

1. Daten sammeln, ordnen und zählen



1.1 Überblick

In diesem Kapitel wird die Grundlage gelegt für alle in diesem Buch behandelten statistischen Fragen.

Nach einem notwendigerweise unvollständigen Hinweis, wo Statistik überall eingesetzt wird, wenden wir uns den Daten zu. Dies sind die Informationen, die für eine statistische Untersuchung gesammelt, geordnet und ausgewertet werden.

Das Gesetz der grossen Zahl gibt eine Ahnung davon, wie viele Untersuchungsergebnisse (z.B. Messungen, befragte Personen) vorliegen müssen, damit die Resultate einigermaßen zuverlässig sind.

Wir sehen ein erstes Mal – und sicher nicht zum letzten Mal, wie klare Zahlen je nach Interesse unterschiedlich gedeutet werden können. Wer nicht auf der Hut ist, kann so leicht in die Irre geführt werden.

Ebenso leicht in die Irre geführt werden kann jemand, der das Paradoxon von Simpson nicht kennt. Es sorgte 1973 an einer renommierten amerikanischen University für ziemlichem Wirbel.

1.2 Statistik im Alltag

Viele Entscheide in Politik und Wirtschaft werden gefällt, nachdem zweckdienliche Daten gesammelt und sorgfältig ausgewertet worden sind. Deshalb ist Statistik in unserem Alltag fast allgegenwärtig, obwohl wir uns dessen oft kaum bewusst sind. Die folgenden Beispiele mögen das illustrieren:

Volkszählung: Der Staat führt alle zehn Jahre eine Volkszählung durch. Die statistische Auswertung der gesammelten Daten liefert den Behörden Grundlagen für vielfältige Entscheidungen und Planungen:

- Die Zahl der Bewohner eines Kantons bestimmt die Zahl seiner Abgeordneten im Nationalrat in Bern und beeinflusst so die Machtverhältnisse in der Schweiz.
- Steigt wegen der immer höheren Lebenserwartung die Zahl der älteren Menschen an, so hat dies Auswirkungen auf die Zahl der benötigten Alters- und Pflegeheimplätze, auf die Altersvorsorge (das Geld aus staatlicher, betrieblicher und privater Altersvorsorge muss für mehr Jahre reichen) usw.

Stromverbrauch: Die Entwicklung des Stromverbrauchs der letzten Jahre gestattet es, den Stromverbrauch in zehn Jahren mit statistischen Methoden abzuschätzen. So kann man vorhersagen, wie viel Strom aus Atomkraftwerken oder anderen Energiequellen nötig sein wird, um den dannzumaligen Stromverbrauch zu decken. Je nach Interessenlage kann man dieselben Daten auch anders interpretieren: Wie viel Strom muss bis in zehn Jahren eingespart werden, damit es kein neues Atomkraftwerk braucht?

Hier zeigt sich deutlich, dass die statistische Berechnung der Prognose das eine, die Deutung der Resultate und die daraus zu ziehenden Folgerungen etwas völlig anderes sind.

Verkaufszahlen: Ein Lehrmittelverlag stellt für die Monate August bis Dezember zusätzliches Personal an. Grund: Die Statistik der Verkaufszahlen zeigt, dass in diesen Monaten wegen des Schuljahresanfangs besonders viele Bücher gekauft werden.

Testauswertung: Die genaue statistische Auswertung eines Tests zeigt, welche Aufgaben von den Prüflingen wie gut gelöst worden und welche Typen von Fehlern häufig begangen worden sind. Diese Erkenntnisse können bewirken, dass einige Themen noch einmal ausführlicher behandelt oder bestimmte Fertigkeiten noch einmal gezielt trainiert werden.

Kleidergrößen: Die Kleiderhersteller haben anhand statistischer Untersuchungen herausgefunden, für welche Beinlängen und Taillenumfänge sie Hosen schneiden müssen, damit diese möglichst vielen Leuten auf Anhieb passen. So praktisch diese Methode für die Mehrheit ist, so unpraktisch ist sie für die Minderheit: Wer zum Beispiel sehr kurze Beine und einen dicken Bauch hat, wird kaum Hosen ab Stange kaufen können oder aber die gekauften Hosen regelmässig nachbearbeiten, z.B. kürzen, lassen müssen.

Finanzielle Vorsorge: Zur Berechnung der Prämien und der Renten benötigen Lebensversicherungen, Pensionskassen und die staatliche Altersvorsorge Angaben darüber, wie viele der bei ihnen versicherten Personen ein bestimmtes Alter erreichen und deshalb eine Altersrente beziehen werden, wie viele Personen vorher sterben und wie viele davon Ehegatten und evtl. Kinder hinterlassen usw.

Fahrplangestaltung: Um den Eisenbahnfahrplan optimal auf die Kundenbedürfnisse abstimmen zu können, führen die Eisenbahngesellschaften Fahrgastzählungen und -befragungen durch.

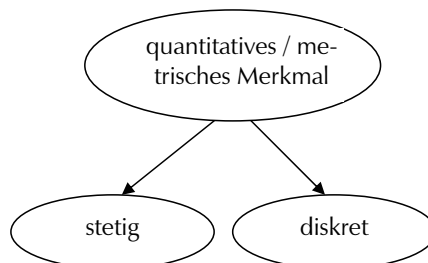
Erprobung neuer Medikamente: Wie wirksam ist ein neu entwickeltes Medikament? Welche Nebenwirkungen treten wie oft auf? Dabei muss einerseits mit grosser Sorgfalt, aber auch mit möglichst grosser Geschwindigkeit vorgegangen werden. Mit grosser Sorgfalt, damit mögliche Nebenwirkungen früh erkannt werden und sich ein Contergan-Debakel nicht wiederholen kann. (Dieses Schlafmittel kam 1957 auf den Markt und steht im Ver-

dacht, bei Neugeborenen zu schweren Missbildungen zu führen. Deshalb wurde es später vom Markt zurückgezogen.) Mit möglichst grosser Geschwindigkeit, damit ein wirksames Medikament so rasch als möglich allen Kranken zur Verfügung steht. Bis zur Entdeckung von Penicillin verliefen Lungenentzündungen aufgrund einer Pneumokokkeninfektion beinahe immer tödlich. Als das Penicillin zur Behandlung von Lungenentzündungen eingeführt wurde, ging die Zahl der Todesfälle drastisch zurück. Es wäre deshalb kaum verantwortungsbewusst gewesen, eine zeitaufwendige Studie durchzuführen und dadurch einigen Kranken die lebensrettende Behandlung vorzuenthalten.

1.3 Arten von Daten

Jede statistische Erhebung befasst sich mit einem Gegenstand, zum Beispiel der „Gesundheit der Bewohnerinnen und Bewohner“ eines Landes. Diejenigen Eigenschaften des Gegenstandes, die man untersucht, nennt man Merkmale oder Variablen, zum Beispiel Körpergrösse und Körpergewicht, Zigaretten- und Alkoholkonsum. Vielleicht werden auch die Merkmale „Geschlecht“ und „Beruf“ erfasst, damit eventuell vorhandene Unterschiede nach Geschlechtern oder Berufsgruppen entdeckt werden können.

Gut geeignet für vielfältige statistische Auswertungen sind „quantitative Variablen“ oder „quantitative Merkmale“. Manchmal werden sie auch „metrische Merkmale“ genannt, weil sie oft durch Messen gewonnen werden (griechisch $\mu\epsilon\tau\rho\epsilon\iota\nu$ „metrein“ = messen). Es handelt sich um Merkmale, deren Ausprägungen man mit Hilfe von Zahlen beschreiben kann: Körpergrösse und Körpergewicht von Menschen, Jahresgehalt, Einwohnerzahl einer Stadt, derzeitige Spannung im Stromnetz usw.. Quantitative Variablen kann man sortieren, um das grösste und das kleinste Jahresgehalt zu bestimmen. Oder man kann mit ihnen Berechnungen anstellen – etwa das durchschnittliche Jahresgehalt ermitteln.



Man kann die quantitativen oder metrischen Merkmale weiter unterteilen, nämlich in stetige Merkmale und diskrete Merkmale. Stetige Merkmale können innerhalb gewisser Grenzen theoretisch jeden Wert annehmen; diskrete Merkmale dagegen können auch innerhalb gewisser Grenzen nur ganz bestimmte Werte annehmen. Beispiele für stetige Merkmale sind Körpergrösse und Körpergewicht, ein diskretes Merkmal ist zum Beispiel die Schulnote der letzten Prüfung. Die Körpergrösse kann jeden Wert zwischen 0.30 und 3.00 Metern annehmen, das Körpergewicht jeden Wert zwischen 1.000 kg und 500.000 kg; dagegen können die Schulnoten nur die Werte 1, 1.5, 2, 2.5, ..., 5.5 und 6 annehmen.

Oft erhält man bei stetigen Merkmalen durch Runden (z.B. auf Zentimeter, Gramm oder Rappen) „diskretisierte“ Werte, welche aber sehr nahe beieinander liegen. Deshalb gelten Körpergrösse, Körpergewicht und Jahreseinkommen als stetige Merkmale einer Person.

1. Daten sammeln, ordnen und zählen

Neben den quantitativen Merkmalen gibt es „qualitative Variablen“ oder „qualitative Merkmale“. Dabei handelt es sich um Merkmale, mit deren Ausprägungen man nicht rechnen kann: Beruf, Herkunftsland, Blutgruppe usw. Bei solchen Informationen kann man oft nur zählen, wie oft ein bestimmter Wert auftritt, weitergehende Berechnungen sind nicht sinnvoll: Was soll der Durchschnitt aller Herkunftsländer sein? Oder die Summe zweier Berufe?

Zwar verwendet man manchmal Zahlen, um die einzelnen Ausprägungen eines Merkmals abzukürzen:

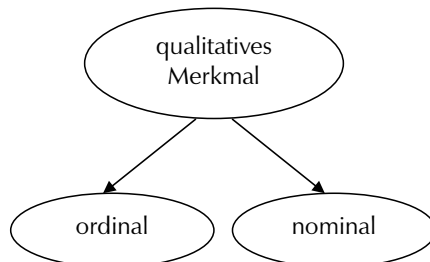
Merkmal „Herkunftsland“: 0 = Schweiz
1 = Deutschland
2 = Italien
3 = Österreich
4 = Frankreich
5 = Liechtenstein
6 = Grossbritannien
usw.

Aber mit diesen Zahlen darf man weder rechnen (Deutschland + Italien \neq Österreich), noch ist in ihnen eine Wertung enthalten (Italien gilt doppelt so viel wie Deutschland), sie sind einfach ein anderer „Name“ für die betreffende Ausprägung.

