

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

QSB 1: Epidemiologie, Medizinische Biometrie und Medizinische Informatik

Übung 6
Statistisches Testen I und II (Mittelwertvergleiche)

Wintersemester 2026/27

imise.

Lernziele

Die Studierenden sollen

- statistische Tests im Kontext klinischer und epidemiologischer Studien sehen,
- t-Tests und nichtparametrische Tests zum Lagevergleich von Verteilungen passend zu den Daten auswählen,
- t-Tests auf praktische Fälle anwenden und
- Testergebnisse im Kontext der geschätzten Effekte interpretieren können.

Auswertung klinischer Studien

1. Auswertungspopulation festlegen
   - Welche Patienten werden in welcher Analyse berücksichtigt?
2. Analyse der Probandencharakteristika
   - mit Methoden der deskriptiven Statistik (Übung 3)
3. Analyse der primären und sekundären Endpunkte
   - zur Entscheidung bzgl. der Studienfragestellung: Punkt- und Intervallschätzung von Effekten sowie statistische Testverfahren
4. Beschreibung von Sicherheitsaspekten
   - mit Methoden der deskriptiven Statistik und ggf. statistischen Tests
5. Weitere explorative Analysen
   - mit Methoden der deskriptiven Statistik und statistischen Modellen

Herzsportstudie – primärer Endpunkt

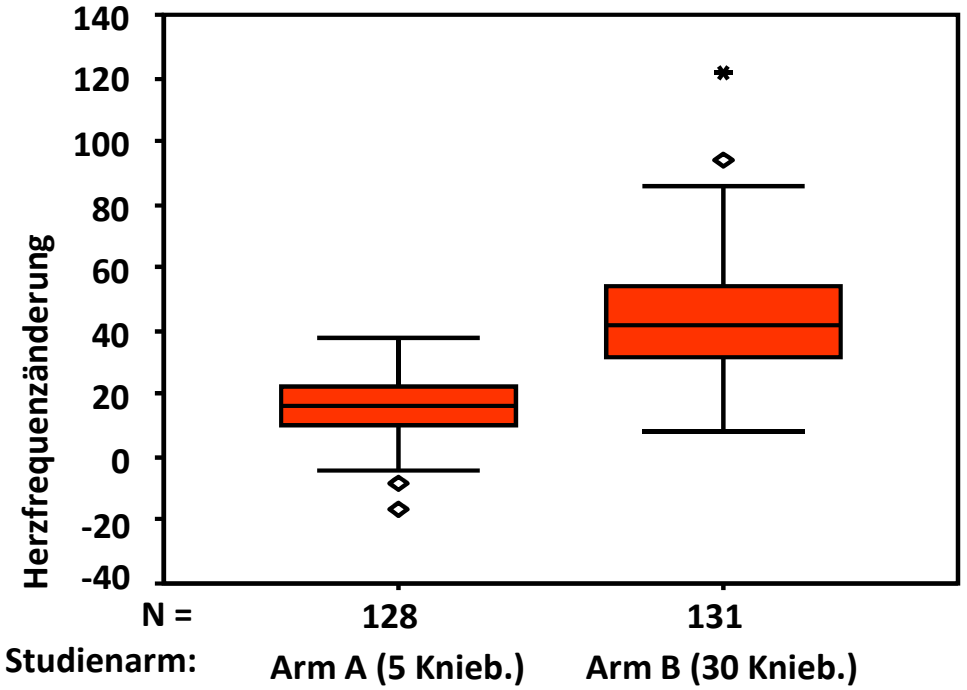
Wissenschaftliche Fragestellung (vgl. Übung 1), primäres Studienziel:

Unterscheidet sich die mittlere Änderung der Herzfrequenz nach 30 Kniebeugen von der mittleren Änderung der Herzfrequenz nach 5 Kniebeugen?

$$HF \quad = HF \quad - HF$$

(HF... Herzfrequenz)

