

M24 Statistik 1: Wintersemester 23/24

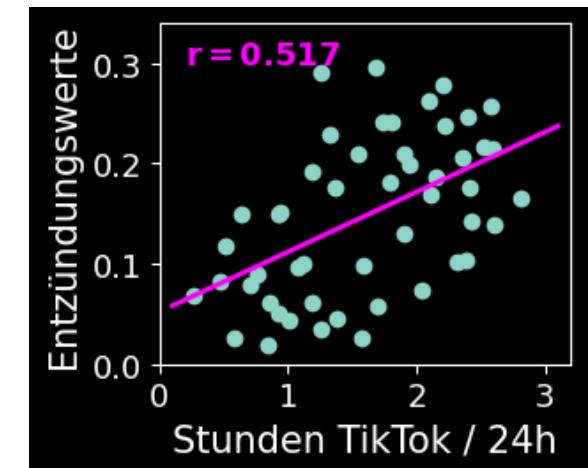
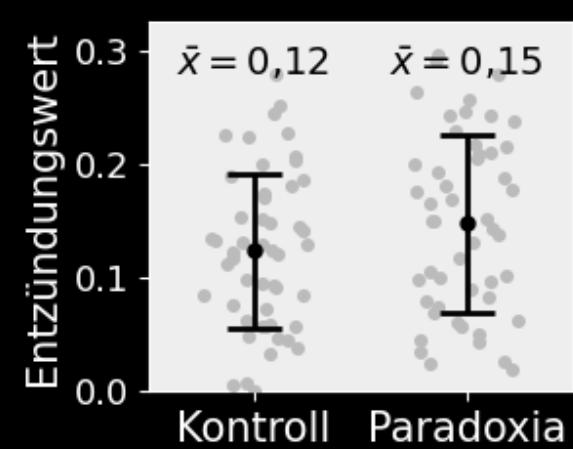
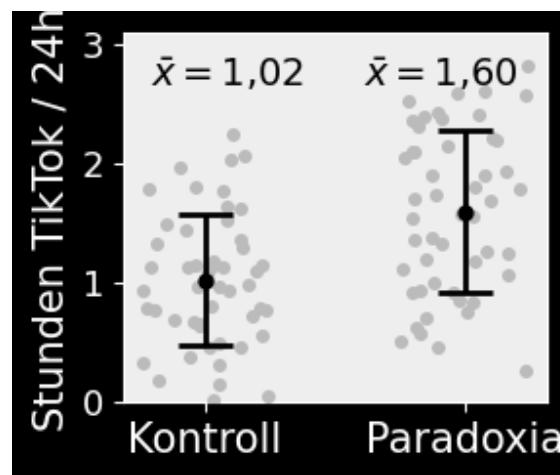
Vorlesung 10: Signifikanztestung

Prof. Matthias Guggenmos

Health and Medical University Potsdam

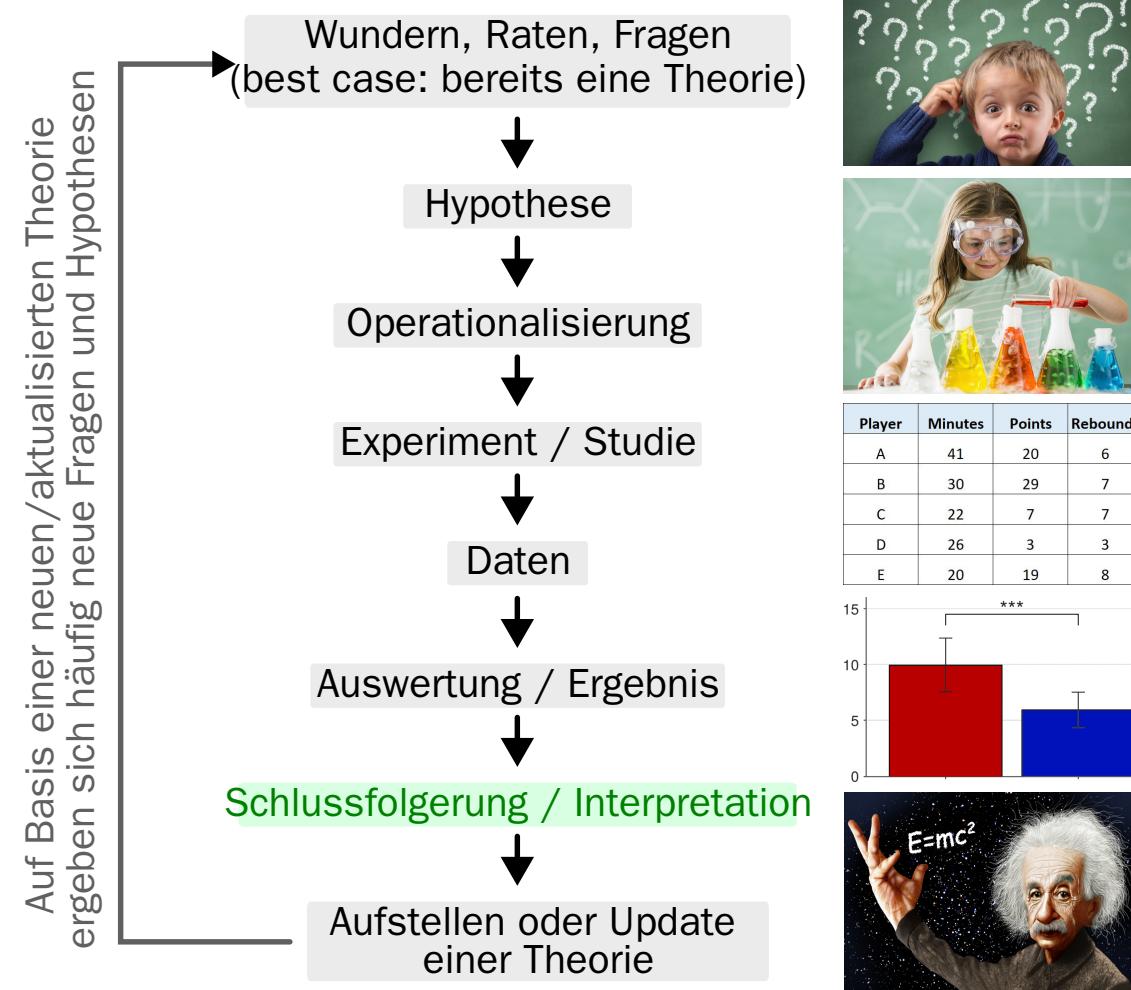


Kurze Zwischenbilanz: Sie haben Mittelwertsunterschiede zwischen Paradoxikern und Kontrollen sowohl für TikTok-Online-Zeit gefunden, als auch für Entzündungswerte (allerdings mit einer geringeren Effektstärke). Dazu gibt es einen mysteriösen Zusammenhang zwischen beiden abhängigen Variablen: mehr TikTok-Zeit = höherer Entzündungswert.



Die finale Frage lautet nun: welcher dieser Effekte ist **statistisch signifikant** und welcher Effekt ist vermutlich eher eine Zufallsbeobachtung?

Der Forschungsprozess



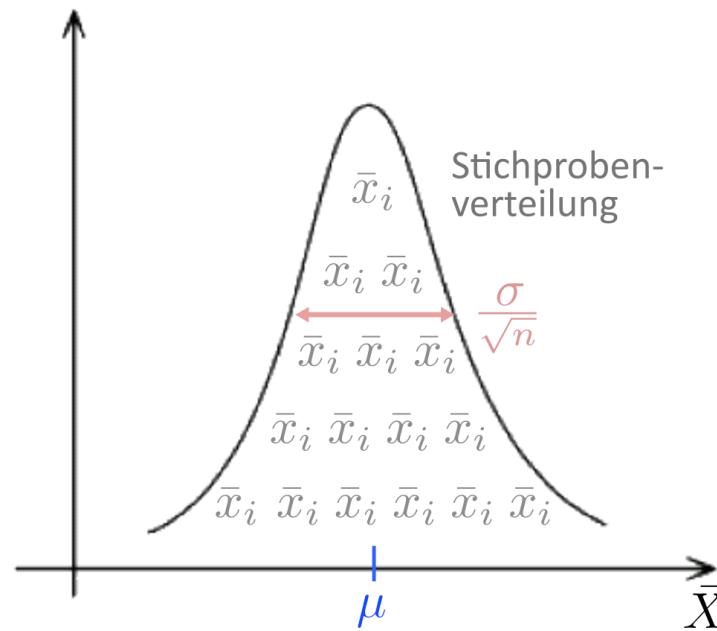
Player	Minutes	Points	Rebounds
A	41	20	6
B	30	29	7
C	22	7	7
D	26	3	3
E	20	19	8



Theoretische Stichprobenverteilung

Zurück zur **theoretischen Stichprobenverteilung**. Alle Erkenntnisse über die Normalverteilung lassen sich auch auf die theoretische Stichprobenverteilung übertragen. Dies wird uns bis zum Ende des Semesters zwei wesentliche Methoden der Inferenzstatistik eröffnen:

- **Hypothesentestung bzw. Signifikanztestung** (u.a. auch Idee des p-Wertes)
- **Konfidenzintervalle** (Verallgemeinerung des Standardfehlers)



Signifikanztestung

Ziel der **Signifikanztestung** ist zu klären, ob ein Effekt wahrscheinlich durch Zufall erklärt werden kann (und daher statistisch unbedeutsam ist) oder Zufall als alleinige Ursache unwahrscheinlich ist (und der Effekt daher statistisch bedeutsam — signifikant — ist).

Signifikanztests bieten ein Kriterium, um wissenschaftliche **Hypothesen** zu überprüfen. Dabei unterscheidet man zwischen zwei Arten von Hypothesen 

Nullhypothese und Alternativhypothese

Nullhypothese (H_0)

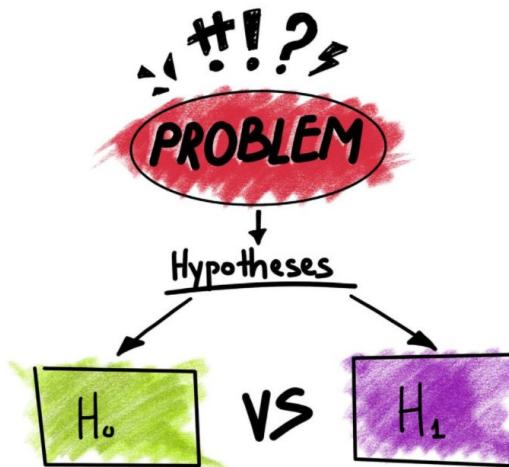
Besagt, dass ein bestimmter Effekt *nicht vorliegt*.

- Die Häufigkeit des Haarschneidens und Haardichte **hängen nicht** zusammen.
- Männer und Frauen **unterscheiden sich nicht** in ihrer emotionalen Intelligenz.

Alternativhypothese (H_1)

Besagt, dass ein bestimmter Effekt *vorliegt*.

- Die Häufigkeit des Haarschneidens und Haardichte **hängen** zusammen.
- Männer und Frauen **unterscheiden sich** in ihrer emotionalen Intelligenz.

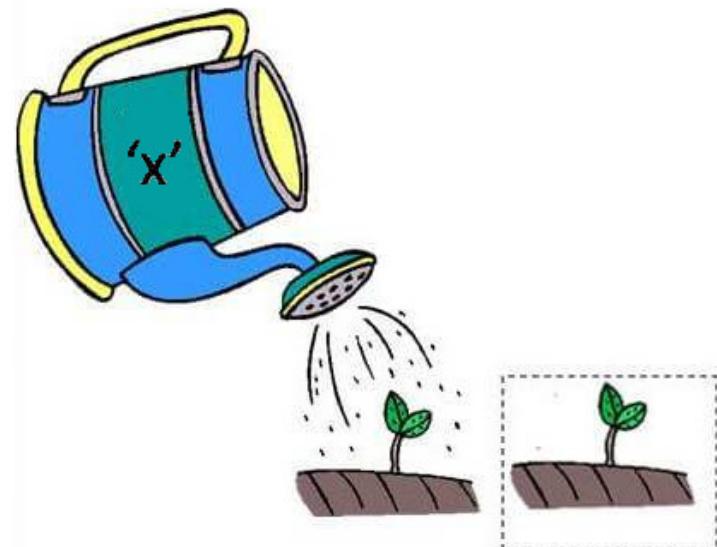


Nullhypothese und Alternativhypothese: Beispiel

Effect of Bio-fertilizer 'x' on Plant growth

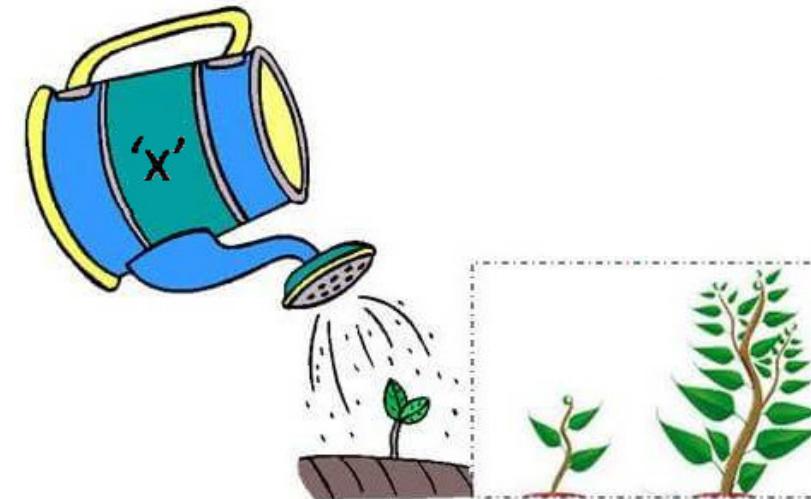
Null Hypothesis

H_0 : Application of bio-fertilizer 'x'
does not increase plant growth.



Alternative Hypothesis

H_1 : Application of bio-fertilizer 'x'
increases plant growth.



Bildnachweis¹

Nullhypotesentestung nach Fisher

- Betrachtet wird ausschließlich die Nullhypothese.
- Ziel ist es zu zeigen, dass das **Stichproben-Ergebnis unter der Annahme dieser Nullhypothese so unwahrscheinlich ist, dass man die Nullhypothese „verwerfen“ kann.**
- Bei der Nullhypotesentestung wird kein Schluss in Bezug auf eine Alternativhypothese gezogen (sie der muss hier nicht einmal formuliert werden!).



