M24 Statistik 1: Wintersemester 2024 / 2025

Vorlesung 12: Power und Fallzahlschätzung

Prof. Matthias Guggenmos

Health and Medical University Potsdam

Statistische Power

Ausgangspunkt

Sie wollen in einer Stichprobe die Hypothese H_1 untersuchen, dass der Mittelwert μ_{med} von Medizinernasenlängen größer ist, als der Bundesdurchschnitt $\mu_0 = 5cm$.

$$H_1 : \mu_{\text{med}} > \mu_0 \qquad H_0 : \mu_{\text{med}} \leq \mu_0$$

Nehmen wir an, dass es den Effekt gibt, H_1 also korrekt ist. Wie groß ist die Wahrscheinlichkeit, die Nullhypothese H_0 auch tatsächlich ablehnen zu können?

- Diese Wahrscheinlichkeit bezeichnen wir als **statistische Power**, **Power** oder **Teststärke**.
- Sie gibt die “Stärke” des Testes an, einen wahren Effekt auch tatsächlich nachzuweisen.

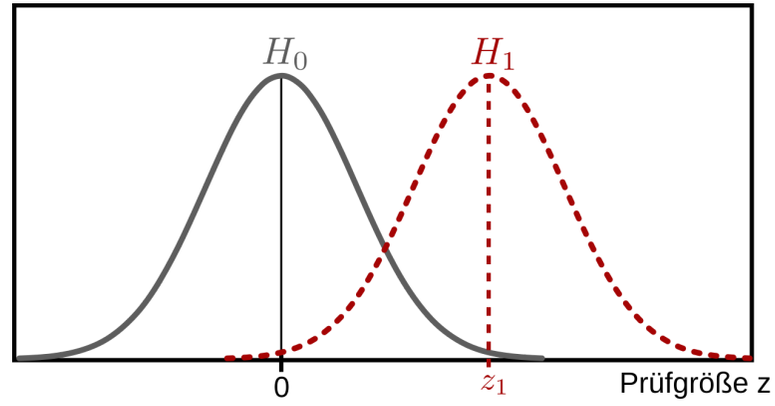
Statistische Power

Definition **Teststärke** oder **statistische Power** bezeichnet die Wahrscheinlichkeit, mit der ein statistischer Test die Nullhypothese **richtigerweise** ablehnt.

- "richtigerweise" heißt hierbei, dass die Nullhypothese H_0 tatsächlich nicht zutrifft, also H_1 gilt.
- Ist die statistische Power des Testes hoch, so können wir zuversichtlich sein, dass wir die Nullhypothese H_0 erfolgreich ablehnen (und H_1 annehmen) werden, wenn sie auch tatsächlich unzutreffend ist.
- Statistische Power ist also eine Wahrscheinlichkeit unter Annahme einer Hypothese — das kennen wir bereits vom Signifikanzniveau α bzw. p-Wert!
  - Wie auch α / p ist statistische Power eine Fläche unter einer Wahrscheinlichkeitsverteilung — in diesem Fall unter der H_1-Verteilung.
  - … was aber ist die H_1-Verteilung?

Statistische Power: Alternativhypothese H_1

- Im ersten Schritt nehmen wir vereinfacht an, dass die statistische Nullhypothesentestung mithilfe eines **z-Testes** durchgeführt werden kann.
- Wir können daher die Wahrscheinlichkeitsdichteverteilung der Nullhypothese H_0 wie gewohnt als Standardnormalverteilung im z-Raum einzeichnen.
- Die H_1-Verteilung unterscheidet sich von der H_0-Verteilung nur durch ihren Mittelwert.
- Ist die Annahme von H_1, dass der untersuchte Effekt *größer 0* ist, so ist die Verteilung der Alternativhypothese nach rechts verschoben.
- Wir bezeichnen den Mittelwert der Wahrscheinlichkeitsverteilung von H_1 vorläufig mit z_1.



Statistische Power: Alternativhypothese H_1

- Über mögliche konkrete (Mittel-)Werte der H_1 haben wir uns im Framework der Nullhypothesentestung keine Gedanken gemacht — hier reichte es, die Nullhypothese zu spezifizieren.
- Für die Berechnung der statistischen Power sind Annahmen über die H_1 nun allerdings notwendig, denn um Wahrscheinlichkeiten unter der H_1-Wahrscheinlichkeitsverteilung zu berechnen, muss ein konkreter Wert für den Mittelwert z_1 dieser Verteilung angenommen werden.
- Im Nasenlängenbeispiel ist der betrachtete Effekt die Differenz $\mu_1 - \mu_0$ und damit $z_1 = \frac{\mu_1 - \mu_0}{se}$.
- Wir verwenden an dieser Stelle das allgemeinere Symbol θ als Platzhalter für unstandardisierte Effekte jeder Art:

$$z_1 = \frac{\theta}{se}$$

- Bei Einzelmessung an einer Stichprobe wie im Nasenlängenbeispiel gilt $se = \sigma/\sqrt{n}$ und damit:

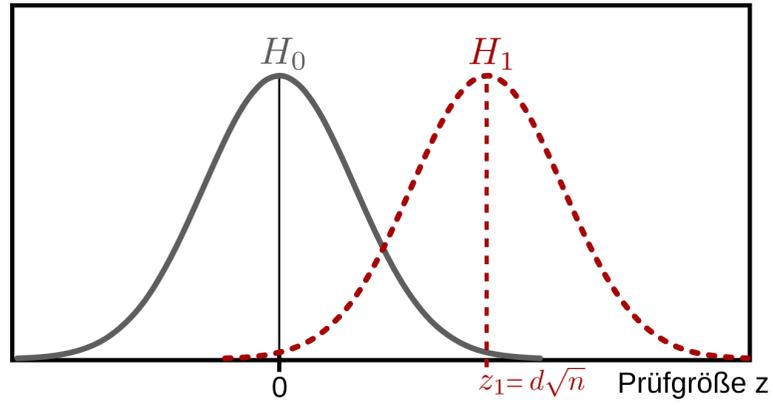
$$z_1 = \frac{\theta}{\sigma/\sqrt{n}} = \frac{\theta}{\sigma}\sqrt{n}$$

Statistische Power: Alternativhypothese H_1

Wir erinnern uns, dass die **Effektstärke** definiert ist als $d = \frac{\theta}{\sigma}$ ist, und so folgt:

$$z_1 = \frac{\theta}{\sigma}\sqrt{n} = d\sqrt{n}$$

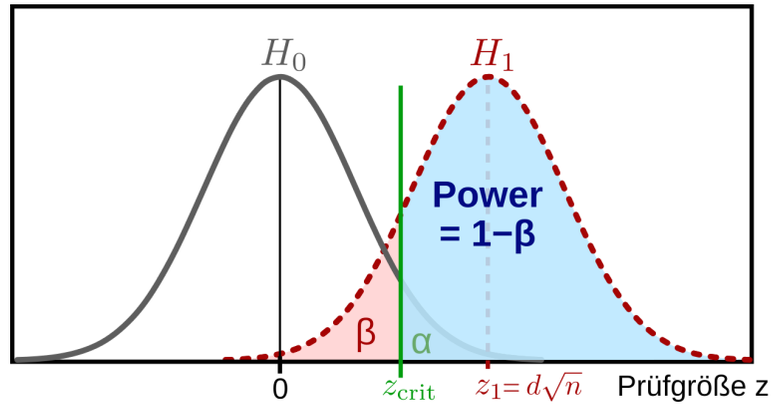
… mit anderen Worten: wir können die H_1 nicht nur durch Angabe eines unstandardisierten Effektes θ definieren, sondern auch durch Angabe einer Effektstärke d!



Nun ist die Wahrscheinlichkeitsverteilung der H_1 im z-Raum definiert (Normalverteilung, Mittelwert $z_1 = d\sqrt{n}$, Standardabweichung 1) und wir können Flächen unter dieser Verteilung berechnen.

Berechnung der statistischen Power

- Wir erinnern uns: statistische Power ist die Wahrscheinlichkeit, dass unter Annahme der Alternativhypothese die Nullhypothese auch tatsächlich abgelehnt wird, also dass gilt: $z > z_{\text{crit}}$.
- In mathematischer Notation:



$$\text{Power} = P(z > z_{\text{crit}} \mid H_1)$$

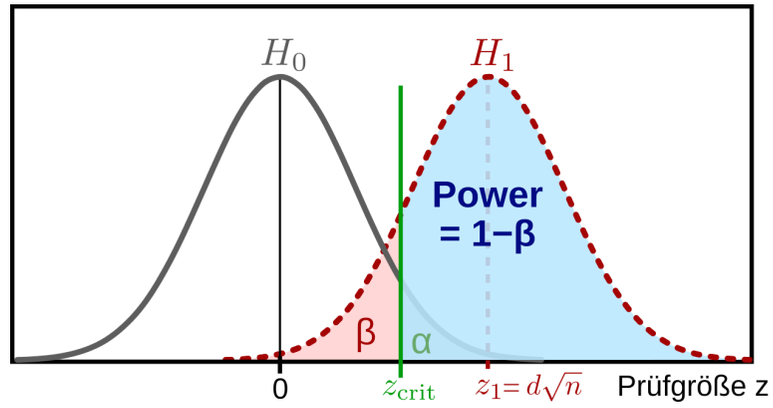
- Statistische Power ist also die Fläche **rechts** von z_{crit} unter der H_1-Verteilung.
- Die Fläche **links** von z_{crit} wird als β bezeichnet (zur Bedeutung von β kommen wir gleich). Entsprechend gilt für die statistische Power:

$$\text{Power} = P(z > z_{\text{crit}} \mid H_1) = 1 - \beta$$

Berechnung der statistischen Power

- Wie groß ist nun die Fläche rechts von z_{crit} unter der H_1-Verteilung? Dazu benötigen wir die kumulative Verteilungsfunktion Φ der um z_1 nach rechts verschobenen Standardnormalverteilung:

$$\Phi(x - z_1)$$



- Die Fläche <u>rechts von einem Wert</u> x ist gegeben als 1 *minus der Wert der Verteilungsfunktion* $\Phi(x)$.
- Entsprechend gilt für die Fläche rechts von z_{crit} (also für die Power):

$$\text{Power} = 1 - \Phi(z_{\text{crit}} - z_1) = 1 - \Phi(z_{\text{crit}} - d\sqrt{n})$$