Grafische Darstellung



Ablaufschema bei Hypothesentests

vgl. Skriptenheft Kapitel 5.6, S. 59

1. Wissenschaftliche Fragestellung → statistisches Modell
2. Ziehen der Stichprobe und Erfassung der Daten
3. Formulierung der Hypothesen H_0 / H_A
4. Wahl des Signifikanzniveaus α
5. Wahl des Tests / der Prüfgröße
6. Prüfung der Testvoraussetzungen
7. Ermittlung des Akzeptanzbereiches
8. Berechnung der Prüfgröße
9. Testentscheidung
10. Interpretation der Ergebnisse im Kontext der Fragestellung (inkl. Angabe eines Effektschätzers)

Anwendung auf die Herzsportstudie

Mathematische Einordnung

- Fragestellung: Es sollen zwei Mittelwerte verglichen werden
- Datentyp: kontinuierliches Merkmal
- 2 unabhängige Stichproben

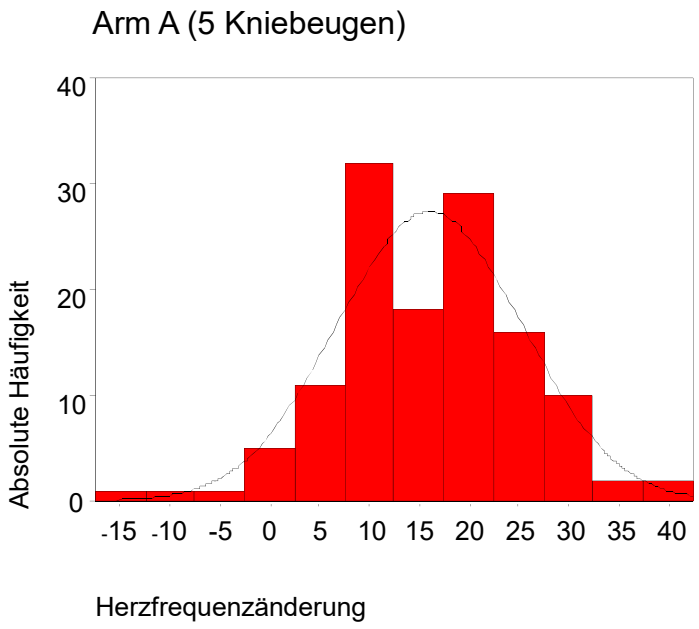
- Fallzahl ist „groß“ (N = 128 + 131)
- Ausreißer sind eher nicht zu erwarten

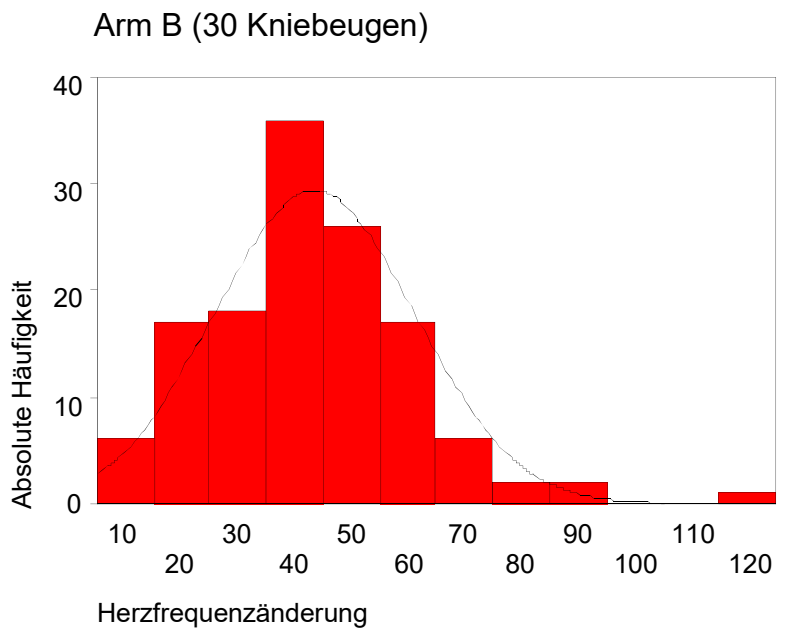
→ t-Test für unabhängige Stichproben ("2-Stichproben-t-Test")

Testvoraussetzungen

Prüfung der Testvoraussetzungen

- Verteilung anhand von Histogrammen überprüfen
- Fallzahl N = 128 + 131 ist "groß"





Statistische Hypothesen

Mit der Formulierung der Hypothesen H_0 und H_A entscheiden wir uns für ein statistisches Modell:

Nullhypothese H_0: $\mu = \mu$

- Die mittlere Herzfrequenzänderung μ_A nach Durchführung von 5 Kniebeugen (Arm A) ist gleich der mittleren Herzfrequenzänderung μ_B nach 30 Kniebeugen (Arm B).

