LEARNING STATISTICS IN THE SENIOR SCHOOL USING ILLUSTRATED REAL DATA IN CONTEXT

HARRAWAY , John A
The University of Otago
New Zealand

Data generating project work in middle school is a successful component of statistics training at this level. CensusAtSchool, GAISE, CAUSE projects and Statistics New Zealand SURFs also provide data for investigation in middle school and beyond. To motivate students further, real data from current research should be used in the classroom to help take statistics training to the next level. Such data are described. Eleven research projects are presented in DVDs with matching data provided on CDs. Context is outlined, study design and data collection are discussed, and proposals are made for analyses to answer questions. The DVDs, developed for use in 2009, feature illustrated researcher talks. The DVDs and CDs are also available for viewing on the web. A range of studies are described along with tasks to be carried out. Feedback from school teachers and students on relevance of this resource will be discussed.

INTRODUCTION

The teaching of statistical literacy, statistical thinking and design of studies in schools in New Zealand is developing in a learning stream called Mathematics and Statistics that extends from new entrants at primary school to the end of secondary schooling (Ministry of Education, 2007). Teaching resources are needed to help the new syllabus and to maintain interest of students. This is important in the last two years of school because it is often at this stage that career objectives are considered. Resources described target this level and extend concepts developed earlier. The new syllabus emphasizes the statistical enquiry cycle. This has five components, problem, plan, data, analysis and conclusions. These reflect the four steps, formulate questions, collect data, analyze data and interpret results, listed in the Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2005).

The statistics syllabus in the last two school years in New Zealand has three themes, statistical investigation, statistical literacy and probability (Ministry of Education, 2007). The syllabus covers surveys that require random sampling, experiments using experimental design principles, use of existing data sets, choice of measures for variables, choice of data collection procedures, exploratory data analysis and statistical inference. Models are assessed using linear regression for bivariate data and additive models for time series data, seeking explanations, making predictions and communicating findings. In the final year at school the inferences made from surveys and experiments involve determining estimates and confidence intervals for means, proportions and differences. Both normal and bootstrap confidence intervals are developed.

Statistical literacy involves the evaluation of statistically based reports including surveys and polls, experiments and observational studies, critiquing causal-relationship claims, interpreting margins of error, interpreting risk and relative risk, and identifying sampling and non-sampling errors in surveys and polls.

Probability investigates situations that involve chance, calculating probabilities using tools such as two-way tables, tree diagrams and simulations, calculating probabilities of independent, combined and conditional events, calculating and interpreting expected values and standard deviations of discrete random variables, and applying distributions such as the Poisson, binomial and normal.

REAL DATA: CONTEXT, DESIGN AND ANALYSIS FOR ELEVEN STUDIES

Real data can enhance student learning provided the data are genuine, motivated well and available for school students to work with. Case studies now described illustrate these objectives. Context is discussed in edited DVD presentations by the researchers describing their current investigation. A CD with the data is also provided along with task suggestions using Excel. These packs are available to high schools. Each DVD recording lasts between 12 and 20 minutes. The DVDs, the CDs and task suggestions for teachers can also be viewed at the website (http://www.maths.otago.ac.nz/video/JAH).

Example 1: Effect of circumcision on herpes (and other sexually transmitted diseases).

This observational study is based on a cohort of children born in 1972/73 in Dunedin who are being followed throughout their lives. When aged 26 the cohort was questioned about presence of herpes, circumcision and sexual partners along with many other variables. The data showed circumcision was protective of herpes but, when allowance was made for the number of sexual partners, herpes was found not to be protective. The DVD is a 10 minute presentation featuring the researching epidemiologist (Dickson et. al., 2005). The CD has data for 1890 cases and nine variables with definitions of categorical variable classes.

Tasks involve analysing observational data resulting from a large cohort, confidentiality issues, causal relationship claims illustrated by investigating cross tabulations for different subgroups or strata, proportions and confidence intervals, interpreting risk and relative risk, and the importance of confounders when working with observational data. This leads to the idea of a designed clinical trial and can be illustrated by reference to trials in Africa to assess the impact of circumcision on HIV/Aids (Bailey et. al., 2007). It is noted that observational studies can suggest pattern in data and relationships between variables but are not used for statistical inference.

Example 2: The side effects of a drug for lowering cholesterol.

