Statistische Verfahren in der Geographie

Skript für den Theorieteil

Till Straube
straube@geo.uni-frankfurt.de

Institut für Humangeographie
Goethe-Universität Frankfurt

Sommersemester 2021

Inhaltsverzeichnis

Terminüberblick **3**

Vorbesprechung **4**

1 Datenerhebung und Häufigkeiten **6**
1.1 Statistische Praxis . . . 6
1.2 Grundlagen der Datenerhebung . . . 9
1.3 Häufigkeitsverteilungen . . . 11
Tipps zur Vertiefung . . . 15
Übungsaufgaben . . . 17

2 Maßzahlen **20**
2.1 Einleitende Bemerkungen . . . 20
2.2 Lagemaße . . . 21
2.3 Streumaße . . . 23
2.4 Boxplot . . . 27
Tipps zur Vertiefung . . . 27
Übungsaufgaben . . . 28

3 z-Werte und Normalverteilung **32**
3.1 Variationskoeffizient . . . 32
3.2 z-Transformation . . . 33
3.3 Normalverteilung . . . 33
3.4 Standardnormalverteilung . . . 35
3.5 Crash-Kurs Wahrscheinlichkeitsrechnung . . . 35
3.6 Wahrscheinlichkeitsdichtefunktionen . . . 35
3.7 Wahrscheinlichkeitsrechnung mit Standardnormalverteilung . . . 36
Tipps zur Vertiefung . . . 42
Übungsaufgaben . . . 43

4 Schätzstatistik **46**
4.1 Stichprobenverteilung . . . 46
4.2 Punktschätzung . . . 49
4.3 Intervallschätzung . . . 50
Tipps zur Vertiefung . . . 56
Übungsaufgaben . . . 56

5 Grundlagen der Teststatistik **58**

5.1 Statistische Tests . . . 58
5.2 z-Test . . . 59
5.3 Die t-Verteilung . . . 64
5.4 1-Stichproben-t-Test . . . 65
Tipps zur Vertiefung . . . 69
Übungsaufgaben . . . 70

6 Testverfahren mit zwei Stichproben 73
6.1 Statistische Tests . . . 73
6.2 2-Stichproben-t-Test {#2-stichproben-t-test} . . . 73
6.3 Die F-Verteilung . . . 77
6.4 F-Test . . . 77
6.5 Fehlerarten . . . 82
Tipps zur Vertiefung . . . 82
Übungsaufgaben . . . 82

7 Korrelation 85
7.1 Bivariate Statistik . . . 85
7.2 Kovarianz s_{xy} . . . 86
7.3 Korrelationskoeffizient r . . . 88
Tipps zur Vertiefung . . . 90
Übungsaufgaben . . . 90

8 Lineare Regression 94
8.1 Regresssionsanalyse . . . 94
8.2 Bestimmung der Regressionsgeraden . . . 96
8.3 Residuen . . . 97
8.4 Determinationskoeffizient R^2 . . . 99
Tipps zur Vertiefung . . . 100
Übungsaufgaben . . . 100

9 Kreuztabellen 103
9.1 Bivariate Verteilungen mit nominalen Variablen . . . 103
9.2 Kreuztabelle . . . 104
9.3 Erwartungswerte . . . 105
9.4 Kontingenzkoeffizent χ^2 . . . 106
9.5 ϕ-Koeffizient . . . 108
9.6 Cramér-Index . . . 108
Tipps zur Vertiefung . . . 110
Übungsaufgaben . . . 110

10 χ^2-Tests 113
Anwendungsbereich . . . 113
10.1 χ^2-Unabhängigkeitstest . . . 114
10.2 χ^2-Anpassungstest . . . 117
Tipps zur Vertiefung . . . 119

Formelsammlung und Wertetabellen 120

Lösungen der Übungsaufgaben 121

Sitzung 1 121
Sitzung 2 123
Sitzung 3 132
Sitzung 4 140
Sitzung 5 143
Sitzung 6 146
Sitzung 7 150
Sitzung 8 155
Sitzung 9 159

Quellenverzeichnis **165**

Terminüberblick

Alle Sitzungen finden von 14 bis 16h c.t. statt

Datum	Sitzung	Inhalt
13. April 2021		Vorbesprechung
20. April 2021	1	Datenerhebung und Häufigkeiten
27. April 2021	2	Maßzahlen
4. Mai 2021	3	[z-Werte und Normalverteilung]
11. Mai 2021	4	Schätzstatistik
18. Mai 2021	5	Grundlagen der Teststatistik
25. Mai 2021	6	Testverfahren mit zwei Stichproben
1. Juni 2021	7	Korrelation
8. Juni 2021	8	Lineare Regression
15. Juni 2021	9	Kreuztabellen
22. Juni 2021	10	[Chi-Quadrat-Tests]
29. Juni 2021		Klausurvorbereitung
6. Juli 2021		Klausurvorbereitung
13. Juli 2021		Klausur

Vorbesprechung

Aufzeichung der Vorbesprechung am 13. April

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

Sitzungsvorbereitung

- Materialien werden zur eigenständigen Vorbereitung bereit gestellt
- Dieses Online-Skript mit den Kerninhalten
- Darin: Videos (aus dem Vorjahr) 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.. auf Zoom (Link in OLAT)
- Übungsaufgaben (und Lösungen) werden online bereit gestellt
- Teilnehmer*innen bearbeiten die Aufgaben in Break-Out-Sessions
- Bei Problemen fragen Sie sich erstmal gegenseitig
- Sonst bin ich ansprechbar (Zoom-Funktion: Um Hilfe bitten)

Empfehlungen

- Lassen Sie sich auf den wöchentlichen Rhythmus ein
- Bereiten Sie die Sitzungen vor und nach
- Bilden Sie Lerngruppen
- Melden Sie mir gerne Break-Out-Wünsche per E-Mail (aber keine Garantie)

- 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:
  - 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)

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!

Klausur

- Termin: 13. Juli 2021, 14 h s. t.
- Präsenz oder online möglich
- Berechnung der Aufgaben „von Hand“ auf Papier
- Bearbeitungsdauer: 60 Minuten
- Bei Online-Klausur wird zusätzliche Zeit für technische Abwicklung gewährt
- Hilfsmittel: Taschenrechner, Formelsammlung
- Vier Teilaufgaben, immer nach demselben Schema (dazu im Laufe des Semesters mehr)
- Viele Probeklausuren zur Vorbereitung

Nachklausur

- Termin: 12. Oktober 2021, 14 h s. t.
- Gleiches Schema wie die reguläre Klausur (mit anderen Aufgaben)
- Nicht einfacher als die reguläre Klausur

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 et al. (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 et al. 1968: 27)

Grundbegriffe der Statistik

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.

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.

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).

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?“

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.

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.

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.

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.

Deskriptive und schließende Statistik

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

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.

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:

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

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

Ablauf einer statistischen Untersuchung

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

Datenerhebung

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

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)

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

Typen von Stichproben

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.

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.

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.

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)

Variablentypen

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

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).

Quantitative Variablen

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

Quantitative Variablen können diskret oder stetig sein:

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).

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).

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.

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

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)

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:

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.

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

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

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 G H H H I 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 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 P P R R R S S S S S S T T T T T T T T T T
V V V V V V V V Y Y

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.

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()`.

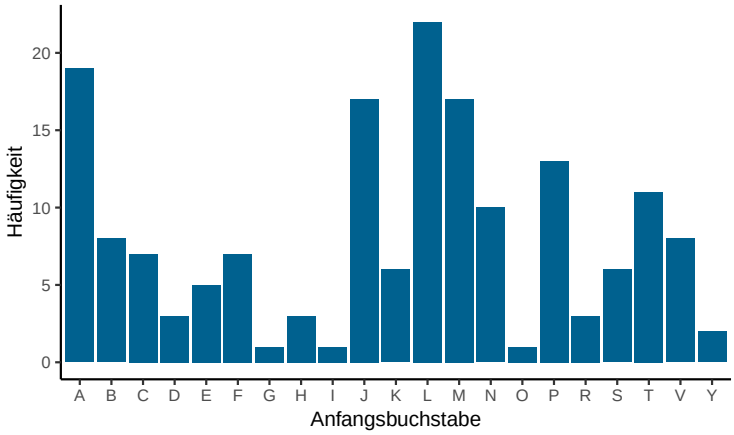


Abbildung 1.1: Stabdiagramm

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%

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.

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).

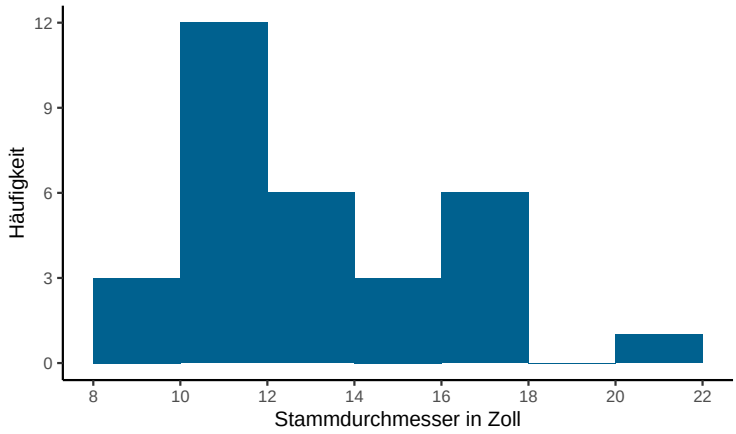


Abbildung 1.2: Histogramm

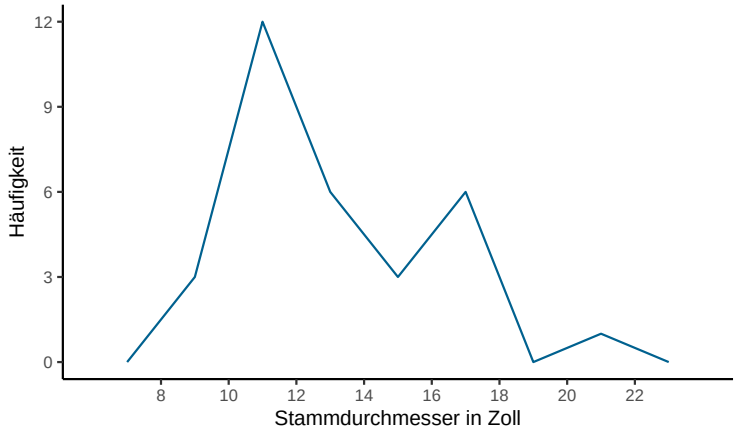


Abbildung 1.3: Polygonzug

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

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)

Stichproben

- Kapitel 6.1 in Bortz und Schuster (2010)

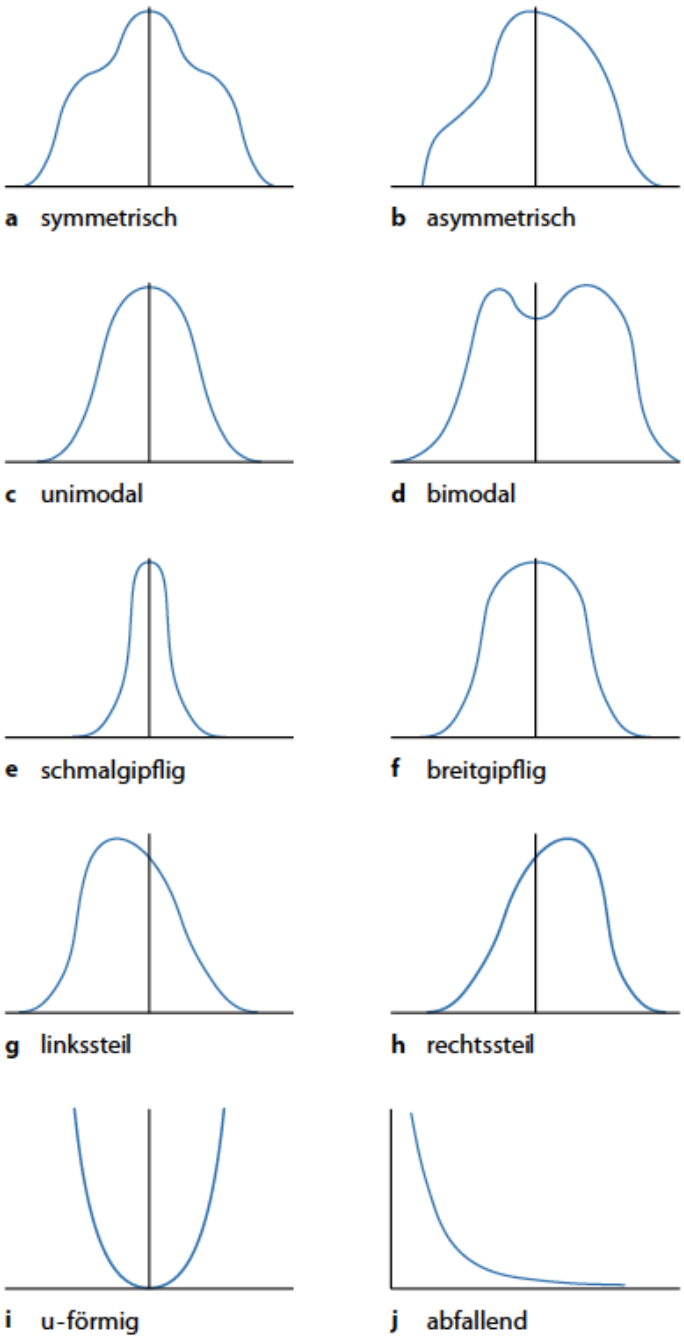


Abbildung 1.4: Merkmale von Verteilungen [aus: @bortz: 42]

- 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)

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)

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

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?

Aufgabe 1-2

zur Lösung

Finden Sie als Gruppe jeweils zwei Beispiele für:

- systematische Zufallsstichproben
- geschichtete Zufallsstichproben
- Klumpenstichproben

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 km²
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

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.

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.

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

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.

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 $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ bezeichnet (also mit Klammern). Für unsere Beispielverteilung ergibt sich Tabelle 2.3.

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

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) wie folgt:

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

Softwarehinweis

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

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.

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

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):

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

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.

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.

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:

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

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

Die oberen 7 Werte lauten:

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

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

Für den Quartilsabstand ergibt sich durch einsetzen in Formel (2.3):

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

Softwarehinweis

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

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!

Varianz

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

Die Formel lautet:

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

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.

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

Schließlich werden die Werte in Formel (2.3) eingesetzt:

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

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()`.

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.

2.4 Boxplot

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

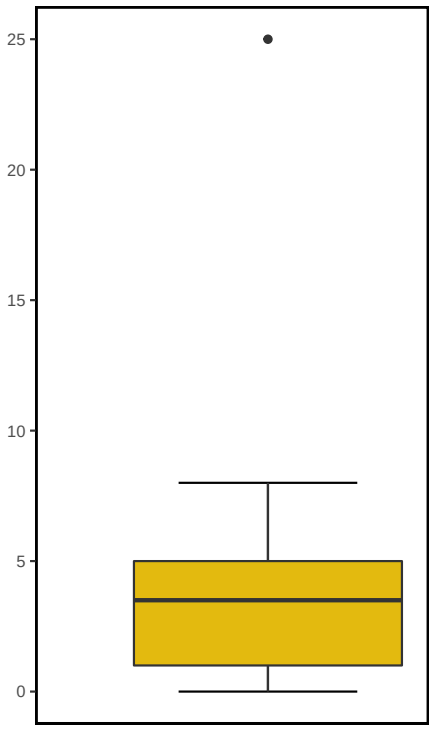


Abbildung 2.1: Boxplot der Beispielverteilung

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

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)

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)

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

Aufgabe 2-1

zur Lösung

Berechnen Sie das arithmetische Mittel für die folgenden Verteilungen:

a)

```
72 55 69 69 30 61
```

b)

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

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?

Aufgabe 2-2

zur Lösung

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

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.

Aufgabe 2-4

zur Lösung

Eine Messreihe der Körperlänge weiblicher Beutelratten hat folgende Werte in cm erfasst (Beispieldatensatz `fossum` aus Maindonald und Braun 2015):

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

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

d) die Standardabweichung.

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:

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

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

b) Vergleichen Sie außerdem die Quartilsabstände der beiden Messreihen.

Aufgabe 2-6

zur Lösung

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

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

a) Welches Skalenniveau liegt vor? (Sitzung 1)

[1] Auszug aus dem Datensatz `bomsoi` in Haseloff et al. (1968)

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.

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:

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

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.

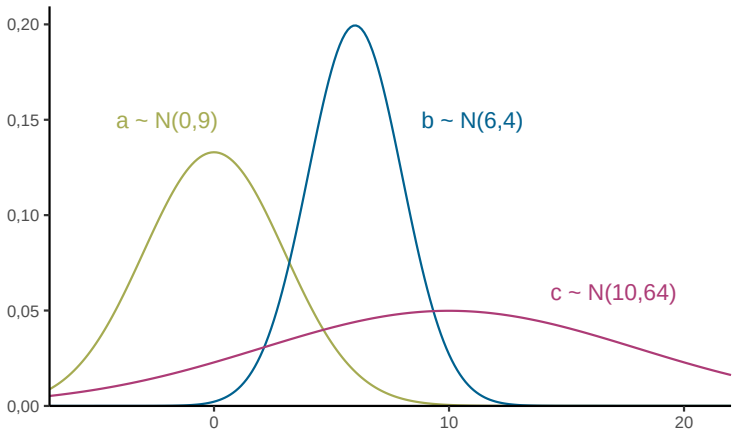


Abbildung 3.1: Dichtefunktionen verschiedener Normalverteilungen

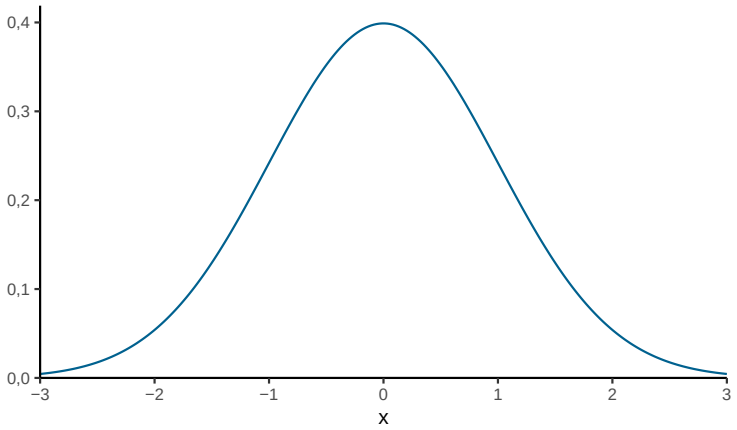


Abbildung 3.2: Dichtefunktion der Standardnormalverteilung

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 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 $z_p \rightarrow P(z < z_p)$ für normalverteilte Werte auch mit $\Phi(z)$ bezeichnet (z. B. in Bahrenberg, Giese und Nipper 2010).

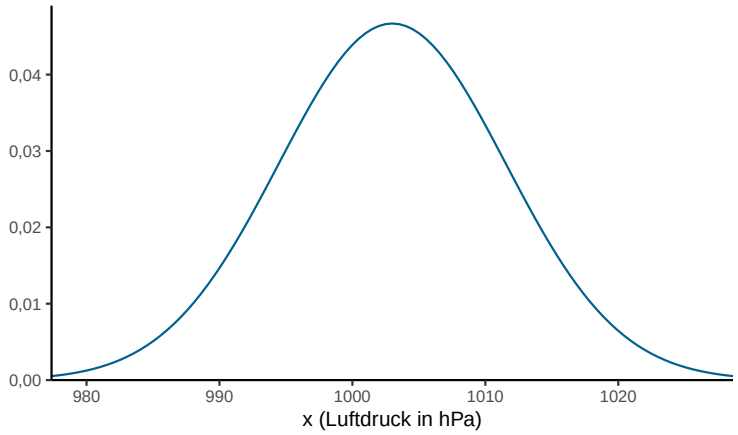


Abbildung 3.3: Theoretische Wahrscheinlichkeitsdichtefunktion des Luftdrucks

Daraus ergibt sich für die Standardabweichung σ:

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

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).

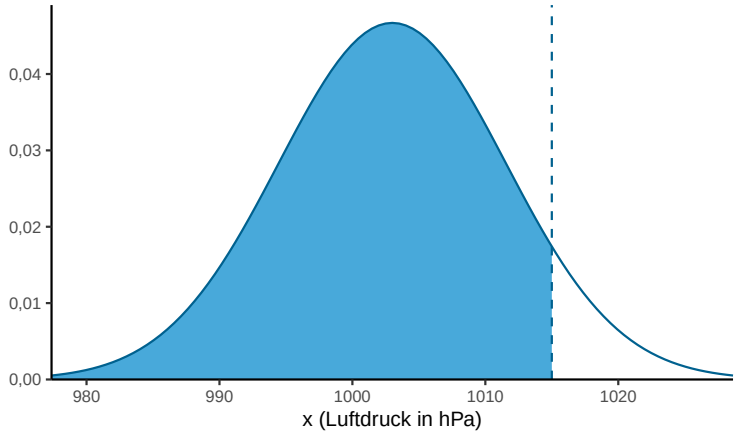


Abbildung 3.4: Unterschreitung eines Messwerts

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.

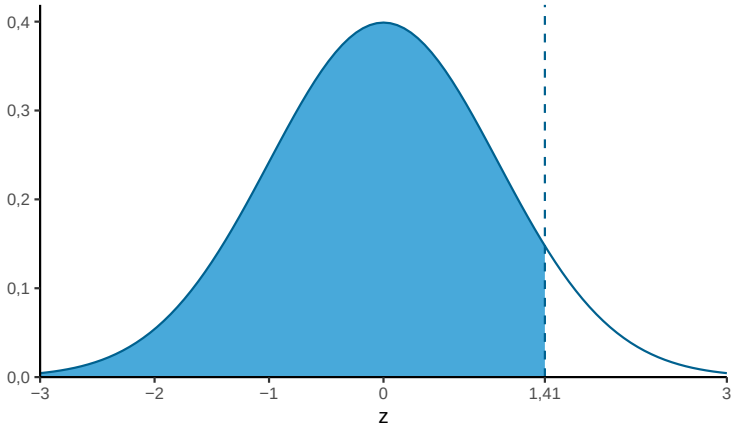


Abbildung 3.5: Standardnormalverteilung des Luftdrucks

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

Der Formelsammlung 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.

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

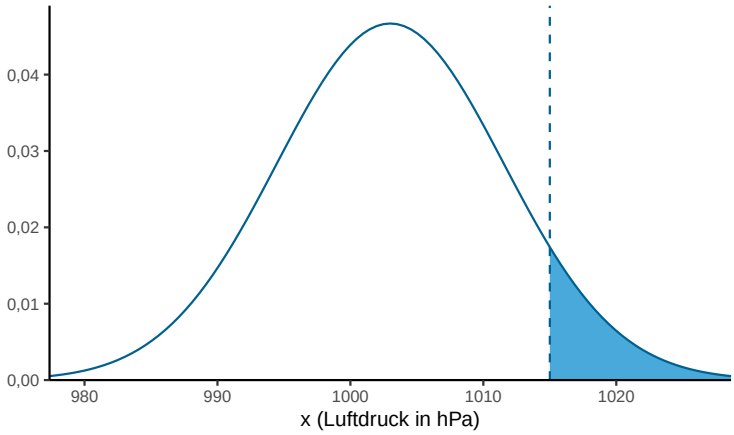


Abbildung 3.6: Überschreitung eines Messwerts

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.

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 Formelsammlung 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.

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.

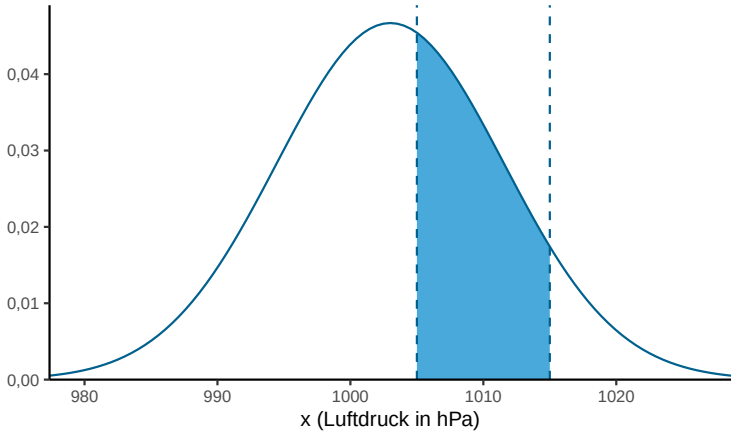


Abbildung 3.7: Messwertintervall

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 Formelsammlung 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.

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.

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.

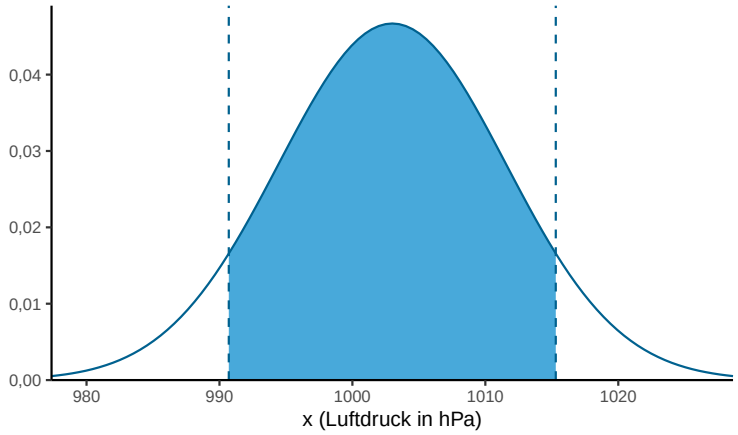


Abbildung 3.8: Die mittleren 85% der Normalverteilung

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

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)

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)

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)

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

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`

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?

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)?

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?

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?)

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?

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.

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.

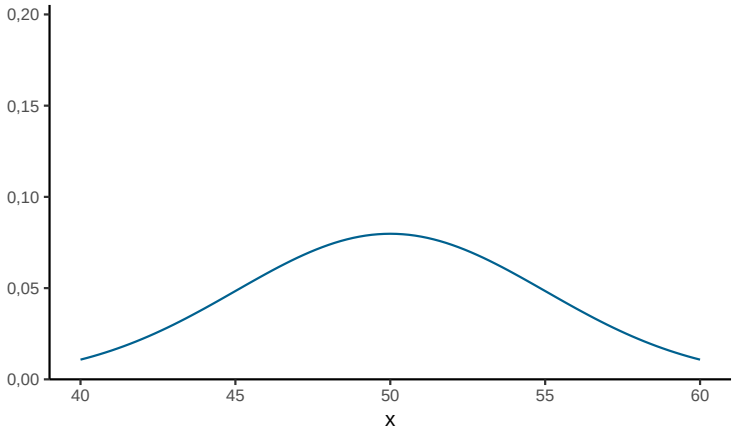


Abbildung 4.1: Dichtefunktion der Grundgesamtheit

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).

