



Bern University  
of Applied Sciences

Bern  
Movement  
Lab

# BSc PHY – Angewandte Statistik

## LE4 Statistische Tests

Patric Eichelberger & Aglaja Busch  
aF&E Physiotherapie



patric.eichelberger@bfh.ch | aglaja.busch@bfh.ch



Moodlekurs

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Intro

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Rückblick Schätzung

- ▶ Es gibt Punkt- und Intervallschätzer.
- ▶ Die Intervallschätzung liefert einen Bereich in dem der wahre Wert mit einer bestimmten Wahrscheinlichkeit liegt und ist zu bevorzugen.
- ▶ Auswahl der richtigen Schätzmethode ist wichtig:
  - ▶ Populationsmittelwert
    - ▶  $n > 100$  oder Daten normalverteilt: t-Test (one-sample)
  - ▶ Abhängige Daten
    - ▶  $n > 100$  oder Differenzen normalverteilt: t-Test (gepaart)
  - ▶ Unabhängige Daten
    - ▶  $n > 100$  oder Daten normalverteilt in beiden Gruppen: t-Test (ungepaart)
    - ▶ Falls Varianzen ungleich sind: Welch-Test
- ▶ Nicht normalverteilt: Behandlung in der Lerneinheit zu statistischen Tests!
- ▶ Ungleiche Varianzen: Behandlung in der Lerneinheit zu statistischen Tests!

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

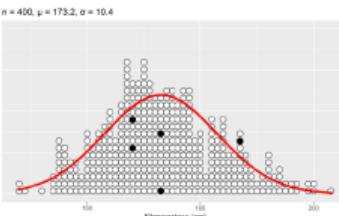
Stichprobengröße

Zusammenfassung

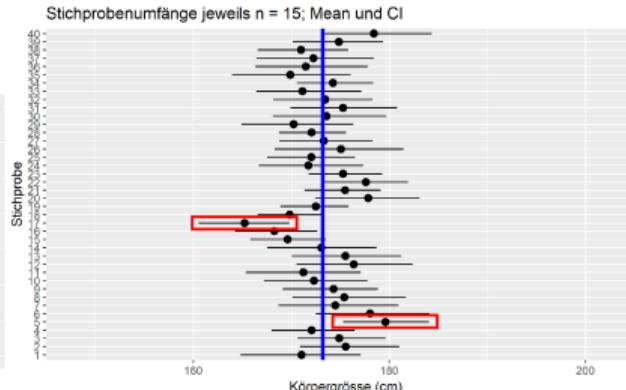
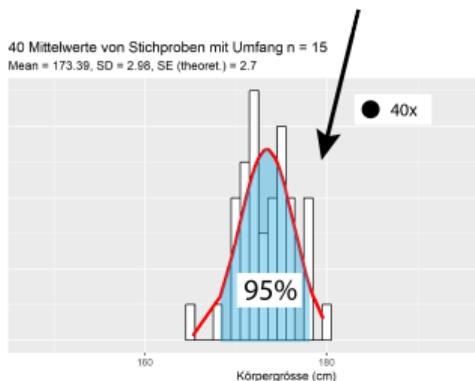


Bern University  
of Applied Sciences

# Rückblick Schätzung (cont.)



Der Mittelwert einer genügend grossen Anzahl Zufallsexperimente ist näherungsweise normalverteilt.



$$(1 - \alpha) \cdot 40 = 0.95 \cdot 40 = 38$$

Wenn wir das Experiment (Stichprobenmittelwert schätzen) 40x wiederholen, dann liegt der wahre Wert 38 Mal im Vertrauensintervall drin.

- Intro
- Fragestellung und Hypothesen
- Versuchsdesign und Daten
- Statistischer Test
- Wahrheit und Wahrscheinlichkeit
- Praxis mit R
  - Unabhängig parametrisch
  - Unabhängig nicht-parametrisch
  - Stichprobengrösse
- Zusammenfassung

# Lernziele Statistische Tests

- ▶ Die Studierenden kennen den Ablauf eines statistischen Tests.
- ▶ Die Studierenden können zu einer Fragestellung die Null- und Alternativhypothese formulieren und dir überlegen, wie man die Daten am besten erheben würde.
- ▶ Die Studierenden wissen, was ein p-Wert ist und kannst aus einem gegebenen p-Wert die richtige Schlussfolgerung über die Null- und Alternativhypothese ziehen.
- ▶ Die Studierenden können zwischen statistischer Signifikanz und praktischer Relevanz eines Testergebnisses unterscheiden und dieses kritisch reflektieren.
- ▶ Die Studierenden können zu einer gegebenen Fragestellung und einem gegebenen Datensatz den richtigen statistischen Test auswählen.
- ▶ Die Studierenden kennen den Zusammenhang zwischen statistischem Test und Vertrauensintervall.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

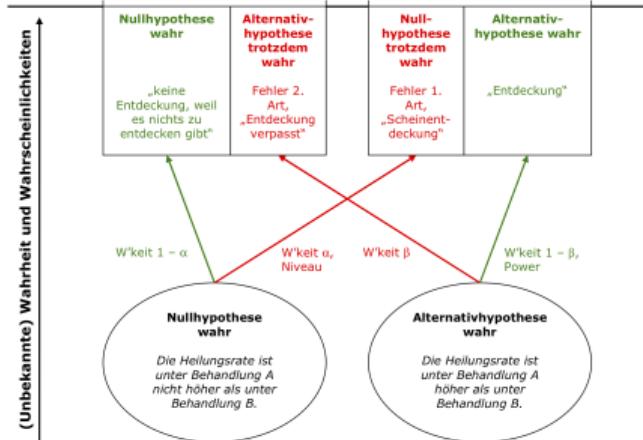
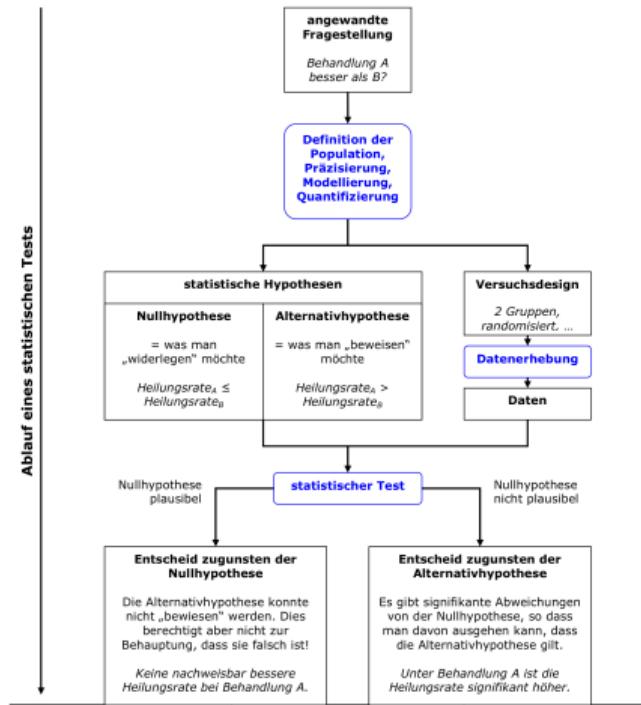


Bern University  
of Applied Sciences

# Statistischer Test

# Wahrheit und Wahrscheinlichkeiten

## Statistischer Test: Ablaufschema, Fehlerwahrscheinlichkeiten



Quelle: Michael Vock, IMSV, Universität Bern

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Fragestellung und Hypothesen

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Angewandte Fragestellung

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

- ▶ Ausgangspunkt ist die wissenschaftliche Fragestellung, die man untersuchen möchte. Zum Beispiel: Sind Frauen beweglicher als Männer?
- ▶ Dazu folgt eine Behauptung, die man mit Hilfe von Daten und statistischen Methoden untersuchen möchte.

