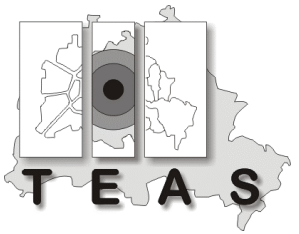


Arbeitsbereich TEAS
Institut für Geographische Wissenschaften
Freie Universität Berlin

Aufbereitung empirischer Daten

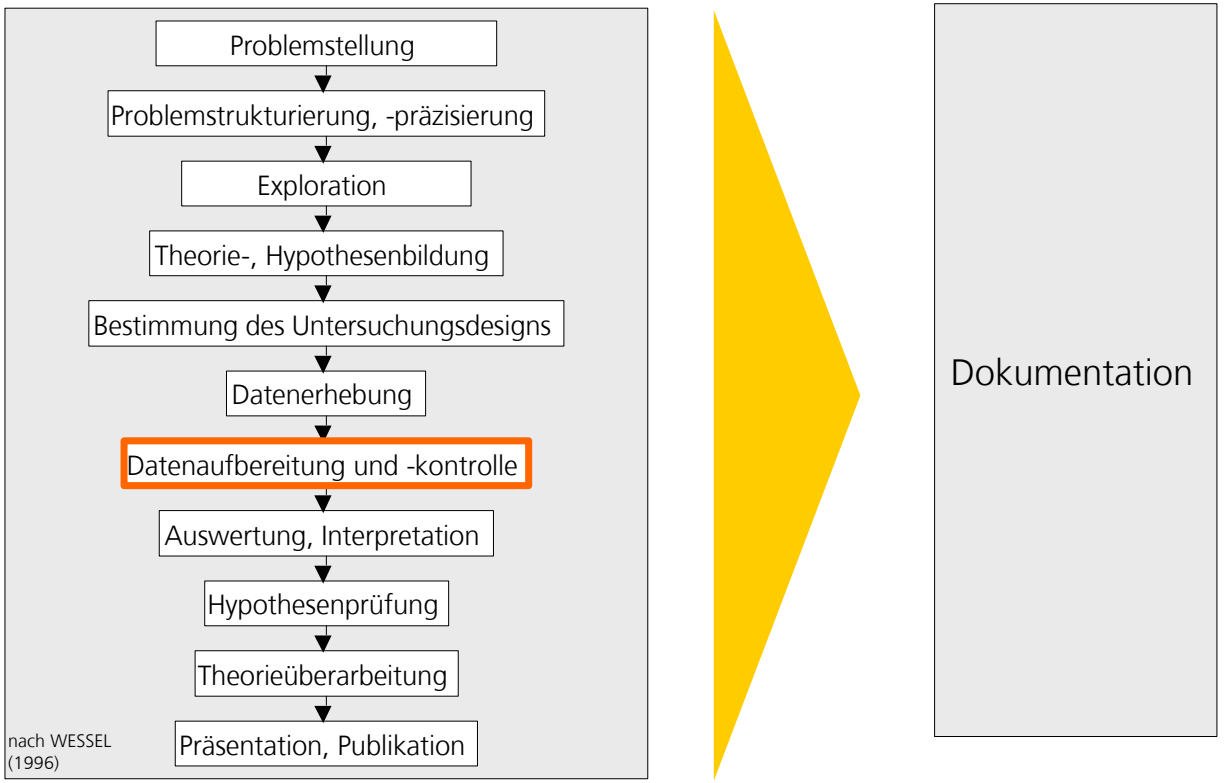
Dozentin: Dipl.-Geogr. Angelika Schulz

Titelseite

INHALT

- Datenmatrix
- Definition der Variablen
- Dateneingabe (I-III)
- Fehlende Werte (I-III)
- Datensatzprüfung
- Datenschutz (I-II)
- Codeplanerstellung
- Primärauswertung

