S-052: Intermediate and Advanced Statistical Methods
Spring 2024

Instructor: Zach Himmelsbach (he/him)
PhD Candidate, Lecturer, Adjunct
zah972@g.harvard.edu

Faculty Assistant: Wendy Angus (she/her)
Senior Coordinator for Faculty Support
Wendy_Angus@gse.harvard.edu

All course materials available on Canvas

Weekly Meeting Times

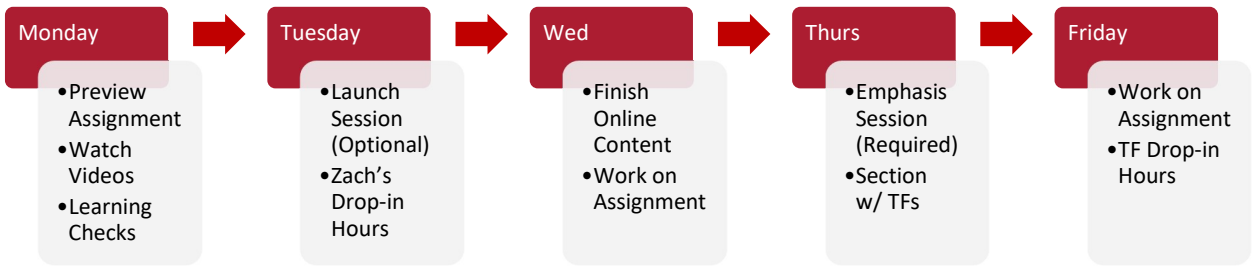
Launch Sessions
Required Week 1 and additional weeks, otherwise recommended.
Tuesdays, 10:30-11:45am, Location TBD

Emphasis Sessions (Required)
Thursdays, 10:30-11:45am, Location TBD

Zach's Drop-in hours (Optional)
Tuesdays in Gutman Library, 1st floor.
Before and after Launch Sessions (10am-4pm)

Sections with Teaching Fellows (Recommended)
Thursdays, 12:00-1:15 and 1:30-2:45

Typical Week in S-052:



Course Overview

This course deepens and extends the statistical skills learned in introductory statistics course(s). We have designed *S-052* to contribute to the diverse data-analytic toolkit that you will need to perform sensible and useful analyses of complex educational, psychological, and social data. The course also emphasizes the development of communication skills for sharing analyses with a wide range of audiences (researchers, policymakers, community members, etc.).

Regression analysis, introduced in prior coursework, remains central in *S-052*. However, we extend the basic techniques to cover many conditions encountered in the real world of data-analysis, including causal inference, multilevel models, and selected multivariate methods. Major course topics are listed later in this document.

S-052 is an *applied* (not a *theoretical*) course focused on real-world data analysis. We model the use of new statistical techniques in class, and you apply them to real problems using real data. Assignments include data-analytic memos (DAMs) and a significant final take-home assignment affectionately known as The Celebration of Learning.

In all assignments, we ask you to communicate your analyses in clear and concise writing. You will also acquire the basic programming skills necessary for hands-on data analysis in R. (Stata code is also provided for the assignments.)

We offer the course twice per year: in the fall for students with a previous semester of applied regression analysis, and in the spring as an integrated continuation of the fall course, *S-040*.

Library reserves and other text resources (on Canvas)

Textbooks are an excellent way to learn statistics and data science. Unfortunately, few textbooks cover the range of methods we teach in this course. Considering this, we provide a range of resources you can find on Canvas's library reserves.

Below are our recommended textbooks, by topic. Some of them, or sections of them, are available in the course library reserves on Canvas.

Andrew Ho recommends the following books:

For causal inference and quasi-experimental methods in education (M&W):
Murnane, R. J., & Willett, J. B. (2010). *Methods matter.* Oxford University Press.

For linear and logistic regression and random effects modeling in Stata (RH&S):
Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using Stata, Volumes I and II* (4th ed.). College Station, TX: Stata Press. Only Volume I is recommended for this course. One chapter in Volume II is relevant and available freely here.

For event history analysis and longitudinal models (S&W):

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis.* Oxford University Press.

For a general text covering most topics:
Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). Applied regression analysis and other multivariable methods (5th ed.). Cengage Learning.

Zach Himmelsbach recommends the following books:

For regression, generalized linear models, and multilevel/hierarchical/random effects models,
Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
N.B. They've released a new version of the first half of this book, called Regression and Other Stories, which is excellent, but it emphasizes a Bayesian view of regression, which we do not consider in this course.

