
**Forschungsmethodik:
Interventions- und Evaluationsforschung
theoretisch begründen**

Vorlesung Biostatistik, WS 2026/2027

Dr. Marcus Wagner, IMISE

Vorlesung III am 27.10.2026

Inhaltsverzeichnis

5. Statistischer Hypothesentest nach Neymar/Pearson

5.1. Grundgedanken

5.2. Interpretation

6. Beziehungen zwischen metrischen Variablen, parametrische Tests

6.1. Mittelwertvergleich für normalverteilte Variablen

6.2. Effektschätzung mit Cohen's d

6.3. Verteilung der Stichproben-Mittelwertdifferenzen

6.4. Der t -Test für unabhängige Stichproben

6.5. Erweiterungen und Zusätze

6.6. Der t -Test für verbundene Stichproben

6.7. Zusammenfassung: Alle Fragestellungen

Wiederholung: Grundkonzepte der Statistik

- **Zweites Grundkonzept:** *In einer Gesamtpopulation $\mathcal{P}$ gibt es objektiv bestimmbare Merkmale, die wir aus praktischen Gründen nicht für jede Person ermitteln können. Die vollständige Information ist nur für eine Stichprobe $S \subset \mathcal{P}$ verfügbar. Können wir (ggf. mit Hilfe eines statistischen Hypothesentests) aus den Merkmalen der Stichprobe auf die Merkmale der Gesamtpopulation schließen?*
- **Drittes Grundkonzept:** *In zwei Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$ gibt es objektiv bestimmbare Merkmale; Informationen besitzen wir aber nur für zwei Stichproben $S' \subset \mathcal{P}'$ und $S'' \subset \mathcal{P}''$, die jeweils aus einer der beiden Populationen kommen. Zum Merkmalsvergleich führen wir anhand der Stichproben einen statistischen Hypothesentest durch. Können wir das Ergebnis des Vergleichs auf die Gesamtpopulationen übertragen?*
- **Antwort in beiden Fällen:** *Ja, aber nur mit begrenzter Tragweite. Die prinzipielle Unsicherheit bei der Auswertung eines statistischen Hypothesentests läßt sich allerdings — wieder mit Begriffen der Wahrscheinlichkeitstheorie — zahlenmäßig beschreiben.*

Hypothesentest mit zwei Stichproben: Grundlagen

- **Ausgangssituation:** *Wir untersuchen zwei Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$.*
- **Zwei Hypothesen:** *Über $\mathcal{P}'$ und $\mathcal{P}''$ werden zwei entgegengesetzte Aussagen (Hypothesen) aufgestellt,*
 - *die “Nullhypothese” $\mathcal{H}_0$, und*
 - *ihr exaktes Gegenteil, die “Alternativhypothese” $\mathcal{H}_A$ (logische Verneinung von $\mathcal{H}_0$).*
- **Ziel des Hypothesentests:** *Es besteht darin, die Nullhypothese $\mathcal{H}_0$ (die “zunichte zu machende Hypothese”) mit einer bestimmten Sicherheit zu widerlegen. Achtung: Ein statistischer Hypothesentest kann **niemals** verwendet werden, um die Nullhypothese $\mathcal{H}_0$ zu beweisen — egal, wie er ausgeht. Im günstigsten Fall liefert der Test Evidenz gegen die Nullhypothese.*
- **Sprachregelung:** *“Der Hypothesentest wird gegen die Nullhypothese $\mathcal{H}_0$ durchgeführt.”*

Hypothesentest mit Stichproben: Grundlagen (II)

- **Fünf Zutaten zum Hypothesentest:**

1) Zwei Stichproben S', S'' aus $\mathcal{P}', \mathcal{P}''$.

2) Das Signifikanzniveau α , eine Zahl zwischen Null und Eins.

3) Eine Prüfgröße $\hat{T}$, die aus den beiden Stichproben unter der Annahme berechnet wird, daß die Nullhypothese $\mathcal{H}_0$ über $\mathcal{P}'$ und $\mathcal{P}''$ richtig ist.

4) Die stetige Wahrscheinlichkeitsverteilung für die Werte von $\hat{T}$, die sich ergibt, wenn man immer wieder Stichproben vom gleichen Umfang zieht und die Prüfgröße $\hat{T}$ jedesmal neu berechnet. Die Dichte f dieser Verteilung muß bekannt sein.

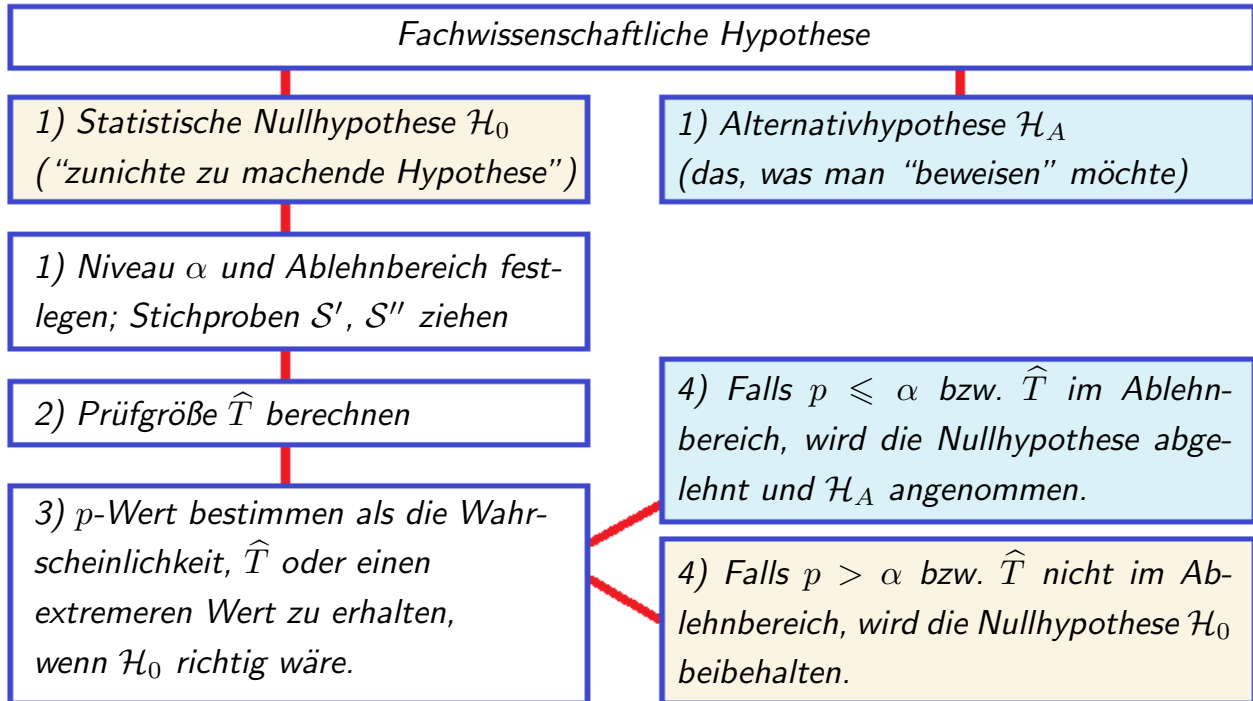
5) Der Ablehnbereich des Tests: Der Wertebereich für $\hat{T}$ wird so zerlegt, daß die Gesamtfläche Eins unter der Dichte f im Verhältnis von $\alpha : (1 - \alpha)$ aufgeteilt wird. Die Gestalt des Ablehnbereichs (einfach oder zweiteilig) hängt von $\mathcal{H}_0$ ab.