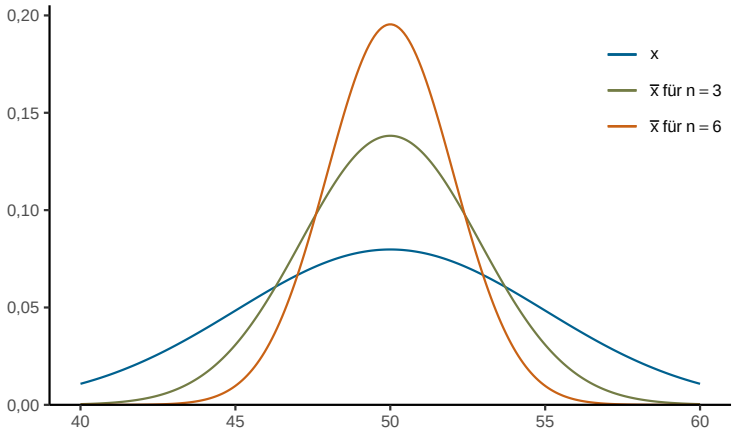


Abbildung 4.2: Dichtefunktionen der Stichprobenverteilungen

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.

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:

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

Szenario 2: Nicht normalverteilte Grundgesamtheit

Die Gleichungen (4.1), (4.1) und (4.1) gelten uneingeschränkt auch für die Stichprobenverteilungen von nicht normalverteilten Populationen. Nur die Normalverteilung der Stichprobenverteilung (Gleichung (4.1)) 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.

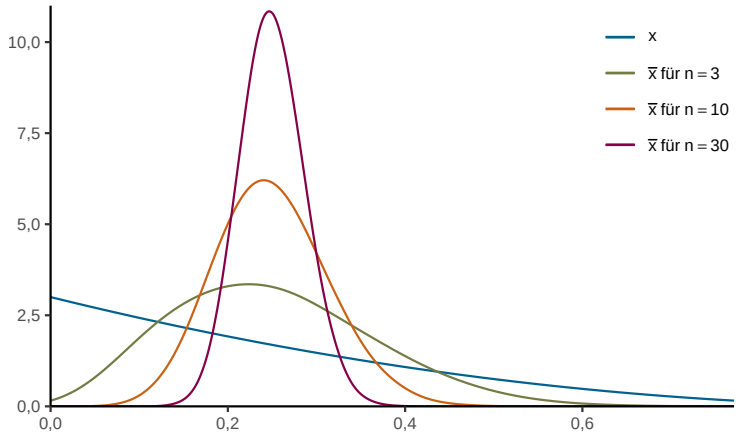


Abbildung 4.3: Stichprobenverteilung bei nicht normalverteilter Population

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.

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) 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)?

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

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

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) 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) 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/m^2)
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).

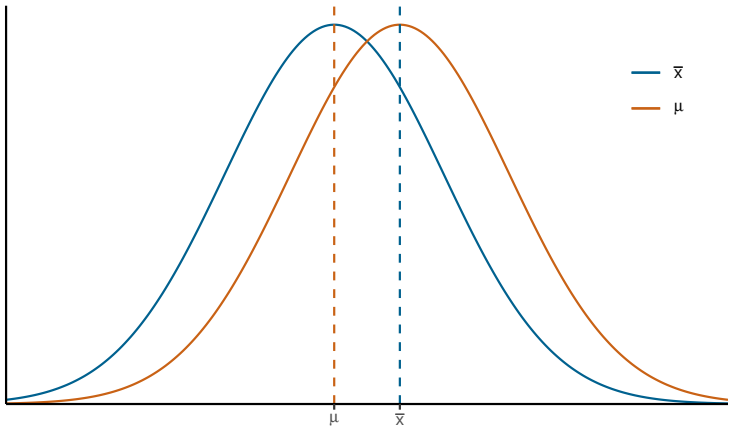


Abbildung 4.4: Konstruierte Verteilung um $\bar{x}$

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).

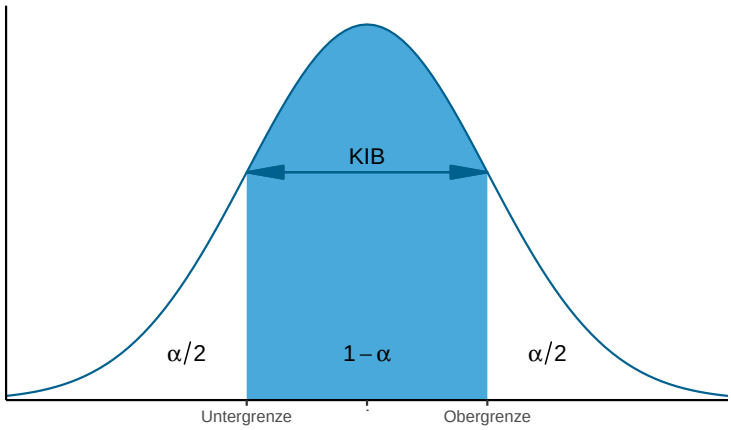


Abbildung 4.5: Konfidenzintervall

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) den Standardfehler der theoretischen Verteilung von $\bar{x}$:

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

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:

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

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

[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…“

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.

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) ein:

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

Dies lässt sich umformen in:

$$1 - \alpha \approx P(z_\mu < 1{,}08) - P(z_\mu < -1{,}08)$$

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

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

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

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:

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

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 $\bar{x} \pm \frac{KIB}{2}$.

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) 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) 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.

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?

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?

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?

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?

Sitzung 5

Grundlagen der Teststatistik

Lernziele dieser Sitzung

Sie können…

- Hypothesen formulieren.
- einen z-Test durchführen.
- einen 1-Stichproben-t-Test durchführen.

Lernvideos (Sommersemester 2020)

- 5a) z-Test
- 5b) 1-Stichproben-t-Test

5.1 Statistische Tests

Gemeinsam mit der Schätzstatistik bildet die Test- bzw. Prüfstatistik jenen Teil statistischer Verfahren, die ausgehend von einer Stichprobenverteilung Rückschlüsse auf die Beschaffenheit von Grundgesamtheiten anstreben (schließende Statistik).

Dabei haben Schätz- und Teststatistik jedoch grundlegend verschiedene Vorgehensweisen. Wie in Sitzung 4 besprochen ermöglicht die Schätzstatistik die Angabe statistischer Parameter einer Grundgesamtheit anhand von Stichprobenwerten, und unter Angabe von Wahrscheinlichkeiten.

Ziel statistischer Tests hingegen ist es, mit Hilfe von Stichproben Hypothesen (also Vermutungen) über die Grundgesamtheit zu prüfen. Geprüft wird dabei ein empirischer Sachverhalt gegen die Zufälligkeit seiner Realisierung. Ein statistischer Test fragt, ab welcher Größenordnung ein Stichprobenergebnis nicht mehr als zufällig, sondern als *signifikant* anzusehen ist.

Dabei folgt die grundsätzliche Vorgehensweise von (hier behandelten) statistischen Tests immer diesem Schema:

1. Test wählen und Voraussetzungen prüfen
2. Hypothesen formulieren
3. Signifikanzniveau entscheiden
4. Ablehnungsbereich bestimmen
5. Prüfgröße berechnen

6. Ergebnis interpretieren

Die einzelnen Schritte werden im Folgenden direkt anhand des z-Tests besprochen.

5.2 z-Test

Die mathematischen Grundlagen des z-Tests leiten sich direkt aus der in Sitzung 4 besprochenen Stichprobenverteilung für $\bar{x}$ ab.

Ein illustrierendes Beispiel: Wir wissen, dass die Anzahl der täglichen Besucher*innen einer Eissporthalle annähernd normalverteilt ist, und zwar mit dem arithmetischen Mittel $\mu = 94{,}2$ und der Standardabweichung $\sigma = 11{,}8$. Wir vermuten, dass die Anzahl der Besucher*innen an bewölkten Tagen größer ist, weil an sonnigen Tagen andere Freizeitbeschäftigungen attraktiver sind.

An fünf zufälligen bewölkten Tagen zählen wir die Besucher*innen und kommen auf einen Mittelwert der Stichprobe von $\bar{x} = 103{,}0$.

Dieser Wert ist höher als das arithmetische Mittel der Grundgesamtheit (μ). Aber heißt das auch, dass unsere Vermutung stimmt? Wir wissen aus Sitzung 4, dass die Stichprobenverteilung einem Standardfehler ($\sigma_{\bar{x}}$) unterliegt (s. Abbildung 5.1).

Ist das Ergebnis also nur zufällig zustande gekommen, oder liegt ein *statistisch signifikantes* Ergebnis vor? Mit anderen Worten: Ist die Stichprobe überhaupt der Verteilung x_0 um μ_0 entnommen, oder gibt es eine *andere* Verteilung (x um ein anderes μ) für bewölkte Tage, denen unser Stichprobenmittelwert $\bar{x}$ entstammt? Genau diese Art von Frage versuchen statistische Tests zu beantworten.

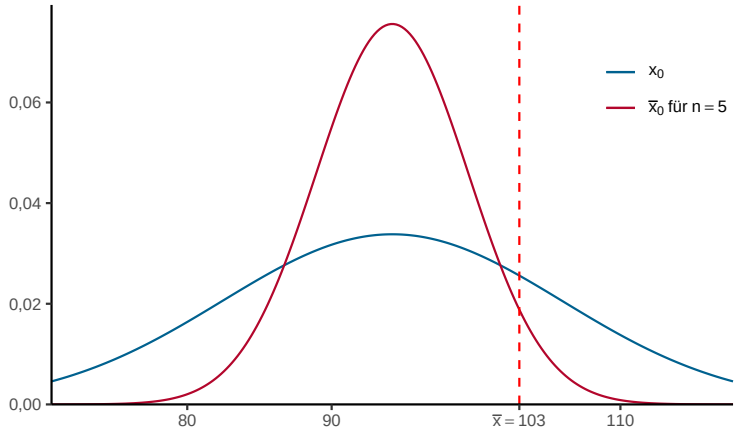


Abbildung 5.1: Theoretische Stichprobenverteilung (unter Annahme der Nullhypothese)

Test wählen und Voraussetzungen prüfen

Je nachdem, was überprüft werden soll, was über die Grundgesamtheit bekannt ist und wie die Stichprobe beschaffen ist, müssen verschiedene Testverfahren angewendet werden.

Statistische Tests unterscheiden sich zunächst in Bezug auf ihre Prüfgröße (und sind auch nach ihrer Prüfgröße benannt). Wir werden zunächst den z-Test kennenlernen, der mit dem (uns seit Sitzung 3 bekannten) z-Wert als Prüfgröße arbeitet.

Der z-Test hat zum Ziel, den Mittelwert einer Stichprobe mit den zu erwartenden Werten bei einer bekannten Verteilung zu vergleichen.

Um den z-Test anwenden zu können, müssen also folgende Voraussetzungen gegeben sein:

- Das Ziel der Untersuchung ist es, eine signifikante Abweichung des Mittelwerts festzustellen.
- Das arithmetische Mittel μ und die Standardabweichung σ der (ursprünglichen) Grundgesamtheit müssen bekannt sein.
- Der Test muss anhand einer reinen Zufallsstichprobe erfolgen.
- Die Stichprobenverteilung muss (annähernd) normalverteilt sein, das heißt:
  - *entweder* die Grundgesamtheit ist (annähernd) normalverteilt,
  - *oder* die Stichprobe hat die Größe $n \geq 30$.

Beispiel

In unserem Beispiel (Besuchszahlen der Eissporthalle) sind diese Voraussetzungen gegeben. Wir können und wollen also einen z-Test durchführen.

Hypothesen formulieren

Es müssen immer zwei Hypothesen formuliert werden: die Nullhypothese und die Alternativhypothese. Die Nullhypothese geht immer davon aus, dass es keine Abweichung gibt, die Alternativhypothese formuliert eine Abweichung.

Dabei werden zwei Verteilungen konstruiert: Die bekannte Grundgesamtheit (in unserem Beispiel: Besuchszahlen insgesamt) x_0 mit Mittelwert μ_0 und eine neue Verteilung (Besuchszahlen an bewölkten Tagen) x mit Mittelwert μ.

Die Hypothesen sind theoriegeleitet (formulieren also eine begründete Vermutung) und stehen stets am Anfang der statistischen Untersuchung. Es ist unzulässig, sie im Nachhinein anzupassen.

Nullhypothese

Die Nullhypothese (engl. *null hypothesis*) geht immer davon aus, das die forscherische Vermutung nicht stimmt. Im z-Test besagt die Nullhypothese, dass es zwischen dem Mittelwert μ_0 und dem Mittelwert μ keinen Unterschied gibt. Generell heißt die Nullhypothese:

$$H_0 : \mu = \mu_0$$

Alternativhypothese

Die Alternativhypothese (engl. *alternative hypothesis*) stellt die Vermutung dar, die überprüft werden soll. Dabei gibt es zwei unterschiedliche Möglichkeiten: ungerichtete und gerichtete Alternativhypothesen.

Ungerichtete Alternativhypothese Die ungerichtete Alternativhypothese besagt nur, *dass* es einen Unterschied zwischen μ und μ_0 gibt, aber nicht in welche Richtung (größer oder kleiner). Sie lautet daher:

$$H_1 : \mu \neq \mu_0$$

Gerichtete Alternativhypothese Die gerichtete Alternativhyptothese gibt eine Richtung des vermuteten Unterschieds (nach oben oder unten) vor. Sie lautet entweder:

$$H_1 : \mu < \mu_0 \quad (\text{abwärts gerichtet})$$

oder:

$$H_1 : \mu > \mu_0 \quad (\text{aufwärts gerichtet})$$

Beispiel

In unserem Beispiel geben wir eine Richtung vor, denn wir vermuten ja, dass die Besuchszahlen an bewölkten Tagen *höher* sind. Wir schreiben also:

$$H_0 : \mu = 94{,}2$$
$$H_1 : \mu > 94{,}2$$

Signifikanzniveau entscheiden

Das Signifikanzniveau α (engl. *significance level*) entscheidet, wie *unwahrscheinlich* eine Prüfgröße unter Annahme der Nullhypothese sein muss, damit wir die Nullhypothese ablehnen können (und damit unsere Annahme bestätigen).

Übliche Werte für das Signifikanzniveau sind $\alpha = 0{,}05$ oder $\alpha = 0{,}01$.

Für die Wahl des Signifikanzniveaus ist jeweils der Kontext entscheidend: Wenn die irrtümliche Bestätigung der forscherischen Annahme gravierende Auswirkungen hat, möchte man das Signifikanzniveau besonders niedrig wählen um diese Art von Fehler auszuschließen.

Auch das Signifikanzniveau muss vor der statistischen Erhebung formuliert werden, und es ist unzulässig, es im Nachhinein an das Ergebnis anzupassen.

Beispiel

Ein Irrtum in der statistischen Signifikanz der Besucherzahl hat vermutlich keine gravierenden Folgen. Wir legen das Signifikanzniveau auf $\alpha = 0{,}05$ fest.

Ablehnungsbereich bestimmen

Zusammen mit der (Un-)Gerichtetheit der Alternativhypothese bestimmt das Signifikanzniveau α den *Ablehnungsbereich* – also den Bereich für die zu errechnende Prüfgröße z, in dem die Nullhypothese abgelehnt würde.

Der Ablehnungsbereich für die ungerichtete Alternativhypothese ist $\frac{\alpha}{2}$ auf beiden Seiten (s. Abbildung 5.2). Die kritischen Werte sind dann die Schwellen des Ablehnungsbereich auf beiden Seiten:

$$z \leq z_{\alpha/2} \quad \text{und} \quad z \geq z_{(1-\alpha/2)} \quad \text{für} \quad H_1 : \mu \neq \mu_0$$

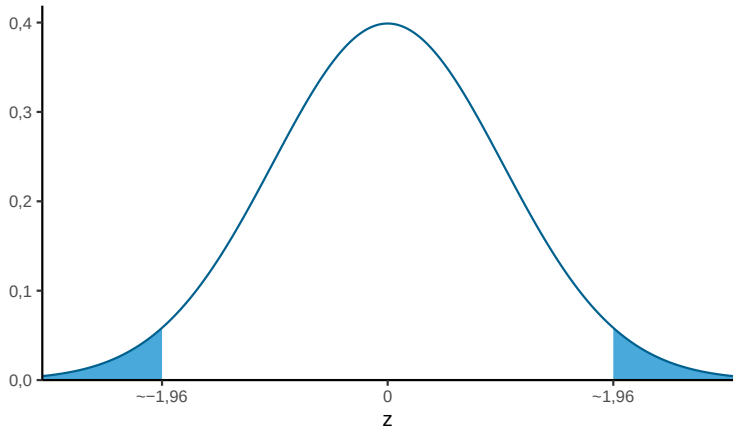


Abbildung 5.2: Kritische Werte für z bei ungerichteter Alternativhypothese und $\alpha = 0{,}05$

Bei den gerichteten Alternativhypothesen ist der Ablehnungsbereich jeweils nur auf einer Seite (s. Abbildungen 5.3 und 5.4). Die kritischen Werte ergeben sich aus:

$$z \leq z_{\alpha} \quad \text{für} \quad H_1 : \mu < \mu_0$$
$$z \geq z_{(1-\alpha)} \quad \text{für} \quad H_1 : \mu > \mu_0$$

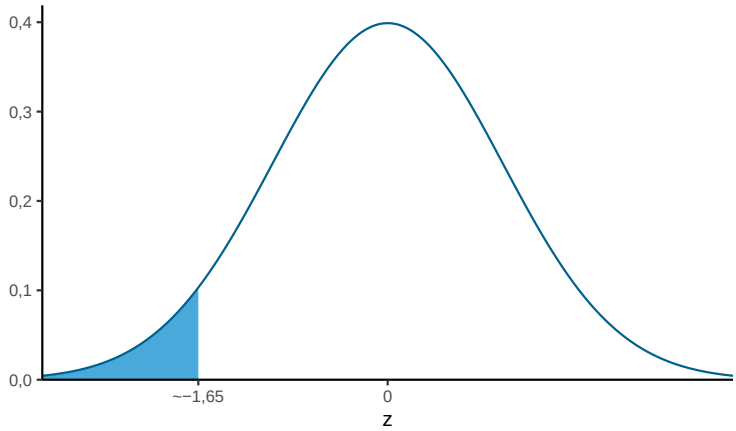


Abbildung 5.3: Kritischer Wert für z bei gerichteter Alternativhypothese nach unten und $\alpha = 0{,}05$

Beispiel

In unserem Beispiel haben wir eine gerichtete Alternativhypothese nach oben und ein Signifikanzniveau von $\alpha = 0{,}05$ verwendet. Der kritische Wert (bei dessen Überschreitung wir die Nullhypothese ablehnen und unsere Vermutung bestätigt sehen) lautet also:

$$z \geq z_{95\%} \approx 1{,}65$$

Der Mittelwert unserer Stichprobe fällt höher aus als μ. Aber übersteigt er auch den kritischen Wert (und ist damit statistisch signifikant)?

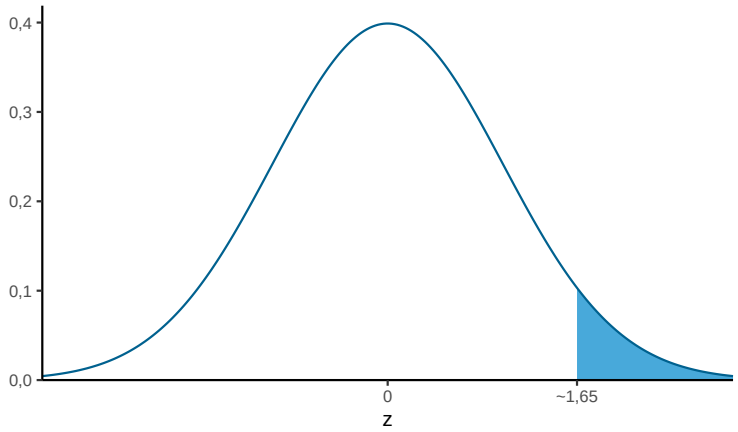


Abbildung 5.4: Kritischer Wert für z bei gerichteter Alternativhypothese nach oben und $\alpha = 0{,}05$

Prüfgröße berechnen

Für den z-Test ist die Prüfgröße der z-Wert der Stichprobe, und zwar standardisiert in Bezug auf μ_0 und den Standardfehler ($\sigma_{\bar{x}}$):

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

Wie wir bereits wissen, ergibt sich der Standardfehler ($\sigma_{\bar{x}}$) wiederum aus der Stichprobengröße (n) und der Standardabweichung der Grundgesamtheit (σ):

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

Durch einsetzen ergibt sich die generelle Formel für die Prüfgröße des z-Tests:

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

Das grundsätzliche Schema dieser Formel werden wir in anderen Tests wiedererkennen.

Beispiel

An dieser Stelle (also *nachdem* wir uns für einen Test und ein Signifikanzniveau entschieden und den kritischen Wert berechnet haben) dürften wir streng genommen erst die Stichprobe erheben.

Diese ergibt bei $n = 5$ den Mittelwert $\bar{x} = 103{,}0$. Die Verteilung x_0 (also unter Annahme der Nullhypothese) hatte die Kennwerte $\mu_0 = 94{,}2$ und $\sigma = 11{,}8$.

Wir setzen ein in die Formel aus Gleichung (5.2):

$$
\begin{aligned}
z &= \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\sigma} \\
&\approx \sqrt{5} \cdot \frac{103{,}0 - 94{,}2}{11{,}8} \\
&\approx 1{,}67
\end{aligned}
$$

Ergebnis inerpretieren

Je nachdem, ob die Prüfgröße in den Ablehnungsbereich fällt (ob der kritische Wert also unter- bzw. überschritten wird), können wir die Nullhypothese ablehnen (und damit unsere Alternativhypothese bestätigen) oder nicht.

Eine Ablehnung der Nullhypothese bedeutet, dass wir ein *statistisch signifikantes Ergebnis zugunsten unserer Vermutung* vorliegen haben.

Diese Art von Ergebnis wird oft falsch interpretiert. Bei einem Signifikanzniveau von $\alpha = 0{,}01$ heißt das zum Beispiel, dass die beobachteten Werte nur mit 1% Wahrscheinlichkeit vorkommen, wenn unsere Vermutung *nicht* stimmt. Wichtig dabei: Das ist etwas ganz anderes als zu behaupten, dass unsere Vermutung zu 99% stimme. Über die Wahrscheinlichkeit, dass eine Hypothese stimmt (oder nicht) können wir mit den Methoden der klassischen Statistik keine Aussage machen!

Beispiel

In unserem Beispiel liegt der z-Wert knapp über dem kritischen Wert von 1,65. Wir können also die Nullhypothese ablehnen und unsere Alternativhypothese annehmen. Unsere statistische Untersuchung hat gezeigt, dass die Eissporthalle an bewölkten Tagen besser besucht ist als an sonnigen (und zwar mit Signifikanzniveau $\alpha = 0{,}05$).

Gut, dass wir eine gerichtete Alternativhypothese aufgestellt haben. Hätten wir nur vermutet, dass sich die Besuchszahlen je nach Wetter unterscheiden (ohne Angabe einer Richtung), dann wäre der kritische Wert nicht erreicht worden und wir hätten die Nullhypothese beibehalten müssen. Hinterher die Hypothesen anzupassen ist natürlich nicht zulässig!

Softwarehinweis

R hat in der Grundversion keinen dezidierten Befehl für einen z-Test. Mit der Funktion `qnorm()` können kritische Werte jedoch einfach bestimmt werden.

5.3 Die t-Verteilung

Wenn die Standardabweichung σ eines Merkmals in der Grundgesamtheit *unbekannt* ist, kann sie durch die Standardabweichung s der Stichprobe geschätzt werden (s. Sitzung 4). Dann ist die Stichprobenverteilung für $\bar{x}$ jedoch nicht mehr normalverteilt, sondern sie folgt einer t-Verteilung.

Im Gegensatz zur Standardnormalverteilung (die wir für den z-Test benutzen) gibt es aber nicht nur eine t-Verteilung, sondern die Form der t-Verteilung hängt von so genannten Freiheitsgraden (engl. *degrees of freedom*) ab. Mit steigender Zahl der Freiheitsgrade nähert sich die t-Verteilung einer Standardnormalverteilung an (s. Abbildung 5.5).

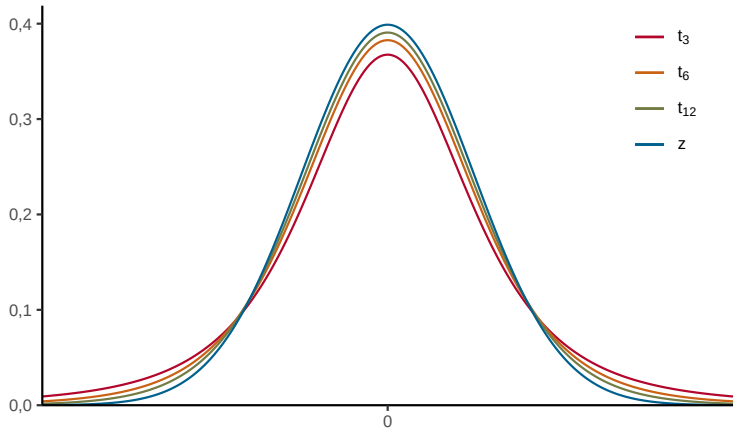


Abbildung 5.5: t-Verteilungen mit verschiedenen Freiheitsgraden

Freiheitsgrade

In Anlehnung an Bortz und Schuster (2010) kürzen wir Freiheitsgrade mit df ab. Dort findet sich auch eine brauchbare Erklärung dieses Phänomens:

> Die Freiheitsgrade, welche mit einem Kennwert verbunden sind, entsprechen der Anzahl der Werte, die bei seiner Berechnung frei variieren können. Der Mittelwert $\bar{x}$ besitzt beispielsweise n Freiheitsgrade, weil es keinerlei Bedingung gibt, der die n Werte genügen müssen. Dies ist für die Varianz $s^2 = QS/(n-1)$ nicht der Fall. Nur $n-1$ Abweichungen, welche in die Berechnung der Quadratsumme $QS = \sum_i (x_i - \bar{x})^2$ eingehen, können frei variieren. [D]ie Summe der Abweichungen von ihrem Mittelwert [ist] null, d.h. $\sum_i (x_i - \bar{x}) = 0$. Von n Abweichungen können deshalb nur $n-1$ frei variieren. Ergeben sich beispielsweise bei einer Stichprobe aus drei Werten die Abweichungen $x_1 - \bar{x} = -4$ und $x_2 - \bar{x} = 0$, muss zwangsläufig $x_3 - \bar{x} = 4$ sein, damit die Summe aller Abweichungen null ergibt. Bei der Varianzberechnung ist eine der n Abweichungen festgelegt, d.h. die Varianz hat nur $n-1$ Freiheitsgrade. Man schreibt die Stichprobenvarianz deshalb gelegentlich auch als $s^2 = QS/df$. Da die Varianz mit $n-1$ Freiheitsgraden verbunden ist, gilt dies auch für die Standardabweichung s. (Bortz und Schuster 2010: 121)

5.4 1-Stichproben-t-Test

Der 1-Stichproben-t-Test vergleicht (wie der z-Test) die Werte einer Stichprobe mit der Grundgesamtheit. Das Vorgehen ist dabei analog zum z-Test, mit dem einzigen Unterschied, dass eine t-Verteilung mit $(n-1)$ Freiheitsgraden herangezogen wird.