Manche qualitativen Merkmale besitzen Ausprägungen, die man immerhin in irgendeiner sinnvollen Form ordnen kann, zum Beispiel die zuletzt besuchte Schulstufe:

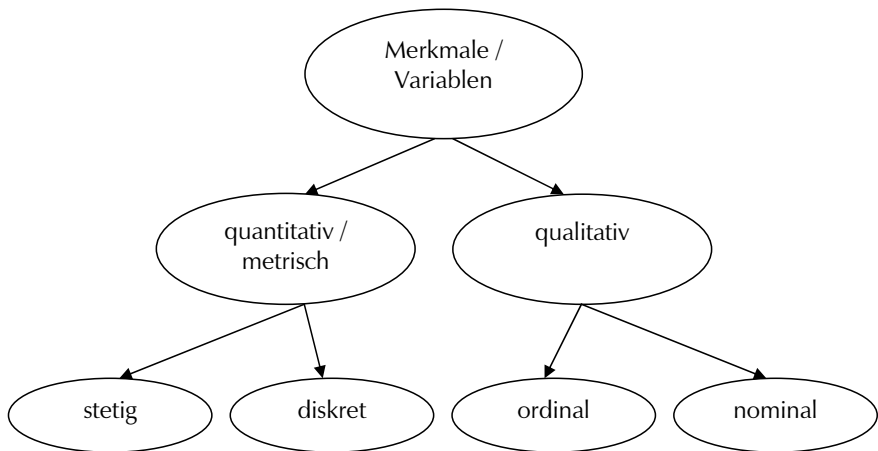
Stufe 0 = keine Schulbildung
Stufe 1 = Primarschule
Stufe 2 = Oberschule
Stufe 3 = Realschule
Stufe 4 = Sekundarschule
Stufe 5 = Gymnasium
Stufe 6 = Hochschule

Ein Eheinstitut achtet vielleicht darauf, dass sich die Schulbildung bei zwei möglichen Partnern um nicht mehr als zwei Schulstufen unterscheidet. Die Zahlen 0, 1, ..., 6 beschreiben die Reihenfolge oder die Rangfolge der Ausprägungen des Merkmals „Schulbildung“. Sie sagen aber zum Beispiel nicht aus, dass ein Mensch der Stufe 6 einen dreimal so hohen Wert hat wie ein Mensch der Stufe 2!



Qualitative Merkmale, deren Ausprägungen man ordnen kann, nennt man „ordinal“ (lateinisch „ordo“ = Ordnung, Reihenfolge, Rang); wenn man die Ausprägungen nicht ordnen kann, ist ein qualitatives Merkmal „nominal“ (lateinisch „nominalis“ = zum Namen gehörig).

Das folgende Schema fasst die Bezeichnungen zusammen:



Beispiele:

Untersuchter Gegenstand	Merkmal oder Variable	Mögliche Ausprägungen oder mögliche Werte	Art des Merkmals
Mensch	Jahresgehalt	Zahl: 0.00, 0.05, 0.10, ..., 100'000.00, ...	metrisch / quantitativ, stetig
	Körpergrösse [cm]	Zahl: 40.0, 40.1, 40.2, ..., 165.0, ..., 300.0	metrisch / quantitativ, stetig
	Körpergewicht [kg]	Zahl: 1.000, 1.001, 1.002, ..., 65.354, ..., 500.000	metrisch / quantitativ, stetig
	Geschlecht	Mann, Frau	qualitativ (nominal)
	Anzahl Geschwister	0, 1, 2, 3, ..., 10	metrisch / quantitativ, diskret
	Beruf	Bauer, Schreiner, Sekretärin, Ärztin, Lehrer, ...	qualitativ (nominal)
	Herkunftsland	Schweiz, Deutschland, Italien, Österreich, ...	qualitativ (nominal)
	Blutgruppe	0, A, B, AB	qualitativ (nominal)
	Fitness	schlecht, recht, mittelmässig, gut, vorzüglich	qualitativ (ordinal)
	Schulbildung	keine, Primarschule, Oberschule, Realschule,	qualitativ (ordinal)
Meinung zum EU-Beitritt	keine Angabe, dagegen, unentschieden, dafür	qualitativ (ordinal)	
Akte der Firma ABC	Aktueller Kurs	Zahl: 0.00, 0.05, 0.10, ..., 45.00, ...	metrisch / quantitativ, stetig
	Ausgeschüttete Dividende	Zahl: 0.00, 0.05, 0.10, ..., 5.00, ...	metrisch / quantitativ, stetig
	Kursgewinn im Vorjahr [%]	Zahl: -100.0, -99.9, ..., +7.9, ...	metrisch / quantitativ, stetig
	An der Börse erhältliche Aktien	Zahl: 0, 1, 2, 3, ...	metrisch / quantitativ, diskret
	Grossaktionäre	Familie D., Bank E., Firma F., Pensionskasse G., Staat I.	qualitativ (nominal)

1. Daten sammeln, ordnen und zählen

	Empfehlung der Bank XYZ	Aktie verkaufen, Aktie behalten, Aktie kaufen	qualitativ (ordinal)
Chemische Substanz	Schmelzpunkt [°C]	-273.15°, -273.14°, ..., 0.0°, ...	metrisch / quantitativ, stetig
	Siedepunkt [°C]	-273.15°, -273.14,, 100.0°, ...	metrisch / quantitativ, stetig
	Aggregatzustand	fest, flüssig, gasförmig	qualitativ (ordinal)
	Giftigkeit	überhaupt nicht, wenig, mittel, hoch	qualitativ (ordinal)

Wir befassen uns in den nächsten Kapiteln häufig mit quantitativen oder metrischen Merkmalen, weil bei diesen mehr mathematische Untersuchungen möglich sind als bei den qualitativen Merkmalen.

1.4 Vollerhebung und Teilerhebung, Grundgesamtheit und Stichprobe

Manchmal ist man in der komfortablen Situation, dass Informationen über *alle* interessierenden Personen oder Dinge – die *Grundgesamtheit* – mit vertretbarem Aufwand erhoben werden können. Man spricht in diesem Fall von einer *Vollerhebung*. Die statistischen Kennzahlen, die man daraus berechnet, sind also grundsätzlich zuverlässig, sofern zum Beispiel bei einer Volkszählung alle Zählbogen ehrlich und fehlerfrei ausgefüllt worden sind. In den folgenden Beispielen von Kapitel 1.2 liegt eine Vollerhebung vor: Volkszählung, Stromverbrauch, Verkaufszahlen, Testauswertung.

Oft kann man aber keine Vollerhebung durchführen, weil diese zu lange dauert oder zu teuer ist. Um dennoch statistische Grundlagen für Entscheidungen zur Verfügung zu haben, führt man eine *Teilerhebung* durch und erhebt die gewünschten Informationen bei einem Teil der interessierenden Personen oder Dinge, bei einer *Auswahl* oder *Stichprobe*. Man hofft, dass die getroffene Auswahl für die Gesamtheit *repräsentativ* ist, d.h. ein exakt verkleinertes Abbild mit denselben Eigenschaften darstellt und deshalb zu denselben Resultaten führt wie die Untersuchung der Grundgesamtheit. Eine Teilerhebung wird bei den folgenden Beispielen von Kapitel 1.2 durchgeführt:

Kleidergrößen: Die benötigten Daten haben die Kleiderhersteller bei vielen Personen erhoben, und nun hoffen sie, dass die gewonnenen Daten nicht nur die Bedürfnisse der „ausgemessenen“ Personen (Stichprobe) wiedergeben, sondern der ganzen Bevölkerung (Grundgesamtheit).

Finanzielle Vorsorge: Eine Vollerhebung für die ganze Schweizer Wohnbevölkerung wäre für eine Versicherungsgesellschaft zu aufwendig und zu teuer. Stattdessen werden diese Untersuchungen für eine möglichst repräsentative Stichprobe durchgeführt, vielleicht für alle bei dieser Gesellschaft versicherten Personen.

Fahrplangestaltung: Es wäre viel zu aufwendig, wenn die Schweizerischen Bundesbahnen SBB jeden ihrer durchschnittlich rund 300'000'000 Passagiere pro Jahr bei jeder Reise befragen würden – und ginge den Reisenden sicher auch sehr rasch auf die Nerven. Durch regelmässige stichprobenweise stattfindende Zählungen und Befragungen auf dem ganzen Streckennetz sammeln die Eisenbahngesellschaften fleissig Daten, von denen sie hoffen, dass sie nicht nur für die (verhältnismässig wenigen) befragten Fahrgäste gelten, sondern für die Grundgesamtheit aller ihrer Fahrgäste zutreffen.

Erprobung neuer Medikamente: Es ist unmöglich, ein Medikament an allen Patienten und Patientinnen mit dieser Krankheit zu testen. Man wird es deshalb an einer ausgewählten Testgruppe ausprobieren und hoffen, dass die dabei gewonnenen Erkenntnisse nicht nur

gerade für die Testgruppe, sondern für alle gegenwärtigen und zukünftigen Menschen mit dieser Krankheit gültig sind.

Das Hauptproblem bei einer Teilerhebung besteht darin, eine *repräsentative* Auswahl von Personen (Stichprobe) zu treffen. Das ist keine einfache Sache! Wir behandeln dieses Problem im Zusammenhang mit Meinungsumfragen in Exkurs B weiter hinten in diesem Buch. Dort gehen wir auch der Frage nach, wie gross eine repräsentative Stichprobe sein muss, damit die Resultate einigermassen zuverlässig auch für die ganze Bevölkerung (Grundgesamtheit) richtig sind.

1.5 Daten sammeln

Wie kommt man zu den gewünschten Daten? Das ist nur in den wenigsten Fällen rasch und einfach möglich. Man greift beim Datensammeln oft auf eines der folgenden Hilfsmittel zurück:

Zählungen: Um abzuklären, ob eine Umfahrungsstrasse ein verkehrsgeplagtes Dorf auch tatsächlich entlastet, sind Verkehrszählungen notwendig. Insbesondere muss ermittelt werden, wie gross der das Dorf durchfahrende Durchgangsverkehr ist. Vor allem der Durchgangsverkehr wird die Umfahrungsstrasse benutzen. Wer im Dorf wohnt oder arbeitet, wird trotz Umfahrungsstrasse weiterhin die Dorfstrassen befahren.

Tests: Um Kenntnisse und praktische Fähigkeiten von Studierenden zu erforschen, muss man Tests sorgfältig planen, durchführen und auswerten.

Experimente: Um naturwissenschaftlichen Gesetzmässigkeiten auf die Spur zu kommen, sind viele Experimente und Messungen nötig. Auch hier ist bei Planung, Durchführung und Auswertung grösste Sorgfalt erforderlich, wenn man aussagekräftige Resultate wünscht.

Fragebogen, Interviews: Die Meinung einer Gruppe oder der ganzen Bevölkerung kann mit Hilfe von Fragebogen oder Interviews (persönlich, telefonisch) untersucht werden. Worauf man hier speziell achten muss, wird in Exkurs B weiter hinten behandelt.

Trotzdem gibt es genügend Beispiele, wo zuverlässige Daten mit vertretbarem Aufwand erhältlich sind.

Prüfung: Hier sind die Antworten der Lernenden und die verteilten Noten der zuständigen Lehrperson ohne weiteres zugänglich und stehen für eine Auswertung zur Verfügung: Welche Fehler traten häufig auf? War eine Aufgabe unerwartet leicht oder unerwartet schwierig? Wie gut fiel die Prüfung insgesamt aus? usw.

Offene Abstimmung: Die Meinung einer nicht allzu grossen Gruppe kann mithilfe einer offenen Abstimmung ermittelt werden. An einer Gemeindeversammlung wird gezählt, wie viele der anwesenden Stimmberechtigten einer Vorlage durch Erheben der Hand zustimmen und wie viele die Vorlage ebenfalls durch Erheben der Hand ablehnen. Vorausgesetzt, die Stimmberechtigten erheben ihre Hand im richtigen Moment und die Stimmen werden richtig gezählt, gibt das Abstimmungsergebnis die Meinung der Anwesenden auf die Stimme genau richtig wieder.

