population der Nicht-Depressiven $n_2 = 51$.
- Als Mittelwerte in den Stichproben ergeben sich $\bar{x}_1 = 165$ und $\bar{x}_2 = 170$
- Als Schätzwert für σ^2 ergibt sich $s^2_{pool} = 87.33$
- Damit ist der Schätzwert für δ

$$\hat{\delta}_{Wert} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{pool}}} = \frac{165 - 170}{\sqrt{87.33}} = -0.54$$

Intervallschätzung

Konfidenzintervalle für δ I

- Noch besser ist es, Konfidenzintervalle für δ zu berechnen.
- Dies ist in unabhängigen und abhängigen Stichproben möglich.
- Die Konstruktion ist mathematisch sehr aufwendig, weswegen wir sie nicht besprechen werden.
- Die praktische Berechnung in R ist jedoch sehr leicht.
- Bemerkung: Die Voraussetzungen dafür, dass das Konfidenzintervall für δ das gewünschte Konfidenzniveau aufweist, sind identisch wie bei den jeweiligen Konfidenzintervallen für $\mu_1 - \mu_2$.

Konfidenzintervalle für δ II

- Beispiel Folie 15: Unterschied zwischen Depressiven und Nicht-Depressiven in der durchschnittlichen Konzentrationsleistung.
- Als Schätzwert für δ hatte sich $\hat{\delta}_{Wert} = -0.54$ ergeben.
- Wir wollen nun ein 0.95-Konfidenzintervall für δ berechnen.
- Um das Konfidenzintervall in R zu berechnen, benötigen wir zudem die beiden Stichprobengrößen: $n_1 = 101$ und $n_2 = 51$.

Konfidenzintervalle für δ III

- Berechnung in R (MBESS Paket):

```
> ci.smd(smd = -0.54, n.1 = 101, n.2 = 51)

$Lower.Conf.Limit.smd
[1] -0.8813043

$smd
[1] -0.54

$Upper.Conf.Limit.smd
[1] -0.1969517
```

- Das Konfidenzinterval ist $[-0.88, -0.20]$. Die plausiblen Werte für δ liegen also zwischen -0.88 und -0.2. Dies schließt sowohl kleine, mittlere, als auch große Effekte ein. Wir können also auf Basis der Daten keine präzisere Aussage über die Größe des Effekts treffen. Wir können jedoch aufgrund des negativen Vorzeichens aller Werte im KI davon ausgehen, dass Depressive eine geringere Konzentrationsleistung als Nicht-Depressive aufweisen.

Stichprobenplanung für Konfidenzintervalle für Effektstärken

Stichprobenplanung für Konfidenzintervalle I

- Neben der unter Umständen besseren Interpretierbarkeit von δ im Vergleich zur einfachen Parameterdifferenz $\mu_1 - \mu_2$ gibt es aus statistischer Sicht noch einen weiteren Grund für die Bevorzugung von δ:

- Im Fall eines Konfidenzintervalls für δ können wir nämlich vor der Datenerhebung eine Stichprobengröße n berechnen, so dass das **zufällige** Konfidenzintervall eine bestimmte, von uns vorgegebene erwartete Länge aufweist.

- Wir können also vorgeben, wie präzise unsere Intervallschätzung für δ sein soll und dann bestimmen, wie groß unsere Stichprobe sein muss, damit diese Vorgabe erfüllt ist.

- Eine Stichprobe dieser Größe erheben wir dann und berechnen in dieser das **konkrete** Konfidenzintervall.

- Dies ist für Konfidenzintervalle für $\mu_1 - \mu_2$ nicht möglich (bzw. nur mit weiteren unrealistischen Annahmen).

Stichprobenplanung für Konfidenzintervalle II

- Mathematisch ist die Berechnung von n noch aufwendiger als die Konstruktion des Konfidenzintervalls selbst.
- Wir beschränken uns daher wieder auf die Anwendung.
- Praktisches Problem: Obwohl uns bei der Stichprobenplanung primär die Präzision und damit die Länge des zufälligen Konfidenzintervalls interessiert, müssen wir eine Schätzung für δ selbst angeben um die benötigte Stichprobengröße n berechnen zu können.
- Woher sollen wir diese nehmen?

Stichprobenplanung für Konfidenzintervalle III

- Wir könnten z.B. als Daumenregel einfach einen mittleren Effekt, also $\delta = 0.5$ verwenden.

- Ein sinnvolles Vorgehen wäre auch, die Stichprobenplanung einfach für mehrere Werte von δ durchzuführen, um zu sehen, ob sich die benötigten Stichprobengrößen überhaupt stark unterscheiden. Oft sind die Unterschiede nicht besonders groß.

- Man kann zeigen, dass die **erwartete Länge des zufälligen** Konfidenzintervalls umso größer wird, je größer das wahre δ ist. Eine konservative Methode wäre also, ein möglichst großes δ vorzugeben, z.B. $\delta = 1$. Dann kann man relativ sicher sein, dass das zufällige Konfidenzintervall höchstens die gewünschte erwartete Länge aufweist.

Stichprobenplanung für Konfidenzintervalle IV

- Beispiel: Wir wollen ein 0.95-Konfidenzintervall für δ berechnen und möchten, dass die erwartete Länge des Konfidenzintervalls klein genug ist, damit wir einen kleinen von einem mittleren Effekt unterscheiden können.

- Die erwartete Länge des Konfidenzintervalls sollte also 0.29 betragen, da in diesem Fall die Grenzen des erwarteten Konfidenzintervalls nicht gleichzeitig die beiden Grenzen 0.2 (kleiner Effekt) und 0.5 (mittlerer Effekt) überdecken können.

- Wir geben für die Berechnung der benötigten Stichprobengröße einen mittleren Effekt von $\delta = 0.5$ vor.

- Berechnung in R:

```
> ss.aipe.smd(0.5, conf.level = 0.95, width = 0.29)
[1] 377
```

- Der ausgegebene Wert 377 entspricht der benötigten Stichprobengröße **pro Stichprobe**. Wir müssen also insgesamt mindestens 377 + 377 = 754 Personen erheben.

Power

Power I

- Zur Erinnerung: Die Power $1 - \beta$ ist neben dem Signifikanzniveau α eines der beiden Gütekriterien eines statistischen Hypothesentests.
- Sie ist die Wahrscheinlichkeit dafür, dass wir uns für die H_1 entscheiden, falls diese tatsächlich wahr ist.
- Wie wir wissen, entscheiden wir uns im Rahmen eines statistischen Hypothesentests genau dann für die H_1, falls die Realisation der Teststatistik im kritischen Bereich liegt.
- Die Power eines statistischen Tests ist also die Wahrscheinlichkeit dafür, dass sich die Teststatistik im kritischen Bereich realisiert, falls die H_1 gilt.
- **Problem:** In allen Fällen, die wir besprochen haben, sind unter der H_1 **unendlich** viele Parameterwerte möglich. Zur Bestimmung einer Wahrscheinlichkeit brauchen wir aber immer eine ganz konkrete Wahrscheinlichkeitsverteilung.
- Im Fall von $H_1: \mu > \mu_0$ sind unter der H_1 z.B. alle Parameterwerte größer als μ_0 möglich.
- Das heißt: Wir können die Power eines statistischen Tests immer nur für einen bestimmten festgelegten Parameterwert unter der H_1 bestimmen.

