

M24 Statistik 1: Wintersemester 2024 / 2025

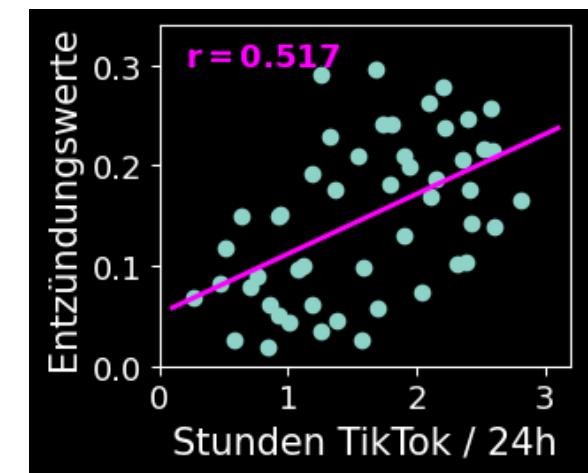
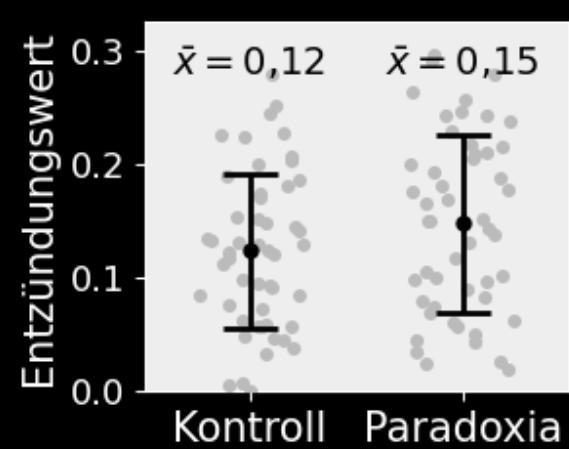
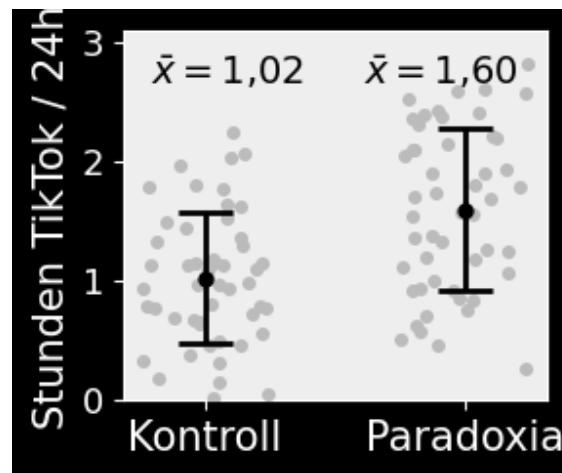
Vorlesung 08: Signifikanztestung

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?

Signifikanztestung

Ziel der **Signifikanztestung** ist zu klären, inwieweit ein empirischer Effekt $\hat{\theta}$ statistisch bedeutsam ist – oder aber durch Zufall erklärt werden kann. Signifikanztests bieten damit ein **Kriterium, um wissenschaftliche Hypothesen zu überprüfen**.

Die Signifikanztestung grenzt sich zur praktischen Signifikanz im Rahmen der Effektstärke ab, da Effekte stark sein können, ihre Beobachtung aber unter Umständen dennoch durch Zufall erklärt werden kann.

Zwei etablierte Frameworks zur Signifikanztestung haben sich im Lauf der Zeit entwickelt: die **Bayesianische Testung** und die **Nullhypotesentestung**. In Statistik 1 betrachten wir ausschließlich das Framework der Nullhypotesentestung.

Nullhypothese und Alternativhypothese

Nullhypothese (H_0)

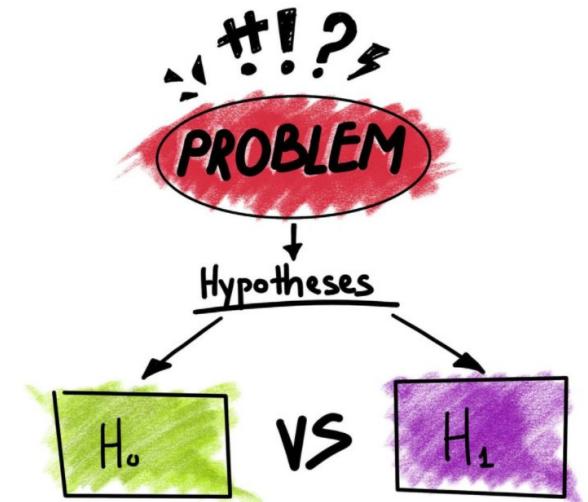
Annahme, dass ein Effekt θ in der Population *nicht* vorliegt.

- Die Häufigkeit des Haarschneidens und Haardichte **hängen nicht** zusammen.
- Frauen haben **keine höhere** emotionale Intelligenz als Männer.

Alternativhypothese (H_1)

Annahme, dass ein Effekt θ in der Population *vorliegt*.

- Die Häufigkeit des Haarschneidens und Haardichte **hängen** zusammen.
- Frauen haben **eine höhere** emotionale Intelligenz als Männer.

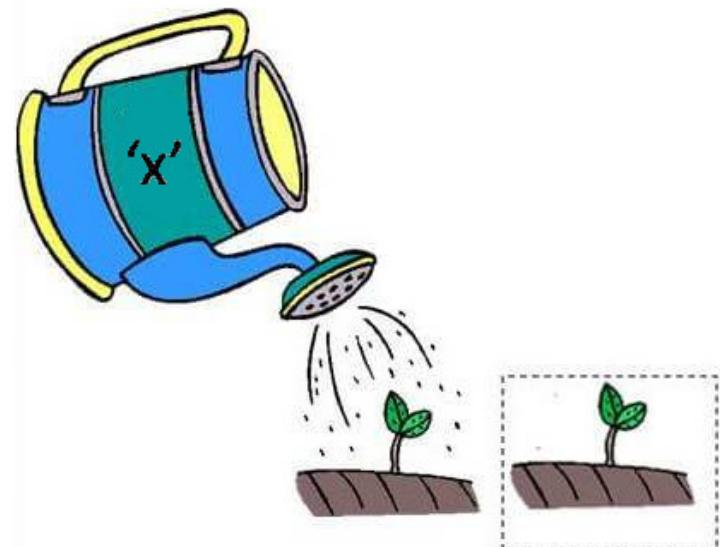


Nullhypothese und Alternativhypothese: Beispiel

Effect of Bio-fertilizer 'x' on Plant growth

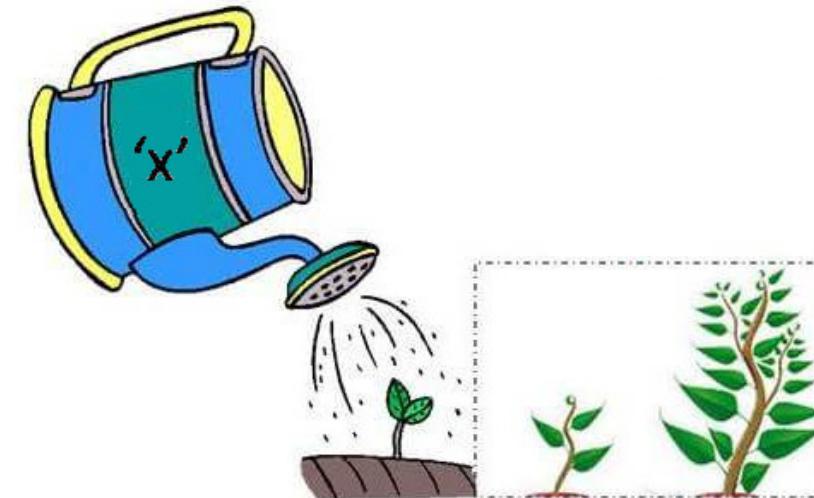
Null Hypothesis

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

p-Wert

Nullhypotesentestung nach Fisher

- Betrachtet wird ausschließlich die Nullhypothese.
- Ziel ist es zu zeigen, dass der Stichprobeneffekt $\hat{\theta}$ unter der Annahme dieser Nullhypothese so unwahrscheinlich ist, dass man die Nullhypothese „verwerfen“ oder „ablehnen“ kann.
- In diesem Fall wird der Stichprobeneffekt $\hat{\theta}$ als **signifikant** gewertet.
- Bei der Nullhypotesentestung wird kein Schluss in Bezug auf eine Alternativhypothese gezogen (sie muss hier nicht einmal formuliert werden!).



