MAT102: Statistische Datenauswertung und -analyse

Martin Staudinger
FERNFH

Version 8.2, WiSe 2025

A judicious man uses statistics, not to get knowledge,
but to save himself from having ignorance foisted upon him.

— Thomas Carlyle (1795 - 1881)

Inhaltsverzeichnis

1 Einleitung

1.1 Worum geht es?

Außenhandel: Stärkste Exportzuwächse in Wien und der Steiermark
Im 1. Halbjahr 2019 erzielten laut vorläufigen Ergebnissen von Statistik Austria sieben Bundesländer sowohl in der Einfuhr als auch in der Ausfuhr höhere Ergebnisse als im Vorjahreszeitraum. Die stärksten absoluten Zuwächse in der Ausfuhr gab es in der Steiermark (+0,94 Mrd. Euro) gefolgt von Wien (+0,75 Mrd. Euro) und Oberösterreich (+0,74 Mrd. Euro); die größten relativen Zuwachsraten in dieser Verkehrsrichtung erzielten ebenfalls Wien (+7,8%) und die Steiermark (+7,6%). Die Ausfuhrwerte von Niederösterreich (-1,8% bzw. -0,21 Mrd. Euro) und Kärnten (-4,3% bzw. -0,17 Mrd. Euro) zeigten einen Rückgang. [...]
Wie im 1. Halbjahr 2018 verbuchten auch im 1. Halbjahr 2019 fünf Bundesländer einen Handelsbilanzüberschuss; das heißt, es wurden mehr Waren von diesen Bundesländern aus- als eingeführt. Das höchste Aktivum entfiel dabei auf Oberösterreich mit 4,77 Mrd. Euro, gefolgt von der Steiermark mit 3,28 Mrd. Euro und Vorarlberg mit 1,27 Mrd. Euro. Das deutlichste Passivum verzeichnete Wien mit 8,76 Mrd. Euro. [...] In den meisten Bundesländern dominierte sowohl ein- als auch ausfuhrseitig der Außenhandel mit Maschinen (Warenkapitel 84, 85 und 87 der Kombinierten Nomenklatur).

Quelle: Statistik Austria, www.statistik.at/web_de/presse/122357.html, abgerufen am 7.1.2020

Statistik besteht manchmal aus vielen Zahlen und nach zwei, drei Zeilen weiß man nicht mehr, was weiter oben gestanden ist. Aber vermutlich haben all diese Zahlen einen Sinn.

Wenn man ein bisschen länger über «Statistik» nachdenkt wird man feststellen, dass sie uns häufiger begegnet als uns bewusst ist. Denk zum Beispiel an den Ski-Weltcup oder die Fußball-Bundesliga. Würden nicht Aufzeichnungen und Auswertungen darüber geführt, mag zwar jedes Rennen oder Match vielleicht für sich spannend sein, aber am Ende der Saison eine Entscheidung darüber, wer denn nun insgesamt der oder die Beste war, unmöglich. Oder wenn wir einschätzen wollen, mit welcher Wahrscheinlichkeit ein neues Medikament einen gesundheitsfördernden Effekt oder ob oder wie lange eine Impfung die gewünschte Wirkung haben wird, oder mit welchen Angeboten Facebook-User beworben werden sollen: Die Bandbreite der Anwendungsfälle für Statistik ist beinahe unbegrenzt. Das Ziel dieser Lehrveranstaltung ist es, sie nachvollziehen und verstehen zu können. Und auch zu erkennen, wo die Grenzen der Statistik sind und welche Schlüsse nicht zulässig sind.

Egal, ob du die heutige Tageszeitung aufschlägst, die Nachrichten im Fernsehen verfolgst, oder auch durch das Internet surfst: Du wirst darin eine Menge Behauptungen, Erklärungen und Schlussfolgerungen vorfinden, die auf der Erhebung, Auswertung und dem Vergleich statistischer Daten beruhen, und auch eine Menge an Graphiken zur Visualisierung von den sich daraus ergebenden Sachverhalten (siehe z.B. Abb.1.1 oder 1.2).

Abb. 1.1: Statistische Diagramme finden sich nicht nur im Internet oder in einschlägigen Statistikbüchern …

Es geht in der Statistik also um *Daten* und die *Analyse von Daten*. Und wir haben heute Zugriff auf eine wirklich riesige Menge von Daten.

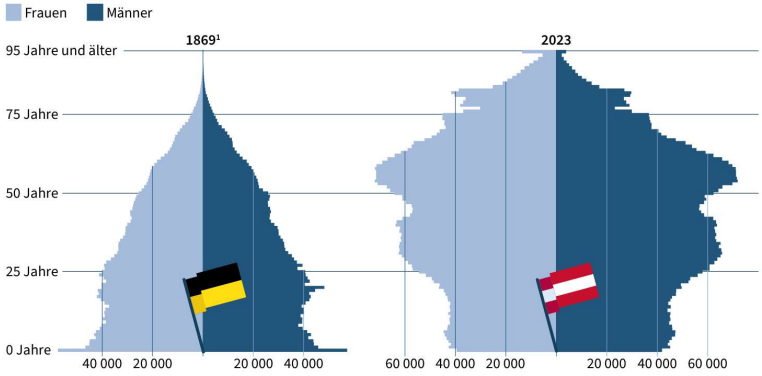


Abb. 1.2: Bevölkerungspyramide Österreichs 1869 und 2023. (Quelle: STATISTIK AUSTRIA: Infografiken 2023)
[1)] Die Bevölkerungszahlen 1869 beinhalten nur Daten aus dem heutigen Gebietsstand des Bundesgebiets

Der 19.11.1999 hatte eine interessante Besonderheit: Es war dies das letzte Datum für eine sehr lange Zeit, das sich nur aus ungeraden Ziffern zusammensetzt. Das nächste Mal wird das erst wieder 1111 Jahre (genauer: 405.827 Tage) später, am 1.1.3111 der Fall sein.
Umgekehrt war der 2.2.2000 seit Langem der erste Tag aus geraden Ziffern (inklusive Null), und zwar seit dem 28.8.888. Vom 29.8.888 bis zum 1.2.2000 befanden sich in jedem Datum ungerade Ziffern. In den Jahren 2000, 2002, 2004, 2006 und 2008 gab es dann ein Datum nur aus geraden Ziffern sehr häufig, genauer: 280 Mal, nämlich an jedem geraden Tag im 2., 4., 6. und 8. Monat. Nach dem 28.8.2008 war wieder eine Weile Pause – bis zum 2.2.2022. Die letzte rein gerade Datumsangaben gab es dann am 28.8.2024, das wird 2026 und 2028 ect. wieder der Fall sein und sich alle 200 Jahre wiederholen. Nach dem 28.8.2888 wird dann wieder für 405.941 Tage kein «gerades» Datum mehr auftreten.

Wir gehen davon aus, dass jede:r eine ungefähre Vorstellung vom Begriff «Daten» hat, und dabei können wir es im Moment auch belassen bzw. auf die Lehrveranstaltung *DAT101 Data and Information Literacy* verweisen. Auch den Begriff Information wollen wir in der dort angegebenen Form verwenden («Daten gewinnen erst an Bedeutung, wenn ich eine konkrete Frage habe und weiß, wie ich aus den Daten eine Antwort auf die Frage finden kann. Erst dann ist es eine *Information* für mich: Eine konkrete Antwort auf eine konkrete Frage. Die Daten selbst sind nur der Rohstoff für Informationen.»)

In der Statistik geht es nun darum, wie wir aus dem Sammeln, Analysieren und Interpretieren von Daten Informationen erhalten können. Wir könnten auch sagen: Wie wir aus Zahlen und Daten «Fakten» herauslesen können.

Ein kleines Rätsel:
85% der Weltbevölkerung, 71% der Menschen in Österreich und 80% ihrer Familie sind Personen, die jünger sind als Emilia.
Was sagt das über die Altersverhältnisse der drei Gruppen Weltbevölkerung, lokale Bevölkerung und Emilias Familie im Verhältnis zueinander aus?
Und wieviele Personen in ihrer Familie sind älter als Emilia?

Datenquelle über Welt- und Österreichische Bevölkerung: population.io.

Ein wichtiger Grundsatz: Keine voreiligen Schlüsse ziehen! Lösung siehe S.??

Im Allgemeinen benutzen wir dabei auch Programme wie *Microsoft Excel* (für Studierende: www.fernfh.ac.at/fernstudium/services-fuer-studierende/office-365), *LibreOffice Calc* (de.libreoffice.org), *JASP* (jasp-stats.org) oder *R* (www.r-project.org) gemeinsam mit *RStudio* (posit.co/downloads). Es geht also im wahrsten Sinn des Wortes um EDV, um *elektronische Datenverarbeitung*.

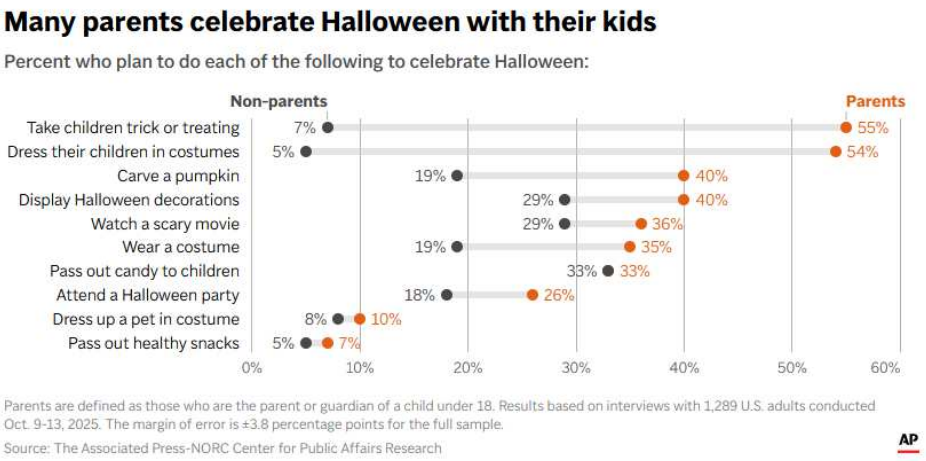


Abb. 1.3: Nicht nur Kinder feiern Halloween. Das Ergebnis einer statistischen Auswertung kann allerdings unterschiedlich ausfallen, je nachdem, ob man einfach alle Antworten zusammenzählt, oder gruppenweise auswertet. (Hier: Sind die Erwachsenen Eltern oder nicht?). Auch das sieht man mit statistischen Methoden.
(Quelle: https://apnews.com/article/halloween-poll-trickortreat-costumes-candy-decorations)

Es gibt fast nichts, was wir nicht statistisch untersuchen können und es gibt eine Menge an Anwendungsfeldern, die mittlerweile eine eigene Unterkategorie der Statistik bilden, darunter zum Beispiel *Wirtschaftsstatistik*, *Betriebsstatistik*, *Ökonometrie*, *Umweltstatistik*, *Biostatistik*, *Geostatistik*, *Sozialstatistik*, *Bevölkerungsstatistik*, *Versicherungsmathematik* (heißt zwar «Mathematik», dahinter steckt aber vor allem Statistik), *Stellarstatistik*, *Statistische Mechanik* etc.

Wir werden in dieser Lehrveranstaltung auch ein Geheimnis der WIBA-Prüfungsordnung lüften. Dort heißt es:
«Mit ausgezeichnetem Erfolg werden Bachelorprüfungen bestanden, wenn die Gesamtbeurteilung eine herausragende Leistung der Kandidatin oder des Kandidaten bescheinigt. Herausragend ist eine Note (gewichtetes Mittel), deren Zahlenwert kleiner oder gleich dem 10%-Quantil der Zahlenwerte der Noten aller Kandidat:innen des Hauptprüfungstermins ist».
Was aber ist ein *gewichtetes Mittel*? Und vor allem: Was ist ein *10%-Quantil*?

Sogar die Studien- und Prüfungsordnung ist ohne Statistik-Kenntnisse nicht lesbar…

1.2 Also was ist jetzt Statistik?

Das Wort «Statistik» selbst kommt aus dem Lateinischen[1] und bedeutet wörtlich übersetzt «(Zu-)Stand, Verfassung, Beschaffenheit». Heute können wir den Begriff in zwei Bedeutungen verwenden:

Einerseits bezeichnet «Statistik» die aus einer Bestandsaufnahme hervorgehende *Datensammlung*. Zum Beispiel werden Daten über die wirtschaftlichen, demografischen, sozialen, ökologischen und kulturellen Gegebenheiten eines Landes gesammelt und veröffentlicht[2].

Im Bundesstatistikgesetz (Ja, das gibt es!) ist der Begriff so definiert: Eine Statistik ist «*die Quantitative Beschreibung und Beurteilung von Massenerscheinungen*»[3].

Man könnte auch sagen: «*In der Statistik geht es um die quantitative Beschreibung von uns und unserer Welt*», wobei «uns» nicht bedeutet, dass es um einzelne Individuen sondern in der Regel um eine Gruppe von Individuen geht, siehe z.B. Abb.1.3.

Aber auch die Gesamtheit der *Methoden*, mit denen die Daten gesammelt und dann die Datensammlungen ausgewertet, analysiert, zusammengefasst, interpretiert, dargestellt und weiterverarbeitet werden, sodass daraus eine Information abgeleitet werden kann, wird als «Statistik» bezeichnet.

In der vorliegenden Lehrveranstaltung werden wir *Statistik* vor allem in diesem Sinn verstehen.

[1] lat. *status*. Ursprünglich ging es dabei vor allem um die Beschreibung und Darstellung geografischer, wirtschaftlicher und politischer Zustände eines «Gemeinwesens», also des Staates.
[2] Für Daten über Österreich siehe z.B. www.statistik.at, für Deutschland www.destatis.de oder für europäische Daten ec.europa.eu/eurostat/de.
[3] §3 Z 1 Bundesstatistikgesetz 2000. BGBl. I Nr. 163/1999

Die Anwendung statistischer Methoden hat dabei die Ziele,

- Daten transparent zu machen,
- die zugrunde liegende Struktur zu finden,
- wichtige Variablen und Kennzahlen dieser Struktur anzugeben,
- Anomalien und Ausreißer herauszufinden,
- Schlüsse aus den Daten zu ziehen und
- diese Schlüsse auch auf ihre Plausibilität hin zu überprüfen.

Letztlich möchten wir mit der Statistik

- ein *Modell* finden, mit dem wir besser verstehen können, wie oder warum bestimmte Phänomene in der *realen* Welt funktionieren,
- ausgehend von real beobachteten Daten für dieses Modell wichtige Faktoren, Kennzahlen und Parameter bestimmen,
- damit wir es aus der *Modellwelt* wieder zurück in die *Realwelt* holen und hier anwenden können und
- dabei auch auf zukünftige Ereignisse schließen oder Neues entdecken können.

«Modelle» gibt es ja in vielen Bereichen. Zum Beispiel kennen wir in der Physik das Newtonsches Gravitationsgesetz[4], das wir als Modell für viele Anwendungen verwenden können. Interessanterweise müssen wir gar nicht wissen, *warum* es so etwas wie Gravitation gibt oder wodurch sie verursacht wird[5]. Wir können das Modell dennoch verwenden, um zum Beispiel Satelliten in einer geostationären Umlaufbahn zu halten und damit Fernsehprogramme übertragen, um miteinander zu kommunizieren oder um das Wetter zu beobachten. Auch in den Wirtschaftswissenschaften bilden wir Modelle (genannt *ökonometrische* Modelle), in den Sozial- und Humanwissenschaften arbeiten wir mit Erklärungs- und Vorhersagemodellen … die Liste ließe sich noch lange fortsetzen: Wenn wir eine Abmagerungskur versuchen oder uns gegen die Auswirkungen eines Virus impfen lassen, gehen wir von einer bestimmten Modellvorstellung aus, nämlich darüber, wie unser Körper funktioniert und auf bestimmte «Eingangsgrößen» reagiert. Jeder der schon einmal versucht hat abzunehmen weiß aber auch, dass es offenbar gar nicht so einfach ist, ein zuverlässiges Modell dafür zu finden. Was für die Umlaufbahn eines Satelliten vielleicht noch gut funktioniert, wird immer schwieriger, je mehr «Mensch» dahintersteckt – und auch: je mehr der *Zufall* dabei zum Zug kommt und je heterogener das Phänomen ist, das wir modellieren wollen.

[4] OK, selbst wenn du es jetzt nicht wirklich im wörtlichen Sinn «kennst», weißt du sicher, dass es so was wie ein Gravitationsgesetz gibt…

[5] Auch Newton wusste noch nicht alles darüber (Wie wir zum Beispiel in www.spektrum.de/lexikon/astronomie/gravitation/150 nachlesen können, was übrigens im Schnitt 24 Minuten dauern wird).

Letztlich geht es immer um dasselbe Ziel: Aus der Beobachtung der Welt (oder eines Ausschnittes daraus) wollen wir ein Verständnis dafür bekommen, wie sie funktioniert. Das soll letztlich dazu führen, dass wir fähig sind, Ergebnisse zukünftiger Beobachtungen vorherzusagen. Am Ende können wir die Beobachtung weglassen und dennoch das Verhalten (eines physikalischen oder technischen Prozesses, des Verlaufs einer wirtschaftlichen Entwicklung, den Verlauf einer Epidemie abhängig vom Verhalten der Menschen, …) vorhersagen – zumindest für die Mehrzahl der Fälle.

Dabei geht es bei den Modellen der Statistik immer auch darum, dabei die *Variabilität* der Daten mit zu berücksichtigen. Während beispielsweise die Biologie davon ausgeht, dass der Mensch ein Zweibeiner ist, müssen wir aus Sicht der Statistik feststellen, dass das zwar eine gewisse Idealvorstellung ist, statistisch gesehen die Menschen im Mittel aber weniger als zwei Beine haben.

Methodisch werden wir uns mit mehreren Teilbereichen der Statistik beschäftigen: Den Methoden der *beschreibenden Statistik*, der *explorativen Statistik* und jenen der *schließenden Statistik*[6].

Die **beschreibende Statistik** hat zum Ziel, aus umfangreichen, unübersichtlichen und komplizierten Datensätzen Informationen zu generieren. Dabei bedienen wir uns numerischer und grafischer Methoden, mit denen wir die Daten zusammenfassen und möglichst anschaulich darstellen wollen. Es geht um Fragen nach den auftretenden Häufigkeiten und der Verteilung der Daten bzw. um Kenngrößen dieser Verteilungen. Dazu benutzen wir Tabellen, Diagramme und aus den Daten (mathematisch) abgeleitete Größen, die als Repräsentanten für die jeweilige Datenmenge und eine bestimmte Fragestellung dienen können. Manchmal besteht die Anwendung der beschreibenden Statistik auch ganz einfach darin, einen Sachverhalt im wahrsten Sinn des Wortes zu «beschreiben». Dabei handelt es sich immer um einen Blick zurück und wir beschreiben, was in der Vergangenheit «passiert» ist oder «war», hin und wieder vielleicht auch ein Beschreibung von Dingen, die jetzt gerade passieren. Letztlich liegt aber dann, wenn wir die Daten analysiert haben, alles bereits in der Vergangenheit.

Die **explorative Statistik** hat zum Ziel, in den beschriebenen Daten Strukturen und Zusammenhänge zu finden und daraus Schlüsse zu ziehen und Hypothesen zu generieren. Diese auf so genannten «Stichproben»[7] beruhenden Hypothesen können dann im Rahmen der **schließenden Statistik** mittels wahrscheinlichkeitstheoretischer Methoden und Testverfahren auf ihre – statistische – All-

[6] Es gibt noch mehr Methoden als diese drei, zum Beispiel die Anwendung verschiedener statistischer Methoden auf sehr große Datenbestände, das so genannte *Data mining*. Aber in diesem Einführungskurs reichen uns die beschreibende, die explorative und die schließende Statistik völlig. Und auch daraus nur Ausschnitte…

[7] Siehe Seite 20

Der Tiergarten Schönbrunn beherbergte im Februar 2025 6.043 Tiere aus 518 verschiedenen Arten und Haustierrassen. Die dem *Institutional Collection Plan* (ICP) folgende Tierbestandsliste weist konkret 393 Wirbeltierarten (Säugetiere, Vögel, Reptilien, Amphibien und Fische) und 125 Wirbellose (Quallen, Insekte, Korallen, etc.) aus. Es wurde auch jedes einzelne individuelle Tier gezählt und festgestellt, dass insgesamt 6043 Tiere in Schönbrunn leben. Die größte Anzahl an Individuen gibt es unter den Fischen (2607), die kleinste Gruppe bilden die 568 Säugetiere, weiters 645 Vögel, 704 Reptilien, 625 Amphibien und 894 einzelne wirbellose Tiere.

Datenquelle: https://www.zoovienna.at/de/news/inventur-abgeschlossen/. 06.03.2025

Beschreibende Statistik hat mit Zahlen und Zählen und dem Angeben von *Häufigkeiten* zu tun

gemeingültigkeit untersucht werden. Ziel ist es also, aus einigen wenigen Daten, die uns zur Verfügung stehen, auf «das große Ganze» zu schließen.

Mit den Ergebnissen der Methoden der beschreibenden, explorativen und schließenden Statistik wollen wir letztlich *Diagnosen* erstellen, *Vorhersagen* treffen und daraus *präskriptive* Vorgaben für bestimmte Verhaltensweisen vorschlagen.

1.3 Einige Begriffe

In der Statistik hat sich – wie in anderen Wissensgebieten – im Laufe der Zeit eine eigene Fachbegriffs-Welt herausgebildet. Oft basieren diese Begriffe auf lateinischen oder (alt-)griechischen Wörtern. Die beschreibende Statistik wird zum Beispiel auch als *deskriptive Statistik*[8] bezeichnet, die schließende Statistik als *induktive*[9] oder *analytische Statistik*, manchmal auch als *Inferenzstatistik*[10]. Bei der *explorativen Statistik*[11] haben wir von vornherein gleich das Fremdwort verwendet.

[8] lat. *describere* = beschreiben; auch: ordnen, einteilen
[9] lat. *inducere* = hin(ein)führen
[10] Laut Duden: *aufbereitetes Wissen, das aufgrund von logischen Schlussfolgerungen gewonnen wurde*
[11] lat. *explorare* = erkunden, erforschen, prüfen, untersuchen

Charakterzüge sind nicht so stabil wie oft angenommen Im Laufe des Lebens ruhen wir immer mehr in uns selbst und geben weniger auf die Meinung anderer, so das Ergebnis der internationalen Studie «*A Coordinated Analysis of Big-Five Trait Change Across 14 Longitudinal Studies*». Tendenziell ziehen wir uns aber auch mehr zurück, sind weniger offen für Neues und werden etwas nachlässiger und unorganisierter, und zwar *statistisch signifikant*. Das zeigt der Vergleich von mehreren Langzeitstudien aus Europa.
Demnach verändert sich der Durchschnitt zwar nicht in großen Sprüngen, sondern etwa alle zehn Jahre ein bisschen. Dennoch seien die Entwicklungen keineswegs trivial.
Ein einziges Merkmal verändert sich in keiner der Studien merklich: die Verträglichkeit. Zwar gibt es Menschen, die ein klein wenig verträglicher werden, also etwas mehr Rücksicht auf andere nehmen und empathischer werden. Andere entwickeln sich ein klein wenig ins Gegenteil - keiner aber verändert sich hier laut den Daten signifikant.
Nun wollen die Psychologen den Datensatz noch genauer untersuchen und weitere Muster erkennen, wonach die Persönlichkeit durch bestimmte soziale Rollen sowie durch gesundheitsbewusstes Verhalten beeinflusst wird und z.B. herausfinden, ob sich Menschen mit Familie anders entwickeln oder welche Rolle das Geschlecht spielt.

Quelle: science.orf.at, 09.02.2018

Mit der *schließenden* Statistik versucht man, aus beobachteten Phänomenen allgemeine *Schlüsse zu ziehen*.

Elemente, Merkmale und Variable

Die Objekte, die Gegenstand unserer Beobachtung und Analyse sind, nennen wir **Merkmalsträger**, **Elemente** oder manchmal auch **Individuen**. Merkmalsträger können Personen oder reale «Dinge» sein, aber auch soziale Systeme oder virtuelle Elemente. Die einzelnen Merkmalsträger müssen anhand mindestens eines sachlichen, räumlichen oder zeitlichen Kriteriums eindeutig identifizierbar und abgrenzbar sein. Merkmalsträger sind zum Beispiel die Teilnehmer:innen dieses Kurses: Sie haben etwas gemeinsam (nämlich die Kursteilnahme), unterscheiden sich aber in vielen Dingen, und diese unterschiedlichen Eigenschaften können wir erheben und (statistisch) auswerten und auswerten. Die Eigenschaft, die wir an den Merkmalsträgern untersuchen, ist das **Merkmal** (oft auch genauer *statistisches Merkmal* genannt). Mathematisch gesehen handelt es sich dabei um eine *Variable*, die unterschiedliche Werte annehmen kann.

Definition der Bevölkerungszahl Österreichs:
«Die Statistik des Bevölkerungsstandes für den 1.1.2021 beruht auf den nach bevölkerungsstatistischen Kriterien aufgearbeiteten Daten über Hauptwohnsitzmeldungen in Österreich laut dem Zentralen Melderegister. In den hier präsentierten vorläufigen Ergebnissen sind statistische Bereinigungen auf Basis der für den Finanzausgleich jährlich zu ermittelnden Einwohnerzahl bereits berücksichtigt, nicht jedoch eine Mindestaufenthaltsdauer in Österreich von drei Monaten.»

Quelle: www.statistik.at/web_de/presse/125347.html

Manchmal ist nicht das (Ab-)Zählen per se schwierig, sondern die Definition der Merkmalsträger, die gezählt werden sollen.

In der Statistik bezeichnen wir Merkmale auch als **Zufallsvariable**[12].

Beispiele für Zufallsvariable: Das Alter oder der Beruf der Teilnehmer:innen in diesem Kurs, die Zeit, die deine Uhr gerade jetzt anzeigt, die Körpergröße oder das Geschlecht von Personen.

Jedes Merkmal kann in verschiedenen, konkreten Erscheinungsformen auftreten; wir nennen das auch die **Merkmalsausprägungen**. Die konkreten Werte, die dann tatsächlich auftreten, nennen wir **Merkmalswert** , in der Statistik auch die *Realisierung der Zufallsvariable* – oder auch einfach: **Daten**.

Beispiele für Merkmalsausprägungen: Die Zufallsvariable «Alter» von Personen kann eine Vielzahl von möglichen Werten haben, von 0 bis etwa 120 Jahre. Genau genommen sind eigentlich unendlich viele Ausprägungen möglich. Wir können ja das Alter nicht nur in Jahren angeben, sondern zum Beispiel auch in Tagen, Stunden, Minuten, Sekunden, Zehntelsekunden, ... Für die Zufallsvariable «Körpergröße» gilt dasselbe; auch hier sind theoretisch unendlich viele Werte möglich und nur eine Frage der Messgenauigkeit. Das «Geschlecht» hingegen kann (in Österreich) nur fünf Ausprägungen haben[13].

[12] Und zwar deshalb weil wir davon ausgehen, dass der Wert, den die Variable zu dem Zeitpunkt hat, wenn wir sie messen oder beobachten, mehr oder weniger zufällig ist. «Zufällig» bedeutet dabei: Es gibt eine bestimmte Wahrscheinlichkeit für die konkrete Merkmalsausprägung, aber auf welche wir gerade treffen, ist zufällig.

[13] Nämlich *männlich*, *weiblich*, *divers*, *inter* und *offen*. Es gibt im Personenstandsregister auch die Möglichkeit der Streichung des Geschlechtseintrags. Der Eintrag *keine Angabe* stellt im Sinne der Statistik aber keine sechste Ausprägung dar, sondern ist quasi eine «Leermeldung».

Daten sind oft in Tabellen angeordnet. Die Spaltenüberschrift entsprechen dabei meist den Zufallsvariablen, die untersucht wurden, und in den einzelnen Zeilen stehen die einzelnen Beobachtungen, also die Realisierungen dieser Zufallsvariablen.
In der Informatik würde man auch sagen: Eine Zufallsvariable ist eine **Klasse** bzw. ein **Objekttyp**, die Daten sind **Instanzen** dieses Objekttyps.

Für alle, die sich in der Informatik leichter tun als in der Statistik und unterscheiden müssen, was Zufallsvariable (Merkmal) und was Realisierung (Merkmalsausprägung) ist.

Empirische Daten

«Empirisch»[14] bedeutet: *auf Erfahrung beruhend* oder *etwas aus der Erfahrung kennen*. Empirische Daten sind demnach solche, die man durch mehr oder weniger systematische (= zielgerichtete) Beobachtung erhalten hat, zum Beispiel durch Abzählen oder Messen, Daten aus **Experimenten**[15], aber auch durch **Befragungen** oder **inhaltsanalytische** Verfahren. Die Daten müssen dabei nicht unbedingt durch dieselbe Person erhoben worden sein, die dann die Auswertung macht. Auch wenn ich auf Datenbestände zugreife, die in irgendeiner Form bereits vorliegen (Studienergebnisse, Dokumente und Datensätze im Internet, Datenbanken, etc.) handelt es sich um empirische Daten – sofern sie durch Beobachtung «der Welt» entstanden sind.

In einer ein Jahr andauernden statistischen Untersuchung stellte Anton Bruszkay am Beginn des 20. Jahrhunderts fest, dass die Donau innerhalb eines Jahres an 11 Tagen im Jahr braun, an 46 Tagen lehmgelb, 59 Tage schmutzig-grün, 45 Tage hellgrün, 5 Tage grasgrün, 69 Tage stahlgrün, 46 Tage smaragdgrün und 64 Tage dunkelgrün, *niemals jedoch BLAU* ist. Das ist sie nur im Walzer des Johann Strauß.

Quelle: Der Standard, Spezial CENTROPE, 21.5.2005

Empirische Daten können zum Beispiel aus Beobachtung, Klassifizierung und Abzählen entstehen.

[14] griech. $\varepsilon\mu\pi\varepsilon\iota\rho\omega\varsigma$ – empeiros
[15] Bei «Experimenten» werden gezielt bestimmte Bedingungen geschaffen und im Laufe des Experiments auch verändert.

Kategoriale und numerische Daten

Kategoriale (auch: *qualitative*[16]) Variable beschreiben die Eigenschaften von Merkmalsträgern durch eine wertmäßige Angabe «mit Worten», d.h. wir können die Merkmalsausprägungen nur *benennen* oder *beschreiben*. Die Werte, die kategoriale Variable annehmen können, sind in der Regel beschränkt auf eine endliche Liste von vorgegebenen Kategorien. Kategoriale Merkmale sind zum Beispiel die Heimatgemeinden der Absolvent:innen des WIBA-Studiengangs, das Schmerzempfinden im Zusammenhang mit der Menstruationsblutung[17] oder die Klassifikation von Wissenschaftszweigen nach der Österreichischen Version der *Fields of Science and Technology Classification*[18].

Numerische (auch: *quantitative*[19]) Variable sind solche, deren Werte wir durch Zählen oder Messen erhalten und dann durch eine mengenmäßige Angabe in Form einer *Zahl* angeben. Beispiele für numerische Merkmale: Der Preis für eine Bahnfahrt von Wien nach Saint-Malo am 12.7.2024; die Anzahl von Frauen in Führungspositionen in österreichischen börsennotierten Unternehmen; die Arbeitsstunden, die Mitarbeiter:innen in ihrer Firma im Jahr 2023 geleistet haben oder der Rangplatz, den Österreich im *Global Gender Gap Index Ranking*[20] eingenommen hat.

Manchmal werden auch kategoriale Merkmale durch numerische Werte repräsentiert. Zum Beispiel könnte beim Münzwurf «Kopf» mit 0 und «Zahl» mit 1 codiert werden. Die Verwendung von Zahlencodes macht die qualitative Variable aber nicht zu einer quantitativen! Arithmetische Operationen wie ein «Durchschnitt» machen keinen Sinn. Die Zahlen stehen hier nur aus praktischen Gründen als Platzhalter für die Wörter «Kopf» und «Zahl».

Kategoriale Daten werden durch Abzählen nicht in numerische Daten «verwandelt», sondern: Wenn wir das Vorkommen einer bestimmten qualitativen Merkmalsausprägung abzählen, schaffen wir eine zweite Zufallsvariable: Die «höchste abgeschlossene Ausbildung» ist eine qualitative Zufallsvariable, die «Anzahl von Personen mit Hochschulabschluss» eine andere, und zwar eine numerische.

---

[16] vom lat. *qualitas* = Beschaffenheit
[17] siehe www.sozialministerium.at/Themen/Gesundheit/Frauen--und-Gendergesundheit.html
[18] www.statistik.at/kdb/downloads/pdf/prod/OEFOS_2012_Alphabetikum_A_EN_20231113.pdf
[19] vom lat. *quantitas* = Größe
[20] siehe www.weforum.org/publications/global-gender-gap-report-2025

Bekannt ist uns die Unterteilung in kategoriale und numerische Daten auch aus Programmiersprachen. In *R* verwenden wir zum Beispiel für kategoriale Variablen den Datentyp `character` und für numerische `numeric`; in Python `str` bzw. `int`, `float` oder `complex`.

Stetige und diskrete Daten

Stetige (auch: *kontinuierliche*) Zufallsvariable können – zumindest theoretisch – innerhalb eines (endlichen oder unendlichen) Intervalls jeden Zahlenwert aus $\mathbb{R}$ und somit unendlich viele beliebige Werte annehmen. Wir können auch sagen: Stetige Daten können in unendlich viele Untereinheiten geteilt werden. Zum Beispiel können wir das Alter einer Person in Jahren angeben, aber auch in Monaten, Tagen, Stunden, Sekunden und so fort. (Auch wenn wir das in der Praxis natürlich nicht «unendlich» klein unterteilen, aber möglich wäre es ...).

Diskrete Zufallsvariable haben eine endliche (*abzählbare*) Anzahl von Ausprägungsmöglichkeiten. Es sind solche mit einer «überschaubaren» Menge von möglichen Ergebnissen, es gibt aber keine Zwischenwerte. Zum Beispiel kann man beim Würfeln nur entweder 1, 2, 3, 4, 5 oder 6 würfeln, aber nicht 3,5. Auch die Anzahl an Menschen, die gerade im Raum sind, können wir nicht in Untereinheiten unterteilen.

Die Unterscheidung in *diskret* und *stetig* ist nur bei numerischen Daten von Bedeutung. Kategoriale Daten sind immer diskret. Es ist manchmal auch möglich, ein und dasselbe Phänomen diskret *oder* stetig zu betrachten. Zum Beispiel beschreiben wir einen Regenbogen meist damit, dass er aus den sieben Farben Rot, Orange, Gelb, Grün, Blau, Indigo und Violett besteht, was aber nur daran liegt, dass Isaac Newton diesen Farben Namen gegeben hat, das Lichtspektrum also kategorisiert und diskretisiert hat. Tatsächlich enthält ein Regenbogen das gesamte für das menschliche Auge sichtbare Lichtspektrum zwischen etwa 380 und 780 *nm* Wellenlänge und als elektromagnetische Welle betrachtet ist «Farbe» eine kontinuierliche Größe.

1.4 Das Skalenniveau von Daten

Den Begriff «Skala» kennen wir wahrscheinlich am Ehesten im Zusammenhang mit Temperaturmessungen. Dort geben wir zum Beispiel die Tageshöchsttemperatur auf einer *Celsius-* oder einer *Fahrenheit-Skala* an. *Skala* bedeutet also, mit welcher «Messlatte» wir Daten messen. In der Statistik verwenden wir statt des

Ausdrucks Messlatte den Begriff **Skalenniveau**. Die Idee dazu geht auf Stanley Stevens[21] zurück.

Die Zuordnung der Daten auf das richtige Skalenniveau spielt sowohl bei der Frage, welche mathematischen Operationen mit den Daten überhaupt zulässig sind, als auch bei der Auswahl der richtigen Visualisierung (also: welches Diagramm wir verwenden dürfen, siehe Kap. 2) eine Rolle.

Wir unterscheiden folgende Skalen: *Nominalskala*, *Ordinalskala* und *Metrische Skala*, wobei wir letztere noch in die *Intervallskala* und die *Rationalskala* unterteilen können[22]. Diese bereits 1946 definierten Skalen ergänzen wir heute auch noch um die *Absolutskala*.

Entsprechend der Skala, mit der Merkmale gemessen werden, sprechen wir in weiterer Folge von *Nominaldaten*, *Ordinaldaten* und *metrischen Daten* (siehe auch Abb. 1.4).

Nominaldaten[23] (auch: *Unterschiedsmerkmale*) sind solche, die nur qualitativ über ein «Etikett», einen *Namen*, angegeben werden. Eine «Beobachtung» besteht dann darin, dass der Merkmalsträger einer bestimmten Kategorie zugeordnet wird oder nicht. In der Regel haben die Merkmale nicht-numerische Werte (bestehend aus Begriffen, Buchstaben oder Symbolen), manchmal auch numerische Werte (Ziffern), die aber eigentlich nur als Namen aufgefasst werden dürfen und keine mathematische Bedeutung haben.

Unterschiedsmerkmale besitzen keine mathematische Ordnung (Reihenfolge); zwischen ihnen kann nur ganz allgemein Gleichheit oder Ungleichheit bestehen. Als Vergleichsoperation ist daher nur das Kriterium «gleich» oder «verschieden» möglich.

Beispiele: Das Geschlecht oder der ausgeübte Beruf von Personen, ihre Nationalität, ihr Familienstand, die Rückennummer auf den Trikots von Sportler:innen, die Matrikelnummer von Studierenden, Postleitzahlen[24], Unfallursachen im Straßenverkehr, die Erzeugnisse, die man laut Konfitürenverordnung beim Einkochen von Obst herstellen darf[25] etc.

[21] Stevens, Stanley Smith. «On the Theory of Scales of Measurement.» *Science, New Series* , Vol. 103, No. 2684 (1946): 677-680.

[22] woraus manchmal auch als Merkregel das Akronym *NOIR* abgeleitet wird.

[23] lat. *nomen* = Namen, Benennung

[24] Auch wenn Rückennummern, Matrikelnummern oder Postleitzahlen dem Namen nach vermeintlich *Zahlen* sind, haben sie keine wie immer geartete numerische Bedeutung – es macht zum Beispiel nicht viel Sinn, eine Summe von Postleitzahlen zu bilden oder eine «mittlere Postleitzahl». Sie sind daher «nur» Nominaldaten.

[25] nämlich: Konfitüre, Konfitüre extra, Leichtkonfitüre, Gelee, Gelee extra, Leichtgelee, Marmelade, Leichtmarmelade, Gelee-Marmelade oder Maronenkrem, siehe BGBl. II Nr. 367/2004 idF

Postleit*zahlen* sind trotz ihres Namens keine wirklichen *Zahlen* sondern eigentlich nur *Namen* und es macht nicht Sinn, mit ihnen zu *rechnen*.
Mathematisch ist es natürlich möglich. Zum Beispiel ist der Median aller 2225 österreichischen Postleitzahlen 5431. Das entspricht der Postleitzahl der Salzburger Ortschaften Kuchl, Gasteig, Moos, Georgenberg, Kellau, Weißenbach, Garnei, Jadorf und Unterlangenberg – prädestiniert aber nicht dazu, diese Orte als «durchschnittlich» zu bezeichnen.

Nominalskalierte Daten können einige wenige oder auch eine sehr große Anzahl von Werten annehmen. Auf die Frage nach dem «Wohnort» gibt es zum Beispiel in Österreich 17 221 verschiedene Antwortmöglichkeiten[26] von A(lfang) bis Z(wölfaxing). Wenn es bei Nominaldaten nur genau zwei mögliche Ausprägungen gibt, dann sprechen wir auch von einem **binären** oder *dichotomen*[27] Merkmal. Zum Beispiel können wir bei einem Münzwurf als Ergebnis nur «Kopf» oder «Zahl» erhalten, oder auf Fragen, die auf ein «ja» oder «nein» (oder `TRUE` oder `FALSE`) hinauslaufen, nur zwei mögliche Antworten[28] (zum Beispiel: *Freust du dich schon, wenn dieses Kapitel endlich zu Ende ist?*).

Ordinaldaten[29] (auch: *Rangmerkmale*) sind Merkmale, die hinsichtlich ihrer Größe (Bedeutung, Rang, ...) unterschieden und durch Rangziffern gekennzeichnet werden können. Es sind jetzt nicht nur die Vergleichsoperationen «gleich» und «ungleich», sondern auch «größer» und «kleiner» möglich. Allerdings ist nicht definiert, «wie viel größer» ein größeres Merkmal ist bzw. «wie viel kleiner» ein kleineres. Das heißt: Eine Größer-Kleiner-Relation kann festgestellt werden, die Abstände dazwischen oder gar ein mathematisches «Ins-Verhältnis-zueinander-Setzen» können hingegen nicht sinnvoll interpretiert werden.

Typische Ordinaldaten sind alle «Rangdaten» im wörtlichen Sinn, also zum Beispiel die Platzierungen, die Hanna Aronsson Elfman im Slalom erreicht hat[30]. Andere Ordinaldaten sind zum Beispiel das Kreditrating auf Staatsanleihen, das von Ratingagenturen wie Standard & Poor's, Moody's oder Fitch Ratings vergeben wird, oder Dienstgrade (z.B. beim Roten Kreuz: Kolonnenkommandant - Rettungsrat - Oberrettungsrat etc.) oder Messungen der Einstellung von Personen zu einem bestimmten Thema, zum Beispiel in Interviews oder Fragebögen:

BGBl. II Nr. 265/2009

[26] Zumindest theoretisch. Praktisch sind es nur 17 082, denn 139 Ortschaften in Österreich haben keine:n einzige:n Einwohner:in.

[27] vom griech. *διχότομος* – dichotomos = in zwei Teile gespalten

[28] Wir lassen jetzt einmal außer Acht, dass die Befragten auch «jein» oder gar keine Antwort geben können...

[29] lat. *ordinare* = reihen, ordnen

[30] Falls dich das interessiert: t1p.de/wiba-stat17

Auf die Frage *Empfindest du die Klimaerwärmung als bedrohlich?* kann als Antwort «überhaupt nicht», «ein wenig», …, «sehr» gegeben werden. Auch *Noten* werden auf einer Ordinalskala gemessen: Wer einen 2er auf eine Prüfung erhält, ist sicher besser als jemand mit einem 4er, aber nicht unbedingt doppelt so gut (nur weil 4 das Doppelte von 2 ist). Man kann das Skalenniveau von Noten auch daran erkennen, dass sie auch mit Wörtern angegeben werden können (Sehr gut, Gut, Befriedigend, …), nicht nur mit den Ziffern 1, 2, 3, 4 und 5. Und da macht «Gut ist die Hälfte von Genügend» nicht einmal mehr sprachlich einen Sinn.

Metrische Daten[31] sind quantitative Daten, bei denen nicht nur die Reihenfolge definiert ist, sondern zumindest auch Abstände eindeutig messbar sind. Wir unterteilen die metrische Skala dabei noch weiter in: *Intervallskala*, *Verhältnisskala* und *Absolutskala* (siehe auch Abb. 1.4).

Intervallskalierte Daten: «Klassische» Beispiele für intervallskalierte Daten sind Jahreszahlen oder Temperaturangaben. Letztere können bekanntlich in Celsius oder Fahrenheit angegeben werden. Je nachdem, welche konkrete Temperaturskala wir verwenden, kommen wir mitunter zu mathematisch unterschiedlichen Aussagen. Mathematisch zeichnen sich intervallskalierte Daten dadurch aus, dass sie auf einer Skala gemessen werden, die keinen streng definierten «Anfangspunkt» haben. Zum Beispiel haben die Celsius- und die Fahrenheit unterschiedliche Nullpunkte. Auch unsere Jahreszählung «nach Christi Geburt» basiert auf einer willkürlichen Festlegung eines Nullpunktes; ebenso die Messung von Längengraden «östlich (oder westlich) von Greenwich».

Rationalskalierte Daten[32] (auch: *Verhältnismerkmale*) sind numerische Messdaten, bei denen auch die Abstände dazwischen eine eindeutig definierte Größe sind und daher mathematisch sowohl Differenzen als auch Verhältnisse (also: Quotienten) gebildet werden können. Mit ihnen können wir alle Grundrechnungsoperationen durchführen, die wir kennen: auf Identität vergleichen (so wie das mit Nominaldaten möglich ist), die Daten in eine Reihenfolge bringen (was auch mit Ordinaldaten möglich ist), mit ihnen addieren oder subtrahieren (was auch mit intervallskalierten Daten möglich ist) und Produkte und Verhältnisse ausrechnen (was nur bei rationalskalierten Daten geht).

Beispiele für rationalskalierte Daten: Studierendenzahlen, das Einkommen, die Energiebilanz Österreichs im Jahre 2021 oder die Zeit, die ein Programm braucht, um einen bestimmten Algorithmus zu durchlaufen.

Rationalskalen haben einen festen Nullpunkt, aber eine offene Wahl der Maßeinheit, d.h. es kann noch festgelegt werden, wie weit die «Einheit 1» geht.

[31] lat. *metor* = (ab)messen
[32] lat. *ratio* = Berechnung, auch: Verhältnis

Wien um 337% zu warm
Die Seite *wetter.at* kam am 7.2.2016 zu folgendem Schluss:
In den letzten 30 Tagen war es in Wien um 337% (!) wärmer als üblich. Statt 0.8 *kletterten die Thermometer im Schnitt auf* 3.5 *Grad. Einen noch höheren Duchschnittswert gab es dabei nur noch in Bregenz.*
Auch wenn 3.5 tatsächlich (annähernd) 337% mehr sind als 0.8 stimmt dieser Schluss nicht. Warum?
Nehmen wir an, ein Engländer würde diese Aussage machen wollen. Er würde zunächst einmal alles in die ihm besser bekannte Fahrenheit-Skala umrechnen: $0.8°C = 33.44°F$ und $3.5°C = 38.3°F$.
Und siehe da: Setzt man 38.3 und 33.44 ins Verhältnis, so erhalten wir «nur» noch 15%. Unser Engländer würde also die Temperaturzunahme längst nicht so dramatisch beschreiben wie *wetter.at*.
Der Grund für diese Verwirrung liegt darin, dass Temperatur-Grade (in Celsius und Fahrenheit) «nur» *intervallskalierte* Daten sind. Mathematisches Merkmal einer Intervallskala ist, dass ihr ein natürlicher Nullpunkt fehlt. Und daher sollte man bei intervallskalierten Daten keine prozentualen Zu- oder Abnahmen rechnen oder sie ins Verhältnis setzen («Es ist heute doppelt so warm wie gestern»).

Bei intervallskalierten Daten kann es leicht zu falschen Interpretationen kommen…

Es gibt auch Merkmale, wo nicht nur der Nullpunkt sondern auch die Einheit 1 absolut vorgegeben sind. Wir haben es dann mit einer **Absolutskala** zu tun. Ein Beispiel dafür ist die Angabe von *Häufigkeiten*, also das System, in dem wir üblicherweise *zählen* (z.B. die Anzahl der Personen in diesem Kurs, die Kurt heißen), ein weiteres Beispiel die Angabe von *Wahrscheinlichkeiten* (z.B. die Wahrscheinlichkeit, dass zwei Personen in diesem Kurs am selben Tag Geburtstag haben). Letztere werden auf einer Absolutskala angegeben, die überhaupt nur Werte zwischen 0 und 1 annehmen kann.

Mathematisch werde Daten auf einer Absolutskala wie rationalskalierte Daten behandelt, d.h. man kann sie auch ins Verhältnis zueinander setzen («Dividieren») etc. Für alle anderen Skalen siehe Tab.1.1.

Neben den oben genannten Grundtypen von Skalenniveaus gibt es auch noch Sonder- und «Zwischenformen», wie zum Beispiel die Likert-Skala:

Likert-Skala: Eine *Likert-Skala*[33] wird verwendet, wenn wir (via Umfrage) die Einstellung und Meinung von Person zu einem bestimmten Thema messen möchten. Sie stellt üblicherweise 5 bis 7 Merkmausausprägungen zur Wahl, die zwar

[33] benannt nach *Rensis Likert*, amerikanischer Sozialforscher (1903 - 1981).

Auf einer **Nominalskala** kann ich	**Ordinalskala**	**Intervallskala**	**Rationalskala**
Elemente *identifizieren*	Elemente in eine *Rang-ordnung* bringen	*Differenzen* zwischen Elementen angeben	*Verhältnisse* zwischen Elementen angeben
und folgende Berechnungen und Vergleiche durchführen:			
$= \neq$	$= \neq$ $< >$	$= \neq$ $< >$ $-$	$= \neq$ $< >$ $-$ $\div$

Tabelle 1.1: Skalenabhängige Rechenoperationen. Bei der Auswertung von Daten sind nicht immer alle Rechenoperationen möglich oder sinnvoll interpretierbar. Die Skalenniveaus bauen dabei aufeinander auf: Jede höhere Stufe übernimmt die Eigenschaften und Möglichkeiten der vorherigen und fügt (mindestens eine) neue hinzu und erlaubt so komplexere statistische Analysen.

aus qualitativen Aussagen bestehen, aber numerisch codiert und damit quasi in quantitative Daten transformiert werden. Die Antwortmöglichkeiten reichen «von einem Extrem zum anderen» und können z.B. lauten:
1 = völlige Zustimmung, 2 = teilweise Zustimmung, 3 = unentschiedene Haltung, 4 = teilweise Ablehnung, 5 = völlige Ablehnung.

Dabei muss es nicht immer wie in obigem Beispiel eine ungerade Anzahl von Auswahlitems geben. Eine ungerade Anzahl (üblich sind dann 5 oder 7) hat den Vorteil, dass es einen neutralen Mittelpunkt gibt, auf den sich alle zurückziehen können, die tatsächlich indifferent oder ambivalent bezüglich der Antwort sind. Manchmal wird dieser Vorteil auch als Nachteil gesehen und man will die Befragten dazu «zwingen», sich für die zustimmende oder ablehnende Seite zu entscheiden. Dann kommt eine gerade Likert-Skala zum Einsatz.

Bei der Auswertung werden Daten auf einer Likert-Skala als metrische Daten, angesehen und angenommen, dass bei allen Befragten der «gefühlte Abstand» zwischen völliger und teilweiser Zustimmung (oder Ablehnung) in Etwa gleich groß ist. Sie können dann wie rationalskalierte Daten ausgewertet werden.

(Abb.1.4) fasst noch einmal einige der bisher eingeführten Begriffe zusammen.

Generell sind Merkmale auf einer der in diesem Abschnitt genannten Skalen so genannte **häufbare Merkmale**. Das heißt, wir können alle Merkmale, egal auf welcher der obigen Art sie gemessen oder beobachtet wurden, *abzählen* und ihre *Häufigkeit* angeben. Dazu werden wir noch im Kapitel 2 (S.29 f.) kommen.

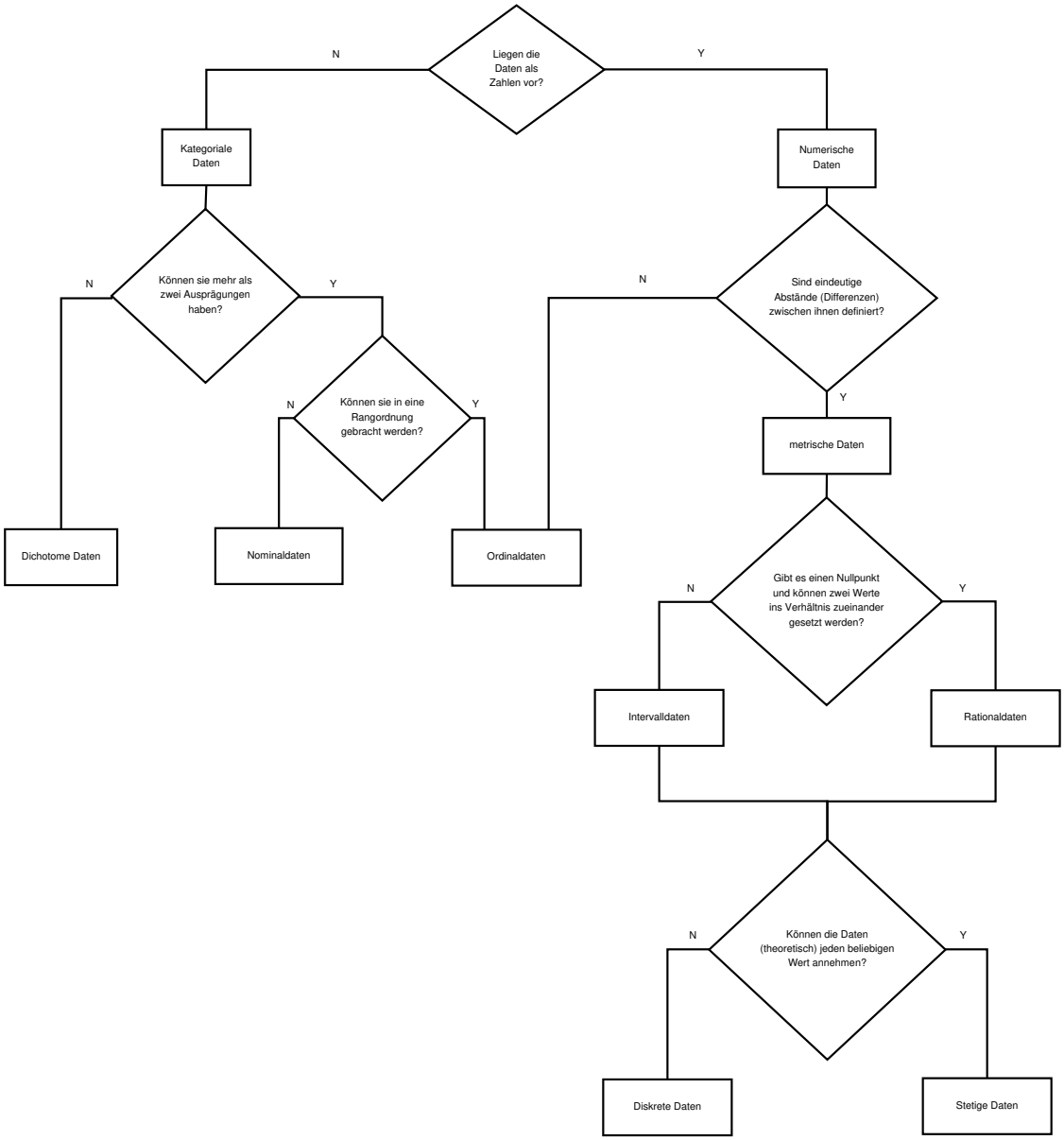


Abb. 1.4: Verschiedene Datentypen können auf verschiedenen Skalen gemessen werden

1.5 Modellwelt und Realwelt

Ergebnisse und Aussagen, zu denen wir mit Methoden der beschreibenden und explorativen Statistik kommen, beziehen sich immer auf die konkret untersuchte Datenmenge der *Realwelt*. Die schließende Statistik hat dann zum Ziel, aus den (oft: wenigen) vorliegenden Daten der Realwelt zu lernen und auf eine *Modellwelt* zu schließen. Es geht dabei um ein generelles Modell der «*Wirklichkeit zufälliger Phänomene*». Schließlich lohnt der ganze Aufwand ja nur, wenn wir damit die Realität möglichst allgemein modellieren können und nicht nur bezogen auf die Daten, die wir mehr oder weniger zufällig erhoben haben.