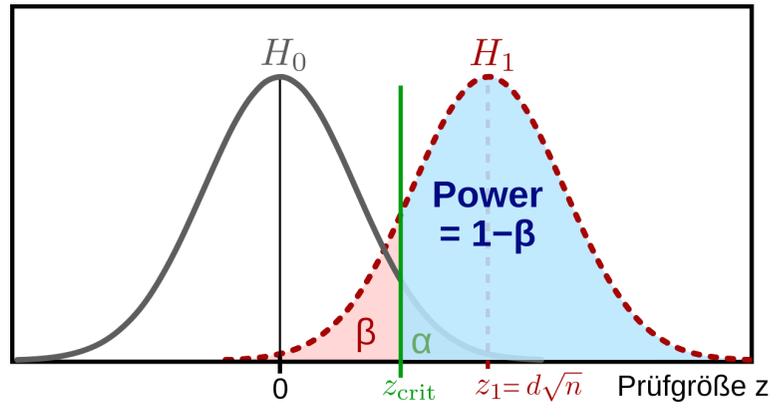
- **Diese Formel gilt für abhängige Messungen und Einzelmessungen.**
- Zur Berechnung der statistischen Power eines Testes benötigen wir also:
  1. Das Signifikanzniveau α zur Bestimmung von z_{crit} (z.B. $z_{\text{crit}} = z_{1-\alpha}$ beim einseitigen Test).
  2. Die angenommene Effektstärke d.
  3. Die Stichprobengröße n.

Berechnung der statistischen Power (unabhängige Messungen)

Auch bei unabhängigen Stichproben / Messungen gilt

$$\text{Power} = 1 - \Phi(z_{\text{crit}} - z_1)$$

Allerdings ist der Standardfehler se, der den Mittelwert $z_1 = \frac{\theta}{se}$ der H_1 bestimmt, ein anderer.



Für unabhängige Stichproben A und B galt:

$$se = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

.. was sich für gleiche Stichprobengrößen $n_A = n_B = n$ und gleiche Stichprobenvarianzen $\sigma_A^2 = \sigma_B^2 = \sigma^2$ vereinfacht zu:

$$se = \sigma\sqrt{\frac{2}{n}}$$

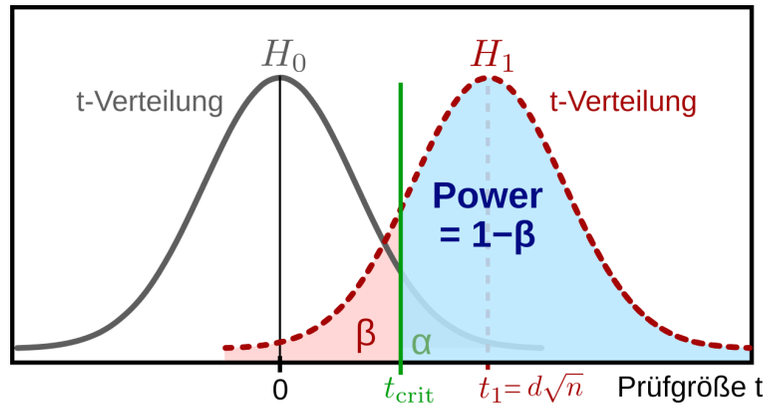
Damit gilt:

$$z_1 = \frac{\theta}{se} = \frac{\theta}{\sigma\sqrt{\frac{2}{n}}} = \frac{\theta}{\sigma}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}}$$

$$\text{Power (unabh. Messungen)} = 1 - \Phi\left(z_{\text{crit}} - d\sqrt{\frac{n}{2}}\right)$$

Berechnung der statistischen Power (t-Test)

- Die statistische Power für einen t-Test berechnet sich analog, indem wir den kritischen z-Wert z_{crit} durch einen kritischen t-Wert t_{crit} ersetzen und die kumulative Normalverteilungsfunktion $\Phi(x)$ durch die kumulative t-Verteilungsfunktion $F_t(x \mid \text{df})$:



$$\text{Power} = 1 - F_t(t_{\text{crit}} - d\sqrt{n} \mid \text{df})$$

$$\text{Power (unabh. Messungen)} = 1 - F_t\left(t_{\text{crit}} - d\sqrt{\frac{n}{2}} \mid \text{df}\right)$$

- Zur korrekten Berechnung der statistischen Power muss die passende Effektstärke d berechnet werden und die Zahl der Freiheitsgrade df entsprechend der Art des t-Testes (⇒ Vorlesung 09 t-Test) gewählt werden.
- Wie üblich gilt bei einseitigen Tests $t_{\text{crit}} = t_{(1-\alpha,\, \text{df})}$ und bei zweiseitigen Tests $t_{\text{crit}} = t_{(1-\frac{\alpha}{2},\, \text{df})}$.

Beispiel

- Sie erhoffen sich, die Nasenlängen von $n = 100$ Medizinstudierende im Rahmen Ihrer Bachelorarbeit zu messen.
- Sie erwarten eine mittlere Effektstärke von $d = 0.5$.
- Wie hoch wäre die statistische Power, wenn Sie zum Vergleich mit dem Bevölkerungsdurchschnitt einen zweiseitigen z-Test mit $\alpha = 0.05$ heranziehen (zweiseitig $\Rightarrow z_{\text{crit}} = z_{1-\frac{\alpha}{2}}$)?

Da es sich um eine Einzelmessung handelt, gilt:

$$\text{Power} = 1 - \Phi(z_{1-\frac{\alpha}{2}} - d\sqrt{n})$$

Aus der z-Tabelle entnehmen Sie $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ und damit:

$$\text{Power} = 1 - \Phi(1.96 - 0.5\sqrt{100}) = 1 - \Phi(1.96 - 5) = 1 - \Phi(-3.04) \overset{\Phi(x)=1-\Phi(-x)}{=}$$

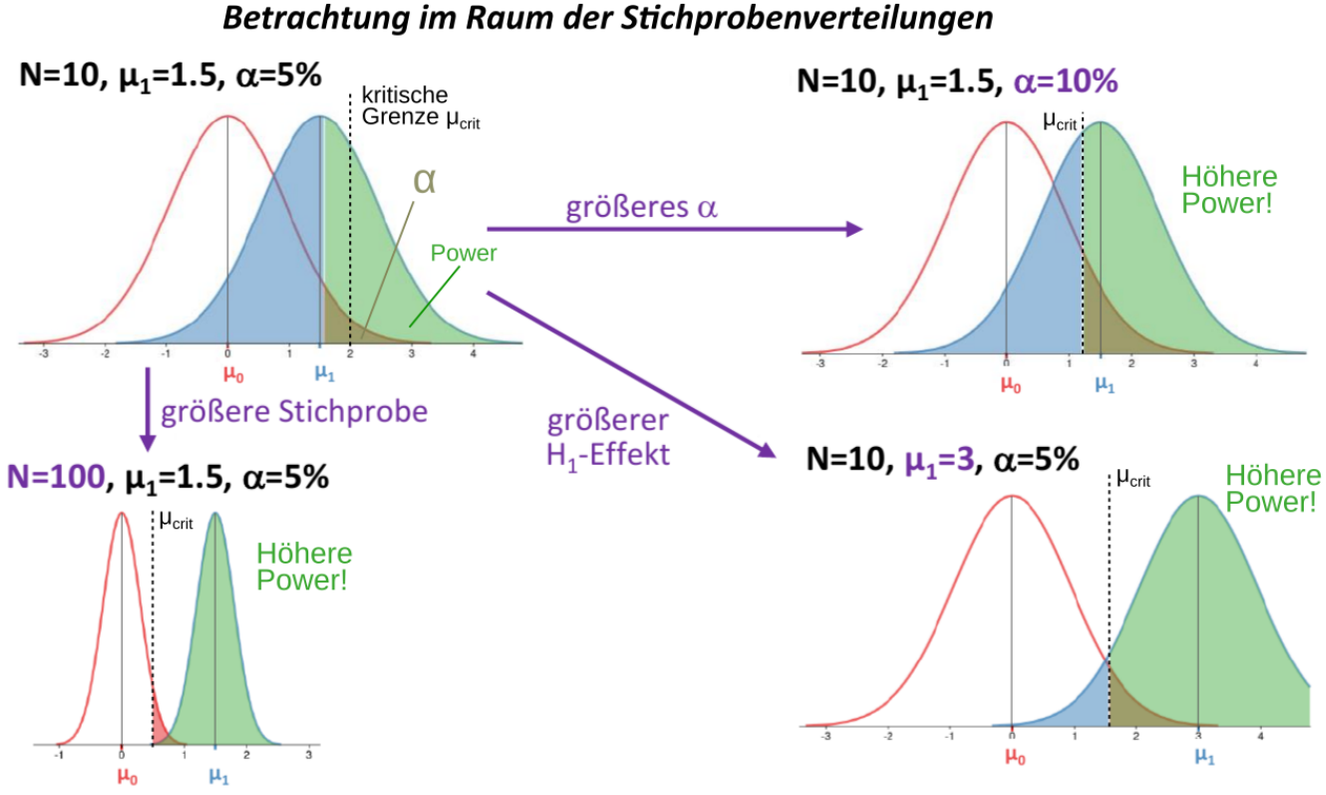
$$= 1 - (1 - \Phi(3.04)) = \Phi(3.04) \overset{\text{(z-Tabelle)}}{=} 0.9988$$

Ihre statistische Power beträgt 99.88% — das ist sehr gut!

