Teil I.

Deskriptive Statistik

Kapitel 1.

Einführung

„*Ein Gast sitzt im Kaffeehaus und trinkt Kaffee.*“ Dieser Satz aus einer berühmten Abhandlung[1] des Schriftstellers und Publizisten F. Torberg (1908–1979) wird von seinem Autor zum Anlass genommen, eine Reihe von Fragen zu erörtern, die die nach seiner Ansicht komplizierteste der funktionierenden Legenden Wiens betreffen – das Wiener Kaffeehaus. Die drei wichtigsten Fragen für ihn sind dabei: Erstens „*Wer ist der Gast?*“, zweitens „*In welcher Art von Kaffeehaus sitzt er?*“ und drittens „*Was ist es für ein Kaffee, den er trinkt?*“. Für Studierende der Wirtschaftswissenschaften dürfte es naheliegend sein, diesen Fragen unmittelbar eine weitere folgen zu lassen, nämlich: „*Was kostet der Kaffee?*“. Das wird auch die erste in einer Reihe interessanter Fragen sein, mit denen wir uns in diesem Buch beschäftigen werden.

Damit sind wir beim Thema dieses Buches angelangt. Einfach ausgedrückt könnte man sagen, dass sich Statistik zunächst vor allem mit der Zusammenfassung von Daten beschäftigt. Diese Formulierung macht deutlich, was im Mittelpunkt der Statistik steht, nämlich Daten. Um es gleich vorwegzunehmen, Daten und Zahlen sind nicht dasselbe! Daten werden zwar häufig zahlenmäßig erfasst, allerdings stehen sie immer in einem sachlichen Kontext. Bei den in einer Klausur erzielten Punktezahlen kann man daher von Daten sprechen, bei beliebig vorgegebenen Zahlen wie 2, 5, 7, 10 dagegen nicht.

Angesichts einer immer komplexer werdenden Welt mit ihrem wachsenden Bedürfnis nach Daten ist es kaum überraschend, dass statistische Themen in Beruf und Alltag mehr und mehr an Bedeutung gewinnen, ob es nun um die aktuelle Preisentwicklung, die Höhe des Wirtschaftswachstums oder um den Ertrag und das Risiko einer Finanzanlage geht. In diesem Buch wird man dabei auch die Erfahrung machen, dass es oft verschiedene Möglichkeiten zur Beantwortung einer statistischen Frage bzw. zur Lösung eines statistischen Problems gibt, die jede für sich einen anderen Aspekt betont. Dieser Umstand hat der Statistik gelegentlich den Vorwurf eingebracht, dass man mit ihrer Hilfe letztlich alles beweisen kann.

TABELLE 1.1. Preise für eine Melange in Wiener Kaffeehäusern (in Euro)

1. Bezirk[2]									
3,10	2,40	3,00	3,20	3,70	4,00	3,20	3,40	2,80	3,20
3,50	3,40	3,10	4,20	3,30	3,00	4,40	3,10	2,90	4,40
3,60	3,30	3,10	2,90	4,40	3,60	3,60	3,50	3,50	3,40
übrige Bezirke (2. bis 23.)									
2,30	3,00	2,90	2,60	3,00	2,90	2,60	2,90	2,90	2,70
2,60	2,90	2,50	3,20	2,80	3,70	2,40	3,00	2,70	3,00
3,00	3,00	2,60	2,40	3,30	2,90	3,10	2,60	3,70	2,60
3,30	3,10	2,50	2,60	2,70	2,40	2,60	2,30	4,40	2,60

Quelle: Eigene Erhebungen, 2009/2010.

Um die Darstellung etwas konkreter zu machen, werden wir am Beginn einen einfachen Datensatz verwenden, anhand dessen wir verschiedene statistische Fragestellungen behandeln werden. Bei diesem Datensatz, der in Tabelle 1.1 wiedergegeben ist, handelt es sich um Preise für eine Tasse Kaffee (genauer: für eine Wiener Melange), die bei 70 Wiener Kaffeehäusern erhoben wurden.[3] Wir werden dabei versuchen, diesen Datensatz „statistisch“ etwas aufzubereiten. Dazu gehört insbesondere die Darstellung der Daten mit Hilfe von Tabellen bzw. Grafiken, aber natürlich auch der Hinweis auf verschiedene Möglichkeiten, wenn es zum Beispiel um die Berechnung eines geeigneten Mittelwerts oder einer Maßzahl zur Beschreibung der Streuung der Preise geht.