Manchmal bezeichnen wir die allgemeinen Modelle auch als *Grundgesamtheit* (auch: *Population* oder *Kollektiv*).

Sehen wir uns zunächst einige Beispiele an:

1. Im November 2019 gab es im Studiengang WIBA 288 aktive ordentliche Studierende.
2. Im gleichen Jahr lebten 2.373 Grauhörnchen im New Yorker Centralpark *(Quelle: 2019.thesquirrelcensus.com)*
3. Am Beginn des Jahres 2025 lebten in Österreich 2 Personen mit einer Staatsangehörigkeit des Inselstaats Tuvalu. *(Quelle: t1p.de/wiba-stat18)*
4. Das beliebteste Schulfach der Österreicher:innen während der Pflichtschulzeit ist «Bewegung und Sport». Bereits an zweiter Stelle kommt «Mathematik». Am Ende der Skala stehen Informatik, Religion und eine zweite Fremdsprache (neben Englisch). *(Quelle: t1p.de/wiba-stat21)*
5. Das Verhältnis zwischen der Masse eines Protons und der Masse eines Elektrons hat sich in den vergangenen sechs Milliarden Jahren nicht geändert und beträgt 1.836,15 zu 1.
   (Quelle: t1p.de/wiba-stat10)

Eine **Grundgesamtheit** besteht aus der Menge aller Objekte, die irgendwelche gemeinsamen Charakteristika aufweisen und Gegenstand unserer Untersuchungen sind. Beispiel: «die Bevölkerung Österreichs» zu einem bestimmten Stichtag oder «die Anzahl der Österreicher:innen, die bei einer europaweiten Volkszählung in einem anderen europäischen Land als Österreich leben».

Eine Grundgesamtheit kann unterschiedliche Größe haben; wir nennen dies den **Umfang der Grundgesamtheit**. Der Umfang reicht von einigen wenigen (alle Personen, die im Wintersemester 2024 im Studiengang WIBA eingeschrieben sind) über eine sehr große Anzahl von Elementen (alle Menschen, die es am 1.1.2024 um 0.00 Uhr auf der Erde gab) bis hin zu unendlich großen Mengen.

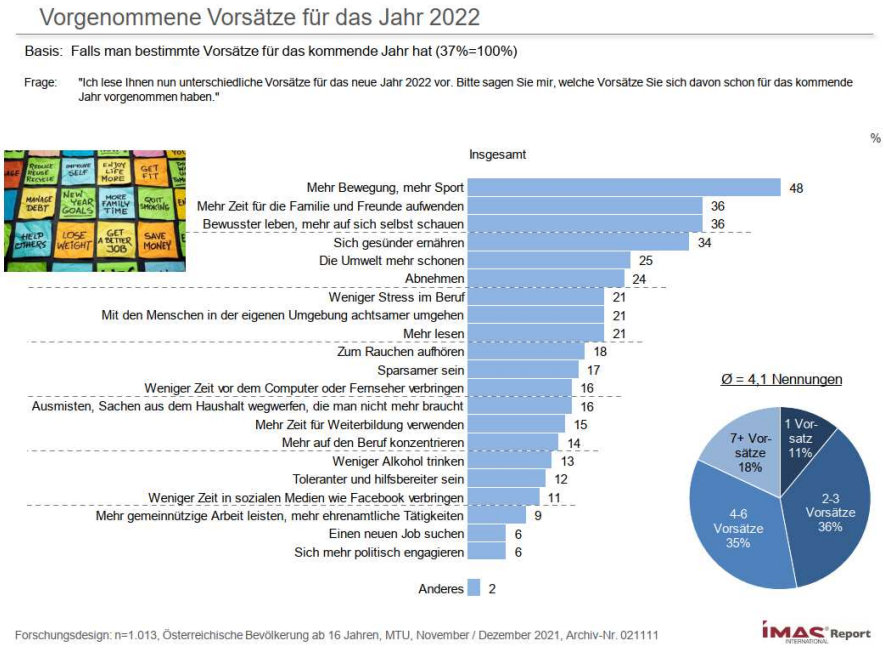


Abb. 1.5: Was sich die «Österreichische Bevölkerung ab 16 Jahren» für 2022 alles vorgenommen hat. Sample: n=1.013 Personen, statistisch repräsentativ für die österreichische Bevölkerung ab 16 Jahren. (Quelle: IMAS International Eigenstudie. Report Nr. 12/2021)

Es kommt nicht so oft vor, dass wir wirklich Zugriff auf die Grundgesamtheit haben. Selbst wenn wir sehr, sehr viele Daten erhoben haben: Die Bevölkerung Österreichs ist – in Hinblick auf die Aufgabe, jede und jeden einzelnen zu befragen – schon ziemlich groß und ändert sich alleine durch Geburten und Sterbefälle ständig. Wir täten uns schwer, hier jeden einzelnen einzubeziehen. Von Angaben über die Weltbevölkerung ganz zu schweigen. Oft haben wir es daher nur mit einer **hypothetischen Grundgesamtheit** zu tun; in vielen Fragestellungen gibt es schon definitionsgemäß nur eine hypothetische Grundgesamtheit, zum Beispiel bei Qualitätsmessungen von technischen Systemen oder im Prozessmanagement: Was könnte die «Grundgesamtheit» von unendlich oft durchgeführten System- oder Prozessdurchläufen sein?

Ein und derselbe Datensatz kann auch manchmal als Grundgesamtheit gesehen werden und ein anderes Mal nicht. Das Beispiel «Studierendenanzahl» ist bezogen auf einen bestimmten Stichtag in einem bestimmten Studienjahr als Grundgesamtheit einzuordnen. Wer nicht für ein ordentliches Studium erfasst und gemeldet wird, ist definitionsgemäß kein:e ordentliche:r Studierende:r. Bei der Anzahl der in Österreich lebenden Person mit tuvaluischer Staatsangehörigkeit ist nicht so eindeutig, weil sich eventuell auch mehr Personen aus diesem Land in Österreich aufhalten könnten, die aber offiziell nicht erfasst wurden. Im Bezug auf ein «allgemeines Modell» sind aber auch die Studierendenzahlen zu einem bestimmten Stichtag nur ein Beispiel von vielen. Wenige Tage davor

kann es eine andere Anzahl gegeben haben, wenige Tage danach kann es wieder anders aussehen.

Im Central-Park-Beispiel ist überhaupt zu hinterfragen, ob bei der Eichhörnchen-Zählung wirklich alle Tiere anwesend waren und gezählt werden wollten. Hier handelt es sich also um eine *hypothetische* Grundgesamtheit.

Weder im vierten Beispiel noch in der Umfrage der Abb.1.5 wurden tatsächlich *alle* Österreicher:innen über die Beliebtheit der Schulfächer bzw. ihre Neujahrsvorsätze befragt, obwohl das sprachlich suggeriert wird («Das beliebteste Schulfach *der Österreicher:innen*», «Die *österreichische Bevölkerung* ab 16 Jahren»).

Für die Umfrage wurde eine **Stichprobe** herangezogen: Jeweils etwas mehr als 1.000 Personen, die als statistisch *repräsentativ* für die österreichische Bevölkerung ab 16 Jahren angesehen werden. Es handelt sich hier nur um eine *Teilmenge* aus der Grundgesamtheit, aus der wir aber auf ein allgemeines Verhalten aller Österreicher:innen geschlossen haben.

Eine Stichprobe können wir uns also ähnlich wie eine Wein-Degustation vorstellen: Wir nehmen nur einen kleinen Schluck. Und dennoch treffen wir dann eine Aussage über die ganze Flasche, vielleicht sogar über ein ganzes Fass: Wir vermuten, die ganze Flasche enthält «lieblichen» oder «blumigen» Wein, der vielleicht einen «anhaltenden Abgang» hat – und das, obwohl wir im Vergleich zum ganzen Fass nur einen ganz kleinen Teil gekostet haben.

In den Sozialwissenschaften spricht man übrigens bei Verwendung einer Stichprobe von einer *Teilerhebung*, bei einer Grundgesamtheit von einer *Vollerhebung*.

Laut *Belastungsbarometer* der Statistik Austria hat im Jahr 2013 das Ausfüllen von Fragebögen Österreichische Unternehmen 774.277 Arbeitsstunden gekostet[a]. Geht man davon aus, dass 2013 die Jahresarbeitszeit eines österreichischen Arbeitnehmers 1.520 Stunden betrug[b], dann haben rein rechnerisch 509 Personen nichts anderes getan als Fragebögen und Statistiken ausgefüllt…

[a] Statistik Austria: *Meldepflichten und Belastung der Wirtschaft durch Erhebungen von Statistik Austria 2001-2013*
[b] OECD: *Average annual hours actually worked per worker*
https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS

Nicht auszudenken was passieren würde, wenn die Statistik Austria nur mehr Vollerhebungen durchführen würde…

Ingenieurwissenschaftlich betrachtet besteht eine Stichprobe aus Daten einer empirischen Beobachtung der Realwelt, mit deren Hilfe wir ein Modell (= die

Grundgesamtheit) aufbauen können. Wir können auch sagen: Eine Stichprobe ist die Momentaufnahme einer Situation, die Grundgesamtheit ist ein Modell für diese Situation. Im fünften Beispiel auf Seite 20 haben wir so einen Fall vorliegen. Wir nehmen Messungen in einer sechs Milliarden Lichtjahre entfernten Galaxie vor und haben dann einen Wert von heute und einen von vor sechs Milliarden Jahren (So einfach kann die Erhebung historischer Daten sein!) und über das Verhalten dazwischen treffen wir irgendwelche modellhaften Annahmen.

Stichproben können unterschiedlich groß sein. Die Anzahl der Elemente einer Stichprobe nennen wir den **Umfang der Stichprobe** und bezeichnen ihn in der Regel mit der Variablen n.

Eine Frage, die bei der statistischen Analyse immer wieder auftaucht, ist die Frage, wie groß eine Stichprobe sein sollte, oder auch allgemein:

Was ist eine «gute» Stichprobe?

Obwohl wir nur einen Teil von «Allen» zur Verfügung haben, wollen wir eine möglichst exakte Schätzung über die Grundgesamtheit erhalten. Bei geschickter Wahl der Stichprobe können wir dann ruhigen Gewissens von der Stichprobe auf die Grundgesamtheit schließen und ein gutes Modell bilden. Man sagt auch: Die Stichprobe muss **repräsentativ** sein. Das heißt: Elemente mit möglichst verschiedenen für die Untersuchung wichtigen Eigenschaften, die für das Ergebnis relevant sein könnten, müssen in der Stichprobe vertreten sein. Die Stichprobe ist idealerweise ein «verkleinertes Abbild» der Grundgesamtheit; dann wird sie sie auch gut widerspiegeln und ich kann von der Stichprobe auf die Grundgesamtheit verallgemeinern.

Um bei bekannter Größe der Grundgesamtheit die notwendige Stichprobengröße berechnen zu können, benötigt man zwei Kennzahlen, die man beim Ergebnis erreichen will: Die Wahrscheinlichkeit, mit der deine Stichprobe die Grundgesamtheit wiedergibt (genannt **Konfidenzniveau**, siehe auch Kap.6) und den Bereich, um den die Grundgesamtheit von der Stichprobe abweichen darf (genannt **Fehlerspanne**). Sie geben zum Beispiel bei Umfragen an, wie genau Umfrageergebnisse die Meinungen der Gesamtpopulation widerspiegeln. Das bedeutet: Wenn zum Beispiel 50% aller Befragten aus der Stichprobe eine bestimmte Frage mit «Ja» beantworten und wir eine Fehlerspanne von ± 5 Prozentpunkten tolerieren, dann könnte es in der Grundgesamtheit irgendein Wert zwischen 45% und 55% sein, die diese Frage bejahen. Und ein Konfidenzniveau von 95% bedeutet, dass diese eben gemachte Aussage («*irgendein Wert zwischen 45% und 55%*») mit einer Wahrscheinlichkeit von 95% stimmt.

Auf der Seite de.surveymonkey.com/mp/sample-size-calculator (und mehreren anderen Seiten im Internet) kann man konkrete Werte für die Stichprobengröße berechnen. Will man demnach die Meinung aller 300 WIBA-Studierender kennen (bei einem Konfidenzniveau von 95% und einer Fehlerspanne von $\pm 5\%$, benötigt man eine (auswertbare) Stichprobengröße von 169 Antworten. Vergrößert man das Konfidenzniveau auf 99% und verringert die Fehlerspanne auf $\pm 1\%$, dann sind es 295 Antworten, also beinahe die gesamte Grundgesamtheit. Bei einer Grundgesamtheit von 1.000 Personen (das entspricht in etwa allen FERNFH-Studierenden) werden 278 Antworten benötigt (95%, $\pm 5\%$), bei 60.000 Personen (= alle FH-Studierenden Österreichs) 382 Antworten – aber natürlich dürfen das nicht nur Studierende der FERNFH sein.

Ob die Stichprobe repräsentativ ist, hängt nämlich nicht nur von der Zahl der Personen ab, die untersucht oder befragt wurde, sondern mehr noch von ihrer methodisch richtigen Auswahl. Um dabei auf Nummer Sicher zu gehen werden zum Beispiel für den so genannten *TELETEST 2.0*, bei dem (mit einem elektronischen Messgerät) die Reichweiten und Marktanteile von TV-Sendern gemessen werden[34], in ganz Österreich 3.253 Menschen (aus 1.474 Haushalten) herangezogen. Dabei stehen diese 3.253 Personen für mehr als 7,5 Millionen Österreicher:innen ab 12 Jahre (und mit Fernsehgerät) und zusätzlich repräsentieren 286 Kinder die etwa 748.500 österreichischen Kinder von 3 bis 11. Insgesamt gibt es ca. 3,872 Millionen Privathaushalte mit einem TV-Gerät (= Grundgesamtheit); 0,038% dieser Haushalte dienen als repräsentative Stichprobe für die Grundgesamtheit. Und aus dieser Stichprobe kann man zum Beispiel ableiten, dass «Bundesland heute» vom 29.09.2024 mit etwa 1,803 Millionen Seher:innen die meistgesehene ORF-Sendung des Jahres 2024 war).

Die Auswahl der konkreten Elemente, die in einer Stichprobe vertreten sein sollen, ist letztlich gar nicht so einfach. Die einfachste und von uns präferierte Möglichkeit, zu einer repräsentativen Stichprobe zu kommen, ist immer noch, bei einfach eine völlige *zufällige* Auswahl vorzunehmen.

1.6 Tipps aus der Mathematik

Statistik bedient sich methodisch sehr oft bei der Mathematik. Die wichtigsten Kapitel und Themen aus der Mathematik, die wir benötigen, sind

- *Algebra*, also der Umgang mit Variablen und das Auflösen von Gleichungen,
- der Umgang mit *Funktionen*,

[34] siehe der.orf.at/medienforschung/fernsehen/teletest

- einige spezielle *Notationen* wie zum Beispiel das Summenzeichen Σ,
- eine ausgeprägte *Abstraktionsfähigkeit*, um die Methoden, Formeln und Algorithmen der Statistik auch praktisch umsetzen zu können,
- und natürlich so einfache Dinge wie *Prozentrechnung* oder
- das richtige *Runden* von Ergebnissen.

Bei der Verwendung von Zahlwörtern zu beachten: Zum Beispiel die Angabe, dass Wien *ca. 1,98 Millionen* Einwohner:innen hat:

1. Verwende für «Millionen» die Abkürzung *Mio.* aber nicht «Mill.» (das könnte nämlich auch als «Milliarde» gelesen werden, was aber mit *Mrd.* abkgekürzt wird).

2. Beachte bei englischsprachigen Texten, dass eine englische *Billion* im Deutschen eine *Milliarde* ist (1.000.000.000), eine *Trillion* (EN) einer *Billion* (DE) entspricht (1.000.000.000.000) und eine *Quadrillion* (EN) einer *Billiarde* (DE) (1.000.000.000.000.000).

Neujahrskonzert verdoppelt Frauenanteil
Fünf Erstaufführungen birgt das Programm für das Neujahrskonzert 2026 im Goldenen Saal des Wiener Musikvereins, darunter zwei Werke von Komponistinnen. Dabei handelt es sich um den «Rainbow Waltz» von Florence Price und die Polka mazur «Sirenen Lieder» von Josefine Weinlich.

Quelle: Salzburger Nachrichten, 30. Oktober 2025

Auch wenn es mathematisch stimmt: Wenn die Wiener Philharmoniker:innen 2025 nur ein Werk einer Komponistin auf dem Programm hatte und jetzt zwei, ist die Überschrift vielleicht ein wenig irreführend…

Verwende immer Einheiten: Beinahe alles, was wir messen, kann in unterschiedlichen *Einheitensystemen* gemessen werden. Rohöl-Mengen können in *Barrel* oder *Litern* angegeben werden; Entfernungen in *Metern* oder *Poronkusema*[35], die Temperatur in Grad *Celsius* oder *Fahrenheit*, etc.

Eine Angabe wie «*Im Jahre 2007 betrug die globale durchschnittliche Temperatur 14.4 Grad.*» ist daher nicht eindeutig, sondern muss richtigerweise lauten: «*Im Jahre 2007 betrug die globale durchschnittliche Temperatur 14.4 Grad Celsius.*»

Verwende Analogien und Vergleiche: Das gilt vor allem für die mündliche Präsentation deiner Forschungsarbeit. Du könntest zum Beispiel nüchtern feststellen:

[35] Das ist nur eine von sehr, sehr vielen alternativen Möglichkeiten zu Metern. Siehe z.B. de.wikipedia.org/wiki/Längenmass

«Monaco ist mit 16 754 *Einwohner:innen pro Quadratkilometer der am dichtesten besiedelte Staat der Welt.»*
oder aber sagen:
«Monaco ist mit 16 754 *Einwohner:innen pro km^2 der am dichtesten besiedelte Staat der Welt. Diese Bevölkerungsdichte entspricht einem Fußballplatz, auf dem sich ständig 119 Personen aufhalten. Im Vergleich dazu hat in Österreich jede:r Einwohner:in 1.4 Fußballplätze als Lebensraum zur Verfügung, jede:r Australier:in sogar 54.»*

Manchmal ist es auch hilfreich anzugeben, ob ein Zahlenwert typisch oder extrem ist. Geben wir zum Beispiel den folgenden Satz in drei Versionen an, werden sie jeweils unterschiedliche Reaktionen beim Zuhören auslösen:

«Im Jahre 2007 betrug die globale durchschnittliche Temperatur 14.4°C. Das sind um 0.4°C mehr als der langjährige Durchschnitt im Referenzzeitraum 1961-1990.»

«Im Jahre 2007 betrug die globale durchschnittliche Temperatur 14.4°C. Das sind um 0.4°C mehr als der langjährige Durchschnitt im Referenzzeitraum 1961-1990, aber 2007 war immerhin das kühlste Jahr des Zeitraums 2001-2007.»

«Im Jahre 2007 betrug die globale durchschnittliche Temperatur 14.4°C. Das sind um 0.4°C mehr als der langjährige Durchschnitt im Referenzzeitraum 1961-1990, aber 2007 war immerhin das kühlste Jahr des Zeitraums 2001-2007. Gleichzeitig war es auch das 8. wärmste Jahr seit Beginn der Aufzeichnungen im Jahre 1850.»

Zur **Genauigkeit**, mit der wir Ergebnisse unserer Berechnungen angeben: Es macht keinen Sinn, Parameter, die wir aus den Daten berechnet haben, auf ein Dutzend Nachkommastellen oder mehr anzugeben, nur weil der Rechner so viele Stellen ausgibt. Es ist üblicherweise ausreichend, die berechneten Parameter mit einer oder maximal zwei Nachkommastellen mehr anzugeben als die Originaldaten. Insbesondere auch bei Zahlen, die wir in Grafiken einfügen.

Wenn du Zahlen rundest: runde immer erst am Schluss beim Ergebnis, nicht schon während der Rechnung. So sollte beispielsweise bei prozentuellen Angaben die Gesamtsumme der gerundeten Werte 100% ergeben. Gib eventuell einen Hinweis an, wenn es sich um gerundete Zahlen handelt, die (scheinbar) in Summe nicht 100 ergeben.

Rundungsregel: Möchte man beispielsweise das Ergebnis auf zwei Stellen runden, dann schneidet man die Zahl zunächst nach der dritten Dezimalstelle ab. Die dritte Dezimalstelle wird in weiterer Folge auch noch weggelassen, allerdings ist dabei Folgendes zu beachten: Ist die dritte Dezimalstelle größer oder gleich 5, dann wird die zweite Dezimalstelle um 1 erhöht; ist sie kleiner oder gleich 4, dann bleibt die zweite Dezimalstelle gleich.

Mathematisch kann man das auch so bewerkstelligen[36]:

Man addiert zu der Dezimalstelle unmittelbar rechts neben der «Abbruchstelle» die Zahl 5. Und dann lässt man alle Ziffern rechts von der Abbruchstelle weg. (Probier das einmal ruhig mit ein paar Beispielen aus!)

Nach dem Runden darf man übrigens nicht willkürlich einfach Nullen dranhängen oder streichen. Damit würde eine andere Genauigkeit vorgegaukelt würde.

Und hier noch der goldene Tipp zum Schluss dieses Kapitels:

> *The only way to really learn statistics is to **do** statistics.*
>
> — Russell A. Poldrack: Statistical Thinking for the 21st Century

Lösung zum Rätsel auf Seite 4

Es mag vielleicht überraschen, aber es lässt sich *kein* klarer logischer Schluss über die relativen Altersverhältnisse der drei Gruppen ziehen. Man kann zwar daraus mutmaßen, dass die Weltbevölkerung im Schnitt am jüngsten ist, gefolgt von Emilias Familie und schließlich der österreichischen Bevölkerung, aber ob das wirklich so ist, ist abhängig davon, um wieviel die jeweiligen Personen jünger sind als Emilia.
Auch die Frage, wieviele Personen in Emilias Familie älter sind als sie, lässt sich aus der Angabe auf Seite 4 alleine nicht lösen. Dazu müsste man wissen, wie groß Emilias Familie ist. Wir wissen nur dass 20% ihrer Familie älter *oder genauso alt* sind wie sie, wobei Emilia selbst bei den 20% eingeschlossen ist. Bei insgesamt fünf Personen wäre dann zum Beispiel niemand älter als Emilia, bei 10 Personen wäre eine Person älter oder genauso alt wie sie…

[36] So ist «Runden» in der DIN 1333 definiert und so lässt es sich auch programmiertechnisch leicht hinkriegen.

2

Die Darstellung empirisch erhobener Daten in Tabellen und Diagrammen

In einem ersten Schritt wollen wir Daten **organisieren**, **strukturieren**, **zusammenfassen** und möglichst anschaulich **darstellen** und übersichtlich **präsentieren**. Ausgangspunkt ist dabei die Erfahrung, dass es für die meisten nicht so leicht ist, sich in einem «Zahlenhaufen» zurecht zu finden, aber einen guten Eindruck von der Verteilung der Daten aus einer grafischen Präsentation in **Diagrammen** ablesen können, zumindest aber die Aufbereitung in **Tabellen** erwarten.

2.1 Klassenbildung

Qualitative Daten repräsentieren eine bestimmte Kategorie eines Merkmals; sie gehören zu einer bestimmten «Klasse». Manchmal werden auch quantitative Daten in mehreren Teilbereiche zusammengefasst und *klassifiziert*:

Klassenbildung bedeutet, den Wertebereich der (möglichen) Merkmalswerte in Teilbereiche (*Klassen*) aufzuteilen, und jede Realisierung (also jede Messung, Beobachtung, ...) einer Klasse zuzuordnen. Die Klassen müssen in ihrer Gesamtheit den Wertebereich vollständig (d.h. auch *lückenlos*) überdecken und außerdem einander ausschließen, d.h. es kann nicht sein, dass ein Messwert in mehrere Klassen hineinpassen würde.

Ergebnis der Klassierung der Daten ist letztlich, dass man nicht mehr jeden einzelnen Merkmalswert und die Häufigkeit seines Auftretens angibt, sondern nur noch für jede Klasse die Gesamtanzahl der in ihr enthaltenen Merkmalswerte.

Je weniger Klassen man bildet, desto übersichtlicher und «einfacher» wird die Stichprobe zwar, es geht aber auch ein mehr oder weniger großes Stück an konkreter Information verloren. Umgekehrt: Je größer die Anzahl der Klassen ist, desto unübersichtlicher bleibt die Stichprobe. Üblich sind, je nach Stichprobengröße, in etwa 5 bis 20 Klassen.

Bei quantitativen numerischen Daten sollten die *Klassengrenzen* «runde» und «einfache» Zahlenwerte sein. Die erste und letzte Klasse werden oft als *offene* Klassen geführt, d.h. von $-\infty$ oder 0 (untere Grenze der ersten Klasse) bzw. $+\infty$ (obere Klasse der letzten Klasse) begrenzt. Die *Klassenbreiten* (= obere minus untere Klassengrenze) werden so gewählt, dass sie möglichst gleich lang und die Klassenhäufigkeiten (Anzahl der Messwerte pro Klasse) nicht extrem unterschiedlich sind. (Die Forderung nach gleich großen Klassenbreiten ist nicht zwingend, in den meisten Anwendungsfällen aber üblich).

Bezeichnen wir den größten beobachteten Wert mit x_{max} und den kleinsten mit x_{min}, so ergibt sich die Klassenbreite d bei einer Einteilung in m Klassen zu:

$$d \approx \frac{x_{\max} - x_{\min}}{m} \tag{2.1}$$

wobei bei offenen Klassen x_{min} und x_{max} in den beiden offenen Klassen liegen sollten (also x_{min} in der ersten und x_{max} in der letzten Klasse). Für die Anzahl m der Klassen kann man als Faustregel[1] heranziehen:

$$m \approx \begin{cases} 1 + 3.322 \cdot {}_{10}\log n & \text{für } n \leq 100 \\ 3.322 \cdot {}_{10}\log n & \text{für } n > 100 \end{cases} \tag{2.2}$$

Jedenfalls sollte gelten:

$$2^m \geq n \tag{2.3}$$

Neben der weiter oben erhobene Forderung nach «runden» Klassengrenzen sollten auch die Klassenbreiten d «runde» Zahlen sein (z.B. 2, 5, 10 oder Vielfache von 5 oder 10).

Damit Messwerte nicht genau auf einer Klassengrenze zu liegen kommen, werden üblicherweise die unteren Klassengrenzen in die jeweilige Klasse eingeschlossen, die oberen hingegen ausgeschlossen und zur nächsten Klasse hinzugezählt. D.h. ein Wert, der genau auf einer Klassengrenze liegt, wird immer zur größeren Klasse gezählt. Oder man legt das gleich explizit fest. Wenn du zum Beispiel das Merkmal «Einwohner:innenzahl» von Städten in Klassen einteilen willst, kannst du als Intervallgrenzen wählen (Tab.2.1):

[1] Woran wir sehen: Nicht jede Faustregel kann leicht im Kopf gerechnet werden. . . .
${}_{10}\log n$ bedeutet übrigens: Logarithmus von n zur Basis 10.

Klassennummer i	explizite Angabe der Klassengrenzen (von-bis)		mathematisch elegantere Angabe der Grenzen
1	0	1999	$x < 2000$
2	2000	4999	$2000 \leq x < 5000$
3	5000	19 999	$5000 \leq x < 20\,000$
4	20 000	99 999	$20\,000 \leq x < 100\,000$
5	100 000	$+\infty$	$100\,000 \leq x$

Tabelle 2.1: Zwei Möglichkeiten zur Angabe von Klassengrenzen (Beispiel: Einwohner:innenzahlen)

Auch bei qualitativen Daten sollte die Anzahl der Klassen überschaubar gehalten und gegebenenfalls übergeordnete Klassen aus mehreren Kategorien gebildet werden. Oft werden auch Elemente, die in der Stichprobe nur selten vorkommen, in einer einzigen Klasse zusammengefasst, die den Namen **«andere»** oder **«sonst»** oder ähnliches trägt.

2.2 Darstellung der Daten in Häufigkeitstabellen

Es mag vielleicht ein wenig verwundern, dass ein Kapitel über *Visualisierung* mit einer *Tabelle* beginnt. Tatsächlich haben Tabellen auch Vorteile gegenüber «bildlichen» Darstellungen: Sie sind meist intuitiv erfassbar und können beliebige Datentypen enthalten, auch (beinahe) beliebig viele Zufallsvariablen (sofern die Tabelle noch lesbar bleibt). Nachteilig ist, dass es letztlich Text bleibt und das visuelle System des Menschen visuelle Informationen und Muster viel schneller erfassen und verarbeiten kann als Text, den man erst lesen muss.

Eine **Häufigkeitstabelle** beinhaltet zumindest zwei Spalten: In der linken Spalte stehen alle *möglichen* Merkmalswerte – entweder in Form von Intervallen (Klassen) oder als explizite Angabe, bei numerischen Daten meist in aufsteigender Form geordnet vom kleinsten zum größten Wert. In der rechten Spalte steht oft die Anzahl, wie oft der jeweilige Datenwert in der Stichprobe vorkommt. Letzteres nennen wir auch die **Häufigkeit**. Häufigkeitszahlen können auch den Wert Null haben – offensichtlich dann, wenn dieser Wert in der konkreten Stichprobe nicht vorkommt.

Wir nennen diese tabellarische Beschreibung auch die **Häufigkeitsverteilung** des Merkmals bzw. der Zufallsvariable.

Die Häufigkeitsverteilung kann außer mit Häufigkeitszahlen auch mit der relativen Häufigkeit (meist in Form von Prozentangaben) beschrieben werden und

darüber hinaus mit der zusätzliche Angabe von absoluten oder relativen Häufigkeitssummen ergänzt werden:

Absolute Häufigkeit und kumulierte Häufigkeit

Die **absolute Häufigkeit** f_i ist die Anzahl der Elemente der Stichprobe, die gleich einem vorgegebenen Wert sind oder in eine bestimmte Klasse i von Werten gehören. Es muss gelten:

$$\sum_{i=1}^{m} f_i = n \tag{2.4}$$

das heißt: Die Summe aller absoluten Häufigkeiten muss die Gesamtanzahl aller Werte – also den Umfang der Stichprobe oder Grundgesamtheit – ergeben.

Sowohl in *MS Excel* als auch in *Libre Office Calc* gibt es verschiedene Möglichkeiten, eine Häufigkeitstabelle zu erstellen, je nachdem, ob die Daten klassifiziert werden sollen oder nicht:
Will man Klassen bilden und zählen, wieviele Elemente in der jeweiligen Klasse vorhanden sind, verwendet man (in Excel und in Calc) die Funktion `=HÄUFIGKEIT(Daten; Klassen)`. Bei unklassifizierten Daten verwendet man zum Erstellen einer Häufigkeitstabelle `=ZÄHLENWENN(Bereich wo die Daten stehen; Kriterium welche Daten daraus mitgezählt werden sollen)`

Die **absolute Häufigkeitssumme** (auch: *kumulierte Häufigkeit*) F_i ist die Anzahl der Beobachtungswerte, die einen vorgegebenen Wert (nämlich die Klassengrenze der $i-$ten Klasse) nicht überschreiten.

Wir erhalten die kumulierten Häufigkeiten, indem wir in der Tabelle neben den absoluten Häufigkeiten eine neue Spalte einfügen und dort in jeder Zeile alle bisherigen absoluten Häufigkeiten (also die f_i) zusammenzählen (siehe Tab. 2.2).

Relative Häufigkeit und relative Häufigkeitssumme

Die **relative Häufigkeit** h_i ist die absolute Häufigkeit dividiert durch den Stichprobenumfang:

$$h_i = \frac{f_i}{n} \tag{2.5}$$

Es muss gelten:

$$\sum_{i=1}^{m} h_i = 1 \tag{2.6}$$

das heißt die Summe aller relativen Häufigkeiten muss 1 ergeben.

Sehr oft geben wir relative Häufigkeiten auch prozentuell an, indem wir jedes h_i mit 100 multiplizieren und das Prozentzeichen % hinzufügen. Die Summe aller relativen Häufigkeiten ist dann 100%.

Die **relative Häufigkeitssumme** (auch: *kumulierte relative Häufigkeit*) H_i ist die jeweilige absolute Häufigkeitssumme dividiert durch den Stichprobenumfang.

Wir erhalten die kumulierten relativen Häufigkeiten, indem wir in der Tabelle neben den relativen Häufigkeiten eine neue Spalte einfügen und dort in jeder Zeile alle bisherigen relativen Häufigkeiten (also die h_i) zusammenzählen (siehe Tab. 2.2).

Beispiel 1 *Die Darstellung der Häufigkeitsverteilung einer Stichprobe in einer Häufigkeitstabelle:*

i	Klassengrenzen	f	F	h	H
1	$160 \leq x < 165$	1	1	0.042	4%
2	$165 \leq x < 170$	3	4	0.125	17%
3	$170 \leq x < 175$	4	8	0.167	33%
4	$175 \leq x < 180$	8	16	0.333	67%
5	$180 \leq x < 185$	3	19	0.125	79%
6	$185 \leq x < 190$	4	23	0.167	96%
7	$190 \leq x < 195$	0	23	0	96%
8	$195 \leq x$	1	24	0.042	100%
Summe		24		1	

Tabelle 2.2: Häufigkeitstabelle zu erhobenen Daten über die Körpergröße (in *cm*)

Wir können aus Tabelle 2.2 zum Beispiel herauslesen:
33% sind kleiner als 175 *cm. Oder:*
8 Personen der Stichprobe sind größer oder gleich 175 *cm aber kleiner als* 180 *cm.*
Wir wissen aber zum Beispiel nicht, wie viele Personen 181 *cm sind; die sind in der Klasse* $180 \leq x < 185$ *«versteckt» und es könnte sein, dass kein einziger* 181 *cm groß ist, oder aber 1, 2 oder 3 Personen.*

Beachte: statt $[160 \leq x < 165]$ *schreiben wir manchmal auch* $[160 - 164]$*, statt* $[165 \leq x < 170]$ $[165 - 169]$ *etc. Zwischen* 164 *aus der ersten und* 165 *aus der*

zweiten Klasse könnte dann aber eine Lücke entstehen. Diese Lücke darf nicht größer sein als die kleinstmögliche Differenz zwischen zwei Datenwerten. Wenn wir also nur auf cm-Genauigkeit messen, passt die verkürzte Angabe $[160 - 164]$*. Wenn wir aber mitunter auch einen mm-genauen Wert erhalten könnten, geht das nicht mehr, weil ja ein Wert* $164,7$ *nirgends in der Tabelle eingetragen werden kann. Dann müsste man* $[160,0 - 164,9]$ *und* $[165,0 - 169,9]$ *etc. schreiben.*

Noch einige Hinweise

...zur Klassifizierung von quantitativen Daten: Wenn es nur wenige mögliche Merkmalsausprägungen gibt (ca. 10-15), ist im Allgemeinen keine Klassifizierung vorzunehmen, sondern es werden gleich die Häufigkeiten der einzelnen Merkmalsausprägungen angegeben.

...zu Häufigkeitssummen: Aufsummiert werden nur die (absoluten oder relativen) Häufigkeiten f oder h, nicht aber die Merkmalswerte x_i. Und: Eine Summenbildung macht nur Sinn, wenn die Daten zumindest ordinalskaliert sind. Bei Nominaldaten ist eine Summenspalte nicht sehr aussagekräftig.

...zu Prozentangaben: Manchmal können Merkmalsträger auch mehr als eine Ausprägung eines Merkmals haben. Auf die Frage nach dem Geburtsort kann üblicherweise nur eine Antwort gegeben werden, aber auf die Frage *«Welche Komponistinnen und Komponisten der klassischen Musik kannst du namentlich benennen?»* kann auch mehr als eine Antwort kommen[2]. In der Statistik dokumentiert man das dann mit *«Mehrfachnennungen möglich»*, und immer, wenn Mehrfachnennungen möglich sind, kann die Summe über 100% liegen; die Angabe von Häufigkeitssummen sind dann nicht mehr sinnvoll. Das schließt auch einige Diagramme wie ein Kreisdiagramm (siehe 2.3, Seite 40) aus.

...ganz allgemein zu Tabellen: Die Tabellen, die wir tatsächlich publizieren und z.B. in wissenschaftlichen Arbeiten oder beruflichen Präsentationen verwenden, müssen nicht unbedingt alle hier angegebenen Spalten und Parameter (f, F, h, H) enthalten. Zu viele Zahlen verwirren eher. Gib nur diejenigen an, die zum Verständnis und zur Nachvollziehbarkeit deiner Aussagen notwendig sind.

[2] Laut Wikipedia ist der Vorrat, aus dem du schöpfen kannst, wirklich sehr, sehr groß: Siehe de.wikipedia.org/wiki/Liste_von_Komponisten_klassischer_Musik

. . . zum Aufbau von Tabellen: Sie sollten auch dann lesbar sein, wenn jemand «nur» die Tabellen anschaut und sich den restlichen Text nicht durchliest. Daher sollten sie einen **Titel** haben (oft steht der auch als «Caption» unter der Tabelle) und jede Spalte sollte eine Kopfzeile enthalten, in dem angegeben wird, welche Elemente in dieser Spalte zu finden sind. Entweder in der Spaltenüberschrift oder in der Beschreibung im Titel sollte auch angegeben sein, welche Einheiten verwendet werden (siehe auch Seite 25).

Sehr umfangreiche Tabellen sind oftmals für diejenigen, für die wir die Daten aufbereiten, zu kompliziert zu lesen und sie verlieren leicht den Überblick. Insbesondere für einen mit visuellen Präsentationsmedien unterstützten Vortrag aber auch für Texte und Berichte ist es in den meisten Fällen hilfreich, die Häufigkeitsverteilung graphisch aufzubereiten und (in schriftlichen Dokumenten, fast nie aber auf Präsentationsfolien) die Tabellen ergänzend hinzuzufügen.

2.3 Graphische Visualisierungen

Die grafische Darstellung der Daten ist meistens sehr hilfreich, um einen guten Eindruck von ihrer Verteilung zu erhalten und um zum Beispiel Häufigkeiten oder Muster in den Daten «auf einen Blick» zu erfassen. Für viele Menschen ist eine Grafik viel einprägsamer als eine Tabelle oder Liste voller Zahlen. Grafiken erlauben auch einen optischen – und damit meist schnelleren – Vergleich zwischen einzelnen Werten. Andererseits stellen Grafiken alleine (ohne die zugrundeliegenden Zahlen) immer auch einen gewissen Informationsverlust dar, weil die ursprünglich beobachteten Werte eventuell nicht mehr erkennbar sind.

Je nachdem, welche Daten wir präsentieren und Information wir dabei herausstreichen wollen, können wir verschiedene Diagrammtypen[3] verwenden.

Säulen- und Balkendiagramm

Säulen- und Balkendiagramme werden verwendet, wenn man die Verteilung von *diskreten* Daten darstellen will. Man sieht gut, in welcher Größenordnung Unterschiede zwischen den einzelnen Klassen bestehen und auch die Rangordnung innerhalb der Daten gut visualisieren.

In einem **Säulendiagramm** (auch: *Stabdiagramm*) werden die *Häufigkeiten* des Auftretens eines bestimmten Merkmalswertes (oder der Elemente einer Klasse)

[3] Unsere Beispiele sind nur eine kleine Auswahl aus unzähligen möglichen Darstellungsformaten und Diagrammtypen.

dargestellt, indem über den jeweils auf der Abszisse eingetragenen Bezeichnungen der Merkmalsträger schmale Rechtecke parallel zur Ordinate eingezeichnet werden, deren Länge proportional zum tatsächlichen Merkmalswert ist. Die Breite der Säulen hat hingegen keine Bedeutung und kann frei (aber gleich breit) gewählt werden – nach Möglichkeit aber so, dass alle vorkommenden Werte auch sinnvoll untergebracht werden können. Für ein Beispiel siehe Abb.2.1.

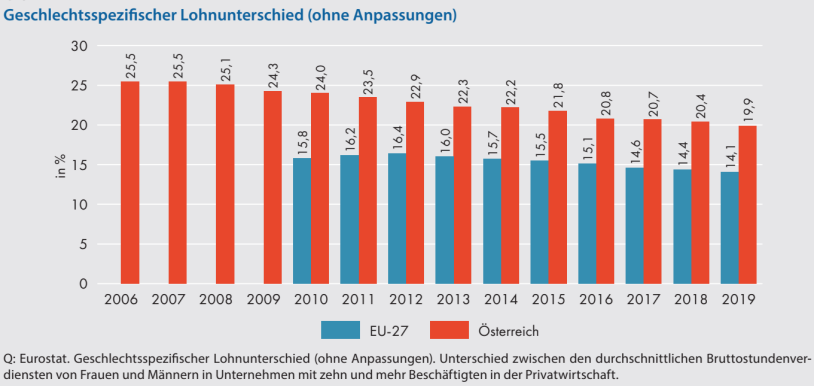


Abb. 2.1: Gender Pay Gap: Differenz zwischen den durchschnittlichen Bruttostundenverdiensten der männlichen und der weiblichen Beschäftigten in Prozent der durchschnittlichen Bruttostundenverdienste der männlichen Beschäftigten. (Quelle: Statistik Austria. 2021. *Wie geht's Österreich?* p.56)

Manchmal wird das Koordinatensystem, in dem das das Säulendiagramm eingebettet ist, auch um 90 Grad gedreht (Merkmalsträger werden auf der senkrechten Achse eingetragen, Merkmalswerte auf der waagerechten) und dann **Balkendiagramm** genannt. Balkendiagramme sind gegenüber Säulendiagramme insbesondere dann im Vorteil, wenn man mehr Klassen darstellen will, als sich nebeneinander ausgehen und auch wenn die Klassenbezeichnungen länger sind und Beschriftungen nur gegen die Leserichtung um 90 Grad gedreht möglich wären. (Abb.2.2).

Wenn in einem Balken- oder Säulendiagramm negative Werte dargestellt werden müssen, werden die negativen Daten (Balken oder Säulen) immer links bzw. unterhalb der Nulllinie platziert (siehe Abb.2.3). Das gilt auch, wenn gegebenenfalls die gesamte Datenreihe nur aus negativen Werten besteht.

Im Säulen- oder Balkendiagramm lassen sich auch zwei oder mehrere Datensätze darstellen, was oft einen anschaulichen Vergleich zwischen den Zufallsvariablen erlaubt. Dabei ist darauf zu achten, dass ein Vergleich zweier oder mehrerer Datensätze auf Basis der absoluten Häufigkeiten nur dann sinnvoll ist, wenn die Datensätze vom gleichen Umfang sind. Bei unterschiedlichem Umfang werden besser die relativen Häufigkeiten repräsentiert.

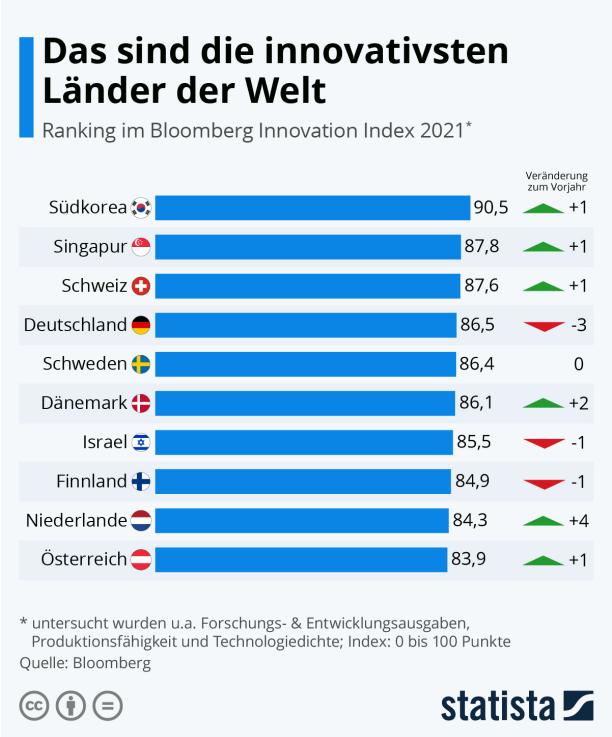


Abb. 2.2: Die innovativsten Länder der Welt (Quelle: de.statista.com/infografik/20548/)

Man sollte mit den Mehrfachsäulendiagrammen aber auch nicht mit der Anzahl der unterschiedlichen Säulen (Kategorien) übertreiben. Fünf oder mehr Säulen können vermutlich nicht mehr verglichen und interpretiert werden.

Und noch ein wichtiger Hinweis: Grundsätzlich sollten sowohl die x- als auch die y-Achse immer bei Null beginnen. Andernfalls werden die Differenzen zwischen den einzelnen Werten (oder Klassen) größer dargestellt, als sie in Wirklichkeit sind. Mitunter können einzelne Säulen überhaupt verschwinden[4] (siehe Abb.2.4).

In der englischsprachigen Literatur werden Säulendiagramme oft als «Histogram» bezeichnet. Im Deutschen sind wir da etwas präziser, und verwenden «Säulendiagramme» für diskrete Daten, «Histogramme» hingegen für stetige.

[4]Vermutlich würde es dir auffallen, wenn ganze Säulen verschwinden – sofern das Diagramm händisch erstellt wird. Im Fall der automatisierten Datenauswertung aber könnte das durchaus passieren, daher sollte man eine Verschiebung der Basislinie wirklich gut überlegen.

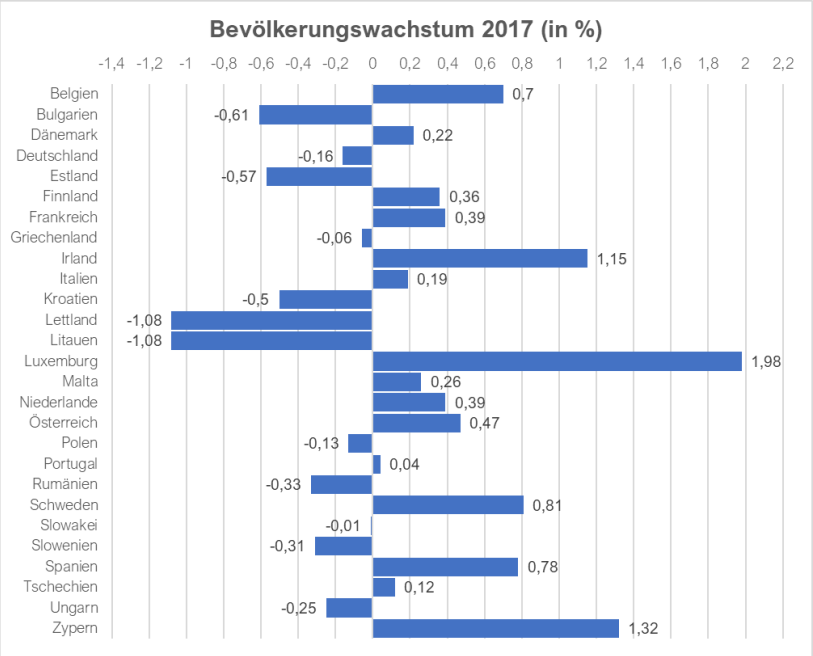


Abb. 2.3: Bevölkerungswachstumsraten der EU-Mitgliedsstaaten: Sie gibt die durchschnittliche jährliche prozentuale Veränderung der Bevölkerung an, die sich aus einem Überschuss oder Defizit zwischen Geburten und Todesfällen und der Differenz der Migrantinnen und Migranten ergibt, die in ein Land ein- oder aus einem Land ausreisen. (Datenquelle: CIA World Factbook, abgerufen: 23.7.2018)

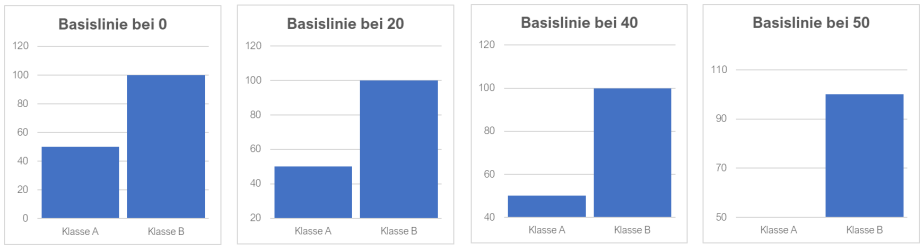


Abb. 2.4: Beginnt die y-Achse nicht bei 0, werden die Differenzen zwischen einzelnen Klassen überdeutlich dargestellt, was manchmal auch dazu führen kann, dass sie überhaupt verschwinden

Histogramm

In einem **Histogramm**[5] werden die Häufigkeiten *stetiger* Daten dargestellt. Stetige Zufallsvariable können ja jeden Zahlenwert aus $\mathbb{R}$ annehmen und so sehen wir die x-Achse des Diagramms als Zahlengerade[6], auf der jede denkbare Zahl abgebildet werden kann und auf der wir auch Intervalle bilden können, die eine

[5] Das Wort hängt vermutlich mit dem griechischen *ἱστός* (*histos*) = Mast(baum) zusammen. Die Bezeichnung wurde um 1895 von *Karl Pearson* eingeführt.

[6] auch unter *Zahlenstrahl* bekannt.

bestimmte Klasse von Merkmalswerten beinhaltet. Abb.2.5 zeigt zum Beispiel das Histogramm zur Tabelle 2.2.

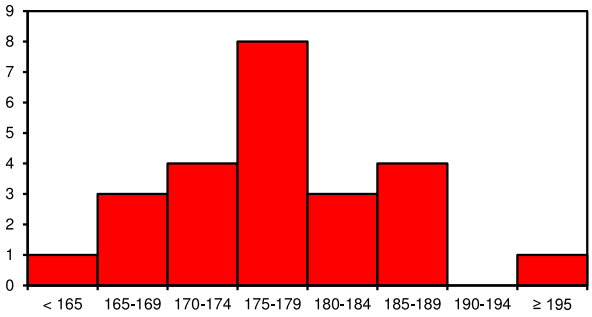


Abb. 2.5: Beispiel für ein Histogramm. Dargestellt sind die Daten aus Tab.2.2.

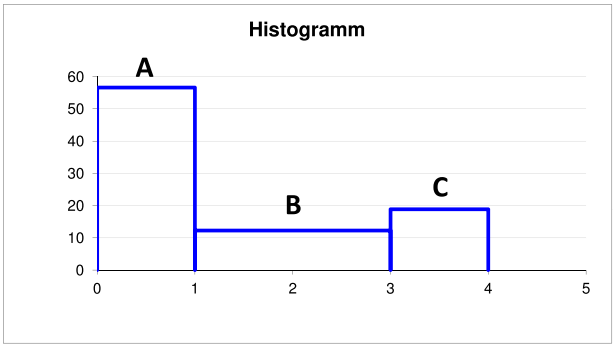
Auf der Abszisse werden im Histogramm die Klassengrenzen aufgetragen und über den Klassenintervallen Rechtecke errichtet, deren **Flächen** (!) proportional zu den Häufigkeiten sind. Beschriftet werden auf der Abszisse entweder die Klassengrenzen, die Klassenindizes oder die Klassenmitten (= obere minus untere Klassengrenze dividiert durch Zwei). Auf der Ordinate (y-Achse) wird die **Häufigkeitsdichte** angegeben, das ist der Quotient

$$\text{Häufigkeitsdichte} = \frac{\text{Häufigkeit}}{\text{Klassenbreite}} \tag{2.7}$$

Auf den ersten Blick schauen Histogramme genau aus wie die auf S.35 beschriebenen Säulendiagramme. Sie unterscheiden sich aber in wichtigen Punkten:

- Zwischen zwischen den Rechtecken (den Säulen) des Histogramms sind keine *Abstände*. Sie repräsentieren ja stetige, also kontinuierliche Daten, und ein Abstand würde der Stetigkeit entgegenstehen.
- Die Verwendung von Säulendiagrammen ist nur für den Fall *gleich breiter* Klassen angeraten. Bei unterschiedlichen Klassenbreiten benötigen wir ein Histogramm.
- Wie schon oben erwähnt: nicht die Höhe sondern die *Fläche* ist das Maß für die Häufigkeit. Nur im Fall gleicher Klassenbreiten spielt dieser Unterschied keine Rolle, dann könnten auf der y-Achse auch direkt die Häufigkeiten aufgetragen werden – streng genommen handelt es sich dann aber nicht mehr um ein Histogramm, sondern um ein Säulendiagramm, denn:
- Im Histogramm wird auf der y-Achse die Häufigkeitsdichte aufgetragen, im Säulendiagramm die (absolute oder relative) Häufigkeit.

Beispiel 2 *Das folgende Histogramm zeigt die Häufigkeitsverteilung einer Zufallsgröße, die in die drei Klassen A, B, und C eingeteilt wurde:*



In welcher Klasse / in welchen Klassen befinden sich die wenigsten Elemente?

Lösung:

In einem Histogramm ist nicht die Höhe der rechteckigen Säulen ausschlaggebend für die Anzahl der in der jeweiligen Klasse enthaltenen Elemente, sondern die Fläche.

In der Klasse A befinden sich demnach knapp unter 60 Elemente, in der Klasse B 24 (Breite = 2 × Höhe = 12) und in der Klasse C 20 (Breite = 1 × Höhe = 20) Elemente.

Somit befinden sich in der Klasse C die wenigsten Elemente (obwohl die Säule selbst höher ist als jene der Klasse B).

