2
Variablen und Skalenniveaus

Im Mittelpunkt des folgenden Kapitels stehen Variablen. Es wird zunächst geklärt, was man unter einer Variablen versteht und was es bei der Messung dieser zu berücksichtigen gilt. Daraufhin werden zwei für das Buch relevante Möglichkeiten vorgestellt, wie man Variablen differenzieren kann. Dabei kommt dem Skalenniveau eine zentrale Stellung zu. Abschließend wird auf die Klassierung von Daten und die Indexbildung eingegangen.

2.1 Variablen

Das zentrale Handwerkszeug eines quantitativen Forschers sind Variablen. Unter einer *Variablen* versteht man ein Merkmal oder eine Eigenschaft von Personen,

K. Völkl und C. Korb, *Deskriptive Statistik*, Elemente der Politik, DOI 10.1007/978-3-658-10675-1_2

Ländern oder anderen Merkmalsträgern. Ein Personenmerkmal wäre z. B. das Geschlecht, die Demokratiezufriedenheit oder das Alter, ein Ländermerkmal das Wahlsystem, der Entwicklungsstand oder die Arbeitslosenquote eines Landes. Per Konvention werden Variablen durch Großbuchstaben gekennzeichnet, ggf. gefolgt von einer Zahl, z. B. X, Y, V1, V2 (vgl. Tab. 2.1).

Unter einem *Merkmalsträger* versteht man die Einheit, auf die sich die Untersuchung bezieht. Man spricht auch von Beobachtung, Fall oder Untersuchungseinheit. Im Rahmen der Politikwissenschaft werden häufig Personen oder Länder als Merkmalsträger untersucht. Mit dem Kleinbuchstaben i wird eine einzelne, aber keine bestimmte Untersuchungseinheit bezeichnet.[4]

Variablen können verschiedene Ausprägungen annehmen, die *Merkmalsausprägungen* oder *Kategorien* genannt werden. Die Kennzeichnung erfolgt mittels Kleinbuchstaben, z. B. x, y, v1, v2. Die spezifische Ausprägung einer Variablen wird abgebildet, indem der Kleinbuchstabe mit dem Code der interessierenden Ausprägung gleichgesetzt wird, z. B. $v2 = 2$. Bei der Konstruktion von Variablen gilt es zu beachten, dass die Ausprägungen disjunkt und erschöpfend sein müssen. *Disjunkt* bedeutet, dass sich die Kategorien nicht überschneiden dürfen. *Erschöpfend* heißt, dass keine Kategorie fehlen darf und

4 Für die Notation von Merkmalen, Merkmalsträgern und Merkmalsausprägungen gibt es keine verbindlichen Regeln. Welche Groß- und Kleinbuchstaben vergeben werden, ist weitgehend frei wählbar.

Tab. 2.1 Variablen und ihre Eigenschaften

Variable	Notation	Merkmals-träger	Merkmalsausprägungen		
			Wertelabel	Code	Notation
Geschlecht	V1	Personen	weiblich	1	V1 = 1
		Personen	männlich	2	V1 = 2
Demokratie-zufriedenheit	V2	Personen	sehr unzufrieden	1	V2 = 1
		Personen	ziemlich unzufrieden	2	V2 = 2
		Personen	teils/teils	3	V2 = 3
		Personen	ziemlich zufrieden	4	V2 = 4
		Personen	sehr zufrieden	5	V2 = 5
		Personen	weiß nicht	8	V2 = 8
		Personen	keine Angabe	9	V2 = 9
Alter (in Jahren)	V3	Personen	z. B. 23	23	V3 = 23
		Personen	keine Angabe	999	V3 = 999
Demokratietyp	X	Länder	parlamen-tarisch	1	X = 1
		Länder	präsidentiell	2	X = 2
		Länder	Sonstiges	3	X = 3
		Länder	keine Angabe	9	X = 9
Entwicklungs-stand	Y	Länder	niedrig	1	Y = 1
		Länder	mittel	2	Y = 2
		Länder	hoch	3	Y = 3
		Länder	keine Angabe	9	Y = 9
Arbeitslosen-quote (in Prozent)	Z	Länder	z. B. 6,1	6,1	Z = 6,1
		Länder	keine Angabe	999	Z = 999

Quelle: Eigene Darstellung.

jeder Fall einer Merkmalsausprägung zugeordnet werden können muss.

