Kapitel 1.

Einführung

> It is easy to lie with statistics. It is hard to tell the truth without statistics.
>
> *(Andrejs Dunkels)*

1.1. Was ist Statistik?

Das Wort *Wahrscheinlichkeit* taucht in der Alltagssprache häufig auf. Hier einige Beispiele:

- Wir hören im Wetterbericht: „Die Wahrscheinlichkeit, dass es heute morgen regnet, liegt bei 60 Prozent“.
- Weiter hört man: „Die Wahrscheinlichkeit, dass ich hundert Jahre alt werde, ist klein“.
- In Basel möchte ein Seismologe bestimmen, wie gross die Wahrscheinlichkeit ist, dass Geothermie-Bohrungen ein Erdbeben von einer bestimmten Grössenordnung auslösen.
- Ein Atomphysiker stellt sich andererseits die Frage: „Wie gross ist die Wahrscheinlichkeit, dass ein Geiger-Zähler in den nächsten 10 Sekunden 20 Zerfälle registriert?“.
- Ein Schweizer Politiker oder Nationalbanker interessiert sich momentan wohl für die Frage: „Wie gross ist die Wahrscheinlichkeit, dass der Wert vom Euro in diesem Jahr wieder über 1.20 Franken steigt?“
- Oder: „Wie gross ist die Wahrscheinlichkeit, dass es einen Börsencrash gibt?“

Wahrscheinlichkeiten geben wir im Zusammenhang mit Vermutungen an. Warum stellen wir Vermutungen an? Wir stellen Vermutungen an, wenn wir eine Aussage oder Vorhersage machen möchten, aber dazu nur über unvollständige Informationen oder

unsichere Kenntnisse verfügen. Wir stellen auch Vermutungen an, weil wir eine Entscheidung fällen möchten: „Soll ich heute morgen einen Regenschirm mitnehmen?“ „Soll ich mich bei einer Bank bewerben, oder selbstversorgender Bio-Bauer werden?“

In den Naturwissenschaften möchten wir mit Hilfe unserer beschränkten oder unvollständigen Kenntnissen ein physikalisches System so allgemein wie möglich beschreiben. Die Beschreibung eines physikalischen Systems stellt aber letztlich nichts anderes als eine Vermutung (Modell) dar, denn wir können ein (realistisches) physikalisches System niemals bis ins letzte Detail beschreiben. Nun gibt es bessere Vermutungen und schlechtere Vermutungen, wie ein physikalisches System beschaffen ist. Die *Stochastik* hilft uns dabei, bessere Vermutungen anzustellen.

Betrachten wir das Beispiel einer Münze: Wir möchten vorhersagen, ob ein Münzwurf das Ergebnis „Kopf“ oder „Zahl“ ergibt. Wüssten wir die genaue Massenverteilung der Münze, die genaue Anfangsgeschwindigkeit und Anfangsposition der Münze und die Positionen und Geschwindigkeiten aller Luftmoleküle zu jedem Zeitpunkt während des Wurfs, könnten wir wohl mit Hilfe der Mechanik vorhersagen, ob der Münzwurf mit Kopf oder Zahl auf dem Boden landet. Nun verfügen wir in der Praxis nie über alle diese Informationen. Aufgrund unserer Unkenntnis stellen wir die Vermutung an, dass die Massenverteilung der Münze dergestalt ist, dass wir diese als fair bezeichnen, d.h., die Anzahl Würfe mit „Kopf“ ist in etwa gleich der Anzahl Würfe mit „Zahl“. Je nach dem, wie stark sich die Anzahl Würfe mit „Kopf“ von der Anzahl Würfe mit „Zahl“ unterscheidet, können wir mit Hilfe der Stochastik aussagen, wie gut unsere Beobachtung mit der Vermutung zusammenpasst, dass die Münze fair ist, und ob wir an unserer Vermutung (dass die Münze fair ist) festhalten sollten.

Auch in der kinetischen Gastheorie, wo wir es mit der Grössenordnung von 10^{22} Gasmolekülen zu tun haben, können wir im besten Fall aussagen, wie wahrscheinlich es ist, dass ein Gasmolekül bei einer bestimmten Temperatur eine Geschwindigkeit in einem bestimmten Intervall hat. Denn es ist nicht realisierbar, jedem Molekül eine genaue Position und Geschwindigkeit zuzuordnen. Dies hat nicht nur mit der Komplexität des Problems zu tun; wir wissen mittlerweile, dass die Quantenmechanik verbietet, die genaue Position und Geschwindigkeit eines Atoms gleichzeitig zu bestimmen (*Heisenbergsche Unschärferelation*). Das Konzept von Wahrscheinlichkeiten ist essentiell, um das atomare Geschehen zu beschreiben.

Stochastik ist ein Teilgebiet der Mathematik und fasst als Oberbegriff die Gebiete Wahrscheinlichkeitsrechnung und Statistik zusammen. In der *Wahrscheinlichkeitsrechnung* geht man von einem Modell aus (man beschreibt einen sogenannten datengenerierenden Prozess) und leitet daraus entsprechende Eigenschaften ab. Wie in Abbildung 1.1 dargestellt, kann man sich unter einem Modell symbolisch eine Urne vorstellen, aus der man Kugeln (Daten) zieht. Zum Beispiel können wir uns die Frage stellen: „Wie gross ist die Wahrscheinlichkeit, eine rote Kugel zu ziehen?“ Diese Frage können wir beantworten, wenn wir wissen, wie viele rote und blaue Kugeln in

der Urne sind. Hat es drei rote und fünf blaue Kugeln in der Urne, so beträgt die Wahrscheinlichkeit, zufällig eine rote Kugel zu ziehen, $^3/_8$.

