An Active Learning Approach to Teach Advanced Multi-predictor Modeling Concepts to Clinicians

Gregory P. Samsa
Laine Thomas
Linda S. Lee
Duke University

Edward M. Neal
University of North Carolina at Chapel Hill

Journal of Statistics Education Volume 20, Number 1 (2012),
www.amstat.org/publications/jse/v20n1/samsa.pdf

Key Words: Active learning; Statistics education; Graduate education; Deconstructing the disciplines; Multi-predictor modeling; Non-statisticians.

Abstract

Clinicians have characteristics – high scientific maturity, low tolerance for symbol manipulation and programming, limited time outside of class – that limit the effectiveness of traditional methods for teaching multi-predictor modeling. We describe an active-learning-based approach that shows particular promise for accommodating these characteristics.

1. Introduction

1.1 Setting

The Duke Clinical Research Training Program (CRTP) is a professional masters' degree program for biomedical researchers (Wilkinson & Oddone, 2002). The curriculum encompasses topics fundamental to clinical investigation including biostatistics, epidemiology, clinical trials, clinical pharmacology, genomics, and comparative effectiveness research, among others. Most students are physicians who are beginning careers in academic medicine; classes are designed with this audience in mind. In addition, a small number of researchers from allied health

professions as well as third-year medical students participate as degree and non-degree students. Students from the National Institutes of Health take CRTP classes remotely and in real time, participating through interactive videoconferencing. Since its inception in 1986 as the Biometry Training Program, the program has awarded over 300 degrees.

CRP245, titled Statistical Analysis, is the second of two core courses in statistical techniques. Topics include the analysis of variance, analysis of covariance, logistic regression, and Cox regression for survival analysis. The goal is to teach principles and techniques of multi-predictor modeling, including selection of an analysis strategy, validation, reporting, interpretation and presentation. Although the course topics are statistically sophisticated, training our students to become statisticians is neither feasible nor desired, a view that is consistent with other programs for similar audiences (Ambrosius & Manatunga 2002; Deutsch, 2002; Supino & Borer, 2007).

Two of the authors (GS, LT) are biostatisticians that teach CRP245. The other two authors (LL, EN) are educational specialists. LL is responsible for, among others, program evaluation. EN serves as a curriculum consultant.

1.2 The case study

This case study describes our experiences in redesigning CRP245 from a traditional to an active-learning-based format.

Given the up-front investment required to redesign a course, the discerning reader might reasonably ask what motivated two busy faculty members to even attempt such a thing. In our case, we were motivated by two observations that we eventually realized were related.

First, despite the presumption that, as dedicated teachers, we should be becoming better at teaching, not worse, we observed that our course lectures were becoming less and less well-received. This observation was based on course evaluations, a decrease in attendance, and a decrease in the quality of the interactions between students and instructors during class. After discussing the matter with colleagues, it became clear that this phenomenon was not limited to us, but instead reflected a shift in student expectations about the educational experience. Put simply: our students, experienced with various technologies that have now become alternatives to transmitting information through lecture (Updike, 2011), now insist that the lecture format be limited to those circumstances where lecture is likely to be the most effective mode of instruction. Moreover, when students perceive lecture to be an ineffective mode of instruction, they do not hesitate to vote with their feet.

The second observation was that our course had not been fully successful in teaching various integrative skills associated with statistical modeling. We viewed this to be a fundamental instructional design problem that undermined achievement of our primary purpose for teaching the course.

Our epiphany occurred when we connected these ideas and asked: Could a different approach to teaching also serve to address the fundamental problem in course design? Fortunately, at least in our setting, the answer appears to be “yes.”

The aims of this paper are twofold. First, we describe the path taken to improve the learning experience of our students by "deconstructing" elements of our discipline relative to the characteristics of our students. Second, we hope to motivate further exploration of active learning strategies in more advanced (as opposed to introductory) statistics courses for non-statistician graduate students.

2. Course design considerations

2. 1 Student characteristics

Our students share characteristics that differentiate them from other graduate students, most notably from graduate students within the discipline of statistics. Our students are exceptionally mature scientifically. All are already consumers of the medical literature and, indeed, many enter our class having already contributed to multiple scientific publications. Our students are also characterized by a highly variable statistical background, some having substantial coursework or applied training and others having much less exposure to statistics.

While acknowledging that one can't be accepted into medical school without high standardized test scores in mathematics, it is nevertheless the case that, relative to their scientific maturity, our students are weaker in mathematics. Because they have not regularly practiced mathematics for some time, we cannot rely on the use of formulae and symbol manipulation as would be the case for similarly talented students that specialize in a mathematical discipline. Consistent with a lack of interest in symbol manipulation, our students are not, in general, skilled programmers. Moreover, they are very busy with clinical responsibilities, and their time for work outside class is limited.

The overall implication of these characteristics is that, as PhD-trained statisticians (GS, LT), we cannot expect our students to learn about multi-predictor modeling in the way that we did. Attempting to teach them in that manner would not be realistic, effective or humane.

2.2 The core design problem

In attempting to define the core problem in redesigning our course, we began with a consistent suggestion for improvement that we received, over the years, on our end-of-course evaluations. We believe that it is a fundamental difficulty that is encountered when attempting to teach advanced statistical concepts to students from non-mathematical disciplines. To paraphrase one of many such comments:

> *"I learned a ton of things in this course and can now more or less follow an analysis protocol if a statistician tells me what to do. However, I'm not confident that I can select what to do on my own, nor am I confident that I can always put my results in the proper perspective."*

In performing our due diligence, we found many resources for teaching multi-predictor modeling to students within the discipline of statistics. However, these either treated the subject

symbolically (e.g., using matrix algebra), or implicitly assumed that the reader possessed "statistical culture" (i.e., experience in thinking about modeling like a statistician, with all the background knowledge and experience that this implies). Thus, these resources were unsuitable for teaching multi-predictor modeling to clinicians.

In addition, we noted the enthusiasm around the concept of active learning. "Active learning" is a general term that refers to pedagogical techniques that require students to take an active role in their own learning in the classroom. There is strong empirical evidence that active learning methods improve student learning, especially their mastery of problem solving and critical thinking skills (Prince, 2004; Michael, 2006). Publications in the mathematically-related disciplines provide many examples of the practical application of these methods in undergraduate and graduate courses (Felder, 1990; Van Heuvelen, 1991; Paulson, 1999).

More specifically, we discovered that there are many active-learning-based resources for teaching *introductory* statistics to non-statisticians (Steinhorst & Keeler, 1995; Bradstreet, 1996; Chance, 2002; Bland, 2004; Enders & Diener-West, 2006; Supino & Borer, 2007; Carlson & Winquist, 2011). However, multi-predictor modeling is not only more complex but is also a substantially different type of topic than those covered in introductory statistics courses. It soon became apparent that the available active-learning-based approaches for teaching basic statistics to clinicians and other non-statisticians would not necessarily extend from introductory statistics to multi-predictor modeling.

Nevertheless, our due diligence did serve to clarify the redesign task. We were now in a position to ask: *Could we design a course that is as intuitive to clinicians as an introductory course in statistics, while covering the more complex topic of multi-predictor modeling?* In other words: *Could we extend "T-tests for clinicians: an intuitive approach", something that educators have discovered how to effectively teach, into "Multi-predictor modeling for clinicians: an intuitive approach", something that educators have not?*

Having made a commitment to explore active-based learning, we also asked: *How could principles of active-based learning be applied to students with a set of characteristics (e.g., talented, scientifically mature, little time, little tolerance for formulae, little skill in programming) that precluded the usual methods of teaching multi-predictor modeling?*

3. Course design

3.1 Course goals

With the core design problem in mind, we reformulated the overall course goal as follows: *The overall course goal is to develop an effective working knowledge of common statistical techniques for models involving one or more predictors.* The relevant paragraph from the course syllabus provides further clarification:

> We don't propose to teach you everything there is to know about multi-predictor modeling, but we will teach you enough to form sound conclusions from most data sets. Moreover, we want you to "really know what you know." Operationally, this means that

> you can select, perform, and interpret the analyses without assistance from a statistician. In practice, this might mean performing an initial analysis yourself, and then consulting with a statistician about the finer points. Nevertheless, we have framed the overall course goal as if you were performing all analyses yourself.

During the first class session, we verified that our students believed this to be an accurate statement of their (perhaps initially unconscious) learning objectives.