- **Anmerkung:** Hinter 3) und 4) stehen mathematische Aussagen ähnlich wie Lehrsatz 2.9. Daraus entstehen technische Voraussetzungen für die Anwendbarkeit des Tests.

Hypothesentest mit Stichproben: Ablauf

- **Schritt 1: Vorbereitung.** Formulierung von Null- und Alternativhypothese $\mathcal{H}_0$ und $\mathcal{H}_A$, Festlegung des Signifikanzniveaus α und des Ablehnbereichs für $\hat{T}$. Prüfung der technischen Voraussetzungen.
- **Schritt 2: Berechnung.** Aus den Stichproben S' , S'' wird die Prüfgröße $\hat{T}$ berechnet, und zwar unter der Annahme, daß die Nullhypothese zutrifft (die Formeln sind bekannt).
- **Schritt 3: Bestimmung des p -Wertes.** In der Fläche unter der Dichte f definiert der Wert $\hat{T}$ selbst einen Schnitt und eine Restfläche. Ist der Ablehnbereich einteilig, benutzt man den einfachen Inhalt der Restfläche als p -Wert. Ist er zweiteilig, so nimmt man den doppelten Inhalt der Restfläche als p -Wert.
- **Schritt 4: Testentscheidung.**
 - Wenn der berechnete Wert $\hat{T}$ in den Ablehnbereich fällt (d.h. $p \leq \alpha$), so entscheidet man sich, die Nullhypothese $\mathcal{H}_0$ abzulehnen und die Alternativhypothese $\mathcal{H}_A$ anzunehmen. Man sagt dann, der Test fällt zum Niveau α signifikant aus.
 - Wenn nicht (d.h. $p > \alpha$), dann entscheidet man sich, die Nullhypothese $\mathcal{H}_0$ beizubehalten.

Hypothesentest mit Stichproben: Ablauf (II)



Testergebnis und Wirklichkeit: Fehler I. und II. Art

		Objektive Realität in $\mathcal{P}'$, $\mathcal{P}''$	
		$\mathcal{H}_0$ richtig	$\mathcal{H}_0$ falsch und $\mathcal{H}_A$ richtig
Test- entscheidung anhand von S' , S''	$\mathcal{H}_0$ wird beibehalten	O.K. $Prob = 1 - \alpha$	Fehler II. Art: Test falsch-negativ $Prob = \beta$
	$\mathcal{H}_0$ wird abgelehnt, $\mathcal{H}_A$ wird angenommen	Fehler I. Art: Test falsch-positiv $Prob = \alpha$	O.K. Power des Tests: $Prob = (1 - \beta)$

- Die "Power" $(1 - \beta)$ ist die Wahrscheinlichkeit, mit dem Hypothesentest einen objektiv vorhandenen Effekt wirklich zu erkennen.

Hypothesentest mit Stichproben: Interpretation

- **Neymar/Pearson (1933):** Nur *“in the long run”*, wenn ein Test immer wieder unter gleichen Bedingungen mit immer neuen Stichproben aus $\mathcal{P}'$, $\mathcal{P}''$ wiederholt wird, kann man garantieren, daß die Wahrscheinlichkeiten für das Auftreten von Fehlern I. und II. Art tatsächlich α und β betragen. Dabei ist α das festgelegte Signifikanzniveau.

- **Was kann man also aus einem einzelnen Testergebnis schließen?**

- Wenn der Test nicht signifikant ausfällt (und die Nullhypothese $\mathcal{H}_0$ beibehalten wird), bedeutet das nicht, daß $\mathcal{H}_0$ tatsächlich zutrifft, denn es könnte sich um einen Fehler II. Art handeln. (Hoffentlich ist die Wahrscheinlichkeit β dafür nicht zu groß!)

- Wenn der Test signifikant ausfällt (und die Alternativhypothese $\mathcal{H}_A$ angenommen wird), bedeutet das nicht, daß $\mathcal{H}_A$ tatsächlich zutrifft, denn es könnte sich um einen Fehler I. Art handeln. (Gleichgültig, wie klein der p -Wert ist!)

- **Konsequenzen der “in-the-long-run”-Logik:** Erst nach mehrfacher Wiederholung dürfen wir schlußfolgern: je öfter der Test signifikant ausfällt, desto unwahrscheinlicher wird die Annahme, daß die Nullhypothese richtig ist (daher “Test gegen $\mathcal{H}_0$ ”).

Hypothesentest mit Stichproben: Stärken und Schwächen

- **Stärken:**

- *weltweit etabliertes Standardwerkzeug*
- *konsistente Fehlerabschätzungen I. und II. Art im mathematischen Modell*

- **Schwächen:**

- *Ein großer p -Wert beweist nicht, daß die Nullhypothese tatsächlich zutrifft!*
- *Ein kleiner p -Wert beweist nicht, daß die Alternativhypothese tatsächlich zutrifft!*
- *Statistische Signifikanz bedeutet nicht automatisch fachliche Relevanz.*
- *Die Voraussetzungen für die "in-the-long-run"-Logik (zufällige Auswahl der Stichproben, beliebige Wiederholbarkeit des Experiments, ...) sind praktisch fast unerfüllbar.*

- **Fazit:**

- *Ergebnisse von Hypothesentests müssen immer sachgerecht interpretiert werden*
- *Testergebnis mit Effektschätzung verbinden (z.B. Konfidenzintervall)*
- *Praktische Konsequenzen: erst aufgrund mehrfach bestätigter Testergebnisse*

Mittelwertvergleich: zwei Populationen, zwei Stichproben

- Gegeben sind zwei Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$. In $\mathcal{P}'$ unterliegt die metrische Variable x der Normalverteilung $\mathcal{N}(\mu', \sigma')$, in $\mathcal{P}''$ der Normalverteilung $\mathcal{N}(\mu'', \sigma'')$ mit $\sigma' = \sigma'' = \sigma$. Die Mittelwerte und Standardabweichungen sind unbekannt.
- Wir untersuchen zwei Stichproben $S' \subset \mathcal{P}'$ und $S'' \subset \mathcal{P}''$ mit evtl. unterschiedlichen Umfängen N' , N'' und beobachten Stichproben-Mittelwerte $\hat{\mu}'$, $\hat{\mu}''$ und empirische Standardabweichungen $\hat{\sigma}'$, $\hat{\sigma}''$.

