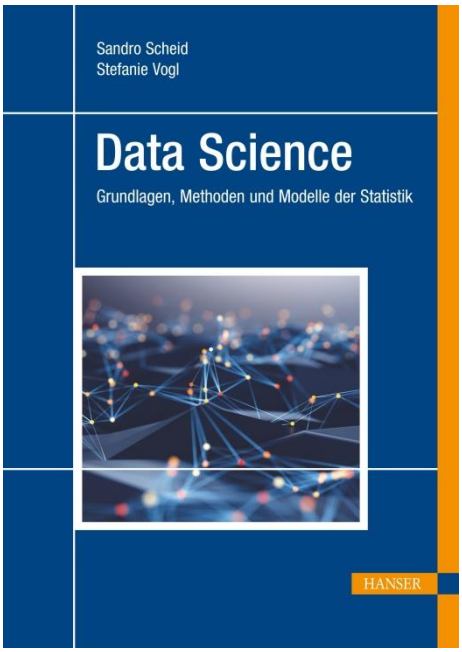
HANSER



Leseprobe

zu

Data Science

von Sandro Scheid and Stefanie Vogl

Print-ISBN: 978-3-446-46663-0
E-Book-ISBN: 978-3-446-47001-9

Weitere Informationen und Bestellungen unter

https://www.hanser-kundencenter.de/fachbuch/artikel/9783446466630

sowie im Buchhandel

© Carl Hanser Verlag, München

Inhalt

Vorwort

In diesem Buch, das aus Vorlesungen für Studierende der Betriebswirtschaftslehre entstanden ist, werden wichtige Methoden der Datenanalyse vorgestellt. Statistische Verfahren werden stets dann benötigt und eingesetzt, wenn im Rahmen empirischer Fragestellungen Daten erhoben, dargestellt und analysiert werden sollen. In allen empirischen Wissenschaften – wir nennen beispielhaft die Wirtschaftswissenschaften – hat die Statistik daher eine große praktische Bedeutung. Zum Studium dieser Wissenschaften gehört deshalb auch eine intensive Beschäftigung mit Statistik.

Dieses Buch ist ein idealer Begleiter zu den Statistikvorlesungen des Bachelor-Studiums an Hochschulen und Universitäten und für das Nacharbeiten statistischer Themen im Masterstudium. Für Praktiker bietet es die Gelegenheit, sich im Selbststudium mit statistischen Fragestellungen zu befassen.

Das Buch setzt keine besonderen mathematischen Kenntnisse voraus. Der Leser benötigt die üblichen Grundkenntnisse der Elementarmathematik, wie sie an Fachoberschulen und Gymnasien unterrichtet wird. Die darüber hinausgehenden Anforderungen beschränken sich im Wesentlichen auf das Rechnen mit dem Summenzeichen.

Das Buch behandelt im ersten Teil (Kapitel 1 bis 3) die deskriptive Statistik, mit der die Datenanalyse beginnen sollte. Im zweiten Teil (Kapitel 4 bis 7) wird die Wahrscheinlichkeitsrechnung behandelt, die als Grundlage für die weiteren Kapitel benötigt wird. Teil drei behandelt die klassischen Themen der induktiven Statistik (Kapitel 8 bis 11). Danach werden im vierten Teil (Kapitel 12 bis 16) verschiedene weiterführende Methoden der Datenanalyse behandelt.

Zu beachten ist, dass die Rechnungen zu den komplexeren Beispielen teilweise mit Hilfe von Software durchgeführt wurden. Hier sind, um die Kumulation von Rundungsfehlern zu vermeiden, Zwischenergebnisse gerundet angegeben. Intern wurde aber mit einer höheren Genauigkeit weitergerechnet. Endergebnisse können sich also beim Rechnen mit gerundeten Zwischenergebnissen leicht unterscheiden.

An dieser Stelle wollen wir unseren Dank aussprechen. Unser Dank gilt zunächst den Studierenden der Hochschule München, die durch zahlreiche Fragen und Anmerkungen zur Gestaltung des Lehrtextes beigetragen haben. Weiter danken wir ganz herzlich Frau Julia Pelger, die durch ihre Korrekturvorschläge eine unersetzliche Arbeit geleistet hat, und Herrn Oliver Hien, durch dessen konstruktive Anregungen und Kommentare der Text weiter wesentlich verbessert werden konnte.

