Brief notes on statistics
Practical aspects of using statistics in research

Michael Wood (Michael.wood@port.ac.uk)

15 December 2012

Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. There is a video on How to use the statistical package SPSS at http://youtu.be/5lchPbq7Tec .

I am assuming here that you have a rough idea of the research you want to do, and a reasonable understanding of the relevant statistical concepts. If you want to bring statistics into your analysis (and I would say you should with most projects!), what do you need to check, and how should you proceed? My suggestions are …

Data

You may start from **primary data** (that you've collected yourself), or you may be using **secondary data** (collected by someone else). In either case, you should think hard about where it comes from and any reasons for suspecting bias or inaccuracies. (It's probably a good idea to consult a book on research methods, particularly if you are collecting primary data yourself.) Secondary data can come from a wide variety of sources (typically websites such as http://www.statistics.gov.uk for UK National Statistics), but you should check the source carefully to ensure it is credible. It is always worth thinking about:

1. **Where your data comes from and the method of sampling.** Remember that some (most?) samples are biased or may not be representative of your target population – see any book on research methods. Most statistical techniques assume your sample is a random one, or similar to a random one – so it's important to discuss any discrepancies between this ideal and the sample on which your data is based. One common problem is non-response to a survey. Can you be sure that those who respond will give similar answers to those who don't?

2. **The size of the sample.** Try to make sure that your sample is large enough to meet your research aims, but not so large that you are wasting time. Normally the main constraint is your time and availability of respondents: in this case the sample should be as large as possible. Don't forget that non-response may make your final sample smaller than you expect, and also that if you are interested in particular categories of respondent, the numbers may be even smaller. Statistical theory can,

in principle, tell you how large a sample you need for a particular purpose (see http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf).

3 **The appropriateness of any measures used.** You may be using a particular way of measuring profits or job satisfaction. You may be taking the average of a number of ratings as an indicator of something. You should check all such measures are sensible. Common sense should be your first guide and you will find more discussion in research methods texts (check under validity and reliability). It is usually a good idea to use a measure somebody else has used (with an acknowledgment, obviously) because then you can cite any checks they have done to justify it.

If you want to analyze it statistically, your data should be organized by ***cases*** (units of analysis) and ***variables***. The cases may be people, organizations, accidents, etc. A variable is anything that varies between cases – these may include responses to questions on a rating (Likert) scale, age, salary, job title, sex, etc.

Variables may be ***numerical*** (e.g. response on 1-5 rating scale, age) or ***category*** (e.g. sex, department) variables. They are also usually thought of as either ***dependent*** or ***independent*** variables. For example, if you are researching job satisfaction your dependent variables might be various measures of job satisfaction, and the independent variables might be sex, age, job-title, etc. Your interest is then in understanding how the dependent variables depend on the independent variables. (As an example look at the data at http://woodm.myweb.port.ac.uk/nms/accquest.xls.)

If you are doing anything like a questionnaire survey, you should always do a small pilot study to make sure that the questionnaire is satisfactory and there are no hidden snags. *You should try out the statistical analysis as part of any pilot.*

We'll now look at some of the ways you can analyze data, and then at how you can use a computer (Excel and SPSS) to help you.

Easy methods of analysis – graphs, tables, averages, etc

This is mainly **common sense, but it is vital to think about it!** You will find many of the main possibilities explained in books like Saunders et al (2007, chapter 12) and Wood (2003, chapter 3). You will need to include whatever tables, diagrams, and statistics like averages, correlations, etc you need to analyze your data and explain to readers. Diagrams are often more user-friendly so use these when possible. The main tables and diagrams should go in the main text – these are your main results so don't hide them! Put extra details (e.g. some of the data), which particularly keen readers might want to check, in an appendix.

Make sure that all tables and diagrams are **clearly labeled and *easy to understand***. Resist the temptation to make them complicated: if you have lots of information to convey, you should use lots of tables and diagrams, rather than trying to get it all in one. It is a good idea to **ask a friend**

to check that your diagrams and tables are easy to understand: if your friend has difficulties, then other readers are also likely to have difficulties (including the examiners)! If in doubt try to make it clearer and simpler. This is more difficult than it sounds: you may need several drafts before you find the best approach.

Now have a go at Exercise 2 at the end.

