

M24 Statistik 1: Sommersemester 2024

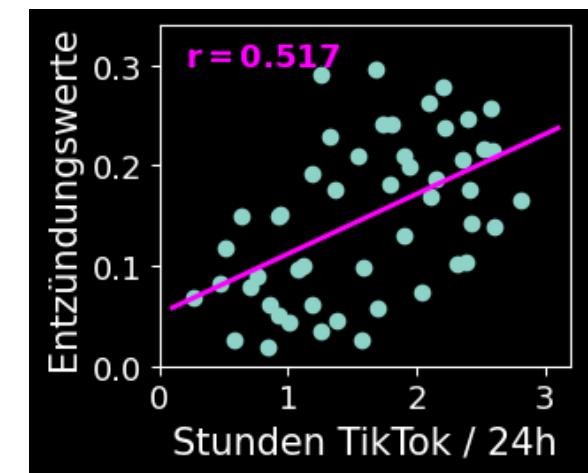
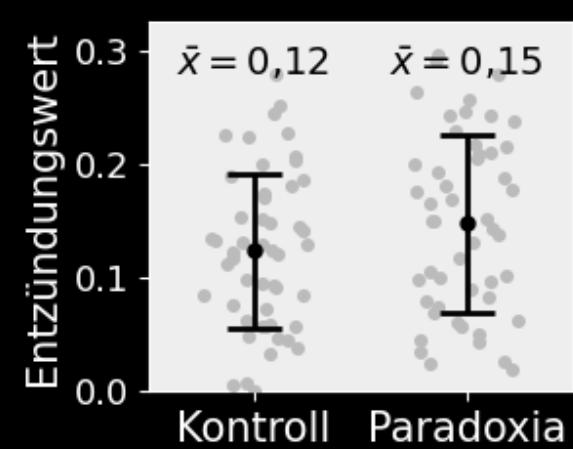
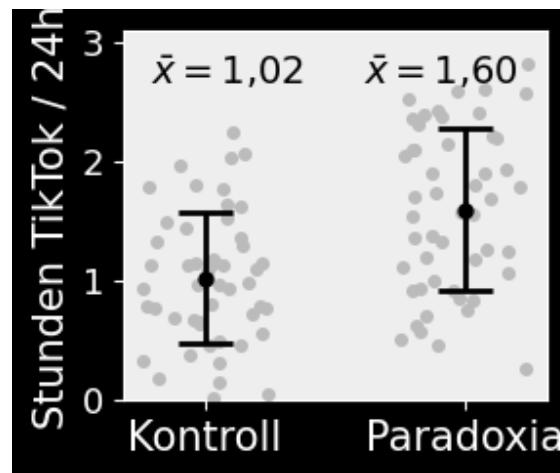
Vorlesung 10: Signifikanztestung

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?

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)

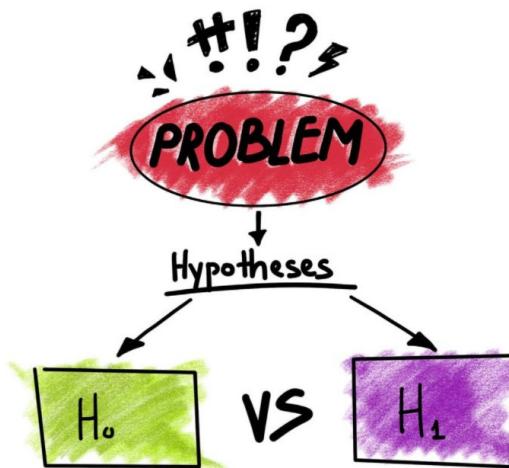
Besagt, dass ein bestimmter Effekt *nicht vorliegt*.

- 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.

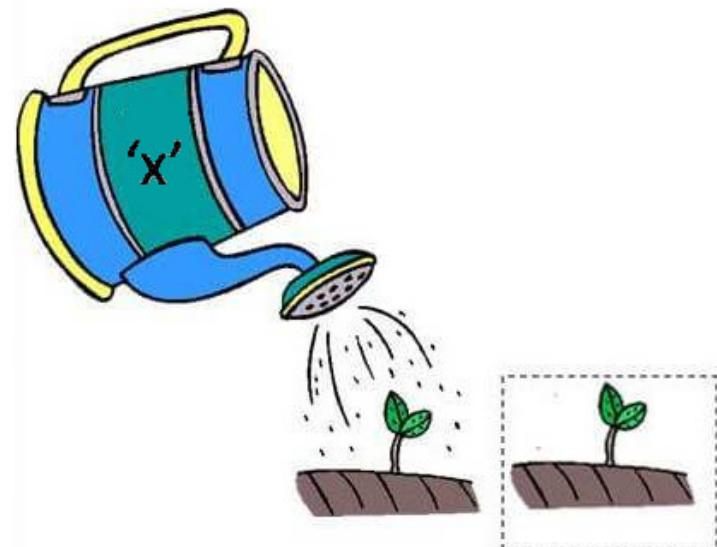


Nullhypothese und Alternativhypothese: Beispiel

Effect of Bio-fertilizer 'x' on Plant growth

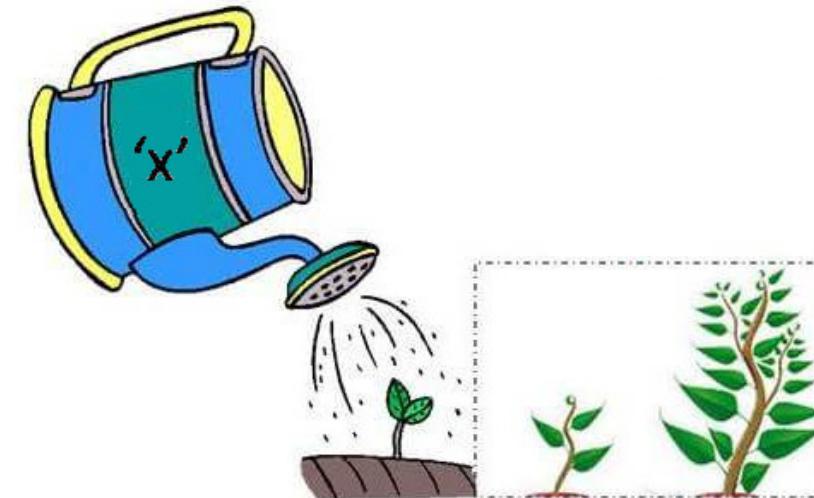
Null Hypothesis

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)



