utb.

Ira Frost

# Statistik für Wirtschafts-wissenschaftler

4. Auflage

utb 5351


**Eine Arbeitsgemeinschaft der Verlage**

Böhlau Verlag · Wien · Köln · Weimar
Verlag Barbara Budrich · Opladen · Toronto
facultas · Wien
Wilhelm Fink · Paderborn
Narr Francke Attempto Verlag / expert Verlag · Tübingen
Haupt Verlag · Bern
Verlag Julius Klinkhardt · Bad Heilbrunn
Mohr Siebeck · Tübingen
Ernst Reinhardt Verlag · München
Ferdinand Schöningh · Paderborn
transcript Verlag · Bielefeld
Eugen Ulmer Verlag · Stuttgart
UVK Verlag · München
Vandenhoeck & Ruprecht · Göttingen
Waxmann · Münster · New York
wbv Publikation · Bielefeld

Ira Frost

# Statistik für Wirtschafts-wissenschaftler

expert

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.


Dischingerweg 5 · D-72070 Tübingen



Internet: www.expertverlag.de
eMail: info@verlag.expert

Einbandgestaltung: Atelier Reichert, Stuttgart
Printed in Germany

utb-Nr.: 5351
ISBN 978-3-8252-5351-6 (Print)
ISBN 978-3-8385-5351-1 (ePDF)

# Vorwort

Es ist offensichtlich, dass Statistik zunehmend in fast alle Disziplinen, ja sogar in den Alltag eindringt. So sind statistische Methoden aus den Wirtschaftswissenschaften nicht mehr wegzudenken. Entsprechend gibt es eine Fülle hervorragender und ausführlicher Lehrbücher zu diesem Fach. Doch gerade diese Fülle scheint viele Studienanfänger zu überfordern. Das vorliegende Buch möchte deshalb insbesondere den *Einstieg* ins Fach Statistik erleichtern und damit den Boden für eine später vertiefende Lektüre bereiten.

Die Grundlage dieses Buches bilden die Vorlesungen und Übungen, die ich für die Studierenden der Betriebswirtschaftslehre an der Hochschule München abgehalten habe. Da dieses Buch Basiswissen vermittelt, genügen in aller Regel die allgemeinen Grundkenntnisse der Schulmathematik. Allerdings ist es häufig die Formelsprache, die „mathematischen Laien“ den Zugang erschwert. Um Zusammenhänge und Vorgänge klar, effizient und universal auszudrücken – gerade in den Wirtschaftswissenschaften – sind Formeln jedoch unverzichtbar.

So finden sich in diesem Buch zahlreiche Formeln. Damit die durch die Formeln dargestellten Zusammenhänge leichter zu erfassen sind, werden diese zusätzlich verbal erläutert bzw. kommentiert. Oft ist es unbefriedigend, fertige Formeln (Ergebnisse) vorgesetzt zu bekommen. Deswegen werden einige dieser Formeln, wie etwa der Verschiebungssatz für die Varianz (Abschnitt 2.5), die Kleinsten-Quadraten-Schätzer (Kapitel 5) sowie ausgewählte Ergebnisse in Schätzen und Testen explizit hergeleitet. Außerdem macht es einfach mehr Spaß, die erworbenen mathematischen Kenntnisse anzuwenden.

Ausführliche, Schritt für Schritt erklärte Beispiele unterstützen das Selbststudium sowie die Vor- und Nachbereitung des Vorlesungsstoffes. Eine richtige Methode zu erkennen und Ergebnisse sachgerecht zu interpretieren setzt voraus, dass man die Instrumente beherrscht. Deswegen werden zusätzlich zu praxisorientierten auch rein technisch ausgerichtete (Rechen-) Beispiele ausgeführt.

Um das Gelernte zu festigen, stehen Übungsaufgaben zum Download unter `https://www.utb-shop.de/9783825253516` bereit. Zu allen Aufgaben können außerdem Musterlösungen heruntergeladen werden. Die Musterlösungen ermöglichen es den Lernenden, ihre *eigenen* Ergebnisse zu überprüfen.

Das Buch gliedert sich in drei Teile und folgt damit den klassischen Statistik-Einführungskursen. Der erste Teil über die *deskriptive Statistik* beginnt mit der Einführung in die Terminologie. Standardverfahren der Datenaufbereitung (Tabellen und Grafiken), Kennzahlen zur Datenbeschreibung sowie Grundlagen der linearen Regression werden eingeführt. Ein Kapitel über Indexzahlen schließt den ersten Teil ab.

Der zweite Teil behandelt die *elementare Wahrscheinlichkeitsrechnung*, die wiederum für die induktive Statistik erforderlich ist. Hier finden sich neben mathematisch anmutenden Grenzwertsätzen auch einige wichtige praxisrelevante Modelle der Wahrscheinlichkeitsverteilung. Wenn auch der Leser die Grenzwertsätze nicht im Detail beherrschen muss, sollte er sich jedoch ihrer Bedeutung bewusst sein. Auf diesen Grenzwertsätzen basieren die praktischen Methoden der Statistik.

