Statistische Verfahren in der Geographie

Skript für den Theorieteil

Till Straube
straube@geo.uni-frankfurt.de

Institut für Humangeographie
Goethe-Universität Frankfurt

Sommersemester 2023

Inhaltsverzeichnis

Terminüberblick

Alle Sitzungen finden von 14 bis 16h c. t. statt, die Klausuren s. t.

Datum	Sitzung	Inhalt	Ort
11. April 2023		Vorbesprechung	HZ10
18. April 2023	1	Datenerhebung und Häufigkeiten	HZ10
25. April 2023	2	Maßzahlen	HZ10
2. Mai 2023	3	z-Werte und Normalverteilung	HZ10
9. Mai 2023	4	Schätzstatistik	HZ10
16. Mai 2023	5	[Grundlagen der Teststatistik]	HZ10
23. Mai 2023	6	[Testverfahren mit zwei Stichproben]	HZ10
30. Mai 2023	7	[Korrelation]	HZ10
6. Juni 2023	8	[Lineare Regression]	HZ10
13. Juni 2023	9	[Kreuztabellen]	HZ10
20. Juni 2023	10	[χ^2-Tests]	HZ10
27. Juni 2023		Klausurvorbereitung	HZ10
4. Juli 2023		Klausur (14h s. t.)	H IV
10. Oktober 2023		Nachklausur (14h s. t.)	TBA

Vorbesprechung

Lernziele der Veranstaltung

Sie können…

- Grundbegriffe der Statistik sinnvoll verwenden.
- die wichtigsten statistischen Kennzahlen berechnen.
- gängige Diagramme interpretieren.
- einfache statistische Schätz- und Prüfverfahren anwenden.
- passende Verfahren für verschiedene Aufgaben wählen.

Konzept der Veranstaltung

- Die gesamte Veranstaltung dient als Klausurvorbereitung
- Die selbständige Anwendung der Verfahren steht im Vordergrund
- Veranstaltung folgt dem Flipped-Classroom-Konzept

Sitzungsvorbereitung

- Materialien werden zur eigenständigen Vorbereitung bereit gestellt
- Dieses Online-Skript mit den Kerninhalten
- Darin: Videos (aus 2020) mit Beispielen und Übungen
- Darin: Verweise auf weiterführende Literatur, YouTube-Videos, etc.
- Fehler und Unklarheiten bitte per E-Mail melden!

Sitzungsablauf

- Dienstags, 14 h c. t. in HZ10
- Übungsaufgaben (und Lösungen) werden online bereit gestellt
- Teilnehmer*innen bearbeiten die Aufgaben in Kleingruppen
- Bei Problemen fragen Sie sich erst mal gegenseitig
- Sonst bin ich immer ansprechbar

Empfehlungen

- Lassen Sie sich auf den wöchentlichen Rhythmus ein
- Bereiten Sie die Sitzungen vor und nach
- Bilden Sie Lerngruppen
- Gleichen Sie in Lerngruppen Ihre Ziele ab
- Machen Sie sich mit Ihrem Taschenrechner vertraut

Literaturempfehlungen

- Ganz besonders:
  - Bortz und Schuster (2010) (als E-Book bei der UB erhältlich; dieselben Notationskonventionen wie in der Veranstaltung)
- Ergänzend:
  - Lange und Nipper (2018) (geographiebezogen)
  - Bahrenberg, Giese und Nipper (2010) (geographiebezogen)
  - Benninghaus (2007) (als E-Book bei der UB erhältlich)
- Bedingt:
  - Zimmermann-Janschitz (2014) (geographiebezogen; als E-Book bei der UB erhältlich)
- Englisch:
  - Burt und Barber (1996)

Taschenrechner

- Zulassungsregeln für Klausur wie für Mathe-Abi (Hessen)
- Also kein „programmierbarer“ Taschenrechner
- Erlaubt ist z. B. CASIO FX-991DE Plus
- „Wissenschaftlicher“ Taschenrechner kann von großem Vorteil sein… aber den statistischen Funktionen nicht blind vertrauen!

Sitzung 1

Datenerhebung und Häufigkeiten

Lernziele dieser Sitzung

Sie können…

- einige Grundbegriffe der Statistik definieren.
- Typen von Stichproben unterscheiden.
- Skalenniveaus von Variablen bestimmen.
- Häufigkeitsverteilungen beschreiben.

Lehrvideos (Sommersemester 2020)

- 1a) Grundbegriffe
- 1b) Skalenniveaus
- 1c) Grundbegriffe

1.1 Statistische Praxis

Was ist Statistik? Je nach Perspektive kann Statistik vieles sein: ein Teilgebiet der Mathematik, ein Untersuchungsobjekt kritischer Forschung oder ein unbeliebtes Studienfach.

Im Rahmen dieser Veranstaltung soll Statistik als eine Zusammenstellung von Praktiken in der quantitativen Forschung verstanden werden, wobei ihre Anwendung stets im Mittelpunkt steht. Eine hilfreiche Definition findet sich bei Haseloff u. a. (1968):

> „Allgemein kann gesagt werden: Die Statistik hat es mit Zahlen zu tun, die entweder aus Abzählvorgängen oder aus Messungen gewonnen wurden. Ihre Aufgabe ist es, ein solches Zahlenmaterial in eine optimal übersichtliche und informationsreiche Form zu bringen, aus ihnen methodische Schlußfolgerungen zu ziehen und gegebenfalls auch die Ursachen der analysierten Zahlenverhältnisse mit sachlichen Methoden aufzudecken.“ (Haseloff u. a. 1968: 27)

1.1.1 Grundbegriffe der Statistik

1.1.1.1 Untersuchungselement

Untersuchungselemente (auch Untersuchungseinheiten, Merkmalsträger, bei Personen: Proband*innen, engl. *sampling unit*) sind die individuellen Gegenstände empirischer Untersuchungen. Bei einer Hochrechnung zur Bundestagswahl ist dies z. B. eine befragte Wählerin.

1.1.1.2 Stichprobe

Eine Stichprobe (engl. *sample*) ist die Menge aller Untersuchungselemente, deren Daten direkt erhoben werden. Die Anzahl der Untersuchungselemente in der Stichprobe wird in Formeln mit n bezeichnet. Bei einer Hochrechnung z.B. bilden alle tatsächlich befragten Wähler*innen die Stichprobe.

1.1.1.3 Grundgesamtheit

Die Grundgesamtheit (auch Population, engl. *population*) ist die Menge aller potentiell untersuchbaren Elemente, über die Aussagen getroffen werden sollen. Die Stichprobe ist eine Teilmenge der Grundgesamtheit. Die Anzahl der Elemente in der Grundgesamtheit wird in Formeln mit N bezeichnet. Bei einer Hochrechnung zur Bundestagswahl sind dies z.B. alle Wähler*innen (bzw. alle Wahlberechtigten, wenn Wahlbeteiligung von Interesse ist).

1.1.1.4 Variable

Variablen (auch Merkmale, engl. *variable*) sind Informationen über die Untersuchungselemente, die in einer Untersuchung von Interesse sind. Typischerweise unterscheiden sie sich von Untersuchungselement zu Untersuchungseelement, sind also variabel. Bei einer Hochrechnung ist dies die Antwort auf die Frage: „Welche Partei haben Sie gerade gewählt?“

1.1.1.5 Wert

Ein Wert (auch Merkmalsausprägung, engl. *observation*) ist die erfasste Ausprägung einer Variable bei einem Untersuchungselement. In Formeln werden Werte mit $x_1, x_2, x_3, ..., x_n$ durchnummeriert. Bei einer Hochrechnung kann die Variable „gewählte Partei“ für ein Untersuchungselement z.B. den Wert „CDU“ annehmen.

1.1.1.6 Kennwert

Kennwerte (auch Maßzahlen, Kennzahlen, engl. *summary statistics*) sind Zahlen, die aus den beobachteten Werten errechnet werden. Sie können beispielsweise Aufschluss über Mittelwerte und Verteilung einer Variable oder den Zusammenhang mehrerer Variablen geben. Bei einer Hochrechnung sind z.B. die relativen Häufigkeiten (in Prozent) der Variable „gewählte Partei“ von besonderem Interesse.

1.1.2 Taxonomien statistischer Verfahren

Statistische Verfahren werden in mehrerlei Hinsicht unterschieden, wie im Folgenden beschrieben. Dabei schließen sich verschiedene Kategorien nicht unbedingt aus, es gibt also durchaus statistische Verfahren, die z.B. als univariat *und* deskriptiv bezeichnet werden.

1.1.2.1 Uni-, bi- und multivariate Statistik

Bei diesen Bezeichnungen ist entscheidend, wie viele Variablen bei den jeweiligen Verfahren zum Einsatz kommen. Im Allgemeinen spricht man bei einer Variable von univariater Statistik, bei zwei Variablen von bivariater Statistik und bei mehr als zwei Variablen von multivariater Statistik. (Manchmal werden allerdings auch Verfahren mit nur zwei Variablen als multivariat bezeichnet.)

In dieser Veranstaltung beschäftigen wir uns zunächst mit univariaten, dann mit bivariaten Verfahren. Verfahren mit mehr als zwei Variablen werden nicht behandelt.

1.1.2.2 Deskriptive und schließende Statistik

Unabhängig von der Anzahl der Variablen unterscheidet man auch nach der Art und Weise des Vorgehens:

1.1.2.2.1 Deskriptive Statistik Die deskriptive Statistik (auch: beschreibende Statistik) dient der Beschreibung der Verteilung von Merkmalen, indem sie z. B. Durchschnittswerte bildet, Häufigkeiten bestimmt oder etwas über die Streuung eines Merkmals aussagt. Sie kann so große Datenmengen übersichtlicher machen, indem sie diese ordnet, gruppiert oder verdichtet. Sie erleichtert es also, das Charakteristische, Wichtige zu erkennen.

1.1.2.2.2 Schließende Statistik Die schließende Statistik (auch: analytische, operative Statistik, Inferenzstatistik, Prüfstatistik) verhilft dazu, von Eigenschaften einer Stichprobe auf Eigenschaften der Grundgesamtheit verallgemeinern bzw. schließen zu können (deshalb eben auch: schließende Statistik) und diese Einschätzung überprüfen zu können.

Die schließende Statistik wird weiter unterteilt in Schätz- und Teststatistik:

1.1.2.2.2.1 Schätzende Statistik Die Schätzstatistik schätzt Kennwerte der Grundgesamtheit aus den Kennwerten einer Stichprobe.

1.1.2.2.2.2 Testende Statistik Die Teststatistik überprüft, als wie wahrscheinlich oder unwahrscheinlich gemachte Schätzungen bzw. Hypothesen gelten können.

1.1.3 Ablauf einer statistischen Untersuchung

Eine typische Anwendung statistischer Verfahren in der Forschung folgt diesem Schema:

1.1.3.1 Datenerhebung

- Eigene Erhebung z.B. durch Zählen, Messen, Befragung (primärstatistische Daten)
  - Auswahl von Untersuchungseinheiten
  - Wahl der Datenniveaus
- Rückgriff auf vorhandenes Datenmaterial (sekundärstatistische Daten)

1.1.3.2 Datenaufbereitung

- Verdichtung des gewonnenen Datenmaterials und Digitalisierung in Form einer Datenmatrix
- Verschneidung von mehreren Datensätzen

- Vereinheitlichung und Säuberung der Daten
- Überblick verschaffen durch einfache Beschreibung von Häufigkeiten und Maßzahlen (deskriptive Statistik)

1.1.3.3 Datenauswertung

- Verdichtete Beschreibung von Verteilungsmustern einer Variable (univariate deskriptive Statistik)
- Verdichtete Beschreibung der Beziehung zwischen zwei Variablen (bivariate deskriptive Statistik)
- Schluss von Stichprobe auf Grundgesamtheit (Schätzstatistik)
- Testen von Hypothesen über die Grundgesamtheit (Teststatistik)

1.2 Grundlagen der Datenerhebung

1.2.1 Typen von Stichproben

1.2.1.1 Reine Zufallsstichprobe

Bei endlichen Grundgesamtheiten können Lotterieverfahren angewendet werden. Dabei wird allen Elementen der Grundgesamtheit eine Zahl zwischen 1 und N zugeordnet. Anschließend werden Zufallszahlen ausgewählt und die entsprechenden Elemente in die Stichprobe übernommen.

1.2.1.2 Systematische Zufallsstichprobe

Die Elemente einer endlichen Grundgesamtheit werden in eine Rangordnung gebracht (Nummerierung 1 bis N). Anschließend wählt man jedes (N/n)-te Element aus. So entsteht eine Stichprobe der Größe n.

1.2.1.3 Geschichtete Zufallsstichprobe

Die Elemente einer endlichen Grundgesamtheit werden in Schichten (Klassen) zusammengefasst. Anschließend zieht man eine Zufallsstichprobe aus jeder Schicht. Geschichtete Stichproben setzen die Kenntnis einiger Parameter der Grundgesamtheit voraus. Zur Aufteilung des Stichprobenumfangs auf die einzelnen Schichten wird in der Regel die proportionale Aufteilung gewählt.

1.2.1.4 Klumpenstichprobe

Hier ist die Grundgesamtheit schon in „natürliche" Gruppen aufgeteilt (z.B. Schulklassen) und es werden mehrere dieser Gruppen (Klumpen, engl. *cluster*) nach einem Zufallsverfahren als Stichprobe gewählt.

> „Man beachte, dass ein einzelner Klumpen (…) keine Klumpenstichprobe darstellt, sondern eine Ad-hoc-Stichprobe, bei der zufällige Auswahlkriterien praktisch keine Rolle spielen. Die Bezeichnung „Klumpenstichprobe" ist nur zu rechtfertigen, wenn mehrere zufällig ausgewählte Klumpen vollständig untersucht werden." (Bortz und Schuster 2010: 81)

Tabelle 1.1: Die vier wichtigsten Skalenniveaus

Skalenart	Beispiel	mögliche Aussagen	gültige Lagemaße
Nominalskala	Postleitzahl	Gleichheit, Verschiedenheit	Modus
Ordinalskala	Militärischer Rang	+ Größer-kleiner-Relationen	+ Median
Intervallskala	Temperatur in °C	+ Gleichheit von Differenzen	+ arithmetisches Mittel
Verhältnisskala	Körpergröße	+ Gleichheit von Verhältnissen	+ geometrisches Mittel

1.2.2 Variablentypen

1.2.2.1 Qualitative Variablen

Qualitative Variablen können nicht der Größe nach, sondern nur im Hinblick auf ihre Eigenschaft/Art („Qualität") unterschieden werden (z.B. Parteizugehörigkeit, Telefonnummer, Automarke).

Qualitative Variablen, die nur zwei mögliche Werte annehmen können, nennt man „dichotome" Variablen (etwa Antworten auf Ja-Nein-Fragen).

1.2.2.2 Quantitative Variablen

Quantitative Variablen können der Größe nach unterschieden werden (Bsp. Geburtenzahl, Arbeitslosenzahl).

Quantitative Variablen können diskret oder stetig sein:

1.2.2.2.1 Diskrete Variablen Diskrete Variablen (auch diskontinuierliche Variablen) können nur endlich viele, ganzzahlige Werte annehmen. Zwischen zwei Ausprägungen befindet sich eine abzählbare Menge anderer Ausprägungen (z.B. Anzahl eigener Kinder, Haushaltsgröße in Personen).

1.2.2.2.2 Stetige Variablen Stetige Variablen (auch: kontinuierliche Variablen) können in einem bestimmten Bereich jede beliebige Ausprägung annehmen. Der Ausdehnungsbereich kennt keine Lücken, sondern ist als ein fortlaufendes Kontinuum vorstellbar: Bei stetigen Variablen können zwischen zwei Werten oder Ausprägungen unendlich viele weitere Ausprägungen oder Werte liegen (z.B. Körpergröße, Längengrad in Dezimalform).

1.2.3 Skalenniveaus

Eine Variable lässt sich aufgrund ihrer Eigenschaften einem Skalenniveau (auch Skalentyp, Messniveau, Datenniveau, engl. *level of measurement*) zuordnen. Bestimmte Rechenoperationen und statistische Verfahren setzen bestimmte Skalenniveaus voraus. Deshalb ist es wichtig zu wissen, welchem Skalenniveau eine Variable zuzuordnen ist.

Variablen lassen sich immer auch einem niedrigeren Skalenniveau zuordnen. Dies geht allerdings mit Informationsverlust einher.

Die im Folgenden beschriebenen Skalenniveaus sind nicht deckungsgleich mit den o.g. Variablentypen. Intervall- und Verhältnisskalen können z.B. jeweils diskret oder stetig sein.

In Tabelle 1.1 sind die wichtigsten Skalenniveaus im Überblick aufgeführt. „Gültige Lagemaße" sind dabei als Zusatzinformation aufgelistet und werden erst in der nächsten Sitzung behandelt.

1.2.3.1 Nominalskala

Die Merkmalsausprägungen einer Variable stehen je ‚für sich'; sie lassen sich nicht sinnvoll in eine Rangordnung bringen oder gar miteinander verrechnen.

Die einzige Aussage, die sich über zwei Werte in einer Nominalskala treffen lässt, ist dass sie gleich oder nicht gleich sind.

Beispiele: Postleitzahlen, Telefonnummern, Staatsangehörigkeit, Krankheitsklassifikationen

1.2.3.2 Ordinalskala

Die Merkmalsausprägungen einer Variablen lassen sich sinnvoll in eine Rangordnung bringen, die Abstände zwischen den Merkmalsausprägungen aber lassen sich nicht sinnvoll quantifizieren.

Über zwei Werte in einer Ordinalskala lässt sich nicht nur sagen, ob sie gleich oder verschieden sind (wie in der Nominalskala), sondern darüber hinaus, welcher Wert bei Verschiedenheit größer ist.

Beispiele: Militärische Ränge, Windstärken, pauschale Häufigkeitsangaben (sehr oft … nie), Zufriedenheitsangaben (sehr zufrieden … unzufrieden)

1.2.3.3 Metrische Skalen (oder Kardinalskalen)

Abstände zwischen den Merkmalsausprägungen lassen sich exakt angeben.

Zusätzlich zu den Möglichkeiten der Ordinalskala können auf einer metrischen Skala Rechenoperationen auch sinnvoll auf die Differenzen zwischen den Merkmalsausprägungen angewendet werden.

Metrische Skalen werden unterteilt in Intervall- und Verhältnisskalen:

1.2.3.3.1 Intervallskala Maßeinheit und Wahl des Nullpunktes sind willkürlich gewählt.

Beispiele: Grad Celsius, Geburtsjahr als Jahreszahl („1961"), in der Praxis häufig: subjektive Bewertung auf einer Skala von 1 bis 10.

1.2.3.3.2 Verhältnisskala (auch Ratioskala) Es gibt einen invarianten (absoluten, natürlichen) Nullpunkt.

In einer Verhältnisskala lassen sich über alle o.a. Möglichkeiten hinaus auch Aussagen über Verhältnisse zwischen Werten treffen (z.B. „x_1 ist doppelt so groß wie x_2").

Beispiele: Lebensalter in Jahren, Haushaltsgröße, Köpergröße, Körpergewicht

1.3 Häufigkeitsverteilungen

1.3.1 Urliste

Die Urliste ist eine ungeordnete Liste aller erfassten Werte.

Für die statistische Erhebung „Anfangsbuchstaben der Vornamen von Teilnehmenden an einer Statistikvorlesung" könnte die Urliste z.B. so aussehen:

```
T J D T E N D F F M A J V T T V A L V P J K P M F M A J N A C I T P B A P H T L
N S P C K J K L J R E Y M K H M N L A A L L M L J G P L B F L J J V M P C J M J
```

```
S A M M M P A A L L O C J L P L V F J R M A V K S B B B N C A A T J P C F L E B
L C A K A L T V Y P F L J S T T N R J A S E L M L T A E B M N M V D P P L N L B
A A J M L N N S H M
```

1.3.2 Geordnete Liste

Die geordnete Liste bringt die Werte der Urliste in eine geeignete Reihenfolge, so dass die unterschiedlichen Werte leicht gezählt werden können:

```
A A A A A A A A A A A A A A A A A A A B B B B B B B B C C C C C C C D D D E E E
E E F F F F F F F G H H H I J J J J J J J J J J J J J J J J J K K K K K K L L L
L L L L L L L L L L L L L L L L L L L M M M M M M M M M M M M M M M M M N N N N
N N N N N N O P P P P P P P P P P P P P R R R S S S S S S T T T T T T T T T T T
V V V V V V V V Y Y
```

1.3.3 Häufigkeiten

Die absoluten Häufigkeiten erhält man durch einfaches Abzählen der jeweiligen Werte. Für die relativen Häufigkeiten teilt man diese Zahl durch n. Kumulierte Häufigkeiten zählen die bisherigen Summen bzw. Anteile zusammen (s. Tabelle 1.2).

Softwarehinweis

In R lässt sich mit dem Befehl `table()` eine einfache Häufigkeitstabelle aus Rohdaten erstellen.

1.3.4 Stabdiagramme

Die so ermittelten Häufigkeiten lassen sich als Stabdiagramm (auch Säulen-, Streifen-, Balkendiagramm, engl. *bar chart*) darstellen (s. Abbildung 1.1).

Softwarehinweis

In R lautet der Standardbefehl zur Erstellung eines Stabdiagramms `barplot()`.