Ronald Aylmer Fisher
(1890-1962)



Bei der Nullhypotesentestung nach Fisher werden keine Wahrscheinlichkeiten von Hypothesen berechnet — weder $P(H_0)$ noch $P(H_1)$ — sondern die Wahrscheinlichkeit von Stichprobendaten unter Annahme der Nullhypothese.

Die Wahrscheinlichkeit von Daten gegeben eine Hypothese, $P(D|H)$, heißt auch **Likelihood**.

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

→ **p-Wert**

p-Wert

Sie überprüfen die Hypothese, dass die mittlere Nasenlänge von Medizinstudierenden (μ_{med}) größer ist als die mittlere Nasenlänge von Psychologiestudierenden (μ_{psych}).

Die zugehörige Nullhypothese H_0 lautet: *Psychologiestudierende haben gleich lange oder kürzere Nasen als Medizinstudierende.*

Oder mathematisch:

$$\mu_{\text{psych}} \leq \mu_{\text{med}} \quad \text{bzw.} \quad \theta = \mu_{\text{psych}} - \mu_{\text{med}} \leq 0$$

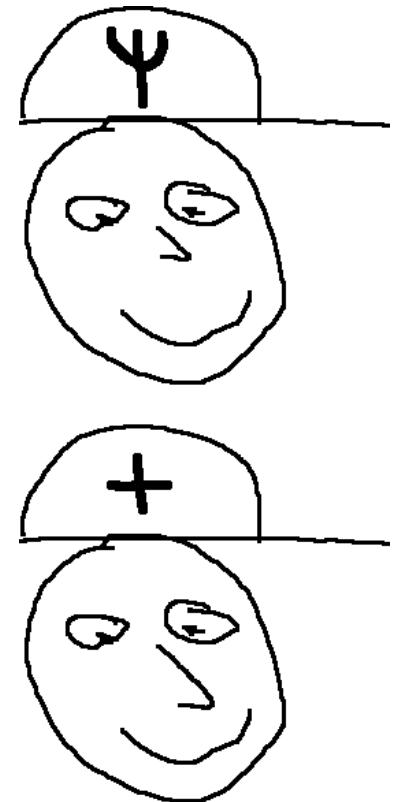
Sie führen eine Studie durch ($n_{\text{med}} = n_{\text{psych}} = n = 30$) und messen folgenden Gruppenunterschied:

$$\hat{\theta} = \Delta \bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}} = 0.2 \text{ cm}$$

Nach Fisher stellen wir nun folgende präzise Frage:

Was ist die Wahrscheinlichkeit dieses Effektes ($\Delta \bar{x} = 0.2 \text{ cm}$) oder eines noch extremeren Effektes ($\Delta \bar{x} > 0.2 \text{ cm}$), obwohl die Nullhypothese H_0 in Wahrheit zutrifft?

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

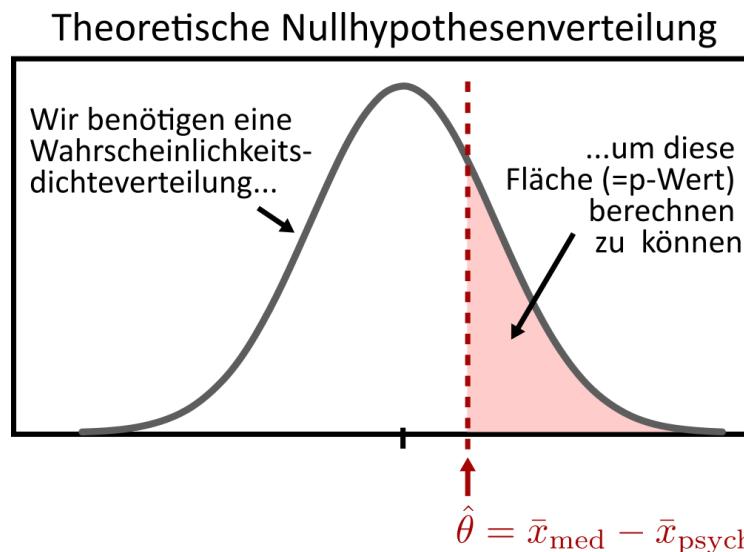


p-Wert

Wir definieren:

Defin ition Der **p-Wert** ist die Wahrscheinlichkeit den Stichprobeneffekt $\hat{\theta}$ oder einen noch extremeren Effekt zu erhalten, obwohl für die Population die Nullhypothese angenommen wird.

- Um die Wahrscheinlichkeit zu berechnen, dass eine Stichprobe einen Effekt mit dem Wert $\hat{\theta}$ oder einem extremeren Wert erzielt, müssen wir die Fläche unter einer Art Stichprobenverteilung berechnen – allerdings unter Annahme der Nullhypothese!
- Wir bezeichnen diese Verteilung als **Theoretische Nullhypotesenverteilung** oder kurz **Nullhypotesenverteilung** oder noch kürzer **Nullverteilung**.

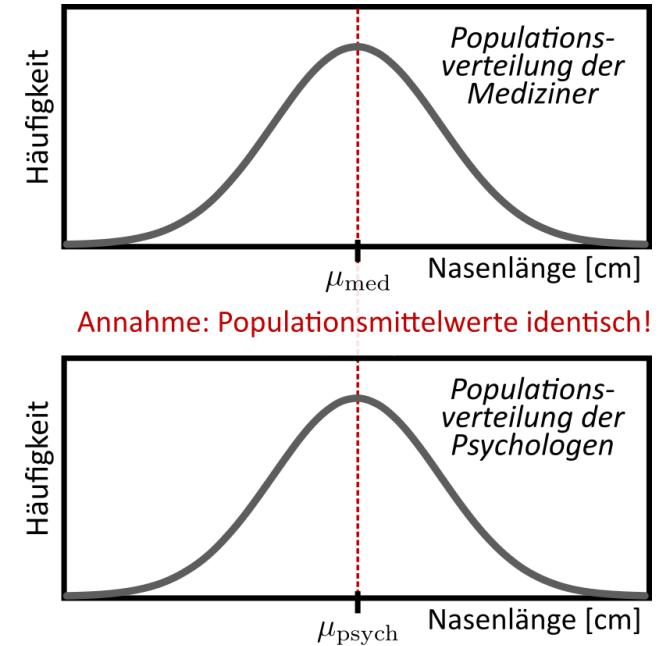


Konstruktion der Nullverteilung

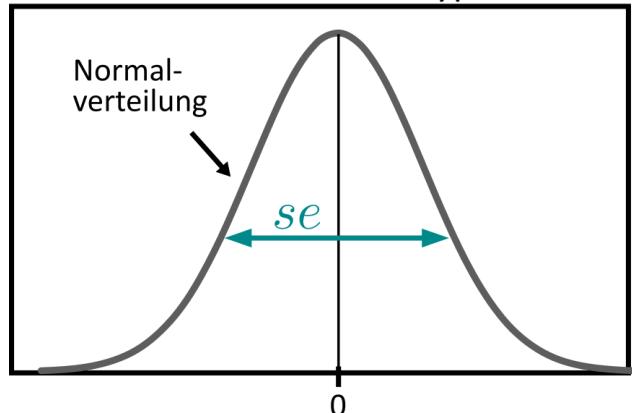
Konstruktion der Nullverteilung

Die Nullhypothese besagt, dass es den vermuteten Effekt in der Population **nicht gibt**. In unserem Nasenlängenbeispiel wäre dies beispielsweise der Fall, wenn sich die wahren Nasenlängenpopulationsmittelwerte zwischen Psychologen und Medizinern **nicht unterscheiden**, also in Wahrheit gilt:

$$\mu_{\text{med}} = \mu_{\text{psych}}$$



Theoretische Stichprobenverteilung bei Annahme der Nullhypothese



*Wenn diese Annahme zuträfe, wissen wir, dass für die Differenz der Nasenlängenmittelwerte gilt: $\mu_{\text{med}} - \mu_{\text{psych}} = 0$. Damit wäre auch klar, wie die theoretische Erwartung für die Stichprobenverteilung aussähe, nämlich eine **Normalverteilung mit Mittelwert 0**. Das ist die Nullverteilung!*

Die Streuung der Nullverteilung kennen wir auch. Da es sich um eine Stichprobenverteilung handelt, entspricht die Standardabweichung der Verteilung dem **Standardfehler se**.