Der dritte Teil über die *induktive Statistik* präsentiert ausgewählte anwendungsorientierte Themen aus dem klassischen Bereich der induktiven Statistik, nämlich aus Schätz- und Testverfahren. Für die Auseinandersetzung mit diesen Methoden sind Ergebnisse aus der Wahrscheinlichkeitsrechnung erforderlich.

Ich möchte an dieser Stelle nicht versäumen, Herrn Dr. Arnulf Krais vom expert verlag meinen Dank auszusprechen. Ohne seine Unterstützung wäre dieses Projekt nicht möglich gewesen. Zudem haben viele Personen an diesem Buch mittelbar oder unmittelbar mitgewirkt: die Studierenden der Hochschule München, die durch Gespräche innerhalb und außerhalb der Vorlesungen zahlreiche Anregungen gegeben haben, Harald Frost, Markus Wessler, Helge Röpcke, Alexandra Fuchs-Würth, Alexandra und Lydia Frost. Ihnen allen danke ich sehr. Schließlich danke ich Herrn Hans Wolfertstetter, meinem ehemaligen Lehrer, der trotz seiner Lehrverpflichtungen das Manuskript durchgesehen hat.

## Vorwort zur zweiten Auflage

Die vorliegende Auflage ist gegenüber der ersten im wesentlichen unverändert. Um Missverständnisse zu vermeiden, wurden Textkorrekturen vorgenommen. Ergänzt wurde die neue Auflage durch Konzentrationsmessung (Lorenzkurve, Gini-Koeffizient).

Ich bedanke mich bei Kollegen und Studierenden für Hinweise auf Fehler und Verbesserungsvorschläge. Mein besonderer Dank gilt Herrn Dr. Josef Dietl für das Durchlesen und wertvolle Anregungen.

## Vorwort zur dritten Auflage

In der dritten Auflage wurden einige Druckfehler korrigiert. Vielen Dank für hilfreiche Hinweise von Studierenden und Lesern.

## Vorwort zur vierten Auflage

Diverse Tippfehler wurden in der vierten Auflage beseitigt; das Beispiel in der Einführung ist aktualisiert worden. Ich bedanke mich für die hilfreichen Hinweise von Studierenden und Lesern. Ein besonderer Dank gilt Herrn Patrick Sorg vom *expert verlag*.

# Inhaltsverzeichnis

# Einführung

Statistik ist ein Instrument zur Gewinnung von Informationen aus Daten. Sie beschäftigt sich mit Methoden der Datenaufbereitung und -analyse. Man unterteilt Statistik im Allgemeinen in zwei Teilgebiete: **deskriptive** oder **beschreibende** Statistik und **induktive** oder **schließende** Statistik - auch **Inferenzstatistik** genannt.

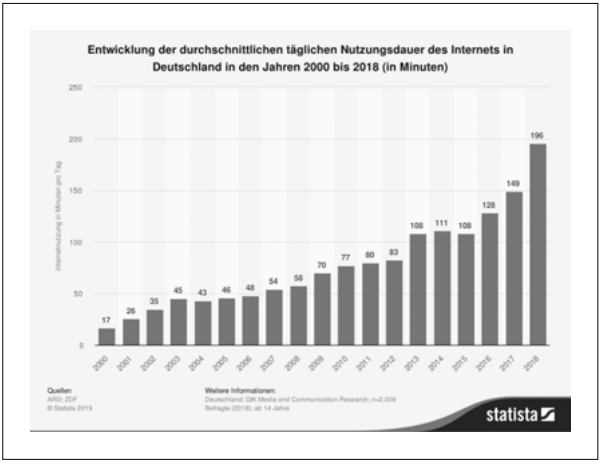
Die Aufgabe der deskriptiven Statistik besteht darin, Informationen aus Daten zu filtern; sie knapp, dennoch aussagekräftig, durch Kennzahlen, Tabellen und Grafiken darzustellen. Jeder von uns kennt sicherlich ein ähnliches Beispiel wie das folgende Ergebnis der Befragung über die durchschnittliche Nutzung des Internets in Minuten pro Tag in den Jahren 2000 bis 2018 in Deutschland. Die Studie wurde von GfK Media and Communication Research durchgeführt; insgesamt haben 2009 Personen ab 14 Jahren in Deutschland daran teilgenommen.[1] Das von ARD/ZDF veröffentlichte Ergebnis wird in Form einer Tabelle (siehe Tabelle 0.1) und einer Grafik (siehe Abbildung) präsentiert. Die im Beobachtungszeitraum jährlich errechnete Kennzahl *Durchschnittswert der Internetnutzung* bildet die Grundlage der beiden Darstellungsformen. An der Grafik erkennt man, dass die durchschnittliche Nutzung des Internets jedes Jahr (mit einer kleinen Unterbrechung in den Jahren 2004 und 2015) zunimmt. Das erneute Wachstum ab 2004 verlief etwas langsamer, bis ein Sprung von 2012 auf 2013 stattfand. Die durchschnittliche Nutzungsdauer bleibt etwa auf diesem Niveau und steigt sichtbar von 2015 bis 2018.