Ronald Aylmer Fisher (1890-1962)



Merke: bei Signifikanztests werden keine Wahrscheinlichkeiten für Hypothesen berechnet (weder H_0 noch H_1), sondern nur die Wahrscheinlichkeit von Daten unter einer Hypothese (bei Fisher: *unter der Nullhypothese*).

Die Wahrscheinlichkeit von Daten, gegeben eine Hypothese, heißt auch **Likelihood**.

Wie kann diese “Wahrscheinlichkeit der Daten unter der Nullhypothese” bestimmt werden?

→ p-Wert

Nullhypotesentestung nach Fisher: p-Wert

Sie überprüfen die Hypothese, dass Medizinstudierende längere Nasen haben als Psychologiestudierende.

Die zugehörige Nullhypothese H_0 lautet also: *kein Unterschied in der (durchschnittlichen) Nasenlänge von Medizin- und Psychologiestudierenden.*

Sie führen eine Studie durch ($n_{\text{med}} = n_{\text{psych}} = n = 30$) und messen folgenden Gruppenunterschied:

$$\Delta \bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}} = 0,2\text{cm}$$

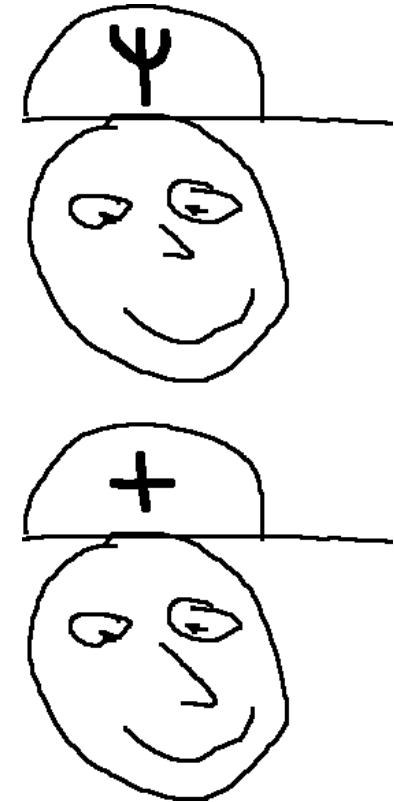
Nach Fisher stellen wir nun folgende präzise Frage:

Angenommen, die Nasenlänge unterscheidet sich nicht zwischen den Gruppen (H_0). Wie wahrscheinlich ist unter dieser Annahme das Ergebnis unserer Daten oder ein noch *extremeres Ergebnis*?

In kurz:

Angenommen H_0 trifft zu und es gilt $\Delta \bar{x} = 0$, wie wahrscheinlich ist es dennoch $\Delta \bar{x} \geq 0,2\text{cm}$ zu messen?

Diese Wahrscheinlichkeit wird als **p-Wert** bezeichnet (*p* für *probability*).

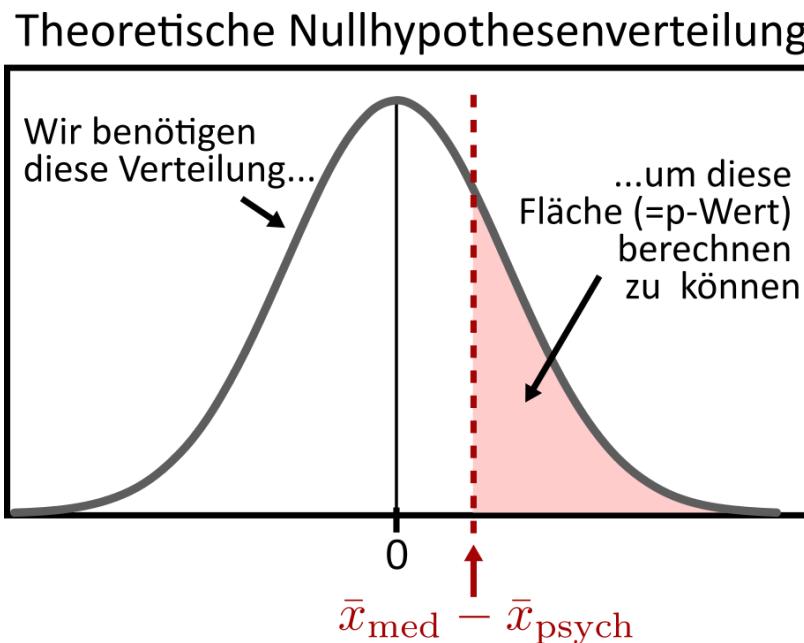


p-Wert und Konstruktion der Nullhypothese

Wir definieren:

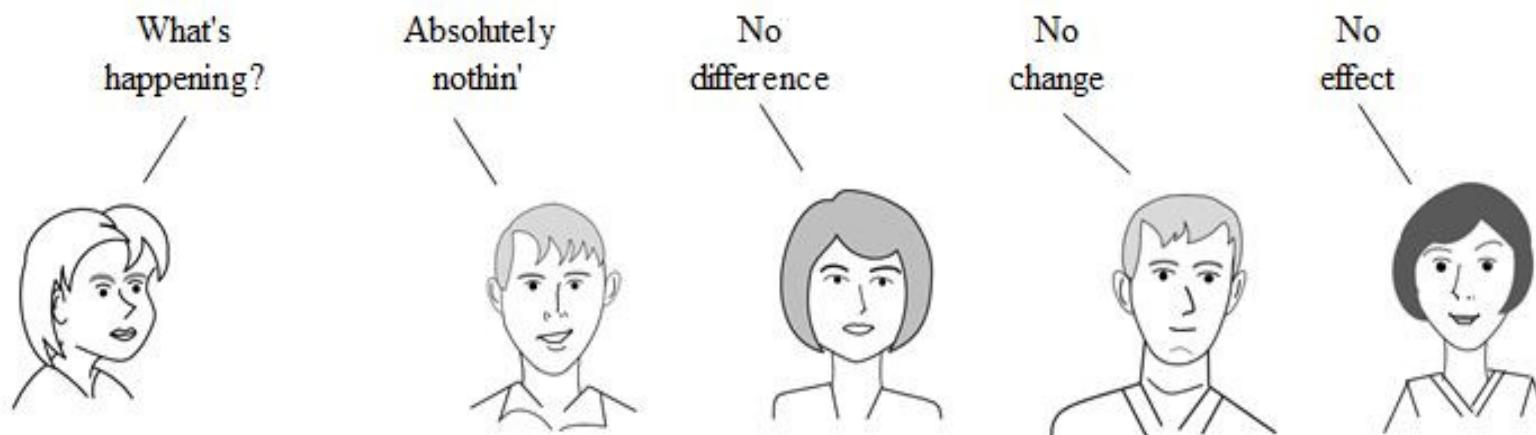
Definition Der p-Wert ist die Wahrscheinlichkeit den Stichprobeneffekt oder einen noch extremeren Effekt zu erhalten, obwohl für die Population die Nullhypothese angenommen wird.

Um diese Wahrscheinlichkeit zu berechnen, benötigen wir die Wahrscheinlichkeitsverteilung aller verschiedenen möglichen Stichprobeneffekten *unter der Annahme der Nullhypothese*. ☺



Konstruktion der Nullhypothese

- Das Ziel “*die Wahrscheinlichkeitsverteilung aller verschiedenen möglichen Stichprobeneffekten unter der Annahme der Nullhypothese*” zu finden, impliziert, dass wir auch hier eine Art von **Stichprobenverteilung** suchen.
- Da wir mithilfe der Inferenzstatistik **Effekte** testen wollen, sind hier im Besonderen Stichprobenverteilung von Effekten gemeint.
- Zu Beginn konzentrieren wir uns hier auf den Effekt des **Mittelwertsunterschiedes**.
- Wichtig ist hier ein weiteres Mal die Feststellung, dass per zentralem Grenzwertsatz nicht nur die Stichprobenverteilung des Mittelwertes per se normalverteilt ist, sondern auch die Stichprobenverteilung von *Mittelwertsunterschieden* (die Differenz zweier normalverteilter Variablen ist normalverteilt).

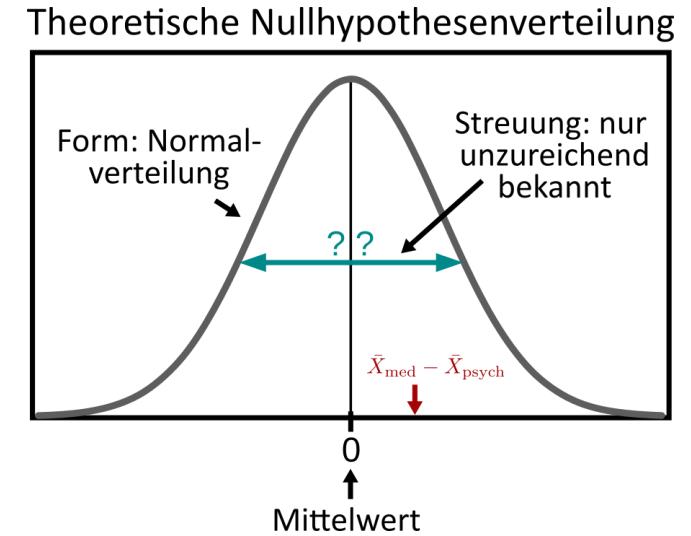


Bildnachweis²

Konstruktion der Nullhypothese

Wie schon bei der Konstruktion der Stichprobenverteilung, benötigen wir für die Konstruktion der Nullhypotesenverteilung drei Informationen: Form, Mittelwert und Streuung der Verteilung.

- Für die **Form** können wir uns wieder auf den zentralen Grenzwertsatz beziehen und eine **Normalverteilung** annehmen.
- Der **Mittelwert** ist auch klar, für die Nullhypotesenverteilung muss er **Null** sein (bzw. das, was als Nulleffekt definiert wird).
- Bei der **Streuung** stoßen wir allerdings auf ein Problem: wir kennen nur die **Streuung unseres Effektes auf Basis Stichprobe** und wissen, dass dies eine **unpräzise Schätzung der Streuung in der Population** ist.



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Formel für die Dichtefunktion der unstandardisierten Normalverteilung mit Mittelwert μ und Standardabweichung σ .

Vereinfachung: Populationsstreuungen σ bekannt

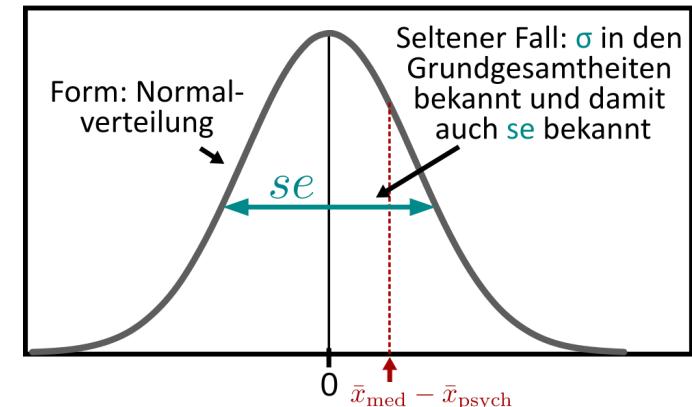
Im ersten Schritt betrachten wir daher den vereinfachten (in der Praxis sehr seltenen) Fall, dass wir die Streuung σ_{med} und σ_{psych} in den beiden Populationen der Medizin und Psychologie kennen, und damit den Standardfehler se unseres Effektes $\Delta\bar{x}$:

$$se = \sqrt{\frac{\sigma_{\text{med}}^2 + \sigma_{\text{psych}}^2}{n}}$$

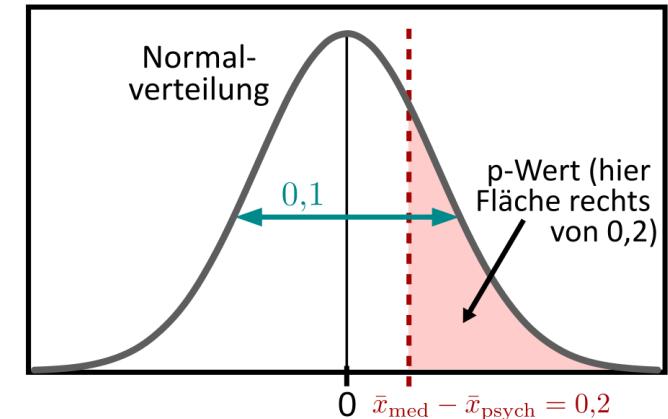
Hinweis: dies ist die Formel für den Standardfehler des Differenzeffektes $\bar{x}_{\text{med}} - \bar{x}_{\text{psych}}$. Nehmen wir an, der Standardfehler ergibt sich nach dieser Formel zu $se = 0,1$.

Nun ist alles angerichtet und wir können unseren ersten p-Wert berechnen! Dazu muss die Fläche unter der Nullhypothesenverteilung rechts von $\Delta\bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}}$ berechnet werden (Erinnerung: wir sind an der Wahrscheinlichkeit interessiert, dass die Daten unseren Effekt annehmen *oder einen noch größeren Effekt*).

Theoretische Nullhypothesenverteilung falls σ bekannt



Theoretische Nullhypothesenverteilung falls σ bekannt



Vereinfachung: Populationsstreuungen σ bekannt

Die Fläche bzw. der p-Wert berechnet sich wie folgt:

$$\begin{aligned} p &= \int_{\Delta\bar{x}}^{\infty} f(x)dx = 1 - F(\Delta\bar{x}) = \\ &= 1 - F(0,2) \stackrel{(Computer)}{=} 0,023 \end{aligned}$$

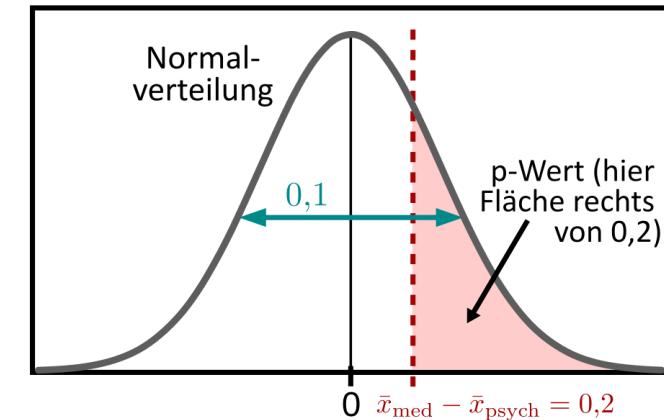
wobei $f(x)$ bzw. $F(x)$ hier die Dichte bzw. Verteilungsfunktion der gegebenen Normalverteilung $n(0, 0,1)$ ist.

