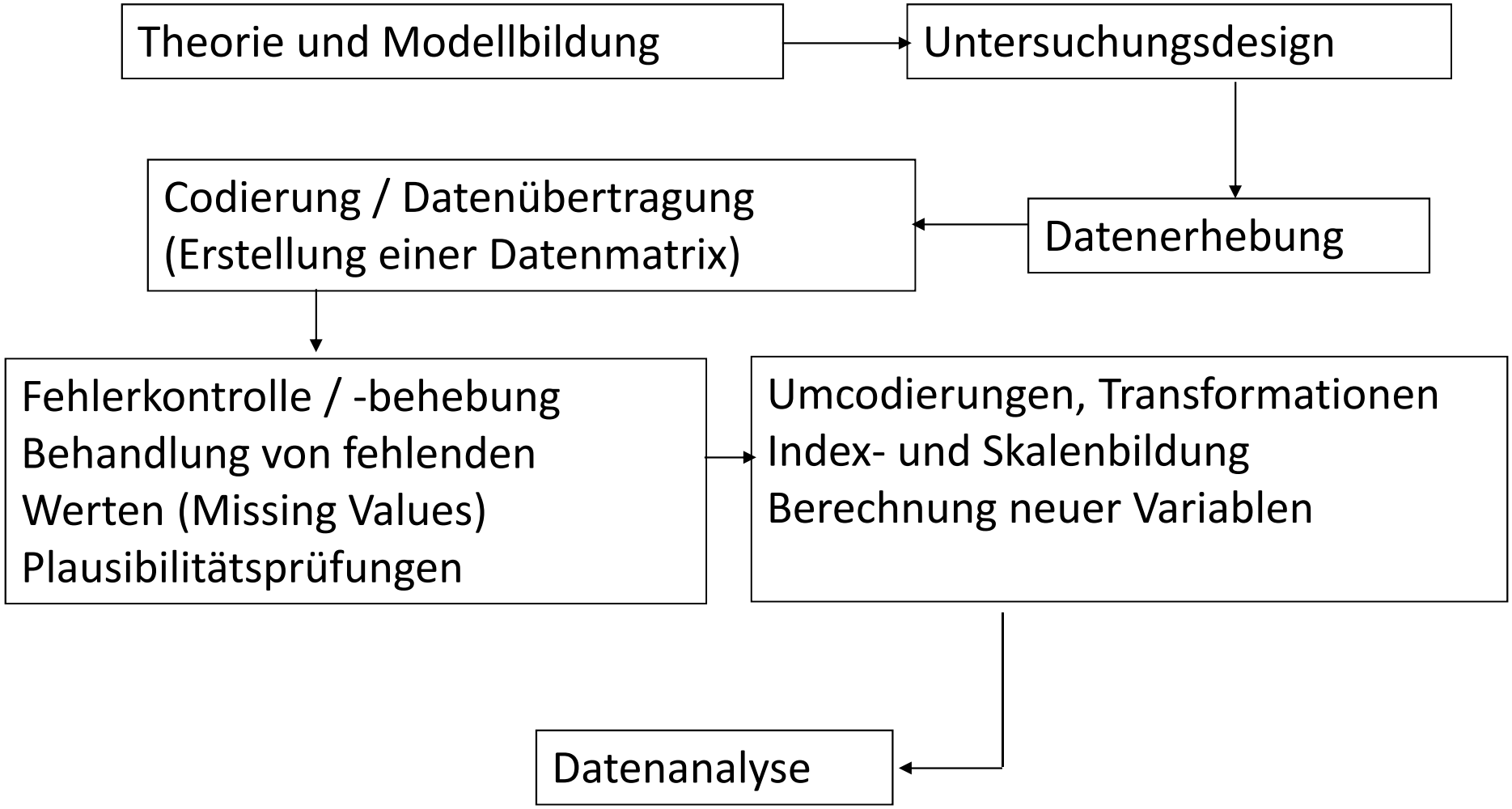
Grundlagen der Datenanalyse

Schematischer Überblick zur Behandlung quantitativer Daten



Variable

Messbare Eigenschaft (Merkmal) von Objekten, die verschiedene Ausprägungen (Merkmalsausprägungen) annehmen kann.

