Kapitel 1

Gesamtheiten und Variablen

In diesem Kapitel beschäftigen wir uns mit dem Ansatz statistischer Begriffsbildungen. Zwei Begriffe stehen im Mittelpunkt: 'Gesamtheit' und 'Variable'.[1] Die Grundidee ist, daß sich statistische Begriffsbildungen auf Gesamtheiten von Dingen, Menschen oder Situationen beziehen und daß man Eigenschaften der einzelnen Mitglieder solcher Gesamtheiten durch Variablen repräsentieren kann.

1.1 Einleitende Bemerkungen

1. Die eben angedeutete Grundidee geht davon aus: Man hat es mit einer Gesamtheit von Dingen, Menschen oder Situationen zu tun, die als Gesamtheit nicht ohne weiteres überschaubar ist, die man aber dennoch als eine Gesamtheit betrachten und beschreiben möchte. Man denke an einen Lagerverwalter, der ein großes Lager zu betreuen hat, in dem es sehr viele Dinge gibt, die sich durch gewisse Eigenschaften charakterisieren lassen. Um sich einen Überblick zu verschaffen, fertigt der Lagerverwalter eine Liste an und trägt ein, wieviele Dinge von jeder Art es in seinem Lager gibt. Dazu muß er natürlich zunächst einmal die Dinge, die sich in seinem Lager befinden, der Reihe nach betrachten und entsprechende Eintragungen in die Liste vornehmen. Ist er schließlich fertig, hat er in Gestalt der Liste ein übersichtliches Bild des Lagers: eine statistische Häufigkeitsverteilung für die Dinge, die sich in seinem Lager befinden. Dann kann er auch auf einfache Weise erreichen, daß seine Liste stets auf einem aktuellen Stand ist; er muß nur Eintragungen vornehmen, wenn Dinge aus seinem Lager herausgenommen werden oder wenn neue Dinge dazukommen.

2. Eine der Quellen der gegenwärtigen statistischen Methodenlehre ist die Sozialstatistik. Ihre Entwicklung verdankt sich einer Betrachtungsweise, die mit derjenigen unseres Lagerverwalters vergleichbar ist. Regierungen oder ihre Verwaltungsbeamten möchten sich einen Überblick über den Zustand ihres Herrschaftsgebiets verschaffen.[2] Von Anfängen einer Sozialstatistik kann man dementsprechend dort sprechen, wo zum erstenmal

[1] Wir verwenden einfache Anführungszeichen, um auf sprachliche Ausdrücke zu verweisen; doppelte Anführungszeichen werden verwendet, um Zitate kenntlich zu machen oder um anzudeuten, daß ein Ausdruck metaphorisch und/oder unklar ist.

[2] „Nur Menschengruppen, die in Staatsgesellschaft leben, sind einer Statistik fähig und würdig. Wilde Völker haben bloss eine Naturkunde.“ heißt es bei August Ludwig von Schlözer (1735–1809), einem Vertreter der „deutschen Universitätsstatistik“; vgl. John 1884, S. 105.

Volkszählungen durchgeführt wurden.[3] Erst in der frühen (europäischen) Neuzeit haben sich allerdings Ansätze zu einer systematischen Staatsbeschreibung entwickelt. Eine Dokumentensammlung für den Zeitraum von 1456 bis 1813, ergänzt durch einen ausführlichen Kommentar, wurde von M. Rassem und J. Stagl (1994) herausgegeben. Zumindest eine wesentliche Intention dieser frühen Bestrebungen kann exemplarisch anhand einer Schrift von G. W. Leibniz aus dem Jahr 1685 verdeutlicht werden:

„Ich nenne *Staats-Tafeln* eine schriftliche kurze Verfassung des Kerns aller zu der Landesregierung gehörigen Nachrichtungen, so ein gewisses Land insonderheit betreffen, mit solchen Vorteil eingerichtet, daß der hohe Landesherr alles darin leicht finden, was er bei jeder Begebenheit zu betrachten, und sich dessen als eines der bequemsten Instrumente zu einer löblichen Selbst-Regierung bedienen könne."

„Durch *Nachrichtungen* verstehe ich nicht allerhand Vernunftschlüsse und Regeln, so verständige Leute bei Gelegenheit und wenn sie darauf zu denken Ursach haben, selbsten leicht finden können, sondern was mehr in facto als Nachsinnen beruhet und daher nicht erfunden, sondern erfahren, erhöret und erlernet werden muß, zum Exempel was in einem Lande für eine Quantität seidener Zeuge oder wüllene Tücher jährlich konsumiert oder vertan werden, das ist eine Nachrichtung und beruhet in facto, es kann es auch keiner erraten, er sei so verständig als er wolle; ob aber ratsam, solche consumation vor sich gehen zu lassen, oder zu verbieten, und enger zu spannen, und ob man solche Manufakturen im Lande selbst einzuführen habe oder nicht, bestehet in ratiocinatio und gehöret nicht zu unserer Staatstafel, sondern kann vielmehr aus denen in der Staatstafel befindlichen Nachrichtungen von verständigen Leuten leicht geschlossen werden." (Zitiert nach Rassem und Stagl 1994, S. 321–329.)

Hier wird auch deutlich: den Zustand einer Gesamtheit von Dingen, Menschen oder Situationen kann man nicht erraten; sondern man muß zunächst ihre einzelnen Elemente untersuchen und kann erst dann versuchen, aus diesen Einzelbeobachtungen ein die Gesamtheit repräsentierendes Bild zu erzeugen.

3. Der Gedanke, daß es die Statistik mit Begriffsbildungen und Methoden zu tun hat, um Bilder zur Repräsentation von Gesamtheiten zu erzeugen, hat sich seit den Anfängen der Sozialstatistik bis heute erhalten. Es ist auch heute noch der Grundgedanke für die Wirtschafts- und Sozialstatistik, wie sie von den statistischen Ämtern und einer Vielzahl sozialwissenschaftlicher Einrichtungen betrieben wird. In besonders einflußreicher Weise wurde eine solche Konzeption der Statistik Ende des 19., Anfang des 20. Jahrhunderts von Georg v. Mayr (dem Begründer des heute noch bestehenden „Allgemeinen Statistischen Archivs") vertreten; hier ist eine seiner Definitionen:

[3] Hinweise finden sich bei John 1884, S. 15ff, Tyszka 1924, S. 83ff. Eine umfassende Sammlung von Informationen zur Geschichte von Volkszählungen gibt es bei Alterman, 1969. Interessante Informationen über sozialgeschichtliche Hintergründe finden sich auch bei Lindner, Wohak und Zeltwanger, 1984.

„Die Wissenschaft von den sozialen Massen. Aus der geordneten Beobachtung der einzelnen Elemente der vorbezeichneten sozialen Massen erwächst der wissenschaftliche Gewinn des Einblicks in den Bestand und die Gliederung der Massen und damit der Erkenntniß aller in diesen Massen zum Ausdruck kommenden gesellschaftlichen Zustände und Erscheinungen. Am vollständigsten ist diese Erkenntniß einer sozialen Masse dann, wenn erschöpfende Beobachtung sämmtlicher Elemente, aus denen sie zusammengesetzt ist, vorliegt, und wenn diese Beobachtung zugleich in exakter Weise, d.h. mittelst Zählens und Messens, durchgeführt wird. Der Grundgedanke der so gearteten wissenschaftlichen Arbeit ist, aus der exakten Beobachtung der Elemente, aus deren angemessener Gruppierung, Gliederung und Vergleichung die Erkenntniß des Ganzen, d.i. der sozialen Masse zu gewinnen. [...]