$F(x)$ berechnet die Fläche bis zum Punkt x , “1 minus diese Fläche” gibt uns also die Fläche rechts vom angegebenen Punkt x – im vorliegenden Fall die Fläche rechts von $\Delta\bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}} = 0,2$.

Der p-Wert ist 0,023.

In Worten können wir feststellen: die Wahrscheinlichkeit, dass unser Effekt $\Delta\bar{x}$ durch Zufall entstanden ist – also unter der Annahme, dass die Nullhypothese gilt – beträgt (nur) 2,3%.

Theoretische Nullhypotesenverteilung falls σ bekannt



Standardfehler bei Mittelwertdifferenzen

Messungen in unabhängigen Gruppen mit unterschiedlichen Varianzen

Sind die Standardabweichungen in beiden Gruppen unterschiedlich (Faustregel: um mehr als einen Faktor 2), kann keine gepoolte Standardabweichung für beide Gruppen berechnet werden, und die Standardfehler werden für beide Gruppen separat berechnet:

$$se_A = \frac{\sigma_A}{\sqrt{n_A}} \quad se_B = \frac{\sigma_B}{\sqrt{n_B}}$$

Die Varianzen der Mittelwerte in beiden Gruppen sind gleich dem quadrierten Standardfehler der Mittelwerte:

$$se_A^2 = \frac{\sigma_A^2}{n_A} \quad se_B^2 = \frac{\sigma_B^2}{n_B}$$

Nun sind wir ja an der Differenz der Mittelwerte $\Delta\bar{x}$ interessiert. Allgemein gilt, dass sich die Varianzen von Summen oder Differenz zweier unabhängiger Zufallsvariablen addieren:

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

Daher gilt für den **Standardfehler der Mittelwertsdifferenz unabhängiger Gruppen A und B**:

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

Standardfehler bei Mittelwertdifferenzen

Messungen in unabhängigen Gruppen mit ähnlichen Varianzen

Gilt $\frac{1}{2} < \frac{\sigma_A}{\sigma_B} < 2$, die Standardabweichungen sind also ähnlich, wird der Standardfehler auf Basis der gepoolten Standardabweichung berechnet:

$$se = \sigma_{\text{pooled}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \quad \text{mit} \quad \sigma_{\text{pooled}} = \sqrt{\frac{n_A \sigma_A^2 + n_B \sigma_B^2}{n_A + n_B}}$$

Messungen in abhängigen Bedingungen einer Gruppe

Zur Herleitung des Standardfehlers der Mittelwertdifferenz zweier abhängiger Messungen A und B *in einer Stichprobe* können wir die Variablen X_A und X_B im ersten Schritt direkt voneinander abziehen:

$$\Delta X = X_A - X_B$$

.. und mit der Standardformel die Standardabweichung σ_Δ dieser Differenz berechnen.

Der Standardfehler der Mittelwertsdifferenz abhängiger Bedingungen A und B ergibt sich daher bei einer Stichprobengröße n als:

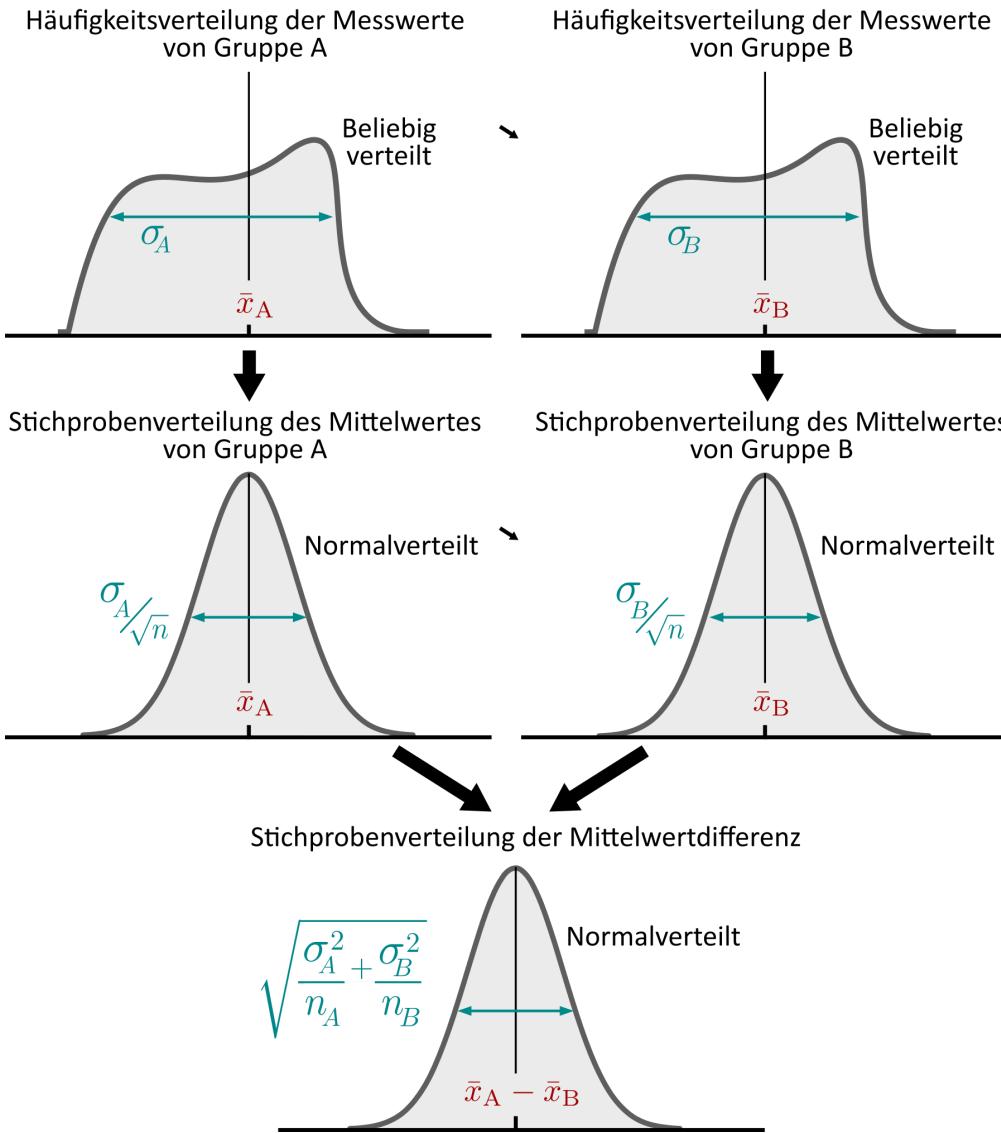
$$se = \frac{\sigma_\Delta}{\sqrt{n}} \quad \text{mit} \quad \sigma_\Delta = \sqrt{\frac{1}{n} \sum (\Delta x_i - \Delta \bar{x})^2}$$

Standardfehler bei Mittelwertdifferenzen: Cheat sheet

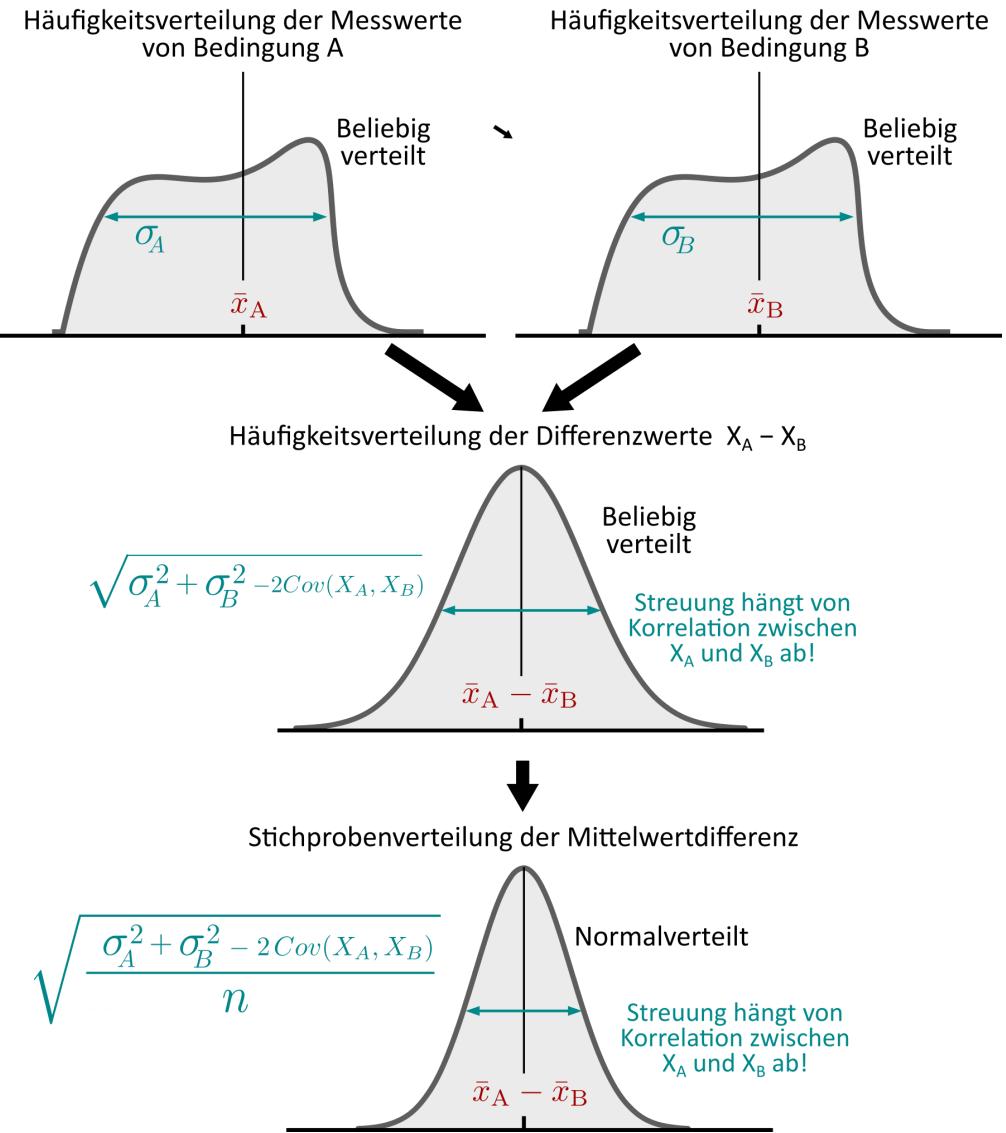
Fall	Berechnung des Standardfehlers se
Differenz des Mittelwertes <u>einer Gruppe</u> und einem Referenzwert μ_0	$se = \frac{\sigma}{\sqrt{n}}$
Mittelwertdifferenz zweier Bedingungen A und B in <u>einer Gruppe</u>	$se = \frac{\sigma_\Delta}{\sqrt{n}}$ mit $\sigma_\Delta = \sqrt{\frac{1}{n} \sum (\Delta x_i - \Delta \bar{x})^2}$ oder $\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \operatorname{Cov}(X_A, X_B)}$
Mittelwertdifferenz <u>zweier unabhängiger Gruppen</u> A und B (ähnliche Varianzen)	$se = \sigma_{\text{pooled}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$ mit $\sigma_{\text{pooled}} = \sqrt{\frac{n_A \sigma_A^2 + n_B \sigma_B^2}{n_A + n_B}}$
Mittelwertdifferenz <u>zweier unabhängiger Gruppen</u> A und B (unterschiedliche Varianzen)	$se = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$
 <p>Dieses Cheat sheet gilt für den Fall, dass die Varianzen σ_A^2 bzw. σ_B^2 in den Populationen bekannt sind! Falls dies nicht der Fall ist, sehen die Formeln aufgrund der Besselkorrektur subtil anders aus (vgl. Vorlesung 11).</p>	

Stichprobenverteilungen von Mittelwertdifferenzen (σ bekannt)

Mittelwertdifferenz: Unabhängige Messungen



Mittelwertdifferenz: Abhängige Messungen



z-Test

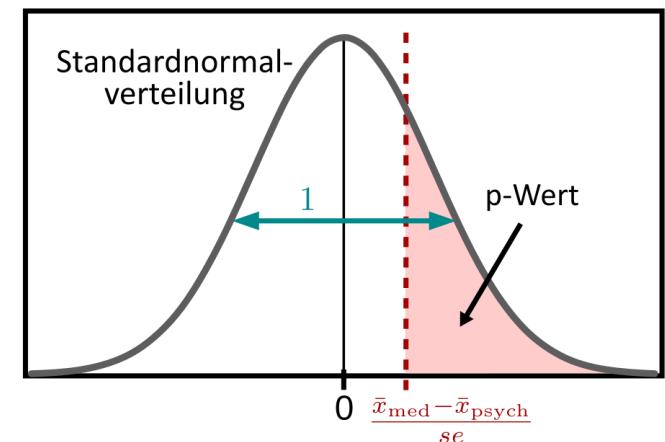
In der Praxis wird die Berechnung von p-Werten um das Konzept der **standardisierten Prüfgröße** erweitert.

- Um das Konzept zu verstehen, erinnern wir uns, dass wir im Beispiel eine Fläche unter der Normalverteilung $\mathcal{N}(0, 0,1)$ berechnet haben. Für die nächste statistische Analyse müssten wir sehr wahrscheinlich die Fläche unter einer Normalverteilung mit einer anderen Streuung als 0,1 berechnen.
- Gerade im Vorcomputerzeitalter war die Berechnung von Flächen unter beliebigen Normalverteilungen eine Herausforderung.
- Abhilfe schafft die **Standardisierung** des Effektes:

$$z = \frac{\Delta\bar{x}}{se}$$

Nach Teilen von $\Delta\bar{x}$ durch die Streuung se , hat die Nullverteilung nicht mehr die Streuung se , sondern *immer* die Streuung $\frac{se}{se} = 1$!

