

# Statistische Methoden der Datenaufbereitung und -auswertung

## Datenaufbereitung

# Inhalt

- ▶ Datenmatrizen
- ▶ Codebuch
- ▶ Anonymisierung
- ▶ Fehlerbereinigung & Plausibilitätsprüfung
- ▶ Umgang mit fehlenden Werten
- ▶ Umkodieren von Variablen
- ▶ Zusammenführen von Datensätzen

# Datenaufbereitung

- ▶ Rohdaten in ein einheitliches und strukturiertes Format überführen
- ▶ Datensätze
  - ▷ anonymisieren
  - ▷ bereinigen
  - ▷ transformieren

Fokus in dieser Veranstaltung:

Daten aus (voll-)strukturierten,  
schriftlichen Befragungen

um Datenqualität für nachfolgende Analyse zu erhöhen.

# Erfassung quantitativer Daten in Tabellen oder Datenmatrizen

The diagram illustrates a data table with the following structure:

ID	Variable 1	Variable 2	Variable 3
Untersuchungseinheit 1			
Untersuchungseinheit 2			
Untersuchungseinheit 3		Merkmalsausprägung (numerisch)	
...			
Untersuchungseinheit n			

Callouts and annotations:

- Variablen in Zeilen:** Points to the header row (ID, Variable 1, Variable 2, Variable 3).
- Untersuchungseinheiten in Spalten (z.B. Teilnehmer einer Befragung):** Points to the first column (ID).
- Merkmalsausprägung (numerisch):** Points to the cell containing the value 'Merkmalsausprägung (numerisch)' in the third column of the third row.
- Einzelne Beobachtung in Zellen:** Points to the cell containing the value 'Merkmalsausprägung (numerisch)' in the third column of the third row.

# Codeplan/Codebuch

\*bcsopen.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	rowlabel	Numeric	12	0	case identifier (8 digi...	None	None	8	Right	Nominal
2	sex	Numeric	8	0	Respondent sex	{1, male}...	None	8	Right	Nominal
3	age	Numeric	8	0	Respondent age	None	998, 999	8	Right	Scale
4	livharm1	Numeric	10	0	ons harmonised mari...	{1, married}...	None	8	Right	Nominal
5	ethgrp2	Numeric	8	0	Respondent ethnic o...	{1, white}...	None	8	Right	Nominal
6	educat3	Numeric	8	0	respondent educatio...	{1, none}...	None	8	Right	Nominal
7	work	Numeric	8	0	any paid work in last...	{1, yes}...	8, 9	8	Right	Nominal
8	ysarea	Numeric	8	0	how long have you liv...	{1, less than 12 mo...	8, 9	8	Right	Nominal
9	resyrago	Numeric	8	0	were you living at thi...	{1, yes}...	8, 9	8	Right	Nominal
10	tenure1	Numeric	8	0	in which way do you ...	{1, own it outright}...	None	8	Right	Nominal
11	rural2	Numeric	8	0	type of area 2004: ur...	{1, urban}...	None	8	Right	Nominal
12	rubbccomm	Numeric	8	0	in the immediate are...	{1, very common}...	None	8	Right	Ordinal
13	vandcomm	Numeric	8	0	how common is vand...	{1, very common}...	None	8	Right	Ordinal
14	poorhou	Numeric	8	0	how common are ho...	{1, very common}...	None	8	Right	Ordinal
15	tcemdiqu2	Numeric	9	2	Index of multiple dep...	None	None	8	Right	Nominal
16	tcwmdiqu2	Numeric	9	2	Index of multiple dep...	None	None	8	Right	Nominal
17	causem	Numeric	8	0	one main cause of cr...	{1, a. too lenient se...	None	8	Right	Nominal
18	walkdark	Numeric	8	0	how safe do you feel...	{1, very safe}...	8, 9	8	Right	Ordinal
19	walkday	Numeric	8	0	how safe do you feel...	{1, very safe}...	8, 9	8	Right	Ordinal
20	homealon	Numeric	8	0	how safe do you feel...	{1, very safe}...	8, 9	8	Right	Ordinal
21	tcviolent	Numeric	9	2	Respondent level of ...	None	None	8	Right	Nominal
22	tcsteal	Numeric	9	2	Respondent level of ...	None	None	8	Right	Nominal

Data View Variable View

IBM SPSS Statistics Processor is ready

Cumulative  
Percent

0.00

100.00

Help

Missing	Columns	Align	Measure
e	8	Right	Nominal
e	8	Right	Nominal
999	8	Right	Scale
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Ordinal
e	8	Right	Ordinal
e	8	Right	Ordinal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Ordinal
e	8	Right	Ordinal
e	8	Right	Nominal
e	8	Right	Nominal
e	8	Right	Nominal

Statistics Processor is ready

statistics-part-1/

# Beispiel

## Codebuch

### Fehlercodes

Die Aufschlüsselung der Fehlercodes ist wie folgt:

-77 = Teilnehmer hat die Frage nicht gesehen, beispielsweise weil er ein Abbrecher war oder weil auf Grund der Filterführung die betreffende Seite oder Frage nicht angezeigt wurde.