Während sich der Begriff Merkmalsausprägungen auf die theoretisch möglichen Ausprägungen einer Variablen bezieht, spricht man bei den tatsächlich beobachteten Ausprägungen von *Realisationen.* Merkmalsausprägungen erhalten einerseits eine inhaltliche Bezeichnung (z. B. weiblich, männlich), das sogenannte *Wertelabel,* und andererseits eine numerische Bezeichnung, bei der den Ausprägungen Zahlencodes (z. B. 1, 2) zugewiesen werden, kurz *Code* genannt. Bei bestimmten Variablen, wie dem Alter von Personen oder der Arbeitslosenquote in Ländern, erfolgt die Bezeichnung ausschließlich numerisch, da die Angabe selbst bereits mittels Zahlen erfasst wird. Neben den inhaltlich gültigen Angaben müssen zudem fehlende Werte ausgewiesen werden. Üblich ist es, für fehlende Angaben den Code 9 zu vergeben und für Befragte, die die Antwort auf eine Frage nicht wussten, den Code 8. Sind diese Zahlen schon für gültige Werte vergeben, wird auf den nächsthöheren Wertebereich ausgewichen, z. B. 98 und 99.[5]

Der tief gestellte Buchstabe i einer Variablen (x_i) wird als *Index* oder *Subskript* bezeichnet. Durch den Index i wird kenntlich gemacht, welcher der Fälle in einem Datensatz konkret gemeint ist. x_4 bedeutet beispielsweise, dass es sich um den Befragten mit der Nummer $i = 4$ in einer unsortierten Datenliste handelt. Liegt bereits eine sortierte Datenliste vor, verwendet man eine Klammer

5 Für die Standardkonventionen bezüglich der Kennzeichnung fehlender Werte vgl. Kühnel/Krebs (2012: 39).

Tab. 2.2 Illustration von Index- und Buchstabenschreibweise

i: einzelner, aber kein bestimmter Fall	1	2	3	4	5
x_i: konkrete Bestimmung eines Falls in einer unsortierten Datenliste	x_1	x_2	x_3	x_4	x_5
$x_{(i)}$: konkrete Bestimmung eines Falls in einer sortierten Datenliste	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
Ausprägungen einer Variablen, z. B. Geschlecht: weiblich = 1, männlich = 2	$x = 1$	$x = 2$	$x = 2$	$x = 1$	$x = 1$
konkrete Ausprägung des i-ten Falls in einer unsortierten Datenliste	$x_1 = 1$	$x_2 = 2$	$x_3 = 2$	$x_4 = 1$	$x_5 = 1$

Quelle: Eigene Darstellung.

bei der Indexangabe: $x_{(i)}$. Auf diese Weise ist gewährleistet, dass aus der Zahlenangabe jeder Befragte eindeutig bestimmbar ist. Zusätzlich kann über die Codeangabe für jeden Fall ausgewiesen werden, welche Merkmalsausprägung er für die Variable X aufweist, z. B. $x_4 = 1$ (vgl. Tab. 2.2). Demnach wäre der Befragte mit der Nummer $i = 4$ weiblich. Die Schreibweise mit Buchstaben und Indizes hat den Vorteil, dass sie in Formeln integrierbar ist, was zu einer besseren Lesbarkeit und mehr Übersichtlichkeit führt.

2.2 Exkurs: Messung

Bei der Vergabe von Wertelabels und Codes für Variablen müssen grundlegende Erkenntnisse der Messung berücksichtigt werden. Nach Stevens (1946: 677) versteht man unter *Messung* die Zuordnung von Ziffern zu Objekten oder Ereignissen nach bestimmten Regeln. Präziser formuliert muss die Zuordnung auf die Weise erfolgen, dass die *empirische Relation,* die zwischen den Objekten hinsichtlich eines bestimmten Merkmals besteht, durch die *numerische Relation,* also die Beziehung der Zahlen, abgebildet wird. Demnach muss für eine präzise und verlässliche Messung eine eindeutige Zuordnung von realen Eigenschaften des Objekts zu Zahlenwerten gewährleistet sein, damit eine *strukturtreue Abbildung* vorliegt.