Ausprägungen müssen

- *disjunkt* und
- *erschöpfend* sein

Disjunkt bedeutet, dass die Merkmalsausprägungen einander ausschließen, d.h. sich nicht *überlappen* dürfen.

Erschöpfend heißt, dass alle möglichen Ausprägungen einer Variablen bekannt sein müssen, sodass *jeder* Merkmalsträger einer Kategorie zugewiesen werden kann.

Da eine Variable bei unterschiedlichen Objekten (i.d.R. Personen) unterschiedliche Ausprägungen annehmen, weist sie Streuung (Varianz) auf. Wäre das nicht der Fall, dann würde es sich um eine Konstante handeln.

Beispiel: Folgende Frage wäre bei direkter Abbildung als Variable nicht erschöpfend und nicht disjunkt:

Wie hoch ist Ihr monatliches Nettoeinkommen ?

- ❒ bis 1.000 €
- ❒ 1.000 – 1.200 €
- ❒ 1.200 – 1.500 €

Eine Person mit einem Einkommen von mehr als 1.500 € kann nicht zugeordnet werden (= da das Kategoriensystem unvollständig ist), während eine Person mit einem Verdienst von genau 1.200 € gleich in zwei Kategorien fällt (da sich das Kategoriensystem überlappt).

Variable – Eigenschaften

Variable können

- ***diskret*** (= in Schritten) oder
- ***kontinuierlich***

messen

Nach der *Anzahl* der Ausprägungen kann zwischen

- ***dichotomen*** (= 2 Ausprägungen) und
- ***polytomen*** (= mehr als 2 Ausprägungen)

Variablen unterschieden werden

Messen und Messniveaus

Messen heißt die nach bestimmten Regeln vollzogene Zuordnung von **Merkmalsträgern** zu beobachteten **Merkmalsausprägungen** auf den zu untersuchenden **Merkmalsdimensionen**.

Der Informationsgehalt von Daten wird wesentlich durch das ***Mess-* oder *Skalenniveau*** festgelegt, mit dem die Merkmalsausprägungen der Untersuchungsobjekte gemessen werden.

Nominales Messen

Klassifizieren von Untersuchungsobjekten hinsichtlich ihres Besitzes oder Nichtbesitzes einer bestimmten (qualitativen) Merkmalsausprägung.

Die Ausprägungen schließen sich nur logisch aus (Beispiele: Geschlecht, Studienrichtung, Wohnort, ...)

Mögliche Vergleiche zwischen Objekten: gleich/ungleich

Ordinales Messen

Die Ausprägungen lassen sich in eine Rangordnung bringen. Gemessen wird auf diesem Niveau die Intensität, Stärke oder Größe, mit der eine bestimmte Eigenschaft bei den einzelnen Untersuchungsobjekten auftritt: z.B. Schulnote, politisches Interesse

Es lässt sich also nur angeben, dass das Untersuchungsobjekt A größer als B ist, aber man kann nicht sagen, um wie viel größer es ist.

Mögliche Vergleiche: gleich/ungleich, größer/kleiner

Metrisches Messen

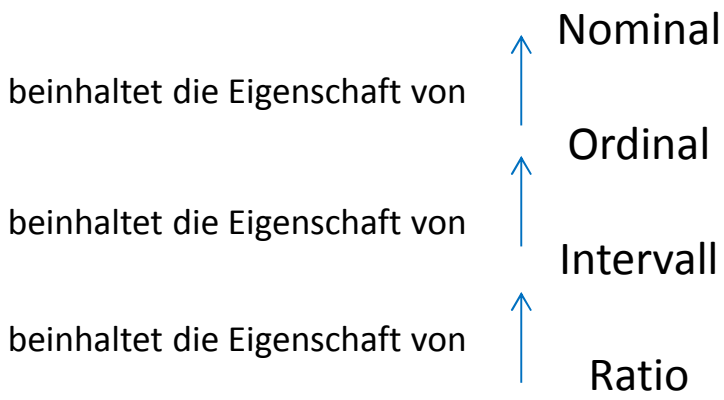
Intervallskala: Abstände bekannt, aber kein natürlicher Nullpunkt (z.B. Celsiusskala)

Mögliche Vergleiche/Operationen: gleich/ungleich, größer/kleiner, Addition/Subtraktion

Ratioskala: Abstände bekannt und natürlicher Nullpunkt (Einkommen, Körpergröße, ...)