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

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

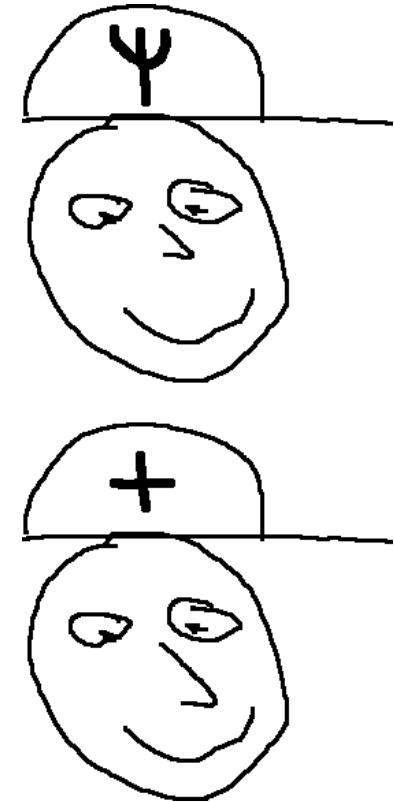
Nach Fisher stellen wir nun folgende präzise Frage:

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

In kurz:

Angenommen H_0 trifft zu und es gilt $\mu_{\text{med}} = \mu_{\text{psych}}$, 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*).

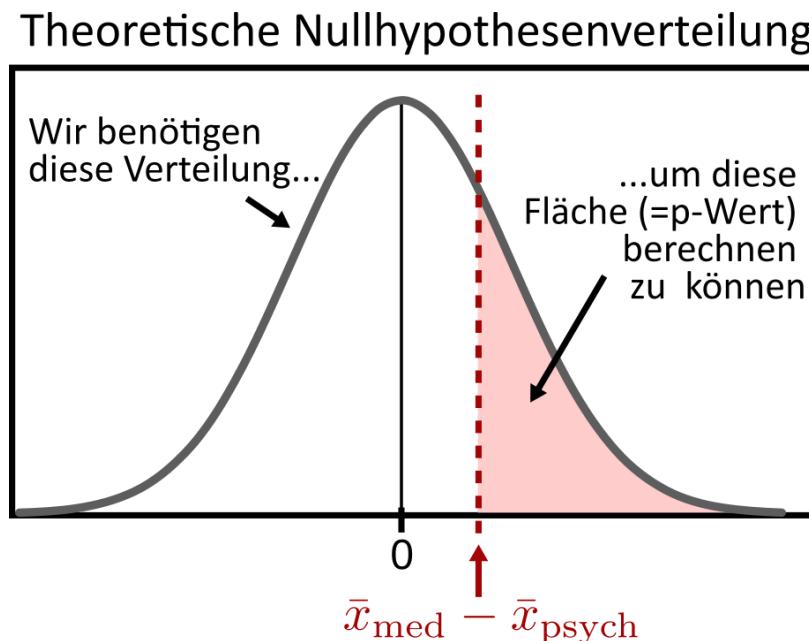


p-Wert und Konstruktion der Nullhypothese

Wir definieren:

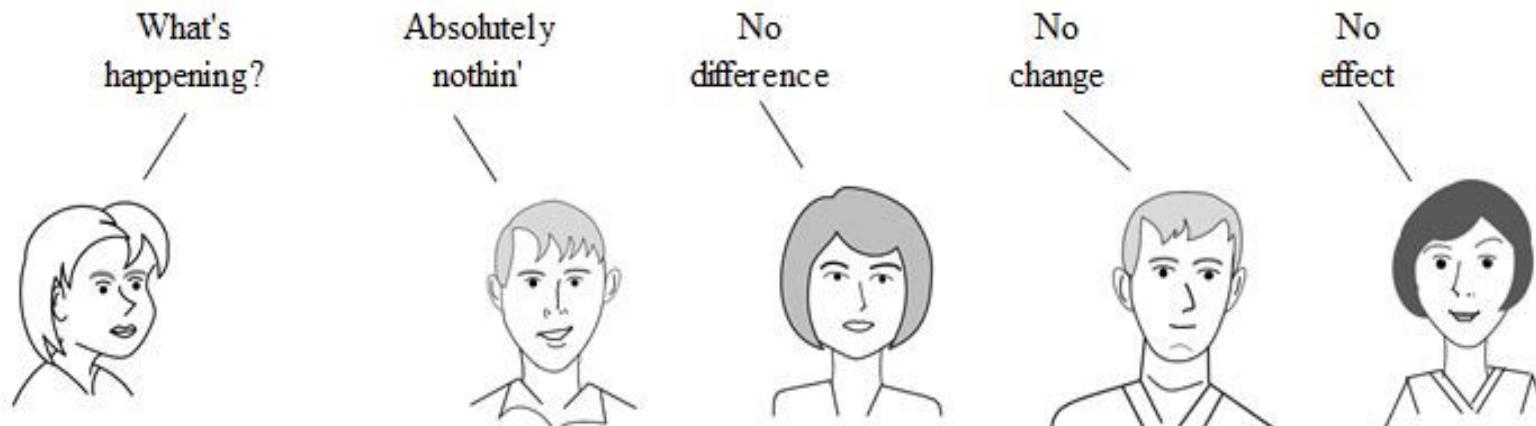
Definition Der p-Wert ist die Wahrscheinlichkeit den Stichprobeneffekt oder einen noch extremeren Effekt zu erhalten, obwohl für die Population die Nullhypothese angenommen wird.

Um diese Wahrscheinlichkeit zu berechnen, benötigen wir die Wahrscheinlichkeits(dichte)verteilung aller theoretisch möglicher Stichprobeneffekte *unter Annahme der Nullhypothese*. Wir bezeichnen sie als **Theoretische Nullhypotesenverteilung** oder kurz **Nullhypotesenverteilung**.



Konstruktion der Nullhypothese

- Das Ziel “*die WahrscheinlichkeitsdichteVerteilung 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**.
- Beachte: nicht nur die Stichprobenverteilung von Mittelwerten ist normalverteilt ist, sondern auch die Stichprobenverteilung von Mittelwertsunterschieden (→ die Differenz zweier normalverteilter Variablen ist ebenfalls normalverteilt).