Der Kanton Glarus mit knapp 40'000 Einwohnerinnen und Einwohnern erlässt seine neuen Gesetze einmal jährlich an der „Landsgemeinde“. Abgestimmt wird durch Handerheben. Natürlich ist ein genaues Auszählen bei rund 6'000 Stimmenden nicht mehr möglich. In den meisten Fällen sind die Mehrheitsverhältnisse aber auf einen Blick klar erkennbar. Wenn die Mehrheit schwer abzuschätzen ist, zieht der Landammann, welcher der Landsgemeinde vorsteht, vier Regierungsmitglieder bei. Diese schätzen die Mehrheitsverhältnisse ab, und die getroffene Entscheidung gilt. Es ist also grundsätzlich möglich,

1. Daten sammeln, ordnen und zählen

das die getroffene Entscheidung knapp nicht die Mehrheitsmeinung widerspiegelt. Eine Urnenabstimmung wäre sicher zuverlässiger. Aber dafür bietet die Glarner Landsgemeinde die Möglichkeit, dass eine einzelne Bürgerin die Änderung eines vorgelegten Gesetzes beantragen kann, und über diesen Antrag muss abgestimmt werden. So viel Bürgernähe ist bei einer Urnenabstimmung nicht möglich, und deshalb dürfte die Glarner Landsgemeinde noch einige Zeit überdauern.

Die erhobenen, noch nicht bearbeiteten Daten heissen *Rohdaten*. Oftmals liegen sie in Form einer Liste vor, der *Urliste*.

- Wir nehmen an, bei einer Prüfung würden folgende Noten erzielt: 6, 4, 4, 5, 4½, 5, 4½, 6, 4, 3, 1, 3, 5, 5½, 4½, 5½, 4½, 4½, 2½, 5, 3½, 2½, 3, 3½, 5½. (Die Noten orientieren sich am Schweizer Schulsystem: 6 bezeichnet die beste, 1 die schlechteste Leistung; Noten von 4 an aufwärts sind genügend). Diese 25 Zahlen sind die „Rohdaten“, sie bilden zusammen die „Urliste“.
- Bei einer Meinungsumfrage sind die Rohdaten die ausgefüllten Fragebogen oder die Notizen zu den durchgeführten Interviews.
- Bei einer Abstimmung und bei einer Wahl sind die Rohdaten die von den Stimmberechtigten ausgefüllten, noch unsortierten Stimm- oder Wahlzettel.

1.6 Daten ordnen

1.6.1 Strichliste

Die Reihenfolge, in der die Rohdaten in der Urliste erscheinen, ist oftmals bedeutungslos. Es ist deshalb zulässig, die Daten zu ordnen, um so einen ersten Überblick zu gewinnen und die nachfolgende Auswertung vorzubereiten. Ein einfaches, aber praktisches Hilfsmittel ist die Strichliste. Im oben erwähnten Prüfungsbeispiel sieht die Strichliste so aus:

Prüfung	
Note	Häufigkeit
6	II
5½	III
5	IIII
4½	IIIII
4	III
3½	II
3	III
2½	II
2	
1½	
1	I

Bereits diese einfache Strichliste zeigt, dass bestimmte Werte deutlich häufiger vorkommen als andere und dass ein Wert deutlich von den anderen abfällt. Diese Informationen waren aus den Urlisten kaum so einfach herauszulesen.

Wenn man schon von vornherein ungefähr weiss, welche Werte auftreten können, wird man die Daten von allem Anfang an in einer Strichliste erfassen und sich so das nachträgliche Erfassen der Rohdaten in einer Strichliste ersparen. Ein gutes Formular bei der Datenerfassung kann die Auswertung sehr erleichtern!

1.6.2 Tabelle

Mehrere Urlisten zu derselben Fragestellung kann man in einer Tabelle übersichtlich zusammenfassen. Beispielsweise kann dieselbe Prüfung in mehreren Klassen durchgeführt werden.

Prüfung			
Note	Häufigkeit		
	Klasse A	Klasse B	Klasse C
6			
5½			I
5			
4½			
4			
3½			
3		I	
2½			
2			I
1½			I
1	I		

Auch diese Tabellen offenbaren wieder einiges: In der relativ kleinen Klasse B hat niemand eine Spitzenleistung erbracht, aber es ist auch keine Prüfung massiv ungenügend ausgefallen. In Klasse C gibt es zwei Lernende, welche in der sonst recht guten Klasse deutlich abfallen.

1.7 Klassen

Im Beispiel „Prüfung“ tritt das untersuchte Merkmal „Note“ in 11 Ausprägungen auf: 1, 1½, 2, 2½, ..., 5½, 6. Weil die Anzahl der Ausprägungen nicht besonders gross ist, wird die Häufigkeit jeder einzelnen Ausprägung bestimmt.

Manchmal ist dies nicht möglich oder nicht sinnvoll. Bei einer Untersuchung über das Jahreseinkommen von Herrn und Frau Schweizer kann theoretisch jeder Betrag zwischen CHF 0.00 und CHF 20'000'000.00 vorkommen. Wenn wir das Einkommen auf ganze Franken runden, sind dies immerhin 20'000'001 mögliche Werte des Merkmals „Jahreseinkommen“. Es ist viel zu aufwendig, über so viele Werte Buch zu führen. Da in der Schweiz ziemlich genau 4'000'000 Menschen arbeiten, werden die meisten der möglichen Werte überhaupt nie oder höchstens einmal auftreten.

Jahreseinkommen (Nettolohn in CHF)	Häufigkeit
.....	
48'273	I
48'274	
48'275	
48'276	
48'277	
48'278	I
48'279	
.....	

Man bildet deshalb Klassen, in denen diverse Jahreseinkommen zusammengefasst werden. Die Befragung von 50 zufällig ausgewählten Personen könnte zu folgender Strichliste führen:

Jahreseinkommen (Nettolohn in CHF)	Häufigkeit
0 ... 24'000	
24'000 ... 48'000	

1. Daten sammeln, ordnen und zählen

48'000 ... 72'000	
72'000 ... 96'000	
96'000 ... 120'000	
120'000 ... 144'000	
über 144'000	

Dabei unterscheidet man nicht zwischen einem Einkommen von CHF 48'000 und einem von CHF 71'900, beide gehören zu derselben Klasse. Beim Bilden von Klassen gehen also Informationen verloren!

Die Klasseneinteilung muss deshalb sorgfältig vorgenommen werden. Sie richtet sich nach den möglichen Ausprägungen des untersuchten Merkmals. Oft achtet man darauf, dass die Klassen gleich gross sind. Im Beispiel fassen alle Klassen 24'000 mögliche Ausprägungen zusammen – mit Ausnahme der nach oben offenen letzten Klasse.

Damit die Strichliste sauber erstellt werden kann, muss klar sein, wie Löhne gehandhabt werden, die genau auf die Klassengrenzen fallen wie CHF 24'000, CHF 48'000 usw. In diesem Beispiel zählen wir die Klassengrenze bereits zur oberen Klasse. In die mit 24'000 ... 48'000 bezeichnete Klasse fallen also alle Nettoeinkommen von CHF 24'000 bis und mit CHF 47999. Solche Feinheiten sind für die Praxis oft nicht von grosser Bedeutung, weil ja kaum ein Nettolohn auf *ganz genau* CHF 48'000 fallen dürfte ...

Sowohl zu viele, zu wenige als auch ungünstig gewählte Klassen verunmöglichen den Blick auf das Wesentliche. Es gilt folgende Faustregel:

Bei n erfassten Werten gilt für die Anzahl k der Klassen:

$$k \approx \sqrt{n}, \text{ aber } k \leq 20.$$

[1.1]

Bei n=50 erfassten Werten sind etwa k=7 Klassen sinnvoll. Mehr als 20 Klassen dürften auch bei riesigen Grundgesamtheiten oder Stichproben kaum noch sinnvoll sein.

Beispiel Nettolöhne: Bildet man Klassen der Breite 48'000 oder gar der Breite 72'000, so erhält man die folgenden Strichlisten:

Jahreseinkommen (Nettolohn in CHF)	Häufigkeit
0 ... 48000	
48'000 ... 96'000	
96'000 ... 144'000	
über 144'000	

Jahreseinkommen (Nettolohn in CHF)	Häufigkeit
0 ... 72'000	
72'000 ... 144'000	
über 144'000	

Beim Betrachten dieser letzten Strichliste mit Klassenbreite 72'000 erhält man einen anderen Eindruck als bei der ersten Strichliste mit Klassenbreite 24'000. Wer die Jahreseinkommen nicht in Bezug zu den Lebenshaltungskosten setzen kann, könnte durchaus den Eindruck gewinnen, dass die grosse Mehrheit der in der Schweiz lebenden Menschen – über drei Viertel! – relativ wenig verdient, dass es eine „Mittelschicht“ von rund einem Fünftel und eine dünne „Oberschicht“ gibt. Noch deutlicher tritt dieser Effekt bei grafischen Darstellungen zutage. Wir kommen deshalb dort noch einmal auf dieses Problem zurück (→ 2.4).

1.8 Absolute Häufigkeit

Einen ersten Überblick gewinnt man über die Daten bereits anhand dieser Strichlisten. Wer es genauer wissen möchte, zählt, wie oft jeder Wert auftritt. Im Fachjargon heisst das: „Man bestimmt die absolute Häufigkeit jedes Wertes.“ Bei unserem Prüfungsbeispiel (Klasse A, → 1.6.2) sieht das so aus:

Prüfung		
Note	Absolute Häufigkeit	
6		2
5½		3
5		4
4½		5
4		3
3½		2
3		3
2½		2
2		0
1½		0
1		1

Die absolute Häufigkeit des Wertes 5 ist 4, weil die Note 5 insgesamt 4mal vorkommt; die absolute Häufigkeit des Wertes 2 ist 0, weil niemand eine 2 geschrieben hat.

Um leichter mathematische Untersuchungen durchführen zu können, verwendet man einige Bezeichnungen und Abkürzungen.

Prüfung			
Nr. i	Note x_i	Absolute Häufigkeit H_i	
11	$x_{11} = 6$		$H_{11} = 2$
10	$x_{10} = 5\frac{1}{2}$		$H_{10} = 3$
9	$x_9 = 5$		$H_9 = 4$
8	$x_8 = 4\frac{1}{2}$		$H_8 = 5$
7	$x_7 = 4$		$H_7 = 3$
6	$x_6 = 3\frac{1}{2}$		$H_6 = 2$
5	$x_5 = 3$		$H_5 = 3$
4	$x_4 = 2\frac{1}{2}$		$H_4 = 2$
3	$x_3 = 2$		$H_3 = 0$
2	$x_2 = 1\frac{1}{2}$		$H_2 = 0$
1	$x_1 = 1$		$H_1 = 1$

- Die *möglichen Werte des untersuchten Merkmals* (z.B. Note) werden im allgemeinen mit x_1, x_2, x_3, \dots usw. bezeichnet.
- Die *Anzahl der möglichen Werte* wird oft mit k bezeichnet. Im Beispiel „Prüfung“ ist $k=11$, weil es insgesamt 11 verschiedene Noten gibt: 1, 1½, 2, 2½, ..., 5½ und 6.
- Die *absolute Häufigkeit* des Wertes x_i wird mit H_i bezeichnet. H_7 bezeichnet also die absolute Häufigkeit des Wertes x_7 .

1. Daten sammeln, ordnen und zählen

- Die Anzahl der untersuchten Objekte heisst *Umfang* der Grundgesamtheit (oder Umfang der Stichprobe) und wird mit n bezeichnet. Im Beispiel „Prüfung“ ist $n=25$, weil 25 Noten vorliegen.