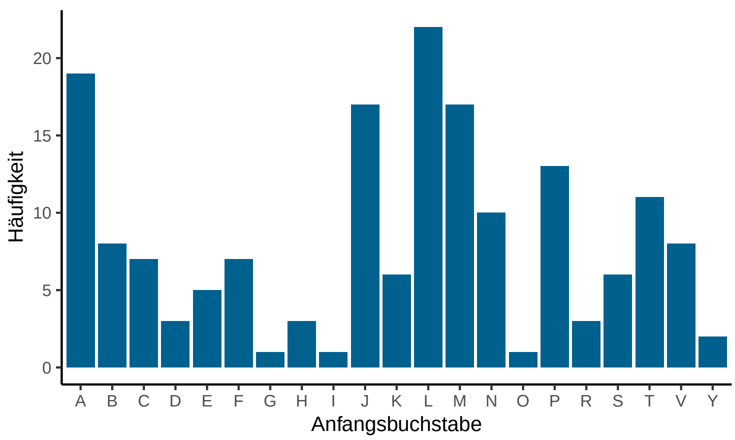


Abbildung 1.1: Stabdiagramm

Tabelle 1.2: Tabelle mit kumulierten Häufigkeiten

Buchstabe	**Absolute Häufigkeit** f	f_{kum}	**Relative Häufigkeit**	$\%_{kum}$
A	19	19	11,2%	11,2%
B	8	27	4,7%	15,9%
C	7	34	4,1%	20%
D	3	37	1,8%	21,8%
E	5	42	2,9%	24,7%
F	7	49	4,1%	28,8%
G	1	50	0,6%	29,4%
H	3	53	1,8%	31,2%
I	1	54	0,6%	31,8%
J	17	71	10%	41,8%
K	6	77	3,5%	45,3%
L	22	99	12,9%	58,2%
M	17	116	10%	68,2%
N	10	126	5,9%	74,1%
O	1	127	0,6%	74,7%
P	13	140	7,6%	82,4%
R	3	143	1,8%	84,1%
S	6	149	3,5%	87,6%
T	11	160	6,5%	94,1%
V	8	168	4,7%	98,8%
Y	2	170	1,2%	100%

Tabelle 1.3: Häufigkeitstabelle mit klassierten Werten

Durchmesser	**Absolute Häufigkeit** f	f_{kum}	**Relative Häufigkeit**	$\%_{kum}$
über 8 bis 10 Zoll	3	3	9,7%	9,7%
über 10 bis 12 Zoll	12	15	38,7%	48,4%
über 12 bis 14 Zoll	6	21	19,4%	67,7%
über 14 bis 16 Zoll	3	24	9,7%	77,4%
über 16 bis 18 Zoll	6	30	19,4%	96,8%
über 18 bis 20 Zoll	0	30	0%	96,8%
über 20 bis 22 Zoll	1	31	3,2%	100%

1.3.5 Quantitative Variablen

Das oben beschriebene Verfahren funktioniert gut für qualitative Variablen (und diskrete Variablen mit wenigen unterschiedlichen Werten). Für quantitative Variablen wird ein anderes Verfahren empfohlen.

Zur Veranschaulichung soll diese geordnete Liste von Messwerten des Stammdurchmessers von Schwarzkirschen (Beispieldatensatz `trees` aus R Core Team 2018) dienen:

```
8,3  8,6  8,8 10,5 10,7 10,8 11,0 11,0 11,1 11,2 11,3 11,4 11,4 11,7 12,0 12,9
12,9 13,3 13,7 13,8 14,0 14,2 14,5 16,0 16,3 17,3 17,5 17,9 18,0 18,0 20,6
```

Für solche Verteilungen müssen zuerst Klassen (engl. *bins*) gebildet werden, in denen die Werte dann zusammengefasst werden (s. Tabelle 1.3).

Für die Wahl der Klassengrenzen gibt es zwei feste Regeln:

- Alle Werte müssen abgedeckt sein.
- Die Klassen dürfen sich nicht überlappen.

Zusätzlich sollten die folgenden Konventionen nach Möglichkeit befolgt werden:

- Klassen sollten gleich große Wertebereiche abdecken.
- Alle Klassen sollten besetzt sein.
- Klassengrenzen sollten möglichst glatte Zahlen sein.
- Aus Gründen der Übersichtlichkeit sollten nicht mehr als 20 Klassen gewählt werden.
- Klassengrenzen sollten „Klumpen“ mit ähnlichen Werten nicht trennen.

Die Darstellung erfolgt in so genannten Histogrammen (engl. *histogram*). Abbildung 1.2 enthält ein Beispiel für ein Histogramm.

Softwarehinweis

In R können Histogramme mit `hist()` erstellt werden.

1.3.6 Polygone

Statt ausgefüllten Flächen wie im Histogramm lassen sich für die Häufigkeiten auch Punkte setzen, die dann mit Linien verbunden werden. So entsteht ein Häufigkeitspolygon (s. Abbildung 1.3).

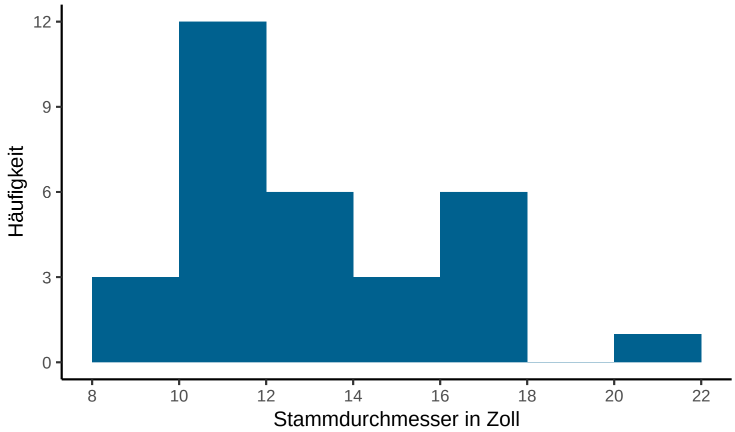


Abbildung 1.2: Histogramm

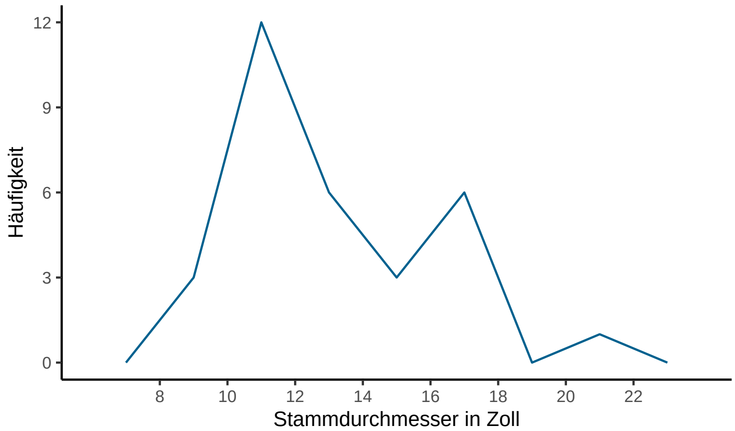


Abbildung 1.3: Polygonzug

1.3.7 Eigenschaften von Häufigkeitsverteilungen

Polygone von Häufigkeitsverteilungen (insbesondere in geglätteter Form) ergeben Annäherungen an so gennannte Dichtefunktionen (engl. *density functions*). Diese lassen sich mit Attributen (uni-/bimodal, schmal-/breitgipflig, etc.) beschreiben, wie in Abbildung 1.4 veranschaulicht.

Tipps zur Vertiefung

1.3.8 Grundbegriffe

- YouTube-Kanal „Kurzes Tutorium Statistik“: Statistische Grundbegriffe
- Kapitel 1.1 in Bortz und Schuster (2010)
- Kapitel 1.1 in Benninghaus (2007)
- Kapitel 2.1 in Bahrenberg, Giese und Nipper (2010)
- *Englisch:* Kapitel 1 in Burt und Barber (1996)

1.3.9 Stichproben

- Kapitel 6.1 in Bortz und Schuster (2010)

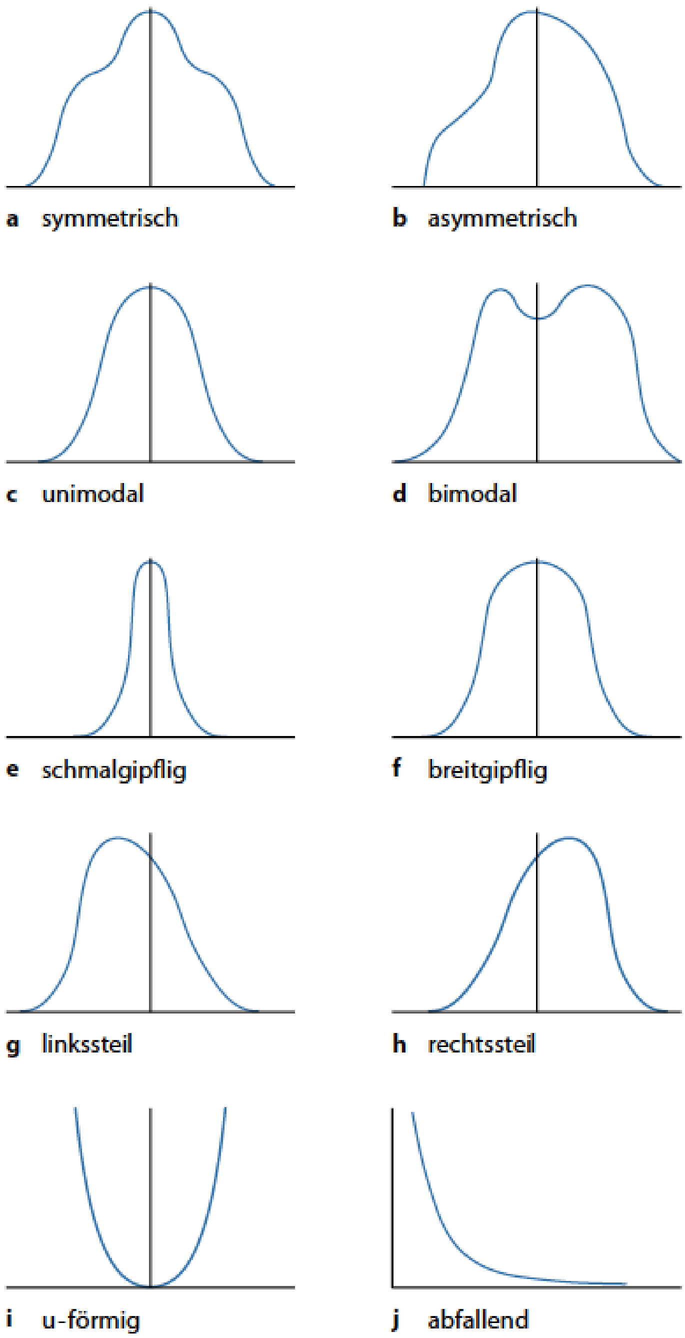


Abbildung 1.4: Merkmale von Verteilungen [aus: @bortz: 42]

- Kapitel 2.5 in Lange und Nipper (2018)
- Kapitel 2.3 in Bahrenberg, Giese und Nipper (2010)
- *Englisch:* Kapitel 1 in Burt und Barber (1996)

1.3.10 Skalenniveaus

- Kapitel 1.2 in Bortz und Schuster (2010)
- Kapitel 2.5 in Lange und Nipper (2018)

- Kapitel 2.1 in Benninghaus (2007)
- Kapitel 2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Skalenniveaus
- *Englisch:* Kapitel 1.3 in Burt und Barber (1996)

1.3.11 Häufigkeiten und Diagramme

- YouTube-Kanal „Kurzes Tutorium Statistik“: Stabdiagramme und Histogramme
- Kapitel 3.1 und 3.2 in Bortz und Schuster (2010)
- Kapitel 2.5 in Lange und Nipper (2018)
- Kapitel 1.2 in Benninghaus (2007)
- Kapitel 4.1 in Bahrenberg, Giese und Nipper (2010)
- *Englisch:* Kapitel 2.1 in Burt und Barber (1996)

Übungsaufgaben

1.3.12 Aufgabe 1-1

zur Lösung

Teilen Sie in Ihrer Kleingruppe folgende Begriffe untereinander auf:

- Variable
- Kennwert
- Wert
- Grundgesamtheit
- Stichprobe
- Untersuchungselement

Gehen Sie nun für jeden Begriff wie folgt vor:

1. Erklären Sie der Reihe nach „Ihren“ Begriff den anderen Gruppenmitgliedern, gerne auch mit Beispielen.
2. Die anderen Gruppenmitglieder nehmen die Rolle von unwissenden Dritten ein und stellen bei Bedarf Nachfragen.
3. Die anderen Gruppenmitglieder geben direkt danach Feedback auf die Erklärung:
    - Was fanden Sie gut erklärt?
    - Was fanden Sie unverständlich?
    - Was hat Ihnen gefehlt?

1.3.13 Aufgabe 1-2

zur Lösung

Finden Sie als Gruppe jeweils zwei Beispiele für:

- systematische Zufallsstichproben
- geschichtete Zufallsstichproben
- Klumpenstichproben

1.3.14 Aufgabe 1-3

zur Lösung

Bestimmen Sie das Skalenniveau der folgenden Variablen. Kennzeichnen Sie darüber hinaus, ob die Variable qualitativ, diskret oder stetig ist.

a) Lebensalter in Jahren
b) Regenmenge in mm
c) Güteklasse
d) Passagieraufkommen
e) Baujahr
f) Geschwindigkeit in km/h
g) Sozialstatus (Unter-, Mittel und Oberschicht)
h) Temperatur in °F
i) Fläche eines Bundeslands in km^2
j) Temperatur in K
k) Einwohnerzahl
l) Pegelstand
m) Staatsangehörigkeit
n) Interesse an Statistik (gering bis hoch)
o) Klausurnote
p) Bodentyp
q) Entfernung zum Stadtzentrum in km
r) Körpergröße
s) Kleidergröße (S bis XXL)
t) Monatliches Nettoeinkommen

1.3.15 Aufgabe 1-4

zur Lösung

Folgende Werte seien erfasst über die Lebensdauer von Klimaanlagen in Stunden (Beispieldatensatz `aircondit7` aus R Core Team 2018):

`14 23 15 139 13 39 188 22 50 3 36 46 30 5 102 5 88 22 197 72 210 97 79 44`

a) Erstellen Sie eine Häufigkeitstabelle. Welche Klassen wählen Sie und warum?
b) Zeichnen Sie ein Histogramm.
c) Beschreiben Sie die Verteilung.

1.3.16 Aufgabe 1-5

zur Lösung

Sind die folgenden Aussagen wahr oder unwahr?

a) Die Auswahl z. B. jedes 100. Merkmalsträgers nennt man „systematische Stichprobe".
b) Eine Stichprobe kann eine Grundgesamtheit niemals völlig richtig repräsentieren, es gibt immer einen Zufallsfehler.
c) Die Größe der Stichprobe wird auch mit N bezeichnet.
d) Klassengrenzen müssen so gewählt werden, dass alle Werte abgedeckt sind.

e) Je stärker die Werte der Variablen streuen, desto kleiner sollte die Stichprobe sein.
f) Variablen auf der Verhältnisskala sind immer metrisch und stetig.
g) Verhältnisskala und Intervallskala unterscheiden sich durch den natürlichen Nullpunkt.
h) Intervallskalierte Daten können immer auf die Nominalskala transformiert werden.
i) Ordinalskalierte Daten können immer auf die Intervallskala transformiert werden.
j) Eine stetige Variable ist nicht zwingend auch metrisch.
k) Im Gegensatz zu nominalskalierten Variablen lassen sich Werte von ordinalskalierten Variablen in eine sinnvolle Reihenfolge bringen.
l) Die relative Häufigkeit eines Werts ist nie größer als 100%.
m) Verfahren der deskriptiven Statistik sind immer auch univariat.
n) Klassengrenzen dürfen sich in Ausnahmefällen überlappen.
o) x_3 ist immer kleiner als x_4.
p) Variablen auf der Verhältnisskala haben einen natürlichen Nullpunkt.
q) Die absolute Häufigkeit eines Werts ist immer eine positive ganze Zahl.
r) Wenn man die Urliste ordnet, erhält man die geordnete Liste.

Sitzung 2

Maßzahlen

Lernziele dieser Sitzung

Sie können…

- die wichtigsten Lagemaße von Stichproben bestimmen.
- die wichtigsten Streumaße von Stichproben bestimmen.
- Boxplots interpretieren.

Lehrvideos (Sommersemester 2020)

- 2a) Lagemaße
- 2b) Streumaße
- 2c) Klassierte Verteilungen
  - In diesem Video ist mir ein Fehler unterlaufen: Bei Minute 6:30 muss das arithmetische Mittel $\bar{x} \approx 4{,}59$ betragen. Daraus ergibt sich ein Folgefehler: Die Varianz müsste den Wert $s^2 \approx 14{,}56$ haben.

2.1 Einleitende Bemerkungen

Die im Folgenden besprochenen Maßzahlen (oder Kennzahlen, Parameter) verdichten (oder aggregieren) Häufigkeitsverteilungen einer Variable. Durch diese Parameter kann das Charakteristische einer Verteilung schnell erfasst und vergleichbar gemacht werden. Die Verdichtung auf Maßzahlen geht jedoch immer auch mit Informationsverlust einher.

Die Möglichkeit der Angabe statistischer Maßzahlen ist abhängig vom Skalenniveau der Daten, wie der Überblick in Tabelle 2.1 zeigt.

2.1.1 Beispielverteilung

Alle Berechnungen von Maßzahlen werden am folgenden Beispiel illustriert: Für die 14 Gemeinden im Landkreis Rothenberge wurde die jeweilige Anzahl an Gaststätten erhoben. Die Zählung ergab die Wertereihe in Tabelle 2.2.

Tabelle 2.1: Die wichtigsten Maßzahlen

Parameter	**Typ**	**Mindestes Skalenniveau**	**Formel**
Modalwert	Lagemaß	nominal	Mo
Median	Lagemaß	ordinal	$Md = \begin{cases} \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2} & \\ x_{(\frac{n+1}{2})} & \text{fa} \end{cases}$
Arithmetisches Mittel	Lagemaß	metrisch	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Spannweite	Streumaß	ordinal	$R = x_{(n)} - x_{(1)}$
Quartilsabstand	Streumaß	ordinal	$IQR = Q_3 - Q_1$
Varianz	Streumaß	metrisch	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Standardabweichung	Streumaß	metrisch	$s = \sqrt{s^2}$

Tabelle 2.2: Beispielverteilung

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
4	1	4	1	5	5	0	1	8	5	1	25	3	3

Tabelle 2.3: Sortierte Wertereihe

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
0	1	1	1	1	3	3	4	4	5	5	5	8	25

2.2 Lagemaße

Lagemaße (auch Maße der Zentraltendenz, Lokalisationsparameter, Mittelwerte, engl. *measures of central tendency*) bezeichnen alle statistischen Maßzahlen, die eine Verteilung repräsentieren, indem sie die Lage der mittleren oder häufigsten Variablenwerte angeben.

Im Falle einer unimodalen, perfekt symmetrischen Verteilung (z. B. Glockenform) haben alle drei Lageparameter den gleichen Wert. Je weiter Verteilungen von dieser Form abweichen – durch Mehrgipfligkeit oder Asymmetrie – desto unpräziser ist die Beschreibung der Verteilung durch einen einzigen Parameter.

2.2.1 Median

Der Median (engl. *median*) einer Verteilung ist der Wert, der größer als genau 50% aller Werte ist.

Da dies eine Größer-kleiner-Relation der Werte voraussetzt, kann der Median nur für ordinale und metrische Skalenniveaus angegeben werden.

Im Folgenden wird die (einfachere) Bestimmung des Medians nach Bortz und Schuster (2010) verwendet. Benninghaus (2007) beschreibt ein anderes Verfahren, welches zu anderen Ergebnissen kommen kann.

Um den Median zu bestimmen, wird zunächst eine geordnete Liste angefertigt, indem die Werte aufsteigend sortiert werden. Diese sortierten Werte werden mit $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$ bezeichnet (also mit Klammern). Für unsere Beispielverteilung ergibt sich Tabelle 2.3.

Bei einer ungeraden Stichprobengröße n teilt der $(\frac{n+1}{2})$-te Wert (also der Wert genau in der Mitte) die Stichprobe in zwei Hälften, weshalb gilt:

$$Md = x_{(\frac{n+1}{2})} \quad \text{falls } n \text{ ungerade.}$$

Bei geradem n entstehen zwei gleich große Hälften der Stichprobe: $x_{(1)}$ bis $x_{(\frac{n}{2})}$ einerseits, und $x_{(\frac{n}{2}+1)}$ bis $x_{(n)}$ andererseits. Der Durchschnitt zwischen $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$ teilt die Stichprobe in zwei Hälften. Es gilt:

$$Md = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} \quad \text{falls } n \text{ gerade.}$$

In unserem Beispiel ist $n = 14$ und damit gerade. Der Median errechnet also nach Formel (2.2.1) wie folgt:

Tabelle 2.4: Häufigkeiten der Beispielverteilung

Wert x_i	**Häufigkeit** f_i
0	1
1	4
3	2
4	2
5	3
8	1
25	1

$$\begin{aligned} Md &= \frac{x_{(7)} + x_{(8)}}{2} \\ &= \frac{3+4}{2} \\ &= 3{,}5 \end{aligned}$$

Softwarehinweis

In R gibt die Funktion `median()` den Median einer Verteilung aus.

2.2.2 Modalwert

Der Modalwert Mo (auch Modus, engl. *mode*) gibt den häufigsten Wert oder die häufigsten Werte einer Verteilung an.

Der Modalwert kann so auch (als einziger Mittelwert) für nominalskalierte Variablen angegeben werden.

Bei ordinalen und metrischen Skalenniveaus sind folgende Besonderheiten zu beachten:

- Wird der Modus einer Verteilung durch unmittelbar benachbarte Werte gebildet, wird er als Kombination (bei metrischen Variablen als arithmetisches Mittel) dieser Werte angegeben.
- Bei bimodalen (multimodalen) Verteilungen werden beide (alle) Modalwerte angegeben.

Hierzu müssen die Häufigkeiten der Werte bekannt sein, bzw. bestimmt werden (s. Tabelle 2.4).

Der Modalwert der Beispielverteilung beträgt 1, da der Wert 1 am häufigsten (viermal) vorkommt.

