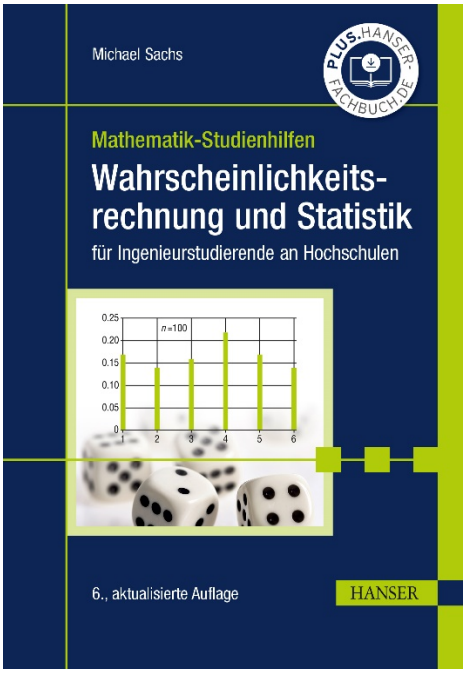
HANSER



Leseprobe

zu

Wahrscheinlichkeitsrechnung und Statistik

von Michael Sachs

Print-ISBN: 978-3-446-46943-3
E-Book-ISBN: 978-3-446-46962-4

Weitere Informationen und Bestellungen unter
https://www.hanser-kundencenter.de/fachbuch/artikel/9783446469433
sowie im Buchhandel

Vorwort zur 6. Auflage

Für die 6. Auflage habe ich zahlreiche Beispiele und Aufgaben mithilfe aktueller Veröffentlichungen auf den neuesten Stand gebracht und hartnäckige Druckfehler beseitigt. Neu aufgenommen wurden Boxplots und aus aktuellem Anlass die Begriffe Sensitivität und Spezifität von Schnelltests. In der Wahrscheinlichkeitsrechnung habe ich die bedingten Wahrscheinlichkeiten vorgezogen und daraus den Begriff der Unabhängigkeit von Ereignissen abgeleitet, weil dieser Zugang nach meiner langjährigen Erfahrung den Studierenden plausibler ist. Ich danke allen Leserinnen und Lesern, die mir Verbesserungshinweise gegeben haben, sowie den Mitarbeiterinnen des Hanser Verlags für die stets angenehme Zusammenarbeit.

München, im März 2021 Michael Sachs

Aus dem Vorwort zur 5. Auflage

„Wozu brauchen wir das alles?“

So lautet eine häufig gestellte Frage von Studierenden, besonders an Hochschulen für angewandte Wissenschaften, wo immer der Aspekt der Anwendung im Vordergrund steht. Der vorliegende Band aus der Reihe „Studienhilfen Mathematik“ versucht, für den Bereich der Statistik eine befriedigende Antwort auf diese Frage zu geben. In allen technischen Studiengängen machen Studierende bei der Durchführung von Versuchen die Erfahrung von zufälligen Einflüssen. Die Erforschung von deren Gesetzen ist Gegenstand der **Wahrscheinlichkeitsrechnung**, ihre Anwendung Gegenstand der **schließenden Statistik**. Am Anfang steht ein Kapitel über **beschreibende Statistik**, das vollkommen ohne den Wahrscheinlichkeitsbegriff auskommt.

Es kommen keine schwierigen Beweise oder umfangreiche Theorien vor. Ingenieurinnen und Ingenieure dürfen sich hier getrost auf die gesicherten Ergebnisse der Mathematik verlassen und sollen vielmehr die Aussagen verstehen und richtig einschätzen lernen, die hinter solchen Sätzen stehen, sowie die Methoden sinnvoll anwenden und ihre Ergebnisse korrekt interpretieren können. Aus der höheren Mathematik werden Kenntnisse der elementaren Funktionen einer reellen Veränderlichen und ihrer Ableitungen sowie des Riemannschen Integrals vorausgesetzt.

Jeder Abschnitt stellt in Lehrsätzen und gelösten Aufgaben die Hilfsmittel bereit, mit denen man die anschließenden Übungsaufgaben lösen kann. Die Ergebnisse der Aufgaben sind zur Kontrolle im Anhang angegeben.

