Basic Computer Skills and Statistical Methods for Analysis of Survey Data

Article · January 2011

Source: OAI

CITATIONS
0

READS
2,472

3 authors, including:

Nicholas Emtage
James Cook University
53 PUBLICATIONS **629** CITATIONS
SEE PROFILE

John Herbohn
University of the Sunshine Coast
285 PUBLICATIONS **2,678** CITATIONS
SEE PROFILE

4. Basic Computer Skills and Statistical Methods for Analysis of Survey Data

Nick Emtage, John Herbohn and Steve Harrison

This module provides an introduction to the use of spreadsheet software packages, to enter, organise and report data from attitudinal and behavioural surveys. In particular, application of the Excel spreadsheet for these purposes is illustrated. The data used for illustration purposes drawn from a survey of landholders' attitudes to forest plantation establishment in north Queensland, Australia. To ensure comprehensive and accurate reporting of the responses to a survey, it is necessary to carry out a carefully designed series of procedures. The basic stages are data entry, reduction and transformation, analysis and reporting. Figure 1 illustrates the methodology adopted to analyse a survey of landholders attitudes to tree planting and management.

The specific procedures which are discussed in this module include:

1. data entry (spreadsheet formatting, data encoding, data entry, data categorisation and transformation);
2. data summary (development of descriptive statistics such as means and measure of variance, summary tables, error checking);
3. data categorisation and transformation (re-categorising nominal data, transforming data to fit normal distributions);
4. data analysis (Chi-square analysis, one-way ANOVA's); and
5. data reporting (presentation of results of analyses).

1. DESIGNING OF THE SURVEY INSTRUMENT TO MAXIMISE DATA UTILITY

The steps taken following data entry depend on the project duration and budget and on the researchers' aims, experience, training and skill. There is no 'right' way to analyse data from surveys, although the formats or types of data collected in the survey and the way they are recorded does determine the types of statistical analysis that can be undertaken. Compiling descriptive statistics of the variables in the data set is the first step and many survey reports fail to go beyond this and analyse the relationships between the variables. The depth of data analysis required will determine the further actions which must be undertaken. If analysis of the relationships between variables is planned, some form of data reduction and transformation is typically needed. Different data types are reduced and transformed in different ways, as illustrated in Figure 1.

It is critically important that the survey instrument (i.e. questionnaire) be designed to provide data in a form that is appropriate for entry into the computer and analysis. Important decisions about the analysis and intended uses of the survey data need to be made prior to the design of the questionnaires. The format of the questions used affects the types of analysis that can be later undertaken. Those designing the survey instrument need to understand the limitations of different formats of data. Data types include nominal data, ordinal data, scales and interval data. If data are collected in nominal (i.e. categorical) form, this limits the way that analyses can be undertaken. Data collected in an ordinal form (i.e. ranked observations) allow the use of more powerful statistical analysis techniques and the data can be collapsed into categories of the analyst's choosing should this be required.

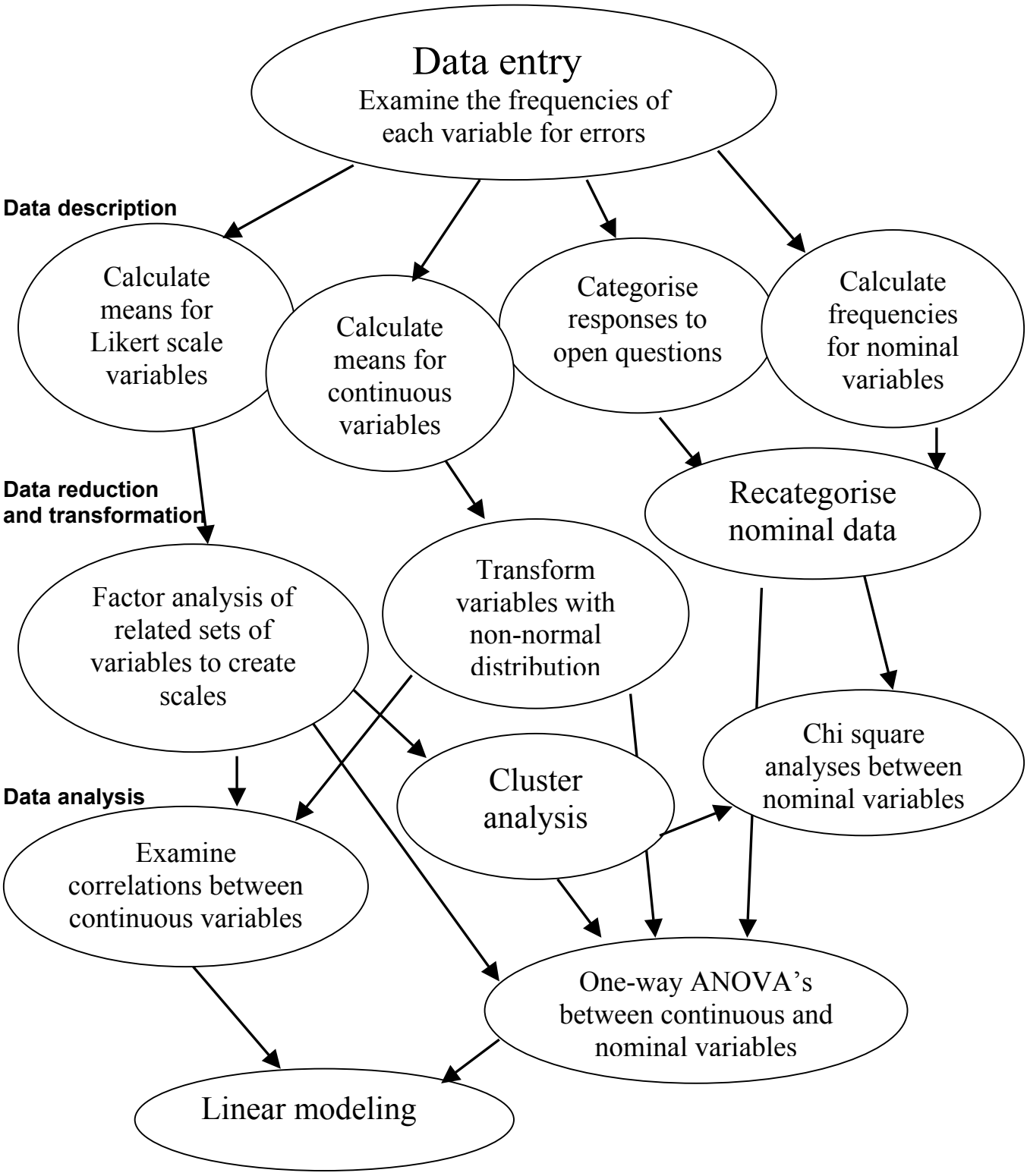


Figure 1. Methodology for analysing the responses to a survey of landholders in the Far North Queensland region of Australia

Source: Emtage *et al*., in prep.

