Leitfaden zur Datenaufbereitung

Inhalt

Beispiele

Dieser Leitfaden stellt eine Orientierungshilfe dar und beinhaltet Regeln die uns die Auswertung im Beratungsalltag erleichtern. Einige Regeln werden vereinfacht dargestellt.
Fragen Sie bei Unklarheiten oder im Zweifelsfall den beratenden Statistiker.

Allgemeines

Zur Datenerfassung sind alle Tabellenkalkulationsprogramme (Microsoft **Excel**, OpenOffice Calc, LibreOffice Calc, ... ; *im Folgenden mit „Excel und Co" bezeichnet*) gut geeignet. Die Tabellen können von den meisten Programmpaketen gelesen werden. Alternativ können Datensätze auch direkt in statistischer Auswertungssoftware wie **SPSS** oder **SAS** eingegeben werden.

Grundsätzlich erfolgt die Dateneingabe in Form **einer** Tabelle (Datenmatrix), wobei die Zeilen der Tabelle den Untersuchungseinheiten (z.B. Patienten) und die **Spalten den Variablen (Merkmalen)** entsprechen. Dabei wird in aller Regel **eine Zeile pro Beobachtung/Patient** angelegt (wiederholte Messungen siehe Eingabe/Kodierung der Antworten).

Benennung der Variablen (Merkmale)

1. Benennen Sie die Variablen mit aussagekräftigen aber kurzen Namen. Z.B. ist "01" weit weniger aussagekräftig als "Frage_01".
2. Nehmen Sie dabei möglichst Bezug auf Ihren Fragebogen / Datenerfassungsbogen. Achten Sie dabei jedoch bitte darauf, dass die Variablennamen nicht zu lang werden.
3. Beginnen Sie die Variablennamen mit einem Buchstaben.
4. Benutzen Sie dabei möglichst keine Sonderzeichen und Umlaute (Unterstrich _ ist erlaubt).
5. Schreiben Sie die Variablennamen in die erste Zeile Ihres Datensatzes (nur bei Excel und Co).

Eindeutige Zuordnung von Fragebögen

Die **erste Spalte** sollte die Nummerierung (ID) der Untersuchungseinheiten (Personen) enthalten. Um später Datenprobleme zu klären, ist es sinnvoll, die Nummer auch auf den Fragebögen (bzw. Datenerfassungsbögen) festzuhalten.

Wichtig:
Aus Datenschutzgründen ist, insbesondere bei personenbezogenen Daten, auf Anonymisierung zu achten! Entfernen Sie deshalb bitte Namen und ähnliche identifizierende Merkmale (z.B. Adressen) aus dem Datensatz! Mit Hilfe der vergebenen ID können Sie die Daten bei Bedarf den Patienten zuordnen.

Eingabe/Kodierung der Antworten

1. Bei **stetigen Größen** (z.B. Größe, Gewicht, Konzentrationsmessungen etc.) müssen *Dezimalzahlen immer einheitlich* entweder durch Punkt oder Komma getrennt werden.
2. Das Geburtsdatum oder **Datumsangaben** allgemein in einheitlichem Format (z.B. TT.MM.JJJJ) angeben.
3. Kodierung von binären oder kategoriellen Größen
   (= **zwei oder mehrere Ausprägungen**, z.B. Geschlecht oder Nationalität):
   a. Allgemein ist bei **Excel und Co** die direkte Eingabe der Kategorien (Wertelabels) in die Zellen zu bevorzugen, da diese somit direkt in Grafiken und Tabellen übernommen werden können. Fügen Sie in jedem Fall unbedingt eine Beschreibung der Kodierung bei (siehe Datensatzbeschreibung).
   In **SPSS** und **SAS** ist es besser die Antworten durchzunummerieren (tippt sich in der

Regel einfacher) und dann Wertelabels zu vergeben. (Eine zusätzliche Variablenbeschreibung ist somit in der Regel unnötig.)

b. Bei Fragen **ohne Mehrfachantworten** sind die Antworten in einer Variable (Spalte) aufzulisten.

c. Bei Fragen, bei **denen Mehrfachantworten** möglich sind, geht man wie folgt vor: Für jede Antwortmöglichkeit wird eine eigene Variable (Spalte) angelegt und die Antwortmöglichkeiten mit "nein" und "ja" oder mit 0 und 1 kodiert. Fügen Sie auch hier unbedingt eine Beschreibung der Kodierung bei (oder vergeben Wertelabels in SPSS/SAS).