Mögliche Vergleiche/Operationen: gleich/ungleich, größer/kleiner, Addition/Subtraktion, Multiplikation/Division

Hierarchie der Messniveaus



Da die möglichen Vergleiche und Rechenoperationen vom Messniveau abhängen, sind auch die Analysemöglichkeiten durch das Messniveau determiniert.

Aus diesem Grund ist jeweils hohes Messniveau bei der Messung anzustreben!

In den Sozialwissenschaften werden häufig Ratingformate zur Messung verwendet.

Beispiele:

trifft überhaupt nicht zu	gar nicht	lehne vollständig ab
trifft eher nicht zu	sehr wenig	lehne ab
trifft gelegentlich zu	wenig	lehne eher ab
trifft eher zu	etwas	stimme eher zu
trifft voll und ganz zu	viel	stimme zu
	sehr viel	stimme vollständig zu

Diese Formate entsprechen natürlich streng genommen ordinalem Messniveau.

Häufig werden diese jedoch in der Praxis wie metrische Messungen behandelt.

Darüber besteht jedoch nicht in allen Wissenschaftsbereichen Konsens.

Argumentiert werden könnte damit, dass diese Ratingskalen so konstruiert wurden, dass die Abstände zwischen den Ausprägungen gleich erscheinen (bzw. das Ausprägungsspektrum in gleichgroße Bereiche geteilt wird.

Dichotome Merkmale und Messniveau

Betrachten wir beispielsweise die Variable Geschlecht mit den Ausprägungen männlich und weiblich.

Das Merkmal hat natürlich grundsätzlich nominalen Charakter (es besteht keine Ordnungsrelation zwischen den Ausprägungen).
Wir können somit auch beliebige Zahlenwerte dafür vercoden (z.B 1=männlich, 2=weiblich oder 0=weiblich, 1=männlich etc.)

Betrachten wir die Variante: 0=weiblich, 1=männlich:

Wenn wir die Variable nun nicht Geschlecht sondern „männliches Geschlecht“ nennen, so könnte man (zurecht) behaupten, dass die Variable ordinales Messnivau aufweist: Eine höhere Ausprägung korrespondiert mit stärkerer „männlicher Eigenschaft“, eine niedrigere Ausprägung korrespondiert mit geringerer „männlicher Eigenschaft“.

Die Ausprägungen bilden somit eine ordinale Rangreihe ab.

Da die Variable zudem auch nur eine mögliche Einheit umfasst (die Distanz von 0 zu 1), spricht auch nichts gegen die Annahme, dass es sich sogar um eine intervallskalierte Messung des Merkmals handelt.

Fazit: Dichotome Variable können im Rahmen der meisten Auswertungsverfahren wie metrische (intervallskalierte) Variablen behandelt werden.

Dafür ist es zwar grundsätzlich egal, wie die Variable codiert wurde (z.B. 1, 2 oder 3,10…), die Interpretation der Auswertungsergebnisse ist jedoch manchmal besser interpretierbar, wenn sie mit 0, 1 codiert ist.

z.B. wir können den Durchschnitt (arithmetisches Mittel) einer dichotomen Variable für folgende Messreihe berechnen (das Verfahren setzt metrisches Messniveau voraus):

