S-052: Intermediate and Advanced Statistical Methods for Applied Educational Research

Harvard Graduate School of Education, Spring 2019

Class meets Tuesdays and Thursdays from 10:10AM to 11:30AM (Likely location: Larsen G08)
Section Meetings to be arranged (Likely location: Gutman 302)

Instructor: Andrew Ho
455 Gutman Library
Andrew_Ho@gse.harvard.edu
617-496-2408

Faculty Assistant: Wendy Angus
444 Gutman Library
Wendy_Angus@gse.harvard.edu
617-496-4802

Course overview

Welcome to *S-052: Intermediate and Advanced Statistical Methods for Applied Educational Research*. This course is an integrated continuation of the fall course, *S-040*, and is part of the *HGSE* school-wide network of courses in quantitative methods. The *S-040* and *S-052* courses form the cornerstone of a sequence of courses in applied statistical methods for consumers and producers of rigorous educational, social, and psychological research.

The course is designed to develop and extend the data-analytic skills that you began to acquire in earlier courses and to help you learn to communicate your findings clearly to audiences of other empirical researchers, scholars, policy-makers, practitioners, students, and parents. We have designed *S-052* to contribute to the diverse data-analytic toolkit that you will need in order to perform sensible and useful analyses of complex educational, psychological, and social data.

Core topics such as multiple regression analysis, introduced in *S-040*, continue to be the foundation of *S-052*. However, we extend your use of these techniques to cover a wider variety of conditions encountered in the world of real data-analysis, including multilevel models and selected multivariate methods. A listing of major course topics is provided later in this document.

True to its name, *S-052* is an *applied* (not a *theoretical*) course in which you will learn by observing and engaging in the authentic activities of real applied data analysis. We will model the use of new statistical techniques in class, and then you will apply these new techniques to real problems using real data in "data-analytic memos" and a take-home examination.

We will also ask you to interpret the outcomes of your data-analyses in words, and to communicate these interpretations clearly and concisely in writing. Finally, you will acquire the basic programming skills necessary for hands-on data analysis in *Stata* or *R*.

Presentational structure

As a rough guide, presentation of each statistical technique will contain *seven components*. For each technique, we will:

1. ***Provide A Relevant Research Question***. All analytic methods are secondary to, and exist because of, questions of substantive importance. So, stating the specific research question that is to be addressed is a critical precursor to any effective data-analysis.
2. ***Obtain a Suitable Dataset***. For each new method described, we will provide one or more real datasets and use them to address the research question. We will introduce each dataset by describing its origins and by providing connections to related internet resource materials where they are available. Our presentations may also include comments on the utility and reasonableness of the research design that led to the data-collection.
3. ***Specify an Appropriate Statistical Model***. The basis of any effective data analysis is the specification of a statistical model that represents credibly the substantive process under investigation and embodies the researcher's hypotheses. We will describe the specification of such statistical models and the meaning and role of critical parameters in the model in terms of the stated research question.
4. ***Fit the Statistical Model to Data***. Fitting the specified statistical model to data will provide the vehicle by which we will discuss the application and functioning of appropriate statistical methods. We will also comment on elements of computer programming for data analysis in Stata or R.
5. ***Describe and Assess the Assumptions Underlying the Statistical Method***. All analytic methods are underpinned by assumptions. Assumptions underlying a statistical technique effectively constitute an additional source of information that is "input" implicitly by the technique into the data-analysis. This makes the adequacy of the assumptions essential to check. If they are violated, then the additional "information" that they have passed into the analysis will be incorrect and the findings dubious. We will describe the assumptions that underpin each new technique and illustrate how the credibility of the assumptions can be assessed using relevant diagnostics.
6. ***Estimate, Test and Interpret Central Parameters of the Statistical Model***. At the end of any data analysis, estimates of model parameters and their associated tests provide the formal answers to the research questions. We will describe the estimation, testing, and interpretation of critical model parameters, along with the use of appropriate statistics for summarizing model goodness-of-fit.
7. ***Answer the Research Question***. At its end, statistical analysis is not worth anything if its findings cannot be translated into common-sense conclusions for an audience of intelligent consumers, whether they are other scholars, practitioners, policy makers, parents, or students. We will emphasize the authoring of cogent substantive interpretations of findings, based on the careful identification and translation of relevant parts of the data-analytic output.

