Frank Siegmann

Statistik verstehen, nicht rechnen

Band 1: Beschreibende Statistik

Kohlhammer

150 Jahre
Kohlhammer

Frank Siegmann

Statistik verstehen, nicht rechnen

Band 1: Beschreibende Statistik

Verlag W. Kohlhammer

Für Marvin, Barbara und Simon

1. Auflage 2017

Gesamtherstellung: W. Kohlhammer GmbH, Stuttgart

Print:
ISBN 978-3-17-031013-1

E-Book-Formate:
pdf: ISBN 978-3-17-031014-8
epub: ISBN 978-3-17-031015-5
mobi: ISBN 978-3-17-031016-2

Inhaltsverzeichnis

Ein Vorwort, das man lesen sollte

Noch ein Buch über Statistik. Muss das sein? Muss man sich mit Statistik beschäftigen?

Vor mehreren tausend Jahren war die Antwort einfach: Nein! Erste Statistiken beschäftigten sich vor allem mit der Wehrfähigkeit der Untertanen. Der König (Fürst, Feldherr, Emir) wollte wissen, wie viele Söldner er rekrutieren kann. Später wollten die Oberen wissen, wie viele Bürger es gibt und wer wie viel Steuern zahlen muss. Viele Jahrhunderte oblag es also dem Staat, Statistiken zu erheben und zu pflegen.

Seit einigen Jahrzehnten ist die Antwort aber komplizierter geworden. Nicht zuletzt durch den Einsatz der elektronischen Datenverarbeitung und des Internets sind Zahlen nahezu unbegrenzt verfügbar. Wir werden mit Statistiken überhäuft; es gibt scheinbar nichts, wofür man nicht gleich eine passende heranziehen kann, und kein Medium, in dem nicht zahlreiche Statistiken zitiert oder interpretiert werden. Man verfolge nur einmal eine Fussballübertragung im Fernsehen, da wird berichtet von der stärksten Heimmannschaft der Rückrunde oder in diesem Jahr, dem ersten Sieg seit 15 Spielen, den vielen Minuten ohne Torschuss, den drei Unentschieden bei Flutlicht oder dass noch nie eine rote Karte gegen Spieler A verhängt wurde, wenn der Trainer B hieß. Zu oft haben wir nun den Eindruck, dass eine Statistik nicht zu uns passt oder wie bei den Sportübertragungen willkürlich ist. Offenbar leidet die Glaubwürdigkeit der Statistik unter dieser Masse an Informationen. Uns werden immer mehr Zahlen angeboten, die uns aber immer weniger glaubhaft erscheinen. Viele Nutzer misstrauen deshalb statistischen Aussagen, was in überspitzten Formulierungen wie »Traue keiner Statistik, die Du nicht selbst gefälscht hast« oder »Die Steigerung von Lüge und Notlüge ist Statistik« ihren Ausdruck findet.

Diese kritische Haltung liegt sicher zum großen Teil darin begründet, dass unsere Umwelt immer komplexer geworden ist; so differenziert, dass sie sich nicht mehr wie früher in einfache Statistiken pressen lässt; es gibt eben nicht mehr nur »wehrfähig oder nicht«.

Ein Beispiel unter vielen ist die Berechnung des Preisniveaus. Statt einer Kennziffer für die Lebenshaltungskosten in Deutschland müsste es mehrere geben, getrennt für Männer und Frauen, Arbeiter und Rentner (die in vielen Bereichen sehr unterschiedliche Güter konsumieren), eine für Singles und Paare (die für Einzel- und Doppelzimmer sehr unterschiedliche Preise berappen müssen), für Kinderlose und Eltern (die Windeln, Kindersitze, Kinderwagen und Kinderbetten kaufen, später als Rentner und Eltern sogar beides), für Autofahrer (denen der Benzinpreis aufs Gemüt schlägt) und diejenigen, die den ÖPNV nutzen (und die dortigen regelmäßigen Preissteigerungen verkraften müs-

sen), für Vegetarier, Bioladen- und Großmarktkunden, für Hauseigentümer (mit Kreditkosten) und Mieter.

Müsste nicht auch die Lebenserwartung der Deutschen differenziert ausgewiesen werden, getrennt für Männer und Frauen (wird sie), für jetzt Neugeborene und schon 50jährige (wird sie, sogar für jedes denkbare Alter), für Raucher und Nichtraucher (das wird sie, wie auch in den folgenden Abgrenzungen, aber nicht), für Schwerstarbeiter oder Büroangestellte, für Menschen mit höherem oder niedrigeren Bildungsabschluss, für Heimatliebende und Vielreisende (Probleme Malaria, Dengue-Fieber)? Und müsste man nicht das mögliche spätere Einkommen von Studierenden in Abhängigkeit vom abgelegten Studienfach, dem Abschluss als Bachelor oder Master, der Hochschule (Universität, Fachhochschule, Berufsakademie) und dem Bildungsniveau vor dem Studium (Berufskolleg, Gymnasium, Gesamtschule, Meisterprüfung, Abendschule, private oder öffentliche Einrichtung) betrachten? Alles Möglichkeiten, die sich beträchtlich auf das erreichbare Einkommen auswirken können.

Je komplexer also unsere Umwelt ist, desto schwieriger ist sie durch »eine« Statistik abzubilden. Zu oft passt eine Statistik nicht zu den individuellen Lebensverhältnissen, womit sie dann schnell als überflüssig oder sogar als falsch abgetan wird. Wir fragen uns, ob diese Statistik auch auf die eigene Person zutrifft. Wer nicht raucht, täglich Sport treibt und sich gesund ernährt, findet sich in einer Statistik, in der die durchschnittliche Lebenserwartung aller Bürger angegeben wird, nicht wieder (oder will sich nicht wiederfinden). Vielleicht haben wir uns auch eine kritische Haltung angewöhnt, die durchaus ihre Berechtigung hat (denn wie viele unsinnige medizinische Ratschläge findet man im Internet?) und überträgt diese Vorbehalte pauschal auf jede Statistik, die uns nicht passt oder nicht unseren Erwartungen entspricht. Wer z. B. viel raucht, ignoriert die Statistik zur Lebenserwartung, hält sie sogar für falsch und verweist auf die Raucher, die dennoch sehr alt geworden sind. »Altbundeskanzler Helmut Schmidt (1918 - 2015) hat geraucht und wurde 96« ist die Standardantwort vieler Raucher.