More specific points to remember include:

- Take care with pie diagrams. These may look pretty but they are only sensible where you want to draw attention to different categories as a proportion of a total. If you have a number variable, or if you want to draw attention to how different categories compare with each other, a bar chart or a histogram will make more sense. If in doubt use a bar chart (or a histogram).
- Histograms (see http://woodm.myweb.port.ac.uk/stats/StatNotes1.pdf) and scatter diagrams (see http://woodm.myweb.port.ac.uk/stats/StatNotes2.pdf) are useful and often forgotten.
- Averages are often helpful to compare something in different groups (e.g. Table 5 in http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf).
- Numbers in tables are often clearer if you round the numbers off to avoid giving too many decimal places. For example, if you have a table giving mean results from rating scales (e.g. Table 5 in http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf), these results are likely to be clearer if you round the numbers to one decimal place (you can use Format – Cells in Excel for this).
- Don't forget to check your research aims to make sure that your analysis is meeting them. In many projects, it is useful to find the difference between a number of groups (Compare means in SPSS), or the relationship between a number of variables (Correlate in SPSS). Make sure you don't forget to do this!
- If you've got data on independent variables like sex and age, don't forget to see how they relate to your dependent variables: e.g. you might compare the views of males and females (otherwise there was little point in finding out about the independent variables!).
- And, of course, make sure everything is clear, well labeled and as simple as possible.

More advanced techniques

There are too many of these to cover in detail here. The most frequently used are

- (Null) Hypothesis testing (see http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf). Don't forget to distinguish clearly between null hypotheses, and the interesting hypotheses for which you may be trying to find supporting evidence.

- Regression (see http://woodm.myweb.port.ac.uk/stats/StatNotes2.pdf and http://woodm.myweb.port.ac.uk/stats/StatNotes4.pdf),

although I would like to see …

- Confidence intervals (also discussed in http://woodm.myweb.port.ac.uk/stats/StatNotes3.pdf) used more widely.

To use any of these, you obviously need to understand the underlying concepts.

If you want to use hypothesis tests in your research I would use SPSS. For regression, Excel is usually adequate. If you want to use confidence intervals, SPSS will provide them in a few situations, but in general you will need to do some more advanced reading (see below).

Unfortunately, much of advanced statistics is complicated and acquiring a detailed knowledge quickly is difficult. However, the techniques above are the main building blocks; if you understand these you should be able to work out what the various techniques are doing, in rough terms at least.

If you want to read up on particular techniques, you could try

- Using Google

- The online textbook at http://www.statsoft.com/textbook/stathome.html

- One of the many books on SPSS in the library (use a keyword search) – these are a good source of a non-technical introductions to statistical techniques even if you don't want to use SPSS.

- Robson (2002) – a research methods text with a better-than-average section on statistics.

Data entry and analysis with a computer

When you've got your data your first step should be to enter it into a spreadsheet using this format:

- Variable headings on the top row. If you want to go on to use SPSS (next section) it's easiest if you use short headings (8 characters maximum) without spaces that start with a letter. If the data comes from questionnaires, it's also a good idea to put a

number on each questionnaire and then have this number as the first variable. This will make it easier to check the entries and to find specific questionnaires.

- Each case should go on a new row
- Leave blank cells for data that is missing (do *not* use 0)
- Don't leave blank rows or columns
- For category variables use a convenient abbreviation for each category (and make sure you always use the same abbreviation). Keep a record in another sheet of the spreadsheet of your coding scheme (i.e. these abbreviations and what they refer to).
- With questions whose answers range from "strongly disagree" to "strongly agree" (or similar), you should always use a big number to represent the positive end and a small number to represent the negative end (e.g. 1 for "strongly disagree" and 5 for "strongly agree"; "don't know" responses should be coded by leaving the cell blank). If you do the opposite, your results are likely to confuse everyone, including yourself.
- Where there are two possible entries (e.g. a yes/no question) it will make things easier if you use 0 for one (e.g. no) and 1 for the other (yes). Then the average of all the responses will give you the proportion who answered (e.g.) yes.

There is an example of this format at http://woodm.myweb.port.ac.uk/nms/accquest.xls.

You can analyze data in spreadsheets like this either using Excel, or you can transfer the data to specialist packages such as SPSS. *Even if all the analysis you want to do is very simple, it is a good idea to start by entering your data in a spreadsheet like this. Otherwise you may waste a lot of time!*

Using Excel to analyze data

If you are familiar with Excel you will know much of what it can do. If you aren't familiar with Excel I would suggest that you

- Become familiar with it (may take some time!), or
- Get some help from someone who can use Excel well, or
- Use SPSS (see next section). As well as being more powerful, this is easier to use if you only want to do standard analyses.