Power II

- Für diesen bestimmten Parameterwert ist die Power zudem nur bestimmbar, falls wir die Wahrscheinlichkeitsverteilung der Teststatistik unter der Voraussetzung bestimmen können, dass dieser Parameterwert der wahre Parameterwert ist.
- Wie wir diese Wahrscheinlichkeitsverteilung bestimmen, unterscheidet sich je nach statistischem Hypothesentest.

Binomialtests I

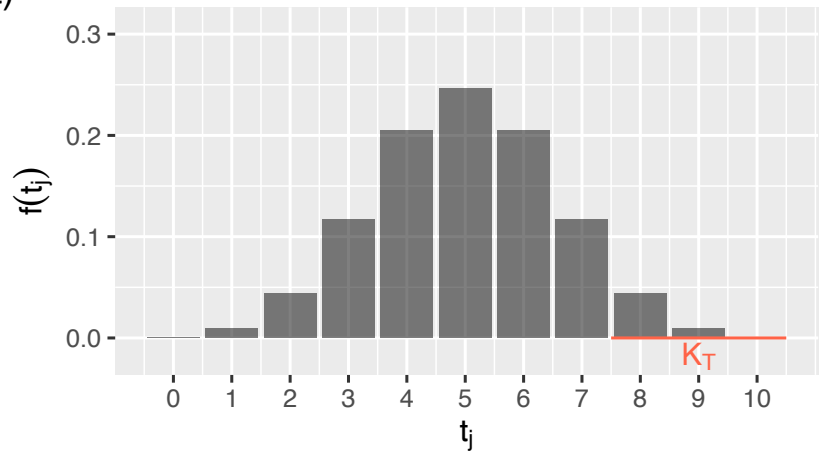
- Wir betrachten zunächst die Binomialtests.
- Hier ist die Teststatistik eine Summe von Bernoulli-Variablen:

$$T = \sum_{i=1}^{n} X_i$$

- Wir wissen, dass diese Teststatistik unter der Voraussetzung $\pi = \pi_0$ - also unter der H_0 bzw. dem extremsten Wert unter der H_0 - einer Binomialverteilung mit Parametern n und π_0 folgt. Auf der Basis dieser Verteilung können wir den kritischen Bereich für ein vorgegebenes Signifikanzniveau α bestimmen.
- Sei nun π_{H_1} ein bestimmter Parameterwert unter der H_1.
- Unter der Voraussetzung $\pi = \pi_{H_1}$ ist die Teststatistik ebenfalls binomialverteilt, und zwar mit Parametern n und π_{H_1}.
- Wir können also für einen vorher auf der Basis von π_0 bestimmten kritischen Bereich K_T für jeden Parameterwert π_{H_1} unter der H_1 die Wahrscheinlichkeit $P(T \in K_T)$, also die Power, berechnen.

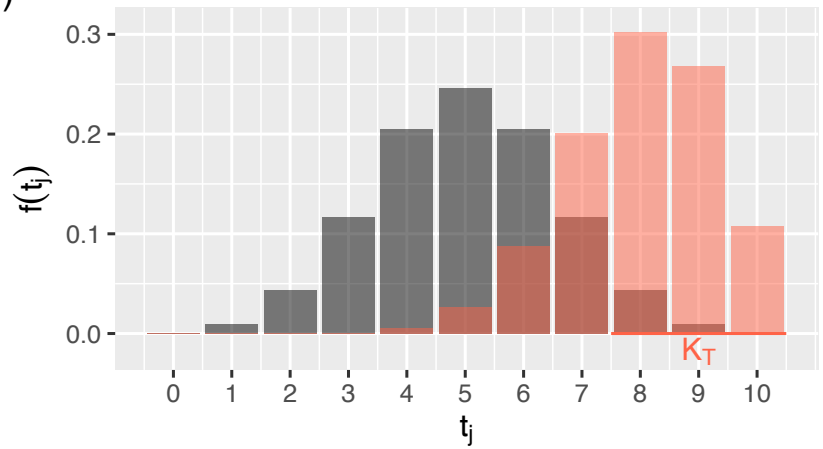
Binomialtests II

(a)



Wahrscheinlichkeitsfunktion der Teststatistik unter der H_0: $\pi = 0.5$

(b)



In rot: $f(t_j)$ falls die spezifische H_1 mit $\pi = 0.8$ gilt.

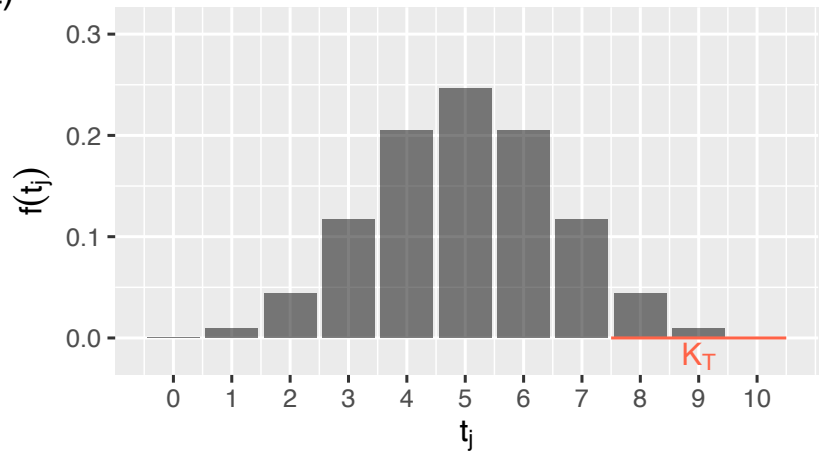
- Falls sich zum Beispiel für einen Binomialtest mit rechtsseitiger Alternativhypothese $H_1: \pi > 0.5$ auf der Basis eines von uns gewählten Signifikanzniveaus α bei einer Stichprobengröße von $n = 10$ ein kritischer Bereich von $K_T = \{8, 9, 10\}$ ergeben hätte, wäre die Power unter der Voraussetzung $\pi = \pi_{H_1} = 0.8$

$$P(T \in K_T) = P(T > 7) = 1 - P(T \leq 7)$$
$$= 1 - F(7)$$

wobei jetzt P eine Binomialverteilung mit Parametern $n = 10$ und $\pi = 0.8$ ist und F deren Verteilungsfunktion.

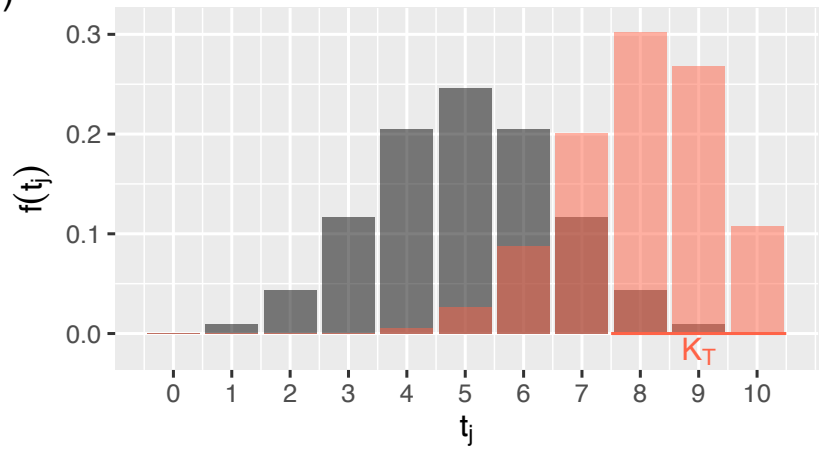
Binomialtests III

(a)



Wahrscheinlichkeitsfunktion der Teststatistik unter der H_0: $\pi = 0.5$

(b)



In rot: $f(t_j)$ falls die spezifische H_1 mit $\pi = 0.8$ gilt.

