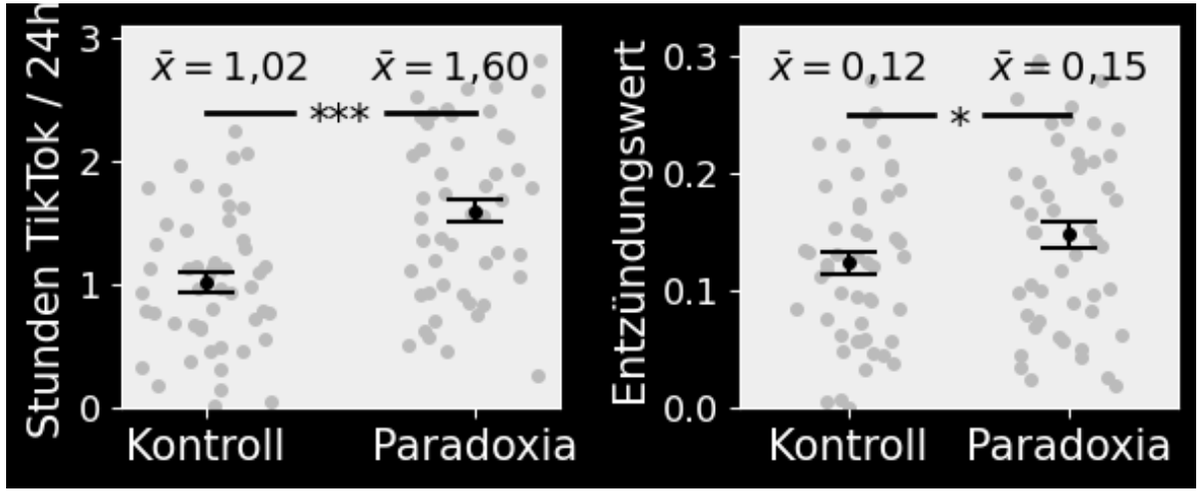
M24 Statistik 1: Wintersemester 23/24

Vorlesung 11: t-Test

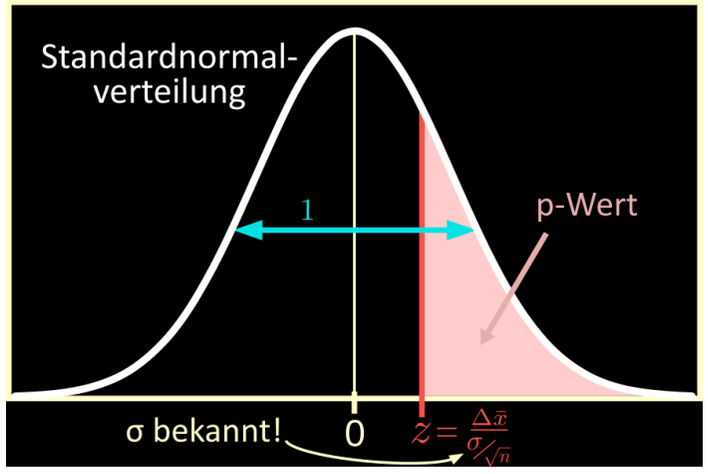
Prof. Matthias Guggenmos

Health and Medical University Potsdam

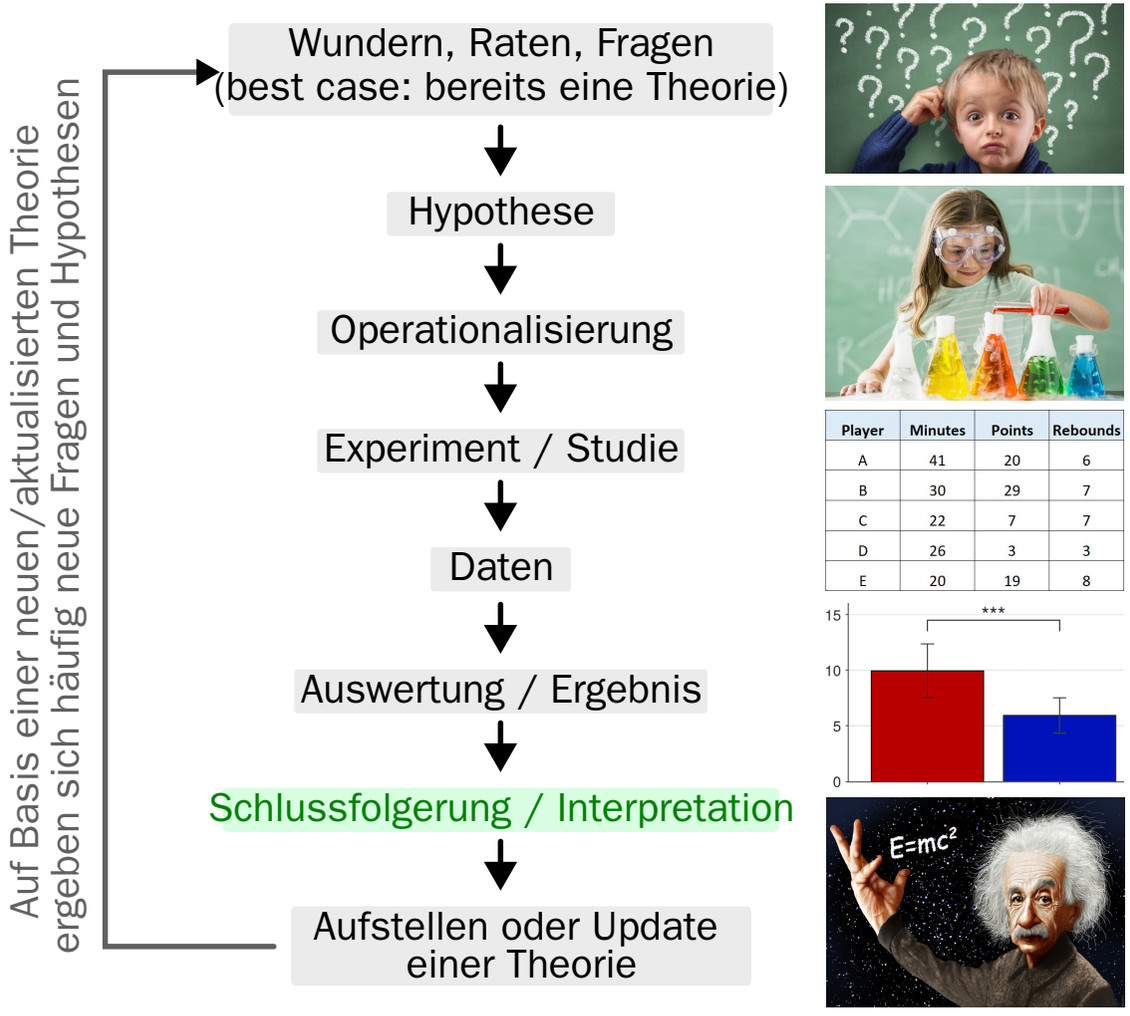
Erinnerung: den z-Test haben Sie unter der Voraussetzung durchgeführt, dass die Streuung σ in der Population bekannt ist (bzw. die Streuung*en* bei zwei Populationen).



Wie schon angesprochen, kommt dieser Fall in der Praxis allerdings nahezu nie vor und gilt auch für den vorliegenden Fall nicht. What shall you do?



Der Forschungsprozess



Einschub: Anmerkungen zur Klausur

Infos zur Klausur: Allgemeines

- **Struktur:** 50% Verständnisfragen (Großteil *multiple choice*), 50% Rechnen
- **Hilfsmittel:** Formelsammlung, einfacher Taschenrechner ($+/-/\cdot/:$), ein Lineal oder Geodreieck wird nicht benötigt (ist aber erlaubt)
- **Wichtig:** die Formelsammlung druckt das Prüfungsbüro für Sie aus. Sie dürfen keinen eigenen Ausdruck in die Klausur mitnehmen.
- **Klausurinhalt:** Stoff der Vorlesungsfolien (Ausnahme: rot markiert)
- **Rechnen:** welche Formeln muss ich beherrschen?
  - Die Formelsammlung bietet einen sehr guten Überblick über mögliche Formeln in der Klausur.
  - Bis auf rot gefärbte Formeln sind alle Formeln klausurrelevant.
  - Sie können und sollen die Formelsammlung in der Klausur anwenden.
  - Das Ziel von Statistik 1 ist explizit nicht, dass Sie Formeln auswendig lernen; Ziel ist die Kompetenz, passende Formeln nachzuschlagen und konzeptuell zu verstehen.
  - Achten Sie darauf, dass Sie die aktuelle Version der Formelsammlung von der Website hmu-stats.github.io verwenden:
    - Version vom 19.1.2024 — siehe Datum in der Formelsammlung
    - Im Fall einer Aktualisierung der Formelsammlung vor der Klausur gebe ich Ihnen in jedem Fall Bescheid.

Infos zur Klausur: spezifische Bemerkungen

- **z-Test:** es werden keine *Rechen*aufgaben zum z-Test in der Klausur gestellt, da dieser eine geringe Praxisrelevanz hat und hauptsächlich aus pädagogischen Gründen gelehrt wird.
  - Aber: Sie sollten konzeptuell verstehen, was der prinzipielle Einsatzzweck des z-Testes ist (verglichen mit dem t-Test).
- **Bessel-Korrektur:**
  - Es soll verstanden werden, warum es sie gibt und wann sie prinzipiell angewendet (Stichprobe -> Population) oder nicht angewendet (deskriptive Beschreibung der Stichprobe) wird.
  - Bei Rechenaufgaben wird die Besselkorrektur keine Rolle spielen, d.h. es wird nicht als Fehler gewertet, wenn Sie die Besselkorrektur irrtümlich anwenden oder nicht anwenden.
  - Das impliziert auch, dass Sie bei Rechenaufgaben wahlweise die Stichprobenvarianz s^2 oder die Populationsschätzung $\hat{\sigma}^2$ angeben bzw. verwenden können.
  - Der Grund ist, dass die Anwendung der Besselkorrektur selbst in akademischen Veröffentlichungen und Lehrbüchern sehr inkonsistent ist. Auf den folgenden beiden Folien finden Sie zwei Beispiele.

Besselkorrektur meets practise: z-Transformation

- Die **z-Standardisierung** oder **z-Transformation** ist eine häufig genutzte Methode, um Variablen zwischen verschiedenen Studien vergleichbar zu machen.
- *Streng genommen* nimmt die z-Transformation an, dass Populationsmittelwert μ und die Populationsstreuung σ bekannt sind. Die Formel lautet daher:

$$\text{z-Transformation:} \qquad Z = \frac{X - \mu}{\sigma}$$

- In der Praxis sind σ und μ aber nahezu nie bekannt und daher werden häufig der empirische Mittelwert $\Delta\bar{x}$ und die Populationsschätzung $\hat{\sigma}$ mit Besselkorrektur verwendet.
- Seien Sie sich aber bewusst, dass dies aus theoretischer Sicht eine Ungenauigkeit darstellt, denn bei unbekannter Populationsstreuung handelt es sich bei Z nicht um einen normalverteilten Wert, sondern um einen t-verteilten Wert (und damit eigentlich um eine *t-Transformation*).
- Erst ab ca. $N \geq 30$ sind die Stichprobenschätzungen so gut, dass die z-Transformation guten Gewissens auch bei unbekannten Populationsparametern angwendet werden sollte.

Besselkorrektur meets practise: Korrelation

Für die Formel der Pearson-Korrelation hatten wir kennengelernt:

$$r = \frac{1}{s_X s_Y} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Dies ist **Stichprobenkorrelationskoeffizient**.
- In der Regel sind wir aber an einer Schätzung der Korrelation in der Population interessiert → dies führt zum **empirischen Korrelationskoeffizienten**:

$$\hat{\rho} = \frac{1}{\hat{\sigma}_X \hat{\sigma}_Y} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

(beachten Sie den Übergang $s \to \hat{\sigma}$ und $n \to n-1$, d.h. die Besselkorrektur)

- In der Praxis herrscht hier leider keine Einheitlichkeit. De facto wird in der Literatur standardmäßig die Bezeichnung r für den Korrelationskoeffizienten verwendet, aber mit der Formel des empirischen Korrelationskoeffizienten.

Der t-Test

Problemstellung

- Beim **z-Test** für Mittelwertunterschiede haben wir für die Verteilung der Prüfgröße $z = \frac{\Delta\bar{x}}{se}$ eine Normalverteilung annehmen können. Grund:
  - Aufgrund des zentralen Grenzwertsatzes können viele statistische Kennwerte, darunter auch die Mittelwertdifferenz $\Delta\bar{x}$, bei ausreichender Stichprobengröße als normalverteilt angenommen werden.
  - Beim z-Test wird die Standardabweichung σ der Population, und damit auch der Standardfehler $se = \frac{\sigma}{\sqrt{n}}$, als bekannt angenommen.
  - De facto wird also bei der Berechnung der Prüfgröße $z = \frac{\Delta\bar{x}}{se}$ eine normalverteilte Variable ($\Delta\bar{x}$) durch eine Konstante ($\sigma/\sqrt{n}$) geteilt, so dass auch die Prüfgröße normalverteilt ist.
- Ist σ *nicht bekannt*, muss die Streuung auf Basis der Stichprobe als $\hat{\sigma}$ geschätzt werden.
- Die Schätzung von $\hat{\sigma}$ ist mit Unsicherheit behaftet und daher selbst eine **Zufallsvariable**.
- Wir teilen also eine normalverteilte Zufallsvariable $\Delta\bar{x}$ durch eine wie-auch-immer-verteilte (*) zweite Zufallsvariable $\hat{\sigma}$. (* die wie-auch-immer-Verteilung ist bekannt: $\hat{\sigma}$ folgt der Chi-Verteilung bzw. χ-Verteilung)
- **In diesem Fall können wir nicht mehr davon ausgehen, dass die Prüfgröße normalverteilt ist!**

Die Gretchen-Frage ist daher: welcher Verteilung folgt die Prüfstatistik

$$\frac{\Delta\bar{x}}{\hat{se}} = \frac{\Delta\bar{x}}{\hat{\sigma}/\sqrt{n}} \quad ? \quad (\text{mit Betonung auf dem } \hat{} \text{ über } \hat{\sigma})$$

Statistischer Kennwert $\hat{\theta}$

- Bevor wir uns der eigentlichen Frage widmen, führen wir ein neues Symbol ein, das als Platzhalter für einen statistischen Kennwert steht, der auf Basis einer Stichprobe berechnet wurde:

$$\text{Statistischer Kennwert:} \qquad \hat{\theta}$$

- Das Symbol $\hat{\theta}$ ist ein allgemeines Symbol für alle statistischen Kennwerte, die wir bisher kennengelernt haben: Mittelwert $\bar{x}$, Mittelwertdifferenz $\Delta\bar{x}$, Korrelation r, Regressionskoeffizient b_1,...

Wir formulieren unsere Frage also etwas allgemeiner: welcher Verteilung folgt die Größe

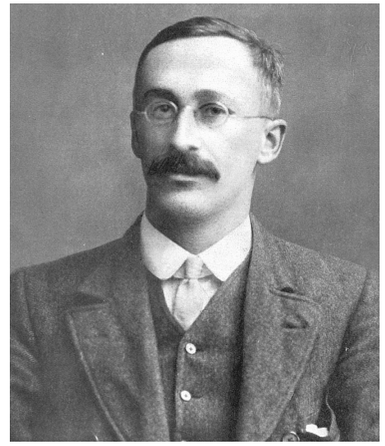
$$\frac{\hat{\theta}}{\hat{se}} \qquad ?$$

Streng genommen lautet die fragliche Prüfgröße $\frac{\hat{\theta}-\theta_0}{\hat{se}}$. Von einigen Ausnahmen abgesehen gilt jedoch zumeist $\theta_0 = 0$ und wir lassen θ_0 aus diesem Grund auch im Folgenden weg.