The desire to collect data in formats that allow greater analytical power has to be balanced against ethical concerns and the need to maximise responses. In Australia many university ethics committees will not approve research that asks for a large amount of detail about an individual. For example, questions about respondents' age are often required to be formatted as class intervals rather than specific number of

years. Such formatting may also be more comfortable for respondents than asking them to state their exact age.

It is important to test the survey instrument to ensure that it is well designed, that questions are clear, and the range of responses can be accurately assessed. It is also important to test the data entry and analysis. This provides the researcher the opportunity to set up the data-entry spreadsheets, and to assess what statistical tests can be legitimately used given the types and formats of data being collected. It also allows the researchers to assess the numbers of responses that may be required to run various statistical tests if the number of categories used for nominal variables is known.

2. DATA ENTRY

A number of factors require consideration at the time of data entry. They include choosing which software package to use for the data analysis, setting up the data entry spreadsheet, and setting up data categorisation and transformations.

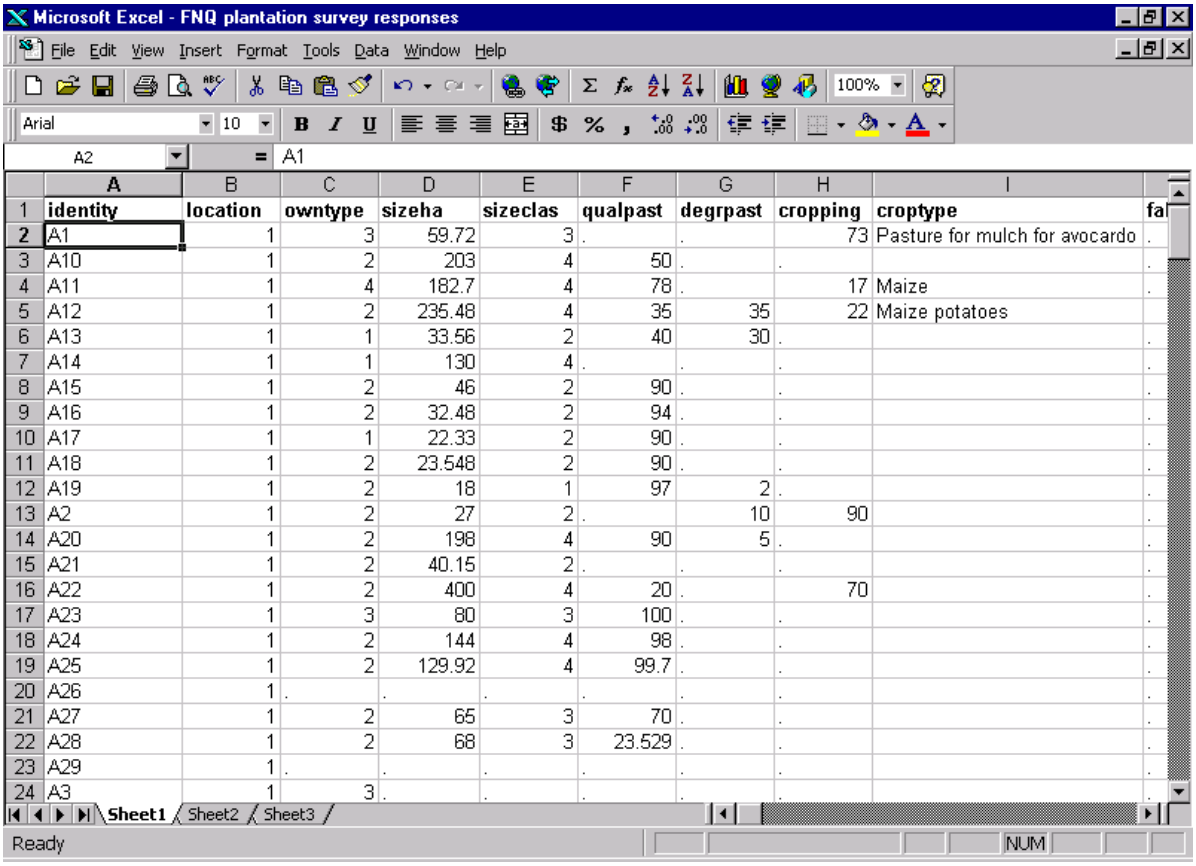
Choosing the software package to use

When entering data from survey responses the researcher needs to consider the types of analysis they wish to undertake and the availability of software packages. If the researcher plans to undertake advanced statistical analysis using multivariate analysis of variance, multiple regressions, factor analysis, cluster analysis or discriminant analyses, then specialist statistical packages such as SPSS (Statistical Package for the Social Sciences) or SAS (Statistical Analysis System) will probably be required. Unless the researcher understands advanced mathematics and statistical theory and can write their own formulae, entering data directly into these specialist programs can save time. If the researcher does not plan to undertake sophisticated statistical analyses or does not have access to such specialist packages then basic analyses can be undertaken using spreadsheet packages such as Microsoft's' Excel or Lotus 1-2-3. The SPSS package allows users to import data directly from Excel if the user later decides to undertake more advanced statistical analyses or the package becomes available. It is recommended that researchers use the specialist packages for all analyses where possible, even the most basic, because of the greater ease of analysis and reporting from statistical programs relative to spreadsheet programs. It should be remembered that all software packages take time to learn. Basic familiarity with large programs such as SPSS and Excel can take months while a high level of expertise may take years of experience to acquire. For the purposes of this module, data entry and analysis is illustrated with reference to Excel spreadsheets, because this package is widely available (as part of the Microsoft Office software) and most researchers have some familiarity with it.

