Statistical Guidelines for Natural Resources Projects

March 1998

The University of Reading
Statistical Services Centre

Biometrics Advisory and Support Service to DFID

Contents

1. Introduction

This is the first in a series of guidelines for staff involved in the development and presentation of research projects. The guidelines are intended to help researchers to identify their biometric or statistical needs. This introductory guide gives general information, while other guides examine individual topics in more detail.

Our basic premise is that research projects can often be enhanced by improvements in the statistical components of the work. Areas where an understanding of statistical ideas is important include the following:

- A clear definition of the objectives of the study and the way in which these objectives should determine the design of the research.
- The design of the research. In an experiment the design includes its location(s), the treatments, the size and layout of the plots and the measurements to be taken. In a survey, research design includes the sampling plan and the questionnaire.
- The entry, management and archiving of the data.
- The analysis of the data.
- The presentation of the results.

Without sufficient confidence in statistics, researchers plan designs that are often conservative, primarily to ensure a simple analysis. While this is sometimes appropriate, simple improvements can often result in more informative experiments, or surveys, for the same cost. Measurements may be made that are inappropriate for the objectives of the study, and are subsequently not analysed. Data entry can become a very time-consuming part of the study if it is not planned in advance. The analysis may be rushed, using inappropriate software, because of the pressure to produce results quickly. The study is then concluded, with the realisation that there is much more that can still be learned from the data. However, funding is at an end, and problems in the data management have made it difficult to allow easy access to the data, even for future researchers within the country where the study was made.

In this guide we consider these stages, from planning to presentation, in turn. Its aim is to encourage the researcher to think about the crucial aspects of planning, analysing and interpreting data. You may find methods referred to that are unfamiliar – this should certainly indicate that you might need advice. We conclude by describing further support that is possible by the involvement of a biometrician within the research team.

2. The Planning Phase

2.1 Specifying the Objectives

An initial step is to specify the areas where there are gaps in the existing knowledge and hence determine the types of research to be used. If there is insufficient knowledge of the constraints to adoption of a new technology, then a survey or an on-farm, participative experiment may be indicated. Lack of information, for example on critical processes affecting water use by proposed crops, might necessitate on-station, laboratory work or the use of a crop simulation model, or both. "Brain-storming" sessions among interested parties are often useful components of this initial process of identification of the areas and types of research that are needed.

In the protocol for the research study, gaps in current knowledge provide the basis for the background / justification section. A statement of the objectives usually follows this section. Sometimes there will be an overall objective, followed by a series of specific objectives. The objectives must be formulated with care, because they determine key features of the study design. For example, the treatments in an experiment follow directly from the objectives, as should the structure of the questionnaire in a survey.

2.2 Units of Observation

The units of observation are the individual items on which measurements are made. Examples include:

- Farmer
- Household
- Community
- Group of plants on an area of land (plot) or in a controlled environment
- Individual plant, single leaf or section of leaf
- Tissue culture dish
- Individual animal or group of animals (grazing flock, pen, hive)
- Fish pond
- Individual tree, sample plot or area of forest

Some studies involve more than one type of unit. For example, a survey may collect data on households and individual farmers; an agroforestry experiment may apply treatments to whole plots and make some of the measurements on individual trees.

2.3 Scope of the Study

The scope of the study includes the population from which the units of observation should be selected. The concept of a "recommendation domain" – the population for which the conclusions of the study are to be relevant – is crucial.

Another important aspect is the size of the investigation. It is essential, before embarking on an investigation, to have an estimate of the precision of the answers that will be obtained from the investigation. At the simplest level the precision could be measured by the standard error of a difference between mean values, or between two proportions. However, it could also include precision of estimates of model parameters (rates of growth, dependence on time or on chemical concentration) or population parameters (proportion of arable land used effectively, percentage increase in uptake of new technologies by farmers).

Unless the investigation is expected to be capable of providing answers to an acceptable degree of precision it should not be started.

