Statistical Education with Official Statistics on the Internet

K. Laurence Weldon
Simon Fraser University, Statistics and Actuarial Science
8888 University Drive
Burnaby, Canada V5A 1S6
weldon@sfu.ca

1. Introduction

In this paper I address the question "How should the modern developments in official statistics, and the explosion of use of the internet, affect statistical education?" The increasing importance of official statistics for business, government, and education, needs recognition in our assessment of what topics are "basic" for statistical education. Similarly, the internet has greatly increased the feasibility of easy communication of huge data sets at all levels of summary: Not only does this provide an opportunity for enriching application examples, it also increases the importance of certain tools associated with large data sets, and raises problems of data management that expand the boundaries of the discipline statistics. I will suggest how these issues could be responded to in basic statistics courses. I will also allude to the potential for online statistical education which makes use of official statistics via the internet.

2. Extension of basic statistics topics motivated by official statistics applications

The Journal of Official Statistics is full of interesting methodological problems that a arise from these applications - for example, a recent issue (v. 16, 2000) included papers concerned with the following problems: computer-assisted personal interviewing, post-stratification, estimation for highly skewed populations, data editing, and computerized survey software evaluation. How often are these topics mentioned in undergraduate courses? Current popular course textbooks do not discuss them (e.g. Moore and McCabe (1999), Wild and Seber (2000)). It might be argued that these topics are "advanced" topics unsuited to undergraduate courses. But while the details in these papers are advanced, the issues addressed are both commonly met in practice and very straightforward to describe, and students receiving an introduction to statistics should be exposed to these issues.

The issues in this recent journal are typical of issues arising in the official statistic area. Some other fairly well-known ones are:

- issues of confidentiality of data and its management by data providers
- standardization requirements for comparison across geographic units and across time
- non-response bias (biased sample) and response bias (wrong responses)
- specious significance from very large data sets
- measurement of data quality
- survey sampling design

While some courses include some of these topics, I think it is fair to say that most instructors would choose to present details of inference techniques to the exclusion of these topics from official

statistics. My suggestion is to compromise in the direction of sacrificing some inference details in favour of some of these other issues.

One way to expose students to issues associated with official statistics, without distracting them too much from the mainstream topics of data analysis and inference, is to give them projects involving official data sources. Use of the internet allows easy access to large data sets in electronic format. In Canada, the Data Liberation Initiative (www.statcan.ca/english/Dli/dli.htm) arranges for free internet access for students and academics to most official data sources. The web site of Statistics Canada (www.statcan.ca) provides a link to "education resources" which provides many attractive data-based problems. For example, data is readily available on:

- television viewing time by anglophones and francophones of Canadian and foreign programs
- revenue from the sales of sound recordings by region and type of music
- most popular sports
- places Canadians travel and countries sending tourists to Canada
- global warming and greenhouse gases.
- employment and salary by occupation

Some data is based on sample surveys, and some on census surveys. Use of census data helps to make the point that sampling variation is not the only thing that statistical techniques address. Questions of efficient and informative display, data quality, describing relationships between variables, lurking variables, correlation and causation, spatial and temporal variation, levels of measurement, construction of indices, etc can all be raised in the context of census data. A student on a cooperative education work term will likely have as great a need for an appreciation of these ideas as they would have of sampling variation and its associated inference. Our basic courses must give adequate emphasis to this non-sampling material. The context of official statistics is a good way to provide this emphasis.

Most official data sets are large. In fact statistical methods research has recently expanded in the area of analysis of large data sets, with a particular emphasis on data mining. The issue of whether such research is really "statistics" is raised by Friedman (1999). He notes that several areas of information science have been lost to statistics because of the professions attitude that if it is not traditional inference, it is not statistics! He mentions pattern recognition, data-base management, neural networks, machine learning, data visualization, and others as such areas that have been lost to statistics because of the attitude of .statistics academics. A concentration in statistics pedagogy on official statistics would be a partial antidote to this narrow view.

Students using the internet for any purpose must have computer access, and it seems a small step to require some statistical software on that computer, especially when adequate software for student use is now very inexpensive (e.g. MINITAB (2000) or XLStats (2000). When students rely on a university lab computer, there can be a complication with adding the software, and server licensing needs to be arranged. However, for analysis of large data sets, statistical software does seem absolutely necessary for a useful course in statistics, and all the more so if internet sources of data are to be employed.

An issue that arises with using internet-based data as sources of project material for

students is the issue of plagiarism. Web-based material is easily searched for project reports that are ready-made. The problem is not too different from traditional plagiarism, although the ease of access to internet-based plagiarism may make it a more tempting alternative. Instructors have to be prepared to discuss reports with students when the reports submitted are suspiciously erudite. Students who do poorly on hard-to-plagiarize material are both most likely to plagiarize and easiest to detect when they do plagiarize. Instructors may have to pay attention to such evidence when assuring internet-based reports of data analysis are original.

