Statistik 1

Daniel J. F. Gerber

23 April, 2025

Inhaltsverzeichnis

Vorwort

Dieses Buch ist im Rahmen meiner Lehrtätigkeit an der FHNW entstanden und frei verfügbar.

Kapitel 1

Einleitung

1.1 Worum geht es?

1.2 Inhaltlicher Aufbau

Dieses Buch umfasst die untenstehenden Inhalte. Die Inhalte wurden hier nach Zwecken sortiert angeordnet:

Stichprobe beschreiben (**deskriptive Statistik**):

- Arithmetisches Mittel
- Median
- Quantile
- Anteil
- Odds Ratio
- Relatives Risiko

Population beschreiben (**Wahrscheinlichkeitslehre**):

- Zufallsvariable
- Erwartungswert
- Standardabweichung
- Varianz
- Wahrscheinlichkeitsdichte
- Wahrscheinlichkeitsverteilung
- Verteilungen

Populationsparameter aus Stichproben schätzen (**Konfidenzintervalle** + Stichprobengrösse):

- Mittelwert
- Standardabweichung
- Anteil
- Berichten
- Darstellen

Aussagen auf die Population aufgrund von Stichproben machen (Test-Theorie):

- Effektstärke
- Berichten
- T-Test (1 Stichprobe)
- T-Test (2 Stichproben), Welch-Test
- Welch Test
- U-Test
- Korrelation absichern gegen 0
- Vierfelder/Mehrfeldertest

Zusammenhänge beschreiben (Zusammenhangsmasse):

- Pearsons r
- Spearmans rho
- Vierfelderkorrelation / Phi
- Punktbiseriale Korrelation
- Kontingenzkoeffizient
- Cramérs V

Die Inhalte nach Zweck zu gruppieren ist eine Option, die andere ist die Verfahren der Skalierung der Variablen folgend aufzubauen. Bei dieser Gruppierung ist der Zweck nicht direkt ersichtlich, dafür ist einfacher zu begreifen welches Verfahren für welche Ausgangslage geeignet ist. Diese Gruppierung wurde für die Präsentation der Inhalte in diesem Buch gewählt.

1.3 Wie soll ich dieses Buch lesen?

Dieses Buch enthält zu jedem Thema eine kurze Beschreibung der Theorie, Beispiele und Übungen. Das selbstständige Lösen der Übungen ist unerlässlich für das Verständnis und die Emanzipation im korrekten Umgang mit Daten. Ohne Übungen fehlt die Auseinandersetzung mit dem Unterrichtsstoff und ohne diese fällt es den allermeisten schwer sogar einfachste Zusammenhänge zu begreifen. Es wird deshalb empfohlen, dass die Übungen zum jeweiligen Thema zeitnah zur Theorie gelöst werden. Damit überprüft werden kann, ob die Übungen richtig gelöst wurden, ist zu jeder Übung eine kurze Lösung hinterlegt. Wer beim ersten selbstständigen Versuch der Übungslösung scheitert - was garantiert den meisten Lesenden hier ein oder mehrmals passieren wird -, kann die Übung mit Hilfe der Lösung lösen und zu einem späteren Zeitpunkt die Übung selbstständig nochmal machen ohne Lösung. Für die Statistik ist es also *nicht* genug den Stoff einmal auswendig zu lernen, Übung ist unerlässlich. Übungen oder Teilaufgaben welche mit einem $\star$ gekennzeichnet sind, dienen zwar dem tieferen Verständnis der Materie, sind aber für den Kurs nicht absolut notwendig.

Weiter sind Testaufgaben, analog zu einem Multiple-Choice-Test verfügbar, welche zur Überprüfung des Wissens und der Emulation einer Prüfungssituation dienen sollen. Auch diese Fragen werden am besten mehrmals wiederholt bis die Lösung selbstständig korrekt erarbeitet werden kann.

1.4 Formeln, Symbole und Zahlen

Die Statistik bedient sich der universellen Sprache der Formeln. Es ist deshalb unerlässlich einige Formeln zu verstehen. Das Verständnis von Formeln ist für ungeübte Lesende verwirrend und schwierig. Deshalb wird dieses Verständnis in diesem Buch nach und nach aufgebaut. Dazu werden Teilformeln isoliert und erklärt und die Einflüsse der verschiedenen Kenngrössen in der Formel exploriert.

Es gibt eine vielzahl von Möglichkeiten Formeln und Zahlen in einem Manuskript niederzuschreiben. Um die Formeln, Symbole und Zahlen verständlich und vergleichbar zu halten wurden verschiedene Standards definiert. In diesem Buch wird der Standard Richtlinien zur Manuskriptgestaltung der Deutschen Gesellschaft für Psychologie verwendet (**?**). Dieser ist wiederum stark an den Standard der American psychological assiciation angelehnt.

1.5 Software

Für die Lösung der Übungen wird oft die freie Software `Jamovi` verwendet. Den Lesenden wird deshalb empfohlen `Jamovi` zu installieren. Für die Erstellung dieses Buches wurden ferner die folgenden Softwareprodukte verwendet:

- Jamovi software (Version 2.3.21.0)
- Jamovi R-package (**?**)
- R (**?**)
- Tidyverse (**?**)
- Bookdown (**?**)

Teil I

Eine intervallskaliertes Merkmal

Kapitel 2

Intervallskalierte Merkmale

2.1 Was ist ein intervallskaliertes Merkmal?

Ein Merkmal ist dann **intervallskaliert**, wenn die einzelnen Beobachtungen in eine natürliche Reihenfolge gebracht werden können und zwischen dem tiefsten und höchsten möglichen Wert, alle erdenklichen Zwischenwerte möglich sind.

Beispiel 2.1 (Körpertemperatur)**.** Ein Beispiel für ein intervallskaliertes Merkmal ist die Körpertemperatur. Beobachtungen der Körpertemperatur einer lebenden Person sind Werte zwischen ungefähr 10°C und 42°C. Es ist möglich zu sagen, dass eine Person mit 40°C Körpertemperatur eine höhere Temperatur hat als eine mit 38°C Körpertemperatur. Ausserdem sind alle erdenklichen Zwischenwerte möglich, so auch dass bei einer Person eine Körpertemperatur von 37.821239°C gemessen wird.

Beispiel 2.2 (Intelligenzquotient)**.** Ein weiteres Beispiel für ein intervallskaliertes Merkmal ist der Intelligenzquotient IQ. Der IQ bewegt sich normalerweise zwischen 50 und 150, eine Person mit einem IQ von 105 hat einen höheren IQ als eine Person mit einem IQ von 103. Ausserdem sind IQ-Werte von 103.12 oder 118.9182 durchaus möglich.

Klicke hier, falls dir verhältnisskalierte Merkmale bekannt sind

Die folgende Diskussion ist auch auf verhältnisskalierte Merkmale anwendbar. Letztere sind intervallskalierte Merkmale, welche einen absoluten Nullpunkt aufweisen.

2.2 Wie kann ein intervallskaliertes Merkmal beschrieben werden?

Beispiel 2.3 (Körpertemperatur Enten)**.** Eine Veterinärin möchte herausfinden, welche Körpertemperatur Enten aufweisen. Dazu untersucht sie 40 Enten und misst die Körpertemperaturen 42.01, 41.72, 41.51, 41.52, 41.5, 41.6, 41.46, 41.81, 42.14, 41.82, 42.06, 41.53, 41.66, 41.65, 41.46, 41.48, 41.92, 41.58, 41.32, 41.58, 41.81, 41.7, 41.62, 41.52, 41.89, 41.53, 41.67, 41.43, 42.18, 41.52, 41.82, 41.96, 41.8, 41.54, 41.88, 41.69, 41.92, 41.35, 41.07 und 41.67.

Für einen Menschen ist es schwierig direkt aus der Sichtung dieser Zahlen zu begreifen, welche Körpertemperatur Enten haben. Ein Mensch kann sich jedoch helfen, indem er die Zahlen zusammenfasst.

2.2.1 Verteilung

Um die Zahlen zusammenzufassen, kann die Veterinärin zum Beispiel Temperaturabschnitte von 0.2°C betrachten und zählen wie viele Beobachtungen sie in den jeweiligen Abschnitten gemacht hat. Diese Zähldaten können tabellarisch oder grafisch mit einem Balkendiagramm dargestellt werden. Letzteres wird ein **Histogramm** genannt.

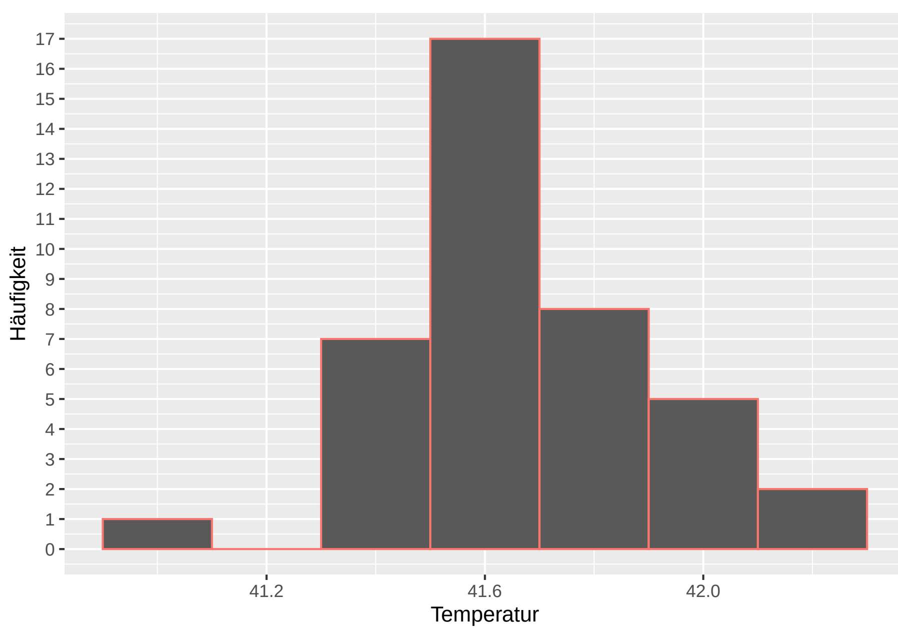


Abbildung 2.1: Histogramm Körpertemperatur Enten.

Aufgrund dieser Darstellung kann die Veterinärin nun sehen, wie häufig welche Körpertemperaturen sind. Dies wird die **Verteilung** des Merkmals genannt. Sie bemerkt zum Beispiel, dass Beobachtungen der Körpertemperatur rund um 41.6°C am häufigsten sind und tiefere und höhere Temperaturen seltener vorkommen. Auf einen Blick sieht sie auch, dass die Temperatur aller Enten zwischen 41°C und 42.2°C war.

Die Verteilung eines Merkmals zu kennen ist hilfreich, jedoch in vielen Situationen (z. B. in der Kommunikation) noch zu komplex. Einfacher ist es die Komplexität einer Verteilung auf zwei Faktoren herunterzubrechen: Die Zentralität und die Variabilität eines Merkmals.

2.2.2 Zentralität

Mit der Zentralität ist ein Wert gemeint, welcher die zentrale Tendenz des Merkmals abbildet. Um die Zentralität zu messen, gibt es drei Möglichkeiten:

- Der **Modus** ist der am häufigsten vorkommende Wert. Im Beispiel ist das der Wert 41.52, welcher 3 mal und damit am häufigsten vorkommt. In Jamovi wird der Modus mit `Modalwert` bezeichnet.
- Wenn die Werte des Merkmals aufsteigend sortiert werden und der Wert betrachtet wird, welcher die Beobachtungen in eine tiefere und eine höhere Hälfte teilt, dann wird dieser Wert als **Median** (abgekürzt *Mdn*, Symbol $\tilde{x}$) bezeichnet. Bei einer geraden Anzahl Beobachtungen, wird in der Regel der Durchschnittswert der beiden mittigsten Beobachtungen verwendet. Im Beispiel haben wir 40 Beobachtungen. Der Median entspricht also dem Durchschnittswert zwischen dem 20. und dem 21. der aufsteigend sortierten Werte 41.07, 41.32, 41.35, 41.43, 41.46, 41.46, 41.48, 41.5, 41.51, 41.52, 41.52, 41.52, 41.53, 41.53, 41.54, 41.58, 41.58, 41.6, 41.62, 41.65, 41.66, 41.67, 41.67, 41.69, 41.7, 41.72, 41.8, 41.81, 41.81, 41.82, 41.82, 41.88, 41.89, 41.92, 41.92, 41.96, 42.01, 42.06, 42.14 und 42.18, also 41.655. In Jamovi wird der Median mit `Median` bezeichnet.
- Das **arithmetische Mittel** (abgekürzt *M*, Symbol $\bar{x}$) bezeichnet, was gemeinhin mit Durchschnitt gemeint ist. Wenn wir die erste von insgesamt n Beobachtung mit x_1 und die letzte Beobachtung mit x_n bezeichnen, so ist das arithmetische Mittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

  Im Beispiel ist das arithmetische Mittel der Körpertemperaturen 41.6725. In Jamovi wird das arithmetische Mittel als `Mittelwert` bezeichnet.

Achtung

Hinweis. Erklärung der Formel: Hier wird zum ersten Mal eine Formel verwendet. $\sum$ steht für die Summe von allen Beobachtungen x_i, wenn der Index i in 1-Schritten von der Zahl unter dem Summenzeichen $i = 1$ bis zu der Zahl oben am Summenzeichen $i = n$ läuft. In unserem Beispiel ist $n = 40$, also ist $i = 1, 2, 3, 4, \ldots, 39, 40$. Der Teil $\sum_{i=1}^{n} x_i$ bedeutet also nichts anderes als $x_1 + x_2 + \ldots + x_{39} + x_{40}$, also die Summe aller Beobachtungen. $\frac{1}{n}$ bedeutet, dass wir diese Summe jetzt noch durch die Anzahl Beobachtungen teilen.
Welchen Einfluss haben die verschiedenen Einflussgrössen: Dies wird in Übung **??** erklärt.

Jedes dieser Masse für die Zentralität hat Vor- und Nachteile und sie werden dementsprechend in unterschiedlichen Situationen eingesetzt, siehe Übungen.

2.2.3 Variabilität

– Die **Spannweite** (abgekürzt R aus dem englisch *range*) ist der höchste beobachtete Wert minus der kleinste beobachtete Wert. Im Beispiel ist der höchste beobachtet Wert $42.18°C$ und der kleinste Beobachtete Wert $41.07°C$. Also ist die Spannweite $42.18 - 41.07 = 1.11°C$. Die Spannweite wird in Jamovi mit `Wertebereich` bezeichnet.

– Wenn die Werte des Merkmals aufsteigend sortiert werden und der Wert betrachtet wird, welcher die Beobachtungen in eine $P\%$ tiefere und $(100\% - P\%)$ höhere Hälfte teilt, dann wird dieser Wert als **Perzentil** bezeichnet. Das 5%-Perzentil zum Beispiel teilt die beobachteten Werte in 5% kleinere und 95% grössere Werte. Im Beispiel haben wir 40 Beobachtungen. 5% davon sind demnach 2 Beobachtungen die tiefer sind als das 5% Perzentil und 95% also 38 Beobachtungen die höher sind als das 5% Perzentil. Das 5% Perzentil liegt also zwischen $41.32°C$ und $41.35°C$. In diesem Fall wird ein Mittelwert der beiden nächsten Werte verwendet, hier $(41.32 + 41.35)/2 = 41.34°C$. Das $P\%$-Perzentil kann in Jamovi bei `Perzentil` gefolgt von der Zahl `P` ermittelt werden. Ein Perzentil alleine gibt jedoch noch keinen Hinweis auf die Streuung der Werte. Werden aber zwei Perzentile zusammen betrachtet, z. B. das 5% und das 95% Perzentil, dann geben diese Werte und der Abstand dazwischen einen Hinweis auf die Streuung der Beobachtungen. Im Beispiel ist das 5% Perzentil bei $41.34°C$ und das 95%-Perzentil bei $42.1°C$. Hier befinden sich also 90% aller Beobachtungen zwischen diesen Werten. Mehrere Perzentile können in Jamovi gleichzeitig angezeigt werden indem die Perzentil-Werte mit Komma getrennt werden, für die Perzentile hier im Beispiel `0.05, 0.95`. Weitere beliebte Werte sind das 25% und das 75%-Perzentil (auch **Quartile** genannt, da sie die beobachteten Werte vierteln), im Beispiel bei $41.52°C$ und $41.82°C$ respektive. Die Differenz dieser Perzentile wird als **Interquartilabstand** (abkürzung IQR von interquartile range) bezeichnet und ist im Beispiel $0.3°C.$. Der Interquartilabstand wird in Jamovi mit `IQR` bezeichnet.

– Die **Standardabweichung** (abgekürzt *SD*, Symbol s) ist die durchschnittliche Abweichung jeder Beobachtung vom arithmetischen Mittel. Wenn wir die erste von insgesamt n Beobachtung mit x_1 und die letzte Beobachtung mit x_n bezeichnen, so ist die Standardabweichung

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

Im Beispiel ist die Standardabweichung der Körpertemperaturen $0.233°C$. In Jamovi wird die Standardabweichung mit `Std.-abweichung` bezeichnet.

Achtung

Hinweis. Erklärung der Formel: $(x_i - \bar{x})$ bezeichnet den Abstand von jeder Beobachtung zum arithmetischen Mittel. Dieser Abstand kann positiv (wenn x_i grösser ist als $\bar{x}$) oder negativ (wenn x_i kleiner ist als $\bar{x}$) ausfallen. Damit diese positiven und negativen Abstände sich in der Summe nicht ausgleichen und eine Standardabweichung von 0 entsteht, werden diese Abstände quadriert $(x_i - \bar{x})^2$ bevor sie summiert werden. Anschliessend wird diese Summe durch $n-1$ geteilt, um den durchschnittlichen Abstand pro Beobachtung zu ermitteln. Intuitiv würde man hier durch n teilen. Statistiker:innen haben jedoch herausgefunden, dass es einige Vorteile hat, wenn durch $n-1$ statt n geteilt wird. Das Quadrat wird nach der Aufsummierung wieder aufgehoben indem die Quadratwurzel gezogen wird.

2.3 Übungen

Übung 2.1.

Mit den Daten `02-exm-ducktemp.sav` aus Beispiel **??**:

(a) Erstellen Sie selbst ein Histogramm mit Jamovi und begründen Sie, weshalb es nicht gleich aussieht wie das Histogramm in Abbildung **??**.

(b) Berechnen Sie Modus, Median und arithmetisches Mittel der Körpertemperaturen der Enten mit `Jamovi` und berichten Sie diese mit der angemessenen Symbolik.

(c) (⋆) Reproduzieren Sie das Histogramm in Abbildung **??** genau mithilfe der Histogramm-Funktion des `Jamovi`-Moduls `JJStatsPlot` (**?**).

(d) Berechnen Sie IQR, 25%- und 75%-Perzentil, sowie 2.5%- und 97.5%-Perzentil, sowie die Spannweite und die Standardabweichung der Körpertemperaturen der Enten mit `Jamovi` und berichten Sie diese mit der angemessenen Symbolik.

Lösung.

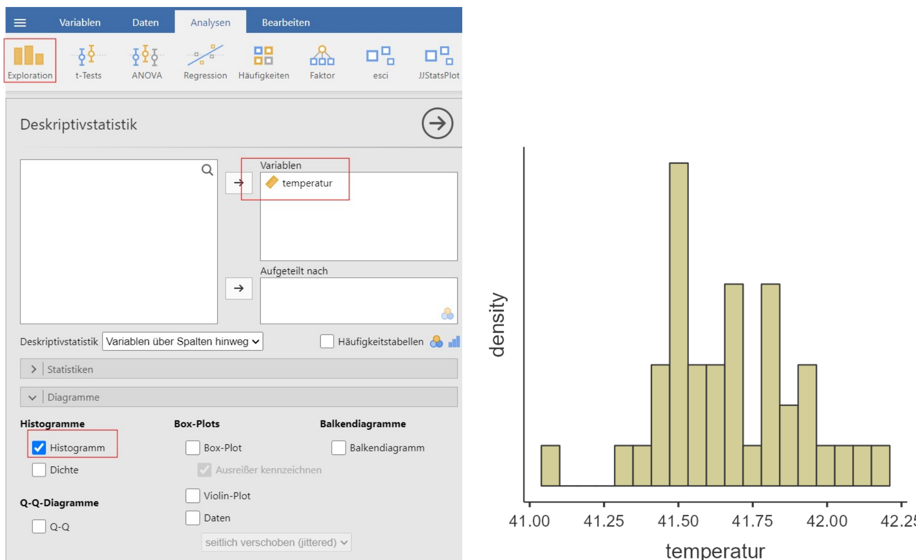
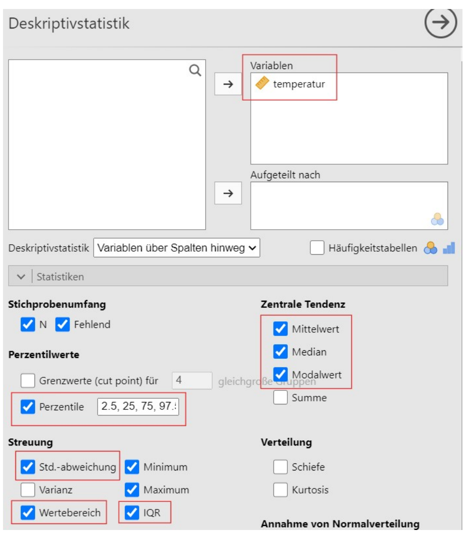


Abbildung 2.2: Links: Jamovi-Anleitung zur Erstellung des Histogramms; rechts: Histogramm der Temperatur.

(a) Das Histogramm, siehe Abbildung **??** sieht nicht gleich aus, da Jamovi die Temperaturabschnitte mit Korbbreite 0.125°C kürzer gewählt hat als die in Abbildung Abbildung **??** dargestellte Korbbreite von 0.2°C. Ein Histogramm sieht immer anders aus je nach ausgewählter Abschnittsweite.

(b) Eine Anleitung zur Berechnung in `Jamovi` sowie die berechneten Werte können in Abbildung **??** abgelesen und sind Modus = 41.5°C, Median $Mdn = 41.7$ und arithmetisches Mittel $M = 41.7$.

(c) Im Modul `JJStatsPlot` kann die Korbgrösse mit `Change Bin Width` angepasst werden. In Abbildung Abbildung **??** kann beobachtet werden,



Deskriptivstatistik

	temperatur
N	40
Fehlend	0
Mittelwert	41.7
Median	41.7
Modalwert	41.5
Standardabweichung	0.233
IQR	0.300
Wertebereich	1.11
Minimum	41.1
Maximum	42.2
2.5. Perzentil	41.3
25. Perzentil	41.5
75. Perzentil	41.8
97.5. Perzentil	42.1

Abbildung 2.3: Links: Jamovi-Anleitung zur Berechnung der gewünschten Parameter; rechts: Parameterwerte.

dass die Balken und also auch die Körbe 0.2 Einheiten breit sind. Dies wird so eingestellt, siehe Abbildung **??**.

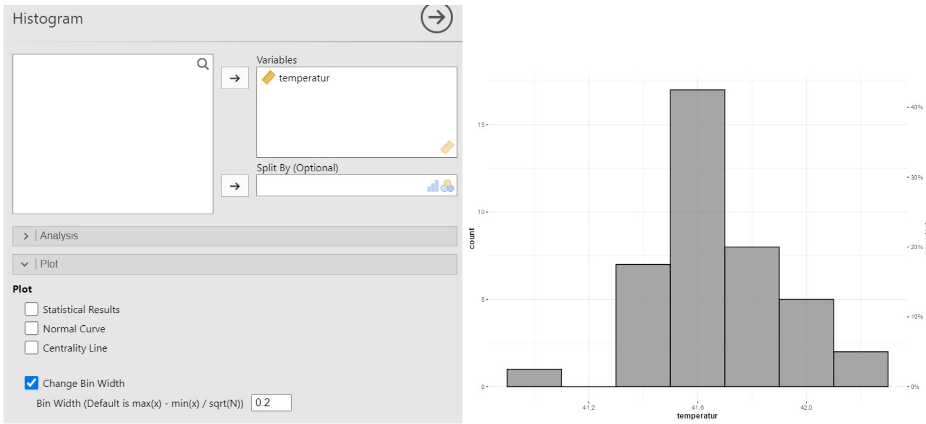


Abbildung 2.4: Links: Jamovi-Anleitung zur Berechnung des gewünschten Histogramms; rechts: Jamovi-Ausgabe.

Es entsteht dabei glücklicherweise genau die gewünschte Darstellung. Es wäre auch möglich gewesen, dass die Körbe auf der x-Achse verschoben sind, zum Beispiel ein Korb 41.2 bis 41.4. Diese Verschiebung könnte nicht mit `JJStatsPlot` behoben werden und müsste mit einer anderen Statistiksoftware bearbeitet werden. (d) Eine Anleitung zur Berechnung in `Jamovi` sowie die berechneten Werte können in Abbildung **??** abgelesen und sind Interquartilabstand $IQR = 0.3$°C, 25%-Perzentil $= 41.5$, 75%-Perzentil $= 41.8$, 2.5%-Perzentil $= 41.3$, 97.5%-Perzentil $= 42.1$, Spannweite $R = 1.11$ und die Standardabweichung $SD = 0.233$.

Übung 2.2.

TODO: Exercise body

Lösung. TODO: solution body

Übung 2.3.

In einem psychologischen Test machen 5 Probandinnen die Werte $18, 21, 20, 19, 22$. Um mit einer Zahl zu sagen, wo die Testresultate liegen, wird ein zentraler Wert berechnet.

(a) Wie gross ist das arithmetische Mittel und der Median dieser Werte? Rechnen Sie im Kopf oder mit einem Taschenrechner.
(b) Nehme an, der Testleiter hat den Wert der ersten Probandin falsch in seine Tabelle übertragen - statt 18 hat er 81 geschrieben. Wie gross ist das arithmetische Mittel und der Median dieser Werte in diesem Fall?

(c) Gleich wie (a), aber führen Sie die Berechnungen aus indem die Zahlen manuell bei `Jamovi` eingegeben. Tipp: Die Messskala muss manuell auf kontinuierlich gestellt werden.
(d) Gleich wie (b), aber führen Sie die Berechnungen aus indem die Zahlen manuell bei `Jamovi` eingegeben.
(e) Was sagt dies über den Median und das arithmetische Mittel aus?

Lösung.

(a) Wir haben hier $n = 5$ Beobachtungen, nämlich $x_1 = 18, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Wird dies in die Formel (**??**) eingesetzt, so gibt dies das arithmetische Mittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1+x_2+x_3+x_4+x_5) = \frac{1}{5}(18+21+20+19+22) = 20.$$

Um den Median zu berechnen, werden die Werte zuerst aufsteigend sortiert $18, 19, 20, 21, 22$. Der Wert, welcher die Werte in eine grössere und eine kleinere Hälfte teilt, ist hier 20, was dem Median entspricht.
(b) Die Beobachtungen sind jetzt $x_1 = 81, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Analog wie in (a) kann demnach das arithmetische Mittel als $\bar{x} = 32.6$ bestimmt werden. Die aufsteigend sortierten Beobachtungen sind nun $19, 20, 21, 22, 81$. Der Median ist also 21.

Für c und d wird der Datensatz bei `Jamovi` eingegeben, siehe Abbildung **??**, und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

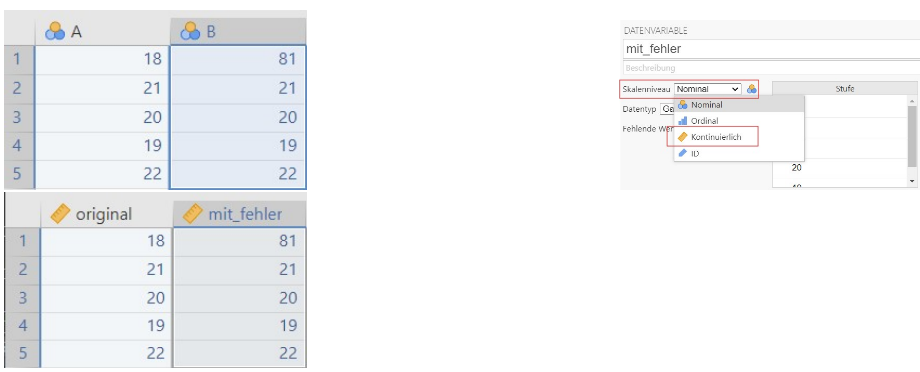


Abbildung 2.5: Jamovi Dateneingabe.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die beiden nächsten Teilfragen beantwortet werden.

(c) Das Resultat in `Jamovi` ist genau gleich wie das händisch berechnete.

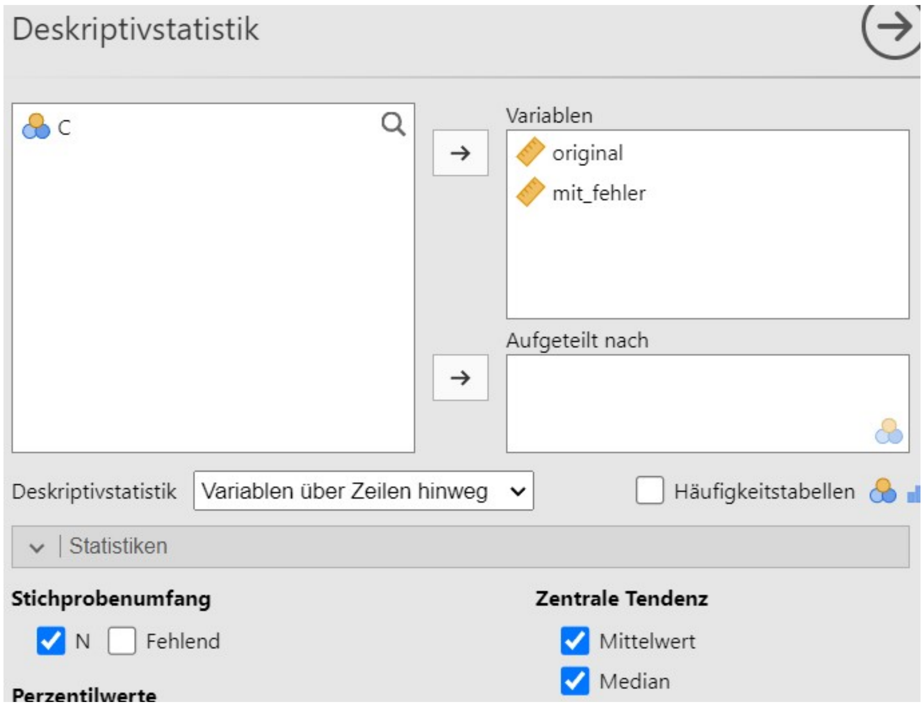


Abbildung 2.6: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

	N	Mittelwert	Median
original	5	20.0000	20
mit_fehler	5	32.6000	21

Abbildung 2.7: Jamovi Ausgabe.

(d) Das Resultat in `Jamovi` ist genau gleich wie das händisch berechnete.
(e) Durch die fälschliche Übertragung eines Wertes, ist das arithmetische Mittel sehr stark und der Median fast gar nicht beeinflusst werden. Wenn die Daten wenige fehlerhafte Beobachtungen enthalten, ist der Median das bessere Mass für den zentralen Wert als das arithmetische Mittel. Wenn die Daten keine Fehler enthalten, ist das arithmetische Mittel gleich gut geeignet wie der Median.

Übung 2.4.

In einem psychologischen Test machen 5 Probandinnen die Werte $18, 21, 20, 19, 22$. Um mit einer Zahl zu sagen, wie stark die Testresultate streuen, wird die Variabilität berechnet.

(a) Wie gross ist die Spannweite und die Standardabweichung dieser Werte? Rechnen Sie im Kopf oder mit einem Taschenrechner.
(b) Nehme an, der Testleiter hat den Wert der ersten Probandin falsch in seine Tabelle übertragen - statt 18 hat er 81 geschrieben. Wie gross ist die Spannweite und die Standardabweichung dieser Werte jetzt? Rechnen Sie im Kopf oder mit einem Taschenrechner.
(c) Gleich wie (a), aber führen Sie die Berechnungen aus indem die Zahlen manuell bei `Jamovi` eingegeben. Tipp: Die Messskala muss manuell auf kontinuierlich gestellt werden.
(d) Gleich wie (b), aber führen Sie die Berechnungen aus indem die Zahlen manuell bei `Jamovi` eingegeben.
(e) Vergewissern Sie sich, dass der Interquartilabstand in jedem Fall dem Abstand zwischen dem 25% und dem 75%-Perzentil entspricht. Vergewissern Sie sich zusätzlich, dass in jedem Fall der Median dem 50%-Perzentil entspricht.
(f) Schliessen Sie aus dieser Übung auf das Verhalten der verschiedenen Variabilitätsmasse bei fehlerhaften Daten?

Lösung.

(a) Die Spannweite entspricht dem höchsten minus dem kleinsten beobachteten Wert, also $R = 22 - 18 = 4$. Wir haben hier $n = 5$ Beobachtungen, nämlich $x_1 = 18, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Wird dies in die Formel (**??**) eingesetzt, so gibt dies die Standardabweichung

$$
\begin{aligned}
s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} && (2.3)\\
&= \sqrt{\frac{1}{n-1}\left((x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2\right)} && (2.4)\\
&= \sqrt{\frac{1}{5-1}\left((18-20)^2+(21-20)^2+(20-20)^2+(19-20)^2+(22-20)^2\right)} && (2.5)\\
&= \sqrt{\frac{1}{4}(4+1+0+1+4)} && (2.6)\\
&= 1.58. && (2.7)
\end{aligned}
$$

(b) Wir haben hier $n = 5$ neue Beobachtungen, nämlich $x_1 = 81, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Die Spannweite entspricht dem höchsten minus dem kleinsten beobachteten Wert, also $R = 81-19 = 62$. Wird dies in die Formel (**??**) eingesetzt, so gibt dies die Standardabweichung

$$
\begin{aligned}
s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} && (2.8)\\
&= \sqrt{\frac{1}{n-1}\left((x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2\right)} && (2.9)\\
&= \sqrt{\frac{1}{5-1}\left((81-32.6)^2+(21-32.6)^2+(20-32.6)^2+(19-32.6)^2+(22-32.6)^2\right)} && (2.10)\\
&= \sqrt{\frac{1}{4}(2342.56+134.56+158.76+184.96+112.36)} && (2.11)\\
&= 27.08. && (2.12)
\end{aligned}
$$

Je nach Rundungsverfahren können hier kleinere Werteunterschiede im Nachkommabereich resultieren.

Für c und d wird der Datensatz bei `Jamovi` eingegeben. Die Variablen werden bearbeitet wie in **??** beschrieben. Die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die beiden nächsten Teilfragen beantwortet werden.

(c) Tatsächlich ist die Spannweite gemäss Jamovi auch $R = 4$ (siehe `Wertebereich`) und die Standardabweichung ist $SD = 1.58$.

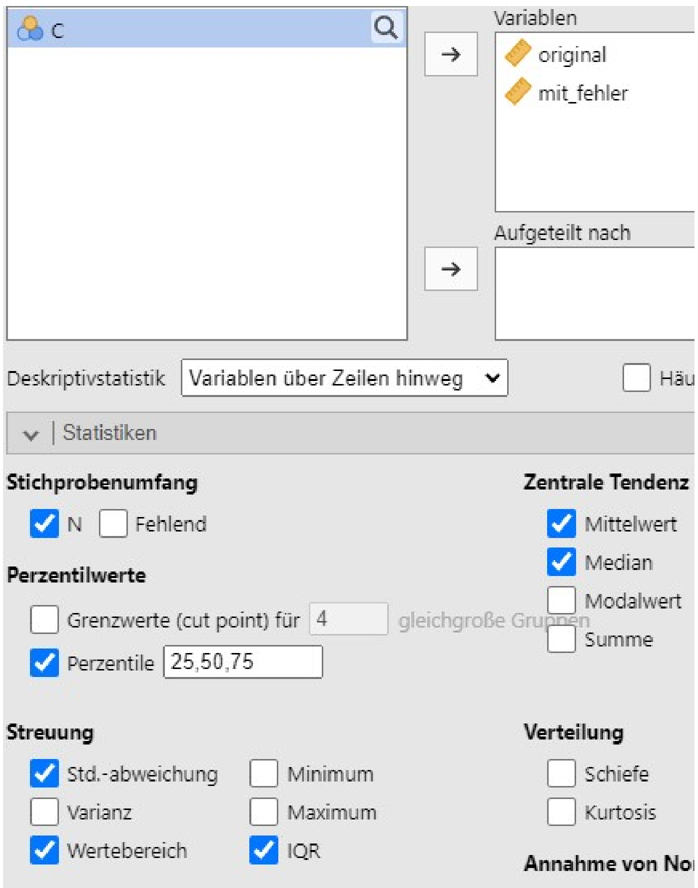


Abbildung 2.8: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

							Perzentile		
	N	Mittelwert	Median	Std.-abw.	IQR	Wertebereich	25th	50th	75th
original	5	20.0000	20	1.5811	2.0000	4	19.0000	20.0000	21.0000
mit_fehler	5	32.6000	21	27.0795	2.0000	62	20.0000	21.0000	22.0000

Abbildung 2.9: Jamovi Ausgabe.