-66 = Projektvariablen vom Typ v\_000, die sich auf Textfelder beziehen: Der Teilnehmer hat die jeweilige Frage nicht gesehen, weil sie ausgeblendet war.

0 = Der Teilnehmer hat die jeweilige Frage gesehen, aber nicht bearbeitet. Gilt nicht für Textfelder.

-99 = Projektvariablen vom Typ v\_000, die sich auf Textfelder beziehen: Der Teilnehmer hat die jeweilige Frage gesehen, aber nicht bearbeitet.

lfdn

		Wert
Standardattribute	Label	number
N	Gültig	2084
	Fehlend	0

### Was ist Ihr höchster akademischer Abschluss? v\_1054

		Wert	Anzahl	Prozent
Standardattribute	Label	Akademischer Abschluss (Art)		
Gültige Werte	0		7	,3%
	1	Bachelor (Uni/FH)	43	2,0%
	2	Master (Uni/FH)	308	14,8%
	3	Diplom (Uni/FH)	525	25,2%
	4	Staatsexamen	59	2,8%
	5	Magister	100	4,8%
	6	Promotion	847	40,7%
	7	Habilitation	191	9,2%
	8	Sonstiges, und zwar:	4	,2%

Quelle: Pscheida, Daniela; Albrecht, Steffen; Herbst, Sabrina; Minet, Claudia; Köhler, Thomas (2015): Nutzung von Social Media und onlinebasierten Anwendungen in der Wissenschaft 2014. GESIS Datenarchiv, Köln. ZA5972 Datenfile Version 1.0.0,

<https://doi.org/10.4232/1.12262>

# Aufgabe

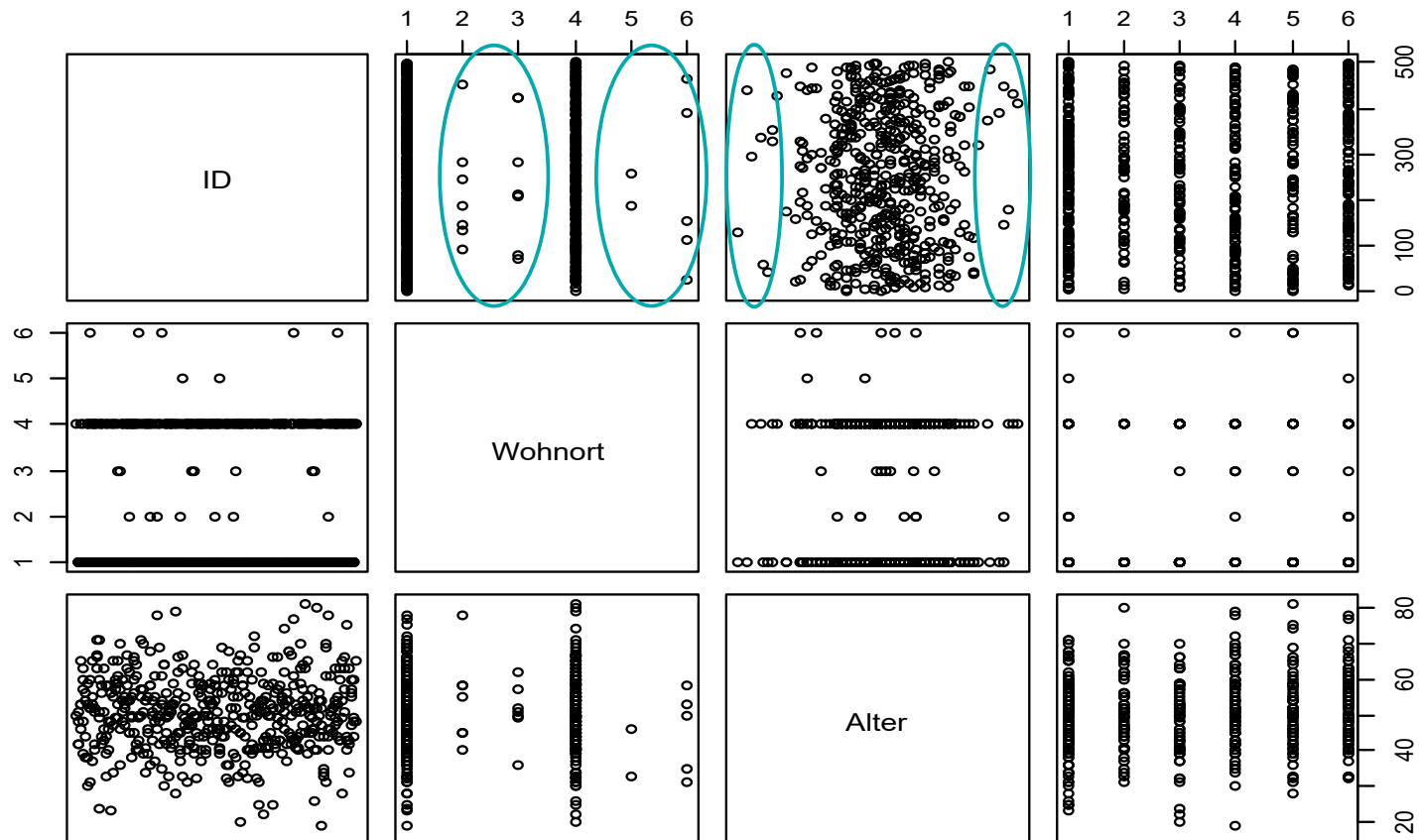
- ▶ Erstellen Sie unter Zuhilfenahme des Fragebogens, der für Ihren Forschungsdatensatz vorliegt, sowie des Datensatzes ein Codebuch mit Angaben zum Variablennamen (möglichst sprechend), Datentyp, Wertebereich bzw. Ausprägungen mit der jeweiligen Kodierung und Skalenniveau.
- ▶ Lesen Sie den Datensatz ein und kodieren Sie die Variablen entsprechend der Vorgaben in Ihrem Codebuch. Kontrollieren Sie dabei insbesondere Reihenfolgen von automatisch angelegten Faktor-Variablen und korrigieren Sie diese gegebenenfalls. Bei sehr komplexen oder umfangreichen Fragebögen, treffen Sie nach Rücksprache eine Auswahl.