2.2.3 Arithmetisches Mittel

Das arithmetische Mittel (auch Mittelwert, Durchschnitt, engl. *mean*) ist das gebräuchlichste Lagemaß und Grundlage für viele statistische Verfahren.

Das arithmetische Mittel setzt ein metrisches Skalenniveau voraus.

Die Berechnung des arithmetischen Mittels einer Stichprobe erfolgt durch die Formel:

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

Für unsere Beispielverteilung ergibt sich durch einsetzen in Formel (2.2.3):

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{14} x_i}{14} \\
&= \frac{4+1+4+1+5+5+0+1+8+5+1+25+3+3}{14} \\
&= \frac{63}{14} \\
&\approx 4{,}71
\end{aligned}$$

> **Softwarehinweis**
>
> Der Befehl für die Ermittlung des arithmetischen Mittels in R lautet `mean()`.

2.3 Streumaße

Streumaße (auch Streuungs-, Variabilitäts-, Dispersionswerte, engl. *measures of variability*) geben Auskunft darüber, wie heterogen die Werte einer Verteilung sind, d. h. wie breit sie gestreut sind. Während Lagemaße den typischen Wert einer Verteilung ermitteln, zeigen Streumaße, wie gut (oder eigentlich: wie schlecht) dieser typische Wert die Verteilung repräsentiert.

2.3.1 Spannweite

Die Spannweite (engl. *range*) gibt Auskunft darüber, wie groß der Wertebereich ist, der von einer Verteilung abgedeckt wird. Sie wird (für metrische Skalen) als die Differenz vom größten zum kleinsten Wert (also vom letzten zum ersten Wert einer geordneten Werteliste) angegeben:

$$R = x_{(n)} - x_{(1)}$$

Für unsere Beispielstichprobe ergibt sich (mit Blick auf Tabelle 2.3):

$$\begin{aligned}
R &= x_{(14)} - x_{(1)} \\
&= 25 - 0 \\
&= 25
\end{aligned}$$

> **Softwarehinweis**
>
> In R gibt die Funktion `range()` die Werte für $x_{(1)}$ und $x_{(n)}$ aus.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
0	1	1	1	1	3	3

$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$
4	4	5	5	5	8	25

2.3.2 Quartilsabstand

Der Quartilsabstand (auch Interquartilsabstand, engl. *interquartile range, IQR*) gibt die Größe des Wertebereichs der mittleren 50% einer Verteilung an.

Genau so wie der Median eine Messwertreihe in zwei gleich große Hälften „schneidet“, schneiden die Quartile die Werte in Viertel. Dabei liegt der so genannte untere Angelpunkt Q_1 genau über 25% der Werte, Q_2 ist identisch mit dem Median und der obere Angelpunkt Q_3 liegt genau über 75% der Werte.

Der Angelpunkt Q_1 wird ermittelt, indem der Median für die unteren 50% (Q_3: die oberen 50%) der Werte bestimmt wird – also jener Werte, die theoretisch unterhalb des Medians der Gesamtverteilung liegen.

Dabei folgen wir Bortz und Schuster (2010) und nehmen im Fall eines ungeraden n den Median auf beiden Seiten hinzu.

Die Formel für den Quartilsabstand lautet:

$$IQR = Q_3 - Q_1$$

Der Quartilsabstand ist Ausreißern gegenüber stabiler als die Spannweite, da extreme hohe oder niedrige Wert nicht in die Berechnung einfließen.

In unserem Beispiel (mit $n = 14$) ist die untere Hälfte der Verteilung:

Q_1 ist der Median dieser Werte, also $x_{(4)} = 1$.

Die oberen 7 Werte lauten:

Q_3 ist also $x_{(11)} = 5$.

Für den Quartilsabstand ergibt sich durch einsetzen in Formel (2.3.2):

$$\begin{aligned} IQR &= 5 - 1 \\ &= 4 \end{aligned}$$

Softwarehinweis

In R werden die Quartile üblicherweise mit `quantile()` und der Quartilsabstand mit `IQR()` bestimmt.

Tabelle 2.5: Häufigkeitstabelle zur Berechnung der Varianz

Werte x_i	**Häufigk.** f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
0	1	-4,71	22,18	22,18
1	4	-3,71	13,76	55,04
3	2	-1,71	2,92	5,84
4	2	-0,71	0,50	1,00
5	3	0,29	0,08	0,24
8	1	3,29	10,82	10,82
25	1	20,29	411,68	411,68

Achtung: Genau wie für den Median gibt es auch für die Ermittlung der Quartile bzw. des Quartilsabstands unterschiedliche Verfahren. Die Ergebnisse dieser R-Funktionen weichen hier deshalb meist leicht vom hier besprochenen Verfahren ab!

2.3.3 Varianz

Die Varianz einer Messwertreihe (engl. *variance*) kann verstanden werden als der durchschnittliche quadrierte Abstand der Werte zum arithmetischen Mittel.

Die Formel lautet:

$$s^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Die Quadrierung der Differenz hat dabei einen doppelten Effekt: Zum einen bekommen auch negative Differenzen ein positives Vorzeichen, so dass sich positive und negative Differenzen nicht neutralisieren. Zum anderen werden hierdurch besonders große Abweichungen zum arithmetischen Mittel stärker gewichtet als dies ohne Quadrierung der Fall wäre.

Zudem fällt auf, dass im Gegensatz zur Formel für das arithmetische Mittel im Nenner $n - 1$ steht und nicht etwa n. Dies hat mit so genannten Freiheitsgraden zu tun, die wir allerdings erst in Sitzung 5 genauer kennenlernen.

Für unsere Beispielstichprobe wird die Berechnung für alle einzelnen $(x_i - \bar{x})^2$ schnell aufwendig und unübersichtlich. Deshalb berechnen wir ihre Summe hier mit Hilfe einer Häufigkeitstabelle (s. Tabelle 2.5). Dabei werden alle distinkten Werte einzeln transformiert und in der letzten Spalte mit ihrer Häufigkeit multipliziert.

Schließlich werden die Werte in Formel (2.3.3) eingesetzt:

$$
\begin{aligned}
s^2 &= \frac{\sum\limits_{i=1}^{14}(x_i - \bar{x})^2}{14-1} \\
&\approx \frac{22{,}18 + 55{,}04 + 5{,}84 + 1 + 0{,}24 + 10{,}82 + 411{,}68}{13} \\
&= \frac{506{,}80}{13} \\
&\approx 38{,}98
\end{aligned}
$$

Eine solche Tabelle lässt sich analog auch für die Berechnung von Summen größerer Messwertreihen für das arithmetische Mittel verwenden.

Zudem lässt dieses Verfahren sich auf klassierte Daten anwenden, wenn für x_i der Mittelwert der Klassen eingesetzt wird (womit allerdings Informations- und Präzisionsverlust einhergeht).

Softwarehinweis

In R lautet der Befehl für die Errechnung der Varianz `var()`.

2.3.4 Standardabweichung

Die Standardabweichung (engl. *standard deviation*) ist das gebräuchlichste Streumaß und spielt eine herausragende Rolle in den allermeisten statistischen Verfahren.

Die Standardabweichung einer Messwertreihe ist definiert als die Quadratwurzel ihrer Varianz:

$$s = \sqrt{s^2}$$

Indem hier die Wurzel gezogen wird, wird in gewisser Weise die Quadrierung der Differenzen für die Varianz wieder „korrigiert“. Insbesondere wird die Quadrierung der Maßeinheit wieder aufgehoben – die Standardabweichung hat also die gleiche Einheit wie die Messreihe selbst.

In unserem Beispiel beträgt die Standardabweichung also:

$$s \approx \sqrt{38{,}98} \approx 6{,}24$$

Softwarehinweis

Die Standardabweichung wird in R mit der Funktion `sd()` berechnet.

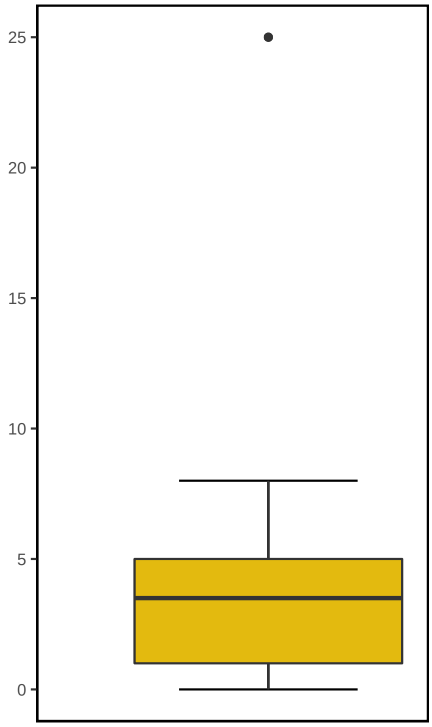


Abbildung 2.1: Boxplot der Beispielverteilung

2.4 Boxplot

Der Boxplot (auch Box-and-whisker-plot) kombiniert einige der gebräuchlichsten Maßzahlen in einer übersichtlichen Grafik (s. Abbildung 2.1).

Die Höhe der „Box" definiert sich durch den Quartilsabstand, der mittlere Strich markiert den Median und die „Whisker" markieren den Wertebereich insgesamt – wobei Ausreißer, deren Abstand zur Box mehr als das 1,5-Fache des Quartilsabstands beträgt, üblicherweise gar nicht oder (wie hier) gesondert mit Punkten markiert werden.

Softwarehinweis

In R lässt sich ein Boxplot mit dem Befehl `boxplot()` ausgeben.

Tipps zur Vertiefung

2.4.1 Lagemaße

- Kapitel 2.1 in Bortz und Schuster (2010)
- Kapitel 3.3.2 in Lange und Nipper (2018)
- Kapitel 3.3.1 in Benninghaus (2007)
- Kapitel 4.2.1 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik": Arithmetisches, harmonisches und geometrisches Mittel
- YouTube-Kanal „Kurzes Tutorium Statistik": Boxplots, Median, Quartile

- *Englisch:* Kapitel 2.2 in Burt und Barber (1996)

2.4.2 Streumaße

- Kapitel 2.2 in Bortz und Schuster (2010)
- Kapitel 3.3.3 in Lange und Nipper (2018)
- Kapitel 3.1.2 in Benninghaus (2007)
- Kapitel 4.2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Streumaße - Varianz, Standardabweichung, Variationskoeffizient und mehr!
- *Englisch:* Kapitel 2.3 in Burt und Barber (1996)

2.4.3 Boxplot

- Kapitel 3.4 in Bortz und Schuster (2010)
- Kapitel 5.3.1 in Lange und Nipper (2018)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Boxplots, Median, Quartile
- *Englisch:* Kapitel 16.3 in Burt und Barber (1996)

Übungsaufgaben

2.4.4 Aufgabe 2-1

zur Lösung

Berechnen Sie das arithmetische Mittel für die folgenden Verteilungen:

2.4.4.1 a)

`72 55 69 69 30 61`

2.4.4.2 b)

`0,759 0,296 0,687 0,7 -0,418 0,459 -0,4 -0,008`

2.4.4.3 c)

`951,73 859,29 937,4 939,96 716,45 891,83 719,92 798,38 864,21 670,99`

Tauschen Sie sich danach in der Lerngruppe darüber aus …

- Was schreiben Sie wann auf?
- Wie geben Sie die Zahlen und Rechenschritte in den Taschenrechner ein?
- Wie überprüfen Sie ggf. Ihr Ergebnis mit Hilfe des Taschenrechners?

2.4.5 Aufgabe 2-2

zur Lösung

Wiederholen Sie Aufgabe 1, aber berechnen Sie statt des arithmetischen Mittels die Standardabweichung (und tauschen sich darüber aus).

x	k_i	f_i	f_{kum}	$f_i \cdot k_i$
von 75 bis unter 77,5 cm	76,25	1	1	76,25
von 77,5 bis unter 80 cm	78,75	0	1	0,00
von 80 bis unter 82,5 cm	81,25	3	4	243,75
von 82,5 bis unter 85 cm	83,75	5	9	418,75
von 85 bis unter 87,5 cm	86,25	7	16	603,75
von 87,5 bis unter 90 cm	88,75	14	30	1242,50
von 90 bis unter 92,5 cm	91,25	9	39	821,25
von 92,5 bis unter 95 cm	93,75	2	41	187,50
von 95 bis unter 97,5 cm	96,25	2	43	192,50

2.4.6 Aufgabe 2-3

zur Lösung

Bei einer Befragung jedes 500. Studierenden im Matrikel einer privaten Hochschule wurden folgende Angaben zur Haushaltsgröße gemacht:

`1 4 4 2 3 2 3 5 2 7 2 1 1`

a) Welches Skalenniveau liegt vor? (Sitzung 1)
b) Berechnen Sie Modalwert,
c) Median und
d) arithmetisches Mittel der Stichprobe.
e) Berechnen Sie außerdem die Spannweite,
f) den Quartilsabstand,
g) die Varianz und
h) die Standardabweichung der Stichprobe.
i) Zeichnen Sie einen Boxplot der Stichprobenverteilung.

2.4.7 Aufgabe 2-4

zur Lösung

Eine Messreihe der Körperlänge weiblicher Beutelratten hat folgende Werte in cm erfasst (Beispieldatensatz `fossum` aus Maindonald und Braun 2020):

a) Wie groß ist der Quartilsabstand?
b) Bestimmen Sie das arithmetische Mittel der Reihe.
c) Berechnen Sie auch die Varianz und
d) die Standardabweichung.

2.4.8 Aufgabe 2-5

zur Lösung

In Wiesbaum soll ein Kulturzentrum entstehen. Zwei leerstehende Industriegebäude – eine Ziegelei und ein Möbellager – kommen für eine Umnutzung in Frage. Bei der Entscheidung, welches Gebäude umfunktioniert werden soll, spielt auch eine Rolle, welcher Ort ohnehin schon mehr Fußverkehr aufweist. Für beide Gebäude wurden daher jeweils die Anzahl der Passant*innen an sechs zufälligen Tagen erfasst:

Jahr	Niederschlag (mm)
1970	384,52
1971	493,65
1972	364,65
1973	661,32
1974	785,27
1975	603,45
1976	527,75
1977	471,81
1978	525,65
1979	455,64
1980	433,01
1981	535,12
1982	421,36
1983	499,29
1984	555,21
1985	398,88
1986	391,96
1987	453,41
1988	459,84
1989	483,78

$$\begin{array}{rrrrrrr} \text{Ziegelei}: & 75 & 91 & 86 & 77 & 78 & 104 \\ \text{Möbellager}: & 109 & 68 & 37 & 78 & 103 & 51 \end{array}$$

a) Welches Gebäude weist im Durchschnitt die höhere Passant*innenzahl auf?

b) Vergleichen Sie außerdem die Quartilsabstände der beiden Messreihen.

2.4.9 Aufgabe 2-6

zur Lösung

In Australien betrug die durchschnittliche Niederschlagsmenge in den 1970er- und 80er-Jahren [1]:

a) Welches Skalenniveau liegt vor? (Sitzung 1)
b) Legen Sie eine klassierte Häufigkeitstabelle an. Begründen Sie die Wahl der Klassen. (Sitzung 1)
c) Was ist der Modalwert der klassierten Verteilung?
d) Wie groß ist der Quartilsabstand?
e) Bestimmen Sie das arithmetische Mittel der klassierten Verteilung.
f) Berechnen Sie die Standardabweichung.
g) Zeichnen Sie einen Boxplot für die Verteilung.

[1] Auszug aus dem Datensatz `bomsoi` in Haseloff u. a. (1968)

Sitzung 3

z-Werte und Normalverteilung

Lernziele dieser Sitzung

Sie können…

- z-Werte ermitteln.
- Merkmale der Normalverteilung wiedergeben.
- anhand einer normalverteilten Dichtefunktion…
  - Wahrscheinlichkeiten errechnen.
  - Perzentile errechnen.

Lehrvideos (Sommersemster 2020)

- 3a) z-Transformation
- 3b) Normalverteilung
- 3c) Quantile der Normalverteilung

3.1 Variationskoeffizient

Die Berechnung von Maßzahlen (Sitzung 2) vereinfacht es uns, auch große Verteilungen miteinander zu vergleichen. Voraussetzung dafür ist jedoch, dass die Kennwerte (wie arithmetisches Mittel, Standardabweichung) in derselben Maßeinheit (kg, cm, °C, etc.) vorliegen und einen vergleichbaren Maßstab haben.

Eine Möglichkeit, unabhängig hiervon eine Aussage über die *relative* Streuung zu treffen, ist der Variationskoeffizient (engl. *coefficient of variation*) v. Er ist definiert als das (prozentuale) Verhältnis von Standardabweichung zu Mittelwert:

$$v = \frac{s}{|\bar{x}|} \cdot 100\%$$

Zur Illustration: An zufälligen Tagen hat die Wetterstation auf dem Feldberg folgende Luftdruckwerte gemessen (in hPa):

```
1007,1  1003,4   990,7   994,2  1000,9   993,0  1016,0   983,9  1007,4   997,8
 997,9  1000,2
```

Mit den bekannten Methoden (Sitzung 2) können wir das arithmetische Mittel $\bar{x} \approx 999{,}37$ und die Standardabweichung $s \approx 8{,}56$ der Stichprobe bestimmen. Durch einsetzen dieser Werte in Formel (3.1) ergibt sich:

$$v \approx \frac{8{,}56}{999{,}37} \cdot 100\%$$
$$\approx 0{,}86\%$$

Da die Standardabweichung im Vergleich zu den absoluten Werten sehr klein ist, ist der Variationskoeffizient hier sehr klein.

Ein Problem ergibt sich, wenn der Mittelwert einer Verteilung nahe Null liegt (z. B. wenn die Reihe auch negative Messwerte enthält). Der Variationskoeffizient wird in diesem Fall sehr groß und verliert stark an Aussagekraft.

3.2 z-Transformation

Ein weiterer Ansatz, Verteilungsmuster vergleichbar zu machen, ist die z-Transformation (auch Standardisierung, engl. *standardization*).

Für jeden der Messwerte lässt sich ein entsprechender z-Wert mit dieser Formel errechnen:

$$z = \frac{x - \bar{x}}{s}$$

Der z-Wert eines Werts x ist also der Abstand des Werts zum arithmetischen Mittel $\bar{x}$ der Verteilung, ausgedrückt im Verhältnis zu ihrer Standardabweichung s.

Die einzelnen z-Werte für die Luftdruckmessungen ergeben sich wie in Tabelle 3.1 dargestellt.

Eine so z-transformierte Verteilung hat *immer* automatisch das arithmetische Mittel $\bar{z} = 0$ und die Standardabweichung $s_z = 1$. Außerdem haben z-Werte keine Maßeinheit. So kann jede Verteilung „standardisiert“ und systematisch vergleichbar gemacht werden.

Softwarehinweis

In R kann eine empirische Verteilung mit dem Behfehl `scale()` z-transformiert werden.

Andersherum lassen sich z-Werte folgendermaßen wieder umwandeln in x-Werte:

$$x = s \cdot z + \bar{x}$$

3.3 Normalverteilung

Die Normalverteilung (auch: Gaußverteilung, engl. *normal distribution*) ist unimodal und symmetrisch. Die Normalverteilung ist eine theoretische Verteilung, für die bekannt ist, mit welcher Wahrscheinlichkeit bestimmte Werte unter- und überschritten werden bzw. mit welcher Wahrscheinlichkeit Werte in einem bestimmten Intervall liegen.

Tabelle 3.1: z-Transformation

x_i	**Berechnung**	z_i
1007,1	$z_1 = \frac{1007{,}1-999{,}37}{8{,}56}$	0,90
1003,4	$z_2 = \frac{1003{,}4-999{,}37}{8{,}56}$	0,47
990,7	$z_3 = \frac{990{,}7-999{,}37}{8{,}56}$	-1,01
994,2	$z_4 = \frac{994{,}2-999{,}37}{8{,}56}$	-0,60
1000,9	$z_5 = \frac{1000{,}9-999{,}37}{8{,}56}$	0,18
993,0	$z_6 = \frac{993-999{,}37}{8{,}56}$	-0,74
1016,0	$z_7 = \frac{1016-999{,}37}{8{,}56}$	1,94
983,9	$z_8 = \frac{983{,}9-999{,}37}{8{,}56}$	-1,81
1007,4	$z_9 = \frac{1007{,}4-999{,}37}{8{,}56}$	0,94
997,8	$z_{10} = \frac{997{,}8-999{,}37}{8{,}56}$	-0,18
997,9	$z_{11} = \frac{997{,}9-999{,}37}{8{,}56}$	-0,17
1000,2	$z_{12} = \frac{1000{,}2-999{,}37}{8{,}56}$	0,10

Tabelle 3.2: Bezeichnung von Parametern in Stichprobe und Grundgesamtheit

Parameter	**Stichprobe**	**Grundgesamtheit**
Anzahl Elemente	n	N
Arithmetisches Mittel	$\bar{x}$	μ
Varianz	s^2	σ^2
Standardabweichung	s	σ

Die Dichtefunktion einer Normalverteilung hat eine markante Glockenform (s. Abbildungen 3.1 und 3.2). Die beiden Wendepunkte einer Normalverteilung (also dort, wo die Steigung zwischen zu- und abnehmend wechselt; oder mathematisch: wo die Ableitung der Dichtefunktion einen Extremwert annimmt) sind je eine Standardabweichung vom Mittelwert entfernt.

Die Dichtefunktion nimmt nie den Wert Null an – Extremwerte sind also sehr selten bzw. unwahrscheinlich, aber nie unmöglich. Perfekte Normalverteilungen kommen in empirischen Beobachtungen nicht vor, sondern nur Annäherungen.

Da es sich um eine *theoretische* Verteilung handelt, ist die Normalverteilung zunächst insbesondere in Bezug auf die Grundgesamtheit interessant. Im Kontext der Grundgesamtheit wird das arithmetische Mittel mit μ („Mü“) und die Standardabweichung mit σ („Sigma“) bezeichnet (s. Tabelle 3.2).