Theoretische Nullhypotesenverteilung
falls σ bekannt



z wird als **standardisierte Prüfgröße** bezeichnet, während der einfache Effekt $\Delta\bar{x}$ eine **unstandardisierte Prüfgröße** darstellte.

z-Test

Es gibt auch im Computerzeitalter noch Vorteile für diese Art der Standardisierung:

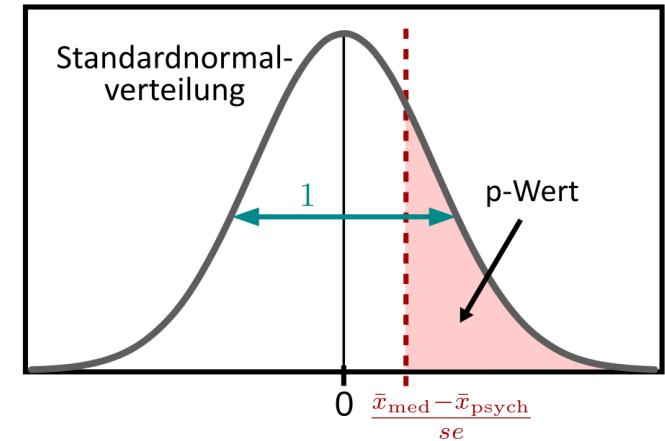
- Der resultierende z-Wert ist vergleichbar zwischen Studien, im Gegensatz zum Mittelwertsunterschied $\Delta\bar{x}$, der von der spezifischen Skala abhängt.
- Der z-Wert hat eine intuitive Interpretation: er gibt an, *wie viele Standardabweichungen der Effekt vom Mittelwert der angenommenen (Null-)Normalverteilung entfernt ist*.

Mit dem Z-Wert als Prüfgröße benötigten wir also für alle Tests dieser Art nur noch eine einzige Verteilung — die Normalverteilung mit Mittelwert 0 und Streuung 1. Man bezeichnet sie als **Standardnormalverteilung**:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \begin{matrix} \mu=0 \\ \sigma=1 \end{matrix} \quad \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Die Art der Berechnung des p-Wertes ändert sich nicht — wir berechnen in unserem Beispiel weiterhin die Fläche rechts von unserem Effekt, nur dass der “Effekt” jetzt standardisiert und als $z = \frac{\Delta\bar{x}}{se} = \frac{\bar{x}_{med} - \bar{x}_{psych}}{se}$ definiert ist. Der resultierende p-Wert ist derselbe.

Theoretische Nullhypotesenverteilung falls σ bekannt



Einschub: Das ABC der Normalverteilung

Aufgrund ihrer Bedeutung in der Statistik, haben sich für die (Standard)Normalverteilung bestimmte Bezeichnung eingebürgert, auf die wir ab jetzt zugreifen werden.

Normalverteilung mit Angabe des Mittelwertes μ und der Varianz σ^2

$$\mathcal{N}(\mu, \sigma^2)$$

Standardnormalverteilung (Mittelwert 0, Varianz 1)

$$\mathcal{N}(0, 1)$$

Dichtefunktion der Normalverteilung

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Dichtefunktion der Standardnormalverteilung

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ (sprich "Klein Phi")}$$

$$\text{Es gilt: } f(x) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$$

Verteilungsfunktion der Normalverteilung

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{x'-\mu}{\sigma}\right)^2} dx$$

Verteilungsfunktion der Standardnormalverteilung

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}x'^2} dx \quad \text{(sprich "Groß Phi")}$$

$$\text{Es gilt: } F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

z-Test

Berechnen wir nun den p-Wert im Nasenlängenbeispiel mithilfe des z-Tests.

Wir bestimmen die standardisierte Prüfgröße z :

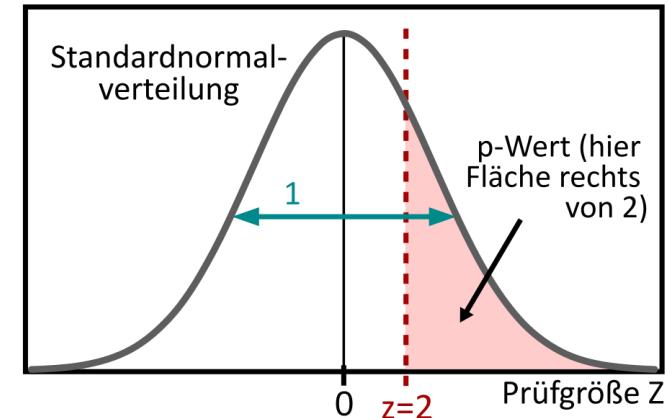
$$z = \frac{\Delta \bar{x}}{se} = \frac{0,2}{0,1} = 2$$

Und berechnen die Fläche rechts von z :

$$p = \int_z^\infty \varphi(x)dx = 1 - \Phi(Z) = 1 - \Phi(2) \stackrel{(Computer)}{=} 0,023$$

Beachte, dass wir hier statt $f(x)$ die Nomenklatur $\varphi(x)$ für die Dichte der Standardnormalverteilung verwenden, und statt $F(x)$ den Ausdruck $\Phi(x)$ für die Verteilungsfunktion der Standardnormalverteilung.

Theoretische Nullhypotesenverteilung falls σ bekannt



z-Test

Der Signifikanztest auf Basis der Standardnormalverteilung – und bei bekannten Streuungen in den relevanten Populationen – wird als **z-Test** bezeichnet.

Der z-Test kann auf alle Arten von Mittelwertsvergleichen angewendet werden:

Fall	Z-Wert	Bekannt	Berechnung von <i>se</i>
Einzelmessung: Vergleich von \bar{x} mit einem Referenzwert μ_0	$z = \frac{\bar{x} - \mu_0}{se}$	σ	$se = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Abhängige Messungen: Vergleich der Mittelwerte \bar{x}_A und \bar{x}_B von zwei Bedingungen A und B in einer Gruppe	$z = \frac{\bar{x}_A - \bar{x}_B}{se} = \frac{\Delta \bar{x}}{se}$	σ_A, σ_B $Cov(X_A, X_B)$	$se = \sqrt{\frac{\sigma_A^2 + \sigma_B^2 - 2 Cov(X_A, X_B)}{n}}$
Unabhängige Messungen: Vergleich der Mittelwerte \bar{x}_A und \bar{x}_B von zwei unabhängigen Gruppen A und B	$z = \frac{\bar{x}_A - \bar{x}_B}{se} = \frac{\Delta \bar{x}}{se}$	σ_A, σ_B	$se = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

Einstichprobentest versus Zweistichprobentest

Statistische Tests, die sich auf eine Stichprobe beziehen (Einzelmessung oder Vergleich zweier abhängiger Messungen) werden auch als **Einstichprobentest** bezeichnet.

Beispielhypothesen für den Einstichprobentest:

- Einzelmessung: Psychologiestudierende sind überdurchschnittlich intelligent ($\overline{IQ} > 100$)
 - Nullhypothese: $\overline{IQ} = 100$
 - Beachte: μ_0 ist in diesem Fall 100 und nicht 0!
- Vergleich zweier abhängiger Messungen: Psychologiestudierende sind morgens intelligenter als abends ($\overline{IQ}_{\text{Morgen}} > \overline{IQ}_{\text{Abend}}$)
 - Nullhypothese: $\overline{IQ}_{\text{Morgen}} - \overline{IQ}_{\text{Abend}} = \Delta \overline{IQ} = 0$

Demgegenüber werden statistische Tests, die sich auf zwei unabhängige Stichproben beziehen, als **Zweistichprobentest** bezeichnet.

Beispielhypothese für den Zweistichprobentest:

- Studierende der HMU sind intelligenter als Studierende der MSB ($\overline{IQ}_{\text{HMU}} > \overline{IQ}_{\text{MSB}}$)
 - Nullhypothese: $\overline{IQ}_{\text{HMU}} - \overline{IQ}_{\text{MSB}} = \Delta \overline{IQ} = 0$

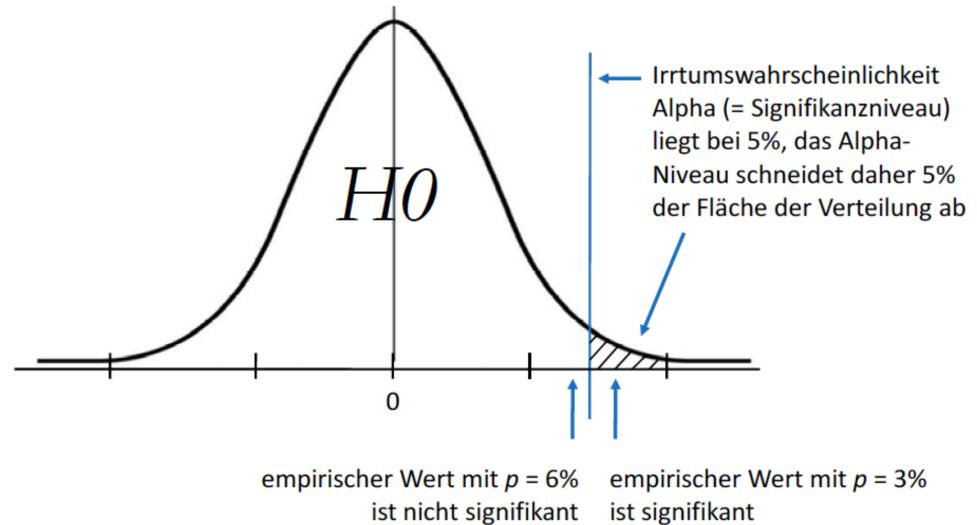
Signifikanzniveau ($p < \alpha$?)

Bislang haben wir die Berechnung des p-Wertes besprochen, aber nicht, wie klein der p-Wert sein sollte, um die Nullhypothese abzulehnen (und im Umkehrschluss den Effekt signifikant zu werten).

Um diese Entscheidung zu treffen, legen wir ein **Signifikanzniveau α** fest. **Unterschreitet der p-Wert das Signifikanzniveau α , lehnen wir die Nullhypothese ab und werten den Effekt als statistisch signifikant.**

Der Wert α wird auch als **Irrtumswahrscheinlichkeit** bezeichnet. Ist beispielsweise $\alpha = 0,1$ so wären wir bereit einen p-Wert von $p < 0,1$ als signifikant zu werten, gehen dabei aber ein 10%-iges Risiko ein, dass die Nullhypothese in Wahrheit doch korrekt ist. D.h. wir nehmen in Kauf, dass wir uns mit 10% Wahrscheinlichkeit irren und fälschlicherweise die Nullhypothese ablehnen.

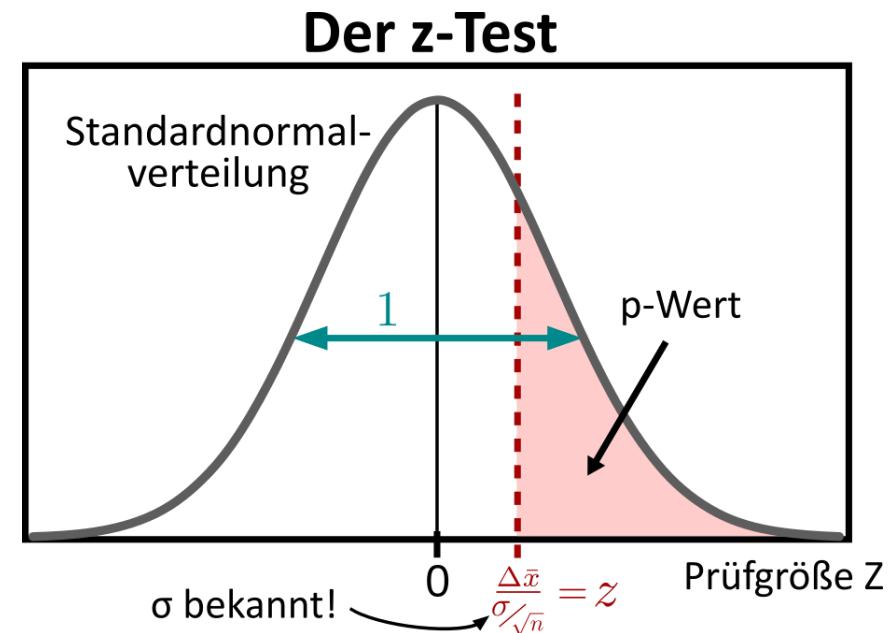
Auf welchen Wert das Signifikanzniveau festgelegt werden sollte ist seit langer Zeit Gegenstand von kontroversen Debatten in der Psychologie. Als de facto Standard-Signifikanzniveau hat sich jedoch der **Wert $\alpha = 0,05$** eingebürgert.



Ein- und zweiseitiges Testen

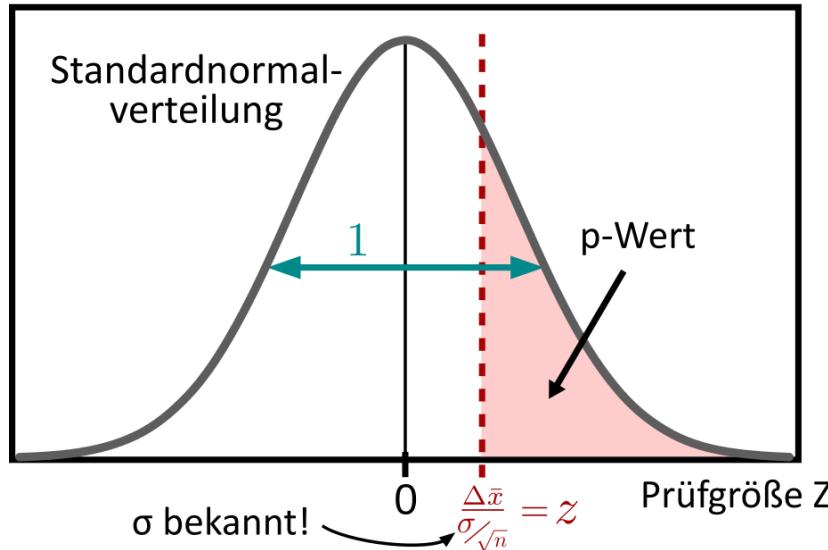
- Im Nasenlängenbeispiel haben wir bislang eine **gerichtete Hypothese betrachtet**, indem wir bei der Hypothese festgelegt haben, welche der beiden Gruppen durchschnittlich eine längere Nase hat (Med-Nasenlänge > Psych-Nasenlänge oder $\Delta\bar{x} > 0$).
- Wir hätten auch die **umgekehrte gerichtete Hypothese** betrachten können:
Med-Nasenlänge < Psych-Nasenlänge oder $\Delta\bar{x} < 0$
- Eine **ungerichtete Hypothese** gibt dagegen keine Richtung vor, d.h. die Hypothese würde schlicht lauten: Med-Nasenlängen und Psych-Nasenlängen sind **unterschiedlich** ($\Delta\bar{x} \neq 0$)

Dies wirkt sich auf die Flächen unter der Nullverteilung aus, die wir betrachten.