Datenaufbereitung im wissenschaftlichen Forschungsprozess

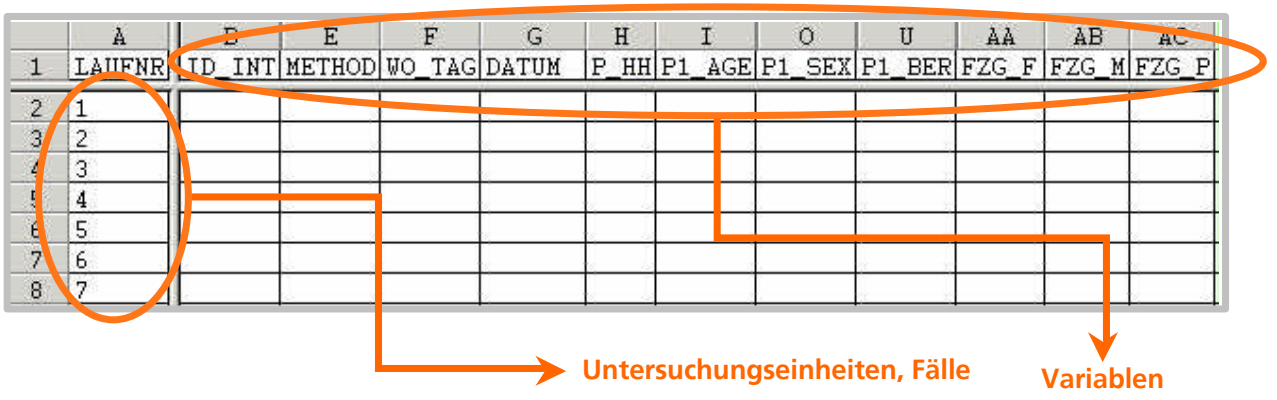


Datenaufbereitung_Forschungsprozess

Erstellung einer Datenmatrix

Vorbereitung der informationstechnischen Verarbeitung der originalen Rohdaten

- Erstellung einer **Dummy-Matrix vor Beginn der Feldphase** zur Überprüfung der geplanten Vorgehensweise:
  - Umsetzung der Variablen
  - Erfassung aller Antwortmöglichkeiten mithilfe eines Codeplans



ebenfalls **vor** Beginn der Feldphase durchzuführen:

- **Pre-Test**
- **Probeauswertung**

Datenmatrix

Definition der Variablen

- Ableitung der Variablen aus dem Erhebungsinstrument
- **eindeutige** Benennung nach einheitlichem Prinzip, z.B. ...
  - anhand der Fragennummer
  - Verwendung geeigneter von Abkürzungen
  - Kennzeichnung zusätzlich erstellter Variablen
- Festlegung eines **gültigen Wertebereichs** für jede Variable:
  - Berücksichtigung **aller** theoretisch möglichen Antworten
  - **eindeutige** Antwortkategorien ohne Überlappungsbereiche
- Berücksichtigung **softwareseitiger Einschränkungen**

H = Haushaltsfragebogen
1 = Haushaltsmitglied 1
2 = Haushaltsmitglied 2
Alter = Gegenstand der Variable

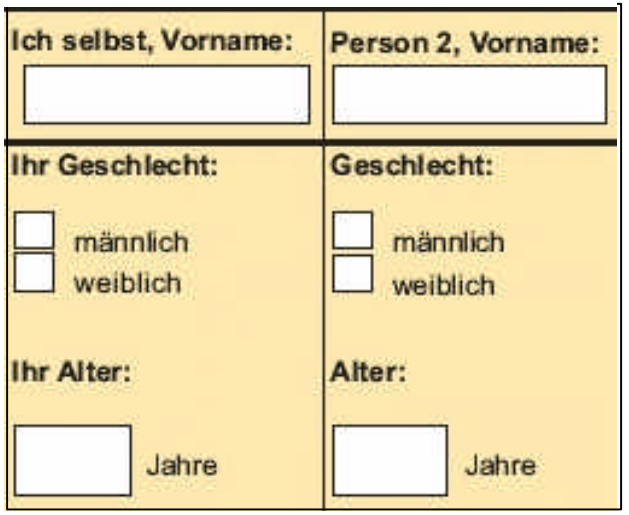
```
H1ALTER   Alter der Person 1
          Druckformat: F8.2
          Schreibformat: F8.2

          Wert     Label

        997,00     Verweigert
        998,00     Weiß nicht
        999,00     Keine Angabe

H2ALTER   Alter der Person 2
          Druckformat: F8.2
          Schreibformat: F8.2
          Missing-Bereich: -21,00

          Wert     Label

        -21,00 M   1-Personen Haushalt
        997,00     Verweigert
        998,00     Weiß nicht
        999,00     Keine Angabe
```

