0 Statistik ist überall – eine Einführung

Nach 12-mal Rot muss es gelingen:
»ich setz auf Schwarz und werd‹ gewinnen,
Roulette ist doch ein Kinderspiel,
man muss nur wissen, wie die Kugel fiel«,
denkt Heinz und setzt nun Haus und Hof.

Tatsächlich ist der Heinz recht doof.

Die Kugel hat nicht mitgezählt,
weshalb sie keine Farbe wählt,
sie fällt auf Schwarz und Rot ganz nach Belieben,
entscheidet neu, wo will sie liegen,
sie weiß nicht, was sie vorher tat,
deshalb an Heinz der gute Rat:

Mit einer Chance eins zu eins,
verlierst erneut nur Du, Du Heinz.

Dieses kleine Gedicht von mir wird Sie vermutlich enttäuschen. Leider ist Statistik so und nicht anders. Sie kann uns nicht helfen, beim Roulette zu gewinnen. Es gibt keine Strategie, und das muss ich in dieser Deutlichkeit vorausschicken, mit der man Glücksspiele überlisten kann. Säße ich dann hier am Schreibtisch und würde über Statistik schreiben?

Bevor wir diesbezüglich in die Tiefe gehen: Auch eine wiederholte Verdoppelung des Einsatzes scheitert früher oder später am Limit, das Sie setzen dürfen. Ganz abgesehen davon verlangt die Bank Bares, haben Sie so viel davon? Rechnen wir nach:

- 1. Fall: 10 Euro gesetzt, verloren, Einsatz verdoppelt, also 20 Euro gesetzt, dann gewonnen. Sie sind wieder da, wo Sie angefangen haben. Und mehr erreichen Sie mit dieser Strategie nicht.
- 2. Fall: 10 Euro gesetzt, gewonnen, nochmal 10 Euro gesetzt, gewonnen. Mit dieser Strategie gewinnen Sie nach 10 Runden 100 Euro. Nicht vergessen: Wenn Sie sehr, sehr viel Glück hatten.
- 3. Fall: 10 Euro gesetzt, verloren, Einsatz verdoppelt, also 20 Euro gesetzt, wieder verloren, Einsatz erneut verdoppelt, also 40 Euro gesetzt, verloren, verdoppelt, 80 Euro …Schon nach 10 Runden verlieren Sie 5120 Euro. Zugegeben, es passiert selten, dass dieselbe Farbe 10mal kommt (so selten wie Fall 2). Aber es passiert und Sie sind alle

Einsätze los (also 5120 Euro). Und es kommt ein weiteres Problem hinzu: Sie erreichen vielleicht schon vorher das Tischlimit, dürfen also nicht mehr verdoppeln.
- Natürlich gibt es noch weitere denkbare Gewinn- und Verlustfolgen, aber glauben Sie mir, wirklich besser wird es nicht.

Fazit: Entweder Sie fallen auf Ihren Einsatz zurück oder Sie gewinnen ein bisschen oder Sie scheitern am Limit. Sie gewinnen also im ersten Fall nichts, im zweiten ein wenig, aber im dritten verlieren Sie richtig viel.

Wir können schon jetzt festhalten, dass das, was in der Beschreibenden Statistik die relativen Häufigkeiten waren (wie oft fällt die Kugel auf Rot), in der Schließenden Statistik zu Wahrscheinlichkeiten wird (die Kugel fällt mit einer Wahrscheinlichkeit von 0,5 auf Rot). Das Ergebnis einer Untersuchung lautet also nicht, dass 4 von 52 Karten Asse sind (7,7 %), sondern wir können mit einer Wahrscheinlichkeit von 0,077 erwarten, dass die nächste Karte ein Ass ist.