Kreisdiagramme

Beim **Kreisdiagramm** (auch: *Tortendiagramm*) wird jeder Ausprägung des Merkmals ein Kreissektor zugewiesen. Auch hier geht es also um die Fläche: Die Fläche jedes Sektors spiegelt die *relative Häufigkeit* seines Auftretens wider. Die Sektorgrenzen können berechnet werden, indem die relativen Häufigkeiten jeweils mit 360° multipliziert werden. Damit erhält jeder Merkmalswert ein «Tortenstück», dessen Größe der relativen Häufigkeit entspricht. Die einzelnen Kreissektoren erhalten zur besseren Lesbarkeit meist unterschiedliche Färbungen oder Grafikmuster. Abb.2.6 zeigt ein Beispiel für ein Kreisdiagramm:

Kreisdiagramme eignen sich besonders auch für nominalskalierte, kategorische Werte. Man erhält mit ihnen einen guten Gesamtüberblick über die Daten. Insgesamt sollten aber nicht mehr als 7 bis 9 Segmente (Klassen, Kategorien) vorliegen, damit es noch lesbar ist. Außerdem ist ein direkter Vergleich zweier Merkmale schwierig, wenn die betroffenen «Tortenstücke» nicht zufällig benachbart sind. Und selbst dann kann es sein, dass der Unterschied so gering ist, dass man

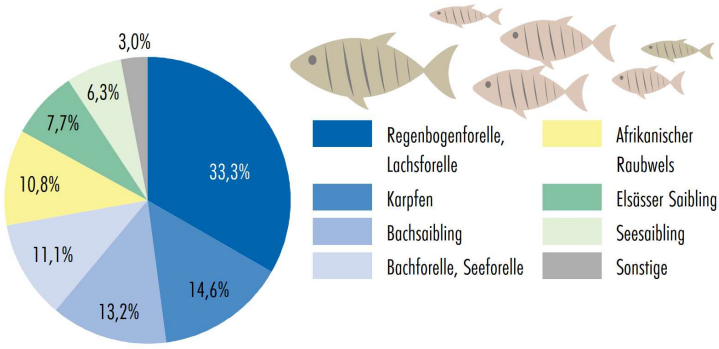


Abb. 2.6: Kreisdiagramme zur Produktion von Speisefischen in Österreich 2019 (Quelle: Statistik Austria: Österreichischer Zahlenspiegel Jänner 2021, p.7)

das aus der Größe der Tortenstücke alleine (also ohne Datenbeschriftung) nicht erkennen kann.

Tortendiagramme können übrigens auch mit einem «Loch» in der Mitte dargestellt werden (und ähneln dann eher einem *Donut* als einer Torte). Sie werden dann als *Ringdiagramme* bezeichnet. Siehe Abb.2.7.

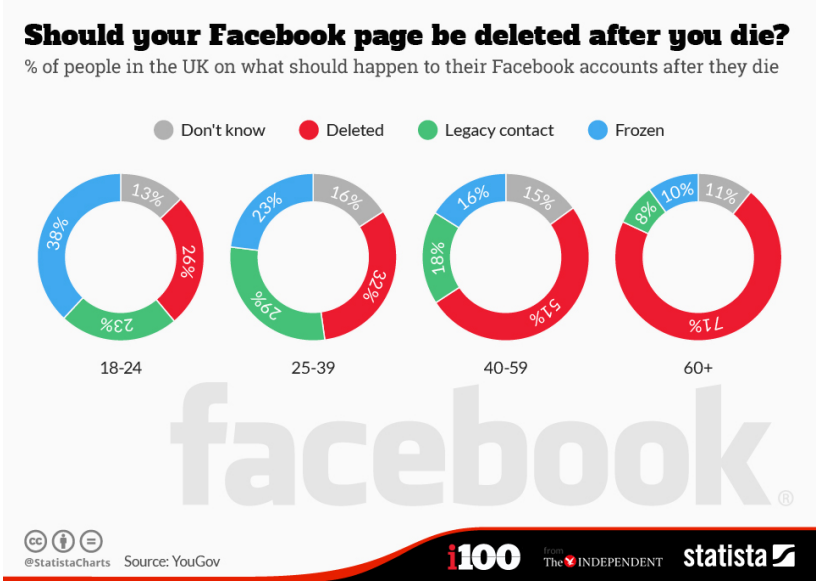


Abb. 2.7: Beispiel für ein «Donut-Diagramm»
(Quelle: http://www.statista.com/chart/3403/should-your-facebook-page-be-deleted-after-you-die/)

In klassischen Tortendiagrammen wie jedem der Abb.2.6 können keine negativen Werte visualisiert werden – dazu müssten Kreissegmente mit einer «negative Flächen» dargestellt werden. Donut-Diagramme können da Abhilfe schaffen, indem negative Werte «nach innen» in das Loch in der Mitte ausgerichtet werden.

Streifendiagramm

Ein **Streifendiagramm** (auch: **gestapeltes Säulendiagramm** oder **gestapeltes Balkendiagramm**) ist von der Aussage einem Tortendiagramm sehr ähnlich; es handelt sich aber nicht um einen Kreis, sondern um Balken oder Säulen, wo mehrere Merkmalswerte je Variable neben- oder übereinandergestapelt werden (siehe Abb.2.8). Im Gegensatz zu einem Tortendiagramm können hier nicht nur Relativ- sondern auch Absolutwerte dargestellt werden.

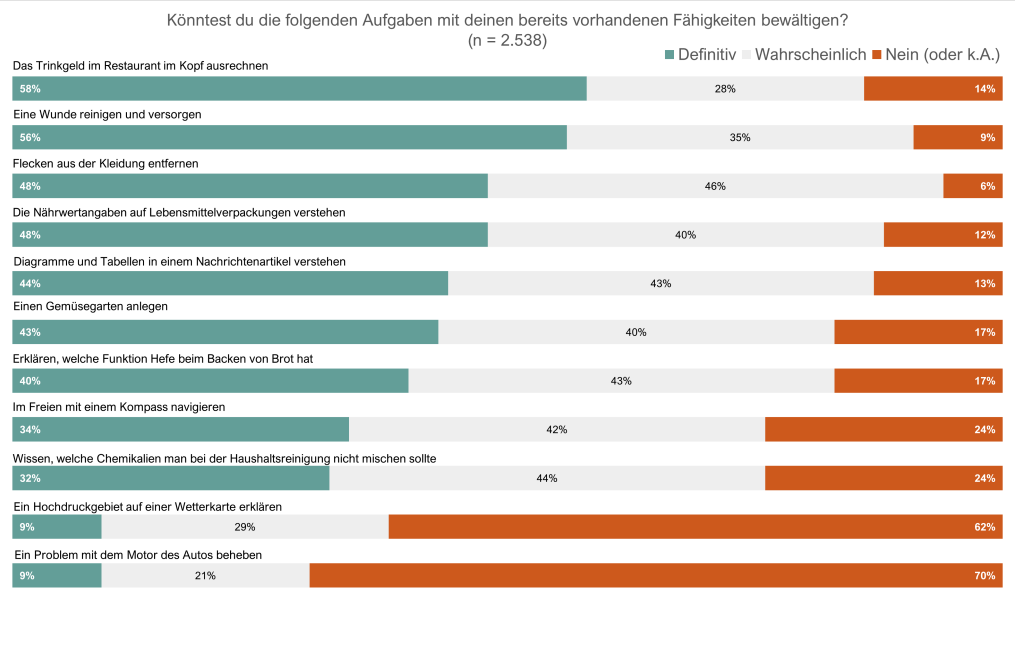


Abb. 2.8: Was US-Amerikaner:innen (Personen ab 18 Jahren, die in den Vereinigten Staaten leben) über ihre praktischen Fähigkeiten denken. (n = 2.538, Konfidenzniveau: 95%, Fehlerspanne: $\pm 1,5$ Prozentpunkte) (Datenquelle: https://www.pewresearch.org/wp-content/uploads/sites/20/2025/10/SR_25.10.10_science-skills_topline.pdf)

Statt einfacher rechteckiger Balken können in einem Diagramm übrigens auch Piktogramme verwendet werden. Das eignet sich insbesondere für die Darstellung der Häufigkeit von nominalskalierten Daten, siehe zum Beispiel Abb.2.9.

Linien- und Flächendiagramm

Liniendiagramme eignen sich vor allem dann, wenn mehrere Datenreihen verglichen werden sollen und wenn wir Trends (zeitliche Änderungen) darstellen wollen. Ein Beispiel ist in Abb.2.10 zu sehen.

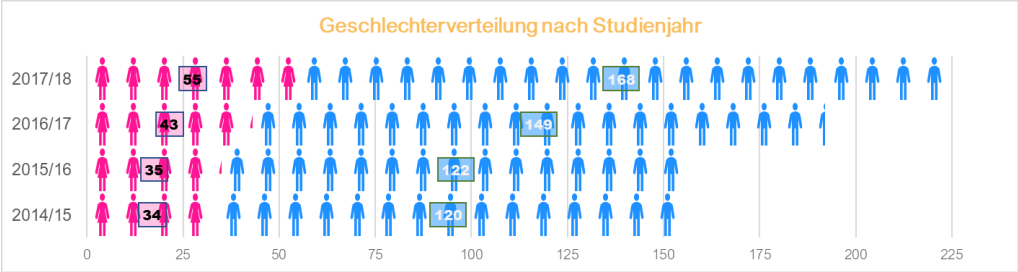


Abb. 2.9: WIBA-Studierende 2015-18 nach Geschlecht

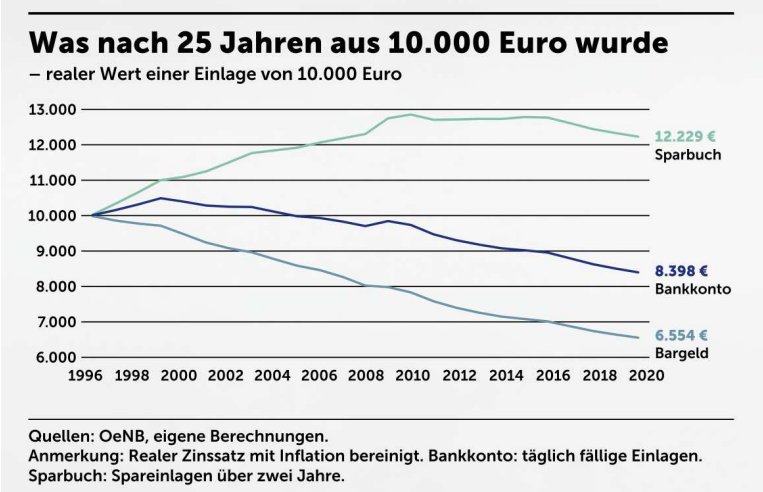


Abb. 2.10: Was über die vergangenen zweieinhalb Jahrzehnte aus 10.000 Euro wurde. (Quelle: Agenda Austria: Balken, Torten, Kurven Zweitausendeinundzwanzig. 2022. p.105)

Zur Erstellung des Liniendiagramms werden zunächst ähnlich wie beim Säulendiagramm über jedem Punkt der x-Achse die Datenwerte eingetragen, allerdings nicht in Form einer Säule, sondern nur in Form eines Punktes. Diese Punkte werden dann mit einer Linie verbunden. Die Punkte können anschließend angezeigt oder auch wieder weggelassen werden – Informationsträger ist ja jetzt die «Daten-Linie».

In Liniendiagrammen können relativ große Datenmengen auf relativ kleinem Raum abgebildet werden. Beachte aber, dass nach ca. 5-7 Linien die Übersichtlichkeit verloren geht. Zur Unterscheidung der einzelnen Linien verwendet man am besten unterschiedliche Farben (Achtung auf Personen mit Farbenfehlsichtigkeiten!), falls mit einem Schwarz-Weiß-Ausdruck zu rechnen ist, dann verschiedene Graustufen. «Gestrichelte» oder «gepunktete» schwarze Linien werden hingegen heutigen ästhetischen Ansprüchen nicht mehr ganz gerecht.

Ein artverwandtes Diagramm zum Liniendiagramm ist das **Flächendiagramm**. Dabei wird die Fläche zwischen der Linie des Liniendiagramms und der x-Achse

noch «ausgemalt». Beispiel: Abb.2.11.

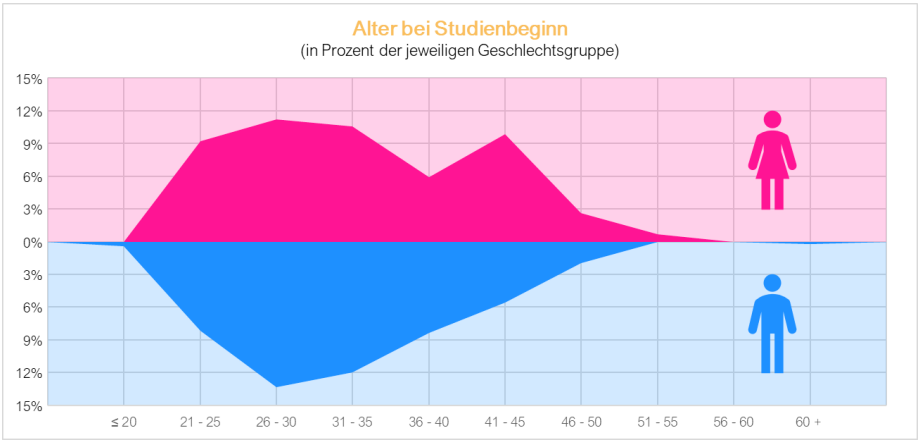


Abb. 2.11: Altersverteilung der WIBA-Studierenden bei Studieneintritt in Prozent der jeweiligen Geschlechtsgruppe

Linien- und Flächendiagramme sind nur schlecht geeignet, wenn Daten klassifiziert wurden.

Punktdiagramm

Streng genommen können wir Liniendiagramme nur für stetige Daten verwenden, weil wir ja meistens gar nicht kontinuierlich gemessen haben. Vermutlich wurde zum Beispiel der Realwert des Euros nicht jede Minute eruiert, wie Abb.2.10 suggeriert, sondern nur an bestimmten Stichtagen (z.B. am Beginn jeden Jahres), und diese Messpunkte einfach durch Linien verbunden. Will man «auf Nummer sicher» gehen und wirklich nur die Daten darstellen, für die gemessene Werte vorliegen, verwendet man ein **Punktdiagramm**, wie in Abb.2.12 dargestellt.

Die hier genannten Diagrammtypen stellen nur einen kleinen Ausschnitt der Möglichkeiten dar. Es gibt unzählige «Unterarten» oder Mischformen, wie das Beispiel der Abb.2.13.

Eine interessante Übersicht über alle möglichen Diagramme findest du zum Beispiel auf der Seite www.data-to-viz.com.

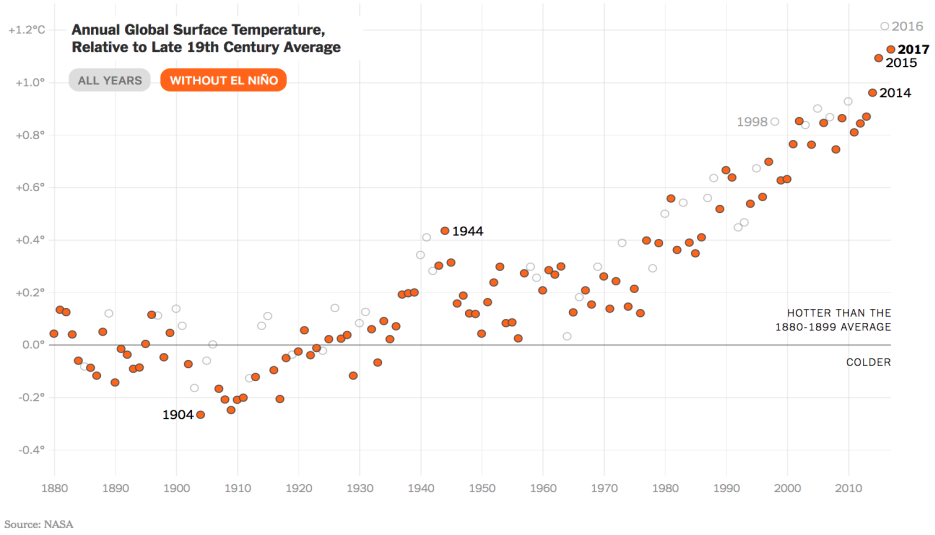


Abb. 2.12: Die durchschnittliche Temperatur an der Erdoberfläche von 1880 bis 2017 in Relation zur Durchschnittstemperatur der letzten 20 Jahre des 19. Jahrhunderts. (Bildquelle: The New York Times. 18.1.2018. http://www.nytimes.com/interactive/2018/01/18/climate/hottest-year-2017.html)

2.4 Hinweise für die Erstellung von Tabellen und Diagrammen

Computerprogramme wie MS Excel machen es ziemlich einfach, Grafiken zu erstellen und Daten zu visualisieren. Allerdings haben sie auch die Tendenz, mit (vermeintlichen) optischen Raffinessen zu übertreiben und die eigentliche Informationsdarstellung dem «Styling» unterzuordnen oder diese sogar zu verfälschen oder Betrachter:innen zumindest zu verwirren. Hier daher einige Tipps:

- Verschiedene Statistikprogramme bieten die oben genannten Diagramme und Histogramme auch in einer dreidimensionalen Ausprägung an. Dies kann eventuell dann Verwendung finden, wenn wir die statistische Verteilung zweier Merkmale zugleich darstellen wollen. In der Regel muss man aber darauf achten, dass durch den 3D-Effekt die Informationen auch verzerrt dargestellt werden können. Weniger ist oft mehr.
- Bei Histogrammen: Achtung auf unterschiedliche Klassenbreiten!
- Für alle Diagramme in einem Koordinatensystem: Die Ordinate (y-Achse) sollte ungefähr $2/3$ bis $3/4$ der Länge der Abszisse (x-Achse) haben. Sie sollte im Ursprung des Koordinatensystems mit dem Wert 0 beginnen, da es sonst zu irreführenden Maßstabsverzerrungen kommen kann.
- Vermeide unnötige «Dekorationen» oder Illustrationen im Hintergrund einer grafischen Darstellung sowie «Zierrahmen».

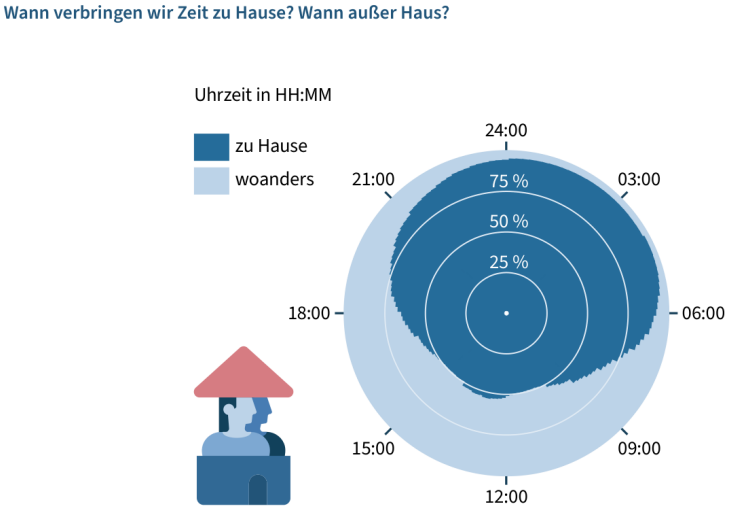


Abb. 2.13: Wann verbringen wir wo unsere Zeit?
(Quelle und Grafik: STATISTIK AUSTRIA, Zeitverwendungserhebung)

- Vergiss nicht: Jedes Diagramm braucht einen Titel, eine Beschriftung der Achsen und eine Beschriftung (möglichst direkt an Ort und Stelle – eine Legende ist nur die zweitbeste Lösung). Und einen Quellenverweis auf die Herkunft der Daten.
- Denke auch an eine ansprechende Schriftart und Farbgebung. Beachte dabei aber auch die Möglichkeit der Farbenfehlsichtigkeit mancher Menschen[7].
- Verwende nach Möglichkeit keine gedrehten oder schrägstehenden Beschriftungen (also keinen Text in anderer als horizontaler Ausrichtung), weder aus «künstlerischen» Gründen, noch um Platz zu sparen. Falls du dann zum Beispiel Bezeichnungen in einem Säulendiagramm nicht mehr unterbringen kannst, verwende – wenn nichts anderes dagegenspricht – Balkendiagramme.
- Fettschrift dient dazu, etwas hervorzuheben, weil es entweder eine Überschrift oder eine Kernaussage ist, oder aber, um es auf einem eingefärbten Hintergrund lesbarer zu machen. Wenn du *alles* fett schreibst, geht dieser Effekt verloren.
- Du musst nicht alle Farben verwenden, die dein Monitor oder Drucker schafft. Bedenke auch: Eventuell läuft deine Präsentation über einen kaputten Beamer oder deine Datei wird auf einem Schwarz-Weiß-Drucker gedruckt.
- Verwende bei Säulen- und Balkendiagrammen möglichst keine «Schatten» hinter den Säulen, außer du zeichnest von Hand auf ein Flipchart (falls das heute noch vorkommt). Schatten beinhalten keine besondere Information.

[7] siehe www.datylon.com/blog/data-visualization-for-colorblind-readers

- Bei Tabellen ist es nicht ratsam, alle möglichen Rasterlinien (Zellrahmen) einzuzeichnen; besser z.B. nur nach jeder dritten oder fünften Zeile eine Linie zeichnen (und nach der ersten Zeile mit den Spaltenüberschriften).
- Ganze Zahlen sollten in Tabellen nicht linksbündig, sondern rechtsbündig gesetzt werden; Dezimalzahlen weder links- noch rechtsbündig, sondern am Komma ausgerichtet.

Und das Wichtigste: Verwende immer *aktuelle* Daten aus zuverlässigen Quellen.
Und überleg immer *zuerst*, welche *Frage* du eigentlich mit deiner Visualisierung beantworten möchtest. Dann wähl den Diagrammtyp aus.
Nicht umgekehrt.

3

Kennwerte empirischer Häufigkeitsverteilungen

Im letzten Kapitel haben wir die Daten tabellarisch oder grafisch dargestellt und dabei auch den Begriff der *Häufigkeit* kennengelernt und *Häufigkeitstabellen* zusammengestellt. Wir wollen uns nun die Daten in so einer Häufigkeitstabelle etwas näher anschauen und insbesondere etwas über die Art und Weise sagen, wie die Daten *verteilt* sind; auf Seite 31 haben wir dafür bereits den Begriff **Häufigkeitsverteilung** verwendet. Mit der etwas genaueren Bezeichnung als *empirische*[1] *Häufigkeitsverteilung* wollen wir betonen, dass es sich bei unseren Daten und ihrer Verteilung um beobachtete oder gemessene Werte (einer Stichprobe) handelt und nicht um eine «theoretische» Verteilung, nach der wir die Daten *modellieren*[2].

Wie können wir nun die Daten und ihre Verteilung charakterisieren und durch aussagekräftige Kennwerte zusammenfassen?

Es gibt eine Reihe von statistischen Kennwerten, mit denen wir die Verteilung der Daten einer Stichprobe beschreiben können. Die bekanntesten sind das *arithmetische Mittel*, mit dem wir alle Daten einer Stichprobe durch einen einzigen Wert repräsentieren wollen (nämlich jenen, der «im Zentrum» der Daten steht), sowie die *Standardabweichung*, mit der wir angeben, wie stark die einzelnen Werte im Schnitt von diesem Mittel (und voneinander) abweichen. Statistiker:innen sagen, das arithmetische Mittel beschreibt die *Lage* einer Verteilung, die Standardabweichung ihre *Streuung*. Es gibt noch weitere Lage- und Streuungsmaße, und einige davon wollen wir hier angeben.

Zuvor können wir auch noch jedem beobachteten oder gemessenen Wert eine *Rangzahl* zuordnen:

[1] zum griech. $\varepsilon\mu\pi\varepsilon\iota\rho\omega\varsigma$ (empeiros): etwas aus der Erfahrung kennen

[2] Solche theoretischen Modelle werden wir in einem späteren Kapitel auch noch kennen lernen.

Rangzahl

Bevor wir zusammenfassende Kennwerte bilden, ist es manchmal für einen ersten Überblick sinnvoll, die Daten entsprechend der Größe ihrer Merkmalswerte zu sortieren und entsprechend des Ranges, den sie dabei einnehmen, zu indizieren. Üblicherweise wird dabei mit dem kleinsten Wert begonnen und dieser mit x_1 bezeichnet. Sind zwei (oder mehr) Daten gleich groß, erhalten sie nicht denselben Rang (es gibt also keine Ex aequo-Plätze), sondern der Index wird einfach weiter fortlaufend gezählt[3].

Damit Daten in eine Rangordnung gebracht werden können, müssen sie nicht unbedingt numerisch, aber zumindest Ordinaldaten sein.

Achtung bei Daten, bei denen die Chronologie eine Rolle spielt, also die zeitliche Reihenfolge, in der sie aufgetreten sind: Diese dürfen natürlich nicht der Größe nach geordnet werden, sondern müssen ihre ursprüngliche Abfolge beibehalten und werden auch in dieser Reihenfolge indiziert.

3.1 Lagekennwerte empirischer Häufigkeitsverteilungen

Minimaler und maximaler Wert

Man kann für jede Stichprobe, in der die Elemente zumindest ordinalskaliert sind, einen **Maximalwert** x_{max} und einen **Minimalwert** x_{min} angeben. Sind die Daten entsprechend ihrer Rangzahl indiziert, so ist

$$x_{\min} = x_1 \tag{3.1}$$

$$x_{\max} = x_n \tag{3.2}$$

Beispiel 3 *Gegeben ist die Kaffeemenge (Tassen pro Tag), die von einer Testperson innerhalb von vierzehn Tagen während des Lesens und Durcharbeitens dieser Unterlagen konsumiert wurde. Gesucht sind der größte und kleinste Wert.*

[3] Diese Vorgangsweise gilt nicht bei der Berechnung der Rangkorrelation, siehe Seite 88

Sowohl in *MS Excel* als auch in *LibreOffice Calc* können wir das Minimum mit dem Befehl `=MIN(Zahl1;Zahl2;...)` berechnen, wobei die Argumente `(Zahl1;Zahl2;...)` die Zahlen sind, von denen wir den kleinsten Wert wissen wollen.
Der entsprechende Befehl für das Maximum lautet in beiden Programmen: `=MAX(Zahl1;Zahl2;...)`
In *R* lauten die Funktionen `min(x)` bzw. `max(x)`, wobei als Argument (also für `x`) der Vektor eingesetzt wird, der die Daten enthält, deren Minimum bzw. Maximum wir ausrechnen wollen[a].

[a] Zur Syntax in *R* siehe die Lehrveranstaltung *PR122: Einführung in die Programmierung*, oder auch direkt in *R* durch Aufruf der «Hilfe» zu den einzelnen Funktionen, z.B. mit `?max()` (Also Eingabe eines Fragezeichens und des Funktionsnamens, ohne Argumente in den Klammern)

1, 3, 1, 3, 2, 2, 5, 4, 3, 2, 3, 4, 6, 3

Der Einfachheit halber ordnen wir die Daten unserer Stichprobe der Größe nach:

1	1			
2	2	2		
3	3	3	3	3
4	4			
5				
6				

Damit lassen sich Minimum und Maximum ganz leicht angeben:

$$x_{\min} = 1$$
$$x_{\max} = 6$$

Um das Beispiel in R zu rechnen, müssen wir zunächst einen Vektor bilden, der die Ausgangsdaten enthält. Wir geben ihm den Namen `bsp3`:

```
bsp3 <- c(1, 3, 1, 3, 2, 2, 5, 4, 3, 2, 3, 4, 6, 3)
```

Dann können wir das Minimum und Maximum ausrechnen und erhalten:

```
min(bsp3)
[1] 1
max(bsp3)
[1] 6
```

Modalwert

Der **Modalwert** (auch *Modus* genannt) ist jener Wert, der in einer Stichprobe am häufigsten vorkommt. In der Stichprobe $(1, 1, 3, 5, 6, 6, 6)$ ist zum Beispiel der Modalwert 6, weil der 6er dreimal vorkommt und keine andere Zahl an diese Häufigkeit herankommt.

Es kann auch mehr als einen Modalwert geben[4]. In $(1, 1, 1, 1, 3, 5, 5, 5, 5, 6)$ zum Beispiel gibt es zwei Modalwerte: 1 und 5. Beide kommen viermal vor.

Gibt es nur einen einzigen Modalwert, so spricht man auch von einer *unimodalen* Verteilung und bezeichnet den Modalwert selbst als *häufigsten Wert* oder auch als *wahrscheinlichsten Wert*: Wenn wir uns das Beispiel der Daten aus Abb.2.6 ansehen und uns 2019 ein Freund erzählt hätte, dass er gestern einen in Österreich produzierten Fisch gegessen hat, dann war das am wahrscheinlichsten eine Regenbogen- oder Lachsforelle – denn das ist der Modalwert dieser Verteilung.

Modalwerte können wir sowohl für qualitative als auch für quantitative Daten angeben. Er ist für alle Skalenniveaus möglich.

Beispiel 4 *Die folgende Tabelle enthält Ortsnamen, die in Österreich mehrfach vorkommen. Welches ist der häufigste Wert?*

Name	Anzahl	Name	Anzahl
Au	57	Aigen	37
Berg	41	Grub	50
Hart	38	Hof	31
Moos	40	Reith	36
Straß	31	Winkl	30

Tabelle 3.1: Die beliebtesten Ortsnamen Österreichs

Bei diesem einfachen Beispiel können wir das Ergebnis gleich durch einen Blick auf die Tabelle herauslesen. Wir könnten die Häufigkeiten auch in einem Diagramm darstellen und dann schauen, welches die höchste Säule ist. In jedem Fall kommen wir zum Ergebnis: Der häufigste in Österreich vorkommende Ortsname ist Au *(nämlich* $57\times$*).*

*(**Hinweis:** Beachte, dass der Modalwert in diesem Beispiel* Au *ist, und nicht 57.)*

Aufgabe 1 Gib zur Stichprobe des Beispiels 3 den oder die Modalwert(e) an.

[4]Wenn man die Bezeichnung *Modus* bevorzugt heißt die Mehrzahl: *Modi*.

Der *Excel*-Befehl für den Modalwert lautet
`=MODUS.VIELF((Zahl1;Zahl2;...)`. Gibt es mehrere Modalwerte, dann werden mit diesem Befehl auch alle zurückgegeben – das macht es aber notwendig, dass die Funktion als so genannte «Arrayformel» eingegeben werden (Siehe dazu die Hilfe-Funktion von Excel).
LibreOffice Calc hat einen ähnlichen Befehl:
`=MODALWERT(Zahl1;Zahl2;...)`. Gibt es hier mehrere Modalwerte, wird nur der kleinste von ihnen zurückgegeben und die anderen ignoriert!
Sowohl *Excel* als auch *Calc* setzen voraus, dass mindestens ein Wert der Stichprobe mindestens zweimal vorkommt, ansonsten gibt es eine Fehlermeldung.
In *R* gibt es keinen vordefinierten Befehl, um den Modalwert auszurechnen.

Mittelwerte

Der **arithmetische Mittelwert** (auch: das *arithmetische Mittel*) ist ein sehr gebräuchliches Maß für die Angabe des «Durchschnitts» der Verteilung von numerischen Daten. Dabei wird sprachlich die Spezifizierung «arithmetisch»[5] auch meist weggelassen und nur vom *Mittelwert* oder *Mittel* gesprochen.

Mathematisch ist das arithmetische Mittel der Quotient der Summe der Beobachtungswerte dividiert durch die Anzahl der Beobachtungswerte:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.3}$$

In *Excel* und auch in *LibreOffice Calc* wird das arithmetische Mittel mit dem Befehl `=MITTELWERT(Zahl1;Zahl2;...)` berechnet.
In *R* lautet der Befehl `mean(x)`.

Beispiel 5 *Das arithmetische Mittel der Daten aus Beispiel 3 beträgt:*

$$\bar{x} = \frac{1+3+1+3+2+2+5+4+3+2+3+4+6+3}{14} = \frac{42}{14} = 3$$

was wir auch aus R erhalten:

[5] griech. $\alpha\rho\iota\theta\mu\eta\tau\iota\kappa o\varsigma$ (arithmetikos) = im Zählen oder Rechnen geschickt

```
mean(bsp3)
[1] 3
```

Liegen die Daten in Form einer Häufigkeitstabelle vor, so erhalten wir den arithmetischen Mittelwert aus:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{m} \left(f_j \cdot x_j \right)$$

mit: (3.4)

$m \ldots$ Anzahl der unterschiedlichen Merkmalswerte

$f_j \ldots$ Häufigkeit des Auftretens dieses Merkmalswertes

Beispiel 6 *Die Daten aus Beispiel 3, in einer Häufigkeitstabelle angegeben:*

x	f	$f \cdot x$
1	2	2
2	3	6
3	5	15
4	2	8
5	1	5
6	1	6
	$n = 14$	$\Sigma = 42$

Aus n und der Summe rechts unten können wir nach Formel 3.4 das arithmetische Mittel berechnen:

$$\bar{x} = \frac{42}{14} = \underline{\underline{3}}$$

Wenn wir unsere Daten in Klassen eingeteilt haben, ist die Bildung des arithmetischen Mittels nicht mehr so einfach, weil wir ja die einzelnen Merkmalswerte, die wir für die Mittelwertsberechnung benötigen, nicht mehr zur Verfügung haben. Wir verwenden dann für das arithmetische Mittel die *Klassenmitten* als Eingangswerte in Formel 3.4. Die Klassenmitten sind jene Werte, die genau in der Mitte zwischen oberer und unterer Klassengrenze liegen.

Das arithmetische Mittel lässt sich dann nach 3.5 berechnen:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{m} \left(f_j \cdot x'_j \right)$$

mit: (3.5)

$m \ldots$ Anzahl der Klassen

$f_j \ldots$ Häufigkeit der Elemente in der j-ten Klasse

$x'_j \ldots$ Klassenmitte der j-ten Klasse

Beispiel 7 *Zur Berechnung des arithmetischen Mittels der Stichprobe aus Tabelle 2.2 ergänzen wir noch die jeweiligen Klassenmitten. Dabei unterstellen wir der letzten, an sich offenen Klasse, die gleiche Breite wie allen anderen:*

j	Klassengrenzen	Klassenmitte x'_j	f_j	$f \cdot x'_j$
1	$160 \leq x < 165$	162.5	1	162.5
2	$165 \leq x < 170$	167.5	3	502.5
3	$170 \leq x < 175$	172.5	4	690.0
4	$175 \leq x < 180$	177.5	8	1 420.0
5	$180 \leq x < 185$	182.5	3	547.5
6	$185 \leq x < 190$	187.5	4	750.0
7	$190 \leq x < 195$	192.5	0	0.0
8	$195 \leq x$	197.5	1	197.5
			$n = 24$	$\Sigma = 4\,270.0$

Tabelle 3.2: Häufigkeitstabelle einer Stichprobe mit klassifizierten Merkmalswerten

Das arithmetische Mittel ist dann

$$\bar{x} = \frac{4\,270}{24} = \underline{\underline{177.9}}$$

Hätten wir in obigem Beispiel nicht nur die klassierten Werte, sondern die Originaldaten zur Verfügung, würde vermutlich nicht genau 177.9 herauskommen. Aber das stört uns als Statistiker:in nicht wirklich. Statistik ist nicht Buchhaltung. Es geht eher darum, Modelle zu finden, wie die (messbare) Welt um uns herum *vermutlich* aussieht[6].

Neben dem arithmetischen Mittel sind in unseren Anwendungen manchmal auch das *gewichtete arithmetische Mittel* und das *geometrische Mittel* von Bedeutung:

Das **gewichtete arithmetische Mittel** wird verwendet, wenn Werte mit unterschiedlicher «Wichtigkeit» in die Berechnung des Mittels einfließen soll. Du erhältst dann unterschiedliche *Gewichte*, also Faktoren, mit denen sie multipliziert werden. Anwendung findet das zum Beispiel, wenn wir drei Noten haben (Bachelorarbeit 1, Bachelorarbeit 2 und Bachelorprüfung), für eine Gesamtbeurteilung diese drei Noten aber prozentuell nicht gleich einfließen sollen, sondern zum Beispiel im Verhältnis 20 : 20 : 60. Für die Berechnung des gewichteten arithmetischen Mittels wird dann jede Note vor dem Addieren mit ihrem Gewicht multipliziert (also mit 20 oder 60). Dividiert wird dann nicht durch die

[6] Daher verwenden wir im Zusammenhang mit Zufallsvariablen manchmal auch das Fremdwort *Stochastik*, aus dem griech. στοχαστικός (stochastikos) = *im Vermuten geschickt*

Anzahl der Elemente (in unserem Beispiel also nicht durch 3), sondern durch die Summe der Gewichte (also durch 100).

Etwas allgemeiner können wir angeben[7]:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i} \tag{3.6}$$

In *Excel* und *LibreOffice Calc* kann man ein gewichtetes arithmetisches Mittel mithilfe der Befehle `SUMMENPRODUKT` und `SUMME` errechnen (indem man die Summe aus den Produkten zwischen den Einzelwerten und ihren Gewichten durch die Summe der Gewichte dividiert).
In *R* gibt es einen direkten Befehl dafür: `weighted.mean(x)`

Das **geometrische Mittel** benötigt man zur Mittelung von Steigerungsfaktoren, zum Beispiel wenn sich bei der Zinseszinsrechnung der Zinsfaktor jährlich ändert und du einen durchschnittlichen Zinsfaktor angeben willst.

Das geometrische Mittel aus n Zahlen ist die n-te Wurzel des Produkts der n Zahlen:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{3.7}$$

Dabei ist zu beachten, dass das geometrische Mittel nur für positive Zahlen definiert ist. Es müssen also alle Elemente der Stichprobe durch numerische Werte ≥ 0 repräsentiert sein.

Der Umgang mit dem geometrischen Mittel ist nicht immer einfach, wie wir an folgendem Beispiel sehen.

Beispiel 8 *(Wir rechnen dieses Beispiel mit Absicht mit etwas übertriebenen, realitätsfernen Zahlen, damit wir den mathematischen Effekt sehen):*

Nehmen wir an, jemand legt 1 000 EUR in einer Veranlagungsform an, die im ersten Jahr 40% *Zinsen erbringt. Im zweiten Jahr gibt es nur noch* 10%*. Welcher durchschnittlichen jährlichen Verzinsung entspricht das?*

[7] Das w in obiger Formel steht für das englische Wort *weight* = Gewicht

Zunächst einmal rechnen wir uns das Endkapital aus:
Im ersten Jahr ergibt die Verzinsung ein Kapital von $1\,000 \cdot 1.4 = 1\,400$.
Im zweiten Jahr dann: $1\,400 \cdot 1.1 = 1\,540$ EUR.

Der durchschnittliche Zinsfaktor beträgt dann entsprechend Formel 3.7

$$\bar{x}_g = \sqrt[2]{1.4 \cdot 1.1} = \underline{\underline{1.24097}}$$

Zur Probe wenden wir jetzt jedes Jahr diesen durchschnittlichen Zinsfaktor an und erhalten tatsächlich:

$$(1\,000 \cdot 1.24097) \cdot 1.24097 = 1.540 \; EUR$$

(Hinweis: *Den genauen Wert erhältst du nur, wenn du statt des gerundeten Wertes* 1.24097 *mit dem genauen Wert* 1.240967365 . . . *rechnest, der beim Wurzelziehen herauskommt).*

Hier jetzt noch zwei Möglichkeiten, wie man dieses Beispiel ***falsch*** *rechnen könnte:*

Wenn du statt des geometrischen Mittels das arithmetische Mittel verwendest, erhältst du $\frac{1.4+1.1}{2} = 1.25$ *und damit dann:* $(1000 \cdot 1.25) \cdot 1.25 = 1562.5$ EUR.

Und wenn du statt der Zinsfaktoren 1.4 *und* 1.1 *die Zinssätze* 40% *und* 10% *in die Formel einsetzt (also die Zahlenwerte* 0.4 *und* 0.1*), erhältst du* $\bar{x}_g = \sqrt{0.4 \cdot 0.1} = 0.2$, *was überhaupt nur ein Endkapital von* 1.440 EUR *ergäbe.*

Du musst also in die Formel für das geometrische Mittel immer den ***Veränderungsfaktor*** *einsetzen, nicht den prozentuellen Veränderungswert!*

In *Excel* und auch in *LibreOffice Calc* wird das geometrische Mittel mit dem Befehl `=GEOMITTEL(Zahl1;Zahl2;...)` berechnet.
In *R* gibt es zwar keine vordefinierte Funktion dafür, wir können aber mathematisch ein wenig «tricksen» und das geometrische Mittel der Daten, die sich in `x` befinden, mit `exp(mean(log(x)))` ausrechnen.

Im Übrigen gilt: Der Wert des geometrischen Mittels ist immer kleiner als der Wert des arithmetischen Mittels derselben Werte[8].

[8] Für zwei positive Zahlen x und y lässt sich das relativ einfach zeigen, für den allgemeinen Fall von n Zahlen ist aber ungleich komplizierter und wir werden das an dieser Stelle nicht *beweisen*. Wir verlassen uns einfach darauf, dass Mathematiker:innen hier ganze Arbeit geleistet haben, siehe zum Beispiel de.wikipedia.org/wiki/Ungleichung_vom_arithmetischen_und_geometrischen_Mittel

Das letzte Mittelmaß, das wir noch anschauen wollen, ist das **harmonische Mittel**. Es wird verwendet, um den Mittelwert von Verhältniszahlen zu berechnen:

$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{3.8}$$

Ein einfaches Beispiel für die Verwendung des harmonischen Mittels ist die Berechnung der durchschnittlichen Geschwindigkeit für den Hin- und Rückweg einer bestimmten Strecke. Geschwindigkeiten sind ja Verhältniszahlen (nämlich das Verhältnis Weg:Zeit). Fährt man zum Beispiel mit 100 km/h von Scheibbs nach Hamburg (Entfernung = 1000 km), aber mit nur 80 km/h retour, so ist man den Gesamtweg von 2000 km nicht mit durchschnittlich $(100 + 80)/2 = 90$ km/h gefahren, sondern mit $88,89$ km/h, also jenen Wert, der dem harmonischen Mittel von 100 und 80 entspricht[9].

Ein weiteres Beispiel ist der so genannte *F1-Score*, der für die Beurteilung der Leistungsfähigkeit eines Large Language Modells verwendet wird. Der F1-Score gibt ein mittleres Maß für die beiden Kennzahlen *Precision* und *Recall* an. Diese beiden Kennzahlen sind selbst Verhältniszahlen, daher muss für den Mittelwert das harmonische Mittel verwendet werden.

In *Excel* und in *LibreOffice Calc* wird das harmonische Mittel mit dem Befehl `=HARMITTEL(Zahl1;Zahl2;...)` berechnet.
In *R* gibt es wieder keine vordefinierte Funktion dafür. Wir können uns aber zunutze machen, dass das harmonische Mittel aus mehreren Zahlen der Kehrwert des arithmetischen Mittels der Kehrwerte dieser Zahlen ist (OK, wer diesen Satz jetzt nicht verstanden hat, wendet einfach Formel 3.8 an...).

Beispiel 9 *Bei der Beurteilung der Leistungsfähigkeit von KI-Systemen lässt man das System eine bestimmte Anzahl von Vorhersagen machen, von denen man selbst die richtige Antwort kennt. Angenommen, es gibt nur zwei Antwortmöglichkeiten, zum Beispiel: Dieses Bild zeigt eine Katze oder nicht. Dann zählt man mit:*

$TP =$ True positive: Eine Katze wird richtigerweise als Katze erkannt
$TN =$ True negative: Es wird richtigerweise erkannt, dass keine Katze zu sehen ist
$FP =$ False positive: Es wird eine Katze erkannt, obwohl keine zu sehen ist
$FN =$ False negative: Es wird keine Katze erkannt, obwohl eine zu sehen ist

[9] Was sich leicht nachprüfen lässt: Für den Hinweg benötigt man 10 Stunden, für den Rückweg 12,5 Stunden, insgesamt also für 2000 km 22,5 Stunden, was $2000/22,5 = 88,89$ km/h ergibt.

Damit können folgende Kennwerte berechnet werden:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \text{harmonisches Mittel aus Precision und Recall}$$

Sei: $TP = 642$, $TN = 279$, $FP = 277$, $FN = 85$.

Welchen Wert hat der F1-Score?

$$\text{Precision} = \frac{642}{642 + 277} = 0,699$$

$$\text{Recall} = \frac{642}{642 + 85} = 0,883$$

$$\text{F1-Score} = \bar{x}_h = \frac{2}{\frac{1}{0,699} + \frac{1}{0,883}} = \underline{\underline{0,780}}$$

Quantile

Mittelwerte sind nicht das einzige Maß, die sich zur Angabe eines «Durchschnitts» eignen. Wir können unsere Daten auch der Größe nach ordnen (falls wir das nicht schon ohnehin gemacht haben) und die geordnete Beobachtungsreihe in zwei Teile zerlegen. Gesucht ist jener Wert, der definiert, wo wir die Stichprobe «durchschneiden» müssen, damit ein bestimmter Anteil der Beobachtungen unterhalb dieser Trennlinie liegt.

Beispiel 10 *Die schwedische Schirennläuferin Frida Hansdotter ist in ihrer Karriere (2005 - 2019) bei insgesamt 130 Slalom-Rennen im Weltcup oder bei Olympischen Spielen oder Weltmeisterschaften gestartet. Dafür hat sie eine Gesamtzeit von 3 Stunden 24 Minuten und 41,20 Sekunden benötigt. (Das inkludiert auch die Zeit der 20 Rennen, bei denen sie nur einen Durchgang ins Ziel geschafft hat (in 12 hat sie sich nicht für den 2. Durchgang qualifiziert, in 6 ist sie im 2. Durchgang ausgeschieden), nicht aber die 8 Rennen, bei denen sie bereits im 1. Durchgang ausgeschieden ist).*
In der folgenden Tabelle sind die Platzierungen der bei den letzten 50 von ihr in Angriff

date	place	position	date	place	position
16.02.2018	Pyeongchang	1	28.11.2015	Aspen	3
10.01.2017	Flachau	1	13.12.2014	Are	3
29.12.2015	Lienz	1	02.02.2019	Maribor	4
13.01.2015	Flachau	1	05.01.2019	Zagreb	4
28.01.2018	Lenzerheide	2	17.11.2018	Levi	4
07.01.2018	Kranjska Gora	2	11.11.2017	Levi	4
15.11.2017	Levi	2	03.01.2017	Zagreb	4
15.01.2016	Flachau	2	11.12.2016	Sestriere	4
13.12.2015	Are	2	04.01.2015	Zagreb	4
29.11.2015	Aspen	2	29.12.2014	Kühtai	4
21.03.2015	Meribel	2	16.03.2019	Soldeu	5
14.02.2015	Beaver Creek	2	16.02.2019	Are	5
30.11.2014	Aspen	2	08.01.2019	Flachau	5
22.12.2018	Courchevel	3	26.11.2017	Killington	5
25.11.2018	Killington	3	05.01.2016	Santa Caterina	5
17.03.2018	Are	3	15.02.2016	Crans Montana	6
10.03.2018	Ofterschwang	3	14.03.2015	Are	6
09.01.2018	Flachau	3	09.03.2019	Spindleruv Mlyn	7
03.01.2018	Zagreb	3	29.12.2016	Semmering	7
28.12.2017	Lienz	3	22.02.2015	Maribor	9
18.03.2017	Aspen	3	27.11.2016	Killington	10
18.02.2017	St. Moritz	3	06.03.2016	Jasna	10
08.01.2017	Maribor	3	29.12.2018	Semmering	DNF2
19.03.2016	St. Moritz	3	11.03.2017	Squaw Valley	DNQ2
12.01.2016	Flachau	3	12.11.2016	Levi	DNF2

Tabelle 3.3: Frida Hansdotters Platzierungen in ihren letzten 50 Slaloms im Weltcup, bei Weltmeisterschaften oder Olympischen Spielen. Für alle im folgenden durchgeführten Berechnungen, die sich auf diese Tabelle beziehen, werden wir nur $n = 47$ Werte berücksichtigen und DNQ2 und die beiden DNF2 weglassen.

genommenen Rennen gegeben. Die Daten sind nicht chronologisch sondern der Größe nach geordnet, Ordnungskriterium ist die Platzierung.

Wir können nun aus dieser geordneten Liste zum Beispiel feststellen, dass Frida Hansdotter in ihren 27 besten Rennen (der Saisonen 2014/15 bis 2018/19) nie schlechter als Dritte geworden ist oder in 40 aus 50 Rennen nie schlechter als Fünfte. In 94% der Rennen hat sie mindestens den 10. Platz erreicht (in den restlichen 3 Rennen gar keine Platzierung).

Etwas systematischer und «statistischer» betrachtet besteht unsere Aufgabe darin, eine geordnete Beobachtungsreihe so in zwei Teile zu zerlegen, dass ein bestimmter Prozentsatz der Daten vom Rest getrennt wird. Wir nennen die Stelle, an der wir diesen Trennstrich einziehen, das **p-Quantil** der Verteilung, wobei p den anteilmäßigen Umfang der abgeteilten Daten angibt.

p kann zwischen 0 und 1 liegen (bzw. zwischen 0% und 100%).

Das p-Quantil ist nun definiert als jener Wert, für den gilt: $(n \cdot p)$ aller Stichprobenelemente sind kleiner oder gleich dem p-Quantil und $(n \cdot (1-p))$ aller Elemente größer als das p-Quantil. Haben wir zum Beispiel 200 der Größe nach geordnete Daten und suchen das Quantil für $p = 0.25$, dann ist das jener Wert, bei dem unsere gesamte Stichprobe so in zwei Teile geteilt wird, dass 50 Werte unterhalb des Quantils liegen, und 150 darüber.

Wie finden wir nun den konkreten Wert des p-Quantils einer empirischen Häufigkeitsverteilung?

Um es vorweg zu nehmen: Die Bestimmung des exakten Wert ist ein wenig kompliziert: Zunächst müssen wir die zum p-Quantil zugehörige Rangzahl bestimmen:

$$i_p = p(n-1) + 1 \tag{3.9}$$

Der Wert, der an der i_p.ten Stelle liegt, ist dann das gesuchte Quantil.

Wenn i_p keine ganze Zahl ist (was ziemlich oft vorkommt), muss zwischen den Werten an der Stelle $int(i_p)$ und $(int(i_p)+1)$ linear interpoliert werden:

$$x_p = x_{int(i_p)} + (i_p - int(i_p))(x_{int(i_p)+1} - x_{int(i_p)}) \tag{3.10}$$

wobei die Funktion *int* die *Integer-Funktion* ist, das ist jene Funktion, die nur den ganzzahligen Anteil einer Zahl zurückgibt.

Zum Glück hat *Excel* die beschriebene Prozedur zur Berechnung des Quantils eingebaut: Mit der Funktion `=QUANTIL.INKL(Array;p)` (mit den Daten der Stichprobe im Datenbereich `Array`) kann es ziemlich einfach berechnet werden, ohne sich über ein i_p, eine Interpolation oder die Integer-Funktion Gedanken machen zu müssen.
In *LibreOffice Calc* heißt der Befehl: `=QUANTIL(Daten;p)`.
In *R*: `quantile(x,p)`

Beispiel 11 *Aus den Daten der Tabelle 3.3 können wir ausrechnen:* $x_{0.25} = 2$, $x_{0.5} = 3$ *und* $x_{0.75} = 4.5$

Näherungsverfahren zur Berechnung von Quantilen

Die oben beschriebene, etwas komplizierte Prozedur werden wir nur verwenden, wenn wir ein (Rechen-)Programm verwenden können. Für die Bestimmung

des Quantils «von Hand» reicht es, wenn wir folgendes **Näherungsverfahren** verwenden:

Um das p-Quantil einer Stichprobe mit n Elementen näherungsweise auszurechnen, multiplizieren wir zunächst

$$k = n \cdot p \tag{3.11}$$

Wenn k eine ganze Zahl ist, dann ist unser p-Quantil der Mittelwert zwischen x_k und dem nächsten Beobachtungswert x_{k+1}:

$$x_p = \frac{x_k + x_{k+1}}{2} \tag{3.12}$$

Wenn k nicht ganzzahlig ist, dann runden wir es auf die nächste ganze Zahl auf. Der Wert, der an dieser Stelle steht, ist dann das gesuchte p-Quantil:

$$\begin{aligned} &x_p = x_{\lceil k \rceil} \\ &\text{mit: } \lceil k \rceil = \text{kleinste ganze Zahl größer oder gleich } k \end{aligned} \tag{3.13}$$

Wir wollen das nun in einem Beispiel für einige p-Werte durchrechnen:

Gegeben sind in Tabelle 3.4 für alle[10] Nachbarländer der Nachbarländer Österreichs die Anzahl der Ausbildungsjahre, die ein Kind im Schuleintrittsalter in diesem Land im Schnitt vor sich hat. Die Werte wurden auf halbe Jahre gerundet. Die Tabelle ist bereits der Größe nach geordnet.

Beispiel 12 *Gib für die Daten der Tabelle 3.4 das Quantil für $p = 25\%$ an.*

$$k = n \cdot p = 20 \cdot 0.25 = 5$$

Das ist eine ganze Zahl, daher müssen wir den Mittelwert aus x_5 und x_6 bilden:

$$\frac{1}{2}(x_5 + x_6) = \frac{1}{2}(14 + 14.5) = 14.25$$

Das ist unser gesuchtes 0.25-Quantil, was wir auch so schreiben: $\underline{\underline{x_{0.25} = 14.25}}$.

(Der exakte Wert nach Formel (3.10) bzw. mit EXCEL ausgerechnet ergibt 14.375)

[10] Vatikanstadt ist in dieser Aufzählung nicht angegeben, weil es dort keine Schulen gibt. Kinder von Einwohnern des Vatikans gehen in der Regel im benachbarten Italien zur Schule.

Land	Schuljahre	Land	Schuljahre
Liechtenstein	12	Tschechien	15.5
San Marino	12.5	Ungarn	15.5
Luxemburg	13.5	Schweiz	15.5
Serbien	13.5	Frankreich	16
Kroatien	14	Italien	16
Rumänien	14.5	Belgien	16.5
Slowakei	14.5	Deutschland	16.5
Ukraine	15	Dänemark	17
Polen	15	Niederlande	17
Österreich	15.5	Slowenien	17

Tabelle 3.4: Expected Years of Schooling of children in years: Number of years of schooling that a child of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates persist throughout the child's life. Source: UNESCO Institute for Statistics (2012), http://stats.uis.unesco.org

Beispiel 13 *Gib für die Daten der Tabelle 3.4 das Quantil für $p = 1/3$ an.*

$$k = n \cdot p = 20 \cdot 1/3 = 6.67$$

Das ist nicht ganzzahlig, daher runden wir auf die nächste ganze Zahl auf:

$$\lceil 6.67 \rceil = 7$$

Der an siebter Stelle stehende Wert ist gleich 14.5*, daher ist* $x_{0.33} = \underline{\underline{14.5}}$.

(Der exakte Wert nach Formel (3.10) ergibt 14.667*)*

Aufgabe 2 Gib für die Daten der Tabelle 3.4 das 0.65-Quantil an (berechnet nach obigem Näherungsverfahren) und vergleiche dein Ergebnis mit dem exakten Wert aus *R*.

Einige Quantile (mit einem bestimmten p-Wert) spielen in der Datenauswertung eine besondere Rolle, daher haben sie eigene Namen bekommen:

Mediane, Quartile und Ähnliches

Wir sind manchmal daran interessiert, die geordneten Daten nicht nur an einer bestimmten Stelle zu teilen, sondern gleich in q gleich große Gruppen zu unterteilen. q kann zum Beispiel 2 sein; wir teilen dann unsere Daten in zwei gleich

große Gruppen. Mit $q = 4$ bilden wir vier Gruppen, mit $q = 10$ zehn Gruppen, mit $q = 100$ hundert Gruppen. Um diese Teilungen zu bewerkstelligen, benötigen wir jeweils $(q - 1)$ Teilungspunkte (Mit einem Punkt können wir eine Datenmenge in 2 Gruppen teilen, mit 3 Punkten in vier, mit neun Punkten in 10 Gruppen und mit neunundneunzig Punkten in 100 Gruppen).