Ein Bespiel für eine Ausnahme ist, wenn $\hat{\theta}$ ein einfacher Mittelwert ist und gegen einen Referenzwert wie den Durchschnitts-IQ $\theta_0 = 100$ getestetet wird; in diesem Fall muss also $\theta_0 = 100$ zunächst vom Mittelwert $\hat{\theta}$ abgezogen werden.

Die t-Verteilung

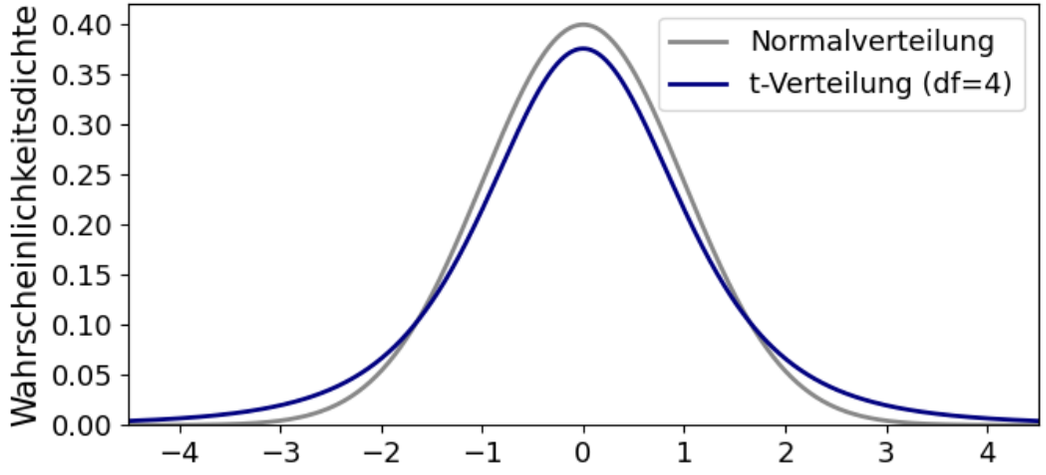
- An dieser Stelle kommt der englische Statistiker **William Sealy Gosset** ins Spiel.
- Er postulierte 1908, dass das Verhältnis eines normalverteilten Kennwertes $\hat{\theta}$ und einer Chi-verteilten Variable einer Wahrscheinlichkeitsverteilung folgt, die seither als **t-Verteilung** (oder auch **Studentsche t-Verteilung**) bezeichnet wird.
- Die Formel der Verteilung ist vergleichsweise kompliziert — für die Praxis entscheidend ist, dass sie **durch einen einzigen Parameter (df) definiert wird**:



William Sealy Gosset (1876-1937)

$$t\text{-Verteilung:} \qquad f_t(x|\mathrm{df})$$

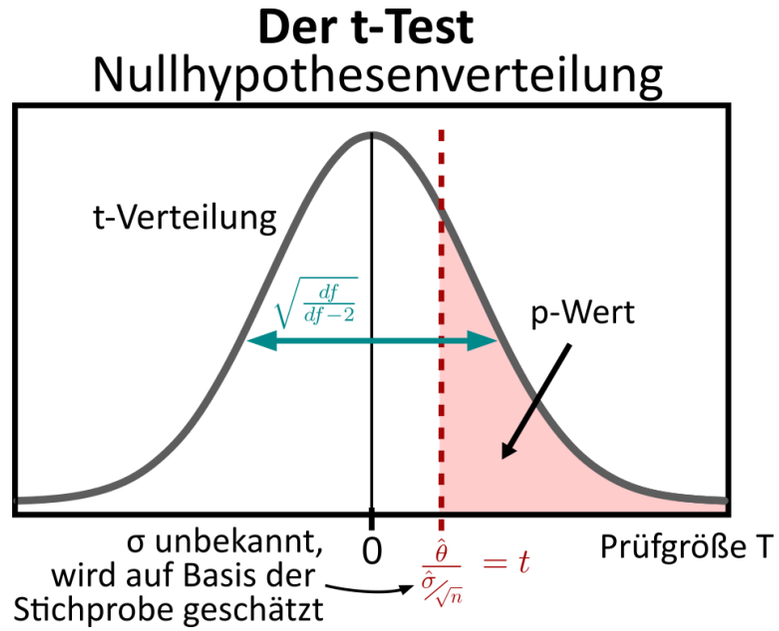
- Der Parameter df steht für die **Zahl der Freiheitsgrade** (*degrees of freedom*) und hängt eng mit der Stichprobengröße n zusammen.
- Im Bild rechts ist eine t-Verteilung mit 4 Freiheitsgraden im Vergleich zur Normalverteilung aufgetragen.
- Erste Erkenntnis: die t-Verteilung hat etwas dickere Flanken (*fat tails*)!



t-Wert

- Wir kennen nun also die Form der Nullverteilung, wenn die Streuung als $\hat{\sigma}$ auf Basis der Stichprobe geschätzt werden muss: **t-Verteilung**
- In Korrespondenz mit dem Namen der Verteilung wird die Prüfgröße auf Basis von $\hat{\sigma}$ als **t-Wert** bezeichnet:

$$t = \frac{\hat{\theta}}{\hat{se}} = \frac{\hat{\theta}}{\hat{\sigma}/\sqrt{n}}$$



- Das Prinzip der p-Wert-Bestimmung ist exakt wie beim z-Test: es wird die Fläche unter der t-Verteilung relativ zum Prüfgrößenwert t bestimmt:
  - Gerichtete Hypothese $\hat{\theta} > 0$: Fläche rechts von t
  - Gerichtete Hypothese $\hat{\theta} < 0$: Fläche links von t
  - Ungerichtete Hypothese $\hat{\theta} \neq 0$: Fläche links von $-|t|$ **PLUS** Fläche rechts von $|t|$
- Merke: der *t-Wert* ist zur *t-Verteilung* wie der *z-Wert* zur *Standardnormalverteilung*!

Die t-Verteilung ist bereits standardisiert

Vor der Einführung des z-Tests hatten wir zunächst die **unstandardisierte Normalverteilung** kennengelernt, die durch zwei Parameter definiert ist: Mittelwert μ und Standardabweichung σ:

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

Verwenden wir statt x die standardisierte Variable $\frac{x-\mu}{\sigma}$, vereinfacht sich die Nullhypothesen-Verteilung zur **Standardnormalverteilung**:

$$f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Die Standardnormalverteilung hat keinen Parameter mehr (nur noch die *Variable* x). Da es sich beim z-Wert auch um eine standardisierte Variable handelte, konnten wir auch für z eine Standardnormalverteilung annehmen:

$$f(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Die **t-Verteilung** hatten wir dagegen direkt auf Basis der standardisierten Prüfgröße $\frac{\hat{\theta}}{\hat{se}}$ eingeführt. **Die klassische t-Verteilung ist aus diesem Grund bereits eine *standardisierte Verteilung*, die nicht mehr vom Mittelwert oder Streuung abhängt.** Im Gegensatz zur Standardnormalverteilung hat sie aber noch einen Parameter: die Zahl der Freiheitsgrade df.

Unstandardisierte t-Verteilung

Die Formel für die klassische (standardisierte) t-Verteilung lautet:

$$f_t(x|\mathrm{df}) = \frac{\Gamma\left(\frac{\mathrm{df}+1}{2}\right)}{\sqrt{\mathrm{df}\,\pi}\,\Gamma\left(\frac{\mathrm{df}}{2}\right)}\left(1+\frac{1}{\mathrm{df}}x^2\right)^{-\frac{\mathrm{df}+1}{2}} \qquad (\Gamma \text{ ist die Gamma-Funktion})$$

Wie die Standardnormalverteilung hat die t-Verteilung Mittelwert 0 (daher kann sie als Nullhypothesenverteilung fungieren). Ihre Streuung ist jedoch nicht exakt 1, sondern hängt von der Zahl der Freiheitsgrade ab: $\sigma^2 = \frac{\mathrm{df}}{\mathrm{df}-2}$.

Tatsächlich gibt es auch zur t-Verteilung ein unstandardisiertes Pendant, die **nicht-standardisierte t-Verteilung**, die von Mittelwert μ und Streuung σ abhängt:

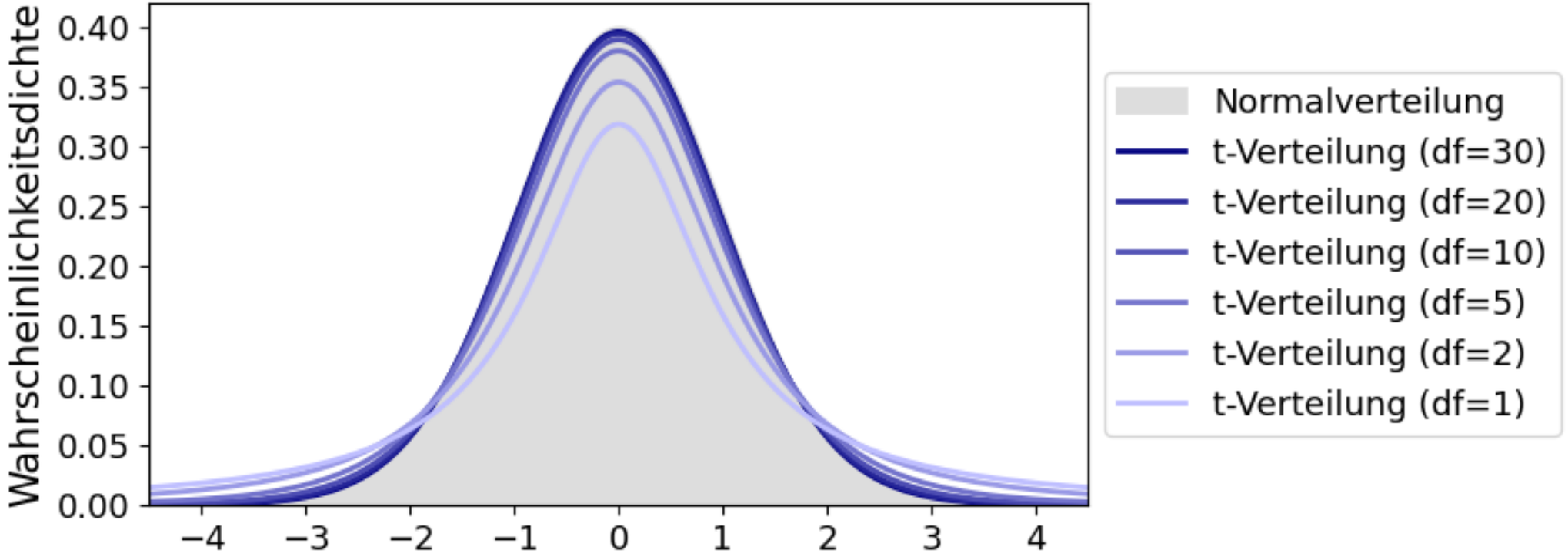
$$f_t(x|\mu,\sigma,\mathrm{df}) = \frac{\Gamma\left(\frac{\mathrm{df}+1}{2}\right)}{\sigma\sqrt{\mathrm{df}\,\pi}\,\Gamma\left(\frac{\mathrm{df}}{2}\right)}\left(1+\frac{1}{\mathrm{df}}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\mathrm{df}+1}{2}}$$

Ähnlich wie beim z-Test hat es sich in der Praxis aber durchgesetzt immer mit standardisierten Prüfgrößen (wie z oder t) zu arbeiten. Daher finden die unstandardisierte t- und Normalverteilung im Kontext der Hypothesentestung selten eine Anwendung.

Der **Vorteil standardisierter Prüfgrößen** ist, dass diese vergleichbar zwischen Studien sind. Ein bestimmter z- oder t-Wert hat eine Aussagekraft, ohne den Standardfehler einer Studie zu kennen.

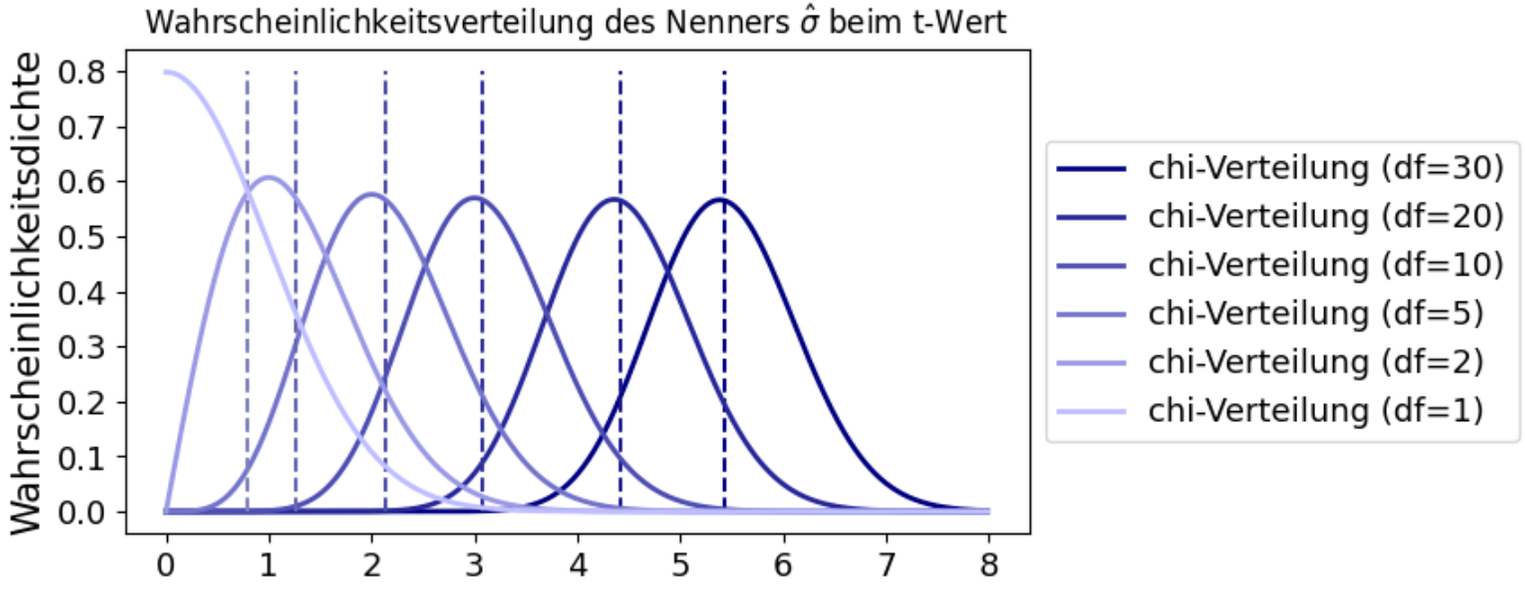
t-Verteilung versus Normalverteilung

- Der einzige Parameter der t-Verteilung, die Zahl der Freiheitsgrade df, bestimmt die Form der Verteilung.
- Die Zahl der Freiheitsgrade df hängt eng mit der Stichprobengröße n zusammen (z.B. $\mathrm{df} = n - 1$ bei einer Mittelwertdifferenz abhängiger Messungen).
- Je größer df/Stichprobengröße, desto ähnlicher wird die t-Verteilung der Normalverteilung!