Am vollkommensten wird diese Erforschung der sozialen Masse durch die erschöpfende Massenbeobachtung ihrer Elemente in Zahl und Maß bewerkstelligt. Die so geartete wissenschaftliche Erforschung der sozialen Masse nennen wir *Statistik*. Die Statistik ist hiernach recht eigentlich die Wissenschaft von den sozialen Massen." (v. Mayr 1895, S. 5)

In diesem Zitat wird auch deutlich, daß v. Mayr noch die Auffassung vertreten hat, daß nach Möglichkeit Vollerhebungen durchgeführt werden sollten (etwa so wie heute noch bei Volkszählungen), um statistische Bilder für Gesamtheiten von Dingen, Menschen oder Situationen zu gewinnen.

4. Auf die Geschichte der sozialwissenschaftlichen Statistik werden wir in diesem Text nicht näher eingehen. An dieser Stelle soll zunächst nur festgehalten werden, daß sich statistische Begriffsbildungen und Überlegungen stets auf Gesamtheiten (von Dingen, Menschen, Situationen, ...) beziehen.[4] Auch auf die Frage, wie man sich Informationen (Daten) über die Mitglieder solcher Gesamtheiten verschaffen kann, soll zunächst nicht näher eingegangen werden. Sie ist natürlich wichtig, und wir werden uns später mit einigen Aspekten dieser Frage genauer beschäftigen. Aber bevor man das tun kann, benötigt man begriffliche Hilfsmittel, um sinnvoll über Gesamtheiten sprechen zu können; denn die primäre Motivation verdankt sich ja gerade der Tatsache, daß es sich um Gesamtheiten handelt, die nicht unmittelbar überschaubar und persönlich erfahrbar sind. Insofern impliziert bereits das Reden von Gesamtheiten gedankliche Konstruktionen.

1.2 Zur Konzeption von Gesamtheiten

1. Denken wir an ein Beispiel: wir möchten etwas über die Lebenssituationen der gegenwärtig in Deutschland arbeitslosen Menschen herausfinden. Die Fragestellung setzt voraus, daß man sich sinnvoll auf eine durch das Merkmal 'Arbeitslosigkeit' charakterisierbare Gesamtheit von Menschen

[4] R. A. Fisher, ein wichtiger Wegbereiter der modernen Statistik, bemerkte hierzu: „Nevertheless, in a real sense, statistics is the study of populations, or aggregates of individuals, rather than of individuals." (Fisher 1970, S. 1f)

beziehen kann. Aber niemand kann eine solche Gesamtheit von Menschen sehen. Sicherlich kennen wir einzelne Mitglieder; mit ihnen kann man reden und dadurch Kenntnisse über ihre gegenwärtige Lebenssituation gewinnen. Aber die Gesamtheit aller gegenwärtig arbeitslosen Menschen ist weder für die Anschauung noch für die Kommunikation unmittelbar gegeben (im Unterschied z.B. zu einem Seminar, bei dem es nur wenige Teilnehmer gibt); es handelt sich vielmehr um eine gedankliche Konstruktion, die sich dem Wunsch verdankt, den engen Umkreis persönlicher Erfahrungen durch eine gedankliche Bezugnahme auf „gesellschaftliche Verhältnisse“ zu erweitern.[5]

2. An dieser Idee einer gedanklichen Konstruktion von Gesamtheiten setzt die Begriffsbildung an. Leitend ist die Vorstellung, daß man sich ein symbolisches Bild für eine Gesamtheit, auf die man sich gedanklich beziehen möchte, machen kann. Wir verwenden die Schreibweise:[6]

$$\Omega := \{\omega_1, \ldots, \omega_n\}$$

Ω ist der symbolische Name für die Gesamtheit, die aus den Mitgliedern $\omega_1, \ldots, \omega_n$ besteht. Die geschweiften Klammern deuten an, daß man diese Mitglieder gedanklich zu einer Menge zusammenfassen möchte. Die Auslassungspunkte sollen die Vorstellung andeuten, daß man die Mitglieder der Menge aufzählen kann:

$$\omega_1,\ \omega_2,\ \omega_3,\ \omega_4,\ \omega_5,\ \ldots$$

und infolgedessen auch fragen kann, wieviele Mitglieder die Gesamtheit enthält. Das wird in der oben verwendeten Schreibweise durch den Index n angedeutet, der auf irgendeine bestimmte natürliche Zahl verweist; z.B. auf die Zahl 5, wenn die Gesamtheit 5 Mitglieder umfaßt, oder auf die Zahl 500, wenn es 500 Mitglieder gibt. Es sei angemerkt, daß die Vorstellung, daß man die Mitglieder einer Gesamtheit „aufzählen“ kann, hier nicht voraussetzt, daß sie bereits in irgendeiner Weise geordnet sind (etwa so, wie man sich die natürlichen Zahlen nach ihrer Größe geordnet vorstellen kann). Den Begriff einer statistischen Gesamtheit verstehen wir so, wie in der Mengenlehre von Mengen gesprochen wird, also ohne irgendeine Art von Ordnung zu unterstellen. Z.B. handelt es sich bei $\{1, 2, 3\}$ und

[5] Sich gesellschaftliche Verhältnisse einfach als Mengen von Menschen vorzustellen, ist natürlich eine ziemlich verkürzte Sichtweise. Aber sie ist für die Statistik charakteristisch und liefert immerhin einen ersten begrifflichen Ansatz. In Teil III werden wir besprechen, wie man diesen Ansatz zu einer dynamischen Betrachtungsweise erweitern kann.

[6] Wir unterscheiden in diesem Text die Zeichen ‘=’ und ‘:=’. Ein Gleichheitszeichen mit vorangestelltem Doppelpunkt wird verwendet, um anzudeuten, daß eine definitorische Gleichsetzung vorgenommen wird, d.h. der Ausdruck auf der linken Seite wird durch den Ausdruck auf der rechten Seite definiert. Dagegen setzt ein einfaches Gleichheitszeichen voraus, daß beide Seiten schon definiert sind.

$\{1, 3, 2\}$ nur um unterschiedliche Darstellungen der gleichen Menge, die aus den Elementen 1, 2 und 3 besteht.

3. Wieviele Mitglieder kann eine solche Gesamtheit umfassen? Für die allgemeinen Begriffsbildungen muß diese Frage nicht unbedingt beantwortet werden. Natürlich, mindestens 1 Mitglied, aber es können auch 10 oder 1000 oder noch viel mehr sein, im Prinzip „beliebig viele“. D.h. die Begriffsbildung verlangt nicht, irgendeine bestimmte obere Grenze anzunehmen. Sie setzt auch nicht voraus, daß man die genaue Anzahl der Mitglieder kennt, sondern nur die Annahme, daß die Gesamtheit tatsächlich aus einer bestimmten Anzahl von Mitgliedern besteht.[7] Weiterhin setzt die Begriffsbildung nicht voraus, daß eine Gesamtheit „viele“ Mitglieder haben muß. Aus der Perspektive der abstrakten Begriffsbildung genügt es, daß sie mindestens ein Mitglied hat. Hier wird aber das eingangs genannte Motiv wichtig: daß die Verwendung statistischer Begriffsbildungen dann sinnvoll wird, wenn es sich um Gesamtheiten handelt, die nicht mehr unmittelbar überschaubar sind. Aber ‘nicht überschaubar’ kann nicht durch irgendeine bestimmte Mindestgröße für eine Gesamtheit definiert werden.