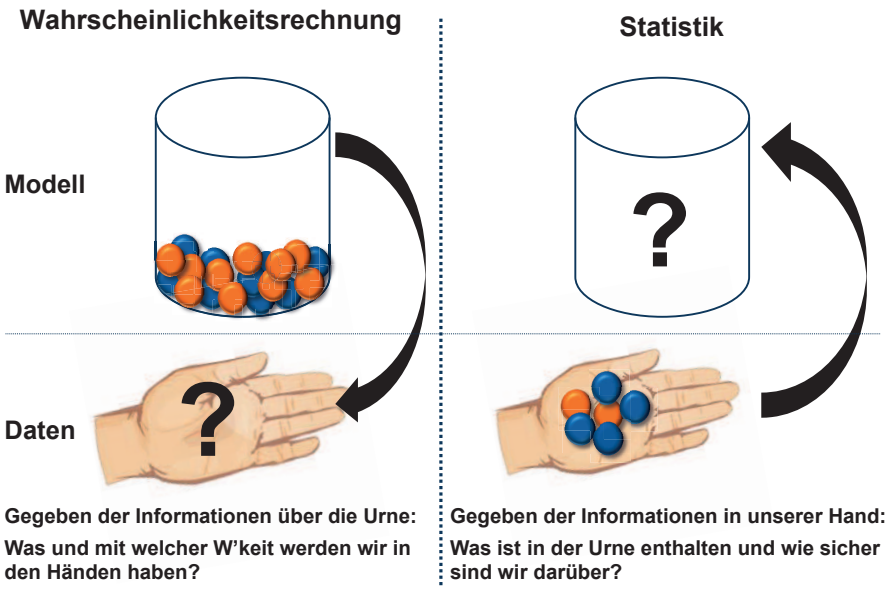

Abbildung 1.1.: Darstellung der Konzepte der Wahrscheinlichkeitsrechnung und der Statistik. Das Modell wird hier durch eine Urne symbolisiert

Betrachten wir ein anwendungsrelevantes Beispiel: Eine Gemeinde will für einen Bach einen Damm bauen, da dieser Bach oft über die Ufer tritt. Die Frage ist nun, wie hoch der Damm gebaut werden soll. Ein zu niedriger Damm ist zwar kostengünstig, dafür werden die Ausgaben für die Überschwemmungen sehr hoch. Auf der anderen Seite ist ein zu hoher Damm sehr teuer. Wie können wir nun versuchen, eine „gute" Dammhöhe zu ermitteln? Eine „gute" Dammhöhe zeichnet sich dadurch aus, dass der Damm genügend Sicherheit bietet, aber gleichzeitig auch noch finanzierbar ist. Hierzu müssen wir die Unsicherheit in Bezug auf den jährlichen maximalen Wasserstand quantifizieren können (z.B. in einer 30-Jahr Periode). Dazu stützen wir uns auf die Wahrscheinlichkeitsrechnung. Mit Hilfe der Wahrscheinlichkeitsrechnung können wir zum Beispiel die Wahrscheinlichkeit berechnen, dass in einer 30-Jahr Periode der maximale Wasserstand gewisse Höhen überschreitet. Oder wir können die zu erwartenden Kosten in einer 30-Jahr Periode aufgrund von Überschwemmungen bei einer gegebenen Dammhöhe berechnen. Solche Zahlen ermöglichen dann eine sinnvolle Kosten-Nutzen-Rechnung für den Bau eines Damms.

In der *Statistik* geht es darum, aus vorhandenen Daten auf das Modell zu schliessen, mit dem wir dann wieder Vorhersagen machen können. Wir denken also gerade „in die andere Richtung". Ist in Abbildung 1.1 rechts die Anzahl der blauen und roten Kugeln in der Urne unbekannt, so können wir zum Beispiel 100-mal eine Kugel ziehen, die nach jeder Ziehung wieder zurückgelegt wird. Ziehen wir 40 rote Kugeln,

so können wir *vermuten*, dass in der Urne 40 % der Kugeln rot sind. Diese Vermutung können wir überprüfen, indem wir weitere 100 Kugeln ziehen. Stimmt das Resultat mehr oder weniger mit dem ersten Versuch überein, so halten wir an unserer Vermutung fest. Ziehen wir aber 60 rote Kugeln, so müssen wir unsere Vermutung überprüfen. Die Statistik hilft uns, quantitative Aussagen zu machen, wie gut eine Beobachtung mit einer solchen Vermutung (Modell) zusammenpasst.

Für unser Dammbeispiel sehen die entsprechenden Überlegungen wie folgt aus: Wir analysieren Datenpunkte (z.B. jährliche Wasserstandsmessungen in den letzten 50 Jahren) und versuchen mit diesem beschränkten Wissen herauszufinden, was wohl ein gutes Modell für die „wahren" jährlichen Höchststände des Wasserpegels ist. Das Modell sollte Vorhersagen über die höchstmöglichen Wasserstände machen können. Diese Vorhersagen werden, wie oben erwähnt, in der Regel in Wahrscheinlichkeiten angegeben. Ein solches Wahrscheinlichkeitsmodell ist dann die Grundlage, aufgrund derer wir die optimale Dammhöhe bestimmen werden.

In der Statistik können wir zusätzlich auch Angaben darüber machen, wie plausibel ein Modell aufgrund von Beobachtungen ist (was auf den ersten Blick erstaunlich erscheint). Werfen wir eine Münze 100-mal und erhalten 70-mal „Kopf", können wir dann immer noch behaupten, dass die Münze fair ist? Theoretisch müssten wir 50-mal „Kopf" erhalten, aber es wäre ja möglich, dass zufälligerweise 70-mal „Kopf" geworfen wurde. Wann können wir den Zufall ausschliessen? Solche Fragen können wir mit der Statistik beantworten.