Intuition: verdickte Flanken der t-Verteilung

- Warum sind die Flanken der Verteilung der Testgröße $t = \frac{\hat{\theta}}{\hat{\sigma}/\sqrt{n}}$ stärker ausgeprägt, wenn $\hat{\sigma}$ auf Basis der Stichprobe geschätzt werden muss?
- Der Grund liegt in der (Chi-)Verteilung der Zufallsvariable $\hat{\sigma}$ im Nenner:
  - Bei kleinen Stichprobengrößen (kleines df), gibt es eine Assymmetrie der Verteilung hin zu Werten kleiner dem Mittelwert (welcher die korrekte Schätzung von σ repräsentiert — in der Abbildung gestrichelte Linien)
  - D.h. wir *teilen $\hat{\theta}$ überproportional häufig durch zu kleine Werte*, wodurch die *Teststatistik* $t = \frac{\hat{\theta}}{\hat{\sigma}/\sqrt{n}}$ *überproportional häufig zu extreme (negative oder positive) Werte* liefert.
  - Dies führt zu den stärkeren Flanken der t-Verteilung!



Funktionen der t-Verteilung

Wie die Normalverteilung wird auch die t-Verteilung durch eine **Wahrscheinlichkeitsdichte** definiert:

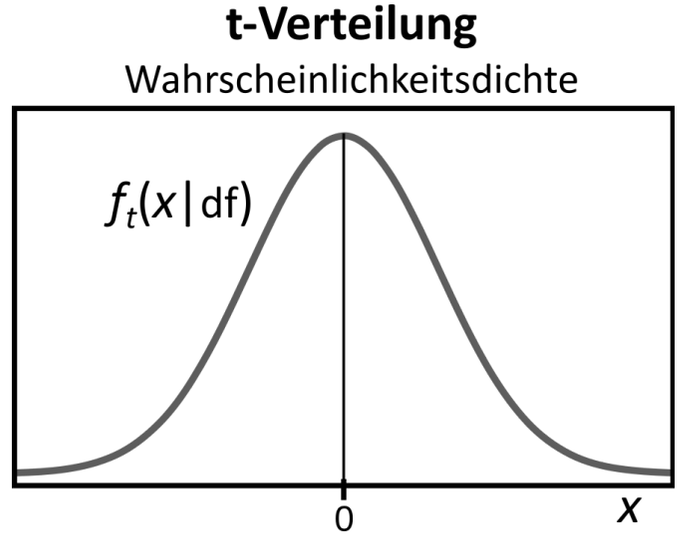
Variable (Achtung: für *x* werden in der Praxis *t*-Werte eingesetzt!)

Parameter der *t*-Verteilung (df = Zahl der Freiheitsgrade)

kleines "*f*" zeigt an, dass es sich um eine Dichte handelt

$$f_t(x|\mathrm{df})$$

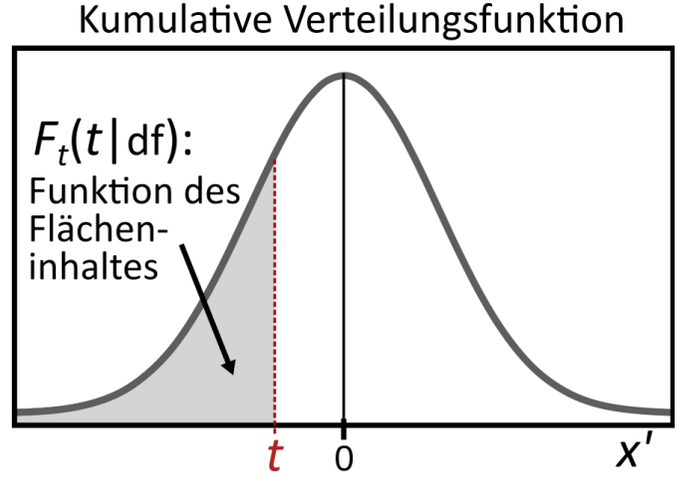
Subscript zeigt an, dass es sich um eine *t*-Verteilung handelt



Die zugehörige **kumulative Verteilungsfunktion** wird der Konvention entsprechend mit einem großen F denotiert:

$$\text{Kumulative Verteilungsfunktion:}$$

$$F_t(x|\mathrm{df}) = \int_{-\infty}^{x} f_t(x'|\mathrm{df})\,dx'$$



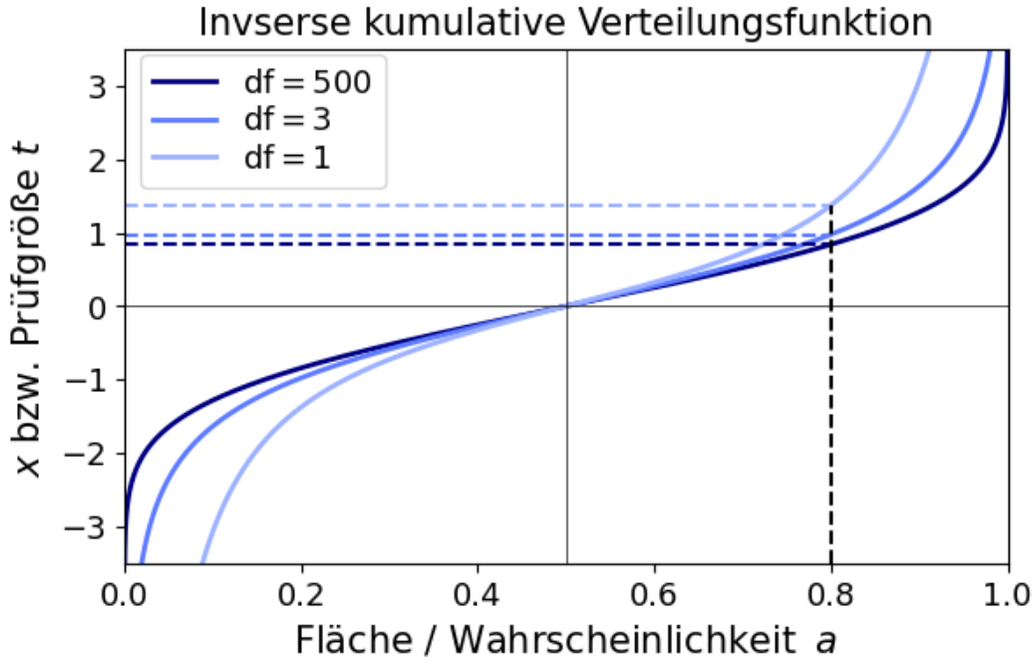
Die kumulative Verteilungsfunktion ordnet jedem t-Wert den Flächeinhalt unter der t-Verteilung im Bereich $[-\infty; t]$ zu.

Funktionen der t-Verteilung

Zuweilen stellt sich die umgekehrte Frage:
Wie lautet der t-Wert für einen bestimmten Flächeninhalt (z.B. einen bestimmten p-Wert) unter der t-Verteilung?

Diesen Zusammenhang stellt die **inverse kumulative Verteilungsfunktion** her:

$$t = F_t^{-1}(a|\mathrm{df})$$



wobei a die Fläche (area) bezeichnet. Da die t-Verteilung eine Wahrscheinlichkeitsverteilung ist, kann die Fläche a nur Werte zwischen 0 und 1 annehmen.

Die Funktion wird auch **Quantilfunktion** genannt, da sie z.B. für $a = 0{,}8$ den Wert t zurück gibt, der 80% der t-Verteilung (von $-\infty$ gerechnet) umfasst.

Sämtliche Funktionen (Wahrscheinlichkeitsdichte, kumulative Verteilungsfunktion, inverse kumulative Verteilungsfunktion) sind zu kompliziert zur manuellen Berechnung und werden mit dem Computer ausgewertet. Zusätzlich bietet die t-Tabelle die Möglichkeit, kritische t-Werte für ausgewählte Signifikanzniveaus nachzuschlagen.

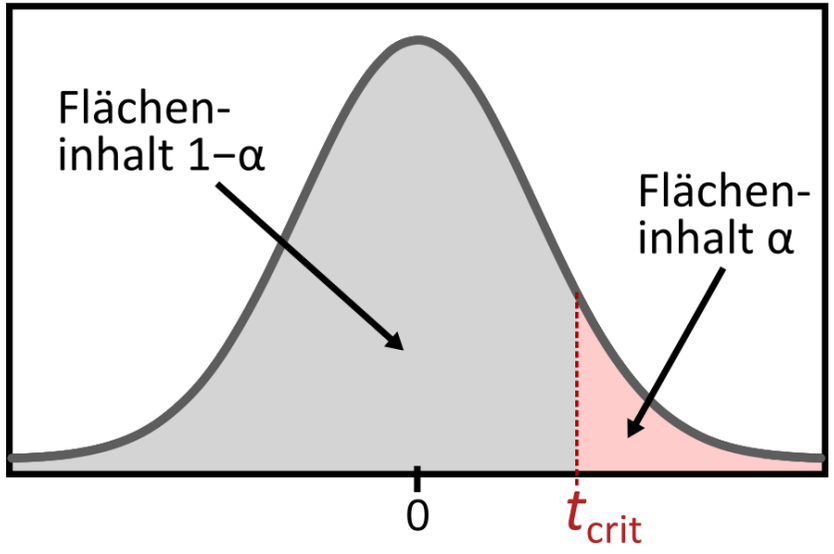
Funktionen der t-Verteilung

Mit der inversen kumulativen Verteilungsfunktion können **kritische t-Werte** für gegebene Signifikanzniveaus α berechnet werden, wie sie von der **t-Tabelle** bereitgestellt werden.

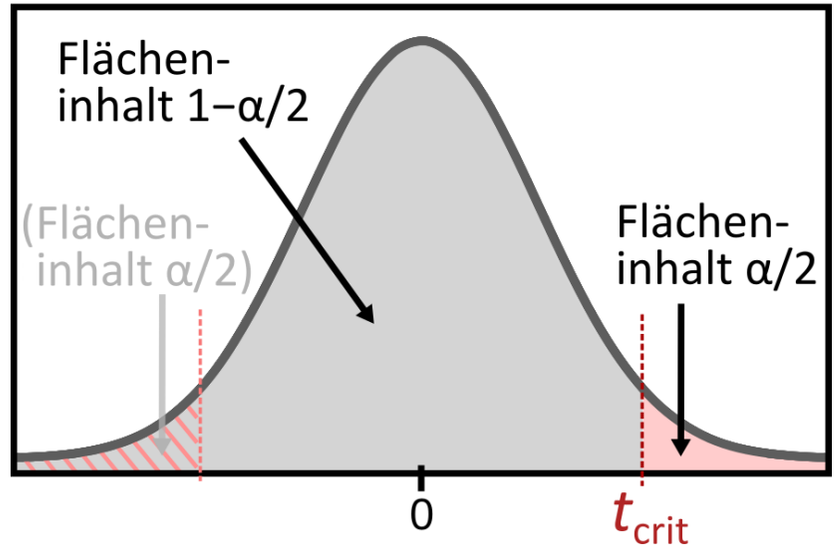
Bei gerichteter Hypothese ist der kritische t-Wert bei einem Signifikanzniveau von α gleich $F_t^{-1}(1-\alpha|\mathrm{df})$, bei ungerichteter Hypothese gleich $F_t^{-1}(1-\frac{\alpha}{2}|\mathrm{df})$. Für die Angabe eines kritischen t-Wertes wird als Subscript die kritische Fläche und die Zahl der Freiheitsgrade df angegeben:

$$\text{Gerichtet:} \quad t_{\text{crit}} = t_{(1-\alpha,\,\mathrm{df})} \qquad \text{Ungerichtet:} \quad t_{\text{crit}} = t_{(1-\frac{\alpha}{2},\,\mathrm{df})}$$