- Berechnung in R:

```
> 1 - pbinom(7, 10, 0.8)
[1] 0.6777995
```

- Für den konkreten Parameterwert $\pi_{H_1} = 0.8$ unter der $H_1: \pi > 0.5$ wäre die Power unseres Hypothesentests also 0.68.

- Frage: Warum wählen wir genau $\pi_{H_1} = 0.8$? Hierzu später mehr.

t-Tests I

- Im Fall der t-Tests ist die Berechnung der Power etwas schwieriger.
- Hier hängt die Wahrscheinlichkeitsverteilung der Teststatistik für bestimmte Parameterwerte unter der H_1 auch wieder von dem unbekannten Parameter σ^2 ab.
- Wir können jedoch die Wahrscheinlichkeitsverteilung der Teststatistik für bestimmte Effektstärken unter der H_1 bestimmen. Diese Effektstärken δ_{H_1} sind je nach Hypothesentest unterschiedlich definiert:

Eine Stichprobe:	Zwei unabhängige Stichproben	Zwei abhängige Stichproben
$\delta_{H_1} = \frac{\mu_{H_1} - \mu_0}{\sqrt{\sigma^2}}$	$\delta_{H_1} = \frac{(\mu_{1H_1} - \mu_{2H_1}) - \mu_0}{\sqrt{\sigma^2}}$	$\delta_{H_1} = \frac{(\mu_{1H_1} - \mu_{2H_1}) - \mu_0}{\sqrt{\sigma^2_{Diff}}}$

- Bemerkung: Um δ_{H_1} zu berechnen ist es im Zweistichprobenfall ausreichend, anstatt der einzelnen Werte für μ_{1H_1} und μ_{2H_1} direkt die Differenz $(\mu_{1H_1} - \mu_{2H_1})$ festzulegen.

t-Tests II

- Da die Formeln für δ_{H_1} den Parameter μ_0 enthalten, wird δ_{H_1} auch dann größer, je stärker sich ein angenommener Effekt von der Nullhypothese unterscheidet.

- Dementsprechend können δ_{H_1} und δ unterschiedliche Werte annehmen, obwohl es eigentlich um den gleichen Effekt geht:

- Ein Beispiel für einen t-test für zwei unabhängige Stichproben mit gerichteter H_1:

  - Umrechnung eines Mittelwertsunterschieds von 3 in die Effektstärke δ:

$$\mu_1 - \mu_2 = 3, \sqrt{\sigma^2} = 10 \Longrightarrow \delta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2}} = \frac{3}{10} = 0.3$$

  - Annahme einer rechtsgerichteten Alternativhypothese:

$$H_0: \mu_1 - \mu_2 \leq 1$$
$$H_1: \mu_1 - \mu_2 > 1$$

  - Berechnung von δ_{H_1} aus dem angenommenen Mittelwertsunterschied und der entsprechenden Nullhypothese:

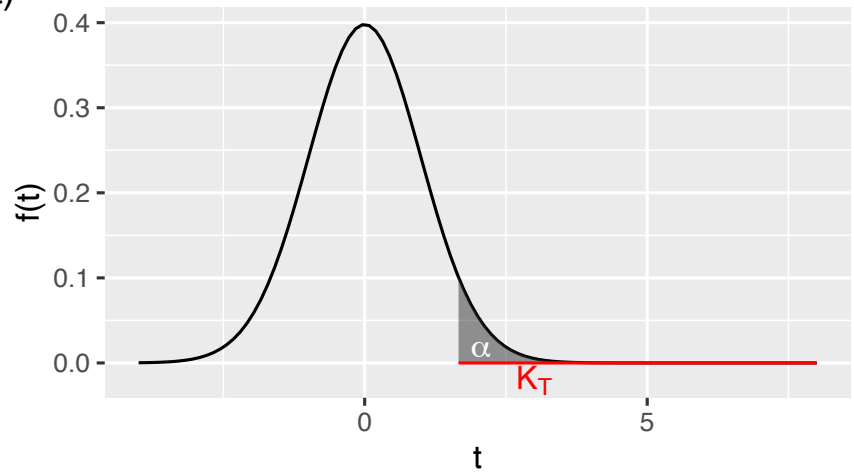
$$\mu_{1H_1} - \mu_{2H_1} = 3 \Longrightarrow \delta_{H_1} = \frac{(\mu_{1H_1} - \mu_{2H_1}) - \mu_0}{\sqrt{\sigma^2}} = \frac{3-1}{10} = \frac{2}{10} = 0.2$$

t-Tests III

- Das heißt: Um die Power eines t-Tests bestimmen zu können, müssen wir eine Effektstärke δ_{H_1} unter der H_1 festlegen.
- Für eine gegebene Effektstärke δ_{H_1} folgt die Teststatistik in allen t-Tests einer sogenannten **nonzentralen t-Verteilung.**
- Bemerkung: „nonzentral“ bezieht sich hier vereinfacht gesagt darauf, dass die Verteilung u.a. im Vergleich zur Verteilung der Teststatistik unter Annahme der H_0 verschoben ist.
- Die genaue Form dieser Verteilung und ihre Parameter werden wir nicht besprechen.
- Wichtig ist, dass wir auf ihrer Basis für jeden kritischen Bereich K_T und jede Effektstärke δ_{H_1} die Wahrscheinlichkeit $P(T \in K_T)$, also die Power , in R berechnen können.

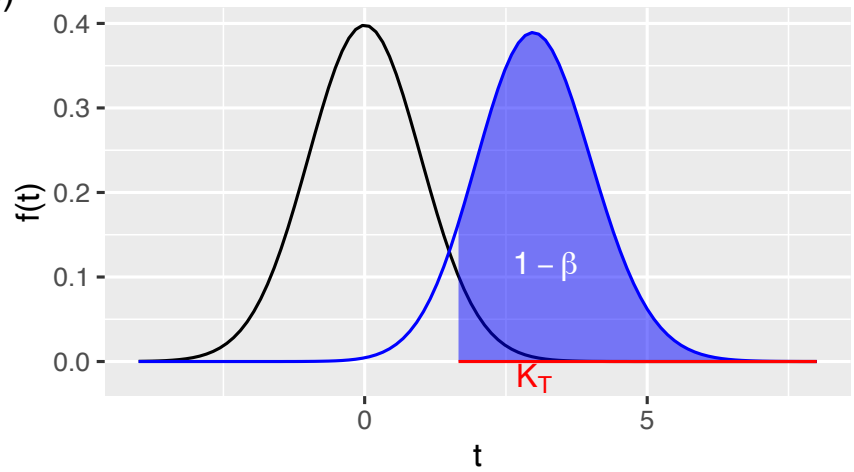
t-Tests IV

(a)



Dichte der Teststatistik unter der extremsten H_0, also falls $\mu = \mu_0$

(b)



In blau: Dichte der Teststatistik, falls δ_{H_1} gilt.