Inhaltsverzeichnis

1 Wozu Statistik?

„Statistik infomiert. Alle.“

Mit diesen Worten beginnt eine Broschüre aus dem Jahr 1990, herausgegeben vom Bayerischen Landesamt für Statistik und Datenverarbeitung[1]. Die dort getroffenen Aussagen gelten vorwiegend für den wirtschaftlichen und sozialen Bereich, lassen sich jedoch unschwer auch auf das Arbeitsumfeld von Ingenieurinnen und Ingenieuren übertragen. In einer sich rasch verändernden Welt, in der Massendaten anfallen und mit Computerhilfe auch schnell verarbeitet werden können, ist es wichtig

- zu wissen, welchen Zustand wir heute erreicht haben,
- mit gestern zu vergleichen,
- zukünftige Zustände abzuschätzen,
- Aussagen zu machen über die Auswirkungen getroffener Maßnahmen.

Diese Aufgaben müssen gelöst werden auf der Basis meist riesiger Datenmengen, die wegen ihres Umfanges keinem Entscheidungsträger unbearbeitet vorgelegt werden können. Aufgabe der **beschreibenden** oder **deskriptiven Statistik** (Kapitel 2) ist es daher, große und unübersichtliche Datenmengen so aufzubereiten, dass wenige aussagekräftige Kenngrößen und/oder Grafiken entstehen, mit denen dann die genannten Fragestellungen gelöst werden können. In diesen Kenngrößen ist dann die gesamte Datenmenge gewissermaßen „fokussiert“.

Beispiel 1.1

Eine Zeugnisnote in einem Diplomzeugnis errechnet sich aus sehr vielen Einzelnoten, die evtl. sogar mit unterschiedlicher Gewichtung in die Gesamtnote eingehen. Die beschreibende Statistik liefert die Formel, mit der man aus den Einzelnoten die Gesamtnote berechnet. ■

Die zur Berechnung der Kenngrößen zugrunde gelegte Datenmenge kann zwar sehr groß werden, ist aber prinzipiell immer endlich und spricht nur für sich selbst. Es werden also keinerlei weiterführende Rückschlüsse gezogen auf irgendwelche Einheiten, die nicht untersucht wurden. Bereits im 19. Jahrhundert hat man jedoch erkannt, dass sehr große statistische Massen, wie z. B. die Bevölkerung eines Landes, nicht mehr durch eine vollständige Erhebung zu erfassen sind. Der Aufwand hierfür wäre technisch und organisatorisch viel

[1] heute: Bayerisches Landesamt für Statistik

zu hoch. Während hier eine Totalerhebung zwar schwierig, aber prinzipiell immerhin noch möglich wäre, verhält es sich ganz anders, wenn man Aussagen treffen will z. B. über alle Teile, die eine bestimmte Maschine jemals produziert, oder gar über die Menge aller möglichen Würfe mit einem bestimmten Würfel. In beiden Fällen ist die Grundgesamtheit, also die Menge aller Teile bzw. aller Würfe, von vorneherein überhaupt nicht bekannt, sie liegt nicht so greifbar vor uns wie im Falle der Noten eines Faches.

In solchen Fällen muss man sich auf **Stichproben**, also Teilerhebungen beschränken. Die gesuchten Kenngrößen können dann nicht mehr exakt bestimmt, sondern nur noch geschätzt werden. Hier kommt aber der Begriff des **Zufalls** ins Spiel: Ziehen zwei Leute eine Stichprobe aus der gleichen Grundgesamtheit und berechnet jeder das arithmetische Mittel $\bar{x}$ seiner Stichprobe, so werden sie im Allgemeinen zu unterschiedlichen Ergebnissen kommen. Welche Elemente in die jeweilige Stichprobe gelangen, hängt nämlich von Faktoren ab, die der Stichprobenzieher nicht bestimmen kann.

Kapitel 4 ist daher der **schließenden** oder **induktiven Statistik** gewidmet, die Methoden entwickelt, um von einer Stichprobe auf die Gesamtheit schließen zu können.

Beispiel 1.2

Eine große Firma kauft 100 Personal Computer bei Händler A und 200 bei Händler B. Im Laufe des ersten Jahres gehen bei den A-Computern zwölf Netzteile kaputt, bei den B-Computern 16 Netzteile. Der Beitrag der beschreibenden Statistik ist hier eher gering: Sie gibt uns die Formel, um die Anteile p_A und p_B der Computer mit defektem Netzteil zu berechnen:

$$p_A = \frac{12}{100} = 12\,\% \quad \text{und} \quad p_B = \frac{16}{200} = 8\,\%.$$

