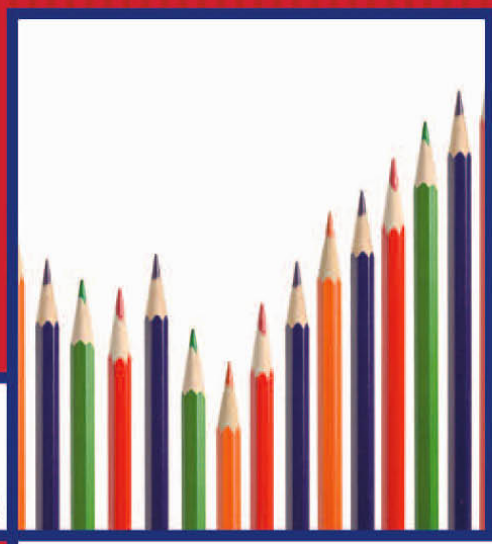
WWW.PEARSON-STUDIUM.DE
CWS
COMPANION WEBSITE

wi
wirtschaft



Andreas Quatember

Statistik ohne Angst vor Formeln

Das Studienbuch für Wirtschafts- und Sozialwissenschaftler

3., aktualisierte Auflage

PEARSON
Studium

Statistik ohne Angst vor Formeln

Unser Online-Tipp
für noch mehr Wissen …

informit.de

Aktuelles Fachwissen rund um die Uhr
– zum Probelesen, Downloaden oder
auch auf Papier.

www.informit.de

Andreas Quatember

Statistik ohne Angst vor Formeln

Das Studienbuch für Wirtschafts- und Sozialwissenschaftler

3., aktualisierte Auflage

PEARSON Studium

ein Imprint von Pearson Education
München • Boston • San Francisco • Harlow, England
Don Mills, Ontario • Sydney • Mexico City
Madrid • Amsterdam

Bibliografische Information Der Deutschen Nationalbibliothek

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

Die Informationen in diesem Buch werden ohne Rücksicht auf einen eventuellen Patentschutz veröffentlicht.
Warennamen werden ohne Gewährleistung der freien Verwendbarkeit benutzt.
Bei der Zusammenstellung von Texten und Abbildungen wurde mit größter Sorgfalt vorgegangen. Trotzdem können Fehler nicht ausgeschlossen werden.
Verlag, Herausgeber und Autoren können für fehlerhafte Angaben und deren Folgen weder eine juristische Verantwortung noch irgendeine Haftung übernehmen.
Für Verbesserungsvorschläge und Hinweise auf Fehler sind Verlag und Autor dankbar.

Alle Rechte vorbehalten, auch die der fotomechanischen Wiedergabe und der Speicherung in elektronischen Medien.
Die gewerbliche Nutzung der in diesem Produkt gezeigten Modelle und Arbeiten ist nicht zulässig.

Fast alle Produktbezeichnungen und weitere Stichworte und sonstige Angaben, die in diesem Buch verwendet werden, sind als eingetragene Marken geschützt.
Da es nicht möglich ist, in allen Fällen zeitnah zu ermitteln, ob ein Markenschutz besteht, wird das ® Symbol in diesem Buch nicht verwendet.

10 9 8 7 6 5 4 3 2 1

13 12 11

ISBN 978-3-86894-055-8

© 2011 Pearson Studium
ein Imprint der Pearson Education Deutschland GmbH,
Martin-Kollar-Straße 10-12, D-81829 München/Germany
Alle Rechte vorbehalten
www.pearson-studium.de
Lektorat: Martin Milbradt, mmilbradt@pearson.de;
Alice Kachnij, akachnij@pearson.de
Korrektorat: Barbara Decker, München
Einbandgestaltung: Thomas Arlt, tarlt@adesso21.net
Herstellung: Elisabeth Prümm, epruemm@pearson.de
Satz: mediaService, Siegen (www.media-service.tv)
Druck und Verarbeitung: Kösel, Krugzell (www.KoeselBuch.de)

Printed in Germany

Inhaltsverzeichnis

Vorwort zur 3. Auflage

Wenn Sie, werte(r) Leser(in), einmal irgendein Wirtschafts- oder Nachrichtenmagazin oder eine beliebige Tageszeitung bewusst nach jenen Artikeln durchsehen, in denen Statistiken vorkommen, dann werden Sie einen Eindruck davon erhalten, welch bedeutende Rolle der Statistik bei der Informationsvermittlung in unserer Wissensgesellschaft zukommt. Ebenso schnell werden die meisten von Ihnen zustimmen, dass das Image des Faches Statistik in völligem Kontrast zu dieser offenkundigen Bedeutung ein denkbar schlechtes ist. Dies verwundert nicht allzu sehr! Findet man doch unter den bewussten Zeitungsartikeln zu oft völlig unsinnige oder falsch interpretierte oder sogar bewusst manipulierte Statistiken (siehe dazu: www.ifas.jku.at → Unsinn in den Medien). Dadurch erhält der aufmerksame Beobachter den Eindruck, dass man mit Statistik tatsächlich alles beweisen könne.