## Beispiel

Behauptung: Frauen sind beweglicher als Männer.

# Präzisierung

Die Fragestellung bzw. Behauptung muss man meistens noch konkretisieren und präzisieren:

- ▶ Über welche Gruppe von Personen will man etwas aussagen?
- ▶ Wie kann man Beweglichkeit messen? Was sind die interessierenden Variablen?

## Beispiel

- ▶ Eine Idee wäre zum Beispiel das Merkmal Hüfte INNEN Rotation passiv in Flexion zu messen.
- ▶ Wir haben nicht die Möglichkeit eine zufällige Stichprobe aus der ganzen Schweiz oder der ganzen Welt zu ziehen.
- ▶ Wir beschränken uns auf die Studierenden der Physiotherapie an der BFH Gesundheit.

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung



Bern University  
of Applied Sciences

## ► Nullhypothese $H_0$

- ▶ Darunter versteht man die Behauptung die man ablehnen möchte.
- ▶ In unserem Beispiel: Das Merkmal Hüfte INNEN Rotation passiv in Flexion bei weiblichen und männlichen Studierenden der PHY an der BFH G ist gleich gross.

## ► Alternativhypothese $H_1$

- ▶ Diese wird der Nullhypothese gegenüber gestellt.
- ▶ Sie enthält die Motivation des Experiments.
- ▶ In unserem Beispiel: Das Merkmal Hüfte INNEN Rotation passiv in Flexion bei weiblichen und männlichen Studierenden der PHY an der BFH G ist nicht gleich gross.

# Statistische Hypothesen (cont.)

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengröße

Zusammenfassung

Die Hypothesen können mathematisch formuliert werden:

$$H_0 : \mu_w = \mu_m$$

und

$$H_1 : \mu_w \neq \mu_m$$

wobei  $\mu_w$  und  $\mu_m$  die mittlere Rotation der Hüfte innen passiv in Flexion weiblicher bzw. männlicher Studierender in der Population (Studierende der PHY an der BFH G) bezeichnen.



Bern University  
of Applied Sciences

# Versuchsdesign und Daten

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Versuchsdesign

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

- ▶ Passend zu den formulierten Hypothesen entwickelt man ein Versuchsdesign.
- ▶ Möglichst viele Einflussgrößen die einem nicht interessieren konstant halten.
  - ▶ Ermüdungsstatus zum Messzeitpunkt
  - ▶ Akute Infekte zum Messzeitpunkt
  - ▶ ...
- ▶ Ist dies nicht möglich, so muss man solche Störgrößen im Design und bei der Auswertung beachten.

# Ein- oder Mehrgruppenfall / abhängig oder unabhängig

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

- ▶ Ja nach Fragestellung hat man einen Ein- oder Mehrgruppenfall.
- ▶ In unserem Beispiel haben wir zwei Gruppen: Frauen und Männer
- ▶ Hat man mehrere Gruppen, können diese abhängig oder unabhängig (auch verbunden/unverbunden; gepaart/ungepaart) sein.
- ▶ In unserem Beispiel: unabhängige Gruppen, da die zu vergleichenden Messwerte von unterschiedlichen Personen stammen.
- ▶ Hätte man Messungen wiederholt an gleichen Personen gemacht (z.B. vor und nach Behandlung) hätte man abhängige Gruppen.

# Stichprobengrösse

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung

Die ideale Stichprobengrösse hängt von verschiedenen Faktoren ab:

- ▶ der Streuung der Daten
- ▶ dem relevanten Unterschied
- ▶ dem statistischen Test der benutzt wird
- ▶ dem Signifikanzniveau
- ▶ der Power

Ebenfalls müssen ökonomische Faktoren berücksichtigt werden, da Studien mit hohen Fallzahlen teuer sind.



Bern University  
of Applied Sciences

# Datenerhebung

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung

Schliesslich sammelt man die Daten, indem man Versuche bzw. Beobachtungen durchführt.

## Beispiel

- ▶ In unserem Beispiel müssten wir jetzt zufällig Studierende der PHY an der BFH G auswählen und deren Rotation der Hüfte innen passiv in Flexion messen.
- ▶ Unser Datensatz ist anders entstanden, der Übung wegen arbeiten wir aber trotzdem damit weiter.

# Daten

Wenn man die Daten schliesslich erhoben hat:

- ▶ Daten mit einer geeigneten Darstellung visualisieren
- ▶ Kennzahlen berechnen
- ▶ Für die statistischen Tests beurteilen ob die Daten normalverteilt sind

## Beispiel

Dies haben wir ja schon im Hinblick auf die Vertrauensintervalle gemacht und sind zum Schluss gekommen:

- ▶ Das Merkmal könnte bei den weiblichen Studierenden normalverteilt sein.
- ▶ Bei den männlichen Studierenden ist der Stichprobenumfang ( $n = 23$ ) zu klein um die Normalverteilung zu überprüfen ( $n < 30$ ).
- ▶ Die Varianz ist in beiden Gruppen etwa gleich.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung



Bern University  
of Applied Sciences

# Statistischer Test

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Statistischer Test

- Um den statistischen Test durchzuführen, muss zunächst die sogenannte **Teststatistik** aus den Daten berechnet werden.

$$\text{Teststatistik} = \frac{\text{Durch Modell erklärte Varianz}}{\text{Durch Modell nicht erklärte Varianz}} = \frac{\text{Effekt}}{\text{Fehler}}$$

- Im Falle des t-Tests zum Vergleich von Mittelwerten

$$t = \frac{\text{Differenz zw. Gruppenmittelwerten}}{\text{Standardfehler}}$$

## Merke

$t$  wird grösser falls:

- Die Differenz, d.h. der Effekt, grösser wird.
- Der Standardfehler kleiner wird: Stichprobenstreuung wird kleiner und/oder der Stichprobenumfang wird grösser.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

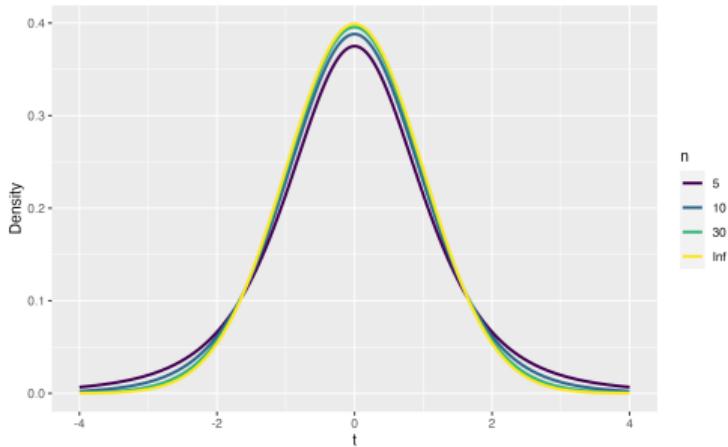
Stichprobengrösse

Zusammenfassung



Bern University  
of Applied Sciences

# Exkurs: t-Verteilung und t-Wert



- ▶ Beim Schätzen von Mittelwerten in normalverteilten Populationen.
- ▶ Wenn Stichproben klein sind.
- ▶ Wenn die Populations Standardabweichung nicht bekannt ist.

$$t = \frac{\text{Differenz zw. Gruppenmittelwerten}}{\text{Standardfehler}} = \frac{d}{SE} = \frac{d}{\frac{SD}{\sqrt{n}}}$$

## Merke

Wenn t-Verteilung vorliegt müssen auf der x-Achse andere Werte gewählt werden als bei Normalverteilung um die 95%-Fläche zu berechnen.

