Statistische Methoden der Datenaufbereitung und -auswertung

Datenaufbereitung

Inhalt

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

Datenaufbereitung

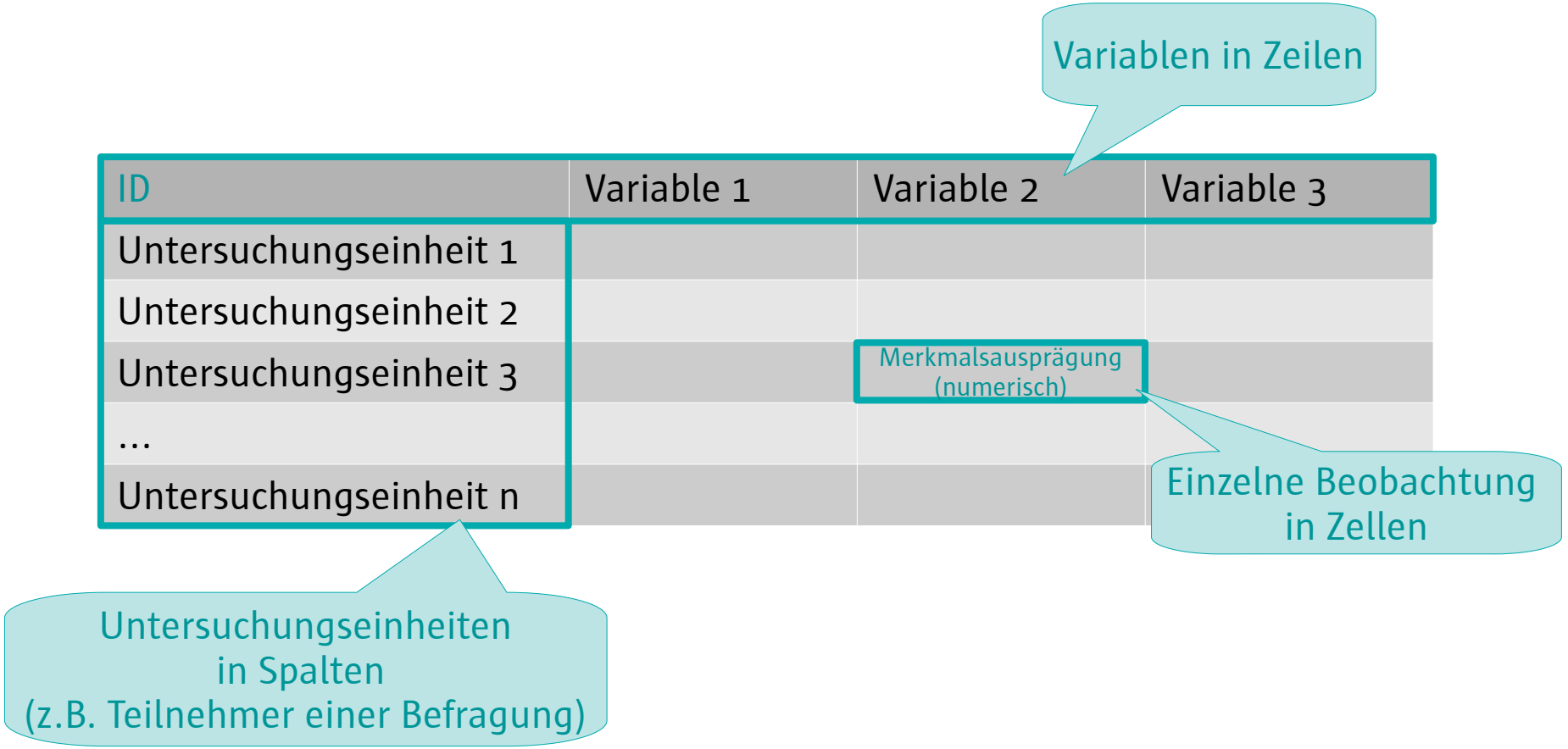
- Rohdaten in ein einheitliches und strukturiertes Format überführen
- Datensätze
  - anonymisieren
  - bereinigen
  - transformieren

  um Datenqualität für nachfolgende Analyse zu erhöhen.