For causal inference,
Huntington-Klein, N. (2022) The Effect: An introduction to research design and causality. Chapman & Hall.
The link above goes to a free online version of the book.

For general statistics and statistical thinking,
Freedman, D., Pisani, R., & Purves, R. (1998). Statistics (3rd ed.) W.W. Norton.

For a wonderfully readable history of statistics,
Salsburg, D. (2002). The lady tasting tea: how statistics revolutionized science in the twentieth century (1st. Owl Books ed.). W.H. Freeman/Owl Books.

Overview of course topics

The schedule of planned topics for S-052 follows. Expect some adjustment, particularly towards the end of the semester.

Week 1: All of Statistics
- The big picture of applied statistics
- "Verbatim" and "Gist" interpretations of descriptive statistics.
- Five ways to interpret means, SDs, standard errors, scatterplots, correlations.
- Interpretation and explanation of null hypothesis significance tests.

Week 2: Model Building and Interaction Terms (RH&S Ch. 1, Beck&Beck)
- Building a model is like building an argument
- Interpreting and explaining models with interaction terms
- Five regression assumptions
- When and why the five assumptions matter for practical purposes
- The Frisch–Waugh–Lovell Theorem

Week 3: Causal Inference, Potential Outcomes, and Experiments (M&W Ch. 4&8)

- The Neyman-Rubin Causal Model
- The logic/magic of random assignment
- Average Treatment Effects (ATE)
- Internal and external validity

Week 4: Regression Discontinuity (M&W Ch. 9)

- Assumptions
- Local Average Treatment Effects (LATE)
- Bandwidth selection and validation checks
- Connection to interaction terms

Week 5: Difference-in-Differences

- The logic of DiD
- Parallel Trends
- Connection to Interaction Terms
- Recent Complications

Week 6: Nonlinearity

- Polynomial regression
- Connections to interaction terms
- Logarithmic transformations of predictors or outcomes

Week 7: Logistic Regression and Generalized Linear Models (Pampel, 2000; RH&S Ch. 10)

- Logistic regression vs. the Linear Probability Model
- Log-odds units (logits)
- Interpreting and avoiding odds-ratios

Week 8: Interpreting Sociodemographic Variables (Castillo & Gillborn, 2022). Maximum Likelihood.

- Critical perspectives on the use of sociodemographic variables
- Maximum likelihood
- Chi-square tests for model comparisons
- Poisson regression

Week 9: Event History Analysis (S&W Ch. 10)

- Questions about whether and when an event occurs
- Time fixed-effects and Discrete Time Survival Analysis
- Analyses concerning persistence and attrition

Week 10: Multilevel Models/Random Effects (RH&S Ch. 2, 3)

- Multilevel models and cluster-robust standard errors approaches to dependent errors
- Within-group and between-group variance
- Intraclass correlations
- Within vs. between relationships

Week 11: Multivariate Methods: Measurement & Reliability (AERA/APA/NCME, RH&S Ch. 9)

- Measurement, validation, and interrater reliability
- The 5 Cs of Validation
- Crossed random effects, reliability, and Cronbach's alpha

Week 12: Multivariate Methods: Principal Components Analysis (Dunteman)

- Comparing regression, measurement, and compositing
- Eigenvalues, eigenvectors, and scree plots
- Assessing dimensionality
- A potato with a bunch of flags stuck in it

Week 13: Review: The Seven Levels of S-052

- Effects
- Control
- Interaction
- Nonlinearity
- Likelihood
- Context
- Combination.

Some topics may appear disjoint, but we will discover their deep connections. Throughout the course, we emphasize some broad concepts and approaches repeatedly, deepening our understanding each time. These include:

- The Big Picture of Statistics (How all the material fits together)
- The 5 Gs of Statistics Communication
    - Greek (mathematical notation)
    - Graphs (visual representations)
    - Grammar (writing and speaking)
    - Gadgets (physical representations)
    - Games (puzzles, demonstrations, competitions)
- Real-world examples and data from education, health, and other fields
- Model comparisons and tradeoffs
- Counterfactuals
- Null Hypotheses
- Evaluating effects both statistically and substantively

Course activities and participation

Most of our time—in and out of the classroom—will be spent learning practical material: how to analyze data and interpret results. At times, we will undertake some study of the mathematics and algorithms that underpin our statistical methods. Alongside the mathematical view, we offer straightforward conceptual explanations that do not sacrifice intellectual rigor.

