Teaching Sampling Statistics:

Experience from the MSc in Official Statistics Programme

Brown, James J
University of Southampton, School of Social Sciences
University Road
Southampton SO17 1BJ, UK
E-mail: jjb1@soton.ac.uk

Introduction

The MSc in Official Statistics Programme was started in September 1999 as a unique (in the UK) collaboration between the University of Southampton and the Office for National Statistics (ONS). The aim of the programme was to offer academic training that would enhance the skills of those working within the UK Government Statistical Service, of which the ONS is a major component. As such the programme was set-up with a much greater emphasis on surveys (their design and application) than a standard Statistics MSc in the UK would usually have. There is also a strong applied element to the programme, although it is recognized that a certain amount of foundation theory is also required. During the last eight years around 10 students have applied each year to study for the full MSc with around 80 individuals per year have opted to take individual modules from the programme appropriate to their particular training needs.

In the next section we will look at the overall structure of the programme and how the sampling component fits into that structure. We will then consider the sampling component in detail looking at the modules that make-up the component. We will consider how the collaboration between the University and ONS enhances the learning and teaching on these modules and how they fit together to give the student an overall knowledge of sampling statistics.

Structure of the Programme

The MSc in Official Statistics Programme (MOffStat) is made-up of 23 modules covering a wide range of topics that a Government Statistician would wish to know about. Each module is designed to be taught intensively over a period of three to five days with students doing an additional three days to complete assessments. Students can choose to register for the MSc or just choose to study modules on a one-off basis. Those doing individual modules can choose to do the assessment for credit points or simply attend the modules for professional development. Those that take the credit can then at a later date use that credit towards gaining the full MSc if they wish. Students on the full MSc must complete 16 modules from the 23 available in up to 4.5 years of registration. Under the current structure (this is being reviewed by the University in consultation with ONS) 11 modules are compulsory with students choosing the remaining five as options. In any one year 16 or 17 modules are offered by the University with optional modules being rotated from year to year.

The modules on the programme tend to be grouped into sets covering similar areas. There are five modules covering regression analysis. The two compulsory modules cover linear regression and standard GLMs. Students can then do optional modules covering complex survey data analysis, multilevel modeling, and longitudinal data analysis. There are two modules on time series; the compulsory one covering standard seasonal adjustment and ARIMA modeling using the X12 software and an optional module looking at more advanced analysis techniques. Likewise there are two modules covering demographic methods; the compulsory one covering population structures, fertility, mortality, and population estimates with an optional module covering the use of multiple decrement life tables and population projections. There are a set of modules that cover more specific areas of interest to a Government Statistician. The compulsory 'elements of official statistics' module covers the structure of Government Statistics in the UK putting the current position in its historical and international context. There are also compulsory modules on 'index numbers' and 'evaluation and monitoring' with optional modules covering 'national accounts', 'disclosure control', and 'statistical computing'. An overview of the programme can be got by referring to the MOffStat website

(http://www.socstats.soton.ac.uk/moffstat/).

In particular, the Programme Specification and marketing booklet give details of the individual modules.

The final two sets cover the survey component of MOffStat and these will be covered in more detail in the following sections. In the world of sampling and survey statistics there tends to be a separation between the statistical theory behind survey design and estimation and the practicalities of actually doing a survey – often referred to as survey methods. Cochran (1977) is perhaps still the classic text on sampling theory but it contains nothing on questionnaire design and only one chapter at the end deals with survey errors coming from issues such as non-response and interviewer effects. On the other side a book such as Biemer and Lyberg (2003) covers the errors that occur through-out the survey process but does not show the reader why the sample mean is an unbiased estimate of the population mean under simple random sampling. This is not a criticism of either book (or those like them) but it does demonstrate that traditionally the two sides of surveys and sampling have tended to be quite separate and this is reflected in the underlying structure of the MOffStat modules. Of the eight sampling modules, five cover the area of sampling theory dealing with sample design and estimation with the remaining three covering survey methods. We will first look in detail and the sampling theory modules, and then consider the survey methods modules, before looking at how they integrate together.