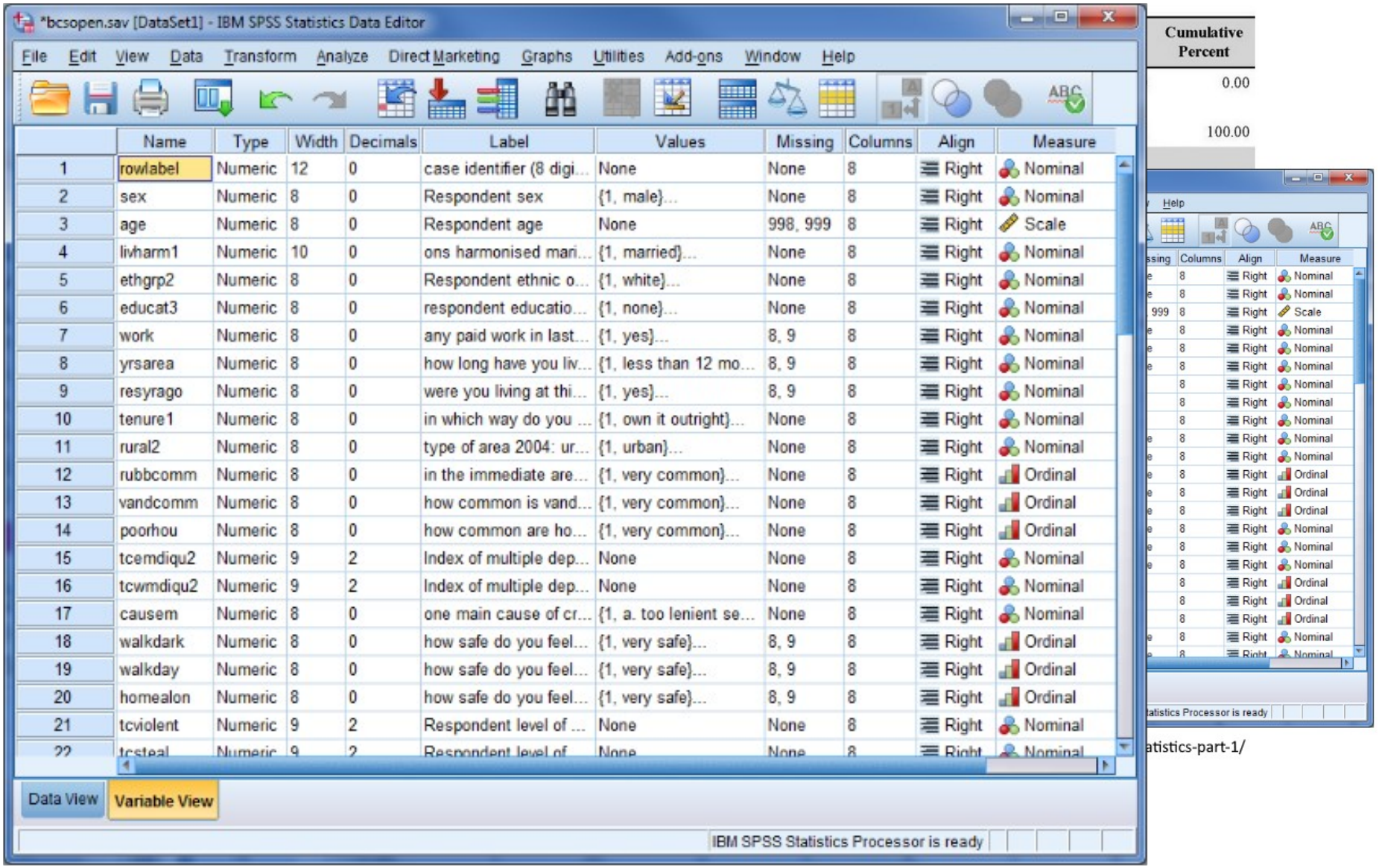
Fokus in dieser Veranstaltung:

Daten aus (voll-)strukturierten, schriftlichen Befragungen

Erfassung quantitativer Daten in Tabellen oder Datenmatrizen



Codeplan/Codebuch



atistics-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

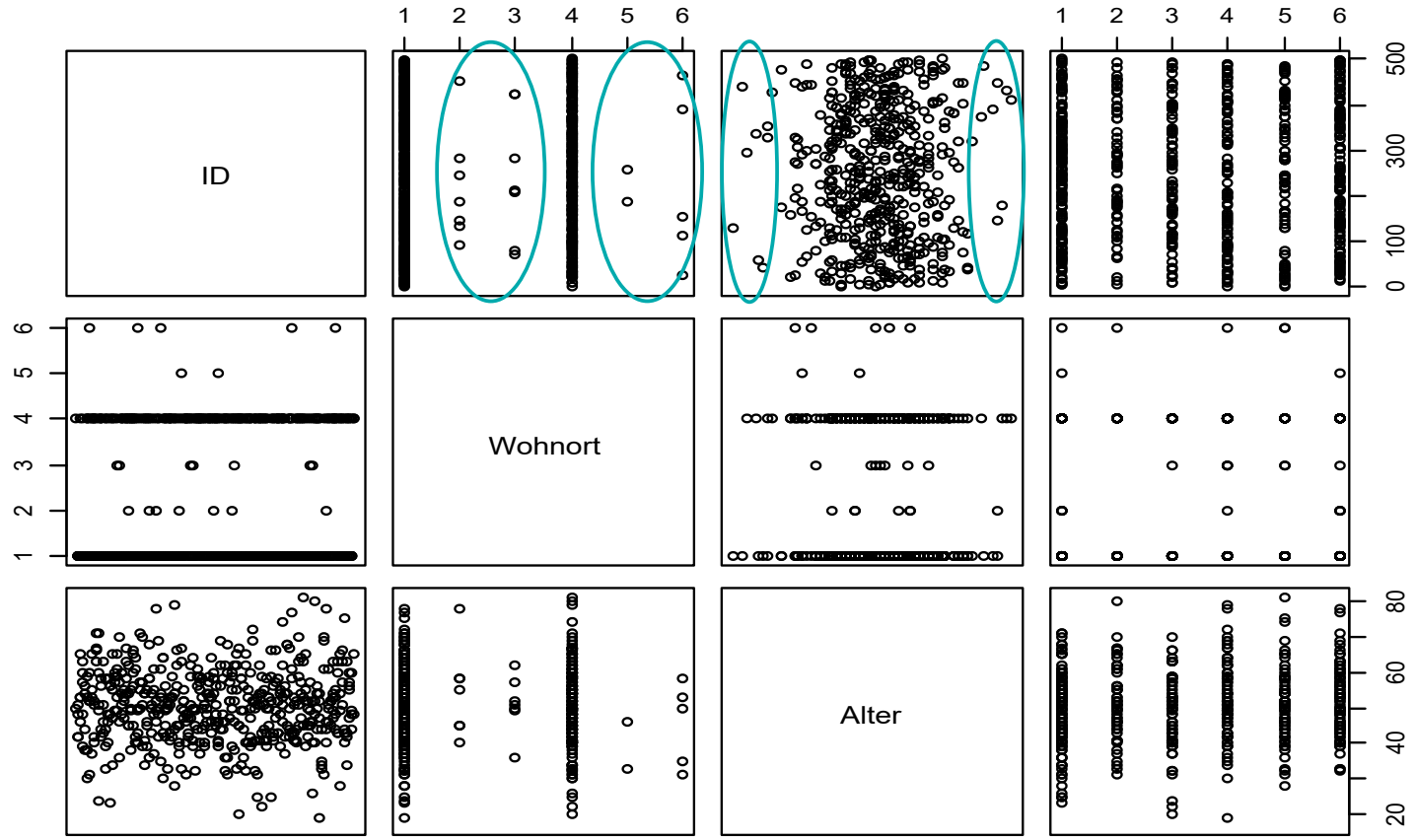
Aufgabe

- 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.