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

# Exkurs: t-Verteilung und t-Wert (cont.)

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

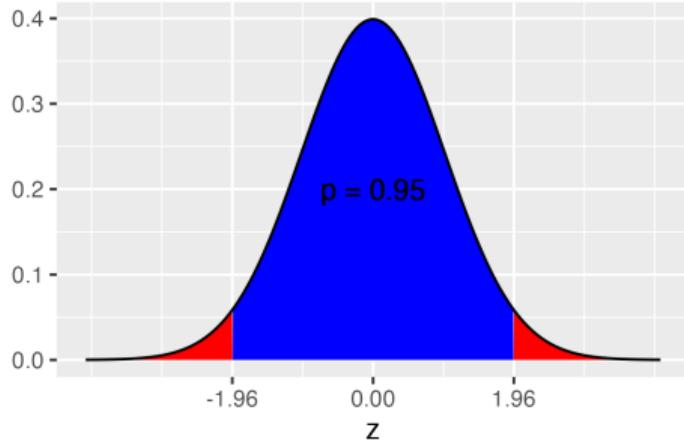
Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

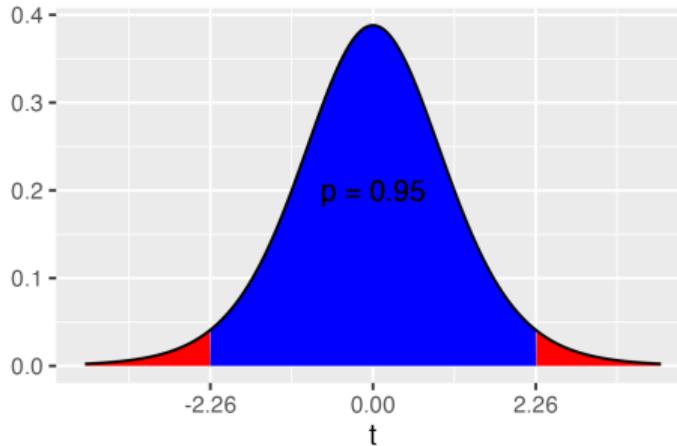
Zusammenfassung

Standard-Normalverteilung



$$CI = \bar{x} \pm 1.96 \cdot SE$$

t-Verteilung, n = 10



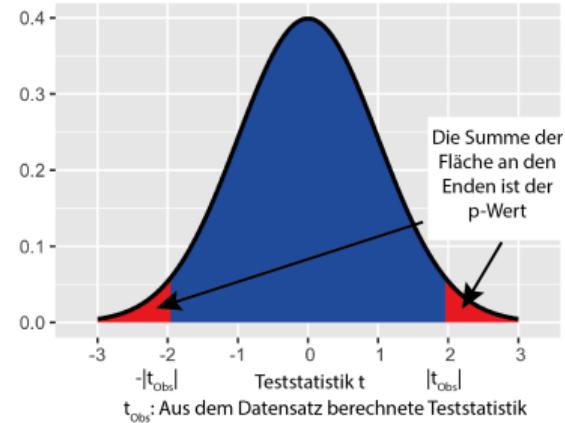
$$CI = \bar{x} \pm 2.26 \cdot SE$$

# p-Wert

- Man berechnet den sogenannten **p-Wert**: die Wahrscheinlichkeit, dass der beobachtete Wert der Teststatistik unter der Nullhypothese vorkommen kann.
- Ist diese Wahrscheinlichkeit genügend klein, so ist die Nullhypothese nicht plausibel.
- Als übliche Grenze (sogenanntes **Signifikanzniveau**) nimmt man meistens 5%.
- Man betrachtet dann also eine Nullhypothese gerade noch als plausibel, wenn sie in 5% der Fälle zufällig zu einer extremeren Teststatistik führen würde.

## Merke

Die Nullhypothese wird abgelehnt wenn der beobachtete t-Wert  $t_{\text{Obs}}$  den dem Signifikanzniveau zugehörigen kritischen t-Wert  $t_{\text{Crit}}$  überschreitet.



$$t_{\text{Obs}} = \frac{d}{SE}$$

# Kritischer t-Wert $t_{\text{Crit}}$ (A. Field, Miles und Z. Field 2012)

## A.2 Critical values of the $t$ -distribution

df	Two-Tailed Test		One-Tailed Test	
	0.05	0.01	0.05	0.01
1	12.71	63.66	6.31	31.82
2	4.30	9.92	2.92	6.96
3	3.18	5.84	2.35	4.54
4	2.78	4.60	2.13	3.75
5	2.57	4.03	2.02	3.36
6	2.45	3.71	1.94	3.14
7	2.36	3.50	1.89	3.00
8	2.31	3.36	1.86	2.90
9	2.26	3.25	1.83	2.82
10	2.23	3.17	1.81	2.76
26	2.06	2.78	1.71	2.48
27	2.05	2.77	1.70	2.47
28	2.05	2.76	1.70	2.47
29	2.05	2.76	1.70	2.46
30	2.04	2.75	1.70	2.46
35	2.03	2.72	1.69	2.44
40	2.02	2.70	1.68	2.42
45	2.01	2.69	1.68	2.41
50	2.01	2.68	1.68	2.40
60	2.00	2.66	1.67	2.39
70	1.99	2.65	1.67	2.38
80	1.99	2.64	1.66	2.37
90	1.99	2.63	1.66	2.37
100	1.98	2.63	1.66	2.36
$\infty$ (z)	1.96	2.58	1.64	2.33

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

## p-Wert kleiner als Signifikanzniveau

- ▶ Wenn der p-Wert kleiner ist als das Signifikanzniveau (sogenannt signifikanter p-Wert), also  $t_{\text{Obs}} > t_{\text{Crit}}$ , entscheidet man sich also zugunsten der Alternativhypothese.
- ▶ Es gibt in diesem Fall signifikante Abweichungen von der Nullhypothese, so dass man davon ausgehen kann, dass die Alternativhypothese gilt.

## p-Wert nicht kleiner als Signifikanzniveau

- ▶ Die Alternativhypothese kann nicht „bewiesen“ werden.
- ▶ Dies berechtigt aber nicht zur Behauptung, dass sie falsch ist!

# Entscheid für Beispiel

- ▶ Aus unseren Daten erhalten wir einen p-Wert von 0.000869.
- ▶ Dieser p-Wert ist natürlich kleiner als das Signifikanzniveau von 0.05.
- ▶ Man entscheidet sich zugunsten der Alternativhypothese.
- ▶ Die weiblichen Studierenden der PHY an der BFH G unterscheiden sich im Merkmal Hüfte INNEN Rotation passiv in Flexion signifikant von den männlichen Studierenden.
- ▶ Die weiblichen Studierenden sind also beweglicher.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung



Bern University  
of Applied Sciences

# Wann benutze ich welchen Test?

## Quantitative Daten

eine Gruppe	grosser Umfang ( $>100$ )		Student's t (one sample)
	kleiner Umfang	normalverteilt	Student's t (one sample)
		nicht normalverteilt	Wilcoxon Rank (WVR-Test)
zwei Gruppen	grosse Umfänge (in jeder Gruppe $>100$ )		Student's t (paired)
		unabhängig	Varianzen gleich: Student's t (unpaired)
			Varianzen ungleich: Welch's t (unpaired)
	kleine Umfänge	normalverteilt	Student's t (paired)
			Varianzen gleich: Student's t (unpaired)
			Varianzen ungleich: Welch's t (unpaired)
		nicht normalverteilt	Wilcoxon-Rank (WVR-Test)
mehrere Gruppen	grosse Umfänge (in jeder Gruppe $>100$ )		unabhängig
		abhangig	ANOVA (repeated-measures)
		unabhängig	ANOVA
	kleine Umfänge	normalverteilt	abhangig
			ANOVA (repeated-measures)
			ANOVA
	nicht normalverteilt		abhangig
			Friedmann-Test
			unabhängig
			Kruskal-Wallis-Test

