

## 2 Grundzüge der Datenerhebung

### 2.1 Merkmale, statistische Einheit, statistische Masse

Vor jeder statistischen Analyse muss das Untersuchungsziel genau angegeben sein. Obwohl die Vorgabe dieses Zieles nicht zum unmittelbaren Problemkreis der Statistik, sondern zum Anwendungsbereich der jeweiligen Substanzwissenschaft gehört, hat die Formulierung des Ziels so zu erfolgen, dass seine statistische Bearbeitung möglich wird. Das bedeutet zunächst die genaue Festlegung des zu quantifizierenden Phänomens. Diese Festlegung kann bereits so erfolgt sein, dass eine unmittelbare Beobachtung möglich wird. Ist das nicht der Fall, müssen die im Untersuchungsziel enthaltenen theoretischen Konstrukte identifiziert werden. **Theoretische Konstrukte** sind fachwissenschaftliche Bezeichnungen, die nicht beobachtbare Sachverhalte festlegen. Beispiele für theoretische Konstrukte sind aus der Physik: Atom, Gravitation, (Magnet-) Feld; aus der Psychologie: Intelligenz, Liebe, Bewusstsein, Deprivation; aus den Wirtschaftswissenschaften: Kosten, Kapazität, Wohlstand, Konjunktur, Inflation; aus der Soziologie: Bildung, (berufliche) Stellung, Akzeptanz. Da theoretische Konstrukte nicht beobachtbar sind, müssen operationale Definitionen entwickelt werden, die den Übergang von der theoretischen Sprache zur Beobachtungssprache leisten. **Operationale Definitionen** ordnen theoretischen Konstrukten Zählbegriffe der Statistik zu. Die mit **Adäquation** bezeichnete Zuordnung gelingt nicht immer ohne Verlust: Das Bedeutungsfeld des theoretischen Konstrukts ist oft allgemeiner als das des Zählbegriffs. Eine solche Diskrepanz bezeichnet man als **Adäquationsproblem**, das zwar nicht gänzlich beseitigt werden kann, jedoch auf jeden Fall zu minimieren ist. Denn Fehler in dieser frühen Phase einer statistischen Untersuchung lassen sich auch mit ausgereiften statistischen Verfahren nicht mehr kompensieren. Sind die Zählbegriffe festgelegt, muss das Untersuchungsziel noch in zeitlicher und räumlicher Hinsicht präzisiert werden.

Der **statistische Zählbegriff** definiert eine beobachtbare Eigenschaft, die **statistisches Merkmal** genannt wird. Die möglichen Erscheinungsformen eines Merkmals heißen **Merkmalsabstufungen**, **Merkmalswerte**, **Merkmalsausprägungen** oder kurz nur **Ausprägungen**, die in endlicher oder unendlicher Anzahl vorliegen können. Die Objekte, an denen das Merkmal in Erscheinung tritt und die der räumlichen und zeitlichen Abgrenzung des Untersuchungsziels genügen, heißen **statistische Einheit**, **Untersuchungseinheit**, **Merkmalsträger** oder kurz **Element**.

Die praktisch unbegrenzte Menge statistischer Merkmale lässt sich nach verschiedenen Kriterien gruppieren. Die für statistische Untersuchungen grundlegende Klassifikation trennt zwischen qualitativen (**klassifikatorischen** bzw. **kategorialen**), ordinalen (**komparativen**) und quantitativen (**metrischen** bzw. **kardinalen**) Merkmalen.

Ein **qualitatives Merkmal** liegt vor, wenn sich seine Ausprägungen nur durch ihre Art unterscheiden. Es gibt daher höchstens abzählbar viele, d.h. endlich viele oder abzählbar unendlich viele Merkmalsausprägungen. Beispiele für qualitative Merkmale sind (natürliche) Haarfarbe (Ausprägungen: blond, schwarz, rot, grau, weiß) oder Beschäftigungsverhältnis (Ausprägungen: Arbeiter, Angestellte, Beamter,...). Bei einem **ordinalen Merkmal** lassen sich die Merkmalsausprägungen intensitätsmäßig abstufen, also in eine Rangordnung bringen. Beispiele sind die Merkmale: Zensuren, Motivation, Windstärke oder Nutzen eines Warenkorbs. Ein **quantitatives Merkmal** besitzt Merkmalsausprägungen, die gezählt oder durch Vergleich mit einem vorgegebenen Maßstab gemessen werden können. Beispiele hierfür sind die Merkmale: Güterproduktion, Einkommen, Beschäftigte oder Körpergröße.