(d) Tatsächlich ist die Spannweite gemäss Jamovi auch $R = 62$ (siehe `Wertebereich`) und die Standardabweichung ist $SD = 27.08$.
(e) Tatsächlich ist der $IQR = 2$ in beiden Beispielen. Dies entspricht genau den Perzentildifferenzen $21-19$ für den original und $22-20$ für den fehlerhaften Datensatz. Dass zweimal genau derselbe Wert resultiert ist Zufall. In beiden Fällen entspricht der Median dem 50%-Perzentil. Dies sollte immer der Fall sein, da sowohl der Median, wie auch das 50%-Perzentil die Beobachtungen in eine höhere und eine tiefere Hälfte teilen.
(f) Diese Übung zeigt, dass die Standardabweichung und die Spannweite durch fehlerhafte Beobachtungen stark beeinflusst werden. Der Interquartilabstand ist hingegen relativ stabil, solange nicht viele Beobachtungen fehlerhaft sind.

Übung 2.5.

Bei einer Befragung wurden die Körpergrösse und das Geschlecht im Datensatz `02-exr-koerpergroesse-sex.sav` festgehalten.

(a) Stellen Sie die Körpergrösse in einem Histogramm dar und berechnen sie alle bekannten Zentralitäts- und Variabilitätsmasse und berichten Sie mit dem korrekten Symbol.
(b) Wiederholen Sie die Übung aber teilen Sie die Daten nach Geschlecht auf. Was fällt auf?
(c) Was bedeutet der Kommentar *Es gibt mehr als einen Modalwert, nur der erste wird berichtet* und welche Bedeutung hat er für die Interpretation des Modus?

Lösung.

(a) Der Datensatz wird bei `Jamovi` eingelesen und die Analyseparameter wie in Abbildung **??** gesetzt.

Dies produziert das Analyseergebnis in Abbildung **??**.

Die Körpergrösse ist demnach $N = 326$ mal beobachtet worden. Die Zentralitätmasse sind $M = 173.13$ cm, $Mdn = 172$ (Rundung nach 2 Kommastellen), Modus 144.77. Die Variabilitätsmasse sind $SD = 12.09$ cm, $IQR = 17.11$ und $R = 57.78$. Auf dem Histogramm ist ausserdem ersichtlich, dass die meisten Leute zwischen 160 und 180 cm gross sind und dass nur weniger unter 155 oder über 200 cm gross sind.

(b) Die Analyse wird mit Gruppierungsvariable Geschlecht wiederholt wie in Abbildung **??**

Dies produziert das Analyseergebnis in Abbildung **??**.

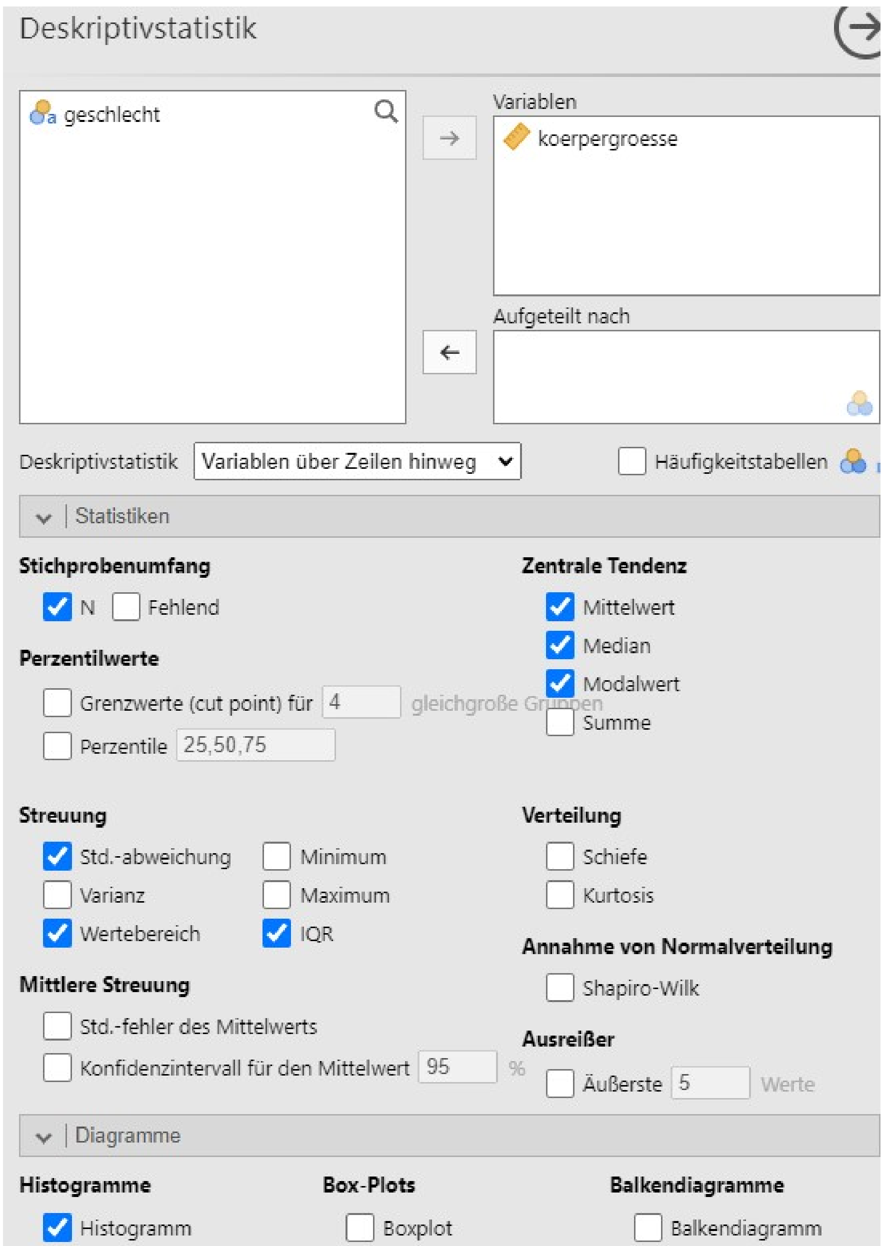


Abbildung 2.10: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

	N	Mittelwert	Median	Modalwert	Std.-abw.	IQR	Wertebereich
koerpergroesse	326	173.1348	171.9953	144.7707 [a]	12.0893	17.1101	57.7808

[a] Es gibt mehr als einen Modalwert, nur der erste wird berichtet

Diagramme

koerpergroesse

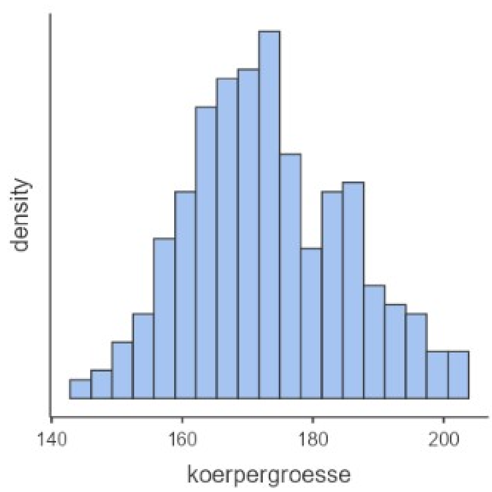


Abbildung 2.11: Jamovi Ausgabe.

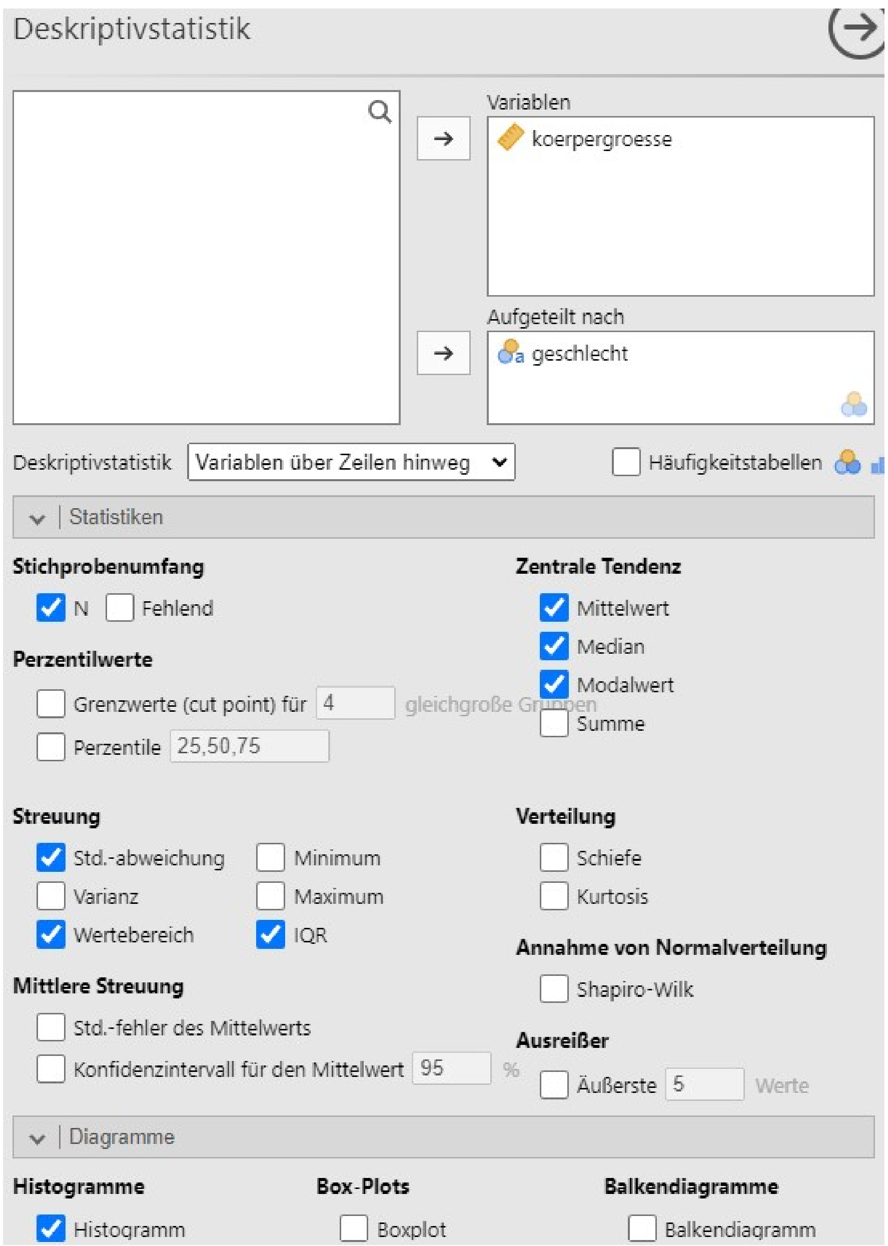


Abbildung 2.12: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

	geschlecht	N	Mittelwert	Median	Modalwert	Std.-abw.	IQR	Wertebereich
koerpergroesse	w	163	166.0030	166.0918	144.7707 [a]	8.4052	11.2749	41.4469
	m	163	180.2666	180.5412	149.1652 [a]	10.9639	15.3514	53.3863

[a] Es gibt mehr als einen Modalwert, nur der erste wird berichtet

Diagramme

koerpergroesse

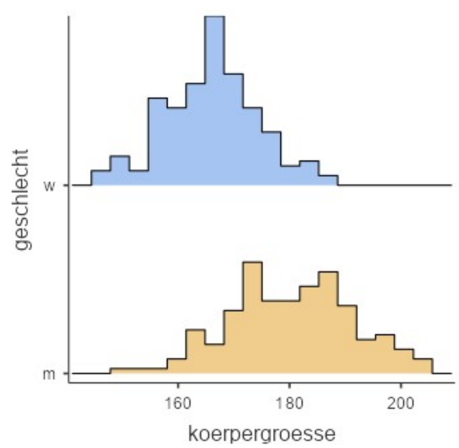


Abbildung 2.13: Jamovi Ausgabe.

Die Körpergrösse ist demnach bei den Frauen $N = 163$ mal beobachtet worden. Die Zentralitätmasse sind $M = 166.00$ cm, $Mdn = 166.09$, Modus 144.77. Die Variabilitätsmasse sind $SD = 8.41$ cm, $IQR = 11.27$ und $R = 41.45$. Die Körpergrösse ist bei den Männern auch $N = 163$ mal beobachtet worden. Die Zentralitätmasse sind $M = 180.27$ cm, $Mdn = 180.54$, Modus 149.17. Die Variabilitätsmasse sind $SD = 10.96$ cm, $IQR = 15.35$ und $R = 53.39$. Auf dem Histogramm ist ausserdem ersichtlich, dass die beiden Gruppen eine Spitze rund um den Mittelwert bei der Häufigkeit aufweisen. Beobachtungen, welche von der Spitze weiter weg sind werden seltener. Diese zwei Spitzen und somit unterschiedliche Körpergrössenverteilungen nach Geschlecht war aus (a) nicht ersichtlich. Es ist immer möglich, dass bei nicht-Experimenten ein zusätzliches Merkmal (hier das Geschlecht) ganz neue Erkenntnisse bringen kann. (c) In den Daten der Körpergrösse ist ersichtlich, dass aufgrund der detaillierten Aufzeichnung der Körpergrössen im hundertstel Millimeterbereich keine Beobachtung zweimal vorkommt. Jede Beobachtung ist somit die häufigste Beobachtung. Der Modus ist hier also bedeutungslos. Um einen sinnvolleren Wert für den Modus zu erhalten könnten die Körpergrössen vorab auf Zentimeter gerundet werden. In `Jamovi` kann dies mit der Funktion `ROUND` gemacht werden. Der Modus ist dann 165 cm für die Frauen und 172 für die Männer, was sich mit den Erwartungen aus dem Histogramm deckt.

Übung 2.6.

Für eine Studie werden Studierende gebeten eine Aufgabe zu lösen, bei welcher Sie eine gewisse Anzahl `Punkte` erzielen. Über jede Proband:in sind ausserdem folgende Eigenschaften bekannt:

- `IQ`: Intelligenzquotient
- `Aufgeschlossenheit`: Likert von 1-7
- `Wartezeit_min`: Wartezeit vor beginn des Experiments in Minuten
- `Wartezeit_std`: Wartezeit vor beginn des Experiments in Stunden
- `Geburtzeit_std_ab_mitternacht`: Geburtszeit in Stunden ab Mitternacht. Wenn jemand um 13h30 auf die Welt kam, ist dieser Wert 13.5.
- `Geburtzeit_std_ab_mittag`: Geburtszeit in Stunden ab Mittag. Wenn jemand um 13h30 auf die Welt kam, ist dieser Wert 1.5.

Die Daten sind in Jamovi unter `02-exr-diverse-distrib.sav` verfügbar.

Analysieren Sie alle erhobenen Merkmale indem Sie ein Histogramm erstellen und die zentralen Tendenzen sowie die Variabilität analysieren.

a. Wie viele Personen nahmen an der Studie teil?
b. Vergleichen Sie Ihre Ergebnisse für die Merkmale IQ und Aufgeschlossenheit. Was für Zusammenhänge fallen auf?
c. Vergleichen Sie Ihre Ergebnisse für die Wartezeiten Merkmale. Was für Zusammenhänge fallen auf?

d. Vergleichen Sie Ihre Ergebnisse für die Merkmale Punkte und Wartezeiten. Was für Zusammenhänge fallen auf?
e. Geburtszeit. TODO.

Lösung.

Die Merkmale werden mit den Befehlen in Abbildung **??** analysiert.

a. Es gibt gemäss **??** genau 500 Studienteilnehmende (siehe N).
b. Die Histogramme für den IQ und die Aufgeschlossenheit weisen eine ähliche Form auf. Viele Beobachtungen sind um eine Mitte zentriert. Je weiter weg von der Mitte, desto seltener sind die Beobachtungen. Das Histogramm des IQ zeigt, dass die Verteilung rund um 100 zentriert ist und ca. von 60 bis 140 reicht. Je weiter entfernt von 100, desto weniger Beobachtungen wurden gemacht. Das Histogramm der Aufgeschlossenheit stellt dar, dass diese rund um 4 zentriert ist mit Werten von 1 bis 7. Je weiter die Werte von 4 entfernt sind, desto weniger häufig sind die Beobachtungen. Der vom Histogramm abgeleitete vorher genannte zentrale Wert entspricht ungefähr dem Mittelwert und dem Median für beide Merkmale. Für die Aufgeschlossenheit hat der Modalwert ebenfalls einen ähnlichen Wert. Der Modus für den IQ ist nicht belastbar, da die Fussnote besagt, dass mehrere Werte als Modus in Frage kommen. Eine genauere Durchsicht der IQ-Werte lässt folgern, dass aufgrund der vielen Nachkommastellen jeder IQ-Wert nur genau einmal vorkommt. Der angebene Modalwert des IQs entspricht also einfach einer zufälligen Beobachtung. Die Kennwerte für die Variabilität lassen ebenfalls auf Unterschiede zwischen den beiden Merkmalen schliessen. Die höheren Werte Standardabweichung, IQR und Wertebereich des IQ im Vergleich zur Aufgeschlossenheit legen nahe, dass die Streuung der Werte für den IQ viel grösser ist. Zum Beispiel ist eine durchschnittliche IQ-Beobachtung 15.5 IQ-Werte weg vom durchschnittlichen IQ und eine durchschnittliche Aufgeschlossenheits-Beobachtung nur 1.3 Aufgeschlossenheits-Werte weg von der duchschnittlichen Aufgeschlossenheit. Dies ist auf dem Histogramm zu erkennen, wenn die Skala der horizontalen Achse betrachtet wird. Für den IQ reicht diese von 50 bis 125 und für die Aufgeschlossenheit von 2 bis 6.
c. Die Wartezeiten wurden einmal in Minuten und einmal in Stunden abgespeichtert. Die resultierenden Histogramme sind deshalb genau identisch bis auf die Werte der horizontalen Achse, welche von 0 bis 0.6 Stunden und von 0 bis 40 Minuten reicht. Im Vergleich zu den Histogrammen des IQ und der Aufmerksamkeit kann für die Wartezeit und eine asymetrische Verteilung beobachtet werden. Kurze Wartezeiten werden demnach häufiger beobachtet als längere Wartezeiten. Die meisten Wartezeiten liegen unter 10 Minuten, sehr selten kommt es zu Wartezeiten über 20 Minuten. Die Kennzahlen für die Wartezeit in Stunden können aus den Kennzahlen der Wartezeit in Stunden hergeleitet werden indem die Werte durch 60 geteilt werden. Es reicht deshalb die Kennzahlen für die Wartezeit in

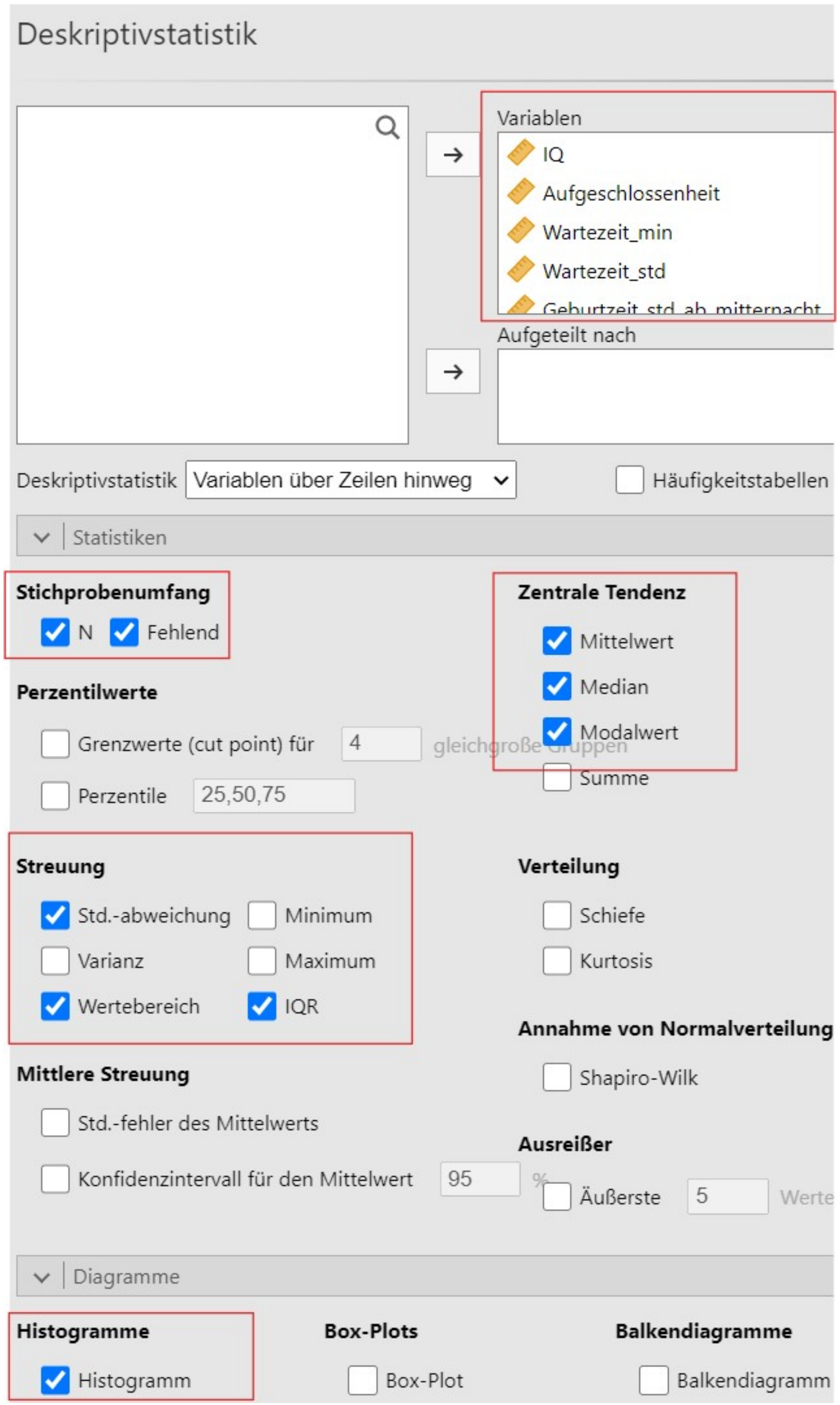


Abbildung 2.14: Jamovi Eingabe.

Deskriptivstatistik

	N	Fehlend	Mittelwert	Median	Modalwert	Std.-abw.	IQR	Wertebereich
IQ	500	0	99.6923	99.8294	53.7478 [a]	15.5131	21.2798	82.285
Aufgeschlossenheit	500	0	3.9768	4.0000	3.7000	1.0429	1.3000	6.000
Wartezeit_min	500	0	5.0735	3.3462	0.0148 [a]	5.2642	5.1358	42.356
Wartezeit_std	500	0	0.0846	0.0558	2.47e−4 [a]	0.0877	0.0856	0.706
Geburtzeit_std_ab_mitternacht	500	0	11.8111	12.1759	0.1157 [a]	6.9384	11.8936	23.758
Geburtzeit_std_ab_mittag	500	0	11.6191	11.5215	0.0310 [a]	6.8985	11.9415	23.953
Punkte	500	0	18.0860	18.0000	19.0000	1.2845	2.0000	6.000

[a] Es gibt mehr als einen Modalwert, nur der erste wird berichtet

Abbildung 2.15: Deskriptive Statistiken.

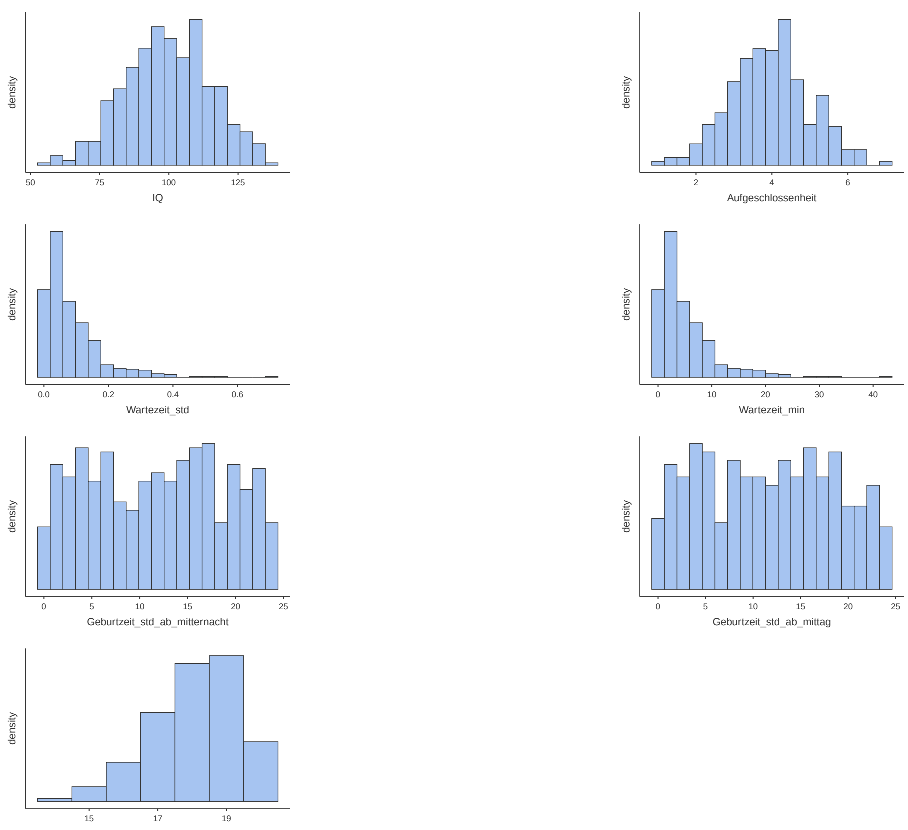


Abbildung 2.16: Histogramme.

Minuten zu betrachten. Die durchschnittliche Wartezeit liegt bei $M = 5.07$ Minuten, $Mdn = 3.35$. Der Modalwert ist wiederum nicht interpretierbar aus demselben Grund wie oben. Der Median bedeutet, dass 50% der Wartezeiten kleiner und 50% der Wartezeiten grösser waren als 3.35 Minuten. Das arithmetische Mittel ist höher als der Median. Die einigen wenigen Beobachtungen mit sehr langen Wartezeiten haben also das arithmetische Mittel im Vergleich zum Median stärker beeinflusst.

d. TODO.
e. TODO: Zentraler Wert hier nicht identifiziertbar, Streuung auch nicht.

2.4 Test

Übung 2.7.

Welche der folgenden Merkmalen sind mindestens intervallskaliert?

a) Verkaufspreise einer Kunstauktion.
b) Eine Person stimmt ja, nein oder enthält sich bei einer Abstimmung.
c) Beobachtungen des Intelligenzquotienten.
d) Reaktionszeit.

Lösung.

a) Ja
b) Nein
c) Ja
d) Ja

Übung 2.8.

Welche der folgenden Aussagen sind wahr, welche falsch?

a) Der Median ist immer kleiner als das arithmetische Mittel.
b) Das arithmetische Mittel ist anfälliger für Messfehler als der Median.
c) Die Balkenhöhe eines Histogramms steht für die Anzahl Beobachtungen.
d) Bei einem Histogramm ist steht das beobachtete Merkmal auf der x-Achse.

Lösung.

a) Falsch
b) Richtig, siehe Übung **??**
c) Richtig
d) Richtig

Übung 2.9.

Von einem intervallskalierten Merkmal wurden folgende fünf Beobachtungen gemacht: $12, 23, 15, 12, 7$. Welche der folgenden Aussagen sind wahr, welche falsch?

a) Der Median liegt bei 15.
b) Der Modus ist 12.
c) Das arithmetische Mittel ist kleiner als der Median.
d) $\sum_{i=1}^{n} x_i$ entspricht der Summe der Beobachtungen, also 69.

Lösung.

a) Falsch
b) Richtig
c) Falsch
d) Richtig

Übung 2.10.

Es wird beobachtet wie viele Autos ein Haushalt hat. Die Daten sind in `02-exr-autos-haushalt.sav` abgelegt. Welche der folgenden Aussagen sind wahr, welche falsch?

a) Die durchschnittliche Anzahl Autos pro Haushalt liegt bei $M = 0.87$.
b) Der Modus liegt bei 1.
c) Der Median liegt bei $M = 1$.
d) Es wurden $N = 92$ Personen beobachtet.

Lösung.

a) Richtig
b) Falsch, siehe `Modalwert`.
c) Falsch, richtig wäre $Mdn = 1$.
d) Falsch, es wurden Haushalte beobachtet nicht Personen.

Übung 2.11.

Von einem intervallskalierten Merkmal wurden folgende fünf Beobachtungen gemacht: $12, 23, 15, 12, 7$. Welche der folgenden Aussagen sind wahr, welche falsch?

a) $SD = 5.89$.
b) $R = 5$.
c) $IQR = 3$.
d) $s = 5.89$.

Lösung.

a) Richtig
b) Falsch
c) Richtig
d) Richtig

Übung 2.12.

Welche der folgenden Aussagen sind wahr, welche falsch?

a) Die Spannweite ist immer grösser als der Interquartilabstand.
b) 50% der Beobachtungen sind auf einer Distanz ausgebreitet, welche dem Interquartilabstand entspricht.
c) Der Interquartilabstand entspricht der Datenstreuung von der kleinsten Beobachtung bis zum Median.
d) Die Standardabweichung wird durch Messfehler weniger beeinflusst als der Interquartilabstand.

Lösung.

a) Richtig
b) Richtig
c) Falsch
d) Falsch

Kapitel 3

Stichprobenziehung

Beispiel 3.1 (Angst)**.** Forschende haben das Messinstrument State-Trait Anxiety Inentory *STAI* entwickelt, welches Angst misst (**?**). Sie unterscheiden dabei zwischen Zustandesangst und dem Persönlichkeitszug Ängslichkeit. Hier interessiert uns nur die Zustandesangst, welche fortan Angst genannt wird und misst wie grosse Angst aktuell empfunden wird. Die so gemessene Angst enstpricht einem Wert zwischen 20 und 80. A priori haben die Forschenden keine Ahnung, wie viel Angst eine Person im Durchschnitt hat und ob die ganze Skala der Werte genutzt wird. Die Forschenden machen deshalb eine kleine Befragung mit $n = 30$ zufällig ausgewählten Studierenden. Die Forschenden finden die zusammenfassenden Werte $M = 43.34, s = 9.72, n = 30$ für die Angst in ihren Beobachtungen.

Zufällig ausgewählte Beobachtungen eines Merkmals werden als **Stichprobe** bezeichnet. Die Auswahl der Beobachtungen für die Stichprobe ist die **Stichprobenziehung**. Ist mit diesen Beobachtungen die Aussage beschränkt auf die Stichprobe oder kann damit auch eine Aussage zur Angst für alle Personen getroffen werden? Alle Personen, oder generell alle möglichen Beobachtungen eines Merkmals, werden als **Population** oder **Grundgesamtheit** bezeichnet. Eine Stichprobe ist für viele Analyseverfahren repräsentativ für eine Population, wenn sie zufällig aus dieser Population gezogen. Ist dies gegeben, wird die Stichprobe auch als **Zufallsstichprobe** bezeichnet.

Hinweis. Viele Studien basieren auf Testresultaten von Studierenden, weil diese nahe am Forschungsbetrieb sind und damit über Studien informiert sind oder für wenig Geld oder Bildungsanerkennung an Studien teilnehmen. Einige dieser Studien generalisieren ihre Forschungsresultate nachher auf alle Personen. Dies ist in der Regel falsch, da Studierende nicht repräsentativ für die Gesamtbevölkerung sind (Altersstruktur, Geschlechtsverteilung, Vermögen, usw.). Die Frage, wie am besten eine repräsentative Stichprobe gezogen werden kann, kann hier aus Platzgründen nicht diskutiert werden.

3.1 Was ist das Problem der Stichprobenziehung?

Es wird angenommen, dass sich alle Personen der Population in einem Zimmer befinden. In Abbildung **??** ist dieses Zimmer aus der Vogelperspektive dargestellt, wobei jeder Punkt im schwarzen Kasten einer Person der Population entspricht. Von den Personen im Zimmer, respektive die Beobachtungen in der Population, ist die Angst nicht bekannt (Punkte in grau). Aus diesem Zimmer wurden zufällig 30 Personen geholt und befragt also sichtbar gemacht, was der Zufallsstichprobe entspricht. Die Zufallsstichprobe ist gekennzeichnet durch die farbigen Punkte über dem Zimmer, oberhalb des Pfeils. Die Farben der Punkte sind jetzt bekannt und entsprechen der jeweiligen Zustandesangst der beobachteten Personen.

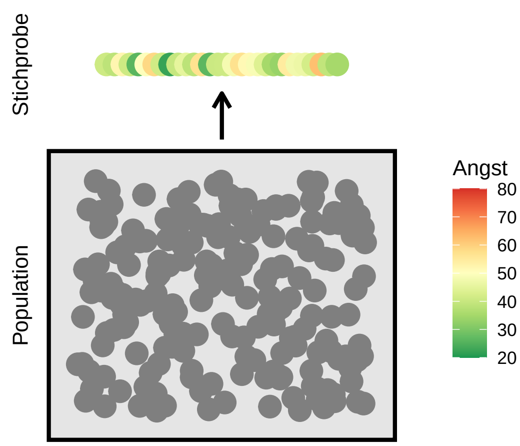


Abbildung 3.1: Population mit unbekannter Angst.

Da die Stichprobe zufällig gezogen wurde, das heisst zufällig Personen aus dem Zimmer geholt wurden, kann es nun sein, dass die Stichprobe einer Population wie in Abbildung **??** entstammt.

Es könnte aber auch sein, dass die Stichprobe einer Population mit viel höherer Zusatandsangst, wie in Abbildung **??** dargestellt, entstammt. Dies wird zwar weniger häufig vorkommen als der Fall oben, aber ist trotzdem möglich.

Das Problem der zufälligen Stichprobenziehung ist also, dass nie ganz klar ist,

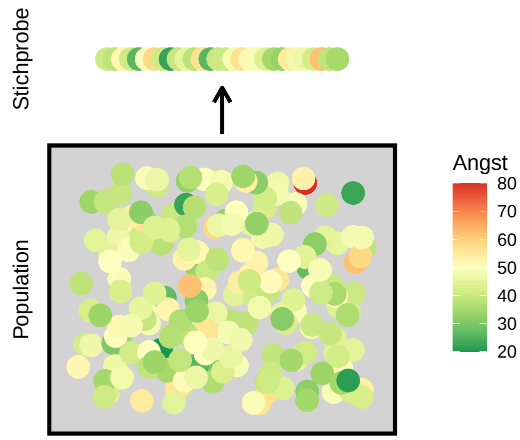


Abbildung 3.2: Population mit ähnlichen Angst-Werten wie in der Stichprobe.

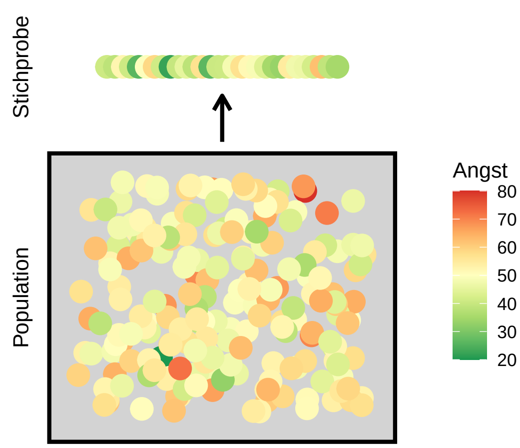


Abbildung 3.3: Population mit höheren Angst-Werten als in der Stichprobe.

wie die darunterliegende Population aussieht. Sind die Werte der Stichprobe tief, weil zufällig gerade Studierende mit tiefer Angst beobachtet wurden, oder haben tatsächlich die meisten Studierenden eine tiefe Angst?

3.2 Wie kann man Aussagen über die Grundgesamtheit machen?

Die Lösung dieses Problems funktioniert intuitiv wie folgt: Man stellt sich vor, die Stichprobenziehung würde erneut gemacht, und dann nochmal und dann nochmal. So oft, bis man einen guten Eindruck davon hat, wie häufig eine Stichprobe mit eher tiefen Angst-Werten wie bei der Stichprobe im Beispiel vorkommt. Im Szenario, in welchem in der Population tatsächlich tiefe Werte häufig vorkommen, kann dies aussehen wie in Abbildung **??**. Stichproben mit eher tiefen Angst-Werten kommen hier häufig vor.

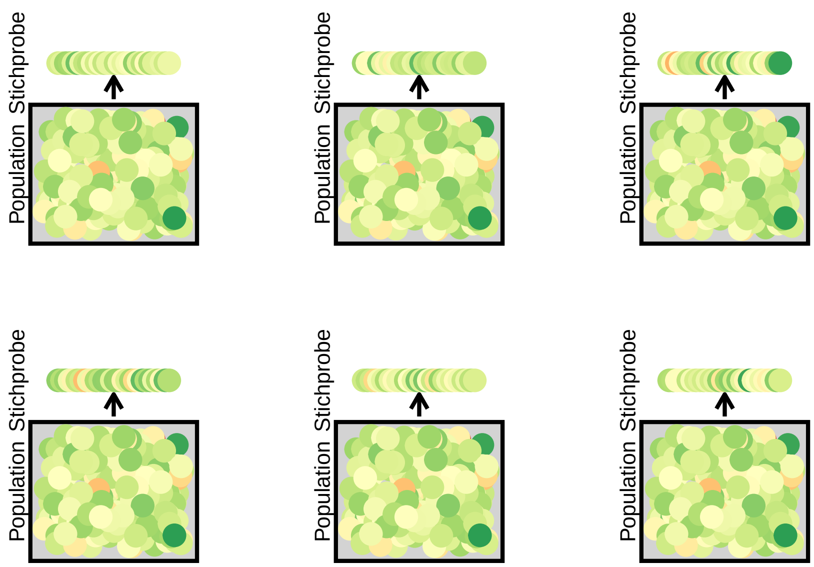


Abbildung 3.4: Wiederholte Stichprobenziehung bei gleichbleibender Population mit eher tiefen Angst-Werten.