4. Es bleibt die Frage, wie die Symbole $\omega_1, \ldots, \omega_n$ für die Mitglieder der Gesamtheit zu verstehen sind. Wofür sie verwendet werden sollen, erscheint einigermaßen klar: zum Verweis auf die *Objekte*, die als Mitglieder einer Gesamtheit betrachtet werden sollen. Natürlich muß in jedem Einzelfall geklärt werden, um welche Art von Objekten es sich handelt. Wir haben bisher von Dingen, Menschen oder Situationen gesprochen, um dadurch anzudeuten, daß es sich um unterschiedliche Arten von Objekten handeln kann. Aber auch diese Aufzählung erschöpft nicht alle Möglichkeiten; man kann z.B. auch an Haushalte oder Schulklassen oder Unternehmen denken. Das Wort ‘Objekt’ verwenden wir als unspezifischen Oberbegriff. Somit kann man die Symbole $\omega_1, \ldots, \omega_n$ als *fiktive Namen* für die Objekte auffassen, die zu einer Gesamtheit zusammengefaßt werden sollen.[8] Fiktiv deshalb, weil nicht vorausgesetzt wird, daß man die korrespondierenden Objekte wirklich kennt. Nicht einmal wird vorausgesetzt, daß man weiß, wie man zu jedem Namen ein korrespondierendes Objekt in der empirischen Realität tatsächlich finden kann. Diese fiktiven Namen haben nur die Aufgabe, gedankliche Operationen zu unterstützen: sie dienen zur Unter-

[7] Diese Anzahl zu ermitteln, ist eine empirische Aufgabe, die sich z.B. dann stellt, wenn man herausfinden möchte, wieviele arbeitslose Menschen es gegenwärtig gibt. Allerdings ist es meistens nicht möglich, den Umfang einer gedanklich konzipierten Gesamtheit empirisch genau zu ermitteln. Dann stellt sich die Frage, wie man eine sinnvolle Schätzung finden kann.

[8] Zwischen Namen und Objekten zu unterscheiden, ist natürlich wichtig. Wenn dies jedoch klargeworden ist, kann man sich auch verkürzter Redeweisen bedienen; z.B. von einer Gesamtheit Ω sprechen, deren Elemente Menschen sind. Jeder versteht dann, daß die Elemente von Ω tatsächlich (fiktive) Namen sind, mit denen auf Menschen verwiesen werden soll.

scheidung und Identifizierung der Mitglieder der Gesamtheit. Die Frage, wie man korrespondierende Objekte in der empirischen Realität finden kann, stellt sich erst, wenn man über die Mitglieder der gedanklich konstruierten Gesamtheit Kenntnisse (Daten) gewinnen möchte. Wir werden jedoch in Kapitel 3 sehen, wie diese Frage dadurch eine eigentümliche Veränderung erfährt, daß man in der Statistik gar nicht an individuell zurechenbaren Beschreibungen interessiert ist.

5. Noch eine weitere Frage sollte überlegt werden: *worauf* kann mit den Symbolen $\omega_1, \ldots, \omega_n$, die für die gedankliche Bildung einer Gesamtheit vorausgesetzt werden, sinnvoll verwiesen werden? Um zu einer Antwort zu gelangen, muß man sich überlegen, wofür statistische Begriffsbildungen verwendet werden sollen. In der sozialwissenschaftlichen Statistik geht es zunächst (wenn auch nicht nur) darum, gesellschaftliche Verhältnisse begrifflich zu repräsentieren und den repräsentierenden Anspruch durch empirisch gewonnene Daten auszuweisen. Somit bezieht man sich in diesem Kontext auf Objekte, die in der gesellschaftlichen Realität identifiziert werden können. Das sind in erster Linie die einzelnen Menschen, die in den gegebenen gesellschaftlichen Verhältnissen leben. Dann handelt es sich bei den Symbolen $\omega_1, \ldots, \omega_n$ um fiktive Namen für Menschen. Aber es muß sich nicht unbedingt um einzelne Menschen handeln; zum Beispiel kann man sich auch auf Haushalte oder Schulklassen oder Unternehmen beziehen, um daraus Gesamtheiten zu bilden. Wichtig ist nur, daß es sich um Objekte handelt, die in unserer Erfahrungswelt fixiert werden können; wichtig deshalb, weil man nur dadurch begründen kann, worüber mit statistischen Begriffsbildungen gesprochen werden soll.

6. Schließlich soll noch einmal betont werden, in welcher Weise es sich bei der Bildung von Gesamtheiten – als Ausgangspunkt für sich anschließende statistische Begriffsbildungen – um gedankliche Konstruktionen handelt. Wie unsere Erläuterungen deutlich gemacht haben sollten, ist damit nicht gemeint, daß es sich bei den Objekten, die als Mitglieder einer Gesamtheit vorstellbar gemacht werden sollen, um fiktive Objekte handelt. Im Gegenteil, es wird angenommen, daß es die Mitglieder statistischer Gesamtheiten in unserer Erfahrungswelt tatsächlich gibt oder gegeben hat. Aber es handelt sich um eine gedankliche Konstruktion, weil nicht vorausgesetzt wird, daß man zunächst eine bestimmte Anzahl von Objekten tatsächlich empirisch identifiziert hat, um dann hinterher aus diesen Objekten eine Gesamtheit zu bilden. Somit sollten zwei Fragen deutlich unterschieden werden. Einerseits die Frage, wie man sinnvolle Gesamtheiten gedanklich konzipieren kann; und andererseits die Frage, wie man über alle oder einen Teil der Mitglieder einer vorab konzipierten Gesamtheit Kenntnisse (Informationen, Daten) gewinnen kann. Im nächsten Kapitel werden wir darauf etwas genauer eingehen.

1.3 Zur Verwendung symbolischer Schreibweisen

Um Überlegungen zur Bedeutung statistischer Begriffsbildungen übersichtlich und geordnet darzustellen, ist es hilfreich, einige symbolische Schreibweisen zu verwenden. Als ein Beispiel kann man an das Symbol Ω denken, das im vorangegangenen Abschnitt zum Verweis auf statistische Gesamtheiten eingeführt wurde. Die meisten symbolischen Schreibweisen, die wir in diesem Text verwenden werden, stammen aus der Mengenlehre, was verständlich ist, weil sich statistische Gesamtheiten als Mengen auffassen lassen. In den folgenden zwei Abschnitten werden die hauptsächlich verwendeten Schreibweisen kurz erläutert.