Quantitative Merkmale lassen sich noch gemäß der Anzahl möglicher Ausprägungen weiter in diskrete oder stetige (**kontinuierliche**) Merkmale unterteilen. Ein **diskretes Merkmal** liegt vor, wenn die Anzahl seiner Ausprägungen endlich oder abzählbar unendlich ist. Ein **stetiges Merkmal** besitzt

überabzählbar viele Merkmalsausprägungen. Ist die Anzahl der Ausprägungen bei einem diskreten Merkmal sehr groß, bezeichnet man es als **quasistetig**. Die „Belegschaft einer Unternehmung“ ist ein diskretes, das Gewicht eines Menschen ein stetiges Merkmal. Das Merkmal „verkaufte Brötchen in deutschen Städten an einem bestimmten Tag“ ist zwar diskret, lässt sich aber durchaus als quasistetiges Merkmal auffassen.

Die Möglichkeit einer sinnvollen Interpretation der Summe von Merkmalsausprägungen verschiedener Merkmalsträger erlaubt eine Klassifikation in intensive und extensive Merkmale. Kann die Summe der Merkmalsausprägungen bei verschiedenen Merkmalsträgern nicht sinnvoll interpretiert werden, wohl aber ihr Durchschnitt, liegt ein **intensives Merkmal** vor. Intensive Merkmale sind z.B. der Intelligenzquotient, die Körpergröße oder der Preis eines Gutes zu verschiedenen Zeitpunkten bzw. an unterschiedlichen Orten. Lassen sich die Summe der Merkmalsausprägungen über verschiedene Merkmalsträger und damit auch ihr Durchschnitt sinnvoll interpretieren, spricht man von einem **extensiven Merkmal**. Ein extensives Merkmal ist z.B. das Jahreseinkommen eines Haushalts; summiert man über alle Haushalte einer Volkswirtschaft, erhält man das Volkseinkommen eines Jahres.

Können die Merkmalsausprägungen direkt am Merkmalsträger beobachtet werden, spricht man von einem **manifesten Merkmal**; ist dies nicht möglich, liegt ein **latentes Merkmal** vor. Manifeste Merkmale sind z.B. die Regenmenge an einem Ort zu einer bestimmten Zeit oder die verkaufte Warenmenge in einer Periode. Ein latentes Merkmal ist im statistischen Sinne noch nicht hinreichend operationalisiert. Seine häufig vorgenommene Ersetzung durch ein geeignetes manifestes Merkmal behebt zwar diesen Mangel, hat aber auch das Adäquationsproblem zur Folge. Die Ersetzung des latenten Merkmals „Bildung“ durch das manifeste Merkmal „Schulabschluss“ verdeutlicht dies.

Kann eine statistische Einheit gleichzeitig Träger mehrerer Merkmalsausprägungen desselben Merkmals sein, handelt es sich um ein **häufbares Merkmal**. Beispiele hierfür sind: Staatsangehörigkeit, Beruf oder Studienfach. Nimmt ein Merkmal nur zwei verschiedene Ausprägungen an, heißt es **binär** oder **dichotom**. Da seine Werte meist mit den Ziffern 0 oder 1 kodiert werden, spricht man auch von einer (0,1)-Variablen. Analog hierzu bezeichnet man ein Merkmal mit nur drei Ausprägungen als **trichonom** bzw. **trinär**.

Die Gesamtheit aller hinsichtlich eines Untersuchungszieles relevanten statistischen Einheiten (Merkmalsträger) bildet die **statistische Masse**, die auch **Grundgesamtheit**, **Untersuchungsgesamtheit**, **Auswahlgesamtheit** oder **Population** genannt und mit dem griechischen Großbuchstaben Omega  $\Omega$  gekennzeichnet wird. Eine statistische Masse ist demnach als eine sachlich, zeitlich und räumlich wohl abgegrenzte Menge von Merkmalsträgern  $\omega_j$  (griechischer Kleinbuchstabe Omega) definiert:  $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$ . Gehört ein Merkmalsträger  $\omega_j$  zu einer Grundgesamtheit  $\Omega$ , schreibt man dafür:  $\omega_j \in \Omega$ . Ist die Anzahl der Merkmalsträger endlich, liegt eine endliche Gesamtheit vor. Dies ist bei der deskriptiven Statistik der Regelfall; man bezeichnet endliche Gesamtheiten auch als **Realgesamtheiten**. Eine unendliche Gesamtheit besitzt unendlich viele Elemente. Hierzu zählen die hypothetischen Grundgesamtheiten der induktiven Statistik. Teilgesamtheiten entstehen, wenn ausgewählte Elemente einer Grundgesamtheit zu Teilmengen zusammengefasst werden.

