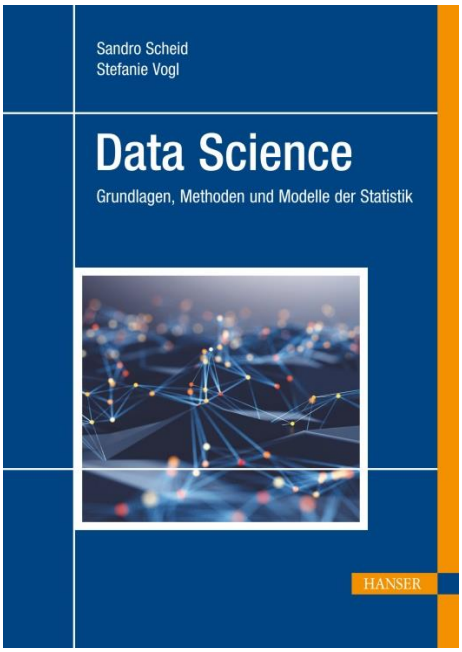
HANSER



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

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.

■

2 Betrachtung eines Merkmals

Um sich einen Überblick bezüglich wesentlicher Eigenschaften eines Merkmals anzueignen, beginnt man mit der Häufigkeitsverteilung. Diese Verteilung beschreibt, wie häufig die einzelnen Merkmalsausprägungen in der Grundgesamtheit zu finden sind. Häufigkeiten lassen sich für jedes Merkmal und jedes Skalenniveau ermitteln. In diesem Kapitel werden – getrennt für diskrete und stetige Merkmale – Häufigkeitsbegriffe erörtert und graphische Darstellungen vorgestellt.
Ferner werden Methoden erarbeitet, mit denen sich die charakteristischen Eigenschaften eines einzelnen Merkmals beschreiben lassen. Die geeigneten Methoden sind abhängig von der Art des jeweiligen Merkmals, insbesondere von dessen Skalenniveau. Man unterscheidet bei diesen statistischen Kenngrößen oder Maßzahlen hierbei Lagemaße und Streuungsmaße. Den Abschluss des Kapitels bildet die Fragestellung der Konzentration von Merkmalen auf Merkmalsträger.

2.1 Häufigkeitsverteilungen bei diskreten Merkmalen

2.1.1 Absolute und relative Häufigkeitsverteilung

Zu den diskreten Merkmalen können wir hier alle qualitativen sowie die quantitativ-diskreten Merkmale zählen. Die Anzahl der unterschiedlichen Merkmalsausprägungen ist in der Regel wesentlich kleiner als der Umfang der Grundgesamtheit und damit überschaubar. So gehören beispielsweise zum qualitativen Merkmal „Geschlecht“ die Ausprägungen „weiblich“, „männlich“ und „divers“ . Durch einfaches Abzählen lässt sich ermitteln, wie häufig die Ausprägungen in der Grundgesamtheit vertreten sind.

Wir bezeichnen mit

N den Umfang der Grundgesamtheit, bestehend aus N Untersuchungseinheiten,

x_i die Merkmalsausprägung des Merkmals X, die bei der i-ten Untersuchungseinheit beobachtet wurde.

Nummeriert man die Untersuchungseinheiten fortlaufend von 1 bis N durch, dann enthält die Urliste die statistischen Daten so, dass jeder Untersuchungseinheit i die Merkmalsausprägung x_i zugeordnet ist.

Beispiel 2.1 Haushaltsgröße von Privathaushalten

Für $N = 40$ Privathaushalte liegen für das Merkmal „Anzahl Personen“ folgende Daten vor:

3	5	2	3	3	2	3	5	3	5	5	3	2	1	4	3	4	4	2	4
3	3	1	4	5	3	2	4	1	4	1	2	2	3	1	3	1	4	2	2

Die Daten liegen in Form einer Urliste vor, die man auch in standardisierter Form mit 40 Zeilen und zwei Spalten notieren könnte, wobei die erste Spalte lediglich zur Numerierung der Untersuchungseinheiten (Objekte) dient. Einem Objekt entspricht ein Haushalt, der das Merkmal X = „Anzahl Personen“ besitzt. Beispielsweise besitzt der erste Haushalt (erstes Objekt) die Merkmalsausprägung $x_1 = 3$, der zweite Haushalt (zweites Objekt) die Merkmalsausprägung $x_2 = 5$, usw., wie man der ersten Zeile in der obigen Urliste entnimmt. ■

Um Fragen der Form „Wie viele Haushalte sind 4-Personen-Haushalte in dieser Grundgesamtheit?“ beantworten zu können, ordnet man den in der Grundgesamtheit vorkommenden unterschiedlichen Merkmalsausprägungen ihre absoluten Häufigkeiten zu, d.h. man ermittelt durch einfaches Abzählen, wie viele 4-Personen-Haushalte es in der Grundgesamtheit gibt.

