M24 Statistik 1: Sommersemester 2024

Vorlesung 14: Power und Fallzahlschätzung

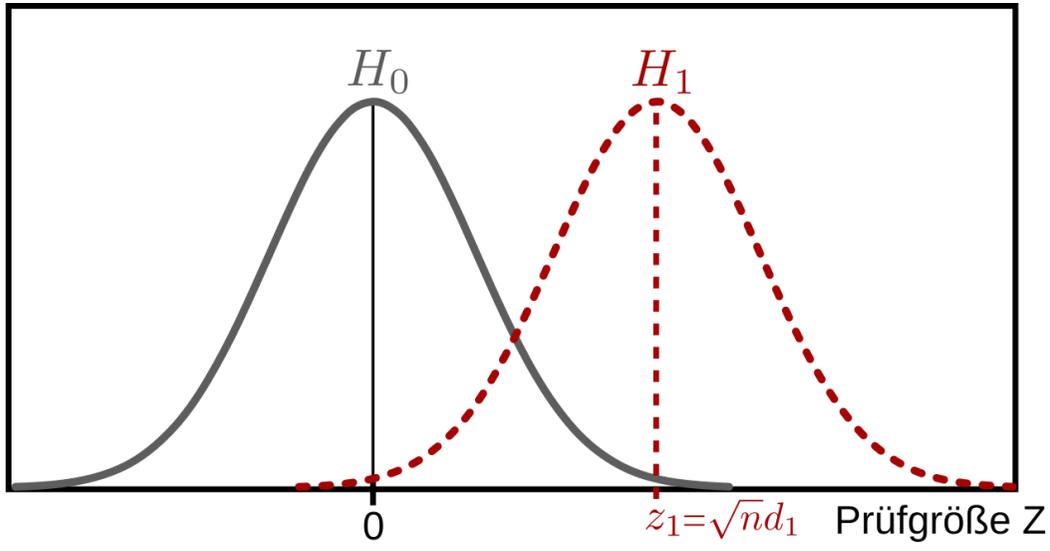
Prof. Matthias Guggenmos

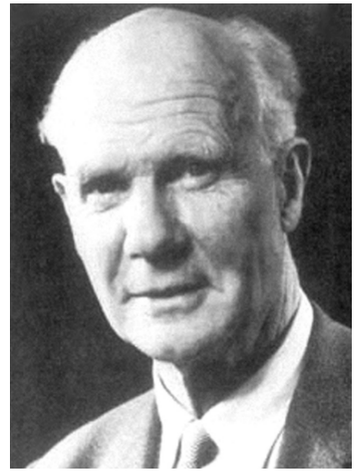
Health and Medical University Potsdam

Binäres Entscheidungskonzept von Neyman und Pearson

Idee

- **Alternatives Framework zur Hypothesentestung** von Jerzy Neyman und Egon Pearson
- Argument: man ist i.d.R. nicht spezifisch an der Nullhypothese interessiert, sondern möchte einen **Test, der zwischen der Nullhypothese H_0 und der Alternativhypothese H_1 „entscheidet“**.
- Im einfachsten Fall legt man dafür eine **minimale Effektstärke d_1** fest, bei der die Alternativhypothese H_1 noch praktische oder konzeptionelle Relevanz hätte.