Statistische Massen können hinsichtlich der zeitlichen Abgrenzung als Bestands- oder Bewegungsmasse vorliegen. Eine **Bestandsmasse (stock)**, auch **Streckenmasse** genannt, enthält Elemente mit bestimmter zeitlicher Verweildauer. Die Elemente treten zu einem Zeitpunkt in die Masse ein und verlassen diese wieder nach einer bestimmten Dauer. Deshalb sind Bestandsmassen stets zeitpunktbezogen definiert. Der Kapitalstock einer Volkswirtschaft zum 31.12. eines Jahres ist eine Bestandsmasse und umfasst alle Inve-

stitutionsgüter, die zu früheren Zeitpunkten installiert wurden, aber am 31.12. noch in Betrieb sind. Weitere Beispiele für Bestandsmassen sind: Wohnbevölkerung eines Landes oder Vermögen eines Haushalts, jeweils zu bestimmten Stichtagen.

Eine **Bewegungsmasse (flow)** bzw. **Ereignis-** oder **Punktmasse** liegt vor, wenn erst die Festlegung eines Zeitintervalls die Zusammenfassung zeitpunktbezogener statistischer Einheiten zu einer Masse ermöglicht. Da solche Massen für vorgegebene Zeitspannen definiert sind, variieren sie auch mit diesen. Das Bruttoinlandsprodukt einer Volkswirtschaft in einem Jahr ist eine Bewegungsmasse. Ausschlaggebend hierfür ist, dass jede Einheit der für einen Endzweck geschaffenen Güter und Dienstleistungen zu einem Zeitpunkt (Ereigniszeitpunkt) des vorgegebenen Jahres den Produktionsprozess verlässt. Aus den gleichen Gründen bilden die Käufe eines privaten Haushalts in einer Woche eine Bewegungsmasse: Obwohl jeder Kauf eine gewisse Zeit bindet, kann er doch als punktuelles Ereignis innerhalb der Woche aufgefasst werden.

Enthält eine Bewegungsmasse die Zugänge, Abgänge oder die saldiereten Zu- und Abgänge (Nettozugänge) einer Bestandsmasse, bezeichnet man beide Massen wegen ihres sachlogischen Zusammenhangs als **korrespondierende Massen**. Addiert man zu einer Bestandsmasse für den Stichtag  $t_1$  die korrespondierende Bewegungsmasse des Zeitintervalls  $\Delta t = t_2 - t_1 > 0$ , resultiert die neue Bestandsmasse zum Stichtag  $t_2$ . Diese Verknüpfung heißt **Fortschreibung** und bietet eine einfache Möglichkeit, umfangreiche Bestandsmassen zu aktualisieren. Fügt man beispielsweise zum Kapitalstock einer Volkswirtschaft zu Jahresbeginn die korrespondierende Bewegungsmasse „Nettoinvestitionen dieses Jahres“ hinzu, erhält man den Kapitalstock der Volkswirtschaft, der am Anfang des nächsten Jahres vorhanden ist.

## Übungsaufgaben zu 2.1

2.1.1 Was versteht man unter Adäquation?

2.1.2 Geben Sie für folgende Merkmale an, ob sie qualitativ, ordinal, oder quantitativ und diskret oder stetig sind!

Gewicht, Körpergröße, Haarfarbe, Preis, Qualität, Volumen, Tagesumsatz, Steuerklasse, Staatsangehörigkeit, Erwerbsstatus, Lagerbestand.

Nennen Sie zu jedem Merkmal mögliche Ausprägungen!

2.1.3. Welche der folgenden Merkmale sind intensiv, extensiv, manifest, latent oder häufbar?

Einkommen, Zensuren, Kosten, Körpergröße, Haarfarbe, Studienfach.

