Anforderungen an einen Datensatz zur statistischen Analyse

Departement Biostatistik
Institut für Epidemiologie, Biostatistik und Prävention
Universität Zürich

Wir analysieren Datensätze nach den Guidelines für Reproduzierbare Forschung in der Biostatistik. Für das Einlesen Ihres Datensatzes in die Programmiersprache R benötigt der Datensatz eine gewisse Struktur. Daher bitten wir Sie, die folgenden Anforderungen zu beachten und möglichst gut umzusetzen. Falls Sie damit Schwierigkeiten haben, können wir Sie dabei evtl. unterstützen.

1. Der Datensatz sollte nur diejenigen Variablen enthalten, die effektiv für die statistische Auswertung benötigt werden. Zusätzliche Variablen sollten entfernt, bzw. in einer anderen Version des Datensatzes gespeichert werden.

2. Die Daten müssen in einer Tabelle angeordnet werden. Zeilen repräsentieren die Beobachtungseinheiten, Spalten die beobachteten Merkmale (Variablen). Ein Datensatz mit wiederholten Messungen über die Zeit kann in einem sog. `long` oder `wide` Format dargestellt werden, Details siehe unter Punkt 10.

3. Die erste Spalte ist üblicherweise die Identifikationsvariable (ID) der Beobachtungseinheit.

4. Die Variablennamen müssen in der ersten Zeile stehen. Sie sollten kurz, eindeutig und verständlich sein. Das erste Zeichen jedes Variablennamens muss ein Buchstabe sein. Nicht erlaubt sind Leerschläge, Umlaute, Sonderzeichen und Satzzeichen (ausgenommen Punkt).

   Erlaubt: `ID`, `Height`, `Weight.pre`, `pCO2`, `T1a`, ...

   Nicht erlaubt: `Gebärmutterhalsumfang`, `Alk‰`, `lebt/tot`, `#Infarkte`, `5-Pkt.Fixierung`, `Beck'scher Wert`, ...

5. Dieselben Regeln gelten für die Bezeichnungen der Stufen bei Faktoren (kategorielle Variablen). Sie sollten kurz, eindeutig und verständlich sein. Nicht erlaubt sind Umlaute.

   Erlaubt: `female`, `male`; `no`, `yes`; `alive`, `dead`; `0-10mm`, `10-20mm`, `>20mm`; ...

   Nicht erlaubt: `männlich`; ...

6. Numerische Variablen dürfen nur Ziffern, das negative Vorzeichen (−) und den Dezimalpunkt enthalten. Angaben wie `>10000` oder `<0` bitte korrigieren.

   Hinweis: In Excel werden Zahlen rechtsbündig dargestellt; sind sie fälschlicherweise als Text formatiert (oder enthalten Buchstaben oder Zeichen wie `>` oder `<`), stehen sie meist linksbündig.

7. Falls der Datensatz fehlende Werte enthält, diese Felder oder Zellen bitte leer lassen und nicht stillschweigend mit einem Leerschlag, 0, 99, 9999, o. ä. Werten auffüllen. Falls ein Wert oder eine alphabetische Zeichenfolge eingesetzt wurde, bitten wir Sie, diese klar zu deklarieren. Besonders bei SPSS Datensätzen kommt es häufig vor, dass fehlende Werte durch 99, oder 9999 ersetzt werden.

8. Auf Daten oder Zeiten basierende Grössen wie z. Bsp. Alter oder OP-Dauer bitte vorab berechnen und die Resultate im Datensatz liefern. Falls Kalenderdaten wichtig sind, bitten wir Sie, diese als Text darzustellen. Bitte Vorsicht mit der amerikanischen Darstellung Monat/Tag/Jahr.

   Empfohlen: `2014-11-23`

   Nicht empfohlen: `2014-November-23`, `2-August-2011`, `07.05.2012`, `10/06/11`.

9. Bitte schicken Sie keine Datensätze mit Patientennamen, alle Angaben sollten durch eindeutige Identifikationsnummern anonymisiert werden. Bitte entfernen Sie auch Adressen und Ortsangaben (ausser für räumliche Analysen, hier gelten spezielle Regeln).

   Achtung: Patientenlisten befinden sich oft in separaten Tabellenblättern oder in ausgeblendeten Spalten von Excel-Dokumenten.

10. Wird ein Merkmal zu mehreren Zeiten erhoben, kann jeder Messzeitpunkt eine eigene Spalte (und damit einen eigenen Variablennamen) erhalten.

ID	Geschlecht	Alter	Gewicht.vor	Gewicht.nach
1	m	27	82	79
2	f	39	71	66
3	f	31	77	72
⋮	⋮	⋮	⋮	⋮

   Falls eine Messung sehr oft und/oder in unregelmässigen Abständen wiederholt wird, oder falls sehr viele der Variablen mehrfach erhoben worden sind, kann stattdessen pro Beobachtungseinheit und Zeitpunkt eine eigene Zeile erfasst werden. Dadurch werden zwei Identifikations-Variablen nötig: eine für die Beobachtungseinheit (ID) und eine für den Messzeitpunkt. Bitte wiederholen Sie sich nicht ändernde Angaben in jeder Zeile.

ID	Geschlecht	Alter	Gewicht	Behandlung	Zeit	Plasmakonz	BD_sys	Puls
1	m	27	79	A	0	124	138	109
1	m	27	79	A	10	102	133	97
1	m	27	79	A	30	88	129	82
2	f	39	66	B	0	119	142	110
2	f	39	66	B	20	96	128	101
2	f	39	66	B	35	89	121	92
2	f	39	66	B	60	79	118	83
3	f	31	72	A	2	128	143	115
3	f	31	72	A	30	101	127	99
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

11. Bitte erstellen Sie ein separates Dokument oder Tabellenblatt mit Erklärungen zu den Variablen, ein sog. Codebook:

    a) Dieses beinhaltet eine Liste, welche die Namen aller Variablen, deren Masseinheit sowie eine kurze Beschreibung (in Worten) enthält.

    b) Bei *stetigen* und *Zählgrössen* bitte die untere und obere mögliche Grenze (Minimum, Maximum) angeben.

       Beispiele:
       - Visual Analog Scale (VAS): 0–10 cm.

    c) Bei *nominalen* Grössen bitte sämtliche möglichen Ausprägungen und Code angeben.

       Beispiele:
       - Schwanger: 0 (nein), 1 (ja)
       - Status: 0 (lebt), 1 (tot)
       - Zelltyp: Ec (Erythrozyt), Bs (basophiler Granulozyt), Eo (eosinophiler Gr.), Neu (neutrophiler Gr.), Mn (Monozyt), Lym (Lymphozyt), Tc (Thrombozyt).

    d) Bei *ordinalen* Grössen bitte sämtliche möglichen Ausprägungen und Ordnungsstruktur angeben.

       Beispiele:
       - NYHA Klasse: 1 (keine Symptome), 2 (milde Symptome), 3 (Einschränkungen im Alltag), 4 (schwerst eingeschränkt)
       - Opioide (in aufsteigender analgetischer Potenz): Mo (Morphium), Oxy (Oxycodon), Fen (Fentanyl), Rem (Remifentanil), Suf (Sufentanil)

12. Keine Informationen mittels Formatierungen (Schriftart, Schriftfarbe, Zellenfarbe) markieren, sondern als eigenständige Variable codieren.

13. Als Dateiformat bitten wir Sie, wenn möglich `.xlsx`, `.csv` oder `.txt` zu verwenden. Andere Formate sind nach Rücksprache ggf. auch möglich.