



New York State Agricultural Experiment Station
Geneva, New York

Computer Centre: 373 (replaces 352)/ 21-Oct-98

Preparation of data for statistical analysis

This note describes data formats that are compatible with most large-scale statistical analysis programmes (e.g. Genstat, Minitab, SAS, S-PLUS). If data is prepared in accordance with these instructions, it will expedite communication between research and support service staff. Any departure therefrom will delay analysis since time and staff will have to be assigned for additional editing.

Of course, data should be collected in the form that is most convenient to the research project. When submitting data for analysis, these guidelines should be followed.

- **In general**, a spreadsheet programme will be most convenient for preparing data. Microsoft Word, Excel, or any 'text only' (ASCII) file is acceptable as long as it is appropriately formatted.
- **Do not include header, trailer information, subtotals, or other extraneous information.** For descriptive purposes you may, if you wish, include one row giving variate names.
- **Format data so that each variate is in its own single column.** For example,

<i>Variety</i>	<i>Rep</i>	<i>Response</i>
ANJOU	1	183
ANJOU	2	177
ANJOU	3	192
BARTLETT	1	186
BARTLETT	2	183
BARTLETT	3	174

and not

<i>Variety</i>	<i>Rep 1 response</i>	<i>Rep 2 response</i>	<i>Rep 3 response</i>
ANJOU	183	177	192
BARTLETT	186	183	174

and not

<i>Variety</i>	<i>Rep</i>	<i>Response</i>	<i>Variety</i>	<i>Rep</i>	<i>Response</i>
ANJOU	1	183	BARTLETT	1	186
ANJOU	2	177	BARTLETT	2	183
ANJOU	3	192	BARTLETT	3	174

There may be occasional exceptions to this rule (for example repeated measurement designs): if in doubt, see a statistician beforehand.

- **Columns must be explicitly filled with data:**

<i>Variety</i>	<i>Rep</i>	<i>Response</i>
ANJOU	1	183
ANJOU	2	177
ANJOU	3	192
BARTLETT	1	186
BARTLETT	2	183
BARTLETT	3	174

and not

<i>Variety</i>	<i>Rep</i>	<i>Response</i>
ANJOU	1	183
	2	177
	3	192
BARTLETT	1	186
	2	183
	3	174

- **All columns must have the same number of rows.**
- **Missing data** (empty cells) should not be left blank. Use a * or a . to indicate missing data.
- **Do not embed weird characters in names.** In particular, spaces should be avoided. Use of underbar (_) in place of space is recommended. For example, *Red_Bartlett*.
- **Make sure that names are consistent** with respect to spelling and case. No programmes will know that *Anjou* and *Anjoo* are the meant to be the same thing; some programmes will not know that *Anjou* and *anjou* are meant to be the same thing.
- **If there are multiple data files**, make sure that variates are in the same relative columns. If one file looks like

<i>Variety</i>	<i>Rep</i>	<i>Diameter</i>	<i>Height</i>
ANJOU	1	1.17	183
ANJOU	2	1.12	177
ANJOU	3	1.30	192
BARTLETT	1	1.32	186
BARTLETT	2	1.22	183
BARTLETT	3	1.19	174

and another looks like

<i>Variety</i>	<i>Rep</i>	<i>Height</i>	<i>Diameter</i>
ANJOU	1	183	1.17
ANJOU	2	177	1.12
ANJOU	3	192	1.30
BARTLETT	1	186	1.32
BARTLETT	2	183	1.22
BARTLETT	3	174	1.19

presupposing that the difference are not overlooked (hmm), clerical rectification will be required.

- **If there are multiple data files**, do not rely on the file names to carry variate information. For example, if separate files are used for the results of two treatments, include a column in each file containing the name of the treatment. For example,

Filename: *Anjou*

<i>Variety</i>	<i>Rep</i>	<i>Height</i>	<i>Diameter</i>
ANJOU	1	183	1.17
ANJOU	2	177	1.12
ANJOU	3	192	1.30

Filename: *Bartlett*

<i>Variety</i>	<i>Rep</i>	<i>Height</i>	<i>Diameter</i>
BARTLETT	1	186	1.32
BARTLETT	2	183	1.22
BARTLETT	3	174	1.19

and not

Filename: *Anjou*

<i>Rep</i>	<i>Height</i>	<i>Diameter</i>
1	183	1.17
2	177	1.12
3	192	1.30

Filename: *Bartlett*

<i>Rep</i>	<i>Height</i>	<i>Diameter</i>
1	186	1.32
2	183	1.22
3	174	1.19

- **STATISTICAL ISSUES**

The following recommendations are concerned with statistical analysis rather than clerical matters.

- **Counted proportion data.** If data consists of counted proportions, e.g. number of individuals responding out of total number of individuals, do not reduce the data to percentages or proportions beforehand. It is recommended that both numerator and denominator of the proportion be entered as separate columns. For example,

<i>Dose</i>	<i>Total</i>	<i>Dead</i>		<i>Dose</i>	<i>% Dead</i>
1	90	5	not	1	5.5
2	85	30		2	35.3
3	93	60		3	64.5

It is easy to compute proportions during the analysis if they are required, but alternative analyses such as logistic regression may be precluded if original counts are unavailable.

- **Polytomous data.** [A generalisation of the comments for counted proportions] If data consists of numbers falling into a number of mutually exclusive classes, do not reduce to proportions or percentages beforehand, but enter the integer counts.

<i>Red</i>	<i>White</i>	<i>Blue</i>		<i>Red</i>	<i>White</i>	<i>Blue</i>
10	25	2	not	27.0	67.5	5.4
5	50	1		8.9	89.3	1.8