4. Verwenden Sie in einer Variable **nicht Zahlen und Text gemischt**. Verwenden Sie in der gesamten Spalte das gleiche Muster. Z.B.
   - *Geschlecht:* maennlich / weiblich ***oder*** 0 / 1
   - *OP-Methode:* Methode 1 / Methode 2 / Methode 3 ***oder*** 1 / 2 / 3

   wobei die Texteingaben in aller Regel den Zahlen vorzuziehen ist (siehe 3.a).

   Achten Sie bei der Texteingabe unbedingt auf eine einheitliche Schreibweise (auch bzgl. Leerzeichen und Sonderzeichen) und **vermeiden Sie Leerzeichen** insbesondere am Anfang oder Ende der Kategorie. Zum Beispiel sind die Labels

   Methode␣1 (← bitte so)

   Methode␣1␣

   Methode␣1␣␣

   ␣Methode␣1

   Methode␣␣1

   nicht identisch (vgl. Stellung und Anzahl der Leerzeichen, dargestellt durch ␣).

5. Vermeiden Sie in den Kategorien **Sonderzeichen** (insb. Umlaute und ß).
6. **Fehlende Werte** werden am besten *einheitlich* mit einem . (Punkt) oder auch durch keine Eingabe (leeres Feld) kodiert.
   Bei unterschiedlichen Quellen für fehlende Werte können diese auch durch verschiedene Zahlencodes (z.B. -99, -999) kodiert werden. Das muss dann aber unbedingt in der Datensatzbeschreibung erwähnt werden.
7. Keine **Kommentare** in derselben Spalte dazu schreiben. Verwenden Sie eine extra Spalte (z.B. „Kommentar_zu_Var1“). Vermeiden Sie auch in den Kommentaren Sonderzeichen (insb. Umlaute und ß).
8. Bei **Mehrfachmessungen** (also Fragen / Variablen, die mehrmals über die Zeit erhoben werden) gibt es mehrere Möglichkeiten: Wurde lediglich eine (oder wenige) Variable(n) mehrfach erhoben, so kann man für jede Messung eine Spalte einfügen und den Zeitpunkt in den Variablennamen aufnehmen (z.B. Blutdruck_T0, Blutdruck_T1, Blutdruck_T2).
   Wurden mehrere Variablen oder gar alle Variablen mehrfach erhoben (i.d.R. alle bis auf die soziodemographischen Daten), dann fügt man am besten pro Messzeitpunkt eine Zeile in den Datensatz ein (siehe Beispiel 2a: Klinische Studie mit Messwiederholung). Im Zweifel wenden Sie sich bitte direkt an uns.

Datensatzbeschreibung

Die Beschreibung der Daten sollte folgende Informationen beinhalten:

1. **Variablenname**
2. **Variablenlabel**: Bedeutung der Variable (z.B. Langform des Variablennamens)
3. (unter Umständen: Variablentyp)
4. **Einheiten** (*bei stetigen Variablen*; evtl. auch direkt in Variablenlabel, z.B. „Gewicht (in kg)“)
5. **Kodierung** unter Berücksichtigung der Reihenfolge der Kategorien (*bei kategoriellen Variablen*)
6. (unter Umständen: Kodierung der fehlenden Werte (z.B. -999))

Besonderheiten:

- **Excel und Co:** Fügen Sie eine Beschreibung der Kodierung in einer gesonderten Datei oder auf einem gesonderten Datenblatt bei. *Unter dem Datensatz ist die Beschreibung nicht sinnvoll!*
  Am besten sollte diese Beschreibung auch eine (separate) **Excel-Tabelle** sein, die in der ersten Spalte die Variablennamen und in den weiteren Spalten die Variablenlabels, Kodierungen etc. enthält.
- **SPSS:** In der Regel reicht es, wenn Sie den Datensatz mit Hilfe der *Variablenansicht* beschreiben (nicht in der *Datenansicht* möglich). Geben Sie hier unbedingt die Variablenlabel, Wertelabels (für kategorielle Variablen) und bei Bedarf die Kodierung der fehlenden Werte an. Eine *zusätzliche* Datensatzbeschreibung ist gegebenenfalls hilfreich.
- **SAS:** In SAS können ebenfalls Variablen- und Wertelabels (PROC FORMAT) vergeben werden, sowie spezielle Kodierungen für fehlende Werte spezifiziert werden. Unter Umständen wird für die weitere Analyse das verwendete SAS-Skript zur Formatierung benötigt oder die Daten müssen mit den entsprechenden Kodierungen abgespeichert werden (kategorielle Variablen als Characters). Eine *zusätzliche* Datensatzbeschreibung ist gegebenenfalls hilfreich.

Tipp: Wenn Sie die Analysen und Grafiken in Englisch benötigen (für Publikationen), dann empfiehlt es sich, die Variablen- und Wertelabels (= Kodierung) auf Englisch zu verfassen (und evtl. die Variablennamen).

Wichtig für Benutzer von Excel und Co

Speichern Sie alle Daten in **einer** Tabelle (sprich **einem Datenblatt**) ab. Fügen Sie also alle Datenblätter in einem zusammen und ergänzen Sie (wenn nötig) eine zusätzliche Variable, welche die Datenblätter kennzeichnet (z.B. Untersuchungsgruppe).

Verwenden Sie **keine Einfärbung** von Zellen, **keine Formeln, keine Pfeile, keine Notizen** oder **Zellformat-Änderungen** (wie z.B. Zellen verbinden). Grafiken oder erste Analysen haben ebenfalls nichts in der Datentabelle zu suchen. Halten Sie die Tabelle so einfach wie möglich!

Beispiele zur Datenaufbereitung

Beispiel 1:
Einfacher Fragebogen mit Mehrfachantworten

Datensatz zu einer Studie mit Fragebogen. Dabei gibt es Beispiele für kategorielle Antwortmöglichkeiten *mit* Mehrfachantwortmöglichkeiten (Frage 2) und *ohne* Mehrfachantwortmöglichkeiten (Fragen 1 und 3). Die Kodierung im Datensatz erfolgt in diesem Beispiel über Zahlen. **In Excel und Co ist Beispiel 2 bzgl. der Kodierung zu bevorzugen.**

Fragebogen

Laufende Nummer: _____

Frage 1: Geschlecht

- o weiblich
- o männlich

Frage 2: Was sind Ihre Hobbies? (Mehrfachantworten möglich)

- o Fragebögen gestalten
- o Datensätze erstellen
- o Texte verfassen

Frage 3: Welche Fragebögen füllen Sie am liebsten aus?

- o kurze
- o sehr kurze
- o keine

Variablenbeschreibung

Extra Dokument (vgl. Datensatzbeschreibung)

Variablenname	Label	Kodierung
ID	Fragebogennummer	
Geschlecht	Geschlecht	1 = männlich 0 = weiblich
Hobby_Fragebogen	Fragebögen gestalten	1 = ja (angekreuzt) 0 = nein (nicht angekreuzt)
Hobby_Daten	Datensätze erstellen	1 = ja (angekreuzt) 0 = nein (nicht angekreuzt)
Hobby_Texte	Texte verfassen	1 = ja (angekreuzt) 0 = nein (nicht angekreuzt)
Fragebogenart	Welche Fragebögen füllen Sie am liebsten aus?	1 = kurze 2 = sehr kurze 3 = keine

Datensatz

ID	Geschlecht	Hobby_Fragebogen	Hobby_Daten	Hobby_Texte	Fragebogenart
01	1	1	0	1	1
02	0	0	0	0	2
03	1	1	1	0	2
04	1	0	1	0	2
05	1	0	0	1	2
06	0	1	1	1	3
07	1	1	0	1	3
08	0	0	0	0	1

Beispiel 2a:
Einfache klinische Studie mit stetigen und kategoriellen Merkmalen

Ein neues Medikament zur Senkung des Blutdrucks soll getestet werden. Dazu wird eine klinische Studie durchgeführt, bei der das neue Medikament A mit einem etablierten Medikament B verglichen wird. Die Analysen sollen dabei in Englisch erfolgen (sinnvoll für spätere Publikation).