Im Falle der Variablen Geschlecht ließe sich die Relation, in der die beiden Merkmalsausprägungen weiblich und männlich zueinander stehen, so beschreiben, dass auf eine Person entweder die eine oder die andere Kategorie zutrifft. Demzufolge wären die empirischen Relative „weiblich“ und „männlich“. Die numerische Bezeichnung erfolgt mittels zweier unterschiedlicher Zahlenwerte, wovon ebenfalls entweder der eine oder der andere auf die Person zutrifft. Als numerische Relative könnte beispielsweise für Frauen der Wert 1 und für Männer der Wert 2 vergeben werden. Im Sinne einer strukturtreuen Abbildung muss sichergestellt sein, dass einer weiblichen Person immer ein anderer Wert zugeordnet werden muss als einer männlichen.

Zu beachten ist, dass die Bezeichnung von Variablen und insbesondere die Festlegung der Ausprägungen eine Frage der jeweiligen Definition des Merkmals ist. Die bloße Bezeichnung Geschlecht ist nämlich keinesfalls hinreichend eindeutig. Unbeantwortet bleibt die Frage, welche Definition von Geschlecht zugrunde gelegt wird: das biologische, das soziale oder das psychologische Geschlecht. Je nach Definition würde man zumindest teilweise andere Antwortmöglichkeiten auf die Frage nach dem Geschlecht in einem Fragebogen verwenden. Deshalb ist es für eine verlässliche Messung wichtig, dass eine genaue Definition der Begriffe bzw. der theoretischen Konstrukte erfolgt.

2.3 Differenzierungsmöglichkeiten von Variablen

Es gibt verschiedene Möglichkeiten, Variablen zu differenzieren (vgl. z. B. Diekmann 2007: 123). Im Folgenden werden diejenigen vorgestellt, die relevant für die weiteren Ausführungen sind. Hierzu zählen die Anzahl der Merkmalsausprägungen einer Variablen und die Skalenniveaus, denen ein zentraler Stellenwert in der deskriptiven Statistik zukommt. Anschließend wird auf das in den Sozialwissenschaften häufig bestehende Ordinalskalenproblem eingegangen.

2.3.1 Anzahl der Ausprägungen

Je nachdem, wie viele Ausprägungen eine Variable hat, unterscheidet man dichotome bzw. binäre und polytome Variablen. *Dichotome* bzw. *binäre* Variablen haben nur zwei Merkmalsausprägungen, wie z. B. die Frage nach der Mitgliedschaft in einer Gewerkschaft (ja, nein). *Polytome* Variablen haben immer mehr als zwei Merkmalsausprägungen, z. B. das Bildungsniveau.

Dichotome Merkmale können entweder *natürlichen* Ursprungs sein oder *künstlich* erzeugt werden. Eine natürliche Dichotomie liegt vor, wenn das Merkmal von Natur aus nur zwei Ausprägungen aufweist (z. B. Frage nach Vorhandensein einer Parteimitgliedschaft). Eine künstliche Dichotomie wird erzeugt, indem man z. B. bei einer polytomen Variablen wie dem Bildungsniveau einen Grenzwert festlegt: Alle Personen, die mindestens die Fachhochschulreife erreicht haben, erhalten den Wert 1, Personen, die keinen Schulabschluss, einen Hauptschul- oder Realschulabschluss haben, werden unter dem Wert 0 zusammengefasst. Die Dichotomisierung von Variablen wird häufig verwendet, wenn es um die Klassifizierung von Objekten geht.

Eine binäre Variable, die abbildet, ob ein Objekt eine bestimmte Ausprägung aufweist oder nicht, wird auch als *Dummy-Variable* bezeichnet. Der Unterschied zwischen einer dichotomen bzw. binären Variablen und einer Dummy-Variablen besteht darin, dass Erstere nicht zwingend mit den Variablenwerten 0 und 1 codiert werden müssen, während dies bei Dummy-Variablen obligatorisch ist. Dummy-Variablen werden bei verschiedenen Analyseverfahren, u. a. der Regressionsanalyse, verwendet (vgl. Kap. 4.7.4).