2.4 Planning an Experiment

Key characteristics of an experiment are the choice of experimental treatments to satisfy the study objectives and of the units to which the treatments are to be applied. There should be some control of sources of variation between the units – this is usually achieved by blocking. A randomisation scheme is used to allocate the treatments to the experimental units. For the treatments, the questions include

- what treatment structure, if any, is to be used?
- can a factorial treatment structure be used to answer questions efficiently, and if so, how?
- how should levels of a quantitative factor be chosen?

Control treatments provide baselines: the comparison of other treatments with a baseline is often an objective. However, controls are simply treatments, and their presence and specification must be justified just like any other treatment. On controlling variation, the questions include

- what form of blocking should be used (blocks of more than eight units are often too big to be efficient)?
- what additional information about the experimental units (plots) should be recorded?

In general a good experimental design ensures that the effect of treatments can be separated from the "nuisance" effects of environments with maximum efficiency. In designing an experiment, factorial treatment structure should be regarded as the norm.

Also, designs using blocks with fewer units than there are treatments should be in common use.

In certain types of trial there are particular aspects of design which are important. For example:

- Crop or forestry variety trials – consider using α-designs.
- Animal pasture trials – these are likely to involve multiple levels of variation: groups, individuals, time sets of observations within animals.
- Laboratory experiments – special attention should be given to factorial structures for treatments.
- On-farm trials – ensure representative and randomly selected sites, enough treatments, enough overall replication; but there is no requirement for each site to have the same design structure.

2.5 Planning a Survey

The main elements of the design of a survey are a well-designed sample and a questionnaire (or other data collection procedure) which satisfies the study objectives. Crucial requirements of the sample design are representativeness and some element of randomness in the selection procedure. These usually imply a need for some structuring of the survey, often involving stratification, to ensure representation of the major divisions of the population. Deliberate (systematic) selection of samples can give, in general, the greatest potential accuracy of overall answers but has the disadvantage of giving no information on precision.

Some random element of selection is needed if we are to know how precise the answers are. In most practical situations clustering or multistage sampling will be the most cost-effective method of sampling. A balance of systematic and random elements in sampling strategy is usually necessary. In multistage sampling the largest or primary units often have to be selected purposively, but the ultimate sampling units are then selected at random within primary units.

For baseline studies the definition of sensible sampling areas, stratification scheme and the capacity for integrating data from varied sources are important.

In environmental sampling, where the spatial properties of variation are an important consideration, the spatial distribution of samples should provide information about variation at very small distances, at large distances and at one or two intermediate distances.

2.6 Other Types of Study

In observational studies, pilot studies and more empirical appraisal systems the general concepts of experiments and surveys are relevant although the particular detailed methods may not be applicable. Thus, representativeness and some element of random selection are desirable. It is also important to have some control, or at least recognition of potential sources of major variation.

Some studies also require access to routinely collected data, such as climatic records or aerial photographs. It is important to verify that these data are appropriate for the research task.

3. The Data

3.1 Types of Measurement

Measurements can be made in many different forms ranging from continuous measurements of physical characteristics (weight, length, cropped areas) through counts (insects, surviving trees), scores of disease intensity or quality of crops, to yes/no assessments (dead, germinated) or attitudes (like, prefer). The importance of including a particular measurement has to be assessed in the context of the objectives of the research.

There are some general points which apply to all investigations.

- For a given design, assessment in the form of continuous measurement will give greater precision than ordered scores, which will in turn give greater precision than yes/no responses (e.g. weights are more precise than low/medium/high weight classes).
- The form(s) of measurement selected must be capable of giving answers of acceptable accuracy to the questions asked.
- The relative precision of different, alternative designs for an investigation is not changed by the particular form of measurement.

3.2 Collecting the Data

Data collection forms will usually have to be prepared for recording the observations. In social surveys, the design and field testing of the questionnaire are critical components of the study plan. In experiments and observational studies, simple data collection forms often suffice.

3.3 Data Entry and Management

Part of the consideration of measurements is the question of how the data are to be managed. At some stage, the data will normally be held in one or more computer data files. It is important that the form of the data files and the software to be used for data entry, management and analysis are determined before data collection begins.