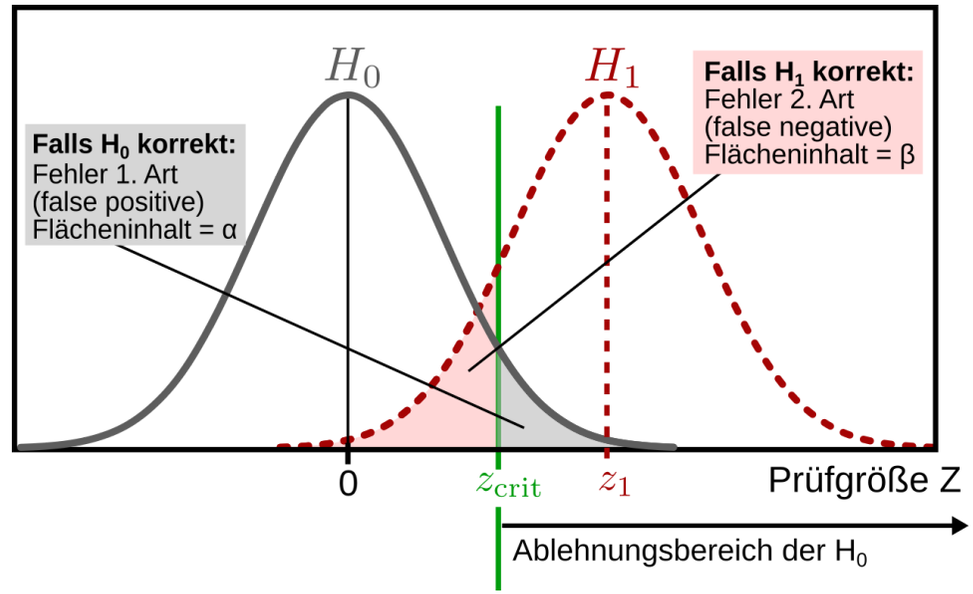


Egon Sharpe Pearson (1895-1980), Sohn von Karl Pearson

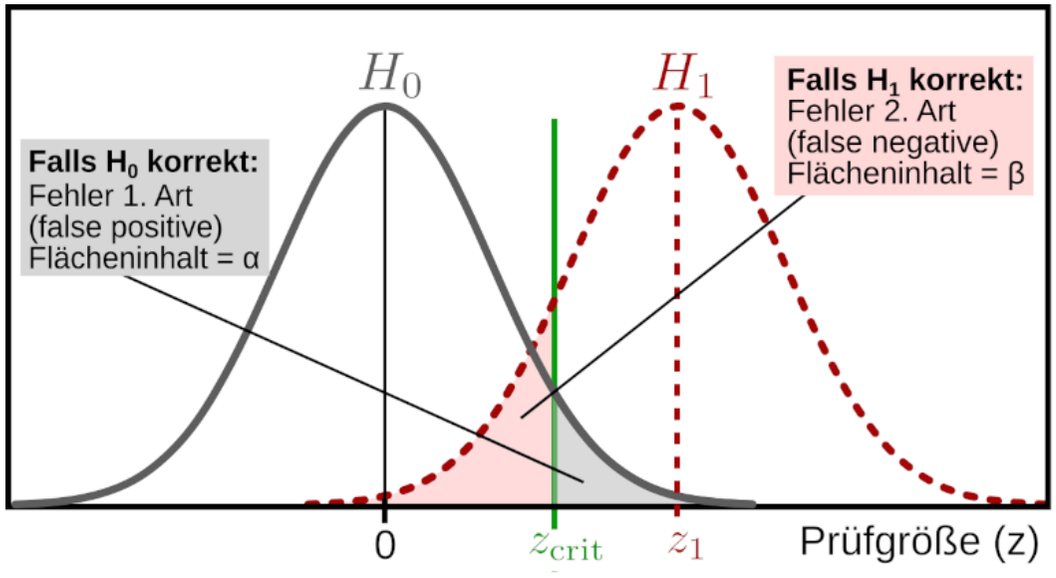
Jerzy Neyman (1894-1981)

Idee

- Durch das Aufstellen der Alternativhypothese gibt es nun zwei Möglichkeiten sich zu irren:
  - Die Nullhypothese ist wahr und man entscheidet sich irrtümlich für die Alternativhypothese (Fehler “erster Art” oder α-Fehler).
  - Die Alternativhypothese ist wahr und man entscheidet sich irrtümlich für die Nullhypothese (Fehler “zweiter Art” oder β-Fehler).
- Nehmen wir an, wir legen zur Entscheidung zwischen H_0 und H_1 einen kritischen Entscheidungswert z_{crit} fest, d.h.
  - Entscheidung für H_0 und gegen H_1 wenn $z \leq z_{\text{crit}}$
  - Entscheidung gegen H_0 und für H_1 wenn $z > z_{\text{crit}}$
- ..dann können beide Fehler als Flächen unter den Hypothesenverteilungen eingetragen werden.
- α und β sind also Flächeninhalte!



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*

Fisher versus Neyman/Pearson

- Nullhypothesenframework nach Fisher:
  - Betrachtet wird nur eine Nullhypothese
  - Fokus: p-Wert
  - p stellt ein **kontinuierliche** Evidenzmaß dar (je kleiner, desto mehr Evidenz)
  - Signifikanztests nach Fisher sind **Ablehnungstests** (Nullhypothese wird abgelehnt oder nicht, nie angenommen)
- Framework nach Neyman und Pearson:
  - Betrachtet wird sowohl eine Null- als auch eine Alternativhypothese
  - Fokus: liegt die Prüfgröße rechts oder links von der kritischen Prüfgröße z_{crit}?
  - Binäres Entscheidungskonzept (Entscheidung für H_0 oder H_1); zwar wird auch ein z-Wert (und assoziierter p-Wert) berechnet, diese dienen aber nur zur Entscheidungsfindung und nicht als kontinuierliches Evidenzmaß.
  - Signifikanztests nach Neyman/Pearson sind **Akzeptanztests** (wir entscheiden uns *für* H_0 oder *für* H_1)