Gerichtete Hypothese

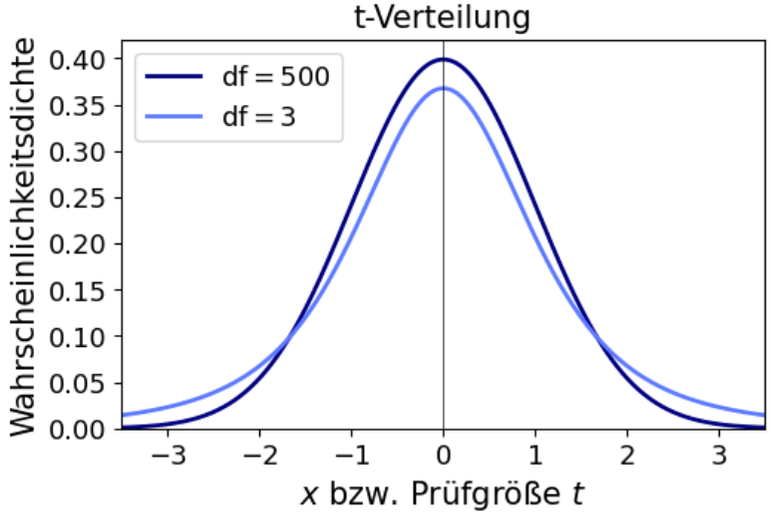


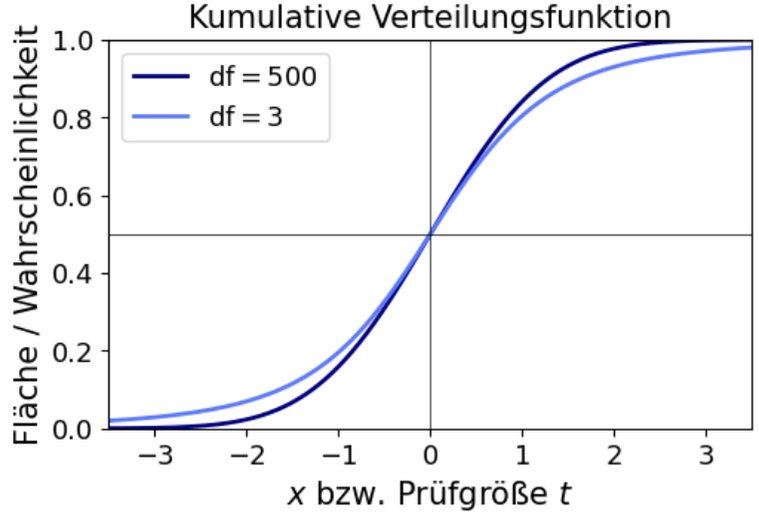
Ungerichtete Hypothese

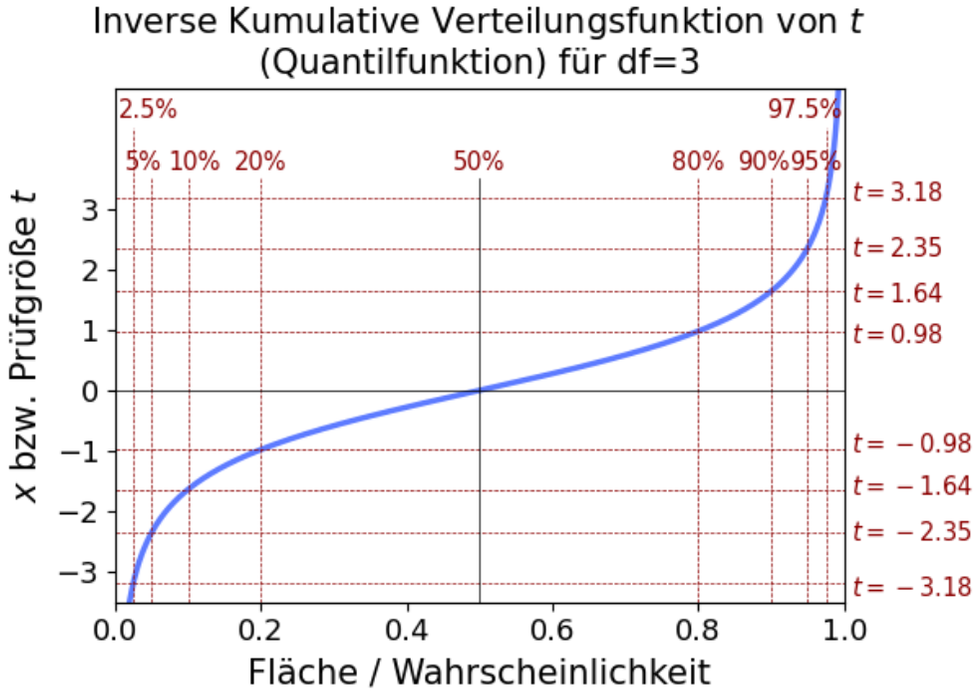


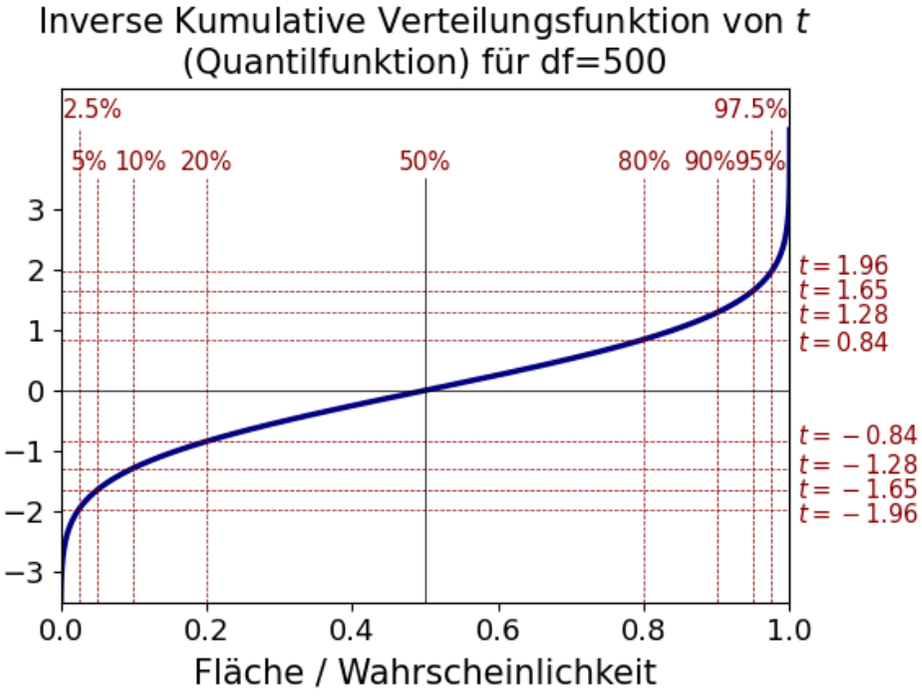
t_{crit} gibt den Wert an, den die Prüfgröße t mindestens erreichen muss, damit der zugehörige p-Wert kleiner als das Signifikanzniveau α ist.

Funktionen der t-Verteilung









Durchführung eines t-Testes

t-Test

Zur Durchführung des t-Tests müssen wir uns über folgende Faktoren Gedanken machen:

0. Sind die Voraussetzungen für einen t-Test gegeben?

Jeder statistische Test basiert auf bestimmten Annahmen. Der Test sollte daher nur dann angewendet werden, wenn diese Annahmen erfüllt sind (oder sich eine Verletzung der Annahme in der Praxis als wenig problematisch herausgestellt hat).

1. Welche Art von t-Test benötige ich?

Für verschiedene Szenarien gibt es unterschiedliche t-Tests, die sich unterscheiden in a) dem verwendeten Standardfehler und b) der Zahl der Freiheitsgrade df.

2. Zahl der Freiheitsgrade

Die Zahl der Freiheitsgrade df ist als Parameter für die t-Verteilung notwendig, und damit auch notwendig um Flächen (wie p-Werte) unter der Verteilung zu berechnen. De facto übernehmen das heute Computerprogramme, jedoch ist es nach wie vor Usus die Zahl der Freiheitsgrade eine statistischen Testes in wissenschaftlichen Veröffentlichungen anzugeben.

3. Signifikanzniveau & ein/zweiseitige Testung

Welchen Wert bestimme ich als Signifikanzniveau α? Ist mein Test einseitig (gerichtete Hypothese) oder zweiseitig (ungerichtete Hypothese)? → siehe auch Vorlesung 10

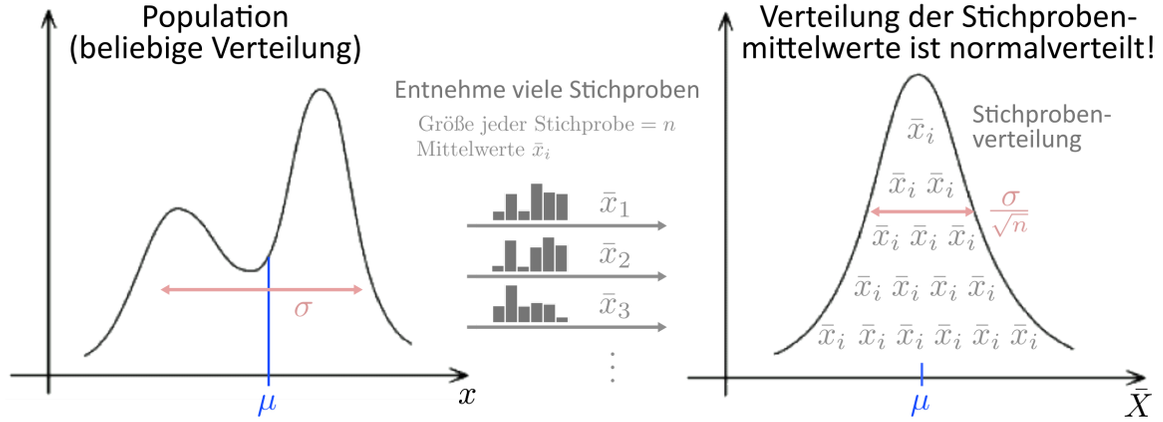
0. Sind die Voraussetzungen für einen t-Test gegeben?

- Mindestens **intervallskalierte Daten**.
- **Hinreichende Normalverteilung** des gemessenen Merkmals X in der Population (Zweistichproben-t-Test: Normalverteilung von X_A und X_B in den beiden Gruppen A und B!)
  - Allerdings zeigen Simulationsstudien, dass der t-Test sehr „robust“ gegen Verletzungen der Normalverteilungsannahme ist.
  - Problematisch sind stark schiefe Verteilungen (rechts- oder linksschief) bei kleinen Stichprobengrößen.
  - Kann von keiner Normalverteilung ausgegangen werden: nicht-parametrische Testverfahren (z.B. Mann-Whitney-U-Test).

Mr. T

But wait..

.. hatten wir nicht festgestellt, dass die Stichprobenverteilung von Kennwerten $\hat{\theta}$ gemäß dem zentralen Grenzwertsatz bei *beliebigen* Populationsverteilungen des gemessenen Merkmals X normalverteilt ist? (vgl. Vorlesung 08)



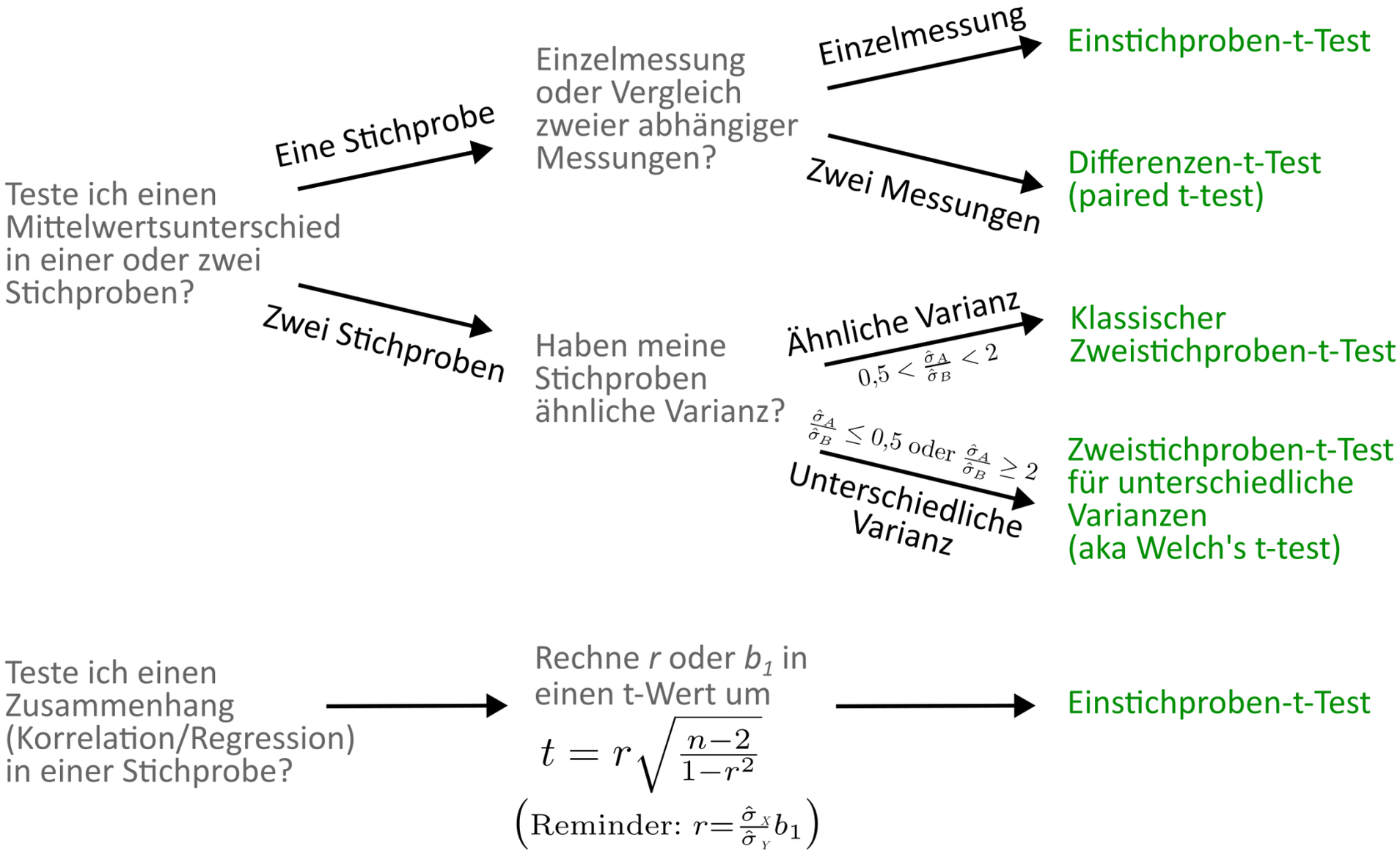
Warum wird also plötzlich eine Normalverteilung des Merkmals X in der Population vorausgesetzt?

Zwei Gründe:

- Damit $t = \hat{\theta} / \hat{se}$ einer t-Verteilung folgt, müssen der Stichprobenkennwert $\hat{\theta}$ und sein Standardfehler $\hat{se}$ unabhängig sein (dürfen nicht korrelieren). Dies ist theoretisch nur erfüllt, wenn das gemessene Merkmal X einer Normalverteilung folgt — tatsächlich lässt sich genau mit dieser Bedingung die Normalverteilung ableiten.
- Der zentrale Grenzwertsatz für die Stichprobenverteilung von $\hat{\theta}$ gilt streng genommen nur für $n \to \infty$, was in empirischen Studien nie gegeben ist. Ist die Stichprobengröße n begrenzt, kann die Stichprobenverteilung von $\hat{\theta}$ nur dann als normalverteilt angenommen werden, wenn auch das gemessene Merkmal X in der Population normalverteilt ist (Satz von Cramer).

Aber wie erwähnt: trotz dieser theoretischen Fallstricke ist der t-Test in der Praxis auch bei nicht-normalverteilten Populationsvariablen X recht robust.

1. Welche Art von t-Test benötige ich?



2. Zahl der Freiheitsgrade

- Die **Zahl der Freiheitsgrade** gibt die **Zahl der frei variierbaren Werte** bei der Berechnung eines statistischen Kennwertes an.
- Häufig kann die Zahl der Freiheitsgrade recht einfach berechnet werden als Stichprobengröße n *minus* die Anzahl der Zwischenparameter, die für die Berechnung des statistischen Kennwertes geschätzt werden müssen.

Beispiel Varianz: Angenommen, wir wollen für einer Stichprobe aus vier Werten die Varianz bestimmen:

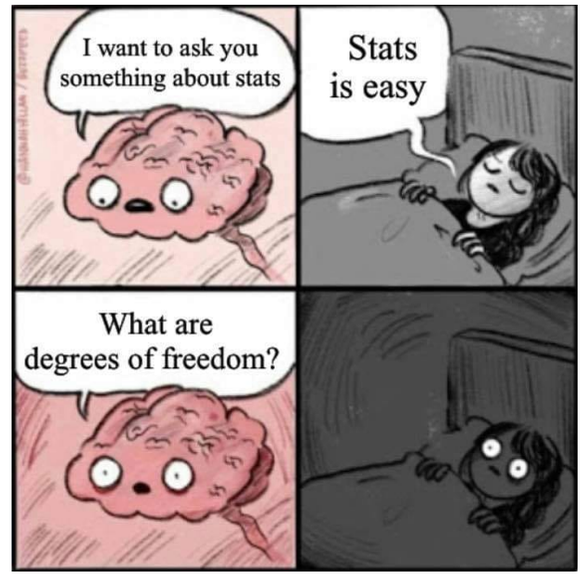
$$\sigma^2 = \frac{1}{4}\sum_{1}^{4}(x_i - \bar{x})^2$$

Aus der Formel der Varianz erkennen wir, dass wir zur Berechnung der Varianz **einen Parameter** aus den Daten bestimmen müssen, nämlich der Mittelwert $\bar{x}$. Die Zahl der Freiheitsgrade ergibt sich damit plain & simple nach obiger Definition als **Stichprobengröße n minus 1**, also in diesem Beispiel $\mathrm{df} = 3$.

Dadurch, dass die Formel der Varianz den Mittelwert aller Datenpunkte als Parameter enthält, können wir nicht mehr alle 4 Datenpunkte beliebig frei variieren. Wir könnten drei beliebige Datenwerte frei bestimmen, der vierte Wert aber wäre durch den Mittelwertparameter automatisch festgelegt. Sagen wir der Mittelwert sei $\bar{x} = 2$ und wir würden frei drei Werte als (2, 4, 1) frei wählen. Der vierte Wert ist nun nicht mehr frei, denn um die Randbedingung $\bar{x} = 2$ zu garantieren, muss der vierte Wert gleich 1 sein. Diese Logik gilt nicht nur für den Mittelwertsparameter, sondern für jeden Parameter der "auf dem Weg" zur Berechnung einer statistischen Größe geschätzt werden muss (bei der Varianz ist der Mittelwert aber der einzige Zwischenparameter).

