Analyse experimenteller Daten

-

Testverfahren

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany

Vergleich von zwei Systemen?

- Ist Interface A bedienfreundlicher als Interface B?

Time Stack

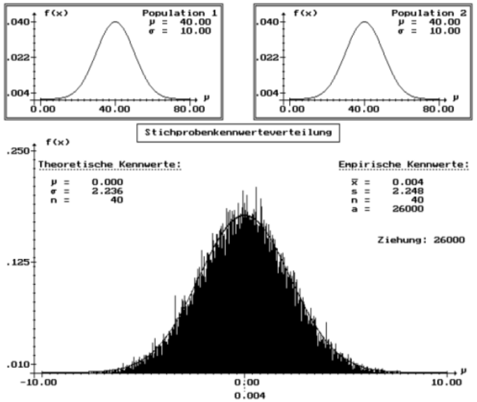
Time Pie

t-Test:

- Entscheidungsregel auf mathematischer Grundlage, mit deren Hilfe ein Unterschied zwischen den empirisch gefundenen Mittelwerten zweier Gruppen näher analysiert werden kann.
- Schätzt die Populationsparametern basierend auf der Streuung und des arithmetischen Mittels, die mit Hilfe der Stichprobe geschätzt werden.
- Liefert eine Entscheidungshilfe dafür, ob ein gefundener Unterschied zwischen den Mittelwerten rein zufällig entstanden ist, oder ob es wirklich bedeutsame Unterschiede zwischen den zwei untersuchten Gruppen gibt.

t-Test

- Nullhypothese: Die Populationsmittelwerte von zwei Gruppen sind identisch und deshalb ist eine Mittelwertdifferenz von Null zu erwarten.
- Erklärung:
  - Wird aus zwei Populationen mit identischen Mittelwerten je eine Stichprobe gezogen, so kann die Differenz der Mittelwerte der Stichproben theoretisch jeden beliebigen Wert annehmen.
  - Die Stichprobenmittelwerte sind aber normalverteilt um den jeweiligen Erwartungswert, dem Populationsmittelwert.
  - Da die Populationsmittelwerte identisch sind, wird sich die Mehrzahl der gefundenen Differenzen folglich in der Nähe von Null befinden.



t-Test

- Aus diesen Überlegungen resultiert nach unendlich vielen Ziehungen von Stichproben eine **Normalverteilung** der **Mittelwertdifferenzen** mit dem arithmetischen Mittel Null und einer von der Populationsstreuung und den Stichprobenumfängen abhängigen Streuung.
- Diese Verteilung heißt Stichprobenkennwerteverteilung von Mittelwertdifferenzen unter der Nullhypothese.

  Schätzung der Stichprobenkennwerteverteilung mit Hilfe der Stichprobe

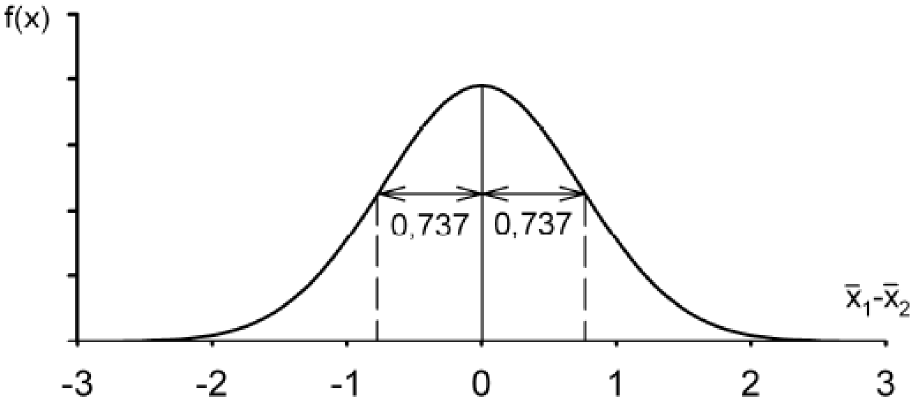
$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$\hat{\sigma}_{\bar{x}_1-\bar{x}_2}$	Standardfehler der Mittelwertdifferenz
n_1	Anzahl der Vpn bzw. Beobachtungen in Sp 1
$\hat{\sigma}_1^2$	Geschätzte Varianz der Population 1
n_2	Anzahl der Vpn bzw. Beobachtungen in Sp 2
$\hat{\sigma}_2^2$	Geschätzte Varianz der Population 2

Beispiel: Vergleich von Interfaces

- Verarbeitungsbedingung — Anzahl von Fehlern des Nutzers
  - Time Stack — $\bar{x}_{Stack} = 7{,}2$ — $\hat{\sigma}_{Stack} = 3{,}162$
  - Time Pie — $\bar{x}_{Pie} = 11$ — $\hat{\sigma}_{Pie} = 4{,}14$
- 50 Versuchspersonen pro Verarbeitungsbedingung
- Standardfehler der Mittelwertdifferenz

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{4{,}14^2}{50} + \frac{3{,}162^2}{50}} = 0{,}737$$

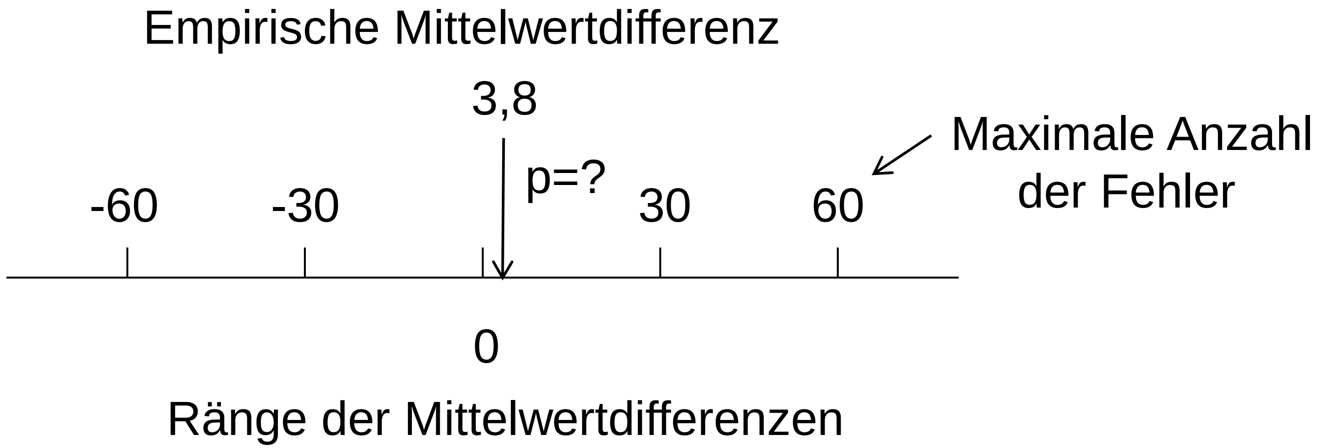


Fragestellung des t-Tests

- Wie wahrscheinlich ist die gefundene Mittelwertdifferenz unter allen möglichen Differenzen?

- Beispiel:

  Versuchspersonen müssen sich 60 Wörter merken



Die t-Verteilung

- Für die Bewertung der **Auftrittswahrscheinlichkeit** einer **empirisch gefundenen Differenz** ist ein **standardisiertes Maß** für eine Mittelwertdifferenz sehr hilfreich.
- Die Standardisierung der Stichprobenkennwerteverteilung erfolgt ähnlich wie bei den z-Werten an der geschätzten Streuung.
- Die empirische **Mittelwertdifferenz** wird unter Kenntnis der entsprechenden **Streuung** in einen **t-Wer**t umgerechnet.
- Die standardisierten Stichprobenkennwerte heißen t-Werte, die standardisierten Verteilungen t-Verteilungen.