Class participation is an important part of learning, even in a relatively large lecture course like S-052. If you have a question, others likely do as well. We encourage you to ask questions in class and/or on Slack. Student efforts – in and out of class – to engage actively with course content may factor into grades that fall near grading cut-points, especially if these efforts benefit other members of the course.

Good quantitative analysis is craft knowledge. It involves more than using software to generate reams of output. High-quality analysis requires thoughtfully handling difficult problems of model specification and parameter interpretation. We confront these issues directly, offering concrete advice for sound decision making.

Meeting times and the attendance policy

Class will start promptly at the official time. Please arrive a bit early to ensure you are settled in before we start. We often take attendance digitally through in-class online "learning checks." We use attendance data formatively, to ensure that students are staying on pace and to direct additional support.

Class meetings are recorded and available on Canvas. Tuesday Launch Sessions are recommended and sometimes required, for example in Week 1. Thursday Emphasis Sessions are always required. Your commitment to in-person attendance will help build a lasting community. If you have an unexpected conflict, simply let us know by email. Class may occasionally meet on Zoom due to the instructor's travel constraints.

Professional behavior in a digital age

All of us are likely to face distractions during meetings, whether in-person or online. Sometimes these distractions are within our control, and other times they are not. We encourage you to skim or read this review article by our HGSE colleague Susan Dynarski about the risks of distraction to learning. Please do your best to monitor your engagement accordingly.

Some Tuesday Launch Sessions (or portions of them) will be laptop free.

Assessments: Learning Checks, Learning Submissions, and Data Analytic Memos

We use **three types** of assessments to monitor and support your learning.

Learning Checks

- What: 1-3 short questions you discuss with other students
- When: In class and between Canvas Videos
- Feedback: Model answers are released to you as soon as you submit your responses. Full credit for all effortful responses
- Instructional Goal: Stimulate your learning and give us a check on your understanding and engagement
- Note: occasionally, to stimulate your thinking, these will be puzzles that go beyond what we've covered. It's okay to be wrong 😊

Learning Submissions

- What: Series of questions in a word document. For some questions, you'll run – and sometimes edit – code that we provide. These questions focus on the careful interpretation and communication of analytic results.
- When: Available every Monday (except DAM weeks); Due the following Sunday (9PM EST). You should open them at the start of the week (and make sure you can run the code)
- Feedback: Model answers are released to you as soon as you submit your responses. Full credit for all effortful submissions
- Instructional Goal: Practice thinking and writing about quantitative analyses. Give us a check on your understanding

Data Analytic Memos (DAMs)

- What: Data analyses that you conduct and interpret. You will work on and submit these in pairs.
- When: Five DAMs, due on Sundays at 9PM EST. DAMs are in weeks 2, 5, 7, 10, and 12. Open each DAM early. They are released at least two weeks before they are due.
- Feedback: The teaching team will provide detailed feedback on your submission. Grades based on quality of submission.
- Instructional Goal: Give you hands-on data analytic experience. This is the real deal.
- Note: The name DAM comes from John Willet, who referred to these assignments, lovingly, as "those DAM things."

Learning Submission Weeks (individual, Sundays 9pm ET): 1, 3, 4, 6, 8, 9, 11

DAM Due Weeks (partnered, Sundays 9pm ET): 2, 5, 7, 10, 12

Statistical computing with R (or Stata 18); the Programming vs. Critical Reading Pathways

Writing code is an important practical skill. It is also a helpful way of learning applied statistics. At the same time, learning to code can be immensely frustrating in practice, with hours spent looking for a missing comma or quotation mark. At worst, programming tasks encourage rote "copy-paste-find-replace" behavior rather than authentic practice. In S-052, we provide statistical code in R (and Stata) for all assignments. You will run this code and interpret the output. We also provide some optional online videos that walk through the intricacies of code we use to generate results in lecture slides and assignments.

For assignments, we give students the choice to learn programming or spend additional time practicing critical reading of quantitative work. We do this via a branching pathway toward the end of each graded assignment that leads to either a statistical programming task or a critical reading task. For each assignment, students can select the programming or critical reading path.

If you are interested in the coding path, we recommend attempting it from the first assignment.

This course primarily supports coding in R. Assignment code is also provided for Stata, but the optional videos that go deeper into coding are only available in R. You are welcome to use any other programming language you like to complete the assignments (e.g. SPSS, python), but we will only offer feedback on your results and interpretations, not your code.

We do not assume that you have used R or Stata before. For Stata users, please check that you have Stata version 12 or later.