Und genauso wie relative Häufigkeiten erscheinen auch statistische Wahrscheinlichkeiten oft widersprüchlich: »Die Wahrscheinlichkeit, mit dem Auto zu verunfallen, ist wahrscheinlicher als die, dass es mit dem Flugzeug oder der Bahn passiert«. Die Aussage ist korrekt. Aber ihr fehlt wieder das Wesentliche: Wie ist diese Wahrscheinlichkeit zustande gekommen, was ist die Bezugsgröße? Denn was ist mit Wahrscheinlichkeit gemeint? Die Wahrscheinlichkeit, dass wir bei dem Autounfall sterben werden? Oder zählt dazu auch die Wahrscheinlichkeit, im Rollstuhl zu landen? In Deutschland ereignen sich ca. 2,5 Millionen Unfälle im Straßenverkehr; bei über 43 Millionen zugelassenen Autos liegt die Wahrscheinlichkeit, einen Unfall zu haben, also etwas unter 6 %. Ca. 4000 Getötete (mehr als 10 pro Tag!) ergeben einen Anteil von 0,01 %, 70000 Schwerverletzte 0,16 %. Zum Vergleich: Die Wahrscheinlichkeit bei einem Flugzeugabsturz ums Leben zu kommen, beträgt je nach Quelle etwa 0,00003 % (und das sogar weltweit).

Das, was für die Beschreibende Statistik gilt, gilt also auch für die Schließende, sie ist allgegenwärtig. Schlagen wir dazu einfach eine beliebige medizinische Zeitung auf, Ihnen werden auf Anhieb viele statistische Aussagen auffallen. »Im Schnitt beträgt die Wahrscheinlichkeit sich nach einem Zeckenstich mit Borreliose zu infizieren bei 1,5 bis 6 Prozent. Je länger die Zecke saugt, desto höher ist die Wahrscheinlichkeit einer Ansteckung.« »Studien mit Zwillingen weisen darauf hin, dass es genetische Faktoren gibt, die eine Depression wahrscheinlicher machen«. »Man geht davon aus, dass Dengue-Fieber bald auch bei uns in Deutschland, vor allem in wärmeren Gebieten wie zum Beispiel am Oberrhein, übertragen werden kann.« »Fast alle Fälle von Chorea Huntington entstehen durch Vererbung: Eltern mit entsprechendem Erbmerkmal geben dieses an ihre Nachkommen weiter; dabei reicht es aus, dass nur ein Elternteil die Genveränderung in sich trägt. Chorea Huntington wird über einen sogenannten autosomal-dominanten Erbgang an die Kinder übertragen: die Kinder erben das veränderte Gen somit mit einer Wahrscheinlichkeit von 50 Prozent.« Oder eine naheliegende Frage, die viele junge (und vielleicht auch ältere wohlhabende Herren mit einer jüngeren Partnerin) interessiert: Wie groß ist die Wahrscheinlichkeit, schwanger zu werden? Dazu gibt es massenweise Studien, aus denen ich nur zwei

Ergebnisse zitieren will: »Frauen zwischen 35 bis 39 Jahren haben eine 50 % geringere Chance schwanger zu werden.« »Je Zyklus liegt die Wahrscheinlichkeit bei ca. 25 %, wer genau die Tage zählt und die Körpertemperatur misst, kann diese auf fast 40 % steigern.« Auch die in diesem Zusammenhang eher traurige Frage nach der Wahrscheinlichkeit einer Fehlgeburt wurde berechnet und interessiert sicher vor allem Betroffene: »15 % aller Schwangerschaften enden mit einer Fehlgeburt, das Risiko einer zweiten sinkt auf 5 % und 1 % aller Paare erleben sogar 3 davon.« Oder z. B. in folgender Aussage: »Übergewicht erhöht nicht nur das Diabetesrisiko, sondern auch die Möglichkeit, sich einer Knieersatztherapie unterziehen zu müssen: Bei schwer übergewichtigen Frauen ist das Risiko – in beiden Fällen – zwölf Mal höher. Die Wahrscheinlichkeit für hohen Blutdruck ist fünf Mal so groß wie bei normalgewichtigen Frauen. Männer in einer hohen Gewichtskategorie haben acht Mal so oft Diabetes und sechs Mal so oft Knie-Operationen und hohen Blutdruck.« Wahrscheinlichkeiten über Wahrscheinlichkeiten.