1.2. Kann ich Statistik überhaupt brauchen?

Die Statistik hat ihren Ursprung in der Mathematik, greift aber in viele Bereiche der modernen Wissenschaften über. Grosse Teile der Biologie, der Medizin, der Ingenieurwissenschaften und der Umweltforschung wären heute ohne Statistik undenkbar. Der Chefökonom von Google, Hal Varian, sagte vor einigen Jahren: Der sexieste Beruf des kommenden Jahrzehnts ist der des Statistikers. In der Vergangenheit bis heute stammen viele Anregungen und Problemstellungen für die Statistik aus den Bereichen Biologie und Pharmazie.

Eine häufige Frage bei Tier und Mensch lautet: wie bestätigt man die Wirksamkeit eines neuen Medikamentes. Dabei erhält eine zufällig ausgewählte Gruppe (Medikamentengruppe) von Patienten das neue Medikament in Form einer Tablette. Eine andere zufällig ausgewählte Gruppe von Patienten (Kontrollgruppe) erhält ein Placebo[1]. Die Medikamentengruppe hat gegenüber der Kontrollgruppe nach zwei Wochen eine deutliche Verbesserung der Symptome gezeigt. Das Medikament wirkt also. Oder doch nicht? Kann es sein, dass das Medikament gar nicht wirkt und alle

[1] Tablette mit gleichem Aussehen und Geschmack wie das Medikament aber ohne Wirkstoff.

Personen der Medikamentengruppe unabhängig vom Medikament *zufällig* eine Verbesserung der Symptome hatte? Wie können wir „Zufall“ ausschliessen? Mit Statistik lässt sich diese Frage beantworten.

Oder in der Genetik stellt sich zum Beispiel folgendes Problem: Man beobachtet, dass in einer Gruppe von Krebspatienten gewisse Gene stärker aktiv sind als in einer Kontrollgruppe. Könnte es Zufall sein, dass alle Personen, bei denen diese Gene aktiver sind in derselben Gruppe gelandet sind? Hat also das aktive Gen gar nichts mit der Krebserkrankung zu tun, obwohl es bei allen Krebspatienten vorkommt? Wie können wir hier mathematisch Zufall ausschliessen?

Aber auch die Ingenieurswissenschaften liefern sehr interessante Aufgabenstellungen für die Statistik. Ein typisches Beispiel, das vielen Leuten gar nicht bekannt sein dürfte: Spracherkennungsprogramme oder Programme, mit denen Roboter visuell ihre Umwelt erkennen können – Stichwort „Machine Learning“ –, funktionieren nur mit Statistik. Auch da geht es um den Umgang mit Unsicherheiten und Variabilität. Eine Silbe oder ein Wort wird von jeder Person leicht anders ausgesprochen. Ein gutes System muss trotz diesen Variationen ein Wort oder einen Text erkennen können. Und ein Roboter wird kaum zweimal die genau gleiche Situation antreffen. Trotzdem muss er entscheiden können, ob eine angetroffene Situation einer gespeicherten Standardsituation ähnlich ist.

Diese Vorlesung soll Ihnen helfen, ein Fundament zu legen und die Grundbegriffe in der Statistik verstehen und anwenden zu können.

1.3. Was ist der Inhalt dieses Kurses?

In Kapitel 2 werden wir uns mit den sehr wahrscheinlich bereits bekannten Grössen der Statistik wie arithmetisches Mittel, Standardabweichung, Varianz, Quantil, Median und Korrelation auseinandersetzen. Dabei werden wir lernen, wie Daten mit der Statistiksoftware R graphisch dargestellt werden können.

In Kapitel 3 geht es zunächst darum, den Begriff „Zufall“ mathematisch genau zu definieren. Wenn das geschehen ist, kann man mit Wahrscheinlichkeiten einfach rechnen. Wir werden zunächst nur Zähldaten behandeln, also Situationen, in denen nur ganze Zahlen auftreten (z.B. die Anzahl Gewinne, wenn man 20 Lose kauft). Wir werden die Begriffe Unabhängigkeit, bedingte Wahrscheinlichkeit, Zufallsvariable, Erwartungswert und Standardabweichung kennenlernen. Ausserdem werden wir verschiedene Arten von Zufall gruppieren und dabei den Begriff der Verteilung kennenlernen.

In Kapitel 4 lernen wir die drei Grundfragen der Statistik - Punktschätzung, Hypothesentest, Vertrauensintervall - anhand von Zähldaten kennen. Damit können wir folgende generische Frage beantworten: Angenommen wir haben ein Medikament

an 100 Testpersonen ausprobiert. 67 Personen wurden gesund. Bei welchem Anteil der Gesamtbevölkerung wirkt das Medikament? In welchem Bereich liegt wohl der wahre Anteil mit grosser Wahrscheinlichkeit (z.B. mit 95%-Wahrscheinlichkeit zwischen 60% und 70%)? Der **Binomialtest** wird uns hier eine Antwort liefern.

In Kapitel 5 werden wir kontinuierliche Wahrscheinlichkeitsverteilungen kennenlernen, die wir zur Modellierung von kontinuierlichen, reellen Daten (z.B. Grösse, Gewicht, etc.) benötigen. Wir werden uns mit der berühmtesten Wahrscheinlichkeitsverteilung, der sogenannten **Normalverteilung** (auch bekannt als Gauss'sche Verteilung), eingehend beschäftigen. Dabei interessiert uns auch, wie der arithmetische Mittelwert einer Messreihe verteilt ist. Denn bei der Angabe des Mittelwertes geben wir in der Regel auch den Messfehler an (entweder den absoluten oder relativen Fehler).