Data are sometimes collected directly into a portable computer. Where data collection sheets are used, the data entry should be done directly from these sheets. Copying into treatment order or "hand calculation" of plot values or values in kg/ha should not be permitted, prior to the data entry. All data should be entered: if measurements are important enough to be made, they are important enough to be computerised. Studies where "just the most important variables are entered first" inevitably result in a more difficult data entry process and the remaining variables are then rarely computerised.

Data entry should normally use a system that has facilities for validation. Double entry should be considered, because it is often less time consuming and less error-prone than other systems for data checking.

It is desirable to follow the basic principles of good database management. In some studies the management is a trivial stage, involving simple transformations of the data into the units for analysis. It can, however, present real challenges, particularly in multistage surveys or in animal, agroforestry or mixed cropping experiments. A general rule is to base the management on a series of programming commands, rather than using cut-and-paste methods. The latter often result in multiple copies of the data in different stages of the process, which make it very difficult to correct any errors that are later discovered when the data are analysed.

4. Analysis

In this section, more than previous ones, you may encounter unfamiliar statistical terms. These should emphasise that there are aspects of new statistical methodology from which you can benefit, with appropriate advice.

4.1 Initial analysis: Tabulation and Simple Graphs

For experiments, initial analyses usually include simple tabulation of the data in treatment order, with summary statistics, such as mean values. For surveys, simple tables are produced, often showing the results to each question in turn. These initial results are only partly for analysis, they are also a continuation of the data checking process.

It is important to distinguish between “exploratory graphics”, which are undertaken at this stage, and “presentation graphics”, mentioned in the next section. Exploratory graphics, such as scatterplots or boxplots, are to help the analyst in understanding the data, while presentation graphics are to help to present the important results to others.

With some surveys, most of the analysis may consist of the preparation of appropriate multiway tables, giving counts or percentages, or both. Caution must be exercised when presenting percentages, making clear both the overall number of survey respondents and the number who responded to the specific question, as well as indicating which is used as the denominator.

4.2 Analysing Sources of Variation

Particularly for experimental data, an important component of the analysis of measurements is often an analysis of variance (ANOVA), the purpose of which is to sort out the relative importance of different causes of variation. For simple design structures, the ANOVA simply calculates the sums of squares for blocks, treatments, etc. and then provides tables of means and standard errors, the pattern of interpretation being signposted by the relative sizes of mean squares in the ANOVA. For more complex design structures (incomplete block designs, multiple level information) the concept of ANOVA remains the same, providing relative variation attributable to different sources and treatment means adjusted for differences between blocks. In studies with less formal structure it may be appropriate to split up the variation between the various causes by regression or, when there are multiple levels of variation, by using a powerful new method, known as REML.

The particular form of measurement will not tend to alter this basic structure of the analysis of variation, although where non-continuous forms of measurement are used the use of generalised linear models (a family of models which includes loglinear models and logistic regression) will usually be appropriate. Such methods are particularly appropriate for binary (yes/no) data and for many data in the form of counts. These methods are useful for all types of study – experiments, surveys and observational studies.

4.3 Modelling Mean Response

There are two major forms of modelling which may occur separately or together. The first form, which has been used for a long time, is the modelling of the mean response (for example, the response of plots of a crop to different amounts of fertiliser, or the response of an animal's blood characteristic through time). The objective of such modelling is to summarise the pattern of results for different input levels, or times, in a

mathematical form which is consistent with the biological understanding of the underlying mechanisms. Frequently, the objectives also indicate the need to estimate particular comparisons or contrasts between treatments or groups. LSDs and other multiple comparison procedures should normally be avoided.

Modelling of mean response may also include multiple regression modelling of the dependence of the principal variable on other measured variables, though care should be taken to ensure that the experimental structure is properly reflected in the model. Note that, in modelling the mean response, the use of R^2 to summarise the success of the response model is usually not adequate. Large R^2 values need not reflect success in model fitting, which should be measured by the error mean square about the fitted model, relative to the expected random variation.

4.4 Modelling Variance

More recently modelling of the pattern of variation and correlation of sets of observations has led to improved estimation of the treatment comparisons or the modelling of mean response. Particular situations where modelling variation has been found to be beneficial are when variation occurs at different levels of units, for spatial interdependence of crop plots units or arrays of units in laboratories, or for temporal correlations of time sequences of observations on the same individual animals or plants.