Mit den soeben eingeführten Bezeichnungen kann man bereits eine Regel formulieren:

$$H_1 + H_2 + H_3 + \dots + H_k = n \quad [1.2]$$

1.9 Relative Häufigkeit

Die absolute Häufigkeit allein ist oftmals nicht besonders aussagefähig. Was nützt es zu wissen, dass 400 Schülerinnen und Schüler einer bestimmten Schule für die Einführung der Fünftageweche sind? Herzlich wenig, weil Sie nicht wissen, wie viele Schülerinnen und Schüler insgesamt ihre Meinung abgegeben haben. Wenn es insgesamt 500 Lernende gewesen sind, dann gibt es eine klare Mehrheit für die Einführung der Fünftageweche; wenn es dagegen 810 Lernende gewesen sind, sind das Nein- und das Ja-Lager fast gleich gross, und wenn es 1600 Lernende gewesen sind, bilden die 400 Befürwortenden eine klare Minderheit.

Viel nützlicher wäre es in diesem Fall zu wissen, welcher Anteil (oder welcher Prozentsatz) der die Meinung äussernden Schülerinnen und Schüler für bzw. gegen die Einführung der Fünftageweche ist. Genau dazu dient die relative Häufigkeit.

Die *relative Häufigkeit* h_i (kleines h) des Merkmalswertes x_i gibt an, welchen Anteil zum Umfang der Grundgesamtheit (bzw. der Stichprobe) der Merkmalswert x_i beiträgt.

Wenn 400 von 500 Lernenden für die Einführung der Fünftageweche sind, so machen die Befürwortenden $\frac{400}{500} = \frac{4}{5}$ aller Studierenden aus. Manchmal gibt man dieses Resultat nicht als gewöhnlichen Bruch an, sondern als Dezimalbruch oder in Prozenten. Dann liegt der Anteil der Befürwortenden bei 0.8 oder bei 80%. Und diese Zahl sagt Ihnen nun, dass eine klare Mehrheit die Einführung der Fünftageweche wünscht. Allgemeiner:

Die relative Häufigkeit h_i des Merkmalswertes x_i wird berechnet nach der Formel

$$h_i = \frac{H_i}{n} \quad [1.3]$$

oder – wenn das Resultat in Prozenten erwünscht ist –

$$h_i = \frac{H_i}{n} \cdot 100\%.$$

Wir berechnen die relativen Häufigkeiten der einzelnen Werte im Beispiel „Prüfung“:

Prüfung						
Nr. i	Note x_i	Absolute Häufigkeit H_i		Relative Häufigkeit h_i als Bruch / Dezimalbruch / in Prozenten		
11	$x_{11} = 6$	II	$H_{11} = 2$	$h_{11} = \frac{2}{25}$	$h_{11} = 0.08$	$h_{11} = 8\%$
10	$x_{10} = 5\frac{1}{2}$	III	$H_{10} = 3$	$h_{10} = \frac{3}{25}$	$h_{10} = 0.12$	$h_{10} = 12\%$
9	$x_9 = 5$	IIII	$H_9 = 4$	$h_9 = \frac{4}{25}$	$h_9 = 0.16$	$h_9 = 16\%$
8	$x_8 = 4\frac{1}{2}$	IIIIII	$H_8 = 5$	$h_8 = \frac{5}{25} = \frac{1}{5}$	$h_8 = 0.2$	$h_8 = 20\%$
7	$x_7 = 4$	III	$H_7 = 3$	$h_7 = \frac{3}{25}$	$h_7 = 0.12$	$h_7 = 12\%$
6	$x_6 = 3\frac{1}{2}$	II	$H_6 = 2$	$h_6 = \frac{2}{25}$	$h_6 = 0.08$	$h_6 = 8\%$

5	$x_5 = 3$	III	$H_5 = 3$	$h_5 = \frac{3}{25}$	$h_5 = 0.12$	$h_5 = 12\%$
4	$x_4 = 2\frac{1}{2}$	II	$H_4 = 2$	$h_4 = \frac{2}{25}$	$h_4 = 0.08$	$h_4 = 8\%$
3	$x_3 = 2$		$H_3 = 0$	$h_3 = \frac{0}{25}$	$h_3 = 0$	$h_3 = 0\%$
2	$x_2 = 1\frac{1}{2}$		$H_2 = 0$	$h_2 = \frac{0}{25}$	$h_2 = 0$	$h_2 = 0\%$
1	$x_1 = 1$	I	$H_1 = 1$	$h_1 = \frac{1}{25}$	$h_1 = 0.04$	$h_1 = 4\%$

Es ist stets $h_1 + h_2 + h_3 + \dots + h_k = 1$ bzw. 100%. [1.4]

Grund: $h_1 + h_2 + h_3 + \dots + h_k = \frac{H_1}{n} + \frac{H_2}{n} + \frac{H_3}{n} + \dots + \frac{H_k}{n} = \frac{H_1 + H_2 + H_3 + \dots + H_k}{n} = \frac{n}{n} = 1$.

Das zweitletzte Gleichheitszeichen gilt wegen [1.2].

Relative Häufigkeiten können oft auch als Wahrscheinlichkeiten aufgefasst werden. Einige Beispiele dazu:

Prüfung: Wenn man einen der 25 Prüflinge zufällig herausgreift, so ist die Wahrscheinlichkeit, dass man einen mit Note $4\frac{1}{2}$ „gezogen“ hat, $\frac{5}{25} = \frac{1}{5}$ bzw. 0.2 bzw. 20%.

Würfeln: Wenn man 120mal einen ungefälschten Würfel wirft und dabei 24mal eine 2 würfelt, so ist die relative Häufigkeit für eine 2 eben $h_2 = \frac{24}{120} = \frac{1}{5} = 0.2 = 20\%$. Das ist eine *Näherung* für die Wahrscheinlichkeit, eine 2 zu würfeln. Diese Wahrscheinlichkeit liegt bekanntlich bei $\frac{1}{6} \approx 0.167 = 16.7\%$.

US-Präsidenten: Von den bisherigen 43 US-Präsidenten (bis und mit George W. Bush) wurden 4 während ihrer Amtszeit ermordet: Abraham Lincoln 1865, James Garfield 1881, William McKinley 1901 und John F. Kennedy 1963. Die relative Häufigkeit beträgt $\frac{4}{43} \approx 0.093 = 9.3\%$. Wenn man also zufällig einen amerikanischen Präsidenten nennt, so ist die Wahrscheinlichkeit, dass dieser während seiner Amtszeit ermordet wurde, $\frac{4}{43} \approx 0.093 = 9.3\%$. Makabre Fortsetzung der Überlegung: Dies ist eine *Näherung* für die Wahrscheinlichkeit, dass der jetzige Präsident noch in seiner Amtszeit ermordet wird.

1.10 Das Gesetz der grossen Zahl

Bei einem normalen Würfel erscheinen die Zahlen 1, 2, 3, 4, 5 und 6 alle ungefähr gleich oft, nämlich bei etwa einem Sechstel aller Würfe. Diese Aussage gilt aber erst *auf lange Sicht*, d.h. nach *vielen* Würfeln oder eben nach *einer grossen Zahl* von Würfeln.

Sicher kennen Sie eine Spielsituation, wo Sie auf eine bestimmte Zahl gewartet haben, diese aber einfach nie erschienen ist. Der Zufall hat eben kein Gedächtnis und weiss deshalb nicht, wann eine bestimmte Zahl wieder einmal erscheinen sollte. Daher kann dieselbe Zahl viermal hintereinander geworfen werden und dann 30mal hintereinander überhaupt nie. Im ersten Fall kommt sie viel zu oft vor, im zweiten Fall viel zu selten. Erst auf lange Sicht wird sie ziemlich genau in einem Sechstel aller Würfe erscheinen.

Mit anderen Worten: Wenn man nur wenige Würfe ausführt, regiert der reine Zufall: Es sind keine Prognosen möglich. Wenn man aber sehr viele Würfe ausführt, kann man im Zufall Gesetzmässigkeiten erkennen – zum Beispiel, dass jede Zahl ungefähr gleich oft erscheint. Diese Erkenntnis stammt von Jacob Bernoulli und wurde um 1700 formuliert:

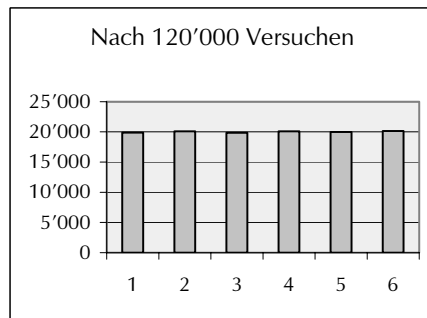
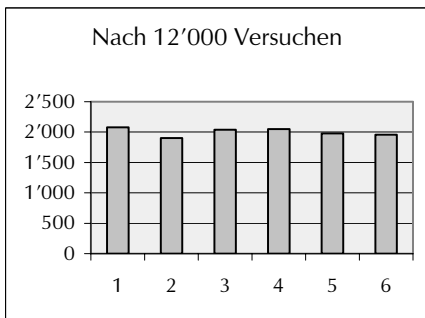
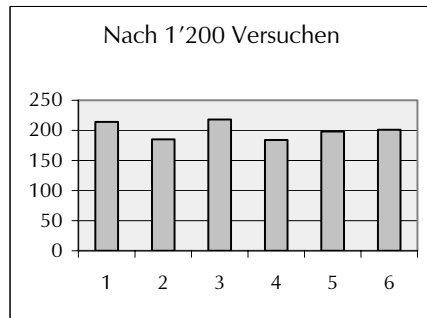
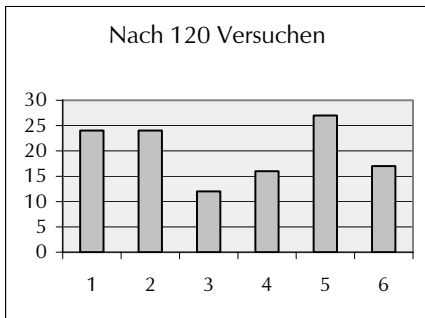
„Gesetz der grossen Zahl“: Die Abweichung zwischen dem wahren (theoretisch erwarteten) Wert und dem beobachteten Wert eines Experimentes nimmt ab, je grösser die Zahl der Beobachtungen ist.

1. Daten sammeln, ordnen und zählen

Das bedeutet: Je mehr Experimente (z.B. Würfelwürfe, Fahrgastbefragungen, ...) für statistische Untersuchungen durchgeführt werden, desto zuverlässiger und genauer stimmt der experimentell gefundene Wert mit dem tatsächlichen Wert überein. Wir illustrieren dies an einer Computersimulation. Ein PC „würfelt“ und zählt, wie oft jede Zahl erscheint.

Versuche	1	2	3	4	5	6
12	4	3	3	0	1	1
120	24	24	12	16	27	17
1'200	214	185	218	184	198	201
12'000	2'077	1'901	2'039	2'049	1'978	1'956
120'000	19'883	20'097	19'846	20'073	19'968	20'133
1'200'000	200'316	199'635	200'007	200'799	200'314	198'929
12'000'000	1'997'239	2'001'286	2'001'533	2'001'573	1'999'619	1'998'750
120'000'000	20'001'720	19'994'638	20'002'414	20'002'363	20'000'843	19'998'022

Nach nur 12 Versuchen ist noch keine Regelmässigkeit zu erkennen: Die 1 erschien bei 4 von 12 Würfeln, die 4 dagegen überhaupt nie. Später holt die 4 aber gewaltig auf, und nach 12'000'000 Versuchen kommt sie sogar am häufigsten vor. Je mehr Versuche durchgeführt werden, desto klarer wird aber, dass jede Zahl ungefähr gleich oft erscheint. Noch offensichtlicher wird dies mithilfe einer grafischen Darstellung. Man erkennt, dass die anfänglich sehr grossen Unterschiede später soweit verschwinden, dass im Rahmen der Zeichengenauigkeit alle Säulen praktisch gleich gross sind und jede Zahl mit der relativen Häufigkeit $\frac{1}{6}$ vorkommt.