Außer tabellarischen und grafischen Darstellungen von Daten umfasst die deskriptive Statistik Themenbereiche, die aus Veröffentlichungen der Wirtschaftswelt bzw. allgemein aus den Medien vertraut sind, wie etwa Korrelation, Regression und Indexzahlen. Statistische Methoden, die in der induktiven Statistik vorgestellt werden, ermöglichen uns, aus dem Ergebnis der obigen Studie mit 1252 Personen Aussagen über die durchschnittliche Nutzung des Internets in der Bundesrepublik

[1] `https://de.statista.com/statistik/daten/studie/1388/umfrage/taegliche-nutzung-des-internets-in-minuten/` (Stand: 09.09.2019)

| | Werte | | Werte | | Werte |
|---|---|---|---|---|---|
| 2000 | 17 | 2006 | 48 | 2012 | 83 |
| 2001 | 26 | 2007 | 54 | 2013 | 108 |
| 2002 | 35 | 2008 | 58 | 2014 | 111 |
| 2003 | 45 | 2009 | 70 | 2015 | 108 |
| 2004 | 43 | 2010 | 77 | 2016 | 128 |
| 2005 | 46 | 2011 | 80 | 2017 | 149 |
| | | | | 2018 | 196 |

Tabelle 0.1.: Tägliche Nutzung des Internets in Minuten




Deutschland zu treffen[2]. Wir können unter bestimmten Bedingungen auch beurteilen, ob ein Stichprobenergebnis eher als zufällig anzusehen ist oder nicht. Kurz gesagt: Die Inferenzstatistik beschäftigt sich mit Methoden, die Schlüsse aus einer Teilgesamtheit (Stichprobe) auf die Grundgesamtheit ermöglichen. Es liegt auf der Hand, dass solche Aussagen mit Unsicherheiten verbunden sind. Eine Wissenschaft, die sich Unsicherheit und Zufall zu eigen macht, ist die Wahrscheinlichkeitstheorie. Deshalb ist es nur verständlich, dass zahlreiche Ergebnisse aus der Wahrscheinlichkeitstheorie in der induktiven Statistik intensiv genutzt werden.

[2] „Es ist mir noch heute schleierhaft, daß man herausbringt, was sechzig Millionen Menschen denken, wenn man zweitausend Menschen befragt. Erklären kann ich das nicht. Es ist eben so.“ (Elisabeth Noelle-Neumann, Meinungsforscherin). Zitat aus [16]

# Teil I.

# Deskriptive Statistik

# 1. Grundbegriffe

Wir haben in der Einführung die Begriffe *Grundgesamtheit* und *Stichprobe* bereits erwähnt. Wie sie genau definiert sind, erfahren wir jetzt. Eine **Grundgesamtheit** oder **Population** ist eine Gruppe aller uns interessierenden Einheiten, auch **statistische Einheiten** genannt. So bilden beispielsweise alle Internetnutzer ab 14 Jahren in der Bundesrepublik Deutschland eine Grundgesamtheit. Eine Grundgesamtheit muss nicht unbedingt aus Personen bestehen. Die Einheiten können Länder, Gebäude, Unternehmen, Maschinen, Waren etc. sein.

Eine **Stichprobe** ist ein Teil der Grundgesamtheit, der, nach einem bestimmten Verfahren ausgewählt, tatsächlich untersucht wird. Grundsätzlich gibt es zwei Auswahlverfahren: die **bewusste Auswahl** und die **Zufallsauswahl**. Bei einer Zufallsauswahl besitzt jedes Element der Grundgesamtheit die gleiche Chance, in die Stichprobe zu gelangen. Bei einer bewussten Auswahl wie etwa der Quotenstichprobe erfolgt die Auswahl nur teilweise zufällig. Sind beispielsweise 21% der bayerischen Bevölkerung evangelisch, so sollen ebenso 21% der Personen in der Stichprobe der evangelischen Kirchen angehören. Unter Einhaltung dieser Quoten hat ein Interviewer freie Hand.

In der Gfk-Studie zur Internetnutzung bilden die 2009 befragten Personen ab 14 Jahren eine Stichprobe. Die Anzahl der Einheiten in der Stichprobe nennt man **Stichprobenumfang**; er wird in der Regel mit $n$ bezeichnet. (In der GfK-Studie ist $n = 2009$.) Eine **Vollerhebung** liegt vor, wenn alle Einheiten der Grundgesamtheit untersucht werden.

