Submitting data files for analysis

Ideally, we would like to receive one data file, populated with numeric data (and their definitions), from one person. A great way to provide definitions for your data is via a data dictionary provided separately to your data. An example of a data dictionary is as follows:

Variable name	Label	Type (Width)	Value codes	Missing Code
ID	Identification number	String (8)	none	none
Gender		Numeric (1.0)	1=Female	999
			2=Male	
Age	Age at study commencement	Numeric (3.0)	none	999
DOB	Date of birth	Date (10) (dd/mm/yyyy)	none	none
Height	Height (m)	Numeric (4.2)	none	999
TestResults	Test score	Numeric (5.0)	none	none
BMI	Body mass index	Numeric (1.0)	1=<25kg/m2	888
			2=25-29.9kg/m2	
			3=30-34.9kg/m2	
			4=35+kg/m2	

Data must be submitted in electronic format. The data can be in text, Excel, Access, SAS, or SPSS format. If you convert your file to a different format please check that it has not been corrupted in the process.

The subject's personal details should either be removed or numerically coded in the final data file. A subject should be identified by their primary ID number or another numeric code. Consider using Year of Birth instead of Date of Birth as this further ensures confidentiality. A subject's personal information should not be sent via email. You might not be able to release certain types of data without consent.

If you do have to transfer confidential information consider password protecting a file or table and passing on the password by phone or mail.

Data files should be screened for errors before being submitted. It is not the statistician's job to screen and clean data.

Do not email files over 5 megabites. Larger files should be transferred by USB or Q-DOCS. Please run a virus check before attaching your file(s). The QIMR Berghofer server blocks files that it deems to be suspicious and does not immediately inform the intended receiver. Please precede all emails with attached files, with a short warning email so that we are aware that a file is coming and can respond if it does not.

Good data practice

Many readers seem to assume that articles published in peer-reviewed journals are scientifically sound, despite much evidence to the contrary
- Altman (2002)

The following data set has some good qualities and bad:

Patient	Procedure	Survival_time	age	gender	no_of_prior_heart_conds	kidney_disease
1	1	4	999	f	0	no
2	2	1	86	m	3	yes
3	1	5	62	f	1	no
4	1	4	59	f	3	No
5	2	1	61	f	2	yes
6	2	5	86	m	1	yes
7	1	5	60	F	3	Yes
8	1	3	57	m	1	no
9	1	1	56	m	1	no
10	1	5	56	m	1	yes
11	2	20	90	m	1	no
12	2	1	89	f	3	no
13	1	4	62	f	1	no
14	1	3	75	m	0	yes
15	2	5	76	M	2	no
16	2	4	81	f	0	no
17	2	5	71	fm	3	no
18	1	4	66	m	0	no
19	2	3	57	m	3	no
20	2	3	58	f	10	yes
21	2	3	78	m	1	no
22	1	5	60	f	0	no
23	2	5	84	f	2	no
24	0	1	84	m	0	no
25	1	3	61	m	2	yes
26	2	2	77	m	2	no
27	1	5	84	f	0	No
28	2	5	62	m	2	yes
29	1	1	73	f	1	no
30	2	3	85	m	1	Yes

The good:

- All information about one patient is contained in a row
- The columns are different variables
- All the variable names are clear and have NO spaces (some programs do not like spaces).

However, the data set fails the first statistical test: can we trust the data?

The bad:

- Column 2 (Procedure): Person 24 has a 0 for procedure - is this correct?
    - Inform us what 0,1,2 represent in a separate file.
- Column 3 (Survival_time): Person 11 has a survival time of 20 - is this correct?
    - Are all measurements recorded with the same units? Or is this an error? Or is it just representative of natural variability?
- Column 4 (age): Is an age of 999 correct?
    - Go back and check all your data points are possible.
- Column 5 (gender): What is the difference between m and M or f and F?
- Column 7 (kidney_disease): What is the difference between yes and Yes or no and No?
    - The statistics programs are case sensitive. Be consistent.
- Sometimes we observe unusual values - this does not mean they are incorrect it just means they are unusual. Do not delete them without strong justification.

To avoid these problems you can check:

- data ranges (are max and min values plausible?)
- data summaries for continuous variables (mean, standard deviation, range)
- frequencies of categorical variables (are there any incorrect categories?)
- cross tabulations of variables (there should be no pregnant males for example)
- missing values (why are they missing? How have you demonstrated that they are missing in your data set?)