SMN405 Research Methods for the Public Services
Quantitative Methods Refresher Notes

Graham Cookson*

May 2007

1 Introduction

Weeks 2-6 of your Research Methods module will cover quantitative methods or statistics. It assumes no prior training in mathematics so there will be no algebra or calculus. The course will focus on the most commonly used technique in the social sciences: the linear regression model. It will introduce methods in an intuitive and wherever possible non-technical way. After each lecture there will a practical session in the computer lab to practice and reinforce the material covered in the lecture.

This course will have succeeded if you can apply statistical reasoning in analysing data to answer questions in your work and research. By the end of this course you should:

- Have an understanding of basic statistical concepts.
- Be familiar with some of the statistical terms commonly used in the analysis, interpretation, and presentation of data.
- Be familiar with some of the basic statistical techniques, and be able to apply them to problems you will encounter in your studies and careers.
- Be able to assess information and statistical arguments commonly found in journal articles in your chosen discipline and in the media.
- Become sceptical about data and understand its limitations, understand your limitations and recognise the need to consult a statistician when you meet more complex statistical problems.

Once you have achieved these objectives you should have a basic understanding of the subject referred to as statistics. In becoming proficient at the simpler statistical methods you will need to learn (and probably in this order):

*T: 0207 594 9318 and E: graham.cookson@imperial.ac.uk

- How to apply them
- What they can tell you
- Why they are useful, and
- When they are applicable

2 What is statistics?

Statistics is the science of collecting, organising, and interpreting numerical facts, which we call *data*. We are bombarded by data in our everyday lives: the Prime Minister's (un)popularity, cricket batting averages, today's average highest temperature, polls on who is getting kicked out of the Big Brother house.

Even though you may not have realized it, you have also (probably) made some statistical statements in conversation. Statements like *"I sleep for about eight hours per night on average"* and *"You are more likely to pass the exam if you start preparing earlier"* are actually statistical in nature.

Statistics is a discipline concerned with:

- Designing experiments and other data collection methods,
- Summarizing information to aid understanding,
- Drawing conclusions from data, and
- Estimating the present or predicting the future.

In making predictions, statistics uses the companion subject of probability, which models chance mathematically and enables calculations of chance in complicated cases.

3 Why study statistics?

Most people do not realise how important the subject of statistics is to their studies or future career, nor do they realise that they require, and typically use, statistical skills to interpret information that we are bombarded with almost daily by television, newspapers, books and people. You may have heard phrases implying that "statistics lie", but many of these instances occur because of mis-interpretations of information by the statistically illiterate. But why should *you* study statistics? I can think of five key reasons.

The first reason is to be able to effectively conduct research. Without the use of statistics it would be very difficult to make decisions based on the data collected from a research project. Statistics provides us with a tool to make an educated decision. It is extremely important for a researcher to know what statistics they want to use before they collect their data. Otherwise data might be collected that is uninterpretable. While you may never plan to be involved in research, it may find its way into your life. More

and more organisations are conducting internal research or are becoming part of broader research studies. This is particularly true of governments and the public sector. Finally, you may decide to do research for your dissertation that requires the application of statistical methods - better to learn now rather than then!

The second reason to study statistics is to be able to read journals. Most technical journals you will read contain some form of statistics. Without an understanding of statistics, the information will be meaningless. An understanding of basic statistics will provide you with the fundamental skills necessary to read and evaluate most research. The ability to extract meaning from journal articles and the ability to critically evaluate research from a statistical perspective are fundamental skills that will enhance your knowledge and understanding in related coursework.

The third reason is to further develop critical and analytical thinking skills. The study of statistics will serve to enhance and further develop these skills. To do well in statistics one must develop and use formal logical thinking abilities that are both high level and creative.

The fourth reason to study statistics is to be an informed consumer. Like any other tool, statistics can be used or misused. Some individuals actively lie and mislead with statistics. More often well meaning individuals unintentionally report erroneous statistical conclusions. If you know some of the basic statistical concepts, you will be in a better position to evaluate the information you have been given.

The final reason to have a working knowledge of statistics is to know when you need to hire a statistician. Most of us know enough about our cars to know when to take it to a mechanic. Usually, we don't attempt the repair ourselves because we don't want to cause any irreparable damage. Also, we try to know enough to be able to carry on an intelligible conversation with the mechanic to make sure that we don't get a whole new engine (at huge expense) when all we need is a new fuel filter (a few pounds). We should be the same way about hiring a statistician.

To summarize, the five reasons to study statistics are to be able to effectively conduct research, to be able to read and evaluate journal articles, to further develop critical thinking and analytic skills, to act as an informed consumer, and to know when you need to hire outside statistical help.

