Statistics 100b/500b: Introductory Statistics

Jay Emerson
Assistant Professor of Statistics
Yale University
john.emerson@yale.edu

TIME: Monday, Wednesday, and Friday 10:30-11:20.

CLASSROOM: Mason 211, subject to change based on course enrollment.

TEACHING FELLOWS (as of 1/7/06):

<table>
<thead>
<tr>
<th>Chandra Erdman</th>
<th>Walton Green</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:chandra.erdman@yale.edu">chandra.erdman@yale.edu</a></td>
<td><a href="mailto:walton.green@yale.edu">walton.green@yale.edu</a></td>
</tr>
</tbody>
</table>
Every day we are inundated with data. How do we recognize dishonest or even unintentionally distorted representations of quantitative information? How can we reconcile two medical studies with seemingly contradictory conclusions? How many observations do we need in order to make a sound decision? This course introduces statistical reasoning, emphasizing how Statistics can help us understand the world. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation and regression. Students will learn to apply statistical concepts to data and reach conclusions about real-world problems.

Students interested in a course that covers similar material at a faster pace and then focuses on applications to a particular field of study (such as Biology or Political Science) should consider taking one of the Stat 101a-Stat 106a courses instead. Students who already have been exposed to the course topics or who have extensive quantitative backgrounds should ask me about other course opportunities. For example, students seeking a more mathematically rigorous or computationally intensive introduction to probability and statistics could consider taking Stat 238a. Students with previous exposure to the concepts of statistics (for instance through AP Statistics) interested in focusing on the practice of data analysis should consider Stat 230b.

Do you remember anything at all from your high school algebra? If so, have no fear! This course is intended to be accessible to all students having a basic knowledge of high school algebra. Students will be challenged by statistical reasoning, not by mathematical manipulations. If you have any concerns about whether you are ready (or over-prepared) for this course, please speak with me.

This is an introductory course with no prerequisites. It is not intended for students majoring in mathematics or the sciences, or for students who have strong mathematical or computational experience.

**Topics**

The topics to be covered largely coincide with those of STAT 101-106a. Here, I provide a list of general topics (with a selection of specific topics and examples included) that reflects my philosophical approach to introductory statistics: statistics can be fun, useful, and accessible to all students.

- **Exploratory data analysis:** graphical and numerical exploration of real data with a heavy dose of common sense. Measures of center (mean, median) and spread (quartiles, variance and standard deviation).
- **Simulation** (used throughout the course to develop an understanding of randomness, even before an introduction to probability). As an example, students might consider the tradeoff between risk and reward in investments, using simulations to compare a portfolio of risky stocks with a savings account.
• Exploring relationships between variables (exploratory data analysis and an introduction to linear regression); correlation and causation.
• Gathering data (understanding randomness, probability, surveys and experiments).
• Randomness and probability (including the useful discrete and continuous probability distributions, expectations, conditional probability, Bayes Rule, and the Central Limit Theorem). These are excellent topics for use of simulation techniques.
• Sampling distributions: a healthy mix of probability theory and simulation helps students strengthen their understanding of randomness.
• Statistical inference: confidence intervals and hypothesis tests. Power, sample size calculations, and the associated tradeoffs.
• Inference for counts (were “women and children first” when the Titanic sank?), tables, regression, and analysis of variance.
• Multiple regression and logistic regression.
• Other topics motivated by real-world problems as time permits.

Reading and Supplementary Material

I will use a new book by Richard D. De Veaux, Paul Velleman, and David Bock, *Stats: Data and Models*. I will provide lecture notes for supplementary material. Students will regularly use Excel and/or Minitab and/or R on homework assignments. A nice review of this text is available online: [http://www.techbookreport.com/tbr0107.html](http://www.techbookreport.com/tbr0107.html).

I think this will be a terrific book – I hope you enjoy it! I will often refer to the book as *DVB* (for the authors).

*The Cartoon Guide to Statistics*, by Gonick and Smith, is a nice, fun, light, inexpensive optional text.
Teaching Fellows

Teaching Fellows will hold regular help sessions in the Statlab (see below).

Tutors

I have requested two peer tutors for the course. Details TBA.

Labs (Designed for, but not limited to, statistical computing help)

Call them what you will. Labs? TF office hours? Sections? Early in the semester, these will be held in the Statlab and will concentrate on statistical computing. Later, based on student demand, we may offer a more traditional TF review session instead. I have times reserved at the Statlab, Tuesday and Friday afternoons 3-5 PM. I expect these will both be impossible for some of you, and I will consult with the students and the TFs on the possibility of an alternate time, perhaps one evening a week.

