

Datenanalyse

Wie bekomme ich einen schnellen Überblick über die bereits eingegebenen Daten?

In SPSS:

Die Variablenansicht sollte bereits ordentlich ausgefüllt sein!

Daten öffnen:

Doppelklick auf den bestehenden SPSS-Datensatz (Dateiendung .sav)

Oder:

Datei -> Öffnen -> Daten

Einfache Analyse-Methoden

Analyse-Methoden, die geeignet sind, um sich einen Überblick über die Daten zu verschaffen und evtl. Fehler in den Daten zu finden:

- Häufigkeitstabellen
- Statistische Maßzahlen
- Kreuztabellen
- Korrelationen

Häufigkeitstabellen

Beispiel: Geschlecht

Ergebnis: absolute und relative Häufigkeiten

- Häufigkeit
- Prozent (relative Häufigkeiten, wobei auch fehlende Werte berücksichtigt werden)
- Gültige Prozente (relative Häufigkeiten für alle nicht-fehlenden Werte)
- Kumulierte Prozente

Vorgehensweise in SPSS

- Analysieren -> Deskriptive Statistiken -> Häufigkeiten
- Bei Statistiken: kein Häkchen notwendig
- Bei „Häufigkeitstabellen anzeigen“: Häkchen notwendig
- Für eine oder mehrere Variablen ausführen

Hinweise

- Besonders geeignet für nominale und ordinale Variablen, weniger geeignet für metrische Variablen mit sehr vielen verschiedenen Ausprägungen

Statistische Maßzahlen

Beispiel: Alter

Ergebnis

- Lagemaße:
 - Mittelwert (arithmetisches Mittel, Durchschnitt)
 - Median („mittlerer“ Wert in den geordneten Daten)
 - Modalwert (Modus, häufigster Wert)
- Streuungsmaße:
 - Standardabweichung
 - Varianz
- 5-Punkte-Zusammenfassung:
 - Minimum (0% Quantil / Perzentil)
 - unteres Quartil (25% Quantil / Perzentil)
 - Median (50% Quantil / Perzentil)
 - oberes Quartil (75% Quantil / Perzentil)
 - Maximum (100% Quantil / Perzentil)

(teilen die Daten in 4 Blöcke, die jeweils (ca.) 25% der Daten enthalten;
entspricht den Werten, die in einem Boxplot dargestellt werden)

Vorgehensweise in SPSS

- Analysieren -> Deskriptive Statistiken -> Häufigkeiten
- Bei Statistiken: mehrere Häkchen notwendig
- Bei „Häufigkeitstabellen anzeigen“: kein Häkchen notwendig
- Für eine oder mehrere Variablen ausführen

Hinweise

- Der Modus ist ab Nominalskala geeignet
- Die 5-Punkte-Zusammenfassung ist ab Ordinalskala geeignet
- Mittelwert, Varianz und Standardabweichung sind nur für metrische Variablen geeignet

Kreuztabellen

Beispiel: Geschlecht und Raucherstatus

Ergebnis: Zusammenhang zwischen zwei kategorialen Variablen (für nicht-fehlende Werte)

- absolute Häufigkeiten: Anzahl
- relative Häufigkeiten : % der Gesamtzahl
- bedingte relative Häufigkeiten: % innerhalb von Variable

Vorgehensweise in SPSS

- Analysieren -> Deskriptive Statistiken -> Kreuztabellen
- Eine Variable für die Zeilen, eine Variable für die Spalten
- Zellen: Häkchen setzen bei
 - Häufigkeiten: Beobachtet
 - Prozentwerte: Zeilenweise, Spaltenweise, Gesamtsumme

Hinweise

- Geeignet für kategoriale Variablen

Korrelationen

Beispiel: systolischer Blutdruck und diastolischer Blutdruck

Ergebnis: Zusammenhang zwischen zwei mindestens ordinalskalierten Variablen

- Korrelation nach Pearson (für zwei metrische Variablen, keine Ausreißer)
- (Kendall's Tau b)
- Korrelation nach Spearman (für ordinale Variablen oder metrische Variablen mit extremen Werten)

Vorgehensweise in SPSS

- Analysieren -> Korrelation -> Bivariat
- Variablen: zwei (oder mehr) Variablen angeben

Hinweise

- Im Output ist hauptsächlich der Korrelationskoeffizient von Interesse.