This experiment contrasts with the previous observational study. It is a randomised controlled trial evaluating the effects of a cholesterol lowering drug which has potential side effects. Patients are allocated randomly to control and treated groups with about 560 patients in each group. Two years later the reductions in cholesterol in the groups are compared. The distributions of differences in both groups are close to normal. Responses to thirty questions measuring psychological effects are scored 0, 1, 2, or 3 for each person. The maximum score is 90 and the minimum score is zero with the pattern of scores for all patients close to normal. There are five categorical variables in the study and six continuous variables. The categorical variables are treatment classification, sex, socioeconomic status at the start of the study, smoking status at the start of the study and extent of prior heart problems. The continuous variables are age, body mass index, cholesterol level and psychological score at the start of the study and cholesterol level and psychological score after two years. Obtaining the psychological score is an important procedure and the questionnaire document is available.

Questions proposed ask if the drug lowers cholesterol, if it has side effects and if the two groups are equal at the start of the study. The fact that randomisation has removed potential confounding effects can be discussed. This can be tested by examining bar charts, cross tabulations and histograms at baseline to establish if there are differences. There are missing data values to be investigated. The analysis therefore includes graphical procedures to summarize data, investigation of variability, calculation of point estimates, and construction of confidence intervals for means, differences between means and proportions. Further analysis could involve regression procedures to analyse relationships between the continuous variables and predict values for missing data.

Example 3: Maui dolphins – uncovering a new species.

This study proposes that Hector and Maui dolphins are different species even though they are small dolphins with similar appearance. Earlier genetic evidence suggests this to be the case (Baker et. al., 2002). The DVD discusses the ecology of the dolphins then describes six measurements recorded on 59 Hector and Maui dolphin skulls found in museum collections in New Zealand, 46 Hector skulls in South Island and 13 Maui skulls in North Island. The 59 sets of measurements are available in the CD.

Procedures carried out involve summary tables and graphs, means and confidence intervals for North and South Island comparisons, correlation between the six measurements, bootstrap confidence intervals, regression for estimating missing values, and exploring relationships between continuous variables. Five tasks are suggested with instructions for teachers. Students will see clear evidence that North Island Maui skulls are larger than South Island Hector skulls, providing evidence of a new sub species. The explanation goes back 15000 years when the two main islands of New Zealand separated. The Maui and Hector dolphins are in-shore animals and there was no contact between the two animals once the islands separated.

Example 4: Measuring public opinion on support for public funding of the Otago Stadium.

A survey of public opinion on building a sports stadium costing $188 million using ratepayer funds is carried out. Two DVD presentations describe question formulation, data collection, data analysis and conclusions reached after discussing validity and other issues that could bias results. A random sample of 5000 questionnaires was mailed to members of the target population. Data from 2100 respondents to this emotive issue are available. This study reflects the five steps of the enquiry cycle. The research question is formulated given conflicting claims by the football union that the new stadium is essential and by the anti stadium citizens that the existing sports field is adequate for international fixtures. The plan involves defining a target population, a sampling frame, and a sampling method including selection of a random sample after establishing the size of the sample given an acceptable margin of error. Data gathering uses a mail out self-completion questionnaire with 10 questions, one relating to support or no support for the stadium and the others demographic questions relating to sex, age, financial circumstances, home ownership and education background.

Data analysis involves cross tabulations, a crude confidence interval for the proportion of residents opposing the stadium, and confidence intervals for the proportion after stratifying for demographic factors. The crude proportion shows 66% (margin of error 2.1%) oppose the stadium with this value increasing to 78% (margin of error 3.5%) after adjusting for sex, age and income. Tasks outlined show students how to calculate results after stratification. Conclusions discuss threats to validity resulting from unit non-response bias, item non-response bias, instrument bias in relation to the question asked, and data entry errors.

Example 5: Exploring data on a rare threatened bird – the rock wren.

The rock wren is a native bird inhabiting the Southern mountains of New Zealand. This DVD describes the habitat of the bird and poses the question that rock wren numbers are decreasing. The data for rock wren counts in 1984/85 compared with 2005 in one region of Fiordland are recorded and are available in an accompanying CD. A set of tasks for students to perform is suggested.

The analysis explores the data graphically and sets up confidence intervals for differences using both the normal and bootstrap confidence intervals. The results from these procedures are compared. One approach to resampling using the rock wren counts is to ask each student to generate several resamples then combine these in class. An Excel routine for the bootstrap is also provided which allows students to perform the task on their own data and also explore the consequence of altering the number of samples chosen.

