

# Statistik 1

Daniel J. F. Gerber

26 May, 2025



# Inhaltsverzeichnis

<b>Vorwort</b>	<b>7</b>
<b>1 Einleitung</b>	<b>9</b>
1.1 Worum geht es? . . . . .	9
1.2 Inhaltlicher Aufbau . . . . .	9
1.3 Wie soll ich dieses Buch lesen? . . . . .	10
1.4 Formeln, Symbole und Zahlen . . . . .	11
1.5 Software . . . . .	11
<b>I Eine intervallskaliertes Merkmal</b>	<b>13</b>
<b>2 Intervallskalierte Merkmale</b>	<b>15</b>
2.1 Was ist ein intervallskaliertes Merkmal? . . . . .	15
2.2 Wie kann ein intervallskaliertes Merkmal beschrieben werden? . .	16
2.3 Übungen . . . . .	19
2.4 Test . . . . .	38
<b>3 Stichprobenziehung</b>	<b>41</b>
3.1 Was ist das Problem der Stichprobenziehung? . . . . .	42
3.2 Wie kann man Aussagen über die Grundgesamtheit machen? . .	44
3.3 Übungen . . . . .	45
3.4 Test . . . . .	46

<b>4 Durchschnitt und Standardabweichung schätzen</b>	<b>49</b>
4.1 Wo liegt der Durchschnitt der Grundgesamtheit? . . . . .	49
4.2 Übungen . . . . .	58
4.3 Test . . . . .	67
<b>5 Zentrale Tendenz testen</b>	<b>71</b>
5.1 Entspricht der Erwartungswert einem gewissen Wert? . . . . .	72
5.2 Weicht der gefundene Durchschnitt stark vom hypothetischen Wert ab? . . . . .	84
5.3 Testvoraussetzungen . . . . .	85
5.4 Übungen . . . . .	86
5.5 Test . . . . .	95
<b>II Zwei Gruppen vergleichen</b>	<b>97</b>
<b>6 Gruppenmittelwertunterschied bei einem intervallskalierten und normalverteilten Merkmal</b>	<b>99</b>
6.1 Was ist das Problem der Stichprobenziehung? . . . . .	100
6.2 Effektstärken . . . . .	106
6.3 Testvoraussetzungen . . . . .	109
6.4 Übungen . . . . .	109
<b>7 Gruppenmittelwertunterschied bei einem mindestens ordinal-skalierten Merkmal</b>	<b>121</b>
7.1 Wie stark unterscheiden sich die Mediane? . . . . .	121
7.2 Effektstärke . . . . .	125
7.3 Übungen . . . . .	127
7.4 Test . . . . .	135
<b>III Zusammenhang zweier Merkmale</b>	<b>139</b>
<b>8 Korrelation</b>	<b>141</b>
8.1 Wie stark ist der Zusammenhang zwischen zwei intervall-skalierten und normalverteilten Variablen? . . . . .	145

<b>INHALTSVERZEICHNIS</b>	<b>5</b>
8.2 Wie stark ist der Zusammenhang zwischen zwei mindestens ordinalskalierten Merkmalen? . . . . .	146
8.3 Wie stark ist der Zusammenhang zwischen einem intervall-skalierten und normalverteilten Merkmal und einem dichotomen Merkmal? . . . . .	148
8.4 Wie stark ist der Zusammenhang zwischen einem mindestens ordinalskalierten Merkmal und einem dichotomen Merkmal? . . . . .	149
8.5 Was ist eine Störfaktor und wie wird damit umgegangen? . . . . .	150
8.6 Übungen . . . . .	151
8.7 Test . . . . .	165
<b>9 Zusammenhang dichotomer Merkmale</b>	<b>167</b>
9.1 Zusammenhang dichotomer Merkmale beschreiben . . . . .	167
9.2 Zusammenhang dichotomer Merkmale testen . . . . .	172
9.3 Übungen . . . . .	178
9.4 Test . . . . .	187
<b>10 Zusammenhang nominalskalierter Merkmale</b>	<b>189</b>
10.1 Zusammenhang nominalskalierter Merkmale beschreiben . . . . .	189
10.2 Zusammenhang nominalskalierter Merkmale testen . . . . .	192
10.3 Übungen . . . . .	194
10.4 Test . . . . .	194
<b>Begriffsverzeichnis</b>	<b>195</b>
<b>Literaturverzeichnis</b>	<b>197</b>



# **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
- Punktbiserial 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 (Deutsche Gesellschaft für Psychologie, 2019). Dieser ist wiederum stark an den Standard der American psychological association 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 (Selker et al., 2024)
- R (R Core Team, 2024)
- Tidyverse (Wickham et al., 2019)
- Bookdown (Xie, 2016)



# **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  $10^{\circ}\text{C}$  und  $42^{\circ}\text{C}$ . Es ist möglich zu sagen, dass eine Person mit  $40^{\circ}\text{C}$  Körpertemperatur eine höhere Temperatur hat als eine mit  $38^{\circ}\text{C}$  Körpertemperatur. Ausserdem sind alle erdenklichen Zwischenwerte möglich, so auch dass bei einer Person eine Körpertemperatur von  $37.821239^{\circ}\text{C}$  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  $0.2^{\circ}\text{C}$  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.

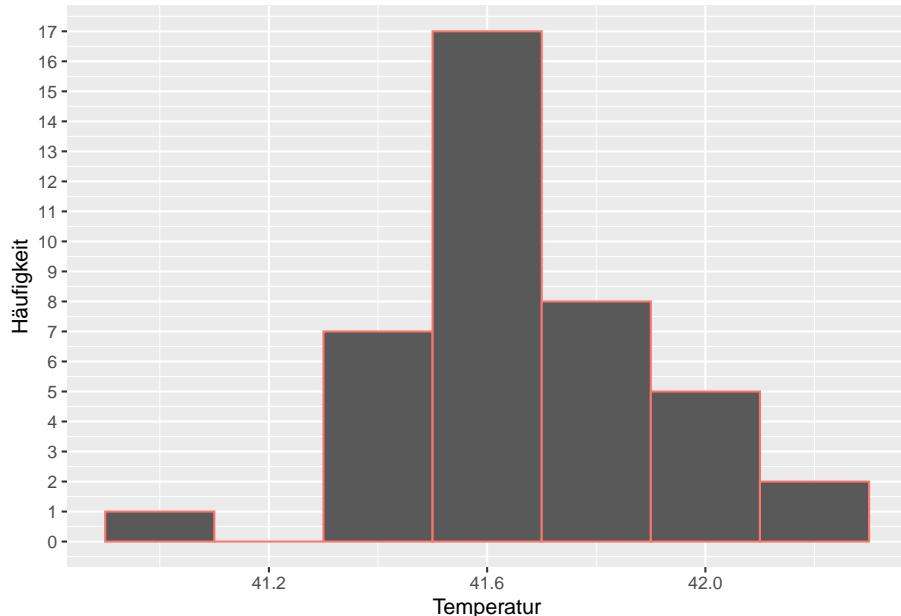


Abbildung 2.1: Histogramm Körpertemperatur Enten.

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  $41.6^{\circ}\text{C}$  am häufigsten sind und tiefere und höhere Temperaturen seltener vorkommen. Auf einen Blick sieht sie auch, dass die Temperatur aller Enten zwischen  $41^{\circ}\text{C}$  und  $42.2^{\circ}\text{C}$  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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.1)$$

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  $i = 1, 2, 3, 4, \dots, 39, 40$ . Der Teil  $\sum_{i=1}^n x_i$  bedeutet also nichts anderes als  $x_1 + x_2 + \dots + x_{39} + x_{40}$ , 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 2.3 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 beobachtete Wert  $42.18^\circ C$  und der kleinste beobachtete Wert  $41.07^\circ C$ . Also ist die Spannweite  $42.18 - 41.07 = 1.11^\circ C$ . 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  $41.32^\circ C$  und  $41.35^\circ C$ . In diesem Fall wird ein Mittelwert der beiden nächsten Werte verwendet, hier  $(41.32 + 41.35)/2 = 41.34^\circ C$ . 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  $41.34^\circ C$  und das 95%-Perzentil bei  $42.1^\circ C$ . 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  $41.52^{\circ}\text{C}$  und  $41.82^{\circ}\text{C}$  respektive. Die Differenz dieser Perzentile wird als **Interquartilabstand** (abkürzung IQR von interquartile range) bezeichnet und ist im Beispiel  $0.3^{\circ}\text{C}$ . 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

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

Im Beispiel ist die Standardabweichung der Körpertemperaturen  $0.233^{\circ}\text{C}$ . 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 2.3:

- Erstellen Sie selbst ein Histogramm mit Jamovi und begründen Sie, warum es nicht gleich aussieht wie das Histogramm in Abbildung 2.1.

- (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 2.1 genau mithilfe der Histogramm-Funktion des **Jamovi**-Moduls **JJStatsPlot** (Balci, 2025).
- (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.*

- (a) Das Histogramm, siehe Abbildung 2.2 sieht nicht gleich aus, da Jamovi die Temperaturabschnitte mit Korbbreite  $0.125^{\circ}\text{C}$  kürzer gewählt hat als die in Abbildung Abbildung 2.1 dargestellte Korbbreite von  $0.2^{\circ}\text{C}$ . Ein Histogramm sieht immer anders aus je nach ausgewählter Abschnittsweite.
- (b) Eine Anleitung zur Berechnung in **Jamovi** sowie die berechneten Werte können in Abbildung 2.3 abgelesen und sind Modus =  $41.5^{\circ}\text{C}$ , 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 2.1 kann beobachtet werden, dass die Balken und also auch die Körbe 0.2 Einheiten breit sind. Dies wird so eingestellt, siehe Abbildung 2.4.

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 2.3 abgelesen und sind Interquartilabstand  $IQR = 0.3^{\circ}\text{C}$ , 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.

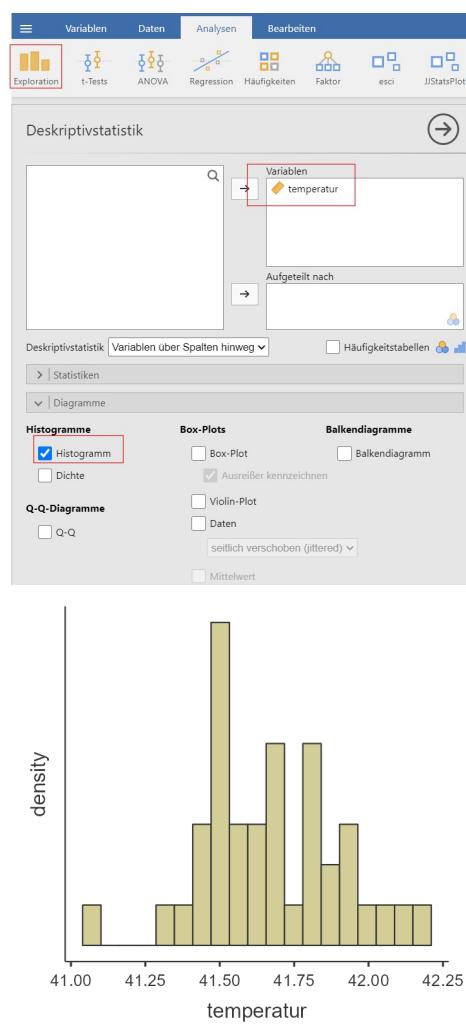


Abbildung 2.2: Links: Jamovi-Anleitung zur Erstellung des Histogramms; rechts: Histogramm der Temperatur.

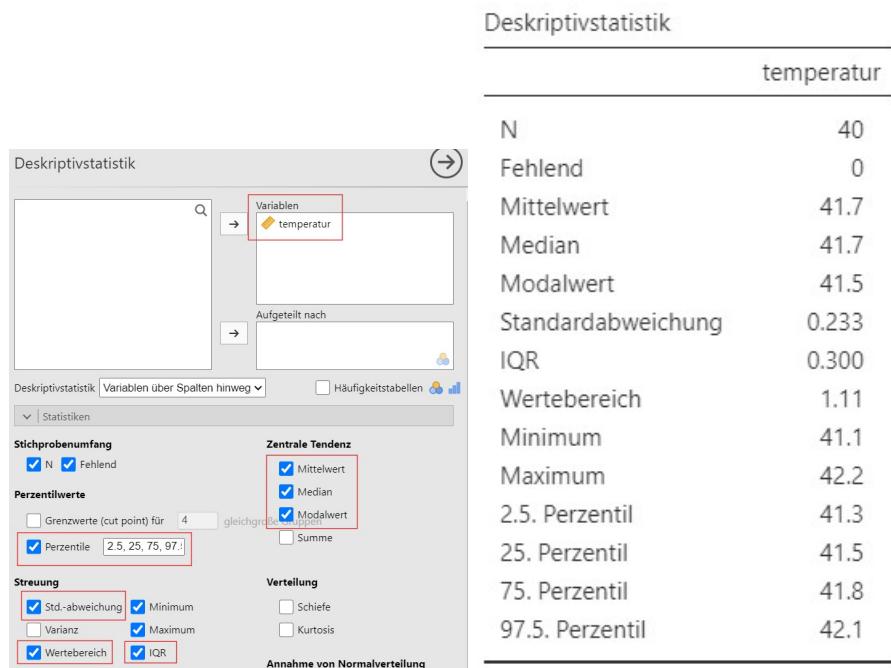


Abbildung 2.3: Links: Jamovi-Anleitung zur Berechnung der gewünschten Parameter; rechts: Parameterwerte.

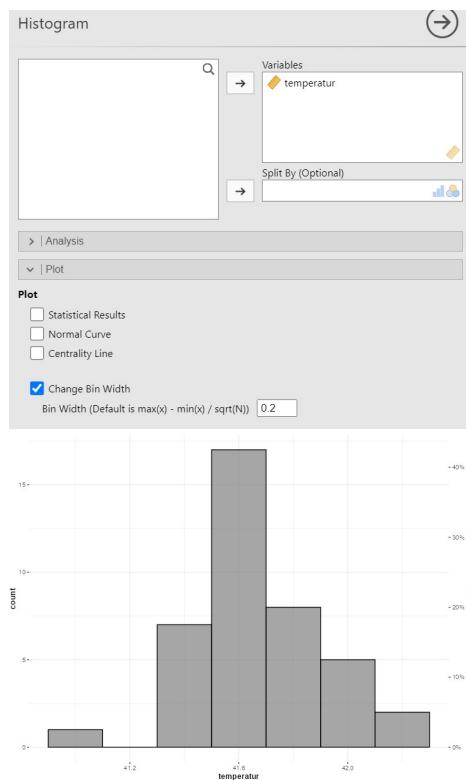


Abbildung 2.4: Links: Jamovi-Anleitung zur Berechnung des gewünschten Histogramms; rechts: Jamovi-Ausgabe.

- Wie gross ist das arithmetische Mittel und der Median dieser Werte? Rechnen Sie im Kopf oder mit einem Taschenrechner.
- 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?
- 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.
- Gleich wie (b), aber führen Sie die Berechnungen aus indem die Zahlen manuell bei **Jamovi** eingegeben.
- Was sagt dies über den Median und das arithmetische Mittel aus?

*Lösung.*

- 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 (2.1) eingesetzt, so gibt dies das arithmetische Mittel

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

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.

- Die Beobachtungen sind jetzt  $x_1 = 18, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$ . Analog wie in (a) kann demnach das arithmetische Mittel als  $\bar{x} = 20$  bestimmt werden. Die aufsteigend sortierten Beobachtungen sind nun 18, 19, 20, 21, 22. Der Median ist also 20.

Für c und d wird der Datensatz bei **Jamovi** eingegeben, siehe Abbildung 2.5, und die Analyseparameter werden gesetzt, siehe Abbildung 2.6.

	A	B
1	18	81
2	21	21
3	20	20
4	19	19
5	22	22

	original	mit_fehler
1	18	81
2	21	21
3	20	20
4	19	19
5	22	22

Abbildung 2.5: Jamovi Dateneingabe.

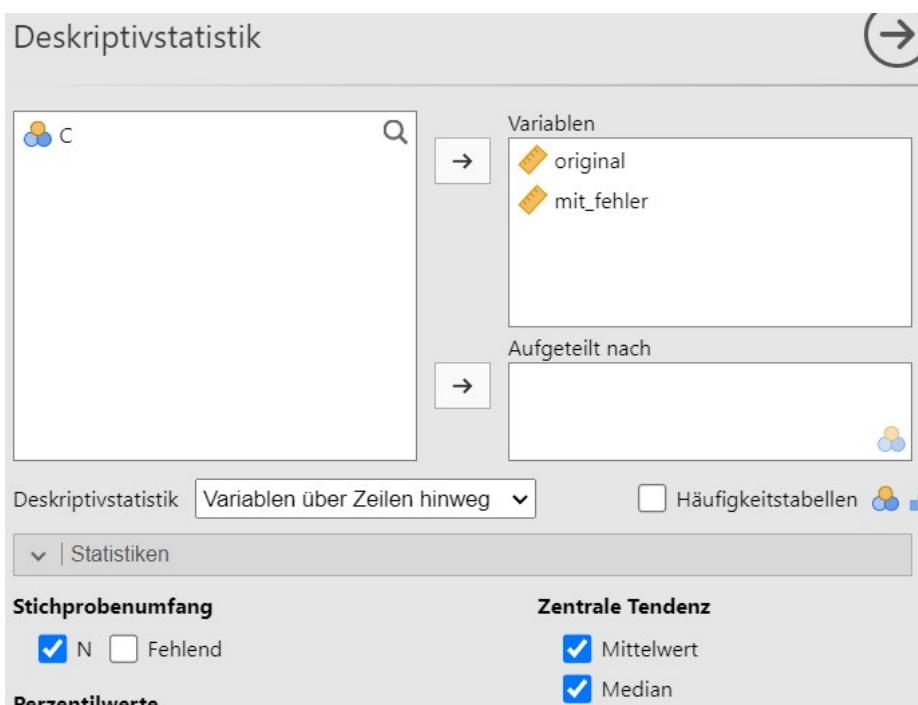


Abbildung 2.6: Jamovi setzen der Analyseparameter.

Deskriptivstatistik			
	N	Mittelwert	Median
original	5	20.0000	20
mit_fehler	5	32.6000	21

Abbildung 2.7: Jamovi Ausgabe.

Dies produziert das Analyseergebnis in Abbildung 2.7.

Damit können die beiden nächsten Teilfragen beantwortet werden.

- (c) Das Resultat in **Jamovi** ist genau gleich wie das händisch berechnete.
- (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 (2.2) eingesetzt, so gibt dies die Standardabwe-

ichung

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.3)$$

$$= \sqrt{\frac{1}{n-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2)} \quad (2.4)$$

$$= \sqrt{\frac{1}{5-1} ((18 - 20)^2 + (21 - 20)^2 + (20 - 20)^2 + (19 - 20)^2 + (22 - 20)^2)} \quad (2.5)$$

$$= \sqrt{\frac{1}{4} (4 + 1 + 0 + 1 + 4)} \quad (2.6)$$

$$= 1.58. \quad (2.7)$$

- (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 (2.2) eingesetzt, so gibt dies die Standardabweichung

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.8)$$

$$= \sqrt{\frac{1}{n-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2)} \quad (2.9)$$

$$= \sqrt{\frac{1}{5-1} ((81 - 32.6)^2 + (21 - 32.6)^2 + (20 - 32.6)^2 + (19 - 32.6)^2 + (22 - 32.6)^2)} \quad (2.10)$$

$$= \sqrt{\frac{1}{4} (2342.56 + 134.56 + 158.76 + 184.96 + 112.36)} \quad (2.11)$$

$$= 27.08. \quad (2.12)$$

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 2.3 beschrieben. Die Analyseparameter werden gesetzt, siehe Abbildung 2.8.

Dies produziert das Analyseergebnis in Abbildung 2.9.

Damit können die beiden nächsten Teilfragen beantwortet werden.

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

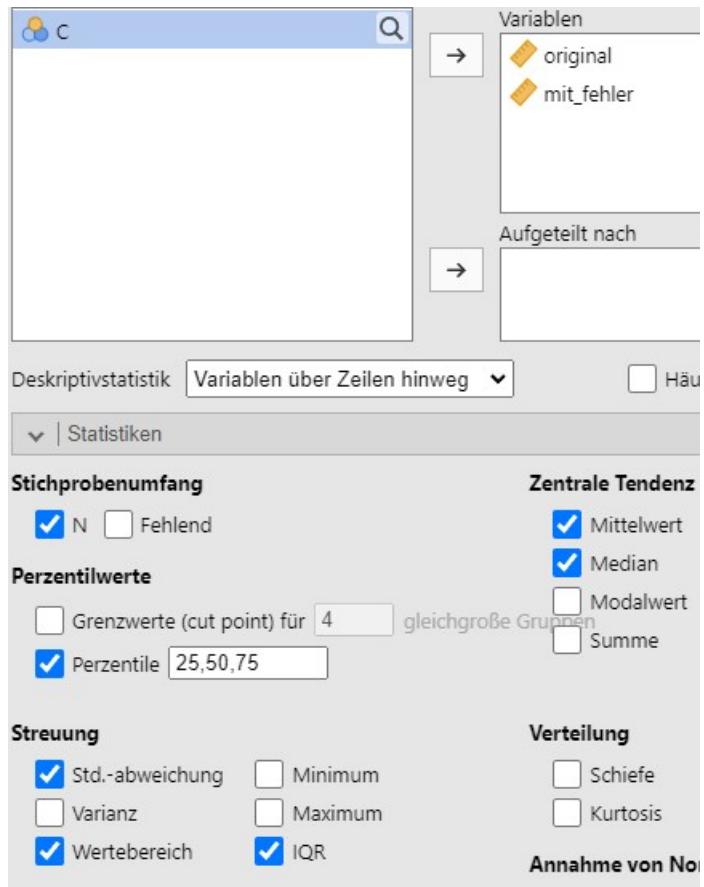


Abbildung 2.8: Jamovi setzen der Analyseparameter.

Deskriptivstatistik									
	N	Mittelwert	Median	Std.-abw.	IQR	Wertebereich	Perzentile		
							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 2.10 gesetzt.

Dies produziert das Analyseergebnis in Abbildung 2.11.

Die Körpergrösse ist demnach  $N = 326$  mal beobachtet worden. Die Zentralitätsmasse 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 2.12

Dies produziert das Analyseergebnis in Abbildung 2.13.

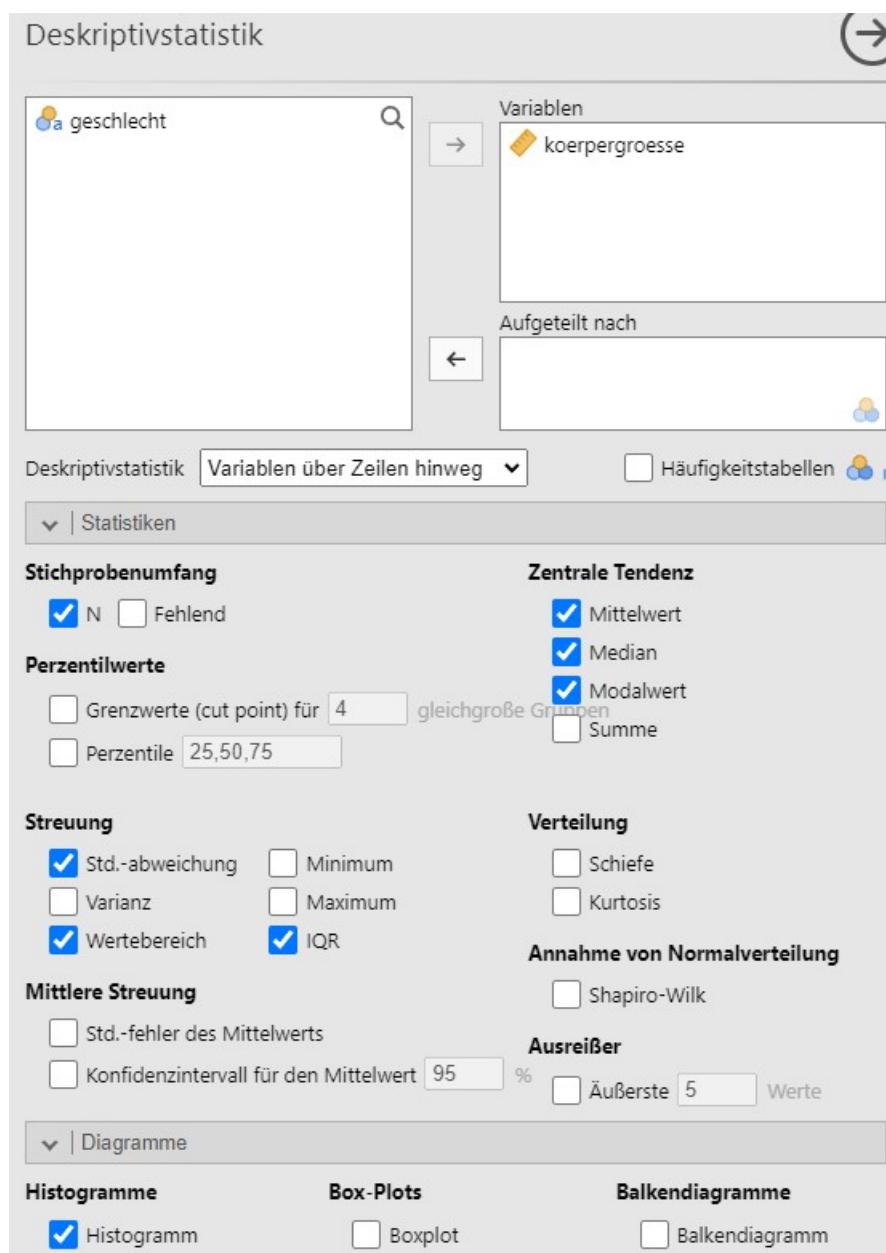


Abbildung 2.10: Jamovi setzen der Analyseparameter.

## Deskriptivstatistik

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

\* Es gibt mehr als einen Modalwert, nur der erste wird berichtet

## Diagramme

## koerpergroesse

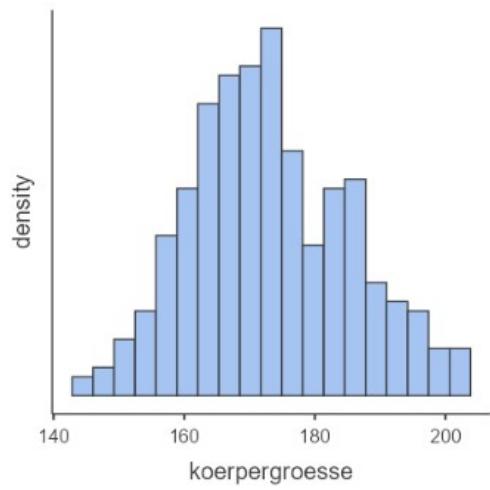


Abbildung 2.11: Jamovi Ausgabe.

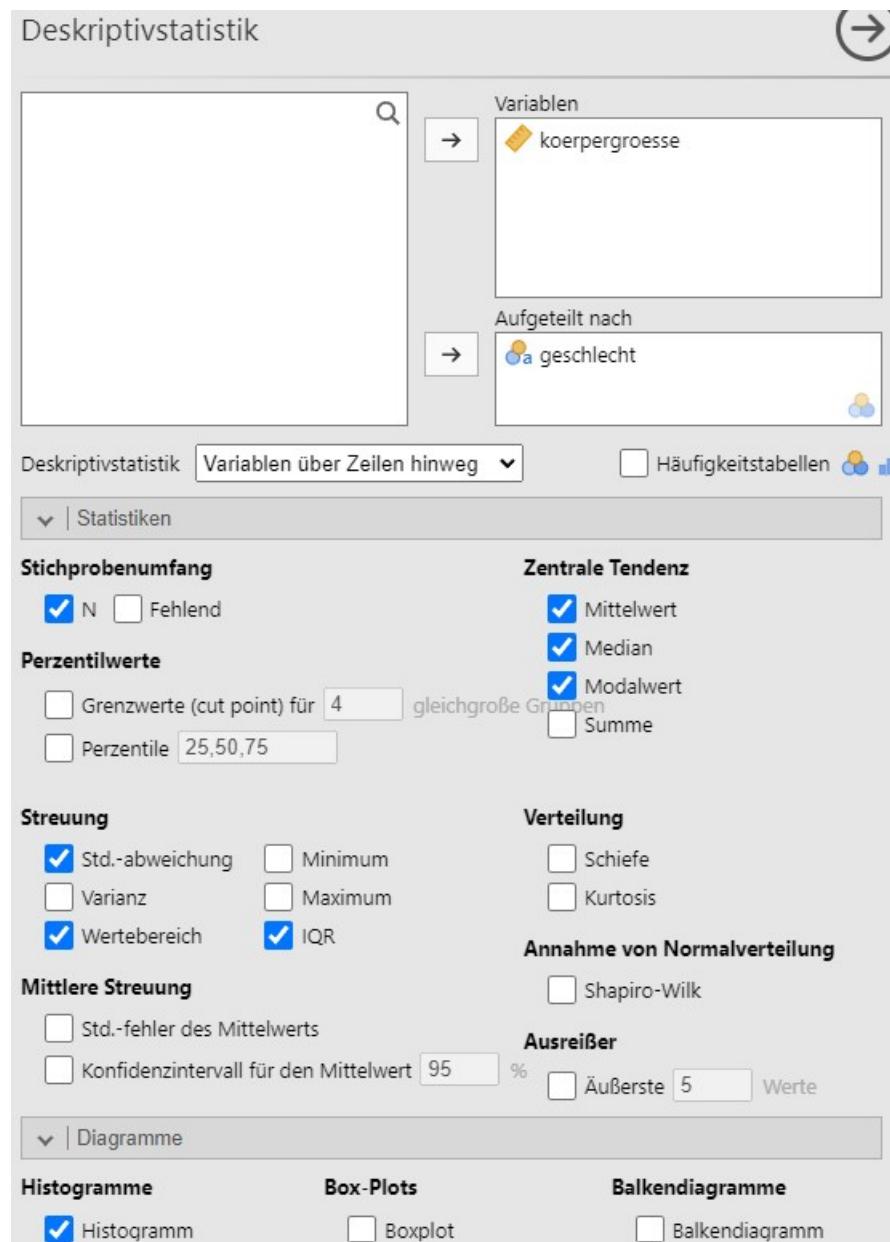


Abbildung 2.12: Jamovi setzen der Analyseparameter.

## Deskriptivstatistik

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

\* Es gibt mehr als einen Modalwert, nur der erste wird berichtet

## Diagramme

koerpergroesse

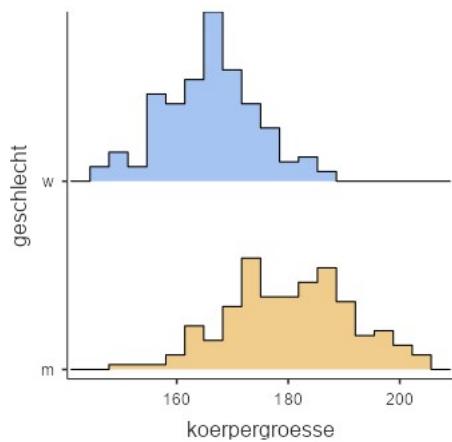


Abbildung 2.13: Jamovi Ausgabe.

Die Körpergrösse ist demnach bei den Frauen  $N = 163$  mal beobachtet worden. Die Zentralitätsmasse 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 Zentralitätsmasse 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 2.14 analysiert.

- a. Es gibt gemäss 2.15 genau 500 Studienteilnehmende (siehe  $N$ ).
- b. Die Histogramme für den IQ und die Aufgeschlossenheit weisen eine ähnliche 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 durchschnittlichen 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 abgespeichert. 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 asymmetrische 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 Minuten hergeleitet werden indem die Werte durch 60 geteilt werden. Es reicht deshalb die Kennzahlen für die Wartezeit in

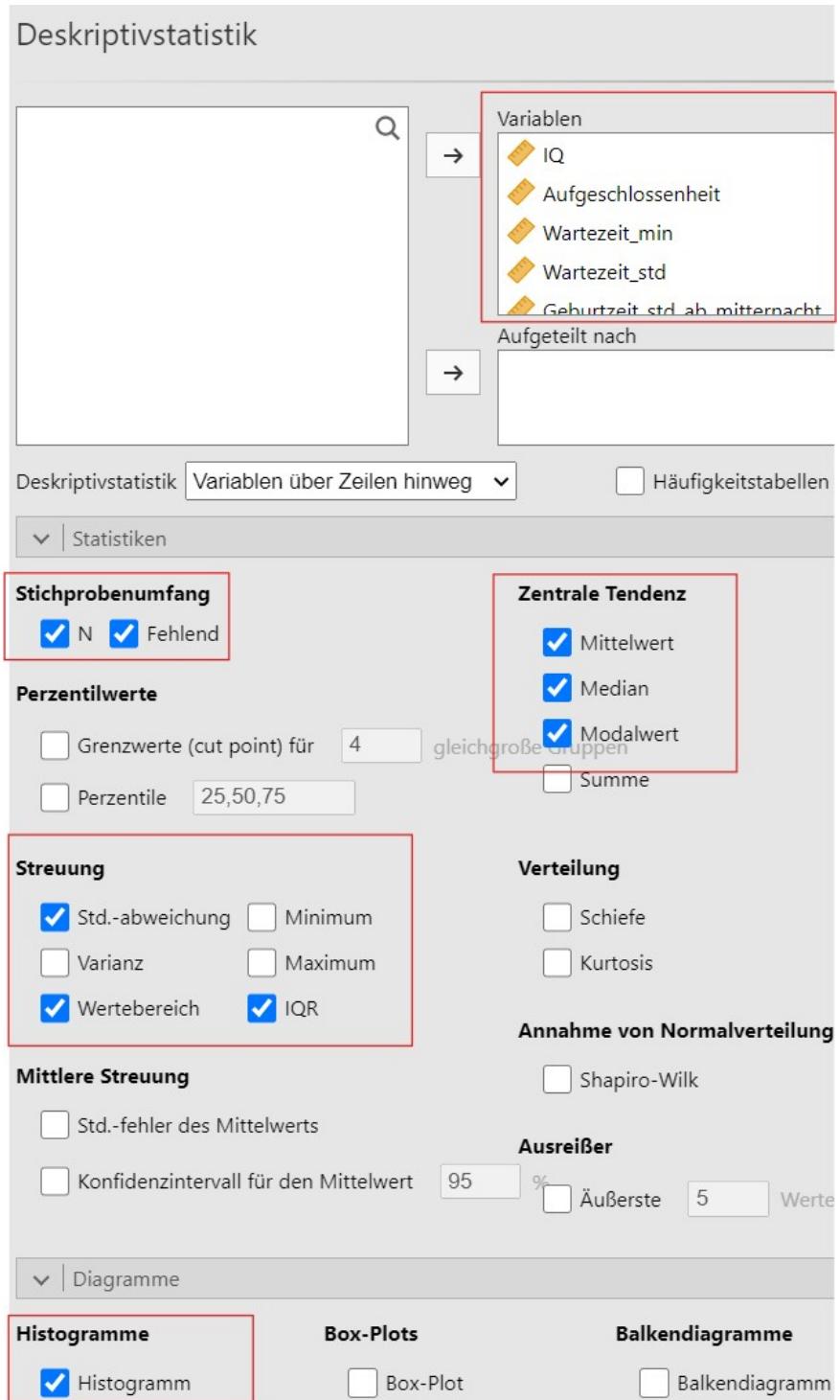


Abbildung 2.14: Jamovi Eingabe.