Nachvollziehbares Arbeiten in SPSS – Arbeiten mit Syntax

- Neue Syntax-Datei öffnen: Datei -> Neu -> Syntax
- Beim Zusammenklicken von Analysen:
 - Statt auf „OK“ auf „Einfügen“ klicken.
 - Dann wird die Syntax für die jeweilige Prozedur im Syntax-Fenster unten ergänzt.
 - (Manchmal heißt die Option zum Erzeugen der Syntax auch anders.)
 - Den relevanten Teil der Syntax markieren.
 - Den markierten Teil der Syntax durch Klicken auf „Auswahl ausführen“ (den grünen Pfeil / das grüne Dreieck) ausführen.
- Die Syntax **kommentieren!**
 - In jeder Kommentarzeile: vorne ein Stern (*), hinten ein Punkt (.)
 - Dementsprechend sollte innerhalb des Kommentares kein Punkt verwendet werden.
 - Beispiele:
 - * Dies ist ein Beispiel-Kommentar.
 - * Häufigkeitstabelle von Geschlecht.
 - * Statistische Kennzahlen von Alter.
 - * Kreuztabelle von Geschlecht und Raucherstatus.
 - * Korrelation von systolischem und diastolischem Blutdruck.
- Die Syntax (immer wieder) **aufräumen!**
 - Nur Syntax zu gelungenen Analysen behalten, damit die Datei übersichtlich bleibt
- Die Syntax **regelmäßig speichern!**
 - Beim ersten Mal: Datei -> Speichern unter
 - Danach: Klick auf Disketten-Symbol
 - Dateiendung .sps
- Die Syntax kann bei aktualisierten Daten neu ausgeführt werden, z.B.
 - Nach dem Ergänzen neuer Fälle
 - Nach der Korrektur der Angaben in der Variablenansicht
 - Nach dem Korrigieren „unplausibler“ Werte
 - Nach dem Ausschließen einzelner Fälle
- Vorteile beim Verwenden von Syntax
 - Zeitersparnis
 - Dokumentation der Analysen
- Was sollte man in Form eines Kommentars in die Syntax schreiben?
 - Alles, was man dokumentieren will ...
 - Welchen Datensatz man verwendet? (evtl. welche Version, welcher Stand)
 - „Was will man machen?“ (am besten vorher) oder „Was hat man gemacht?“ (nachher)

- Überlegungen / Begründungen, warum man etwas auf eine bestimmte Weise gemacht hat, z.B. Variablen umkodieren, Gruppen bilden, Fälle ausschließen usw.

Hilfreiches beim Analysieren mit SPSS

- Oft kann man **mehrere Variablen gleichzeitig analysieren**, z.B.
 - alle kategorialen Variablen (Häufigkeitstabellen)
 - alle metrischen Variablen (Mittelwert, 5-Punkte-Zusammenfassung)und so die Werte auf Plausibilität prüfen

- Man kann Analysen nur **für einen Teil der Daten** ausführen
 - Daten -> Fälle auswählen
 - Alle Fälle:
Zum Zurücksetzen der Einstellung
 - Falls Bedingung zutrifft (z.B. nur die Männer)
 - Zurücksetzen der Einstellung nicht vergessen!

- Man kann Analysen **getrennt nach Gruppen** ausführen
 - Daten -> Aufgeteilte Datei
 - Alle Fälle analysieren, keine Gruppen bilden:
Zum Zurücksetzen der Einstellung
 - Gruppen vergleichen:
Die Gruppen werden in den Tabellen direkt verglichen
 - Ausgabe nach Gruppen aufteilen:
Es erscheint erst der Output für die erste Gruppe, dann die zweite, usw.
 - Zurücksetzen der Einstellung nicht vergessen!

- **Daten umstrukturieren**
 - Wenn man pro Patient mehrere Erhebungszeitpunkte hat, ist es manchmal notwendig, die Daten für die Analyse in eine andere Form zu bringen
 - Format: lang, d.h. für jeden Patienten gibt es mehrere Zeilen, nämlich eine Zeile pro Erhebungszeitpunkt
 - Format: breit, d.h. für jeden Patienten gibt es eine Zeile, die Daten zu den verschiedenen Erhebungszeitpunkten stehen in verschieden benannten Variablen hintereinander
 - Daten -> Daten umstrukturieren
 - Wichtig: Daten vorher unter einem neuen Namen speichern!!!
Die alte Datei wird hier überschrieben. Gefahr von Datenverlust!

Einschub: Arbeiten am Computer

Hilfreich: Sinnvolle Ordnerstruktur anlegen

Welche Ordner brauche ich?

- Hauptorder für die Doktorarbeit / das Projekt
- Unterordner
 - Original-Daten (zum Sichern der Daten)
 - Daten (zum Arbeiten mit den Daten)
 - Syntax / Programme
 - Ergebnisse
 - Dokumentation

 - Literatur
 - fachlich
 - methodisch
 - Doktorarbeit (der Text)
 - Ethikantrag / Datenschutzantrag
 - Organisatorisches
 - ...
- Evtl. Dateien mit Datum im Dateinamen speichern

Regelmäßig die Daten sichern!

Datenqualität

Wie stelle ich eine hohe Datenqualität sicher?

