

Statistik 1

Daniel J. F. Gerber

15 December, 2024

Inhaltsverzeichnis

Vorwort	5
1 Einleitung	7
1.1 Worum geht es?	7
1.2 Inhaltlicher Aufbau	7
1.3 Wie soll ich dieses Buch lesen?	8
1.4 Formeln, Symbole und Zahlen	9
1.5 Software	9
I Eine intervallskaliertes Merkmal	11
2 Intervallskalierte Merkmale	13
2.1 Was ist ein intervallskaliertes Merkmal?	13
2.2 Wie kann ein intervallskaliertes Merkmal beschrieben werden? . .	14
2.3 Übungen	17
3 Stichprobenziehung	25
3.1 Was ist das Problem der Stichprobenziehung?	26
3.2 Wie kann man Aussagen über die Grundgesamtheit machen? . .	28
3.3 Übungen	29
4 Durchschnitt und Standardabweichung schätzen	31
4.1 Wo liegt der Durchschnitt der Grundgesamtheit?	31
4.2 Übungen	39

5	Zentrale Tendenz testen	45
5.1	Entspricht der Erwartungswert einem gewissen Wert?	46
5.2	Weicht der gefundene Durchschnitt stark vom hypothetischen Wert ab?	55
5.3	Übungen	57
II	Gruppenvergleich einer intervallskalierten Variable	65
6	Gruppenvergleich einer intervallskalierten Variable	67
6.1	Zwei Gruppen vergleichen	67
6.2	Was ist das Problem der Stichprobenziehung?	67
6.3	Wie kann man Aussagen über die Grundgesamtheit machen? . .	67
6.4	Übungen	67
7	Welch-Test	69
7.1	Zwei Gruppen vergleichen	69
7.2	Sind die Durchschnitte der beiden Gruppen in der Grundgesamtheit gleich?	69
7.3	Wie stark unterscheiden sich die Durchschnitte?	69
7.4	Übungen	69
	Begriffsverzeichnis	71

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 ρ
- Vierfelderkorrelation / Φ
- Punktbiseriale Korrelation
- Kontingenzkoeffizient
- Cramér's 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.

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. Dem Leser wird deshalb empfohlen diese Software 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°C und 42°C . Es ist möglich zu sagen, dass eine Person mit 40°C Körpertemperatur eine höhere Temperatur hat als eine mit 38°C Körpertemperatur. Ausserdem sind alle erdenklichen Zwischenwerte möglich, so auch dass bei einer Person eine Körpertemperatur von $37.821239^{\circ}\text{C}$ 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.

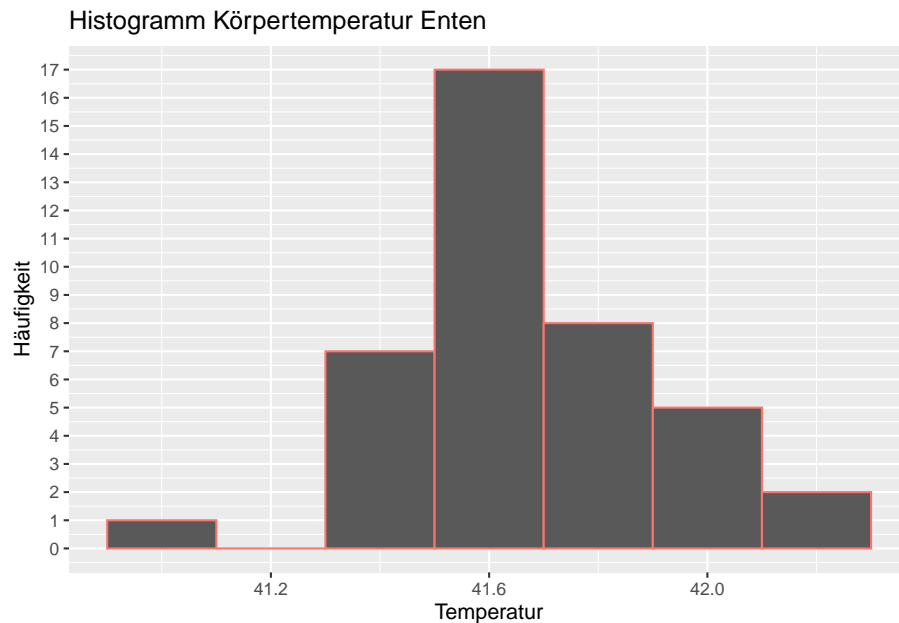
2.2 Wie kann ein intervallskaliertes Merkmal beschrieben werden?

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°C betrachten und zählen wie viele Beobachtungen sie in den jeweiligen Abschnitten gemacht hat. Diese Zählzeiten können tabellarisch oder grafisch mit einem Balkendiagramm dargestellt werden. Letzteres wird ein **Histogramm** genannt.



Aufgrund dieser Darstellung kann die Veterinärin nun sehen, wie häufig welche Körpertemperaturen sind. Dies wird die **Verteilung** des Merkmals genannt.

2.2. WIE KANN EIN INTERVALLSKALIERTES MERKMAL BESCHRIEBEN WERDEN?15

Sie bemerkt zum Beispiel, dass Beobachtungen der Körpertemperatur rund um 41.6°C am häufigsten sind und tiefere und höhere Temperaturen seltener vorkommen. Auf einen Blick sieht sie auch, dass die Temperatur aller Enten zwischen 41°C und 42.2°C war.

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

2.2.2 Zentralität

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

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

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

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°C und der kleinste Beobachtete Wert 41.07°C . Also ist die Spannweite $42.18 - 41.07 = 1.11^\circ\text{C}$. Die Spannweite wird in Jamovi mit **Wertebereich** bezeichnet.
- Wenn die Werte des Merkmals aufsteigend sortiert werden und der Wert betrachtet wird, welcher die Beobachtungen in eine $P\%$ tiefere und $(100\% - P\%)$ höhere Hälfte teilt, dann wird dieser Wert als **Perzentil** bezeichnet. Das 5%-Perzentil zum Beispiel teilt die beobachteten Werte in 5% kleinere und 95% grössere Werte. Im Beispiel haben wir 40 Beobachtungen. 5% davon sind demnach 2 Beobachtungen die tiefer sind als das 5% Perzentil und 95% also 38 Beobachtungen die höher sind als das 5% Perzentil. Das 5% Perzentil liegt also zwischen 41.32°C und 41.35°C . In diesem Fall wird ein Mittelwert der beiden nächsten Werte verwendet, hier $(41.32 + 41.35)/2 = 41.34^\circ\text{C}$. Das $P\%$ -Perzentil kann in Jamovi bei **Perzentil** gefolgt von der Zahl P ermittelt werden. Ein Perzentil alleine gibt jedoch noch keinen Hinweis auf die Streuung der Werte. Werden aber zwei Perzentile zusammen betrachtet, z. B. das 5% und das 95% Perzentil, dann geben diese Werte und der Abstand dazwischen einen Hinweis auf die Streuung der Beobachtungen. Im Beispiel ist das 5% Perzentil bei 41.34°C und das 95%-Perzentil bei 42.1°C . Hier befinden sich also 90% aller Beobachtungen zwischen diesen Werten. Mehrere Perzentile können in Jamovi gleichzeitig angezeigt werden indem die Perzentil-Werte mit Komma getrennt werden, für die Perzentile hier

im Beispiel 0.05, 0.95. Weitere beliebte Werte sind das 25% und das 75%-Perzentil (auch **Quartile** genannt, da sie die beobachteten Werte vierteln), im Beispiel bei 41.52°C und 41.82°C respektive. Die Differenz dieser Perzentile wird als **Interquartilabstand** (Abkürzung IQR von interquartile range) bezeichnet und ist im Beispiel 0.3°C . Der Interquartilabstand wird in Jamovi mit **IQR** bezeichnet.

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

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

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

Achtung



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

2.3 Übungen

Übung 2.1.

Mit den Daten `02-exm-ducktemp.sav` aus Beispiel 2.3:

- Erstellen Sie selbst ein Histogramm mit Jamovi und begründen Sie, weshalb es nicht gleich aussieht wie das Histogramm oben.

- (b) Berechnen Sie zusätzlich Modus, Median, arithmetisches Mittel, IQR, 25%- und 75%-Perzentil, sowie 2.5%- und 97.5%-Perzentil, sowie die Spannweite und die Standardabweichung der Körpertemperaturen der Enten mit Jamovi.

Lösung.