Quelle: MiD 2002 – Mobilität in Deutschland 2002
(Haushaltsfragebogen und Datensatzbeschreibung)

Variablendefinition

Beispiel für Variablennamen

Ausschnitt aus einem Haushaltsfragebogen (Angaben zu Haushaltsmitglied 1 und 2)



H = Haushaltsfragebogen

1 = Haushaltsmitglied 1

2 = Haushaltsmitglied 2

Alter = Gegenstand der Variable

```
H1ALTER   Alter der Person 1
          Druckformat: F8.2
          Schreibformat: F8.2

          Wert     Label

         997,00     Verweigert
         998,00     Weiß nicht
         999,00     Keine Angabe

H2ALTER   Alter der Person 2
          Druckformat: F8.2
          Schreibformat: F8.2
          Missing-Bereich: -21,00

          Wert     Label

         -21,00 M   1-Personen Haushalt
         997,00     Verweigert
         998,00     Weiß nicht
         999,00     Keine Angabe
```

Quelle: MiD 2002 – Mobilität in Deutschland 2002
(Haushaltsfragebogen und Datensatzbeschreibung)

Bsp. 1 Variablendefinition

Beispiel für Variablennamen

Ausschnitt aus einem Personenfragebogen (Frage Nr. 2)

2. Besitzen Sie zurzeit ein verkehrstüchtiges Fahrrad?

☐ ja
☐ nein

Codeplan zu Frage 2 des Personenfragebogens

```
P02       Besitzen Sie zur Zeit ein verkehrstüchtiges Fahrrad?
          Druckformat: F8.2
          Schreibformat: F8.2
          Missing-Bereich: -200,00

          Wert    Label

       -200,00 M  Proxy ab 14 Jahre werden nicht befragt
          1,00    Ja
          2,00    Nein
          7,00    Verweigert
          8,00    Weiß nicht
          9,00    Keine Angabe
```

P = Personenfragebogen

02 = Frage Nr. 2

Quelle: MiD 2002 – Mobilität in Deutschland 2002
(Personenfragebogen und Datensatzbeschreibung)

Bsp. 2 Variablendefinition

Dateneingabe I

- **extrem fehleranfälliger Arbeitsschritt**
- **eindeutige** Kennzeichnung aller originalen Erhebungsunterlagen vor Beginn der Dateneingabe, z.B. durch laufende Nummerierung (eigene Variable im Datensatz)
  - erlaubt die Trennung anderer identifizierender Variablen vom Datensatz (z.B. Name, Adresse etc.)
  - erste – nicht hinreichende – Maßnahme des Datenschutzes (Anonymisierung)
- **verschiedene Verfahrensweisen** für Dateneingabe:
  - manuelle Eingabe (direkt vom Erhebungsinstrument oder anhand einer Datenmatrix)
  - maschinelles Einlesen (Voraussetzung: geeigneter Fragebogen)

