

2.1 Möglichkeiten der Datenbeschaffung

Beginnen wir mit dem ersten Schritt des Intelligence Cycle, der Datengewinnung: In vielen Unternehmen liegen wichtige Daten häufig bereits vor. So werden Absatz- und Verbrauchszahlen erfasst, aber leider in den wenigsten Fällen auch systematisch und zentral gesammelt, sodass eine Auswertung unter Zugriff auf eine zentrale Datenbank nur selten möglich ist. Die Aufgabe, die ein Statistiker somit zunächst zu erfüllen hat, ist die „Bergung dieser Schätze“ im Unternehmen. Auch hier ist nicht selten kommunikative Kompetenz gefragt, um Mitarbeiter davon zu überzeugen, das „Eigentum“ der Daten – die Leistungen und vielleicht auch Fehlleistungen offenbaren könnten – für eine systematische Auswertung freizugeben.

Selbst wenn ein Unternehmen den strategischen Entschluss getroffen hat, bestimmte Daten systematisch zu sammeln, heißt das noch lange nicht, dass sofort mit der Auswertung begonnen werden kann. Wer darf die Daten auswerten, wer ist dazu überhaupt in der Lage, wer hat die Zeit dazu? Diese Fragen begegnen einem in der Praxis immer wieder. So entstehen im Rahmen der Sammlung von Einkaufsdaten durch Kundenkarten (*engl.*: *fidelity cards*) außergewöhnlich große Datensätze, deren Verwaltung nicht selten die Arbeit einer ganzen Abteilung in Anspruch nimmt, ohne dass eine systematische Auswertung durchgeführt werden kann.

Neben den Daten, die den Unternehmen aufgrund eigener Datensammlung zur Verfügung stehen, existieren viele *öffentlich zugänglichen Datenbanken*: Unter den Anbietern finden sich private Dienstleistungsunternehmen wie AC-Nielsen, die Gesellschaft für Konsumforschung (GfK) etc., die ihre Datenreihen in der Regel kostenpflichtig zur Verfügung stellen. Aber auch bei Forschungsinstituten, statistischen Landes- und Bundesämtern sowie im internationalen Bereich bei Eurostat, bei der OECD, der Weltbank etc. sind Daten verfügbar, aus denen sich z. T. lehrreiche Schlüsse für unternehmerisches Handeln ziehen lassen. Interessante Links sind in Tab. 2.1, 2.2 und 2.3 zusammengestellt:

Tab. 2.1 Amtliche Statistiken nationaler Institutionen

Statistisches Bundesamt	www.stabu.de	Detaillierte Zeitreihen (z. T. Monatsbasis)
	Statistisches Jahrbuch	Jährliche Informationen über die Bundesrepublik, z. T. auch internationale Daten
	Fachserien	Unregelmäßige detailliertere Fachinformationen über Teilbranchen/Teilregionen
Statistische Landesämter	Statistische Berichte	Regionale Daten
Bundesbank	www.bundesbank.de	Berichte zu Kapital-, Wertpapiermärkten, Zahlungsbilanzen, Banken, Währungen
Bundesregierung	Homepage von Ministerien	Jahreswirtschaftsbericht, Finanzbericht etc. unterschiedlicher Ministerien (z. B. BMWi oder BFA)

Tab. 2.2 Nichtamtliche Statistiken nationaler Institutionen

Forschungsinstitute	www.bundes-regierung.de	Halbjahresgutachten: Informationen zur Gesamtwirtschaft
Sachverständigenrat	www.sachver-staendi-genrat-wirtschaft.de	Jahresgutachten mit nationalen und internationalen makroökonomischen Daten
Ifo Institut	www.ifo.de	Ifo Spiegel der Wirtschaft, Konjunkturtest, Investitionstest (Panel seit 1949)
DIW Berlin	www.diw.de	DIW Wochenberichte: Aktuelle Daten und Prognosen; SOEP Haushaltspanel
ZEW Mannheim	www.zew.de	Finanzmarkttest; Innovationspanel
Verbandsstatistiken	Homepage/ Publikationen	Fachverbände stellen Daten zur Verfügung z. B. VCI, VDMA, VDA etc.
GfK Nürnberg	www.gfk.de	Handelspanel, TV-Panel, GfK ConsumerScan, GfK ConsumerScope, Konsumindex etc.
AC Nielsen	www.acnielsen.de	Handelspanel, Haushaltspanel

Tab. 2.3 Statistiken internationaler Institutionen

Nationale Statistikämter	www.destatis.de/allg/d/sitemap/sitemap7.htm	Siehe Linkservice International unter www.stabu.de
Eurostat	www.europa.eu.int/comm/eurostat/	Informationen in der Gesamtdarstellung sowie Länder- und Branchenberichte
OECD	www.oecd.org	Länderdaten über Wirtschaft, Arbeitsmarkt, Gesundheit, Handel, Entwicklung, etc.
Weltbank	www.worldbank.org	Länderdaten über Kapital- und Finanzmärkte
UN	www.un.org	Politikdaten, Wirtschaftsdaten
ILO	www.ilo.org	Arbeitsmarktdaten
IMF (IWF)	www.imf.org	Finanzmärkte, Währungsstabilität, Entwicklungsberichte
Internationale Verbände	Homepage/ Publikationen	Branchendaten