The Sampling Component – Sampling Theory

The five sampling theory modules cover what we consider to be foundation theory in two compulsory modules and then offer students an additional three optional modules. The first compulsory module is a classic look at sampling theory from a design-based perspective and while Cochran (1977) is not a core text for this module the aim is to familiarize students with the main theory covered in the book. Students often prefer Lohr (1999) as a companion to this module finding it more readable. The module is taught by a member of the University using a traditional mix of lectures and workshops spread over five days. Significant amounts of time are devoted to students doing examples that demonstrate the basics of design-based sample inference under a variety of standard sample designs. Although the examples are often not realistic (samples of size two from populations of size four) their role is to demonstrate the concept of the randomization distribution and its role in survey inference across a range of sample designs. The module is assessed by an open-book exam (we do not expect students working in a professional environment to learn large numbers of variance formulae) to see if students have understood and can apply the basic design-based theory to solve sampling problems.

The second compulsory module looks in more detail at survey estimation and inference. In particular it extends the design-based framework to include the model-assisted approach (see Särndal, Swensson and Wretman, 1992). It focuses on GREG estimation (generalized regression estimation) as this is used in the UK by ONS on business surveys and will form the basis of social survey estimation in the future. This is followed by contrasting the approach with a full model-based approach for standard stratified, ratio, and regression situations. The module is taught by a member of the University and teaching is focused in the first three days. There are lectures mixed in with a set of exercises that are designed to demonstrate the differences between the approaches to survey inference. The assessment for the second module is via coursework. The main component involves the student first designing a survey and selecting a sample from a population; they are then give the sample values and asked to produce a set of estimates with appropriate standard errors. The aim is not to say which approach to inference is best but to get the student to make a decision at both the design and estimation stage, which they are able to justify.

The role of the two compulsory modules is to provide a solid foundation in the basics of sampling theory that students can then choose to build upon by taking some (or all) of the three optional modules. The first of the three optional modules deals with aspects of advanced sample design. How to design rotating surveys, multi-phase samples, and particular aspects of business survey design. This module is taught by members of the University with practical sample-design experience and is augmented by input from ONS on particular sample-design projects. The module is assessed by coursework and students are asked to review a sample design and describe how they would go about re-designing it. In particular, what analysis would they do and what would be the potential benefits. This module aims to build on the foundation theory and demonstrate how it is applied when designing real surveys. The second of the three optional modules deals with non-response in surveys. As this module is part of the sampling theory set it looks at estimation techniques to deal with non-response. Imputation techniques are covered for item non-response and weighting techniques for unit non-response. The weighting methods particularly build on the content of the second compulsory module by covering the use of calibration methods (standard for UK social surveys) on multiple auxiliaries. Again this module is taught by members of the University with practical experience and the coursework assessment also includes a section asking students to review critically the non-response methods of a current survey. The final of the three compulsory modules deals with the specific topic of small area estimation. The module is taught by a member of the University who is a specialist in the methodological developments that have taken place in this area. This is a slightly more theory based module than the other two options focusing on the development of model-based approaches to the problem over the past few years. Given its focus, the module is assessed by examination but as with the first sampling module the exam is open-book. It is not about 'wrote learning' of theory but about being able to utilize and manipulate the theory.

The aim of this set of modules is to give all the students a good grounding in the background theory behind sampling designs and inference. This is then extended into the more practical aspects of actually carrying-out sample design and estimation through the optional modules. We feel that the latter is not really possible without the former although it is a careful balancing act to make sure that the compulsory modules are not too abstract as that tends to discourage students from choosing the more 'practical' optional modules.

The Sampling Component – Survey Methods

The three survey methods modules cover an overview of the whole survey process followed by questionnaire design and data collection as two compulsory modules, with an optional module on the measurement and reporting of overall survey quality. All three modules are taught by members of ONS who have experience in the relevant areas. The style of these modules is somewhat different in that they often utilize several lecturers doing sessions on specific components of the modules. This is again a trade-off to get the right balance between the continuity of a 'small number' (one or two) of lecturers versus the specialist expertise in specific areas from many different lecturers. In more recent years there has been a move towards a single lecturer giving the bulk of the lectures with a few 'guest lecturers' in specific areas and this seems to give the right balance. The teaching on all three modules is focused into three days and lectures are interspersed with group work that often uses case studies based on real surveys.