Deskriptivstatistik

	N	Fehlend	Mittelwert	Median	Modalwert	Std.-abw.	IQR	Wertebereich
IQ	500	0	99.6923	99.8294	53.7478*	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*	5.2642	5.1358	42.356
Wartezeit_std	500	0	0.0846	0.0558	2.47e-4*	0.0877	0.0856	0.706
Geburzzeit_std_ab_mitternacht	500	0	11.8111	12.1759	0.1157*	6.9384	11.8936	23.758
Geburzzeit_std_ab_mittag	500	0	11.6191	11.5215	0.0310*	6.8985	11.9415	23.953
Punkte	500	0	18.0860	18.0000	19.0000	1.2845	2.0000	6.000

\* Es gibt mehr als einen Modalwert, nur der erste wird berichtet

Abbildung 2.15: Deskriptive Statistiken.

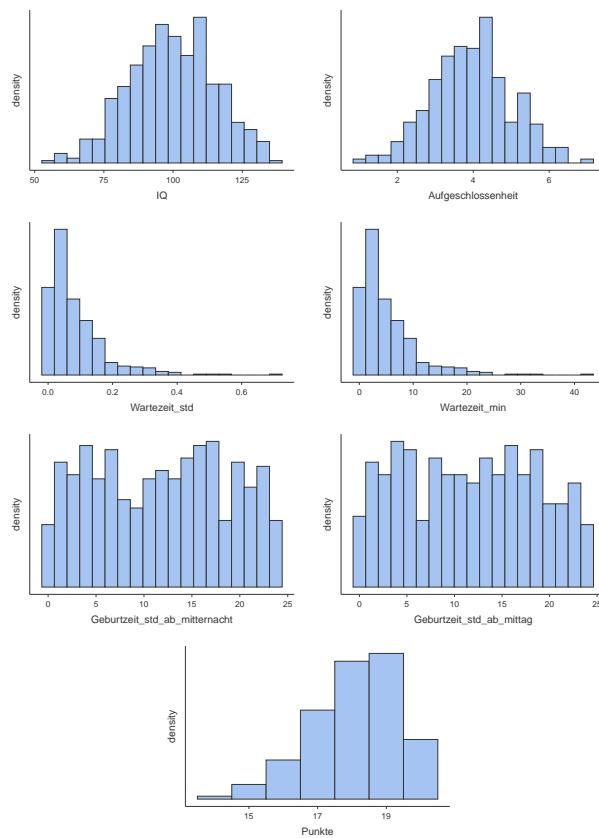


Abbildung 2.16: Histogramme.

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 identifizierbar, Streuung auch nicht.

## 2.4 Test

### Übung 2.7.

Welche der folgenden Merkmale 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 2.3
- 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 Inventory *STAI* entwickelt, welches Angst misst (Spielberger et al., 1983). Sie unterscheiden dabei zwischen Zustandesangst und dem Persönlichkeitszug Ängstlichkeit. Hier interessiert uns nur die Zustandesangst, welche fortan Angst genannt wird und misst wie grosse Angst aktuell empfunden wird. Die so gemessene Angst entspricht 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 3.1 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.

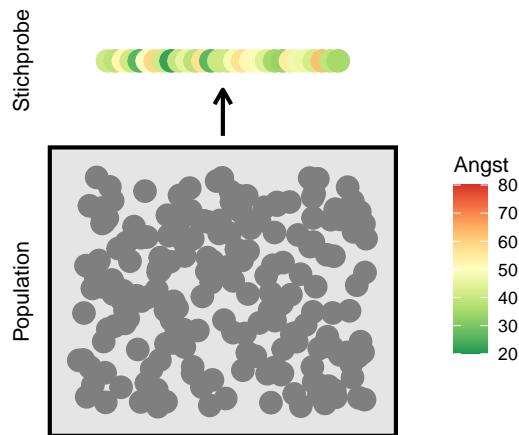


Abbildung 3.1: Population mit unbekannter Angst.

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 3.2 entstammt.

Es könnte aber auch sein, dass die Stichprobe einer Population mit viel höherer Zusatzzustandsangst, wie in Abbildung 3.3 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,

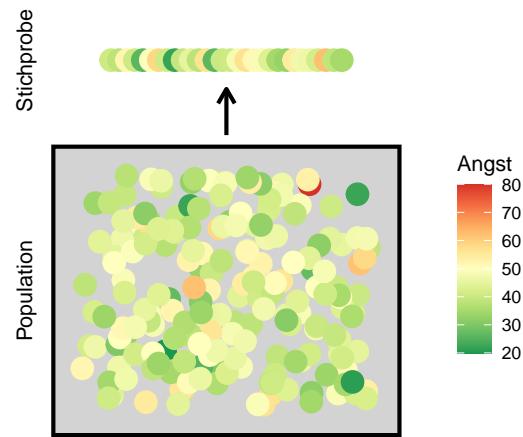


Abbildung 3.2: Population mit ähnlichen Angst-Werten wie in der Stichprobe.

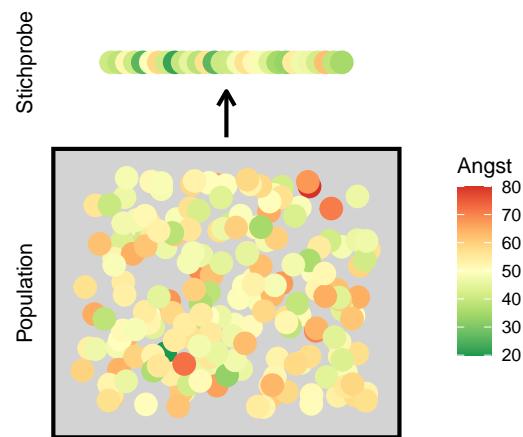


Abbildung 3.3: Population mit höheren Angst-Werten als in der Stichprobe.

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 3.4. Stichproben mit eher tiefen Angst-Werten kommen hier häufig vor.

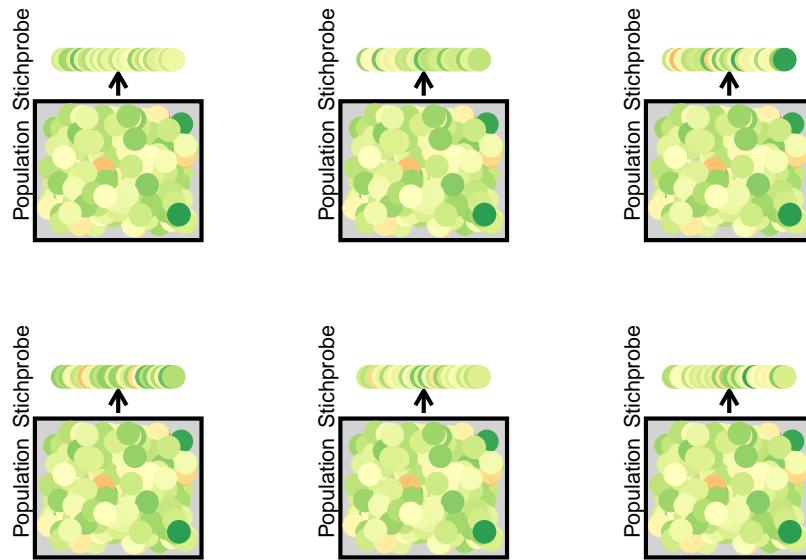


Abbildung 3.4: Wiederholte Stichprobenziehung bei gleichbleibender Population mit eher tiefen Angst-Werten.

Im Szenario, in welchem in der Population tatsächlich höhere Werte häufig vorkommen, kann dies aussehen wie in Abbildung 3.5. 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

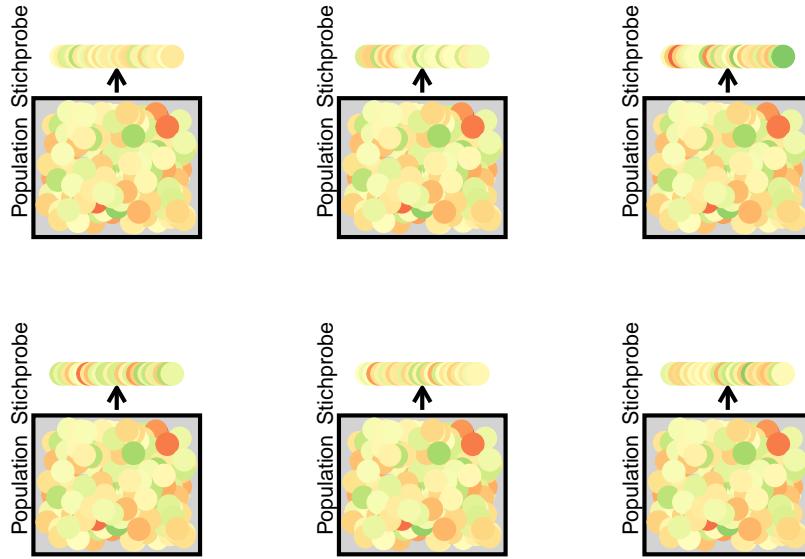


Abbildung 3.5: Wiederholte Stichprobenziehung bei gleichbleibender Population mit eher hohen Angst-Werten.

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 Phänotyp auftreten) werden 20 rothaarige Studierende und 54 Studierende mit anderer 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 ausgegangen 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
- Teilnehmende 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 Armdrücke 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 Adressen 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 3.2 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  $\mu$  [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 3.1 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.

## 50 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

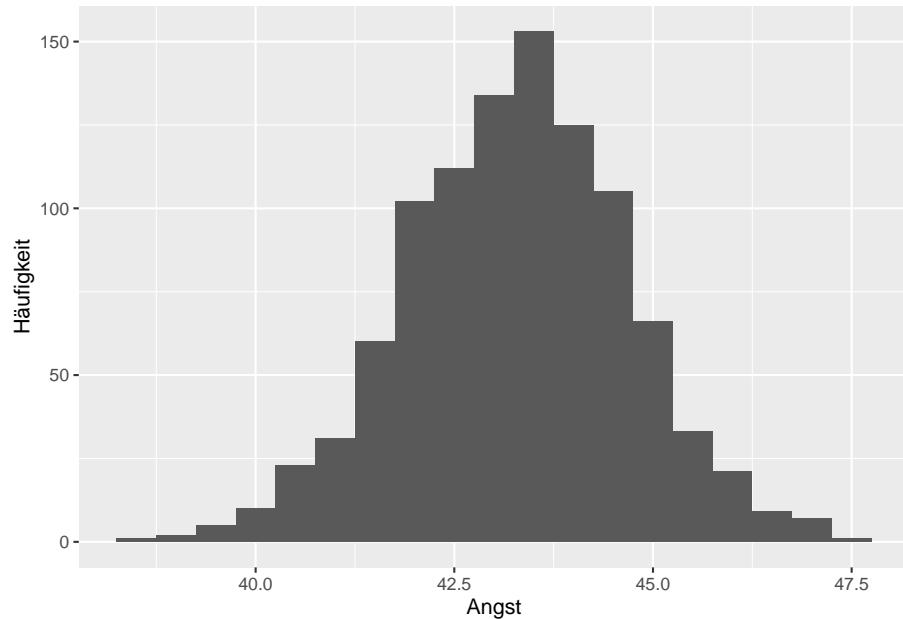


Abbildung 4.1: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

Die Häufigkeitsverteilung der berechneten arithmetischen Mittel in Abbildung 4.1 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 4.2 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* (Gosling et al., 2003). 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  $M = 3.91, s = 1.73$ .

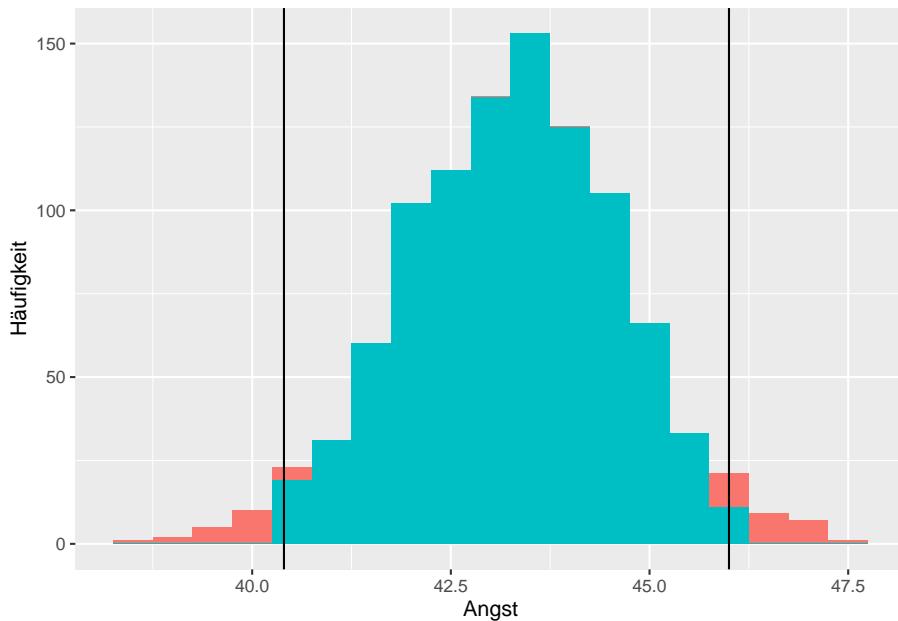


Abbildung 4.2: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

Die Verteilung der Beobachtungen, siehe Abbildung 4.3, 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 4.4 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 4.2 und 4.4 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 Gauss'sche Glockenkurve nach Carl F. Gauss (1777-1855) benannt. Die Normalverteilung kann mit nur zwei Parametern beschrieben werden.

## 52 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

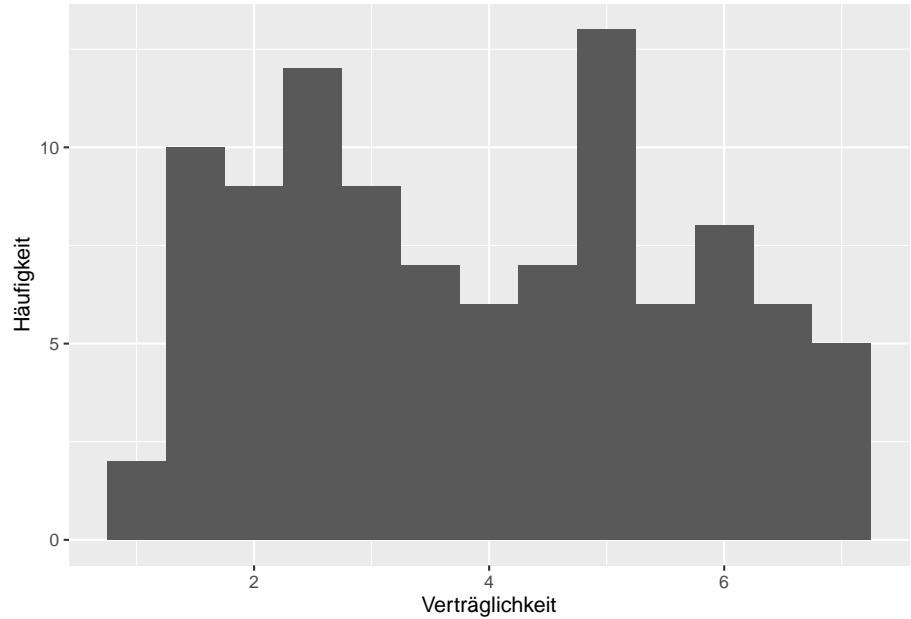


Abbildung 4.3: Verteilung der 100 beobachteten Verträglichkeitswerte einer zufällig gezogenen Stichprobe.

- $\mu_g$  gibt an, wo auf der x-Achse der höchste Punkt der Glocke liegt
- $\sigma_g$  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](http://seeing-theory.brown.edu) > Continuous > Normal kann der Einfluss von  $\mu_g$  und  $\sigma_g$  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  $\mu$  und Standardabweichung  $\sigma$ , der Durchschnitt aller Stichprobenwerte einer Normalverteilung mit  $\mu_g = \mu$  und  $\sigma_g = \frac{\sigma}{\sqrt{n}}$  entspricht, wobei  $n$  die Stichprobengrösse und  $\sigma$  die Standardabweichung des Merkmals in der Population bezeichnet.

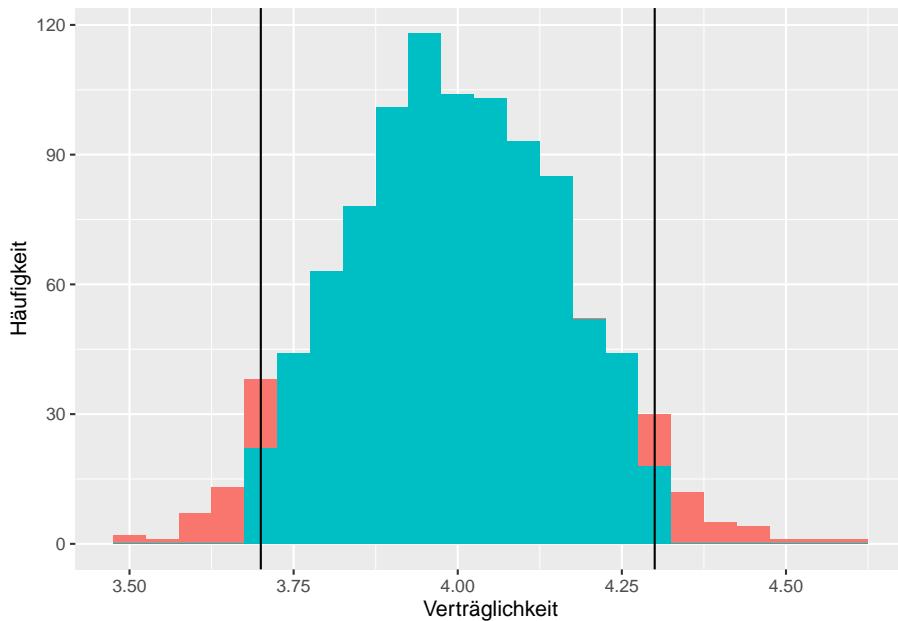


Abbildung 4.4: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Verträglichkeit.

**Achtung**

**Hinweis.**

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

Die Abbildungen 4.5 und 4.6 illustrieren den zentralen Grenzwertsatz für

## 54 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

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

Beispiel	Stichprobe		Normalverteilung		t-Verteilung	
	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

Beispiel 3.1 und 4.1 respektive, wobei die Normalverteilung der roten Linie entspricht. Dabei wird einstweilen angenommen, dass  $\mu$  und  $\sigma$  bekannt sind. Diese Annahme wird später aufgelöst und dient hier lediglich der Illustration.

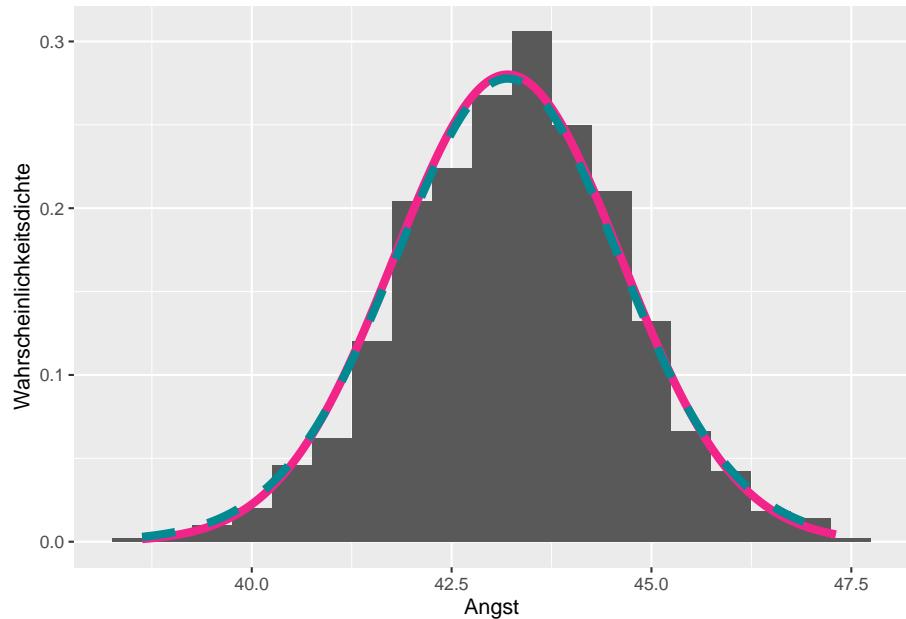


Abbildung 4.5: Die arithmetischen Mittel sind Normalverteilt mit Parametern  $\mu_g = 43.34$  und  $\sigma_g = 9.72/\sqrt{30}$ .

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 4.1 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 röhrt daher, dass der zentrale Grenzwertsatz nur dann exakt funktioniert, wenn die Anzahl Beobachtungen (unendlich) gross ist.

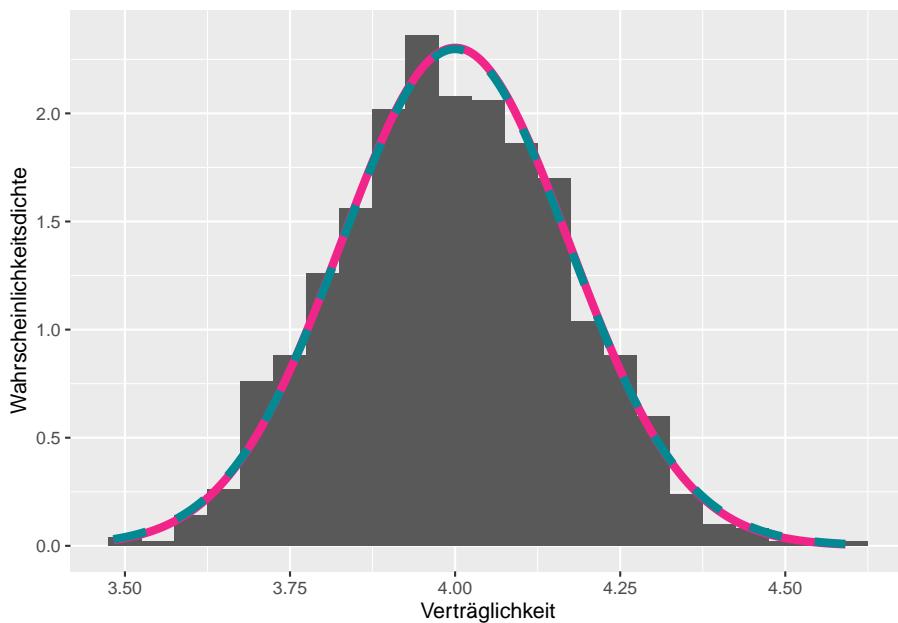


Abbildung 4.6: Die arithmetischen Mittel sind Normalverteilt mit Parametern  $\mu_g = 3.91$  und  $\sigma_g = 1.73/\sqrt{100}$ .

## 56 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

Einstweilen wurde hier angenommen, dass die Streuung des Merkmals  $\sigma$  bekannt ist. Dies ist in der Realität nie der Fall und eine weitere, wenn auch weniger grosse Ungenauigkeitsquelle. Wenn  $\sigma$  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 4.5 und 4.6 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 3.1 mit  $n = 30$  ist ein kleiner Unterschied, in Beispiel 4.1 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 4.7 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.

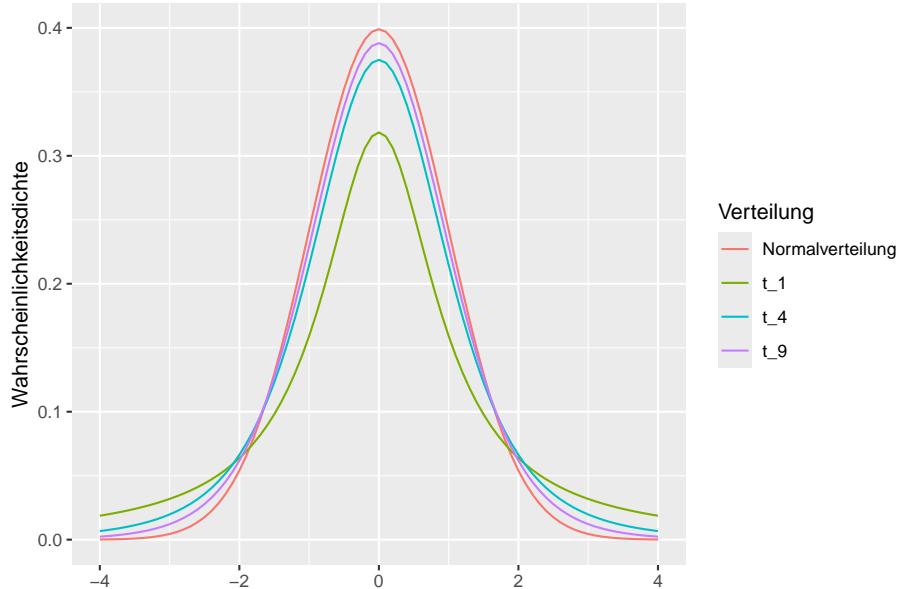


Abbildung 4.7: Student-t-Verteilungen mit 1, 4 und 9 Freiheitsgraden im Vergleich zu der Normalverteilung.

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

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

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. Letztere 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 Erwartungswert 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  $\alpha$  bezeichnet. Es ist demnach äquivalent von einem 99% Konfidenzintervall oder von einem Konfidenzintervall mit 1% Irrtumswahrscheinlichkeit zu sprechen.

In Beispiel 3.1, kann aus der Tabelle 4.1 entnommen werden, dass die Angst in der Population bei  $M = 43.34$  95% KI [42.18, 44.5] liegt. In Beispiel 4.1, kann aus der Tabelle 4.1 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 4.4 und 4.5 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 4.8 gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung 4.9.

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.

- 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?
- Vergleichen Sie die Längen der errechneten Konfidenzintervalle.
- 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 Sandstrom et al. (2022).

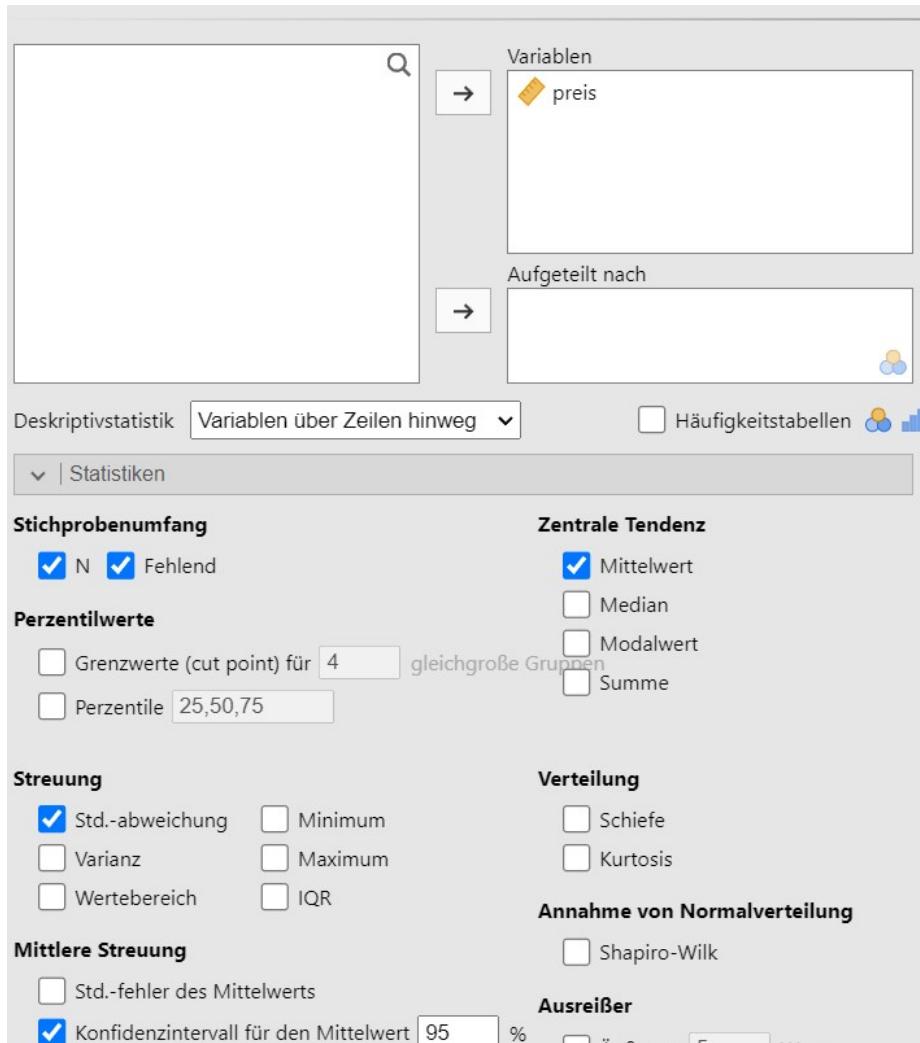


Abbildung 4.8: Jamovi setzen der Analyseparameter.

## 60 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

Deskriptivstatistik

	N	Fehlend	Mittelwert	95% Konfidenzintervall		
				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 4.10 gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung 4.11.

- (a) Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei  $M = 3.784, 95\% \text{ KI } [3.764, 3.803]$  Punkten und am Ende der Woche bei  $M = 3.328, 95\% \text{ KI } [3.309, 3.348]$  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 4.10 dargestellten Maske der Wert für Konfidenzintervall für den Mittelwert auf 90 gesetzt werden. Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei  $M = 3.784, 90\% \text{ KI } [3.767, 3.800]$  Punkten und am Ende der Woche bei  $M = 3.328, 90\% \text{ KI } [3.312, 3.345]$  Punkten. Die durchschnittliche Unangenehmheit lag am Anfang der Woche bei  $M = 3.784, 99\% \text{ KI } [3.758, 3.809]$  Punkten und am Ende der Woche bei  $M = 3.328, 99\% \text{ KI } [3.303, 3.354]$  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.

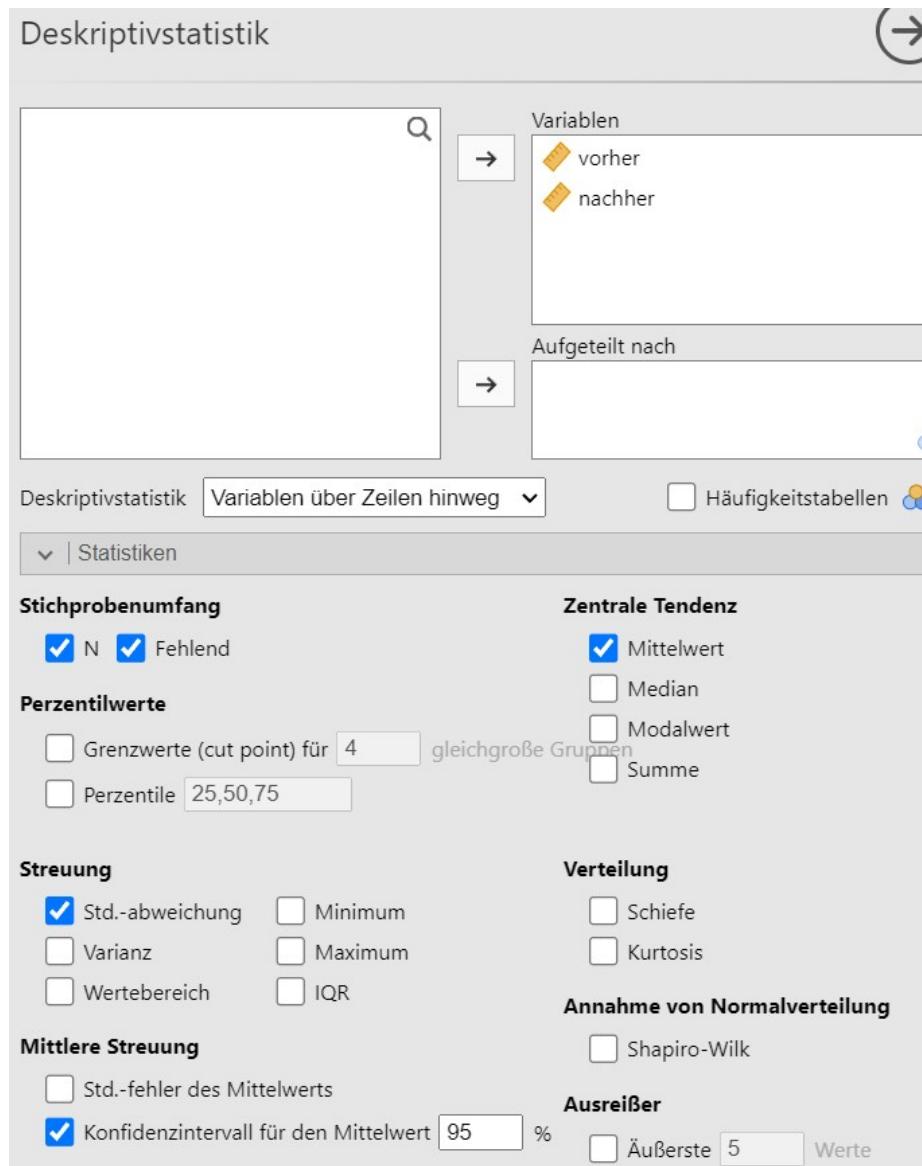


Abbildung 4.10: Jamovi setzen der Analyseparameter.