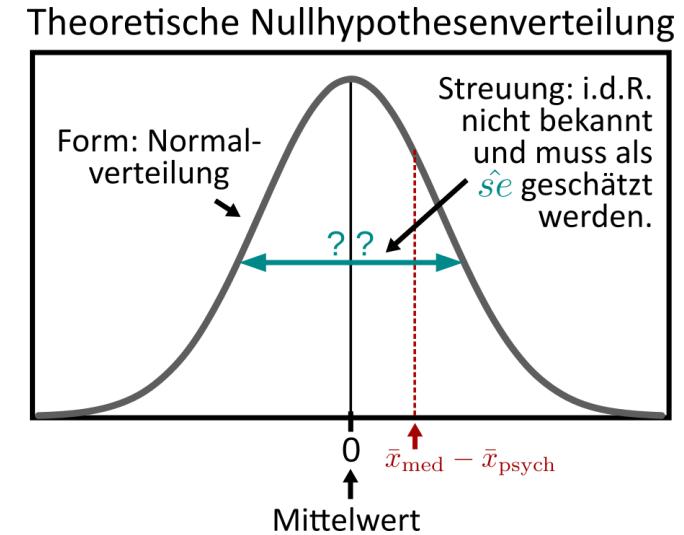


Bildnachweis²

Konstruktion der Nullhypothese

Da die Nullhypotesenverteilung ihrer Natur nach eine normalverteilte Stichprobenverteilung ist (unter der Annahme der Nullhypothese), benötigen wir wie schon bei der Konstruktion der Stichprobenverteilung zwei Informationen: Mittelwert und Streuung der Normalverteilung.

- Der **Mittelwert** der Nullhypotesenverteilung entspricht einem Nulleffekt (i.d.R. der Wert 0).
- Die **Streuung** entspricht wie bei der Stichprobenverteilung einem Standardfehler se – hier dem Standardfehler des Effektes.
 - Problem: se ist in der Regel unbekannt und muss als \hat{se} auf Basis der Stichprobe geschätzt werden.



$$f(x) = \frac{1}{\hat{se}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{\hat{se}}\right)^2}$$

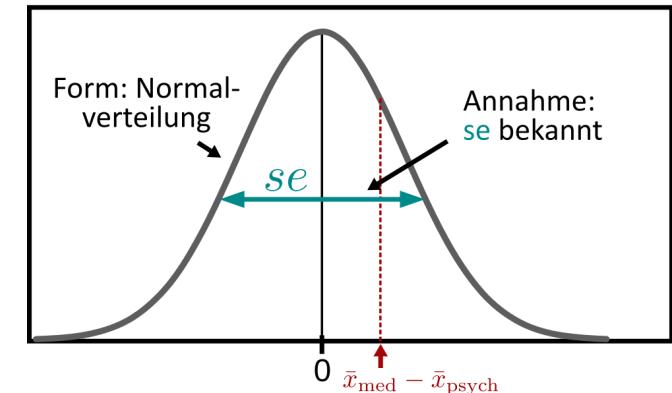
Formel für die Dichtefunktion der Normalverteilung mit Mittelwert 0 und geschätzter Streuung \hat{se} .

Vereinfachung: Populationsstreuungen bekannt

- In einem ersten Schritt betrachten wir daher den **vereinfachten Fall, dass die Streuung se bekannt ist**. D.h. 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 (Herleitung in Kürze):

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

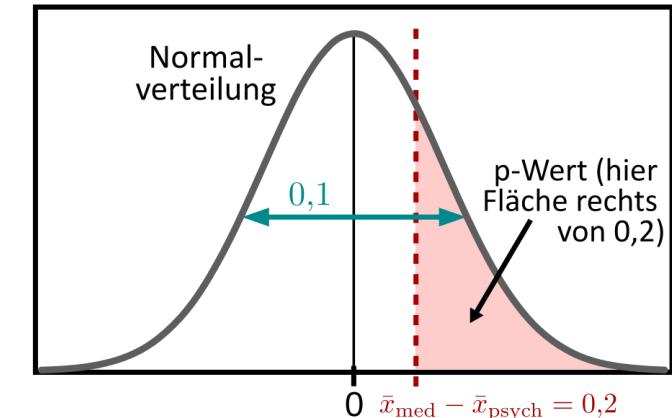
Theoretische Nullhypotesenverteilung falls σ bekannt



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, dass der Stichprobeneffekt (Nasenlängendifferenz) unseren gemessenen Wert $\Delta\bar{x}$ annimmt *oder einen noch größeren Effekt*.
- Dazu muss die Fläche unter der Nullhypothesenverteilung rechts von $\Delta\bar{x} = \bar{x}_{\text{med}} - \bar{x}_{\text{psych}}$ berechnet werden.

Theoretische Nullhypothesenverteilung falls σ bekannt



Vereinfachung: Populationsstreuungen bekannt

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

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

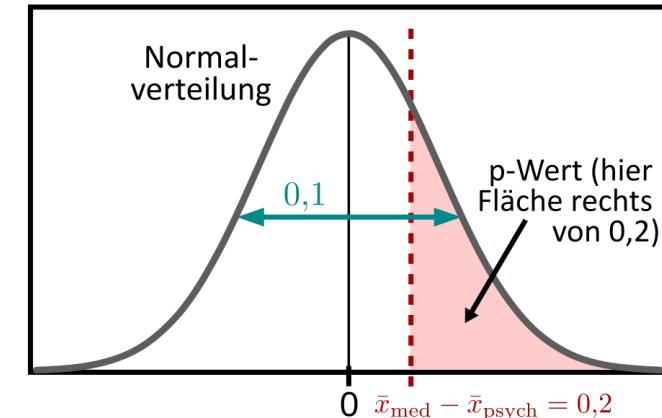
wobei $f(x)$ bzw. $F(x)$ hier die Dichte bzw. Verteilungsfunktion der gegebenen Normalverteilung $\mathcal{N}(0, 0.1)$ ist.

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

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



z-Test

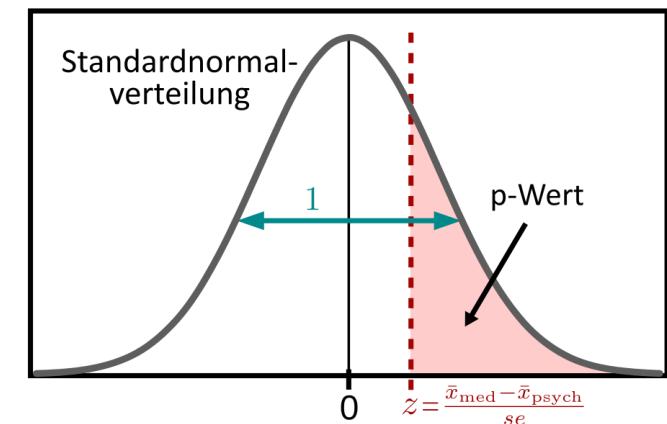
In der Praxis wird die Berechnung von p-Werten um das Konzept der **standardisierten Prüfgröße** erweitert.