1.3.1 Notationen aus der Mengenlehre

1. Als ein Grundbegriff dient das Wort 'Menge' im Sinne einer Gesamtheit von Elementen. Zur Erläuterung verwenden wir hier Großbuchstaben für Mengen und Kleinbuchstaben für Elemente; z.B. $A := \{a_1, a_2, a_3\}$, um eine Menge mit dem Namen A zu definieren, die aus den drei Elementen a_1, a_2 und a_3 besteht. Dieser Konvention werden wir, soweit es möglich ist, im gesamten Text folgen.[9]

2. Um von einem Objekt zu sagen, daß es Element einer Menge ist, wird das Zeichen $\in$ verwendet. Z.B. könnte man sagen: $a \in A$; dann ist gemeint, daß a ein (irgendein) Element der Menge A ist, und aus der vorangegangenen Definition von A folgt, daß a entweder gleich a_1 oder gleich a_2 oder gleich a_3 ist. Entsprechend wird das Zeichen $\notin$ verwendet, um zu sagen, daß etwas kein Element einer Menge ist oder sein soll. Zwei Mengen werden als gleich angesehen, wenn jedes Element der einen auch ein Element der anderen Menge ist, und umgekehrt. Der Begriff einer Menge impliziert also nicht, daß es irgendeine Art von Ordnung für ihre Elemente gibt; z.B. gibt es im Sinne der Gleichheit von Mengen keinen Unterschied zwischen $\{a_2, a_1, a_3\}$ und der oben angegebenen Menge A.

3. Gelegentlich kommt es jedoch auch auf die Reihenfolge an; dann werden runde Klammern verwendet, z.B. in der Form

$$(a_1, a_2, a_3)$$

In diesem Beispiel werden drei Elemente zu einer Gesamtheit zusammengefaßt, bei der es auf die Reihenfolge ankommt, d.h. es ist z.B.

$$(a_1, a_2, a_3) \neq (a_2, a_1, a_3)$$

Enthält eine solche Gesamtheit zwei Elemente, spricht man von einem

[9] Vollständig konsequent kann man sich nicht an diese Konvention halten, weil Elemente von Mengen selbst wieder Mengen sein können.

Paar, bei drei Elementen von einem *Tripel*. Allgemein wird eine geordnete Gesamtheit

$$(a_1, \ldots, a_n)$$

die aus n Elementen besteht, ein *n-Tupel* genannt.

4. Hat man eine Menge eingeführt, kann man aus ihr neue Mengen bilden. Hat man z.B. bereits eine Menge B eingeführt, kann man daraus mit der folgenden Formulierung eine neue Menge bilden:

$$C := \{b \in B \,|\, \text{für } b \text{ gilt die Eigenschaft } \ldots\}$$

Es wird hierdurch eine neue Menge mit dem Namen C gebildet, die aus allen Elementen von B besteht, für die die hinter dem senkrechten Bedingungsstrich angegebene Eigenschaft zutrifft. Die neue Menge C ist infolgedessen eine *Teilmenge* der Menge B, wofür man auch schreibt: $C \subseteq B$. Mit dieser Schreibweise ist gemeint: jedes Element von C ist auch ein Element von B. Die Definition impliziert, daß auch die Aussage $B \subseteq B$ richtig ist. Manchmal möchte man diesen Fall ausschließen und sich nur auf *echte Teilmengen* beziehen; dafür wird die Schreibweise $C \subset B$ verwendet. Sie besagt: C ist eine Teilmenge von B und nicht mit B identisch.

5. Hat man zwei Mengen, kann man aus ihnen auch mit den Operationen 'Vereinigung' und 'Durchschnitt' neue Mengen bilden. Hat man etwa bereits Mengen A und B definiert, kann man daraus ihre *Vereinigungsmenge* $A \cup B$ bilden. Sie besteht aus allen Objekten, die Element von A *oder* Element von B sind (wobei hier ein nicht-ausschließendes 'oder' gemeint ist). Analog kann man die *Durchschnittsmenge* (oder kurz: den *Durchschnitt*) von A und B bilden. Dafür wird die Schreibweise $A \cap B$ verwendet. Diese Menge besteht aus allen Objekten, die sowohl Element von A als auch Element von B sind. Hierbei kann es natürlich vorkommen, daß es überhaupt kein Objekt gibt, das sowohl in der einen als auch in der anderen Menge enthalten ist. Man nennt die beiden Mengen dann *disjunkt*. Um trotzdem davon ausgehen zu können, daß in jedem Fall eine neue Menge entsteht, wird der Begriff einer *leeren Menge* eingeführt. Um auf sie zu verweisen, dient das Symbol $\emptyset$. Somit kann man sagen: zwei Mengen A und B sind genau dann disjunkt, wenn $A \cap B = \emptyset$ ist.

6. Für die Verknüpfungen 'Vereinigung' und 'Durchschnitt' gelten einige einfache Rechenregeln. Zunächst ist evident, daß die Verknüpfungen kommutativ sind:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Weiterhin gibt es zwei Distributivgesetze:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

7. Hier schließt der Begriff einer *Partition* an, den wir oft verwenden werden. Ist eine Menge A gegeben, besteht eine Partition von A aus einer Menge von Teilmengen von A, etwa aus den Mengen $A_1, \ldots, A_m$, so daß folgende Bedingungen erfüllt sind: die Mengen $A_1, \ldots, A_m$ sind paarweise disjunkt und ihre Vereinigung ist mit der Menge A identisch. Ist z.B. $A := \{a_1, a_2, a_3\}$, dann wäre die Menge

$$\{\{a_1\}, \{a_2, a_3\}\}$$

eine Partition von A. Partitionen sind also Mengen, deren Elemente wiederum Mengen sind. Es ist auch offensichtlich, daß es im allgemeinen viele unterschiedliche Partitionen einer Menge geben kann.

8. Weiterhin wird oft der Begriff einer *Potenzmenge* verwendet. Ist eine Menge A gegeben, versteht man unter ihrer Potenzmenge die Menge aller ihrer Teilmengen. Als Schreibweise wird $\mathcal{P}(A)$ verwendet, um auf die Potenzmenge von A zu verweisen. Man beachte, daß insbesondere die leere Menge $\emptyset$ und die Menge A selbst Elemente von $\mathcal{P}(A)$ sind. Ist z.B. wieder $A := \{a_1, a_2, a_3\}$, findet man:

$$\mathcal{P}(A) = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}\}$$

9. Ebenfalls sehr oft wird der Begriff eines *kartesischen Produkts* von zwei oder mehr Mengen verwendet. Zur Erläuterung soll ein kleines Zahlenbeispiel dienen. Es seien zwei Mengen

$$A := \{1, 2\} \quad \text{und} \quad B := \{3, 4, 5\}$$

gegeben. Dann besteht das kartesische Produkt von A und B (geschrieben: $A \times B$) aus der Menge aller geordneten Paare, die man durch Kombination der Elemente von A und B bilden kann. In unserem Beispiel:

$$A \times B = \{(1,3), (1,4), (1,5), (2,3), (2,4), (2,5)\}$$

Diese Begriffsbildung ist sehr allgemein; z.B. kann man auch das kartesische Produkt von drei (im Prinzip beliebig vielen) Mengen bilden. Angenommen, man hat noch eine dritte Menge, die nur aus einem Element besteht, etwa $C := \{6\}$, dann findet man:

$$A \times B \times C = \{(1,3,6), (1,4,6), (1,5,6), (2,3,6), (2,4,6), (2,5,6)\}$$

Man kann auch das kartesische Produkt einer Menge mit sich selbst bilden; zum Beispiel:

$$A \times A \times A = \{(1,1,1),(1,1,2),(1,2,1),(1,2,2),\\ (2,1,1),(2,1,2),(2,2,1),(2,2,2)\}$$

Wenn man das kartesische Produkt einer Menge mit sich selbst bildet, wird oft eine abkürzende Schreibweise verwendet:

$$A^n := \underbrace{A \times \cdots \times A}_{n\text{-mal}}$$

Weiterhin wird folgende Konvention verwendet:

$$A \times \emptyset = \emptyset \times A = \emptyset$$

wobei A eine beliebige Menge ist.