Es liegt auch daran, dass Statistiken zu politischen oder gesellschaftspolitischen Themen zwingend der Meinung eines Teils der Bevölkerung widersprechen. Die Aussage »5 % der Menschen sind homosexuell« stößt schon deshalb bei einigen auf strikte Ablehnung, weil sie Homosexualität nicht mögen. »In der Rangfolge der korrupten Länder liegt Deutschland auf Platz 13 (1 bedeutet geringste Korruption)«, diese Aussage wird von denen als Unsinn eingestuft, die Korruption noch nicht erlebt haben und sie deshalb nicht als Problem ansehen. Die in diesen Studien im Wesentlichen bemängelte Bestechung und die fehlende Transparenz der Nebeneinkünfte der Abgeordneten hat für den Normalbürger schließlich keine unmittelbare Relevanz. Wer Ausländer nicht mag, glaubt eher der Statistik, nach der bestimmte Migrationsgruppen straffälliger sind als andere und überträgt die Ergebnisse auf »den Ausländer« an sich. Es wird immer Statistiken geben, die ein Einzelner bezweifelt, und die damit für ihn ein Beispiel für den Unsinn statistischer Aussagen insgesamt sind.

Hinzu kommt: Informationen sind immer schneller und überall verfügbar. Wir sind es gewohnt, in elektronischen Nachschlagewerken oder mit Hilfe von Suchmaschinen scheinbar jede Information schnell abzurufen, machen uns aber nicht mehr die Mühe, die Quelle der Statistik zu hinterfragen. Was oder wer wurde gemessen? Wann wurde

gemessen? Konnten alle Betroffenen in der Statistik erfasst werden? Ist die Polizeiliche Kriminalstatistik (PKS) geeignet, die Frage nach der Ausländerkriminalität zu beantworten? Denn diese Statistik zählt Tatverdächtige, nicht verurteilte Täter, sie zählt als Ausländer diejenigen ohne deutschen Pass, also nicht die mit Migrationshintergrund, und sie stellt fest, dass vor allem junge Männer in Großstädten strafverdächtig sind, und dorthin zieht es vor allem auch Ausländer. Kein Wunder, dass solche Statistiken dann oft nicht auf den Sachverhalt zutreffen, den man beurteilen wollte. Letztendlich ist es bequemer, Statistiken abzulehnen, als sich die Mühe zu machen, über ihr Zustandekommen nachzudenken. Natürlich wirkt es nicht gerade vertrauensbildend, wenn Medien zusätzlich angewiesen werden, die Nationalität von Straftätern in ihren Artikeln zu verschweigen.

Außerdem interessieren wir uns häufig erst dann intensiver für Statistiken, wenn wir in einer aktuellen Situation Orientierung suchen. Eine Statistik über die Überlebenschance im Falle einer schweren Krankheit beschäftigt uns. Die Lebenserwartung im Allgemeinen interessiert uns (und die kreditgebende Bank), wenn wir noch einen Kredit über 30 Jahre für ein neues Haus aufnehmen wollen. Denn uns allen ist bekannt, dass wir nur etwa 70, 80 Jahre leben werden und die Zeit für die Rückzahlung des Kredites knapp werden kann. Lohnt sich die private Altersvorsorge, wenn die demografische Entwicklung so weitergeht wie bisher? Wird die nahegelegene Grundschule geschlossen, weil immer weniger Kinder unterrichtet werden müssen? Muss ich mein Kind also demnächst durch die halbe Stadt zum Unterricht kutschieren? Bei welchem Professor sind die Klausuren einfach und die Durchfallquoten gering? Das sind durchaus Statistiken, die, manchmal eher unbewusst, auch Ihr Interesse wecken.

Und noch ein Problem verschärft sich. Immer weniger kann man Menschen für überwiegend formale Denkweisen und die intensive und langwierige Beschäftigung mit einem Thema begeistern. Und da die Statistik oftmals sehr abstrakt mit vielen Formeln und Rechenvarianten daher kommt, möchte man sich lieber nicht mit ihr beschäftigen. Und leider ist es schick geworden, von Mathematik oder Statistik keine Ahnung zu haben.

Vor diesem Hintergrund ist die Idee entstanden, ein sehr einfach gehaltenes Buch zur Statistik zu schreiben, das auch ohne Kenntnisse der Mathematik lesbar und verständlich bleibt. Und das vor allem die oben beschriebene kritische Haltung aufgreift. Eine Einführung zur Statistik, die sich nicht nur an die Studierenden aller Fachrichtungen richtet, sondern auch den interessierten Laien ansprechen möchte, der »schon immer mal etwas Statistik verstehen wollte«. Oder es bisher vergeblich versucht hat, Statistik zu verstehen.

Ich werde daher in dieser Einführung zeigen, dass für das Verständnis der statistischen Methodenlehre die Auseinandersetzung mit komplizierten mathematischen Ansätzen, vielen Formeln und Rechenvarianten nicht zwingend notwendig ist. Fast ohne Formeln und ohne Taschenrechner werden wir uns mit dem Thema auseinandersetzen. Versprochen! Für den, der dann doch einmal mit komplexeren Zahlen rechnen möchte, werden im Anhang zusammenfassende Aufgaben besprochen; erst diese können nur mit Hilfe eines Taschenrechners gelöst werden.