My Review Sessions

I will offer weekly review sessions on Sunday evenings, 7-8:00 PM. I’m not insulted if you walk out at 7:30, and I’ve been known to stay as long as students have questions. I sometimes bring munchies, and coffee is available. Very casual.

Office Hours

TBA

Homework

Regular reading and homework is required. Homework may consist of online questions (distinct from the reading assessment questions), more traditional written (or electronically submitted) homework, and/or computational exercises. Homework must be submitted promptly (online assignments will have strict submission deadlines, and homework collected in class must be turned in at the beginning of class); late homeworks receive at most half credit unless accompanied by a Dean’s excuse. The homework assignments are an essential part of the course; they are the primary way of learning the material.
Grading

Homework 25%
Midterm exams or quizzes (during class hours) 30%
Final Exam (Tuesday, May 2, 2:00 PM) 30%
Submitting Reading Assessments 10%
Polling System Participation 5%

Please note, these last two “activities” are strongly encouraged but not strictly required. If you object to either of them, your grade will simply be based on homework and exam scores. However, I believe you will get the most out of this course by actively participating (in terms of quality of learning and the ultimate grade). Your final grade will not be lower than your combined homework and exam scores, and could very well be higher.

Homework will be a fairly broad category, with several different types of assignments. I’m planning three midterm exams (perhaps calling them quizzes is more appropriate). There will be a comprehensive, 3 hour final exam.

Reading assessments will be used regularly, to be completed online by 3 AM the morning before class, covering material in the assigned reading. They will be designed to take no more than 10 minutes of your time, once you have completed the assigned reading. If your participation percentage on the reading assessments is less than your combined homework, midterm and final exam averages, your grade will be based only on your exam and homework scores. Any participation percentage greater than 90% will receive full 100% credit; I reserve the right to lower this threshold (to your advantage) as I see fit, particularly if I use fewer reading assessments than originally intended.

The polling system is a new addition to the course – interactive learning is more effective (and, I think, enjoyable) than traditional lectures alone. The good news is that I convinced Yale to provide the “clickers” to students for free. Last semester, students were required to purchase them for use in “Natural Hazards” (Geology & Geophysics). As with the reading assessments, if your participation percentage is less than your combined midterm and final exam averages, your grade will be based only on your exam and homework scores. Again, any participation percentage greater than 90% will receive full 100% credit, with the same qualification as to the adjustment of this threshold based on our success using this polling system.

Simply, if you participate in the lectures and complete most of the reading assessments, 15% of your final grade will be an automatic 100%. For most students, this could make a difference of half a letter grade. For some students, the impact could be even greater.

The Course Website

http://classesv2.yale.edu
Computing

I have two documents, “Getting Started with R” and “Getting Started with Excel,” which explain the choice of these software packages. I expect that most of you have easy access to Excel, either on your own computer or in a cluster. R is even better: free! You can use either one. You can use both. You can use another package, like Minitab, Stata, SPSS, or SAS, but you’ll be on your own. We’ll teach you everything you need to know in Excel and R. For now, let me just make a few points:

- R is not a menu-driven interface, and at first it may be unsettlingly unfamiliar to you. Think of it as a glorified calculator. We’ll teach you everything you need to know.
- You will not be expected to do any serious “programming.” All my examples will be extremely well documented in my notes, and computational homework problems should be straightforward modifications of my examples.
- If you do pick up a little about R, don’t forget to add this to your CV/resume!

Expectations, hints, and other information and comments.

1. Regular reading, homework, class attendance and participation (via reading assessments and polling questions) is required. This is a lot of “little things,” I realize. Treat it like a foreign language course, where you need to improve steadily throughout the term: cramming at the last minute will not be successful.
2. Here is a reasonable guideline: spend at least 2 hours on your own for every class meeting, doing the reading and the problem sets. A steady, consistent effort is the key to doing well, and should make the class more enjoyable.
3. Working together: I encourage you to work in groups or with a friend, but all problem set solutions must be your own (written and computational).
4. I am confident that the material learned will be useful in future courses.
5. Many students use statistical research as part of their senior essays.
6. Knowledge of statistics can be invaluable in summer internships.
7. Knowledge of statistics and statistical computing is one of the most practical skills to have in today’s world. It will help you on the job market, either in academia, industry, government, or business. The National Science Foundation predicts a shortage of statisticians and a high demand for statistical services!

Citation and Thanks

I would like to thank my father/colleague, John Emerson of Middlebury College, for his teaching, support, and guidance over the years. I have adopted some of his ideas, organization, and a few of the hints, above, in my syllabus.