

Statistik 1

Daniel J. F. Gerber

2024-10-05

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	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 eine intervallskaliertes Merkmal beschrieben werden?	14
2.3 Übungen	16
3 Stichprobenziehung	19
3.1 Was ist das Problem der Stichprobenziehung?	20
3.2 Wie kann man Aussagen über die Grundgesamtheit machen?	22
3.3 Übungen	22
4 Durchschnitt und Standardabweichung schätzen	23
4.1 Wo liegt der Durchschnitt der Grundgesamtheit?	23
4.2 Wo liegt der Durchschnitt der Standardabweichung?	23
4.3 Übungen	23

5	Durchschnitt testen	25
5.1	Entspricht der Durchschnitt der Grundgesamtheit einem gewissen Wert?	25
5.2	Weicht der gefundene Durchschnitt stark vom hypothetischen Wert ab?	25
5.3	Übungen	25
II	Gruppenvergleich einer intervallskalierten Variable	27
6	Gruppenvergleich einer intervallskalierten Variable	29
6.1	Zwei Gruppen vergleichen	29
6.2	Was ist das Problem der Stichprobenziehung?	29
6.3	Wie kann man Aussagen über die Grundgesamtheit machen? . .	29
6.4	Übungen	29
7	Welch-Test	31
7.1	Zwei Gruppen vergleichen	31
7.2	Sind die Durchschnitte der beiden Gruppen in der Grundgesamtheit gleich?	31
7.3	Wie stark unterscheiden sich die Durchschnitte?	31
7.4	Übungen	31

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

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.

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.

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.

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.

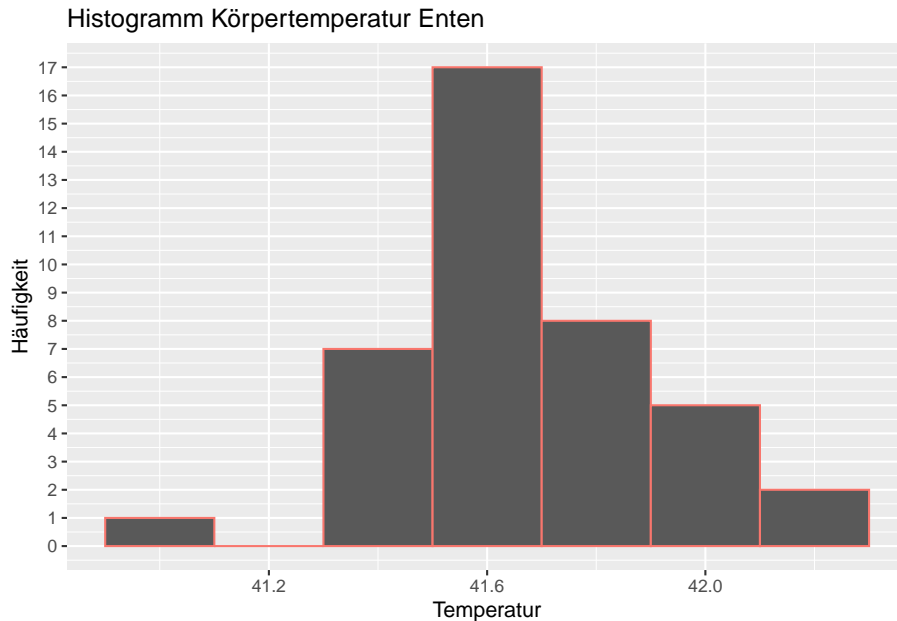
2.2 Wie kann eine intervallskaliertes Merkmal beschrieben werden?

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, 41.67.

Für einen Menschen ist es ziemlich 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 EINE 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.
- 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, 42.18, also 41.655.
- 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.

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 2.2 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

TODO: Quantile TODO: Standardabweichung

2.3 Übungen

Übung 2.1.

- (a) Versuch selbst ein Histogramm der Daten oben (*Enten_n40.sav*) mit Jamovi zu erstellen und begründe, weshalb es nicht gleich aussieht wie das Histogramm oben.
- (b) Berechne zusätzlich das arithmetische Mittel und die Standardabweichung des Merkmals.

Lösung.

- (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) TODO

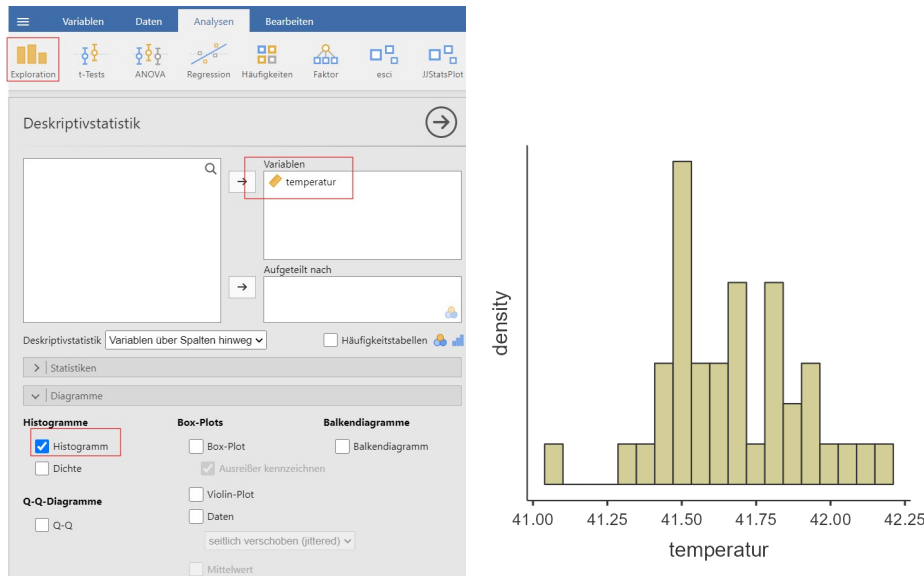


Abbildung 2.1: Links: Jamovi-Anleitung zur Erstellung des Histogramms; rechts: Histogramm der Temperatur.

