® ASSESSING LARGE SECOND YEAR UNDERGRADUATE SERVICE COURSES IN DATA ANALYSIS

FORSTER, Mike & SMITH, David P.
The University of Auckland
New Zealand

The guiding principle in the assessment of any course is that we must assess what we teach. We begin by outlining the assessment instruments we use and discuss how we use these instruments to assess what we teach. We then look at the following assessment considerations: firstly, two specific types of questions we use and why we use them, then equity for students across semesters, the time and cost associated with assessment, some strategies and administrative tools and finally, one of the biggest challenges, finding enough suitable data sets to use.

INTRODUCTION

The assessment of a large regularly offered service course presents some unique challenges. At Auckland, the statistics department's guiding principles are to build up student numbers, retain as many students as we can and establish and maintain a good reputation. The key to achieving these aims is to provide service courses that are well resourced, well taught, well administered, and fairly assessed. With service courses, we not only gain a large captive audience (commerce, social sciences, …) and the funding associated with it, but they give us the opportunity to convince that audience of the benefits of studying statistics and becoming statistically competent.

One of the major challenges associated with providing large service courses is their assessment. While our prime concern must always be that we assess what we teach, equity in the assessment across semesters is also important. The inevitable budgetary and time constraints present a major challenge, which can be mitigated by using a variety of assessment instruments and developing suitable administrative tools and procedures. Another challenge in assessing courses that are regularly offered is finding enough suitable data sets to use in the assessment instruments.

In our first year service course in Introductory Statistics, with close to 4500 students per annum over the three semesters of our academic year, the time constraint leaves us little option but to use a set of assignments, marked by final year undergraduate and postgraduate students, and a mid-semester test and final examination that are entirely multiple-choice questions. In our second year service course in data analysis, with around 1300 students per annum, the time constraint is somewhat less of an issue.

ASSESSING SECOND YEAR SERVICE COURSES IN DATA ANALYSIS

In our second year course in Data Analysis we aim to teach our students to recognise the type of analysis appropriate for a given set of data (spot the analysis), to analyse the data in a modern computer package (R), to interpret their results, present their findings in a two part written report and finally, to have a basic understanding of the underlying theory and concepts of statistical analysis.

The assessment instruments we use are a set of four assignments (20% of their final grade), a mid-semester test (10% or 20%) and the final examination (70% or 60%).

The majority of the assignment questions are simply a description of a data set, including how and why it was collected, and the variable names with brief explanations. The students are then required to choose the appropriate form of analysis to use, analyse the data in R and write up a set of Technical Notes on their analysis and an Executive Summary of their main findings. These questions are designed to address all the aims of the course, and make up about 55% of the total assignment marks. They require around 14 different data sets to cover the main techniques we teach (one, two and paired sample t-tests, one and two-way anova, one and two-way tables of counts, odds ratios, multiple regression, and ancova, including data transformations).

A student's writing style in the Executive Summaries and their assignment presentation make up 5% each of the total assignment marks. The remaining 35% includes completing a class

web survey (a valuable source of data as well as a useful teaching tool), revision of key first year concepts, plotting in R, using simulated data on questions designed to help students understand key concepts and techniques in the course (e.g. transforming data) and identifying the type of analysis appropriate for a given description of some data.

The mid-semester test is entirely multiple-choice questions, a decision that is largely dictated by the time constraints that operate at that time of the semester. Around 75% of these questions, requiring another four data sets, relate to the interpretation of output from R, while the remaining questions are designed to test theory and concepts and to get the students to identify the appropriate form of analysis to use.

The final examination is 30% multiple-choice questions, including interpretation of output from R, largely on analyses from the end of the course (where we introduce our students to time series and logistic regression) and with the remainder being theory and concept questions. Another 20% focuses on short answers, including calculations (odds ratios and fitted values in regression) and identifying the appropriate form of analysis to use. The remaining 50% requires the students to write Technical Notes and Executive Summaries on analyses that have been run in R and given in a data appendix. In the examination, we typically require another eight data sets. The assessment instruments require 26 data sets, per semester.