4 Helpful information

4.1 A rose by any other name

According to Romeo would still be a rose. And the same goes for statistics. Subjects such as econometrics, biostatistics or biometrics, environmetrics and epidemiology are all essentially statistical disciplines. What differs them from statistics is that the people doing them typically come from a background within the core subject (e.g. economics, biology, environmental science) and take up advanced study of statistics applied to their discipline later, rather than study a statistics undergraduate degree. And the disciplines are related solely to statistical analysis of the parent subject i.e. econometrics is the

application of statistics to economic data.

4.2 Mathematics

There is no need to have any mathematical training to do well in this course by which I mean there is no need for any undergraduate modules in algebra, probability, calculus, etc. However, if you are uncomfortable with interpreting information provided in a graph, do not know what the equation of a straight line is ($y = a + bx$), or generally feel apprehensive about this course then you may want to do a *very* basic maths refresher. The following websites will provide you with the necessary material:

- http://www.gcse.com/maths/graphs.htm
- http://www.gcse.com/maths/averages.htm
- http://www.gcse.com/maths/algebra.htm
- http://www.gcse.com/maths/equations.htm
- http://www.gcse.com/maths/algebra5.htm

For those of you who have some undergraduate mathematical training and would like a refresher so that you can access the more technical statistical textbooks, one can be found in Appendix A to these notes.

There is absolutely no point in reading or learning this material if you have no prior mathematical training. You have been warned!

4.3 Notation

For those of you who haven't studied any mathematics since you were 16 you may be unfamiliar (or can't remember) what some of the common notation employed in statistics means. So, I've outlined some of the notation and conventions used in Appendix B.

5 Statistical inference

5.1 Simple random sampling

Most of this course deals with the analytical and statistical tools used to analyse data once it has been collected. This is partly because you are not experimental scientists and as such will rarely need to conduct your own surveys and experiments, but partly because understanding statistical methods is necessary in order to understand how to design and conduct experiments and surveys in the first place. In other parts of SMN405 you will learn about survey design and interview technique that will enable you to collect your own data.

In my opinion, statistics is really about answering general questions with limited data. It's about trying to discover the truth without being able to know everything

there is to know about a subject. Statistics is the science of uncertainty. It's uncertain because we don't have the time, money or inclination (even if it were possible) to get totally accurate data on every subject in our study. There would be no need to study statistics (beyond descriptive statistics) if I could costlessly and accurately survey every member of my population under study.

A quick example will suffice to illustrate this. Suppose that you're Ken Livingstone (the London Mayor) and want to know what Londoners earn. You're interested in understanding the average as well as the distribution of income. Is it highly skewed i.e. are there vastly more poor people earning very little than the rest? There are approximately 7,500,000 people living in London. To survey the whole population of London would take a long time, cost a lot of money and involve a lot of people. Imagine it takes 10 minutes to telephone everybody in turn and you paid the enumerators the minimum wage (£5.35 per hour from October 1st 2006). Working 35 hours a week, it would take almost 700 people a year to survey them all and cost the London taxpayer almost £7 million. And that's only to collect the data. You'd then have to enter that data into a computer (at great cost) and analyse it. Given the Government's record on outsourcing computer systems were going to pour millions of pounds down the drain, be a few years older and no closer to finding out about London salaries.

But there is another way. If we randomly sample 50,000 people (i.e. if the probability of being included in the sample is equal for all individuals in the population) then our 700 enumerators could conduct the sample in a very long working day at a cost of less than £50,000. From this sample of 50,000 people we can make inferences about the underlying population. It is possible (but unlikely) that the sample contained the richest individuals in London. Equally possible is the situation where the poorest individuals were included in the survey. These possibilities exist through no fault of the procedure utilized. They are said to be due to chance.

The beauty of inferential statistics is that the amount of probable error, or likelihood of either of the above possibilities, may be specified. In this case, the possibility of either of the above extreme situations actually occurring is so remote that they may be dismissed. However, the chance that there will be some error in our estimation procedure is pretty good. Inferential statistics will allow us to specify the amount of error with statements like, "I am 95 percent sure that the estimate will be within £200 of the true value. *" We are willing to trade the risk of error and inexact information because of the savings in time, effort, and money are so great."* As a result, when we report our results we can stipulate how certain (or uncertain) we are about the probability of making these errors.

In summary, we use statistical inference to make statements about a population that interests us by analysing a random sample from that population. Each subject from the population is chosen randomly such that each subject has the same probability of being chosen at any stage during the sampling process. It is critically important, therefore, that the sampling design used to collect the data is robust and applicable to the statistical methods you wish to use.