Variablenbeschreibung

Extra Dokument (vgl. Datensatzbeschreibung)

Variablenname	Label	Typ	Kodierung / Einheiten
id	Patient ID	Numerisch	
sex	Sex	Kategoriell	Male Female
group	Therapy Group	Kategoriell	Medication A Medication B
blood_sys	Systolic Blood Pressure	Stetig	mmHg
blood_dia	Diastolic Blood Pressure	Stetig	mmHg

Datensatz

id	sex	group	blood_sys	blood_dia
1	Male	B	152	89
2	Male	A	145	85
3	Female	A	147	82
4	Male	A	142	78
5	Male	B	154	93
6	Female	B	151	90

Beispiel 2b:
Klinische Studie mit Messwiederholung

Um den Effekt der Senkung des Mittels besser einschätzen zu können werden zusätzlich die Baseline Werte der Patienten erhoben. Für die restlichen Daten siehe Beispiel 2: Einfache klinische Studie mit stetigen und kategoriellen Merkmalen.

Möglichkeit 1

Als erste Möglichkeit kann man für die mehrfach gemessenen Variablen einfach weitere Variablen angelegen.

Variablenbeschreibung

Variablenname	Label	Typ	Kodierung / Einheiten
id	Patient ID	Numerisch	
sex	Sex	Kategoriell	Male Female
group	Therapy Group	Kategoriell	Medication A Medication B
blood_sys_t0[1]	Systolic Blood Pressure (T0, pre treatment)	Stetig	mmHg
blood_dia_t0	Diastolic Blood Pressure (T0, pre treatment)	Stetig	mmHg
blood_sys_t1	Systolic Blood Pressure (T1, post treatment)	Stetig	mmHg
blood_dia_t1	Diastolic Blood Pressure (T1, post treatment)	Stetig	mmHg

Datensatz

id	sex	group	blood_sys_t0[1]	blood_dia_t0	blood_sys_t1	blood_dia_t1
1	Male	B	160	94	152	89
2	Male	A	152	90	145	85
3	Female	A	165	91	147	82
4	Male	A	143	80	142	78
5	Male	B	165	98	154	93
6	Female	B	163	99	151	90

[1] **Anmerkung:** Die neuen Variablen sind (hier) rot eingefärbt um die Änderung im Vergleich zu Beispiel 2 leichter zu erkennen. In echten Datensätzen bitte keine Einfärbungen. Diese können von Statistik-Software ohnehin nicht erkannt werden.

Möglichkeit 2

Als zweite Möglichkeit kann eine weitere Variable für den Zeitpunkt eingeführt werden und für jeden Patienten mehrere Zeilen angelegt werden. Dabei unterscheiden sich diese nur in den zeitlich veränderlichen Variablen, nicht aber in den soziodemographischen Daten und anderen Baseline Messungen.

Variablenbeschreibung

Variablenname	Label	Typ	Kodierung / Einheiten
id	Patient ID	Numerisch	
time[2]	Time	Kategoriell	T0 (Pre Treatment) T1 (Post Treatment)
sex	Sex	Kategoriell	Male Female
group	Therapy Group	Kategoriell	Medication A Medication B
blood_sys	Systolic Blood Pressure	Stetig	mmHg
blood_dia	Diastolic Blood Pressure	Stetig	mmHg

Datensatz

id	time[2]	sex	group	blood_sys	blood_dia
1	T0	Male	B	160	94
1	T1	Male	B	152	89
2	T0	Male	A	152	90
2	T1	Male	A	145	85
3	T0	Female	A	165	91
3	T1	Female	A	147	82
4	T0	Male	A	143	80
4	T1	Male	A	142	78
5	T0	Male	B	165	98
5	T1	Male	B	154	93
6	T0	Female	B	163	99
6	T1	Female	B	151	90

[2] **Anmerkung:** Neue Werte (Pre Treatment) und die neue Variable sind (hier) rot eingefärbt um die Änderung im Vergleich zu Beispiel 2 leichter zu erkennen. In echten Datensätzen bitte keine Einfärbungen. Diese können von Statistik-Software ohnehin nicht erkannt werden.