Wir besprechen den 1-Stichproben-t-Test direkt an einem Beispiel:

Beim Frankfurter Amt für Wohnungswesen betrage die durchschnittliche Bearbeitungsdauer von Anträgen auf Wohngeld 30,2 Tage und sei normalverteilt. Wir vermuten, dass die Bearbeitungszeit zu Anfang des Wintersemesters höher ist als im Jahresdurchschnitt und planen eine zufällige Stichprobe von 12 Anträgen mit Einreichungsdatum im Oktober.

Test wählen und Voraussetzungen prüfen

Um den 1-Stichproben-t-Test durzuführen müssen folgende Voraussetzungen erfüllt sein:

- Das Ziel der Untersuchung ist es, eine statistisch signifikante Abweichung des Mittelwerts einer Stichprobe im Vergleich zu einer Grundgesamtheit festzustellen.
- Das zu untersuchende Merkmal ist in der Grundgesamtheit normalverteilt.
- Das arithmetische Mittel (μ) des Merkmals in der Grundgesamtheit ist bekannt. (Im Gegensatz zum z-Test ist σ hier unbekannt!)
- Der Test erfolgt anhand einer reinen Zufallsstichprobe.

Beispiel

In unserem Beispiel (Bearbeitungszeit Wohngeldanträge) sind diese Bedingungen erfüllt und wir können einen 1-Stichproben-t-Test durchführen.

Hypothesen formulieren

Die Hypothesen werden genauso wie beim z-Test formuliert:

Nullhypothese

$$H_0 : \mu = \mu_0$$

Alternativhypothese

$$H_1 : \mu \neq \mu_0 \quad (\text{ungerichtet})$$

oder

$$H_1 : \mu < \mu_0 \quad (\text{abwärts gerichtet})$$

oder

$$H_1 : \mu > \mu_0 \quad (\text{aufwärts gerichtet})$$

Beispiel

In unserem Beispiel geben wir eine Richtung vor, denn wir vermuten ja, dass die Bearbeitungsdauer zu Semesteranfang *höher* ist. Wir schreiben also:

$$H_0 : \mu = 30{,}2$$
$$H_1 : \mu > 30{,}2$$

Signifikanzniveau entscheiden

Wie beim z-Test entscheidet das Signifikanzniveau α, wie *unwahrscheinlich* eine Prüfgröße unter Annahme der Nullhypothese sein muss, damit wir die Nullhypothese ablehnen können (und damit unsere Annahme bestätigen).

Übliche Werte für das Signifikanzniveau sind auch beim t-Test $\alpha = 0{,}05$ oder $\alpha = 0{,}01$.

Beispiel

Ein Irrtum zugunsten der Alternativhypothese hat bei unserer Untersuchung keine gravierenden Folgen. Angenommen, wir wollen uns in der Analyse trotzdem ganz sicher sein. Dann entscheiden wir uns für das Signifikanzniveau $\alpha = 0{,}01$.

Ablehnungsbereich bestimmen

Genau wie beim z-Test bestimmt das Signifikanzniveau α den *Ablehnungsbereich* – also den Bereich für die zu errechnende Prüfgröße t, in dem die Nullhypothese abgelehnt würde.

Der Ablehnungsbereich für die ungerichtete Alternativhypothese ist $\frac{\alpha}{2}$ auf beiden Seiten. Die kritischen Werte sind dann die Schwellen des Ablehnungsbereich auf beiden Seiten:

$$t \leq t_{df;\alpha/2} \quad \text{und} \quad t \geq t_{df;(1-\alpha/2)} \quad \text{für} \quad H_1 : \mu \neq \mu_0$$

Bei den gerichteten Alternativhypothesen ist der Ablehnungsbereich jeweils nur auf einer Seite. Die kritischen Werte ergeben sich aus:

$$t \leq t_{df;\alpha} \quad \text{für} \quad H_1 : \mu < \mu_0$$

$$t \geq t_{df;(1-\alpha)} \quad \text{für} \quad H_1 : \mu > \mu_0$$

Die kritischen Werte für t bei gegebenem Freiheitsgrad $(n-1)$ und Flächenabschnitt lassen sich aus der Formelsammlung ablesen. Dabei ist zu beachten, dass aufgrund der Symmetrie die Werte für Flächenanteile unter 50% nicht in der Tabelle verzeichnet sind. Es gilt die Formel:

$$P(-t_{df}) = 1 - P(t_{df})$$

So ist zum Beispiel der Wert für $t_{5;1\%} = -t_{5;99\%} = -3{,}365$.

Beispiel

In unserem Beispiel haben wir eine gerichtete Alternativhypothese nach oben und ein Signifikanzniveau von $\alpha = 0{,}01$ verwendet. Wir haben uns zudem für eine Stichprobengröße von $n = 12$ entschieden, woraus der Freiheitsgrad $df = n - 1 = 11$ resultiert.

Der kritische Wert (bei dessen Überschreitung wir die Nullhypothese ablehnen und unsere Vermutung bestätigt sehen) lautet also:

$$t \geq t_{df;(1-\alpha)}$$
$$t \geq t_{11;99\%}$$
$$t \geq 2{,}718$$

Graphisch ist der Ablehnungsbereich für unser Beispiel in Abbildung 5.6 dargestellt.

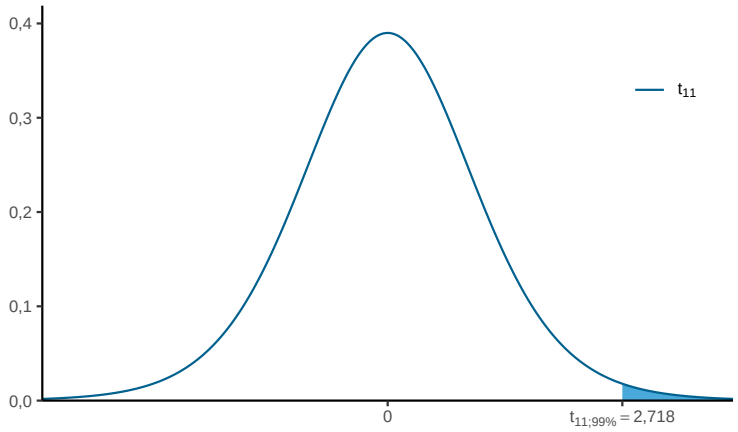


Abbildung 5.6: Ablehnungsbereich bei gerichteter Alternativhypothese nach oben, $n = 12$ und $\alpha = 0{,}01$

Prüfgröße berechnen

Die Formel für die Berechnung der Prüfgröße t im 1-Stichproben-t-Test lautet ganz ähnlich wie die für die Prüfgröße z im z-Test – mit dem Unterschied, dass statt der (hier unbekannten) Standardabweichung der Grundgesamtheit (σ) die Standardabweichung der Stichprobe (s) eingesetzt wird:

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

Zum direkten Vergleich noch einmal die Prüfgröße im z-Test:

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

Beispiel (ausführlich)

Wir erheben die Stichprobe von $n = 12$ Anträgen im Oktober und erhalten folgende Werte für die Bearbeitungsdauer (in Tagen):

$$45 \quad 41 \quad 37 \quad 41 \quad 35 \quad 44 \quad 34 \quad 44 \quad 38 \quad 41 \quad 39 \quad 36$$

Wir errechnen zunächst das arithmetische Mittel $\bar{x}$ (s. Sitzung 2):

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{45 + 41 + 37 + 41 + 35 + 44 + 34 + 44 + 38 + 41 + 39 + 36}{12} \\ &\approx 39{,}58\end{aligned}$$

Damit können wir die Standardabweichung s berechnen:

$$
\begin{aligned}
s &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \\
&\approx \sqrt{\frac{29{,}38 + 2{,}02 + 6{,}66 + 2{,}02 + 20{,}98 + 19{,}54 + 31{,}14 + 19{,}54 + 2{,}5 + 2{,}02 + 0{,}34 + 12{,}82}{11}} \\
&\approx 3{,}67
\end{aligned}
$$

Schließlich setzen wir diese Werte in die Formel für die Prüfgröße t (5.4) ein:

$$
\begin{aligned}
t &= \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s} \\
&\approx \sqrt{12} \cdot \frac{39{,}58 - 30{,}2}{3{,}67} \\
&\approx 8{,}854
\end{aligned}
$$

Ergebnis interpretieren

Genau wie beim z-Test kommt es darauf an, ob die Prüfgröße in den Ablehnungsbereich fällt (ob der kritische Wert also unter- bzw. überschritten wird). Wenn dies der Fall ist, können wir die Nullhypothese ablehnen (und damit unsere Alternativhypothese bestätigen). Wenn nicht, müssen wir die Nullhypothese beibehalten.

Beispiel

In unserem Beispiel liegt der t-Wert deutlich über dem kritischen Wert von 2,718. Wir können also die Nullhypothese ablehnen und unsere Alternativhypothese annehmen. Unsere statistische Untersuchung hat gezeigt, dass die Bearbeitungsdauer von Anträgen, die im Oktober eingehen, länger ist als im Jahresdurchschnitt (und zwar mit Signifikanzniveau $\alpha = 0{,}01$).

Softwarehinweis

In R kann ein t-Test mit dem Befehl `t.test()` durchgeführt werden. Neben der Prüfgröße t gibt der Befehl einen p-Wert aus – ist dieser kleiner als α, so liegt eine signifikante Abweichung vor.

Tipps zur Vertiefung

- YouTube-Kanal „Kurzes Tutorium Statistik“: p-Wert, Nullhypothese, Signifikanzniveau - die Idee erklärt
- YouTube-Kanal „Benedict K“: p-Wert: einseitiger und beidseitiger Hypothesentest / Signifikanztest - erklärt
- YouTube-Kanal „Kurzes Tutorium Statistik“: Einstichproben t-Test
- Kapitel 7, 8.1 in Bortz und Schuster (2010)
- Kapitel 8.2.2.1 und 8.2.3 in Lange und Nipper (2018)
- Kapitel 5.5.2 in Bahrenberg, Giese und Nipper (2010)

- Kapitel 9 in Klemm (2002)
- *Englisch:* Kapitel 9.1 in Burt und Barber (1996)

Übungsaufgaben

Aufgabe 5-1

zur Lösung

Sie interessieren sich für die durchschnittliche Haushaltsgröße in Frankfurt im europäischen Vergleich. In der EU sei die durchschnittliche Haushaltsgröße 2,30 Personen mit einer Standardabweichung von 1,42.

Sie vermuten, dass Frankfurter Haushalte sich in ihrer Größe vom europäischen Durchschnitt unterscheiden, können aber nicht sagen, in welche Richtung.

a) Welche Stichprobengröße ist für einen z-Test in diesem Fall nötig und warum?
b) Formulieren Sie Null- und Alternativhypothese.
c) Sie entscheiden Sich für ein Signifikanzniveau von $\alpha = 0{,}05$. Notieren Sie die kritischen Werte.
d) Eine Stichprobe von 40 Frankfurter Haushalten ergibt eine durchschnittliche Größe von 1,82. Berechnen Sie die Prüfgröße z.
e) Wie bewerten Sie das Ergebnis?

Aufgabe 5-2

zur Lösung

Bestimmen Sie die folgenden kritischen Werte:

a) $t_{4;0{,}5\%}$
b) $t_{19;0{,}1\%}$
c) $t_{7;2{,}5\%}$
d) $t_{13;5\%}$
e) $t_{11;97{,}5\%}$
f) $t_{3;95\%}$
g) $t_{6;99{,}5\%}$
h) $t_{16;99{,}9\%}$
i) $t_{5;99\%}$
j) $t_{20;1\%}$

Aufgabe 5-3

zur Lösung

Die Prüfungsergebnisse für eine Klausur im Geographiestudium seien normalverteilt mit einer mittleren Punktzahl von 61,5 und einer Standardabweichung von 10,3. Sie vermuten, dass berufstätige Studierende im Durchschnitt schlechter abschneiden, weil ihnen die Vorbereitungszeit fehlt. Eine Zufallsstichprobe berufstätiger Studierender ergibt die Prüfungsergebnisse: 42 78 46 65

Prüfen Sie Ihre Vermutung. Begründen Sie die Wahl des Tests und des Signifikanzniveaus.

Aufgabe 5-4

zur Lösung

Sie vermuten, dass Angestellte mit Migrationshintergrund in einem bestimmten Betrieb weniger als das Durchschnittsgehalt verdienen. Die Personalabteilung bestätigt Ihnen gegenüber die annähernde Normalverteilung der Bruttogehälter mit Mittelwert $\mu = 3042{,}43$ (in EUR). Sie planen, das Bruttogehalt von sechs zufälligen Angestellten mit Migrationshintergrund direkt zu ermitteln.

a) Welchen Test führen Sie durch?
b) Formulieren Sie die Hypothesen.
c) Bestimmen Sie den kritischen Wert bei Signifikanzniveau $\alpha = 0{,}01$.

Aufgabe 5-5

zur Lösung

(Fortführung von Aufgabe 4)

Sie ermitteln die folgenden Werte (in EUR):

$$2927{,}35 \quad 2930{,}68 \quad 2903{,}58 \quad 3032{,}59 \quad 3013{,}37 \quad 2979{,}4$$

a) Berechnen Sie die Prüfgröße.
b) Welche Schlüsse ziehen Sie aus der Untersuchung?

Aufgabe 5-6

zur Lösung

In Ermberg ist die Verteilung der Mietpreise für Ladenflächen pro Quadratmeter (in €) annähernd normalverteilt mit Mittelwert 11,8 und Varianz 5,2.

Die Baudezernentin sagt, dass die Ladenmieten im Neubaugebiet Auwiese deutlich günstiger seien als im Gemeindedurchschnitt. Um die Behauptung zu überprüfen, erheben Sie die folgende Zufallsstichprobe von Ladenmieten im Neubaugebiet (pro m² in €):

$$8{,}54 \quad 7{,}16 \quad 14{,}47 \quad 11{,}84 \quad 10{,}27$$

Prüfen Sie die Behauptung. Schließen Sie einen Fehler 1. Art zu 95% aus.

Aufgabe 5-7

zur Lösung

In einem landwirtschaftlichen Großbetrieb wird ein neues Düngemittel für Zuckerrüben getestet. Zunächst wird es nur auf sechs zufällig ausgewählten Feldern (von 60) eingesetzt.

Der durchschnittliche Ertrag aller 60 Felder beträgt 69 Tonnen pro Hektar (t/ha). Für die sechs Felder mit dem neuen Düngemittel wurden folgende Ertragswerte erhoben:

Feld	Ertrag in t/ha
1	93
2	74
3	65
4	69
5	89
6	85

Prüfen Sie, ob der Einsatz des neuen Düngemittels zu einem signifikanten Unterschied im Ertrag der Felder geführt hat. Akzeptieren Sie in Ihrer Analyse 5% als Wahrscheinlichkeit für einen Fehler 1. Art.

Sitzung 6

Testverfahren mit zwei Stichproben

Lernziele dieser Sitzung

Sie können…

- einen 2-Stichproben-t-Test durchführen.
- einen F-Test durchführen.
- Fehler 1. und 2. Art unterscheiden.

Lehrvideos (Sommersemester 2020)

- 6a) 2-Stichproben-t-Test
- 6a) F-Test
- 6c) Fehler 1. und 2. Art
  - Ich rede am Ende über mögliche Klausurformen. Letztes Jahr haben wir eine „Online-Papierklausur“ geschrieben, das machen wir dieses Jahr auch.

6.1 Statistische Tests

In Sitzung 5 haben wir mit dem z-Test und dem 1-Stichproben-t-Test die ersten Testverfahren kennengelernt. In dieser Sitzung kommt der 2-Stichproben-t-Test sowie der F-Test dazu.

Das grundsätzliche Verfahren bleibt dabei stets das gleiche. Zur Erinnerung noch einmal die sechs Schritte:

1. Test auswählen und Voraussetzungen prüfen
2. Hypothesen formulieren
3. Signifikanzniveau entscheiden
4. Ablenhnungsbereich bestimmen
5. Prüfgröße berechnen
6. Ergebnis interpretieren

6.2 2-Stichproben-t-Test {#2-stichproben-t-test}

Bei der folgenden Variante des t-Tests (und beim F-Test) wird nicht wie gehabt *eine* Stichprobe auf signifikante Abweichungen *von der Grundgesamtheit* überprüft, sondern *zwei* Stichproben auf signifikante

Abweichungen *voneinander.* An den sechs Schritten ändert sich nichts.

Den 2-Stichproben-t-Test gibt es je nach Voraussetzungen bzw. Annahmen in vielen unterschiedlichen Varianten. In dieser Veranstaltung wird nur eine bestimmte (vergleichsweise einfache) Variante behandelt. In der Praxis geht es aber oft darum, für ganz bestimmte empirische Bedinungen den „richtigen" 2-Stichproben-t-Test auszuwählen.

Die hier behandelte Variante soll mit folgendem Beispiel illustriert werden: Wir interessieren uns für die Mietpreise von kleine Gewerbeflächen in den beiden Frankfurter Stadtteilen Höchst und Praunheim. Wir vermuten, dass es einen signifikanten Unterschied gibt, wissen aber nicht in welche Richtung. Wir planen eine Befragung von je 6 Mieter*innen von kleinen Gewerberäumen, die nach Zufallsprinzip ausgewählt werden.

Test wählen und Voraussetzungen prüfen

Der hier behandelte 2-Stichproben-t-Test hat folgende Voraussetzungen (bzw. Annahmen):

- Es soll untersucht werden, ob ein Merkmal in zwei Stichproben signifikant voneinander abweicht.
- Die Stichproben sind einfache Zufallsstichproben und unabhängig voneinander erhoben.
- Die Stichproben haben dieselbe Anzahl an Elementen ($n_1 = n_2$).
- Das Merkmal ist grundsätzlich (annähernd) normalverteilt.
- Die Varianzen der zu vergleichenden Populationen sind gleich ($\sigma_1^2 = \sigma_2^2$).

Die letzte Voraussetzung ist etwas merkwürdig, denn beim t-Test kennen wir ja die Varianzen der Grundgesamtheiten gar nicht. Der F-Test kann diese Voraussetzung anhand der Stichprobenverteilungen prüfen.

Beispiel

Probleme bereiten hier die Voraussetzungen der Normalverteilung und der gleichen Varianzen. Mit der Annahme der Normalverteilung können wir leben (weil wir uns mit Statistik gut auskennen und wissen, dass der t-Test „robust" auf nicht-ganz-normalverteilte Merkmale reagiert). Wenn sich während des Tests jedoch herausstellen sollte, dass die Varianzen zu unterschiedlich sind, müssten wir das Vorgehen neu überdenken.

Hypothesen formulieren

Im Unterschied zu zuvor besprochenen Verfahren gibt es hier keine übergeordnete Grundgesamtheit, und damit kein μ_0. Stattdessen werden Hypothesen über die Populationen der beiden Stichproben (μ_1 und μ_2) formuliert.

Nullhypothese

Die Nullhypothese geht davon aus, dass es keinen Unterschied zwischen den beiden Populationen gibt. Sie lautet daher:

$$H_0 : \mu_1 = \mu_2$$

Alternativhypothese

Die Alternativhypothese stellt üblicherweise die forscherische Vermutung dar, die überprüft werden soll. Dabei gibt es auch hier zwei unterschiedliche Möglichkeiten: ungerichtete und gerichtete Alternativhypothesen.

Ungerichtete Alternativhypothese Die ungerichtete Alternativhypothese besagt nur, *dass* es einen Unterschied zwischen μ_1 und μ_2 gibt, aber nicht in welche Richtung (größer oder kleiner). Sie lautet daher:

$$H_1 : \mu_1 \neq \mu_2$$

Gerichtete Alternativhypothese Die gerichtete Alternativhyptothese gibt eine Richtung des vermuteten Unterschieds vor. Sie lautet entweder:

$$H_1 : \mu_1 < \mu_2$$

oder:

$$H_1 : \mu_1 > \mu_2$$

Beispiel

Wir vermuten zwar einen Unterschied, wissen aber nicht in welche Richtung. Deshalb formulieren wir neben der Nullhypothese eine ungerichtete Alternativhypothese:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Signifikanzniveau entscheiden

Wie auch sonst sind übliche Werte hier $\alpha = 0{,}01$ und $\alpha = 0{,}05$.

Beispiel

Wir entscheiden uns für das Signifikanzniveau $\alpha = 0{,}05$.

Ablehnungsbereich bestimmen

Der kritische Wert wird genau wie bei dem 1-Stichproben-t-Test aus der Formelsammlung abgelesen. Der einzige (wichtige!) Unterschied ist die Bestimmung der Freiheitsgrade: Bei zwei Stichproben der Größe n werden die Freiheitsgrade bestimmt durch:

$$df = 2 \cdot n - 2$$

Beispiel

Wir planen mit je 6 Stichproben. Deswegen berechnen wir die Freiheitsgrade:

$$\begin{aligned} df &= 2 \cdot n - 2 \\ &= 2 \cdot 6 - 2 = 10 \end{aligned}$$

Kritische Werte gibt es nun in beide Richtungen. Aufgrund der Symmetrie der t-Verteilung reicht es, wenn wir einen Wert (mit $\alpha = 0{,}05$) nachschlagen:

$$\begin{aligned} t \leq t_{df;\alpha/2} \quad &\text{und} \quad t \geq t_{df;(1-\alpha/2)} \\ t \leq t_{10;2{,}5\%} \quad &\text{und} \quad t \geq t_{10;97{,}5\%} \\ t \leq -2{,}228 \quad &\text{und} \quad t \geq 2{,}228 \end{aligned}$$

Prüfgröße berechnen

Bei zwei Stichproben mit Mittelwert $\bar{x}_1$ bzw. $\bar{x}_1$ und Varianz s_1^2 bzw. s_2^2 lautet die Formel zur Bestimmung der Prüfgröße t:

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

Beispiel

Wir erheben folgende Werte für die Kaltmiete pro m²:

$\text{Höchst}(x_1):$	7,96	8,46	7,13	8,95	7,62	8,22
$\text{Praunheim}(x_2):$	5,54	5,80	8,70	7,99	6,23	6,75

Für die arithmetischen Mittel ergibt sich (s. Sitzung 2):

$$\begin{aligned} \bar{x}_1 &\approx 8{,}06 \\ \bar{x}_2 &\approx 6{,}84 \end{aligned}$$

Die Varianzen (s. Sitzung 2):

$$\begin{aligned} s_1^2 &\approx 0{,}41 \\ s_2^2 &\approx 1{,}59 \end{aligned}$$

Diese Varianzen sehen auf den ersten Blick sehr unterschiedlich aus, was ein Problem ist: Der 2-Stichproben-t-Test hat ja zur Annahme, dass die Varianzen in den beiden Populationen gleich sind.

Andererseits sind ja auch die Stichprobenvarianzen zu einem gewissen Grad Zufallsprodukte, und diese beiden Varianzen bewegen sich auch irgendwie noch in der selben Größenordnung – schließlich könnten sie auch 0,1 und 20 lauten.

Wir entscheiden uns zunächst dazu, den Test fortzuführen und lernen gleich eine Methode kennen, wie wir überprüfen können, ob das auch gerechtfertigt ist.

Um die Prüfgröße t zu bestimmen, setzen wir einfach unsere Stichprobenwerte in Formel (6.2) ein:

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2+s_2^2}{n}}} \\ &\approx \frac{8{,}06 - 6{,}84}{\sqrt{\frac{0{,}41+1{,}59}{6}}} \\ &\approx 2{,}113 \end{aligned}$$

Ergebnis interpretieren

Genau wie bei den anderen Tests wird je nach erreichen des kritischen Werts (des Ablehnungsbereichs) die Nullhypothese verworfen oder beibehalten.

Beispiel

Der kritische Wert von $t \geq 2{,}228$ wurde nicht überschritten. Wir müssen die Nullhypothese beibehalten, d.h. wir konnten keinen signifikanten Unterschied zwischen den Mietpreisen in Höchst und Praunheim feststellen ($\alpha = 0{,}05$).

Softwarehinweis

In R wird auch der 2-Stichproben-t-Test mit dem Befehl `t.test()` durchgeführt. Im Gegensatz zum 1-Stichproben-t-Test werden dabei zwei Verteilungen als Argumente eingegeben.

6.3 Die F-Verteilung

Die Prüfgröße F im F-Test ist unter Annahme der Nullhypothese F-verteilt. Im Gegensatz zu den Verteilungen von z und t ist die F-Verteilung nicht symmetrisch und nimmt nur positive Werte an (s. Abbildung 6.1).

Dazu ist die F-Verteilung nicht wie t von einem, sondern von zwei Freiheitsgraden abhängig. Die Reihenfolge dieser Freiheitsgrade ist auch wichtig: Wir sprechen vom Zähler-Freiheitsgrad (df_1) und vom Nenner-Freiheitsgrad (df_2). Die F-Verteilung wird also notiert mit: $F_{df_1;df_2}$

6.4 F-Test

Auch der F-Test untersucht zwei unabhängige Stichproben. Er unterscheidet sich jedoch insofern grundlegend von den zuvor besprochenen Testverfahren, als dass sein Untersuchungsgegenstand nicht der Mittelwert (μ) sondern die Varianz (σ^2) der beiden Populationen ist.

Die Prüfgröße F ist dann unter Annahme der Nullhypothese F-verteilt.

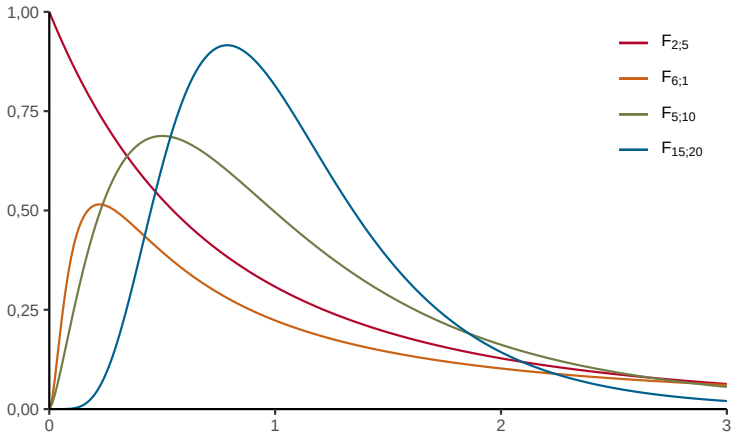


Abbildung 6.1: F-Verteilungen mit verschiedenen Freiheitsgraden

Unser Beispiel ist eine Fortführung des vorigen Beispiels für den 2-Stichproben-t-Test (Mietpreise für Gewerbeflächen). Uns interessiert: Sind die Varianzen eventuell so unterschiedlich, dass wir den obigen t-Test gar nicht hätten durchführen dürfen?

Test wählen und Voraussetzungen prüfen

Das Ziel des F-Tests ist die Feststellung eines signifikanten Unterschieds in der Varianz von zwei Populationen. Die Voraussetzungen lauten:

- Ausgangspunkt sind zwei unabhängig voneinander erhobene Stichproben (die aber grundsätzlich unterschiedlich groß sein dürfen).
- Das Merkmal ist in beiden Populationen (annähernd) normalverteilt.