---

2. Zahl der Freiheitsgrade: t-Test



- Für die Zahl der Freiheitsgrade beim t-Test gilt, dass die Zahl der Freiheitsgrade ausschließlich auf Basis der Streuung $\hat{\sigma}$ im Nenner bestimmt wird.

$$t = \frac{\hat{\theta}}{\hat{\sigma}/\sqrt{n}}$$

- Häufig reicht es, zu zählen, wie viele Mittelwerte für die Berechnung von $\hat{\sigma}$ im Nenner des t-Wertes für die verschiedenen Tests verwendet werden:
  - Einstichproben-t-Test: 1 Mittelwert für 1 Stichprobe → $\mathrm{df} = n - 1$
  - Differenzen-t-Test: 1 Mittelwert für 1 Stichprobe (hier: Mittelwert $\Delta\bar{x}$ der Differenzen $\Delta\bar{x}_i$) → $\mathrm{df} = n - 1$
  - Zweistichproben-t-Test (ähnliche Varianzen): 2 Mittelwerte für 2 Stichproben → $\mathrm{df} = n_A + n_B - 2$
  - (Ausnahme: Zweistichprobentest mit unterschiedlichen Varianzen → Welch–Satterthwaite-Gleichung)
  - Korrelation und Regression: 2 Mittelwerte für die zwei Zusammenhangsvariablen X und Y → $\mathrm{df} = n - 2$

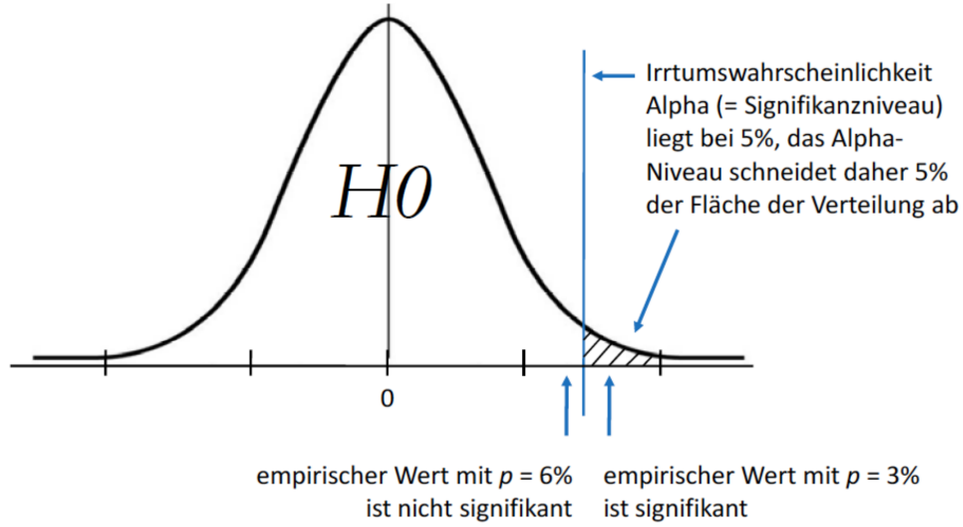
3. Signifikanzniveau & ein/zweiseitige Testung

Erinnerung:

Das **Signifikanzniveau** α gibt legt ein Kriterium für die Ablehnung der Nullhypothese fest.

Es gilt: ist $p < \alpha$ lehnen wir die Nullhypothese ab und werten unseren Effekt als statistisch signifikant.

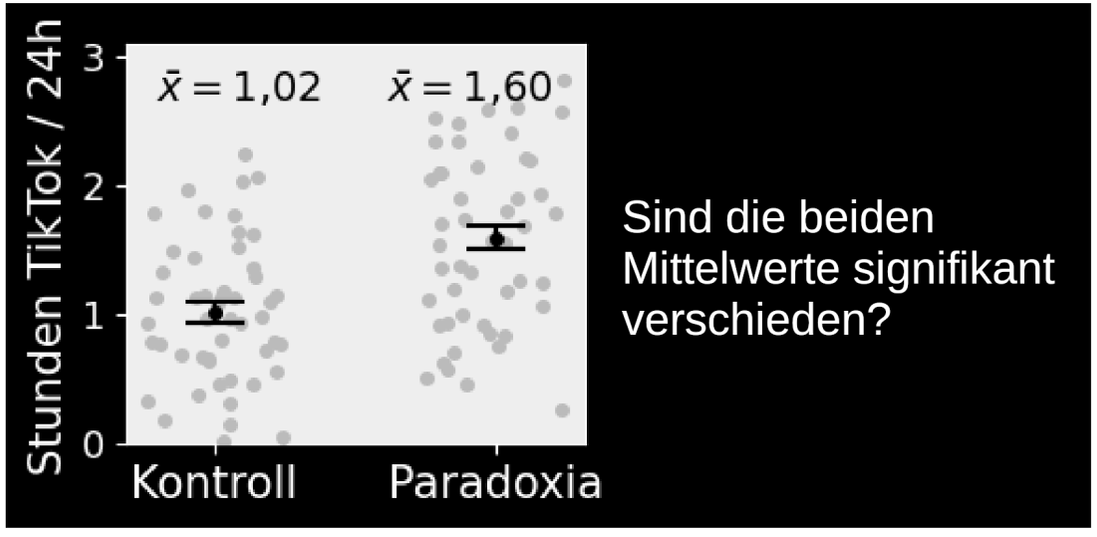
- Der Wert α wird auch Irrtumswahrscheinlichkeit genannt — sie gibt das Risiko an, mit dem wir bereit sind, die Nullhypothese fälschlicherweise abzulehnen.
  - Beispiel: ist $\alpha = 0{,}05$, so gibt es eine 5%-ige Wahrscheinlichkeit, einen signifikanten Effekt zu erhalten *obwohl die Nullhypothese zutrifft*.
  - Je kleiner α, desto "konservativer" der Test, d.h. desto eher will ich vermeiden, einen Effekt fälschlicherweise als signifikant zu werten (erhöhe aber dabei mein Risko, einen wahren Effekt zu "verpassen").



t-Test für Mittelwertdifferenzen

t-Test für Mittelwertdifferenzen

- Zur Erinnerung, bei einem statistischen Test zu Mittelwertdifferenzen stellen wir folgende Fragen:
  - Unterscheiden sich die Mittelwerte von zwei unabhängigen Stichproben signifikant voneinander? (Zweistichproben-Test)
  - Unterscheiden sich die Mittelwerte von zwei abhängigen Bedingungen signifikant voneinander? (Einstichproben-Test mit zwei abhängigen Messungen)
  - Unterscheidet sich ein Mittelwert signifikant von einem fixen Referenzwert? (Einstichproben-Test mit Einzelmessung)
- Beispiel Paradoxia:



t-Test für Mittelwertdifferenzen

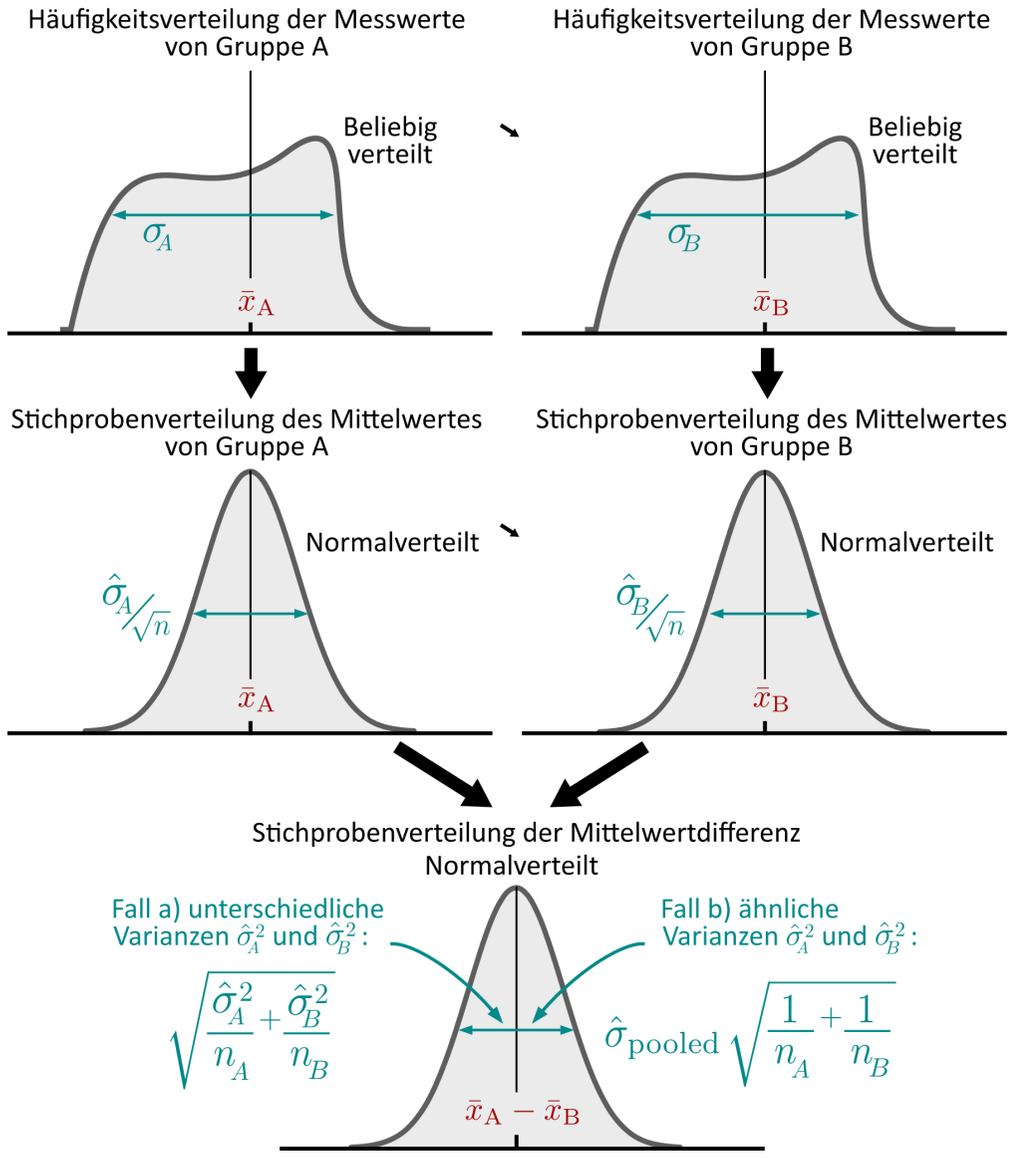
- Um den t-Wert für eine Mittelwertdifferenz $\Delta\bar{x}$ zu erhalten, ersetzen wir $\hat{\theta}$ durch $\Delta\bar{x}$:

$$\text{t-Wert für Mittelwertdifferenz:} \qquad t = \frac{\hat{\theta}}{\hat{se}} = \frac{\Delta\bar{x}}{\hat{se}}$$

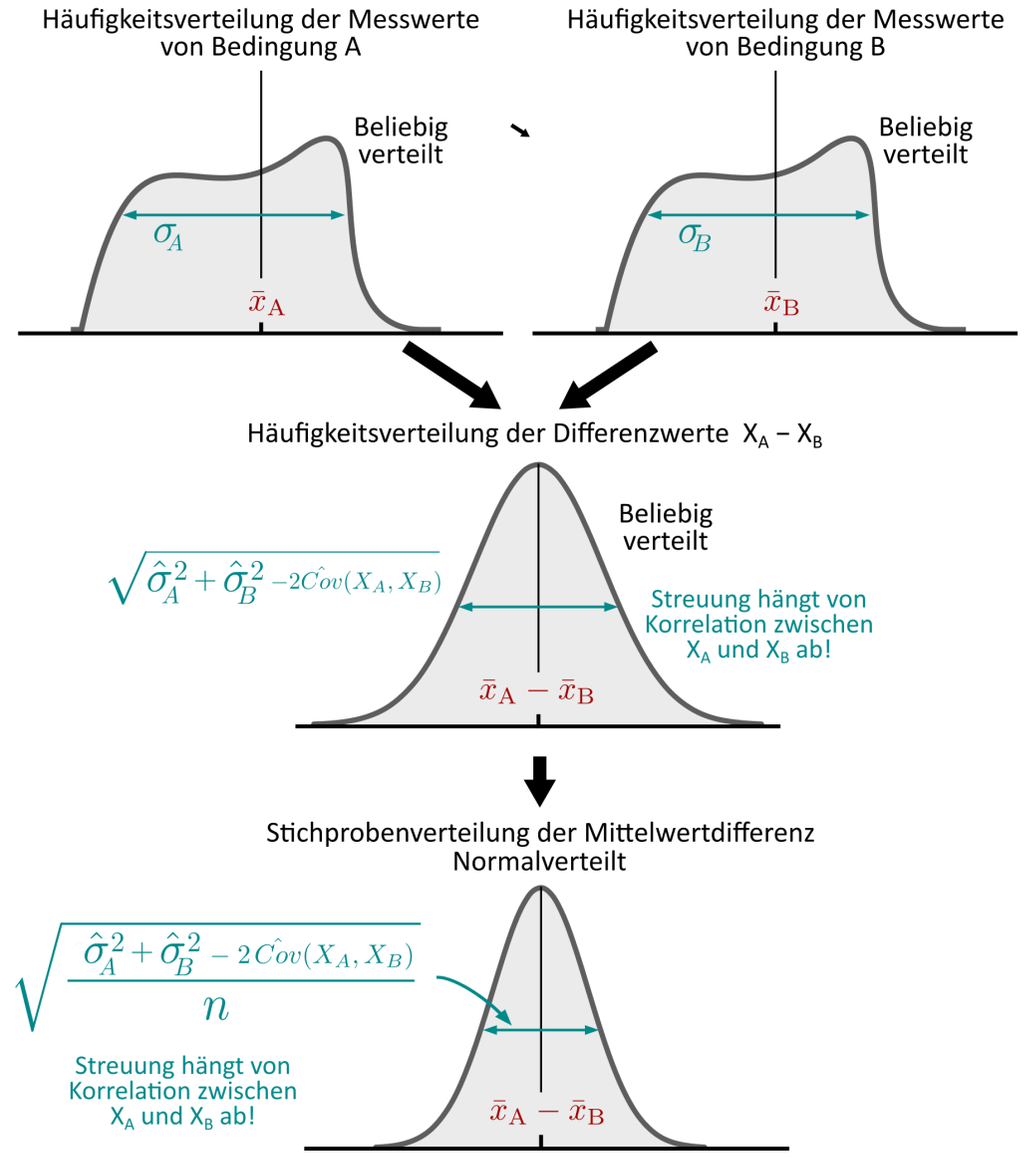
- Die Mittelwertdifferenz $\Delta\bar{x}$ ist einfach zu bestimmen, die zentrale Frage lautet daher: **wie berechnet sich der Standardfehler $\hat{se}$ einer Mittelwertdifferenz $\Delta\bar{x}$?**
- **Merke:** der Standardfehler im Nenner des t-Wertes bezieht sich immer auf die Prüfgröße $\hat{\theta}$. Die Frage lautet also bei jedem t-Test: was ist der Standardfehler von $\hat{\theta}$ bzw. was ist die Standardabweichung der Stichprobenverteilung von $\hat{\theta}$?
- Wie beim z-Test hängt die Berechnung des Standardfehlers von Mittelwertdifferenzen von verschiedenen Faktoren ab.

