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.

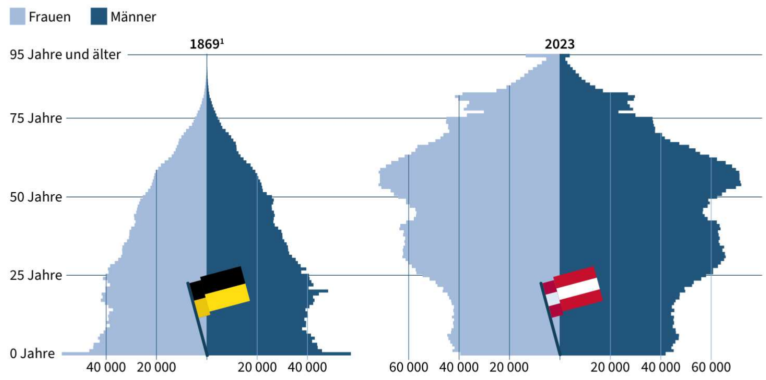

Abb. 1.2: Bevölkerungspyramide Österreichs 1869 und 2023. (Quelle: STATISTIK AUSTRIA: Infografiken 2023)
[1)] Die Bevölkerungszahlen 1869 beinhalten nur Daten aus dem heutigen Gebietsstand des Bundesgebiets

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*.

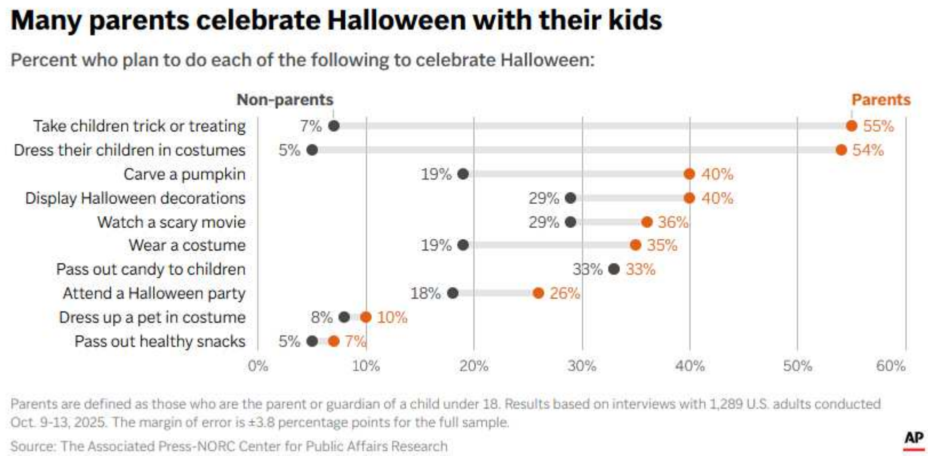

Abb. 1.3: Nicht nur Kinder feiern Halloween. Das Ergebnis einer statistischen Auswertung kann allerdings unterschiedlich ausfallen, je nachdem, ob man einfach alle Antworten zusammenzählt, oder gruppenweise auswertet. (Hier: Sind die Erwachsenen Eltern oder nicht?). Auch das sieht man mit statistischen 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. *εμπειρως* – 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

Postleitzahlen 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(alfang) 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.