In the rest of this document, we describe textual resources, list the course content in greater detail, spell out prerequisites, review dates, and clarify policies.

Recommended Text

We recommend using the first volume of Sophia Rabe-Hesketh and Anders Skrondal's textbook (RH&S), *Multilevel and longitudinal modeling using Stata, Volume I*, 3rd edition. You can find it at the Stata Bookstore for around $65, cheaper if you have a discount code. Amazon and the Coop tend to be more expensive. It is a slightly advanced but highly applied text that offers excellent support for about 2/3 of the course material, including linear, logistic, and multilevel regression. You can see the coverage here. We will also be using a free online chapter from their second volume on logistic regression, here. I review overlap with course content in the next section.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata, Volumes I and II* (3rd ed.). College Station, TX: Stata Press. [Only Volume I is recommended for this course.]

Overview of course content (This is subject to adjustment, particularly toward the course's end)

1. Fitting Sensible Taxonomies of Multiple Regression Models (RH&S Ch. 1):

(a) *Deciding Which Regression Models to Fit*. Addressing research questions by fitting taxonomies of multiple regression models and determining a sensible "final" model. Reviewing the specification of regression models in which a single substantive construct is represented by a system of "dummy" predictors. Reviewing the notion of a statistical interaction between predictors. This introductory section provides an opportunity to review your prior learning about multiple regression analysis.

(b) *Testing Complex Hypotheses About Regression Parameters*. Comparing nested multiple regression models. Using the *General Linear Hypothesis Tess* (ΔR^2 *test*) to conduct formal tests of the joint effect of several predictors simultaneously on an outcome.

(c) *Detecting Influential Observations and Assessing their Effects on Model Fit and Parameter Estimation*. Introducing the notion of influence statistics. Conducting sensitivity analyses.

(d) *Checking the Assumptions on the Residuals*. Understanding the importance of the assumptions on the residuals in a multiple regression analysis. Graphical methods for assessing distributional assumptions.

(e) *Interpreting and Reporting Findings*. Using fitted plots to display and interpret the size and direction of detected effects for prototypical individuals in the population, especially in the presence of statistical interactions.

(f) *Dealing Empirically with Non-Linear Outcome/Predictor Relationships*. Using power transformations to linearize the outcome/predictor relationship. Introducing Tukey's *Ladder of Transformations*.

(g) *Causal Inference and Counterfactual Reasoning:* Introducing the fundamental problem of causal inference and distinguishing between causal vs. predictive interpretation of regression coefficients.

2. Basic Logistic ("Binomial Logit") Regression Analysis (RH&S Ch. 10, online):

(a) *Modeling the Relationship Between a Dichotomous Outcome and Predictors using a Linear Probability Model*. The problematic impact of specifying a linear regression model when the outcome is dichotomous – appropriate interpretation of fitted values and problems in the residual distribution.

(b) *Modeling the Relationship Between a Dichotomous Outcome and Predictors with Logistic Regression ("Logit") Analysis*. Using a non-linear logistic (or "logit") function to represent the hypothesized relationship between a dichotomous outcome and predictors. Goodness-of-fit statistics for logistic regression analysis.

(c) *Fitting Taxonomies of Nested Logistic Regression Models*. Addressing research questions about the prediction of dichotomous outcomes by fitting and comparing nested logistic regression models using a *General Linear Hypothesis* (χ^2) *Test*.

(d) *Interpreting Fitted Logistic Regression Models*. Using fitted *odds, odds-ratios,* and fitted trend lines plotted for prototypical individuals in the population to demonstrate the size and direction of an effect detected via logistic regression analysis.

3. Survival Analysis (Singer & Willett, 2003, Chapter 10, iPac):

(a) *Discrete-Time Survival Analysis*. Using logit analysis to examine the occurrence and timing of events in a person's life. Introducing the concepts of hazard and survivor probability, and the discrete-time hazard model.

(b) *Adding Predictors to the Discrete-Time Hazard Model.* Using prototypical fitted hazard and survivor functions, and predicted median lifetimes, to interpret findings.