One of the implications of reframing the course goal in this fashion was to raise the stakes. In particular, by explicitly tasking the students with the responsibility to select an appropriate statistical analysis we were asking them to achieve something that our previous students had been systematically unable to do. As instructors, we set ourselves the challenge of meeting a goal that was different and more substantial.

Reflecting upon these course goals, we subsequently recognized that "knowing what you know" doesn't necessarily imply "selecting, performing and interpreting analyses without a statistician". Indeed, this latter phrase would be better stated as "with limited assistance or review from a statistician". This is consistent with our teaching philosophy, which is to empower our students as much as is realistically feasible, while simultaneously making them aware of the limits of their knowledge. A restatement of the course goals is as follows:

> We don't propose to teach you everything there is to know about multi-predictor modeling, but we will teach you enough to form sound conclusions from most data sets. Moreover, we want you to "really know what you know." Operationally, this means that you can select, perform, and interpret the analyses – perhaps on your own and perhaps with assistance from a statistician. Moreover, you should know when to ask for input and/or assistance from a statistician and you should be able to communicate with statistical literacy. In practice, this might mean performing an initial analysis yourself, and then consulting with a statistician about the finer points. Nevertheless, for clarity of exposition we have framed the overall course goal as if you were performing the analyses yourself.

3.2 Some design principles

Although the preceding discussion has focused on the challenges that we faced, we also had various resources to bring to bear. Our students are bright, scientifically mature, hard-working and motivated to learn the material. Over the years, we had accumulated experience in understanding their characteristics. And, we had the bounty of having an experienced educator and master teacher (EN) who served as a consultant during the process of course redesign. Thus, we had access to current thinking about educational pedagogy; for example, to such concepts as "backward design (Wiggins & McTighe, 2005)," "active-learning," "evaluation and assessment (Kellaghan & Stufflebeam, 2003; Rossi, Lipsey, & Freeman, 2004)," "learning outcomes (Harden, Davis, & Friedman, 1999)," and "constructivism (Phillips, 1995; Fosnot, 1996)". The instructors' operational interpretation of some of the key educational principles relevant to the course redesign was as follows:

- We should clearly define what we want our students to be able to do -- rather than what we want our students to know.
- Learning should be active -- students must practice the behaviors in question in order to master them.
- Feedback to students should be rapid -- and need not always be provided by the instructors.
- Not everything has to be graded -- non-graded assessment should be regular, while graded evaluation using circulated rubrics can occur later in the learning process.
- Evaluation should be integral to course design -- not an afterthought.
- The course syllabus should be used as an organizing document -- discussing course goals, course organization and typical teaching methods in as much detail as possible so as to encourage transparency.

In retrospect, many of the above ideas seem to be no more than common sense – although not so "common" as to have been prominent in our own experiences as students. These ideas did, however, serve to reframe our thinking from "What ***topics***, in general, should we ***teach***?" to "What ***tasks***, specifically, should our students ***learn*** to accomplish?"

3.3 Overall course design

Our solution to the core design problem (developing an intuitive approach to learning multi-predictor modeling) was first to organize the course around the following topic-specific modules:

- 1-way ANOVA
- 2-way ANOVA
- Simple (i.e., 1-predictor) linear regression & ANCOVA
- Pre-post designs
- Multiple predictors
- Repeated measures and longitudinal data
- Logistic regression
- Survival analysis

Labeling the pre-class student preparation as "day 0" and the post-class student assignment as "day 4", each of these topic-specific modules typically required three in-class days (i.e., 90-minute class sessions) to cover and included the following segments:

- Day 0 (pre-class): Pre-class preparation
- Day 1 (in-class): Review and demonstration
- Day 2 (in-class): Directed data analysis
- Day 3 (in-class): Integrative ("stretch") assignment
- Day 4 (post-class): Reflection reports

We attempted to apply our understanding of the above pedagogical principles to the design of each "day" of the modules.

The course concluded with a final exam consisting of an analysis of a real data set reported in the format of a scientific manuscript.

3.4 Module day 0 – pre-class student preparation

Day 0 Overview: Traditional college and professional school teaching assumes text or journal readings prior to each individual class on a particular topic or concept. Given our students' growing aversion to our classroom lectures and their experience with alternative technologies as noted in the introduction, we recorded and archived video mini-lectures of five to ten minutes each, and requested that students view them before the first class (i.e., "day 1") of each module. For example, titles of the mini-lectures for the 1-way analysis of variance module were:

- Overview and summary
- Presenting model inputs
- Fixed versus random effects
- Comparisons and contrasts
- Multiple comparisons techniques
- Implementation using indicator variables
- The ANOVA table
- Presenting the results of an ANOVA

The total pre-class preparation time for the students was approximately one hour, and students could view the mini-lectures on their computers through links posted in the Blackboard course website at their leisure.

Instructor Preparation: We prepared our mini-lectures using a software program, Camtasia®, to record their audio portion and to integrate it with prepared PowerPoint® slides. Other CRTP faculty members have used different software to produce similar recordings and software tutorials. We spoke from prepared scripts, which provided reassurance that we had covered everything that we intended. The prepared scripts also assisted us in integrating material across modules. In large part because we removed topics that were of marginal importance, we were able to cover a more focused but still extensive amount of material in approximately half the time.

Making the videos into short segments simplified matters for everyone involved. For the students, this organization made it easy to retrieve information. Indeed, students could tailor their use of the videos to their background – those already familiar with the topic in question could skim the slides and only view selected portions of the videos, while those that found the content to be more challenging could view the videos repeatedly. This is in contrast to a traditional lecture, which risks boring those students with the best level of preparation and confusing those with the worst. For the instructors, this organization made it easy to develop and revise the videos.

Preparation time for the videos was extensive – approximately 15 hours per module. Our hope is that some of this time can be recouped when the videos require revision. When taping the videos, we quickly learned that the perfect is the enemy of the good.

One potential problem with this approach is that, when first encountering the material, students would not be able to ask questions of the instructors in real time. We ameliorated this difficulty somewhat by being available by telephone and email to answer questions but -- practically speaking -- in order to make this arrangement work it was critical that the videos be exceptionally clear when viewed as self-contained entities.

Student response: Student receptivity to the mini-lectures as reported in our end-of-course evaluations was generally positive. In particular, there were few reported concerns about the absence of an instructor when first encountering the material.

Indeed, in their reflection reports students shared various strategies for utilizing the videos, of which the following seemed particularly appealing and were recommended to the class as a whole:

- View the lectures before class in order to obtain a "birds-eye" view of the material.
- Write down questions and, if not answered during the day 1 demonstration (see below) ask those questions during the demonstration.
- If unable to attend the day 2 and 3 classes (see below), refer to the lectures to help complete the day 2 and 3 assignments.
- After day 3, view selected lectures again, in order to place the information into better context.
- Email the instructors with any final questions.

3.5 Module day 1 – review and demonstration

Day 1 Overview: Day 1 is a review and software demonstration. Typically, we analyzed a real data set, thus not only demonstrating how to use statistical software but also reviewing salient points about the analytical methods within their usual context. Questions were encouraged.

Instructor preparation: In designing the day 1 class sessions we believed that we needed to review some of the material on the videos, especially since it is not realistic to expect that all of the students would have viewed them before class. On the other hand, a completely comprehensive review would duplicate the content of the videos and thus discourage students from performing the desired pre-class preparation. Moreover, we worried that the traditional review/summary format would be uninteresting, and also worried that a software demonstration that only focused on how to use the software would be out of context and thus ineffective. Accordingly, we made every effort to integrate the review of content with the demonstration of the software.

Preparation time was modest to moderate – approximately 5 hours per module. Much of this preparation time involved translating our usual analysis protocols from the software packages that we typically use to a software package with which our students were familiar.

Student response: Students were generally, although not unanimously, engaged by this format. Some believed the demonstrations to be too simplistic. According to our qualitative assessment,

student questions were usually on target. For example, some of the student questions within the 1-way analysis of variance module were:

- How extreme does an outlier have to be for me to worry?
- How do I decide whether I want to look at contrasts that are more complex than simple pair-wise comparisons?
- Should I always perform separate analyses of a data set with and without outliers?
- When is the usual 1-way analysis of variance robust to violation of the assumption of normality?
- How do I demonstrate that a statistically significant result is clinically significant?
- How do I derive the inputs to the power calculation?