Am Beispiel des Ifo-Tests wollen wir die für betriebswirtschaftliche Entscheidungen wichtigen Erkenntnisgewinne einmal genauer betrachten. Stellen wir uns hierzu die Abteilung Einkauf/Lagerhaltung eines Unternehmens zur Herstellung von Vorprodukten aus dem Maschinenbau vor. Um Lagerungs-, Fehlmengen-, Bestell- und Produktionskosten im Unternehmen optimieren zu können, muss die stochastische Nachfrage nach Roh-, Hilfs- und Betriebsstoffen – also Warenabflüsse aus dem Eingangslager – im Hinblick auf die optimale Bestellmenge und den optimalen Bestellzeitpunkt für das Eingangslager prognostiziert werden. Sicher könnte man an dieser Stelle die eigene Vertriebsabteilung nach den künftigen Auftragseingängen befragen, denn hieraus ließen sich die künftige Produktion und der künftige Materialbedarf abschätzen. Erfahrungsgemäß geben Vertriebsabteilungen aus dem Interesse der Sicherstellung der eigenen Lieferfähigkeit die künftigen Verkäufe immer weit überschätzt an. Deshalb entscheidet sich die Abteilung für die Nutzung der im IFO-Konjunkturtest erhältlichen Informationen:¹ Im einfachsten Fall könnte die Abteilung aus einer der monatlich erhobenen Informationen *über die Geschäftslage der Abnehmerindustrie für die nächsten 6 Monate* eine sicherlich valide Prognose erstellen. Schätzt die Abnehmerindustrie die Geschäftslage in den nächsten sechs Monaten nämlich rückläufig ein, wird auch der Absatz unseres Zulieferunternehmens in der nächsten Zeit zurückgehen und vice versa. Ohne eine eigene Befragung lassen sich somit aus öffentlich zugänglichen Quellen Bestellmengen-Entscheidungen treffen.²

Die Daten können in unterschiedlichen Aggregationszuständen vorliegen. Sie beziehen sich dabei fast nie auf ein einzelnes Unternehmen oder eine einzelne Person, sondern immer auf das Aggregat bestimmter Unternehmenskategorien oder Personengruppen. So finden sich in den Auswertungen des ZEW Innovationspanel niemals Informationen über das Innovationsverhalten eines einzelnen Unternehmens, aber Angaben über die durchschnittlichen F&E Ausgaben bzw. die Innovationsintensität (Innovationsausgaben/Umsatz) von Chemieunternehmen in der Größenklasse zwischen 20 und 49 Mitarbeitern. Diese Informationen ermöglichen dem Einzelunternehmen allerdings ein Benchmarking mit den eigenen Kennziffern. Auch beim GfK Haushaltspanel geht es nicht um die Kaufaktivitäten einzelner Personen, sondern immer um die Betrachtung von Haushalten. Selbst Daten von individuellen Kundenkarten lassen realistischerweise nicht auf Einzelpersonen rückschließen, denn es bleibt unklar, ob der Ehemann bei seinem Einkauf nicht doch die Kundenkarte seiner Ehefrau genutzt hat. Sachlich würde man auch hier von einer Haushaltsbetrachtung ausgehen.

Um an Informationen über eine Einzelperson bzw. ein Einzelunternehmen zu kommen, muss man auf eine eigene *Befragung* (engl.: survey) zurückgreifen. Diese ist in der Regel

¹ Aus einer monatlichen Befragung geben rund 7000 Unternehmen des Verarbeitenden Gewerbes, des Bauhauptgewerbes, des Großhandels und des Einzelhandels Auskunft über die derzeitige Geschäftslage, die inländische Produktionstätigkeit, den Bestand an Fertigwaren, die Nachfragesituation, die Inlandsverkaufspreise und den Auftragsbestand im Vergleich zum Vormonat, die Auslandsaufträge, das Exportgeschäft, die Entwicklung der Beschäftigtenzahl und die Verkaufspreise für die nächsten 3 Monate sowie die Geschäftslage für die nächsten 6 Monate.

² Vergleiche hierzu besonders die in Kap. 5 beschriebene Methode.

die kostenintensivste Form der Datenbeschaffung, erlaubt es auf der anderen Seite aber, die genauen Fragestellungen der Untersuchung selbst spezifizieren zu können. Die Befragung kann – abhängig vom jeweiligen Thema – mündlich oder schriftlich durchgeführt werden. Neben die traditionellen Papierfragebögen treten dabei immer häufiger Telefon- und Internetbefragungen.

2.2 Die Entscheidung für ein Skalenniveau

Es ginge in diesem Lehrbuch über die Aufgabe eines Statistikbuches hinaus, alle Regeln der richtigen Konstruktion von Fragebögen zu behandeln. Hier sei auf die dafür einschlägige Literatur verwiesen (vgl. z. B. Bühner 2010). Deshalb werden wir uns an dieser Stelle auf die Grundlagen beschränken, die für die Entscheidung für oder gegen eine bestimmte quantitative Auswertungsmethodik von Bedeutung sind.

Hierzu wollen wir zunächst ein Beispiel betrachten: Stellen Sie sich vor, Sie seien Besitzer eines Einzelhandels in einer Kleinstadt. Einige Kunden sind an Sie herangetreten, weil sie eine Erhöhung der Auswahlvielfalt der angebotenen Margarine- bzw. Buttersorten wünschen. Da Ihnen allerdings nur begrenzt Ausstellungs- und Lagerraum zur Verfügung steht, möchten Sie wissen, ob die Meinung der Anfragenden *repräsentativ* für alle Kunden ist. Sie beauftragen eine Gruppe von Studierenden mit der Durchführung einer kleinen schriftlichen Kundenbefragung unter Verwendung des Fragebogens in Abb. 2.1.