Beispiel

Die Voraussetzung der Normalverteilung ist hier besonders wichtig, denn der Test wird bei anderen Verteilungen stark verfälscht. (Der F-Test ist also nicht „robust“, was die Normalverteilung angeht.)

Wir müssen also explizit die Annahme treffen, dass die Mietpreise annähernd normalverteilt sind. Das ist einerseits nicht ganz abwegig, andererseits würden wir in der Praxis unsere statistische Untersuchung dadurch angreifbar machen.

Hypothesen formulieren

Alles wie gehabt – nur, dass es um die Varianz σ^2 der jeweiligen Populationen geht.

Nullhypothese

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Alternativhypothesen

Ungerichtet

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Gerichtet

$$H_1 : \sigma_1^2 > \sigma_2^2$$

oder

$$H_1 : \sigma_1^2 < \sigma_2^2$$

Beispiel

Die Nullhypothese ist einfach:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Bei der Alternativhypothese ist die Ausgangslage, dass wir empirisch einen Unterschied zwischen $s_1^2 \approx 0{,}41$ und $s_2^2 \approx 1{,}59$ festgestellt haben. Die Frage, ob die Varianz der Mietpreise in Höchst *tatsächlich signifikant kleiner* ist, wird übersetzt in die Alternativhypothese:

$$H_1 : \sigma_1^2 < \sigma_2^2$$

Interessanterweise wäre hier (zur Abwechslung) das forscherische Interesse, die Nullhypothese beizubehalten – denn wir wollen ja den t-Test durchführen dürfen.

Signifikanzniveau entscheiden

Die Logik ist hier genau dieselbe: Wie unwahrscheinlich muss das empirische Ergebnis unter Annahme der Nullhypothese sein, damit wir diese ablehnen (müssen)?

Beispiel

Wir entscheiden uns für das (für unsere Zwecke sehr übliche) Signifikanzniveau von $\alpha = 0{,}05$.

Ablehnungsbereich bestimmen

Für die ungerichtete Alternativhypothese sind die kritischen Werte:

$$F \leq F_{df_1;df_2;\alpha/2} \quad \text{und} \quad F \geq F_{df_1;df_2;(1-\alpha/2)}$$

Für die gerichtete Alternativhypothese:

$$F \leq F_{df_1;df_2;\alpha}$$

bzw.

$$F \geq F_{df_1;df_2;(1-\alpha)}$$

Die Besonderheit der F-Verteilung ist, dass sie gleich von zwei Freiheitsgraden abhängt: dem Zähler-Freiheitsgrad df_1 und dem Nenner-Freiheitsgrad df_2.

Dabei bestimmen sich die Freiheitsgrade wieder durch die Stichprobengrößen:

$$df_1 = n_1 - 1$$
$$df_2 = n_2 - 1$$

In der Formelsammlung sind nur die Werte für Flächenanteile von 0,95 vermerkt. Die Werte für Flächenanteile von 0,05 (also am linken Rand) können durch Gleichung (6.4) bestimmt werden:

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

Dabei ist zu beachten, dass im Nenner die Reihenfolge der Freiheitsgrade getauscht wird!

Zur Verdeutlichung könnte – losgelöst von unserem Beispiel – ein unterer kritischer Wert berechnet werden durch:

$$\begin{aligned} F_{13;20;5\%} &= \frac{1}{F_{20;13;95\%}} \\ &\approx \frac{1}{2{,}46} \approx 0{,}41 \end{aligned}$$

Beispiel

Die Freiheitsgrade berechnen sich durch Gleichung (6.4):

$$df_1 = n_1 - 1 = 5$$
$$df_2 = n_2 - 1 = 5$$

Durch unsere gerichtete Alternativhypothese ergibt sich der kritische Wert aus Gleichung (6.4) (unter Anwendung des Tricks aus Gleichung (6.4)):

$$\begin{aligned} F &\leq F_{df_1;df_2;\alpha} \\ F &\leq F_{5;5;5\%} \\ F &\leq \frac{1}{F_{5;5;95\%}} \\ F &\leq \frac{1}{5{,}05} \\ F &\leq 0{,}20 \end{aligned}$$

Der so berechnete Ablehnungsbereich ist grafisch in Abbildung 6.2 aufbereitet.

Prüfgröße berechnen

Die Formel für die Prüfgröße F ist denkbar einfach:

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

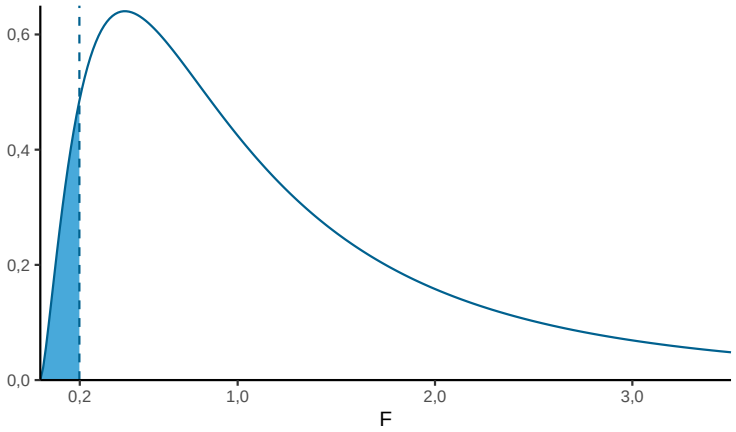


Abbildung 6.2: Ablehnungsbereich für $F \leq F_{5;5;5\%}$

Beispiel

Wir hatten die Varianzen der Stichproben berechnet mit:

$$s_1^2 \approx 0{,}41$$
$$s_2^2 \approx 1{,}59$$

Einsetzen in die Formel aus (6.4) ergibt:

$$F = \frac{s_1^2}{s_2^2}$$
$$= \frac{0{,}41}{1{,}59} \approx 0{,}26$$

Nullhypothese ablehnen oder beibehalten

Auch hier gilt dasselbe wie bei allen Tests.

Beispiel

Der kritische Wert von 0,20 müsste *unterschritten* werden, um die Nullhypothese abzulehnen. Das ist nicht passiert – wir „dürfen“ die Nullhypothese also beibehalten: Es gibt keinen statistisch signifikanten Unterschied in den beiden Varianzen ($\alpha = 0{,}05$).

Damit haben wir im vorherigen Beispiel die Voraussetzungen des 2-Stichproben-t-Tests also nicht verletzt.

Softwarehinweis

In R lautet der Befehl für den F-Test `var.test()`.

6.5 Fehlerarten

Bei statistischen Tests sind „Fehler“ nicht etwa Rechenfehler, sondern Angaben über die Wahrscheinlichkeit, die Nullhypothese aufgrund des Zufalls, dem die Stichprobe ja unterliegt, fälschlicherweise beizubehalten oder abzulehnen. Dabei wird unterschieden zwischen Fehlern 1. und 2. Art.

Fehler 1. Art

Der Fehler 1. Art (engl. *type I error*) steht für die Wahrscheinlichkeit, dass die Nullhypothese fälschlicherweise abgelehnt wird. Das passiert, wenn die Ergebnisse nur zufällig in den Ablehnungsbereich fallen. Konsequenz ist, dass eine Vermutung statistisch belegt wird, obwohl sie gar nicht stimmt. Die Wahrscheinlichkeit dafür ist also gleich dem Signifikanzniveau (α).

Fehler 2. Art

Der Fehler 2. Art (engl. *type II error*) ist die Wahrscheinlichkeit, dass die Nullhypothese fälschlicherweise beibehalten wird. Das passiert immer dann, wenn die Vermutung also eigentlich stimmt, die Stichprobenwerte aber zufällig so ausfallen, dass der Ablehnungsbereich nicht erreicht wird. Konsequenz ist, dass eine korrekte Vermutung statistisch nicht belegt werden kann. Die Warscheinlichkeit für einen Fehler 2. Art wird mit β gekennzeichnet.

Tipps zur Vertiefung

- YouTube-Kanal „Methodenlehre Mainz“: Inferenzstatistik (Playlist) 3.2–3.7
- YouTube-Kanal "Methodenlehre Mainz: Irren ist statistisch: Fehler 1. und 2. Art
- Kapitel 8 in Bortz und Schuster (2010)
- Kapitel 8.2.2; 8.2.4–8.2.6 in Lange und Nipper (2018)
- Kapitel 9.5.1, 10.1.3 und 10.3 in Klemm (2002)
- Kapitel 5.3.3 in Bahrenberg, Giese und Nipper (2010)
- *Englisch:* Kapitel 10 in Burt und Barber (1996)

Übungsaufgaben

Aufgabe 6-1

zur Lösung

Bestimmen Sie die folgenden kritischen Werte:

a) $F_{4;1;5\%}$
b) $F_{8;9;95\%}$
c) $F_{7;10;95\%}$
d) $F_{9;4;95\%}$
e) $F_{3;15;95\%}$
f) $F_{5;6;5\%}$
g) $F_{2;2;5\%}$
h) $F_{100;100;5\%}$
i) $F_{1;20;95\%}$
j) $F_{20;50;95\%}$

Aufgabe 6-2

zur Lösung

Sie wissen, dass in städtischen Freibädern die Wassertemperatur an verschiedenen Tagen normalverteilt ist. Sie vermuten, dass die Temperatur in zwei Bädern unterschiedlich stark variiert. Sie planen zwei unabhängige Erhebungen an zufälligen Tagen während der Badesaison. Aus organisatorischen Gründen beträgt die Stichprobengröße in „Schwimmbad 1“ $n_1 = 5$ und in „Schwimmbad 2“ $n_2 = 7$.

a) Welchen Test führen Sie durch?
b) Formulieren Sie die Hypothesen.
c) Sie wählen das Signifikanzniveau $\alpha = 0{,}1$. Was bedeutet diese Zahl?
d) Bestimmen Sie den Ablehnungsbereich.

Aufgabe 6-3

zur Lösung

(Fortführung von Aufgabe 2)

Sie erheben folgende Werte für die Wassertemperatur:

Schwimmbad 1 :	23,3	21,4	20,9	19,4	21,6		
Schwimmbad 2 :	21,5	21,7	21,5	21,4	22,0	20,9	21,8

a) Berechnen Sie die Prüfgröße.
b) Welche Schlüsse ziehen Sie aus der Untersuchung?

Aufgabe 6-4

zur Lösung

Sie interessieren sich für das Kommunikationsverhalten von Jugendlichen über WhatsApp. Sie vermuten, dass Nutzer*innen, die die „Gelesen-Benachrichtigung“ deaktiviert haben, im Durchschnitt langsamer antworten als diejenigen, die die Benachrichtigung aktiviert lassen.

Sie finden je Einstellung sechs freiwillige Schüler*innen, die Sie ihre WhatsApp-Protokolle auf die Durchschnittliche Antwortzeit auswerten lassen (natürlich unter Einwilligung der Eltern).

a) Welchen Test wollen Sie durchführen? Prüfen Sie die Voraussetzungen. Was könnte hier problematisch sein?
b) Formulieren Sie die Hypothesen.
c) Bestimmen Sie den Ablehnungsbereich bei Signifikanzniveau $\alpha = 0{,}05$.

Aufgabe 6-5

zur Lösung

Sie ermitteln die folgenden durchschnittlichen Antwortzeiten der individuellen Nutzer*innen (in Minuten):

Ohne Benachrichtigung :	24,7	32,0	48,9	23,7	23,0	10,0
Mit Benachrichtigung :	18,2	14,3	23,4	31,6	36,4	9,2

a) Berechnen Sie die Prüfgröße.

b) Welche Schlüsse ziehen Sie aus der Untersuchung?

Aufgabe 6-6

zur Lösung

In zwei Naturschutzgebieten werden zu zufälligen Zeitpunkten die Storchenpopulationen gezählt:

Naturschutzgebiet Esselrode :	17	15	16	22	17	21
Naturschutzgebiet Albwald :	23	17	13	20	19	19

Sie berechnen Mittelwerte, die sich sehr ähneln: 18,0 für Esselrode und 18,5 für Albwald. Sie haben jedoch die Vermutung, dass die Varianzen signifikant voneinander abweichen. Prüfen Sie die Vermutung mit $\alpha = 0{,}1$. (Es sei bekannt, dass die Populationen annähernd normalverteilt sind.)

Aufgabe 6-7

zur Lösung

Ein Musikstreaming-Portal zeichnet die Anzahl der Aufrufe einzelner Tracks nach Aufenthaltsort auf. Sie sind beauftragt nachvollziehen, ob sich die täglichen Aufrufzahlen eines angehenden Sommerhits in Hessen und in Niedersachsen – zwei Länder mit vergleichbaren Abonnent*innenzahlen – signifikant voneinander unterscheiden.

In der letzten Woche waren folgende Zahlen zu verzeichnen:

Hessen :	1172	1180	1307	1178	1156	1205	1212	1150	1114
Niedersachsen :	1337	1178	1230	1594	1658	1274	1094	1617	1056

Wählen Sie $\alpha = 0{,}01$.

Sitzung 7

Korrelation

Lernziele dieser Sitzung

Sie können…

- ein Streudiagramm interpretieren.
- die Kovarianz von zwei Variablen berechnen.
- den Korrelationskoeffizienten von zwei Variablen berechnen.

Lehrvideos (Sommersemester 2020)

- 7a) Kovarianz
- 7b) Korrelation

7.1 Bivariate Statistik

Grundlage der bivariaten Statistik ist es, dass für eine Reihe von Untersuchungseinheiten jeweils zwei Merkmale erfasst sind.

Diese Merkmale werden üblicherweise mit x und y gekennzeichnet. Für jedes i (laufende Nummer der Merkmalsträger*innen) gibt es dann ein x_i (Ausprägung des Merkmals x) und ein y_i (Ausprägung des Merkmals y).

Das Streudiagramm (engl. *scatter plot*) stellt alle erfassten Werte dar, indem es die Untersuchungseinheiten als Punkte arrangiert – und zwar anhand ihres jeweiligen Werts der Variable x entlang der x-Achse und entlang der y-Achse anhand des y-Werts (s. Abbildung 7.1).

Beispiel

Die statistischen Verfahren dieser Sitzung sollen wieder an einem Beispiel illustriert werden.

Wir fragen uns, ob der jährliche Ertrag in einem bestimmten Anbaugebiet für Klebreis in Nordostthailand mit dem jährlichen Niederschlag zusammenhängt. Die erfassten Werte sind in Tabelle 7.1 festgehalten („Rai“ ist ein in Thailand übliches Flächenmaß).

In einem Streudiagramm können diese Werte veranschaulicht werden. Dabei ist es üblich, die unabhängige Variable auf der x-Achse und die abhängige Variable auf der y-Achse einzutragen. Im Beispiel

Tabelle 7.1: Niederschlag und Ertrag im Reisanbau

Laufende Nr.	Jahr	Niederschlag (mm)	Ertrag (kg/Rai)
i		x_i	y_i
1	2008	1449	1860
2	2009	1472	2118
3	2010	1607	2225
4	2011	1494	2172
5	2012	1390	1816
6	2013	1764	2430
7	2014	1767	2580
8	2015	1765	2563
9	2016	1671	2276
10	2017	1838	2455

liegt nahe, dass der Ertrag vom Regen abhängt, und nicht etwa umgekehrt.

Abbildung 7.1 ist das Streudiagramm für unser Beispiel. Es fällt schon rein optisch auf, dass ein Zusammenhang zu bestehen scheint: Je mehr Regen, desto reicher die Ernte. Doch wie lässt sich dieser Zusammenhang beziffern?

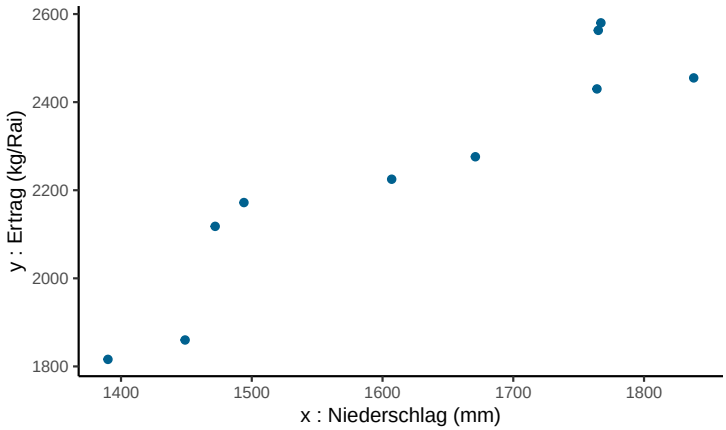


Abbildung 7.1: Streudiagramm zum Reisanbau

7.2 Kovarianz s_{xy}

Die Kovarianz (engl. *covariance*) s_{xy} gibt an, inwiefern die beiden Variablen x und y *gemeinsam variieren*. Die Kovarianz ergibt sich durch die Summe der jeweiligen Produkte der Differenzen zu den Mittelwerten $(x_i - \bar{x})$ und $(y_i - \bar{y})$, geteilt durch $(n - 1)$. Die Formel lautet also:

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

Gleichung (7.2) lässt erahnen: Wenn sowohl x als auch y in die gleiche Richtung vom jeweiligen Mittel-

Tabelle 7.2: Hilfstabelle für die Berechnung der Kovarianz

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	1449	1860	-172,7	-389,5	67266,65
2	1472	2118	-149,7	-131,5	19685,55
3	1607	2225	-14,7	-24,5	360,15
4	1494	2172	-127,7	-77,5	9896,75
5	1390	1816	-231,7	-433,5	100441,95
6	1764	2430	142,3	180,5	25685,15
7	1767	2580	145,3	330,5	48021,65
8	1765	2563	143,3	313,5	44924,55
9	1671	2276	49,3	26,5	1306,45
10	1838	2455	216,3	205,5	44449,65
Summe:	**16217**	**22495**			**362038,5**

wert abweichen (also beide Differenzen positiv oder beide Differenzen negativ), dann ist das Produkt positiv, sonst ist es negativ. Eine positive Kovarianz lässt also auf einen positiven Zusammenhang schließen (je größer x, desto größer auch y), eine negative Kovarianz auf einen negativen Zusammenhang (je größer x, desto *kleiner* y).

Softwarehinweis

Der Befehl `cov()` berechnet die Kovarianz einer bivariaten Verteilung in R.

Beispiel

Es macht Sinn, eine Tabelle anzulegen, in der Teilrechenschritte durchgeführt werden. Tabelle 7.2 veranschaulicht dies.

Als Zwischenschritt müssen die Mittelwerte $\bar{x}$ und $\bar{y}$ berechnet werden, wofür die Summen der ersten beiden Spalten herangezogen werden können:

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{16217}{10} = 1621{,}7 \\ \bar{y} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ &= \frac{22495}{10} = 2249{,}5\end{aligned}$$

Schließlich ergibt Einsetzen der Produktsumme in Gleichung (7.2) die Kovarianz:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$
$$\approx \frac{362038{,}5}{9} = 40226{,}5$$

Die Kovarianz ist also $s_{xy} = 40226{,}5$. Was sagt uns diese Zahl? Zunächst ist sie positiv, womit wir von einer positiven Korrelation (je mehr Regen, desto mehr Ertrag) ausgehen können. Sie ist auch „irgendwie“ ziemlich groß, was einen deutlichen Zusammenhang nahelegt. Aber die Kovarianz ist abhängig vom Maßstab – wäre der Ertrag nicht in Kilogramm pro Rai, sondern (wie in Deutschland üblich) in Dezitonnen pro Hektar angegeben, dann wäre die Zahl deutlich kleiner (2514,156 um genau zu sein). Wie lässt sich die Stärke der Korrelation also unabhängig von den Maßeinheiten angeben?

7.3 Korrelationskoeffizient r

Der Korrelationskoeffizient r (auch Produkt-Moment-Korrelation, Bravais-Pearson-Korrelation, Pearsons r, engl. *correlation coefficient*) standardisiert die Kovarianz s_{xy} anhand der Standardabweichungen s_x und s_y. Die Formel lautet:

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

Durch diese Standardisierung kann der Korrelationskoeffizient nur noch Werte zwischen $r = -1$ (perfekte negative Korrelation) und $r = 1$ (perfekte positive Korrelation) annehmen. Ein Korrelationskoeffizient nahe $r = 0$ bedeutet, dass es keinen Zusammenhang zwischen den Variablen x und y gibt (s. Abbildung 7.2).

Softwarehinweis

In R kann der Korrelationskoeffizient von zwei Merkmalen mit dem Befehl `cor()` bestimmt werden.

Beispiel

In der Formel für den Korrelationskoeffizienten r (7.3) werden die Standardabweichungen s_x und s_y benötigt. Es ist daher sinnvoll, die Hilfstabelle um die Quadrate der Differenzen (und deren Summen) zu erweitern (s. Tabelle 7.3).

Die Standardabweichungen ergeben sich nun wie gewohnt aus:

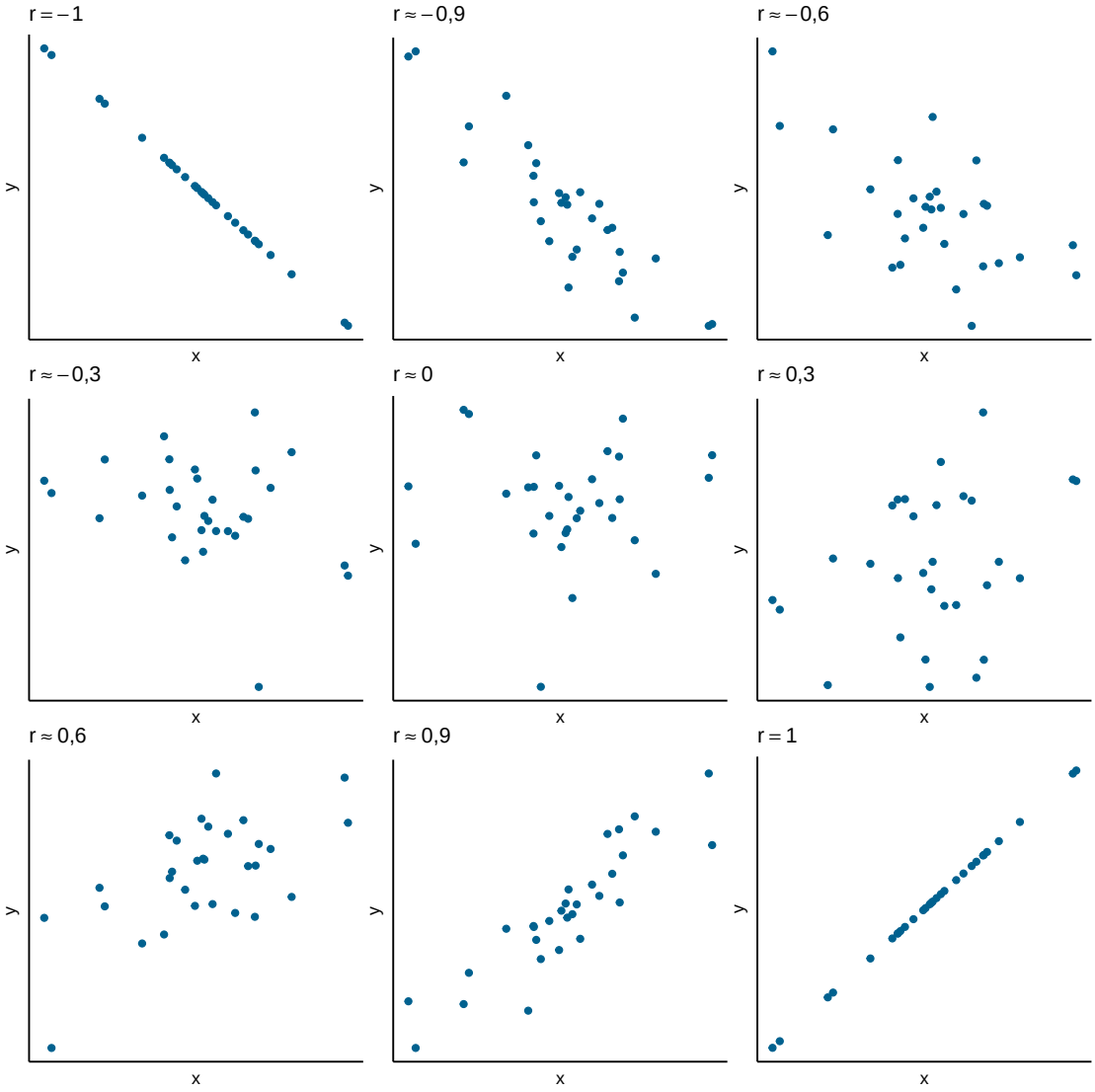


Abbildung 7.2: Verschiedene Korrelationskoeffizienten

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

$$= \sqrt{\frac{233556{,}1}{9}} = \sqrt{25950{,}68} \approx 161{,}09$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$= \sqrt{\frac{646556{,}5}{9}} = \sqrt{71839{,}61} \approx 268{,}03$$

Tabelle 7.3: Hilfstabelle für die Berechnung des Korrelationskoeffizienten

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	1449	1860	-172,7	-389,5	29825,29	151710,25
2	1472	2118	-149,7	-131,5	22410,09	17292,25
3	1607	2225	-14,7	-24,5	216,09	600,25
4	1494	2172	-127,7	-77,5	16307,29	6006,25
5	1390	1816	-231,7	-433,5	53684,89	187922,25
6	1764	2430	142,3	180,5	20249,29	32580,25
7	1767	2580	145,3	330,5	21112,09	109230,25
8	1765	2563	143,3	313,5	20534,89	98282,25
9	1671	2276	49,3	26,5	2430,49	702,25
10	1838	2455	216,3	205,5	46785,69	42230,25
Summe:	**16217**	**22495**			**233556,1**	**646556,5**

Nun lassen sich die errechneten Werte in Gleichung (7.3) einsetzen:

$$\begin{aligned} r &= \frac{s_{xy}}{s_x \cdot s_y} \\ &\approx \frac{40226{,}5}{161{,}09 \cdot 268{,}03} \approx 0{,}93 \end{aligned}$$

Wir können bei einem Korrelationskoeffizienten $r \approx 0{,}93$ von einem deutlichen positiven Zusammenhang zwischen Niederschlag und Ertrag ausgehen.