Jede Normalverteilung lässt sich anhand von zwei Parametern beschreiben: ihr arithmetisches Mittel und ihre Standardabweichung. Normalverteilte Grundgesamtheiten werden so notiert:

$$x \sim N(\mu,\ \sigma^2)$$

Der Mittelwert μ bestimmt die Lage der Kurve auf der x-Achse, die Varianz σ^2 bestimmt die „Stauchung“ der Kurve (je größer desto flacher). Es gibt also unendlich viele verschiedene Normalverteilungen (s. Abbildung 3.1).

3.4 Standardnormalverteilung

Die Standardnormalverteilung (engl. *standard normal distribution*) ist sozusagen das Grundmuster aller Normalverteilungen. Sie hat den Mittelwert $\mu = 0$ und die Standardabweichung $\sigma = 1$ (s. Abbildung 3.2).

Alle Normalverteilungen lassen sich durch die z-Transformation auf die Standardnormalverteilung standardisieren.

3.5 Crash-Kurs Wahrscheinlichkeitsrechnung

Ein Zufallsexperiment ist ein beliebig oft wiederholbarer, nach bestimmten Vorschriften ausgeführter Versuch, dessen Ergebnis zufallsbedingt ist (d. h. nicht eindeutig voraussagbar ist).

Jedem zufälligen Ereignis A ist eine bestimmte „Wahrscheinlichkeit des Auftretens“ (engl. *probability*) $P(A)$ zugeordnet, die der Ungleichung $0 \leq P(A) \leq 1$ genügt (d. h. zwischen 0 und 1 liegt).

Die Wahrscheinlichkeit eines sicheren Ergebnisses A ist $P(A) = 1$. Hingegen würde $P(B) = 0$ bedeuten, dass das Ereignis B nicht eintreten kann. Die Summe der Wahrscheinlichkeiten aller möglichen Ereignisse beträgt 1.

Der *Additionssatz* besagt: Die Wahrscheinlichkeit, dass eins von verschiedenen zufälligen, sich gegenseitig ausschließenden Ereignissen eintritt, ist die Summe ihrer Wahrscheinlichkeiten.

Der *Multiplikationssatz* besagt: Die Wahrscheinlichkeit für das Eintreten zweier voneinander unabhängiger Ereignisse ist gleich dem Produkt der Einzelwahrscheinlichkeiten.

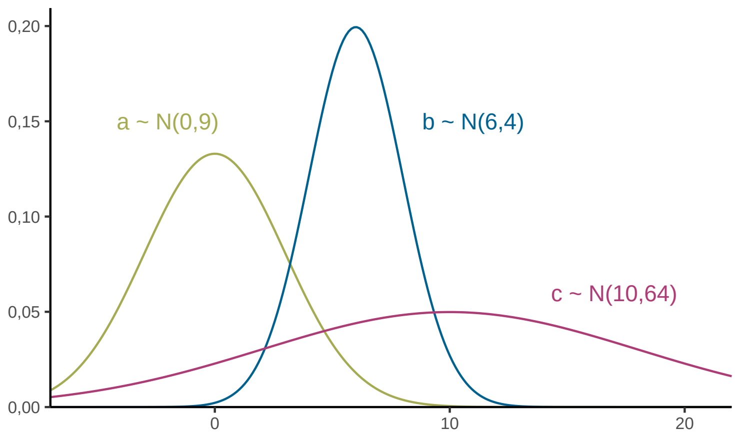


Abbildung 3.1: Dichtefunktionen verschiedener Normalverteilungen

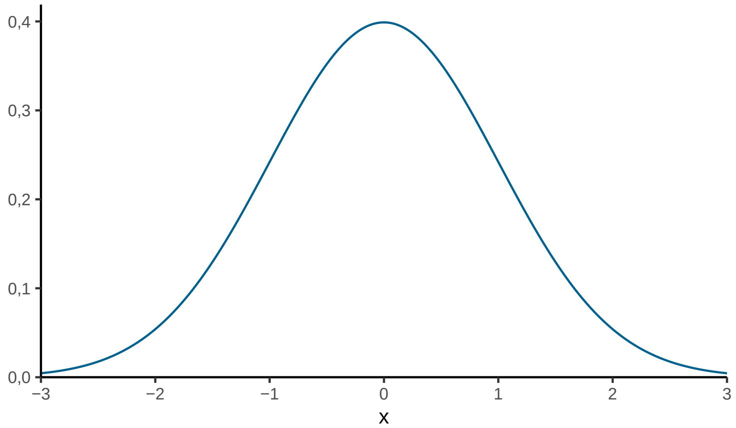


Abbildung 3.2: Dichtefunktion der Standardnormalverteilung

3.6 Wahrscheinlichkeitsdichtefunktionen

Die Fläche unter einer Wahrscheinlichkeitsdichtefunktion (engl. *probability density function*) beträgt genau 1.

Das Perzentil x_p (engl. *percentile*) ist definiert als der Wert, unter dem der Anteil p der Verteilung liegt. In Sitzung 2 haben wir also bereits den Median $x_{50\%}$ sowie die Angelpunkte $Q_1 = x_{25\%}$ und $Q_3 = x_{75\%}$ kennengelernt.

Die Fläche unter einer Wahrscheinlichkeitsdichtefunktion innerhalb der Limits $-\infty$ und x_p beträgt p. Für einen zufälligen Wert x ist die Wahrscheinlichkeit $P(x < x_p) = p$, dass er kleiner als x_p ausfällt. Für die Standardnormalverteilung finden sich die p-Werte für positive z in der Wertetabelle in der Formelsammlung.[1]

3.7 Wahrscheinlichkeitsrechnung mit Standardnormalverteilung

Für die im Rest dieser Sitzung vorgestellten Verfahren müssen folgende Voraussetzungen gegeben sein:

- Die Grundgesamtheit ist (annähernd) normalverteilt.
- Arithmetisches Mittel μ und Standardabweichung σ der Grundgesamtheit sind bekannt.

Die Verfahren sollen anhand eines Beispiels illustriert werden: Es sei bekannt, dass der Luftdruck auf dem Feldberg annähernd normalverteilt ist, und zwar mit dem arithmetischen Mittel $\mu = 1003$ und Varianz $\sigma^2 = 73$. Graphisch stellt sich die Wahrscheinlichkeitsdichtefunktion wie in Abbildung 3.3 dar.

Wir können auch (analog zu Formel (3.3)) schreiben:

$$x \sim N(1003,\ 73)$$

[1]Manchmal wird die Funktion $z_p \to P(z < z_p)$ für normalverteilte Werte auch mit $\Phi(z)$ bezeichnet (z. B. in Bahrenberg, Giese und Nipper 2010).

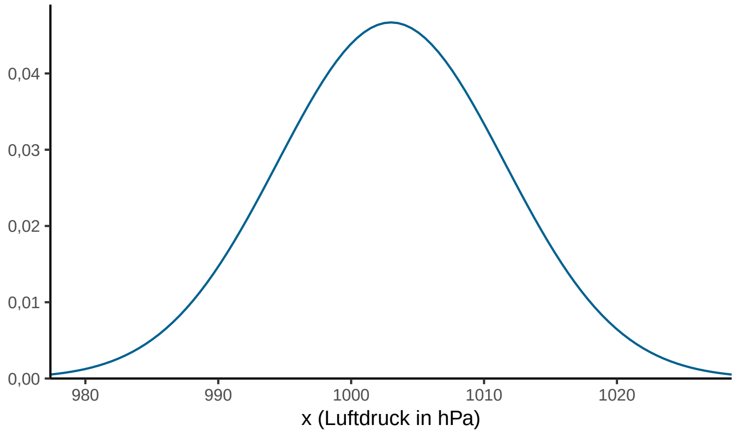


Abbildung 3.3: Theoretische Wahrscheinlichkeitsdichtefunktion des Luftdrucks

Daraus ergibt sich für die Standardabweichung σ:

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{73} \\ &\approx 8{,}54\end{aligned}$$

3.7.1 Unterschreitungswahrscheinlichkeit

Die einfachste Art der Fragestellung ist nun, mit welcher Wahrscheinlichkeit ein bestimmter Wert x_p unterschritten wird.

Nehmen wir an, es sei gefragt, mit welcher Wahrscheinlichkeit zu einem beliebigen Zeitpunkt der Luftdruck weniger als 1015 hPa beträgt. Anders gesagt interessiert uns der Anteil der Fläche unter der Verteilung, der zwischen $-\infty$ und $x_p = 1015$ liegt (s. Abbildung 3.4).

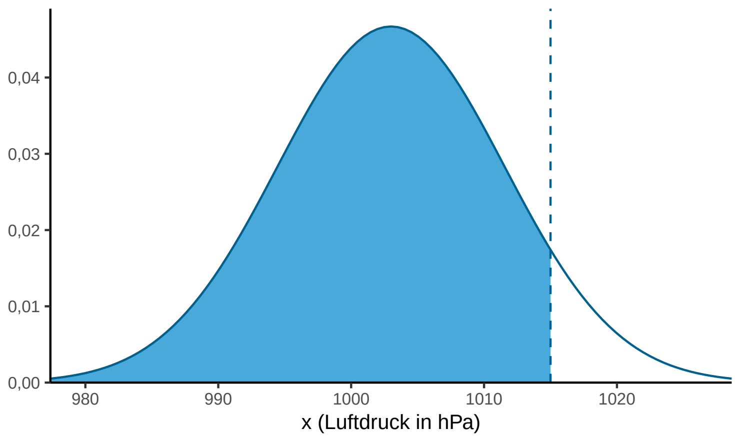


Abbildung 3.4: Unterschreitung eines Messwerts

Um den entsprechenden Wert für $P(x < x_p)$ (also die Wahrscheinlichkeit, dass ein zufälliges x unser Perzentil x_p unterschreitet) in Erfahrung zu bringen, müssen wir die Verteilung zunächst standardisieren.

Der Wert z_p ergibt sich aus der Formel für die z-Transformation, diesmal jedoch mit μ statt $\bar{x}$ und σ statt s, da es sich um die Grundgesamtheit handelt:

$$\begin{aligned} z_p &= \frac{x_p - \mu}{\sigma} \\ &\approx \frac{1015 - 1003}{8{,}54} \\ &\approx 1{,}41 \end{aligned}$$

Graphisch ist das standardisierte Perzentil in Abbildung 3.5 dargestellt.

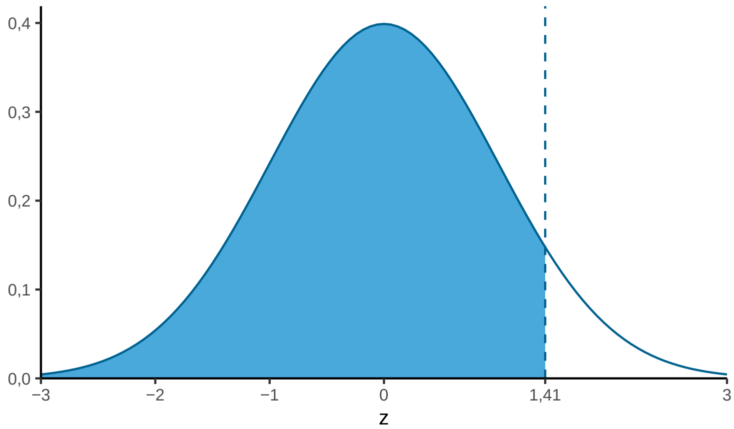


Abbildung 3.5: Standardnormalverteilung des Luftdrucks

Die Wertetabelle für die Standardnormalverteilung gibt für z-Werte die Wahrscheinlichkeit ihrer Unterschreitung in ener Normalverteilung an. Diese Wahrscheinlichkeit kann notiert werden als $P(z < z_p)$.

Der Wertetabelle können wir den Wert $P(z < 1{,}41) \approx 0{,}9207$ entnehmen. Die Wahrscheinlichkeit, dass der Luftdruck zu einem zufälligen Zeitpunkt weniger als 1015 hPA beträgt, ist somit 92,07%.

Softwarehinweis

In R lässt sich die Unterschreitungswahrscheinlichkeit eines z-Werts mit dem Befehl `pnorm()` ermitteln.

3.7.1.1 Überschreitungswahrscheinlichkeit

Wird nach der Wahrscheinlichkeit der Überschreitung eines Werts gefragt, ist in anderen Worten die Fläche unter der Wahrscheinlichkeitsdichtefunktion zwischen x_p und ∞ gemeint. Wir bleiben bei unserem Beispiel $x_p = 1015$ (s. Abbildung 3.6).

Hier können wir genauso wie bei der Unterschreitung $z_p = 1{,}41$ errechnen.

Jetzt stehen wir zunächst vor dem Problem, dass die p-Werte in der Tabelle immer die Wahrscheinlichkeit der Unterschreitung darstellen. Wir wissen jedoch: Die gesamte Fläche unter der Verteilung ist 1, und die Wahrscheinlichkeiten der Unter- und Überschreitung sind komplementär, d. H. einer von

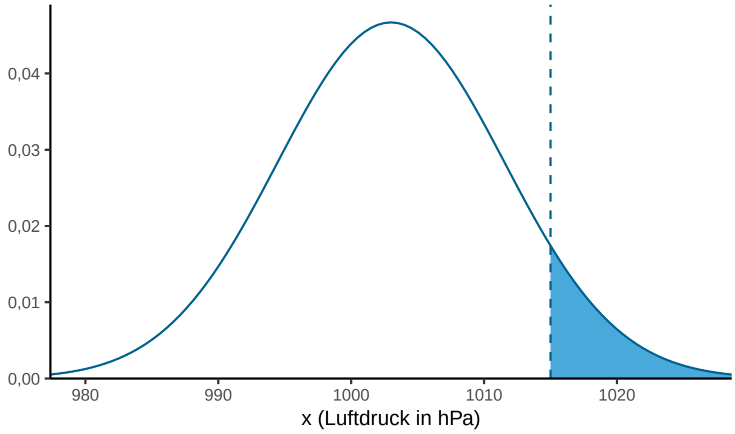


Abbildung 3.6: Überschreitung eines Messwerts

beiden Fällen tritt sicher (mit einer Wahrscheinlichkeit von 100%) ein. (Den Sonderfall $x = x_p$ können wir bei stetigen Variablen vernachlässigen.)

Hieraus ergibt sich ganz allgemein:

$$P(x \geq x_p) = 1 - P(x < x_p)$$

Und für unser Beispiel:

$$\begin{aligned} P(x \geq 1015) &= 1 - P(x < 1015) \\ &\approx 1 - P(z < 1{,}41) \\ &\approx 1 - 0{,}9207 \\ &= 0{,}0793 \end{aligned}$$

In 7,93% der Fälle beträgt der Luftdruck also über 1015 hPA.

3.7.1.2 Negativer z-Wert

Wenn nach der Unterschreitungswahrscheinlichkeit eines unterdurchschnittlichen Werts gefragt ist (z. B. 990 hPA), dann ergibt sich ein negativer Wert für z_p:

$$\begin{aligned} z_p &= \frac{x_p - \mu}{\sigma} \\ &= \frac{990 - 1003}{8{,}54} \\ &\approx -1{,}52 \end{aligned} \tag{3.1}$$

Die Wertetabelle enthält keine p für negative z_p. Da die Standardnormalverteilung jedoch um $z = 0$ symmetrisch ist, gilt ganz allgemein:

$$P(z < -z_p) = 1 - P(z < z_p)$$

Für unser Beispiel ergibt sich (mit dem Wert $P(z < 1{,}52) = 0{,}9357$ aus der Tabelle):

$$\begin{aligned} P(z < -1{,}52) &= 1 - P(z < 1{,}52) \\ &\approx 1 - 0{,}9357 \\ &= 0{,}0643 \end{aligned}$$

Ein Luftdruck von 990 hPa wird also nur in ca. 6,43% der Fälle unterschritten.

Softwarehinweis

Der Befehl `pnorm()` funktioniert auch mit negativen z-Werten.

3.7.1.3 Wert in einem Intervall

Nun wollen wir wissen, mit welcher Wahrscheinlichkeit ein zufälliger Meßwert zwischen 1005 und 1015 hPa liegt. Graphisch ist dies in Abbildung 3.7 aufbereitet.

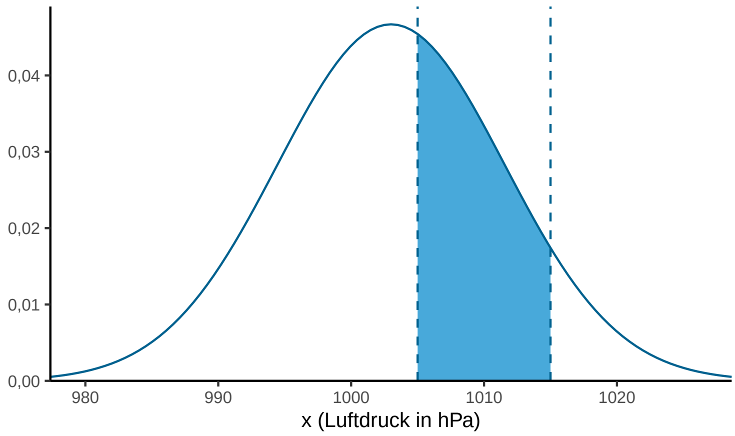


Abbildung 3.7: Messwertintervall

Rechnerisch müssen wir also von den (günstigen) Fällen, in denen 1015 hPA unterschritten werden, noch jene (ungünstige) Fälle abziehen, in denen die 1005 hPA *ebenfalls* unterschritten werden.

Ganz allgemein heißt das für die Untergrenze x_u und die Obergrenze x_o:

$$P(x_u \leq x < x_o) = P(x < x_o) - P(x < x_u)$$

Für unseren Fall ist $x_u = 1005$ und $x_o = 1015$. In den vorherigen Aufgaben haben wir $z_o \approx 1{,}41$ bereits ermittelt. Wir müssen aber noch z_u ermitteln:

$$\begin{aligned} z_u &= \frac{x_u - \mu}{\sigma} \\ &= \frac{1005 - 1003}{8{,}54} \\ &\approx 0{,}23 \end{aligned}$$

Dann können wir die entsprechende Wahrscheinlichkeit berechnen, indem wir wieder die Werte aus der Wertetabelle einsetzen:

$$
\begin{aligned}
P(1005 \leq x < 1015) &= P(x < 1015) - P(x < 1005) \\
&\approx P(z < 1{,}41) - P(z < 0{,}23) \\
&\approx 0{,}9207 - 0{,}5910 \\
&= 0{,}3297
\end{aligned}
$$

Der Luftdruck liegt also mit einer Wahrscheinlichkeit von 32,97% zwischen 1005 und 1015 hPa.

3.7.1.4 Gesuchter Wert bei gegebener Wahrscheinlichkeit

Die Fragerichtung lässt sich umdrehen: Welche Marke wird beim Messen des Luftdrucks nur in 5% der Fälle überschritten?

5% Überschreitungswahrscheinlichkeit entsprechen einer Unterschreitungswahrscheinlichkeit von 95%. Welcher Wert wird also mit 95% Wahrscheinlichkeit unterschritten?

Der Tabelle entnehmen wir, dass einer Unterschreitungswahrscheinlichkeit von 0,95 ein z-Wert zwischen 1,64 und 1,65 entspricht. Da es bei dieser Fragestellungen oft darum geht, einen „kritischen" Wert zu nennen, der nur in Ausnahmefällen überschritten wird, nehmen wir hier üblicherweise den extremeren Wert, also $z_{95\%} \approx 1{,}65$.

Mit der umgekehrten z-Transformation erhalten wir:

$$
\begin{aligned}
x_{95\%} &= z_{95\%} \cdot \sigma + \mu \\
&\approx 1{,}65 \cdot 8{,}54 + 1003 \\
&\approx 1017{,}10
\end{aligned}
$$

Die Marke von 1017,10 hPa wird also nur in 5% der Fälle überschritten.

Softwarehinweis

Das Perzentil für eine gegebene Unterschreitungswahrscheinlichkeit lässt sich in R mit `qnorm()` bestimmen.

3.7.1.5 Gesuchte Grenzwerte eines Intervalls

Eine übliche Art der Fragestellung ist auch: Zwischen welchen beiden Werten liegen die mittleren 85% der Fälle (s. Abbiddung 3.8)?

Da die Verteilung symmetrisch ist, teilen sich die ungünstigen 15% der Fälle gleichmäßig an den oberen und unteren Rand der Verteilung auf. Die Obergrenze x_o ist also der Wert, der zu 7,5% über- und damit zu 92,5% unterschritten wird.

Der Tabelle entnehmen wir den Wert $z_o = z_{92{,}5\%} \approx 1{,}44$.

Die Untergrenze ist entsprechend der Wert, der in 7,5% der Fälle unterschritten wird.

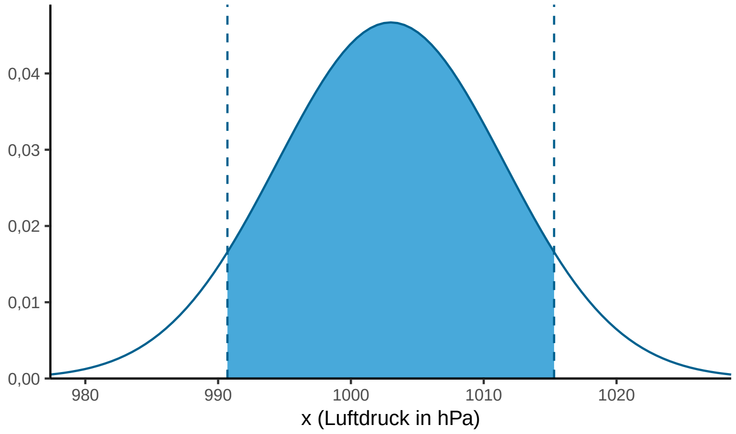


Abbildung 3.8: Die mittleren 85% der Normalverteilung

Der Wert für $z_u = z_{7{,}5\%}$ ist in der Tabelle nicht enthalten. Weil die Verteilung aber symmetrisch ist, wissen wir uns zu helfen:

$$z_u = z_{7{,}5\%} = -z_{92{,}5\%} \approx -1{,}44$$

Die absoluten Werte ergeben sich schließlich aus:

$$\begin{aligned} x_u &= z_u \cdot \sigma + \mu \\ &\approx -1{,}44 \cdot 8{,}54 + 1003 \\ &\approx 990{,}70 \end{aligned}$$

Und:

$$\begin{aligned} x_o &= z_o \cdot \sigma + \mu \\ &\approx 1{,}44 \cdot 8{,}54 + 1003 \\ &\approx 1015{,}30 \end{aligned}$$

Die mittleren 85% der Messwerte liegen also zwischen 990,7 und 1015,3 hPa.