In Kapitel 6 erweitern wir die statistischen Methoden, die wir in Kapitel 4 kennengelernt haben, auf kontinuierliche, reelle Daten. Angenommen, wir haben neue Augentropfen entwickelt, die den Augeninnendruck senken sollen. Wir wählen zufällig 20 Testpersonen und teilen sie zufällig in zwei Gruppen mit je 10 Personen auf. Gruppe N wird mit den neuen Augentropfen behandelt, die andere Gruppe A mit herkömmlichen Augentropfen. In Gruppe N scheint der Augeninnendruck stärker zu sinken als in der Gruppe A. Die zentrale Frage, die wir in diesem Kapitel beantworten, ist folgende: Könnte es sein, dass beide Medikamente gleich gut wirken, aber die Personen, die besonders gut auf Augentropfen ansprechen, zufällig in Gruppe N zusammengefasst wurden? Der **t-Test** und der **Wilcoxon-Test** werden uns hier eine Antwort liefern.

In Kapitel 7 beschäftigen uns **stochastische Prozesse** wie die Brownsche Bewegung oder weisses Rauschen. Diese finden häufige Anwendung in den Ingenieur- und Naturwissenschaften, z.B. für die Beschreibung von thermischem Rauschen an elektrischen Widerständen oder für die Erkennung eines deterministischen Signals in einem verrauschten Signal.

1.4. Software

Wir werden die Statistiksoftware `R` verwenden, insbesondere `R-Studio`. Die Beherrschung von `R` ist wesentlicher Bestandteil dieses Stochastik Moduls und auch prüfungsrelevant. `R-Studio` können Sie kostenlos über folgenden Link beziehen und auf Ihrem Laptop installieren. Unter Docs finden Sie auch eine sehr gut gestaltete Benutzungsanleitung zu `R-Studio`, die wir sehr empfehlen können. Andere weit verbreitete Statistikprogramme sind SPSS und SAS. Alle Methoden, die wir in diesem Kurs besprechen, sind in jeder Statistiksoftware implementiert. In der ersten Unterrichtseinheit wird es eine Einführung in die Benützung von `R` geben. [Dal08] ist ein

hervorragendes Buch, das auf Ilias abgelegt wird und das Sie sowohl als Nachschlagewerk zur Benützung von R als auch zu allen in der Vorlesung behandelten Themen verwenden können.

1.5. Literaturverzeichnis

[Dal08] P. Dalgaard. *Introductory Statistics with R*. Statistics and Computing. Springer, 2008.

[Kal] M. Kalisch. Statistik für Biologie und Pharmazeutische Wissenschaften. Vorlesungsskript.

[MR14] D.C. Montgomery and G.C. Runger. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2014.

[Pap06] Lothar Papula. *Mathematische Formelsammlung für Ingenieure und Naturwissenschaftler: mit zahlreichen Rechenbeispielen und einer ausführlichen Integraltafel*. Lothar Papula. Vieweg, 2006.

[Pap08] L. Papula. *Mathematik für Ingenieure und Naturwissenschaftler 3: Vektoranalysis, Wahrscheinlichkeitsrechnung, Mathematische Statistik, Fehler- und Ausgleichsrechnung*. Viewegs Fachbücher der Technik. Vieweg + Teubner, 2008.

[Ric06] J. Rice. *Mathematical Statistics and Data Analysis*. Number S. 3 in Advanced series. Cengage Learning, 2006.

Dieses Vorlesungsskript beruht auf dem von Markus Kalisch, Peter Bühlmann und Hansruedi Künsch verfassten und an der ETH Zürich verwendeten Skript Statistik für Biologie und Pharmazeutische Wissenschaften [Kal].

Wir empfehlen das bewährte Lehrbuch von Lothar Papula [Pap08].

Als Ergänzung oder als ausführlichere Alternative zum Skript empfehlen wir für besonders Interessierte die im angelsächsischen Raum verbreiteten und in englischer Sprache geschriebenen Lehrbücher [MR14] und [Ric06]. Zusätzlich empfehlen wir die Formelsammlung [Pap06], die auch an der Modulendprüfung benützt werden darf.

Kapitel 2.

Deskriptive Statistik

> Definition of Statistics: The science of producing unreliable facts from reliable figures.
>
> *(Evan Esar)*

2.1. Deskriptive Statistik eindimensionaler Daten

Die deskriptive Statistik befasst sich mit der Darstellung von Datensätzen (Zusammenstellung verschiedener Daten). Dabei werden diese Datensätze durch gewisse Zahlen charakterisiert (zum Beispiel den Mittelwert) und graphisch in einem Koordinatensystem dargestellt. Wir befassen uns zunächst mit *eindimensionalen* Daten, wo *eine* Messgrösse an einem Untersuchungsobjekt ermittelt wird. Anhand des folgenden Beispieles werden wir die wichtigen Begriffe und Vorgehensweisen genauer kennenlernen.

2.1.1. Messungen der Schmelzwärme von Eis

Als Einführungsbeispiel betrachten wir zwei *Datensätze*, bei welchen zwei Methoden zur Bestimmung der latenten Schmelzwärme von Eis verglichen werden. Wiederholte Messungen der freigesetzten Wärme beim Übergang von Eis bei $-0.7\,^{\circ}\mathrm{C}$ zu Wasser bei $0\,^{\circ}\mathrm{C}$ ergaben die Werte (in $\mathrm{cal/g}$), die in Tabelle 2.1 aufgeführt sind. Obwohl die Messungen mit der grösstmöglichen Sorgfalt durchgeführt und alle Störeinflüsse ausgeschaltet wurden, variieren die Messwerte innerhalb beider Methoden. Es stellen sich hier nun die folgenden Fragen:

- Gibt es einen Unterschied zwischen der Methode A und der Methode B?
- Falls ja, wie können wir diesen Unterschied ermitteln?

Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Methode A	80.03	80.02	80.00	80.02					
Methode B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

Tabelle 2.1.: Messungen zur Bestimmung der latenten Schmelzwärme von Eis anhand von zwei Methoden.