2.3.2 Skalenniveaus

Ein für die deskriptive Statistik sehr wichtiges Unterscheidungskriterium von Variablen ist ihr *Skalenniveau.* In der Regel unterscheidet man vier Skalenniveaus: Nominal-, Ordinal-, Intervall- und Ratioskalen. Letztere wird auch als Verhältnisskala bezeichnet. Gelegentlich wird noch die Absolutskala als fünftes Skalenniveau und Sonderfall der Ratioskala angeführt (vgl. Tab. 2.3). In-

Tab. 2.3 Eigenschaften von Skalenniveaus

Skalenniveau	Nominalskala	Ordinalskala	Intervallskala	Ratioskala	Absolutskala
empirische Relation	Unterscheidbarkeit/ Gleichheit	+ Rangordnung	+ Linearität/ konstante Abstände	+ natürlicher Nullpunkt	+ natürliche Einheiten
Term	$a \neq b$, $a = b$	$a < b$, $a > b$	$a < b < c < d$ und damit $b - a = d - c$	$a/b = c/d$	$a = a$ Wert an sich
Beispiel	Geschlecht	höchster Schulabschluss	Jahr der Geburt	Einnahmen im Monat (in €)	Anzahl der Geschwister

→ zunehmender Informationsgehalt

Quelle: Eigene Darstellung.

tervall-, Ratio- und Absolutskalen werden oft unter dem Oberbegriff *metrische Skalen* oder *Kardinalskalen* zusammengefasst, während es sich bei Nominal- und Ordinalskala demzufolge um *nicht metrische* Skalen handelt. Häufig findet in Anlehnung an die Skalenniveaus auch die Differenzierung zwischen *kategorialen* und *nicht kategorialen* Variablen Verwendung. Zu kategorialen Variablen zählen alle Variablen mit nur wenigen Merkmalsausprägungen, was insbesondere auf nominale und ordinale Variablen zutrifft. Manchmal werden auch Variablen als kategorial angesehen, die durch Klassierung aus metrischen Variablen entstanden sind. Nicht kategoriale Variablen sind dagegen Variablen mit sehr vielen Merkmalsausprägungen, was kennzeichnend für metrische Variablen ist.

Eine *Skala* im Sinne der Messtheorie zeichnet sich durch die bei der Messung genannten Bestandteile aus: ein empirisches Relativ, ein numerisches Relativ und eine strukturtreue Abbildung, die die beiden Relative miteinander verknüpft. Messungen werden danach klassifiziert, welche Arten empirischer Relationen von Objekten sie abbilden. Mithilfe dieser Relationen lassen sich wiederum die Skalenniveaus beschreiben, die im Folgenden anhand verschieden skalierter Merkmale erläutert werden.

Bei einer Variablen auf *Nominalskalenniveau* (z. B. Geschlecht) können die Merkmalsausprägungen bzw. die Variablenwerte nur im Hinblick auf gleich ($a = b$) oder ungleich ($a \neq b$) interpretiert werden. Der Umstand, dass der Wert 2 für „männlich“ größer ist als der Wert 1 für „weiblich“ bildet keine entsprechende Relation bei den

Personen im empirischen Relativ. Nominale Variablen dienen lediglich der Klassifikation von Untersuchungsobjekten. Folglich können zwei Personen z. B. entweder das gleiche Geschlecht haben oder ein unterschiedliches.

Die Merkmalsausprägungen einer Variable, die auf *Ordinalskalenniveau* gemessen wurde (z. B. höchster Schulabschluss), können zusätzlich zu gleich oder ungleich im Sinne von kleiner ($a < b$) oder größer ($a > b$) interpretiert werden. Eine Person verfügt beispielsweise über einen höheren Schulabschluss als eine andere.

Das *Intervallskalenniveau* zeichnet aus, dass zusätzlich zur (Un-)Gleichheit und Reihenfolge der Werte die Differenzen ($b - a = d - c$) interpretiert werden können (z. B. Geburtsjahr). Die Abstände zwischen zwei Geburtsjahren sind beispielsweise identisch. Allerdings ist kein natürlicher Nullpunkt gegeben. Der Nullpunkt ist weitgehend willkürlich festgelegt. So hat die Oktoberrevolution nach dem Julianischen Kalender im Oktober stattgefunden, nach dem Gregorianischen Kalender allerdings im November.

Auf *Ratioskalenniveau* gemessene Variablen (z. B. Einkommen) ermöglichen zusätzlich zur (Un-)Gleichheit, der Reihenfolge und den Differenzen der Werte die Interpretation von Verhältnissen. Eine Person kann beispielsweise doppelt so viel oder ein Viertel weniger verdienen als eine andere. Die Besonderheit dieses Skalentyps ist der natürliche Nullpunkt, der vom Menschen nicht beeinflussbar ist. Der Wert 0 kann im Sinne von „nicht vorhanden“ verstanden werden (z. B. beim Einkommen). Weniger als 0 Euro kann man monatlich nicht einnehmen.