Wir wollen uns dem Themengebiet also mit unserem gesunden Menschenverstand nähern und es auf diesem Wege erschließen. Es geht um das grundsätzliche Verstehen, nicht um das Rechnen. Daher verspreche ich auch, dass jeder versteht, was er hier liest. Ich kann aber nicht versprechen, dass man Statistik mal so eben in wenigen Minuten, quasi nebenbei, begreift.

Also: Butter bei die Fische. Lassen Sie sich auf das Thema ein und beurteilen Sie, ob ich zu viel versprochen habe.

Mein ältester Sohn hat das Ganze Korrektur gelesen und mir zahlreiche Hinweise gegeben. Dafür danke ich ihm herzlich. Und meiner Frau und meinem kleinen Sohn danke ich für die ungestörte Ruhe in meinem Arbeitszimmer. Insbesondere möchte ich mich bei Herrn Dr. Fliegauf vom Verlag Kohlhammer bedanken, der dieses Projekt sofort mit Begeisterung aufnahm, das Manuskript intensiv gelesen hat und zahlreiche Anregungen und Verbesserungsvorschläge einbrachte.

Bochum, im Mai 2016 Frank Siegmann

0 Statistik ist überall – eine Einführung

Das Korbballmatch neigt sich dem Ende
die Heimmannschafft erhofft die Wende
nur einen Punkt liegt sie zurück
das kann noch klappen, mit viel Glück

Der Schiri stoppt das Spiel
»Das letzte Foul war nun zu viel«
zur Strafe gibt's der Würfe zwei
(das Match ist danach wohl vorbei)

Der Werfer hat die Wurfbahn frei,
den ersten wirft er links vorbei,
um diesen Fehlwurf aufzuheben,
wirft er bewusst nun rechts daneben,
Im Schnitt, sagt ihm sein schlichter Sinn,
war'n beide Würfe mittig drin.

Der Schiri wird den Spieler rügen,
»Du sollst nicht mit Statistik lügen«.

Diesem kleinen Gedicht von mir möchte ich natürlich direkt widersprechen. So einfach (und so einfältig) ist die statistische Methodik nicht.

Ich möchte lieber den Schriftsteller Herbert George Wells (1866-1946) zitieren, der sinngemäß gesagt haben soll: Neben Lesen und Schreiben erfordert eine technologisch geprägte Gesellschaft vor allem statistische Kenntnisse des Einzelnen. Vielleicht etwas übertrieben, aber man kann sicher sagen: Das Themengebiet der Statistik ist heute aus vielerlei Gründen und nicht zu Unrecht in allen Wissenschaftsdisziplinen fest verankert. Denn überall werden statistische Methoden eingesetzt: im Marketing (Stichworte: Markt- und Meinungsforschung, Konsumentenpanels), in der Volkswirtschaftlichen Gesamtrechnung (z. B. bei der Input/Output-Analyse), bei der Berechnung von Arbeitslosenzahlen (strukturell und regional differenziert), in der Sozialhilfe (Höhe der Hartz IV-Leistung), an der Börse (bei der Chartanalyse und bei Reaktionen auf neueste Konjunkturdaten), bei Banken (Einstufung der Kreditwürdigkeit), in der Qualitätskontrolle (Messverfahren in der Produktion), in der Wirtschaftspolitik (bei der Abschätzung von Konsequenzen steuerlicher Maßnahmen), im Gesundheitswesen (Kostendämpfung), bei der Europäischen Zentralbank (zur Kontrolle der Inflation und der Vorbereitung von Zinssatzänderungen), in der Bildungspolitik (PISA-Studie), bei der

Bekämpfung der Kriminalität (Herkunft und Häufigkeit), bei Mietspiegeln, bei den internationalen Berechnungen über Lebensstandard, -qualität und Armut, in der Medizin (Ursachen von Krebserkrankungen, Heilungschancen), in der Psychologie (Wirkung von Psychotherapien), in der Verhaltensforschung. Sogar in früher vernachlässigten Bereichen wie den Rechtswissenschaften (Rückfälligkeit von Tätern) gewinnt die Statistik immer mehr an Bedeutung.

Diese Liste der Einsatzgebiete ließe sich beliebig fortführen. Kein Soziologe, Pädagoge, Psychologe erhält seinen Abschluss ohne Nachweis zumindest grundlegender Statistikkenntnisse. Wüsste man ohne den Einsatz von Statistiken etwas über die Anzahl von über 21 verschiedenen bösartigen Tumorneubildungen beim Menschen, von der Lippe und der Mundhöhle über den Magen, die Leber, die Haut bis zur Harnblase, aufgeteilt in 10 Altersklassen?

Statistik ist überall. Z. B. auch in folgender Aussage: »Laut einer neuen Studie bekommen 1,2 Millionen Schüler Nachhilfe-Unterricht, also jeder siebte Schüler bis zum Alter von 16 Jahren. Dafür zahlen Eltern im Schnitt 87 Euro im Monat, insgesamt macht das 879 Millionen Euro in Deutschland pro Jahr. Und auch das ist interessant: Jeder dritte Schüler bekommt Nachhilfe, obwohl er beispielsweise in Mathe drei oder sogar besser steht.« Statistiken über Statistiken.

Schlagen Sie einfach eine beliebige Zeitung auf, ich wette, dass Ihnen dort auf Anhieb zahlreiche statistische Aussagen auffallen werden. Denn wir werden permanent mit Statistiken konfrontiert, alle Medien sind voll davon. 28 % der Führungskräfte gaben an, dass ihnen das Aussehen ihrer Angestellten wichtig ist. 1970 pilgerten 68 Menschen auf dem Jakobsweg, 2007 waren es schon 114026. Der Betrag, den ein Tierarzt für den Besuch bei einem Mastschwein erhält, beträgt 43 Euro, ein Hausarzt bekommt für den Besuch bei einem Patienten aber nur 14 Euro.

Und mehr als die Hälfte aller Nobelpreisträger in den Wirtschaftswissenschaften hat sich auch mit statistischen Methoden beschäftigt. Das sollte doch eigentlich Motivation genug sein.