Das schlechte Image basiert demnach zu einem Gutteil auf dem fundamentalen Irrtum, die Qualität der *Anwendungen* der statistischen Methoden mit der Qualität der *Methoden* zu verwechseln. Die Korrektheit der Methoden ist jedoch im besten naturwissenschaftlichen Sinne beweisbar. Sofern nämlich die für die jeweiligen Fragestellungen geeigneten Verfahren ausgewählt, die Berechnungen korrekt durchgeführt und die Ergebnisse richtig interpretiert werden, ist die statistische Analyse von Daten ein wichtiges Instrument zur Gewinnung objektiver Informationen über einen interessierenden Sachverhalt. Dazu genügt es aber nicht, lediglich den richtigen Knopf eines Statistikprogrammpaketes drücken zu können. Vielmehr ergibt sich aus der leichten Durchführbarkeit der Berechnungen die geradezu zwingende Notwendigkeit, sich ein grundlegendes Methodenverständnis anzueignen. Im vorliegenden Buch soll das Verständnis vor allem durch die – durch anschauliche Beispiele unterstützte – Beschreibung der Ideen, die hinter den Methoden stehen, gefördert werden. Die Formeln verstehen sich hierbei nicht als Basis der Ideenbeschreibung, sondern als deren formaler Abguss.

Als „Statistik-Programmpaket“ begleitet dieses Buch das Statistik-Modul im Tabellenkalkulationsprogramm Excel. Der Grund dafür ist einfach, dass es auf beinahe jedem Rechner zur Verfügung steht. Hinweise darauf sind im erklärenden Text bewusst sparsam gesetzt. An jeden Abschnitt schließen sich Übungsaufgaben zum selbstständigen Lösen an. Ein Teil dieser Übungsaufgaben ist zur Förderung des Methodenverständnisses nicht in Excel, sondern mit Hilfe eines Taschenrechners zu lösen. Für jene Übungsaufgaben aber, die durch das CWS-Logo der Companion-Website gekennzeichnet sind (siehe nebenan), steht eine Excel-Lerndatei zur Verfügung. In dieser wird dem Anwender beziehungsweise der Anwenderin das Lösen der betreffenden Übungsaufgaben mit Excel Schritt für Schritt erklärt. Diese Lerndatei findet sich genauso wie die Beschreibungen des Rechengangs der grundlegenden und der Lösungen aller Übungsbeispiele auf der Companion-Website von Pearson Studium (siehe Umschlag).

Für die vorliegende 3. Auflage wurden Daten aktualisiert, Übungsbeispiele neu verfasst und auch zwei komplett neue Abschnitte in den Text aufgenommen. Ferner wurden ungenaue Passagen weiter präzisiert. Das Buch lernt, passt sich an und verändert sich. Kurz: Es lebt!

Ich möchte mich bei allen beteiligten Mitarbeiterinnen und Mitarbeitern von Pearson Studium, insbesondere bei meinem direkten Ansprechpartner beim Verlag, Herrn Martin Milbradt, für die erfreuliche und gedeihliche Zusammenarbeit bedanken. Ebenso danke ich jenen Menschen in meinem beruflichen wie auch privaten Umfeld, die dafür sorgen, dass meine Begeisterung fürs Leben täglich anhält.

Viele interessante Stunden beim Lesen des Buches wünscht Ihnen

Andreas Quatember

Internet: *www.ifas.jku.at*

E-Mail: *andreas.quatember@jku.at*

Beschreibende Statistik

1

ÜBERBLICK

1.1 Grundbegriffe

Was ist eigentlich Statistik? Im Begriff **Statistik** (*lat. status = Stand, Umstände*) werden alle Methoden der Analyse von Daten mit dem Ziel einer Informationsbündelung subsumiert. Anstelle der Betrachtung aller vorliegenden Daten (zum Beispiel der Punktezahlen aller Prüflinge bei der letzten Statistikklausur auf einer im Internet veröffentlichten Liste) werden diese Daten tabelliert, grafisch aufbereitet und durch Kennzahlen (wie etwa dem Mittelwert der Punktezahlen der Prüflinge) charakterisiert. In jedem dieser Schritte gehen dabei zwar Informationen verloren (zum Beispiel, wie viele Punkte ein ganz bestimmter Prüfling erhalten hat) und dennoch gewinnt man gerade mit jedem dieser Schritte weg vom Detail an Über- und Einblick. Und genau das ist der Zweck der Informationsbündelung und somit der Statistik.