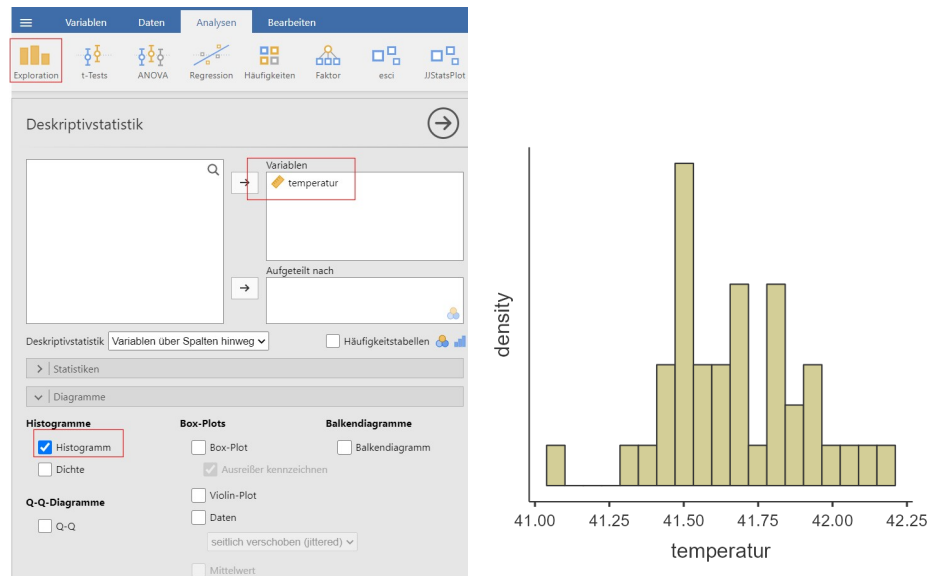


Abbildung 2.1: Links: Jamovi-Anleitung zur Erstellung des Histogramms; rechts: Histogramm der Temperatur.

- (a) Das Histogramm, siehe Abbildung 2.1 sieht nicht gleich aus, da Jamovi die Temperaturabschnitte kürzer gewählt hat nämlich bei 0.125°C statt 0.2°C wie oben im Text. In Jamovi gibt es aktuell keine Möglichkeit die Abschnittsweite anzupassen. Ein Histogramm sieht immer anders aus je nach ausgewählter Abschnittsweite.
- (b) Eine Anleitung zur Berechnung in Jamovi sowie die berechneten Werte können in Abbildung 2.2 abgelesen werden.

Übung 2.2.

TODO: Exercise body

Lösung. TODO: solution body

Übung 2.3.

Deskriptivstatistik	
	temperatur
N	40
Fehlend	0
Mittelwert	41,7
Median	41,7
Modalwert	41,5
Standardabweichung	0,233
IQR	0,300
Wertebereich	1,11
Minimum	41,1
Maximum	42,2
2,5. Perzentil	41,3
25. Perzentil	41,5
75. Perzentil	41,8
97,5. Perzentil	42,1

Abbildung 2.2: Links: Jamovi-Anleitung zur Berechnung der gewünschten Parameter; rechts: Parameterwerte.

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

- (a) Wie gross ist das arithmetische Mittel und der Median dieser Werte?
- (b) Nehme an, der Testleiter hat den Wert der ersten Probandin falsch in seine Tabelle übertragen - statt 18 hat er 81 geschrieben. Wie gross ist das arithmetische Mittel und der Median dieser Werte in diesem Fall?
- (c) Was sagt dies über den Median und das arithmetische Mittel aus?

Lösung. Die Aufgabe kann im Kopf gelöst werden, oder mithilfe eines Taschenrechners, oder indem die Zahlen manuell bei Jamovi eingegeben werden.

- (a) Wir haben hier $n = 5$ Beobachtungen, nämlich $x_1 = 18, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Wird dies in die Formel (2.1) eingesetzt, so gibt dies das arithmetische Mittel

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

Um den Median zu berechnen, werden die Werte zuerst aufsteigend sortiert 18, 19, 20, 21, 22. Der Wert, welcher die Werte in eine grössere und eine kleinere Hälfte teilt, ist hier 20, was dem Median entspricht.

- (b) Die Beobachtungen sind jetzt $x_1 = 81, x_2 = 21, x_3 = 20, x_4 = 19, x_5 = 22$. Analog wie in (a) kann demnach das arithmetische Mittel als $\bar{x} = 32.6$ bestimmt werden. Die aufsteigend sortierten Beobachtungen sind nun 19, 20, 21, 22, 81. Der Median ist also 21.
- (c) Durch die fälschliche Übertragung eines Wertes, ist das arithmetische Mittel sehr stark und der Median fast gar nicht beeinflusst worden. 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.

TODO: Exercise body

Lösung. TODO: solution body

Übung 2.5.

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
- **Geburtszeit_std_ab_mitternacht**: Geburtszeit in Stunden ab Mitternacht. Wenn jemand um 13h30 auf die Welt kam, ist dieser Wert 13.5.
- **Geburtszeit_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 2.3 analysiert.

- a. Es gibt gemäss 2.4 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 angegebene 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

Deskriptivstatistik

→

Variablen
☒ IQ
☒ Aufgeschlossenheit
☒ Wartezeit_min
☒ Wartezeit_std
☒ Geburtszeit_std_ab_mitternacht

→

Aufgeteilt nach

Deskriptivstatistik Variablen über Zeilen hinweg ☐ Häufigkeitstabellen

▼ | Statistiken

Stichprobenumfang
☒ N ☒ Fehlend

Zentrale Tendenz
☒ Mittelwert
☒ Median
☒ Modalwert
☐ Summe

Perzentilwerte
☐ Grenzwerte (cut point) für gleichgroße Gruppen
☐ Perzentile

Streuung
☒ Std.-abweichung ☐ Minimum
☐ Varianz ☐ Maximum
☒ Wertebereich ☒ IQR

Verteilung
☐ Schiefe
☐ Kurtosis

Annahme von Normalverteilung
☐ Shapiro-Wilk

Mittlere Streuung
☐ Std.-fehler des Mittelwerts
☐ Konfidenzintervall für den Mittelwert %

Ausreißer
☐ Äußerste Werte

▼ | Diagramme

Histogramme
☒ Histogramm

Box-Plots
☐ Box-Plot

Balkendiagramme
☐ Balkendiagramm

Abbildung 2.3: 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
Geburtzeit_std_ab_mitternacht	500	0	11.8111	12.1759	0.1157 *	6.9384	11.8936	23.758
Geburtzeit_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 2.4: Deskriptive Statistiken.

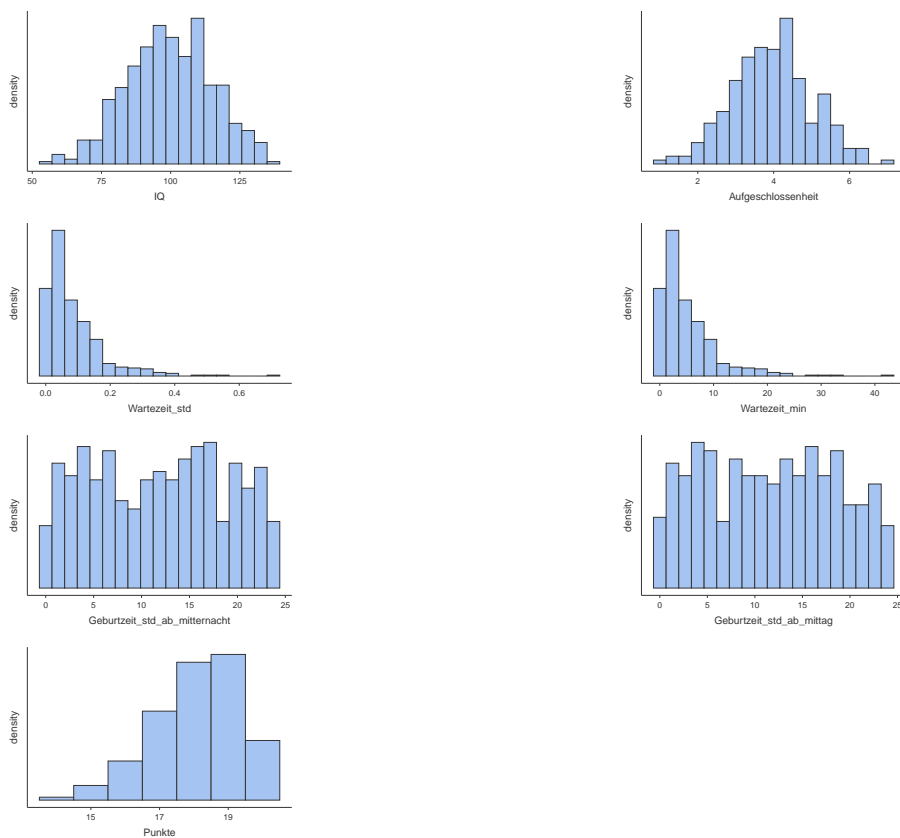


Abbildung 2.5: Histogramme.

