UNIVERSITÄT
BIELEFELD
Medical School OWL

Biostatistics and Medical Biometry

Guidelines for entering data

Compliance with the following instructions enables efficient data management and systematic evaluation by third parties. All information is provided without any claim to completeness.

General information

- for efficient evaluation, the focus is on the data, not the design, as data are usually exported and analyzed in other software
- content of the rows: units (e.g., `patients`, `cells`, `mice`)
- content of the columns: features of the units (e.g., `height`, `age`)
- every file should only contain a single tabular
- for every dataframe there should be a code book, denoting the encoding of all variables along with a short description

Naming the variables

- first column: unique `ID` of the unit
- first row: name of the variable
- informative, **short** variable names
- variable names should not contain any special characters (e.g., `*`, `-`, `/`, spaces). An exception is the underscore (“_”)
- variable names should not start with a number
- the dataframe should not contain additional rows to the data, e.g., a title
- the dataframe should not contain empty rows

Content of the dataframe

- numbers are easier to handle than words in the analysis and more clear in the tabular as well. Therefore, categorical variables should be encoded using numbers instead of characters (e.g., sex `0` or `1` instead of “male” or “female”)
- for binary variables (two possible outcomes): `0` or `1` instead of (e.g.) “no” or “yes”
- enter dates in the form `YYYY-MM-DD`
- if a variable in text form is unavoidable: if possibly, use only a single word without special characters (pay attention capitalization and uniformity)
- avoid free-text information if possible (danger of typos or unclear information)
- all information have to be linked to the `ID` of the unit. Multiple tabulars should be linked exclusively through the unit `ID`
- no calculations should be conducted in the raw data files

- in cases of multiple measurements per unit, every time of measurement should be denoted in a new row
- avoid entering several entries in a single cell/column. Example: Blood pressure (systolic/diastolic) 120/80 in two columns instead of one
- units of measurement should not be specified in the data file, but in the codebook
- each column may only contain values in one format. For example, numeric and alphanumeric (character) values should not be confused

Missing values

- avoid missing values if possible
- if a value is missing, leave the corresponding cell in the tabular **empty**. No “missing”, “N/A” or similar

Example of a good data structure

ID	sex	time	insulin
321	0	0	
321	0	5	0.205
321	0	10	0.129
322	1	0	0.251
322	1	5	2.228
322	1	15	2.078

- all data unchanged over time should still be reported in every row

Example of a bad data structure

		Week 4			Week 6			Week 8		
Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
3434	M	11/22/2006	26.6	238.9	12/6/2006	45.9	378	4/27/2007	56.2	409.8