A hat also schlechtere Qualität in Bezug auf das Merkmal „Lebensdauer des Netzteils“ geliefert als B. Aber kann man daraus schließen, dass A **generell** einen höheren Ausschussanteil hervorbringt als B, oder könnte das schlechte Ergebnis für A auch einfach Zufall gewesen sein? Diese Frage kann nur die schließende Statistik beantworten. Wir werden auf dieses Beispiel im Abschnitt 4.3 zurückkommen. ■

Beispiel 1.3

Um eine Prognose über den Ausgang einer Wahl treffen zu können, werden 100 ausgewählte Personen befragt. 45 antworten, sie würden die A-Partei wählen. Die beschreibende Statistik kann dann nur sagen,

dass 45 % der Befragten A-Anhänger sind. Die schließende Statistik dagegen versucht, ausgehend von diesem Ergebnis Aussagen über **alle** Wahlberechtigten zu machen (die berühmte Hochrechnung). Kann man ausgehend von der Stichprobe mit Umfang 100 schon sagen, dass Partei A unter allen Wahlberechtigten 45 % der Stimmen hat? Rein gefühlsmäßig erscheinen 100 als zu wenig, aber mit welcher Genauigkeit kennen wir den wahren Anteil der A-Wähler, wenn wir 1 000 oder gar 10 000 Personen befragen? Hängt die Anzahl der zu befragenden Personen von der Gesamtzahl der Wahlberechtigten ab oder nicht? Was heißt überhaupt in diesem Zusammenhang „Genauigkeit“? Dies sind Fragen, die die schließende Statistik präzisieren und beantworten muss. ■

Als Grundlage für die schließende Statistik müssen wir also zunächst den Begriff des Zufalls präzisieren. Hierfür steht eine mathematische Theorie zur Verfügung, die **Wahrscheinlichkeitsrechnung**. Ihre wichtigsten Erkenntnisse stehen in Kapitel 3.

Beschreibende und schließende Statistik werden auch zusammengefasst unter dem Begriff **statistische Methodenlehre**. Darüber hinaus wird der Begriff **Statistik** aber auch angewendet für die einzelne Tabelle, („die Arbeitslosenstatistik“, „die Unfallstatistik“). Schließlich meint Statistik auch noch die praktische Tätigkeit des Registrierens und Auswertens von Daten, z. B. in dem Satz „Die amtliche Statistik in Bayern wird vom Landesamt für Statistik wahrgenommen“.

Wir fassen zusammen:

Definition 1.1

Die statistische Methodenlehre ist eine Hilfswissenschaft. Ihre Aufgabe ist es, Methoden für die Erhebung, Aufbereitung, Analyse und Interpretation von Daten bereitzustellen mit dem Ziel, Strukturen in Massenerscheinungen zu erkennen.

2 Beschreibende Statistik

2.1 Grundbegriffe

Die **statistische Masse** besteht aus allen denjenigen Elementen, denen das Interesse des Statistikers gilt. Ihre einzelnen Mitglieder heißen **statistische Einheiten**, **statistische Elemente** oder **Merkmalsträger**. Vor jeder ernst zu nehmenden statistischen Tätigkeit muss die statistische Masse in räumlicher, zeitlicher und sachlicher Hinsicht präzise definiert werden.

Beispiel 2.1

Mögliche statistische Massen sind:

- Natürliche Personen (z. B. Studierende, Kunden, Mitarbeiter),
- Sachen (z. B. Maschinen, Produkte),
- Institutionen, Körperschaften (z. B. Betriebe, Städte, Länder),
- Ereignisse (z. B. Maschinenausfälle, Geburten, Todesfälle, Firmengründungen, Konkurse).

Eine korrekt beschriebene statistische Masse wäre etwa: „Alle Studierenden des Studienganges Bioingenieurwesen an der Hochschule München im WS 2017/2018“. Räumliche Eingrenzung: Hochschule München. Zeitliche Eingrenzung: WS 2017/2018. Sachliche Eingrenzung: Studierende des Bioingenieurwesens. ■

An den statistischen Elementen interessieren uns bestimmte Eigenschaften oder **Merkmale**. Die Werte, die ein Merkmal annehmen kann, heißen **Merkmalsausprägungen**. Ein sinnvolles Merkmal muss mindestens zwei verschiedene Ausprägungen haben.