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 Minuten zu betrachten. Die durchschnittliche Wartezeit liegt bei $M = 5.07, Mdn = 3.97$ Minuten. 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 3.97 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.

Kapitel 3

Stichprobenziehung

Beispiel 3.1 (Angst). Forschende haben ein Messinstrument State-Trait Anxiety Inventory *STAI*, welches Angst misst (Spielberger et al., 1983). Sie unterscheiden dabei zwischen Zustandesangst und dem Persönlichkeitszug Ängslichkeit. Hier interessiert uns nur die Zustandesangst, welche fortan Angst genannt wird und misst wie grosse Angst man aktuell empfindet. 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. Nach der Auswahl der 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 Forschungsergebnisse 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 eine repräsentative Stichprobe würde den Rahmen dieses Buches sprengen.

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. Die Personen im Zimmer, respektive die Beobachtungen in der Population sind normalerweise nicht sichtbar. Aus diesem Zimmer wurden also 30 Personen geholt und befragt also sichtbar gemacht, was der Zufallsstichprobe entspricht. Die Zufallsstichprobe ist gekennzeichnet durch die Punkte über dem Zimmer, oberhalb des Pfeils. Die Farben der Punkte sind jetzt bekannt und entsprechen der jeweiligen Zustandesangst der beobachteten Personen.

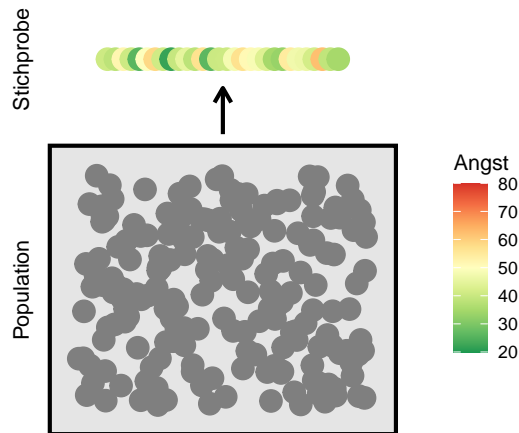


Abbildung 3.1: Population mit unbekannter Angst.

Da die Stichprobe nun eben 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 Zustandsangst, 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, wie die darunterliegende Population aussieht. Sind die Werte der Stichprobe

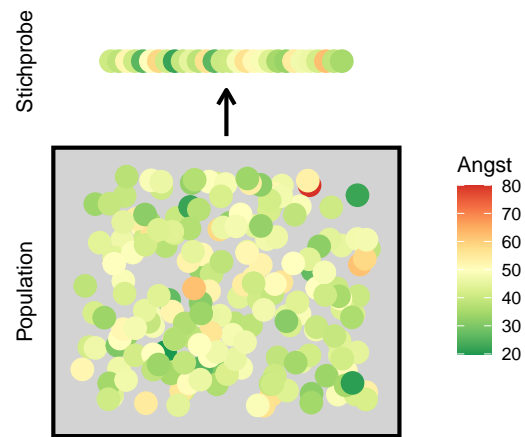


Abbildung 3.2: Population mit ähnlichen Angst-Werten wie in der Stichprobe.

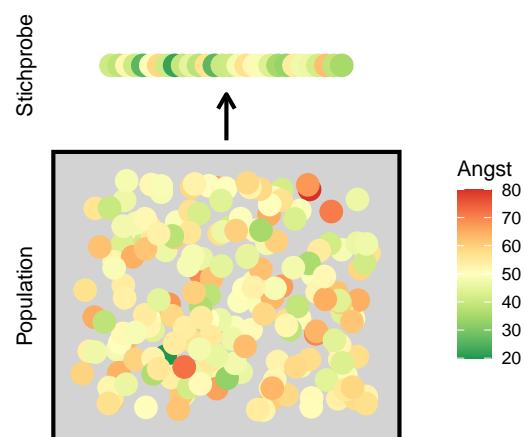


Abbildung 3.3: Population mit höheren Angst-Werten als in der Stichprobe.

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.

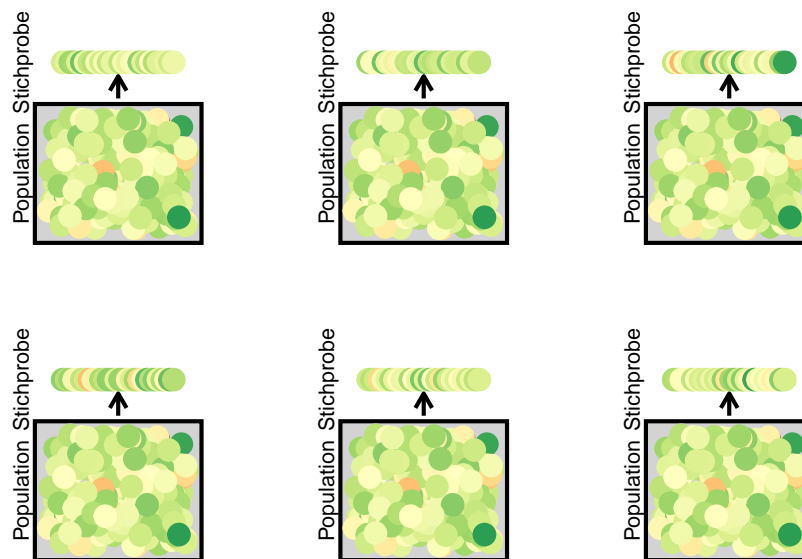


Abbildung 3.4: TODO.

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 einer Population mit eher höheren Angst-Werten. Ganz sicher kann man jedoch

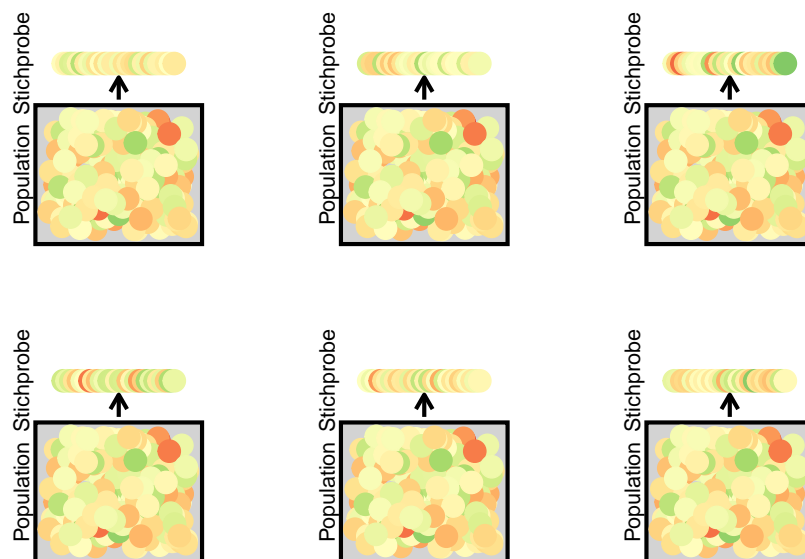


Abbildung 3.5: TODO.

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.

TODO: Exercise body

Lösung. TODO: solution body

Übung 3.2.

TODO: Exercise body

Lösung. TODO: solution body

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 μ [gr.: mü]) genannt. Wenn das arithmetische Mittel der Stichprobe berechnet wird, ergibt dies auch ein Schätzwert für besagten Erwartungswert. Aufgrund der zufälligen Stichprobenziehung ist jedoch auch klar, dass dieser Schätzwert nie genau dem wahren Erwartungswert entspricht.

In Beispiel 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.