Setting-up the data entry spreadsheet

Just as it is important to know what types of analysis will be attempted when designing the survey instrument, it is also important to keep the intended analysis in mind when entering the data into the software program. As a general principle, a master spreadsheet (and back-up copies!) should be used to enter the data where attempts are made to capture the greatest possible details in the survey responses. For ease of analysis the detail can be summarised or reduced in later copies of the spreadsheet. It is inconvenient to add detail at a later stage and the data entry has to be finished before analyses can commence so it is best to start by entering all available information.

In the north Queensland forestry survey, the survey instrument was a self-administered (i.e. postal) questionnaire. Respondents sent the questionnaires back to the research team using pre-paid and self-addressed envelopes. A master recording spreadsheet was set-up in Excel with the respondents labeled using an identifying code in the first column, and with their responses to each question recorded in subsequent columns (Figure 2).



Microsoft Excel - FNQ plantation survey responses

File Edit View Insert Format Tools Data Window Help

Arial 10 B I U

A2 = A1

	A	B	C	D	E	F	G	H	I	
1	identity	location	owntype	sizeha	sizeclas	qualpast	degrpast	cropping	croptype	fa
2	A1	1	3	59.72	3	.	.	73	Pasture for mulch for avocardo	.
3	A10	1	2	203	4	50	.	.		.
4	A11	1	4	182.7	4	78	.	17	Maize	.
5	A12	1	2	235.48	4	35	35	22	Maize potatoes	
6	A13	1	1	33.56	2	40	30	.		.
7	A14	1	1	130	4	.	.	.		.
8	A15	1	2	46	2	90	.	.		.
9	A16	1	2	32.48	2	94	.	.		.
10	A17	1	1	22.33	2	90	.	.		.
11	A18	1	2	23.548	2	90	.	.		.
12	A19	1	2	18	1	97	2	.		.
13	A2	1	2	27	2	.	10	90		.
14	A20	1	2	198	4	90	5	.		.
15	A21	1	2	40.15	2	.	.	.		.
16	A22	1	2	400	4	20	.	70		.
17	A23	1	3	80	3	100	.	.		.
18	A24	1	2	144	4	98	.	.		.
19	A25	1	2	129.92	4	99.7	.	.		.
20	A26	1	.	.	.	.	.	.		.
21	A27	1	2	65	3	70	.	.		.
22	A28	1	2	68	3	23.529	.	.		.
23	A29	1	.	.	.	.	.	.		.
24	A3	1	3	.	.	.	.	.		.

Sheet1 Sheet2 Sheet3

Ready NUM

Figure 2. Extract from data entry spreadsheet for north Queensland forestry survey

Data categorisation and transformation

In the example presented in Figure 2, to maintain confidentiality the respondents have been labeled using a code (in column A). Coding is used not only to maintain confidentiality, as in this case, but also to speed up data entry. Note that the responses to some questions are already coded.

For example the responses to the question about the ownership type (which included 'partnerships', 'sole trader', 'business' and others) has already been coded into a numerical format rather than writing the full category title for each respondent. This is easily done when there are a limited number of categories. In column 'I' (croptype) the full text of responses has been entered because this question was framed as an 'open' question. Once all of the responses have been entered the range of responses can be assessed and a decision made about how to collapse or reduce the data. In SPSS, labels can be applied to categories which are then shown in reports of analyses to aid interpretation of the data. An important part of pilot testing a survey instrument is to identify the likely range of responses to such a question in order to determine whether to include a discrete range of responses in the questionnaire (plus an 'other' category), or frame the question in an open format.

An example of the categorisation of continuous data is provided in columns 'D' and 'E' relating to the size of the property operated. In this case the range of responses in column 'D' were examined and size classes were determined and computed as a new variable in column 'E' (i.e. less than 20ha = 1, 20-<50 ha = 2, 50-<100 ha = 3, and >100ha = 4). This is one example of transforming variables to create new variables to assist in summary and analysis of the survey responses. In other cases, transformation of responses may be necessary because of the assumption of normal distribution required for some statistical tests, including one-way ANOVA, as discussed later.