An jeder ausgewählten Einheit wird eine bestimmte *Eigenschaft* beobachtet. Diese Eigenschaft nennen wir **Merkmal** oder **Variable**. Wir bezeichnen sie mit Großbuchstaben wie $X, Y, Z$ o. a.. Die möglichen (unterschiedlichen) Werte eines Merkmals heißen **Merkmalsausprägungen**. In der GfK-Studie wurde das Merkmal *Tägliche Nutzung des Internets in Minuten* untersucht. Die Ausprägungen dieses Merkmals sind nicht negative ganze Zahlen.

Unter **Daten** verstehen wir die beobachteten Ausprägungen in der Stichprobe. Um Daten zu beschaffen, kann man beispielsweise eine Befragung durchführen; sie kann schriftlich, mündlich, telefonisch oder online erfolgen. Auch interne Firmenunterlagen, die amtliche Statistik (`www.destatis.de`) oder Veröffentlichungen der Deutschen Bundesbank (`http://www.bundesbank.de`) können als Datenquelle fungieren. (Ausführlicher zur Stichprobenverfahren und Datenbeschaffung siehe [2], [6], [11].)

Merkmalsausprägungen und Daten bezeichnen wir mit Kleinbuchstaben wie $a_j, x_i, \ldots$. Dabei weist der Index $i$ oder $j$ auf den $i$-ten Datenpunkt oder die $j$-te Ausprägung hin. Um die eingeführten Begriffe zu festigen, betrachten wir nun den folgenden Ausschnitt eines Fragebogens[1]:

Alter: ..... Jahre

Geschlecht Männlich... ☐
Weiblich ... ☐

Wie schätzen Sie Ihre eigenen Mathematik-Kenntnisse ein?
(1 = sehr gut bis 6 = sehr schlecht)
1 ... ☐ 2 ... ☐ 3 ... ☐ 4 ... ☐ 5 ... ☐ 6 ... ☐

Bei dieser Untersuchung werden die Merkmale *Alter* ($X$), *Geschlecht* ($Y$) und *Einschätzung der eigenen Mathematik-Kenntnisse* ($Z$) erhoben. Die Ausprägungen des Merkmals $X$ sind beispielsweise Zahlen zwischen 18 und 34, die des Merkmals $Y$ sind m = männlich oder w = weiblich, und schließlich hat das Merkmal $Z$ die Ausprägungen 1 = sehr gut bis 6 = sehr schlecht. Tabelle 1.1 gibt ein mögliches Ergebnis einer solchen Befragung wieder.

## 1.1. Merkmalsarten

Merkmale, die numerischer Natur sind, heißen **quantitativ**. Beispiele sind u. a. *Alter*, *Einkommen*, *Wohnfläche*. Merkmale mit verbal for-

[1] Eine solche Umfrage könnte beispielsweise im Rahmen einer Verbesserung der Studienbedingungen durchgeführt werden.

| Stud. Nr. $i$ | Alter $x_i$ | Geschlecht $y_i$ | Einschätzung der Mathe.-Kenntnisse $z_i$ |
|---|---|---|---|
| 1 | 20 | m | 1 |
| 2 | 19 | m | 3 |
| 3 | 21 | m | 5 |
| 4 | 20 | w | 1 |
| 5 | 28 | m | 2 |
| 6 | 28 | w | 2 |
| 7 | 34 | w | 5 |
| 8 | 25 | m | 5 |
| 9 | 25 | w | 5 |
| 10 | 25 | m | 3 |
| 11 | 18 | w | 2 |
| 12 | 24 | m | 4 |
| 13 | 19 | w | 3 |
| 14 | 19 | w | 3 |
| 15 | 24 | w | 4 |
| 16 | 21 | w | 2 |
| 17 | 22 | m | 2 |
| 18 | 22 | w | 2 |
| 19 | 20 | w | 1 |
| 20 | 18 | w | 2 |

Tabelle 1.1.: Ergebnis der Befragung unter $n = 20$ Studierenden des ersten Semesters

mulierten Ausprägungen nennt man **qualitativ**. Beispiele sind u. a. *Geschlecht*, *Religionszugehörigkeit*, *Nationalität*. Aber auch das Merkmal *Einschätzung der eigenen Mathematik-Kenntnisse* ist qualitativ, obwohl seine Ausprägungen Ziffern sind. Diese Ziffern stellen eine (im Prinzip) willkürliche Kodierung dar. Man kann sie beliebig ändern, wenn man nur die Ordnung beibehält. Statt die Kodierung „1 = sehr gut“, „2 = gut“ bis „6 = sehr schlecht“ könnte man zum Beispiel auch „10 = sehr gut“, „9 = gut“ bis „5 = sehr schlecht“ wählen.

Bei einer Datenanalyse kann man die verschiedenen Merkmale nicht gleich behandeln. Für das Merkmal *Alter* ergibt zum Beispiel die Aussage „A ist doppelt so alt wie B“ einen Sinn, während eine solche Aussage für die Merkmale *Geschlecht* oder *Einschätzung der Mathematik-Kenntnisse* sinnlos ist.