4. Multilevel Models (RH&S Ch. 2, 3):

(a) *Introducing the "Fixed Effects" Model*. A useful and robust approach using dummy variables for groups to account for multilevel structures.

(b) *Introducing the "Random Effects" Multilevel Model*. Contrasting a "fixed effects" model with a multilevel "random effects" regression model to account for the grouping of individuals within higher-level "units."

(c) *Interpreting variance components and intraclass correlations.* Fitting the multilevel model using random-effects regression analysis. Partitioning residual variance into its within-group and between-group components, estimating and interpreting the intraclass correlation.

(d) *Using the Multilevel Regression Model to Analyze Longitudinal Data*. Extending the random intercepts model to a random slopes model for analyzing individual change over time.

5. Multivariate Methods—Reliability and Principal Components Analysis

(a) *Reliability and Classical Test Theory*. Traditional strategies for forming data-composites -- standardization of indicators, creating a weighted linear composite. Measurement error and internal-consistency reliability (Cronbach's α).

(b) *Using Principal Components Analysis (PCA) To Form An "Ideal" Data-Composite*. Introducing principal components analysis as an alternative to classical item-analysis in data-compositing. Creating a composite that accounts for maximum variance.

(c) *Using PCA To Evaluate the Multivariate Structure of Data*. Using the eigenvalues and a scree plot to estimate how many "dimensions of information" underlie a given set of indicators. Interpreting the underlying dimensions numerically, graphically and substantively.

6. Cluster Analysis (time permitting)

(a) *Exploratory Cluster Analysis of Variables*. Automated exploratory analyses to extend PCA so that interesting "coherent" clusters of indicators can be detected within a larger defined group of variables.

(b) *Exploratory Cluster Analysis of Individuals*. Automated exploratory analyses that group individuals together into discrete clusters so that "profiles" and "types" of behavior are revealed. Using the *Pseudo-F Statistic* to determine the number of clusters, and the *Tree Diagram* to display the clustering detected.

(c) *Using the Cluster Analysis of Individuals to Detect Multivariate Outliers*. Which of these people are most like the others, and which of these people don't belong?

7. Factor Analysis and Structural Equation Models (time permitting):

(a) *Basic Factor Model*. Using path diagrams to describe multivariate hypotheses about relationships among multiple indicators, constructs and measurement errors in a first-order factor model. Representing the path model as a system of inter-linked statistical models and understanding the role played by the model parameters.

(b) *Structural Equation Model (SEM) and Confirmatory Factor Analysis (CFA)*. Fitting the hypothesized model to data and assessing its fit. Testing and interpreting critical parameters in the model.

Frequently Asked Questions: Is S-052 right for me?

Can I attend lecture or section as a guest or an auditor?
In-person attendance at lectures and sections is restricted to enrolled students. Due to room size constraints, I do not allow formal auditing or guest attendance. Students interested in electronic resources may email my assistant in mid-January to receive guest access to the course website, which I allow in certain cases, such as for students with conflicts and alumni.

I am not a HGSE student. When should I file a cross-registration petition?
This course does not have an enrollment limit. If you are reading this before January 25, there is no immediate rush. After January 18, cross-register here: https://courses.harvard.edu/search?q=180866. I ask that you file your petition by Friday, January 25, at 12PM.

I intend to cross-register. Do I need to demonstrate my prerequisites now?
No. You should decide whether this course is best for you, including whether you meet prerequisites, below and then you should file a cross-registration petition by Friday, January 25, at 12PM. I will then contact you to determine your eligibility. If you meet the requirements, I will approve you. If we decide that another course is better for you, I will not.

What are the prerequisites for S-052?
Successful completion of S-030 or S-040 (A- or A) or the equivalent. We expect you to have successfully, a) fit a regression model, b) with an interaction term, c) to real data, d) with a computer program, and e) interpreted the statistical significance of the coefficient for the interaction term as well as, f) written out the meaning of the coefficient for the interaction term in writing, g) in the context of a research question.

