A Guide to Good Research Planning in Biological Research

H M Dicks

Former Senior Lecturer, Statistics & Biometry, University of Natal

Pietermaritzburg, South Africa

Summary

Good quality statistical design and analysis of agricultural research is a prerequisite to meeting the future demands for food and improving the incomes and livelihoods of poor people.

The agricultural research focus in developing countries is changing and is no longer based just on the old commodity and disciplinary boundaries.

Integrated approaches to solving problems and understanding agricultural systems are increasing in importance at all stages of research, with farmers and scientists often becoming research partners.

The future teaching of biometrics and research methods needs to recognise this changing focus in order to ensure that future graduates/researchers are equipped with the appropriate tools to meet the challenge.

[Acknowledgements: ILRI 2006. Biometrics and Research Methods Teaching Resource Version 1. (Editor John Rowlands). ILRI (International Livestock Research Institute), Nairobi, Kenya]

1. Introduction

With the increasing multi-disciplinary approach to agricultural research, to which other disciplines such as environmental and social science are increasingly being added, the discipline of 'biometrics'; the statistical or quantitative study of biology, is becoming increasingly important. Indeed, one could say that biometrics is now an essential discipline within the makeup of a multi-disciplinary research team. 'Biostatistics' is an alternative word, commonly used in the medical field. 'Econometrics' is the statistical study of economics. 'Statistics' is a more general term that encompasses all three.

The discipline of biometrics (or biostatistics) contributes to the overall field of 'research methodology' that is applied in the design and analysis of research studies. Indeed the term 'research methods specialist' is now sometimes used instead of the word 'biometrician' because of the fear that the subject 'biometrics' can sometimes instil in people

A research investigator, whatever his/her discipline, needs to have some understanding of the principals of study design and methods of data analysis for getting reliable results from research. Some knowledge of research methods is, therefore, essential.

A research investigator can be any one of a number of people working in a range of development and research fields: scientist, biologist, agriculturalist, extension worker (or even in some cases the biometrician himself/herself!). In order to maintain a constant nomenclature in this Teaching Resource to cover all these categories we use the word 'researcher'.

Each member of a multi-disciplinary team needs to have some involvement from the beginning, whether it be at the project proposal writing stage or at the beginning of the research project itself once funding has been approved. This means that clearly defined plans with clearly stated milestones need to be established so that all members know what is expected of them. This is where 'research strategy' comes in.

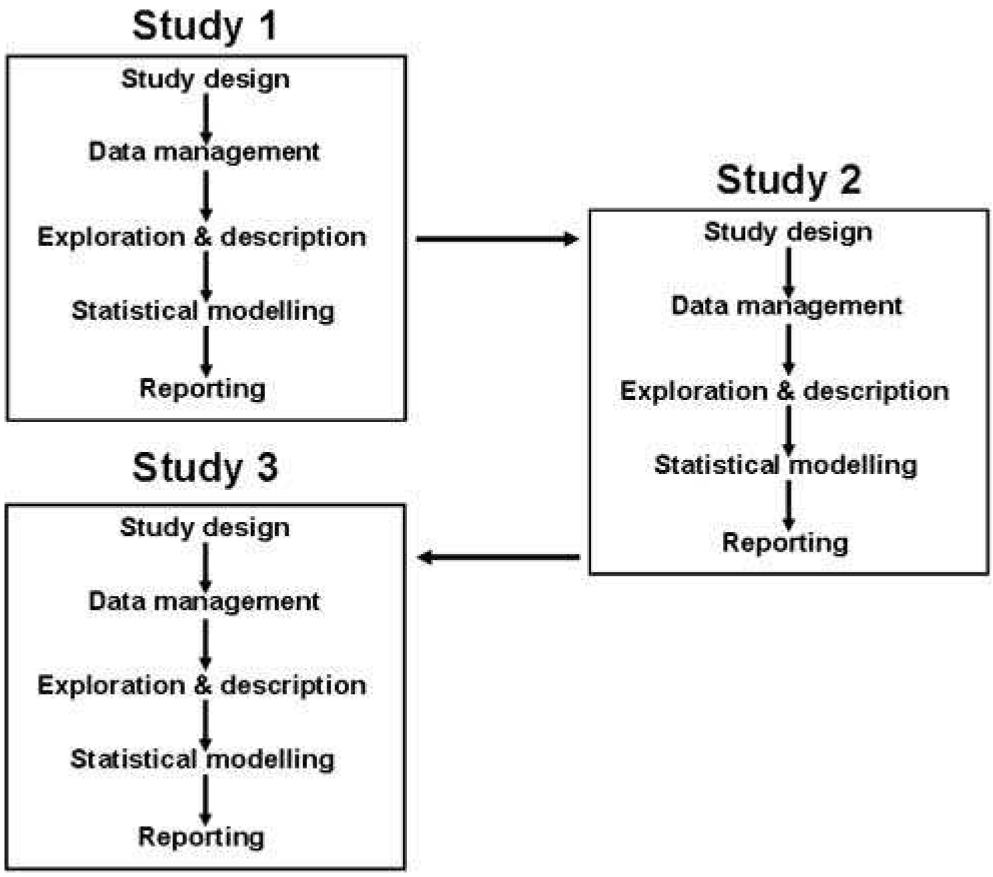
Every research project will encompasses a range of activities in the research process - namely from study design to data management, to data exploration to statistical modelling and, finally, to the reporting of results. These five stages each require important research method inputs all of which need to be considered at the outset when putting together a research strategy.

So what is a research strategy? It is the activity that needs to be undertaken to ensure that there are adequate resources available to complete the study in the time available, to make sure that the approach to the design of the study is the appropriate one to achieve the study's objectives, that suitable softwares are available to manage and analyse the data, that sensible sets of data are collected to ensure that analysis will allow the required information to be extracted, and so on.

A research strategy can evolve and is not necessarily cast in stone. Things can go wrong; for example, unexpected field conditions (drought or flood) may cause a study to fail. Thus, the original strategy may need to be thought through again and revised. Alternatively, execution of exploratory analysis may suggest a different approach to formal statistical analysis than had originally been envisaged. This will cause a modification to the original strategy that had been proposed for the statistical analysis of the data.