Unser ganz besonderer Dank gilt Herrn Frank Katzenmayer und Frau Christina Kubiak, Hanser Verlag, für die exzellente Betreuung des Buchprojekts.

Kochel am See und Konzell, Sommer 2021 Sandro Scheid und Stefanie Vogl

1

Deskriptive Statistik

1 Grundlagen

Die deskriptive Statistik stellt Methoden bereit, um Untersuchungsgegenstände, die aus einer Vielzahl von einzelnen Objekten bestehen, zu beschreiben. Hinter dieser abstrakten Beschreibung stehen vielfältige unterschiedliche Anwendungen. Beispielsweise soll das Wahlverhalten der deutschen Bürger beschrieben werden oder die Nebenwirkungen bei der Einnahme eines Medikamentes oder das Interesse der Konsumenten an einem neuen Produkt. Die drei genannten Beispiele haben gemeinsam, dass das konkrete Interesse auf eine Gesamtheit bezogen ist. Dies sind die wahlberechtigten Personen in Deutschland, die Patienten, die das Medikament einnehmen oder der Kundenkreis des neuen Produkts. Diese Gesamtheiten sollen durch die statistische Analyse quantitativ beschrieben werden. Dazu werden an den Einheiten – in den genannten Beispielen sind dies Personen – Untersuchungsmerkmale betrachtet. Eine deskriptive Auswertung sollte das Untersuchungsziel und auch die Gesamtheit genauer beschreiben. Wichtige Begriffe, die dabei nützlich sind, werden in diesem Kapitel behandelt.

1.1 Aufgaben der deskriptiven Statistik

Die deskriptive Statistik beschäftigt sich mit der Aufbereitung und Komprimierung von Daten, die in den verschiedensten Bereichen anfallen oder erhoben werden. Mithilfe der deskriptiven Statistik können Sachgebiete quantitativ beschrieben werden. Dabei werden viele Daten im Hinblick auf die Eigenschaften, die von Interesse sind, zusammengefasst. Zum Aufgabengebiet der deskriptiven Statistik zählen im Einzelnen:

- das Zusammenfassen und Ordnen der Daten in Tabellen,
- das Erstellen von Diagrammen und
- das Berechnen charakteristischer Kenngrößen oder Maßzahlen.

Beispiel 1.1

- Der Absatz der verschiedenen Modelle eines Kfz-Herstellers soll graphisch dargestellt werden.
- Die Einkommensstruktur der Bundesbürger soll komprimiert dargestellt werden. Von Interesse ist hier der Anteil der Bevölkerung an der Gesamtbevölkerung, der zu einer Einkommensklasse gehört, wobei das Einkommen in verschiedene Größenklassen aufgeteilt ist.
- Ein Kraftfahrzeugversicherer möchte die eingegangenen Schadensfälle verschiedenen Merkmalen zuordnen. So ordnet er sie dem Kraftfahrzeugtyp und dem Alter der Versicherungsnehmer zu.

■

Deskriptive Statistiken werden in unterschiedlichsten Bereichen benötigt. So stellen etwa die staatlichen Statistikämter unter anderem Statistiken der Bevölkerungsstruktur, der Erwerbstätigkeit, des produzierenden Gewerbes, der Sozialleistungen für Handel und Verkehr oder für das Gesundheitswesen zur Verfügung. Im Bereich Betriebsstatistik erfassen Unternehmen etwa Auftragseingänge, Produktion, Erzeugerpreise sowie Lohn- und Materialkosten. In der Forschung werden mittels Statistik Forschungsgegenstände wie z. B. die Verträglichkeit von Medikamenten, das Wahlverhalten, die Belastungsfähigkeit unterschiedlicher Materialien, das Kaufverhalten der Geschlechter oder das Armutsrisiko von Patchworkfamilien beschrieben.

1.2 Grundgesamtheit und Stichprobe

Ausgangspunkt einer statistischen Fragestellung ist ein Untersuchungsgegenstand, der sich in der Regel auf viele einzelne Objekte bezieht. Dabei sollte der Untersuchungsgegenstand präzise formuliert und außerdem sachlich, räumlich und zeitlich abgegrenzt sein, damit eindeutig feststeht, auf welche konkreten Objekte er sich bezieht. Für die Vielzahl an möglichen Untersuchungsgegenständen sollen nun exemplarisch drei Fälle illustriert werden.