Beispiel 2.2

Eine Computerzeitschrift testet verschiedene Drucker. Die statistische Masse sind also die getesteten Drucker. Bei jedem Drucker interessieren den Leser folgende Merkmale:

- Hersteller (alle Herstellernamen),
- Bezeichnung des Gerätes (alle Bezeichnungen),
- Preis (0 € – 10 000 €),
- Gewicht (0 kg – 10 kg),

- Seitenzahl pro Minute (0 – 100 Seiten),
- Gesamturteil (sehr gut, gut, mittel, schlecht, sehr schlecht).

In Klammern stehen jeweils die möglichen Merkmalsausprägungen. ■

Bezüglich der Art ihrer Merkmalsausprägungen (Werte) lassen sich Merkmale wie folgt in Kategorien einteilen:

- Ein **qualitatives Merkmal** liegt vor, wenn die Werte keine physikalische Einheit brauchen. Man unterscheidet dabei **qualitativ-nominale** Merkmale von **qualitativ-ordinalen** Merkmalen: Bei einem nominalen Merkmal sind die Merkmalsausprägungen nur dem Namen nach unterscheidbar, sie drücken aber keinerlei Wertung oder Intensität aus. Bei einem ordinalen Merkmal[1] unterscheiden sich die Ausprägungen nicht nur hinsichtlich ihrer Namen, sondern können zusätzlich noch in eine (inhaltlich sinnvolle) Rangordnung gebracht werden.
- Ein **quantitatives Merkmal**[2] liegt vor, wenn die Nennung des Wertes allein noch keinen Sinn ergibt, weil die Einheit fehlt. Man unterscheidet hier zwischen **quantitativ-diskret** und **quantitativ-stetig**: Diskrete Merkmale haben Werte, die durch einen Zählprozess entstehen. Die Werte sind dann meist die natürlichen Zahlen 0, 1, 2, ..., die Einheit ist 1 oder bei größeren Anzahlen auch oft 1 000. Zwischen den einzelnen Ausprägungen können keine Werte angenommen werden. Stetige Merkmale dagegen werden durch Messung gewonnen und können (theoretisch, je nach Genauigkeit des Messgerätes) jeden Wert innerhalb eines sinnvollen Intervalls annehmen.

Beispiel 2.3

Im letzten Beispiel sind die Merkmale „Hersteller“, „Bezeichnung“ und „Gesamturteil“ qualitativ, denn ihre Werte benötigen keinerlei Einheit. „Gesamturteil“ ist darüber hinaus noch ordinal, denn die Werte können in eine Rangskala von „sehr gut“ bis „sehr schlecht“ geordnet werden. Die übrigen Merkmale sind quantitativ, „Preis“ und „Gewicht“ stetig, „Seitenzahl“ diskret. (Streng genommen ist „Preis“ ebenfalls ein diskretes Merkmal, denn Preise können keine Werte zwischen zwei aufeinanderfolgenden Cent annehmen. Hier ist die Einteilung aber so fein, dass man „Preis“ meist als stetiges Merkmal behandelt.) ■

[1] auch *Rang*-Merkmal
[2] auch *metrisches* oder *kardinales* Merkmal

Die Kategorie eines Merkmals hat Einfluss auf die Gestaltung von Fragebögen und grafischen Bildschirmoberflächen: Bei einem qualitativen Merkmal mit nur wenigen Ausprägungen (z. B. Familienstand, Geschlecht) wird man alle Alternativen auflisten mit der Möglichkeit, die zutreffende anzukreuzen. Ein qualitatives Merkmal mit sehr vielen möglichen Ausprägungen (z. B. Name, Wohnort, Staatsangehörigkeit) wird man als Freitextfeld zum Ausfüllen realisieren. Quantitative Merkmale schließlich haben sehr viele Werte und werden daher als Freitextfeld dargestellt, oder man fasst die Werte zu wenigen Intervallen zusammen (Klassierung, siehe Abschnitt 2.2) mit der Möglichkeit anzukreuzen.

Zum Zwecke der effizienten Abspeicherung von Merkmalsausprägungen in DV-Anlagen werden die Ausprägungen oft **verschlüsselt**, d. h., den einzelnen Werten werden Zahlen zugeordnet, z. B. beim qualitativ-nominalen Merkmal Familienstand: „0“ für „ledig“, „1“ für „verheiratet“, „2“ für „verwitwet“ usw. Das heißt aber **nicht**, dass daraus jetzt ein quantitatives Merkmal entstanden ist, was man schon allein daran erkennt, dass „verwitwet“ nicht doppelt soviel ist wie „verheiratet“.