2. Research process

Studies are rarely done in isolation. One study builds upon another. Each study is composed of a set of different stages that can be broadly summarised into: **Study design**, **Data management**, data analysis (in the form of **Exploration & description** followed by some form of **Statistical modelling**) and finally **Reporting** of the results. Results obtained from one study, even before it is written up, will help to define the next Figure. The figure shows how plans can be initiated as soon as some preliminary results are known. On other occasions the researcher may prefer to wait until the results are written up and firm conclusions from the study are available before deciding what to do next.



Study design

The researcher's first task, once a research proposal is approved, is to plan the first study. Biometric inputs are essential to ensure that sample size, land availability, etc., is adequate to meet the study objectives, and that an appropriate study is planned that makes optimal use of resources.

It is also important at this stage to make sure that study can be done with the resources available, and to be able to define a time frame within which the study can be completed.

The process in planning a study often becomes an iterative one, involving an assessment of objectives and resources, and usually needs more time for discussion than is often appreciated. The biometrician, for example, may propose a scaling down of objectives, for example, a reduction in the number of treatments. Alternatively, the researcher may agree to a biometrician's recommendation to increase the size of the study beyond that originally considered.

It is important when designing a study to look both backwards, to consider what has been done before and forwards, to consider what might be investigated next. Study design is a major component of the overall research strategy and the biometrician's contribution to study design will be greatest when he/she understands the overall strategy.

A clear understanding of the conclusions that might be drawn at the end of the study is also important at this stage. The study needs to be relevant to the problem in hand and appropriate to the population for which its eventual impact is intended. Basic questions need to be asked.

Sooner or later study design may need to cater for some form of participation by farmers. Participation can vary from simple participation, to farmer management of a trial, to joint researcher/farmer trial design. The research strategy needs to decide when an element of farmer participation is required and what form it should take.

DATA MANAGEMENT

The importance of ensuring good study design as discussed in the previous section cannot be overemphasised. If the design of the study is inadequate and the data produced for analysis have little chance of meeting the desired objectives then the study will be wasted.

The discipline of biometrics is often viewed as having two components - study design and statistical analysis - but this an oversimplification. There are other components in the research process. The one that is often overlooked, but one that is as important as any other, is that of data management. This is the component that can take the most time and, if not attended to carefully, can lead to errors in data recording, computer entry, storage and management that can spoil the subsequent statistical analysis.

Data are often expensive to collect. There is no point, therefore, in putting a lot of effort into making sure that the methods used for obtaining the data are of high scientific analytical quality if the same stringent quality controls are not maintained during data recording and computer entry.

Procedures in data management begin at the start of the study design stage and cover the following range of activities:

- deciding how observational units are to be identified
- planning how data are to be collected and recorded
- budgeting and planning for how the data will be managed
- deciding what software to use
- planning how data are to be entered into the computer and stored
- devising schemes for data entry and checking
- undertaking data entry and validation
- calculating variables derived from input variables
- organising the data for analysis
- tabulating data for reports
- archiving the data once they have been analysed and reported.

It can be seen that these activities begin right at the beginning and finish when everything else is completed. Thus, the whole data management process, even that of archiving, needs to be planned at the start of the project. Indeed, unless proper attention is paid to this subject it is easy to make mistakes and reach wrong conclusions.

Some aspects of data management, such as data entry and verification, documentation and archiving, are often seen to be boring. But others, such as data system design, data handling and programming can be as interesting as other statistical topics covered under **Study design** and **Statistical modelling**.

1. How many data?

Data management is time consuming. It is therefore essential that plans are drawn up at the outset to set milestones for the completion of each task, and to ensure that the necessary resources are in place to meet these objectives.

Many projects are now funded to be completed within a given time-frame. Data entry can be particularly time-consuming and estimates need to be made on the number of data items that will be involved in computer entry and the length of time envisaged to do this. The actual time involved will invariably be more than estimated. Having designed a questionnaire or data recording sheet it is a good idea to prepare some dummy data to time how long it takes to enter the data into the computer. Then

- Enter the data to calculate how long it takes to enter a questionnaire or section of a data sheet.
- Calculate the hourly rate (making an hour equivalent to 50 minutes to account for interruptions due to phone calls, power cuts etc.)
- Calculate the total number of hours required to enter all the data.
- Multiply by 2 to allow for verification.
- Multiply by 2 again to allow for checking and correcting errors by researcher and for other unforeseen problems.

The time to be taken to do a job of work is often underestimated. As indicated in the last bullet above 'doubling the first estimate' is often a good rule of thumb.

Having estimated the time that is likely to be involved one can then calculate the length of the job from the total number of hours anticipated, and the number of people to be used in data entry and checking.

Practical exercises such as this will bring home to the research the time-consuming aspects of data management and the importance of deciding how many data should be collected in the first place, both in terms of the total sample size and the number of variables recorded per sample. Study design and data management are thus closely linked.

Decisions will also be required on whether additional computer programming is necessary, whether interim reports are required at strategic times during the course of a project and for whom they may be required. For example, haematological and parasitological data collected in a clinical trial may need to be summarised, not only to report on the results at the end of the experiment, but also to monitor the health of individual subjects as the experiment progresses. The suitability of alternative data storage systems may need to be considered to ensure simple but regular data retrieval.

It may be that the estimates suggest that the quantity of data to be collected may be too large to handle with existing resources. If so, consideration should be given to reducing the number of variables to be measured or reducing the overall size of the study.

Certainly, there is a tendency in survey work to collect more data than are warranted, and it is tempting to use the opportunity to collect as much information as possible. As discussed in **Study design** it is important when planning a study to ensure that only those data, which are relevant to the objectives of the survey and the hypotheses to be evaluated, are collected.

In summary, a researcher must ensure that as far as possible only data pertinent to the objectives of the study are collected so as to ensure that computer entry can be done within the required time-frame. Good filing systems, well designed data sheets and questionnaires, as discussed later, will help with the overall efficiency.