Tipps zur Vertiefung

- YouTube-Kanal „Kurzes Tutorium Statistik“: Streudiagramm und Korrelation
- YouTube-Kanal „Methodenlehre Mainz“: Bivariate Daten (Playlist)
- Kapitel 10 in Bortz und Schuster (2010)
- Kapitel 4.5.1 – 4.5.6 in Lange und Nipper (2018)
- Kapitel 6.1, 6.3 und 6.4 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 16 in Klemm (2002)
- *Englisch:* Kapitel 13.1 – 13.4 in Burt und Barber (1996)

Übungsaufgaben

Aufgabe 7-1

zur Lösung

Zeichnen Sie ein Streudiagramm und berechnen Sie die Kovarianz sowie den Korrelationskoeffizenten für die folgenden Messreihen.

a) Messreihe:

x_i	y_i
14,21	134
10,32	131
13,82	134
15,79	135
14,70	134
17,23	137
14,84	136
14,96	135

b) Messreihe:

x_i	y_i
-1,17	14,40
-0,10	2,31
-0,15	2,95
0,46	-1,39
0,34	-2,96
-0,44	2,44
2,13	-20,47
0,66	-10,51
-1,37	11,81
0,56	-4,05

Aufgabe 7-2

zur Lösung

Sie erheben für zufällige Wasserhäuschen in Frankfurt die Entfernung zur nächsten Haltestelle der S- oder U-Bahn sowie den durchschnittlichen Tagesumsatz. Die Erhebung ergibt:

Entfernung (m)	**Umsatz (€/Tag)**
35	394,61
79	468,92
234	385,75
105	376,17
318	283,26
31	342,77

Gibt es einen Zusammenhang zwischen Entfernung und Umsatz? Wenn ja: Wie hängen die Variablen zusammen? Wie stark ist der Zusammenhang?

Aufgabe 7-3

zur Lösung

Für eine Umnutzung zu Kulturzentren sollen zwei Gebäude kernsaniert werden. Um die Kosten schätzen zu können, werden die Nutzflächen und Kosten von fünf ähnlichen Sanierungsprojekten herangezogen. Sie berechnen zunächst die Mittelwerte und Varianzen der erfassten Merkmale.

Projekt	Nutzfläche (m²)	Kosten (Tsd. €)
1	456	264
2	628	306
3	497	348
4	275	202
5	549	322
6	313	99

Wie deutlich fällt der Zusammenhang zwischen Fläche und Kosten aus?

Aufgabe 7-4

zur Lösung

Eine Stadtverwaltung möchte die Mietpreisentwicklung für Gewerbeimmobilien in der innerstädtischen Einkaufspassage abschätzen. Sie folgt dabei der These: Entscheidend für die Höhe der monatlichen Mietpreise (in Euro pro Quadratmeter) sei die Entfernung zur nächstgelegenen Haltestelle des ÖPNV: Je näher an der Haltestation gelegen, desto höher der Mietpreis.

Für Aussagen über den angenommenen Zusammenhang stehen die Daten von sechs zufällig ausgewählten Gewerbeimmobilien in der Einkaufspassage zur Verfügung.

Immobilie	Entfernung (m)	Quadratmeterpreis (€)
1	1141	30
2	850	49
3	862	40
4	1000	39
5	783	51
6	890	42

Die (gerundeten) arithmetischen Mittel betragen $\bar{x} = 921{,}00$ Meter und $\bar{y} \approx 41{,}83$ Euro, und die (gerundeten) Standardabweichungen liegen bei $s_x = 128{,}97$ Meter und $s_y = 7{,}57$ Euro.

Wie groß ist der Zusammenhang zwischen der Entfernung zur nächstgelegenen Haltestelle und dem gemessenen Mietpreis pro Quadratmeter? Berechnen Sie den angemessenen Korrelationskoeffizienten und interpretieren Sie das Ergebnis.

Aufgabe 7-5

zur Lösung

(weiterführend, nicht klausurrelevant)

a) Zeigen Sie, dass der Korrelationskoeffizent r ein standardisierter Wert ist, indem Sie ihn in z-Werten ausdrücken.

b) Überprüfen Sie die Formel anhand Aufgabe 1 a).
c) Angenommen, Sie wollen r angeben, ohne die Koviarianz berechnet zu haben. Wie lassen sich die Rechenschritte dann vereinfachen?
d) Überprüfen Sie den Rechenweg anhand Aufgabe 2.

Sitzung 8

Lineare Regression

Lernziele dieser Sitzung

Sie können…

- eine Regressionsgerade berechnen.
- Werte aus der Regressionsgerade ableiten.
- Residuen errechnen.
- den Determinationskoeffizienten R^2 berechnen und interpretieren.

Lehrvideos (Sommersemester 2020)

- 8a) Regressionsgerade
- 8b) Residuen und Determinationskoeffizient
  - Beim Teil „Klausur-Update“ gilt der **Ablauf** und die **Struktur** der Klausur auch dieses Semester.
  - Die administrative Anmeldung für die Theorieklausur ist auf OLAT bis … möglich.
  - Probeklausuren sind hier zu finden.
  - Zur formalen Anmeldung und zur Versuchsregelung kann ich dieses Jahr keine Angaben machen. (Fragen Sie im Zweifel Ihr Prüfungsamt!)

8.1 Regresssionsanalyse

> Sind zwei stochastisch abhängige Variablen x und y durch eine Regressionsgleichung miteinander verknüpft, kann die eine Variable zur Vorhersage der anderen eingesetzt werden. (Bortz und Schuster 2010: 183)

Es gibt viele Möglichkeiten, Regressionen zu modellieren. Im Rahmen dieser Veranstaltung wird nur die lineare Regression (engl. *linear regression*) behandelt. Lineare Regressionsmodelle werden immer durch eine lineare Gleichung des Formats

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

ausgedrückt, wobei a der Achsenabschnitt ist und b die Steigung. Ist die Gleichung bekannt, so können wir für jeden Wert x einen entsprechenden Wert y „vorhersagen“.

Abbildung 8.1 zeigt ein solches lineares Regressionsmodell als Gerade durch ein Streudiagramm.

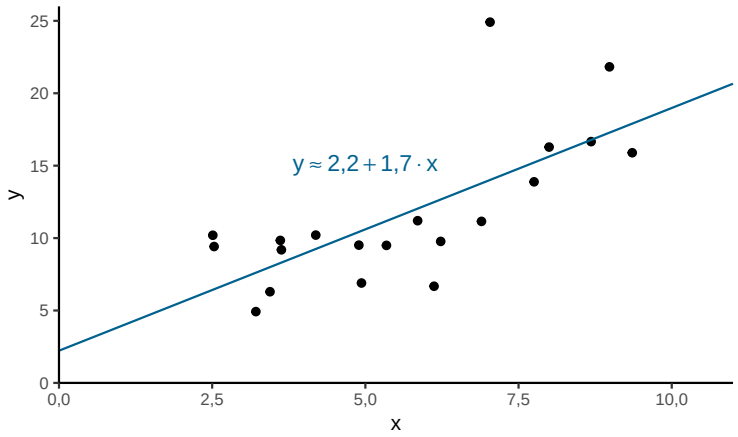


Abbildung 8.1: Regressionslinie durch ein Streudiagramm

Der Achsenabschnitt $a \approx 2{,}2$ bedeutet, dass die Regressionsgerade die y-Achse etwa auf der Höhe 2,2 schneidet (bei $x = 0$). Die Steigung $b \approx 1{,}7$ heißt, dass für jede zusätzliche Einheit der Variable x ca. 1,7 zusätzliche Einheiten der Variable y erwartet werden können.

Wenn die Regressionsgleichung bekannt ist, kann für jedes gültige (grundsätzlich: jedes beliebige) x ein erwarteter Wert $\hat{y}$ berechnet werden. So könnte uns bei der Beispielregression interessieren, welchen Wert $\hat{y}_i$ im Modell annimmt, wenn $x_i = 20$ beträgt:

$$\begin{aligned}\hat{y}_i &= a + b \cdot x_i \\ &\approx 2{,}2 + 1{,}7 \cdot 20 \\ &= 36{,}2\end{aligned}$$

Bei solchen Schätzungen *außerhalb* des bekannten Wertebereichs spricht man auch vom „Extrapolieren“, sonst – für fehlende Werte innerhalb des bekannten Wertebereich – vom „Interpolieren“.

Umgekehrt könnte die Frage lauten: Wie groß muss ein x_i sein, damit (im Modell) $\hat{y}_i = 12$ beträgt? Dies lässt sich durch eine einfache Umformung der Gleichung (8.1) berechnen:

$$\begin{aligned}\hat{y}_i &= a + b \cdot x_i \\ x_i &= \frac{\hat{y}_i - a}{b} \\ &= \frac{12 - 2{,}2}{1{,}7} \\ &\approx 5{,}8\end{aligned}$$

Bei der Regressionsanalyse wird ein gerichtetes Abhängigkeitsverhältnis der Variablen impliziert: y hängt hier von x ab. Daher wird x auch die „Prädiktorvariable“ und y die „Kriteriumsvariable“ genannt.

Softwarehinweis

Wenn in R ein lineares Modell (eine Regressionsgerade) vorliegt, können Werte mit `predict()` geschätzt werden.

Es ist also für derartige Fragestellungen nötig, die Gleichung der Regressionsgeraden zu kennen. Im Folgenden wird gezeigt, wie diese anhand einer bivariaten Verteilung bestimmt werden kann.

8.2 Bestimmung der Regressionsgeraden

Der Koeffizient b (also die Steigung der Regressionsgeraden) lässt sich berechnen, indem man die Kovarianz s_{xy} durch die Varianz von x dividiert:

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

Der Koeffizient a (also der Achsenabschnitt) ergibt sich wiederum aus b und den Mittelwerten $\bar{x}$ und $\bar{y}$:

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

Softwarehinweis

In R lässt sich ein lineares Regressionsmodell mit dem Befehl `lm()` erstellen.

Die Bestimmung der Regressionsgeraden soll nun mit einem Beispiel illustriert werden.

Beispiel

Wir fragen uns, wie die Aufenthaltszeit von Passagieren am Frankfurter Flughafen mit dem Betrag zusammenhängt, den sie in den dortigen Geschäften ausgeben. Eine Zufallserhebung habe die Werte in Tabelle 8.1 ergeben.

Mit den Methoden aus Sitzung 2 und 7 können wir folgende Werte für die Mittelwerte $\bar{x}$ und $\bar{y}$, die Varianz s_x^2 sowie die Kovarianz s_{xy} berechnen:

$$\begin{aligned} \bar{x} &= 256{,}25 \\ \bar{y} &\approx 32{,}56 \\ s_x^2 &\approx 9340{,}93 \\ s_{xy} &\approx 1062{,}50 \end{aligned}$$

Für die Steigung der Regressionsgeraden b setzen wir die entsprechenden Werte in Gleichung (8.2) ein:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ &\approx \frac{1062{,}50}{9340{,}93} \\ &\approx 0{,}114 \end{aligned}$$

Tabelle 8.1: Messwerte am Frankfurter Flughafen

Aufenthaltszeit (min)	Ausgaben (€)
x_i	y_i
121	17,94
125	23,15
293	44,31
370	42,46
246	35,51
281	28,46
169	18,47
328	56,77
388	40,11
131	12,64
299	24,54
324	46,37

Die Steigung von 0,114 bedeutet, dass – im linearen Regressionsmodell – Passagiere in jeder zusätzlichen Minute, die sie am Flughafen verbringen, in etwa 11,4 zusätzliche Cent ausgeben.

Der Achsenabschnitt a berechnet sich dann gemäß Gleichung (8.2):

$$\begin{aligned} a &= \bar{y} - b \cdot \bar{x} \\ &\approx 32{,}56 - 0{,}114 \cdot 256{,}25 \\ &\approx 3{,}35 \end{aligned}$$

Dieser Wert ergibt nur einen abstrakt-mathematischen Sinn – es dürfte in der Praxis wohl kaum Passagiere geben, die 0 Minuten am Flughafen verbringen und € 3,35 ausgeben.

Mit dem Achsenabschnitt a und der Steigung b lässt sich folgende Gleichung für die Regressionsgerade aufstellen (s. Gleichung (8.1)):

$$\begin{aligned} y &= a + b \cdot x \\ y &\approx 3{,}35 + 0{,}114 \cdot x \end{aligned}$$

Graphisch ist diese lineare Regression in Abbildung 8.2 dargestellt.

8.3 Residuen

Residuen (engl. *residuals*) werden mit e bezeichnet und sind die Differenzen zwischen den tatsächlichen y-Werten und den im Modell erwarteten $\hat{y}$-Werten für die jeweiligen x-Werte:

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

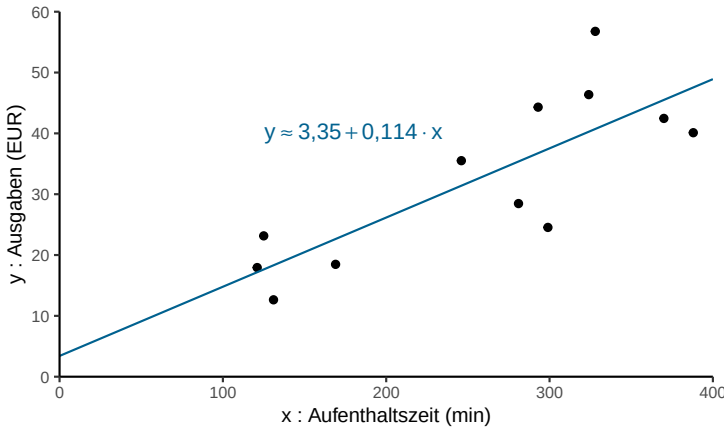


Abbildung 8.2: Regressionslinie durch ein Streudiagramm

Residuen sind also – auch dem Wortstamm nach – das, was nach der Vorhersage durch das Modell „übrig bleibt“ von den tatsächlich beobachteten Werten (also der Teil des Werts, der *nicht* durch das Regressionsmodell erklärt wird).

Softwarehinweis

Residuen lassen sich in R durch den Befehl `resid()` errechnen.

Beispiel

Graphisch sind die Residuen für unser Beispiel in Abbildung 8.3 dargestellt (positive Werte in grün, negative Werte in rot), tabellarisch in Tabelle 8.2.

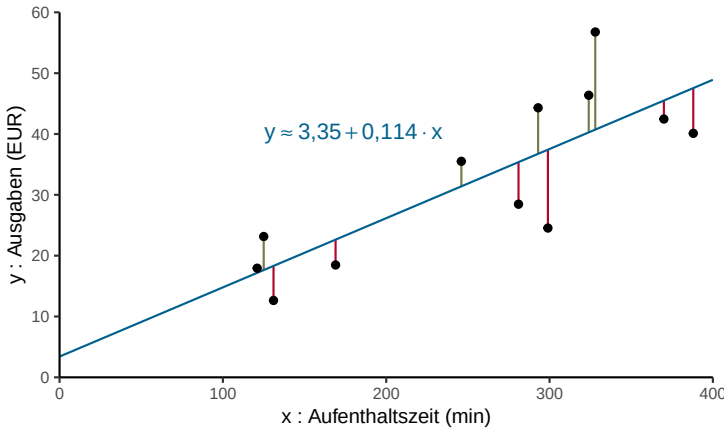


Abbildung 8.3: Graphische Darstellung der Residuen

Residuen spielen in vielen statistischen Verfahren eine Rolle, z.B. in der Residuenanalyse. Diese Verfahren werden im Rahmen dieser Veranstaltung jedoch nicht behandelt.

Tabelle 8.2: Residuen der Beispielwerte

Aufenthaltszeit (min)	Ausgaben (€)	Erwartete Ausgaben (€)	Residuen (€)
x_i	y_i	$\hat{y}_i \approx 3{,}35 + 0{,}114 \cdot x_i$	$e_i = y_i - \hat{y}_i$
121	17,94	17,144	0,796
125	23,15	17,600	5,550
293	44,31	36,752	7,558
370	42,46	45,530	-3,070
246	35,51	31,394	4,116
281	28,46	35,384	-6,924
169	18,47	22,616	-4,146
328	56,77	40,742	16,028
388	40,11	47,582	-7,472
131	12,64	18,284	-5,644
299	24,54	37,436	-12,896
324	46,37	40,286	6,084

8.4 Determinationskoeffizient R^2

Der Determinationskoeffizient R^2 (engl. *coefficient of determination*) ist formal definiert als das Verhältnis der Varianz der vorhergesagten $\hat{y}$-Werte zur Varianz der tatsächlich beobachteten y-Werte (wobei sich der Term $[n-1]$ auskürzt):

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Da Zähler und Nenner als Quadratsummen stets positiv sind und die Varianz der $\hat{y}$-Werte immer *kleiner oder gleich* der Varianz der y-Werte ist, nimmt der Determinationskoeffizient immer einen Wert zwischen 0 und 1 an.

Je größer R^2, desto besser erklärt das lineare Regressionsmodell die tatsächlich beobachteten Werte. $R^2 = 1$ bedeutet, dass das Modell die Werte perfekt erklärt.

Für lineare Regressionsmodelle (also für die einzige Regression, die im Rahmen dieser Veranstaltung behandelt wird) lässt sich R^2 auch berechnen, indem wir den Korrelationskoeffizienten r quadrieren:

$$R^2 = r^2$$

Softwarehinweis

In R wird mit dem Befehl `summary()` unter anderem der Determinationskoeffizient eines linearen Regressionsmodells ausgegeben.

Beispiel

Mit den Methoden aus Sitzung 7 können wir den Korrelationskoeffizienten für unser Beispiel errechnen:

$$\begin{aligned} r &= \frac{s_{xy}}{s_x \cdot s_y} \\ &\approx \frac{1062{,}50}{96{,}65 \cdot 13{,}68} \\ &\approx 0{,}804 \end{aligned}$$

Der Determinationskoeffizient ergibt sich dann mit Gleichung (8.4):

$$\begin{aligned} R^2 &= r^2 \\ &\approx 0{,}804^2 \\ &\approx 0{,}646 \end{aligned}$$

Tipps zur Vertiefung

- Kapitel 11 in Bortz und Schuster (2010)
- Kapitel 4.5.1 – 4.5.6 in Lange und Nipper (2018)
- Kapitel 6.2 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 17 in Klemm (2002)
- *Englisch:* Kapitel 13.1 – 13.4 in Burt und Barber (1996)

Übungsaufgaben

Aufgabe 8-1

zur Lösung

Sie haben für eine bivariate Verteilung die folgende Regressionsgleichung bestimmt:

$$y = -1{,}48 - 0{,}975 \cdot x$$

a) Bestimmen Sie die erwarteten $\hat{y}_i$-Werte für diese x_i-Werte:

$$0{,}3 \quad -18{,}5 \quad -13{,}5 \quad -17{,}2 \quad 29{,}8 \quad 25{,}6 \quad -36{,}4 \quad -26{,}2$$

b) Für welche Werte x_i sagt das Regressionsmodell diese Werte $\hat{y}_i$ voraus?

$$-10 \quad 15 \quad -50 \quad -10 \quad -60 \quad -55 \quad -20 \quad 0$$

c) Bestimmen Sie die Residuen für die tatsächlich beobachtete Messreihe:

x_i	y_i
-11,49	6,82
8,22	-8,59
-25,66	25,92
23,81	-26,91
-3,14	4,41
-1,52	-3,39
20,15	-19,89
-10,22	9,30

Aufgabe 8-2

zur Lösung

Eine bivariate Verteilung sei gekennzeichnet durch die folgenden Parameter:

$$\bar{x} = 157{,}5$$
$$\bar{y} = 156{,}7$$
$$s_x^2 = 1080{,}94$$
$$s_y^2 = 884{,}46$$
$$s_{xy} = 869{,}83$$

a) Bestimmen Sie die Regressionsgleichung im linearen Modell.

b) Bestimmen Sie den Determinationskoeffizienten R^2.

Aufgabe 8-3

zur Lösung

Für die Messreihe der Aufgabe 3 aus Sitzung 7 sei gefragt:

a) Welche Gleichung beschreibt ein geeignetes lineares Regressionsmodell?

b) Wenn die Nutzfläche für Objekt A 318 m² und für Objekt B 380 m² beträgt, wie hoch können dann jeweils die Kosten für die Sanierung geschätzt werden?

Aufgabe 8-4

zur Lösung

Für die Messreihe der Aufgabe 4 aus Sitzung 7 sei gefragt:

a) Die Stadtverwaltung hat unter Rückgriff auf diese Daten ein einfaches lineares Modell entwickelt, das eine Prognose der Mietpreise der Gewerbeimmobilien in Abhängigkeit von ihrer Entfernung zur nächstgelegenen Haltestelle des ÖPNV erlaubt. Wie lautet die Regressionsgleichung?

b) Wie hoch fällt laut Modell der Mietpreis pro Quadratmeter für eine 500 Meter von der nächstgelegenen ÖPNV-Haltestelle entfernte Gewerbeimmobilie aus?

Aufgabe 8-5

zur Lösung

Sie fragen sich, wie die erreichte Punktzahl in einer Klausur mit der Vorbereitungszeit der geprüften Studierenden zusammenhängt. Sie erheben die folgende Messreihe:

Vorbereitungszeit (min)	Erreichte Punktzahl
834	88
17	41
519	75
253	39
739	77
844	100

a) Welche Punktzahl ist mit einer Vorbereitungszeit von sechs Stunden zu erwarten?

b) Ab welcher Vorbereitungszeit ist im Modell zu erwarten, dass ein*e Studierende die Klausur besteht ($\geq$ 50 Punkte)?

c) Ab welcher Vorbereitungszeit kann laut Modell mit der vollen Punktzahl (100 Punkte) gerechnet werden?

d) Wie gut erklärt ein lineares Modell die Prüfungsleistungen anhand der Vorbereitungszeit?

e) Welche Limitationen hat das Modell? Denken Sie an extreme Werte.

Sitzung 9

Kreuztabellen

Lernziele dieser Sitzung

Sie können…

- eine Kreuztabelle erstellen und interpretieren.
- den Kontingenzkoeffizienten χ^2 errechnen.
- die Maßzahlen ϕ bzw. CI errechnen und interpretieren.

Lehrvideos (Sommersemester 2020)

- 9a) Kreuztabellen
- 9b) Kontingenzkoeffizenten

9.1 Bivariate Verteilungen mit nominalen Variablen

In der bivariaten Statistik (Sitzung 7 und Sitzung 8) ging es bisher um Zusammenhänge zwischen zwei metrischen Variablen. In dieser Sitzung geht es um statistische Verfahren der bivariaten Statistik, bei denen für beide Variablen nur das Nominalskalenniveau vorausgesetzt ist. (Für Skalenniveaus s. Sitzung 1.)

Mit den Werten von nominalskalierten Variablen lassen sich die in Sitzung 7 und Sitzung 8 besprochenen Parameter (z. B. Kovarianz) nicht errechnen, weil wir mit ihnen nicht die notwendigen Rechenoperationen (Addition, Subtraktion) durchführen können. Stattdessen sind die beobachteten Häufigkeiten Ausgangslage für die im Folgenden besprochenen Verfahren.

Beispiel

Wir fragen uns, ob es einen Zusammenhang zwischen dem Studienfach von Studierenden an einer Universität und ihrem präferierten Transportmittel für den Pendelweg zum Campus gibt. Insbesondere interessiert uns, ob ein Zusammenhang zwischen dem Studium der Geistes- und Sozialwissenschaften und der Fahrradnutzung besteht.

Beide Variablen sind nur nominalskaliert: Die erhobenen Werte können in Kategorien eingeordnet werden, die aber keine inhärente Hierarchie aufweisen (Studienfach: Geographie, Politikwissenschaft, BWL, …; Transportmittel: Bus, Fahrrad, zu Fuß, …).

Tabelle 9.1: Ungeordnete Rohdaten der Erhebung

i	**Studienfach**	**Transportmittel**
1	Geistes-/Sozialwissenschaft	anderes Transportmittel
2	anderes Studienfach	anderes Transportmittel
3	anderes Studienfach	anderes Transportmittel
4	anderes Studienfach	Fahrrad
5	Geistes-/Sozialwissenschaft	anderes Transportmittel
6	anderes Studienfach	anderes Transportmittel
...	...	...
85	Geistes-/Sozialwissenschaft	anderes Transportmittel
86	anderes Studienfach	anderes Transportmittel
87	anderes Studienfach	Fahrrad
88	anderes Studienfach	anderes Transportmittel
89	Geistes-/Sozialwissenschaft	anderes Transportmittel
90	Geistes-/Sozialwissenschaft	anderes Transportmittel

Tabelle 9.2: Kreuztabelle für die Beispielerhebung

	Fahrrad	**anderes Transportmittel**	
Geistes-/Sozialwissenschaft	11	28	**39**
anderes Studienfach	9	42	**51**
	20	**70**	**90**

Um die Variablen im Sinne unserer Fragestellung zu vereinfachen, wandeln wir beide Variablen in *dichotome* Variablen um (die dann nur zwei Werte annehmen können). Wir beschränken uns auf die Erhebung von „Fahrrad“ oder „anderes Transportmittel“ einerseits und „Geistes-/Sozialwissenschaft“ oder „anderes Studienfach“ andererseits. Die (verkürzte) Tabelle der Rohdaten einer Zufallsstichprobe der Größe $n = 90$ könnte dann so aussehen wie 9.1.

9.2 Kreuztabelle

Die Kreuztabelle (auch Kontingenztabelle, Kontingenztafel, engl. *contingency table*) ist eine übersichtliche Zusammenfassung der Rohdaten. Sie spannt die beiden Variablen in Spalten- und Zeilenrichtung auf, so dass in jeder Zelle die Häufigkeit einer bestimmten Wertekombination steht.

Bei zwei dichotomen Variablen ergeben sich zwei Spalten und zwei Zeilen, also vier Tabellenfelder. Wir sprechen in diesem Fall auch von einer 2×2-Tabelle.

Beispiel

Die Kreuztabelle für unser Beispiel ist in 9.2 dargestellt. Die Spaltenüberschriften sind die beiden Werte der dichotomen Variable „Transportmittel“, und die Zeilennamen sind die beiden Werte für „Studienfach“. In den Zellen stehen die Häufigkeiten. Es lässt sich also z. B. ablesen, dass die Kombination „Fahrrad“ und „anderes Studienfach“ neun mal vorkommt.

Tabelle 9.3: Allgemeine Bezeichnungen in der Kreuztabelle

	Spalte 1	**Spalte 2**	**...**	**Spalte ℓ**	
Zeile 1	n_{11}	n_{12}	...	$n_{1\ell}$	$n_{1\cdot}$
Zeile 2	n_{21}	n_{22}	...	$n_{2\ell}$	$n_{2\cdot}$
...	...	...	...	...	**...**
Zeile k	n_{k1}	n_{k2}	...	$n_{k\ell}$	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	**...**	$n_{\cdot\ell}$	n

Am rechten Rand der Tabelle stehen die Summen für die Zeilen, am unteren Rand die Summen der Spalten. Ganz unten rechts steht die Gesamtsumme (Größe der Stichprobe).

Softwarehinweis

In R kann eine einfache Kreuztabelle mit dem Befehl `table()` ausgegeben werden.

Verallgemeinerung

In Tabelle 9.3 ist das allgemeingültige Format für Kreuztabellen festgehalten. Dabei sind folgende Besonderheiten zu beachten:

- Das Symbol k steht für die Anzahl der Zeilen, ℓ für die Anzahl der Spalten.
- Die Häufigkeiten für Merkmalskombinationen in den Tabellenfeldern werden durch n_{ij} symbolisiert, wobei i für die laufende Nummer der Zeile steht, und j für die laufende Nummer der Spalte.
- Die Teilsummen an den Rändern werden mit Punktnotation bezeichnet. Dabei steht die Zeilensumme $n_{i\cdot}$ für die Summe *aller* Felder in Zeile i (Zeilensumme) und $n_{\cdot j}$ für die Summe *aller* Felder in Spalte j (Spaltensumme).
- Die Gesamtsumme unten rechts wird hier mit n gekennzeichnet und steht wie gewohnt für die Gesamtgröße der Stichprobe.