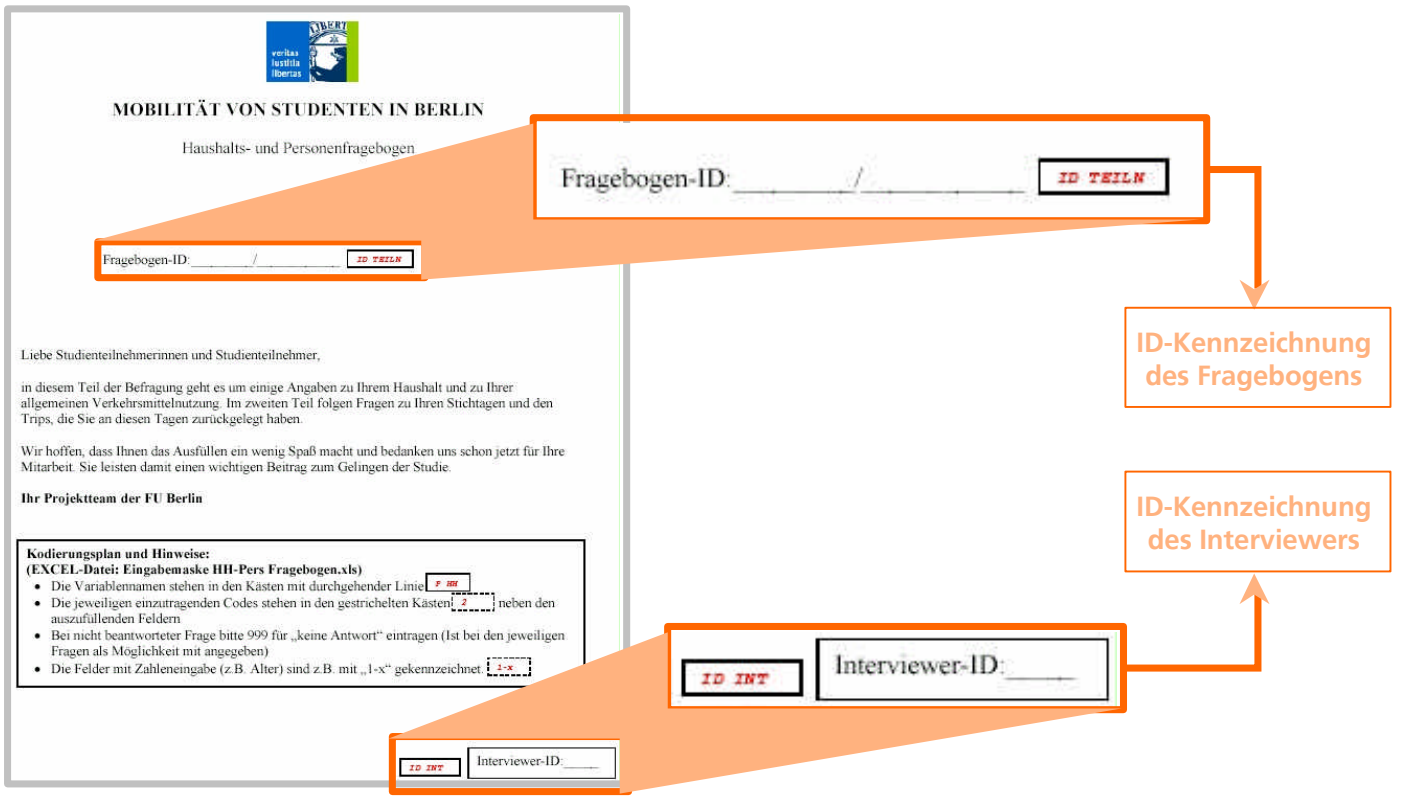
  } herkömmliche Erhebungsverfahren, PAPI

  - computergestützte Datenerfassung

  } CATI, elektronische Fragebögen, online-Erhebung
- **verschiedene Auswirkungen** des gewählten Verfahrens auf ...
  - Bearbeitungsdauer
  - Kosten
  - Qualität der Daten