Konstruktion der Nullverteilung



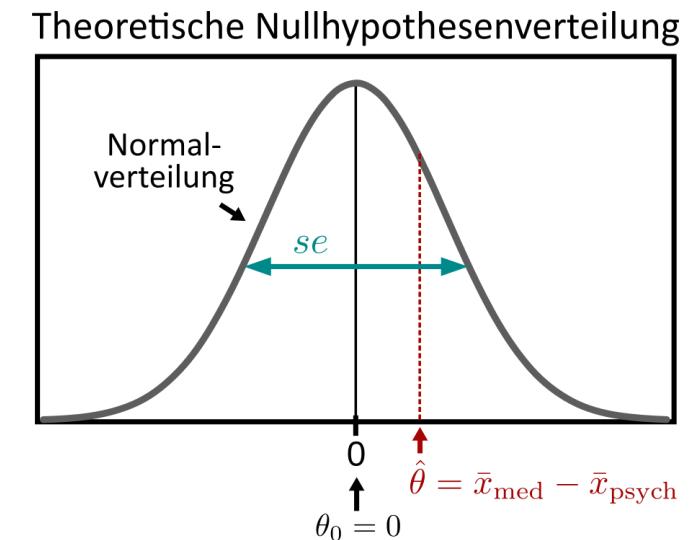
In der Praxis ist die Streuung se idR nicht bekannt und muss als \hat{se} geschätzt werden. Wir betrachten diesen etwas komplexeren Fall hier noch nicht. Jedoch ist diese Problematik größer, als auf den ersten Blick angenommen, und führt uns in der Vorlesung zum t-Test zur Erkenntnis, dass die Annahme einer Normalverteilung für die Nullhypotesenverteilung nicht mehr gerechtfertigt ist (stattdessen benötigen wir dann die t-Verteilung). Behalten Sie dies im Hinterkopf!

Konstruktion der Nullverteilung: Hinweis zum Mittelwert θ_0

In unserem Beispiel hatten wir als Nullhypothese:

$$\theta = \mu_{\text{psych}} - \mu_{\text{med}} \leq 0$$

... also die Null-Annahme, dass die Nasenlängen der Psychologen *nicht größer bzw. kleiner gleich* sind als die Nasenlängen der Mediziner. Die zugehörige Nullhypotesenverteilung ist eine Normalverteilung mit Mittelwert $\theta_0 = 0$.



Frage: warum eigentlich gerade $\theta_0 = 0$ und nicht irgendein Wert $\theta_0 < 0$? Schließlich besagt die Nullhypothese im Beispiel ja gerade $\theta \leq 0$, also die Annahme, dass der Effekt θ in der Population **kleiner** oder gleich Null ist?

Antwort: $\theta_0 = 0$ ist der “unvorteilhafteste” Mittelwert für die Ablehnung der Nullhypothese, da $\theta_0 = 0$ dem Stichprobeneffekt $\hat{\theta}$ näher ist als alle anderen potentiellen H_0 -Mittelwerte $\theta_0 < 0$. Wir bewerten also einen empirischen Effekt $\hat{\theta}$ als signifikant, wenn er so stark ist, dass wir die Nullhypothese sogar unter Annahme des Mittelwertes $\theta_0 = 0$ ablehnen können. Für $\theta_0 < 0$, kann in diesem Fall die Nullhypothese erst recht abgelehnt werden.

Kurzum: es genügt für $\theta_0 = 0$ zu testen, alle anderen Fälle $\theta_0 < 0$ sind damit abgedeckt.

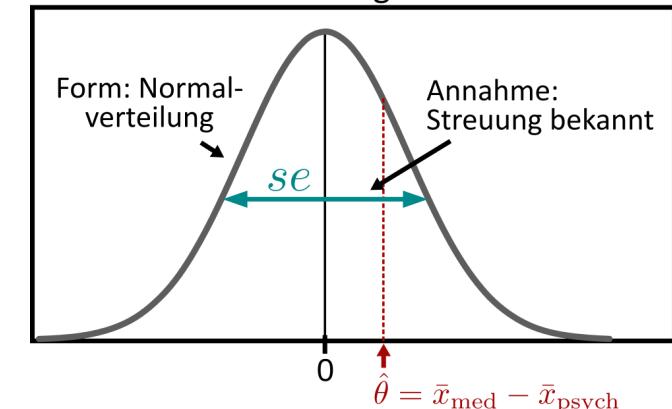
z-Test

Vereinfachung: Populationsstreuungen bekannt

- In einem ersten Schritt betrachten wir den **vereinfachten Fall, dass die Streuung se bekannt ist.**
- Entweder ist se selbst bekannt, oder aber die Streuungen σ der Merkmale, die dem Effekt zugrunde liegen, denn aus diesen lässt sich se ableiten.
- Im Nasenlängenbeispiel etwa können wir annehmen, dass die Populationsstreuungen σ_{med} und σ_{psych} des Merkmals Nasenlänge bekannt sind. Der Standardfehler se der Mittelwertdifferenz lässt sich aus diesen Populationsstreuungen wie folgt ableiten (siehe [Cheatsheet](#)):

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

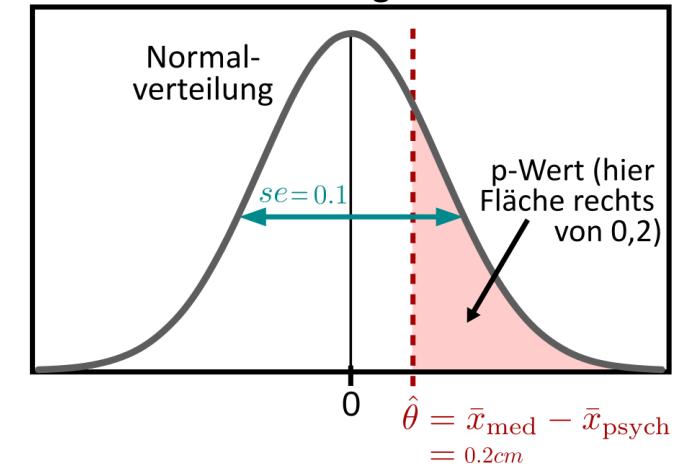
Theoretische Nullhypotesenverteilung falls Streuung bekannt



Vereinfachung: Populationsstreuungen bekannt

- Nehmen wir an, der Standardfehler der Mittelwertdifferenz von Psychologen- und Medizinernasenlängen ist als $se = 0.1$ bekannt.
- Mit dieser Information können wir unseren ersten p-Wert berechnen!
- Der p-Wert ist im Beispiel die Wahrscheinlichkeit der gemessenen Nasenlängendifferenz $\Delta\bar{x}$ oder eines noch größeren Effektes unter der Nullhypotesenverteilung.
- Dazu muss die Fläche unter der Nullhypotesenverteilung rechts von $\hat{\theta} = \Delta\bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}}$ berechnet werden.

Theoretische Nullhypotesenverteilung falls Streuung bekannt



Vereinfachung: Populationsstreuungen bekannt

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

$$\begin{aligned} p &= \int_{\hat{\theta}}^{\infty} f(x)dx = 1 - F(\hat{\theta}) = \\ &= 1 - F(0.2) \stackrel{(Computer)}{\approx} 0.023 \end{aligned}$$

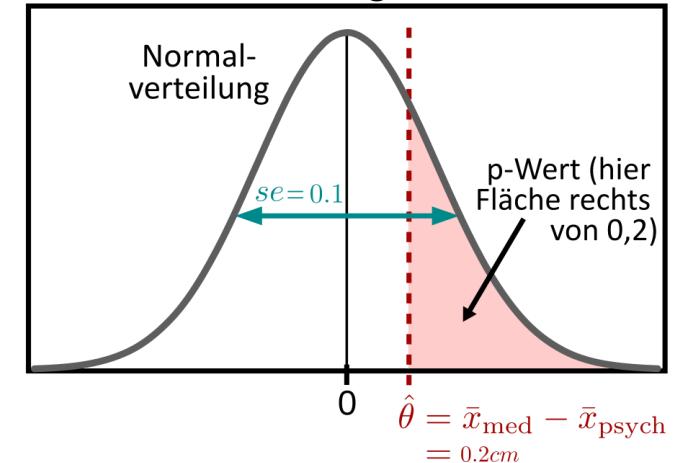
wobei $f(x)$ bzw. $F(x)$ hier die Dichte bzw. Verteilungsfunktion der gegebenen Normalverteilung $\mathcal{N}(0, 0.1)$ ist.

$F(x)$ 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 $\hat{\theta} = \Delta\bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}} = 0.2$.

Der p-Wert ist 0.023.

In Worten können wir feststellen: die Wahrscheinlichkeit, dass unser Effekt $\hat{\theta} = \Delta\bar{x}$ durch Zufall entstanden ist – also unter der Annahme, dass die Nullhypothese gilt – beträgt (nur) 2.3%.

Theoretische Nullhypotesenverteilung falls Streuung bekannt



z-Wert – eine standardisierte Prüfgröße

- 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.