Ein- und zweiseitiges Testen

Der z-Test



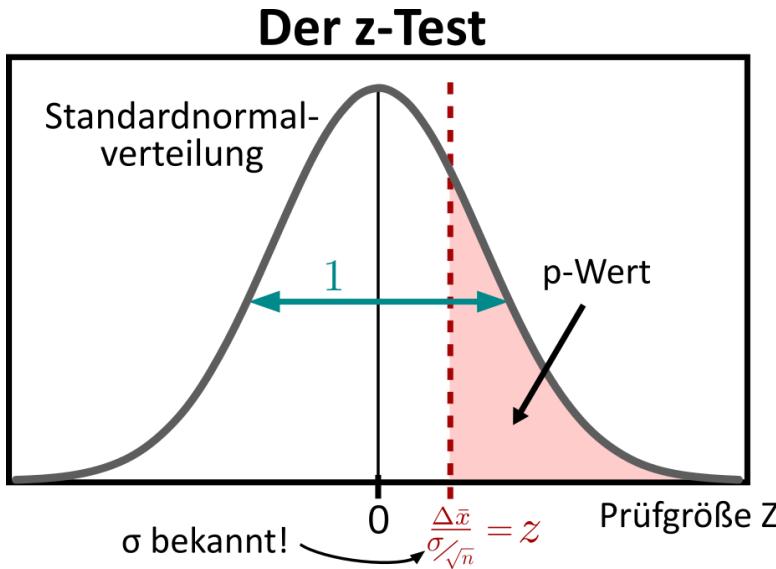
Wie sehen folgende Zusammenhänge:

- Die p-Werte der beiden gerichteten Hypothesen (“größer als” oder “kleiner als”) sind genau **invers**:

$$p(\bar{x}_{\text{med}} > \bar{x}_{\text{psych}}) = 1 - p(\bar{x}_{\text{med}} < \bar{x}_{\text{psych}})$$
- Der p-Wert der ungerichteten Hypothese (“unterschiedlich”) ist gleich **2 x der kleinere p-Wert der beiden gerichteten Hypothesen**:

$$p(\bar{x}_{\text{med}} \neq \bar{x}_{\text{psych}}) = 2 \times \min(p(\bar{x}_{\text{med}} > \bar{x}_{\text{psych}}), p(\bar{x}_{\text{med}} < \bar{x}_{\text{psych}}))$$

Ein- und zweiseitiges Testen



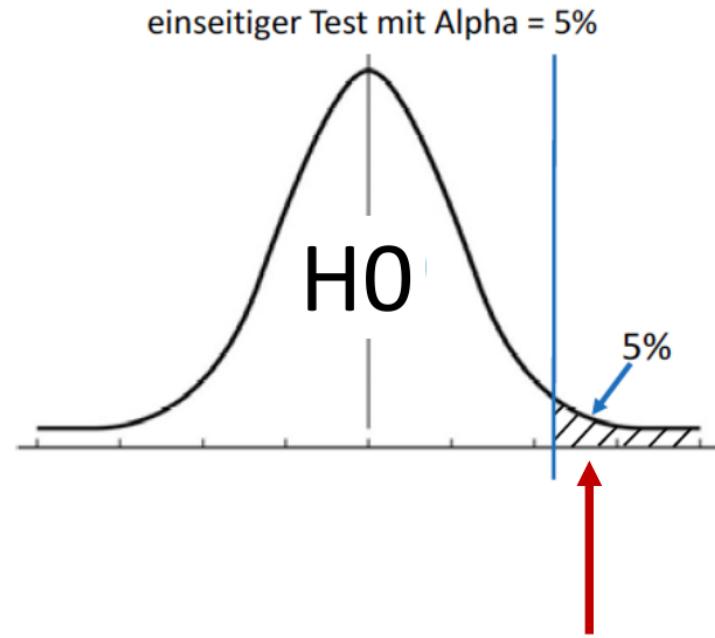
Der p-Wert der ungerichteten Hypothese ist damit immer größer (doppelt so groß) wie der kleinere p-Wert der gerichteten Hypothesen.

Intuition: bei der ungerichteten Hypothese legen wir uns weniger fest, denn unsere Hypothese wäre sowohl erfüllt wenn \bar{x}_{med} signifikant **größer** ist als \bar{x}_{psych} , als auch wenn \bar{x}_{med} signifikant **kleiner** ist als \bar{x}_{psych} . Wir machen uns das Leben (bzw. unsere Vorhesagen) also “einfacher”.

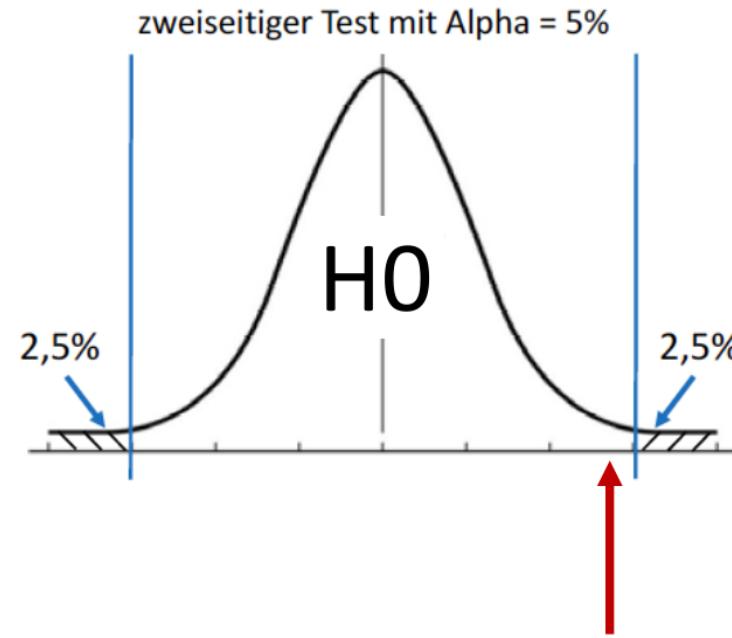
Genauer gesagt verdoppeln wir unsere Chance, mit der Hypothese richtig zu liegen, da a priori beide ungerichtete Hypothesen gleich wahrscheinlich sind. Die Verdopplung des p-Wertes kompensiert dies (man könnte auch sagen “bestraft unsere schwammige Vorhersage”) – nun wird es schwieriger für p das Signifikanzniveau zu unterschreiten.

Ein- und zweiseitiges Testen

Es ist zusätzlich gewinnbringend, sich den Vergleich von gerichteten und ungerichteten Hypothesen aus der Perspektive des Signifikanzniveaus α zu betrachten:



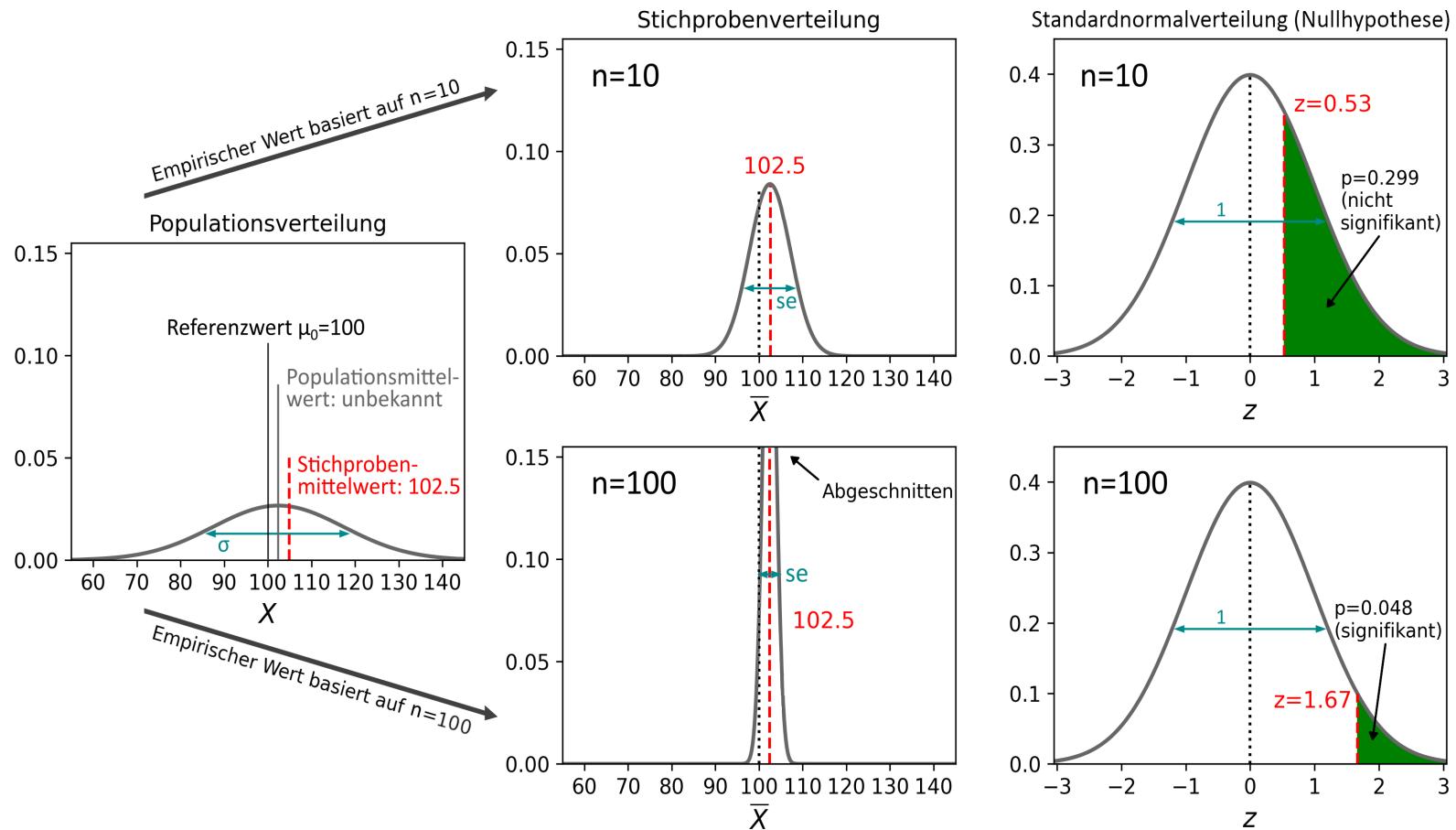
dieser Effekt wäre hier signifikant



hier wäre derselbe Effekt nicht signifikant

Beim zweiseitigen Test verteilt sich die Irrtumswahrscheinlichkeit (im Bild 5%) *auf beide Flanken* der Nullverteilung. Es wird damit auf beiden Seiten schwieriger für unseren Effekt einen noch extremeren Wert aufzuweisen, als durch die Irrtumswahrscheinlichkeit vorgegeben.

p-Wert versus Stichprobengröße



Getestet wird, ob der IQ der betrachteten Population (hier 102,5) größer ist, als der Referenzwert $\mu_0 = 100$, der idealerweise den durchschnittlichen IQ aller Menschen darstellt. Im Bild ist zu sehen, dass die betrachtete Population (z.B. alle Brandenburger:innen) tatsächlich einen etwas höheren Mittelwert als 100 aufweisen, was sich auch im Stichprobenmittelwert niederschlägt. Wie verändert sich der p-Wert abhängig davon, ob die Stichprobe $n = 10$ oder $n = 100$ umfasst? Schritt 1: der Standardfehler se ist bei der höheren Stichprobenzahl wesentlich kleiner (aufgrund von $\hat{\sigma}/\sqrt{n}$) und damit die Stichprobenverteilung bei $n = 100$ wesentlich schmäler. Schritt 2: der kleinere Standardfehler wirkt sich auf den z-Wert $z = \frac{102,5}{se}$ aus: kleinerer Standardfehler bedeutet größerer z-Wert. Schritt 3: größerer z-Wert bedeutet kleinerer p-Wert, da kleinere Fläche rechts von z.

p-Wert versus Stichprobengröße/Populationsstreuung

Gibt es den untersuchten Effekt in der Population tatsächlich (Alternativhypothese wahr), gilt:

- Größere Stichprobengrößen führen im Schnitt zu kleineren p-Werten.
- Kleinere Populationsstreuungen führen zu kleineren p-Werten.

