

# 1 Die Aufbereitung der Daten für die statistische Analyse

Bevor Daten ausgewertet werden können, müssen sie zunächst so aufbereitet werden, dass sie mit statistischen Analyseprogrammen bearbeitet werden können. Hat man etwa eine Face-to-face-Befragung<sup>1</sup> – also eine mündliche persönliche Befragung – durchgeführt und die Antworten in einem Papierfragebogen notiert, so muss man nun den Transfer der Daten von den vielen einzelnen Fragebögen in eine einzige übersichtliche Datentabelle organisieren. Eine solche Datentabelle für die statistische Analyse besitzt einen rechteckigen Aufbau und sieht im Prinzip folgendermaßen aus:

Tab. 1-1: Die Datentabelle als Ergebnis der Datenaufbereitung

ID	Geschlecht	Note	Religion	Zufriedenheit	Beruf
...					
101	m	3,2	2	2	Lehrer
102	w	2,1	1	3	Ärztin
103	m	2,3	2	3	Schreiner
104	w	3,2	3	4	Pädagogin
105		1,9	1	0	Anwalt
106	m	2,9	3	2	Verkäufer
107	m	1,6	0	1	Krankenpfleger
...					

Die erste Zeile enthält die Namen der Variablen, hier z.B. Geschlecht, Note etc. Die fertige Datentabelle, die häufig auch Datenmatrix genannt wird, besteht aus n Zeilen, also genau so vielen Zeilen, wie es Befragte gibt, und m Spalten, d.h. so vielen Spalten, wie der Fragebogen Fragen enthält bzw. um es genau zu formu-

<sup>1</sup> Im Fall von Online-Befragungen muss man sich viele der hier folgenden Überlegungen bereits vor der Datenerhebung machen (vgl. hierzu z.B. Kuckartz u.a. 2009).

lieren: so viele Spalten, wie Variablen definiert werden müssen, um die Befragung adäquat auswerten zu können.

#### **Was ist eigentlich eine Variable?**

Der Begriff „Variable“ wird in den Sozialwissenschaften für ein Merkmal oder eine Eigenschaft verwendet. Eine Variable besitzt verschiedene Ausprägungen, z.B. hat das „Geschlecht“ die Ausprägungen „männlich“ und „weiblich“ und die Ausprägungen der Variable „Alter in Jahren“ sind die Jahre. Häufig werden die Begriffe „Variable“ und „Merkmal“ synonym verwendet.

In der ersten Spalte der oben dargestellten Datentabelle (Tab. 1-1) steht eine Identifikationsnummer (Spaltenbenennung „ID“), die es ermöglichen soll, schnell auf den Originalfragebogen zurückzugreifen. Wenn auf den zu erfassenden Fragebögen nicht bereits eine eindeutige Kennung abgedruckt war, muss man also vor der Dateneingabe einen Stift zur Hand nehmen und alle ausgefüllten Fragebögen mit einer laufenden Nummer versehen. Eine solche Identifikationsnummer ist vor allem dann wichtig, wenn sich später bei der Kontrolle der eingegebenen Daten herausstellt, dass offenbar ein Eingabefehler vorliegen muss, weil die Datentabelle z.B. Variablenwerte enthält, die es aufgrund des Codeplans gar nicht geben kann oder die sehr unwahrscheinlich sind (Alter = 200 Jahre, 20-Jährige mit sieben Kindern etc.).

## **1.1 Der Codeplan**

Wenn man die Daten in Tabelle 1 näher betrachtet, wird man höchstwahrscheinlich die Tabellenwerte der Variablen „Geschlecht“ intuitiv mit den tatsächlichen Ausprägungen der Variable in Verbindung bringen. Man vermutet wohl zurecht, dass der Code „m“ männlich bedeutet und es sich bei ID = 101 um einen männlichen Befragten handelt. Dementsprechend bedeutet die Eingabe „w“, dass die befragte Person „weiblich“ ist. Anders verhält es sich bei der vierten Spalte, die Angaben über die Religionszugehörigkeit enthält. Diese ist hier nicht im Klartext eingetragen, sondern wir finden dort nur Zahlenangaben, die wir nicht direkt in Verbindung mit den möglichen Ausprägungen „katholisch“, „evangelisch“, „keine Religionszugehörigkeit“ etc. bringen können. Hier bedarf es also einer entsprechenden Korrespondenztabelle, in der die Bedeutung eines Variablenwertes eindeutig festgelegt wird. Eine solche Korrespondenztabelle bezeichnet man auch als *Codeplan*, *Codierschema* oder englisch als *Codebook*. Betrachten wir Tab. 1-2 als Beispiel:

Tab. 1-2: Beispiel-Codeplan für sechs Variablen

Variablenname	Variablenlabel	Wertelabel	Variablentyp	Skalenniveau
ID	Identifikationsnummer	Wert eingeben	Numerisch	nominal
Geschlecht	Geschlecht	m = männlich w = weiblich 0 = k.A.	String	nominal
Note	Durchschnittsnote im Abitur	Wert eingeben 0 = k.A.	Numerisch, 1 Dezimalstelle	intervall
Religion	Religionszugehörigkeit	1 = katholisch 2 = protestantisch 3 = nicht-christl. Religion 0 = k.A.	Numerisch	nominal
Zufriedenheit	Zufriedenheit mit Einkommenshöhe	1 = sehr zufrieden 2 = eher zufrieden 3 = eher nicht zufrieden 4 = sehr unzufrieden 0 = k.A.	Numerisch	ordinal
Beruf	Welchen Beruf üben Sie aus?	Text eingeben	String	nominal

Der Codeplan ordnet den Fragen und Teilfragen eines Fragebogens *Variablennamen* (engl. *Variable Names*) und den möglichen Ausprägungen einer Variablen *Wertelabels* (*Value Values*) zu. Im folgenden Abschnitt werden die einzelnen Spalten von Tab. 1-2 erläutert:

#### *Variablenname*

Die Variablennamen dienen einem Statistikprogramm zur eindeutigen Identifizierung der einzelnen Variablenspalten. Variablennamen bestehen aus einem Wort bzw. einer Zeichenkette ohne Leerzeichen. Wenn der Fragebogen kurz ist und nur wenige Variablen zu definieren sind, kann man wie in Tab. 1-2 eine einfache, gegebenenfalls verkürzte Klartextbezeichnung verwenden. In den meisten Codeplänen enthält der Variablenname jedoch die Nummer der korrespondierenden Frage im Fragebogen, also bspw. „F1“ oder „v23“, wobei das „F“ als Abkürzung für Frage und das „v“ für Variable stehen. Diese Methode hat insbesondere bei längeren Fragebögen den Vorteil, dass man leichter den Überblick behält und im Statistikprogramm auf gesuchte Variablen schnell zugreifen kann. Zu beachten ist ferner, dass Statistikprogramme häufig nur bestimmte Zeichenkombinationen als Variablennamen erlauben (die Vorschriften werden im Detail weiter unten dargestellt).

### *Variablenlabel*

In das Feld „Variablenlabel“ lässt sich für jede Variable eine detaillierte Beschreibung der Variablen eintragen. Während die Variablennamen meist Beschränkungen unterliegen, hat man bei der Definition von Variablenlabels „freie Hand“. So bietet es sich an, die vollständige Formulierung der Frage aus dem Fragebogen als Variablenlabel zu übernehmen. Statistikprogramme benutzen das Variablenlabel später bei der Ausgabe zur Beschriftung von Tabellen und Grafiken.

### *Wertelabel*

Die dritte Spalte ist für die Ausprägungen der jeweiligen Variablen vorgesehen. Es wird festgehalten, wie die Antworten auf eine Frage in die Datenmatrix eingegeben werden. Enthält eine Frage im Fragebogen Antwortvorgaben (z.B. männlich und weiblich), so informiert die Spalte Wertelabel darüber, mit welcher Zahl oder mit welchem Zeichen die unterschiedlichen Antwortmöglichkeiten in der Datenmatrix erfasst werden.

### *Variablentyp*

Beim Anschauen der Tab. 1-1 stellt man auf den ersten Blick fest, dass es offenbar Variablen verschiedenen Typs geben kann, bspw. findet man in der Spalte Geschlecht nur einzelne Buchstaben („w“ und „m“) und in der Spalte Beruf ganze Wörter („Lehrerin“). Solche Variablen, die nicht nur Zahlen, sondern auch Buchstaben und andere Zeichen enthalten, bezeichnet man als Zeichenketten- oder Stringvariable. Die Variablen Religion und Zufriedenheit enthalten hingegen ganze Zahlen und die Variable Note (das ist die Durchschnittsnote der Befragten im Abitur) enthält Zahlen mit Nachkommastellen. Derartige Variablen werden als „numerisch“ bezeichnet.