2. Observational unit identification

Each observational, sampling or experimental unit needs a unique identifier. In an experiment or an observational study it may often be wise to give two separate identifiers to the main units under investigation, such as a plot in a field or an animal, just in case one identifier is lost.

3. Data management software

One of the first decisions is to decide on the appropriate software to use. This will depend on the complexity of the study and the different types of observational units at different hierarchical levels. So far we are considering simple data structures, essentially at one layer. The different types of software that can be used for handling such data include:

- Statistical packages, e.g. GenStat, SAS, SPSS, Status

- Spreadsheet packages, e.g. Excel.
- Relational database management systems, e.g. Access, dBase.

Geographic information systems are also available for storing spatial data.

Some of the advantages and disadvantages of the three types of packages for data management are given in the following table.

Package	Advantages	Disadvantages
Statistical	Data management and data analysis can be done within the same package.	Usually unsuitable for multi-level data structures
	Can do most that a spreadsheet package can do.	Lacks security of a relational database management system.
	Will usually have programming capabilities.	Graphical facilities may not be as good as in specialised graphical software.
Spreadsheet	User friendly, well known and widely used	Unsuitable for multi-level data structures
	Good exploratory facilities	Lacks security of a relational database management system
		Statistical analysis capabilities limited to simple descriptive methods
Database management	Secure	Needs computer expertise in database development
	Can handle complex multi-level data structures	Graphical facilities may not be as good as in specialised graphical software.
	Allows screen design for data input	Statistical analysis capabilities limited to simple descriptive methods
	Will generally have standard facilities for reporting	

Each type of package has its own special use. Nevertheless, statistical, spreadsheet and database management packages have overlapping facilities for data management, and all can now 'talk' to each other. In other words a data file stored in one package can be exported (transferred) to another. The researcher needs to anticipate at the start of a project how data entry, management and analysis will proceed, and plan accordingly.

Excel is popular among researchers and is suitable for storing data from simple studies. But care is needed in handling spreadsheets as it is easy to make mistakes. Excel does not offer the same data security as a data base management system and so it is easy to lose and spoil data.

There is no reason nowadays why data entry, management and analysis cannot all be done in the one statistical package, especially for the more simple studies in which data can be stored in standard spreadsheets. The problems of data security remain but the advantage, as indicated in the above table, is that all the stages from data entry to analysis and interpretation can be done within the same package.

Data validation

1. At the time of data entry

It is helpful to detect unusual values during the data entry phase, especially for complex data sets. This may, for instance, be due to errors in writing down the values when recording the raw data.

The simplest approach is to set upper and lower values within which it would be biologically reasonable to expect the data to lie. This is not routinely possible in GenStat (except by writing data checking procedures once the data have been entered). Excel, on the other hand, does have a simple validation facility that can allow upper and lower limits to be defined for whole numbers, decimal numbers and dates, set limits to the lengths of text and calculate values for formulae to determine what is allowed.

Similar facilities to capture errors during data entry can be programmed into relational data base management systems.

As well as handling multi-level data structures, such as arise in surveys, data base management systems are also especially suited for handling data collected during longitudinal studies when different types of data are recorded. For instance, data on certain events that take place during the study may be required in addition to the routinely recorded measurements that occur at regular intervals. For cattle, for example, this could include information on dates of calving, dates of weaning, dates of treatments, dates of disposals and so on.

A database management system is able to manage and assemble data in several files simultaneously, and so data validation checks, such as checking that intervals between successive calvings fall within reasonable limits or that a date of disposal does not occur while measurements are still being recorded on the live animal, can be readily programmed.

When measurements are made in random order over time it is important to ensure that each plot or animal code that is entered from the recording sheet (along with the measurements for the monitored variables) matches one of the of the permitted plot or animal codes previously stored in a master file within the data system.

A livestock data base management system LIMMS, written by the International Livestock Centre for Africa (one of the two institutes merged to form ILRI) provides a comprehensive system for handling livestock data; it is written in dBase.

2. During the preliminary exploration phase

Many recording errors can be captured during data entry. However, some may escape. Once data entry is complete a variety of summary tables, box plots, histograms or scatter diagrams, as discussed in the **Exploration & description** guide, can be produced to identify potentially extreme or unusual values that fall within the preset limits, but nevertheless fall outside the distribution shown by the majority of values for the other observational units. These values should first be checked against the original raw data to see whether there has been a mistake in data entry.

If the value is the value written on the recording sheet the researcher will then need to decide what to do. The decision will be based on biological considerations. If a data value is changed in the database then a note should be made to that effect in a recording book or in a separate spreadsheet.

Statistical and spreadsheet packages contain good exploratory techniques and it is sometimes preferable to transfer data files from a database management system to one of these packages to commence the analytical phase.

Excel has procedure known as the 'pivot table' facility that allows data to be tabulated in different ways. This is an attractive and easy-to-use facility. As well as being able to calculate means and frequency totals, the tables can be used to highlight unusual values. For example, by creating tables of frequency counts checks can be made of potential errors in coding of classification variables.

A count other than 1 in a cell in a two-way pivot table by block x treatment for a randomised block could signify that the code for a particular plot had either been omitted or entered incorrectly. This can similarly be done using statistical, spreadsheet or database management packages, but Excel has the additional feature whereby, on clicking a cell of particular interest, the observational units that contributed to the contents of the cell are displayed.

A log book is recommended to keep a record of all data handling events, including the dates when data were first entered into the computer and when they were verified. If data values are altered from those recorded on the data sheet then these should also be noted. In this way the researcher, at any subsequent time during statistical analysis or writing up, can look to see what changes may have been made to the data. He/she will then be able to see what influence any alterations of the data may have had on the results. The need for keeping such logs is often overlooked. Although tedious it is an essential ingredient of good data management and should be budgeted for at the planning stage.

3. Missing values

Data may sometimes be missing. An animal may be ill during routine herd monitoring and kept behind at the homestead, or an animal may die. One plant may fail to grow in a glasshouse experiment or a plot in an experiment may be destroyed. There may be a fault in analytical equipment or a sample is mislaid.