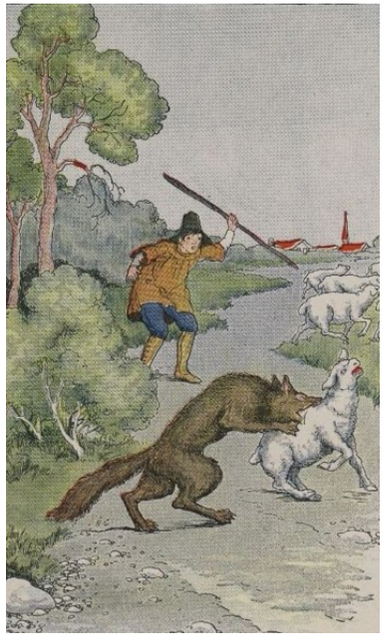
Beachte: **auch bei Neyman & Pearson wird nur die Nullhypothese getestet** und wie bei Fisher die Wahrscheinlichkeit (p-Wert) überprüft, dass der Effekt (oder ein noch extremerer Effekt) unter der Nullhypothese H_0 entstanden ist. Die **Alternativhypothese spielt bei der Testprozedur selbst keine Rolle**. Sie kommt an zwei Stellen ins Spiel: bei der Power-Berechnung *vor Beginn der Studie* (s. nächste Folien) und bei der wissenschaftlichen Entscheidung *nach der Testprozedur* (Entscheidung für H_1 falls $z > z_{\text{crit}}$).

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.

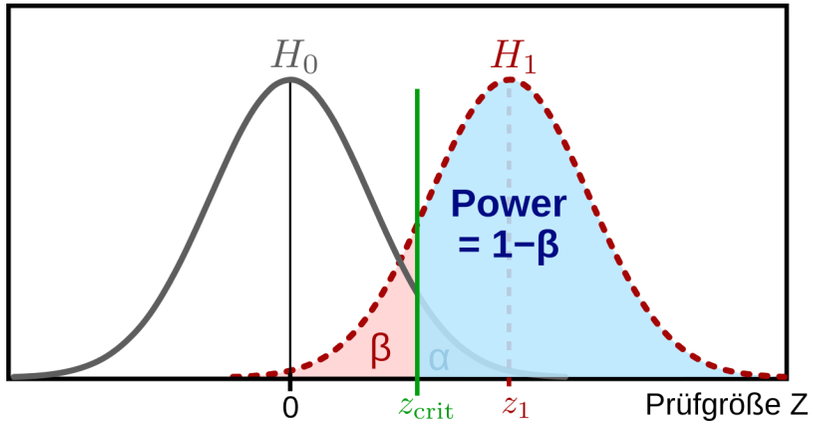
Quelle: Wikipedia[1]



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

Relevanz für Berechnung der Stichprobengröße

- Die häufigste Anwendung findet das Neyman-Pearson-Framework bei der Planung der Stichprobengröße.
- Der Vorteil des Frameworks ist, dass es nicht wie bei Fisher nur die Irrtumswahrscheinlichkeit α für die H_0 berücksichtigt ("Wahrscheinlichkeit das Ergebnis als signifikant zu werten obwohl H_0 gilt"), sondern auch die Irrtumswahrscheinlichkeit β für die H_1 ("Wahrscheinlichkeit das Ergebnis als nicht signifikant zu werten obwohl H_1 gilt").
- Beide Fehlerarten sollten gegeneinander abgewogen werden (welcher Fehler ist wichtiger/fataler?)
  - Konvention ist es, die Hypothese als H_0 festzulegen, die mit geringerer Wahrscheinlichkeit fälschlich abgelehnt werden soll (das impliziert $\alpha < \beta$).
- Im Kontext der Stichprobenberechnung wird statt der Irrtumswahrscheinlichkeit β häufig $1 - \beta$ betrachtet.
  - $1 - \beta$ entspricht der Wahrscheinlichkeit die (wahre) Alternativhypothese zu bestätigen, d.h. einen vorhandenen Effekt auch tatsächlich zu finden.
  - Diese Wahrscheinlichkeit wird als **Teststärke** oder **Power** bezeichnet und die Prozedur daher auch **Power-Berechnung**.



Eine häufige Wahl für die Teststärke/Power ist 80% (entspricht $\beta = 0.2$).