Alternativhypothese H_A: $\mu_{\square} \neq \mu_{\square}$

- Die mittleren Herzfrequenzänderungen μ_A und μ_B unterscheiden sich.

Festlegung des Signifikanzniveaus

Das Signifikanzniveau α wird im Studienprotokoll **vor** Beginn der Auswertungen festgelegt.

- α ist die zugelassene Wahrscheinlichkeit, die Nullhypothese abzulehnen, obwohl sie richtig ist (Fehler I. Art).
- Der Test ist so konstruiert, das α (theoretisch) eingehalten wird.
- typische Konvention: α = 0,05; besser α = 0,005 (JASA 2019)

Für Übungszwecke sei in dieser Aufgabe α = 0,001.

Prüfgröße (Teststatistik)

Die Prüfgröße misst den **beobachteten Effektunterschied** zwischen den Gruppen und setzt ihn ins Verhältnis zur Streuung.

Eigenschaften:

- Sie ist eine Zufallsgröße,
- die aus beobachteten Daten berechnet und
- deren Wahrscheinlichkeitsverteilung unter der Nullhypothese (H_0) bekannt ist.

Prinzip der Falsifizierung (Vorgehen ähnlich wie beim indirekten Beweis):

- Man postuliert "Null-Effekt" (Nullhypothese).
- Übersteigt die Prüfgröße einen kritischen Wert, verwirft man die Nullhypothese, weil es ziemlich unwahrscheinlich ist, diesen beobachteten Wert (oder einen extremeren) zu erhalten.

Prüfgröße (Teststatistik)

Prüfgröße des t-Tests für 2 unabhängige Gruppen

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$t \sim$ t-Verteilung mit f Freiheitsgraden (FG)

(Berechnung der Freiheitsgrade siehe Folie 14)

	Arm A	Arm B
Mittelwert HF_{diff}	$\bar{x}_A$	$\bar{x}_B$
empir. Standardabw. HF_{diff}	s_A	s_B
Fallzahl	n_A	n_B

Prüfgröße (Herzsportstudie)

Aufgabe 1

Berechnung der Prüfgröße für den primären Endpunkt der Herzsportstudie

Benötigte Informationen (aus den Daten)

		Studienarm	
		Arm A (5 Knieb.)	Arm B (30 Knieb.)
Herzfrequenzänderung	Anzahl	128	131
	Mittelwert	16,02	43,40
	Standardabweichung	9,32	17,74

Berechnen Sie die Prüfgröße des t-Tests für 2 unabhängige Gruppen!

Berechnung der Freiheitsgrade

Freiheitsgrade f (Formel rein informativ)
nach der **Formel von Welch**

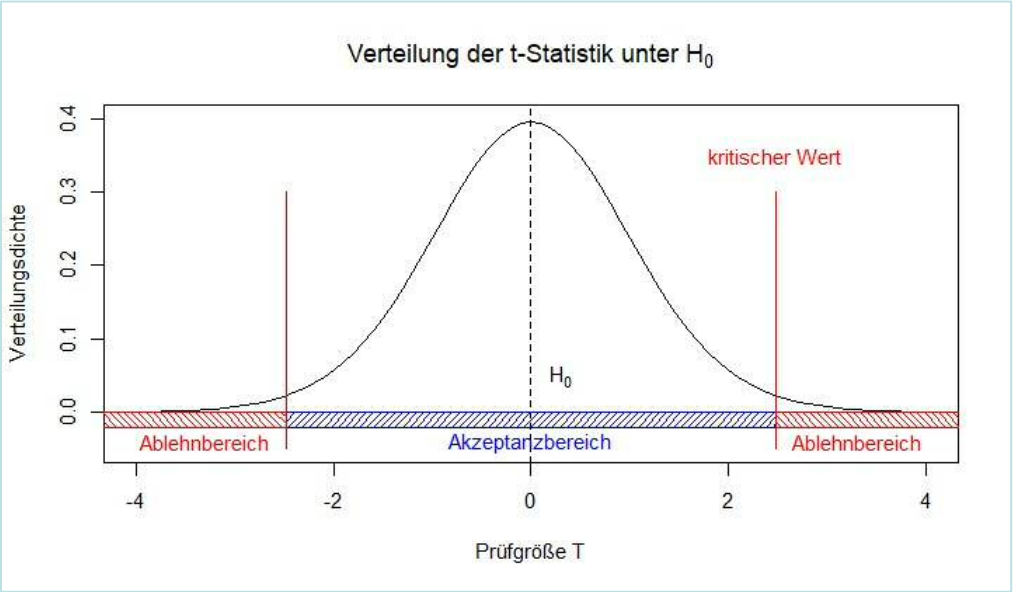
(setzt keine Gleichheit der Varianzen s_A und s_B voraus
und ist robust gegenüber Varianzinhomogenität)