Nur: Bis es soweit ist, sind 120'000 Versuche nötig! Wenn man dasselbe Experiment mit den deutschen Lottozahlen 1, 2, 3, ..., 49 durchspielt, braucht es etwa 10'000'000 Ziehungen zu je 6 Zahlen, bis alle 49 Zahlen im Rahmen der Zeichengenauigkeit gleich oft gezogen worden sind. Beim „Gesetz der grossen Zahl“ kann es also um wirklich sehr grosse Zahlen gehen!

Was für Lottozahlen gilt, gilt ebenso für Roulettezahlen 0, 1, 2, ..., 36. Auch Spielcasinos basieren letztlich genau auf dem Gesetz der grossen Zahl. Für den Spieler ist es reizvoll, gewinnen oder verlieren zu können, weil er dem blanken Zufall ausgeliefert ist. Auch wenn er einen Abend und eine Nacht lang durchgehend sein Glück probiert, ist dies doch viel zu wenig oft, als dass bereits das Gesetz der grossen Zahl gelten würde.

Das Casino dagegen hält wenig vom blanken Zufall, es möchte mit dem Spielbetrieb regelmässig Geld verdienen. Wenn an vielen Spieltischen viele Gambler oft spielen, gilt das Gesetz der grossen Zahl. Dieses besagt, dass beim Roulette langfristig $1/37$ (in Europa) bzw. $2/38 = 1/19$ (in den USA) aller Einsätze ans Casino geht. Und in Anbetracht all der vielen Einsätze ergibt dies auf lange Sicht ein hübsches Sümmchen ...

1.11 Vorsicht im Umgang mit Häufigkeiten!

Von Winston Churchill (1874 - 1965) stammt das Bonmot „Ich glaube nur jenen Statistiken, die ich selber manipuliert habe.“ Manipulation im Reich der Mathematik, wo doch alles streng logisch zugeht und klar berechenbar ist?

Im Bereich des Berechnens gibt es tatsächlich keinen Spielraum für Manipulation. Spielraum gibt es aber bei den Entscheidungen, was berechnet wird und wie die Resultate zu deuten sind. Die folgenden Beispiele aus der Wirtschaft zeigen das deutlich:

Arbeitslosigkeit: Angenommen, in einem Land gibt es 10'000'000 Arbeitswillige, von denen 500'000 keine Arbeit finden. Die Arbeitslosenquote, d.h. die relative Häufigkeit der Arbeitslosen, liegt also bei $\frac{500'000}{10'000'000} \cdot 100\% = 5\%$. Nun steigt die Arbeitslosenzahl auf 600'000 an.

Reaktion von Partei A: Die Arbeitslosenquote liegt neuerdings bei $\frac{600'000}{10'000'000} \cdot 100\% = 6\%$.

Folge: In der Partei A freundlich gesinnten Zeitung erscheint eine kleine Randnotiz unter dem Titel „Die Arbeitslosigkeit hat um 1% zugenommen“.

Reaktion von Partei B: Die Zahl der Arbeitslosen hat sich um 100'000 erhöht, das sind $\frac{100'000}{500'000} \cdot 100\% = 20\%$ aller bisherigen Arbeitslosen. Folge: Eine fette Schlagzeile auf der Titelseite der Partei B nahe stehenden Zeitung: „Die Arbeitslosigkeit hat um 20% zugenommen“.

Genau dieselben Zahlen kann man also je nach Interesse so oder anders interpretieren. Entsprechend ergeben sich daraus verschiedene Konsequenzen. Das funktioniert in diesem Beispiel deshalb, weil nicht klar ist, was 100% sind. Bei den Arbeitgebern sind 100% die 10'000'000 Arbeitswilligen, bei den Gewerkschaften sind 100% die 500'000 Arbeitslosen.

Verkehrsrisiko: Ist die Eisenbahn das gefährlichere Verkehrsmittel als das Flugzeug? Da stellt sich als erstes die Frage, wie man das Risiko messen will. Zählt man die tödlich verunglückten Flugzeug- oder Eisenbahnpassagiere? Oder zählt man auch durch Trümmerteile getötete Menschen ausserhalb des Flugzeuges oder des Zuges? Setzt man diese Zahl in Bezug zur Anzahl der insgesamt zurückgelegten Reisen? Oder in Bezug zur zurückgelegten Strecke? Oder zur Reisezeit? Je nach Wahl erfolgt die Angabe dann zum Beispiel in Anzahl Toten pro 1'000'000'000 Personenreisen, in Anzahl Toten pro 1'000'000'000 Personenkilometer oder in Anzahl Toten pro 1'000'000'000 Personenstunden. Es besteht also durchaus ein gewisser Ermessensspielraum bei der Wahl der Messmethode, mit der man das Verkehrsrisiko bestimmt. Ein Beispiel soll den Unterschied verdeutlichen.

Ein vollbesetzter Schnellzug mit 1000 Passagieren an Bord legt in 20 Stunden Fahrzeit eine Strecke von 2000 Kilometern zurück. Bilanz dieser Fahrt:

1. Daten sammeln, ordnen und zählen

1000 Personen · 1 Reise = 1000 Personenreisen,
1000 Personen · 20 Stunden = 20'000 Personenstunden,
1000 Personen · 2'000 Kilometer = 2'000'000 Personenkilometer.

Und nun zur ursprünglichen Frage, ob die Eisenbahn oder das Flugzeug sicherer sei.

Antwort 1: Die Eisenbahn ist sicherer. Begründung: Pro Milliarde Personenstunden verunglücken beim Fliegen 240 Passagiere tödlich, beim Eisenbahnfahren dagegen 70.

Antwort 2: Das Flugzeug ist sicherer. Begründung: Pro Milliarde Personenkilometer gibt es beim Fliegen im Mittel 0.3 tote Passagiere, beim Eisenbahnfahren dagegen 1.0 tote Passagiere.

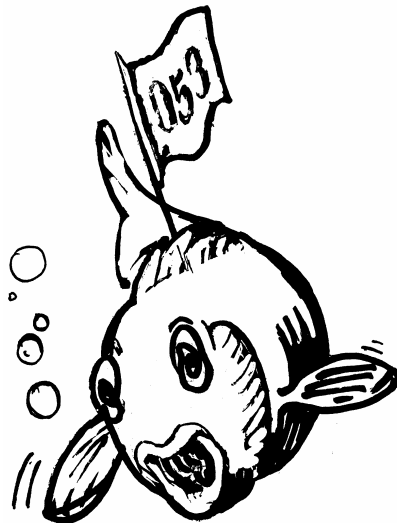
Antwort 3: Die Eisenbahn ist sicherer. Begründung: Pro Milliarde Personenreisen sterben im Mittel 47 Eisenbahnpassagiere, aber 550 Flugzeugpassagiere.

Wer gibt wohl welcher Antwort den Vorzug? Je nach Interessenlage kann man die eine oder die andere Meinung vertreten. Und beide lassen sich statistisch einwandfrei untermauern. Vielleicht war dies einer der Gründe, weshalb Churchill der Statistik kritisch gegenüberstand. Wir werden auch in den nächsten Kapiteln ab und zu auf solche Beispiele stossen.

1.12 Ergänzung: Anwendungen der relativen Häufigkeit

1.12.1 Bestimmung des Fischbestandes eines Sees

Um die Zahl der in einem See lebenden Fische abschätzen zu können, verwendet man einen Kniff. Man fängt – sagen wir einmal – 100 Fische, markiert sie und lässt sie wieder frei. Nach einiger Zeit haben sich die 100 Fische über den ganzen See verteilt und sich gleichmässig unter die nicht markierten Fische gemischt.



Dann fängt man noch einmal Fische – sagen wir wieder 100. Von diesen 100 Fischen seien 12 markiert. Die relative Häufigkeit der markierten Fische unter den gefangenen Fischen beträgt also $\frac{12}{100} = 0.12$.

Wenn man annimmt, dass die markierten Fische im See gleichmässig unter den nicht markierten verteilt waren, dann ist auch die relative Häufigkeit aller markierten Fische unter allen Fischen im See gleich 0.12:

$$\frac{\text{insgesamt markierte Fische}}{\text{alle Fische im See}} = 0.12$$

Insgesamt gibt es 100 markierte Fische. Also gilt

$$\frac{100}{\text{alle Fische im See}} = 0.12,$$

und die Zahl aller Fische im See ist ungefähr $\frac{100}{0.12} = 833$.

Natürlich funktioniert dieser Kniff nur dann, wenn die markierten Fische einigermaßen gleichmässig über den See verteilt sind und auch sonst eine möglichst repräsentative Auswahl des ganzen Fischbestandes darstellen. Wenn man diesen Kniff mit der gebotenen Sorgfalt anwendet, ist es ein praktisches Verfahren, um den Fischbestand einigermaßen abschätzen zu können.

1.12.2 Sterbetafeln

Frau Meyer ist heute 32 Jahre alt. Sie wünscht, dass ihre derzeit 5-jährige Tochter bis zu deren 25. Geburtstag monatlich eine Rente von CHF 3'000 bekommt, falls Frau Meyer vorher sterben sollte.

Die von Frau Meyer um ein Angebot gebetene Versicherungsgesellschaft berechnet die Prämie, die Frau Meyer jährlich zu zahlen hat. Dazu muss sie abschätzen können, mit welcher Wahrscheinlichkeit Frau Meyer vor dem 25. Geburtstag ihrer Tochter stirbt. Wenn Frau Meyer kurz nach Vertragsunterzeichnung im Alter von 32 Jahren stirbt, muss die Versicherung während 20 Jahren pro Monat CHF 3'000 auszahlen, also insgesamt $20 \times 12 \times \text{CHF } 3'000 = \text{CHF } 720'000$. Falls Frau Meyers Tod mit 33 Jahren eintritt, kostet das die Versicherung $19 \cdot 12 \cdot \text{CHF } 3'000 = \text{CHF } 684'000$. Es ist aber auch gut möglich, dass sich Frau Meyer in 20 Jahren bester Gesundheit erfreut und die Versicherung alle von Frau Meyer einbezahlten Prämien behalten kann.

Für das Abschätzen der benötigten Wahrscheinlichkeiten benötigt die Versicherungsgesellschaft Sterbetafeln. Diese geben an, wie viele von 100'000 neugeborenen Mädchen oder Knaben ein bestimmtes Alter erreichen. Das könnte aussehen wie in der Tabelle rechts.

Diese besagt, dass von 100'000 neugeborenen Mädchen 99'415 ihren ersten Geburtstag erleben, 585 Mädchen sterben vorher. Die Wahrscheinlichkeit für ein Neugeborenes, vor seinem ersten Geburtstag zu sterben, ist nach dem Gesetz der grossen Zahl ungefähr gleich der relativen Häufigkeit $\frac{585}{100'000} \approx 0.006 = 0.6\%$.