WVR-Test: Wilcoxon-Vorzeichen-Rangsummentest

WR-Test: Wilcoxon-Rangsummentest, Mann-Whitney-U-Test

ANOVA: Varianzanalyse

## Kategoriale Daten

kleine Häufigkeiten (<5)			Fisher's Exact Test, Binomial Test
große Häufigkeiten	eine Variable		Chi^2-Test
	zwei Variablen	abhängig	McNemar-Test
		unabhängig	Chi^2-Test
	>2 Variablen		Log-Linear Analyse

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

# Parametrische Tests

Bei den sogenannten parametrischen Tests (t-Test, ANOVA, usw.) müssen folgende Annahmen beachtet werden:

► Unabhängige Daten

- ▶ Bei kleinen Stichprobenumfängen (für uns  $n < 100$ ) müssen die **Daten beider Gruppen normalverteilt** sein.
- ▶ Die **Streuung** in den verschiedenen **Gruppen** etwa gleich gross sein (für einige Tests gibt es aber Korrekturen dafür).

► Abhängige Daten

- ▶ Bei kleinen Stichprobenumfängen (für uns  $n < 100$ ) müssen die **Differenzen zwischen den Gruppen** normalverteilt sein.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

# Nichtparametrische Test

Ist die **Verteilung der Daten nicht bekannt**, müssen andere Testverfahren verwendet werden, nämlich die sogenannten **nichtparametrischen** Methoden.

- ▶ Die meisten nichtparametrischen Verfahren sind Rangtests.  
(Wilcoxon-Vorzeichen-Rangsummen-Test, Wilcoxon-Rangsummen-Test, Friedmann-Test, usw.)
- ▶ Die Teststatistik ist nur eine Funktion der rangierten Beobachtungswerte.
- ▶ Diese Verfahren sind somit robuster und sind auch auf ordinal skalierte Daten anwendbar.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

## Beispiel

- ▶ In unserem Beispiel hatten wir zwei unabhängige Gruppen mit kleinen Umfängen.
- ▶ Da wir keine Möglichkeit haben, die Normalverteilung der Daten der männlichen Studierenden zu überprüfen, müssen wir davon ausgehen, dass keine Normalverteilung vorliegt.
- ▶ In diesem Fall ist ein Wilcoxon-Rangsummen-Test, auch Mann-Whitney-U-Test genannt, der richtige Test.

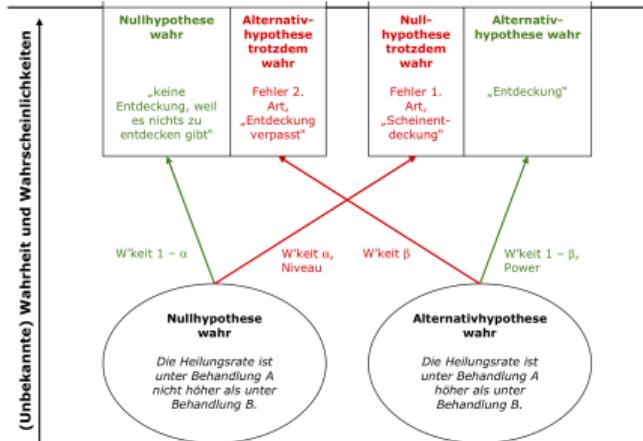
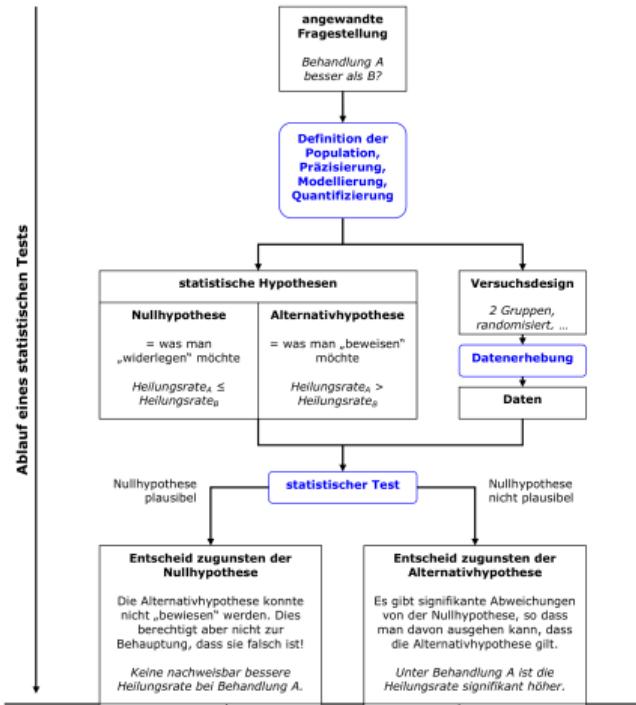
# Wahrheit und Wahrscheinlichkeit

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

# Statistischer Test

# Wahrheit und Wahrscheinlichkeit

## Statistischer Test: Ablaufschema, Fehlerwahrscheinlichkeiten



Quelle: Michael Vock, IMSV, Universität Bern

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Binäre Klassifikation - Entscheid und Irrtum

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

Decision	True state	
	Not different - H0 True	Different - H1 True
Not different	Correct decision Frequency $1 - \alpha$ True negative <b>Specificity</b>	Incorrect decision Frequency $\beta$ False negative <b>Type 2 error</b>
Different	Incorrect decision Frequency $\alpha$ False positive <b>Type 1 error</b>	Correct decision Frequency $1 - \beta$ True positive <b>Sensitivity (Power)</b>

# Fehler 1. Art

Decision	True state	
	Not different - H0 True	Different - H1 True
Not different	Correct decision Frequency $1 - \alpha$ True negative <b>Specificity</b>	Incorrect decision Frequency $\beta$ False negative <b>Type 2 error</b>
Different	Incorrect decision Frequency $\alpha$ False positive <b>Type 1 error</b>	Correct decision Frequency $1 - \beta$ True positive <b>Sensitivity (Power)</b>

- Falls die Nullhypothese wahr ist und der p-Wert kleiner als das Signifikanzniveau ist, wird fälschlicherweise die Alternativhypothese angenommen.
- Diese Fehlentscheidung nenn man **Fehler 1. Art**.
- Die Wahrscheinlichkeit eines Fehlers 1. Art nennt man  **$\alpha$ -Fehler**.
- Dieser Fehler kann kontrolliert werden, er entspricht dem gewählten Signifikanzniveau

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit

Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße

Zusammenfassung



# Fehler 2. Art

Decision	True state	
	Not different - H0 True	Different - H1 True
Not different	Correct desicion Frequency $1 - \alpha$ True negative Specificity	Incorrect desicion Frequency $\beta$ False negative <b>Type 2 error</b>
Different	Incorrect desicion Frequency $\alpha$ False positive <b>Type 1 error</b>	Correct desicion Frequency $1 - \beta$ True positive <b>Sensitivity (Power)</b>

- ▶ Falls die Alternativhypothese wahr ist und der p-Wert grösser als das Signifikanzniveau ist, wird fälschlicherweise die Nullhypothese beibehalten.
- ▶ Diese Fehlentscheidung nenn man **Fehler 2. Art**.
- ▶ Die Wahrscheinlichkeit eines Fehlers 2. Art nennt man  **$\beta$ -Fehler**.
- ▶ Dieser Fehler kann mit der passenden Wahl des Stichprobenumfangs kontrolliert werden.