Vielleicht hilft auch folgendes kleines Beispiel, Sie für Statistik zu interessieren. Angenommen, Ihr Chef bietet Ihnen (Variante A) 600 Euro mehr Gehalt zu jedem Jahresersten oder (Variante B) 200 Euro zu jedem Halbjahresbeginn; zusätzlich zu Ihrem bisherigen Grundgehalt von z. B. 20000 Euro im Jahr. Was wählen Sie? Natürlich die 200 Euro. Erstaunt?

Betrachten wir Ihr Gehalt in den folgenden Jahren, damit klar wird, warum das die richtige Entscheidung ist. Vergleichen wir dazu die Entwicklung in den einzelnen Halbjahren (Grundgehalt pro Halbjahr 20000 / 2 = 10000 Euro; 600 Euro mehr pro Jahr bedeutet 300 Euro mehr pro Halbjahr).

	Variante A: 600 Euro mehr pro Jahr	**Variante B: 200 Euro mehr je Halbjahr**
1. Halbjahr	10000	10000
2. Halbjahr	10000	10200

	Variante A: 600 Euro mehr pro Jahr	Variante B: 200 Euro mehr je Halbjahr
3. Halbjahr	10300	10400
4. Halbjahr	10300	10600
5. Halbjahr	10600	10800

Wer hätte das gedacht? In jedem Halbjahr verdienen Sie mit der Variante B mehr.

In dieser Einführung zur Deskriptiven Statistik lernen Sie einige Grundbegriffe der statistischen Methodenlehre kennen. Ausgehend vom Ziel der Statistik werden Begriffe wie Beobachtungswert, Untersuchungsmerkmal, Statistische Einheit und Statistische Masse abgegrenzt.

Ebenfalls unterschieden wird zwischen

- Querschnittsanalysen (Analysen zu einem bestimmten Zeitpunkt, z. B. wie viele Studierende im Sommersemester die Klausur in Statistik bestanden haben) und
- Längsschnittanalysen (Analysen, die regelmäßig und immer wieder durchgeführt werden und beispielsweise herausfinden wollen, ob sich die Ergebnisse durch den Einsatz von Tutoren verbessert haben).

Wenn also in einer Zeitung berichtet wird, dass immer mehr Eltern die Ferien ihrer Kinder eigenmächtig verlängern (um z. B. Reisekosten zu sparen) und dass deshalb immer mehr Bußgelder gegen sie verhängt werden (bis zu 90 Euro pro Kind und abwesendem Schultag), dann kann es sich sowohl um eine Querschnittsanalyse handeln (wenn zum Beginn der Sommerferien Zahlen aus verschiedenen Schulbezirken genannt werden) als auch um eine Längsschnittanalyse (wenn diese Zahlen mit denen aus den Vorjahren verglichen werden).

Übrigens werden es immer mehr: Eltern und Bußgelder. Wobei wir bei der zweiten Unterscheidung wären, der zwischen

- eindimensionalen Verteilungen; untersucht wird nur eine Größe, z. B. die Anzahl der verhängten Bußgeldbescheide in einem Schulbezirk, und
- mehrdimensionalen Verteilungen; untersucht wird, ob es einen Zusammenhang zwischen der Anzahl der Bußgeldbescheide und der Anzahl der Eltern gibt, die ihre und die Ferien ihrer Kinder verlängern (nicht jede unerlaubte Verlängerung muss ja zu einem Bußgeldbescheid führen, da diesbezüglich sicher Vermeidungsstrategien getestet werden).

Wir beginnen also mit Statistiken (oder sogenannten Verteilungen), bei denen wir nur eine Größe untersuchen, z. B. das Alter von Studienabbrechern, das Geschlecht von Teilnehmern auf der Kennenlernparty, die Anzahl von unehelichen Kindern oder die

Noten in der Statistikklausur (eine Statistik, die regelmäßig auch die Studierenden interessiert, die das Fach als eher bedrohlich einschätzen). In Deutschland gibt es 7 Mio. Analphabeten, fast 400.000 neue Abiturienten und über 27 Mio. Schweine, Häufigkeiten also, die wir da betrachten.

Wir beschäftigen uns dann damit, für solche Verteilungen Kennziffern zu berechnen. Statt der vielen einzelnen Daten wollen wir einen einzigen charakteristischen Wert angeben. Dieser Wert soll sozusagen ein Stellvertreter sein, typisch für alle Werte der Verteilung. Ein Wert, der uns etwas über die Verteilung sagt, obwohl wir viele der Ursprungswerte nicht mehr kennen.

So hat man beispielsweise ausgerechnet, dass der durchschnittliche Deutsche in seinem Leben 0,4 Häuser baut, 1,4 Kinder hat und 10,8 Autos kauft. Angegeben werden soll ein Wert, der eine gewisse Vorstellung von der Größenordnung vermittelt und unter Umständen Vergleiche ermöglicht. Wie sehen die entsprechenden Werte in Italien aus? Man kann mit solchen Parametern die Statistik natürlich ad absurdum führen, denn natürlich wird man niemanden finden, der tatsächlich genau 0,4 Häuser baut, 1,4 Kinder hat und 10,8 Autos in seinem Leben kauft. Geradezu unsinnig ist die Aussage, dass in Vatikanstadt 2 Päpste pro km^2 leben (dieser kleinste europäische Staat, er heißt tatsächlich korrekterweise Vatikanstadt, ist gerade einmal 0,44 km^2 groß, rein formal passt das also).