Sehen wir zunächst uns wieder die – der Größe nach geordneten – Daten aus Beispiel 3 an:

1 1 2 2 2 3 3 3 3 3 4 4 5 6

Sei zunächst $q = 2$, d.h. wir wollen in zwei Gruppen aufteilen und benötigen dazu $q - 1 = 1$ Teilungspunkt. Zur Bestimmung dieses Punktes bedienen wir uns der Quantile aus dem letzten Abschnitt und bestimmen das Quantil mit

$$p = \frac{q-1}{q} = \frac{1}{2} = 0.5 \tag{3.14}$$

Das ist mittlerweile ja ziemlich einfach für uns:

Beispiel 14

$k = 14 \cdot 0.5 = 7$ ganzzahlig, daher:
$x_{0.5} = \frac{x_7 + x_8}{2} = \frac{3+3}{2} = 3$

Das 0.5-Quantil teilt unsere Stichprobe in genau zwei Teile. Der «mittelste» Datenwert ist 3. Oberhalb und unterhalb liegen je 50% der Werte. Das 0.5-Quantil wird auch **Median** (oder *Zentralwert*) genannt. Der Median ist jener Wert, der von mindestens der Hälfte der Merkmalswerte nicht unterschritten wird. (*Mindestens* deshalb, weil es auch sein könnte, dass der Median ein Merkmalswert ist, der mehrfach vorkommt).

Wir können das auch in unseren Daten visualisieren und einzeichnen, wo wir die Stichprobe teilen müssen:

1 1 2 2 2 3 3 | 3 3 3 4 4 5 6

In Excel lautet der Befehl für den Median: `=MEDIAN(Zahl1;Zahl2;...)`.
In *R* gibt der Befehl `median(x)` den Median der Daten in `x` aus.

Für den Median können wir die Formeln 3.13 und 3.12 auch so angeben:

Wenn n eine gerade Zahl ist, dann ist der Median der Mittelwert zwischen dem Wert an der Stelle $\frac{n}{2}$ und dem nächsten Beobachtungswert an der Stelle $\frac{n}{2}+1$:

$$x_{0.5} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \tag{3.15}$$

Wenn n ungerade ist, dann ist der Median der Wert an der Stelle $\frac{n+1}{2}$:

$$x_{0.5} = x_{\frac{n+1}{2}} \tag{3.16}$$

Nachdem wir unsere Daten mit dem Median in zwei gleiche Hälften geteilt haben, wiederholen wir das und teilen die beiden Hälften wieder genau in der Mitte:

Für eine Unterteilung in $q = 4$ Teile benötigen wir $q - 1 = 3$ Teilungspunkte, die wir auch **Quartile** (oder *Viertelwerte*) nennen. Das **1. Quartil** ist das $\frac{1}{4}$-Quantil, das **2. Quartil** das $\frac{2}{4}$-Quantil und das **3. Quartil** das $\frac{3}{4}$-Quantil. Das 1. Quartil wird auch *unteres Quartil* genannt, das 3. Quartil *oberes Quartil* – und das 2. Quartil ist nichts anderes als der *Median* (siehe oben).

Oberhalb des *oberen Quartils* und unterhalb des *unteren Quartils* liegen je 25 % der Elemente, dazwischen die restlichen 50 %.

Wir verwenden wieder das Näherungsverfahren:

Beispiel 15 *Gib das nach dem Näherungsverfahren bestimmte 1., 2. und 3. Quartil der Daten aus Beispiel 3 an:*

1. Quartil:

$$\begin{aligned} &n = 14, \quad q = 4, \quad p = \tfrac{1}{4} = 0.25 \\ &k = 14 \cdot 0.25 = 3.5 \\ &\lceil 3.5 \rceil = 4 \\ &x_{0.25} = \boxed{2} \end{aligned}$$

2. Quartil:

$$\begin{aligned} &n = 14, \quad q = 4, \quad p = \tfrac{2}{4} = 0.5 \\ &k = 14 \cdot 0.5 = 7 \\ &x_{0.5} = \tfrac{x_7 + x_8}{2} = \boxed{3} \end{aligned}$$

3. Quartil:

$$n = 14, \quad q = 4, \quad p = \tfrac{3}{4} = 0.75$$
$$k = 14 \cdot 0.75 = 10.5$$
$$\lceil 10.5 \rceil = 11$$
$$x_{0.75} = \boxed{4}$$

Anm.: Wenn wir nicht das Näherungsverfahren verwenden, sondern zum Beispiel R, dann erhalten wir:

```
quantile(bsp3,0.25)
2

quantile(bsp3,0.5)
3

quantile(bsp3,0.75)
3.75
```

Manchmal wird ergänzend auch noch ein **0. Quartil** und ein **4. Quartil** angegeben: Das ist nichts anderes als das *Minimum* und das *Maximum*, die unsere Daten am Anfang und Ende «einrahmen», siehe Tab.3.5.

Quantil	**weitere Bezeichnungen**
0.00-Quantil	*0.Quartil* oder *Minimum*
0.25-Quantil	*1.Quartil* oder *unteres Quartil*
0.50-Quantil	*2.Quartil* oder *Median*
0.75-Quantil	*3.Quartil* oder *oberes Quartil*
1.00-Quantil	*4.Quartil* oder *Maximum*

Tabelle 3.5: Quartile und ihre synonymen Bezeichnungen

In *R* gibt der Befehl `summary(x)` die in Tab.3.5 angeführten Werte einer Stichprobe an, dazu auch noch das arithmetische Mittel. Sind zum Beispiel die Platzierungen aus Tab. 3.3 (ohne die letzten drei, nicht numerischen Werte DNF2 und DNQ2) im Vektor `frida` gespeichert, dann ergibt `summary(frida)` das Ergebnis:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	3.745	4.500	10.000

Aufgabe 3 Gegeben sei die Körpergröße der 7 Zwerge. Bestimme das Minimum, Maximum sowie das $1. - 3.$ Quartil sowohl näherungsweise als auch deren exakten Werte:

Name	Größe (in *cm*)
Doc	85
Grumpy	80
Happy	62
Sleepy	81
Bashful	70
Sneezy	80
Dopey	88

Aufgabe 4 Gegeben sei die Größe der 7 roten Zwerge, die innerhalb einer Entfernung von 10 Lichtjahren zur Erde liegen. Bestimme das Minimum, Maximum sowie das $1. - 3.$ Quartil sowohl «visuell» als auch deren exakten Werte:

Bezeichnung	Durchmesser (in *Tausend km*)
Luyten 726-8 A	195
Luyten 726-8 B	195
Proxima Centauri	196.4
Wolf 359	222.8
Barnards Stern	273
Ross 154	334.2
Lalande 21185	547.4

Neben Median und Quartilen sind manchmal auch noch Unterteilungen in $q = 10$ Gruppen (zu je 10%) interessant – wir nennen die Teilungspunkte dann **Dezile**, sowie jene mit $q = 100$ (also eine Einteilung in Gruppen zu je 1%), die so genannten **Perzentile**. Die Berechnung dürfte aber hoffentlich klar sein: Für das 8. Dezentil berechnen wir zum Beispiel das 0.8-Quantil, für das 5. Perzentil das 0.05-Quantil etc.

Anmerkungen zur Verwendung von Mittelwert, Median und Modalwert

Mittelwert und Median werden beide verwendet, um eine umfangreiche Datenmenge durch einen einzigen Wert möglichst gut zu repräsentieren. Im allgemeinen Sprachgebrauch sagen wir auch: wir suchen den *Durchschnitt*. Mittelwert und Median haben dabei unterschiedliche Eigenschaften, die sie – je nach Anwendungsfall – geeigneter erschienen lassen, diese Aufgabe zu erfüllen.

Sie zeigen zum Beispiel unterschiedliches *Resistenzverhalten* (Widerstandsfähigkeit) gegenüber Ausreißern:

Der Mittelwert ist sehr empfindlich gegenüber Ausreißern. Nachdem jeder einzelne Wert in seiner vollen Höhe in die Berechnung des Mittelwerts einfließt, kann jeder einzelne Wert $\bar{x}$ auch bedeutend verändern.

Der Median hingegen wird durch einzelne Ausreißer kaum verändert. Ändert sich ein Datenwert – egal um wie viel – so ändert der Median seinen Wert nur dann, wenn dieser Datenwert von der einen Hälfte der geordneten Daten in die andere Hälfte wandert.

Trittst du in Gehaltsverhandlungen mit deiner Chefin und nimmst einen «mittleren Wert» aus allen Gehältern innerhalb der Firma als Grundlage, dann verwende den arithmetischen Mittelwert, weil dann das überproportionale Gehalt deiner Chefin als «Ausreißer» den Mittelwert erhöhen wird.
Deine Chefin wird hingegen versuchen, den Median als Basis heranzuziehen, weil dann die Höhe ihres Gehalts keinen Einfluss hat.
Letztendlich ist es aber ziemlich wahrscheinlich, dass du den Modalwert erhalten wirst, also das, was die meisten kriegen …

Gut zu wissen: Der Unterschied zwischen Mittelwert, Median und Modalwert

Bei der praktischen Berechnung gibt es auch einen Unterschied zwischen Mittelwert und Median: Während für den arithmetischen Mittelwert die (ungeordnete) Urliste herangezogen werden kann, müssen zur Berechnung des Medians die Daten zuerst in eine (der Größe nach geordnete) Rangliste gebracht werden.

Ein weiterer Unterschied zwischen Median und Mittelwert ist der, dass für das arithmetische Mittel numerische Daten notwendig sind. Für den Median reicht es hingegen, wenn wir die Daten der Größe nach ordnen können. Und das geht bereits mit *ordinalskalierten* Daten, also *Rangmerkmalen*. **Hinweis:** Diese Behauptung gilt nur mit Einschränkungen: Formel 3.15 können wir nicht anwenden, wenn wir nichtnumerische ordinalskalierte Daten haben. Dann müssen wir immer, unabhängig davon ob wir eine gerade oder ungerade Anzahl von Elementen haben, Formel 3.16 verwenden.

Der Modalwert (häufigste Wert) kann als einziger auch für nominal skalierte Merkmale angegeben werden. Er gibt ein «typisches» Ergebnis an. Ein Unterschied des Modalwertes ist auch, dass es sich dabei immer um einen tatsächlich beobachteten Wert handelt. Median und Mittelwert hingegen sind errechnete

Größen, die als Beobachtung in der Stichprobe gar nicht vorkommen müssen.

3.2 Streuungskennwerte empirischer Häufigkeitsverteilungen

Lage-Kennzeichen geben noch kein vollständiges Bild der Daten und ihrer Verteilung wieder. So können zum Beispiel verschiedene Daten alle denselben Mittelwert haben, die Histogramme und Häufigkeitssummenkurven hingegen sehen alle anders aus. Offensichtlich gibt es noch ein anderes wichtiges Unterscheidungsmerkmal.

Es sind dies die so genannten **Streuungskennwerte**. Sie charakterisieren die Schwankungen der Daten und geben Auskunft darüber, wie stark die einzelnen Werte voneinander abweichen beziehungsweise wie weit sie vom Durchschnitt abweichen. Je weniger die einzelnen Merkmalswerte mit dem Durchschnitt übereinstimmen, umso wichtiger ist die Angabe von Streuungskennwerten. Einfach anzugebende Streuungsmaße sind

Spannweite und Interquartilsabstand

Die **Spannweite** (auch: *Variationsbreite*) ist die Distanz (mathematisch: die Differenz) zwischen dem größten und dem kleinsten Beobachtungswert:

$$\Delta = x_{\max} - x_{\min} \tag{3.17}$$

Der **Interquartilsabstand** (eng.: *interquartile range*) ist die Distanz zwischen dem 1. und 3. Quartil:

$$IQR = x_{0.75} - x_{0.25} \tag{3.18}$$

und beinhaltet die mittleren 50% der Daten. Zur grafischen Darstellung des Quartilsabstands dienen so genannte **Boxplots**. Dabei wird zwischen dem 1. und 3. Quartil ein Rechteck – die «Box» – gezeichnet, mit einer «Mittellinie» in der Höhe des Medians. An der Box hängt noch oben und unten je eine durch einen waagrechten Strich abgeschlossene Linie; sie werden auch «Whisker» genannt und reichen bis zum Minimum bzw. Maximum der Daten. Abb.3.1 zeigt ein Boxplot-Diagramm der Daten aus Beispiel 3.

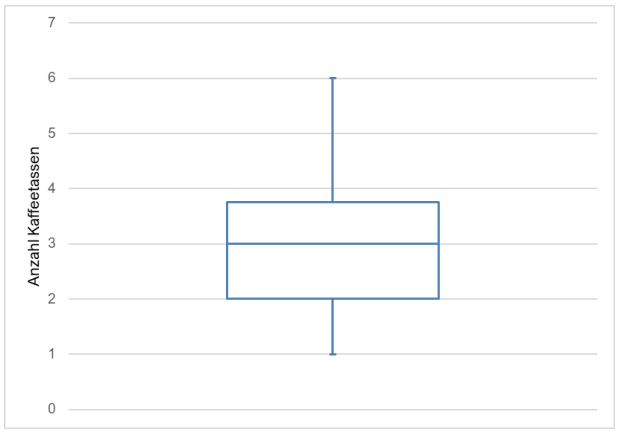


Abb. 3.1: Das Boxplot-Diagramm zu den Daten aus Beispiel 3: 50% der Daten liegen innerhalb der «Box», die beiden «Antennen» reichen vom Minimum bis zum Maximum und umfassen somit 100% der Daten

In *R* erhalten wir mit dem Befehl `range(x)` nicht die Spannweite, sondern das Minimum und das Maximum der Daten.
Der Interquartilsabstand kann mit dem Befehl `IQR(x)` berechnet werden. `IQR(frida)` ergibt zum Beispiel `2.5`

Der Quartilsabstand kann auch dazu verwendet werden, um in einer ersten Näherung *Ausreißer-Grenzen* festzulegen:

$$A_u = x_{0.25} - 1.5 \cdot IQR \tag{3.19}$$
$$A_o = x_{0.75} + 1.5 \cdot IQR \tag{3.20}$$

Datenwerte, die außerhalb des Intervalls $[A_u, A_o]$ liegen, können als extreme Werte (Ausreißer) angesehen und eventuell gestrichen werden. *Achtung:* Dies ist nur ein näherungsweises Vorgehen. Für unsere Zwecke aber meist ausreichend.

Aufgabe 5 Bestimme zu den Daten der Tabelle 3.4 die Spannweite und den Quartilsabstand. Gibt es auf Grund dieser Streuungswerte Anzeichen, dass die Daten Ausreißer enthalten?

Variationsbreite und Quartilsabstand sagen zwar schon einiges über die Verteilung der Daten aus, berücksichtigen aber nur einige wenige Werte (eben Maximum und Minimum und die Quartile). Noch informativer wäre ein Kennwert, der alle Messwerte berücksichtigt. Das machen die Folgenden:

Empirische Varianz und Standardabweichung

Die **empirische Varianz** (auch: *Stichprobenstreuung*) charakterisiert die Abweichungen der Daten von ihrem Mittelwert. Es ist die Summe der quadrierten Abweichungen der Beobachtungswerte von ihrem arithmetischen Mittelwert dividiert durch die Anzahl aller Werte minus Eins. Sie wird auch *mittlere quadratische Abweichung* genannt:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.21}$$

Die Verwendung der Quadrate in Formel (3.21) hat zwei Vorteile: Durch das Quadrieren werden alle Abweichungen positiv und können sich nicht gegenseitig aufheben; außerdem werden größere Abweichungen stärker berücksichtigt als kleinere.

Es gibt aber auch einen Nachteil: Die Varianz ist nicht sehr anschaulich und damit praktisch nicht verwendbar. Wenn wir zum Beispiel irgendeine Zufallsvariable untersuchen, die wir in Euro messen, hat die Varianz die Einheit Quadrateuro (die Einheiten werden ja mitquadriert); es ist nicht so leicht, sich darunter auch etwas vorzustellen. Besser wäre ein Maß, das in derselben Einheit wie die Messwerte angegeben werden kann:

Die **Standardabweichung** ist die positive Quadratwurzel[11] aus der Varianz:

$$s = +\sqrt{s^2} \tag{3.22}$$

Jetzt kann man sich auch leichter vor Augen halten, was dieses Maß repräsentiert: Die Standardabweichung gibt an, um wieviel ein einzelner Messwert durchschnittlich (also sozusagen «standardmäßig») vom Mittelwert abweicht. Eine geringe Standardabweichung bedeutet, die Daten liegen eher enger um den Mittelwert; eine hohe Standardabweichung weist auf eine stärkere Streuung um den Mittelwert hin.

Beispiel 16 *Berechne die Standardabweichung der Körpergröße der 7 Zwerge (Aufgabe 3):*

Dazu müssen wir zunächst den Mittelwert ausrechnen:

$$\bar{x} = \frac{85 + 80 + 62 + 81 + 70 + 80 + 88}{7} = \frac{546}{7} = 78\ cm$$

[11] normalerweise kann die Wurzel einer Zahl positiv oder negativ sein. 4 zum Beispiel hat die beiden Wurzeln $+2$ und -2. Laut Definition verwenden wir für die Standardabweichung aber nur die positive Wurzel.

und daraus die Varianz:

$$s^2 = \frac{(85-78)^2+(80-78)^2+(62-78)^2+(81-78)^2+(70-78)^2+(80-78)^2+(88-78)^2}{(7-1)} =$$
$$= \frac{7^2+2^2+(-16)^2+3^2+(-8)^2+2^2+10^2}{6} = \frac{49+4+256+9+64+4+100}{6} = \frac{486}{6} = 81$$

und schließlich die Standardabweichung:

$$s = \sqrt{81} = 9\ cm$$

Aufgabe 6 Wie groß ist die Varianz der in Beispiel 3 gegebenen Daten? (Didaktischer Hinweis: Bevor du den Rechner anwirfst: Versuch einmal, diese Aufgabe mit «Papier und Bleistift» zu lösen.).

In *Excel* wird die Standardabweichung einer Stichprobe mit der Funktion `=STABW.S(Zahl1;Zahl2;...)` berechnet.
In *LibreOffice Calc* heißt der Befehl: `=STABW(Zahl1;Zahl2;...)`
In *R* erhalten wir die Standardabweicheung mit `sd(x)` bzw. die Varianz mit `var(x)`.

Aufgabe 7 Der folgende Datensatz besteht aus elf Elementen und hat einen arithmetischen Mittelwert von 25. Außerdem wissen wir, dass die Daten völlig symmetrisch um den Mittelwert gestreut sind.

Welche Werte musst du für x_1 und x_{11} einsetzen, damit die Varianz 26 beträgt?

$$x_1, 21, 22, 23, 24, 25, 26, 27, 28, 29, x_{11}$$

Aufgabe 8 Kannst du – ohne konkret jede Kennzahl auszurechnen – «auf einen Blick» sagen und begründen, welche der drei folgenden Datensätze die größte Standardabweichung hat und welche die kleinste?

A) 0, 20, 40, 50, 60, 80, 100
B) 0, 48, 49, 50, 51, 52, 100
C) 0, 1, 2, 50, 98, 99, 100

Variationskoeffizient

Der **Variationskoeffizient** ist der Quotient der Standardabweichung dividiert durch den Betrag des arithmetischen Mittelwerts. Er wird meistens in Prozent angegeben:

$$v_x = \frac{s}{|\bar{x}|} \cdot 100\% \tag{3.23}$$

Der Variationskoeffizient ist demnach eine Art *relative Standardabweichung*. Er wird verwendet, wenn man Standardabweichungen miteinander vergleichen will.

3.3 Zentrierter, normierter und standardisierter Beobachtungswert

Manchmal müssen Werte aus unterschiedlichen Verteilungen miteinander verglichen werden. Zum Beispiel wird der IQ (Intelligenzquotient) oft auf einer Skala gemessen, die den Mittelwert 100 IQ-Punkte und die Standardabweichung 15 IQ-Punkte hat. Mitunter gibt es aber auch andere Tests, die auf einer Standardabweichung von 16 oder 24 IQ-Punkten basieren. Der unmittelbare Vergleich der Absolutwerte der Testergebnisse ist in diesem Fall nicht sehr aussagekräftig[12] und entspricht in etwa dem sprichwörtlichen Vergleich von Äpfeln und Birnen.

Besser ist es dann, die Einzelwerte zuerst zu «relativieren» und auf eine einheitliche Basis zu bringen. Diesen Vorgang nennt man *Standardisieren*. Er läuft in zwei Schritten ab: einer Zentrierung und einer Normierung.

Der **zentrierte Beobachtungswert** ist der Beobachtungswert minus des arithmetischen Mittelwerts:

$$x_i - \bar{x} \tag{3.24}$$

Zentriert man einen gesamten Datensatz, dann ist das arithmetische Mittel der zentrierten Daten gleich Null.

Der **normierte Beobachtungswert** ist der Beobachtungswert dividiert durch die Standardabweichung:

$$\frac{x_i}{s} \tag{3.25}$$

[12] Das gilt nicht nur für IQ-Tests sondern für alle Vergleich von Daten aus unterschiedlichen Verteilungen

Der **standardisierte Beobachtungswert** ist der zentrierte Beobachtungswert dividiert durch die Standardabweichung, es wird also zuerst zentriert und anschließend normiert:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.26}$$

Dieser Wert (auch als **z-Wert** bezeichnet) gibt an, «wie viele Standardabweichungen» der Messwert x_i vom Mittelwert $\bar{x}$ entfernt ist. Der *z-Wert* ist dimensionslos. Er kann positiv, negativ oder Null sein. Das Vorzeichen gibt Auskunft darüber, ob der zugehörige Messwert über- oder unterdurchschnittlich ist. Ein *z-Wert* von 2 gibt zum Beispiel an, dass der zugehörige Messwert 2 Standardabweichungen oberhalb des Mittelwertes liegt; ein *z-Wert* von -1.7 bedeutet, dass der zugehörige Messwert 1.7 Standardabweichungen unterhalb des Mittelwertes liegt; ein *z-Wert* von 0 bedeutet, dass der zugehörige Messwert genauso groß ist, wie der Mittelwert.

Bildet man von allen Messwerten die *z-Werte*, und wertet diese statistisch aus, so zeigt sich, dass der Mittelwert der *z-Werte* gleich 0 ist, und ihre Standardabweichung 1.

Aufgabe 9 Gib zu den Daten aus Aufgabe 7 die standardisierten $z-$Werte an.

Für einen besseren Überblick

hier noch einmal eine Übersicht über die Kennwerte für empirische Häufigkeitsverteilungen und für welche Skalen wir sie verwenden können:

Datenart	**Kennwerte**
Nominaldaten	Modalwert
Ordinaldaten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand
Intervallskalierte Daten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand, arithmetisches Mittel, Standardabweichung, standardisierter Beobachtungswert
Rationalskalierte Daten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand, arithmetisches und geometrisches Mittel, Standardabweichung, Variationskoeffizient, standardisierter Beobachtungswert

Tabelle 3.6: Übersicht: je nach Datenart mögliche Kennwerte

4 Merkmalszusammenhänge

In diesem Kapitel geht es um die Beziehung zwischen zwei Zufallsvariablen. Wir sprechen dabei auch von **bivariaten** Daten[1], d.h. dass wir gleichzeitig *zwei* Merkmale untersuchen. Wir wollen dabei herausfinden, ob oder wie stark die beiden Zufallsvariablen einander beeinflussen. Wenn uns das gelingt, können wir für einige Phänomene der Wirklichkeit – zumindest statistisch – erklären, warum sie sich im Ergebnis unterscheiden, in manchen Fällen sogar in gewisser Weise Vorhersagen treffen. Nicht 100%ig perfekte Vorhersagen, aber immerhin. Umgekehrt können wir auch mitunter zeigen, dass an manchen scheinbaren Zusammenhängen nichts dran ist und nur einer getäuschten Intention (oder unserem Wunschdenken) entspricht. Gesucht sind letztlich Art und Stärke des Zusammenhangs.

Manchmal unterscheiden wir dabei in eine **Zielvariable** (auch: *interessierende* Variable, eng.: *response variable*) und eine **Einflussvariable** (auch: *erklärende* Variable, eng.: *explanatory variable*). Die Zielvariable lässt sich dabei aus der Einflussvariablen ableiten. Manchmal ist es aber auch so, dass es zwar einen Zusammenhang gibt, aber beide Variablen gleichberechtigt sind. Wir können also nicht immer genau sagen, welches die Ziel- und welches die Einflussvariable ist, oder ob sie nicht zum Beispiel beide von einer dritten beeinflusst werden.

Mathematisch geben wir die Beziehung zwischen zwei Zufallsvariablen an, indem wir eine Variable mehr oder weniger als Funktion der anderen darzustellen versuchen. *«Mehr oder weniger»* bedeutet dabei, dass es nicht um eine strenge Funktion im mathematischen Sinn geht (siehe Abb.4.1).

[1] vom. lat. *bis* = zweimal und *variare* = (sich) verändern. Bei der gleichzeitigen Betrachtung von mehr als zwei Zufallsvariablen sprechen wir von *multivariaten* Verfahren, betrachten wir jeweils nur einzelne Variable (wie in den Kapiteln bisher), von *univariaten*.

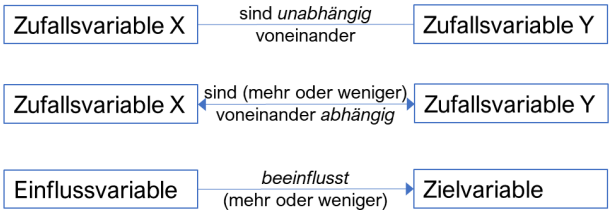


Abb. 4.1: Zufallsvariable können unterschiedlich zusammenhängen

4.1 Streu- und Bubblediagramme

Ein Beispiel: Sehen wir uns zunächst ein einfaches Beispiel an: Tabelle 4.1 zeigt das Ergebnis der Messung von Größe und Gewicht zwanzig zufällig ausgewählter Personen:

X Größe [cm]	Y Gewicht [kg]	X Größe [cm]	Y Gewicht [kg]
188	83	170	68
183	88	187	92
183	81	177	85
185	85	178	78
178	70	180	75
198	94	182	75
163	55	189	88
164	57	173	68
174	80	176	77
185	78	177	78

Tabelle 4.1: Größe und Gewicht 20 zufällig ausgewählter Personen

Wir können nun die beiden Zufallsgrößen Größe und Gewicht *gemeinsam* betrachten und in einem **Streudiagramm** (auch: *Punktdiagramm* oder *«Punktwolke»)* darstellen (Abb.4.2). Dazu stellen wir die beiden Variablen X und Y in einem Koordinatensystem dar und zeichnen für jeden Merkmalsträger einen Punkt an den Koordinaten (X,Y) ein.

Jeder Punkt im Streudiagramm repräsentiert somit Informationen über die Kombination aus zwei Merkmalen. Aus dem Diagramm können wir in weiterer Folge gut eventuelle «Muster» in unseren Daten visuell ablesen und Trends und augenscheinliche Zusammenhänge (und auch: Nicht-Zusammenhänge) erkennen. Möglich ist die Visualisierung in einem Punktdiagramm in der «klassischen» Form aber nur für metrische, unklassierte Daten.

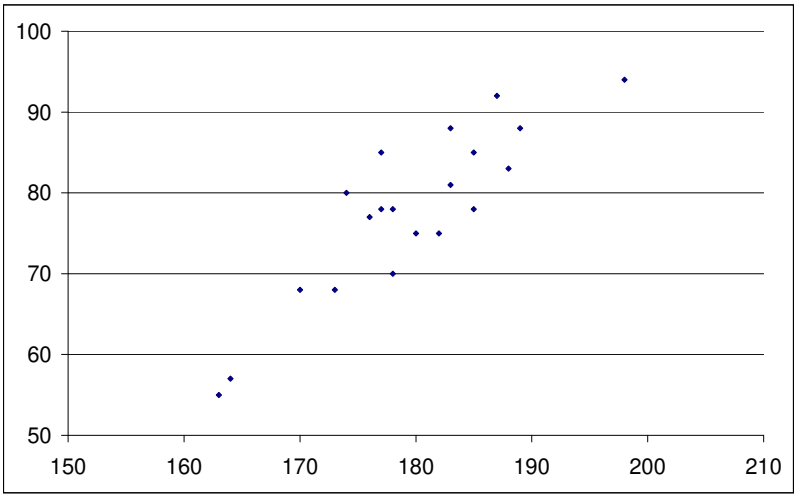


Abb. 4.2: Streudiagramm zu den Daten aus Tab.4.1

Bubble-Diagramme (*Blasendiagramme*) werden ähnlich wie Streudiagramme dafür verwendet, Zusammenhänge zwischen Zufallsgrößen zu visualisieren. Dabei werden zunächst zwei Merkmale in einem Streudiagramm eingezeichnet. Anstelle von einfachen, gleich großen Punkten verwendet man aber Punkte mit unterschiedlichen Durchmessern – dadurch werden aus den Punkten «Blasen» (eng. *bubble*).

Damit ist es möglich, noch eine dritte Variable und somit zusätzliche Information im Diagramm darzustellen. Verwendet man darüber hinaus auch noch unterschiedliche Farben für die Blasen, kann man auch noch ein viertes Merkmal in die grafische Darstellung hineinpacken.

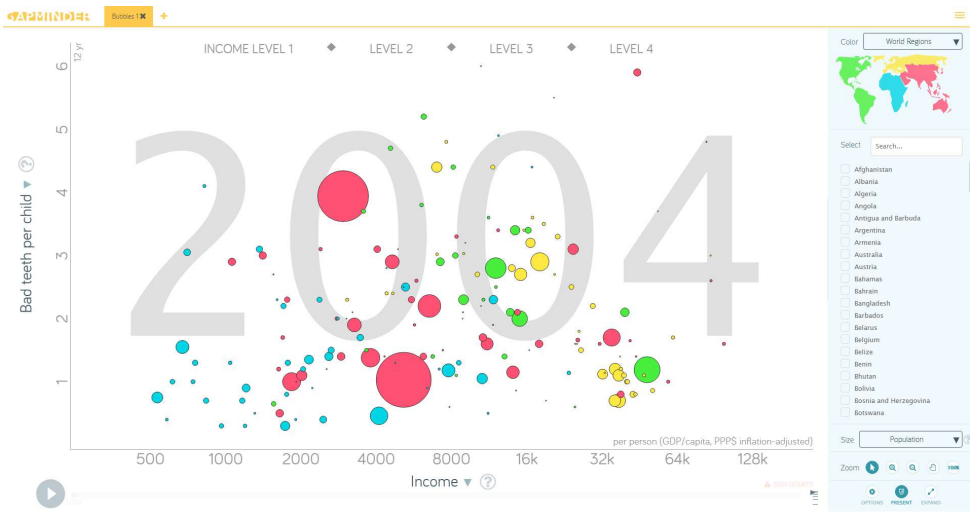


Abb. 4.3: Ein Bubblediagramm (Quelle: www.gapminder.org)

Abbildung 4.3 zeigt ein Beispiel dazu. Auf der x-Achse ist länderweise das «Bruttoinlandsprodukt pro Einwohner» eingezeichnet, auf der y-Achse der Mittelwert der «Anzahl an schlechten Zähnen», die ein zwölfjähriges Kind in diesen Ländern hat. Zusätzlich ist über die Größe der Blasen die Größe des jeweiligen Länder repräsentiert und über die Farbgebung die Zugehörigkeit zu einer bestimmten Region.

4.2 Regressionsrechnung

In der Physik beschreiben wir Zusammenhänge durch Formeln. Zum Beispiel ist beim Autofahren der Zusammenhang zwischen dem Anhalteweg s und der gefahrenen Geschwindigkeit v, der Reaktionszeit t und der Bremsverzögerung a gegeben durch:

$$s = t \cdot v + \frac{v^2}{2a}$$

So exakt die Formel auch aussehen mag: Nur wenn wir exakte Werte für t, a und v kennen, erhalten wir einen exakten Wert für s. Meist «schätzen» wir das Ergebnis, indem wir für $t = 0.8\,s$ Reaktionszeit und $a = 8.0\,m/s^2$ Bremsverzögerung einsetzen. Damit ist zum Beispiel bei einer Geschwindigkeit von $v = 50\,km/h$ der Anhalteweg $s = 23\,m$ lang, bei $v = 130\,km/h$ ist er $110\,m$ lang, etc.

Andere Zusammenhänge lassen sich zwar auch eindeutig abbilden, allerdings nicht immer durch eine in eine mathematische Formel gegossene Funktion. Zum Beispiel wird jedem Platz in einem Ski-Weltcuprennen ein eindeutiger Punktegewinn zugeschrieben. Abb. 4.4 zeigt diesen Zusammenhang.

Und dann gibt es Beispiele von Merkmalszusammenhängen, die sich eben nicht auf mathematische Funktionen oder eindeutige Zuordnungen zurückführen lassen, sondern eher *statistischer* Natur sind. Dazu betrachten wir noch einmal die beiden Abbildungen 4.2 und 4.3. Aus dem Bubblediagramm (Abb.4.3) lässt sich kein Zusammenhang zwischen den beiden Zufallsgrößen ableiten[2]. In Abb.4.2 hingegen können wir augenscheinlich feststellen, dass mit zunehmen- dem X auch die Variable Y tendenziell zunimmt. Das legt den Schluss nahe, dass sich das Körpergewicht aus der Körpergröße erklären lässt[3]. Dieser Zusammenhang ist natürlich kein streng *deterministischer*, d.h. es gibt kein naturwissenschaftliches Gesetz oder Funktion, nach dem man aus der Körpergröße das exakte

[2] Ob ein solcher überhaupt zu erwarten gewesen wäre, wollen wir hier nicht weiter erörtern...

[3] Zumindest teilweise. Wir wissen, dass die Größe nur eine Variable ist, die das Gewicht beeinflusst und noch andere Parameter eine Rolle spielen. Aber in dieser einfachen statistischen Untersuchung betrachten wir nur den Zusammenhang bivariater Daten.

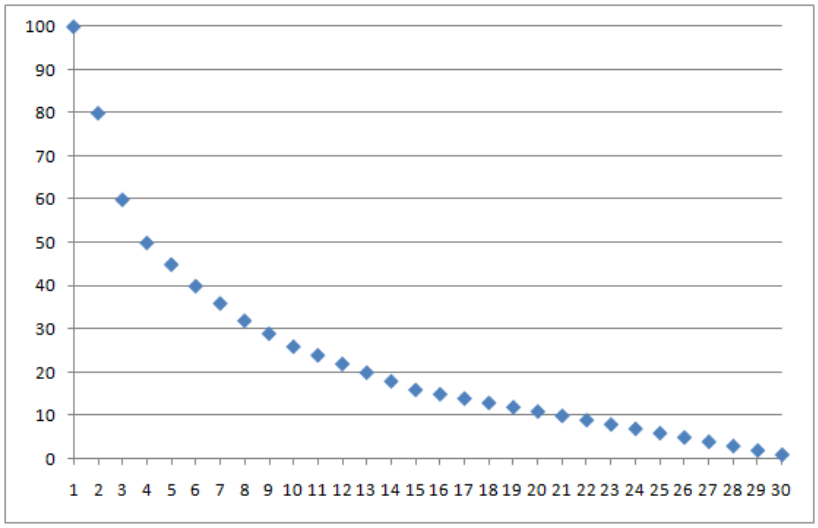


Abb. 4.4: Punktezuteilung zur erreichten Platzierung in einem Ski-Worldcuprennen

Gewicht errechnen kann. Es gibt aber einen *tendenziellen* Zusammenhang; wir nennen das auch einen *statistischen* bzw. einen *stochastischen* Zusammenhang. Den Beinamen «stochastisch» erhält er, weil wir ihn immer nur mit einer gewissen *Unschärfe* angeben können[4]. Aufgabe der **Regressionsrechnung** ist es, die Art des stochastischen Zusammenhangs zu beschreiben.

Zunächst einmal können wir in Abb.4.2 ein bestimmtes Muster erkennen, das von links unten nach rechts oben verläuft. Niedrigen Werten auf der x-Achse entsprechen niedrige Werte auf der y-Achse; steigt der x-Wert, dann steigt auch der y-Wert. Wir sprechen in diesem Fall von einem *positiven* Zusammenhang. Andernfalls – wenn das Muster also von links oben nach rechts unten läuft und niedrige Werte auf der x-Achse mit hohen Werten auf der y-Achse korrespondieren (und umgekehrt) – von einem *negativen*. Es kann natürlich auch sein, dass wir wirklich im wahrsten Sinn des Wortes einen Punkt-*Haufen* vor uns haben und zunächst einmal überhaupt kein nennenswerter Zusammenhang oder Muster erkennbar ist. Diese drei grundsätzlichen Möglichkeiten (positiver, negativer und kein Zusammenhang) sind in Abb.4.5 dargestellt.

Die nächste Frage, die wir uns stellen, ist: Von welchem Typ könnte eine Funktion sein, die wir in die Punktwolke hineinlegen können, und die als charakteristischer Repräsentant der Punktwolke gelten kann?

Prinzipiell unterscheiden wir dabei zwischen *linearen* und *nicht-linearen* Funktionen. Lineare Funktionen (z.B. Gerade) sind einfacher zu handhaben; nichtlineare Regressionszusammenhänge benötigen kompliziertere Funktionen. Wir

[4]Zum Wort *stochastisch* siehe Fußnote 6 auf Seite 55.

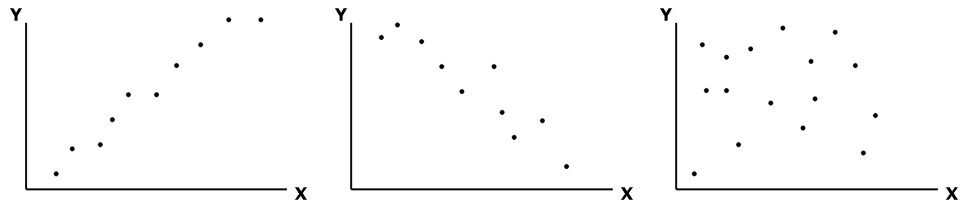


Abb. 4.5: Streudiagramme mit verschiedenen Mustern (positiv, negativ und «zusammenhangslos»)

werden uns in diesem Kurs auf lineare Zusammenhänge beschränken, also solche, die sich mit Geraden darstellen lassen – die so genannte *Regressionsgerade*. Das ist jene Gerade, die einen Punkthaufen wie jenen in Abb.4.2 «am besten» repräsentiert. Wie können wir die Parameter dieser Regressionsgeraden bestimmen?

Die Regressionsgerade

Eine Gerade (und ihre Gleichung) ist – wie wir aus der Mathematik wissen – durch zwei Parameter eindeutig bestimmt: den *Anstieg* der Geraden und den *Achsenabschnitt* auf der y-Achse (= die «Verschiebung» entlang der y-Achse relativ zum Ursprung des Koordinatensystems). Die Geradengleichung heißt dann:

$$y = kx + d \tag{4.1}$$

Für die Regressionsgerade gehen wir so vor: Zunächst berechnet man für jede Zufallsvariable den jeweiligen Mittelwert sowie die Varianz der Zufallsgröße X:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.2}$$

und anschließend eine weitere Größe, die wir mit s_{xy} bezeichnen:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \tag{4.3}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}\right) \tag{4.4}$$

Dann erhält man die Parameter der Regressionsgeraden aus

$$k = \frac{s_{xy}}{s_x^2} \qquad d = \bar{y} - k\bar{x} \tag{4.5}$$

Der Achsenabschnitt d der Regressionsgeraden wird auch als **Niveaufaktor** bezeichnet.

Der Anstieg k der Regressionsgeraden wird auch als **Regressionskoeffizient** bezeichnet. Er kann positiv oder negativ sein und dementsprechend sprechen wir von *positiver* bzw. *negativer linearer Regression*

In *MS Excel* und *LibreOffice Calc* können wir den Anstieg k der Regressionsgeraden mit dem Befehl `=STEIGUNG(Y-Werte; X-Werte)` berechnen, den Achsenabschnitt d mit `=ACHSENABSCHNITT(Y-Werte;X-Werte)`.
In *R* lautet der Befehl `lm(Y~X)`.

Beispiel 17 *Für unser Eingangsbeispiel erhalten wir:*

$$k = 1.08 \qquad d = -116.10$$

was wir auch gleich grafisch umsetzen können und in das Streudiagramm 4.2 die Regressionsgerade einzeichnen (Abb.4.6).

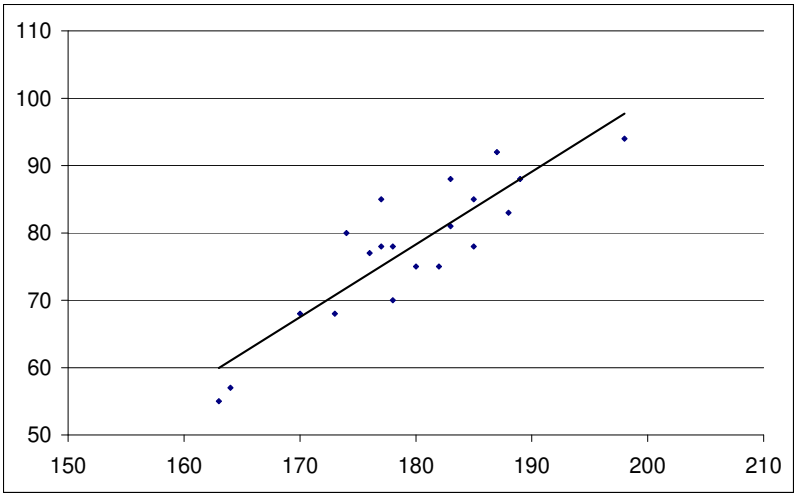


Abb. 4.6: Regressionsgerade zu den Daten aus Tab.4.1

Mit Hilfe der Regressionsgeraden sind durch einfaches Einsetzen nun auch Prognosen für nicht empirisch bestimmte Merkmalsausprägungen möglich. Wir können zum Beispiel angeben, welches Körpergewicht für einen Erwachsenen mit einer Körpergröße von 196 *cm* statistisch zu erwarten ist, nämlich:

$$y = kx + d = 1.08 \cdot 196 - 116.10 = 96\,kg$$

Achtung: Die «Vorhersage», die wir eben über eine 196 *cm* große Person getroffen haben, ist mathematisch gesehen eine **Interpolation**, d.h. wir haben für x einen Wert angegeben, der innerhalb des Wertebereiches liegt, mit dem wir die Parameter k und d berechnet haben. (Der kleinste Wert war 163, der größte 198). Dem gegenüber liegt eine **Extrapolation** vor, wenn wir für x einen Wert einsetzen, der außerhalb dieses Wertebereichs liegt.
Extrapolationen sind immer mit Vorsicht zu genießen. Setzt man in unserem Beispiel für $x = 100$ ein, käme für y ein negativer Wert heraus ($y = 1.08 \cdot 100 - 116.10 = -8.1$). Offensichtlich kann aber selbst ein Kind mit einer Körpergröße von 1 *m* kein negatives Körpergewicht haben. . . .

Aufgabe 10 In Tabelle 4.2 sind für sieben in der Vergangenheit in Wien abgehaltene Wahlen die Mittagstemperatur am jeweiligen Wahltag (x) sowie das Verhältnis der abgegebenen Stimmen zur Anzahl der Wahlberechtigten, also die Wahlbeteiligung (y) gegeben. Gib den Regressionskoeffizienten an.

	28.09.08	07.06.09	10.10.10	29.09.13	25.05.14	11.10.15	24.04.16
x (Temperatur °C)	15	22	12	12	24	7	7
y (Wahlbeteiligung)	0.74	0.43	0.68	0.70	0.35	0.75	0.64

Tabelle 4.2: «Wahltemperatur» und Wahlbeteiligung in Wien

Aufgabe 11 (Fortsetzung zu Aufgabe 10): In obiger Tabelle ist die Bundespräsidentenwahl 2010 nicht enthalten. Die Mittagstemperatur am Wahltag (25.4.2010) betrug 18°. Welche Wahlbeteiligung war bei dieser Temperatur zu erwarten?

Der Weg zur Mittelmäßigkeit

An dieser Stelle noch ein weiterer Hinweis: Das Wort *Regression*[5] ist an sich keine sehr aussagekräftige Bezeichnung für diese Methode; sie wurde von ihrem Erfinder, *Francis Galton*[6], auf Grund eines einzigen Beispiels geprägt: Galton, ein Cousin von Charles Darwin, versuchte, die Evolutionstheorie seines Cousins

[5] vom lat. *regredior* = zurückgehen

[6] Sir Francis Galton, 1822-1911, englischer Arzt und Biologe. Er verfasste zahlreiche Arbeiten über Anthropologie und Vererbung und sammelte dazu Daten über verschiedene Merkmalsausprägungen der Menschen. Anschließend entwickelte er statistische Methoden zu ihrer Auswertung.

durch quantitative Beispiele zu untermauern. In einer großangelegten experimentellen Studie untersuchte er, ob es eine Beziehung zwischen der Körpergröße der Eltern und der ihrer Kinder gibt. Er fand heraus, dass zwar große Eltern tendenziell auch große Kinder haben und kleine Eltern kleine Kinder, allerdings in der Weise, dass die Kinder großer Eltern eher kleiner sind als ihre Eltern und umgekehrt. Eltern haben also meistens Kinder, deren Größe näher am Durchschnitt liegt als ihre eigene Größe. Er nannte diesen Zusammenhang «*regression* to mediocrity» – den «Rückschritt zum Mittelmaß»[7].

4.3 Korrelationsrechnung

Die Regressionsgerade beschreibt zwar die *Art* des statistischen Zusammenhangs, sagt aber nichts über seine *Stärke* aus. Wir werden umso ungenauere Prognosen abgeben, je geringer der statistische Zusammenhang der beiden Variablen ist. Eine Regressionsgerade lässt sich nach obigen Formeln ja in jedem Fall berechnen, auch wenn so gut wie kein Zusammenhang vorliegt. Die Frage ist aber, wie eng oder weit die Punktwolke um die erhaltene Regressionsgerade streut. Dies beantwortet die **Korrelationsrechnung**.

Dazu rufen wir uns zunächst die Größe in Erinnerung, die wir in Formel (4.4) auf Seite 80 verwendet haben. Es ist dies die

Kovarianz

Zwischen je zwei statistischen Variablen X und Y können wir einen Parameter für die «gemeinsame Streuung» angeben, genannt die «*Kovarianz* von X und Y». Sie lässt sich mit Hilfe der beiden Mittelwerte $\bar{x}$ und $\bar{y}$ ausrechnen:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{4.6}$$

Die Kovarianz ist also das *mittlere Abweichungsprodukt* und ist ein Maß für den wechselseitigen Zusammenhang der beiden Zufallsgrößen X und Y.

Ist die Kovarianz positiv, so sind die Zufallsgrößen X und Y tendenziell eher gleich, d.h. mit großer Wahrscheinlichkeit nimmt die eine zu, wenn auch die andere zunimmt, beziehungsweise ab, wenn die andere abnimmt.

[7] Galton, Francis. *Regression Towards Mediocrity in Hereditary Stature.* The Journal of the Anthropological Institute of Great Britain and Ireland 15 (1886): 246–63. archive.org/details/journalroyalant15irelgoog/page/244/mode/1up

Ist die Kovarianz hingegen negativ, verhalten sich die Zufallsgrößen tendenziell eher reziprok, d.h. mit großer Wahrscheinlichkeit nimmt die eine ab, wenn die andere zunimmt, beziehungsweise zu, wenn die andere abnimmt.

Zufallsgrößen, deren Kovarianz gleich Null ist, bezeichnen wir als *statistisch unabhängig* voneinander.

Der Korrelationskoeffizient

Der Wert der Kovarianz ist abhängig von der Dimension der beiden Zufallsgrößen X und Y. Beschreibt zum Beispiel X die Körpergröße und Y das Gewicht, so ist der Wert von s_{xy} unterschiedlich, je nachdem ob die Größe in *cm* oder *m* angegeben wird bzw. das Gewicht in *dag* oder *kg*. Das ist nicht besonders praktisch. Die Kovarianzen können aber *normiert* werden, indem sie durch die jeweiligen Standardabweichungen dividiert werden. Damit schafft man ein dimensionsloses Maß. Der entsprechende Quotient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{4.7}$$

ist der **Korrelationskoeffizient**, genauer auch: der *Pearson-Korrelationskoeffizient* oder manchmal auch die *Produkt-Moment-Korrelation nach Bravais und Pearson*[8] genannt. Er ist ein Maß für den *linearen* statistischen Zusammenhang zwischen zwei Zufallsvariablen.

Zur Erinnerung: s_{xy} ist die Kovarianz – vgl. Formel (4.6), s_x die Standardabweichung der Variablen X und s_y die Standardabweichung der Variablen Y – beide werden mit der Formel (3.21) bzw. (3.22) ausgerechnet. Es gilt:

$$-1 \leq r \leq 1 \tag{4.8}$$

d.h. dass der Korrelationskoeffizient nie kleiner als minus Eins und nie größer als plus Eins werden kann, ganz egal, wie groß X oder Y sind.

Eine positive Korrelation bedeutet, dass eine Vergrößerung der Werte der einen Zufallsgröße auch eine Vergrößerung der Werte der anderen Zufallsgröße zur Folge hat, bzw. eine Verkleinerung der einen Zufallsgröße eine Verkleinerung der anderen Zufallsgröße. Eine negative Korrelation hingegen bedeutet, dass eine Vergrößerung der Werte der einen Zufallsgröße eine Verkleinerung der Werte der anderen Zufallsgröße bewirkt und vice versa.

[8] *Karl Pearson*, 1857 - 1936, britischer Mathematiker und Statistiker; wir haben ihn bereits beim Histogramm auf Seite 40 kennengelernt. *Auguste Bravais*, 1811 - 1863, französischer Astronom und Physiker, hatte schon eine Zeit vor Pearson grundlegende theoretische Überlegungen zur Korrelationsrechnung veröffentlicht.

Ein Korrelationskoeffizient von *exakt* $+1.0$ oder -1.0 bedeutet, dass nicht nur ein statistischer linearer Zusammenhang besteht, sondern die Punkte tatsächlich auch mathematisch auf einer Geraden liegen und die Veränderungen streng äquivalent erfolgen. Alle anderen Werte von r lassen auf einen mehr oder weniger starken linearen Zusammenhang schließen. Ganz grob könnten wir sagen: Ein Korrelationkoeffizent bis etwa 0.3 bedeutet einen kleinen statistischen Zusammenhang. bei einem (Absolut-)Wert des Korrelationskoeffzienten von mindestens 0.3 und maximal 0.5, sprechen wir von einer mittleren Korrelation und ab 0,5 von einem großen Zusammenhang. Etwas feiner granuliert ergibt sich eine Zuordnung wie sie zum Beispiel in Tabelle4.3 vorgeschlagen wird.

Korrelationskoeffizient	Bedeutung
$r = -1$	vollständige lineare Abhängigkeit
$-1 < r \leq -0.8$	starker negativer linearer Zusammenhang
$-0.8 < r \leq -0.6$	mäßig starker negativer linearer Zusammenhang
$-0.6 < r \leq -0.4$	mittlerer negativer linearer Zusammenhang
$-0.4 < r \leq -0.2$	geringer negativer linearer Zusammenhang
$-0.2 < r < 0$	sehr schwache Korrelation
$r = 0$	keine lineare statistische Abhängigkeit
$0 < r < 0.2$	sehr schwache Korrelation
$0.2 \leq r < 0.4$	geringer positiver linearer Zusammenhang
$0.4 \leq r < 0.6$	mittlerer positiver linearer Zusammenhang
$0.6 \leq r < 0.8$	mäßig starker positiver linearer Zusammenhang
$0.8 \leq r < 1$	starker positiver linearer Zusammenhang
$r = 1$	vollständige lineare Abhängigkeit

Tabelle 4.3: Aus dem Korrelationskoeffizienten lässt sich die Stärke des linearen Zusammenhangs ablesen.

In *MS Excel* und *LibreOffice Calc* erhalten wir den Korrelationskoeffizienten mit `=KORREL(Y-Werte; X-Werte)`, wobei es – im Gegensatz zu Steigung und Achsenabschnitt der Regressionsgeraden – nicht darauf ankommt, welche Zufallsgröße als Y-Werte und welche als X-Werte bezeichnet werden. In *R* lautet der Befehl zur Berechnung des Korrelationskoeffizienten `cor(Y,X)`.

Beispiel 18 *Berechne zu den Daten aus Tab. 4.1 den Korrelationskoeffizienten und interpretiere den erhaltenen Wert.*

Zur Berechnung verwenden wir MS Excel und erhalten: $r_{xy} = 0.88$.

Das ist ein positiver Wert, was darauf hindeutet, dass eine Vergrößerung der Werte

der einen Zufallsgröße tendenziell auch eine Vergrößerung der Werte der anderen Zufallsgröße zur Folge hat (Was nicht weiter überraschend ist: Je größer jemand ist, desto schwerer ist er oder sie im Allgemeinen auch . . .).
Der konkrete Wert von 0.88 ist zudem relativ groß und lässt auf einen starken linearen statistischen Zusammenhang zwischen Größe und Gewicht schließen.

Übrigens geben wir üblicherweise von einem Korrelationskoeffizienten nicht mehr als zwei Nachkommastellen an – natürlich unter Beachtung üblicher Rundungsregeln.

Aus Formel (4.7) kann man erkennen, dass für den Korrelationskoeffizienten – im Gegensatz zur Regression – eine Unterscheidung in eine *Ziel-* und eine *Einflussvariable* nicht möglich ist. Es spielt keine Rolle, was wir als X und was als Y bezeichnen – die Formel ist bezüglich X und Y symmetrisch. Genauer müssen wir daher sagen: Die Regression beschreibt die Abhängigkeit einer Zufallsvariablen von einer anderen, der Korrelationskoeffizient die Stärke der *wechselseitigen* (linearen) Abhängigkeit.

An dieser Stelle noch ein Hinweis auf die Berechnung des Korrelationskoeffizienten, wenn wir X und Y in Form von standardisierten Messwerten vorliegen haben, also die «z-Werte» z_x und z_y (jeweils berechnet nach Formel 3.26 auf Seite 74): Der Korrelationskoeffizient kann dann auch berechnet werden aus

$$r = \frac{\sum z_x \cdot z_y}{n-1} \tag{4.9}$$

Aufgabe 12 In welchem mathematischen Zusammenhang stehen der Korrelationskoeffizient r_{xy} und der Regressionskoeffizient k? (Hinweis: Gib eine mathematische Gleichung an, die sowohl r_{xy} als auch k enthält)

Aufgabe 13 Besteht zwischen der in Tab.4.2 gegebenen Wahlbeteiligung der Wienerinnen und Wiener und der am Wahltag vorherrschenden Temperatur eine hohe Korrelation?
Wie groß ist der Korrelationskoeffizient?