Bei einer *Absolutskala* können zusätzlich zu allen anderen Eigenschaften die Werte „an sich" interpretiert werden. Das bedeutet, dass diese Variablen nur in der gegebenen Einheit gemessen werden können. Bei der Anzahl der Geschwister wäre es z. B. nicht sinnvoll, eine andere Maßeinheit (etwa Doppelgeschwister) zu verwenden. Die Maßeinheit ist natürlich gegeben. Wenn jemand zwei Geschwister hat, dann ist ihm auch im numerischen Relativ der Wert 2 und kein anderer zuzuweisen.

Die Skalenniveaus sind in der angeführten Reihenfolge hierarchisch geordnet. Das heißt, jedes höhere Skalenniveau weist auch die Eigenschaften der niedrigeren Skalenniveaus auf. Beispielsweise besitzen intervallskalierte Merkmale nicht nur die ihnen zugewiesene Eigenschaft der konstanten Abstände zwischen den Ausprägungen, sondern zusätzlich auch die der Ordinal- und Nominalskala zugeordnete Eigenschaft der Rangordnung und der Gleichheit bzw. Ungleichheit. Daraus folgt, dass Variablen je nach Skalenniveau unterschiedlich interpretiert werden können. Dabei gilt, dass die Anzahl der Interpretationen und der Informationsgehalt der Messwerte in der oben angeführten Reihenfolge der Skalenniveaus steigen. Umgekehrt können Variablen durch Umcodierung, z. B. Zusammenfassung von Ausprägungen, abwärts transformiert werden. So kann die metrische Altersvariable durch Klassierung in eine ordinale Variable (z. B. < 18, 18 bis < 30, 30 bis < 60, 60+ Jahre) oder durch Festlegung eines Schwellenwerts (z. B. jünger als 50 Jahre vs. 50 Jahre und älter) in eine nominale Variable überführt werden.

Skalenniveaus sind sowohl für die Untersuchungsplanung als auch die Datenanalyse sehr wichtig, da sie festlegen, welche Berechnung von Maßzahlen zulässig ist. Dies trifft sowohl für die Beschreibung von Merkmalsverteilungen im Rahmen von univariaten Analysen als auch für die Quantifizierung von Zusammenhängen in bivariaten und multivariaten Analysen zu. Dabei gilt die Faustregel: Je höher das Skalenniveau, desto mehr Auswertungsoptionen gibt es.

2.3.3 Exkurs: Ordinalskalenproblem

Die Relevanz, die den Skalenniveaus beigemessen wird, unterscheidet sich je nach Fachdisziplin. In manchen sozialwissenschaftlichen Disziplinen wie in Teilen der Psychologie wird die „liberale" Auffassung vertreten, auf eine empirische Überprüfung des Skalenniveaus in der Forschungspraxis verzichten zu können. Anhänger dieser Position gehen davon aus, dass es sich bei den meisten Messungen um *Per-Fiat-Messungen* handelt, was Messung „durch Vertrauen" bedeutet. Dieser Praxis liegt die Annahme zugrunde, dass Messungen mit Erhebungsformaten (z. B. Ratingskalen) erfolgen, die das interessierende Merkmal auf Intervallskalenniveau abbilden. Der Vorteil hierbei ist, dass unter dieser Prämisse die deutlich differenzierteren Auswertungsmöglichkeiten für intervallskalierte Variablen auch auf anders skalierte Variablen anwendbar sind. Dieser „liberalen" Auffassung stehen die Anhänger der strengeren Position gegenüber, die entschieden eine Überprüfung der messtheoretischen Annahmen fordern.

In der sozialwissenschaftlichen Praxis wird auch der Begriff von *quasi-* bzw. *pseudo-metrischen* Variablen verwendet. Damit wird zum Ausdruck gebracht, dass die Merkmalsausprägungen einer Variablen den Ansprüchen des metrischen Skalenniveaus zwar nicht gerecht werden, aber unter Vorbehalt wie ein solches behandelt werden.