10. Für kartesische Produkte gelten die folgenden Distributivgesetze:

$$A \times (B \cup C) = (A \times B) \cup (A \times C)$$
$$A \times (B \cap C) = (A \times B) \cap (A \times C)$$

Man beachte jedoch, daß die kartesische Produktbildung im allgemeinen nicht kommutativ ist, d.h. im allgemeinen führt $B \times A$ zu einer anderen Menge als $A \times B$. In unserem Beispiel:

$$B \times A = \{(3,1),(4,1),(5,1),(3,2),(4,2),(5,2)\}$$

Der Unterschied entsteht daraus, daß die Elemente eines kartesischen Produkts *geordnete* Paare (oder n-Tupel) der Elemente der Ursprungsmengen sind.

11. Die meisten Mengen, mit denen wir uns beschäftigen werden, sind endlich, d.h. haben nur eine endliche Anzahl von Elementen. Insbesondere beschäftigen wir uns nur mit endlichen statistischen Gesamtheiten. Ist A eine endliche Menge, verwenden wir die Schreibweise $|\,A\,|$ für die Anzahl ihrer Elemente. Ist z.B. $A := \{a_1, a_2, a_3\}$, dann ist $|\,A\,| = 3$. Als Konvention wird vereinbart: $|\,\emptyset\,| = 0$.

1.3.2 Erläuterungen zum Funktionsbegriff

1. Von grundlegender Bedeutung für die Statistik (wie auch für viele andere Gebiete) ist der Funktionsbegriff. Wir verwenden diesen Begriff so, wie er in der Mathematik verwendet wird, und beziehen ihn auf eine vorgängige Einführung von Mengen. Wenn zwei Mengen A und B gegeben sind, ist eine *Funktion* (auch *Abbildung* genannt) eine Regel, durch die jedem

Element $a \in A$ genau ein Element $b \in B$ zugeordnet wird. Wir verwenden die Schreibweise

$$f : A \longrightarrow B$$

f ist der Name der Funktion (wofür auch beliebige andere Buchstaben und Symbole verwendet werden können); A wird *Definitionsbereich* und B wird *Wertebereich* der Funktion genannt. Ist $a \in A$ ein Element aus dem Definitionsbereich der Funktion f, wird mit $f(a)$ dasjenige Element aus dem Wertebereich B bezeichnet, das dem Element a durch die Funktion f zugeordnet wird. In dieser Schreibweise wird a als ein *Argument* der Funktion verwendet, was durch runde Klammern kenntlich gemacht wird.

2. Zur Illustration betrachten wir Mengen $A := \{1,2\}$ und $B := \{3,4,5\}$. Eine Funktion $f : A \longrightarrow B$ könnte z.B. durch folgende Festlegung eingeführt werden: $f(1) = 3, f(2) = 4$. Hier sollte man sich überlegen, wann zwei Funktionen als gleich angesehen werden können. Wir verwenden folgende Vereinbarung: Zwei Funktionen $f : A \longrightarrow B$ und $g : C \longrightarrow D$ werden als gleich angesehen, wenn gilt: $A = C$, $B = D$ und $f(a) = g(a)$ für alle $a \in A$. Würde man z.B. eine zweite Funktion

$$g : \{1,2\} \longrightarrow \{3,4\}$$

einführen, wobei $g(1) = 3$ und $g(2) = 4$ ist, wäre sie von der oben als Beispiel verwendeten Funktion f verschieden.

3. Ist eine Funktion $f : A \longrightarrow B$ eingeführt worden, kann man als Argumente zunächst Elemente ihres Definitionsbereichs verwenden, also z.B. den Ausdruck $f(a)$ verwenden, wobei a ein Element des Definitionsbereichs A der Funktion ist. Es ist jedoch oft zweckmäßig, als Argumente auch Teilmengen des Definitionsbereichs zuzulassen. Dies bedeutet, daß f als eine *Mengenfunktion*

$$f : \mathcal{P}(A) \longrightarrow \mathcal{P}(B)$$

verwendet wird, die jeder Teilmenge $C \subseteq A$ eine Teilmenge

$$f(C) := \{b \in B \,|\, \text{es gibt ein } a \in C \text{ mit } f(a) = b\}$$

im Wertebereich von f zuordnet. Gleichbedeutend ist die Schreibweise

$$f(C) = \{f(a) \,|\, a \in C\}$$

Insbesondere ist auch $f(A)$ eine Teilmenge des Wertebereichs von f und wird das *Bild von A unter der Funktion f* oder auch *Bildmenge von f* genannt. Offenbar gilt stets: $f(A) \subseteq B$; wie jedoch das oben angeführte Beispiel zeigt, ist es durchaus möglich, daß $f(A) \neq B$ ist.

4. Faßt man eine Funktion $f : A \longrightarrow B$ als eine Mengenfunktion auf, kann auch stets eine inverse Funktion gebildet werden. Wir verwenden folgende Definition: Die zu f *inverse Mengenfunktion* ist die Funktion

$$f^{-1} : \mathcal{P}(B) \longrightarrow \mathcal{P}(A)$$

die jeder Teilmenge des Wertebereichs von f eine Teilmenge aus dem Definitionsbereich von f zuordnet, und zwar nach folgender Vorschrift:

$$f^{-1}(C) := \{a \in A \,|\, f(a) \in C\}$$

wobei C ein beliebiges Element von $\mathcal{P}(B)$, also eine beliebige Teilmenge von B ist. $f^{-1}(C)$ wird auch das *Urbild* von C (bzgl. f) genannt. Ist z.B. eine Funktion $f : \{1, 2\} \longrightarrow \{3, 4, 5\}$ durch $f(1) = 3$ und $f(2) = 4$ gegeben, findet man für die Teilmengen des Wertebereichs $\{3, 4, 5\}$:

$$f^{-1}(\{3\}) = \{1\},\ f^{-1}(\{4\}) = \{2\},\ f^{-1}(\{5\}) = \emptyset,$$

$$f^{-1}(\{3, 4\}) = \{1, 2\},\ f^{-1}(\{3, 5\}) = \{1\},\ f^{-1}(\{4, 5\}) = \{2\},$$

$$f^{-1}(\{3, 4, 5\}) = \{1, 2\},\ f^{-1}(\emptyset) = \emptyset$$

Es gelten folgende Rechenregeln:

$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D)$$

$$f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D)$$

wobei C und D beliebige Teilmengen von B sind.

5. Es sollte deutlich geworden sein, daß der hier verwendete mathematische Funktionsbegriff sich grundsätzlich von Redeweisen unterscheidet, in denen von 'Funktion' im Sinne von 'Zweck' gesprochen wird. Es bleiben natürlich Fragen übrig. Zunächst kann man sich fragen, wie Funktionen zustande kommen. Die allgemeine Antwort auf diese Frage ist, daß Funktionen durch Menschen gemacht werden. Es sind Menschen, die Mengen konzipieren und Zuordnungen zwischen ihren Elementen vornehmen und diese Zuordnungen als Funktionen darstellen. Funktionen sind keine empirischen Sachverhalte, die in unserer Erfahrungswelt wahrgenommen werden können. Dennoch gibt es einen wichtigen Unterschied zwischen Mathematik und Statistik. In der Mathematik kann man Mengen und Funktionen ohne Rücksicht auf empirische Sachverhalte konstruieren. Statistische Begriffsbildungen sollen aber helfen, empirisch gewonnenes Wissen darstellbar und reflektierbar zu machen. Wenn im Rahmen der Statistik Mengen und Funktionen konstruiert werden, sind deshalb nicht allein ihre formalen Eigenschaften, sondern in erster Linie die jeweils intendierten Bedeutungen wichtig. Allerdings liefern die formalen Implikationen von Begriffsbildungen oft wichtige Hinweise oder zumindest einen Ausgangspunkt, um auch genauer über mögliche Bedeutungen nachdenken zu können.