Für eine adäquate Datenanalyse werden Merkmale nach zwei weiteren Kriterien in Kategorien aufgeteilt. Bezüglich ihrer quantitativen Eigenschaften unterscheidet man drei Skalenniveaus:

1. Auf dem niedrigsten Skalenniveau befindet sich die **Nominalskala**. Alle Ausprägungen nominal skalierter Merkmale lassen sich nur nach ihrer Art unterscheiden und sind gleichwertig. Die Merkmale *Geschlecht*, *Religionszugehörigkeit* oder *Nationalität* sind einige Beispiele dafür.

2. Auf der nächsthöheren Skalenstufe steht die **Ordinalskala**. Zwischen den Ausprägungen ordinal skalierter Merkmale gibt es eine Rangordnung. Beispiele sind u. a. *Einschätzung der eigenen Mathematik-Kenntnisse*, *Beurteilung der Geschäftslage* (zum Beispiel im Rahmen des ifo-Konjunkturtests).

3. Die **metrische** oder **kardinale** Skala befindet sich auf dem höchsten Skalenniveau. In der Literatur findet man eine weitere Unterteilung der Kardinalskala in **Intervall-** und **Verhältnisskala**. Die Verhältnisskala besitzt einen natürlichen Nullpunkt. Beispiele dafür sind *Alter*, *Einkommen* oder *Wohnfläche*. Quotientenbildung ist für die Verhältnisskala sinnvoll, d. h., es lassen sich Aussagen treffen wie „Herr A verdient doppelt so viel wie Frau A. Dafür ist sie halb so alt wie er“. Solche Aussagen sind für intervallskalierte Merkmale nicht sinnvoll, denn die Intervallskala basiert auf Differenzen. Sie besitzt keinen natürlichen Nullpunkt. Ein klassisches Beispiel dazu ist *Temperatur in °C bzw. Fahrenheit*. Wir lassen im Folgenden die Unterscheidung außer Acht und sprechen einfach von der Kardinal- oder metrischen Skala.

Eine weitere Unterscheidung der Merkmale erfolgt nach der Abzählbarkeit ihrer Ausprägungen. Merkmale heißen **diskret**, wenn ihre Ausprägungen *abzählbar* sind, d. h. man kann sie „durchnummerieren“. Beispiele sind u. a. *Haushaltsgröße*, *Anzahl der Kinder*, *Alter in Jahren*. Kann man die Ausprägungen eines Merkmals nicht abzählen, sondern nur messen, dann heißt das Merkmal **stetig**. Beispiele sind u. a. *Alter*, *Wohnfläche*, *Lebensdauer*.

Je nach Ziel einer Studie können stetige in diskrete Merkmale überführt werden. So kann man beispielsweise das stetige Merkmal *Alter* (das Altern kann man als einen kontinuierlichen Prozess in der Zeit ansehen) in Jahren, wie etwa 1 Jahr, 2 Jahre usw., angeben, wodurch es einen diskreten Charakter erhält.

## 1.2. Zusammenfassung

Wichtige Begriffe

| | |
|---|---|
| Merkmale (Variable) | Qualitatives Merkmal |
| Merkmalsausprägungen | Quantitatives Merkmal |
| Statistische Einheiten (Merkmalsträger) | Nominalskala |
| Grundgesamtheit (Population) | Ordinalskala |
| Stichprobe | Kardinalskala (Metrische Skala) |
| Vollerhebung | Diskret |
| Daten | Stetig |

# 2. Eindimensionale Daten

Daten heißen eindimensional, wenn sie nur Werte eines einzigen Merkmals darstellen. Untersucht man an jeder Einheit zwei Merkmale gemeinsam, z. B. $X$: *Alter* und $Y$: *Gewicht*, dann heißen sie zweidimensional. Wir werden uns mit den beiden genannten Fällen beschäftigen und beginnen in diesem Kapitel mit dem eindimensionalen Fall. Die hier gewonnenen Ergebnisse werden in der zweidimensionalen Datenanalyse im nachfolgenden Kapitel verwendet und erweitert.

Sei $X$ ein Merkmal mit $m$ Ausprägungen $a_1, \ldots, a_m$, das an $n \in \mathbb{N}$ Merkmalsträgern gemessen wird. Die Daten über $X$, die an den $n$ Merkmalsträgern beobachtet werden, seien $x_1, \ldots, x_n$. Man nennt diese Daten **Rohdaten** oder **Urliste**[1].

**Beispiel 2.1**
Betrachten wir die folgende Urliste des Merkmals $X$: *Alter* (in Jahren), beobachtet an $n = 10$ Personen:

$$20 \quad 19 \quad 21 \quad 20 \quad 28 \quad 28 \quad 34 \quad 25 \quad 25 \quad 25.$$

Nach unserer Notation lauten die $n = 10$ Beobachtungen (Daten)

$$x_1 = 20,\ x_2 = 19, \ldots, x_{10} = 25$$

und die $m = 6$ Ausprägungen sind:

$$a_1 = 19,\ a_2 = 20,\ a_3 = 21,\ a_4 = 25,\ a_5 = 28,\ a_6 = 34.$$

(Ordnet man sie der Größe nach, so verringert man das Risiko, Ausprägungen zu vergessen.)