Vier Werkzeuge zum Mittelwertvergleich:

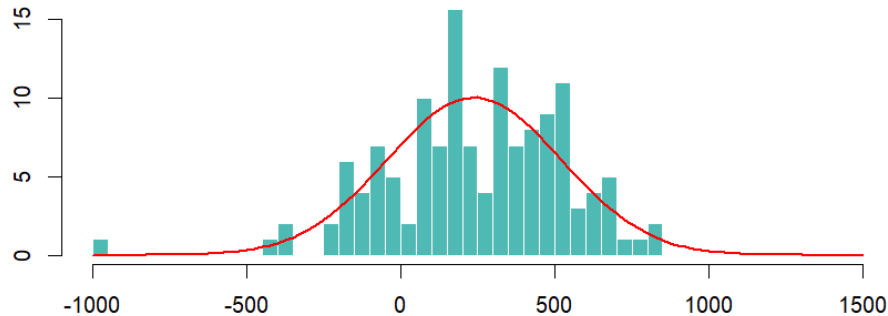
- Effektschätzung mit Cohen's d
- Zweiseitiger t -Test für unabhängige Stichproben
- Einseitiger t -Test für unabhängige Stichproben
- Fehlerbalkendiagramm mit Konfidenzintervallen

Schülersportstudie: sekundärer Endpunkt

- **Beispiel 3.1.:** Schülersportstudie: zufällige Zuordnung von Schülern zur Interventions- (A) oder Kontrollgruppe (B); in Gruppe A zwei Schuljahre lang fünf Sportstunden pro Woche / in Gruppe B zwei Sportstunden pro Woche (siehe Beispiel 2.15.)
- Sekundärer Endpunkt: maximale O_2 -Aufnahme VO_2MAX , diesmal nicht aufs Körpergewicht, sondern auf die Körperoberfläche normiert, Zuwachs Δ nach einem Jahr
- Das Ergebnis soll mit einem Hypothesentest beurteilt werden:
 $\Delta = + 238.2 \text{ ml}/(\text{min} \cdot \text{m}^2)$ (Gruppe A) / $\Delta = + 157.0 \text{ ml}/(\text{min} \cdot \text{m}^2)$ (Gruppe B)
- **Bezeichnungen:**
 - Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$ — alle Schulkinder von etwa 11 Jahren, die ein Jahr lang ausgeweiteten bzw. normalen Sportunterricht erhalten,
 - Stichproben S' , S'' — Gruppen A und B,
 - μ' , μ'' — mittlere tatsächliche Zuwüchse von VO_2MAX ,
 - $\hat{\mu}'$, $\hat{\mu}''$ — Stichproben-Mittelwerte der Zuwüchse von VO_2MAX

Daten im Überblick

- *Ein-Jahres-Zuwüchse von VO_2MAX in S' (Gruppe A — Intervention)*

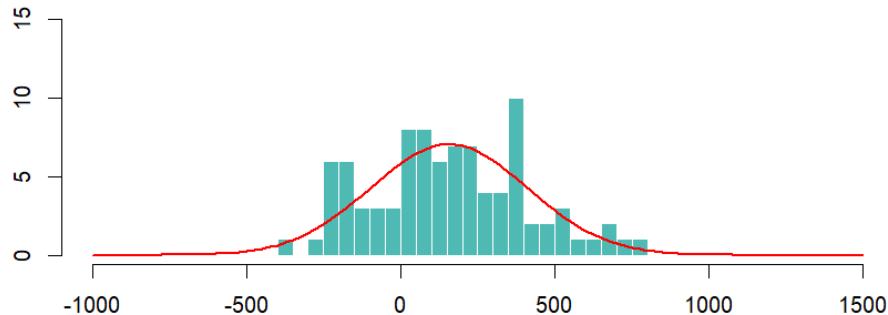


$$N' = 142$$

$$\hat{\mu}' = 238.2 \text{ ml}/(\text{min} \cdot \text{m}^2)$$

$$\hat{\sigma}' = 281.7 \text{ ml}/(\text{min} \cdot \text{m}^2)$$

- *Ein-Jahres-Zuwüchse von VO_2MAX in S'' (Gruppe B — Kontrolle)*



$$N'' = 90$$

$$\hat{\mu}'' = 157.0 \text{ ml}/(\text{min} \cdot \text{m}^2)$$

$$\hat{\sigma}'' = 254.0 \text{ ml}/(\text{min} \cdot \text{m}^2)$$

Cohen's d als Effektschätzer

• **Grundidee:** Eine metrische Variable x sei in den Populationen $\mathcal{P}'$, $\mathcal{P}''$ annähernd normalverteilt mit Mittelwerten μ' , μ'' und gleicher Standardabweichung $\sigma' = \sigma'' = \sigma$. [COHEN (1977)] schlägt vor, die Mittelwertdifferenz auf Vielfache von σ zu normieren: $d = (\mu' - \mu'') / \sigma$. Wir benötigen:

• **Definition 3.2.:** (gepoolte empirische Standardabweichung $\hat{\sigma}_p$) Eine Art gewichtetes Mittel aus $\hat{\sigma}'$, $\hat{\sigma}''$ als Schätzung für die unbekannte gemeinsame Standardabweichung $\sigma' = \sigma'' = \sigma$:

$$\hat{\sigma}_p = \sqrt{\frac{(N' - 1)}{(N' - 1) + (N'' - 1)} \cdot (\hat{\sigma}')^2 + \frac{(N'' - 1)}{(N' - 1) + (N'' - 1)} \cdot (\hat{\sigma}'')^2}.$$

• **Definition 3.3.:** (Cohen's d) In der beschriebenen Situation seien zwei Stichproben S' , S'' aus $\mathcal{P}'$, $\mathcal{P}''$ mit Umfängen N' , N'' , Stichproben-Mittelwerten $\hat{\mu}'$, $\hat{\mu}''$ und empirischen Standardabweichungen $\hat{\sigma}'$, $\hat{\sigma}''$ gegeben. Dann wird Cohen's d geschätzt durch

$$\hat{d} = \frac{\hat{\mu}' - \hat{\mu}''}{\hat{\sigma}_p}.$$

Cohen's d als Effektschätzer (II)

- **Effektstärken** nach Cohen: kein Effekt — $|\hat{d}| < 0.2$ /
klein — $0.2 \leq |\hat{d}| < 0.5$ / moderat — $0.5 \leq |\hat{d}| < 0.8$ / stark — $|\hat{d}| \geq 0.8$
- **Beispiel 3.4.:** Berechnung von $\hat{d}$ für die Ein-Jahres-Zuwächse von VO_2MAX aus
Beispiel 3.1.: $\hat{\mu}' = 238.2 \text{ ml}/(\text{min} \cdot \text{m}^2)$, $\hat{\mu}'' = 157.0 \text{ ml}/(\text{min} \cdot \text{m}^2)$,

$$\hat{\sigma}_p = \sqrt{\frac{141}{230} \cdot (281.7)^2 + \frac{89}{230} \cdot (254.0)^2} = 271.3 \text{ ml}/(\text{min} \cdot \text{m}^2)$$
$$\implies \hat{d} = 0.299 \text{ — kleiner Effekt}$$

- **Beispiel 3.5.:** Berechnung von $\hat{d}$ für die Ruhe-Herzfrequenz nach 1 Jahr in der
Schülersportstudie: $\hat{\mu}' = 98.8 \text{ min}^{-1}$, $\hat{\mu}'' = 101.9 \text{ min}^{-1}$

$$\hat{\sigma}_p = \sqrt{\frac{141}{230} \cdot (11.2)^2 + \frac{89}{230} \cdot (11.9)^2} = 11.5 \text{ min}^{-1},$$
$$\implies \hat{d} = -0.270 \text{ — kleiner Effekt}$$

- **Fazit:** Cohen's d, eine dimensionslose Größe, ermöglicht den Vergleich von Effekten für
verschiedene metrische Variablen unabhängig von deren Dimensionen und Maßeinheiten.
Allerdings erlaubt $\hat{d}$ noch keinen Rückschluß auf die wahre Mittelwertdifferenz.