0, 1, 1, 1, 0, 0, 0, 1, 1, 0

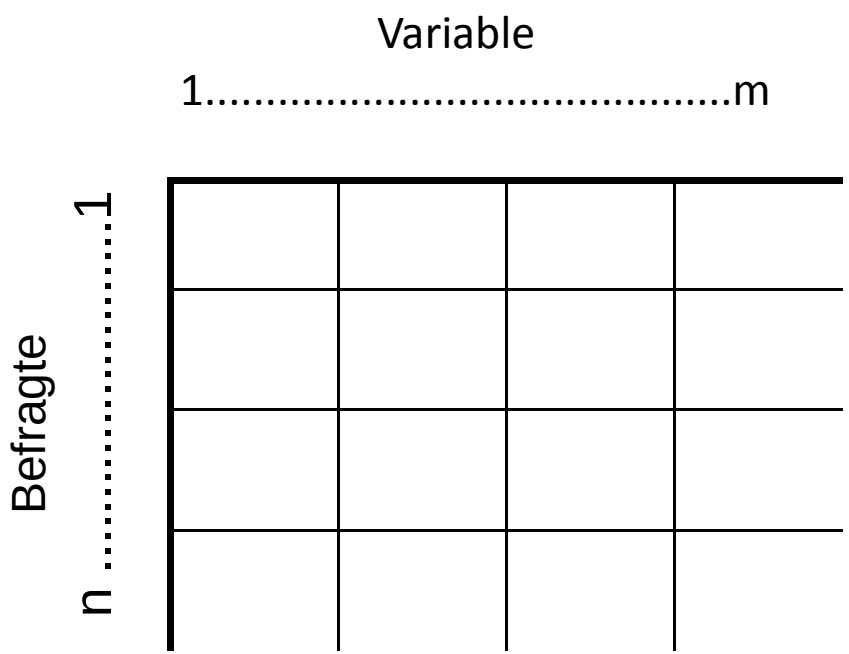
Mittelwert = 5/10 = 0.5. Bei 0/1 codierten Variablen entspricht dieser Durchschnitt dann der relativen Häufigkeit der 1-Ausprägung.

Arbeitsschritte einer computerunterstützten Datenauswertung

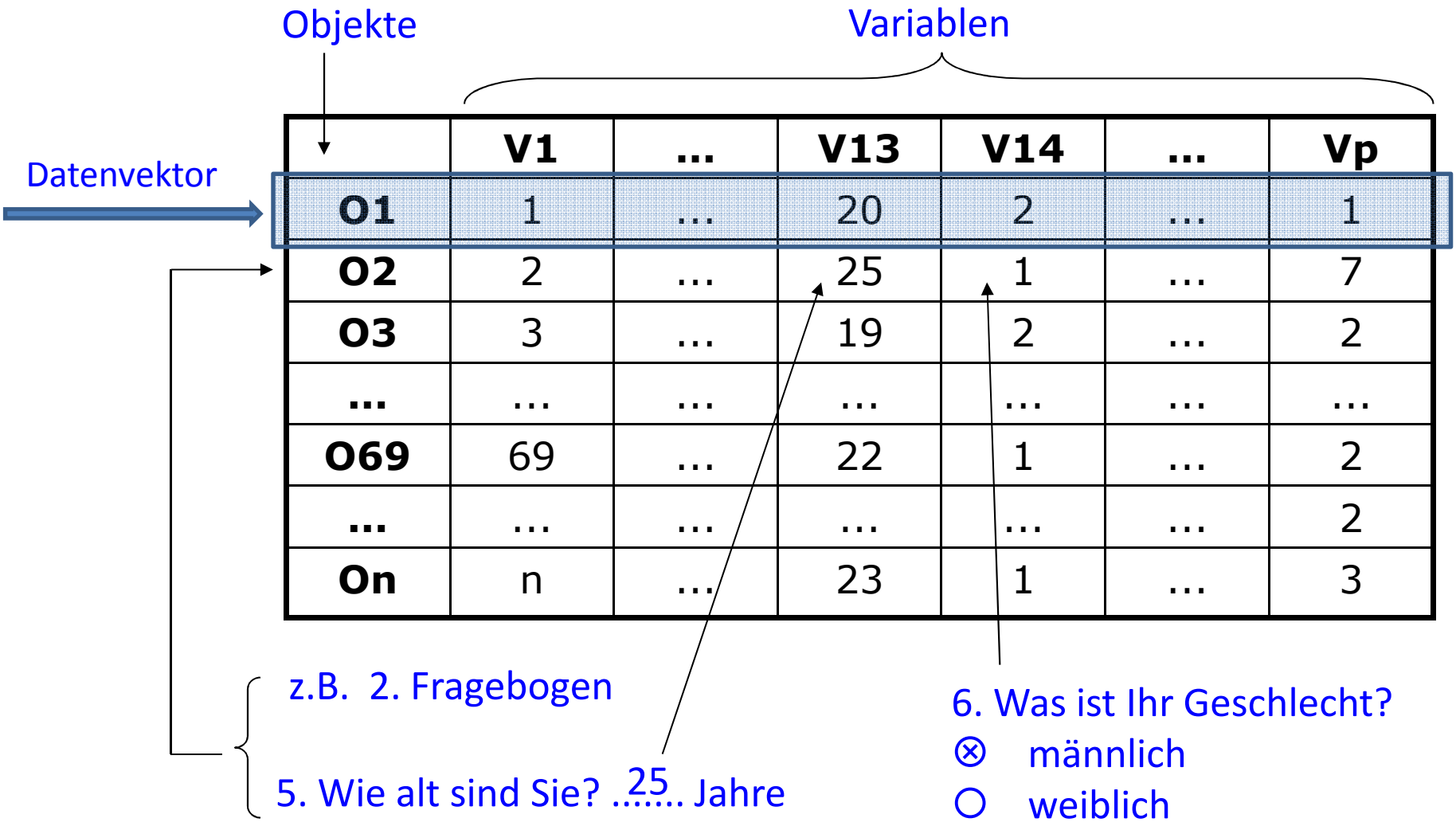
Ausgangspunkt ausgefüllter Fragebogen (oder andere vergleichbare Datenbasis)