9.3 Erwartungswerte

Bestünde *kein* Zusammenhang zwischen den Variablen, dann wäre zu erwarten, dass sich die Kombinationen gleichmäßig auf die Tabellenfelder aufteilen, und zwar ausgehend von den Teilsummen für die Zeilen und Spalten.

Der Erwartungswert für ein Tabellenfeld (also der „durchschnittliche“ Wert, wenn es keinen Zusammenhang zwischen den beiden Variablen gibt) berechnet sich durch die Formel:

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

Es wird also das Produkt der Zeilen- und der Spaltensumme geteilt durch die Gesamtsumme.

Tabelle 9.4: Kreuztabelle der Beispieldaten mit Erwartungswerten

	Fahrrad	**anderes Transportmittel**	
Geistes-/Sozialwissenschaft	11	28	**39**
	(8,67)	(30,33)	
anderes Studienfach	9	42	**51**
	(11,33)	(39,67)	
	20	**70**	**90**

Beispiel

Die beobachtete Häufigkeit für die Kombination „Geistes-/Sozialwissenschaft“ (Zeile 1) und „anderes Transportmittel“ (Spalte 2) ist 28. Aber was wäre der Erwartungswert bei den gegebenen Summen? Wir setzen einfach die entsprechenden Werte in Gleichung (9.3) ein:

$$\begin{aligned} m_{12} &= \frac{n_{1\cdot} \cdot n_{\cdot 2}}{n} \\ &= \frac{39 \cdot 70}{90} \\ &\approx 30{,}33 \end{aligned}$$

Diese Rechnung lässt sich für alle Tabellenfelder durchführen. Die Kreuztabelle kann dann um diese erwarteten Werte in Klammern ergänzt werden (s. Tabelle 9.4).

9.4 Kontingenzkoeffizent χ^2

Sind für alle Tabellenfelder die Beobachtungs- und Erwartungswerte gegeben, lässt sich für jedes Tabellenfeld ein Wert berechnen, der diese Werte in Relation setzt. Die Summe dieser Werte über die gesamte Tabelle hinweg wird Kontingenzkoeffizient genannt und mit χ^2 („Chi-Quadrat“) abgekürzt.

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

Bei der Formel steht k wieder für die Anzahl der Zeilen (und i für ihre laufende Nummer) und ℓ für die Anzahl der Spalten (und j für ihre laufende Nummer).

Das doppelte Summenzeichen mag etwas verwirrend sein, bedeutet aber nur, dass die Zeilen spaltenweise summiert werden, und dann die Summe dieser Zeilensumme genommen wird – d.h. dass einfach alle Tabellenfelder aufsummiert werden.

Der χ^2-Wert kann (ähnlich wie der F-Wert aus Sitzung 6) nur positive Werte annehmen. Er bildet die Grundlage für die im Folgenden besprochenen Kennwerte ϕ und CI sowie für den in Sitzung 10 zu besprechenden χ^2-Test.

Tabelle 9.5: Kreuztabelle der Beispieldaten mit Teilwerten für χ^2

	Fahrrad	**anderes Transportmittel**	
Geistes-/Sozialwissenschaft	11	28	**39**
	(8,67)	(30,33)	
	0,626	0,179	
anderes Studienfach	9	42	**51**
	(11,33)	(39,67)	
	0,479	0,137	
	20	**70**	**90**

Beispiel

Ein möglicher Zwischenschritt ist es, diese Teilwerte von χ^2 für die einzelnen Tabellenfelder auszurechnen und in der Kreuztabelle zu notieren. Die Teilwerte werden dann für jedes Tabellenfeld mit der Formel

$$\frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

berechnet und sind in Tabelle 9.5 in blau dargestellt.

Zum Beispiel ergibt sich der Teilwert für χ^2 für die Kombination „anderes Studienfach“ – „Fahrrad“ durch Einsetzen in Gleichung (9.4):

$$\begin{aligned}\frac{(n_{21} - m_{21})^2}{m_{21}} &\approx \frac{(9 - 11{,}33)^2}{11{,}33} \\ &= \frac{-2{,}33^2}{11{,}33} \\ &\approx \frac{5{,}43}{11{,}33} \\ &\approx 0{,}479\end{aligned}$$

Der χ^2-Wert lässt sich nun bestimmen, indem diese Teilwerte aufsummiert werden:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij} - m_{ij})^2}{m_{ij}} \\ &\approx 0{,}626 + 0{,}179 + 0{,}479 + 0{,}137 \\ &= 1{,}421\end{aligned}$$

Mit diesem Wert $\chi^2 \approx 1{,}421$ können wir noch nicht so viel anfangen – wir wissen aber, dass er ein Maß dafür ist, wie sehr unsere beobachtete Verteilung von einer zu erwarteten Verteilung (vorausgesetzt, es gibt keinen Zusammenhang) abweicht.

9.5 ϕ-Koeffizient

Der ϕ-Koeffizient ist der Korrelationskoeffizient für zwei dichotome Variablen (wobei er in der hier besprochenen Version nur positive Werte annehmen kann). Er ist jedoch *nicht* ohne weiteres mit dem Korrelationskoeffizienten r (aus Sitzung 7) vergleichbar.

Der Wert für ϕ kann aus χ^2 berechnet werden mit:

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

Beispiel

In unserem Beispiel ergibt sich also für ϕ durch Einsetzung in Gleichung (9.5):

$$\begin{aligned}\phi &= \sqrt{\frac{\chi^2}{n}} \\ &\approx \sqrt{\frac{1{,}421}{90}} \\ &\approx 0{,}126\end{aligned}$$

Es wird ersichtlich, dass es eine leichte Korrelation der Variablen gibt. Aber in welche Richtung? Dafür müssen wir auf die Kreuztabelle 9.5 blicken: Der beobachtete Wert für die Wertekombination „Fahrrad“ und „Geistes-/Sozialwissenschaft“ beträgt $n_{11} = 11$ und liegt über dem Erwartungswert $m_{11} = 8{,}67$. Damit ist klar: Das Studium von Geistes- und Sozialwissenschaften korreliert *positiv* mit der Fahrradnutzung für den Pendelweg.

Ob diese Korrelation auch statistisch relevant ist, kann mit dem χ^2-Test (Sitzung 10) überprüft werden.

9.6 Cramér-Index

Bisher wurden in dieser Sitzung nur Verteilungen von zwei dichotomen Variablen besprochen. Nun gibt es aber auch nominalskalierte bivariate Verteilungen, in denen die Merkmale mehr als zwei Werte annehmen können (also nicht dichotom sind). In diesem Fall ist der Cramér-Index (auch Cramérs v, engl. *Cramér index*) ein geeigneter Kennwert für die Abhängigkeit der Variablen.

Die Formel für den Cramér-Index lautet

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

wobei der Ausdruck $\min(k,\ell)$ für den *kleineren* Wert aus Zeilenanzahl k und Spaltenanzahl ℓ steht.

In einer 2×2-Tabelle ist dieser Wert identisch mit dem ϕ-Koeffizienten.

Tabelle 9.6: Kreuztabelle des Beispiels ohne Dichotomisierung

	→ Transportmittel			
Studienfach ↓	**Fahrrad**	**Auto**	**Öffentliche**	
Geisteswissenschaft	5	5	9	**19**
	(4,22)	(8,02)	(6,76)	
	0,144	1,137	0,742	
Sozialwissenschaft	6	6	8	**20**
	(4,44)	(8,44)	(7,11)	
	0,548	0,705	0,111	
Naturwissenschaft	5	9	9	**23**
	(5,11)	(9,71)	(8,18)	
	0,002	0,052	0,082	
Ingenieurswissenschaft	4	18	6	**28**
	(6,22)	(11,82)	(9,96)	
	0,792	3,231	1,574	
	20	**38**	**32**	**90**

Beispiel

Hätten wir im Beispiel die Erhebung nicht auf dichotome Variablen reduziert, sondern die Wissenschaftsdisziplinen und Verkehrsmittel direkt erhoben, so würde sich die Kreuztabelle vielleicht wie in Tabelle 9.6 darstellen.

Dabei werden die Erwartungswerte wie gehabt mit Gleichung (9.3) und die Teilwerte für χ^2 mit Gleichung (9.4) errechnet.

Der χ^2-Wert ergibt sich wieder aus der Summe (s. Gleichung (10.1)):

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\\ &\approx 0{,}144 + 1{,}137 + 0{,}742 + 0{,}548 + 0{,}705 + 0{,}111\\ &\quad + 0{,}002 + 0{,}052 + 0{,}082 + 0{,}792 + 3{,}231 + 1{,}574\\ &= 9{,}120\end{aligned}$$

Mit diesem Wert kann der Cramér-Index anhand von Gleichung (9.6) berechnet werden.

Die Zeilenanzahl ist $k = 4$ und die Spaltenanzahl $\ell = 3$. Der Ausdruck $\min(k,\ell)$ ergibt den kleineren dieser Werte, also 3:

$$
\begin{aligned}
CI &= \sqrt{\frac{\chi^2}{n \cdot (\min(k, \ell) - 1)}} \\
&\approx \sqrt{\frac{9{,}122}{90 \cdot (3 - 1)}} \\
&\approx 0{,}225
\end{aligned}
$$

Dieser Wert ist größer als der oben berechnete ϕ-Koeffizient. Das ist nicht besonders überraschend: Eine detailliertere Erfassung der Variablen führt zu einem deutlicheren Zusammenhang.

Tipps zur Vertiefung

- Kapitel 9.1, 10.3.4 und 10.3.7 in Bortz und Schuster (2010)
- Kapitel 4.2.2 – 4.2.3 in Lange und Nipper (2018)
- Kapitel 6.7.2 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 2.3 in Klemm (2002)
- *Englisch:* Kapitel 12.2 in Burt und Barber (1996)

Übungsaufgaben

Aufgabe 9-1

zur Lösung

Sie fragen sich, wie die Wohnumgebung einer Person (Stadt oder Land) damit zusammenhängt, ob die Person ein eigenes Auto besitzt. Sie erheben die folgende Messreihe:

Wohnort	Autobesitz
Land	Ja
Land	Ja
Stadt	Nein
Stadt	Nein
Stadt	Ja
Stadt	Nein
Land	Ja
Land	Nein
Land	Ja
Land	Ja
Stadt	Nein
Land	Ja
Land	Ja
Land	Ja
Stadt	Nein
Land	Ja
Stadt	Nein
Land	Nein
Stadt	Ja
Stadt	Nein

a) Überführen Sie die Daten in eine Kreuztabelle.

b) Berechnen Sie die Erwartungswerte für jedes Tabellenfeld.

c) Berechnen Sie χ^2.

d) Berechnen Sie den ϕ-Koeffizienten.

e) Besteht eine Korrelation? In welche Richtung?

Aufgabe 9-2

zur Lösung

Sie interessieren sich dafür, ob zwei „Ja/Nein“-Fragen auf einem Fragebogen korrelieren.

Sie ermitteln folgende Häufigkeiten:

	→ Frage 2	
Frage 1 ↓	**Ja**	**Nein**
Ja	5	28
Nein	40	72

a) Vervollständigen Sie die Kreuztabelle um ihre Summen und die Erwartungswerte.

b) Berechnen Sie χ^2 und den ϕ-Koeffizienten.

c) Wie würden Sie den Zusammenhang beschreiben?

Aufgabe 9-3

zur Lösung

Sie möchten überprüfen, ob auf dem Arbeitsmarkt anhand von Namen diskriminiert wird, die auf einen Migrationshintergrund schließen lassen. Sie antworten als fiktive Bewerber*innen mit vergleichbaren Qualifikationen auf zufällige Stellenanzeigen und halten fest, ob die jeweilige Bewerbung in einer Einladung zum Vorstellungsgespräch resultiert.

Sie erheben diese Daten:[1]

	→ Ergebnis	
Herkunft des Namens ↓	**eingeladen**	**nicht eingeladen**
deutsch	36	64
italienisch	23	77
slawisch	9	91
türkisch	11	89

Können Sie einen Zusammenhang zwischen Namensherkunft und Erfolg der Bewerbung feststellen? Begründen Sie Ihre Antwort.

Aufgabe 9-4

zur Lösung

In einer breit angelegten Befragung von Haushalten wird erhoben,

1) ob die Proband*innen zur Miete wohnen und
2) welchen Internetanschluss sie nutzen.

Sie fassen die Ergebnisse in einer Kreuztabelle zusammen:

	→ Wohnverhältnis	
Internetanschluss ↓	**Miete**	**Eigentum**
Glasfaser	1926	1567
DSL	2758	3686
Koaxialkabel	3002	1903
Kein fester Anschluss	1277	167

Berechnen Sie den Cramér-Index und interpretieren Sie das Ergebnis.

[1] Diese Zahlen sind fiktiv. Echte Ergebnisse sogenannter Korrenspondenztests zu ähnlichen Fragestellungen sind bei Veit (2020) zusammengefasst.

Sitzung 10

χ^2-Tests

Lernziele dieser Sitzung

Sie können…

- einen χ^2-Unabhängigkeitstest durchführen.
- einen χ^2-Anpassungstest durchführen.

Lernvideos (Sommersemester 2020)

- 10a) χ^2-Unabhängigkeitstest
- 10a) χ^2-Anpassungstest
  - Der Hinweis am Ende auf die Vorbereitungssitzung ist natürlich nicht mehr aktuell. Wir treffen uns am 29. Juni und am 6. Juli 2021 zur Klausurvorbereitung.

Anwendungsbereich

In Sitzung 9 haben wir gelernt, wie für bivariate Verteilungen Korrelationen beschrieben werden können, wenn beide Variablen nominalskaliert sind. Grundlage dafür waren die Häufigkeiten von Wertekombinationen in der Kreuztabelle.

Auch für χ^2-Tests sind beobachtete Häufigkeiten in einer Kreuztabelle unser Ausgangspunkt. Wir fragen jedoch nicht nach einem Kennwert für die Stärke der Korrelation, sondern wollen wissen, ob es einen statistisch signifikanten Zusammenhang zwischen den beiden Variablen gibt – also einen Zusammenhang, der höchstens mit einer Wahrscheinlichkeit α (Signifikanzniveau) zufällig zustande gekommen sein kann.

Um den Unterschied zu verdeutlichen: Bei sehr großen Fallzahlen kann auch eine leichte Korrelation statistisch signifikant sein, bei kleinen Fallzahlen wird es es selbst für starke Korrelationen schwierig, eine statistische Signifikanz nachzuweisen.

Mit dem χ^2-Unabhängigkeitstest und dem χ^2-Anpassungstest lernen wir im Folgenden zwei unterschiedliche Varianten des χ^2-Tests kennen. Beide sollen direkt an Beispielen ausgeführt werden.

Tabelle 10.1: Kreuztabelle der Beispieldaten

	→ Dienst	
Wohnort ↓	**Grundwehrdienst**	**Zivildienst**
Land	18	11
Stadt	10	23

10.1 χ^2-Unabhängigkeitstest

Grundlage sind bivariate Häufigkeiten, die in einer Kreuztabelle dargestellt werden können (s. Tabelle 10.1). Wie Kreuztabellen erstellt werden, haben wir bereits in Sitzung 9 behandelt.

Unser Beispieldatensatz beschäftigt sich mit Kriegsdienstverweigerern. Zwischen 1956 und 2011 galt in der BRD die Wehrpflicht, d. h. alle vom Staat als „männlich“ erfassten und als „tauglich“ gemusterte jungen Menschen mussten Dienst an der Waffe leisten – es sei denn, sie verweigerten den Kriegsdienst und leisteten stattdessen Zivildienst (z. B. in sozialen Einrichtungen).

Zusätzlich zur Frage der Kriegsdienstverweigerung sei in einer Zufallsstichprobe von als tauglich gemusterten erhoben, ob der Wohnort eine Gemeinde mit über oder unter 20 000 Einwohner*innen („Stadt“ oder „Land“) ist.[1] Die Ergebnisse sind in Tabelle 10.1 zusammengefasst.

Wir interessieren uns für den statistischen Zusammenhang dieser beiden Variablen, und zwar möchten wir die Hypothese prüfen, dass Menschen aus der Stadt eher den Kriegsdienst verweigerten als Menschen vom Land. Der Test wird entlang der bekannten sechs Schritte ausgeführt.

Test wählen und Voraussetzungen prüfen

Für den χ^2-Unabhängigkeitstest müssen folgende Voraussetzungen erfüllt sein:

- Ziel ist die Überprüfung einer bivariaten Verteilung auf einen statistisch signifikanten Zusammenhang zwischen zwei nominalskalierten Variablen.
- Grundlage sind beobachtete Häufigkeiten aus einer einfachen, unabhängigen Zufallsstichprobe.
- Alle Tabellenfelder enthalten beobachtete Häufigkeiten $(n_{ij} \geq 5)$.

Für unsere Beispieldaten sind diese Voraussetzungen gegeben.

Hypothesen formulieren

Wir haben wieder zwei Möglichkeiten: die gerichtete und die ungerichtete Alternativhypothese.

Ungerichtete Alternativhypothese

Wir verzichten an dieser Stelle auf mathematische Notationen und würden bei ungerichteter Alternativhypothese im Klartext schreiben:

[1] Hier wird also eine verhältnisskalierte Variable (Bevölkerungszahl der Gemeinde) in eine nominalskalierte Variable transformiert. In Fällen wie diesen, wo die Variable nach der Transformation nur zwei Werte annehmen kann, sprechen wir auch von der „Dichotomisierung“ einer Variable.

H_0 : Es gibt keinen Zusammenhang zwischen Wohnort und Verweigerungsentscheidung.
H_1 : Es gibt einen Zusammenhang zwischen Wohnort und Verweigerungsentscheidung.

Gerichtete Alternativhypothese

Im Falle einer gerichteten Alternativhypothese bleibt die Nullhypothese bestehen, aber die Alternativhypothese gibt eine bestimmte Richtung des Zusammenhangs vor.

H_0 : Es gibt keinen Zusammenhang zwischen Wohnort und Verweigerungsentscheidung.
H_1 : Es gibt einen positiven Zusammenhang zwischen Wohnort in der Stadt und Kriegsdienstverweigerung.

Gerichtete Alternativhypothesen sind im χ^2-Unabhängigkeitstest *nur* für 2×2-Tabellen möglich.

Im Beispiel entscheiden wir uns für die gerichtete Alternativhypothese, denn wir vermuten einen Zusammenhang in diese bestimmte Richtung.

Signifikanzniveau entscheiden

Wie in anderen Tests ist ein Signifikanzniveau von $\alpha = 0{,}05$ üblich, wofür wir uns auch im Beispiel entscheiden.

Kritischen Wert bestimmen

Bei χ^2-Tests gibt es immer nur einen kritischen Wert. Zunächst müssen beim χ^2-Unabhängigkeitstest die Freiheitsgrade bestimmt werden mit der Formel:

$$df = (k-1) \cdot (\ell - 1)$$

wobei auch hier wieder k für die Zeilenanzahl und ℓ für die Spaltenanzahl steht.

Im Beispiel also:

$$\begin{aligned} df &= (k-1) \cdot (\ell - 1) \\ &= (2-1) \cdot (2-1) = 1 \end{aligned}$$

Damit lässt sich der kritische Wert aus der Formelsammlung ablesen, die allerdings für *ungerichtete* Alternativhypothesen ausgelegt ist.

Hätten wir eine ungerichtete Alternativhypothese gewählt, würde der Ablehnungsbereich also definiert durch:

$$\begin{aligned} \chi^2 &\geq \chi^2_{df;(1-\alpha)} \\ \chi^2 &\geq \chi^2_{1;95\%} \\ \chi^2 &\geq 3{,}841 \end{aligned}$$

Tabelle 10.2: Kreuztabelle mit Erwartungswerten und Teilwerten für χ^2

	→ Dienst		
Wohnort ↓	**Grundwehrdienst**	**Zivildienst**	
Land	18	11	**29**
	(13,1)	(15,9)	
	1,833	1,51	
Stadt	10	23	**33**
	(14,9)	(18,1)	
	1,611	1,327	
	28	**34**	**62**

Für unsere *gerichtete* Alternativhypothese „dürfen“ wir den Ablehnungsbereich jedoch verdoppeln (müssen aber einem späteren Schritt unbedingt auch prüfen, ob die Richtung stimmt):

$$\begin{aligned}\chi^2 &\geq \chi^2_{df;(1-2\cdot\alpha)}\\ \chi^2 &\geq \chi^2_{1;90\%}\\ \chi^2 &\geq 2{,}706\end{aligned}$$

Prüfgröße berechnen

Wie in Sitzung 9 besprochen, wird die Prüfgröße χ^2 anhand der Formel

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

errechnet. Dabei ist die Ermittlung der Randsummen und Erwartungswerte m_{ij} ein notwendiger Schritt, und auch die Teilwerte für χ^2 können wieder direkt in die Kreuztabelle eingetragen werden.

Für unser Beispiel erfolgt die Berechnung anhand Tabelle 10.2.

Zunächst muss dabei geprüft werden, ob die Richtung unserer Alternativhypothese stimmt. Die beobachtete Häufigkeit der Zivildienstleistenden in der Stadt $n_{22} = 23$ ist größer als der Erwartungswert $m_{22} = 18{,}1$. Wenn eine Signifikanz nachgewiesen werden kann, dann also für den *positiven* Zusammenhang zwischen Wohnort in der Stadt und Kriegsdienstverweigerung (wie in unserer Alternativhypothese spezifiziert).

Für χ^2 ergibt sich im Beispiel:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\\ &= 1{,}833 + 1{,}51 + 1{,}611 + 1{,}327\\ &= 6{,}281\end{aligned}$$

Tabelle 10.3: Beispielwerte für Verspätungen nach Wochentagen

Montag	Dienstag	Mittwoch	Donnerstag	Freitag
409	387	437	414	459

Ergebnis interpretieren

Der Wert der Prüfgröße $\chi^2 = 6{,}281$ liegt deutlich im Ablehnungsbereich $\chi^2 \geq 2{,}706$. Die Nullhypothese kann abgelehnt werden. Es wurde ein statistisch signifikanter positiver Zusammenhang zwischen Wohnort in Gemeinden mit über 20 000 Einwohner*innen und Kriegsdienstverweigerung festgestellt ($\alpha = 0{,}05$).

Softwarehinweis

In R lässt sich ein χ^2-Unabhängigkeitstest mit dem Befehl `chisq.test()` durchführen.

10.2 χ^2-Anpassungstest

Beim χ^2-Anpassungstest geht es um die Häufigkeiten *eines* nominalskalierten Merkmals – er ist deshalb der univariaten Teststatistik zuzuordnen. Der Test überprüft, ob das Merkmal entlang einer vorgegebenen Verteilung (im Normalfall gleichmäßig) verteilt ist, oder ob es signifikante Abweichungen von dieser erwarteten Verteilung gibt.

Ein Beispiel: Für größere Verspätungen ($\geq$ 10 Minuten) beim ÖPNV einer Großstadt wird festgehalten, an welchen Wochentagen sie auftreten. Wir ignorieren Wochenenden und Feiertage und fragen uns, ob sich die Verzögerungen gleichmäßig auf Werktage verteilen, oder ob es signifikante Abweichungen in Bezug auf den Wochentag gibt. Die Werte in Tabelle 10.3 seien über drei Monate hinweg erhoben worden.

Wir befolgen wieder die sechs Schritte für statistische Testverfahren.

Test wählen und Voraussetzungen prüfen

Für den χ^2-Anpassungstest müssen folgende Voraussetzungen erfüllt sein:

- Ziel ist die Überprüfung *einer* nominalskalierten Variable auf eine statistisch signifikante Abweichung von einer vorgegebenen Verteilung.
- Grundlage sind beobachtete Häufigkeiten aus einer einfachen, unabhängigen Zufallsstichprobe.
- Alle Tabellenfelder enthalten beobachtete Häufigkeiten ($n_i \geq 5$).

In unserem Beispiel sind diese Voraussetzungen gegeben.

Hypothesen formulieren

H_0 : Starke Verspätungen sind an allen Werktagen gleich wahrscheinlich.

H_1 : Starke Verspätungen sind an manchen Werktagen wahrscheinlicher als an anderen.

Gerichtete Hypothesen dürften hier wieder nur bei dichotomen Variablen formuliert werden (also bei genau zwei Tabellenfeldern) – denn sonst können wir die Richtung der Vermutung nicht genau genug formulieren.

Signifikanzniveau entscheiden

Wie gewohnt: $\alpha = 0{,}05$

Kritischen Wert bestimmen

Die Freiheitsgrade bestimmen sich aus

$$df = k - 1$$

wobei k hier einfach die Anzahl der Katorien ist.

In unserem Beispiel (bei fünf Werktagen) also:

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

Der kritische Wert für den Ablehnungsbereich ist der Formelsammlung zu entnehmen.

$$\begin{aligned} \chi^2 &\geq \chi^2_{df;(1-\alpha)} \\ \chi^2 &\geq \chi^2_{4;95\%} \\ \chi^2 &\geq 9{,}488 \end{aligned}$$

Auch hier dürften wir bei einer gerichteten Hypothese den Ablehnungsbereich verdoppeln, d. h. der kritische Wert $\chi^2_{df;(1-2\cdot\alpha)}$ wäre anzuwenden – dies ist allerdings wie bereits erwähnt nur für dichotome Variablen möglich.

Prüfgröße berechnen

Die Prüfgröße χ^2 berechnet sich analog zu vorherigen Beispielen. Einzige Besonderheit: Die Erwartungswerte werden direkt anhand der zu erwartenden (im unserem Fall: gleichmäßigen) Verteilung bestimmt.

Im Beispiel ergibt sich in den fünf Kategorien jeweils ein Erwartungswert von

$$\frac{n}{k} = \frac{2106}{5} = 421{,}2$$

Dann nehmen wir wieder eine Tabelle zu Hilfe um die Prüfgröße χ^2 zu berechnen (s. Tabelle 10.4). Wie gehabt werden einfach die Teilwerte zusammengezählt:

Tabelle 10.4: Tabelle für den χ^2-Anpassungstest

Montag	**Dienstag**	**Mittwoch**	**Donnerstag**	**Freitag**	
409	387	437	414	459	**2106**
(421,2)	(421,2)	(421,2)	(421,2)	(421,2)	
0,353	2,777	0,593	0,123	3,392	

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i} \\ &\approx 0{,}353 + 2{,}777 + 0{,}593 + 0{,}123 + 3{,}392 \\ &= 7{,}238\end{aligned}$$

Ergebnis interpretieren

Der Ablehnungsbereich $\chi^2 \geq 9{,}488$ wurde nicht erreicht. Die Nullhypothese muss beibehalten werden. Eine statistisch signifikante Abweichung von einer gleichmäßigen Verteilung konnte nicht nachgewiesen werden ($\alpha = 0{,}05$).

Softwarehinweis

Mit einer univariaten Verteilung als Eingabe führt der Befehl `chisq.test()` einen χ^2-Anpassungstest durch.