Ein Standardinstrument, das eine solche Sonderstellung einnimmt, ist die *Likert-Skala.* Mit dieser Methode wird ein theoretisches Konstrukt (z. B. Ausländerfeindlichkeit) mit Hilfe mehrerer Variablen gemessen. Die Befragten können ihre Antwort zu vorgegebenen Aussagen, z. B. bezüglich Zustimmung, abstufen. Oft handelt es sich um Ratingskalen mit fünf Merkmalsausprägungen: stimme voll ganz zu, stimme eher zu, teils/teils, stimme eher nicht zu, stimme überhaupt nicht zu. Die Werte der einzelnen Antworten werden dann zu einer Likert-Skala summiert. Messtheoretisch betrachtet sind die Antworten auf einer Likert-Skala ordinalskaliert, da nicht angenommen werden kann, dass die Befragten die Abstände zwischen den einzelnen Antworten als gleich ansehen. Um jedoch bei der Datenanalyse die gesamten statistischen Operationen einsetzen zu können, wird die Likert-Skala gerne als quasi-metrisch bezeichnet und wie eine Intervallskala behandelt. Voraussetzung für eine Interpretation als quasi-metrisch ist, dass die Variablen mindestens fünf Ausprägungen haben und die Abstände zwischen den Antworten semantisch und durch numerische Wertzuweisung als gleich groß interpretiert werden können (vgl. Urban/Mayerl 2011: 275).

Letztlich muss jeder Forscher selbst entscheiden, welche Position er beim Umgang mit Skalenniveaus präferiert. Beide Positionen haben Vor- und Nachteile, denen man sich als Forscher bewusst sein sollte (vgl. Baur 2011: 213 f.).

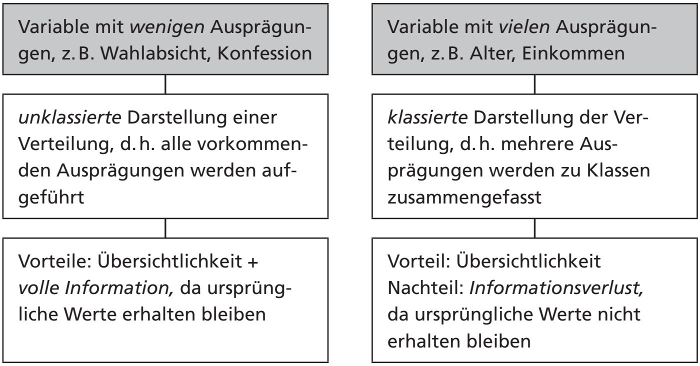
2.4 Klassierung von Daten

Generell sollte versucht werden, Variablen auf dem höchstmöglichen Skalenniveau zu messen, da die Daten auf diese Weise am meisten Informationen enthalten. Manchmal kann es allerdings sinnvoll sein, ein niedrigeres Skalenniveau zu verwenden. Beispielsweise wird die Frage nach dem Einkommen gelegentlich mit Hilfe von *Klassen* erhoben, da auf diese Weise weniger Befragte die Antwort verweigern. Dabei werden die möglichen Ausprägungen einer Variablen zusammengefasst. Die Befragten werden also nicht gebeten, ihr Einkommen direkt anzugeben (offene Frage), sondern bekommen bestimmte Einkommensgruppen vorgegeben, um sich einzuordnen (geschlossene Frage). Der Vorteil einer geringeren Anzahl von Antwortverweigerern auf die Frage geht bei dieser Vorgehensweise zulasten des Informationsniveaus. Die Entscheidung für eine *Klassierung* fällt in diesem Fall *vor* der Datenerhebung und ist folglich im Nachhinein nicht mehr rückgängig zu machen. Ein höheres Messniveau kann zwar immer in ein niedrigeres Messniveau überführt werden, aber nicht umgekehrt. In der praktischen Arbeit sollte daher sorgfältig abgewogen werden, welche Operationalisierung für das Untersuchungsziel angemessen ist.

Häufig ist es bei der Arbeit mit Daten auch notwendig, diese im Rahmen der Datenaufbereitung zunächst geeignet zu transformieren. Das bedeutet, die numerischen Werte einer Variablen werden nach einer bestimmten Regel in andere Werte umgeformt. Dies kann für die Interpretation einer Variablen von Vorteil sein. Es empfiehlt sich aber auch, wenn man metrische Variablen mit sehr vielen Merkmalsausprägungen als Tabelle oder als Abbildung darstellen möchte. Um die Übersichtlichkeit zu erhalten, sollte man die Daten klassieren. Da im Normalfall auf die ursprünglichen Werte wieder zugegriffen werden kann, ist eine Klassierung von Daten im Rahmen der Datenaufbereitung weniger folgenschwer, als wenn die Entscheidung vor der Datenerhebung gefällt wird bzw. werden muss. Hinzu kommt, dass der mit der Klassierung einhergehende Informationsverlust für deskriptive Analysen in der Regel zu vernachlässigen ist.