$$f = \frac{\left(\frac{s}{n} + \frac{s}{n}\right)}{\frac{\left(\frac{s}{n}\right)}{n\ -1} + \frac{\left(\frac{s}{n}\right)}{n\ -1}} = \frac{\left(\frac{9{,}32}{128} + \frac{17{,}74}{131}\right)}{\frac{\left(\frac{9{,}32}{128}\right)}{128-1} + \frac{\left(\frac{17{,}74}{131}\right)}{131-1}} \approx 198$$

Verteilung der Prüfgröße

Dichte der Prüfgröße unter H_0:

Die kritischen Werte (Tafelwerte) werden so berechnet, dass die Prüfgröße unter H_0 mit Wahrscheinlichkeit von je $\alpha/2$ in den linken oder rechten Ablehnbereich fällt. Beim t-Test ergeben sich beim zweiseitigen Test die Grenzen durch das t -Quantil der t-Verteilung mit f Freiheitsgraden.



Quantile der t-Verteilung: FG ≥ 30

Hinweis für Übungen und Klausur:
Ist FG nicht explizit gelistet, dann die Angabe von dem FG nutzen, der am nächsten gelegen ist (kleiner oder größer).

Irrtumswahrscheinlichkeit α für den zweiseitigen Test									
FG \ α	0,50	0,20	0,10	0,05	0,02	0,01	0,002	0,001	0,0001
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385	3,646	4,482
32	0,682	1,309	1,694	2,037	2,449	2,738	3,365	3,622	4,441
34	0,682	1,307	1,691	2,032	2,441	2,728	3,348	3,601	4,405
35	0,682	1,306	1,690	2,030	2,438	2,724	3,340	3,591	4,389
36	0,681	1,306	1,688	2,028	2,434	2,719	3,333	3,582	4,374
38	0,681	1,304	1,686	2,024	2,429	2,712	3,319	3,566	4,346
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307	3,551	4,321
42	0,680	1,302	1,682	2,018	2,418	2,698	3,296	3,538	4,298
45	0,680	1,301	1,679	2,014	2,412	2,690	3,281	3,520	4,269
47	0,680	1,300	1,678	2,012	2,408	2,685	3,273	3,510	4,251
50	0,679	1,299	1,676	2,009	2,403	2,678	3,261	3,496	4,228
55	0,679	1,297	1,673	2,004	2,396	2,668	3,245	3,476	4,196
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232	3,460	4,169
70	0,678	1,294	1,667	1,994	2,381	2,648	3,211	3,435	4,127
80	0,678	1,292	1,664	1,990	2,374	2,639	3,195	3,416	4,096
90	0,677	1,291	1,662	1,987	2,368	2,632	3,183	3,402	4,072
100	0,677	1,290	1,660	1,984	2,364	2,626	3,174	3,390	4,053
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160	3,373	4,025
200	0,676	1,286	1,653	1,972	2,345	2,601	3,131	3,340	3,970
500	0,675	1,283	1,648	1,965	2,334	2,586	3,107	3,310	3,922
1000	0,675	1,282	1,646	1,962	2,330	2,581	3,098	3,300	3,906
∞	0,675	1,282	1,645	1,960	2,326	2,576	3,090	3,290	3,891