Verteilung der Mittelwert-Differenzen

• **Frage:** *Wie sind die Differenzen der Stichproben-Mittelwerte für Paare von unabhängigen Stichproben verteilt? Die Antwort gibt*

• **Lehrsatz 3.6.:** *(Verteilung der Differenzen der Stichproben-Mittelwerte) Eine metrische Variable x sei in beiden Populationen $\mathcal{P}'$, $\mathcal{P}''$ normalverteilt mit Mittelwerten μ' , μ'' und gleicher Standardabweichung $\sigma' = \sigma'' = \sigma$. Aus $\mathcal{P}'$, $\mathcal{P}''$ ziehen wir immer wieder unabhängige Stichproben-Paare S' , S'' mit stets denselben Umfängen N' , N'' .*

1) Dann ist die Größe $v = \frac{(\hat{\mu}' - \hat{\mu}'') - (\mu' - \mu'')}{\sigma \cdot \sqrt{\frac{1}{N'} + \frac{1}{N''}}}$ Standard-normalverteilt.

2) Ersetzt man σ durch die gepoolte empirische Standardabweichung $\hat{\sigma}_p$, so unterliegt

die Größe $v = \frac{(\hat{\mu}' - \hat{\mu}'') - (\mu' - \mu'')}{\hat{\sigma}_p \cdot \sqrt{\frac{1}{N'} + \frac{1}{N''}}}$ einer t -Verteilung mit $df = (N' + N'' - 2)$

Freiheitsgraden.

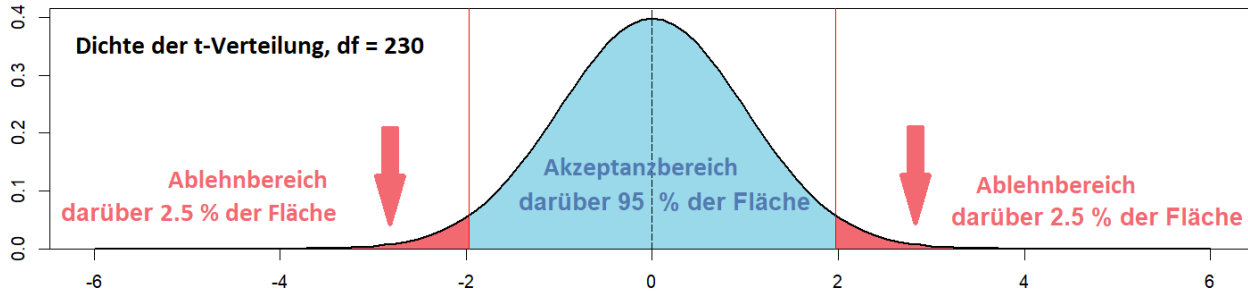
Zweiseitiger t-Test am Beispiel

- Auf dem mathematischen Hintergrund von Lehrsatz 3.6., Teil 2), wird der bekannteste statistische Hypothesentest aufgebaut — der zweiseitige t-Test. Erklärung am Beispiel:
- Zweiseitige Nullhypothese $\mathcal{H}_0$: “Die mittleren Zuwüchse von VO₂MAX in $\mathcal{P}'$ und $\mathcal{P}''$ stimmen überein, also $\mu' = \mu''$.”
- Zweiseitige Alternativhypothese $\mathcal{H}_A$: “Die mittleren Zuwüchse von VO₂MAX in $\mathcal{P}'$ und $\mathcal{P}''$ sind verschieden, also entweder $\mu' < \mu''$ oder $\mu' > \mu''$.”
- Stichproben: S' , S'' mit Umfängen $N' = 142$, $N'' = 90$, Stichproben-Mittelwerten $\hat{\mu}' = 238.2$, $\hat{\mu}'' = 157.0$ und empir. Standardabweichungen $\hat{\sigma}' = 281.7$, $\hat{\sigma}'' = 254.0$
- Signifikanzniveau: $\alpha = 0.05$ (wird nahezu standardmäßig verwendet).
- Die Prüfgröße $\hat{T}$ aus Lehrsatz 3.6., Teil 2), unterliegt einer t-Verteilung mit $df = (N' + N'' - 2) = 230$ Freiheitsgraden. Berechnung unter der Annahme, daß $\mathcal{H}_0$ gilt:

$$\hat{T} = \frac{(\hat{\mu}' - \hat{\mu}'') - \overbrace{(\mu' - \mu'')}^{0 \text{ wegen } \mathcal{H}_0}}{\hat{\sigma}_p \cdot \sqrt{\frac{1}{N'} + \frac{1}{N''}}} = \frac{238.2 - 157.0}{271.3 \cdot \sqrt{\frac{1}{142} + \frac{1}{90}}} = 2.2213.$$

Zweiseitiger t-Test am Beispiel (II)

- Zur zweiseitigen Nullhypothese gehört ein symmetrischer, zweiteiliger Ablehnbereich. Die zugehörigen Quantile der t -Verteilung mit $df = 230$ Freiheitsgraden sind $Q_{2.5}(t, 230) = -1.97$ und $Q_{97.5}(t, 230) = 1.97$.



Die Fläche unter der Dichte wird im Verhältnis $\alpha : (1 - \alpha) = 0.05 : 0.95$ aufgeteilt.

- Technische Voraussetzungen für den t -Test: Für VO_2MAX werden ungeprüft Normalverteilungen in $\mathcal{P}'$, $\mathcal{P}''$ angenommen. Man darf von annähernd gleichen Standardabweichungen ausgehen, falls $0.5 \leq \hat{\sigma}' / \hat{\sigma}'' \leq 2$ (Faustregel). Hier: $\hat{\sigma}' / \hat{\sigma}'' = 1.11$.