Beispiel für Kennzeichnung des Fragebogens

Deckblatt eines Fragebogens



Bsp. Fragebogenkennzeichnung

Dateneingabe II

- **einheitliche** und **eindeutige Codieranweisungen** erstellen:
  - Umgang mit nicht vorgesehenen oder uneindeutigen Antworten
  - insb. wenn mehrere Personen an der Dateneingabe beteiligt sind
- Die **Eingabe ungültiger Werte** lässt sich mit Hilfe entsprechender Software, z.B. Microsoft Access©, Microsoft Excel©, SPSS Data Entry©, von vornherein **vermeiden**:
  - fehlende Werte
  - Werte, die nicht codiert sind und somit außerhalb des Gültigkeitsbereiches liegen
- Das **Einlesen von Datendateien**, die im ASCII-Format vorliegen, erfordert einen **vollständigen** und **aktuellen Codeplan**, der alle Informationen über die Variablenformate enthält:
  - Bei der Festlegung der Variablenbreite ist besondere Sorgfalt hinsichtlich der Spaltentrennung erforderlich.
  - Bei numerischen Variablen ist auf korrekte Dezimaltrennung zu achten. Sonderformate wie Datum oder Währung sind zu berücksichtigen.
  - Nach dem Import sind **alle (!) Variablen** einzeln auf Richtigkeit zu **überprüfen** !

Beispiel für eine Codieranleitung

Ausschnitt aus einer Codierungsanleitung – allgemeine Hinweise

Kodierungsplan und Hinweise:
(EXCEL-Datei: Eingabemaske HH-Pers Fragebogen.xls)

- Die Variablennamen stehen in den Kästen mit durchgehender Linie [F_HH]
- Die jeweiligen einzutragenden Codes stehen in den gestrichelten Kästen [2] neben den auszufüllenden Feldern
- Bei nicht beantworteter Frage bitte 999 für „keine Antwort" eintragen (Ist bei den jeweiligen Fragen als Möglichkeit mit angegeben)
- Die Felder mit Zahleneingabe (z.B. Alter) sind z.B. mit „1-x" gekennzeichnet. [1-x]

Ausschnitt aus einer Codierungsanleitung – Spezifikation der Codes

Benennung der Variablen

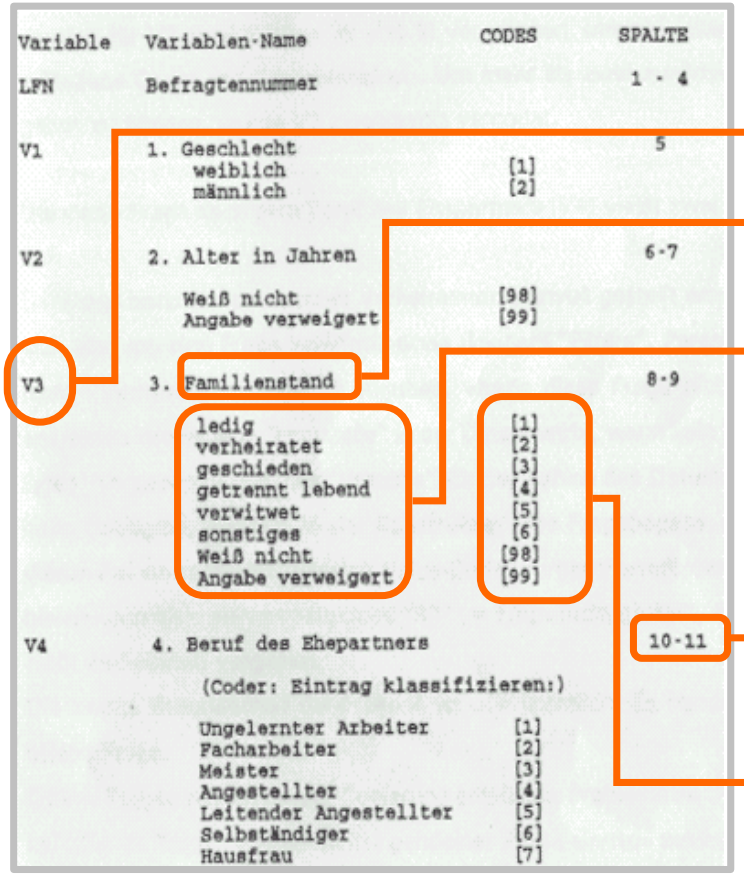
Festlegung eines Wertebereiches

Codierung der Antwortkategorien (inkl. "999" = "keine Antwort")