In response to initial student comments that reflected varied level of confidence about being above to navigate the software, during the first few modules we placed a heavy emphasis on technical and mechanical issues, in order to verify students could successfully perform the requested data analyses. Over time, the content of the demonstration day became less focused on mechanical issues, and more focused on issues of strategy and interpretation as per the course goals.

3.6 Module day 2 – directed data analysis

Day 2 Overview: The day 2 sessions are organized around a straightforward data analysis. Students were encouraged to work on the assignment in groups of two. The assignment usually began with some factual questions, intended to clarify some of the concepts covered in the videos. The instructors circulated among the students while they discussed the questions. (We circulated physically for the Duke students located in Durham, and used an audio link to check in with the NIH students located in Bethesda.) We planned to bring the class together to de-brief at least every 20 minutes or so, but would do so sooner if we discovered that a question was causing particular problems.

The next component of the assignment was a directed data analysis, which was intended to serve as a model for a "standard analytic protocol". Questions pertained to reading numbers off the computer output, reporting and interpretation. See Appendix B for the day 2 assignment for the 1-way analysis of variance module.

We debriefed as a class at the conclusion of the directed data analysis. Answers were posted on the course website at the conclusion of the exercise.

Instructor preparation: In preparing the day 2 exercises, we were often able to make use of assignments taken from the previous iteration of the class. In the absence of such materials, the time required to prepare the day 2 exercises would probably have been moderate to heavy, depending on the degree of integration across modules that was desired. This might turn 2-5 hours of preparation into 5-10 hours.

In response to student reaction to the first modules, we revised the exercises so that each group of questions began with an introduction describing their purpose. We quickly discovered that doing so helped to clarify the assignment, and also to better direct our students' efforts.

We also discovered that the day 2 exercises tended to take longer than we expected, and we sometimes had to extend these exercises into the following class period. Often, the delays involved the technical mechanics of downloading the data and using the statistical software to perform analyses. Because we want to have more of a focus on concepts than on mechanics, we are actively pursuing ways to make our software instruction more streamlined and more effective.

Student response: Our students commented on how much more they learned by having to execute specifics, rather than just passively listening to a lecture. Breaking the analytic task into small steps isolated what the students did and did not functionally understand. Often, when questions arose they could be answered immediately either by other students or by the circulating instructors. Students emphasized the importance of debriefing regularly, to provide reassurance that they were on the right track, rather than allowing extended periods of self-guided work. It was particularly important to hold regular debriefings with the NIH students in Bethesda, in order to compensate for the inability of the instructors to circulate among these students physically.

3.7 Module day 3 – integrative assignment

Day 3 Overview: Day 3 is a "stretch assignment" that encourages students to extend what they have learned, apply what they learned in a new context, and integrate what they have learned in a new or more complicated context. Some topics for day 3 assignments included a critical review of various journal articles that claimed to "validate" multi-predictor models, performing a sample size calculation including the derivation of its inputs, writing a statistical methods section of a grant, working through various study design issues such as the selection of the study groups and the analytical variables, and so forth. See Appendix C for the day 3 assignment for the 1-way analysis of variance module.

Day 3 assignments varied quite a bit from module to module, and were often based on issues of study design, analysis strategy and/or interpretation that the instructors were facing in their actual practice of statistics. We encouraged students to form teams of 4-5 to discuss the questions. The layout of the Durham site classroom is more conducive to triads, which is how many of the students arranged themselves. Because the day 3 questions were so substantial, we held debriefings more regularly than we did during the day 2 exercises. Although there weren't always "correct" answers to each of the questions, we did post comments (based on the class discussion) on the course website after the exercise.

Instructor preparation: The preparation time for the day 3 assignments was heavy (at least 8 hours per module) and, indeed, this is perhaps the element of the class that we expect to expend the most effort on between now and the next time CRP245 is offered. It turned out that making the day 3 sessions effective rather than chaotic required that two criteria be met.

The first criterion for an effective session is that the exercises had to have a clearly stated educational objective. The groundwork for determining that objective was laid during our initial discussions about the topics that were important enough to include in the videos. The objective-setting process continued for the day 1 and 2 sessions, although usually in abbreviated fashion as the objectives for these days tended to be relatively straightforward and consistent across modules. For day 3, however, we had many potentially interesting examples from our own practice of statistics, but on reflection these examples had varying amounts of pedagogic value. Clarifying what we wanted our students to get out of the day 3 sessions was critical to which topics to ultimately include, and also how those topics could be best presented.

The second criterion for an effective session was the proper use of facilitation techniques. These are important to utilize when circulating among individual students, and also when managing the group-level debriefing and discussion. We recommend that some form of training in facilitation techniques, whether through coursework, mentoring, self-study, or another method, is definitely worth the time for those that plan to teach active-learning-based courses. These techniques are extensively documented in the literature on college teaching and many of these sources provide excellent guidelines for practical implementation (Johnson & Johnson, 1999; Tiberius, 1999; Davis, 2009). For example, one effective technique when responding to a student question is to help identify the underlying statistical principle involved (ideally, the instructor assists students to do this on their own), restate the principle in terms that are understandable to the student, propose how that principle can be applied, and then encourage the student to suggest a specific course of action.

We initially underestimated the difficulty of the day 3 sessions, but over time learned to develop assignments that were more pedagogically focused and within our students' capabilities.

Student response: Students were highly engaged with this exercise, which most closely approximated the high-level tasks that they would have to accomplish in their academic careers. A few of the students were unnerved by not fully understanding the discussion during day 3, by not having a sufficient level of statistical background knowledge to derive all the answers on their own, and by the fact that not all of the questions had a single "correct" answer. We reassured these students that the intellectual stretching in and of itself would be beneficial – and, moreover, that the exercises were partially intended to illustrate how statisticians think about complex statistical issues, with a view towards making the students' subsequent interactions with statisticians more fruitful, rather than to derive a single definitive answer. In retrospect, these exercises also served to acquaint students with the limits of their knowledge, something that was quite consistent with our stated course goals.

One qualitative metric that we used to assess the effectiveness of the day 3 sessions was through the grant proposals and manuscripts on which the students asked us to comment – these topics closely paralleling the content of the day 3 sessions. Over the course of the semester, the statistical sophistication with which students discussed these topics increased markedly. Also, the fact that students solicited our comments was encouraging because it demonstrated that they were recognizing the limits of their knowledge and the importance of collaborating with statisticians.

3.8 Module day 4 – post-class reflection reports

Day 4 Overview: In order to encourage unworried participation from the students, modules were ungraded. Instead, we asked students to submit reflection reports that verified that they had in fact completed the modules; noted whether the work was performed in class or outside class (an informal way of encouraging attendance); asked for comments on the videos, course notes, and assignments; asked for any remaining questions; and invited general observations about the module. Formal evaluation of our students' efforts was postponed until the final examination.

Instructor preparation: No preparation was required, beyond setting the initial guidelines.

Student response: The reflection component of the report served as our formative evaluation, allowing us to make changes to the course in real time. Finally, for grading purposes the reflection report allowed us to officially verify that students had completed their assignments.

We received a large amount of commentary in the reflection reports for the first few modules, almost all quite enthusiastic about active-based learning and some suggesting modest changes to the in-class exercises. After this, the reports mostly settled into reporting that the modules had been completed, with perhaps 5 out of 50 students for any particular module reporting a residual source of confusion or a suggestion for improvement. The instructors responded to each comment by email, and attempted to do so promptly in order to encourage additional suggestions.

Although the reports were intended to be both status reports (e.g., to track completed modules) and reflection reports (e.g., to encourage students to actively reflect on their educational experience), they devolved into the former almost exclusively. We are reconsidering how to better encourage active reflection among our students.

3.9 Final examination

The final examination was an analysis of an actual data set, to be presented in the format of a scientific manuscript. The details were provided early in the semester, and students were invited to begin work on the final examination as soon as they wished. Consistent with actual practice, they could ask questions of other students and the instructors, but were asked to perform the analyses themselves and also to write up the results themselves.

Students were provided with a link to a published manuscript containing the study background and design, and also reporting one potential set of analyses from their data set. However, it was noted for the students that there is no single correct answer and that it is quite appropriate for them to draw different conclusions than did the manuscript's authors. Appendix D illustrates the grading rubric that we provided to the students.

The final examination was intended to evaluate students' abilities to select, implement and interpret the results of a non-trivial statistical analysis using real data. It allowed us to close the loop by assessing the degree to which the student critiques that provided the impetus to redesign the course had been successfully addressed.