Beispiel 1.2

Eine Firma überlegt, ein neues Produkt einzuführen. Sie möchte wissen, wie es auf dem Markt ankommen würde. Um das herauszufinden, führt das Unternehmen eine Umfrage durch.
Untersuchungsgegenstand ist also die Akzeptanz eines neuen Produkts auf dem Markt. Dazu muss der potenzielle Absatzmarkt festgelegt werden. Welche geographische Größe ist relevant, ist der Markt national, europaweit oder global definiert? Sollen nur Ballungszentren bedient werden oder die ganze Fläche? Welches Geschlecht und welche Altersgruppen kommen infrage? Welche Einkommensklassen? Welche Vertriebswege sollen genutzt werden?

■

Beispiel 1.3

Ein Kfz-Versicherer möchte die Schadensfälle des abgelaufenen Geschäftsjahres nach Schadenshöhe und geografischer Häufigkeit charakterisieren.
Untersuchungsgegenstand ist die Gesamtheit der im abgelaufenen Geschäftsjahr eingegangenen Schadensfälle. Somit müssen Beginn und Ende des Geschäftsjahres durch konkrete Zeitpunkte markiert werden. Sollen sämtliche Versicherungstarife analysiert werden oder gilt das Interesse nur einer Auswahl?

■

Wie an den Beispielen deutlich wird, bezieht sich der Untersuchungsgegenstand stets auf eine Menge von Objekten. Diese werden in der Statistik als Untersuchungseinheiten bezeichnet. Die Menge aller Untersuchungseinheiten bildet dann die Grundgesamtheit.

Untersuchungseinheit
Untersuchungseinheiten sind die Objekte, auf die sich die statistische Analyse bezieht.

Grundgesamtheit
Die Menge aller Untersuchungseinheiten wird als Grundgesamtheit oder Population bezeichnet.

Meist werden in der Praxis nicht alle Untersuchungseinheiten, auf die sich eine Fragestellung bezieht, analysiert. Dies wäre aus organisatorischen und zeitlichen Gründen oft viel zu aufwendig oder sogar vollkommen unmöglich. Häufig ist das auch gar nicht notwendig. Die moderne Statistik stellt nämlich Methoden zur Verfügung, die es ermöglichen, basierend auf einer relativ kleinen Auswahl von Untersuchungseinheiten allgemein gültige Aussagen bezüglich einer weitaus größeren Grundgesamtheit herzuleiten.

Stichprobe
Eine Stichprobe bezeichnet eine Auswahl an Untersuchungseinheiten einer Grundgesamtheit.

Die Vorgehensweise, Ergebnisse einer Teilgesamtheit auf eine übergeordnete Gesamtheit zu verallgemeinern, ist Aufgabe der induktiven Statistik. Die deskriptive Statistik dient ausschließlich nur zur Beschreibung einer vollständig bekannten Grundgesamtheit oder zur Beschreibung einer Stichprobe. Interpretationen, die darüber hinausgehen, sind nicht Gegenstand der Untersuchungen und bleiben der induktiven Statistik vorbehalten.

1.3 Merkmale und Skalenniveaus

Bei einer Untersuchung werden an den Untersuchungseinheiten Merkmale betrachtet, die dem jeweiligen Interesse entsprechen. Merkmale werden mit Großbuchstaben $X, Y, Z, \ldots$ notiert. Allgemein sind an einer statistischen Einheit in der Regel mehrere Merkmale beobachtbar.

Beispiel 1.4

Sind die Untersuchungseinheiten Personen, so könnten folgende Merkmale von Interesse sein:
X = das Monatseinkommen,
Y = die bei der letzten Bundestagswahl gewählte Partei,
Z = das Geschlecht.

■

Beispiel 1.5

Sind die Untersuchungseinheiten Schadensereignisse, die bei einer Versicherung innerhalb eines Zeitraums eingehen, so könnten folgende Merkmale von Interesse sein:
X = die Schadenshöhe,
Y = der Zeitpunkt des Schadens,
Z = die Versicherungshöhe des Versicherungsnehmers.