- ← Kategorienbildung bei offenen Fragen
- ← Codeplanerstellung
- ← Codierung
- ← Datenerfassung

Datendatei
(Datensatz als Datenmatrix)



Datenmatrix



	V1	…	V13	V14	…	Vp
O1	1	…	20	2	…	1
O2	2	…	25	1	…	7
O3	3	…	19	2	…	2
…	…	…	…	…	…	…
O69	69	…	22	1	…	2
…	…	…	…	…	…	2
On	n	…	23	1	…	3

Codierung der Fragen

Informationen des Fragebogens müssen verschlüsselt (codiert) werden.

Möglichkeiten:

- Die Codieranweisungen werden direkt am Fragebogen vermerkt (ungünstig).

- Erstellung eines Codeplans (Codebook), der die Verschlüsselungsanweisung für jede Frage enthält

- Codieranweisungen im Rahmen der Erfassungssoftware

Codierung - Vorgangsweise

- Der Fragebogen wird in Variablen aufgelöst.
- Als erste Variable wird üblicherweise eine Personen-Id (Fragebogennummer) eingefügt
- Fragen und Unterfragen werden üblicherweise durchnummeriert (V1, V2, V3,...)
- Den Antwortkategorien werden Codeziffern zugeordnet.

Codierung von Fragen mit Einfachnennungen

Frage 6: Was ist Ihr Geschlecht?

○ männlich
○ weiblich

Codierung: Hier eigentlich beliebig (nominales Merkmal)
männlich = 1; weiblich = 2 oder männlich 0; weiblich 1.

Nicht ratsam, aber möglich wären Codierungen der Art
männlich = 10, weiblich = 15

Empfehlenswert ist eine einheitliche Zuordnung von Zahlenwerten entsprechend der Darstellung im Fragebogen. Umcodierungen können im Nachhinein erfolgen.

Item-Nonresponse

Im Codeplan muss eine Nonresponse-Codierung berücksichtigt werden

Nonresponse-Codierung:
(„keine Antwort oder keine eindeutige Antwort“):

⊗ männlich ⊗ weiblich	oder	○ männlich ○ weiblich

(1) Nonresponse kann mit einer dafür reservierten Ausprägung codiert werden (z.B. Code 9). Manchmal werden auch mehrere Ausprägungen dafür reserviert (z.B. 7=verweigert, 8=weiß nicht, 9=ungültig). Wichtig ist natürlich: Der Nonresponse-Code darf sich nicht im möglichen validen Ausprägungsbereich befinden.
Die Missing-Deklaration erfolgt im jeweiligen Statistik-Programm (Benutzer-definierte fehlende Werte).