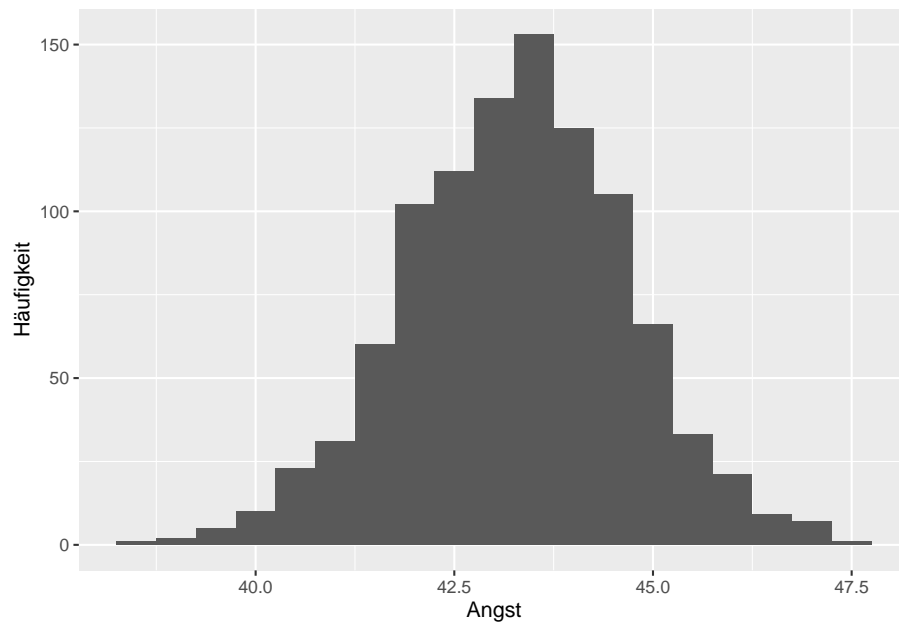


Abbildung 4.1: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

Die Häufigkeitsverteilung der berechneten arithmetischen Mittel in Abbildung 4.1 lässt nun Aussage über die Häufigkeit und damit über die Wahrscheinlichkeit von gewissen Werten als Erwartungswert zu. Ein Durchschnittswert der Zustandesangst um die 30 ist hier am wahrscheinlichsten und ein Wert tiefer als 27 oder höher 33 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 $M = 3.91$, $s = 1.73$.

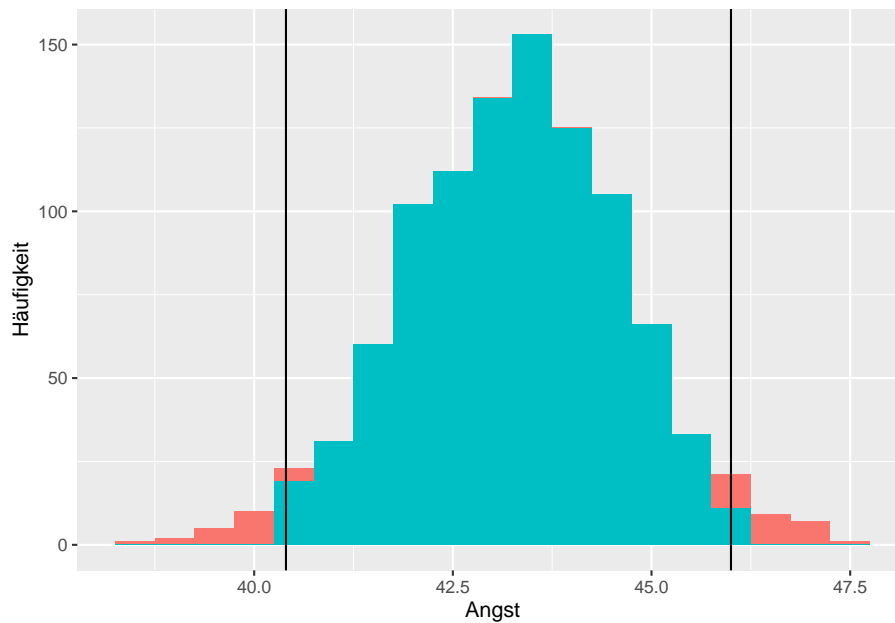


Abbildung 4.2: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Angst.

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 Gausssche Glockenkurve nach Carl F. Gauss (1777-1855) benannt. Die Normalverteilung kann mit nur zwei Parametern beschrieben werden.

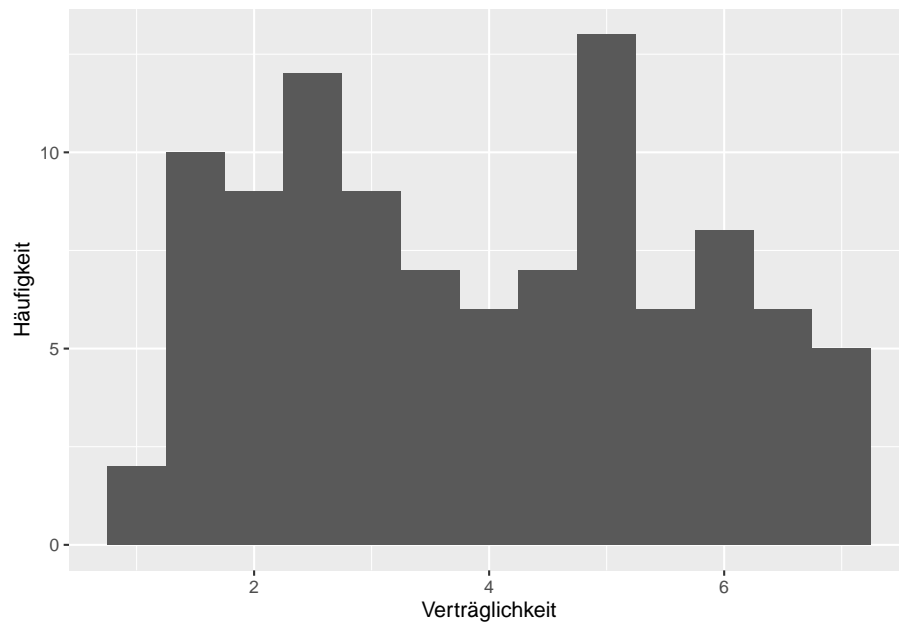


Abbildung 4.3: Verteilung der 100 beobachteten Verträglichkeitswerte einer zufällig gezogenen Stichprobe.

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

Auf seeing-theory.brown.edu > Continuous > Normal kann der Einfluss von μ und σ 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 geneauer, dass bei einem Merkmal mit Erwartungswert μ und Standardabweichung σ , der Durchschnitt aller Stichprobenwerte einer Normalverteilung mit $\mu_g = \mu$ und $\sigma_g = \frac{\sigma}{\sqrt{n}}$ entspricht, wobei n die Stichprobengrösse bezeichnet.

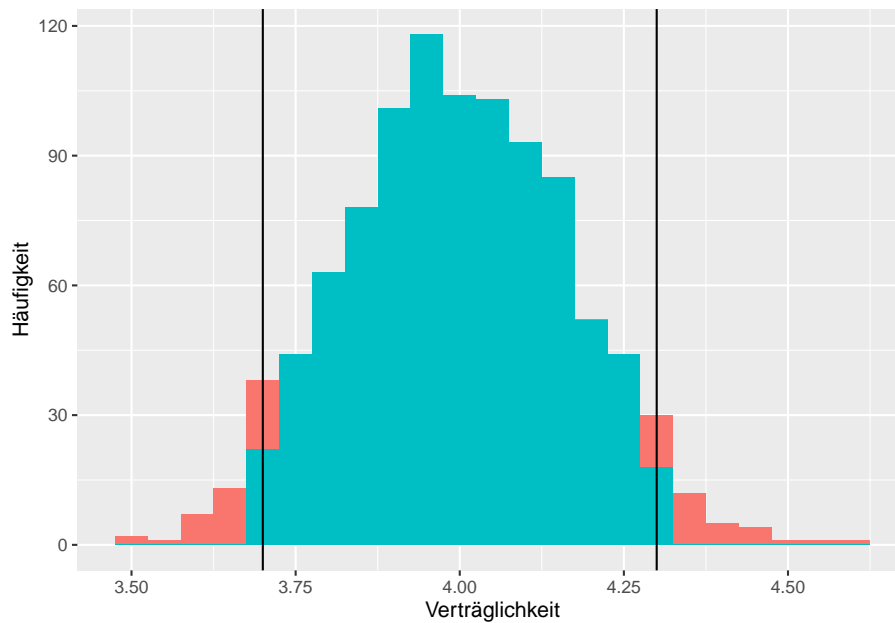


Abbildung 4.4: Verteilung der arithmetischen Mittel von 1000 zufällig gezogenen Stichproben der Verträglichkeit.

Achtung



Hinweis.