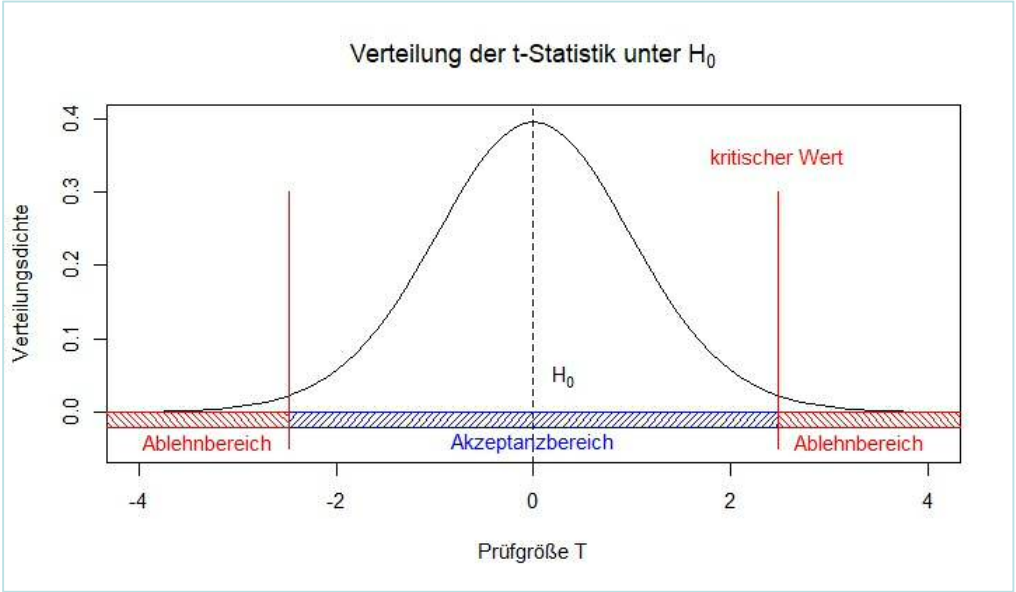
$$t \approx t = 3{,}340$$

= Quantile der standardisierten Normalverteilung

Ermittlung des Akzeptanzbereichs

H_0 wird beibehalten,
solange die Prüfgröße in den Akzeptanzbereich fällt.
Für den t-Test ergibt sich das Intervall
$(-t \quad ; t \quad)$.

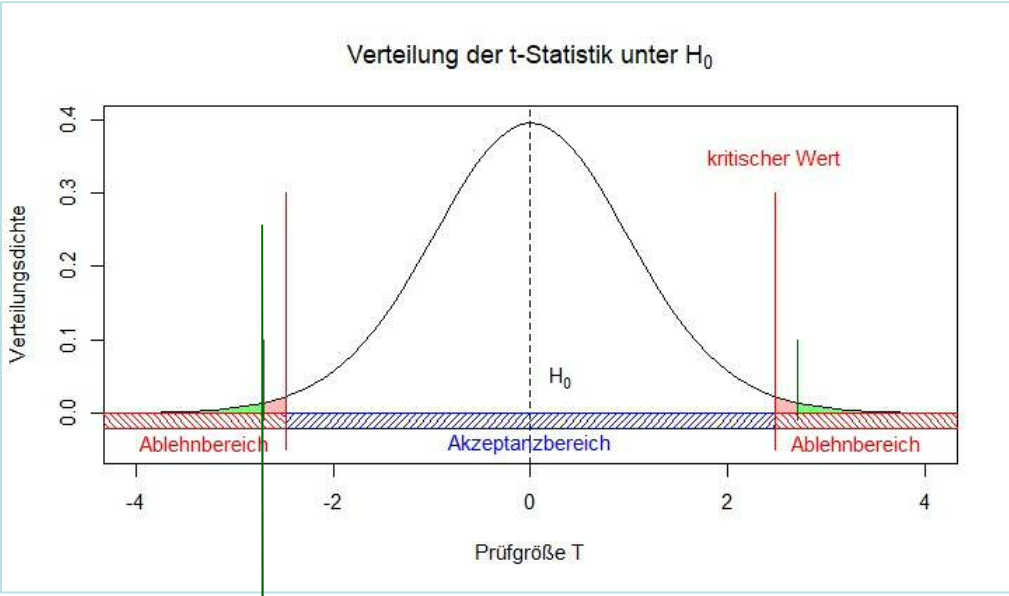
Der Ablehnbereich ist komplementär dazu.