Zweiseitiger t-Test am Beispiel (III)

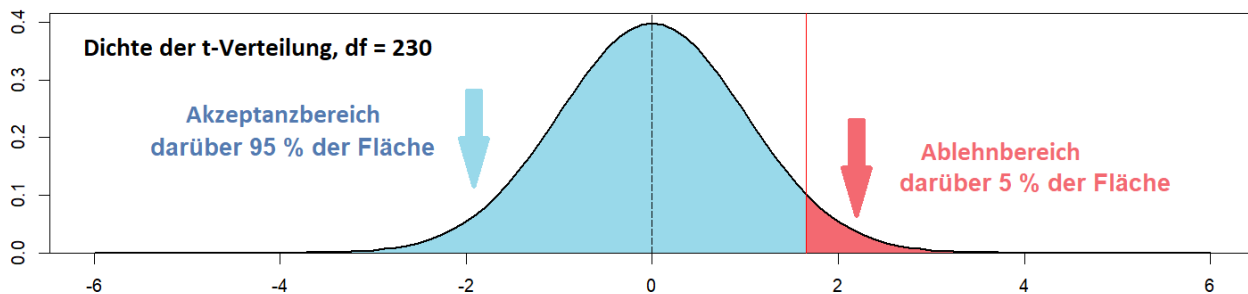
- **Berechnung der Prüfgröße unter $\mathcal{H}_0$:** $\hat{T} = 2.2213 = Q_{98.63}(t, 230)$ (Tabelle!).
- **Bestimmung des p -Werts:** $p = 2 \cdot 0.0137 = 0.0274$ (Verdopplung der Restfläche).
- **Testentscheidung:** Wegen $p = 0.0274 < \alpha$ wird die zweiseitige Nullhypothese $\mathcal{H}_0$ verworfen und die zweiseitige Alternativhypothese $\mathcal{H}_A$ angenommen.
- **95%-Konfidenzintervall** zur Differenz $\hat{\mu}' - \hat{\mu}'' = \hat{\Delta} = 81.2$ der Stichproben-Mittelwerte; wird mit Hilfe der t -Verteilung gebildet (Achtung, neue Definition!):

$$\left[\hat{\Delta} + Q_{2.5}(t, 230) \cdot \hat{\sigma}_p \cdot \sqrt{\frac{1}{N'} + \frac{1}{N''}}, \hat{\Delta} + Q_{97.5}(t, 230) \cdot \hat{\sigma}_p \cdot \sqrt{\frac{1}{N'} + \frac{1}{N''}} \right] \\ = [9.2, 153.2].$$

- **Interpretation:** "Für etwa elfjährige Schulkinder, die ausgeweiteten Sportunterricht erhalten, und solche, die normalen Sportunterricht erhalten, unterscheiden sich die mittleren Zuwüchse von $VO_2\text{MAX}$ nach einem Jahr um $\Delta = 81.2 \text{ ml}/(\text{min} \cdot \text{m}^2)$ (95%-K.I. $[9.2, 153.2]$). Der Unterschied ist signifikant zum 5%-Niveau ($p = 0.027$)."

Einseitiger t-Test

- **Beispiel 3.7.:** Wenn wir uns nur dafür interessieren, ob in Beispiel 3.1. die Interventions- der Kontrollgruppe beim Zuwachs von VO_2MAX überlegen ist, so kann ein einseitiger t -Test durchgeführt werden. Jetzt kommt es aber auf die Reihenfolge der Gruppen an!
- Einseitige Nullhypothese $\mathcal{H}_0$: "Der mittlere Zuwachs von VO_2MAX in $\mathcal{P}'$ übersteigt nicht den in $\mathcal{P}''$, also $\mu' \leq \mu''$."
- Einseitige Alternativhypothese $\mathcal{H}_A$: "Der mittlere Zuwachs von VO_2MAX ist in $\mathcal{P}'$ größer als in $\mathcal{P}''$, also $\mu' > \mu''$."
- Zur einseitigen Nullhypothese gehört ein einteiliger Ablehnbereich, dessen Lage durch die Relation in $\mathcal{H}_0$ bestimmt wird. Das Quantil der t -Verteilung mit $df = 230$ Freiheitsgraden zu $\alpha = 0.05$ lautet $Q_{95}(t, 230) = 1.65$.



Einseitiger t-Test (II)

- *Voraussetzungen und Rechnung unverändert: $\hat{T} = 2.2213 = Q_{98.63}(t, 230)$. Dazu gehört $p = 0.0137$ (einfache Restfläche). Auch das 95%-K.I. für $\hat{\Delta}$ bleibt dasselbe.*
- **Testentscheidung:** *Wegen $p = 0.0137 < \alpha$ wird die einseitige Nullhypothese $\mathcal{H}_0$ verworfen und die einseitige Alternativhypothese $\mathcal{H}_A$ angenommen.*
- **Interpretation:** *“Für etwa elfjährige Schulkinder, die ausgeweiteten Sportunterricht erhalten, ist der mittlere Zuwachs von $VO_2\text{MAX}$ nach einem Jahr größer als bei solchen, die normalen Sportunterricht erhalten. Das Ergebnis mit einer Differenz von $\Delta = 81.2 \text{ ml}/(\text{min} \cdot \text{m}^2)$ (95%-K.I. $[9.2, 153.2]$) ist signifikant zum 5%-Niveau ($p = 0.0137$).”*
- **Fazit:** *Einseitiges Testen gegen $\mu' \leq \mu''$ anstatt $\mu' = \mu''$ mit den gleichen Daten halbiert den p -Wert — aber um den Preis, daß ein möglicher Effekt $\mu' < \mu''$ in umgekehrter Richtung nicht mehr erkannt werden kann.*

Das Fehlerbalken-Diagramm

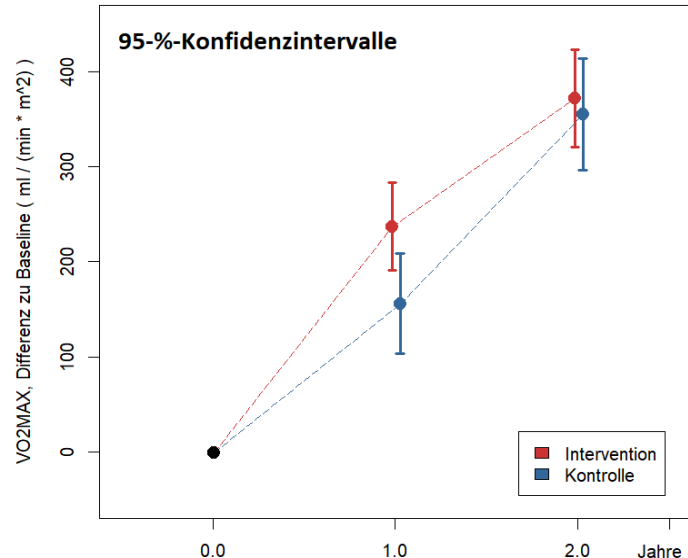
- **Elemente im Fehlerbalken-Diagramm:**

- *Skala: Wertebereich der metrischen Variable (Einheit angeben!)*

- *Punkte: Stichproben-Mittelwerte*

- *Balken: Konfidenzintervalle (alle zum gleichen Level, hier 95%)*

- *Verbindungslinien: veranschaulichen den Zuwachs der Mittelwerte im zeitlichen Verlauf*



- *Man sollte diejenigen Merkmale darstellen, deren Mittelwerte im Hypothesentest verglichen wurden!*

Wie groß ist der Fehler II. Art?