Example 6: Property sales in North Shore City – 1999 to December 2008.

The analyses of house sales on the North Shore are described in an 18 minute DVD. Data provided by the Real Estate Institute of New Zealand are on a CD and tasks which students can perform are described. There are two data files. The first discussed in the DVD has 3788 residential property sales recorded for the month of December from 1999 to 2007. The second records 9281 sales each month from December 2006 to December 2008. The first file allows investigation of trends on the North Shore before the property crash while the second file shows the impact of the property crash. For each sale several categorical and continuous variables are recorded, namely area of property, list price, selling price, listing date, selling date, number of days to sell, number of bedrooms, type of property, land area and government valuation. An inspection of missing values is carried out and data cleaning instituted.

Many questions can be posed. These could investigate if the number of sales per month is decreasing, if the number of days to sell a property is changing and if the median price is rising. Tasks described set up descriptive statistics and graphs for the continuous variables, explore how the property market is performing using StatPro, and investigate differences between list price and selling price and between government valuation and selling price. The data are appropriate for time series analysis and regression if edited appropriately. Random subsets of the data could be investigated.

Example 7: Effects of satellite transmitters on titi (sooty shearwater).

This DVD describes the ecology of the titi which has one of the longest migration routes on the planet. The bird breeds on islands of the South of New Zealand. The harvest is declining indicating the resource may not remain sustainable. Transmitters have been attached to birds to track their flight path during migration north. This study assesses the effect of transmitters on birds. An experiment is described and some birds are randomly allocated an imitation transmitter or not when nesting in a local colony.

The two groups are compared by recording the effect of the imitation transmitter on colony attendance. In particular, the daily proportions of nest attendances for the sampled birds over a period of 41 days are recorded. Results are compared using simple linear regression of the proportion of nests attended against day for the birds with and without imitation. A difference is observed from the two scatter plots. The proportions of nest attendance for each day are also compared directly but confidence intervals are wide given the small samples. Bootstrap confidence intervals can also be established. Those with the transmitters are more likely to be attending the colony.

Example 8: An exploration of global rotavirus deaths.

The World Health Organisation monitors health issues in countries of the world. Diarrhoea is a serious condition which kills many children under five. In this DVD deaths due to rotavirus are investigated to obtain an estimate of numbers affected by rotavirus in the world. Health issues are discussed and the accompanying CD has detail for 175 countries. Variables reported are the estimated proportion of diarrhoea cases due to rotavirus from a large number of country surveys, diarrhoea deaths for children under five and four health indicators.

Countries are grouped or stratified separately on the four health indicators and each of the four stratifications is used to find an estimate for the number of deaths. Confidence intervals are established either by normal methods or by the bootstrap. Different stratifications are compared.

Example 9: Stress in cattle.

Currently being recorded and edited, the DVD describes the problem of stress in cattle and outlines a method for measuring stress using data from the eye of the cattle beasts. This is an important problem given the dependency of New Zealand on agriculture. The study is a designed study. It does not depend on random samples which arise from the likes of a target population in a survey, instead relying on random assignment of "treatments" to preselected cattle beasts, the experimental units.

Example 10: Who will consume alcohol when pregnant?

Currently being recorded and edited, the DVD examines the profiles of women who would consume alcoholic drinks while pregnant. This study relies on a random telephone survey of 2083 women aged 16 to 40 years with responses from 1327 women. The sample reflects the population profile of New Zealand women with respect to age, ethnicity, education, employment, household income, marital status and type of drinker. Both ethnicity and prior history of consuming alcohol are significant predictors. (Parackal et.al., 2009)

Example 11: Health issues of the Sherpas in the Himalayas.

This study compares health issues at high altitude with health at sea level. It is a designed study involving matched data for paired tests and also unpaired tests. A conclusion is reached about sleep apnoea at high altitude.