Die Studierenden beginnen sofort mit ihrer Arbeit und nach einer Woche haben sie 850 Kunden befragt. Jede der befragten Personen ist dabei ein sog. *Merkmalsträger*, der die (den Auftraggeber) interessierenden *Merkmale* besitzt. So ist darunter auch der Merkmalsträger Herr Hinz mit den *Merkmalsausprägungen* *männlich, 67 Jahre, 74 kg, Margarine* und *mittelmäßig* für die Merkmale Geschlecht, Alter, Körpergewicht, Brotaufstrich und Wahrnehmung der Angebotsvielfalt. Vor jeder Befragung muss also definiert werden,

Geschlecht: ☐ männlich ☐ weiblich

Alter: _____ Jahre

Körpergewicht: _____ kg

Welchen Brotaufstrich verwenden Sie? (Nur eine Antwort möglich)

☐ Butter ☐ Margarine ☐ Sonstiges

Wie empfinden Sie die Angebotsvielfalt des von Ihnen bevorzugten Brotaufstriches in unserem Geschäft auf einer Skala von 1 (= sehr schlecht) bis 5 (=sehr gut)?

sehr schlecht ☐ ☐ ☐ ☐ ☐ sehr gut

1 2 3 4 5

Abb. 2.1 Fragebogen Kundenbefragung Einzelhandel

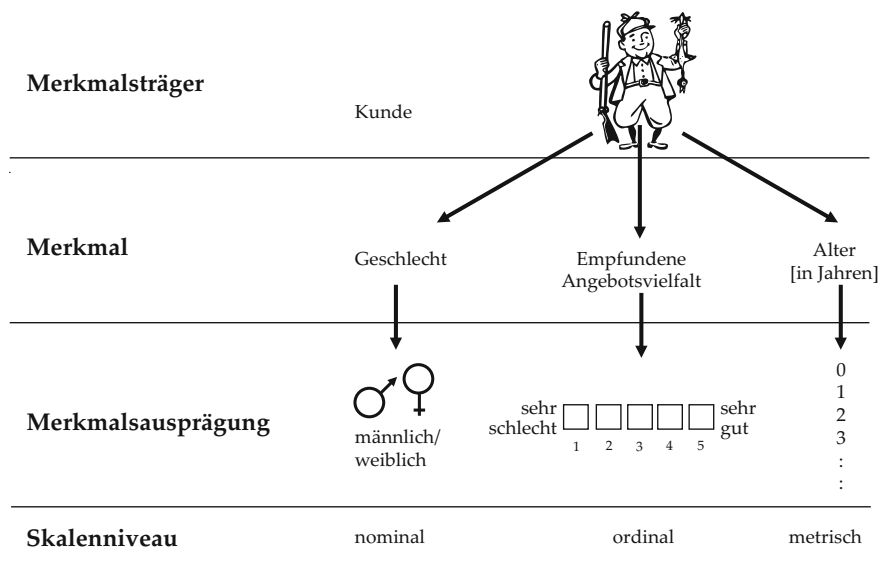


Abb. 2.2 Merkmalsträger/Merkmale/Merkmalsausprägung/Skalenniveau

welche Merkmalsträger von Interesse sind, welche Personen befragt werden sollen. Danach müssen die interessierenden Merkmale (Fragen) und deren Merkmalsausprägungen festgelegt werden.

Verallgemeinernd kann man sagen, dass es sich bei Merkmalsträgern um die Subjekte oder Objekte der Untersuchung handelt, die sich hinsichtlich ihrer Merkmalsausprägungen für bestimmte Merkmale unterscheiden. Die in Abb. 2.2 dargestellten Merkmale Geschlecht, Angebotsvielfalt und Alter stellen dabei beispielhaft die *drei existierenden Skalenniveaus* in der quantitativen Datenanalyse dar: Nominalskalen, Ordinalskalen und metrische Skalen.

Die niedrigste Form der Skala ist die *Nominalskala*. Den einzelnen Merkmalsausprägungen werden dabei Zahlen zugeordnet, wie z. B. $x_i = 1$ für *männlich* oder eine $x_i = 2$ für *weiblich*. Diese Zahlen dienen aber ausschließlich dazu, die einzelnen Untersuchungsobjekte der jeweiligen Gruppe zuzuordnen (z. B. der Gruppe der männlichen Befragten) und sie von anderen Gruppen (z. B. der Gruppe der weiblichen Befragten) unterscheidbar zu machen. Jedes Untersuchungsobjekt kann dabei nur einer Gruppe zugeordnet werden und alle Untersuchungsobjekte mit der gleichen Merkmalsausprägung erhalten die gleiche Zahl. Da die zugeordneten Zahlen somit nichts anderes als Gruppenbezeichnungen sind, kann es kein sinnvolles *größer/kleiner*, *weniger/mehr* oder *besser/schlechter* geben. Lediglich die Unterscheidung *Zugehörigkeit* oder *Keine Zugehörigkeit* zu einer Gruppe kann getroffen werden ($x_i = x_j$ versus $x_i \neq x_j$). Bezüglich des Merkmals Geschlecht bedeutet dies, dass eine *Eins* für die Ausprägung männlich nicht besser oder schlechter ist als eine *Zwei* für weiblich, sondern lediglich den Datensatz hinsichtlich männlicher und weibli-

cher Befragter segmentiert. Auch für die nominalen Merkmale Beruf (z. B. 1 = Metzger; 2 = Bäcker; 3 = Schornsteinfeger), Nationalität, Studiengangzugehörigkeit, etc. sind keine Bildungen von Rangfolgen möglich.