■

Neben dem Begriff Merkmal wird häufig auch der Begriff Variable verwendet. Beide Begriffe werden synonym angewandt. Der Begriff Variable deutet schon darauf hin, dass die Variable oder das Merkmal für die verschiedenen Untersuchungseinheiten verschiedene Werte annehmen kann. Die Werte, die die Merkmale annehmen, werden als Merkmalsausprägungen bezeichnet. Die möglichen Ausprägungen eines Merkmals bilden den Merkmals- oder Zustandsraum. Die folgende Übersicht enthält nochmals eine Zusammenfassung der beschriebenen Begriffe.

Merkmal
Unter einem Merkmal versteht man bestimmte Aspekte oder Eigenschaften einer Untersuchungseinheit.

Merkmalsausprägung
Den konkreten Wert, den eine Untersuchungseinheit für ein Merkmal annimmt, bezeichnet man als Merkmalsausprägung.

Merkmalsraum
Die Menge der möglichen Ausprägungen eines Merkmals wird mit Merkmals- oder Zustandsraum bezeichnet.

In der Literatur ist es üblich, folgende Begriffe synonym zu verwenden:
Untersuchungseinheit = Merkmalsträger = Objekt = Fall (Case),
Merkmal = Variable = Attribut.

Beispiel 1.6

Im *Beispiel 1.4* wurden Personen als Einheiten betrachtet. Die genannten Merkmale könnten die Werte
X = Monatseinkommen: 1 700 €, 2 100 €, 1 900 €, ...
Y = gewählte Partei: CDU, SPD, Grüne, FDP, ...
Z = Geschlecht: männlich, weiblich
annehmen.

■

Das Beispiel illustriert, dass die Art der Ausprägungen unterschiedlich sein kann. Im Folgenden sollen die verschiedenen möglichen Arten von Merkmalsausprägungen beschrieben werden. Je nach Art der Merkmalsausprägungen werden sich später die möglichen statistischen Methoden, die zur Auswertung der Daten herangezogen werden können, unterscheiden.
Eine grobe Einteilung der Merkmalsausprägungen ergibt sich durch die Unterscheidung in zahlenmäßige und nicht zahlenmäßige Werte. Die verwendeten Begriffe sind:

Qualitative Merkmale
Die Ausprägungen unterscheiden sich in ihrer Art. Sie können nicht durch eine Größenordnung beschrieben werden, aber in verschiedene Kategorien eingeteilt werden.

Quantitative Merkmale
Die Ausprägungen lassen sich durch Zahlen beschreiben. Sie besitzen eine Ausprägung, bei der die Größe interpretiert werden kann.

Oft werden die Kategorien von qualitativen Merkmalen im praktischen Gebrauch mit Zahlen belegt. So stehen z. B. Steuer- oder Kundennummern für verschiedene Personen oder Steuerklassen für unterschiedliche Personengruppen. Bankleitzahlen stehen für Banken, Postleitzahlen für Orte. Wichtig für das Verständnis des Begriffs des qualitativen Merkmals ist hierbei, dass die jeweiligen Zahlen in diesen Fällen nur der Unterscheidung dienen. Eine Anordnung der Merkmale im Sinne einer Größe ist hier nicht sinnvoll.

Beispiel 1.7

- Qualitative Merkmale: Geschlecht, Familienstand, Konfession, Steuerklasse, Farbe eines Autos, Rechtsform eines Unternehmens.
- Quantitative Merkmale: Körpergröße, Alter, Einkommen, Umsatz, Temperatur.

■

Quantitative Merkmale lassen sich weiter unterteilen, je nachdem, ob die Merkmalsausprägungen nur diskrete oder kontinuierliche Werte annehmen können.

Diskrete Merkmale
Die Merkmalsausprägungen nehmen nur bestimmte, separate Zahlenwerte an. Es werden nur endlich viele oder abzählbar unendlich viele Werte angenommen.

Stetige Merkmale
Es können Werte aus einem reellen Zahlenintervall angenommen werden. Die Werte sind kontinuierlich. Man stellt sich vor, dass sie auf beliebig viele Nachkommastellen messbar sind.