Vorher

Die Basis: Das Codebuch

- Ein Codebuch erstellen und pflegen (gelegentlich ändert man Kodierungen während der Planung)
- Das Codebuch in SPSS übertragen (in die Variablenansicht), ebenfalls pflegen
- (Zusätzliche Variablen (z.B. aus bestehenden Variablen berechnete neue Variablen) ebenfalls dokumentieren)

Sorgfalt

- Die Daten sorgsam eingeben – lieber mehr Zeit investieren

Wie prüfe ich die eingegebenen Daten auf Plausibilität?

Nachher

Prüfen, ob bei kategorialen Variablen nur zulässige Werte eingegeben wurden.

- Drüber schauen:
In der Datenansicht auf „Wertbeschriftungen“ (rosa Pfeil mit A oben und 1 links) klicken und über die Daten schauen. (Voraussetzung: Die Variablenansicht in SPSS muss ordentlich ausgefüllt sein.)
- **Häufigkeitstabellen** anschauen
- Evtl. auch Kreuztabellen verwenden

Prüfen, ob sich der Wertebereich von metrischen Variablen in einem plausiblen Bereich befindet

- Drüber schauen (nach Klicken auf „Wertbeschriftungen“)
- **Statistische Maßzahlen** anschauen, insbesondere Minimum, Maximum und Mittelwert (müssen im zulässigen Bereich liegen)
- Evtl. auch Korrelationen verwenden

Gute Praxis: Doppelte Dateneingabe

- Die Daten bzw. einen Teil der Daten (einen Teil der Fälle) doppelt eingeben (falls möglich, von zwei verschiedenen Personen).
- Dann die beiden Datensätze miteinander vergleichen:
Dazu beide Datensätze in SPSS öffnen.
Dann: Daten -> Datasets vergleichen
 - Zu vergleichende Felder: alle Nicht-ID Variablen
 - Fall-IDs: alle ID-Variablen

Sonstiges

- Wie erstelle ich aus mehreren ID-Variablen eine Gesamt-ID?
 - In Excel: Funktion „Verketten“
 - In SPSS: Transformieren -> Variable berechnen
 - Evtl. muss das Format der Variablen dazu angepasst werden.
 - Nicht nochmal eintippen, zu fehleranfällig!

- Wie mache ich aus einer kategorialen Variable mehrere?
 - Transformieren -> Umkodieren in andere Variable
 - z.B. Familienstand_ledig

- Wie mache ich aus mehreren kategorialen Variablen eine?
 - Transformieren -> Variable berechnen
 - z.B. für Variablen mit möglichen Mehrfachantworten, wie Verkehrsmittel

- Wie erstelle ich eine Grafik?
 - Grafik -> Diagrammerstellung

- Wie gebe ich kategoriale Variablen mit sehr vielen Antwortkategorien (z.B. ICD-10 Codes) sinnvoll ein? Wie kann ich diese dann analysieren?
 - Vorher überlegen, wie man diese Variable(n) analysieren möchte und die Daten so eingeben, dass diese Analysen durchgeführt werden können
 - Vorschläge:
 - Alle Antwortkategorien in einer Variable als Freitext eingeben (als Trennzeichen weder Komma noch Strichpunkt verwenden, da dies die typischerweise verwendeten Spaltentrennzeichen sind)
 - Evtl. die wichtigsten Antwortkategorien (zusätzlich) als einzelne Variablen anlegen
 - Evtl. Antwortkategorien zusammenfassen / vergrößern (nur höhere Ebene im ICD-10 Code erfassen, z.B. nur den Buchstaben und die erste Zahl)

SPSS-Lizenz

- Lizenz beim LRZ (Leibniz-Rechenzentrum)
- <https://www.lrz.de/services/swbezug/lizenzen/spss/>
- Eine SPSS-Lizenz kostet 60,-
- Der Lizenzzeitraum läuft vom 01.04. des aktuellen Jahres bis zum 31.03. des Folgejahres.

Wo ist SPSS installiert?

- In vielen medizinischen Instituten
- Klinikum Großhadern, Hörsaaltrakt 2. OG, Kursraum 5

Möglichkeiten, um SPSS zu lernen

- SPSS-Kurse am IBE:
http://www.ibe.med.uni-muenchen.de/studium_lehre/kurse/spss-ss_17/index.html
- Workshops am IBE zu I1 und I2:
http://www.ibe.med.uni-muenchen.de/studium_lehre/mecum/wahlpflichtkurse/index.html
 - "Statistisches Testen in SPSS"
 - „Korrelation und Regression in SPSS“
 - „Risiken analysieren in SPSS“
- Online, z.B. Webseite des UCLA (Institute for Digital Research and Education)
<https://stats.idre.ucla.edu/spss/>