Teaching environment for survey sampling based on a textbook and its Web extension

Risto Lehtonen
University of Jyväskylä, Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FIN-40014 University of Jyväskylä, Finland
risto.lehtonen@maths.jyu.fi

1. Introduction

Development of information technology in an information society has invoked educators in universities and elsewhere to modify the prevailing teaching methods. At the same time calls for restructuring the way students learn come from a variety of institutions. Web-based learning has been one possibility to restructure the way of learning. (Brandon 1996; Dodge 1995.). Educators agree that we must help students learn to solve problems and think independently. Web tools provide a potential option for active learning, since these tools are believed to approach education more in a learner-focused way than in a teacher-focused way (Brandon 1996). Also a potential advantage of Web-based learning is that it supports the "learning by doing" approach, which is highly appreciated in modern pedagogical thinking (e.g. Smith 1998).

A "hybrid" solution constituting of a combination of a paper-printed textbook and its Web extension that supports the use of the textbook in teaching and learning provides a challenging option. Our Web application "VLISS-Virtual Laboratory in Survey Sampling" has been prepared to support the use of the textbook of Lehtonen and Pahkinen, "Practical Methods for Design and Analysis of Complex Surveys", in university-level teaching of basic and more advanced survey sampling. The textbook itself is not included in the VLISS application (only the list of the contents of the book is visible for the user). We believe that at least at the current stage of educational practices and cultures, a paper-printed textbook still provides a very powerful educational tool.

The textbook and its Web extension are designed to serve learning purposes in survey sampling. Methods of survey sampling are routinely used in official statistics, and the methods are increasingly used in empirical research in different disciplines. Teaching of survey sampling usually takes place in university departments of statistics. The potential audience of the VLISS application covers university students, teachers and instructors (especially in statistical science), researchers in various disciplines dealing with empirical research, and junior and senior statisticians working in official statistics, research institutes and business firms.

VLISS can be accessed via Internet (Web reference 1). Our main aim in building the application has been to take advantage of the special features of Internet, that is, dynamic and interactive features, updating possibilities and wide access. Access to VLISS is free.

VLISS covers to some extent the following aspects of survey sampling: Finite population sampling and estimation, adjustment for nonresponse, approximate variance estimation in complex surveys, and design-based multivariate survey analysis. Additional components include such options as Teacher's Corner, Student's Corner, Course Materials, Help Desk, Links and FAQ pages. Indexes of program codes and concepts are included. From a more technical point of view, VLISS is designed by using PHP, HTML, JavaScript, SAS (Web reference 2) and SUDAAN (Web reference 3) programming languages so that the coding supports Internet Explorer 4 and Netscape 6 or higher browsers. A limited number of applications are written in R language (Web reference 4). Certain functions programmed by the PERL language are provisionally implemented. PHP (Web reference 5) is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML. For browsing VLISS from CD or hard disk, the user is advised to use the Apache Web server (Web reference 6).

This paper is an update of Kiviniemi and Lehtonen (2002).

2. Components of the VLISS application

The core of the VLISS is the concept of "Training Key" making a bridge between the printed textbook and the Web extension. A Training Key includes a fully worked case study covering data specification, statistical analysis (including access to data and program codes) and display as well as interpretation of results. Graphical presentations and Monte Carlo simulation techniques are used when appropriate.

The main pedagogical idea in a Training Key is the "learning by doing" approach. In a Training Key, the relevant materials of the textbook are first worked out (problem setting, data display, computation, interpretation). This is followed by an extended example (going beyond the textbook materials) or a Monte Carlo simulation exercise, which are worked out under the guidance of the application. Finally, an option for interactive further training is offered by downloading the data file and program code on a personal computer.

Each Training Key includes a figure, which refers to the respective page number of the book (i.e. Training Key 101 refers to page 101 of the book). By activating a Training Key the user will enter to the start page of the corresponding training session. Instructions on the use and interpretation of the materials are provided.

Training Keys cover the following aspects of survey sampling and analysis:

- Simple random sampling and the design effect: Analyzing a simple random sample (SRS) drawn without replacement.
- Use of auxiliary information in the sampling design: PPS sampling (sampling with probabilities proportional to size).
- Use of auxiliary information in the estimation design: Regression estimation (with simple and multiple regression), poststratification, calibration.
- Adjusting for unit nonresponse: Reweighting techniques.
- Adjusting for item nonresponse: Single and multiple imputation (Rubin 1987).
- Estimation for domains: Design-based model-assisted techniques using generalized regression (GREG) estimators (Särndal, Swensson and Wretman 1992).
- Approximate variance estimation of non-linear statistics in complex surveys: Linearization, jackknife and bootstrap for a stratified multi-stage sampling design.
- Design-based multivariate survey analysis: Logistic analysis of variance (ANOVA) and the analysis of covariance (ANCOVA) for data collected by stratified multistage sampling design. Pseudolikelihood and generalized weighted least squares estimation are used.
- Accounting for sampling design complexities (weighting, stratification and clustering) in multivariate survey analysis: Design-based and model-based techniques by using pseudolikelihood and GEE methods (generalized estimation equations; Diggle, Liang and Zeger 1994) and generalized linear mixed models (McCulloch and Searle 2001).