Is S-030 a better choice for me?
Maybe! S-030, offered in the same semester as S-052 by Professor Hadas Eidelman, is designed precisely for students with limited exposure to multiple regression. The depth of coverage of multiple regression in S-030 is considerable. In contrast, coverage of multiple regression in S-052 is limited to a brief review. Students with limited past exposure who wish to develop real comfort and expertise with multiple regression will be better off in S-030 than in S-052. S-052 offers a much broader introduction to more advanced statistical methods but cannot compensate for a student's limited exposure to multiple regression on its own.

I just took an introductory statistics course (e.g., in SPH, BIO201 or ID201) this fall. Is that sufficient?
An A- or an A in a recent, rigorous introductory statistics course like BIO201 or ID201 will suffice as a prerequisite for S-052 if you have estimated and interpreted regression models with interaction terms as described above. However, S-030 remains complementary, and we strongly suggest that you consider it. Again, S-030 is a deep dive into applied multiple regression, and it builds analytical and interpretive skills that a typical introductory statistics class does not.

I earned a B+ in an introductory statistics course like BIO201 or ID201. Should I take S-052?
No. S-052 moves quickly, demands deep conceptual understanding, and covers several advanced topics. Students at this level should take advantage of S-030 as an option to truly master foundational regression analysis first.

I don't meet the prerequisites, but I really want to learn survival analysis/multilevel modeling/principal components analysis. Should I take the course?
No. Our goal is not superficial acquisition of methods but deep conceptual mastery. For this, the foundations are necessary, and we strongly recommend S-030 as a rigorous alternative.

Course activities and participation

Most of our time—both inside and outside of class—will be spent learning how to do data analysis. When I believe that your understanding will be enhanced by knowing more about the mathematical underpinnings, I'll offer (what I hope are) straightforward conceptual explanations that do not sacrifice intellectual rigor.

We will devote time to illustrating how to present results in words, tables and figures. Good data analysis is craft knowledge. It involves more than using software to generate reams of output. Thoughtful analysis can be difficult and messy, raising delicate problems of model specification and parameter interpretation. We'll confront such issues directly, offering concrete advice for sound decision making.

Class participation is an important part of learning, even in a relatively large lecture course like S-052. If you have a question, it's likely that others do as well. I encourage active participation with live, open-edit google documents as well as raised hands. If students make efforts in and out of class to engage actively with course content to the benefit of themselves and their fellow students, I may factor this into grades that fall near grading cutpoints. For in-class participation, please do not be offended if I defer your contribution to another time, if I feel addressing the question may take us too far astray.

Course website: https://canvas.harvard.edu/courses/55166/

Bookmark the course website and check it often (especially in advance of every class and sometimes more frequently). The website is my primary means of taking care of "housekeeping" matters (eliminating the need to discuss deadlines, etc. in class). It also has resources designed to enhance your learning, including handouts, homework assignments, datasets, and some web-based materials that help further explain statistical concepts.

Meeting times and the attendance policy

Consistent with HGSE policy, class will start promptly 10 minutes after the official start time. Thus, we will begin 10:10 and end at 11:30 every Tuesday and Thursday. Please be seated and ready at the appointed time. **I expect all students to attend every class meeting, on time.** I will take attendance digitally through in-class online quizzes. All students must bring a digital device with an internet-connected online browser, such as a laptop or smartphone, to every class. Students who do not have such a device should contact me for alternative arrangements before enrolling in the course.

Online class videos

Each class meeting will be digitally recorded and streamable online. We endeavor to have the videos ready by the end of the day or midday the next day, but we cannot guarantee this. I provide the videos so that you can review the class material at your own pace. There are occasional glitches, so please do not rely on videos exclusively. **Do not abuse the system: videos should supplement, not supplant, lectures.**

Professional behavior in a digital age

S-052 is technologically intensive and many students bring laptops to class to take notes. Personally, I find it difficult to take notes online, because the notes I write would be less text-based and more graphical (equations, graphs, and other sketches), but I will leave that up to you. **What I do expect is professional behavior—that means no email, web surfing, instant messaging, or any other unrelated electronic activity during class.** It's not only rude, it is distracting to your classmates. For more about the negative effects of digital distractions on your learning and those of your peers, I encourage you to read this review article by Susan Dynarski. Cell phones should be completely silenced, including loud vibrations, and they should not be used for texting in class.