Modelling of multilevel variation is beneficial when there is information about treatments from differences within blocks and between blocks, or when data from different trials with some common treatments are being analysed together allowing for variation between trials and within trials. The REML method, mentioned earlier, is relevant in all such cases. Multilevel modelling is also important for the correct analysis of survey data when the sample has a hierarchical structure.

Spatial analysis models have been found to increase information from field plot variety trials by between 20% and 80%. Essentially each plot is used to assess the information from adjacent plots.

When observations are made at several times for each animal in an observation trial, or for each plot in a crop growth study, it is crucial to recognise that variation between observations for the same animal is almost always much less than variation between animals. Analysis of such “repeated measurements” data must separate the between-animal variation from the within-animal variation in separate sections of the analysis. Several different approaches are available for the analysis of repeated measurement data, and they require thinking first about the general pattern of response through time, and analysing variation of that pattern.

5. Presentation of Results

The appropriate presentation of the results of the analysis of an experiment will usually be in the form of tables of mean values, or as a response equation. A graph of change of mean values with time or with different levels of quantitative input can be informative. Standard errors (and degrees of freedom) should generally accompany tables of means or response equations.

6. Biometric Support

Whenever a project involves the collection, analysis or interpretation of quantitative information it should be assumed that a biometrician or statistician may be able to help. This help is to make both the planning of the data collection, and the analysis, more efficient (in the sense of maximising the information per unit of resource).

The ideal situation is to include a biometrician as part of the research team. This is feasible if the research organisation includes a biometric support group. Ideally a named biometrician should be associated with each research project and should be involved from the earliest stage of planning in identifying the critical stages for biometric input. If no biometrician is available locally, then the leader of the research project should seek advice from a university, research institute or private consultancy or from DFID's biometric advisers at the Statistical Services Centre.

The ideal is to use a biometrician or statistician familiar with the particular area of research, who will therefore have the experience of the particular scientific concepts and practical problems. If the biometrician does not have that detailed experience, both scientist and biometrician will have to explain concepts to each other.

Sometimes local biometricians are available, but they lack the necessary experience to make a substantial contribution unaided. There is no reason to expect more from a biometrician who has recently graduated with a Masters degree, than from an agronomist or soil scientist with a similar qualification. It remains important to try to use local expertise; the DFID biometric advisers can backstop such staff where necessary.

– – – – ⚑ ⚑ ⚑ ⚑ – – – –

The Statistical Services Centre is attached to the Department of Applied Statistics at The University of Reading, UK, and undertakes training and consultancy work on a non-profit-making basis for clients outside the University.

These statistical guides were originally written as part of a contract with DFID to give guidance to research and support staff working on DFID Natural Resources projects.

The available titles are listed below.

- *Statistical Guidelines for Natural Resources Projects*
- *On-Farm Trials – Some Biometric Guidelines*
- *Data Management Guidelines for Experimental Projects*
- *Guidelines for Planning Effective Surveys*
- *Project Data Archiving – Lessons from a Case Study*
- *Informative Presentation of Tables, Graphs and Statistics*
- *Concepts Underlying the Design of Experiments*
- *One Animal per Farm?*
- *Disciplined Use of Spreadsheets for Data Entry*
- *The Role of a Database Package for Research Projects*
- *Excel for Statistics: Tips and Warnings*
- *The Statistical Background to ANOVA*
- *Moving on from MSTAT (to Genstat)*
- *Some Basic Ideas of Sampling*
- *Modern Methods of Analysis*
- *Confidence & Significance: Key Concepts of Inferential Statistics*
- *Modern Approaches to the Analysis of Experimental Data*
- *Approaches to the Analysis of Survey Data*
- *Mixed Models and Multilevel Data Structures in Agriculture*

The guides are available in both printed and computer-readable form. For copies or for further information about the SSC, please use the contact details given below.

Statistical Services Centre, The University of Reading
P.O. Box 240, Reading, RG6 6FN United Kingdom

tel: SSC Administration	+44 118 931 8025
fax:	+44 118 975 3169
e-mail:	statistics@reading.ac.uk
web:	http://www.reading.ac.uk/ssc/