Im Szenario, in welchem in der Population tatsächlich höhere Werte häufig vorkommen, kann dies aussehen wie in Abbildung **??**. Stichproben mit eher tiefen Angst-Werten kommen hier selten oder gar nicht vor.

Es kann also zusammenfassend gesagt werden, dass die gezogene Stichprobe wohl eher aus einer Population mit tiefen Angst-Werten gezogen wurde als aus

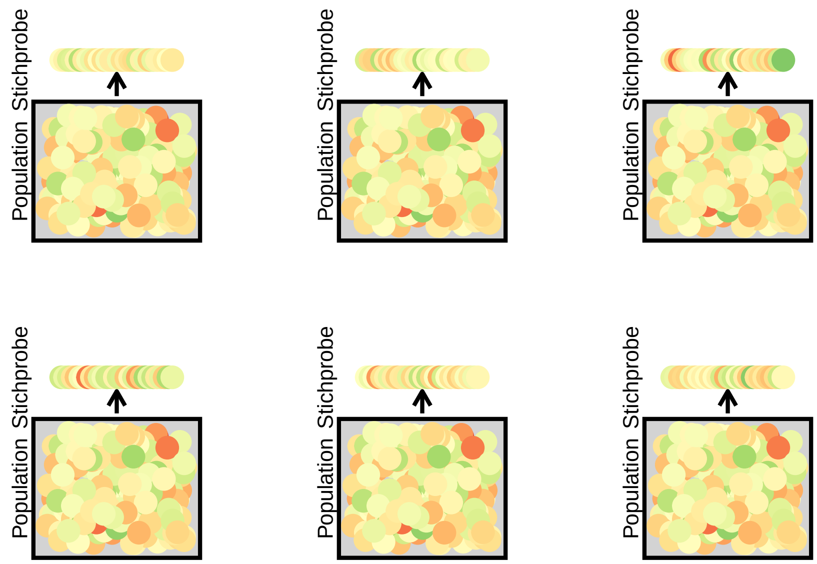


Abbildung 3.5: Wiederholte Stichprobenziehung bei gleichbleibender Population mit eher hohen Angst-Werten.

einer Population mit eher höheren Angst-Werten. Ganz sicher kann man jedoch nie sein, da die Werte in der Population eigentlich unbekannt sind. Eine genaue Quantifizierung dieser Unsicherheit kann mit Hilfe der Statistik erreicht werden und wird in den folgenden Kapiteln dieses Buches erläutert.

3.3 Übungen

Übung 3.1.

In einer Studie zum Schmerzempfinden von Personen mit dem Gen MC1R (welche meistens als rothaariger Phenotyp auftreten) werden 20 rothaarige Studierende und 54 Studierende mit andere Haarfarbe auf ihr Schmerzempfinden getestet. Identifizieren Sie die Population und die Stichprobe und erklären Sie ob es sich um eine Zufallsstichprobe handelt.

Lösung. Die Studie will eine Aussage über alle Personen mit dem Gen MC1R treffen im Vergleich zu Personen ohne das MC1R Gen. Alle Personen mit diesem Gen sind also der eine Teil der Population und alle Personen ohne das Gen sind der andere Teil der Population. Die beobachteten 20 rothaarigen Studierenden und die 54 anderen Studierenden sollen eine Stichprobe von dieser Population darstellen. Dies ist den Studienleitenden jedoch nicht gelungen, da bei der Auswahl der Studierenden auf ihre Haarfarbe geachtet wurde und nicht auf die Ausprägung des MC1R-Gens. Da auch blonde Personen das MC1R-Gen in sich tragen können und sich auch Personen die Haare rot färben können, ist hier nicht davon auszugehen, dass es sich um eine Stichprobe der erwünschten Population handelt. Wenn davon ausgegangen wird, dass die rothaarigen tatsächlich alle das erwünschte Gen in sich tragen und die anderen Teilnehmenden nicht, dann kann zusätzlich bemängelt werden, dass eine Aussage über alle Personen getroffen werden soll, sich aber nur Studierende in der Stichprobe befinden - also vorwiegend junge Personen mit wenig Geld. Mit dieser Stichprobe kann dann also eigentlich nur eine Aussage zu allen jungen Rothaarigen und jungen Leuten mit anderer Haarfarbe getroffen werden. Zudem muss davon ausgangen werden, dass die Studierenden aus folgenden möglichen Gründen nicht zufällig ausgewählt wurden:

- Teilnehmende brauchen das Geld oder Punkte für ihr Studium
- Teilnehmende tendieren dazu zusammen mit Freunden an Studien teilzunehmen
- Teilnhemende müssen Zeit haben, sind also tendenziell weniger durch Erwerbstätigkeit oder Care-Arbeit blockiert als andere
- etc.

Übung 3.2.

Anna und Isabel wollen mit einmaligem Armdrücken herauszufinden, wer die stärkere Person ist. Identifizieren Sie Population und Stichprobe. Isabel gewinnt. Können Anna und Isabel nachher mit Gewissheit sagen, dass Isabel stärker ist? Wie könnte mehr Sicherheit erlangt werden?

Lösung. Die Population beinhaltet in diesem Fall alle hypothetischen je gemachten oder noch durchzuführende Armdrücken zwischen Anna und Isabel. Die Stichprobe ist das jetzt durchgeführte Armdrücken, bei welchem Isabel gewonnen hat. Anna könnte zu recht behaupten, dass nur weil Isabel einmal gewonnen hat, dies relativ wenig aussagekräftig ist für die Population also für die Frage, ob Isabel stärker ist. Vielleicht hatte Isabel einfach einen sehr starken Tag und Anna ist aber normalerweise eigentlich stärker. Um diese zufälligen Effekte zu minimieren, könnten Anna und Isabel das Armdrücken regelmässig wiederholen.

3.4 Test

Übung 3.3.

Eine Aktiengesellschaft hat die Addressen aller ihrer 10000 Aktionär:innen. Mit einer Umfrage soll herausgefunden werden, ob die Aktionär:innen den neuen umweltfreundlichen Unternehmenskurs begrüssen. Dazu werden per losverfahren 100 Aktionär:innen bestimmt und an der Generalversammlung kurz um eine Stellungnahme gebeten. Welche der folgenden Aussagen sind wahr, welche falsch?

a) Die Stichprobengrösse ist 10000.
b) Die Grundgesamtheit sind alle Aktionär:innen.
c) Es handelt sich um eine Zufallsstichprobe.
d) Die Population umfasst 100 Aktionär:innen.

Lösung.

a) Falsch
b) Richtig
c) Richtig
d) Falsch

Kapitel 4

Durchschnitt und Standardabweichung schätzen

Wie die in Abschnitt **??** skizzierte Lösung für das Problem der zufälligen Stichprobe konkret umgesetzt wird, hängt von der Problemstellung ab. Im Folgenden wird ein Verfahren zur Generalisierung der Schätzung der zentralen Tendenz basierend auf einer Stichprobe präsentiert.

4.1 Wo liegt der Durchschnitt der Grundgesamtheit?

Ein Parameter, über welchen wir gerne eine Aussage treffen würden, ist die zentrale Tendenz in der Grundgesamtheit. Diese wird **Erwartungswert** (Symbol μ [gr.: mü]) genannt. Wenn das arithmetische Mittel der Stichprobe berechnet wird, ergibt dies auch ein Schätzwert für besagten Erwartungswert. Aufgrund der zufälligen Stichprobenziehung ist jedoch auch klar, dass dieser Schätzwert nie genau dem wahren Erwartungswert entspricht.

In Beispiel **??** liegt das arithmetische Mittel in der Stichprobe der Studierenden bei $M = 43.2$. Dieser Wert entspricht nun auch der Schätzung des Erwartungswertes, also der geschätzten durchschnittlichen Angst aller Menschen. Die Folgefrage ist also wie genau unsere Schätzung ist. Um dies zu quantifizieren, wiederholen wir die Stichprobenziehung und berechnen das arithmetische Mittel dieser zweiten Stichprobe. Dann wiederholen wir diesen Prozess, zum Beispiel 1000 mal.

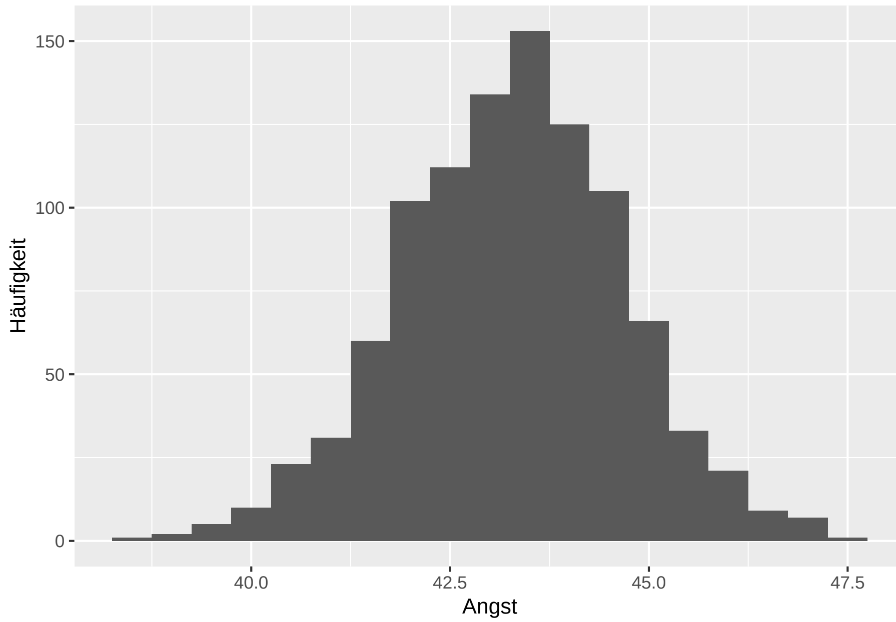


Abbildung 4.1: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

Die Häufigkeitsverteilung der berechneten arithmetischen Mittel in Abbildung **??** lässt nun eine Aussage über die Häufigkeit und damit über die Wahrscheinlichkeit von gewissen Werten als Erwartungswert zu. Ein Durchschnittswert der Zustandesangst um die 43 ist hier am wahrscheinlichsten und ein Wert tiefer als 41 oder höher 45 eher selten. Um diese Aussage präziser zu gestalten, werden konventionell die 95% häufigsten Werte (die höchsten Balken im Histogramm) als wahrscheinlich betrachtet. Die 5% verbleibenden Werte, verteilt auf das untere und obere Extrem, werden als unwahrscheinlich betrachtet. Das 2.5% Perzentil trennt die 2.5% tiefsten arithmetischen Mittel ab und liegt im Beispiel bei 40.4. Das 97.5%-Perzentil trennt die höchsten 2.5% (oder eben die tiefsten 97.5%) arithmetischen Mittel ab und liegt bei 46. Dies ist in Abbildung **??** ersichtlich.

Beispiel 4.1 (Verträglichkeit)**.** Einer der Big-5 Persönlichkeitszüge ist die Verträglichkeit. Eine einfache Art die Big-5 zu messen ist mit den 10 Fragen aus dem ten-item personality inventory *TIPI* (**?**). Für die Verträglichkeit müssen zwei Items (Item 1: Critical, quarrelsome; Item 2: Sympathetic, warm) auf einer Likert-Skala von 1 bis 7 eingeordnet werden. Anschliessend werden die Antworten gemittelt. Ein Student möchte herausfinden, ob mit diesem Messinstrument die durchschnittliche Verträglichkeit aller Menschen mittig also bei 4 liegt. Dafür befragt er $n = 100$ Personen und findet die Werte $M = 3.91, s = 1.73$.

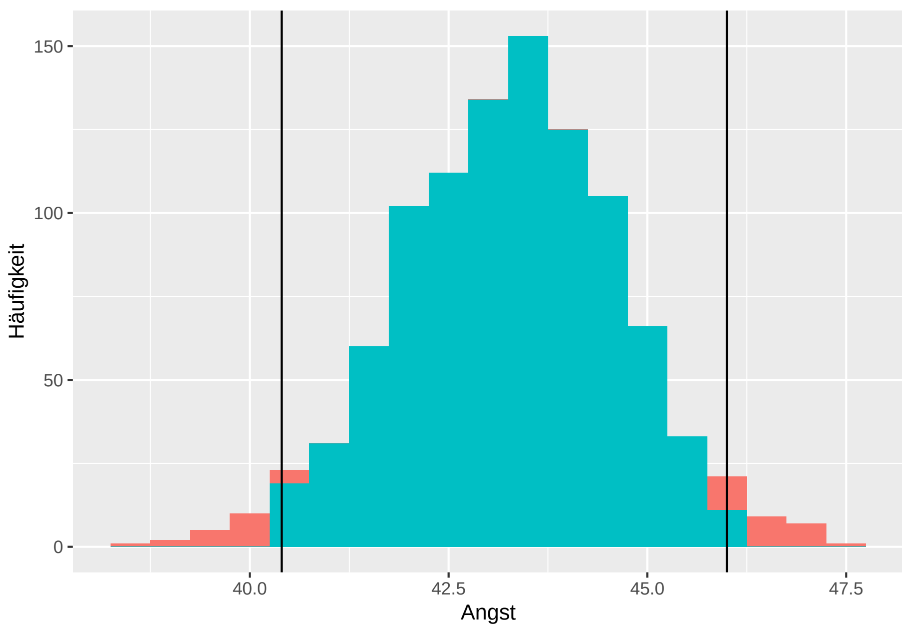


Abbildung 4.2: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

Die Verteilung der Beobachtungen, siehe Abbildung **??**, zeigt, dass alle Werte zwischen 1 und 7 vorkommen, aber keine zentrale Tendenz greifbar ist. Um herauszufinden wie zutreffend die Schätzung des Erwartungswertes der Verträglichkeit von $M = 3.91$ ist, stelle man sich wieder vor, dass der Student 1000-mal die Stichprobenziehung wiederholt und jedes Mal das arithmetische Mittel M von neuem berechnet. Die Verteilung der arithmetischen Mittel dieser Stichproben ist in Abbildung **??** dargestellt. Bei dieser Verteilung kann erneut links und rechts 2.5% der Werte abgeschnitten werden, um zum Schluss zu gelangen, dass das arithmetische Mittel in 95% der Fälle zwischen 3.7 und 4.3 zu liegen kommt.

Das Problem mit diesem Vorgehen ist, dass es aus finanziellen oder technischen Gründen selten möglich ist mehrere Stichproben aus derselben Population zu ziehen. Glücklicherweise haben Statistiker:innen herausgefunden, dass die Häufigkeitsverteilungen wie in Abbildungen **??** und **??** immer dieselbe Verteilung haben und dies unabhängig davon wie die ursprüngliche Verteilung des Merkmals aussah. Diese Verteilung ist eine sogenannte **Normalverteilung**.

Die Normalverteilung sieht eine Glocke ähnlich. Deshalb wird sie auch Gausssche Glockenkurve nach Carl F. Gauss (1777-1855) benannt. Die Normalverteilung kann mit nur zwei Parametern beschrieben werden.

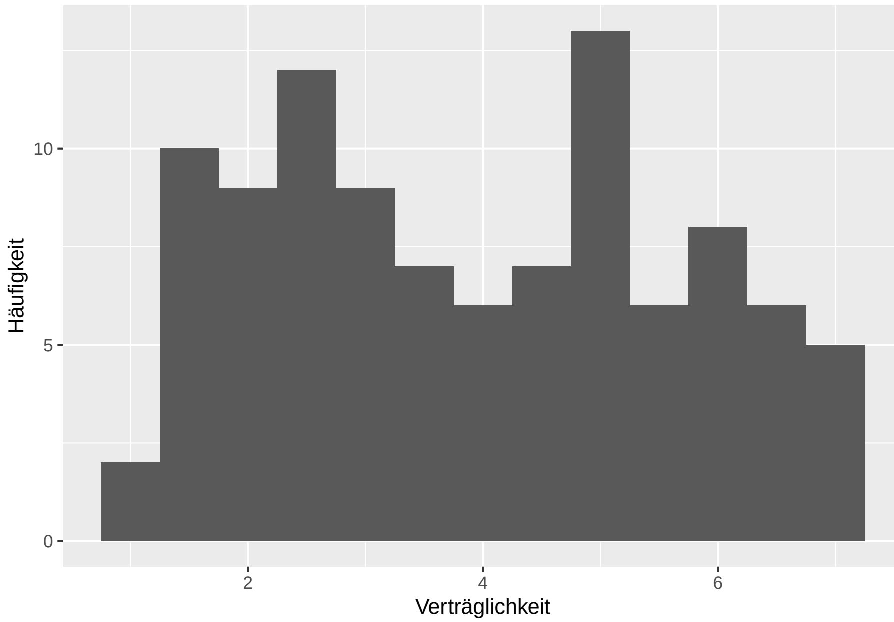


Abbildung 4.3: Verteilung der 100 beobachteten Verträglichkeitswerte einer zufällig gezogenen Stichprobe.

- μ_g gibt an, wo auf der x-Achse der höchste Punkt der Glocke liegt
- σ_g gibt an, wie flach die Glockenform ist (ein grosser Wert entspricht einer flachen Glockenform, ein tiefer Wert einer steilen Glockenform).

Auf seeing-theory.brown.edu > Continuous > Normal kann der Einfluss von μ_g und σ_g auf die Normalverteilung erfahren werden.

Diese Tatsache, dass die Durchschnitte aller Merkmale normalverteilt sind, ist so zentral für die Statistik, dass sie **Zentraler Grenzwertsatz** genannt wurde. Der zentrale Grenzwertsatz besagt genauer, dass bei einem Merkmal mit Erwartungswert μ und Standardabweichung σ, der Durchschnitt aller Stichprobenwerte einer Normalverteilung mit $\mu_g = \mu$ und $\sigma_g = \frac{\sigma}{\sqrt{n}}$ entspricht, wobei n die Stichprobengrösse und σ die Standardabweichung des Merkmals in der Population bezeichnet.

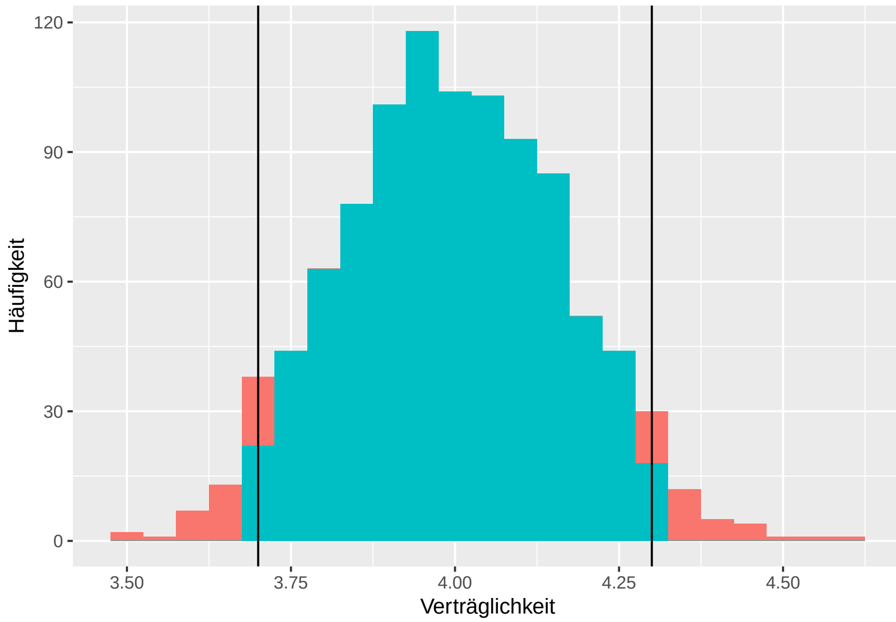


Abbildung 4.4: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Verträglichkeit.

Achtung

Hinweis.

- $\mu_g = \mu$ bedeutet, dass der Wert, welcher unter der Normalverteilung am wahrscheinlichsten ist, genau dem Erwartungswert des untersuchten Merkmales entspricht.
- $\sigma_g = \frac{\sigma}{\sqrt{n}}$ hat zwei Implikationen:
  - je grösser die Streuung des Merkmals (grosses σ) desto breiter ist auch die Streuung der arithmetischen Mittel (grosses σ_g). Dies bedeutet, je weniger Streuung das Merkmal aufweist, desto genauer ist die Bestimmung des Erwartungswertes des Merkmales.
  - je grösser die Anzahl Beobachtungen n, desto kleiner die Streuung der arithmetischen Mittel (kleines σ_g). Dies bedeutet, je grösser die Stichprobe ist, desto genauer ist die Bestimmung des Erwartungswertes des Merkmales.

Die Abbildungen **??** und **??** illustrieren den zentralen Grenzwertsatz für Beispiel

Tabelle 4.1: Vergleich Perzentile der Stichprobe und der theoretischen Verteilung.

	Stichprobe		Normalverteilung		t-Verteilung	
Beispiel	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Angst	40.4	46.0	42.23	44.46	42.18	44.50
Vertraeglichkeit	3.7	4.3	3.66	4.34	3.66	4.34

?? und **??** respektive, wobei die Normalverteilung der roten Linie entspricht. Dabei wird einstweilen angenommen, dass μ und σ bekannt sind. Diese Annahme wird später aufgelöst und dient hier lediglich der Illustration.

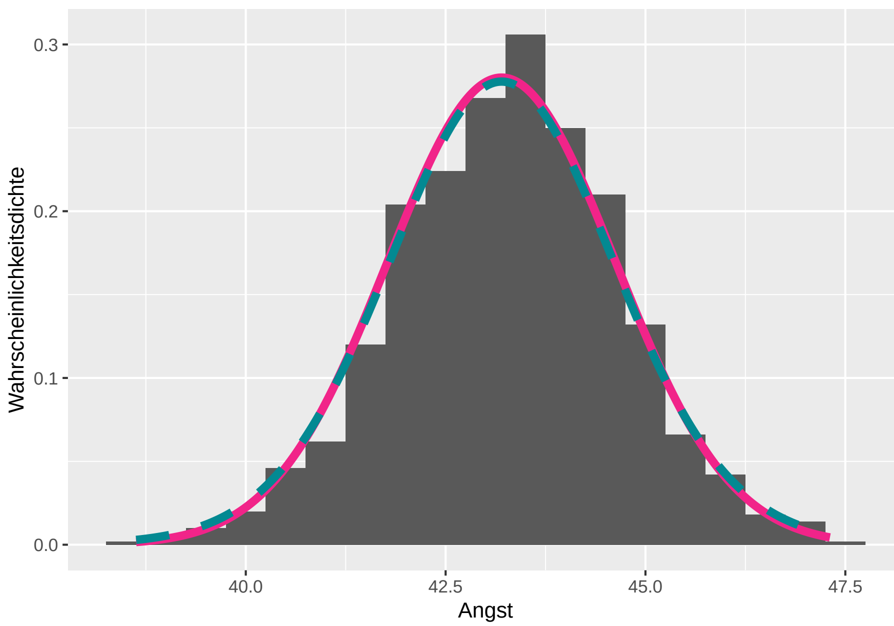


Abbildung 4.5: Die arithmetischen Mittel sind Normalverteilt mit Parametern $\mu_g = 43.34$ und $\sigma_g = 9.72/\sqrt{30}$.

Die Erkenntnis des zentralen Grenzwertsatz macht also das wiederholte Ziehen von Stichproben unnötig. Die Normalverteilung ist theoretisch konstruiert und ihr 2.5%- und 97.5%-Perzentil können theoretisch hergeleitet werden. Tabelle **??** wird kann beobachtet werden, dass für unsere zwei Beispiele die Perzentile der Stichprobe und der Normalverteilung sehr ähnlich, wenn auch nicht exakt gleich sind. Die Ungenauigkeit rührt daher, dass der zentrale Grenzwertsatz nur dann exakt funktioniert, wenn die Anzahl Beobachtungen (unendlich) gross ist.

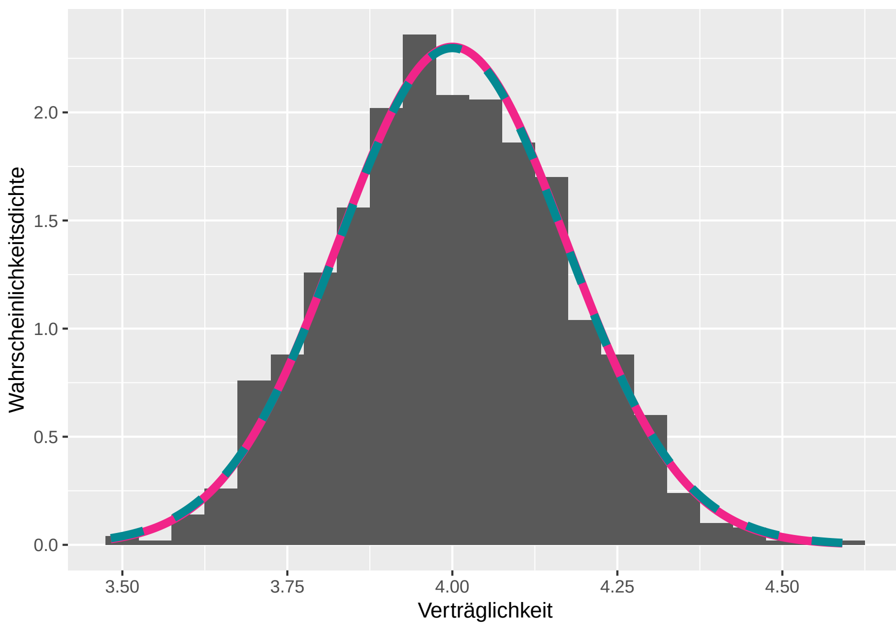


Abbildung 4.6: Die arithmetischen Mittel sind Normalverteilt mit Parametern $\mu_g = 3.91$ und $\sigma_g = 1.73/\sqrt{100}$.

Einstweilen wurde hier angenommen, dass die Streuung des Merkmals σ bekannt ist. Dies ist in der Realität nie der Fall und eine weitere, wenn auch weniger grosse Ungenauigkeitsquelle. Wenn σ also auch aus der Stichprobe geschätzt werden muss, ist die Annäherung der Verteilung der arithmetischen Mittel besser gegeben mit einer **Student-t-Verteilung** oder kurz t-Verteilung. Die grüne gestrichelte Linie in den Abbildungen **??** und **??** entspricht der t-Verteilung im jeweiligen Beispiel.

Der Unterschied zwischen der Normalverteilung und der t-Verteilung ist nur sichtbar, wenn n klein ist. In Beispiel **??** mit $n = 30$ ist ein kleiner Unterschied, in Beispiel **??** mit $n = 100$ ist kein Unterschied zwischen der Normalverteilung und der t-Verteilung sichtbar. Tatsächlich wird die t-Verteilung mit einem Parameter charakterisiert, welcher **Freiheitsgrade** (eng. degrees of freedom, df) genannt wird. In Abbildung **??** wird die t-Verteilung mit verschiedenen Freiheitsgraden mit der Normalverteilung verglichen. Bei der t-Verteilung mit den kleinsten Freiheitsgraden sind extremere Werte wahrscheinlicher als t-Verteilungen mit grösseren Freiheitsgraden.

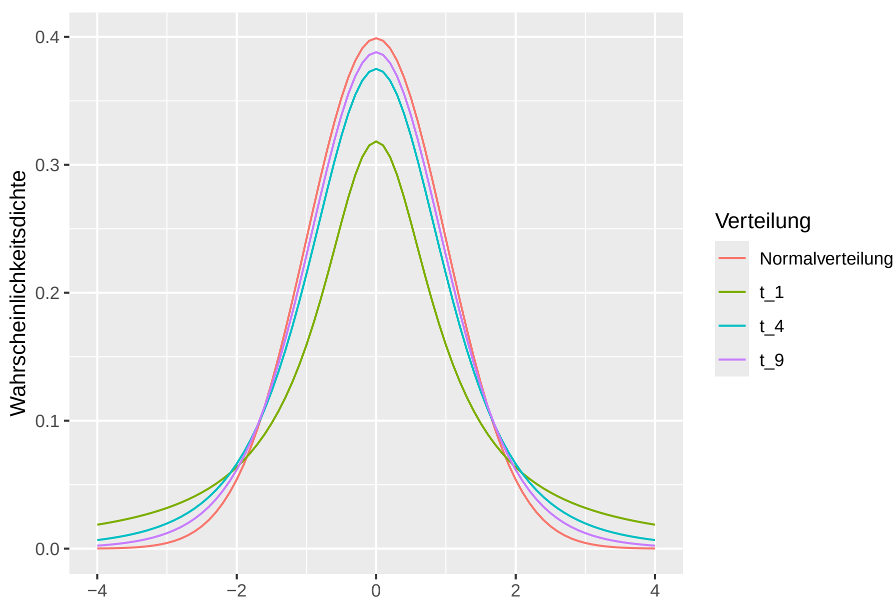


Abbildung 4.7: Student-t-Verteilungen mit 1, 4 und 9 Freiheitsgraden im Vergleich zu der Normalverteilung.

Die Freiheitsgrade der t-Verteilung in der Annäherung oben entsprechen der Anzahl Beobachtungen minus 1, also $df = n-1$. Die höhere Wahrscheinlichkeit von extremeren Werten bei kleinen Freiheitsgraden spiegelt die grössere Unsicherheit der Schätzung des Erwartungswertes wider, wenn die Standardabweichung unbekannt und damit auch geschätzt werden muss. Je kleiner n ist, desto stärker fällt diese Unsicherheit aus.

Die arithmetischen Mittel bei unbekannter Standardabweichung sind bei wiederholter Stichprobenziehung genau t-verteilt. Um die Genauigkeit der Schätzung des Erwartungswertes zu bestimmen, genügt es folglich, das 2.5% und das 97.5% Perzentil der t-Verteilung mit $n-1$ Freiheitsgraden zu bestimmen. Diese Perzentile können mit

$$\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{97.5\%,n-1} < \mu < \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{97.5\%,n-1} \tag{4.1}$$

berechnet werden, wobei $\bar{x}$ das arithmetische Mittel, s die Standardabweichung und $t_{97.5\%,n-1}$ dem Wert des 97.5%-Perzentil einer auf 0 zentrierten t-Verteilung mit $n-1$ Freiheitsgraden entspricht. Letzere Perzentile der t-Verteilung können bei Bedarf in entsprechenden Tabellen nachgeschlagen werden. Als Gedankenstütze kann für $t_{97.5\%,n-1}$ immer 2 gedacht werden, da dies ungefähr dem wahren Wert entspricht, wenn n grösser als 50 ist.

Das 2.5% und das 97.5% Perzentil der Verteilung der arithmetischen Mittel ergeben nun die untere respektive obere Schranke eines **Intervalls**. Ein Intervall bezeichnet durch die Symbolik [untere Schranke, obere Schranke] beinhaltet alle Zahlen zwischen der unteren und der oberen Schranke. Ein Intervall mit den oben beschriebenen Perzentilen als Schranken wurde so berechnet, dass bei wiederholter Stichprobenziehung der wahre Erwartungswert in 95% der Fälle umschlossen wird. Grob übersetzt bedeutet dies, dass wir zu 95% sicher oder *konfident* sind, dass der Erwarungswert in diesem Intervall liegt. Dieses Intervall wird deshalb als 95%-**Konfidenzintervall** (Symbol KI) bezeichnet. Als Sicherheit wird konventionell oft 95% gewählt, andere Sicherheitswerte sind aber ebenfalls möglich und sinnvoll. Diese Werte heissen **Vertrauenswahrscheinlichkeit**. Ein 95%-Konfidenzintervall ist also ein Konfidenzintervall mit 95% Vertrauenswahrscheinlichkeit. Andersherum betrachtet kann auch festgestellt werden, dass bei einem 95%-Konfidenzintervall die Wahrscheinlichkeit sich zu irren bei 5% liegt. Irren bedeutet hier, dass der wahre Erwartungswert bei wiederholter Stichprobenziehung von 5% der Konfidenzintervallen nicht überdeckt wird. Dieser Wert wird demnach **Irrtumswahrscheinlichkeit** genannt und mit α bezeichnet. Es ist demnach äquivalent von einem 99% Konfidenzintervall oder von einem Konfidenzintervall mit 1% Irrtumswahrscheinlichkeit zu sprechen.

In Beispiel **??**, kann aus der Tabelle **??** entnommen werden, dass die Angst in der Population bei $M = 43.34$ 95% KI $[42.18, 44.5]$ liegt. In Beispiel **??**, kann aus der Tabelle **??** entnommen werden, dass die Verträglichkeit in der Population bei $M = 3.91$ 95% KI $[3.66, 4.34]$ liegt. Wann immer eine Schätzung eines zentralen Wertes berichtet wird, soll dies ab jetzt in der soeben gezeigten Darstellung inklusive Angabe des Konfidenzintervalls erfolgen. Damit wird der Leserin aufgezeigt, wo der Schätzwert der zentralen Tendenz liegt und gleichzeitig wird intuitiv vermittelt, wie genau die Schätzung ist.

Es ist nun spannend zu explorieren, wie sich die Stichprobengrösse n oder die geschätzte Standardabweichung s auf die Länge des Konfidenzintervalls auswirkt. Dies kann in den Übungen **??** und **??** selbst erforscht werden.

4.2 Übungen

Übung 4.1.

Die Firma Pear bringt ein neues Smartphone das F42 der Reihe Supernova X auf den Markt. Das Smartphone ist für Jugendliche im Alter von $15-20$ Jahre konzipiert. Um herauszufinden, welcher Marktpreis für das F42 verlangt werden kann, erfragt Pear bei 70 Jugendlichen die Zahlbereitschaft. Die Daten stehen unter `04-exr-marktpreisanalyse.sav` zur Verfügung. Wie gross ist die durchschnittliche Zahlbereitschaft der Jugendlichen? Berichten Sie die Ergebnisse der Marktanalyse mit einem 95%-Konfidenzintervall.

Lösung. Der Datensatz wird bei `Jamovi` eingelesen und die Analyseparameter wie in Abbildung **??** gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung **??**.

Die Marktanalyse mit $N = 70$ Befragten hat ergeben, dass Jugendliche im Alter von $15-20$ Jahren bereit sind durchschnittlich $M = 288.34$ CHF 95%-KI $[260.92, 315.76]$ auszugeben für das neue Supernova X F42 von Pear.

Übung 4.2.

In einer Studie werden 421 Probandinnen über eine Woche lang beauftragt immer wieder fremde Personen anzusprechen. Dabei wird unter anderem am Anfang und am Ende der Woche gemessen, wie unangenehm auf einer Skala von 1 bis 5 dies für die Probandinnen ist. Die Daten stehen unter `04-exr-stranger.sav` zur Verfügung. Verwenden Sie für drei Nachkommastellen in `Jamovi` für die folgenden Teilaufgaben.

(a) Berechnen Sie das 95%-Konfidenzintervall für die durchschnittliche Unangenehmheit in der Grundgesamtheit für die Situation am Anfang und am Ende der Studie und berichten und interpretieren Sie das Resultat. Denken Sie die Intervention hat die Unangenehmheit, welche durch das Ansprechen von Fremden entsteht, in der Grundgesamtheit durchschnittlich gesenkt?
(b) Vergleichen Sie die Längen der errechneten Konfidenzintervalle.
(c) Wiederholen Sie die Aufgabe und berechnen Sie jetzt das 90% und das 99%-Konfidenzintervall. Wie verhält sich die Länge des Konfidenzintervalls bei unterschiedlichen Vertrauenswahrscheinlichkeiten?

Diese Aufgabe ist angelehnt an **?**.