Wir können deshalb festhalten: Auch das Themengebiet der Schließenden Statistik ist heute aus vielerlei Gründen und nicht zu Unrecht in allen Wissenschaftsdisziplinen fest verankert. Denn überall werden schließende statistische Methoden eingesetzt: in der Medizin (wenn es darum geht, die Wahrscheinlichkeit einer zweiten Fehlgeburt zu berechnen), in der Psychologie (wenn man die Heilungschancen beim Einsatz verschiedener Psychotherapien vergleicht), in der Wahlnacht (wenn man die Prognosen zu ersten Hochrechnungen verdichtet), bei Banken (Stresstests zur Vorhersage von Reaktionen auf Marktturbulenzen), in der Qualitätskontrolle (Stichprobenverfahren in der Produktion), in der Wirtschaftspolitik (Wahrscheinlichkeit von Steuervermeidungsstrategien), im Gesundheitswesen (bei der Abschätzung von Risiken und Nebenwirkungen, Stichwort seltene und sehr seltene Nebenwirkungen), in den Rechtswissenschaften (Wahrscheinlichkeit eines Rückfalls bei Straftätern).

Grundsätzlich gilt: Wäre alles auf der Welt vorhersehbar, also determiniert, bräuchte man die Schließende Statistik nicht mehr. Ob dieser Zustand durch fortschreitenden Erkenntnisstand jemals erreicht wird, darf wohl bezweifelt werden. Denn letztendlich lassen sich alle Ereignisse unseres Lebens und unserer Umwelt vereinfacht in zwei Kategorien unterteilen, über die sich schon zahlreiche Philosophen den Kopf zerbrochen haben: determinierte und stochastische Ereignisse. Die Entfernung zwischen Erde und Mond ist ein streng determinierter Prozess, man weiß immer sehr genau, wann welcher Abstand gemessen werden kann. Das Gleiche gilt für viele andere physikalische Gesetzmäßigkeiten: den Zeitpunkt des Sonnenaufgangs, die nächsten Gezeiten, die Beschleunigung beim freien Fall, der Verbrauch eines Motors bei konstanter Last. Aber Wahrscheinlichkeiten sind nicht wie Gleichungen, die Zahl ist nicht exakt. Sicheres Wissen gibt es in der Wirtschaft und Soziologie nicht (dort nennt man so etwas auch ceteris paribus). Wir müssen mit der Unsicherheit umgehen und diese richtig kalkulieren. Wenn eine amerikanische Radiostation den Wetterbericht verliest, gibt sie Prozentzahlen für bestimmte Arten von Niederschlägen an (z. B. »twenty percent chance of rain«). Dass es im Winter kälter wird, ist vorhersehbar (manchmal scheint selbst das nicht mehr »wahrscheinlich«); wie kalt es tatsächlich wird, entzieht sich aber einer sicheren Vorhersage. Zugvögel fliegen im Herbst in den Süden, ein genaues Datum ist

aber auch hier nicht vorhersehbar. Gleichwohl ist es »besser« zu wissen, dass es dann so ungefähr geschieht.

Schließen wir die einführenden Bemerkungen mit einem netten Spruch, der uns Mut machen soll: »Da lernt man also Mittelwerte und berechnet Wahrscheinlichkeiten und dann steht man trotzdem grübelnd vor dem Backofen und fragt sich, welche der vier Schienen mit der größten Wahrscheinlichkeit die mittlere ist.«

0.0 Prolog: Binomialkoeffizient und Co.

Im Rahmen der Schließenden Statistik gibt es wie in der Beschreibenden Statistik nichts Mathematisches, vor dem man wirklich Angst haben müsste. Auch hier sind »Addition«, »Subtraktion«, »Multiplikation« und »Division« die einzigen arithmetischen Operationen, die man kennen muss. Nur vier Dinge könnten einen Leser dieses Statistikbuches beunruhigen:

$$N^n,\ N!\ \text{bzw.}\ n!\ \text{und}\ \binom{N}{n}.$$

N soll die Anzahl der Elemente in einer Grundgesamtheit und n die Menge in einer Stichprobe sein (wir werden den Unterschied zwischen Grundgesamtheit und Stichprobe später genauer auflösen). Nehmen wir nun an, dass wir 4 verschiedene Cocktailzutaten im Kühlschrank haben (N) und daraus einfach 2 beliebige auswählen möchten. N^n ist damit $4^2 = 16$; es gibt also 16 Möglichkeiten, 2 von 4 Cocktailzutaten aus dem Kühlschrank zu holen. Aus den 4 Zutaten ABCD also AA, AB, AC, AD, BA, BB, BC, BD, CA, CB, CC, CD, DA, DB, DC, DD.