Andere Verteilungen

Die theoretische Verteilung, von der eine signifikante Abweichung festgestellt werden soll, ist im obigen Beispiel uniform, d. h. die Erwartungswerte sind gleichmäßig über die Wochentage verteilt. Allerdings kann beim Anpassungstest auch von anderen Verteilungen ausgegangen werden – so könnte eine (begründete) Nullhypothese auch lauten, dass Kategorie A doppelt so viele Fallzahlen aufweist wie Kategorie B und C.

In der Praxis wird der χ^2-Anpassungstest oft verwendet, um nachzuweisen, dass *keine* signifikante Abweichung von der Normalverteilung zu beobachten ist – nur dann dürfen nämlich viele statistische Verfahren durchgeführt werden.

Tipps zur Vertiefung

- Kapitel 9 in Bortz und Schuster (2010)
- Kapitel 8.2.7 in Lange und Nipper (2018) (χ^2-Anpassungstest)
- Kapitel 5.3.4 in Bahrenberg, Giese und Nipper (2010)
- Kapitel 13 in Klemm (2002)
- *Englisch:* Kapitel 11.3 in Burt und Barber (1996)

Formelsammlung und Wertetabellen

Die Formelsammlung mit Wertetabellen liegt als PDF vor. So (oder ähnlich) formatiert würden Ihnen diese Informationen auch in einer Präsenzklausur zur Verfügung stehen. Ich empfehle deshalb, das Dokument herunterzuladen und auszudrucken – so gewöhnen Sie sich gleich an das Format.

Als PDF herunterladen

Lösungen der Übungsaufgaben

Sitzung 1

Lösung 1-1

zur Aufgabenstellung

– keine Musterlösung –

Lösung 1-2

zur Aufgabenstellung

– keine Musterlösungen –

Lösung 1-3

zur Aufgabenstellung

	Variable	Skalenniveau	Variablentyp	Anmerkungen
a)	Lebensalter in Jahren	Verhältnisskala	diskret	ganze Zahlen vorausgesetzt
b)	Regenmenge in mm	Verhältnisskala	stetig	
c)	Güteklasse	Ordinalskala	qualitativ	
d)	Passagieraufkommen	Verhältnisskala	diskret	
e)	Baujahr	Intervallskala	diskret	
f)	Geschwindigkeit in km/h	Verhältnisskala	stetig	bei ganzzahligen Werten: dis
g)	Sozialstatus (Unter-, Mittel und Oberschicht)	Ordinalskala	qualitativ	
h)	Temperatur in °F	Intervallskala	stetig	
i)	Fläche eines Bundeslands in km²	Verhältnisskala	stetig	
j)	Temperatur in K	Verhältnisskala	stetig	0 K ist ein natürlicher Nullpu
k)	Einwohnerzahl	Verhältnisskala	diskret	
l)	Pegelstand	Intervallskala	stetig	willkürlicher Nullpunkt
m)	Staatsangehörigkeit	Nominalskala	qualitativ	
n)	Interesse an Statistik (gering bis hoch)	Ordinalskala	qualitativ	
o)	Klausurnote	Ordinalskala	qualitativ	wird jedoch oft metrisch ver
p)	Bodentyp	Nominalskala	qualitativ	
q)	Entfernung zum Stadtzentrum in km	Verhältnisskala	stetig	
r)	Körpergröße	Verhältnisskala	stetig	
s)	Kleidergröße (S bis XXL)	Ordinalskala	qualitativ	
t)	Monatliches Nettoeinkommen	Verhältnisskala	stetig	oder diskret für Cent-Beträg

Lösung 1-4

zur Aufgabenstellung

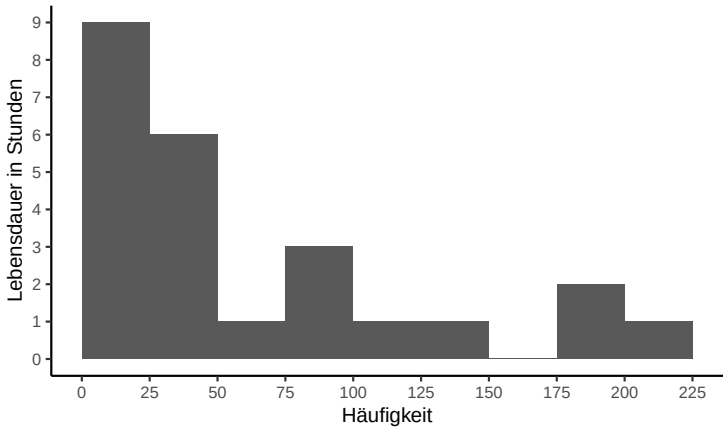
a)

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

b)

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



c)

Die Verteilung ist unregelmäßig abfallend.

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

Lösung 2-1

zur Aufgabenstellung

a)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{356{,}00}{6}$
Ergebnis	$\bar{x} = 59{,}33$

b)

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

c)

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

Lösung 2-2

zur Aufgabenstellung

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	$s \approx 15{,}68$

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	$s \approx 0{,}48$

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	$s \approx 102{,}92$

Lösung 2-3

zur Aufgabenstellung

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:

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

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:

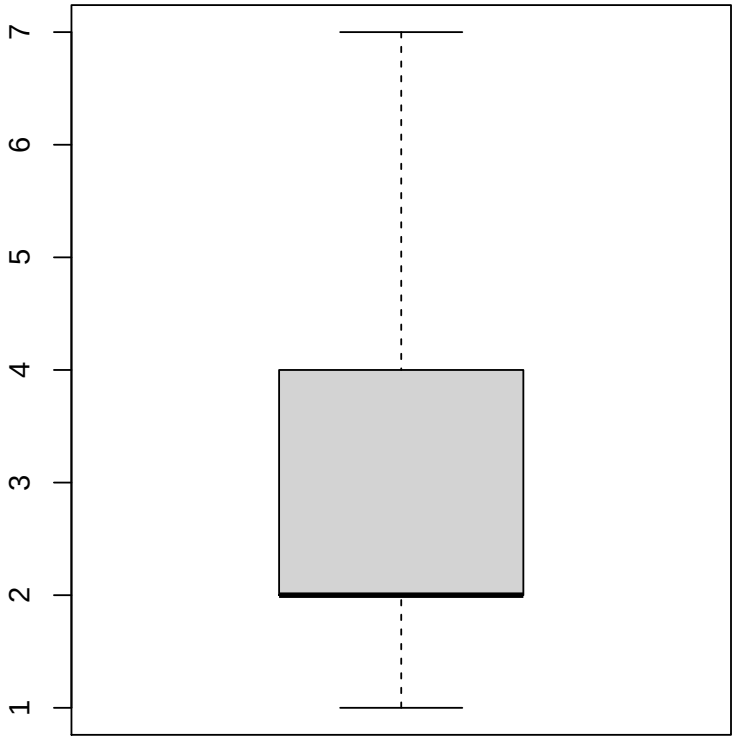
$$s^2 = \frac{\sum_{x=1}^{n} (x_i - \bar{x})^2}{n-1} \approx \frac{10{,}22 + 2{,}86 + 0{,}05 + 2{,}66 + 4{,}64 + 17{,}25}{13-1} = \frac{37{,}68}{12} = 3.14$$

Schließlich ist die Standardabweichung:

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

c)

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



Lösung 2-4

zur Aufgabenstellung

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

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

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:

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

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:

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

d)

Somit beträgt die Standardabweichung

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

Lösung 2-5

zur Aufgabenstellung

a)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{511{,}00}{6}$
Ergebnis	$\bar{x} = 85{,}17$
Einsetzen	$\bar{y} = \frac{446{,}00}{6}$
Ergebnis	$\bar{y} = 74{,}33$
Antwortsatz	Die Ziegelei weist im Mittel die größere Passantinnenzahl auf.

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.

Lösung 2-6

zur Aufgabenstellung

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.

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

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.

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.

e)

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

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

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:

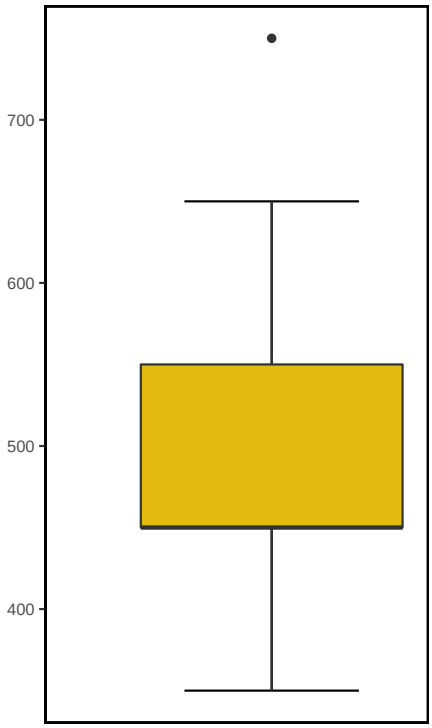
$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{72900 + 11025 + 16900 + 54450 + 70225}{20 - 1} \\ &= \frac{225500}{19} \\ &\approx 11868{,}42 \end{aligned}$$

Somit beträgt die Standardabweichung

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

g)

Auch der Boxplot lässt sich anhand der klassierten Werte zeichnen:



Sitzung 3

Lösung 3-1

zur Aufgabenstellung

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	$z_7 = \frac{-23{,}61 + 18{,}94}{7{,}82} \approx -0{,}60$
-25,90	$z_8 = \frac{-25{,}9 + 18{,}94}{7{,}82} \approx -0{,}89$
-27,08	$z_9 = \frac{-27{,}08 + 18{,}94}{7{,}82} \approx -1{,}04$

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	$x_7 = 2{,}87 \cdot 3{,}61 + 221{,}54 \approx 231{,}90$
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	$x_{12} = 0{,}68 \cdot 3{,}61 + 221{,}54 \approx 223{,}99$

Lösung 3-2

zur Aufgabenstellung

a)

σ lässt sich berechnen durch:

Schritt	**Lösung**
Formel	$\sigma = \sqrt{\sigma^2}$
Einsetzen	$\sigma = \sqrt{19{,}36}$
Lösung	$\sigma \approx 4{,}40$

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	$z_2 = \frac{20{,}77 - 32{,}2}{4{,}4} \approx -2{,}60$
33,41	$z_3 = \frac{33{,}41 - 32{,}2}{4{,}4} \approx 0{,}27$
44,95	$z_4 = \frac{44{,}95 - 32{,}2}{4{,}4} \approx 2{,}90$
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%

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**	**....**	**Ergebnis**
0,015	0,985	$P(z \leq z_1) = 0{,}985$		$z_1 \approx 2{,}17$
0,025	0,975	$P(z \leq z_2) = 0{,}975$		$z_2 \approx 1{,}96$
0,050	0,950	$P(z \leq z_3) = 0{,}95$		$z_3 \approx 1{,}64$
0,130	0,870	$P(z \leq z_4) = 0{,}87$		$z_4 \approx 1{,}13$
0,500	0,500	$P(z \leq z_5) = 0{,}5$		$z_5 \approx 0{,}00$
0,900	0,100	$P(z \leq -z_6) = 1 - 0{,}1 = 0{,}9$	$-z_6 \approx 1{,}28$	$z_6 \approx -1{,}28$
0,990	0,010	$P(z \leq -z_7) = 1 - 0{,}01 = 0{,}99$	$-z_7 \approx 2{,}33$	$z_7 \approx -2{,}33$
0,995	0,005	$P(z \leq -z_8) = 1 - 0{,}005 = 0{,}995$	$-z_8 \approx 2{,}58$	$z_8 \approx -2{,}58$

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	$x_5 = 0 \cdot 4{,}4 + 32{,}2$	$x_5 \approx 32{,}20$
-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	$x_8 = -2{,}58 \cdot 4{,}4 + 32{,}2$	$x_8 \approx 20{,}85$

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.

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	$z_u = \frac{30 - 32{,}2}{4{,}4} \approx -0{,}50$
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 zwische

Lösung 3-3

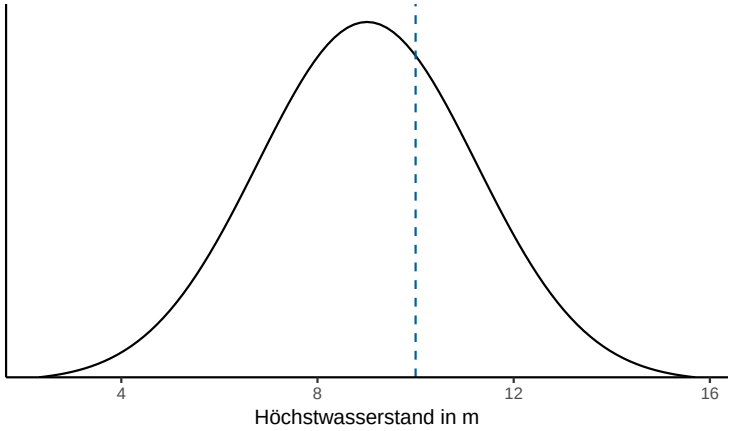
zur Aufgabenstellung

a)

Siehe b)

b)

Die Dichtefunktion mit kritischem Wert sollte in etwa so aussehen:



c)

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

d)

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

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

Lösung 3-4

zur Aufgabenstellung

a)

Die Übertretungswahrscheinlichkeit beträgt:

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

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:

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

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:

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

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.

Lösung 3-5

zur Aufgabenstellung

a)

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

b)

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

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.

Lösung 3-6

zur Aufgabenstellung

a)

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

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.

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 $P(z < z_o) \approx 0{,}9332$
- $P(200 < x < 550) = P(x < 550) - P(x < 200) \approx 91{,}04\%$

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 $z_p \approx 2{,}05$
- $x_p = 605$

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.

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	$v \approx 38{,}13\%$

Sitzung 4

Lösung 4-1

zur Aufgabenstellung

a) $\mu = \bar{x} = 162$

$\sigma = s \approx 13{,}30$

b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{13{,}30}{\sqrt{6}} \approx 5{,}43$

Lösung 4-2

zur Aufgabenstellung

a) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{9}} \approx 1{,}33$

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

$\frac{KIB}{2} = z_{97{,}5\%} \cdot \sigma_{\bar{x}}$

$\frac{KIB}{2} \approx 1{,}96 \cdot 1{,}33 \approx 2{,}61$

$KIB = 5{,}22$

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

$z_{(1-\alpha/2)} = \frac{KIB}{2 \cdot \sigma_{\bar{x}}} \approx \frac{1}{2 \cdot 1{,}33} \approx 0{,}38$

$1 - \frac{\alpha}{2} \approx 0{,}648$

$-\frac{\alpha}{2} \approx 0{,}648 - 1$

$\frac{\alpha}{2} \approx 0{,}352$

$\alpha \approx 0{,}704$

Das Konfidenzniveau beträgt ca. 70,4%.

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

$\sigma_{\bar{x}} = \frac{KIB}{2 \cdot z_{95\%}}$

$\sigma_{\bar{x}} = \frac{2}{2 \cdot z_{95\%}}$

$\sigma_{\bar{x}} \approx \frac{2}{2 \cdot 1{,}65}$

$\sigma_{\bar{x}} \approx 0{,}61$

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

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

$n \approx \left(\frac{4}{0{,}61}\right)^2 \approx 43$

Lösung 4-3

zur Aufgabenstellung

a)

$\alpha = 0{,}1$

$\sigma = \sqrt{\sigma^2} = \sqrt{4096} = 64$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{64}{\sqrt{40}} \approx 10{,}12$

$\frac{KIB}{2} = z_{95\%} \cdot \sigma_{\bar{x}}$

$\frac{KIB}{2} \approx 1{,}65 \cdot 10{,}12 \approx 16{,}70$

$\text{Untergrenze} = \bar{x} - \frac{KIB}{2} \approx 2650 - 16{,}70 = 2633{,}30$

$\text{Obergrenze} = \bar{x} + \frac{KIB}{2} \approx 2650 + 16{,}70 = 2666{,}70$

b)

$KIB = 20$

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

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

$z_{(1-\alpha/2)} = \frac{20}{2 \cdot 10,12} \approx 0,99$

$1 - \frac{\alpha}{2} \approx 0,8389$

$\alpha \approx 0,3222$

Das Konfidenzniveau beträgt ca. 67,78%.

Lösung 4-4

zur Aufgabenstellung

a)

Schritt	Lösung
Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Einsetzen	$\bar{x} = \frac{759,50}{7}$
Ergebnis	$\bar{x} = 108,50$

b)

Schritt	Lösung
Formel	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Einsetzen	$\sigma_{\bar{x}} = \frac{11,5}{\sqrt{7}}$
Ergebnis	$\sigma_{\bar{x}} \approx 4,35$

c)

Schritt	Lösung
Formel	$\frac{KIB}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}}$
Einsetzen	$\frac{KIB}{2} = z_{97,5\%} \cdot \sigma_{\bar{x}} \approx 1,96 \cdot 4,35$
Ergebnis	$\frac{KIB}{2} \approx 8,53$
Antwortsatz	Die tatsächliche durchschnittliche Lieferzeit liegt mit 95% Wahrscheinlichkeit zwischen 99,97 und 11

d)

Schritt	Lösung
Standardfehler: Formel	$\frac{KIB}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}}$
Standardfehler: Umformen	$\sigma_{\bar{x}} = \frac{KIB}{2} \cdot \frac{1}{z_{(1-\alpha/2)}}$
Standardfehler: Einsetzen	$\sigma_{\bar{x}} = \frac{KIB}{2} \cdot \frac{1}{z_{99,5\%}} = 8{,}53 \cdot \frac{1}{2{,}58}$
Standardfehler: Ergebnis	$\sigma_{\bar{x}} \approx 3{,}31$
n: Formel	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
n: Umformen	$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$
n: Einsetzen	$n = \left(\frac{11{,}5}{3{,}31}\right)^2$
n: Ergebnis	$n \approx 12{,}07$
Antwortsatz	Es müssten 6 zusätzliche Messungen vorgenommen werden (13 insgesamt).

Sitzung 5

Lösung 5-1

zur Aufgabenstellung

a) Ob die Grundgesamtheit normalverteilt ist oder nicht, ist nicht bekannt. (Vermutlich ist das sogar nicht der Fall.) Deshalb muss die Stichprobengröße mindestens 30 betragen.

b) $H_0 : \mu = 2{,}30$

$H_1 : \mu \neq 2{,}30$

c) $z \leq -1{,}96$ und $z \geq 1{,}96$

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

$z = \sqrt{40} \cdot \frac{1{,}82-2{,}30}{1{,}42} \approx -2{,}14$

e) Der z-Wert ist mit -2,14 kleiner als der kritische Wert -1,96 und damit im Ablehnungsbereich. Die Nullhypothese kann verworfen werden. Die Vermutung, dass sich die Frankfurter Haushaltsgröße vom europäischen Durchschnitt unterscheidet, ist damit bestätigt.

Lösung 5-2

zur Aufgabenstellung

a) -4,604
b) -3,579
c) -2,365
d) -1,771
e) 2,201
f) 2,353
g) 3,707
h) 3,686
i) 3,365

j) -2,528

Lösung 5-3

zur Aufgabenstellung

1. Voraussetzungen prüfen (Test wählen):

   z-Test, da σ bekannt

2. Hypothesen formulieren:

   $H_0 : \mu = 61{,}5$

   $H_1 : \mu < 61{,}5$

3. Signifikanzniveau entscheiden:

   Signifikanzniveau z.B. $alpha = 0{,}05$, weil ein zu großes $alpha$ hier nicht in besonderer Weise problematisch ist.

4. Kritischen Wert bestimmen:

   $z \leq -1{,}65$

5. Prüfgröße berechnen:

   Zunächst muss $\bar{x} = 57{,}75$ berechnet werden (s. Sitzung 2)

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

   $z \approx \sqrt{4} \cdot \frac{57{,}75 - 61{,}5}{10{,}3} \approx -0{,}73$

6. Nullhypothese ablehnen oder beibehalten:

   Der kritsche Wert wurde nicht erreicht. Die Nullhypothese muss beibehalten werden, eine systematisch schlechtere Prüfungsleistung von berufstätigen Studierenden ließ sich hier nicht bestätigen.

Lösung 5-4

zur Aufgabenstellung

a) Es geht um den Vergleich des Mittelwerts einer Stichprobe mit dem Mittelwert der Grundgesamtheit bei unbekanntem σ,s deshalb 1-Stichproben-t-Test.

b) Gerichtete Alternativhypothese nach unten:

$$H_0 : \mu = 3042{,}43$$
$$H_1 : \mu < 3042{,}43$$

c) Stichprobengröße 6, also 5 Freiheitsgrade:

$$t \leq t_{5;1\%}$$
$$t \leq -3{,}365$$

Lösung 5-5

zur Aufgabenstellung

a) Wir berechnen zunächst die Parameter der Stichprobe (s. Sitzung 2):

$$\bar{x} \approx 2964{,}50$$
$$s \approx 51{,}93$$

Und setzen anschließend ein:

$$\begin{aligned} t &= \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s} \\ &= \sqrt{6} \cdot \frac{2964{,}50 - 3042{,}43}{51{,}93} \\ &\approx -3{,}676 \end{aligned}$$

b) Der kritische Wert wurde unterschritten, die Nullhypothese wird abgelehnt. Wir haben gezeigt, dass in diesem Betrieb Angestellte mit Migrationshintergrund schlechter bezahlt werden ($\alpha = 0{,}01$).

Lösung 5-6

zur Aufgabenstellung

Schritt	Lösung
Test wählen	Varianz bekannt, deshalb z-Test
Nullhypothese	$H_0 : \mu = \mu_0$
Alternativhypothese	$H_1 : \mu < \mu_0$
Signifikanzniveau	$\alpha = 0{,}05$
Ablehnungsbereich	$z \leq z_\alpha$
Ablehnungsbereich	$z \leq z_{5\%}$
Ablehnungsbereich	$z \leq -1{,}65 \quad \text{oder} -1{,}65$
Mittel: Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Mittel: Einsetzen	$\bar{x} = \frac{52{,}28}{5}$
Mittel: Ergebnis	$\bar{x} = 10{,}46$
Standardabweichung	$\sigma \approx 2{,}28$
Prüfgröße: Formel	$z = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\sigma}$
Prüfgröße: Einsetzen	$z = \sqrt{5} \cdot \frac{10{,}46 - 11{,}8}{2{,}28}$
Prüfgröße: Ergebnis	$z \approx -1{,}31$
Interpretieren	Der Ablehnungsbereich wurde nicht erreicht.
Interpretieren	Die Nullhypothese muss beibehalten werden.
Interpretieren	Die Behauptung, im Neubaugebiet seien die Mietpreise günstiger, konnte nicht bestätigt w

Lösung 5-7

zur Aufgabenstellung

Schritt	Lösung
Test wählen	Der Mittelwert einer Stichprobe soll auf signifikante Abweichung von der Grur
Nullhypothese	$H_0 : \mu = \mu_0$
Alternativhypothese	$H_0 : \mu \neq \mu_0$
Signifikanzniveau	$\alpha = 0{,}05$
Freiheitsgrade	$df = n - 1 = 6 - 1 = 5$
Ablehnungsbereich: Formel	$t \leq t_{df;\alpha/2} \quad \text{oder} \quad t \geq t_{df;(1-\alpha/2)}$
Ablehnungsbereich: Einsetzen	$t \leq t_{5;2{,}5\%} \quad \text{oder} \quad t \geq t_{5;97{,}5\%}$
Ablehnungsbereich: Ergebnis	$t \leq -2{,}571 \quad \text{oder} \quad t \geq 2{,}571$
Mittel: Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Mittel: Einsetzen	$\bar{x} = \frac{469{,}00}{6}$
Mittel: Ergebnis	$\bar{x} = 78{,}17$
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Varianz: Einsetzen	$s^2 = \frac{90{,}83}{5}$
Varianz: Ergebnis	$s^2 = 18{,}17$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung: Einsetzen	$s = \sqrt{18{,}17}$
Standardabweichung: Ergebnis	$s \approx 4{,}26$
Prüfgröße: Formel	$t = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s}$
Prüfgröße: Einsetzen	$t = \sqrt{6} \cdot \frac{78{,}17 - 73}{4{,}26}$
Prüfgröße: Ergebnis	$t \approx 2{,}97$
Interpretieren: Ablehnungsbereich	Der Ablehnungsbereich wurde erreicht.
Interpretieren: Hypothese	Die Nullhypothese wird abgelehnt.
Interpretieren: Inhalt	Der Ertrag mit dem neuen Düngemittel ist signifikant höher als ohne ($\alpha = 0{,}0$

Sitzung 6

Lösung 6-1

zur Aufgabenstellung

a) 0,13
b) 3,23
c) 3,14
d) 6,00
e) 3,29
f) 0,20
g) 0,05
h) 0,72
i) 4,35
j) 1,78

Lösung 6-2

zur Aufgabenstellung

a) Es geht um den Vergleich der Varianzen von zwei Stichproben, deshalb F-Test.

b) Ungerichtete Alternativhypothese:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

c) Ein Signifikanzniveau von $\alpha = 0{,}1$ bedeutet, dass wir die Nullhypothese genau dann verwerfen, wenn das empirische Ergebnis unter Annahme der Nullhypothese eine Wahrscheinlichkeit von 10% oder weniger hat.

d) Bei ungerichteter Hypothese:

$$F \leq F_{4;6;5\%} \quad \text{und} \quad F \geq F_{4;6;95\%}$$
$$F \leq 0{,}16 \quad \text{und} \quad F \geq 4{,}53$$

Lösung 6-3

zur Aufgabenstellung

a) Die Varianzen lauten:

$$s_1^2 = 1{,}967$$
$$s_2^2 \approx 0{,}123$$

F berechnet sich durch:

$$\begin{aligned} F &= \frac{s_1^2}{s_2^2} \\ &\approx \frac{1{,}967}{0{,}123} \\ &\approx 15{,}992 \end{aligned}$$

b) Der kritische Wert wurde deutlich übertroffen. Ein Unterschied in der Streuung der Wassertemperaturen konnte nachgewiesen werden ($\alpha = 0{,}1$).

Lösung 6-4

zur Aufgabenstellung

a) Es geht um den Vergleich von Mittelwerten von zwei Stichproben, also ist der 2-Stichproben-t-Test angedacht.

Die Normalverteilung des Merkmals „durchschnittliche Antwortzeit“ ist nicht gesichert, (aber auch nicht ganz abwegig).

Ein weiteres Problem stellt die Bedingung der reinen Zufallsstichprobe dar, was hier allerdings auch nur sehr schwer zu konstruieren wäre (also zufällig ausgewählte Proband*innen aus *allen* WhatsApp-Nutzer*innen im relevanten Alter).