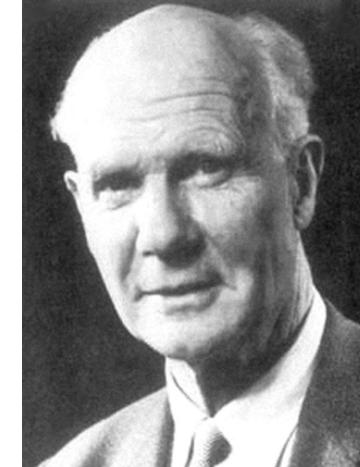
Gibt es den untersuchten Effekt in der Population nicht (Nullhypothese wahr), gilt:

- Größere Stichprobengrößen haben keine Auswirkung auf den p-Wert.
- Kleinere Populationsstreuungen haben keine Auswirkung auf den p-Wert.
- Alle p-Werte sind gleich wahrscheinlich (d.h. p-Werte haben eine uniforme Verteilung auf dem Intervall [0; 1])

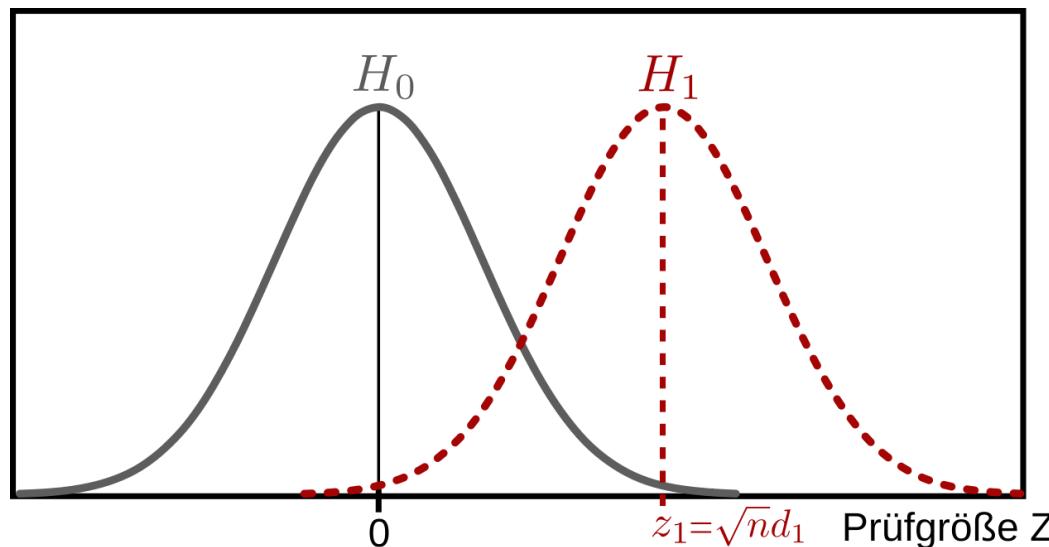
Binäres Entscheidungskonzept von Neyman und Pearson

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.



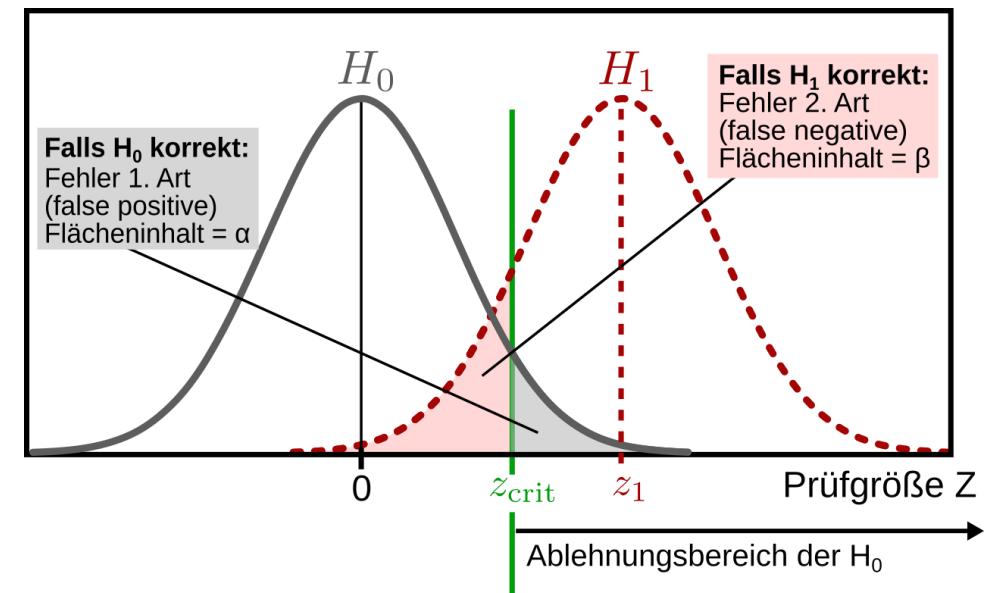
Egon Sharpe Pearson
(1895-1980), Sohn von
Karl Pearson



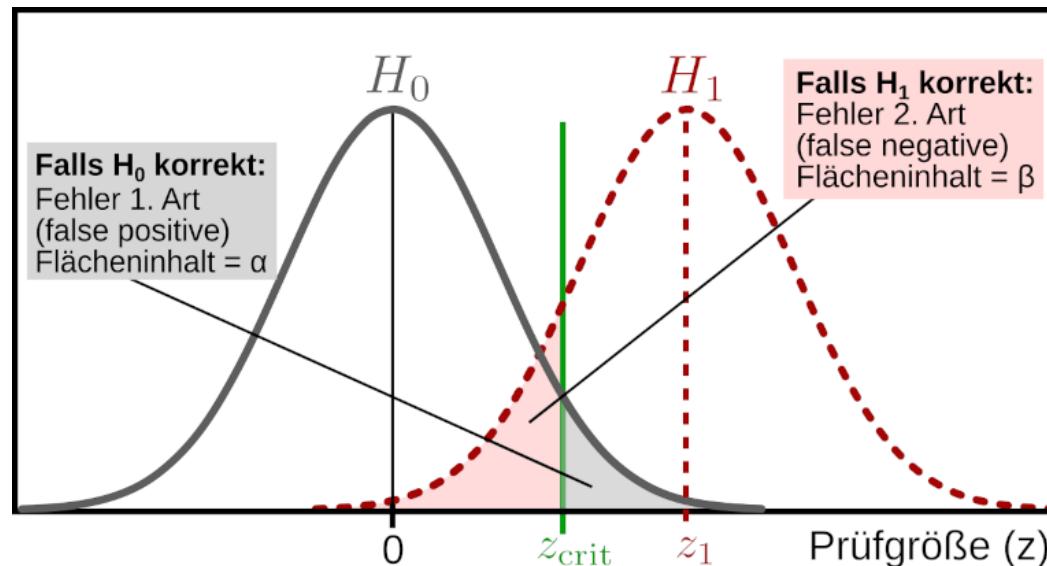
Jerzy Neyman (1894-1981)

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 <i>true negative</i> α-Fehler („Fehler erster Art“) false positive
In der Population gilt die H_1	β-Fehler („Fehler zweiter Art“) false negative Korrekte Entscheidung <i>true positive</i>

Fisher versus Neyman/Pearson

- Nullhypotesenframework 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.

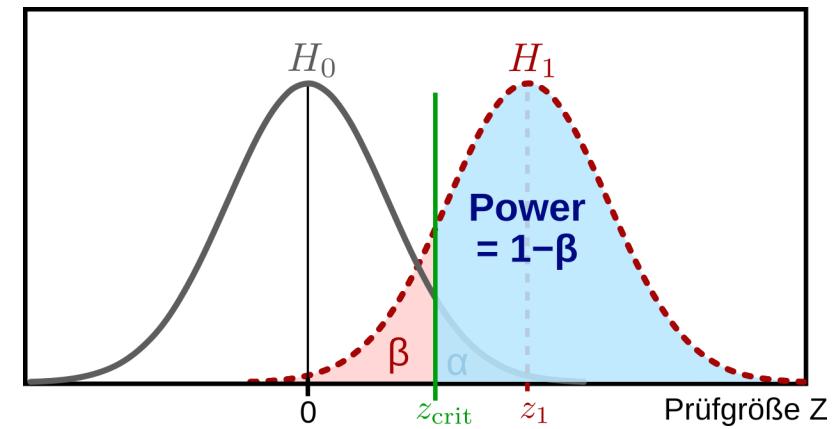
Quelle: Wikipedia³



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

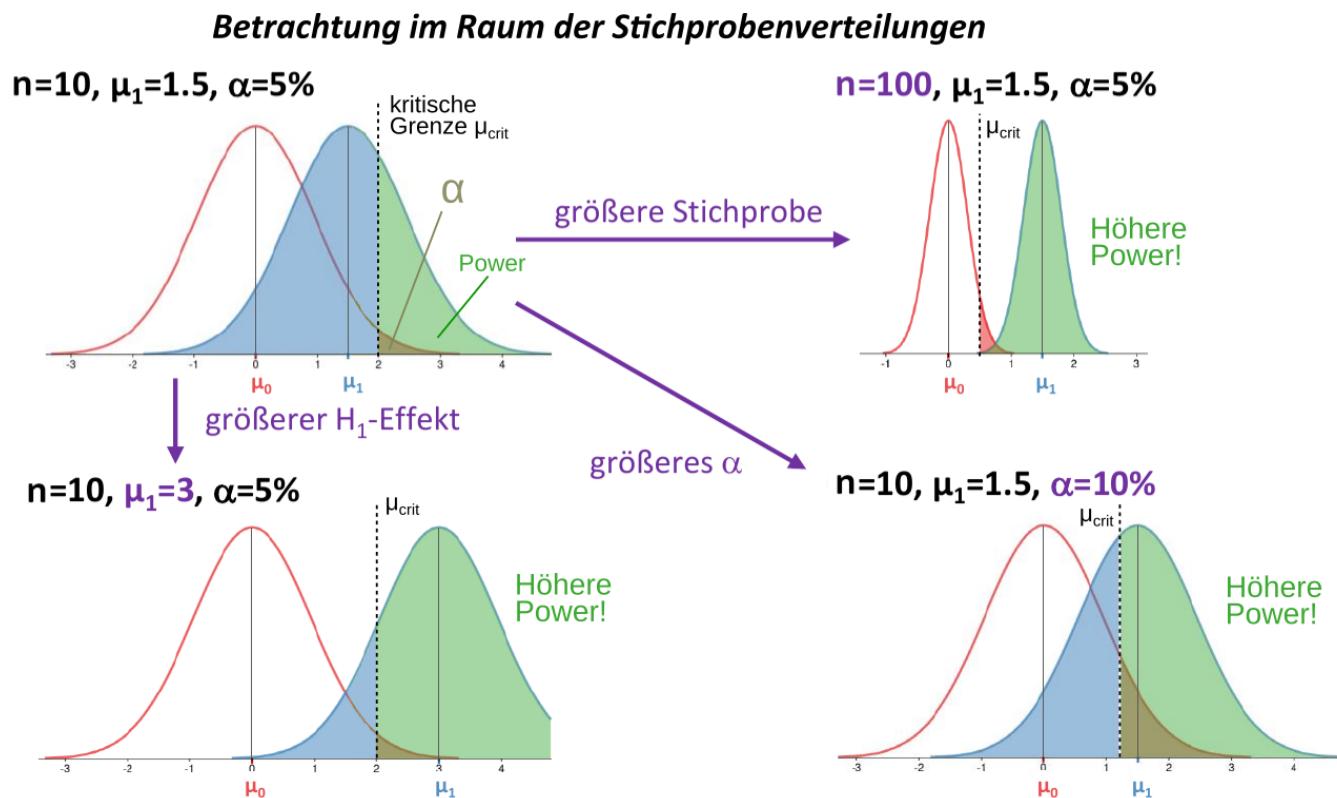
Im Kontext der Stichprobenberechnung, gibt das Neyman-Pearson-Framework eine Antwort auf folgende Frage:

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?

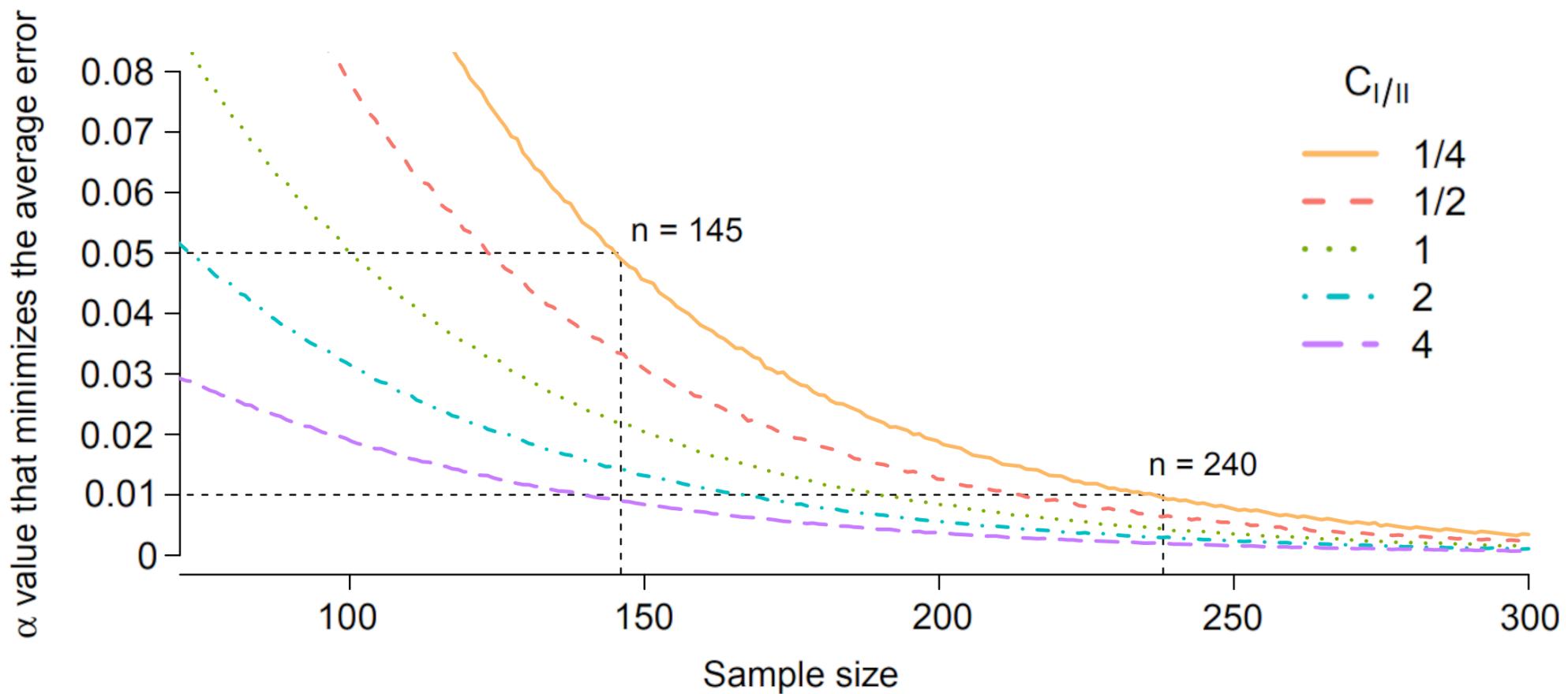
Teststärke (Power)