Eine weitere Urliste stellt das Ergebnis einer Beobachtung des Merkmals $Y$: *Geschlecht* ($n = 10$) dar:

$$\text{m, m, m, w, m, w, w, m, w, w.}$$

Nach unserer Notation sind:

$$y_1 = \text{m},\ y_2 = \text{m}, \ldots, y_{10} = \text{w}.$$

Die $m = 2$ Ausprägungen lauten: $b_1 = \text{m},\ b_2 = \text{w}$.

[1] $m$ = Anzahl der Ausprägungen, $n$ = Anzahl der Daten = Anzahl der Merkmalsträger, ($m \leq n$)

## 2.1. Häufigkeitstabelle und Grafiken

In der Urliste (Rohdaten) befinden sich Werte, die mehrfach vorkommen. Urlisten sind in der Regel unübersichtlich. Die einfachste Methode, sie zu ordnen, ist das Erstellen einer **Häufigkeitstabelle**. In eine Häufigkeitstabelle trägt man zu jeder Ausprägung[2] $a_j$ die Anzahl ein, wie oft diese beobachtet wurde. Die Häufigkeit, mit der $a_j$ in der Urliste vorkommt, wird mit $f(a_j)$ oder kurz $f_j$ bezeichnet. Man nennt $f_j$ auch die **absolute Häufigkeit** von $a_j$. Die absoluten Häufigkeiten addieren sich zu $n$ (im Anhang finden Sie eine ausführliche Erklärung zum *Summenzeichen* $\sum$):

$$\sum_{j=1}^{m} f_j = n \tag{2.1}$$

Der Anteilswert

$$\boxed{h_j = h(a_j) = \frac{f_j}{n}} \tag{2.2}$$

heißt entsprechend die **relative Häufigkeit** von $a_j$. Die Anteilswerte summieren sich zu Eins:

$$\sum_{j=1}^{m} h_j = 1 \tag{2.3}$$

**Beispiel 2.2**
Zählt man für den Datensatz im Beispiel 2.1, wie oft jeweils eine Ausprägung $a_j$ erscheint, so erhält man die absoluten Häufigkeiten:

$$f_1 = 1,\ f_2 = 2,\ f_3 = 1,\ f_4 = 3,\ f_5 = 2,\ f_6 = 1.$$

Summenbildung dieser Werte ergibt

$$\sum_{j=1}^{6} f_j = 1 + 2 + 1 + 3 + 2 + 1 = 10 = n.$$

Die relativen Häufigkeiten (Anteilswerte) sind

$$h_1 = \frac{1}{10},\ h_2 = \frac{2}{10},\ h_3 = \frac{1}{10},\ h_4 = \frac{3}{10},\ h_5 = \frac{2}{10},\ h_6 = \frac{1}{10}.$$

[2]Der Klarheit halber vereinbaren wir: Der Laufindex $i$ wird für die Werte einer Urliste und der Laufindex $j$ für die Ausprägungen verwendet. Somit gilt: $i = 1, \ldots, n$ und $j = 1, \ldots, m$.

Die Anteilswerte addieren sich zu Eins

$$\sum_{j=1}^{6} h_j = 0,1 + 0,2 + 0,1 + 0,3 + 0,2 + 0,1 = 1,0.$$

| Ausprägung Nr. $j$ | Ausprägung $a_j$ | Absolute H. $f_j$ | Relative H. $h_j$ |
|---|---|---|---|
| 1 | 19 | 1 | 0,1 |
| 2 | 20 | 2 | 0,2 |
| 3 | 21 | 1 | 0,1 |
| 4 | 25 | 3 | 0,3 |
| 5 | 28 | 2 | 0,2 |
| 6 | 34 | 1 | 0,1 |
| | | 10 | 1,0 |

Tabelle 2.1.: Häufigkeitstabelle des Merkmals *Alter* aus Beipiel 2.1

Grafisch kann man die absoluten oder relativen Häufigkeiten als **Stab-** oder **Säulendiagramm** darstellen. Man zeichnet über $a_1, \ldots, a_m$ jeweils einen zur $x$-Achse senkrechten Stab der Höhe $f_j$ bzw. $h_j$ (siehe Abbildung 2.1). Stabdiagramme stellt man auf, wenn lediglich die Häufigkeiten der Ausprägungen dargestellt oder verglichen werden sollen. Sie sind sehr einfach zu konstruieren und für Merkmale mit wenigen Ausprägungen geeignet.