Werden die benötigten Daten durch eine eigens für statistische Zwecke organisierte Erhebung gewonnen, sprechen wir von einer **Primärstatistik**, werden sie dagegen von bereits vorhandenen Verwaltungs- oder Unternehmensdateien für die Statistik „abgezweigt“, von einer **Sekundärstatistik**.

Beispiel 2.4

Ermittelt die „Arbeitsgemeinschaft Fernsehforschung“ (AGF) die Fernseh-Einschaltquoten, so entsteht eine Primärstatistik, dasselbe gilt z. B. für die Qualitätskontrolle in Betrieben. Die monatliche Arbeitslosenstatistik der Bundesagentur für Arbeit dagegen ist eine Sekundärstatistik, da hier auf das bereits in den Arbeitsämtern vorliegende Material zurückgegriffen wird. ■

Werden alle Elemente einer statistischen Masse in die Erhebung einbezogen, liegt eine **Totalerhebung** oder **Vollerhebung** vor, ansonsten eine **Teilerhebung** oder **Stichprobe**. Elemente in Stichproben können zufällig ausgewählt werden (**Zufallsstichprobe**, sie wird im Abschnitt 4.1 behandelt), oder bewusst zusammengesetzt werden, sodass in der Stichprobe die Werte gewisser Merkmale mit den gleichen relativen Häufigkeiten („Quoten“) repräsentiert sind wie in der Gesamtheit (**repräsentative Stichprobe**). Um repräsentative Stichproben zusammensetzen zu können, braucht man Informationen über die zugrunde liegende Grundgesamtheit aus früheren Untersuchungen.

Beispiel 2.5

Die AGF führt Teilerhebungen zur Fernsehzuschauerforschung durch. Dieses Fernsehpanel umfasst ca. 5 400 täglich berichtende Haushalte, in denen rund 11 000 Personen leben (Stichprobe). Damit wird die Fernsehnutzung von 75,3 Mio. Personen ab drei Jahren (Grundgesamtheit) abgebildet (Stand 04.01.2021, Quelle: [12]). Sind also etwa in der Gesamtbevölkerung 12 % zwischen 14 und 29 Jahre alt, so müssen in einer repräsentativen Stichprobe von 10 000 Testpersonen ca. 1 200 Testpersonen zwischen 14 und 29 Jahre alt sein. ■

Für die anschließende Veröffentlichung der Daten in Form einer Tabelle gibt es die DIN-Norm 55 301 ([9]). Auch wenn sie inzwischen zurückgezogen wurde, ist es sicher nicht verkehrt, sich an einige Grundregeln dieser Norm zu halten. Bild 2.1 zeigt die vier Hauptbestandteile einer Tabelle. Der Tabellenkopf

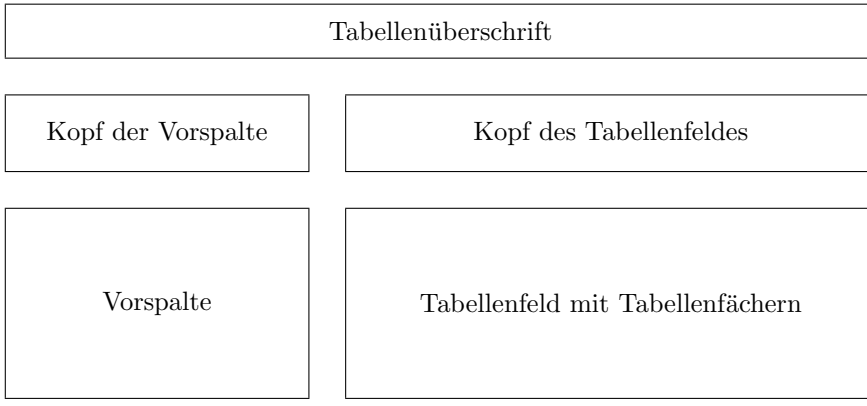


Bild 2.1: Bestandteile einer Tabelle

kann dabei Unterstrukturen enthalten. Bei quantitativ-stetigen Merkmalen müssen die Einheiten im Kopf oder in der Überschrift stehen. Tabellenfächer sollten niemals leer sein, es können aber u.a. folgende Symbole verwendet werden (siehe [9], 10.6):

– nichts vorhanden
... Angabe fällt später an
/ Zahlenwert nicht sicher genug
. Zahlenwert unbekannt oder geheimzuhalten
x Tabellenfach gesperrt, weil Aussage nicht sinnvoll

Sachwortverzeichnis