Fasst man mehrere Merkmalsausprägungen einer Variablen zusammen, sollte man darauf achten, dass die Klassen disjunkt und erschöpfend sind (vgl. Kap. 2.1). Die Faustregel bei der Festlegung von *Klassengrenzen* lautet, dass die Klassen so zu wählen sind, dass die Abstände regelmäßig bzw. alle Klassen gleich breit sind (ggf. mit Ausnahme der ersten und letzten Kategorie). In besonderen Fällen kann jedoch von dieser Regel abgewichen werden und eine Klasseneinteilung auf Basis inhaltlicher Kriterien (z. B. historische Jahrgangskohorten) erfolgen. Ebenso ist zu beachten, dass alle Klassen mit einer ausreichenden Zahl von Fällen besetzt sind. *Offene Randklassen* (z. B. < 20 Jahre, > 90 Jahre) sind zu

Abb. 2.1 Klassierung von Daten



Quelle: Eigene Darstellung.

vermeiden, da ansonsten bestimmte statistische Kennwerte und Grafiken nicht umsetzbar sind. Außerdem sollte die „Grundform" der Verteilung durch die Klassierung erhalten bleiben. Eine zweigipflige Verteilung sollte beispielsweise nicht zu einer eingipfligen werden (vgl. Kap. 3.6.1). Einen Überblick, für welche Variablen sich eine klassierte oder unklassierte Darstellung anbietet und welche Vor- und Nachteile damit verbunden sind, gibt Abb. 2.1.

2.5 Indexbildung

Bei einem *Index* werden mehrere Variablen zu einer neuen zusammengerechnet. Die *Indexbildung* stellt ebenfalls ein Hilfsinstrument dar, bei dem metrisches

Skalenniveau unterstellt wird, obwohl in der Regel nicht alle Informationen auf metrischem Skalenniveau erfasst wurden. Indizes sind in den Sozialwissenschaften weit verbreitet, da häufig theoretische Konstrukte Analysegegenstand sind, die *latent* und folglich nicht direkt messbar sind (z. B. politisches Kompetenzgefühl, Ausländerfeindlichkeit). Um das nicht direkt beobachtbare (latente) theoretische Konstrukt möglichst strukturtreu abzubilden, erfolgt die Operationalisierung über mehrere beobachtbare (*manifeste*) Einzelindikatoren. Oft wird in diesem Zusammenhang die Likert-Skala verwendet.

Indizes können auf unterschiedliche Art und Weise berechnet werden: Im Falle des (ungewichteten) *additiven Indexes,* auch *Summenscore* genannt, werden die Werte der Einzelindikatoren aufsummiert. Wenn ein bestimmter Indikator aufgrund theoretischer Überlegungen oder empirischer Vorkenntnisse stärker in den Indexwert eingehen soll als die anderen Indikatoren, wird ein *gewichteter additiver Index* gebildet. Beim *multiplikativen Index* werden die Einzelindikatoren multipliziert. Dieser kommt zum Einsatz, wenn intendiert ist, dass das theoretische Konstrukt jeden der Einzelindikatoren erfordert. Eine weitere Möglichkeit der Indexbildung ist der *Mittelwertindex.* Für diesen wird der additive Index durch die Anzahl der eingegangenen Indikatoren geteilt. Das hat den Vorteil, dass die neu gebildete Variable den Wertebereich der ursprünglichen Indikatoren aufweist. Voraussetzung hierfür ist, dass die Einzelindikatoren die gleichen Antwortmöglichkeiten haben und in die gleiche Richtung weisen. Das heißt, hohe Werte dürfen nicht gleichzeitig Zustimmung und Ablehnung bedeuten, wie es bei positiv und negativ formulierten Aussagen der Fall wäre.