Zwischen einzelnen Variablen (Alter, Anzahl der Kinder) können Abhängigkeiten bestehen, z. B.: »Je höher das Lebensalter, desto größer die Anzahl der Kinder«. Oder: »Je weniger Haare, desto mehr Gehalt« (weil die Haare und die Zahl der Vorgesetzten im Laufe der Lebensjahre abnehmen). Dies sind dann zweidimensionale Verteilungen. Ebenso wie: »Hat das Alter einen Einfluss auf den Notendurchschnitt der Teilnehmer? Und wenn ja, welchen?«

In diesem Zusammenhang werden wir Größen im Zeitablauf betrachten (Längsschnittanalysen). Fragen können wir beispielsweise, ob die Kenntnisse im Fach Mathematik seit Beginn der gymnasialen Oberstufe G8 besser oder schlechter geworden sind. Genauso kann die Frage beantwortet werden, ob sich der Anteil der Teilnehmer im Masterstudiengang, die bereits eine kaufmännische Ausbildung abgeschlossen haben, im Zeitablauf verändert hat. Gibt es eine Funktion, die uns zuverlässig prognostiziert, wie der CO_2-Ausstoß in 30 Jahren sein wird? Lässt sich die globale Durchschnittstemperatur zuverlässig vorhersehen (wir erinnern uns dazu gerne an die Prognosen zum Waldsterben, die in den 80er Jahren des vorigen Jahrhunderts vorhersagten, dass heute kein Baum mehr in Deutschland stehen würde). Wann setzt die Wirkung eines Medikaments gegen Demenz ein? Die zweite, mögliche erklärende Größe ist manchmal einfach nur die Zeit.

Erinnern wir uns in diesem Zusammenhang an einen Slogan, mit dem die deutsche Milchwirtschaft geworben hatte: »Milch macht müde Männer munter«. Das sollten Studien belegen, die immer nach dem gleichen Schema ablaufen: Die Bewohner von Milchland trinken 10 % mehr Milch pro Kopf als die Bewohner von Bierland. Gleichzeitig stellte man fest, dass Krebs in Milchland seltener auftrat als in Bierland. Milch schützt also vor Krebs (noch schlimmer ist die Interpretation: Bier verursacht Krebs). Dass in den beiden Ländern vielleicht völlig unterschiedliche Lebensbedingungen vorherrschen, hat man außer Acht gelassen. Milchland liegt vielleicht in der zivilisierten

nördlichen Hemisphäre, Bierland vielleicht in Mittelafrika. Offensichtlich gibt es noch weitere, nicht berücksichtigte Faktoren.

Glaubt man übrigens einer anderen Studie, die von der »American Association for Cancer« veröffentlicht wurde, kann man nur hoffen, dass der zünftige Spruch möglichst unbeachtet bleibt. Denn in dieser Studie fand man heraus, dass Männer sich mit Milch ein erhöhtes Risiko antrinken, an Prostatakrebs zu erkranken. Deshalb nochmal: Es reicht in der Regel nicht aus, nur einen Grund zu suchen (Milch), der eine andere Größe beeinflusst (Krebs). Vielmehr ist es häufig ein Bündel von Ursachen, die eine Veränderung bewirken und die es herauszufinden gilt.

Nicht unerwähnt lassen wollen wir an dieser Stelle die Aussage, die oben so nebenbei getroffen wurde: »Bier verursacht Krebs«. Daraus, dass Milch im Gegensatz zu Bier keinen Krebs verursacht, kann eben nicht im Umkehrschluss abgeleitet werden, dass Bier Krebs verursacht. Oft verwechselt man Ursache und Wirkung. »Wer jünger ist, rennt schneller« bedeutet nicht: »Wer schneller rennt, ist jünger«. »Wer nicht raucht, lebt länger« bedeutet nicht: »Wer länger lebt, raucht nicht«. Oder um es noch deutlicher zu machen. »Wer häufig Sex hat, ist glücklicher« bedeutet nicht: »Wer glücklicher ist, hat häufiger Sex«. Oder auch: »Wer betrunken war, bekommt Kopfschmerzen« bedeutet nicht: »Wer Kopfschmerzen bekommt, war betrunken«.

Mit dem Verhältnis einzelner Größen zueinander beschäftigten wir uns im letzten Kapitel. Wenn in der Vatikanstadt die meisten Kriminalitätsdelikte pro Einwohner gemessen werden, dann liegt das weniger an den Einwohnern als an den vielen Touristen (bei ca. 1000 Einwohnern und ca. 20 Millionen Besuchern nachvollziehbar). Oder ein anderes Beispiel: Der Arbeitsplatz des US-Präsidenten ist der sicherste der Welt, in über 225 Jahren gab es nur 4 berufsbedingte Todesfälle im Amt (Lincoln, Garfield, McKinley, Kennedy) neben 4 weiteren, die zwar in ihrer Amtszeit starben, aber nicht durch Fremdeinwirkung (Harrison, Taylor, Harding, Roosevelt). Das ist nur ein toter Präsident alle 28 Jahre. Andererseits: Der Arbeitsplatz des US-Präsidenten ist auch gleichzeitig der gefährlichste: Immerhin wurden 4 von ihnen (fast 10 %) ermordet, mehr Tote als bei so gefährlichen Berufsgruppen wie Fischer oder Bergmann. Es kommt also darauf an, auf welche Zahl man die 4 ermordeten Präsidenten bezieht, um das Ergebnis gut oder schlecht aussehen zu lassen.

Sehr oft hört man deshalb, man solle bei statistischen Zusammenhängen eben nicht mit Prozentzahlen agieren. Teilweise richtig. Aber ist die Aussage, dass in China mehr Menschen Englisch sprechen als in England, dann noch eine Notiz wert? Wenn jeder 4., also ca. 25 % der Weltbevölkerung Chinesen sind, muss man schon länger suchen, um etwas zu finden, das sie nicht besser oder am häufigsten können.

Häufig werden wir sogar mit statistischen Aussagen konfrontiert, die widersprüchlich erscheinen: »Die Bahn ist das sicherste Verkehrsmittel«. »Das Flugzeug ist das sicherste Verkehrsmittel«. Beide Aussagen sind korrekt. Beiden Aussagen fehlt aber das Wesentliche: Wie sie zustande gekommen sind, was die Bezugsgröße ist.