ASSESSMENT CONSIDERATIONS

The main consideration for any course is to assess what we teach. Large regularly offered service courses impose some additional constraints. We begin by discussing two types of questions we use, multiple choice and "A" grade sorter questions, including some examples. We then discuss equity for students across semesters, economic considerations in writing and grading the assessment, administrative tools and finally, finding enough suitable data sets to use in the assessment instruments.

MULTIPLE-CHOICE QUESTIONS

With around 35% of a student's final grade being assessed using multiple-choice questions, we have put considerable effort into designing questions that not only assess a student's learning, but also act as a resource for future semesters. Writing good multiple-choice questions is a very time consuming task. The advantage is that the writing time is traded-off against quick, relatively cheap grading and the creation of a future course resource.

Good multiple-choice questions are typically ones that students should know the answer to as soon as they have read the question, and that the correct option should be a positive, or correct statement. Our multiple-choice questions conform to neither of these ideals. We typically ask: Which one of the following statements concerning the output on Page X of Appendix Y is false? (Figure 1)

Our reasons for this approach are threefold. Firstly, it is much easier to think up four correct options and one false option, than the other way round, vastly reducing the writing time. Secondly, when we are getting students to interpret a page of statistical computer output in a single multiple-choice question, designing questions that the student knows the answer to once they have read the question is almost impossible. Lastly, a well resourced service course that is equitable for students across semesters will provide examples of previous test and exam questions for the students to use in their study. We believe that it is infinitely better in this situation that our students see four correct statements rather than four incorrect statements. Incorrect statements do not reinforce what it is we want our students to learn, correct statements do. Therefore, using "which one of the following is false" provides a much better study resource for the students, as it is one that focuses on positive reinforcement. We also use multiple-choice questions to test a student's understanding of the underlying theory and concepts of statistical analysis (Figure 2).

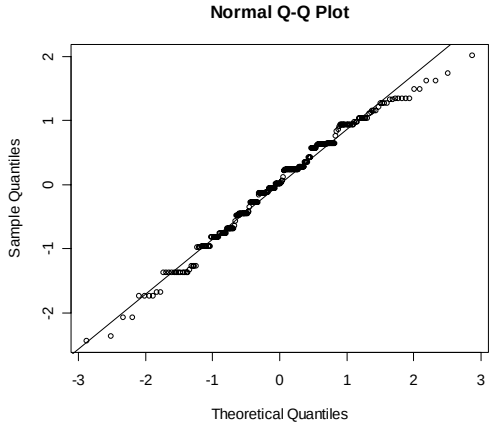
```
> levene.test(study~status)
             Df  Sum Squares  Mean Square  F-statistic  p-value
Between Gps  2   570.25618    285.12809    3.71252      0.0258
Within Gps   245 18816.4172   76.8017
Total        247 19386.67339

> log.study<-log(study)

> levene.test(log.study~status)
             Df  Sum Squares  Mean Square  F-statistic  p-value
Between Gps  2   0.04936      0.02468      0.09745      0.90719
Within Gps   245 62.05657     0.25329
Total        247 62.10593

> study.fit<-lm(log.study~status)

> qqnorm(study.fit$residuals)
> abline(0,summary(study.fit)$sigma)
```




```
> shapiro.test(study.fit$residuals)

        Shapiro-Wilk normality test

data:  study.fit$residuals
W = 0.9855, p-value = 0.01297
```

Which one of the following statements concerning the output on page 3 (Appendix A) is FALSE?

(1) We could have faith in the F-test P-value if we performed ANOVA on the untransformed data, since we can rely on the Central Limit Theorem.
(2) For the transformed values, we have no evidence against the hypothesis that the underlying population variances are the same.
(3) The Normal Q-Q plot indicates that the residuals could have come from a normal distribution.
(4) We have evidence against the hypothesis that the underlying errors for the log model come from a normal distribution.
(5) For the untransformed values, we have evidence against the hypothesis that the underlying population variances are the same.

Figure 1: R Output and Multiple Choice Question

Which one of the following statements is FALSE?

(1) R^2 always increases if we add another explanatory variable to our model.
(2) An R^2 of 89% implies that 11% of the variation in the response is not explained by the fitted model.
(3) If we obtain a high R^2, then we do not need to worry about the other regression assumptions.
(4) In simple linear regression, R^2 is the square of the sample correlation coefficient.
(5) If there is a large amount of scatter about the trend, R^2 will tend to be low.