Es folgt ein kurzer Überblick über den Inhalt dieses Kapitels. Der erste Abschnitt beschäftigt sich mit einem elementaren statistischen Konzept – der Grundgesamtheit. Weitere wichtige Konzepte werden im zweiten Abschnitt behandelt, und zwar Merkmale (Variablen) und ihre Verteilungen. Dabei wird insbesondere auf die Unterschiede zwischen quantitativen und qualitativen Merkmalen bzw. zwischen diskreten und stetigen Merkmalen eingegangen. Im dritten Abschnitt werden, anhand verschiedener Beispiele, zwei Darstellungsformen betrachtet, die typisch für den Umgang mit statistischen Daten sind: Tabellen und Grafiken. Der vierte Abschnitt präsentiert den ersten Teil einer empirischen (das heißt datenbasierten) Untersuchung. Dabei werden Daten und Maßzahlen verschiedener Aktienindizes aus Deutschland (DAXK), Österreich (ATX), der Schweiz (SMI) und den USA (Dow-Jones Index) betrachtet.

1.1. Die Grundgesamtheit

Wenn man im Rahmen der Mengenlehre, in einem bestimmten Zusammenhang, verschiedene Mengen betrachtet, dann wird dabei stets vorausgesetzt, dass diese Teilmengen einer entsprechend vorgegebenen Grundmenge sind. Andernfalls könnte man zum Beispiel das Komplement einer Menge gar nicht bilden. Ähnlich geht man bei statistischen Untersuchungen in der Regel von der Abgrenzung der sogenannten Grundgesamtheit aus. Es handelt sich dabei um die einer Untersuchung zugrunde liegende Gesamtheit von Individuen oder Objekten. Beispiele für Grundgesamtheiten könnten etwa sein

Wahlberechtigte eines Staates
Einwohner einer Stadt
Kunden einer Bank
Studierende eines Studiengangs
PCs eines Unternehmens

Betrachtet man die Gesamtheit der Wahlberechtigten, so erhält man an einem Wahltag natürlich Informationen über deren Beteiligung an der Wahl und über die jeweiligen Parteipräferenzen. Oft besteht aber auch ein weiter reichendes Interesse an der Einstellung der Bevölkerung im Hinblick auf persönliche Ansichten, politische Vorhaben oder Entscheidungen. Dabei wird versucht, mit Hilfe von Umfragen durch Markt- und Meinungsforschungsinstitute, entsprechende Antworten von den Befragten zu erhalten.

An dieser Stelle kommt ein wichtiger Begriff ins Spiel, der in gewisser Hinsicht einen Gegensatz zur Grundgesamtheit bildet: die Stichprobe. Es handelt sich hierbei um eine Auswahl von „Objekten“ einer Grundgesamtheit, wobei der Auswahlprozess häufig zur Gänze oder zumindest teilweise zufällig ist. Zweck einer Stichprobenerhebung ist es, Informationen über verschiedene Aspekte einer Grundgesamtheit zu erhalten, ohne allerdings die Grundgesamtheit als Ganzes untersuchen zu müssen. Letzteres wird zum Beispiel vor allem aus Zeit- oder Kostengründen zweckmäßig sein. Die mit Hilfe einer Stichprobe gewonnene Information ist naturgemäß unvollständig und daher mit einer entsprechenden Unsicherheit behaftet. Auf dieses Problem der Unsicherheit im Zusammenhang mit Stichproben werden wir im Rahmen der Induktiven Statistik zurückkommen.

Natürlich verwendet man auch häufig den Begriff der Grundgesamtheit, ohne dass irgendeine Bezugnahme zu einer Stichprobe besteht. Eine Bestandsaufnahme oder Überprüfung sämtlicher PCs eines Unternehmens wäre dafür ein einfaches Beispiel. Ein weiteres wären die zu Beginn dieses Kapitels beschriebenen Melange-Preise von Wiener Kaffeehäusern.