1.4 Statistische Variablen

Wir knüpfen jetzt an den Begriff einer Gesamtheit an und besprechen den für alles weitere grundlegenden Begriff einer statistischen Variablen.

1.4.1 Zur Konzeption von Merkmalsräumen

1. Hat man eine Gesamtheit Ω gedanklich konzipiert, hat man dadurch einen Ausgangspunkt, um über Eigenschaften der Mitglieder von Ω nachzudenken. Um dies zu präzisieren, dient der Begriff 'statistische Variable'. Die Idee ist, daß durch eine statistische Variable den Mitgliedern einer Gesamtheit Eigenschaften zugeordnet werden. Um statistische Variablen zu definieren, muß man sich also zunächst überlegen, welche Eigenschaften man bei den Mitgliedern einer Gesamtheit sinnvoll feststellen kann. Wenn es sich um eine Gesamtheit von Menschen handelt, kann man sie z.B. danach unterscheiden, ob es sich um Männer oder Frauen handelt; dagegen macht diese Unterscheidung bei Schulklassen oder Unternehmen keinen Sinn. Statt von 'Eigenschaften' spricht man in der Statistik auch von 'Merkmalen'. Dem liegt die Vorstellung zugrunde, daß man die Mitglieder einer Gesamtheit durch individuell zurechenbare Merkmale charakterisieren kann. Hierbei bedeutet 'individuell zurechenbar' jedoch nur, daß man sich bei der Zurechnung von Merkmalen auf die individuellen Mitglieder einer Gesamtheit beziehen kann. Z.B. kann man die gegenwärtig in Deutschland arbeitslosen Menschen danach unterscheiden, in welchem Bundesland sie leben; somit ist z.B. 'lebt in Hamburg' eine individuell zurechenbare Eigenschaft, obwohl sie natürlich vielen Menschen gleichermaßen zukommt.

2. Eine Menge von Eigenschaften, die man den Mitgliedern einer Gesamtheit zurechnen kann, nennen wir einen *Eigenschafts-* oder auch *Merkmalsraum*. Als Symbole für Merkmalsräume verwenden wir kalligraphische Großbuchstaben, die mit einer Tilde versehen sind, z.B. $\tilde{\mathcal{X}}$. Zum Verweis auf Elemente eines Merkmalsraums, die auch *Merkmalsausprägungen* oder *Merkmalswerte* genannt werden, verwenden wir korrespondierende lateinische Kleinbuchstaben, die ebenfalls mit einer Tilde gekennzeichnet werden. Die allgemeine Schreibweise, um sich auf einen Merkmalsraum zu beziehen, ist also

$$\tilde{\mathcal{X}} := \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots\}$$

In konkreten Anwendungsfällen ist es natürlich erforderlich, die Bedeutung der Symbole $\tilde{x}_1, \tilde{x}_2, \ldots$ zu erklären. Die Auslassungspunkte sollen andeuten, daß ein Merkmalsraum beliebig viele Merkmalsausprägungen umfassen kann; es wird also nicht vorausgesetzt, daß Merkmalsräume stets endliche Mengen sind.

3. An einen Merkmalsraum, der zur Definition statistischer Variablen verwendet werden kann, werden zwei Anforderungen gestellt:

a) Die Merkmalswerte (Eigenschaften), die in einem Merkmalsraum zusammengefaßt werden, müssen sich wechselseitig ausschließen. Als Beispiel kann man den Merkmalsraum

$$\tilde{\mathcal{X}} := \{\,\text{‘männlich’}, \text{‘weiblich’}\,\}$$

betrachten, dessen Elemente sich ausschließen. Würde man die Eigenschaft ‘verheiratet’ hinzufügen, wäre die Bedingung, daß sich die Merkmalswerte wechselseitig ausschließen, nicht mehr erfüllt.

b) Jedem Mitglied einer Gesamtheit muß genau ein Merkmalswert aus dem Merkmalsraum zurechenbar sein.

Hier zeigt sich erneut, daß man bereits bei der Bildung von Merkmalsräumen auf Gesamtheiten Bezug nehmen muß. Wir fassen deshalb die zuvor genannten Bedingungen zu folgender Definition zusammen: Ein Merkmalsraum $\tilde{\mathcal{X}}$ heißt *konsistent bzgl. einer Gesamtheit* Ω, wenn sich die Merkmalswerte in $\tilde{\mathcal{X}}$ wechselseitig ausschließen und jedem Mitglied von Ω genau ein Merkmalswert aus $\tilde{\mathcal{X}}$ zugewiesen werden kann. Damit ein Merkmalsraum sinnvoll für empirische Untersuchungen verwendet werden kann, wird man allerdings noch etwas weiteres fordern: Daß man bei den Mitgliedern einer Gesamtheit Ω, für die man sich einen konsistenten Merkmalsraum $\tilde{\mathcal{X}}$ gemacht hat, auch *feststellen* kann, welche Eigenschaft aus $\tilde{\mathcal{X}}$ ihnen jeweils zukommt.

4. Man beachte, daß bei unserer Definition eines Merkmalsraums $\tilde{\mathcal{X}}$ nicht gefordert wird, daß es für jede Eigenschaft in $\tilde{\mathcal{X}}$ auch mindestens ein Mitglied von Ω gibt, dem diese Eigenschaft zukommt. Als Beispiel kann man daran denken, daß Ω auf eine Gesamtheit von Menschen verweist und daß man u.a. über ihr Alter sprechen möchte. Dann benötigt man einen Merkmalsraum, dessen Elemente sinnvoll zurechenbare Altersangaben sind. Man kann z.B. ‘vollendete Lebensjahre’ verwenden und folgenden Merkmalsraum definieren:

$$\tilde{\mathcal{Y}} := \{0, 1, 2, \ldots, 200\}$$

wobei die Zahlen im Merkmalsraum sinngemäß für eine entsprechende Anzahl vollendeter Lebensjahre stehen. Wenn man sich dann auf eine reale Gesamtheit von Menschen bezieht, wird sie vermutlich niemanden enthalten, der 200 Jahre alt geworden ist.