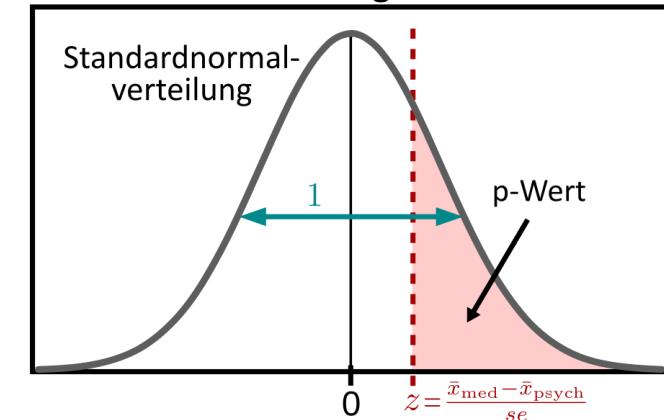
Abhilfe schafft die **Standardisierung des Effektes**:

$$z = \frac{\hat{\theta}}{se}$$

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

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

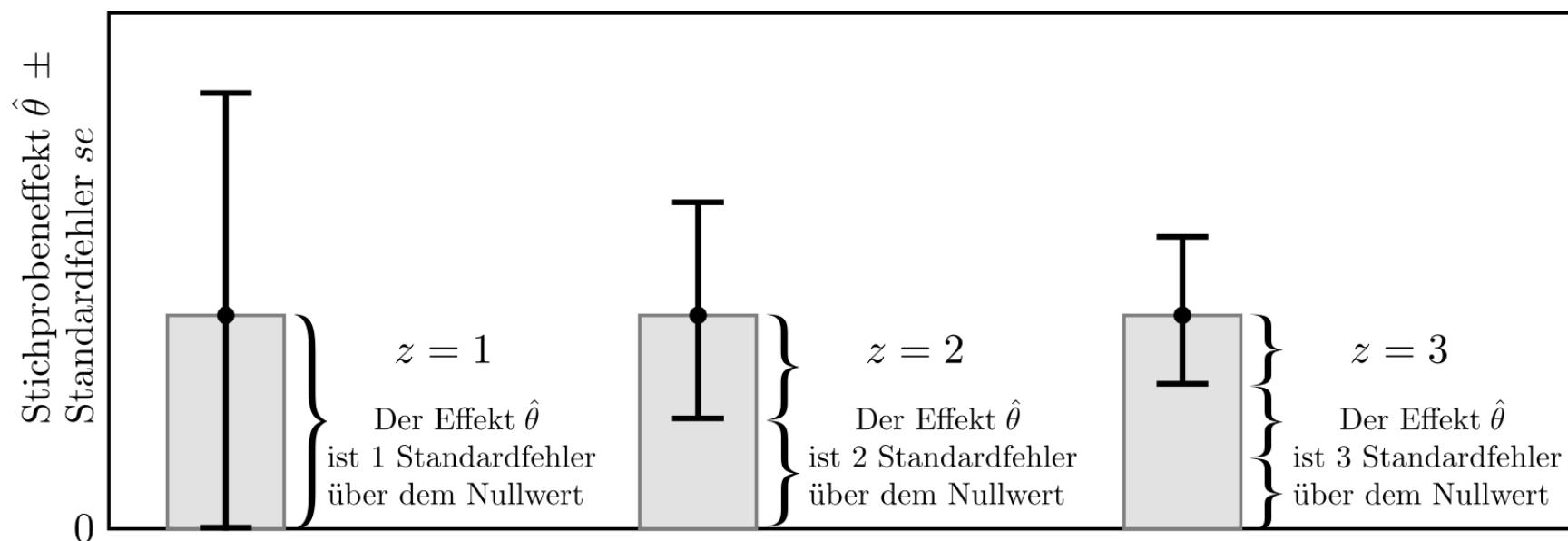
Theoretische Nullhypotesenverteilung falls Streuung bekannt



z-Wert – eine standardisierte Prüfgröße

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

- Der resultierende z-Wert ist vergleichbar zwischen Studien, im Gegensatz zu unstandardisierten Effekten $\hat{\theta}$ wie dem Mittelwertsunterschied $\Delta\bar{x}$, die von der spezifischen Skala abhängen.
- Der z-Wert hat eine (halbwegs) intuitive Interpretation: er gibt an, *wie viele Standardfehler der Effekt von einem Nulleffekt entfernt ist*.



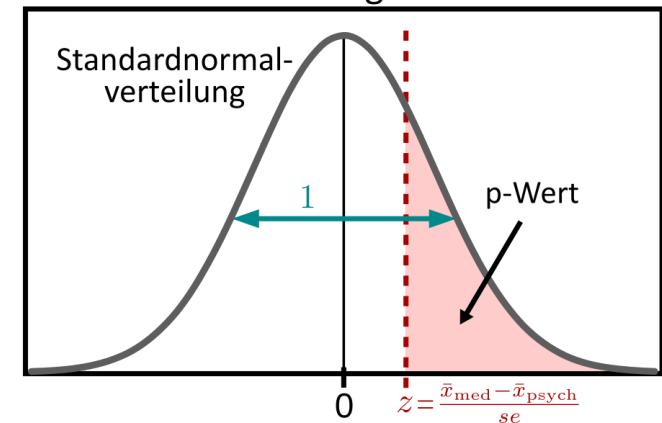
z-Test

Mit dem z-Wert als Prüfgröße benötigten wir für alle Tests dieser Art nur noch eine einzige Verteilung – die Normalverteilung mit Mittelwert 0 und Streuung 1, also die **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 $z = \frac{\hat{\theta}}{se} = \frac{\Delta\bar{x}}{se} = \frac{\bar{x}_{med} - \bar{x}_{psych}}{se}$ definiert ist. Der resultierende p-Wert ist derselbe.

Theoretische Nullhypotesenverteilung
falls Streuung bekannt



Der **z-Test** ist ein statistischer Nullhypotesentest, der auf Basis der standardisierten Prüfgröße $z = \frac{\hat{\theta}}{se}$ durchgeführt wird.

z-Test

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

Wir bestimmen die standardisierte Prüfgröße z :

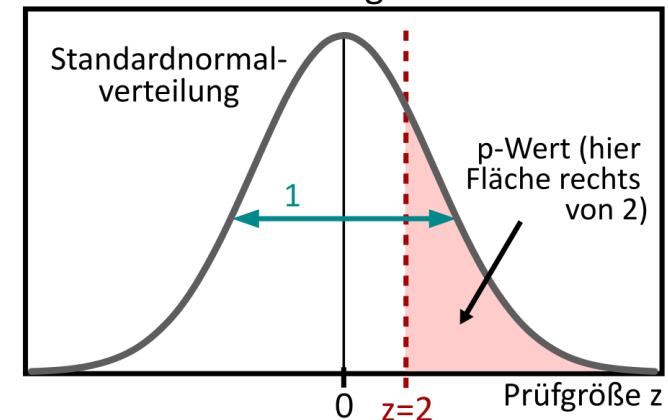
$$z = \frac{\hat{\theta}}{se} = \frac{\Delta\bar{x}}{se} = \frac{0.2}{0.1} = 2$$

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 Streuung bekannt



z-Test: Cheat Sheet für Mittelwertdifferenzen

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

Fall	z-Wert	Bekannt	Standardfehler
Einzelmessung: Vergleich von \bar{x} mit einem Referenzwert μ_0	$z = \frac{\bar{x} - \mu_0}{se}$	σ	$se = \frac{\sigma}{\sqrt{n}}$
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, ρ	$se = \frac{\sigma_\Delta}{\sqrt{n}} \text{ mit}$ $\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \rho \sigma_A \sigma_B}$
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	$se = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$



Die Formeln für die Standardfehler gelten für den Fall, dass die **Varianzen σ_A^2 bzw. σ_B^2 in den Populationen bekannt sind** (wie beim z-Test). Siehe Bonuscontent für eine **Herleitung der Standardfehler von Mittelwertdifferenzen**.

Sind die Populationsstreuungen nicht bekannt, sehen die Standardfehlerformeln aufgrund der Besselkorrektur etwas anders aus (→ siehe t-Test in Vorlesung 11).