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengrösse  
Zusammenfassung



# Power: Stichprobenumfang, Streuung und Differenzen

Die **Power**, **Macht** oder **Güte** eines Tests, ist die Wahrscheinlichkeit, bei wahrer Alternativhypothese die Nullhypothese zu verwerfen, und ist durch

$$\gamma = 1 - \beta$$

gegeben.

## Merke

- Die Power wird grösser wenn...
  - Die Stichproben grösser sind.
  - Die Streuungen der Stichproben kleiner sind.
  - Die Differenz zwischen den Stichproben grösser ist.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

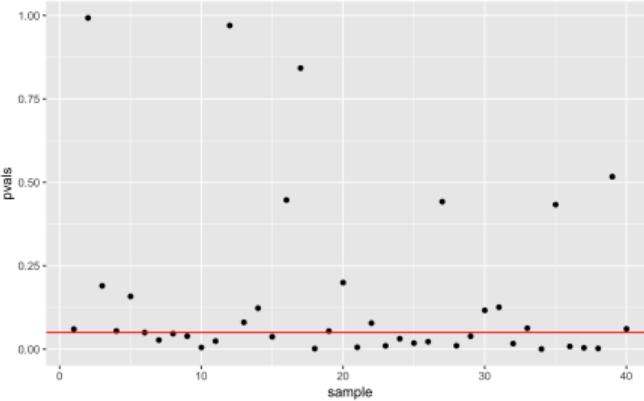
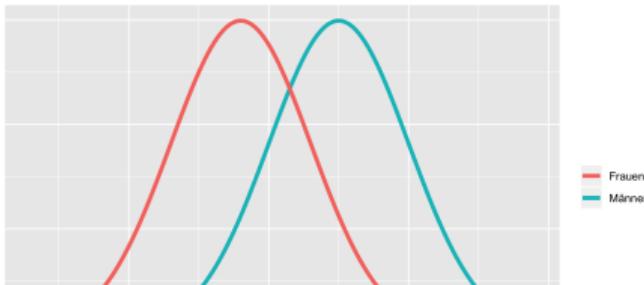
Zusammenfassung



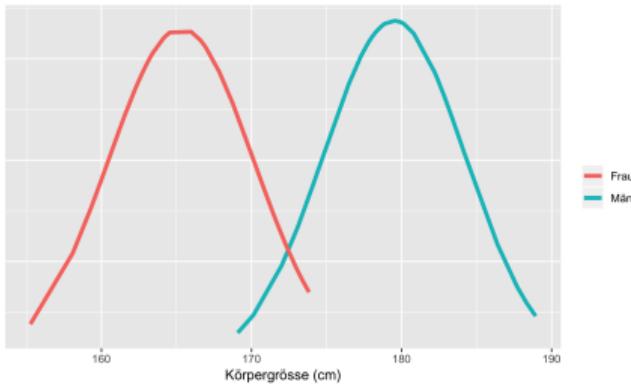
Bern University  
of Applied Sciences

# Beispiel Fehler 2. Art $n = 5$

$n = 1000, \sigma = 10$



40 Mittelwerte von Stichproben mit Umfang  $n = 5$



- ▶ Theoretische Power: 0.495
- ▶ True (+): 19 / False (-): 21
- ▶  $(1 - \beta) = \frac{19}{40} = 0.475$

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

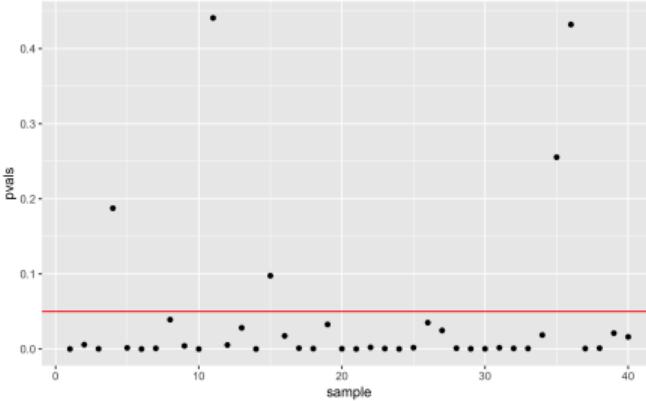
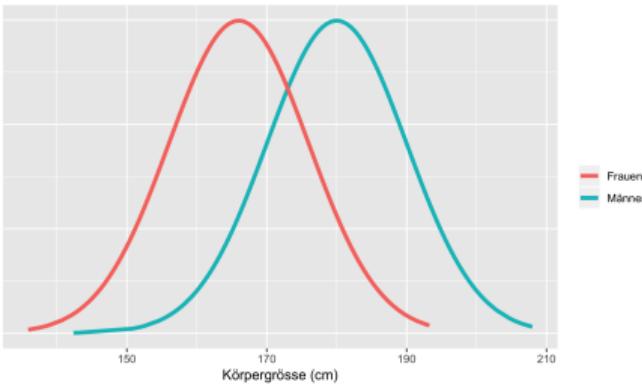
Zusammenfassung



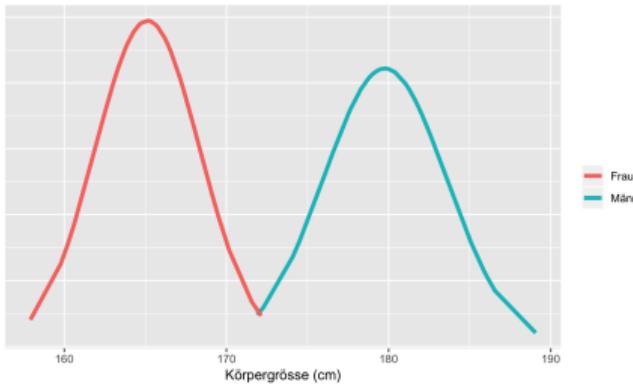
Bern University  
of Applied Sciences

# Beispiel Fehler 2. Art $n = 10$

$n = 1000, \sigma = 10$



40 Mittelwerte von Stichproben mit Umfang  $n = 10$



- ▶ Theoretische Power: 0.841
- ▶ True (+): 35 / False (-): 5
- ▶  $(1 - \beta) = \frac{35}{40} = 0.875$

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

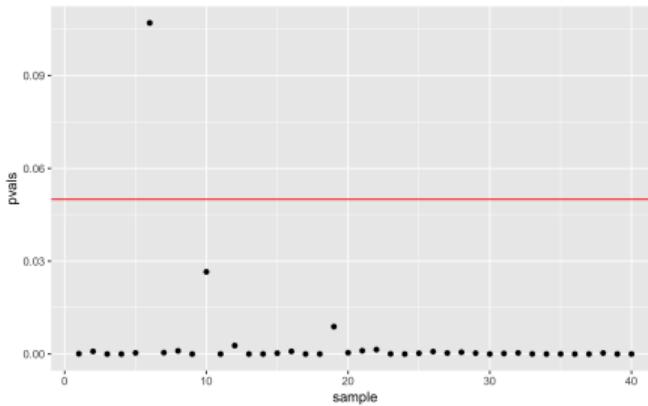
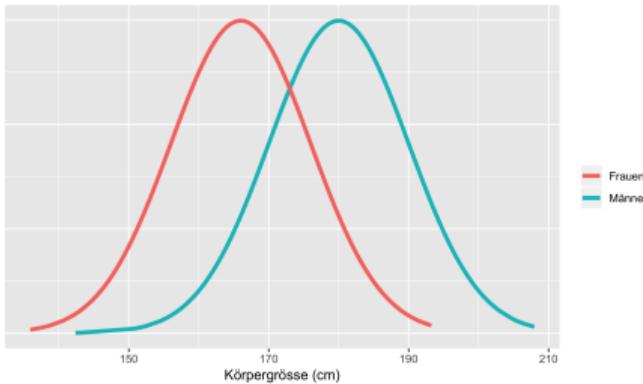
Zusammenfassung