Denn was ist ein Verkehrsmittel? Auto, Bus, Bahn, Flugzeug? Oder auch Fahrrad, Moped, Bobby Car? Welche Flugzeuge wurden in der Erhebung berücksichtigt? Jets und/oder Propellermaschinen? Wurden auch private und Sportflugzeuge betrachtet? Auch die Luftverkehrsgesellschaften, die auf der so genannten schwarzen Liste der EU stehen und in Europa keine Landeerlaubnis erhalten? Auf die man aber angewiesen ist, wenn

man im Kongo weiter fliegen möchte als die Lufthansa es tut? Und überhaupt: Was ist mit Sicherheit gemeint? Sicherheit, dass die Maschine nicht abstürzt? Oder zählt dazu auch die Sicherheit, sich nicht anzustecken? In Flugzeugen kommt man mit vielen, teils sogar lebensgefährlichen Bakterien in Kontakt, denn in Flugzeugen reisen sehr viele Menschen mit sehr unterschiedlichen Vorstellungen von Hygiene. Colibakterien tummeln sich im Waschraum, auf Klinken und auf den Klapptischchen, und das sind noch die harmlosesten Angreifer. Was nützt also die Aussage: »Das Flugzeug ist das sicherste Verkehrsmittel«, wenn ich nicht weiß, wie diese Aussage zustande gekommen ist?

Außerdem sollten wir immer versuchen, solche Aussagen in einem Gesamtzusammenhang zu sehen. Denn jährlich sterben mehr Menschen durch wildgewordene Esel als durch Flugzeugabstürze. Und jährlich sterben auf deutschen Straßen weit mehr als 3000 Menschen. Würden die alle an einem Tag sterben, gäbe es sofort eine Reisewarnung für Deutschland. Wie sicher ist jetzt die Bahn im Vergleich? Oder gar das Flugzeug? Egal, was Statistiker sagen: Wer eben Flugangst hat, vertraut eher auf die Sicherheit der Bahn oder seine eigenen Fahrkünste.

Solche schon auf den ersten Blick unplausiblen Aussagen wie »Schieße links vorbei und rechts vorbei und Du hast die Mitte getroffen«, die es in allerlei Versionen gibt, wollen wir aber erst gar nicht näher betrachten. Bei solchen Beispielen ist jedem Leser klar, dass die Urheber sich nicht wirklich mit den Methoden der Statistik beschäftigt haben.

Schließen wir unsere ersten Erläuterungen mit einem schönen Spruch:

»Geburtstage sind gut für Dich. Statistiken zeigen, dass die Menschen, welche die meisten haben, auch am längsten leben.« Damit kann man doch sofort etwas anfangen!

0.0 Prolog: Das Summenzeichen

Im Rahmen der Statistik, und hier insbesondere in dem der sogenannten Beschreibenden oder Deskriptiven Statistik, gibt es eigentlich nichts Mathematisches, vor dem

man Angst haben müsste. »Addition«, »Subtraktion«, »Multiplikation« und »Division« sind die einzigen arithmetischen Operationen, die man kennen muss; und das sollte nun einmal jeder. Das einzige Zeichen, das Sie beunruhigen könnte ist das hier:

$$\sum \text{ (sprich: Summe)}.$$

Es wird Ihnen aus Tabellenkalkulationsprogrammen bekannt vorkommen, wo es so häufig verwendet wird, dass es sogar im Standardmenü permanent zur Verfügung steht.

Ist eine Wassermelone 3 und eine andere 6 Kilogramm schwer, dann wiegen beide zusammen 9 Kilogramm. Wie immer liebt es der Mathematiker etwas formaler. Mit Recht, denn je klarer etwas formuliert ist, desto weniger gibt es unterschiedliche Meinungen über die Interpretation. Deshalb nennt er die beiden Werte nicht nur 3 und 6, sondern ordnet ihnen eine allgemeine Variable zu, z. B. x. Und da es mehrere x gibt, werden diese indiziert, z. B. mit i. In unserem Beispiel würde der Mathematiker also die beiden Werte $x_1 = 3$ und $x_2 = 6$ nennen; beim ersten Wert ist also $i = 1$, beim zweiten ist $i = 2$. Und nun bedeutet $\sum x_i$ nichts anderes, als diese beiden Werte zu addieren: $x_1 + x_2 = 3 + 6 = 9$. $\sum$ heißt also nichts anderes als »Summe« und bedeutet, dass man alles, was dahinter genannt wird, aufaddiert.

Man kann es auch allgemeiner formulieren: Alle Werte vom ersten bis zum letzten i werden addiert. Häufig gibt man deshalb nicht nur an, wie der Index lautet, sondern auch, von welchem ersten bis zu welchem letzten Wert er laufen kann. In unserem einfachen Beispiel nur von $i = 1$ bis $i = 2$:

$$\sum_{i=1}^{2} x_i.$$

Hätte man 5 Gewichtsangaben, würde das Summenzeichen so aussehen:

$$\sum_{i=1}^{5} x_i.$$

Wenn man vorher nicht genau weiß, wie viele Werte da eigentlich addiert werden sollen, gibt man als Obergrenze einen allgemeinen Wert vor, meist durch Verwendung der Variable n:

$$\sum_{i=1}^{n} x_i.$$

Man lässt also die Obergrenze offen und präzisiert sie erst im Rahmen weiterer Überlegungen. Damit hat man das Summenzeichen so allgemein formuliert, dass man es auf verschiedene Fälle anwenden kann, ohne am grundsätzlichen Konstrukt etwas ändern zu müssen.

Untersucht man also 2 Wassermelonen, lautet die Formel allgemein $\sum_{i=1}^{n} x_i$, im speziellen Fall $\sum_{i=1}^{2} x_i$; untersucht man 5 Wassermelonen, lautet die allgemeine Form natürlich immer noch $\sum_{i=1}^{n} x_i$, aber nun ist die Obergrenze 5, also präziser $\sum_{i=1}^{5} x_i$.