Die t-Verteilung

- Allgemeine Definition des t-Wertes:

$$\mathrm{t}_{\mathrm{df}} = \frac{empirische \quad Mittelwertsdifferenz \quad - \quad theoretische \quad Mittelwertsdifferenz}{geschätzter \quad Standardfehler \quad der \quad Mittelwertsdifferenz}$$

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

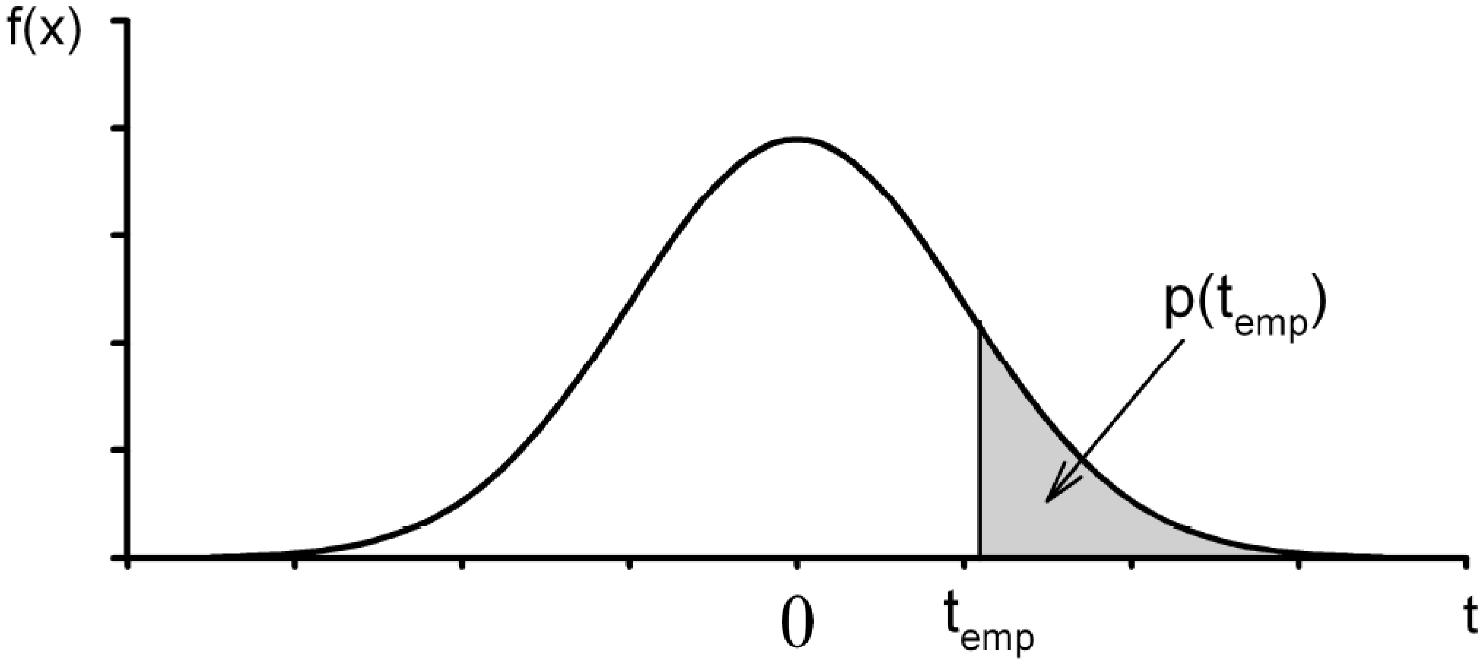
- Theoretische Mittelwertdifferenz unter der Nullhypothese: $\mu_1 - \mu_2 = 0$
- Vereinfachte Formel:

$$t_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

- Anhand der t-Verteilung kann einem empirischen t-Wert eine Wahrscheinlichkeit zugeordnet werden, mit der exakt dieser oder ein größerer t-Wert unter der Annahme der Nullhypothese auftritt.

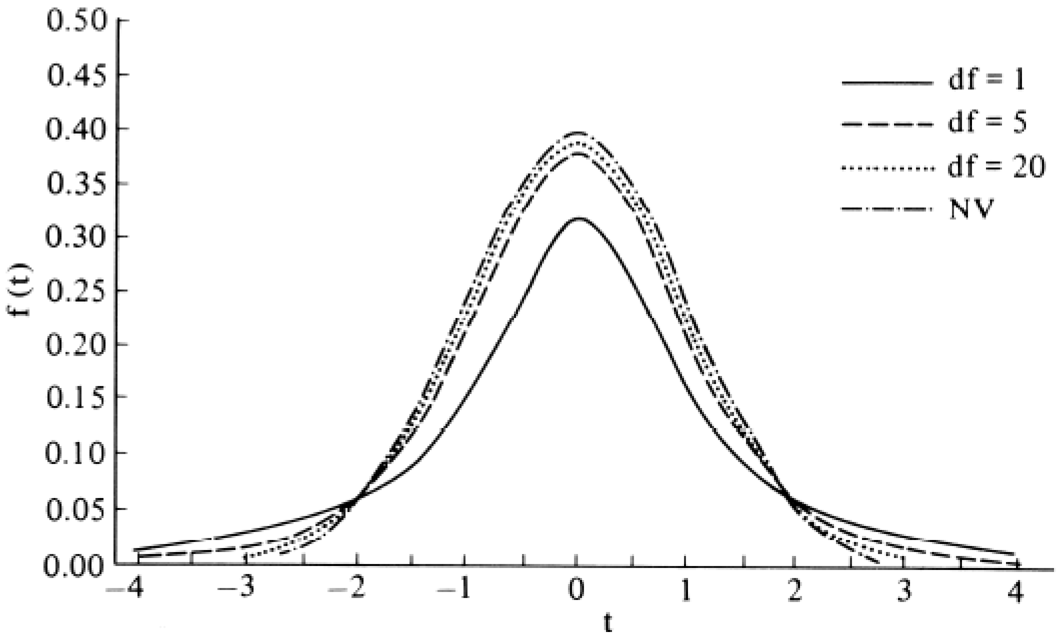
Die t-Verteilung

- Die Auftrittswahrscheinlichkeit eines positiven t-Werts entspricht dem Anteil der Fläche unter der Kurve, den der t-Wert rechts abschneidet.



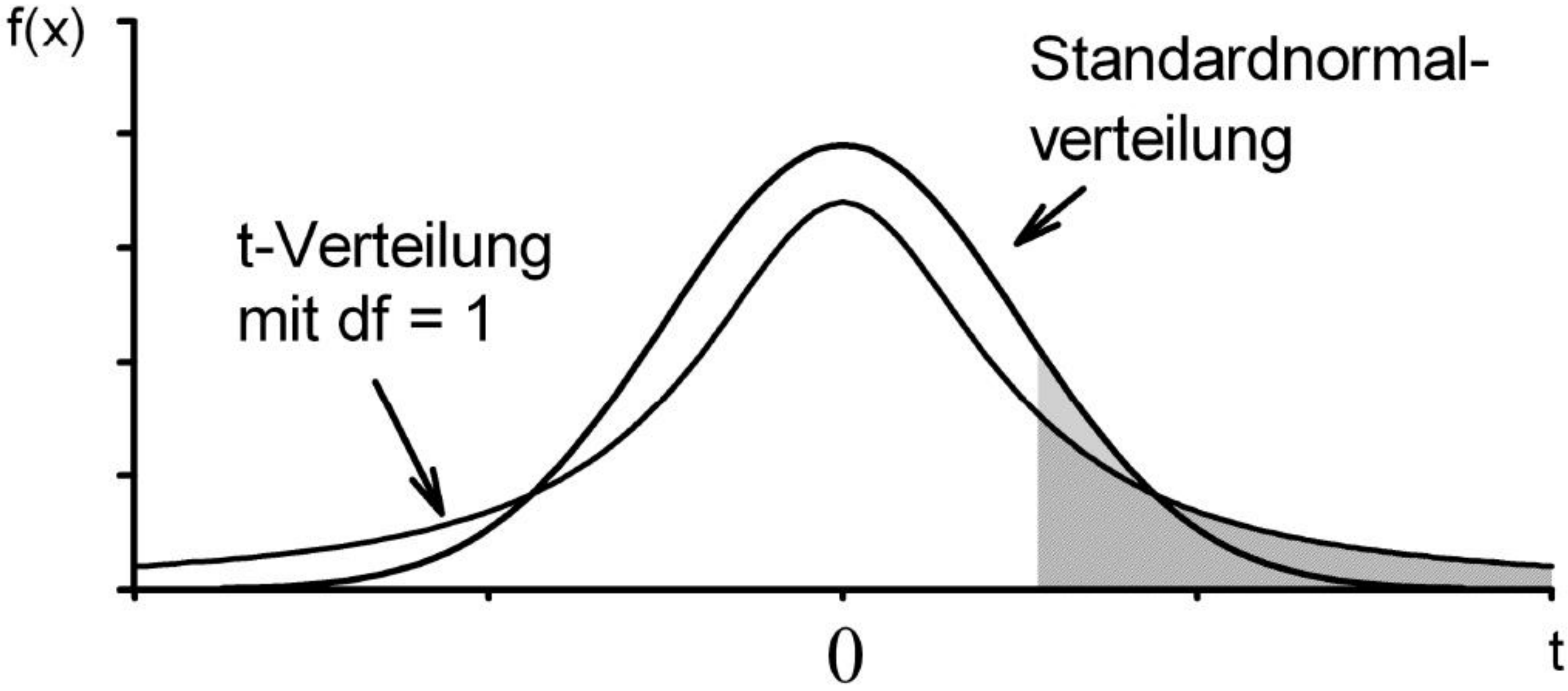
Die t-Verteilung

- In Abhängigkeit von der Anzahl der Freiheitsgrade (hängen von der Anzahl der Versuchspersonen bzw. Beobachtungen ab) haben die t-Verteilungen unterschiedliche Formen.
- Bei einer hohen Anzahl der Freiheitsgrade entspricht die t-Verteilung nahezu einer Normalverteilung.