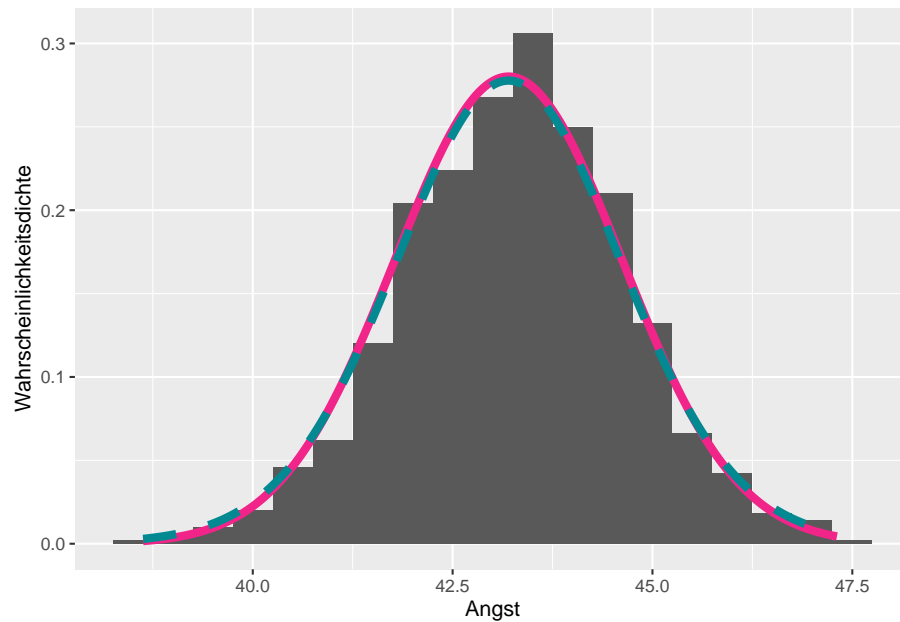
- $\mu_g = \mu$ bedeutet, dass der Wert, welcher unter der normalverteilung am wahrscheinlichsten ist, genau dem Erwartungswert des untersuchten Merkmals entspricht.
- $\sigma_g = \frac{\sigma}{\sqrt{n}}$ hat zwei Implikationen:
 - je grösser die Streuung des Merkmals (grosses σ) desto breiter ist auch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

Die Abbildungen 4.5 und 4.6 illustrieren den zentralen Grenzwertsatz für

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 μ und σ bekannt sind. Diese Annahme wird später aufgelöst und dient hier lediglich der Illustration.

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

Die Erkenntnis des zentralen Grenzwertsatz macht also das wiederholte ziehen von Stichproben unnötig. Die Normalverteilung ist theoretisch konstruiert und ihr 2.5%- und 97.5%-Perzentil können theoretisch hergeleitet werden. Tabelle 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 rührt daher, dass der zentrale Grenzwertsatz nur dann exakt funktioniert, wenn die Anzahl Beobachtungen (unendlich) gross ist.

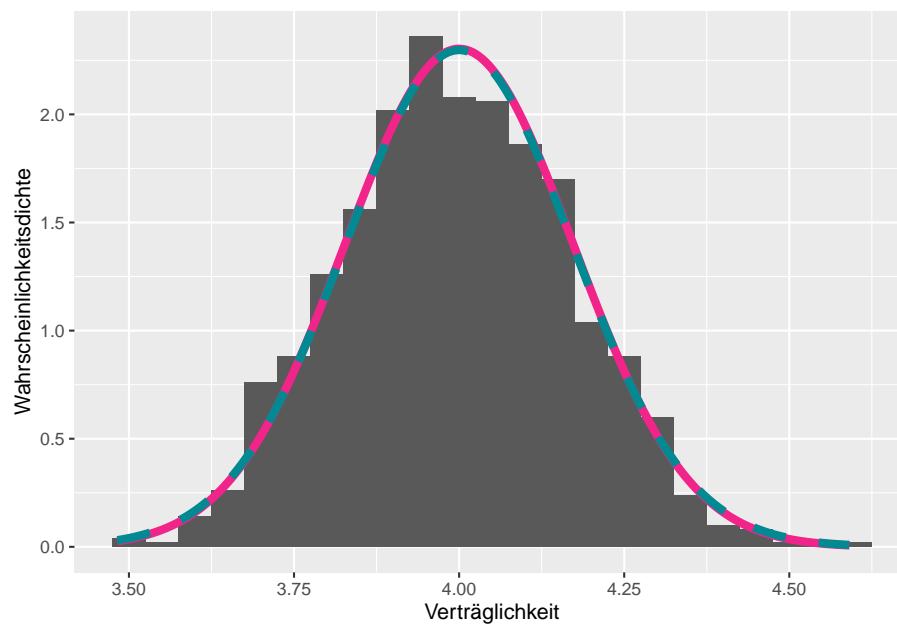


Abbildung 4.6: Die arithmetischen Mittel sind Normalverteilt mit Parametern $\mu_g = 3.91$ und $\sigma_g = 1.73/\sqrt{100}$.

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

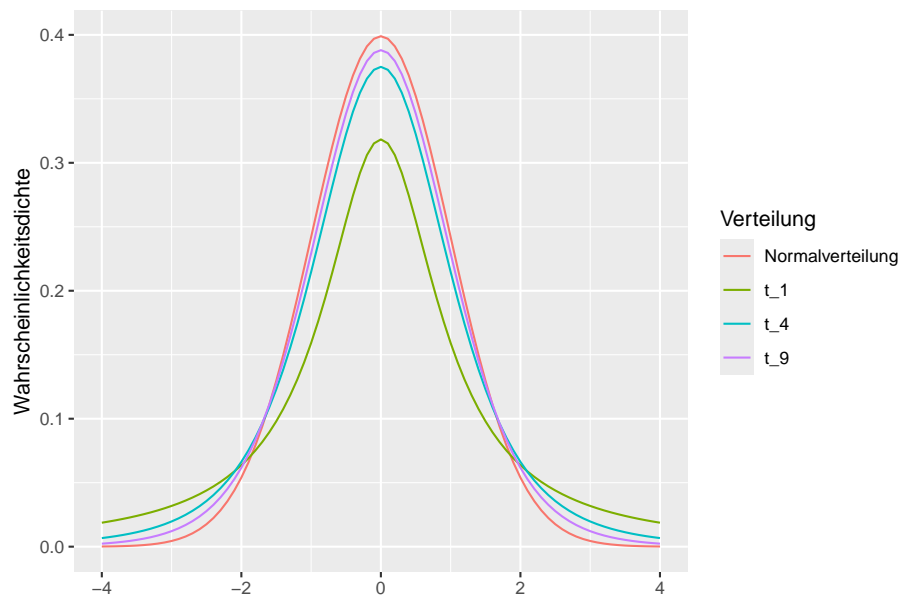


Abbildung 4.7: Student- t -Verteilungen mit 1, 4 und 9 Freiheitsgraden im Vergleich zu der Normalverteilung.

Die Freiheitsgrade der t -Verteilung in der Annäherung oben entsprechen der Anzahl Beobachtungen minus 1, also $df = n - 1$. Die höhere Wahrscheinlichkeit von extremen Werten bei kleinen Freiheitsgraden spiegelt die grössere Unsicherheit der Schätzung des Erwartungswertes wieder, 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

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

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

Das 2.5% und das 97.5% Perzentil der Verteilung der arithmetischen Mittel ergeben nun die untere respektive obere Schranke eines **Intervalles**. 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.

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.

```
## [1] "hi"
```

Lösung. TODO

Übung 4.2.

```
## [1] "hi"
```

Lösung. TODO

Übung 4.3.

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 4.8 dargestellt. Die Daten sind nicht normalverteilt, weil keine Glockenkurve wie oben beschrieben das Histogramm gut abdecken würde.

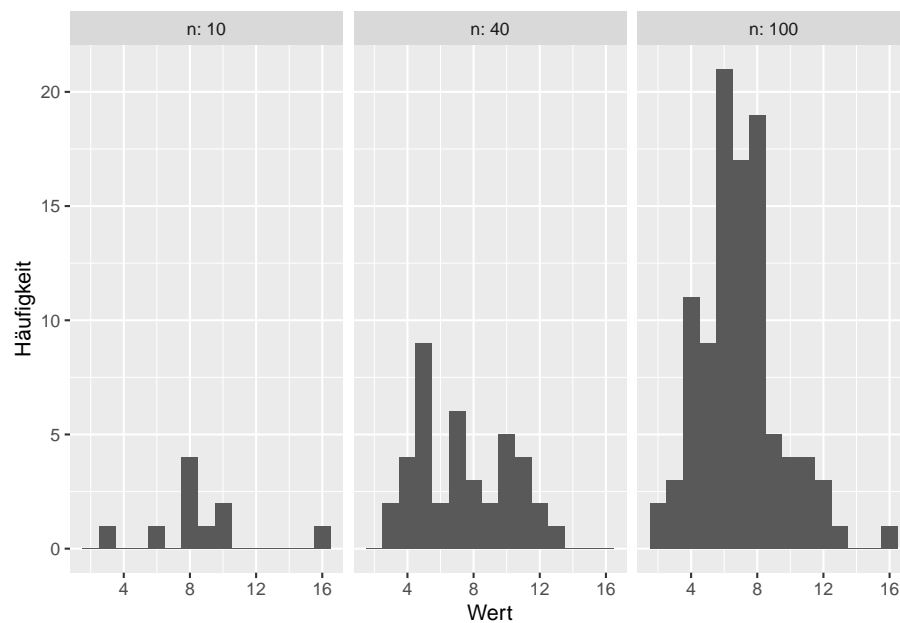


Abbildung 4.8: TODO.