Eine andere grafische Darstellung ist das **Kreis-** oder **Tortendiagramm**. Dabei werden die Häufigkeiten als Kreissektoren abgebildet. Der Kreis stellt die Gesamtheit dar. Kreisdiagramme sind besonders zur Darstellung von Anteilen geeignet. Der Winkel des $j$-ten Kreissektors $\alpha_j$ wird so bestimmt, dass

$$\alpha_j = h_j \cdot 360^\circ. \tag{2.4}$$

Abbildung 2.2 stellt die Daten aus Tabelle 2.1 als Kreisdiagramm dar. Die Winkel der Sektoren sind:

$$\alpha_1 = 0,1 \cdot 360^\circ = 36^\circ,\ \alpha_2 = 0,2 \cdot 360^\circ = 72^\circ,\ \alpha_3 = 0,1 \cdot 360^\circ = 36^\circ,$$

$$\alpha_4 = 0,3 \cdot 360^\circ = 108^\circ,\ \alpha_5 = 0,2 \cdot 360^\circ = 72^\circ,\ \alpha_6 = 0,1 \cdot 360^\circ = 36^\circ.$$

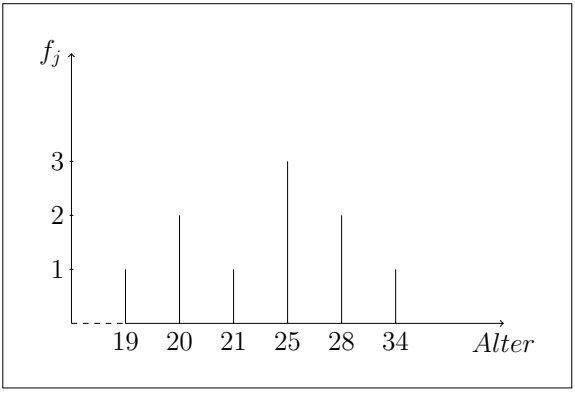



Abbildung 2.1.: Stabdiagramm für das Merkmal *Alter* aus Tabelle 2.1

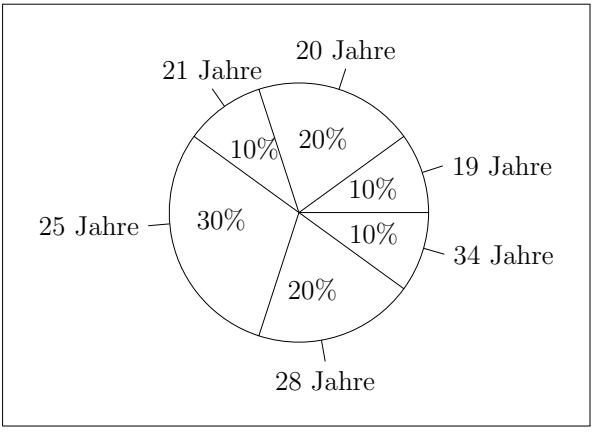



Abbildung 2.2.: Kreisdiagramm für das Merkmal *Alter* aus Tabelle 2.1

## 2.2. Empirische Verteilungsfunktion

**Definition 2.1**
Seien $a_1 < \ldots < a_m$ die Ausprägungen des Merkmals $X$. Die Funktion

$$H(x) = \sum_{a_j \leq x} h(a_j) \tag{2.5}$$

nennt man **empirische Verteilungsfunktion** von $X$. $H(x)$ gibt den Anteil der Werte an, die kleiner oder gleich $x$ sind.

Zur Konstruktion von $H(x)$ kumuliert man zunächst die relativen Häufigkeiten, d. h. für jedes $j = 1, \ldots, m$ bestimmt man

$$H_j = H(a_j) = \sum_{i=1}^{j} h(a_i) \tag{2.6}$$

Diese bilden die Werte von $H(x)$ für $a_j \leq x < a_{j+1},\ j = 1, \ldots, m-1$; für $x < a_1$ ist $H(x) = 0$ und für $x \geq a_m$ ist $H(x) = 1$.

**Beispiel 2.3**
Wir bestimmen für das Merkmal *Alter* aus Tabelle 2.1 die empirische Verteilungsfunktion. Dazu werden zunächst die relativen Häufigkeiten kumuliert (siehe Tabelle 2.2).

| $j$ | Alter $a_j$ | Rel. Häufigkeit $h_j$ | Kumulierte rel. Häufigkeit $H_j = \sum_{i=1}^{j} h(a_i)$ |
|---|---|---|---|
| 1 | 19 | 0,1 | 0,1 |
| 2 | 20 | 0,2 | 0,3 (= 0,1 + 0,2) |
| 3 | 21 | 0,1 | 0,4 (= 0,1 + 0,2 + 0,1) |
| 4 | 25 | 0,3 | 0,7 (= 0,1 + 0,2 + 0,1 + 0,3) |
| 5 | 28 | 0,2 | 0,9 (= 0,1 + 0,2 + 0,1 + 0,3 + 0,2) |
| 6 | 34 | 0,1 | 1,0 (= 0,1 + 0,2 + 0,1 + 0,3 + 0,2 + 0,1) |