Die Wahrscheinlichkeit, dass die heute 32-jährige Frau Meyer ihren 52. Geburtstag noch erlebt, wird entsprechend berechnet und beträgt $\frac{95'826}{98'345} \approx 0.974 = 97.4\%$. Die Wahrscheinlichkeit, dass die Versicherungsgesellschaft überhaupt eine Rente auszahlen muss, ist also nur gerade 2.6%.

Zur Bestimmung der Prämie stellt die Versicherungsgesellschaft selbstverständlich genauere Berechnungen an. Sie berücksichtigt, dass die auszuzahlende Rente kleiner wird, je länger Frau Meyer lebt. Sie berücksichtigt auch, dass keine Rente ausbezahlt werden muss, wenn nicht nur die Mutter innerhalb der nächsten 20 Jahre stirbt, sondern auch die Tochter. Sie berücksichtigt die Provision des Versicherungsagenten und vergisst auch die eigenen Gewinne nicht.

Die Sterbetafel muss wegen der stets wachsenden Lebenserwartung der Bevölkerung regelmässig angepasst werden. Die Sterbetafel gibt im Einzelfall – zum Beispiel bei Frau

Alter	Überlebende
0	100'000
1	99'415
...	...
32	98'345
33	98'284
34	98'221
...	...
52	95'826
...	...

1. Daten sammeln, ordnen und zählen

Meyer – natürlich keine Auskunft über den Zeitpunkt des Todes. Aber bei vielen Personen – zum Beispiel bei allen Versicherten einer grossen Gesellschaft – geben diese Sterbetafeln Auskunft darüber, wie viele Todesfälle zu erwarten sind. Die Versicherungsgesellschaft kann zum Beispiel angeben, dass 97.4% ihrer 1000 32-jährigen Kundinnen in 20 Jahren noch leben werden. Das bedeutet, dass 26 Todesfälle zu erwarten sind, und diese Zahl dürfte mit bemerkender Genauigkeit stimmen. Es geht hier letztlich wieder um das Gesetz der grossen Zahl. Dieses ermöglicht es der Versicherungsgesellschaft, recht zuverlässig zu kalkulieren.

1.13 Ergänzung: Das Paradoxon von Simpson

Das 1951 vom amerikanischen Mathematiker E. H. Simpson erstmals beschriebene Paradoxon stammt aus dem Gebiet der Statistik und zeigt, dass das nahe Liegende nicht immer richtig ist ...

1.13.1 Tuberkulose-Todesfälle in New York und Richmond (1910)

In welcher Stadt ist das Risiko grösser, an Tuberkulose zu sterben: in New York oder in Richmond?

	New York		Richmond	
	Bevölkerung	Todesfälle	Bevölkerung	Todesfälle
Weisse	4'675'174	8'365	80'895	131
Farbige	91'709	513	46'733	155
Total	4'766'883	8'878	127'628	286

Wir bilden die entsprechenden relativen Häufigkeiten. Gemäss dem Gesetz der grossen Zahl sind die gefundenen relativen Häufigkeiten Näherungen für die gesuchten Wahrscheinlichkeiten.

Zunächst berechnen wir die relative Häufigkeit der Weissen, die an Tuberkulose sterben:

$$\text{In New York: } \frac{8'365}{4'675'174} \approx 1.79\% \quad \text{In Richmond: } \frac{131}{80'895} \approx 1.62\%$$

Nun die relative Häufigkeit der Schwarzen, die an Tuberkulose sterben:

$$\text{In New York: } \frac{513}{91'709} \approx 5.59\% \quad \text{In Richmond: } \frac{155}{46'733} \approx 3.32\%$$

Zwischenbilanz: Egal ob man weisser oder schwarzer Hautfarbe ist: In New York stirbt man häufiger an Tuberkulose als in Richmond.

Nun betrachten wir die Gesamtbevölkerung der beiden Städte, also die Summe von Weissen und Schwarzen. Die relative Häufigkeit der Personen, die an Tuberkulose erkranken, beträgt:

$$\text{In New York: } \frac{8'878}{4'766'883} \approx 1.86\% \quad \text{In Richmond: } \frac{286}{127'628} \approx 2.24\%$$

Bilanz: Insgesamt stirbt man in Richmond häufiger an Tuberkulose als in New York!

Widerspricht das nicht der Zwischenbilanz? Und wo ist man nun wirklich sicherer vor Tuberkulose?

Vorsicht: Eine Gesamtdatenmenge kann andere Eigenschaften haben als alle untersuchten Teildatenmengen. Wenn man Datenmengen also zusammenfasst oder zerlegt, können sich manche Eigenschaften ändern!

1.13.2 Sterblichkeit von Männern pro Jahr (1901)

Die folgende Tabelle gibt Auskunft über das Risiko (genauer: die relative Häufigkeit dafür, gemessen in Promillen), dass ein lediger bzw. verheirateter Mann stirbt.

Alter	Ledig	Verheiratet	Insgesamt
22 – 26	6.70	3.80	6.13
27 – 31	7.80	4.19	5.89
32 – 36	8.63	4.86	5.88

Bei den Ledigen nimmt die Sterblichkeit mit dem Alter zu, bei den Verheirateten ebenso. Insgesamt aber nimmt sie mit dem Alter ab ... Ein Erklärungsversuch:

- Grundsätzlich nimmt die Sterblichkeit mit dem Alter zu: Je älter jemand ist, desto eher wird er sterben. Das gilt gleichermassen für Singles wie Verheiratete und wird durch obige Tabelle bestätigt.
- Vermutlich gehen Verheiratete aus Verantwortungsbewusstsein gegenüber Frau und Kindern weniger Risiken ein als Ledige. Das äussert sich vielleicht in der Wahl der Hobbies, der Fahrweise als Automobilist usw. Deshalb ist die Sterblichkeit bei Verheirateten geringer als bei Singles.
- Der Grund für die Abnahme der Gesamtsterblichkeit zwischen Alter 22 und 36 ist der, dass in diesem Alter viele Singles ins Lager der Verheirateten übertreten und gemäss der vorherigen Überlegung etwas vorsichtiger leben. Dadurch sinkt ihre persönliche Sterblichkeit. Und wenn viele Singles heiraten, kann durchaus die Sterblichkeit von einer Altersgruppe zur nächsten insgesamt abnehmen.

Die folgende Tabelle gibt an, wie gross der Anteil der Ledigen resp. der Verheirateten in der jeweiligen Altersklasse ist:

Alter	Ledig	Verheiratet
22 – 26	80%	20%
27 – 31	47%	53%
32 – 36	27%	73%

Da in der Altersgruppe der 32 – 36-jährigen bereits eine grosse Mehrheit verheiratet ist, dürfte in der Altersgruppe der 37 – 41-jährigen die Sterblichkeit insgesamt höher sein als bei den 32 – 36-jährigen, auch wenn noch ein paar Singles heiraten sollten.

In diesem Beispiel kommt das scheinbar paradoxe Ergebnis also vermutlich durch die Heiratsfreudigkeit der 22 – 36-jährigen zustande.

1.13.3 Auswahl von Studierwilligen

Die folgende Geschichte ereignete sich 1973 an der renommierten University of California at Berkeley. Diese Universität ist bei den Studierwilligen offenbar so begehrt, dass sie es sich leisten kann auszuwählen, welche der Studierwilligen sie aufnehmen will. Dabei darf sie jedoch keinesfalls Frauen oder Männer, Weisse oder Schwarze usw. bevorzugen.

Wir nehmen einmal an, es bewarben sich 1000 Frauen und 1000 Männer um einen Studienplatz. Aufgenommen wurden 530 Frauen, aber 640 Männer. Ein Sturm der Entrüstung brach los, *die Frauen würden benachteiligt*.

Es wurde untersucht, in welchem der beiden Fachbereiche Natur- und Gesellschaftswissenschaften die Frauen wie stark diskriminiert worden waren. Was stellte sich heraus? Im Fachbereich Naturwissenschaften wurden 70% der Männer aufgenommen und 80% der Frauen, im Fachbereich Gesellschaftswissenschaften 40% der Männer und 50% der Frauen. Also *wurden in jedem Fachbereich die Männer benachteiligt!* Die Zahlen:

1. Daten sammeln, ordnen und zählen

		Fachbereich Naturwissen- schaften	Fachbereich Gesellschafts- wissenschaften	Ganze Universität
Studierwillige	Männer	800	200	1000
	Frauen	100	900	1000
Aufgenommene, Anteil an den Stu- dierwilligen	Männer	560 = 70%	80 = 40%	640 = 64%
	Frauen	80 = 80%	450 = 50%	530 = 53%

Rechnen Sie nach!

1.13.4 Hinweise zum Simpson-Paradoxon

Was passiert, wenn man die Zahlen für die einzelnen Fachbereiche zusammenzählt, um die Zahlen für die gesamte Universität zu berechnen?

Im Bereich der absoluten Häufigkeiten hat alles seine Richtigkeit: $560 + 80 = 640$, $80 + 450 = 530$.

Im Bereich der relativen Häufigkeiten wird es spannender: Wenn 560 von 800 Männern in den naturwissenschaftlichen Fachbereich aufgenommen werden und 80 von 200 Männern in den gesellschaftswissenschaftlichen, dann werden insgesamt $(560+80)$ von $(800+200)$ an der Universität aufgenommen. Hier wird also wie folgt gerechnet:

$$\frac{560}{800} + \frac{80}{200} = \frac{560+80}{800+200} = \frac{640}{1000}$$

Man rechnet also Zähler + Zähler, Nenner + Nenner! Diese Art der Addition ist in diesem Zusammenhang sicher richtig, weicht aber von der gebräuchlichen Art ab, zwei Brüche zu addieren. Zur Unterscheidung von der gewöhnlichen Addition mit dem Zeichen $+$ verwenden wir für diese aussergewöhnliche Addition das Zeichen \oplus . Es ist also

$$\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

(gewöhnliche Addition), aber

$$\frac{1}{2} \oplus \frac{1}{4} = \frac{2}{6} = \frac{1}{3}$$

(aussergewöhnliche Addition.) Das obige Beispiel muss natürlich mit dem Zeichen \oplus geschrieben werden:

$$\frac{560}{800} \oplus \frac{80}{200} = \frac{560+80}{800+200} = \frac{640}{1000}.$$

Für die Addition \oplus gelten andere Regeln als für die Addition $+$. Und genau aus dieser nicht offensichtlichen Verschiedenheit der Rechenregeln für $+$ und \oplus entsteht das Paradoxon von Simpson.

Für Interessierte zwei Beispiele für die Verschiedenheit der Rechenregeln:

- Sind h_1 und h_2 zwei relative Häufigkeiten, dann gilt einerseits

$$h_1 + h_2 \geq h_1 \quad \text{und} \quad h_1 + h_2 \geq h_2,$$

weil sowohl h_1 als auch h_2 grösser oder gleich Null sind. Die Summe $h_1 + h_2$ ist also mindestens so gross wie h_1 und mindestens so gross wie h_2 .

Andererseits kann man beweisen, dass $h_1 \oplus h_2$ zwischen h_1 und h_2 liegt:

$$h_1 \leq h_1 \oplus h_2 \leq h_2 \quad \text{oder} \quad h_1 \geq h_1 \oplus h_2 \geq h_2.$$

Illustration: $\frac{640}{1000} = 0.64$ liegt tatsächlich zwischen $\frac{560}{800} = 0.7$ und $\frac{80}{200} = 0.4$. Die „Summe“ $h_1 \oplus h_2$ ist also nur noch mindestens so gross wie h_1 oder h_2 .