Relevanz für Berechnung der Stichprobengröße

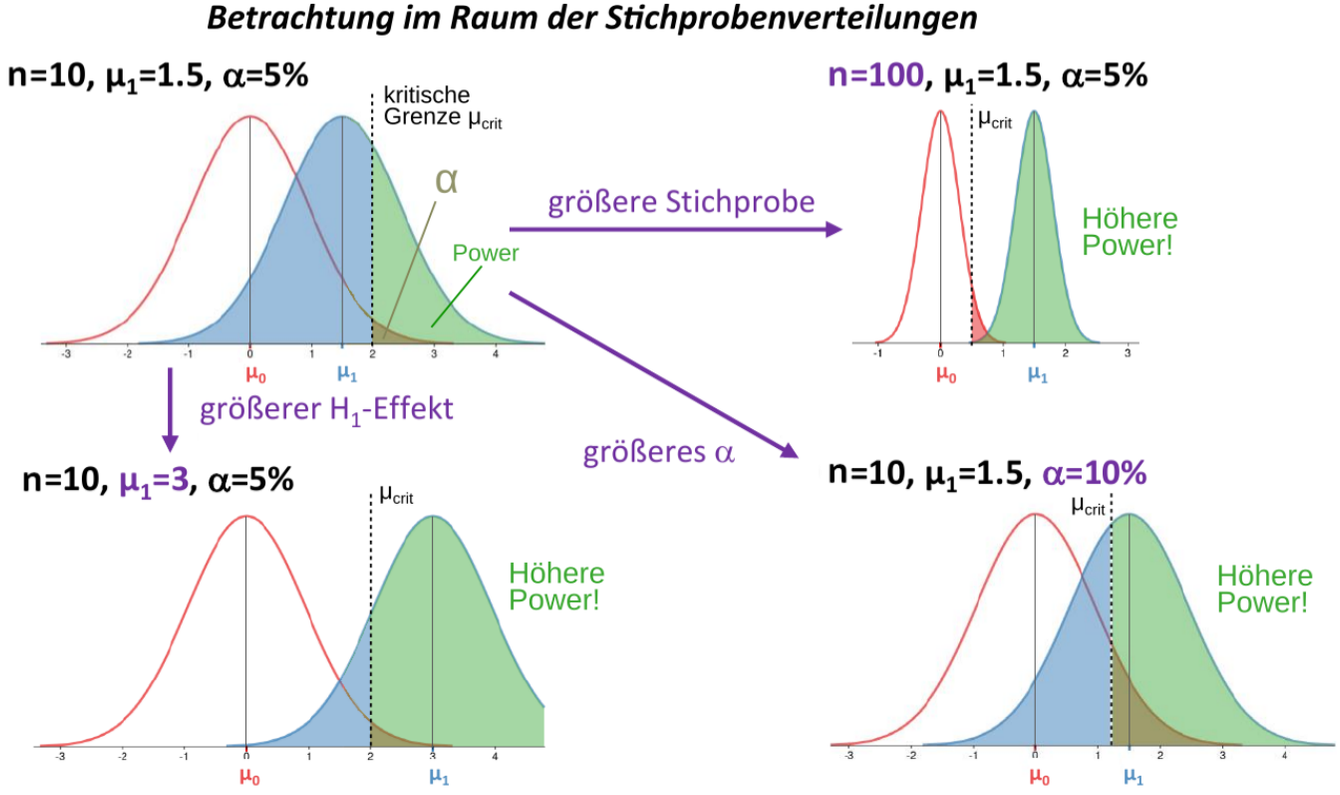
Im Kontext der Stichprobenberechnung, gibt das Neyman-Pearson-Framework eine Antwort auf folgende Frage:

Meine Studie soll einen **Fehler erster Art von höchstens** α haben. Ich erwarte, dass mein Effekt eine **Effektstärke** d_1 aufweist. Wie groß muss ich meine **Fallzahl** n wählen, damit ich mit einer **Wahrscheinlichkeit von** $1 - \beta$ **(=Teststärke/Power)** einen tatsächlich vorhandenen Effekt finden würde?