## 2.2 Messen und Skalieren

Die Festlegung der beobachtbaren Merkmalsausprägungen geschieht in der Statistik unabhängig von der Art des Merkmals durch Zählen oder Messen. Unter **Messen** versteht man die nach einer angegebenen Regel vorgenommene eindeutige Zuordnung von Zahlen zu den Merkmalsausprägungen. Damit nach dem Messen dieselbe Ordnung der Merkmalsträger gemäß ihrer Merkmalsausprägungen vorliegt, muss eine Skala verwendet werden. Durch eine **Skala** gelingt die relationstreue Abbildung der Merkmalsausprägungen in ein Zahlensystem, das meist durch die Menge der reellen Zahlen gegeben ist. Messvorschrift und geeigneter Skalentyp sind bereits durch die mit der operationalen Definition vorgenommenen Zuordnung von Zählbegriffen zu theoretischen Konstrukten festgelegt. Die dort angestrebte Minimierung des Adäquationsproblems führt zu empirisch sinnvollen Messvorschriften. Beispielsweise könnten die Monatseinkommen von Haushalten durch die Höhe des ihnen entsprechenden Centstapels in Meter gemessen werden; informationsreicher und damit sinnvoller ist aber eine Messung in Geldeinheiten. Die

grundlegende Klassifikation in qualitative, ordinale oder quantitative Merkmale legt die geeignete Skala fest. Damit sind auch diejenigen mathematischen Transformationen bestimmt, denen die Messungen unterzogen werden können, ohne dass sich dadurch die vorgegebene, natürliche Ordnung der Merkmalsausprägungen ändert. Die Kenntnis ordnungserhaltender Transformationen ist für Maßeinheitsänderungen bedeutsam.

Bei qualitativen Merkmalen bedeutet die Zuordnung von Zahlen zu den einzelnen Merkmalsausprägungen lediglich eine neue Kennzeichnung. Die verwendete Skala bezeichnet man deshalb als **Nominalskala**. Die Zahlenzuordnung heißt **Kodierung**, die Zahlen selbst heißen **Kennzahlen**. Da die einzige Funktion in der Unterscheidung der Merkmalsausprägungen besteht, kann jede getroffene Zahlenzuordnung durch eine eindeutige Transformation in eine andere Zahlenzuordnung überführt werden. Die Ausprägungen des qualitativen Merkmals Haarfarbe könnte mit der Kodierung: 1 = rot, 2 = braun, 3 = blond, 4 = schwarz, 5 = grau und 6 = weiß, genauso gut aber mit sechs anderen Zahlen unterschieden werden.

Ist bei der Messung von Merkmalsausprägungen nur ihre Rangordnung, nicht aber der Abstand zwischen benachbarten Rängen relevant, kommt eine **Ordinalskala** zur Anwendung. Alle komparativen Merkmale sind ordinal skaliert. Da die zugeordneten Zahlen nur die Rangordnung wiedergeben, können sie mit **streng monoton steigenden (isotonen) Transformationen** in andere Zahlen abgebildet werden. Eine Transformation  $T$  heißt isoton, wenn aus  $x_1 < x_2$  immer folgt:  $T(x_1) < T(x_2)$ . Die in der Wirtschaftstheorie verwendete Nutzenfunktion ist ein weiteres Beispiel für ordinal skalierte Messung.

Lassen sich Merkmalsausprägungen in eine Rangfolge bringen und ist der Abstand zwischen je zwei Ausprägungen definiert, bilden die zugeordneten Zahlen eine **Intervallskala**. Alle Intervallskalen können durch die Funktion  $y = ax + b, a > 0$  transformiert werden, ohne dass sich der Skalentyp ändert.

Zum Beispiel ist die Temperaturmessung in Grad Celsius oder in Grad Fahrenheit intervallskaliert. Bei einer Temperatur von  $4^{\circ}\text{C}$  ist es nicht doppelt so warm wie bei  $2^{\circ}\text{C}$ , jedoch liegt derselbe Temperaturunterschied wie bei  $18^{\circ}\text{C}$  und  $20^{\circ}\text{C}$  vor.

Können Merkmalsausprägungen in eine Rangordnung gebracht werden und sind Abstand sowie Verhältnis zweier Merkmalsausprägungen definiert, erfolgen die Messungen der Ausprägungen mit einer **Verhältnisskala (Ratioskala)**. So skalierte Merkmale besitzen einen sachlogisch begründbaren natürlichen Nullpunkt, aber die Maßeinheit ist noch willkürlich. Da der natürliche Nullpunkt durch eine Transformation nicht verschoben werden darf, sind nur linear homogene Transformationen wie  $y = ax, a > 0$  zulässig. Verhältnisskalierte Merkmale sind z.B. (Güter-) Preise, Länge, Gewicht oder Temperatur in Grad Kelvin. Während der natürliche Nullpunkt der ersten drei angegebenen Merkmale intuitiv einleuchtet, wird er bei Grad Kelvin durch die niedrigste mögliche Temperatur auf der Erde festgelegt: 0 Grad Kelvin entspricht  $-273,15^{\circ}$  Celsius. Mit der Transformation  $y = ax$  lassen sich Änderungen der Maßeinheit erreichen. Ist  $x$  der in Cent gemessene Preis eines Gutes, stellt  $y = \frac{1}{100}x$  den Güterpreis in der Maßeinheit EUR dar.