Eigenschaften der statistischen Power

Drei Größen beeinflussen die statistische Power, also die Wahrscheinlichkeit einen wahren Effekt (H_1) auch tatsächlich als signifikant zu werten (H_0 abzulehnen):

- Signifikanzniveau α
- Angenommene Effektstärke d (hängt ab von μ_1 und σ)
- Stichprobengröße n

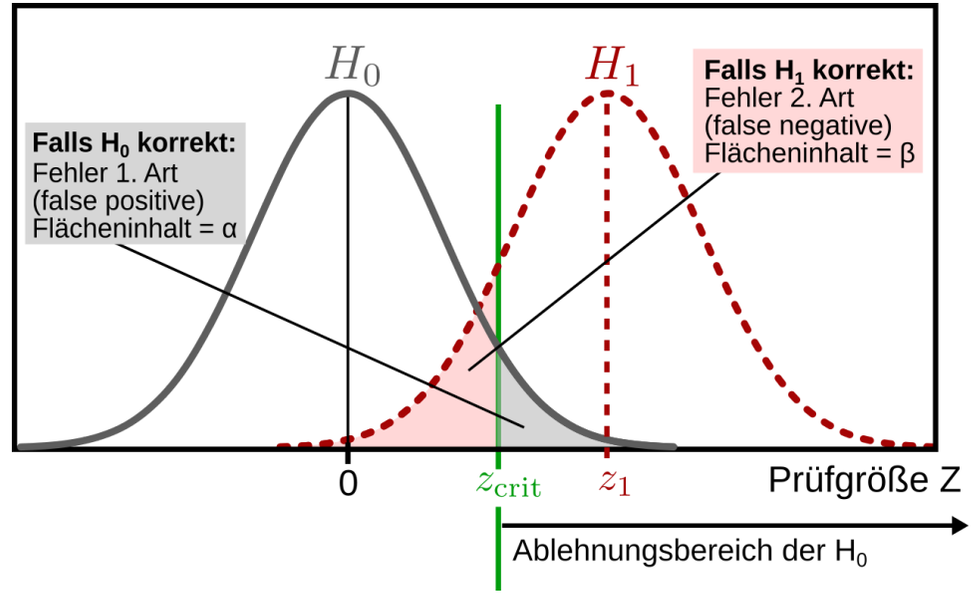


Fehler erster und zweiter Art

Fehler erster und zweiter Art

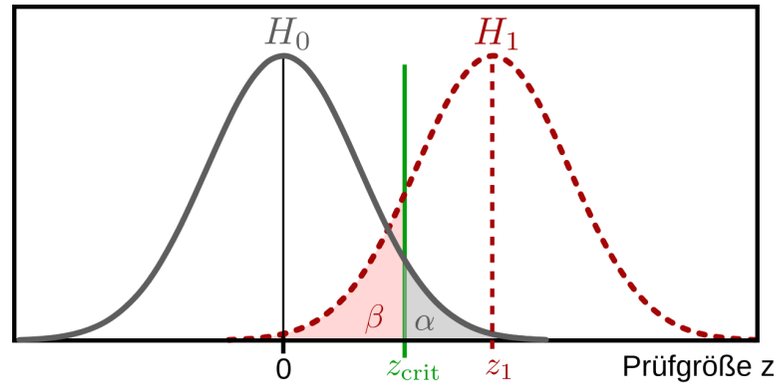
Die Parameter α (Signifikanzniveau) und β (von Power $1 - \beta$) bemessen zwei unterschiedliche Irrtumswahrscheinlichkeiten:

- α-**Fehler:** die Nullhypothese ist wahr und man lehnt sie irrtümlich ab (auch **Fehler erster Art** oder engl. *type 1 error*). Das kennen wir bereits von den Signifikanztests.
  - "Es gibt keinen Effekt, aber der Test ist signifikant."
- β-**Fehler:** die Nullhypothese ist unwahr und man akzeptiert sie irrtümlich (auch **Fehler zweiter Art** oder engl. *type 2 error*).
  - "Es gibt einen Effekt, aber der Test ist nicht signifikant."
- Sowohl α als auch β sind Flächeninhalte unter den Hypothesenverteilungen.

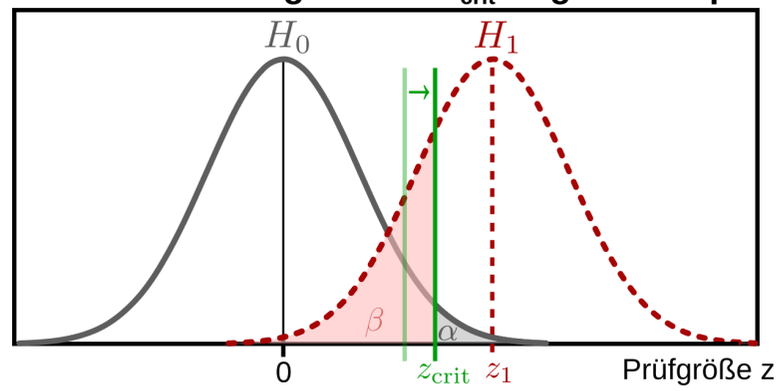


Fehler erster und zweiter Art

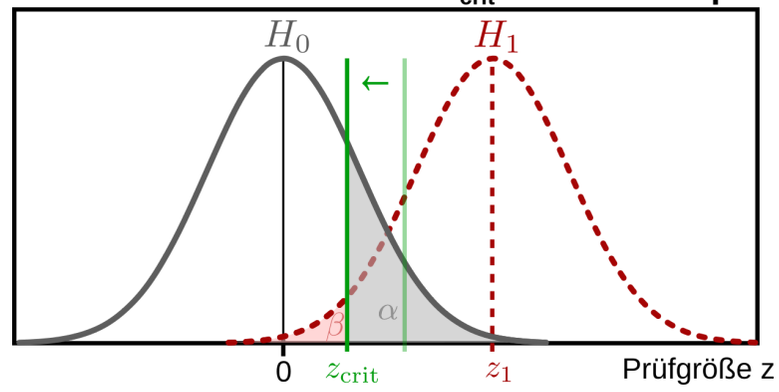
- Wir erkennen: Fehler erster (α) und zweiter (β) Art sind Antagonisten.
- Wählen wir die Irrtumswahrscheinlichkeit α möglichst klein, wollen also ein falsch-positives Ergebnis vermeiden, so erhöht sich die Irrtumswahrscheinlichkeit β automatisch, und ein falsch-negativer Befund wird wahrscheinlicher. Und umgekehrt.
- Ziel muss also sein, ein ausgewogenes Verhältnis beider Fehlerarten zu erreichen.
- Eine häufige Wahl in der Psychologie ist $\alpha = 0.05$ und $\beta = 0.2$.
  - $\beta = 0.2$ entspricht $1 - \beta = 0.8$ und damit einer statistischen Power von 80%.



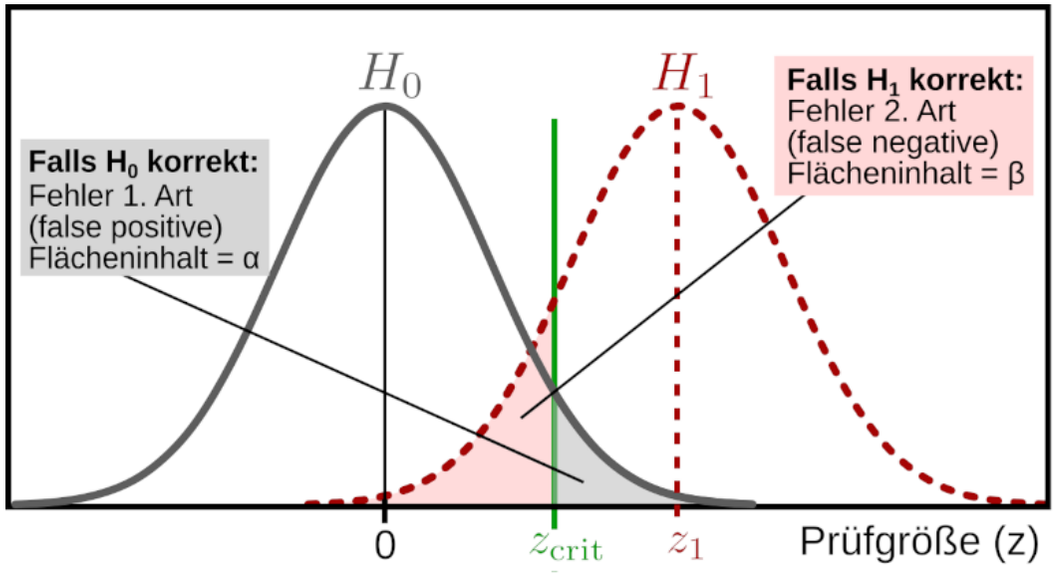
Kleineres α → größeres z_{crit} → größeres β