- Graphische Veranschaulichung für einen Einstichproben t-Test mit rechtsseitiger Alternativhypothese.
- (a): Dichte der Teststatistik unter der Voraussetzung, dass $\mu = \mu_0$, also unter der extremsten H_0
- (b): Dichte der Teststatistik unter der Voraussetzung, dass δ_{H_1} der wahre Effekt ist, also unter einem spezifischen Parameterwert unter der H_1

t-Tests V

- Beispiel: Zweistichproben t-Test für unabhängige Stichproben mit

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

- Wir haben ein Signifikanzniveau von $\alpha = 0.005$ gewählt und zwei Stichproben mit $n_1 = n_2 = 100$ erhoben.
- Falls die H_1 wahr wäre und ein kleiner Effekt $\delta_{H_1} = 0.2$ vorliegen würde, wäre die Power (pwr Paket):

```
> pwr.t.test(n = 100, d = 0.2, sig.level = 0.005, type = 'two.sample',
             alternative = 'greater')

     Two-sample t test power calculation

              n = 100
              d = 0.2
      sig.level = 0.005
          power = 0.1203114
    alternative = greater
```

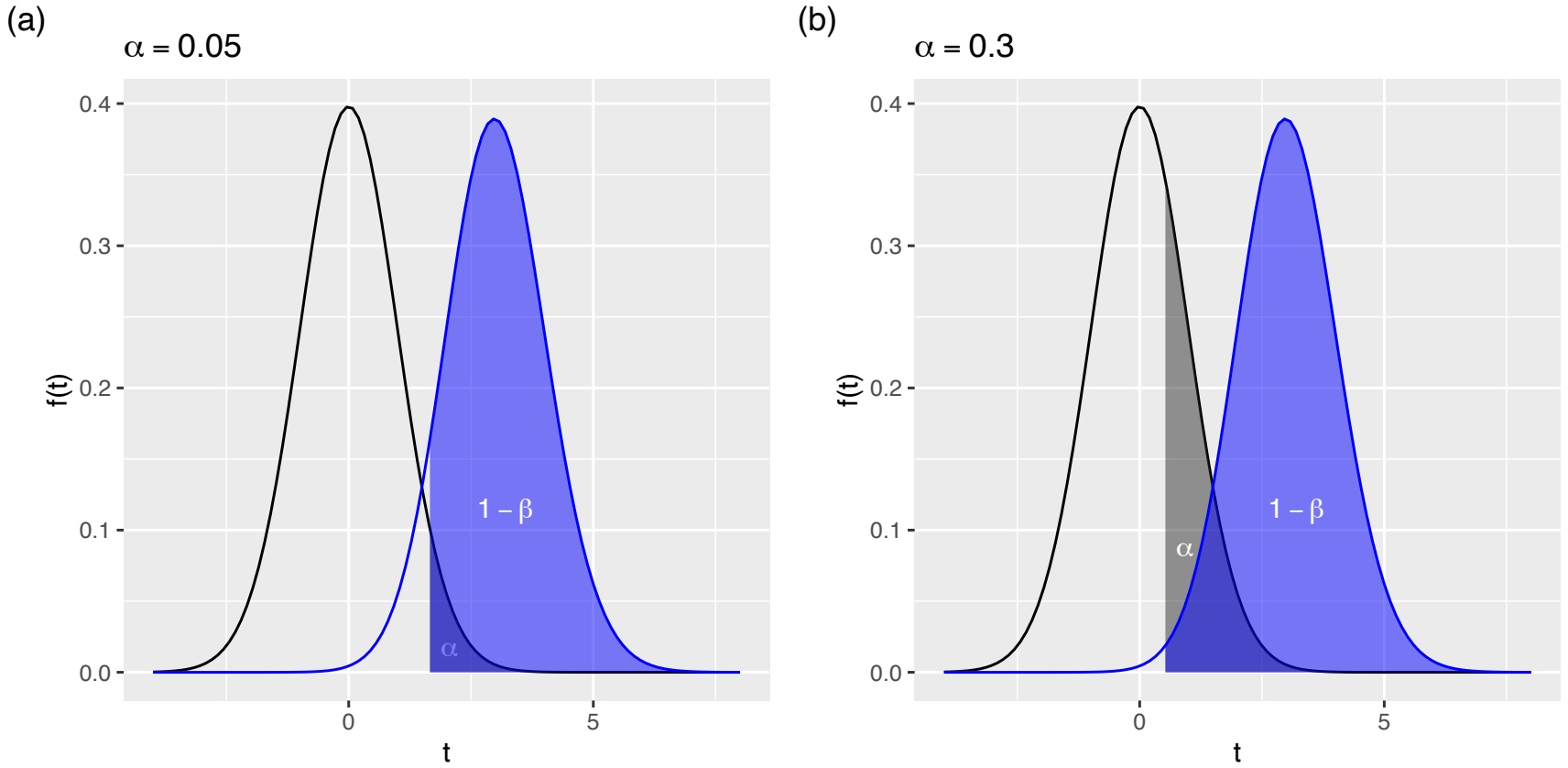
Die Power unseres Hypothesentests wäre in diesem Fall also gleich 0.12 und somit sehr niedrig. Unser Hypothesentest wäre sehr schlecht.

Einflussgrößen der Power

- Die Power eines statistischen Hypothesentests hängt von folgenden Größen ab:
  - dem Signifikanzniveau α,
  - der Stichprobengröße,
  - im Fall der Binomialtests von der Differenz $\pi_{H_1} - \pi_0$, also zwischen dem wahren Parameterwert π_{H_1} und dem (extremsten) Parameterwert π_0 unter der H_0;
    im Fall der t-Tests von der Effektstärke δ_{H_1}.
- Bemerkung: Sowohl $\pi_{H_1} - \pi_0$ als auch δ_{H_1} werden wir im Folgenden als „wahren Effekt“ bezeichnen.