Besitzen Merkmale zusätzlich zu den Eigenschaften, die zu einer Verhältnisskala führen, noch eine natürliche Skaleneinheit, verwendet man bei ihrer Messung eine **Absolutskala**, die nicht transformiert werden kann. Beispiele für absolut skalierte Merkmale sind die Bevölkerung einer Region und Stückzahlen.

Nominal- und Ordinalskala heißen **topologische Skalen**; Intervall-, Verhältnis- und Absolutskala bezeichnet man als **Kardinal-** bzw. **metrische Skalen**. Alle quantitativen Merkmale sind metrisch skaliert. Daher liegen bei einem diskreten Merkmal in jedem Intervall  $(a, b) \subset \mathbb{R}, a < b, \mathbb{R}$ : Menge der reellen Zahlen, nur endlich viele Messungen. Ist das Merkmal hingegen stetig, bilden seine Ausprägungen ein Kontinuum, das entweder durch die

Menge der reellen Zahlen selbst oder durch eine geeignete Teilmenge gegeben wird. Endliche Messgenauigkeiten führen aber dazu, dass in der Realität jedes stetige Merkmal „nur“ diskret vorliegt.

Die Skalen und damit auch die Merkmale sind gemäß der zu erfüllenden Bedingungen hierarchisch aufsteigend geordnet als: Nominal-, Ordinal- und Kardinalskala. Mit aufsteigender Ordnung nimmt der Informationsgehalt der Merkmale zu. Während der Übergang von einer höheren zu einer niedrigeren Stufe der Skalenhierarchie mit Informationsverlust möglich ist, gelingt der umgekehrte Übergang — wenn überhaupt — erst nach Änderung der operationalen Definition.

Ein Merkmal bildet durch Messen seiner Ausprägungen jeden Merkmalsträger  $\omega_j \in \Omega$  in eine Skala  $S$  ab, die Teilmenge der reellen Zahlen  $\mathbb{R}$  ist:  $S \subset \mathbb{R}$ . Kommt es auf eine sachliche Spezifikation des Merkmals nicht an, sondern steht nur der Abbildungsaspekt im Vordergrund, bezeichnet man das Merkmal als „**statistische Variable  $X$** “. Der Abbildungsvorgang stellt sich formal dann dar als:

$$X : \Omega \longrightarrow S \subset \mathbb{R}. \quad (2.1)$$

Mit der Definition (2.1) ist ausgeschlossen, dass  $X$  ein häufbares Merkmal sein kann. Bei häufbaren Merkmalen könnte es vorkommen, dass ein Merkmalsträger  $\omega_j$  mindestens zwei Merkmalsausprägungen aufweist. Die statistische Variable  $X$  würde  $\omega_j$  mindestens zwei Zahlen zuordnen;  $X$  wäre dann aber keine Abbildung mehr.

Das Bild von  $\omega_j \in \Omega$  unter  $X$  heißt Beobachtung von  $X$  und wird mit  $x_j$  bezeichnet:  $x_j = X(\omega_j)$ . Die Gesamtheit aller Beobachtungen  $x_j$  sind die statistischen Daten (Datensatz). Sie müssen nicht alle verschieden sein, da mehrere Merkmalsträger dieselbe Merkmalsausprägung und damit denselben Messwert aufweisen können. Hingegen sind alle Elemente der Menge  $\{X(\omega_j), \omega_j \in \Omega\}$  wegen der Mengendefinition verschieden. Diese Menge stellt

die unterschiedlichen Ausprägungen von  $X$  dar, die im Datensatz vorkommen. Zur Unterscheidung von den Beobachtungen werden die Elemente dieser Menge mit  $x_i$  bezeichnet:  $x_i \in \{X(\omega_j), \omega_j \in \Omega\}$ . In den wenigen Fällen, in denen die Verwendung eines Wertes, z.B.  $x_5$ , als Ausprägung oder als Beobachtung nicht klar aus dem Zusammenhang hervorgeht, wird zur Verdeutlichung  $x_{i=5}$  oder  $x_{j=5}$  geschrieben.