Die t-Verteilung

- Bei einer geringen Zahl von Freiheitsgraden sind große t-Werte unter der Nullhypothese wahrscheinlicher.



- Vergleicht die Mittelwerte von zwei Stichproben.

- Zeigt die Wahrscheinlichkeit p, dass beide Testreihen den gleichen Mittelwert haben (Mittelwertdifferenz = 0)

- Zeigt, ob die Unterschiede der Mittelwerte per Zufall entstanden sind
  - Wenn p = 0.05 ist, dann ist zu 5% der Mittelwert der beiden Testreihen gleich.
  - ➢ Es ist relativ wahrscheinlich (95%), dass ein Unterschied der Mittelwerte nicht durch Zufall entstanden ist.

Signifikanzstufen:

- Ein Testergebnis heißt statistisch signifikant, wenn der p-Wert unterhalb des vorgegebenen Fehlers liegt.
- Dabei gibt es klassischerweise drei Signifikanzstufen:
  - $p \leq 0{,}05$ signifikant *
  - $p \leq 0{,}01$ sehr signifikant **
  - $p \leq 0{,}001$ höchst signifikant ***

- Es gibt T-Tests für **Within-Group Experimente** und **Between-Group Experimente**

- Es wird unterschieden, ob bei den Werten eine **Abhängigkeit** von den Testpersonen besteht oder nicht.
  - **Within-Group Experimente** benötigen **abhängige T-Tests**, da jede Testperson alle Level durchläuft.
  - **Between-Group Experimente** benötigen **unabhängige T-Tests**, da jede Testperson nur einen Level durchläuft.

- Mittelwert vor Verwendung des Systems:

$$\bar{x}_1 = {}^{291}\!/_{10} = 29{,}1$$

- Mittelwert nach Verwendung des Systems:

$$\bar{x}_2 = {}^{248}\!/_{10} = 24{,}8$$

- Mittelwert der Differenzen:

$$\bar{d} = {}^{43}\!/_{10} = 4{,}3$$

- Standardabweichung der Differenzen

$$s = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}}{n-1}} = \sqrt{\frac{359 - \frac{43^2}{10}}{9}} = 4{,}398$$

Vp	vor	nach	d	d^2
1	30	20	10	100
2	22	24	-2	4
3	38	31	7	49
4	34	28	6	36
5	25	20	5	25
6	28	28	0	0
7	33	27	6	30
8	21	24	-3	9
9	29	21	8	64
10	31	25	6	36
Σ	291	248	43	359

1. Berechnung der Prüfgröße t und der Freiheitsgrade df
   - Prüfgröße t

$$t = \frac{\left|\bar{d}\right| \cdot \sqrt{n}}{s} = \frac{4{,}3 \cdot \sqrt{10}}{4{,}398} = 3{,}092$$

   - Anzahl der Freiheitsgrade

$$df = n - 1 = 10 - 1 = 9$$

df	p=0,05	p=0,01	p=0,001
1	12,706	63,657	636,619
2	4,303	9,925	31,599
…	…	…	…
9	2,262	3,250	4,781
10	2,228	3,169	4,587

2. Prüfgröße t mit Wert in Signifikanz-Tabelle vergleichen
   - Wert in der Tabelle für 9 Freiheitsgrade und p = 0,05: 2,262.
   - Der errechnete Wert von 3,092 liegt über dem Wert aus der Tabelle.
   - Damit ist der Unterschied signifikant nachgewiesen.
     t(9)= 3,092, p<=0.05

Anwendung:

- Vergleich von zwei unabhängigen Stichproben hinsichtlich ihrer Mittelwerte (Between-Group)

Voraussetzung:

- Normalverteilung der Werte der Stichproben
- Wissen über die Varianzen nötig!

Vorgehen:

1. Überprüfung auf Varianzhomogenität
2. Berechnung der Prüfgröße t und der Freiheitsgrade df
3. Errechneten Wert t mit Wert in Signifikanz-Tabelle vergleichen

1. Überprüfung auf Varianzhomogenität

- Berechne Prüfgröße

$$F = \frac{s_{major}^2}{s_{\min or}^2}$$

s_{major}: größere Standardabweichung
s_{minor}: kleinere Standardabweichung

- Die Prüfgröße F ist F-verteilt mit (df ist die Anzahl der Freiheitsgrade)

$$df = (n_{major} - 1, n_{\min or} - 1)$$

- Varianzenheterogenität wird bei einem Signifikanzniveau p < 0,05 angenommen.
  - d.h. die Varianz (Streuung) unterscheidet sich signifikant

Universität Augsburg University

2. Berechnung der Prüfgröße t und des Freiheitsgrads df

- Bei Varianzhomogenität

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

$$df = n_1 + n_2 - 2$$

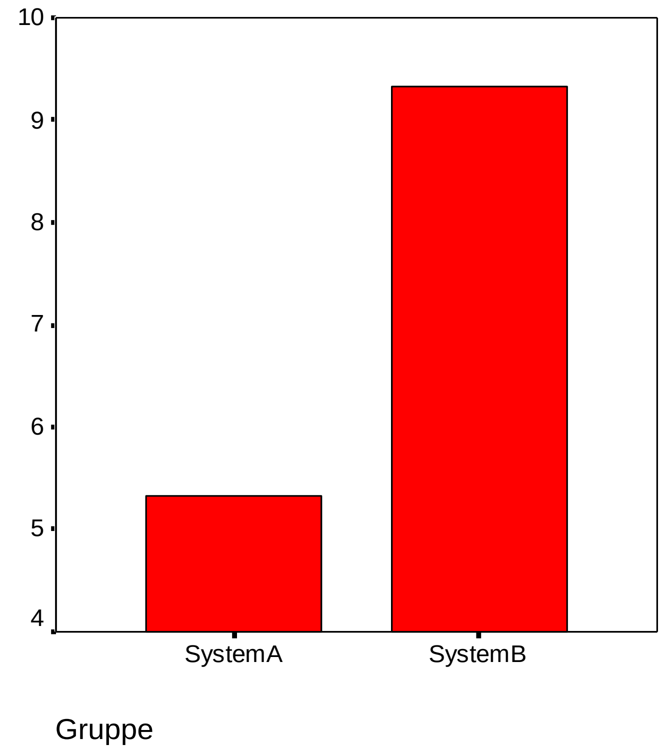
- Bei Varianzheterogenität

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{n_1 + n_2 - 2}{2}$$

Beispiel 1:

- **Situation:**
  - Zwei Systeme A und B werden hinsichtlich der Antwortzeiten miteinander verglichen.
  - Die Mittelwerte der dabei auftretenden Messwerte zeigen Unterschiede.



Universität Augsburg University

1. Überprüfung auf Varianzhomogenität

SystemA	SystemB
12	8
10	3
6	5
Summe: 28	Summe: 16
$\bar{x}_1$=9,333	$\bar{x}_2$=5,33333
s_1=3,05505	s_2=2,51661

- Berechne Prüfgröße

$$F = \frac{s^2_{major}}{s^2_{minor}} = \frac{3{,}05505^2}{2{,}51661^2} = 1{,}47369$$

- laut F-Tabelle ist dies bei (2,2) Freiheitsgraden ein **nicht signifikanter** Wert, da 1,47 < 19

	df1		
df2	**1**	**2**	**…**
1	162	200	…
2	18,51	19	…
…	…	…	…

- Für eine Signifikanz müsste der Wert größer als 19 sein.

➢ **Varianzhomogenität**

2. Berechnung der Prüfgröße t und der Freiheitsgrade df

$$t = \frac{|9{,}33333 - 5{,}33333|}{\sqrt{\frac{2 \cdot 3{,}05505^2 + 2 \cdot 2{,}51661^2}{4}}} \cdot \sqrt{\frac{9}{6}} = 1.7504 \qquad df = n_1 + n_2 - 2 = 3 + 3 - 2 = 4$$

3. Prüfgröße t mit Wert in Signifikanz-Tabelle vergleichen

- Der Wert in der Tabelle für vier Freiheitsgrade (df = 4) und p = 0.05 beträgt 2,776 (siehe nächste Folie)
- Für die Signifikanz von p = 0.05 müsste der berechnete t-Wert also **größer als 2,776** sein.
- Der Wert ist also für df = 4 **nicht signifikant**, da 1,7504 < 2,776

3. Prüfgröße t mit Wert in Signifikanz-Tabelle vergleichen

- Der Wert ist also nicht signifikant für 4 Freiheitsgrade. Für die Signifikanz von p = 0.05 müsste der berechnete t-Wert größer als 2,776 sein!