As with the first sampling theory module, the first module in survey methods aims to lay the foundation by covering the whole survey process from the formulation of a research topic to the production of a final report. The purpose is not to cover topics in depth but to give students an understanding of how complex the actual survey process is; something that students do not appreciate so easily as survey methods modules tend to not be full of formulae. The module is assessed by a closed-book exam. Students tackle essay type questions on a range of topics from data collection methods to the practical implications of types of sample design such as clustering to manage interviewer workloads to the use of other research methods apart from standard surveys. The second module (also compulsory) focuses on the specific area of data collection. It covers the design of the questionnaire from the formulation of the research questions to the definition of concepts to the creation of the questions and their testing. This is linked to a wide range of data collection methods with an emphasis on those commonly used in UK Government surveys (CAPI and CATI for social surveys, postal for business surveys) as the data collection methodology and questionnaire design of a survey cannot be separated. The module is assessed by coursework and students are asked to design the data collection methodology and questionnaire for a research problem that is based on a real-life example. The third module on survey methods is optional and covers the area of survey data quality. (We are currently reviewing whether to broaden this to cover data quality for administrative sources as well given their growing importance within Government Statistics.) In a sense, this module deals with what happens after everything else has taken place. You designed and selected your sample, you design and tested your questionnaire, you undertook your data collection, and you applied an estimation strategy to come up with your outputs. How do you now tell users whether what you have produced is actually any good? Therefore, it reviews where and how errors occur during the survey process and then how these can be presented through quality reports and quality profiles. For the coursework students are required to produce a quality report for a data collection exercise that is undertaken within their work area.

Fitting the Two Components Together – The Student Experience

The teaching of these two components in complete isolation from each other is not possible. You cannot talk about designing a sample from the sampling theory perspective or talk about planning and undertaking a sample from the survey methods perspective without talking about the population you wish to sample and its encapsulation in a sample frame. You cannot separate decisions on the data collection strategy from design choices. If you are using CAPI it will almost definitely imply a geographically clustered design while CAPI will generally not. This is reflected in the fact that the first modules in each component contain a small element of over-lap, with both discussing the basic definition of sampling and non-sampling errors along with stratification and clustering. The overlap is intentional as it should help students realize that both components are equally important when under-taking a real survey.

Another way is to see how an issue like non-response is dealt with by the modules. The first compulsory module in each component gives students the basic foundation in sample design and survey methods. The second compulsory module in the survey methods component looks at effective questionnaire design and data collection to minimize the amount of non-response at both the item and unit level. The second compulsory module in the sampling theory component expands the estimation framework to explicitly include ‘models’ and this is then utilized in the optional module on non-response. This optional module looks at estimation techniques (and the assumptions they are based on) to adjust for the non-response that will inevitably occur. The optional module in the survey methods component then considers how the impact of the non-response and its adjustments will have affected the quality of the data that has been produced. All these modules in combination allow the student see how they attempt to avoid non-response in the first place, how they correct for it when it happens, and finally how it impacts on their data quality.

Inevitably students will not choose all the optional modules from both components. They often choose the options from one or other. However, by having compulsory modules covering theory and methods students should get an appreciation of the importance of the entire survey process when using a sample to measure a population. The link between the University and ONS in the provision of the teaching also provides the student with a balance between the need to understand sampling theory and the realities of carrying out survey research in the real world. By putting both together we aim to ensure that the learning of theory is integrated with its practical application and not seen by students as an irrelevant and abstract mathematical exercise.

REFERENCES

Biemer, P. and Lyberg, L. (2003) Introduction to Survey Quality. Wiley: New Jersey.

Cochran, W. (1977) Sampling Techniques (3rd edition). Wiley: New York.

Lohr, S. (1999) Sampling: Design and Analysis. Duxbury Press: California

Särndal, C. E., Swensson, B. and Wretman, J. (1992) Model Assisted Survey Sampling. Springer: New York.