3. DATA SUMMARY

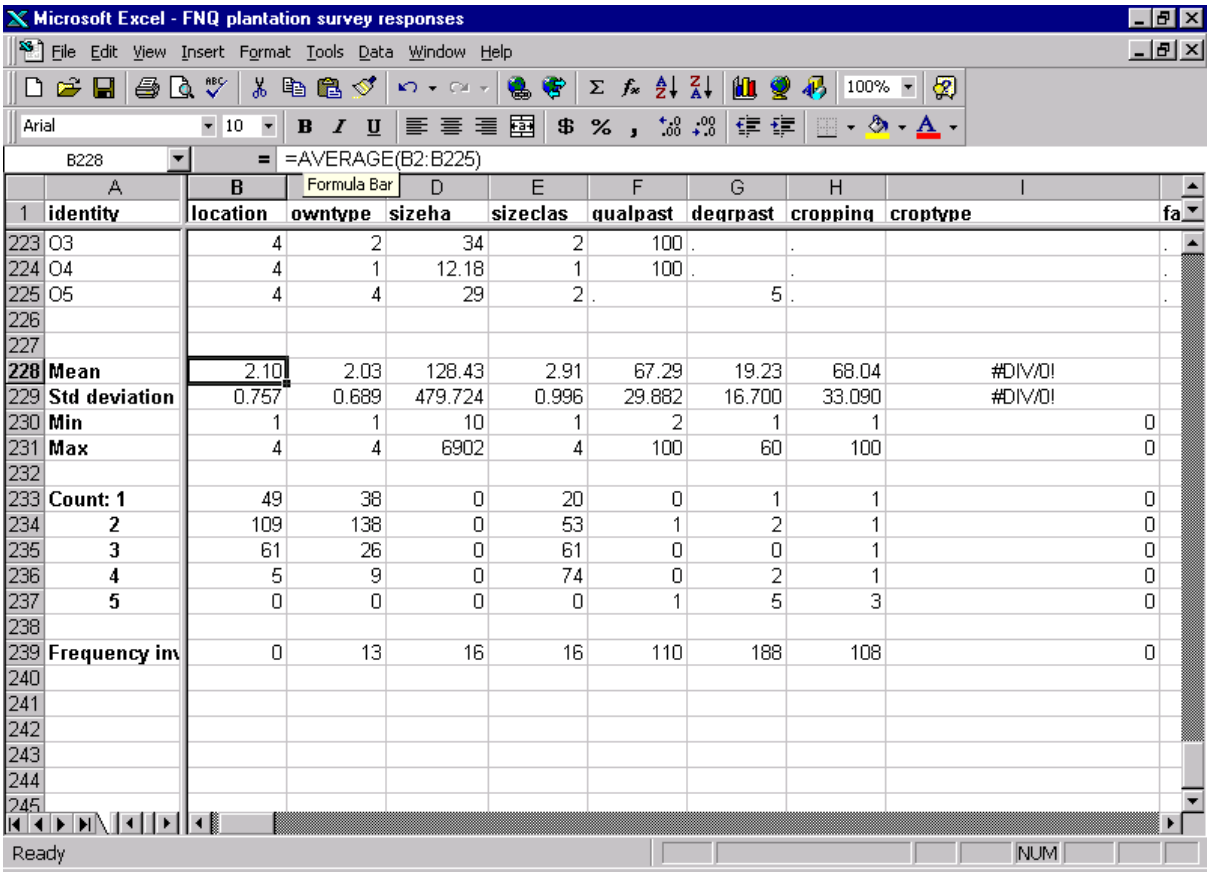
Part of the advantage of using spreadsheets to enter and organise data from surveys is the potential to calculate quickly descriptive statistics of responses to various questions. The specialist statistical software packages such as SPSS are designed for this task and are easier to use than spreadsheet programs such as Excel for this purpose although Excel is relatively simple to use. The development of descriptive statistics by writing formulae into cells is illustrated in Figure 3. Note that the different data types or formats require different summary measures. The calculation of means for categorical variables such as 'location' (column B in Figure 3) is meaningless while the 'count' of the number of responses in each category is valid. It is quicker to type a formula into a cell (e.g. cell B228 in Figure 3) then copy it across the spreadsheet than to enter formulae into each cell individually depending on the data type. Users can go through the columns and delete the irrelevant statistics if they wish to avoid confusion. Organisation of the data for analysis and reporting is necessary. This can be done through categorisation of the sheets in a spreadsheet. Data entry is made onto a 'master' spreadsheet, then copies of this are and are used to carry out data transformation and analyses. The separate sheets in the workbook can be organised to summarise data by topics, organised as summaries of the statistical tests used in analyses, or both can be used. The filing system used to manage the volumes of data generated by surveys and their analysis is up to the researcher.

Some of the summary statistics that can be developed using functions in Excel are illustrated in Figure 4 that shows the 'Paste Function' dialog box. Clicking on the '*fx*' button on the 'standard' toolbar at the top of the screen when Excel is running (as shown in Figures 2 and 3) accesses these functions. The dialog box then prompts users to enter the required parameters for a function. Once the user knows the syntax for these functions they can be typed directly into the formula bar (as shown in Figure 3). An alternative to generating summary statistics using the calculation functions is to use the 'Pivot Table Function' that is available under the 'Data' menu in Excel. This function is discussed further below.

The summary statistics serve three functions. First, they illustrate the types of respondents in terms of their land size, average age, education, land use activities and so on. Second, these averages can be compared to regional or national averages to assess if the respondents to the survey are representative of the broader community (non-response bias tests should also be used). Third, examining the summary statistics helps to identify if there are recording errors in the database. It is easy to make typographic errors that can seriously affect later statistical tests and examination of the database prior to running statistical tests is essential.

Another powerful feature of Excel that can be used to help analyse and report data is the 'macro'. Macros allow users to write their own functions in Visual Basic computer code for specific applications. Like the use of the Excel program generally, it takes time to become familiar with the use of macros and to set-up new code.

If users only need to undertake an operation such as categorising an ordinal variable several times it is probably more efficient to do these tasks manually. If a task is repetitive and needs to be carried out many times it can be more efficient to record and alter a macro to automate the task. Following data entry, macros can be used to automate virtually any of the tasks involved in transforming, analysing and reporting data from a survey. Whether developing macros is more efficient than manually carrying out these tasks depends upon the size of the database being used, the repetition involved in the tasks, and the skills of the researcher as a programmer using Visual Basic.



	A	B	C	D	E	F	G	H	I	
1	identity	location	owntype	sizeha	sizeclas	qualpast	degrpast	cropping	croptype	fa
223	O3	4	2	34	2	100	.	.		.
224	O4	4	1	12.18	1	100	.	.		.
225	O5	4	4	29	2	.	5	.		.
226										
227										
228	**Mean**	2.10	2.03	128.43	2.91	67.29	19.23	68.04	#DIV/0!	
229	**Std deviation**	0.757	0.689	479.724	0.996	29.882	16.700	33.090	#DIV/0!	
230	**Min**	1	1	10	1	2	1	1	0	
231	**Max**	4	4	6902	4	100	60	100	0	
232										
233	**Count: 1**	49	38	0	20	0	1	1	0	
234	**2**	109	138	0	53	1	2	1	0	
235	**3**	61	26	0	61	0	0	1	0	
236	**4**	5	9	0	74	0	2	1	0	
237	**5**	0	0	0	0	1	5	3	0	
238										
239	**Frequency inv**	0	13	16	16	110	188	108	0	
240										
241										
242										
243										
244										
245										

Figure 3. Descriptive statistics developed in Excel

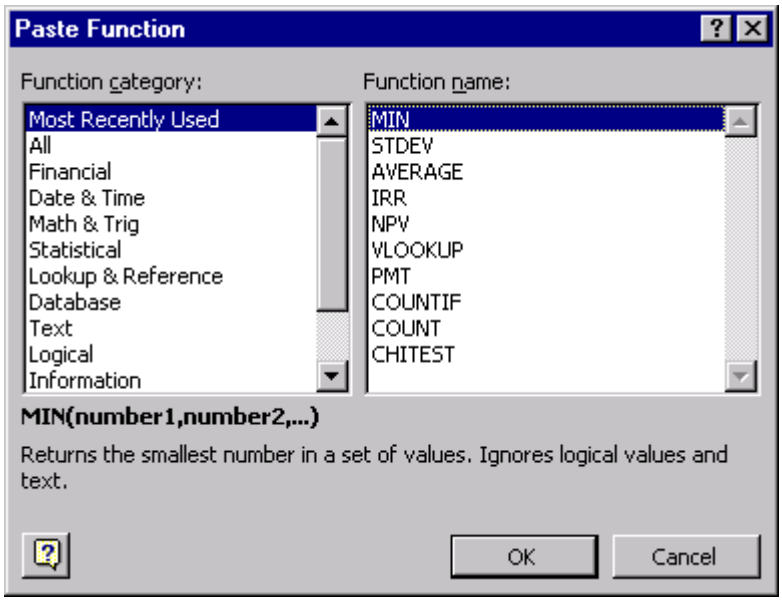


Figure 4. The 'Paste function' dialog box

Once the summary statistics have been computed they can be entered into tables to aid the interpretation of the data. The tables can be organised to contain only related variables, i.e. those related to a particular subject. Two such tables are Tables 1 and