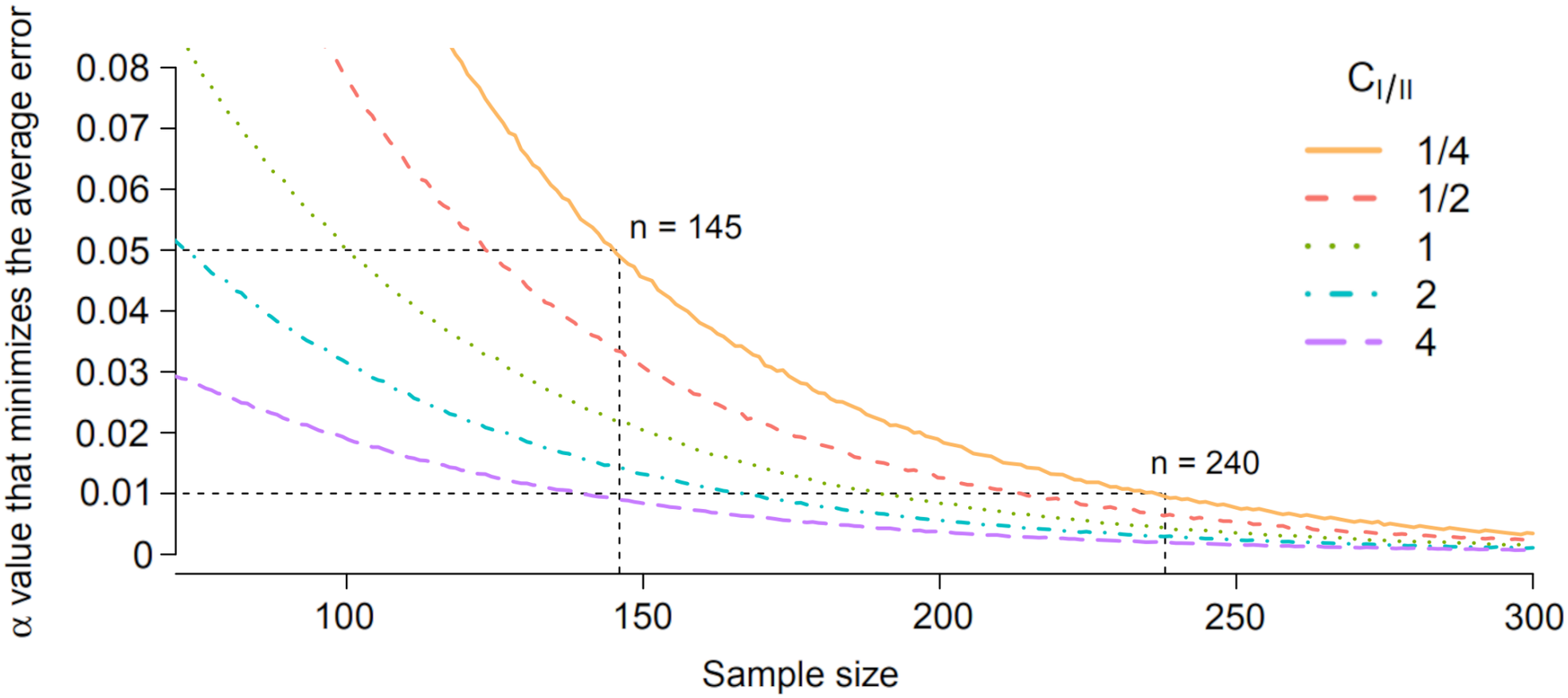
Teststärke (Power)

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

- Effektstärke von H_1 (hängt ab von μ_1 und σ)
- Stichprobengröße n
- Fehlerrate erster Art α