# Anonymisierung bei vollstrukturierten Fragebogenstudien

- ▶ Hoher Standardisierungsgrad der Antworten trägt zu Anonymisierung bei
- ▶ Qualitative Fragen prüfen
- ▶ Kombinationen von demographischen Fragebogenitems prüfen (insbes. bei bekanntem Personenkreis)
  - ▷ Löschen ganzer Variablen
  - ▷ Verallgemeinerung von Daten
  - ▷ Neukodieren von Variablen
- ▶ Entfernen von Identifikatoren (z.B. IP-Adressen)



# Graphische Darstellungen als Hilfsmittel



# Fehlerbereinigung

Fehlerhafte Werte (z.B. durch manuelle Dateneingabe) identifizieren

- ▶ Überprüfen von
  - ▷ Wertebereichen
  - ▷ Häufigkeitsverteilungen
- ▶ Graphische Exploration

Alle Veränderungen dokumentieren und begründen!

# Plausibilitätsprüfung

- ▶ Identifizieren von Fragebögen oder Antworten, die nicht wahrheitsgemäß, ernsthaft oder vollständig ausgefüllt/beantwortet wurden.
  - ▷ Betrachten von Antwortzeiten und -mustern sowie Freitextfeldern
  - ▷ Betrachten von Antworten auf verschiedene Items in Kombination
- ▶ Möglicher Umgang: Ausschluss von einzelnen Datensätzen (erhobene Daten eines Probanden), ggf. Korrekturen vornehmen (begründen, dokumentieren!)

# Fehlende Werte

- ▶ Gründe: Antwortverweigerung, unabsichtliche Auslassung, fehlende Antwortoptionen, keine Meinung, aufgrund von Filterfrage nicht gezeigt, Fehler beim einlesen der Daten ...
- ▶ Kodierung von Gründen für fehlende Werte möglich
- ▶ Umgang:
  - ▷ Ausschluss von fehlenden Werten (z.B. fallweise oder paarweise)
  - ▷ Imputation

# Umkodierung

- ▶ Zusammenfassen metrischer Variablen (Bilden von Kategorien)
- ▶ Aggregation kategoriale Variablen
- ▶ Variablen für komplexe Konstrukte

Originaldaten  
erhalten und neue  
Variablen bilden

# Datensätze zusammenführen

- ▶ Horizontal (z.B. aus verschiedenen Erhebungsinstrumenten oder Erhebungszeitpunkten)
- ▶ Vertikal (mehrere Stichproben mit unterschiedlichen Probanden)

# Aufgabe

Explorieren Sie Ihren Datensatz graphischen und bereiten Sie den eigenen Datensatz gemäß der besprochenen Aspekte auf. Die Anonymisierung wird an dieser Stelle höchstwahrscheinlich wegfallen können, da diese vor der Bereitstellung der Datensätze passiert sein sollte.