Applied Statistics for Accounting

David Ramsey

Room: B2-026
e-mail: david.ramsey@ul.ie
website: www3.ul.ie/ramsey

January 17, 2012

Recommended Reading

The lecture notes, tutorial sheets, laboratory notes and supplementary material can be found on my website www3.ul.ie/ramsey (just click on "lectures" and then "statistics").

Prime Texts: Available in library

For lectures Berenson, M. L., 2002, Basic business statistics: concepts and applications, 8th Edition, Prentice Hall (at 519.502465).

For Computer Labs Field A. P. Discovering statistics using SPSS (and sex and drugs and rock 'n' roll), 3rd Edition (at 519.50285536)

Other Relevant Texts

Daniel, W. W., 1992, Business statistics: for management and economics, 6th Edition, Houghton Mifflin.

Sincich, T. 1996, Business statistics by example, 5th Edition, Prentice Hall

McClave, J. T. 1998, First course in business statistics, 7th Edition, Prentice Hall.

Neter, J. 2005, Fundamental statistics for business and economics, 4th Edition.

Other Relevant Texts

Ross, S. M., 1987, Introduction to probability and statistics for engineers and scientists, Wiley.

Walpole, R. E., 1998, Probability and statistics for engineers and scientists, 6th Edition, Prentice Hall

Stuart, M. 2003, An introduction to statistical analysis for business and industry: a problem solving approach, Arnold.

Outline of Course

1. Data Collection
2. Presentation of data
3. Probability
4. Regression and Prediction
5. Statistical Inference

1. Data Collection - 1.1 Types of Data

The methods of statistical analysis available depend on the type of data we are dealing with. Data can be broadly classified into two types:

1. Quantitative Data: Numeric data that indicates how much or how many: e.g. height, mass, number of children.
2. Qualitative Data: Normally classifications or groupings, e.g. gender, university department, social class. Note that classifications may use numeric labels. However, such variables are still qualitative, e.g. Department: 1-Maths, 2- Equine studies, 3-Sociology (i.e. the numbers here are not counted or measured).

Types of Quantitative Data

Quantitative data may be further split into

1. **Continuous** variables. Such variables can take any value in a certain range. They are usually **measured** according to some scale, e.g. age, height, mass.
2. **Discrete** variables. Such variables take values from a set that can be listed (commonly integer values). Such variables are often **counted**, e.g. number of children, number of subjects passed at leaving certificate

Types of Quantitative Data

Note that continuous variables are only measured to a given accuracy.

e.g. Age is normally given to the closest year. However, in theory it could be measured much more accurately.

When a discrete variable takes a very large number of values , e.g. the number of individuals employed by a firm, it may be treated for practical purposes as a continuous variable.

Types of Qualitative Data

Qualitative data may be further split into

1. **Nominal** classifications. Such data is defined by a pure classification, in which the order of the classes has no practical interpretation, e.g. Department: 1-Maths, 2-Equine studies, 3-Sociology.
2. **Ordinal** classifications. The order of the classification is important, e.g. 1. Non-smoker, 2. Light Smoker, 3. Heavy Smoker, i.e. the higher a number the more an individual smokes.

It is important to distinguish between quantitative variables and classifications using numeric labels, e.g. the mean of a discrete variable has a sensible interpretation, but not the mean of a nominal, or even ordinal, variable.

1.2 Sampling

Suppose we wish to study a particular group of individuals, e.g. Irish teenagers, Irish voters, or European students.

The **population** in a study is the entire group of individuals that we wish to investigate (as above).

Since it is impractical to observe all the individuals in a population, we observe a **sample** of the population in question.

A **unit** is any individual member of the population.

Sampling

A **sampling frame** is a list of members of a population. It may be used to choose a sample.

A sampling frame may be incomplete or inaccurate. For example, the Irish electoral register will be a complete sampling frame for the population of eligible voters in Ireland. However, it will not be a complete sampling frame for the population of adult Irish residents.

Choice of an inappropriate sampling frame may well lead to **systematic errors** in estimates obtained from sampling (bias).

If I tried to measure the mean IQ (or height) of the whole Irish population by just observing a sample of Irish students, I would tend to overestimate this population mean.

Example

Suppose we wish to do a study of recent immigrants to Ireland and classify their occupation, e.g. managerial, professional, retail, unemployed.

We may use the register of PPS numbers given to non-nationals in the past 3 years as a sampling frame.

Such a sampling frame will not be completely accurate as some of these immigrants will have already left Ireland and some immigrants will not have registered.

A sample from this population is the set of individuals we choose to observe.

The main variable observed in this study is the class of occupation.

1.3 Parameters and Statistics

A **parameter** is a numeric characterisation of the **population** (its value is usually unknown).

e.g. 13% of all recent immigrants have professional occupations.

A **statistic** is a numeric characterisation of the **sample** (observed).

12.4% of individuals in our sample have professional occupations.

Statistics can be used to estimate the parameters of the population, e.g. here we would estimate that 12.4% of all recent immigrants have professional occupations.