Das ABC der Hypothesentestung

- Einstichprobentest und Zweistichprobentest
- Signifikanzniveau
- Ein- und zweiseitiges Testen

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 (μ_{IQ} = mittler IQ in der Population)
 - Nullhypothese: $\mu_{IQ} \leq 100$ Alternativhypothese: $\mu_{IQ} > 100$
 - Beachte: der “Nullwert” μ_0 ist in diesem Fall 100 und nicht 0!
- Abhängige Messungen: Psychologiestudierende sind morgens intelligenter als abends ($\Delta\mu_{IQ} = \mu_{IQ, Morgen} - \mu_{IQ, Abend}$)
 - Nullhypothese: $\Delta\mu_{IQ} \leq 0$ Alternativhypothese: $\Delta\mu_{IQ} > 100$

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 ($\Delta\mu_{IQ} = \mu_{IQ, HMU} - \mu_{IQ, MSB}$)
 - Nullhypothese: $\Delta\mu_{IQ} \leq 0$ Alternativhypothese: $\Delta\mu_{IQ} > 100$

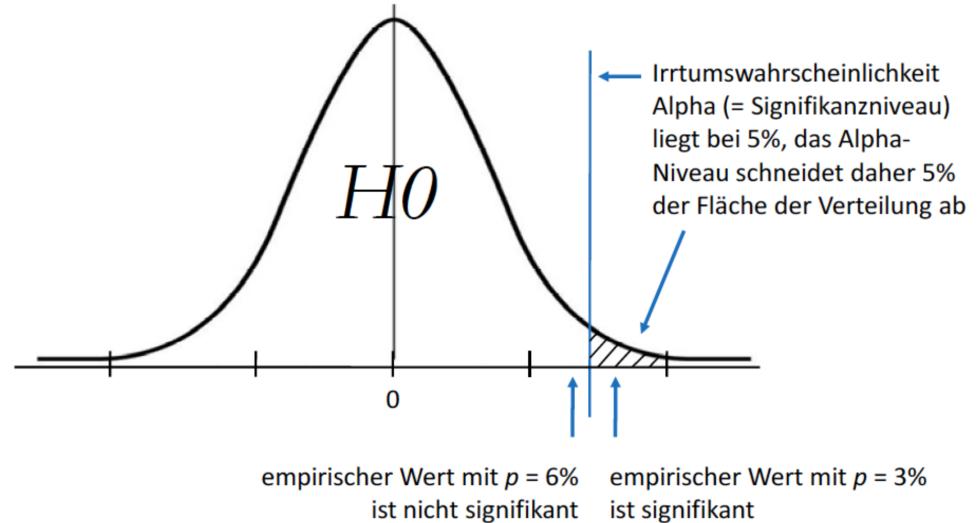
Signifikanzniveau ($p < \alpha$?)

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

Bislang haben wir konkrete **gerichtete Hypothese** betrachtet, etwa indem bei der Nasenlängenhypothese festgelegt wurde, welche der beiden Gruppen im Mittel eine längere Nase hat. Es gibt jedoch insgesamt **drei Fälle**.

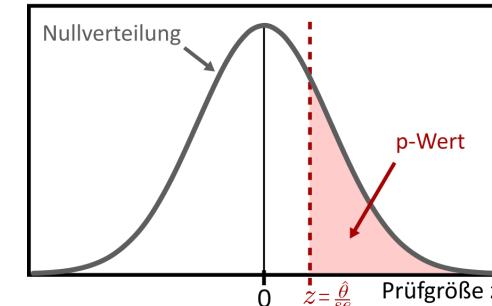
Rechtsseitige gerichtete Hypothese:
gibt Richtung vor

Hypothese: $\theta > 0$

Beispiel: Medizinernasen sind länger als Psychologennasen

$$\Delta\mu > 0 \quad \text{mit} \quad \theta = \Delta\mu = \mu_{\text{med}} - \mu_{\text{psych}}$$

Einseitiger Test rechts ($\theta > 0$)



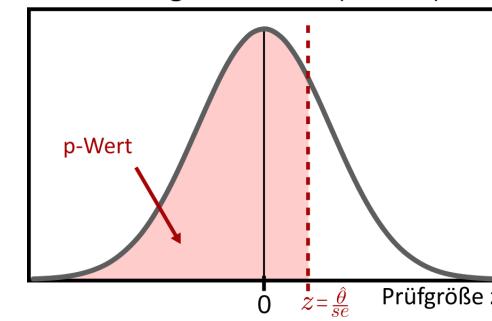
Linksseitige gerichtete Hypothese:
gibt Richtung vor

Hypothese: $\theta < 0$

Beispiel: Medizinernasen sind kürzer als Psychologennasen

$$\Delta\mu < 0 \quad \text{mit} \quad \theta = \Delta\mu = \mu_{\text{med}} - \mu_{\text{psych}}$$

Einseitiger Test links ($\theta < 0$)



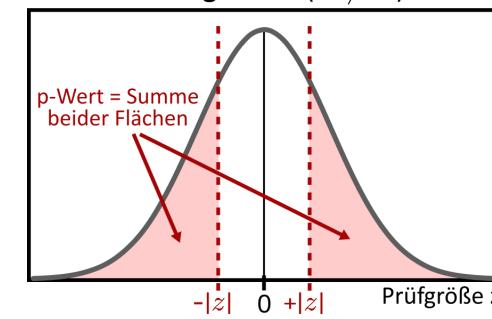
Ungerichtete Hypothese:
gibt keine Richtung vor

Hypothese: $\theta \neq 0$

Beispiel: Medzinernasen und Psychologennasen sind unterschiedlich lang

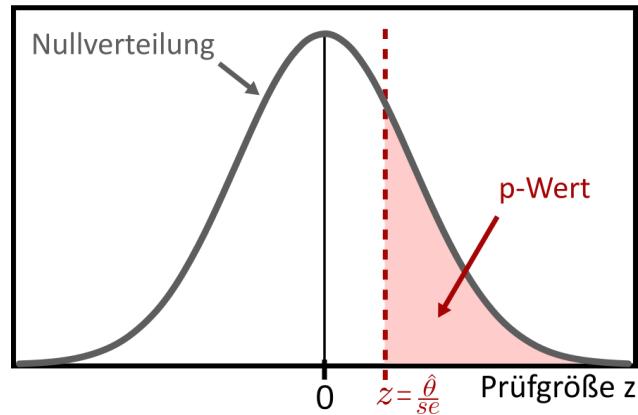
$$\Delta\mu \neq 0 \quad \text{mit} \quad \theta = \Delta\mu = \mu_{\text{med}} - \mu_{\text{psych}}$$

Zweiseitiger Test ($\theta \neq 0$)

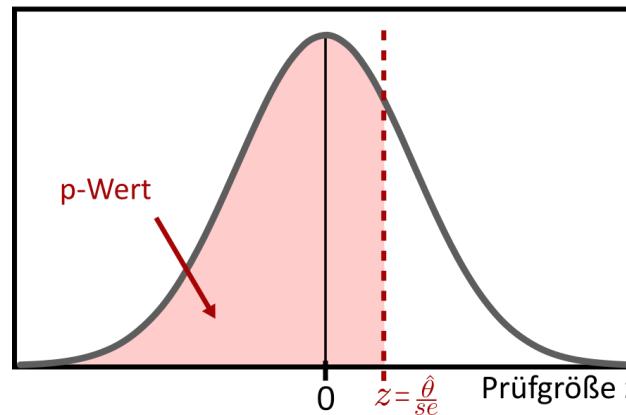


Ein- und zweiseitiges Testen

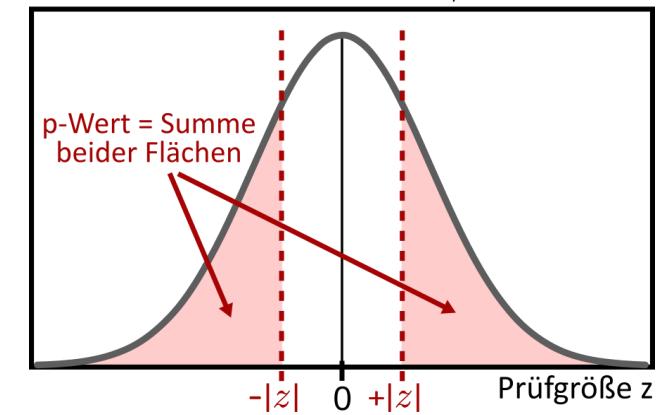
Einseitiger Test rechts ($\theta > 0$)



Einseitiger Test links ($\theta < 0$)



Zweiseitiger Test ($\theta \neq 0$)



Wie sehen folgende Zusammenhänge:

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

$$p(\theta > 0) = 1 - p(\theta < 0)$$

Beispiel: $p(\mu_{\text{med}} > \mu_{\text{psych}}) = 1 - p(\mu_{\text{med}} < \mu_{\text{psych}})$