This page offers installation instructions and learning resources for R and Stata.

Culminating Assessments: Final "Celebration of Learning" and Optional Course Projects

The final, our Celebration of Learning, is a partly collaborative and partly individual affair that will be posted during the examination period. It consists of two documents. The first document, the "evidentiary materials," contains an extended amount of R/Stata code and output. You may discuss this document in groups of any size. However, once you open the second document, no further discussion of *any part of the final* is allowed. To be clear: you may discuss the evidentiary materials in the first document with anyone until you open the second document. Once you open the second document, you should no longer discuss any of the Celebration of Learning materials with any other students.

The second document contains a set of questions like those in your other assignments. This document, once opened, must be completed individually, and no subsequent discussion of the evidentiary materials nor the content of the questions is allowed until the celebration is complete.

Most students begin the final by individually reviewing the first document for a short time. After individual review, they transition to group discussions, working together to ensure that all

members understand the evidentiary materials. Once they are comfortable with the evidentiary material, students open the second document of the celebration.

The questions sum to the magnitude of a large DAM, without coding. To provide flexibility, the assessment window is open for four days. We encourage you to plan so that final celebrations will be submitted on time. Please contact me early if you have any conflicts with this window.

For students who wish to apply the skills developed in this class to their own data – especially those interested in careers involving quantitative analysis – we encourage the submission of a course project. Projects will usually be 5-10 written pages. They should include a statement of the research questions, a methods section describing the analysis, and a writeup of the results and conclusions. Students who submit a course project can complete a shorter version of the celebration of learning (around 60% of the full version). Students who wish to develop their project over the course of the semester may also offset some DAM parts with progress reports on their project. Criteria for the optional final project follow:

1. Projects should demonstrate fluent application or conceptualization of course content beyond S-040 (i.e. you should use methods from this class not covered in intro courses)
2. Students may take on projects individually or in partnerships. Partnerships need not be the same as DAM partnerships.
3. Examples of projects include analysis of your own data and interpretation of results, extensions of DAMs, and development (including schematics if construction is not possible) of teaching tools.
4. Individuals or partners who take on a final project will complete ~60% of the final celebration of learning.
5. Projects should be around 5 pages double spaced, not including tables, figures, code, and references. Stylistically, many final projects will look like DAMs, with an introduction of the data and the context, a motivation of the question, exploratory data analysis, model building, testing of relevant hypotheses, and interpretation.

Slack, collaboration, and study groups

We use Slack for collaboration and communication. You can access the S-052 Slack by clicking on the link in the left-side navigation bar on Canvas. Standard Slack norms are here:

- Be good Slack citizens: answer questions, assume best intent, try not to sidetrack conversations.
- Understand **public and private channels**, and try to communicate with the proper audience.
- Use **threads** to organize smaller group discussions around specific topics.
- **Format messages** so that they're easy to read.
- **Use emoji reactions** liberally to acknowledge messages while keeping channels from becoming inundated with reaction messages 😊
- Use the search function to try to find an answer, reducing duplicate questions.
- Manage **notifications** to reduce information overload and maintain focus.
- Remember that there are situations when it's best to move the conversation off Slack messaging. Students can meet face-to-face using Zoom or **Slack calls.**

To mimic statistical work in the real world and to provide a chance for you to use statistical language actively, we mandate completion of DAMs in partnerships throughout the course.

Collaboration is mandated for three reasons. First, learning statistics is like learning a language. To learn it, one must speak it actively and in a genuine context with others. Second, collaborative statistical analysis is the norm and individual work is the exception in the world of statistical practice. Third, our experience has been that, on average, students who work in partnerships and groups 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. Note that the Week 7 DAM is an individual submission. We do this to ensure that individual as well as partnership learning goals are advancing at the midterm mark.

Larger study groups can be helpful as you prepare to do the assignments. **However, students must submit work as pairs, not work from a larger group. Papers should be written in your pair's 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 responses to members outside of the partnership. Second, never compose collaborative documents with members outside of the partnership; beyond the partnership, do not 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 require not only statistical techniques but also skills in analyzing and reporting the material.
- Use study groups to ask questions, try out interpretations, and so on. Often, a peer can explain something that makes you see the material in a new way. 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.
- 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. If the distinction begins to feel murky, refocus your group's work on lecture content and course materials.

Grading

You will be evaluated based on your engagement with learning checks and learning submissions (15%), the learning you demonstrate on DAMs (50%), and the final celebration plus (if applicable) your final project (35%).