2. It is likely that a reasonable sized survey covering several topics will require the construction of many such tables. Graphs are another way to present data, as discussed in a later section.

In some cases it is useful to present summaries of data using two categories such as land size classes by location as illustrated in Table 3. The 'Pivot Table Report' function in Excel (available under the 'Data' menu) allows users to put together quickly tables that summarise one or more than one variable.

Another Excel function that can be used to construct summary tables for numerical data is the 'descriptive statistics' function. This function is located under 'Data analysis' which is in the 'Tools' menu. The dialog box shown after following the above steps guides users through the use of the function. The table produced is like Table 4.

4. DATA CATEGORISATION AND TRANSFORMATION

Once the responses to the survey have been entered into a database and the database has been examined for errors, the next step toward data analysis involves categorising and transforming the data into formats suitable for analyses. In the case of nominal data, particularly with questions that have been framed in an open format, the researcher often has to re-categorise the initial responses before analyses are possible.

A trade-off is usually necessary between maintaining the details of the responses and being able to analyse and report them.

Table 1. Land uses as a proportion of the total landholding for all respondents (%)

Statistical measure	Quality pasture	Degraded pasture	Cropping	Fallow	Forest	Other
Average	67.29	19.23	68.04	11.85	26.89	12.72
Standard deviation	29.88	16.70	33.09	11.38	28.41	17.17
Minimum	2	1	1	1	0.3	1
Maximum	100	60	100	50	100	100

Table 2. Ratings of importance (on 5-point scale) for various reasons for planting trees by all respondents

Reason for planting trees	Average	Standard deviation	Minimum rating	Maximum rating	n
Other reasons	4.39	0.839	2	5	23
Protect land resources	3.98	1.157	1	5	172
Protect water resources	3.96	1.193	1	5	170
Provide fauna habitat	3.64	1.256	1	5	169
Personal reasons	3.44	1.301	1	5	170
Aesthetic reasons	3.35	1.327	1	5	168
Increase value of land	3.16	1.362	1	5	166
Windbreak	3.15	1.483	1	5	168
Legacy for children	3.13	1.514	1	5	166
To make money	2.66	1.472	1	5	167
Diversification of income	2.39	1.492	1	5	163
Superannuation	2.16	1.483	1	5	164
Fenceposts	1.52	0.975	1	5	161

Table 3. Size classes of respondents by location

Location	10 – 20 ha	20 – 50 ha	50 – 100 ha	>100 ha	Missing	Total
Atherton	6	13	12	13	5	49
Johnstone	1	26	30	44	8	109
Eacham	12	11	19	16	3	61
Unknown	1	3		1		5
Totals	20	53	61	74	16	224

Table 5 presents the results of applying the pivot table function to count responses to an open-ended question that asked landholders what types of crops they grow on their land. It can be seen that a number of the categories are really the same (e.g. Banana and cane; or Cane, bananas, or cane and bananas), but slight differences in the way they have been entered means that the pivot table function reads them as different categories.

There two steps to addressing this problem. The first is to be consistent when entering the responses into the database. A hard (i.e. paper) copy of the questionnaire can be used to record new responses to each question as they are being entered.

This copy can be consulted when recording responses to open-ended questions, or categorising responses to nominal questions that have an 'other' category that is effectively open ended. This ensures that consistent names are given to the same responses. The second step once responses have been entered into the database is to define categories based on examination of a range of responses, like those presented in Table 5.

Table 4. Descriptive statistics for selected land use variables

Statistic	Quality pasture	Degraded pasture	Cropping
Mean	68.2368	23.6667	68.4741
Standard Error	2.69739	3.27375	3.02579
Median	80	20	83.5
Mode	100	30	100
Standard Deviation	28.8003	19.6425	32.5887
Sample Variance	829.457	385.829	1062.03
Kurtosis	-0.754	-0.1582	-0.9025
Skewness	-0.6836	0.86976	-0.744
Range	98	70	99
Minimum	2	1	1
Maximum	100	71	100
Sum	7779	852	7943
Count	114	36	116

Table 5. Initial crop types in the responses database

Crop type	Frequency	Crop type	Frequency
None	107	Cane, bananas	2
Aloe Vera, maize and taro	1	Cane, bananas, nursery	1
Avocados	1	fruit trees	1
Banana and cane	2	Hay	1
banana, pawpaw	1	Maize	2
Bananas	15	Maize Peanuts Potatoes	1
Bananas, pawpaw	1	Maize, potatoes	1
Beans and zucchini	1	Maize, peanuts, vegetables	1
Cane	62	Mangoes	1
Cane & banana	7	Orchid	1
Cane & exotic fruit	1	Pasture seed	1
Cane & pawpaws	2	Peanuts, cane	2
Cane and pawpaws	1	Sorghum, oats and hay crops	1
Cane pawpaw	2	Sorghum, oats, rye and grass for silage	1
cane, bananas	1	Tea, cane	1
Total			223

The definition of categories is up to the researcher and depends upon the number of responses to the questionnaire and the variation in the data. Categorical data are more limiting than ordinal data in terms of the statistical analyses that can be used. One question facing researchers that wish to analyse relationships between variables defined using categorical data is how to establish a series of categories that maintain the diversity in the data yet still have sufficient responses in each category to allow the use of statistical analyses like the chi-squared test and one-way ANOVA. When carrying out chi-square tests, each cell in the table of expected responses should have at least five respondents. If more than 25% of the cells in the expected frequency table do not have five responses the test results may be unreliable.