1.4. Accuracy of Estimates - Bias and Precision

As described above, if we use an inappropriate sampling frame, our estimates of parameters may have a systematic **bias**.

Bias that results from our method of sampling is called **sampling bias**. Other sources of bias exist (see later).

Also, we have random errors depending on the sample actually observed. This determines the **precision** of a study.

Consider the hypothetical situation in which we have a large number of small samples. The mean heights in these samples will be rather variable, i.e. a small sample leads to low precision.

Bias and Precision

Suppose we had many large samples. The mean heights in these samples will be similar to each other (and if the sampling frame is appropriate will also be similar to the population mean).

In this case we have low bias and high precision (ideal).

Now suppose we observe the heights of Irish students in order to estimate the mean height of the population of Irish adults. When we have a large number of small samples, the sample means will be rather variable and tend to be larger than the mean height of all Irish adults (large bias and low precision).

Increasing the size of these samples would increase the precision (the sample means would be more similar), but the bias is unaffected (the mean height will still tend to be overestimated as the sampling frame is inappropriate).

Bias and Precision

Hence, increasing the sample size will increase the precision of a study.

However, increasing sample size leaves the bias unaffected.

Moral: Results may be misleading, even when we have large sample sizes, as inappropriate methods for choosing samples may lead to systematic bias.

Non-sampling Bias

This is a form of bias that occurs when certain groups of individuals have a tendency to give inaccurate responses or not give an answer.

For example, it was noticed that UK political polls systematically underestimated the support of the Conservative party in the 80s and 90s.

This may well have been due to the fact that Conservative supporters were more likely to hide their preferences than supporters of other parties.

Non-sampling Bias

The wording of a questionnaire, who the interviewer is and what the interviewee perceives to be the "right" answer in a given situation may also lead to bias.

For example, suppose a questionnaire is carried out on how willing people are to pay extra for "ecologically friendly goods".

Such a survey will tend to overestimate the proportion of individuals willing to pay extra, as it is politically correct to express such a willingness.

Suppose in survey I there are the options a) not willing to pay more, b) willing to pay 10% more. Suppose in survey II, option b) is replaced by "willing to pay 20% more". It is likely that survey II would indicate that on average people are willing to pay more for "ecologically friendly goods" (this is also an example of why you should not calculate a mean for data categorised in such a way).

1.5 Types of Sampling - 1. Simple Random Sampling

A sample is a simple random sample (SRS) if the probability of picking any collection of n units from the sampling frame as a sample does not depend on the collection of units, i.e. individuals are picked completely "at random" from the population.

Advantages: 1. Simple, 2. If the sampling frame is appropriate, then there is no sampling bias.

Disadvantages: If the population can be divided into units called strata, which have different characteristics, more precision can be obtained by using Stratified Random Sampling (see below).

2. Stratified Random Sampling

The population is divided into strata according to various characteristics, e.g. sex, occupation, age.

A simple random sample is taken from each stratum, so that the proportions of individuals taken from each stratum equals the proportion of individuals in the population as a whole within that stratum.

Stratified Random Sampling

Advantages: 1. If the sampling frame is appropriate and the proportions of individuals in each stratum known, then there is no sampling bias.

2. If the strata are different in their characteristics, this method leads to increased precision.

Disadvantages: 1. Complexity, increased time required to collect samples.

2. The proportions of individuals in each stratum should be known accurately for the method to be effective.

3. Systematic sampling

Suppose we wish to sample a fraction $1/k$ of the population. One individual is chosen from the first k units listed in the sampling frame and each k-th individual is chosen from then onwards, i.e. if I want to choose 1% of the population, I may choose number 53 from the first 100, then numbers 153, 253,... are also picked.

Advantages and Disadvantages: Same as simple random strategies, except that it is more difficult to control the precision if the list is not ordered randomly with respect to the variables of interest.

e.g. If one wanted to study earnings in a given population, it would be reasonable to choose subjects from an alphabetic list using systematic sampling.
However, it would not be reasonable to choose the subjects in such a way from a list of highest earners (i.e. with the highest earners at the top).

4. Convenience Sampling

The sample is chosen so as to minimise the costs of obtaining a sample, i.e. questioning friends and colleagues, inviting volunteers etc.

Advantages: 1. Simplicity, 2. Low sampling costs.

Disadvantages: 1. The sample may not be representative (i.e. there is sampling bias). For example, when we invite responses on some political issue (e.g. immigration), we are more likely to get responses only from those who have rather extreme views on that issue. If we invite respondents to a study on the mathematical abilities of students, mathematically gifted individuals are more likely to volunteer than non-gifted individuals.

5. Expert Sampling

An expert chooses a sample that he feels is representative of the population as a whole.

Advantages: 1. If the method employed by the expert is appropriate, this may well result in increased precision.

Disadvantages: 1. An inappropriate method will lead to bias.

Probability Sampling

These methods can be more broadly defined into probability and non-probability sampling.