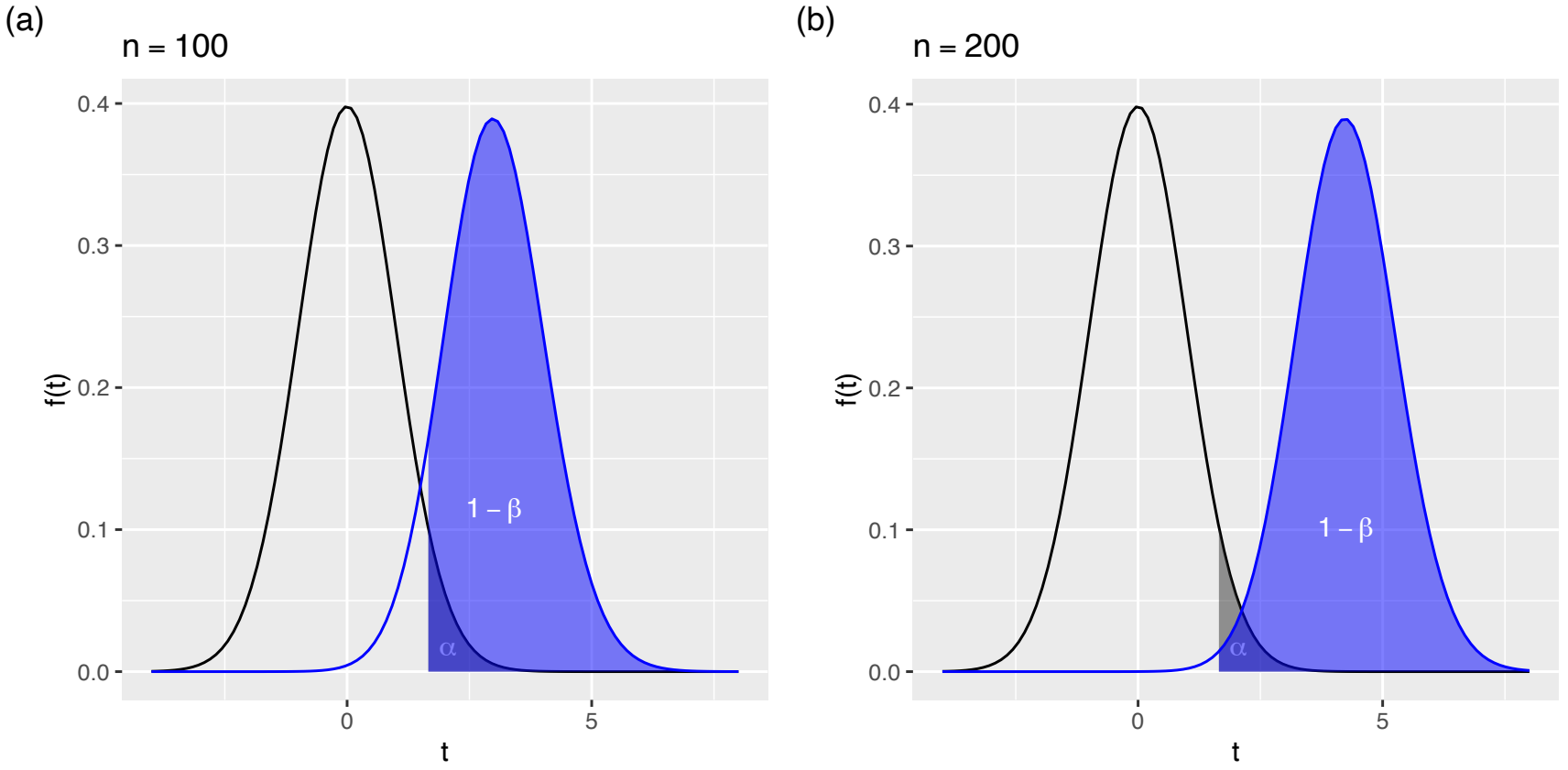
Power und Signifikanzniveau

- **Je größer** das Signifikanzniveau α, **desto größer** die Power.
- Graphische Veranschaulichung für einen rechtsgerichteten Einstichproben t-Test:



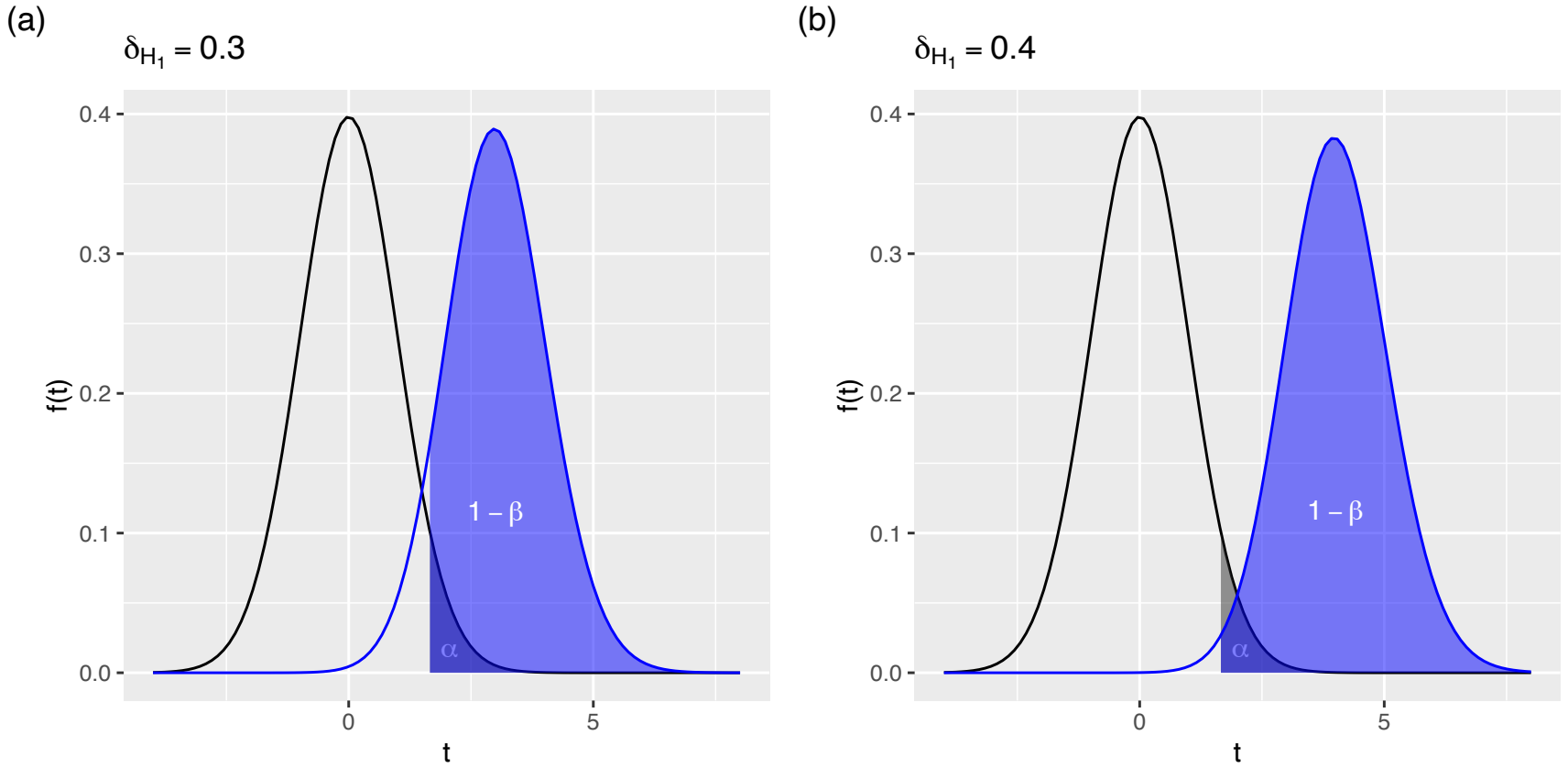
Power und Stichprobengröße

- **Je größer** die Stichprobengröße, **desto größer** die Power.
- Graphische Veranschaulichung für einen rechtsgerichteten Einstichproben t-Test:



Power und wahre Effektgröße

- **Je größer** der wahre Effekt, **desto größer** die Power.
- Graphische Veranschaulichung für einen rechtsgerichteten Einstichproben t-Test:



Zusammenfassung

- **Je größer** das Signifikanzniveau α, **desto größer** die Power.
- **Je größer** die Stichprobengröße, **desto größer** die Power.
- **Je größer** der wahre Effekt, **desto größer** die Power.