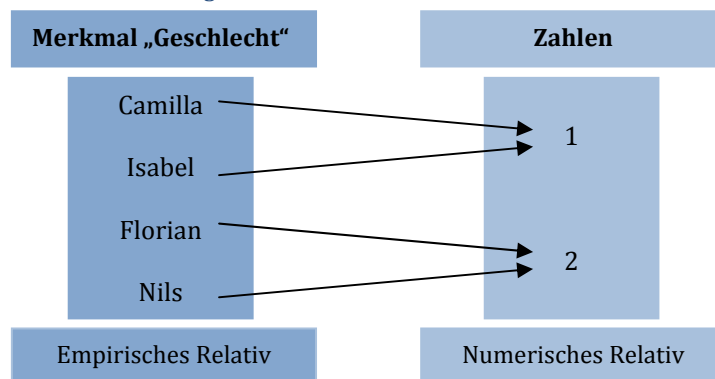
### *Skalenniveau*

Welche Operationen man mit Variablen durchführen kann, hängt von ihrem Skalenniveau (Messniveau) ab. Man unterscheidet zwischen Nominalskala, Ordinalskala und Intervallskala<sup>2</sup>, wobei die Nominalskala das geringste Skalenniveau und die Intervallskala das höchste aufweisen. Beim Vorgang des Messens werden den Merkmalen des empirischen Relativs Zahlen so zugeordnet, dass die ursprünglichen Relationen möglichst erhalten bleiben. Für das einfachste Skalenniveau, die *Nominalskala*, sieht dies etwa wie in Abb. 1-1 aus.

---

2 Die vierte Skalenart, die Verhältnisskala oder Ratioskala, spielt in der empirischen Forschung so gut wie keine Rolle und wird hier nicht berücksichtigt.

Abb. 1-1: Veranschaulichung der Nominalskala



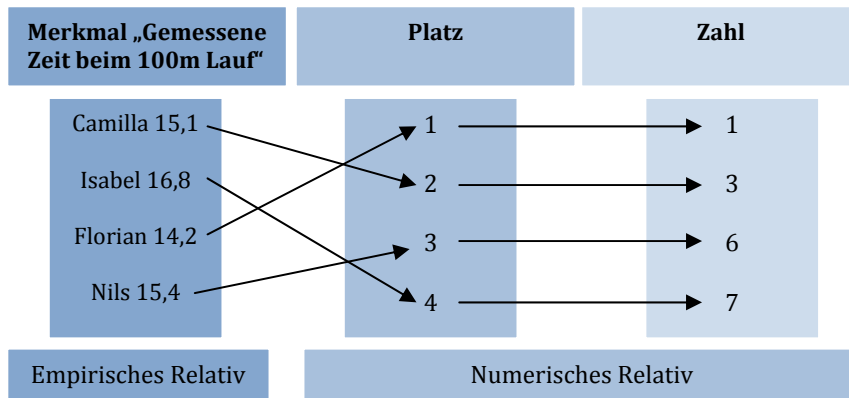
Anstelle der Werte „1“ und „2“ hätte man auch andere Werte zur Bezeichnung des Geschlechts wählen können (etwa „7“ und „12“). Für Variablen mit Nominalskalenniveau lassen sich nur Aussagen über Gleichheit bzw. Ungleichheit treffen. Jemand ist männlich oder weiblich; deutscher, italienischer, französischer oder anderer Nationalität. Dementsprechend ist es auch ohne Belang, welche Zahlen man den verschiedenen Ausprägungen einer nominalskalierten Variable zuordnet, ob man bei der Frage nach der Parteipräferenz der SPD „1“ oder „4“ zuordnet, spielt keine Rolle.

Anders verhält es sich bei der *Ordinalskala*. Hier ist es erforderlich, dass im empirischen Relativ eine Ordnungsrelation besteht und diese Relation muss bei der Zuordnung von Zahlen erhalten bleiben. Beispiele für ordinalskalierte Variablen sind Gehaltsstufen, Bildungsabschlüsse, soziale Schicht und alle Arten von Rangfolgen. Sind Objekte äquivalent – etwa Personen, die beide nach der Gehaltsgruppe TVÖD 13 bezahlt werden – erhalten sie eine identische Zahl zugeordnet. Die schematische Darstellung in Abb. 1-2 verdeutlicht, dass die Information über den Rangplatz (Platzierung) auch erhalten bleibt, wenn anstelle von „1“, „2“, „3“ und „4“ die Zahlen „1“, „3“, „6“ und „7“ zugeordnet werden.