Da Unter- und Obergrenze sich in unseren Beispielen, wie allgemein in der Statistik, aus dem Sachverhalt oder Zusammenhang immer direkt ergeben (sollten), verzichtet man oft ganz auf die Angabe von Unter- und Obergrenze (man vergisst natürlich keine Werte, die man einmal erhoben hat, da man, vereinfacht gesagt, »einfach alle« in die Addition übernimmt). Wir schreiben daher nur $\sum x_i$ und wissen, dass wir die Werte mit

dem Index i aufsummieren sollen, und zwar vom ersten bis zum letzten bekannten Wert. Bei 2 Wassermelonen wissen wir (bezogen auf den hier vorliegenden Zusammenhang), dass wir beide Gewichtsangaben aufaddieren sollen, bei 5 Wassermelonen eben alle 5. Und wären es 10 Gewichtsangaben, dann bedeutet $\sum x_i$ eben, addiere alle 10.

Statt x mit dem Index i kann man natürlich auch eine Variable y mit einem Index j, also $\sum_{j=1}^{m} y_j$ oder jede anders bezeichnete Variable mit jedem anderen Index betrachten.

Soweit ganz einfach. Ein nachvollziehbares Problem haben Einsteiger aber immer dann, wenn hinter dem Summenzeichen nicht nur eine Größe steht, die man aufaddieren soll, sondern mehrere »verbundene«. Eine solche Formel könnte also so aussehen:

$$\sum x_i y_j \text{ oder } \sum x_i \sum y_j.$$

Das Ergebnis ist sehr unterschiedlich. Merken wir uns dazu: Immer was direkt hinter dem Summenzeichen steht, wird ausgeführt.

$\sum x_i \sum y_j$ bedeutet also, wir addieren erst alle x_i, dann alle y_j und multiplizieren dann diese beiden Summen (wie in der Mathematik üblich, lässt man das Multiplikationszeichen gerne mal weg).

Ein Beispiel: Wir haben 2 Werte für x und 5 Werte für y in tabellarischer Form also:

x_1	x_2	y_1	y_2	y_3	y_4	y_5
3	6	3	4	5	4	4

Dann ist die Summe aller x (wegen $3 + 6$, genauer $x_1 = 3$ und $x_2 = 6$)

$$\sum x_i = 9,$$

die Summe aller y (wegen $3 + 4 + 5 + 4 + 4$, genauer $y_1 = 3$, $y_2 = 4$, $y_3 = 5$, $y_4 = 4$ und $y_5 = 4$)

$$\sum y_j = 20.$$

Und folglich ist wegen $9 \cdot 20$

$$\sum x_i \sum y_j = 180.$$

Auch der Fall $\sum x_i + \sum y_j = 29$ wäre denkbar, hat aber in der Statistik keine weitere Bedeutung.

Da wir anhand der Datenlage wissen, dass die x aus 2 und die y aus 5 Werten bestehen, hätte man neben der Angabe der Unter- und Obergrenze auch noch die Angabe des Indizes weglassen können, ohne Verwirrung zu stiften:

$$\sum x = 9,$$

$$\sum y = 20,$$

$$\sum x \sum y = 180.$$

Folgerichtig ist, dass wir $\sum x_i y_j$ nicht berechnen können; die Anzahl i (= 2) ist nämlich nicht gleich der von j (= 5). Um den Unterschied zwischen $\sum x_i y_j$ bzw. $\sum xy$ zu erläutern, bemühen wir deshalb ein Beispiel, in dem die Anzahl von x und y gleich groß ist – dazu die folgende Wertetabelle:

	x	y	$x \cdot y$
1	3	3	9
2	6	4	24
3	3	5	15
4	6	4	24
5	3	4	12
$\sum$	**21**	**20**	**84**

Berechnen wir erst $x_1y_1 = 9$, dann $x_2y_2 = 24$, $x_3y_3 = 15$, $x_4y_4 = 24$, $x_5y_5 = 12$, erhalten wir:

$$\sum x_i y_i = 9 + 24 + 15 + 24 + 15 = 84.$$

Oder einfacher dargestellt:

$$\sum xy = 84.$$

Zum Vergleich:

$$\sum x = 21 \,(\text{wegen } 3 + 6 + 3 + 6 + 3),$$

$$\sum y = 20 \,(\text{wegen } 3 + 4 + 5 + 4 + 4).$$

Im Gegensatz zu $\sum xy = 84$ ist $\sum x \sum y = 21 \cdot 20 = 420$.

Gleich noch ein weiteres einfaches Beispiel:

x	y	$x \cdot y$	$y \cdot x$
1	4	4	4
2	5	10	10
3	6	18	18
$\sum x = 6$	$\sum y = 15$	$\sum xy = 32$	$\sum yx = 32$

Hier erhalten wir:

$$\sum xy = 32,$$

$$\sum x \sum y = 6 \cdot 15 = 90.$$

Natürlich ist $\sum xy = \sum yx$ und $\sum x \sum y = \sum y \sum x$.

Wie gesagt, wäre auch die Berechnung $\sum x + \sum y$ möglich (6 + 15), aber wenig sinnvoll.

Betrachten wir nun noch ein Beispiel, das den höchsten Ansprüchen genügt, die wir im Laufe dieses Buches erfüllen müssen:

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}.$$

Wie so oft hilft es, auf den Taschenrechner zu verzichten. Der ist zwar schlau, aber nicht jeder Nutzer (auch ich nicht) beherrscht die hohe Kunst der Klammerbildung, so wie sie der Rechner verlangt. Denn wir wissen alle, dass Punktrechnung vor Strichrechnung kommt, wir also in der Formel nicht einfach durch $n\sum x^2$ teilen können, weil wir von diesem $n\sum x^2$ erst etwas abziehen müssen, nämlich $(\sum x)^2$. Und das geht nicht, ohne vorher eine Klammer zu öffnen.