Wird die Skala  $S$  einer statistischen Variablen  $X$  in abzählbar viele halboffene Intervalle zerlegt, spricht man von **Klassierung** bzw. **Klasseneinteilung**. Die Klassenbildung kann entweder durch rechtsgeschlossene  $(x'_{k-1}, x'_k]$  oder linksgeschlossene  $[x'_{k-1}, x'_k)$  Intervalle mit  $k \in \mathbb{N}$ ,  $\mathbb{N}$ : Menge der natürlichen Zahlen, erfolgen. Die Klassengrenzen  $x'_{k-1}$  und  $x'_k$  müssen nicht notwendigerweise zu der Menge der Ausprägungen gehören.

## Übungsaufgaben zu 2.2

2.2.1 Mit welchen Skalen sind folgende Merkmale zu messen?

- (1) Gewicht, (2) Körpergröße, (3) Haarfarbe, (4) Preis, (5) Qualität, (6) Volumen, (7) Tagesumsatz, (8) Steuerklasse, (9) Staatsangehörigkeit, (10) Erwerbsstatus, (11) Lagerbestand.

2.2.2 Die Temperaturmessung in Grad Fahrenheit ( $y$ ) erhält man aus der Temperaturmessung in Grad Celsius ( $x$ ) durch die Lineartransformation  $y = 32 + \frac{9}{5}x$ . Zeigen sie, dass  $y$  intervallskaliert ist!

## 2.3 Datengewinnung

Datenerhebungen sind meistens mit umfangreichen praktischen Problemen verbunden. Es soll daher hier nur die allgemeine Vorgehensweise skizziert werden. Die Gewinnung von Daten erfolgt durch die Datenerhebung, kurz

Erhebung genannt. Bevor sie durchgeführt wird, müssen die in den Abschnitten 2.1 und 2.2 aufgezeigten Probleme geklärt sein. Die hierzu notwendigen Entscheidungen bilden zusammen mit der Festlegung der Erhebungstechnik den **Erhebungsplan**. Bei Erhebungen ist zwischen **Erhebungs-** und **Untersuchungseinheit** (Merkmalsträger) zu unterscheiden. Als Erhebungseinheit bezeichnet man diejenige Einheit, bei der die Erhebung im technischen Sinne durchgeführt wird. Geschieht dies bei den Merkmalsträgern direkt, fallen Erhebungs- und Untersuchungseinheit zusammen und eine Unterscheidung ist überflüssig. Die Erhebungseinheiten gehören dann zur statistischen Masse  $\Omega$ . Bei einer Volkszählung z.B. wählt man gewöhnlich Haushalte als Erhebungseinheit, während die Untersuchungseinheit die Haushaltsmitglieder sind. Bei dieser Vorgehensweise gehören die Erhebungseinheiten nicht zu  $\Omega$ . Will man dagegen die Personenzahl von Haushalten ermitteln, stellen Haushalte Erhebungs- und Untersuchungseinheit (Merkmalsträger) dar; die Erhebungseinheit ist daher Element von  $\Omega$ .

Eine Erhebung kann als Voll- bzw. Totalerhebung oder als Teilerhebung angelegt sein. Bei einer **Vollerhebung** werden alle Merkmalsträger einer statistischen Masse erfasst. Bei einer **Teilerhebung** werden nur bestimmte Merkmalsträger aus  $\Omega$  untersucht. Teilerhebungen können durch begriffliches Ausgliedern nach bestimmten Merkmalsausprägungen (z.B. Bevölkerung unter 40 Jahren) oder durch Zufallsauswahl entstehen. Eine Teilerhebung ist leichter, schneller und vor allem billiger als eine Totalerhebung durchzuführen; dafür sind die Ergebnisse bei Zufallsauswahlen aber auch unsicherer als bei Vollerhebungen.