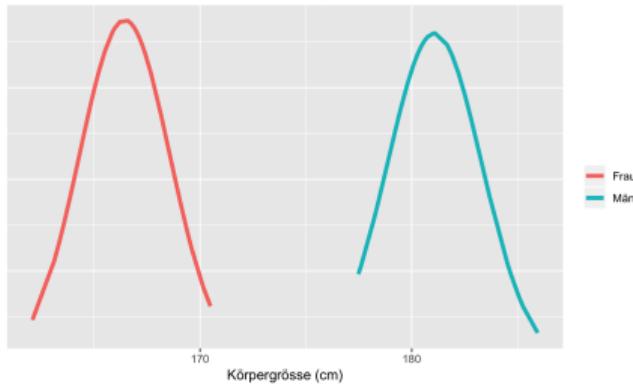
Bern University  
of Applied Sciences

# Beispiel Fehler 2. Art $n = 20$

$n = 1000, \sigma = 10$



40 Mittelwerte von Stichproben mit Umfang  $n = 20$



- ▶ Theoretische Power: 0.991
- ▶ True (+): 39 / False (-): 1
- ▶  $(1 - \beta) = \frac{39}{40} = 0.975$

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

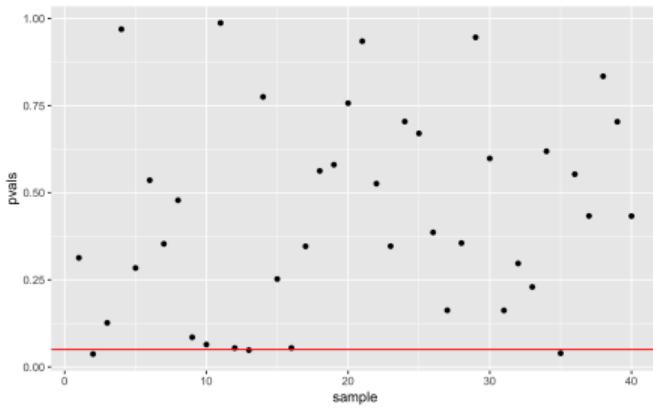
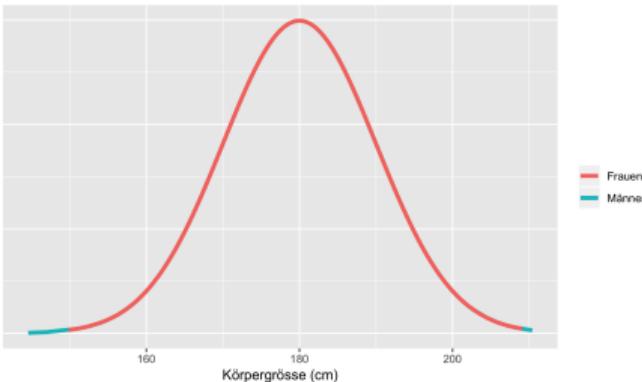
Zusammenfassung



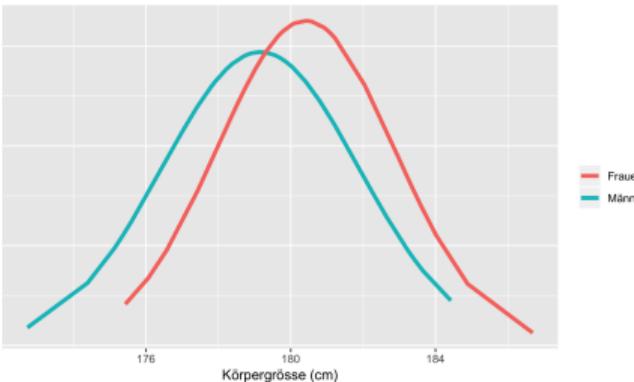
Bern University  
of Applied Sciences

# Beispiel Fehler 1. Art

$n = 1000, \sigma = 10$



40 Mittelwerte von Stichproben mit Umfang  $n = 20$



- ▶ Signifikanzniveau  $\alpha = 0.05$
- ▶ True (-): 38 / False (+): 2
- ▶  $\alpha = \frac{2}{40} = 0.05$

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

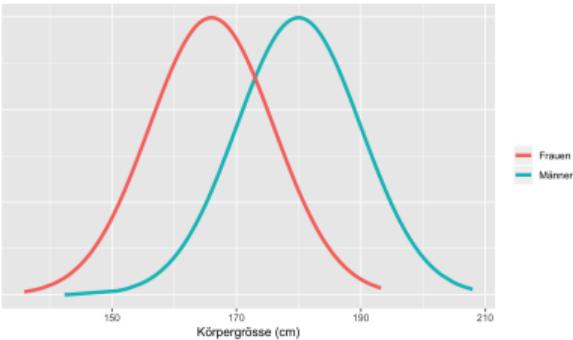
Zusammenfassung



Bern University  
of Applied Sciences

# Bestimmen des Stichprobenumfangs

$n = 1000, \sigma = 10$



## Effektstärke (Cohen's d)

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

- ▶ Festlegen von...
  - ▶ Signifikanzniveau
  - ▶ Power. Üblich sind 0.8 oder 0.9.
  - ▶ Effektstärke. Oft 0.5.
- ▶ Berechnen der Stichprobengrösse um den Effekt mit der gewünschten Power bei gegebenem Signifikanzniveau nachzuweisen.

## Effektstärke für Beispiel

- ▶ Klein: 0.2 bis 0.3
- ▶ Mittel: 0.5 bis 0.8
- ▶ Stark: > 0.8

$$d = \frac{180 - 166}{10} = 1.4$$

# Statistischer Test vs. Vertrauensintervall

- ▶ Zu vielen statistischen Tests gibt es ein entsprechendes Vertrauensintervall.
- ▶ Mit einem statistischen Test kommt man immer zur gleichen Schlussfolgerung wie mit dem entsprechenden Vertrauensintervall.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Praxis mit R

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

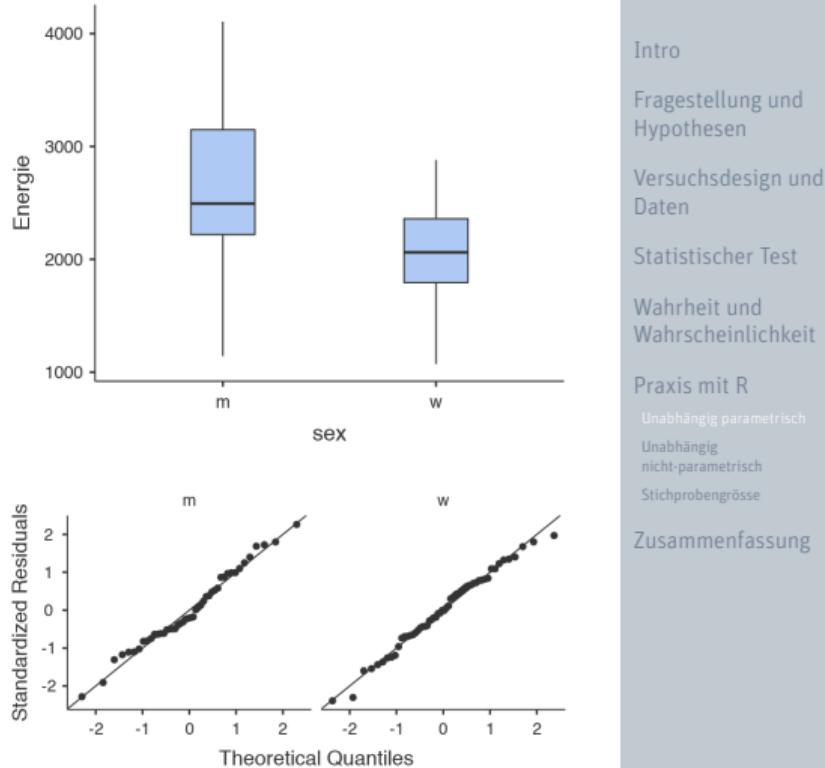
# Beispiel 1: Vergleich unabhängig parametrisch