Am Beginn dieses Abschnitts haben wir geschrieben: *«Eine Regressionsgerade lässt sich [. . .] in jedem Fall berechnen, auch wenn so gut wie kein Zusammenhang vorliegt»*. Wir können nun umgekehrt näher ausführen: Nur wenn es sich über den Korrelationskoeffizienten zeigen lässt, dass ein linearer statistischer Zusammenhang vorliegt, macht es Sinn, eine Regressionsgerade aufzustellen.

Und: Auf Seite 68 haben wir darauf hingewiesen, dass der Mittelwert ziemlich anfällig auf Ausreißer reagiert. An dieser Stelle müssen wir ergänzen: Das

gilt auch für den Korrelationskoeffizienten, insbesondere für kleine Stichproben: Ausreißer können nicht vorhandene lineare Zusammenhänge vorgaukeln – oder vorhandene verschleiern.

Der Determinationskoeffizient

Wenn wir den Korrelationskoeffizienten r quadrieren, erhalten wir den so genannten **Determinationskoeffizienten** (auch: *Bestimmtheitsmaß*):

$$r_{xy}^2 = \left(\frac{s_{xy}}{s_x s_y} \right)^2 \tag{4.10}$$

Er wird in der Regel in Prozent angegeben (d.h. mit 100 multipliziert und mit einem %-Zeichen versehen) und man kann ihn folgendermaßen deuten:

Der Determinationskoeffizient gibt an, zu wieviel Prozent sich eine Änderung der einen Zufallsvariable durch eine Änderung der anderen Zufallsvariable erklären lässt, also zu wieviel Prozent die eine die andere Zufallsvariable «determiniert» (Daher auch der Name).

Beispiel 19 *Berechne zu den Daten aus Tab. 4.1 den Determinationskoeffizienten und interpretiere den erhaltenen Wert.*

Nachdem wir im letzten Beispiel bereits den Korrelationskoeffizienten berechnet haben, brauchen wir ihn jetzt nur noch quadrieren (wobei wir den auf vier Stellen gerundeten Wert 0.8809 *verwenden):*

$$r^2 = 0.8809^2 = 0.7760 \approx 78\%$$

D.h. zu etwa 78% lässt sich das Gewicht durch den linearen Zusammenhang zwischen Körpergröße und Körpergewicht ableiten.

In *MS Excel* und *LibreOffice Calc* erhalten wir den Determinationskoeffizienten mit `=BESTIMMTHEITSMASS(Y-Werte; X-Werte)`
In *R* gibt es keinen Befehl für den Determinationskoeffizienten alleine. Mit dem Befehl `summary(lm(Y X))` werden aber alle möglichen Regressionsergebnisse angezeigt, darunter auch der Wert `Multiple R-squared`. Das ist das Bestimmtheitsmaß.

Formel 4.6 – uns somit Formel 4.7 – lassen sich nur anwenden, wenn sich auch die arithmetischen Mittelwerte $\bar{x}$ und $\bar{y}$ berechnen lassen. Das geht aber laut Tab.3.6 nur für metrische Merkmalswerte auf einer Intervall- oder Rationalskala. Wenn wir eine Korrelation zwischen zwei Zufallsgrößen angeben wollen, von denen eine oder beide «nur» ordinalskaliert sind, müssen wir die **Rangkorrelation** verwenden.

Für die Untersuchung des Zusammenhangs zweier Rangmerkmale müssen wir zunächst schauen, ob es auch «ex aequo-Plätze» gibt, ob es also Werte gibt, die mehrfach auftreten. Ist das nicht der Fall, ist die Berechnung sehr einfach: In Formel 4.6 werden einfach für x_i und y_i (und $\bar{x}$ und $\bar{y}$) die Rangplätze eingesetzt und dann nach 4.7 der Korrelationskoeffizient berechnet. Wir können diese einfache Vorgangsweise in der Praxis auch anwenden, wenn es nur einige wenige Mehrfachvorkommen gibt. (Sie hat nämlich den großen Vorteil, dass es eine Excel-Funktion gibt, in die sich direkt einsetzen lässt).

Gibt es hingegen mehr als «einige wenige» ex aequo-Ränge, dann müssen wir anstelle von Formel 4.7 den **Rangkorrelationskoeffizient nach Spearman**[9] verwenden:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n^3 - n} \tag{4.11}$$

wobei die D_i die Differenzen zwischen den beiden Rangzahlen des i-ten Elements sind.

Ist eine Variable ordinalskaliert, die andere aber rationalskaliert, muss vor der Berechnung des Korrelationskoeffizienten auch die rationalskalierte Variable auf eine Ordinalskala «herabskaliert» werden.

Beispiel 20 *Die folgende Tabelle zeigt das Körpergewicht und die Platzierung von 20 Teilnehmern an einem Laufbewerb. Gibt es zwischen diesen beiden Größen einen statistischen Zusammenhang? Gib den entsprechenden Korrelationskoeffizienten an:*

[9] Charles Edward Spearman, 1863 - 1945, britischer Psychologe

X Gewicht [kg]	Y Platzierung	X Gewicht [kg]	Y Platzierung
58	2	55	6
84	15	85	12
92	14	95	20
70	7	58	4
86	16	63	1
81	17	87	19
63	3	61	10
90	13	64	5
97	18	71	8
89	11	63	9

Wenn wir die «normale» Pearson-Korrelation nach 4.7 ausrechnen, erhalten wir einen Korrelationskoeffizient von 0.86. Allerdings haben wir dabei nicht beachtet, dass das Gewicht eine rationalskalierte Zufallsgröße ist und die Platzierung ordinalskaliert. Formel 4.7 darf daher nicht unmittelbar angewandt werden; die Zufallsgröße «Gewicht» muss zuvor ebenfalls ordinalskaliert werden. Daraus ergibt sich in diesem Beispiel:

X Gewichts-Rang	Y Lauf-Rang	X Gewichts-Rang	Y Lauf-Rang
2	2	1	6
12	15	13	12
18	14	19	20
9	7	2	4
14	16	5	1
11	17	15	19
5	3	4	10
17	13	8	5
20	18	10	8
16	11	5	9

und daraus für die Rangkorrelation, nach 4.7 ein Korrelationskoeffizient von 0.82.

Es gibt allerdings in den Originaldaten auch ex aequo-Plätze: Zwei Personen haben je 58 kg, drei Personen 63 kg. Der Pearson-Korrelationskoeffizient ist für die Rangkorrelation daher nur ein Näherungswert. Für die exakte Berechnung müssen wir den Spearman-Koeffizienten verwenden und dazu die Differenzen zwischen den jeweiligen Rangzahlen (und in weiterer Folge deren Quadratsumme) angeben:

X Gewichts-Rang	*Y Lauf-Rang*	D_i	D_i^2
2	*2*	*0*	*0*
12	*15*	*-3*	*9*
18	*14*	*4*	*16*
9	*7*	*2*	*4*
14	*16*	*-2*	*4*
11	*17*	*-6*	*36*
5	*3*	*2*	*4*
17	*13*	*4*	*16*
20	*18*	*2*	*4*
16	*11*	*5*	*25*
1	*6*	*-5*	*25*
13	*12*	*1*	*1*
19	*20*	*-1*	*1*
2	*4*	*-2*	*4*
5	*1*	*4*	*16*
15	*19*	*-4*	*16*
4	*10*	*-6*	*36*
8	*5*	*3*	*9*
10	*8*	*2*	*4*
5	*9*	*-4*	*16*
			$\Sigma = 246$

Daraus ergibt sich der Spearman-Rangkorrelationskoeffizient:

$$r_s = 1 - \frac{6 \cdot 246}{20^3 - 20} = 1 - \frac{1476}{7980} = \underline{\underline{0.82}}$$

Für das Beispiel 20 haben wir für die Berechnung des «Gewichts-Ranges» die *Excel*-Funktion `=RANG.GLEICH(Zahl;Bezug;1)` verwendet. In *LibreOffice Calc* lautet der entsprechende Funktionsaufruf ebenfalls `=RANG.GLEICH(Zahl;Bezug;1)`.

Aufgabe 14 Welche Werte kann ein Rangkorrelationskoeffizient annehmen?

4.4 Zusammenhänge kategorischer Merkmale

Bei kategorischen Merkmalen, zum Beispiel beim Vorliegen qualitativer Nominaldaten (z.B. Geschlecht, Beruf, Herkunftsland,...) aber auch quantitativen metrischen Daten, die in Klassen eingeteilt wurden (z.B. nach Altersgruppen), können die bisherigen Methoden dieses Kapitels nicht so einfach angewandt werden. Mit Kategorien oder Klassenintervallen können wir ja nicht wirklich gut einen Korrelationskoeffizienten berechnen. Für Nominaldaten können wir keine Rangkorrelation angeben, sie können auch in keinem Streudiagramm dargestellt werden ☹.

Ihre Häufigkeitsverteilung können wir aber in einer Art «gemeinsame Häufigkeitstabelle» darstellen, genannt **Kontingenztabelle** (auch: *Kreuztabelle*). Das sehen wir uns am besten an Hand eines Beispiels an:

Beispiel 21 *In einem Unternehmen gibt es drei «Verwendungsgruppen» für die Mitarbeiterinnen und Mitarbeiter:*

- *A : Arbeitnehmer:innen, die qualifizierte Tätigkeiten aufgrund ihrer Kenntnisse und Erfahrungen im Rahmen an sie erteilter Aufträge weitgehend selbstständig erledigen*
- *B : Arbeitnehmer:innen, die verantwortungsvolle Expert:innen-Tätigkeiten mit entsprechendem Entscheidungsspielraum selbstständig verrichten*
- *C : Arbeitnehmer:innen mit erhöhtem Verantwortungsbereich in leitenden Stellungen, inkl. Mitarbeiter:innenführung*

Im konkreten Fall der Firma ABC gibt es $n = 51$ Mitarbeiter:innen, darunter in der Verwendungsgruppe A 25 weibliche und 5 männliche Angestellte, in der Verwendungsgruppe B 5 weibliche und 7 männliche Angestellte und in Verwendungsgruppe C 4 Frauen und 5 Männer.

Das können wir auch in einer Tabelle darstellen:

	Geschlecht	
Verwendungsgruppe	*weiblich*	*männlich*
A	*25*	*5*
B	*5*	*7*
C	*4*	*5*

Für eine Kontingenztabelle ergänzen wir nun sowohl die Spalten als auch die Zeilen in obiger Tabelle um eine Summenspalte bzw. -zeile:

Beispiel 21 *(Fortsetzung)*

Verwendungsgruppe	*Geschlecht* *weiblich*	*männlich*	Σ
A	*25*	*5*	*30*
B	*5*	*7*	*12*
C	*4*	*5*	*9*
Σ	*34*	*17*	*51*

Anschließend bestimmen wir die so genannten **Randverteilungen**. Dazu geben wir zunächst die relativen Häufigkeiten (in Prozent) an, d.h. wir dividieren einfach jeden Wert in der Tabelle durch n (in unserem Beispiel: durch 51)

Beispiel 21 *(Fortsetzung)*

Verwendungsgruppe	*Geschlecht* *weiblich*	*männlich*	*RV*
A	*49.0%*	*9.8%*	***58.8%***
B	*9.8%*	*13.7%*	***23.5%***
C	*7.8%*	*9.8%*	***17.6%***
Σ	***66.7%***	***33.3%***	*100.0%*

Die Randverteilungsspalte *bedeutet: 58.8% der Mitarbeiter:innen sind in Verwendungsgruppe A angestellt, 23.5% in Verwendungsgruppe B und 17.6% in C. Und die* Randverteilungszeile*: 66.7% der Angestellten sind weiblich und 33.3% männlich.*

Unser Ziel ist, eventuelle *Zusammenhänge* zwischen den auftretenden Zufallsvariablen zu untersuchen. In unserem Beispiel können wir hinterfragen, ob es einen Zusammenhang zwischen dem Geschlecht und der Verwendungsgruppe gibt, oder ob diese beiden Merkmale unabhängig voneinander sind.

Die weitere Vorgangsweise schaut zugegebenermaßen auf den ersten Blick etwas kompliziert aus, tatsächlich lässt sie sich aber z.B. in EXCEL ziemlich einfach bewerkstelligen. Zunächst einmal überlegen wir, welche Häufigkeitsverteilung wir in den einzelnen Verwendungsgruppen erwarten würden, wenn es eine vom Geschlecht unabhängige Verteilung gäbe. Offensichtlich wären das sowohl unter den 34 Frauen als auch unter den 17 Männern jeweils 58.8% in Gruppe A, 23.5% in Gruppe B und 17.6% in Gruppe C.

Beispiel 21 *(Fortsetzung) Die* ***erwartete*** *Verteilung sieht so aus:*

Verwendungsgruppe	*Geschlecht*	
	weiblich	*männlich*
A	*20*	*10*
B	*8*	*4*
C	*6*	*3*

Wir sehen: Zwischen Realität und erwarteter Verteilung besteht ein Unterschied. Diese Differenzen rechnen wir zunächst aus und quadrieren sie anschließend. Und dann dividieren wir noch alle quadratischen Differenzen durch die erwarteten Werte:

Beispiel 21 *(Fortsetzung)*

$$\frac{(25-20)^2}{20} = 1.25 \quad \frac{(5-10)^2}{10} = 2.5$$
$$\frac{(5-8)^2}{8} = 1.13 \quad \frac{(7-4)^2}{4} = 2.25$$
$$\frac{(4-6)^2}{6} = 0.67 \quad \frac{(5-3)^2}{3} = 1.33$$

Wenn wir jetzt die **Summe** der eben berechneten Werte bilden, sind wir schon fast am Ziel:

Beispiel 21 *(Fortsetzung)*

$$\chi^2 = 1.25 + 2.5 + 1.13 + 2.25 + 0.67 + 1.33 = 9.13$$

Das Formelzeichen, das wir dafür verwendet haben, ist übrigens ein griechisches *Chi* (bzw. ein Chi zum Quadrat, ausgesprochen «Ki quadrat»); der Wert selbst ist ein Maß für die *Quadratische Kontingenz*. Letztendlich können wir aus dem χ^2 dann den **korrigierten Kontingenzkoeffizienten** ausrechnen:

$$C_{korr} = \sqrt{\frac{k}{k-1} \cdot \frac{\chi^2}{\chi^2+n}}$$
mit:
$k = \min(i; j)$ also die kleinere der beiden Zahlen i und j
$j =$ Anzahl der unterschiedlichen Kategorien der Zufallsvariablen X
$i =$ Anzahl der unterschiedlichen Kategorien der Zufallsvariablen Y (4.12)

Beispiel 21 *(Fortsetzung)*

In unserem Beispiel gibt es $j = 2$ *Geschlechter und* $i = 3$ *Verwendungsgruppen, die kleinere der beiden Zahlen ist* 2, *daher* $k = \min(3; 2) = 2$ *und*

$$C_{korr} = \sqrt{\frac{2}{2-1} \cdot \frac{9.13}{9.13+51}} = \underline{\underline{0.55}}$$

Der Wert deutet darauf hin, dass die beiden Merkmale nicht unabhängig voneinander sind.

Ob es sich dabei um eine signifikante *Abhängigkeit handelt, darauf werden wir in einem späteren Kapitel noch einmal zurückkommen.*

Zusammenfassend können wir sagen: Welches Maß wir für die Angabe des statistischen Zusammenhangs angeben können, ist abhängig vom Skalenniveau der Zufallsvariablen:

- Für *metrische* Daten können wir den **Bravais-Pearson Korrelationskoeffizienten** verwenden,
- für *Ordinaldaten* den **Rangkorrelationskoeffizienten nach Spearman**; und
- für *Nominaldaten* den **korrigierten Kontingenzkoeffizienten**.

Sind die beiden Zufallsvariablen auf unterschiedlichem Niveau, dann können wir nur jenes Merkmalsmaß verwenden, das für die Zufallsvariable auf niedrigerem Niveau möglich ist. Für den Zusammenhang zwischen einer metrischen Variable und einer rangskalierten, kann «nur» der Spearman-Koeffzient angegeben werden, nicht aber ein Bravais-Pearson-Koeffizient; für den Zusammenhang zwischen einer metrischen Variable und einer kategorialen nur ein Kontingenzkoeffizient.

4.5 Statistische und kausale Zusammenhänge

Die Korrelation beschreibt per se zunächst *statistische* und nicht unbedingt *kausale* Zusammenhänge. Das heißt selbst ein sehr, sehr hoher Wert des Korrelations- oder Kontingenzkoeffizienten (nahe ± 1) sagt nichts darüber aus, dass das eine Merkmal die *Ursache* für die Größe des anderen Merkmals ist. Natürlich *kann* eine kausale Beziehung bestehen, das muss aber nicht der Fall sein. Hier muss man unterscheiden, ob die Daten, die wir ausgewertet haben, aus einer reinen *Beobachtung* stammen oder aus einem gezielten *Experiment*.

Der Unterschied sei am Beispiel des «Mozarteffekts» erläutert:

Bei einer reinen Beobachtung fragen wir Studierende, wie oft und wie lange sie während des Lernens Musik von Mozart hören. Das vergleichen wir – individuell – mit der Anzahl der Punkte, die diese Studierenden auf die Tests bekommen haben, für die sie (mit oder ohne Mozart) gelernt haben. Das ergibt zwar statistische Zusammenhänge (vielleicht), aber keine kausalen.

Frances Rauscher, Gordon Shaw und Katherine Ky von der Universität von Irvine, Kalifornien, berichteten 1993 im Wissenschaftsjournal *Nature*, dass Studierende nach dem Anhören von Mozarts Sonate für zwei Klaviere, KV 448, in einem anschließenden Test über ihre Fähigkeiten zum räumlichen Denken signifikant höhere Leistungen erzielt hatten als ihre Kollegen, die entweder Entspannungsmusik zu hören bekamen oder überhaupt in aller Ruhe den Test absolvierten. (Rauscher Frances, Gordon Shaw und Katherine Ky. 1993. «Music and spatial task performance». In: *Nature* Vol.365, Oktober 1993, S.611). Nachdem das Thema von diversen Medien mit Begeisterung aufgenommen und verbreitet wurde, ließ sich ein geschäftstüchtiger Autor den Begriff «Mozart Effect» schützen und verdiente gut mit einem Buch und Vorträgen, in denen er der Macht Mozarts Musik gleich auch die Linderung von körperlichen Beschwerden und heilende Effekte im Fall von Aids, diversen Allergien und Diabetes versprach. (Mozart selbst war übrigens von Kindheit an immer wieder kränklich und starb 1791 mit nur 36 Jahren. Er hätte öfter seine eigene Musik hören sollen).
Hast du die Mozartsonate angehört (zum Beispiel unter t1p.de/mozartkv448, aber es hat nicht mit dem Zuwachs räumlicher Intelligenz oder der Verbesserung deiner Gesundheit geklappt, kannst du sie auch anderweitig verwenden: Ein Milchbauer aus der Nähe von Madrid beschallt seine 700 Kühe jeden Tag mit Mozart. «Es klappt nur mit Mozart», schwört Nicolas Siebert. Die Kühe seien nicht nur ausgeglichener und einfacher im Umgang, jede einzelne produziere auch ein bis sechs Liter mehr Milch pro Tag.

(Quellen: Swartz, Luke. 2000. *The Mozart Effect: Does Mozart Make You Smarter?* http://xenon.stanford.edu/~lswartz/mozarteffect.pdf. Sowie: Driessen, Barbara. 2008. *Mozart-Sonaten beruhigen Kinder und Kühe.* WELT ONLINE 28.2.2008. http://www.welt.de/wissenschaft/article1735411/)

Der *Mozart-Effekt:* Statistischer oder kausaler Zusammenhang?

In diesem Zusammenhang spricht man auch oft von einer **Scheinkorrelation**.

Als Statistiker:innen wissen wir, dass Zusammenhänge wie beim Mozart-Effekt zwar vielleicht tatsächlich aufzeigbar sind, dass es sich dabei aber eben um *statistische* Zusammenhänge handelt und nicht um *kausale*. Es kann zum Beispiel sein, dass Menschen, die intelligenter sind, auch eher klassische Musik hören, als Menschen mit einem niedrigen Intelligenzquotienten. Daraus kann aber nicht abgeleitet werden, dass ein wenig Mozart-Hören praktisch ohne sonstigen Aufwand die Intelligenz steigert. Auch zwischen dem gesundheitlichen Wohlbefinden und der Vorliebe für bestimmte Musik *kann* ein Zusammenhang bestehen, aber auch hier ist – zumindest mit der statistischen Methode der Korrelationsrechnung – keine Kausalitätsrichtung auszumachen.

Bei reiner Beobachtung gilt: Es lässt sich nichts über den kausalen Zusammenhang sagen.

Es gibt daher drei Möglichkeiten: Das Hören von Mozart führt zu besseren Lernergebnissen. Oder: Wer gut lernt, hört auch gerne Mozart. Oder: Es gibt eine dritte, uns unbekannte Variable, die zu einer Verbesserung der Lernergebnisse bei denen, die gerne Klassik hören, geführt hat. Diese dritte Variable nennen wir auch *Störfaktor* (engl.: *confounder*)[10].

Bei einem Experiment hingegen ginge das so: Zufällig ausgewählte Menschen bekommen Mozart vorgespielt (während des Lernens); eine andere Gruppe bekommt keine Musik. Wenn die beiden Gruppen wirklich zufällig ausgewählt wurden, finden sich in beiden Gruppen ungefähr gleich viele Klassikliebhaber wie solche, denen diese Musik nicht besonders gefällt. Wenn jetzt trotzdem die eine Gruppe einen besseren Lernerfolg zeigt, zeigt das einen kausalen Zusammenhang.

Korrelation bedeutet nicht Kausalität. Ob ein kausaler Zusammenhang besteht, ist nur aus der Art der Datenerhebung ableitbar: Aus einer Beobachtung alleine ist keine Kausalität ableitbar, aus einem Experiment hingegen kann eine Kausalität vermutet werden.

Aufgabe 15 In einem bestimmten Jahrgang wurden bei einer Analyse der Test- und Prüfungsergebnisse aus MAT101 (Mathematik) und MAT102 (Statistik) u.a. folgende Zusammenhänge beobachtet:

Korrelation zwischen den Gesamtpunkten aus MT122 und den im Vorsemester erreichten Punkten aus MAT101: $r = 0.37$ (Determinationskoeffizient: $r^2 = 14\%$).
Korrelation zwischen den aus den Online-Tests in MT122 erreichten Punkten und der Zeit, die im Durchschnitt für die Bearbeitung der Online-Tests aufgewandt wurde: $r = -0.08$ (Determinationskoeffizient: $r^2 = 1\%$).

Was lässt sich daraus über den Zusammenhang zwischen Mathematik- und Statistik-Kenntnissen bzw. den Zeitaufwand, den Studierende für die Online-Tests aufwenden, sagen?

[10] Wobei die deutsche Bezeichnung *Störfaktor* ein wenig unglücklich ist, weil ja keine Störung im Sinne eines *Störenfrieds* vorliegt, der das Ergebnis des Experiments verunmöglicht, sondern lediglich ein Faktor von außen einen Einfluss nimmt, an den wir nicht gedacht haben.

5 Zufälliges, Wahrscheinliches und Normales

In diesem Kapitel wollen wir uns den *Zufall* ein wenig näher anschauen und auch die Theorie der damit im Zusammenhang stehenden *Wahrscheinlichkeit* für das Eintreffen eines Ereignisses, das wir als «zufällig» einstufen.

5.1 Zufall

Beim Finale der UEFA Champions League am 19. Mai 2012 in München spielten in der Mannschaft des Chelsea FC gleich zwei «Geburtstagspaare», also je zwei Spieler, die am selben Tag Geburtstag feiern: Salomon Kalou und Ryan Bertrand (5. August) sowie David Luiz und John Obi Mikel (22. April). Die beiden letztgenannten sind sogar nicht nur am selben Tag sondern auch im selben Jahr (1987) auf die Welt gekommen.
Inklusive des Schiedsrichters und der (tatsächlich zum Einsatz gekommenen) Ersatzspieler waren 27 Personen am Spielfeld tätig. Würdest du – hätte ich es nicht bereits verraten – darauf wetten, dass mindestens zwei von ihnen am selben Tag Geburtstag haben?

So ein *Zufall*: Geburtstags(zu)fälle: Intuitiv empfinden wir das als außergewöhnliches Zusammentreffen – und als *großen Zufall*.

Der Begriff **Zufall** beim Beobachten einer *Zufalls*größe soll unterstreichen, dass das Ergebnis dieser Beobachtung nicht vorhersehbar oder *deterministisch*[1] ist.

[1] vom lat. *determinare* = bestimmen, festsetzen. Ein «deterministisches» Ergebnis bedeutet: Es gibt einen funktionalen Zusammenhang zwischen den Eingangsparametern und dem Ergebnis und wenn wir alle Eingangsparameter kennen, kennen wir auch das Ergebnis.

Wenn wir zum Beispiel eine Münze werfen, wissen wir nicht, ob wir Kopf oder Zahl erhalten. Wenn wir auf der Hauptseite von *de.wikipedia.org* auf den Link «Zufälliger Artikel» klicken (oder gleichzeitig die Tasten `alt-shift-x`), rufen wir irgendeinen, zufällig ausgewählten Artikel auf. Alles zufällige Ereignisse. Aber auch andere, kompliziertere Dinge wie der Wert eines Aktienindex oder die Paketumlaufzeit in einem IP-Netzwerk sind – im Sinne der Statistik – zufällige Ereignisse.

Trotz aller Zufälligkeit würden wir gerne herausfinden, ob es nicht doch irgendwelche «Gesetzmäßigkeiten» gibt, wie wir dem Zufall auf die Schliche kommen und ihn in den Griff bekommen könnten.

Wenn wir die Geburtstage von Fußballspielern «beobachten», ist das aus statistischer Sicht dasselbe, wie wenn wir mit einem 365-seitigem Würfel würfeln oder aus einer Urne mit 365 verschiedenen Kugeln (mit jeweils einem aufgedruckten Tagesdatum) eine beliebige Kugel herausziehen. Beides – «Würfeln» und «Ziehen aus einer Urne» – wird oft als anschauliches Denkmodell für ein *Zufallsexperiment* (auch: *zufälliger Versuch*) verwendet. Ein **Zufallsexperiment** ist ein Vorgang, der – zumindest im Prinzip – beliebig oft wiederholbar ist und die jeweiligen Ergebnisse sind innerhalb einer Menge möglicher Ausgänge ungewiss, eben zufällig. Das Ergebnis eines zufälligen Versuches bezeichnen wir dann als ein **Zufallsereignis** E.

Der einzelne Wert, den die Zufallsgröße nach der Beobachtung (als Ergebnis des Zufallsexperiments) annimmt, ist die **Realisierung** x der Zufallsgröße X. Beim Zufallsexperiment «Würfeln» und Beobachtung der Zufallsvariable $X = \mathit{Augenzahl}$ können wir zum Beispiel im 3. Versuch die Zahl 4 erhalten. $x_3 = 4$ ist dann eine Realisierung der Zufallsvariable $\mathit{Augenzahl}$. Ein anderes Beispiel für die Realisierung einer Zufallsgröße ist der tägliche Schlusskurs des *Dow Jones Industrial Average*, nämlich der Zufallsgröße «Kursindex der dreißig größten US-Unternehmen am Ende eines Börsentages an der New York Stock Exchange».

Realisierungen von Zufallsgrößen sind selbst übrigens nicht mehr zufällig. Sie haben ja einen bestimmten Wert.

In der Alltagssprache bezeichnen wir meist Ereignisse, die ohne offensichtlichen Grund eintreten, und deren Eintritt uns als ziemlich unwahrscheinlich scheint, als «zufällig». Zum Beispiel wenn in einer Fußballmannschaft zwei Spieler am selben Tag Geburtstag haben, oder wenn man sechsmal hintereinander würfelt und dabei genau die Abfolge $1, 2, 3, 4, 5, 6$ erhält. Tatsächlich ist aber zum Beispiel auch die Abfolge $3, 3, 5, 2, 6, 1$ genauso «zufällig».

Um den Zufall mathematisch-statistisch beschreiben und modellieren zu können, benötigen wir die Wahrscheinlichkeitsrechnung.

Zur erstmaligen Verwendung des Wortes «Lockdown» kam es laut dem deutschen Sprachwissenschafter Anatol Stefanowitsch bei der Beschreibung eines Vorfalls in einem US-Gefängnis 1973: Am 6. Dezember dieses Jahres wurde durch mehrere Mithäftlinge 32 mal auf einen Häftling eingestochen, der einen «Kollegen» unabsichtlich angefahren und sich dafür nicht entschuldigt hatte. Daraufhin wurden zunächst alle Gefangenen für einige Zeit in ihren Zellen eingesperrt, was als «Lockdown» bezeichnet wurde.
Wirklich spooky an der Sache aber: Der Gefangene, der Opfer dieser Messerattacke wurde, hieß Juan Vallejo *Corona*.

Was 1973 aus Sicht der restlichen Welt vermutlich nur eine unbedeutende Episode war, klingt beinahe 50 Jahre später nach einem aberwitzigen Zufall.

5.2 Ein bisschen Wahrscheinlichkeitsrechnung

Was hinter einer «Wahrscheinlichkeit» steckt, davon hat vermutlich jeder seine subjektive Vorstellung. Man überlegt beispielsweise, wie wahrscheinlich es ist, dass man durch die Abschlussprüfungen am Ende des Semesters kommt, dass im Urlaub im Salzkammergut eine Woche lang die Sonne scheint, oder dass man im Lotto gewinnt. Und obwohl die Wahrscheinlichkeit für Letzteres wirklich sehr klein ist[2], gibt es mehr Lottospieler als Urlauber im Salzkammergut.

Auch beim Festlegen von Börsenstrategien und Devisengeschäften rechnet man – bewusst oder unbewusst – mit Wahrscheinlichkeiten (dort nennt man es «Spekulieren»), aber auch wenn man die Erfolgschancen für ein neues Produkt einschätzt oder bei anderen Marketingentscheidungen. Die Wahrscheinlichkeitsrechnung spielt in der Versicherungsmathematik eine Rolle, bei der Qualitätskontrolle, bei der Optimierung von Produktion und Lagerhaltung und so fort. Trotz ihrer Wichtigkeit ist sie bei Wirtschafts- und Informatikstudierenden eher unbeliebt (insbesondere als Prüfungsgegenstand), was vielleicht auch daran liegt, dass es mehrere Definitionen für sie gibt. Die «subjektive Wahrscheinlichkeitsdefinition» haben wir bereits angesprochen, es gibt aber natürlich auch eine «richtige» Definition:

Sie stammt von *Laplace*[3] und gibt folgendes Verhältnis wieder:

$$P(E) = \frac{\textit{Zahl der möglichen Eintrittsfälle von } E}{\textit{Gesamtzahl aller überhaupt möglichen Ausgänge}} \tag{5.1}$$

[2] Die Chance auf einen 6er bei «6 aus 45» beträgt 1:8.145.060. Dennoch hat 2008 eine Kärntnerin bei einem Sechser rund 1.7 Millionen Euro gewonnen, nachdem ihr Ehemann bereits 1999 einen Sechser mit umgerechnet rund 1 Million Euro getippt hatte.

[3] *Pierre-Simon Marquis de Laplace*, frz. Mathematiker, Astronom und Physiker, 1749-1827

In *MS Excel* und *LibreOffice Calc* können wir auf zwei Arten Zufallszahlen erzeugen:
Die Funktion `=ZUFALLSZAHL()` gibt eine reelle Zufallszahl größer oder gleich 0 und kleiner als 1 zurück.
Mit `=ZUFALLSBEREICH(UntereZahl; ObereZahl)` erhalten wir eine *ganze* Zahl, die (zufällig) irgendwo zwischen den beiden angegebenen Grenzen liegt. (Dabei sind die untere und obere Grenze in den möglichen Zahlen ebenfalls inkludiert). Die Zufallszahl kann (bei entsprechender Angabe der beiden Grenzen) auch negativ sein, ist aber in jedem Fall eine ganze Zahl.
Will man eine reelle Zufallszahl zwischen den Grenzen a und b haben, muss man ein wenig kreativ sein und angeben: `=ZUFALLSZAHL()*(b-a)+a`.

In *R* gibt es mehrere (und «ausgefeiltere») Methoden, um Zufallszahlen zu erzeugen. Mit `sample(10:20,2)` erhalten wir zum Beispiel 2 zufällige Zahlen aus dem Intervall von 10 bis 20. Weitere Funktionen sind zum Beispiel `rnorm` oder `runif`, zu deren Bedeutung wir später kommen.

Falls du zufällig einmal eine Zufallszahl brauchst: Der PC kann aushelfen. Oder das Internet unter *www.random.org* oder *www.randomnumbers.info*

$P(E)$ ist die **Wahrscheinlichkeit**[4] für das Eintreten des Zufallsereignisses E.

Formel 5.1 sieht sehr einfach aus und wir können gleich ein Beispiel rechnen:

Beispiel 22 *Bei der Fußball-Europameisterschaft 1968 in Italien wurde nach einem 0:0 im Halbfinale zwischen Italien und der UdSSR per Münzwurf entschieden, dass Italien ins Endspiel aufstieg (Das es schließlich auch gewann). Und auch bei der Fußball-EM der Frauen 2013 in Schweden musste nach den Vorrundenspielen per Münzwurf entschieden werden, ob Dänemark oder Russland ins Viertelfinale einziehen wird (Der Zufall brachte Dänemark Glück).*

Unter der Annahme, dass sich die Münze bei einem Losentscheid nicht buchstäblich in Luft auflöst oder in ein Erdloch fällt (und auch nicht auf der Kante zu stehen kommt[5]): Wie groß ist bei einem Münzwurf die Wahrscheinlichkeit für «Kopf»?

Es gibt 2 mögliche Ausgänge (nämlich «Kopf» oder «Zahl»), und einen Eintrittsfall von

[4] Das P kommt vom lat. *probabilitas* = Wahrscheinlichkeit
[5] wie zum Beispiel 1965 beim Europapokal-Entscheidungsspiel zwischen dem 1. FC Köln und dem FC Liverpool, siehe https://youtu.be/-P8zft3Yc1M

Im 17. Jhdt. wurde *Blaise Pascal* (frz. Mathematiker, 1623-1662) vom frz. Schriftsteller und Berufsspieler *Antoine Gombaud Chevalier de Méré* (1607-1684) mit der Frage konfrontiert, wie der Einsatz bei einem bestimmten Würfelspiel fairerweise aufzuteilen ist, wenn das Spiel vorzeitig abgebrochen werden muss. Es ging also um die Frage nach der Wahrscheinlichkeit, mit der jeder Teilnehmer das Spiel gewinnen würde, wenn es fortgesetzt werden würde. Pascal beriet sich daraufhin in mehreren Briefwechseln mit seinem Kollegen *Pierre de Fermat* (frz. Mathematiker und Jurist, 1607-1665). Damit war die *Wahrscheinlichkeitsrechnung* geboren.

Zurück zum Ursprung: Die Anfänge der Wahrscheinlichkeitsrechnung

In Vernon River-Stratford, einem Wahlbezirk in der ostkanadischen Provinz Prince Edward Island, erreichten am 4.5.2015 Mary Ellen McInnis und Alan McIsaac jeweils 1.173 Stimmen. Daraufhin wurde per Münzwurf entschieden, wer den Sitz im Provinzparlament erhalten sollte.
Dabei sieht die Wahlordnung vor, dass sich aus der alphabetischen Reihenfolge der Nachnamen ergibt, welchem Kandidaten «Kopf» und welchem «Zahl» zugeordnet wird. Und selbst da wurde der Zufall ziemlich herausgefordert. Unterscheiden sich die beiden Namen McInnis und McIsaac doch erst ab dem vierten Buchstaben.
Letztlich fiel die Münze auf «Zahl» und bescherte somit Alan McIsaac das Mandat.

Der Zufall gibt sich manchmal auch demokratisch.

E (nämlich «Kopf»). Somit:

$$P(E) = \frac{\textit{Eintrittsfälle}}{\textit{Ausgangsmöglichkeiten}} = \frac{1}{2} = 0.5 = 50\%$$

Aufgabe 16 In Oberösterreich gibt es insgesamt 6 630 Orte. 40 davon heißen «Au». Angenommen alle 6 630 Orte werden auf Kärtchen geschrieben und daraus eine beliebige Karte herausgezogen. Wie groß ist die Wahrscheinlichkeit, dass darauf *Au* steht?

Aufgabe 17 Unter der Annahme, dass beim Würfeln jede Augenzahl gleich wahrscheinlich ist: Wie groß ist die Wahrscheinlichkeit, eine gerade Zahl zu würfeln?

In der Bulgarischen Lotterie, in der man 6 aus 49 Zahlen tippt, wurden am 6. September 2009 und bei der darauffolgenden Ziehung am 10. September 2009 exakt dieselben Zahlen gezogen (4, 15, 23, 24, 35, und 42). Mag das schon unglaublich genug sein, ist es mindestens ebenso unerwartet, dass beim zweiten mal gleich 18 Spieler einen Sechser erzielten, also entgegen jeglicher landläufiger Vorstellung von Wahrscheinlichkeiten dieselben Zahlen wie bei der vorherigen Ziehung getippt hatten. (Für ihren Sechser gewannen sie umgerechnet jeweils ca. 5 200 EUR).

Auch Ereignisse, die nur mit einer Wahrscheinlichkeit von 1:4.2 Millionen auftreten, treten unwahrscheinlicherweise irgendwann auf.

Neben einer Definition benötigen wir für das Berechnen von und Rechnen mit Wahrscheinlichkeiten noch Rechenregeln. Die einfachsten[6] sind:

$$P(E) + P(\textit{not } E) = 1 \tag{5.2}$$

$$P(\textit{not } E) = 1 - P(E) \tag{5.3}$$

$$P(E_1 \textit{ and } E_2) = P(E_1) \cdot P(E_2) \tag{5.4}$$

$$P(E_1 \textit{ or } E_2) = P(E_1) + P(E_2) - P(E_1) \cdot P(E_2) \tag{5.5}$$

Hinweis: Wenn du die Verwendung von $+$ und $\cdot$ an die Symbolik der Booleschen Algebra aus MAT101 erinnert, ist das kein Zufall. So wie dort verwenden wir auch hier das Pluszeichen für *oder* (OR) und das Produkt für *und* (AND) und haben sogar ein *Komplement*: die Wahrscheinlichkeit, dass E *nicht* eintrifft. Wir bezeichnen sie mit $P(\textit{not } E)$.

Konkret bedeuten die Formeln (5.2) bis (5.5) somit:

1. Gehen wir davon aus, dass E entweder eintreffen kann oder nicht (es aber keinen dritten Fall geben kann[7], dann ist die Summe aus $P(E) + P(\textit{not } E)$ gleich 1.

2. Kennen wir die Wahrscheinlichkeit für das Eintreffen des Ereignisses E und sie beträgt $P(E)$, so können wir aus Formel (5.2) unmittelbar Formel (5.3) ableiten und damit die Wahrscheinlichkeit angeben, dass E *nicht* eintrifft: Wir rechnen uns einfach die Ergänzung auf 1 aus. Wir nennen das auch die **Gegenwahrscheinlichkeit**.

[6] Das sind wirklich nur die einfachsten Regeln. Es gibt dann zum Beispiel noch welche für eine *bedingte Wahrscheinlichkeit*, für die *totale Wahrscheinlichkeit*, oder für den so genannten *Satz von Bayes*, und noch vieles mehr. Für uns sind die obigen Regeln aber zunächst ausreichend.

[7] Ein «dritter Fall» wäre zum Beispiel *E trifft gleichzeitig ein und nicht*.

3. Kennen wir die Wahrscheinlichkeiten für das Eintreffen der Ereignisse E_1 und E_2 und betragen sie jeweils $P(E_1)$ und $P(E_2)$, so können wir mit Formel (5.4) die Wahrscheinlichkeit angeben, dass *sowohl* E_1 *als auch* E_2 eintreffen: Dazu multiplizieren wir die Einzelwahrscheinlichkeiten.

4. Kennen wir die Wahrscheinlichkeiten für das Eintreffen der Ereignisse E_1 und E_2 und betragen sie jeweils $P(E_1)$ und $P(E_2)$, so können wir mit Formel (5.5) die Wahrscheinlichkeit angeben, dass *entweder* E_1 *oder* E_2 eintreffen: Wir addieren dazu einfach die Einzelwahrscheinlichkeiten und ziehen davon die Wahrscheinlichkeit, dass E_1 *und* E_2 eintreffen wieder ab.

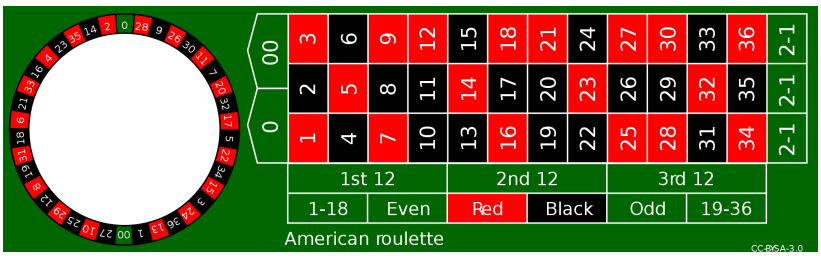
Alle genannten Regeln gelten für *unabhängige* Ereignisse und von solchen unabhängigen Ereignissen gehen wir hier aus. Das bedeutet, dass die Wahrscheinlichkeit für das Eintreffen von E_1 unabhängig davon ist, ob E_2 eingetroffen ist oder nicht und umgekehrt. Das ist zugegebenermaßen manchmal auf den ersten Blick nicht ganz offensichtlich. Würden heute die exakt selben sechs Lottozahlen gezogen werden wie bei der letzten Ziehung vorige Woche, würde das vermutlich viel Erstaunen auslösen (siehe S.102). Tatsächlich ist es aber völlig egal, welche Zahlen vorige Woche gezogen wurden. Die Lottomaschine hat kein «Gedächtnis», sondern geht nächste Woche wieder völlig unbedarft an ihre Arbeit.

Beispiel 23 *Wir wissen: Beim Würfeln mit einem Würfel beträgt die Wahrscheinlichkeit, einen 6er zu würfeln,* $1/6$*. Wie groß ist die Wahrscheinlichkeit, keinen 6er zu würfeln?*

Wenn wir schon wissen, dass die Wahrscheinlichkeit für einen 6er gleich $1/6$ *ist, dann ist die Wahrscheinlichkeit, keinen 6er zu würfeln nach Formel 5.3 gleich*

$$P(\text{not 6er}) = 1 - P(\text{6er}) = 1 - 1/6 = 5/6 = 83.3\%$$

Aufgabe 18 Beim American Roulette gibt es je 18 rote und schwarze Nummernfelder sowie zwei grüne Felder mit einer Null (siehe nachfolgende Abbildung[8]).



[8]Bildquelle: American roulette table and wheel layout. First published by Betzaar.com under the CreativeCommons Attribution+ShareAlike licence

Gib an:

1. Wie groß ist die Wahrscheinlichkeit auf eine «rote» Zahl?
2. Wie groß ist die Wahrscheinlichkeit auf eine Primzahl?
3. Wie groß ist die Wahrscheinlichkeit auf eine rote Primzahl?

Beispiel 24 *Im einleitenden Beispiel auf Seite 97 haben wir uns gefragt, wie groß die Wahrscheinlichkeit ist, dass zwei Personen (aus einer Gruppe von 27) am selben Tag Geburtstag haben.*

Mit der Definition (5.1) und den Formeln (5.3) und (5.4) können wir das nun angeben (wobei wir der Einfachheit halber Schaltjahre nicht berücksichtigen):

Bei zwei Personen beträgt die Wahrscheinlichkeit, dass sie nicht *am selben Tag Geburtstag haben:* $\frac{364}{365} = 0.997$. *(Die zweite Person hat unter allen 365 Tagen des Jahres noch 364 Möglichkeiten, die nicht mit dem Tag der ersten Person übereinstimmen). Also ist die Wahrscheinlichkeit, dass sie* am selben Tag *Geburtstag haben die Gegenwahrscheinlichkeit* $(1 - 0.997) = 0.003$ *oder* 0.3%.

Bei drei Personen darf die dritte weder mit der ersten noch mir der zweiten Person am gleichen Tag Geburtstag feiern. Diese Wahrscheinlichkeit ist $\frac{364}{365} \cdot \frac{363}{365} = 0.992$*; die gesuchte Gegenwahrscheinlichkeit ist dann* 0.8%.

Für 27 Personen gilt: $\frac{364}{365} \cdot \frac{363}{365} \cdot \ \cdots \ \cdot \frac{339}{365} = 37\%$. *Das ist die Wahrscheinlichkeit, dass es unter den 27 Personen kein «Geburtstagspaar» gibt, bzw. umgekehrt: Die Wahrscheinlichkeit, dass unter 27 Personen zwei am selben Tag Geburtstag haben, beträgt* 63%.

Nimmt man noch alle Ersatzspieler, die beiden Trainer und die beiden Schiedsrichterassistenten dazu, kommt man auf 41 Personen, die beim Champions League Finale 2012 am Fußballfeld herumliefen, und da ist die Wahrscheinlichkeit, dass zwei am selben Tag Geburtstag haben, bereits über 90%*, wie sich leicht ausrechnen lässt.*

Vermutlich bist du ein wenig überrascht, dass bei 41 zufällig anwesenden Personen die Wahrscheinlichkeit, dass zwei an einem beliebigen aber selben Tag Geburtstag haben, so hoch ist – schließlich hat das Jahr ja 365 Tage. Aber du bist mit deiner Überraschung nicht alleine: Das Phänomen trägt auch den Namen «Geburtstagsparadoxon».

Beispiel 24 zeigt uns aber noch ein weiteres Dilemma: Während Formel 5.1 an sich ja ziemlich einfach ist, ist es manchmal schwierig bis unmöglich, die exakte Anzahl der Eintrittsfälle und Ausgangsmöglichkeiten herauszufinden. Im Geburtstagsbeispiel war es mit viel Nachdenken gerade noch möglich, aber Vieles,

worüber wir Wahrscheinlichkeiten angeben wollen, ist überhaupt nicht zählbar. Wenn man sich zum Flughafen aufmacht, um eine Urlaubsreise zu beginnen, will man auf keinen Fall zu spät kommen und schätzt daher die Wahrscheinlichkeit, mit dem Auto in einen Stau zu kommen. Aber was sind da die Eintrittsfälle und die Ausgangsmöglichkeiten, die man ins Verhältnis zueinander setzen kann? Oder wenn für einen bestimmten Ort eine Regenwahrscheinlichkeit berechnet wird (vgl. Abb.5.1) – was wurde da abgezählt?

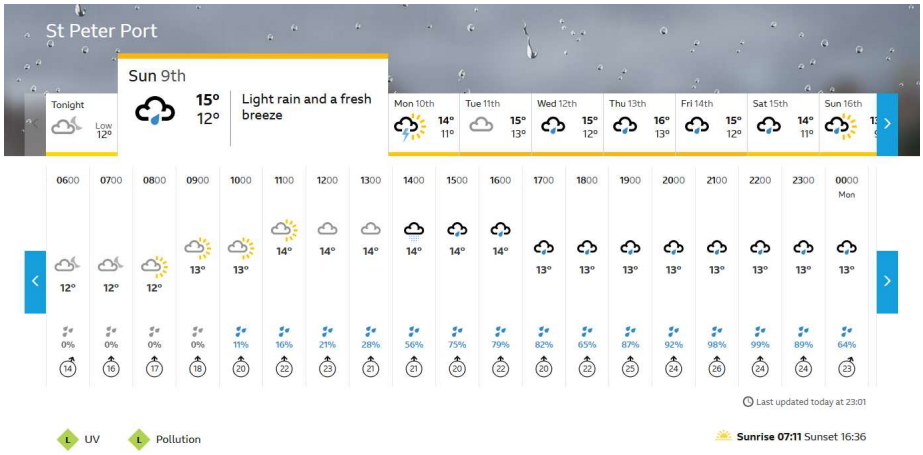


Abb. 5.1: Offenbar kann für jede Stunde eines bestimmten Tages eine Regenwahrscheinlichkeit angegeben werden. Aber wie kann man sie berechnen? (Quelle: www.bbc.com/weather)

Wir benötigen offenbar noch eine weitere Definition für die Wahrscheinlichkeit. Dazu kannst an dieser Stelle selbst ein kleines Experiment starten und eine Münze werfen und notieren, wie oft sie auf «Kopf» und wie oft auf «Zahl» fällt. Nach der klassischen Definition der Formel 5.1 beträgt die (theoretische) Wahrscheinlichkeit für «Kopf» $\frac{1}{2} = 0.5 = 50\%$. Das bedeutet aber nicht, dass wir im praktischen Experiment in genau 50% der Fälle Kopf erhalten, und schon gar nicht, dass wir dies genau bei jedem zweiten Wurf tun. Die Abfolge könnte zum Beispiel so aussehen (25 Würfe):

$$\{Z, Z, K, K, Z, Z, Z, Z, Z, Z, K, K, Z, K, Z, K, K, K, K, K, Z, Z, Z, Z, K\}$$

Im nächsten Schritt erinnern wir uns an die relative Häufigkeit h aus Formel 2.5 (Seite 32). Wir erhalten sie, wenn wir die absolute Häufigkeit durch die Gesamtzahl der Elemente in der Stichprobe dividieren. Wir geben für unser Experiment die relative Häufigkeit für «Kopf» an:

Nach dem ersten Wurf (Z) beträgt sie 0, ebenso nach den ersten beiden (Z, Z). Nach dem dritten Wurf (Z, Z, K) 0.33, nach dem vierten (Z, Z, K, K) 0.5 etc. Nach dem 10. Wurf beträgt die relative Häufigkeit für «Kopf» 0.2, nach allen 25 Würfen 0.44. In Abb. 5.2 sind die relativen Häufigkeiten eingezeichnet.

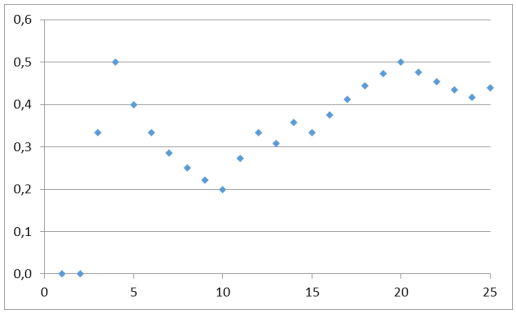


Abb. 5.2: Die relative Häufigkeit für «Kopf» bei 25maligem Münzwurf variiert in unserem Experiment anfänglich ziemlich stark.

Wir setzen das fort und werfen 50, 100, 500 und 1000 mal eine Münze und stellen die relativen Häufigkeiten wieder jeweils in einem Diagramm dar (Abb.5.3).

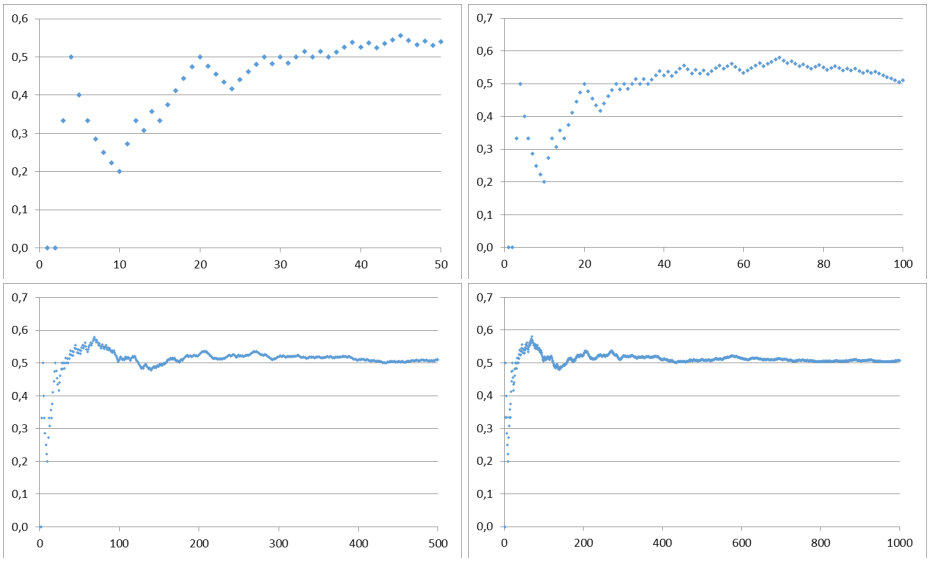


Abb. 5.3: Relative Häufigkeit für «Kopf» bei 50-/100-/500- und 1000maligem Münzwurf

Wir sehen: Je öfter wir die Münze werfen, desto eher nähert sich die relative Häufigkeit an die 50%-Linie an, also an die weiter oben (Beispiel 22) angegebene theoretische Wahrscheinlichkeit für das Auftreten von «Kopf».

Stellen wir uns jetzt vor, dass im Quotienten der Formel (2.5) das n überhaupt unendlich groß wird, wir also unendlich viele Elemente in der Stichprobe (in unserem Beispiel: unendlich viele Münzwürfe) hätten. Mathematisch bedeutet dies, dass wir den *Grenzwert* der relativen Häufigkeit für «n gegen ∞» angeben. Wir nennen diesen Grenzwert P, genauer $P(E)$:

$$\lim_{n\to\infty} \frac{k}{n} = P(E) \tag{5.6}$$

$P(E)$ ist eine Maßzahl für die Charakterisierung der Häufigkeit des Auftretens des Zufallsereignisses E – die **Wahrscheinlichkeit** für E. In Worten ausgedrückt lautet die Definition 5.6:

Mit wachsender Größe der Stichprobe konvergiert die relative Häufigkeit gegen die Wahrscheinlichkeit.

Diese Definition der Wahrscheinlichkeit stammt von *Richard von Mises*[9].

Er beruft sich dabei auf das **Gesetz der großen Zahlen**, das besagt, dass die theoretische Wahrscheinlichkeit $P(E)$ umso besser geschätzt werden kann, je mehr unabhängige Ausführungen des Zufallsexperimentes durchgeführt werden. Oder anders ausgedrückt: Wenn man weiß, dass zum Beispiel beim Werfen einer Münze die Wahrscheinlichkeit auf «Kopf» 50% beträgt, heißt das bei 10 oder 20 Münzwürfen noch nicht viel. Aber umso öfter man die Münze wirft, desto eher wird sie tatsächlich in der Hälfte der Fälle auf «Kopf» fallen.