Standardfehler bei Mittelwertdifferenzen (σ unbekannt)

Mittelwertdifferenz: Unabhängige Messungen



Mittelwertdifferenz: Abhängige Messungen



Standardfehler bei Mittelwertdifferenzen (σ unbekannt): Cheat sheet

Das folgende Cheat sheet gibt Ihnen eine Übersicht über die Berechnung des Standardfehlers $\hat{se}$ von Mittelwertdifferenzen im Nenner der Prüfgröße $t = \frac{\Delta\bar{x}}{\hat{se}}$:

Fall	Berechnung des Standardfehlers $\hat{se}$
Differenz des Mittelwertes einer Stichprobe und einem Referenzwert μ_0 (→ Einstichproben-t-Test)	$\hat{se} = \frac{\hat{\sigma}}{\sqrt{n}}$
Mittelwertdifferenz unabhängiger Messungen in zwei Stichproben A und B (ähnliche Varianzen) (→ Klassischer Zweistichproben-t-Test)	$\hat{se} = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$ mit $\quad \hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_A-1)\hat{\sigma}_A^2 + (n_B-1)\hat{\sigma}_B^2}{n_A+n_B-2}}$
Mittelwertdifferenz unabhängiger Messungen in zwei Stichproben A und B (unterschiedliche Varianzen) (→ Welch's Zweistichproben-t-Test)	$\hat{se} = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$
Mittelwertdifferenz abhängiger Messungen A und B in einer Stichprobe (→ Differenzen-t-Test)	$\hat{se} = \frac{\hat{\sigma}_\Delta}{\sqrt{n}}$ $\quad$ mit $\hat{\sigma}_\Delta = \sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2 - 2\,\hat{Cov}(X_A, X_B)}$ oder $\quad \hat{\sigma}_\Delta = \sqrt{\frac{1}{n-1}(\Delta x_i - \Delta\bar{x})^2}$

Dieses Cheat Sheet gilt für den Fall, dass die Populationsvarianzen σ_A^2 bzw. σ_B^2 **nicht bekannt** sind. In diesem Sinne ist es das in der Praxis relevante Cheat Sheet, da die Populationsvarianzen nahezu nie bekannt sind.

Einschub: gepoolter Standardfehler

- Im Cheat sheet auf der vorherigen Folie ist folgender Standardfehler für den Fall unabhängiger Messungen und ähnlicher Varianzen angegeben:

$$\hat{se} = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

- Zur Herleitung starten wir wieder mit der allgemeinen Varianzsummenformel: wir wissen, dass sich die Varianzen der beiden Stichprobenmittelwerte (=quadrierte Standardfehler $\hat{se}_A^2$ und $\hat{se}_B^2$) aufaddieren:

$$\hat{se}^2 = \hat{se}_A^2 + \hat{se}_B^2 \quad \rightarrow \quad \hat{se} = \sqrt{\hat{se}_A^2 + \hat{se}_B^2} = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$$

- Können die Varianzen $\hat{\sigma}_A^2$ und $\hat{\sigma}_B^2$ als ähnlich angenommen werden, gilt folgende Überlegung: statt beide Varianzen einzeln zu schätzen, wird eine einheitliche **gepoolte Varianz** $\hat{\sigma}_{\text{pooled}}$ auf Basis beider Stichproben geschätzt (vgl. Vorlesung 07).
  - Vorteil: diese Varianz kann präziser geschätzt werden als die Einzelvarianzen, da die kombinierte Stichprobenzahl $n_A + n_B$ zugrundegelegt wird.
- Wir ersetzen also $\hat{\sigma}_A^2$ und $\hat{\sigma}_B^2$ jeweils durch $\hat{\sigma}_{\text{pooled}}^2$:

$$\hat{se} = \sqrt{\frac{\hat{\sigma}_{\text{pooled}}^2}{n_A} + \frac{\hat{\sigma}_{\text{pooled}}^2}{n_B}} = \hat{\sigma}_{\text{pooled}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \qquad \text{q.e.d.}$$

Zahl der Freiheitsgrade für t-Tests von Mittelwertdifferenzen

Test	Frage	Zahl der Freiheitsgrade
Einstichprobentest mit Einzelmessung	Ist $\bar{x}$ größer als ein Referenzwert μ_0?	$\mathrm{df} = n - 1$, da für die Berechnung von t genau ein Zwischenparameter bestimmt werden muss (der Mittelwert $\bar{x}$)
Einstichprobentest mit zwei abhängigen Messungen	Unterscheiden sich die Mittelwerte $\bar{x}_A$ und $\bar{x}_B$ zwischen zwei Bedingungen A und B? In diesem Fall kann man auch fragen: ist der Mittelwert der Differenzvariable ($\Delta\bar{x} = \overline{X_A - X_B}$) verschieden von Null?	$\mathrm{df} = n - 1$, da für die Berechnung von t genau ein Zwischenparameter bestimmt werden muss (der Mittelwert $\overline{X_A - X_B}$)
Zweistichprobentest (ähnliche Varianzen)	Unterscheiden sich die Mittelwerte $\bar{x}_A$ und $\bar{x}_B$ zwischen zwei Gruppen A und B? ($\Delta\bar{x} = \bar{x}_A - \bar{x}_B$)	$\mathrm{df} = n_A + n_B - 2$, da für die Berechnung von t genau zwei Zwischenparameter bestimmt werden müssen (die Mittelwerte $\bar{x}_A$ und $\bar{x}_B$)
Zweistichprobentest (unterschiedliche Varianzen, aka Welch's t-Test)	Unterscheiden sich die Mittelwerte $\bar{x}_A$ und $\bar{x}_B$ zwischen zwei Gruppen A und B? ($\Delta\bar{x} = \bar{x}_A - \bar{x}_B$)	**Welch–Satterthwaite-Gleichung:** $\mathrm{df} = \frac{\left(\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}\right)^2}{\frac{\left(\hat{\sigma}_A^2/n_A\right)^2}{n_A - 1} + \frac{\left(\hat{\sigma}_B^2/n_B\right)^2}{n_B - 1}}$ Falls $n_A = n_B = n$: $\mathrm{df} = (n-1)\left(1 + \frac{2}{\left(\frac{\hat{\sigma}_A}{\hat{\sigma}_B}\right)^2 + \left(\frac{\hat{\sigma}_B}{\hat{\sigma}_A}\right)^2}\right)$

Intuition zur Welch–Satterthwaite-Gleichung

Auch wenn die Welch–Satterthwaite-Gleichung für die Freiheitsgrade beim Zweistichproben-t-Test mit unterschiedlicher Varianz recht kompliziert aussieht, ist eine nähere Betrachtung aufschlussreich.

Für $n_A = n_B = n$ gilt:

$$\mathrm{df} = (n-1)\left(1 + \frac{2}{\left(\frac{\hat{\sigma}_A}{\hat{\sigma}_B}\right)^2 + \left(\frac{\hat{\sigma}_B}{\hat{\sigma}_A}\right)^2}\right)$$

Zwei Fälle sind interessant:

Fall 1: Sind die Varianzen gleich (entgegen der Annahme), d.h. $\hat{\sigma}_A = \hat{\sigma}_B$, so vereinfacht sich die Formel zu

$$\mathrm{df} = (n-1)(1 + \frac{2}{2}) = 2n - 2,$$

also exakt die Formel des klassichen Zweistichprobentests.

Intuition zur Welch–Satterthwaite-Gleichung

Fall 2: Sind die Varianzen extrem unterschiedlich, so wird entweder $\left(\frac{\hat{\sigma}_A}{\hat{\sigma}_B}\right)^2$ oder $\left(\frac{\hat{\sigma}_B}{\hat{\sigma}_A}\right)^2$ extrem groß und die Formel vereinfacht sich zu

$$\mathrm{df} = (n-1)(1+\frac{2}{\infty}) = (n-1)(1+0) = n-1$$

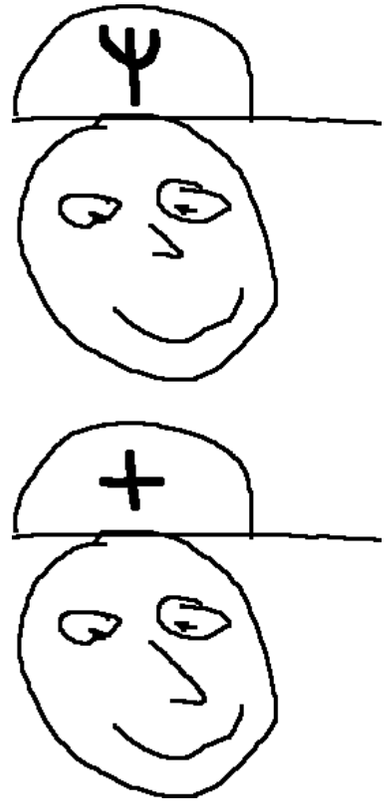
also exakt die Formel des Einstichprobentests.

Für extrem ungleiche Varianzen reduzieren sich also die Freiheitsgrade — und damit die *effektive Stichprobengröße* — auf die Größe einer der beiden verglichenen Gruppen.

Das ist durchaus intuitiv: ist z.B. $\hat{\sigma}_A$ extrem viel kleiner als $\hat{\sigma}_B$, so spielt die Varianz der Gruppe A nahezu keine Rolle mehr. Die Versuchspersonen dieser Gruppe gehen also für die Berechnung des Standardfehlers “verloren”, was sich entsprechend auf die Freiheitsgrade auswirkt.

Im Grenzfall spielt die Varianz dieser Gruppe überhaupt keine Rolle mehr, sondern nur noch ihr Mittelwert. Die Gruppe hat damit die gleiche Funktion wie der Referenzwert μ_0 beim Einstichproben-t-Test.

Beispiel: t-Test für unabhängige Gruppen



Betrachten wir das Beispiel Med- versus Psych-Nasen für den Fall, dass wir die Standardabweichungen σ_{med} und σ_{psych} von Nasenlängen in der Med- und Psych-Population nicht kennen.

Es gilt immer noch:

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

.. aber σ_{med} und σ_{psych} müssen jetzt aus unseren Messwerten geschätzt werden. Für das Beispiel nehmen wir an:

$$\hat{\sigma}_{\text{med}} = 0{,}5 \qquad \hat{\sigma}_{\text{psych}} = 0{,}25$$

Wir schlagen die Formel für den Standardfehler bei unabhängigen Messungen und ungleichen Varianzen nach und setzen ein:

$$\hat{se} = \sqrt{\frac{\hat{\sigma}^2_{\text{med}}}{n_{\text{med}}} + \frac{\hat{\sigma}^2_{\text{psych}}}{n_{\text{psych}}}} \overset{(Computer)}{=} 0{,}102$$

Erinnerung: es galt $n_{\text{med}} = n_{\text{psych}} = 30$.

Beispiel: t-Test für unabhängige Gruppen

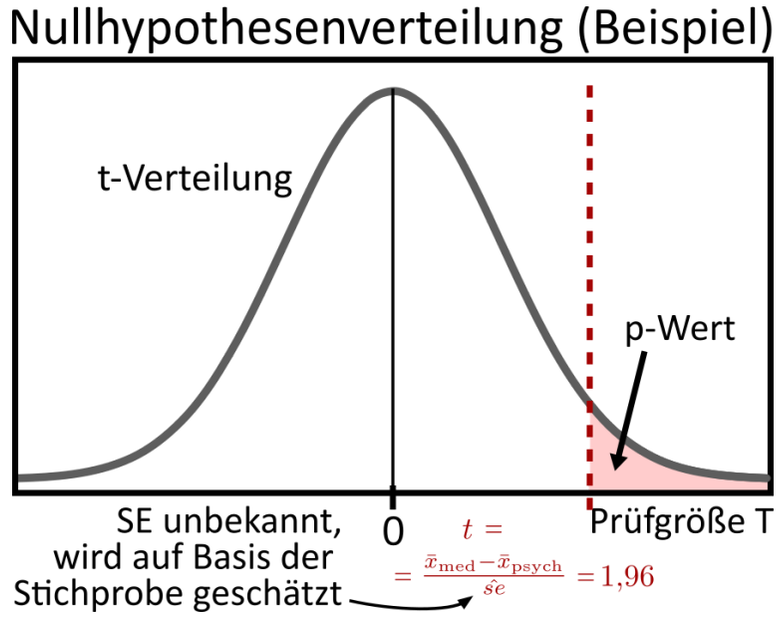
Mit dem Standardfehler bewaffnet, können wir nun den t-Wert berechnen:

$$t = \frac{\Delta \bar{x}}{\hat{se}} = 1{,}96$$

Es fehlen noch die Freiheitsgrade. Bei ungleichen Varianzen ist die Formel recht sperrig:

$$\mathrm{df} = (n-1)\left(1 + \frac{2}{\left(\frac{\hat{\sigma}_{\mathrm{med}}}{\hat{\sigma}_{\mathrm{psych}}}\right)^2 + \left(\frac{\hat{\sigma}_{\mathrm{psych}}}{\hat{\sigma}_{\mathrm{med}}}\right)^2}\right) =$$

$$= (30-1)\left(1 + \frac{2}{\left(\frac{0{,}5}{0{,}25}\right)^2 + \left(\frac{0{,}25}{0{,}5}\right)^2}\right) = 42{,}6$$



Aufgrund der gerichteten (positiven) Hypothese ist der p-Wert die Fläche *rechts des t-Wertes* unter der t-Verteilung:

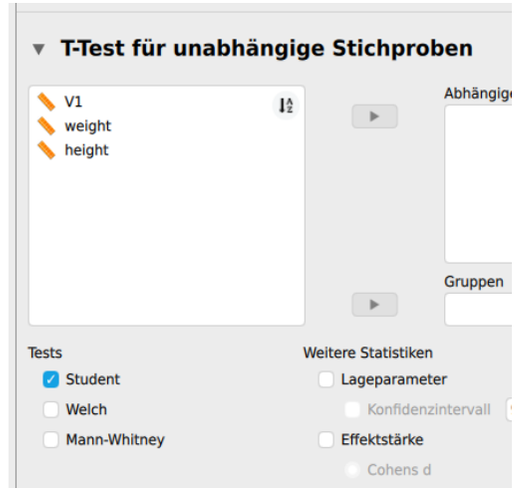
$$p = \int_t^{\infty} f_t(x|\mathrm{df})dx = \int_{1{,}96}^{\infty} f_t(x|42{,}6)dx = 1 - F_t(1{,}96|42{,}6) \overset{(Computer)}{=} 0{,}028$$

Der p-Wert ist dem p-Wert des z-Tests ($0{,}023$) sehr ähnlich. Dies ist wenig überraschend, da a) der Standardfehler einen ähnlichen Wert aufwies ($se = 2$ beim z-Test) und b) die Zahl er Freiheitsgrade so hoch ist, dass der Unterschied Normalverteilung vs. t-Verteilung marginal ist.

t-Test für unabhängige Stichproben

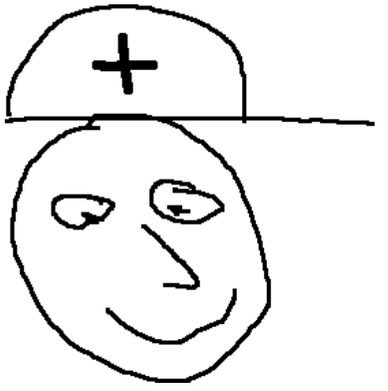
Klassischer Zweistichproben-t-Test (ähnliche Varianzen) versus Welch's t-Test (unterschiedliche Varianzen) — welcher der beiden Tests sollte nun verwendet werden?

- In der Psychologie sind ungleiche Varianzen die realistischere Annahme. Selbst bei randomisiert-kontrollierten Studien, bei denen Versuchspersonen den Gruppen aus einer identischen Ursprungs-Population zugewiesen werden, kann das Treatment selbst einen Einfluss auf die Varianz haben.
- Neben der Faustformel $0{,}5 < \frac{\hat{\sigma}_A}{\hat{\sigma}_B} < 2$ gibt es auch Testverfahren zur Überprüfung der Varianzgleichheit zwischen Gruppen, allerdings sind diese nicht unproblematisch
  - Ein Grund:auf Varianzgleichheit wird idR auf Basis eines nicht-signifikanten Ergebnis in einem solchen Test angenommen — "absence of evidence is not evidence of absence"[1]
- Herrscht Varianzgleichheit vor und ist die Stichprobengröße nicht extrem klein in einer Gruppe, kommen der Student'sche t-Test und Welch's t-Test zu sehr ähnlichen Ergebnissen.
- Vor diesem Hintergrund wird empfohlen[2], bereits ab moderaten Gruppengrößen (ca. $n \geq 8$ pro Gruppe) *immer* Welch's t-Test anzuwenden. In den meisten Statistikprogrammen ist dieser Test implementiert.