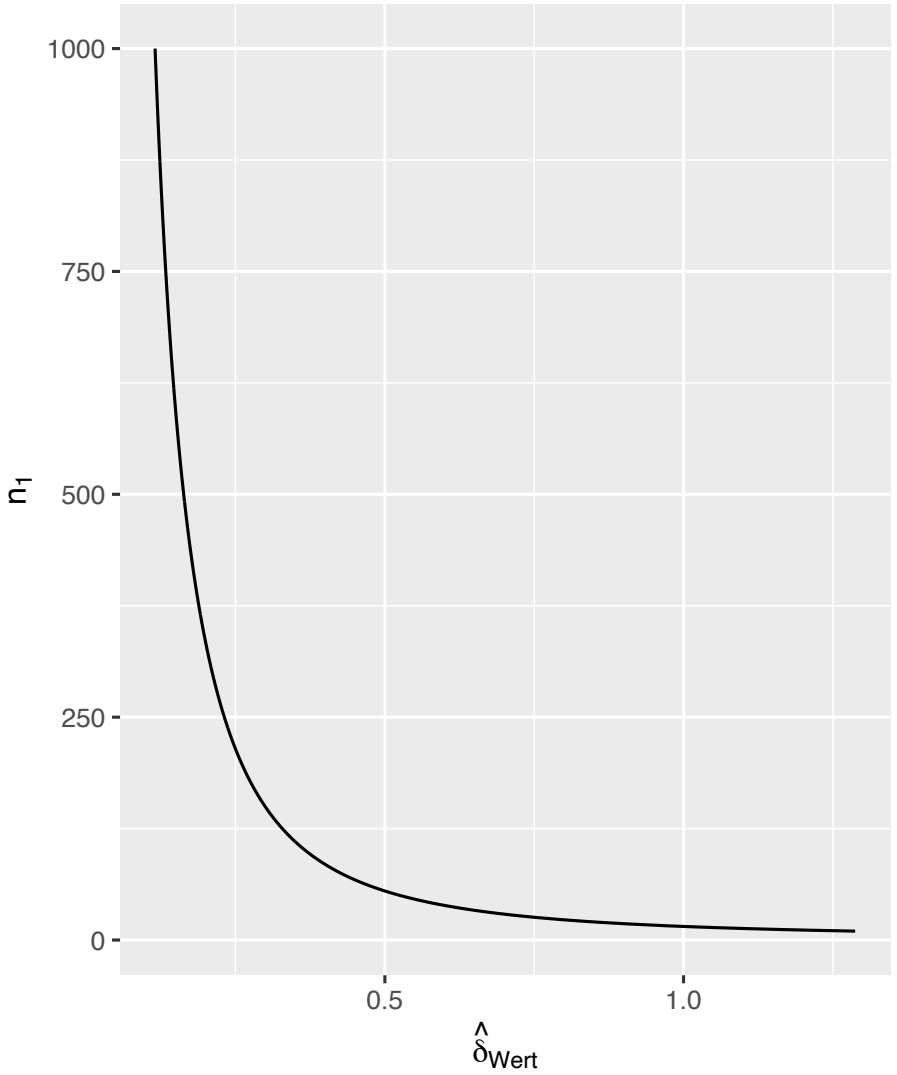
Wahl des Effekts unter der H_1 bei der Berechnung der Power

- Zur Erinnerung: Bei der konkreten Berechnung der Power müssen wir einen Effekt vorgeben.
- Die von uns unter dieser Voraussetzung berechnete Power entspricht nur dann der tatsächlichen Power des Hypothesentests, falls dieser Effekt dem wahren Effekt entspricht.
- Was der wahre Effekt ist, können wir natürlich nie wissen.
- Wir wissen aber: Je größer der wahre Effekt, desto größer die Power.
- Das heißt: Wenn der wahre Effekt größer ist als der von uns für die Berechnung gewählte Effekt, ist die tatsächliche Power größer als die von uns berechnete.
- Falls wir also eine konservative Abschätzung der Power haben wollen, sollten wir für die Berechnung einen kleinen Effekt z.B. $\pi_{H_1} - \pi_0 = 0.1$ oder $\delta_{H_1} = 0.2$ wählen.
- Falls die mit diesem Effekt berechnete Power hoch ist, können wir davon ausgehen, dass die tatsächliche Power für alle größeren Effekte mindestens genauso hoch ist.

Post-hoc Power

- Aber warum ein kleines δ_{H_1} annehmen, wenn sich in unserer Studie vielleicht ein viel größeres $\hat{\delta}_{Wert}$ gezeigt hat?
- Wäre es nicht einfacher, das $\hat{\delta}_{Wert}$ aus unserer Stichprobe zu nehmen und danach zu berechnen, was die Power für einen solchen Effekt gewesen wäre?
- Gerade wenn meine Stichprobe nur klein war, könnte ich doch so vielleicht zeigen, dass die Power für den beobachteten Effekt trotzdem ausreichend groß gewesen wäre?
- Vorsicht! Gerade dann, wenn ein solches Vorgehen gewählt wird, sobald ein Hypothesentest signifikant geworden ist, leitet eine solche nachträglich (post-hoc) berechnete Power in die Irre.
- Unabhängig davon, wie groß ein Effekt wirklich ist, müssen gerade bei kleinen Stichproben große Effekte beobachtet werden damit der Hypothesentest signifikant wird. Bei großen Stichproben reichen bereits kleine Effekte.
- Warum?

Zusammenhang der Teststatistik mit $\hat{\delta}_{Wert}$



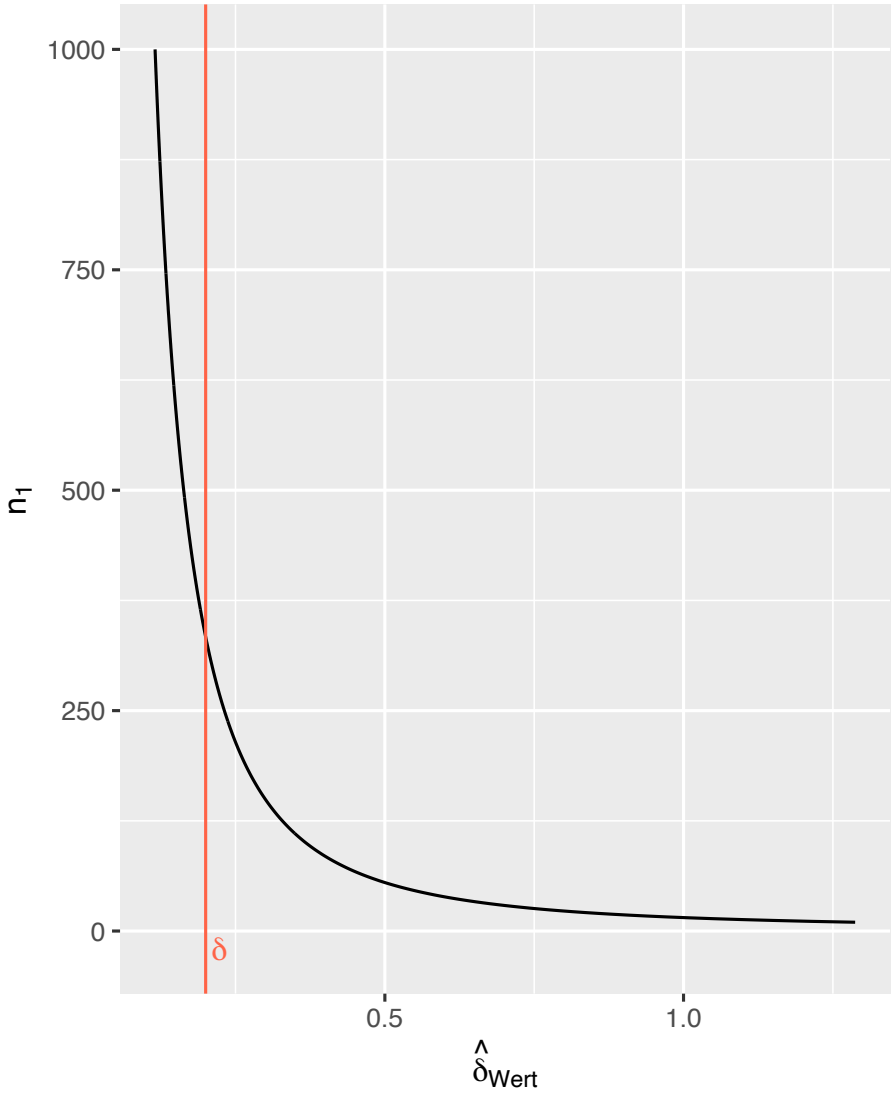
Untergrenze von $\hat{\delta}_{Wert}$ für $t \geq t_{krit}$.