We use arithmetic computations to arrive at a first approximation of your course grade. We then conduct several checks to ensure that no individual assignment or score takes on undue weight. We consider your growth as well as your average performance. And 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 celebration.

Avoiding plagiarism and ensuring appropriate use of Artificial Intelligence (e.g., ChatGPT)

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.

Please also read the school's policy on the appropriate use of generative artificial intelligence (AI), including tools like ChatGPT. Acknowledge and document any use of these tools in your DAM. You are welcome to use these tools to refine or translate your code and to refine or translate your ideas. However, you should "lead" and not "follow" any AI tool. The "seed" of every answer in any submitted assignment must be your own, and you must acknowledge and cite any way that AI has helped that seed to grow.

If you have any questions about what constitutes appropriate collaboration or use of AI, or how to define what constitutes your own work, please see me or a Teaching Fellow.

Accommodating Students with Disabilities

I and the Harvard Graduate School of Education strive to make all learning experiences accessible by providing reasonable accommodations for students with disabilities. If you anticipate or experience academic barriers based on your disability (including mental health, cognitive, learning, sensory, physical, chronic or temporary medical conditions) please contact our staff in the Office of Student Affairs, as soon as possible, to explore what arrangements need to be made to assure access to course work and the classroom learning experience. Please contact me if there are additional accommodations you would find useful.

Our use of electronic data on Google and Canvas

I am always trying to improve my teaching and your learning. As part of my effort to improve my teaching and your learning, I will use data from online 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 Harvard's Canvas Privacy Policy, linked here.

(Boilerplate) Required Information for Students Included in a Zoom Classroom Recording

Note that my travel may require me to conduct a class session by Zoom. If I do, I typically record the class so that students can review the lecture later. The following boilerplate applies to those recordings:

- Instructors may use Zoom to record class sessions.
- If an instructor uses Zoom to record a class session, Zoom provides audio and visual indicators to inform you when the recording starts, stops, is in progress, and is paused or unpaused.
- You, as a student, may not yourself record a class session. (More generally, you should not record a class session using any other technology.)
- You have the option to appear in an audio-only mode, such that your webcam is disabled (turned off) during the class.
- You have the option to access Zoom class sessions using a pseudonym.
- In order to facilitate class participation, you are expected to communicate any pseudonym to your instructor in advance of the class.
- Links to class session recordings, if available, will be posted in the Zoom meetings section of the Canvas course webpage. More generally, any class recording must only be posted inside of the Canvas site for the course.
- Links to Zoom class session recordings will be removed and videos deleted at the end of the academic term.
- You may not disclose the link to a class session recording or copies of recordings to anyone, for any reason. It is available to your class only.

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

Note: Some answers below reference dates for the fall version of the course.

Can I attend synchronous sessions as a guest or an auditor?
Unfortunately not. Synchronous attendance at lectures and sections is restricted to enrolled students. Because I have explicit learning goals related to creating community among enrolled students, I do not allow formal auditing or guest attendance. Students interested in asynchronous resources may email my assistant to receive guest access to the course website. I allow this access in certain cases, such as for alumni and for interested students with conflicts who cannot otherwise enroll.

I am not an HGSE student. When should I file a cross-registration petition?
Fall Cross-registration opens on August 25 at http://my.harvard.edu. I ask that you file your petition by Friday, September 1, at 12PM. When you cross-register, you may email me to let me know how you have satisfied the prerequisites, or I will reach out to you directly. There is no enrollment limit in the course.

I intend to cross-register. Do I need to demonstrate my prerequisites now?
Not before August 25. You should decide whether this course is best for you, including whether you meet prerequisites, below. Then you should file a cross-registration petition by Friday, September 1, at 12PM. I will then contact you to determine your eligibility. If you meet the requirements, I will approve you.

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 or S-040 a better choice for me?
Maybe! S-040 is offered in the fall by Dr. Joe McIntyre and assumes no statistical background. S-030 is offered in the spring by Dr. Hadas Eidelman and supports students with limited exposure to multiple regression. The depth of coverage of multiple regression in S-030 and S-040 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. Note: Students who take S-040 in the fall should enroll in S-052 in the spring, not S-030. You can see more here.

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?
We do not recommend this. S-052 moves quickly, demands deep conceptual understanding, and covers several advanced topics. We recommend that students at this level 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?
We do not recommend this. Our goal is not superficial acquisition of methods but deep conceptual mastery. For this, the foundations are necessary, and we recommend S-030 as a rigorous alternative.