Die Statistik wird nach ihren Aufgabenbereichen in **beschreibende** (oder **deskriptive**) und **schließende** (oder **inferenzielle**) **Statistik** gegliedert. Die Erstere beschränkt sich auf die Beschreibung einer Grundgesamtheit von Erhebungseinheiten (zum Beispiel der Prüflinge) durch die Analyse der in einer **Vollerhebung** gewonnenen Daten, während sich die Letztere mit den Rückschlüssen von lediglich in Stichproben aus solchen Grundgesamtheiten (etwa aus der wahlberechtigten Bevölkerung) gewonnenen Daten auf solche Grundgesamtheiten beschäftigt. Für diese Rückschlüsse wird die **Wahrscheinlichkeitstheorie** benötigt, weshalb deren Grundlagen ebenfalls ein unverzichtbarer Bestandteil der statistischen Grundausbildung sind.

Die Statistik beschäftigt sich also mit der Analyse von Daten und den dazu benötigten Methoden. In der **beschreibenden Statistik**, mit der wir uns in diesem ersten Kapitel beschäftigen werden, liegen uns also vollständige Daten über jene Gesamtheit vor, die wir mit Hilfe statistischer Methoden hinsichtlich bestimmter interessierender Eigenschaften beschreiben wollen.

Wie jedes Fach bedient sich auch die Statistik eigener Begriffe, deren Kenntnis die Kommunikation zwischen dem Anwender der statistischen Methoden und den Statistik-Experten wesentlich erleichtert: So nennt man etwa jene Objekte, über die wir Daten erhalten (zum Beispiel die einzelnen Prüflinge oder in einem anderen Fall einzelne Schrauben), die **Erhebungseinheiten** der Erhebung. Die Gesamtheit aller potenziellen Erhebungseinheiten bildet die **Grundgesamtheit** der Erhebung (also etwa alle Prüflinge eines Klausurtermins oder die Schrauben einer Packung).

Die interessierende Eigenschaft, die an den Erhebungseinheiten beobachtet werden soll, über die also Daten gewonnen werden sollen, ist das interessierende **Merkmal** der Erhebung (etwa die Punktezahl bei der Klausur oder die Schraubenlänge). Bei vielen Erhebungen interessiert man sich nicht nur für ein Merkmal, sondern für eine Vielzahl von Merkmalen der Erhebungseinheiten. Jedes dieser Merkmale besitzt verschiedene mögliche Werte (zum Beispiel eine Punktezahl von 0 oder 1 oder 2 und so weiter beziehungsweise eine Länge von 6,04 cm, von 5,99 cm und so weiter). Diese möglichen Werte sind die so genannten **Merkmalsausprägungen** der Merkmale. Alle Merkmalsausprägungen zusammen bilden den **Wertebereich** des Merkmals.

Beispiel 1: Grundbegriffe einer statistischen Erhebung

Betrachten wir folgende Erhebungen:

A: Erhebung der Punkteverteilung bei der Statistikklausur am Ende des vergangenen Semesters.

B: Erhebung der Zufriedenheit der Kunden eines Sportartikelhändlers mit der Beratung.

C: Erhebung des besten Kinofilms des vergangenen Jahres aus einer Auswahl von zehn Filmen unter den teilnahmebereiten Lesern einer Kinozeitschrift.

Tabelle 1.1

Begriffe\Erhebung	A	B	C
Grundgesamtheit	alle Prüflinge	alle Kunden	teilnahmebereite Leser
Merkmal	Punkte	Zufriedenheit mit der Beratung	bester Film
Merkmals-ausprägungen	0, 1, 2, ..., 7 Punkte	sehr zufrieden, eher zufrieden, teils-teils, eher unzufrieden, sehr unzufrieden	die 10 angegebenen Filme der Auswahl

Diese interessierenden Merkmale werden nach der Art ihrer Ausprägungen unterschieden. Dies ist deshalb notwendig, weil sich für die verschiedenen Merkmalstypen – wie sich später zeigen wird – nicht die gleichen Methoden zur Datenanalyse eignen. So ist intuitiv einleuchtend, dass kein Mittelwert eines Merkmals wie Geschlecht berechnet werden kann, sondern nur von Merkmalen, deren Ausprägungen Zahlen sind (also etwa der Punktezahlen bei einer Statistikklausur).

Wir unterscheiden insofern nominale, ordinale und metrische Merkmale. Die Zuordnung erfolgt durch die Betrachtung der Merkmalsausprägungen eines Merkmals. Unter einem **nominalen** (oder auch qualitativen) Merkmal versteht man ein Merkmal, dessen Ausprägungen sich nicht zwingend ordnen lassen und sich nur durch ihren Namen (*lat. nomen = Name*) unterscheiden. Dazu gehören etwa die Merkmale Geschlecht (die beiden Merkmalsausprägungen weiblich und männlich lassen sich vielleicht von Männern höflichkeitshalber in der Reihenfolge weiblich/männlich oder alphabetisch, aber jedenfalls nicht zwingend ordnen), Parteipräferenz (die Reihenfolge der Parteien auf dem Stimmzettel ergibt sich in Österreich aus dem Stimmenanteil bei der letzten Wahl, ist aber nicht zwingend so vorzunehmen) oder auch Staatsbürgerschaft.