1.2. Merkmale und Verteilungen

Im Hinblick auf die Grundgesamtheit 'Einwohner einer Stadt' könnte man an der Beantwortung von Fragen wie zum Beispiel nach dem Geschlecht, dem Alter, dem Familienstand oder der Stellung im Beruf interessiert sein. Eine Bank wird sich unter anderem für das Einkommen, das Vermögen oder die Altersstruktur ihrer Kunden interessieren. Dies führt uns zum wichtigen Begriffspaar Merkmal (auch Variable genannt) und Merkmalsausprägungen.

Merkmal und Merkmalsausprägungen

Ein Merkmal ist eine Zusammenfassung von Merkmalsausprägungen. Darunter versteht man Zahlenwerte oder Attribute, die den Objekten der Grundgesamtheit zugeordnet werden.

Bei den Melange-Daten handelt es sich somit um eine Untersuchung des Merkmals 'Preis für eine Melange', wobei die angegebenen Preise konkret beobachtete Ausprägungen dieses Merkmals darstellen.

Verständlicherweise soll eine eindeutige Zuordnung der Objekte der Grundgesamtheit zu den Merkmalsausprägungen erreicht werden. Dazu ist es erforderlich, dass die Ausprägungen eine Zerlegung (Partition) des Merkmals bilden. Das bedeutet, dass die einzelnen Ausprägungen sich gegenseitig ausschließen und jedem Objekt der Grundgesamtheit genau eine dieser Ausprägungen zugeordnet wird. Es könnte zum Beispiel sein, dass bei einer Umfrage das Merkmal 'Familienstand' erhoben wird. Sinnvollerweise sollten dann nicht nur die üblichen Ausprägungen ledig, verheiratet, geschieden usw. vorgesehen sein, sondern etwa auch die Möglichkeit, dass die Antwort verweigert wird.

Quantitative und qualitative Merkmale

Quantitative Merkmale sind Merkmale, deren Ausprägungen durch Zahlenwerte beschrieben werden.

Qualitative Merkmale sind Merkmale, deren Ausprägungen durch Attribute beschrieben werden.

Es gibt verschiedene Möglichkeiten, Merkmale einzuteilen. Eine wichtige Unterteilung ist dabei die in quantitative und qualitative Merkmale. Dies hat vor allem Konsequenzen für die Anwendung statistischer Konzepte und Methoden.

Typische Beispiele für quantitative Merkmale sind etwa das Einkommen oder das Lebensalter von Personen, der Kurs einer Aktie oder die Zahl der täglich eintreffenden E-Mails. Die Melange-Preise gehören natürlich ebenfalls zu den quantitativen Merkmalen. Beispiele für qualitative Merkmale sind das Geschlecht, die Stellung im Beruf (Arbeiter, Angestellter, Beamter, Selbständiger), der Familienstand oder der ordentliche Wohnsitz. Quantitative Merkmale lassen sich noch weiter unterteilen in diskrete und stetige Merkmale.

Diskrete und stetige Merkmale

Diskrete Merkmale sind Merkmale, deren Ausprägungen sich nacheinander aufzählen lassen.

Stetige Merkmale sind Merkmale, deren Ausprägungen (zumindest prinzipiell) jeden Wert innerhalb eines bestimmten Intervalls annehmen können.

Beispiele für diskrete Merkmale sind etwa die Anzahl der innerhalb eines bestimmten Zeitraums eintreffenden Ereignisse (zum Beispiel Geburten, Todesfälle), die Kinderzahl pro Familie usw. Zu den stetigen Merkmalen gehören etwa das Alter, die Größe einer Person oder auch die Zeit, die Sie jede Woche vor dem PC verbringen. In diesem Zusammenhang spricht man auch von diskreten und stetigen Daten. Merkmale, die sich in Geldeinheiten ausdrücken lassen (monetäre Größen), werden häufig wie stetige Merkmale behandelt, obwohl sie eigentlich zu den diskreten Merkmalen gehören. Beispiele dafür sind das Einkommen privater Haushalte, der Umsatz von Unternehmen oder auch die Kurse von Aktien.