Dies führt über zum nächst höheren Skalenniveau, der sog. *Ordinalskala*. Bei dieser werden den einzelnen Merkmalsausprägungen ebenfalls Zahlen zugeordnet, diese stellen nun aber eine Rangfolge dar. Typische Beispiele hierfür sind Fragen eines Fragebogens, die Antworten auf einer Skala von 1 bis X vorsehen, wie beispielsweise das obige Merkmal der Wahrnehmung der Angebotsvielfalt, welches auf einer Skala von eins bis fünf gemessen wird. Dies ermöglicht dem Forscher die Feststellung der Intensität einer Merkmalsausprägung eines Untersuchungsobjektes im Vergleich zu anderen. Haben Frau Koslowski und Frau Neumüller bei der Angabe zur Angebotsvielfalt das Kästchen drei angekreuzt, so kann davon ausgegangen werden, dass beide die gleiche Wahrnehmung bezüglich der Angebotsvielfalt haben. Wie bei der nominalen Skala erhalten Untersuchungsobjekte mit gleichen Ausprägungen auch den gleichen Wert zugeordnet. Kreuzt Herr Martin das Kästchen vier an, so bedeutet dies nicht nur, dass er eine andere Wahrnehmung als die Damen Koslowski und Neumüller hat, sondern auch, dass er die Angebotsvielfalt als besser einschätzt. Ein ordinales Skalenniveau erlaubt also eine Rangbildung und somit ein sinnvolles *größer/kleiner*, *weniger/mehr* oder *besser/schlechter* ($x_i = x_j$; $x_i > x_j$; $x_i < x_j$).

Was hingegen nicht gesagt werden kann ist, wie groß der Abstand zwischen diesen beiden Rangplätzen drei und vier ist. Mehr noch, es kann streng genommen nicht einmal davon ausgegangen werden, dass der Abstand zwischen dem ersten und dem zweiten Rang genau so groß ist, wie der zwischen anderen benachbarten Rangplätzen. Dies wird an Platzierungen bei sportlichen Wettkämpfen, einem weiteren Beispiel für eine ordinale Skalierung, deutlich: Hierbei müssen zwischen zwei Rangplätzen nicht zwangsläufig gleiche Leistungsabstände liegen. So kann in einem Schwimmwettbewerb zwischen dem ersten und zweiten Rang eine 1000-stel Sekunde liegen, während der dritte erst zwei Sekunden später ins Ziel gekommen ist. Dennoch liegt zwischen allen nur ein Rangplatz.

Die höchste Form der Skalierung stellt die *metrische* oder *kardinale Skalierung* dar. Sie enthält nicht nur die Informationen ordinaler Skalen mit den Aussagen *größer/kleiner*, *weniger/mehr* oder *besser/schlechter* ($x_i = x_j$; $x_i > x_j$; $x_i < x_j$), sondern es kann auch etwas über den Abstand der Merkmalsausprägungen zweier Merkmalsträger gesagt werden. Auf das Beispiel *Alter* bezogen heißt dies, dass eine 20-jährige Person nicht nur älter ist als eine 18-jährige Person, sondern es kann genau spezifiziert werden, dass sie um zwei Jahre älter ist. Zudem ist der Abstand zwischen einem 20- und einem 30-Jährigen genau so groß wie der Abstand zwischen einem 80- und einem 90-Jährigen. Die Abstände sind im metrischen Skalenniveau tatsächlich äquidistant. Typische Beispiele für metrische Skalen sind Alters-, Währungs-, Gewichts-, Längen- und Geschwindigkeitseinheiten.

In der Literatur werden die metrischen Skalen häufig nochmals unterschieden in Absolutskalen³, Verhältnisskalen⁴ und Intervallskalen⁵. Diese Unterscheidung ist dabei eher akademischer Natur, da sie für die Entscheidung, welches statistische Verfahren angewendet werden soll, in fast allen Fällen unerheblich ist.

Dies kann bei der Unterscheidung zwischen metrischen und ordinalskalierten Variablen nicht behauptet werden. Aufgrund der Verfahrensvielfalt für metrische Variablen im Vergleich zu den ordinalen Verfahren, versuchen Forscher gerne, eine eigentlich ordinalskalierte Variable als metrisch skaliert zu betrachten. So wird die empfundene Angebotsvielfalt für Butter – sie liegt auf einer fünfstufigen Skala zwischen sehr schlecht und sehr gut und ist somit ordinalskaliert – in der Praxis häufig als metrische Variable behandelt, indem die Forscher annehmen, dass der Abstand zwischen allen Rangplätzen identisch ist. Es wird in der empirischen Praxis also nicht selten von gleichen Abständen ausgegangen. In seriöseren Studien wird dabei zumindest in einem Nebensatz erwähnt, dass *Äquidistanz* unterstellt wird bzw. begründet, warum in diesem Fall davon ausgegangen werden kann. In der Literatur lassen sich zudem Bedingungen finden, unter denen auch bei Ordinalskalierung durchaus Verfahren für metrisch skalierte Variablen angewendet werden dürfen. Schmidt und Opp (1976, S. 35) formulieren als Daumenregel das Vorliegen einer ordinalskalierten Variablen mit mehr als vier Rangstufen und einer Stichprobe von mehr als 100 Beobachtungen. Auch Pell (2005) und Carifio und Perla (2008) diskutieren die Zulässigkeit metrischer Verfahren bei der Verwendung von Likert Skalen. Eine Interpretation einer Differenz von 0,5 zwischen zwei ordinalen Durchschnittswerten bleibt meines Erachtens dennoch schwierig und führt bei Statistikern nicht selten zu „Bauchschmerz“.