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

- ▶ Daten: `ERB_Daten.csv`
- ▶ Wir möchten die Kalorienaufnahme von weiblichen und männlichen Jugendlichen vergleichen.
- ▶ Wir haben die Messungen von  $n = 56$  weiblichen und  $n = 46$  männlichen Jugendlichen und nehmen wiederum an, die Stichprobe sei zufällig gezogen worden.
- ▶ Wir müssen die Annahme der Normalverteilung in beiden Gruppen überprüfen.
- ▶ Zudem müssen wir beurteilen ob von gleichen Varianzen ausgegangen werden kann.

# Beispiel 1: Vergleich unabhängig parametrisch (cont.)

- ▶ In der Mitte des QQ-Plots der männlichen Jugendlichen ist eine Kurve zu sehen, die Daten liegen nicht schön auf einer Gerade.
- ▶ Auch im Boxplot der männlichen Jugendlichen sieht man, dass die Daten nicht schön symmetrisch sind.
- ▶ Die Daten sind also bei den männlichen Jugendlichen nicht normalverteilt, aber der Übung halber machen wir jetzt doch weiter, als wäre es so.



Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

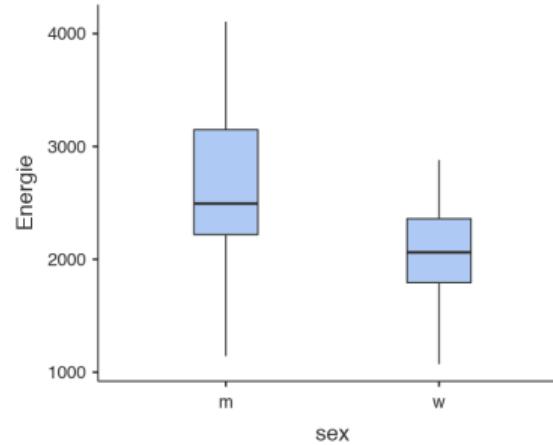
Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Beispiel 1: Vergleich unabhängig parametrisch (cont.)

- ▶ Bei den männlichen Jugendlichen beträgt die Standardabweichung 651 kcal, bei den weiblichen Jugendlichen ist sie mit 415 kcal deutlich tiefer.
- ▶ Auch im Boxplot sehen wir deutlich, dass bei den männlichen Jugendlichen eine höhere Streuung vorliegt, als bei den weiblichen Jugendlichen. Also ist die Streuung nicht in beiden Gruppen gleich.



Descriptives

	sex	Energie
N	m	46
	w	56
Missing	m	0
	w	0
Mean	m	2630
	w	2064
Standard deviation	m	651
	w	415

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

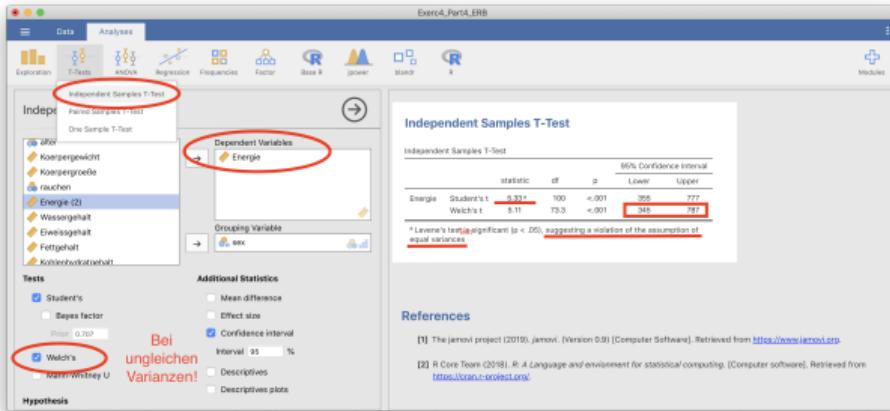
Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Beispiel 1: Vergleich unabhängig parametrisch (cont.)



- Wegen ungleichen Varianzen den Test Welch's t zur Bestimmung des Vertrauensintervalls verwenden.
- Wir erhalten ein 95%-Vertrauensintervall von [345,787] kcal.
- 0 ist nicht im Vertrauensintervall. Die männlichen Jugendlichen scheinen also eine höhere Kalorienaufnahme zu haben als die weiblichen Studierenden.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

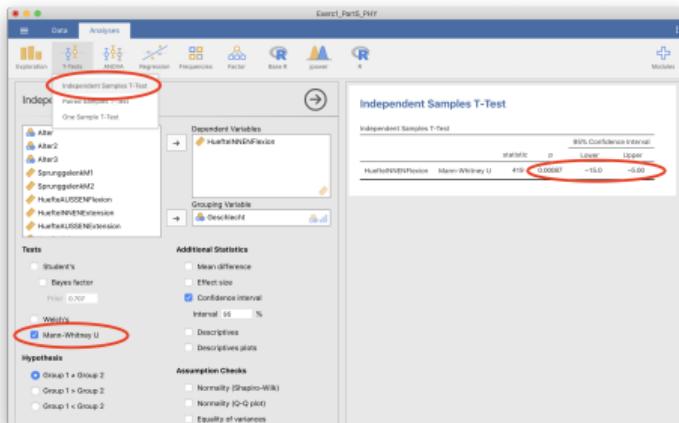
Zusammenfassung

# Beispiel 2: Vergleich unabhängig nicht-parametrisch

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

- ▶ Daten: `HipData.csv`
- ▶ Nullhypothese Das Merkmal Hüfte INNEN Rotation passiv in Flexion bei weiblichen und männlichen Studierenden der PHY an der BFH G ist gleich gross.
- ▶ Alternativhypothese Das Merkmal Hüfte INNEN Rotation passiv in Flexion bei weiblichen und männlichen Studierenden der PHY an der BFH G ist nicht gleich gross.
- ▶ Anzuwendender Test: Wilcoxon-Rangsummen-Test (Mann-Whitney U in Jamovi)

# Beispiel 2: Vergleich unabhängig nicht-parametrisch (cont.)



- Für den Vergleich des Merkmals Hüfte INNEN Rotation passiv in Flexion erhalten wir einen p-Wert von 0.00087 und wir entscheiden zugunsten der Alternativhypothese.
- Die weiblichen Studierenden haben also eine grössere passive Hüft-Innenrotation in Flexion als die männlichen Studierenden.