DISCUSSION

Other sources of data for school use are available but some like CensusAtSchool are aimed currently at the middle or lower school. In 2007 in New Zealand, 505 schools covering 25048 students participated in CensusAtSchool. A random sampler provides access to samples of participants, a table-maker allows users to create tables which may be viewed graphically and activities are found at CensusAtSchool website (http://www.censusatschool.org.nz/). Expansion is intended to the final two years at school building on earlier ideas of the statistics enquiry cycle. This requires a wider range of examples if the interest of older students is to be retained. Examples outlined here are appropriate. The idea is to move from self generated data sets constructed by students to real data with the context of each study discussed by the researcher. Adding the data files and DVD presentations to the CensusAtSchool site is being investigated to keep resources in one easily accessible location.

The GAISE project develops statistics education from early through to late school years. There are three levels to the programme based not on age but on development of statistical literacy. Students begin with the first level regardless of age before moving to the second and third levels. The third level reflects the syllabus in the last two years of school in New Zealand and proposes a range of interesting case studies of a deeper statistical nature which build on earlier concepts. The case studies in the DVDs would be equivalent to some examples in third level GAISE but extend to new areas. The Consortium for the Advancement of Undergraduate Statistics Education (CAUSEweb.org) supports undergraduate statistics education in areas including resource provision. Lecture examples, laboratories, links to useful data sets and webinars describing classroom activities motivate active learning in a way similar to the DVD case studies but representing a wider range of target groups.

Another source of real data for school use is the synthetic unit record file (SURF) created from official statistics survey data by Statistics New Zealand. Three files have been constructed based on official statistics survey data at the national statistics office (Statistics New Zealand). Official Statistics agencies need to keep information confidential. The first SURF has been created from the New Zealand Income Survey and has 200 records with 7 variables. The second is from the Household Savings Survey with 300 records and 11 variables while the third is from the Census and has 300 records with 11 variables. The SURFs are downloadable from the website

(http://www.stats.govt.nz/schoolscorner). Included with the datasets are suggested classroom activities at different levels of schooling with some activities appropriate for the final year at school.

In 2006 a first DVD/CD was developed by recording talks at an education day during the New Zealand Statistical Association Conference. Nine talks were well received by teachers who requested that the video recordings with data on a CD be made available to schools (Otago University, 2007). 210 packs were distributed to schools in 2007. Feedback suggested the studies were interesting but that the data sets were complicated without adequate explanation of tasks to be carried out. Some of the presentations were long and it was suggested 15 minutes was desirable for a DVD. It was agreed, however, that real data from current research is a useful teaching resource which will motivate interest in statistics. Therefore, it was decided to produce a second DVD/CD pack in 2008 with shorter talks, and suggest a set of appropriate tasks. The new DVD/CD packs will be made available to schools in 2009 and are also be available at the website (http://www.maths.otago.ac.nz/video/JAH). Some of the DVDs are being trialled in local schools and feedback will be reported.

ACKNOWLEDGEMENTS

The Campbell Fund of the New Zealand Statistical Association supported this project which allowed researchers to travel and be recorded in Dunedin. The Higher Education Development Centre at the University of Otago filmed and edited the DVDs free of charge as a research project to assess use and relevance of such a teaching aid for final year high school.

REFERENCES

Baker, A.N., Smith, A.N.H. and Pichler, F.B. Geographical variation in Hector's dolphin: recognition of new subspecies of Cephalorynchus hectori. *Journal of the Royal Society of New Zealand.* 32: 713-727 (2002).

Bailey, R.C., Moses, S., Parker, C.B., Agot, K., McLean, A., Krieger, J.N., Williams, C.F.M., Campbell, R.T. and Ndinya-Achola, J.O. Male Circumcision for HIV Prevention in Young Men in Kisumu, Kenya: a Randomised Controlled Trial. *Lancet* (2007), 369, 643-656.

Dickson, N., van Roode, T. and Paul, C. Herpes Simplex Virus Type 2 Status at Age 26 is not Related to Early Circumcision in a Birth Cohort. *Sexually Transmitted Diseases* (2005), 32, 8, 517-519.

GAISE (2005) http://www.amstat.org/education/gaise .

Ministry of Education. The New Zealand Curriculum. Learning Media Limited, Box 3293, Wellington, New Zealand. (2007).

Otago University. Statistics in research: a visual teaching resource. Ed. J. A. Harraway and R. van der Vyver

Parackal, S.M., Parackal, M.K., Harraway, J.A. and Ferguson, E.L. Opinions of non-pregnant New Zealand women aged 16-40 years about alcohol consumption during pregnancy. *Drug and Alcohol review*. 28, 135-141 (2009).