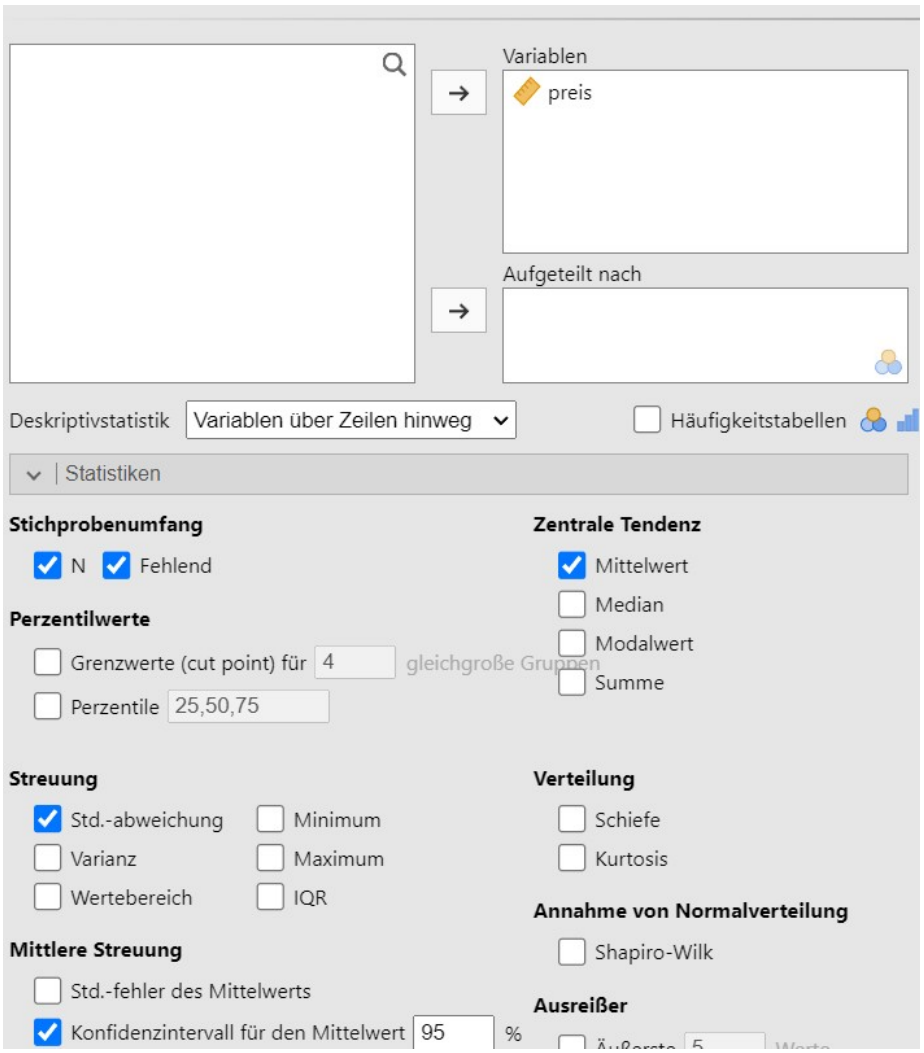


Abbildung 4.8: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

				95% Konfidenzintervall		
	N	Fehlend	Mittelwert	Untere	Obere	Std.-abw.
preis	70	0	288.338	260.915	315.761	115.009

Anmerkung. Das Konfidenzintervall des Mittelwerts nimmt an, dass die Mittelwerte einer t-Verteilung mit N - 1 Freiheitsgraden folgen

Abbildung 4.9: Jamovi Ausgabe.

Lösung. Der Datensatz wird bei `Jamovi` eingelesen und die Analyseparameter wie in Abbildung **??** gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung **??**.

(a) Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei $M = 3.784, 95\%$ KI $[3.764, 3.803]$ Punkten und am Ende der Woche bei $M = 3.328, 95\%$ KI $[3.309, 3.348]$ Punkten. Wenn die Studie 100 mal wiederholt wird und jedes Mal ein 95% Konfidenzintervall für den Erwartungswert der Unangenehmheit berechnet wird, so wird der tatsächliche Erwartungswert in 95% der Fälle also ungefähr 95 mal vom Konfidenzintervall überdeckt. Da die Konfidenzintervalle weit auseinander liegen, kann davon ausgegangen werden, dass die Unangenehmheit durchschnittlich tatsächlich nach dem Versuch tiefer liegt als vor dem Versuch. Die Unangenehmheit kann also durch Training vermindert werden.

(b) Die Länge der Konfidenzintervalle betragen am Anfang der Woche $3.803 - 3.764 = 0.039$ und am Ende der Woche $3.348 - 3.309 = 0.039$.

(c) Um das 90% Konfidenzintervall zu berechnen kann in der in Abbildung **??** dargestellten Maske der Wert für `Konfidenzintervall für den Mittelwert` auf 90 gesetzt werden. Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei $M = 3.784, 90\%$ KI $[3.767, 3.800]$ Punkten und am Ende der Woche bei $M = 3.328, 90\%$ KI $[3.312, 3.345]$ Punkten. Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei $M = 3.784, 99\%$ KI $[3.758, 3.809]$ Punkten und am Ende der Woche bei $M = 3.328, 99\%$ KI $[3.303, 3.354]$ Punkten. Die Länge der 90% Konfidenzintervalle ist $3.800 - 3.767 = 0.033$ und $3.345 - 3.312 = 0.033$. Die Länge der 99% Konfidenzintervalle ist $3.809 - 3.758 = 0.051$ und $3.354 - 3.303 = 0.051$. Es kann also hier empirisch festgestellt werden, dass das Konfidenzintervall grösser wird, je höher die Vertrauenswahrscheinlichkeit sein soll.

Übung 4.3.

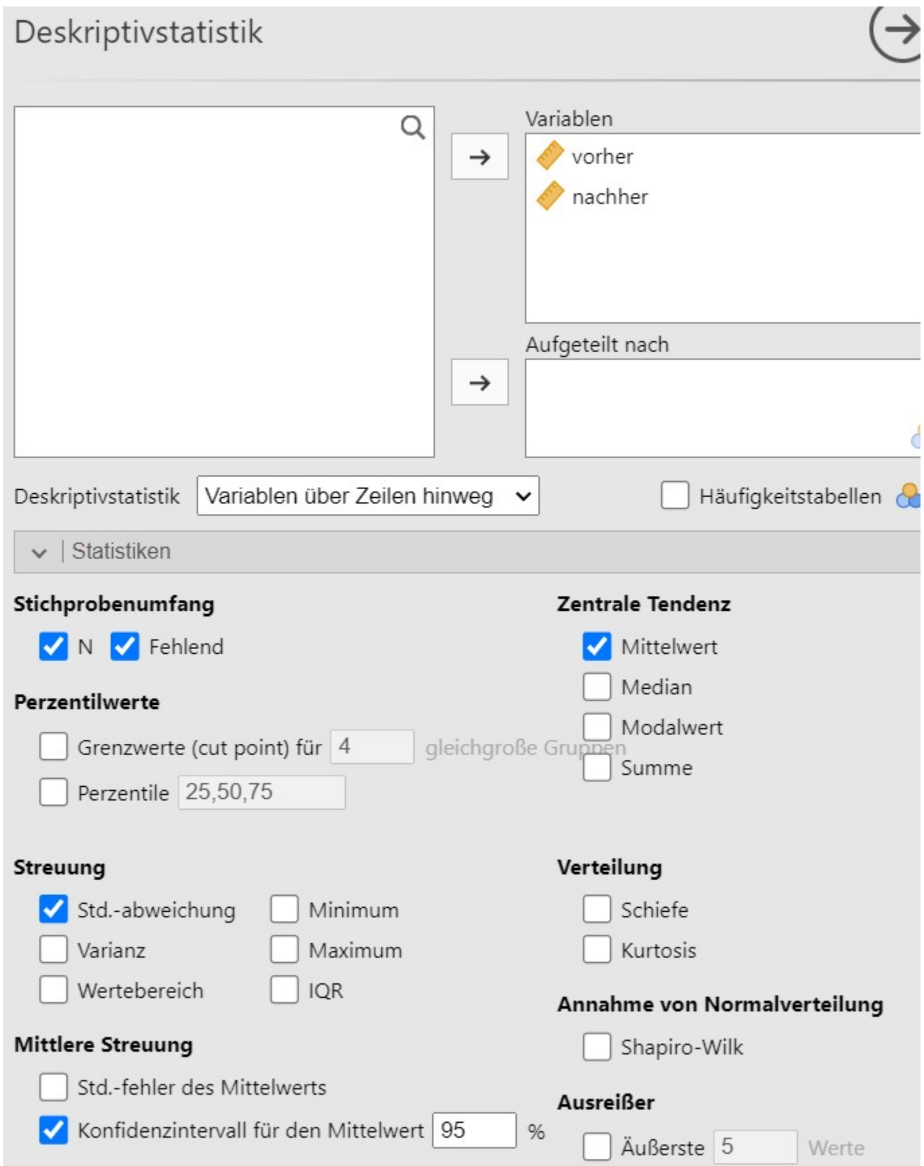


Abbildung 4.10: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

				95% Konfidenzintervall		
	N	Fehlend	Mittelwert	Untere	Obere	Std.-abw.
vorher	421	0	3.784	3.764	3.803	0.204
nachher	421	0	3.328	3.309	3.348	0.203

Anmerkung. Das Konfidenzintervall des Mittelwerts nimmt an, dass die Mittelwerte einer t-Verteilung mit N - 1 Freiheitsgraden folgen

Abbildung 4.11: Jamovi Ausgabe.

(TODO, Achtung: Diese Augabe funktioniert je nach Version von Jamovi und JJStatsPlot nicht. Wenn die Normalverteilungskurve trotz Anwählen nicht angezeigt wird, kann die Lösung der Aufgabe nachgelesen werden. Diese Aufgabe ist wichtig für das Verständnis, nicht aber zur Nachahmung an der Prüfung.) Für ein Experiment werden in drei Runden jeweils 10000 Zufallsstichproben erhoben mit respektive 10, 40 und 100 Beobachtungen pro Zufallsstichprobe. Die Verteilung der jeweils ersten Zufallsstichprobe für eine Stichprobengrösse ist in Abbildung **??** dargestellt. Die Daten sind nicht normalverteilt, weil keine Glockenkurve wie oben beschrieben das Histogramm gut abdecken würde.

Die arithmetischen Mittel der $10'000$ Stichproben sind im Datensatz `04-exr-zentraler-grenzwertsatz.sav` festgehalten. In der Spalte `n_10` zum Beispiel steht jede Zeile für das arithmetische Mittel eine Zufallsstichprobe mit 10 Beobachtungen. Der zentrale Grenzwertsatz besagt, dass diese arithmetischen Mittel normalverteilt sind mit zunehmender Stichprobengrösse n. Erstellen Sie ein Histogramm mit der Erweiterung `JJStatsPlot` und zeichnen Sie eine Normalverteilung darüber. Interpretieren Sie das Resultat.

Lösung. Das Übereinanderlegen des jeweiligen Histogramms und der Wahrscheinlichkeitsdichte der Normalverteilung wird in Abbildung **??** gezeigt. Es ist deutlich zu sehen, dass die Linie nur bei $n = 100$ die Häufigkeitsverteilung der arithmetischen Mittel gut nachbilden kann. Bei $n = 10$ und $n = 50$ ist ein grosser Unterschied zwischen Häufigkeitsverteilung und Linie sichtbar. Das genaue n ab welchem eine Häufigkeitsverteilung gut durch die Normalverteilung angenähert wird hängt von der ursprünglichen Verteilung der Daten ab, d.h. der Verteilung in Abbildung **??**. Es kann deshalb nicht generell gesagt werden, dass ab $n = 100$ die Annäherung immer gut sei, so wie in diesem Beispiel. Der zentrale Grenzwertsatz besagt demnach auch lediglich, dass man immer ein grosses n wählen kann, so dass die Annäherung gut ist. Er besagt nichts darüber, wie gross n sein muss.

Übung 4.4.

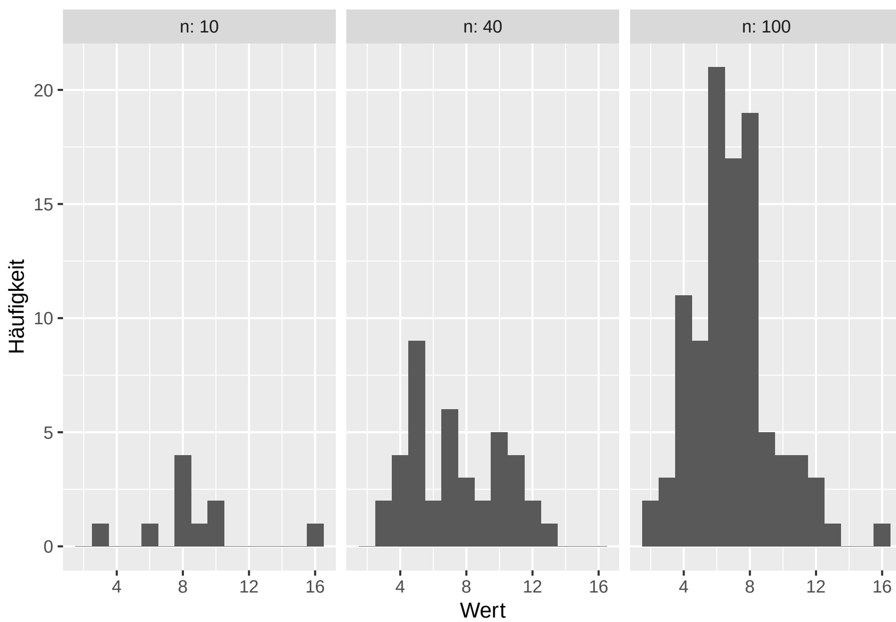


Abbildung 4.12: Verteilung der Werte einer Stichprobe. (Achtung es handelt sich hierbei nicht um Mittelwerte, sondern die Effektiv gemessenen Beobachtungen einer Stichprobe.)

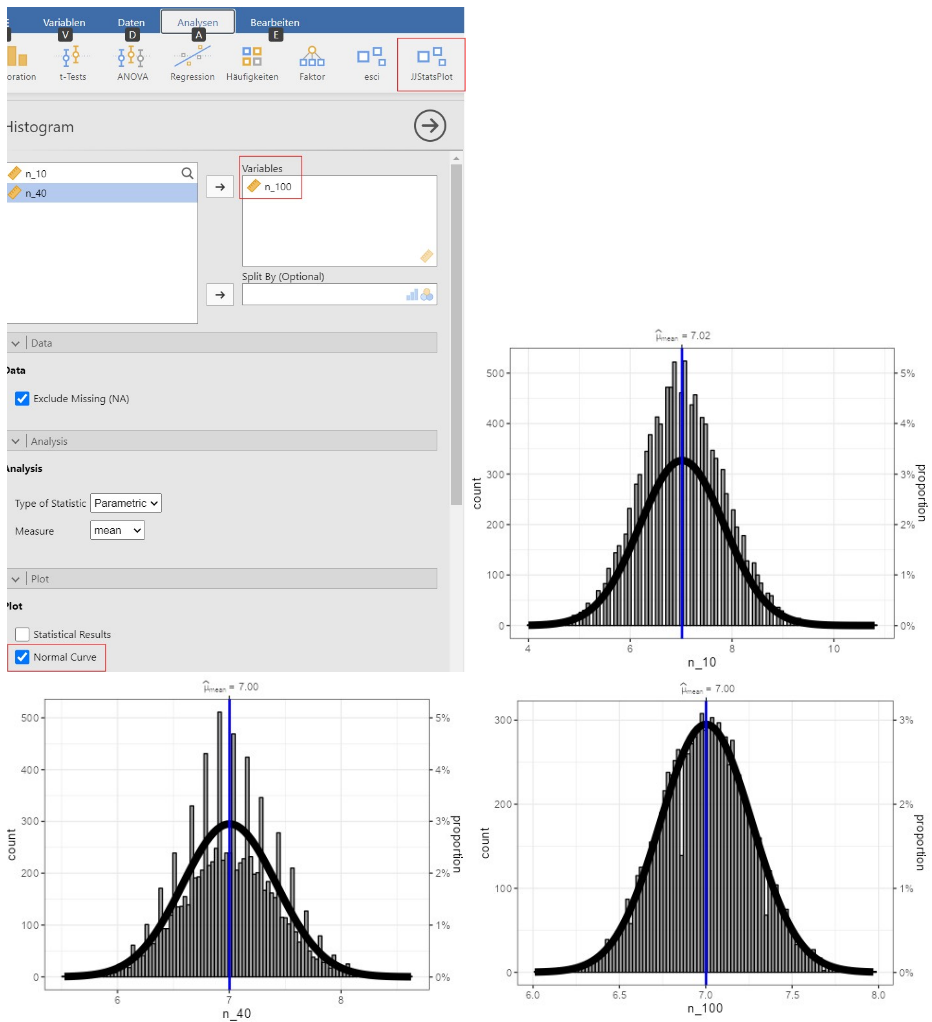


Abbildung 4.13: Jamovi-Eingabeeinstellungen und die Histogramme der Mittelwerte für die Stichprobengrössen 10, 40 und 100.

Eine Mensa will herausfinden, wie lange die Leute um 12h durchschnittlich anstehen müssen. Dazu befragt sie 5 Kund:innen. Das Resultat der Untersuchung ist, dass die Kund:innen im Durchschnitt 0.4 Stunden anstehen müssen. Leider ist das Konfidenzintervall sehr gross. Da die Mensa nicht weiss, wie viele Leute befragt werden müssen, um ein kleineres Konfidenzintervall zu erhalten befragt sie in 4 weiteren Runden jeweils 20, 50, 100 und 1000 Kund:innen. Die Daten aller 5 Untersuchungen sind unter `04-exr-stichprobengroesse.sav` abgelegt. Für jede der 5 Stichproben:

a. Was ist die Schätzung des Erwartungswertes der Wartezeit?
b. Wie gross ist die Standardabweichung der Wartezeit?
c. Wie gross ist die Standardabweichung der arithmetischen Mittel?
d. Bestimmen Sie die 95%-Konfidenzintervalle.
e. Berechnen Sie die Länge jedes Konfidenzintervalls.

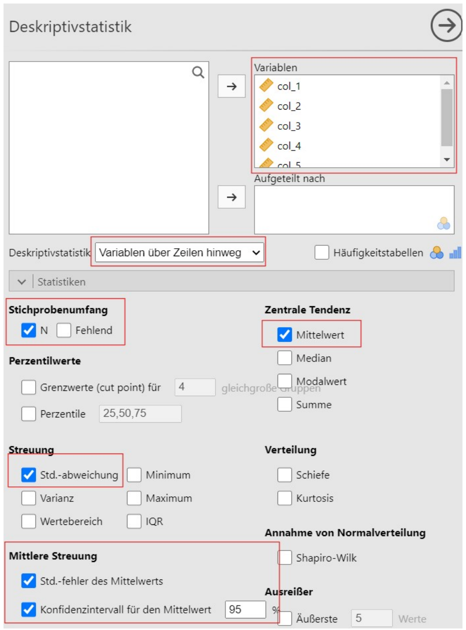
Vergleichen Sie die Resultate der Berechnungen für jede Stichprobe:

f. Weshalb ist die Schätzung für den Erwartungswert für jede Stichprobe unterschiedlich?
g. Was lässt sich über den Zusammenhang zwischen Stichprobengrösse und der Länge des Konfidenzintervalls sagen?

Lösung. Abbildung **??** zeigt die Berechnungsanweisungen für Jamovi und die resultierende Tabelle daraus.

a. Der Erwartungswert der Wartezeiten (das heisst der Populationsmittelwert der Wartezeiten) wird mit dem arithmetischen Mittel der Stichprobe geschätzt und kann in der Tabelle bei `Mittelwert` abgelesen werden. Der Erwartungswert der Wartezeiten beträgt bei allen Stichproben ausser bei der ersten ungefähr 0.22 Stunden, also ein bisschen weniger als eine Viertelstunde.
b. Der Standardabweichung der Wartezeiten der Stichprobe sind in der Tabelle bei `Std.-abw.` abzulesen. Die Standardabweichungen sind für alle Stichproben ausser der ersten ungefähr bei 0.23.
c. Die Standardabweichung der arithmetischen Mittel liegt bei $s/\sqrt{n}$. Für die erste Stichprobe ist dies $0.157/\sqrt{5} = 0.0702125$. Diese Werte werden auch als Standardfehler bezeichnet und sind in der Tabelle bei `Std.-fehler` ablesbar.
d. Die untere und obere Schranke der 95%-Konfidenzintervalle sind bei `Untere` und `Obere` respektive abzulesen.

e. Die Länge des Konfidenzintervalls entspricht jeweils dem höheren Wert minus dem tieferen Wert. Für die erste Stichprobe ist dies $0.597 - 0.208 = 0.389$, für die anderen 0.23, 0.13, 0.09 und 0.03.



Deskriptivstatistik

				95% Konfidenzintervall		
	N	Mittelwert	Std.-fehler	Untere	Obere	Std.-abw.
col_1	5	0.402	0.07000	0.2080	0.597	0.157
col_2	20	0.214	0.05511	0.0987	0.329	0.246
col_3	50	0.227	0.03230	0.1625	0.292	0.228
col_4	100	0.240	0.02241	0.1953	0.284	0.224
col_5	1000	0.241	0.00767	0.2263	0.256	0.242

Anmerkung. Das Konfidenzintervall des Mittelwerts nimmt an, dass die Mittelwerte einer t-Verteilung mit N - 1 Freiheitsgraden folgen

Abbildung 4.14: Links: Jamovi-Anleitung zur Erstellung der Tabelle mit den relevanten Kenngrössen; rechts: Tabelle mit relevanten Kenngrössen.

f. Die Schätzung des Erwartungswertes ist das arithmetische Mittel der Stichprobe. Da jedes Mal eine neue Zufallsstichprobe gezogen wurde und diese nicht dieselben Beobachtungen enthalten, ergeben sich auch jedes Mal andere Stichprobenmittelwerte.
g. Je grösser n, desto kleiner ist das Konfidenzintervall. Wenn man also ein kleines Konfidenzintervall erreichen will, braucht man eine grössere Stichprobe.

Übung 4.5.

Eine Klasse bringt bei einem Biologietest eine durchwachsene Leistung. Die Lehrkraft entscheidet sich die genau gleichen Test zu wiederholen. Berichten Sie das durchschnittliche Resultat der beiden Tests und schätzen Sie den Einfluss der Standardabweichung auf die Länge des Konfidenzintervalls ein.

Lösung. Der Datensatz wird bei `Jamovi` eingelesen und die Analyseparameter wie in Abbildung **??** gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung **??**.

Die Klasse mit $N = 20$ Lernenden hat beim ersten Biologietest eine durchschnittliche Punktzahl von $M = 14.6$ Punkten 95% KI $[11.9, 17.3]$ erzielt. Bei der Wiederholung des Tests wurde eine durchschnittliche Punktzahl von $M = 14.8$

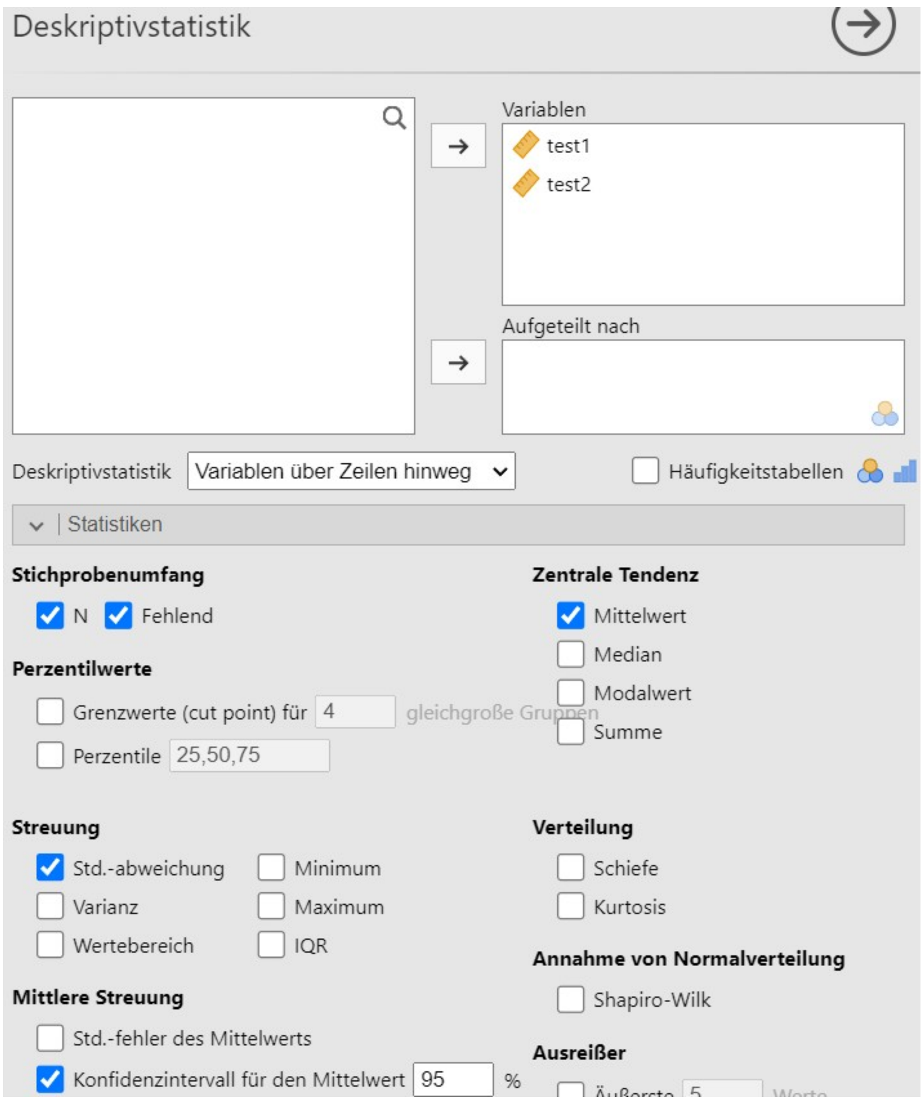


Abbildung 4.15: Jamovi setzen der Analyseparameter.

Deskriptivstatistik

				95% Konfidenzintervall		
	N	Fehlend	Mittelwert	Untere	Obere	Std.-abw.
test1	20	0	14.605	11.895	17.314	5.790
test2	20	0	14.793	13.984	15.603	1.730

Anmerkung. Das Konfidenzintervall des Mittelwerts nimmt an, dass die Mittelwerte einer t-Verteilung mit N - 1 Freiheitsgraden folgen

Abbildung 4.16: Jamovi Ausgabe.

Punkten 95% KI $[14.0, 15.6]$ erzielt. Die Standardabweichung des Testergebnisses war beim ersten Mal $SD = 5.8$ Punkte und bei der Wiederholung $SD = 1.7$ Punkte. Die Länge des Konfidenzintervalls war bei der ersten Durchführung mit $17.3 - 11.9 = 5.4$ Punkten bedeutend grösser als bei der zweiten Durchführung mit $15.6 - 14.0 = 1.6$. Eine grössere Standardabweichung führt also zu einer grösseren Länge des Konfidenzintervalls. Dies kann auch durch Durchprobieren von Testwerten in Gleichung (**??**) festgestellt werden.

4.3 Test

Übung 4.6. Welche der folgenden Aussagen zum Konfidenzintervall des Erwartungswertes sind wahr, welche falsch?

a) Je mehr Personen befragt werden, desto grösser wird das Konfidenzintervall.
b) Je grösser die Standardabweichung des Merkmals, desto grösser wird das Konfidenzintervall.
c) Um ein kleineres Konfidenzintervall zu erreichen, können mehr Beobachtungen gemacht werden.
d) Je grösser die Irrtumswahrscheinlichkeit, desto grösser das Konfidenzintervall.

Lösung.

a) Falsch
b) Richtig
c) Richtig
d) Falsch

Übung 4.7. Im Datensatz `02-exr-koerpergroesse-sex.sav` wurden Körpergrössen von Versuchsteilnehmenden erfasst. Welche der folgenden Aussagen sind wahr, welche falsch?

a) Die durchschnittliche Körpergrösse der Frauen liegt bei $M = 166.0$ cm 90% KI $[164.7, 167.3]$.
b) Die durchschnittliche Körpergrösse der Männer liegt bei $M = 180.3$ cm 95% KI $[178.6, 182.0]$.
c) Es wurden $N = 163$ Frauen beobachtet.
d) Die durchschnittliche Körpergrösse der Männer und Frauen liegt bei $M = 173.1$ cm 99% KI $[171.4, 174.9]$.

Lösung.

a) Falsch
b) Richtig
c) Richtig
d) Richtig

Kapitel 5

Zentrale Tendenz testen

Eine andere Fragestellung, welche mit Daten beantwortet werden soll, ist, ob eine gewisse Aussage wahr ist oder falsch. Eine solche Aussage wird **Hypothese** (Symbol: H) genannt. Eine Hypothese könnte zum Beispiel sein:

$$H : \text{Es regnet.}$$

Ist die Hypothese einmal gefunden, können Daten gesammelt werden, um diese Hypothese zu bestätigen oder zu falsifizieren. Das heisst man geht nach raus ins Feld. Spürt man Regen auf der Haut bedeutet dies, dass H wahr ist. Spürt man keinen Regen, so ist H falsch.

Wenn eine Hypothese wahr ist, dann ist das Gegenteil der Hypothese falsch. Weil oft über die Hypothese und ihr Gegenteil debattiert wird, ist es nützlich die beiden auch terminologisch auseinanderhalten zu können. Die Hypothese, welche den bisherigen Informationsstand reflektiert wird **Nullhypothese** (Symbol H_0) genannt. War es draussen bei der letzten Messung vor einer Stunde schönes Wetter, dann ist die Nullhypothese

$$H_0 : \text{Es regnet nicht.}$$

Das Gegenteil der Nullhypothese wird **Alternativhypothese** (Symbol H_1) genannt. Im Beispiel ist die Alternativhypothese

$$H_1 : \text{Es regnet.}$$

Die Nullhypothese bleibt der Stand der Wahrheit, bis sie durch Daten widerlegt wurde. Wenn man noch drinnen ist, kann keine Aussage über die Wahrheit von H_0 und H_1 gemacht werden, da die Daten fehlen. In diesem Fall wird angenommen, dass H_0 weiterhin wahr ist. Wenn man draussen Regen auf der Haut spürt, deutet dieser Datenpunkt darauf hin, dass H_0 nicht länger wahr ist und jetzt wahrscheinlich H_1 wahr ist. In diesem Fall spricht man davon, dass H_0 **abgelehnt** und H_1 **angenommen** wird.

5.1 Entspricht der Erwartungswert einem gewissen Wert?

Um eine Hypothese mit Daten überprüfbar zu machen, muss diese in eine Form gebracht werden, welche Daten einbezieht. Eine einfache Form einer solchen überprüfbaren Hypothese ist

> H : Das durchschnittliche Vermögen einer in der Schweiz lebenden Person beträgt $100'000$ CHF.

Wenn die Population alle in der Schweiz lebenden Personen sind, dann entspricht dies also der Nullhypothese

$$H_0 : \mu = 100'000.$$

Abstrahiert, soll bei dieser Problemstellung herausgefunden werden, ob der Erwartungswert einer Population einem gewissen Wert entspricht. Das Gegenteil dieser Nullhypothese ist die Alternativhypothese

$$H_1 : \mu \neq 100'000.$$

Dies bedeutet, dass das durchschnittliche Vermögen der Population nicht bei $100'000$ CHF liegt. Weil die Alternativhypothese hier zwei Ausgänge zulässt, nämlich kleiner oder grösser als $100'000$ CHF wird diese Art der Hypothesenstellung als **zweiseitige Hypothese** bezeichnet.

Eine weitere Form der Hypothese wäre

> H : Das durchschnittliche Vermögen einer in der Schweiz lebenden Person beträgt weniger als oder genau $100'000$ CHF.

In Formelsprache übersetzt entspricht dies

$$H_0 : \mu \leq 100'000.$$

Das Gegenteil davon ist, wenn das durchschnittliche Vermögen grösser und ungleich 100’000 CHF ist, also

$H_1 : \mu > 100'000.$

Weil die Alternativhypothese hier nur einen Ausgang zulässt, nämlich grösser als $100'000$ CHF wird dies als **einseitige Hypothese** bezeichnet. Eine einseitige Hypothese kann auf beide Seiten formuliert sein: $H_0 : \mu \leq 100'000$ und $H_1 : \mu > 100'000$, wie eben erwähnt oder auch $H_0 : \mu \geq 100'000$ und $H_1 : \mu < 100'000$.

Achtung

Hinweis. Die verwendeten Zeichen in den Formeln sind

- $=$: Gleichheit, sprich "gleich". Beispiele:
  - $3 = 3$ (3 gleich 3) ist eine wahre Aussage.
  - $3 = 4$ (3 gleich 4) ist eine falsche Aussage.
- $\neq$: Ungleichheit, sprich "ungleich" oder "nicht gleich". Beispiele:
  - $3 \neq 3$ (3 ist nicht gleich 3) ist eine falsche Aussage.
  - $3 \neq 4$ (3 ist nicht gleich 4) ist eine wahre Aussage.
- $<$: Kleiner, sprich "kleiner". Beispiele:
  - $4 < 3$ (4 ist kleiner als 3) ist eine falsche Aussage.
  - $3 < 3$ (3 ist kleiner als 3) ist eine falsche Aussage.
  - $3 < 4$ (3 ist kleiner als 4) ist eine wahre Aussage.
- $\leq$: Kleiner gleich, sprich "kleiner gleich". Beispiele:
  - $4 \leq 3$ (4 ist kleiner oder gleich wie 3) ist eine falsche Aussage.
  - $3 \leq 3$ (3 ist kleiner oder gleich wie 3) ist eine wahre Aussage.
  - $3 \leq 4$ (3 ist kleiner oder gleich wie 4) ist eine wahre Aussage.
- $>$: Grösser, sprich "grösser". Beispiele:
  - $4 > 3$ (4 ist grösser als 3) ist eine wahre Aussage.
  - $3 > 3$ (3 ist grösser als 3) ist eine falsche Aussage.
  - $3 > 4$ (3 ist grösser als 4) ist eine falsche Aussage.
- $\geq$: Grösser gleich, sprich "grösser gleich". Beispiele:
  - $4 \geq 3$ (4 ist grösser oder gleich wie 3) ist eine wahre Aussage.
  - $3 \geq 3$ (3 ist grösser oder gleich wie 3) ist eine wahre Aussage.
  - $3 \geq 4$ (3 ist grösser oder gleich wie 4) ist eine falsche Aussage.

Beispiel 5.1 (Vermögen)**.**

Eine Sozialpolitikberatungsfirma will herausfinden, ob das durchschnittliche Vermögen der in der Schweiz lebenden Personen im letzten Jahr gestiegen ist. Sie stellen dazu basierend auf dem aktuellen Wissensstand die Nullhypothese auf, dass das durchschnittliche Vermögen nicht gestiegen ist, und die

Alternativhypothese, dass das durchschnittliche Vermögen gestiegen ist:

$$H_0 : \mu \leq 100'000 \text{ CHF}$$

$$H_1 : \mu > 100'000 \text{ CHF}$$

Um die Hypothesen auf einer Datengrundlage zu evaluieren, erfragt es das Vermögen von $n = 20$ zufällig ausgewählten Personen und findet ein durchschnittliches Vermögen von $M = 119853$ CHF.

Es kann nun schnell gesagt werden, dass das durchschnittliche Vermögen in der Population gestiegen ist, weil 119853 CHF grösser ist als $100'000$ CHF. Dies so zu behaupten wäre jedoch falsch, weil nicht alle Personen in der Population befragt wurden, sondern lediglich eine Zufallsstichprobe. Wie in Kapitel **??** muss hier für eine Generalisierung der Stichprobe auf die Population der Effekt der zufälligen Stichprobenziehung miteinbezogen werden.

Aufgrund der Zufallsstichprobe ist es unmöglich zu sagen, ob unsere Stichprobe eine eher seltene Stichprobenziehung aus einer Population mit unverändertem durchschnittlichen Vermögen von $100'000$ CHF ist (Abbildung **??** links) oder ob es eine eher häufig vorkommende Stichprobenziehung aus einer Population mit höherem durchschnittlichen Vermögen ist (Abbildung **??** rechts).

Es kann jedoch ausgesagt werden, mit welcher Wahrscheinlichkeit der gefundene Stichprobenmittelwert realisiert wird, gegeben dass die Nullhypothese wahr ist. Hier wird also angenommen, dass eine Population mit Erwartungswert $\mu = 100'000$ CHF vorliegt und dass anschliessend zum Beispiel 3000 Stichproben an je 20 Beobachtungen pro Stichprobe gezogen werden. Von jeder dieser Stichproben wird das arithmetische Mittel berechnet. In der Verteilung dieser Mittelwerte, siehe Abbildung **??**, wird nun der tatsächliche Mittelwert der Stichprobe $\bar{x} = 119853$ verortet.

Der beobachtete Mittelwert ist zwar nicht genau bei $100'000$ CHF, aber trotzdem noch einigermassen plausibel, wenn die Nullhypothese stimmt. Um diesen Gedanken zu formalisieren, gibt es zwei Denkweisen, welche nun vorgestellt werden.