4. Course evaluation

We had planned to gather formative evaluation data from our students through the reflection reports at the end of each module to guide real-time course revision. However, we sought additional data sources to triangulate with our subjective observations that we were moving in the desired direction. Additional data sources included our end-of-course evaluation surveys and the final examination.

4.1 Summary of formative data

Our formative evaluation has been discussed previously. Briefly, feedback from the first few modules suggested that students liked the active-learning-based approach, and also suggested various mid-course corrections. Some of the student comments included personal encouragement, which were greatly appreciated and helped to maintain the energy of the instructors.

4.2 Summary of end-of-course data

Table 1 below presents relevant items from the redesign year and the year immediately preceding it. In view of the moderate power induced by the sample size, we were less concerned with the formal statistical significance of individual items, and more concerned with interpreting the overall pattern of results in order to help improve the course for next year.

Overall satisfaction with the course increased from 77% to 84%. Moreover, the six items which were our primary focus all improved in absolute terms from 12-18%. Table 1 also provides hypotheses about why the satisfaction scores for these six items improved – for example, we hypothesize that course organization improved in the redesign year because it was now discussed in the syllabus more clearly and in more detail.

Of particular note is the item on the relevance of the course to students' career goals. Although we hoped that satisfaction with this item would increase, this has historically been a sticking point for statistics courses within the CRTP. We were greatly encouraged to observe that satisfaction on this item improved from 74% to 92%.

Table 1: Total student satisfaction before and after course redesign

Total Satisfaction Scores (Somewhat Satisfied plus Very Satisfied)	Baseline Year (N=35)	Redesign Year (N=37)	Possible reason for change
Response rate to course survey	74%	70%	
Faculty availability for questions and feedback	89%	89%	Not expected to change
Course organization	71%	89%	Syllabus used as an organizing document
Relevance of activities to course goals	83%	97%	Explicit justification of course activities in syllabus and in modules
Integration of course with others in curriculum	74%	86%	Syllabus explicitly discusses integration with other core statistics course
Effectiveness of technologies used in presentations	77%	92%	Use of videotaped lectures
Course resources	69%	86%	Use of videotaped lectures
Relevance of course to career goals	74%	92%	Revised course goal provides rationale for developing integrative understanding of multi-predictor modeling
Overall satisfaction with course	77%	84%	Impact of above aspects of course redesign

In addition, students responded favorably in their assessment of the course format. These questions were posed for the first time in the redesign year; comparative data are not available for the previous year.

Course Format Questions – Total Agreement Scores (Somewhat Agree plus Agree Strongly)	Redesign Year (N=37)
The format encouraged me to prepare for class.	78%
The format engaged me in the material.	89%
The format helped me master course concepts.	78%

4.3 Summary of manuscript ratings

Our primary evaluation of student learning and achievement of the overall course goals was a qualitative rating of the quality of the manuscripts produced as part of the final examination. In general, student performance on this examination compared favorably with that of students from previous years.

The quality of the examinations of the third-year medical students in the course was, on average, lower than the rest of the students -- although this phenomenon was not uniform, with some of the manuscripts being notably outstanding. The primary difference between the third-year medical students and the others is their level of experience. We are presently reviewing our course materials to make as explicit as possible things that the other students will have already learned from experience before beginning our class.

5. Discussion

Using active-learning-based techniques to teach multi-predictor modeling has been enjoyable for both instructors and students. It is fortunate that students find the experience enjoyable and engaging, as our expectations of the quality of their work have increased. Moreover, it is fortunate that the instructors have found the experience to be enjoyable as the preparation time, particularly during the transition away from a more traditional lecture-based format, was considerable.

The redesign of CRP245 had numerous risks. Students might not attend the in-class exercises and thus undermine their effectiveness. The active-learning-based learning exercises might be so chaotic as to lead to frustration rather than illumination. Perhaps most worrisome of all, it might not really be possible to teach clinicians the high-level skills of model selection and interpretation that statisticians learn through specialized training and years of experience.

We were not alone in this concern and, indeed, how much clinicians should be empowered to perform their own analyses is an ongoing topic of debate within the community of statisticians that participate in biomedical research (Berwick, Fineberg, & Weinstein, 1981; Mintz & Ostbye, 1992; Windish, Huot, & Green, 2007; Swift, Miles, Price, Shepstone, & Leinster, 2009). This debate can be summarized by competing analogies. One analogy is to open-heart surgery: physicians require years of training and apprenticeship to learn to perform this complex surgery safely. Why should these same physicians, without years of statistical training and apprenticeship, expect to do the same for something as complex as multi-predictor modeling? The competing analogy replaces open-heart surgery with home glucose monitoring. Many diabetic patients can test their own blood sugar and make basic medical management decisions in response – so long as they are provided with sound training and understand the limitations of their knowledge. While acknowledging and in no way minimizing the concerns of the former statisticians, our experience places us into the latter camp.

Our contribution to this debate is the empirical observation that clinicians can in fact learn to apply something approaching statistical intuition to the task of multi-predictor modeling. The primary basis for this observation is our assessment of the quality of the final examinations. While this assessment isn't quantitatively-based, neither is it entirely subjective, as it was derived from the grading rubric of Appendix D.

Indeed, the most common impression of the final examination manuscripts was that as data analysts our students were inexperienced but well-grounded. Much of this grounding appears to have been based on students' abilities to appropriately use simple tools to visualize their data. We recommend that visualization techniques, many of which are technically and conceptually

simpler than full multi-predictor modeling, be a particular area of emphasis when teaching multi-predictor modeling to students outside the discipline of statistics.

Our observation does, however, come with a critical caveat; namely, that "something approaching statistical intuition" is not the same as "statistical intuition". Accordingly, it is crucial that, in addition to learning the techniques of multi-predictor modeling, our students understand the limits of their knowledge. Students tended to encounter these limits in a non-systematic fashion as part of the day 3 exercises, but on reflection there is much to be said for being more systematic in helping them to identify when they are swimming toward the deep end of the pool.

We speculate that what was most critical to the success of the redesigned CRP245 was not the specific form that the active-learning-based approach to teaching used – indeed, another form of course organization might have worked better for our students, and the current form of organization might fail when applied within another context. However, upon reflection and quite by accident, the 3-module format might have been particularly natural for this audience as it parallels the "see one, do one, teach one" model of medical education already familiar to our students. Here, the "see one" translates into the day 1 demonstration. The "do one" translates into the day 2 exercises that allow students to practice their craft on a "straightforward case". The "teach one" translates into the day 3 integrative stretching exercise – especially once it is understood that within the medical context "teaching one" is a common form of integrative application.

We also speculate that one of the factors contributing to the successful redesign of CRP245 was that we had the experience with our target audience to be able to "deconstruct our discipline" (Middendorf & Pace, 2004; Diaz, Middendorf, Pace, & Shopkow, 2008) in terms that were understandable to our students. One of the elements that make this case study noteworthy was the strong connection between the characteristics of our students and the curriculum: ***because our students were unusual, the resulting curriculum was unusual as well***. We learned that deconstructing our discipline is not only an exercise that is intellectually quite profound – it is also one that depends on its context. The results of deconstructing multi-predictor modeling for a clinician look very different than the results of deconstructing multi-predictor modeling for a statistician-in-training, but both tasks are eminently accomplishable.