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



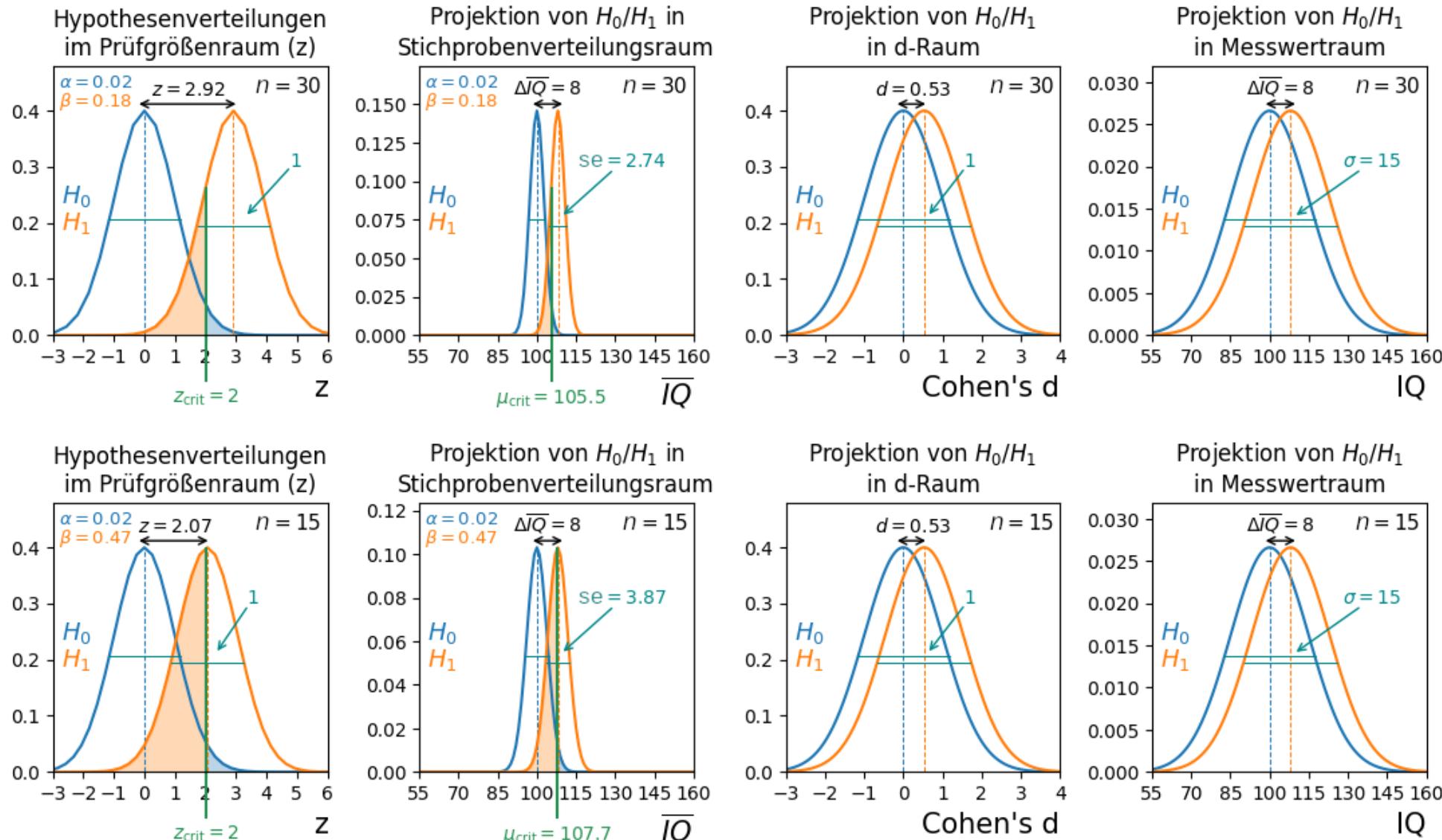
Relevanz für Berechnung der Stichprobengröße



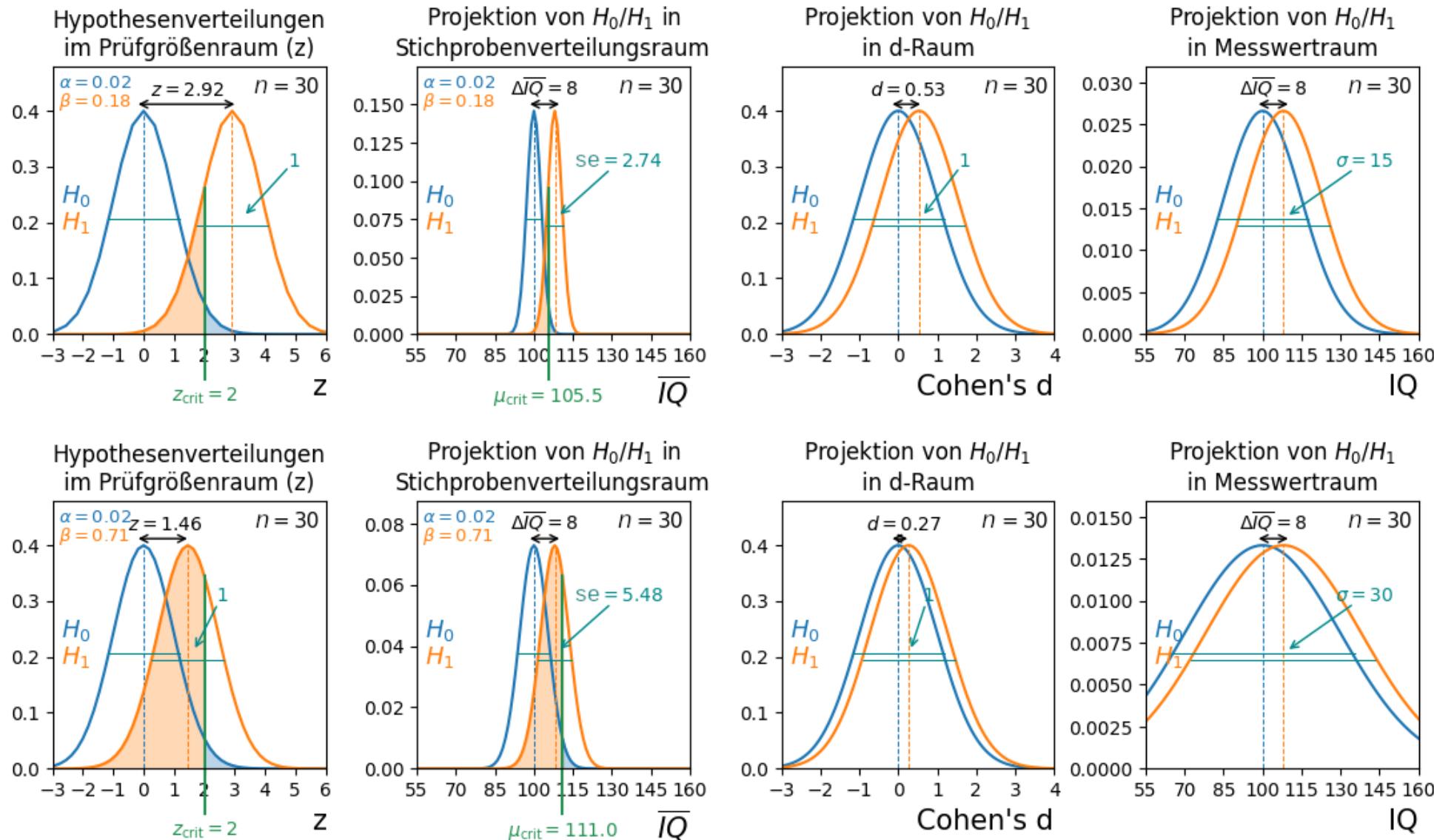
Die Abbildung zeigt den optimalen α -Wert für verschiedene Stichprobengrößen und verschiedene Verhältnisse von Fehler 1. und 2. Art (in der Legende bedeutet z.B. $C_{I/II} = 1/4$, dass die Fehlerrate β zweiter Art 4x so klein sein soll, wie die Fehlerrate α erster Art). Die "Optimalität" des α -Wertes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



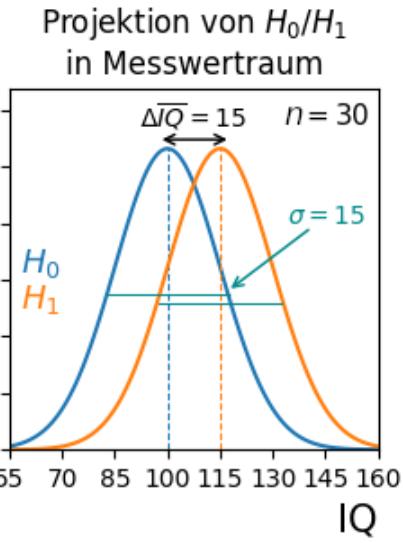
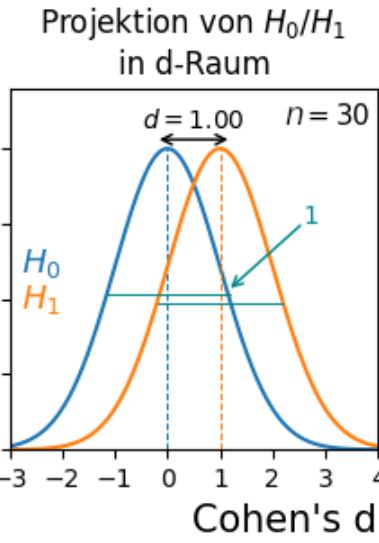
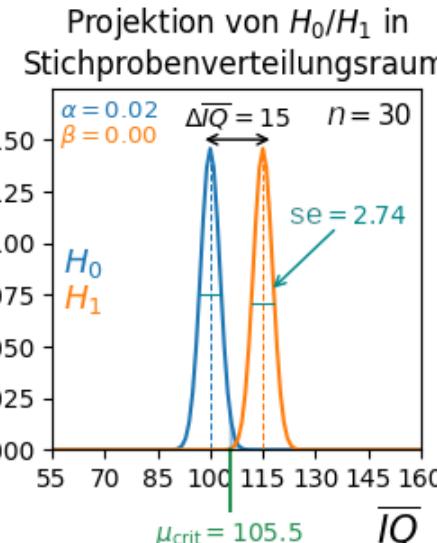
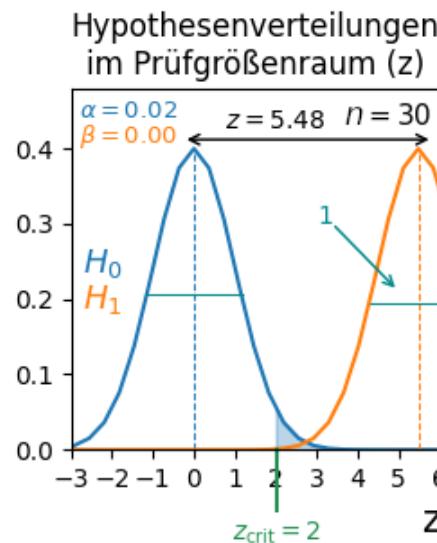
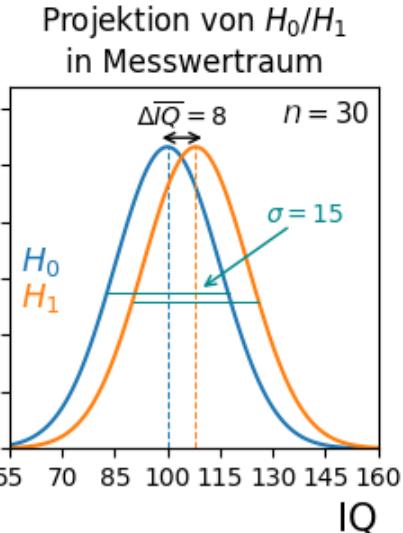
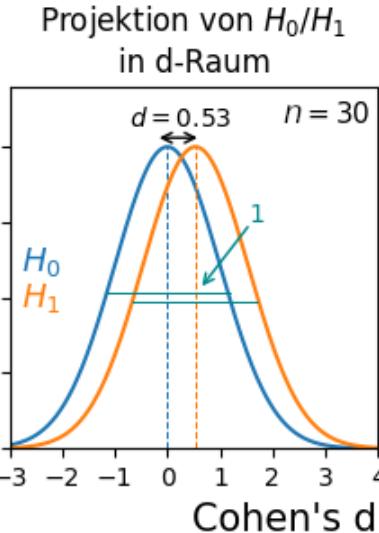
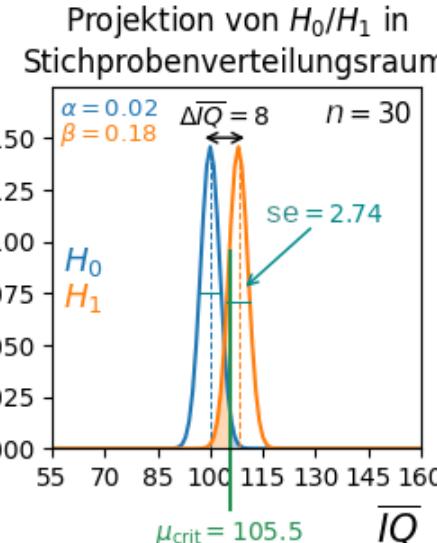
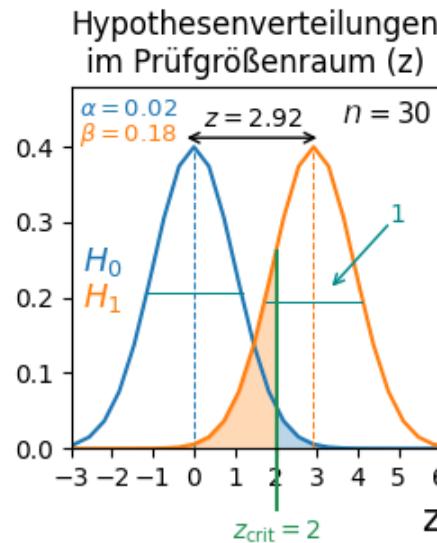
Veränderung der Stichprobengröße n