N! (sprich N Fakultät) bzw. n! bedeutet, dass man das Produkt aller Zahlen von N bzw. n bis 1 bildet. Bei $N = 4$ ist das also:

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24,$$

bei $n = 2$ ist das:

$$2! = 2 \cdot 1 = 2.$$

Der sogenannte Binomialkoeffizient (man spricht »N über n«)

$$\binom{N}{n}$$

ist eine verkürzte Form von

$$\frac{N!}{n! \cdot (N-n)!}$$

bzw.

$$\frac{N \cdot (N-1) \cdot (N-2) \cdot (N-3) \cdot \ldots \cdot (N-n+1)}{n!}.$$

Die verschiedenen Möglichkeiten, 2 verschiedene (!) Zutaten zufällig aus dem Kühlschrank zu holen, sind:

$$\binom{N}{n} = \binom{4}{2} = \frac{N!}{n! \cdot (N-n)!} = \frac{4!}{2! \cdot (4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1}$$

bzw.

$$= \frac{4 \cdot 3}{2 \cdot 1} = 6.$$

Zur Verdeutlichung: Es sind die Zutaten AB, AC, AD, BC, BD, CD, da im Gegensatz zu oben nicht zweimal die gleiche Zutat gewählt werden darf.

Auch die Wahrscheinlichkeit, im Lotto 6 Richtige zu erzielen, lässt sich über den Binomialkoeffizienten leicht berechnen; exakt ist sie:

$$\binom{N}{n} = \binom{49}{6}$$

$$= \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot \ldots \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot \ldots \cdot 3 \cdot 2 \cdot 1}$$

bzw.

$$= \frac{49 \cdot (49-1) \cdot (49-2) \cdot (49-3) \cdot \ldots \cdot (49-6+1)}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

$$= \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13.983.816.$$

Verlieren wir noch ein Wort zu einer der unbeliebtesten Rechenmethoden, die Schüler jemals gequält hat, dem Integral. Vereinfacht gesagt: Was in der Beschreibenden Statistik das Summenzeichen ist, ist in der Schließenden Statistik (manchmal) das Integral. Manchmal deshalb, weil es nur dann verwendet werden muss, wenn wir stetige Variablen betrachten (wir erinnern uns: stetige Variable können einen Tatbestand beliebig genau messen, während diskrete Merkmale nur bestimmte Werte annehmen können).

Summiert man die Werte $x_1 = 110$, $x_2 = 111$ und $x_3 = 112$ auf, so erhält man im diskreten Fall $\sum x_i = 333$; beim Integral wäre das Ergebnis kein anderes. Addiert man die Körpergrößen von 3 Jugendfußballern, die 110, 111 und 112 cm groß sind, so erhält man also:

$$\sum_{110}^{112} x_i = 333.$$

Allerdings ist dieser Wert auf ganze cm gerundet; in Wirklichkeit ist das erste Kind genau 110,25 oder noch genauer 110,2538 oder noch genauer …cm groß. Es gibt also keinen exakten Wert, sondern immer nur einen, den man noch genauer fassen könnte. Deshalb arbeitet man mit Wertebereichen, z. B. den Körpergrößen zwischen 110,00000 und 111,00000…, also Bereichen, die sehr viele einzelne Werte enthalten könnten. Und diese Bereiche schließt ein Integral ein. Wenn wir dann wissen wollen, wie groß die Wahrscheinlichkeit dafür ist, dass unter den Kindern nur diejenigen berücksichtigt wurden, die zwischen 110 und 112 groß sind, dann berechnen wir dazu ein Integral, in das alle diese Kinder fallen, egal, wie exakt gemessen wird. Wir berechnen

$$W(x) = \int_{110}^{112} f(x_i)dx.$$

Folgerichtig ist ein Kind, das in einem Jahr (z. B. zwischen seinem 5. und 6. Lebensjahr) von 101,0000…. auf 113,0000…. Zentimeter gewachsen ist, 12,0000…. Zentimeter größer geworden. Deshalb schreibt man dann

$$\int_{101}^{113} f(x_i)dx.$$

Ein Fußballspiel dauert in der Regel 90 Minuten, der Zeitraum zwischen der 50. und 60. Minute lässt sich auch über das entsprechende Integral berechnen; ein Tor, das in diesem Zeitraum fällt, fällt genaugenommen in der 56,0523…-ten Minute.