- Für beliebige Brüche gilt bei der normalen Addition + die Regel, dass man gleichartige Ungleichungen addieren darf und wieder eine richtige Ungleichung erhält. Wenn

$$h_1 > h_1^*$$

und

$$h_2 > h_2^*$$

ist, dann folgt daraus

$$h_1 + h_2 > h_1^* + h_2^*.$$

Bei der speziellen Addition \oplus braucht diese Regel nicht mehr unbedingt zu gelten. Für h_1, h_1^*, h_2 und h_2^* kann man gemäss Beispiel 1.13.3 folgende Zahlen wählen:

$$h_1 = \frac{80}{100}, h_1^* = \frac{560}{800}, h_2 = \frac{450}{900}, h_2^* = \frac{80}{200}.$$

Diese Zahlen erfüllen die Bedingungen $h_1 > h_1^*$ und $h_2 > h_2^*$. Bei der speziellen Addition \oplus erhält man aber

$$h_1 \oplus h_2 = \frac{530}{1000}$$

und

$$h_1^* \oplus h_2^* = \frac{640}{1000}.$$

Also ist in diesem Beispiel im Gegensatz zur gewöhnlichen Addition $h_1 \oplus h_2 < h_1^* \oplus h_2^*$.

Es soll nicht im Detail untersucht werden, wann das Simpson-Paradoxon eintritt und wann nicht. Es wird jedoch begünstigt, wenn zwei oder mehrere Teilmengen mit unterschiedlichen Eigenschaften zu einer Gesamtmenge zusammengefasst werden. In diesem Beispiel sind die beiden Teilmengen die männlichen und weiblichen Studierwilligen, und die unterschiedlichen Eigenschaften sind die gewünschten Studienrichtungen: Männer ziehen die Naturwissenschaften den Gesellschaftswissenschaften deutlich vor, und bei den Frauen ist es gerade umgekehrt.

Wer es noch genauer wissen möchte und über Vorkenntnisse in Wahrscheinlichkeitsrechnung verfügt, findet eine Antwort auf die Frage, wann das Simpson-Paradoxon eintritt, in Quelle [W4].

1.14 Taschenrechner

TI-89 / TI-92 Plus / Voyage 200

<p>Eine (Ur-)Liste eingeben und speichern</p>	<p>Speichern Sie die beiden Stichproben 6, 5, 3, 3, 5, 2 und a, b, c:</p> <p>{ 6, 5, 3, 3, 5, 2 } [STO] liste1 [ENTER] → {6 5 3 3 5 2} {a, b, c} [STO] liste2 [ENTER] → {a b c}</p>
<p>Eine Liste sortieren Aufsteigend sortieren</p> <p>Absteigend sortieren</p>	<p>Sortieren Sie die beiden Listen aus dem letzten Abschnitt aufsteigend:</p> <p>sorta liste1 [ENTER] → Done liste1 [ENTER] → {2 3 3 5 5 6} sorta liste2 [ENTER] → Done liste2 [ENTER] → {a b c}</p> <p>Sortieren Sie die beiden Listen aus dem letzten Abschnitt absteigend:</p>

1. Daten sammeln, ordnen und zählen

	<pre>sortd liste1 [ENTER] → Done liste1 [ENTER] → {6 5 5 3 3 2} sortd liste2 [ENTER] → Done liste2 [ENTER] → {c b a}</pre>
Zusammensetzen und Zerlegen von Listen Zwei Listen zusammensetzen Ein Listenelement abrufen Die ersten Elemente einer Liste angeben Die letzten Elemente einer Liste angeben Einen Auszug aus einer Liste angeben Die Länge der Liste abfragen	<pre>Hängen Sie an die Liste {6, 5, 3, 3, 5, 2} die Liste {4, 5, 6} und speichern Sie das Resultat in liste3: augment({6, 5, 3, 3, 5, 2}, {4, 5, 6}) [STO] liste3 [ENTER] → {6 5 3 3 5 2 4 5 6} Welches ist das 7. Element von liste3? liste3[7] [ENTER] → 4 Welches sind die ersten 4 Elemente von liste3? left(liste3, 4) [ENTER] → {6 5 3 3} Welches sind die letzten 4 Elemente von liste3? right(liste3, 4) [ENTER] → {2 4 5 6} Nenne vom 2. Element an 4 Elemente von liste3: mid(liste3, 2, 4) [ENTER] → {5 3 3 5} Wie viele Elemente enthält liste3? dim(liste3) [ENTER] → 9</pre>
Eine Stichprobe mit Klasseneinteilung eingeben und speichern	<pre>Bei einer Stichprobe treten die Ereignisse bzw. Klassen {1, 2, 3, 4, 5, 6} mit den Häufigkeiten {0, 1, 2, 0, 2, 1} auf: {1, 2, 3, 4, 5, 6} [STO] klassen [ENTER] → {1 2 3 4 5 6} {0, 1, 2, 0, 2, 1} [STO] oft [ENTER] → {0 1 2 0 2 1} Bei einer Klasseneinteilung sind bei der Liste klassen jeweils die Klassenmitten anzugeben.</pre>

1.15 Übungen

A. Daten sammeln und ordnen

- Geben Sie bei jedem der folgenden Merkmale seinen Typ an:
 - Rückennummer eines Fussballers
 - Schuhgröße
 - Haarfarbe
 - Hautfarbe
 - Augenfarbe
 - Beruf
 - Wohnort
 - Schulbildung
 - Leistung im Hochsprung
 - Körpergewicht
 - Ferienziel
 - Name des Partners
 - Zeugnisnote in Deutsch
 - Fehlerzahl in der letzten Prüfung
 - Automarke
 - gegenwärtige Temperatur
 - Postleitzahl des Wohnorts
 - Beliebtheit (klein, mittel, gross)
 - Jahreseinkommen
 - Höhe eines Berggipfels über Meer
 - Zahl der Bücher in einem Gestell
 - Stilnote in einem Aufsatz
- Geben Sie je 3 Merkmale an, die
 - quantitativ / stetig
 - quantitativ / diskret
 - qualitativ / ordinal
 - qualitativ / nominalsind.

3. Angenommen, ein Wirt führe Buch darüber, welcher Wein von den Gästen wie oft getrunken wird. Welche Variablen wird er dann erheben, und von welchem Typ sind diese Variablen?
4. Welche der folgenden Aussagen sind richtig?
 - a) Eine Grundgesamtheit liegt dann vor, wenn man nicht alle zugehörigen Elemente statistisch erfassen kann.
 - b) Qualitative Merkmale sind nominal oder ordinal.
 - c) Diskrete Merkmale können nur ganzzahlige Werte annehmen.
5. Klassen bilden:
Bei den folgenden Aufgaben sollen jeweils n Werte, deren kleinster x_{\min} und deren grösster x_{\max} ist, sinnvoll in Klassen eingeteilt werden. Legen Sie eine vernünftige Klasseneinteilung fest durch Angabe der Untergrenze der ersten Klasse und der Klassenbreite.

a) $n=50, x_{\min}=7, x_{\max}=58$	b) $n=250, x_{\min}=7, x_{\max}=58$
c) $n=20, x_{\min}=41, x_{\max}=96$	d) $n=37, x_{\min}=36.5, x_{\max}=37.4$
e) $n=4'380'000, x_{\min}=0, x_{\max}=75$	
6. Weitsprung:
Am Sporttag erzielte eine Klasse im Weitsprung folgende Leistungen [in cm] :
452, 514, 372, 502, 401, 406, 350, 564, 605, 375, 423, 564, 649, 589, 465, 489, 392, 424, 498, 482, 513, 567, 456, 565, 633
Fassen Sie diese Leistungen zusammen in
a) 10 b) 6 c) 5 d) 3
gleich breiten Klassen.
7. Sport in der Freizeit:
Die Studierenden einer Klasse geben an, wie viel ihrer Freizeit sie wöchentlich für sportliche Aktivitäten einsetzen. Hier die Antworten in Minuten:
0, 150, 320, 745, 215, 0, 30, 40, 70, 120, 0, 90, 160, 210, 220, 360, 480, 0, 90, 100.
Fassen Sie diese Angaben zusammen in
a) 15 gleich breiten Klassen,
b) 5 gleich breiten Klassen,
c) folgender Klasseneinteilung mit unterschiedlicher Klassenbreite:
 $0 \leq x < 30, 30 \leq x < 60, 60 \leq x < 120, 120 \leq x < 180, 180 \leq x < 240, 240 \leq x < 360, 360 \leq x < 780$
8. Erheben Sie von allen Mitstudierenden die Körpergrösse, und fassen Sie die Resultate zusammen in
a) 10 b) 6 c) 4
gleich breiten Klassen.
9. Messen Sie bei sich und allen Mitstudierenden den Puls, und fassen Sie die Resultate zusammen in
a) 10 b) 6 c) 4 d) 3
gleich breiten Klassen. Welche Klasseneinteilung ist am sinnvollsten?

B. Absolute und relative Häufigkeit ohne Klasseneinteilung

1. Bei welchen Aussagen liegen absolute Häufigkeiten vor, wo relative?
 - a) Verunfallte Personen in der Schweiz pro Jahr ...
... durch Umhergehen in Haus und Garten: 157'400,

1. Daten sammeln, ordnen und zählen

- b) ... durch Zubereitung oder Einnahme von Mahlzeiten: 15'850
- c) ... durch Haushaltapparate, Steckdosen, Kabel: 450
- d) Personen pro Haushalt in der Schweiz: 2.4; im Iran: 4.8
- e) Haushalte mit mehr als 5 Personen in der Schweiz: 1.7%; im Iran: 34%
- f) Frauen in Prozent, die 1968 täglich die Unterhose wechselten: 59
- g) Männer in Prozent, die 1968 täglich die Unterhose wechselten: 5
- h) Frauen in Prozent, die 1988 täglich die Unterhose wechselten: 70
- i) Männer in Prozent, die 1988 täglich die Unterhose wechselten: 45
- j) WC-Benutzende in Prozent, die das Toilettenpapier vor Gebrauch ...
... falten: 40
- k) ... knüllen: 45
- l) ... um die Hand wickeln: 15
- m) Schweizer in Prozent, die im Stehen pinkeln: 72
- n) Schweizerinnen und Schweizer in Prozent, die auf der Toilette lesen: 40
- o) Österreicher, die beim Nachhausekommen denken: „Hoffentlich ist das Abendessen schon fertig“: die Hälfte,
- p) ... dasselbe bei deutschen Männern: 28%
- q) ... dasselbe bei Schweizer Männern: 19%

[Quelle: NZZ Folio, Februar 2003]

2. Blutgruppen:

Beim Blutspenden wird bei jedem Spender und jeder Spenderin die Blutgruppe bestimmt. Bestimmen Sie die absolute und relative Häufigkeit jeder Blutgruppe gemäss der folgenden Strichliste:

A	
B	
AB	
O	

3. Tour de France:

Zwischen 1952 und 2001 endeten 22 Etappen der Tour de France auf Alpe d'Huez. Um das Ziel zu erreichen, müssen die Radfahrer den letzten 13 Kilometern 1128 Höhenmeter überwinden. Der Italiener Marco Pantani benötigte dafür 1997 nur gerade 37:35 Minuten.

Von den 22 Etappensiegen gingen 8 an einen Niederländer, 7 an einen Italiener, 2 an einen US-Amerikaner und je einer an einen Franzosen, Kolumbianer, Portugiesen, Schweizer und Spanier.