Die eine Denkweise wurde von Ronald Fisher propagierte. Sie stellt die Frage nach der Wahrscheinlichkeit, dass zufällig der beobachtete Wert oder ein noch extremerer Wert in Richtung der Alternativhypothese resultiert, gegeben die Nullhypothese ist wahr. Im Beispiel entspricht dies der Wahrscheinlichkeit den Wert 119853 oder einen grösseren Wert zu beobachten, wenn der Erwartungswert tatsächlich bei $100'000$ CHF liegt. Um diese Wahrscheinlichkeit zu bestimmen, kann einfach gezählt werden, welcher Anteil der Stichprobenmittelwerte grösser oder gleich 119853 CHF ist. Im Beispiel sind dies $0.143 = 14.3\%$. Dieser Wert wird, abgeleitet vom englischen *probability*,

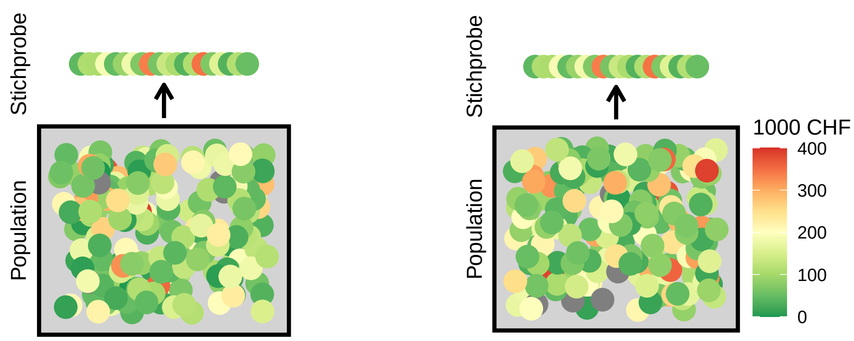


Abbildung 5.1: Vorgestellte Zufallsstichprobenziehung. Links: Nullhypothese ist wahr. Rechts: Nullhypothese ist falsch. Die grauen Punkte entsprechen Vermögen über 400’000 CHF.

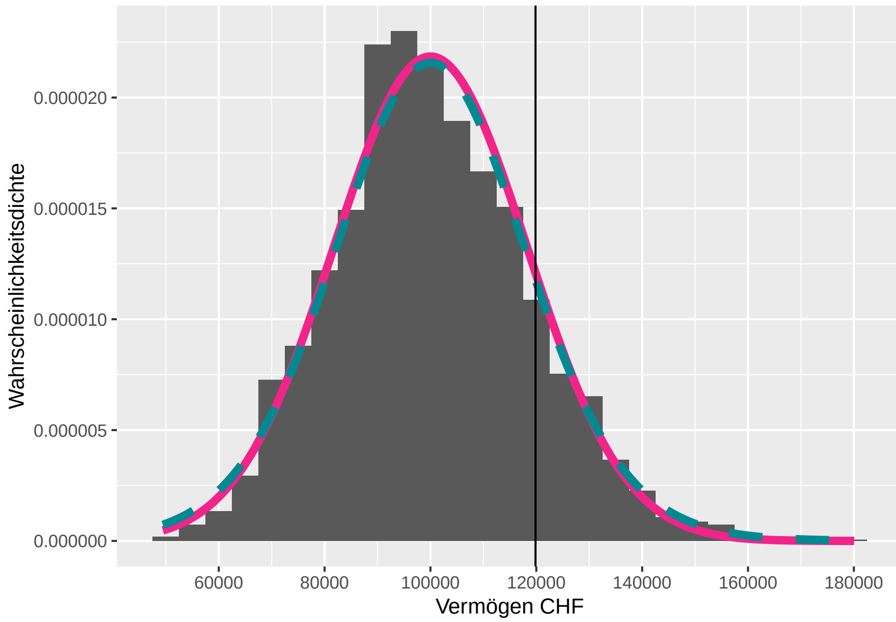


Abbildung 5.2: TODO.

p-Wert (Symbol: p) genannt. Beim Berichten des p-Werts wird normalerweise die führende 0 nicht geschrieben, also $p = .143$.

Bei der anderen von Neyman und Pearson propagierten Denkweise muss noch vor der Datenerhebung ein sogenanntes **Signifikanzniveau** (Symbol α, sprich 'alpha') bestimmt werden. Dieser Wert entspricht der Wahrscheinlichkeit, dass der statistische Test die Nullhypothese verwirft, obwohl diese wahr gewesen wäre. Normalerweise wird $\alpha = 5\%$ gesetzt. Es wird also akzeptiert, dass ein statistischer Test in 5% der Fälle gegen die Nullhypothese entscheidet, obwohl diese wahr wäre. In einem zweiten Schritt wird bestimmt, welches die 5% unwahrscheinlichsten Werte sind, wenn die Nullhypothese wahr ist. Diese Werte werden **Ablehnungsbereich** genannt. Im Beispiel sind dies die 5% höchsten Werte, nämlich Vermögen von 131511 CHF und grössere Vermögen. Nun wird bestimmt, ob der tatsächliche beobachtete Wert im Ablehnungsbereich liegt oder nicht. Im Beispiel liegt der Stichprobenmittelwert 119853 CHF nicht im Ablehnungsbereich. In diesem Fall wird die Nullhypothese nicht verworfen und das Testresultat erhält das Prädikat **nicht signifikant**. Läge der Stichprobenmittelwert im Ablehnungsbereich, so wäre das Testresultat als **signifikant** einzustufen.

Achtung

Hinweis. Ein signifikanter Unterschied bedeutet im allgemeinen Sprachgebrauch ein *bedeutsamer, substanzieller* Unterschied. Im statistischen Kontext bedeutet ein *signifikanter Unterschied*, wie oben beschrieben, dass ein Unterschied bis auf eine gewisse Irrtumswahrscheinlichkeit (angegeben durch das Signifikanzniveau) *nicht zufällig* zustande gekommen ist. Ein *nicht signifikanter Unterschied* bedeutet dagegen, dass die Beobachtung *zufällig* zustande gekommen sein könnte. Für letzteres gibt es zwei Erklärungen: (1) H_0 ist tatsächlich wahr. (2) H_0 ist zwar falsch, aber die Stichprobenziehung hat zufällig zu einem ähnlichen Resultat geführt, wie wenn H_0 wahr wäre. Ist ein Testresultat nicht signifikant, so kann also nicht genau gesagt werden, ob H_0 wahr ist oder nicht. Ist das Testresultat signifikant, so ist H_0 eher unwahrscheinlich.
In manchen Texten werden allgemeine und auch statistische Fragen bearbeitet. Hier empfiehlt sich für den allgemeinen Sprachgebrauch *substanziell* und für die statischen Aussagen *statistisch signifikant* zu verwenden.
Es wird ausserdem empfohlen, das Wort signifikant immer nur als Prädikat für eine Qualifizierung der Nullhypothese zu verwenden. Im Beispiel war $H_0 : \mu \leq 100'000\text{CHF}$. Korrekte Aussage sind: - Das durchschnittliche Vermögen ist im letzten Jahr nicht signifikant gewachsen. - Das durchschnittliche Vermögen ist in diesem Jahr nicht signifikant grösser als $100'000$ CHF.

Die beiden Denkarten entsprechen sich insofern, als ein p-Wert kleiner als 5% ein signifikantes Resultat bei Signifikanzniveau $\alpha = 5\%$ bedeutet. In der Praxis werden beide Methoden verwendet. Im Beispiel liegt der p-Wert bei $p = .143$. Dies bedeutet, dass die Wahrscheinlichkeit zufällig den realisierten Stichprobenmittelwert zu erhalten, gegeben, dass die Nullhypothese stimmt, grösser als 5% ist und demnach auch der Unterschied nicht signifikant ist.

Ein noch zu lösendes Problem ist, dass normalerweise Geld, Zeit und Nerven fehlen, um eine Stichprobenziehung 3000-mal zu wiederholen. Hier hilft es wieder zu beobachten, dass die Verteilung der Werte des Histogramms in Abbildung **??** wieder mit zunehmender Stichprobengrösse immer genauer einer Normalverteilung folgen. Tatsächlich trifft es aufgrund des zentralen Grenzwertsatzes immer zu, dass wenn ein Merkmal mit N Beobachtungen, Erwartungswert μ und Standardabweichung σ hat, der Wert

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

normalverteilt ist, wobei μ hier dem Wert der Nullhypothese entspricht, also

100′000 CHF. Dies entspricht der roten Linie in Abbildung **??**. Ist die Standardabweichung des Merkmals σ in der Population unbekannt, so wird diese mit der Standardabweichung in der Stichprobe s geschätzt. Diese zusätzliche Unsicherheit führt dazu, dass

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{5.1}$$

nicht mehr normal-, sondern t-verteilt ist bei $n - 1$ Freiheitsgraden (grüne Linie, Abbildung **??**). Die t-Verteilung mit allen Freiheitsgraden ist in `Jamovi` hinterlegt und es kann der Software überlassen werden den p-Wert und den Ablehnungsbereich genau zu bestimmen. In Abbildung **??** wurde nochmal illustriert, dass es bei vielen Beobachtungen der theoretische p-Wert (Kurve) mit dem empirischen p-Wert der Simulationen (Histogramm) übereinstimmt respektive der Ablehnungsbereich der t-Verteilung (Kurve) gleich ist, wie der simulierte Ablehnungsbereich (Histogramm).

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

Die Berechnung des für den Test relevanten Wertes, hier des t-Wertes wird **Teststatistik** (oder auch *Prüfgrösse* oder nur *Statistik*) genannt. Eine Teststatistik hat normalerweise eine bekannte theoretische Verteilung, welcher die Teststatistik folgt, wenn die Nullhypothese wahr ist. Aufgrund der theoretischen t-Verteilung der vorliegenden Statistik und der einen Stichprobe (vgl. nächstes Kapitel) wird dieser Test **Einstichproben-t-Test** genannt.

Das oben gefundene Resultat wird in der folgenden Form berichtet:

> Ein Einstichproben-t-Test ergibt, dass das durchschnittliche Vermögen ($M = 119853$ CHF, $SD = 88528$, $N = 20$) in diesem Jahr nicht signifikant grösser als 100′000 CHF ist, $t(19) = 1.003$, $p = .164$.

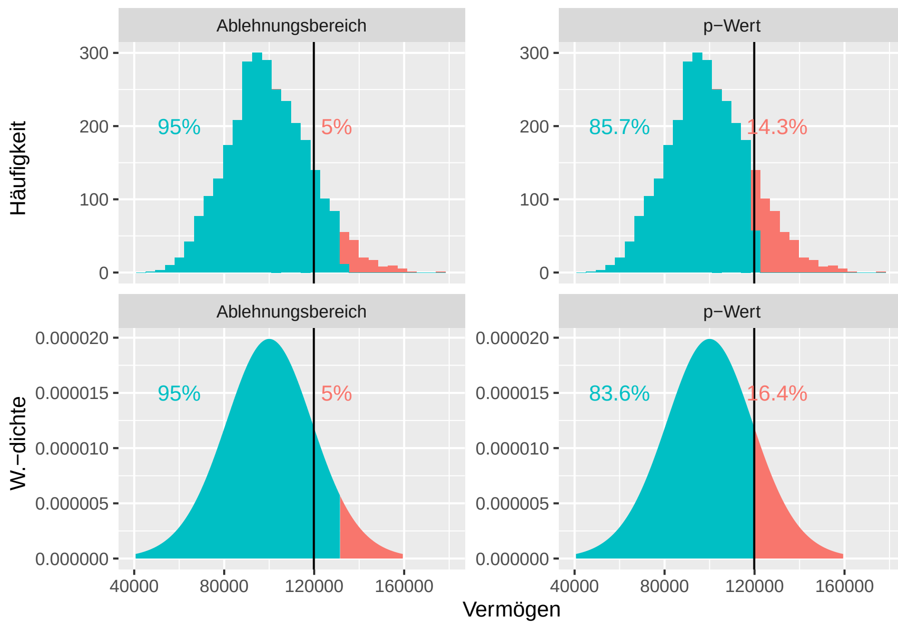


Abbildung 5.3: Oben: Histogramm der simulierten Verteilung; unten: theoretische t-Verteilung; links: Illustration p-Wert; rechts: Illustration Ablehnungsbereich. Die Linie entspricht dem beobachteten Stichprobenmittelwert.

Achtung

Hinweis. Folgende Begriffe und Zahlen werden dabei verwendet:

- Das *durchschnittliche* Vermögen (fehlt durchschnittlich ist die Aussage falsch).
- M, SD, N entsprechen dem arithmetischen Mittel, der geschätzten Standardabweichung und der Anzahl Beobachtungen in der Stichprobe. Die Einheit muss nicht wiederholt werden.
- Signifikanz (siehe letzter Hinweis)
- grösser als $100'000$ CHF ist die Referenz zur Alternativhypothese
- $t(19)$ bedeutet, dass die Teststatistik t-verteilt ist mit 19 Freiheitsgraden.
- 1.003 ist der Wert der Teststatistik berechnet mit Formel (**??**) aus der Stichprobe. Dieser Wert ist skaliert und muss im Kontext der standardisierten t-Verteilung wie in Abbildung **??** interpretiert werden.
- $p = .164$ entspricht dem p-Wert. Es wird normalerweise die führende 0 weggelassen (also nicht 0.164), da es sich um eine Zahl handelt, welche nie kleiner als 0 oder grösser als 1 sein kann.

Beispiel 5.2 (Alexithymie)**.** Mit Gefühlsblindheit oder *Alexithymie* (griechisch: a = ohne, lexis= lesen, sprechen, thymie = Gefühle) werden Einschränkungen bei der Fähigkeit Emotionen wahrzunehmen, zu erkennen und zu beschreiben bezeichnet. Es gibt ein online Messinstrument, welches die Alexithymie auf einer Skala von 37 Punkten (kleine Gefühlsblindheit) bis 185 (grosse Gefühlsblindheit) misst. Die Skala wurde so gewählt, dass die durchschnittliche Alexithymie aller Menschen bei 100 liegt. Eine Psychologin interessiert sich nun dafür, ob junge Menschen unter 25 durchschnittlich andere Alexithymie-Werte aufweisen als die Gesamtbevölkerung. Um dies zu testen, befragt sie $N = 391$ unter 25-jährige mit besagtem Messinstrument. In dieser Gruppe wurde eine durchschnittliche Alexithymie von $M = 96.7$ Punkten festgestellt.

Der erste Schritt ist auch hier die Null- und Alternativhypothesen aufzustellen. Die Psychologin stellt die Frage, ob sich die durchschnittliche Alexithymie in der Grundgesamtheit, in der Folge mit μ bezeichnet, von 100 unterscheidet oder nicht. Es ist zu beobachten, dass sie keine Annahme über die Richtung der Abweichung trifft (eine höhere oder eine tiefere Alexithymie wären denkbar) und es sich deshalb um eine zweiseitige Hypothesenstellung handelt.

Die Nullhypothese beschreibt den bisherigen Informationsstand, also dass die durchschnittliche Alexithymie der Population bei 100 Punkten liegt, oder kurz

$H_0 : \mu = 100$ Punkte.

Die Alternativhypothese besagt das Gegenteil davon, also hier, dass die durchschnittliche Alexithymie nicht mehr bei 100 Punkten liegt, oder kurz

$H_1 : \mu \neq 100$ Punkte.

Um die Wahrscheinlichkeit des beobachteten arithmetischen Mittels der Stichprobe von $M = 96.7$ Punkten zu ermitteln, gegeben, dass die Nullhypothese wahr ist, kann erneut auf den Gedanken der wiederholten Stichprobenziehung zurückgegriffen werden. Bei diesem Gedankenexperiment wird angenommen, dass Nullhypothese wahr ist und dass das die Untersuchung 4000-mal wiederholt wurde mit jeweils 391 Beobachtungen. Von jeder dieser Stichproben kann wiederum das arithmetische Mittel berechnet werden. Die Verteilung dieser arithmetischen Mittel ist in Abbildung **??** oben dargestellt.

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).

## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

Der p-Wert, also die Wahrscheinlichkeit, dass der beobachtete Wert oder ein noch extremerer Wert in Richtung der Alternativhypothese resultiert, wird hier aufgrund der zweiseitigen Hypothesenstellung auch zweiseitig ausgelegt. Extremer in Richtung der Alternativhypothese meint hier alle Werte, die weiter weg als der beobachtete Durchschnittswert 96.7 vom hypothetischen Erwartungswert $\mu = 100$ sind. Konkret sind dies alle Werte, welche kleiner als 96.7, und alle Werte, welche grösser als 103.3 sind (roter Bereich in Abbildung **??** oben rechts). Der Anteil der Werte, welche diese Bedingung erfüllen liegt bei $p = 0.7\%$. Es ist demnach recht unwahrscheinlich, dass die Nullhypothese stimmt und zufällig ein Stichprobendurchschnittswert von 96.7 Alexithymie-Punkten herauskommt.

Aufgrund der zweiseitigen Hypothesenstellung beinhaltet auch der Ablehnungsbereich sowohl die tiefsten 2.5% und höchsten 2.5%, also insgesamt die 5% extremen Durchschnittswerte. Dies sind alle Werte tiefer als 97.48 und alle Werte höher als 102.45 (roter Bereich in Abbildung **??** oben links). Da das arithmetische Mittel der Stichprobe 96.7 im Ablehnungsbereich liegt, liegt hier ein signifikantes Resultat vor bei Signifikanzniveau 5%.

Auch in diesem Fall kann die Verteilung der Stichprobenmittelwerte mit dem zentralen Grenzwertsatz angenähert werden. Es ergeben sich annähernd dieselben Resultate für den p-Wert (roter Bereich in Abbildung **??** unten rechts) und für den Ablehnungsbereich (roter Bereich in Abbildung **??** unten links).

Die Psychologin kann nun wie folgt berichten:

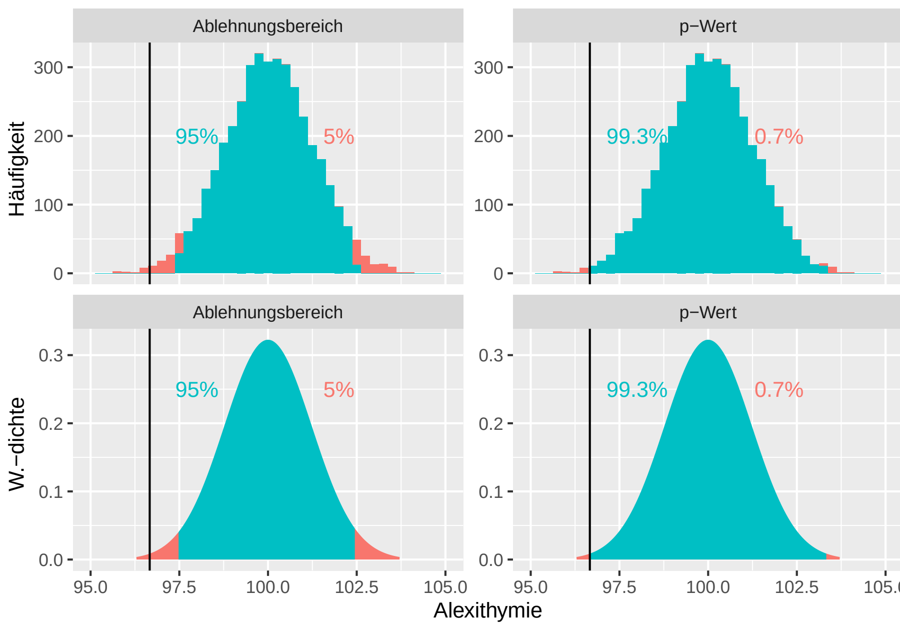


Abbildung 5.4: Oben: Histogramm der simulierten Verteilung der Alexithymie-Mittelwerte; unten: theoretische t-Verteilung; links: Illustration p-Wert; rechts: Illustration Ablehnungsbereich. Die Linie entspricht dem beobachteten Stichprobenmittelwert.

> Ein Einstichproben-t-Test ergibt, dass die durchschnittliche Alexithymie ($M = 96.7$ Punkte, $SD = 24.4$, $N = 391$) sich bei den unter 25-jährigen signifikant vom Populationsdurchschnitt von 100 Punkten unterscheidet, $t(390) = -2.698$, $p = .007$.

5.2 Weicht der gefundene Durchschnitt stark vom hypothetischen Wert ab?

In einem so berichteten Testresultat sind essenziell zwei Informationen enthalten: (1) was sind die getesteten Hypothesen und (2) wie wahrscheinlich es ist, dass das gefundene Resultat eine Folge der Zufallsstichprobenziehung ist. Was hier noch fehlt ist eine Angabe darüber, wie gross die praktische Relevanz dieses Testresultates ist.

Um eine solche Relevanz zu messen wurde der Begriff der Effektstärke eingeführt. Eine Effektstärke ist eine Zahl ohne Einheit (Meter, Franken, …), welche unabhängig von der Stichprobengrösse ist und nahe bei null liegt, wenn die Nullhypothese nicht abgelehnt wurde.

Wird im Vermögensbeispiel **??** die Differenz zwischen geschätztem Erwartungswert und hypothetischem Erwartungswert

$$\bar{x} - \mu = 119853\text{CHF} - 100000\text{CHF} = 19853$$

betrachtet, so fällt auf, dass dieser Wert bereits zwei der oben genannten Eigenschaften aufweist. Tatsächlich ist dieser Wert unabhängig von der Stichprobengrösse und er liegt nahe bei 0, wenn das Testresultat nicht signifikant war. Letzteres kann beobachtet werde indem in der Formel (**??**) verschiedene Differenzen eingesetzt werden und mit der Abbildung **??** verglichen werden.

Wenn jetzt ein anderer Sozialpsychologe die Auswertung wiederholen würde, aber statt in CHF in Rappen Rp rechnet, dann erhält er den Wert

$$\bar{x} - \mu = 11985300\text{Rp} - 10000000\text{Rp} = 1985300.$$

Dass mit den gleichen Zahlen je nach Einheit eine andere Effektstärke gefunden wird, ist unpraktisch für den Vergleich der Testresultate. Die Lösung in diesem Fall ist diese Differenz durch die geschätzte Standardabweichung zu rechnen. Dies ergibt

- in CHF: $d = \frac{\bar{x}-\mu}{s} = \frac{119853\text{CHF}-100000\text{CHF}}{88528\text{CHF}} = 0.22$
- in Rp: $d = \frac{\bar{x}-\mu}{s} = \frac{11985300\text{Rp}-10000000\text{Rp}}{8852800\text{Rp}} = 0.22.$

Mit dieser Formel werden für beide Einheiten derselbe Wert berechnet. Effektiv dient jetzt als Einheit die Standardabweichung: Eine grosse Differenz bei einer grossen Standardabweichung des Merkmals führt zur selben Effektstärke wie eine kleine Differenz bei kleiner Standardabweichung eines Merkmals. Da Menschen sich nicht gewohnt sind Zahlen als Standardabweichungen zu interpretieren hat (?) folgende Richtwerte entwickelt:

- $|d| \approx 0.3$: schwacher Effekt
- $|d| \approx 0.5$: mittlerer Effekt
- $|d| \approx 0.8$: starker Effekt

Cohen selbst hat davor gewarnt diese Werte als absolut darzustellen. Vielmehr sollte die Interpretation der Effektstärke vom Forschungsgebiet und dem Messinstrument abhängen. Um im Unterricht eine beurteilbare Praxis zu etablieren, sollen folgende Regeln gelten:

- $0 < |d| \leq 0.4$: schwacher Effekt
- $0.4 < |d| \leq 0.65$: mittlerer Effekt
- $0.65 < |d|$: starker Effekt

Das Berichten der Testresultate wird mit der Effektstärke ergänzt:

> Ein Einstichproben-t-Test ergibt, dass das durchschnittliche Vermögen ($M = 119853$ CHF, $SD = 88528$, $N = 20$) in diesem Jahr nicht signifikant grösser als $100'000$ CHF ist, $t(19) = 1.003$, $p = .164, d = 0.22$.

> Ein Einstichproben-t-Test ergibt, dass die durchschnittliche Alexithymie ($M = 96.7$ Punkte, $SD = 24.4$, $N = 391$) sich bei den unter 25-jährigen signifikant vom Populationsdurchschnitt von 100 Punkten unterscheidet, $t(390) = -2.698$, $p = .007, d = -0.14$.

In beiden Fällen liegt ein schwacher Effekt vor. Der Effekt bei der Alexithymie ist schwächer als der Effekt bei der Vermögensstudie. Der p-Wert sagt aber aus, dass der Effekt beim Vermögen durch die Zufallsstichprobe zustande gekommen ist, während es bei der Alexithymie unwahrscheinlich ist, dass der Effekt durch die Zufallsstichprobe zustande gekommen ist.

5.3 Testvoraussetzungen

Damit der Einstichproben-t-Test durchgeführt werden dürfen, müssen einige Voraussetzungen eingehalten werden.

1. Das Merkmal muss intervallskaliert sein.

2. Die Beobachtungen müssen einer Zufallsstichprobe der Population entsprechen.
3. Die Beobachtungen müssen einer Normalverteilung entstammen oder die Anzahl der Beobachtungen muss gross genug sein. Häufig wird die Faustregel mehr als 30 Beobachtungen verwendet.

5.4 Übungen

Übung 5.1.

Reproduziere das Beispiel Vermögen **??** mit `Jamovi` indem folgende Teilschritte durchgeführt werden:

- Datensatz `05-exm-vermoegen.sav` in `Jamovi` einladen.
- Wähle `Analysen > t-Tests > t-Test mit einer Stichprobe`.
- Definiere die Hypothese wie im Beispiel und wähle die Testoptionen so, dass du alle Zahlen des Testberichts wiederfindest.

Lösung.

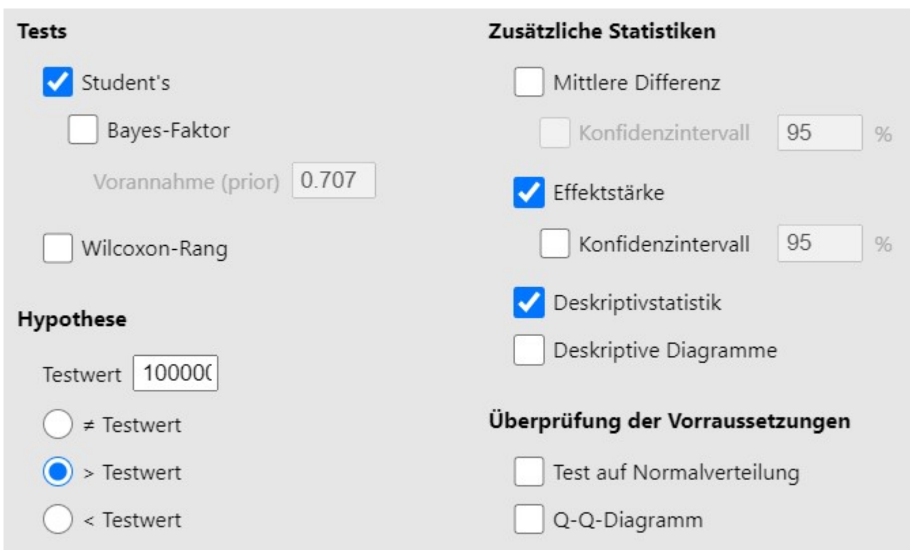


Abbildung 5.5: Jamovi Eingabe.

Übung 5.2.

Reproduziere das Beispiel Alexithymie **??** mit `Jamovi` indem folgende Teilschritte durchgeführt werden:

t-Test mit einer Stichprobe

t-Test mit einer Stichprobe

		Statistik	df	p		Effektstärke
vermoegen	Student's t	1.00	19.0	0.164	Cohens d	0.224

Anmerkung. H_a $\mu > 100000$

Deskriptivstatistik

	N	Mittelwert	Median	Std.-abw.	Std.-fehler
vermoegen	20	119853	105787	88528	19795

Abbildung 5.6: Testresultat Einstichproben-t-Test und deskriptive Statistiken.

- Datensatz `05-exm-alexithymie.sav` in `Jamovi` einladen.
- Wähle `Analysen > t-Tests > t-Test mit einer Stichprobe`.
- Definiere die Hypothese wie im Beispiel und wähle die Testoptionen so, dass du alle Zahlen des Testberichts wiederfindest.

Lösung.

Übung 5.3.

Es soll überprüft werden, ob der 24-stündige Tagesrhythmus, auch *zirkadianer Rhythmus* genannt, des Menschen auch ohne Tageslicht aufrechterhalten wird. Eine solche Untersuchung wird von **?** berichtet. Wir gehen von folgendem fiktiven Versuch aus: Freiwillige werden für vier Tage in einer Kellerwohnung ohne jedes Tageslicht einquartiert. Jede Versuchsperson ist während der vier Tage allein, darf die Wohnung nicht verlassen und erhält keinerlei Hinweise auf die aktuelle Tageszeit. Die Person muss unmittelbar vor dem Zu-Bett-Gehen, einen Knopf betätigen, wodurch die Uhrzeit festgehalten wird. Als Variable wird die Dauer der `tageslaenge` (in Stunden) zwischen dem Zu-Bett-Gehen am dritten Versuchstag und dem Zu-Bett-Gehen am vierten Versuchstag verwendet. Die erhobenen Daten sind in `05-exr-circadian.sav` abgelegt.

a) Ohne einen Test durchzuführen, haben die Proband:innen einen anderen zirkadianen Rhythmus als Menschen die nicht am Experiment teilnehmen? Weshalb es hier sinnvoll ist einen statistischen Test zu verwenden?
b) Stellen Sie mit einem Einstichproben-t-Test fest, ob der zirkadiane Rhythmus durch das Tageslicht beeinflusst wird. Stellen Sie insbesondere die Hypothesen auf und berichten Sie das Testresultat adäquat.

Abbildung 5.7: Jamovi Eingabe.

t-Test mit einer Stichprobe

t-Test mit einer Stichprobe

		Statistik	df	p		Effektstärke
alexithymie	Student's t	−2.70	390	0.007	Cohens d	−0.136

Anmerkung. H_a $\mu \neq 100$

Deskriptivstatistik

	N	Mittelwert	Median	Std.-abw.	Std.-fehler
alexithymie	391	96.7	94.9	24.4	1.24

Abbildung 5.8: Testresultat Einstichproben-t-Test und deskriptive Statistiken.

c) Erklären Sie alle Zahlen und Symbole im Testbericht.

Lösung. Für diese Übung werden die Daten in `Jamovi` wie in Abbildung **??** analysiert. Das Resultat der Analyse ist in Abbildung **??** festgehalten.

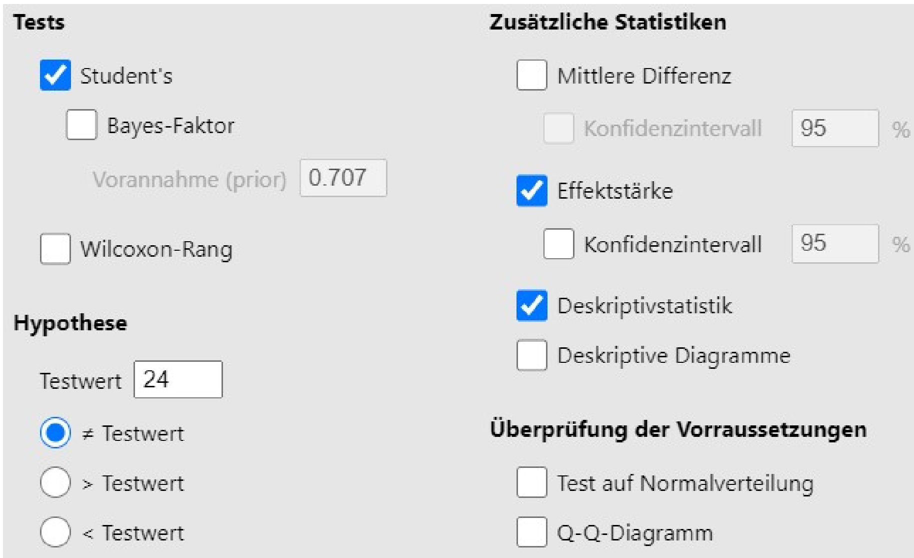


Abbildung 5.9: Jamovi Eingabe.

t-Test mit einer Stichprobe

		Statistik	df	p		Effektstärke
tageslaenge	Student's t	2.41	56.0	0.019	Cohens d	0.319

Anmerkung. H_a $\mu \neq 24$

Deskriptivstatistik

	N	Mittelwert	Median	Std.-abw.	Std.-fehler
tageslaenge	57	24.6	24.3	1.83	0.242

Abbildung 5.10: Jamovi Ausgabe.

a) Die Versuchpersonen haben einen durchschnittlichen zirkadianen Rhythmus von $M = 24.6$ Stunden. Dies ist länger als die regulären 24 Stunden. Es ist unklar, ob hier gerade zufällig Personen beobachtet wurden bei welche sich der zirkadiane Rhythmus verlängert. Um die Wahrscheinlichkeit dieses Zufalls zu quantifizieren wird ein statistischer Test durchgeführt.

b) Die Nullhypothese geht vom aktuell bekannten aus, also in diesem Fall, dass sich der durchschnittliche zirkadiane Rhythmus unter den Versuchsbedinungen nicht verändert. Der normale zirkadiane Rhythmus ist Sonnenbedingt 24 Stunden lang, also wird die Nullhypothese $H_0 : \mu = 24$ Stunden aufgestellt. μ ist hier die durchschnittiche Dauer des zirkadianen Rhythmus in der Population. Im Versuch geht es darum festzustellen, ob der normale zirkadiane Rhythmus gehalten wird oder nicht. Ein nicht gehaltener zirkadianer Rhythmus würde bedeuten, dass sich die Tagesdauer verkürzt oder verlängert gegenüber der Nullhypothese. Es ist hier also eine zweiseitige Hypothesenstellung und die Alternativhypothese lautet $H_1 : \mu \neq 24$ Stunden.

   Ein Einstichproben-t-Test ergibt, dass die durchschnittliche Tageslänge ($M = 24.6$ Stunden, $SD = 1.8$, $N = 57$) unter Experimentalbedingungen sich signifikant von 24 Stunden unterscheidet, $t(56) = 2.41$, $p = .019, d = 0.319$.

c) M, SD, und N sind das arithmetische Mittel, die geschätzte Standardabweichung und die Anzahl Beobachtungen der Stichprobe. p ist die Wahrscheinlichkeit, zufällig den Stichprobenmittelwert oder einen noch extremeren Wert im Sinne der Alternativehypothese zu beobachten, falls die Nullhypothese stimmt. Dieser Wert ist kleiner als 5%. Deswegen wird von einem signifikanten Unterschied der durchschnittlichen Tageslänge zum Erwartungswert gesprochen. 24 Stunden ist der Vergleichswert der Nullhypothese. $t(56)$ bedeutet, dass die Teststatistik t-verteilt ist mit 56 Freiheitsgraden, sofern die Nullhypothese wahr ist. Mit der aktuellen Stichprobenziehung wurde ein Wert von 2.41 realisiert. Dieser Wert ist mit der t-Verteilung in Abbildung **??** zu vergleichen. Der Wert entspricht einer eher unwahrscheinlichen Beobachtung dieser Verteilung. $d = 0.319$, schliesslich, bezieht sich auf die Effektstärke. Das Testresultat entspricht einem mittleren Effekt.

Übung 5.4.

Im Schwimmclub Neustadt erreichen neue Schwimmer nach einem Jahr Training eine Kraul-Schwimmzeit von durchschnittlich 1.58 Minuten für 100 Meter. Eine Sportstudentin will eine neue Trainingsmethode ausprobieren und herausfinden, ob die Methode bessere Ergebnisse erzielt. Dazu trainiert neue Schwimmer ein Jahr lang mit dieser Methode und misst anschliessend deren Kraul-Schwimmzeit über 100 Meter. Die Daten sind in `05-exr-schwimmen.sav` abgelegt.

a) Wie viele Schwimmer hat die Sportstudentin trainiert?
b) Ist die neue Trainingsmethode besser als die bisherige? Erklären Sie die Signifikanz und Relevanz des Experimentresultats.

Lösung. Für diese Übung werden die Daten in `Jamovi` wie in Abbildung **??** analysiert. Das Resultat der Analyse ist in Abbildung **??** festgehalten.

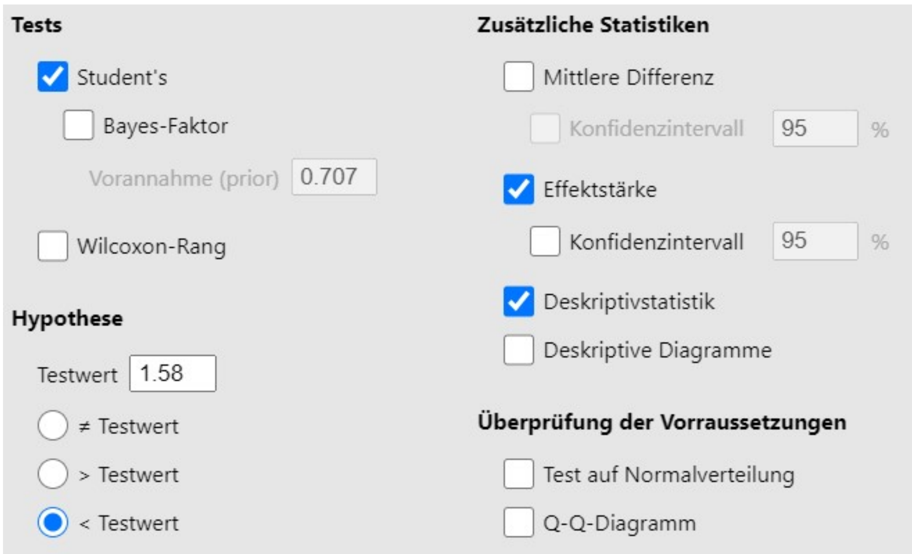


Abbildung 5.11: Jamovi Eingabe.

t-Test mit einer Stichprobe

		Statistik	df	p		Effektstärke
kraul_zeit	Student's t	−1.54	12.0	0.075	Cohens d	−0.426

Anmerkung. H_a $\mu < 1.58$

Deskriptivstatistik

	N	Mittelwert	Median	Std.-abw.	Std.-fehler
kraul_zeit	13	1.31	1.20	0.629	0.175

Abbildung 5.12: Jamovi Ausgabe.

a) Die Sportstudentin hat $N = 13$ Schwimmer trainiert.
b) Die Forschungsfrage ist hier, ob die neue Trainingsmethode besser ist. Besser meint hier, dass die mit dieser Trainingsmethode trianierten Schwimmer nach dem Training durchschnittlich schneller schwimmen als die anderen. Die Alternativhypothese ist also $H_1 : \mu < 1.58$. Die Nullhypothese sagt genau das Gegenteil davon aus, nämlich, dass die durchschnittliche Schimmzeit mit der neuen Methode gleich bleibt oder sogar noch länger wird $H_0 : \mu \geq 1.58$. Das Testresultat lässt sich wie folgt berichten:

> Ein Einstichproben-t-Test ergibt, dass die durchschnittliche Schwimmzeit ($M = 1.31$ Minuten, $SD = 0.63$, $N = 13$) mit der neuen Trainingsmethode nicht signifikant tiefer als 1.58 Minuten ist, $t(12) = -1.54$, $p = .075, d = -0.426$.

