STATS2: AN APPLIED STATISTICS MODELING COURSE

Jeffrey A Witmer
Oberlin College, United States of America
Jeff.Witmer@oberlin.edu

The typical Stats1 introductory course at the tertiary level covers one-sample and two-sample inference and ends with regression or perhaps one-way ANOVA. We propose that a second course in statistics be built around the idea of statistical modeling ("data = model + error"), beginning with a review of simple linear regression and continuing through two-way ANOVA and logistic regression. Unlike the situation in Stat1, students in Stat2 require access to powerful and flexible computing, which suggests that R be used to fit models. We discuss a Stat2 course that includes both traditional, normal-theory based, inference and randomization tests.

INTRODUCTION

The introductory, one semester course in statistics that students take while in college–let's call it Stat1–has a fairly well accepted syllabus (at least in the United States): some coverage of descriptive statistics and exploratory data analysis; some discussion of data collection; perhaps some probability, particularly the normal and binomial distributions; confidence intervals, for means and for proportions, for one and for two samples; hypothesis testing, for means and for proportions, for one and for two samples; regression and correlation; perhaps chi-square tests; perhaps analysis of variance. It wasn't always the case that statisticians agreed on this syllabus, and there is still some debate, and some teachers cover more than others cover, but Stat1 is reasonably consistent from one campus to another. Moreover, the Advanced Placement exam in statistics tests student mastery of this syllabus and a good score is widely accepted at US college and universities as a substitute for taking the local Stat1 course.

The second course–Stat2–is another story. Some colleges do not offer a Stat2 course, while those that do offer Stat2 do it in many ways. On some campuses Stat2 is a regression course. For others it is an ANOVA course. For some it is a course on categorical data analysis, or perhaps on nonparametric methods, while others present design of experiments. Often there is a textbook chosen that focuses on the selected topic (e.g., regression), but some teachers simply continue with the Stat1 textbook and use the second semester to push farther into the material than was possible in a single semester.

A NEW COURSE

I am part of SLAW[1]: the Statistics in the Liberal Arts Workshop, a group that has been meeting for over 20 years. Together we have developed a somewhat different Stat2 course, along with text and computing materials to support the course. Rather than thinking about a set of topics, we are focused on how statisticians function as they work with clients and analyze data. Stat1 covers a great many important ideas, but typically the concept of a model is beneath the surface, if its presence can be felt at all. The Stat2 course that we have developed takes modeling as its theme, placing the course above Stat1 in sophistication, but below Mathematical Statistics in theoretical complexity. We have nothing against calculus, probability theory, and mathematical statistics–indeed, we enjoy teaching those subjects–but Stat2 is a course for students who got their feet wet in Stat1 and want to dive into the subject without first acquiring the underwater breathing apparatus needed for deep sea diving.

The use of computers and flexible software is essential to the application of modeling ideas. Indeed, computing power allows us to include logistic regression rather easily, which would not have been possible in the past. We use software to make graphs, to fit models, to help assess those models, and to use the models in making predictions. We have settled on Minitab (which is fairly popular and is easy to use) and R (which is free and powerful, although harder to use); other software choices certainly are possible. We also use software to help with some optional enhancements, such as bootstrapping and randomization tests. These augment, rather than replace, our modeling and normal-theory based approach, but we want to expose students to such ideas along the way.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

The mantra Choose, Fit, Assess, Use is repeated throughout the course. First we consider the nature of the data and choose a modeling family (e.g., regression). Then we fit a model and assess the fit (e.g., using residual plots). Next, we modify the model as needed (e.g., by transforming a variable or adding or deleting terms from the model). Finally, we use the fitted model to summarize the situation and to make predictions.

WHAT TO COVER?

When using a model we want to explain (or "model") a response variable by using information typically recorded in one or more predictor or explanatory variables. Table 1 shows the possibilities.

	Predictor Variable(s)	
Response Variable	Quantitative	Categorical
Quantitative	(1) Regression	(2) ANOVA
Categorical	(3) Logistic Regression	(4) Chi-square

Table 1. Possibilities for predictor or explanatory variables

Stat1 usually covers cells (1) and (4) at some depth, cell (2) lightly, and cell (3) not at all. By the end of Stat2 we want students to feel comfortable in each of the four cells. Moreover, we want to allow for multiple predictors in each case, so that the basic idea of, say, logistic regression can be expanded to allow the user to build and use models of growing complexity.

The three major themes for the course correspond to cells (1), (2), and (3), starting with the simple case of a single predictor, moving on to the case of multiple predictors, and including some optional material (e.g., bootstrapping) that some instructors will want to cover and some will omit. We envision many paths through the material, with the default path being (1), (2), (3), but some will prefer (1), (3), (2), for example. Some will want to spend a lot of time on ANOVA while others will want to spend more time on logistic regression. We think that's fine, but we do want the student to see that models can be fit and used in each of the four cells.

WHERE TO START?

We see overlap between Stat1 and Stat2 and expect that the Stat2 course will begin with some review, in part so that students can refresh their knowledge of older material, in part to assure that students who took different versions of Stat1 are "on the same page," and in part so that the concept of statistical modeling can be introduced in a familiar setting. We begin Stat2 with a review of the two-sample t-test, presenting it as a simple model:

$$Y = \mu_i + e$$

where μ_i is the population mean for the ith group and $e \sim N(0, \sigma_i)$ is the random error term.

With two groups this model becomes:

$Y = \mu_1 + e \sim N(\mu_1, \sigma_1)$ for individuals in the first group.

$Y = \mu_2 + e \sim N(\mu_2, \sigma_2)$ for individuals in the second group.

We fit the model by using each sample mean as an estimate of the corresponding population mean and assess the fit by computing residuals and examining them. We consider the simple (reduced) model in which $\mu_1 = \mu_2$ and conduct the usual t-test to see whether this model is sufficient. Again, graphs such as normal probability plots help in assessing the appropriateness of the model. We complete this excursion into Stat1 land by using the chosen model to predict future observations.

WHICH PATH TO TAKE?

One route through the material is to start with a review of two-sample inference, then to cover simple linear regression, followed by multiple regression, then to cover one-way and two-way analysis of variance, and to end with logistic regression (both simple and multiple). We find it helpful to show the connection that a two-sample t-test is a special case of ANOVA, which can also be conducted by fitting a regression model with an indicator variable for group membership. We also like to take the occasional diversion into bootstrapping, for example. But others will prefer to talk about logistic regression immediately following multiple regression. Some may wish to handle the single predictor setting for regression, ANOVA, and logistic regression before working with multiple predictors.

CONCLUSION

We are excited about the growth of interest in statistics over the past several years and have responded by creating a second course that we believe shows students a feature of the discipline–the power of modeling–that may have been missing in their first course. While we recognize that there are many good Stat2courses available, we hope that the course that we have developed will help educators as they consider how best to present the evolving profession of statistics to undergraduates.

ACKNOWLEDGEMENT

[1] SLAW members working on this project are Tom Moore, Grinnell College; Robin Lock, St. Lawrence University; George Cobb, Mt. Holyoke College; Allan Rossman, Cal Poly San Luis Obispo; Brad Hartlaub, Kenyon College; Julie Legler, St. Olaf College; Ann Cannon, Cornell College; and Jeff Witmer, Oberlin College.