Die Darstellung der drei Skalenniveaus macht deutlich, warum es so wichtig ist, sich über das jeweils vorliegende *Skalenniveau* einer Variablen genau bewusst zu sein. Von diesem hängt letztlich auch *das zu wählende statistische Verfahren* ab: Man kann nämlich für die nominale Variable Beruf keinen Mittelwert aus drei Bäckern, fünf Metzgern und zwei Schornsteinfegern bilden. Im weiteren Verlaufe des Buches werden wir deshalb lernen, bei welchem Skalenniveau bzw. welcher Kombination von Skalenniveaus welches statistische Verfahren zulässig ist und welches nicht.

Zunächst müssen die vorliegenden Daten aber von der Papierform in eine für den Computer les- und verarbeitbare Form gebracht werden. Dies soll anhand des obigen Beispiels der 850 – durch die Studenten erhobenen – Fragebögen exemplarisch gezeigt werden.

³ Eine metrische Skala mit natürlichem Nullpunkt und natürlicher Einheit (z. B. *Alter*).

⁴ Eine metrische Skala mit natürlichem Nullpunkt ohne natürliche Einheit (z. B. *Flächen*).

⁵ Eine metrische Skala ohne natürlichen Nullpunkt und ohne natürliche Einheit (z. B. *geografische Längengrade*).

2.3 Datenerfassung mit dem Computer: Skalierung und Kodierung

Zunächst müssen die Skalenniveaus der Merkmale festgestellt werden. Bei der Auswertung einer bereits durchgeführten Befragung lassen sich Skalenniveaus im Nachhinein nicht erhöhen. Wird beispielsweise das Alter nicht in Jahren, sondern in Altersklassen abgefragt, hat diese Variable zwangsläufig ordinales Skalenniveau. Dies ist insofern „ärgerlich“, als damit das Durchschnittsalter nur mehr unter Annahmen bestimmt werden kann. Steht man noch vor der Durchführung einer Befragung, sollte man deshalb versuchen, immer das höchstmögliche Skalenniveau einer Variablen zu erlangen, das Alter also beispielsweise in Jahren zu erfragen, Ausgaben für bestimmte Konsumgüter in vollen Geldbeträgen, etc.

Aus Gründen der Anonymisierung von Befragungen können der Auftraggeber oder andere Beteiligte – z. B. der Betriebsrat bei Befragungen im Unternehmen – die Abfrage auf einem niedrigeren Skalenniveau verlangen, wenn hierdurch eine direkte Zurückführbarkeit eines Fragebogens auf eine bestimmte Person unmöglich wird. In diesem Fall muss sich der Forscher den Wünschen in der Regel beugen.

Im Rahmen der obigen Befragung liegen folgende Skalenniveaus vor:

- Nominal: *Geschlecht, verwendeter Brotaufstrich,*
- Ordinal: *Empfundene Angebotsvielfalt,*
- Metrisch: *Alter, Körpergewicht.*

Wie lassen sich nun die Informationen dem Computer mitteilen? Jedes Statistikpaket verfügt über ein Excel-ähnliches Arbeitsblatt, in das Daten direkt eingegeben werden können (vgl. beispielsweise Abb. 3.1.) Während in Excel die einzelnen Spalten mit A, B, C etc. benannt sind, können die *Spalten* in professionelleren Statistikpaketen direkt mit einem *Variablennamen* (engl.: variable name) bezeichnet werden. Als Variablennamen können häufig nur Folgen von bis zu acht Zeichen vergeben werden, so zum Beispiel der Variablenname *Angebot* für die Variable der Angebotsvielfalt. Um dennoch Klarheit zu schaffen, lassen sich diesen Variablennamen längere Textpassagen zuordnen, die sogenannten *variable labels*. Der Variablen *Angebot* könnte beispielsweise das Label *Empfundene Angebotsvielfalt* oder der gesamte Fragetext aus dem Fragebogen zugeordnet werden. Die Rechenbefehle werden mit den Variablennamen gesteuert – also z. B.: Berechne eine Grafik für die Variable *Angebot* – während im zugehörigen Ergebnisausdruck das verständlichere Label erscheint.

Nun gibt man die *einzelnen Ergebnisse der Befragung in den Zeilen* ein. Fragebogen 1 wird in der ersten Zeile erfasst, Fragebogen 2 in der zweiten etc. Ein Computer „versteht“ dabei nur Zahlenwerte. Für metrisch skalierte Variablen ist dies kein Problem, da die Variablenausprägungen in jedem Fall Zahlen darstellen. Person 1 ist 31 Jahre alt und wiegt rund 63 kg. Demnach können die Zahlen 31 und 63 in die entsprechenden Spalten eingetragen werden. Schwieriger gestaltet sich die Vorgehensweise bei nominalen Variablen. Bei der Variable *Geschlecht* lässt sich nun nicht einfach *männlich* oder *weiblich* eintragen.