Testentscheidung

Prüfung: $|t| \geq t$

- $|t| < t$:
  (t im Akzeptanzbereich)
  → H_0 beibehalten
- $|t| \geq t$:
  (t im Ablehnbereich,
  |t| übersteigt den kritischen Wert)
  → H_0 ablehnen



t im Ablehnbereich: H_0 ablehnen

Testentscheidung: Handrechnung

Herzsportstudie:

Für den t-Test, Skriptenheft S. 178:

$t_{198;1-0,001/2} \approx 3{,}340$; also wird H_0 abgelehnt,

wenn $t \leq -3{,}340$ oder $t \geq 3{,}340$

Testentscheidung mittels berechneter Prüfgröße

$$|t| = |-15{,}6| = 15{,}6 \geq 3{,}340 \approx t_{198;\,1-\,0{,}001/2}$$

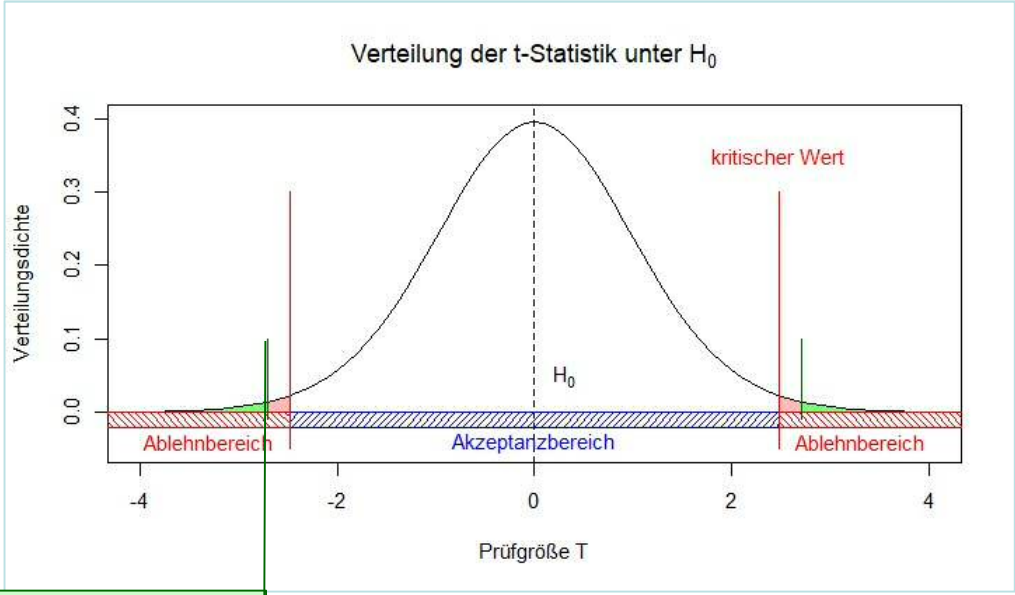
→ H_0 wird auf dem 0,1% Signifikanzniveau abgelehnt

Testentscheidung: Software

- Aus dem Wert der Prüfgröße t ergibt sich direkt die Fläche unter der Kurve: der **p-Wert**.
- Er ist die Wahrscheinlichkeit, den beobachteten Wert t der Prüfgröße zu erhalten oder einen extremeren, wenn H_0 wahr wäre.

Prüfung:

- $p \leq \alpha$:
  → H_0 wird abgelehnt
- $p > \alpha$:
  → H_0 wird beibehalten



p-Wert: Summe der grünen Flächen

Testentscheidung – Software

SPSS – Ausgabe:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Herzfrequenzänderung	Equal variances assumed	27,810	,000	-15,495	257	,000	-27,38132	1,76705	-30,86107	-23,90158
	Equal variances not assumed			-15,598	197,701	,000	-27,38132	1,75544	-30,84310	-23,91954

Vergleich des *p*-Wertes mit dem Signifikanzniveau α

- Anzeige: ,000 (auf 3 Nachkommastellen gerundet)

→ also $p < 0{,}0005$

- $p < 0{,}0005 < \alpha = 0{,}001$ → also $p \leq \alpha$

Kleine Wahrscheinlichkeit p spricht gegen den Zufall.