	P1 SEX	P2 SEX	P3 SEX	P4 SEX	P5 SEX	P6 SEX
	Sie Selbst	Zweite Person	Dritte Person	Vierte Person	Fünfte Person	Sechste Person
	Ihr Geschlecht	Geschlecht	Geschlecht	Geschlecht	Geschlecht	Geschlecht
2	☐ männlich	☐ männlich	☐ männlich	☐ männlich	☐ männlich	☐ männlich
1	☐ weiblich	☐ weiblich	☐ weiblich	☐ weiblich	☐ weiblich	☐ weiblich
999						
	P1 AGE	P2 AGE	P3 AGE	P4 AGE	P5 AGE	P6 AGE
	Ihr Alter	Alter	Alter	Alter	Alter	Alter
1-x	___ Jahre	___ Jahre	___ Jahre	___ Jahre	___ Jahre	___ Jahre
999						
	P1 BER	P2 BER	P3 BER	P4 BER	P5 BER	P6 BER
	Berufstätigkeit	Berufstätigkeit	Berufstätigkeit	Berufstätigkeit	Berufstätigkeit	Berufstätigkeit
1	☐ Vollzeit	☐ Vollzeit	☐ Vollzeit	☐ Vollzeit	☐ Vollzeit	☐ Vollzeit
2	☐ Teilzeit	☐ Teilzeit	☐ Teilzeit	☐ Teilzeit	☐ Teilzeit	☐ Teilzeit
3	☐ Azubi/ Student(in)	☐ Azubi/ Student(in)	☐ Azubi/ Student(in)	☐ Azubi/ Student(in)	☐ Azubi/ Student(in)	☐ Azubi/ Student(in)
4	☐ nicht berufstätig	☐ nicht berufstätig	☐ nicht berufstätig	☐ nicht berufstätig	☐ nicht berufstätig	☐ nicht berufstätig
999						

Bsp. 1 Codieranleitung

Beispiel für eine Codieranleitung



Variable	Variablen-Name	CODES	SPALTE
LFN	Befragtennummer		1 - 4
V1	1. Geschlecht		5
	weiblich	[1]	
	männlich	[2]	
V2	2. Alter in Jahren		6-7
	Weiß nicht	[98]	
	Angabe verweigert	[99]	
V3	3. Familienstand		8-9
	ledig	[1]	
	verheiratet	[2]	
	geschieden	[3]	
	getrennt lebend	[4]	
	verwitwet	[5]	
	sonstiges	[6]	
	Weiß nicht	[98]	
	Angabe verweigert	[99]	
V4	4. Beruf des Ehepartners		10-11
	(Coder: Eintrag klassifizieren:)		
	Ungelernter Arbeiter	[1]	
	Facharbeiter	[2]	
	Meister	[3]	
	Angestellter	[4]	
	Leitender Angestellter	[5]	
	Selbständiger	[6]	
	Hausfrau	[7]	

Ausschnitt aus einer Codierungsanleitung – Spezifikation der Codes

Variablenname

Variablenlabel

Variablenausprägung (Antwortkategorien)

laufende Nummer der Spalten in einer ASCII-Datei

Codierung der Antwortkategorien (inkl. "999" = "keine Antwort")

Bsp. 2 Codieranleitung

Dateneingabe III – Datenblatt

Bei der Erstellung eines Datenblattes werden die Antworten zunächst manuell vom Fragebogen auf dieses übertragen; erst danach erfolgt die endgültige Eingabe in das gewünschte Datenverarbeitungsprogramm.