Freiheitsgrade	p=0,05	p=0,01	p=0,001
4	2,776	4,604	8,610
…			
10	2,228	3,169	4,587

- ➢ Damit kann die Nullhypothese **nicht** verworfen werden
- ➢ Nicht einmal mit 95% Sicherheit kann auf die bessere Leistung von System B geschlossen werden

Beispiel 2: Varianzenhomogenität

- Mehr Testwerte (6) als in Beispiel 1

1. Überprüfung auf Varianzhomogenität

- Berechne Prüfgröße

$$F = \frac{s^2_{major}}{s^2_{\min or}} = \frac{2{,}73252^2}{2{,}25093^2} = 1{,}47368$$

SystemA	SystemB
12	8
10	3
6	5
12	8
10	3
6	5
Summe: 56	Summe: 32
$x_1=\boxed{9{,}33333}$	$x_2=\boxed{5{,}33333}$
$s_1=\boxed{2{,}73252}$	$s_2=\boxed{2{,}25093}$

- laut F-Tabelle ist dies bei (5,5) Freiheitsgraden ein **nicht signifikanter** Wert, da 1,47 < 5,05
- **Varianzhomogenität** ist also gegeben

	df1		
df2	**1**	**…**	**5**
1	162	…	230
…	…	…	…
5	6,61	…	5,05

Universität Augsburg University

2. Berechnung der Prüfgröße t und der Freiheitsgrade df

$$t = \frac{|9{,}33333 - 5{,}33333|}{\sqrt{\frac{5 \cdot 2{,}73252^2 + 5 \cdot 2{,}25093^2}{10}}} \cdot \sqrt{\frac{36}{12}} = 2{,}7676 \qquad df = n_1 + n_2 - 2 = 10$$

3. Prüfgröße t mit Wert in Signifikanz-Tabelle vergleichen

- Der Wert in der Tabelle für 10 Freiheitsgrade (df = 10) und p = 0,05 beträgt 2,228 (siehe nächste Folie)
- Für die Signifikanz von p = 0.05 müsste der berechnete t-Wert also **größer als 2,2278** sein.
- Der Wert ist also für df = 10 **signifikant**, da 2,7676 > 2,2278

3. Prüfgröße t mit Wert in Signifikanz-Tabelle vergleichen

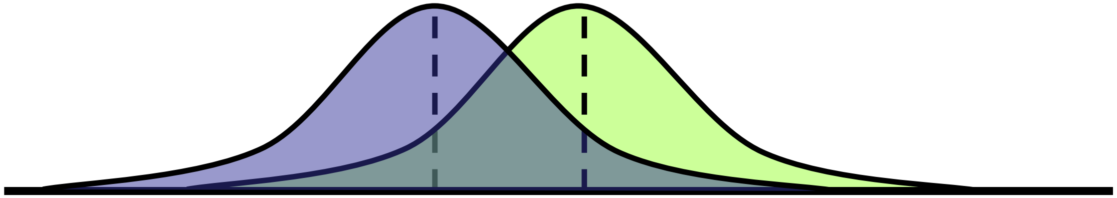
- t = 2,7676 ist größer als 2,2278 => Signifikanz.

Freiheitsgrade	p=0,05	p=0,01	p=0,001
4	2,776	4,604	8,610
…			
10	2,228	3,169	4,587

- Damit kann die Nullhypothese verworfen werden
- D.h. mit 95% Sicherheit kann auf die bessere Leistung von System B geschlossen werden. Formal: t(10)=2.7676, p<=0.05

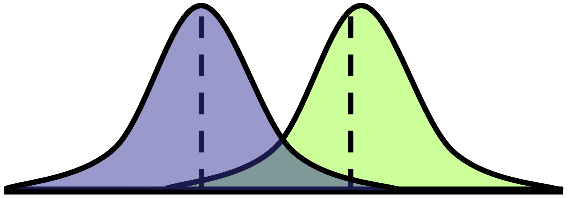
Vergleich der Mittelwerte:

- Beispiel 1: Die Spitzen trennen sich nicht klar voneinander.



  - Keine Signifikanz

- Beispiel 2: Die Spitzen trennen sich jetzt (mit mehr Versuchspersonen) klarer voneinander.



  - Signifikanz

Globalübung

-

Beispiel für ein Laborexperiment

Beispiel für ein Laborexperiment –
Problemstellung

Ist die Usability der neuen Webseite des HCM-Lehrstuhls besser, als die Usability der alten Webseite?

Aus welchen drei Kriterien setzt sich Usability zusammen? (ISO Norm)

- Effizienz
- Effektivität
- Zufriedenheit

Beispiel für ein Laborexperiment – Variablen

- Unabhängige Variable: Eingesetzte Webseite
  - Level 1: Alte Webseite
  - Level 2: Neue Webseite

- Abhängige Variable (hängen von der eingesetzten Webseite ab):
  - Effizienz
  - Effektivität
  - Zufriedenheit

Hypothesen:

- H1: Die Effizienz der neuen Webseite ist besser als die der alten.
  - H0-1: Es gibt keinen Unterschied bei der Effizienz.
- H2: Die Effektivität der neuen Webseite ist besser als die der alten.
  - H0-2: Es gibt keinen Unterschied bei der Effektivität.
- H3: Die Zufriedenheit mit der neuen Website ist besser als die mit der alten.
  - H0-3: Es gibt keinen Unterschied bei der Zufriedenheit.

Beispiel für ein Laborexperiment – Messung der abhängigen Variablen

- Wie misst man die abhängigen Variablen?
  - Effizienz
    - Zeitmessung für die Erledigung einer Aufgabe
  - Effektivität
    - Messung der Fehleranzahl bei der Erledigung einer Aufgabe
  - Zufriedenheit
    - Messung der Zufriedenheit mittels eines standardisierten Fragebogens

Mögliche Messtechniken:

- Effizienz und Effektivität
  - Interaktionslogging
  - Videoaufzeichung
  - Screen-Recording

  - ➢ Beobachtungstechnik
  - ➢ Objektive Daten
  - ➢ Quantitative (Interaktionslogging) und Qualitative Daten (Videoaufzeichnung und Screen-Recording)

Mögliche Messtechniken:

- Zufriedenheit
  - Geschlossene Fragen (SUS-Fragebogen)
  - Offene Fragen zu Problemen und sonstigem Feedback

  - ➢ Befragungstechnik
  - ➢ Subjektive Daten
  - ➢ Quantitative (SUS) und Qualitative Daten (offene Fragen)

- Within-Group Ansatz
  - Jede Testperson durchläuft jeden Level, d.h. jeder Teilnehmer benutzt beide Webseiten
  - Reihenfolge wird beachtet:
    - Die eine Hälfte der Testpersonen fängt mit der neuen Webseite an und benutzt dann die alte Seite.
    - Die andere Hälfte benutzt erst die alte Seite und dann die neue Seite.

Beispiel für ein Laborexperiment – Durchführung

- 20 Testpersonen, die auf die Spezifikation der Personas passen, werden eingeladen. (z.B. Studenten)

- Durchlauf von drei Tasks für jede Webseite:
  1. Suchen sie nach Informationen zur Vorlesung „Multimedia 1: Usability Engineering“
  2. Laden sie sich den aktuellen Foliensatz herunter.
  3. Informieren sie sich über ihren Prüfungstermin der 2. mündlichen Prüfung

- Beantwortung des Fragebogens jeweils direkt nach der Nutzung einer Webseite

Beispiel für ein Laborexperiment – Auswertung

1. Nötige Vorarbeit für die Auswertung
   - Annotation der Video- und Screen-Recordings
   - Annotation der offenen Fragen

2. Durchführung einer statistischen Analyse der Ergebnisse

Beispiel für ein Laborexperiment –
Auswertungsbeispiel: Effizienz

- Nötige Zeit **in Sekunden** für Task 1:

Versuchsperson	Alte Webseite	Neue Webseite
VSP1	23	15
VSP2	43	32
VSP3	23	14
VSP4	43	23
VSP5	22	10
VSP6	34	26
VSP7	33	27
VSP8	24	24
VSP9	44	12

Beispiel für ein Laborexperiment – Auswertungsbeispiel: Effizienz

- Nötige Zeit in Sekunden für Task 1:
  - Mittelwert
    - Alte Webseite: **32,11 Sekunden**
    - Neue Webseite: **20,33 Sekunden**
  - Standardabweichung
    - Alte Webseite: **9,47 Sekunden**
    - Neue Webseite: **7,73 Sekunden**
  - T-Test
    - **p = 0.0051**
    - ➢ Die neue Webseite ist **sehr signifikant** effizienter für Task 1 als die alte Webseite, da die Wahrscheinlichkeit, dass die beiden Mittelwerte der Testreihen gleich sind, **kleiner als 0,01** ist.

