A COURSE ON SAMPLE SURVEYS FOR STATISTICS STUDENTS

Flavia Jolliffe, University of Greenwich, UK

Statisticians should be involved at all stages of sample surveys and courses on surveys need to reflect this by covering both theoretical and practical aspects. Teaching methods could include some hands-on experience, directed reading, and use of software designed for teaching or professional use, as well as more traditional lecturing. Suggestions are given for a course of about fifty hours.

INTRODUCTION

Courses on sample surveys come in a variety of forms and are delivered to groups with a variety of backgrounds. At one extreme there is the one session input to a longer course given to nonspecialists and at the other there is a full year's course of several hours a week given at postgraduate level to statistics specialists and short intensive courses designed for professionals, often concentrating on only one aspect of surveys. Somewhere between these are courses of about fifty hours in length given to undergraduates who are taking single or joint honours degrees in statistics and who have already studied a basic course in probability and statistics. It is a course of this nature which is considered here.

Statisticians should, of course, be involved at all stages of sample surveys and courses on surveys need to cover all of - methods of collecting information, questionnaire design, sampling theory, non-response and other non-sampling errors, coding and analysis, and report writing. The amount of emphasis given to each will depend on such factors as the general aims of the course, the interests and background of the students, and the time and resources available. It will also, unfortunately, be determined to some extent by assessment methods. Not only is this because students tend to value only what is assessed (Garfield 1995) but often there are also constraints imposed by the degree scheme, for example the length of an examination and the proportion of assessment which comes from coursework, and it can be difficult to change these. It can in fact sometimes be a challenge to convince students that it is important to take a holistic approach to survey design and analysis.

Many of the skills needed for the successful implementation of a survey are essentially practical and are best learned by doing. Some can be undertaken by students but in many other cases students will have to be content with second-hand experience either by reading or by contact with practitioners. Courses and books on sample surveys designed for statisticians have traditionally tended to concentrate on sampling theory, dealing mainly with estimation and one variable at a time, for example Barnett (1991), Hansen et al (1953). As sampling is one of the more statistical aspects of surveys it rightly continues to be a main component of courses on surveys, but use of computers has resulted in a move away from tedious algebra, made it easier to explore concepts, and enabled more complex issues to be considered than was previously possible.

A suitable plan for a course where the emphasis is on breadth of coverage rather than mathematical depth in a few topics, as advocated in Fesco et al (1996), is to start with two or three weeks on general issues relating to the stages of a survey, looking in more detail at methods of collecting information, then to spend six to eight weeks on sampling, perhaps including some time on non-response and other non-sampling errors within this block. The course could end with two to three weeks on more advanced analysis and a session drawing everything together with a discussion on report writing. In line with current thinking (Moore 1997), a variety of activities and teaching methods would be used.

THE FIRST FEW WEEKS

In any course the first few weeks are perceived by students as indicative of the nature of the course as a whole and of the assessment tasks they will be set. Sadly, many statistics students tend to think that they are not doing statistics if no formulae or calculations are involved and feel that writing words or thinking are definitely not statistics. Thus the lecturer’s attempts to stimulate an interest in surveys and make the course enjoyable by starting with an overview of surveys and methods of collecting information and questionnaire design runs the risk of achieving the opposite effect to that intended, especially as this material is not dependent on statistical knowledge.

However, launching straight into sampling notation and theory without some background discussion can convey an impression that, as far as statisticians are concerned, sampling is the most important topic in surveys. In outlining the course the lecturer might give a flavour of later sections by giving summaries of one or two calculations of the type students would be expected to perform, and perhaps interpret some output from a categorical data analysis. It is also important to indicate how long will be spent on each section and the teaching methods to be used. Indicating the types of assessment task and their importance can reassure students (or not) as to whether they think they can do well on the course.

After giving an overview of the course the lecturer could talk round headings relating to survey design, and data collection and analysis, stressing the interdependence of topics. The lecturer would need to give references to further reading and make clear the extent to which it is important for students to do this. It is appropriate to mention ethical issues here and distributing copies of actual codes of ethics indicates their importance. It is not good use of time, and can be incredibly boring, to talk at length about such things as the advantages and disadvantages of using interviewers rather than self-administered questionnaires, though getting the students to generate ideas can be very rewarding. Sometimes this is more successful if students work in small groups initially and share their ideas with the whole group at a later stage. Inviting an enthusiastic practitioner in to talk to the group is another possibility.

Questionnaire design is fairly central to any discussion of surveys and statisticians giving advice on a survey nearly always find themselves commenting on the design of the questionnaire or explaining the concept of coding or discussing what estimation or analysis is possible with the questions asked. Questionnaire design and coding are practical skills and are ideally learnt by doing. The lecturer’s role here is mainly to give examples and to help, encourage, and give constructive criticism while students perform tasks, but students might also be given information about some of the key points of questionnaire design, for example by reference to the workbook developed by Wilson and McClean (1994). It is far easier to see what is wrong with someone else’s questionnaire than to design one from scratch and a useful exercise which students enjoy is to get them to give critical evaluations of questionnaires which have been used in surveys. It does not take long to build up a suitable collection. Another possibility is to have students explore the CASS (Centre for Applied Social Surveys) question bank on the World Wide Web (http://www.scpr.ac.uk /cass) perhaps as an assignment. The question bank reproduces questionnaires which have been used in several major UK social surveys and provides backup material.