## 62 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

Deskriptivstatistik

	N	Fehlend	Mittelwert	95% Konfidenzintervall		
				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 4.12 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 einer 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 4.13 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 4.12. 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.

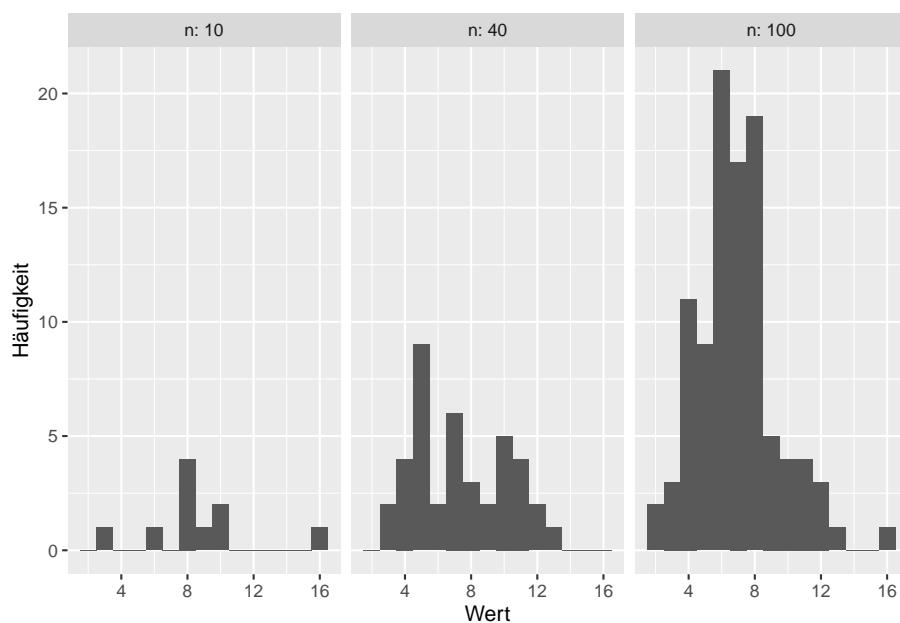


Abbildung 4.12: Verteilung der Werte einer Stichprobe. (Achtung es handelt sich hierbei nicht um Mittelwerte, sondern die Effektiv gemessenen Beobachtungen einer Stichprobe.)

## 64 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

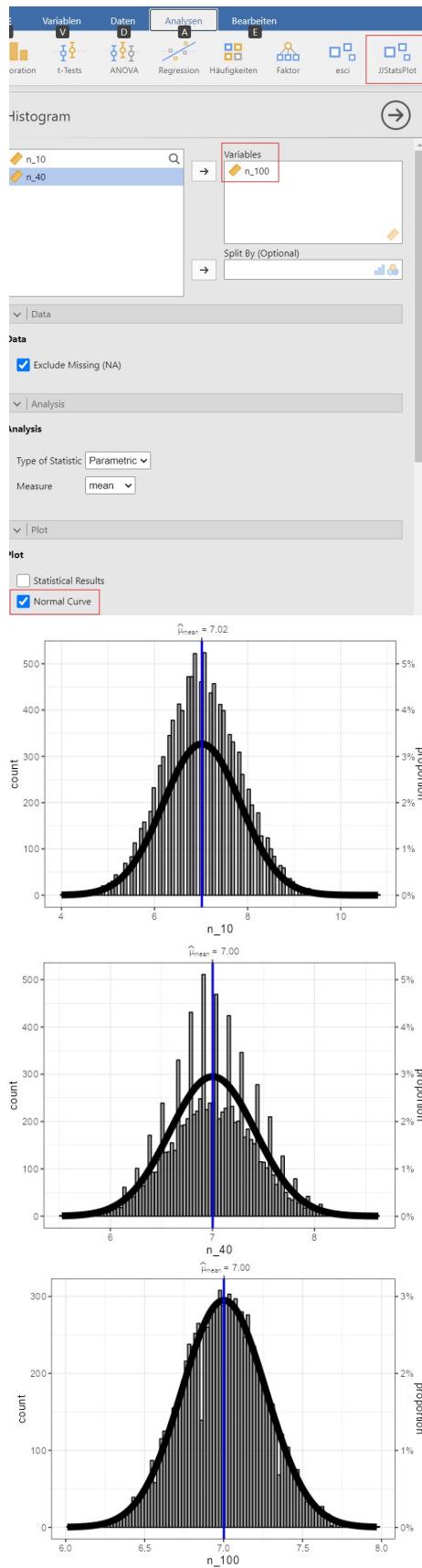


Abbildung 4.13: Jamovi-Eingabeeinstellungen und die Histogramme der Mittel-

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 4.14 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.

## 66 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

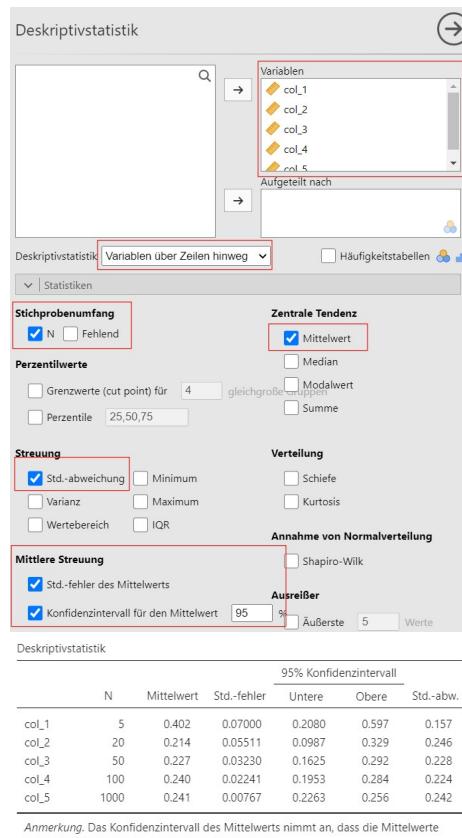


Abbildung 4.14: Links: Jamovi-Anleitung zur Erstellung der Tabelle mit den relevanten Kenngrößen; rechts: Tabelle mit relevanten Kenngrößen.

- 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 4.15 gesetzt. Die Nachkommastellen können im Menu oben rechts bei den drei vertikalen Punkten eingestellt werden.

Dies produziert das Analyseergebnis in Abbildung 4.16.

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  $M = 14.8$  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 (4.1) 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.*

68 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

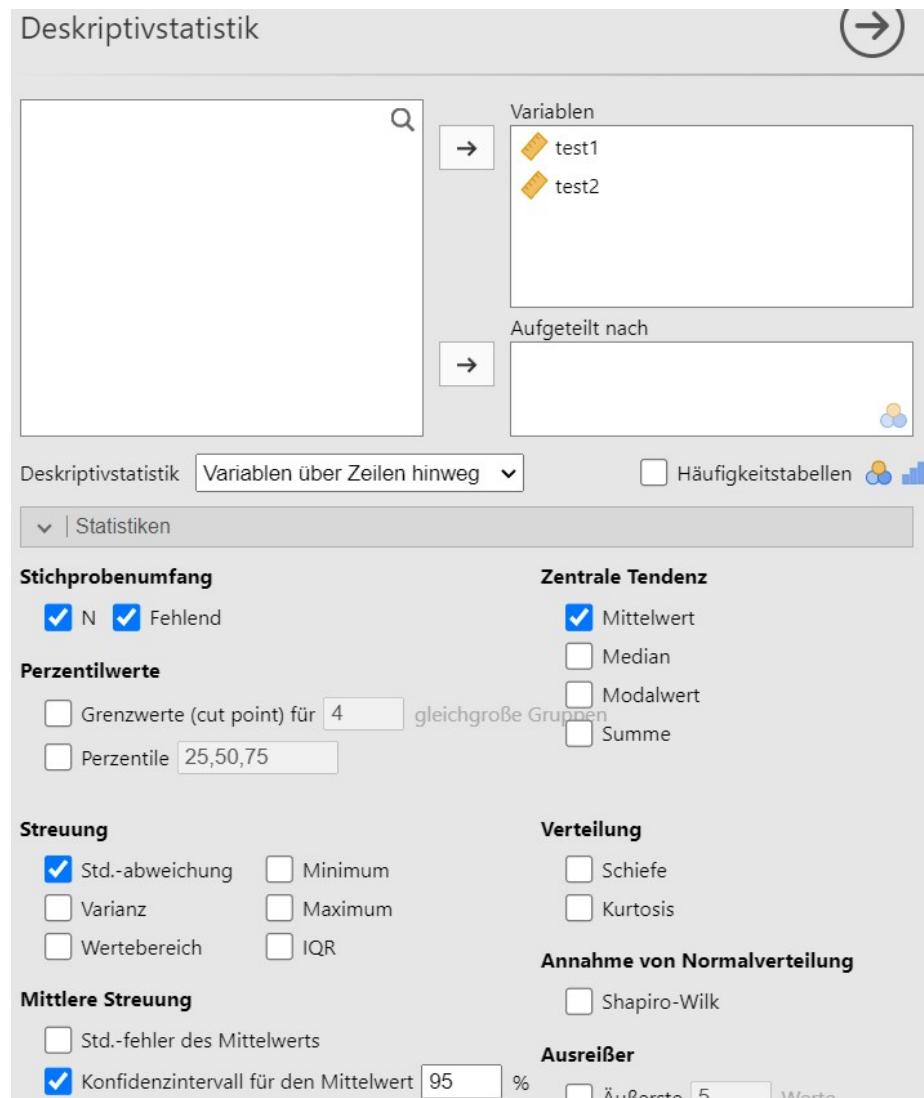


Abbildung 4.15: Jamovi setzen der Analyseparameter.

## Deskriptivstatistik

	N	Fehlend	Mittelwert	95% Konfidenzintervall		
				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.

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

**Übung 4.7.** Im Datensatz 02-exr-koerpergrösse-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

70 KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

# 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

## 5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT? 75

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 3 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 5.1 links) oder ob es eine eher häufig vorkommende Stichprobenziehung aus einer Population mit höherem durchschnittlichen Vermögen ist (Abbildung 5.1 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 5.2, 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*,

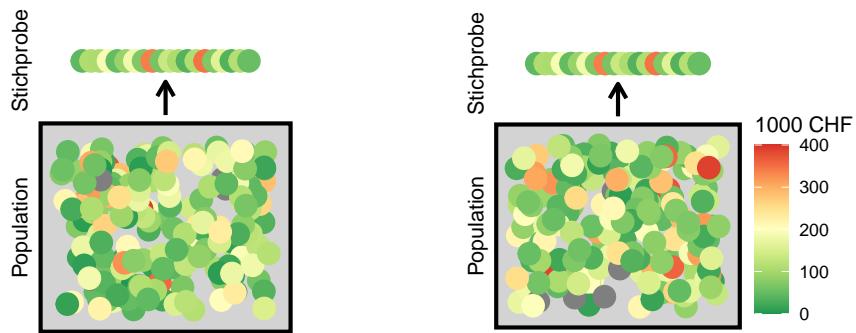


Abbildung 5.1: Vorgestellte Zufallsstichprobenziehung. Links: Nullhypothese ist wahr. Rechts: Nullhypothese ist falsch. Die grauen Punkte entsprechen Vermögen über 400'000 CHF.

### 5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT? 77

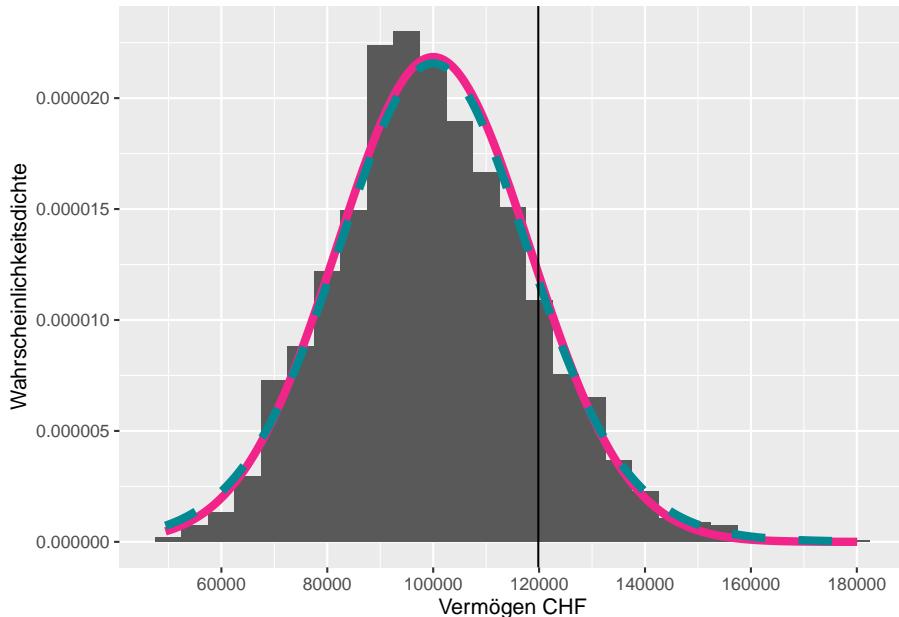


Abbildung 5.2: TODO.

**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  $\alpha$ , 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 Denkartens 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 5.2 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  $\mu$  und Standardabweichung  $\sigma$  hat, der Wert

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

normalverteilt ist, wobei  $\mu$  hier dem Wert der Nullhypothese entspricht, also

## 5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT?79

100'000 CHF. Dies entspricht der roten Linie in Abbildung 5.2. Ist die Standardabweichung des Merkmals  $\sigma$  in der Population unbekannt, so wird diese mit der Standardabweichung in der Stichprobe  $s$  geschätzt. Diese zusätzliche Unsicherheit führt dazu, dass

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (5.1)$$

nicht mehr normal-, sondern  $t$ -verteilt ist bei  $n - 1$  Freiheitsgraden (grüne Linie, Abbildung 5.2). 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 5.3 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,  $t(19) = 1.003$ ,  $p = .164$ .

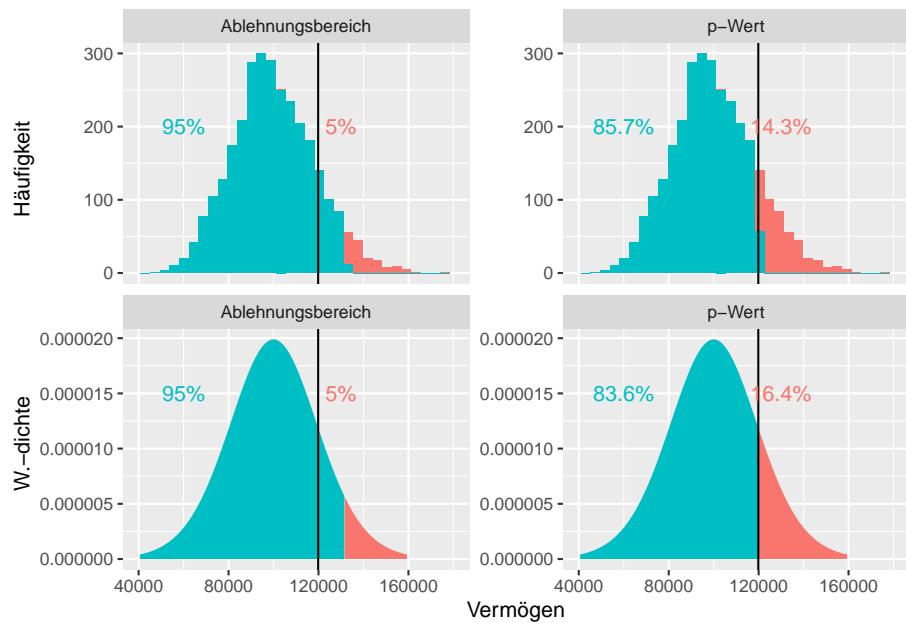


Abbildung 5.3: Oben: Histogramm der simulierten Verteilung; unten: theoretische t-Verteilung; links: Illustration p-Wert; rechts: Illustration Ablehnungsbereich. Die Linie entspricht dem beobachteten Stichprobenmittelwert.

**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 (5.1) aus der Stichprobe. Dieser Wert ist skaliert und muss im Kontext der standardisierten  $t$ -Verteilung wie in Abbildung 4.7 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:  $\alpha$  = ohne,  $\lambda\epsilon\xi\theta\mu\nu$  = lesen, sprechen,  $\theta\mu\nu$  = 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 Alternativhypotesen aufzustellen. Die Psychologin stellt die Frage, ob sich die durchschnittliche Alexithymie in der Grundgesamtheit, in der Folge mit  $\mu$  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

$$H_0 : \mu = 100 \text{ Punkte.}$$

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

$$H_1 : \mu \neq 100 \text{ Punkte.}$$

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 5.4 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 5.4 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 5.4 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 5.4 unten rechts) und für den Ablehnungsbereich (roter Bereich in Abbildung 5.4 unten links).

Die Psychologin kann nun wie folgt berichten:

## 5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT? 83

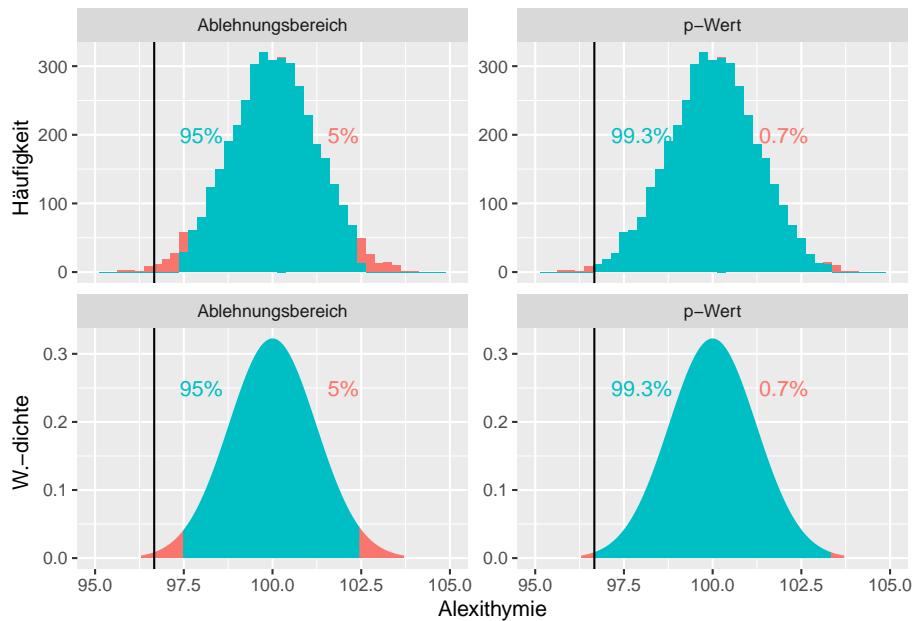


Abbildung 5.4: Oben: Histogramm der simulierten Verteilung der Alexithymie-Mittelwerte; unten: theoretische t-Verteilung; links: Illustration p-Wert; rechts: Illustration Ablehnungsbereich. Die Linie entspricht dem beobachteten Stichprobenmittelwert.

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 5.1 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 werden indem in der Formel (5.1) verschiedene Differenzen eingesetzt werden und mit der Abbildung 4.7 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 (Cohen, 1988) 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 5.1 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 5.2 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.  $H_0: \mu > 100000$

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 Czeisler et al. (1999) 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?



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.  $H_0: \mu = 100$

### 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.

- 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.  
 c) Erklären Sie alle Zahlen und Symbole im Testbericht.

*Lösung.* Für diese Übung werden die Daten in Jamovi wie in Abbildung 5.9 analysiert. Das Resultat der Analyse ist in Abbildung 5.10 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.  $H_0: \mu \neq 24$

#### 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 Versuchspersonen 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 Versuchsbedingungen nicht verändert. Der normale zirkadiane Rhythmus ist Sonnenbedingt 24 Stunden lang, also wird die Nullhypothese  $H_0 : \mu = 24$  Stunden aufgestellt.  $\mu$  ist hier die durchschnittliche 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 Tagedauer 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 4.7 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 5.11 analysiert. Das Resultat der Analyse ist in Abbildung 5.12 festgehalten.



Abbildung 5.11: Jamovi Eingabe.

- 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

## 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.  $H_0: \mu < 1.58$

## Deskriptivstatistik

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

Abbildung 5.12: Jamovi Ausgabe.

die Zufallsstichprobenziehung zustande kommt. Kurz gesagt, das Resultat könnte 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 signifikant nachweisen zu können. Falls der gefundene Effekt nur zufällig zustande gekommen ist und er nicht existiert, 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.

- Stellen Sie die oben formulierte Hypothese mit mathematischer Schreibweise dar.
- Testen Sie die Hypothese.
- Berichten Sie die Testergebnisse.
- Was bedeuten die Werte Statistik,  $df$ ,  $p$  und Effektstärke.
- 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 5.13 analysiert. Das Resultat der Analyse ist in Abbildung 5.14 festgehalten.

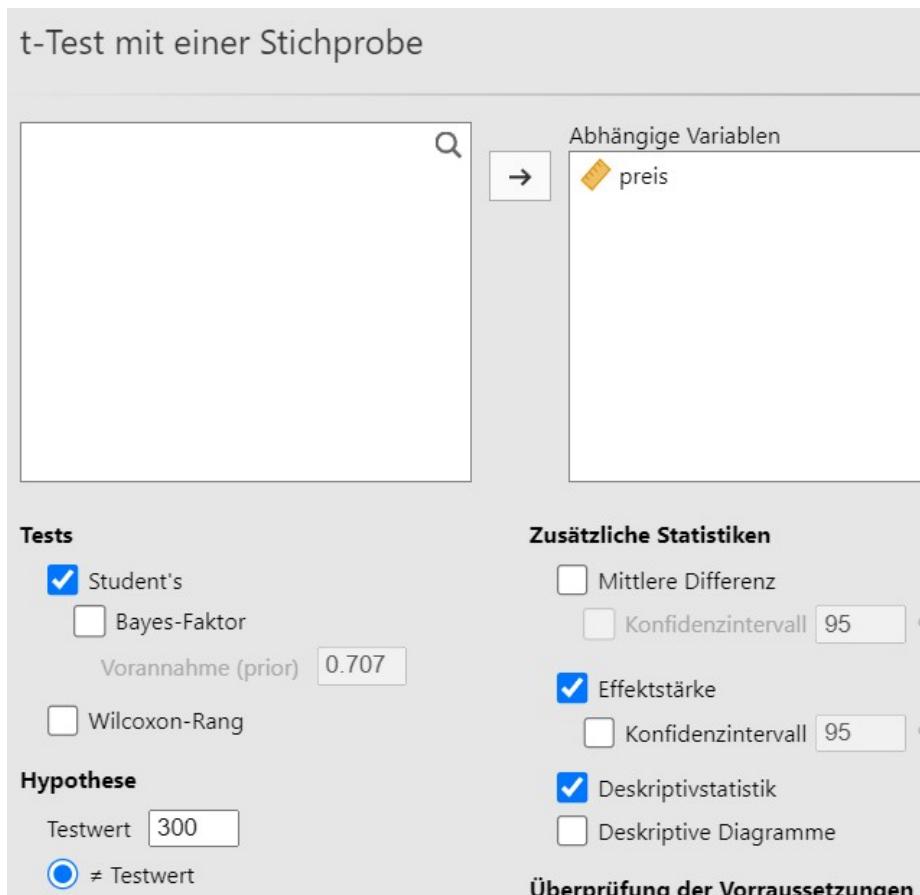


Abbildung 5.13: Jamovi Eingabe.

- Es gibt zunächst keinen Anhaltspunkt, weshalb sich die Zahlbereitschaft geändert haben sollte. Deshalb ist die Nullhypothese  $H_0 : \mu = 300$  CHF, wobei  $\mu$  für den Erwartungswert des Merkmals Preis ist. Pear fragt sich, ob der  $\mu$  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.
- Das Testen erfolgt wie oben in den Bildschirmaufnahmen von Jamovi dargestellt.
- 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,  $t(69) = -0.848$ ,  $p = .399$ ,  $d = -0.101$ .

## t-Test mit einer Stichprobe

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

Anmerkung.  $H_0: \mu = 300$

## Deskriptivstatistik

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

Abbildung 5.14: Jamovi Ausgabe.

- d) Statistik entspricht der beobachteten Teststatistik in der Stichprobe. Der Wert kann im Vergleich zur  $t$ -Verteilung in Abbildung 4.7 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 4.7 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 4.7. 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 Nullhypothese 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). Morris et al. (2015) 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 6.1 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 6.1 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 6.2 dargestellt.

Das Testprinzip funktioniert genau gleich wie beim *t*-Test für eine Stichprobe wie in Kapitel 5. 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

$$\begin{aligned} H_0 : \mu_{\text{Frau}} &= \mu_{\text{Mann}} \\ H_1 : \mu_{\text{Frau}} &\neq \mu_{\text{Mann}}. \end{aligned}$$

Dies entspricht, einfacher Arithmetik folgend,

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

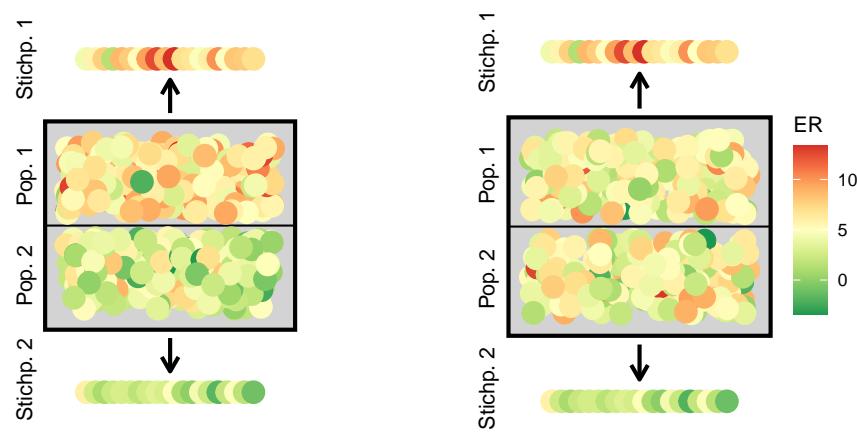


Abbildung 6.1: Links: Zwei Stichprobenziehungen aus zwei Populationen mit unterschiedlichen Mittelwerten. Rechts: Zwei Stichprobenziehungen aus einer Population, bzw. aus zwei Populationen die sich bezüglich ihrer Werte nicht unterscheiden.

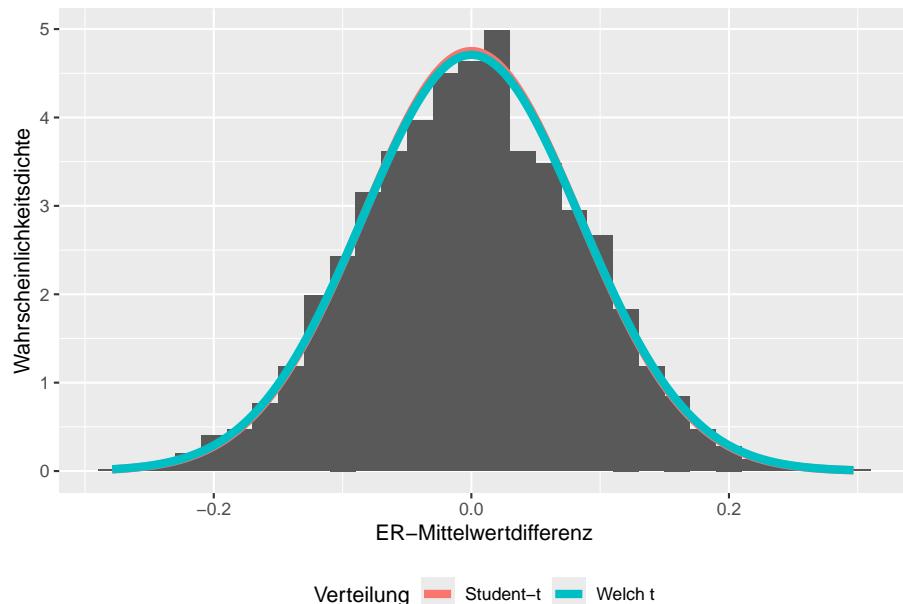


Abbildung 6.2: Verteilung simulierter ER-Mittelwertdifferenzen bei wiederholten Zufallsstichprobenziehung. Rot: Annäherung der Verteilung mit dem Student t-Test; grün: Annäherung der Verteilung durch den Welch-Test.

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 6.3 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 unterscheiden.

```
## 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()`).
```

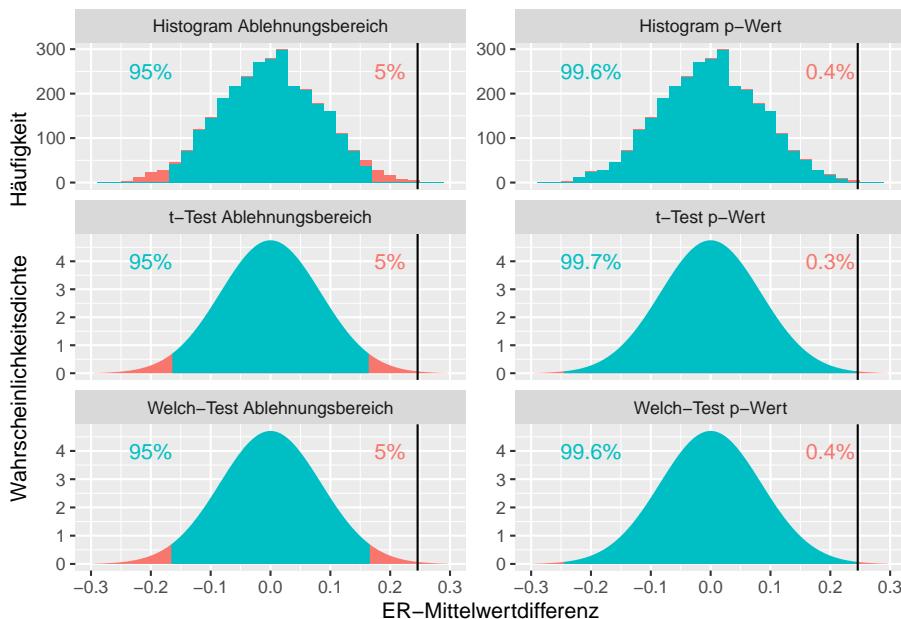


Abbildung 6.3: TODO

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 6.3 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

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \omega_0}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (6.1)$$

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 6.2 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 (Macleod et al., 1996). Gemessen wird dabei die Reaktionsgeschwindigkeit *RT* in Millisekunden. In einem Versuch wollten Khanna et al. (2017) 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

$$\begin{aligned} H_0 : \mu_{\text{PTSD}} &= \mu_{\text{non-PTSD}} \\ H_1 : \mu_{\text{PTSD}} &\neq \mu_{\text{non-PTSD}}. \end{aligned}$$

In diesem Beispiel sind die Standardabweichungen und demnach auch die Varianzen der Reaktionszeiten in den beiden Gruppen sehr unterschiedlich. Wenn das Experiment 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 6.4 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.

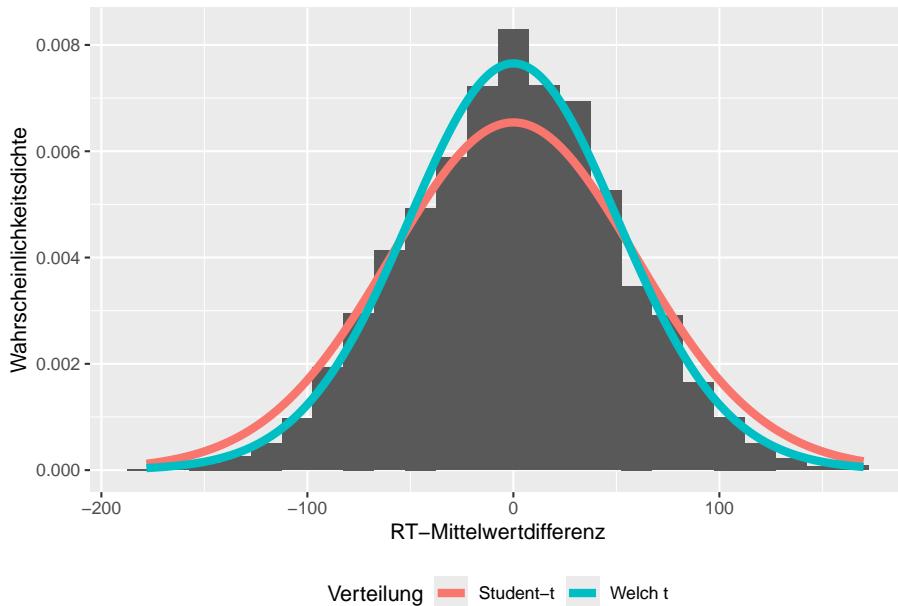


Abbildung 6.4: Verteilung simulierter RT-Mittelwertdifferenzen bei wiederholten Zufallsstichprobenziehungen. Rot: Annäherung der Verteilung mit dem Student *t*-Test; grün: Annäherung der Verteilung durch den Welch-Test.

Für diesen Fall wurde von Welch (1947) eine alternative Annäherung an die Verteilung der Mittelwertdifferenzen gefunden, nämlich

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \omega_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad (6.2)$$

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 6.4.