Appendix A
Course syllabus (revised for purposes of presentation here)

Syllabus for CRP245 (Statistical Modeling)

Overall course goal

The overall course goal is to develop an effective working knowledge of common statistical techniques for models involving one or more predictors. We don't propose to teach you everything there is to know about multi-predictor modeling, but we will teach you enough to form sound conclusions from most data sets. Moreover, we want you to "really know what you know". Operationally, this means that you can select, perform, and interpret the analyses without assistance from a statistician. In practice, this might mean performing an initial analysis yourself, and then consulting with a statistician about the finer points. Nevertheless, we have framed the overall course goal as if you were performing all analyses yourself. (See **Supplement**: *"What do we mean by "effective working knowledge?"*)

Specific course objectives

Our more specific objectives (i.e., sub-goals) follow from the primary goal. As CRP241 and CRP245 are intended to be an integrated sequence of courses, some of these objectives will have already been addressed in 241, some of these objectives will be new, and some will overlap. Some objectives are:

- ***Frame scientific questions in statistical terms.***
- ***Be able to critique advantages and disadvantages of study designs encountered in the literature.***
- ***Select an appropriate design.***
- ***Determine which statistical models are good candidates for your data.***
- ***Select a reasonable analysis plan, consistent with the study question and design. (We will define what we mean by "analysis plan" during the course.)***
- ***While acknowledging that it will sometimes be preferable for analyses to be performed by specialists, perform standard statistical analyses in order to determine the basic message of the data.***
- ***Draw conclusions that appropriately reflect the available data and model.***
- ***Present the results of statistical analyses in a coherent and persuasive fashion.***
- ***Recognize the limits of your knowledge, and appropriately determine when additional statistical consultation is necessary.***

When "performing standard analyses", this usually involves the following:

- ***Perform basic descriptive analyses of single variables and pairs of variables to orient yourself (and your audience) to the basic features of the data.***
- ***Perform multi-predictor modeling according to a reasonable analysis plan.***

Comment on goals and objectives

The goals and objectives have been framed as if you were a "primary producer" of statistical analyses. In practice, at times you will also be a "consumer" of and a "seller" of those analyses. To become an informed consumer of statistical analysis, the primary objective is:

- ***To be able to judge the soundness of statistical analyses, especially those that utilize multiple predictor variables, performed by others.***

To become an effective seller, the primary objective is to:

- ***Gain experience in making and assessing effective presentations*** (e.g., manuscripts, grants, proposals, talks at scientific meetings) ***that include statistics.***

Accordingly, the assignments will also give you opportunities to practice these skills as well.

Content

Our topic area is statistical modeling, especially models with multiple predictor variables. We will begin with the analysis of continuous outcome variables (linear regression, analysis of variance, analysis of covariance), then proceed to dichotomous outcomes (logistic regression) and survival outcomes (survival analysis). The organization of the course is very much "front loaded". For example, more class time is devoted to the analysis of continuous outcomes than to the others. Our intention is to present new ideas in the setting that is simplest and most likely to be familiar to you, and then to take advantage of the fact that these same ideas will recur in other settings. Following this strategy, we will discuss categorical predictors before continuous ones, even though the continuous case is the basis of the mathematical theory and is, accordingly, presented first in most statistics texts.

Class organization and teaching methods

Each three-session "module" (see Class Schedule, below) will be presented in a 4-step sequence. ***The first step is for you to view the lectures, through Blackboard, before the first class on that topic.*** Class 1 will include a brief summary of the material, with time for questions. We'll assume that you have already viewed the lectures, so won't recap all of their content, but instead will comment on what we believe to be most important, and how that information can be applied in practice. Class 1 will also include a demonstration of how to use Enterprise Guide to perform the analyses in question. We'll make every effort to integrate the summary and the demonstration – for example, by analyzing a data set and making comments on the underlying statistical issues as we go along.

Class 2 will involve answering some simple questions and performing a directed data analysis. The first component consists of factual questions and simple calculations intended to reassure yourself that you have mastered some of the technical mechanics of the course material. The directed data analysis can either be performed individually or – as we recommend -- as part of a study group. The questions for this directed data analysis range from simply finding numbers in the computer output to finer points of interpretation. The intention is for this exercise to provide you with some hands-on experience in modeling, with a minimum of headaches. The answers will be posted on the website one week after the completion of the module in question.

The ideal size for the groups during class 2 is 2. What we'd really like is for everyone to work their way through the directed data assignments themselves. However, we don't want anyone to be frustrated by hitting a snag, and our hope is that a size of 2 is small enough for everyone to actually do the work, yet be large enough that there is someone to ask in case of trouble. By the end of classes 1 and 2 you should have a basic understanding of how to perform analyses using the statistical model being studied.

Class 3 will involve a "stretch assignment" where you practice applying the course content in a more sophisticated fashion. Most of you will eventually use the knowledge gained from this course to write grants, write manuscripts, give presentations, plan and critique research studies, and interpret published articles and other forms of presentation. Accordingly, the "stretch assignment" will focus on the above; typically, preparing presentations and critiquing articles. We recommend groups of approximately 5 people for this assignment, as these are assignments that will benefit from a diversity of perspectives. Often, we'll ask you to work for 15-20 minutes on an element of the assignment, break to have one or more of the groups present their answers and discuss, work for 15-20 minutes on another element of the assignment, and so forth.

There are no "correct answers" to this component, so nothing will be posted on the web. One report per group is sufficient.

Please bring your laptops for classes 2 and 3 (and also for class 1 if you'd like to perform the demonstration yourself in real time).

Examinations and grading

Our view is that ranking students is not an integral part of adult education, but that grades can be helpful in providing feedback to students about what they do and do not currently understand. Accordingly, anyone who satisfactorily completes the above assignments, plus the final examination described below, has demonstrated adequate proficiency in the course content and will receive a "P". Please feel free to nominate members of your study group who seem particularly insightful and/or put particular effort into the group assignments for an "H". (These will be considered on a case-by-case basis.) Also, we will assign a grade of "H" to anyone that performs particularly well on the final examination.

To document that you have completed each module, please prepare a reflection report. In that report, state that you completed the day 2 and day 3 assignments, and also whether you performed the work in class or outside class. Also, please provide any comments that you might have on how the lectures, demonstration, assignments, and course notes can be improved. Forward any remaining questions that you have about the module content. Finally, feel free to share any insights about statistics that the educational experience has induced.

The final examination will consist of the analysis of data from a headache management trial. The trial itself, and also the data elements, are described in another document. The primary results of the trial are reported in an article whose link you will find on the website. You are welcome to review the analyses within that manuscript, but should not feel any urgency to repeat them exactly since: (a) your data set might not be identical to the one used in the manuscript; and

(b) there are a number of plausible approaches to the analysis, of which the manuscript only illustrates one. For the final examination you should perform the data analyses yourself, and write up the statistical methods, results, and conclusions sections yourself as well. The final examination is intended to assess how well that you "not only know how to implement an analysis strategy but can reasonably select one as well". You are welcome to discuss the final examination with your colleagues and/or instructors at any time.

Text

There is no ideal text for a course of this type. Well written "how to" manuals for introductory statistics (e.g., how to interpret a histogram, how to perform a chi-squared test on a 2x2 table) abound, but we are unaware of such a text for multivariable modeling. Accordingly, we will use course notes. These notes include a light discussion of statistical theory, intended to be just enough to get by, and are arranged in more or less the same order as the content modules. The notes are written in a "show and tell" style that seems consistent with the way that many physicians learn statistics.

Your comments about the notes are welcomed – we do want to revise and improve them over time.

Preliminary tasks

One of the things you should do early in the semester is to verify that you can find and download the class data sets (i.e., from Blackboard), and also to verify that you can run the archived Enterprise Guide projects. To run one of these archived projects, you should first copy the project and the input SAS data set onto your computer. Next, you need to change the pointers so that Enterprise Guide can find the input file. Do this by right-clicking on the data set icon, selecting "properties", and then pointing to the appropriate folder on your computer.

Because we'll be providing demonstrations of analyses using Enterprise Guide, we won't be creating separate tutorials about software. Most of you will find Enterprise Guide to be quite intuitive. On the other hand, it is frustrating when you aren't in sync with your software. If you start to have trouble with Enterprise Guide, you might consider the following: (a) buy The Little SAS Book for Enterprise Guide 4.1 by S. J. Slaughter and L. D. Delwiche, which is very clearly written, recommended by Doc Muhlbaier, and has as perhaps its only serious flaw that it sometimes tells you more than you need to know; and (b) join a work group containing at least one person who does understand Enterprise Guide.

Schedule

We have approximately 30 classes, roughly organized as follows.

Introduction – 1 session
1-way ANOVA – 3 sessions
2-way ANOVA – 3 sessions
Simple (i.e., 1-predictor) linear regression & ANCOVA – 3 sessions
Pre-post designs – 3 sessions
Multiple predictors – 3 sessions
Repeated measures and longitudinal data – 3 sessions

Logistic regression – 3 sessions
Survival analysis – 3 sessions
Special topics – 3 sessions
Time to work on final exam in class – 2 sessions

This schedule is tentative – in particular, we won't worry if some of the modules require more than 3 class sessions since there is time built in at the end of the semester.

SUPPLEMENT

What do we mean by "effective working knowledge?"