Größeres α → kleineres z_{crit} → kleineres β



Fehler erster und zweiter Art



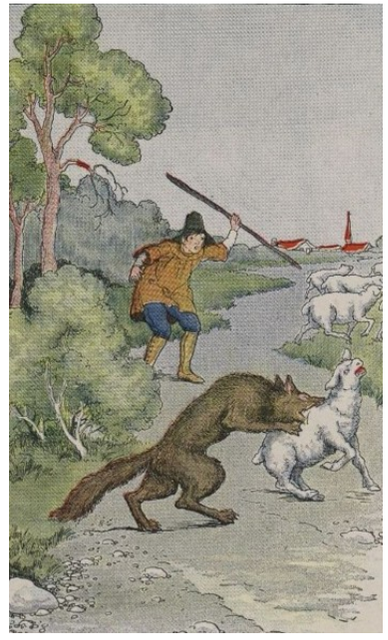
		Entscheidung aufgrund der Stichprobe	
		Entscheidung für die H_0	**Entscheidung für die H_1**
Verhältnisse in der Population (unbekannt)	**In der Population gilt die H_0**	Korrekte Entscheidung *true negative*	**α-Fehler („Fehler erster Art“)** *false positive*
	In der Population gilt die H_1	**β-Fehler („Fehler zweiter Art“)** *false negative*	Korrekte Entscheidung *true positive*

Fehler erster und zweiter Art

Merkregel auf Basis der Fabel “Der Hirtenjunge und der Wolf”:

> Die Hauptperson der Fabel ist ein Hirtenjunge, der aus Langeweile beim Schafehüten laut „Wolf!“ brüllt. Als ihm daraufhin Dorfbewohner aus der Nähe zu Hilfe eilen, finden sie heraus, dass falscher Alarm gegeben wurde und sie ihre Zeit verschwendet haben (**Fehler erster Art**). Als der Junge nach einiger Zeit wirklich einem Rudel Wölfe begegnet, nehmen die Dorfbewohner die Hilferufe nicht mehr ernst und bleiben bei ihrem Tagwerk (**Fehler zweiter Art**). Die Wölfe fressen die ganze Herde und in manchen Versionen der Fabel auch den Jungen.

Quelle: Wikipedia[1]



Die Fabel stammt von Äsop, einem griechischen Dichter (6.Jhd. v. Chr.).

Fallzahlschätzung

Ausgangspunkt

- Zurück zum Ausgangspunkt, dem Nasenlängenbeispiel.
- Eine zentrale Frage der Studienplanung ist die Stichprobengröße n (auch Fallzahl).
- Wie hoch sollte n gewählt werden? $n = 5$? $n = 20$? $n = 1000$?

- Um die optimale Stichprobengröße n zu schätzen, müssen die Irrtumswahrscheinlichkeiten α und β festgelegt werden.
- Es scheint ein nahezuliegenes Ziel, eine möglichst hohe statistische Power $1 - \beta$ zu erreichen. Aber wir erinnern uns: **je höher die statistische Power $1 - \beta$, desto kleiner der Fehler zweiter Art β, desto größer der Fehler erster Art α**.
- Man erkauft sich also höhere statistische Power mit einer höheren Irrtumswahrscheinlichkeit α (=Signifikanzniveau).
- Das bedeutet: es ist ein sinnvoller Kompromiss zwischen α und β notwendig.
  - z.B. $\alpha = 0.05$ und $\beta = 0.2$

Berechnung der notwendigen Fallzahl (z-Test)

- Wir betrachten zunächst den Fall abhängige Messungen oder Einzelmessung.
- Durch Festlegung der Power auf einen Wert $1-\beta$ gilt:

$$\text{Power} = 1 - \Phi(z_{\text{crit}} - d\sqrt{n}) \overset{!}{=} 1 - \beta$$

- Oder äquivalent:

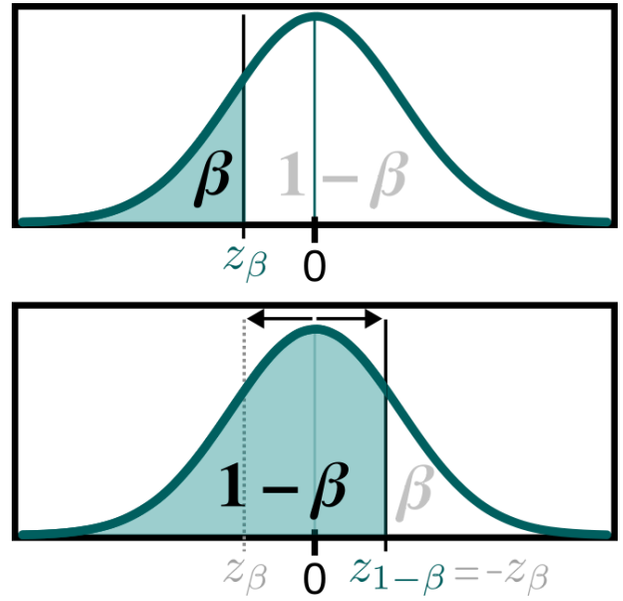
$$\Phi(z_{\text{crit}} - d\sqrt{n}) = \beta$$

- Ziel muss nun sein, die Gleichung nach n aufzulösen.
- Damit $\Phi(\,.\,.\,) = \beta$, muss das Argument $(\,.\,.\,)$ gleich z_β sein:

$$z_{\text{crit}} - d\sqrt{n} = z_\beta \overset{(Symmetrie)}{=} -z_{1-\beta}$$

$$\text{Es gilt also:} \quad d\sqrt{n} = z_{\text{crit}} + z_{1-\beta}$$

$$\text{Nach } n \text{ aufgelöst:} \quad n = \left(\frac{z_{\text{crit}} + z_{1-\beta}}{d}\right)^2$$



Berechnung der notwendigen Fallzahl (z-Test)

Damit sind wir am Ziel. Die Gleichung

$$n = \left(\frac{z_{\text{crit}} + z_{1-\beta}}{d} \right)^2$$

gibt die notwendige Stichprobenzahl für einen **z-Test bei abhängigen Messungen oder Einzelmessungen**.

Die Werte z_{crit} und $z_{1-\beta}$ können beispielsweise in der z-Tabelle nachgeschlagen werden. Wie immer gilt bei einseitigen Tests $z_{\text{crit}} = 1 - \alpha$ und bei zweiseitigen Tests $z_{\text{crit}} = 1 - \frac{\alpha}{2}$.

Anhand der Formel zeigt sich der **enge Zusammenhang der Fehlerraten α und β und der Stichprobengröße n**.

Abhängige Messungen: Einfluss der Korrelation ρ

Beim Vergleich zweier abhängigen Messungen A und B nimmt die Formel $n = \left(\frac{z_{\text{crit}}+z_{1-\beta}}{d}\right)^2$ an, dass die Effektstärke anhand der Differenzenstandardabweichung σ_Δ standardisiert ist:

$$d = \frac{\theta}{\sigma_\Delta}$$

Unter Annahme gleicher Varianzen $\sigma_A^2 = \sigma_B^2 = \sigma^2$ gilt für σ_Δ:

$$\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B} = \sqrt{2\sigma^2 - 2\rho\sigma^2} = \sigma\sqrt{2(1-\rho)}$$

Damit gilt:

$$d = \frac{\theta}{\sigma_\Delta} = \frac{\theta}{\sigma\sqrt{2(1-\rho)}} = \frac{\theta}{\sigma}\frac{1}{\sqrt{2(1-\rho)}}$$

Und:

$$n = \left(\frac{z_{\text{crit}} + z_{1-\beta}}{\frac{\theta}{\sigma}\frac{1}{\sqrt{2(1-\rho)}}}\right)^2 = 2(1-\rho)\left(\frac{z_{\text{crit}} + z_{1-\beta}}{\frac{\theta}{\sigma}}\right)^2$$

Abhängige Messungen: Einfluss der Korrelation ρ

- In der Form

$$n = 2(1-\rho)\left(\frac{z_{\text{crit}} + z_{1-\beta}}{\frac{\theta}{\sigma}}\right)^2$$

- … verdeutlicht die Formel einen wichtigen Aspekt bei abhängigen Messungen: **bei gegebener unstandardisierter Effektstärke θ, hängt die Stichprobengröße von der Korrelation ρ der beiden abhängigen Messungen ab**.
- **Je höher die Korrelation ρ**, also je gleichsinniger sich die Versuchspersonen in den Bedingungen A und B verändern, **desto geringer die erforderliche Stichprobengröße**.
- Wie in der Vorlesung zu Effektstärken thematisiert, schlagen manche Autoren vor, die Effektstärke beim Vergleich abhängiger Messungen an der einfachen Standardabweichung σ zu standardisieren, also $d^* = \frac{\theta}{\sigma}$.
- In diesem Fall gilt für die Stichprobenschätzung bei abhängigen Messungen:

$$n = 2(1-\rho)\left(\frac{z_{\text{crit}} + z_{1-\beta}}{d^*}\right)^2$$

- Bei Verwendung von d^* ist der Einfluss der Korrelation ρ auf die Fallzahl n transparenter.

Berechnung der notwendigen Fallzahl (z-Test): Beispiel

- Sie wollen nun die Fallzahl in Ihrer Nasenlängenstudie so bestimmen, dass gilt: $\alpha = 0.05, \beta = 0.33$
- Sie erwarten eine mittlere Effektstärke von $d = 0.5$.
- Sie entscheiden sich für einen zweiseitigen z-Test (d.h. $z_{\text{crit}} = z_{1-\frac{\alpha}{2}}$).
- Im ersten Schritt müssen Sie nun die kritischen z-Werte $z_{1-\frac{\alpha}{2}}$ und $z_{1-\beta}$ bestimmen, also $z_{0.975}$ und $z_{0.67}$.
- Aus der z-Tabelle entnehmen Sie: $z_{0.975} = 1.96$ und $z_{0.67} = 0.44$.

Alle notwendigen Parameter sind bestimmt und Sie können die Fallzahl berechnen:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{d} \right)^2 = \left(\frac{1.96 + 0.44}{0.5} \right)^2 = 23.04$$

Es ist gute Praxis den Wert aufzurunden (außer Sie finden eine 0.04-Versuchsperson 😉). Damit haben Sie also die notwendige Fallzahl zu $n = 24$ bestimmt!