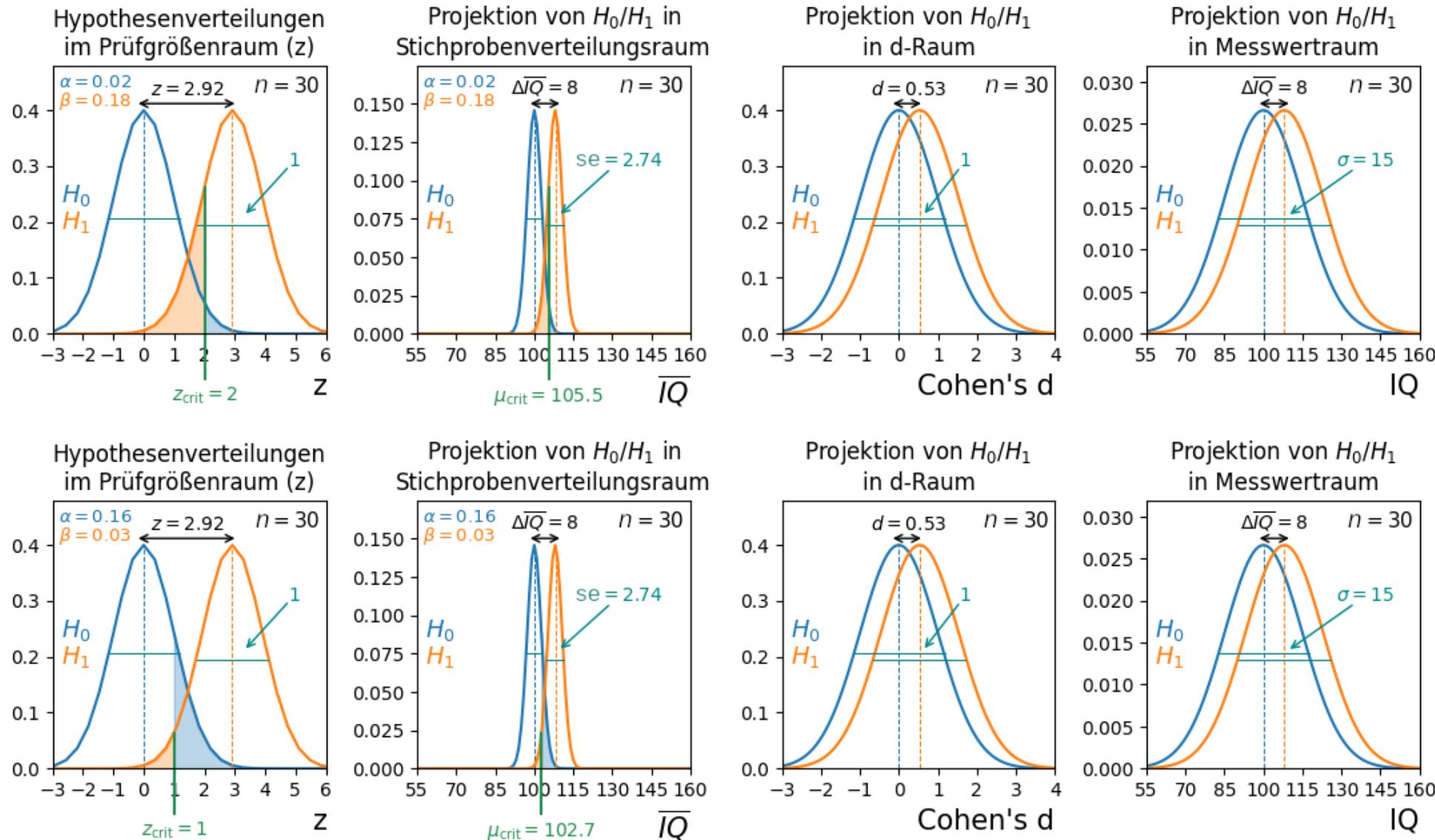
Veränderung der Populationsstreuung σ



Veränderung des Mittelwertsunterschiedes ΔIQ



Veränderung des kritischen Entscheidungswertes z_{crit}



[Zusammenfassung]

- **Signifikanztests** prüfen die Wahrscheinlichkeit, mit der unser experimenteller Effekt (oder ein noch extremerer Effekt) durch Zufall entstanden sein könnte (d.h. für die Annahme dass die *Nullhypothese tatsächlich zutreffend* ist).
- Signifikanztests nach Fisher basieren auf der **Stichprobenverteilung um den Wert der Nullhypothese**.
- Signifikanztests treffen Entscheidungen für (Neyman & Pearson) oder gegen (Fisher) eine Hypothese, basierend auf dem berechneten p-Wert und dem festgelegten **Signifikanzniveau α** .
- Das **Binäres Entscheidungskonzept von Neyman und Pearson** findet häufig bei der **Power-Berechnung** in der Planungsphase einer Studie Anwendung.



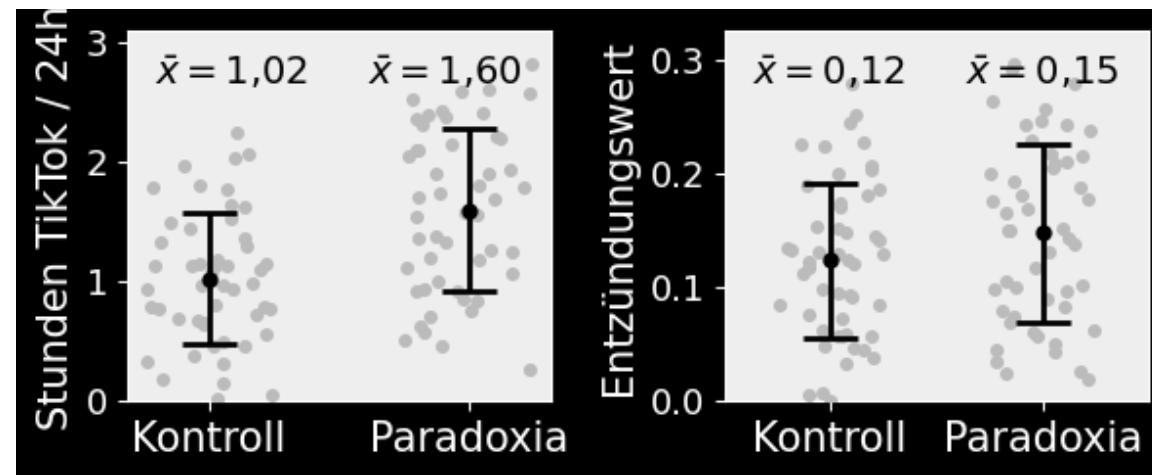
Nach kurzer Beratung in der Task Force sind Sie sich einig, dass die Voraussetzungen für einen z-Test nicht optimal sind. Die Populationsstreuung lässt sich nicht aus vorhergehenden Studien ableiten, da die untersuchten Phänomene brandneu sind. Da die Fallzahl von $n = 50$ gemäß der Faustregel “ $n \geq 30$ ” aber ausreichend wäre, rechnen Sie aus reiner Neugier dennoch z-Tests für die Gruppenvergleiche.

Sie erstellen sich eine Tabelle mit den relevanten Parametern für den z-Test:

<u>TikTok</u>	Fallzahl n	Mittelwert \bar{x}	Standardabweichung σ
Kontroll	50	1,022	0,545
Paradoxia	50	1,599	0,677



<u>Entzündung</u>	Fallzahl n	Mittelwert \bar{x}	Standardabweichung σ
Kontroll	50	0,1233	0,0682
Paradoxia	50	0,1477	0,0783



Um den z-Wert zu bestimmen, benötigen Sie den Standardfehler. Da es sich um unabhängige Messungen in zwei Gruppen handelt, verwenden Sie die Formel, die auf der mittleren Varianz beider Gruppen basiert:

$$se = \sqrt{\frac{\sigma_{\text{control}}^2}{n_{\text{control}}} + \frac{\sigma_{\text{paradox}}^2}{n_{\text{paradox}}}}$$



Sie erhalten:

TikTok: $se = \sqrt{\frac{0,545^2}{50} + \frac{0,677^2}{50}} = 0,123$ $\Delta\bar{x} = \bar{x}_{\text{paradox}} - \bar{x}_{\text{control}} = 1,599 - 1,022 = 0,577$

Entzündung: $se = \sqrt{\frac{0,0682^2}{50} + \frac{0,0783^2}{50}} = 0,0147$ $\Delta\bar{x} = \bar{x}_{\text{paradox}} - \bar{x}_{\text{control}} = 0,1477 - 0,1233 = 0,0243$

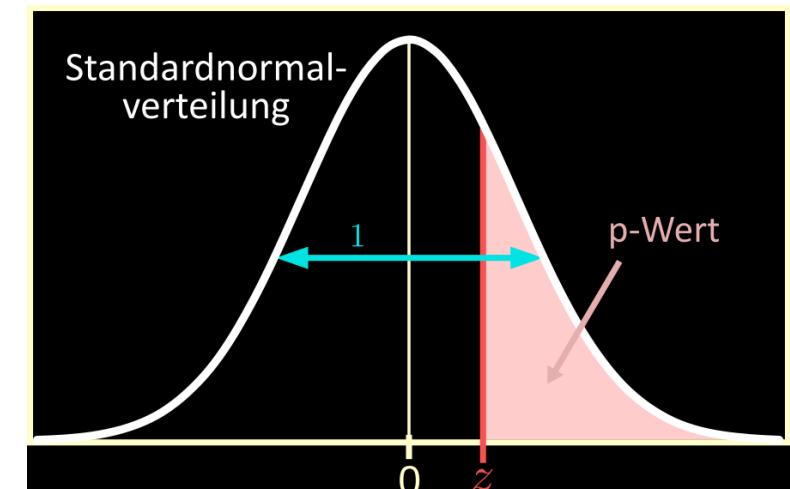
Der z-Wert ergibt sich als das Verhältnis aus dem “Effekt” (hier $\Delta\bar{x}$) und dem Standardfehler:

$$\text{TikTok: } z = \frac{\Delta\bar{x}}{se} = \frac{0,577}{0,123} = 4,69$$

$$\text{Entzündung: } z = \frac{\Delta\bar{x}}{se} = \frac{0,0243}{0,0147} = 1,65$$

In beiden Fällen haben Sie eine **gerichtete** Hypothese, nämlich dass Pardoxiker höhere Werte als Kontrollen aufweisen. Der p-Wert entspricht also der Fläche unter der Standardnormalverteilung *rechts* vom z-Wert und damit dem Integral:

$$p = \int_z^{\infty} \varphi(x)dx = 1 - \Phi(z)$$



Die Spannung steigt, als Sie nun zum ersten Mal die Signifikanz Ihrer Ergebnisse beurteilen können. Sie erhalten folgende p-Werte:

$$\text{TikTok: } p = 1 - \Phi(4,69) \stackrel{\text{(Computer)}}{=} 0,000001$$



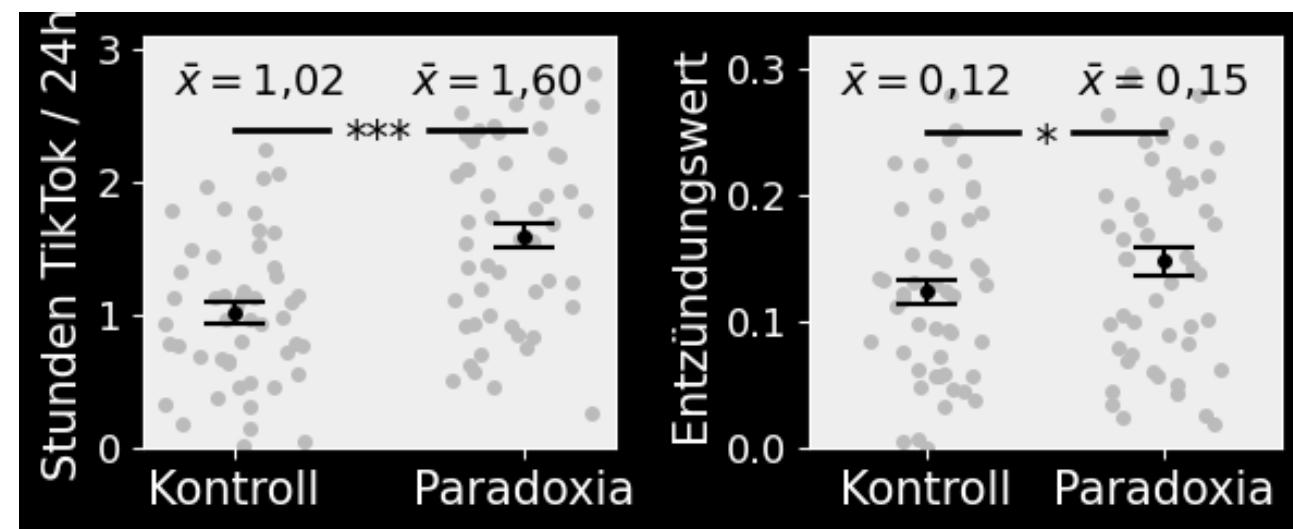
$$\text{Entzündung: } p = 1 - \Phi(1,67) \stackrel{\text{(Computer)}}{=} 0,049$$

Dieser erste Signifikanztest ergibt also, dass die TikTok-Hypothese mit überwältigender Signifikanz bestätigt wird. Aber auch die Entzündungs-Hypothese wird mit einem p-Wert von 0,049 – und einem Signifikanzniveau von $\alpha = 0,05$ – um Haarsbreite bestätigt.

Noch genießen Sie die Ergebnisse aber mit Vorsicht, da Ihnen bewusst ist, dass der z-Test nicht der ideale Test für Ihre Studie ist.

Sie aktualisieren vorläufig Ihre vorherige Abbildung in zweierlei Hinsicht:

- Sie ersetzen die Standardabweichung durch den Standardfehler. Dieser legt die Betonung auf den Effekt der Sie interessiert: den Unterschied der Mittelwerte.
- Sie fügen Signifikanzsternchen ein. Eine gängige Notation ist ein Sternchen (*) für p-Werte kleiner 0,05, zwei Sternchen (**) für p-Werte kleiner 0,01 und drei Sternchen (***) für p-Werte kleiner 0,001.



Fußnoten

1. <https://www.majordifferences.com/2016/10/5-differences-between-null-and.html>
2. <https://www.statisticsfromatoz.com/blog/new-video-null-hypothesis>
3. https://de.wikipedia.org/wiki/Der_Hirtenjunge_und_der_Wolf
4. 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.