5.2 Sampling bias

A sample is biased if some members of the population are more likely to be chosen in the sample than others. A biased sample will generally give you a wrong estimate of the quantity being estimated, such as the average London salary. For example, if your sample contains members with a higher or lower value of the quantity being estimated, the outcome will be higher or lower than the true value. A famous case of what can go wrong when using a biased sample is found in the 1936 US presidential election polls. The *Literary Digest* held a poll that forecast that Alfred M. Landon would defeat Franklin Delano Roosevelt by 57% to 43%. George Gallup, using a much smaller sample (300,000 rather than 2,000,000), predicted Roosevelt would win, and he was right. What went wrong with the *Literary Digest* poll? They had used lists of telephone and automobile owners to select their sample. In 1936 they were luxuries, so their sample consisted mainly of middle- and upper-class citizens who voted, in the majority, for Landon. The lower classes, however, voted for Roosevelt. As their sample was biased towards wealthier citizens *Literary Digest's* result was incorrect.

If we realize during our analysis that we don't have a truly random sample then we can correct this by weighting the variables in our analysis. Using statistical software packages this is very easy.

6 Hypotheses

Does coffee make me more alert? If you all drank a cup of coffee before class, would the time spent sleeping in class decrease? These questions may be answered using an experimental methodology and hypothesis testing procedures. The purpose of hypothesis testing is perhaps best illustrated by an example.

To test the effect of caffeine on alertness in people, one experimental design would divide you into two groups; one group receiving coffee with caffeine, the other coffee without caffeine (i.e. decaffeinated). The second group gets coffee without caffeine rather than nothing to drink because the effect of caffeine is the effect of interest, rather than the effect of ingesting liquids. The number of minutes that students sleep during that class would be recorded. Suppose the group, which got coffee with caffeine, sleeps less on average than the group which drank coffee without caffeine. On the basis of this evidence, we conclude that caffeine had the predicted effect - it makes you more alert.

A statistician, learning of the study, is mortified and argues that such a conclusion is unwarranted without performing a hypothesis test. The reasoning for this argument goes as follows: Suppose that caffeine really had no effect. Isn't it possible that the difference between the average alertness of the two groups was due to chance? That is, the individuals who belonged to the caffeine group had gotten a better night's sleep, were more interested in the class, etc., than the no caffeine group? If the class were divided in a different manner the differences would disappear.

The purpose of the hypothesis test is to make a rational decision between the hypotheses of real effects and chance explanations. The scientist is never able to totally

eliminate the chance explanation, but may decide that the difference between the two groups is so large that it makes the chance explanation unlikely. If this is the case, the decision would be made that the effects *are* real. A hypothesis test specifies how large the differences must be in order to make a decision that the effects are real.

At the conclusion of the experiment, then, one of two decisions will be made depending upon the size of the differences between the caffeine and no caffeine groups. The decision will either be that caffeine has an effect, making people more alert, or that chance factors (the composition of the group) could explain the result. The purpose of the hypothesis test is to eliminate false scientific conclusions as much as possible. Hypothesis tests are, therefore, about showing how *statistically significant* the difference is between the groups. The results are presented in terms of critical values and p-values which we will cover later in the course. For the moment remember that whenever you see these things presented in papers or research, the author is demonstrating how significant (or not) the findings of their research are. Invariably, results published in academic journals are significant because few people are interested in experiments that have failed.

Hypothesis tests are rampant in the experimental sciences because they are genuinely interested in the results. They are also common in the social sciences but are often hidden. For instance, when we conduct our linear regression we will investigate the relationship between a dependent or response variable (e.g. income) with some explanatory or independent variable (e.g. level of education) in an attempt to explain one by the other. In doing so, our statistical package (SPSS) will report a coefficient for each of our variables (the amount by which our dependent variable income is increased when we increase our explanatory variable "years of education" by one). Among the results presented by the software package will be a hypothesis test on whether the coefficient is significantly different from zero i.e. whether we get our results just by chance. We will discuss more about this at the time. Personally, I think it's tacky to state a formal "hypothesis" especially in the social sciences partly because, in most research, the most interesting finding was not anything hypothesized about beforehand. There's a problem with the whole mode of research that focuses on "rejecting hypotheses" using statistical significance which you will learn more about during this course, but it is important to understand where hypotheses tests lurk in your work.

Hypothesis testing rests upon the notion of falsifiability first postulated by Karl Popper. Essentially, we can never prove something to be true. It will only remain true until we prove otherwise because we will never know everything (be omniscient). We can demonstrate something to be false and know that once proven to be false it will always remain false i.e. it can never become true again. Based on this fact and the notion of symmetry if we prove the opposite of what we're interested in to be false, then what we're interested in must be true. An example will make this clear. I am interesting in proving that caffeine has an effect on alertness, we call this the alternative hypothesis. In order to prove this I should disprove (falsify) the hypothesis that caffeine *does not* make people more alert. This is called the null hypothesis.

Don't get too hung up on hypothesis testing and significance, as we will later come to discover during our studies that the null hypothesis can nearly always be proven to

be false if we have a big enough sample size. Additionally, there is a huge difference between statistical significance and practical importance i.e. we may not be interested in something even though it is statistically significant.