Prüfen Sie das Ergebnis mit dem Power-Modul in JASP nach!

Unabhängige Messungen

Wir erinnern uns, dass sich bei der Power-Formel unabhängiger Messungen / Stichproben ein Faktor 2 eingeschlichen hatte:

$$\text{Power (unabh. Messungen)} = 1 - \Phi\left(z_{\text{crit}} - d\sqrt{\frac{n}{2}}\right)$$

Diesen Faktor finden wir auch in der Fallzahl-Formel für unabhängige Messungen an zwei Gruppen A und B wieder:

$$n(\text{unabh. Messungen}) = 2\left(\frac{z_{\text{crit}} + z_{1-\beta}}{d}\right)^2$$

Zu beachten ist, dass n hier die Stichprobengröße einer Gruppe ist (wir hatten angenommen $n_A = n_B = n$)! Insgesamt müssen also $2n$ Versuchspersonen getestet werden.

Der Faktor 2 in der Formel für die Fallzahlschätzung unabhängiger Stichproben zeigt, dass bei einem between-subject Design mit zwei Stichproben etwa **doppelt so viele Versuchspersonen pro Gruppe** notwendig sind (für eine gegebene Effektstärke d). Da es beim between-subject Design zudem **zwei Gruppen** gibt, kann die **notwendige Stichprobengröße sogar um einen Faktor 4 höher** sein!

Berechnung der notwendigen Fallzahl (t-Test)

Beim t-Test abhängiger Messungen gilt analog für ein- und zweiseitige Tests:

$$n_{\text{einseitig}} = \left(\frac{t_{(1-\alpha,\text{df})} + t_{(1-\beta,\text{df})}}{d} \right)^2 \qquad n_{\text{zweiseitig}} = \left(\frac{t_{(1-\frac{\alpha}{2},\text{df})} + t_{(1-\beta,\text{df})}}{d} \right)^2$$

Allerdings: die Zahl der Freiheitsgrade df hängt selbst von n ab (z.B. $\text{df} = n - 1$ beim Einstichproben-t-Test)!

- Beim t-Test kann die Gleichung daher nicht analytisch nach n aufgelöst werden.
- Statistikprogramme verwenden aus diesem Grund numerische / iterative Verfahren um n zu bestimmen.
- Grundsätzlich gilt: bei gegebenem α sind die kritischen t-Werte immer etwas größer als die kritischen z-Werte und die notwendige Fallzahl n beim t-Test ist damit größer als beim z-Test!
  - Das macht Sinn, denn beim t-Test besteht zusätzliche Unsicherheit bezüglich des Standardfehlers $\hat{se}$, die durch eine höhere Fallzahl n "kompensiert" werden muss.
  - Der Unterschied zwischen kritischen t- und z-Werten ist um so größer, je kleiner die Fallzahl n. Bei kleinen Studien kann somit die Fallzahlberechnung für den t- und z-Test sehr unterschiedliche Werte liefern.
  - Umgekehrt gilt: ist ohnehin klar, dass die Fallzahl recht groß sein wird (Faustregel $n > 30$), so kann auch im Fall des t-Testes näherungsweise die analytische Fallzahlformel des z-Testes verwendet werden.

Festlegung der Effektstärke

Festlegung der Effektstärke d

- Eine Frage haben wir bislang aufgeschoben: zur Berechnung der Fallzahl n muss die angenommene Effektstärke d festgelegt werden.
  - Die Effektstärke d bestimmt den Mittelwert der H_1-Verteilung, z.B. beim t-Test: $z_1 = d\sqrt{n}$.
- Gretchenfrage: woher soll man die Effektstärke d ***vor*** der Studie kennen — ist nicht gerade das ***Ziel*** von Wissenschaft, die Stärke von Effekten zu untersuchen?
- Dies ist tatsächlich ein Dilemma! Drei Ansätze haben sich in der Forschung herauskristallisiert:
  1. **Schätzung der erwarteten Effektstärke auf Basis vorheriger Forschung**.
  2. **Festlegung einer smallest effect size of interest (SESOI)**, also der minimalen Effektstärke, die noch von wissenschaftlicher (Grundlagenforschung) oder praktischer (Anwendung, klinisch) Relevanz wäre.
  3. **Pilotstudie**: seltener, da Pilotstudien selten groß genug sind, um Effektstärken mit adäquater Präzision bestimmen zu können.

Schätzung der Effektstärke auf Basis vorheriger Forschung

- In manchen Fällen ist es möglich, die erwartete Effektstärke in einer Studie auf Basis früherer identischer oder ähnlicher Studien zu schätzen.
- Beispiel: nehmen wir an, es gab bereits drei vorherige Studien, die Mediziner-Nasenlängen mit dem Bevölkerungsdurchschnitt verglichen haben. Die Studien berichten folgende Effektstärken:

	Studie 1	**Studie 2**	**Studie 3**
Stichprobengröße n	20	80	50
Effektstärke d	0.8	0.1	0.2

- Bei mehreren Studien bietet es sich an, die erwartete Effektstärke auf Basis eines an der Stichprobengröße n gewichteten Mittelwertes zu bestimmen:

$$\bar{d} = \frac{\sum n_i \cdot d_i}{\sum n_i} = \frac{20 \cdot 0.8 + 80 \cdot 0.1 + 50 \cdot 0.2}{20 + 80 + 50} \approx 0.23$$

- Für Ihre Fallzahlberechnung würden Sie in diesem Fall also eine Effektstärke $d = 0.23$ zugrundelegen.

Schätzung der Effektstärke auf Basis vorheriger Forschung

Die Schätzung der Effektstärke auf Basis vorheriger Studien ist häufig problematisch:

- **Vergleichbare Studien zu finden ist schwierig**.
- Selbst wenn auf dem Papier vergleichbare Studien gefunden werden, ist es häufig **schwierig zu beurteilen, ob die berichteten Effektstärken auf die eigene Studie übertragbar sind**.
  - Sind die Interventionen wirklich vergleichbar? Sind die Untersuchungspopulationen vergleichbar? Wenn vorherige Studien älter sind: sind die Effekte auf die "heutige Zeit" übertragbar? Usw.
- Der **Publikationsbias** führt dazu, dass Effekte in veröffentlichten Studien die **wahren Effekte häufig überschätzen**. Werden überschätzte Effekte zugrundegelegt, so liefert die Fallzahlberechnung erwartbar zu kleine Stichprobengrößen n!
- **Effektschätzungen vorheriger Studien sind häufig selbst mit großer Unsicherheit verbunden**. Tatsächlich sind Effektstärken notorisch unpräzise, da Sie i.d.R. auf dem Verhältnis eines geschätzen unstandardisierten Effektes $\hat{\theta}$ (z.B. Mittelwertdifferenz) UND einer geschätzten Streuung $\hat{\sigma}$ basieren.
  - Beispiel: eine Studie berichtet eine Effektstärke von $d = 0.2$. Ein genauerer Blick verrät aber, dass diese Effektstärke mit einem Konfidenzintervall von CI95=[0; 0.4] verbunden ist. D.h. sowohl $d = 0$ als auch $d = 0.4$ wären statistisch kompatibel mit der berichteten Effektstärke der Studie. Legen Sie auf Basis dieser Studie $d = 0.2$ zugrunde, so unterschätzen (oder überschätzen) Sie die notwendige Fallzahl ggf. massiv.

Festlegung einer smallest effect size of interest (SESOI)

- Die *smallest effect size of interest* (SESOI) ist diejenige Effektstärke, die noch als relevant für eine wissenschaftliche Theorie oder eine praktische Anwendung eingeschätzt wird.

Ein Bürgerbüro testet ein neues Zuordnungssystem von Kunden zu Sachbearbeitern, um Wartezeiten zu verkürzen. Aktuell beträgt die mittlere Wartezeit 10 Minuten mit einer Standardabweichung $\sigma = 5\,min$. Da eine Zeitersparnis unter $\theta_{\text{SESOI}} = 1\,min$ pro Kunde als praktisch nicht relevant eingeschätzt wird, wird dies als der kleinste (unstandardisierte) Effekt von Relevanz gewertet. Der standardisierte Effekt d_{SESOI} lautet damit:

$$d_{\text{SESOI}} = \frac{\theta_{\text{SESOI}}}{\sigma} = \frac{1\,min}{5\,min} = 0.2$$

(es wird hierbei angenommen, dass sich die Streuung $\sigma = 5\,min$ nicht verändert)

Beck's Depression Inventory (BDI)

Das National Institute for Health and Care Excellence (NICE) in England gibt an, dass eine Reduktion des BDI-Scores *um mindestens 3 Punkte* als klinisch gerade noch signifikant einzuschätzen ist (d.h. $\theta_{\text{SESOI}} = 3\,Punkte$). Die Streuung von BDI-Scores in Patientenstichproben beträgt ca. $\sigma = 10\,Punkte$. Damit ergibt sich:

$$d_{\text{SESOI}} = \frac{\theta_{\text{SESOI}}}{\sigma} = \frac{3\,Punkte}{10\,Punkte} = 0.3$$

Festlegung einer smallest effect size of interest (SESOI)

- Häufig ist es allerdings schwer, eine SESOI zu berechnen bzw. zu begründen.
  - Hinzu kommt, dass die Standardabweichung σ häufig nicht bekannt ist, die benötigt wird um das unstandardisierte θ_{SESOI} in eine standardisierte Effektstärke d_{SESOI} zu überführen.
- Als Notnagel kann man sich in diesem Fall an den Effektstärken-Niveaus nach Cohen orientieren:

Cohen's d	< 0.2	ab 0.2	ab 0.5	ab 0.8
Interpretation	Trivialer Effekt	Kleiner Effekt	Mittlerer Effekt	Großer Effekt

- Sie könnten sich beispielsweise dafür entscheiden, Ihre Stichprobengröße so zu wählen, dass Sie einen "mittleren" Effekt $d_{\text{SESOI}} = 0.5$ mit einer Power von 80% detektieren (Signifkanzniveau z.B. $\alpha = 0.05$ einseitig):

$$n_{\text{SESOI}} = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{d_{\text{SESOI}}}\right)^2 \underset{(\beta=0.2)}{\overset{(\alpha=0.05)}{=}} \left(\frac{z_{0.95} + z_{0.8}}{d_{\text{SESOI}}}\right)^2 = \left(\frac{1.64 + 0.84}{0.5}\right)^2 = 24.6$$

- Diese "Poor man's" Fallzahlschätzung gibt Ihnen zumindest einen groben Eindruck bezüglich der erforderlichen Fallzahl n.
- Es gilt: *(Beinahe) Jede Fallzahlschätzung ist besser als keine Fallzahlschätzung.*

Festlegung der Stichprobengröße anhand des Konfidenzintervalls

- Eine weitere Alternative zur Festlegung der Stichprobengröße, ist nicht einen statistischen Test zum Kriterium zu machen, sondern die Präzision des gemessenen Effektes $\hat{\theta}$.