Statistical computing with Stata 15 or R

Statistical computing is an integral part of S-052. To support your learning, the quantitative methods sequence at HGSE uses Stata 15 for Windows. I assume that everyone is comfortable using a computer to perform basic statistical analysis, although I do not assume that you have used Stata. For partners who agree to use the free statistical program, R, we will not provide support, but we will allow you to submit assignments for which you have used R to fit models and estimate statistics. We recommend Stata for all but the most ambitious or already-R-fluent students, as you will essentially be "on your own" for programming support.

We will cover Stata programming in sections more than we will in lecture, although code is threaded through the lecture slides and available online. You may find the recommended course text useful for coding. If an additional reference is desired, I recommend this text: Kohler, U., & Kreuter, F. (2012). Data analysis using Stata (3rd ed.). College Station, TX: Stata Press. You can search for prices here: http://www.addall.com/New/compare.cgi?dispCurr=USD&isbn=1597181102

There are two ways you can access Stata. The least expensive option is to use one of the networked workstations available in the Learning Technology Center (LTC) on the 3rd or 4th floors of Gutman Library and elsewhere on the HGSE campus. For students who would like to use Stata on their own PCs, you may purchase Stata following this link: http://www.stata.com/order/new/edu/gradplans/student-pricing/. Stata/IC, which will be sufficient for this course, is available for $45 for a 6-month license and $198 for a perpetual license. Note that "Small Stata" will be sufficient for most but not all of this course.

Homework assignments

I believe that the only way to understand statistical analysis is to actually conduct statistical analysis. To help you develop your skills, we will administer and grade **seven homework assignments.** In honor of my predecessor, John Willett, we call these "Data Analytic Memos" (DAMs) which he referred to lovingly as "those DAM things." A tentative schedule for homework assignments follows; these dates may change. All assignments must be submitted on Canvas by the date and time specified. Late assignments will not be graded and will contribute 0 to your course grade, so please submit with time to spare in anticipation of unforeseen technical issues. To avoid last-minute panicking, I strongly encourage you to have the assignment complete or near-complete by the Thursday class period before the Friday due date.

Tentative Assignment Schedule

Assignment	Available on or about	Due by 1PM on	Collaboration Format
Assignment #1	Monday, January 28	Friday, February 8	Individual
Assignment #2	Monday, February 4	Friday, February 15	Pairs
Assignment #3	Monday, February 18	Friday, March 1	Pairs
Assignment #4	Monday, March 4	Friday, March 15	Pairs
Assignment #5	Monday, March 25	Friday, April 5	Pairs
Assignment #6	Monday, April 8	Friday, April 19	Pairs
Assignment #7	Monday, April 22	Wednesday, May 1	Pairs
Final Exam	Monday, May 6, 9AM	Wednesday, May 8, 5PM	Individual

Collaboration and study groups

Many people learn best when working in a group, and I encourage collaborative learning. My primary goal in teaching S-052 is to help students improve their understanding of applied statistics and data analysis, and collaborative learning is a great way to achieve this goal. To mimic statistical work in the

real world and to provide a chance for you to use statistical language actively, I mandate completion of assignments in pairs throughout the course, excepting only the initial assignment and the final exam.

We mandate collaboration for at least three reasons. First, learning statistics is like learning a language. To learn it, one must "speak" it actively and in a genuine context with other individuals. Second, collaborative statistical analysis is the norm and individual work is the exception in the world of statistical practice. Third, my experience has been that, on average, students who work in pairs and groups both perform better and enjoy themselves more than students who work individually. Statistical collaboration is a case where the whole is greater than the sum of its parts.

Beyond pairs, larger study groups can be helpful to you as you prepare to do the assignments, both in terms of how to approach the work (including how to use Stata effectively) and in terms of how to think about important concepts. **However, students must turn in work as pairs or individuals where specified above, not group work. Papers should be written in your own words—your text should reflect your own understanding of the material.**

A couple of rules will help to avoid misunderstandings and violations. First, never send electronic documents with your answers to members outside of the partnership. Second, never sit at the same computer with members outside of the partnership and cowrite answers to be shared.