Several new variables could be created from the data in Table 5. The simplest variable would record the presence or absence of cropping as shown in Table 6. This variable would have the advantage of having many respondents in each category, and the disadvantage of losing a lot of information about the types of crops that are grown.

Another way to classify the data could include some more details about the types and mixtures of crops commonly grown (Table 7).

Table 6. Number of respondents growing crops on their land

If crops grown	Frequency
Crops	121
No crops	103

Table 7. Number of respondents growing crops on their land

If crops grown	Frequency
No crops	107
Cane only	62
Cane and other crops	22
Crops other than cane	33

The resulting classification scheme has four categories and reasonable numbers of respondents in each category. The implications of different classification schemes for categorical data will be further examined in the following section.

5. DATA ANALYSIS

The Excel program contains a number of basic data analysis functions including chi-square tests for independence. An 'add-in' can be loaded with additional statistical functions including t-tests, z-tests, correlation, covariance, regression and ANOVA. In this section the chi-square test is examined.

The relevant application of the chi-square for this discussion is to assess whether there is a relationship between two sets of nominal (categorical) data, known as the *chi-squared test of independence*.

The null hypothesis for this test is that there is no relationship between the two data categories[1]. To run the test in Excel the user has to calculate the expected frequencies of values under the null hypothesis in a table and compare these values with the distribution of observed frequencies. The Pivot Table function makes it easy to compile the table of actual values. An example is provided in Tables 8 and 9. The expected frequencies are calculated by multiplying the row total by the column total then dividing the result by the grand total. Thus the expected frequency of those who have primary school education and have not planted is calculated as (33 x 123)/196 = 20.71.

The chi-square test for independence is performed using the CHITEST function in Excel. The chi-square statistic is calculated as the sum over the rows and columns of: (observed frequency – the expected frequency)^2 / expected frequency. The calculated statistic is then compared to a critical value for the chi-square statistic for the relevant number of degrees of freedom (the product of number of rows less 1 and number of columns less 1). The CHITEST function returns the probability for a chi-square statistic for the relevant number of degrees of freedom. If the probability of the statistic is less than the designated significance level (usually set at 0.05), then the null hypothesis is rejected and it is concluded that there is a relationship between the two variables or categories. In the above example, with the probability of the chi-square statistic of 1.3^{-5} or 0.000013, it is concluded that there is a difference in planting behaviour between those with different levels of formal education. In other words, those with diplomas and degrees are more likely to plant trees than those with primary and secondary education.

As mentioned in the preceding section, the categorisation scheme used to reclassify data for analyses has important implications for the types of statistical tests that can legitimately be carried out.

Difficulties may arise in surveys with relatively small samples if researchers attempt to test relationships between ordinal variables with more than a few categories each.

Consider the example of the different ways of categorising the types of crops grown by landholders in Tables 6 and 7. The data set of responses to the survey does not have sufficient information to legitimately test the relationship between the crop types grown by respondents and their level of formal education (Tables 10 and 11). More than 25% (5/16) of the cells in the table of expected values (Table 11) have a value of less than 5. The probability of obtaining the chi-square statistic in this case is 0.02, which is less than 0.05, but the result should not be reported since the test is invalid.

In the example below there are too many categories in each variable to carry out a chi-square test. The alternative is to reduce the number of categories in one or both of the variables. An example of this procedure is illustrated in Tables 12 and 13.

[1] Technically, this is a test of whether the joint probability distribution is the product of the univariate probability distributions for each of the variables. Further details can be found in Harrison and Tamaschke (1993, pp. 222-224).

Table 8. Actual frequency of respondents who have planted more than 30 trees by education classes

Education category	If planted		Total
	No	Yes	
Primary school	23	10	33
Secondary school	82	31	113
Diploma	12	11	23
Degree	6	21	27
Total	123	73	196

Table 9. Expected frequency of respondents who have planted more than 30 trees by education classes

Education category	If planted		Total
	No	Yes	
Primary school	20.71	12.29	33
Secondary school	70.91	42.09	113
Diploma	14.43	8.57	23
Degree	16.94	10.06	27
Total	122.99	73.01	196

In the second example (Tables 12 and 13,) the reduction in categories of the cropping variable means that there is sufficient responses in each cell to use a chi-square test. For this example the probability of the chi-square statistic returned by the test is less than 0.0001. Thus the statistical decision can be made to reject the null hypothesis, with the practical inference that there are different proportions of the population growing crops when comparing those with different levels of formal education. Inspection of the observed and expected frequencies used in the test tells us that those with lower levels of formal education are more likely to grow crops than those with higher levels of formal education. The combining of categories involves some loss of information about relationships between the variables and thus diminishes our understanding about the relationships.

It can be seen from Table 10 that no respondent with a degree reported growing only sugarcane as a crop. If the researcher thinks that this point is important and worth pursuing then it possible to construct another variable for the types of crops grown by respondents, with three categories.

As the survey has sufficient respondents who report growing sugarcane only this category can be retained, as can the category of respondents who grow no crops. The third category combines those who grow sugarcane and other crops, and those who grow other crops but no sugarcane. The observed frequency table of those with different levels of education by different crop growing categories would then appear as in Table 14, and the expected frequencies would be as in Table 15.

Table 10. Actual frequency of cropping categories by education classes

Cropping category	Education category				Total
	Primary	Secondary	Diploma	Degree	
No crops	12	42	15	21	90
Cane only	14	39	3		56
Cane and …	2	15	1	1	19
Other	5	17	4	5	31
Total	33	113	23	27	196

Table 11. Expected frequency of cropping categories by education classes

Cropping category	Education category				Total
	Primary	Secondary	Diploma	Degree	
No crops	15.2	51.9	10.6	12.4	90
Cane only	9.4	32.3	6.6	7.7	56
Cane + other	3.2	11.0	2.2	2.6	19
Other	5.2	17.9	3.6	4.3	31
Total	33.0	113.1	23.0	27.0	196

Table 12. Actual frequency of crop growing categories by education classes

Education category	Crops	No crops	Total
Primary school	21	12	33
Secondary school	72	41	113
Diploma	8	15	23
Degree	6	21	27
Total	107	89	196