Genauer brauchen wir uns hier über Integrale nicht den Kopf zu zerbrechen, denn zum Glück können wir immer dann, wenn integriert werden muss, auf Tabellen zugreifen, aus denen wir den Wert nur ablesen müssen. Perfekt.

0.1 Warum man nicht einfach nur zählen kann: die Abgrenzung der Schließenden von der Beschreibenden Statistik

Statt von der Abgrenzung der Deskriptiven von der Induktiven, Analytischen Statistik zu sprechen, wollen wir von der Abgrenzung der Beschreibenden von der Schließenden Statistik reden, was mir persönlich besser gefällt, weil es den Unterschied deutlich macht, ohne dass man sich in einer doch nicht so ganz geläufigen (früheren) Fremdsprache zurechtfinden muss.

In der Beschreibenden Statistik untersuchen wir eine statistische Masse vollständig so wie sie ist. Wir berechnen, wie viele Arbeitslose es zu einem bestimmten Stichtag in Deutschland gibt (beispielsweise am 31.3. des Jahres); von der örtlichen Arbeitsagentur bis hin zur Bundesagentur für Arbeit werden dazu an verschiedenen Erhebungsstellen die Zahlen im Einzelnen erhoben und zusammengezählt. Man spricht dann von einer Vollerhebung, weil alle, in diesem Fall alle Arbeitslosen, voll erhoben wurden. Auch die Durchführung einer Bundestagswahl erfordert Methoden der Beschreibenden Statistik. Die Stimmen, die einzelnen Parteien gegeben wurden (Häufigkeiten), werden von den Wahlbüros an die städtischen Wahlämter und von dort an die Landeswahlleiter und letztendlich an den Bundeswahlleiter weitergegeben. Auch das ist eine Vollerhebung, denn wirklich alle Stimmen aller Wahlberechtigten (ob sie nun wählen oder nicht, also auch die der Nichtwähler) werden erfasst und finden zumindest in der Wahlbeteiligung Berücksichtigung.

Gleichzeitig kennen wir im Zusammenhang mit Wahlen aber auch eine andere Methode, eine aus der Schließenden Statistik. Vor der Wahl nämlich und auch gleich nach Schließung der Wahllokale werden Prognosen und Hochrechnungen veröffentlicht. Prognosen darüber, wie viele Stimmen denn welche Partei vermutlich insgesamt erzielen wird. Oft sind diese Zahlen recht nah am Endergebnis. Sie sind nicht auf den Punkt genau, aber doch so gut, dass sie für einen ersten Jubel oder erste Enttäuschungen sorgen. Aber natürlich sind sie auf einem anderen Weg zustande gekommen als die Zahlen einer Vollerhebung. Solche Teilerhebungen sind Stichproben, die hochgerechnet werden. Wenn also in Deutschland 60 Millionen Bürger wahlberechtigt sind und man etwa 20.000 (typische) Bürger sozusagen als Stellvertreter befragt, dann rechnet man das

Ergebnis entsprechend hoch. 20.000 von 60.000.000 sind 0,03 %. Wählen von diesen 20.000 befragten 2.500 die Partei A, dann sind das hochgerechnet 250/0,0003 = 8.333.333 Wähler insgesamt (= 13,89 % der Stimmen).