Das Testresultat ist nicht signifikant, da der p-Wert grösser als 5% ist. Tatsächlich bedeutet $p = .075$, dass, wenn die Nullhypothese wahr ist, das gefundene Testresultat oder dass die Schwimmer noch schneller sind in 7.5% zufällig durch die Zufallsstichprobenziehung zustande kommt. Kurz gesagt, das Resulat könnt auch Zufall sein.

Die gefundene Effektstärke ist mittel. Wenn das Resultat nicht zufällig wäre, dann würde die Trainingsmethode immerhin einen mittleren Effekt erzielen. Wenn es tatsächlich einen mittleren Effekt gibt, dann könnte die Sportstudentin das Experiment nochmal mit mehr Probanden wiederholen, um den Effekt auch als statistisch signifikan nachweisen zu können. Falls der gefundene Effekt nur zufällig zustande gekommen ist und er nicht exisitiert, wird auch eine Experimentwiederholung mit mehr Probandinnen immernoch kein signifikantes Testergebnis liefern.

Übung 5.5.

Die Firma Pear bringt ein neues Smartphone das F42 der Reihe Supernova X auf den Markt. Das Smartphone ist für Jugendliche im Alter von $15 - 20$ Jahre konzipiert. Das Vorgängermodell F41 wurde für durchschnittlich 300 CHF verkauft. Um herauszufinden, ob sich die durchschnittliche Zahlbereitschaft des neuen Modells von der Zahlbereitschaft für das alte Modell abweicht, erfragt Pear bei 70 Jugendlichen die Zahlbereitschaft. Die Daten stehen unter `04-exr-marktpreisanalyse.sav` zur Verfügung.

a) Stellen Sie die oben formulierte Hypothese mit mathematischer Schreibweise dar.
b) Testen Sie die Hypothese.
c) Berichten Sie die Testergebnisse.
d) Was bedeuten die Werte Statistik, df, p und Effektstärke.

e) Der Stichprobenmittelwert liegt tiefer als 300 CHF. Hätte man bereits hier feststellen können, dass sich die durchschnittliche Zahlungsbereitschaft verändert hat?

Lösung. Für diese Übung werden die Daten in `Jamovi` wie in Abbildung **??** analysiert. Das Resultat der Analyse ist in Abbildung **??** festgehalten.

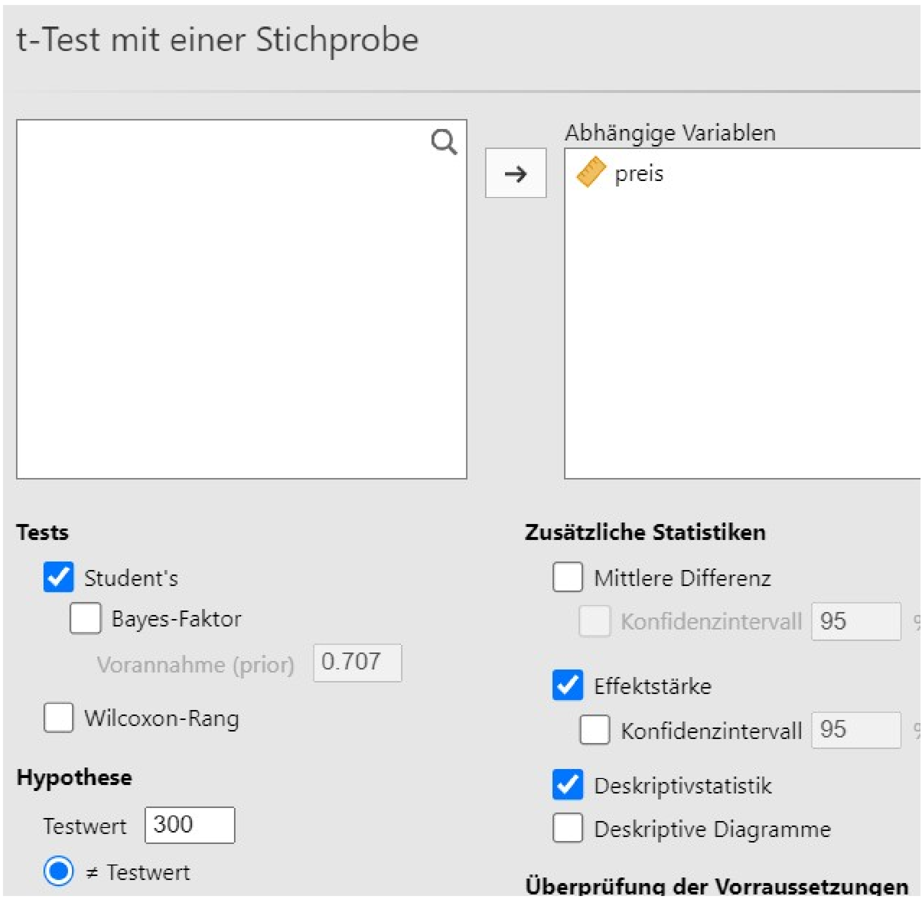


Abbildung 5.13: Jamovi Eingabe.

a) Es gibt zunächst keinen Anhaltspunkt, weshalb sich die Zahlbereitschaft geändert haben sollte. Deshalb ist die Nullhypothese $H_0 : \mu = 300$ CHF, wobei μ für den Erwartungswert des Merkmals Preis ist. Pear fragt sich, ob der μ von 300 abweicht, gibt aber keine Richtung vor. Deshalb wurde H_0 zweiseitig formuliert. Das Gegenteil der Nullhypothese ist die Alternativhypothese $H_1 : \mu \neq 300$ CHF.
b) Das Testen erfolgt wie oben in den Bildschirmaufnahmen von `Jamovi` dargestellt.

t-Test mit einer Stichprobe

		Statistik	df	p		Effektstärke
preis	Student's t	−0.848	69.000	0.399	Cohens d	−0.101

Anmerkung. H_a $\mu \neq 300$

Deskriptivstatistik

	N	Mittelwert	Median	Std.-abw.	Std.-fehler
preis	70	288.338	296.089	115.009	13.746

Abbildung 5.14: Jamovi Ausgabe.

c) Ein Einstichproben-t-Test ergibt, dass sich die durchschnittliche Zahlbereitschaft ($M = 288.34$ CHF, $SD = 115.01$, $N = 70$) nicht signifikant von 300 CHF unterscheidet, $t(69) = -0.848, p = .399, d = -0.101$.

d) Statistik entspricht der beobachteten Teststatistik in der Stichprobe. Der Wert kann im Vergleich zur t-Verteilung in Abbildung **??** gelesen werden. -0.848 ist bei allen dargestellten Verteilungen kein seltener Wert, wenn die Nullhypothese stimmt. Dieser Wert der Statistik deutet also nicht darauf hin, dass die Nullhypothese falsch ist. Die Freiheitsgrade df bestimmen die genaue Form der t-Verteilung. In Abbildung **??** sind die genauen Formen für $df = 1$, $df = 4$ und $df = 9$ dargestellt. Die t-Verteilung, welche die Verteilung der Mittelwerte am besten abbildet ist die mit $df = n - 1$, wobei n die Anzahl Beobachtungen ist. Es sind 70 Beobachtungen gemacht worden, also ist $df = 69$. Die t-Verteilung sieht in diesem Fall ungefähr aus wie die Normalverteilung in Abbildung **??**. Der p-Wert von 0.399 bedeutet, dass die Wahrscheinlichkeit diesen Stichprobenmittelwert oder einen extremeren im Sinne der Alternativhypothese bei 39.9% liegt und damit ziemlich wahrscheinlich ist, gegeben dass die Nullhypohthese wahr ist. Auch dies reflektiert also, dass aufgrund der Stichprobe nicht geschlossen werden kann, dass der Erwartungswert von 300 CHF abweicht. Die Effektstärke $d = -0.101$ ist gemäss Cohen als schwach einzustufen.

e) Der Stichprobenmittelwert sagt aus, dass in dieser Stichprobe die Zahlungsbereitschaft nicht gleich war wie für das Modell F41. Diese Aussage ist jedoch limitiert auf die Stichprobe und kann nur auf die Population ausgeweitet werden, wenn ein statistischer Test durchgeführt wurde. Es könnte ja sein, dass es einen Unterschied im Populationsmittelwert gibt, dieser aber aufgrund einer seltenen Zufallsstichprobenziehung nicht offenbar wird.

Übung 5.6.

TODO

Lösung. TODO

5.5 Test

Übung 5.7. Welche der folgenden Aussagen zum Einstichproben-t-Test sind wahr, welche falsch?

a) Der Einstichproben-t-Test überprüft, ob der Stichprobenmittelwert einer bestimmten Zahl entspricht.
b) Beim Einstichproben-t-Test ist die Teststatistik t-verteilt mit $n-1$ Freiheitsgraden.
c) Der p-Wert ist immer kleiner als das Signifikanzniveau.
d) $H_1 : \mu > 50$ ist eine mögliche Formulierung für die Alternativhypothese des Einstichproben-t-Test.

Lösung.

a) Falsch
b) Richtig
c) Falsch
d) Richtig

Übung 5.8. In der Schweiz wird empfohlen 3 Liter Flüssigkeit pro Tag zu sich zu nehmen. Auf einer Reise in die USA fragt Karin zufällige Leute nach ihrer Flüssigkeitsaufnahme. Die Daten notiert sie im Datensatz `05-exr-drink-usa`. Sie will nun testen, ob alle Leute in den USA durchschnittlich mehr Flüssigkeit pro Tag zu sich nehmen, als es in der Schweiz empfohlen ist. Testen Sie die Hypothese einem Einstichproben-t-Test, stellen Sie dabei `Jamovi` auf 3 Nachkommastellenrundung ein. Welche der folgenden Aussagen sind wahr, welche falsch.

a) Die durchschnittliche Flüssigkeitsaufnahme ist in den USA signifikant grösser als in der Schweiz empfohlen.
b) Der gefundene Effekt ist gemäss Cohen als gross einzustufen.
c) Karin hat 14 Personen befragt.
d) Die Nullhypothese lautet $H_0 : \mu \leq 3$ Liter.

Lösung.

a) Richtig
b) Falsch
c) Falsch
d) Richtig

Übung 5.9.

TODO

Lösung. TODO

Teil II

Zwei Gruppen vergleichen

Kapitel 6

Gruppenmittelwertunterschied bei einem intervallskalierten und normalverteilten Merkmal

Bislang wurde versucht mithilfe *einer* Stichprobe eine Aussage über *eine* Population zu treffen. Dies setzt voraus, dass der Erwartungswert bereits aus früheren Untersuchungen bekannt ist oder theoretisch hergeleitet werden kann (Beispiel zirkadianer Rhythmus). In der Realität ist dies oft nicht der Fall. Es muss also gleichzeitig etwas über eine potenziell veränderte Population und über die Referenzpopulation herausgefunden werden. Im experimentellen Kontext entspricht dies dem Vergleich der Experimental- mit der Kontrollgruppe. Im observationellen Kontext wird die Referenzgruppe willkürlich bestimmt.

Beispiel 6.1 (Trennungsschmerz)**. ?** haben untersucht, ob das Geschlecht einen Einfluss auf den Schmerz bei der Auflösung einer romantischen Beziehung hat. Die Autoren unterscheiden dabei zwischen emotionaler (Angst, Wut, Depression, Taubheit, usw.) und physischer Reaktion (Essgewohnheit, Schlaf, Gewicht, Panik, Immunsystem). Hier wird nur auf erstere fokussiert, welche mit *ER* abgekürzt wird. Dazu wurde mit erlösfreien Online-Umfragen unter anderem erfragt, ob die Person eine Trennung erlebt hat und wie sie ihren emotionalen Trennungsschmerz von 0 (keine Schmerzen) bis 10 (unerträglich) einstuft. An der Studie haben $N_{\text{Frau}} = 2695$ Frauen und $N_{\text{Mann}} = 1409$ Männer mitgemacht, welche eine ER von $M_{\text{Frau}} = 6.81, SD_{\text{Frau}} = 2.53$ und $M_{\text{Mann}} = 6.56, SD_{\text{Mann}} = 2.6$ respektive aufwiesen.

6.1 Was ist das Problem der Stichprobenziehung?

In der Stichprobe kann also ein kleiner geschlechterspezifischer Mittelwertunterschied der ER beobachtet werden. Dieser Mittelwertunterschied könnte nun einerseits auf einen Mittelwertunterschied in der Population zurückzuführen sein, wie in Abbildung **??** links dargestellt. Hier gibt es zwei Populationen: Frauen-Population mit höheren und Männer-Population mit tieferen ER-Werten. Dies führt dazu, dass der Erwartungswert der Frauen-Population höher ist als bei Männer-Population und eine zufällige gezogene Stichprobe aus Frauen-Population auch ein höheres arithmetisches Mittel aufweist als Männer-Population.

Andererseits könnte der Mittelwertunterschied auch auf die zufällige Stichprobenziehung zurückzuführen sein, siehe Abbildung **??** rechts. In dieser Situation haben die Frauen- und die Männer-Populationen ähnliche Werte und demnach auch einen ähnlichen Erwartungswert. Beim Ziehen der Stichproben spielt der Zufall hier so, dass aus der Frauen-Population einige Beobachtungen mehr mit hohen ER-Werten ausgewählt wurden als bei der Männer-Population. Dies führt dazu, dass in den zwei Stichproben ein Unterschied im arithmetischen Mittel der ER beobachtet werden kann.

Welche dieser Situationen zutrifft kann nicht genau herausgefunden werden, da die Population nie vollständig beobachtet werden kann.

Um trotzdem eine Aussage über die Population zu treffen, kann wie bereits mehrmals gemacht, die Stichprobenziehung oft - beispielsweise 3000-mal - wiederholt werden. Dies wird unter der Annahme gemacht, dass es keinen ER-Erwartungswertunterschied zwischen der Frauen- und Männer-Population gibt. Die Verteilung der ER-Mittelwertdifferenzen dieser Stichproben ist in Abbildung **??** dargestellt.

Das Testprinzip funktioniert genau gleich wie beim t-Test für eine Stichprobe wie in Kapitel **??**. Zunächst werden die Hypothesen aufgestellt. A priori liegt keine Vermutung darüber vor, ob Männer oder Frauen eine stärkere ER zeigen. Die Null- und Alternativhypothese sind deshalb

$$H_0 : \mu_{\text{Frau}} = \mu_{\text{Mann}}$$
$$H_1 : \mu_{\text{Frau}} \neq \mu_{\text{Mann}}.$$

Dies entspricht, einfacher Arithmetik folgend,

$$H_0 : \mu_{\text{Frau}} - \mu_{\text{Mann}} = 0$$
$$H_1 : \mu_{\text{Frau}} - \mu_{\text{Mann}} \neq 0.$$

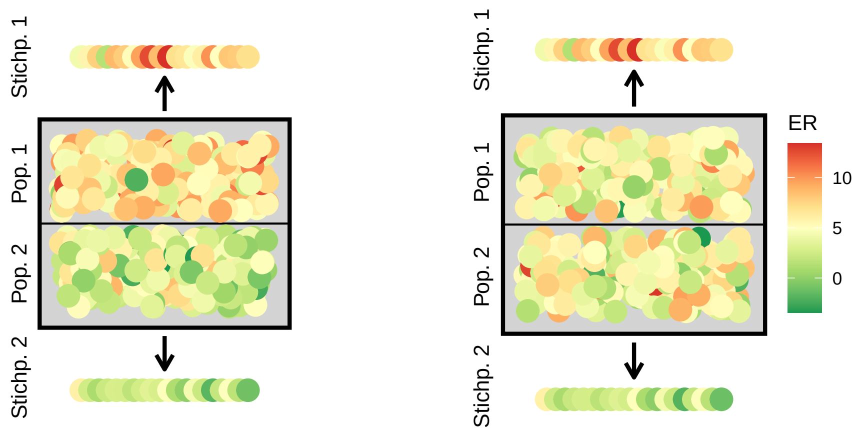


Abbildung 6.1: Links: Zwei Stichprobenziehungen aus zwei Populationen mit unterschiedlichen Mittelwerten. Rechts: Zwei Stichprobenziehungen aus einer Population, bzw. aus zwei Populationen die sich bezüglich ihrer Werte nicht unterscheiden.

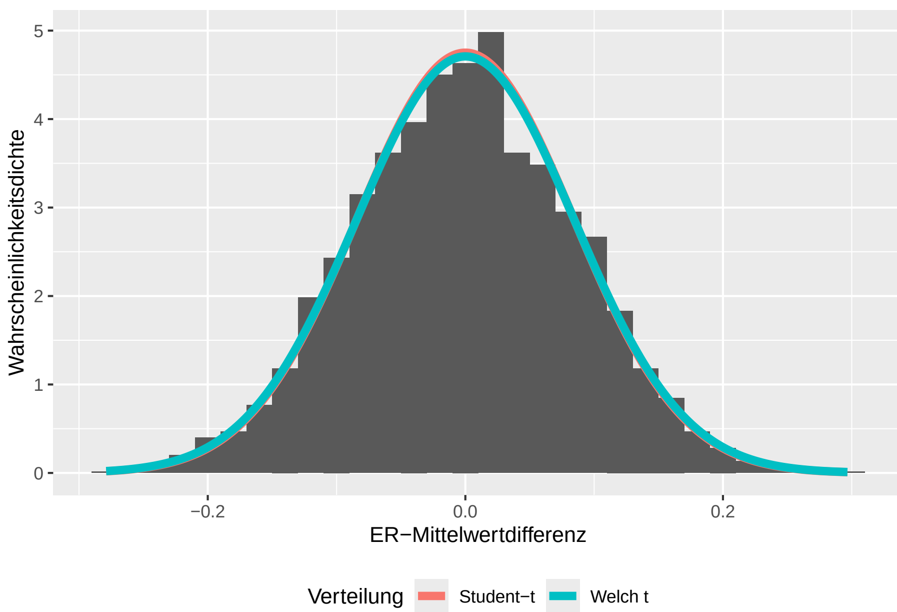


Abbildung 6.2: Verteilung simulierter ER-Mittelwertdifferenzen bei wiederholten Zufallsstichprobenziehung. Rot: Annäherung der Verteilung mit dem Student t-Test; grün: Annäherung der Verteilung durch den Welch-Test.

Es kann beobachtet werden, dass, wenn es keine Erwartungswertdifferenz gibt, die Mittelwertdifferenzen der Stichproben am häufigsten bei 0 liegen und mit zunehmender Entfernung von 0 unwahrscheinlicher werden. Dies kann wieder formalisiert werden indem die 5% unwahrscheinlichsten Werte (2.5% links und 2.5% rechts) zum Ablehnungsbereich erklärt werden und entspricht der roten Fläche in Abbildung **??** links. Die tatsächlich beobachtete Mittelwertdifferenz (schwarze Linie) liegt im Ablehnungsbereich. Dies bedeutet dass sich die Erwartungswertdifferenz bei Signifikanzniveau 5% signifikant von 0 unterscheidet. Dies ist äquivalent zu der Aussage, dass sich die ER-Erwartungswerte der Männer und Frauen signifikant unterscheidet.

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

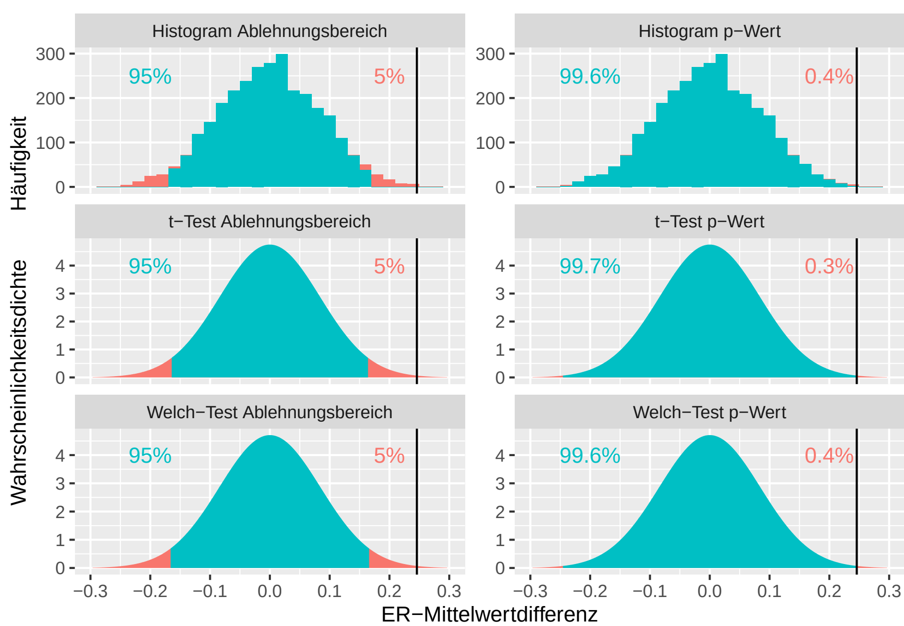


Abbildung 6.3: TODO

Ebenfalls kann erneut der p-Wert berechnet werden. Dieser entspricht hier allen ER-Mittelwertdifferenzen, welche *extremer* als die beobachtete Mittelwertdifferenz 0.25 sind. Da die Hypothesenstellung hier zweiseitig ist, bedeutet extremer hier wieder grösser als 0.25 oder kleiner als −0.25. Der *p*-Wert entspricht dem Anteil der roten Fläche in Abbildung **??** rechts an der Gesamtfläche und beträgt 0.004.

Die Verteilung der Mittelwertdifferenzen unter der Annahme, dass die Nullhypothese wahr ist, kann wieder mit einer Kurve angenähert werden. Diese Annäherung hat den Vorteil, dass der Ablehnungsbereich und der p-Wert abgeschätzt werden kann, ohne dass dazu das Experiment wiederholt werden muss. Für die Annäherungskurve gibt es zwei Optionen, welche dann entsprechenden Tests ihre Namen geben: der Zweistichproben-t-Test nach Student und der Welch Test.

6.1.1 Erwartungswertunterschied Zweistichproben-t-Test nach Student

Der **Zweistichproben-t-Test** setzt voraus, dass die beiden Populationen eine ähnliche Varianz oder äquivalent eine ähnliche Standardabweichung haben. Dazu später mehr. Ist dies gegeben, so kann die Teststatistik mit

$$t = \frac{\bar{x_1} - \bar{x_2} - \omega_0}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{6.1}$$

berechnet werden, wobei $\omega_0 = \mu_1 - \mu_2$ der Erwartungswertdifferenz entspricht und in unserem Fall 0 beträgt. Wenn die Nullhypothese wahr ist, so ist diese Teststatistik bei wiederholter Stichprobenziehung t-verteilt bei $df = n_1+n_2-2$ Freiheitsgraden.

Die rote Linie in Abbildung **??** zeigt, dass die Annäherung durch den Zweistichproben-t-Test nach Student die Verteilung der Mittelwertdifferenzen ziemlich gut trifft.

6.1.2 Erwartungswertunterschied Welch Test

Beispiel 6.2 (Emotionaler Stroop-Test bei posttraumatischer Belastungsstörung.)**.** Analog zum klassischen Stroop-Test werden bei einem emotionalen Stroop-Test *EST* Testpersonen gebeten die Farben verschiedener ausgeschriebener Wörter zu erkennen. Die Wörter sind beim emotionalen Stroop-Test entweder emotional aufgeladen (Bombe, Schweiss, Faustschlag, ...) oder neutral (Tisch, Weg, Bahn, ...) für die Testpersonen (**?**). Gemessen wird dabei die Reaktionsgeschwindigkeit *RT* in Millisekunden. In einem Versuch wollten **?** herausfinden, ob von posttraumatischer Belastungsstörung betroffene Veteranen *PTSD* andere EST-Resultate erzielen als nicht betroffene *non-PTSD*. Die durchschnittliche Reaktionszeit der 26 von PTSD betroffenen Veteranen lag bei $M = 741$ ms ($SD = 226.8$) und bei den 16 nicht von PTSD betroffenen Veteranen bei $M = 636.9$ ms ($SD = 106.1$).

Es wird keine Annahme über die Richtung einer eventualen Mittelwertdifferenz angenommen. Die Hypothesen sind deshalb zweiseitig formuliert und lauten

$$H_0 : \mu_{\text{PTSD}} = \mu_{\text{non-PTSD}}$$
$$H_1 : \mu_{\text{PTSD}} \neq \mu_{\text{non-PTSD}}.$$

In diesem Beispiel sind die Standardabweichungen und demnach auch die Varianzen der Reaktionszeiten in den beiden Gruppen sehr unterschiedlich. Wenn das Experiment wiederum wiederholt wird, kann der Verteilung der Mittelwertdifferenzen entnommen werden, dass der Zweistichproben-t-Test nach Student diese Verteilung nicht gut abbildet. Die rote Linie in Abbildung **??** liegt mittig zu hoch und an den Enden zu tief Wird diese Annäherung in diesem Fall verwendet, dann besteht die Gefahr, dass ein signifikanter Mittelwertunterschied nicht erkannt wird.

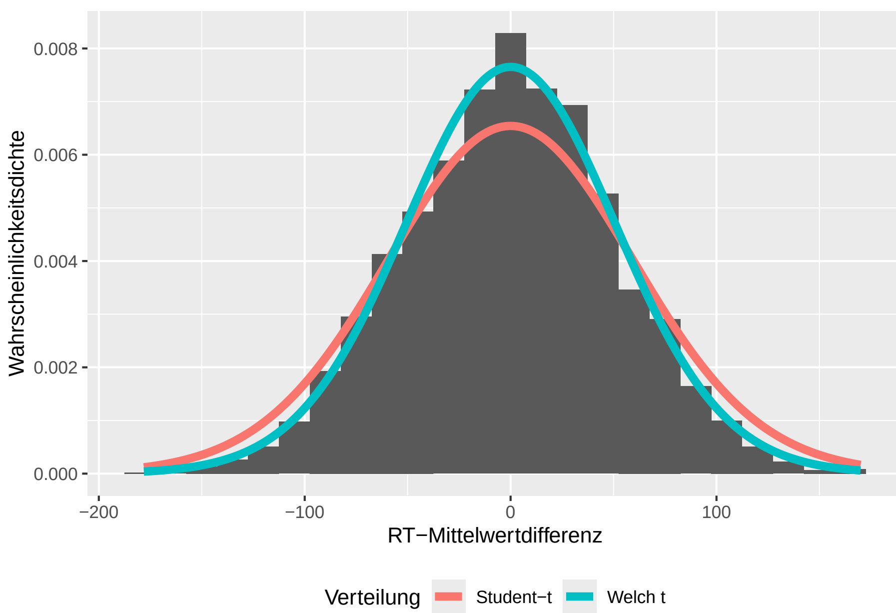


Abbildung 6.4: Verteilung simulierter RT-Mittelwertdifferenzen bei wiederholten Zufallsstichprobenziehung. Rot: Annäherung der Verteilung mit dem Student t-Test; grün: Annäherung der Verteilung durch den Welch-Test.

Für diesen Fall wurde von **?** eine alternative Annäherung an die Verteilung der Mittelwertdifferenzen gefunden, nämlich

$$t = \frac{\bar{x_1} - \bar{x_2} - \omega_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{6.2}$$

Die so berechnete Teststatistik t ist t-Verteilt bei approximativ

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{(n_1-1)n_1^2} + \frac{s_2^4}{(n_2-1)n_2^2}}$$

Freiheitsgraden und ein damit durchgeführter Test wird **Welch-Test** genannt. Sie nähert die Verteilung der Mittelwertdifferenzen trotz unterschiedlicher Gruppenvarianzen gut an, siehe grüne Linie in Abbildung **??**.

Der Zweistichproben-t-Test und der Welch-Test sind also zwei Testvarianten, um zu testen, ob der Erwartungswert in zwei Gruppen unterschiedlich ist. Dabei hat sich gezeigt, dass der Welch-Test die wahre Verteilung besser annähert als der Zweistichproben-t-Test, wenn die beiden Gruppen unterschiedliche Varianzen aufweisen, siehe Abbildung **??**. Wenn beide Varianzen ungefähr gleich sind, so geben beide Tests jedoch ähnlich gute Resultate, siehe Abbildung **??**. Es wird deshalb empfohlen immer den Welch-Test durchzuführen (**?**). Ein Vergleich der Abbildungen **??** und **??** zeigt auch, dass der Unterschied von Ablehnungsbereich und p-Wert beim im Falle der ähnlichen Varianzen gering und im Falle der unterschiedlichen Varianzen augenscheinlich wird.

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

6.2 Effektstärken

In den Formeln (**??**) und (**??**) kann beobachtet werden, dass mit zunehmenden Stichprobengrössen der Gruppen der Nenner immer kleiner und damit die Teststatistik t für eine gleichbleibende Mittelwertdifferenz immer grösser wird. Dies bedeutet, dass auch kleine Mittelwertdifferenzen bei grossen Stichprobengrösse signifikanten - also nicht auf die zufällige Stichprobenziehung zurückzuführenden - Unterschied darstellen. Beim Trennungsschmerzbeispiel ist der Mittelwertunterschied von 0.25 gering. Dies trotz dem p-Wert des Welch-Test von $p = .004$, welcher auf einen stark signifikanten Mittelwertunterschied hindeutet. Umgekehrt bei der posttraumatischen Belastungsstörung: Hier ist der Mittelwertunterschied mit 104 ms substanziell, aber der p-Wert des Welch-Test von $p = .052$ deutet knapp auf keine signifikante Mittelwertdifferenz hin.

Würde die Relevanz des beobachteten Effekts mit der Mittelwertsdifferenz gemessen, dann wäre, analog zu Kapitel **??**, dieses Mass wieder abhängig von der Einheit. Um dies zu verhindern, wird die Mittelwertdifferenz wieder durch die Standardabweichung geteilt. Für die konkrete Berechnung der Effektstärke gibt es verschiedene Methoden, wovon drei hier vorgestellt werden:

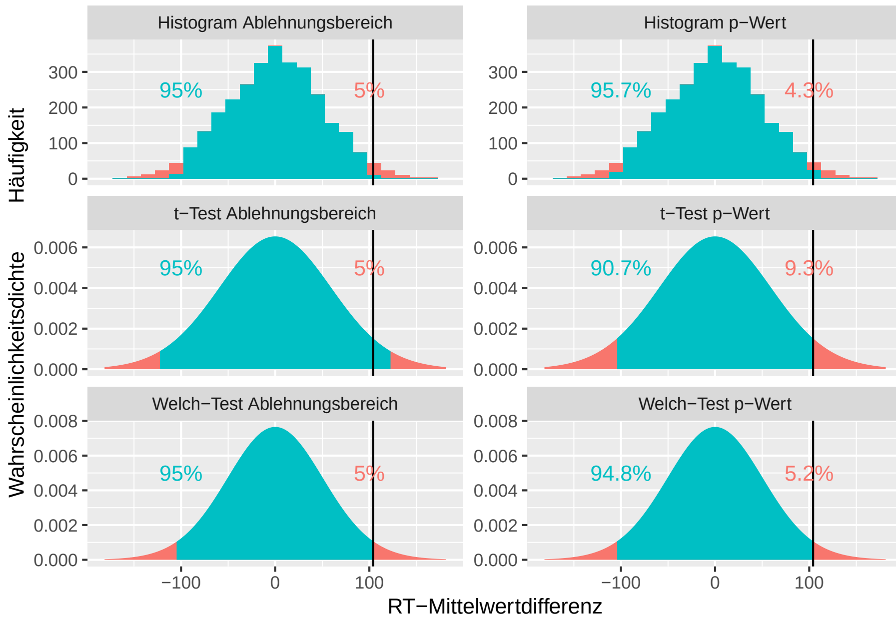


Abbildung 6.5: TODO

- Bei **Cohens d für Zweistichproben-t-Test** (Symbol d) wird die Mittelwertdifferenz durch das gewichtete Mittel der Standardabweichungen geteilt.

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

  Diese Formel entpricht dem `Cohens d` für den Zweistichproben-t-Test in `Jamovi`.
- Bei **Hedges** g (Symbol g) handelt es sich um eine um einen Faktor korrigierte Version von Cohens d für den Zweistichproben-t-Test.

$$g = \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) d$$

  Hedges g ist genauer als Cohens d bei kleinen Stichprobengrössen und gleich wie Cohens d für grosse Stichproben. Es kann daher immer Hedges g verwendet werden. Diese Formel wird für den Zweistichproben-t-Test verwendet und ist besser geeignet als d oben - ein Unterschied ist jedoch nur bei kleinen Stichproben ersichtlich. Hedges g wird in `Jamovi` nicht standardmässig ausgegeben und muss händisch berechnet werden.
- Bei **Cohens d für den Welch-Test** (Symbol d) wird die Mittelwertdif-

ferenz durch die mittlere Standardabweichungen geteilt.

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

Diese Formel entpricht dem `Cohens d` für den Welch-Test in `Jamovi`.

– **Glass** Δ (gr. delta, Symbol Δ) wir nur bei Experimenten verwendet. Dabei wird die Mittelwertdifferenz durch die Standardabweichung der Kontrollgruppe dividiert, weil angenommen wird, dass die Kontrollgruppe repräsentativer für die Popoulation ist.

$$\Delta = \frac{\bar{x}_{\text{Experiment}} - \bar{x}_{\text{Kontroll}}}{s_{\text{Kontroll}}}$$

Glass Δ wird in `Jamovi` nicht standardmässig ausgegeben und muss händisch berechnet werden.

Da es sich bei beiden Beispielen nicht um Experimente handelt, weil weder das Geschlecht noch die posttraumatische Belastungsstörung zufällig zugeorndet wurde, ist hier Glass Δ keine sinnvolle Effektgrösse. Aus diesem Grund wird für die Effektstärkenberechnung bei beiden Beispielen Cohens d für den Welch-Test verwendet. Das berichten der Testresultate kann deshalb wie folgt aussehen:

> Ein zweiseitiger Welch-Test ergibt, dass die durchschnittliche emotionale Antwort ER bei einer Trennung bei Männern ($M = 6.56$, $SD = 2.6$, $N = 1409$) signifikant anders ist als bei Frauen ($M = 6.81$, $SD = 2.53$, $N = 2695$), $t(2786.7) = -2.9$, $p = .004, d = -0.1$.

> Ein zweiseitiger Welch-Test ergibt, dass die durchschnittliche Reaktionszeit beim emotionalen Stroop-Test bei Veteranen ohne PTSD ($M = 636.86$, $SD = 106.08$, $N = 16$) nicht signifikant anders ist als bei Menschen mit PTSD ($M = 740.98$, $SD = 226.81$, $N = 26$), $t(37.9) = -2.01$, $p = .052, d = -0.588$.

Beim Trennungsschmerz handelt es sich um einen schwachen, bei der Reaktionszeit auf den EST um einen mittleren Effekt.

Achtung

Hinweis.

- Die Namensgebung von diesen Berechnungsarten und insbesondere, was unter Cohens d verstanden wird variiert beträchtlich und es empfiehlt sich immer die genaue Berechnungsart zu überprüfen.
- In `Jamovi` wird für den Zweistichproben-t-Test und den Welch-Test eine unterschiedliche Effektstärke angegeben.

6.3 Testvoraussetzungen

Damit der Zweistichproben-t-Test und der Welch Test durchgeführt werden dürfen, müssen einige Voraussetzungen eingehalten werden.

1. Das Merkmal muss intervallskaliert sein.
2. Die Beobachtungen müssen einer Zufallsstichprobe der jeweiligen Gruppe entsprechen.
3. Die Beobachtungen beider Gruppen müssen einer Normalverteilung entstammen oder die Anzahl der Beobachtungen muss gross genug sein. Häufig wird die Faustregel mehr als 30 Beobachtungen pro Gruppe verwendet.
4. Für den Zweistichproben-t-Test müssen die Varianzen gleich sein. Für den Welch Test gilt diese Voraussetzung nicht.

6.4 Übungen

Übung 6.1.

Mit dem Bobo-Doll-Experiment sollte die Übertragung von Aggression durch Imitation aggressiver Modelle nachgewiesen werden. An der Studie nahmen 48 Kinder im Alter von drei bis sechs Jahren teil. Die Kinder wurden in zwei Gruppen eingeteilt: eine mit aggressivem Modell und eine mit nicht-aggressivem Modell. In der aggressiven Bedingung sahen die Kinder, wie eine erwachsene Person (das Modell) eine Bobo-Puppe aggressiv behandelte, während in der nicht-aggressiven Bedingung das Modell ruhig mit der Puppe spielte. Nach der Beobachtungsphase wurden die Kinder einzeln in einen Spielraum geführt, der ähnliche Spielzeuge wie im Experiment enthielt, einschliesslich der Bobo-Puppe. Die Forscher beobachteten und notierten die Anzahl gezeigter aggressiven Handlungen gegenüber der Bobo-Puppe. Inspiriert von **?**.