Es fällt auf, dass bei beiden Methoden die Messwerte um 80 herum liegen. Bei Methode A liegen aber nur 2 Werte von 13 *unter* 80, während bei der Methode B nur 2 von 8 Werten *über* 80 liegen. Die Werte der Methode A sind also eher grösser als die der Methode B. Was heisst hier aber „eher"? Es ist also von Interesse, die Messreihen irgendwie so zusammenzufassen, dass wir die beiden Methoden miteinander vergleichen können.

Die *deskriptive Statistik* beschäftigt sich damit, auf welche Weisen (quantitative) Daten organisiert und zusammengefasst werden können. Dies hat zum Ziel, dass die Interpretation und darauffolgende statistische Analyse dieser Daten vereinfacht werden. Wir machen dies mit Hilfe von

- graphischen Darstellungen
- Zusammenfassungen von Daten, die die wichtigen Merkmale der Daten hervorheben sollen, wie eben zum Beispiel die mittlere Lage der Messwerte und die Streuung dieser Messwerte um die mittlere Lage.

Diese sogenannten *Kennzahlen* sollen die Daten numerisch zusammenfassen und grob charakterisieren.

Bei statistischen Analysen, wie wir sie im Laufe der Vorlesung kennenlernen werden, ist es ausserordentlich wichtig, nicht einfach blind ein Modell anzupassen oder ein statistisches Verfahren anzuwenden. Die Daten sollten immer mit Hilfe von geeigneten graphischen Mitteln *und* den Kennzahlen dargestellt werden, da man nur auf diese Weise (teils unerwartete) Strukturen und Besonderheiten entdecken kann.

Im Folgenden werden die Daten mit $x_1, \ldots, x_n$ bezeichnet, wobei n der *Umfang* der Messreihe genannt wird. Im Fall der Messreihe der Methode A ist dies für $n = 13$:

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \ldots, \quad x_{13} = 80.02$$

2.1.2. Darstellung von Messwerten

Bevor wir uns mit Kennzahlen und graphischen Darstellungen von Datensätzen auseinandersetzen, müssen wir Regeln für die Darstellung von Messwerten festlegen. Dazu benötigen wir die Begriffe *Nachkommastellen* und *signifikante Stellen*.

Als **Nachkommastellen** werden die in der dezimalen Darstellung einer Zahl verwendeten Ziffern rechts des Kommas bezeichnet. Im obigen Beispiel haben die Messpunkte

$$x_1 = 79.98, \quad x_2 = 80.04, \quad \ldots, \quad x_{13} = 80.02$$

zwei Nachkommastellen.

Die **signifikanten Stellen** werden als die erste von Null verschiedene Stelle bis zur Rundungsstelle definiert. Die Rundungsstelle ist die letzte Stelle, die nach dem Runden noch angegeben werden kann. Im obigen Beispiel haben wir also **vier** signifikante Stellen.

Beispiel 2.1.1

Zahl	Anzahl Signifikante Stellen	Anzahl Nachkommastellen
98.76	4	2
0.009876	4	6
$987.6 \cdot 10^4$	4	1
$9.876 \cdot 10^6$	4	3

□

Bemerkungen:

i. Ganze Zahlen haben keine Nachkommastellen.

ii. In manchen Fällen ist die Bestimmung der signifikanten Stellen unklar: Besitzt 20 eine, zwei oder sogar mehr signifikante Stellen? Je nach Zusammenhang ist eine Zahl exakt zu werten, wenn sie z. B. als natürliche Zahl verwendet wird; oder sie ist als gerundete Zahl zu werten, wenn sie als Zahlenwert zu einer physikalischen Grösse verwendet wird. Zu einer exakten Zahl stellt sich die Frage nach der Signifikanz nicht, da sie mit beliebig vielen Nachkomma-Nullen verlängert werden kann.

iii. Um zu einer mittels Messtechnik ermittelten Grösse beim Zahlenwert 20 eine Mehrdeutigkeit zu vermeiden, soll man die wissenschaftliche Schreibweise mit Zehnerpotenz-Faktor wählen. Im Fall von einer signifikanten Stelle also $2 \cdot 10^1$; im Fall von drei signifikanten Stellen $2.00 \cdot 10^1$.

Darstellung Rechenergebnis

Bei der Darstellung eines Rechenergebnis von Messwerten gelten folgende zwei Regeln:

1. Das Ergebnis einer **Addition/Subtraktion** bekommt genauso viele Nachkommastellen wie die Zahl mit den wenigsten Nachkommastellen.
2. Das Ergebnis einer **Multiplikation/Division** bekommt genauso viele signifikante Stellen wie die Zahl mit den wenigsten signifikanten Stellen.

Beispiel 2.1.2

Zahlen	Kleinste Anzahl Signifikante Stellen	Kleinste Anzahl Nachkommastellen	Ergebnis
$20.567 + 0.0007$		3	20.568
$12 + 1.234$		0	13
$12.00 + 1.234$		2	13.23
$12.000 + 1.234$		3	13.234
$1.234 \cdot 3.33$	3		4.11
$1.234 \cdot 0.0015$	2		0.0019

□

Bemerkungen:

i. Eine Rundung sollte erst möglichst spät innerhalb des Rechnungsgangs durchgeführt werden. Sonst können sich mehrere Rundungsabweichungen zu einer grösseren Gesamtabweichung zusammensetzen. Um diese Vergrösserung zu vermeiden, sollen in Zwischenrechnungen bekannte Grössen mit mindestens einer Stelle mehr eingesetzt werden als im Ergebnis angegeben werden kann.

2.1.3. Kennzahlen

Häufig ist es sinnvoll, Datensätze *numerisch* zusammenzufassen. Die Datensätze werden dabei auf eine oder mehrere Zahlen reduziert. Dazu verwenden wir meistens zwei *Kenngrössen*: Eine beschreibt die mittlere Lage der Messwerte und die andere die Variabilität oder Streuung dieser Messwerte. Mit Streuung meinen wir die „durchschnittliche“ Abweichung der Messwerte von der mittleren Lage.