- Vorteile:
  - Identifizierung nicht vorhergesehener Antwortkategorien; Möglichkeit zur Ergänzung fehlender Codierungen
  - Konzentration auf jeweils eine einzige Aufgabe (a. "Entziffern" der Antworten und Codierung, b. Eintippen der codierten Antworten)
  - Gelegenheit zur ersten Konsistenzprüfung
- Nachteile:
  - zusätzlicher Bearbeitungsschritt erhöht die Wahrscheinlichkeit von Fehleingaben
  - erhöhter Zeitaufwand

Fehlende Werte (missing values) I – Entstehung

- Fehlende Werte in Datensätzen ergeben sich **aus unterschiedlichen Gründen**:
  - Frage trifft für Probanden nicht zu
  - Filterfragen werden übersehen/ ignoriert oder Filterhinweise fehlen im Fragebogen
  - Probanden finden keine für sie zutreffende Antwortkategorie, sie weichen entweder auf "weiß nicht" aus oder antworten gar nicht.
  - Frage ist missverständlich gestellt
  - Eingabefehler
  - Beim Zusammenführen verschiedener Quellen gibt es keine "matches".
- Fehlende Werte können **vermieden** werden:
  - Fragebogen klar und übersichtlich gestalten (graphisch, sprachlich).
  - Fragen und Antwortkategorien eindeutig formulieren, ggf. mit Erläuterungen versehen.
  - Bei Verzweigungen (Filterfragen) eindeutige Sprungbefehle geben ("Sind sie nicht berufstätig, fahren Sie bitte fort mit Frage 10").

Fehlende Werte (missing values) II - Umgang

Für den Umgang mit fehlenden Werten gibt es verschiedene Möglichkeiten:

- Datenzellen frei lassen (**blanks**)
  - Vorteil: Zahl der Fälle ist in jedem Fall korrekt zu bestimmen.
  - Nachteil: Beim Export in andere Datenformate werden leere Zellen manchmal mit "0" gefüllt. Das Ergebnis kann verfälscht werden (falsche Fallzahlen).
- Fehlende Werte mit **speziellen Codes** versehen, z.B. "999"; in keinem Fall Werte aus dem gültigen Wertebereich verwenden !
  - Vorteil: Fehlende Werte sind eindeutig zu identifizieren.
  - Nachteil: Werden die entsprechenden Codes (z.B. "999") nicht korrekt als "missing" definiert, werden sie bei Analysen als gültige Werte mit allen negativen Folgen mit berücksichtigt (falsche Fallzahlen, Mittelwerte etc.).
- **verschiedenen Ursachen** für fehlende Werte können erfasst werden (z.B. "77"="weiß nicht", "88"="Antwort verweigert", "99"="Frage trifft nicht zu (Filter)")
- variablenübergreifend **einheitliche Codierung** für fehlende Werte verwenden

Fehlende Werte II

Fehlende Werte (missing values) III - Bereinigung

Die Bereinigung fehlender Werte ist immer **problematisch**, da der Originalzustand der Rohdaten nachträglich nicht mehr rekonstruiert werden kann.

Fehlende Werte können auf folgende Weise bereinigt werden:

- **intervallskalierte** Variablen: leere Zellen mit entsprechendem Mittelwert auffüllen
- **kategoriale** Variablen: leere Zellen anhand der Verteilung der Werte in den anderen Zellen füllen
- Fälle mit vielen fehlenden Werten können komplett entfernt werden ("listwise deletion").

- Alle durchgeführten **Bereinigungsverfahren** sind in jedem Fall **ausführlich** zu **dokumentieren** !

Datensatzprüfung

Im Anschluss an die Dateneingabe können Datensätze **formal** und **inhaltlich** geprüft werden.

- **Aufwand** und **Nutzen** abwägen (Zeit, Kosten, Datenqualität)
- **Formale Prüfung** (fehlende Werte, nicht codierte Werte, ungültige Werte):
  - Erstellung von Häufigkeitstabellen (sehr einfach); wichtig: Filterkriterien berücksichtigen
  - gezielte Suchabfragen, z.B. nach Werten außerhalb des gültigen Wertebereichs (eher umständlich)
  - Ausdruck der Rohdaten als Liste, dann Sichtprüfung (nur bei überschaubaren Datensätzen mit vergleichsweise geringer Fallzahl sinnvoll)
- **Inhaltliche Prüfung**:
  - Kreuztabellierung von Variablen zur Identifizierung inkompatibler Wertebereiche (z.B. "Alter"="< 18" und "Pkw-Führerscheinbesitz"="ja")
  - ausführliche Einzelprüfung von 5-10% aller Datensätze hinsichtlich Plausibilität und Konsistenz zur Identifizierung von Fehlermustern