• **Beispiel 3.8.:** *Nochmals die Schülersportstudie: Wir testen die Mittelwertdifferenz bei den Zuwüchsen von VO_2MAX , auf die Körperoberfläche normiert, diesmal nach zwei Jahren.*

• **Daten:** $N' = 142$, $N'' = 90$, $\hat{\mu}' = 373.2$, $\hat{\mu}'' = 356.2$, $\hat{\sigma}' = 311.6$, $\hat{\sigma}'' = 283.2$, $\hat{\sigma}_p = 300.9$ (mehr Informationen braucht man nicht!)

• **Zweiseitiger t -Test:** Hypothesen $\mathcal{H}_0$ und $\mathcal{H}_A$ wie in Beispiel 3.1. (zweiseitig), Signifikanzniveau $\alpha = 0.05$, $df = 230$. Folgt: $\hat{T} = 0.4192$, dazu gehören $Q_{33.8}(t, 230)$, $Q_{66.2}(t, 230)$ und $p = 2 \cdot 0.338 = 0.676 > \alpha$.

• **Testentscheidung:** Test fällt nicht signifikant aus, also muß $\mathcal{H}_0$ beibehalten werden. Wie groß ist die Wahrscheinlichkeit β , daß wir uns damit irren ("Fehler II. Art")?

• **Problem:** Wir kennen weder die wahren Mittelwerte μ' , μ'' noch die wahren Standardabweichungen σ' , σ'' in den Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$!

Wovon hängt β außerdem ab?

Schätzung des Fehlers II. Art durch Simulation

- Der Fehler II. Art hängt ab von: α , N' , N'' , μ' , σ' , μ'' , σ'' .
- Wir fixieren $\alpha = 0.05$, $N' = 142$, $N'' = 90$ und betrachten die unbekannte Mittelwert-Differenz $\delta = (\mu' - \mu'')$ und die unbekannte Standardabweichung $\sigma = \sigma' = \sigma''$ als Parameter.
- **Simulation:** zu festem δ und σ jeweils $K = 1000$ Stichproben S' , S'' mit $N' = 142$ und $N'' = 90$ aus normalverteilten Populationen mit Parametern $\mu' = \delta$, $\sigma' = \sigma$ bzw. $\mu'' = 0$, $\sigma'' = \sigma$. Dann führen wir jeweils einen zweiseitigen t -Test zum Signifikanzniveau $\alpha = 0.05$ durch. Wieviel Prozent der Tests fallen falsch-negativ aus?

	$\delta = 10$	$\delta = 30$	$\delta = 50$	$\delta = 70$	$\delta = 90$	$\delta = 110$	$\delta = 130$	$\delta = 150$
$\sigma = 250$	94.6	86.3	71.3	47.1	24.8	10.3	2.2	0.6
$\sigma = 270$	95.3	87.3	74.1	51.6	30.7	15.1	4.3	1.1
$\sigma = 290$	95.6	89.1	75.4	58.8	35.5	19.2	8.1	3.0
$\sigma = 310$	97.2	87.6	77.3	59.3	42.5	24.2	10.9	6.1

Fallzahlplanung durch Simulation

- **Beispiel 3.9. (Fallzahlplanung):** Wir wollen anhand von zwei gleichgroßen Stichproben einen Effekt von $\delta = (\mu' - \mu'') \geq 60$ durch einen zweiseitigen t -Test zum Signifikanzniveau $\alpha = 0.05$ mit mindestens 80% Wahrscheinlichkeit aufdecken. Welcher Stichproben-Umfang N wird mindestens benötigt?
- Wir fixieren $\alpha = 0.05$ und $\delta = 60$ und betrachten den unbekannten Stichproben-Umfang N und die unbekannte Standardabweichung σ als Parameter.
- **Simulation:** zu festem N und σ jeweils $K = 1000$ Stichproben S', S'' mit $N' = N'' = N$ aus normalverteilten Populationen mit Parametern $\mu' = \delta = 60$, $\sigma' = \sigma$ bzw. $\mu'' = 0$, $\sigma'' = \sigma$. Dann führen wir jeweils einen zweiseitigen t -Test zum Signifikanzniveau $\alpha = 0.05$ durch. Wieviel Prozent der Tests fallen falsch-negativ aus?

	$N = 50$	$N = 100$	$N = 150$	$N = 200$	$N = 250$	$N = 300$	$N = 350$
$\sigma = 250$	78.7	60.7	45.8	36.0	22.6	16.7	11.6
$\sigma = 280$	80.7	68.2	52.6	45.7	33.0	24.4	18.1
$\sigma = 310$	85.4	71.7	60.1	52.4	39.1	33.4	25.9

Der Welch-Test

- **Variante des t -Tests:** falls deutliche Abweichung von Normalverteilung in $\mathcal{P}'$, $\mathcal{P}''$ zu erwarten ist oder $\hat{\sigma}'/\hat{\sigma}'' < 0.5$ bzw. $\hat{\sigma}'/\hat{\sigma}'' > 2$ ausfällt. Durchführung wie klassischer t -Test, aber:

- Neue Prüfgröße (ohne $\hat{\sigma}_p$):
$$\hat{T} = \frac{\hat{\mu}' - \hat{\mu}''}{\sqrt{\frac{(\hat{\sigma}')^2}{N'} + \frac{(\hat{\sigma}'')^2}{N''}}}$$

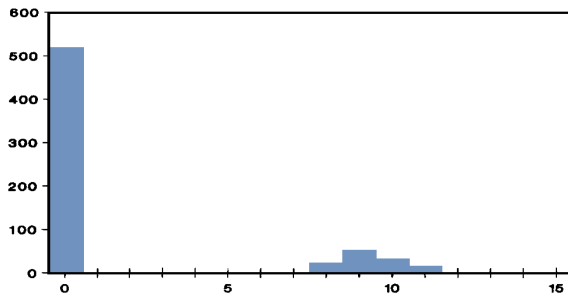
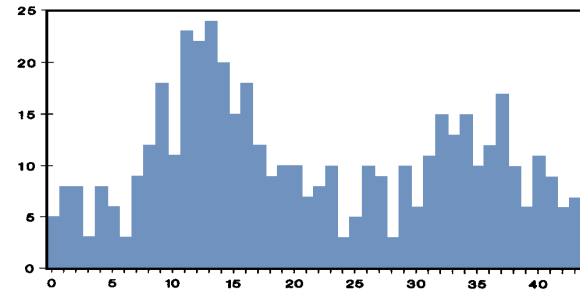
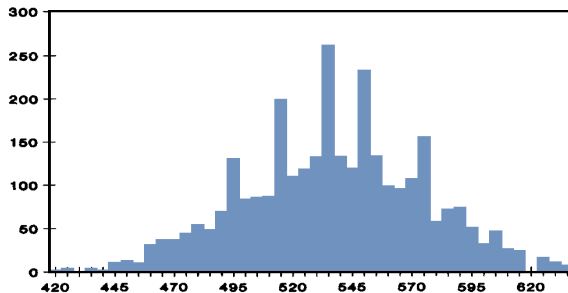
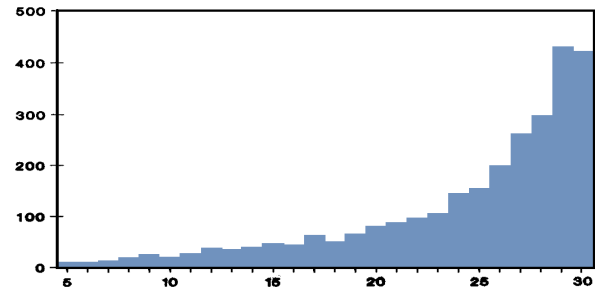
- Neue Formel für die Freiheitsgrade df der t -Verteilung:

$$df = \text{ganzer Anteil von } \frac{\left[\left(\frac{(\hat{\sigma}')^2}{N'} \right) + \left(\frac{(\hat{\sigma}'')^2}{N''} \right) \right]^2}{\frac{1}{(N' - 1)} \cdot \left(\frac{(\hat{\sigma}')^2}{N'} \right)^2 + \frac{1}{(N'' - 1)} \cdot \left(\frac{(\hat{\sigma}'')^2}{N''} \right)^2}$$

- In Beispiel 3.1.: $\alpha = 0.05$, $df = \text{ganzer Anteil von } 206.11 = 206$, $\hat{T} = 2.2769$, dazu gehören $Q_{1.19}(t, 206)$ und $Q_{98.81}(t, 206)$ sowie $p = 2 \cdot 0.0119 = 0.0238 < \alpha$.

Robustheit des t-Tests

- [SAWILOWSKY/BLAIR (1992)] : Der *t*-Test hält die Fehlerwahrscheinlichkeiten I. und II. Art auch dann noch ein, wenn x in $\mathcal{P}'$, $\mathcal{P}''$ ganz und gar nicht normalverteilt ist!



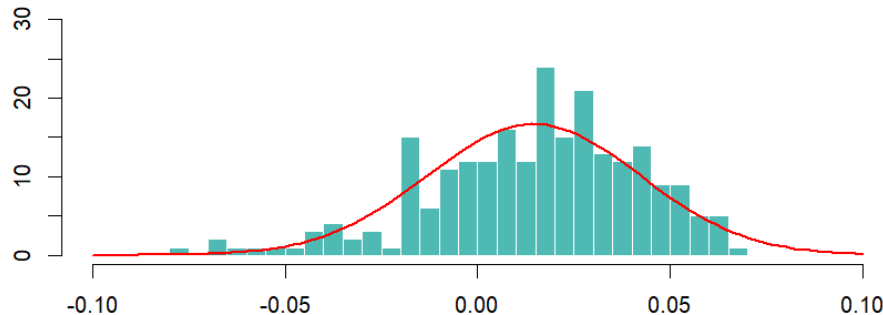
- Bedingungen für Robustheit: $N' \approx N''$; $N', N'' \geq 30$; Welch-Test zweiseitig

Mittelwertvergleich bei verbundenen Stichproben

- **Beispiel 3.10.:** (*“Vorher-Nachher”*) 47 Personen unterziehen sich 3 Monate lang der sog. *Drei-Äpfel-Diät*. Zu Beginn und am Ende der Diät wird der HDL-Cholesterolverwert gemessen. Gibt es einen signifikanten Unterschied zwischen den mittleren Werten am Anfang und nach 3 Monaten?
- **Beispiel 3.11.:** (*Fallkontrollstudie*) Ein Sportmediziner möchte zur Rehabilitation nach Knieverletzungen ein neues Trainingsprogramm erproben und wendet es bei den nächsten 30 Patienten an. Dann wird jedem einzelnen Patienten eine nach Geschlecht, Alter, BMI, ... übereinstimmende *“Kontrollperson”* fest zugeordnet, die mit dem bisherigen Programm behandelt worden ist. Gibt es bei der Wiederherstellung der Beweglichkeit des Kniegelenks in beiden Gruppen einen signifikanten Unterschied?
- **Beispiel 3.12.:** (*Methodenvergleich*) Für 232 Schulkinder wird die Körperoberfläche KOF_{Mo} nach der Formel von Mosteller (die von Größe und Gewicht abhängt) und nochmals als KOF_{Li} nach der Formel von Livingston-Lee berechnet (die nur vom Gewicht abhängt). Gibt es einen signifikanten Unterschied zwischen den Ergebnissen, oder sind die beiden Berechnungsmethoden äquivalent?

Zweiseitiger t-Test für Paardifferenzen

- **Beispiel 3.12. (Fortsetzung):** Daten aus der Schülersportstudie (Baseline); aus den verbundenen Stichproben S' ($N = 232$, KOF_{Mo}) und S'' ($N = 232$, KOF_{Li}) bilden wir die Stichprobe S^Δ der Paardifferenzen ($N = 232$, $KOF_\Delta = KOF_{Mo} - KOF_{Li}$).



$$N^\Delta = 232$$

$$\hat{\mu}^\Delta = 0.0145 \text{ m}^2$$

$$\hat{\sigma}^\Delta = 0.0277 \text{ m}^2$$

- Zweiseitige Nullhypothese $\mathcal{H}_0$: “Der Mittelwert der Paardifferenzen KOF_Δ beträgt Null, also $\mu^\Delta = 0$.”
- Zweiseitige Alternativhypothese $\mathcal{H}_A$: “Der Mittelwert der Paardifferenzen KOF_Δ ist von Null verschieden, also entweder $\mu^\Delta < 0$ oder $\mu^\Delta > 0$.”
- **Frage:** Wie sind die Stichproben-Mittelwerte der Paardifferenzen verteilt?

Zweiseitiger t-Test für Paardifferenzen (II)

- Prüfgröße für Paardifferenzen:
$$\hat{T} = \frac{\hat{\mu}^{\Delta} - 0}{\hat{\sigma}^{\Delta} \cdot \sqrt{\frac{1}{N^{\Delta}}}} = \hat{\mu}^{\Delta} \cdot \sqrt{N^{\Delta}} / \hat{\sigma}^{\Delta}.$$

Diese Prüfgröße ergibt sich, wenn man in die bisherigen Formeln für $\hat{\sigma}_p$ und $\hat{T}$ formal zwei Stichproben mit den Kenngrößen $(N^{\Delta}, \hat{\mu}^{\Delta}, \hat{\sigma}^{\Delta})$ und $(N^{\Delta}, 0, 0)$ einsetzt.

• **Lehrsatz 3.13.:** (Verteilung der Stichproben-Mittelwerte der Paardifferenzen) Eine metrische Variable x in der Gesamtpopulation $\mathcal{P}$ sei normalverteilt mit Mittelwert μ und Standardabweichung σ . Ziehen wir immer wieder verbundene Stichproben-Paare S', S'' mit stets denselben Umfängen $N' = N'' = N^{\Delta}$, so unterliegt die Prüfgröße $\hat{T}$ für die Paardifferenzen einer t -Verteilung mit $df = (N^{\Delta} - 1)$ Freiheitsgraden.