Beispiel für ein Laborexperiment – Auswertungsbeispiel: Zufriedenheit

- Nach dem SUS-Fragebogen zum Statement „I thought the system was easy to use.” (10 Punkte Skala):

Versuchsperson	Alte Webseite	Neue Webseite
VSP1	4	7
VSP2	5	8
VSP3	6	2
VSP4	4	5
VSP5	5	4
VSP6	3	6
VSP7	5	7
VSP8	5	3
VSP9	5	5

Beispiel für ein Laborexperiment –
Auswertungsbeispiel: Zufriedenheit

- Nach dem SUS-Fragebogen zum Statement „I thought the system was easy to use.” (10 Punkte Skala):
  - Mittelwert
    - Alte Webseite: **4,67**
    - Neue Webseite: **5,22**
  - Standartabweichung
    - Alte Webseite: **0,87**
    - Neue Webseite: **1,99**
  - T-Test
    - **p = 0.525**
    - ➢ Die neue Webseite ist **nicht signifikant** zufriedenstellender für das Statement „I thought the system was easy to use.” als die alte Webseite, da die Wahrscheinlichkeit, dass die beiden Mittelwerte der Testreihen gleich sind, **größer als 0,05** ist.

- Statistische Analysen sind nötig, um sicherzustellen, dass der Unterschied zwischen Mittelwerten zweier Stichproben kein Zufall ist.
- Null-Hypothesen werden widerlegt bzw. die Hypothesen werden belegt, wenn die Mittelwertvergleiche signifikant unterschiedlich sind.

- Der t-Test erlaubt die Betrachtung von 2 Mittelwerten.
- Was, wenn 3 oder mehr Mittelwerte vorliegen?
  - Beispiel: 3 Gesten zur Datenübertragung mit einer App werden miteinander verglichen?

➢ Varianzanalyse

Varianzanalyse –
ohne Messwiederholung (Between-Group)

- Bei unabhängigen Stichproben mit mehr als 2 Mittelwerten, kommt die einfache Varianzanalyse mittels ANOVA zum Einsatz (ANalysis Of VAriance)

- Die unabhängige Variable nennt man auch Faktor, welcher die Daten in die einzelnen Faktorstufen gruppiert.

- Null-Hypothese: Alle Gruppenmittelwerte der Variablen in der Grundgesamtheit sind identisch.

- Voraussetzung: Normalverteilung der Grundgesamtheit, Varianzenhomogenität

Einfaktorielle ANOVA

- Erweiterung des t-Test für mehr als 2 Mittelwerte.
- Wird auf eine Testvariable angewendet, bei der die Werte verschiedenen Fallgruppen (mehr als zwei Level) angehören.

- **Grob erläutert:**
  - Das Prinzip der Varianzanalyse ist eine Zerlegung der Gesamtvarianz (über alle Gruppen) in eine Varianz innerhalb der Gruppen und eine Varianz zwischen den Gruppen.
  - Die Betrachtung der Abweichungen der verschiedenen Varianzen führt zu einer Prüfgröße, welche einer F-Verteilung folgt.
  - ➢ Signifikante Unterschiede der Mittelwerte, falls die Varianzen innerhalb und zwischen den Gruppen nicht zufällig entstanden sind.

Einfaktorielle ANOVA

- Die Berechnungen hierfür sind sehr rechenintensiv.
- Als Ergebnis wird die **Prüfgröße F** (bei gegebenen **Freiheitsgraden**) und ein **p-Signifikanzniveau** erwartet.

- Beispiel: F = 4.467, df_innerhalb=27, df_zwischen=2, p=0.021
  Ergebnis formuliert als: $F(2,27)=4.467$, $p<0.05$

- Bei signifikantem Unterschied der Mittelwerte ist lediglich bekannt, dass es einen Unterschied gibt. Nicht, aber welche Mittelwerte bzw. Level der unabhängigen Variable er betrifft.
- Hierfür gibt es **Post-Hoc-Analysen**, welche die Zwischenergebnisse der Varianzanalyse nutzt und die Unterschiede zwischen den Gruppen aufzeigt. (ähnlich einem paarweisen Vergleich)

Effektgröße

- **Achtung:**
  - Die Tatsache, dass ein Unterschied signifikant ist, heißt nicht unbedingt, dass er auch bedeutsam ist.
  - Die Größe eines Effekts ist für die inhaltliche Bewertung eines signifikanten Ergebnisses wichtig, da durch eine Erhöhung des Stichprobenumfangs (N) theoretisch jeder noch so kleine Effekt signifikant gemacht werden kann.
- **Beispiel:**
  - Vergleich der Intelligenzleistung von Kindern
  - Unabhängige Variable: Lehrmethode (Level 1: neu, Level 2: alt)
  - Bei einer sehr großen Anzahl von Kindern pro Stichprobe (N = 5.000), können schon Unterschiede von beispielsweise 0.1 IQ-Punkten zwischen den Gruppen zu signifikanten Unterschieden führen.
  - Ganz klar bedeuten 0.1 IQ-Punkte Unterschied aber trotz eines signifikanten Testergebnisses kaum eine Verbesserung.

Effektgröße

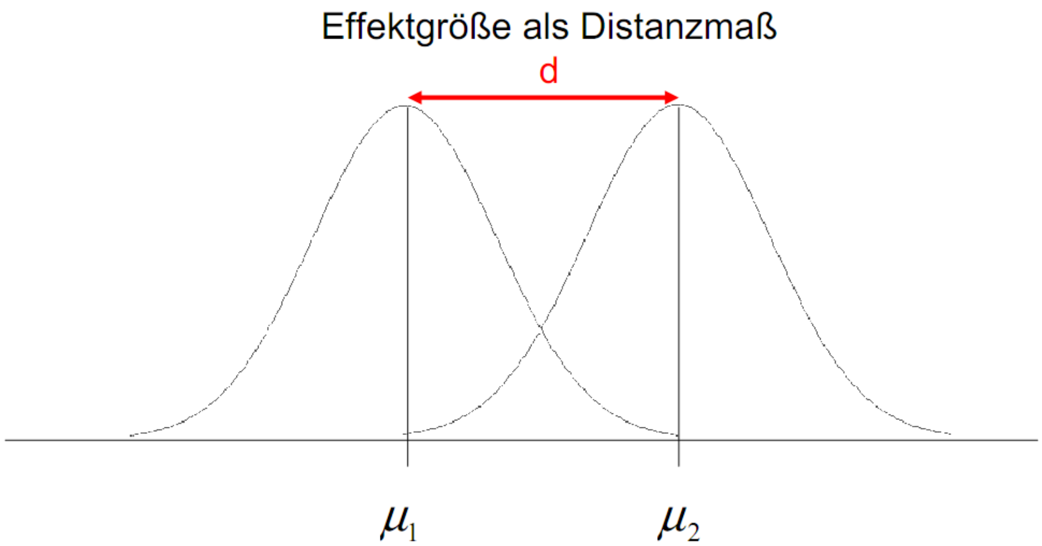
- Effekte auf zwei Ebenen:
  - **Empirische Effekte**, die das Ergebnis einer Untersuchung beschreiben
  - **Populationseffekte**, die entweder angenommen oder aus den empirischen Daten geschätzt werden müssen.