1.4.2 Definition des Variablenbegriffs

1. Wenn eine Gesamtheit Ω und ein für sie konsistenter Merkmalsraum $\tilde{\mathcal{X}}$ gegeben sind, besteht eine statistische Variable in einer Funktion, die

jedem Mitglied aus Ω ihren Merkmalswert in $\tilde{\mathcal{X}}$ zuordnet. Wir verwenden dafür die symbolische Schreibweise

$$X : \Omega \longrightarrow \tilde{\mathcal{X}}$$

Hier ist X der symbolische Name der statistischen Variablen, und der Pfeil soll andeuten, daß X eine Funktion ist, die jedem Mitglied aus Ω einen bestimmten Wert im Merkmalsraum $\tilde{\mathcal{X}}$ zuordnet. Wenn also ω irgendein Mitglied von Ω ist, dann ist $X(\omega)$ ein Element von $\tilde{\mathcal{X}}$, und zwar derjenige Merkmalswert, der dem Objekt ω durch die statistische Variable X zugeordnet wird. Handelt es sich z.B. um eine Menge von Menschen und ist $\tilde{\mathcal{X}}$ der im vorangegangenen Abschnitt definierte Merkmalsraum für das Geschlecht, hätte man:

$$X(\omega) = \begin{cases} \text{‘männlich’}, & \text{wenn } \omega \text{ männlich ist, und} \\ \text{‘weiblich’}, & \text{wenn } \omega \text{ weiblich ist.} \end{cases}$$

2. Die in Abschnitt 1.2 betonte Unterscheidung zwischen gedanklichen Konstruktionen und Daten gilt sinngemäß auch für statistische Variablen. Die Annahme, daß es für eine Gesamtheit Ω eine statistische Variable X gibt, die jedem Mitglied von Ω einen Merkmalswert in einem Merkmalsraum $\tilde{\mathcal{X}}$ zuordnet, impliziert nicht, daß man die Merkmalswerte tatsächlich kennt, also über entsprechende Daten verfügt. Trotzdem gibt es eine wesentliche Implikation. Der Begriff einer statistischen Variablen impliziert nämlich die Annahme, daß entsprechende Merkmalswerte für die Mitglieder einer Gesamtheit tatsächlich bestehen und fixiert sind. Eine solche Annahme ist oft sinnvoll. Z.B. kann man sinnvoll annehmen, daß alle gegenwärtig lebenden Menschen ein bestimmtes Alter haben; und zwar unabhängig davon, ob wir ihr Alter kennen oder nicht. Andererseits gibt es Merkmalswerte, die gegenwärtig noch nicht fixiert sind. Als Beispiel kann man daran denken, daß man Menschen u.a. durch ihr Heiratsalter charakterisieren möchte. Aber für diejenigen Menschen, die noch nicht geheiratet haben, gibt es ein solches Heiratsalter noch nicht. Man muß nicht nur bedenken, daß einige Menschen vielleicht nie heiraten werden, sondern daß – solange ein Mensch noch nicht geheiratet hat – auch noch nicht feststeht, ob und ggf. in welchem Alter er heiraten wird. Würde man in diesem Fall die Existenz einer statistischen Variablen annehmen, würde dies auch die Annahme implizieren, daß zum Zeitpunkt der gedanklichen Fixierung der statistischen Variablen bereits für jedes Mitglied der Gesamtheit feststeht, ob und ggf. in welchem Alter eine Heirat stattfinden wird.

3. Es sei schließlich noch einmal betont, daß wir statistische Variablen *als Funktionen* auffassen, durch die Mitgliedern einer Gesamtheit Eigenschaften zugerechnet werden. Sie dürfen also nicht mit logischen Variablen verwechselt werden, wie sie z.B. in der Arithmetik und Algebra verwendet werden. Wenn man z.B. sagt: „für alle reellen Zahlen x gilt: wenn $x > 1$,

dann ist $x^2 > x$“, dann ist x eine logische Variable, für die man beliebige reelle Zahlen einsetzen darf, ohne daß die Aussage falsch wird. Eine logische Variable ist also durch eine Menge möglicher Werte charakterisierbar; im eben angeführten Beispiel die Menge aller reellen Zahlen. Auch in der sozialwissenschaftlichen Methodenliteratur wird der Variablenbegriff oft in Analogie zu logischen Variablen verwendet. Das kommt darin zum Ausdruck, daß viele Autoren glauben, Variablen könnten allein durch die Angabe eines Merkmalsraums definiert werden, so daß dann ein Sprachgebrauch entsteht, bei dem die Worte ‘Variable’ und ‘Merkmalsraum’ weitgehend synonym verwendet werden. Zur Begründung statistischer Grundbegriffe erscheint es uns jedoch sinnvoller, den Begriff ‘statistische Variable’ von Anfang an so zu konzipieren, daß sichtbar wird, worüber man sprechen möchte: nämlich über Eigenschaften von in unserer Erfahrungswelt vorhandenen oder vorstellbaren Dingen, Menschen oder Situationen. Es wird dann deutlich, daß man nicht nur Merkmalsräume benötigt, sondern zunächst und vor allem Mengen von Objekten, über die man etwas aussagen möchte. Unser Begriff statistischer Variablen stellt die Verbindung her. Wenn aus dem Kontext hervorgeht, was gemeint ist, kann man natürlich das Attribut ‘statistisch’ weglassen und einfach von ‘Variablen’ sprechen.

1.4.3 Mehrdimensionale Variablen

1. Wie im vorangegangenen Abschnitt erklärt worden ist, verlangt die Definition einer statistischen Variablen X, daß man zunächst sowohl eine Gesamtheit Ω als auch einen Merkmalsraum $\tilde{\mathcal{X}}$ konzipiert. Dann kann man die Variable X als eine Funktion definieren, die jedem Mitglied von Ω einen bestimmten Merkmalswert in $\tilde{\mathcal{X}}$ zuordnet. Es spricht natürlich nichts dagegen, gleichzeitig mehrere Merkmalsräume zu verwenden. Wenn sich Ω auf eine Gesamtheit von Menschen bezieht, kann man z.B. sowohl einen Merkmalsraum $\tilde{\mathcal{X}}$ für ihr Geschlecht als auch einen Merkmalsraum $\tilde{\mathcal{Y}}$ für ihr Alter konzipieren; und im Anschluß daran kann man eine statistische Variable definieren, die jedem Mitglied aus Ω sowohl ein Geschlecht als auch ein Alter zuordnet.

2. Um diese Überlegung zu präzisieren, muß man sich überlegen, wie man Merkmalsräume kombinieren kann. Dafür eignet sich der in Abschnitt 1.3.1 erläuterte Begriff des kartesischen Produkts. Als Beispiel betrachten wir die in Abschnitt 1.4.1 angeführten Merkmalsräume $\tilde{\mathcal{X}}$ (für das Geschlecht) und $\tilde{\mathcal{Y}}$ (für das Alter). Ihr kartesisches Produkt kann folgendermaßen geschrieben werden:

$$\tilde{\mathcal{X}} \times \tilde{\mathcal{Y}} = \{(\text{‘männlich’},0), (\text{‘weiblich’},0), \ldots, (\text{‘männlich’},200), (\text{‘weiblich’},200)\}$$

Offenbar kann man die resultierende Menge $\tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$ wieder als einen Merkmalsraum auffassen, dessen Elemente jetzt kombinierte Merkmalsausprägungen sind. Z.B. kann man für den neuen Merkmalsraum den Namen $\tilde{\mathcal{Z}}$ verwenden und seine Bedeutung durch

$$\tilde{\mathcal{Z}} := \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$$

festlegen. $\tilde{\mathcal{X}}$ und $\tilde{\mathcal{Y}}$ werden die *Komponenten* des kombinierten Merkmalsraums $\tilde{\mathcal{Z}}$ genannt. Da es in diesem Fall zwei Komponenten gibt, wird $\tilde{\mathcal{Z}}$ auch ein *zweidimensionaler* Merkmalsraum genannt.