- 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** darstellt.

z-Test

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

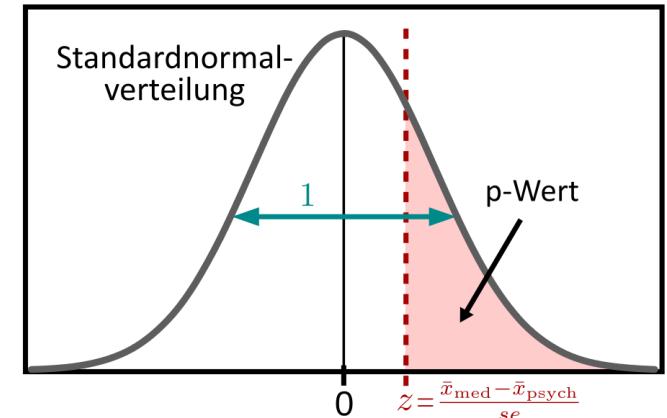
- Der resultierende z-Wert ist vergleichbar zwischen Studien, im Gegensatz zum Mittelwertsunterschied $\Delta\bar{x}$, der von der spezifischen Skala abhängt.
- Der z-Wert hat eine intuitive Interpretation: er gibt an, *wie viele Standardabweichungen der Effekt von einem Nulleffekt entfernt ist*.

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, 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 $z = \frac{\Delta\bar{x}}{se} = \frac{\bar{x}_{med} - \bar{x}_{psych}}{se}$ definiert ist. Der resultierende p-Wert ist derselbe.

Theoretische Nullhypotesenverteilung
falls σ bekannt



z-Test

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

Wir bestimmen die standardisierte Prüfgröße z :

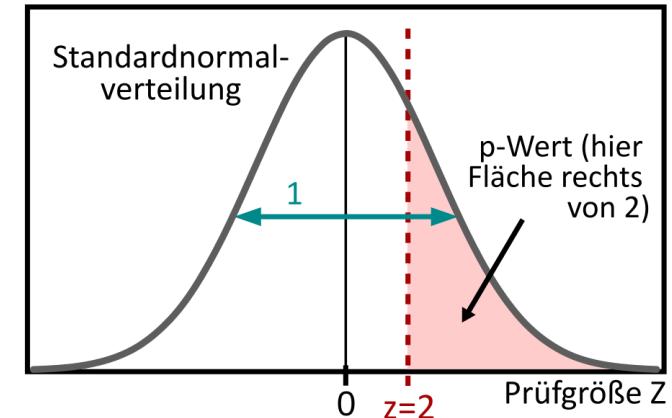
$$z = \frac{\Delta \bar{x}}{se} = \frac{0.2}{0.1} = 2$$

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 se
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	$z = \frac{\bar{x}_A - \bar{x}_B}{se} = \frac{\Delta\bar{x}}{se}$	$\sigma_A, \sigma_B, (\rho)$	$se = \frac{\sigma_{\Delta}}{\sqrt{n}} \text{ mit } \sigma_{\Delta} = \sqrt{\frac{1}{n} \sum (\Delta x_i - \Delta \bar{x})^2}$ oder $\sigma_{\Delta} = \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \rho \sigma_A \sigma_B}$
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	$se = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

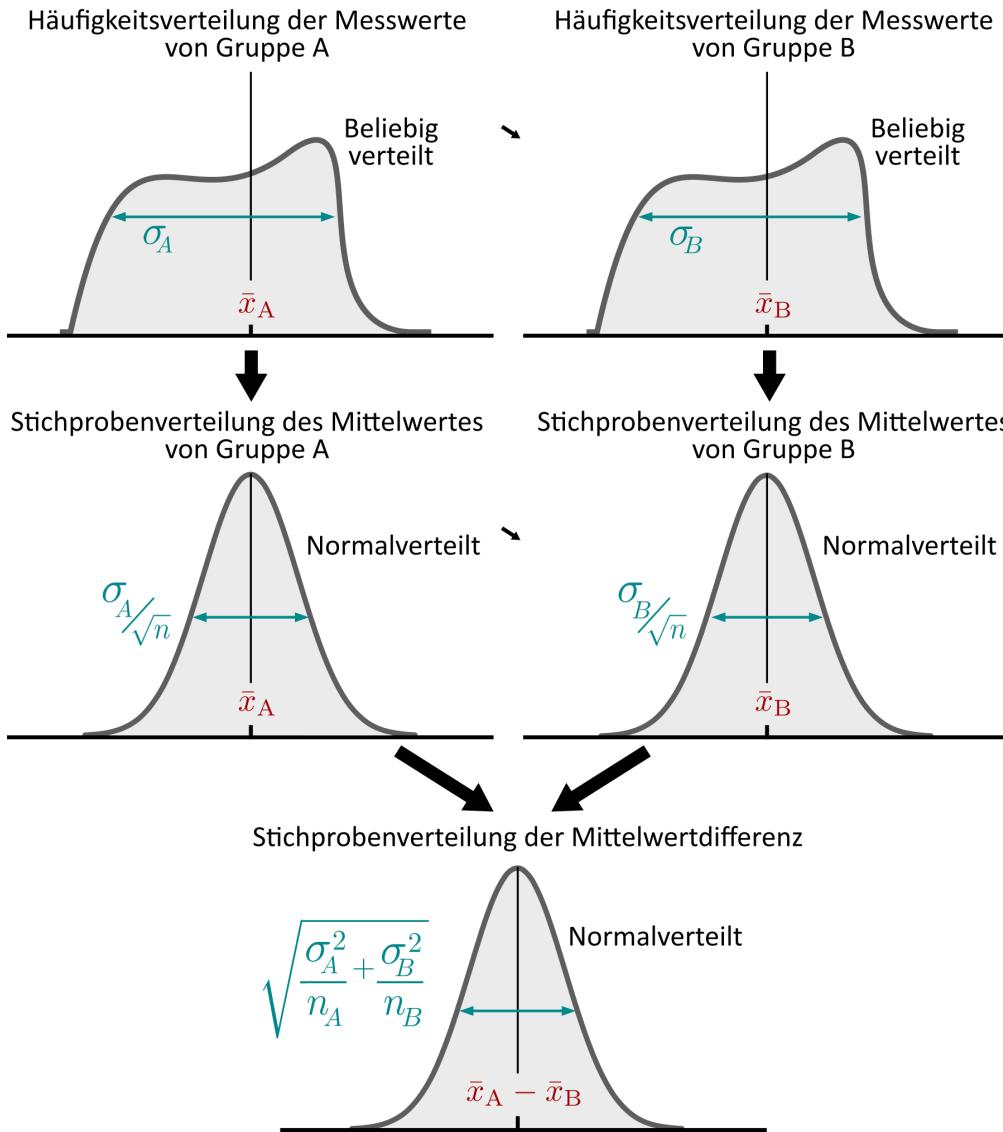


Die Standardfehler gelten für den Fall, dass die Varianzen σ_A^2 bzw. σ_B^2 in den Populationen bekannt sind (wie beim z-Test). Siehe Bonuscontent für eine Herleitung der **Standardfehler bei Mittelwertdifferenzen**.