Effektmaße

- Effekt als absolute Größe
  - Unterschied zwischen gemessenen Mittelwerten zweier Stichproben
  - Unterschied ist eine Schätzung für die Größe des systematischen Effekts

- Effektmaße sollten standardisiert sein, um die Effekte unterschiedlicher Untersuchungen miteinander vergleichen zu können.
- Standardisierte Effektmaße
  - Effekt als Distanz zwischen Populationsmittelwerten
  - Effekt als Varianzquotient

Effektmaße –
Distanz zwischen Populationsmittelwerten



Abstand der Mittelwerte (d) zweier unterschiedlicher Populationen normiert an der mittleren Standardabweichung

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 + s_2^2)/2}}$$

Konventionen für die Effektgröße d für t-Tests mit unabhängigen Stichproben:

- Die Beurteilung, ob ein Effekt eher groß oder klein zu bewerten ist, unterliegt den inhaltlichen Überlegungen des Forschers.
- Mit der Zeit etablierte Konventionen für die Größe von Effekten sind:
  - Kleiner Effekt: d=0,20
  - Mittlerer Effekt: d=0,50
  - Großer Effekt: d=0,80
- Beispiel: Experiment mit TimeStack und TimePie (Anzahl der Fehler)

$$\bar{x}_{Stack} = 7{,}2 \qquad \hat{\sigma}_{Stack} = 3{,}162 \qquad \bar{x}_{Pie} = 11 \qquad \hat{\sigma}_{Pie} = 4{,}14$$

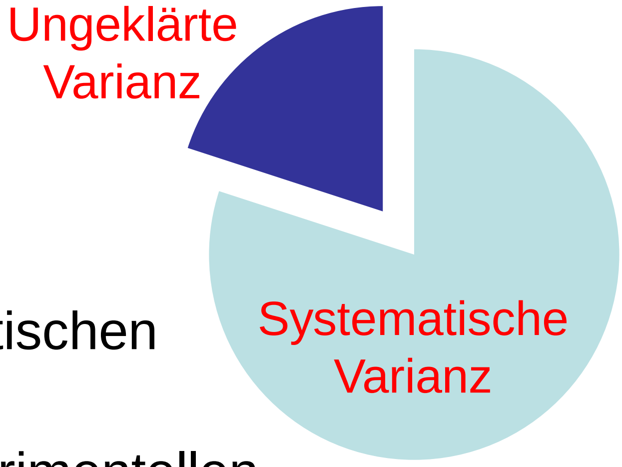
$$\hat{\sigma}_x = \sqrt{(3{,}162^2 + 4{,}14^2)/2} = 3{,}684$$

$$d = (11 - 7{,}2)/3{,}684 = 1{,}03$$

➢ Es handelt sich um einen großen Effekt.

Effektmaße –
Effekt als Varianzquotient

- Zwei Gründe für Variabilität zwischen Stichproben
  - Manipulation durch das Experiment
  - Individuelle Unterschiede



- **Effektstärkemaß Ω^2**

$$\Omega^2 = \frac{\sigma^2_{sys}}{\sigma^2_{Gesamt}}$$

  - drückt aus, wie groß der Anteil der systematischen Varianz an der Gesamtvarianz ist.
  - erfasst die Größe des Einflusses einer experimentellen Manipulation auf die Gesamtvarianz.
- Bewertung von Effekten hängt von inhaltlichen Überlegungen ab
- Richtwerte:
  - Kleiner Effekt: $\Omega^2 = 0{,}01$
  - Mittlerer Effekt: $\Omega^2 = 0{,}06$
  - Großer Effekt: $\Omega^2 = 0{,}14$

Effektstärkemaß Ω²

- Schätzung anhand empirischer Daten

$$\hat{\Omega}^2 = \frac{f^2}{1+f^2} \quad mit \quad f^2 = \frac{t^2-1}{N}$$

N Anzahl aller Versuchspersonen

- Beispiel: Experiment mit bildhaftem und textuellen Interface

$$t(df = 98) = 5{,}16 \qquad df = n1 + n2 - 2$$

$$f^2 = \frac{5{,}16^2 - 1}{100} = 0{,}256$$

$$\hat{\Omega}^2 = 0{,}256 / (1 + 0{,}256) = 0{,}20$$

- Der geschätzte Effekt zwischen den Bedingungen beträgt 20%.

Fehlentscheidungen beim Testen

Wiederholung:

- **Fehler 1. Art (Signifikanzniveau): das unberechtigte Ablehnen der Nullhypothese**

  p(Fehler 1. Art) = α
- **Fehler 2. Art: das unberechtigte Beibehalten der Nullhypothese**

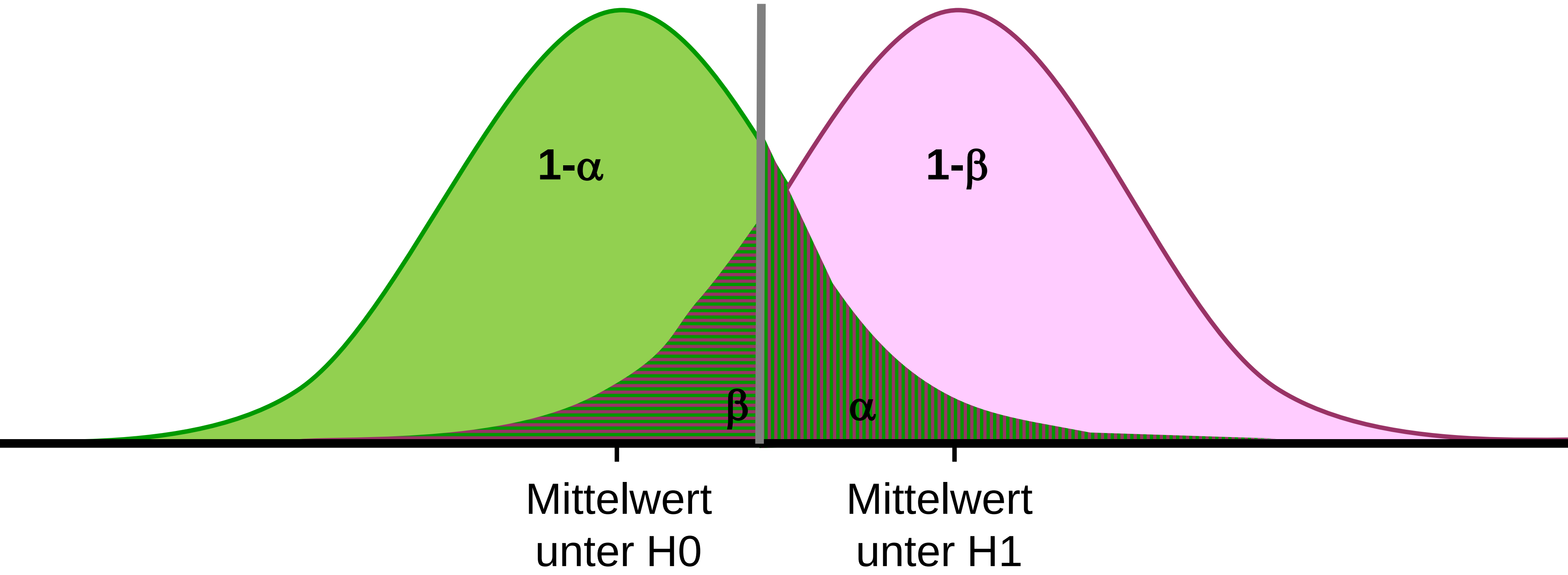
  p(Fehler 2. Art) = β

α	Nicht existierender Unterschied wird als Effekt ausgegeben
$1-\alpha$	Nicht existierender Unterschied wird auch als solcher erkannt
β	Vorhandener Effekt wird nicht entdeckt
$1-\beta$	Vorhandener Effekt wird auch entdeckt

Die 4 Möglichkeiten des Entscheidungsproblems

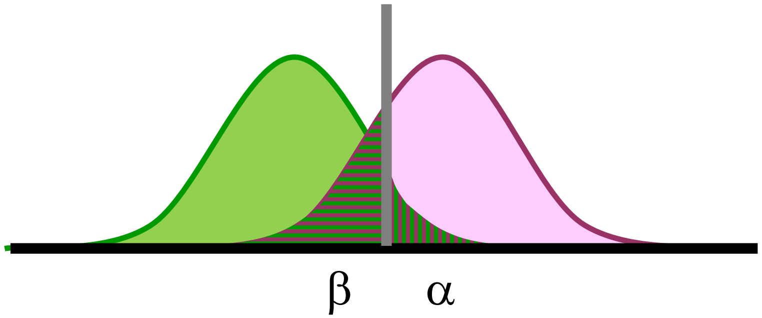
Fehlentscheidungen beim Testen

- Falls die Alternativhypothese gilt, dann machen wir in α der Fälle einen Fehler, in 1- β der Fälle liegen wir richtig.
- Falls die Nullhypothese gilt, dann machen wir in β der Fälle einen Fehler, in 1- α der Fälle liegen wir richtig.



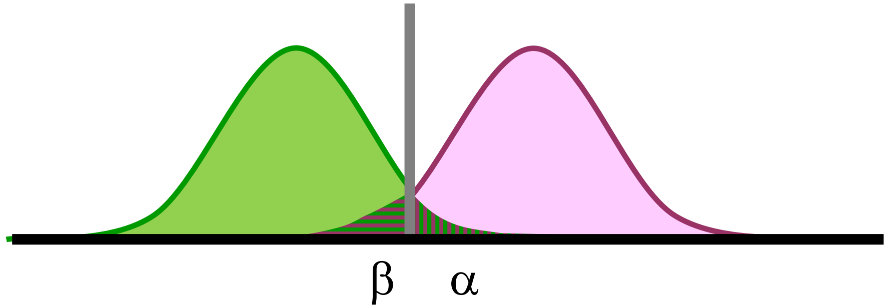
Fehlentscheidungen beim Testen - Der Einfluss des Effekts

- kleiner angenommener Effekt: Verteilungen der H_0 und H_1 liegen eng zusammen und überschneiden sich in der Regel stark.