Tipps zur Vertiefung

3.7.2 Variationskoeffizient

- Kapitel 3.3.4 in Lange und Nipper (2018)
- Kapitel 4.2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Streumaße - Varianz, Standardabweichung, Variationskoeffizient und mehr!
- *Englisch:* Kapitel 2.3 in Burt und Barber (1996)

3.7.3 z-Transformation

- Kapitel 2.4 in Bortz und Schuster (2010)
- Kapitel 3.5.2 in Lange und Nipper (2018)
- Kapitel 4.2.2 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 3.3.3 in Benninghaus (2007)
- YouTube-Kanal „Methodenlehre Mainz“: WT.012.09 Äpfel mit Birnen vergleichen: Die z-Standardisierung
- *Englisch:* Kapitel 6.3 in Burt und Barber (1996)

3.7.4 Normalverteilung

- Kapitel 5.4 in Bortz und Schuster (2010)
- Kapitel 7.3.2.2 und 7.3.2.3 in Lange und Nipper (2018)
- Kapitel 5.2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Mathe by Daniel Jung“: Was ist die Normalverteilung, Gauß-Verteilung, Schaubilder, Übersicht
- *Englisch:* Kapitel 6.3 in Burt und Barber (1996)

3.7.5 Wahrscheinlichkeitsdichtefunktion

- Kapitel 5.3 in Bortz und Schuster (2010)
- Kapitel 7.3.2.1 in Lange und Nipper (2018)
- Kapitel 5.2.2 in Bahrenberg, Giese und Nipper (2010)
- YouTube-Kanal „Kurzes Tutorium Statistik“: Zufallsvariable, Massenfunktion, Dichtefunktion und Verteilungsfunktion
- *Englisch:* Kapitel 6.1 in Burt und Barber (1996)

Übungsaufgaben

3.7.6 Aufgabe 3-1

zur Lösung

a) Führen Sie eine z-Transformation der folgenden Verteilung durch:

   `-16,93  -16,09  -10,97  -3,77  -25,55  -20,57  -23,61  -25,9  -27,08`

b) Sie kennen das arithmetische Mittel (221,54) und die Varianz (13,02) einer Verteilung. Welche x-Werte entsprechen diesen z-Werten?

   `0,9  -1,4  1,12  -0,33  2,22  0,15  2,87  0,4  -1,54  0,13  -0,17  0,68`

3.7.7 Aufgabe 3-2

zur Lösung

Gegeben sei eine Normalverteilung beschrieben durch:

$$x \sim N(32{,}2,\ \ 19{,}36)$$

a) Mit welcher Wahrscheinlichkeit werden die folgenden Werte unterschritten?

   `40,63   20,77   33,41   44,95   41,91   32,95`

b) Welche Werte werden jeweils mit der folgenden Wahrscheinlichkeit über(!)schritten?

   `1,5%   2,5%   5%   13%   50%   90%   99%   99,5%`

c) In welchem Bereich liegen die mittleren 95% der Werte?

d) Wie wahrscheinlich ist es, dass ein Wert zwischen 30 und 40 liegt?

3.7.8 Aufgabe 3-3

zur Lösung

Deiche werden durch Wasserdruck bei Hochwasser belastet und dadurch beschädigt. Bei einem 12 m hohen Deich gilt als kritische Marke ein Wasserstand von 10 m. Die jährlichen Höchstwasserstände des Flusses sind normalverteilt mit einem Mittelwert von 9,01 m und einer Standardabweichung von 2,23 m.

In den folgenden Teilaufgaben beantworten wir Schritt für Schritt die Frage, wie wahrscheinlich es (für ein beliebiges Jahr) ist, dass der Deich das jährliche Hochwasser ohne Beschädigung übersteht, d. h. dass ein Höchstwasserstand von 10 m oder weniger eintritt.

a) Zeichnen Sie die Wahrscheinlichkeitsdichtefunktion (ganz grob, ohne y-Achse).
b) Markieren Sie den kritischen Wert 10 m.
c) Welchem z-Wert entspricht die kritische Marke von 10 ?
d) Mit welcher Wahrscheinlichkeit bleibt der Deich in einem gegebenen Jahr unbeschädigt (Höchstwasserstand unter der kritischen Marke von 10 m)?

3.7.9 Aufgabe 3-4

zur Lösung

Wir bleiben beim Deich aus Aufgabe 3.

a) Mit welcher Wahrscheinlichkeit wird der Deich beschädigt (Wasserstand über 10 m)?
b) Mit welcher Wahrscheinlichkeit wird der Deich nicht nur beschädigt, sondern läuft über (Wasserstand über 12 m)?
c) Mit welcher Wahrscheinlichkeit wird der Deich beschädigt, läuft aber nicht über (Wasserstand zwischen 10 und 12 m)?
d) In welchen Grenzen liegen die mittleren 80% der Hochwasserstände?

3.7.10 Aufgabe 3-5

zur Lösung

Es ist ein neuer Deich zu bauen, der so sicher sein soll, dass er nur alle 200 Jahre vom Hochwasser übertreten wird.

a) Welcher Wahrscheinlichkeitswert $p = P(x < x_p)$ ist anzuwenden, d. h. wie wahrscheinlich ist die *Unterschreitung* eines „zweihundertjährigen Hochwassers“?
b) Mit welchem z-Wert korrespondiert der gesuchte Wert x_p?
c) Wie hoch muss dieser Deich sein? (Welcher Wert x_p entspricht diesem z_p?)

3.7.11 Aufgabe 3-6

zur Lösung

Die jährlichen Niederschlagsmengen in Mittelstedt betragen im Durchschnitt 400 mm bei annähernder Normalverteilung und einer Standardabweichung von 100 mm.

a) Wie groß ist die Wahrscheinlichkeit, dass mehr als 500 mm Niederschlag fallen?
b) Wie oft pro hundert Jahre kann mit weniger als 200 mm Niederschlag gerechnet werden?
c) Mit welcher Wahrscheinlichkeit fallen zwischen 200 und 550 mm Niederschlag?
d) Welche Niederschlagsmenge wird wahrscheinlich in nur 2 von 100 Jahren übertroffen?
e) In welchen Grenzen liegen die mittleren 75% der jährlichen Niederschlagsmenge?

3.7.12 Aufgabe 3-7

zur Lösung

Errechnen Sie für die Verteilungen in Aufgabe 5 aus Sitzung 2 jeweils den Variationskoeffizienten.

Sitzung 4

Schätzstatistik

Lernziele dieser Sitzung

Sie können…

- eine Punktschätzung für μ und σ durchführen.
- den Standardfehler der Stichprobenverteilung von $\bar{x}$ bestimmen.
- eine Intervallschätzung für μ durchführen.

Lehrvideos (Sommersester 2020)

- 4a) Alphafehler
  - In diesem Video gibt es einen Fehler: In Schritt c) der Übungsaufgabe setze ich den falschen Wert für μ ein. Die Werte müssten stattdessen $x_{(1-\alpha/2)} = 27{,}84$ und $x_{\alpha/2} = 20{,}16$ betragen.
- 4b) Stichprobenverteilung
- 4c) Schätzungen

4.1 Stichprobenverteilung

> Die Stichprobenverteilung ist eine theoretische Verteilung, welche die möglichen Ausprägungen eines statistischen Kennwertes (z. B. $\bar{x}$) sowie deren Auftretenswahrscheinlichkeit beim Ziehen von Zufallsstichproben des Umfanges n beschreibt. (Bortz und Schuster 2010: 83)

Hier ist zunächst die theoretische Verteilung des Mittelwerts einer Stichprobe relevant. Insbesondere interessiert uns, wie sich die theoretische Verteilung des Mittelwerts abhängig von der Stichprobengröße verhält.

4.1.1 Szenario 1: Normalverteilte Grundgesamtheit

Die Grundgesamtheit (Population) einer Variable x sei normalverteilt mit $\mu = 50$ und $\sigma^2 = 25$. Wir können also schreiben:

$$x \sim N(50,\ \ 25)$$

Die Standardabweichung der Population beträgt entsprechend:

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{25} = 5\end{aligned}$$

Graphisch ist die Dichtefunktion der Verteilung in Abbildung 4.1 veranschaulicht.

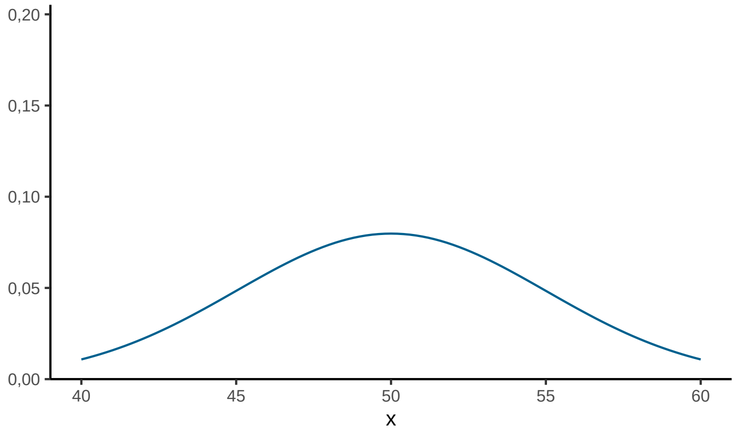


Abbildung 4.1: Dichtefunktion der Grundgesamtheit

Wenn eine einzelne Stichprobe der Größe $n = 3$ aus dieser Verteilung gezogen würde, hätte sie drei konkrete Werte (x_1, x_2 und x_3) sowie ein konkretes arithmetisches Mittel ($\bar{x}$).

Es lässt sich jedoch auch eine Wahrscheinlichkeitsdichtefunktion der Mittelwerte *aller theoretisch möglichen Stichproben* der Größe $n = 3$ (und zusätzlich der Größe $n = 6$) zeichnen (s. Abbildung 4.2).

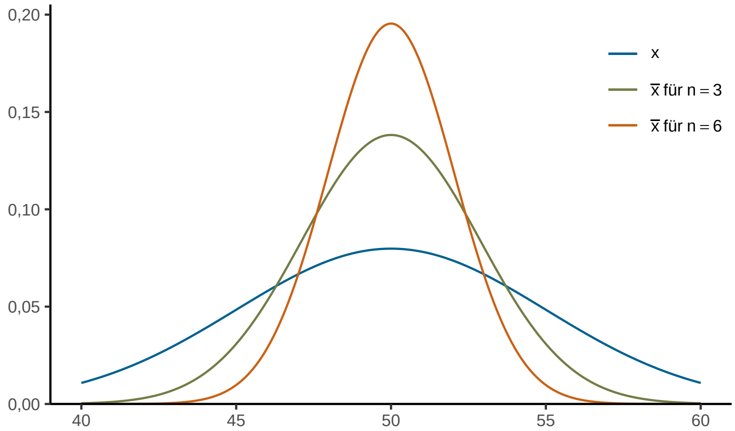


Abbildung 4.2: Dichtefunktionen der Stichprobenverteilungen

4.1.1.1 Erwartungswert

Es fällt auf, dass die Stichprobenverteilungen für $\bar{x}$ normalverteilt sind und um das arithmetische Mittel der Grundgesamtheit (μ) symmetrisch sind.

Das arithmetische Mittel der Stichprobenverteilung $\mu_{\bar{x}}$ wird auch als **Erwartungswert** (engl. *expected value*) von $\bar{x}$ bezeichnet. Es gilt:

$$\mu_{\bar{x}} = \mu$$

Wir können auch sagen: $\bar{x}$ ist ein „erwartungstreuer" Schätzparameter für μ; nicht weil er in der Empirie zwangsläufig identisch mit μ wäre, sondern weil er mit zunehmender Stichprobengröße immer stärker zu μ tendiert.

4.1.1.2 Standardfehler

Zusätzlich fällt in Abbildung 4.2 auf: Je größer die Stichprobe, desto gestauchter die Dichtekurve der Stichprobenverteilung: Die theoretische Verteilung von $\bar{x}$ bei $n = 6$ weist eine kleinere Varianz auf als bei $n = 3$. Das ist einigermaßen intuitiv, denn wir können uns vorstellen, dass das arithmetische Mittel $\bar{x}$ bei steigender Stichprobengröße ein immer präziserer Schätzwert für μ wird.

Die Varianz der Stichprobenverteilung für $\bar{x}$ bezeichnen wir mit $\sigma^2_{\bar{x}}$. Sie hängt von der Varianz der Population ab und ist invers proportional zur Stichprobengröße. Es gilt:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

Die Standardabweichung der Stichprobenverteilung ($\sigma_{\bar{x}}$) wird auch Standardfehler (engl. *standard error*) genannt. Durch Wurzelziehen ergibt sich:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Zusammenfassend lässt sich sagen:

$$\bar{x} \sim N(\mu, \tfrac{\sigma^2}{n}) \quad \text{für} \quad x \sim N(\mu, \sigma^2)$$

4.1.2 Szenario 2: Nicht normalverteilte Grundgesamtheit

Die Gleichungen (4.1.1.1), (4.1.1.2) und (4.1.1.2) gelten uneingeschränkt auch für die Stichprobenverteilungen von nicht normalverteilten Populationen. Nur die Normalverteilung der Stichprobenverteilung (Gleichung (4.1.1.2)) ist bei nicht normalverteilten Grundgesamtheiten nicht automatisch gegeben.

Das zentrale Grenzwerttheorem (engl. *central limit theorem*) besagt jedoch:

> Die Verteilung von Mittelwerten aus Stichproben des Umfangs n, die derselben Grundgesamtheit entnommen wurden, geht mit wachsendem Stichprobenumfang in eine Normalverteilung über. (Bortz und Schuster 2010: 86)

Abbildung 4.3 veranschaulicht diesen Effekt für eine nicht normalverteilte Grundgesamtheit.

In der Praxis gilt die Faustregel: Ab einer Stichprobengröße von $n = 30$ können wir statistische Verfahren anwenden, die von einer theoretischen Normalverteilung von $\bar{x}$ ausgehen – und zwar *unabhängig* von der Verteilung der Grundgesamtheit.

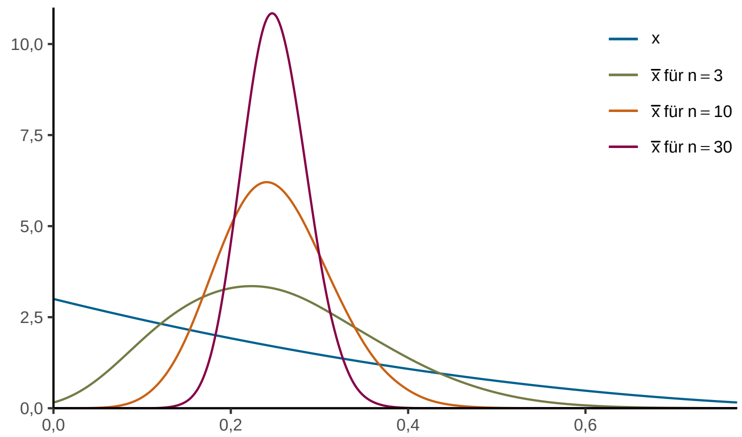


Abbildung 4.3: Stichprobenverteilung bei nicht normalverteilter Population

4.2 Punktschätzung

Bei statistischen Untersuchungen geht es oft darum, ausgehend von der empirischen Verteilung einer Stichprobe auf Parameter der Grundgesamtheit zu schließen.

Die Punktschätzung (engl. *point estimation*) ist dabei eine vergleichsweise einfache und intuitive Vorgehensweise.

4.2.1 Punktschätzung des arithmetischen Mittels

Wenn eine Stichprobe vorliegt, dann ist ihr arithmetisches Mittel ($\bar{x}$) als erwartungstreuer Punktschätzer der wahrscheinlichste Wert für das arithmetische Mittel der Grundgesamtheit (μ). Es gilt

$$\hat{\mu} = \bar{x}$$

wobei das „Dach“ auf dem μ dafür steht, dass es sich nur um eine Schätzung handelt.

Beispiel:

- Zehn Studierende der Humangeographie werden zufällig ausgewählt, um ihre Pendelzeit zum IG-Farben-Campus zu erfassen.
- Die Angaben in Minuten lauten: 22 26 12 23 48 31 15 71 17 35
- Das arithmetische Mittel der Messreihe lässt sich – wie in Sitzung 2 ausführlich besprochen – berechnen: $\bar{x} = 30$
- Da es sich um eine erwartungstreue Schätzgröße (und eine valide Zufallsstichprobe) handelt, kann die durchschnittliche Pendelzeit *aller* Studierenden der Humangeographie gemäß Gleichung (4.2.1) auf $\hat{\mu} = \bar{x} = 30$ Minuten geschätzt werden.

Gleichzeitig wissen wir jedoch, dass diese Punktschätzung des arithmetischen Mittels vermutlich nicht ganz präzise ist, sondern einem Standardfehler ($\sigma_{\bar{x}}$) unterliegt. Woher wissen wir, wie groß dieser Standardfehler ist (und wie unpräzise damit unsere Schätzung)?

4.2.2 Punktschätzung der Varianz und der Standardabweichung

Bei der Varianz einer Stichprobe s^2 handelt es sich ebenfalls um einen erwartungstreuen Punktschätzer für die Varianz der Grundgesamtheit σ^2.

Es gilt also

$$\hat{\sigma^2} = s^2$$

und damit natürlich auch

$$\hat{\sigma} = s$$

4.2.3 Schätzung des Standardfehlers

Wir führen das obige Beispiel fort:

- Die Varianz der Stichprobe können wir berechnen: $s^2 \approx 319{,}78$ (s.Sitzung 2).
- Die Varianz der Grundgesamtheit kann also mit Gleichung (4.2.1) auch auf $\hat{\sigma^2} = s^2 \approx 319{,}78$ geschätzt werden.
- Analog können wir die Standardabweichung der Population auf $\hat{\sigma} = s \approx 17{,}88$ schätzen.
- Den Standardfehler können wir mit diesem Schätzwert anhand Gleichung (4.1.1.2) berechnen. Allerdings benutzen wir statt $\sigma_{\bar{x}}$ das Symbol $s_{\bar{x}}$, da es sich um einen Schätzwert handelt:

$$\begin{aligned} s_{\bar{x}} &= \frac{s}{\sqrt{n}} \\ &\approx \frac{17{,}88}{\sqrt{10}} \approx 5{,}65 \end{aligned}$$

Je größer die Stichprobe, desto genauer lassen sich also Parameter der Population schätzen. Die statistische Antwort auf die Frage, wie groß die Stichprobe denn sein müsse, lautet demnach zunächst immer: Möglichst groß!

Bemerkenswert ist jedoch, dass dabei die Größe der Grundgesamtheit (N, im Beispiel die Anzahl aller Studierenden der Humangeographie) bei diesen Überlegungen überhaupt keine Rolle spielt.

4.3 Intervallschätzung

Um eine Intervallschätzung durchführen zu können, muss:

- die Standardabweichung der Grundgesamtheit σ bekannt und
- die theoretische Verteilung von $\bar{x}$ normalverteilt sein. Das bedeutet:
  - *Entweder* es ist bekannt, dass die Grundgesamtheit normalverteilt ist
  - *Und/oder* die Stichprobengröße ist $n \geq 30$

Für das obige Beispiel der Pendelzeiten wissen wir nicht, wie die Verteilung der Grundgesamtheit aussieht, und die Stichprobengröße ($n = 10$) ist kleiner als 30. Eine Intervallschätzung können wir hier also nicht durchführen!

Tabelle 4.1: Jahresniederschlag in Hessen

Jahr	Niederschlag (l/m²)
2011	855,3
2012	839,5
2013	850,6
2014	873,1
2015	858,3
2016	857,1
2017	861,4

Auch bei der Intervallschätzung (engl. *interval estimation*) geht es darum, das arithmetische Mittel der Population (μ) zu schätzen. Allerdings geben wir nicht einfach nur den wahrscheinlichsten Wert an, sondern einen Bereich (ein *Intervall*), in dem μ mit einer bestimmten Wahrscheinlichkeit liegt.

Die Grundüberlegung ist dabei folgende:

- Wir haben eine *empirische* Stichprobe vorliegen (und können ihren Mittelwert $\bar{x}$ und ihre Standardabweichung s berechnen).
- Wir wissen dass die *theoretische* Verteilung aller möglichen Stichproben normalverteilt ist, und um den gesuchten Wert μ symmetrisch ist.
- Den Mittelwert unserer empirischen Stichprobe $\bar{x}$ können wir uns als zufälligen Wert der theoretischen Stichprobenverteilung von $\bar{x}$ vorstellen.
- Wo genau in dieser theoretischen Verteilung wir mit unserem empirischen Wert „gelandet“ sind, wissen wir nicht.
- Wenn wir den Wert μ kennen würden, könnten wir (mit den Methoden aus Sitzung 3) die Wahrscheinlichkeit für einen beliebeigen Bereich angeben, in den ein zufälliges $\bar{x}$ fällt.
- Der entscheidende Trick: Weil die Normalverteilung symmetrisch ist, sind diese Wahrscheinlichkeiten analog anzuwenden auf die Bereiche einer konstruierten Verteilung mit gleichem $\sigma_{\bar{x}}$ um unser $\bar{x}$, in die der wirkliche Wert μ fällt. (s. Abbildung 4.4).