One element of an effective working knowledge is being able to separate what you know from what you don't know. This will be critical in determining whether and when you need to seek assistance from statistical specialists. Presumably, even in those cases where you must seek assistance, your understanding of the basic rationale behind these statistical techniques, and also the preliminary analyses that you will have performed to understand the data set, will make you a much more effective participant in the resulting collaboration.

To comment: if we are successful in meeting this goal, it will also address one of the most consistent student concerns regarding CRP245 and, indeed, the CRP curriculum as a whole. Loosely stated, the concern is that "I've learned a lot and can follow an analysis protocol if you tell me what to do, but I'm not confident that I can decide what to do on my own". What makes addressing this concern such a challenge for designing this course is that in most texts on multi-predictor modeling "deciding what to do" is taught using mathematics as the primary language for communicating the statistical ideas in question. Recognizing that you aren't a mathematical specialist and would find the usual treatment of this subject far too symbol-intensive, we won't rely on mathematics in this way. Instead, we will rely on the fact that you have high levels of scientific maturity, and base our treatment around this great strength of yours.

To illustrate the application of these ideas, and to clarify our expectations about what you have learned in CRP241, consider an observational study within a health system trying to determine whether to roll out specialized services for anticoagulation management, or to have this management performed in the more usual fashion by physicians. From administrative files of visits during the last 12 months, all of the INR test values are available for patients managed by the 2 groups, as well as the target ranges. Some simple demographic variables are available as well.

We assume that you can more or less come up with the following:

> The scientific question is "does anticoagulation service management (ACS) lead to better outcomes than usual physician management (UC)"? Because anticoagulation services require specialized personnel and are thus more costly, the organization might have performed a cost-effectiveness analysis and determined that the services must improve

the time in target range (TTR) by 5% in absolute terms (e.g., from 55% time in range to 60%) in order to be worth rolling out more generally. This minimum clinically important difference would help provide context when interpreting the results. Without necessarily being able to estimate the MCID for this study, you should at least know to ask about its likely value.

In statistical terms, the scientific question is "does the continuous outcome of TTR, calculated on a per-patient basis, differ when comparing ACS and UC"?

The main disadvantage of the present observational study design is that, because of the lack of randomization, the two groups might not be comparable at baseline. If the groups are comparable, then a t-test is a reasonable first step. What to do if the groups aren't comparable is one of the topics of this course (as are more efficient alternatives to the t-test). No matter what, the inference won't be as strong as that provided by randomization.

A reasonable analysis strategy, based on knowing nothing more than rudimentary statistical techniques, is to perform t-tests (on continuous variables such as age) and chi-square tests (on categorical variables such as gender, and the presence of various diagnoses) to determine whether the groups are comparable on those demographic variables that are present on the administrative files. If the groups aren't comparable, call a statistician. If they are comparable, perform a t-test on TTR supplemented by a non-parametric Wilcoxon test.

At this point, an unsophisticated analyst would immediately enter the usual protocol for the t-test and Wilcoxon test. A more sophisticated analyst would worry about instrumentation. Specifically, is the TTR for a patient with 2 observations taken 8 months apart as accurate as the TTR for a patient that is tested monthly without fail? To get a handle on the magnitude of the potential problem, the analyst might calculate various descriptive statistics on a per-patient basis – for example, the number of observations, the average time between observations, the maximum time between observations, etc. The analyst might also plot the time course of observations for various patients. If a problem exists, the analyst might try a simple solution – for example, to compare the results using all patients against those results for the subset of patients with at least 4 records – or the analyst might choose to call a statistician. For simplicity, we'll assume that TTR is roughly comparable across patients.

The next step is for the analyst to create a descriptive presentation of TTR – for example, box-plots for the 2 study groups. These box-plots should be sufficient to provide a qualitative impression of whether the observed study means are a fair representation of the central tendencies of the groups, and thus whether the t-test is likely to be appropriate. Normal probability plots are also helpful at this point.

In any event, both the t-test and the Wilcoxon test are statistically significant, thus reassuring the analyst that the results of the t-test are likely to be robust. The analyst combines statistical significance, precision, and clinical interpretation using the following

statement: "ACS care was consistent with that reported in the literature – average TTR was 70.5% (s.d., 3.6%). Usual care was not as good as ACS care – average TTR was 60.5% (s.d., 7.2%). The higher variability among usual care patients might reflect that some physicians provided care comparable to the ACS whereas others did not – the current analysis does not take physician into account. The mean difference between the groups was 10.0% (95% c.i., 8.5% to 11.5%), this entire confidence interval falling within the range of clinically important differences between the groups. Assuming that the groups in fact treated comparable patients (something that we could not fully assess), our recommendation is to either roll out the ACS more generally or else to allocate resources toward determining why some physicians' outcomes appeared to be better than others".

We assume that you can run Enterprise Guide or similar software to produce the above results, including the graphics. Short-cuts are acceptable. For example, if you had trouble getting the graphics to work it would be fine to select 20 observations from each group and plot the results by hand.

To comment: The above analysis illustrates each of the course objectives. It also illustrates the benefits of integrating clinical content into the execution of the analysis plan and the interpretation of its results, and the benefits of descriptive presentations (including graphics) in helping analysts to maintain their grounding.

Appendix B
Day 2 assignment for 1-way analysis of variance

1-way ANOVA – Exercises

We will perform these exercises in class. After each section, we'll take a few moments to discuss the results.

This section asks you to perform some hand calculations and create some graphs to verify that you have an intuitive understanding of what it means to "explain the variation in the outcome".

Address the following questions in pairs:

Group A -- 17,18,19,20,21
Group B -- 18,19,20,21,22
Group C -- 19,20,21,22,23

Calculate the following:
1. Overall mean (i.e., mean of the entire 15 patients)
2. Group means (i.e., means of groups A, B and C separately)

3. Plot the data points. Use a separate column for each, and use A, B and C as the plotting symbols. (Please see the Blackboard document for clarification.) Now, draw a horizontal line at the overall mean. The differences between these points and this line contribute to the sum of squares total, (i.e., the total variability in data points around the overall mean). Calculate the SST. (Please note that, for purposes of calculating SST, the labels of the data points don't matter. Instead, the calculation of SST just treats them as outcomes for 15 different patients.)

4. Now draw a horizontal line at each group mean. Notice that it is closer to the data points than the overall mean. The differences between the points, and this line, contribute to the sum of squares error, (i.e., the variability that is left over, after getting closer to the points, by using their own group mean). Calculate the SSE.

5. Note: The differences between group means and overall mean contribute to the model sum of squares, (i.e., the variability in data points due to the model effects (group membership)). Calculate the SSM.

6. *Challenging*. Label the plots (a) "within group variability"; (b) "between group variability"; and (c) "overall variability".

Calculate the following:
7. MSM
8. MSE
9. F

10. R2

11. Check your calculations of SST, SSE, SSM, MSM, MSE, F and R2 using Enterprise Guide. Blackboard has an Excel file, called CALCULATION_CHECK, containing the data.

12. Suppose group A, B, and C represent three U.S. States. Are these fixed or random effects?

This question is intended to be a realistic scenario within which researchers should at least consider the question of multiple comparisons and, indeed, often do the wrong thing.

13. Formal hypotheses are often formulated after considering preliminary data. Suppose a researcher is interested in comparing the relationship between 5 diets and weight loss. The statistician prepares a summary of the mean weight loss on each diet and presents the following average weight loss statistics to the researcher:

Group A: 2.6
Group B: 2.7
Group C: 3.3
Group D: 2.8
Group E: 2.2

The researcher notices that weight loss in groups A, B and D looks pretty similar, group E looks a bit lower and group C looks a bit higher. The researcher hypothesizes that C may work better.

a. She asks the statistician to conduct a t-test comparing groups C and E, using alpha = 0.05 for significance. Is this valid? Is the probability of a Type 1 error, 5%?
b. Suppose, instead, that she asks the statistician to conduct a 1-way ANOVA, using all of the data, but comparing groups C and E via a contrast. Is this valid? Is the probability of a Type 1 error, 5%?
c. Suppose, instead, that prior data suggested a priori that groups C and E would be different. Thus, the above hypotheses were created before any knowledge of the current data. Would your answers to a and b change?

This section is intended as practice working through the 1-way ANOVA software. It is realistic, in the sense that it has a similar structure as the analysis of data from an actual study.