Arithmetisches Mittel

Die bekannteste Grösse für die mittlere Lage ist der wohlbekannte Durchschnitt oder das

Arithmetische Mittel $\overline{x}_n$

$$\overline{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Beispiel 2.1.3 Messung der Schmelzwärme von Eis mit Methode A

Das arithmetische Mittel der $n = 13$ Messungen ist

$$\overline{x}_{13} = \frac{79.98 + 80.04 + \cdots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.02$$

Wir summieren also alle Werte auf und dividieren die Summe durch die Anzahl der Werte.

□

Mit R berechnen wir den Mittelwert wie folgt.

```
methodeA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
    80.04, 79.97, 80.05, 80.03, 80.02, 80, 80.02)

mean(methodeA)

## [1] 80.02077
```

Empirische Varianz und Standardabweichung

Obwohl das arithmetische Mittel schon einiges über einen Datensatz aussagt, beschreibt er diesen aber nur unvollständig. Wir betrachten als Beispiel die folgenden beiden Datensätze von (fiktiven) Schulnoten:

$$2;\ 6;\ 3;\ 5 \quad \text{und} \quad 4;\ 4;\ 4;\ 4$$

Beide haben denselben Mittelwert 4, aber die Verteilung der Daten um den Mittelwert ist sehr unterschiedlich. Im ersten Fall gibt es zwei gute und zwei schlechte Schüler

und im zweiten Fall sind alle Schüler gleich gut. Wir sagen, die Datensätze haben eine unterschiedliche Streuung um die Mittelwerte.

Wir wollen diese Streuung numerisch erfassen. Ein erster Ansatz besteht darin, dass man den Durchschnitt der *Unterschiede zum Mittelwert* nimmt. Im ersten Fall wäre dies

$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4} = \frac{-2+2-1+1}{4} = 0$$

Im zweiten Fall gibt dies auch 0. Diese Methode trägt also nicht viel zur Beschreibung der Streuung bei, da die Unterschiede zum Mittelwert *negativ* werden können und sich diese wie im obigen Fall aufheben können.

Der nächste Ansatz geht dahin, dass wir die Unterschiede zum Mittelwert durch die Absolutwerte der Unterschiede zum Mittelwert ersetzen. Im ersten Fall erhalten wir dann

$$\frac{\big|(2-4)\big|+\big|(6-4)\big|+\big|(3-4)\big|+\big|(5-4)\big|}{4} = \frac{2+2+1+1}{4} = 1.5$$

Die Noten weichen nun im Schnitt 1.5 Noten vom Mittelwert ab. Im zweiten Fall ist dieser Wert natürlich 0. Je grösser dieser Wert (der immer grösser gleich 0 ist), desto mehr unterscheiden sich die Daten bei gleichem Mittelwert voneinander. Dieser Wert für die Streuung heisst auch *mittlere absolute Abweichung*.

Da es sich mit Absolutwerten nicht einfach rechnen lässt (zum Beispiel Ableitungen), wird die (scheinbar kompliziertere) *empirische Varianz* und *empirische Standardabweichung* für das Mass der Variabilität oder Streuung der Messwerte verwendet. Diese sind definiert durch

Empirische Varianz $\mathrm{Var}(x)$ und Standardabweichung s_x

$$\mathrm{Var}(x) = \frac{(x_1-\overline{x}_n)^2+(x_2-\overline{x}_n)^2+\cdots+(x_n-\overline{x}_n)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2$$

$$s_x = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}\,.$$

Bemerkungen:

i. Bei der Varianz quadrieren wir die Abweichungen $x_i - \overline{x}_n$, damit sich die Abweichungen vom Mittelwert nicht gegenseitig aufheben können. Der Nenner $n-1$, anstelle von n, ist mathematisch begründet[1].

[1] Eine genaue Begründung finden Sie im Anhang in Kapitel A.7

ii. Die Standardabweichung ist die Wurzel der Varianz. Da wir für die Berechnung der Varianz die Quadrate der Abstände zum Mittelwert verwendet haben, bekommen wir durch das Wurzelziehen wieder dieselbe Einheit wie bei den Daten selbst. Der Wert der empirischen Varianz hat keine physikalische Bedeutung. Wir wissen nur, je grösser der Wert, desto grösser die Streuung.

Beispiel 2.1.4 Messung der Schmelzwärme von Eis mit Methode *A*

Das arithmetische Mittel der $n = 13$ Messungen ist $\overline{x}_{13} = 80.02$ (siehe oben) und die empirische Varianz ergibt

$$\begin{aligned}\mathrm{Var}(x) &= \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \cdots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1} \\ &= 0.000\,574\,4\end{aligned}$$

Die empirische Standardabweichung ist dann

$$s_x = \sqrt{0.000574} = 0.02397$$

Somit ist die „mittlere“ Abweichung vom Mittelwert $0.023\,97\,\mathrm{cal/g}$.

Für Methode B finden wir $\overline{x}_8 = 79.98$ und $s_x = 0.03137$ mit der analogen Interpretation.

□

Die empirische Varianz bzw. Standardabweichung (englisch: standard deviation) ist von Hand mühsam auszurechnen, deswegen benützen wir R:

```
var(methodeA)

## [1] 0.000574359

sd(methodeA)

## [1] 0.02396579
```

2.1.4. Weitere Kennzahlen

Im Folgenden werden wir zwei alternative Kenngrössen studieren, und zwar den *Median* als Lagemass und die *Quartilsdifferenz* als Streuungsmass.

Median

Ein weiteres Lagemass für die mittlere Lage ist der *Median*. Es handelt sich dabei um den Wert, bei dem rund die Hälfte der Messwerte unterhalb von diesem Wert liegen. Ist beispielsweise bei einer Prüfung der Median 4.6, dann hat die Hälfte der Klasse eine Note unterhalb von 4.6. Umgekehrt liegen die Noten der anderen Hälfte *oberhalb* dieser Note.