Ein Merkmal heißt hingegen **ordinal** (oder auch Rangmerkmal), wenn seine Ausprägungen in einer Ordnungsrelation zueinander stehen (*lat. ordinare = ordnen*). Das heißt, dass die Ausprägungen eine natürliche Reihenfolge besitzen. Dazu gehören etwa Schulnoten (1 ist besser als 2, 2 besser als 3), Platzierungen in irgendwelchen Wettbewerben

(1. vor 2., 2. vor 3.) oder Zustimmungsgrad zu einer Frage (1 = volle Zustimmung, ... 10 = volle Ablehnung).

Schließlich nennt man ein Merkmal **metrisch** (oder auch quantitativ), wenn seine Merkmalsausprägungen nicht nur, wie dies bei ordinalen Merkmalen der Fall ist, der Größe nach geordnet werden können, sondern auch noch Vielfaches einer Einheit sind (*gr. metron = Maß*). Das ist beispielsweise der Fall bei den Merkmalen Körpergröße (180 cm ist nicht nur größer als 170, sondern auch das Vielfache der Einheit Zentimeter, was dieses Merkmal eben von einem ordinalen Merkmal unterscheidet), Schraubenlänge (4,019 oder 4,222 cm) oder Punktezahl bei einer Statistikklausur (0, 1, ... 7) (siehe dazu: Beispiel 2).

Eine andere Frage betrifft die **Kodierung** (= Verschlüsselung) der Merkmalsausprägungen eines Merkmals zum Zwecke der Datenverarbeitung (etwa in Excel). Hierbei werden den auftretenden Merkmalsausprägungen der Einfachheit halber (ganzzahlige) Zahlenwerte zugewiesen, selbst wenn die Ausprägungen selbst nicht Zahlen sind. Wenn man einem nominalen Merkmal Zahlenwerte zuordnet (zum Beispiel: 1 = weiblich, 2 = männlich), so spricht man naheliegenderweise von einer künstlichen **Metrisierung** eines nominalen Merkmals. Dabei darf man jedoch niemals vergessen, dass man eigentlich ein Merkmal eines anderen Typs vor sich hat, als es die kodierten Ausprägungen anzuzeigen scheinen.

Dazu betrachten wir folgenden Ausschnitt eines Fragebogens über die Zufriedenheit von Studierenden mit einer Lehrveranstaltung:

Geben Sie bitte Ihr Geschlecht an:

O *weiblich (=1)* O *männlich (=2)*

Wie alt sind Sie (in vollendeten Lebensjahren)?

..... Jahre

Wie beurteilen Sie das persönliche Engagement, mit dem der Kursleiter den Kurs bestreitet?

O *sehr gut (=1)* O *gut (=2)*

O *mittelmäßig (=3)* O *mangelhaft (=4)*

O *schlecht (=5)*

Wurden Sie vom Kursleiter zu Fragen ermuntert?

O *oft (=1)* O *manchmal (=2)*

O *selten (=3)* O *nie (=4)*

In einer **Excel-Tabelle** etwa sind die Ergebnisse einer solchen Befragung für die statistische Auswertung so zu organisieren, dass in jede Zeile der Tabelle eine Erhebungseinheit und in jede Spalte dieser Zeile die Antwort dieser Erhebungseinheit auf eine bestimmte Frage eingegeben werden:

2	21	1	3
1	38	2	2
...	...	...	...

Die erste Befragungsperson hatte also auf ihrem Fragebogen angegeben, dass sie männlichen Geschlechts und 21 Jahre alt ist, das persönliche Engagement des Lehrenden mit sehr gut beurteilt und von diesem selten zu Fragen ermuntert wurde. Die zweite Befragungsperson war weiblich, 38 Jahre alt, beurteilte das Engagement mit gut und fühlte sich manchmal zu Fragen ermuntert. Nach Eingabe aller Fragebögen einer solchen Befragung liegt zur methodischen Bearbeitung der Daten eine Liste (= Datenmatrix) vor, deren Zeilenanzahl der Anzahl der befragten Erhebungspersonen und deren Spaltenanzahl der Anzahl der erhobenen Merkmale entspricht.

Ein zweites, im Hinblick auf die Anwendung geeigneter statistischer Methoden notwendiges Einteilungsprinzip von Merkmalen wird folgendermaßen eingeführt: Wenn ein Merkmal nur bestimmte mögliche Merkmalsausprägungen besitzt, dann handelt es sich bei einem solchen Merkmal um ein so genanntes **diskretes** Merkmal (zum Beispiel Schulnoten, Fehlerzahlen, Geschlecht), während Merkmale, deren Ausprägungen alle reellen Werte eines Intervalls annehmen können, als **stetig** bezeichnet werden. Beispiele dafür sind die Länge von Schrauben, die Körpergröße und ebenso das Alter, denn wir werden leider jeden Augenblick älter, auch wenn wir unser Alter meist in vollendeten (= ganzen) Lebensjahren angeben, wodurch man den Eindruck gewinnt, dass man erst am jeweiligen Geburtstag um ein ganzes Jahr älter wird (oder bei einem so genannten runden Geburtstag sogar um zehn Jahre auf einmal!). Bei einer solchen Vorgehensweise spricht man von einer (künstlichen) **Diskretisierung** eines stetigen Merkmals (andere Beispiele sind Körpergröße in ganzen cm oder Gewicht in ganzen kg). Das Merkmal selbst bleibt aber stetig.