Option für Welch's t-test in JASP.

Beispiel: t-Test für abhängige Messungen



Sie führen ein Experiment *innerhalb der Med-Gruppe* (n=32) durch. In einer experimentellen Intervention stellen Sie den Med-Studierenden die Frage

> Studieren Sie Medizin, weil es der Wunsch Ihrer Eltern ist?

Sie messen die Nasenlängen dabei sowohl vor ("pre"), als auch nach ("post") der Intervention. Ihre ungerichtete Hypothese ist, dass sich die Nasenlängen vor und nach der Intervention auf einem Signifikanzniveau $\alpha = 0{,}05$ unterscheiden.

Die Mittelwertdifferenz betrage $\Delta\bar{x} = \bar{x}_{\text{post}} - \bar{x}_{\text{pre}} = 0{,}1cm.$

Wir nehmen der Einfachheit halber an, dass die beiden Messungen X_{pre} und X_{post} unkorreliert sind, d.h. $\hat{Cov}(X_{\text{pre}}, X_{\text{post}}) = 0$. Die Standardabweichungen seien $\hat{\sigma}_{\text{pre}} = \hat{\sigma}_{\text{post}} = 0{,}5$. Die Standardabweichung $\hat{\sigma}_\Delta$ der Differenzvariable ΔX ist damit:

$$\hat{\sigma}_\Delta = \sqrt{\hat{\sigma}^2_{\text{pre}} + \hat{\sigma}^2_{\text{post}} - 2\,\hat{Cov}(X_{\text{pre}}, X_{\text{post}})} = \sqrt{(0,5)^2 + (0,5)^2 - 2 \cdot 0} = \frac{1}{\sqrt{2}}$$

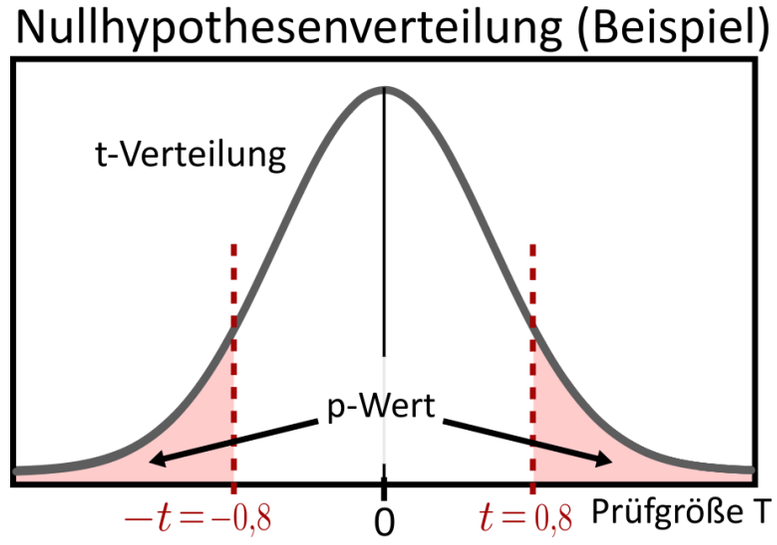
$$\text{Standardfehler:} \qquad \hat{se} = \frac{\hat{\sigma}_\Delta}{\sqrt{n}} = \frac{1}{\sqrt{2}\sqrt{32}} = \frac{1}{8}$$

Beispiel: t-Test für abhängige Messungen

Nun können wir den t-Wert berechnen:

$$t = \frac{\Delta\bar{x}}{\hat{se}} = \frac{0{,}1}{1/8} = 0{,}8$$

Aufgrund der ungerichteten Hypothese ist der p-Wert die Summe der Flächen unter der t-Verteilung links von $t = -0{,}8$ und rechts von $t = +0{,}8$. Bei $\mathrm{df} = n - 1 = 31$ Freiheitsgraden gilt:



$$p = \int_{-\infty}^{-t} f_t(x|\mathrm{df})dx + \int_{t}^{\infty} f_t(x|\mathrm{df})dx \overset{(Symmetrie)}{=} 2 \cdot \int_{-\infty}^{-t} f_t(x|\mathrm{df})dx =$$

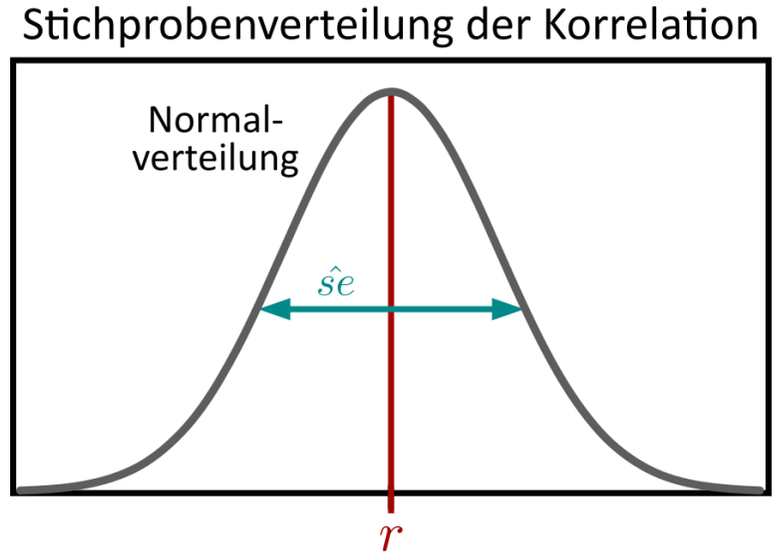
$$= 2 \cdot F_t(-t|\mathrm{df}) \overset{(einsetzen)}{=} 2 \cdot F_t(-0{,}8|31) \overset{(Computer)}{=} 0{,}43$$

Der p-Wert ist also deutlich größer als unser Signifikanzniveau $\alpha = 0{,}05$ und wir können die Nullhypothese $\Delta\bar{x} = 0$ nicht ablehnen. More research is needed!

t-Tests für Zusammenhänge

Stichprobenverteilung der Korrelation

- Analog zu Mittelwertsunterschieden kann auch die **Stichprobenverteilung der Korrelation** aufgestellt werden: als **Normalverteilung** mit dem **Mittelwert unserer Kennwertschätzung (hier r)** und einer Streuung, die dem **Standardfehler** ($\hat{se}$) des Kennwertes entspricht.



- Den Standardfehler der Korrelation haben wir bereits in Vorlesung 08 kennengelernt:

$$\hat{se}(r) = \sqrt{\frac{1 - r^2}{n - 2}}$$

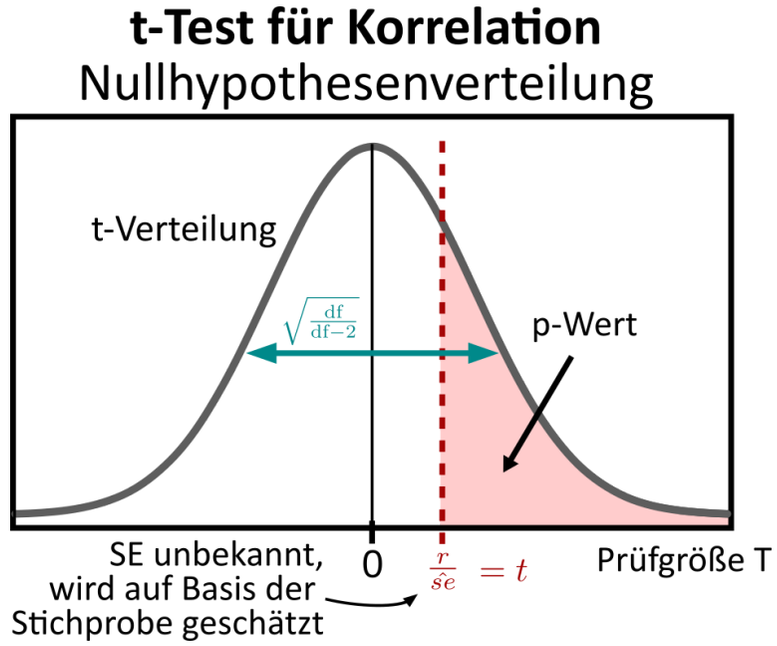
- Auch dieser Standardfehler ist eine Schätzung auf Basis der Stichprobe und entsprechend stellt die standardisierte Prüfgröße einen t-Wert dar:

$$t = \frac{r}{\hat{se}(r)} = r\sqrt{\frac{n - 2}{1 - r^2}}$$

(Präziser gesagt handelt es sich auch hier wieder um das Verhältnis einer normalverteilten Variable (r) und einer chi-verteilten Variable ($\hat{se}(r)$) — ein solches Verhältnis führt zu einer t-verteilten Prüfgröße)

Nullhypothese und Hypothesentestung bei Korrelationen

- Die **Nullhypothesenverteilung** der Korrelation ist also eine **t-Verteilung**.
- Die Nullhypothese entspricht der Annahme, dass der wahre Zusammenhang in der Population $\rho = 0$ ist.
- Auch hier können wir gerichtete und ungerichtete Hypothesen testen:
  - Gerichtete Hypothesen:
    - die Korrelation ist größer 0 ($\rho > 0$)
    - die Korrelation ist kleiner 0 ($\rho < 0$)
  - Ungerichtete Hypothese: die Korrelation ist ungleich 0 ($\rho \neq 0$)



- Die Berechnung der p-Werte erfolgt analog wie bei Mittelwertunterschieden.

Die Zahl der Freiheitsgrade der Korrelation ist $\mathrm{df} = n - 2$. Grund: die Streuung $\hat{se}$ im Nenner des t-Wertes enthält den Korrelationskoeffizient r und dieser involviert die Berechnung von **zwei** Mittelwerten (Mittelwert $\bar{x}$ der Variable X und Mittelwert $\bar{y}$ der Variable Y).

Beispiel: Bestimmung des p-Wertes bei der Korrelation

- Nehmen wir an, die Korrelation von Größe und Gewicht in einer Stichprobe von $n = 12$ betrage $r = 0{,}4$.
- Wir wollen testen, ob die Korrelation auf einem Signifikanzniveau $\alpha = 0{,}05$ signifikant größer als 0 ist.
- Berechnung von der Prüfgröße t:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}4\sqrt{\frac{12-2}{1-0{,}4^2}} \overset{(Computer)}{=} 1{,}38$$

- Wir sehen in der Tabelle, dass der kritische t-Wert für $\mathrm{df} = n - 2 = 10$ Freiheitsgrade bei einem (einseitigen!) Signifikanzniveau von $\alpha = 0{,}05$ gleich $1{,}813$ beträgt.
- Die Nullhypothese wird also nicht abgelehnt, und der Effekt als nicht signifikant gewertet.

Tabelle der t-Verteilung

	Fläche				
df	0,85	0,9	0,95	0,975	0,99
1	1,964	3,078	6,314	,2,706	31,821
2	1,386	1,886	2,92	4,303	6,965
3	1,25	1,638	2,353	3,182	4,541
4	1,19	1,533	2,132	2,776	3,747
5	1,156	1,476	2,015	2,571	3,365
6	1,134	1,44	1,943	2,447	3,143
7	1,119	1,415	1,895	2,305	2,998
8	1,108	1,397	1,86	2,306	2,896
9	1,1	1,383	1,833	2,262	2,821
10	1,093	1,372	1,813	2,228	2,764
30	1,055	1,31	1,697	2,042	2,459
40	1,05	1,303	1,684	2,021	2,423
60	1,046	1,296	1,071	1,997	2,39
120	1,041	1,289	1,658	1,98	2,358
∞	1,039	1,282	1,645	1,96	2,326

Beispiel: Bestimmung des p-Wertes bei der Korrelation

- Statistische Programme geben den p-Wert in der Regel automatisch mit an und es ist somit ersichtlich, ob es sich um eine statistisch signifikante Korrelation handelt.
- **Standardmäßig** handelt es sich dabei immer um einen **Test für eine ungerichtete Hypothese**.
- War die Hypothese dagegen gerichtet, gilt:
  - Halbiere den p-Wert der ungerichteten Hypothese, falls das Vorzeichen von r in Richtung der Hypothese ist.
  - Verdoppele den p-Wert der ungerichteten Hypothese, falls das Vorzeichen von r in gegensätzlicher Richtung der Hypothese ist.
  - Im Beispiel wäre also der Korrelationskoeffizient $2 \times 0{,}198 = 0{,}396$, falls wir einen negativen Zusammenhang vorhergesagt hätten, und $0{,}198/2 = 0{,}099$, falls wir einen positiven Zusammenhang vorhergesagt hätten.

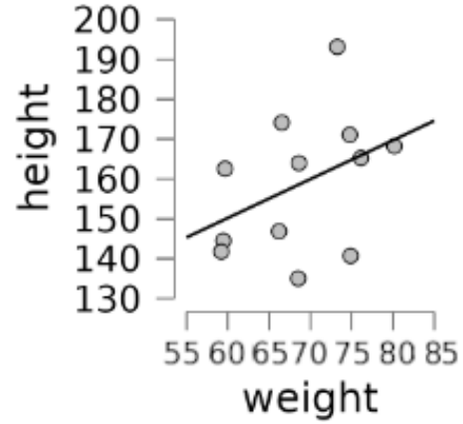
Korrelation

```
jaspRegression::Correlation(
    version = "0.17.2",
    scatterPlot = TRUE,
    variables = list("height", "weight"))
```

Pearsons Korrelationen

Variable		height	weight
1. height	Pearsons r	—	
	p-Wert	—	
2. weight	Pearsons r	0.400	—
	p-Wert	0.198	—

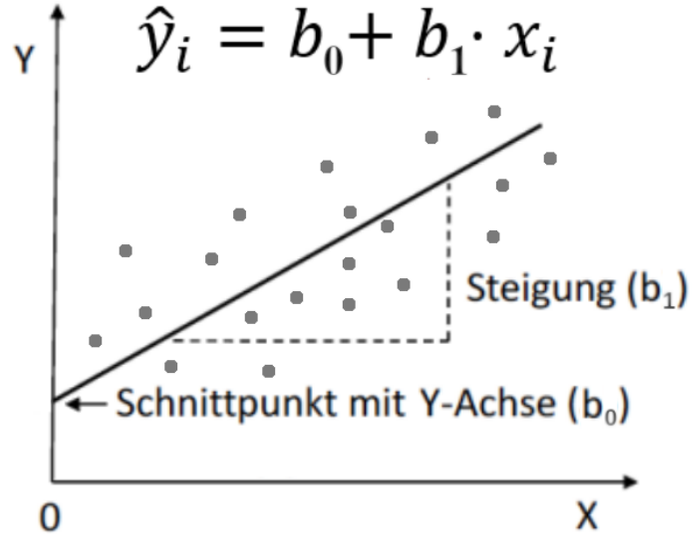
Streudiagramm



Auswahl in JASP: Regression → Klassisch → Korrelation. Zusätzliche Selektion des Streudiagramms.

Inferenzstatistik für Regressionskoeffizienten

- Die Hypothesentestung für den Regressionskoeffizienten der einfachen Regression (d.h. 1 UV) unterscheidet sich nicht von der Korrelation.
- Auch für den Regressionskoeffizienten können wir einen Standardfehler definieren:



$$\hat{se}(b_1) = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\sqrt{\frac{1-r^2}{n-2}}$$

- .. und einen darauf basierenden t-Wert:

$$t = \frac{b_1}{\hat{se}(b_1)} \overset{\text{(s. nächste Folie)}}{=} r\sqrt{\frac{n-2}{1-r^2}}$$

- Die Zahl der Freiheitsgrade ist wie bei der Korrelation $\mathrm{df} = n - 2$, da auch hier für die Berechnung des Standardfehlers die beiden Mittelwerte $\bar{x}$ und $\bar{y}$ bestimmt werden müssen.
- Der Rest ist bekannt.

