A Second Course in Statistics

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

Abstract

The typical introductory statistics course at the tertiary level covers one-sample and two-sample inference and ends with regression or perhaps one-way ANOVA. Although a bit of modeling appears in the regression and ANOVA units, the typical introductory course does not have modeling as a unifying theme. My colleagues and I have developed a second course in statistics (and have written a book for that course) that is built around the idea of statistical modeling ("data = model + error"). Our course begins with a review of simple linear regression and continues through two-way ANOVA and multiple logistic regression. We use R to create graphs, to fit models, and to make traditional, normal- theory based inferences. We also use computing power to conduct randomization tests and to create bootstrap intervals.

Keywords: modeling, regression

1. INTRODUCTION

Introductory statistics is typically a one-semester course for undergraduate students. This course, which I'll refer to as Stat1, has a reasonably well accepted syllabus (at least in the United States). I handle transfer credit requests for statistics courses for my school and what I see over and over again is that Stat1 includes some coverage of descriptive statistics and exploratory data analysis, some discussion of sampling and other data collection methods, some probability (particularly the normal and binomial distributions), confidence intervals for means and for proportions (usually for one and for two samples), a fair amount of hypothesis testing, for means and for proportions (again, usually for one and for two samples), and some treatment of correlation and simple linear regression. Some Stat1 courses include analysis of variance and some touch on two-way ANOVA. Other topics vary from course to course, but there is very high overlap among courses which means that but Stat1 is reasonably consistent from one university or college to another, with differences between versions of Stat1 being primary due to how much is covered (i.e., how many chapters of the textbook are covered). The Advanced Placement examination in statistics tests students on the commonly covered topics and a good AP score is widely accepted at US colleges and universities as a substitute for Stat1 course.

There might be agreement on what to teach in Stat1, but there has not been such agreement on the second course, Stat2. Not long ago it was a rather uncommon to find a Stat2 course at a US college or university and where there was a clear Stat2 follow-up to Stat1 the second course was handled differently from school to school. On some campuses Stat2 was, and still is, a regression course. For others it is an ANOVA course. Some statisticians prefer to follow Stat1 with a course on nonparametric methods, while others teach design of experiments. Often there is a textbook chosen that focuses on the selected topic (e.g., regression), but some faculty continue with the Stat1 textbook and use the second semester to push farther into the material than was possible in a single semester.

2. STAT2: A NEW COURSE

I am a member of SLAW[1]: the Statistics in the Liberal Arts Workshop, a group that has been meeting for about 25 years to discuss statistics education for undergraduates at small colleges. After many years of discussion of the question "What should a Stat2 course look like?" we have developed a somewhat different Stat2 course, along with text and other materials to support the course. Rather than thinking about a set of topics, our course is 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 not discussed, let alone emphasized, in Stat1 except perhaps when regression is presented. The Stat2 course that we have developed takes modeling as its theme. This sets Stat2 above Stat1 in sophistication, but below Mathematical Statistics in theoretical complexity. It is common to find a Mathematical Statistics course offered at a college as a deeper exposure to statistics then is given in Stat1. However, Mathematical Statistics follows a semester of probability theory, which follows multivariable calculus. In short, although Math Stat is a great course, it is not an appropriate second course for the typical student who has taken Stat1. We have nothing against calculus, probability theory, and mathematical statistics – indeed, we enjoy teaching those important 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. Computing power has changed the practice of statistics – and arguably should lead to greater changes in Stat1, with permutation tests supplanting t tests, but that's the topic of another paper. We can now fit logistic regression models quite easily, which was not feasible in the past. In our Stat2 course we use software to make graphs, to fit models, to help assess those models, and to use the fitted models in making predictions. We use R – which is free, powerful, and increasingly popular – as one option for Stat2; we present Minitab – which is also popular and is easy to use – as an alternative. The use of software is integral to our course and thus we have written an R Companion and a Minitab Companion to accompany our textbook, since we expect students to be using software pretty much every day for the entire semester.

We repeat the slogan "Choose, Fit, Assess, and Use" throughout Stat2. First we consider the nature of the research question that led to a dataset and we choose a modeling family (e.g., regression) accordingly. Then we fit a model and assess the fit (e.g., using residual plots and other diagnostics). If the assessment of the model shows a problem then we modify the model as needed (e.g., by transforming a variable or perhaps by adding a quadratic term to the model). After we have a satisfactory model we use it to summarize the data and to make predictions.

3. TOPICS FOR STAT2

When using a model we want to explain (or "model") a response variable by using information that has been measured on one or more predictor (or explanatory) variables. The table below shows four possibilities for combinations of the predictor(s) and of the response.

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

Stat1 usually covers cell (1) and often covers cell (2) or cell (4) or both. However, cell (3) is rarely touched upon in Stat1, although this is not due to a lack of importance of logistic regression in practice. By the end of Stat2 we want students to feel comfortable analyzing data in each of the four cells. Moreover, we want to allow for multiple predictors in each case, so that simple linear regression is expanded to multiple regression, simple logistic regression and multiple logistic regression are covered, and both one-way and two-way analysis of variance are covered. In short, we want students to be able to build and to use models at a level of complexity that they will see in real-world applications across many disciplines.

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. (Note that cell (4) can be considered a special case of cell (3).)

4. 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 and in part to assure that students who took different versions of Stat1 have a common understanding of notation and terminology. We also like to use this review period to introduce software so that, for example, students who have never used R can gain some experience with it while working with familiar material. We also want to help students become comfortable with the concept of statistical modeling in a familiar setting. Thus, 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 reduced model in which $\mu_1 = \mu_2$ and conduct the standard two-sample t-test to judge whether this simple model is sufficient. We use graphs such as normal probability plots to help assess the appropriateness of the model. We complete this review of Stat1 by using the chosen model to predict future observations. Note that prediction is not commonly presented in Stat1 except in the regression chapter, perhaps because models are beneath the surface when the Stat1 course covers t-tests.

5. MANY PATHS

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). Indeed, this is how our textbook is organized, which corresponds to moving along the path (1), (2), (3) in the previous table. 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. Rather than the path (1), (2), (3), some of us prefer to talk about logistic regression immediately following multiple regression. Some of us like to present nonparametric methods when discussion ANOVA or cross-validation and added variable plots when discussing regression. These and other optional topics are sprinkled into the textbook so that they can easily be added to the course if the instructor wants to cover them.

Thus, we present more than one path through the material, with the textbook being organized along path (1), (2), (3) but allowing for the choice of path (1), (3), (2) and for other possibilities. (As another possibility, some instructors may wish to handle the single predictor setting for regression, ANOVA, and logistic regression before working with multiple predictors.) Some instructors prefer 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.

We also use software to help with some optional enhancements, including bootstrapping and randomization tests. These augment, rather than replace, our modeling and normal-theory based approach. We don't expect that every instructor will want to cover these topics, which appear as optional sections in our book, but we to give instructors the option of presenting these

tools so that students can be exposed to computer-intensive data-analysis methods that are becoming more common in practice.

6. CONCLUSION

We have seen growth in the number of students who are interested in statistics and who want to take a course that goes beyond the usual Stat1 material. After years of discussion and debate we have created a second course that we believe shows students a feature of our discipline – the power of modeling – that may have been missing in their first course. We recognize that many colleges and universities have been offering good Stat2 courses for many years, but these have varied greatly from institution to institution. We hope that the course that we have developed will help educators as they consider how best to present the growing and evolving field of statistics to undergraduates.

1. SLAW members who worked 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.