3. Example 1: Bias and accuracy of estimators of population totals in a descriptive survey

In Training Key 101 (referring to page 101 of Lehtonen and Pahkinen 2004), Monte Carlo simulation techniques are used to study the bias and accuracy of certain estimators of a population total. The aim is to demonstrate the effect of the incorporation of population-level auxiliary information in the sampling design or in the estimation design for a given sampling design (we use the concept "strategy" to refer to a combination of a sampling design and an estimation design). The strategies examined in the Training Key 101 section "Monte Carlo Simulation" are:

(1) SRS-WOR-HT (simple random sampling without replacement with a standard SRS estimation design) where no auxiliary information is used,

(2) PPS-WOR-STR (stratified PPS sampling with Horvitz-Thompson estimation) using in the sampling design the information on size measures of population elements, and

(3) SRS-WOR-REG (regression estimation for a given SRS-WOR sample) using auxiliary information in the estimation design for a given SRS-WOR sample.

In this Training Key, after examining the relevant textbook materials the user starts working out the extended materials. To examine empirically the bias and accuracy properties of the three different strategies in more detail, the user is guided to simulate several independent samples (10, 100, 500 or 1000) from the given population with a given sampling design. The measures used to examine the relative behaviour of the estimators are the Monte Carlo mean and standard deviation, bias, absolute relative bias (ARB) and root mean squared error (RMSE), calculated on the basis of the distributions of the estimators generated by the Monte Carlo experiments. The distributions of the estimators can also be examined graphically. Finally, a summary table can be displayed containing the results from all simulation experiments.

The effectiveness of the strategies that incorporate auxiliary information either in the sampling design or in the estimation design appears to be very good when compared to the pure SRS strategy. For further training, Training Key 101 provides an option for interactive analysis where the user can download the population frame data set and the program code. The user can then select the sample size and the number of simulated samples, and study more closely the behaviour of the estimators.

4. Example 2: Design-based multivariate analysis in an analytical survey

Training Key 277 provides an access to design-based multivariate modelling of a binary response variable in a complex survey. The main aim of this exercise is to familiarize the user with stepwise model building in a situation where all predictor variables are categorical. Special interests are in demonstrating the role of interaction terms in a logit ANOVA model. The effect of the removal of an interaction term is examined by graphical presentations. The analysis is essentially design based due to the complex structure of the data set used. The data for this Training Key is taken from a real survey (over 7800 observations), which is based on stratified cluster sampling. The phenomenon under study (psychic strain) appears to be positively intra-cluster correlated, as is indicated for example by the design effect estimates of estimated model coefficients, which tend to be greater than one.

Also in this case, the basic example is introduced in the textbook. The Web materials in Training Key 277 provide an extended treatment of the modelling problem. The user is encouraged to work out a step-by-step model building procedure. The model selection procedure begins from the saturated model (which includes all the main effects and interaction terms). The aim is to end up with a reduced model that is parsimonious and fits reasonably well. In all phases, odds ratio statistics are estimated and fitted proportions are calculated and displayed graphically. The system is built to guide interactively the navigation of the user. To make progress, the user must make sensible selections on the model terms to be removed. Once the exercise is completed, the final reduced model can be used for the interpretation of the relationships of the predictors with the response variable.

In the model selection exercise, pseudolikelihood estimation is used to account for the sampling design complexities. In the second part of this Training Key, a more technical treatment of design-based estimation is carried out. Program code written in SAS/IML (Interactive matrix Language) is provided and used for generalized weighted least squares estimation of parameters of a logit ANOVA model.

For further training, the Training Key provides an option for interactive analysis where the user can download the data file and program code for a more detailed examination of the modelling problem.

5. Challenges and further development

The VLISS project has been started in 2002. Our user feedback indicates that the application has been widely used in teaching of survey sampling. VLISS is under continuous updating and new Training Keys will be added. Computer simulation provides a fruitful approach to teach statistics (Mills 2002) and we plan to include additional learning components using simulation techniques.

There are certain limitations in the VLISS application. Currently, SAS and SUDAAN are mainly used as analysis tools. This is due to the fact that SAS and SUDAAN are widely used and they provide good tools for survey sampling and analysis. We plan to include other software (Stata, SPSS and R codes). We are open for proposals for development and other user feedback.

ACKNOWLEDGEMENT

The VLISS project has been supported by a grant of the Virtual University Project of the University of Jyväskylä. I am thankful to Vesa Kiviniemi, Antti Pasanen and Juha Kajava, all students of statistics at the University of Jyväskylä, for their enthusiastic technical work and support in building and developing the VLISS application.

REFERENCES

Brandon, B. (1996). Computer Trainer's Personal Trainer's Guide. Indianapolis: Que Education & Training.

Diggle P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Dodge, B. J. (1995). WebQuests: A technique for Internet-based learning. *The Distance Educator*, 1, 10-13.

Kiviniemi V. and Lehtonen R. (2002). Web tools in teaching and learning of survey sampling: the VLISS application. *Statistics in Transition*, 5, 995-1011.

Lehtonen R. and Pahkinen E. (2004*). Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons, Ltd. (With a Web extension).

McCulloch, C., and Searle, S. (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.

Mills, J.D. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education* (Online), 10. http://www.amstat.org/publications/jse/v10n1/mills.html

Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

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

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education* (Online), 6. http://www.amstat.org/publications/jse/v6n3/smith.html

WEB REFERENCES

Web reference 1: The VLISS Web site. http://www.stat.jyu.fi/mpss/VLISS
Web reference 2: The SAS Web site. http://www.sas.com
Web reference 3: The SUDAAN Web site. http://www.rti.org/sudaan
Web reference 4: The R Project Web site. http://www.r-project.org/
Web reference 5: PHP Web site. http://www.php.net/
Web reference 6: Apache Web site. http://www.apachefriends.org/en/xampp.html

RÉSUMÉ

L'article présente une «solution hybride» pour assister les études universitaires portant sur les enquêtes par sondage. La solution est basée sur un livre déjà paru et son extension WEB.