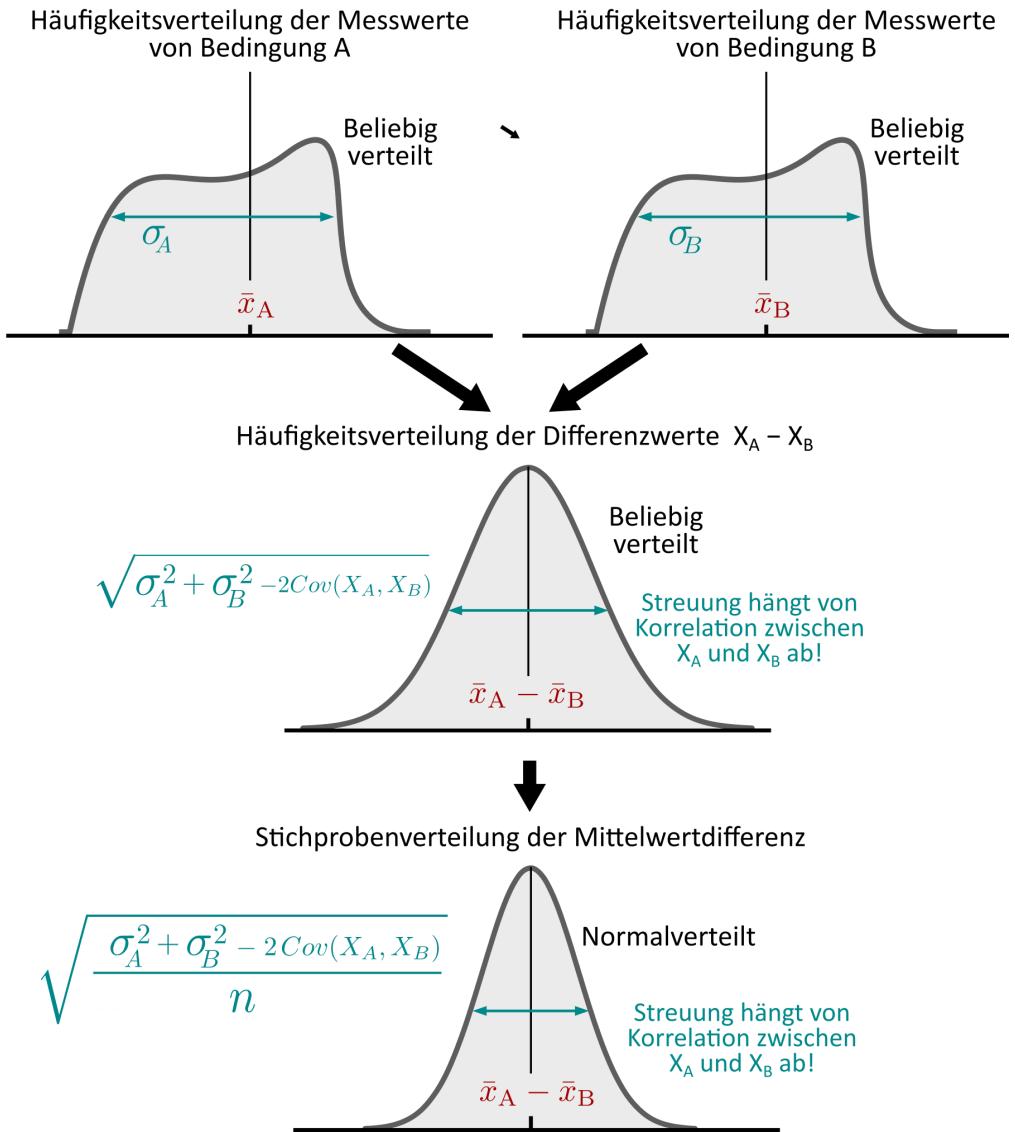
Sind die Populationsstreuungen nicht bekannt, sehen die Standardfehlerformeln aufgrund der Besselkorrektur etwas anders aus (→ siehe t-Test in Vorlesung 11).

Stichprobenverteilungen von Mittelwertdifferenzen (σ bekannt)

Mittelwertdifferenz: Unabhängige Messungen



Mittelwertdifferenz: Abhängige Messungen



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: $\overline{IQ}_{\text{HMU}} - \overline{IQ}_{\text{MSB}} = \Delta \overline{IQ} = 0$

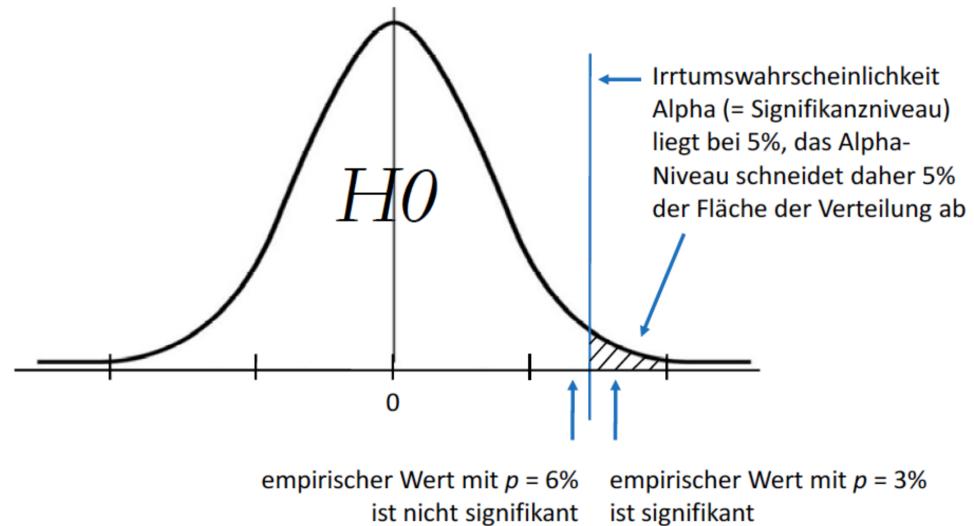
Signifikanzniveau ($p < \alpha$?)

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.