Um den *Median* zu bestimmen, müssen wir die Daten zuerst der Grösse nach ordnen:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Für die Daten der Methode A ergibt dies

$$79.97;\ 79.98;\ 80.00;\ 80.02;\ 80.02;\ 80.02;\ 80.03;\ 80.03;\ 80.03;\ 80.04;\ 80.04;\ 80.04;\ 80.05$$

Der Median von diesen 13 Messungen ist dann der Wert der mittleren Beobachtung. Dies ist in diesem Fall der Wert der 7. Beobachtung:

$$79.97; 79.98;\ 80.00;\ 80.02;\ 80.02;\ 80.02;\ \boxed{80.03}\ ;\ 80.03;\ 80.03;\ 80.04;\ 80.04;\ 80.04;\ 80.05$$

Der Median des Datensatzes der Methode A lautet somit 80.03. 6 Beobachtungen sind kleiner oder gleich 80.03 und 6 Messwerte sind grösser oder gleich 80.03. In diesem Beispiel ist die Anzahl der Daten ungerade, und somit gibt es eine mittlere Beobachtung. Ist die Anzahl der Daten gerade, so gibt es zwei gleichwertige mittlere Beobachtungen. Als Median benützen wir in diesem Fall den Mittelwert der beiden mittleren Beobachtungen. Der Datensatz der Methode B hat 8 Beobachtungen. Wir ordnen den Datensatz der Grösse nach und definieren als Median den Durchschnitt von der 4. und 5. Beobachtung:

$$79.94;\ 79.95;\ 79.97;\ \boxed{79.97;\ 79.97}\ ;\ 79.94;\ 80.02;\ 80.03$$

$$\frac{79.97 + 79.97}{2} = 79.97$$

Der Median der Daten der Methode B ist 79.97. Das heisst, die Hälfte der Messwerte ist kleiner oder gleich diesem Wert und die andere Hälfte ist grösser oder gleich diesem Wert. Die Werte der beiden mittleren Beobachtungen sind hier zufällig gleich, dies ist aber im Allgemeinen nicht so.

Mit R bestimmen wir den Median wie folgt:

```
median(methodeA)

## [1] 80.03
```

```
methodeB <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
    79.95, 79.97)

median(methodeB)

## [1] 79.97
```

Wir haben nun zwei Lagemasse für die Mitte eines Datensatzes: das arithmetische Mittel und den Median. Welches sind die Vorzüge der jeweiligen Lagemasse? Eine Eigenschaft des Medians ist die *Robustheit*. Der Median wird weniger stark durch extreme Beobachtungen beeinflusst als das arithmetische Mittel.

Beispiel 2.1.5 Messung der Schmelzwärme von Eis mit Methode A

Bei der grössten Beobachtung ($x_9 = 80.05$) ist ein Tippfehler passiert und $x_9 = 800.5$ eingegeben worden. Das arithmetische Mittel ist dann

$$\overline{x}_{13} = 135.44$$

Der Median ist aber nach wie vor

$$x_{(7)} = 80.03$$

Das arithmetische Mittel wird also durch Veränderung einer Beobachtung sehr stark beeinflusst, während der Median hier gleich bleibt – er ist *robust*.

□

Quartile

Das **untere Quartil** ist derjenige Wert, bei welchem 25 % aller Beobachtungen kleiner oder gleich gross und 75 % grösser oder gleich gross wie dieser Wert sind. Dementsprechend ist das **obere Quartil** derjenige Wert, bei dem 75 % aller Beobachtungen kleiner oder gleich gross und 25 % grösser oder gleich gross sind wie dieser Wert.

Allerdings gibt es für die meisten Datensätze nicht *exakt* 25 % der Anzahl Beobachtungen, wie folgendes Beispiel zeigt: Die Methode A hat $n = 13$ Messpunkte und 25 % dieser Anzahl ist 3.25. Wir *wählen* in diesem Fall den nächstgrösseren Wert $x_{(4)}$ als unteres Quartil:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

Das untere Quartil ist somit 80.02. Rund ein Viertel der Messwerte sind kleiner oder gleich gross wie dieser Wert. Für das obere Quartil wählen wir $x_{(10)}$, da für $0.75 \cdot 13 = 9.75$ die Zahl 10 der nächsthöhere Wert ist.

$$79.97; 79.98;\ 80.00;\ 80.02;\ 80.02;\ 80.02; 80.03;\ 80.03;\ 80.03;\ \underline{80.04};\ 80.04;\ 80.04;\ 80.05$$

Rund drei Viertel aller Messwerte sind also kleiner oder gleich 80.04. Bei der Methode B sind 25 % von 8 Werten zwei Werte. Dies ist eine ganze Zahl, und wir wählen den Durchschnitt von $x_{(2)}$ und $x_{(3)}$ als unteres Quartil. Dann sind 2 Beobachtungen kleiner und 6 Beobachtungen grösser als dieser Wert.

$$79.94;\ \underline{79.95;\ 79.97};\ 79.97;\ 79.97;\ 79.94;\ 80.02;\ 80.03$$

$$\frac{79.95 + 79.97}{2} = 79.96$$

Das untere Quartil der Methode B ist also 79.96.

Bemerkungen:

i. Wir haben hier für den Fall, dass die Anzahl Beobachtungen keine ganze Zahl ist, jeweils aufgerundet. Somit liegt das untere Quartil für Methode A bei der 4. Beobachtung, was etwa 31 % der Beobachtungen entspricht. Hätten wir abgerundet, so entspräche dies der 3. Beobachtung, also ungefähr 23 %, was eigentlich näher bei 25 % liegt als 31 %. Der Unterschied ist hier allerdings nur deswegen recht gross, da der Datensatz mit $n = 13$ ziemlich klein ist. Für grosse Datensätze spielt es praktisch keine Rolle, ob wir auf- oder abrunden.

ii. Es existieren in der Statistik mehrere Definitionen für die Quartile und Quantile. Allerdings sind die Unterschiede für grosse Datensätze klein.