  - Die Wahrscheinlichkeit eines β-Fehlers, d.h. dass ein vorhandener Unterschied nicht erkannt wird, ist relativ hoch.

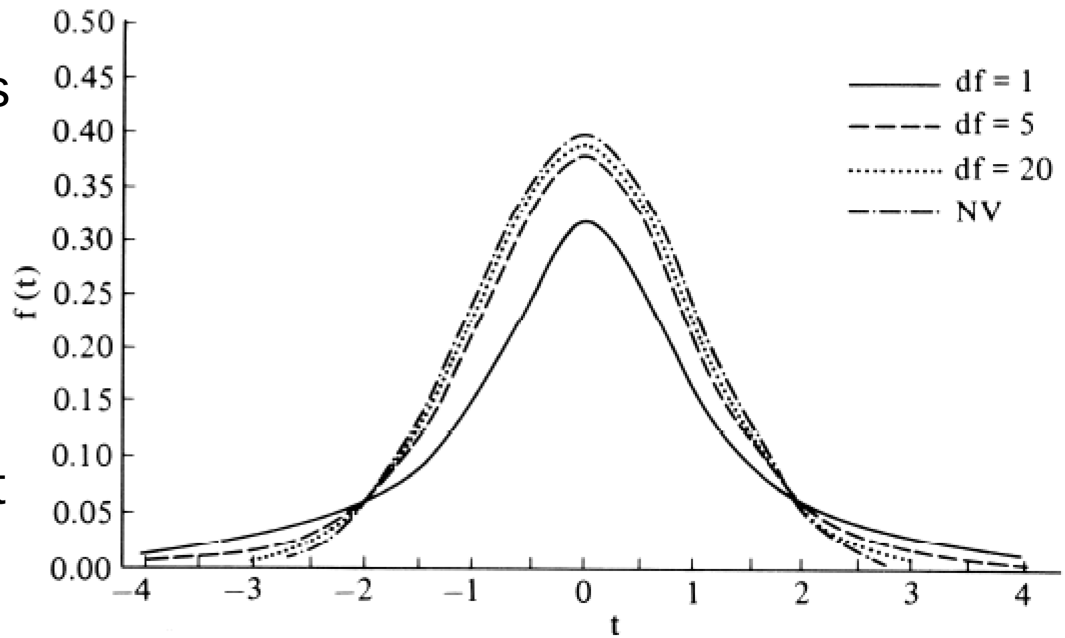
- größerer angenommener Effekt: die β-Fehler-Wahrscheinlichkeit bei gleichem α und gleicher Streuung wird kleiner, die Teststärke größer.



Fehlentscheidungen beim Testen - Einfluss des Stichprobenumfangs

- Je größer die Stichprobe, desto schneller wird ein bestimmter t-Wert signifikant, da der kritische t-Wert für ein bestimmtes Signifikanzniveau von der Freiheitsgradzahl abhängt.

- Dies gilt aber nur für Stichproben, die kleiner als 30 sind.
- Bei größeren Stichproben schmiegt sich die t-Verteilung bereits eng an eine Normalverteilung an und die Wahrscheinlichkeit für die t-Werte verändert sich nur noch geringfügig.



β-Fehler und Teststärke

- **Achtung:**
  - Ein nicht signifikantes Ergebnis nach einem t-Test erlaubt nicht unbedingt die Entscheidung für die Nullhypothese.
  - Die Wahrscheinlichkeit für den β-Fehler, d.h. dass man einen vorhandenen Unterschied nicht erkennt bzw. die Nullhypothese fälschlicherweise annimmt, sollte bei 10% oder weniger liegen.
  - Ist sie größer, so spricht ein nicht signifikantes Ergebnis für keine der beiden Hypothesen.
  - Es ist keine Entscheidung möglich.

➢ Aufgewendete Zeit und Mühe waren umsonst, da keine weiterführende Erkenntnis durch das Experiment gewonnen wurde.

➢ Sowohl die Nullhypothese als auch die Alternativhypothese sind immer noch möglich.

β-Fehler und Teststärke

- Mit Hilfe des β-Fehlers kann eine Aussage darüber getroffen werden, wie gut ein t-Test konstruiert ist.

- **Teststärke oder Power eines t-Tests**
  - Fähigkeit eines Tests, einen Effekt zu finden, falls dieser tatsächlich existiert.
  - Wahrscheinlichkeit, die Alternativ-Hypothese H1 anzunehmen, wenn sie auch in Wirklichkeit gilt.
  - Wird mit 1-β bezeichnet, da sie die Gegenwahrscheinlichkeit zu der β-Fehler-Wahrscheinlichkeit ist.

Teststärke oder Power eines t-Tests

- Spielt bei der Planung und Beurteilung von t-Tests eine große Rolle.
- Sollte mindestens $1-\beta=0{,}9$ betragen.
- Wird abgeschätzt durch den Wert λ:

$$\lambda = \frac{\Omega^2}{1-\Omega^2} * N$$

Ω^2: Effektstärkemaß,

N: Stichprobenumfang

Beispiel:

Test-stärke	**0,1**	**0,5**	**0,6667**	**0,75**
λ	0,0	2,7	4,31	5,30

➢ Je größer der Effekt ist, desto weniger Versuchspersonen sind nötig, um eine Entscheidung für bzw. gegen die Nullhypothese treffen zu können.

Zwei Arten der Teststärkebestimmung

$$\lambda = \frac{\Omega^2}{1-\Omega^2} * N$$

Ω^2: Effektstärkemaß,
N: Stichprobenumfang

- **A priori:**
  - Vor der Berechnung eines t-Tests ist es notwendig, eine gewünschte Teststärke festzulegen.
  - Die a priori Bestimmung der Teststärke führt zusammen mit der Entscheidung für einen bestimmten inhaltlich relevanten Effekt zu der Berechnung des Stichprobenumfangs.

- **A posteriori:**
  - Die Berechnung der Teststärke eines bereits durchgeführten t-Tests ist dann notwendig, wenn ein nicht signifikantes Ergebnis auftritt und der Stichprobenumfang nicht im Vorfeld geplant wurde.
  - Die Annahme der Nullhypothese ist nur dann möglich, wenn die Teststärke ausreichend hoch ist.

β-Fehler und Teststärke

Beispiel:

- Nach einem einseitigen t-Test mit $n_1 = n_2 = 15$ ergibt sich bei einem Signifikanzniveau von $\alpha=0{,}05$ ein nicht signifikantes Ergebnis.
- Der Forscher erklärt einen mittleren Effekt von $\Omega^2=0{,}1$ als inhaltlich relevant.
- Die Berechnung von λ ergibt:

$$\lambda = \frac{\Omega^2}{1-\Omega^2} * N = \frac{0{,}1}{1-0{,}1} * 30 = 3{,}33$$

Test-stärke	0,1	0,5	0,6667	0,75
λ	0,0	2,7	4,31	5,30

- Die Teststärke dieses t-Tests, den Effekt von $\Omega^2=0{,}1$ zu finden, falls er wirklich existiert, liegt zwischen $50\% < 1-\beta < 66{,}7\%$.
- Die Entscheidung für die Nullhypothese, dass kein Effekt von mind. $\Omega^2=0{,}1$ vorliegt, wäre also mit einer β-Wahrscheinlichkeit von 33%-50% behaftet. In einem solchen Fall erlaubt das Ergebnis keine eindeutige Entscheidung. Der Test war also schlecht konstruiert.