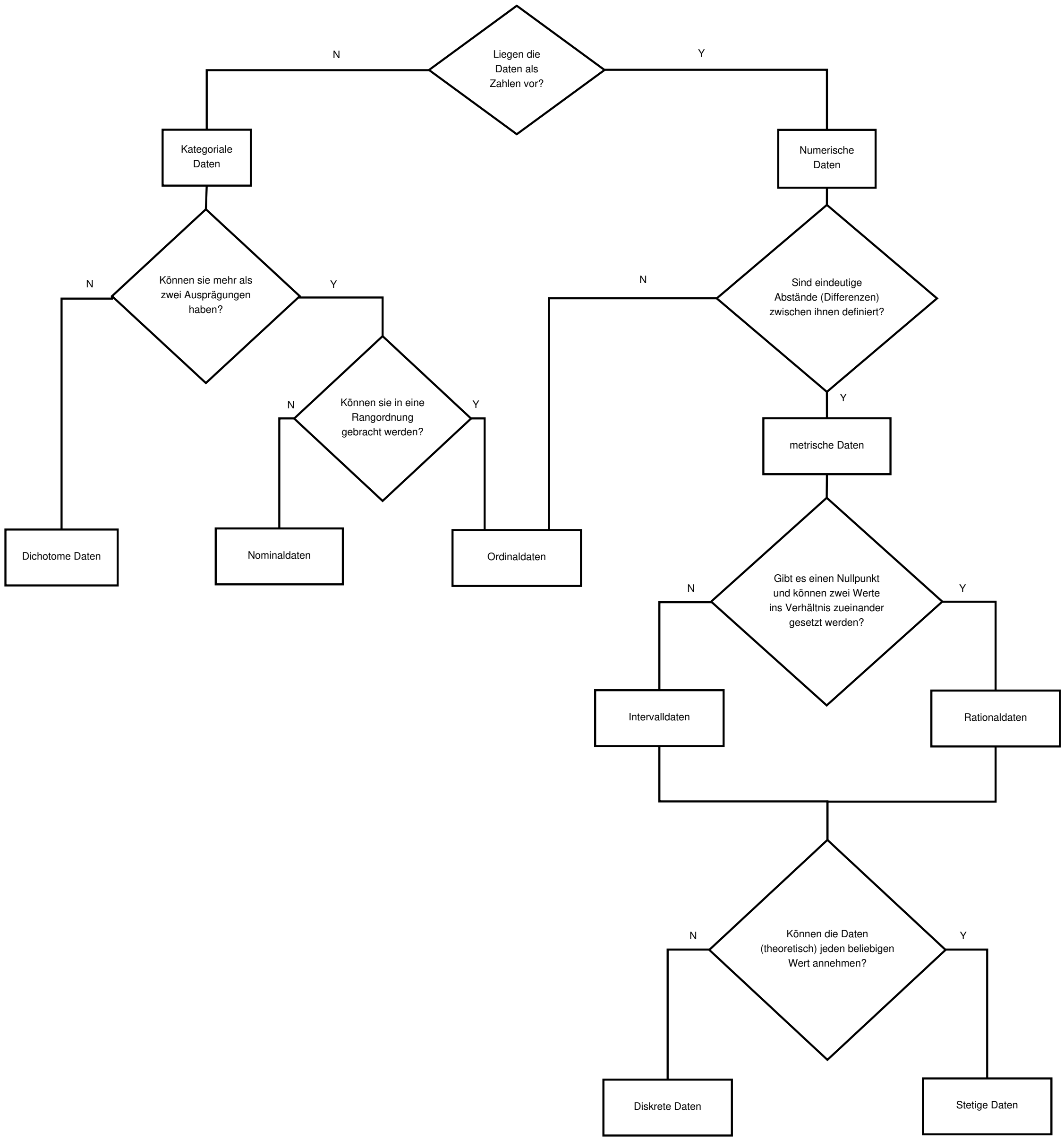

Abb. 1.4: Verschiedene Datentypen können auf verschiedenen Skalen gemessen werden

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.

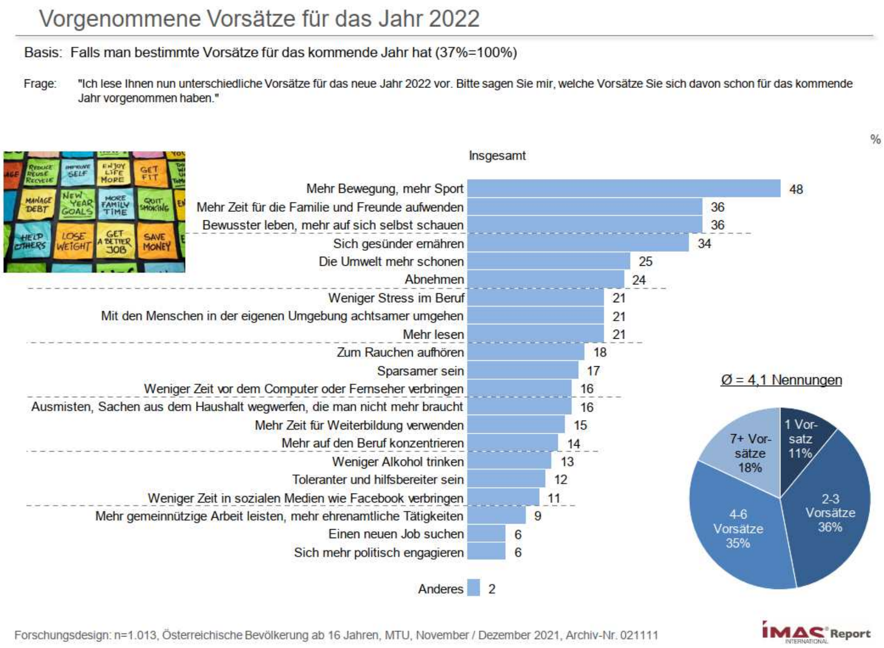

Abb. 1.5: Was sich die «Österreichische Bevölkerung ab 16 Jahren» für 2022 alles vorgenommen hat. Sample: n=1.013 Personen, statistisch repräsentativ für die österreichische Bevölkerung ab 16 Jahren. (Quelle: IMAS International Eigenstudie. Report Nr. 12/2021)

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.