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 6.4. Wenn beide Varianzen ungefähr gleich sind, so geben beide Tests jedoch ähnlich gute Resultate, siehe Abbildung 6.2. Es wird deshalb empfohlen immer den Welch-Test durchzuführen (Zimmerman, 2004). Ein Vergleich der Abbildungen 6.3 und 6.5 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 (6.1) und (6.2) 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 Stichprobengrössen 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 5, 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:

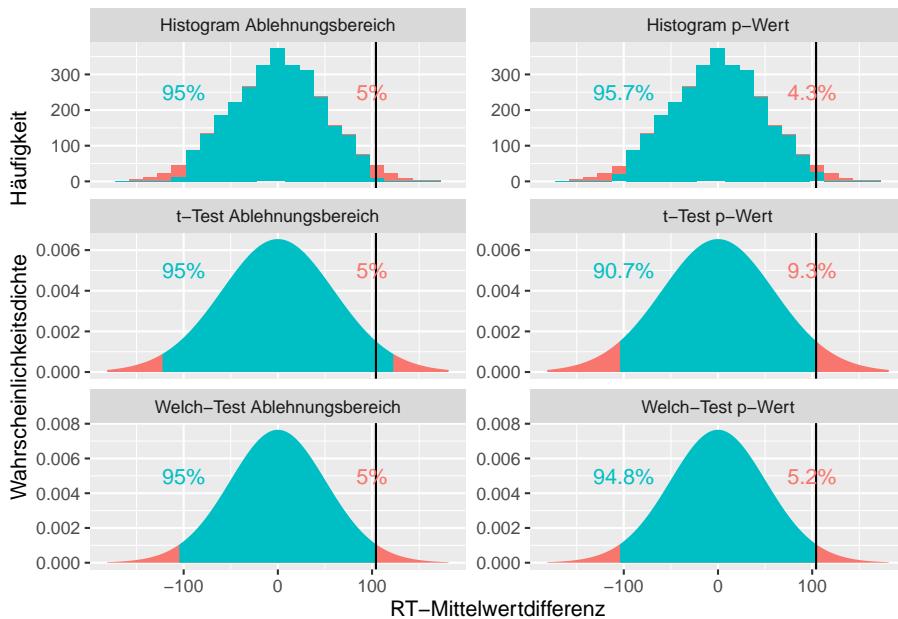


Abbildung 6.5: TODO

- 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 entspricht 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 entspricht dem **Cohens d** für den Welch-Test in **Jamovi**.

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

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

Glass Δ wird in **Jamovi** nicht standardmäßig ausgegeben und muss handisch berechnet werden.

Da es sich bei beiden Beispielen nicht um Experimente handelt, weil weder das Geschlecht noch die posttraumatische Belastungsstörung zufällig zugeordnet 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.

<b>Achtung</b>
 <p><i>Hinweis.</i></p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• In <b>Jamovi</b> wird für den Zweistichproben-t-Test und den Welch-Test eine unterschiedliche Effektstärke angegeben.</li> </ul>

## 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 Bandura et al. (1961).

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 Testhypotesen 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 6.6.

## 110 KAPITEL 6. GRUPPENMITTELWERTUNTERSCHIED BEI EINEM INTERVALLSKALIERTEN

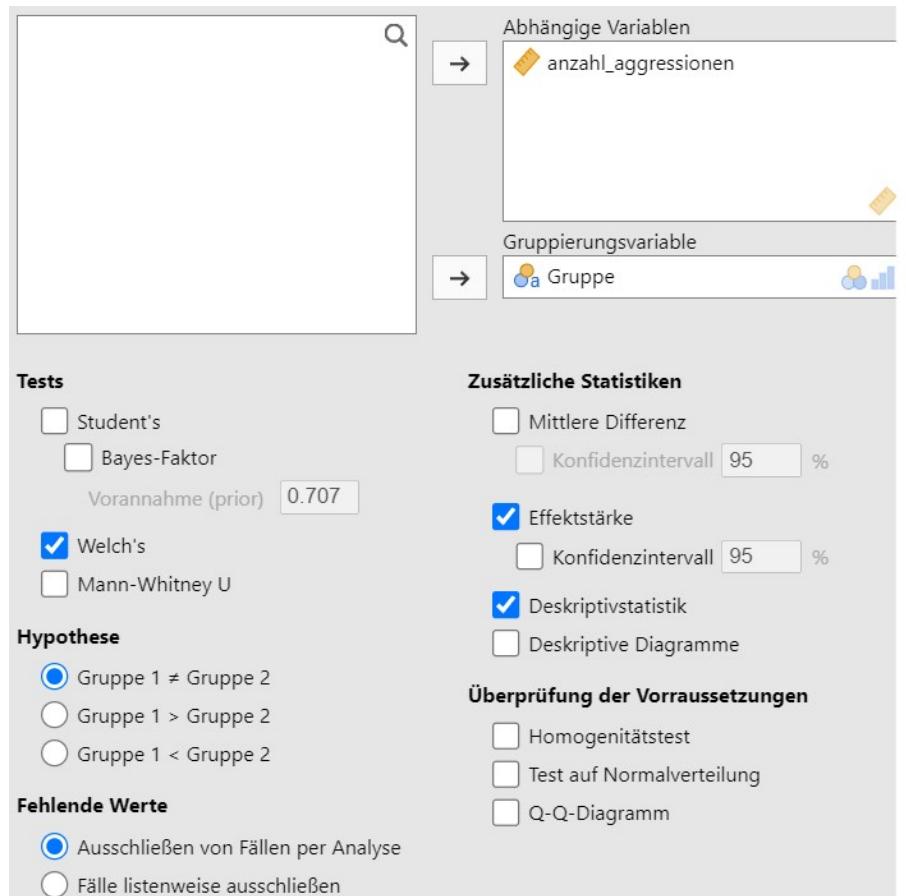


Abbildung 6.6: Jamovi Eingabe.

### t-Test für unabhängige Stichproben

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

Anmerkung.  $H_0: \mu_{\text{aggressiv}} = \mu_{\text{nicht aggressiv}}$

### 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 6.7.

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  $\Delta$  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 Blaney et al. (1977) 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 Testhypotesen 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.

## 112 KAPITEL 6. GRUPPENMITTELWERTUNTERSCHIED BEI EINEM INTERVALLSKALIERTEN

*Lösung.* Zuerst wird der Datensatz mit Jamovi eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 6.8.

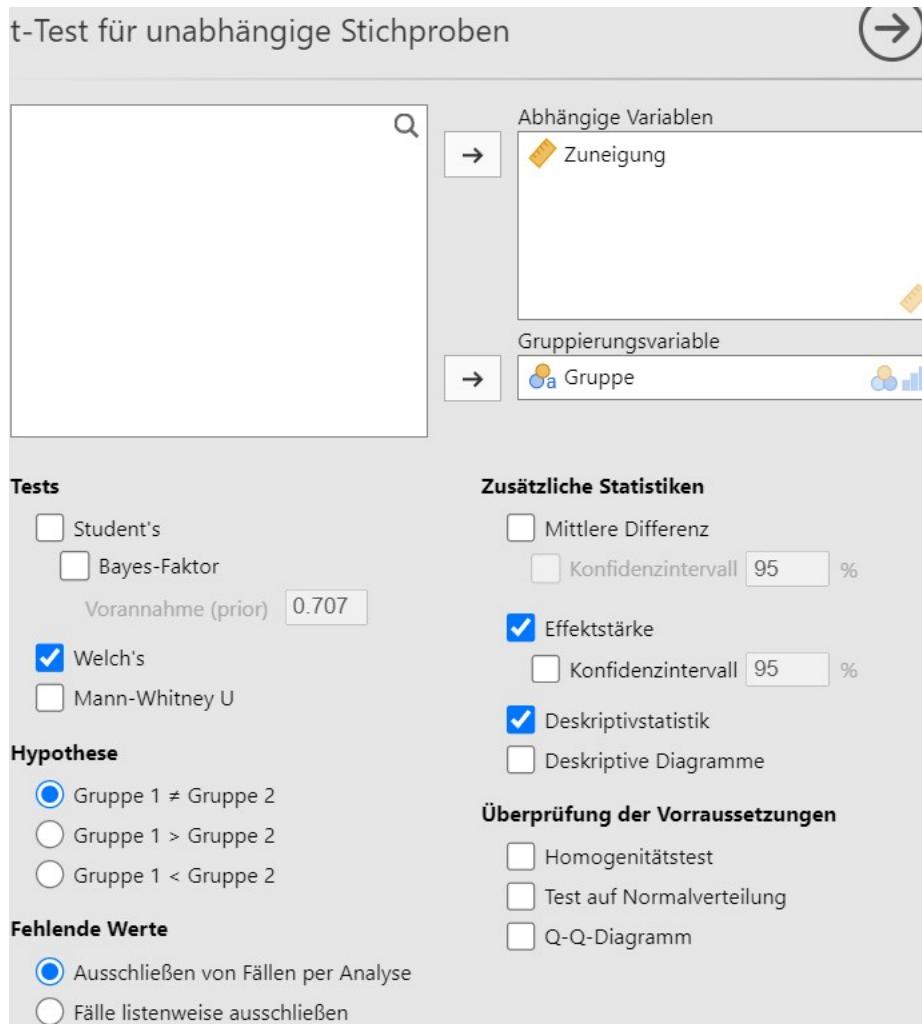


Abbildung 6.8: Jamovi Eingabe.

Dies produziert das Analyseergebnis in Abbildung 6.9.

Damit können nun die Teilfragen beantwortet werden.

- Angenommen die Zuneigung zur Personen der einen Gruppe ist unabhängig von der Lehrmethode, dann sollten beide Gruppen im durchschnitt denselben Erwartungswert  $\mu$  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.  $H_0: \mu_{\text{gruppenpuzzle}} = \mu_{\text{traditionell}}$

## 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 Gruppenpuzzleguppe ( $M = 5.11, SD = 0.25, N = 35$ ) signifikant anders ist als in der traditionell unterrichteten Gruppe ( $M = 4.82, SD = 0.31, N = 21$ ),  $t(35.6) = 3.69, p < 0.001, \Delta = 0.95$ .
- c) Die Statistik von 3.69 ist ein Wert, welcher eine Verteilung wie in 4.7 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.

## 114 KAPITEL 6. GRUPPENMITTELWERTUNTerschied BEI EINEM INTERVALSKALIERTEN

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 wieder. Die Beschallungsart wird den Studienteilnehmenden zufällig zugeordnet. Die Anzahl korrekt memorisierte Wortsilben sind im Datensatz **06-exr-music-memory.sav** verfügbar.

- Stellen Sie die Testhypthesen auf für einen zweiseitigen Welch-Test.
- Führen Sie den Test durch und berichten Sie das Resultat.
- Erklären Sie den Wert der Statistik, des  $p$ -Werts und der Effektstärke Cohen's  $d$  respektive.

*Lösung.* Zuerst wird der Datensatz mit **Jamovi** eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 6.10.

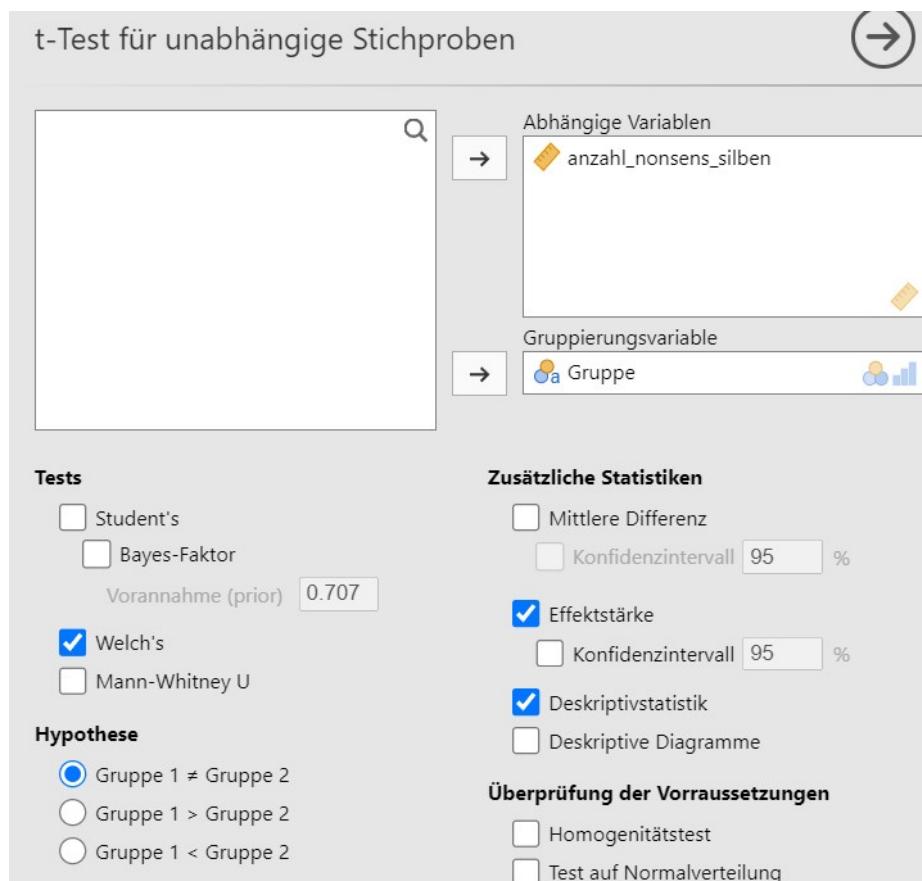


Abbildung 6.10: Jamovi Eingabe.

Dies produziert das Analyseergebnis in Abbildung 6.11.

t-Test für unabhängige Stichproben

		Statistik	df	p	Effektstärke	
anzahl_nonsense_silben	Welch's t	-0.6561	54.6398	0.514	Cohens d	-0.1527

Anmerkung.  $H_0: \mu_{\text{mit\_text}} = \mu_{\text{ohne\_text}}$

Deskriptivstatistik für die Gruppen

	Gruppe	N	Mittelwert	Median	Std.-abw.	Std.-fehler
anzahl_nonsense_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.

- 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 Wortsilben mit oder ohne Musik unterscheiden  $H_1: \mu_{\text{Musik mit Text}} \neq \mu_{\text{Musik ohne Text}}$ .
- 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 ( $M = 8.22$ ,  $SD = 1.06$ ,  $N = 43$ ),  $t(54.6) = -0.66$ ,  $p = .514$ ,  $d = -0.153$ .
- Die Statistik von  $-0.66$  ist ein Wert, welcher eine Verteilung wie in 4.7 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 aussergewöhnlich 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 Wortsilben 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.

- 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?
- Stellen Sie die Hypothesen für einen einseitigen Test auf.
- Prüfen Sie die Hypothesen mit einem geeigneten einseitig durchgeföhrten 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 6.12.

Dies produziert das Analyseergebnis in Abbildung 6.13.

Damit können und beide Teilfragen beantwortet werden.

- 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 zurückzuföhren ist. Um dieses Risiko zu quantifizieren und damit einzuschätzen, ob das Ergebnis auch für die Population gelten könnte, kann ein statistischer Test durchgeführt werden.
- 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  $H_0 : \mu_{\text{Hypnose}} \geq \mu_{\text{Placebo}}$ .
- 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 ( $M = 5.81$ ,  $SD = 0.87$ ,  $N = 15$ ),  $t(20.6) = -3.5$ ,  $p = .001$ ,  $\Delta = -1.022$ .

### Übung 6.5.

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

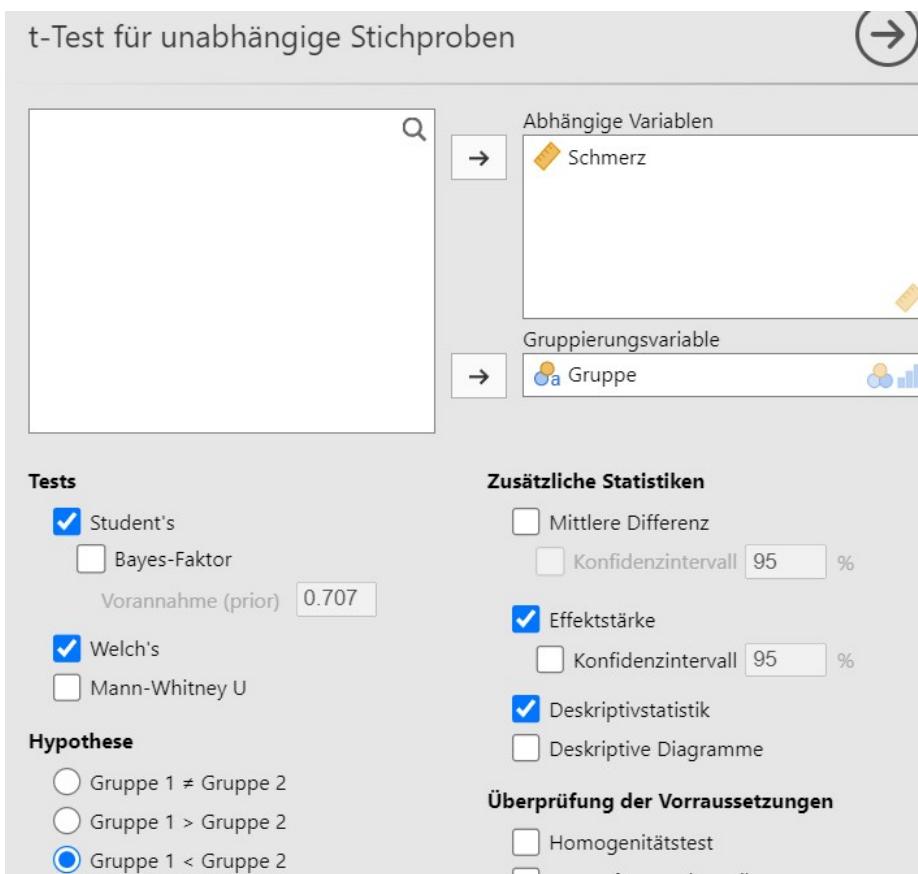


Abbildung 6.12: Jamovi Eingabe.

## 118 KAPITEL 6. GRUPPENMITTELWERTUNTERSCHIED BEI EINEM INTERVALSKALIERTEN

t-Test für unabhängige Stichproben

		Statistik	df	p	Effektstärke	
Schmerz	Student's t	-3.1486*	24.0000	0.002	Cohens d	-1.2499
	Welch's t	-3.5006	20.6278	0.001	Cohens d	-1.3170

Anmerkung.  $H_0: \mu_{\text{hypnose}} = \mu_{\text{placebo}}$

\* 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 6.14.

Dies produziert das Analyseergebnis in Abbildung 6.15.

Damit können die Teilfragen beantwortet werden.

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

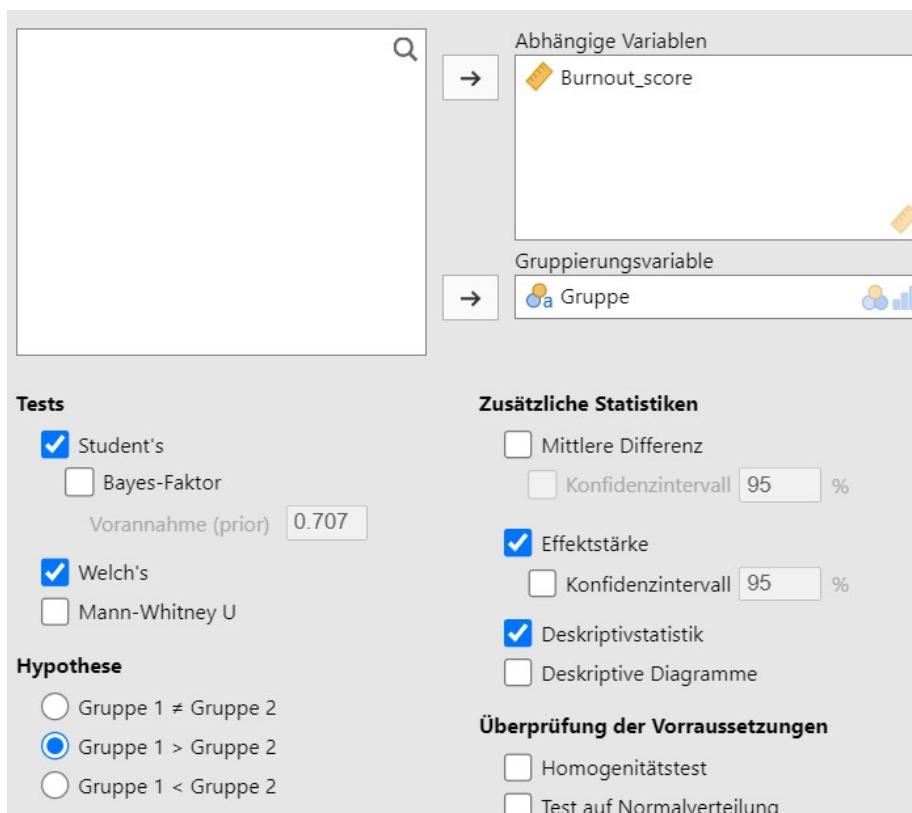


Abbildung 6.14: Jamovi Eingabe.

## 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.  $H_0: \mu_{\text{Unverheiratet}} \geq \mu_{\text{Verheiratet}}$

## 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 durchschnittlich höheres Burnout-Risiko haben als verheiratete und zwar nicht nur in der Stichprobe sondern auch in der Population. Das durchschnittliche Burnout-Risiko in der Population ist der Erwartungswert des Burnout-Risiko und wird mit  $\mu$  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 Robinson et al. (2021). 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 7.1. Die so sortierten Beobachtungen werden nummeriert, was den sogenannten Rangnummern entspricht. Für gleiche Ränge wird den betreffenden Rängen ein mittlerer Rang zugewiesen.

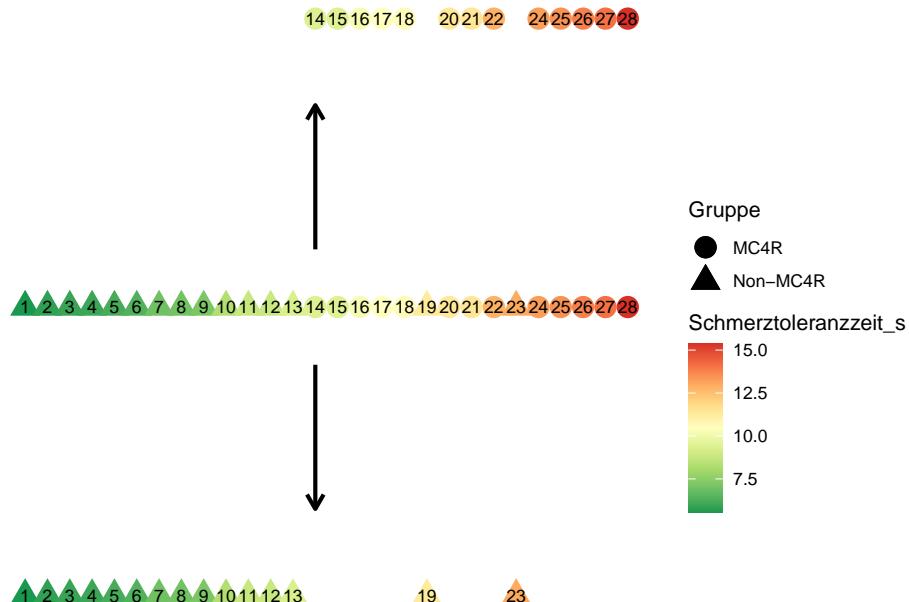


Abbildung 7.1: Mitte: Aufsteigend sortierte Beobachtungen des Merkmals Schmerztoleranz für MC4R und Non-MC4R Mäuse. Die Zahlen stehen für die Rangnummern. Oben (MC4R) und unten (Non-MC4R) stellt die Gruppenaufteilung der Rangnummern dar.

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  $1 + 2 + \dots + 15 = 120$  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 Referenztafel 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 sich die durchschnittliche Aufgeschlossenheit von jüngeren Menschen unter 30 Jahren und älteren Menschen mit über 30 oder genau 30 Jahren unterscheidet. 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  $N = 'rexm_aufgeschlossenheit_jung_a.ltn2'$  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 7.2 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

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

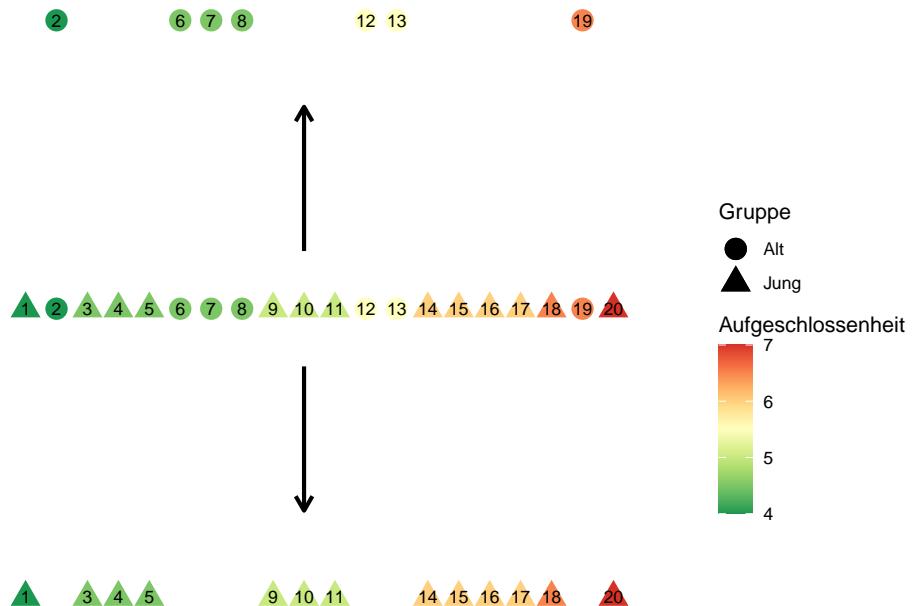


Abbildung 7.2: Mitte: Aufsteigend sortierte Beobachtungen des Merkmals Schmerztoleranz für Junge und Alte. Die Zahlen stehen für die Rangnummern. Oben (Jung) und unten (Alt) stellt die Gruppenaufteilung der Rangnummern dar.

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

## 126 KAPITEL 7. GRUPPENMITTELWERTUNTERSCHIED BEI EINEM MINDESTENS ORDINALSKA

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. Cohen (1988) 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 Rankkorrelation unter **Zusätzliche Statistiken** > **Effektstärke** eingeblendet werden. Für die beiden Beispiele 7.1 und 7.2 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 7.3.

Dies produziert das Analyseergebnis in Abbildung 7.4.

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.

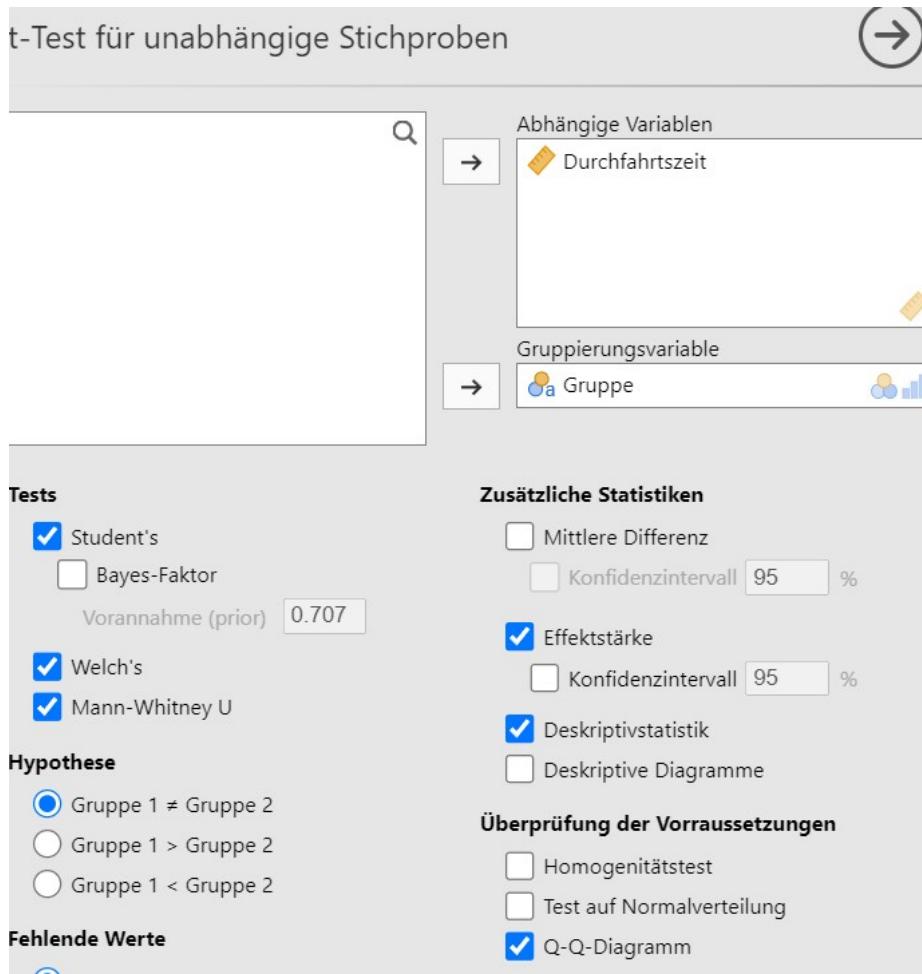


Abbildung 7.3: Jamovi Eingabe.

## t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Durchfahrtszeit	Student's t	2.251 *	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	Biserialle Rangkorrelation	-0.227

Anmerkung.  $H_0: \mu_{\text{Tempo } 50} = \mu_{\text{Tempo } 30}$

\* 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

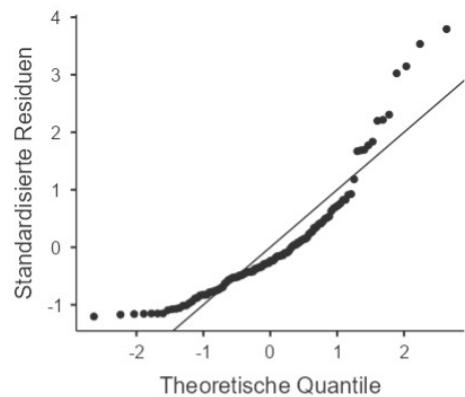


Abbildung 7.4: Jamovi Ausgabe.

- 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 ( $\text{Mdn} = 3.05\text{Min}, N = 54$ ) und Tempo 50 ( $\text{Mdn} = 4.27\text{Min}, N = 65$ ) 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 7.5.

Dies produziert das Analyseergebnis in Abbildung 7.6.

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:

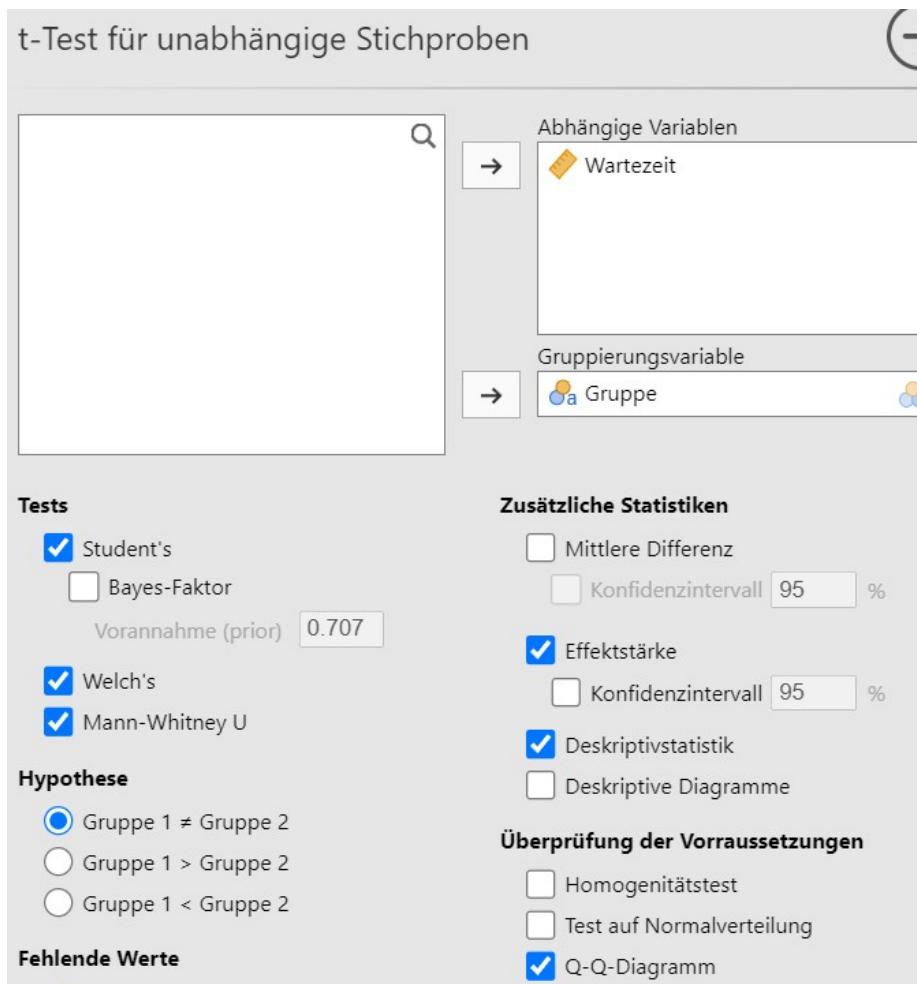


Abbildung 7.5: Jamovi Eingabe.

## t-Test für unabhängige Stichproben

		Statistik	df	p		Effektstärke
Wartezeit	Student's t	-1.844*	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	Biserialle Rangkorrelation	0.141

Anmerkung.  $H_0: \mu_{\text{Konzept 1}} = \mu_{\text{Konzept 2}}$

\* 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

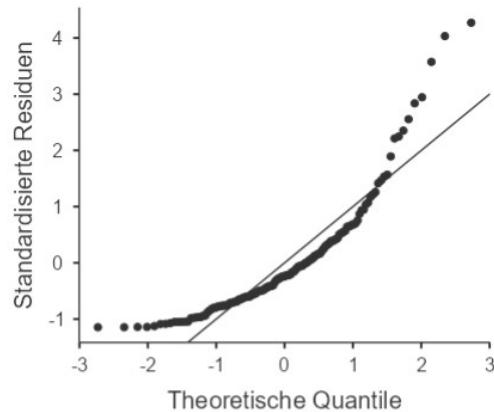


Abbildung 7.6: Jamovi Ausgabe.

Ein zweiseitiger  $U$ -Test nach Mann und Whitney ergibt, dass sich die mediane Wartezeit am neuen Terminal bei Konzept 1 ( $Mdn = 3.41\text{Min}$ ,  $N = 51$ ) und Konzept 2 ( $Mdn = 4.18\text{Min}$ ,  $N = 107$ ) 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?

- Stellen Sie die Testhypotesen auf.
- Weshalb wird hier ein  $U$ -Test gegenüber einem Welch Test bevorzugt?
- 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 7.7.