Decisions need to be made on how to record such missing observations. One solution is to leave the data field blank. However, it should be remembered that a 'blank' may be expected to be different from the value zero and that a computer program may not distinguish between them. Packages sometimes use a '*' to signify a missing value. Alternatively, an unusual value, e.g. -9999, can be stored. Provided a consistent method of definition is followed during data entry, it will always be possible to recode missing data later to suit a particular software package. The method used for identifying a missing value should be documented.

GenStat has a useful facility whereby cells can temporally be defined as missing, yet retaining the original value so that it can be recalled later. This is particularly useful in analysis of variance for balanced designs where deleting a record can cause an experiment to become unbalanced, whereas temporally specifying an observation to be missing allows the record to be retained and the balance maintained.

GenStat often identifies potential outliers or observations exerting particular influence on the results of analysis. By temporally allowing these values to be missing analyses can be rerun to see what the results look like without them.

10. Data manipulation

Data management does not necessarily stop when all the data have been entered and checked, guide patterns in the data often need to be explored. This can be through summary tables or scatter or histogram plots. Summary tables of totals or means may require to be prepared for subsequent analysis.

The different types of data manipulation facilities that are likely to be needed are listed below.

1. Calculating new variables

As has already been mentioned calculations needed to derive variables from the raw data are best done by computer. This saves errors in calculation. GenStat also has data manipulation features that allow the formation of new groupings from existing values, such as ages of experimental subjects. GenStat distinguishes variables that are used to classify other variables and refers to them as 'factors'. Sometimes variables need to be used in either form, as a factor or variable (referred to by GenStat as 'variate'), and this is allowed.

GenStat also has the useful facility whereby factor levels that occur within definite patterns (e.g. treatment order within blocks) can be entered using a shortcut method that describes the pattern.

Most of the calculations needed to derive new variables can be done within the dialog box facilities of GenStat. In some cases a series of programming steps are necessary. The flexibility of the programming capabilities of GenStat are demonstrated in **Case Study 2** in which lactation yields to a fixed length are derived from 2-weekly offtake recordings and **Case Study 6**, in which a statistical analysis for one variable is extended to several variables.

Transformation of variables to values that follow more closely normal distributions for analysis are also easily done.

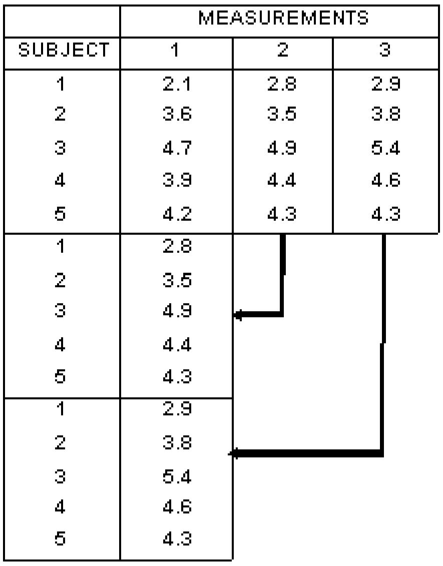
2. Residuals

Residuals from a statistical model may be required to investigate their distribution. This may be to validate that their distribution is appropriate for statistical analysis. Alternatively residuals may be required for subsequent data analysis with the effects of certain model parameters removed. GenStat allows residuals to be stored in the spreadsheet alongside the original data.

3 Stacking and unstacking

In studies involving repeated measurements (for example) stacking and unstacking data are often required. Stacking allows blocks of variables into single columns, often to facilitate statistical analysis (Figure. 2).

The opposite, unstacking, allows data to be displayed in 'matrix' form with columns representing the repeated measurements for a given variable and rows containing the observational units. This method of data presentation provides a useful structure for monitoring purposes. Averages for each column can be calculated over rows (observational units) and the mean values used for studying changes over time.



	MEASUREMENTS		
SUBJECT	1	2	3
1	2.1	2.8	2.9
2	3.6	3.5	3.8
3	4.7	4.9	5.4
4	3.9	4.4	4.6
5	4.2	4.3	4.3
1	2.8		
2	3.5		
3	4.9		
4	4.4		
5	4.3		
1	2.9		
2	3.8		
3	5.4		
4	4.6		
5	4.3		

Fig. 2 Stacking of data

6. Transposing data

GenStat can also convert values stored in standard vector form into matrices, and by doing so use the matrix transpose function to convert rows into columns and columns into rows. This might be useful when one wishes to calculate the means (or some other function) of a group of observational units. Statistical programs are generally set up to apply calculations to columns rather than rows. Thus, by transposing observational units from rows to columns it becomes easier to apply the desired calculations.

7. Restricting/filtering

Variables often need to be restricted or filtered for statistical analysis purposes in order to include only a subset of the data set defined by certain values of a factor, say. For example, data were collected in the Swaziland livestock survey from both commercial and smallholder farms. GenStat has simple facilities to do this. Excel has similar facilities but it is preferable to perform such functions within the package that is being used for the statistical analysis.

8. Appending/merging

Facilities will also be required for appending data from one spreadsheet to the bottom of another or for merging data from two spreadsheets side by side into one. When spreadsheets are merged in GenStat columns can be specified for matching purposes. Usually this would be the identification code for observational units. Missing matches appear at the foot of the merged spreadsheet.

Most of the above data handling facilities should be available in most statistical packages. Students may be familiar with similar facilities in Excel. As already stated, however, restriction of the teaching of data management and data handling aspects during the early part of the course can have advantages. The student will gain greater confidence in the use of the package and have the opportunity to learn what is available without being distracted by other available software.

11. Multiple-level data structures

1. Understanding data structures and their linkages

Individual spreadsheets become more difficult to manage when a study involves the collection of data from more than one layer (or level). The use of a comprehensive, relational database management system, such as Access or dBase, will, as already indicated, become necessary. Such data management systems are described as relational database management systems because they have built-in structures that allow links to be made to relate one data set with another. The advantage of such a system is that the full details of the study design do not need to be held at each layer, only at the top.

Although, this guide has so far promoted the concept of teaching data management using the same statistical package that is used for the analysis, such programs are inadequate for the handling multiple layered structures. In practice this does not matter because transfer of data between packages is fairly simple.