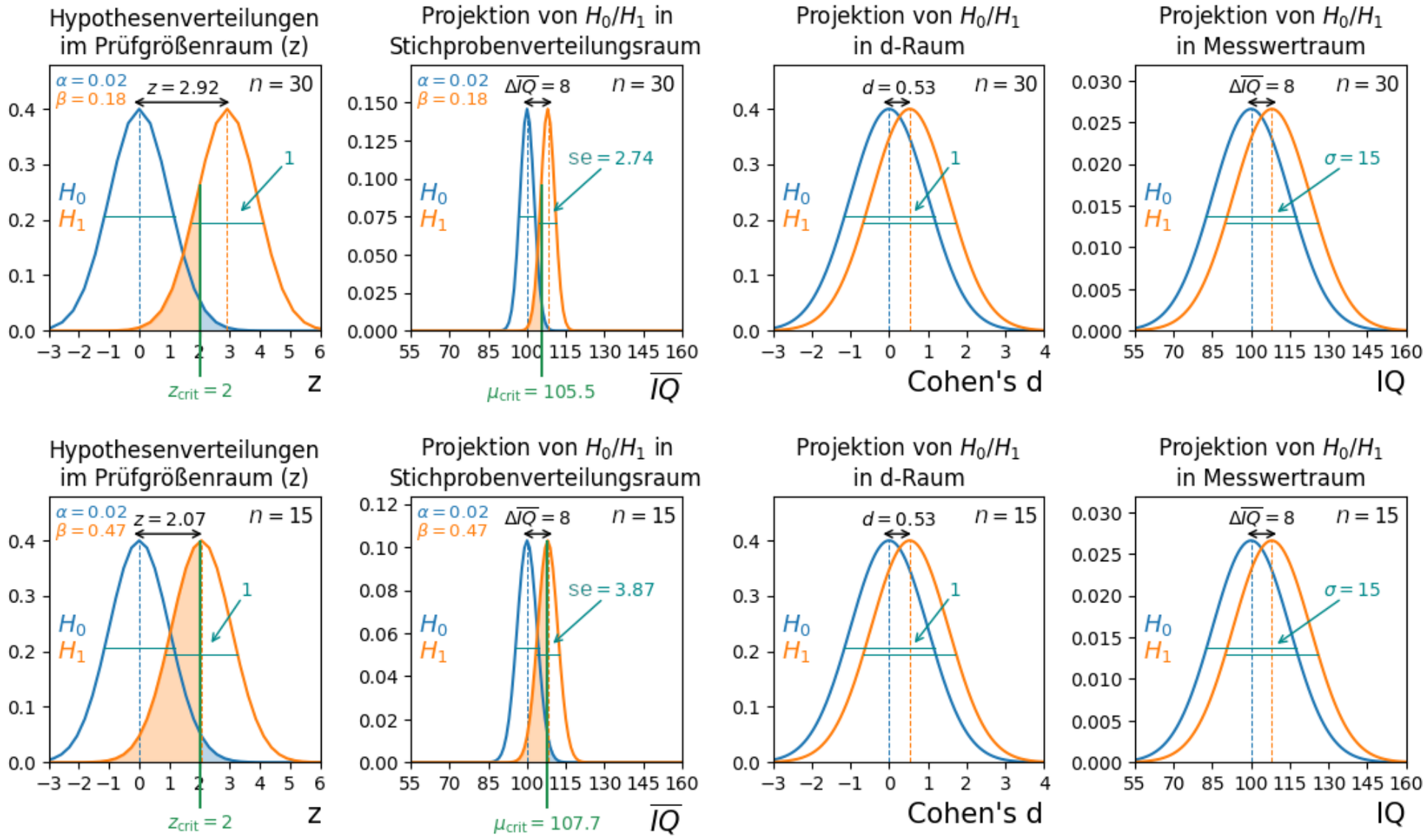


Relevanz für Berechnung der Stichprobengröße

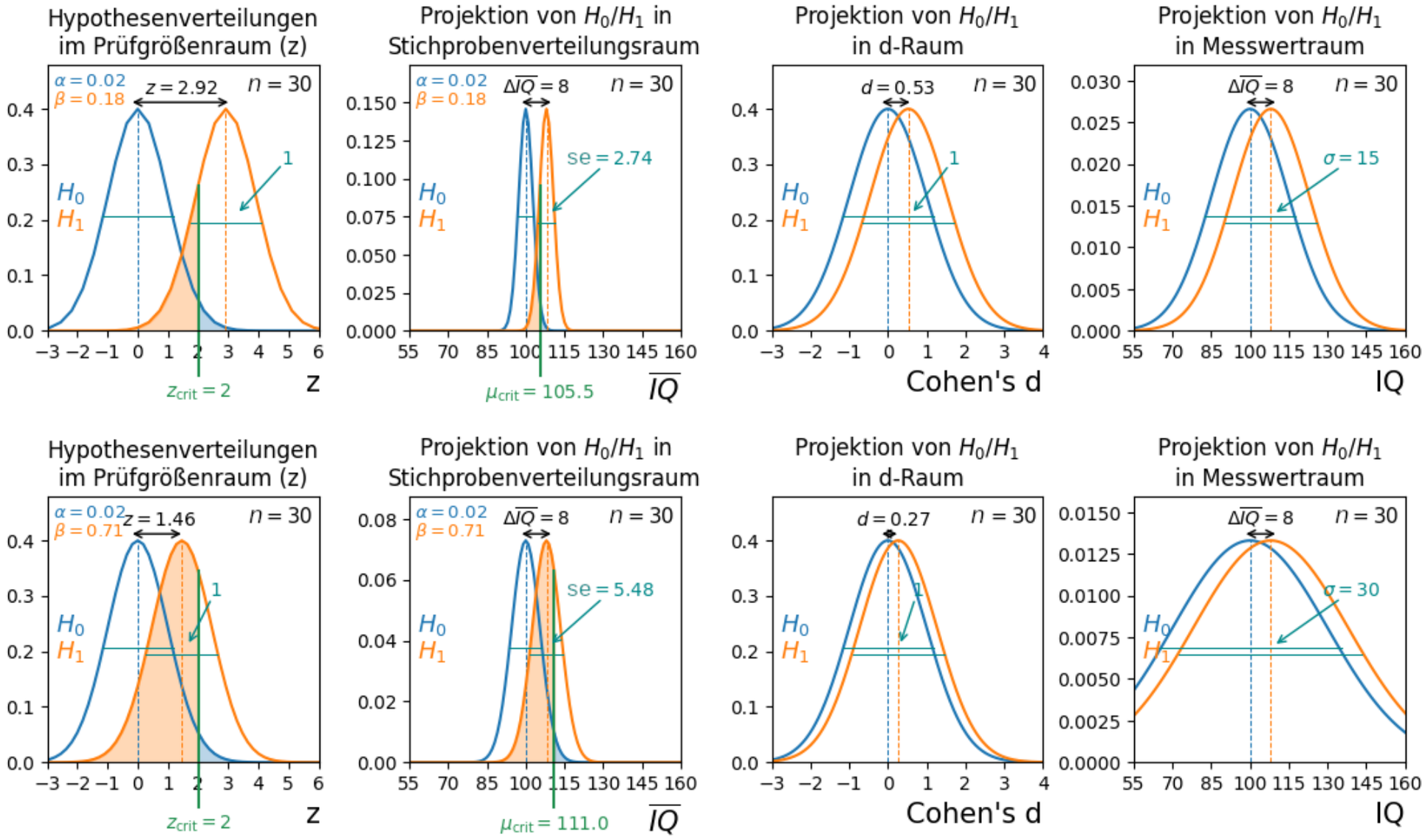


Die Abbildung zeigt den optimalen α-Wert für verschiedene Stichprobengrößen und verschiedene Verhältnisse von Fehler 1. und 2. Art (in der Legende bedeutet z.B. $C_{I/II} = 1/4$, dass die Fehlerrate β zweiter Art 4x so klein sein soll, wie die Fehlerrate α erster Art). Die "Optimalität" des α-Wertes 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]