(2) System-Fehlende Werte: Sie verwenden ein Eingabeprogramm, welches „Null-Werte“ zulässt. (In SPSS, Excel oder WinMask kann die Ausprägung „leer“ gelassen werden, wobei in diesem Fall das leere Feld von SPSS als Missing-Wert erkannt wird.)

Codierung von Fragen mit Mehrfachnennungen

Welche der folgenden LVA haben Sie in vorherigen Semestern schon besucht?

VU Grundlegende ...	☒
PS Fallstudien	☒
Kurs Statistik	☐

Mehrfachantworten werden bei der Codierung wie „Einzelfragen“ (mit Antwortvorgaben „ja – nein“) behandelt.

Codierung üblicherweise mit

genannt = 1,
nicht genannt = 0

Anmerkung: Eine Missing-Codierung ist i.d.R. nicht sinnvoll, wenn keine explizite „nein“-Kategorie vorhanden ist. Im Rahmen der Auswertung (nicht bei der Datenerfassung) ist es manchmal sinnvoll, jene Fälle als Nonresponse zu behandeln, bei denen keine der Optionen gewählt wurden.

Codierung von Fragebatterien

Wie gut treffen folgende Aussagen auf Sie zu:

		trifft voll zu	trifft eher zu	trifft eher nicht zu	trifft gar nicht zu
+	Ich habe eine Menge gute Eigenschaften.	①	②	③	④
-	Es gibt nicht viel, worauf ich stolz sein kann.	①	②	③	④
-	Ich wünschte, ich hätte mehr Achtung vor mir selbst.	①	②	③	④
-	Manchmal denke ich, dass ich für überhaupt nichts gut bin.	①	②	③	④

Richtung der Formulierung:
+: in Richtung hoher Selbstwert
-: in Richtung niedriger Selbstwert

Zur Vermeidung von Eingabefehlern ist – unabhängig von der Richtung der Formulierung – eine konstante Codierweise sinnvoll.

Eventuell nötige Umcodierungen erfolgen im Zuge der Auswertung im Statistikprogramm.

Codierung von offenen Fragen:

(1) Qualitativ offene Frage:

22. Welche Veränderungen bei Weiterbildungsmöglichkeiten, die für Lehrer zur Verfügung stehen, halten Sie für nötig? (Mehrfachantworten möglich)

- ❍ Verbesserung der Qualität der angebotenen Weiterbildungsmöglichkeiten
- ❍ Verbesserung der Angebotsvielfalt zur fachlichen Weiterbildung
- ❍ Verbesserung der Angebotsvielfalt zur Persönlichkeitsbildung
- ❍ Bessere zeitliche Koordination
- ❍ Bessere finanzielle Unterstützung bei der Weiterbildung
- ❍ Mehr Information über Möglichkeiten zur Weiterbildung
- ❍ Mehr regionale Angebote (kürzere Anfahrtswege)
- ❍ Zusätzliches Angebot, nämlich: _____

mehrere Möglichkeiten:

-in einem Textfeld (Stringvariable) bei der Dateneingabe erfassen.

-zunächst nur als dichotome Variable (Nennung / keine Nennung) erfassen. Die weitere Behandlung der offenen Frage erfolgt in Abhängigkeit der Anzahl der Nennungen.

-Die Antworten werden zunächst kategorisiert. Die jeweilige Kategorie wird vercodet.

Codierung von offenen Fragen:

(1) Quantitativ offene Frage:

24. An wie vielen Schulen unterrichten Sie? _____ *(Bitte Anzahl angeben)*

Ausprägung wird als Zahlenwert codiert.
Häufig werden Zusatzvereinbarungen getroffen (z.B. 9=9 und mehr)

Ergänzende Literaturempfehlung:

Benninghaus, Hans (2001). Einführung in die Sozialwissenschaftliche Datenanalyse. München, Oldenbourg, 6.Auflage oder höher.