Inferenzstatistik für Regressionskoeffizienten

- Während die Regressionssteigung abhängt davon, welche Variable als UV (bzw. x) und welche als AV (bzw. y) definiert wird, sind die t-Werte (und damit auch die p-Werte) unabhängig von der Rollenverteilung der Variablen.
- Dies ergibt sich direkt aus dem Zusammenhang von r und b_1 (vgl. Vorlesung 06):

$$b_1 = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} r$$

- Damit ist die Prüfgröße t:

$$t = \frac{b_1}{\hat{se}(b_1)} = \frac{\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} r}{\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \sqrt{\frac{1-r^2}{n-2}}} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}$$

- Der t-Wert bei der einfachen Regression ist also identisch zum t-Wert der Korrelation und insbesondere nur noch abhängigg von r und nicht mehr b_1
- Da für die Korrelation r die Rollenverteilung der beiden Zusammenhangsvariablen unerheblich ist, folgt, dass der p-Wert bei der Regression ebenso wenig von der Zuordnung der Variablen als UV und AV abhängt.

Testung des y-Achsenabschnitts bei der Regression

- Wir haben bislang den Achsenabschnitt b_0 aus der Diskussion außen vor gelassen.
- Im Kontext der Regression wird der Achsenabschnitt b_0 idR nicht getestet.
- Grund: die Frage, ob Variable Y bei $X = 0$ einen y-Achsenabschnitt aufweist, der signifikant von Null verschieden ist, ist im Kontext der Regression selten interessant — schließlich ist es ja der ganze Zweck der Regression die systematische Verändeurng von Y in X zu analysieren und dabei gerade nicht nur einen bestimmten Wert von X zu betrachten.
- Nichts desto trotz ist auch die Schätzung von b_0 mit Unsicherheit verbunden, die durch folgenden Standardfehler definiert ist:

$$\hat{se}(b_0) = \underbrace{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}_{\text{Standardabweichung der Residuen}} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

Testung des y-Achsenabschnitts bei der Regression

- Ein Spezialfall ist das lineare Modell ohne Steigung (bzw. ohne x!):

$$\hat{y} = b_0$$

- In diesem Modell kommt dem Regressionskoeffizienten b_0 und dessen Unsicherheit eine interessantere Bedeutung zu: die Frage ob b_0 signifikant verschieden von 0 ist, ist hier gleichbedeutend mit der Frage ob die Zufallsvariable Y **signifikant verschieden von Null ist**.
- Mit anderen Worten: dieses Modell ist nichts anderes als ein Einstichproben-t-Test!
- Dies beinhaltet auch den Vergleich zweier abhängiger Messungen A und B, wenn wir zuvor Y als Differenzvariable definieren: $Y = Y_A - Y_B$
- .. dann testet

$$\hat{y} = b_0$$

- die Frage, ob die Mittelwertsdifferenz von A und B signifikant verschieden von Null ist.

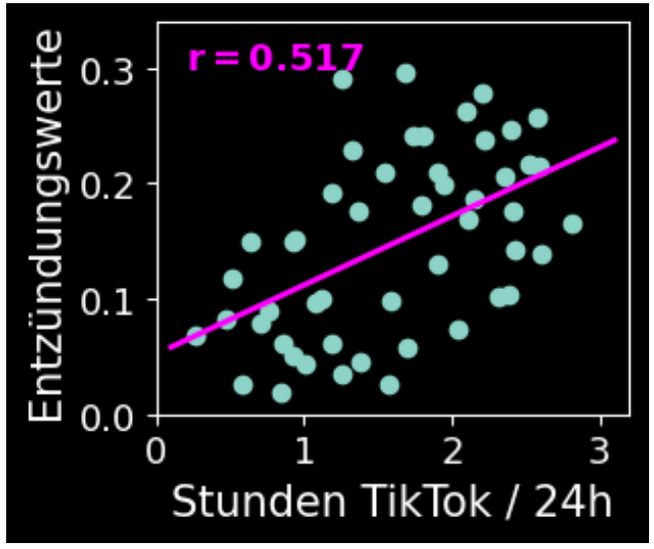
“Common statistical tests are linear models”

Die Parallele von bekannten statistischen Tests und (generalisierter) linearer Regression lässt sich auf alle Tests erweitern, die wir in Statistik 1 und Statistik 2 kennenlernen.

Hier eine Auswahl der Tests aus Statistik 1:

Test	Lineares Modell	Spezifierung	Nullhypothese
Pearson-Korrelation	$\hat{y} = \beta_0 + \beta_1 x$	-	$\mathcal{H}_0 : \beta_1 = 0$
Spearman-Korrelation	$\text{Rang}(\hat{y}) = \beta_0 + \beta_1 \text{Rang}(x)$	-	$\mathcal{H}_0 : \beta_1 = 0$
Einstichproben-t-Test	$\hat{y} = \beta_0$	-	$\mathcal{H}_0 : \beta_0 = 0$
Differenzen-t-Test	$\hat{y_A} - \hat{y_B} = \beta_0$	-	$\mathcal{H}_0 : \beta_0 = 0$
Zweistichproben-t-Test	$\hat{y} = \beta_0 + \beta_1 x$	$x = 0$ für alle Datenpunkte von Gruppe A und $x = 1$ für alle Datenpunkte von Gruppe B → Regression auf binäre x-Variable!	$\mathcal{H}_0 : \beta_1 = 0$

Die Analogie von bekannten statistischen Tests und linearem Modell ist lange bekannt, ging aber 2021 durch Beiträge von Jonas Lindeløv viral[3] [4] [5].

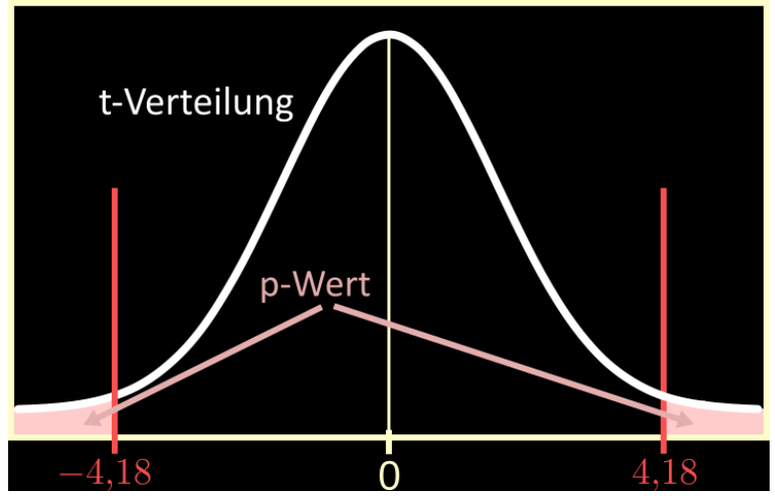


Stichwort Signifikanz von Zusammenhängen: die Signifikanz des unerwarteten Zusammenhangs von TikTok-Onlinezeit und Entzündungswerten haben Sie bislang nicht getestet.

Basierend auf dem Korrelationskoeffizienten $r = 0{,}517$ und der Stichprobengröße $n = 50$ ergibt sich die Prüfgröße t direkt:

$$t = \frac{r}{\hat{se}(r)} = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}517\sqrt{\frac{50-2}{1-0{,}517^2}} = 4{,}18$$

Der Zusammenhang war unerwartet und es gab dementsprechend keine gerichtete Hypothese. Der p-Wert ist in diesem Fall also der doppelte Wert der Fläche rechts des (positiven) t-Werts von $4{,}18$, oder analog, der doppelte Wert links des negierten t-Werts $-4{,}18$.



$$p = 2 \int_{-\infty}^{-4{,}18} f_t(x|\mathrm{df})dx = 2F_t(-4{,}18|48) = 0{,}0001$$

(Die Zahl der Freiheitsgrade beim t-Test der Korrelation ist $n - 2$)

Der Zusammenhang von TikTok-Onlinezeit und Entzündungswerten ist also deutlich signifikant.

Trotz des hochsignifikanten Effektes bleiben Sie skeptisch — Ihr größtes Fragezeichen: was ist Ursache, was ist Wirkung? Folgen erhöhte Entzündungswerte tatsächlich auf TikTok-Konsum (einschlägiger Kanäle)? Oder ist es einfach so, dass Erkrankte (mit erhöhten Entzündungswerten) Rat und Solidarität auf TikTok suchen?

Um die Gruppen*unterschiede* nun ebenfalls mithilfe des t-Tests zu überprüfen, listen Sie nochmals alle relevanten Werte auf, die Sie zum Teil bereits beim z-Test bestimmt hatten:

	Fallzahl n	$\Delta\bar{x}$	**Standardfehler** $\hat{se}$	**Freiheitsgrade** df
TikTok	2×50	$0{,}577$	$0{,}123$	$93{,}7$
Entzündung	2×50	$0{,}0243$	$0{,}0147$	$96{,}2$

- Neu hinzugekommen sind die Freiheitsgrade, die Sie in der Tabelle mit der Formel für ungleiche Varianzen bestimmt haben:

$$\mathrm{df} = (n-1)\left(1 + \frac{2}{\left(\frac{\hat{\sigma}_{\text{control}}}{\hat{\sigma}_{\text{paradoxia}}}\right)^2 + \left(\frac{\hat{\sigma}_{\text{paradoxia}}}{\hat{\sigma}_{\text{control}}}\right)^2}\right)$$

- Zur Erinnerung:
  - TikTok: $\hat{\sigma}_{\text{control}} = 0{,}55; \hat{\sigma}_{\text{paradoxia}} = 0{,}68$
  - Entzündung: $\hat{\sigma}_{\text{control}} = 0{,}068; \hat{\sigma}_{\text{paradoxia}} = 0{,}078$

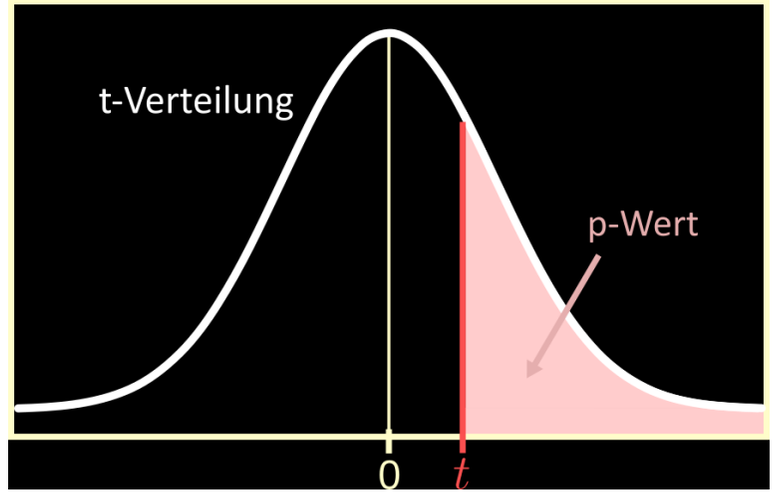
Mit diesen Information erhalten Sie folgende t-Werte (* in diesem Fall identisch mit den z-Werten!):

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

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

Statt der Standardnormalverteilung integrieren Sie nun einfach die entsprechende Fläche unter der t-Verteilung:

$$p = \int_t^{\infty} f_t(x|\mathrm{df})dx = 1 - F_t(t|\mathrm{df})$$



(*) Warum identisch? Weil wir für die z-Werte — in Abwesenheit anderweitig bekannter Werte — ebenfalls die Stichprobenstreuung verwendet hatten.

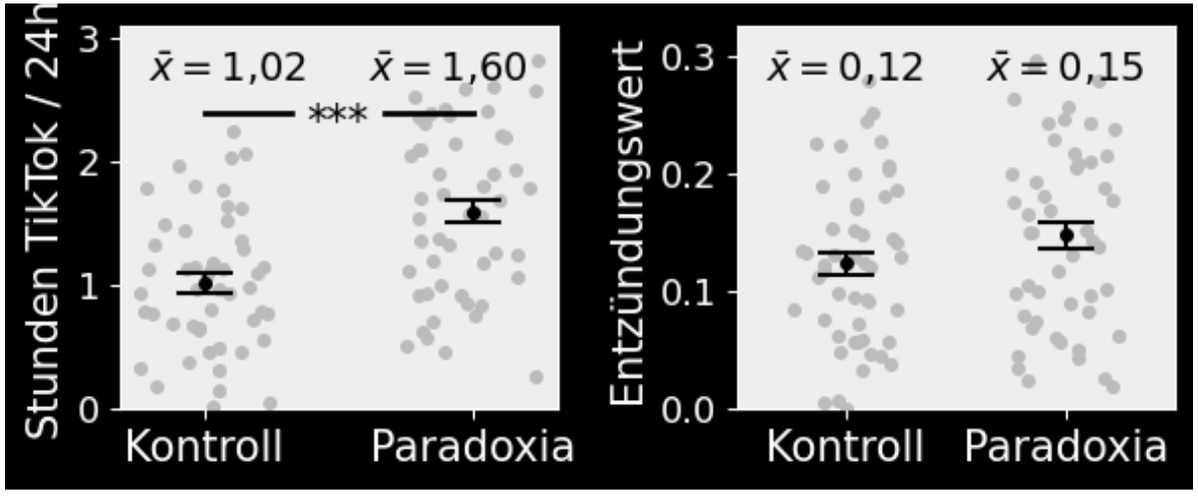
Die große Frage ist: würden weiterhin beide Effekte signifikant bleiben? Antwort:

$$\text{TikTok:} \qquad p = 1 - F_t(4{,}83|93{,}7) \overset{\text{(Computer)}}{=} 0{,}000003$$

$$\text{Entzündung:} \qquad p = 1 - F_t(1{,}65|96{,}2) \overset{\text{(Computer)}}{=} 0{,}051$$

Kurios — das Pendel beim Entzündungseffekt schlägt exakt auf der anderen Seite des Signifikanzniveaus aus! Der Effekt ist nun nicht mehr signifikant. Der TikTok-Effekt bleibt stabil.

Bei beiden Effekten zeigt sich die leicht konservativere Natur der t-Verteilung. Die stärkeren Flanken im Vergleich zur Normalverteilung führen zu geringfügig höheren p-Werten. Notgedrungen müssen Sie Ihre zentrale Grafik anpassen und das Signifikanzsternchen beim Entzündungseffekt entfernen:



Vor dem Hintergrund dieser neuen Ergebnisse, stellen sich eine Reihe von Fragen:

- Ist die Interpretation der Signifikanz fundamental verschieden zwischen dem z- und dem t-Test?
- Ist die Entzündungshypothese eindeutig widerlegt?
- Können Sie aus dem hochsignifikanten Effekt der TikTok-Hypothese und dem nicht-signifikanten Effekt der Entzündungshypothese schlussfolgern, dass der TikTok-Effekt signifikant stärker ist?
- Können Sie schlusfolgern, dass Paradoxia eindeutig durch den TikTok-Effekt verursacht wird?

Fußnoten

1. Altman DG, Bland JM (1995) Absence of evidence is not evidence of absence. BMJ 311:485.
2. Delacre M, Lakens D, Leys C (2017) Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. International Review of Social Psychology 30:92.
3. https://stats.stackexchange.com/questions/303269/common-statistical-tests-as-linear-models
4. https://lindeloev.github.io/tests-as-linear/
5. https://twitter.com/jonaslindeloev/status/1110907133833502721