Beantworten Sie die Frage, ob aggressives Verhalten Erwachsener von Kindern imitiert wird anhand der folgenden Teilfragen:

a) Die Kinder welcher Gruppe zeigen ein aggressiveres Verhalten? Argumentieren Sie mit zahlen.
b) Kann die Aussage aus a) von der Stichprobe auf die Population verallgemeinert werden? Stellen Sie zweiseitige Testhypothesen für den Erwartungswert auf.
c) Führen Sie den statistischen Test mit Jamovi durch und berechnen Sie eine angemessene Effektstärke. Berichten und interpretieren Sie das Testresultat.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

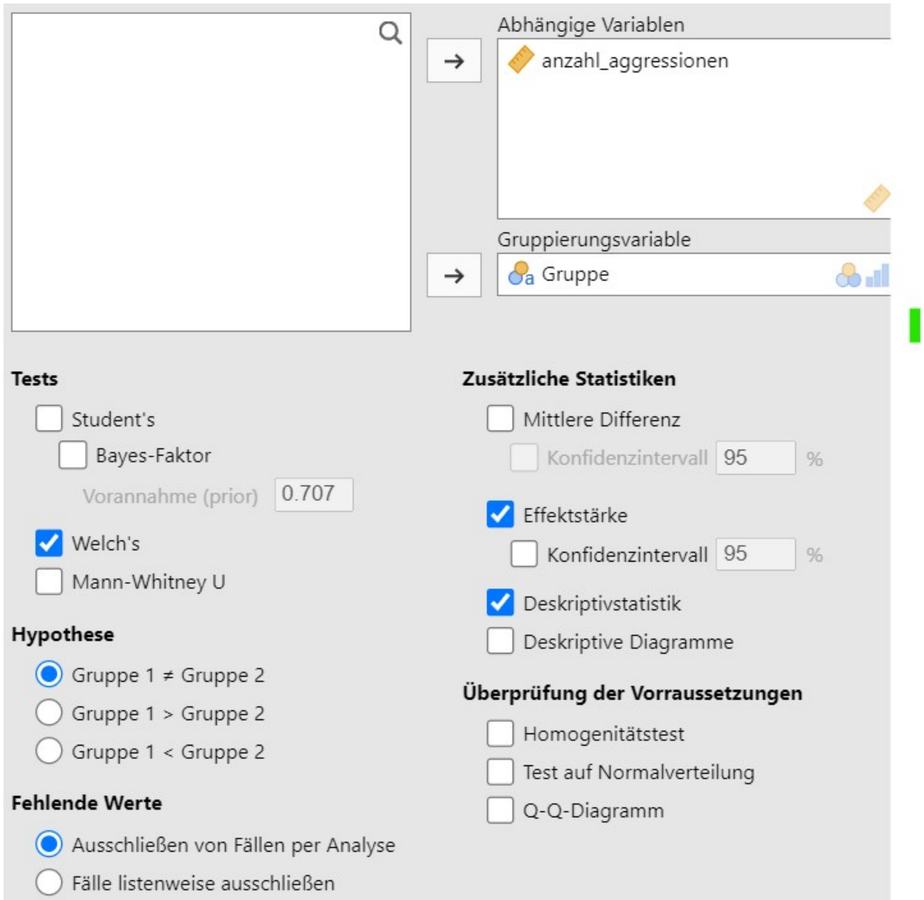


Abbildung 6.6: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
anzahl_aggressionen	Welch's t	10.4544	24.3589	< .001	Cohens d	3.0179

Anmerkung. H_a $\mu_{\text{aggressiv}} \neq \mu_{\text{nicht aggressiv}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
anzahl_aggressionen	aggressiv	24	10.6266	10.6173	2.6231	0.5354
	nicht aggressiv	24	4.9467	4.9942	0.4510	0.0921

Abbildung 6.7: Jamovi Ausgabe.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können und beide Teilfragen beantwortet werden.

a) In der Stichprobe ist der durchschnittliche Anzahl gezählter Aggressionen in der Gruppe mit aggressiven Modellen mit $M = 10.63$ höher als in der Gruppe mit nicht Aggressiven Modellen $M = 4.95$. Es könnte sein, dass der gefundene Mittelwertunterschied auf die zufällige Stichprobenziehung zurückzuführen ist. Um dieses Risiko zu quantifizieren und damit einzuschätzen, ob das Ergebnis auch für die Population gelten könnte, kann ein statistischer Test durchgeführt werden.
b) Es soll gezeigt werden, dass sich der durchschnittlich beobachtete Anzahl aggressiver Handlungen der Kinder in der Gruppe mit aggressivem Modell anders ist als in der Gruppe mit nicht aggressivem Modell. Die Alternativhypothese lautet also $H_1 : \mu_{\text{Aggressiv}} \neq \mu_{\text{Nicht aggressiv}}$. Die Nullhypothese dagegen sagt, dass beide Gruppen durchschnittlich gleich viele aggressive Handlungen begehen, also $H_0 : \mu_{\text{Aggressiv}} = \mu_{\text{Nicht aggressiv}}$.
c) Es werden Mittelwerte von einer intervallskalierten Variabel über zwei Gruppen verglichen. Als statischer Test kommt demnach der Zweistichproben-t-Test oder der Welch-Test infrage. Aufgrund der genaueren Testergebnisse wird immer der Welch-Test bevorzugt und dieser in folge durchgeführt und berichtet. Ein zweiseitiger Welch-Test ergibt, dass der durchschnittliche Anzahl Aggressionen in der Gruppe mit aggressivem Modell ($M = 10.63$, $SD = 2.62$, $N = 24$) signifikant anders ist als in der Gruppe mit nicht aggressivem Modell ($M = 4.95$, $SD = 0.45$, $N = 24$), $t(24.4) = 10.45$, $p < 0.001, \Delta = 12.593$. Da es sich um ein Experiment handelt ist hier die Effektstärke Glass Δ angebracht. Als Kontrollgruppe wurde die nicht aggressive Gruppe verwendet. Die Effektstärke ist als gross einzustufen.

Übung 6.2.

In den 1970er Jahren hat eine Gruppe um **?** Versuche durchgeführt zu neuen Lehrmethoden. Insbesondere wurde dabei das sogenannte Gruppenpuzzle `gruppenpuzzle`, eine Lernform bei welcher die Lernenden den Inhalt mit und in Abhängigkeit voneinander erarbeiten, mit dem traditionellen Frontalunterricht `traditionell` verglichen. Die Forschenden wollten unter anderem Herausfinden, ob sich die Gruppenpuzzleteilnehmende nach dem Unterricht besser oder schlechter mochten (*liking*), als traditionell unterrichtete Lernende. Fiktive Daten zu dem Experiment sind als `06-exr-gruppenpuzzle.sav` verfügbar.

a) Stellen Sie die Testhypothesen auf für einen zweiseitigen Welch-Test.
b) Führen Sie den Test durch und berichten Sie das Resultat.
c) Erklären Sie den Wert der Statistik, der Freiheitsgrade, des p-Werts und der Effektstärke respektive.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

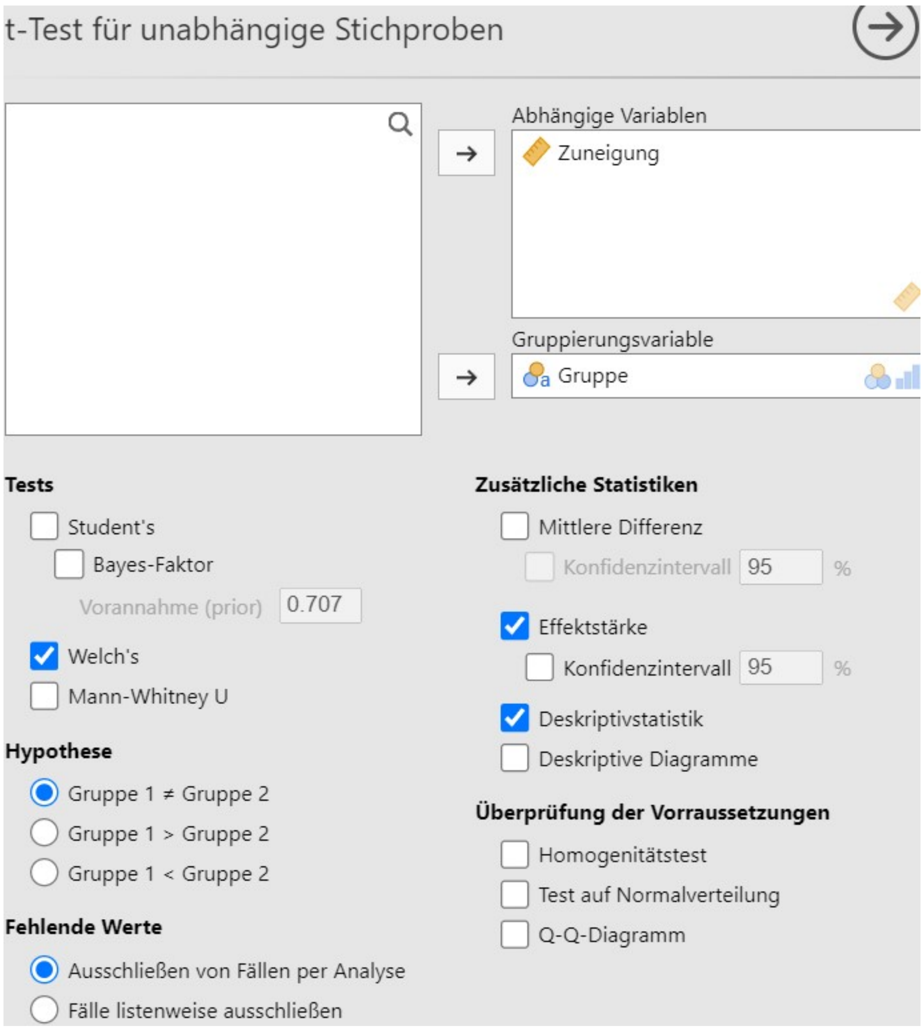


Abbildung 6.8: Jamovi Eingabe.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können und die Teilfragen beantwortet werden.

a) Angenommen die Zuneigung zur Personen der einen Gruppe ist unabhängig von der Lehrmethode, dann sollten beide Gruppen im durchschnitt denselben Erwartungswert μ bei der Zuneigung haben. Die Nullhypothese ist also $H_0 : \mu_{\text{Gruppenpuzzle}} = \mu_{\text{Traditionell}}$. Ein Unterschied dazu wäre,

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Zuneigung	Welch's t	3.6867	35.6158	< .001	Cohens d	1.0436

Anmerkung. H_a $\mu_{\text{gruppenpuzzle}} \neq \mu_{\text{traditionell}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Zuneigung	gruppenpuzzle	35	5.1143	5.0997	0.2538	0.0429
	traditionell	21	4.8168	4.8533	0.3132	0.0683

Abbildung 6.9: Jamovi Ausgabe.

wenn es die Lernenden der beiden Gruppen einen unterschiedlichen Erwartungswert aufweisen, formell $H_1 : \mu_{\text{Gruppenpuzzle}} \neq \mu_{\text{Traditionell}}$.

b) Ein zweiseitiger Welch-Test ergibt, dass die durchschnittliche Zuneigung in der Gruppenpuzzlegruppe ($M = 5.11$, $SD = 0.25$, $N = 35$) signifikant anders ist als in der traditionell unterrichteten Gruppe($M = 4.82$, $SD = 0.31$, $N = 21$), $t(35.6) = 3.69$, $p < 0.001, \Delta = -1.172$.

c) Die Statistik von 3.69 ist ein Wert, welcher eine Verteilung wie in **??** aufweist. Diese Verteilung weist die Statistik auf, wenn das Experiment oft wiederholt wird und die Nullhypothese wahr ist. Die Verteilung zweigt, dass der beobachtete Wert 3.69 selten zufällig vorkommt (tiefer Wert der Linie weist auf eine tiefe Wahrscheinlichkeit der Statistik hin). Die Freiheitsgrade 35.6 bestimmen die Form der oben referenzierten Verteilung. Wo bei kleinen Freiheitsgraden die beobachtete Statistik noch mit einer nicht allzukleinen Wahrscheinlichkeit beobachtet werden kann (vgl. $df = 1$ in der Abbildung), so ist es bei dieser Anzahl Freiheitsgrade sehr selten (vgl. `Normalverteilung` in der Abbildung). $p < 0.001$ bedeutet, dass der p-Wert kleiner als $0.001 = 0.1\%$ ist. Damit ist die Wahrscheinlichkeit den Statistik-Wert 3.69 oder einen extremeren Wert im Sinne der Alternativhypothese zu beobachten, gegeben dass die Nullhypothese wahr ist, kleiner als 0.1% also sehr selten. Da der p-Wert kleiner ist als 5% ist wird geschlossen, dass die Annahme, dass die Nullhypothese wahr ist, wahrscheinlich falsch ist. Die Effektstärke von $\Delta = 0.95$ bedeutet, dass hier ein Mittelwertunterschied von ungefähr 0.95 Standardabweichungen des Merkmals Zuneigung entspricht. Dies heisst, auf der Skala des Merkmals ist der Mittelwertunterschied gross oder anders gesagt: es handelt sich um einen starken Effekt. Das Vorzeichen hängt von der Gruppenbeschriftung ab und hat keine spezielle Bedeutung.

Übung 6.3.

Studierende wollen herausfinden, ob Entspannungsmusik ohne Text oder Musik mit Text einen unterschiedlichen Einfluss auf die Merkfähigkeit haben. Dazu lernen die Studienteilnehmenden während 10 Minuten Wortsilben ohne semantische Bedeutung auswendig und geben diese nach einer Latenzzeit wider. Die Beschallungsart wird den Studienteilnehmenden zufällig zugeordnet. Die Anzahl korrekt memorisierte Wortsilben sind im Datensatz `06-exr-music-memory.sav` verfügbar.

a) Stellen Sie die Testhypothesen auf für einen zweiseitigen Welch-Test.
b) Führen Sie den Test durch und berichten Sie das Resultat.
c) Erklären Sie den Wert der Statistik, des p-Werts und der Effektstärke respektive.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

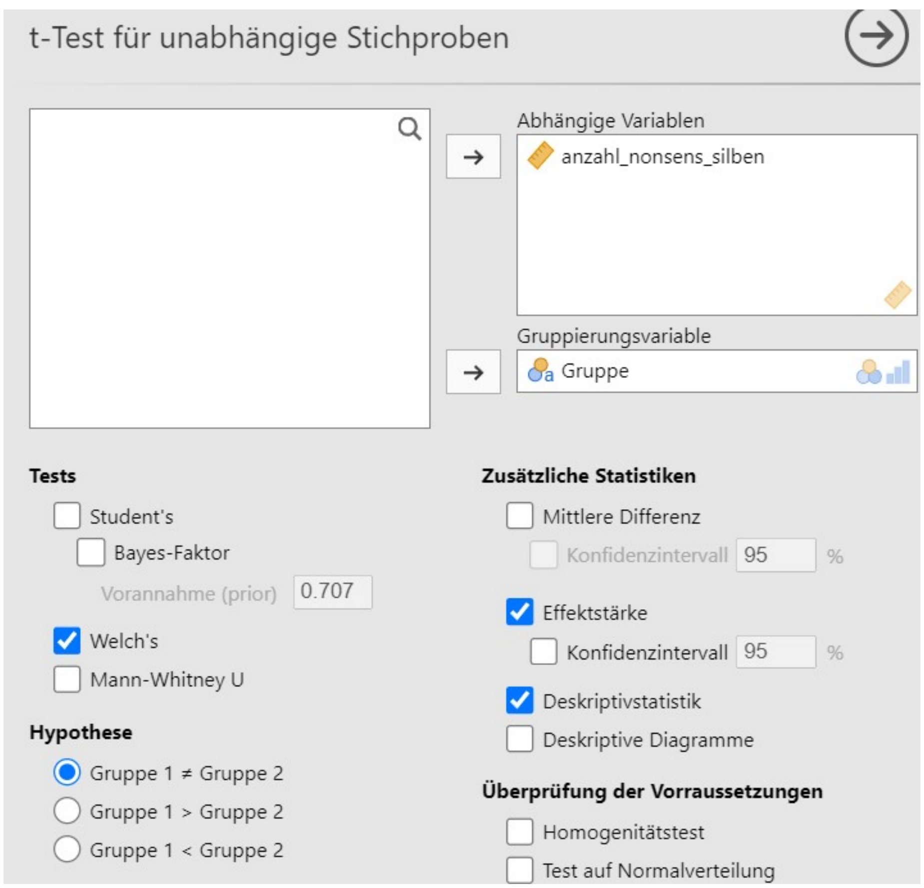


Abbildung 6.10: Jamovi Eingabe.

Dies produziert das Analyseergebnis in Abbildung **??**.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
anzahl_nonsens_silben	Welch's t	−0.6561	54.6398	0.514	Cohens d	−0.1527

Anmerkung. H_a $\mu_{\text{mit\_text}} \neq \mu_{\text{ohne\_text}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
anzahl_nonsens_silben	mit_text	35	8.0019	8.0678	1.6931	0.2862
	ohne_text	43	8.2176	8.1990	1.0596	0.1616

Abbildung 6.11: Jamovi Ausgabe.

Damit können die Teilfragen beantwortet werden.

a) Die Nullhypothese besagt, dass die durchschnittliche Anzahl gemerkter Wortsilben beim Lernen mit oder ohne Musik gleich ist, also $H_0 : \mu_{\text{Musik mit Text}} = \mu_{\text{Musik ohne Text}}$. Die Alternativhypothese besagt, dass sich die durchschnittliche Anzahl gemerkter Wortsiblen mit oder ohne Musik unterscheiden $H_1 : \mu_{\text{Musik mit Text}} \neq \mu_{\text{Musik ohne Text}}$.
b) Ein zweiseitiger Welch-Test ergibt, dass die durchschnittliche Anzahl gemerkter Wortsilben beim Lernen mit Musik mit Text ($M = 8$, $SD = 1.69$, $N = 35$) nicht signifikant anders ist als beim Lernen mit Musik ohne Text ($M = 8.22$, $SD = 1.06$, $N = 43$), $t(54.6) = -0.66$, $p = .514, d = -0.153$.
c) Die Statistik von -0.66 ist ein Wert, welcher eine Verteilung wie in **??** aufweist. Diese Verteilung weist die Statistik auf, wenn das Experiment oft wiederholt wird und die Nullhypothese wahr ist. Die Verteilung zeigt, dass der beobachtete Wert -0.66 oft zufällig vorkommt (hoher Wert der Linie weist auf eine hohe Wahrscheinlichkeit der Statistik hin). $p = .514$ bedeutet, dass die Wahrscheinlichkeit den Statistik-Wert -0.66 oder einen extremeren Wert im Sinne der Alternativhypothese zu beobachten, gegeben dass die Nullhypothese wahr ist, nicht aussergewönlich erscheint. Da der p-Wert grösser ist als 5% ist, kann keine Aussage zur Wahrheit oder Falschheit der Nullhypothese getroffen werden. Die Effektstärke von $d = -0.153$ bedeutet, dass hier ein Mittelwertunterschied von ungefähr -0.153 Standardabweichungen des Merkmals Anzahl gemerkter Wortsiblen entspricht. Dies heisst, auf der Skala des Merkmals ist der Mittelwertunterschied klein oder anders gesagt: es handelt sich um einen schwachen Effekt. Das Vorzeichen hängt von der Gruppenbeschriftung ab und hat keine spezielle Bedeutung.

Übung 6.4.

Die Gesellschaft für Hypnose will unter Beweis stellen (Signifikanzniveau $\alpha = 5\%$), dass ein neues Hypnoseverfahren eine schmerzlindernde Wirkung hat. Dazu werden Probanden zufällig und doppelblind in zwei Gruppen eingeteilt. Eine Gruppe erhält die Behandlung mit dem neuen Hypnoseverfahren, die andere wird einer Placebo-Behandlung unterzogen. Nach der Behandlung wird das Schmerzempfinden auf einer Skala von 1 bis 10 gemessen. Die Daten beider Versuchsgruppen stellen sich als normalverteilt heraus. Die erhobenen Daten sind unter `06-exr-hypnose.sav` abgelegt.

a) Beschreiben Sie die beiden Stichproben deskriptiv. Hat die neue Behandlungsmethode einen Vorteil gegenüber der Placebo-Behandlung in der Stichprobe? Weshalb ist es sinnvoll danach noch einen statistischen Test durchzuführen?
b) Stellen Sie die Hypothesen für einen einseitigen Test auf.
c) Prüfen Sie die Hypothesen mit einem geeigneten einseitig durchgeführten statistischen Test, ob das Resultat auch auf die Population übertragen werden kann. Berichten Sie das Ergebnis.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können und beide Teilfragen beantwortet werden.

a) In der Stichprobe ist der durchschnittliche Schmerz (arithmetisches Mittel) in der Hypnose-Gruppe mit $M = 5.81$ tiefer als in der Placebo-Gruppe mit $M = 4.92$. Es könnte sein, dass der gefundene Mittelwertunterschied auf die zufällige Stichprobenziehung zurückzuführen ist. Um dieses Risiko zu quantifizieren und damit einzuschätzen, ob das Ergebnis auch für die Population gelten könnte, kann ein statistischer Test durchgeführt werden.
b) Es soll gezeigt werden, dass sich der durchschnittlich empfundene Schmerz mit der Hypnose-Behandlung tiefer liegt als mit der Placebo-Behandlung. Die Alternativhypothese lautet also $H_1 : \mu_{\text{Hypnose}} < \mu_{\text{Placebo}}$. Die Nullhypothese dagegen sagt, dass die Hypnose-Behandlung nicht besser oder sogar schlechter ist als die Placebo-Behandlung also $H_0 : \mu_{\text{Hypnose}} \geq \mu_{\text{Placebo}}$.
c) Ein einseitiger Welch-Test ergibt, dass der durchschnittliche erhobene Schmerz bei einer Behandlung mit der neuen Hypnose-Methode ($M = 4.92$, $SD = 0.39$, $N = 11$) signifikant tiefer ist als bei der Placebo-Behandlung ($M = 5.81$, $SD = 0.87$, $N = 15$), $t(20.6) = -3.5$, $p = .001, \Delta = 2.264$.

Übung 6.5.

Eine Forscherin hat die Hypothese, dass unverheiratete Ärztinnen ein weniger stabiles Umfeld haben als ihre verheirateten Kolleginnen. Das Fehlen dieser

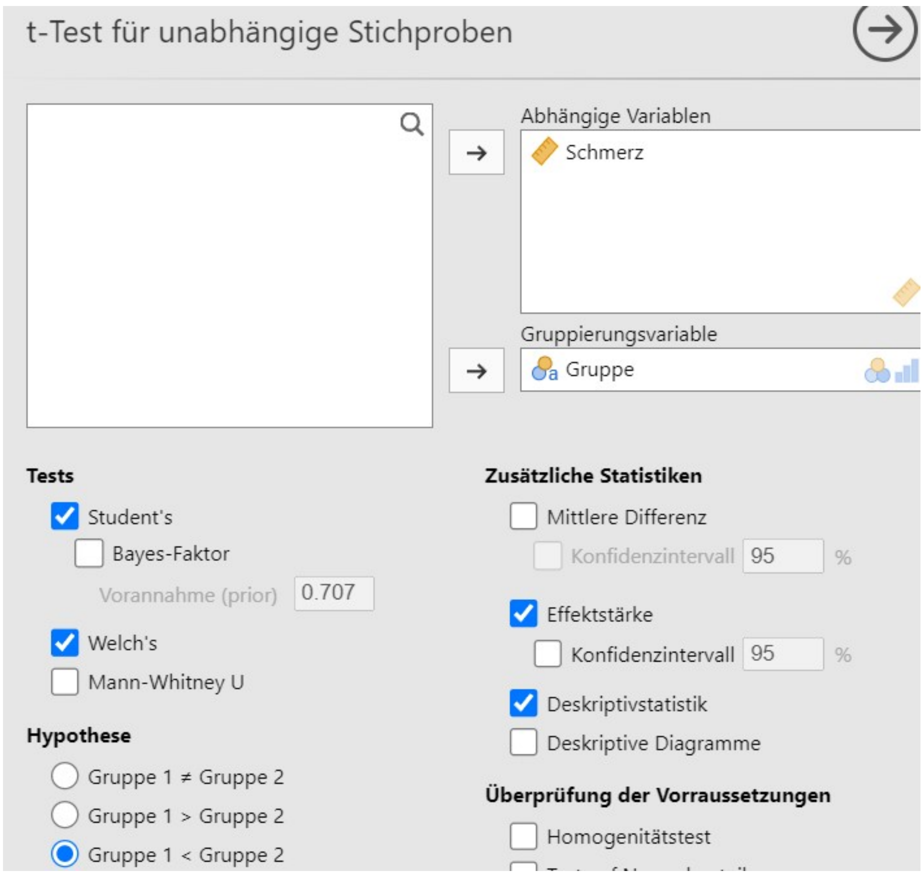


Abbildung 6.12: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Schmerz	Student's t	−3.1486 [a]	24.0000	0.002	Cohens d	−1.2499
	Welch's t	−3.5006	20.6278	0.001	Cohens d	−1.3170

Anmerkung. H_a $\mu_{hypnose} < \mu_{placebo}$

[a] Der Levene-Test ist signifikant (p < 0,05), was auf eine Verletzung der Annahme gleicher Varianzen hindeutet

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Schmerz	hypnose	11	4.9200	4.7788	0.3944	0.1189
	placebo	15	5.8130	5.6932	0.8741	0.2257

Abbildung 6.13: Jamovi Ausgabe.

Ressource führt dazu, dass unverheiratete Ärztinnen eher Burnout gefährdet sind. Um diese Hypothese zu untersuchen befragt die Forscherin in einer Umfrage zufällig verheiratete und unverheiratete Ärztinnen. Diese füllen einen Online-Fragebogen mit einem Burnout-Inventar aus, welches zu einem Burnout-score führt. Die Daten sind unter `06-exr-ehe-burnout.sav` verfügbar.

a) Wie viele verheiratete und unverheiratete haben den Fragebogen abgeschlossen?
b) Welche Gruppe hat in der Stichprobe ein höheres mittleres Burnout-Risiko?
c) Übersetzen Sie die Hypothese der Forscherin in eine Statistische Hypothese.
d) Lässt sich die Hypothese statistisch bestätigen? Berichten Sie das Testresultat.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die Teilfragen beantwortet werden.

a) Aus der Stichprobenbeschreibung kann entnommen werden, dass 51 unverheiratete und 61 verheiratete Ärztinnen den Fragebogen abgeschlossen haben.

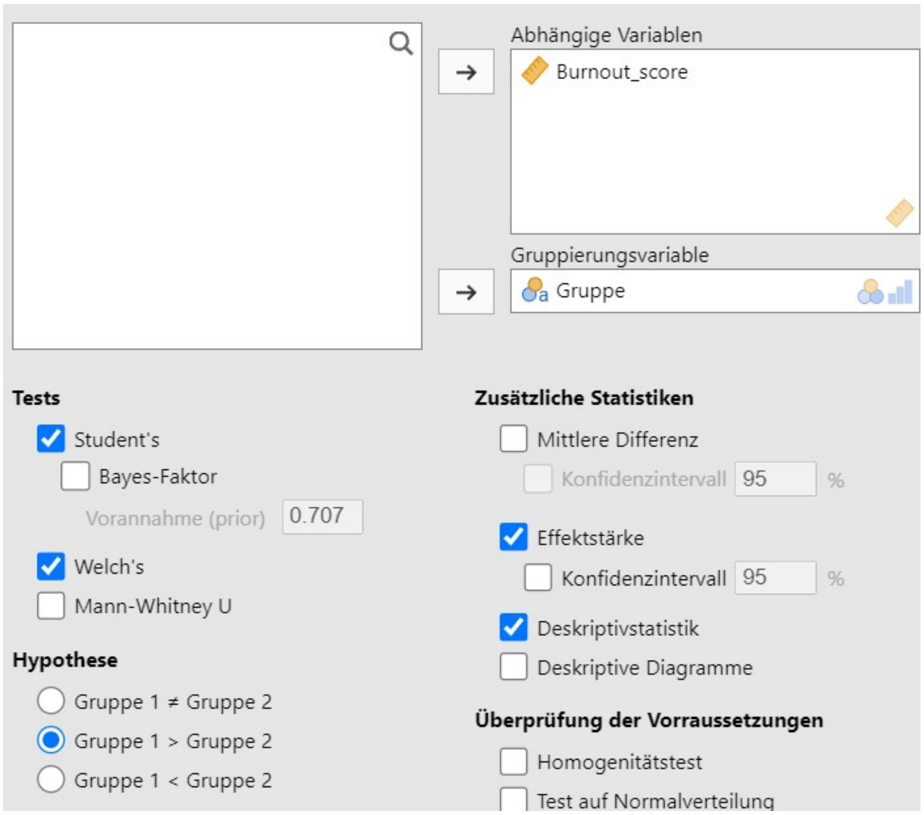


Abbildung 6.14: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Burnout_score	Student's t	2.2458	110.0000	0.013	Cohens d	0.4261
	Welch's t	2.2555	108.0926	0.013	Cohens d	0.4270

Anmerkung. H_a $\mu_{\text{Unverheiratet}} > \mu_{\text{Verheiratet}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Burnout_score	Unverheiratet	51	10.9899	11.0199	1.9173	0.2685
	Verheiratet	61	10.1507	10.1017	2.0118	0.2576

Abbildung 6.15: Jamovi Ausgabe.

b) Die unverheirateten Ärztinnen $M = 10.99$ scheiden durchschnittlich höher ab als die verheirateten Ärztinnen $M = 10.15$. Dieser Befund beschränkt sich ohne statistischen Test auf die Stichprobe. Deshalb wurde darin das Wort signifikant nicht verwendet.

c) Die Forscherin will zeigen, dass unverheiratete Ärztinnen ein durchhschnittlich höheres Burnout-Risiko haben als verheiratete und zwar nicht nur in der Stichprobe sondern auch ich der Population. Das durchschnittliche Burnout-Risiko in der Population ist der Erwartungswert des Burnout-Risiko und wird mit μ bezeichnet. Die Forscherin will also zeigen, dass $H_1 : \mu_{\text{unverheiratet}} > \mu_{\text{verheiratet}}$. Demgegenüber steht die Nullhypothese, dass dies nicht so ist oder das gar das Gegenteil der Fall sein könnte also $H_0 : \mu_{\text{unverheiratet}} \leq \mu_{\text{verheiratet}}$. Die Hypothese ist also einseitig gestellt.

d) Ein einseitiger Welch-Test ergibt, dass der durchschnittliche Burnout-Wert bei unverheirateten Ärztinnen ($M = 10.99$, $SD = 1.92$, $N = 51$) signifikant höher ist als bei verheirateten Ärztinnen ($M = 10.15$, $SD = 2.01$, $N = 61$), $t(108.1) = 2.26$, $p = .013, d = 0.427$.

Übung 6.6.

TODO: Exercise body

Lösung. TODO: solution body

Kapitel 7

Gruppenmittelwertunterschied bei einem mindestens ordinalskalierten Merkmal

Um einen Gruppenmittelwertunterschied mit dem Zweistichproben-t-Test oder dem Welch Test testen zu können muss das betrachtete Merkmal intervallskaliert und (a) die Beobachtungen beider Gruppen einer Normalverteilung entstammen oder (b) genügend Beobachtungen, normalerweise mehr als 30 pro Gruppe, vorhanden sein. Dies ist in der Realität nicht immer gegeben. Eine Alternative zu den oben genannten Tests, welche ohne diese Voraussetzungen auskommt, ist der U-Test nach Mann und Whitney. Dieser kann bei mindestens ordinalskalierten Variablen eingesetzt werden und es wird keine Verteilung der Daten vorausgesetzt.

Da der U-Test keine Verteilung voraussetzt und auch bei nicht intervallskalierten Merkmalen eingesetzt werden kann sind jedoch auch die Hypothesen leicht anders als beim Welch Test. Der U-Test testet in jedem Fall, ob die Verteilungen in den beiden Gruppen gleich sind. Unter ein paar Zusatzannahmen ist dies Äquivalent zur Hypothese, dass die beiden Populationsmediane sich entsprechen. Auf diese letzte Subtilität wird hier nicht eingegangen.

7.1 Wie stark unterscheiden sich die Mediane?

Beispiel 7.1 (Schmerzen bei Rothaarigen.)**.** Beispiel frei nach ?. Viele rothaarige Menschen haben eine höhere Schmerztoleranz. Der Mechanismus dazu ist auf das MC4R-Gen zurückzuführen, welches vor allem bei Rothaarigen vorkommt. Um dies zu testen haben Forschende die Schmerztoleranz von Mäusen mit und ohne MC4R-Gen-Variante untersucht, indem sie den sogenannten *Hot Plate Test* durchgeführt haben. Dabei werden die Mäuse auf eine erhitzte Platte gestellt und die Zeit in Sekunden gemessen, bis die Maus anfängt zu hüpfen oder sich die Pfoten zu lecken, um den Schmerz zu reduzieren. Dies bei einer maximalen Versuchszeit von 20 Sekunden. Die Daten sind unter `07-exm-red-hair-pain.sav` verfügbar. Die Beobachtete Stichprobe ergibt, dass es die $N = 13$ MC4R-Mäuse $M = 12s, SD = 1.93$ und die 15 Non-MC4R-Mäuse $M = 7.8s, SD = 2.16$ auf der heissen Platte ausgehalten haben. Werden die Daten auf die Normalverteilung getestet ergibt sich kein klares Bild. Es könnte sein, dass die Daten nicht normalverteilt sind. Ein Welch Test wäre in diesem Fall nicht angebracht. Da der U-Test keine Normalverteilung voraussetzt, kann dieser hier verwendet werden.

Beim U-Test nach Mann und Whitney, werden zunächst die Beobachtungen ungeachtet der Gruppenzugehörigkeit in eine aufsteigende Reihenfolge gebracht, siehe mittige Punkte in Abbildung **??**. Die so sortierten Beobachtungen werden nummeriert, was den sogenannten Rangnummern entspricht. Für gleiche Ränge wird den betreffenden Rängen ein mittlerer Rang zugewiesen.

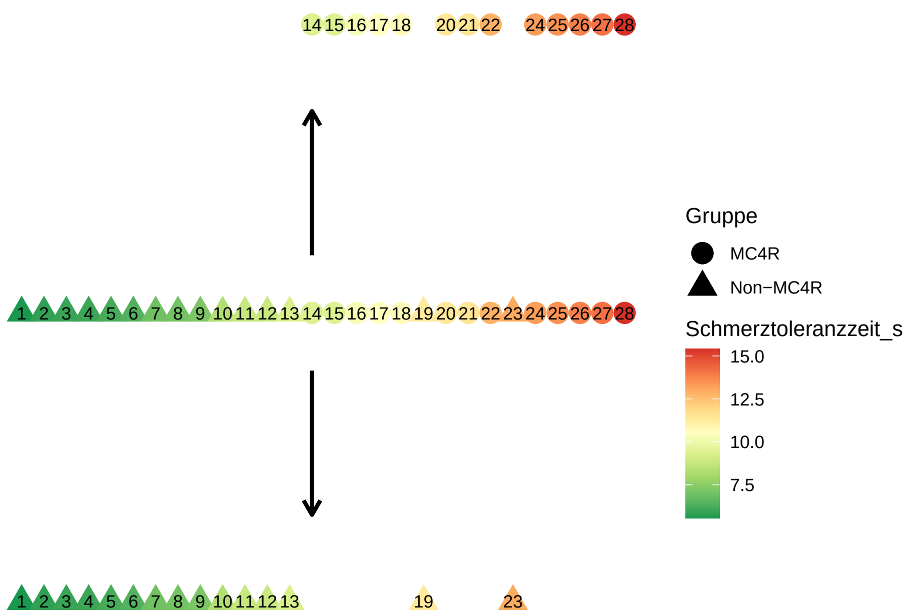


Abbildung 7.1: Mitte: Aufsteigend sortierte Beobachtungen des Merkmals Schmerztoleranz für MC4R und Non-MC4R Mäuse. Die Zahlen stehen für die Rangnummern. Oben (MC4R) und unten (Non-MC4R) stellt die Gruppenaufteilung der Rangnummern dar.