Abb. 2.3 Kodierungsplan
(*engl.*: label book)

value label anbot
definition
1 sehr schlecht
2 schlecht
3 mittelmäßig
4 gut
5 sehr gut
variables: anbot
value label aufstrie
definition
0 butter
1 margarine
2 sonstiges
variables: aufstrie
value label gesch
definition
0 m
1 w
variables: gesch

Vielmehr müssen alle vorkommenden Informationen bzw. Inhalte kodiert werden, d. h. es muss jeder möglichen Ausprägung eine Zahl zugewiesen werden. So steht im Beispieldatensatz für die Ausprägung männlich die Zahl *Null* und für die Ausprägung weiblich die Zahl *Eins*. Die Zahlen sind nur Zuordnungen zu Inhalten. Gleiches gilt übrigens für ordinalskalierte Variablen. Übersichtshalber werden diese Zuordnungen, wie in Abb. 2.3 ersichtlich, in einem Kodierungsplan festgehalten. Anhand dieser Systematik kann eine vollständige Dateneingabe beginnen.

2.4 Fehlende Werte oder Missing Values

Eine bei der Auswertung von Datenmaterial sofort vergegenwärtigte „Fehlreaktion“ von Befragten ist die unvollständige Beantwortung von Fragebogenteilen bzw. das (vermehrte) Äußern von Meinungslosigkeit (*Ich weiß nicht*). Die Gründe hierfür sind vielfältig und gehen von bewussten Antwortverweigerungen, über fehlende Information bzw. fehlende Kompetenz zur qualifizierten Antwort, bis zur unentschlossenen oder fehlenden Meinung zu einem bestimmten Thema.

Faulkenberry und Mason (1978, S. 533 ff.) unterscheiden deshalb zwei Haupttypen von Antwortausfällen:

- a) *No-opinion*: Befragter ist (aufgrund z. B. von Ambiguität der Fragestellung, etc.) tatsächlich unentschlossen zu einer bestimmten Antwort.
- b) *Non-opinion*: Befragter besitzt tatsächlich keine Meinung zum jeweiligen Thema.

Die Autoren stellen fest, dass der erste Befragtentyp (*no-opinion*) im Vergleich zu anderen Antwortausfällen über mehr theoretisches Wissen sowie über eine durchschnittlich höhere Schulbildung verfügt. Darüber hinaus wirken Variablen wie Geschlecht, Alter und Herkunft des Interviewers etc. auf die Neigung zur Nichtbeantwortung von Fragen.

Letztlich liegt genau in diesem Punkt aber die Gefahr einer *systematischen Verzerrung*. Die Vermeidung der Gefahr systematischer Verzerrungen durch den Ausfall von Informationen über bestimmte Subpopulationen ist deshalb ein besonders zu beachtendes Problem. Einige Studien konnten zeigen, dass die Meinungslosigkeit um bis zu 30 Prozent höher liegt, wenn eine *Ich-Weiß-Nicht*-Kategorie als Antwortmöglichkeit vorgegeben ist (Schumann und Presser 1981, S. 117 ff.). Hieraus eine Strategie zur Senkung der Quote der Meinungslosigkeit abzuleiten, ist nicht nur äußerst fragwürdig, sondern ergebnisverzerrend. Die zur *Ich-Weiß-Nicht*-Kategorie neigenden Befragten sehen sich nicht zwangsläufig zu substantiell bzw. inhaltlich richtigen Antworten veranlasst, wenn diese Antwortmöglichkeit nicht vorgegeben ist. Eine zufällige bzw. gar keine Antwortkategorie sind nur einige „Ausweichstrategien“ für die Befragten. Es besteht somit die Gefahr, dass sich ein feststellbarer, systematischer Fehler von der *Ich-Weiß-Nicht*-Kategorie zu einem unentdeckten, systematischen Fehler der inhaltlichen Ebene transformiert (Schnell et al. 2013). Der Rat, *Ich-Weiß-Nicht*-Kategorien als mögliche Antwortkategorien nicht vorzusehen, kann unter diesen Gesichtspunkten schwer nachvollzogen werden. Wichtiger ist in diesem Zusammenhang die Frage, wie mit den Antwortausfällen innerhalb der Datenanalyse umgegangen werden soll.

Prinzipiell sollten Antwortausfälle keine Transformation in inhaltlich interpretierbare Werte erfahren, weshalb einige Analysemethoden die Weiterverarbeitung fehlender Werte nicht zulassen. Fehlende Werte können sogar zu einer Nichtberücksichtigung von Daten führen, für die Informationen vorliegen. Beispielsweise ist im Rahmen der Regressionsanalyse oder der Faktorenanalyse die Berücksichtigung eines Befragten unmöglich, wenn bei einer einzigen Variablen Angaben fehlen. Alle sonst vorhandenen Daten der Untersuchungseinheit bleiben ebenfalls unberücksichtigt. Da Antwortausfälle regelmäßig vorkommen und ein akzelerierter Verlust von Informationen ungen in Kauf genommen wird, liegt die *Substitution von Antwortausfällen* nahe. Es lassen sich fünf Ansätze abgrenzen:

- a) Lassen sich die fehlenden Merkmalsausprägungen *nachrecherchieren*, so ist dies mit Sicherheit die beste, aber wahrscheinlich auch die arbeitsaufwändigste Strategie, um fehlende Werte zu beseitigen. Nicht selten lassen sich in Unternehmensbefragungen mögliche fehlende Angaben zu Umsatz, F&E Ausgaben etc. durch sorgsame Studie von Sekundärliteratur (z. B. veröffentlichte Bilanzen) ermitteln.