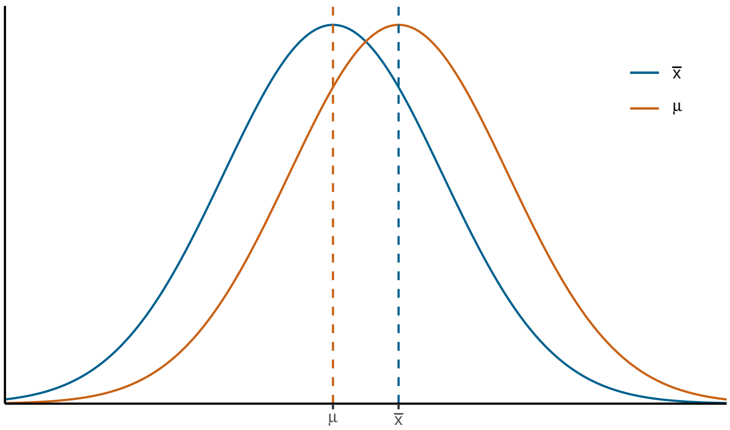


Abbildung 4.4: Konstruierte Verteilung um $\bar{x}$

Dabei heißt der Bereich Konfidenzintervall (engl. *confidence interval*), und seine Breite wird mit KIB abgekürzt. Die Wahrscheinlichkeit, dass wir mit unserer Schätzung *außerhalb* des Konfidenzintervalls liegen wird mit α gekennzeichnet. Ein 95%-Konfidenzintervall hat also ein α von 0,05 (s. Abbildung 4.5).

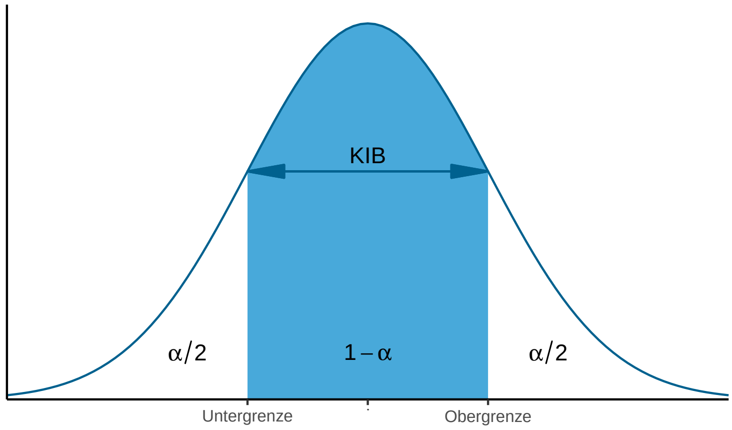


Abbildung 4.5: Konfidenzintervall

Ein Beispiel soll dies verdeutlichen: Wir wissen, dass die jährliche Niederschlagsmenge in Hessen normalverteilt ist mit $\sigma = 10{,}23$. Wir haben die Messwerte in Tabelle 1 erhoben und möchten den Mittelwert (μ) per Intervallschätzung angeben.

Zunächst errechnen wir den Mittelwert unserer empirischen Stichprobe:

$$\bar{x} \approx 856{,}47$$

Dann errechnen wir anhand Gleichung (4.1.1.2) den Standardfehler der theoretischen Verteilung von $\bar{x}$:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &\approx \frac{10{,}23}{\sqrt{7}} \approx 3{,}86 \end{aligned}$$

4.3.1 Gesuchtes α

Nun könnte eine Fragerichtung lauten: Wie groß ist die Wahrscheinlichkeit, dass der Mittelwert der Population μ in einem Korridor von ± 5 l/m² um $\bar{x}$ liegt? [1]

Gesucht ist bei einer Konfidenzintervallbreite von *KIB* $= 10$ also die Wahrscheinlichkeit:

$$1 - \alpha \approx P(851{,}47 < \mu < 861{,}47)$$

Generalisierend lässt sich schreiben:

$$1 - \alpha = P(x_{\alpha/2} < \mu < x_{(1-\alpha/2)})$$

[1] Genau genommen ist das nicht ganz korrekt, „denn tatsächlich kann der Parameter nur innerhalb oder außerhalb des gefundenen Bereichs liegen. Die Wahrscheinlichkeit, dass ein Parameter in einen bestimmten Bereich fällt, ist damit entweder 0 oder 1.“ (Bortz und Schuster 2010: 93). Mathematisch korrekt müsste es heißen: „Die Wahrscheinlichkeit, dass $\bar{x}$ zu einer Population gehört, deren Parameter μ in diesem Bereich liegt…“

…wobei $x_{\alpha/2}$ die Untergrenze darstellt und $x_{(1-\alpha/2)}$ die Obergrenze.

In z-Werten ausgedrückt:

$$1 - \alpha = P(z_{\alpha/2} < z_\mu < z_{(1-\alpha/2)})$$

In Sitzung 3 haben wir bereits gelernt, wie diese Wahrscheinlichkeit berechnet werden kann. Im Folgenden wird der Rechenweg noch einmal am Beispiel dargelegt.

4.3.1.1 Die umständliche Variante

Zunächst müssen wir die Intervallgrenzen in z-Werte umwandeln, um die Unter- bzw. Überschreitungswahrscheinlichkeiten ermitteln zu können. Die z-Transformation muss hier jedoch anhand des Standardfehlers $\sigma_{\bar{x}}$ geschehen, da wir ja an der Stichprobenverteilung interessiert sind. Durch z-Transformation mit $\bar{x}$ und dem Standardfehler $\sigma_{\bar{x}}$ erhalten wir die standardisierten Intervallgrenzen.

Untergrenze:

$$\begin{aligned} z_{\alpha/2} &= \frac{x_{\alpha/2} - \bar{x}}{\sigma_{\bar{x}}} \\ &\approx \frac{851{,}47 - 856{,}47}{3{,}86} \approx -1{,}30 \end{aligned}$$

Obergrenze:

$$\begin{aligned} z_{(1-\alpha/2)} &= \frac{x_{(1-\alpha/2)} - \bar{x}}{\sigma_{\bar{x}}} \\ &\approx \frac{861{,}47 - 856{,}47}{3{,}86} \approx 1{,}30 \end{aligned}$$

Es ist wenig überraschend, dass die z-transformierten Werte symmetrisch sind. Wir setzen in Gleichung (4.3.1) ein:

$$1 - \alpha \approx P(-1{,}30 < z_\mu < 1{,}30)$$

Dies lässt sich umformen in:

$$1 - \alpha \approx P(z_\mu < 1{,}30) - P(z_\mu < -1{,}30)$$

Die jeweiligen Wahrscheinlichkeiten lassen sich in der Tabelle für p-Werte der Normalverteilung nachschauen (bzw. für den negativen z-Wert errechnen):

$$\begin{aligned} 1 - \alpha &\approx 0{,}9032 - 0{,}0968 \\ &= 0{,}8064 \end{aligned}$$

Die Wahrscheinlichkeit, dass μ im Konfidenzintervall 856,47 ± 5 l/m² liegt, beträgt also 80,64%.

4.3.1.2 Die schnelle Variante

Wir können den z-Wert für die Obergrenze des Konfidenzintervalls ganz einfach ausrechnen, weil wir wissen, dass die Obergrenze um 5 größer ist als $\bar{x}$ und dass $z_{\bar{x}} = 0$:

$$\begin{aligned} z_{(1-\alpha/2)} &= \frac{5}{\sigma_{\bar{x}}} \\ &\approx \frac{5}{3{,}86} \\ &\approx 1{,}30 \end{aligned}$$

Oberhalb dieses Werts liegt bekanntermaßen der Anteil $\frac{\alpha}{2}$, woraus sich mit Blick auf die Tabelle ergibt:

$$\begin{aligned} \frac{\alpha}{2} &= 1 - 0{,}9032 \\ \alpha &= 0{,}1936 \end{aligned}$$

4.3.2 Gesuchtes Konfidenzintervall

Eine weitere Möglichkeit der Fragestellung lautet: In welchem Bereich liegt das arithmetische Mittel μ mit einer Wahscheinlichkeit von 90%?

Vorgegeben ist also $\alpha = 0{,}1$, und gesucht sind die Unter- und die Obergrenze des Konfidenzintervalls.

Wir setzen ein:

$$\begin{aligned} 1 - \alpha &= P(z_{\alpha/2} < z_{\mu} < z_{(1-\alpha/2)}) \\ 0{,}9 &= P(z_{5\%} < z_{\mu} < z_{95\%}) \end{aligned}$$

Die entsprechenden z-Werte der Intervallgrenzen lassen sich (in umgekehrter Suchrichtung) aus der Tabelle ablesen:

$$\begin{aligned} z_{5\%} &\approx -1{,}64 \\ z_{95\%} &\approx 1{,}64 \end{aligned}$$

Durch umgekehrte z-Transformation – auch hier weider mit $\bar{x}$ und $\sigma_{\bar{x}}$ – ergeben sich die Intervallgrenzen.

Untergrenze:

$$\begin{aligned} x_{5\%} &= z_{5\%} \cdot \sigma_{\bar{x}} + \bar{x} \\ &\approx -1{,}64 \cdot 3{,}86 + 856{,}47 \\ &\approx 850{,}14 \end{aligned}$$

Obergrenze:

$$\begin{aligned} x_{95\%} &= z_{95\%} \cdot \sigma_{\bar{x}} + \bar{x} \\ &\approx 1{,}64 \cdot 3{,}86 + 856{,}47 \\ &\approx 862{,}80 \end{aligned}$$

Auch hier gibt es wieder eine kleine Abkürzung: Aufgrund der Symmetrie unserer theoretischen Verteilung gilt für die Konfidenzintervallbreite generell:

$$\frac{KIB}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}}$$

Wir setzen einfach unsere Werte ein:

$$\begin{aligned} \frac{KIB}{2} &= z_{95\%} \cdot s_{\bar{x}} \\ &\approx 1{,}64 \cdot 3{,}86 \\ &\approx 6{,}33 \end{aligned}$$

Die Intervallgrenzen ergeben sich dann trivial aus $\bar{x} \pm \frac{KIB}{2}$.

4.3.3 Gesuchtes n

Eine letzte Fragerichtung lautet: Wie viele Messwerte müssten vorliegen, um den durchschnittlichen Niederschlag mit einem Konfidenzniveau von 99% und einer Genauigkeit von ± 5 l/m² schätzen zu können?

Gegeben sind also das Konfidenzintervall und $\alpha = 0{,}01$, gesucht wird n. Wir wissen, dass die Stichprobengröße n den Standardfehler $\sigma_{\bar{x}}$ bestimmt. Also benutzen wir zunächst Gleichung (4.3.2) und formen um:

$$\begin{aligned} \frac{KIB}{2} &= z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}} \\ \sigma_{\bar{x}} &= \frac{KIB}{2 \cdot z_{(1-\alpha/2)}} \end{aligned}$$

Durch Einsetzen und mit Blick auf die Tabelle erhalten wir:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{10}{2 \cdot z_{99{,}5\%}} \\ &\approx \frac{10}{2 \cdot 2{,}58} \\ &\approx 1{,}94 \end{aligned}$$

Dieser Standardfehler $\sigma_{\bar{x}} \approx 1{,}94$ würde unseren Anforderungen genügen. Welches n ist nötig, um diesen Standardfehler zu erreichen? Wir formen Gleichung (4.1.1.2) um…

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$$

…und setzen den angestrebten Standardfehler sowie die Standardabweichung der Population ($\sigma = 10{,}23$) ein:

$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$$
$$n \approx \left(\frac{10{,}23}{1{,}94}\right)^2$$
$$\approx 27{,}80$$

Wir müssten also 28 Stichproben vorliegen haben.

Tipps zur Vertiefung

- YouTube-Kanal „Kurzes Tutorium Statistik“: Intervallschätzungen - Konfidenzintervalle
- Kapitel 6.2–6.4 in Bortz und Schuster (2010)
- Kapitel 8.1.1 – 8.1.4 in Lange und Nipper (2018)
- Kapitel 8 in Klemm (2002)
- Kapitel 5.3.1 in Bahrenberg, Giese und Nipper (2010)
- *English:* Kapitel 8 in Burt und Barber (1996)

Übungsaufgaben

Die folgenden Aufgaben sind zur eigenständigen Überprüfung Ihrer Lernleistung gedacht (als Vor- oder Nachbereitung der Vorlesung, oder als Klausurübung) und nicht etwa als Hausaufgabe.

4.3.4 Aufgabe 4-1

zur Lösung

Eine Messreihe habe die Werte:

`165 173 155 179 158 142`

a) Führen Sie eine Punktschätzung für μ und σ der Grundgesamtheit durch.
b) Welcher Standardfehler für $\bar{x}$ ist zu erwarten?

4.3.5 Aufgabe 4-2

zur Lösung

Die Sonnenstunden auf einer Ferieninsel (pro Tag, im Jahresdurschnitt) sind annähernd normalverteilt mit einer Standardabweichung von vier Minuten. Der Mittelwert μ ist unbekannt, es liegen neun Messwerte vor.

a) Welcher Standardfehler für $\bar{x}$ ist zu erwarten?
b) Welche Konfidenzintervallbreite korrespondiert mit einem Konfidenzniveau von 95%?
c) Mit welchem Konfidenzniveau lässt sich μ „auf die Minute genau" (± 30 Sekunden) schätzen?
d) Welche Stichprobengröße ist nötig um den Mittelwert mit einer Konfidenzintervallbreite von zwei Minuten und -niveau von 90% zu schätzen?

4.3.6 Aufgabe 4-3

zur Lösung

Sie intressieren sich für das Durchschnittseinkommen (in EUR) der Haushalte eines Stadtteils. Die Varianz ist mit $\sigma^2 = 4096$ bekannt. Eine Zufallsstichprobe von 40 befragten Haushalten weist einen Mittelwert von $\bar{x} = 2650$ auf.

a) Wie lautet das 90%-Konfidenzintervall?
b) Mit welcher Wahrscheinlichkeit liegt das Durchschnittseinkommen zwischen 2640 und 2660 EUR?

4.3.7 Aufgabe 4-4

zur Lösung

Es sei bekannt, dass die Lieferzeit eines Bauteils aus Übersee annähernd normalverteilt ist mit einer Standardabweichung von 11,5 Tagen.

Bei sieben Bestellvorgängen werden folgende Lieferzeiten festgestellt (in Tagen):

$$116{,}5 \quad 94{,}5 \quad 101{,}5 \quad 109{,}0 \quad 125{,}0 \quad 112{,}5 \quad 100{,}5$$

Sie interessieren sich für die tatsächliche durchschnittliche Lieferzeit, von der Sie auch in Zukunft ausgehen können.

a) Berechnen Sie das arithmetische Mittel der beobachteten Werte für die Lieferzeit.
b) Was ist der Standardfehler für die Stichprobenverteilung von $\bar{x}$?
c) Zwischen welchen Werten liegt die tatsächliche durchschnittliche Lieferzeit mit 95% Wahrscheinlichkeit?
d) Wie viele zusätzliche Messungen müssten Sie vornehmen, um den tatsächlichen Mittelwert im selben Wertebereich zu 99% verorten zu können?

Formelsammlung und Wertetabellen

Hinweise

- Die im Folgenden dargestellten Informationen werden Ihnen so oder ähnlich auch in der Klausur zur Verfügung stehen.
- Bezeichnungen und Konventionen orientieren sich an Bortz und Schuster (2010), sind aber teilweise abweichend vereinfacht.
- Die Wertetabellen wurden mit den entsprechenden Funktionen in R (R Core Team 2018) automatisch generiert.

Formelsammlung

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

$$R = x_{(n)} - x_{(1)}$$

$$IQR = Q_3 - Q_1$$

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

$$s = \sqrt{s^2}$$

$$Md = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{falls } n \text{ gerade} \\ x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \end{cases}$$

$$v = \frac{s}{|\bar{x}|} \cdot 100\%$$

$$z_i = \frac{x_i - \bar{x}}{s}$$

$$x_i = z_i \cdot s + \bar{x}$$

$$P(x > x_p) = 1 - P(x \leq x_p)$$

$$1 - \alpha = P(z_{\alpha/2} < z_\mu < z_{(1-\alpha/2)})$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\frac{KIB}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}}$$

$$z = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\sigma}$$

$$t = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

$$F = \frac{s_1^2}{s_2^2}$$

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

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$y = a + b \cdot x$$

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

$$e_i = y_i - \hat{y}_i$$

$$m_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

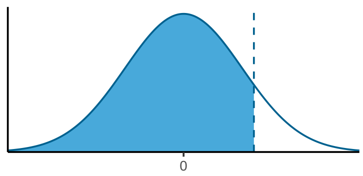
$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

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

$$CI = \sqrt{\frac{\chi^2}{n \cdot (\min(k, \ell) - 1)}}$$

Bestimmung der Freiheitsgrade für…	Formel
1-Stichproben-t-Test	$df = n - 1$
2-Stichproben-t-Test	$df = 2 \cdot n - 2$
F-Test	$df_1 = n_1 - 1; \quad df_2 = n_2 - 1$
χ^2-Unabhängigkeitstest	$df = (k-1) \cdot (\ell - 1)$
Eindimensionaler χ^2-Test	$df = k - 1$

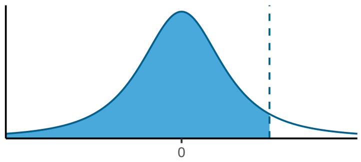
Standardnormalverteilung



$$P(z \leq -z_p) = 1 - P(z \leq z_p)$$

	z (zweite Nachkommastelle)									
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

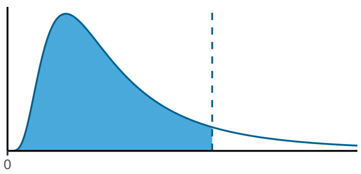
t-Verteilungen



$$P(t \leq -t_p) = 1 - P(t \leq t_p)$$

	Fläche														
df	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995	0,999	0,9995	0,9999
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	318,309	636,619	3183,099
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	22,327	31,599	70,700
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	10,215	12,924	22,204
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	7,173	8,610	13,034
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	5,893	6,869	9,678
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,208	5,959	8,025
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,785	5,408	7,063
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	4,501	5,041	6,442
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,297	4,781	6,010
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,144	4,587	5,694
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,025	4,437	5,453
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,930	4,318	5,263
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,852	4,221	5,111
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,787	4,140	4,985
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,733	4,073	4,880
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,686	4,015	4,791
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,646	3,965	4,714
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,610	3,922	4,648
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,579	3,883	4,590
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,552	3,850	4,539
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,450	3,725	4,352
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,385	3,646	4,234
35	0,127	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	2,030	2,438	2,724	3,340	3,591	4,153
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,307	3,551	4,094
45	0,126	0,255	0,388	0,528	0,680	0,850	1,049	1,301	1,679	2,014	2,412	2,690	3,281	3,520	4,049
50	0,126	0,255	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678	3,261	3,496	4,014
55	0,126	0,255	0,387	0,527	0,679	0,848	1,046	1,297	1,673	2,004	2,396	2,668	3,245	3,476	3,986
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,232	3,460	3,962
65	0,126	0,254	0,387	0,527	0,678	0,847	1,045	1,295	1,669	1,997	2,385	2,654	3,220	3,447	3,942
70	0,126	0,254	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648	3,211	3,435	3,926
75	0,126	0,254	0,387	0,527	0,678	0,846	1,044	1,293	1,665	1,992	2,377	2,643	3,202	3,425	3,911
80	0,126	0,254	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	3,195	3,416	3,899
90	0,126	0,254	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632	3,183	3,402	3,878
100	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	3,174	3,390	3,862
110	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,659	1,982	2,361	2,621	3,166	3,381	3,848
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,160	3,373	3,837
130	0,126	0,254	0,386	0,526	0,676	0,844	1,041	1,288	1,657	1,978	2,355	2,614	3,154	3,367	3,828
140	0,126	0,254	0,386	0,526	0,676	0,844	1,040	1,288	1,656	1,977	2,353	2,611	3,149	3,361	3,820
150	0,126	0,254	0,386	0,526	0,676	0,844	1,040	1,287	1,655	1,976	2,351	2,609	3,145	3,357	3,813
200	0,126	0,254	0,386	0,525	0,676	0,843	1,039	1,286	1,653	1,972	2,345	2,601	3,131	3,340	3,789
300	0,126	0,254	0,386	0,525	0,675	0,843	1,038	1,284	1,650	1,968	2,339	2,592	3,118	3,323	3,765
400	0,126	0,254	0,386	0,525	0,675	0,843	1,038	1,284	1,649	1,966	2,336	2,588	3,111	3,315	3,754
500	0,126	0,253	0,386	0,525	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586	3,107	3,310	3,747
z	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,090	3,291	3,719

F-Verteilungen

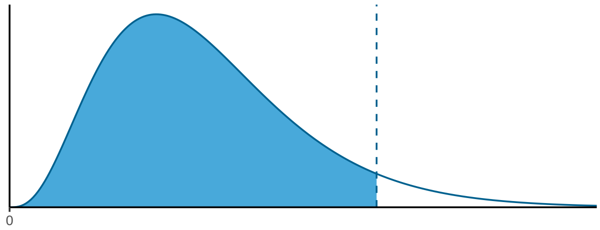


$$F_{df_1;df_2;\alpha} = \frac{1}{F_{df_2;df_1;(1-\alpha)}}$$

Alle Werte für Flächenanteil 0,95

	df_1													
df_2	1	2	3	4	5	6	7	8	9	10	15	20	50	100
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	245,95	248,01	251,77	253,04
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,43	19,45	19,48	19,49
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,58	8,55
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,70	5,66
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,44	4,41
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,75	3,71
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,32	3,27
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,02	2,97
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,80	2,76
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,64	2,59
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,51	2,46
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,40	2,35
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,31	2,26
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,24	2,19
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,18	2,12
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,12	2,07
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,08	2,02
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,04	1,98
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,00	1,94
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	1,97	1,91
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,09	2,01	1,84	1,78
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,76	1,70
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	1,96	1,88	1,70	1,63
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,66	1,59
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05	1,89	1,81	1,63	1,55
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,60	1,52
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,56	1,48
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,81	1,72	1,53	1,45
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,51	1,43
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94	1,78	1,69	1,49	1,41
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,48	1,39
110	3,93	3,08	2,69	2,45	2,30	2,18	2,09	2,02	1,97	1,92	1,76	1,67	1,47	1,38
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,75	1,66	1,46	1,37
130	3,91	3,07	2,67	2,44	2,28	2,17	2,08	2,01	1,95	1,90	1,74	1,65	1,45	1,36
140	3,91	3,06	2,67	2,44	2,28	2,16	2,08	2,01	1,95	1,90	1,74	1,65	1,44	1,35
150	3,90	3,06	2,66	2,43	2,27	2,16	2,07	2,00	1,94	1,89	1,73	1,64	1,44	1,34
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,41	1,32
300	3,87	3,03	2,63	2,40	2,24	2,13	2,04	1,97	1,91	1,86	1,70	1,61	1,39	1,30
400	3,86	3,02	2,63	2,39	2,24	2,12	2,03	1,96	1,90	1,85	1,69	1,60	1,38	1,28
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,38	1,28
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,68	1,58	1,36	1,26