Die Software `R` kennt keine eigenen Befehle für die Quartile. Wir können allerdings den allgemeineren Befehl `quantile` benützen (die Quantile werden gleich genauer behandelt). Damit `R` die Quartile nach unserer Definition berechnet, müssen wir die Option `type=2` hinzufügen.

```
# Syntax für das untere Quartil: p=0.25

quantile(methodeA, 0.25, type = 2)

##   25%
## 80.02

quantile(methodeB, 0.25, type = 2)
```

```
##   25%
## 79.96

# Syntax für das obere Quartil: p=0.75

quantile(methodeA, 0.75, type = 2)

##   75%
## 80.04
```

Quartilsdifferenz

Die Quartilsdifferenz

$$\text{oberes Quartil} - \text{unteres Quartil}$$

ist ein Streuungsmass für die Daten. Es misst die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthält. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte beim Median und umso kleiner ist die Streuung. Dieses Streuungsmass ist robust. So ist die Quartilsdifferenz bei der Methode A

$$80.04 - 80.02 = 0.02$$

```
IQR(methodeA, type = 2)

## [1] 0.02
```

Rund die Hälfte aller Messwerte liegt also in einem Bereich der Länge 0.02.

Quantile

Mit der *Quantile* können wir das Konzept der Quartile auf jede andere Prozentzahl verallgemeinern. So ist das 10 %-Quantil derjenige Wert, wo 10 % der Werte kleiner oder gleich und 90 % der Werte grösser oder gleich diesem Wert sind.

Das *empirische α-Quantil* ist anschaulich gesprochen der Wert, bei dem $\alpha \times 100\,\%$ der Datenpunkte kleiner oder gleich und $(1 - \alpha) \times 100\,\%$ der Punkte grösser oder gleich sind.

Empirische α-Quantile

$$\begin{cases} \frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) \text{ , falls } \alpha \cdot n \text{ eine natürliche Zahl ist,} \\ x_{(k)} \text{ wobei } k \text{ die Zahl } \alpha \cdot n \text{ aufgerundet ist , falls } \alpha \cdot n \notin \mathbb{N} \text{ .} \end{cases}$$

Bemerkungen:

i. Der empirische Median ist das empirische 50 %-Quantil; das empirische 25 %-Quantil ist das untere Quartil und das empirische 75 %-Quantil das obere Quartil.

Beispiel 2.1.6 Messung der Schmelzwärme von Eis mit Methode A

Wir bestimmen Median, unteres und oberes Quartil mit Hilfe der Definition oben.

Es sind $n = 13$ Messwerte, die wir zuerst der Grösse nach ordnen: der kleinste Wert ist $x_{(1)} = 79.97$, der drittgrösste Wert $x_{(3)} = 80.00$, der grösste Wert $x_{(13)} = 80.05$. Wir wollen das 25 %-Quantil, den Median und das 75 %-Quantil bestimmen. Im Fall vom 25 %-Quantil ist dann $\alpha = 0.25$, also

$$\alpha \cdot n = 0.25 \cdot 13 = 3.25$$

was keine natürliche Zahl ist; folglich runden wir 3.25 auf 4 auf und erhalten für das 25 %-Quantil $x_{(4)} = 80.02$. Im Fall vom Median ist $\alpha = 0.5$, also

$$\alpha \cdot n = 0.5 \cdot 13 = 6.5$$

was keine natürliche Zahl ist; folglich runden wir 6.5 auf 7 auf und erhalten für den Median $x_{(7)} = 80.03$. Im Fall vom 75 %-Quantil ist $\alpha = 0.75$, also

$$\alpha \cdot n = 0.75 \cdot 13 = 9.75$$

was keine natürliche Zahl ist; folglich runden wir 9.75 auf 10 auf und erhalten für das 75 %-Quantil den Beobachtungswert $x_{(10)} = 80.04$.

□

Das 10 %- und 70 %-Quantil der Methode A berechnen wir wie folgt:

```
quantile(methodeA, 0.1, type = 2)

##   10%
## 79.98
```

```
quantile(methodeA, 0.7, type = 2)

##    70%
## 80.04
```

Rund 10 % der Messwerte sind kleiner oder gleich 79.98 . Entsprechend sind rund 70 % der Messwerte kleiner oder gleich 80.04.

Beispiel 2.1.7

In einer Schulklasse mit 24 SchülerInnen gab es an einer Prüfung folgende Noten:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

Wir berechnen nun mit R verschiedene Quantile:

```
noten.1 <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9,
    6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2,
    4.9, 5.1)

quantile(noten.1, seq(0.2, 1, 0.2), type = 2)

##  20%  40%  60%  80% 100%
##  3.6  4.2  5.0  5.6  6.0
```

Rund 20 % der SchülerInnen haben also eine 3.6 oder waren schlechter. Genau 20 % der SchülerInnen ist nicht möglich, da dies 4.8 SchülerInnen entsprechen würde. Das 60 %-Quantil besagt, dass 60 Prozent der SchülerInnen eine 5 haben oder schlechter waren. Folglich haben 40 % eine 5 oder sind besser.

□

2.1.5. Graphische Methoden

Histogramm

Einen graphischen Überblick über die auftretenden Werte erhalten wir mit einem sogenannten *Histogramm*. Histogramme helfen uns bei der Frage, in welchem Wertebereich besonders viele Datenpunkte liegen. Ist die Datenmenge gross, so macht es keinen Sinn, alle Werte einzeln zu betrachten. Wir bilden sogenannte *Klassen*, die