Halten wir also fest: In der Schließenden Statistik (manche sagen auch Analytische Statistik, weil man die Daten halt nicht nur beschreibt, sondern auch analysiert) geht es immer um Stichproben aus einer sogenannten Grundgesamtheit. Dabei verwechseln wir bitte nicht die Begriffe Grundgesamtheit und statistische Masse. Wenn man in der Beschreibenden Statistik alle statistischen Einheiten, also alle Schüler einer Schule oder eines Bundeslandes untersucht, dann ist das die statistische Masse. Wenn ich nur einige Klassen auswähle und der Meinung bin, das Ergebnis dieser ausgewählten Klassen stehe stellvertretend für alle Klassen (also die statistische Masse), dann handelt es sich um eine Stichprobe. Dennoch kann die statistische Masse auch Stichprobe sein, wenn man beispielsweise alle statistischen Einheiten (alle Schüler des Bundeslandes) untersucht hat (was in dieser Untersuchung die Zielgröße, die statistische Masse, war) und dieses Ergebnis als Ausgangspunkt einer Hochrechnung auf alle Schüler in Deutschland nutzt. Dann ist die vorherige statistische Masse nun Stichprobe für eine noch größere Grundgesamtheit.

Unterzieht man alle deutschen Banken einem Stresstest, handelt es sich um eine Vollerhebung im Rahmen der Beschreibenden Statistik, die einzelnen Banken sind die statistischen Einheiten, alle zusammen die statistische Masse. Untersucht man nur die Banken und nicht die Sparkassen, erhält man eine (sicher nicht repräsentative) Stichprobe, von der aus man auf die Grundgesamtheit aller Banken und Sparkassen (fehlerhaft) schließen könnte.

Untersucht das Organisationsteam im beschaulichen Südtiroler Ort Sexten die Brenndauer von Wunderkerzen (die man an Sylvester an Luftballone bindet und zu heimeliger Musik zwecks stillem Feuerwerk in die Luft steigen lässt) hat man sicherlich nur einige wenige untersucht (um herauszufinden, ob sie wirklich lange genug für den gewünschten Effekt brennen). Man hat also eine Stichprobe gezogen, eine Frage der Schließenden Statistik. Und man hat eben nicht wie im Rahmen der Beschreibenden Statistik die statistische Masse insgesamt untersucht (alle Wunderkerzen). Hört sich komplizierter an als es ist.

> Letztendlich bestimmt die Fragestellung, was statistische Masse und was Grundgesamtheit ist. Ordnen wir die Begriffe einfach den verschiedenen statistischen Methoden zu. Bewegen wir uns im Rahmen der Beschreibenden Statistik, reden wir von statistischen Einheiten und der statistischen Masse, bewegen wir uns auf dem Gebiet der Schließenden Statistik, untersuchen wir Stichproben und Grundgesamtheiten. Eine Stichprobe kann auch die statistische Masse sein, über die man im Rahmen der Beschreibenden Statistik abschließend urteilen möchte. Aber eine statistische Einheit ist immer Teil der statistischen Masse, Stichprobe immer ein Teil der Grundgesamtheit.

Stellen wir uns zum besseren Verständnis noch ein paar Fragen und ordnen sie der Beschreibenden oder Schließenden Statistik zu:

- Eine Skischule hat 23 Anmeldungen für die morgen beginnenden Anfängerkurse, davon sind 12 Teilnehmer unter 6 Jahre alt, 7 unter 10 und der Rest älter. Das ist eine Häufigkeitsverteilung, die in der Beschreibenden Statistik aufgestellt wird.
- An einer Abfüllanlage wird zwischen 14.00 und 15.00 Uhr gemessen, dass alle Bierflaschen den Mindestinhalt von 0,5 Litern enthalten. Man geht nun davon aus, dass die Anlage Flaschen immer korrekt mit mindestens 0,5 Litern befüllt. Eine Frage der Schließenden Statistik.
- Bei der Wahl zur Miss Ruhrgebiet werden die Kandidatinnen befragt, was die Hauptstadt der Mongolei sei. Nur 2 der insgesamt 24 Kandidatinnen können die Frage korrekt beantworten. Dieses Ergebnis kam im Rahmen einer Beschreibenden Statistik zustande; 2 von 24 (oder 8,33 %) sind die absolute (oder relative) Häufigkeit. Nun vergleicht man das Ergebnis mit den 16 Kandidatinnen aus der Schwäbischen Alb, dort können die Frage ebenfalls 2 Kandidatinnen beantworten. Immer noch bewegen wir uns im Rahmen der Beschreibenden Statistik, in beiden Fällen untersuchen wir (verschieden große) statistische Massen. Wenn wir aber ausgehend von den 8,3 % im Ruhrgebiet oder von den 12,5 % im Schwabenland auf die gesamte Bundesrepublik schließen, dann…
- Am Gambacher Dreieck werden an einem Tag 84 Autofahrer wegen erhöhter Geschwindigkeit geblitzt, das spült der Gemeinde 3500 Euro in die Kasse. Der Stadtkämmerer möchte diese Einnahmen auf das Jahr hochrechnen und fest in seinem Etat einplanen, eine typische Frage der Beschreibenden Statistik.
- Familie Schuster hatte in den letzten Jahren sehr viel Glück. Ihr Bauernhof im Allgäu und die dazu gehörenden landwirtschaftlichen Flächen wurden von der Gemeinde (was auf Grund der familiären Verflechtungen zum Bürgermeister kein großes Problem war) durch einen Skilift erschlossen und als Baugebiet freigegeben. Die Familie baute daraufhin insgesamt 5 Ferienwohnungen, die sie zu je 130 Euro pro Tag vermieten kann. Im ersten Jahr beträgt die Auslastung 80 %. Durch den Klimawandel muss aber in den folgenden Jahren mit einer geringeren Auslastung gerechnet werden, man geht von einem Rückgang der Buchungszahlen von 3 % je Jahr aus. Rechnet sich die Investition? Da hier ein Auslastungsgrad fortgeschrieben wird, handelt es sich um eine Zeitreihenanalyse im Rahmen der Beschreibenden Statistik. Sollte Familie Schuster allerdings auf die Idee kommen, auf Grund der beobachteten Auslastung weitere Ferienwohnungen zu erstellen und möchte sie daher die Auslastung auf diese neuen Ferienwohnungen hochrechnen, bewegt sie sich im Rahmen der Schließenden Statistik.

0.2 Warum man manchmal nicht alle Daten untersuchen kann: das Ziel der Schließenden Statistik

Warum gibt es überhaupt das Themengebiet der Schließenden, Analytischen oder Induktiven Statistik? Versuchen wir es mit folgendem Beispiel:

Das ist Marvin.

Marvin studiert an einer Hochschule im mittlerweile 9. Semester Wirtschaft. Da er weiß, dass neben den Kenntnissen in Statistik, Mathematik, Personalwesen, Organisation, Führungslehre bei Personalchefs auch sogenannte Schlüsselqualifikationen gern gesehen werden, engagiert er sich in einer studentischen Initiative, die Firmenkontaktmessen vorbereitet und durchführt. Täglich isst er in der nahe dem Büro gelegenen sogenannten kleinen Mensa, da es dort nicht so voll ist und er sich dort nur zwischen zwei verschiedenen Gerichten entscheiden muss, nämlich »mit Fleisch« oder »vegetarisch«.

Das alles hat, wie wir bereits wissen, mit Statistik wenig zu tun. Alle Daten, die man über Marvin und die kleine Mensa wissen möchte, könnte man in Erfahrung bringen. Statistik beginnt erst wieder dann, wenn man nicht nur ihn betrachtet, sondern z. B. alle Studierenden (und natürlich auch nur das betrachtet, was für eine Untersuchung wesentlich erscheint).

Wenn man die Daten in eine Tabelle untereinander schreibt (nicht nur die folgenden 9 Datensätze, sondern die der mehreren hundert Besucher der kleinen Mensa, wenn die Tabelle also mehrere Seiten lang ist), geht jeglicher Überblick verloren. Die Frage, wer und vor allem wie viele Studierende jeden Tag in der Mensa vegetarisch oder mit Fleisch essen und dies vermutlich auch morgen tun werden, lässt sich nur mit Mühe beantworten.

Name	Alter	Geschlecht	Studiengang	Semester	Mittagessen
Marvin	23	männlich	BWL	9	mit Fleisch
Lisa	23	weiblich	VWL	8	vegetarisch
Kevin	24	männlich	BWL	7	mit Fleisch
Alex	28	männlich	BWL	16	vegetarisch
Yusuf	25	männlich	BWL	11	vegetarisch
…	…	…	…	…	…
Lenni	25	weiblich	VWL	9	mit Fleisch
Aische	23	weiblich	BWL	7	vegetarisch
Jason	24	männlich	BWL	9	mit Fleisch
Alex	21	männlich	VWL	3	vegetarisch