Manche statistische Daten kann man an einem bestimmten Stichtag erheben, zum Beispiel die Bevölkerungszahl oder die Zahl der Kunden einer Bank. Solche Daten werden auch Bestandsdaten genannt. Andere Daten lassen sich sinnvoll nur über einen gewissen Zeitraum erheben, wie zum Beispiel Geburten, Todesfälle oder auch das Einkommen. Diese Daten nennt man Bewegungs- oder Stromdaten.

Eine weitere wichtige Klassifizierung von Merkmalen stellt das Skalenniveau dar. Dieses sieht eine Unterteilung von Merkmalen gemäß den folgenden Skalen vor: Nominalskala, Ordinalskala (Rangskala) und Verhältnisskala (metrische Skala). Je nachdem, auf welchem Skalenniveau sich eine Variable befindet, lassen sich bestimmte Operationen durchführen. Qualitative Variablen weisen demnach eine Nominalskala auf. Die Ausprägungen stellen in diesem Fall unterschiedliche Bezeichnungen dar, bei denen aber keine Anordnung wie etwa größer/kleiner vorliegt. Bei Ausprägungen, die eine Rangordnung darstellen, zum Beispiel Ratings

von Unternehmen, ist es nicht sehr sinnvoll, diese zu addieren (auch wenn sie in Form von Zahlen vorliegen sollten). Metrische Variablen treten sehr häufig auf, wie zum Beispiel die Preise von Waren oder Dienstleistungen, das Einkommen, die Körpergröße oder das Alter einer Person.

Verteilung

Unter der Verteilung eines Merkmals versteht man die Zuordnung der Merkmalsausprägungen zu den Objekten der Grundgesamtheit.

Der Begriff der Verteilung gehört zu den wichtigsten Begriffen der Statistik. Gelegentlich wurde die Statistik als Lehre von den Verteilungen bezeichnet. Auch wenn eine solche Definition heutzutage überholt ist, unterstreicht dies die zentrale Bedeutung des Begriffs für die Statistik.

Im Rahmen der Deskriptiven Statistik werden wir nur Verteilungen betrachten, die auf Beobachtungen beruhen, das heißt auf beobachteten Daten. Derartige Verteilungen werden auch als empirische Verteilungen bezeichnet. Von besonderem Interesse ist dabei die Frage: Welche Größen und welche Darstellungsformen verwendet man im Zusammenhang mit Verteilungen? Zunächst zum ersten Teil der Frage. Zur Darstellung (empirischer) Verteilungen verwendet man sehr oft absolute oder relative Häufigkeiten. Haben wir ein qualitatives Merkmal oder

Tabelle 1.2. Die Melange-Preise

Preis	absolute Häufigkeit	Preis	absolute Häufigkeit	Preis	absolute Häufigkeit
2,20	0	3,00	8	3,80	0
2,30	2	3,10	6	3,90	0
2,40	4	3,20	4	4,00	1
2,50	2	3,30	4	4,10	0
2,60	9	3,40	3	4,20	1
2,70	3	3,50	3	4,30	0
2,80	2	3,60	3	4,40	4
2,90	8	3,70	3	4,50	0

Quelle: Eigene Erhebungen, 2009/2010.

ein diskretes Merkmal, dann versteht man unter der absoluten Häufigkeit einer Ausprägung die Anzahl der Objekte der Grundgesamtheit, die diese Ausprägung besitzen. Für die Melange-Preise zeigt Tabelle 1.2 die Einzelpreise zwischen 2,20

Euro und 4,50 Euro, wobei 10 Cent-Intervalle verwendet wurden, um etwaige Lücken zu vermeiden.

Symbolisch bezeichnet man die absolute Häufigkeit einer Ausprägung i oft mit h_i. Die relative Häufigkeit f_i erhält man, indem man die absolute Häufigkeit h_i durch die Anzahl n der Objekte der Grundgesamtheit (den Umfang der Grundgesamtheit) dividiert:

$$f_i = \frac{h_i}{n}$$

Die Summe der relativen Häufigkeiten ergibt natürlich den Wert Eins

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{h_i}{n} = \frac{1}{n} \sum_{i=1}^{k} h_i = 1$$

wobei k die Anzahl der verschiedenen Merkmalsausprägungen bedeutet. In der folgenden Tabelle werden die Melange-Preise in einer etwas übersichtlicheren Form dargestellt und zwar mit Hilfe von Intervallen bzw. Klassen. Dabei sind jeweils Preise innerhalb eines bestimmten Bereichs zu einer gemeinsamen Klasse zusammengefasst. Mit anderen Worten, die Ausprägungen im Bereich zwischen 2,20 Euro bis unter 2,60 Euro bilden die erste Preisklasse, usw. Die absolute Häufigkeit einer Klasse ist dann die Zahl derjenigen Preise, die genau in diesen Bereich fallen. Ähnlich verhält es sich mit der relativen Häufigkeit einer Klasse.