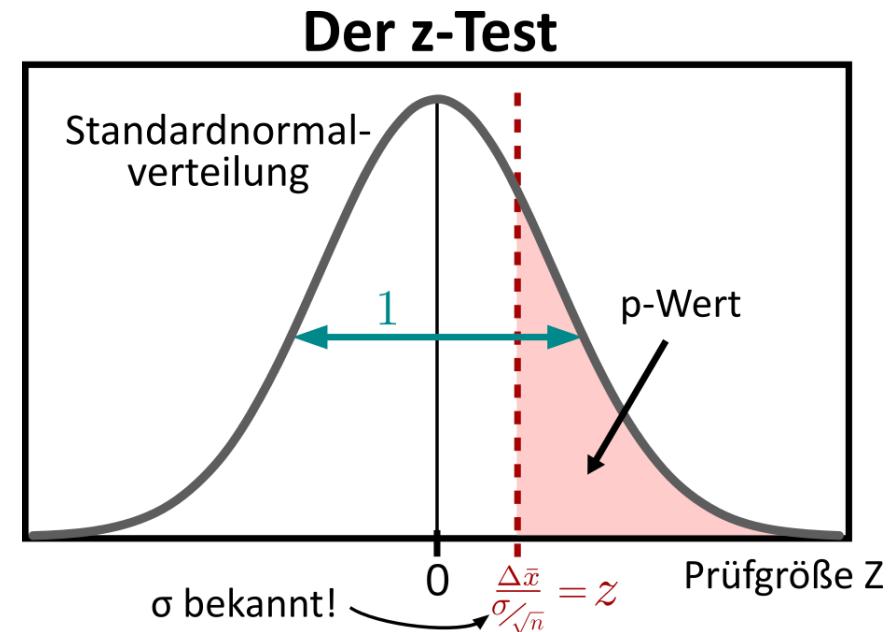
Auf welchen Wert das Signifikanzniveau festgelegt werden sollte ist seit langer Zeit Gegenstand von kontroversen Debatten in der Psychologie. Als de facto Standard-Signifikanzniveau hat sich jedoch der **Wert $\alpha = 0,05$** eingebürgert.



Ein- und zweiseitiges Testen

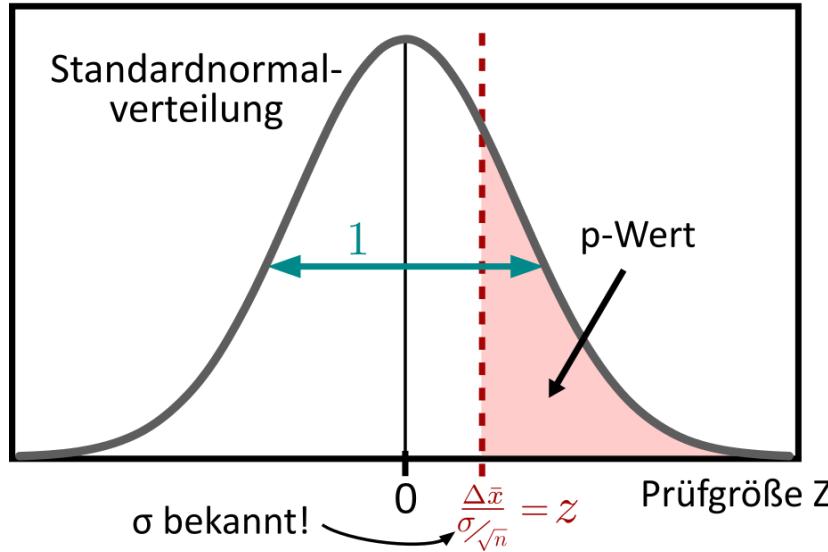
- Im Nasenlängenbeispiel haben wir bislang eine **gerichtete Hypothese betrachtet**, indem wir bei der Hypothese festgelegt haben, welche der beiden Gruppen durchschnittlich eine längere Nase hat (Med-Nasenlänge > Psych-Nasenlänge oder $\Delta\bar{x} > 0$).
- Wir hätten auch die **umgekehrte gerichtete 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



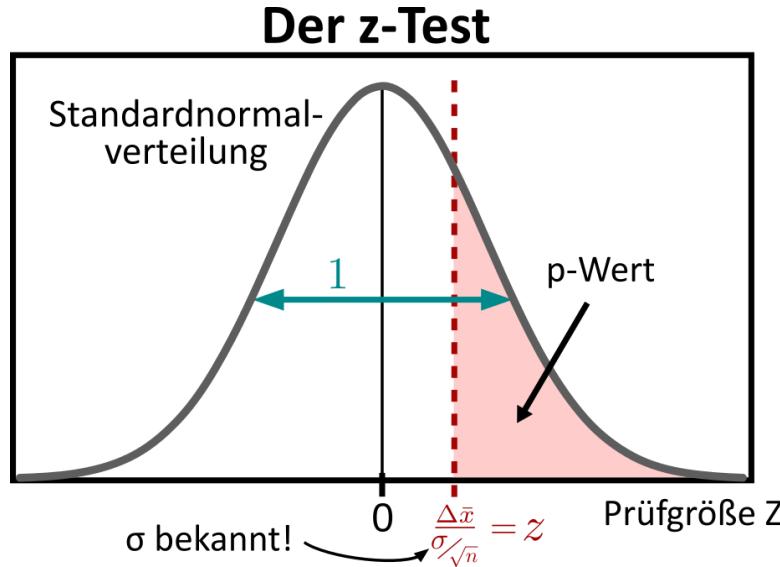
Wie sehen folgende Zusammenhänge:

- 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



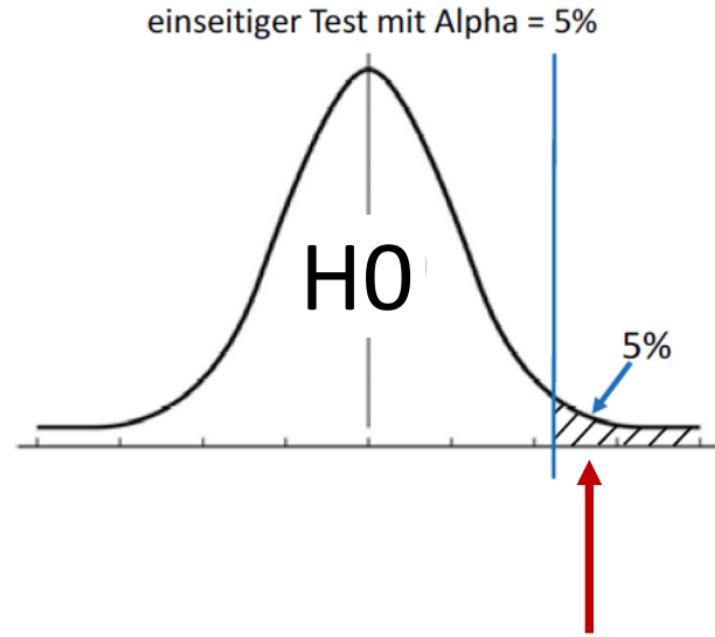
Der p-Wert der ungerichteten Hypothese ist damit immer größer (doppelt so groß) wie der kleinere p-Wert der gerichteten Hypothesen.

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”.