Die *Intervallskala* erlaubt nicht nur Aussagen über die Rangfolge von Objekten, sondern auch über die Größe ihrer Abstände. Während ich bei der Rangskala den zugeordneten Werten nur entnehmen kann, dass Camilla vor Isabel und Florian vor Nils ins Ziel gekommen ist, sind die Werte einer Intervallskala so zugeordnet, dass gleiche Zahlendifferenzen zwischen zwei Objekten gleichen Merkmalsunterschieden entsprechen. Erhalten bspw. unsere vier Läufer/innen die von ihnen benötigte Zeit in Sekunden zugeordnet, so lassen sich auch Aussagen über die Abstände von je zwei Objekten formulieren („Der Abstand zwischen Camilla und Isabel ist größer als der zwischen Florian und Nils“). Beispiele für intervallskalierte Variablen sind „Zeit, die für Zusammenlegen eines Puz-

zles benötigt wurde“, „Zahl der Kinder“, „Einkommen in Euro“, „Entfernung von Wohnung zur Arbeit in Kilometern“. Anstelle von Intervallskalenniveau ist häufig auch vom metrischen Skalenniveau die Rede.

Abb. 1-2: Veranschaulichung der Ordinalskala



Zu unterscheiden sind ferner *stetige* (kontinuierliche) und *diskrete Variablen*: Bei stetigen Variablen existieren im Prinzip zwischen zwei Werten unendliche viele Zwischenwerte (Beispiel: Zeitmessung), während bei diskreten Variablen die Werte abzählbar sind und keine Zwischenwerte auftreten können (Beispiel „Zahl der Kinder“).

Mit dem Begriff *kategoriale Variablen* werden üblicherweise neben nominalskalierten auch ordinalskalierte Variablen mit relativ wenigen Ausprägungen bezeichnet. Anstelle von Ausprägungen ist dann meistens von Kategorien die Rede. In manchen Statistikprogrammen, wie etwa in SYSTAT, wird nur zwischen intervallskalierten und kategorialen Variablen unterschieden.

Häufig findet man in der Literatur auch den Begriff *dichotome Variable*. Dies ist eine Variable mit lediglich zwei Ausprägungen, wobei es sich um eine natürliche oder konstruierte Dichotomie handeln kann. Das Merkmal „Geschlecht“ ist ein Beispiel für eine natürliche Dichotomie, während andere Dichotomien wie etwa die dichotome Variable „Einkommen“ von den Forschenden konstruiert werden, indem ein Schwellenwert definiert wird, der die Werte in lediglich zwei Gruppen aufteilt (erste Gruppe: Einkommen über dem Durchschnitt; zweite Gruppe: Einkommen unter dem Durchschnitt). Prinzipiell können Variablen eines bestimmten Skalenniveaus in solche eines geringeren Skalenniveaus transformiert werden. So kann die intervallskalierte Variable „Einkommen“ nicht nur wie beschrieben in eine dichotome (=nominalskaliert), sondern auch in eine

ordinalskalierte Variable verwandelt werden, indem die Probanden gemäß ihrem Einkommen in eine Rangreihe gebracht werden.

Das Skalenniveau determiniert die Art von möglichen mathematischen Operationen und damit auch die statistischen Verfahren, die mit den so skalierten Variablen durchführbar sind (vgl. Tab. 1-3). So ist es offenkundig unsinnig, einen Mittelwert der Religionszugehörigkeiten zu berechnen, obwohl die Spalte in der Tab. 1-3 nur Zahlenangaben enthält und theoretisch – anders als bei der Stringvariable „Geschlecht“ – die Berechnung eines Mittelwerts denkbar wäre.

Tab. 1-3

Skalenniveau	Erlaubte Operationen	Beispiele
nominal	$a = b$ $a \neq b$	Geschlecht, Beruf, Parteipräferenz, Studienfach, Religionszugehörigkeit
ordinal	$a < b$ $a > b$	Gehaltsstufe, militärischer Rang, Rangliste der besten Freunde/Freundinnen
intervall	$a - b = c - d$	Zahl der Kinder, Einkommen, Durchschnittsnote im Abitur, Zahl der Elektrogeräte im Haushalt,

Die Variablen der Tab. 1-2 besitzen also folgendes Skalenniveau:

Variablenname	Skalenniveau
ID	nominal (zur Identifikation von Fragebögen)
Geschlecht	nominal (dichotom)
Note	intervall
Religion	nominal
Zufriedenheit	ordinal
Beruf	nominal

An der unterschiedlichen Codierung der Variablen Geschlecht, Beruf und Religion lässt sich erkennen, dass nominalskalierte Variablen sowohl als Stringvariable als auch als numerische Variable codiert werden können. Während das Geschlecht in unserem Beispiel als Stringvariable definiert wurde, ist die Religionszugehörigkeit als numerische Variable definiert.

In der Regel ist es bei Benutzung von statistischen Analyseprogrammen vorzuziehen, numerische Variable zu verwenden, obwohl deren Ausprägungen wie gesehen nicht ohne Hinzuziehen einer Korrespondenztabelle interpretiert werden können.