Am einfachsten kommen wir mit einer Hilfstabelle voran. Diese hilft dabei, unsere Fehler auf ein Minimum zu reduzieren; alles lässt sich durch Kopfrechnen erledigen:

	x	y	$x \cdot y$	$y \cdot x$	x^2	y^2
1	3	3	9	9	9	9
2	6	4	24	24	36	16
3	3	5	15	15	9	25
4	6	4	24	24	36	16
5	3	4	12	12	9	16
$\sum$	$\mathbf{x = 21}$	$\mathbf{y = 20}$	$\mathbf{xy = 84}$	$\mathbf{yx = 84}$	$\mathbf{x^2 = 99}$	$\mathbf{y^2 = 82}$

Bedeutend ist der Unterschied zwischen $(\sum x)^2$ und $\sum x^2$:

$$\left(\sum x\right)^2 = \sum x \cdot \sum x = 21^2 = 441$$

ergibt etwas völlig anderes als

$$\sum x^2 = 99.$$

Der Vollständigkeit halber wollen wir unsere komplizierte Formel nun einmal durchrechnen:

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2} = \frac{99 \cdot 20 - 21 \cdot 84}{5 \cdot 99 - 441} = \frac{1980 - 1764}{495 - 441} = \frac{216}{54} = 4.$$

Übrigens: Alle höherwertigen Rechner enthalten vorinstallierte Funktionen, mit denen sich solche bekannte Formeln durch einfache Eingabe der x- und y-Werte leicht lösen lassen. Ein Blick in die Bedienungsanleitung lohnt!

Schließen wir diese Ausführungen zum Summenzeichen mit einem bekannten Beispiel, das dem Mathematiker Carl Friedrich Gauß (1777-1855) zugesprochen wird. Sein Lehrer wollte ein wenig Ruhe und gab der Klasse die Aufgabe, alle Zahlen von 1 bis 100 zu addieren. Gauß kam spontan auf die Idee, dass er mit den Paaren 1 und 100, 2 und 99,

3 und 98 usw. schneller ans Ziel gelangt. Sein Ergebnis also: 50mal kommt dieses Wertepaar vor, also $50 \cdot 101 = 5050$. Wir würden schreiben:

$$\sum_{i=1}^{100} i = 5050.$$

Der Lehrer hat ihm das angeblich nicht übel genommen, ganz im Gegenteil hat er den kleinen Gauß gefördert. Er fand es offensichtlich gut, dass es Schüler gibt, die nachdenken, statt schematisch die vermeintlich naheliegende Lösung zu wählen.

0.1 So werden die Daten übersichtlicher: das Ziel der Statistik

Das ist Simon:

Über ihn weiß man einiges: Er ist 5 Jahre alt und lebt mit seinen Eltern und seinem großen Bruder in einer kleinen Gemeinde am Rande des Ruhrgebiets. Er geht gerne in den Kindergarten, gerade beginnt er dort sein drittes Jahr. Er ist gesund und gegen die üblichen Kinderkrankheiten Masern, Keuchhusten, Kinderlähmung, Diphterie und Wundstarrkrampf geimpft. Außerdem gehört er im Kindergarten zur sogenannten »Igelgruppe« und er wird über Mittag betreut, ist also auch nachmittags im Kindergarten. Offensichtlich mag er auch Luftballons.

Das alles hat mit Statistik erst einmal wenig zu tun. Alle Daten, zumindest die, die man wissen möchte, sind bekannt. Statistik beginnt erst dann, wenn man nicht nur Simon betrachtet, sondern z. B. alle Kinder in seinem Kindergarten und auch nur das betrachtet, was für eine Untersuchung wesentlich erscheint. Vermutlich wird niemand auf die Idee kommen, die getätigten Impfungen zu zählen, denn das sollte im Ermessen der Familien liegen.

Wenn man die Daten in eine Tabelle untereinander schreibt, wird das Ganze sehr schnell unübersichtlich. Die Frage, wer und vor allem wie viele Kinder 5 Jahre alt sind und nächstes Jahr in die Schule kommen, lässt sich nicht auf einen Blick beantworten:

Name	Alter	Geschlecht	Gruppe	Elternbeitrag	Mittagsbetreuung
Simon	5	männlich	Igel	168	nein
Lena	3	weiblich	Igel	168	ja
Max	4	männlich	Schildkröte	168	nein
Lara	5	weiblich	Igel	0	nein
Yusuf	5	männlich	Schildkröte	42	nein
Mehmed	5	männlich	Schildkröte	0	nein
Clara	3	weiblich	Igel	84	nein
Leoni	4	weiblich	Igel	127	ja
Marian	4	männlich	Schildkröte	42	nein
Marvin	3	männlich	Igel	42	nein
Luise	3	weiblich	Schildkröte	0	nein
Luisa	5	weiblich	Schildkröte	0	nein
Kevin	4	männlich	Igel	127	nein
Laura	5	weiblich	Schildkröte	168	ja
Anne	3	weiblich	Schildkröte	84	nein
Christian	5	männlich	Igel	84	nein
Christine	5	weiblich	Schildkröte	42	nein
Martin	4	männlich	Schildkröte	127	nein
Johanna	5	weiblich	Igel	84	nein
Trude	4	weiblich	Schildkröte	84	nein
Claude	3	männlich	Schildkröte	127	nein
Salome	4	weiblich	Igel	127	nein
Stephanie	4	weiblich	Igel	0	ja
Janina	3	weiblich	Schildkröte	42	nein
Murat	5	männlich	Schildkröte	42	ja
Kosima	5	weiblich	Igel	42	nein
Ernst	5	männlich	Igel	168	nein
Manuel	4	männlich	Igel	168	nein
Ralf	5	männlich	Schildkröte	127	nein
Kassandra	4	weiblich	Igel	84	ja