Für die absolute Häufigkeit k gilt:

$$0 \leq k \leq n \tag{5.7}$$

und somit – wenn ich alles durch n dividiere – für die relative Häufigkeit:

$$0 \leq \frac{k}{n} \leq 1 \tag{5.8}$$

Das gilt in jedem Fall, egal wie groß n ist. Auch für $n \to \infty$ ändern sich diese Grenzen nicht, und daher gilt auch für die Wahrscheinlichkeit:

$$0 \leq P(E) \leq 1 \tag{5.9}$$

Die Wahrscheinlichkeit ist also eine reelle Zahl größer gleich Null und kleiner gleich Eins. Mit $P(E)$ können wir dann auf einer Skala von 0 bis 1 (bzw. 0% bis 100%) angeben, wie wahrscheinlich ein bestimmtes Ergebnis für zufällige Ereignisse und Experimente ist, bei denen es mehrere mögliche Ausgänge gibt. Dabei ist ein Ereignis, dem die Wahrscheinlichkeit 1 (bzw. 100%) zugeordnet ist, ein **sicheres Ereignis** ist, jenes mit der Wahrscheinlichkeit 0 (0%) ein **unmögliches Ereignis**. Umso näher $P(E)$ bei 1 liegt, desto wahrscheinlicher ist es, dass das Ereignis E stattfindet; je näher es bei 0 liegt, desto größer ist die Wahrscheinlichkeit, dass es gar nicht eintrifft. Was aber nicht bedeutet, dass es unmöglich ist – schließlich gewinnt auch fast jede Woche jemand beim Lotto, obwohl die Wahrscheinlichkeit dafür nicht sehr hoch ist…

[9] österr.-amerikan. Mathematiker und Philosoph, 1883-1953

5.3 Die Wahrscheinlichkeitsverteilung von Zufallsgrößen

Sehen wir uns zunächst wieder ein Beispiel an: 30 Personen haben jeweils 100mal eine Münze geworfen und notiert, wie oft sie auf «Kopf» oder «Zahl» gefallen ist. Die Zufallsgröße, die wir in Folge näher betrachten wollen, ist die pro Person jeweils erzielte Anzahl des Ereignisses «Kopf». Gefühlsmäßig würden wir wohl erwarten, dass das bei den meisten Personen 50 Mal eintritt[10].

Das tatsächliche Ergebnis des Experiments ist in der Häufigkeitstabelle 5.1 gegeben.

Anz. Kopf	k	h	Anz. Kopf	k	h	Anz. Kopf	k	h
30	0	0	43	1	0.033	56	1	0.033
31	0	0	44	0	0	57	0	0
32	0	0	45	2	0.067	58	2	0.067
33	0	0	46	3	0.100	59	0	0
34	0	0	47	1	0.033	60	1	0.033
35	0	0	48	2	0.067	61	0	0
36	0	0	49	1	0.033	62	1	0.033
37	0	0	50	4	0.133	63	0	0
38	0	0	51	3	0.100	64	0	0
39	0	0	52	3	0.100	65	0	0
40	1	0.033	53	2	0.067	66	0	0
41	0	0	54	1	0.033	67	0	0
42	0	0	55	1	0.033	68	0	0

Tabelle 5.1: Absolute und relative Häufigkeiten eines Zufallsexperiments, bei dem 30 Spieler je 100 Mal eine Münze warfen. Die beobachtete Zufallsgröße ist dabei die Anzahl der «Kopf-Würfe». Kein Spieler hat in diesem Experiment weniger als 40 mal Kopf geworfen und keiner öfter als 62 mal.

In einem nächsten Schritt lassen wir das Experiment von 500 Personen durchführen, die wieder je 100mal eine Münze werfen. Das Ergebnis ist die Häufigkeitstabelle 5.2:

[10] Wobei dieses Gefühl auch trügerisch sein kann. *Bartovs et al.* haben 2023 in 350.757 Experimenten gezeigt, dass eine geworfene Münze eher dazu neigt, auf der gleichen Seite zu landen, auf der sie am Beginn des Wurfes gelegen ist. Konkret war das 178.079 mal der Fall, also in 51% der Würfe. Siehe: browse.arxiv.org/pdf/2310.04153v3.pdf

Anz. Kopf	k	h	Anz. Kopf	k	h	Anz. Kopf	k	h
30	0	0	43	13	0.026	56	17	0.034
31	1	0.002	44	20	0.04	57	17	0.034
32	0	0	45	27	0.054	58	12	0.024
33	2	0.004	46	28	0.056	59	3	0.006
34	0	0	47	37	0.074	60	3	0.006
35	2	0.004	48	27	0.054	61	1	0.002
36	0	0	49	36	0.072	62	1	0.002
37	0	0	50	41	0.082	63	1	0.002
38	5	0.01	51	40	0.08	64	2	0.004
39	3	0.006	52	42	0.084	65	1	0.002
40	4	0.008	53	43	0.086	66	0	0
41	6	0.012	54	29	0.058	67	0	0
42	13	0.026	55	23	0.046	68	0	0

Tabelle 5.2: Absolute und relative Häufigkeiten eines Zufallsexperiments, bei dem 500 Spieler je 100 Mal eine Münze warfen. Keine Person hat weniger als 31 oder öfter als 65 mal Kopf geworfen.

Schließlich lassen wir 10.000 Testpersonen Münzen werfen. Das Ergebnis sehen wir in Tabelle 5.3.

Anz. Kopf	k	h	Anz. Kopf	k	h	Anz. Kopf	k	h
30	0	0	43	301	0.0301	56	398	0.0398
31	2	0.0002	44	385	0.0385	57	321	0.0321
32	2	0.0002	45	471	0.0471	58	245	0.0245
33	5	0.0005	46	575	0.0575	59	168	0.0168
34	5	0.0005	47	632	0.0632	60	105	0.0105
35	12	0.0012	48	751	0.0751	61	67	0.0067
36	14	0.0014	49	818	0.0818	62	43	0.0043
37	32	0.0032	50	817	0.0817	63	32	0.0032
38	36	0.0036	51	776	0.0776	64	15	0.0015
39	77	0.0077	52	750	0.075	65	8	0.0008
40	110	0.011	53	647	0.0647	66	6	0.0006
41	158	0.0158	54	531	0.0531	67	5	0.0005
42	222	0.0222	55	458	0.0458	68	0	0

Tabelle 5.3: Absolute und relative Häufigkeiten eines Zufallsexperiments, bei dem 10 000 Spieler je 100 Mal eine Münze warfen. Minimum an «Kopf-Würfen»: 31, Maximum: 67

Die relativen Häufigkeiten können wir auch in einem Diagramm darstellen (Abb.5.4):

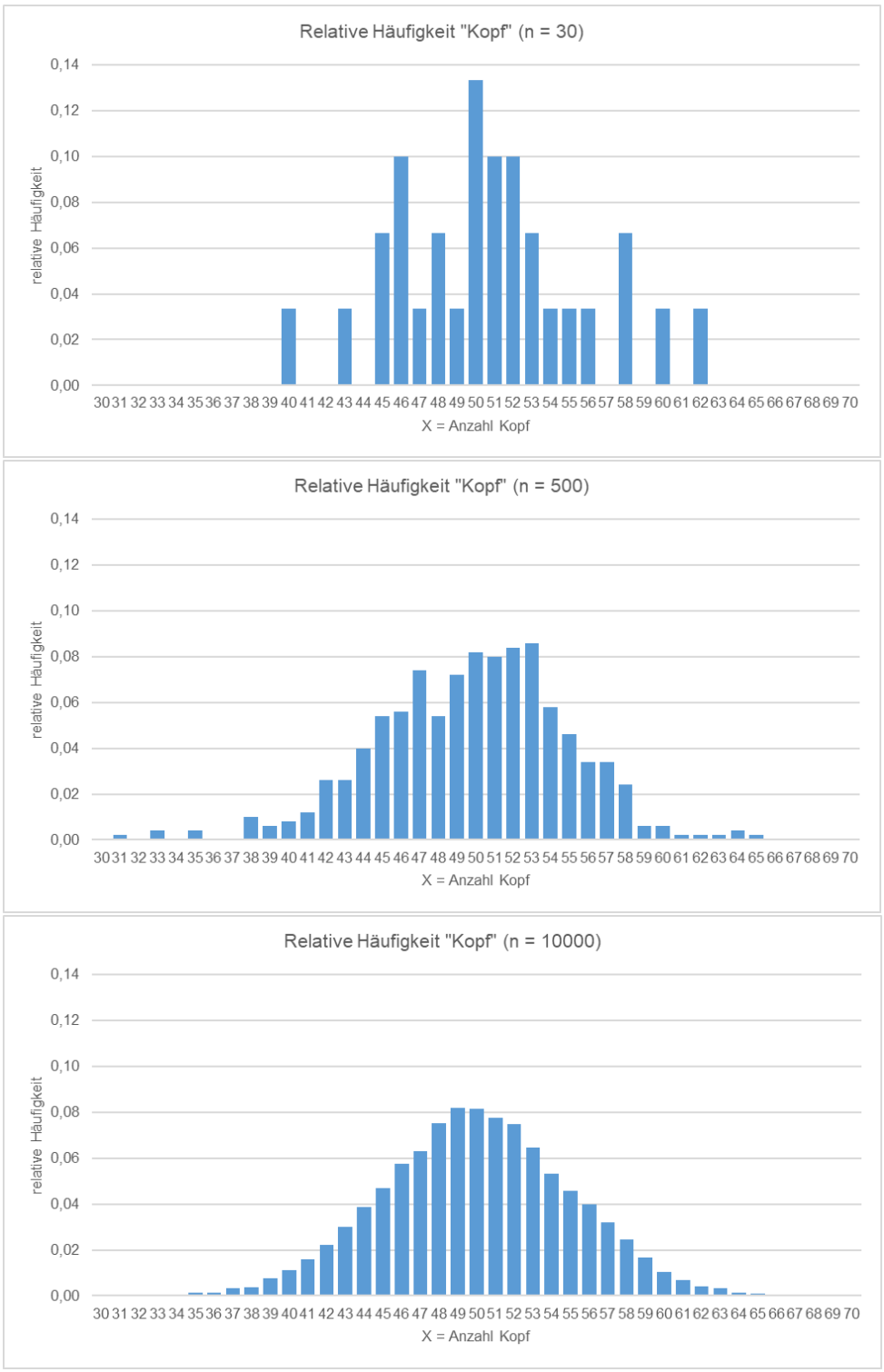


Abb. 5.4: Relative Häufigkeiten zu den Daten aus den Tabellen 5.1 (oben), 5.2 (Mitte) und 5.3 (unten)

Fassen wir noch einmal zusammen: Wir haben 30 Personen 100mal eine Münze werfen lassen; die Zufallsgröße, die wir dabei beobachten, ist die Anzahl, wie oft bei diesen 100 Würfen die Münze auf «Kopf» fällt. Wir erhalten 30 solche Zahlen (von jedem Mitspieler eine Anzahl, wie oft «Kopf» gefallen ist); der Umfang der Stichprobe ist also $n = 30$. Wir können die 30 Stichprobenelemente in einer Häufigkeitstabelle eintragen; eine Klasseneinteilung verwenden wir dabei nicht. Uns interessiert vor allem die relative Häufigkeit, außerdem stellen wir die Häufigkeitsverteilung in einem Säulendiagramm dar. Anschließend haben wir dasselbe Experiment von 500 Personen durchführen lassen und daraus eine neue Stichprobe (mit $n = 500$) erhalten und ausgewertet, und schließlich noch von 10 000 Personen eine Stichprobe mit $n = 10\,000$.

Beim Betrachten der drei Säulendiagramme in Abb.5.4 könnte Folgendes auffallen: Während bei einer Stichprobe aus $n = 30$ Elementen die Verteilung der Anzahl der Fälle, in denen eine 100mal zufällig geworfene Münze auf «Kopf» fiel, relativ willkürlich (und *zufällig*) aussieht, könnte man das Gefühl haben, dass die Diagramme von oben nach unten immer «kompakter» werden und im unteren Fall ($n = 10\,000$) die Verteilung eigentlich schon gar nicht mehr zufällig ausschaut, sondern so, als ob da irgendein Muster dahinter stecken würde. Mathematiker:innen[11] haben versucht, dieses Muster bzw. die dahinterliegende Mathematik zu finden, und zwar für den Fall, der die Häufigkeitsverteilung für den Fall $n = \infty$ beschreibt. Das Ergebnis sieht so aus (Abb. 5.5):

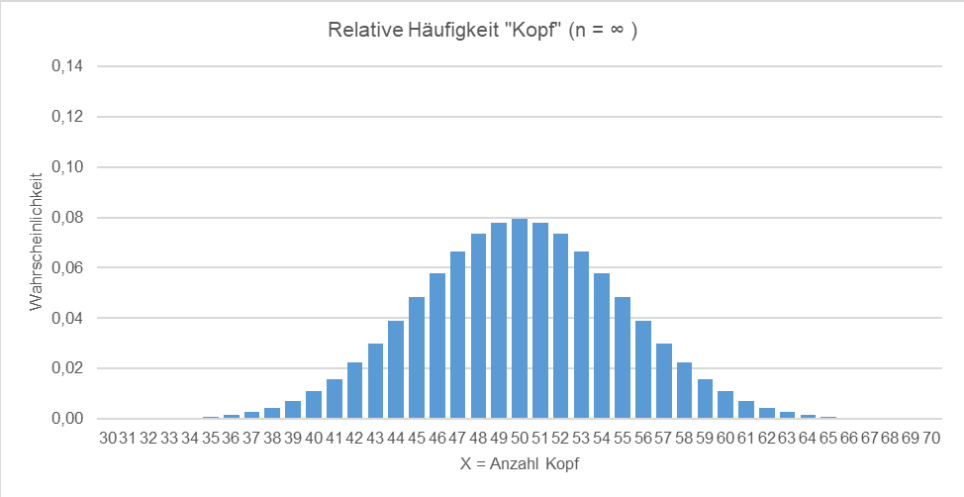


Abb. 5.5: Relative Häufigkeit der Zufallsgröße «Anzahl Kopf» bei je 100 Münzwürfen unendlich vieler Personen

[11] Zum Beispiel der Schweizer Mathematiker und Physiker *Jakob Bernoulli* (1655 - 1705). In seinem Buch *Ars conjectandi* (lat. für «*Die Kunst der Mutmaßung*»), posthum erschienen 1713, hat er den Grundstein für die oben beschriebene Wahrscheinlichkeitsfunktion gelegt.

In Tabelle 5.4 sind einige Zahlenwerte des in Abb.5.5 visualisierten Modells tabellarisch angegeben.

x	$P(X=x)$	x	$P(X=x)$	x	$P(X=x)$	x	$P(X=x)$
30	0.0000	40	0.0108	50	0.0796	60	0.0108
31	0.0001	41	0.0159	51	0.0780	61	0.0071
32	0.0001	42	0.0223	52	0.0735	62	0.0045
33	0.0002	43	0.0301	53	0.0666	63	0.0027
34	0.0005	44	0.0390	54	0.0580	64	0.0016
35	0.0009	45	0.0485	55	0.0485	65	0.0009
36	0.0016	46	0.0580	56	0.0390	66	0.0005
37	0.0027	47	0.0666	57	0.0301	67	0.0002
38	0.0045	48	0.0735	58	0.0223	68	0.0001
39	0.0071	49	0.0780	59	0.0159	69	0.0001

Tabelle 5.4: Wahrscheinlichkeitsverteilung der Zufallsgröße «*Anzahl Kopf*» *bei 100 Münzwürfen. Es handelt sich dabei um die auf* Jacob Bernoulli *(siehe Fußnote 11, S.111) zurückgehende* Binomialverteilung, *auf die wir auf Seite 113 noch zurückkommen werden.*

An dieser Stelle noch ein Geständnis: Wir haben keine 10 000 Menschen gefunden, die für uns für die Daten der Tabelle 5.3
$10\,000 \times 100 = 1\,000\,000$-mal gewürfelt haben – nicht einmal 30. Zum Glück können wir uns heute mit einer Computersimulation weiterhelfen. Das geht für 1 Million Realisierungen einer Zufallsvariable sogar noch in Excel.

Die oben verwendete Verteilung ist nur eines von mehreren möglichen Modellen, das wir verwenden können, um stochastische Phänomene in der Welt zu beschreiben. Schauen wir uns noch einige weitere an:

5.4 Die Modellierung der Verteilung diskreter Zufallsgrößen

Für *diskrete* Zufallsvariable können wir ein Modell angeben, das für jede einzelne mögliche Realisierung die Wahrscheinlichkeit für ihr Auftreten angibt. Dazu ordnen wir jeder Realisierung der Zufallsvariablen x eine Wahrscheinlichkeit $P(x_i)$ zu. Die Funktion, die das leistet, ist die **Wahrscheinlichkeitsfunktion**:

$$f(x) = P(X=x) = \begin{cases} p_i & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases} \tag{5.10}$$

Eine wichtige Frage ist manchmal auch die nach der Wahrscheinlichkeit, dass die Zufallsgröße X kleiner oder gleich einer vorgegebenen Zahl x ist oder zwischen zwei vorgegebenen Werten a und b liegt. Diese Fragen können mit Hilfe der *Verteilungsfunktion* beantwortet werden. Sie ist so definiert:

Der Funktionswert der **Verteilungsfunktion** $F(x)$ an der Stelle x gibt die Wahrscheinlichkeit an, dass X kleiner oder gleich x ist. Im diskreten Fall entspricht das der Summe der Einzelwahrscheinlichkeiten:

$$F(x) = P(X \leq x) = \sum_{i=1}^{k} P(x = x_i) = \sum_{i=1}^{k} f(x_i) \text{ für } x_i \leq x \tag{5.11}$$

Die Verteilungsfunktion ist also das modellhafte Pendant zur kumulierten relativen Häufigkeit bei einer empirischen Häufigkeitsverteilung (vgl. S.33).

Formel 5.10 und 5.11 sind in dieser Form noch sehr abstrakte Definitionen – was konkret aber wird für die einzelnen p_i eingesetzt? Dafür gibt es mehrere Möglichkeiten, von denen wir zwei näher anschauen: Die diskrete *Gleichverteilung* und die *Binomialverteilung*.

Die Binomialverteilung

haben wir bereits in unserem «Münzwurf-Beispiel» (S.108ff.) kennen gelernt. Es handelt sich dabei um das mögliche Modell der Verteilung einer diskreten Zufallsvariable, und zwar für folgenden Fall eines Zufallsexperiments:

- Das Experiment besteht aus N voneinander unabhängigen Versuchen.
- Bei jedem Versuch gibt es nur zwei mögliche Ausgänge. Einen bezeichnen wir – im Sinne des Experiments – als Erfolg, den anderen als Misserfolg
- Die Wahrscheinlichkeit für einen Erfolg beträgt p, für einen Misserfolg $(1 - p)$.
- Diese Wahrscheinlichkeiten p und $(1 - p)$ bleiben bei jedem der N Versuche gleich.

Für unser Beispiel war das: Es gab $N = 100$ Münzwürfe. Bei jedem Wurf gibt es zwei mögliche Ausgänge: Kopf oder Zahl. Die Wahrscheinlichkeit für Kopf beträgt $p = 0.5$ und für Zahl $(1 - p) = 0.5$. Das bleibt bei jedem der 100 Münzwürfe gleich. Die einzelnen Würfe sind auch unabhängig voneinander, d.h. die Münze hat kein «Gedächtnis», das z.B. nach 10 mal Kopf sagt: «Hey, jetzt kommt mal Zahl an die Reihe», sondern die Chancen sind auch nach 10 mal Kopf 50 : 50 für Kopf oder Zahl.

Die Binomialverteilung hat die *Wahrscheinlichkeitsfunktion*

$$f(x) = P(X = x) = \begin{cases} \binom{N}{x} p^x (1-p)^{N-x} & \text{für } x = 0, 1, \ldots, N \\ 0 & \text{sonst} \end{cases} \tag{5.12}$$

und die *Verteilungsfunktion*

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{N}{k} p^k (1-p)^{N-k} \tag{5.13}$$

Wir brauchen diese Formeln nicht auswendig zu können – sie sind hier der Vollständigkeit halber angegeben (damit man z.B. die Werte der Tabelle 5.4 nachvollziehen kann). In unserem Beispiel ist $N = 100$ (weil jeder Teilnehmer 100 mal würfelt), $p = 0.5$ (weil die Wahrscheinlichkeit, dass «Kopf» kommt, 50% beträgt) und für x werden der Reihe nach die x aus der Tabelle 5.4 eingesetzt. Außerdem sehen wir an der Formel, warum die Verteilung «Binomialverteilung» heißt: Der dabei auftretende Koeffizient $\binom{N}{x}$ ist der so genannte *Binomialkoeffizient* (siehe Kap. 1.3 aus MAT101).

In *MS Excel* erhalten wir die Werte der Wahrscheinlichkeitsfunktion der Binomialverteilung mit `=BINOM.VERT(x;N;p;FALSCH)` und die Verteilungsfunktion mit `=BINOM.VERT(x;N;p;WAHR)`.
In *R* lautet der Befehl für die binomiale Wahrscheinlichkeitsfunktion `dbinom(x,N,p)` und `pbinom` für die Verteilungsfunktion.

Werte der Wahrscheinlichkeitsfunktion der Binomialverteilung in *Excel* und *R*

Beispiel 25 *In einer Firma sind täglich zehn Server in Betrieb. Die Wahrscheinlichkeit, dass ein Server ausfällt, beträgt 1%. Durch redundante Serverkomponenten kann «normal» weitergearbeitet werden, solange nicht mehr als zwei Server gleichzeitig ausfallen. Wie hoch ist die Wahrscheinlichkeit für einen ausfallsicheren Betrieb?*

Das Beispiel erfüllt alle Voraussetzungen einer Binomialverteilung: Die «Versuche» entsprechen den $N = 10$ Servern. Für jeden Server gibt es genau zwei Möglichkeiten: Er fällt aus (mit der Wahrscheinlichkeit $p = 0.01$), oder nicht (mit der Wahrscheinlichkeit $(1-p) = 0.99$). Und wir gehen davon aus, dass die Server unabhängig voneinander sind, d.h. auch evtl. Ausfälle unabhängig voneinander sind.

Damit ein ungestörtes Arbeiten möglich ist, dürfen nicht mehr als zwei Server ausfallen, d.h. es könnten 0, 1 oder 2 Server ausfallen.

Wir benötigen zunächst die Werte der Wahrscheinlichkeitsfunktion an den Stellen $0, 1$ *und* 2*, die wir z.B. aus Excel erhalten:* `BINOM.VERT(0;10;0,01;FALSCH)` *ergibt* 90.44%*, für* $x = 1$ *erhalten wir* 9.14% *und für* $x = 2$*:* 0.42%*. In Summe sind das*

$$F(2) = P(X \leq 2) = 0.9044 + 0.0914 + 0.0042 = 0.9999$$

(Wir hätten mit `BINOM.VERT(2;10;0,01;WAHR)` *auch direkt den Wert der Verteilungsfunktion an der Stelle 2 rechnen können und wären auf denselben Wert gekommen).*

D.h. mit einer Wahrscheinlichkeit von 99.99% *fallen nicht mehr als 2 Server aus und es kann in der Firma ungestört gearbeitet werden – also zumindest was die Funktionalität der Server betrifft...*

Aufgabe 19 Ein Statistiktest besteht aus 6 Single-Choice Fragen mit jeweils 4 Antwortmöglichkeiten. (Single-Choice = genau eine Antwort aus den 4 möglichen ist richtig). Für jemanden, der sich nicht auf den Test vorbereitet hat und nach Belieben zufällige Antworten ankreuzt, beträgt die Erfolgswahrscheinlichkeit pro Frage $p = 25\%$. Wie groß ist die Wahrscheinlichkeit, dass diese Person den Test positiv besteht, d.h. mindestens 3 Fragen richtig beantwortet?

Die diskrete Gleichverteilung

Eine andere, einfache Verteilung, die eine Zufallsgröße haben kann, ist die diskrete **Gleichverteilung**. Sie ordnet allen innerhalb des Intervalls $[a, b]$ liegenden Werten einer Zufallsgröße die gleiche Wahrscheinlichkeit zu.

Die *Wahrscheinlichkeitsfunktion* der diskreten Gleichverteilung lautet:

$$f(x) = \begin{cases} \frac{1}{k} & \text{für alle } x_i \text{ mit } i = 1, \ldots, k \\ 0 & \text{sonst} \end{cases} \tag{5.14}$$

Beispiel 26 *Beim Würfeln mit einem «unverfälschten» Würfel gibt es sechs mögliche Ausgänge: Man kann* $1, 2, 3, 4, 5$ *oder* 6 *würfeln. Im Sinne der Formel (5.14) können wir also schreiben:* $k = 6$ *und* $x_1 = 1$*,* $x_2 = 2$*,* $x_3 = 3$*,* $x_4 = 4$*,* $x_5 = 5$*,* $x_6 = 6$*.*

Jetzt können wir den Wert der Wahrscheinlichkeitsfunktion zum Beispiel an der Stelle $x = 2$ *ausrechnen:*

$$f(2) = \frac{1}{6} = 0.167$$

oder an der Stelle 7:

$$f(7) = 0$$

Das stimmt auch mit unserer bisherigen Vorstellung überein: Die Wahrscheinlichkeit, einen 2er zu würfeln, beträgt 16.7%*, die Wahrscheinlichkeit, einen 7er zu würfeln ist hingegen gleich Null. (Jetzt wissen wir auch, warum Mathematiker in den Formeln 5.10 und 5.14 die «sonst»-Zeile eingefügt haben).*

Aufgabe 20 Auf Seite 52 haben wir den Begriff Modalwert kennengelernt und ihn u.a. auch als als *wahrscheinlichsten Wert* bezeichnet. Was ist der wahrscheinlichste Wert der in Abb.5.5 dargestellten Verteilung?

Aufgabe 21 Hat eine Gleichverteilung auch einen Modalwert?

5.5 Die Modellierung der Verteilung stetiger Zufallsgrößen

Bei stetigen Zufallsgrößen müssen wir Folgendes beachten: Die Anzahl aller möglichen Realisierungen einer stetigen Zufallsvariable ist nicht abzählbar sondern unendlich groß – so wurden ja stetige Variable auf Seite 13 definiert. Wenn wir aber in Formel 5.1 im Nenner ∞ einsetzen, erhalten wir: $P(E) = 0$. Das bedeutet, dass wir im stetigen Fall einem bestimmten x keine Wahrscheinlichkeit $P(X = x)$ zuordnen können! Es geht aber für ein Intervall $[a, b]$.

Beginnen wir mit der **Verteilungsfunktion** $F(x)$. Formal sieht sie so ähnlich aus wie 5.11, allerdings wird die Summe durch ein Integral ersetzt:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{5.15}$$

Die in Formel 5.15 auftretende Funktion $f(t)$ nennen wir **Dichtefunktion** der Verteilung (auch: *Wahrscheinlichkeitsdichte* bzw. nur *Dichte*, im Englischen: *probability density function*, abgekürzt PDF). Sie ist so was ähnliches wie die Wahrscheinlichkeitsfunktion im diskreten Fall. Mit dem Unterschied, dass wir – wie bereits oben begründet – nicht die Wahrscheinlichkeit für ein bestimmtes x angeben können, sondern nur für ein Intervall. Abb. 5.6 zeigt ein Beispiel für eine Dichtefunktion und den Zusammenhang zur Verteilungsfunktion $F(x)$. Wir sehen dabei unter anderem: Die Wahrscheinlichkeit $P(a < X \leq b)$ entspricht der

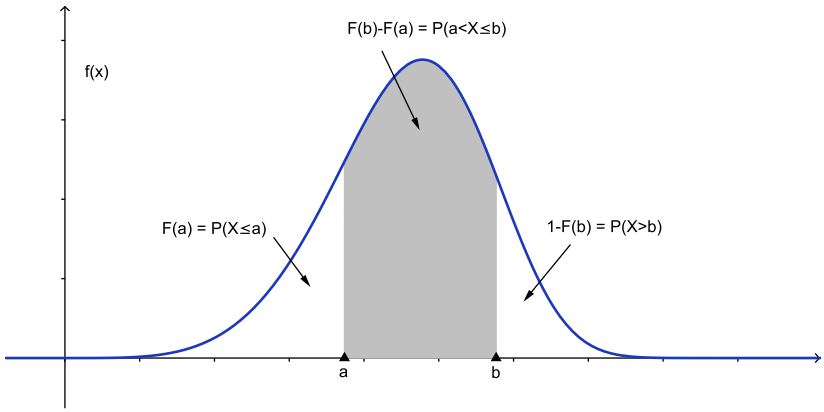


Abb. 5.6: Eine mögliche Dichtefunktion $f(x)$ einer stetigen Zufallsgröße und der Zusammenhang zur Verteilungsfunktion $F(x)$

Fläche zwischen der x-Achse und der Dichtefunktion $f(x)$ von a bis einschließlich b.

Wie aus Abb.5.6 ersichtlich, gelten für die Verteilungsfunktion folgende wichtige Zusammenhänge:

$$P(X \leq a) = F(a) = \int_{-\infty}^{a} f(x)dx \tag{5.16}$$

$$P(X > b) = 1 - F(b) = \int_{b}^{+\infty} f(x)dx \tag{5.17}$$

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx \tag{5.18}$$

Es gibt eine große Anzahl von Modellen für die Verteilung stetiger Zufallsgrößen. Abb.5.7 zeigt ein paar Beispiele für häufig vorkommende Dichtefunktionen. Konkret werden wir uns zwei näher ansehen: Die stetige Gleichverteilung und die Normalverteilung.

Die stetige Gleichverteilung

ist ähnlich definiert wie die diskrete Gleichverteilung: Sie ordnet allen gleichgroßen Intervallen aus dem Bereich von a bis b die gleiche Wahrscheinlichkeit $\neq 0$ zu, außerhalb dieses Bereichs ist die Wahrscheinlichkeit gleich Null. Der

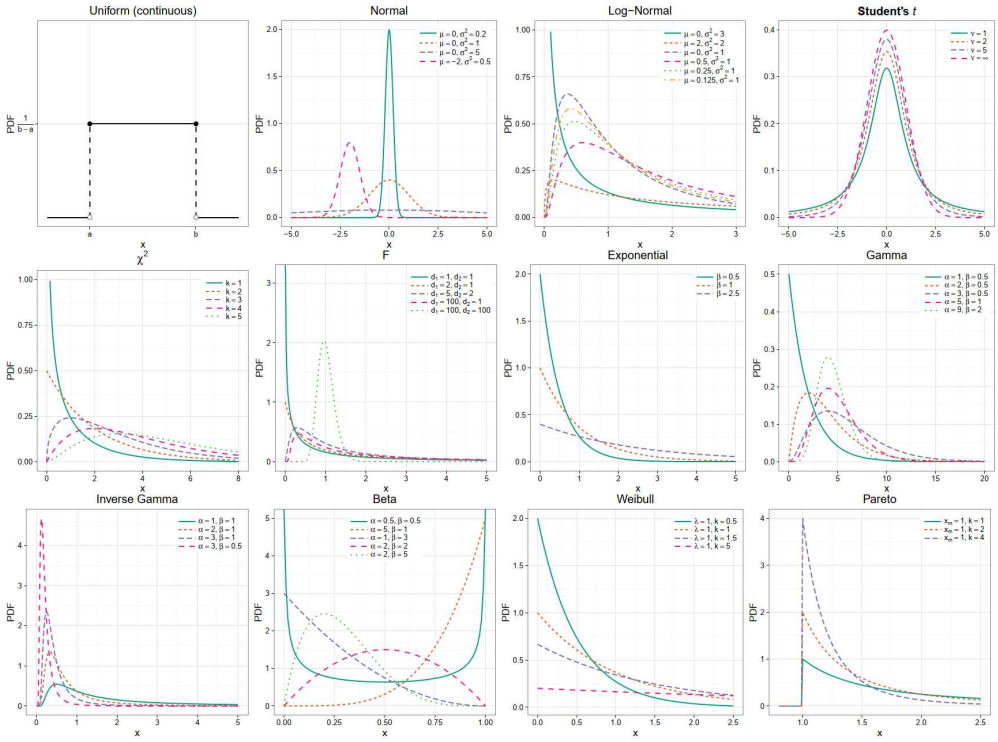


Abb. 5.7: Darstellung der Dichtefunktionen einiger stetiger Wahrscheinlichkeitverteilungen. (Quelle: Vallentin, Matthias. 2017. The Probability and Statistics Cookbook. https://github.com/mavam/stat-cookbook. CC BY-NC-SA 4.0)

Unterschied zur diskreten Gleichverteilung ist der, dass im Bereich $[a, b]$ nicht eine abzählbare Anzahl von Werten liegt, sondern unendlich viele Intervalle mit unendlich vielen Zahlen aus $\mathbb{R}$.

Die stetige Gleichverteilung hat die *Dichtefunktion*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases} \tag{5.19}$$

Die *Verteilungsfunktion* der stetigen Gleichverteilung im Bereich $a \leq x \leq b$ ist gegeben durch:

$$P(X \leq x) = F(x) = \frac{x-a}{b-a} \tag{5.20}$$

Der Graph der Dichtefunktion hat ein rechteckiges Aussehen. Dichte- und Verteilungsfunktion sind in Abb.5.8 dargestellt und wir sehen darin auch: Wenn $x < a$ oder $x > b$, dann ist $F(x) = 0$.

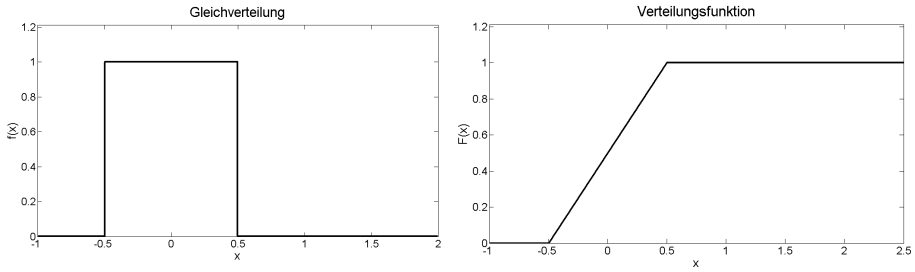


Abb. 5.8: Dichtefunktion und Verteilungsfunktion der stetigen Gleichverteilung in den Grenzen $-0.5 \leq x \leq 0.5$.

Beispiel 27 *Wie groß ist die Wahrscheinlichkeit, dass eine gleichverteilte stetige Zufallsgröße zwischen den Werten x_1 und x_2 liegt?*

Nach Formel 5.18 können wir angeben:

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$$

und hier jetzt die konkreten Funktion aus Formel 5.20 einsetzen:

$$P(x_1 \leq X \leq x_2) = \frac{x_2 - a}{b - a} - \frac{x_1 - a}{b - a} = \frac{x_2 - a - x_1 + a}{b - a} = \underline{\underline{\frac{x_2 - x_1}{b - a}}}$$

Aufgabe 22 Am Bahnhof Meidling wird folgende Information durchgesagt: «Der Zug Richtung Wiener Neustadt hat zwischen 5 und 15 Minuten Verspätung». Angenommen es gibt keinen Grund zur Annahme, dass die Verspätung eher bei 5 oder 15 Minuten liegt, sondern der Zug tatsächlich «irgendwann» in diesem Intervall eintreffen wird: Wie groß ist die Wahrscheinlichkeit, dass die Verspätung maximal 7 Minuten ausmacht?

Die Normalverteilung

Eine für uns sehr wichtige Verteilung von Daten ist die **Normalverteilung**, auch *Gaußsche Verteilung* genannt[12]. Viele Sachverhalte aus den Natur- und Sozialwissenschaften sind annähernd normalverteilt (oder zumindest modellieren wir sie so). In den modernen Wirtschaftswissenschaften folgen die Phänomene zwar manchmal anderen Mustern, es lässt sich aber zeigen, dass immer dann, wenn wir eine entsprechend große Anzahl zufälliger Ereignisse mit einer beliebigen Verteilung zu einer einzigen Zufallsvariable zusammenfassen können, diese (zumindest annähernd) normalverteilt ist[13].

[12] *Johann Friedrich Carl Gauß*, deutscher Mathematiker und Geodät, 1777-1855

[13] Das besagt der so genannte *Zentrale Grenzwertsatz*, auf den wir hier aber nicht näher eingehen.

Die mathematische Funktion der Normalverteilung ist ziemlich kompliziert und für uns nicht weiter wichtig. Wir werden einfach ein Computerprogramm verwenden, wenn wir tatsächlich einen Wert ausrechnen müssen. Wir merken uns nur, dass die Normalverteilung über zwei Parameter definiert ist: den theoretischen Mittelwert μ (der **Erwartungswert** genannt wird) und die theoretische Varianz σ^2. Wie wir zu diesen beiden Parametern kommen, werden wir uns im nächsten Kapitel genauer ansehen. Einstweilen setzen wir, wenn uns die theoretischen Modellparameter nicht bekannt sind, für den Erwartungswert das arithmetische Mittel ein und für die theoretische Varianz den Wert der empirischen Varianz.

Zwei Datensätze, die wir untersuchen, haben üblicherweise – auch wenn sie beide normalverteilt sind – unterschiedliche Erwartungswerte und unterschiedliche Varianzen, und somit eine andere Form der Normalverteilung. Letztlich gibt es unzählig viele Normalverteilungen; einige sind beispielhaft in Abb.5.9 dargestellt. Wenn wir unsere normalverteilten Daten wie auf Seite 74 angegeben *standardisieren* (zu *z-Werten* mit Mittelwert 0 und Standardabweichung 1), so sind diese **standardnormalverteilt**, d.h. verteilt nach einer Normalverteilung mit Erwartungswert $\mu = 0$ und $\sigma^2 = 1$.

Grafisch ähnelt die Dichtefunktion der Normalverteilung der Form einer «Glocke» und wird daher auch *Glockenkurve* (auch: *Gaußsche Glockenkurve*) genannt (Abb.5.9).

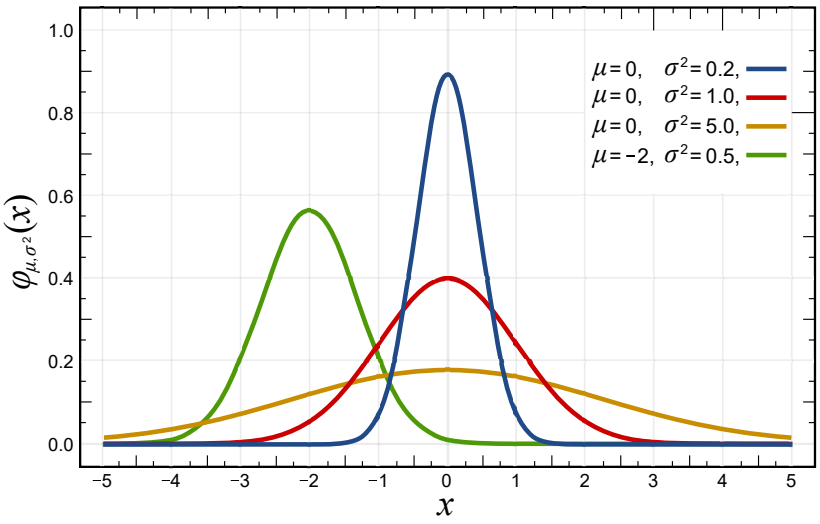


Abb. 5.9: Gaußsche Glockenkurven: Verschiedene Dichtefunktionen zur Normalverteilung mit unterschiedlichen Erwartungswerten und Varianzen. Der Scheitel der jeweiligen Kurven liegt bei $x = \mu$, ihre Wendepunkte im Abstand $\pm\sigma$ vom Scheitelwert. Die Kurve ist umso höher und steiler, je kleiner σ ist. (Quelle: https://commons.wikimedia.org/w/index.php?curid=3817954)

Wie wir aus der Abbildung sehen können, hat die Normalverteilung folgende Eigenschaften:

Der Parameter μ bestimmt das Zentrum der Kurve, σ ihre «Schlankheit» (Breite). Der Scheitel der Dichtefunktion – also das Maximum – liegt bei $x = \mu$. Die Wendepunkte (das ist der mathematische Ausdruck für Punkte, wo die Kurve am «steilsten» ist) liegen im Abstand $\pm\sigma$ von μ. Je kleiner σ ist, desto schmäler ist die Glockenkurve, je größer σ ist, desto breiter und flacher wird sie.

Die beiden Enden der Kurve kommen zwar sehr nahe an die x-Achse heran, berühren sie aber nie. (Mathematisch heißt das: Sie nähern sich *asymptotisch* an die *Abszisse* an). Und obwohl das so ist, hat die Fläche unter der Kurve überraschenderweise einen endlichen Wert, nämlich genau 1. Das gilt für alle einzelnen unterschiedlichen Kurven der Abb.5.9: Die Gesamtfläche unter der orangen, grünen, blauen und magenta-farbenen Kurve ist jeweils genau 1 Einheit groß.

Weiters ist erkennbar, dass die Normalverteilung eine um μ symmetrische Verteilung ist, d.h. betragsmäßig gleich große positive oder negative Abweichungen vom Erwartungswert sind gleich wahrscheinlich. Daher ist der Erwartungswert auch gleichzeitig der Median der Verteilung. Die Wahrscheinlichkeit für eine Abweichung vom Erwartungswert ist umso geringer, je größer diese Abweichung ist. Große Abweichungen sind also weniger wahrscheinlich als kleine.

Und: Für eine normalverteilte Zufallsgröße gilt:

- ca. 68% aller Daten liegen im Intervall $\mu \pm 1 \cdot \sigma$, also innerhalb eines Bereichs im Abstand von maximal einer Standardabweichung um den Erwartungswert.
- ca. 95% aller Daten liegen im Intervall $\mu \pm 2 \cdot \sigma$, also innerhalb eines Bereichs im Abstand von maximal zwei Standardabweichungen um den Erwartungswert.
- ca. 99.7% aller Daten liegen im Intervall $\mu \pm 3 \cdot \sigma$, also innerhalb eines Bereichs im Abstand von maximal drei Standardabweichungen um den Erwartungswert.

Beispiel 28 *Angenommen im Studiengang WIBA beträgt das (langjährig beobachtete) Durchschnittsalter der Studierenden $\bar{x} = 37$ Jahre mit einer Standardabweichung von $s = 5$ Jahren. Wie groß ist die Wahrscheinlichkeit, dass unter der Annahme einer Normalverteilung ein:e (beliebig ausgewählte:r) Studierende:r nicht älter als 40 Jahre alt ist?*

Allgemein finden wir die Antwort in Formel 5.16. Konkret suchen wir den Wert der Verteilungsfunktion an der Stelle $P(X \leq 40) = F(40)$ für eine Normalverteilung mit

In *MS Excel* können wir für eine normalverteilte Zufallsgröße sowohl den Wert der Verteilungs- als auch den der Dichtefunktion ausrechnen:
Den Wert der *Dichtefunktion* einer normalverteilten Zufallsgröße an der Stelle x erhalten wir aus
`=NORM.VERT(x;Erwartungswert;Standabweichung;FALSCH)`
Den Wert der *Verteilungsfunktion* einer normalverteilten Zufallsgröße an der Stelle x erhalten wir aus
`=NORM.VERT(x;Erwartungswert;Standabweichung;WAHR)`

In *R* erhalten wir den Wert der Verteilungsfunktion der Normalverteilung mit `pnorm`, die Dichtefunktion mit `dnorm`.

dem Erwartungswert 37 *und der Standardabweichung* 5*. Laut EXCEL beträgt dieser Wert* 0.7257 *(*`=NORM.VERT(40;37;5;WAHR)`*).*

Die Wahrscheinlichkeit, dass ein Studierender nicht älter als 40 *Jahre ist, beträgt demnach* 72.6%

Aufgabe 23 Angenommen im Studiengang WIBA beträgt das (langjährig beobachtete) Durchschnittsalter der Studierenden $\bar{x} = 37$ Jahre mit einer Standardabweichung von $s = 5$ Jahren. Wie groß ist die Wahrscheinlichkeit, dass unter der Annahme einer Normalverteilung ein:e (beliebig ausgewählte:r) Studierende:r zwischen 32 und 42 Jahre alt ist?

In *R* können wir einen Vektor von Zufallszahlen erzeugen, die einer bestimmten Verteilung folgen. Zum Beispiel erhalten wir 100 *gleichverteilte* Zufallszahlen im Intervall (a, b) mit der Funktion `runif(100,a,b)`. *Binomialverteilte* Zufallszahlen erhalten wir mit dem Befehl `rbinom`, normalverteilte mit `rnorm`. Die genaue Syntax dieser Befehle ist in der R-Hilfe beschrieben (aufrufbar mit dem Befehl `help()`, also z.B. `help(rnorm)`).

Um überhaupt Hilfe über alle in *R* implementierten Verteilungen zu erhalten, gibt man den Befehl `help(Distributions)` ein.

Zusammenfassung

Der eine oder die andere mag sich vielleicht fragen, wozu wir in der Statistik die Wahrscheinlichkeitsrechnung und diese manchmal ziemlich kompliziert anmutenden theoretischen Verteilungsfunktionen brauchen[14]. Dazu erinnern wir uns an das auf Seite 6 Gesagte über das Ziel, das wir mit der Anwendung der Statistik erreichen wollen: Wir wollen ***Modelle*** *finden, mit denen wir besser verstehen können, wie oder warum bestimmte Phänomene in der realen Welt funktionieren*. Die Wahrscheinlichkeitsverteilungen dieses Kapitels liefern uns für diese Modellbildung eine gute Basis. Wir versuchen so gut wie möglich, auf den ersten Blick «regellos» verteilten empirisch gewonnenen Daten eine Struktur zu geben, am besten eine mathematisch formalisierte Struktur. Nur so können wir letztlich die dahinterliegenden Phänomene erklären und sogar einen Nutzen für zukünftige Ereignisse ziehen, indem wir zum Beispiel Computer dazu bringen, das Verhalten von Menschen, Wirtschaftssystemen, dem Klima, Verkehrsströmen, technischen oder Produktions-Abläufen, oder die Ausbreitung von Virus-Epidemien etc. zu simulieren.

Dafür stehen uns eine Menge von Modellen zur Verfügung. Abb.5.10 zeigt eine Übersicht dazu – oder zumindest das, was Statistiker:innen als «Übersicht» bezeichnen. Wenn dir das ziemlich komplex und/oder chaotisch vorkommt, keine Angst: Wir werden uns in dieser Einführungslehrveranstaltung nicht mit mehr als der Normalverteilung auseinandersetzen. Und mit dem nächsten Kapitel gleich die Tür dorthin öffnen und das Feld der *schließenden Statistik* betreten.

Wer sich zuvor noch ein paar anschauliche Beispiele ansehen möchte, in denen Wahrscheinlichkeiten visualiert werden, kann das unter seeing-theory.brown.edu tun.

[14] An dieser Stelle können wir nicht auf die eingehen, die sich schon seit dem ersten Kapitel fragen, wozu wir das Ganze brauchen…

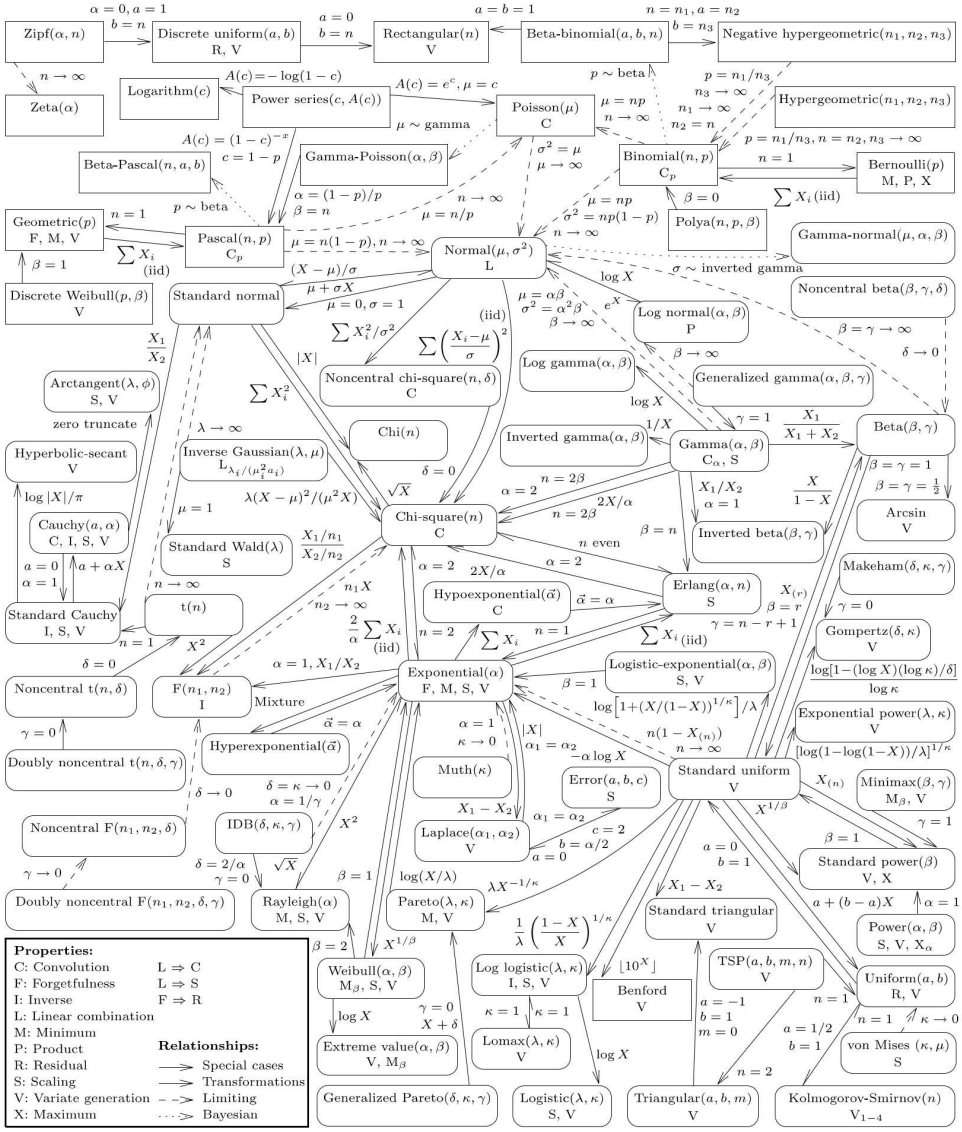


Abb. 5.10: Univariate Wahrscheinlichkeitsverteilungen und wie sie zusammenhängen. (Quelle: Leemis, Lawrence M. and McQueston, Jacquelyn T. 2008. "Univariate Distribution Relationships". *The American Statistician* 62(1) p.47)

6 Ergebnissen vertrauen

6.1 Stichproben und Modelle

Wir haben uns in den bisherigen Überlegungen vornehmlich mit der statistischen Untersuchung von empirisch ermittelten Daten der *Realwelt* (einer *Stichprobe*) beschäftigt und endliche Beobachtungsreihen untersucht. Im letzten Kapitel haben wir auch ein entsprechendes theoretisches *Modell* angeschaut; oft sprechen Statistiker:innen dann von einer *Grundgesamtheit*.

In der Praxis ist es nun so: Die Binomialverteilung beispielsweise ist ein theoretisches Modell, wie die Verteilung einer Grundgesamtheit aussehen könnte. In der Realität werden wir für die empirische Häufigkeitsverteilung bestimmter Zufallsgröße eine Verteilung erhalten, die zwar ungefähr wie eine Binomialverteilung ausschaut, aber eben nur ungefähr – es gibt da und dort kleinere oder größere Abweichungen von der Binomialverteilung. Soweit wir in Abb. 5.4 gesehen haben, können wir zumindest annehmen, dass wir uns an das theoretische Modell umso besser annähern, je größer unsere Stichprobe ist.

In diesem Kapitel fragen wir nun: Inwieweit lassen sich die Ergebnisse einer Stichprobe überhaupt für eine (theoretisch existierende) Grundgesamtheit verallgemeinern und die Ergebnisse aus der Realwelt auf ein Modell übertragen? Wie weit können wir darauf vertrauen, dass die aus den empirischen Daten abgeleiteten Kennwerte auch auf das Modell der Grundgesamtheit zutreffen?

Fragestellungen dieser Art sind Hauptaufgabe der **Induktiven Statistik** (auch: *Schließende* oder *Analytische* Statistik). Zuallererst erinnern wir uns dabei an die wichtigsten Kennwerte einer Stichprobe: Den Mittelwert und die Standardabweichung (siehe Kapitel 3).

Auch für das theoretische Modell, also die Grundgesamtheit, können wir derartige Parameter angeben:

Modell = Grundgesamtheit		**Empirie = Stichprobe**	
Bezeichnung	*Formelzeichen*	*Bezeichnung*	*Formelzeichen*
Erwartungswert	μ	arithmetischer Mittelwert	$\bar{x}$
Standardabweichung	σ	empirische Standardabweichung	s

Tabelle 6.1: Korrespondierende Parameter und Bezeichnungen der (theoretischen) Grundgesamtheit und der (empirischen) Stichprobe

Die Werte in der linken Spalte der Tabelle 6.1 können wir – bis auf wenige Ausnahmefälle – empirisch nicht exakt bestimmen, aber wir können den korrespondierenden Wert der Stichprobe (rechte Spalte) dafür benutzen, die theoretischen Werte zumindest zu *schätzen*. Wenn wir zum Beispiel für μ einen Schätzwert suchen, verwenden wir den empirischen Wert $\bar{x}$, den wir aus einer Stichprobe erhalten haben. Im Sinne der Schätztheorie handelt es sich dabei um einen *Punktschätzer*.

Statistisch gesprochen ist der Erwartungswert μ eine *Zufallsvariable* und der empirische Wert $\bar{x}$ eine *Realisierung* dieser Zufallsvariable.

Wir können uns den «Erwartungswert» auch so vorstellen: Wenn wir eine unendlich große (oder zumindest: ziemlich große) Stichprobe hätten, also zum Beispiel ein Experiment unzählige Male durchgeführt haben, dann sagen wir zum Mittelwert, den wir aus dieser großen Datenmenge erhalten: Erwartungswert.

Aus *einer* Stichprobe erhalten wir zunächst *einen* Schätzwert für den Erwartungswert. Wenn wir *mehrere* Stichproben ziehen, erhalten wir *mehrere* Schätzwerte, sprich: mehrere Realisierungen der Zufallsvariable «Erwartungswert». Wenn wir das oft genug machen (mindestens 30 mal), dann erhalten wir $N \geq 30$ Schätzwerte für den Erwartungswert – und diese Schätzwerte sind normalverteilt[1].

[1] Was wieder mit dem bereits auf Seite 119 erwähnten *Zentralem Grenzwertsatz* zusammenhängt.