3. Hat man, wie in unserem Beispiel, einen kombinierten Merkmalsraum gebildet, kann man auch eine statistische Variable definieren, die jedem Mitglied von Ω einen Wert in diesem kombinierten Merkmalsraum zuordnet. Nennen wir sie Z. Ihr Merkmalsraum $\tilde{\mathcal{Z}}$ besteht jetzt aus $\tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$, so daß man die Schreibweise

$$Z : \Omega \longrightarrow \tilde{\mathcal{Z}} := \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$$

verwenden kann. Verweist z.B. $\omega \in \Omega$ auf eine 25 jährige Frau, liefert die Variable Z den Wert

$$Z(\omega) = (\text{‘weiblich’}, 25) \in \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$$

Das Beispiel zeigt, wie man durch Kombination von Merkmalsräumen neue Merkmalsräume und, davon ausgehend, auch neue statistische Variablen definieren kann. Entsteht der neue Merkmalsraum aus einer Kombination von zwei Merkmalsräumen, wird auch die Variable *zweidimensional* genannt.[10]

4. In unserem Beispiel ist Z eine zweidimensionale Variable, die jedem Element von Ω einen Merkmalswert in $\tilde{\mathcal{Z}}$ zuordnet. Da $\tilde{\mathcal{Z}}$ aus der Kombination von zwei Merkmalsräumen, nämlich $\tilde{\mathcal{X}}$ und $\tilde{\mathcal{Y}}$, hervorgegangen ist, kann man auch sagen, daß Z den Mitgliedern von Ω jeweils zwei Merkmalswerte zuordnet: einen Merkmalswert in $\tilde{\mathcal{X}}$ und einen Merkmalswert in $\tilde{\mathcal{Y}}$. Also kann man Z auch explizit als eine zweidimensionale Funktion darstellen, indem man schreibt:

$$Z = (X, Y)$$

[10] Es sollte betont werden, daß der Begriff ‘Dimension’ hier rein formal verwendet wird, um auf die Anzahl der unterschiedlichen Merkmalsräume hinzuweisen, die man für die Definition statistischer Variablen unterscheiden möchte. Es gibt keinerlei sinnvolle Assoziationen mit räumlichen Dimensionen; auch keinen unmittelbaren Zusammenhang mit in der Methodenliteratur verbreiteten Redeweisen, in denen von „Dimensionen“ im Sinne unterschiedlicher inhaltlicher Aspekte eines Gegenstandsbereichs gesprochen wird. Würden z.B. zwei Merkmalsräume für das Lebensalter konzipiert, wobei das Lebensalter zum einen in Jahren und zum anderen in Monaten erfaßt wird, würde es sich um unterschiedliche Merkmalsräume und infolgedessen unterschiedliche „Dimensionen“ handeln.

So wie Z ist infolgedessen auch (X, Y) eine Funktion. Wendet man sie auf ein Element $\omega \in \Omega$ an, erhält man:

$$Z(\omega) \;=\; (X, Y)(\omega) \;=\; (X(\omega), Y(\omega))$$

X und Y werden *Komponenten* von Z genannt; und offenbar kann man auch diese Komponenten wiederum als statistische Variablen (Funktionen) auffasssen. Manchmal sagt man auch: die zweidimensionale Variable Z besteht aus den beiden eindimensionalen Variablen X und Y.

5. Die Begriffsbildungen können für Kombinationen aus beliebig vielen Merkmalsräumen verallgemeinert werden. Da das Alphabet bei Z zuende ist, verwendet man dann oft Indizes. Wenn man m Merkmalsräume konzipieren möchte, kann man z.B. die Namen $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_m$ verwenden und daraus einen neuen Merkmalsraum

$$\tilde{\mathcal{X}} \;:=\; \tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_m$$

bilden; $\tilde{\mathcal{X}}$ wird dann ein *m-dimensionaler Merkmalsraum* genannt. Man beachte, daß durch diese Schreibweise $\tilde{\mathcal{X}}$ neu definiert, zu einem Namen für die Kombination der Merkmalsräume $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_m$ gemacht wird. Davon ausgehend kann man eine *m-dimensionale statistische Variable* X definieren, die aus m Komponenten besteht:

$$X \;:=\; (X_1, \ldots, X_m)$$

Es ist eine Funktion, die jedem Element $\omega \in \Omega$ einen Merkmalswert im kombinierten Merkmalsraum $\tilde{\mathcal{X}}$, also gleichzeitig in allen seinen Komponenten, zuordnet. Ausführlich geschrieben:

$$X(\omega) \;=\; (X_1, \ldots, X_m)(\omega) \;=\; (X_1(\omega), \ldots, X_m(\omega))$$

Im nächsten Kapitel illustrieren wir diese Begriffsbildungen anhand einfacher Beispiele.

Kapitel 2

Variablen und Daten

In Abschnitt 1.2 wurde betont, daß es bei statistischen Begriffsbildungen zunächst um gedankliche Konstruktionen geht. Daß man auch über Daten verfügen kann, wird allerdings wichtig, sobald man mit den gedanklichen Konstruktionen einen empirischen Anspruch verbinden möchte. Dann muß man Rechenschaft darüber ablegen, wie man Daten gewonnen hat und wie sie mit den theoretischen Begriffsbildungen verknüpft werden können. Mit Fragen dieser Art beschäftigen wir uns in einem separaten Text („Methoden sozialwissenschaftlicher Datenkonstruktion“). Aber bereits um ein Verständnis statistischer Grundbegriffe zu gewinnen, muß daran gedacht werden, daß sie schließlich dazu dienen sollen, Aussagen mit empirischen Ansprüchen formulierbar und begründbar zu machen. In diesem Kapitel besprechen wir deshalb, wie man von statistischen Daten sprechen kann, und geben einige Beispiele an, die auch in späteren Kapiteln zur Illustration von Begriffsbildungen herangezogen werden.

2.1 Definition statistischer Daten

1. Von „Daten“ wird in unserer Gesellschaft sehr oft und in unterschiedlichen Kontexten gesprochen. Gelegentlich wird darauf hingewiesen, daß das Wort ‘Datum’ aus dem Lateinischen kommt und ‘das Gegebene’ bedeutet. Allerdings so wie heute meistens von „Daten“ gesprochen wird, beziehen sie sich keineswegs auf „Gegebenes“, sondern sie werden *gemacht*. Diesen Gesichtspunkt verfolgen wir in unserem Buch über „Methoden sozialwissenschaftlicher Datenkonstruktion“. Dem entspricht, wie in der Statistik von Daten gesprochen wird: *statistische Daten* sind Werte statistischer Variablen. Von statistischen Daten zu sprechen setzt also voraus, daß man zunächst eine Gesamtheit Ω und einen (meistens kombinierten) Merkmalsraum $\tilde{\mathcal{X}}$ konzipiert hat. Dies erlaubt es, eine statistische Variable X zu definieren, die jedem Mitglied der Gesamtheit Ω einen bestimmten Wert im Merkmalsraum $\tilde{\mathcal{X}}$ zuordnet. Von statistischen Daten kann man schließlich sprechen, wenn man für einige oder alle Mitglieder von Ω Werte der Variablen X tatsächlich ermittelt hat.

2. Nehmen wir an, daß es gelungen ist, Werte einer Variablen X mit dem Merkmalsraum $\tilde{\mathcal{X}}$ für alle Mitglieder einer Gesamtheit

$$\Omega \;:=\; \{\omega_1, \ldots, \omega_n\}$$

zu ermitteln. Die Daten können dann in einer *Datenmatrix* dargestellt