The data set ONE-WAY-ANOVA-EXAM presents data from a study of satisfaction with medical care among patients receiving warfarin. The variable GROUP describes the system used for anticoagulation management: A=usual physician care, B=anticoagulation service, C=patient self-management. The variable Y summarizes the result of a satisfaction survey, placed on a 0-100 scale. (100=most satisfied, 0=least satisfied). The satisfaction survey is reasonably well validated, and it is plausible to assume that it represents a continuous variable having an interval scale. A 10-unit difference is clinically significant. Within a certain health care organization, patients under usual physician care were listed, and a random selection made of those to

interview. A similar procedure was followed for patients managed by an anticoagulation service. Relatively few patients were engaging in self-management; accordingly, all of these patients were interviewed. Complete the following questions. The Programming Notes (based on version 4.1 of Enterprise Guide) at the bottom of this assignment may be helpful.

14. Create a presentation that simultaneously summarizes the satisfaction data for the 3 groups (e.g., 3 box plots next to one another). It should be obvious that something is wrong. Find the outlier, and drop it from the data set. What was the value of the outlier?

15. Informally assess the assumptions of normality and equal variances. Are the data sufficiently consistent with these assumptions in order to be able to proceed with the ANOVA?

16. What are the best estimates for the true values of the 3 means to be compared?

17. What are the observed variances for each of the 3 groups? From the ANOVA table, what is the pooled estimate of variance?

18. Using an overall F-test, does satisfaction differ across the groups?

19. How much of the difference in satisfaction is attributable to system of anticoagulation management?

20. Which pairs of groups show statistically significant differences in satisfaction? (Adjust for multiple comparisons, if appropriate.)

21. Estimate a 95% confidence interval for the difference in satisfaction between anticoagulation service management and patient self-management. (Adjust for multiple comparisons, if appropriate). How strong is the evidence that there is a clinically significant difference in satisfaction with these two models of care?

22. An advocate of patient self-management asserts that these data demonstrate that more patients should be placed on self-management because they would become more satisfied. What component of the study design might potentially call this conclusion into question?

Programming notes (works for version 4.1 of Enterprise Guide):
Begin the project, open the data set, and change out of ***read only*** mode using the ***data*** tab. The suspicious observation can be deleted by highlighting, then applying the ***delete*** key.

Now go into the submenu ***analyze***, ***anova***, ***linear models***. In ***task roles*** define Y as a dependent variable and GROUP as a ***classification variable***. In ***model*** click on GROUP, then officially enter this into the model as a predictor variable by clicking on ***main***. GROUP should now be highlighted as an ***effect***. Under ***post hoc tests***, click on ***least squares***, when the new menu appears click on ***add***, then change the value for GROUP in the right-hand portion of the screen from ***false*** to ***true***. The right-hand portion of the screen also contains menus for selecting a ***multiple comparisons*** procedure (here, ask for

all pairwise differences), and for asking for the calculation of ***confidence limits***. Under ***plots***, request ***means***. This submenu contains various other options, of which the most commonly used include the ability to calculate influence statistics, save predicted values into a permanent data set, and produce various plots. Feel free to experiment.

Both this and the t-test project have used the submenu ***analyze***, ***anova***, ***linear models*** rather than the submenu ***analyze***, ***anova***, ***one-way anova***. The former menu is more general, accommodating both categorical and continuous predictors, and is recommended.

Appendix C
Day 3 assignment for 1-way analysis of variance

1 WAY ANOVA - DAY 3

In general, "Day 3" assignments are intended to integrate the techniques and concepts that you've learned up to that point. Various approaches might be equally applicable. This assignment is designed to illustrate complex issues that may arise in a relatively simple analytical setting. In particular, this project is focused on the concept of power and how various modifications of a study design might impact power. The example is derived from a current consulting project, but the details are exaggerated or modified to illustrate various points.

Background:
I've been asked to design a study to assess a community intervention to improve blood pressure in hypertensive patients. Hypertensive patients are diagnosed by having a blood pressure greater than 140/90. When future measurements are reduced below this threshold their hypertension is said to be "controlled". Hypertensive patients who enroll in the study will receive an at-home blood pressure monitoring system to encourage them to keep track of their status. They will receive treatment and follow-up only through normal clinical care. Patients with blood pressure greater than 140/90 would normally be treated with medications to reduce blood pressure, but follow-up visits are usually required to achieve an appropriate dose and affect a change.

The researchers want to quantify the success of the program; whether it results in greater decrease in blood pressure than would have been seen under normal circumstances in this population. There is not enough money to implement a randomized clinical trial. All hypertensive patients attending participating clinics in 2011 will be offered the intervention. Some may refuse to participate if they are not interested, don't have time to enroll, or don't want to be monitored. Patients entering a clinic in 2011 will be compared to similarly eligible patients who entered the clinic in 2010. Thus, the comparison is between a current, intervention group, "2011", and historical control, "2010". Standard information on 2010 subjects is readily available through digital records. The response of interest is "change in blood pressure" (first to last measurement) within a one year period. If the intervention is effective, we expect the intervention group to exhibit a greater "change in blood pressure", on average, than the control group. Missing data will be assumed to remain constant. In other words, patients who don't return for follow-up are assumed to have 0 changes in blood pressure. The researchers hypothesize that the "change in blood pressure" will be greater on average in 2011 than in 2010.

1. What is the response (outcome)? What is the predictor (independent variable)? What scales do these have (continuous, categorical, dichotomous)?
2. In this course, and the previous, you've seen a variety of techniques for analyzing data. These include (1) t-test for a difference of means (2) t-test for a difference of proportions (2) 1-way ANOVA. Which would you use to analyze this data? Does it matter?

3. Which group, intervention or control, do you expect to have higher variability in the response, "change in blood pressure"? Do you think this will compromise the model assumptions?
4. What assumption are we making when we compare 2010 to 2011?
5. Eligible patients, who decline to participate, clearly can't benefit from the intervention. The researchers think that we can increase the treatment effect (difference between 2011 and 2010) by focusing on participants. They suggest that we exclude non-participants from the analysis. What are the pros and cons to this approach? Will it increase power?
6. The researchers are concerned that enrollment may be slow and the sample size may be too small. They suggest that we lower the threshold for eligibility to 130/90. This would increase the number of eligible patients and consequently increase the sample size. Do you think this will increase power? Why or why not?
7. We decide to compare all eligible patients in 2011, to all eligible patients in 2010, regardless of actual participation. Some patients, who decline in 2011, may not actually receive the intervention. If participation is low, and many patients decline the intervention, how will that affect the following?
    a. Sample sizes in each group: n's
    b. Effect size: difference in "change in blood pressure"
    c. Power
8. The initial study protocol calls for staff to enroll patients at participating clinics, with only one staff member at each clinic, 20 hours a week, enrolling over the course of 6 months. Some eligible patients will be missed when the enrollment staff is not there. An alternative strategy would allow 40 hours a week, for a shortened, 3 month enrollment period. Would this increase power? Why?
9. The organizers want to create a media event to recruit participants. They plan to pass out fliers at the state fair. The fliers will explain the study goals to reduce hypertension and refer people to a clinic where they can go to participate. How might this affect the following?
    a. Sample sizes in each group: n's
    b. Effect size: difference in "change in blood pressure"
    c. Power

    Does this have any problems, given the study design?
10. In fact, there are multiple ways that this outcome of blood pressure could be evaluated. Here, only systolic blood pressure is considered. An alternative would focus on diastolic blood pressure, or a combination like the average of systolic and diastolic blood pressure. Yet another alternative would create a dichotomous outcome of hypertension (yes/no).
    a. If you chose to evaluate the average of systolic and diastolic blood pressure would your analysis strategy change (i.e. t-test, ANOVA, regression…)?
    b. If you chose to evaluate the binary variable, hypertension, would your analysis strategy change (i.e. t-test, ANOVA, regression…)?
    c. Can you think of any pros/cons to these alternative outcomes?