Die Weisheit der Vielen
1906 fand im Rahmen der «West of England Fat Stock and Poultry Exhibition», einer regionalen Nutztiermesse in Plymouth, ein Schätzbewerb statt, bei dem das Gewicht eines Ochsen zu raten war. Unter den etwa 800 Personen, die sich daran beteiligten, waren einige Fachleute wie Landwirte und Fleischhauer, aber auch eine große Anzahl von Laien, die einfach ihr Glück im Schätzen versuchen wollten. *Francis Galton* (siehe Fußnote 6, S.82) wertete im Anschluss an den Wettbewerb alle abgegebenen Schätzungen statistisch aus. Eigentlich wollte er zeigen, dass wegen der großen Anzahl von Nicht-Fachleuten das durchschnittliche Ergebnis weit jenseits einer brauchbaren Zahl liegt und man sich daher bei seinen Entscheidungen nicht auf das Urteil von nur durchschnittlich (oder gar unterdurchschnittlich) gebildeten Menschen verlassen sollte. Aber das Gegenteil war der Fall: Das Mittel aus den 787 Schätzwerten betrug mit 542.95 kg beinahe punktgenau das tatsächliche Gewicht von 543.3 kg.

Fundquelle: James Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Schwarmintelligenz anno 1906

6.2 Vertrauensintervalle

Die Abweichung zwischen der Schätzung für einen Parameter und dem wahren Wert dieses Parameters, zum Beispiel die Differenz zwischen Mittelwert der Stichprobe und Erwartungswert der Grundgesamtheit, kann klein oder groß sein. Insbesondere bei einem kleinen Stichprobenumfang wird die Unsicherheit gefühlsmäßig eher größer sein als bei einem großen Stichprobenumfang. Es ist zwar nicht gerade unmöglich aber doch ziemlich unwahrscheinlich, dass wir gleich bei der ersten Stichprobe ins Schwarze treffen.

Um besser auf die Ungenauigkeit der Schätzung eines Parameters des zugrunde liegenden Modells einzugehen, gehen wir so vor:

Wir berechnen aus der Stichprobe nicht nur einen einzigen Wert, z.B. den Mittelwert, sondern wir geben gleich auch ein Intervall rundherum an. Und zwar so, dass das Intervall mit einer Wahrscheinlichkeit von, sagen wir 95%, auch den – uns unbekannten und gar nicht zugänglichen – Erwartungswert der Grundgesamtheit überdeckt. Oder anders ausgedrückt: Wir sind uns zu 95% sicher, dass der Erwartungswert in diesem Intervall enthalten ist.

Zugegebenermaßen ist das so ähnlich wie wenn wir beim Bogenschießen einen

Pfeil auf eine Styroporplatte mit einer «unsichtbaren» Zielscheibe abschießen. Und erst wenn er feststeckt, zeichnen wir die Zielscheibe rundherum. Ausgangspunkt ist der Punkt, wo wir den Pfeil hingeschossen haben. Es ist uns klar, dass wir vermutlich nicht genau ins Zentrum der unsichtbaren Zielscheibe getroffen haben. Daher zeichnen wir einen etwas großzügigeren Kreis, damit wir uns halbwegs sicher sein können, dass das Zentrum der wahren (aber eben unsichtbaren) Zielscheibe auch enthalten ist.

Im Bogensport ist diese Vorgangsweise zwar unüblich, tatsächlich ist es aber eine sehr clevere Methode, aus empirischen Daten auf die Parameter eines unbekannten Modells zu schätzen.

Das gefundene Intervall nennen wir **Konfidenzintervall** (auch: *Vertrauensintervall*)[2]; die Wahrscheinlichkeit, dass wir mit unserer Schätzung richtig liegen, also dass das Intervall, das wir aus der Stichprobe geschätzt haben, den gesuchten Parameter der Grundgesamtheit tatsächlich beinhaltet, ist das **Konfidenzniveau** (auch: *Vertrauenswahrscheinlichkeit*, *statistische Sicherheit* oder *Überdeckungswahrscheinlichkeit* genannt). Das Intervall kann den gesuchten Parameter überdecken oder auch nicht, d.h. wir könnten uns auch irren. Die Wahrscheinlichkeit, dass wir uns irren, wird **Irrtumswahrscheinlichkeit** (auch: *Fehlerwahrscheinlichkeit* oder *Signifikanzniveau*) genannt.

Die Irrtumswahrscheinlichkeit bezeichnen wir üblicherweise mit dem Formelzeichen α und das Konfidenzniveau mit γ, wobei die beiden in Summe immer 1 (bzw. 100%) ergeben müssen[3]: $\gamma = (1 - \alpha)$.

Formal können wir dann ein Konfidenzintervall so ausdrücken:

$$P\,(L \leq \text{Modellparameter} \leq U) = \gamma = (1 - \alpha)$$

mit L . . . *untere* (für «*Lower*») und U . . . *obere Intervallgrenze* («*Upper*»).

Je größer α ist, desto kleiner wird das Konfidenzintervall sein und umgekehrt. Das bringt uns ein bisschen in eine verzwickte Situation: Entweder können wir eine präzise Aussage machen (*Morgen hat es zwischen 10°C und 13°C*), die jedoch höchst unsicher ist, oder eine unscharfe Aussage (*Morgen ist die Temperatur zwischen -10° und +30°*), die sehr zuverlässig eintrifft, aber nicht gerade eine tolle Information enthält.

In der Praxis wählen wir in sozial- und wirtschaftswissenschaftlichen Fragen für α meist 5% oder 10%. (Bei medizinischen oder ingenieurtechnischen Aufgaben

[2] vom lat. *confidere* = vertrauen

[3] Entweder liegen wir richtig, oder nicht. In Summe müssen sich Irrtums- und Vertrauenswahrscheinlichkeit daher auf 1 (bzw. 100%) ergänzen.

wird die Irrtumswahrscheinlichkeit in der Regel kleiner angesetzt, z.B. mit $\alpha = 1\%$ oder 0.1%).

Wenn wir jetzt zum Beispiel aus einer Stichprobe ein Konfidenzintervall für den Erwartungswert μ der zugehörigen Grundgesamtheit schätzen wollen, und wir wählen für $\alpha = 10\%$ bzw. $\gamma = 90\%$, so bedeutet dies: Wir fühlen uns zu 90% «sicher», dass das Intervall, das wir aus der Stichprobe geschätzt haben, den Erwartungswert der Grundgesamtheit enthält. Oder anders ausgedrückt: Wenn wir aus 100 Stichproben jeweils die Konfidenzintervalle bestimmen, ist in 90 derartigen Intervallen damit zu rechnen, dass der Erwartungswert enthalten ist, in 10 Fällen nicht (Leider wissen wir nicht, in welchen 10 Fällen dies eintrifft). $\alpha = 10\%$ ist somit die Angabe eines zehnprozentigen Risikos, dass man bei der Angabe des Konfidenzintervalls eine falsche Aussage tätigt.

Wir können dies auch grafisch veranschaulichen: Nehmen wir an, wir wollen den Erwartungswert μ einer normalverteilten Zufallsgröße bestimmen. Für unser Gedankenbeispiel gehen wir davon aus, dass μ gleich 50 sei (was wir aber in Wirklichkeit nicht wissen – wir wollen μ ja erst bestimmen). Jetzt ziehen wir mehrere Stichproben und bestimmen deren jeweilige Mittelwerte. Jeder dieser Mittelwerte ist dann ein Schätzwert für μ.

In Abb.6.1 sehen wir zehn verschiedene Mittelwerte $\bar{x}_i$, die wir aus den zehn verschiedenen Stichproben erhalten haben. Dazu ist auch jeweils ein Intervall angegeben, in dem unserer Schätzung nach der tatsächliche Wert für μ enthalten sein sollte. Die Intervalle liegen an verschiedenen Stellen (je nachdem, wo der jeweilige Stichprobenmittelwert $\bar{x}_i$ liegt) und sie sind auch unterschiedlich groß (was u.a. mit der Standardabweichung der jeweiligen Stichprobe zusammenhängt). Neun Intervalle liegen nun so, dass μ tatsächlich vom jeweiligen Intervall überdeckt wird. Bei $\bar{x}_5$ hingegen ist μ nicht im Konfidenzintervall enthalten.

Im konkreten Beispiel beträgt also die Wahrscheinlichkeit dafür, ein Intervall wie jenes um $\bar{x}_5$ zu schätzen, $\alpha = 10\%$ (einer von insgesamt zehn Fällen). Die Gegenwahrscheinlichkeit ist $(1 - \alpha) = \gamma = 90\%$. D.h. Die Überdeckungswahrscheinlichkeit beträgt 90%, die Irrtumswahrscheinlichkeit 10%.

Wie können wir jetzt konkrete Grenzen für das Konfidenzintervall angeben? Hier einige Beispiele:

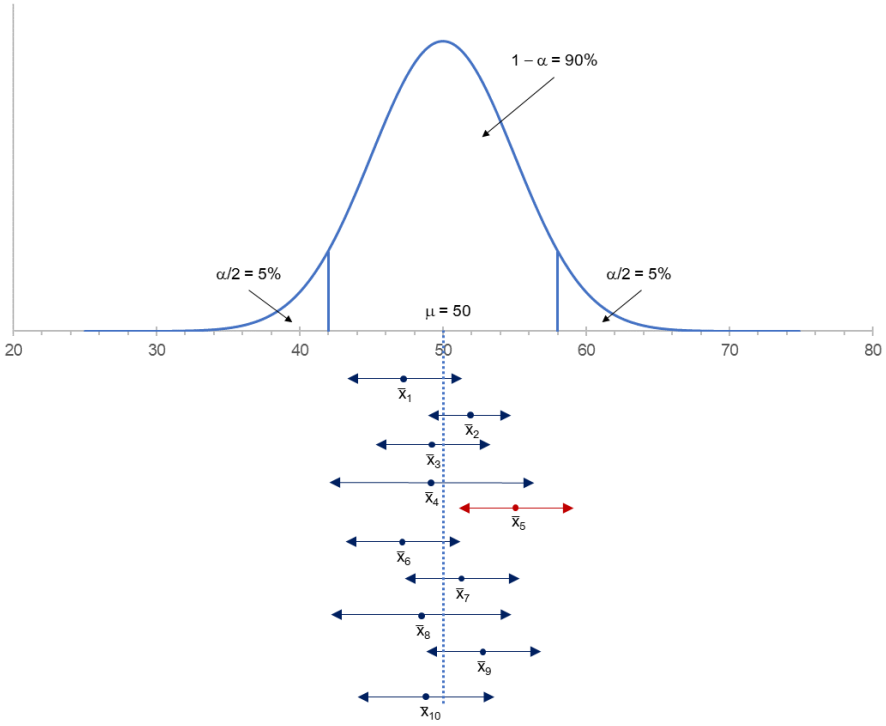


Abb. 6.1: Empirisch erhaltene Konfidenzintervalle einer normalverteilten Zufallsgröße. Deren Erwartungswert beträgt $\mu = 50$ (mit $\sigma = 5$). In einem von 10 Fällen irren wir uns mit dem Konfidenzintervall, aber in 90% liegen wir richtig und μ liegt in dem aus der Stichprobe erhaltenen Intervall. Die Irrtumswahrscheinlichkeit $\alpha = 10\%$ liegt symmetrisch zu jeweils 5% am unteren und oberen Ende der Verteilung.

Schätzung des Konfidenzintervalls für den Erwartungswert μ einer normalverteilten Zufallsgröße

Gegeben sei die Stichprobe einer normalverteilter Zufallsvariablen X. Der Erwartungswert der zugrundeliegenden Grundgesamtheit sei unbekannt und gesucht. Wir bestimmen zunächst einen Schätzwert für den Erwartungswert μ, indem wir uns den arithmetischen Mittelwert $\bar{x}$ ausrechnen. Anschließend berechnen wir auch die empirische Standardabweichung s.

Das Konfidenzintervall für μ ist dann gegeben durch seine untere Grenze L und die obere Grenze U:

$$L = \bar{x} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \tag{6.1}$$

$$U = \bar{x} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \tag{6.2}$$

wobei die Größe $t_{1-\frac{\alpha}{2},n-1}$ das $(1-\frac{\alpha}{2})$-Quantil der so genannten *Studentverteilung* (auch: *t-Verteilung*) ist und aus einem entsprechenden Programm (zum Beispiel *Excel* oder *R*) erhalten werden kann.

Als Eingangsgröße benötigen wir dafür die Anzahl der Elemente in der Stichprobe n sowie die Irrtumswahrscheinlichkeit α. Wählen wir zum Beispiel für $\alpha = 0.05$ und sei $n = 20$, dann beträgt $t = 2.093$.

In *MS Excel* und *LibreOffice Calc* heißt der Befehl zur Berechnung von t `=T.INV.2S(alpha;n-1)`, wobei für `alpha` und `n-1` die jeweiligen Werte für α und n einzusetzen sind.

In *R* lautet der Befehl `qt((1-alpha/2),n-1)`.

Bestimmung der Quantile der T-Verteilung mit *Excel*, *LibreOffice* und *R*.

Der englischer Chemiker und Mathematiker *William Sealey Gosset* (1876-1937) beschäftigte sich um 1900 mit verschiedenen statistischen Analysen und entwickelte die so genannte *Student-* oder *t*-Verteilung.
Die *t*-Verteilung ist – ähnlich der Normalverteilung – über dem Intervall $[-\infty, +\infty]$ definiert, symmetrisch und glockenförmig, hat im Allgemeinen aber eine größere Streuung als die Normalverteilung.
Die *t*-Verteilung ist für uns bei der Bestimmung von Konfidenzintervallen für den Erwartungswert einer Grundgesamtheit wichtig.

Für Fortgeschrittene: Für das Konfidenzintervall benötigen wir mitunter andere Verteilungen als die Normalverteilung.

Das Konfidenzintervall liegt in unserem Fall – wie aus Formeln 6.1 und 6.2 ersichtlich – symmetrisch um den Mittelwert $\bar{x}$ und hat die Länge

$$CI = \frac{2\,t\,s}{\sqrt{n}} \tag{6.3}$$

In *MS Excel* und *LibreOffice Calc* kann mit dem Befehl `=KONFIDENZ.T(alpha;s;n)` (wobei für `alpha` die Irrtumswahrscheinlichkeit, für `s` die Standardabweichung und für `n` die Anzahl der Stichprobenelemente einzugeben sind) die halbe Länge des Konfidenzintervalls ausgerechnet werden, also der Wert $\frac{t\,s}{\sqrt{n}}$. Dieser Wert kann einmal vom Mittelwert abgezogen und einmal dazu addiert werden und wir erhalten nach (6.1) und (6.2) die Grenzen des Konfidenzintervalls.
In *R* können wir den Befehl `t.test(x, conf.level = gamma)` verwenden. Es wird dann gleich mehr berechnet als das Konfidenzintervall (worüber wir uns erst im nächsten Kapitel Gedanken machen), aber unter anderem eben auch das `confidence interval` für die Daten, die im Vektor `x` enthalten sind. Genauer: Je nachdem, was wir für `gamma` eingesetzt haben (z.B. `conf.level = 0.95` oder `conf.level = 0.9`) erhalten wir zum Beispiel das `95 percent confidence interval` oder das `90 percent confidence interval`.
Das 95%-Konfidenzintervall wird dabei als Defaultanwendung gesehen, d.h. wenn wir nur eingeben: `t.test(x)` erhalten wir das 95%-Konfidenzintervall.

Excel-Alternative zur Schätzung des Konfidenzintervalls für den Erwartungswert

Aus Formel 6.3 sehen wir auch: Umso größer der Stichprobenumfang n ist, desto kleiner wird – bei gleichbleibendem α – das Intervall. Kleineres Intervall bedeutet aber: informativere Aussage. Und das bei gleicher Wahrscheinlichkeit.

Es gilt aber auch: Je größer die Standardabweichung s der Stichprobe, desto größer wird das Konfidenzintervall werden.

Für unser Beispiel ($\alpha = 0.05$, $n = 20$, $t = 2.093$) können wir auch leicht rechnen:

$$CI = \frac{2 \cdot 2.093}{\sqrt{20}} \cdot s = 0.936 \cdot s$$

d.h. die Länge des Konfidenzintervalls beträgt 93.6% der Standardabweichung. Verdoppeln wir die Irrtumswahrscheinlichkeit auf $\alpha = 10\%$, dann gilt: $t = 1.729$ und die Länge des Konfidenzintervalls beträgt «nur» mehr 77.3% der Standardabweichung. Wir haben also ein kürzeres Intervall erhalten, aber um den Preis, dass wir uns jetzt in 10 von 100 Fällen irren (vorher war das Intervall größer, aber wir mussten nur in 5 von 100 Fällen mit einem Irrtum rechnen).

Verdoppeln wir hingegen den Stichprobenumfang auf $n = 40$, dann ist t bei

einer Irrtumswahrscheinlichkeit $\alpha = 5\%$ gleich 2.022 und die Länge des Konfidenzintervalls beträgt 64% der Standardabweichung der Stichprobe.

Beispiel 29 *Betrachten wir aus Tabelle 4.1 ausschließlich die Spalte «Größe» und nehmen an, dass diese 20 zufällig ausgewählten Studierenden die übergeordnete Grundgesamtheit «Alle WIBA-Studierenden der FERNFH» repräsentieren.*

Aus den 20 Werten der Stichprobe können wir nach Formel 3.3 einen Mittelwert und nach 3.21 bzw. 3.22 die empirische Standardabweichung ausrechnen. Sie betragen:

$$\bar{x} = 179.5\,cm \quad s = 8.45\,cm$$

Wählen wir für die Irrtumswahrscheinlichkeit $\alpha = 0.05$, dann hat – wie oben angegeben – das entsprechende Quantil der t-Verteilung (für $n = 20$) den Wert $t = 2.093$.

Damit können wir nun die beiden Grenzen des Konfidenzintervalls angeben:

$$L = 179.5 - 2.093 \cdot \frac{8.45}{\sqrt{20}} = 175.5\,cm$$
$$U = 179.5 + 2.093 \cdot \frac{8.45}{\sqrt{20}} = 183.5\,cm$$

D.h. wir nehmen an, dass der tatsächliche Mittelwert aller WIBA-Studierenden irgendwo im Intervall zwischen 175.5 und 183.5 cm liegt.

Mathematisch-statistisch formuliert bedeutet das eben erhalten Ergebnis

$$P\,(175.5\,cm \leq \mu \leq 183.5\,cm) = 95\%$$

was wir «mit Worten» so ausdrücken können:

Erhalten wir für den Parameter μ ein 95%-Konfidenzintervall von $[175.5; 183.5]$, so bedeutet das: *Die Wahrscheinlichkeit, dass der wahre Wert von μ im Intervall $[175.5; 183.5]$ enthalten ist, beträgt 95%*. Oder: *Zögen wir 100 Stichproben und bildeten jeweils das Konfidenzintervall, so würden 95 Intervalle μ enthalten und fünf nicht*.

Diese letzte Aussage lässt uns auch umgekehrt schließen: *Wenn wir aus einer Stichprobe für μ ein 95%-Konfidenzintervall von $[175.5; 183.5]$ erhalten, kann der Erwartungswert der Grundgesamtheit, aus der diese Stichprobe stammt, auch außerhalb dieses Intervalls liegen, also beispielsweise 172 oder 196 cm betragen. Die Wahrscheinlichkeit, dass dies passiert, ist zwar relativ klein (nämlich 5%), aber doch möglich.* Etwas genauer können wir auch sagen: *Im Schnitt wird in 2.5% aller Fälle der wahre Erwartungswert unterhalb von 175.5 cm liegen und in 2.5% aller Fälle oberhalb von 183.5 cm*.

Aufgabe 24 Betrachte aus Tabelle 4.1 ausschließlich die Spalte «Gewicht» und nimm an, dass diese 20 zufällig ausgewählten Personen eine Stichprobe der Grundgesamtheit «Alle WIBA-Studierenden der FERNFH» sind. Gib ein Intervall an, das mit 95%-iger Wahrscheinlichkeit den Erwartungswert für das durchschnittliche Gewicht der WIBA-Studierenden beinhaltet.

Aufgabe 25 Gegeben sind für die Jahre 2005 bis 2014 die Anzahl polizeilich angezeigter Fälle in Österreich:

Jahr	*Anzeigen*	*Jahr*	*Anzeigen*
2005	604 229	2010	534 351
2006	588 229	2011	539 970
2007	592 636	2012	547 764
2008	570 952	2013	546 396
2009	589 961	2014	527 692

In welchen Jahren lag die Anzahl der Anzeigen außerhalb des 95%-Konfidenzintervalls (bezogen auf den Durchschnitt der Jahre 2005-2014)?

Schätzung des Konfidenzintervalls für den Erwartungswert μ einer normalverteilten Zufallsgröße bei Vorliegen einer großen Stichprobe

In den Formeln 6.1 und 6.2 kann man bei großen Stichproben, so ab $n > 30$, die t-Verteilung auch durch eine Normalverteilung ersetzen. Das Konfidenzintervall für μ ist dann gegeben durch die Grenzen

$$L = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{6.4}$$

$$U = \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{6.5}$$

Der Koeffizient z ist der entsprechende $(1 - \frac{\alpha}{2})$-Quantilswert der *Normalverteilung*. Als Eingangsgröße benötigen wir nur die Irrtumswahrscheinlichkeit α.

Die Länge des durch (6.4) und (6.5) gegebenen Intervalls ist:

$$CI = \frac{2\,z\,s}{\sqrt{n}} \tag{6.6}$$

Wir können jetzt auch die Intervalllänge CI vorgeben und einen (Mindest-)Wert für den Stichprobenumfang n abschätzen:

$$n > \left(\frac{2\,z}{CI} \cdot \hat{s}\right)^2 \tag{6.7}$$

In *MS Excel* und *LibreOffice Calc* heißt der Befehl zur Berechnung von z `=NORM.S.INV(1-alpha/2)`.

In *R* lautet der Befehl `qnorm(1-alpha/2)`.

In *MS Excel* und *LibreOffice Calc* kann die halbe Länge des Konfidenzintervalls auch direkt ausgerechnet werden. Der Befehl dazu lautet `=KONFIDENZ.NORM(alpha;s;n)`. Diesen Wert muss man einmal von $\bar{x}$ abziehen und einmal dazuzählen und erhält so die obere und untere Grenze des Konfidenzintervalls.

Bei großen Stichproben kann bei der Berechnung des Konfidenzintervalls die Student-Verteilung durch eine Normalverteilung ersetzt werden.

Wie gut wir mit unserer Schätzung liegen, ist jetzt davon abhängig, wie gut wir – vor der Beobachtung der Stichprobe – die Standardabweichung der Stichprobe abschätzen können. (Weil wir die Standardabweichung selbst auch schätzen, bezeichnen wir sie mit $\hat{s}$). Manchmal ist das gar nicht so einfach. Vielleicht ist es ja einfacher abzuschätzen, wie groß der kleinste und größte Wert der Stichprobe ungefähr sein werden. Daraus können wir nach 3.17 die Spannweite Δ ausrechnen (siehe S.69) und daraus wiederum die Standardabweichung abschätzen:

$$\hat{s} = \frac{\Delta}{6} \tag{6.8}$$

(Wie wir auf diese Faustformel kommen? Auf Seite 121 haben wir festgestellt, dass bei einer Normalverteilung ca. 99.7% aller Daten – also schon ziemlich viele – innerhalb eines Bereichs im Abstand von maximal drei Standardabweichungen links und rechts vom Erwartungswert liegen. Macht zusammen sechs Standardabweichungen...).

Beispiel 30 *Wir wollen das 95%-Konfidenzintervall für den Erwartungswert der Körpergröße aller WIBA-Studierenden angeben und dabei soll die Intervalllänge nicht größer als 5 cm sein, d.h. der Erwartungswert soll auf 2.5 cm genau geschätzt werden. Wie groß sollte die Stichprobe mindestens sein, damit wir das schaffen?*

Aus Erfahrung können wir annehmen, dass kein:e Studierende:r kleiner als 150 cm und keine:r größer als 200 cm sein wird. (Und wenn doch, dann sind es nur ganz, ganz wenige). Daraus ergibt sich:

$$\hat{s} = \frac{200 - 150}{6} = \frac{50}{6} = 8.33$$

Für ein Konfidenzniveau von 95% können wir außerdem angeben (z.B. aus Excel): $z =$

1.96, *und somit:*

$$n > \left(\frac{2 \cdot 1.96}{5} \cdot 8.33\right)^2 = 42.7$$

D.h. damit wir ein 95%-Konfidenzintervall für den Erwartungswert der Körpergröße erhalten, das nicht größer als 5 cm ist, sollten wir mindestens 43 Personen in die Stichprobe aufnehmen.

Aufgabe 26 Wir wollen mittels Umfrage herausfinden, mit wie viel Stunden Schlaf unsere Studierenden während des Semesters pro Nacht auskommen (müssen). Konkret wollen wir letztlich durch ein 90%-Konfidenzintervall den Erwartungswert der Schlafdauer auf eine Viertelstunde genau schätzen können. Methodisch bedienen wir uns dabei einer einfachen WhatsApp-Umfrage, d.h. wir ersuchen die Studierenden, uns ein WhatsApp mit den «Schlafstunden» der letzten Nacht zu schicken. Wir erwarten eine maximale Rücklaufquote von 20%, d.h. nur jede:r fünfte Studierende:r, die wir einladen mitzutun, wird uns eine WhatsApp-Nachricht zurückschicken.
Wie viele Personen sollten wir einladen, an der Untersuchung teilzunehmen?

Schätzung des Konfidenzintervalls für den Anteilswert einer Grundgesamtheit

Nicht immer geht es um die direkte Angabe von Mittel- oder Erwartungswerten. Manchmal interessiert uns auch der *Anteil* oder *Prozentsatz* der Merkmalsträger, die eine bestimmte Eigenschaft haben. Wir nennen das dann auch **Anteilswert** bzw. auch: *Quote*). Beispiele sind Ausschussquoten in Produktionsbetrieben, Arbeitslosenquoten, Marktanteile, Einschaltquoten, der Frauenanteil in den Leitungsgremien österreichischer Unternehmen, der Anteil von Personen mit einer bestimmten Meinung oder der Anteil von roten Schokolinsen in einer Packung m&m etc.

Mathematisch ist der Anteilswert definiert als Quotient «Teilgruppengröße durch Gesamtgruppengröße» und lautet für eine Stichprobe mit $n =$ Elementen:

$$\hat{p} = \frac{x}{n} \tag{6.9}$$

mit $x =$ Anzahl der «Treffer», also die Anzahl der Elemente, die die interessierende Eigenschaft haben[4].

[4]Wenn dir bei dieser Formel eine Ähnlichkeit zur Formel 2.5 für die relative Häufigkeit auffällt, ist das kein Zufall…

$\hat{p}$ wird oft in Prozent angegeben. Das ˆ über dem p soll andeuten, dass es sich hier um einen Kennwert aus einer Stichprobe handelt. In den allermeisten Fällen haben wir ja nur eine Stichprobe zur Verfügung. Aus ihm schätzen wir den Anteilswert p der Grundgesamtheit. Es werden zum Beispiel 1 000 Personen (Eltern von Kindern bis 14 Jahren) zur Wiederaufnahme des Schulbetriebs während der Coronakrise befragt und daraus abgeleitet, wie die Mehrheit der österreichischen Eltern darüber denkt. Angenommen, 520 der Befragten – also mehr als die Hälfte – begrüßen die Öffnung der Schulen, wie «sicher» ist die Aussage «Die Mehrheit der Eltern ist für eine Schulöffnung während Corona» bzw. wie hoch ist der Anteilswert in der Grundgesamtheit, wenn wir einen Schätzwert aus der Stichprobe von $\hat{p} = 0.52$ haben? Um diese Frage zu beantworten, können wir wieder ein Konfidenzintervall heranziehen. In dem Fall hat es die beiden Grenzen:

$$L = \hat{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{6.10}$$

$$U = \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{6.11}$$

Beispiel 31 52 *von* 100 *Befragten sprachen sich in einer Umfrage für eine Öffnung des Schulbetriebs trotz Coronakrise aus. Gib ein* 95%*−Konfidenzintervall für den Anteilswert der zugehörigen Grundgesamtheit an.*

Für $\alpha = 0.05$ *ist* $z = 1.96$ *und mit* $n = 100$ *und* $\hat{p} = \frac{52}{100} = 0.52$ *ergibt sich:*

$$L = 0.52 - 1.96\sqrt{\tfrac{0.52(1-0.52)}{100}} = 0.422$$
$$U = 0.52 + 1.96\sqrt{\tfrac{0.52(1-0.52)}{100}} = 0.618$$

Der Anteilswert der Grundgesamtheit liegt demnach zu 95% *zwischen* 42.2% *und* 61.8%. *D.h. wir können – aus Sicht der Statistik – trotz der* 52% nicht *davon sprechen, dass eine* Mehrheit *der Eltern für eine Schulöffnung ist. Es könnten auch «nur»* 42% *sein.*

Aufgabe 27 Welcher Anteil $\hat{p}$ aus einer Stichprobe mit $n = 183$ Personen müsste sich für einen Gesetzesvorschlag aussprechen, damit wir mit 95%iger Sicherheit darauf schließen können, dass eine (einfache) Mehrheit der Grundgesamtheit für den Beschluss dieses Gesetzes ist?

Im Zusammenhang mit Marktforschung und Meinungsumfragen wird das durch die Grenzen 6.10 und 6.11 definierte Konfidenzintervall auch **Schwankungsbreite** genannt. Konkret ist damit der Wert gemeint, den wir in 6.10 und 6.11

zum empirisch erhaltenen Anteilswert $\hat{p}$ dazuzählen bzw. von ihm abziehen, also:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{6.12}$$

Beispiel 32 *Bei der Befragung von 500 potenziellen Kundinnen und Kunden kommt man zum Ergebnis, dass für ein bestimmtes Produkt ein Marktanteil von 30% zu erwarten ist. Wie groß ist die dabei die Schwankungsbreite?*

Für $\alpha = 0.05$ *ist* $z = 1.96$. $n = 500$ *und* $\hat{p} = 0.30$, *daraus ergibt sich:*

$$d = 1.96 \cdot \sqrt{\frac{0.3\,(1-0.3)}{500}} = 0.040$$

Die Schwankungsbreite beträgt also ± 4 *Prozentpunkte.*

Streng genommen handelt es sich bei 6.10 und 6.11 um Annäherungen an die exakten Formeln. Für unsere Zwecke ist dies aber ausreichend genau, zumindest sofern gilt: $n \geq 100$, $n\hat{p} > 5$ und $n\,(1-\hat{p}) > 5$.

Wenn zwar $n\hat{p} > 5$ und $n\,(1-\hat{p}) > 5$ gilt, aber $n < 100$, dann muss man – sofern das Ergebnis auf 1% genau sein soll – noch eine **Kontinuitätskorrektur** anbringen:

$$L = \left(\hat{p} - \frac{1}{2n}\right) - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{6.13}$$

$$U = \left(\hat{p} + \frac{1}{2n}\right) + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \tag{6.14}$$

Für alle, die von verschiedenen (nicht-industriellen) Arten zu fischen eine Vorstellung haben: Durch einen Punktschätzer (vgl. 126) den wahren Wert eines Modellparameters zu erhalten ist vergleichbar mit Speerfischen. Das Konfidenzintervall hingegen gleicht einem Wurfnetz. Und damit werden die meisten von uns mehr Erfolg haben, insbesondere wenn es darum geht, einen ganz bestimmten Fisch zu fangen.
Dabei ist auch klar: Je größer das Netz ist, das wir dabei verwenden, desto plausibler ist es, dass dieser besondere Fisch im Fang auch enthalten ist…

An dieser Stelle noch der Hinweis, dass wir nicht nur für den Erwartungswert oder den Anteilswert einer Verteilung ein Konfidenzintervall ausrechnen können, sondern auch für jeden anderen Parameter, z.B. für eine Standardabweichung, einen Korrelationskoeffizienten, den Anstieg einer Regressionsgeraden etc. In dieser Lehrveranstaltung genügen uns aber Erwartungs- und Anteilswert als Beispiele.

6.3 Kompatibilitätsintervalle und signifikante Unterschiede

Wir können Konfidenzintervalle auch dazu benutzen, zwei Stichproben dahingehend zu vergleichen, ob sie sich bezüglich eines Parameters nur zufallsbedingt unterscheiden oder **signifikant**.

Wenn letzteres zutrifft (signifikanter Unterschied), dann gehen wir davon aus, dass die beiden Stichproben zu verschiedenen Grundgesamtheiten gehören – sich also auch die dahinterliegenden Grundgesamtheiten unterscheiden. Gehören sie hingegen zur selben Grundgesamtheit, dann unterscheiden sich die Parameter nicht sigifikant sondern **zufällig**.

Wir nennen die Konfidenzintervalle dann auch **Kompatibilitätsintervalle**[5] und entscheiden so:

Wenn sich die Kompatibilitätsintervalle der beiden Stichproben nicht oder nur wenig *überschneiden*, dann stammen die Stichproben vermutlich aus verschiedenen Modellen.

Wenn die Überschneidung größer ist, sind die beiden Stichproben miteinander «kompatibel»[6] und deuten auf dasselbe Modell hin.

[5] Das ist nur ein andere Name. Die Berechnung erfolgt gleich wie in den Abschnitten des Kapitels 6.2 angegebenen Fällen.

[6] Daher der Name Kompatibilitätsintervall

7 Induktive Statistik: Schlüsse ziehen

Da wir bei der Datenanalyse in der Regel Stichproben vorliegen haben, gibt es keine gesicherten Aussagen über die Parameter des theoretische Modells, das wir dem beobachteten Phänomen zugrunde legen, also über die so genannte «Grundgesamtheit». Dennoch wollen wir Aussagen über die Grundgesamtheit tätigen, oder zumindest solche, die für eine größere Anzahl von Stichproben zutreffen. Wir werden dazu im abschließenden Kapitel das Prinzip *statistischer Tests* anschauen, das sind Tests, bei denen wir *Hypothesen* über die Parameter der Grundgesamtheit oder über die Ähnlichkeit zweier Stichproben aufstellen, und die Gültigkeit dieser Hypothesen auch so gut es geht überprüfen.

7.1 Prinzip statistischer Tests

Zunächst einmal einige Beispiele dafür, was wir mit statistischen Tests überprüfen können:

- Eine Imbisskette wirbt damit, dass in ihren Semmeln mindestens 130g Leberkäse enthalten sind. Einige Kunden vermuten aber, dass die Stücke viel kleiner sind. Zehn von ihnen wägen ihren Leberkäse nach. Es ergibt sich, dass *im Durchschnitt* dieser Stichprobe eine Portion Leberkäse nur 129.4g wiegt. Ist das nur ein stichprobenbedingter Zufall? Oder steckt da Methode dahinter, und die Stücke sind zu klein?

- Jemand möchte für eine bestimmte Entscheidung wissen, ob sich die mittlere Jahrestemperatur in Wiener Neustadt von jener in Villach unterscheidet.

- ▷ Ich möchte feststellen, ob sich bei Hamstern in zwei nebeneinanderliegenden Käfigen das Risiko einer kontaktlosen Übertragung des Arenavirus durch das Tragen von FFP2-Schutzmasken verringern lässt.

Es geht im Folgenden also darum, entweder zwei Stichproben miteinander zu vergleichen, oder eine Stichprobe mit der ihr zugrunde liegenden Grundgesamtheit. Für diese Vergleiche können wir die jeweiligen Parameter – in unseren Fällen meist Mittelwerte – heranziehen und sie mit Hilfe so genannter **Hypothesentests** überprüfen (manchmal auch als *Signifikanztest*[1] bezeichnet).

Ausgangspunkt ist dabei zunächst eine bestimmte Frage, wie zum Beispiel: «*Diese Stichprobe ergibt einen Mittelwert von 129.4 g Leberkäse pro Semmel. Erwartet hätte ich 130 g. Ist das nur zufällig oder hat auch die zugehörige Grundgesamtheit einen Erwartungswert ungleich 130 g?*»

Wir treffen dann eine Annahme, die wir **Arbeitshypothese** nennen, und die die Antwort auf die oben gestellte Frage als Behauptung formuliert. Will ich zum Beispiel wissen, ob eine Grundgesamtheit einen Erwartungswert ungleich $130g$ hat, so könnte meine Arbeitshypothese lauten: $H_A : \mu \neq 130$.

Für andere Fragestellungen benötige ich entsprechend andere Hypothesen. Dabei gibt es mathematisch-formal drei Arten von Arbeitshypothesen[2]:

Fall 1a: $H_A : \mu < 130$
Fall 1b: $H_A : \mu > 130$
Fall 2 : $H_A : \mu \neq 130$

(wobei 130 nur ein Beispiel ist und stattdessen auch ein anderer Wert stehen kann.)

Ziel des Hypothesentests ist es, eine gewählte Hypothese zu akzeptieren oder zu verwerfen. Unsere Entscheidung ist abhängig davon, ob die in der Stichprobe beobachteten Werte eher unwahrscheinlich sind, sollte die Hypothese wahr sein. Wir überprüfen also ein *Modell* (die Grundgesamtheit) anhand der *Daten* aus der Stichprobe (mitunter auch aus mehreren Stichproben): Solange Modell und Daten konsistent sind, gibt es keinen guten Grund, die Hypothese nicht zu akzeptieren. Passen sie aber nicht gut zueinander[3], dann lehnen wir sie ab.

Wenn sich unsere Daten «hypothesenkonform» zeigen und wir die Hypothese akzeptieren, heißt das aber nicht, das wir irgendwas «beweisen» konnten. Tat-

[1] vom lat. *significanter* = klar, deutlich
[2] Beachte: In unseren Fällen haben Arbeitshypothesen nie ein Gleichheitszeichen. Die stehen alle in den *Nullhypothesen*, zu denen wir gleich kommen.
[3] Was «*nicht gut zueinander passen*» bedeutet, müssen wir noch definieren…

sächlich lässt sich mit Stichproben gar nichts *beweisen*. Wenn ein Experiment mit den theoretischen Voraussagen übereinstimmt, heißt das noch nicht, dass die Theorie richtig ist. Es könnte ja auch eine andere, uns unbekannte Theorie zu diesen Ergebnissen geführt haben.

Theorien lassen sich allerdings durch ein einziges negatives Experiment widerlegen[4]. Daher gehen wir folgendermaßen vor: Zu jeder Arbeitshypothese formulieren wir noch eine zweite Hypothese, die genau das Gegenteil behauptet. Wir nennen das die **Nullhypothese**. Stellt sich dann heraus, dass die Nullhypothese nicht zutrifft, können wir daraus schließen, dass die Arbeitshypothese richtig sein muss – also genau, was wir insgeheim ohnehin zeigen wollten.

Bei Hypothesentests wird also immer eine Arbeitshypothese H_A aufgestellt, und dann die zugehörige Nullhypothese H_0 getestet. Wenn im Zuge des Tests anhand einer (oder mehrerer) Stichproben H_0 verworfen wird, können wir H_A akzeptieren.

Arbeitshypothese und Nullhypothese: Nicht jedes Ding hat zwei Seiten

Bei der Wahl der Hypothesen müssen wir unterscheiden, ob uns die Abweichungen des getesteten Parameters nach oben und unten gleich wichtig sind oder nur in eine Richtung interessieren.

Hypothesen der Form[5]

$$H_A : \mu \neq 100 \tag{7.1}$$

mit der Nullhypothese

$$H_0 : \mu = 100 \tag{7.2}$$

sind so genannte *zweiseitige Fragestellungen*. Die Abweichungen des Erwartungswertes μ von 100 sind nach oben oder unten gleich wichtig, d.h. alle abweichenden Parameterwerte, seien sie größer als 100 oder kleiner, bringen die Nullhypothese zu Fall.

[4] Von *Karl Popper* (1902-1994) stammt dazu folgendes berühmte Beispiel: Nimm an, du wolltest die Theorie prüfen «Alle Raben sind schwarz». Du beobachtest 100 Raben und stellst tatsächlich fest, dass jeder Rabe schwarz ist. Ist mit diesem Ergebnis die Theorie bewiesen? Popper sagt: Es könnte auch sein, dass der 101. Rabe, den man irgendwo beobachtet, weiß ist, und die Theorie «Alle Raben sind schwarz» wäre mit einem Schlag widerlegt.

[5] Auch hier ist der Wert 100 nur ein Beispiel und an seiner Stelle kann auch ein beliebig anderer konkreter Wert stehen. Und auch das μ kann durch einen anderen Parameter ersetzt werden.

Umgekehrt sind Hypothesentests der Form

$$H_A: \quad \mu < 100 \tag{7.3}$$
$$H_0: \quad \mu \geq 100 \tag{7.4}$$

bzw.

$$H_A: \quad \mu > 100 \tag{7.5}$$
$$H_0: \quad \mu \leq 100 \tag{7.6}$$

einseitige Fragestellungen, d.h. nur die Abweichung in eine Richtung ist interessant. Testen wir zum Beispiel ein bestimmtes Qualitätsmerkmal, so bedeutet die Unterschreitung eines vorgegebenen Sollwertes eine «schlechte» Qualität und das Ausscheiden des untersuchten Merkmalsträgers. Die Überschreitung hingegen hat in dem Fall keine negativen Folgen für die Qualität und ist daher OK.

Verspricht zum Beispiel der Hersteller einer Batterie eine Lebensdauer von «100 Lichtstunden» für die Verwendung in einer bestimmten Taschenlampe, so testen wir die Nullhypothese $H_0 : \mu \geq 100$ gegen die Arbeitshypothese $H_A : \mu < 100$ (einseitiger Test) und nicht $H_0 : \mu = 100$ gegen $H_A : \mu \neq 100$ (zweiseitiger Test). Aus Konsument:innensicht heißt ja «100 Lichtstunden» *mindestens* 100 Stunden, wir sind aber mit 110 oder 130 Stunden auch zufrieden.

Betrachten wir die Abfüllanlage einer Molkerei, die in jede Packung 1 Liter Milch einfüllen soll, so werden die Konsument:innen gegebenenfalls ebenfalls eine einseitige Fragestellung testen, die Molkerei hingegen wird einen zweiseitigen Test durchführen, weil aus ihrer Sicht auch eine Abweichung nach oben (zuviel Milch) negative Konsequenzen hat.

Die möglichen Formen von Nullhypothesen und Arbeitshypothesen für einseitige und zweiseitige Hypothesentests des Erwartungswertes sind in Tab. 7.1 zusammengefasst. (Sie entsprechen den Formeln (7.1) bis (7.6), wobei der konkrete Wert 100 durch die Variable m_0 ersetzt wurde).

Fehler erster und zweiter Art

Wir hoffen natürlich, dass wir uns mit unseren Stichproben ein gutes Spiegelbild der Grundgesamtheit beschafft haben. Trotzdem: Egal wie unsere Entscheidung

H_A	H_0	Art der Fragestellung
$\mu < m_0$	$\mu \geq m_0$	einseitig
$\mu > m_0$	$\mu \leq m_0$	einseitig
$\mu \neq m_0$	$\mu = m_0$	zweiseitig

Tabelle 7.1: Arbeitshypothesen und Nullhypothesen bei ein- bzw. zweiseitigen Hypothesentests des Erwartungswertes, wobei m_0 für eine beliebige konkrete Zahl steht

bezüglich der Nullhypothese ausfällt, es verbleibt immer eine gewisse Unsicherheit. Diese Unsicherheit hängt damit zusammen, dass wir unsere Schlüsse aus einer Stichprobe schließen. Hätten wir eine andere Stichprobe «erwischt», würde das Ergebnis vielleicht ein wenig anders aussehen. Letztlich hängt die Unsicherheit unserer Entscheidung also vom Zufall ab. Damit können wir ihr aber eine Wahrscheinlichkeit zuordnen: Wir nennen sie die **Irrtumswahrscheinlichkeit** α (auch: das *Signifikanzniveau*). α ist die Wahrscheinlichkeit dafür, dass bei einem Hypothesentest die Nullhypothese H_0 abgelehnt wird, obwohl sie wahr ist. Wir nennen dies auch einen *Fehler erster Art* (siehe Tab. 7.2).

Üblicherweise[6] wählen wir für $\alpha = 0.05$. Eine Irrtumswahrscheinlichkeit von $\alpha = 0.05$ bedeutet: Wenn wir den Hypothesentest häufig durchführen, so werden wir in 5 von 100 Fällen die Nullhypothese irrtümlich ablehnen.

Die Gegenwahrscheinlichkeit $(1 - \alpha)$ heißt auch **Sicherheitswahrscheinlichkeit**. Sie gibt an, mit welcher Wahrscheinlichkeit wir eine richtige Nullhypothese als solche erkennen und nicht ablehnen.

Wir können bei einem Hypothesentest auch den Fehler begehen, eine falsche Nullhypothese nicht abzulehnen. Dies nennen wir einen *Fehler zweiter Art* und ordnen ihm die Wahrscheinlichkeit β zu.

Die Gegenwahrscheinlichkeit $(1 - \beta)$ ist die *«Teststärke»* (auch: *Macht des Testes*). Sie gibt an, mit welcher Wahrscheinlichkeit eine falsche Nullhypothese tatsächlich als solche entlarvt und abgelehnt wird. Es ist also die Wahrscheinlichkeit, einen Fehler zweiter Art zu verhindern. (Auch hier gilt wieder: Das ist nicht dasselbe wie die Frage, wie wahrscheinlich es ist, dass die Nullhypothese falsch *ist*).

Tabelle 7.2 fasst dies noch einmal zusammen.

[6] 1931 beschrieb *Ronald Fisher* (1890-1962) in seinem Buch *The Design of Experiments*, dass für viele wissenschaftliche Experimente ein α von 0.05 («1 aus 20») ein angemessener Wert für das Signifikanzniveau sei. Seitdem wurde dieser Wert von vielen Disziplinen ohne weiteres Hinterfragen übernommen. – Wir werden es ebenso tun.

Die Sicherheitswahrscheinlichkeit gibt an, mit welcher Wahrscheinlichkeit wir eine richtige Nullhypothese auch als solche erkennen – was aber nicht gleichbedeutend ist mit der Wahrscheinlichkeit, dass die Nullhypothese richtig *ist*.

Das mag beim ersten Durchlesen verwirrend klingen; vielleicht hilft folgendes Beispiel: Angenommen, du erhältst eine Mail in deinen Posteingang, bei der im Header «vera.wormser@heidelberg.com» als Absenderin angegeben ist. Tatsächlich hattest du vor 18 Jahren einmal in Heidelberg eine Freundin namens Vera Wormser. Es könnte sich aber auch um einen Fall von Mail-Spoofing handeln.

Bevor du jetzt was Unüberlegtes tust, könntest du darüber nachdenken, wie groß die Wahrscheinlichkeit ist, dass sich Vera wirklich bei dir meldet und wenn dir diese Wahrscheinlichkeit hoch genug erscheint, die Mail öffnen.

Oder du schätzt ab, wie groß die Wahrscheinlichkeit ist, dass du an verschiedenen Kriterien und Hinweisen erkennen würdest, ob die Mail tatsächlich von Vera stammt und danach entscheiden, die Mail zu öffnen (oder zu löschen…). Und diese Wahrscheinlichkeit kann ganz anders eingeschätzt werden als die oben angegebene.

Annahme oder Verwerfen der Hypothese

Nach welchen Kriterium entscheiden wir jetzt, ob wir eine Hypothese annehmen oder ablehnen?

Für die Durchführung des Hypothesentests benötigen wir eine *Testfunktion* und Kenntnisse[7] über deren Verteilung unter der Annahme, dass H_0 zutrifft. Wir

[7] *Wir* benötigen diese Kenntnisse zum Glück nicht sehr genau sondern verlassen uns auf die Vorgangsweisen, die Mathematiker:innen und Statistiker:innen in der Vergangenheit gefunden haben.

	H_0 ist richtig	H_0 ist falsch
H_0 annehmen H_A verwerfen	richtige Entscheidung $P = (1 - \alpha) =$ *Sicherheitswahrscheinlichkeit*	Fehler 2. Art H_0 annehmen, obwohl H_A gilt: $P = \beta$
H_0 verwerfen H_A annehmen	Fehler 1. Art H_A annehmen, obwohl H_0 gilt: $P = \alpha =$ *Irrtumswahrscheinlichkeit*	richtige Entscheidung $P = (1 - \beta) =$ *Teststärke*

Tabelle 7.2: Entscheidungsmöglichkeiten und zugehörige Wahrscheinlichkeiten bei einem statistischen Hypothesentest

nennen die Testfunktion allgemein $F(\mathbf{X})$.

Für eine konkrete Stichprobe können wir eine *Realisierung* von $F(\mathbf{X})$ bestimmen, d.h. eine konkrete *Prüfgröße* f ausrechnen. Mit dieser Prüfgröße sind wir nun in der Lage, die Nullhypothese zu beurteilen. Dazu müssen wir zuvor noch ein Intervall dergestalt bestimmen, dass der Wert f mit einer Wahrscheinlichkeit von $(1 - \alpha)$ in diesem Intervall enthalten ist. Die Grenzen dieses Intervalls nennen wir die *Sicherheitsgrenzen*; der Bereich, der außerhalb dieses Intervalls liegt, führt zur Ablehnung von H_0 und wir bezeichnen ihn als *kritischen Bereich*, auch: *Verwerfungsbereich*.

Liegt die Prüfgröße f innerhalb der Sicherheitsgrenzen, so wird die Nullhypothese H_0 angenommen, weil ihr die vorliegenden Stichprobendaten nicht widersprechen. Liegt die Prüfgröße im kritischen Bereich, so verwerfen wir H_0 und akzeptieren die Arbeitshypothese H_A.

Arten statistischer Hypothesen

Parameterhypothesen sind Annahmen über einen (unbekannten) Parameter des Modells wie zum Beispiel in Tabelle 7.1 für solche über den Erwartungswert. (Siehe S.148)

Unterschiedshypothesen beziehen sich auf den Unterschied zweier Stichproben und vergleichen zum Beispiel die aus ihnen erhaltenen Mittelwerte. (Siehe S.154). Wir sprechen dabei auch von Tests für *unabhängige Stichproben*.

Veränderungshypothesen sind Unterschiedshypothesen sehr ähnlich, nur vergleichen sie Daten, die aus der Messung oder Beobachtung derselben Merkmalsträger zu verschiedenen Zeitpunkten. Verglichen wird dabei zum Beispiel ein «Vorher-Zustand» mit einem «Nachher-Zustand» (z.B. $H_0 : \mu_{vorher} = \mu_{nachher}$ gegen $H_A : \mu_{vorher} \neq \mu_{nachher}$). Nachdem wir hier dieselben Merkmalsträger (zweimal) untersuchen, sprechen wir auch von Tests für *verbundene Stichproben*. (Siehe S.156)

Zusammenhangshypothesen postulieren einen Zusammenhang zwischen zwei Zufallsvariablen und testen dann zum Beispiel, ob der Korrelationskoeffizient des Modells gleich Null ist oder nicht, also[8] $H_0 : \rho = 0$ gegen $H_A : \rho \neq 0$, siehe auch S.157.

[8] Den empirischen Korrelationskoeffizienten haben wir in Formel 4.7 angegeben. Für das theoretische Pendant verwenden wir wieder – wie üblich – den entsprechenden griechischen Buchstaben, hier ein «Rho» ρ.

Parameter-, Unterschieds-, Veränderungs- und Zusammenhangshypothesen können sich auf einen einzigen Wert beziehen und werden dann auch *Punkthypothesen* oder **ungerichtete Hypothesen** genannt, oder auf einen ganzen Bereich, so genannte *Bereichshypothesen* oder **gerichtete Hypothesen**[9].

Und dann gibt es noch **Verteilungshypothesen**, das sind Annahmen über die Form der Verteilung des Modells (= der Grundgesamtheit), die wir auf Grund der Verteilungsform der Stichprobe aufstellen. Zum Beispiel könnten wir die Hypothese aufstellen, dass bestimmte Merkmale *normalverteilt* sind. Wir nennen Tests, die Verteilungshypothesen prüfen, **Anpassungstests**. Sie sind aber nicht Gegenstand dieser einführenden Lehrveranstaltung (oder dieser Unterlagen).

Bedenke auch, wenn du zum Beispiel in deiner Abschlussarbeit statistische Hypothesen aufstellst und überprüfst, dass du zuvor je nach zu verwendendem Test vielleicht prüfen musst, ob gleiche Varianzen der Stichproben vorliegen (so genannte *Varianzhomogenität*; getestet werden kann das mit einem *Levene-Test*) oder dass die verwendeten Daten normalverteilt sind – was man übrigens manchmal auch ausreichend durch einen graphischen Vergleich der Dichtefunktion mit einer Glockenkurve (siehe Abb.5.9 auf Seite 120) herausfinden kann. Mathematisch-statistisch genauere Tests auf Normalverteilungen sind z.B. der *Kolmogorov-Smirnov Test* (auch: *KS-Test*), ein *Shapiro-Wilk Test* oder ein *Anderson-Darling Test*). Für konkrete Beispiele dafür sei auf weiterführende Literatur verwiesen.

7.2 Testen von Parameterhypothesen

t-Test eines Mittelwerts aus normalverteilten Daten einer kleinen Stichprobe

Wenn wir zwar eine kleine Stichprobe haben, aber von einer Normalverteilung der Daten ausgehen, dann können wir den Mittelwert dieser Stichprobe mit einem so genannten **t-Test** testen. Das *t* bezieht sich auf den Namen der Wahrscheinlichkeitsverteilung, die wir bei diesem Test verwenden: Die *Student-* oder *t*-Verteilung. Wie sie aussieht, können wir in Abb.5.7 (Seite 118) rechts oben feststellen. Und wie man ihre Funktionswerte berechnet, wissen wir auch schon: Sie ist bereits in Formel 6.1 und 6.2 vorgekommen und auf Seite 131 ihre Berechnung in *Excel* oder *R* angegeben.

Beim t-Test (genauer: beim **einfachen t-Test**) eines Mittelwerts wollen wir über-

[9] Der Unterschied ist ziemlich einfach: Wenn im Hypothesenpaar H_0, H_A nur die Vergleichssymbole $=$ und $\neq$ vorkommen, handelt es sich um *ungerichtete* Hypothesen; wenn die Ungleichheitszeichen $\leq$ und $>$ bzw. $\geq$ und $<$ vorkommen, um *gerichtete*.

prüfen, ob der unbekannte Erwartungswert μ einer normalverteilten Zufallsvariablen X einen bestimmten Wert m_0 besitzt bzw. über- oder unterschreitet. m_0 kann zum Beispiel ein Sollwert bei der Herstellung eines Produkts sein. Dabei kennen wir für μ und σ nur Schätzwerte, nämlich das arithmetische Mittel $\bar{x}$ und die empirische Standardabweichung s.