A diagrammatical illustration of the form that a multi-level data structure can take is given in Figure.3. It is assumed that data are collected at three layers: region, household, then household member. Each record at each layer has a unique identifier (indicated in parentheses in the figure). These identifiers are often described as 'key' or 'index' variables or fields.

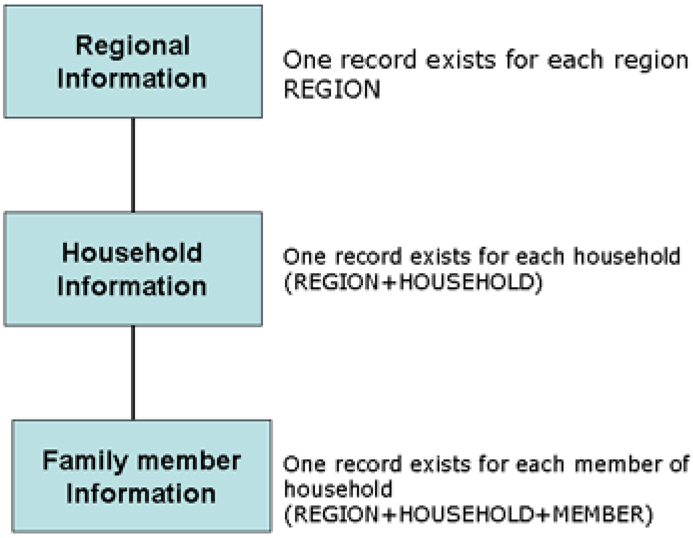

Regional Information
One record exists for each region
REGION
Household Information
One record exists for each household
(REGION+HOUSEHOLD)
Family member Information
One record exists for each member of household
(REGION+HOUSEHOLD+MEMBER)

Figure.3 An illustration of the linking of multiple layers of data.

Data collected at the regional level might include a description of the agro-ecological zone, the general type of farming system applied in the region and average meteorological information such as average rainfall, temperature etc.

Data collected at each household might include the grid reference of the household, information on the number of members belonging to the household, the age and gender of the head of the household, acreage, numbers of different species of livestock, numbers of fields etc.

Within each household data may be collected on individual members of the household (such as position in household, sex, household activities engaged in, etc.) As more than one livestock species will be raised within a household, a further data layer may be required to describe which species are covered by the activities engaged in by each member of the household.

One can thus imagine a further rectangle containing the words 'Species information' under the 'Family member' rectangle.

Provided unique and simple codes for the key variables are carried down to each of the lower layers the study structure definition can be picked up at any time. A well-programmed database management system will take into account how data are to be entered and, where multiple layers of data are being entered at the same time, recall the coded key variable levels used at the higher layer to save re-entry at the next.

11.2 Repeated measurements

Sometimes measurements are taken repeatedly in time from the same sampling units. In this case two layers, each with its own set of observational units, namely experimental subject and time when measurements were taken, can be defined. The first contains the study design with definitions of the experimental subjects and the second a set of files with details of the observations made on each subject at each sampling time. Provided that appropriate linkages are set up the only additional design variable that needs to be entered at the second layer is a code that identifies the sampling unit and the date when the observation was made.

Repeated measurement designs do not need to be held in a data management system when data are collected from a balanced controlled experiment. A well-documented series of Excel spreadsheets can be designed for such a purpose with each spreadsheet holding the data for each time point.

12. Screen display

When designing the format for a questionnaire it is also important to visualise how data entry will be done. There are various facilities that can be used in designing a data input computer screen that make data entry easy. Thus, an answer to the question: age of household <20, 21-25, 26-30, 31-40, 41-50, >50 years, can simply be entered by selecting the appropriate age range on the screen. This saves typing in the appropriate value.

It is advantageous, therefore, especially for entry of data collected from survey questionnaires, or indeed for general data input from long-term projects, to design computer screens for data entry. When a questionnaire is being designed thought needs to be given as to how well the layout lends itself to computer screen formatting.

When the format of the computer screen matches precisely the format of the questionnaire it becomes easy for a technician to enter the data. *Fig.4 Example of computer data entry screen depicting different forms of data entry*.

Different softwares are available to help to design screen layouts for data entry from questionnaires. Computer screen design is perhaps not a necessary component for a data management course in an applied biometrics course but the student should be encouraged to investigate what different software packages can do.

13. Data retrieval

There is no point storing data in a database if they cannot be easily retrieved. The researcher needs to consider ahead of time the information that he/she needs to see. As already described, the Access software provides a facility known a as a 'query' that allows simple summaries of data stored in the system to be viewed, reported or summarised More complicated reports may require some programming.

Projects that continue over time will require some form of interim monitoring and the types of reports required should be anticipated in advance. Data handling can be a major preoccupation and it can be irritating to the researchers if reports cannot be provided as soon as new data have been entered.

Data analysis will also be required, not only at the end of the project, but also at strategic stages of the project. This can be facilitated by anticipating ahead of time the data files that will need to be retrieved and how they are to be analysed. This needs to be planned early during the data collection exercise. It may be useful to do dummy runs with partial data, with the necessary transfers of data between different software packages, to ensure that the output meets the required formats.

When analysing data it is preferable to extract from the original file only those data to be used for the analysis and to store them in a temporary file. If the data analysis picks up further possible data errors and it is decided to change raw data items then it is important that these should be changed in the original file (or a copy of it) and the data extracted again before continuing with the analysis. Thus, it is better if a computer routine (a query in Access) is written (or at least documented) for any transfer of information from one file to another file so that the process can be repeated, if necessary.

15. Backing up and audit trails

Backing up and setting up audit trails can be a drag, but if a procedure for this is set up in advance it can become less of a chore. Data should be regularly backed up (or copied) onto a diskette, CD or other device. This is particularly important for long-term studies where the database is continually being updated. A suitable system of file naming should be organised so that it is clear when each back-up is made. As soon as an experiment has been completed, and all the data have been stored in a computer, a master copy should be made and again stored on diskette or CD in the event that the data kept in the computer and being used for data analysis are spoilt.