Dies produziert das Analyseergebnis in Abbildung 7.8.

Damit können die Teilfragen beantwortet werden.

- 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  $H_0 : \mu_{\text{Statistik 1}} \geq \mu_{\text{Statistik 2}}$
- Hier wird ein  $U$ -Test verwendet, da das Merkmal auf einem einzigen Likertskalierterem Merkmal beruht. Dieses ist demnach ordinalskaliert. Ein Welch Test eignet sich nur für intervallskalierte Merkmale.
- 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

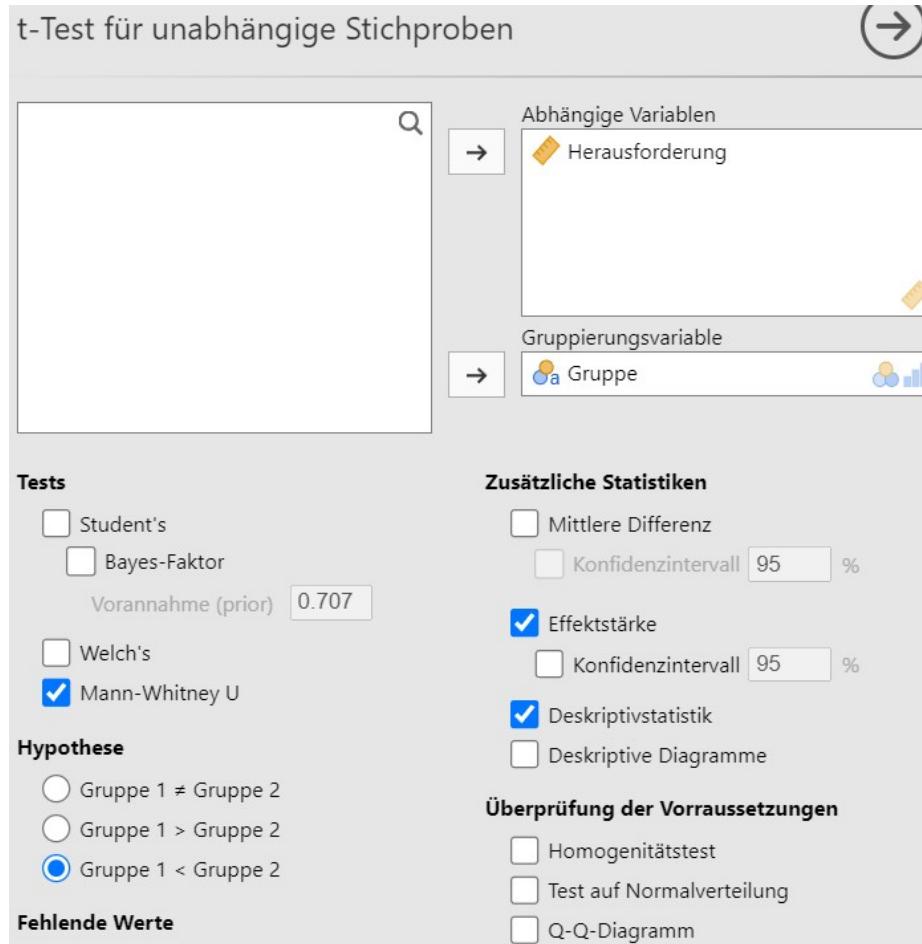


Abbildung 7.7: Jamovi Eingabe.

t-Test für unabhängige Stichproben

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

Anmerkung.  $H_0: \mu_{\text{Statistik1}} < \mu_{\text{Statistik2}}$

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 7.9.

Dies produziert das Analyseergebnis in Abbildung 7.10.

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.

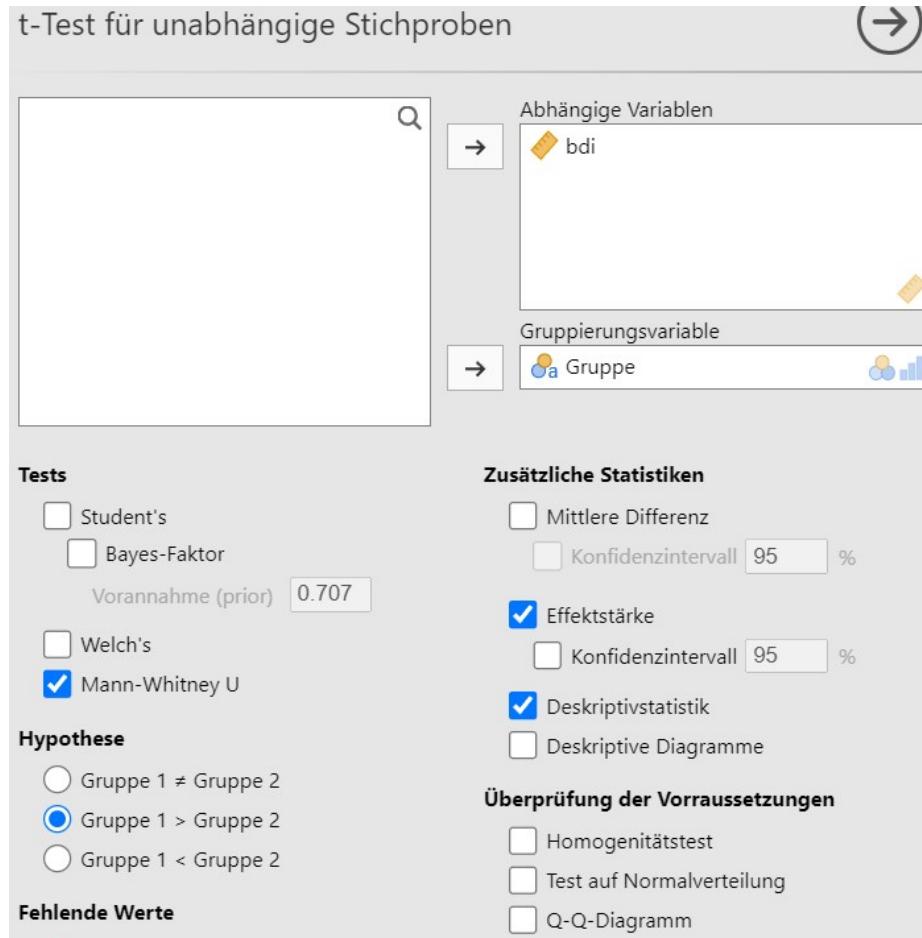


Abbildung 7.9: Jamovi Eingabe.

## t-Test für unabhängige Stichproben

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

Anmerkung.  $H_0: \mu_{\text{Ecstasy}} > \mu_{\text{Alkohol}}$

## 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

## Korrelation

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 8.1 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 weniger klar ist die Beobachtung, dass die Zahlungsbereitschaft mit dem Jahreseinkommen ansteigt.

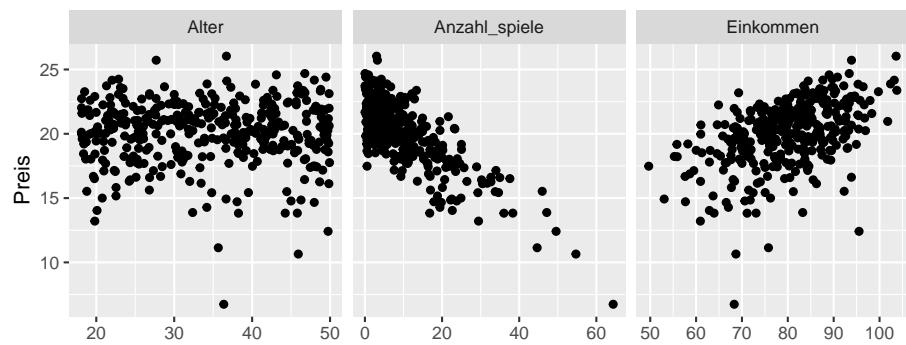


Abbildung 8.1: Die drei Streudiagramme zeigen jeweils die Zahlungsbereitschaft (Preis in CHF) auf der y-Achse und das Alter in Jahren, die Anzahl Spiele im Haushalt und das Jahreseinkommen in 1'000 CHF respektive auf der x-Achse.

Diese drei Beobachtungen können mit der sogenannten **Korrelation** digitalisiert werden. Die Korrelation ist eine Zahl zwischen  $-1$  und  $+1$ . Negative Zahlen bedeuten, dass es einen gegenläufigen Zusammenhang gibt. **Gegenläufig** bedeutet je mehr von Merkmal A, desto weniger von Merkmal B. Positive Zahlen bedeuten, dass es einen gleichläufigen Zusammenhang gibt. **Gleichläufig** bedeutet je mehr von Merkmal A, desto mehr von Merkmal B. Eine Korrelation von 0 schliesslich bedeutet, dass es keinen Zusammenhang zwischen den beiden Merkmalen gibt. Je weiter weg von 0 die Korrelation liegt, desto eindeutiger ist ein Zusammenhang auf dem Streudiagramm erkennbar. Die Korrelationen für die in der Abbildung gezeigten Zusammenhänge liegen bei  $-0.002$  für Alter und Zahlungsbereitschaft,  $-0.770$  für die Anzahl Spiele im Haushalt und Zahlungsbereitschaft und  $0.484$  für das Einkommen und die Zahlungsbereitschaft.

Die Korrelation erfasst nur lineare Zusammenhänge, also nur wenn Punkte entlang einer Linie streuen im Streudiagramm. Je weiter weg die Korrelation von 0 ist, desto weniger streuen die Punkte entlang einer Linie, siehe erste und zweite Zeile der Abbildung 8.2. Die Korrelation sagt lediglich, ob die Linie steigt oder fällt, jedoch nicht wie stark. Nicht lineare Zusammenhänge wie in der dritten Zeile von Abbildung 8.2 können mit der Korrelation nicht richtig erfasst werden und gängige Korrelationsberechnungen geben zufällige Resultate.

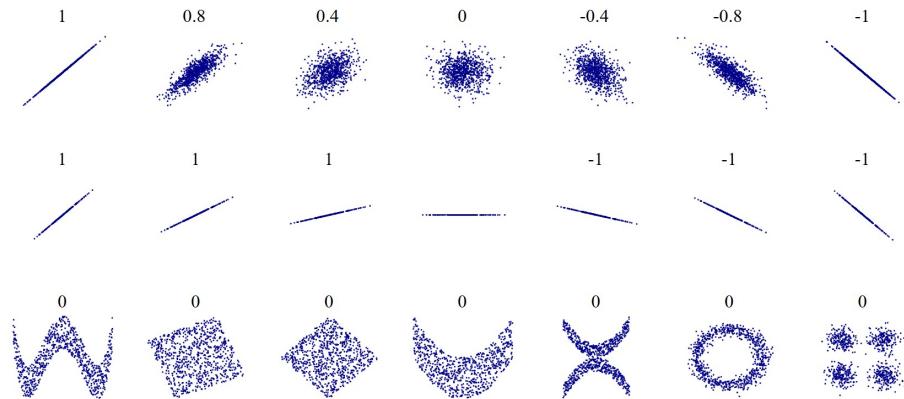


Abbildung 8.2: Streudiagramme und dazugehörige Korrelation.

Liegt eine Korrelation vor bedeutet dies, dass die Merkmale sich gemeinsam verändern. In Beispiel 8.1 scheint intuitiv klar, dass das Einkommen und die Anzahl vorhandene Spiele pro Haushalt die Zahlungsbereitschaft beeinflussen. Das Einkommen und die Anzahl Spiele sind also ursächlich. Dass die Zahlungsbereitschaft ursächlich wäre und zum Beispiel das Einkommen beeinflusst ist eher unwahrscheinlich. Welches von zwei Merkmalen ursächlich ist, bzw. wie die Merkmale kausal zusammenhängen, lässt sich jedoch nicht immer einfach beantworten, wie das folgende Beispiel zeigt.

**Beispiel 8.2** (Depression, Angststörungen und Alkohol). In der Psychotherapie

ist aufgefallen, dass sich Alkoholabhängigkeit, Angststörungen und Depression in der Tendenz wechselwirkend positiv beeinflussen (Schuckit, 1996). Dies soll mit den fiktiven Daten in `08-exm-depression-alkohol-angst.sav` illustriert werden. Für die Messung der Schwere der drei Merkmale Alkoholismus, Angststörung und Depression wurden die folgenden Messinstrumente verwendet: Das Beck Depression Inventory *BDI* für die Depression (Beck et al., 1988), das state trait anxiety inventory *STAI* für die Angststörungen (Spielberger et al., 1983) und das alcohol use inventory für den Alkoholismus (Skinner and Allen, 1982).

Streudiagramme aller möglicher bivariaten Zusammenhänge sind in Abbildung 8.3 dargestellt. Während der lineare Zusammenhang zwischen Alkoholismus und Angststörung kaum erkennbar ist, so kann zwischen Depression und Alkoholismus ein leichter und zwischen Depression und Angststörung ein deutlich gleichläufiger linearer Zusammenhang festgestellt werden. Die geschätzten Korrelationen sind 0.32 zwischen Alkoholismus und Depression, 0.1 zwischen Alkoholismus und Angststörung und 0.46 zwischen Depression und Angststörung.

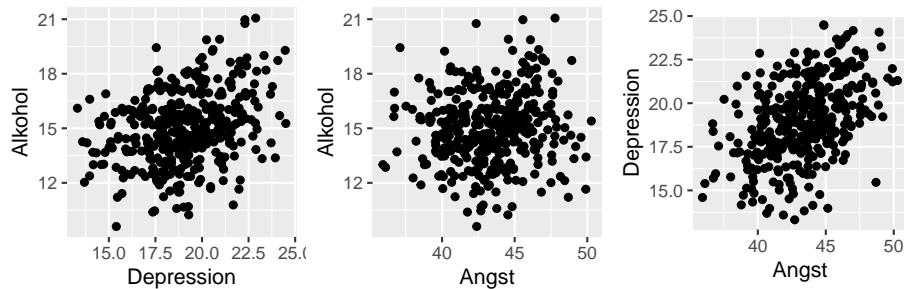


Abbildung 8.3: Wiederholte Stichprobenziehung bei gleichbleibender Population mit eher hohen Angst-Werten.

In diesem Fall ist unklar, ob jemand mit Alkoholismus eher eine Angststörung entwickelt, oder jemand mit Angststörung eher eine Alkoholabhängigkeit entwickelt. Die Korrelation gibt nur einen Anhaltspunkt über die Art des Zusammenhangs, nicht aber über die Ursächlichkeit der Variablen. Es könnte auch sein, dass die beiden Merkmale eigentlich unabhängig voneinander sind, aber

## 8.1. WIE STARK IST DER ZUSAMMENHANG ZWISCHEN ZWEI INTERVALSKALIERTEN UND NORMALVERTEILTEN VARIABLEN

ein drittes Merkmal die beiden Merkmale beeinflusst. Dazu später mehr unter 8.5.

Für die genaue Berechnung der Korrelation gibt es verschiedene Möglichkeiten, wovon auf einige im Verlauf dieses Kapitels eingegangen wird.

### 8.1 Wie stark ist der Zusammenhang zwischen zwei intervallskalierten und normalverteilten Variablen?

Wenn die beiden Merkmale je intervallskaliert sind und deren Beobachtungen  $x_1, \dots, x_n$  und  $y_1, \dots, y_n$  einer normalverteilten Population entstammen, dann ist die Korrelation mit dem **Korrelationskoeffizient nach Pearson**, auch **Produkt-Moment-Korrelation** genannt,

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

geschätzt. Im Zähler steht dabei die sogenannte Kovarianz. Diese misst mit Wert der Summanden, wie stark ein Punkt vom Durchschnitt abweicht und mit dem Vorzeichen der Summanden, ob beide Merkmale in dieselbe Richtung vom Durchschnitt abweichen oder nicht. Der Wert der Kovarianz hängt von der Einheit der Merkmale ab. Um dies zu vermeiden wird die Korrelation durch das Produkt der Standardabweichungen der beiden Merkmale geteilt. So wird  $-1 < r < 1$  erreicht.

Die oben in den Beispielen angegebenen Korrelationen wurden mit dem Korrelationskoeffizient nach Pearson basierende auf einer Zufallsstichprobe geschätzt. Der so berechnete Wert gilt deshalb für die Stichprobe und ist die beste Schätzung für die Korrelation in der Population. Aufgrund der Überlegungen aus Kapitel 3 ist diese Schätzung jedoch mit Unsicherheit behaftet. Diese Unsicherheit kann mit einem Konfidenzintervall abgeschätzt werden. Die genaue Berechnung davon wird hier nicht erläutert.

Wichtig für die Beurteilung der Korrelation ist die Frage, ob eine Korrelation besteht oder nicht. Die Korrelation in der Population wird mit  $\rho$ , sprich 'rho' bezeichnet. Es wird also gefragt, ob  $H_0 : \rho = 0$ ,  $H_1 : \rho > 0$  oder  $H_1 : \rho < 0$  ist. Diese Fragestellungen können mit einem zweiseitigen respektive einseitigen statistischen Test getestet werden. Da getestet wird, ob die Korrelation sich von 0 unterscheidet wird dieser Test **Absicherung gegen Null** genannt.

Die Teststatistik für die Absicherung gegen Null der Korrelation nach Pearson ist

$$t = r \cdot \sqrt{\frac{n - 2}{1 - r^2}}. \quad (8.1)$$

Wenn wiederholt Stichproben gezogen werden und immer wieder die Teststatistik berechnet wird, kann beobachtet werden, dass die Teststatistik  $t$ -verteilt bei  $n - 2$  Freiheitsgraden ist.

In **Jamovi** wird die Korrelation nach Pearson unter **Analysen > Regression > Korrelationsmatrix** und **Korrelationskoeffizienten: Pearson** geschätzt und mit der Zusatzoption **Zusätzliche Optionen: Signifikanzniveau** gegen Null abgesichert. Angenommen, es gäbe kein Vorwissen über die Richtung des Zusammenhangs zwischen den Merkmalen Alkoholismus, Depression und Angst. In diesem Fall werden zweiseitige Tests berechnet und wie folgt berichtet:

Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Alkoholismus und Depression ( $r = 0.32$ ) signifikant von 0 unterscheidet,  $t(373) = 6.41$ ,  $p < 0.001$ .

Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Alkoholismus und Angststörung ( $r = 0.1$ ) nicht signifikant von 0 unterscheidet,  $t(373) = 1.96$ ,  $p = .051$ .

Ist eine klare Hypothese über die Richtung des linearen Zusammenhangs vorhanden, zum Beispiel Leute mit mehr Geld kaufen mehr Geld für Spielzeuge ausgeben wollen, so kann auch einseitig getestet werden:  $H_0 : \rho \leq 0$  und  $H_1 : \rho > 0$  und

Die einseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass die Korrelation zwischen Kaufbereitschaft und Einkommen ( $r = 0.48$ ) signifikant grösser ist als 0,  $t(373) = 10.7$ ,  $p < 0.001$ .

Da die Teststatistik nicht von **Jamovi** direkt ausgegeben wird, muss diese händisch berechnet werden.

## 8.2 Wie stark ist der Zusammenhang zwischen zwei mindestens ordinalskalierten Merkmalen?

Manchmal ist ein Merkmal oder beide Merkmale nicht intervallskaliert und normalverteilt. Sind die Daten beider Merkmale dennoch mindestens ordinalskaliert, so kann die **Korrelation nach Spearman**, auch **Spearmans Korrelationskoeffizient** genannt, verwendet werden.

## 8.2. WIE STARK IST DER ZUSAMMENHANG ZWISCHEN ZWEI MINDESTENS ORDINALSKALIERTEN MERKMALEN

**Beispiel 8.3** (Facebook und Lebenszufriedenheit.). Je intensiver Facebook konsumiert wird, desto tiefer ist die Lebenszufriedenheit (Błachnio et al., 2016). Eine Studie will dieses Resultat reproduzieren. Die Facebook-Nutzungsintensität wurde dafür mit einer Skala von 1 (keine) bis 7 (sehr intensive Nutzung) und die Lebenszufriedenheit mit Punkten von 1 (sehr unzufrieden) bis 10 (sehr zufrieden) gemessen. Es wurden 73 Personen befragt. Ihre Antworten wurden bereinigt im Datensatz `08-exm-fb-life.sav` abgelegt. Nach der Erhebung wird festgestellt, dass die beiden Merkmale nicht normalverteilt sind.

Wenn die Daten beider Merkmale mindestens ordinalskaliert sind, diese aber nicht die Voraussetzungen für die Korrelation nach Pearson erfüllen, kann die Korrelation nach Spearman angewendet werden. Diese wird berechnet indem den Beobachtungen jedes Merkmals aufsteigende sortiert und entsprechend der Reihenfolge Rangplätze vergeben werden. Die Differenz der Ränge der beiden Merkmale für Beobachtung  $i$  wird mit  $d_i$  bezeichnet. Die Korrelation nach Spearman ist

$$r = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}.$$

### Achtung



*Hinweis.* Exploration der Definition der Spearman'schen Korrelation

1. Alle Punkte sind entlang einer aufsteigenden Linie: In diesem Fall entsprechen sich die Ränge der beiden Merkmale genau und  $d_i = 0$  für alle Beobachtungen. Dies führt zu  $d_i^2 = 0$  und demnach zu  $r = 1$ .
2.  $n = 5$  Punkte sind entlang einer absteigenden Linie: In diesem Fall entspricht der kleinste Rang des einen Merkmals dem grössten des anderen Merkmals, der zweitgrösste Rang dem zweitkleinsten, usw. Die Rangdifferenzen sind in dem Fall  $d_1 = 5 - 1 = 4, d_2 = 4 - 2 = 2, d_3 = 3 - 3 = 0, d_4 = 2 - 4 = -2, d_5 = 1 - 5 = -4$ . Dies führt zu

$$r = 1 - \frac{6 \cdot (4^2 + 2^2 + 0^2 + (-2)^2 + (-4)^2)}{5 \cdot (5^2 - 1)} = 1 - \frac{240}{120} = 1 - 2 = -1.$$

Die Formel ist also so ausgestaltet, dass bei perfektem gegenläufigem Zusammenhang eine Korrelation von  $-1$  erreicht wird.

Die Absicherung gegen Null der Korrelation nach Spearman erfolgt gleich wie bei der Korrelation nach Pearson. Die Teststatistik wird also auch mit Gleichung

(8.1) berechnet und ist ebenfalls  $t$ -verteilt bei  $n - 2$  Freiheitsgraden, wobei  $n$  für die Anzahl Beobachtungspaare steht.

In Jamovi wird die Korrelation nach Spearman unter **Analysen > Regression > Korrelationsmatrix** und **Korrelationskoeffizienten: Spearman** geschätzt und mit der Zusatzoption **Zusätzliche Optionen: Signifikanzniveau** gegen Null abgesichert. Angenommen, es gäbe kein Vorwissen über die Richtung des Zusammenhangs zwischen den Merkmalen Facebook-Nutzungsintensität und Lebenszufriedenheit. In diesem Fall werden zweiseitige Tests berechnet und wie folgt berichtet:

Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Spearman ergibt, dass sich die Korrelation zwischen Facebook-Nutzungsintensität und Lebenszufriedenheit ( $r = -0.29$ ) signifikant von 0 unterscheidet,  $t(71) = -2.58$ ,  $p = .007$ .

### 8.3 Wie stark ist der Zusammenhang zwischen einem intervallskalierten und normalverteilten Merkmal und einem dichotomen Merkmal?

Ein mit dem BDI gemessener Wert für Depression kann kategorisiert werden in leichte Depression (1 – 19) Punkte und schwere Depression (20 – 63) Punkte. Es folgt also eine neues dichotomes Merkmal Depressionsdiagnose mit Ausprägungen leichte und schwere Depression. Hier kann ebenfalls die Frage gestellt werden, wie stark der Zusammenhang zwischen Depressionsdiagnose und dem intervallskalierten Wert für den Alkoholismus. Die **punktbiserialen Korrelation** ist dazu geeignet den Zusammenhang zwischen einer dichotomen und einem intervallskalierten normalverteilten Merkmal zu messen. Sie wird berechnet mit

$$r = \frac{\bar{x}_1 - \bar{x}_2}{(n_1 + n_2) \cdot s} \sqrt{n_1 \cdot n_2},$$

wobei  $s$  die Standardabweichung des intervallskalierten Merkmals bezeichnet.

Die Berechnung der punktbiserialen Korrelation ist in Jamovi nicht implementiert. Die Kenngrößen für die Berechnung können jedoch aus der **Erforschung > Deskriptivstatistik** von Jamovi abgelesen werden. Dabei ist zu beachten, dass für  $\bar{x}_1, n_1, \bar{x}_2, n_2$  die Option **Aufgeteilt nach** mit dem dichotomen Merkmal befüllt sein muss. Für die Berechnung von  $s$  muss dieses Feld jedoch leer sein. Im Beispiel ist die punktbiserialen Korrelation

$$r = \frac{\bar{x}_1 - \bar{x}_2}{(n_1 + n_2) \cdot s} \sqrt{n_1 \cdot n_2} = \frac{15.88 - 14.59}{(122 + 253) \cdot 1.97} \sqrt{122 \cdot 253} = 0.31.$$

#### 8.4. WIE STARK IST DER ZUSAMMENHANG ZWISCHEN EINEM MINDESTENS ORDINALSKALIERTEN MERKMALE

Da ein dichotomes Merkmal wie die Depressionsdiagnostik nur zwei Ausprägungen hat, funktioniert die übliche Interpretation dieses gleichläufigen Zusammenhangs mit “je mehr schwere Depression, desto höher der Alkoholismus-Wert” nicht mehr. Stattdessen kann die leicht angepasste Interpretation “je eher eine Person eine schwere Depression hat, desto höher der Alkoholismus-Wert” oder “je höher der Alkoholismus-Wert, desto eher hat eine Person eine schwere Depression” verwendet werden.

Eine genaue Betrachtung dieser Formel lässt eine Ähnlichkeit zur Effektstärke des Zweistichproben-*t*-Tests und Welch Tests erkennen. Tatsächlich kommt die punktbiseriale Korrelation in den gleichen Fällen zur Anwendung wie besagte Tests und ist äquivalent zu deren Effektstärken.

Die punktbiseriale Korrelation wird ähnlich wie die Korrelation nach Pearson abgesichert. Die Teststatistik ist

$$t = r \cdot \sqrt{\frac{n_1 + n_2 - 2}{1 - r^2}}. \quad (8.2)$$

Sie folgt einer *t*-Verteilung bei  $n_1 + n_2 - 2$  Freiheitsgraden. Dieser Wert kann mit der entsprechenden *t*-Verteilung in Abbildung 4.7 verglichen werden. Da die Absicherung nicht in Jamovi implementiert ist, wird die Absicherung gegen null der punktbiserialen Korrelation an der Prüfung nicht erwartet. Um die Zufälligkeit des gefundenen Zusammenhangs zu beurteilen, kann behelfsmässig auf das Resultat des Zweistichproben-*t*-Tests resp. Welch Test abgestellt werden.

### 8.4 Wie stark ist der Zusammenhang zwischen einem mindestens ordinalskalierten Merkmal und einem dichotomen Merkmal?

Genau wie die vorherige Situation äquivalent zum Zweistichproben-*t*-Test und Welch Test war, ist die Situation mit einem mindestens ordinalskaliertem Merkmal und einem dichotomen Merkmal äquivalent zur Effektstärke des *U*-Tests nach Mann und Whitney. Für die Effektstärke des *U*-Test wurde tatsächlich bereits die biseriale Rangkorrelation definiert und in Abschnitt 7.2 hinlänglich beschrieben. Für das Beispiel 8.2 mit Depressionsdiagnose kann die biseriale Rangkorrelation herausgefunden werden, indem ein *U*-Test und dessen Effektstärke berechnet wird, was  $r = 0.357$  ergibt.

Da auch die Absicherung gegen Null der biserialen Rangkorrelation nicht in Jamovi implementiert ist, wird dies an der Prüfung nicht erwartet. Um die Zufälligkeit des gefundenen Zusammenhangs zu beurteilen, kann behelfsmässig auf das Resultat des *U*-Tests abgestellt werden.

## 8.5 Was ist eine Störfaktor und wie wird damit umgegangen?

Im Beispiel 8.2 wurde festgestellt, dass es einen Zusammenhang schwachen Zusammenhang zwischen Alkoholismus und Angststörung gibt. Je stärker der Alkoholismus ausgeprägt ist, desto höher ist auch der Wert für die Angststörung. Dies ist im oberen Teil der Abbildung 8.5 abstrahiert dargestellt.

Es wurde jedoch auch festgestellt, dass die Merkmale Angst und Alkoholismus jeweils mit dem Merkmal Depression korrelieren. Es könnte also sein, dass der Zusammenhang zwischen Alkoholismus und Angststörung lediglich darauf beruht, dass Menschen mit schwerer Depression zu höherem Alkoholismus und ebenfalls zu mehr Angststörungen neigen. Die Angststörung und der Alkoholismus könnten schwächer korreliert oder gar unabhängig voneinander sein. Wenn ein drittes Merkmal den Zusammenhang zweier Merkmale auf irgendeine Weise verfälsch, wird dieses Merkmal **Störfaktor** oder **Störvariable** genannt.

Um den Einfluss einer Störvariable  $X_3$  auf den Zusammenhang zweier Merkmale  $X_1$  und  $X_2$  zu beurteilen, kann die **partielle Korrelation**

$$r_{12.3} = \frac{r_{12} - r_{23} \cdot r_{13}}{\sqrt{(1 - r_{23}^2) \cdot (1 - r_{13}^2)}}$$

berechnet werden. Die  $r$  stehen dabei für die Korrelation nach Pearson zwischen den zwei indexierten Merkmalen.

### Achtung



*Hinweis.* Exploration der Formel der partiellen Korrelation:

1. Das dritte Merkmal korreliert nicht mit den ersten beiden  $r_{23} = r_{13} = 0$ . In diesem Fall ist  $r_{12.3} = r_{12}$ .
2. Das dritte Merkmal korreliert sehr schwach mit den ersten beiden  $r_{23} = r_{13} = 0.01$ . In diesem Fall ist  $r_{12.3} \approx r_{12}$ .

Die partielle Korrelation kann in **Jamovi** unter **Analysen > Regression > Partielle Korrelation** berechnet werden und beträgt  $r = -0.05$ . Unter Berücksichtigung der Störvariable Depression, ist der Zusammenhang zwischen Alkoholismus und Angststörung also nicht nur kleiner geworden, sondern hat sich sogar in einen gegenläufigen Zusammenhang gewandelt. Dies ist abstrahiert im unteren Teil der Abbildung 8.5 dargestellt. Wird die partielle Korrelation unter Berücksichtigung einer Störvariable angegeben, spricht man auch von es wurde für die Störvariable **kontrolliert**.

Zusammenhang zwischen Angst und Alkoholismus ohne (oben) und mit (unten) Berücksichtigung der Störvariable.

Zusammenhang zwischen Angst und Alkoholismus ohne (oben) und mit (unten) Berücksichtigung der Störvariable.

Die Absicherung gegen Null der partiellen Korrelation erfolgt ähnlich wie bei der Korrelation nach Pearson. Die Teststatistik wird mit

$$t = r \cdot \sqrt{\frac{n - 3}{1 - r^2}}. \quad (8.3)$$

berechnet und ist  $t$ -verteilt bei  $n - 3$  Freiheitsgraden, wobei  $n$  für die Anzahl Beobachtungstripel steht. Die Teststatistik wird von **Jamovi** nicht berechnet und muss händisch eruiert werden.

Die zweiseitige Absicherung gegen Null der partiellen Korrelation ergibt, dass sich die Korrelation zwischen Alkoholismus und Angststörung unter Berücksichtigung der Depression als Störfaktor ( $r = -0.05$ ) nicht signifikant von 0 unterscheidet,  $t(372) = -1$ ,  $p = .316$ . Die Korrelation ist als schwach einzustufen.

In diesem Beispiel wurde die Angst arbiträr als Störfaktor gehandelt. Die selben Überlegungen und Berechnungen wären aber auch zum Beispiel für den Alkoholismus als Störfaktor für den Zusammenhang zwischen Angst und Depression zulässig. Die Identifikation von Störvariablen muss also theoriegeleitet erfolgen.

## 8.6 Übungen

### Übung 8.1.

Sichern Sie die Korrelation nach Pearson zwischen Angststörung und Depression in Beispiel 8.2 mit dem Datensatz **08-exm-depression-alkohol-angst.sav** zweiseitig gegen Null ab, berichten Sie das Ergebnis und interpretieren Sie die Stärke des Zusammenhangs.

*Lösung.*

Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Depression und Angststörung ( $r = 0.46$ ) sich signifikant von 0 unterscheidet,  $t(373) = 9.98$ ,  $p < 0.001$ . Die Korrelation deutet auf einen starken Zusammenhang der Art je mehr Depression desto grösser die Angststörung hin.

### Übung 8.2.

Testen Sie mit dem Datensatz **08-exm-zahlungsbereitschaft.sav** aus Beispiel 8.1, ob die Korrelation nach Pearson zwischen der Zahlungsbereitschaft

**Preis** und Alter positiv und ob die Korrelation zwischen der Zahlungsbereitschaft und der Anzahl Spiele im Haushalt negativ ist. Berichten Sie das Ergebnis und interpretieren Sie die Stärke des Zusammenhangs.

*Lösung.*

Die einseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass die Korrelation zwischen Kaufbereitschaft und Alter ( $r = 0$ ) nicht signifikant grösser ist als 0,  $t(373) = -0.04$ ,  $p = .515$ . Der Zusammenhang ist so schwach, dass keine Richtung des Zusammenhangs aus dem Korrelationskoeffizienten abgelesen werden kann.

Die einseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass die Korrelation zwischen Kaufbereitschaft und Anzahl Spiele ( $r = -0.77$ ) signifikant kleiner ist als 0,  $t(373) = -23.33$ ,  $p < 0.001$ . Der Zusammenhang der Art je höher die Kaufbereitschaft, desto tiefer die Anzahl Spiele im Haushalt, ist stark.

### Übung 8.3.

Im Rahmen einer Studie wurde die Big-5-Persönlichkeitszüge von 500 psychologie Studierenden abgefragt (Dolan et al., 2009). Der Datensatz ist unter **Jamovi > Öffnen > Datenbibliothek > Big 5** (Dolan, Oort, Stoel & Wicherts, 2009) verfügbar. Die Merkmale Neuroticism, Extraversion, Openness, Agreeableness und Conscientiousness stehen für Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit respektive.