3. Verbalization of statistical concepts

Data access is an obvious benefit for statistical education provided by the internet. But equally important is the benefit resulting from verbal communication about statistical ideas that the e-mail format encourages. Experience with a completely online statistics course is reported in Weldon (1999a and 1999b). A key feature of the course is the online discussion among students in study groups of certain open questions regarding statistical concepts. This discussion is non-simultaneous, on the internet. Students are asked to discuss such questions as "Which technique is best for graphing a univariate data distribution?" or "How do the residuals from a time series smooth reveal information of interest?" or "What practical use is the Central Limit Theorem? ". If the questions can be framed in the context of an important public data set, more students will be drawn into the discussion. The aim is to get students to verbalize statistical ideas, with expert advice available as needed, provided online by a tutor. A moderator of each group of four or five students reports on the discussion of the open question. Students are graded on the quality of their contribution to the discussion as well as the moderator's final report. The entire discussion is visible to the tutor - control of who sees what is easily controlled by the conferencing software FirstClass (SoftArc, 1998).

The importance of verbalization is suggested by the observation that students often get high marks in the calculation questions of an examination without really understanding the what, whether and why of the context of the calculation. (Lipson, 2000). If they are forced to verbalize their opinion of certain statistical strategies, they bring the ideas into the realm of language, and have a better chance of "owning" the ideas as a result. This makes the material more useful to the student than the quickly forgotten calculation procedures. The internet is involved in this verbalization because words are easily transmitted via e-mail whereas hand-waving and formulas are not so easily transmitted. Even discussion of calculations will require some verbalization.

4. Summary

The well-developed field of official statistics involves many problems, both theoretical and context-specific, that suggest an expanded scope in undergraduate statistics courses. One way to introduce students to these ideas is through data-based projects involving official statistics. The convenience of internet access and the feasibility of using large data sets make this option easy to implement. Another important feature of the internet that can be useful for teaching statistics is the ease with which text-based non-simultaneous communication can be effected. If students are

required to communicate statistical concepts in this way, they link natural language with the language of statistics, and the learning becomes more useful as a result. The combination of internet and official statistics can be used for any courses but is particularly useful for distance education.

REFERENCE

Moore, D.S., and McCabe, G. P.(1999) , *Introduction to the Practice of Statistics* Third Edition. W.H. Freeman.

Wild, C.J. and Seber, G.A.F. (2000) *Chance Encounters: A First Course in Data Analysis and Inference*. Wiley.

Weldon, K.L. (1999a) Experience with an Online Introductory Statistics Course. Bulletin of the International Statistical Institute. Contributed papers. 52nd Session. Book 3. pp 425-426. Helsinki.

Weldon, K.L. (1999b) Seven Trimesters of an Online Introductory Course. Journal of Distance Education 14(2):81-84. Ottawa, Canada.

Softarc Inc (1998) FirstClass Version 5.506 is software puplished by SoftArc Inc. and detailed information can be obtained from info@softarc.com.

Lipson, K. (2000) "Determining the relationship between the concept of sampling distribution and the development of understanding of inferential statistics". PhD thesis,
Swinburne University. Melbourne.

Friedman, J.H. (1999) The Role of Statistics in the Data Revolutuion? Bulletin of the International Statistical Institute, Tome LVIII, Book 1: 121-124.

XLStats(2000). Carr, R. (2000). XLStatistics 5.28. XLent Works, Australia. The URL is
http://www.man.deakin.edu.au/rodneyc/xlstats.htm

MINITAB(2000). Minitab Inc. The URL is http://www.minitab.com/

RÉSUMÉ

En cette papier, j'adresse à la question "Comment devrait les développments modernes du statistique officiel, et l'explosion du usage de l'internet, ont une effet sur l'éducation statistique?" L'importance croissant du statistique officiel pour le commerce, le gouvernment, et l'éducation, ont besoin de récognition dans notre evaluation de quelques topiques sont "fondamental" pour l'éducation statistique. Aussi, l'internet a augmenté beaucoup des possibilités de faire communication facile du collections de données tres grandes à toutes niveaux de résumé. En plus de fournir une occasion de enricher les exemples applications, mais aussi augmente l'importance de certains outils que sont associé avec les collections de données tres grandes, et éleve les problèmes de direction de des données que dilate les limites de la discipline du statistique. Je suggérai un mode utiliser ces questions dans les courses statistiques fondamentals. Je parlerai des potentialités pour l'éducation statistique online qu'utilises statistiques officieles sur l'internet.