Die Datengewinnung kann nach drei Erhebungstechniken erfolgen: (1) Befragung, (2) Beobachtung und (3) Experiment. Bei Experimenten können Größen, die den Ausgang beeinflussen, kontrolliert werden. Während eine Datengewinnung auf experimenteller Basis für weite Teile der Physik, Chemie, Biologie und Medizin typisch ist, stellt sie bei den Wirtschafts- und Sozialwis-

senschaften (noch) die Ausnahme dar. Erste Entwicklungen in diese Richtung finden in den Teilgebieten Marketing, Personalwesen und Spieltheorie statt, die in der an Bedeutung gewinnenden experimentellen Wirtschaftsforschung aufgehen. Ähnliches gilt für die Einsatzmöglichkeiten der Beobachtungstechnik. Diese in den Naturwissenschaften sehr häufig eingesetzte Methode ist bei wirtschafts- und sozialwissenschaftlicher Datengewinnung nur eingeschränkt verwendbar. Dies liegt daran, dass hier Beobachtungen, die nicht mechanisch zu erheben sind, oft ausgeprägte subjektive Komponenten enthalten. Können solche Einflussfaktoren nicht ausgeschaltet bzw. bis zur Unerheblichkeit reduziert werden, ist die Vergleichbarkeit von Beobachtungsdaten zum selben Phänomen, aber von verschiedenen Beobachtern erstellt, kaum gewährleistet. Deshalb beschränkt man sich in den Wirtschafts- und Sozialwissenschaften mit der Erhebungstechnik „Beobachtung“ auf Merkmale, die von subjektiven Elementen weitgehendst unabhängig sind. Ein Beispiel ist die Verkehrszählung, obwohl auch hier die Genauigkeit von der Aufmerksamkeit des Beobachters abhängt. Beobachtung als Technik der Datenerhebung ist von dem einzelnen Ergebnis dieses Vorgangs, das ebenfalls Beobachtung genannt wird, zu unterscheiden. Im Zusammenhang mit statistischen Daten bezeichnet Beobachtung stets den Messwert  $x_j = X(\omega_j)$ , gleichgültig, wie die Beobachtungen erhoben wurden.

In den Wirtschafts- und Sozialwissenschaften dominiert als Erhebungstechnik die Befragung. Befragungen können in mündlicher oder schriftlicher Form oder als Kombination beider Formen durchgeführt werden. Sie haben den Vorteil, dass subjektive Beurteilungen und schwer oder gar nicht beobachtbare Sachverhalte erfasst werden können. Jedoch besteht die Gefahr einer bewussten oder unbewussten Verfälschung durch den Befragten. Diese Gefahr lässt sich durch Kontrollfragen und/oder indirekte Fragestellung verringern. **Kontrollfragen** beinhalten meistens das Gegenteil zu derjenigen Fragestellung, deren wahrheitsgemäße Beantwortung von besonderer Be-

deutung ist. Bei indirekter Fragestellung gewinnt man die eigentlich interessierende Information erst durch Kombination der Antworten zu unverfänglich erscheinenden Fragen. Allgemein sollte jede Frage einfach und präzise formuliert sein. Bei mündlicher Befragung (Interview) können wegen der Erläuterungsmöglichkeiten durch den Interviewer kompliziertere Fragen als bei der schriftlichen Befragung (Fragebogen) gestellt werden. Jedoch dürfen die Erläuterungen nicht suggestiv erfolgen. Wegen der Kosten, die Interviews verursachen, steht für die Beantwortung der Fragen weniger Zeit als bei einem Fragebogen ohne Interviewer zur Verfügung. Deshalb werden bei Interviews spontane Antworten häufiger als beim Fragebogen sein. Spontaneität kann bei der Meinungs- und Motivforschung aufschlussreicher als wohlüberlegtes Antworten sein; bei der Erfassung von Tatsachen dürfte sich diese Bewertung wohl umkehren. Werden Daten für ein bestimmtes Untersuchungsziel erstmalig erhoben, liegt eine **primärstatistische Erhebung** vor. Zieht man für das Untersuchungsziel bereits vorliegende, aber für andere Zwecke erhobene Daten heran, spricht man von **sekundärstatistischer Erhebung**. Sind diese Daten nicht mehr in reiner Form verfügbar, sondern bereits für den anderen Zweck aufbereitet, handelt es sich um eine **tertiärstatistische Erhebung**.

Der zeitliche Bezug der Datenerhebung führt zur Unterscheidung zwischen Längsschnitt- und Querschnitterhebung. Eine **Längsschnitterhebung** liegt vor, wenn die Beobachtungen für aufeinander folgende Zeitpunkte bzw. Perioden erhoben werden. Die gewonnenen Daten bilden dann eine **Zeitreihe**. Bei **Querschnitterhebungen** haben alle Beobachtungen denselben Zeitbezug. Die Daten für die Entwicklung des Inlandsprodukts in der Bundesrepublik Deutschland von 1975 bis 1994 erhält man mit einer Längsschnitterhebung; die Konsumausgaben der Haushalte einer Stadt in der 36. Woche eines Jahres hingegen mit einer Querschnitterhebung.