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. Die Übereinanderlegung des jeweiligen Histogramms und der Wahrscheinlichkeitsdichte der Normalverteilung wird in Abbildung 4.9 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 4.8. 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.

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:

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

Vergleichen Sie die Resultate der Berechnungen für jede Stichprobe:

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

Lösung. Abbildung 4.10 zeigt die Berechnungsanweisungen für Jamovi und die resultierende Tabelle daraus.

- Der Erwartungswert der Wartezeiten (das heisst der Populationmittelerwartungswert 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.

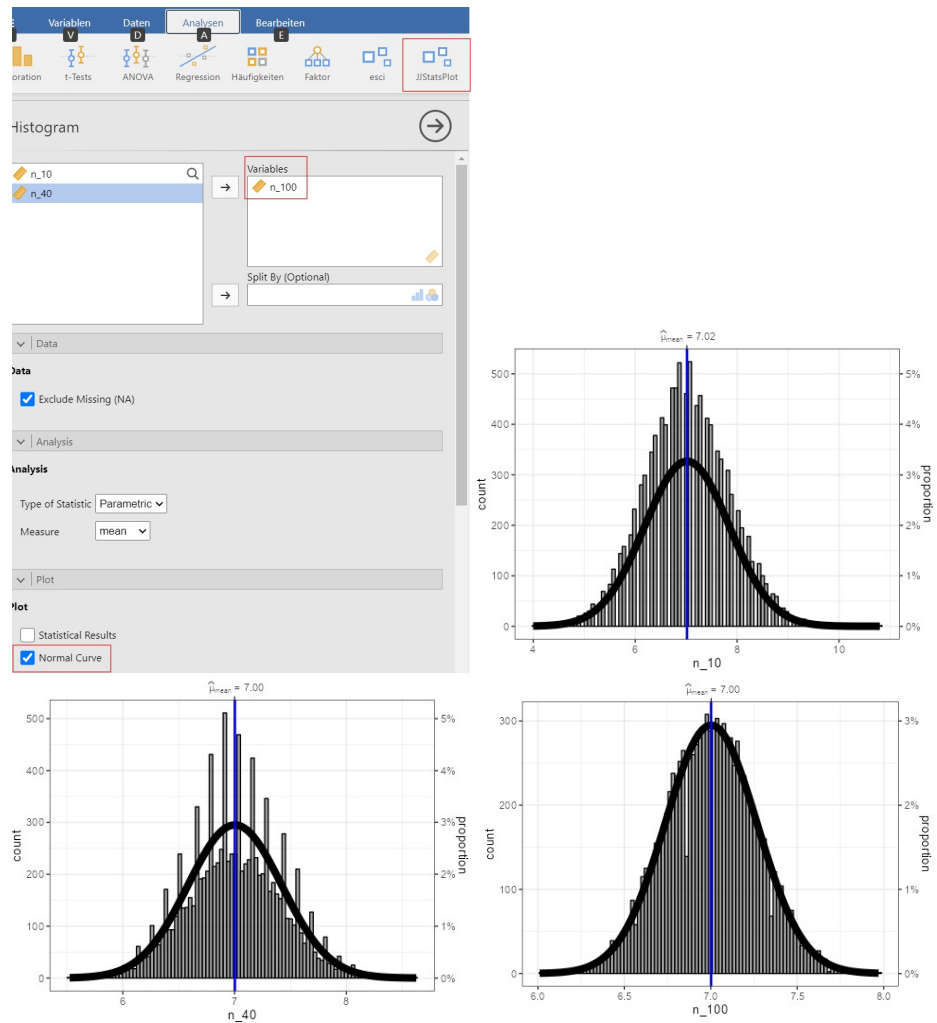


Abbildung 4.9: TODO

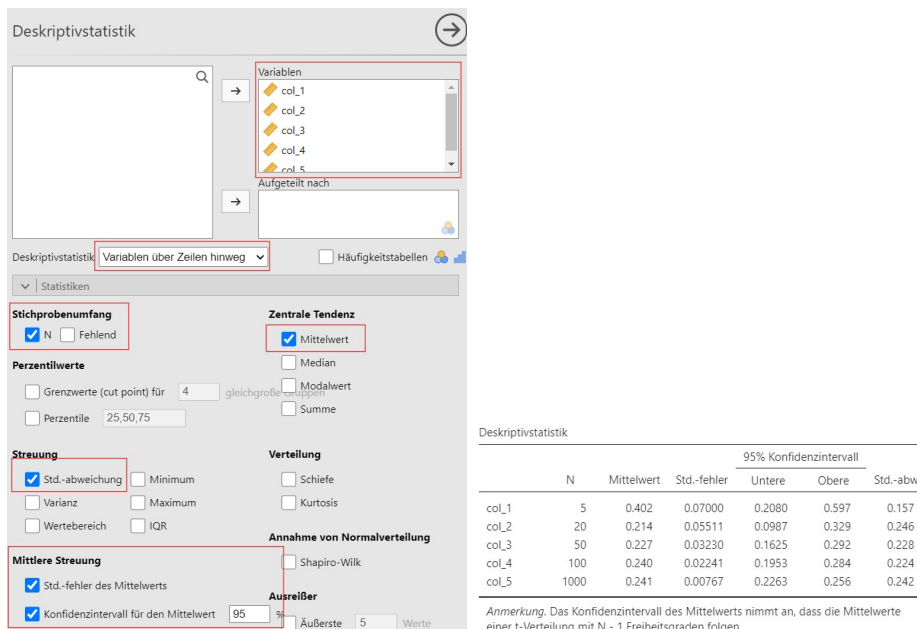


Abbildung 4.10: Links: Jamovi-Anleitung zur Erstellung der Tabelle mit den relevanten Kenngrößen; rechts: Tabelle mit relevanten Kenngrößen.

- 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.
- 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.
- Die untere und obere Schranke der 95%-Konfidenzintervalle sind bei **Untere** und **Obere** respektive abzulesen.
- 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.
- Die Schätzung des Erwartungswertes ist das arithmetische Mittel der Stichprobe. Da jedesmal eine neue Zufallsstichprobe gezogen wurde und diese nicht dieselben Beobachtungen enthalten, ergeben sich auch jedesmal andere Stichprobenmittelewerte.
- Je grösser n , desto kleiner ist das Konfidenzintervall. Wenn man also ein kleines Konfidenzintervall erreichen will, braucht man eine grössere Stichprobe.

Übung 4.5.

44KAPITEL 4. DURCHSCHNITT UND STANDARDABWEICHUNG SCHÄTZEN

```
## [1] "hi"
```

Lösung. TODO

Kapitel 5

Zentrale Tendenz testen

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

H : Es regnet.

Ist die Hypothese einmal gefunden, können Daten gesammelt werden, um diese Hypothese zu bestätigen oder zu falsifizieren. Man geht nach draussen und spürt Regen auf der Haut bedeutet H ist wahr, 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 : Es regnet nicht.

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

H_1 : Es regnet.

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 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 **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

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

Achtung



Hinweis. Die verwendeten Zeichen in den Formeln sind

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

Beispiel 5.1 (Vermögen).

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

Alternativhypothese, dass das durchschnittliche Vermögen gestiegen ist:

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

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

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

Es kann nun schnell gesagt werden, dass das durchschnittliche Vermögen in der Population gestiegen ist, weil 193'000 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 von Ronald Fisher Denkweise ist die Frage nach der Wahrscheinlichkeit, dass der beobachtete Wert oder ein noch extremerer Wert in Richtung der Alternativhypothese resultiert. 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* auch **p-Wert** (Symbol: p) genannt. Beim Berichten des p-Werts wird normalerweise die führende 0 nicht geschrieben, also $p = .143$.