In einem zweiten Schritt werden die Beobachtungen wieder in die Gruppen aufgeteilt (siehe obere und untere Reihe in der Abbildung) und die jeweiligen Rangnummern innerhalb einer Gruppe addiert. Diese Gruppensummen werden Rangsummen genannt und sind hier

$$R_{\text{MC4R}} = 273 \quad R_{\text{Non-MC4R}} = 133.$$

Die Idee dabei ist, dass wenn sich die Messungen in den beiden Gruppen nicht oder kaum unterscheiden, dann müssten auch die Rangnummern mehr oder weniger zufällig auf die beiden Gruppen verteilt sein. Gegeben, dass die beiden Gruppen gleich gross sind, müssten in diesem Fall auch die Rangsummen ungefähr gleich gross sein. Um in einem dritten Schritt für unterschiedliche Gruppengrössen zu korrigieren, wird nun noch die kleinste Rangsumme abgezogen, welche mit den Beobachtungen erreicht werden könnte. Für die 15 Non-MC4R Beobachtungen wäre dies also $1+2+...+15 = 120$ oder kurz $n\cdot(n+1)/2 = 120$. Der so korrigierte Wert wird U-Wert genannt und ist im Beispiel

$$U_{\text{MC4R}} = 273 - 91 = 182 \quad U_{\text{Non-MC4R}} = 133 - 120 = 13.$$

Wenn es keinen Gruppenunterschied gibt, dann wären diese U-Werte nahe beieinander und beide Werte wären nicht nahe bei 0. Wenn es einen Gruppenunterschied gibt, dann sind die U-Werte weit auseinander und der kleinere der beiden Werte läge nahe bei 0. Dies macht sich der U-Test zu Nutze, indem er nun den kleineren der beiden U-Werte, hier 13 - die sogenannte Teststatistik - mit einer Referenztabelle vergleicht, wo die Wahrscheinlichkeiten für einen solchen U-Wert, gegeben dass die Nullhypothese wahr ist, hinterlegt sind. Dieser Prozess ist in `Jamovi` automatisiert und es kann direkt der p-Wert in der Ausgabe abgelesen werden, hier $p < 0.001$.

Beispiel 7.2 (Aufgeschlossenheit bei Jung und Alt.)**.** Eine Studentin will herausfinden, ob jüngere Menschen unter 30 Jahren aufgeschlossener sind als ältere Menschen mit über 30 oder genau 30 Jahren. Dazu befragt sie zufällig Leute der beiden Gruppen mit dem TIPI, welcher die Aufgeschlossenheit auf einer Skala von 1 bis 7 misst, wobei das kleinste Messintervall 0.5 Punkte beträgt. Die Daten sind also eher ordinal als intervallskaliert. In der Stichprobe waren $N = 13$ junge mit einer Aufgeschlossenheit von $M = 5.38, SD = 0.92$ und $N = 7$ alte mit einer Aufgeschlossenheit von $M = 5, SD = 0.87$. Die Daten sind unter `07-exm-aufgeschlossenheit-jung-alt.sav` verfügbar. (Beispiel frei erfunden.)

Auch in diesem Fall werden für den U-Test zunächst die Beobachtungen gruppenunabhängig aufsteigend sortiert wie in der Mitte der Abbildung **??** dargestellt. Danach werden Rangnummern vergeben und die Beobachtungen wieder in ihre Gruppen geteilt, siehe die Reihen oben und unten der Abbildung. Die Ränge scheinen zufällig in die beiden Gruppen zu fallen.

Nun werden die Ränge innerhalb einer Gruppe addiert

$$R_{\text{Jung}} = 143 \quad R_{\text{Alt}} = 67.$$

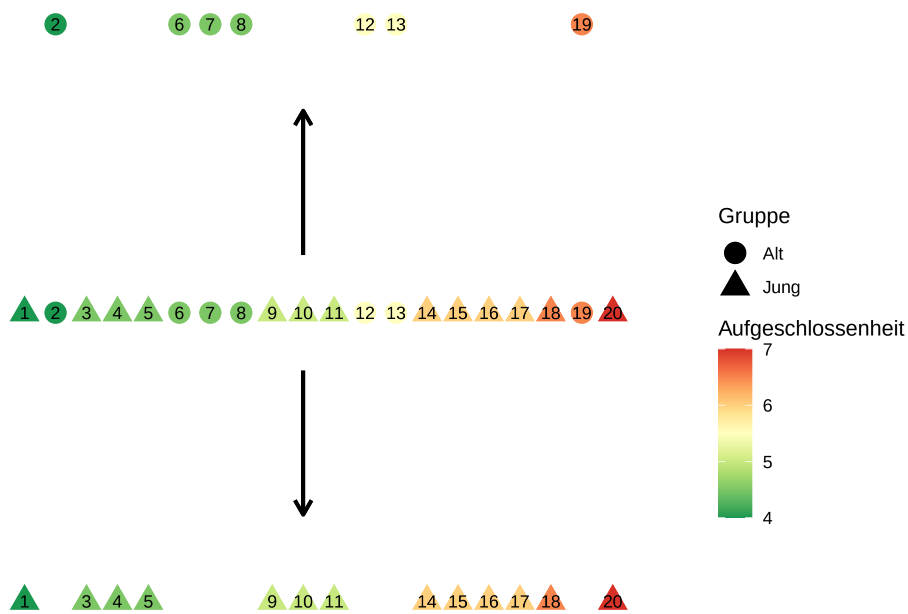


Abbildung 7.2: Mitte: Aufsteigend sortierte Beobachtungen des Merkmals Schmerztoleranz für Junge und Alte. Die Zahlen stehen für die Rangnummern. Oben (Jung) und unten (Alt) stellt die Gruppenaufteilung der Rangnummern dar.

Die Rangsummen unterschieden sich hier nicht, weil eine Gruppe systematisch höhere Ränge erreicht, sondern, weil die Gruppe der Jungen mehr Beobachtungen enthält und damit in dieser Gruppe auch mehr Rangnummern addiert werden.

Wird dies nun korrigiert, ergeben sich die U-Werte

$$U_{\text{Jung}} = 143 - \frac{13 \cdot (13+1)}{2} = 52 \quad U_{\text{Alt}} = 67 - \frac{7 \cdot (7+1)}{2} = 39.$$

Es kann festgestellt werden, dass die U-Werte nicht weit auseinander und damit weit entfernt von 0-liegen. Dies deutet darauf hin, dass es keinen signifikanten Gruppenunterschied in der Aufgeschlossenheit gibt.

Tatsächlich gibt `Jamovi` für dieses Beispiel $p = 0.35$ zurück. Letzteres kann unter `Analysen > t-Tests > t-Test für unabhängige Stichproben` herausgefunden werden, wenn unter `Tests` der Test `Mann-Whitney U` ausgewählt wird.

7.2 Effektstärke

Als Effektstärke können die bisher gesehene Masse wie Cohens d nicht mehr dienen, da diese auf Parametern basieren, welche auf intervallskalierten Merkmalen beruhen. Stattdessen wird als Effektstärke die biserielle Rangkorrelation verwendet (mehr zum Thema Korrelation folgt in den nächsten Kapiteln).

Zur Illustration der Berechnung der biseriellen Rangkorrelation wird folgendes Beispiel verwendet. Es soll getestet werden, ob Hasen oder Schildkröten schneller laufen können. Dazu wird auf einer Rennstrecke die Zeit von 10 Hasen und 10 Schildkröten gestoppt. Um die biserielle Rangkorrelation zu berechnen, werden zwischen den Beobachtungsgruppen alle möglichen Paare aufgelistet - also Hase 1 mit Schildkröte 1, Hase 1 mit Schildkröte 2, usw., Hase 2 mit Schildkröte 1, Hase 2 mit Schildkröte 2, usw., bis Hase 10 mit Schildkröte 10. Dabei ergeben sich in diesem Beispiel genau 100 Paarungen. Nehmen wir nun an, dass für 90 dieser Paare, der Hase schneller lief als die Schildkröte und für 10 Paare die Schildkröte schneller als der Hase. Die biserielle Rangkorrelation entspricht nun der Anzahl Paare f für welche Gruppe 1 höhere Werte hatte minus der Anzahl Paare u für welche Gruppe 2 höhere Werte hatte und ist

$$r = f - u = 0.9 - 0.1 = 0.8.$$

Wären die Schildkröten in der Hälfte der so aufgelisteten Paarungen besser gewesen wäre die biserielle Rangkorrelation $r = f - u = 0.5 - 0.5 = 0$. Wären die Hasen immer besser gewesen, wäre die biserielle Rangkorrelation $r = f - u = 1 - 0 = 1$. Wären die Schildkröten immer besser gewesen, wäre die biserielle Rangkorrelation $r = f - u = 0 - 1 = -1$. Kein Effekt entspricht also $r = 0$, je weiter weg von 0 die biserielle Rangkorrelation, desto stärker ist der Effekt. Es spielt keine Rolle, ob der Effekt positiv oder negativ ist.

Die biserielle Effektstärke kann auch aus der Teststatistik U berechnet werden mit der Formel

$$r = \frac{2 \cdot U}{n_1 \cdot n_2} - 1,$$

wobei hier das Vorzeichen ändert, je nach dem, für welche Gruppen U_1 resp. U_2 stehen.

Die Interpretation einer Korrelation als Effektstärke unterliegt einer anderen Referenz als die bisher gesehenen Effektstärken. **?** schlägt zur Interpretation einer Korrelation die Richtgrössen

- $|r| \approx 0.1$: schwacher Effekt
- $|r| \approx 0.3$: mittlerer Effekt
- $|r| \approx 0.5$: starker Effekt

vor. Um wieder Klarheit im Unterrichtssetting zu schaffen, werden hier die folgenden Abgrenzungen verwendet:

- $0 < |r| \leq 0.2$: schwacher Effekt
- $0.2 < |r| \leq 0.4$: mittlerer Effekt
- $0.4 < |r|$: starker Effekt

In Jamovi kann die biserielle Rankorrelation unter `Zusätzliche Statistiken > Effektstärke` eingeblendet werden. Für die beiden Beispiele **??** und **??** führt dies zu folgenden Effektstärken und Berichtensätzen:

> Ein zweiseitiger U-Test nach Mann und Whitney ergibt, dass sich die mediane Schmerztoleranz von Mäusen mit MC4R Gen (Mdn $= 11.5, N = 13$) und ohne MC4R Gen (Mdn $= 7, N = 15$) signifikant unterscheidet, $U = 13, p < .001, r = -0.87$ Der Effekt ist als gross einzustufen.
>
> Ein zweiseitiger U-Test nach Mann und Whitney ergibt, dass sich die mediane Aufgeschlossenheit von jungen Menschen (Mdn $= 5, N = 13$) und alten Menschen (Mdn $= 5, N = 7$) nicht signifikant unterscheidet, $U = 33.5, p = .35, r = -0.26$. Der Effekt ist als mittel einzustufen.

Achtung

Hinweis. Im Berichtensatz muss Folgendes enthalten sein:

- Test und Art der Hypothesenstellung
- Dass es sich um einen Unterschied der Mediane handelt
- Zusammenfassende deskriptive Statistiken der Stichproben
- Statistiken des Tests inklusive Teststatistik, p-Wert und Effektstärke
- Die Interpretation der Effektstärke kann, wenn gefragt, hinzugefügt werden.

7.3 Übungen

Übung 7.1.

Für die Dorfstrasse in Köniz wurden zwei Verkehrskonzepte verglichen Tempo 50 mit Fussgängerstreifen und Tempo 30 ohne Fussgängerstreifen. Um die beiden Verkehrskonzepte zu evaluieren, wurden verschiedene Zahlen erhoben unter anderem die Durchfahrtszeit von Autos, die Anzahl Strassenquerungen von Fussgängern und die Anzahl Unfälle. Hier soll nur die Durchfahrtszeit von Autos in Minuten betrachtet und herausgefunden werden wie sich die Verkehrskonzepte auf diese ausgewirkt haben. Fiktive Daten dazu wurden als Datensatz `07-exr-tempo30.sav` abgelegt.

a) Prüfen Sie, ob die Testvoraussetzungen für einen Welch Test gegeben sind.
b) Beschreiben Sie die Population und das Merkmal.
c) Testen Sie mit einem U-Test bei Signifikanzniveau 5%, ob sich die medianen Durchfahrtszeiten in der Population unterscheiden. Berichten Sie das Resultat und interpretieren Sie die Effektstärke.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die Teilfragen beantwortet werden.

a) Die Durchfahrtszeit ist ein intervallskaliertes Merkmal. Ob die Beobachtungen einer Zufallsstichprobe entstammen oder nicht muss beim Versuchsaufbau berücksichtigt werden und kann nicht im Nachhinein aus den Daten gelesen werden. Es sind mehr als 30 Beobachtungen pro Gruppe vorhanden. Damit sind alle Voraussetzungen für den Welch Test gegeben und es könnte hier auch ein Welch Test durchgeführt werden.

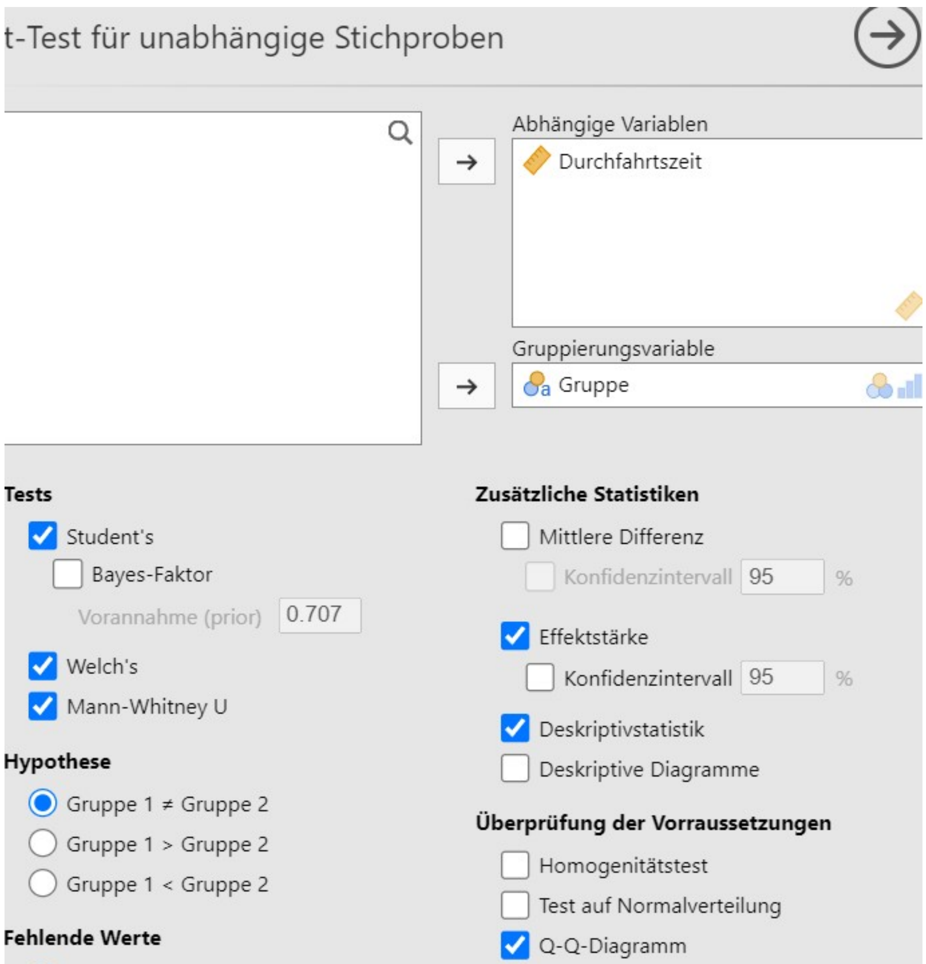


Abbildung 7.3: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Durchfahrtszeit	Student's t	2.251 [a]	117.000	0.02628	Cohens d	0.414
	Welch's t	2.326	113.563	0.02182	Cohens d	0.421
	Mann-Whitney U	1357.000		0.03386	Biserielle Rangkorrelation	−0.227

Anmerkung. H_a $\mu_{\text{Tempo 50}} \neq \mu_{\text{Tempo 30}}$

[a] Der Levene-Test ist signifikant ($p < 0,05$), was auf eine Verletzung der Annahme gleicher Varianzen hindeutet

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Durchfahrtszeit	Tempo 50	65	5.845	4.265	5.368	0.666
	Tempo 30	54	3.899	3.046	3.721	0.506

Diagramme

Durchfahrtszeit

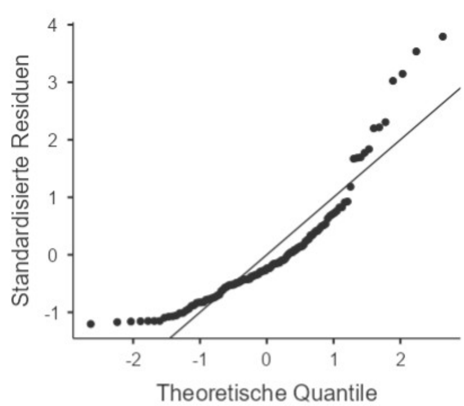


Abbildung 7.4: Jamovi Ausgabe.

b) Die Population sind alle Autos die durch die betroffene Strasse fahren. Das Merkmal ist die Durchfahrtszeit von Autos in Minuten.
c) Es soll getestet werden, ob sich die medianen Durchfahrtszeiten unterscheiden. Die Hypothesen für den U-Test sind also zweiseitig formuliert $H_0 : \mu_{\text{Tempo 30}} = \mu_{\text{Tempo 50}}$ und $H_1 : \mu_{\text{Tempo 30}} \neq \mu_{\text{Tempo 50}}$. Das Testergebnis kann aus Jamovi abgelesen werden und wird wie folgt berichtet und interpretiert:

   Ein zweiseitiger U-Test nach Mann und Whitney ergibt, dass sich die mediane Durchfahrtszeit durch Köniz bei Tempo 30 (Mdn $= 3.05\text{Min}, N = 54$) und Tempo 50 (Mdn $= 4.27\text{Min}, N = 65$) signifikant unterscheidet, $U = 1357, p = .034, r = -0.227$. Der Effekt ist als mittel einzustufen.

Übung 7.2.

Ein Flughafen nimmt ein neues Terminal in betrieb. Um den Betrieb für die Zukunft zu optimieren und die Wartezeit der Gäste möglichst kurz zu halten, werden zwei Warteschlangenkonzepte getestet. Gäste werden dabei per Kamera getrackt und ihre Wartezeit in Minuten wird im Datensatz `07-exr-warteschlangen.sav` festgehalten.

a) Prüfen Sie, ob die Testvoraussetzung für einen Welch Test gegeben ist.
b) Beschreiben Sie die Population und das Merkmal.
c) Testen Sie mit einem U-Test bei Signifikanzniveau 5%, ob sich die medianen Wartezeiten der beiden Gruppen in der Population unterscheiden. Berichten Sie das Resultat und interpretieren Sie die Effektstärke.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die Teilfragen beantwortet werden.

a) Die Wartezeit ist ein intervallskaliertes Merkmal. Ob die Beobachtungen einer Zufallsstichprobe entstammen oder nicht muss beim Versuchsaufbau berücksichtigt werden und kann nicht im Nachhinein aus den Daten gelesen werden. Es sind mehr als 30 Beobachtungen pro Gruppe vorhanden. Damit sind alle Voraussetzungen für den Welch Test gegeben und es könnte hier auch ein Welch Test durchgeführt werden.
b) Die Population sind alle Gäste, die den neuen Terminal benutzen. Das Merkmal ist die Wartezeit der Gäste in Minuten.
c) Es soll getestet werden, ob sich die medianen Durchfahrtszeiten unterscheiden. Die Hypothesen für den U-Test sind also zweiseitig formuliert $H_0 : \mu_{\text{Konzept 1}} = \mu_{\text{Konzept 2}}$ und $H_1 : \mu_{\text{Konzept 1}} \neq \mu_{\text{Konzept 2}}$. Das Testergebnis kann aus `Jamovi` abgelesen werden und wird wie folgt berichtet und interpretiert:

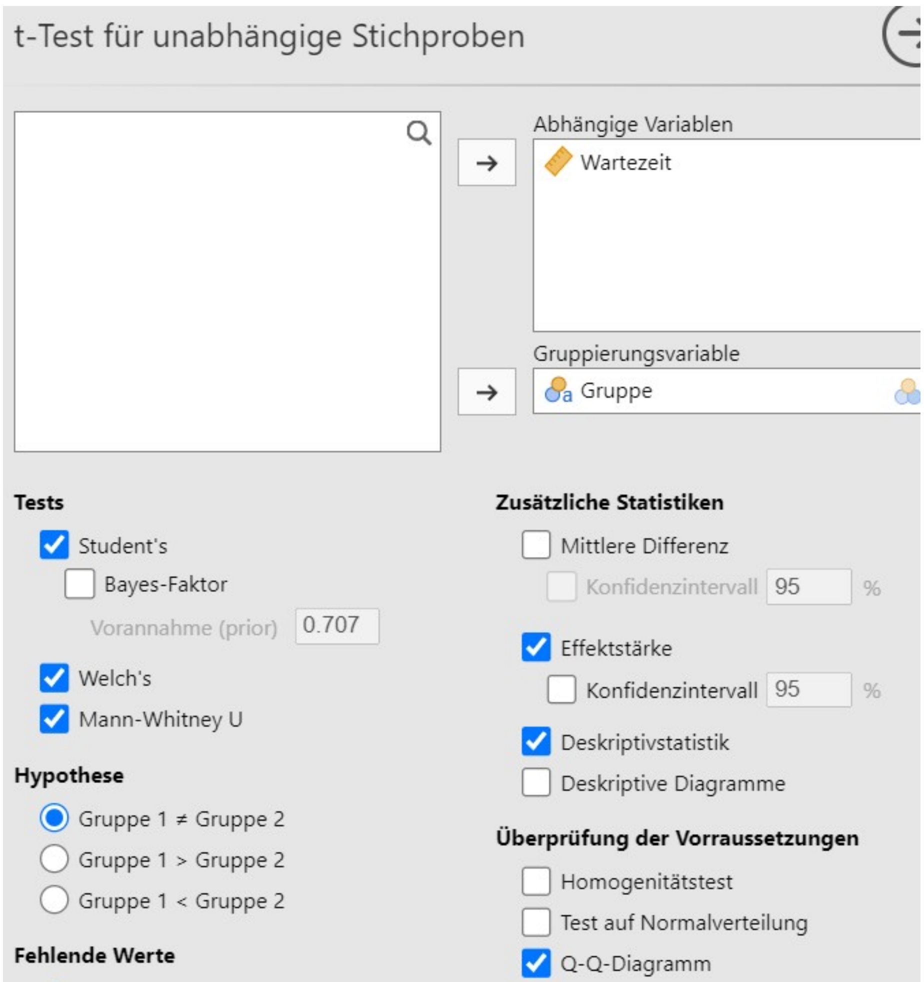


Abbildung 7.5: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Wartezeit	Student's t	−1.844 [a]	156.000	0.06710	Cohens d	−0.314
	Welch's t	−2.120	139.645	0.03581	Cohens d	−0.335
	Mann-Whitney U	2345.000		0.15435	Biserielle Rangkorrelation	0.141

Anmerkung. H_a $\mu_{\text{Konzept 1}} \neq \mu_{\text{Konzept 2}}$

[a] Der Levene-Test ist signifikant ($p < 0{,}05$), was auf eine Verletzung der Annahme gleicher Varianzen hindeutet

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Wartezeit	Konzept 1	51	4.247	3.405	3.727	0.522
	Konzept 2	107	5.843	4.183	5.617	0.543

Diagramme

Wartezeit

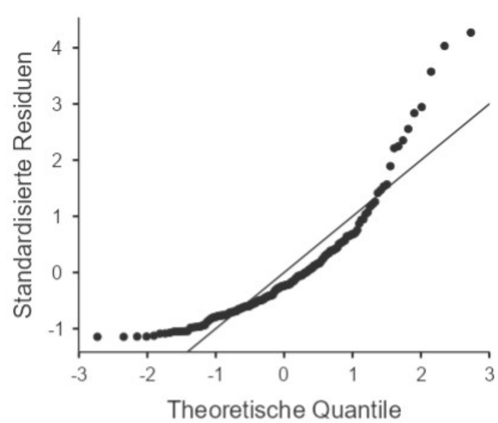


Abbildung 7.6: Jamovi Ausgabe.

> Ein zweiseitiger U-Test nach Mann und Whitney ergibt, dass sich die mediane Wartezeit am neuen Terminal bei Konzept 1 (Mdn $=$ 3.41Min, $N = 51$) und Konzept 2 (Mdn $=$ 4.18Min, $N = 107$) nicht signifikant unterscheidet, $U = 2345, p = .154, r = 0.141$. Der Effekt ist als klein einzustufen.

Übung 7.3.

Studierende der Kurse Statistik 1 und 2 sollen auf einer Skala von 1 bis 10 bewerten (1 = gar nicht herausfordernd, 10 = äusserst herausfordernd), wie herausfordernd der Statistikunterricht für sie ist. Wird von allen Studierenden der Kurs Statistik 2 durchschnittlich als herausfordernder betrachtet als Statistik 1?

a) Stellen Sie die Testhypothesen auf.
b) Weshalb wird hier ein U-Test gegenüber einem Welch Test bevorzugt?
c) Führen Sie einen U-Test durch mit Signifikanzniveau 5%, berichten Sie das Ergebnis in einem Satz und interpretieren Sie die Effektstärke. Die Daten sind unter `07-exr-statistik-herausforderung.sav` verfügbar. (Beispiel frei erfunden.)

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit können die Teilfragen beantwortet werden.

a) Die Frage "Wird von allen Studierenden der Kurs Statistik 2 durchschnittlich als herausfordernder betrachtet als Statistik 1" ist einseitig formuliert. Es soll getestet werden, ob $H_1 : \mu_{\text{Statistik 1}} < \mu_{\text{Statistik 2}}$. Dies entspricht der Nullhypothese $H_0 : \mu_{\text{Statistik 1}} \geq \mu_{\text{Statistik 2}}$
b) Hier wird ein U-Test verwendet, da das Merkmal auf einem einzigen Likertskaliertem Merkmal beruht. Dieses ist demnach ordinalskaliert. Ein Welch Test eignet sich nur für intervallskalierte Merkmale.
c) Das Testergebnis kann aus `Jamovi` abgelesen werden und wird wie folgt berichtet und interpretiert:

> Ein einseitiger U-Test nach Mann und Whitney ergibt, dass die mediane Herausforderung in Statistik 2 (Mdn $= 7, N = 5$) signifikant grösser ist als in Statistik 1 (Mdn $= 3.5, N = 6$), $U = 2, p = .011, r = 0.867$. Der Effekt ist als gross einzustufen.

Übung 7.4. Eine Neurologin sammelt Daten, um die depressive Wirkung bestimmter Freizeitdrogen zu untersuchen. Sie schickt dazu 20 männliche Clubgänger unter kontrollierten Bedingungen während vier Stunden in ein Tanzlokal. Zehn Testpersonen nehmen eine Ecstasy Pille ein, die zehn anderen

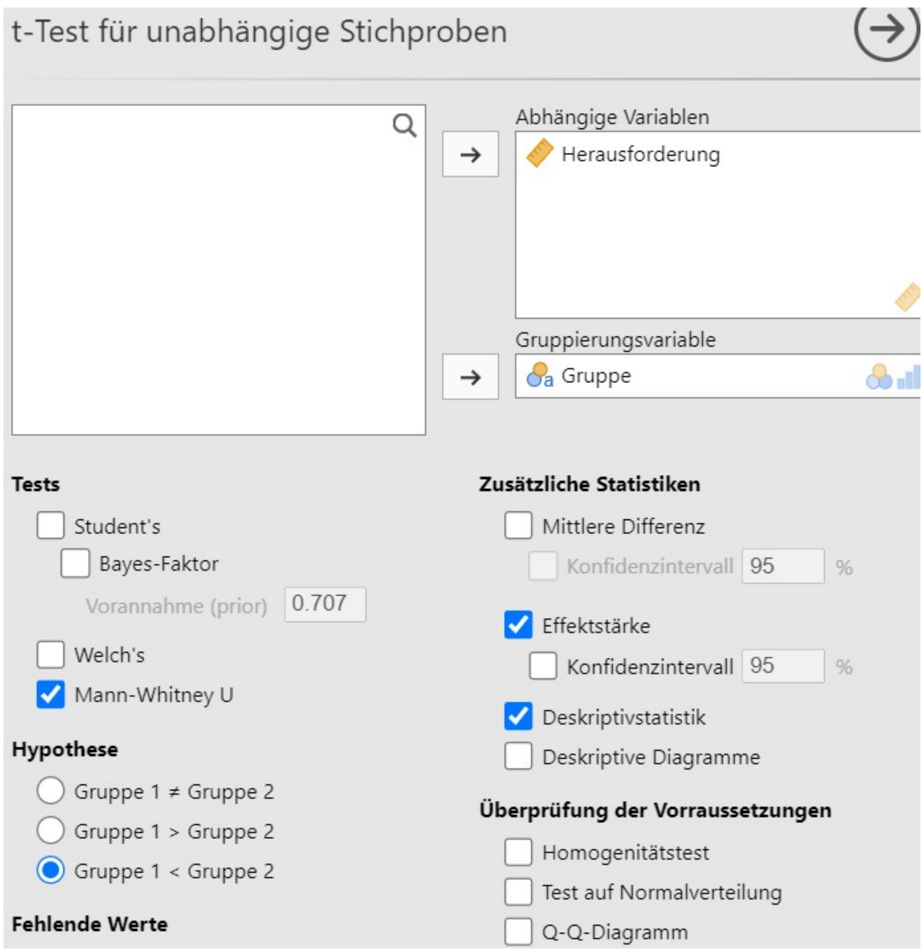


Abbildung 7.7: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	p		Effektstärke
Herausforderung	Mann-Whitney U	2.000	0.01078	Biserielle Rangkorrelation	0.867

Anmerkung. H_a $\mu_{\text{Statistik1}} < \mu_{\text{Statistik2}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
Herausforderung	Statistik1	6	3.667	3.500	1.633	0.667
	Statistik2	5	7.400	7.000	2.074	0.927

Abbildung 7.8: Jamovi Ausgabe.

trinken einen Liter Bier. Der Grad der Depression wird mit dem Beck Depression Inventory (BDI) zwölf Stunden nach dem Verlassen des Tanzlokals gemessen. Die Daten sind unter `07-exr-depression-ecstasy.sav` verfügbar. Ist die durchschnittliche Schwere der Nachtanzdepression bei der Ecstasy-Gruppe schlimmer als bei der Alkohol-Gruppe? Testen Sie mit einem U-Test, berichten Sie das Testresultat und schätzen sie die Effektstärke ein.

Lösung. Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung **??**.

Dies produziert das Analyseergebnis in Abbildung **??**.

Damit kann die Frage nun beantwortet werden:

> Ein einseitiger U-Test nach Mann und Whitney ergibt, dass die mediane Nachtanzdepression in der Ecstasy-Gruppe (Mdn $= 17.5, N = 10$) nicht signifikant grösser ist als in der Alkohol-Gruppe (Mdn $= 16, N = 10$), $U = 35.5, p = .143, r = -0.290$. Der Effekt ist als mittel einzustufen.

7.4 Test

TODO.

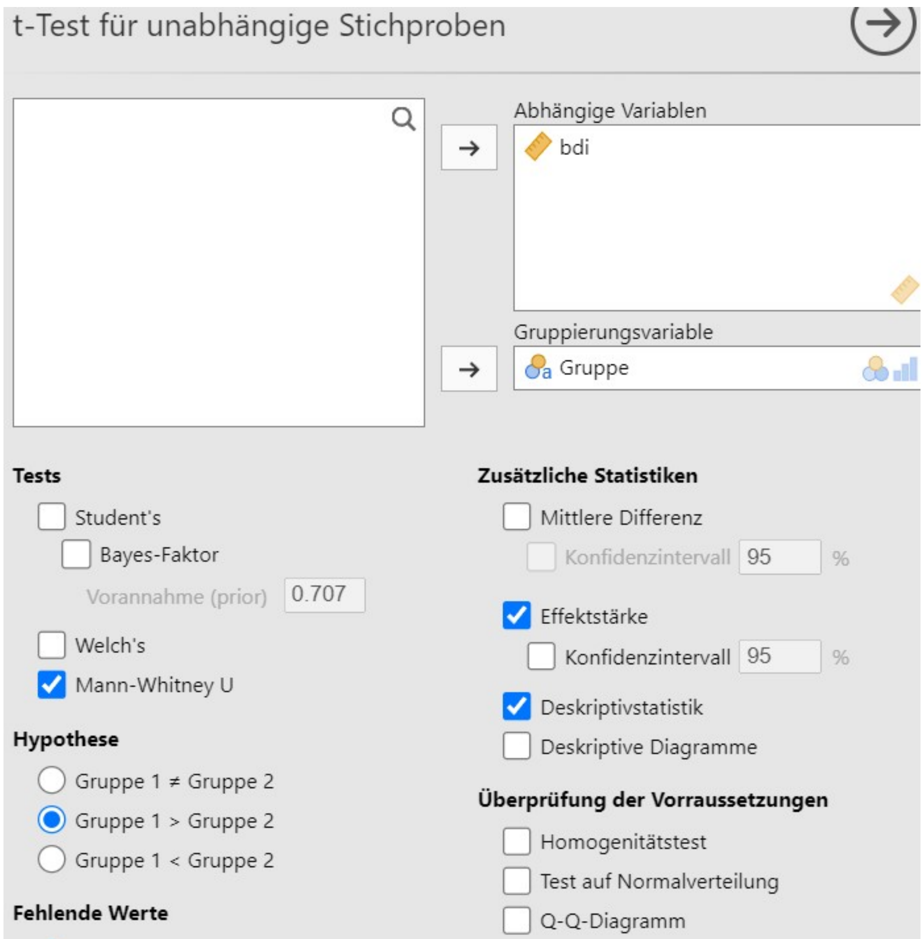


Abbildung 7.9: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	p		Effektstärke
bdi	Mann-Whitney U	35.500	0.14304	Biserielle Rangkorrelation	−0.290

Anmerkung. H_a $\mu_{\text{Ecstasy}} > \mu_{\text{Alkohol}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
bdi	Ecstasy	10	19.600	17.500	6.603	2.088
	Alkohol	10	16.400	16.000	2.271	0.718

Abbildung 7.10: Jamovi Ausgabe.

Teil III

Zusammenhang zweier Merkmale

Kapitel 8

Zusammenhang zweier Merkmale

Bislang wurde immer ein Merkmal separat betrachtet und manchmal wurden Untergruppen verglichen. Dabei ging es um die Frage, wo der Erwartungswert des Merkmals liegt und ob er einem gewissen Wert entspricht respektive für zwei Gruppen identisch ist. Wird jetzt noch ein zweites Merkmal beobachtet stellt sich die Frage, wie sich die Merkmale zueinander verhalten.

Beispiel 8.1 (Zahlungsbereitschaft)**.** Eine Firma will eine Kickstarteridee für ein Kinderspielzeug auf den Markt bringen. Dazu muss sie herausfinden wie viel die Konsumierenden bereit sind für das Spielzeug zu bezahlen. Es werden 375 Konsumierende gefragt, wie viel sie für das Spielzeug zahlen würden. Zusätzlich wurde auch nach dem Jahreseinkommen in CHF und der Anzahl Spielzeuge im Haushalt gefragt.

Um den Zusammenhang zwischen zwei intervallskalierten Merkmalen aufzuzeigen wird ein sogenanntes **Streudiagramm** verwendet. Um den Zusammenhang zwischen zwei intervallskalierten Merkmalen aufzuzeigen wird ein sogenanntes Streudiagramm verwendet. Dabei wird ein Koordinatensystem erstellt mit einem Merkmal auf der x-Achse und dem anderen Merkmal auf der y-Achse. Danach wird für jede Beobachtung ein Punkt bei den entsprechenden Werten für die beiden betrachteten Merkmale in dieses Koordinatensystem eingezeichnet.

In Abbildung **??** sind drei Streudiagramme aufgezeichnet. Sie setzen jeweils die Zahlungsbereitschaft (Preis) mit dem Alter in Jahren, der Anzahl vorhandener Spiele im Haushalt und dem Einkommen respektive in Bezug. Es kann beobachtet werden, dass die Zahlungsbereitschaft unabhängig vom Alter immer ungefähr ähnlich hoch ist. Es ist des Weiteren klar zu sehen, dass die Zahlungsbereitschaft mit der Anzahl im Haushalt vorhandener Spiele sinkt. Ein bisschen

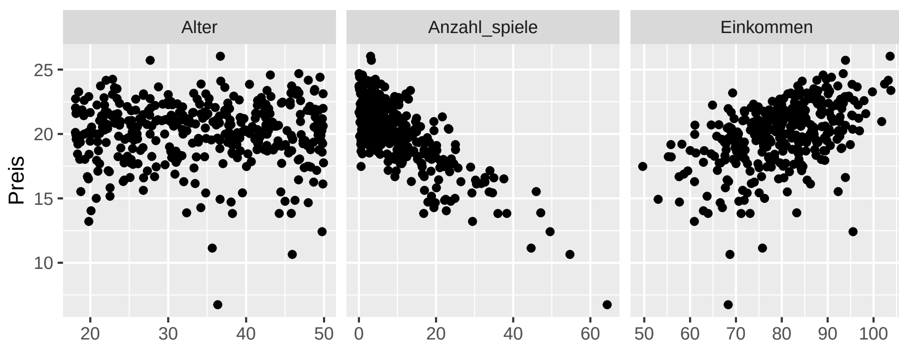


Abbildung 8.1: Die drei Streudiagramme zeigen jeweils die Zahlungsbereitschaft (Preis in CHF) auf der y-Achse und das Alter in Jahren, die Anzahl Spiele im Haushalt und das Jahreseinkommen in 1'000 CHF respektive auf der x-Achse.