An audit trail is a type of log book which contains a complete record of changes that have been made in the data. Just as a researcher is required to keep a note book detailing the various steps taken during his/her experiments so should notes be kept of any decisions taken during data processing. This starts with data entry and continues right through to data analysis. If changes are made to data then it should be possible to repeat the whole process from any point in the trail.

16. Documentation and archiving

Documentation and archiving can be a tedious component of data management but they are nevertheless very important. Data are expensive to collect and they may have use beyond the particular objectives of the study for which they were originally collected. A researcher therefore has a duty to ensure that his work is satisfactorily documented and archived.

Because these aspects of data management are often considered to be tedious they are often not done. However, careful planning at the start of a study can save many headaches later. When the researcher appreciates the importance of documentation and archiving at the outset then much of the work can be done during the course of the study leaving a minimum of work to be done at the end.

When a spreadsheet or a database file is being designed thought needs to be given to the form of documentation that is needed. A well written study protocol will contain much of the information that needs to be archived.

Data documentation files are included with each case study. The student needs to familiarise himself/herself with them to see how data should be documented. **Case Study 3** gives a particular example showing how documentation in simple cases can be contained within the spreadsheet itself by using the top rows to give information on the variables in the columns below.

Once the data analysis has been completed and reports written, the data must be archived. This is just as important as reporting the results. Data do not belong to the individual researcher - they belong foremost to his/her institution. The institution may give permission for the researcher to access the data in future should he leave the institution, but he/she has a responsibility to ensure that a fully documented data file is lodged behind with the institution. Another researcher may wish to study the data to see what has been done, even to reanalyse them. He/she may wish to put similar sets of data together and do an overview.

The archive should include the data sets themselves, a documentation file that describes the information contained in the data sets, a copy of the original study protocol, any maps or pictures made during the study, copies of the final statistical analyses, reports or journal references, and the log book that describes what happened during the course of the experiment, during data entry and during data analysis. Hard copies of data sheets, study protocols, relevant computer outputs and reports should also be retained in the archive together with electronic versions.

A researcher is trained to keep detailed laboratory books - the same rigour needs to be applied to data recording, storage and analysis. Opportunities should be given to students to practise archiving materials from an experiment. It is better that they are taught to appreciate the importance of documentation and archiving while they are being trained, rather than expect them to pick up the knowledge once they are working.

17. Other types of data

Spatial data can be kept in a geographical information system - a specialist form of software for containing in the form of values contained with grid coordinates of a map. Different measurements are stored at different layers so that any can be superimposed one above each other. It is important for researchers to know the principles of such systems, to understand how data are stored and retrieved so that data can be incorporated, where appropriate, within their own studies.

Some types of information are not in a form that can be stored as data in a data file. Maps and pictures, for example, fall into this category. Modern computer scanning facilities, however, can be used to store images and these can be suitably annotated for storage on a CD. Short videos that have been taken during a study can also be stored in this way.

In summary, treat data in the same way that you might treat money.

EXPLORATION & DESCRIPTION

1. Importance of data exploration

Data exploration is the first process in the analytical treatment of data. Insufficient attention is often given to preliminary investigations of data, and researchers often jump straight into the formal statistical analysis phase of analysis of regression, analysis of variance etc. without making sure that data have been entered correctly.

Furthermore, following a series of preliminary procedures the researcher should be able to identify definite patterns in the data, gain insight into the variability contained within the data, detect any strange observations that need following up and decide how to proceed with formal analysis.

When a researcher and biometrician are working together it is often useful for the researcher to undertake some preliminary analyses first which will help him/her in his discussions with the biometrician in describing the patterns which formal data analysis need to take into account.

Sometimes application of suitable exploratory and descriptive techniques is all that is needed for the complete analysis.

This point is often neglected in statistical courses and students are left with the idea that all statistical analysis must end with a statistical test. When a result is obvious, for example when 18/20 untreated plants are infected with a fungus but none of the 20 treated plants are infected, one should not need to provide a P-value to convince a reader that there is difference.

Statistical analysis should be viewed as just one of the tools that the researcher uses to support his/her conclusions. When it becomes ridiculous to carry out a t-test, for example, when the results are so clear cut, one should not need to do so. A simple description of the results, e.g. 0% of treated plants infected versus 90% of untreated plants infected should be sufficient.

The more complicated the data set the more interesting and necessary the exploratory phase becomes. With some expertise in data management the researcher is able to highlight the important patterns within the data and list the types of statistical models and their forms that need to be fitted at the next stage.

Just as for Study Design and for the modelling and inferential ideas that appear in the **Statistical modelling** there are some principles that direct the exploratory and descriptive analysis phase. These are developed below.

2. Aims of exploratory analysis

The aims in undertaking explorative and descriptive analysis are covered by the following activities:

A. Investigate the types of patterns inherent in the data.
B. Decide on appropriate ways to answer the null hypotheses set out at the initiation of the study.
C. Assist in the development of other hypotheses to be evaluated.
D. Detect unusual values in the data and decide what to do about them.
E. Study data distributions and decide whether transformations of data are required.
F. Study frequency distributions for classification variables and decide whether amalgamation of classification levels is needed.

G. Understand the data and their variability better.
H. Decide on which variables to analyse.
I. Provide preliminary initial summaries, arranged in tables and graphs, that provide some of the answers required to meet the objectives of the study.
J. Develop a provisional, draft report.
K. Lay out a plan and a series of objectives for the statistical modelling phase.

This helps to rationalise the further work to be done in the statistical modelling phase. It is possible, especially in a simple study, that the only revisions that may need to be done to the final report, after all data analysis is complete, will be to fill in the contents of the tables and insert appropriate significance levels in the text.

3. Exploratory methods

Methods for exploring explore and displaying data are best illustrated using examples from the case studies. The examples in the case studies provide a suitable teaching aid.

3.1 Means and ranges

A useful place to start when describing data is to obtain lists of overall means, standard deviations, ranges and counts. This gives a useful summary, provides the first step in defining patterns within the data and is a useful stepping point for further investigation. Such summaries can be used to verify that the data set is complete and that there are no unexpected missing values.