Wir erinnern uns an die Definition des Konfidenzintervalls:	$CI = \hat{\theta} \pm t_{\text{crit}} \cdot \hat{se}$
Bei einer Einzelmessung gilt $\hat{se} = \frac{\hat{\sigma}}{\sqrt{n}}$ und damit:	$CI = \hat{\theta} \pm t_{\text{crit}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$
Ist das Ziel, dass das Konfidenzintervall erwartbar die Größe $\pm x$ hat, so gilt:	$x = t_{\text{crit}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$

- Letztere Gleichung können wir nach n auflösen:

$$n = \left(\frac{t_{\text{crit}} \hat{\sigma}}{x} \right)^2$$

- Die Formel gibt die Stichprobengröße n, die notwendig ist, damit das Konfidenzintervall erwartbar die Größe $\pm x$ haben wird, also $[\hat{\theta} - x; \hat{\theta} + x]$.

Die Formel gilt nur für Einzelmessungen. Für andere Designs (abhängige / unabhängige Messungen) muss zur Herleitung der Formel der jeweils passende Standardfehler $\hat{se}$ herangezogen werden.

Fallzahl gegeben: Bestimmung der minimalen detektierbaren Effektstärke

In manchen Fällen ist die Stichprobengröße durch externe Faktoren vorgegeben:

- Sie haben nur X Wochen Zeit in Ihrer Bachelorarbeit und können daher nur n Versuchspersonen testen.
- Sie untersuchen eine bestimmte Patientengruppe mit seltener Erkrankung X und können von dieser an Ihrem Standort nur n Versuchspersonen testen.
- Sie haben nur X Geld in Ihrem Forschungsbudet und können daher nur n Versuchspersonen bezahlen.

In diesem Fall ist es sinnvoll die inverse Frage zu stellen: welche kleinste Effektstärke kann mit der gegebenen Stichprobengröße n mit einer Power von $1-\beta$ und einem Signifikanzniveau α detektiert werden? Man bezeichnet diese Effektstärke als **minimale detektierbare Effektstärke (MDES)**.

Wir können hierzu unsere Formel

$$n = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{d}\right)^2$$

nach d auflösen:

$$d = d_{\text{MDES}} = \frac{z_{1-\alpha} + z_{1-\beta}}{\sqrt{n}}$$

Achtung: gilt für Einzelmessungen bzw. abhängige Messungen. Bei unabhängigen Messungen gilt:

$$d_{\text{MDES}}(\text{unabh. Messungen}) = \frac{z_{1-\alpha} + z_{1-\beta}}{\sqrt{\frac{n}{2}}}$$

Fallzahl gegeben: Bestimmung der minimalen detektierbaren Effektstärke

Beispiel

- Sie konnten exakt 25 Medizinstudierende innerhalb des Zeitraums Ihrer Bachelorarbeit überzeugen, an Ihrer Nasenlängenstudie teilzunehmen.
- Welche Effektstärke können Sie detektieren, wenn Sie bei einem zweiseitigen z-Test $\alpha = 0.05$ und $\beta = 0.33$ zugrundelegen?

Kritische z-Werte (aus z-Tabelle):

$$z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96 \qquad z_{1-\beta} = z_{0.67} = 0.44$$

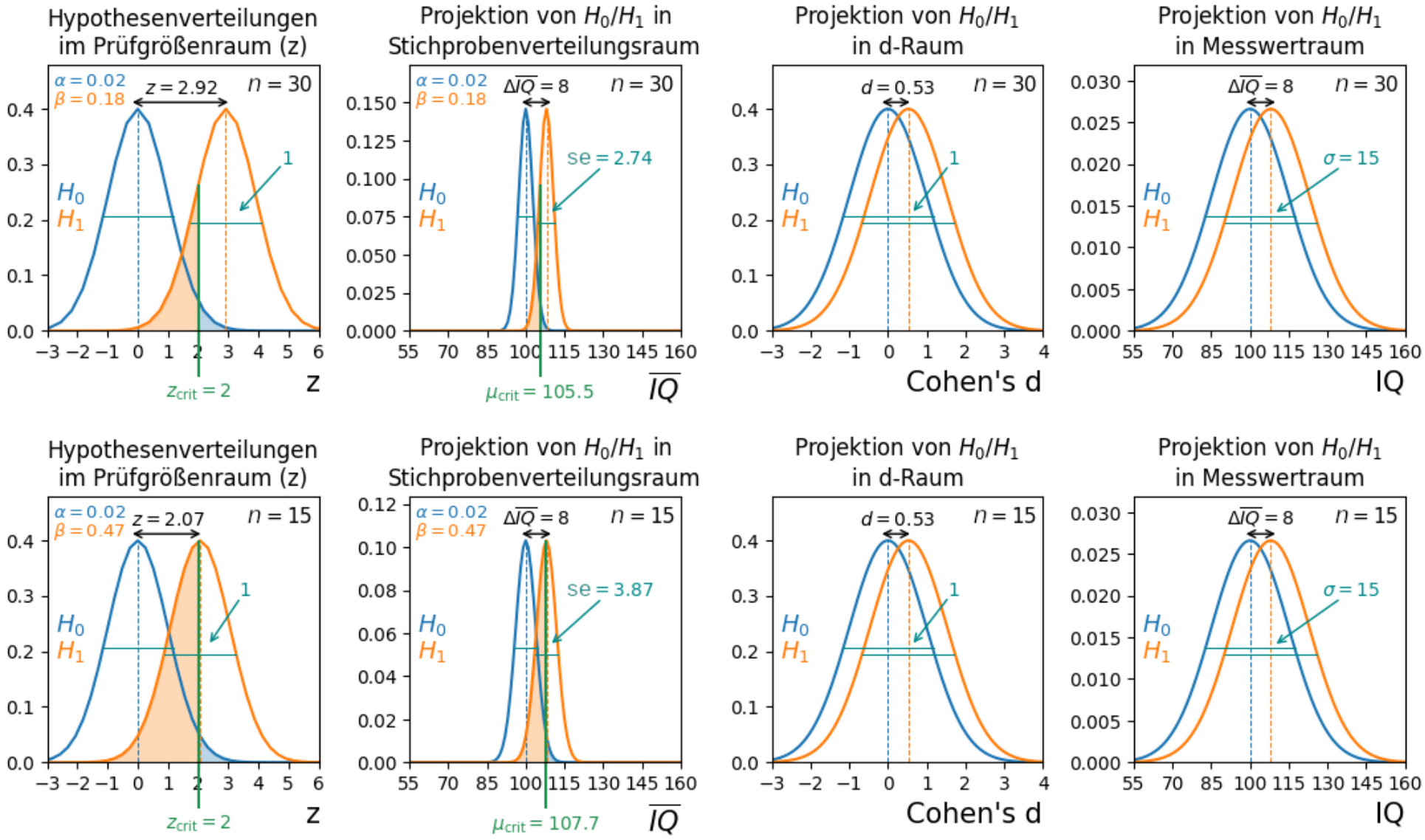
Einsetzen in die Formel für Einzelmessungen / abh. Messungen:

$$d_{\text{MDES}} = \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\sqrt{n}} = \frac{1.96 + 0.44}{\sqrt{25}} = \frac{2.4}{5} = 0.48$$