Übung 2.2. 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.

- Wie gross ist das arithmetische Mittel und der Median dieser Werte?
- 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?
- 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.

- 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 gar keine Fehler enthalten, ist das arithmetische Mittel gleich gut geeignet wie der Median.

Kapitel 3

Stichprobenziehung

Forschende haben ein Messinstrument für Angst entwickelt, welches STAI (State-Trait Anxiety Inventory) (Spielberger et al., 1983). Sie erheben dabei unter anderem einen Wert zwischen 20 und 80 für eine Zustandsangst. A priori haben die Forschenden keine Ahnung, wie ängstlich eine Person im Durchschnitt ist oder ob die ganze Skala der Werte genutzt wird. Die Forscher:innen machen deshalb eine kleine Befragung mit 30 zufällig ausgewählten Studierenden. Zufällig ausgewählte Beobachtungen eines Merkmals werden als **Stichprobe** bezeichnet. Die Forschenden finden die zusammenfassenden Werte $M = 43.2$, $s = 7.8$, $n = 30$ für die Zustandsangst in ihrer Stichprobe.

Anschließend stellt sich die Frage, wie stark diese Werte basierend auf dieser Stichprobe Zustandsangst von allen Personen widerspiegelt. 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 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 Farbe der Punktes ist jetzt bekannt und entspricht der jeweiligen Zustandsangst der beobachteten Person.

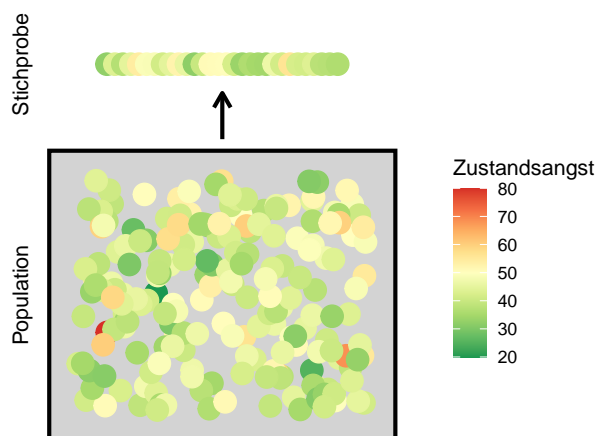


Abbildung 3.1: Personen ohne Angst. TODO, make points black in population.

Da die Stichprobe nun eben zufällig gezogen wurde, dass 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.

Das Problem der zufälligen Stichprobenziehung ist also, dass nie ganz klar ist, wie die darunterliegende Population aussieht. Sind die Werte der Stichprobe tief, weil zufällig gerade Studierende mit tiefer Zustandsangst beobachtet wurden,

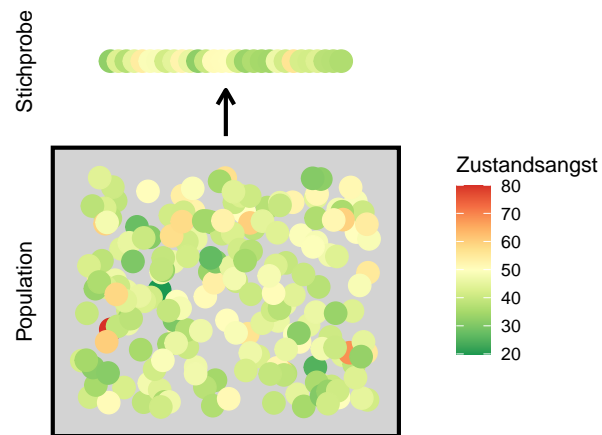


Abbildung 3.2: Population mit ähnlichen Werten wie in der Stichprobe.

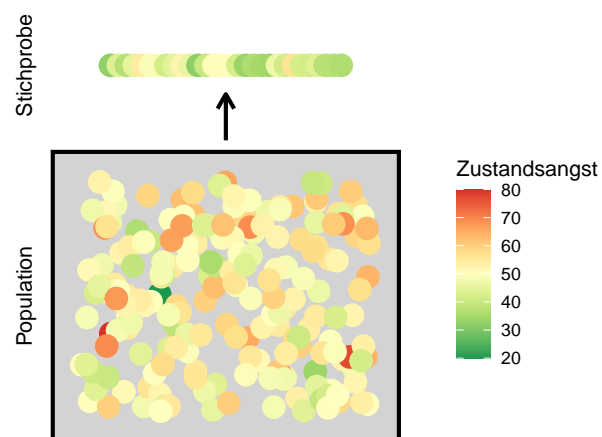


Abbildung 3.3: Population mit höheren Werten als in der Stichprobe.

oder haben tatsächlich die meisten Studierenden eine tiefe Zustandsangst?

3.2 Wie kann man Aussagen über die Grundgesamtheit machen?

3.3 Übungen

Kapitel 4

Durchschnitt und Standardabweichung schätzen

- 4.1 Wo liegt der Durchschnitt der Grundgesamtheit?
- 4.2 Wo liegt der Durchschnitt der Standardabweichung?
- 4.3 Übungen

Kapitel 5

Durchschnitt testen

- 5.1 Entspricht der Durchschnitt der Grundgesamtheit einem gewissen Wert?
- 5.2 Weicht der gefundene Durchschnitt stark vom hypothetischen Wert ab?
- 5.3 Übungen

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 Durch-
schnitte?

7.4 Übungen

Literaturverzeichnis

- 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., Golemund, 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.