Allgemein formuliert man diesen Sachverhalt folgendermaßen:
Ein diskretes Merkmal X habe $M \leq N$ verschiedene Ausprägungen $x_1, \ldots, x_M$. Die absolute Häufigkeit einer Ausprägung x_j wird mit h_j bezeichnet. Der Buchstabe j ist der sogenannte Laufindex, der zwischen 1 und M variiert. Die Summe aller absoluten Häufigkeiten h_j entspricht der Anzahl der Untersuchungseinheiten.

Absolute Häufigkeiten

$h(x_j) = h_j$	absolute Häufigkeit der Ausprägung x_j
$h_1, \ldots, h_M$	absolute Häufigkeitsverteilung
$0 \leq h_j \leq N$	$j = 1, \ldots, M$
$h_1 + \ldots + h_M = \sum_{j=1}^{M} h_j = N$	

Will man den Anteil der 4-Personen-Haushalte in der Grundgesamtheit ermitteln, so benötigt man die Anzahl N der Untersuchungseinheiten. Man spricht dann von der relativen Häufigkeit, mit der die Merkmalsausprägung in der Grundgesamtheit vorkommt.

Relative Häufigkeiten

$f(x_j) = f_j$	relative Häufigkeit der Ausprägung x_j
$f_j = \frac{h_j}{N}$	$j = 1, \ldots, M$
$f_1, \ldots, f_M$	relative Häufigkeitsverteilung
$0 \leq f_j \leq 1$	$j = 1, \ldots, M$
$f_1 + \ldots + f_M = \sum_{j=1}^{M} f_j = 1$	

In der Praxis gewinnt man die Häufigkeiten am einfachsten durch das Erstellen einer Strichliste oder – weniger mühsam – mittels einer geeigneten Statistiksoftware.

Beispiel 2.2 Privathaushalte (Fortsetzung)

Für die Daten der 40 Privathaushalte aus *Beispiel 2.1* ergeben sich folgende Häufigkeiten:

Haushaltsgröße x_j	Strichliste	absolute Häufigkeiten h_j	relative Häufigkeiten f_j
1	//////	6	6/40 = 0,150
2	//////////	9	9/40 = 0,225
3	///////////	12	12/40 = 0,300
4	////////	8	8/40 = 0,200
5	/////	5	5/40 = 0,125
$\sum$		40	1,0

Die Anzahl der 4-Personen-Haushalte unter den betrachteten 40 Privathaushalten beträgt also 8, bzw. 20 % der Haushalte sind 4-Personen-Haushalte.

■

Beispiel 2.3 Studienwahl

Neu eingeschriebene Studierende an einer kleinen Hochschule verteilen sich auf die in der Tabelle aufgeführten Fächergruppen. Folgende Anzahlen ergeben sich für ein ausgewähltes Semester:

Fächergruppe	Neueinschr., Anzahl	Neueinschr. in %
Recht-, Wirtschafts- und Sozialwissenschaften	453	49
Ingenieurwissenschaften	232	25
Mathematik, Naturwissenschaften	147	16
Agrar-, Forst- und Ernährungswissenschaften	88	10
$\sum$	920	100,0

Objekt = neu eingeschriebene Studierende an einer kleinen Hochschule
Grundgesamtheit = alle neu eingeschriebene Studierende an einer kleinen Hochschule in einem ausgewählten Semester
Untersuchungsmerkmal (qualitativ, nominalskaliert) X = Fächergruppe
Merkmalsausprägungen: x_1 = Recht-, Wirtschafts- und Sozialwissenschaften, x_2 = Ingenieurwissenschaften, x_3 = Mathematik, Naturwissenschaften, x_4 = Agrar- Forst- und Ernährungswissenschaften
Die zu den Merkmalsausprägungen $x_1, \ldots, x_4$ gehörenden absoluten Häufigkeiten $h_1, \ldots, h_4$ findet man in der Spalte „Neueinschr., Anzahl“, die relativen Häufigkeiten (in Prozent) $f_1, \ldots, f_4$ in der Spalte „Neueinschr. in %“.

■