THE SECTION ON SAMPLING

In the kind of course under discussion it is envisaged that about half the course, say six to eight weeks, would be spent on sampling and related issues. Derivation of sampling formulae is more properly in the realm of the statistician than some of the other topics suggested for the course, but it is a fairly common experience that the majority of students do not find this very exciting and only like sampling when they have substituted numbers in formulae correctly. In any case there is an argument that those who are going to be practitioners do not need a great deal of exposure to background theory so that the emphasis could shift to understanding and appropriate use.

The initial lecture on sampling could give an overview of those methods of sampling which will be covered, typically simple random, stratified, systematic, cluster and quota, say what population quantities are to be estimated, and set up notation. This last is particularly important as the notation used in sampling from finite populations tends to differ from that used when considering theoretical populations and there is no unique standard.

It is important to try to get students to appreciate what is meant by a sampling distribution, why sampling is useful, and why some estimators and some sampling schemes are better than others. Many suggestions have been made as to how to do this. Classroom based activities include producing a complete sampling distribution by “hand” for a small finite population (Levy and Lemeshow 1991) and finding the average area of a rectangle given a page printed with rectangles of varying sizes (Scheaffer 1997). The use of software for exploration of sampling methods makes the task less laborious though it is questionable whether students really appreciate the process if they have not first done a small exercise by hand. One possibility is to use Minitab to take samples from a small population of choice. With values of a single variable and some identifying information estimation of a mean and a proportion under simple random, stratified, systematic and cluster sampling can be investigated. Another possibility is to use packages which mimic a real survey situation such as the STEPS (Statistical Education through Problem Solving) modules “Unfinished business” and “Market testing Sunjoy” designed for business students or StatVillage (Schwarz 1997). In the STEPS modules (http://www.stats.gla.ac.uk/steps) there is no control over which individual units are selected and summaries of samples are obtained almost automatically. In StatVillage (http://www.stat.ncsu.edu/info/jse/v5n2/schwarz.supp/

index.html) units are selected by clicking on dwellings shown on a map and this produces a data file of results from the 1991 Canadian Census ready to download for analysis.

Exciting though many of these activities are, one problem is that students might well not see the point of them and some will query as to whether they are important as regards assessment. Another problem is that some depend very much on adequate computing provision and a session can be ruined if it takes students fifteen minutes or more to access a facility or if the system crashes. Both problems can be overcome to some extent by setting students exercises as assessed work to do in their own time. This ensures they attempt them and hopefully they would learn something in the process for it is possible to work through even well designed exercises mindlessly.

Students should be told expressions for estimating common parameters such as the mean of a population and the proportion having a characteristic and for estimating the standard errors of estimators under common sampling schemes, and how to estimate sample size. All of this should be illustrated with numerical examples which can be done by hand. Scheaffer et al (1990) is a suitable text for this section. Non-sampling errors should be covered as well, particularly non-response and how to reduce it and deal with it in analysis. The use of deff in estimating sampling errors when the sampling scheme is complex and quota sampling are important topics because they are used in practice. If there is time students might be introduced to software for calculation of sampling errors (Lepkowski and Bowles 1996) but the set-up time of learning the software has to be balanced against the expected short-term and long-term benefits of using it.

THE END OF THE COURSE

The statistical section of the course should also include something on analysis of surveys. Analysis can stop at tabulations and diagrams, but a course for statisticians might also usefully include something on methods of categorical data analysis (Fesco et al 1996) such as loglinear modelling. Here use of software is essential but can be quickly learned in a Windows environment. Lectures would concentrate on general ideas and interpretation of output.

Last, but not least, the course would not be complete without some mention of how to report on the design and analysis of studies. This is a good way to draw the different topics covered in the course together. Showing students reports of actual surveys is helpful here and the ready availability of data sets in computer readable form makes it feasible for students to analyse professional surveys and report on the analysis. This is admittedly second best to some extent but the time and organisation needed to do even a small survey should not be underestimated and a survey undertaken under a university's name needs to be good. Some of the aspects which have not been covered in the course could be mentioned, for example the effect of proceeding as if a complex sampling method were simple random, qualitative research, or the difference between model-based and design-based inference.

REFERENCES

Barnett, V. (1991). Sample survey principles and methods. Edward Arnold.

R.Fesco, W.D.Kalsbeek, S.L.Lohr, R.L.Scheaffer, F.J.Scheuren, E.A.Stasny (1996) Teaching survey sampling. American Statistician, 50,4,328-40.

Garfield, J. (1995). How students learn statistics. *International Statistical Review, 63(1),* 25-34.

Hansen, M. H, Hurwitz, W.N., and. Madow W. G. (1953). Sample survey methods and theory. Wiley.

Lepkowski, J. and Bowles, J. (1996). Sampling error software for personal computers. *The Survey Statistician, 35*, 10-17.

Levy, P. S. and Lemeshow, S. (1991). Sampling of populations: Methods and applications. Wiley.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 65(2),* 123-37.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1990). Elementary survey sampling. 4th ed. PWS-Kent Publishing Co.

Scheaffer, R. L. (1997). Discovery of sampling concepts through activities. Bulletin of the International Statistical Institute. Proceedings of the 51st session of the ISI, Tome LVII, Book 1, 421-4.

Schwarz, C. J. (1997). StatVillage: An on-line, WWW-accessible, hypothetical city based on real data for use in an introductory class in survey sampling. *Journal of Statistical Education, 5(2).* http://www2.ncsu.edu/ncsu/pams/stat/info/jsev5n2/schwarz.html

Wilson, N., and McClean, S. (1994). Questionnaire design: A practical introduction. University of Ulster.