- b) Handelt es sich bei den fehlenden Werten um qualitative bzw. nominalskalierte Variablen, kann ein fehlender Wert durch *Bildung einer entsprechenden Klasse* ersetzt werden. Ist bei einer Kundenbefragung beispielsweise für einen Teil der Befragten die Tatsache *bisher schon Kunde gewesen* und *bisher noch kein Kunde gewesen* bekannt und für einen anderen Teil der Befragten nicht, so kann für letzteren Teil die Klasse *Kundenbeziehung unbekannt* gebildet werden. Im Rahmen von Häufigkeitstabellen erscheint diese dann in einer gesonderten Zeile und kann zudem als *Missing Value* definiert werden. Selbst bei komplizierteren Verfahren wie z. B. der Regressionsanalyse ließen sich die fehlenden Werte in einer gewissen Weise inhaltlich interpretieren. Hierzu im Laufe des Buches mehr.
- c) Ist eine inhaltliche Ergänzung fehlender Werte nicht möglich, könnten fehlende *metrische Werte* durch den *Gesamtdurchschnitt* der Variablen ersetzt werden. Aus dem empirischen Datenmaterial lässt sich ein Gesamtmittelwert auf Basis der tatsächlich vorliegenden Messwerte bestimmen.
- d) Eine weitere Möglichkeit besteht in der Substitution von *metrischen Antwortausfällen* durch den *Mittelwert von Einzelgruppen*. So könnten fehlende Werte bei einer Befragung von Studierenden durch die Mittelwerte der einzelnen Studiengänge und nicht durch die Mittelwerte der gesamten Hochschule ersetzt werden.
- e) Allerdings darf der Beweis nicht unterbleiben, dass die vorausgesetzte Annahme der Unsystematik der Antwortausfälle erfüllt ist, da andernfalls mit schwerwiegenden Ergebnisverzerrungen zu rechnen ist. Selbst bei unsystematischen Antwortausfällen lassen sich die fehlenden Werte zwar relativ gut schätzen, allerdings können die Streuungen der Werte nur unterschätzt und die Zusammenhänge nur verzerrt berechnet werden: „In particular, variances from filled-in data are clearly understated by imputing means, and associations between variables are distorted. Thus, the method yields an inconsistent estimate of the covariance matrix“ (Roderick et al. 1995, S. 45). Der Einsatz komplizierter Schätzverfahren wird somit unausweichlich, wenn die Anzahl fehlender Werte so groß wird, dass deren Ersetzung durch Mittelwerte zu einer signifikanten Veränderung von statistischen Kennziffern führt. Diese Verfahren basieren vor allem auf der Idee *regressionsanalytischer Schätzungen* der fehlenden Werte mit Hilfe von anderen vorhandenen unabhängigen Variablen im Datensatz. Fehlen zum Beispiel bei einer Unternehmensbefragung teilweise die Angaben zu den Ausgaben für F&E, gleichzeitig ist aber bekannt, dass diese von der Branche, der Größe und der Herkunft (Ost-/Westdeutschland) des Unternehmens abhängen, so kann dieser Zusammenhang mit Hilfe der vorliegenden Daten bedingt geschätzt und auf die fehlenden Werte übertragen werden. Die Vorgehensweise der Regressionsanalyse wird in Kap. 5 genauer beschrieben.

Insgesamt sollte mit dem Problem der nachträglichen Ergänzung fehlender Werte sorgsam umgegangen werden. In jedem Fall sollten – wenn möglich – die unterschiedlichen Formen fehlender Werte unterscheidbar bleiben. So kann in einem Telefoninterview beispielsweise unterschieden werden,

- ob der Befragte nicht antworten konnte, weil er die Antwort nicht wusste;
- ob der Befragte nicht antworten wollte, obwohl er die Antwort wusste, diese aber nicht mitteilen wollte oder
- die Frage keine Relevanz für den Befragten hatte, da die Frage nur einer anderen Altersgruppe gestellt wurde.

Während im letzten Fall die Werte häufig einfach ausgelassen werden (systembedingte fehlende Werte), werden für die beiden ersten Fälle zwar Werte vergeben, die aber in der Software wiederum als fehlende Werte definiert werden.

2.5 Ausreißer und offensichtlich falsche Werte

Ein dem Problem der fehlenden Werte ähnliches ist das der offensichtlich falschen Werte. Bei standardisierten Kundenbefragungen erhält man bei der Frage nach dem Einkommen häufig nicht nur fehlende Werte, sondern ebenfalls fantastische Einkommen wie 1.000.000.000 € bei gleichzeitiger Berufsangabe *Arbeitsloser*. Stellt man sich auf den Standpunkt, dass jede Zahl auch so verwendet werden muss, wie sie im Fragebogen ausgefüllt wurde, würde dies für eine Befragung von 500 Personen eine Erhöhung des Durchschnittseinkommens um 2.000.000 € allein aufgrund der obigen Angabe bedeuten. Letztlich muss es also erlaubt sein, *offensichtlich falsche Angaben korrigieren zu dürfen*. Im obigen Fall einer offensichtlichen und gleichzeitig willentlichen Irreführung durch den Befragten kann der entsprechende Wert auf einen fehlenden Wert gesetzt oder durch entsprechende Verfahren ein Schätzwert berechnet werden (vgl. Abschn. 2.4).