Berechnen Sie für jedes Land die relative Häufigkeit, mit der es auf Alpe d'Huez den Sieger stellte.

[Quelle: SonntagsZeitung, 13.7.2003]

4. Zusammensetzung der Milch:

400 g Vollmilch enthalten ca. 19.6 g Kohlenhydrate, 15.6 g Fett, 12.8 g Eiweiss; der Rest ist Wasser. Berechnen Sie die relative Häufigkeit jedes Milchbestandteils.

5. Umfrage:

Im Juli 2002 nahmen 1300 Personen Stellung zur folgenden Frage: „Haben Sie nach der Pannenserie noch Vertrauen in die Swiss?“ 37% sagten Ja, 63% Nein.

- a) Wie viele Personen antworteten mit Ja, wie viele mit Nein?
- b) Was meinen Sie zur Formulierung der Frage?

[Quelle: SonntagsBlick, 4. August 2002]

6. Zahl der Kinder I:

Eine Untersuchung ergab, dass die Familien einer Stadt folgende Kinderzahlen haben:

Kinderzahl	0	1	2	3	4	5
Familien mit ... Kindern (absolute Häufigkeit)	120	150	200	90	50	15

- a) Wie viele Familien wohnen in dieser Stadt?
 - b) Wie viele Kinder wohnen in dieser Stadt?
- Welches ist die relative Häufigkeit aller Familien mit
- c) genau einem Kind?
 - d) höchstens einem Kind?
 - e) mindestens 3 Kindern?
 - f) mindestens 2 und höchstens 4 Kindern?

7. Zahl der Kinder II:

In einer Stadt hat man untersucht, wie kinderreich die Familien sind. Resultate:

Kinderzahl	0	1	2	3	4	ab 5
Familien mit ... Kindern (rel. Häufigkeit)	?	0.25	0.3	0.15	0.06	0.04

100 Familien sind kinderlos.

- a) Wie viele Familien wohnen in dieser Stadt?
- b) Was können Sie über die Zahl der Kinder in dieser Stadt aussagen?

8. AIDS-Test:

In der Schweiz wohnen rund 7'000'000 Menschen. Wir nehmen an, dass davon 0.3% mit dem HI-Virus infiziert sind. Aufgrund von medizinischen Untersuchungen weiss man:

Wenn jemand mit dem HI-Virus infiziert ist, erkennt dies ein bestimmter Test in 999 von 1000 Fällen, d.h. mit der relativen Häufigkeit 0.999. Wenn jemand nicht mit dem HI-Virus infiziert ist, erkennt dies der Test in 997 von 1000 Fällen, also mit der relativen Häufigkeit 0.997.

- a) Wie viele Menschen in der Schweiz sind mit dem HI-Virus infiziert, und wie viele Menschen sind es nicht?
- b) Wir wenden uns den Menschen zu, die das HI-Virus in sich tragen. Wie viele von ihnen werden vom Test als Infizierte erkannt, und wie viele werden fälschlicherweise nicht als Infizierte erkannt?
- c) Nun zu den Menschen, die nicht HIV-Infizierte sind. Wie viele werden von diesem Test als gesund erkannt, und wie viele werden zu Unrecht als HIV-Infizierte ausgewiesen?
- d) Nun betrachten wir jene Menschen, welche *gemäss dem AIDS-Test* HIV-infiziert sind. Wie gross ist die relative Häufigkeit derjenigen, welche tatsächlich infiziert sind, und wie gross ist die relative Häufigkeit derjenigen, die zu Unrecht die erschreckende Diagnose erhalten „Sie sind mit dem HIV-Virus infiziert“?

C. Absolute und relative Häufigkeit mit Klassenbildung

- 1. Beim Weitsprung wurden in einer Klasse die nachfolgend angegebenen Leistungen erzielt. Berechnen Sie die relative Häufigkeit jeder Klasse.

a)

$350 \leq x < 380$	III
$380 \leq x < 410$	III

$410 \leq x < 440$	II
$440 \leq x < 470$	III
$470 \leq x < 500$	III

1. Daten sammeln, ordnen und zählen

$500 \leq x < 530$	III
$530 \leq x < 560$	
$560 \leq x < 590$	IIII
$590 \leq x < 620$	I
$620 \leq x < 650$	II

$470 \leq x < 530$	IIII I
$530 \leq x < 590$	IIII
$590 \leq x < 650$	III

b)

$350 \leq x < 400$	IIII
$400 \leq x < 450$	IIII
$450 \leq x < 500$	IIII I
$500 \leq x < 550$	III
$550 \leq x < 600$	IIII
$600 \leq x < 650$	III

d)

$350 \leq x < 450$	IIII III
$450 \leq x < 550$	IIII III
$550 \leq x < 650$	IIII III

c)

$350 \leq x < 410$	IIII I
$410 \leq x < 470$	IIII

2. Sport in der Freizeit:

Die Studierenden einer Klasse geben an, wie viel ihrer Freizeit sie wöchentlich für sportliche Aktivitäten einsetzen. Nachfolgend die Antworten in Minuten. Berechnen Sie die relative Häufigkeit jeder Klasse.

a)

Klasse	Häufigkeit
$0 \leq x < 50$	IIII I
$50 \leq x < 100$	III
$100 \leq x < 150$	II
$150 \leq x < 200$	II
$200 \leq x < 250$	III
$250 \leq x < 300$	
$300 \leq x < 350$	I
$350 \leq x < 400$	I
$400 \leq x < 450$	
$450 \leq x < 500$	I
$500 \leq x < 550$	
$550 \leq x < 600$	
$600 \leq x < 650$	
$650 \leq x < 700$	
$700 \leq x < 750$	I

b)

Klasse	Häufigkeit
$0 \leq x < 150$	IIII IIII I
$150 \leq x < 300$	IIII
$300 \leq x < 450$	II
$450 \leq x < 600$	I
$600 \leq x < 750$	I

c)

Klasse	Häufigkeit
$0 \leq x < 30$	IIII
$30 \leq x < 60$	II
$60 \leq x < 120$	IIII
$120 \leq x < 180$	III
$180 \leq x < 240$	III
$240 \leq x < 360$	I
$360 \leq x < 780$	III

D. Das Gesetz der grossen Zahl

1. Münzenwerfen I:
 - a) Werfen Sie eine Münze 60-mal, und notieren Sie sich das Ergebnis K (=Kopf) oder Z (=Zahl). Wie oft haben Sie Kopf geworfen, wie oft Zahl? Mit welcher absoluten Häufigkeit und mit welcher relativen Häufigkeit haben Sie Kopf geworfen?
 - b) Wie oft haben die Studierenden, die in derselben Bankreihe sitzen wie Sie, Kopf bzw. Zahl geworfen? Mit welcher absoluten Häufigkeit und mit welcher relativen Häufigkeit haben Sie und Ihre Mitstudierenden in derselben Bankreihe Kopf geworfen?
 - c) Beantworten Sie Frage b) für die ganze Klasse.
 - d) Wenn beim Münzenwerfen nacheinander zwei verschiedene Ergebnisse erzielt wurden (KZ oder ZK), nennen wir dies einen Wechsel. Wie viele Wechsel erwarten Sie in 60 Versuchen? Welches ist die erwartete relative Häufigkeit eines Wechsels?
 - e) Welches ist die tatsächlich aufgetretene relative Häufigkeit der Wechsel – bei Ihnen, bei Ihnen und Ihren Mitstudierenden derselben Bankreihe, bei der ganzen Klasse?
2. Münzenwerfen II:
 - a) Simulieren Sie eine Münze, indem Sie zufällig insgesamt 60 Z und K notieren.
 - b) Zählen Sie die Wechsel (→ Aufgabe D.1.d)) in Ihrer Simulation.
 - c) Wenn viermal nacheinander Kopf (und nachher Zahl) erschien, oder wenn viermal nacheinander Zahl (und nachher Kopf) erschien, nennen wir dies einen Viererblock.
Zählen Sie die in Ihrer simulierten Wurfserie von Aufgabe a) aufgetretenen Einer-, Zweier-, Dreier-, Vierer-, Fünfer-, Sechser-, Siebner-, Achter-, Neuner- und Zehnerblöcke.
 - d) Zählen Sie die in Ihrer tatsächlich durchgeführten Wurfserie von Aufgabe 1. a) aufgetretenen Einer-, Zweier-, Dreier-, Vierer-, Fünfer-, Sechser-, Siebner-, Achter-, Neuner- und Zehnerblöcke.
 - e) Vergleichen Sie die Resultate von c) und d).
 - f) Führen Sie die Aufgaben c), d) und e) für die ganze Klasse durch.
3. Würfeln:

Werfen Sie einen Würfel so oft, bis eine durch 3 teilbare Zahl erscheint, und zählen Sie die Anzahl der dazu notwendigen Würfe.

Führen Sie dieses Experiment in verschiedenen Schülergruppen je 20 mal durch.

- a) Mit welcher relativen Häufigkeit erscheint eine Dreierzahl in Ihrer Schülergruppe im 1., 2., 3., 4., 5., 6. usw. Wurf?
- b) Wie viele Würfe sind im Durchschnitt nötig bis zur ersten Dreierzahl?
- c) Beantworten Sie die Fragen a) und b), indem Sie die Resultate der einzelnen Schülergruppen zu einer Untersuchung zusammenfassen, und vergleichen Sie Ihre Resultate mit den theoretisch errechneten Wahrscheinlichkeiten (siehe Tabelle nebenan).

Die Dreierzahl erscheint im ...	Wahrscheinlichkeit (=ideale relative H.)
1. Wurf	0.3333
2. Wurf	0.2222
3. Wurf	0.1481
4. Wurf	0.0988
5. Wurf	0.0656
6. Wurf	0.0439
7. Wurf	0.0293
8. Wurf	0.0195
9. Wurf	0.0130
10. Wurf	0.0086

1. Daten sammeln, ordnen und zählen

4. Buchstabenhäufigkeiten I:
Bestimmen Sie in einem deutschen Text die relative Häufigkeit
 - a) der Vokale,
 - b) der Konsonanten,
 - c) der Umlaute.
 - d) Welches sind die 7 häufigsten Buchstaben der deutschen Sprache – abgesehen vom Leerzeichen?
 - e) Ein mögliches Verschlüsselungsverfahren für Text besteht darin, jeden Buchstaben durch einen anderen zu ersetzen, aber stets durch denselben. Wie kann ein so verschlüsselter Text geknackt werden?
5. Buchstabenhäufigkeiten II:
Bestimmen Sie in einem fremdsprachigen Text die relative Häufigkeit
 - a) der Vokale,
 - b) der Konsonanten.
 - c) Welches sind die 7 häufigsten Buchstaben der untersuchten Sprache – abgesehen vom Leerzeichen?
6. Zahlenlisten studieren:
 - a) Nehmen Sie eine Tageszeitung zur Hand und notieren Sie sich von *jeder* Zahl, auf die Sie stossen – also von der Seitenzahl, von den Aktienkursen, vom Verkaufs- und Abonnementspreis usw. - die erste Ziffer. Bestimmen Sie anschließend die relative Häufigkeit, mit der als erste Ziffer die 1 erscheint, die 2, 3, 4 usw.
 - b) Führen Sie dasselbe Experiment mit einer Formelsammlung durch, mit einem geografischen Nachschlagewerk usw.
Es ist verblüffend festzustellen, dass Zahlen, die mit kleinen Ziffern beginnen, sehr viel häufiger auftreten als Zahlen, die mit grossen Ziffern beginnen.