Table 13. Expected frequency of crop growing categories by education classes

Education category	Crops	No crops	Total
Primary school	18.0	15.0	33
Secondary school	61.7	51.3	113
Diploma	12.6	10.4	23
Degree	14.7	12.3	27
Total	107.0	89.0	196

Table 14. Actual frequency of those with different levels of education by different crop growing categories

Education category	No crops	Cane only	Cane and other crops	Total
Primary school	12	14	7	33
Secondary school	42	39	32	113
Diploma	15	3	5	23
Degree	21		6	27
Total	90	56	50	196

Table 15. Expected frequency of those with different levels of education by different crop growing categories

Education category	No crops	Cane only	Cane and other crops	Total
Primary school	15.2	9.4	8.4	33
Secondary school	51.9	32.3	28.8	113
Diploma	10.6	6.6	5.9	23
Degree	12.4	7.7	6.9	27
Total	90.1	56.0	50.0	196

The probability for the chi-square statistic for the data in Tables 14 and 15 is 0.010. As this is less than the critical probability of 0.05, the decision is made to reject the null hypothesis, i.e. there is a significant difference in terms of the types of crops grown by respondents with different levels of formal education. It can thus be concluded that this type of difference exists in the underlying population Comparison of the observed and expected frequencies suggests that the likely source of the difference is the lower than expected frequency of those with degrees growing only cane.

6. DATA REPORTING

The preceding section has illustrated some forms of summary tables used to present data. The way in which data are presented depends upon the type of report being compiled and the types of statistical tests performed. When survey data are analysed, the presentation can occur on a number of levels (as illustrated in Figure 1). Reporting of survey responses should cover:

- responses to survey questions;
- transformation of response data in preparation for data analysis; and
- results of all analyses of relationships between variables prepared from the survey responses.

The first stage of reporting is to summarise responses to each question used in the survey before they are modified. Most survey reports have a section describing on the types of respondents to the survey; tables summarising the data collected about the socio-economic characteristics of the respondents can be used to describe respondents as well as discuss the potential of non-response bias. Where the survey is large – in terms of sample size and number of questions – the researcher may use appendices to report large amounts of data and concentrate on those analyses and descriptions that are most relevant to the research questions. In the case of the examples used in this paper (drawn from a survey of landholders tree planting and management attitudes and behaviour), the initial data should include description of the socio-economic characteristics of respondents. The descriptive sections for a report should be organised to present the responses by topics covered in the survey. The various topics in this case included the reasons landholders plant trees, restrictions to tree planting on their land, their past and intended planting behaviour, their attitudes to tree planting on a regional scale, and their attitudes to past and potential tree planting incentive and assistance schemes.

In the initial descriptive reporting of survey findings, the responses should be reported as an average or mean figure for all respondents. Where the survey has covered clearly different political or geographic areas, or clearly different types of people in socio-economic terms, then the descriptions of responses may be organised to illustrate these differences in the respondents. In the case of the north Queensland survey, three local government areas over two distinct bio-geographic regions were included. Two of the government areas are located in an upland area, and the third is coastal. The differences in the two types of areas arise from differences in their climates, topography and soils, as well as the farm sizes and enterprise types. Initial description of the responses to the survey showed the average responses to the various questions for all respondents and for respondents from each local government area. The presentation of these data also described tests for significant differences in characteristics of respondents in the various local government areas. An example of such information is provided in Table 16.

Using graphs is an excellent way to display data for descriptive purposes or to illustrate the results of analyses. Note that graphs in Excel are called ‘charts’. The type of graph used varies according to the type of data involved and the intentions of the researcher. The *pie chart* format can be used to illustrate the average proportion of land used for different activities as shown in Figure 5.

Where the data are in continuous or ordinal form, line graphs or histograms may be used. Line graphs are particularly useful to aid interpretation of relationships between ordinal variables and to assess if the distribution of the variable is ‘normal’ or at least linear. An example of this is shown in

Figure 6, illustrating the initial distribution of land sizes before they are standardised, with Figure 7 illustrating the distribution of the standardised values.

To obtain the graph shown in Figure 6, the raw data were first copied to a new sheet and sorted according to property size (land area). Examination of the maximum value for the variable showed that one respondent reported a property size of 6902 ha which is clearly an extreme case given that the next largest property size is only 500 ha.

Table 16. Importance placed upon various reasons for planting trees by landholders in the Johnstone, Atherton and Eacham shires

Reason for planting	Rating by shire			Sign. diffs.		Mean rating (all shires)	n	Frequency rated 5
	J	A	E	LSD	Bon.			(%)
To protect and restore land	3.9	3.9	4.2	ns	ns	4.0	172	42
To protect the local water catchment	3.8	4.0	4.2	ns	ns	4.0	170	42
To attract wildlife and birds	3.5	3.7	3.8	ns	ns	3.6	169	31
Personal interest in trees	3.3	3.4	3.7	ns	ns	3.4	170	26
To improve the look of the property	3.2	3.5	3.6	ns	ns	3.3	170	26
To increase the value of the farm	3.1	3.2	3.2	ns	ns	3.2	166	19
To create windbreaks	2.8	3.4	3.4	A. E. > J	ns	3.1	168	25
Legacy for children or grand children	3.3	2.7	3.2	J > A	ns	3.1	166	26
To make money in the future	2.9	2.5	2.4	ns	ns	2.7	167	15
To diversify farm business	2.6	2.2	2.2	ns	ns	2.4	163	13
Superannuation or retirement fund	2.3	2.1	2.1	ns	ns	2.2	164	13
To provide fence posts	1.5	1.8	1.4	ns	ns	1.5	161	3

Notes: (1 = not important, through to 5 = very important). ‘J’ = Johnstone, ‘A’ = Atherton, ‘E’ = Eacham. Significant differences between means for each shire were tested using least square difference (lsd) and Bonferroni tests ($P > 0.05$). Significant differences between mean ratings for responses for each question were tested using the Bonferroni test. Overlapping lines indicate means which are not significantly different from each other. The mean rating for all shires includes five responses that could not be classified by shire.