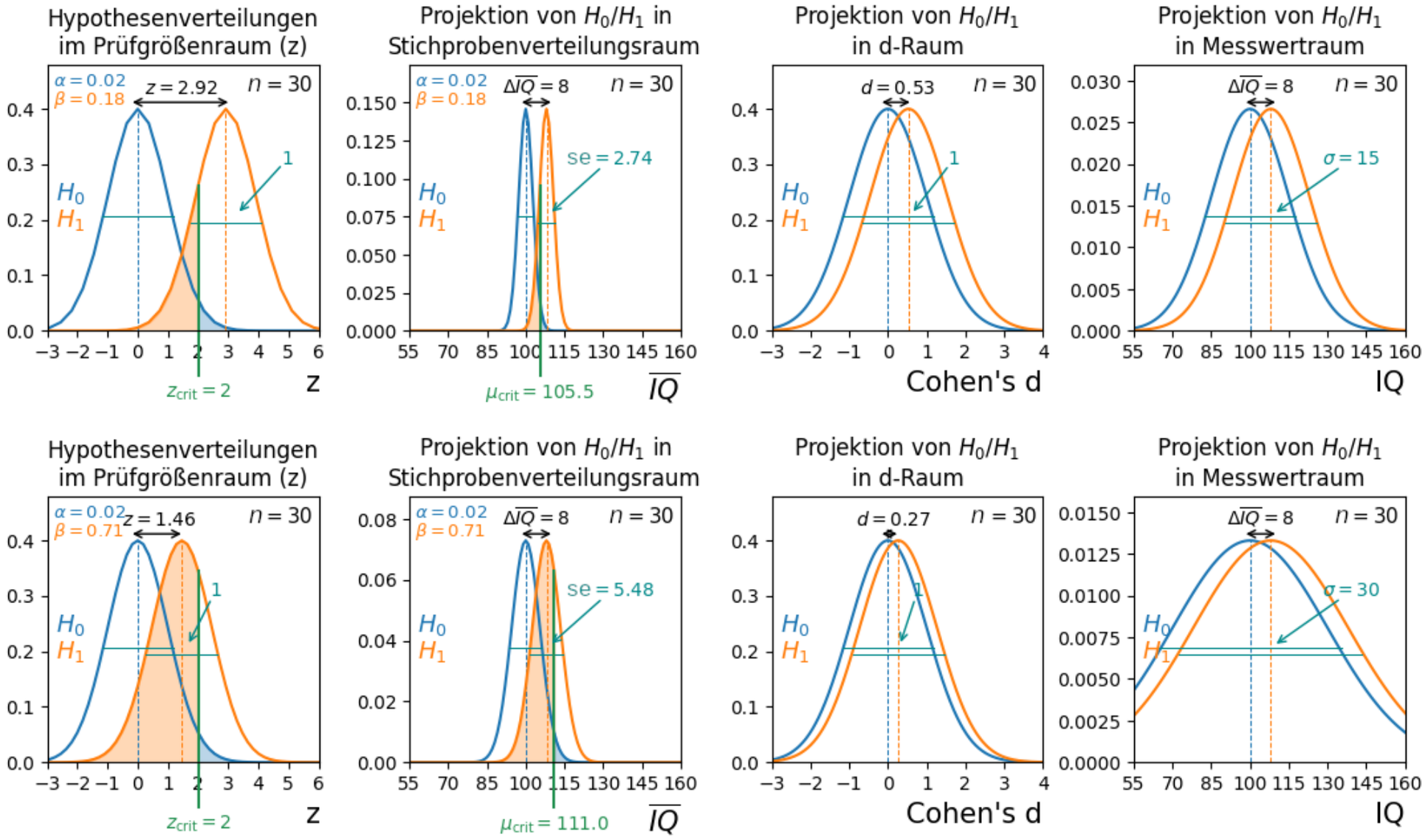
Ihre minimale detektierbare Effektstärke beträgt also $d_{\text{MDES}} = 0.48$ — ein kleiner bis mittlerer Effekt!