Wir wissen: Hier setzt Statistik an. Sie will unübersichtliche Daten komprimieren und dadurch übersichtlich gestalten. Es soll nicht die Frage beantwortet werden, ob Marvin oder auch Lisa im 9. Semester sind, sondern, wie viele Studierende insgesamt im 9. Semester BWL sind. Oder wie viele in der gesamten Hochschule im 9. Semester studieren, denn damit kann man vorausberechnen, wie viele demnächst im 10. sein werden und damit aus der BaföG-Förderung herausfallen. Wir wollen nicht wissen, ob Jason lieber Fleisch und Alex lieber vegetarisch isst, sondern wie viele insgesamt diese Vorlieben haben, um daraus das Angebot der nächsten Tage zusammenzustellen. Statistik ordnet also als erstes Daten, bringt sie in eine übersichtliche Reihenfolge und zählt dann die beobachteten Werte.

Notwendigerweise wiederholen wir an dieser Stelle einige Begriffe der statistischen Methodenlehre: An den statistischen Einheiten (Merkmalsträgern) beobachtet man ein oder mehrere Eigenschaften (Merkmale), deshalb nennt man so etwas das Untersuchungsmerkmal. Wenn wir uns wie hier dafür interessieren, in welchem Semester die Studierenden sind, dann betrachten wir ein bestimmtes Untersuchungsmerkmal X_1 (großer Buchstabe X), in dem vorliegenden Fall das Semester. Weitere Untersuchungsmerkmale wären das Geschlecht (könnte damit X_2 sein), das Studienfach (X_3), das Alter (X_4) und das geliebte Essen (X_5). Wenn man nicht nur Marvin betrachtet (die erste statistische Einheit E_1), sondern alle Besucher dieser Mensa (3240 Studierende), haben wir es mit 3240 statistischen Einheiten E_1 bis E_{3240} und 5 Untersuchungsmerkmalen zu tun (eigentlich 6, denn auch der Name könnte ein Untersuchungsmerkmal sein, wenn wir z. B. die beliebtesten Namen herausfinden wollen). Wenn man vorher nicht genau weiß, wie viele statistische Einheiten man untersucht, endet man nicht bei 3240, sondern allgemein bei n, also E_1, E_2, E_3, ..., Ej, …, E_n; j läuft von 1 bis n ($j = 1, 2, 3, \ldots, n$). Alle statistischen Einheiten zusammen ergeben die statistische Masse, das können alle Mensabesucher einer Hochschule, aber auch alle eines Bundeslandes sein. Wie groß die statistische Masse ist, hängt letztendlich vom Ziel einer Untersuchung ab.

Das, was dabei tatsächlich herauskommt, ist der Beobachtungswert b. »Mit Fleisch« wäre der erste Beobachtungswert (b_1) für das Lieblingsessen des ersten untersuchten Studierenden (E_1 = Marvin), »vegetarisch« (b_2) ein anderer (für E_2 = Lisa), »mit Fleisch« (b_3) ein weiterer (für E_3 = Kevin) usw. (b läuft also auch von 1 bis n). Ein Beobachtungswert ist also das, was man für ein Untersuchungsmerkmal bei einer beliebigen statistischen Einheit tatsächlich gemessen hat. Die Merkmalausprägungen x (jetzt der kleine Buchstabe x) bezeichnen die Werte, die grundsätzlich vorkommen können und zu denen man die b zusammenfassen kann. In dieser Mensa gibt es normalerweise Studierende, die x_1 = »mit Fleisch«, x_2 = »vegetarisch« mögen. Diese möglichen Merkmalsausprägungen können wir nicht mit j indizieren, denn j ist schon für die Beobachtungswerte reserviert; x wird deshalb mit dem Buchstaben i versehen (allgemein $x_1, x_2, x_3, \ldots, x_i, \ldots, x_m$; $i = 1, 2, 3, \ldots, m$). Übrigens müssten nicht alle diese Werte in einer Mensa tatsächlich beobachtet werden; theoretisch gibt es vielleicht keinen Studierenden, der vegetarisch isst (während das in anderen Mensen der Fall sein könnte) oder auch keinen, der 27 Jahre alt ist.