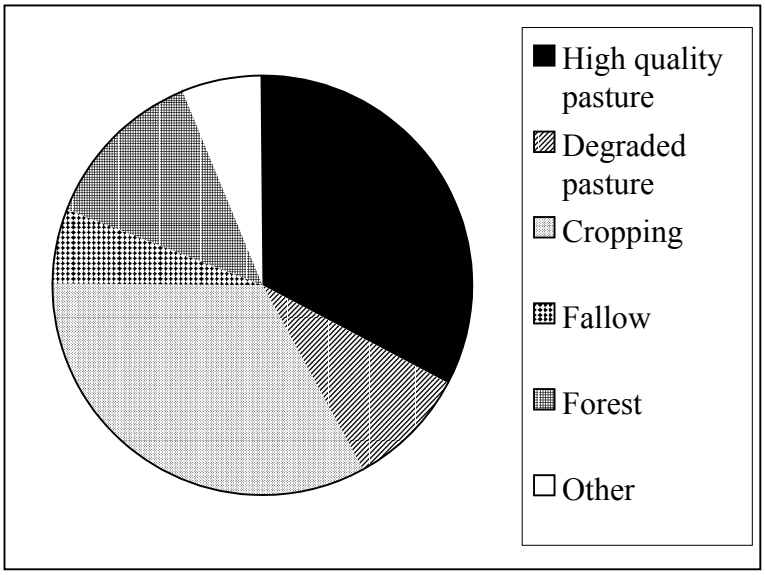


Figure 5. Average proportion of landholding used for various purposes in far north Queensland

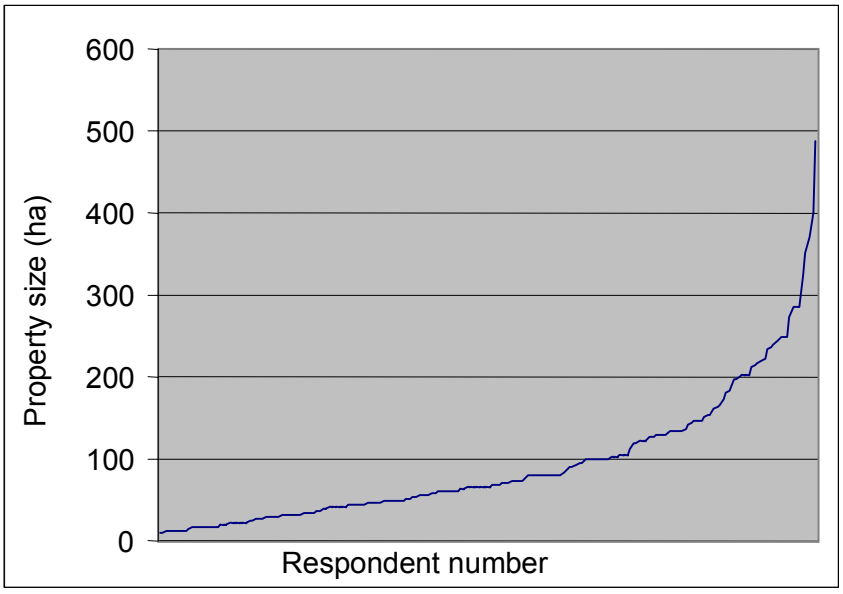


Figure 6. Distribution of values for the variable Landsize

The graph used to illustrate the distribution of the variable therefore dropped the largest value as the graph scale becomes useless when it is included. The shape of the distribution is parabolic indicating that it could be transformed to an approximately linear cumulative distribution using the Log10 function (i.e. which calculates logarithms to the base 10) in Excel. The data for the variable were transformed by taking the Log10 of the initial values and a new variable LogSize was created. The distribution of this new variable is illustrated in Figure 7.

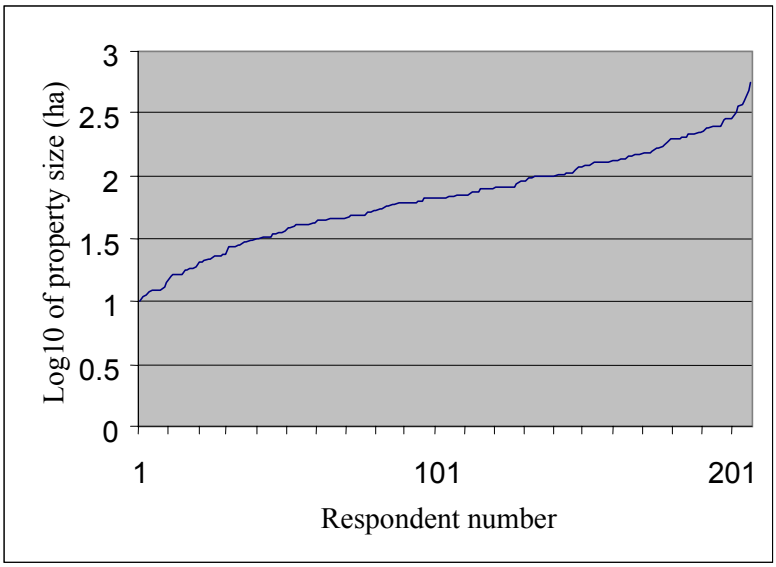


Figure 7. Distribution of values for the variable LogSize

When copying and pasting graphs from Excel to Word (or PowerPoint), open both the Excel file from which the graph is to be taken and the Word file into which it is to be placed. Copy the graph using the 'copy' function under the 'Edit' menu in Excel, then use the 'Paste special' function under the 'Edit' menu in Word to select the format used to save the graph in the Word document. Using the 'picture' format for the graphs creates the smallest file size, but does not maintain a link with the Excel file used to create the graph, and is more difficult to edit than a graph saved as an 'Excel object'.

7. CONCLUDING COMMENTS

Modern statistical packages provide a convenient means to store survey data and powerful facilities of descriptive and statistical analysis. Individual researchers tend to have their favourite data analysis packages, although Microsoft's Excel spreadsheet package and SPSS are widely used. Familiarity with statistical packages requires practice in their use, but some simple steps can be laid down for new users, as set out in this module. It is critical to plan the types of analysis intended when developing the questionnaire for a survey.

REFERENCES

Emtage, N. F., J.L. Herbohn, S.R. Harrison, and D.B. Smorfitt (in prep.), 'Landholders attitudes to farm forestry in far North Queensland: report of a survey of landholders in Eacham, Atherton and Johnstone shires', Rainforest Cooperative Research Centre, James Cook University, Cairns.

Harrison, S.R. and Tamaschke, R.H.U. (1993), *Statistics for Business, Economics and Management*, Prentice-Hall, New York.