Beispiel 33 *Als einfaches Beispiel können wir die Herstellung von Brotlaiben betrachten. Deren (in kg gemessene) Masse X sei normalverteilt. Das angegebene Verkaufsgewicht des Brotes sei $\mu = 2$ kg. Eine Konsumentenschutzorganisation zieht nun eine Stichprobe von $n = 20$ Brotlaiben und stellt einen Stichprobenmittelwert von $\bar{x} = 1.97$ kg und eine empirische Standardabweichung von $s = 0.1$ kg fest. Es soll überprüft werden, ob diese Stichprobe gegen die Hypothese spricht, dass die Brote der Grundgesamtheit mindestens 2 kg wiegen – dass wir also vom Verkäufer nicht bedackelt werden.*

Zunächst sind Arbeits- und Nullhypothese festzulegen (vgl. auch Tab.7.1):

- Für die *einseitige* Fragestellung lauten sie (je nachdem, welche Richtung für uns interessant ist):
  Fall 1a: $H_A : \mu < m_0 \rightarrow H_0 : \mu \geq m_0$ oder
  Fall 1b: $H_A : \mu > m_0 \rightarrow H_0 : \mu \leq m_0$

- Für eine *zweiseitige* Fragestellung:
  Fall 2: $H_A : \mu \neq m_0 \rightarrow H_0 : \mu = m_0$

Beispiel 33 *(Fortsetzung)*

Im konkreten Beispiel geht es um eine einseitige Fragestellung, weil wir ja nur unzufrieden sind, wenn das Brot weniger als 2 kg wiegt. Wir wollen also herausfinden: Deutet der Mittelwert $\bar{x} = 1.97$, den wir aus der Stichprobe erhalten haben, auf eine Grundgesamtheit hin, deren Erwartungswert μ signifikant kleiner als $m_0 = 2$ ist?

Wir wählen als Arbeitshypothese bzw. als Nullhypothese (entsprechend Fall 1a):

$$H_A : \mu < 2$$

$$H_0 : \mu \geq 2$$

Anschließend ist eine Irrtumswahrscheinlichkeit festzulegen. Wir werden den üblichen Wert von $\alpha = 0.05$ wählen.

Als Testfunktion ziehen wir folgende Funktion heran:

$$F(\mathbf{x}) = \frac{\bar{x} - m_0}{s}\sqrt{n} \tag{7.7}$$

Beispiel 33 *(Fortsetzung)*

Aus der Realisierung der Stichprobe unseres Beispiels können wir die konkrete Prüfgröße angeben:

$$f = \frac{\bar{x} - m_0}{s}\sqrt{n} = \frac{1.97 - 2}{0.1}\sqrt{20} = -1.34$$

Nun bestimmen wir den kritischen Bereich. Unter H_0 besitzt die Funktion eine *t-Verteilung* mit $(n-1)$ Freiheitsgraden. Die entsprechenden Werte für den kritischen Bereich erhalten wir aus einer Tabelle oder einem Programm, und dann können wir eine Entscheidung treffen: Die Nullhypothese wird abgelehnt, falls die Testgröße im kritischen Bereich liegt, andernfalls wird H_0 akzeptiert.

H_A	H_0	Prüfgröße	Entscheidung
$\mu < m_0$	$\mu \geq m_0$	$f < -t_{(1-\alpha,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -t_{(1-\alpha,n-1)}$	H_0 akzeptieren, H_A ablehnen
$\mu > m_0$	$\mu \leq m_0$	$f > t_{(1-\alpha,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq t_{(1-\alpha,n-1)}$	H_0 akzeptieren, H_A ablehnen
$\mu \neq m_0$	$\mu = m_0$	$\lvert f\rvert > t_{(1-\alpha/2,n-1)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f\rvert \leq t_{(1-\alpha/2,n-1)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.3: Mögliche Ergebnisse eines t-Tests. Den Wert für f berechnen wir aus Formel 7.7, den Wert für t entnehmen wir aus einer Tabelle oder eine Programm. Beachte: Bei zweiseitiger Fragestellung wird t für $(1-\alpha/2)$ berechnet, bei einseitiger Fragestellung hingegen für $(1-\alpha)$. Die Anzahl der Freiheitsgrade ist immer $n-1$.

In *MS Excel* heißt der Befehl zur Berechnung von t für eine einseitige Fragestellung `=T.INV(1-alpha;n-1)`, wobei für `alpha` und `n-1` die jeweiligen Werte für α und n einzusetzen sind. In *R* lautet der Befehl `qt((1-alpha),n-1)`.

Für eine zweiseitige Fragestellung müssen wir eingeben: `=T.INV(1-alpha/2;n-1)` bzw. `qt((1-alpha/2),n-1)`.

Achtung auf einseitige und zweiseitige Fragestellungen

Beispiel 33 *(Fortsetzung)*

In unserem Beispiel benötigen wir also noch den t-Wert an der Stelle $(0.95, 19)$*. Das Quantil der t-Verteilung ist (laut Excel oder R) an dieser Stelle gleich* 1.729*. Nach Tab.7.3 benötigen wir den negativen Wert davon, also* $-t = -1.729$*.*

Die Prüfgröße f *haben wir schon ausgerechnet; jetzt können wir vergleichen:*

Da $-1.3 > -1.7$*, also* $f > -t$*, werden wir* ***die Nullhypothese annehmen*** *und können* ***die Arbeitshypothese nicht mehr aufrechterhalten****.*

Das bedeutet: Die in der Stichprobe beobachtete mittlere Masse von 1.97 *ist zwar kleiner als der Sollwert* 2 *kg, diese Abweichung ist allerdings statistisch nicht signifikant sondern vermutlich zufällig bedingt. Die Wahrscheinlichkeit, aus einer Grundgesamtheit mit* $\mu = 2$ *eine Stichprobe mit einem Mittelwert von höchstens* 1.97 *zu erhalten, ist größer als* 5%*. Es gibt daher – aus Sicht der Statistik – keinen Grund, das angegebene Verkaufsgewicht von* 2 *kg zu beanstanden.*

An dieser Stelle noch ein Hinweis zum Aufstellen der Hypothesen: Gehen wir, so wie im eben gerechneten Beispiel, von einer gerichteten Hypothese aus, und schon der empirische Wert aus der Stichprobe geht in die andere Richtung, muss man schon sehr gut argumentieren (können), warum man dennoch eine gegenteilige Arbeitshypothese aufstellt. Hätten wir also aus der «Brot-Stichprobe» einen Mittelwert von $\bar{x} = 2.1\ kg$ erhalten, spricht nicht viel dafür, eine Arbeitshypothese der Form $H_A : \mu < 2$ aufzustellen. Bei einer zweiseitigen Fragestellung geht das aber nicht: Hier können wir nicht einfach Null- und Arbeitshypothese umdrehen. Wie schon in der Fußnote auf Seite 142 beschrieben: Das Gleichheitszeichen steht immer bei der Nullhypothese, nie bei der Arbeitshypothese.

Gaußtest eines Mittelwerts aus Daten einer großen Stichprobe

Wir sind beim t-Test davon ausgegangen, dass die Daten normalverteilt sind. Wenn wir darüber nicht sicher sind, aber der Stichprobenumfang n größer als 30 ist, dann können wir einen ähnlichen Test anwenden, den so genannten *Gaußtest* bzw. streng genommen den **approximativen Gaußtest**[10].

Wenn wir den Mittelwert testen, ist die Testfunktion beim Gaußtest dieselbe wie beim t-Test des Mittelwerts (Formel 7.7). Der Vollständigkeit halber schreiben

[10] benannt nach Johann Friedrich Carl Gauß, den wir schon in Fußnote 12 auf Seite 119 kennengelernt haben.

H_A	H_0	Prüfgröße	Entscheidung
$\mu < m_0$	$\mu \geq m_0$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu > m_0$	$\mu \leq m_0$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu \neq m_0$	$\mu = m_0$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.4: Mögliche Ergebnisse eines approximativen Gaußtests. Den Wert für f berechnen wir aus Formel 7.8, den Wert für z entnehmen wir aus einer Tabelle oder eine Programm.

wir sie hier nochmal auf:

$$F(\mathbf{x}) = \frac{\bar{x} - m_0}{s}\sqrt{n} \tag{7.8}$$

Den kritischen Bereich erhalten wir jetzt aber nicht aus der t-Verteilung, sondern aus einer *Standardnormalverteilung* und treffen die Entscheidung über die Annahme der Nullhypothese nach Tab.7.4. Die dafür benötigten Werte z können wir wieder aus einem Programm entnehmen.

In *MS Excel* heißt der Befehl zur Berechnung von z für eine einseitige Fragestellung `=NORM.S.INV(1-alpha)`. In *R* lautet der Befehl `qnorm(1-alpha)`.

Für eine zweiseitige Fragestellung müssen wir eingeben: `=NORM.S.INV(1-alpha/2)` bzw. `qnorm(1-alpha/2)`.

Aufgabe 28 Deine Ärztin empfiehlt dir, aus gesundheitlichen Gründen deinen Kaffeekonsum auf fünf Tassen pro Tag einzuschränken. Du bist dir eigentlich sicher, dass du das im Schnitt ohnehin nicht überschreitest ($H_0 : \mu \leq 5$), dein besorgter Ehepartner schenkt dem aber keinen so rechten Glauben und behauptet, dass es mehr als fünf Tassen pro Tag sind ($H_A : \mu > 5$). Ihr vereinbart, eine Zeitlang darüber Buch zu führen. Nach 40 Tagen hast du 210 Tassen Kaffee getrunken. Im Schnitt ergab das Experiment also 5,25 Tassen pro Tag, und das bei einer Standardabweichung von 0,25. Wessen Hypothese kann bei diesem Ergebnis aufrecht erhalten werden?

Test eines Anteilswertes

Sowohl der einfache t-Test als auch der approximative Gaußtest kann angewendet werden, wenn wir einen *Anteilswert* p testen wollen. Anteilswerte haben wir bereits auf Seite 136 untersucht – dort ging es um Konfidenzintervalle für Anteilswerte[11]. Konkret verwenden wir einen t-Test, wenn $n < 30$ und einen Gaußtest wenn $n \geq 30$.

Am Beginn beider Tests steht wie üblich die Formulierung der Hypothesen. Das Schema möglicher Hypothesenpaare kennen wir ja mittlerweile bereits; wir können einen der folgenden drei Fälle behandeln:

Fall 1a:	$H_A : p < p_0$	$\rightarrow$	$H_0 : p \geq p_0$
Fall 1b:	$H_A : p > p_0$	$\rightarrow$	$H_0 : p \leq p_0$
Fall 2 :	$H_A : p \neq p_0$	$\rightarrow$	$H_0 : p = p_0$

Die Testfunktion für den Hypothesentest des Anteilswertes unterscheidet sich ein wenig von jener für den Mittelwert. Wenn $\hat{p}$ der Anteilswert ist, den wir aus der Stichprobe erhalten haben, dann lautet sie:

$$F(\mathbf{x}) = \frac{\hat{p} - p_0}{\sqrt{p_0 \, (1 - p_0)}} \cdot \sqrt{n} \tag{7.9}$$

Nachdem wir den konkreten Funktionswert f ausgerechnet haben, ziehen wir für den Vergleich wieder entweder die Quantile der t-Verteilung (wenn $n - 1 \leq 30$) oder jene der Normalverteilung heran (wenn $n > 30$), siehe Tab.7.5.

Aufgabe 29 Eine Software-Entwicklerin überlegt, ihre Software als *Donationware* zur Verfügung stellen, d.h. sie kann grundsätzlich kostenlos verwendet werden, sie bittet aber um (freie) Spenden, damit auf Sicht wenigstens die bei ihr entstehenden Drittkosten abgedeckt sind. Sie schätzt: Nur wenn mindestens 25% der User bereit sind, 10 € zu spenden, werde ich kein Geld verlieren.
Sie startet einmal probehalber und stellt nach den ersten 500 Downloads fest: 145 User haben tatsächlich 10 € gespendet. Das sind sogar 29%. Kann sie (aus statistischer Sicht) optimistisch sein, dass der Anteil der spendenfreudigen User tatsächlich größer als 25% ist?

[11] Und wie auf Seite 138 angegeben gilt auch hier: In aller Strenge gelten die Formeln dieses Abschnitts nur, wenn $n \cdot \hat{p} > 5$ und $n \cdot (1 - \hat{p}) > 5$.

H_A	H_0	Prüfgröße wenn $n < 30$	Prüfgröße wenn $n \geq 30$	Entscheidung
$p < p_0$	$p \geq p_0$	$f < -t_{(1-\alpha,n-1)}$	$f < -z_{(1-\alpha)}$	H_0 ablehnen
		$f \geq -t_{(1-\alpha,n-1)}$	$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren
$p > p_0$	$p \leq p_0$	$f > t_{(1-\alpha,n-1)}$	$f > z_{(1-\alpha)}$	H_0 ablehnen
		$f \leq t_{(1-\alpha,n-1)}$	$f \leq z_{(1-\alpha)}$	H_0 akzeptieren
$p \neq p_0$	$p = p_0$	$\|f\| > z_{(1-\alpha/2)}$	$\|f\| > z_{(1-\alpha/2)}$	H_0 ablehnen
		$\|f\| \leq t_{(1-\alpha/2,n-1)}$	$\|f\| \leq z_{(1-\alpha/2)}$	H_0 akzeptieren

Tabelle 7.5: Entscheidungsalternativen beim Test eines Anteilswertes. Dabei gilt: Wenn H_0 abgelehnt wird, kann H_A angenommen werden.
($n < 30$: t-Test. $n \geq 30$: Gaußtest)

7.3 Testen von Unterschiedshypothesen

In diesem Abschnitt geht es um das Testen von Hypothesen über zwei unabhängige Stichproben.

Vergleich zweier Mittelwerte mittels Gaußtest bei großen Stichproben

Wenn wir zwei Stichproben mit Stichprobenumfang n_1 bzw. n_2 (die jeweils ≥ 30 sind) und den Mittelwerten $\bar{x}_1$ und $\bar{x}_2$ erhoben haben, können wir die Differenz $\bar{x}_1 - \bar{x}_2$ bilden und für den Fall, dass diese Differenz ungleich Null ist, untersuchen, ob es auch eine von Null verschiedene Differenz der beiden Erwartungswerte der jeweils zugrundeliegenden Modelle $\mu_1 - \mu_2$ gibt, de facto also zwei unterschiedliche Modelle vorliegen, aus denen die jeweiligen Stichproben stammen.

Die Arbeitshypothesen können dann lauten:

Fall 1a: $H_A : \mu_1 - \mu_2 < 0$
Fall 1b: $H_A : \mu_1 - \mu_2 > 0$
Fall 2 : $H_A : \mu_1 - \mu_2 \neq 0$

wobei wir das meist umformen zu:

Fall 1a: $H_A : \mu_1 < \mu_2$
Fall 1b: $H_A : \mu_1 > \mu_2$
Fall 2 : $H_A : \mu_1 \neq \mu_2$

Zum Beispiel könnten wir das Einkommen der Angestellten in unserem Unter-

H_A	H_0	Prüfgröße	Entscheidung
$\mu_1 < \mu_2$	$\mu_1 \geq \mu_2$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu_1 > \mu_2$	$\mu_1 \leq \mu_2$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$\mu_1 \neq \mu_2$	$\mu_1 = \mu_2$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.6: Entscheidungsalternativen beim Gaußtest zweier Mittelwerte aus Stichproben mit $n \geq 30$

nehmen untersuchen und dabei unter den Frauen und den Männer jeweils eine Stichprobe ziehen. Die Mittelwerte werden da vermutlich nicht genau gleich sein, aber ist dieser Unterschied nur zufällig und wir haben im Grunde eine Equal Pay-Situation im Unternehmen, oder ist der Unterschied so signifikant, dass wir von einem Gender-Pay-Gap sprechen müssen?

Die formale Vorgangsweise beim Testen zweier Mittelwerte läuft «wie immer» ab – mittlerweile kennen wir das ja schon:

Wir entscheiden, welche Arbeits- und Nullhypothese wir aufstellen:

$$\begin{array}{lllll} \text{Fall 1a:} & H_A : \mu_1 < \mu_2 & \rightarrow & H_0 : \mu_1 \geq \mu_2 \\ \text{Fall 1b:} & H_A : \mu_1 > \mu_2 & \rightarrow & H_0 : \mu_1 \leq \mu_2 \\ \text{Fall 2 :} & H_A : \mu_1 \neq \mu_2 & \rightarrow & H_0 : \mu_1 = \mu_2 \end{array}$$

Ist zum Beispiel $\bar{x}_1$ um einiges größer als $\bar{x}_2$, dann gilt vermutlich auch $\mu_1 > \mu_2$; ist die Differenz $\bar{x}_1 - \bar{x}_2 \approx 0$, dann testen wir eher Fall 2.

Die Testfunktion lautet:

$$F(\mathbf{x}) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 n_2 + s_2^2 n_1}} \sqrt{n_1 n_2} \tag{7.10}$$

Und das Ergebnis können wir nach dem Schema in Tabelle 7.6 finden.

Hinweis: Auch für den Zweistichprobentest gibt es wieder eine t-Test Variante, wenn wir nur kleine Stichproben vorliegen haben. Die Formel für die Testfunktion ist dann um einiges komplexer als wir uns in dieser Einführungslehrveranstaltung «zumuten» wollen…

Aufgabe 30 Nach dem ersten Studienjahr wurde für alle Studierenden eines Jahrgangs ein nach ECTS gewichteter Notenschnitt (GPA) berechnet und daraus dann ein Durchschnittswert für jeweils alle männlichen ($n_m = 84$) und alle weiblichen ($n_w = 36$) Studierenden angegeben. Für die männlichen Studierenden beträgt er $\bar{x}_m = 1.6$, für die weiblichen $\bar{x}_w = 1.4$. Ermittelt wurden auch die zugehörigen empirischen Standardabweichungen: $s_m = 0.6$, $s_w = 0.4$.

Der empirisch ermittelte GPA der Frauen unterscheidet sich offenbar von dem der Männer. Ist dieser Unterschied signifikant bzw. inwiefern lässt sich aus den obigen Daten die Hypothese $H_A : \mu_m - \mu_w \neq 0$ behaupten?

Vergleich zweier Anteilswerte

Fall 1a:	$H_A : p_1 < p_2$	$\rightarrow$	$H_0 : p_1 \geq p_2$
Fall 1b:	$H_A : p_1 > p_2$	$\rightarrow$	$H_0 : p_1 \leq p_2$
Fall 2 :	$H_A : p_1 \neq p_2$	$\rightarrow$	$H_0 : p_1 = p_2$

Testfunktion:

$$F(\mathbf{x}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{7.11}$$

Entscheidung:

H_A	H_0	Prüfgröße	Entscheidung
$p_1 < p_2$	$p_1 \geq p_2$	$f < -z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$p_1 > p_2$	$p_1 \leq p_2$	$f > z_{(1-\alpha)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq z_{(1-\alpha)}$	H_0 akzeptieren, H_A ablehnen
$p_1 \neq p_2$	$p_1 = p_2$	$\lvert f \rvert > z_{(1-\alpha/2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq z_{(1-\alpha/2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.7: Entscheidungsalternativen beim Vergleich zweier Anteilswerte

7.4 Testen von Veränderungshypothesen

Dabei handelt es sich um das Testen von Parametern von zwei *verbundenen* Stichproben. Wir könnten zum Beispiel von einer Gruppe von Personen am Faschingdienstag das Gewicht ermitteln und als Stichprobe vorhalten. 40 Tage später ermitteln wir dann das Gewicht *derselben* Personen erneut, das ist dann die mit

der ersten *verbundene* Stichprobe. Uns interessiert jetzt zum Beispiel, ob die Differenz der beiden Stichprobenmittelwerte gleich Null ist. Wir sprechen daher für diese Art von Hypotesentest auch von einem **Differenzentest**.

Die Vorgangsweise ist relativ einfach: Aus den beiden Stichproben mit den jeweils untersuchten Zufallsvariablen X bzw. Y bilden wir die paarweisen Differenzen

$$d_i = x_i - y_i \tag{7.12}$$

und danach aus allen d_i den arithmetischen Mittelwert $\bar{d}$ und die Standardabweichung s_d.

Unsere neue Zufallsvariable $D = X - Y$ hat – wie jede andere Zufallsvariable – einen Erwartungswert μ und wir können verschiedene Hypothesen über μ aufstellen:

Fall 1a: $H_A : \mu < 0 \rightarrow H_0 : \mu \geq 0$
Fall 1b: $H_A : \mu > 0 \rightarrow H_0 : \mu \leq 0$
Fall 2: $H_A : \mu \neq 0 \rightarrow H_0 : \mu = 0$

Als Testfunktion ziehen wir folgende Funktion heran:

$$F(\mathbf{x}) = \frac{\bar{d}}{s_d}\sqrt{n} \tag{7.13}$$

Je nachdem, ob wie eine große oder kleine Stichprobe vorliegen haben, ist die weitere Vorgangsweise dieselbe wie beim *t-Test* eines Mittelwerts aus normalverteilten Daten einer kleinen Stichprobe (S. 148) oder eines approximativen Gaußtests (S. 151).

7.5 Testen von Zusammenhangshypothesen

Test einer Korrelationshypothese

Fall 1a:	$H_A : \rho < 0$	$\rightarrow$	$H_0 : \rho \geq 0$
Fall 1b:	$H_A : \rho > 0$	$\rightarrow$	$H_0 : \rho \leq 0$
Fall 2 :	$H_A : \rho \neq 0$	$\rightarrow$	$H_0 : \rho = 0$

Testfunktion:

$$F(\mathbf{x}) = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \tag{7.14}$$

Entscheidung:

H_A	H_0	Prüfgröße	Entscheidung
$\rho < 0$	$\rho \geq 0$	$f < -t_{(1-\alpha,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$f \geq -t_{(1-\alpha,n-2)}$	H_0 akzeptieren, H_A ablehnen
$\rho > 0$	$\rho \leq 0$	$f > t_{(1-\alpha,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$f \leq t_{(1-\alpha,n-2)}$	H_0 akzeptieren, H_A ablehnen
$\rho \neq 0$	$\rho = 0$	$\lvert f \rvert > t_{(1-\alpha/2,n-2)}$	H_0 ablehnen, H_A akzeptieren
		$\lvert f \rvert \leq t_{(1-\alpha/2,n-2)}$	H_0 akzeptieren, H_A ablehnen

Tabelle 7.8: Entscheidungsalternativen beim Testen einer Korrelation

7.6 Abschließende Hinweise

Zunächst sei noch einmal auf die richtige Reihenfolge beim Hypothesentest verwiesen:

1. Man stellt eine Arbeitshypothese und eine Nullhypothese auf
2. Man gibt die Irrtumswahrscheinlichkeit vor und bestimmt damit einen Ablehnungsbereich
3. *Danach* wird die Stichprobe gezogen
4. Dann wird der Hypothesentest durchgeführt und entweder die Nullhypothese oder die Arbeitshypothese angenommen

Völlig unzulässig ist es, zuerst die Stichprobe zu ziehen, in den Stichprobendaten dann verschiedene Hypothesen auszuprobieren – womöglich unter mehrfacher, abwechslungsreicher Wahl von α , und dann diejenige auszuwählen, die am Besten zu meinen Daten «passt». Statistische Tests dürfen nie so ablaufen, dass die eigentliche Fragestellung erst nach der Beobachtung der Stichprobe aufgestellt wird!

Wir haben in diesem Kapitel «Einstichprobentests» (= eine Stichprobe wird gegen eine Grundgesamtheit getestet) und «Zweistichprobentests» (= jeweils zwei Stichproben werden miteinander verglichen) angesehen. Wollen wir eine Hypothese über mehr als zwei Stichproben testen, so verwenden wir eine **einfache Varianzanalyse**. Dazu sei auf weiterführende Literatur verwiesen.

Zum Begriff «Signifikanz»:

Die Trennung zwischen *signifikant* und *nicht signifikant* (also: *nicht-zufällig* versus *zufällig*) an einer bestimmten, scharfen Grenze festzumachen ist zugegebenermaßen eine etwas vereinfachte Sicht auf die Welt. Während es in technischen Anwendungen wie der Qualitätskontrolle in Produktionsprozessen vielleicht noch nachvollziehbar ist, dass man ab einem bestimmten zahlenmäßigen Wert zum Beispiel eine Toleranzgrenze überschritten hat, ist das bei der Untersuchung menschlichen Verhaltens vielleicht nicht immer ganz argumentierbar. Menschliches Verhalten ist ziemlich komplex und Menschen sind manchmal ziemlich kompliziert, und eine einfache Schwarz-Weiß-Einteilung wird dem vielleicht nicht immer gerecht. Aktuell gibt es aber kein «kontinuierliches» Signifikanzmaß, das zum Beispiel in abgestufter Form die «Plausibilität für die Zufälligkeit» angibt oder ein Intervall, in dem die Daten sowohl mit der Idee kompatibel sind, dass sie zufällig so zustandegekommen sind als auch mit der Idee, dass hier eine Signifikanz vorliegt.

A

Lösungen zu den Aufgaben

Empirischer Häufigkeitsverteilungen

1

Gib zur Stichprobe des Beispiels 3 den oder die Modalwert(e) an.

Am öftesten wurden im Beobachtungszeitraum 3 Tassen Kaffee pro Tag getrunken. Daher lautet der Modalwert: 3.

2

Gib für die Daten der Tabelle 3.4 das 0.65-Quantil **an (berechnet nach dem Näherungsverfahren) und vergleiche dein Ergebnis mit dem exakten Wert.**

$$k = 20 \cdot 0.65 = 13$$

Da dies eine ganze Zahl ist, müssen wir das Mittel aus dem 13. und 14. Element ausrechnen:

$$x_{0.65} = \frac{15.5 + 16}{2} = \frac{31.5}{2} = 15.75$$

Der exakte, mit *R* berechnete Wert ist: 15.675. Unser genäherter Schätzwert ist also ein wenig größer als der exakte Wert.

3

Gegeben sei die Körpergröße der 7 Zwerge. Bestimme das Minimum, Maximum sowie das 1. - 3. Quartil sowohl näherungsweise als auch deren exakten Werte:

Name	Größe (in *cm*)
Doc	85
Grumpy	80
Happy	62
Sleepy	81
Bashful	70
Sneezy	80
Dopey	88

Wenn wir «von Hand» rechnen, ist es praktisch, die Daten der Größe nach zu ordnen:

$$62, 70, 80, 80, 81, 85, 88$$

Damit sind Minimum und Maximum schon klar: $x_{min} = 62$, $x_{max} = 88$.

Für die Quartile von $n = 7$ Werten bilden wir:

p	$n \cdot p$	aufgerundet
0.25	1.75	2
0.5	3.5	4
0.75	5.25	6

Und daraus ergeben sich die Quartile:

$x_{0.25} = x_2 = 70$, $x_{0.5} = x_4 = 80$, $x_{0.75} = x_6 = 85$.

Für die exakte Berechnung verwenden wir die Excel-Funktion `QUARTILE.INKL` und erhalten:

$x_{0.25} = 75$, $x_{0.5} = 80$ und $x_{0.75} = 83$

4

Gegeben sei die Größe der 7 roten Zwerge, die innerhalb einer Entfernung von 10 Lichtjahren zur Erde liegen. Bestimme das Minimum, Maximum sowie das 1. - 3. Quartil sowohl «visuell» als auch deren exakten Werte:

Bezeichnung	Durchmesser (in *Tausend km*)
Luyten 726-8 A	195
Luyten 726-8 B	195
Proxima Centauri	196.4
Wolf 359	222.8
Barnards Stern	273
Ross 154	334.2
Lalande 21185	547.4

Wir ordnen wieder der Größe nach und können gleich die Quartile einzeichnen:

$$195 \quad \boxed{195} \quad 196.4 \quad \boxed{222.8} \quad 273 \quad \boxed{334.2} \quad 547.4$$

Die exakten Werte sind:

$x_{0.25} = 195.7$, $x_{0.5} = 222.8$ und $x_{0.75} = 303.6$

5

Bestimme zu den Daten der Tabelle 3.4 die Spannweite und den Quartilsabstand. Gibt es auf Grund dieser Streuungswerte Anzeichen, dass die Daten Ausreißer enthalten?

Das Minimum der Daten in Tab.3.4 ist 12, das Maximum 17. Die Spannweite beträgt daher $17 - 12 = \underline{5 \text{ Jahre}}$.

Die Quartile kann man exakt oder mit einem Näherungsverfahren berechnen. Bei der exakten Berechnung in *R* erhalten wir $\underline{\underline{IQR = 1.75}}$.

Wenden wir das Näherungsverfahren an, dann lauten die entsprechenden Quartile: $x_{0.25} = 14.25$, $x_{0.75} = 16.25$ und der Quartilsabstand ist 2.

Für den Ausreißertest nach Formel 3.20 und 3.20 gilt dann: $A_u = 11.25$ und $A_o = 19.25$, d.h. dass $\underline{\text{keines}}$ der Elemente als Ausreißer zu betrachten ist.

6

Wie groß ist die Varianz der in Beispiel 3 gegebenen Daten?

In der Stichprobe des Bsp.3 sind $n = 14$ Elemente enthalten. Die Summe ist 42, das Mittel daher $\overline{x} = \frac{42}{14} = 3$.

Die jeweiligen Differenzen zwischen gegebenen Elementen und dem Mittel sind:
$-2, 0, -2, 0, -1, -1, 2, 1, 0, -1, 0, 1, 3, 0$

bzw. deren Quadrate:
$4, 0, 4, 0, 1, 1, 4, 1, 0, 1, 0, 1, 9, 0$

Die Quadratsumme ist $4+0+4+0+1+1+4+1+0+1+0+1+9+0 = 26$ und somit:

$$s^2 = \frac{26}{14-1} = \frac{26}{13} = \boxed{2}$$

7

Der folgende Datensatz besteht aus elf Elementen und hat einen arithmetischen Mittelwert von 25. Außerdem wissen wir, dass die Daten völlig symmetrisch um den Mittelwert gestreut sind.

$$x_1, 21, 22, 23, 24, 25, 26, 27, 28, 29, x_{11}$$

Welche Werte musst du für x_1 und x_{11} einsetzen, damit die Varianz 26 beträgt?

Nachdem die Daten symmetrisch um den Mittelwert gestreut sind, müssen x_1 und x_{11} den gleichen Abstand vom Mittelwert haben.

Für x_1 schreiben wir dann $(25 - a)$ und für $x_{11} = (25 + a)$ und suchen nach einem passenden Wert a.

Die Varianz soll 26 betragen, d.h. die Quadratsumme der Abweichungen vom Mittelwert dividiert durch $(n - 1) = 10$ muss 26 betragen.

Das ergibt folgende Gleichung:

$$26 = \tfrac{1}{11-1}[(25 - a - 25)^2 + (21 - 25)^2 + (22 - 25)^2 + (23 - 25)^2 + (24 - 25)^2 + \\ + (25 - 25)^2 + (26 - 25)^2 + (27 - 25)^2 + (28 - 25)^2 + (29 - 25)^2 + (25 + a - 25)^2]$$

$$260 = \left[a^2 + 60 + a^2\right]$$
$$200 = 2a^2$$
$$100 = a^2$$
$$10 = a$$

Somit ist $x_1 = 15$ und $x_{11} = 35$.

8

Kannst du – ohne konkret jede Kennzahl auszurechnen – «auf einen Blick» sagen und begründen, welche der drei folgenden Datensätze die größte Standardabweichung hat und welche die kleinste?

A) 0, 20, 40, 50, 60, 80, 100
B) 0, 48, 49, 50, 51, 52, 100
C) 0, 1, 2, 50, 98, 99, 100

Die Standardabweichung ist abhängig von der Quadratsumme der Abstände der einzelnen Elemente vom Mittelwert.

A) B) und C) streuen alle um den Mittelwert 50, wobei C) wegen der jeweils größeren Abstände zwischen den einzelnen Werten und dem Mittelwert 50 auf die größte Quadratsumme kommen wird und B) auf die kleinste.

Daher hat C) die größte Standardabweichung und B) die kleinste.

9

Gib zu den Daten aus Aufgabe 7 die standardisierten z-Werte an.

Der Mittelwert der Daten beträgt laut Vorgaben der Angabe $\bar{x} = 25$ und die Varianz $s^2 = 26$. Somit ist die Standardabweichung $s = 5.099$ und daraus erhalten wir die standardisierten Werte:

x_i	15	21	22	23	24	25	26	27	28	29	35
z_i	-1.96	-0.78	-0.59	-0.39	-0.20	0.00	0.20	0.39	0.59	0.78	1.96

10

In der folgenden Tabelle sind für sieben in Wien abgehaltenen Wahlen die Mittagstemperatur am jeweiligen Wahltag (x) sowie das Verhältnis der abgegebenen Stimmen zur Anzahl der Wahlberechtigten, also die Wahlbeteiligung (y) gegeben. Gib den Regressionskoeffizienten an.

	28.09.08	07.06.09	10.10.10	29.09.13	25.05.14	11.10.15	24.04.16
x (Temperatur °C)	15	22	12	12	24	7	7
y (Wahlbeteiligung)	0.74	0.43	0.68	0.70	0.35	0.75	0.64

Der Regressionskoeffizient ist der Anstieg der Regressionsgeraden. Den können wir z.B. mit der Excel-Funktion `STEIGUNG` berechnen und erhalten als Ergebnis:

$$\underline{\underline{k = -0.020}}$$

11

(Fortsetzung zu Aufgabe 10): **In obiger Tabelle ist die Bundespräsidentenwahl 2010 nicht enthalten. Die Mittagstemperatur am Wahltag (25.4.2010) betrug** 18°**. Welche Wahlbeteiligung war bei dieser Temperatur zu erwarten?**

Dafür benötigen wir die Parameter der Regressionsgeraden, also das k und das d, wobei wir den Anstieg k ja gerade im vorigen Beispiel (Aufgabe 10) ausgerechnet haben. Das d erhalten wir mit der Excel-Funktion `ACHSENABSCHNITT` und es ergibt sich insgesamt:

$$k = -0.020,\ d = 0.895$$

Der (statistische) Zusammenhang zwischen Temperatur (x) und Wahlbeteiligung (y) ist also gegeben durch

$$y = -0.020x + 0.895$$

d.h. bei $\underline{\underline{18°C}}$ wäre nach diesem Regressionsmodell eine Wahlbeteiligung von 54% zu erwarten gewesen. (Tatsächlich lag die Wahlbeteiligung bei der Bundespräsidentenwahl 2010 in Wien bei 52%).

12

In welchem mathematischen Zusammenhang stehen der Korrelationskoeffizient und der Regressionskoeffizient?

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \text{und} \quad k = \frac{s_{xy}}{s_x^2}$$

und somit:

$$k \cdot s_x^2 = r_{xy} \cdot s_x s_y$$
$$k = \frac{s_y}{s_x} \cdot r_{xy}$$

13

Besteht zwischen der in Aufgabe 10 gegebenen Wahlbeteiligung der Wienerinnen und Wiener und der am Wahltag vorherrschenden Temperatur eine hohe Korrelation? Wie groß ist der Korrelationskoeffizient? Für den Korrelationskoeffizienten verwenden wir die Excel-Funktion `KORREL` und erhalten:

$$\underline{\underline{r = -0.85}}$$

Dieser Wert deutet auf eine hohe negative Korrelation hin, und somit auf einen starken negativen linearen Zusammenhang.

14

Welche Werte kann ein Rangkorrelationskoeffizient annehmen?

Der Rangkorrelationskoeffizient kann (so wie der «normale» Korrelationskoeffizient) einen beliebigen Wert von -1 bis $+1$ annehmen.

15

In einem bestimmten Jahrgang wurden bei einer Analyse der Test- und Prüfungsergebnisse aus MAT101 (Mathematik) und MAT102 (Statistik) u.a. folgende Zusammenhänge beobachtet:

Korrelation zwischen den Gesamtpunkten aus MT122 und den im Vorsemester erreichten Punkten aus MAT101: $r = 0.37$ (Determinationskoeffizient: $r^2 = 14\%$).
Korrelation zwischen den aus den Online-Tests in MT122 erreichten Punkten und der Zeit, die im Durchschnitt für die Bearbeitung der Online-Tests aufgewandt wurde: $r = -0.08$ (Determinationskoeffizient: $r^2 = 1\%$).

Was lässt sich daraus über den Zusammenhang zwischen Mathematik- und Statistik-Kenntnissen bzw. den Zeitaufwand, den Studierende für die Online-Tests aufwenden, sagen?

Antwort:

Ein Korrelationskoeffizient von $r = 0.37$ bedeutet einen geringen bis mittleren positiven Zusammenhang zwischen den beiden Zufallsgrößen. D.h. tendenziell

werden zwar Studierende, die viele Punkte auf die Mathematikprüfung erhalten haben, auch in Statistik eher gut abschneiden, dieser Zusammenhang ist aber nicht besonders groß, sodass sich nicht von vornherein ein zuverlässige Prognose über das Abschneiden der einzelnen Studierenden in Statistik erstellen lässt, nur weil man ihre Mathematik-Punkte kennt.

Noch weniger ist ein solcher Zusammenhang aus der Zeit abzulesen, die die Studierenden für die Bearbeitung der Online-Tests aufbringen. Ein kurzer Zeitaufwand heißt nicht, dass die oder der Studierende so gut ist, dass sie oder er auch eine gute Note erhalten wird. Und umgekehrt: Länger am Test zu sitzen, garantiert auch keine gute Note.

Wahrscheinlichkeitsverteilungen

16

In Oberösterreich gibt es insgesamt 6 630 Orte. 40 davon heißen «Au». Angenommen alle 6 630 Orte werden auf Kärtchen geschrieben und daraus eine beliebige Karte herausgezogen. Wie groß ist die Wahrscheinlichkeit, dass darauf *Au* steht?

$$P = \frac{40}{6\,630} = 0.006 = 0.6\%$$

17

Unter der Annahme, dass beim Würfeln jede Augenzahl gleich wahrscheinlich ist: Wie groß ist die Wahrscheinlichkeit, eine gerade Zahl zu würfeln?

Es gibt 6 verschiedene Ausgangsmöglichkeiten (nämlich die Augenzahlen 1 bis 6), und drei Eintrittsfälle (nämlich die drei geraden Zahlen 2,4 und 6). Somit:

$$P(E) = \frac{\textit{Eintrittsfälle}}{\textit{Ausgangsmöglichkeiten}} = \frac{3}{6} = 0.5 = 50\%$$

18

Beim American Roulette gibt es je 18 rote und schwarze Nummernfelder sowie zwei grüne Felder mit einer Null. Gib an:

Wie groß ist die Wahrscheinlichkeit auf eine «rote» Zahl?

Es gibt insgesamt $18 + 18 + 2 = 38$ Felder, davon sind 18 rot, daher ist die Wahrscheinlichkeit für eine rote Zahl $P = \frac{18}{38} = \boxed{47.4\%}$.

Wie groß ist die Wahrscheinlichkeit auf eine Primzahl?

Unter den Zahlen von 0 bis 36 sind 11 Primzahlen (nämlich $2, 3, 5, 7, 11, 13, 17, 19, 23, 29$ und 31, daher ist $P = \frac{11}{38} = \boxed{28.9\%}$.

Wie groß ist die Wahrscheinlichkeit auf eine rote Primzahl?

Es gibt 4 rote Primzahlen (nämlich $3, 5, 7, 23$) und wir erhalten: $P = \frac{4}{38} = \boxed{10.5\%}$.

19

Ein Statistiktest besteht aus 6 Single-Choice Fragen mit jeweils 4 Antwortmöglichkeiten. (Single-Choice = genau eine Antwort aus den 4 möglichen ist

richtig). Für jemanden, der sich nicht auf den Test vorbereitet hat und nach Belieben zufällige Antworten ankreuzt, beträgt die Erfolgswahrscheinlichkeit pro Frage $p = 25\%$. Wie groß ist die Wahrscheinlichkeit, dass diese Person den Test positiv besteht, d.h. mindestens 3 Fragen richtig beantwortet?

«Mindestens drei Fragen richtig» bedeutet: Es können drei, vier, fünf oder sechs Fragen richtig sein.

Mit Formel 5.13 können wir zunächst die Gegenwahrscheinlichkeit $P(x \leq 2)$ ausrechnen, also die Wahrscheinlichkeit keine, eine oder zwei Fragen richtig zu haben. Sie beträgt 83% (aus EXCEL: `BINOM.VERT(2;6;0,25;WAHR)`) und somit:

$$1 - 0.83 = 0.17$$

d.h. die Wahrscheinlichkeit, den Test positiv zu bestehen, beträgt $\underline{\underline{17\%}}$

20

Was ist der wahrscheinlichste Wert der in Abb. 5.5 dargestellten Binomialverteilung?

Der Modalwert bzw. wahrscheinlichste Wert einer Verteilung ist derjenige, der am häufigsten vorkommt. In der in Abb.5.5 dargestellten Binomialverteilung ist das der Wert 50. Das sieht man auch in Tab.5.4, wo $x = 50$ mit $P(X = 50) \approx 8\%$ den höchsten Wahrscheinlichkeitswert hat.

21

Hat eine Gleichverteilung auch einen Modalwert?

Nein. Kennzeichen einer Gleichverteilung ist, dass allen Werten die gleiche Wahrscheinlichkeit zugeordnet wird. Es gibt also keinen, der «am wahrscheinlichsten» ist.

22

Am Bahnhof Meidling wird folgende Information durchgesagt: «Der Zug Richtung Wiener Neustadt hat zwischen 5 und 15 Minuten Verspätung». Angenommen es gibt keinen Grund zur Annahme, dass die Verspätung eher bei 5 oder 15 Minuten liegt, sondern der Zug tatsächlich «irgendwann» in diesem Intervall eintreffen wird: Wie groß ist die Wahrscheinlichkeit, dass die Verspätung maximal 7 Minuten ausmacht?

Ausgehend von einer Gleichverteilung mit $a = 5$ und $b = 15$ suchen wir also $P(X \leq 7)$, was nach Formel 5.20 den Wert

$$P(X \leq 7) = \frac{7-5}{15-5} = \frac{2}{10} = 0.2$$

ergibt. Die gesuchte Wahrscheinlichkeit beträgt also 20%.

23

Angenommen im Studiengang WIBA beträgt das (langjährig beobachtete) Durchschnittsalter der Studierenden $\bar{x} = 37$ Jahre mit einer Standardabweichung von $s = 5$ Jahren.

Wie groß ist die Wahrscheinlichkeit, dass unter der Annahme einer Normalverteilung ein:e (beliebig ausgewählte:r) Student:in zwischen 32 und 42 Jahre alt ist?

Für die Berechnung der gesuchten Wahrscheinlichkeit müssen wir entsprechend Formel 5.18) die Differenz bilden:

$$P\,(32 < x \leq 42) = F(42) - F(32)$$

und zwar für eine Normalverteilung mit dem Erwartungswert 37 und einer Standardabweichung 5.

Eingesetzt in EXCEL erhalten wir:

$$0.8413 - 0.1587 = 0.6827$$

Die Wahrscheinlichkeit, dass ein:e WIBA-Studierende:r zwischen 32 und 42 Jahre alt ist, beträgt also 68.3%.

Das stimmt im Übrigen auch mit der Aussage von Seite 121 überein, dass ca. 68% aller Realisierungen einer normalverteilten Zufallsgröße im Intervall $[\mu \pm 1 \cdot \sigma]$ liegen.

Konfidenzintervalle

24
Betrachte aus Tabelle 4.1 ausschließlich die Spalte «Gewicht» und nimm an, dass diese 20 zufällig ausgewählten Personen eine Stichprobe der Grundgesamtheit «Alle WIBA-Studierenden der FERNFH» sind.

Gib ein Intervall an, das mit 95%-iger Wahrscheinlichkeit den Erwartungswert für das durchschnittliche Gewicht der WIBA-Studierenden beinhaltet.

Wir können angeben:

$$\bar{x} = 77.75\,kg \quad s = 10.36\,kg$$

und für eine Irrtumswahrscheinlichkeit $\alpha = 0.05$ das Quantil der t-Verteilung (für $n = 20$) mit $t = 2.093$.

Die Grenzen des Konfidenzintervalls sind somit:

$$L = 77.75 - 2.093 \cdot \frac{10.36}{\sqrt{20}} = 72.9\,kg$$
$$U = 77.75 + 2.093 \cdot \frac{10.36}{\sqrt{20}} = 82.6\,kg$$

25
Gegeben sind für die Jahre 2005 bis 2014 die Anzahl polizeilich angezeigter Fälle in Österreich:

Jahr	*Anzeigen*	*Jahr*	*Anzeigen*
2005	604 229	2010	534 351
2006	588 229	2011	539 970
2007	592 636	2012	547 764
2008	570 952	2013	546 396
2009	589 961	2014	527 692

In welchen Jahren lag die Anzahl der Anzeigen außerhalb (über oder unter) des 95%-Konfidenzintervalls (bezogen auf den Durchschnitt der Jahre 2005-2014)?

Der Durchschnitt der Jahre 2005 bis 2014 beträgt 564 218, die Standardabweichung 28 078.36.

Für die Berechnung des Konfidenzintervalls können wir auch die Excel Funktion `=KONFIDENZ.T(0,05;28078,36;10)` verwenden. Wir erhalten den Wert 20 086.05, den wir nun einmal vom Mittelwert abziehen und einmal dazuzählen:

$L = 564\,218 - 20\,086.05 = 544\,131.95$
$U = 564\,218 + 20\,086.05 = 584\,304.05$

Somit lagen in den Jahren 2005, 2006, 2007 und 2009 die Anzahl der Anzeigen oberhalb der oberen und in den Jahren 2010, 2011 und 2014 unterhalb der unteren Grenze des Konfidenzintervalls.

26

Wir wollen mittels Umfrage herausfinden, mit wie viel Stunden Schlaf unsere Studierenden während des Semesters pro Nacht auskommen (müssen). Konkret wollen wir letztlich durch ein 90%-Konfidenzintervall den Erwartungswert der Schlafdauer auf eine Viertelstunde genau schätzen können. Methodisch bedienen wir uns dabei einer einfachen WhatsApp-Umfrage, d.h. wir ersuchen die Studierenden, uns ein WhatsApp mit den «Schlafstunden» der letzten Nacht zu schicken. Wir erwarten eine maximale Rücklaufquote von 20%, d.h. nur jede:r fünfte Studierende:r, die wir einladen mitzutun, wird uns eine WhatsApp-Nachricht zurückschicken.
Wie viele Personen sollten wir einladen, an der Untersuchung teilzunehmen?

Zunächst müssen wir schätzen, was denn vermutlich die maximale und minimale Schlafdauer sein wird, die uns genannt wird. Anhaltspunkt könnte die Einschätzung unserer eigenen Zeiten sein. Gehen wir einmal davon aus, dass wir in der Regel mindestens 5 und höchstens 11 Stunden pro Nacht schlafen, was einer Spannweite von 6 Stunden entspricht. Eingesetzt in Formel 6.8 ergibt das eine geschätzte Standardabweichung von $\hat{s} = 1\,h$.

Für ein Konfidenzniveau von 90% ist $z = 1.64$.

Eine Genauigkeit von einer Viertelstunde ergibt eine Intervalllänge von einer halben Stunde (= eine Viertelstunde auf oder ab).

Und somit:

$$n > \left(\frac{2 \cdot 1.64}{0.5} \cdot 1\right)^2 = 43.3$$

Im Grunde benötigen wir daher eine Stichprobe von mindestens 44 Antworten. Allerdings beträgt die zu erwartende Rücklaufquote nur 20%, d.h. wir müssen 44 noch mit 100 multiplizieren und durch 20 dividieren und erhalten somit die geforderte Anzahl von 220 Personen, die wir befragen sollten.

27

Welcher Anteil $\hat{p}$ aus einer Stichprobe mit $n = 183$ Personen müsste sich für

einen Gesetzesvorschlag aussprechen, damit wir mit 95%iger Sicherheit darauf schließen können, dass eine (einfache) Mehrheit der Grundgesamtheit für den Beschluss dieses Gesetzes ist?

Die Fragestellung könnte auch so formuliert werden: Wie groß muss $\hat{p}$ sein, damit die untere Grenze des Konfidenzintervalls für p mindestens 50 Prozent beträgt, also $L > 50\%$? Dazu müssen wir die Ungleichung

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{183}} > 0.5$$

nach $\hat{p}$ auflösen:

$$\begin{aligned}
&\hat{p} - 1.96\sqrt{\tfrac{\hat{p}(1-\hat{p})}{183}} > 0.5 \\
&\hat{p} - 0.5 > 1.96\sqrt{\tfrac{\hat{p}(1-\hat{p})}{183}} \\
&(\hat{p} - 0.5)^2 > 1.96^2 \left(\tfrac{\hat{p}(1-\hat{p})}{183}\right) \\
&\hat{p}^2 - \hat{p} + 0.25 > 3.84\tfrac{\hat{p}-\hat{p}^2}{183} \\
&183\hat{p}^2 - 183\hat{p} + 45.75 > 3.84\hat{p} - 3.84\hat{p}^2 \\
&(183 + 3.84)\,\hat{p}^2 - (183 + 3.84)\,\hat{p} + 45.75 > 0 \\
&186.84\hat{p}^2 - 186.84\hat{p} + 45.75 > 0
\end{aligned}$$

Um eine Ungleichung zu lösen, müssen wir zunächst die zugehörige Gleichung auflösen:

$$\begin{aligned}
&186.84\hat{p}^2 - 186.84\hat{p} + 45.75 = 0 \\
&{}_1\hat{p}_2 = \tfrac{186.84 \pm \sqrt{186.84^2 - 4 \cdot 186.84 \cdot 45.75}}{2 \cdot 186.84} \\
&\hat{p}_1 = 0.572 \quad \hat{p}_2 = 0.428
\end{aligned}$$

und dann in die Ungleichung einsetzen:

$$(\hat{p} - 0.572)\,(\hat{p} - 0.428) > 0$$

Ein Produkt ist bekanntlich dann positiv, wenn beide Faktoren positiv oder beide Faktoren negativ sind. Dass beide negativ sind, kann aber ausgeschlossen werden. (Warum? Weil das gesuchte $\hat{p}$ sicher größer als 0.5 ist und somit $(\hat{p} - 0.428)$ jedenfalls positiv ist). Es verbleibt also:

$$\begin{aligned}
&(\hat{p} - 0.572 > 0) \wedge (\hat{p} - 0.428 > 0) \\
&(\hat{p} > 0.572) \wedge (\hat{p} > 0.428)
\end{aligned}$$

und somit erhalten wir das Ergebnis: $\hat{p} > 0.572$, d.h. bei einer Stichprobe mit 183 Elementen bedarf es eines empirischen Anteilswerters von mindestens 57.2%, damit wir uns 95%ig sicher sein können, dass die dahinterliegende Grundgesamtheit einen theoretischen Anteilswert von mindestens 50% hat.

Statistische Tests

28

Deine Ärztin empfiehlt dir, aus gesundheitlichen Gründen deinen Kaffeekonsum auf fünf Tassen pro Tag einzuschränken. Du bist dir eigentlich sicher, dass du das im Schnitt ohnehin nicht überschreitest ($H_0 : \mu \leq 5$), dein besorgter Ehepartner schenkt dem aber keinen so rechten Glauben und behauptet, dass es mehr als fünf Tassen pro Tag sind ($H_A : \mu > 5$). Ihr vereinbart, eine Zeitlang darüber Buch zu führen. Nach 40 Tagen hast du 210 Tassen Kaffee getrunken. Im Schnitt ergab das Experiment also 5,25 Tassen pro Tag, und das bei einer Standardabweichung von 0,25. **Wessen Hypothese kann bei diesem Ergebnis aufrecht erhalten werden?**

Die Testfunktion ergibt:

$$f = \frac{5.25 - 5}{0.25}\sqrt{40} = 6.32$$

und das Quantil der Normalverteilung:

$$z_{1-0.05} = 1.64$$

Nachdem $f > z_{1-\alpha}$ müssen wir die Nullhypothese ablehnen und H_A akzeptieren – und unseren Kaffeekonsum in Zukunft wohl reduzieren. Er ist signifikant höher als die empfohlene Menge von 5 Tassen.

29

Eine Software-Entwicklerin überlegt, ihre Software als *Donationware* zur Verfügung stellen, d.h. sie kann grundsätzlich kostenlos verwendet werden, sie bittet aber um (freie) Spenden, damit auf Sicht wenigstens die bei ihr entstehenden Drittkosten abgedeckt sind. Sie schätzt: Nur wenn mindestens 25% der User bereit sind, 10 € zu spenden, werde ich kein Geld verlieren.
Sie startet einmal probehalber und stellt nach den ersten 500 Downloads fest: 145 User haben tatsächlich 10 € gespendet. Das sind sogar 29%. **Kann sie (aus statistischer Sicht) optimistisch sein, dass der Anteil der spendenfreudigen User tatsächlich größer als 25% ist?**

Wir wählen die Arbeitshypothese, dass der Anteil p größer als 25% ist, also:

$$H_A : p > 0.25 \rightarrow H_0 : p \leq 0.25$$

Aus den ersten 500 Downloads erhalten wir einen Schätzwert für p:

$$\hat{p} = \frac{500}{145} = 0.29$$

und damit die Prüfgröße:

$$f = \frac{0.29 - 0.25}{\sqrt{0.25\,(1 - 0.25)}}\sqrt{500} = 2.066$$

Der kritische Bereich beginnt bei 1.645 (Excel: `=NORM.S.INV(1-0.05)`). Mit 2.066 liegen wir da bereits drüber, d.h. wir lehnen H_0 ab und akzeptieren H_A, also die Hypothese, dass auch in der Grundgesamtheit der Anteil der User, die 10 € spenden, über 25% ist.

30

Nach dem ersten Studienjahr wurde für alle Studierenden eines Jahrgangs ein nach ECTS gewichteter Notenschnitt (GPA) berechnet und daraus dann ein Durchschnittswert für jeweils alle männlichen ($n_m = 84$) und alle weiblichen ($n_w = 36$) Studierenden angegeben. Für die männlichen Studierenden beträgt er $\bar{x}_m = 1.6$, für die weiblichen $\bar{x}_w = 1.4$. Ermittelt wurde auch die zugehörigen empirischen Standardabweichungen: $s_m = 0.6$, $s_w = 0.4$.

Der empirisch ermittelte GPA der Frauen unterscheidet sich offenbar von dem der Männer. Ist dieser Unterschied signifikant bzw. inwiefern lässt sich aus den obigen Daten die Hypothese $H_A : \mu_m - \mu_w \neq 0$ behaupten?

Das zu untersuchende Hypothesenpaar lautet: $H_A : \mu_m \neq \mu_w \rightarrow H_0 : \mu_m = \mu_w$.

Einsetzen der Werte aus den beiden Stichproben ergibt die Testfunktion:

$$f = \frac{1.6 - 1.4}{\sqrt{0.6^2 \cdot 36 + 0.4^2 \cdot 84}}\sqrt{84 \cdot 36} = 2.14$$

Dieser Wert muss bei einem α von 0.05 mit dem Wert $z_{1-\alpha/2} = 1.96$ verglichen werden und ergibt, da $|f| > z_{(1-\alpha/2)}$, die Ablehnung der Nullhypothese und Akzeptanz der Arbeitshypothese. D.h. der Unterschied zwischen dem von den weiblichen und den von den männlichen erzielte GPA ist als signifikant einzustufen.