Bemerkung:

- Alle qualitativen Merkmale sind trivialerweise diskret. Quantitative Merkmale sind dann diskret, wenn die Merkmalsausprägungen durch einen Zählvorgang ermittelt werden können.
- Merkmale, die sich sehr fein unterteilen lassen, aber im Prinzip diskrete Merkmale sind, werden auch quasi-stetige Merkmale genannt und zu den stetigen Merkmalen gezählt.

Beispiel 1.8

- Diskrete Merkmale: Anzahl der Semester eines Studierenden, die Anzahl der Angestellten in einem Betrieb, Anzahl der Fehltage eines Arbeitnehmers.
- Stetige Merkmale: Körpergröße, Temperatur, Zeit, Benzinverbrauch.
- Quasi-stetige Merkmale: Einkommen eines Haushalts, Umsatz einer Firma, Quadratmeterpreis einer Wohnung.

■

Je nach Art des betrachteten Merkmals können die Merkmalsausprägungen anhand verschiedener Skalen gemessen werden. Im Folgenden sind die verschiedenen Skalenniveaus, geordnet nach dem Informationsgehalt, beginnend beim niedrigsten, beschrieben.

Skalenniveaus

- Nominales Skalenniveau
  Bei einem nominalen Skalenniveau lassen sich die verschiedenen Ausprägungen des Merkmals lediglich unterscheiden. Es gibt keine natürliche Anordnung der Merkmalsausprägungen. Man spricht von einem nominalen Merkmal.
- Ordinales Skalenniveau
  Es gibt eine Rangfolge bzw. Ordnung innerhalb der Ausprägungen des Merkmals. Der Abstand zwischen den Ausprägungen ist aber nicht sinnvoll interpretierbar. Man spricht von einem ordinalen Merkmal.
- Metrisches Skalenniveau
  Die Ausprägungen des Merkmals lassen sich der Größe nach anordnen. Zudem sind die Abstände der Ausprägungen interpretierbar. Man spricht von einem quantitativen bzw. metrischen Merkmal.

Beispiel 1.9 Skalen

- Beispiele für Merkmale, die auf einer nominalen Skala gemessen werden, sind: Geschlecht, Familienstand, Konfession, Augenfarbe, Rechtsform eines Unternehmens, Branche eines Unternehmens.
- Beispiele für Merkmale, die auf einer ordinalen Skala gemessen werden, sind: Leistungsbeurteilung, Bewertung bei einem Schönheitswettbewerb, Zufriedenheit mit einem Produkt mit beispielsweise den Ausprägungen: sehr zufrieden, zufrieden, unzufrieden, sehr unzufrieden.

- Beispiele für Merkmale, die auf einer metrischen Skala gemessen werden, sind: Körpergröße, Wartezeit auf den Bus, Gewinn eines Unternehmens, Einkommen eines Arbeitnehmers.

■

Für die metrischen Merkmale sind weitere Unterteilungen möglich.

Arten metrischer Skalen

- Intervallskala
  Nur die Differenzen zwischen Ausprägungen können interpretiert werden.
- Verhältnisskala
  Das Merkmal besitzt einen absoluten Nullpunkt. Als Folge davon ergibt eine Aussage wie „die eine Ausprägung ist doppelt so hoch wie die andere“ Sinn. Allgemein sind die Verhältnisse der verschiedenen Ausprägungen interpretierbar.

Beispiel 1.10 Metrische Skalen

- Ein Merkmal, das auf einer Intervallskala gemessen wird, ist die Temperatur. Abstände sind hier interpretierbar. Eine Aussage wie „heute ist es 3 Grad wärmer als gestern“ ist sinnvoll. Bei null Grad Celsius handelt es sich nicht um einen natürlichen Nullpunkt. Angenommen heute sind 6 °C und gestern 3 °C gemessen worden, so ist es nicht sinnvoll zu sagen, dass es heute doppelt so warm ist wie gestern.
- Ein Beispiel für ein Merkmal, das auf einer Verhältnisskala gemessen wird, ist das Einkommen eines Arbeitnehmers. Hier bietet es sich an, Verhältnisse zu bilden. Aussagen wie „Peter verdient doppelt so viel wie Dieter“ oder „Petra verdient 15 % mehr als Susanne“ sind zulässig.

■

Sachwortverzeichnis