Beispiel 2: Merkmalstypen

Betrachten wir folgende Merkmale von statistischen Untersuchungen und ihre Einteilung nach den beiden genannten Einteilungsprinzipien:

Tabelle 1.2

Merkmal	Merkmals-ausprägungen	nominal/ordinal/ metrisch	diskret/stetig
Familienstand von Befragten	ledig (=1), verheiratet (=2), geschieden (=3), verwitwet (=4), verpartnert (=5), ...	nominal	diskret
Zeiten der Teilnehmer an einem 100-m-Lauf	11,21 sec., 11,24 sec., ...	metrisch	stetig
Preis eines Sportartikels	29,90 €, 34,90 €, ...	metrisch	diskret
Platzierungen in einem 100-m-Lauf	1., 2., 3., ...	ordinal	diskret
Marke verkaufter LCD-Fernsehgeräte	SONY, Philips, ...	nominal	diskret
Einwohnerzahlen verschiedener Bundesländer	2,362.929, 4,746.014, ...	metrisch	diskret
Weitsprungleistung von Schülern (in ganzen cm)	516 cm, 392 cm, ...	metrisch	stetig
Beurteilung der Qualität einer TV-Show durch ausgewählte Konsumenten	1 = sehr gut, 2 = gut, 3 = teils-teils, 4 = schlecht, 5 = sehr schlecht	ordinal	diskret
Gewicht von TV-Geräten im Lager eines Unternehmens	20,426 kg, 22,822 kg, ...	metrisch	stetig

Um einen besseren Überblick über die erhobenen Daten als durch deren Einzelbetrachtung (zum Beispiel der Punktezahlen der einzelnen Prüflinge) zu gewinnen, bildet man **Häufigkeitsverteilungen**, das heißt, wir geben zu jeder Merkmalsausprägung an, wie häufig sie aufgetreten ist. Die Ergebnisse können, mit einem erklärenden Text versehen, in Tabellenform ausgewiesen oder grafisch dargestellt und durch verschiedene Kennzahlen charakterisiert werden.

Übungsaufgaben

Die Lösungen zu den aufgeführten Aufgaben finden Sie unter www.pearson-studium.de.

Ü1

Welchen Merkmalstypen (nach den Einteilungskriterien „metrisch – ordinal – nominal" beziehungsweise „diskret – stetig") gehören die folgenden Merkmale an?

a) Länge von Videobändern einer Produktion in cm
b) Reiseziel von befragten Urlaubsbuchern
c) Güteklassen von Obst am Markt
d) Inflationsrate verschiedener Länder
e) Religion von Befragten
f) Heizungsart von Mietwohnungen in einer Stadt
g) Anzahl an Kinobesuchen von Schülern einer Schule in den Ferien
h) Einstellung von Befragten zur Einführung eines Berufsheeres (Antwortalternativen: sehr skeptisch, eher skeptisch, unentschieden, eher positiv, sehr positiv)

Ü2

Welchen Merkmalstypen gehören die folgenden Merkmale an und wie sind eventuelle Kodierungen der Merkmalsausprägungen vorzunehmen?

a) Fußballinteresse von Befragten (Merkmalsausprägungen: sehr groß, groß, mittel, schwach, gar keines)
b) Einkommen von Erwerbstätigen in ganzen EURO
c) Einstellung der Bevölkerung zu einem EU-Beitritt der Türkei (Merkmalsausprägungen: dafür, teils-teils, dagegen)
d) Temperatur um 12 Uhr
e) Anzahl an in einem Monat in einem Konzern produzierten Autos
f) Militärische Dienstgrade in der deutschen Bundeswehr
g) Lieblingsschauspieler/in aus einer Liste von 20 vorgeschlagenen Personen
h) Seehöhe

1.2 Tabellarische und grafische Darstellung von Häufigkeitsverteilungen

1.2.1 Häufigkeitsverteilungen einzelner Merkmale

Tabellarische Darstellung

Sehr häufig werden einzelne Merkmale (oder eindimensionale Merkmale) betrachtet. Nehmen wir als Beispiel die von den Prüflingen erreichten Punktezahlen bei einer

Statistikklausur. Bei 142 angetretenen Prüflingen kann man sich durch Betrachten der Ergebnisse jedes einzelnen Prüflings in einer Liste nur einen sehr groben Überblick über die Klausurergebnisse beschaffen. Viel besser geeignet ist für diesen Zweck eine **Tabelle** der Häufigkeitsverteilung. Eine solche könnte beispielsweise folgendermaßen aussehen:

Beispiel 3: Tabellarische Darstellung einer Häufigkeitsverteilung für ein diskretes Merkmal