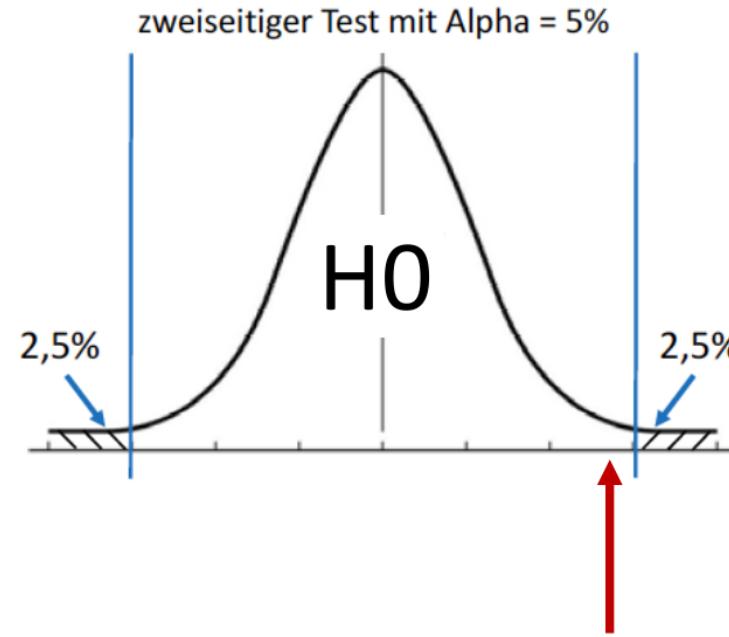
Genauer gesagt verdoppeln wir unsere Chance, mit der Hypothese richtig zu liegen, da a priori beide ungerichtete Hypothesen gleich wahrscheinlich sind. Die Verdopplung des p-Wertes kompensiert dies (man könnte auch sagen “bestraft unsere schwammige Vorhersage”) – nun wird es schwieriger für p das Signifikanzniveau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



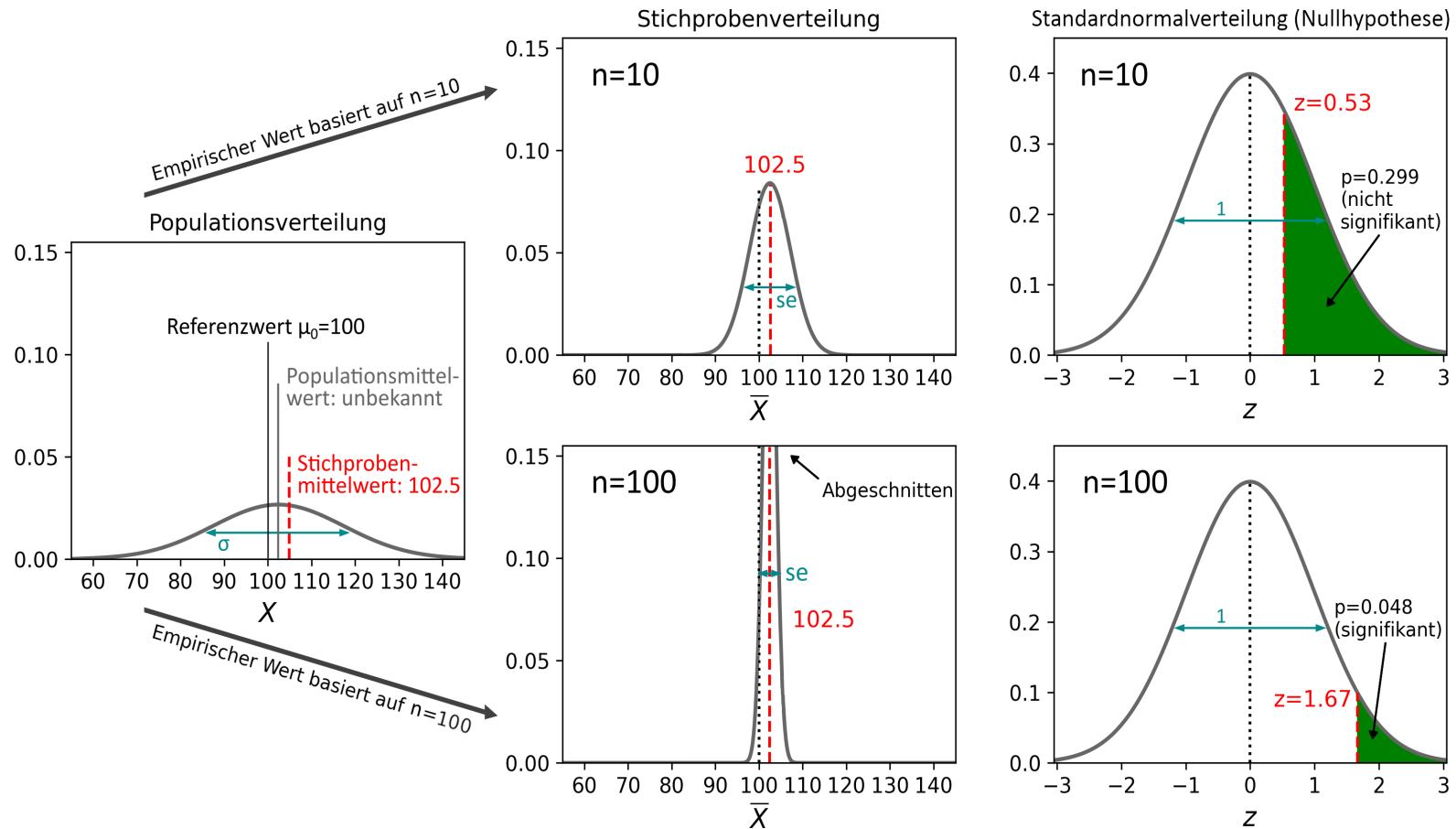
dieser Effekt wäre hier signifikant



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])

[[Zusammenfassung]]

- Bei der **Nullhypotesentestung** nach Fisher wird die Wahrscheinlichkeit geprüft, mit der ein beobachteter Effekt (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 untesuchte Hypothese **gerichtet** oder **ungerichtet** ist, erfolgt die statistische Testung entweder als **einseitiger Test** oder als **zweiseitiger Test**.