Datenschutz I

- Bei der Verarbeitung personenbezogener Daten sind die **regional unterschiedlichen Bestimmungen der Datenschutzgesetze** zu berücksichtigen, insbesondere dann, wenn Dritte Zugang zu den erhobenen Daten haben sollen.
- **Personenbezogene Daten**: Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (in nicht aggregierter Form).
- **Anonymisierte** und **aggregierte** Daten fallen nicht mehr unter die Bestimmungen der Datenschutzgesetze:
  - **Anonymisieren** ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können (faktische Anonymisierung). Ein Restrisiko der Deanonymisierung kann bestehen bleiben.
- Zu berücksichtigen sind ggf. vorhandene **Verknüpfungsverbote**.
- Für die Nutzung durch Dritte (insbesondere im Rahmen von Sekundäranalysen) werden häufig **"public use files"** oder **"scientific use files"** erstellt, welche mindestens faktisch anonymisiert sind.

Datenschutz II - Anonymisierung

- Zu unterscheiden sind Variablen, ...
  - die die **direkte Identifizierung** einer Person, eines Haushalts oder eines Unternehmens ermöglichen (z.B. Namen, Adressen, Telefonnummern, Sozialversicherungsnummern, in Einzelfällen auch PLZ-Gebiete),
  - die eine **indirekte Identifizierung** ermöglichen (z.B. räumliche Einheiten wie amtl. Gemeindeschlüssel, Zensusgebiete oder Wahlbezirke, exakte Daten wie Geburtsdatum, Führerscheinerwerb, Jahr der Unternehmensgründung oder Insolvenz).
- Folgende Verfahren können für die Behandlung identifizierender Variablen angewendet werden:
  - Variable vollständig entfernen
  - Zusammenfassung einzelner Werte einer Variable zu größeren Wertebereichen (Recodierung), die jeweils unteren/ oberen Wertebereiche kappen (z.B. bei genauen Angaben zum Einkommen "Jahreseinkommen in Euro"=" >50000").
  - Zusammenfassung von zwei einzelnen Variablen zu einer neuen Variable
  - Einstreuung von Zufallsfehlern
  - Zerlegung des Datensatzes in Teildatensätze für einzelne Merkmalsbereiche
  - Überführung identifizierender Elemente mittels eines Schlüssels in ein Pseudonym (Pseudoanonymisierung)

Erstellung eines Codeplans

- Der **endgültige Datensatz** ist mittels eines **begleitenden Codeplans** zu dokumentieren.
  Dieser stellt die minimale Form der Dokumentation dar und ist insbesondere bei Dateien im ASCII-Format erforderlich.
- Bestandteile des Codeplans sind folgende Angaben:
  - Anzahl der Zeilen pro Fall
  - Anzahl der Variablen pro Fall
  - Anzahl der Fälle im Datensatz
  - jeweilige Position der Variablen im Datensatz (Startspalte, Endspalte)
  - Breite der Variablen (Anzahl der Zeichen)
  - Namen und Label aller Variablen
  - Schlüssel für gültige Antwortkategorien inkl. fehlender Werte
  - Quellenangabe für verwendete Codierungsschemata (z.B. amtl. Gemeindekennzahlen, Klassifikation der Wirtschaftszweige, Güterverzeichnis für Produktionsstatistiken)

Dokumentation der Primärauswertung

- Primärauswertung als minimale Form der Datenaufbereitung in Form von einfachen Tabellen oder Kreuztabellen der einzelnen Variablen (Publikation i. d. R. in Tabellenbänden)
- Die folgenden Bestandteile einer Tabelle sind **zwingend** erforderlich:
  - Überschrift/ Titel
  - zugrunde liegende Grundgesamtheit
  - bei Prozentwerten: Angabe der zugrunde liegenden Fallzahl (n)
  - Hinweise auf geltende Filterbedingungen
  - ggf. Begriffsdefinitionen
  - Quelle der Daten