Die Punktezahlen von 142 Studierenden bei einer Statistikklausur:

Tabelle 1.3

Punkte-zahlen	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
0	1	0,007	0,7	0,007
1	3	0,021	2,1	0,028
2	10	0,070	7,0	0,098
3	16	0,113	11,3	0,211
4	32	0,225	22,5	0,436
5	44	0,310	31,0	0,746
6	20	0,141	14,1	0,887
7	16	0,113	11,3	1

Früher erhielt man die so genannten **Häufigkeiten** zu jeder Merkmalsausprägung nur durch deren Abzählen in Form einer „Strichliste“. Dabei wurde in eine Liste mit den möglichen Merkmalsausprägungen ein Strich zu einer bestimmten Merkmalsausprägung hinzugefügt, wenn diese auftrat. Am Schluss zählte man die Striche zu jeder Merkmalsausprägung und erhielt so deren Häufigkeit. Heute erledigt diesen Vorgang der Computer per Knopfdruck nach Eingabe aller Punktezahlen in eine Datei. Zum Beispiel kann in Excel eine Spalte mit Häufigkeiten wie in Beispiel 3 durch die korrekte Handhabung der Funktion HÄUFIGKEIT erstellt werden. Doch auch das dabei im Hintergrund ablaufende Computerprogramm macht nichts „Wissenschaftlicheres“ als zu zählen, wie oft die einzelnen Ausprägungen aufgetreten sind. Die Summe der Häufigkeiten der einzelnen Merkmalsausprägungen ergibt jedenfalls die Gesamtzahl der Erhebungseinheiten der Grundgesamtheit.

Wir haben durch die Häufigkeiten in der Tabelle schon einen ersten Überblick über die Klausurergebnisse gewonnen. In Worten lässt sich dieser Überblick folgendermaßen präsentieren: Von den 142 Prüflingen haben 16 die maximale Punktezahl von sieben Punkten bei der Klausur erhalten, 20 haben sechs Punkte erreicht und so weiter. Wie man sieht, muss für eine richtige Einschätzung dieser Häufigkeiten die Gesamt-

zahl der Prüflinge mit angegeben werden. „16 von 142“ bedeutet doch wohl etwas völlig anderes als etwa „16 von 20“.

Um sich diese gewünschte Relation der Häufigkeiten der einzelnen Merkmalsausprägungen zur Gesamtzahl der Erhebungseinheiten zu vergegenwärtigen, werden neben den Häufigkeiten (oder an deren Stelle) gerne die **relativen Häufigkeiten** (oder Anteile) der einzelnen Merkmalsausprägungen angegeben. Diese erhält man, indem man die Häufigkeit jeder Merkmalsausprägung durch die Anzahl aller Erhebungseinheiten dividiert. In Beispiel 3 also beträgt etwa die relative Häufigkeit der Punktezahl 7 (auf drei Stellen gerundet)

$$\frac{16}{142} = 0{,}113,$$

diejenige der Punktezahl 6

$$\frac{20}{142} = 0{,}141$$

und so fort.

Bezeichnet man die Häufigkeit der ersten Merkmalsausprägung mit h_1, der zweiten mit h_2 und so weiter, die relativen Häufigkeiten genauso mit p (*lat. pro portione = im Verhältnis*) und die Gesamtzahl der Erhebungseinheiten mit N, so ergibt sich die relative Häufigkeit irgendeiner i-ten Merkmalsausprägung also durch

$$p_i = \frac{h_i}{N}. \tag{1}$$

Der Ausdruck, den wir mit der Formelnummer (1) kennzeichnen, ist die formale Übersetzung der vorher beschriebenen Überlegung und gibt an, wie man ganz allgemein aus den Häufigkeiten und der Gesamtzahl der Erhebungseinheiten die relativen Häufigkeiten für alle Merkmalsausprägungen berechnet.

Diese Zahlen sind also die in Relation zur Anzahl aller Erhebungseinheiten gesetzten Häufigkeiten (eben: *relative* Häufigkeiten). Multipliziert man die relativen Häufigkeiten noch mit 100, dann erhalten wir die **Prozentzahlen** zu jeder Merkmalsausprägung (*lat. pro centum = im Verhältnis zu hundert*). Die häufig verwendeten Prozentzahlen sind der Grund, warum es meist nützlich ist, die relative Häufigkeit mit drei Stellen nach dem Komma anzugeben. Die Prozentzahlen erhalten auf diese Weise nämlich auch noch eine Stelle nach dem Komma, was zumeist noch als Informationsgewinn im Vergleich zur Rundung auf ganze Prozent empfunden wird. Jede weitere Nachkommastelle bringt jedoch einen immer kleiner werdenden verwertbaren Informationsnutzen. Ein die Prozentzahlen in der Tabelle zu Beispiel 3 erklärender Text könnte also lauten: 11,3 Prozent der Prüflinge erreichten die Maximalzahl von 7 Punkten, 14,1 Prozent erreichten 6 Punkte und so fort.