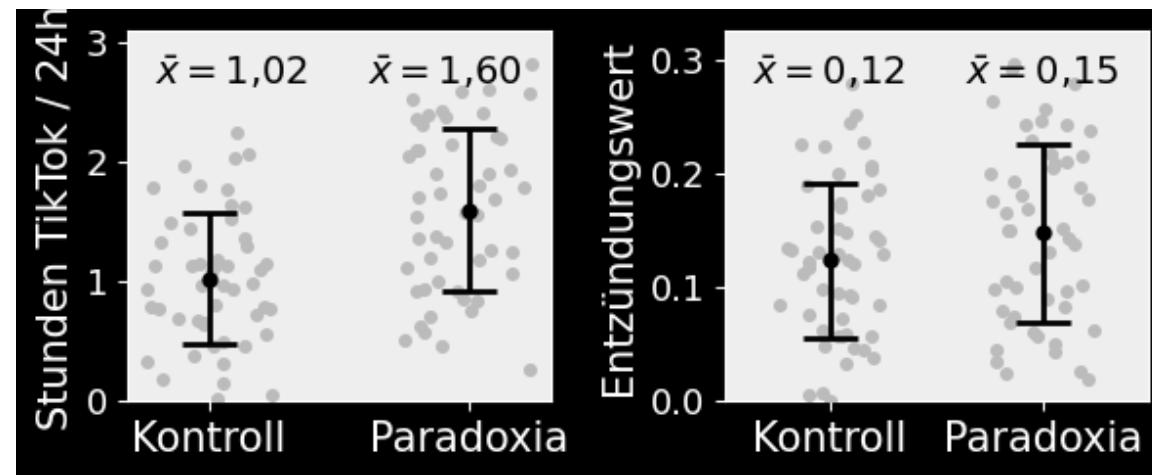
Nach kurzer Beratung in der Task Force sind Sie sich einig, dass die Voraussetzungen für einen z-Test nicht optimal sind. Die Populationsstreuung lässt sich nicht aus vorhergehenden Studien ableiten, da die untersuchten Phänomene brandneu sind. Da die Fallzahl von $n = 50$ gemäß der Faustregel “ $n \geq 30$ ” 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: $se = \sqrt{\frac{0,0682^2}{50} + \frac{0,0783^2}{50}} = 0,0147$ $\Delta\bar{x} = \bar{x}_{\text{paradox}} - \bar{x}_{\text{control}} = 0,1477 - 0,1233 = 0,0243$

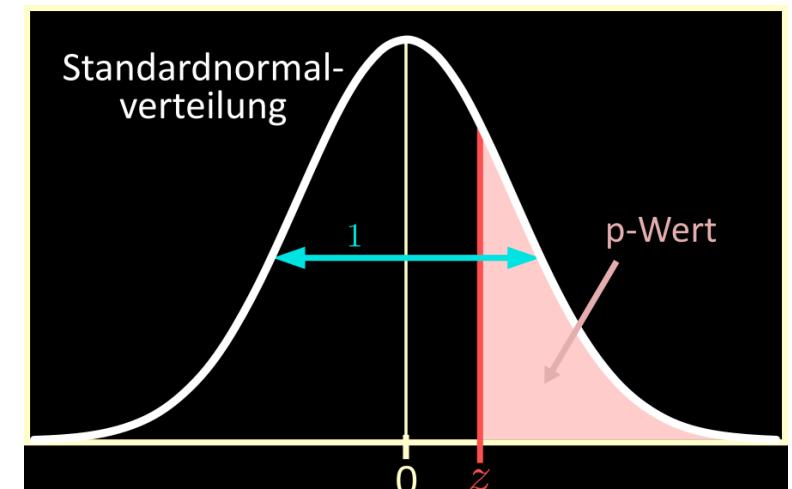
Der z-Wert ergibt sich als das Verhältnis aus dem “Effekt” (hier $\Delta\bar{x}$) und dem Standardfehler:

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

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

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:

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



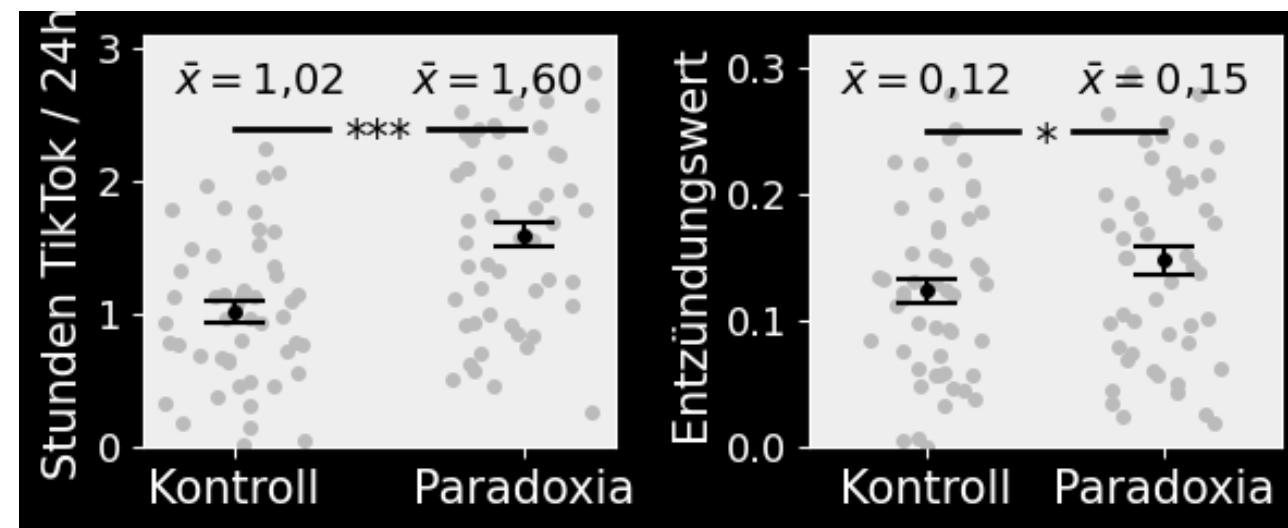
$$\text{Entzündung: } p = 1 - \Phi(1,67) \stackrel{\text{(Computer)}}{=} 0,049$$

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.



Bonuscontent

Standardfehler bei Mittelwertdifferenzen

Zentrale Frage: wie lässt sich auf Basis der (hier als bekannt vorausgesetzten) Populationsstreuungen σ_A und σ_B der Standardfehler der Mittelwertdifferenz berechnen?

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:

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

Standardfehler bei Mittelwertdifferenzen

Abhängige Messungen in einer Stichprobe

Ähnlich wie bei der Standardisierung der Effektstärke abhängiger Messungen, können die Variablen X_A und X_B im ersten Schritt direkt voneinander abgezogen werden:

$$\Delta X = X_A - X_B$$

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

Der Standardfehler der Mittelwertsdifferenz abhängiger Bedingungen A und B ergibt sich daher als:

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



Fußnoten

1. <https://www.majordifferences.com/2016/10/5-differences-between-null-and.html>
2. <https://www.statisticsfromatoz.com/blog/new-video-null-hypothesis>