11. *EXTRA – This topic is beyond your experience. The answers are given. We will discuss this extensively in future modules. This is just intended to give you a first exposure to the issue. Don't worry if it seems challenging.* There's another problem with this study. Blood pressure is a moving target. The observed value is impacted by biological and temporal fluctuations, device error, human error and other sources of noise. Noise causes greater spread and observations move towards the tails of the distribution. The hypertension eligibility criterion requires a blood pressure of 140/90, which is towards the upper tail of the blood pressure range. This will tend to select people with measurement error on the high side, i.e. the sample is biased upward. In subsequent measurements there will be both positive and negative measurement error and the sample will no longer be biased upward. Values that are uniquely identified for being extremely high, will tend to be lower in subsequent measurements. This phenomenon is called regression to the mean. Mathematically, we can expect a reduction in blood pressure without any kind of intervention or treatment whatsoever. Does this matter for the current analysis?
    a. "No": Partially not, because we are comparing the reduction that is observed, on average, in 2010 and 2011. Both years will experience a regression to the mean. Presumably, that phenomenon will be similar in both years, so a greater reduction in 2011 can probably be explained by the intervention.
    b. "Yes": The above arguments are correct if the follow-up measurements were obtained in a similar way in 2010 and 2011. However, the study does not have the budget to mandate follow-up measurements on an intervention and control group. Instead, follow-up is obtained through natural clinical care, and the absence of follow-up is considered to be "no change". The presence of an intervention may change the behavior of both clinicians and patients regarding follow-up for blood pressure. Even in the absence of other improvements, they may schedule and attend more follow-up visits. More follow-up corresponds to more chances to regress to the mean. The intervention group can appear more successful if their follow-up is greater.

    Would you trust a standard ANOVA comparison of this data? What if you knew that the number of follow-up visits and timing of follow-up was similar in the intervention and control groups?

Appendix D
Grading rubric for the final examination

Is there a well-justified analysis strategy? (The justification might include a formal conceptual model / analytic framework, the distinction between variable selection and a critical experiment, the classification of predictor variables as primary and secondary, the determination of which variables should be included in all analyses, etc.).

Do the statistical maneuvers match the analysis strategy (e.g., for variable selection, some sort of stepwise selection should be used)?

Is a distinction made between primary and secondary analyses?

Was there some attempt made to visualize the data, and do the analyses appropriately take into account what was discovered during the visualization process?

Does the reporting and interpretation of the results take into account the variation in the data?

Does the interpretation of the results consider both clinical and statistical significance?

Are the limitations of the analyses adequately described?

Does the analysis plan include an assessment of the robustness of the conclusions (e.g., by implementing different analyses and determining whether the conclusions are similar)?

Is the interpretation of the data consistent with the results?

Could a clinical reader, relatively unacquainted with statistics, explain the analyses in your manuscript?

References

Ambrosius, W. T. & Manatunga, A. K. (2002), "Intensive Short Courses in Biostatistics for Fellows and Physicians," *Statistics in Medicine*, 21, 2739-2756.

Berwick, D. M., Fineberg, H. V., & Weinstein, M. C. (1981), "When Doctors Meet Numbers," *The American Journal of Medicine*, 71, 991-998.

Bland, J. M. (2004), "Teaching Statistics to Medical Students Using Problem-Based Learning: The Australian Experience," *BMC Medical Education*, 4, 31.

Bradstreet, T. E. (1996), "Teaching Introductory Statistics Courses So That Nonstatisticians Experience Statistical Reasoning," *The American Statistician*, 50, 69-78.

Carlson, K. A. & Winquist, J. R. (2011), "Evaluating an Active Learning Approach to Teaching Introductory Statistics: A Classroom Workbook Approach," *Journal of Statistics Education*, 19:1. www.amstat.org/publications/jse/v19n1/carlson.pdf

Chance, B. L. (2002), "Components of Statistical Thinking and Implications for Instruction and Assessment," *Journal of Statistics Education*, 10:3. www.amstat.org/publications/jse/v10n3/chance.html

Davis, B. G. (2009). "Discussion strategies." In B. G. Davis, *Tools for Teaching* (2nd ed.). (pp. 63-95). San Francisco, California: Jossey-Bass.

Deutsch, R. (2002), "A Seminar Series in Applied Biostatistics for Clinical Research Fellows, Faculty and Staff," *Statistics in Medicine*, 21, 801-810.

Diaz, A., Middendorf, J., Pace, D., & Shopkow, L. (2008), "The History Learning Project: A Department "Decodes" Its Students," *Journal of American History*, 94, 1211-1224.

Enders, F. B. & Diener-West, M. (2006), "Methods of Learning in Statistical Education: A Randomized Trial of Public Health Graduate Students," *Statistics Education Research Journal*. 5:1, 5-19. www.stat.auckland.ac.nz/~iase/serj/SERJ_5%281%29_Enders_West.pdf

Felder, R. M. (1990), "Stoichiometry without tears," *Chemical Engineering Education*, 24:4, 188-196.

Fosnot, C. T. (1996) (Ed.), *Constructivism: Theory, Perspectives, and Practice*. New York, New York: Teachers College Press.

Harden, J. C., Davis, M.H., & Friedman, R.M. (1999), "From Competency to Meta-Competency: A Model for the Specification of Learning Outcomes," *Medical Teacher*, 21, 546-552.

Kellaghan, T. & Stufflebeam, D. L. (2003), *International Handbook of Educational Evaluation* (Vol. 1), Kluwer Academic Pub.

Johnson, D. & Johnson, R. (1999), Making cooperative learning work. *Theory into Practice*, 38, 67-73.

Michael, J. (2006), " Where's the evidence that active learning works?," *Advances in Physiology Education* 30:4, 159-167.

Middendorf, J. & Pace, D. (2004), "Decoding the Disciplines: A Model for Helping Students Learn Disciplinary Ways of Thinking," *New directions for teaching and learning*, 1-12.

Mintz, E. & Ostbye, T. (1992), "Teaching Statistics to Health Professionals: The Legal Analogy," *Medical Teacher*, 14, 371-374.

Paulson, D. R. (August, 1999), "Active learning and cooperative learning in the organic chemistry lecture class," *Journal of Chemical Education*, 76:8, 1136-1140.

Phillips, D. C. (1995), "The Good, the Bad, and the Ugly: The Many Faces of Constructivism," *Educational Researcher*, 24, 5-12.

Prince, M. (2004), "Does Active Learning Work? A Review of the Research," *Journal of Engineering Education*, 93(3), 223-232.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004), *Evaluation: A Systematic Approach*, Sage Publications, Inc.

Slaughter, S.J. & Delwiche, L.D. (2006). *The Little SAS® Book for Enterprise Guide® 4.1*, Cary, NC: SAS Institute Inc.

Steinhorst, R. K. & Keeler, C. M. (1995), "Developing Material for Introductory Statistics Courses from a Conceptual, Active Learning Viewpoint," *Journal of Statistics Education*, 3:3. www.amstat.org/publications/jse/v3n3/steinhorst.html

Supino, P. G. & Borer, J. S. (2007), "Teaching Clinical Research Methodology to the Academic Medical Community: A Fifteen-Year Retrospective of a Comprehensive Curriculum," *Medical Teacher*, 29, 346-352.

Swift, L., Miles, S., Price, G. M., Shepstone, L., & Leinster, S. J. (2009), "Do Doctors Need Statistics? Doctors' Use of and Attitudes to Probability and Statistics," *Statistics in Medicine*, 28, 1969-1981.

Tiberius, R. (1999), *Small group teaching: A trouble-shooting guide*. London: Kogan Page.

Updike, S. (2011), "Clinical Research Training Program Student Technology Survey." Unpublished report.

Van Heuvelen, A. (1991), "Learning to think like a physicist: A review of research-based instructional strategies," *American Journal of Physics.* 59, 891.

Wiggins, G. P. & McTighe, J. (2005), *Understanding by Design*, Association for Supervision & Curriculum Development.

Wilkinson, W. E. & Oddone, E. Z. (2002), "Training Physicians for Careers in Clinical Research: A Tailored Educational Experience," *Nature Biotechnology*, 20, 99-100.

Windish, D. M., Huot, S. J., & Green, M. L. (2007), "Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature," *Journal of the American Medical Association*, 298, 1010-1022.

Gregory P. Samsa, PhD
Department of Biostatistics and Bioinformatics
Duke University Medical Center
Durham NC 27705
Mailto:samsa001@mc.duke.edu

Laine Thomas, PhD
Department of Biostatistics and Bioinformatics
Duke University Medical Center
Durham NC 27705
Mailto:laine.thomas@dm.duke.edu

Linda S. Lee, PhD
Department of Biostatistics and Bioinformatics
Duke University Medical Center
Durham NC 27705
Mailto:linda.s.lee@dm.duke.edu

Edward M. Neal, PhD (retired)
Department of Education
University of North Carolina at Chapel Hill
Chapel Hill NC 27515
mailto:ed_neal@unc.edu