5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT?49

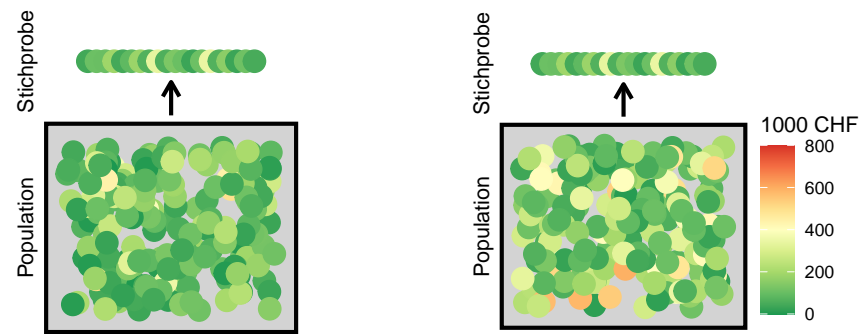


Abbildung 5.1: Vorgestellte Zufallsstichprobenziehung. Links: Nullhypothese ist wahr. Rechts: Nullhypothese ist falsch.

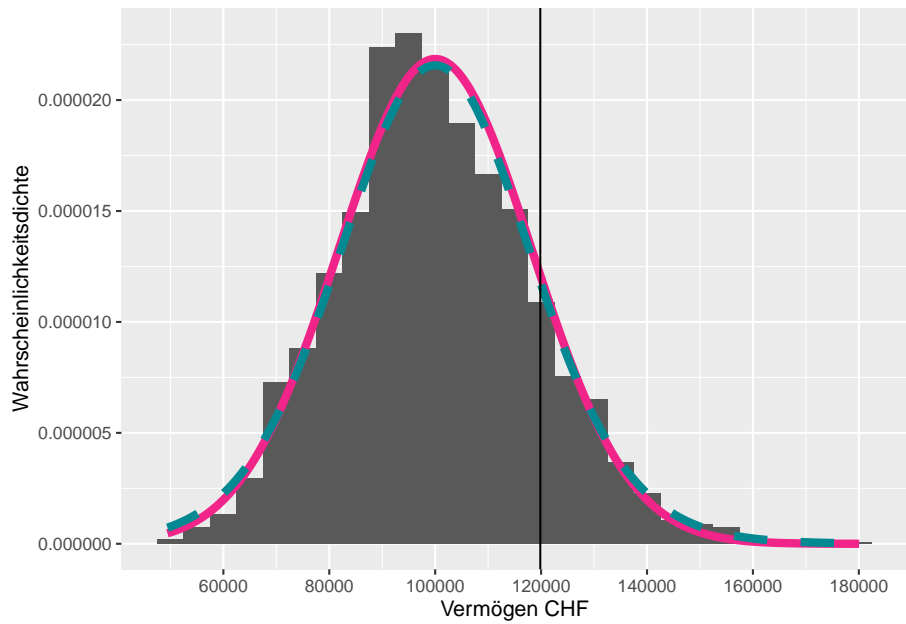


Abbildung 5.2: TODO.

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

Achtung



Hinweis. Ein signifikanter Unterschied bedeutet im allgemeinen Sprachgebrauch ein *bedeutsamer, substanzieller* Unterschied. Im statistischen Kontext bedeutet ein *signifikanter Unterschied*, wie oben beschrieben, dass ein Unterschied bis auf eine gewisse Irrtumswahrscheinlichkeit (angegeben durch das Signifikanzniveau) *nicht zufällig* zustande gekommen ist. Ein *nicht signifikanter Unterschied* bedeutet dagegen, dass die Beobachtung *zufällig* zustande gekommen sein könnte. Für letzteres gibt es zwei Erklärungen: (1) H_0 ist tatsächlich wahr. (2) H_0 ist zwar falsch, aber die Stichprobenziehung hat zufällig zu einem ähnlichen Resultat geführt, wie wenn H_0 wahr wäre. Ist ein Testresultat nicht signifikant, so kann also nicht genau gesagt werden, ob H_0 wahr ist oder nicht. Ist das Testresultat signifikant, so ist H_0 eher unwahrscheinlich.

In manchen Texten werden allgemeine und auch statistische Fragen bearbeitet. Hier empfiehlt sich für den allgemeinen Sprachgebrauch *substanziell* und für die statischen Aussagen *statistisch signifikant* zu verwenden.

Es wird ausserdem empfohlen, das Wort signifikant immer nur als Prädikat für eine Qualifizierung der Nullhypothese zu verwenden. Im Beispiel war $H_0 : \mu \leq 100'000\text{CHF}$. Korrekte Aussage sind:
 - Das durchschnittliche Vermögen ist im letzten Jahr nicht signifikant gewachsen.
 - Das durchschnittliche Vermögen ist in diesem Jahr nicht signifikant grösser als 100'000 CHF.

Die beiden Denkart 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 $p = .143$.

5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT? 51

Dies bedeutet, dass die Wahrscheinlichkeit zufällig den realisierten Stichprobenmittelwert zu kommen, 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 μ und Standardabweichung σ hat, der Wert

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

normalverteilt ist, wobei μ hier dem Wert der Nullhypothese entspricht, also 100'000 CHF. Dies entspricht der roten Linie in Abbildung 5.2. Ist die Standardabweichung des Merkmals unbekannt, so wird diese mit 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 Statistik wird dieser Test **t -Test** genannt.

Das oben gefundene Resultat wird in der folgenden Form berichtet:

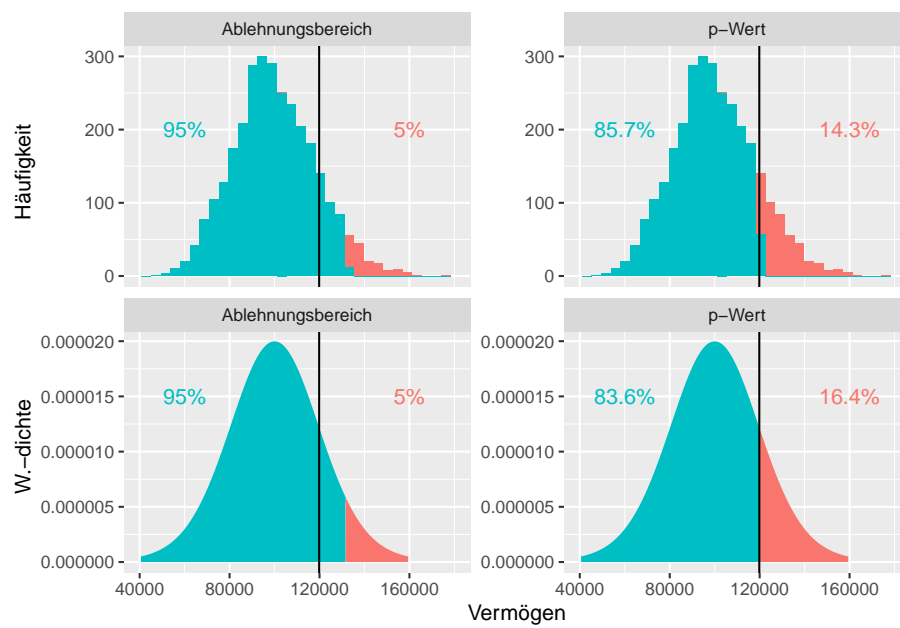


Abbildung 5.3: Oben: Histogramm der simulierten Verteilung; unten: theoretische t-Verteilung; links: Illustration p-Wert; rechts: Illustration Ablehnungsbereich. Die Linie entspricht dem beobachteten Stichprobenmittelwert.

5.1. ENTSPRICHT DER ERWARTUNGSWERT EINEM GEWISSEN WERT?53

Das durchschnittliche Vermögen ($M = 119853$ CHF, $SD = 88528$, $N = 20$) ist in diesem Jahr nicht signifikant grösser als 100'000 CHF, $t(19) = 1.003$, $p = .164$.

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

Der erste Schritt ist auch hier die Null- und Alternativhypothesen aufzustellen. Die Psychologin stellt die Frage, ob sich die durchschnittliche Alexithymie in der Grundgesamtheit, in der Folge mit μ bezeichnet, von 100 unterscheidet oder nicht. Es ist zu beobachten, dass sie keine Annahme über die Richtung der

Abweichung trifft (eine höhere oder eine tiefere Alexithymie wären denkbar) und es sich deshalb um eine zweiseitige Hypothesenstellung handelt.

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

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

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 eine Simulation 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. Extremere 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%.

