Writing statistics project reports

S. Zachary

This is a draft version of this document. It is intended to provide some further guidance on writing reports for statistics practical projects. The need for such additional guidance is apparent from the thousands of reports we have marked in recent years. We are getting desperate!

Any comments are very welcome. Please send them to S. Zachary or J. Hansen.

1 Intended readership

You should assume that your report is to be read by someone who is scientifically and statistically literate. Therefore you do not need to describe standard statistical procedures (for example, the derivation of confidence intervals, or indeed anything else taught in lecture courses). You should, however, understand these yourself! In addition, your analysis must be clear and capable of being reproduced by the reader (see Section 2). This means that you *will* need to describe carefully any *nonstandard* procedures.

2 Structure and content

You should divide your report into clearly-titled sections. It may be helpful to number these to assist in cross-referencing.

In many scientific disciplines it is customary to have separate sections along the following lines: *introduction*, *methods*, *results*, *discussion*, *summary and conclusions*, *appendices* (frequently the *summary and conclusions* section comes first). Sections may be divided into subsections as appropriate. This scheme is perhaps somewhat artificial in the present context. With the advent of modern computing facilities, statistical analysis has become a highly interactive process, and *methods*, *results* and *discussion* are perhaps best replaced by a number of *analysis* sections. Further it is frequently simplest to include the content of the *summary and conclusions* section in the *introduction*. We therefore suggest the following scheme.

Introduction and Summary

This should give a brief description of both the problem and the available data. This should be followed by a brief *summary* of the analysis, any further discussion appropriate at this stage, and a clear and concise statement of your conclusions.

Analysis

Unless the project is very short, you will require several *analysis* sections. Each of these should clearly indicate the statistical methodology used (there is in general no need to describe the underlying theory), and report and discuss the results obtained. Such is the interactive nature of data analysis that you will usually require some iteration of these activities.

It is important that you describe your analysis sufficiently clearly and carefully to enable it to be reproduced by the reader.

Use plots to display patterns and relationships graphically. These are far more effective than numerical summaries. However, avoid repetition (see Section 6).

Do remember to *include* your main results (estimates obtained, etc), as well as discussing them. Any lengthy tables of results may reasonably be relegated to an appendix. Quantities such as fitted values and residuals are usually best reported graphically, although it *may* be appropriate to include, for example, a table of residuals in an appendix.

Appendices

Appendices are for necessary technical detail (for example, anything reasonably required to enable you results to be reproduced) which would otherwise clutter the remainder of your report. In general the number of appendices should be kept to a minimum-possibly zero.

In particular, there is usually no need for the inclusion of computing code.

3 Use of English

This is very important. The object of writing, as of speaking, is to *communicate*. Writing, in particular, requires very careful use of language, especially when abstract and difficult concepts (such as those involved in statistical modelling and analysis) are to be communicated. It is easy to read and understand English which is well written and follows the rules of good usage. However, as the quality of writing deteriorates, it becomes progressively more difficult for the reader to work out its intended meaning. Finally it becomes impossible.

While many project reports are written to a very high standard, many others contain much that is unintelligible. An obvious, and easily corrected, problem occurs when the author fails to use properly constructed sentences. (A good first check here is to identify the verb in each sentence!)

A much more insidious (and very frequent) problem occurs when the author unconsciously assumes that the reader is somehow following his (or her) quite unwritten train of thought. Thus the reader is presumed to know exactly what the author is talking about-for example, what variables are currently being considered, what is being held fixed, and what is being allowed to vary-when in fact none of this has ever been stated.

The Web version of this document contains a link to some illustrative quotations, with commentary where appropriate, which have been taken from recent project reports. (Of course we mark so many projects that occasional lapses such as these are easy to find.) In most cases you will find that the sentences look just fine-until you try to decode the meaning.

How should you attempt to improve the quality of your writing? It is important to understand that, like learning to drive, this is something you have to work at. No doubt an ability to write well comes more easily to some than to others, but everyone needs to put some effort into acquiring this skill, and anyone can improve if they do. Of course there are plenty of books on how to write well, but perhaps the most useful thing you can do is to *read widely* (books, articles in quality newspapers, etc-anything except *The Sun*) and learn to observe how others write.

4 Use of notation

In general you should use mathematical notation to identify and refer to variables. This is typically clearer and less ambiguous than constantly referring to ``the mean monthly temperatures" or whatever.

Even more important is the use of notation to express mathematical relationships between variables. For example, although you could write

> the ratio of the temperatures at any two pressure levels in any one month is the same as the ratio of the temperatures at the same two pressure levels in any other month,

it is both shorter and clearer to write

> T_{p1m}/T_{p2m} is independent of m for all p_1 and p_2.

The advantages of the latter, particularly if you wish to explore the further mathematical consequences of a relationship, are obvious.

5 Tables and figures

Tables and figures should be labelled (usually with consecutive numbers-Table 1, Table 2,..., Figure 1, Figure 2,...) and captioned, for example,

> Figure 1: Plot of temperature against pressure.

In the text you should refer to ``Figure 1'' rather than ``the first plot on page 4''. This is simpler, clearer, and avoids the problems which would otherwise arise when your wordprocessing package moves the figure to page 5! (The best wordprocessing packages automatically renumber tables, figures, and all references to them as necessary.)

Axes should of course be labelled, and an explanation given (usually in the caption) of the meanings of any symbols, line styles, etc.

6 Some things *not* to do

Computer packages
: Do *not* assume that your reader knows all about your favourite computer package. Your choice of ``statistics calculator'' should be as unobtrusive as possible. For example, you should write ``the parameters were fitted by least squares'' rather than ``the parameters were fitted using the S-PLUS function `lm`''.

Repetition
: Do *not* present numerous variations of the analysis (no doubt in the hope that one of them will be optimal and thus collect full marks). Repetition only leaves the reader confused. Decide on the best approach to the data (perhaps after considerable prior experimentation) and present that. (Of course additional analyses which provide additional insights are entirely in order.)

Similarly avoid the use of repeated figures-for example, a histogram followed by a quantile plot-to demonstrate the same conclusions.

``Sample'' analyses
: Do *not* draw conclusions based on ``sample'' analyses of subsets of the data which are apparently chosen arbitrarily. Instead approach all the data even-handedly. (Of course this may well lead you to focus extra attention on the most interesting features of the data.)

Significant figures
: Do *not* report results to an excessive number of significant figures-three (or at most four) is *usually* sufficient.

File translated from TEX by TTH, version 3.01.
On 7 Oct 2001, 21:14.