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengrösse

Zusammenfassung



Bern University  
of Applied Sciences

# Beispiel 2: Vergleich unabhängig nicht-parametrisch (cont.)

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

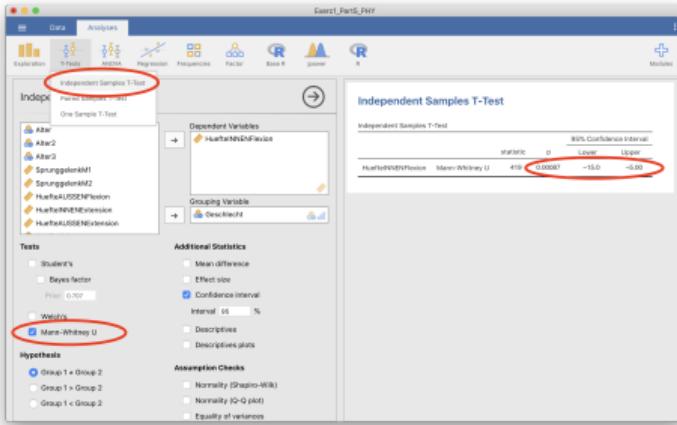
Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



- ▶ Mit dem Vertrauensintervall  $[-15, -5]$  kommt man zum gleichen Schluss.
- ▶ Für die Interpretation des Vorzeichens die deskriptive Statistik beachten!

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Beispiel 2 - Stichprobengrösse

The screenshot shows the RStudio interface with the 'jpower' module selected. The left panel displays the 'Independent Samples T-Test' configuration, and the right panel shows the results of the power analysis.

**Left Panel (Configuration):**

- Calculate: N per group
- Minimally-interesting effect size ( $\delta$ ): 1.4 (highlighted with a red circle)
- Minimum desired power: 0.9
- N for group 1: 20
- Relative size of group 2 to group 1: 1
- $\alpha$  (type I error rate): 0.05
- Tails: two-tailed
- Plots:**
  - Power contour plot
  - Power curve by effect size
  - Power curve by N
  - Power demonstration
- Additional Options:**
  - Explanatory text

**Right Panel (Results):**

### Independent Samples T-Test

The purpose of a power analysis is to evaluate the sensitivity of a design and test. You have chosen to calculate the minimum sample size needed to have an experiment sensitive enough to consistently detect the specified hypothetical effect size.

#### A Priori Power Analysis

		User Defined		
N <sub>1</sub>	N <sub>2</sub>	Effect Size	Power	$\alpha$
12	12	1.40	0.900	0.0500

We would need a sample size of 12 in each group to reliably (with probability greater than 0.9) detect an effect size of  $\delta \geq 1.4$ , assuming a two-sided criterion for detection that allows for a maximum Type I error rate of  $\alpha = 0.05$ .

To evaluate the design specified in the table, we can consider how sensitive it is to true effects of increasing sizes; that is, are we likely to correctly conclude that  $\delta > 0$  when the effect size is large enough to care about?

#### Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.837$	$\leq 50\%$	Likely miss
$0.837 < \delta \leq 1.197$	$50\% - 80\%$	Good chance of missing
$1.197 < \delta \leq 1.541$	$80\% - 95\%$	Probably detect

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

# Zusammenfassung

Intro

Fragestellung und  
Hypothesen

Versuchsdesign und  
Daten

Statistischer Test

Wahrheit und  
Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig  
nicht-parametrisch

Stichprobengrösse

Zusammenfassung

# Zusammenfassung

- ▶ Wir haben gesehen wie ein statistischer Test abläuft.
- ▶ Dabei werden aus einer Fragestellung statistische Hypothesen formuliert.
- ▶ Um die Hypothesen zu testen müssen Daten erhoben werden.
- ▶ Je nach Versuchsdesign und Verteilung der Daten ist ein passender Test auszuwählen.
- ▶ Die Entscheidung anhand eines Tests heisst nicht, dass es in der Wirklichkeit so ist.
- ▶ Statistische Signifikanz heisst nicht klinische Relevanz! Differenzen und Effektgrößen beachten!

## Beispiel

Gerichtsprozess vs. klinische Studie – Übernommen von Terry Mills **Legal vs clinical trials: an explanation of sampling errors and sample size**

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung



Bern University  
of Applied Sciences

# Ausgangspunkt

## Gerichtsprozess

- Unschuldsvermutung

## Klinische Studie

- Nullhypothese ( $H_0$ )
- Neue Behandlung ist nicht besser als Standardbehandlung

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Ausgangspunkt

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

## Gerichtsprozess

- Alternative: Angeklagter ist schuldig

## Klinische Studie

- Alternativhypothese ( $H_1$ )
- Neue Behandlung ist besser als Standardbehandlung

# Prozess

## Gerichtsprozess

### Gerichtsanhörung:

- ▶ Beweise werden gesammelt und präsentiert

## Klinische Studie

### Durchführung der Studie:

- ▶ Daten werden gesammelt und analysiert

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Entscheidung

## Gerichtsprozess

### Urteil:

- ▶ „schuldig“
- ▶ oder...
- ▶ „nicht schuldig“ (Gericht entscheidet nicht ob die Person unschuldig ist!)

## Klinische Studie

### Testergebnis:

- ▶ „Datenlage zeigt Überlegenheit an“ ( $H_0$  verwerfen)
- ▶ oder...
- ▶ „Datenlage zeigt Überlegenheit nicht an“ ( $H_0$  nicht verwerfen; nicht:  $H_0$  annehmen!)

Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

# Mögliche Fehler

Gerichtsprozess  
Urteil:

- ▶ Unschuldige Person schuldig gesprochen
- ▶ oder...
- ▶ Schuldige Person nicht schuldig gesprochen

Klinische Studie  
Testergebnis:

- ▶ Daten zeigen an, dass die neue Behandlung überlegen ist, obwohl es in Wirklichkeit nicht so ist. Wahre  $H_0$  verwerfen (Fehler 1. Art).
- ▶ oder...
- ▶ Daten zeigen nicht an, dass die neue Behandlung überlegen ist, obwohl es in Wirklichkeit so ist. Falsche  $H_0$  nicht verwerfen (Fehler 2. Art).

Intro  
Fragestellung und Hypothesen  
Versuchsdesign und Daten  
Statistischer Test  
Wahrheit und Wahrscheinlichkeit  
Praxis mit R  
Unabhängig parametrisch  
Unabhängig nicht-parametrisch  
Stichprobengröße  
Zusammenfassung

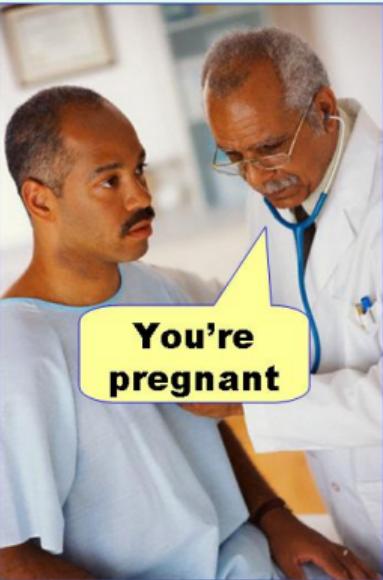


Bern University  
of Applied Sciences

		Realität	
		$H_0$ wahr	$H_1$ wahr
$H_0$ wahr	$H_0$ wahr	Keine Entdeckung $1-\alpha$	Entdeckung verpasst Fehler 2. Art $\beta$
	$H_1$ wahr	Falsche Entdeckung Fehler 1. Art $\alpha$	Entdeckung Power $1-\beta$
$H_1$ wahr	$H_0$ wahr		
	$H_1$ wahr		

- Intro
- Fragestellung und Hypothesen
- Versuchsdesign und Daten
- Statistischer Test
- Wahrheit und Wahrscheinlichkeit
- Praxis mit R
  - Unabhängig parametrisch
  - Unabhängig nicht-parametrisch
  - Stichprobengröße
- Zusammenfassung

## Type I error (false positive)



## Type II error (false negative)



Intro

Fragestellung und Hypothesen

Versuchsdesign und Daten

Statistischer Test

Wahrheit und Wahrscheinlichkeit

Praxis mit R

Unabhängig parametrisch

Unabhängig nicht-parametrisch

Stichprobengröße

Zusammenfassung

Quelle: effectsizefaq.com