Figure 2: Concept Multiple Choice Question

"A"-GRADE SORTER QUESTIONS

With large numbers of students, it is useful to be able to determine who the very top students are. In our final examination, we typically include one question that is designed to identify these students, with an eye to encouraging those who do well to further their statistical studies. We use two approaches: a question that requires the students to apply what they have learnt to a situation they have not specifically seen in the course, but to think laterally (Figure 3), or a question that covers some detailed aspect from the course notes (Figure 4).

The 20x Students' data used in Appendix A was analysed using a Linear Regression model. Selected R output is presented below. (Note: You may wish to look at the boxplot on page 4 and the t-test output on page 5 of Appendix A.)

```
> fit<-lm(log(clothes)~sex)
> summary(fit)

Call:
lm(formula = log(clothes) ~ sex)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.39197    0.06159  71.311  < 2e-16 ***
sexmale     -0.42343    0.09061  -4.673 3.85e-06 ***

> ci.reg(fit)
            95 % C.I.lower    95 % C.I.upper
(Intercept)        4.27095           4.51298
sexmale           -0.60147          -0.24540
```

(a) The estimate for the Intercept is 4.39197. Briefly explain what this is an estimate of.
(b) Sketch a scatter plot of the logged data and draw in the fitted line from the regression model. Label the values of the intercept and slope. Put the minimum and maximum numeric values on the axes.
(c) The confidence interval for sexmale is (0.60147 , 0.24540). Briefly explain why this is a confidence interval for the difference in means (approx = medians) of the transformed data.
(d) Quantify, in a meaningful way, the difference in the median amount spent on clothes per month by male STATS 20x students compared to female STATS 20x students, using the above output.

Figure 3: Lateral Thinking Question

Typically, we never get more than five or six out of 500 odd students in a semester gaining full marks for these questions, and most of them are among the top ten students for the semester. The majority of students get either 0 or 1 mark here.

In Case Study 3 of the Course Notes we performed a Two Sample t-test to determine whether regular lecture attendance affected a student's final examination mark in 20x. We found a highly significant result with the P-value = 0.0000011 and a 95% confidence interval for the difference in the mean exam mark between attenders and non-attenders of [9.6 , 21.5].

(a) Does this result mean that regular lecture attendance causes higher exam marks, on average? Explain your answer.