• **Zurück zu Beispiel 3.12.:** Zweiseitiger Test zum Signifikanzniveau $\alpha = 0.01$, dazu gehören die Quantile $Q_{99.5}(t, 231) = -Q_{0.5}(t, 231) = 2.5973$. Wir erhalten $\hat{T} = 7.9732$ und dazu $Q_{4 \cdot 10^{-14}}(t, 231)$ und $Q_{100 - 4 \cdot 10^{-14}}(t, 231)$ sowie $p = 2 \cdot 0.000 \dots$, kurz " $p \leq 0.0005$ " $< \alpha$. Also wird $\mathcal{H}_0$ abgelehnt und $\mathcal{H}_A$ angenommen.

Signifikant, aber nicht relevant!

- **Frage:** Der mittlere Unterschied von 0.0145 m^2 zwischen den Methoden ist zwar (hoch-)signifikant zum 1-%-Niveau ($p < 0.0005$), aber ist er auch praktisch relevant?
- **Mosteller-Formel:** $KOF_{Mo}(G, H) = \frac{1}{60} \cdot \sqrt{G \cdot H} \text{ m}^2$ (G in kg, H in cm einsetzen)

$$\begin{aligned} Err_{Mo} &= KOF_{Mo}(G + \Delta G, H + \Delta H) - KOF_{Mo}(G, H) \\ &= \frac{1}{120} \cdot \sqrt{H/G} \cdot \Delta G + \frac{1}{120} \cdot \sqrt{G/H} \cdot \Delta H + \dots \end{aligned}$$

Mittelwert **Fehlerterm rot**: 0.0161, Mittelwert **Fehlerterm blau**: 0.0044, Gewicht auf $\Delta G = 0.5 \text{ kg genau}$, Größe auf $\Delta H = 1 \text{ cm genau} \implies Err_{Mo} \sim 0.0124$

- **Livingston-Lee-Formel:** $KOF_{Li}(G) = 0.1173 \cdot G^{0.6466} \text{ m}^2$ (G in kg einsetzen)

$$Err_{Li} = KOF_{Li}(G + \Delta G) - KOF_{Li}(G) = 0.1173 \cdot 0.06466 \cdot G^{-0.3534} \cdot \Delta G + \dots$$

Mittelwert **Fehlerterm**: 0.0205, Gewicht auf $\Delta G = 0.5 \text{ kg genau} \implies Err_{Li} \sim 0.0102$

- **Antwort:** Offenbar nicht praxisrelevant wegen $Err_{Mo} \pm Err_{Li} \sim \pm 0.02$!

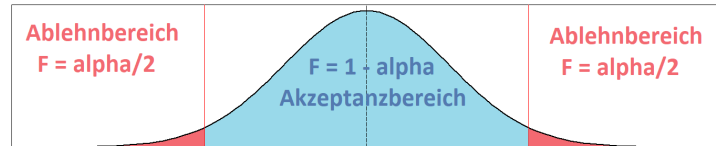
Zusammenfassung: Alle Fragestellungen

(1) zweiseitig — Test gegen $\mathcal{H}_0: \mu' = \mu''$

$$\mathcal{H}_A: \mu' \neq \mu''$$

$$\hat{T} = (\hat{\mu}' - \hat{\mu}'') / \dots$$

AB zweiseitig, symmetrisch

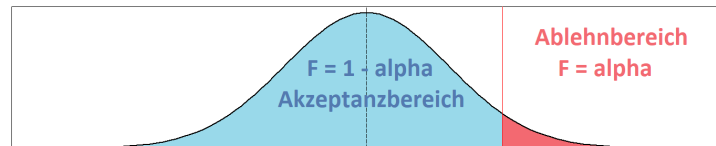


(2) einseitig rechts — Test gegen $\mathcal{H}_0: \mu' \leq \mu''$

$$\mathcal{H}_A: \mu' > \mu''$$

$$\hat{T} = (\hat{\mu}' - \hat{\mu}'') / \dots$$

AB einteilig, rechts

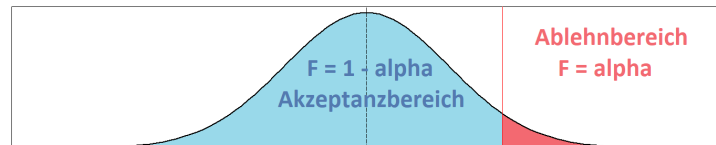


(3) einseitig rechts mit Verschiebung — Test gegen $\mathcal{H}_0: \mu' - \mu'' \leq v$

$$\mathcal{H}_A: \mu' - \mu'' > v$$

$$\hat{T} = (\hat{\mu}' - \hat{\mu}'' - v) / \dots$$

AB einteilig, rechts



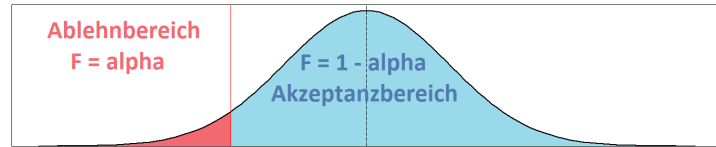
Zusammenfassung: Alle Fragestellungen (II)

(4) einseitig links — *Test gegen $\mathcal{H}_0: \mu' \geq \mu''$*

$$\mathcal{H}_A: \mu' < \mu''$$

$$\hat{T} = (\hat{\mu}' - \hat{\mu}'') / \dots$$

AB einteilig, links

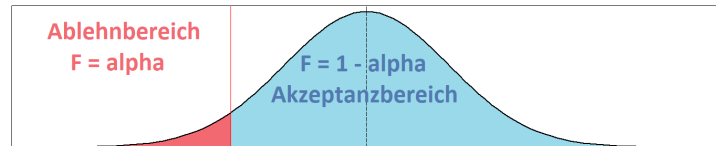


(5) einseitig links mit Verschiebung — *Test gegen $\mathcal{H}_0: \mu' - \mu'' \geq v$*

$$\mathcal{H}_A: \mu' - \mu'' < v$$

$$\hat{T} = (\hat{\mu}' - \hat{\mu}'' - v) / \dots$$

AB einteilig, links



(6) Äquivalenz mit Toleranzbereich (“TOST”) — *Test gegen $\mathcal{H}_0: |\mu' - \mu''| \geq v$*

$$\mathcal{H}_A: |\mu' - \mu''| < v$$

Zwei einseitige Tests zum gleichen Signifikanzniveau α nötig:

(6a) — *Test gegen $\mu' - \mu'' \leq -v \implies$ weiter mit (3), darin $\hat{T} = (\hat{\mu}' - \hat{\mu}'' - (-v)) / \dots$*

(6b) — *Test gegen $\mu' - \mu'' \geq v \implies$ weiter mit (5)*