Beispiel für die Durchführung eines Experiments

Universität Augsburg University

1. **Aufstellung einer Hypothese**
2. **Prüfung der Voraussetzungen**
3. **Festlegung des Populationseffekts**

   Beispiel: Ein Vergleich mit bereits durchgeführten Studien zum Thema „Zeitabhängige Datenvisualisierung“ ergibt die Erwartung eines großen Effekts **Ω^2=0,2**
4. **Festlegung des Signifikanzniveaus**

   Beispiel: Das Signifikanzniveau liegt per Konvention meist bei **α=0,05** und wird daher auch von uns so gewählt.
5. **Stichprobenumfangsplanung**

   Beispiel: Die Teststärke soll **1-β=0,9** betragen → **$\lambda_{90\%}$=8,56**

$$N = \frac{8{,}56}{\left(\frac{0{,}2}{1-0{,}2}\right)} = 34{,}24 \approx 36$$

Test-stärke	0,7500	0,8000	0,8500	0,9000
λ	5,30	6,18	7,19	8,56

6. **Bestimmung des kritischen t-Werts (vgl. Tabelle)**

Beispiel: $t_{krit(df=34)} \approx t_{krit(df=30)} = 1{,}697$ α=0,05 einseitige Fragestellung

7. **Prüfung des empirischen t-Werts auf Signifikanz**

Beispiel: Wir gehen von folgenden Werten aus:

$$N = 36 \quad \bar{x}_{Pie} = 11 \quad \bar{x}_{Stack} = 7{,}2 \quad t_{krit} = 1{,}697$$

Berechnung der Stichprobenstreuung aus den Daten:

$$\hat{\sigma}_1 = 4{,}14 \quad \hat{\sigma}_2 = 3{,}162$$

Schätzung der Streuung der Stichprobenkennwerteverteilung:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{4{,}14^2}{18} + \frac{3{,}162^2}{18}} = 1{,}23$$

Berechnung des empirischen t-Werts:

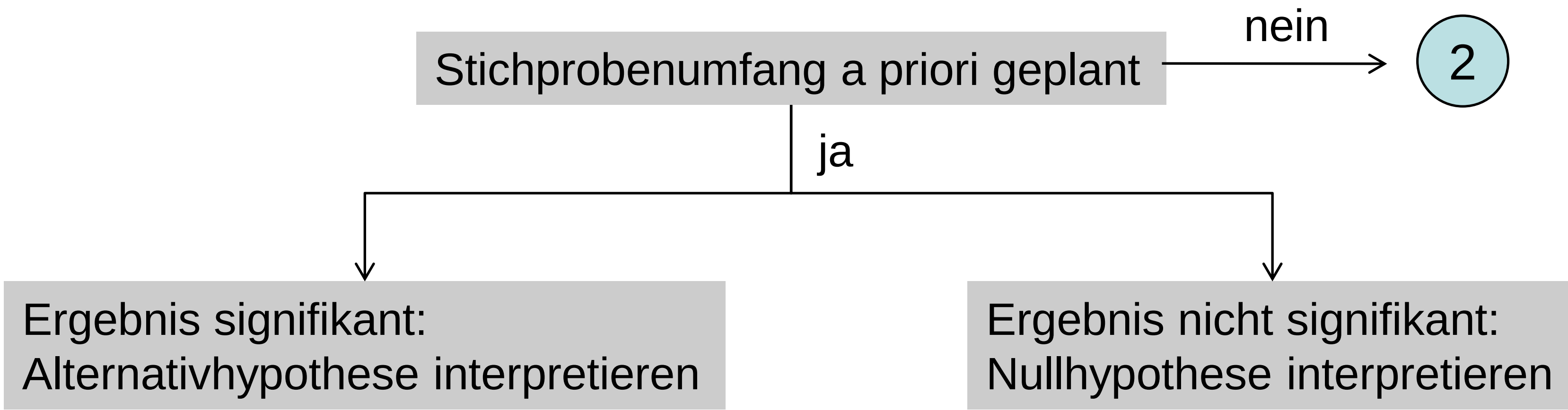
$$t_{(df=34)} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} = \frac{11 - 7{,}2}{1{,}23} = 3{,}1$$

df	0,95	0,975	0,990	0,995	0,9995
30	1,697	2,042	2,457	2,750	3,646
40	1,684	2,021	2,434	2,704	3,551

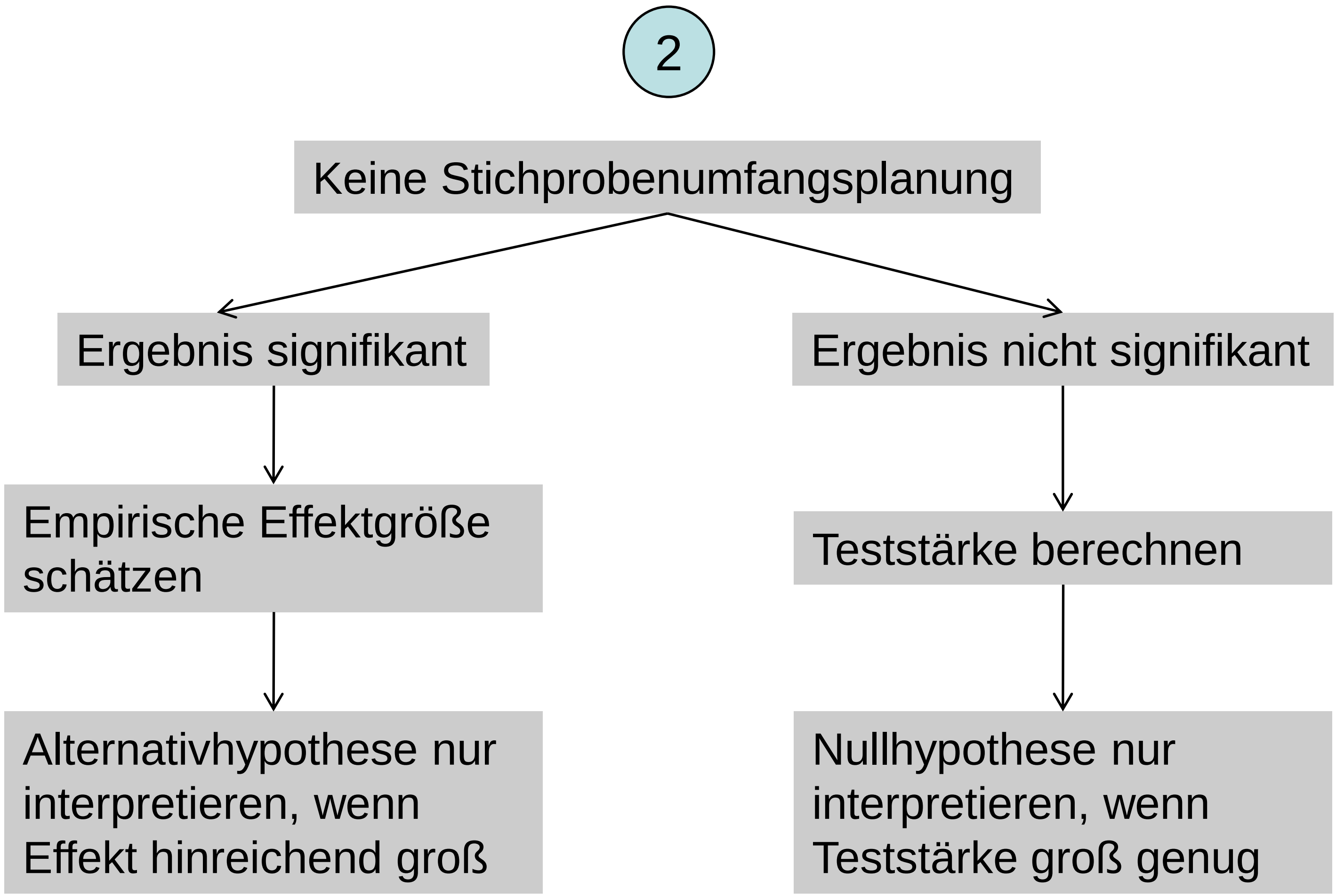
$t_{emp} > t_{krit} \quad p < 0{,}005 \Rightarrow$ Das Ergebnis ist sehr signifikant.

8. Interpretation der Ergebnisse
   - Ein in der beschriebenen Form konstruierter Test erlaubt die eindeutige Interpretation jeder bei der Auswertung auftretenden Mittelwertsdifferenz.
   - Signifikantes Ergebnis:
     - Annahme der Alternativhypothese
     - Fehlerwahrscheinlichkeit beträgt bei Signifikanzniveau von $\alpha = 5\%$ weniger als 5%
   - Nicht signifikantes Ergebnis:
     - Annahme der Nullhypothese
     - Fehlerwahrscheinlichkeit beträgt bei einer festgelegten Teststärke von $1-\beta = 0{,}9$ weniger als 10%

Entscheidungsdiagramm für die Bewertung eines t-Tests



Entscheidungsdiagramm für die Bewertung eines t-Tests



Zusammenfassung

- t-Test ist ein wichtiges Auswertungsverfahren für den Vergleich zweier Gruppenmittelwerte.
- Er liefert eine Entscheidungsgrundlage dafür ob es einen systematischen Unterschied zwischen zwei Gruppen gibt oder ob sich der gefunden Unterschied zufällig ergeben hat.
- Eine auf der Grundlage eines t-Tests getroffene Entscheidung ist mit einer bestimmten Wahrscheinlichkeit falsch.
- Wahrscheinlichkeiten der möglichen Fehler beruht auf ihrer gegenseitigen Abhängigkeit, der Größe des Effekts und dem Stichprobenumfangen.