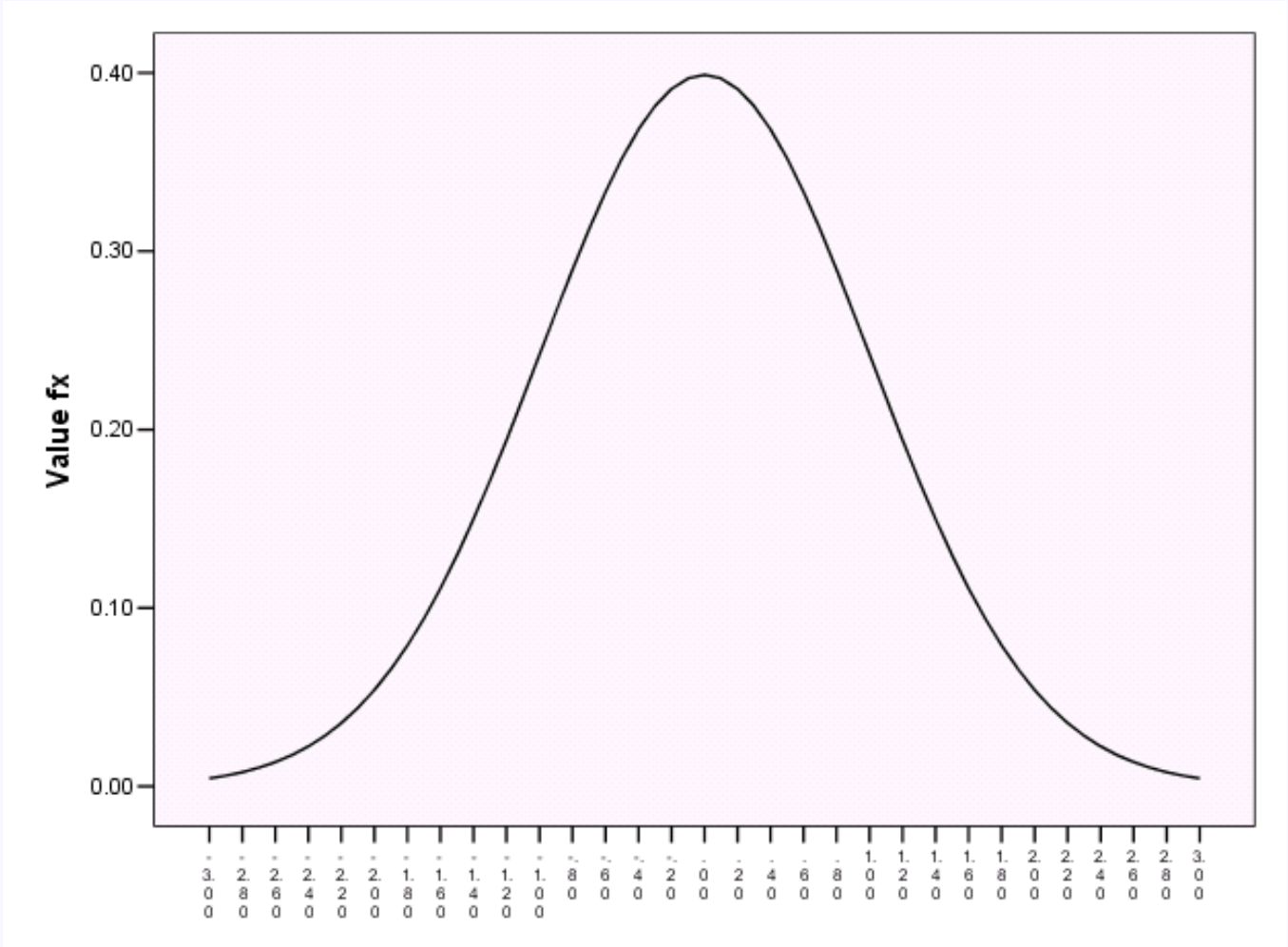
Probability sampling occurs when the probability that an individual is picked to be in a sample does not depend on the individual.

The first 3 methods described above are forms of probability sampling. **Note**: Stratified random sampling is a form of probability sampling as long as the correct proportions of individuals in the various strata are known.

Such methods are advantageous as much can be said about the distribution of the statistics obtained from such a sample (see below).

1.6 Sampling Distribution

Suppose we observed a large number of samples of size n of the height of Irish adults. The distribution of the sample means would look like the distribution below:



Sampling Distribution

If an appropriate sampling frame were used, then this sampling distribution would be centred around the true mean height of the population.

The larger the sample size, the more concentrated the distribution around the true population mean.

1.7 Critical analysis of a poll

To assess the credibility of a study, we should ask the following questions.

1. Who carried out the survey? (This person/organisation might have a hidden agenda).
2. What methods were used in choosing a sample? How large is the sample? (sampling bias, precision)
3. How was the information gained? This may well affect the validity of the data obtained. The reaction of an interviewee to the interviewer is important.
4. How were the questions formulated? (see example on buying environmentally friendly goods).
5. What was the response rate? A low response rate could lead to bias, as non-responders may well differ from respondents.