Tabelle 1.3. Verteilung der Melange-Preise

Preisklasse (von ... bis unter)	absolute Häufigkeit	relative Häufigkeit (in %)
2,20 – 2,60	8	11,4
2,60 – 3,00	22	31,4
3,00 – 3,40	22	31,4
3,40 – 3,80	12	17,1
3,80 – 4,20	1	1,4
4,20 – 4,60	5	7,1
Summe	70	100,0

Quelle: Eigene Berechnungen, 2009/2010.

Eine derartige Darstellung, bei der Merkmalsausprägungen in Klassen zusammengefasst werden und die absoluten und relativen Häufigkeiten auf der Grundlage dieser Klassen gebildet werden, ist bei stetigen Merkmalen allgemein üblich. Eine Altersverteilung könnte zum Beispiel auf der Klasseneinteilung 0–9 Jahre,

10–19 Jahre, 20–29 Jahre usw. beruhen. Absolute und relative Häufigkeiten beziehen sich dann immer auf die jeweiligen Klassen, wobei es dabei auch zu ungleichen Klassenbreiten kommen kann (Einkommensverteilung).

Gelegentlich werden in Tabellen die kumulierten absoluten und relativen Häufigkeiten ausgewiesen, wie dies in Tabelle 1.4 gezeigt wird. Dabei enthält die zweite

Tabelle 1.4. Verteilung der Melange-Preise

Preisklasse (von ... bis unter)	H_i	F_i (in %)
2,20 – 2,60	8	11,4
2,60 – 3,00	30	42,9
3,00 – 3,40	52	74,3
3,40 – 3,80	64	91,4
3,80 – 4,20	65	92,9
4,20 – 4,60	70	100,0

Quelle: Eigene Berechnungen, 2009/2010.

Spalte die kumulierten absoluten Häufigkeiten (hier mit H_i bezeichnet), während die dritte Spalte die entsprechenden kumulierten relativen Häufigkeiten enthält (hier mit F_i bezeichnet). Dieser Tabelle kann man zum Beispiel entnehmen, dass 52 von insgesamt 70 Preisen, das heißt etwa 75 %, unter 3,40 Euro liegen. Entsprechend liegen dann etwa 25 % der Preise bei 3,40 Euro oder darüber.

Abschließend sei noch auf einen wichtigen Aspekt hingewiesen. Wenn Häufigkeiten verwendet werden, dann denkt man wohl in der Regel an absolute Häufigkeiten, die natürlich ihre Berechtigung haben. Allerdings bringt dieses Vorgehen oft Nachteile mit sich. Sollen zum Beispiel zwei Verteilungen auf der Basis von Histogrammen verglichen werden, dann sind deren Gesamtflächen unterschiedlich groß, falls die beiden Gesamthäufigkeiten ungleich sind. Das erschwert natürlich Vergleiche. Für solche Zwecke sind daher relative Häufigkeiten die ideale Alternative, da sich diese immer auf 100 % summieren und somit eine gemeinsame Vergleichsbasis bilden. Dies werden wir im Folgenden stärker berücksichtigen.

Ähnliche Probleme treten übrigens auch bei anderen Situationen auf. Angenommen, man würde erfahren, dass die Zahl der Kriminalfälle in früheren Jahrzehnten niedriger als die derzeitigen Zahlen waren, was einige vielleicht mit der Aussage „*Früher war alles besser*“ kommentieren könnten. Allerdings sollte man bei solchen Vergleichen auch die Bevölkerungszahlen berücksichtigen, zum Beispiel in der Form „Zahl der Kriminalfälle auf 100.000 Einwohner“. Dann hätte man wiederum eine gemeinsame Basis, auf der man dann Vergleiche durchführen könnte.