In Case Study 33 of the Course Notes we built a model to explain or predict 20x students' final exam marks. In the model building process, we dropped the variable attend as it was found to be not significant (P-value = 0.23). The final regression model is presented below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.69567    6.49900   4.723 5.58e-06 ***
Assign        1.56651    0.23068   6.791 2.93e-10 ***
Test         -1.24720    1.12706  -1.107  0.27036
I(Test^2)     0.13846    0.04698   2.947  0.00376 **
Stage1B      -4.45745    2.21620  -2.011  0.04621 *
Stage1C     -11.84911    2.35973  -5.021 1.54e-06 ***
```

(b) The t-test (part a) showed that the difference in mean exam marks between lecture attenders and non-attenders was highly significant. Explain why this variable was not needed in the final regression model.

Figure 4: Detailed Understanding Question

EQUITY

Part of the success of service courses, especially courses that most students would rather not have to take, depends on them developing a positive reputation amongst the wider student population. Critical aspects of this are that students come to see that the course assessment is reasonable in terms of their time and effort, that it assesses what is taught in the course, and that the assessment is equitable across semesters. Our philosophy is that any student who passes a service course in a given semester should pass, with the same or a very similar grade, in any semester. This can be achieved in two ways: making the assessment easy, which defeats the underlying purpose of the course, or making the assessment predictable. We choose predictability.

Predictability can be achieved by fully informing the students of the assessment instruments that are to be used at the beginning of the course, sticking to a set format across semesters for each of the instruments and giving the students examples of previous assessment instruments to use as a resource and study tool. We provide our students with copies of the test and exam from the previous 3 semesters' courses, including model answers to all questions. In addition, over 35 separate case studies in the student's course notes, with the data analysed in R, and examples of Technical Notes and Executive Summaries provide the main resource the students need for most of their assignment work.

Church, Elliot and Gable (2001) discuss methods of teaching and assessment that try and get students to focus on performing well. They believe this can be achieved by making classes engaging and giving the students the impression that the assessment instruments are not difficult. If the assessment is perceived as difficult, students will tend to focus on just passing. Our aim is to try and make the students see the assessment as fair and relevant to what it is we teach. We do try to remove any perception that the assessment is difficult in order to motivate the students to perform well in the course and to master the content, rather than just aim to pass.

With the assignments, we regularly tell the students when they can do questions after the relevant topics have been completed in class. We do not always tell them specifically which

questions can be done, since one focus of our course is getting the students to recognise which form of analysis is appropriate for a given data set. We also advise them in class how to use the study resources we provide to maximize their performance in the test and final examination.

ECONOMIC FACTORS: COST AND TIME

The time involved in preparing assessment instruments for regularly offered service courses is considerable, especially finding appropriate data sets to use. The time and cost of marking the instruments, especially when the courses are large, can also be considerable.

We have already noted that the marking time constraint is the major reason our mid-semester test is entirely multiple choice. The cost of grading assignments is somewhat mitigated by using final year undergraduate and graduate students, but having a sufficient number of these depends on successfully convincing students to study statistics. The final examination must, in our opinion, be fairly and appropriately graded by competent statisticians. We only use staff and doctoral students here.

By combining different types of assessment instruments we can keep the cost of grading reasonably low, make the time required to prepare and write the instruments and model answers acceptable and yet still achieve the key requirement of any course assessment, to assess what we teach. Since we then use the assessment instruments as a self-updating resource for the students, the cost and time involved is further mitigated.

SOME STRATEGIES AND ADMINISTRATIVE TOOLS

The administration of the assessment presents a whole range of issues. When using inexperienced senior students to mark assignments, despite giving them model answers and detailed marking guides, the marks awarded to different students can be variable. Each assignment has a total mark allocation of 70 or 80. To partially overcome this marker variability, the final mark awarded for an assignment is out of ten, with any fraction rounded up. Once this system is known and understood by the students, the number asking for a re-mark of their assignment drops dramatically. The students are given complete model answers to the assignment, which makes it easy to detect a marking error, but unless the error moves them up to the next integer mark, asking for a re-mark is just a waste of their time, and ours.

Large numbers of students produce a large number of examination scripts, and in our course, 70% of each script has to be marked manually. Equity within a given semester, combined with the large number of scripts to mark, lends itself to one individual marking a single question for all the students. It does not matter whether they mark soft or hard, as all students are treated equally, in a given question. In order to minimize the overall time for grading, and the inevitable bottlenecks that emerge when script bundles need to be swapped among markers, we put together a separate stapled answer booklet for our students to write all their answers in. After the examination, all scripts are dismantled so that each marker can take a question away and mark all of the scripts. Once all the marking is done, the scripts are re-assembled.

The cost in time of separating and re-assembling scripts is mitigated by less downtime in the marking process, enabling us to put together the students' final grades much faster. Entering marks into a spreadsheet is also quicker if done question by question, rather than student by student. Once all the marks are entered, they are then cross checked, student by student, after the scripts have been reassembled so that any addition or data entry errors are picked up.