Der Angabe von Prozentzahlen liegt die berechtigte Annahme zu Grunde, dass der Bezug auf eine fiktive Grundgesamtheit von 100 Erhebungseinheiten die Anschaulich-

keit der Ergebnisse erhöht. Denn beispielweise kann man sich die wahre Größenordnung an Prüflingen, die die maximale Punktezahl erhalten haben, durch die Angabe, dass dies 16 von 142 Personen waren, weniger gut „vor Augen führen“ als durch die Vorstellung, dass dies 11,3 von 100 Personen waren. 16 verhält sich zu 142 wie 11,3 zu 100. Absolut nicht notwendig ist es demnach, eine Relation in Prozent anzugeben, deren Größenordnung keinerlei Anschaulichkeitsproblem bietet. Wenn etwa sechs von zwölf Personen eine bestimmte Eigenschaft aufweisen, dann führt die Auskunft, dass dies 50 Prozent tun, wegen der Angabe in Prozent eher zur irrigen Auffassung, dass eine große Grundgesamtheit vorliegen müsse, als dass dadurch die Anschaulichkeit gefördert wird. Sechs von zwölf Personen sind allemal anschaulich genug.

Die Unterschiede von Prozentzahlen werden häufig in **Prozentpunkten** angegeben. Die Merkmalsausprägung 5 kommt in Beispiel 3 natürlich nicht um 8,5 Prozent häufiger vor als die Merkmalsausprägung 4, sondern um 44 : 32 = 37,5 Prozent. Behandelt man die Prozentzahlen jedoch wie Punktezahlen, dann ist die Merkmalsausprägung 5 um 8,5 Prozent*punkte* häufiger als die Merkmalsausprägung 4 vorgekommen. Vor allem beim Vergleich von aktuellen Wahlergebnissen mit denen der letzten Wahl wird dieser Terminus häufig gewählt.

Eine bei metrischen und ordinalen Merkmalen oftmals sehr brauchbare Zusatzinformation lässt sich gewinnen, wenn man die relativen Häufigkeiten verschiedener Merkmalsausprägungen addiert. Insbesondere kann es nützlich sein, wenn man für jede Merkmalsausprägung die relativen Häufigkeiten genau dieser Merkmalsausprägung und aller kleineren addiert – also für die Punktezahl 3 in Beispiel 3 etwa die Summe der relativen Häufigkeiten der Ausprägungen 0, 1, 2 und 3 Punkte. Diese Summe

$$0{,}007 + 0{,}021 + 0{,}070 + 0{,}113 = 0{,}211$$

ist die **relative Summenhäufigkeit** (oder empirische Verteilungsfunktion) der Merkmalsausprägung 3. Sie gibt an, wie groß der Anteil an Erhebungseinheiten ist, die eine Merkmalsausprägung von höchstens 3 aufweisen. Die relative Summenhäufigkeit der Merkmalsausprägung 5 ist demnach die Summe der relativen Häufigkeiten der Merkmalsausprägungen 5 und aller kleineren, also von 0, 1, 2, 3, 4 und 5. Das ist:

$$0{,}007 + 0{,}021 + 0{,}070 + 0{,}113 + 0{,}225 + 0{,}310 = 0{,}746.$$

Der Nutzen liegt auf der Hand: Wenn man zum Beispiel 4 Punkte benötigt, um positiv zu sein, dann gibt die relative Summenhäufigkeit der Punktezahl 3 den Anteil derer an, die 3 oder weniger Punkte erreicht haben. Das sind also diejenigen, welche die Klausur nicht geschafft haben. In Prozenten sind dies 21,1 Prozent. Und auch die umgekehrte Fragestellung lässt sich mittels relativer Summenhäufigkeiten natürlich sofort beantworten. Da die Summe aller relativen Häufigkeiten 1 ergeben muss, ist der Anteil derer, die mindestens 4 Punkte aufweisen, gleich

$$1 - 0{,}211 = 0{,}789.$$

Das wiederum sagt aus, dass 78,9 Prozent der Prüflinge die Prüfung erfolgreich absolviert haben.

Auch Anteile für Intervalle wie „2 bis 6 Punkte“ lassen sich natürlich als Differenz zweier Summenhäufigkeiten bestimmen. Dieser etwa ist konkret die Differenz der relativen Summenhäufigkeiten der Merkmalsausprägungen 6 und 1, also die Differenz aus dem Anteil dafür, höchstens 6 und dem Anteil dafür, höchstens 1 Punkt erhalten zu haben:

$$0{,}887 - 0{,}028 = 0{,}859.$$

Der Wert sagt aus, dass 85,9 Prozent der Prüflinge 2 bis 6 Punkte erreicht haben. Dieses Ergebnis erhalten wir natürlich auch, wenn wir die relativen Häufigkeiten der Ausprägungen 2, 3, 4, 5 und 6 addieren.