→ H_0 wird auf dem 0,1% Signifikanzniveau abgelehnt.

Interpretation

Kontextbezogene Formulierung
(p-Wert nicht bekannt - im Falle der Rechnung mit Tafelwert)

Die mittlere Herzfrequenzänderung nach 30 Kniebeugen ist um 27,4/min signifikant (α = 0,1%) höher als die mittlere Herzfrequenzänderung nach 5 Kniebeugen.

Interpretation

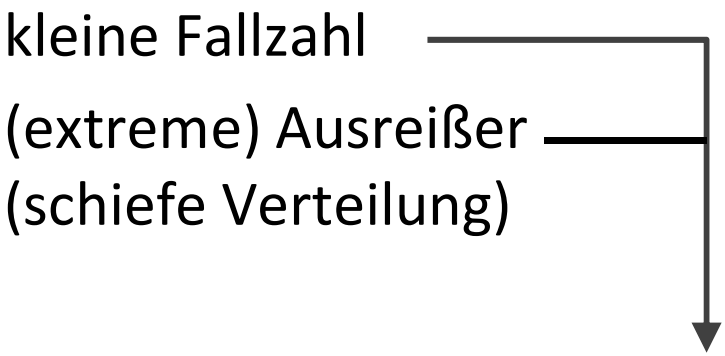
Kontextbezogene Formulierung
(p-Wert bekannt - im Falle der Rechnung mit Statistiksoftware)

Die mittlere Herzfrequenzänderung nach 30 Kniebeugen ist um 27,4/min signifikant höher ($p < 0{,}0005$) als die mittlere Herzfrequenzänderung nach 5 Kniebeugen.

Vergleich der Lage von Verteilungen

Situation:

- kontinuierliches Merkmal
- Vergleich der Lage (location)



a) t-Tests (Normalverteilungsmodell)

b) nichtparametrische Tests (Rangtests)

Tests zum Vergleich von Lagemaßen

Skriptenheft Kapitel 5.7

Stich-proben	Art der Stichproben	NV-Modell	Test
1	----	ja	t-Test (Normwertvergleich)
		nein	Wilcoxon-Test (Normwertvergleich)
2	unabhängig	ja	t-Test für unabhängige Stichproben
		nein	U-Test nach Mann-Whitney
	abhängig	ja	t-Test für Paardifferenzen
		nein	Wilcoxon-Test
> 2	unabhängig	ja	einfaktorielle Varianzanalyse
		nein	H-Test nach Kruskal-Wallis
	abhängig	ja	Varianzanalyse m. Messwiederholung
		nein	Friedman-Test

t-Test für verschiedene Situationen

1) für den Vergleich von 2 unabhängigen Gruppen:

a) bei ungleichen Varianzen ($\sigma \neq \sigma$)

siehe Skript Formel 5.1 (f aufwändiger, siehe Folie 14)

b) bei gleichen Varianzen ($\sigma = \sigma$), siehe Skript Formel 5.3,

Spezialfall $n = n$, Freiheitsgrade $f = n + n - 2$

2) für den Vergleich von 2 abhängigen Gruppen:

t-Test für Paardifferenzen, siehe Skript Formel 5.5,
Freiheitsgrade (FG): $f = n - 1$

3) für den Vergleich einer Gruppe mit einem Normwert:

siehe Skript Formel 5.2, Freiheitsgrade (FG): $f = n - 1$

Aufgabe 2

In der Herzsportsstudie wurden folgende Werte für die Herzfrequenz ermittelt:

	Arm A	Arm B
$\mathbf{HF_{prä}}$ $(\bar{x} \pm s)$	77,07 ± 11,29	76,16 ± 9,97
$\mathbf{HF_{post}}$ $(\bar{x} \pm s)$	93,09 ± 12,83	119,56 ± 17,37
$\mathbf{HF_{diff}}$ $(\bar{d} \pm s)$	16,02 ± 9,32	43,40 ± 17,74
n	128	131

- Prüfen Sie auf dem Signifikanzniveau von 0,2%, ob sich die mittlere Herzfrequenz im Arm A vor und nach den Kniebeugen unterscheidet!