Each group will undoubtedly develop its own structure; nevertheless, here are a few suggestions:

- Groups with six or more members become less useful and may be harder to organize because finding common meeting times becomes increasingly problematic.

- Plan at least one session of 1½ to 2 hours (early enough so that there is sufficient time if an additional session is necessary). After 2 hours of statistics, everyone's eyes will be glazing over.

- Schedule the meetings so that you have sufficient time afterwards to write in pairs or individually. When we read your assignments, we focus on what you say and how you say it. The assignments have been devised to require not only computation and programming skills, but skills in analyzing and reporting the material.

- Use the groups to ask questions, try out interpretations, and so on—you each represent each other's resources. Often one person can explain something that makes you see something in a new way—or the other way around. Different people have different insights and strengths – some are good programmers, some ask good questions, others value contextual analysis—and you can learn from listening to what others in a group have to offer.

- **Do not share text, online or in person, to complete assignments outside of partnerships.** You and your partner must write your own paper, on your own, using your own language. **Your papers should be written in your own words, not those of your study group.**

- Be sensitive to the distinction between collaboration to plan for and interpret the assignment and collaboration to write up the assignment. The former is encouraged; the latter is forbidden beyond, when applicable, your partner. If the distinction begins to feel murky, refocus your group's work on lecture content and course materials.

Final Exam

The final exam is a three-working-day, partly collaborative and partly individual affair that will be posted during reading week. It is like a large assignment, without coding. It will run from Monday, May 6

through Wednesday, May 8. Like assignments, final exams must be submitted on time. Extensions will not be granted, except in the case of personal emergency.

Avoiding plagiarism

Please read the School's policy on plagiarism in the HGSE Student Handbook, which includes the statement, "Students who submit work either not their own or without clear attribution to the original source, for whatever reason, face sanctions up to and including dismissal and expulsion." Attention to this policy is particularly important in a course like S-052, in which collaboration with other students is encouraged. If you work closely with other students during the planning of your analyses—a process that I encourage and fully support—recognize the other students' contributions explicitly in your written account (a footnote is fine for this purpose). This helps avoid the natural questions that arise when similarities are detected at grading. **If you have any questions about what constitutes appropriate collaboration, or how to define what constitutes your own work, please see me or a Teaching Fellow.**

> **I cannot overemphasize the need for all students to monitor their own behavior. Assignments are structured such that you can receive feedback on *your understanding of the material.* The consequences for plagiarism are appropriately severe.**

Grading

You will be evaluated based on your performance on the homework assignments (approximately one half of your grade) and the final exam (approximately one half of your grade). I may also factor in your attendance and data from in-class online quizzes using Google Forms, for no more than 10% of your grade.

While we use arithmetic computations to arrive at a first approximation of your course grade, in the end, no individual assignment takes on undue weight, and we consider the slope of individual trajectories as well as the level. We look at your whole portfolio of work when assigning course grades. For more details, see our handout entitled, *How We Evaluate Assignments*.

Students may choose to take the course on a satisfactory/unsatisfactory basis on the condition that can find another partner who can take it on this basis. Satisfactory performance requires course attendance, an average of B or better, and completion of all assignments and the final exam.

Accommodations

Students needing accommodations in instruction or evaluation must notify me early in the semester, and HGSE's policies must be followed. Late requests for accommodations will not be honored unless there is a pressing reason, such as a recent injury.

Our use of electronic data on Google and Canvas

I am always trying to improve my teaching and your learning, and I often use online resources to aid this effort. We use Google Forms for in-class quizzes and attendance, Google Docs for in-class participation, and Canvas for collecting and grading your assignments. As part of my effort to improve my teaching and your learning, I will use data from these resources to provide feedback to myself and, of course, to you.

You should be aware that all these resources record data from your interactions in their server logs. Sometimes this will seem obvious to you, such as when you submit an assignment, answer an assessment question, or ask a question. Other times it will seem less obvious, such as when you log in and download a handout. It is important for you to understand that, while all these data exist, I will always make it clear when and how I will use this data for grading.

You should also be aware that, like all educational data collected in the natural course of an educational process, these data may support future research endeavors, provided that your identity is masked or exemptions required by federal law apply.

For more details, see the Canvas Privacy Policy, linked here.