χ^2-Verteilungen



	Fläche										
df	0,6	0,7	0,8	0,85	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,708	1,074	1,642	2,072	2,706	3,841	5,024	6,635	7,879	10,828	12,116
2	1,833	2,408	3,219	3,794	4,605	5,991	7,378	9,210	10,597	13,816	15,202
3	2,946	3,665	4,642	5,317	6,251	7,815	9,348	11,345	12,838	16,266	17,730
4	4,045	4,878	5,989	6,745	7,779	9,488	11,143	13,277	14,860	18,467	19,997
5	5,132	6,064	7,289	8,115	9,236	11,070	12,833	15,086	16,750	20,515	22,105
6	6,211	7,231	8,558	9,446	10,645	12,592	14,449	16,812	18,548	22,458	24,103
7	7,283	8,383	9,803	10,748	12,017	14,067	16,013	18,475	20,278	24,322	26,018
8	8,351	9,524	11,030	12,027	13,362	15,507	17,535	20,090	21,955	26,124	27,868
9	9,414	10,656	12,242	13,288	14,684	16,919	19,023	21,666	23,589	27,877	29,666
10	10,473	11,781	13,442	14,534	15,987	18,307	20,483	23,209	25,188	29,588	31,420
11	11,530	12,899	14,631	15,767	17,275	19,675	21,920	24,725	26,757	31,264	33,137
12	12,584	14,011	15,812	16,989	18,549	21,026	23,337	26,217	28,300	32,909	34,821
13	13,636	15,119	16,985	18,202	19,812	22,362	24,736	27,688	29,819	34,528	36,478
14	14,685	16,222	18,151	19,406	21,064	23,685	26,119	29,141	31,319	36,123	38,109
15	15,733	17,322	19,311	20,603	22,307	24,996	27,488	30,578	32,801	37,697	39,719
16	16,780	18,418	20,465	21,793	23,542	26,296	28,845	32,000	34,267	39,252	41,308
17	17,824	19,511	21,615	22,977	24,769	27,587	30,191	33,409	35,718	40,790	42,879
18	18,868	20,601	22,760	24,155	25,989	28,869	31,526	34,805	37,156	42,312	44,434
19	19,910	21,689	23,900	25,329	27,204	30,144	32,852	36,191	38,582	43,820	45,973
20	20,951	22,775	25,038	26,498	28,412	31,410	34,170	37,566	39,997	45,315	47,498
25	26,143	28,172	30,675	32,282	34,382	37,652	40,646	44,314	46,928	52,620	54,947
30	31,316	33,530	36,250	37,990	40,256	43,773	46,979	50,892	53,672	59,703	62,162
35	36,475	38,859	41,778	43,640	46,059	49,802	53,203	57,342	60,275	66,619	69,199
40	41,622	44,165	47,269	49,244	51,805	55,758	59,342	63,691	66,766	73,402	76,095
45	46,761	49,452	52,729	54,810	57,505	61,656	65,410	69,957	73,166	80,077	82,876
50	51,892	54,723	58,164	60,346	63,167	67,505	71,420	76,154	79,490	86,661	89,561
60	62,135	65,227	68,972	71,341	74,397	79,082	83,298	88,379	91,952	99,607	102,695
70	72,358	75,689	79,715	82,255	85,527	90,531	95,023	100,425	104,215	112,317	115,578
80	82,566	86,120	90,405	93,106	96,578	101,879	106,629	112,329	116,321	124,839	128,261
90	92,761	96,524	101,054	103,904	107,565	113,145	118,136	124,116	128,299	137,208	140,782
100	102,946	106,906	111,667	114,659	118,498	124,342	129,561	135,807	140,169	149,449	153,167
110	113,121	117,269	122,250	125,376	129,385	135,480	140,917	147,414	151,948	161,581	165,435
120	123,289	127,616	132,806	136,062	140,233	146,567	152,211	158,950	163,648	173,617	177,603
130	133,450	137,949	143,340	146,719	151,045	157,610	163,453	170,423	175,278	185,571	189,682
140	143,604	148,269	153,854	157,352	161,827	168,613	174,648	181,840	186,847	197,451	201,683
150	153,753	158,577	164,349	167,962	172,581	179,581	185,800	193,208	198,360	209,265	213,613
200	204,434	209,985	216,609	220,744	226,021	233,994	241,058	249,445	255,264	267,541	272,423
300	305,574	312,346	320,397	325,409	331,789	341,395	349,874	359,906	366,844	381,425	387,203
400	406,535	414,335	423,590	429,340	436,649	447,632	457,305	468,724	476,606	493,132	499,666
500	507,382	516,087	526,401	532,803	540,930	553,127	563,852	576,493	585,207	603,446	610,648

Lösungen der Übungsaufgaben

Sitzung 1

4.3.8 Lösung 1-1

zur Aufgabenstellung

– keine Musterlösung –

4.3.9 Lösung 1-2

zur Aufgabenstellung

– keine Musterlösungen –

4.3.10 Lösung 1-3

zur Aufgabenstellung

	Variable	Skalenniveau	Var
a)	Lebensalter in Jahren	Verhältnisskala	disk
b)	Regenmenge in mm	Verhältnisskala	stet
c)	Güteklasse	Ordinalskala	qua
d)	Passagieraufkommen	Verhältnisskala	disk
e)	Baujahr	Intervallskala	disk
f)	Geschwindigkeit in km/h	Verhältnisskala	stet
g)	Sozialstatus (Unter-, Mittel und Oberschicht)	Ordinalskala	qua
h)	Temperatur in °F	Intervallskala	stet
i)	Fläche eines Bundeslands in km^2	Verhältnisskala	stet
j)	Temperatur in K	Verhältnisskala	stet
k)	Einwohnerzahl	Verhältnisskala	disk
l)	Pegelstand	Intervallskala	stet
m)	Staatsangehörigkeit	Nominalskala	qua
n)	Interesse an Statistik (gering bis hoch)	Ordinalskala	qua
o)	Klausurnote	Ordinalskala (Intervall- auch vertretbar)	qua
p)	Bodentyp	Nominalskala	qua
q)	Entfernung zum Stadtzentrum in km	Verhältnisskala	stet
r)	Körpergröße	Verhältnisskala	stet
s)	Kleidergröße (S bis XXL)	Ordinalskala	qua
t)	Monatliches Nettoeinkommen	Verhältnisskala	stet

4.3.11 Lösung 1-4

zur Aufgabenstellung

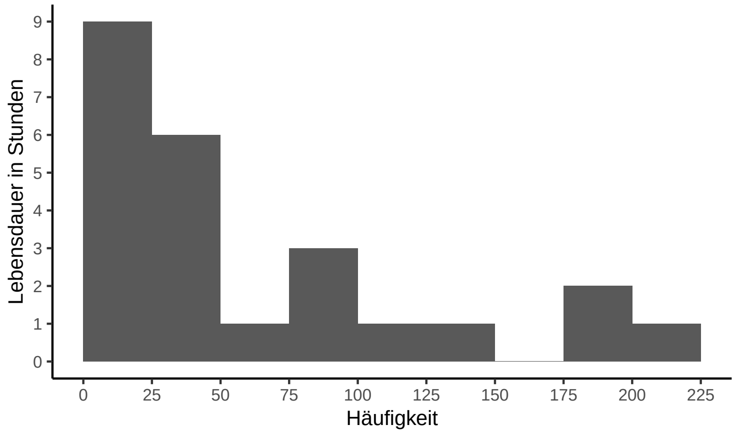
4.3.11.1 a)

Die Werte sind im Bereich zwischen 3 und 210 Stunden. Eine Klassengröße von 25 Stunden bietet sich an, es sind jedoch auch andere Größen denkbar. Da die Variable diskret zu sein scheint, können die Klassengrenzen als ganze Zahlen angegeben werden.

Wert x_i	Häufigkeit f_i
von 0 bis unter 25 h	9
von 25 bis unter 50 h	5
von 50 bis unter 75 h	2
von 75 bis unter 100 h	3
von 100 bis unter 125 h	1
von 125 bis unter 150 h	1
von 150 bis unter 175 h	0
von 175 bis unter 200 h	2
von 200 bis unter 225 h	1

4.3.11.2 b)

Das Resultat sollte je nach gewählter Klassengröße in etwa so aussehen:



4.3.11.3 c)

Die Verteilung ist unregelmäßig abfallend.

4.3.12 Lösung 1-5

zur Aufgabenstellung

Sind die folgenden Aussagen wahr oder unwahr?

a) wahr
b) wahr
c) unwahr
d) wahr
e) unwahr
f) unwahr
g) wahr
h) wahr
i) unwahr
j) unwahr
k) wahr
l) wahr
m) unwahr
n) unwahr
o) unwahr
p) wahr
q) wahr
r) wahr

Sitzung 2

4.3.13 Lösung 2-1

zur Aufgabenstellung

4.3.13.1 a)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{356}{6}$
Ergebnis	$\bar{x} = 59{,}33$

4.3.13.2 b)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{2{,}08}{8}$
Ergebnis	$\bar{x} = 0{,}26$

4.3.13.3 c)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{8350{,}16}{10}$
Ergebnis	$\bar{x} = 835{,}02$

4.3.14 Lösung 2-2

zur Aufgabenstellung

4.3.14.1 a)

Schritt	Lösung
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Varianz: Einsetzen	$s^2 = \frac{1229{,}33}{5}$
Varianz: Ergebnis	$s^2 = 245{,}87$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung: Einsetzen	$s = \sqrt{245{,}87}$
Varianz: Ergebnis	$s \approx 15{,}68$

4.3.14.2 b)

Schritt	Lösung
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Varianz: Einsetzen	$s^2 = \frac{1{,}63}{7}$
Varianz: Ergebnis	$s^2 = 0{,}23$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung: Einsetzen	$s = \sqrt{0{,}23}$
Varianz: Ergebnis	$s \approx 0{,}48$

4.3.14.3 c)

Schritt	Lösung
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Varianz: Einsetzen	$s^2 = \frac{95338{,}94}{9}$
Varianz: Ergebnis	$s^2 = 10593{,}21$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung: Einsetzen	$s = \sqrt{10593{,}21}$
Varianz: Ergebnis	$s \approx 102{,}92$

4.3.15 Lösung 2-3

zur Aufgabenstellung

4.3.15.1 a)

Die geordnete Liste ist:

1 1 1 2 2 2 2 3 3 4 4 5 7

Für das arithmetische Mittel und die Varianz ist diese Tabelle hilfreich:

x_i	f_i	$f_i \cdot x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
1	3	3	-1,85	3,41	10,22
2	4	8	-0,85	0,72	2,86
3	2	6	0,15	0,02	0,05
4	2	8	1,15	1,33	2,66
5	1	5	2,15	4,64	4,64
7	1	7	4,15	17,25	17,25

Der häufigste Wert (und damit der Modalwert) ist 2.

Die Stichprobengröße ist ungerade ($n = 13$), daher ist der Median:

$$x_{(\frac{n+1}{2})} = x_{(7)} = 2$$

Das arithmetische Mittel berechnet sich einfacher mit den Werten aus der Tabelle:

$$\bar{x} = \frac{\sum_{x=1}^{n} x_i}{n} = \frac{3+8+6+8+5+6}{13} = \frac{37}{13} \approx 2.85$$

4.3.15.2 b)

Die Spannweite ist:

$$R = x_{(n)} - x_{(1)} = 7 - 1 = 6$$

Der Quartilsabstand ist:

$$IQR = Q_3 - Q_1 = 4 - 2 = 2$$

Für die Varianz bieten sich ebenfalls die Tabellenwerte an:

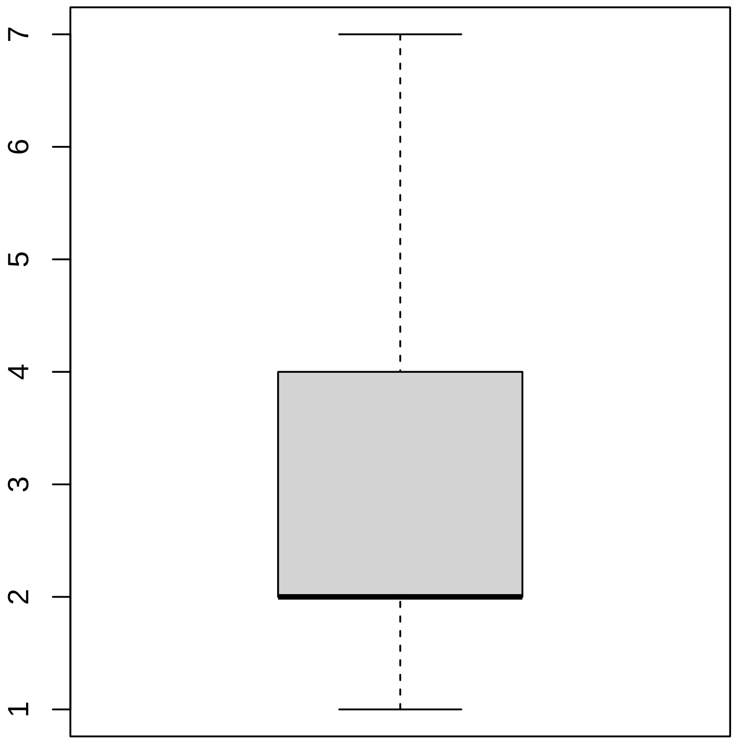
$$s^2 = \frac{\sum_{x=1}^{n} (x_i - \bar{x})^2}{n-1} \approx \frac{10{,}22 + 2{,}86 + 0{,}05 + 2{,}66 + 4{,}64 + 17{,}25}{13-1} = \frac{37{,}68}{12} = 3.14$$

Schließlich ist die Standardabweichung:

$$s = \sqrt{s^2} \approx \sqrt{3{,}14} \approx 1{,}77$$

4.3.15.3 c)

Da der untere Angelpunkt und der Median zusammenfallen, sieht der Boxplot etwas ungewöhnlich aus:



4.3.16 Lösung 2-4

zur Aufgabenstellung

4.3.16.1 a)

Für den Quartilsabstand brauchen wir den Klassendurchschnitt und kumulative Häufigkeiten:

x	k_i	f_i	f_{kum}
von 75 bis unter 77,5 cm	76,25	1	1
von 77,5 bis unter 80 cm	78,75	0	1
von 80 bis unter 82,5 cm	81,25	3	4
von 82,5 bis unter 85 cm	83,75	5	9
von 85 bis unter 87,5 cm	86,25	7	16
von 87,5 bis unter 90 cm	88,75	14	30
von 90 bis unter 92,5 cm	91,25	9	39
von 92,5 bis unter 95 cm	93,75	2	41
von 95 bis unter 97,5 cm	96,25	2	43

Bei $n = 43$ ist $Q_1 = \frac{x_{(11)}+x_{(12)}}{2}$ und $Q_3 = \frac{x_{(32)}+x_{(33)}}{2}$.

Aus der Tabelle mit kumulativen Häufigkeiten können wir $Q_1 = 86{,}25$ und $Q_3 = 91{,}25$ ablesen.

Der Quartilsabstand beträgt dann

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 91{,}25 - 86{,}25 \\ &= 5 \end{aligned}$$

4.3.16.2 b)

Um die Berechnung des arithmetischen Mittels zu vereinfachen berechnen wir den Klassendurchschnitt und Zwischensummen:

x	k_i	f_i	f_{kum}	$f_i \cdot k_i$
von 75 bis unter 77,5 cm	76,25	1	1	76,25
von 77,5 bis unter 80 cm	78,75	0	1	0,00
von 80 bis unter 82,5 cm	81,25	3	4	243,75
von 82,5 bis unter 85 cm	83,75	5	9	418,75
von 85 bis unter 87,5 cm	86,25	7	16	603,75
von 87,5 bis unter 90 cm	88,75	14	30	1242,50
von 90 bis unter 92,5 cm	91,25	9	39	821,25
von 92,5 bis unter 95 cm	93,75	2	41	187,50
von 95 bis unter 97,5 cm	96,25	2	43	192,50

Die Summen für das arithmetische Mittel entnehmen wir dann einfach der letzten Spalte:

$$
\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{76{,}25 + 243{,}75 + 418{,}75 + 603{,}75 + 1242{,}50 + 821{,}25 + 187{,}50 + 192{,}50}{43} \\
&= \frac{3786{,}25}{43} \\
&\approx 88{,}05
\end{aligned}
$$

4.3.16.3 c)

Für die Varianz erweitern wir die Tabelle:

x_i	k_i	f_i	$(k_i - \bar{x})$	$(k_i - \bar{x})^2$	$f_i \cdot (k_i - \bar{x})^2$
von 75 bis unter 77,5 cm	76,25	1	-11,8	139,24	139,24
von 77,5 bis unter 80 cm	78,75	0	-9,3	86,49	0,00
von 80 bis unter 82,5 cm	81,25	3	-6,8	46,24	138,72
von 82,5 bis unter 85 cm	83,75	5	-4,3	18,49	92,45
von 85 bis unter 87,5 cm	86,25	7	-1,8	3,24	22,68
von 87,5 bis unter 90 cm	88,75	14	0,7	0,49	6,86
von 90 bis unter 92,5 cm	91,25	9	3,2	10,24	92,16
von 92,5 bis unter 95 cm	93,75	2	5,7	32,49	64,98
von 95 bis unter 97,5 cm	96,25	2	8,2	67,24	134,48

Die Varianz beträgt:

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \\
&= \frac{139{,}24 + 138{,}72 + 92{,}45 + 22{,}68 + 6{,}86 + 92{,}16 + 64{,}98 + 134{,}48}{43-1} \\
&= \frac{691{,}57}{42} \\
&\approx 16{,}47
\end{aligned}
$$

4.3.16.4 d)

Somit beträgt die Standardabweichung

$$
\begin{aligned}
s &= \sqrt{s^2} \\
&\approx \sqrt{16{,}47} \\
&\approx 4{,}06
\end{aligned}
$$

4.3.17 Lösung 2-5

zur Aufgabenstellung

4.3.17.1 a)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{511}{6}$
Ergebnis	$\bar{x} = 85{,}17$
Einsetzen	$\bar{y} = \frac{446}{6}$
Ergebnis	$\bar{y} = 74{,}33$
Antwortsatz	Die Ziegelei weist im Mittel die größere Passantinnenzahl auf.

4.3.17.2 b)

Schritt	Lösung
Formel	$IQR = Q_3 - Q_1$
Einsetzen	$IQR_x = 91 - 77$
Ergebnis	$IQR_x = 14$
Einsetzen	$IQR_y = 103 - 51$
Ergebnis	$IQR_y = 52$
Antwortsatz	Das Möbellager hat den größeren Quartilsabstand für die Passantinnenzahl.

4.3.18 Lösung 2-6

zur Aufgabenstellung

4.3.18.1 a)

Es gibt eine Hierarchie der Werte (Ordinal-), sinnvolle Abstände (Intervall-) und einen sinnvollen Nullpunkt (Verhältnis-). Deshalb sind die angegebenen Werte als verhältnisskaliert zu verstehen.

4.3.18.2 b)

Klassen könnten z. B. wie in der folgenden Tabelle gewählt werden. Um die Berechnung des arithmetischen Mittels zu vereinfachen berechnen wir gleich den Klassendurchschnitt und Zwischensummen:

x	k_i	f_i	f_{kum}	$f_i \cdot k_i$
von 300 bis unter 400 mm	350	4	4	1400
von 400 bis unter 500 mm	450	9	13	4050
von 500 bis unter 600 mm	550	4	17	2200
von 600 bis unter 700 mm	650	2	19	1300
von 700 bis unter 800 mm	750	1	20	750

4.3.18.3 c)

Der Modalwert der so klassierten Stichprobe ist die Klasse von 400 bis unter 500 mm und kann auch mit dem Klassenmittelwert 450 mm angegeben werden.

4.3.18.4 d)

Bei $n = 20$ ist $Q_1 = \frac{x_{(5)}+x_{(6)}}{2}$ und $Q_3 = \frac{x_{(15)}+x_{(16)}}{2}$.

Aus einer geordneten Liste könnten wir also

$$\begin{aligned} Q_1 &= \frac{x_{(5)} + x_{(6)}}{2} \\ &= \frac{421{,}36 + 433{,}01}{2} \\ &\approx 427{,}19 \end{aligned}$$

und

$$\begin{aligned} Q_3 &= \frac{x_{(15)} + x_{(16)}}{2} \\ &= \frac{527{,}75 + 235{,}12}{2} \\ &\approx 531{,}44 \end{aligned}$$

bestimmen.

Wenn uns nur die klassierte Verteilung zur Verfügung steht oder wenn der Datensatz besonders unübersichtlich ist, ist es auch legitim, aus der kumulativen Häufigkeit $Q_1 = 450$ und $Q_3 = 550$ für die klassierte Verteilung abzulesen.