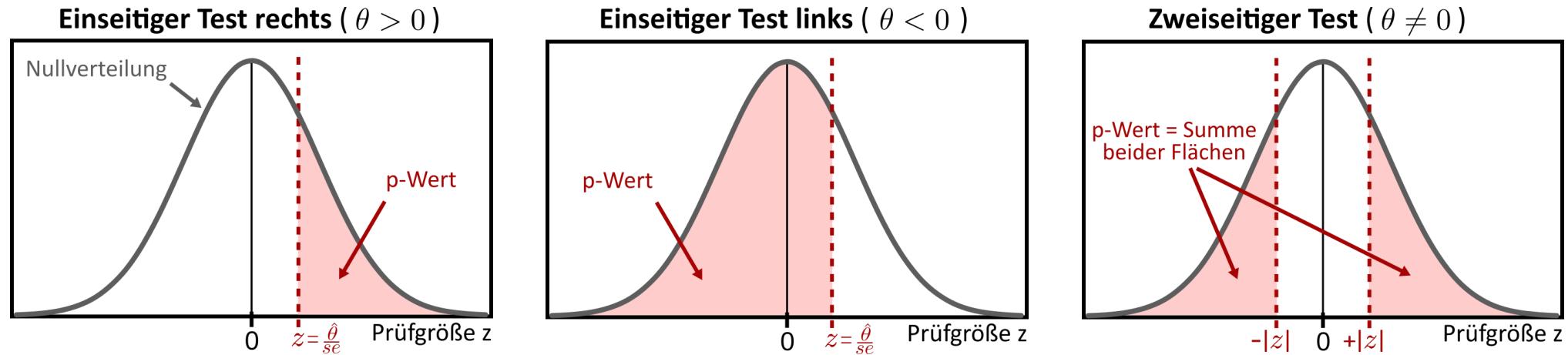
- Der p-Wert der ungerichteten Hypothese ("unterschiedlich") ist gleich **2 x der kleinere p-Wert der beiden gerichteten Hypothesen**:

$$p(\theta \neq 0) = 2 \times \min(p(\theta > 0), p(\theta < 0))$$

Beispiel:

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

Ein- und zweiseitiges Testen



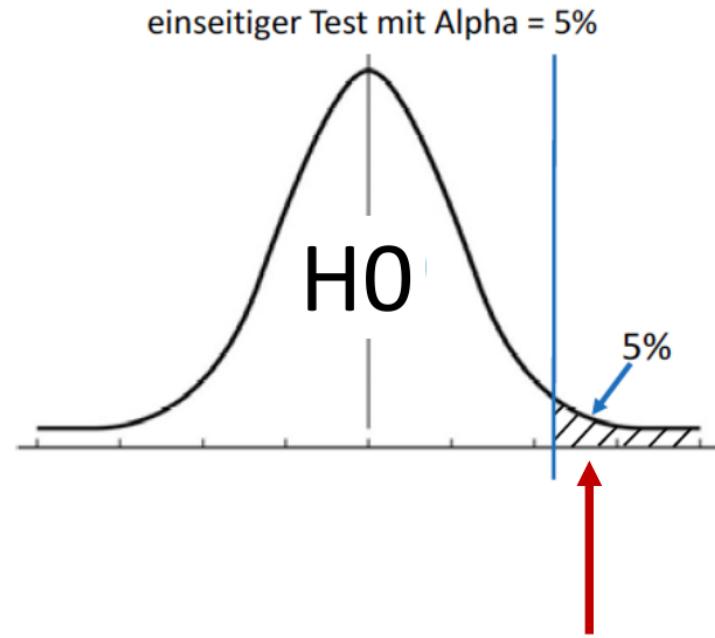
Der p-Wert der ungerichteten Hypothese ist damit immer größer (doppelt so groß) wie der kleinere p-Wert der gerichteten Hypothesen.

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 Vorhersagen) also “einfacher”.

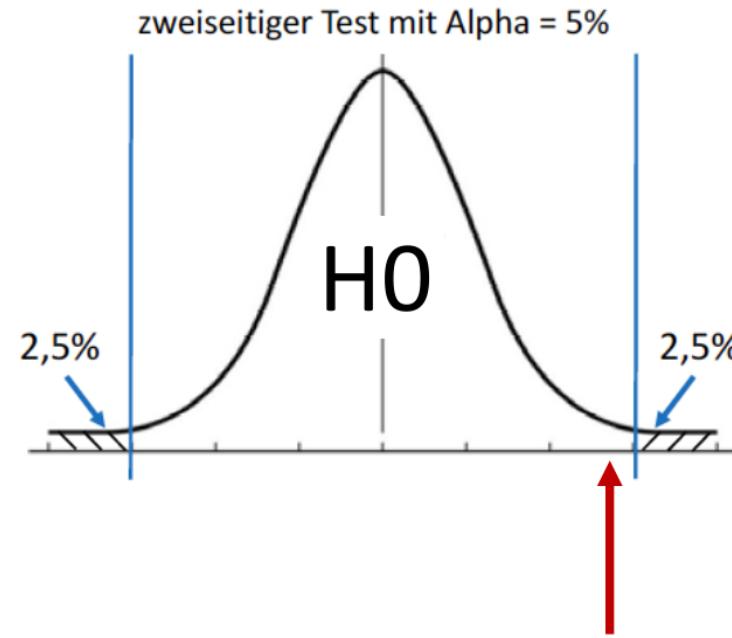
Genauer gesagt verdoppeln wir unsere Chance, die Nullhypothese abzulehnen, da unter der Annahme der Nullhypothese jeglicher beobachteter z-Wert mit gleichen Flächen links von $-|z|$ und rechts von $+|z|$ verbunden ist. Dies wird durch die Verdopplung des p-Wertes der gerichteten Hypothese kompensiert – man könnte auch sagen “die Verdopplung des p-Wertes bestraft unsere schwammige ungerichtete Vorhersage”.

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.

Eigenschaften und Verhalten des p-Wertes

p-Wert versus Stichprobengröße & Populationsstreuung

Wir werten einen Effekt als **signifikant**, wenn der p-Wert kleiner dem Signifkanzniveau α ist.

Es gilt: je kleiner der p-Wert, desto...

... weniger kompatibel ist unser Stichprobeneffekt $\hat{\theta}$ mit der Nullhypothese.

... schärfer lehnen wir also die Nullhypothese H_0 ab.

... stärker ist die Evidenz für unsere Alternativhypothese H_1 .

... statistisch bedeutsamter ist unser Stichprobeneffekt $\hat{\theta}$.

p-Wert versus Stichprobengröße & Populationsstreuung

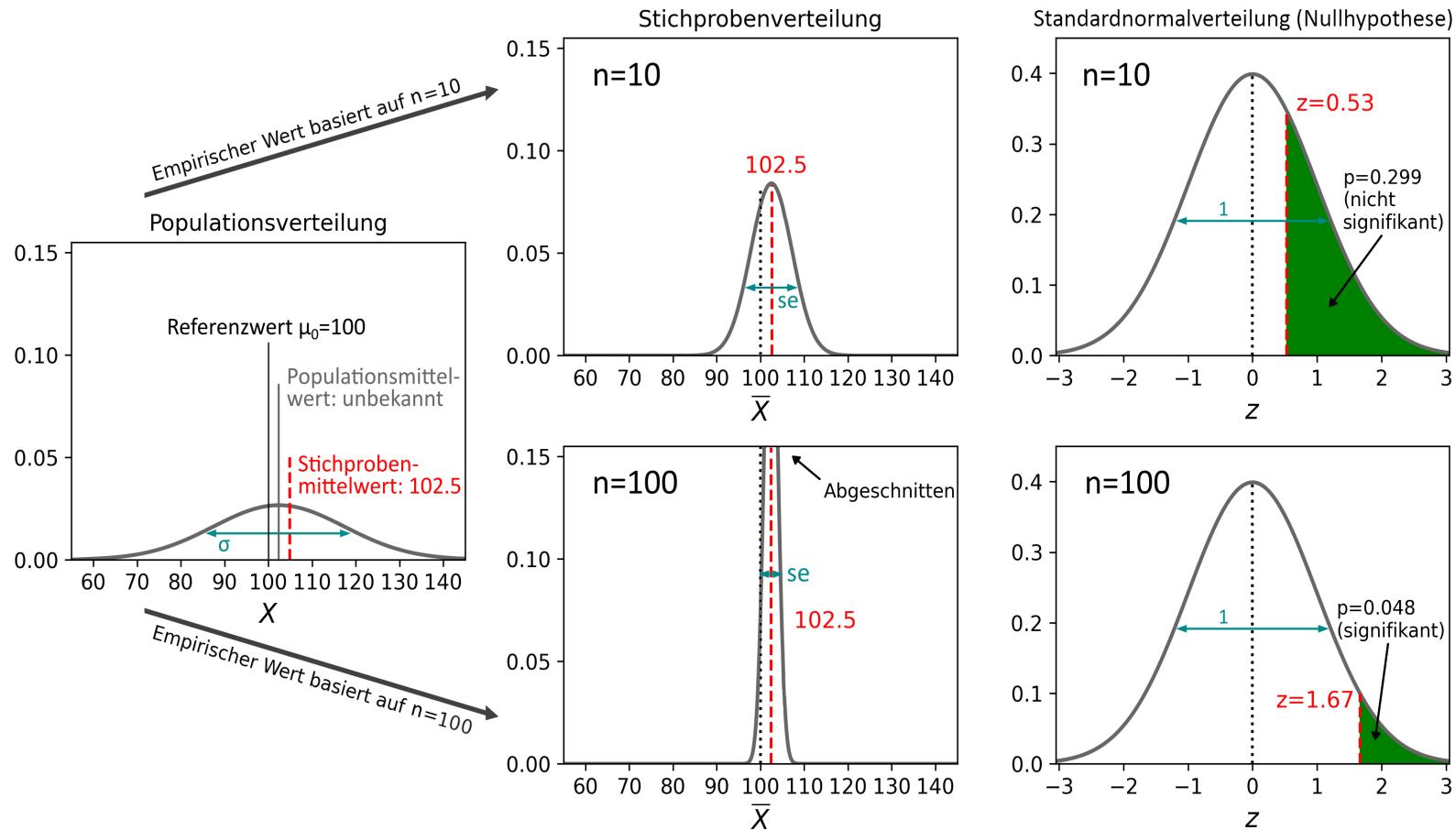
Gibt es den hypothesierten Effekt in der Population tatsächlich (Alternativhypothese wahr), gilt:

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

Gibt es den hypothesierten 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]).

p-Wert versus Stichprobengröße: Beispiel



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.

[Zusammenfassung]

- Bei der **Nullhypotesentestung** nach Fisher wird die Wahrscheinlichkeit geprüft, mit der ein beobachteter Effekt $\hat{\theta}$ (oder ein noch extremerer Effekt) durch Zufall entstanden sein könnte – d.h. für die Annahme dass die *Nullhypothese tatsächlich zutreffend* ist.
- Diese Wahrscheinlichkeit wird als **p-Wert** bezeichnet – je kleiner der p-Wert, desto weniger haltbar ist die Nullhypothese, desto stärker die **statistische Signifikanz**.
- Ein Effekt wird als statistisch signifikant gewertet, wenn der p-Wert unterhalb eines festgelegten **Signifikanzniveaus α** liegt (in diesem Fall wird die “Nullhypothese abgelehnt”).
- Sind die Populationsstreuungen bekannt, so erfolgt die statistische Testung anhand der **Prüfgröße z (z-Wert)** in Verbindung mit der **Standardnormalverteilung**.
- Abhängig davon ob die untersuchte Hypothese **gerichtet** oder **ungerichtet** ist, erfolgt die statistische Testung entweder als **einseitiger Test** oder als **zweiseitiger Test**.



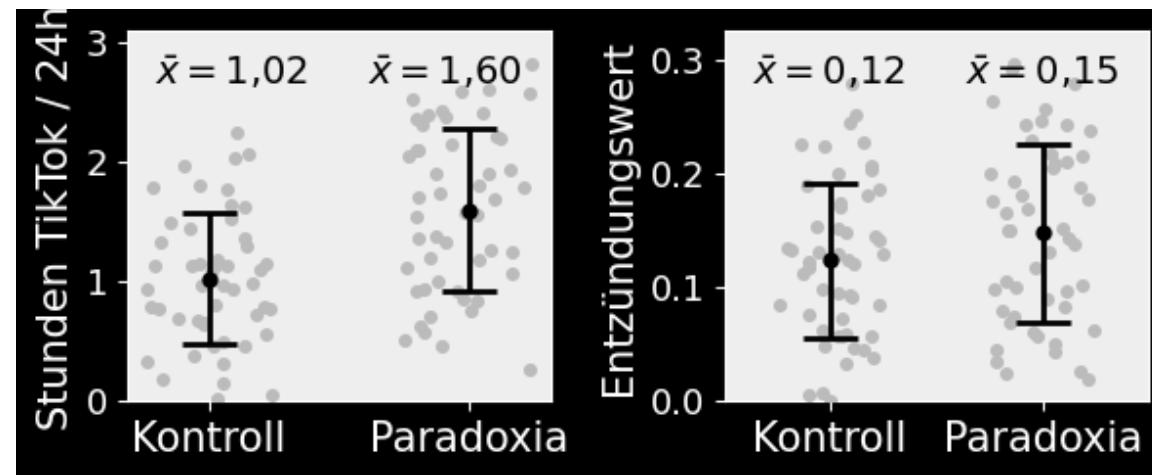
Nach kurzer Beratung in der Task Force ist Ihnen klar, dass die Voraussetzungen für einen z-Test nicht gegeben sind, da die Populationsstreuungen σ nicht bekannt sind, sondern als $\hat{\sigma}$ geschätzt wurden. Aus Neugier rechnen Sie dennoch z-Tests für die Gruppenvergleiche und nehmen dafür an: $\sigma = \hat{\sigma}$.

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

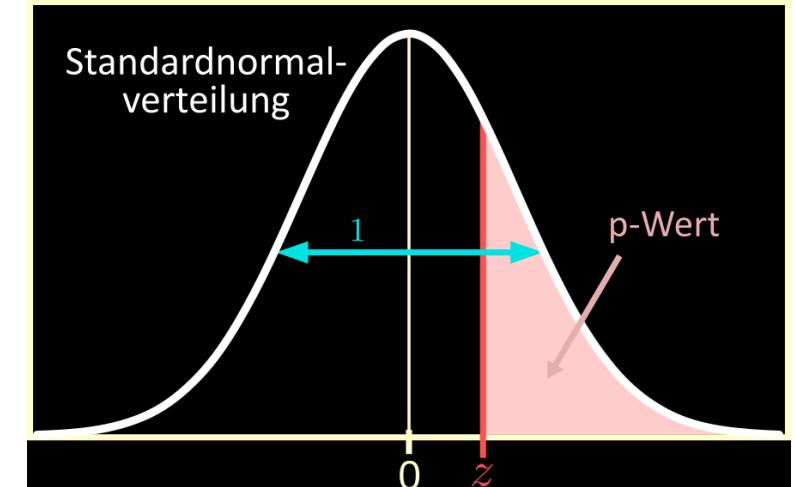
Der z-Wert ergibt sich als das Verhältnis aus dem “Effekt” (hier $\Delta\bar{x}$) und dem Standardfehler:

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

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

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:

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



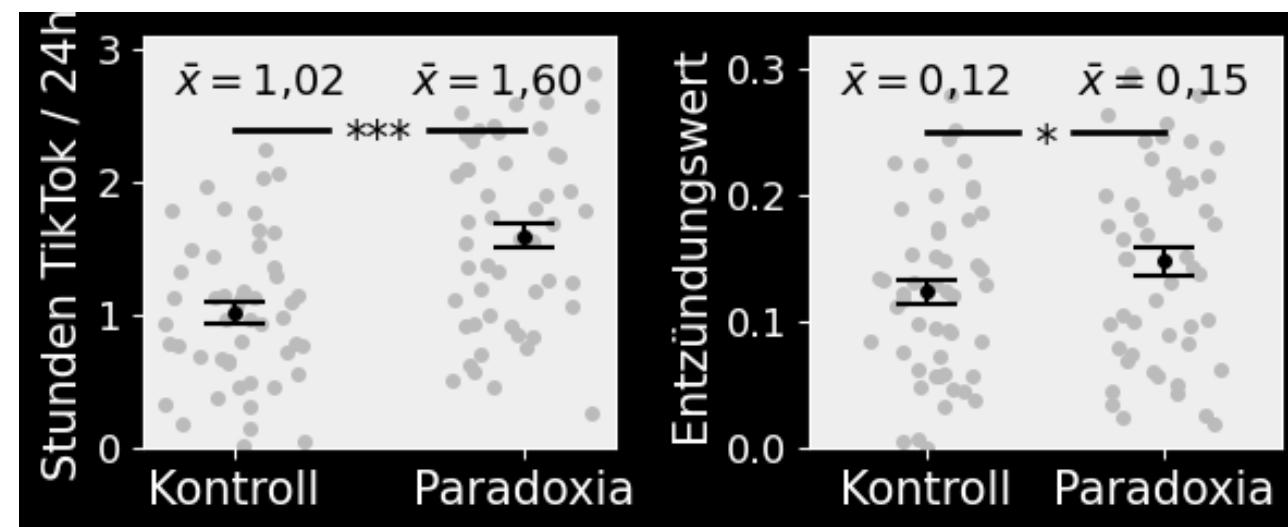
$$\text{Entzündung: } p = 1 - \Phi(1,67) \stackrel{\text{(Computer)}}{=} 0,049$$

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.



Bonuscontent

z-Test: Herleitung der Standardfehler von Mittelwertdifferenzen

Sind die Populationsstreuungen σ_A und σ_B bekannt (wie beim z-Test vorausgesetzt), lässt sich der Standardfehler se mit den Formeln aus dem [Cheatsheet](#) berechnen. Hier betrachten wir die Herleitung dieser Formeln.