- Ausgangspunkt des Beispiels:

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

- $n_1 = n_2$ bei unabhängigen Stichproben.
- $\alpha = 0.005$
- $t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s^2_{pool}}{n_1} + \frac{s^2_{pool}}{n_2}}} \overset{\mu_0 = 0}{\Longrightarrow} (\bar{x}_1 - \bar{x}_2) = t\sqrt{\frac{2}{n_1}}\sqrt{s^2_{pool}}$
- $\hat{\delta}_{Wert} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2_{pool}}} = \frac{t\sqrt{\frac{2}{n_1}}\sqrt{s^2_{pool}}}{\sqrt{s^2_{pool}}} = t\sqrt{\frac{2}{n_1}}$
- Falls $t = t_{krit}$ wird der Test signifikant.

Warnung vor Post-hoc Power I



In Rot: δ = 0.2.

- Falls tatsächlich die H_1 wahr ist und beispielsweise ein Effekt von $\delta = 0.2$ vorliegt, müssten wir diesen Effekt bei kleinen Stichproben überschätzen, um einen signifikanten Hypothesentest zu erhalten.
- Bei großen Stichproben reichen dagegen bereits kleinere Effektschätzungen aus, damit wir aufgrund eines signifikanten Tests die H_1 auch selbst für wahr halten.
- Verwenden wir also bei kleinen Stichproben und einem signifikanten Hypothesentest $\hat{\delta}_{Wert}$ für die Powerberechnung, überschätzen wir damit die tatsächliche Power.

Warnung vor Post-hoc Power II

- Beispiel: *t*-Test für zwei unabhängige Stichproben mit rechtsgerichteter H_1
- Angenommen der wahre Effekt in der Population beträgt $\delta = 0.2$
- Erhebung von zwei kleinen Stichproben mit $n_1 = n_2 = 10$
- Verwendung eines üblichen Signifikanzniveaus von $\alpha = 0.05$
- Es wird ein Effekt von $\hat{\delta}_{Wert} = 1$ beobachtet
  Bemerkung: $P(\hat{\delta} \geq 1) \approx 0.05$, falls $\delta = 0.2$ und $n_1 = n_2 = 10$
- Der Hypothesentest liefert ein signifikantes Ergebnis (bei $\alpha = 0.05$) mit $p = 0.019$
- Berechnung der Post hoc Power ergibt:

```
> pwr.t.test(d = 1, sig.level = 0.05, n = 10, type = 'two.sample',
+            alternative = 'greater')

     Two-sample t test power calculation

              n = 10
              d = 1
      sig.level = 0.05
          power = 0.6935575
    alternative = greater

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

Warnung vor Post hoc Power III

- Die tatsächliche Power ist in diesem Fall allerdings:

```
> pwr.t.test(d = 0.2, sig.level = 0.05, n = 10, type = 'two.sample',
+            alternative = 'greater')

     Two-sample t test power calculation

              n = 10
              d = 0.2
      sig.level = 0.05
          power = 0.1123273
    alternative = greater

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

- Hätte man die Poweranalyse konservativ mit einem kleinen Effekt unter der H_1 von $\delta_{H_1} = 0.2$ durchgeführt, hätte man erkannt, dass es sich hier um einen sehr schlechten Hypothesentest handelt, der das Gütekriterium einer hohen Power nicht erfüllt.

- Fazit des Beispiels: Die aufgrund der kleinen Stichprobe viel zu niedrige tatsächliche Power von 0.11 wird durch die angegebene Post hoc Power von 0.69 verschleiert.

Stichprobenplanung für Hypothesentests

Wdh.: Gütekriterien für statistische Hypothesentests

- Das Signifikanzniveau α eines statistischen Tests können wir direkt durch die Wahl eines geeigneten kritischen Bereichs festlegen.

- Die Power $1 - \beta$ eines statistischen Tests können wir nur indirekt durch die Größe unserer einfachen Zufallsstichprobe festlegen.

- **Sehr, sehr, sehr wichtig**:

  - Die Verwendung eines statistischen Hypothesentests ist nur dann sinnvoll, falls er **sowohl ein geringes Signifikanzniveau als auch eine hohe Power aufweist**.

  - Der Fehler 1. Art ist **nicht** „wichtiger“ als der Fehler 2. Art.
    Ein statistischer Hypothesentest, der lediglich ein geringes Signifikanzniveau, aber keine hohe Power aufweist, ist kein guter Hypothesentest.

- Frage: Wie stellen wir sicher, dass die Power unseres Hypothesentests hoch ist?

Konsequenzen für die Gütekriterien

- Zur Erinnerung:
  - **Je größer** das Signifikanzniveau α, **desto größer** die Power.
  - **Je größer** die Stichprobengröße, **desto größer** die Power.
  - **Je größer** der wahre Effekt, **desto größer** die Power.
- Ein niedriges Signifikanzniveau ist genau wie eine hohe Power ein Gütekriterium von statistischen Hypothesentests. Dieses können wir also nicht einfach erhöhen, um eine hohe Power zu erhalten.
- Wie groß der wahre Effekt ist, falls die H_1 gilt, können wir nicht beeinflussen.
- Das einzige, was wir dafür tun können, dass unser Hypothesentest eine hohe Power aufweist, ist also, eine **große Stichprobe** zu erheben.
- Frage: Wie groß muss die Stichprobe sein, damit unser Hypothesentest eine von uns vorgegebene Power aufweist?

Stichprobenplanung I

- Genau wie wir für ein gegebenes Signifikanzniveau, eine gegebene Stichprobengröße und einen wahren Effekt die Power eines statistischen Tests bestimmen können, können wir für ein gegebenes Signifikanzniveau, eine gegebene Power und einen gegebenen wahren Effekt die benötigte Stichprobengröße bestimmen.

- Die mathematische Herleitung hierfür ist sehr aufwendig.

- Wir beschränken uns daher wieder auf die praktische Anwendung mit R.