- Berechnen Sie paarweise die Korrelation nach Pearson zwischen allen fünf Merkmalen. Erklären und interpretieren Sie die stärkste negative, die stärkste positive Korrelation und die schwächste Korrelation.
- Zeichnen Sie die Korrelationsmatrix als Diagramm, inklusive Variablenliste und Statistiken und erklären Sie, was für Diagrammtypen Sie sehen.
- Testen Sie, ob die Korrelationen von Null abweichen und berichten Sie das Ergebnis für die stärkste negative, die stärkste positive Korrelation und die schwächste Korrelation.

*Lösung.* Zuerst wird der Datensatz mit **Jamovi** eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 8.4.

Dies produziert das Analyseergebnis in Abbildungen 8.5 und 8.6.

Damit können die Teilfragen beantwortet werden.

- Die Korrelationen können direkt aus Abbildung 8.5 entnommen werden. Die stärkste positive Korrelation beschreibt den Zusammenhang zwischen

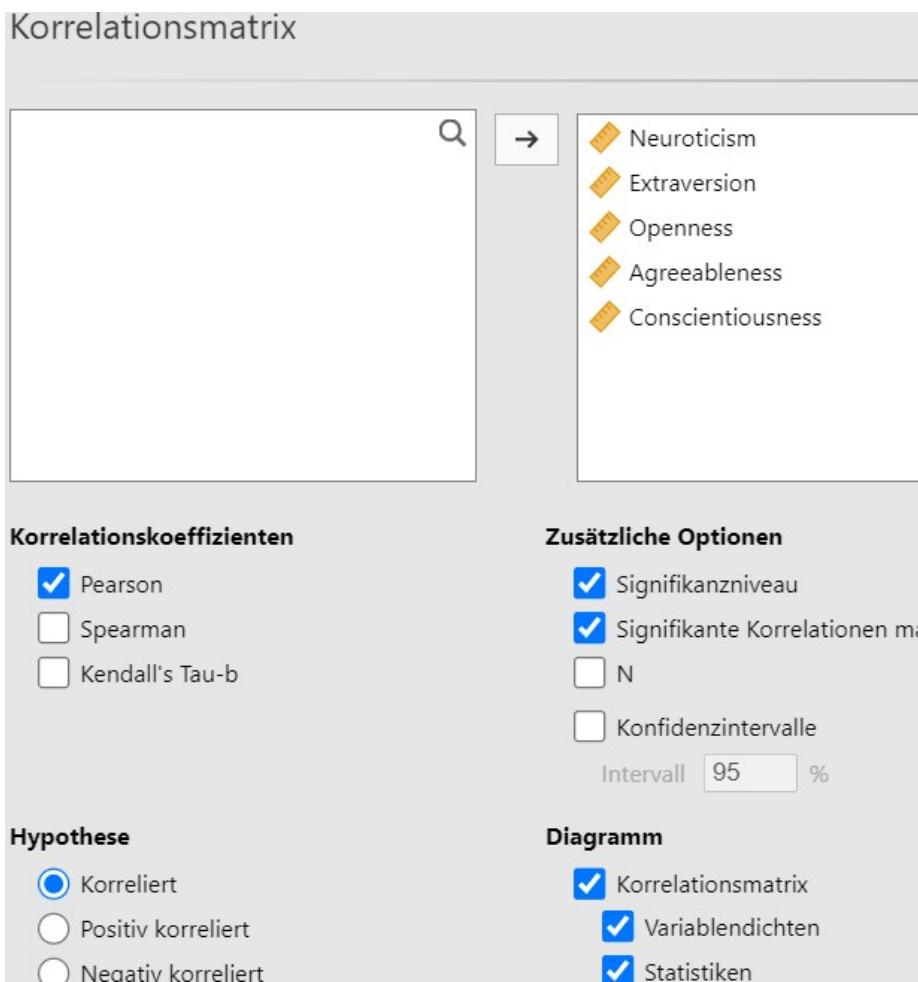


Abbildung 8.4: Jamovi Eingabe.

Korrelationsmatrix

		Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Neuroticism	Pearson's r	—				
	df	—				
	p-Wert	—				
Extraversion	Pearson's r	-0.35 ***	—			
	df	498	—			
	p-Wert	< .00001	—			
Openness	Pearson's r	-0.01	0.27 ***	—		
	df	498	498	—		
	p-Wert	0.81686	< .00001	—		
Agreeableness	Pearson's r	-0.13 **	0.05	0.16 ***	—	
	df	498	498	498	—	
	p-Wert	0.00262	0.22339	0.00035	—	
Conscientiousness	Pearson's r	-0.37 ***	0.06	-0.01	0.16 ***	—
	df	498	498	498	498	—
	p-Wert	< .00001	0.14924	0.76419	0.00037	—

Anmerkung. \* p < .05, \*\* p < .01, \*\*\* p < .001

Abbildung 8.5: Jamovi Ausgabe.

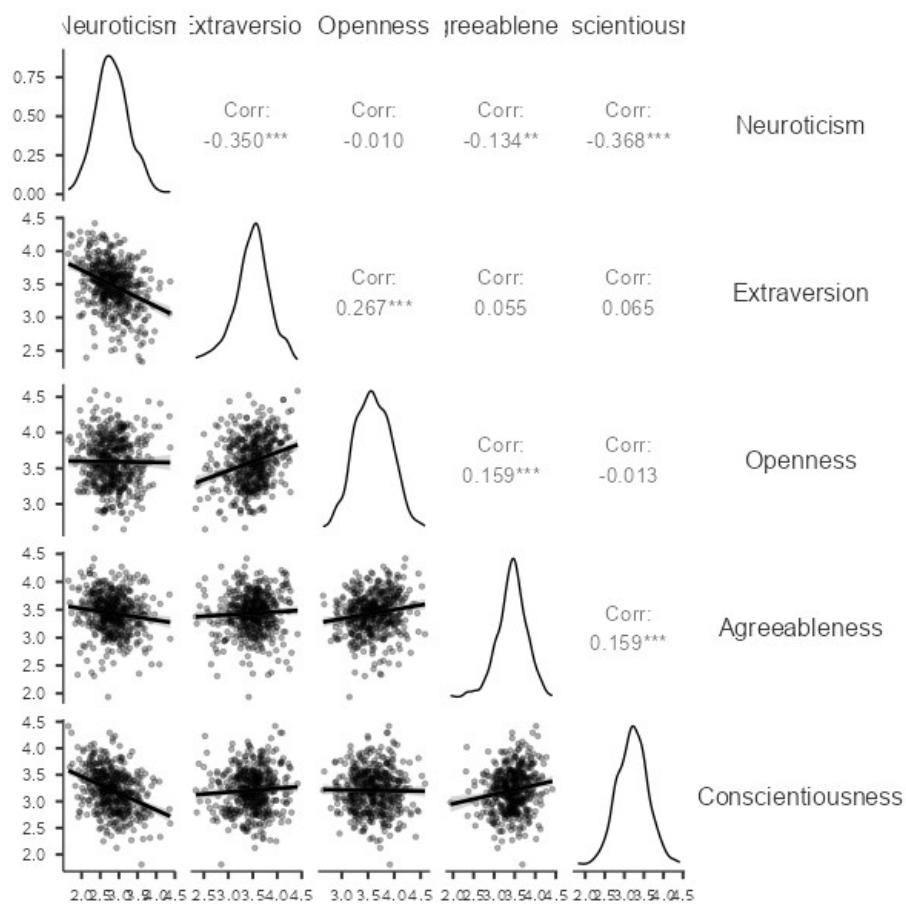


Abbildung 8.6: Jamovi Ausgabe.

Extraversion und Openness ( $r = 0.267$ ). Diese Korrelation ist als mittel einzustufen und bedeutet je extravertierter eine Person ist desto höhere Werte hat sie tendenziell auch bei der Offenheit. Die stärkste negative Korrelation beschreibt den Zusammenhang zwischen Neuroticism und Conscientiousness ( $r = -0.368$ ). Dies ist ebenfalls eine mittlere Korrelation. Sie bedeutet je neurotischer jemand ist, desto tiefere Werte für Conscientiousness hat die Person. Die schwächste Korrelation beschreibt den Zusammenhang zwischen Neuroticism und Openness ( $r = -0.010$ ). Es handelt sich um einen schwachen Zusammenhang. Da die Korrelation quasi Null ist, bedeutet dies, dass Neuroticism und Openness unabhängig voneinander sind. Das Vorzeichen der Korrelation deutet darauf hin, dass je höhere Neuroticism-Werte jemand hat, desto tiefere Openness-Werte hat die Person tendenziell.

- b) Die Korrelationsmatrix als Diagramm in Abbildung 8.6 hat folgende Bestandteile: Auf der Diagonalen befinden sich die geschätzten Verteilungen der Merkmale. Je höher die Linie bei einem gewissen Wert, desto wahrscheinlicher ist eine Beobachtung an diesem Punkt. Im oberen Dreieck sind die paarweise Korrelationen abgetragen. Die Sterne stehen für den  $p$ -Wert der Absicherung gegen Null, wobei ein Stern einer Korrelation entspricht, welche bei Signifikanzniveau 5% signifikant anders ist als Null. Im unteren Dreieck sind paarweise Streudiagramme aufgeführt und eine sogenannte Regressionsgerade, welche die Punkte am besten linear abbildet. Hier können Korrelationen grafisch erkannt werden und Anomalien wie zum Beispiel Aussreißer oder nicht lineare Zusammenhänge erkannt werden.
- c) Die entsprechenden Zusammenhänge wurden bereits in a) identifiziert.  
 > Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Openness und Extraversion ( $r = 0.267$ ) signifikant von 0 unterscheidet,  $t(498) = 6.18$ ,  $p < 0.001$ .> Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Neuroticism und Conscientiousness ( $r = -0.368$ ) signifikant von 0 unterscheidet,  $t(498) = -8.83$ ,  $p < 0.001$ .> Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Neuroticism und Openness ( $r = -0.010$ ) nicht signifikant von 0 unterscheidet,  $t(498) = -0.22$ ,  $p = .817$ .

#### **Übung 8.4.**

Mit einer Studie soll der Zusammenhang zwischen Haarausfall und erfülltem Sexualleben bei männlichen Patienten eruiert werden (Tas et al., 2018). Die Schwere des Haarausfalls wird mit der Hamilton-Norwood-Schema HNS ganzzahlig von 1 bis 7 gemessen, wobei jede höhere Zahl für ein neues zunehmendes Haarausfallstadium steht (Hamilton, 1951). Die Dysfunktionalität des Sexuallebens wird mit der Arizona Sexual Experience Scale ASES von 5 bis 30 Punkten gemessen, wobei höhere Zahlen auf eine sexuelle Dysfunktion hindeuten (McGahuey et al., 2016). Bei diesem Instrument führen mehrere

Likert-skalierte Items zum Endpunkteergebnis. Es wurden Daten erhoben, welche unter `08-exr-haarausfall-sex.sav` verfügbar sind.

- a) Welcher Korrelationskoeffizient ist hier am ehesten angebracht?
- b) Berechnen Sie die Korrelation nach Spearman, sichern Sie diese zweiseitig gegen Null ab und interpretieren Sie die Stärke des Zusammenhangs.
- c) Angenommen alle Merkmale wären intervallskaliert und normalverteilt. Berechnen Sie die partielle Korrelation zwischen Haarausfall und der Dysfunktionalität des Sexlebens unter Berücksichtigung des Störfaktor Alter, sichern sie diese gegen Null ab und interpretieren Sie die Stärke des Zusammenhangs vor und nach dem Herausrechnen des Störfaktors.

*Lösung.*

- a) Die Messung mit der ASES ist intervallskaliert. Die Messung des Haarausfalls dagegen kann sowohl als ordinalskaliert oder intervallskaliert betrachtet werden. Einerseits entsprechen die Zahlen auf der HNS verschiedenen abgegrenzten und klassifizierten Stadien der Progression des Haarausfalls. Dies deutet auf eine ordinalskaliertes Merkmal hin. Andererseits braucht es hier keine grosse Vorstellungskraft, sich eine 3.23 vorzustellen. Das wäre wie Stadium 3 mit ein wenig zusätzlichem Haarausfall. Aufgrund dieser Überlegung wäre hier sowohl der Korrelationskoeffizient nach Pearson wie der nach Spearman angebracht. Für letzteren müsste noch die Normalverteilung überprüft werden, was hier nicht gemacht wird.
- b) Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 8.7.

Dies produziert das Analyseergebnis in Abbildung 8.8

Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Spearman ergibt, dass sich die Korrelation zwischen Haarausfall und Dysfunktion der Sexualität ( $r = 0.24$ ) signifikant von 0 unterscheidet,  $t(373) = 4.7, p < 0.001$ . Die Korrelation ist als mittel einzustufen.

- c) Zuerst wird der Datensatz mit `Jamovi` eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 8.9.

Dies produziert das Analyseergebnis in Abbildung 8.10

Die zweiseitige Absicherung gegen Null der partiellen Korrelation ergibt, dass sich die Korrelation zwischen Haarausfall und Dysfunktion der Sexualität unter Berücksichtigung des Alters als Störfaktor ( $r = 0.02$ ) nicht signifikant von 0 unterscheidet,  $t(372) = 0.35, p = .728$ . Die Korrelation ist als schwach einzustufen.

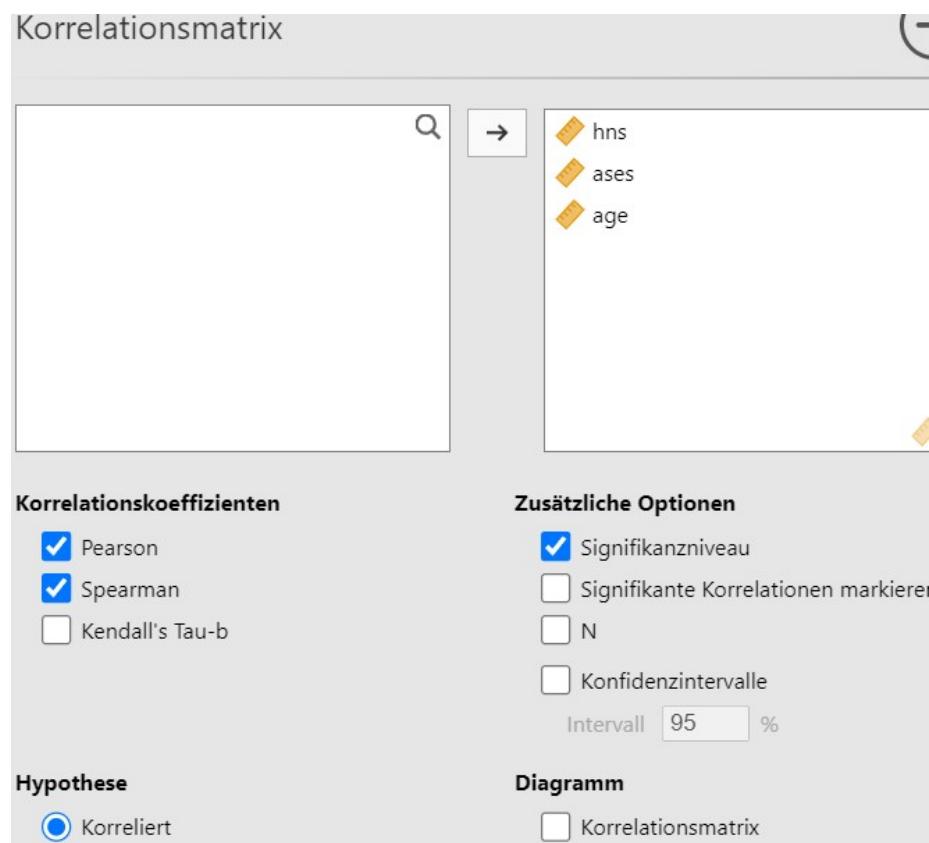


Abbildung 8.7: Jamovi Eingabe.

		hns	ases	age
hns	Pearson's r	—		
	df	—		
	p-Wert	—		
	Spearman's Rho	—		
ases	Pearson's r	0,244	—	
	df	373	—	
	p-Wert	< .00001	—	
	Spearman's Rho	0,236	—	
	df	373	—	
	p-Wert	< .00001	—	
age	Pearson's r	0,464	0,495	—
	df	373	373	—
	p-Wert	< .00001	< .00001	—
	Spearman's Rho	0,454	0,485	—
	df	373	373	—
	p-Wert	< .00001	< .00001	—

Abbildung 8.8: Jamovi Ausgabe.

**Übung 8.5.**

Berechnen Sie (a) Punktbiseriale Korrelation und (b) die biserielle Rangkorrelation für die Übung 6.5, wo das Burnout-Risiko von verheirateten und unverheirateten Ärztinnen analysiert wird und interpretieren Sie die Werte.

*Lösung.*

- a) Zuerst wird der Datensatz mit **Jamovi** eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 8.11.

Dies produziert das Analyseergebnis in Abbildung 8.12

Daraus ergibt sich

$$r = \frac{10.99 - 10.15}{(51 + 61) \cdot 2.01} \sqrt{51 \cdot 61} = 0.21.$$

Hier wurden die Zwischenresultate mit zwei Nachkommastellen aus **Jamovi** gegeben. Je eher eine Ärztin unverheiratet ist, desto höher ist der Burnout-Wert. Der Zusammenhang ist als mittel einzustufen.

- b) Die Analyseparameter in **Jamovi** werden nun geändert, siehe Abbildung 8.13.

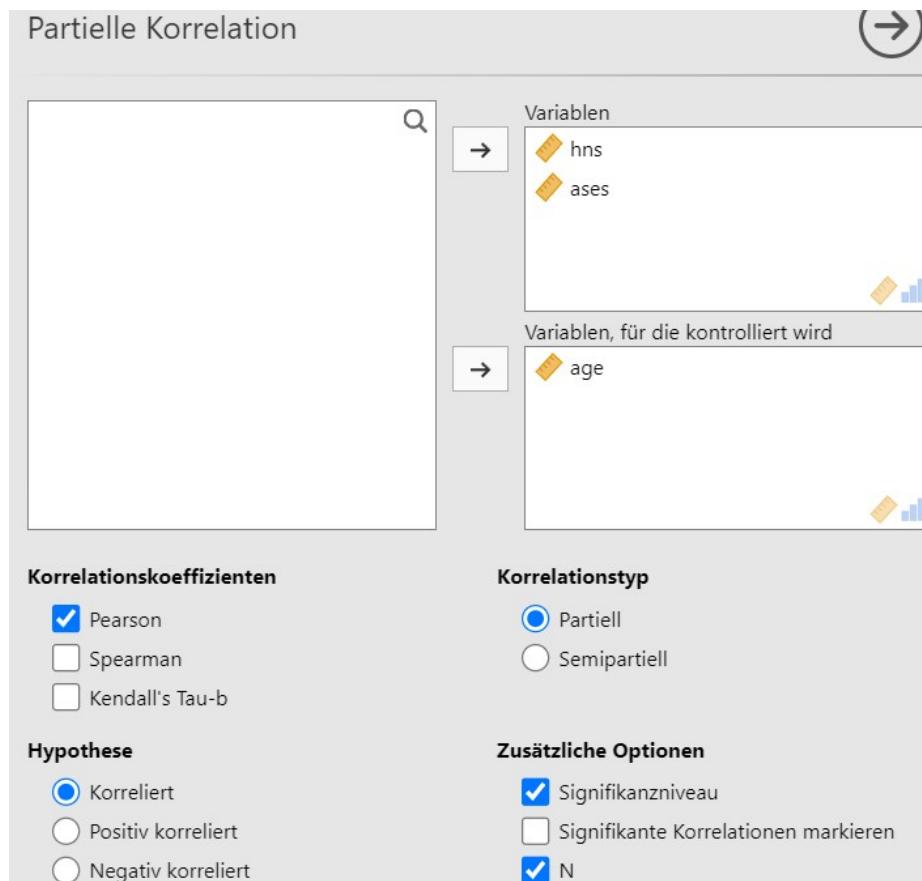


Abbildung 8.9: Jamovi Eingabe.

Partielle Korrelation		
	hns	ases
hns	Pearson's r	—
	p-Wert	—
ases	Pearson's r	0.018
	p-Wert	0.72818

Anmerkung. Kontrolliert für 'age'

Abbildung 8.10: Jamovi Ausgabe.

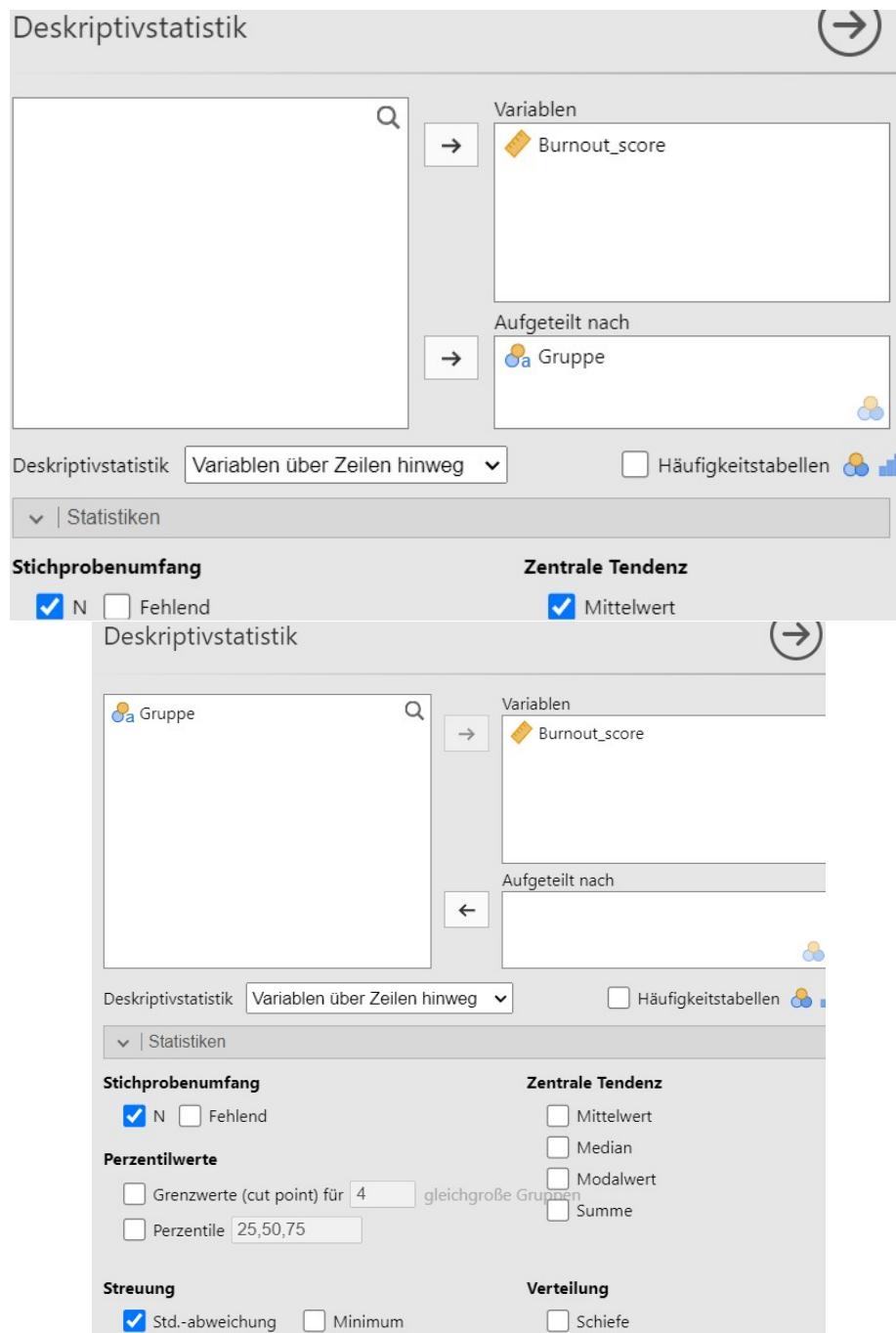


Abbildung 8.11: Jamovi Eingabe.

Deskriptivstatistik			
	Gruppe	N	Mittelwert
Burnout_score	Unverheiratet	51	10.990
	Verheiratet	61	10.151

Deskriptivstatistik		
	N	Std.-abw.
Burnout_score	112	2.005

Abbildung 8.12: Jamovi Ausgabe.

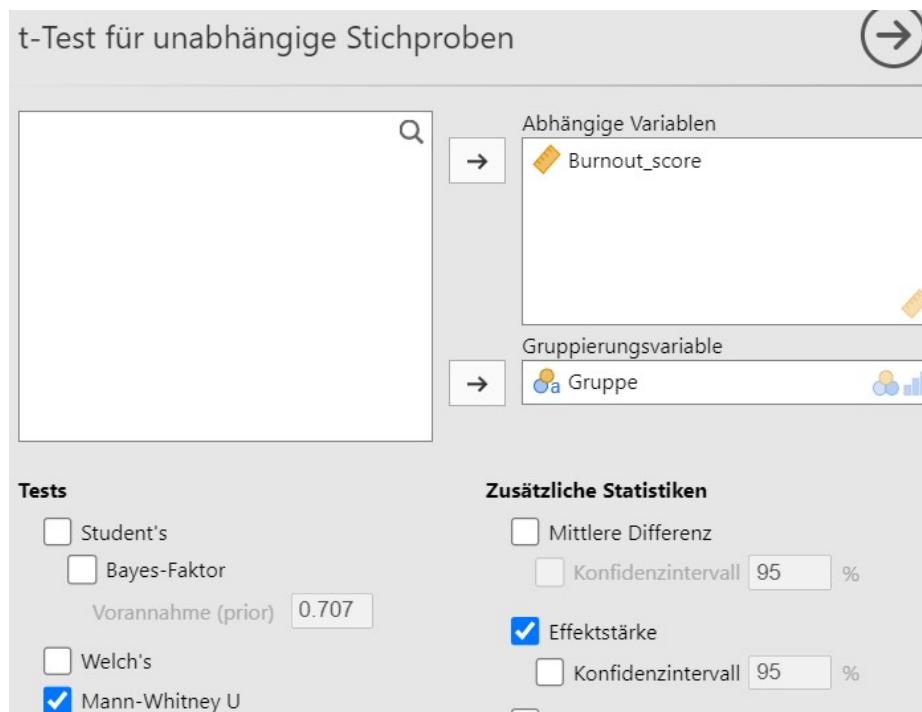


Abbildung 8.13: Jamovi Eingabe.

t-Test für unabhängige Stichproben

		Statistik	p	Effektstärke	
Burnout_score	Mann-Whitney U	1226.000	0.05458	Biserielle Rangkorrelation	-0.212

Anmerkung.  $H_0: \mu_{\text{Unverheiratet}} \neq \mu_{\text{Verheiratet}}$

Abbildung 8.14: Jamovi Ausgabe.

Dies produziert das Analyseergebnis in Abbildung 8.14

Die biseriale Rangkorrelation ist also  $r = -0.212$ . Je eher eine Ärztin verheiratet ist, desto tiefer ist der Burnout-Wert. Der Zusammenhang ist als mittel einzustufen. Achtung das Vorzeichen der biseriellen Rangkorrelation hängt einzig davon ab, welche Gruppe (verheiratet/unverheiratet) als erste Gruppe in **Jamovi** hinterlegt ist. In diesem Datensatz ist dies, sofern nichts geändert wird die unverheiratete Gruppe.

### Übung 8.6.

Es ist bekannt, dass Prüfungsangst zu schlechteren Prüfungsergebnissen führt. In einer Studie soll herausgefunden werden, wie dieser Zusammenhang durch die Vorbereitung beeinflusst wird (Yusefzadeh et al., 2019). Dafür wurde die Prüfungsangst mit dem test anxiety inventory TAI (von 0 – 75), das Prüfungsergebnis (von 0 – 100 Punkte) und die Vorbereitung (von 0 – 10) gemessen. Letztere beruhte darauf, wie viele Prüfungsvorbereitungsgelegenheiten wahrgenommen wurde.

- a) Berechnen Sie die Korrelation zwischen Prüfungsangst und Prüfungsergebnis und sichern Sie diese zweiseitig gegen Null ab. Berichten Sie das Ergebnis.
- b) Berechnen Sie die partielle Korrelation indem Sie für die Vorbereitung kontrollieren und sichern Sie diese zweiseitig gegen Null ab. Berichten Sie das Ergebnis.

*Lösung.* Die Analyseparameter in **Jamovi** werden nun geändert, siehe Abbildung 8.15.

Dies produziert das Analyseergebnis in Abbildung 8.16

- a) Die zweiseitige Absicherung gegen Null des Korrelationskoeffizienten nach Pearson ergibt, dass sich die Korrelation zwischen Prüfungsangst und Prüfungsresultat ( $r = -0.17$ ) signifikant von 0 unterscheidet,  $t(168) = -2.22$ ,  $p = .028$ . Die Korrelation deutet auf einen schwachen Zusammenhang der Art je mehr Prüfungsangst desto schlechter das Prüfungsresultat.

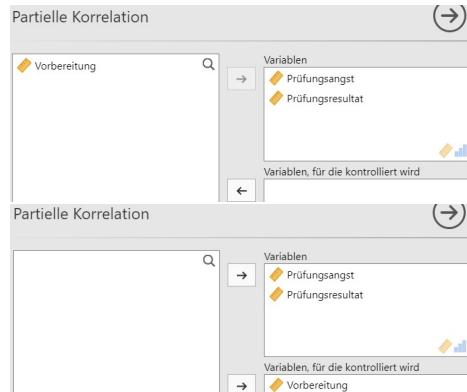


Abbildung 8.15: Jamovi Eingabe.

Korrelation		
	Prüfungsangst	Prüfungsresultat
Prüfungsangst	Pearson's r — p-Wert —	
Prüfungsresultat	Pearson's r —0.169 p-Wert 0.02777	—

Partielle Korrelation		
	Prüfungsangst	Prüfungsresultat
Prüfungsangst	Pearson's r — p-Wert —	
Prüfungsresultat	Pearson's r —0.102 p-Wert 0.18883	—

*Anmerkung.* Kontrolliert für 'Vorbereitung'

Abbildung 8.16: Jamovi Ausgabe.

- b) Die zweiseitige Absicherung gegen Null der partiellen Korrelation ergibt, dass sich die Korrelation zwischen Prüfungsangst und Prüfungsresultat unter Berücksichtigung der Vorbereitung als Störfaktor ( $r = -0.1$ ) nicht signifikant von 0 unterscheidet,  $t(167) = -1.32$ ,  $p = .189$ . Die partielle Korrelation deutet auf einen schwachen Zusammenhang der Art je mehr Prüfungsangst desto schlechter das Prüfungsresultat.

## 8.7 Test

**Übung 8.7.** Welche der folgenden Aussagen zum Zusammenhang zwischen zwei Merkmalen sind wahr, welche falsch?

- a) Eine Korrelation von  $r = -0.23$  deutet auf einen mittleren Zusammenhang der Art je mehr desto weniger hin.
- b) Ein Grund für eine Korrelation von  $r = 0$  kann sein, dass der Zusammenhang nicht linear ist.
- c) Je weiter die Korrelation von 0 weg ist, desto eher ist der Zusammenhang von der Stichprobe auf die Grundgesamtheit übertragbar, gegeben dass die Stichprobengröße gleich bleibt.
- d) Eine Korrelation von  $r = 0.2$  bedeutet, dass das erste Merkmal ursächlich für den Zusammenhang ist.

*Lösung.*

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

**Übung 8.8.** Welche der folgenden Aussagen zum Zusammenhang zwischen zwei Merkmalen sind wahr, welche falsch?

- a) Um den Zusammenhang zwischen Gästezufriedenheit (Likert 1 bis 5) und der Zimmerart (basic oder superior) zu ermitteln, kann die Korrelation nach Spearman verwendet werden.
- b) Um den Zusammenhang zwischen Anzahl Konflikten um Wasserrechte (intervallskaliert, nicht normalverteilt) und dem Wohlstand der beiden Gemeinden (intervallskaliert, normalverteilt) zu ermitteln, kann die Korrelation nach Pearson verwendet werden.
- c) Um den Zusammenhang zwischen Alphabetisierungsrate (intervall-skaliert, normalverteilt) und Bruttoinlandsprodukt (intervallskaliert, normalverteilt) von Ländern zu ermitteln, kann die Korrelation nach Pearson verwendet werden.

- d) Die partielle Korrelation ist immer weniger oder gleich stark, als die direkt gemessene Korrelation.

*Lösung.*

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

### Übung 8.9.

Im folgenden wird eine Korrelation zwischen der Wartezeit am Flughafen und der Gästezufriedenheit berechnet. Dabei wird festgestellt, dass beide Merkmale intervallskaliert aber nicht normalverteilt sind. Jamovi liefert die Ausgabe in Abbildung 8.17.

Korrelationsmatrix		wartezeit	zufriedenheit
wartezeit	Pearson's r	—	—
zufriedenheit	Pearson's r	-0.302 ***	—
df		120	—
p-Wert		0.00072	—

Anmerkung. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

Abbildung 8.17: Jamovi Eingabe.

Welche der folgenden Aussagen sind wahr welche falsch?

- a) Es ist richtig die Korrelation hier mit dem Korrelationskoeffizienten nach Pearson zu berechnen.
- b) Die Korrelation sagt aus, dass die Gäste mit höherer Zufriedenheit weniger lange warten.
- c) Der Zusammenhang ist schwach.
- d) Es wurden 120 Personen befragt.
- e) Die Nullhypothese, dass die Korrelation in der Population bei -0.302 liegt, kann aufgrund des vorliegenden Testresultats abgelehnt werden.
- f) Die Korrelation unterscheidet sich signifikant von 0.

*Lösung.*

- a) Falsch
- b) Richtig

- c) Falsch
- d) Falsch
- e) Falsch
- f) Richtig



# Kapitel 9

## Zusammenhang dichotomer Merkmale

### 9.1 Zusammenhang dichotomer Merkmale beschreiben

**Beispiel 9.1** (Alkohol und Bildung). Vom Bundesamt für Statistik BFS werden regelmässig Daten zum Alkoholkonsum in der Schweiz erhoben. Dabei wird ermittelt, welcher Anteil der Bevölkerung weniger als einmal pro Woche und welcher Anteil mehr als einmal pro Woche Alkohol konsumiert. Diese Anteile werden anschliessen für verschiedene Untergruppen ausgewiesen, zum Beispiel für Leute mit Tertiärbildung und anderem Bildungsabschluss. Es werden hier also zwei dichotome Merkmale (Bildung: Tertiär/nicht Tertiär und Alkoholkonsum: mind. 1x / Woche, weniger als 1x / Woche) und deren Zusammenhang betrachtet. Um diese Anteile abzuschätzen werden 1000 Personen befragt. Der Datensatz ist als `09-exm-alcohol-edu.sav` verfügbar.