3.2 One and two-way tables of means

As well as providing an overall descriptive table of means and distributions it is sometimes also useful to produce one- or two-way tables of means for the key response variables. The classification variables or factors for the tables will be those associated with the primary objectives for the study. These tables will provide the first insight into the influence of these factors on the response variables of interest. Such tables can also be produced to provide maximum or minimum values or show how the standard deviation varies across different classification levels.

3.3 Frequency tables

Once a good preliminary idea has been obtained of the distributions in mean values, one might then have a look at a series of frequency tables. Once again these help to identify any missing data. They also provide further understanding of how the data are distributed.

The numbers of observations represented by different parameters in a statistical model are often required for reporting purposes. Frequency counts are not always provided by the software being used to carry out a statistical analysis for unbalanced data. These counts may need to be calculated separately; if so, it is useful to keep the frequency tables obtained during exploratory analysis to be referred to later.

The Excel Pivot Table facility is an excellent device for developing frequency tables, and by clicking a particular cell the records that have contributed to it can be listed. This is a good way of detecting individual outliers.

3.4 Histograms

A one-way frequency table can be presented pictorially in the form of a histogram. A histogram helps to visualise the shape of the distribution.

If the histogram is more or less symmetrical, with the bulk of the data gathered near the centre and the proportions of data on each side roughly balancing each other, then we can assume the data to be normally distributed. But if the data are bunched up to one side then we can say that the distribution is skewed. In this case some special action may be required at the formal statistical analysis stage.

The following histogram shows how the numbers of cattle owned by homesteads in a livestock breed survey in Swaziland (**Case Study 11**) are skewed with the majority of homesteads having few cattle and a few homesteads several cattle.

3.5 Scatter plots

Associations between two (or more) variables can be studied pictorially by means of an (x, y) scatter diagram (or scatter plot) with or without a straight line included. Scatter plots can also be used to indicate outliers and potential data errors. Excel has a useful feature for identifying outliers. By clicking a point in a scatter diagram the original record to which the point belongs can be seen.

Individual data points in a scatter plot can be numbered according to some factor or other grouping. This provides the means for assessing whether relationships are similar or parallel for different groups.

e researcher to investigate the distributions of individual points for each of the levels.

3.6 Box plots

Box plots provide another useful way of summarising data. They display in diagrammatical form the range, mean, median, upper and lower quartiles (containing between them three quarters of the data). They also identify individual outlying points that lie beyond the range covered by most of the data. Depending on the distance of outlying points from the bulk of the data, the researcher can decide whether to investigate them further, or whether to discard or to retain them.

Box plots can also be used to assess how data are distributed and whether the distributions appear to be normal or not. By producing box plots of different groups of data side by side, an understanding can also be obtained of the ways observations are distributed not only within and but also across groups.

3.7 Bar charts and pie charts

A bar chart is a suitable way of expressing frequencies, mean values or proportions pictorially. These values appear as series of columns (or bars) above the horizontal axis with the height of the bar indicating the size of the value.

A pie chart, as the name suggests, is a circle in the shape of a pie, cut into sections to represent the proportions of an attribute assigned to different categories.

3.8 Trend graphs

Rather than just plotting the individual data points it is sometimes useful to join them up. This helps to describe trends especially when data are collected over time. Plotting the data in this way may help to decide how to summarise data for further analysis.

.9 Survival curves

These are used to examine survival rates calculated for data based on times to a specified event (e.g. death or disease). They provide a preliminary guide in the development of models to assess survival rates. Such curves, known as Kaplan-Meier curves, show how the occurrences of events (e.g. deaths) are distributed over time. The natures of the patterns in survival rate can influence the type of survival model fitted. Such curves can also be used to identify covariates that can be included in the subsequent model.

4. Outliers

The various methods described above may identify outliers, i.e. observations (or group of observations) that do not conform to the general pattern determined by the other data points. There are different forms of outliers. Some can justifiably be excluded and others not. If, having checked the data sheets, and no obvious reasons for an unusual data values indicated, then a decision will need to be made on whether to exclude an outlier or not.

If a data value is biologically nonsensical then it should be omitted. But in other cases the decision may be difficult. Just because a point does not conform to the pattern dictated by the other points does not mean that it is wrong. Thus, to remove a data item, just because it spoils the pattern expressed by the others and is likely to reduce the chance of a null hypothesis being rejected, is ill-advised. Identification of data variation can be as informative as the data mean.

Having decided to delete an outlier from a data set the researcher must record the fact in a note book and also state why certain observations were omitted when reporting the results of data analysis. Such a log book can also be used for writing down the various steps undertaken during the exploratory phase and for reporting the general observations made from each tabular or graphical presentation. This is essential when dealing with a complicated data set because not all the outputs will necessarily be kept. Without such care the researcher may find himself repeating earlier work.

5. Zeros

Zeros can occur in data for a variety of reasons. They could represent missing values. For example, a yield from a particular plot in a crop experiment might be missing (perhaps recording of this particular plot was omitted by mistake or the measurement failed to be copied to the recording sheet). Or maybe an animal is removed from an experiment or a farmer does not bring his/her animal to be weighed.

It is important to distinguish between a zero that represents a genuine missing value and a zero that is a real value. It is also important to distinguish between a zero and a blank'. Sometimes a value such as -9999. that is unlikely to occur in the data is used to indicate a missing value. GenStat recognises and uses a *', for example, to indicate a missing value.

Sometimes, especially when dealing with cross-classified categorical data, one may find that cells are empty due to some structural nature of the design. Again it is important not to confuse a missing cell with a zero.

Sometimes the numbers of zeros may exceed what was expected. For example 20% of farmers may have a crop failure. It is not sensible to hide this feature and might indicate that subsequent data analysis should be split into two: a study of incidence of crop failure and a study of the yields of farmers for whom crops did not fail.

6. Measures of variation

A number of the outputs produced during the exploratory phase will be useful in providing basic summary descriptive statistics for the final report. Such descriptions will almost certainly include some measure of variation, typically a variance, standard deviation and/or standard error. An understanding of the calculation of the variance expressed by a number of data points is important in understanding how to express the variability displayed by the data. This topic will, of course, feature in any basic course in statistics. Nevertheless, it is an important feature to understand when presenting descriptive statistics, and those who may wish to refresh their knowledge in measures of variation may find the following description helpful.