Schließlich ist die Voraussetzung $\sigma_1^2 = \sigma_2^2$ nicht unbedingt gegeben. Bei sehr unterschiedlichen Varianzen der Stichproben sollte daher der Test abgebrochen werden.

b) Wenn Nutzer*innen ohne Benachtigungsfunktion die Population x_1 darstellen und jene mit x_2, dann lauten die Hypothesen:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 > \mu_2 \end{aligned}$$

c) Freiheitsgrade:

$$\begin{aligned} fg &= 2 \cdot n - 2 \\ &= 2 \cdot 6 - 2 \\ &= 10 \end{aligned}$$

Kritischer Wert:

$$\begin{aligned} t &\geq t_{10;95\%} \\ t &\geq 1{,}812 \end{aligned}$$

Lösung 6-5

zur Aufgabenstellung

a) Zunächst die Mittelwerte und Varianzen:

$$\begin{aligned} \bar{x}_1 &= 27{,}05 \quad s_1^2 \approx 165{,}16 \\ \bar{x}_2 &\approx 22{,}18 \quad s_2^2 \approx 107{,}77 \end{aligned}$$

Dabei fällt auf, dass die Varianzen gar nicht so unterschiedlich sind (was ja beim 2-Stichproben-t-Test vorausgesetzt ist. In der Praxis sollte dies aber noch mit einem F-Test abgesichert werden.

b) Durch Einsetzen in die Formel für t ergibt sich:

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \\ &\approx \frac{27{,}05 - 22{,}18}{\sqrt{\frac{165{,}16 + 107{,}77}{6}}} \\ &\approx 0{,}722 \end{aligned}$$

c) Der kritische Wert (1,812) wurde nicht überschritten. Die Nullhypothese muss beibehalten werden. Dass jugendliche Nutzer*innen mit Benachrichtigungsfunktion schneller antworten, konnte in dieser Untersuchung nicht belegt werden ($\alpha = 0{,}05$).

Lösung 6-6

zur Aufgabenstellung

Schritt	Lösung
Test wählen	Zwei Stichproben sollen auf einen signifkanten Unterschied in der Varianz überpr
Nullhypothese	$H_0 : \sigma_1^2 = \sigma_2^2$
Alternativhypothese	$H_1 : \sigma_1^2 \neq \sigma_2^2$
Signifikanzniveau entscheiden	$\alpha = 0{,}1$
Freiheitsgrade 1	$df_1 = 6 - 1 = 5$
Freiheitsgrade 2	$df_2 = 6 - 1 = 5$
Ablehnungsbereich: Formel	$F \leq F_{df_1;df_2;\alpha/2} \quad \text{oder} \quad F \geq F_{df_1;df_2;(1-\alpha/2)}$
Ablehnungsbereich: Einsetzen	$F \leq F_{5;5;5\%} \quad \text{oder} \quad F \geq F_{5;5;95\%}$
Ablehnungsbereich: Ergebnis	$F \leq 0{,}20 \quad \text{oder} \quad F \geq 5{,}05$
Varianzen: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Varianzen: Esselrode	$s_1^2 = \frac{40{,}00}{5}$
Varianzen: Esselrode	$s_1^2 = 8{,}00$
Varianzen: Albwald	$s_2^2 = \frac{55{,}50}{5}$
Varianzen: Albwald	$s_2^2 = 11{,}10$
Prüfgröße: Formel	$F = \frac{s_1^2}{s_2^2}$
Prüfgröße: Einsetzen	$F = \frac{8}{11{,}1}$
Prüfgröße: Ergebnis	$F = 0{,}72$
Interpretieren	Der Ablehnungsbereich wurde nicht erreicht.
Interpretieren	Die Nullhypothese wird beibehalten.
Interpretieren	Die Vermutung konnte nicht bestätigt werden: Die Storchpopulationen weichen n

Lösung 6-7

zur Aufgabenstellung

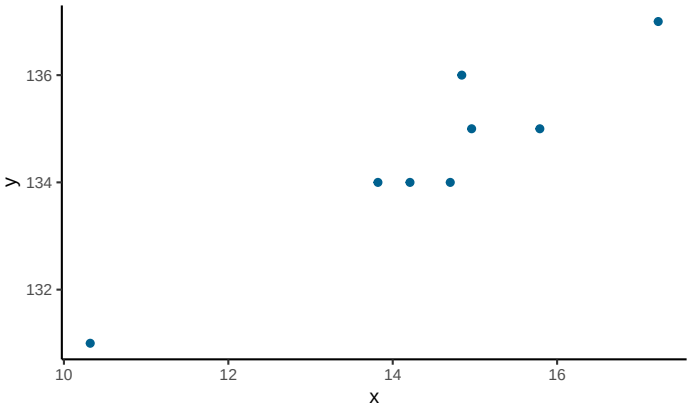
Schritt	Musterlösung
Test wählen	2-Stichprobem-t-Test
Nullhypothese	$H_0 : \mu_1 = \mu_2$
Alternativhypothese	$H_1 : \mu_1 \neq \mu_2$
Signifikanzniveau entscheiden	$\alpha = 0{,}01$
Freiheitsgrad: Formel	$df = 2 \cdot n - 2$
Freiheitsgrad: Ergebnis	$df = 16$
Ablehnungsbereich: Formel	$t \leq t_{df;\alpha/2} \quad \text{und} \quad t \geq t_{df;(1-\alpha/2)}$
Ablehnungsbereich	$t \leq t_{16;0{,}5\%} \quad \text{und} \quad t \geq t_{16;99{,}5\%}$
Ablehnungsbereich	$t \leq -2{,}921 \quad \text{und} \quad t \geq 2{,}921$
Mittelwert: Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Mittelwert 1: Ergebnis	$\bar{x} = 1186{,}00$
Mittelwert 2: Ergebnis	$\bar{y} = 1337{,}56$
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Varianz 1: Ergebnis	$s_x^2 = 2919{,}25$
Varianz 2: Ergebnis	$s_y^2 = 53274{,}52$
Prüfgröße: Formel	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$
Prüfgröße: Einsetzen	$t = \frac{1186{,}00 - 1337{,}56}{\sqrt{\frac{2919{,}25 + 53274{,}52}{9}}}$
Prüfgröße: Ergebnis	$t = -1{,}92$
Interpretieren	Der Ablehungsbereich wurde nicht erreicht.
Interpretieren	Die Nullhypothese muss beibehalten werden.
Interpretieren	Die Abrufzahlen zwischen Hessen und Niedersachsen unterscheiden sich nicht sig

Sitzung 7

Lösung 7-1

zur Aufgabenstellung

a) Streudiagramm:



Berechnungstabelle:

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	14,21	134	-0,27	-0,5	0,14	0,07	0,25
2	10,32	131	-4,16	-3,5	14,56	17,31	12,25
3	13,82	134	-0,66	-0,5	0,33	0,44	0,25
4	15,79	135	1,31	0,5	0,66	1,72	0,25
5	14,7	134	0,22	-0,5	-0,11	0,05	0,25
6	17,23	137	2,75	2,5	6,88	7,56	6,25
7	14,84	136	0,36	1,5	0,54	0,13	2,25
8	14,96	135	0,48	0,5	0,24	0,23	0,25
Summe:	**115,87**	**1076**			**23,24**	**27,51**	**22**

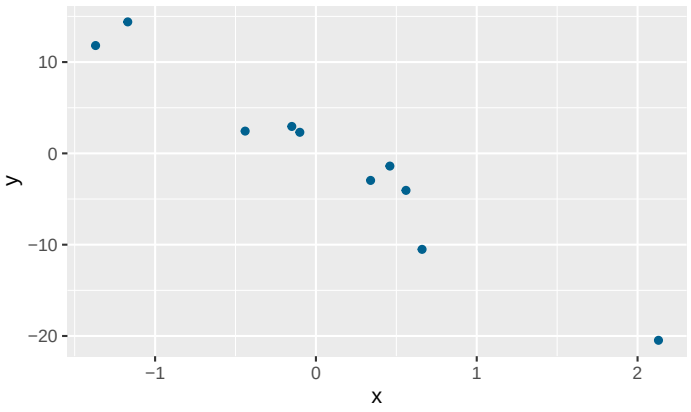
Kovarianz:

$$\bar{x} \approx 14{,}48$$
$$\bar{y} = 134{,}5$$
$$s_{xy} \approx 3{,}32$$

Korrelationskoeffizient:

$$s_x \approx 1{,}98$$
$$s_y \approx 1{,}77$$
$$r \approx 0{,}95$$

b) Streudiagramm:



Berechnungstabelle:

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	-1,17	14,4	-1,26	14,95	-18,84	1,59	223,5
2	-0,1	2,31	-0,19	2,86	-0,54	0,04	8,18
3	-0,15	2,95	-0,24	3,5	-0,84	0,06	12,25
4	0,46	-1,39	0,37	-0,84	-0,31	0,14	0,71
5	0,34	-2,96	0,25	-2,41	-0,6	0,06	5,81
6	-0,44	2,44	-0,53	2,99	-1,58	0,28	8,94
7	2,13	-20,47	2,04	-19,92	-40,64	4,16	396,81
8	0,66	-10,51	0,57	-9,96	-5,68	0,32	99,2
9	-1,37	11,81	-1,46	12,36	-18,05	2,13	152,77
10	0,56	-4,05	0,47	-3,5	-1,65	0,22	12,25
Summe:	**0,92**	**-5,47**			**-88,73**	**9**	**920,42**

Kovarianz:

$$\bar{x} \approx 0{,}09$$
$$\bar{y} \approx -0{,}55$$
$$s_{xy} \approx -9{,}86$$

Korrelationskoeffizient:

$$s_x \approx 1{,}00$$
$$s_y \approx 10{,}11$$
$$r \approx -0{,}98$$

Lösung 7-2

zur Aufgabenstellung

Berechnungstabelle:

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	35	394,61	-98,67	19,36	-1910,25	9735,77	374,81
2	79	468,92	-54,67	93,67	-5120,94	2988,81	8774,07
3	234	385,75	100,33	10,5	1053,46	10066,11	110,25
4	105	376,17	-28,67	0,92	-26,38	821,97	0,85
5	318	283,26	184,33	-91,99	-16956,52	33977,55	8462,16
6	31	342,77	-102,67	-32,48	3334,72	10541,13	1054,95
Summe:	**802**	**2251,48**			**-19625,91**	**68131,34**	**18777,09**

Kovarianz:

$$\bar{x} \approx 133{,}67$$
$$\bar{y} \approx 375{,}25$$
$$s_{xy} \approx -3925{,}18$$

Korrelationskoeffizient:

$$s_x \approx 116{,}73$$
$$s_y \approx 61{,}28$$
$$r \approx -0{,}55$$

Es lässt sich eine schwache (bis mäßige) negative Korrelation zwischen Entfernung und Umsatz feststellen.

Lösung 7-3

zur Aufgabenstellung

Berechnung Mittelwerte

Schritt	Lösung
Arithm. Mittel: Formel	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Arithm. Mittel x: Einsetzen	$\bar{x} = \frac{2718{,}00}{6}$
Arithm. Mittel x: Ergebnis	$\bar{x} = 453{,}00$
Arithm. Mittel y: Einsetzen	$\bar{y} = \frac{1541{,}00}{6}$
Arithm. Mittel y: Ergebnis	$\bar{y} = 256{,}83$

Tabelle

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
456	264	3	7,17	21,51
628	306	175	49,17	8604,75
497	348	44	91,17	4011,48
275	202	-178	-54,83	9759,74
549	322	96	65,17	6256,32
313	99	-140	-157,83	22096,20

Berechnung

Schritt	Lösung
Kovarianz: Formel	$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$
Kovarianz: Einsetzen	$s_{xy} = \frac{50750}{5}$
Kovarianz: Ergebnis	$s_{xy} = 10150{,}00$
Varianz: Formel	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Varianz x	$s_x^2 = 18614{,}00$
Varianz y	$s_y^2 = 8588{,}97$
Standardabweichung: Formel	$s = \sqrt{s^2}$
Standardabweichung x	$s \approx 136{,}43$
Standardabweichung y	$s_y \approx 92{,}68$
Korrelationskoeff.: Formel	$r = \frac{s_{xy}}{s_x \cdot s_y}$
Korrelationskoeff.: Einsetzen	$r = \frac{10150{,}00}{136{,}43 \cdot 92{,}68}$
Korrelationskoeff.: Ergebnis	$r = 0{,}80$
Antwortsatz	Es lässt sich eine mäßig starke positive Korrelation zwischen Fläche und Kosten fes

Lösung 7-4

zur Aufgabenstellung

Tabelle

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1141	30	220	-11,83	-2602,60
850	49	-71	7,17	-509,07
862	40	-59	-1,83	107,97
1000	39	79	-2,83	-223,57
783	51	-138	9,17	-1265,46
890	42	-31	0,17	-5,27

Berechnung

Schritt	Lösung
Kovarianz: Formel	$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$
Kovarianz: Einsetzen	$s_{xy} = \frac{-4498}{5}$
Kovarianz: Ergebnis	$s_{xy} = -899{,}60$
Korr.koeff.: Formel	$r = \frac{s_{xy}}{s_x \cdot s_y}$
Korr.koeff.: Einsetzen	$r = \frac{-899{,}60}{128{,}97 \cdot 7{,}57}$
Korr.koeff.: Ergebnis	$r = -0{,}92$
Interpretieren	Mit dem Korrelationskoeffizienten $r \approx -0{,}92$ konnte eine starke negative Korrelation festge

Lösung 7-5

zur Aufgabenstellung

a)

$$
\begin{aligned}
r &= \frac{s_{xy}}{s_x \cdot s_y} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y} \\
&= \frac{\sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y}}{n-1} \\
&= \frac{\sum_{i=1}^{n} z_{xi} \cdot z_{yi}}{n-1}
\end{aligned}
$$

b) hier nicht ausgeführt

c) nachzulesen bei Bortz und Schuster (2010: 157)

d) hier nicht ausgeführt

Sitzung 8

Lösung 8-1

zur Aufgabenstellung

a) Berechnung durch die Formel

$$\hat{y}_i = -1{,}48 - 0{,}975 \cdot x_i$$

ergibt folgende Werte für $\hat{y}_i$:

$$-1{,}77 \quad 16{,}56 \quad 11{,}68 \quad 15{,}29 \quad -30{,}54 \quad -26{,}44 \quad 34{,}01 \quad 24{,}07$$

b) Umformen der Regressionsgleichung ergibt:

$$x_i = \frac{\hat{y}_i + 1{,}48}{-0{,}975}$$

Einsetzen ergibt die Werte:

$$8{,}74 \quad -16{,}9 \quad 49{,}76 \quad 8{,}74 \quad 60{,}02 \quad 54{,}89 \quad 18{,}99 \quad -1{,}52$$

c) Tabellarische Berechnung:

x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
-11,49	6,82	9,72	-2,90
8,22	-8,59	-9,49	0,90
-25,66	25,92	23,54	2,38
23,81	-26,91	-24,69	-2,22
-3,14	4,41	1,58	2,83
-1,52	-3,39	0,00	-3,39
20,15	-19,89	-21,13	1,24
-10,22	9,30	8,48	0,82

Lösung 8-2

zur Aufgabenstellung

a) Schritt 1: Steigung b

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ &= \frac{869{,}83}{1080{,}94} \\ &\approx 0{,}81 \end{aligned}$$

Schritt 2: Achsenabschnitt a

$$\begin{aligned} a &= \bar{y} - b \cdot \bar{x} \\ &\approx 156{,}7 - 0{,}805 \cdot 157{,}5 \\ &\approx 29{,}91 \end{aligned}$$

Schritt 3: Regressionsgleichung

$$\begin{aligned} y &= a + b \cdot x \\ y &\approx 29{,}91 + 0{,}805 \cdot x \end{aligned}$$

b) Schritt 1: Bestimmung r (s. Sitzung 7)

$$\begin{aligned} r &= \frac{s_{xy}}{s_x \cdot s_y} \\ &\approx \frac{869{,}83}{\sqrt{1080{,}94} \cdot \sqrt{884{,}46}} \\ &\approx 0{,}89 \end{aligned}$$

Schritt 2: Bestimmung R^2

$$\begin{aligned} R^2 &= r^2 \quad \text{(für lineare Regression)} \\ &\approx 0{,}89^2 \\ &\approx 0{,}79 \end{aligned}$$

Lösung 8-3

zur Aufgabenstellung

(Fortführung von Lösung 7-3)

a) Welche Gleichung beschreibt ein geeignetes lineares Regressionsmodell?

Schritt	Lösung
Regressionsgleichung: Formel	$y = a + b \cdot x$
Steigung: Formel	$b = \frac{s_{xy}}{s_x^2}$
Steigung: Einsetzen	$b = \frac{10150}{18614}$
Steigung: Ergebnis	$b = 0{,}55$
Achsenabschnitt: Formel	$a = \bar{y} - b \cdot \bar{x}$
Achsenabschnitt: Einsetzen	$a = 256{,}83 - 0{,}545 \cdot 453$
Achsenabschnitt: Ergebnis	$a = 9{,}94$
Regressionsgleichung: Ergebnis	$y \approx 9{,}94 + 0{,}545 \cdot x$

b) Wenn die Nutzfläche für Objekt A 318 m² und für Objekt B 380 m² beträgt, wie hoch können dann jeweils die Kosten für die Sanierung geschätzt werden?

Schritt	Lösung
Formel	$\hat{y} = 9{,}94 + 0{,}545 \cdot x$
Objekt A: Einsetzen	$\hat{y}_A \approx 9{,}94 + 0{,}545 \cdot 318$
Objekt A: Ergebnis	$\hat{y}_A \approx 183{,}25$
Objekt B: Einsetzen	$\hat{y}_B \approx 9{,}94 + 0{,}545 \cdot 380$
Objekt B: Ergebnis	$\hat{y}_B \approx 217{,}04$
Antwortsatz	Die Kosten der Sanierung können auf 183.250 € für Objekt A und auf 217.040 € für Objekt B ge

Lösung 8-4

zur Aufgabenstellung

Schritt	Lösung
Steigung: Formel	$b = \frac{s_{xy}}{s_x^2}$
Steigung: Einsetzen	$b = \frac{-899{,}6}{128{,}97^2}$
Steigung: Ergebnis	$b = -0{,}054$
Achsenabschnitt: Formel	$a = \bar{y} - b \cdot \bar{x}$
Achsenabschnitt: Einsetzen	$a = 41{,}83 + 0{,}054 \cdot 921$
Achsenabschnitt: Ergebnis	$a = 91{,}56$
Regression: Formel	$y = a + b \cdot x$
Regression: Ergebnis	$y \approx 91{,}56 - 0{,}054 \cdot x$

a)

Schritt	Lösung
Formel	$\hat{y} = 91{,}56 - 0{,}054 \cdot x$
Einsetzen	$\hat{y} \approx 91{,}56 - 0{,}054 \cdot 500$
Ergenbis	$\hat{y} \approx 64{,}56$
Antwortsatz	Der Quadradmeterpreis für die Immoblie beträgt laut Modell 64,56.

b)

Lösung 8-5

zur Aufgabenstellung

Es wird eine Regressionsgleichung benötigt. Dazu müssen zunächst einige Kennwerte der Verteilung berechnet werden:

$$\begin{aligned} \bar{x} &\approx 534{,}33 \\ s_x^2 &= 114919{,}9 \\ \bar{y} &= 70 \\ s_y^2 &= 620 \\ s_{xy} &= 7952{,}8 \end{aligned}$$

Dann lauten Regressionskoeffizienten und -gleichung:

$$\begin{aligned} b &\approx 0{,}0692 \\ a &\approx 33{,}13 \\ y &\approx 33{,}13 + 0{,}0692 \cdot x \end{aligned}$$

a)

$$\begin{aligned} \hat{y}_i &= a + b \cdot x \\ &\approx 33{,}13 + 0{,}0692 \cdot (6 \cdot 60) \\ &\approx 58{,}04 \end{aligned}$$

b)

$$\begin{aligned} x_i &= \frac{\hat{y}_i - a}{b} \\ &\approx \frac{50 - 33{,}13}{0{,}0692} \\ &\approx 243{,}79 \end{aligned}$$

c)

$$\begin{aligned} x_i &= \frac{\hat{y}_i - a}{b} \\ &\approx \frac{100 - 33{,}13}{0{,}0692} \\ &\approx 966.33 \end{aligned}$$

d) Gefragt ist nach R^2

$$\begin{aligned} r &\approx 0{,}94 \quad (\text{s. Sitzung 7}) \\ R^2 &\approx 0{,}88 \end{aligned}$$

e) Das Modell kann nur gültig sein für Wertebereiche $x > 0$ und $0 < y < 100$. Darüber hinaus ist eigentlich zu erwarten, dass in der ersten Stunde Vorbereitungszeit die Punktezahl stärker verbessert wird als in der 10. oder 11. Stunde. Diese Abflachung der Kurve ist jedoch im linearen Modell nicht vorgesehen.

Sitzung 9

Lösung 9-1

zur Aufgabenstellung

a) Überführung in Kreuztabelle

	→ Autobesitz		
Wohnort ↓	**Ja**	**Nein**	
Land	9	2	**11**
Stadt	2	7	**9**
	11	**9**	**20**

b) Erwartungswerte

	→ Autobesitz		
Wohnort ↓	**Ja**	**Nein**	
Land	9 (6,05)	2 (4,95)	**11**
Stadt	2 (4,95)	7 (4,05)	**9**
	11	**9**	**20**

c) Teilwerte für χ^2 berechnet durch

$$\frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

z. B. für die Kombination „Autobesitz“ und „Stadt“:

$$\frac{(n_{21} - m_{21})^2}{m_{21}} = \frac{(2 - 4{,}95)^2}{4{,}95} \approx 1{,}758$$

	→ Autobesitz		
Wohnort ↓	**Ja**	**Nein**	
Land	9 (6,05) 1,438	2 (4,95) 1,758	**11**
Stadt	2 (4,95) 1,758	7 (4,05) 2,149	**9**
	11	**9**	**20**

Die Summe der Teilwerte ergibt χ^2:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij} - m_{ij})^2}{m_{ij}} \\ &\approx 1{,}438 + 1{,}758 + 1{,}758 + 2{,}149 \\ &= 7{,}103\end{aligned}$$

d) ϕ-Koeffizient:

$$\begin{aligned}\phi &= \sqrt{\frac{\chi^2}{n}} \\ &\approx \sqrt{\frac{7{,}103}{20}} \\ &\approx 0{,}596\end{aligned}$$

e) Es besteht ein deutlicher Zusammenhang ($\phi = 0{,}596$). Dabei übersteigt die beobachtete Kombination „Land/Auto“ ihren Erwartungswert. Die Wohnumgebung „Land“ korreliert also mit Autobesitz.

Lösung 9-2

zur Aufgabenstellung

a) Vervollständigung der Kreuztabelle:

	→ Frage 2		
Frage 1 ↓	**Ja**	**Nein**	
Ja	5 (10,24)	28 (22,76)	**33**
Nein	40 (34,76)	72 (77,24)	**112**
	45	**100**	**145**

b) Teilwerte χ^2:

	→ Frage 2		
Frage 1 ↓	**Ja**	**Nein**	
Ja	5 (10,24) 2,681	28 (22,76) 1,206	**33**
Nein	40 (34,76) 0,79	72 (77,24) 0,355	**112**
	45	**100**	**145**

Summe ergibt χ^2:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\\ &\approx 2{,}681 + 1{,}206 + 0{,}79 + 0{,}355\\ &= 5{,}032\end{aligned}$$

Berechung ϕ:

$$\begin{aligned}\phi &= \sqrt{\frac{\chi^2}{n}}\\ &\approx \sqrt{\frac{5{,}032}{145}}\\ &\approx 0{,}186\end{aligned}$$

c) Es gibt eine mäßige Korrelation der beiden Antworten ($\phi \approx 0{,}186$). Die Bejahung beider Fragen liegt unter dem Erwartungswert und korreliert also leicht negativ.

Lösung 9-3

zur Aufgabenstellung

Erwartungswerte und Teilwerte für χ^2:

	→ Ergebnis		
Herkunft des Namens ↓	**eingeladen**	**nicht eingeladen**	
deutsch	36	64	**100**
	(19,75)	(80,25)	
	13,37	3,29	
italienisch	23	77	**100**
	(19,75)	(80,25)	
	0,535	0,132	
slawisch	9	91	**100**
	(19,75)	(80,25)	
	5,851	1,44	
türkisch	11	89	**100**
	(19,75)	(80,25)	
	3,877	0,954	
	79	**321**	**400**

Summe der Teilwerte ergibt χ^2:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{\ell}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\\ &\approx 13{,}37 + 3{,}29 + 0{,}535 + 0{,}132 + 5{,}851 + 1{,}44 + 3{,}877 + 0{,}954\\ &= 29{,}449\end{aligned}$$

Cramér-Index (wobei $k = 4$ und $\ell = 2$ und damit $\min(k,\ell) = 2$):

$$\begin{aligned}CI &= \sqrt{\frac{\chi^2}{n\cdot(\min(k,\ell)-1)}}\\ &\approx \sqrt{\frac{29{,}449}{400\cdot(2-1)}}\\ &\approx 0{,}271\end{aligned}$$

Der Cramér-Index weist auf einen leichten Zusammenhang zwischen Namensherkunft und Bewerbungsergebnis hin ($CI \approx 0{,}217$). Dabei lag die Anzahl erfolgreicher Bewerbungen bei Namen deutscher Herkunft deutlich über dem Erwartungswert und bei Namen türkischer und slawischer Herkunft deutlich unter dem Erwartungswert.

Lösung 9-4

zur Aufgabenstellung

Die Kreuztabelle lässt sich mit folgenden Werten vervollständigen:

	→ Wohnverhältnis		
Internetanschluss ↓	**Miete**	**Eigentum**	
Glasfaser	1926	1567	**3493**
	(1922,37)	(1570,63)	
	0,007	0,008	
DSL	2758	3686	**6444**
	(3546,46)	(2897,54)	
	175,293	214,551	
Koaxialkabel	3002	1903	**4905**
	(2699,47)	(2205,53)	
	33,905	41,498	
Kein fester Anschluss	1277	167	**1444**
	(794,71)	(649,29)	
	292,69	358,243	
	8963	**7323**	**16286**

Dann berechnet sich der Cramér-Index wie folgt:

Schritt	**Lösung**
Kontingenzkoeffizent: Formel	$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(n_{ij}-m_{ij})^2}{m_{ij}}$
Kontingenzkoeffizent: Einsetzen	$\chi^2 \approx 0{,}007 + 175{,}293 + 33{,}905 + 292{,}690 + 0{,}008 + 214{,}551 + 41{,}498 + 358{,}2$
Kontingenzkoeffizent: Ergebnis	$\chi^2 \approx 1116{,}195$
Cramér-Index: Formel	$CI = \sqrt{\frac{\chi^2}{n \cdot (\min(k,\ell)-1)}}$
Cramér-Index: Einsetzen	$CI \approx \sqrt{\frac{1116{,}195}{16286 \cdot (2-1)}}$
Cramér-Index: Ergebnis	$CI \approx 0{,}262$
Antwortsatz	Es gibt einen leichten Zusammenhang zwischen Wohnverhältnis und Internetan

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. 2015. 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).

Veit, Susanne. 2020. Feldexperimentelle Forschung zu ethnischer Diskriminierung auf dem Arbeitsmarkt: „Alle sind gleich, aber manche sind gleicher“. In: *Handbuch Stress und Kultur*, hg. von Tobias Ringeisen, Petia Genkova, und Frederick T. L. Leong, 1–22. Wiesbaden: Springer Fachmedien Wiesbaden. doi:10.1007/978-3-658-27825-0_25-1, http://link.springer.com/10.1007/978-3-658-27825-0_25-1 (zugegriffen: 10. Mai 2021).

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