Bonuscontent

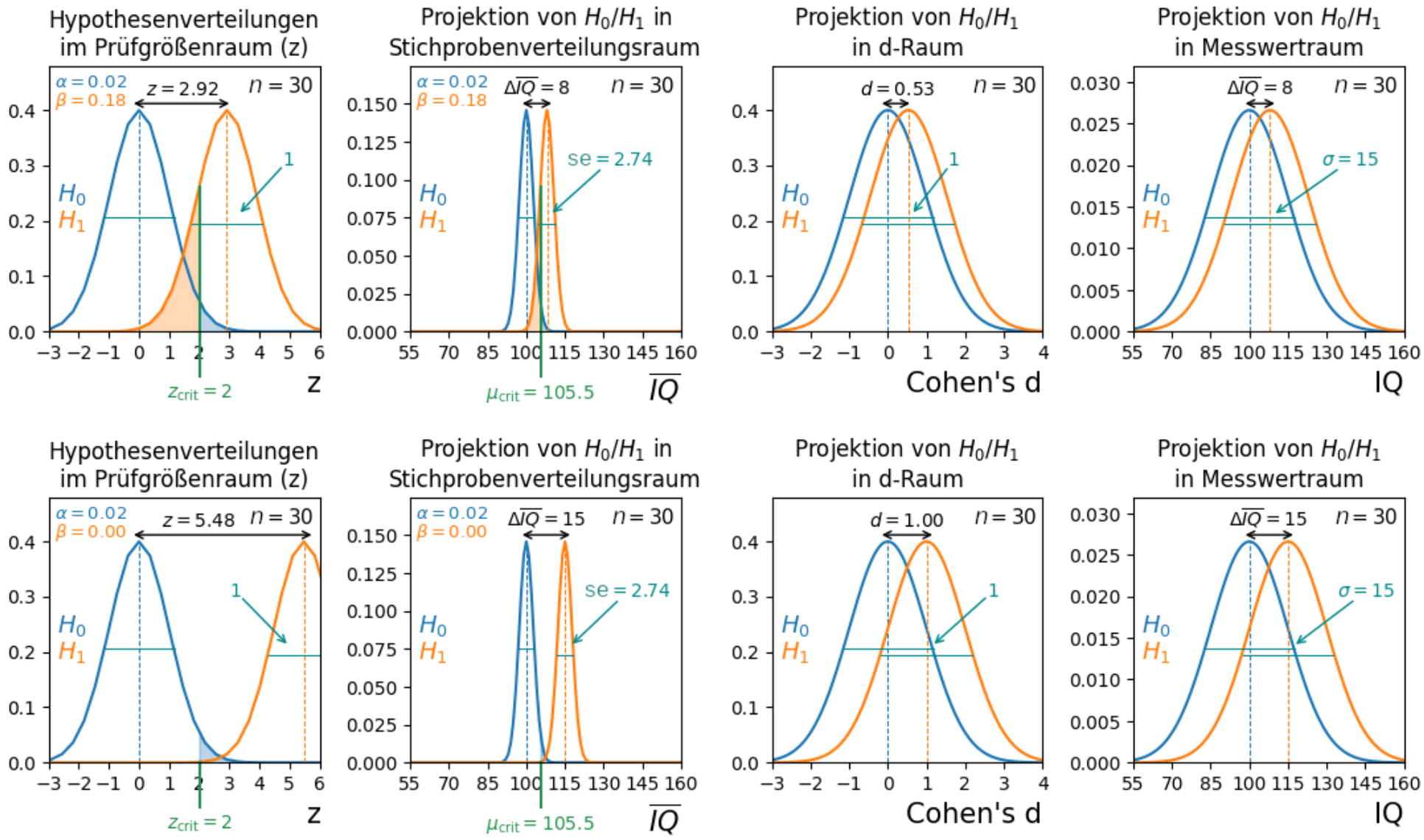
Veränderung der Stichprobengröße n



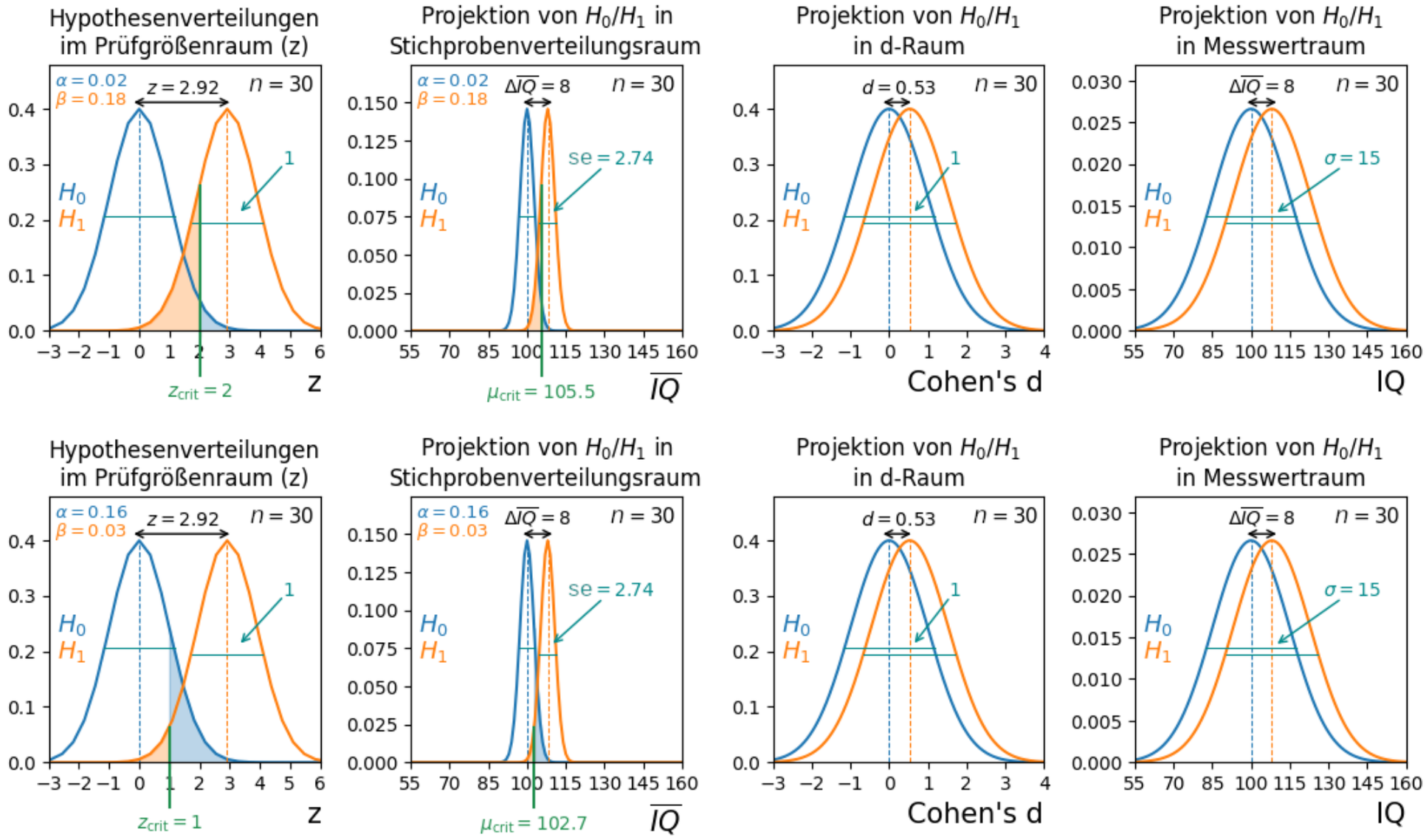
Veränderung der Populationsstreuung σ



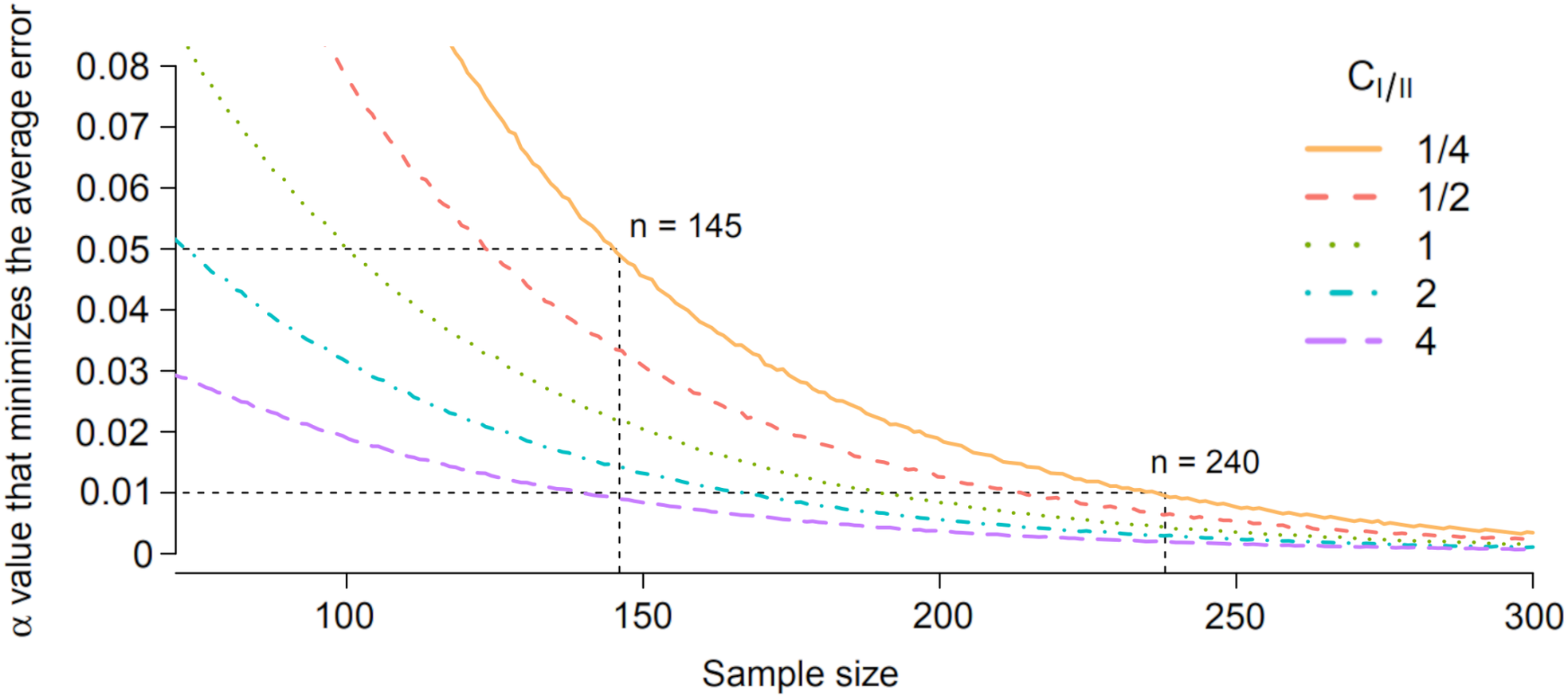
Veränderung des Mittelwertsunterschiedes $\Delta\overline{IQ}$



Veränderung des kritischen Entscheidungswertes z_{crit}

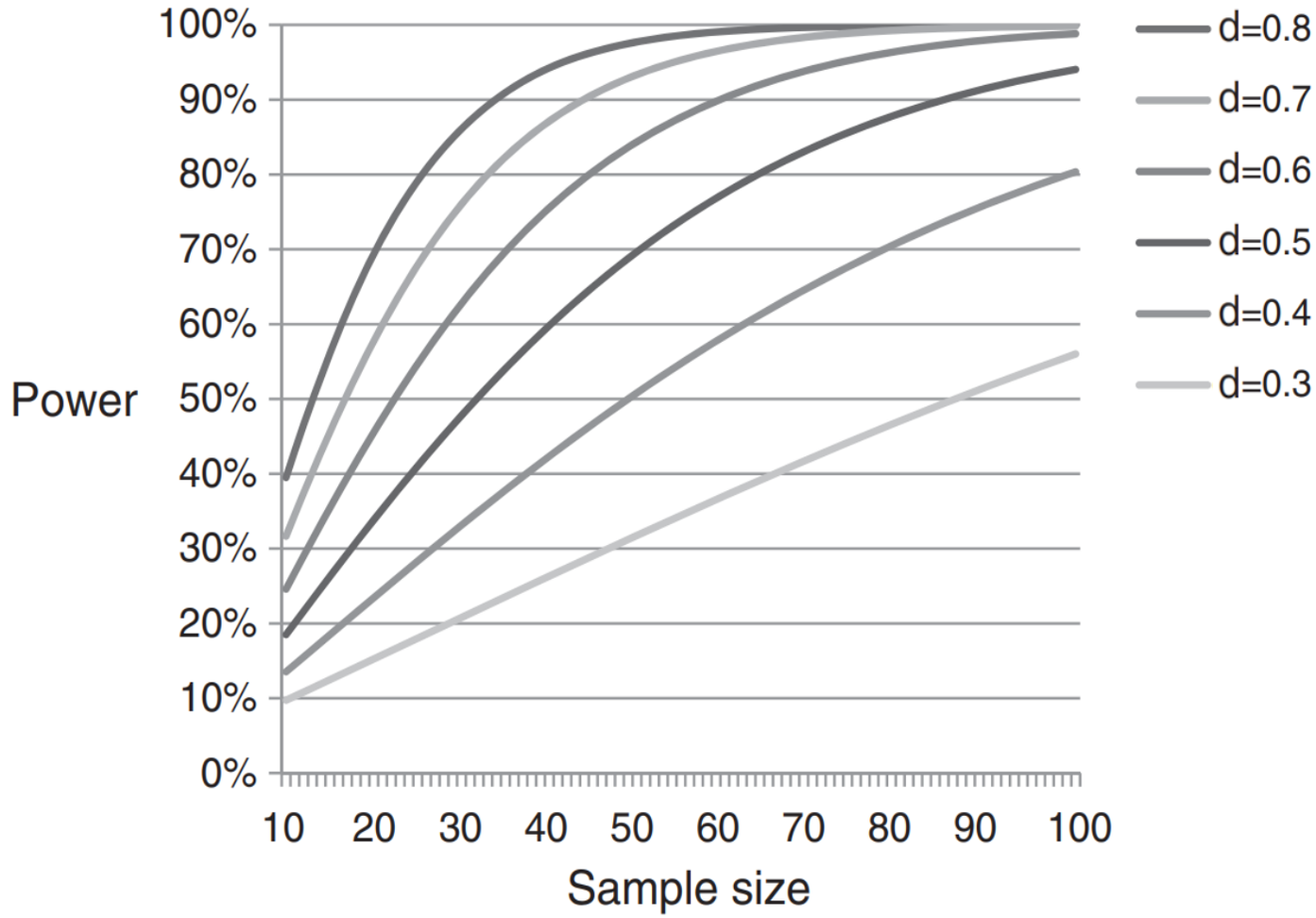


Optimaler α-Fehler



Die Abbildung zeigt den optimalen α-Wert für verschiedene Stichprobengrößen und verschiedene Verhältnisse von Fehler 1. und 2. Art (in der Legende bedeutet z.B. $C_{I/II} = 1/4$, dass die Fehlerrate β zweiter Art 4x so klein sein soll, wie die Fehlerrate α erster Art). Die "Optimalität" des α-Wertes bezieht sich darauf, dass die Summe der Fehlerraten minimal ist (unter Berücksichtigung des Verhältnisses). Aus der Abbildung ist ersichtlich, dass die Konvention $\alpha = 0{,}05$ nicht aus der Luft gegriffen ist, sondern bei typischen Stichprobengrößen in der Nähe des optimalen Wertes liegt.[2]

Zusammenhang Stichprobengröße und Power



Power ist eine konkave Funktion der Stichprobengröße, d.h. jedes 1% Inkrement an Power benötigt ein zunehmendes Inkrement in der Stichprobengröße. Die Abbildung gilt für einen zweiseitigen Zweistichproben-t-Test mit Signifikanzniveau $\alpha = 0.05$. Adaptiert von Lakens 2014[3].

Fußnoten

1. https://de.wikipedia.org/wiki/Der_Hirtenjunge_und_der_Wolf
2. Wulff J, Taylor L (2023) How and Why Alpha Should Depend on Sample Size: A Bayesian-frequentist Compromise. Academy of Management Annual Meeting Proceedings 2023:12131.
3. Lakens D (2014) Performing high-powered studies efficiently with sequential analyses. Euro J Social Psych 44:701–710.