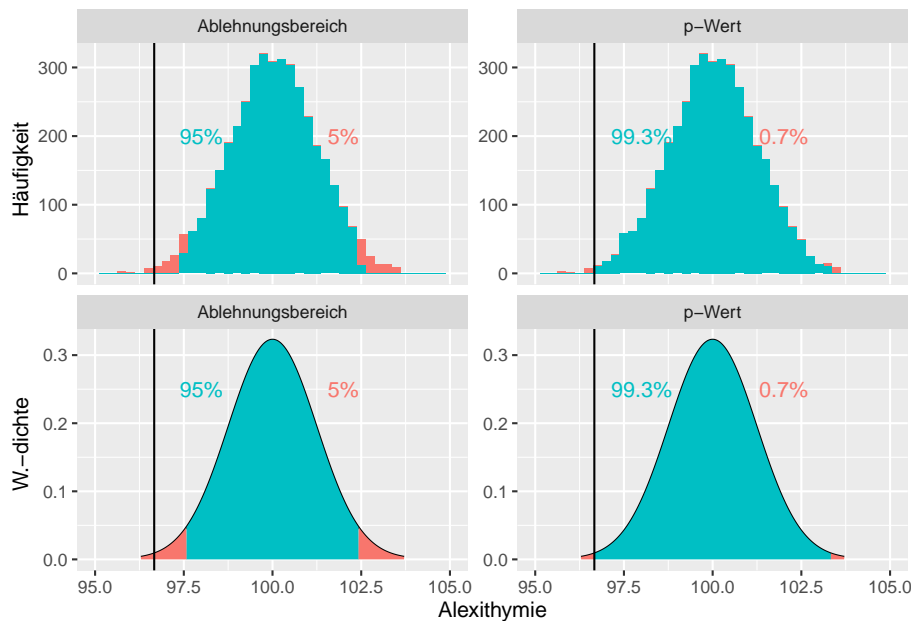


Abbildung 5.4: TODO:

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:

Die durchschnittliche Alexithymie ($M = 96.7$ Punkte, $SD = 24.4$, $N = 391$) unterscheidet sich bei den unter 25-jährigen signifikant vom Populationsdurchschnitt von 100 Punkte, $t(390) = -2.698$, $p = .007$.

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

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

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) ein 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:

Das durchschnittliche Vermögen ($M = 119853$ CHF, $SD = 88528$, $N = 20$) ist in diesem Jahr nicht signifikant grösser als 100'000 CHF, $t(19) = 1.003$, $p = .164$, $d = 0.22$.

Die durchschnittliche Alexithymie ($M = 96.7$ Punkte, $SD = 24.4$, $N = 391$) unterscheidet sich bei den unter 25-jährigen signifikant vom Populationsdurchschnitt von 100 Punkte, $t(390) = -2.698$, $p = .007$, $d = -0.14$.

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 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.

Übung 5.2.

Reproduziere das Beispiel Alexithymie 5.2 mit **Jamovi** indem folgende Teilschritte durchgeführt werden:

- 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.

Tests

☒ Student's

☐ Bayes-Faktor

Vorannahme (prior)

☐ Wilcoxon-Rang

Hypothese

Testwert

☐ ≠ Testwert

☒ > Testwert

☐ < Testwert

Zusätzliche Statistiken

☐ Mittlere Differenz

☐ Konfidenzintervall %

☒ Effektstärke

☐ Konfidenzintervall %

☒ Deskriptivstatistik

☐ Deskriptive Diagramme

Überprüfung der Voraussetzungen

☐ Test auf Normalverteilung

☐ Q-Q-Diagramm

Abbildung 5.5: Jamovi Eingabe.

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: Deskriptive Statistiken.

Tests

☒ Student's

☐ Bayes-Faktor

Vorannahme (prior)

☐ Wilcoxon-Rang

Hypothese

Testwert

☒ ≠ Testwert

☐ > Testwert

☐ < Testwert

Zusätzliche Statistiken

☐ Mittlere Differenz

☐ Konfidenzintervall %

☒ Effektstärke

☐ Konfidenzintervall %

☒ Deskriptivstatistik

☐ Deskriptive Diagramme

Überprüfung der Voraussetzungen

☐ Test auf Normalverteilung

☐ Q-Q-Diagramm

Abbildung 5.7: Jamovi Eingabe.

t-Test mit einer Stichprobe

t-Test mit einer Stichprobe

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

Anmerkung. $H_a: \mu \neq 100$

Deskriptivstatistik

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

Abbildung 5.8: Deskriptive Statistiken.

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.

- 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?
- 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.
- 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.

Tests	Zusätzliche Statistiken
<input checked="" type="checkbox"/> Student's	<input type="checkbox"/> Mittlere Differenz
<input type="checkbox"/> Bayes-Faktor	<input type="checkbox"/> Konfidenzintervall 95 %
Vorannahme (prior) 0.707	<input checked="" type="checkbox"/> Effektstärke
<input type="checkbox"/> Wilcoxon-Rang	<input type="checkbox"/> Konfidenzintervall 95 %
Hypothese	<input checked="" type="checkbox"/> Deskriptivstatistik
Testwert 24	<input type="checkbox"/> Deskriptive Diagramme
<input checked="" type="radio"/> ≠ Testwert	Überprüfung der Voraussetzungen
<input type="radio"/> > Testwert	<input type="checkbox"/> Test auf Normalverteilung
<input type="radio"/> < Testwert	<input type="checkbox"/> Q-Q-Diagramm

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. $H_1: \mu \neq 24$

Deskriptivstatistik

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

Abbildung 5.10: Jamovi Ausgabe.

- 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.
- 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. μ 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 Tagesdauer verkürzt oder verlängert gegenüber der Nullhypothese. Es ist hier also eine zweiseitige Hypothesenstellung und die Alternativhypothese lautet $H_1: \mu \neq 24$ Stunden.

Die durchschnittliche Tageslänge ($M = 24.6$ Stunden, $SD = 1.8$, $N = 57$) unter Experimentalbedingungen unterscheidet sich signifikant von 24 Stunden, $t(56) = 2.41$, $p = 0.019$, $d = 0.319$.

- 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 Alternativhypothese 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.

- Wie viele Schwimmer hat die Sportstudentin trainiert?
- 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.

The image shows the Jamovi software interface for setting up a one-sample t-test. The 'Tests' section on the left has 'Student's' selected with a blue checkmark. Below it, 'Bayes-Faktor' is unchecked, and 'Vorannahme (prior)' is set to 0.707. 'Wilcoxon-Rang' is also unchecked. The 'Hypothese' section shows 'Testwert' set to 1.58. Under the hypothesis options, the radio button for '< Testwert' is selected. On the right, the 'Zusätzliche Statistiken' (Additional Statistics) section has 'Mittlere Differenz' and 'Konfidenzintervall' (95%) unchecked, while 'Effektstärke' (Effect Size) is checked. Below that, 'Konfidenzintervall' (95%) is also unchecked. 'Deskriptivstatistik' (Descriptive Statistics) is checked, and 'Deskriptive Diagramme' (Descriptive Diagrams) is unchecked. The 'Überprüfung der Voraussetzungen' (Check Assumptions) section at the bottom right has 'Test auf Normalverteilung' (Test for Normality) and 'Q-Q-Diagramm' (Q-Q Plot) both unchecked.

Abbildung 5.11: Jamovi Eingabe.

- Die Sportstudentin hat $N = 13$ Schwimmer trainiert.
- Die Forschungsfrage ist hier, ob die neue Trainingsmethode besser ist. Besser meint hier, dass die mit dieser Trainingsmethode trainierten

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. $H_1: \mu < 1.58$

Deskriptivstatistik

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

Abbildung 5.12: Jamovi Ausgabe.

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 Schwimmzeit mit der neuen Methode gleich bleibt oder sogar noch länger wird $H_0: \mu \geq 1.58$. Das Testresultat lässt sich wie folgt berichten:

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

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

Teil II

Gruppenvergleich einer intervallskalierten Variable

Kapitel 6

Gruppenvergleich einer intervallskalierten Variable

6.1 Zwei Gruppen vergleichen

6.2 Was ist das Problem der Stichprobenziehung?

6.3 Wie kann man Aussagen über die Grundgesamtheit machen?

6.4 Übungen

Kapitel 7

Welch-Test

7.1 Zwei Gruppen vergleichen

7.2 Sind die Durchschnitte der beiden Gruppen in der Grundgesamtheit gleich?

7.3 Wie stark unterscheiden sich die Durchschnitte?

7.4 Übungen

Begriffsverzeichnis

- Alternativhypothese
- Erwartungswert
- Freiheitsgrade
- Grundgesamtheit
- Hypothese
- Interquartilabstand
- Intervalles
- Konfidenzintervall
- Median
- Modus
- Normalverteilung
- Nullhypothese
- Perzentil
- Population
- Spannweite
- Standardabweichung
- Stichprobe
- Stichprobenziehung
- Student- t -Verteilung
- Zentraler Grenzwertsatz
- Zufallsstichprobe
- arithmetische Mittel
- intervallskaliert
- zweiseitige Hypothese

Literaturverzeichnis

- 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.
- 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.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Selker, R., Love, J., and Dropmann, D. (2024). *jmv: The jamovi Analyses*. R package version 2.5.6.
- 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.
- 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.