Achtung!

- Statistische Signifikanz bedeutet nicht automatisch klinische Relevanz!

- Fehlende Signifikanz ist kein Beweis für die Gültigkeit der Nullhypothese!

- Kleiner p-Wert bedeutet nicht „großer Therapieeffekt“, sondern nur: die beobachtete Statistik ist unter H_0 ziemlich unwahrscheinlich.
  Deshalb sollte immer auch ein Maß für die Effektgröße angegeben werden.

Hausaufgabe 1a

Überlegen Sie sich, für welchen Test Sie sich in den folgenden Situationen entscheiden! Nutzen Sie dafür die Tabelle auf Folie 25!

a) In einer Studie mit Herzinsuffizienten sollen die Werte des Markers NT-proBNP (pg/ml) zwischen Patienten mit erhaltener Pumpfunktion (HFpEF) und solchen mit reduzierter Pumpfunktion (HFrEF) verglichen werden.
   - Die Verteilung von NT-proBNP ist schief mit einzelnen sehr hohen Werten.
   - Fallzahl: N = 479

Hausaufgabe 1b

b) In einer Pilotstudie soll die Wirksamkeit eines Antihypertensivums an 20 Probanden geprüft werden.
   - Dazu wird der systolische Blutdruck (RRsys) zu Baseline und nach der Einstellphase nach 6 Wochen (primärer Endpunkt) gemessen.
   - 2 Probanden müssen die Behandlung wegen Nebenwirkungen absetzen (fehlende Werte).
   - Extreme Ausreißer sind nicht zu erwarten.
   - Es sollen die Werte zu Baseline und nach 6 Wochen verglichen werden.

Hausaufgabe 1c

c) N = 1853 Herzinsuffiziente werden in einer Trainingsstudie in den Arm A: Standardbehandlung bzw. den Arm B: Standardbehandlung und angeleitetes Training randomisiert.
   - Ein sekundärer Endpunkt ist die Veränderung der Gehstrecke im 6-Minuten-Gehtest.
   - Dieser wird zu Baseline und nach 3 Monaten durchgeführt.
   - Es soll die mittlere Differenz der Gehstrecken zwischen beiden Armen verglichen werden.
   - Gehstrecken weitgehend symmetrisch verteilt, nur einzelne Ausreißer.

Hausaufgabe 2

In der Herzsportsstudie wurden folgende Werte für die Herzfrequenz ermittelt:

	Arm A	Arm B
$\mathbf{HF_{prä}}$ $(\bar{x} \pm s)$	77,07 ± 11,29	76,16 ± 9,97
$\mathbf{HF_{post}}$ $(\bar{x} \pm s)$	93,09 ± 12,83	119,56 ± 17,37
$\mathbf{HF_{diff}}$ $(\bar{d} \pm s)$	16,02 ± 9,32	43,40 ± 17,74
n	128	131

Prüfen Sie auf dem Signifikanzniveau von 1%, ob sich die mittlere Herzfrequenz im Arm B vor und nach den Kniebeugen unterscheidet! Es gibt keine Ausreißer.

Hausaufgabe 3

Table 1. Demographic and Clinical Characteristics of the Patients at Baseline.*

Characteristic	Canagliflozin (N=2202)	Placebo (N=2199)	All Patients (N=4401)
Age — yr	62.9±9.2	63.2±9.2	63.0±9.2

Publikation Perkovich et al. 2019:

- Prüfen Sie auf dem 0,5%-Niveau, ob sich die beiden Studienarme zu Studienbeginn im Alter unterscheiden!

Hinweise:

- Der Test wird hier zu Übungszwecken durchgeführt. Aus fachlicher Sicht ist er eigentlich nicht sinnvoll (warum?)!
- Wegen der sehr hohen Fallzahl kann das Normalquantil verwendet werden.