Im Datensatz wird für jede Person eine Zeile ausgewiesen, siehe Abbildung 9.1.

Da der Mensch nicht besonders gut darin ist unzählige Zeilen eine Tabelle zu absorbieren, werden die Daten oft in einer Vierfeldertafel zusammengefasst. Die **Vierfeldertafel** ist eine Kreuztabelle bei welcher die Ausprägungen des einen Merkmals als Spalten und des anderen als Zeilen fungieren. Die Zellen werden dann mit der Anzahl Beobachtungen befüllt, für welche die Ausprägungskombination im Datensatz zutrifft, siehe Abbildung 9.1.

Im psychologischen und medizinischen Kontext bezieht sich ein Merkmal oft auf einen schädlichen und einen nicht schädlichen Ausgang. Im Beispiel ist es der Alkoholkonsum. Dieses Merkmal wird **Risikovariable** genannt. In der Vierfeldertafel kann die Risikovariable die Zeilen oder die Spalten bestimmen. Die

	bildung	alkoholkonsum
1	tertiär	selten
2	tertiär	selten
3	andere	oft
4	andere	selten
5	andere	oft
6	tertiär	oft
7	andere	oft
8	andere	oft
9	andere	oft
10	tertiär	oft

Abbildung 9.1: Daten Alkoholkonsum und Bildung.

Kreuztabellen			
bildung	alkoholkonsum		
	oft	selten	Insgesamt
tertiär	308	70	378
andere	476	146	622
Insgesamt	784	216	1000

Abbildung 9.2: Vierfeldertafel Alkoholkonsum und Bildung..

## 9.1. ZUSAMMENHANG DICHOTOMER MERKMALE BESCHREIBEN 171

Risikovariable wird durch andere Merkmale sogenannte ursächliche Variablen erklärt. Im breiteren statistischen Kontext wird die Risikovariable abhängige Variable und die ursächliche Variable als unabhängige Variable bezeichnet.

Mit **Risiko** wird die Wahrscheinlichkeit benannt den schädlichen Ausgang zu erleiden und wird mit dem Anteil des schädlichen Ausgangs an der Gesamtzahl berechnet. Das Risiko kann je nach Ausprägung der ursächlichen Variable unterschiedlich hoch sein. Das Risiko mehr als 1x pro Woche Alkohol zu konsumieren ist demnach  $308/378 = 0.815 = 81.5\%$  für Menschen mit tertiärer Ausbildung und  $476/622 = 0.765 = 76.5\%$  für Menschen mit anderer Ausbildung. Das Risiko kann in Jamovi unter **Zellen > Prozentsätze** und dann **Zeile** (wenn das ursächliche Merkmal die Zeilen bestimmt) oder **Spalte** (wenn das ursächliche Merkmal die Spalte bestimmt) angezeigt werden.

Die Risiken für die tertiär und andere Ausbildungen können nun verglichen werden. Dazu kann die Differenz der beiden Risiken sogenannte **Risikodifferenz** (in Jamovi unter **Statistiken > Unterschiede in den Proportionen**) betrachtet werden. Wenn die Risiken der beiden Gruppen mit  $p_1$  und  $p_2$  bezeichnet werden, entspricht dies schlicht

$$p_1 - p_2 = 0.815 - 0.765 = 0.05.$$

Das Risiko mehr als 1x pro Woche Alkohol zu konsumieren ist also  $0.05 = 5\%$  höher für Personen mit einem tertiären Bildungsabschluss. Da die Risiken immer Werte zwischen 0 und 1 sind, muss dieser Formel nach die Differenz der Risiken zwischen  $-1$  und  $1$  liegen. Wenn die Differenz der beiden Risiken 0 ist, bedeutet dies, dass die Risiken in beiden Gruppen gleich gross sind. Je weiter die Differenz der Risiken von 0 weg ist, desto unterschiedlicher sind die Risiken in den zwei Gruppen.

Eine andere Art die Risiken zu vergleichen ist sie ins Verhältnis zu setzen. Dies wird **relatives Risiko**

$$\text{RR} = \frac{p_1}{p_2} = \frac{0.815}{0.765} = 1.065$$

genannt. Das Risiko mehr als 1x pro Woche Alkohol zu konsumieren ist für Personen mit tertiärer Ausbildung also 1.065 mal so gross wie für Personen mit anderer Ausbildung. Sind  $p_1$  und  $p_2$  gleich gross, so ist das relative Risiko bei 1. Ist  $p_1$  kleiner als  $p_2$ , so ist das relative Risiko kleiner als 1. Ist  $p_1$  grösser als  $p_2$ , so ist das relative Risiko grösser als 1. Insgesamt ist das relative Risiko immer eine Zahl zwischen 0 und  $+\infty$ . Das Tauschen der Gruppennummerierung führt zu einer Umkehr des Wertes rund um 1. Wenn im Beispiel also die andere Bildung als Gruppe 1 bezeichnet würde, so ist das relative Risiko

$$\text{RR} = \frac{p_1}{p_2} = \frac{0.765}{0.815} = 0.939$$

**Achtung**



*Hinweis.* Das hier die relativen Risiken für die Gruppenneumerierung fast gleich weit von 1 entfernt liegen ist im Normalfall nicht so. Dies kann an folgendem Zahlenbeispiel gesehen werden:  $0.75/0.25 = 3$  und umgekehrt  $0.25/0.75 = 0.333$ .

Das relative Risiko sagt im Gegensatz zu Risikodifferenz nichts mehr über das absolute Risiko aus. Wenn ein Medikament, zum Beispiel das Risiko einer Psychose von 1 aus 5000 auf 1 aus 10000 reduziert, so ist die Risikodifferenz bei  $0.0001 - 0.0002 = -0.01\%$ . Das relative Risiko ist in dem Fall jedoch  $0.0001/0.0002 = 50\%$ . Wird nun nur das relative Risiko berichtet, könnten Lesende von einem zu grossen Nutzen des Medikaments ausgehen. Es wird deshalb empfohlen immer die Risikodifferenz und das relative Risiko zu berichten.

Die Risikodifferenz und das relative Risiko können nur berechnet werden, wenn die gesamt Anzahl Fälle repräsentativ für die Population ist. Dies ist bei Fall-Kontroll Studien nicht der Fall, wie folgendes Beispiel zeigt. Für solche Studien können weder die Risikodifferenz noch das relative Risiko sinnvoll berechnet werden.

**Beispiel 9.2** (Krebs bei Hunden). Hayes et al. (1991) haben sich für den Zusammenhang zwischen malignen Lymphomen bei Hunden (ugs. bösartiger Lymphdrüsengeschwülste) und der Anwendung des Herbizids 2,4-Dichlorphenoxyessigsäure in Hausgärten interessiert. Dabei haben Sie in einer Fall-Kontroll Studie (case-control study) die Zahlen im Datensatz **09-exm-dog-cancer.sav** ermittelt.

Die Daten sind hier bereits in aggregierten Fallzahlen präsentiert, siehe Abbildung 9.3.

	Ca Hund	Ca Herbizid	Anzahl
1	Tumor	Angewendet	191
2	Kein Tumor	Angewendet	304
3	Tumor	Nicht Angew...	300
4	Kein Tumor	Nicht Angew...	641

Abbildung 9.3: Daten Malignes Lymphoma bei Hunden.

Um diese Daten korrekt in Jamovi einzulesen, kann unter **Analysen > Häufigkeiten > Kreuztabellen > Unabhängige Stichproben** die Anzahl Fälle bei **Anzahl (optional)** eingelesen werden. Dies resultiert in der Vierfeldertafel in Abbildung 9.4.

Case-control bedeutet, dass Hunde mit Tumor (Fall/case) in einer Tierklinik gegeben waren. Dazu wurde eine gewisse Anzahl (normalerweise zwischen 1 bis 4 mal so viele wie kranke) gesunde Hunde (Kontroll/control) zufällig ausgewählt.

### 9.1. ZUSAMMENHANG DICHOTOMER MERKMALE BESCHREIBEN 173

		Herbizid		Insgesamt
		Angewendet	Nicht Angewendet	
Hund	Tumor	191	300	491
	Kein Tumor	304	641	945
	Insgesamt	495	941	1436

Abbildung 9.4: Vierfeldertafel Malignes Lymphoma bei Hunden.

Bei allen Hunden wurde anschliessend ermittelt, ob die Hunde auf einem mit dem entsprechenden Herbizid belasteten Garten Zeit verbracht haben. Da von den Studienautoren bestimmt wurde, wie viele gesunde Hunde ermittelt werden, kann der Anteil der kranken Hunde nicht als Mass für das Vorkommen der Erkrankung in der jeweiligen Gruppe dienen. Es ist hier also nicht aussagekräftig den Anteil kranker Hunde pro Gruppe oder das daraus folgende relative Risiko zu bestimmen.

Stattdessen wird in diesen Fällen das **Chancenverhältnis** (eng. odds ratio)

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \cdot d}{b \cdot c} = \frac{191 \cdot 641}{300 \cdot 304} = 1.34$$

berechnet, wobei  $a, b, c$  und  $d$  den Anzahl Fällen in der Vierfeldertafel von oben nach unten und von links nach rechts entsprechen. Die **Chance** (eng. odds) ist dabei eine Art Wahrscheinlichkeit auszudrücken. Sie ist definiert als Wahrscheinlichkeit, dass ein Ereignis eintrifft geteilt durch die Wahrscheinlichkeit, dass das Ereignis nicht eintrifft. Für die Herbizid belasteten Hunde ist die Chance ein Tumor zu haben also  $a/c = 191/304 = 0.63$  und für die anderen Hunde  $b/d = 300/641 = 0.47$ . Es gilt, je höher die Chance desto höher die Eintreffwahrscheinlichkeit. Ein Chancenverhältnis von 1.34 schliesslich bedeutet, dass die Chance einen Tumor zu entwickeln für einen Herbizid belasteten Hund 1.34-mal so hoch ist wie für einen nicht Herbizid belasteten.

**Achtung***Hinweis.*

- Es kann festgestellt werden, dass das Chancenverhältnis unabhängig von der Wahl der ursächlichen und Risikovariable ist

$$OR = \frac{a \cdot d}{b \cdot c} = \frac{a \cdot d}{c \cdot b}.$$

- Für kleine Fallzahlen (Risiko < 10%) liegt das Chancenverhältnis nahe am relativen Risiko.
- Das Chancenverhältnis lässt, wie das relative Risiko, keinen Schluss über das absolute Risiko zu.

## 9.2 Zusammenhang dichotomer Merkmale testen

In Jamovi > Analysen > Häufigkeiten > Kreuztabellen > Unabhängige Stichproben > Statistiken > Tests stehen verschiedene Tests zur Verfügung, um zu testen, ob ein in einer Zufallsstichprobe gefundener Zusammenhang zwischen den Risiken  $p_1$  und  $p_2$  auf die Risiken in der Population  $\pi_1$  und  $\pi_2$  (sprich 'pi') übertragen werden darf.

### 9.2.1 $z$ -Test für den Unterschied zwischen den zwei Anteilen

Wenn die Merkmale unabhängig voneinander sind, dann sollte das Risiko in beiden Gruppen ungefähr gleich gross sein. Dies impliziert auch, dass die Risikodifferenz 0 ist. Beim  $z$ -Test für den Unterschied zwischen den zwei Anteilen wird also  $H_0 : \pi_1 - \pi_2 = 0$  und damit äquivalent  $H_0 : \pi_1 = \pi_2$  getestet. Wird die Nullhypothese verworfen, wird fortan an die Alternativhypothese  $H_1 : \pi_1 \neq \pi_2$  geglaubt. Aus Platzgründen werden hier nur die zweiseitigen Hypothesenstellungen erwähnt. Die einseitigen Hypothesenstellungen funktionieren aber analog zu den bisher gesehenen Tests.

Die Teststatistik für den  $z$ -Test der Risikodifferenz ist

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = -1.85,$$

wobei  $n_1$  und  $n_2$  der Anzahl vom Risiko betroffene bzw. nicht betroffene Beobachtungen darstellen. Wenn die Nullhypothese wahr ist, ist diese Test-

statistik ist standardnormalverteilt. Dies entspricht der Normalverteilung in Abbildung 4.7. Je grösser die Differenz der beiden Risiken, desto weiter weg von 0 ist die Teststatistik und desto unwahrscheinlicher ist es, dass die berechnete Teststatistik aufgrund der Zufallsstichprobenziehung zustande kommt. Der genaue  $p$ -Wert kann mit Jamovi unter Tests > z-Test für den Unterschied zwischen den zwei Anteilen ermittelt werden. Mit  $p = .0649$  kann die Nullhypothese knapp nicht verworfen werden. Die beiden Anteile  $\pi_1$  und  $\pi_2$  könnten also gleich sein.

Ein zweiseitiger  $z$ -Test für den Unterschied zwischen den zwei Anteilen ergibt, dass der Anteil der mehr als einmal pro Woche alkoholkonsumierenden Personen mit Tertiärbildung ( $p_1 = 81.5\%$ ), nicht signifikant von ebendiesem Anteil der Personen mit anderer Bildung ( $p_2 = 76.5\%$ ) unterscheidet,  $z = -1.85, p = .0649$ .

### 9.2.2 $\chi^2$ -Test

Der  $\chi^2$ -Test (sprich ‘chi-quadrat-test’) ist eine zweite allgemeinere Variante, um den Zusammenhang zwischen zwei dichotomen Merkmalen zu testen. Die Nullhypothese ist wiederum, dass die beiden Merkmale unabhängig voneinander sind.

Um nun die Teststatistik näher verstehen zu können, muss ein kurzer Exkurs in Wahrscheinlichkeitslehre unternommen werden. Diese besagt nämlich, dass wenn zwei Ereignisse A und B unabhängig voneinander auftreten, dann ist die Auftretenswahrscheinlichkeit genau gleich dem Produkt der Wahrscheinlichkeiten der jeweiligen Ereignisse. In Formeln ausgedrückt

$$p(A \text{ und } B) \stackrel{\text{unabhängig}}{=} p(A) \cdot p(B).$$

Auf das Beispiel 9.1 bezogen bedeutet dies, dass wenn Alkoholkonsum und die Ausbildungsart unabhängig voneinander sind, dann entspricht die Wahrscheinlichkeit mehr als 1x pro Woche Alkohol zu konsumieren und eine tertiäre Ausbildung genossen zu haben

$$p(\text{[alkohol} > 1\text{x pro Woche] und tertiar}) = \frac{476}{784} = 0.60714$$

genau

$$p(\text{[alkohol} > 1\text{x pro Woche]}) \cdot p(\text{tertiär}) = \frac{784}{1000} \cdot \frac{622}{1000} = 0.48765$$

entsprechen. Die Zahlen zeigen an, dass dies für diese Stichprobe klar nicht gegeben ist. Die beobachtete Differenz könnte sich aber auch aus der Zufallsstichprobenziehung ergeben haben. Der  $\chi^2$ -Vierfeldertest vergleicht nun in

jeder Zelle  $i$  der Vierfeldertafel die tatsächlich beobachteten Beobachtungen  $o_i$  (observed) mit den erwarteten Beobachtungen  $e_i$  (expected), wenn die Unabhängigkeit gegeben wäre. Letztere berechnet sich durch  $p(A) \cdot p(B) \cdot n$ , wobei  $n$  die Gesamtzahl Beobachtungen ist. Im Beispiel

$$e_1 = p(\text{alkohol} > 1x \text{ pro Woche}) \cdot p(\text{tertiär}) \cdot n = 0.48765 \cdot 1000 = 487.648$$

Je weiter diese Zahl von der beobachteten Zahl abweicht, desto unwahrscheinlicher ist die Unabhängigkeit. Der  $\chi^2$  trägt dem Rechnung, indem die Teststatistik

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = 3.41,$$

wobei  $k$  hier für die Anzahl Zellen steht. Eine grosse Teststatistik spricht also gegen die Unabhängigkeit.

Wenn die Stichprobenziehung oft wiederholt wird, kann festgestellt werden, dass diese Teststatistik einer bekannten Verteilung, der  $\chi^2$ -Verteilung bei  $df = 1$  Freiheitsgraden folgt. Die  $\chi^2$ -Verteilung ist für verschiedene Freiheitsgrade in Abbildung 9.5 dargestellt. Die Teststatistik des  $\chi^2$ -Vierfeldertests 3.41 wird also mit der roten Verteilung in der Abbildung verglichen. Die Werte rechts auf der Abbildung sind seltener und der beobachtete Wert liegt so, dass er zu den  $p = .065$  seltensten Beobachtungen zählt, sofern die Unabhängigkeit gilt. Dies reicht hier gerade nicht, um die Nullhypothese bei Signifikanzniveau  $\alpha = 5\%$  zu verwerfen.

Um die Effektstärke des  $\chi^2$ -Tests anzugeben kann **Cramérs  $\phi$**

$$\phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{3.41}{1000}} = 0.06$$

Cramérs  $\phi$  immer grösser als 0 und kleiner als 1. Je weiter  $\phi$  von 0 weg ist, desto stärker sind die Merkmale voneinander abhängig. Da dies auch als Zusammenhangsmass gesehen werden kann wird Cramérs  $\phi$  auch **Vierfelderkorrelation** genannt. Die Interpretation als Effektstärke erfolgt dabei wie für die Korrelation. Ein Wert von 0.06 wird demnach als schwach eingestuft.

Ein  $\chi^2$ -Test ergibt, dass der Alkoholkonsum (mehr/weniger als 1x pro Woche) und Bildung (Tertiär/andere Bildung) ( $p_1 - p_2 = 0.05$ ,  $RR = 1.06$ ) nicht signifikant abhängig voneinander sind,  $\chi^2(1) = 3.41$ ,  $p = .0649$ ,  $\phi = 0.06$ .

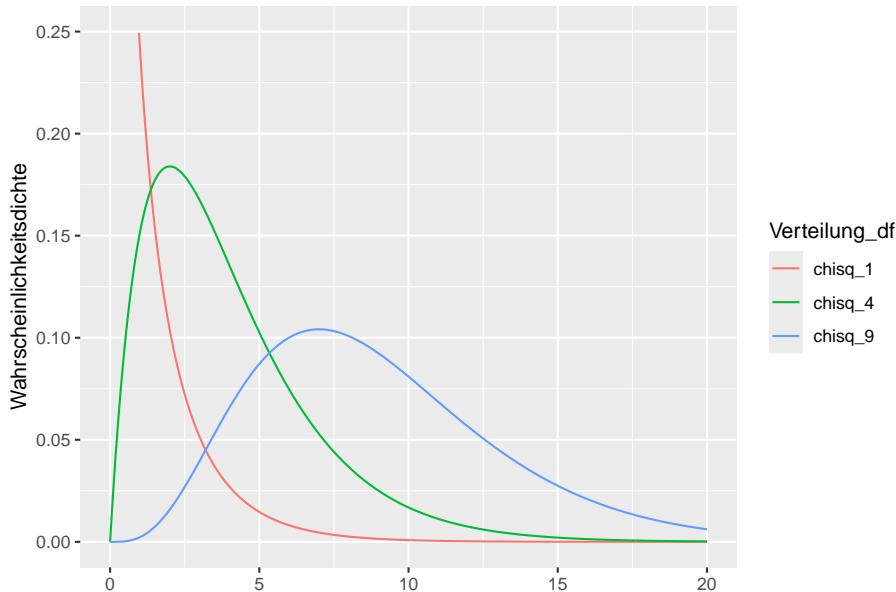


Abbildung 9.5: Chiquadrat-Verteilungen mit 1, 4 und 9 Freiheitsgraden.

Der  $\chi^2$  Test hat für die Vierfeldervariante immer den gleichen  $p$ -Wert wie der Test der Risikodifferenz auf 0. Beide Tests basieren auf dem zentralen Grenzwertsatz und dürfen deshalb nur angewendet werden, wenn gewisse Voraussetzungen erfüllt sind. Die Voraussetzung ist, dass insgesamt 40 oder mehr Beobachtungen vorliegen müssen. Ist dies nicht der Fall, aber in jeder Zelle ist die Anzahl der erwarteten Beobachtungen bei 5 oder grösser, so kann der  $\chi^2$ -Test mit **Kontinuitätskorrektur** oder Yates-Korrektur verwendet werden

$$\chi^2_{Yates} = \sum_{i=1}^k \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

In Beispiel sind mit  $n = 1000$  genügend Beobachtungen vorhanden, um ohne die Kontinuitätskorrektur auszukommen. Der Bericht des Tests ist genau gleich wie der Bericht des  $\chi^2$ -Test ausser, dass der Name des Verfahrens geändert wird:

Ein  $\chi^2$ -Test mit Kontinuitätskorrektur ergibt, dass der Alkoholkonsum ...

### 9.2.3 Exakter Test nach Fisher und Yates

Wenn die erwarteten Beobachtungen in mindestens einer Zelle kleiner als 5 sind, dann ist die Approximation der Verteilung der Teststatistik durch den  $\chi^2$ -Test

sogar mit Kontinuitätskorrektur zu ungenau.

**Beispiel 9.3** (Wette: Unterscheiden von Rot- und Weisswein). Nina behauptet, sie kann Rotwein von Weisswein am Geschmack unterscheiden. Sie bereiten ein Experiment vor, um dies zu testen verbinden Nina die Augen. Als Statistiklernde wissen sie, dass es nicht genügt, dass Nina einmal die richtige Weinsorte wählt - dies könnte ja zufällig richtig sein. Stattdessen werden 5 Gläser Rotwein und 4 Gläser Weisswein vorbereitet und Nina in zufälliger Reihenfolge zum Probieren angeboten. Das Experiment endet mit den Zahlen in Abbildung 9.6.

Kreuztabelle

		sagt		Insgesamt
ist		rotwein	weisswein	
rotwein		4.00	1.00	5.00
weisswein		1.00	3.00	4.00
Insgesamt		5.00	4.00	9.00

Abbildung 9.6: Vierfeldertafel Wette Unterschied Rot- und Weisswein.

Da nur wenige Beobachtungen gemacht wurden, kann die Wahrscheinlichkeit eines Ausgangs des Experiments mit Hilfe der hypergeometrischen-Verteilung genau berechnet werden. Die Wahrscheinlichkeit für den Experimentausgang in Abbildung 9.6 entspricht  $p_0 = 0.159$ .

Der  $p$ -Wert eines Tests ist bekannterweise die Wahrscheinlichkeit die berechnete Teststatistik zu beobachten *oder eine unwahrscheinlichere Situation im Sinne der Alternativhypothese*. Nina will zeigen, dass Sie Rotwein und Weisswein richtig erkennen kann, was der Alternativhypothese entspricht. Ihr ist gedient, wenn sie Rot- und Weisswein überzufällig oft richtig erkennt, nicht aber, wenn sie den Rot- und Weisswein unterdurchschnittlich oft richtig erkennt. Die Alternativhypothese ist also hier einseitig formuliert zu verstehen.

Um dem zweiten Teil dieser Definition des  $p$ -Werts gerecht zu werden, werden nun andere, unwahrscheinlichere Experimentausgänge im Sinne der Alternativhypothese bewertet. Um dies zu erreichen, wird hier angenommen, dass in jedem Fall die Randsummen immer gleichbleiben (also 5 für ist Rotwein, 4 für ist Weisswein, 5 für sagt Rotwein und 4 für sagt Weisswein). Wenn ist und sagt unabhängig voneinander ist und Nina also die Weinfarben nicht auseinanderhalten kann, dann wäre es unwahrscheinlicher, wenn sie alle 5 Rotweine als solche erkannt hätte. Wenn die Randsummen gleich gehalten werden, entspricht dies der Situation in Abbildung 9.7. Die Wahrscheinlichkeit für diese Situation ist  $p_1 = 0.008$ .

Grundsätzlich können hier noch weitere Situationen aufgezählt werden. Im Beispiel sind mit der eben beschriebenen Situation und der ursprünglichen

Kreuztabellen

		sagt		Insgesamt
		rotwein	weisswein	
ist	rotwein	5.00	0.00	5.00
	weisswein	0.00	4.00	4.00
Insgesamt		5.00	4.00	9.00

Abbildung 9.7: Vierfeldertafel Wette Unterschied Rot- und Weisswein.

Situation jedoch die Alternativen erschöpft: Es kann nicht sein, dass Nina 6-mal Rotwein richtig voraussagt, wenn sie insgesamt nur 5-mal Rotwein sagt.

Nun werden die Wahrscheinlichkeiten aller Situation aufsummiert,  $p = p_0 + p_1 + \dots = 0.159 + 0.008 = 0.167$ . Da mit diesem Vorgehen genau der Definition des  $p$ -Wertes gefolgt wurde, stellt dieses  $p$  nun auch einen  $p$ -Wert dar. Der dazugehörige Test wird **exakter Test nach Fisher und Yates** oder einfach Fisher-Yates-Test genannt.

Ein einseitiger exakter Test nach Fisher und Yates ergibt, dass das Ansagen der Weinsorte von Nina nicht signifikant von der tatsächlichen Weinsorte abhängt,  $p = .167$ .

**Achtung***Hinweis.*

- Der Test ist exakt, weil hier die Wahrscheinlichkeiten genau bestimmt wurden. Bei allen anderen bislang gesehenen Tests wird die Wahrscheinlichkeit über eine Verteilung angenähert. Diese Annäherung ist theoretisch nur richtig, wenn unendlich viele Beobachtungen gemacht werden, weshalb diese Art Test auch **asymptotisch** genannt wird. In der Praxis wird jedoch festgestellt, dass asymptotische Tests bereits für eine kleine Anzahl Beobachtungen (z. B. 50 für den Einstichproben-*t*-Test) hinreichend genau sind.
- Da für den exakten Test nach Fisher und Yates die Wahrscheinlichkeiten direkt berechnet werden, wird keine Teststatistik verwendet.
- Um eine zweiseitige Alternative testen zu können, müsste der Test Wahrscheinlichkeiten von gleichen oder extremeren Situationen in die andere Richtung dazusummieren. Solche Situationen können nicht immer eindeutig bestimmt werden. Im Beispiel wäre eine Möglichkeit, dass Nina 1-mal Rotwein sagt, wenn es Rotwein ist. Eine andere Möglichkeit wäre, dass Nina 1-mal Weisswein sagt, wenn es Weisswein ist. Grundsätzlich wird hier nur der einseitige exakte Test nach Fisher verlangt.

## 9.3 Übungen

### Übung 9.1.

Viscidi et al. (2013) wollten herausfinden wie sich das Alter (Kind bis 10 Jahre / Jugendlich älter als 10 Jahre) auf das Auftreten von Epilepsie bei Menschen mit einer Autismus-Spektrum-Störung auswirkt. Dafür wurden zufällig 5185 Menschen mit Autismus-Spektrum-Störung zu ihrem Alter und dem Auftreten von Epilepsie befragt. Simulierte Daten befinden sich im Datensatz **09-exr-autism-epilepsy.sav**.

- a) Lässt das Studiendesign das Berechnen des relativen Risikos zu?
- b) Berechnen Sie das Risiko an Epilepsie zu leiden in den beiden Altersgruppen und anschliessend die Risikodifferenz, das relative Risiko und das Chancenverhältnis.
- c) Sichern Sie die Risikodifferenz zweiseitig gegen Null ab und berichten Sie das Ergebnis.
- d) Testen Sie den Zusammenhang mit einem  $\chi^2$ -Test und berichten Sie das Ergebnis. Schätzen und interpretieren Sie auch die Effektstärke.

*Lösung.* Zuerst wird der Datensatz mit Jamovi eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 9.8.

Dies produziert das Analyseergebnis in Abbildung 9.9.

Damit kann die Frage nun beantwortet werden:

- a) Ja die Leute wurden zufällig ausgewählt und nicht aufgrund der Präsenz oder Absenz eines Epilepsieanfalls.
- b) Das Risiko an Epilepsie zu leiden liegt bei 12.0% für Kinder und 26.0% für Jugendliche. Die Risikodifferenz liegt bei 14.0%. Das Risiko ist für Jugendliche 14.0% höher als für Kinder. Das relative Risiko liegt bei 2.17. Das Risiko ist für Jugendliche also 2.17-mal so hoch wie für Kinder. Das Chancenverhältnis liegt bei 2.58. Die Chance an Epilepsie zu leiden ist für Jugendliche also 2.58-mal so hoch wie für Kinder.
- c) Ein zweiseitiger  $z$ -Test für den Unterschied zwischen zwei Anteilen ergibt, dass sich der Anteil der autistischen Jugendlichen mit Epilepsieanfällen ( $p_1 = 26.0\%$ ), signifikant vom Anteil der Kinder mit Epilepsieanfällen ( $p_2 = 12.0\%$ ) unterscheidet,  $z = 12.8, p < .001$ .
- d) Ein  $\chi^2$ -Test ergibt, dass das Auftreten von Epilepsieanfällen (ja/nein) und die Alterskategorie (Jugendlich/Kind) ( $p_1 - p_2 = 0.14, RR = 2.17$ ) signifikant voneinander abhängig sind,  $\chi^2(1) = 163.16, p < 0.001, \phi = 0.18$ . Der Effekt ist schwach.

### Übung 9.2.

Eine Psychologin hat versucht herauszufinden, wie sich das Trainingsverhalten (oft/selten) auf depressive Stimmungen auswirkt. Dazu hat sie in einem Experiment 77 zufällige Leute befragt und die Resultate in Datensatz 09-exr-depression-training.sav erhalten.

- a) Welches ist die Risikovariable?
- b) Wie gross sind die Risiken in den beiden Gruppen? Wie gross ist die Risikodifferenz, das relative Risiko und das Chancenverhältnis? Welche Wirkung hat das Trainingsverhalten auf das Risiko eine Depression zu erleiden?
- c) Sind die Variablen Depression und Trainingsverhalten voneinander abhängig, wenn bei  $\alpha = 5\%$  mit einem  $\chi^2$ -Test getestet wird?
- d) Ist die Kontinuitätskorrektur angebracht? Wie gross ist der Unterschied zwischen der Teststatistik mit und ohne Yates-Korrektur hier?
- e) Wie gross ist die Vierfelderkorrelation zwischen den beiden Variablen? Interpretieren Sie den Zusammenhang.

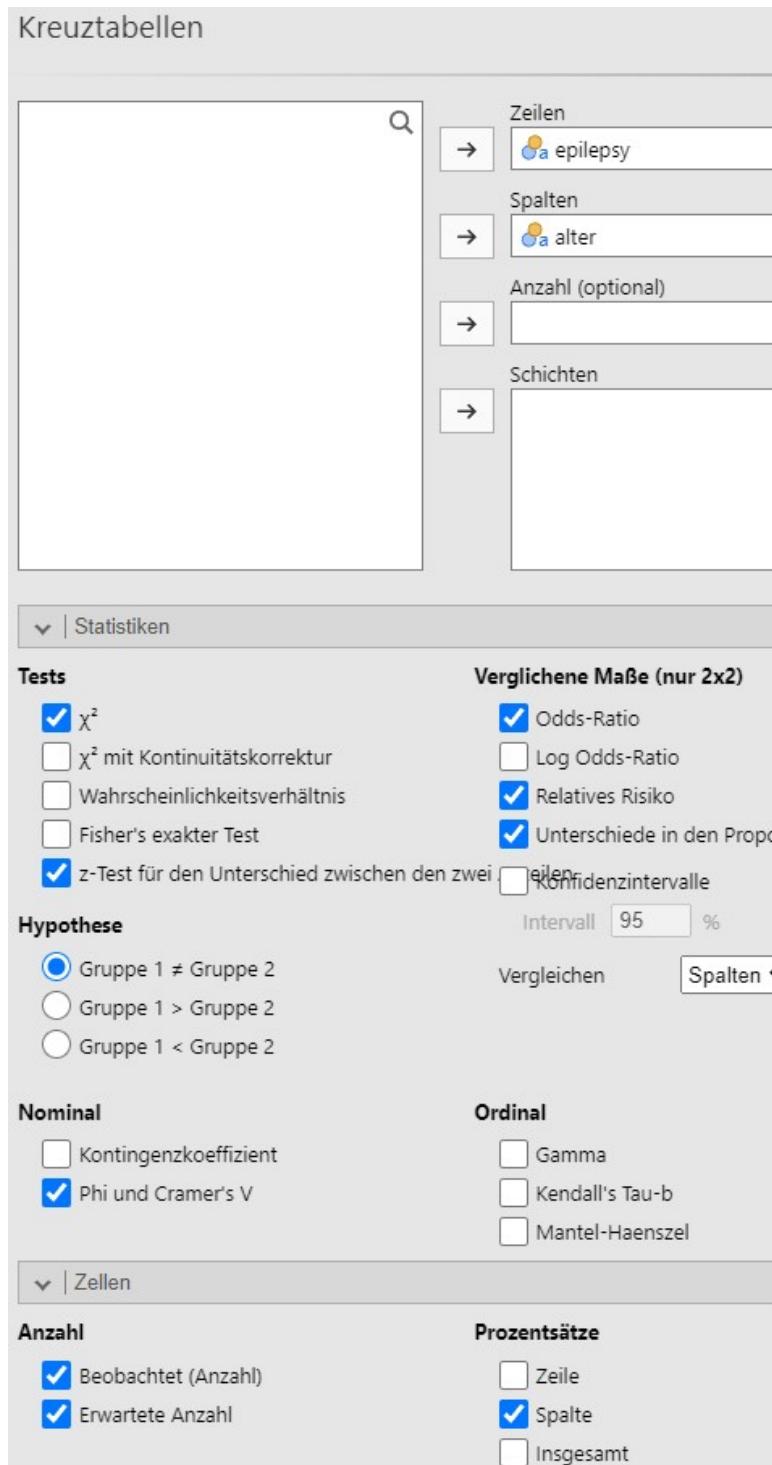


Abbildung 9.8: Jamovi Eingabe.