Wenn man – um einmal ein anderes Merkmal zu betrachten – das Alter von Personen misst, dann gibt die relative Summenhäufigkeit einer bestimmten Ausprägung an, wie groß der Anteil jener in der Erhebung ist, die höchstens dieses Alter aufweisen. Und die Differenz dieser relativen Summenhäufigkeit zu ihrem Maximum 1 gibt den Anteil derer an, die älter sind. Auch Anteile für Intervalle wie „30 bis 60 Jahre“ lassen sich natürlich als Differenz zweier Summenhäufigkeiten leicht bestimmen.

Bei stetigen Merkmalen wie dem genauen Alter oder diskreten bzw. diskretisierten Merkmalen mit sehr vielen unterschiedlichen Merkmalsausprägungen wie dem Alter in vollendeten Lebensjahren ist es im Übrigen zielführend, den Wertebereich zum Zweck des besseren Überblicks in **Intervalle** zu zerlegen und dann die Häufigkeitsverteilung dieses in Intervalle zerlegten Merkmals tabellarisch darzustellen. Würde man das Alter einer Gesamtheit wirklich möglichst genau erfassen (zum Beispiel durch Angabe des genauen Geburtstages und sogar der Geburtsstunde), so entstünden nämlich unzählige Merkmalsausprägungen, die jeweils geringe Häufigkeiten besäßen. Dadurch ließe sich auch durch die tabellarische beziehungsweise grafische Darstellung der Häufigkeitsverteilung kein besserer Überblick über die Daten gewinnen. Durch eine Einteilung des Wertebereichs des betrachteten Merkmals in eine geringe Zahl von Intervallen geht zwar eine Menge (unnötig genauer) Information verloren, aber man gewinnt stattdessen an Überblick über die erhobenen Daten.

Die relativen Häufigkeiten beziehen sich bei einem solchermaßen erfassten Merkmal jeweils auf die Intervalle, und die zu einem solchen Intervall gehörenden relativen Summenhäufigkeiten geben den Anteil der betrachteten Grundgesamtheit an, der in dieses Intervall oder in eines mit kleineren Merkmalsausprägungen fällt. Das heißt, dass die relative Summenhäufigkeit eines Intervalls den Anteil der Erhebungseinheiten angibt, die höchstens eine Merkmalsausprägung aufweisen, die der Obergrenze des betreffenden Intervalls entspricht.

Beispiel 4: Tabellarische Darstellung einer Häufigkeitsverteilung eines in Intervalle zerlegten Merkmals

Im Jahr 2007 veröffentlichte die Statistik Austria über die österreichische Wohnbevölkerung des Jahres 2005 folgende Häufigkeitsverteilung des Merkmals Alter (*www.statistik.at/web_de/services/stat_jahrbuch/index.html*):

Tabelle 1.4

Altersklasse	Häufigkeit	Relative Häufigkeit	Prozent	Relative Summenh.
0 – unter 15	1,317.707	0,160	16,0	0,160
15 – unter 30	1,526.909	0,185	18,5	0,345
30 – unter 45	1,984.501	0,241	24,1	0,586
45 – unter 60	1,596.849	0,194	19,4	0,780
60 – unter 75	1,173.166	0,142	14,2	0,922
75 und mehr	634.174	0,077	7,7	1

Die Gesamtheit, die hier hinsichtlich des Merkmals Alter charakterisiert werden soll, besteht aus $N = 8{,}233.306$ Personen. Wir sehen, dass etwa die zum Intervall „15 – unter 30“ gehörende relative Häufigkeit den Wert 0,185 aufweist, was bedeutet, dass 18,5 Prozent der österreichischen Bevölkerung im Jahr 2005 zwischen 15 und (unter) 30 Jahre alt waren. Dabei darf darauf hingewiesen werden, dass wir es mit einem stetigen Merkmal zu tun haben, so dass man im Allgemeinen zu einem bestimmten Zeitpunkt genau angeben kann, ob man unter oder über 30 Jahre alt ist.

Die relative Summenhäufigkeit der Altersklasse „15 – unter 30“ ist die Summe der relativen Häufigkeiten der betrachteten Altersklasse und aller jüngeren. Die Summe 0,345 bedeutet demnach, dass 34,5 Prozent der Bevölkerung zwischen 0 und (unter) 30 Jahre alt waren. Wollen wir berechnen, wie groß der Anteil derer war, die zwischen 15 und (unter) 60 Jahre alt waren, so können wir aus Tabelle 1.4 ablesen, dass dies die Differenz

$$0{,}780 - 0{,}160 = 0{,}620$$

sein muss. Das ist nämlich die Differenz der relativen Summenhäufigkeiten bei den Ausprägungen 60 und 15.

Nicht stören dürfte dabei, dass die Summe der relativen Häufigkeiten nicht 1, sondern 0,999 ergibt. Ein solcher **Rundungsfehler** kann durch die Rundung auf drei Nachkommastellen entstehen. Dennoch besitzt die relative Summenhäufigkeit an der letzten Obergrenze natürlich den Wert 1. Wie groß diese Obergrenze jedoch ist, geht aus der Tabelle nicht hervor.