6.4 Coefficient of variation

The standard deviation/mean x 100 is known as the coefficient of variation (C.V.). This is a measure of relative variation.

The coefficient of variation of variables measured in controlled studies designed within a laboratory or on station will be found to be generally in the range of 10 and 15% or thereabouts; the more controlled a study the smaller will be the coefficient of variation. This rule does not apply, however, for all variables, especially those such as growth rate, for which individual variation can be large relative to the mean.

When planning (or analysing) an experiment, the researcher should have an idea about the coefficient of variation to expect for a particular variable within the environment that the research will be done. Coefficients of variation will be higher in studies in the field, especially when some form of farmer participation is involved, than on station. The expected magnitude of the coefficient of variation has a bearing on sample size.

8. Transformations

When data are not distributed normally, i.e. in the form of a bell-shaped' curve, it is sometimes possible to transform them and to analyse the transformed values. Certain extreme observations may be accommodated within the main distribution of a transformed variable. Three transformations are commonly applied:

$$\log (y + k)$$

$$\sqrt{y}$$

$$\sin^{-1}\left[\sqrt{y}\right]$$

The first two transformations are applied when the data are skewed to the right. The logarithmic transformation tends to be applied when the variances of different groups of observations are proportional to their means (i.e. $s^2 = k\bar{y}$ for some constant k) and the square root transformation when their standard deviations are proportional to their means (i.e. $s = k\bar{y}$). The square root function transforms data from a Poisson distribution into data that follow a normal distribution - this is useful when the data are counts (of insects, for example). The arcsine function is appropriate for proportions which tend to belong to a binomial distribution. The distribution of the data in this case is not skewed but the shape tends to be narrower than that of the normal distribution.

The section under **Exploratory methods** illustrates the distribution of a variable before and after a logarithmic transformation has been applied. Although still showing some deviation from a perfect normal distribution the transformed distribution is clearly better.

Analysis of variance (see **Statistical modelling**) is a robust tool. It can handle slight departures from normality with little impact on the overall conclusions. Thus, it is important during the exploratory phase of data analysis to decide on how essential a transformation may be. Transformation of data can sometimes result in difficulties in presenting results in a way that they are easily interpretable by the reader. On occasion it might be worth analysing both the untransformed and transformed data to see the consequences of transformation on the results.

↑

9. Residuals

One of the purposes of data exploration is to investigate the manner in which data are distributed, not only to look at the overall patterns induced by a study design (for example, differences between means) but also to examine residual distributions. Various procedures have been described (e.g. box plots) that can separate variations across and within different categories. Sometimes, however, when individual data values are influenced by a number of factors it becomes difficult to separate the different types of variation through these simple exploratory tools. One then needs to resort to fitting a statistical model (**Statistical modelling**) and investigate the distributions of residuals there.

Case Study 1, for example, illustrates an analysis of residuals carried out by GenStat after fitting a regression model. Two comments are made on the pattern of the residuals. Firstly, two observations were identified with large residuals or deviations from their fitted values. Secondly, the assumption that the response variable had a constant variance across all data points may not have been tenable.

Case Study 2 provides another example and shows how different residual patterns can be examined after parameters in the model have been fitted. The output from GenStat is reproduced here. The distribution of the residuals is expressed in different ways: histogram, scatter plot against fitted values and full and half normal plots. When residuals conform precisely to a normal distribution the curves displayed in the latter two graphs will fall on a straight line.

10. Data description

During the process of data exploration one should think ahead of the best ways that results can be described in the final report. This can be in any form of the tabular or graphical methods that have already been presented, for example means, standard deviations and ranges, tables of frequency counts, histograms, bar charts and so on. **Case Study 11** uses, for instance, two-way summary tables, pie-charts and histograms to present results of the livestock breed survey in Swaziland. Indeed much of the data analysis for this case study did not go much further than this.

Methods of presenting results are discussed under **Reporting**. Graphs are often most suitably represented using the raw data and without further data analysis. Tables, however, will generally contain least squares means adjusted for other factors or covariates during the **Statistical modelling** phase. Standard errors will also need to be added once analyses of variance have been completed.

Nevertheless, tables of raw means can be suitably prepared before statistical modelling commences to be replaced by adjusted means later. There is also no reason why the bones of a draft report cannot be put together at this stage. This helps to streamline the process for further data analysis.

Longitudinal studies will require preliminary analyses and reports at strategic times during their execution. These may be simple summaries of means and standard errors to assess how

↑

12. Meeting the objectives

One of the primary aims of the exploratory phase is to define precisely any further formal analyses that need to be done. By doing this well it should be possible to complete the formal data analysis phase quickly and efficiently. Of course, it may be that no further analysis is required. For example, principal component or cluster analysis may be all that is necessary for the analysis of a particular data set.

A second aim is to identify those variables that need to be included in the statistical models to be fitted, and to decide in what form the variables should be presented. At the same time the researcher needs to ensure that the model fulfils the objectives and null hypotheses that were defined when the study was planned.

There may be additional covariates that need to be included in the model to account for some of the residual variation. Decisions also need to be made on how complex a model should be. This will depend on the sample size and the degree of uncontrollable variation evident within the data.

A set of objectives for the statistical modelling should be prepared before moving onto the modelling stage. These will include

- suitable models to test the null hypotheses defined at the initiation of the study
- additional models to evaluate additional hypotheses indicated during the exploratory phase
- procedures to be used for evaluations of the appropriateness of the models (e.g. distributions of residuals)

As already indicated, a list of possible tables and graphs that might be appropriate to include in the final report will have been usefully prepared by this time too.

As has already been mentioned, it is often difficult during the data exploration phase to examine different patterns simultaneously across more than one grouping of the data. It is only when adjustments are made for the presence and levels of other factors or covariates in a model that the real patterns between the response variable and the primary independent variables of interest become absolutely clear. The sizes of different interactions may also be difficult to foresee.

Sometimes completion of the formal statistical modelling phase may suggest new patterns as yet unexplored. The researcher may wish to return to the data exploration phase study these patterns further.

↑