Abhängige Messungen in einer Stichprobe

Werden die Messungen X_A und X_B in denselben Versuchspersonen durchgeführt, können wir – wie schon bei der Effektstärke – die Standardabweichung σ_Δ der Differenzenvariable $\Delta X = X_A - X_B$ bestimmen:

$$\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

Da σ_Δ in diesem Fall bereits die Standardabweichung des Effektes ist, folgt direkt der Standardfehler:

$$se = \frac{\sigma_\Delta}{\sqrt{n}} \quad \text{mit} \quad \sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B}$$

Wichtig: bei abhängigen Messungen ist, zusätzlich zu den Populationsstreuungen σ_A/σ_B , auch Kenntnis über die Populationskorrelation ρ der beiden Variablen X_A und X_B notwendig.



z-Test: Herleitung der Standardfehler von Mittelwertdifferenzen

Unabhängige Messungen in zwei Stichproben

- Bei unabhängigen Gruppen sind die Mittelwerte \bar{x}_A und \bar{x}_B der beiden Stichproben unabhängige Zufallsvariablen.
- Für die Mittelwertdifferenz $\Delta\bar{x} = \bar{x}_A - \bar{x}_B$ gilt daher die allgemeine Varianzsummenformel, der zufolge sich bei der Differenzbildung zweier unabhängiger Zufallsvariablen die Varianzen addieren.
- Die Varianz von Mittelwerten ist nichts anderes als der quadrierte Standardfehler, hier se_A^2 und se_B^2 .
- Für die Varianz se^2 der Mittelwertdifferenz $\Delta\bar{x}$ gilt also:

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

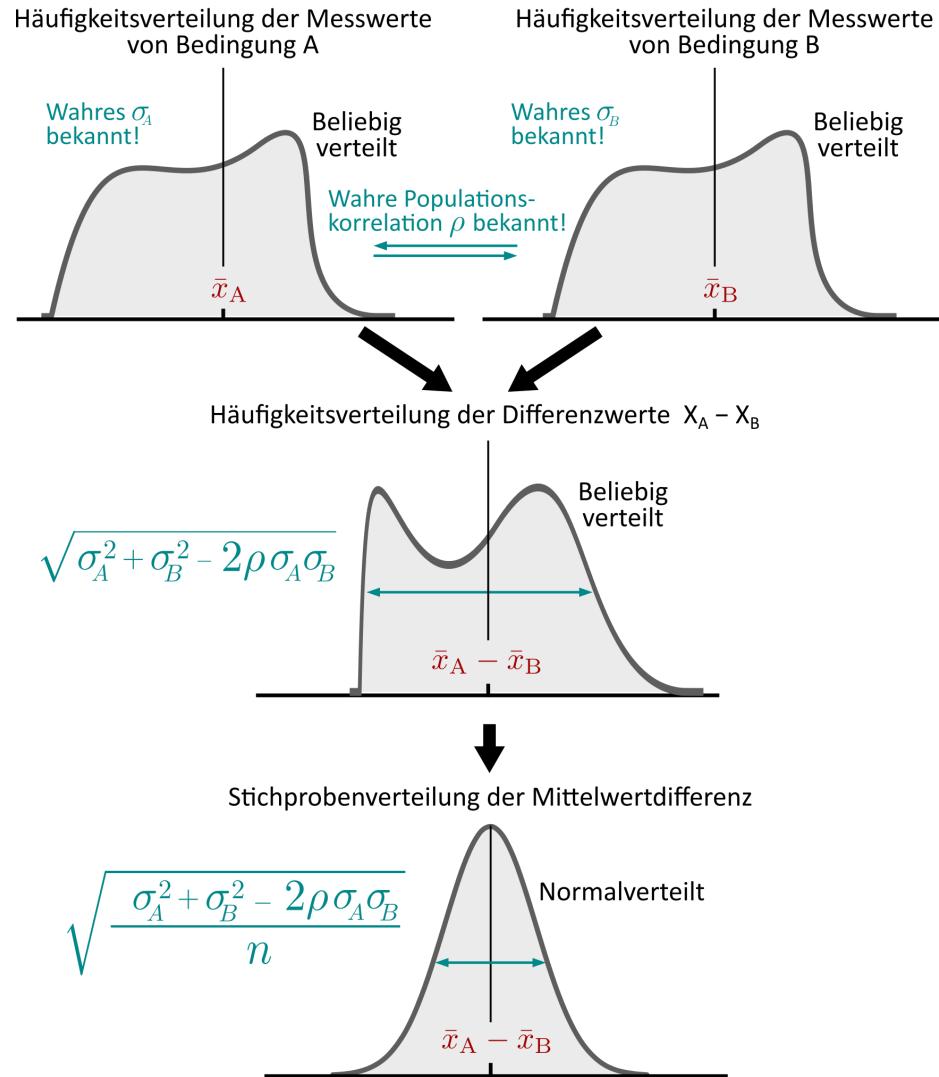
Wurzel ziehen:

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

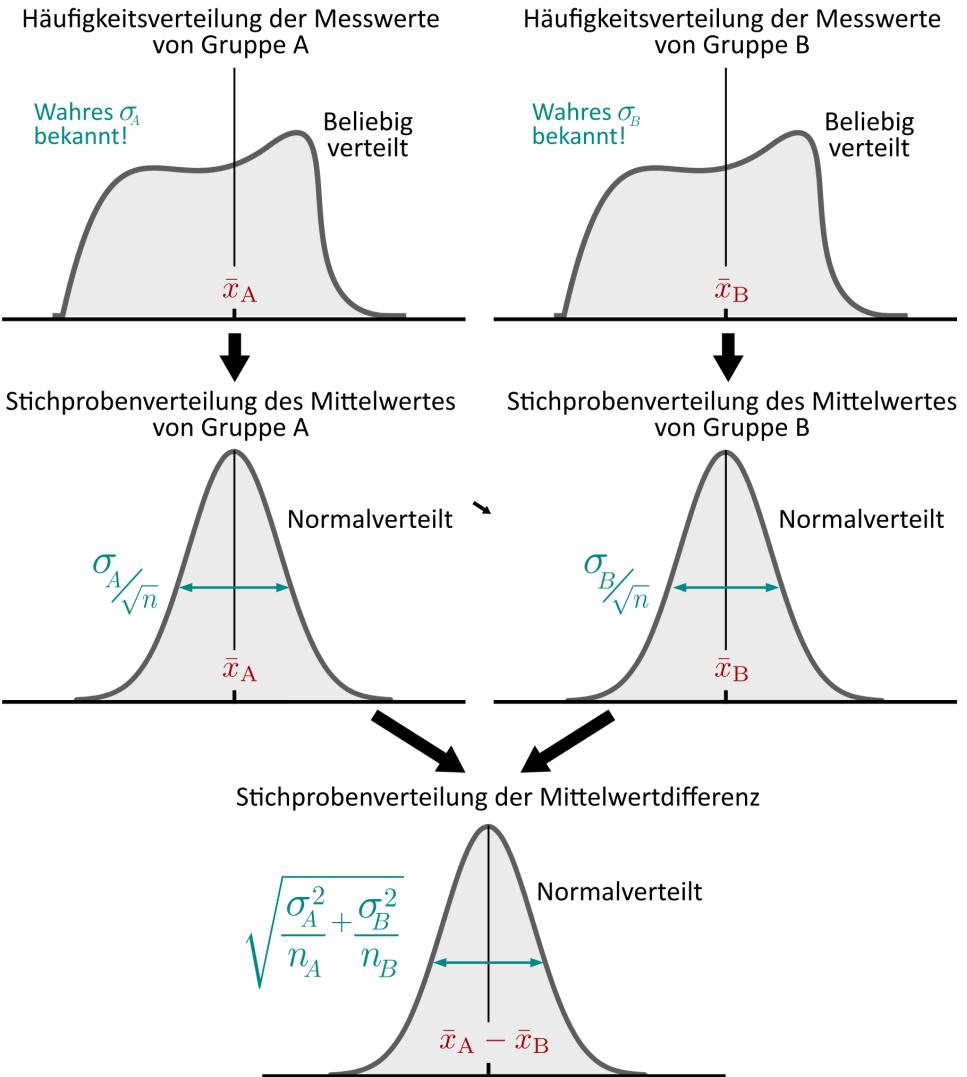


z-Test: Veranschaulichung der Standardfehler von Mittelwertdifferenzen

Mittelwertdifferenz: Abhängige Messungen



Mittelwertdifferenz: Unabhängige Messungen



Fußnoten

1. <https://www.majordifferences.com/2016/10/5-differences-between-null-and.html>