Am Beispiel des Institutionenvertrauens wird in Tab. 2.4 (Seite 26) die Berechnung aller vier Indextypen vorgestellt. Ein Befragter wurde hierzu gebeten, sein Vertrauen in vier politische Institutionen in Deutschland auf einer Skala von 1 bis 7 anzugeben, wobei 1 für gar kein Vertrauen und 7 für sehr großes Vertrauen steht. Es wird jeweils der Indexwert ausgewiesen, der sich für den Befragten ergeben würde. Zusätzlich werden die theoretischen Endpunkte der Indizes in Form des geringsten Wertes (Minimum) und höchsten Wertes (Maximum) angeführt.

2.6 Exkurs: Summenzeichen $\sum$

Die Berechnung von Summen ist in der deskriptiven Statistik nicht nur zur Bildung von Indizes wichtig. Häufig müssen auch Realisationen einer Variablen von unterschiedlichen Untersuchungseinheiten aufaddiert werden, um statistische Kennwerte zu bestimmen. Da in der sozialwissenschaftlichen Forschung oft eine Vielzahl von Fällen vorliegt, ist es nicht zweckmäßig, Formeln mit jedem einzelnen Summanden anzuführen. Stattdessen wird für eine verkürzte Schreibweise das Summenzeichen $\sum$ (griech. Sigma) gefolgt von einer Funktion verwendet. Meist wird dieses noch um einen Laufindex mit einem Startwert unterhalb und einem Endwert oberhalb des Symbols ergänzt.

Tab. 2.4 Indextypen am Beispiel Institutionenvertrauen

Vertrauen in …	Parteien	Bundestag	Bundesregierung	BVerfG
Antwort des Befragten	4	5	5	7

Indextyp	Indexwert Befragter	minimaler Indexwert	maximaler Indexwert
(ungewichtet) additiv	$4 + 5 + 5 + 7 = 21$	$1 + 1 + 1 + 1 = 4$	$7 + 7 + 7 + 7 = 28$
gewichtet additiv*	$(4 \cdot 2) + 5 + 5 + 7 = 25$	$(1 \cdot 2) + 1 + 1 + 1 = 5$	$(7 \cdot 2) + 7 + 7 + 7 = 35$
multiplikativ	$4 \cdot 5 \cdot 5 \cdot 7 = 700$	$1 \cdot 1 \cdot 1 \cdot 1 = 1$	$7 \cdot 7 \cdot 7 \cdot 7 = 2401$
Mittelwertindex	$\frac{(4 + 5 + 5 + 7)}{4} = 5{,}25$	$\frac{(1 + 1 + 1 + 1)}{4} = 1$	$\frac{(7 + 7 + 7 + 7)}{4} = 7$

* doppeltes Gewicht von Parteien

Quelle: Eigene Darstellung.

Ein typisches Beispiel für die Verwendung des Summenzeichens sieht wie folgt aus:

$$\sum_{i=1}^{n} x_i$$

Das bedeutet, dass der Laufindex i einen Startwert von 1 aufweist und die Summenbildung so lange wiederholt wird, bis der Endwert n erreicht ist. Der Term, der zur Summation verwendet wird, lautet x_i. Anstelle des Summenzeichens könnte man die Formel auch auf folgende Weise schreiben:

$$x_1 + x_2 + \ldots + x_n$$

Als konkretes Beispiel für die Berechnung einer solchen Summe ist in Tab. 2.5 für fünf Personen die Anzahl der Vereinsmitgliedschaften wiedergegeben.

Tab. 2.5 Illustration Summenzeichen

unbestimmter Einzelfall i	1	2	3	4	5
konkreter Fall x_i	x_1	x_2	x_3	x_4	x_5
Ausprägungen der Fälle für Variable X „Anzahl Vereinsmitgliedschaft“	$x_1 = 2$	$x_2 = 3$	$x_3 = 0$	$x_4 = 2$	$x_5 = 1$

Quelle: Eigene Darstellung.

Setzt man für n die Zahl der Untersuchungseinheiten ein, ergibt sich folgende Formel:

$$\sum_{i=1}^{5} x_i = 2 + 3 + 0 + 2 + 1 = 8$$

Die Summe der Vereinsmitgliedschaften aller fünf Befragten liegt folglich bei 8.

http://www.springer.com/978-3-658-10674-4

Deskriptive Statistik
Eine Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler
Völkl, K.; Korb, C.
2018, XXI, 325 S. 34 Abb., Softcover
ISBN: 978-3-658-10674-4