Eine für die Wirtschaftswissenschaften typische Unterscheidung ist mit dem Begriffspaar Mikro- und Makrovariablen gegeben. Wird eine statistische Va-

riable für einen Untersuchungsraum, z.B. eine Volkswirtschaft, inhaltlich so definiert, dass pro Periode oder Zeitpunkt nur eine Beobachtung eintreten kann, heißt sie **Makrovariable**; ist ihre Beobachtung an mehreren Merkmalsträgern möglich, liegt eine **Mikrovariable** vor. Das Inlandsprodukt einer Volkswirtschaft stellt demnach eine Makro-, die wöchentliche Konsumausgabe der Haushalte eine Mikrovariable dar. Für Makrovariablen sind nur Längsschnitt-, für Mikrovariablen sowohl Längs- als auch Querschnitterhebungen möglich. Die Kombination beider Erhebungsarten nennt man **Listentechnik**; die damit gewonnenen Beobachtungen heißen **Paneldaten**. Die Listentechnik ist nur bei Mikrovariablen anwendbar.

Die Einteilung in Makro- bzw. Mikrovariable variiert mit der Abgrenzung des Untersuchungsraumes. Wird dieser erweitert, können Makro- in Mikrovariablen übergehen. Ist z.B. der Untersuchungsraum als EURO-Zone festgelegt und das Merkmal wieder als Inlandsprodukt der hierzu gehörenden Volkswirtschaften (Merkmalsträger) definiert, stellt jetzt das Inlandsprodukt eine Mikrovariable dar, für die neben Längsschnitt- auch Querschnitt- bzw. Paneldaten erhoben werden können.

Die Untersuchungsgesamtheiten  $\Omega$  der deskriptiven Statistik sind stets endlich; der Erhebungsumfang wird mit  $n \in \mathbb{N}$  bezeichnet. Der jetzt aus  $n$  Beobachtungen  $x_j, j = 1, \dots, n$  bestehende Datensatz enthält  $m \in \mathbb{N}$  verschiedene Ausprägungen  $x_i, i = 1, \dots, m$ . Da Beobachtungen im Gegensatz zu den Ausprägungen gleich sein können, gilt immer  $m \leq n$ .

Bei einer statistischen Masse lässt sich nicht nur eine statistische Variable, sondern mehrere statistische Variablen beobachten, die zwecks Unterscheidung jetzt mit  $X_1, X_2, \dots, X_g$  bezeichnet werden. Jeder Merkmalsträger  $\omega_j, j = 1, \dots, n$  weist für jede Variable eine Beobachtung auf, es liegen somit insgesamt  $ng$  Beobachtungen vor. Um die Fülle an Beobachtungen zu strukturieren, verwendet man eine Beobachtungsmatrix (siehe Abbildung 2.1).

Man bezeichnet diese Matrix als **multivariaten (mehrdimensionalen)**

**Abb. 2.1: Beobachtungsmatrix**

statistische Merk- malsträger \ Variable	$X_1$	$X_2$	$\dots$	$X_g$
$\omega_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1g}$
$\omega_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2g}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\omega_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{ng}$

**Datensatz.** Wird für eine statistische Masse nur eine statistische Variable erhoben, liegt ein **univariater (eindimensionaler) Datensatz** vor. Die Beobachtungsmatrix geht dann in einen Spaltenvektor über, der — als Zeile geschrieben — ein  $n$ -Tupel ergibt. Für  $X_1$  erhält man:  $(x_{11}, x_{21}, \dots, x_{n1})$ . Da nur ein Merkmal  $X$  beobachtet wird, kann der zweite Index bei den Elementen des Vektors entfallen. Man bezeichnet das  $n$ -Tupel  $(x_1, \dots, x_n)$  als **Urliste** bzw. **Urmaterial**. Bei mindestens ordinal skalierten Merkmalen ist es vorteilhaft, die unterschiedlichen Ausprägungen  $x_i$  eines Datensatzes der Größe nach zu ordnen:  $x_{i=1}$  stellt dann die kleinste,  $x_{i=m}$  die größte Ausprägung dar.