The features of Excel which are particularly useful for data analysis are

- The graphs . Use Insert menu on Excel 2007 or Chart wizard on earlier versions. It's all a bit easier if you select (highlight) the data before inserting the graph. Note that you can't do a histogram like this – you will need to check Help.

- The statistical functions (f_x on Insert menu).

- Sorting data (Data – Sort). If, say, you want to compare the average score on one variable for males and females, you can sort the data on the Sex variable and then all the females will be listed together so you can easily use the =average function to work out the average, and then repeat this for the males. There are more sophisticated ways of doing things like this – e.g. the next two bullet points:

- Pivot Tables (see Help).

- Database functions (e.g. =Daverage). See Help.

- If you want to do regression, the Regression Tool is easy to use and gives a lot of information. See Help.

SPSS (Statistical Package for the Social Sciences)

This is a specialist package that is available on the network (under Academic applications) or on a disc from the library. It has three advantages over Excel:

- SPSS will do many more, and much more advanced, statistical procedures, and

- SPSS will "mass produce" tables and diagrams: e.g. if you want to do a histogram for a lot of variables, this is far easier than with Excel, and

- If you don't know Excel, and want to do something standard you may find it easier.

To load an Excel data file (in the above format) into SPSS you will need to tell SPSS to look for an Excel file (xls), and click the box saying that your variable headings are on the top row. If you are using Excel 2007, it may be best save as Excel 2003 files; when I tried to load Excel 2007 files they were not recognized by SPSS.

The menus in SPSS should be easy to navigate. The most useful options are likely to be:

- Analyze – Descriptive statistics – Frequencies. *This will give you means, standard deviations, bar charts or histograms, etc.*

- Analyze – Compare means – Means (tick the ANOVA box under Options if you want to test the null hypothesis that there is no difference between the means). *This is useful if you want to compare the average of some measurement from several groups of people, organizations or whatever (e.g. you might want to know if one group of firms is more profitable than another, or one group of people is happier than another).*

- Analyze – Descriptive statistics – Crosstabs (for frequency tables for two or more category variables). Click statistics then the box labeled Chi Square to test a null hypothesis of no relationship between the variables.
- Analyze – Custom tables
- Analyze – Correlate – Bivariate
- Graph – Scatter
- Analyze – Regression – Linear
- Analyze – Compare means – Paired-samples T test. *If, for example, you have data on a 1-5 scale on how much people like carrots and cabbages, you can use Frequencies (see above) to work out the mean for carrots and for cabbages and compare. The paired sample T test will also test the null hypothesis that there is no difference in the underlying population.*

In all cases you will need to explore the Statistics and Options available. If in doubt use the Help and experiment with a little bit of data so that you can see what's happening.

You should find that you can copy and paste the results into Word or Excel. Copying them into Excel may be the easiest was of editing them, if this is necessary. You can then copy and paste into Word.

There are many books on using SPSS and Excel in the library (use the catalogue). There is a very brief introduction to both packages in the Appendices of Wood (2003), and a YouTube video summary at http://youtu.be/5lchPbq7Tec .

Exercises

1 The data at http://woodm.myweb.port.ac.uk/nms/accquest.xls comes from members of an accounting society. The variables are explained in the Notes sheet of this spreadsheet. (The spreadsheets contain responses to just a few of the questions in the questionnaire: the actual questionnaire was much longer.)

Suppose you wanted to use this data to find out how important social events and networking are to members with a view to deciding whether the Society needs to make any changes. How would you analyze the data? What graphs, tables and statistics would you produce? Do you think the sample is likely to give a fair picture of the membership as a whole?

2 Suppose you are interested in rates of smoking among males and females in a particular organization. Among the females there are 30 smokers and 10 non-smokers. Among males there are 50 smokers and 50 non-smokers. How would you present and analyze this information?

3 Look at the questionnaire and the bar chart showing responses to it at http://woodm.myweb.port.ac.uk/stats/BarChart.pdf. There is a lot wrong with the bar chart. How would you improve it?

References

Robson, C (2002). *Real world research (2nd edition).* Oxford: Blackwell.

Saunders, M.; Lewis, P. & Thornhill, A. (2007). *Research methods for business students (4th edition).* Harlow: Pearson Education.

Wood, M. (2003). *Making sense of statistics: a non-mathematical approach.* Basingstoke: Palgrave.