## Kreuztabellen

		alter			
		epilepsy	jugendlich	kind	Insgesamt
ja	Beobachtet	698	300	998	
	Erwartet	517	481	998	
	% der Spalte	26.0 %	12.0 %	19.2 %	
nein	Beobachtet	1987	2200	4187	
	Erwartet	2168	2019	4187	
	% der Spalte	74.0 %	88.0 %	80.8 %	
Insgesamt	Beobachtet	2685	2500	5185	
	Erwartet	2685	2500	5185	
	% der Spalte	100.0 %	100.0 %	100.0 %	

 $\chi^2$ -Tests

	Wert	df	p
$\chi^2$	163	1	< .00001
z-Test für den Unterschied zwischen den zwei Anteilen	12.8		< .00001
N	5185		

## Verglichene Maße

	Wert
Differenz von 2 Anteilen	0.140*
Odds-Ratio	2.58
Relatives Risiko	2.17*

\* Verglichene Spalten

## Nominal

	Wert
Phi-Koeffizient	0.177
Cramer's V	0.177

Abbildung 9.9: Jamovi Ausgabe.

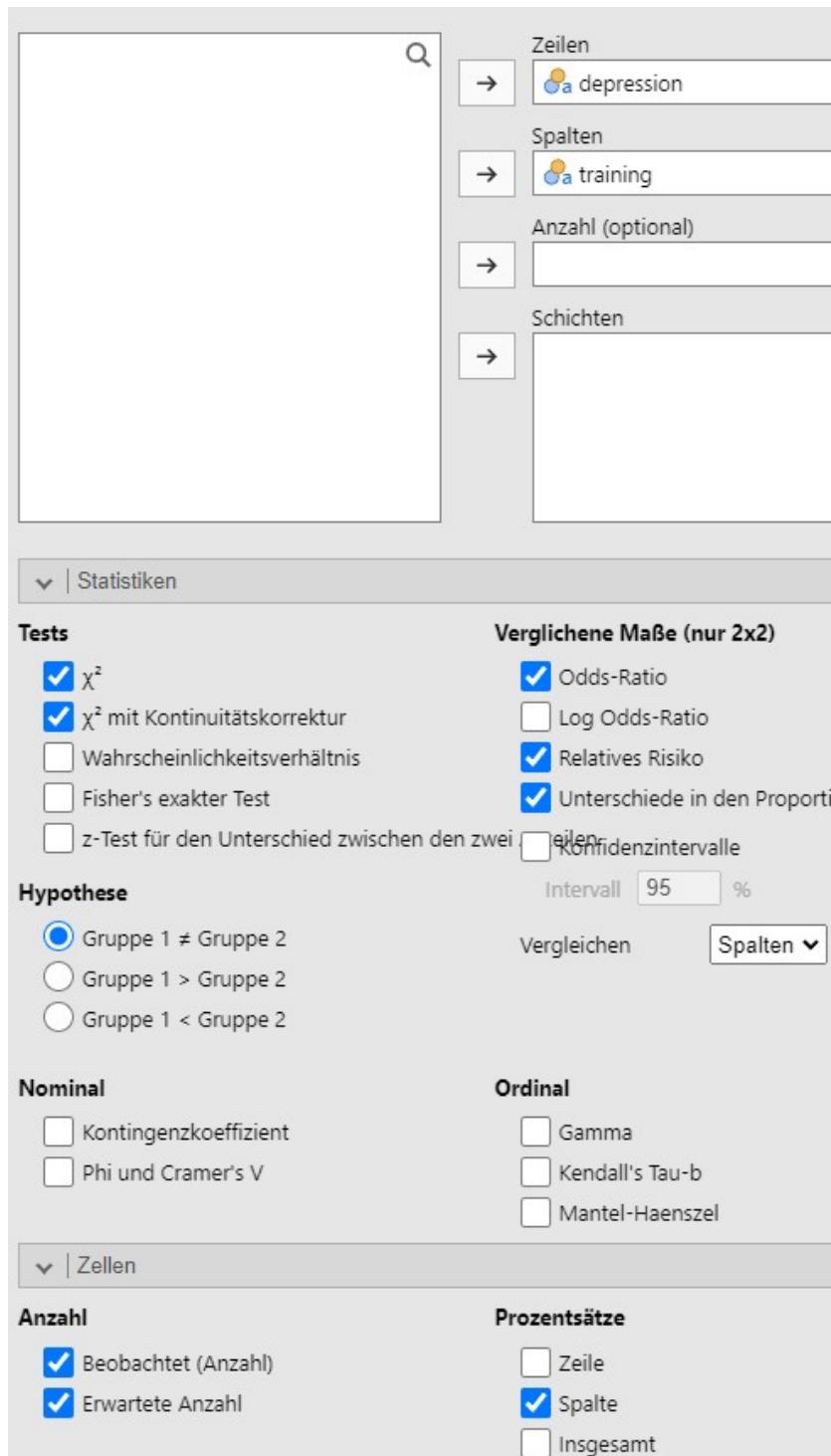


Abbildung 9.10: Jamovi Eingabe.

*Lösung.* Zuerst wird der Datensatz mit Jamovi eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 9.10.

Dies produziert das Analyseergebnis in Abbildung 9.11.

Damit kann die Frage nun beantwortet werden:

- Die Risikovariable ist hier eine Depression zu haben oder nicht.
- Das Risiko beträgt  $p_1 = 48.9\%$  in der oft trainierenden Gruppe. In der anderen Gruppe liegt das Risiko bei  $p_2 = 78.1\%$ . Die Risikodifferenz liegt bei  $-29.2\%$ , das relative Risiko bei 0.626 und das Chancenverhältnis bei 0.268. Oft zu trainieren hat einen protektiven Wirkung auf das Risiko an einer Depression zu erkranken.
- Ein  $\chi^2$ -Test ergibt, dass das Auftreten von Depression (ja/nein) und die Trainingsfrequenz (oft/selten) ( $p_1 - p_2 = -0.29$ ,  $RR = 0.63$ ) signifikant voneinander abhängig sind,  $\chi^2(1) = 6.72$ ,  $p = .01$ ,  $\phi = 0.3$ . Der Effekt ist mittel.
- Die Kontinuitätskorrektur ist nicht nötig da mit 77 Beobachtungen mehr als 40 Beobachtungen vorliegen. Die Teststatistik ohne Kontinuitätskorrektur ist  $\chi^2(1) = 6.72$  und mit Kontinuitätskorrektur  $\chi^2(1) = 5.55$ . Beide Testverfahren deuten auf eine signifikante Abhängigkeit hin.
- Die Vierfelderkorrelation entspricht der Effektstärke  $\phi = 0.3$ . Die Korrelation ist mittel. Die Richtung des Zusammenhangs muss aus den Anteilen abgelesen werden. Je eher jemand unter Depression leidet, desto eher trainiert jemand selten.

### Übung 9.3.

Um den Zusammenhang zwischen Rauchen und Lungenkrebs zu analysieren haben Forschende in einer Studie 200 an Lungenkrebs erkrankte und 397 nicht an Lungenkrebs erkrankte Menschen zu ihrem Rauchverhalten (rauchen ja oder nein) befragt. Die Forschenden haben die Daten im Datensatz **09-exr-cancer-smoking.sav** aggregiert zur Verfügung gestellt. Lose nach Matos et al. (1998).

- Handelt es sich um eine Fall-Kontroll (case-control) Studie.
- Identifizieren Sie abhängige und unabhängige Variable. Wie gross ist das Risiko an Lungenkrebs zu erkranken für rauchende und nicht rauchende Menschen? Können hier die Risikodifferenz, das  $RR$  und der  $OR$  verwendet werden, um den Zusammenhang zu beschreiben?
- Wie gross ist das Chancenverhältnis und was heisst das für die rauchenden Menschen?
- Ist das Lungenkrebsrisiko vom Rauchverhalten abhängig? Führen Sie einen  $\chi^2$ -Test durch.

## Kreuztabellen

		training		
depression		oft	selten	Insgesamt
depressiv	Beobachtet	22	25	47
	Erwartet	27.5	19.5	47.0
	% der Spalte	48.9 %	78.1 %	61.0 %
nicht_depressiv	Beobachtet	23	7	30
	Erwartet	17.5	12.5	30.0
	% der Spalte	51.1 %	21.9 %	39.0 %
Insgesamt	Beobachtet	45	32	77
	Erwartet	45	32	77
	% der Spalte	100.0 %	100.0 %	100.0 %
$\chi^2$ -Tests				
		Wert	df	p
$\chi^2$		6.72	1	0.00953
$\chi^2$ mit Kontinuitätskorrektur		5.55	1	0.01850
N		77		

## Verglichene Maße

	Wert
Differenz von 2 Anteilen	-0.292*
Odds-Ratio	0.268
Relatives Risiko	0.626*

\* Verglichene Spalten

Abbildung 9.11: Jamovi Ausgabe.

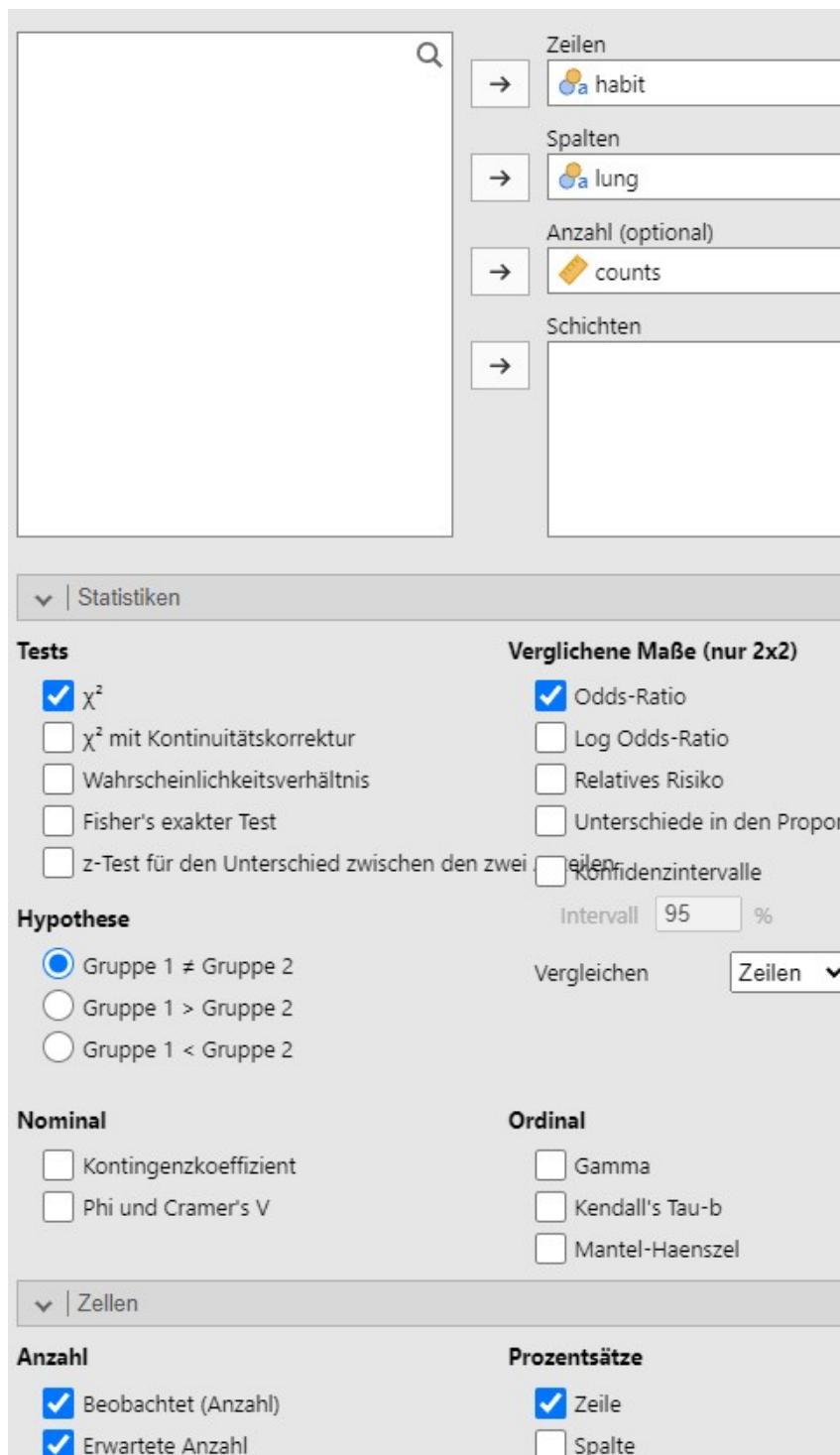


Abbildung 9.12: Jamovi Eingabe.

*Lösung.* Zuerst wird der Datensatz mit Jamovi eingelesen und die Analyseparameter werden gesetzt, siehe Abbildung 9.12.

Dies produziert das Analyseergebnis in Abbildung 9.13.

		lung		
		cancer	no_cancer	Insgesamt
habit				
smoking	Beobachtet	150.0	100	250
	Erwartet	83.8	166	250
	% der Zeile	60.0 %	40.0 %	100.0 %
no_smoking	Beobachtet	50.0	297	347
	Erwartet	116.2	231	347
	% der Zeile	14.4 %	85.6 %	100.0 %
Insgesamt	Beobachtet	200.0	397	597
	Erwartet	200.0	397	597
	% der Zeile	33.5 %	66.5 %	100.0 %

$\chi^2$ -Tests			
	Wert	df	p
$\chi^2$	136	1	<.00001
N	597		

Verglichene Maße	
	Wert
Odds-Ratio	8.91

Abbildung 9.13: Jamovi Ausgabe.

Damit kann die Frage nun beantwortet werden:

- a) Ja. Die Forschenden haben bestimmt wie viele erkrankte und gesunde Personen sie für die Studie anschreiben.

- b) Die abhängige Variable oder Risikovariable ist hier die Erkrankung an Lungenkrebs. Die ursächliche oder unabhängige Variable ist das Rauchverhalten. Da in der Studie der Anteil gesunder und erkrankter Menschen im Vorherein festgelegt wurde, ist eine Berechnung des Risikos nicht sinnvoll. Daraus folgend ist auch die Risikodifferenz oder das relative Risiko nicht sinnvoll. Der OR kann immer verwendet werden, um den Zusammenhang zu beschreiben.
- c) Das Chancenverhältnis liegt bei 8.91. Das bedeutet für die rauchenden Leute ist die Chance an Lungenkrebs zu erkanken 8.91-mal so hoch wie für nicht rauchende.
- d) Ein  $\chi^2$ -Test ergibt, dass das Auftreten von Lungenkrebs (ja/nein) und das Rauchverhalten (ja/nein) ( $OR = 8.91$ ) signifikant voneinander abhängig sind,  $\chi^2(1) = 135.57$ ,  $p < 0.001$ ,  $\phi = 0.48$ . Der Effekt ist stark.

#### **Übung 9.4.**

Eine Freundin von Ihnen behauptet, dass Sie Bio-Milch und nicht Bio-Milch am Geschmack unterscheiden kann. Sie geben ihr 9-mal Bio-Milch zu trinken und 6-mal nicht Bio-Milch zu trinken bei einem doppel-blind Test. Es entstehen die Daten in **09-exr-bio-milch.sav**.

- a) Welcher Test ist bei dieser Datenlage angebracht, um zu testen, ob die Freundin tatsächlich Bio und nicht Bio-Milch am Geschmack unterscheiden kann?
- b) Konnte die Freundin ihre Behauptung im Experiment nachweisen? Führen Sie den angebrachten Test durch und berichten Sie das Resultat inklusive Effektstärke.

*Lösung.* TODO

#### **Übung 9.5.**

Im Zusammenhang mit Covid-19 wurden Daten zu den Todesfällen publiziert (fiktive Zahlen, siehe Daten in **09-exr-covid-sterblichkeit.sav**). Wie gross sind die Risiken, die Risikodifferenz, das relative Risiko und der odds ratio und wie werden diese Grössen interpretiert? Was sind UV und AV?

*Lösung.* TODO

## **9.4 Test**



# Kapitel 10

## Zusammenhang nominalskalierter Merkmale

Bislang wurde der Zusammenhang zwischen zwei dichotomen Merkmalen angeschaut. In Beispiel 9.2 wird die Anwendung eines Herbizides (Ja/Nein) mit Lymphdrüsenerkrankung (Tumor/Kein Tumor) verglichen. Was wäre jedoch, wenn der Lymphdrüsenerkrankung mit drei Ausprägungen (Kein Tumor / gutartiger Tumor / bösartiger Tumor) erhoben wird? Im folgenden wird aufgezeigt, wie sich die bisher kennengelernten Methoden mit wenig Aufwand auf den Fall von einem oder zwei nominalskalierten Merkmalen ausweiten können.

### 10.1 Zusammenhang nominalskaliert Merkmale beschreiben

**Beispiel 10.1** (Akademischer Erfolg bei verschiedenen Lernstilen). Das Lernen wird von Pädagog:innen oft vereinfachend in Lernstile unterteilt. Eine Unterteilung ist in einen visuellen, einen auditiven und einen kinästhetischen Lernstil. Pädagog:innen wollten nur herausfinden, ob der akademische Erfolg vom Lernstil abhängt und haben dazu Pädagogik-Studierende nach der Abschlussprüfung nach ihrem Abschluss (genügend, gut und ausgezeichnet) und ihrem dominanten Lernstil befragt.

Hier sollen also zwei nominalskalierte Merkmale verglichen werden. Die Daten können wie bei der Vierfeldertafel in zwei Formaten gespeichert werden: Entweder wird eine Zeile pro Beobachtung verwendet oder die Zeilen geben die Ausprägungskombinationen zusammen mit der Anzahl Beobachtungen an. Um eine Übersicht über die erhobenen Daten zu erhalten, wird wieder eine Kreuztabelle erstellt. Für diese sogenannte **Mehrfeldertabelle** werden die Zählun-

## 192 KAPITEL 10. ZUSAMMENHANG NOMINALSKALIERTER MERKMALE

gen für jede Ausprägung der beiden Merkmale aufsummiert, siehe Abbildung 10.1. Die Mehrfeldertabelle kann in **Jamovi** unter **Analysen > Häufigkeiten > Kreuztabellen > Unabhängige Stichproben** erstellt werden.

Kreuztabellen

lernstil	abschluss			Insgesamt
	genügend	gut	ausgezeichnet	
Visuell	11	25	16	52
Auditiv	33	22	10	65
Kinästhetisch	13	7	6	26
Insgesamt	57	54	32	143

Abbildung 10.1: Mehrfeldertabelle mit absoluten Häufigkeiten der Daten Lernstil und Abschlusserfolg.

Die Zellen der Mehrfeldertabellen werden allgemein mit einem Zeilenindex  $i$  und einem Spaltenindex  $j$  bezeichnet. Die Zelle  $i = 1$  und  $j = 3$  enthält also Informationen zu Personen mit ausgezeichnetem Abschluss bei visuellem Lernstil. Die Anzahl Beobachtungen in der Zelle werden mit  $o_{ij}$  für *observed* bezeichnet, zum Beispiel  $o_{13} = 16$ . Der Zeilenindex geht von 1 bis  $I$  und der Zeilenindex von 1 bis  $J$ . Da beide Merkmale im Beispiel genau 3 Ausprägungen habe ist hier  $I$  und  $J$  genau 3.

Trotz der verbesserten Übersicht ist es aufgrund der verschiedenen Randhäufigkeiten schwer Auffälligkeiten in der Mehrfeldertabelle zu erkennen. Um dies zu vereinfachen, können der Tabelle die relativen Häufigkeiten (bezüglich Zeile oder Spalte) hinzugefügt werden, siehe Abbildung 10.2. Dies wird in **Jamovi** berechnet indem unter **Zellen > Prozentsätze** die Optionen **Zeile** und **Spalte** angewählt werden.

Es kann zum Beispiel festgestellt werden, dass beim visuellen Lernstil mit 31% ein viel höherer Anteil einen ausgezeichneten Abschluss macht, als bei den Lernstilen auditiv mit 15% oder kinästhetisch mit 23%. Die Frage ist nun, ob dieser Zusammenhang zwischen Lernstil und Abschluss auf die Zufallsstichprobeneziehung zurückzuführen ist oder ob man davon ausgehen kann, dass der Zusammenhang auch in der Population (also allen Pädagogikstudierende) übertragen lässt.

10.1. ZUSAMMENHANG NOMINALSKALIERTER MERKMALE BESCHREIBEN 193

Kreuztabellen

		abschluss			
lernstil		genügend	gut	ausgezeichnet	Insgesamt
Visuell	Beobachtet	11	25	16	52
	% der Zeile	21 %	48 %	31 %	100 %
	% der Spalte	19 %	46 %	50 %	36 %
Auditiv	Beobachtet	33	22	10	65
	% der Zeile	51 %	34 %	15 %	100 %
	% der Spalte	58 %	41 %	31 %	45 %
Kinästhetisch	Beobachtet	13	7	6	26
	% der Zeile	50 %	27 %	23 %	100 %
	% der Spalte	23 %	13 %	19 %	18 %
Insgesamt	Beobachtet	57	54	32	143
	% der Zeile	40 %	38 %	22 %	100 %
	% der Spalte	100 %	100 %	100 %	100 %

Abbildung 10.2: Mehrfeldertabelle mit absoluten und relativen Häufigkeiten der Daten Lernstil und Abschlusserfolg.

## 10.2 Zusammenhang nominalskalierter Merkmale testen

Um dies zu testen, kann der im letzten Kapitel besprochene  $\chi^2$ -Test für dichotome Merkmale erweitert werden. Die Nullhypothese ist, dass die beiden nominalskalierten Merkmale, hier der Lernstil und der Abschluss, unabhängig voneinander sind. Der Test funktioniert nun genau gleich wie der Vierfeldertest, indem erst für jede Zelle  $ij$  eine unter der Annahme der Unabhängigkeit der zwei Merkmale erwartete Anzahl Beobachtungen  $e_{ij}$  berechnet wird. Die erwartete Anzahl Beobachtungen kann in Jamovi unter Zelle > Anzahl > Erwartete Anzahl dazugeschaltet werden, siehe Abbildung 10.3.

Kreuztabellen

		abschluss			
		genügend	gut	ausgezeichnet	Insgesamt
lernstil					
Visuell	Beobachtet	11	25	16	52
	Erwartet	20.73	19.64	11.64	52.00
Auditiv	Beobachtet	33	22	10	65
	Erwartet	25.91	24.55	14.55	65.00
Kinästhetisch	Beobachtet	13	7	6	26
	Erwartet	10.36	9.82	5.82	26.00
Insgesamt	Beobachtet	57	54	32	143
	Erwartet	57	54	32	143

Abbildung 10.3: Mehrfeldertabelle mit beobachteten und unter Unabhängigkeit erwarteten Häufigkeiten der Daten Lernstil und Abschlusserfolg.

Je weiter diese Zahl von der beobachteten Zahl abweicht, desto unwahrscheinlicher ist die Unabhängigkeit der beiden Merkmale. Es werden zum Beispiel  $o_{11} = 11$  genügende Abschlüsse bei visuellem Lernstil beobachtet. Wenn Lernstil und Abschluss unabhängig wären, würden für diese Kombination  $e_{11} = 20.73$  Beobachtungen erwartet.

Der  $\chi^2$ -Mehrfeldertest trägt dem Rechnung, indem die Teststatistik diese Differenz der erwarteten und beobachteten Anzahl austariert und für jede Zelle aufsummiert mit

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \stackrel{Bsp}{=} 12.78.$$

Da die beiden Merkmale nominalskaliert sind, kann  $I$  und  $J$  im Gegensatz zu Kapitel 9 auch einen Wert grösser als 2 annehmen. Eine grosse Teststatistik spricht demnach wieder gegen die Unabhängigkeit.

Wenn die Stichprobenziehung oft wiederholt wird, kann festgestellt werden, dass diese Teststatistik einer  $\chi^2$ -Verteilung bei

$$df = (I - 1) \cdot (J - 1) = (3 - 1) \cdot (3 - 1) = 2 \cdot 2 = 4$$

Freiheitsgraden folgt, siehe Abbildung 9.5. Die Teststatistik des  $\chi^2$ -Mehrfeldertests 12.78 wird also mit der grünen Verteilung in der Abbildung verglichen. Die Werte rechts auf der Abbildung sind seltener und der beobachtete Wert der Teststatistik liegt so, dass er zu den  $p = .012$  seltensten Beobachtungen zählt, sofern die Unabhängigkeit gilt. Der genaue  $p$ -Wert, kann mit Jamovi unter Statistiken > Tests > \$\chi^2\$ bestimmt werden. Da der  $p$ -Wert kleiner als 5% liegt, kann hier die Nullhypothese bei Signifikanzniveau  $\alpha = 5\%$  verworfen werden.

Bei der  $\chi^2$ -Verteilung handelt es sich wieder um eine Annäherung der tatsächlichen Verteilung der Teststatistik. Diese Annäherung ist nur gut, wenn die Anzahl erwartete Beobachtungen in jeder Zelle mindestens 5 beträgt. Ist dies nicht gegeben, kann entweder auch im Mehrfelderkontext Fishers exakter Test verwendet werden - wie letzterer genau erweitert wird, wird hier nicht behandelt - oder es können Ausprägungen zusammengefasst werden. Im Beispiel könnte zum Beispiel die Abschlussbewertung als neues Merkmal mit den zwei Ausprägungen *genügend* und *mehr-als-genügend* betrachtet werden.

Um die Effektstärke des  $\chi^2$ -Tests anzugeben kann Cramérs **V**

$$\phi = \sqrt{\frac{\chi^2}{n \cdot (k - 1)}} = \sqrt{\frac{12.78}{143 \cdot (3 - 1)}} = 0.21$$

verwendet werden. Hier stellt  $k$  die kleinere der beiden Dimensionen  $I$  und  $J$  der Mehrfeldertabelle dar. Da im Beispiel  $I$  und  $J$  gleich gross sind können beide als  $k$  verwendet werden. Also ist  $k = 3$  im Beispiel. Cramérs **V** ist immer grösser als 0 und kleiner als 1. Je weiter **V** von 0 weg ist, desto stärker sind die Merkmale voneinander abhängig. Die Interpretation als Effektstärke erfolgt dabei wie für eine Korrelation mit der Einschränkung, dass hier keine Richtung des Zusammenhangs interpretiert werden kann. Der Wert von 0.42 wird demnach als stark eingestuft. Cramérs **V** ist in Jamovi unter Statistiken > Nominal > Phi und Cramer's **V** zu finden.

Für eine Vierfeldertafel, also  $I = 2, J = 2$  entspricht Cramérs **V** genau Cramérs  $\phi$ .

Eine Alternative zu Cramérs **V** ist der Kontingenzkoeffizient nach Pearson

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{12.78}{12.78 + 143}} = 0.29,$$

wobei  $n$  die Gesamtzahl Beobachtungen ist. Es gilt  $0 < C < \sqrt{(k-1)/k}$ . Der Kontingenzkoeffizient wird also je nach Dimension auf einer unterschiedlichen Skala gemessen, was den intuitiven Vergleich der Werte für verschiedene Anwendungen erschwert. Korrekturmethoden existieren, sind aber weder in **Jamovi** implementiert noch werden sie oft verwendet. Es wird deshalb empfohlen immer Cramérs  $V$  zu verwenden.

Das Testergebnis wird schliesslich wie folgt berichtet:

Ein  $\chi^2$ -Test ergibt, dass der Abschluss (genügend/gut/ausgezeichnet) und der Lernstil (visuell/auditiv/kinästhetisch) signifikant voneinander abhängig sind,  $\chi^2(4) = 12.78, p = .012, V = 0.21$ . Der Zusammenhang ist als mittel einzustufen.

#### Achtung



*Hinweis.* Für die Mehrfeldertabelle gibt es keine standardmässig verwendeten Grössen wie das relative Risiko oder das Chancenverhältnis. Um den Lesenden eine Datenübersicht zu präsentieren, kann in einer Arbeit zusätzlich zum Berichtensatz die Mehrfeldertabelle mit den absoluten oder und den relativen Häufigkeiten dargestellt werden.

## 10.3 Übungen

## 10.4 Test

# Begriffsverzeichnis

- Ablehnungsbereich
- Alternativhypothese
- arithmetische Mittel
- Cohens  $d$  für den Welch-Test
- Cohens  $d$  für Zweistichproben-t-Test
- Cramérs  $\phi$
- einseitige Hypothese
- Einstichproben-t-Test
- Erwartungswert
- Freiheitsgrade
- Glass  $\Delta$
- Grundgesamtheit
- Hedges  $g$
- Hypothese
- Interquartilabstand
- Intervalls
- intervallskaliert
- Irrtumswahrscheinlichkeit
- Konfidenzintervall
- Median
- Modus
- nicht signifikant
- Normalverteilung
- Nullhypothese
- Perzentil
- Population
- Signifikanzniveau
- Spannweite
- Standardabweichung
- Stichprobe
- Stichprobenziehung
- Student- $t$ -Verteilung
- Teststatistik
- Vertrauenswahrscheinlichkeit

## 198 KAPITEL 10. ZUSAMMENHANG NOMINALSKALIERTER MERKMALE

- Welch-Test
- Zentraler Grenzwertsatz
- Zufallsstichprobe
- zweiseitige Hypothese
- Zweistichproben- $t$ -Test

# **Literaturverzeichnis**

200 KAPITEL 10. ZUSAMMENHANG NOMINALSKALIERTER MERKMALE

# Literaturverzeichnis

- Balci, S. (2025). *jjstatsplot: Wrapper for ggstatsplot.* R package version 0.0.2.54, <https://github.com/sbalci/jjstatsplot>, <https://sbalci.github.io/ClinicoPathJamoviModule/>, <https://www.serdarbalci.com/jjstatsplot/>.
- Bandura, A., Ross, D., and Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63(3):575–582.
- Beck, A. T., Steer, R. A., and Carbin, M. G. (1988). Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. 8(1):77–100.
- Blaney, N. T., Stephan, C., Rosenfield, D., Aronson, E., and Sikes, J. (1977). Interdependence in the classroom: A field study. *Journal of Educational Psychology*, 69(2):121–128.
- Blachnio, A., Przepiorka, A., and Pantic, I. (2016). Association between face-book addiction, self-esteem and life satisfaction: A cross-sectional study. 55:701–705.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition.
- Czeisler, C., Duffy, J., Shanahan, T., Brown, E., Mitchell, J., Rimmer, D., Ronda, J., Silva, E., Allan, J., Emens, J., Dijk, D.-J., and Kronauer, R. (1999). Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science*, 284:2177–2181.
- Deutsche Gesellschaft für Psychologie (2019). *Richtlinien zur Manuskriptgestaltung*. Hogrefe Verlag, Göttingen, auflage: 5., aktualisierte auflage edition.
- Dolan, C., Oort, F., Stoel, R., and Wicherts, J. (2009). Testing measurement invariance in the target rotated multi-group exploratory factor model. 16(2):295–314.
- Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

- Hamilton, J. B. (1951). Patterned loss of hair in man: Types and incidence. 53(3):708–728. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1951.tb31971.x>.
- Hayes, H. M., Tarone, R. E., Cantor, K. P., Jessen, C. R., McCurnin, D. M., and Richardson, R. C. (1991). Case-control study of canine malignant lymphoma: Positive association with dog owner's use of 2, 4-dichlorophenoxyacetic acid herbicides. 83(17):1226–1231.
- Khanna, M. M., Badura-Brack, A. S., McDermott, T. J., Embury, C. M., Wiesman, A. I., Shepherd, A., Ryan, T. J., Heinrichs-Graham, E., and Wilson, T. W. (2017). Veterans with post-traumatic stress disorder exhibit altered emotional processing and attentional control during an emotional stroop task. *Psychological Medicine*, 47(11):2017–2027.
- Macleod, C., Williams, J., and Mathews, A. (1996). The emotional stroop task and psychopathology. *Psychological Bulletin*, 120(1):3–24.
- Matos, E., Vilensky, M., Boffetta, P., and Kogevinas, M. (1998). Lung cancer and smoking: a case-control study in buenos aires, argentina. 21(3):155–163.
- McGahuey, C. A., Gelenberg, A. J., Laukes, C. A., Moreno, F. A., Delgado, P. L., McKnight, K. M., and Manber, R. (2016). Arizona sexual experience scale. Institution: American Psychological Association.
- Morris, C., Reiber, C., and Roman, E. (2015). Quantitative sex differences in response to the dissolution of a romantic relationship. *Evolutionary Behavioral Sciences*, 9.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Robinson, K. C., Kemény, L. V., Fell, G. L., Hermann, A. L., Allouche, J., Ding, W., Yekkirala, A., Hsiao, J. J., Su, M. Y., Theodosakis, N., Kozak, G., Takeuchi, Y., Shen, S., Berenyi, A., Mao, J., Woolf, C. J., and Fisher, D. E. (2021). Reduced MC4r signaling alters nociceptive thresholds associated with red hair. 7(14):eabd1310.
- Sandstrom, G. M., Boothby, E. J., and Cooney, G. (2022). Talking to strangers: A week-long intervention reduces psychological barriers to social connection. 102:104356.
- Schuckit, M. A. (1996). Alcohol, anxiety, and depressive disorders. 20(2):81–85.
- Selker, R., Love, J., and Dropmann, D. (2024). *jmv: The jamovi Analyses*. R package version 2.5.6.
- Skinner, H. A. and Allen, B. A. (1982). Alcohol dependence syndrome: Measurement and validation. 91(3):199–209. Place: US Publisher: American Psychological Association.

- Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., and Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press, Palo Alto, CA.
- Tas, B., Kulacaoglu, F., Belli, H., and Altuntas, M. (2018). The tendency towards the development of psychosexual disorders in androgenetic alopecia according to the different stages of hair loss: a cross-sectional study. 93(2):185–190.
- Viscidi, E. W., Triche, E. W., Pescosolido, M. F., McLean, R. L., Joseph, R. M., Spence, S. J., and Morrow, E. M. (2013). Clinical characteristics of children with autism spectrum disorder and co-occurring epilepsy. 8(7):e67797. Publisher: Public Library of Science.
- Welch, B. L. (1947). The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.
- Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida.
- Yusefzadeh, H., Amirzadeh Iranagh, J., and Nabilou, B. (2019). The effect of study preparation on test anxiety and performance: a quasi-experimental study. 10:245–251.
- Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. 57(1):173–181. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1348/000711004849222>.