DATA SETS

As mentioned above, a large second year service course in Data Analysis requires a large number of data sets to use in the assessment instruments: for us, 26 per semester, or 78 per year. Finding sufficient data sets across a wide range of analysis techniques is incredibly time consuming. We have developed some useful strategies that can be used to reduce the time required hunting out data.

At the beginning of the semester, we get our students to fill in an online survey as part of their first assignment. There are 24 questions ranging from their age and gender to their favourite fruit, the number of txt messages they send and receive per day, to the amount they spend getting their hair cut. A further five categorical variables are then created from various questions in the

survey. For example, the question on the average amount spent on tobacco products per week is turned into a categorical variable that indicates whether or not a student smokes.

The resulting data are then used in class to demonstrate some of the problems associated with survey data in general and techniques involved in data cleaning. The data from the previous semester in used in the first assignment to give students practice in plotting with R. Some of the open ended data analysis questions in the assignments and some the data sets used in the test and exam are also taken from the current and/or historical survey data sets.

A data bank of around 250 data sets that have been used in previous assessments has also been created. These data can then be re-used, after a suitable period of time has elapsed. These data sets came from textbooks, journals, colleagues, consulting jobs and even some that have been collected by members of the staff.

Another advantage of using different types of assessment is that data sets can be used more than once in a given semester. If a data set is used in an assignment where the students are required to analyse, interpret and write a report, it can be used again in the test or exam multiple choice questions.

DOES IT ALL WORK?

We believe our service courses are successful. We have developed an enviable reputation among the student population for service courses that are well taught, well resourced, well administered and fairly assessed. The faculties and departments whose students we teach are happy for us to continue teaching basic statistics (first year) and data analysis (second year). Our second year course in data analysis is a pre-requisite for marketing students and those doing postgraduate studies in accounting and finance. The enrolments in both our first and second year service courses continue to increase.

We also use our service courses as a vehicle to convince a captive audience of the benefits of studying statistics. Many of our final year undergraduate and postgraduate students never intended to make statistics their main field of study. We have also been successful in convincing many students to take statistics as a minor in their degree to support a variety of other fields of study from commerce to marine science.

CONCLUSION

In a modern university setting, large service courses can become the foundation of a successful statistics department. Service courses that are well resourced, well taught, well administered and fairly assessed provide us, as statisticians, with an opportunity to convince students of the value to them of studying statistics. The assessment of large numbers of students presents its own set of interesting challenges. Using a variety of assessment instruments that assess what is taught enables scarce data sets to be used more than once in a given semester. Obtaining data directly from the students via a web based survey also alleviates part of the time consuming process of finding suitable assessment data sets. Assessment instruments that can double as a study resource help to create good courses from a student's perspective and save valuable time for teaching staff. Equity across semesters in service courses is essential and predictability in how the students are to be assessed is the best way we have found to achieve this. Carefully designed administrative tools can reduce the burden that large numbers of students can become for staff involved in teaching and running large service courses.

REFERENCES

Bradstreet, T.E. (1996). Teaching Introductory Statistics courses so that non statisticians experience statistical reasoning. *The American Statistician, 50*, 69-78.

Burton, R.F., & Miller, D.J. Statistical modelling of multiple-choice and true/false tests: ways of considering, and at reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, *24*(4) 399.

Bush, M.E. Quality assurance of multiple-choice tests (2006). *Quality assurance in education: An international perspective*, *14*(4), 398-404.

Crawford, E., & Bowman, A. (2002). Web resources for teaching and learning statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics:*

Developing a statistically literate society, Cape Town, South Africa. Glasgow: University of Glasgow.

Church, M.A., Elliot, A.J. & Gable, S.L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology*, *93*(1), 43-54.

Forster, M., Smith, D.P. & Wild, C.J. (2005). Teaching students to write about statistics. In *Proceedings of IASE Conference on Statistics Education and the Communication of Statistics,* Sydney, Australia. Auckland: University of Auckland.

Gal, I. & Garfield, J. (1997). *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Hayden, R. (1989). Using writing to improve student learning of statistics. *Writing Across the Curriculum, 1*(1), 3-9.

Radke-Sharpe, N. (1991). Writing as a component of statistics education. *The American Statistician, 45*(4), 292-293.

Singer, J.M. As good as it gets: challenges in teaching applied statistics. *International Conference on Teaching Statistics – ICOTS 7, Salvador, Bahia, Brazil*. Brazil: Universidale de Sao Paulo.

Smith Jr., E.V. (2006). Developing and validating multiple-choice test items. *Applied Psychological Measurement*, *30*(1), 69 - 71

Wild, C.J. (1994). Embracing the "wider view" of statistics. *The American Statistician, 48*(2), 163-171.

Wild, C.J., Triggs, C. & Pfannkuch, M. (1997) *Assessment on a Budget: Using Traditional Methods Imaginatively* (pp. 205-220). In Gal, I. & Garfield, J.B. (Eds) *The assessment challenge in statistics education*. The Netherlands: IOS Press.