Stichprobenplanung II

- Problem: Wie legen wir den Effekt unter der H_1 für die Stichprobenplanung fest?
- Konservative Lösung: Wir wählen für die Berechnung der benötigten Stichprobengröße einen kleinen Effekt.
- Falls der Effekt unter der H_1 nämlich in Wahrheit größer ist, ist dies nicht so schlimm, da die Power des Hypothesentests für die berechnete Stichprobengröße dann mindestens so groß ist, wie die von uns vorgegebene Power.
- **Wichtig:** Im Gegensatz zur Stichprobenplanung für Konfidenzintervalle hat die Wahl des Effekts δ_{H_1} bei Hypothesentests in vielen Fällen einen starken Einfluss auf die benötigte Stichprobengröße. Da sich die Stichprobenplanung bei Hypothesentests auf ein zentrales Gütekriterium (die Power) auswirkt, sollten wir also auf jeden Fall eine konservative Lösung wählen.
- Bemerkung: Die Wahl des Effekts unter der H_1 kann man sich im Sinne des kleinst-möglichen Effekts vorstellen, der in der konkreten Anwendung praktisch relevant ist. „Falls der tatsächliche Effekt kleiner ist als mein festgelegter Mindesteffekt, dann ist es mir praktisch egal, wenn ich den Effekt mit meiner Studie nicht finde.“

Beispiel I

- Wir hatten auf Folie 35 gesehen, dass für einen Zweistichproben t-Test für unabhängige Stichproben mit Signifikanzniveau $\alpha = 0.005$ und

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

  bei einem kleinen Effekt $\delta_{H_1} = 0.2$ zwei Stichproben mit einer Größe von jeweils 100 zu klein waren, um eine hohe Power zu erhalten.

- Wie groß muss die Stichprobe also sein?

Beispiel II

- Wir wollen, dass unser Hypothesentest mindestens eine Power von 0.8 aufweist.

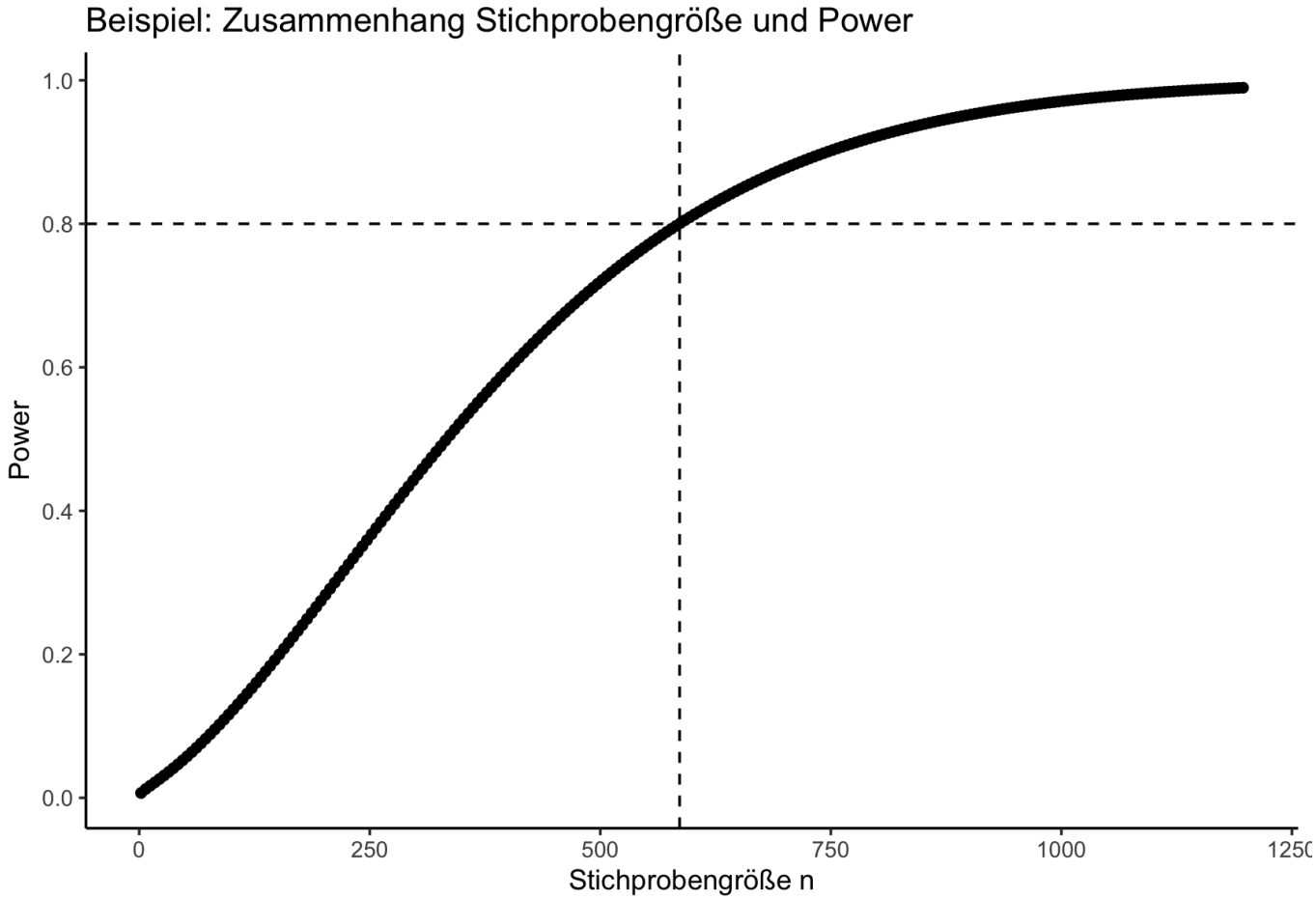
- Berechnung der hierfür benötigten Stichprobengröße in R:

```
> pwr.t.test(d = 0.2, sig.level = 0.005, power = 0.8, type = 'two.sample‘,
             alternative = 'greater')

     Two-sample t test power calculation

              n = 585.6093
              d = 0.2
      sig.level = 0.005
          power = 0.8
    alternative = greater

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

- Der ausgegebene Wert 585.61 entspricht der benötigten Stichprobengröße **pro Stichprobe**. Wir müssen also aufgerundet insgesamt mindestens 586 + 586 = 1172 Personen erheben.

Zusammenhang Power und Stichprobengröße n

- Es besteht ein positiver (monoton steigender) Zusammenhang zwischen der Stichprobengröße und der Power (bei festem α und Effekt δ_{H_1}):



Zusammenfassung

- Mit Effektstärken lassen sich Effekte unabhängig von Einheit und/oder Streuung eines Merkmals leichter interpretieren.
- Die Effektstärke Cohen‘s δ vereinfacht die Interpretation von Mittelwertsunterschieden und kann als Punkt- und Intervallschätzung aus Stichprobendaten geschätzt werden.
- Die Intervallschätzung von Cohen‘s δ erlaubt auch eine Stichprobenumfangsplanung für eine gewünschte erwartete Länge („Präzision“) der Schätzung.
- Die Power ist die Wahrscheinlichkeit dafür, dass wir uns für die H_1 entscheiden, falls diese tatsächlich wahr ist.
- Für konkrete Berechnungen der Power müssen konkrete Parameterwerte unter der H_1 angenommen werden, die aus theoretischen Überlegungen oder Vorstudien abgeleitet werden.
- Die Power wird größer, je größer der angenommene Effekt, das Signifikanzniveau α und/oder die Stichprobengröße ist.
- Um eine gewünschte Power bei gegebenem Signifikanzniveau und gegebenem Effekt zu erhalten, kann eine Stichprobenumfangsplanung durchgeführt werden.