Tabelle 2.2.: Relative und kumulierte relative Häufigkeiten

Die empirische Verteilungsfunktion für das Merkmal $X$: *Alter* ist gegeben durch

$$H(x) = \begin{cases} 0 & \text{für} \quad x < 19 \\ 0,1 & \text{für} \quad 19 \leq x < 20 \\ 0,3 & \text{für} \quad 20 \leq x < 21 \\ 0,4 & \text{für} \quad 21 \leq x < 25 \\ 0,7 & \text{für} \quad 25 \leq x < 28 \\ 0,9 & \text{für} \quad 28 \leq x < 34 \\ 1,0 & \text{für} \quad x \geq 34 \end{cases}$$

Lesebeispiele: Der Funktionswert an der Stelle

- $x = 30:\ H(30) = 0,9$ bedeutet, dass 90 % der Befragten jünger oder gleich 30 Jahre alt sind.
- $x = 20:\ H(20) = 0,3$ bedeutet, dass 30 % der Befragten jünger oder gleich 20 Jahre alt sind.
- $x = 21,6:\ H(21,6) = 0,4$ bedeutet, dass 40 % der Befragten jünger oder gleich 21,6 Jahre alt sind.

Grafisch wird $H(x)$ in Abbildung 2.3 dargestellt.

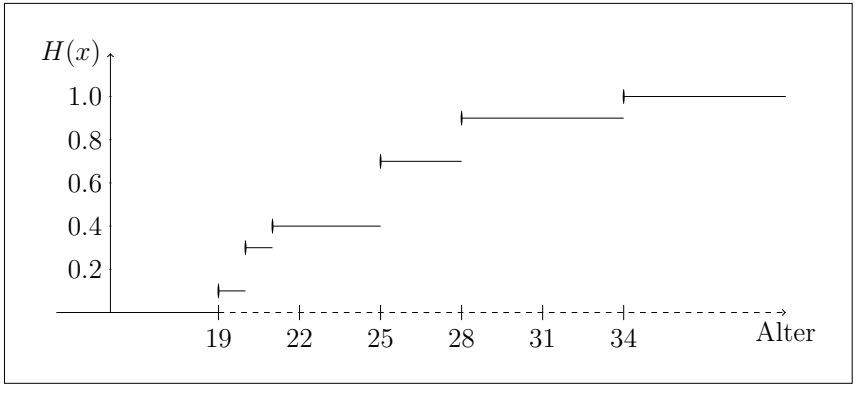



Abbildung 2.3.: Empirische Verteilungsfunktion für das Merkmal *Alter* aus Tabelle 2.1

Eigenschaften der empirischen Verteilungsfunktion:

1. $H(x)$ ist eine monoton steigende Treppenfunktion, d. h. für alle $x_1 \leq x_2$ gilt $H(x_1) \leq H(x_2)$.
2. An den Ausprägungen $a_1 < \ldots < a_m$ springt $H(x)$ um die entsprechende relative Häufigkeit.
3. Der zugehörige Funktionswert an den Sprungstellen ist der obere Wert (d. h. sie ist rechtsseitig stetig).
4. $H(x)$ besitzt die Grenzwerte

$$\lim_{x \to -\infty} H(x) = 0 \qquad \text{und} \qquad \lim_{x \to \infty} H(x) = 1.$$

## 2.3. Klassierte Daten und Histogramm

Häufig ist es sinnvoll, Daten in **Klassen** oder **Gruppen** aufzuteilen, etwa dann, wenn ein Merkmal sehr viele Ausprägungen besitzt. Man spricht dann von klassierten oder gruppierten Daten. Zum Beispiel wurden im Rahmen der Nationalen Verzehrsstudie II (NVS II)[3] die Haushalts-Netto-Einkommen der Teilnehmer in 9 Einkommensklassen aufgeteilt. Von 19.329 Befragten gaben 2137 „Weiß nicht“ an und 1445 machten keine Angabe. Einschließlich dieser Ergebnisse gibt Tabelle 2.3 die absoluten Anzahlen der Beobachtungen in der einzelnen Klassen wieder. Formal werden aus einer Urliste $k \in \mathbb{N}$ Klassen (Gruppen)

$$[b_0, b_1[, \ [b_1, b_2[, \ldots, [b_{k-1}, b_k[$$

gebildet. Jede Klasse $[b_{j-1}, b_j[, \ j = 1, \ldots, k$ ist ein linksabgeschlossenes und rechtsoffenes Intervall, d. h. für alle

$$x \in [b_{j-1}, b_j[$$

gilt

$$b_{j-1} \leq x < b_j.$$

Man nennt $b_{j-1}$ **Klassenuntergrenze** und $b_j$ **Klassenobergrenze**. Die absolute Häufigkeit $f_j$ stellt entsprechend die Anzahl der Beobachtungen dar, die in die Klasse $j$ fallen. (Die Klassenhäufigkeit wird auch **Besetzungszahl** genannt.)

[3] NVS II ist eine Studie des *Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz* zum Ernährungsverhalten der 14- bis 80-jährigen Bevölkerung Deutschlands.