Je nachdem beträgt der Quartilsabstand $IQR = Q_3 - Q_1$ dann 104,24 oder 100 mm.

4.3.18.5 e)

Die Summen für das arithmetische Mittel entnehmen wir der letzten Spalte der Wertetabelle:

$$\begin{aligned} \bar{x} &= \frac{\sum\limits_{i=1}^{n} x_i}{n} \\ &= \frac{1400 + 4050 + 2200 + 1300 + 750}{20} \\ &= \frac{9700}{20} \\ &\approx 485 \end{aligned}$$

4.3.18.6 f)

Für die Standardabweichung erweitern wir die Tabelle:

x_i	k_i	f_i	$(k_i - \bar{x})$	$(k_i - \bar{x})^2$	$f_i \cdot (k_i - \bar{x})^2$
von 300 bis unter 400 mm	350	4	-135	18225	72900
von 400 bis unter 500 mm	450	9	-35	1225	11025
von 500 bis unter 600 mm	550	4	65	4225	16900
von 600 bis unter 700 mm	650	2	165	27225	54450
von 700 bis unter 800 mm	750	1	265	70225	70225

Die Varianz beträgt:

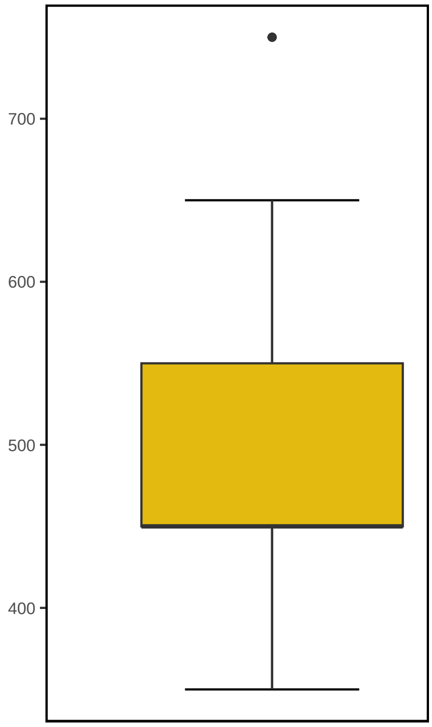
$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{72900 + 11025 + 16900 + 54450 + 70225}{20-1} \\ &= \frac{225500}{19} \\ &\approx 11868{,}42 \end{aligned}$$

Somit beträgt die Standardabweichung

$$\begin{aligned} s &= \sqrt{s^2} \\ &\approx \sqrt{11868{,}42} \\ &\approx 108{,}94 \end{aligned}$$

4.3.18.7 g)

Auch der Boxplot lässt sich anhand der klassierten Werte zeichnen:



Sitzung 3

4.3.19 Lösung 3-1

zur Aufgabenstellung

4.3.19.1 a)

Zunächst brauchen wir das arithmetische Mittel:

Schritt	Musterlösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{-170{,}47}{9}$
Ergebnis	$\bar{x} = -18{,}94$

Und die Standardabweichung:

Schritt	Lösung
Formel	$s = \sqrt{s^2}$
Einsetzen	$s = \sqrt{61{,}08}$
Ergebnis	$s \approx 7{,}82$

Dann lässt sich die Formel bestimmen:

Schritt	Musterlösung
Formel	$z_i = \frac{x_i - \bar{x}}{s}$
Einsetzen	$z_i = \frac{x_i + 18{,}94}{7{,}82}$

Und schließlich die einzelnen Werte berechnen. Hier sind die Berechnungen zum Prüfen ausformuliert, das wird in der Klausur nicht für jeden Wert erwartet.

x_i	Berechnung
-16,93	$z_1 = \frac{-16{,}93+18{,}94}{7{,}82} \approx 0{,}26$
-16,09	$z_2 = \frac{-16{,}09+18{,}94}{7{,}82} \approx 0{,}36$
-10,97	$z_3 = \frac{-10{,}97+18{,}94}{7{,}82} \approx 1{,}02$
-3,77	$z_4 = \frac{-3{,}77+18{,}94}{7{,}82} \approx 1{,}94$
-25,55	$z_5 = \frac{-25{,}55+18{,}94}{7{,}82} \approx -0{,}85$
-20,57	$z_6 = \frac{-20{,}57+18{,}94}{7{,}82} \approx -0{,}21$
-23,61	$z_7 = \frac{-23{,}61+18{,}94}{7{,}82} \approx -0{,}6$
-25,90	$z_8 = \frac{-25{,}9+18{,}94}{7{,}82} \approx -0{,}89$
-27,08	$z_9 = \frac{-27{,}08+18{,}94}{7{,}82} \approx -1{,}04$

4.3.19.2 b)

Zunächst die Standardabweichung:

Schritt	Musterlösung
Formel	$s = \sqrt{s^2}$
Einsetzen	$s = \sqrt{13{,}02}$
Ergebnis	$s \approx 3{,}61$

Dann die Formel:

Schritt	Musterlösung
Formel	$z_i = \frac{x_i - \bar{x}}{s}$
Umformen	$z_i = \frac{x_i - \bar{x}}{s}$
Einsetzen	$x_i = z_i \cdot 3{,}61 + 221{,}54$

Schließlich die einzelnen Werte:

z_i	**Berechnung**
0,90	$x_1 = 0{,}9 \cdot 3{,}61 + 221{,}54 \approx 224{,}79$
-1,40	$x_2 = -1{,}4 \cdot 3{,}61 + 221{,}54 \approx 216{,}49$
1,12	$x_3 = 1{,}12 \cdot 3{,}61 + 221{,}54 \approx 225{,}58$
-0,33	$x_4 = -0{,}33 \cdot 3{,}61 + 221{,}54 \approx 220{,}35$
2,22	$x_5 = 2{,}22 \cdot 3{,}61 + 221{,}54 \approx 229{,}55$
0,15	$x_6 = 0{,}15 \cdot 3{,}61 + 221{,}54 \approx 222{,}08$
2,87	$x_7 = 2{,}87 \cdot 3{,}61 + 221{,}54 \approx 231{,}9$
0,40	$x_8 = 0{,}4 \cdot 3{,}61 + 221{,}54 \approx 222{,}98$
-1,54	$x_9 = -1{,}54 \cdot 3{,}61 + 221{,}54 \approx 215{,}98$
0,13	$x_{10} = 0{,}13 \cdot 3{,}61 + 221{,}54 \approx 222{,}01$
-0,17	$x_{11} = -0{,}17 \cdot 3{,}61 + 221{,}54 \approx 220{,}93$
0,68	$x_{12} = 0{,}68 \cdot 3{,}61 + 221{,}54 \approx 223{,}99$

4.3.20 Lösung 3-2

zur Aufgabenstellung

4.3.20.1 a)

σ lässt sich berechnen durch:

Schritt	**Lösung**
Formel	$\sigma = \sqrt{\sigma^2}$
Einsetzen	$\sigma = \sqrt{19{,}36}$
Lösung	$\sigma \approx 4{,}4$

Dann geht es zunächst darum, die x-Werte in z-Werte zu transformieren:

Schritt	**Lösung**
Formel	$z_i = \frac{x_i - \mu}{\sigma}$
Einsetzen	$z_i = \frac{x_i - 32{,}2}{4{,}4}$

Durch Einsetzen ergeben sich die folgenden Werte. (So ausführlich muss es in der Klausur nicht sein.)

x_i	**Berechnung**
40,63	$z_1 = \frac{40{,}63 - 32{,}2}{4{,}4} \approx 1{,}92$
20,77	$z_2 = \frac{20{,}77 - 32{,}2}{4{,}4} \approx -2{,}6$
33,41	$z_3 = \frac{33{,}41 - 32{,}2}{4{,}4} \approx 0{,}27$
44,95	$z_4 = \frac{44{,}95 - 32{,}2}{4{,}4} \approx 2{,}9$
41,91	$z_5 = \frac{41{,}91 - 32{,}2}{4{,}4} \approx 2{,}21$
32,95	$z_6 = \frac{32{,}95 - 32{,}2}{4{,}4} \approx 0{,}17$

Für die positiven z-Werte können die Unterschreitungswahrscheinlichkeiten direkt in der Wertetabelle nachgeschaut werden. Für negative z-Werte gilt die Formel:

$$P(z \leq -z_p) = 1 - P(z \leq z_p)$$

Die Unterschreitungswerte ergeben:

x_i	z_i	**Formel**	**Ergebnis**	**In Prozent**
40,63	1,92	$p = P(z \leq 1{,}92)$	$p \approx 0{,}9726$	97,26%
20,77	-2,6	$p = 1 - P(z \leq 2{,}6)$	$p \approx 0{,}0047$	0,47%
33,41	0,27	$p = P(z \leq 0{,}27)$	$p \approx 0{,}6064$	60,64%
44,95	2,9	$p = P(z \leq 2{,}9)$	$p \approx 0{,}9981$	99,81%
41,91	2,21	$p = P(z \leq 2{,}21)$	$p \approx 0{,}9864$	98,64%
32,95	0,17	$p = P(z \leq 0{,}17)$	$p \approx 0{,}5675$	56,75%

4.3.20.2 b)

Es handelt sich um Überschreitungswahrscheinlichkeiten, aber aus der Tabelle lassen sich nur Unterschreitungswerte ablesen. Weil die Normalverteilung symmetrisch ist, gilt aber:

$$P(x > x_p) = 1 - P(x \leq x_p)$$

So lässt sich jeweils sagen:

Überschr. p_i	**Unterschr.** $(1 - p_1)$	**Berechnung**	**....**
0,015	0,985	$P(z \leq z_1) = 0{,}985$	
0,025	0,975	$P(z \leq z_2) = 0{,}975$	
0,050	0,950	$P(z \leq z_3) = 0{,}95$	
0,130	0,870	$P(z \leq z_4) = 0{,}87$	
0,500	0,500	$P(z \leq z_5) = 0{,}5$	
0,900	0,100	$P(z \leq -z_6) = 1 - 0{,}1 = 0{,}9$	$-z_6 \approx 1{,}28$
0,990	0,010	$P(z \leq -z_7) = 1 - 0{,}01 = 0{,}99$	$-z_7 \approx 2{,}33$
0,995	0,005	$P(z \leq -z_8) = 1 - 0{,}005 = 0{,}995$	$-z_8 \approx 2{,}58$

Für die Rücktransformation gilt die Formel:

$$x_i = z_i \cdot \sigma + \mu$$

z_i	**Einsetzen**	x_i
2,17	$x_1 = 2{,}17 \cdot 4{,}4 + 32{,}2$	$x_1 \approx 41{,}75$
1,96	$x_2 = 1{,}96 \cdot 4{,}4 + 32{,}2$	$x_2 \approx 40{,}82$
1,64	$x_3 = 1{,}64 \cdot 4{,}4 + 32{,}2$	$x_3 \approx 39{,}42$
1,13	$x_4 = 1{,}13 \cdot 4{,}4 + 32{,}2$	$x_4 \approx 37{,}17$
0	$x_5 = 0 \cdot 4{,}4 + 32{,}2$	$x_5 \approx 32{,}2$
-1,28	$x_6 = -1{,}28 \cdot 4{,}4 + 32{,}2$	$x_6 \approx 26{,}57$
-2,33	$x_7 = -2{,}33 \cdot 4{,}4 + 32{,}2$	$x_7 \approx 21{,}95$
-2,58	$x_8 = -2{,}58 \cdot 4{,}4 + 32{,}2$	$x_8 \approx 20{,}85$

4.3.20.3 c)

Die mittleren 95% der Werte liegen zwischen einem unteren Wert $x_{2{,}5\%}$ (der zu 2,5% unterschritten wird) und einem oberen Wert $x_{97{,}5\%}$ (der zu 2,5% überschritten wird).

Der obere z-Wert lässt sich leicht finden: $z_{97{,}5\%} \approx 1{,}96$

Durch Symmetrie wissen wir dann auch, dass: $z_{2{,}5\%} \approx -1{,}96$

Nun noch rückwärts transformieren:

Schritt	**Lösung**
Formel	$x_i = z_i \cdot \sigma + \mu$
Untergrenze: Einsetzen	$x_u = -1{,}96 \cdot 4{,}4 + 32{,}2$
Untergrenze: Ergebnis	$x_u \approx 23{,}58$
Obergrenze: Einsetzen	$x_o = 1{,}96 \cdot 4{,}4 + 32{,}2$
Obergrenze: Ergebnis	$x_o \approx 40{,}82$
Antwortsatz	Die mittleren 95 Prozent der Werte liegen zwischen 23,58 und 40,82.

4.3.20.4 d)

Es ist immer einfacher, mit Unterschreitungswahrscheinlichkeiten zu arbeiten. Zwischen 30 und 40 heißt auch: unter 40, aber nicht unter 30. Formal sieht das so aus:

$$P(30 < x \leq 40) = P(x \leq 40) - P(x \leq 30)$$

Diese Unterschreitungswahrscheinlichkeiten bestimmen wir wieder über die z-Transformation:

Schritt	Lösung
Formel	$z_i = \frac{x_i - \mu}{\sigma}$
Untergrenze: z-Wert	$z_u = \frac{30-32{,}2}{4{,}4} \approx -0{,}5$
Untergrenze: Unterschr.	$p \approx 0{,}3085$
Obergrenze: z-Wert	$z_o = \frac{40-32{,}2}{4{,}4} \approx 1{,}77$
Obergrenze: Unterschr.	$p \approx 0{,}9616$
Intervall	$P(30 < x \leq 40) = P(x \leq 40) - P(x \leq 30)$
Intervall einsetzen	$P(30 < x \leq 40) \approx P(z \leq 0{,}9616) - P(z \leq 0{,}3085)$
Intervall Ergebnis	$P(30 < x \leq 40) \approx 0{,}6531$
Antwortsatz	Ein zufälliger Wert der Verteilung liegt mit 65,31-prozentiger Wahrscheinlichkeit zwischen 30 und 40.

4.3.21 Lösung 3-3

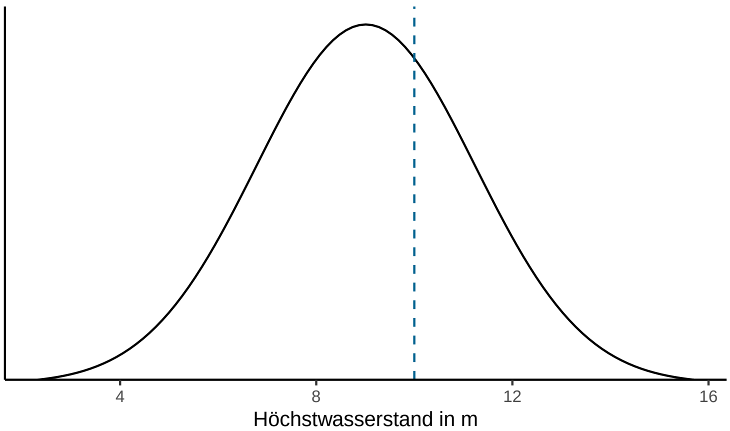
zur Aufgabenstellung

4.3.21.1 a)

Siehe b)

4.3.21.2 b)

Die Dichtefunktion mit kritischem Wert sollte in etwa so aussehen:



4.3.21.3 c)

$$z_p = \frac{x_p - \mu}{\sigma} = \frac{10 - 9{,}01}{2{,}23} \approx 0{,}44$$

4.3.21.4 d)

$$p = P(z < z_p) \approx P(z < 0{,}44) \approx 0{,}6700$$

Die Wahrscheinlichkeit, dass der Deich unbeschädigt bleibt, beträgt 67%.

4.3.22 Lösung 3-4

zur Aufgabenstellung

4.3.22.1 a)

Die Übertretungswahrscheinlichkeit beträgt:

$$P(z > z_p) = 1 - P(z < z_p) \approx 1 - 0{,}6700 = 0{,}3300 = 33\%$$

4.3.22.2 b)

Für $x_p = 12$ ergibt sich:

$$z_p = \frac{x_p - \mu}{\sigma} = \frac{12 - 9{,}01}{2{,}23} \approx 1{,}34$$

Und für die Übertretungswahrscheinlichkeit:

$$P(z > z_p) = 1 - P(z < z_p) \approx 1 - 0{,}9099 = 0{,}0901 = 9{,}01\%$$

4.3.22.3 c)

Wir kennen $P(x < 12) \approx 0{,}9099$ aus Aufgabe 2 b) und $P(x < 10) \approx 0{,}6700$ aus Aufgabe 1 d). Also rechnen wir:

$$P(10 < x < 12) = P(x < 12) - P(x < 10) \approx 0{,}9099 - 0{,}6700 = 0{,}2399$$

4.3.22.4 d)

Für die Obergrenze soll gelten: $P(x < x_o) = 0{,}9$. Der Tabelle entnehmen wir $z_o \approx 1{,}28$. Entsprechend ist $z_u \approx -1{,}28$.

Die Umkehrung der z-Transformation ergibt:

$$x_o = z_o \cdot \sigma + \mu \approx 1{,}28 \cdot 2{,}23 + 9{,}01 \approx 11{,}86$$
$$x_u = z_u \cdot \sigma + \mu \approx -1{,}28 \cdot 2.23 + 9.01 \approx 6{,}16$$

Die mittleren 80% der Werte liegen also zwichen 6,16 und 11,86 m.

4.3.23 Lösung 3-5

zur Aufgabenstellung

4.3.23.1 a)

$$p = P(x < x_p) = 1 - P(x > x_p) = 1 - \frac{1}{200} = 1 - 0{,}005 = 0{,}995$$

4.3.23.2 b)

$$z_{99,5\%} \approx 2{,}58$$

4.3.23.3 c)

$$x_{99,5\%} = z_{99,5\%} \cdot \sigma + \mu \approx 2{,}58 \cdot 2{,}23 + 9{,}01 \approx 14{,}76$$

Der neue Deich muss 14,76 m hoch sein.

4.3.24 Lösung 3-6

zur Aufgabenstellung

4.3.24.1 a)

- $z_p = 1$ und $P(z < 1) \approx 84{,}13\%$, also $P(z > 1) \approx 15{,}87\%$

4.3.24.2 b)

- $z_p = -2$ und $P(z < -2) = 1 - P(z < 2) \approx 1 - 0{,}9772 = 0{,}0228$
- Es kann also 2,28 Mal in 100 Jahren (oder: in etwa 2 von 100 Jahren, in weniger als 3 von 100 Jahren) mit weniger als 200 mm Regen gerechnet werden.

4.3.24.3 c)

- $z_u = -2$ und $P(z < z_u) \approx 0{,}0228$ (siehe b)
- $z_o = \frac{x_o - \mu}{\sigma} = \frac{550-400}{100} = 1{,}5$ und $P(z < z_o) \approx 0{,}9332$
- $P(200 < x < 550) = P(x < 550) - P(x < 200) \approx 91{,}04\%$

4.3.24.4 d)

- Gesucht ist x_p, für das gilt: $P(x > x_p) = \frac{2}{100} = 0{,}02$
- Daraus folgt: $P(x < x_p) = 0{,}98$ und $z_p \approx 2{,}05$
- $x_p = 605$

4.3.24.5 e)

- $z_{12,5\%} \approx -1{,}15$ und $z_{87,5\%} = 1{,}15$
- Die mittleren 75% liegen zwischen $x_u = 285$ und $x_o = 515$ mm.

4.3.25 Lösung 3-7

zur Aufgabenstellung

Für die Ziegelei:

Schritt	Lösung
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Varianz: Einsetzen	$s_x^2 = \frac{610{,}83}{5}$
Varianz: Ergebnis	$s_x^2 = 122{,}17$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung: Ergebnis	$s_x \approx 11{,}05$
Variationskoeffizient: Formel	$v = \frac{s}{\lvert\bar{x}\rvert} \cdot 100\%$
Variationskoeffizient: Einsetzen	$v \approx \frac{11{,}05}{85{,}17} \cdot 100\%$
Variationskoeffizient: Ergebnis	$v \approx 12{,}97\%$

Für das Möbellager:

Schritt	Lösung
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Varianz: Einsetzen	$s_y^2 = \frac{4015{,}33}{5}$
Varianz: Ergebnis	$s_y^2 = 803{,}07$
Standardabweichung: Formel	$s_y = \sqrt{s_y^2}$
Standardabweichung: Ergebnis	$s_y \approx 28{,}34$
Variationskoeffizient: Formel	$v = \frac{s}{\lvert\bar{x}\rvert} \cdot 100\%$
Variationskoeffizient: Einsetzen	$v \approx \frac{28{,}34}{74{,}33} \cdot 100\%$
Variationskoeffizient: Ergebnis	$v \approx 38{,}13\%$

Quellenverzeichnis

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Benninghaus, Hans. 2007. *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler*. Wiesbaden: VS Verlag.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Burt, James E. und Gerald M. Barber. 1996. *Elementary statistics for geographers*. 2nd ed. New York: Guilford Press.

Haseloff, Otto W., Hans-Joachim Hoffmann, John H. Maindonald und W. John Braun. 1968. *Kleines Lehrbuch der Statistik DAAG. Data Analysis and Graphics Data and Functions*. Berlin: de Gruyter.

Klemm, Elmar. 2002. *Einführung in die Statistik. Für die Sozialwissenschaften*. Wiesbaden: Westdeutscher Verlag.

Lange, Norbert de und Josef Nipper. 2018. *Quantitative Methodik in der Geographie*. UTB Geographie, Methoden, Statistische Verfahren 4933. Paderborn: Ferdinand Schöningh.

Maindonald, John H. und W. John Braun. 2020. DAAG: Data Analysis and Graphics Data and Functions. https://CRAN.R-project.org/package=DAAG.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Wien: R Foundation for Statistical Computing. https://www.R-project.org/ (zugegriffen: 9. April 2021).

Zimmermann-Janschitz, Susanne. 2014. *Statistik in der Geographie. Eine Exkursion durch die deskriptive Statistik*. Berlin: Springer.