Die offensichtlich falschen Werte entstehen dabei nicht immer durch willentlich falsches Ausfüllen, sondern vor allem durch Irren des Befragten. So kann bei Unternehmensbefragungen nicht selten festgestellt werden, dass Umsätze in 1000 € erfragt, letztlich doch in vollen Eurobeträgen angegeben und so irrtümlicherweise Umsätze in tausendfacher Höhe ausgewiesen werden. Auch solche Fehler müssen – sollten sie festgestellt werden – selbstverständlich nachträglich korrigiert werden.

Schwieriger ist der Umgang mit offensichtlich falschen, aber nicht nachzurecherchierenden Informationen. Befragt man wiederum Unternehmen – bei der Befragung von Privatpersonen ergeben sich übrigens ähnliche Probleme – nach ihrer anteilmäßigen Aufschlüsselung von Ausgabekategorien, so entstehen bei Fragebögen nach Addition aller Anteile regelmäßig Prozentsätze von mehr als 100 Prozent.

Ähnlich „verzwick“ ist das Problem, wenn man mit Bestimmtheit weiß, dass ein Wert zwar stimmt, es sich bei diesem aber um ein singuläres Ereignis (singulären Ausreißer) handelt. So stellte sich in einem Unternehmen die Personalabteilung die Frage nach dem durchschnittlichen Renteneintrittsalter der Beschäftigten, um auf dieser Basis die Betriebsrente kalkulieren zu können. Aus diesem Grund wurde der entsprechende Durchschnittswert der letzten Jahre ermittelt. Einer der betroffenen Rentenempfänger war der Gründer des Familienunternehmens, der erst im Alter von fast achtzig Jahren aus

dem Betrieb ausgeschieden ist. Obwohl es sich bei ihm um eine reale Beobachtung der Zielgruppe der zu verrentenden Mitarbeiter handelte, würde die Berücksichtigung dieses Wertes die Berechnung des durchschnittlichen Rentenalters wohl verzerren, da künftig kaum ein weiteres Mal ein Unternehmensgründer mit diesem Alter ausscheiden würde. Unter bestimmten Umständen ist es daher sinnvoll, *selbst „stimmende“ Ausreißer aus der Analyse auszuschließen*. Die Entscheidung, wann dies der Fall ist, kann nur inhaltlich getroffen werden. Als Lösung kann das sogenannte *Trimmen* des Datensatzes dienen, bei dem jeweils fünf Prozent der größten und fünf Prozent der kleinsten Beobachtungen aus dem Datensatz entfernt werden. Wir werden hierauf später (vgl. Abschn. 3.2.2) noch eingehen.

2.6 Übungsaufgaben zum Abschnitt

Aufgabe 1

Geben Sie für die folgenden Fragestellungen Merkmalsträger, Merkmal und Merkmalsausprägungen an:

- a) die Todesursache von Patienten,
- b) Studiendauer von Studierenden,
- c) Alkoholgehalt eines Getränkes.

Aufgabe 2

Geben Sie zu folgenden Merkmalen jeweils das geeignete Skalenniveau an:

- a) Nebentätigkeit von Studenten,
- b) Marktanteil eines Produktes zwischen 0 % und 100 %,
- c) Studiengangszugehörigkeit von Studierenden,
- d) Uhrzeit,
- e) Blutalkoholgehalt,
- f) Benzinverbrauch eines PKW,
- g) Intelligenzquotient,
- h) Güte eines Restaurants in Sternen ausgedrückt.

Aufgabe 3

Bereiten Sie Stata, SPSS oder Excel für den in Abb. 2.1 dargestellten Fragebogen vor und geben Sie die Daten aus Abb. 3.1 ein. Sehen Sie ebenfalls *Missing Values* vor.

Literatur

- Bühner M (2010) Einführung in die Test- und Fragebogenkonstruktion, 3. Aufl. Pearson, München
- Carifio J, Perla R (2008) Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42:1150–1152
- Faulkenberry GD, Mason R (1978) Characteristics of nonopinion and no opinion response groups. *Public Opinion Quarterly* 42:533–543
- Pell G (2005) Use and misuse of Likert scales. *Medical Education* 39:970
- Roderick JA, Little, Schenker N (1995) Missing Data. In: Arminger G, Clogg CC, Sobel ME (Hrsg) *Handbook of Statistical Modelling for the Social and Behavioral Sciences*. Plenum Press, London/New York, S 39–75
- Schmidt P, Opp K-D (1976) Einführung in die Mehrvariablenanalyse. Rowohlt, Reinbek/Hamburg
- Schnell R, Hill PB, Esser E (2013) Methoden der empirischen Sozialforschung, 10. Aufl. Oldenbourg, München und Wien
- Schumann H, Presser S (1981) Questions and Answers in Attitude Surveys. Academic Press, New York

Deskriptive Statistik und Explorative Datenanalyse
Eine computergestützte Einführung mit Excel, SPSS und
STATA

Cleff, Th.

2015, XVI, 264 S. 123 Abb., Softcover

ISBN: 978-3-8349-4747-5