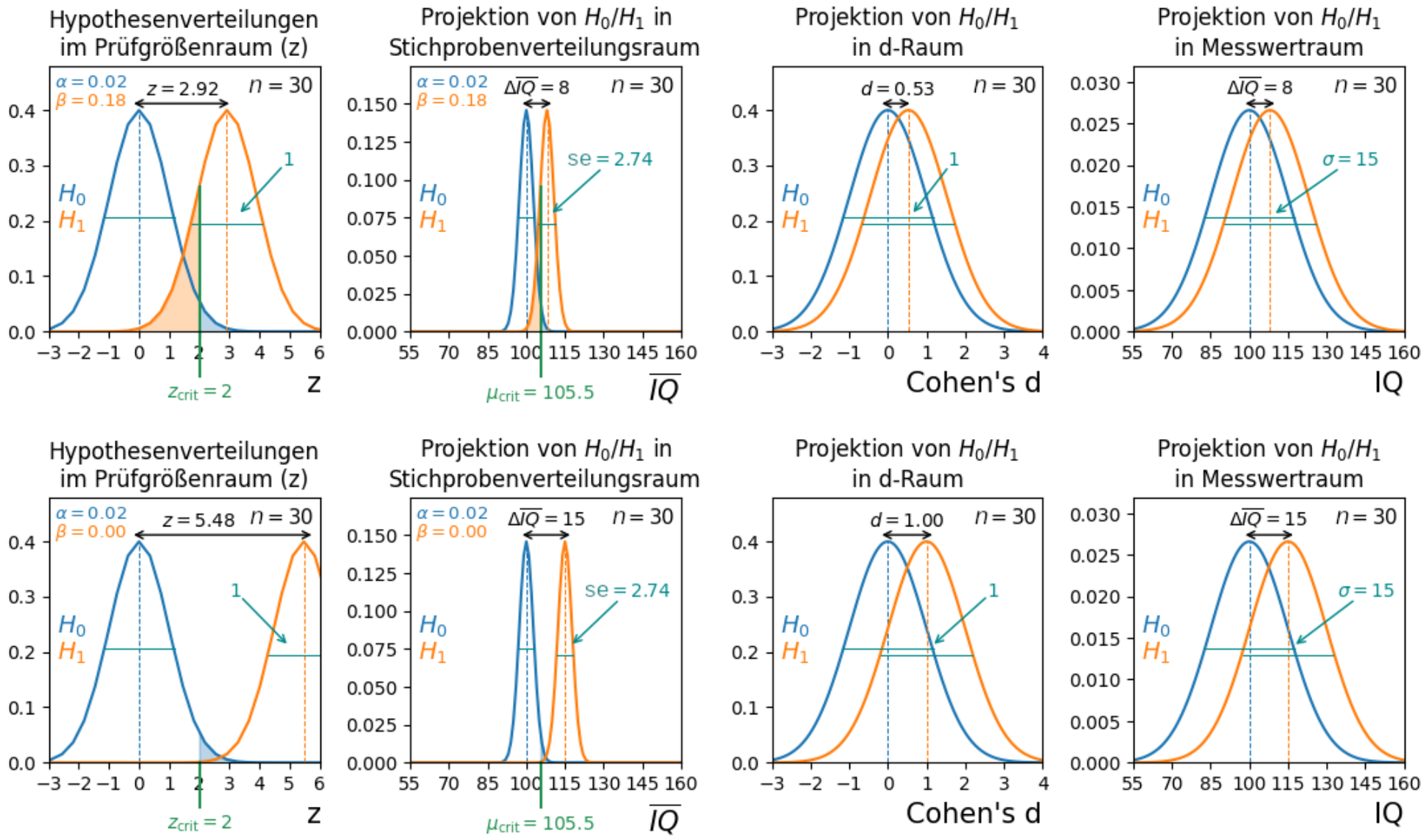
Veränderung der Stichprobengröße n



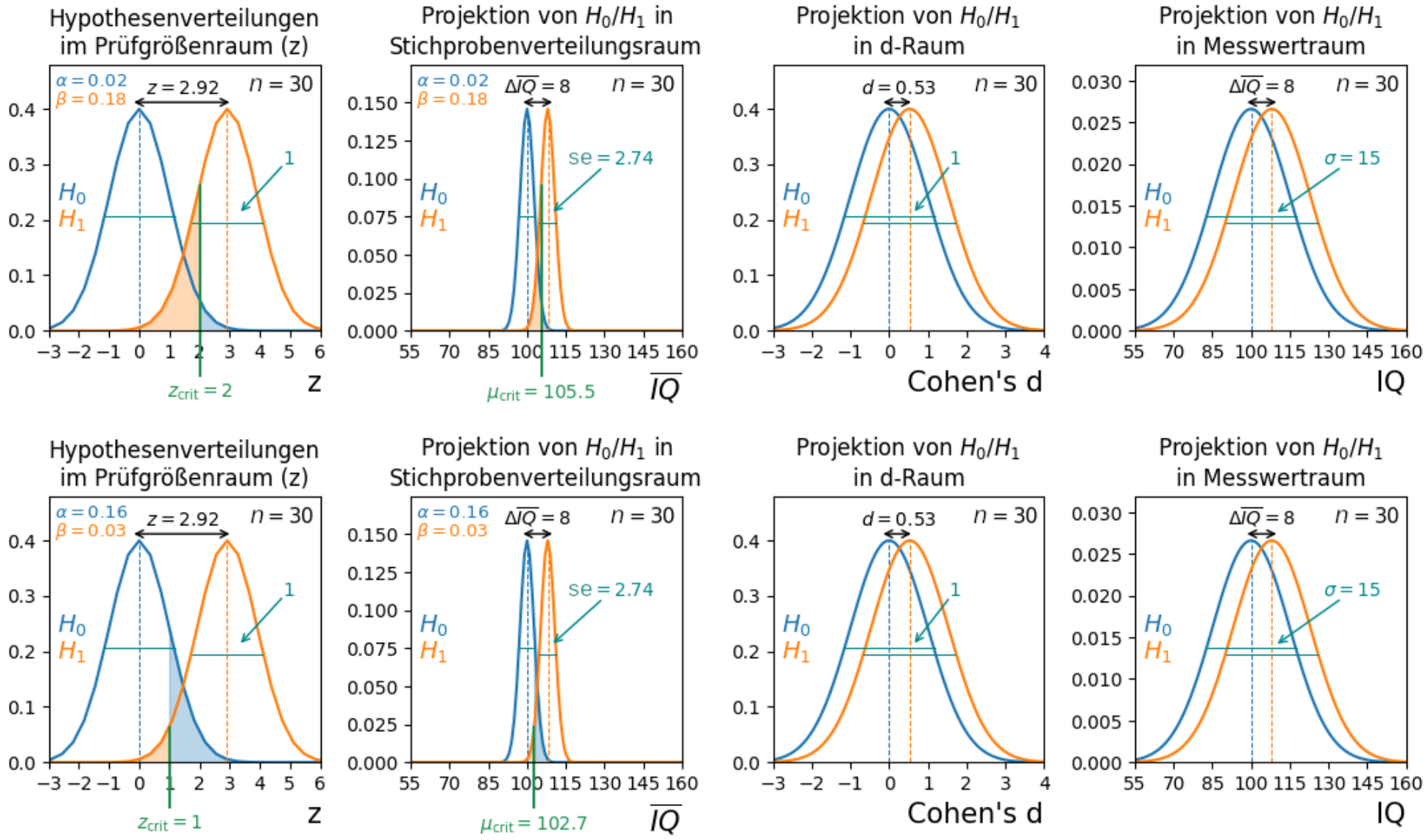
Veränderung der Populationsstreuung σ



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



Veränderung des kritischen Entscheidungswertes z_{crit}



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.