Statistics 100b/500b: Introductory Statistics

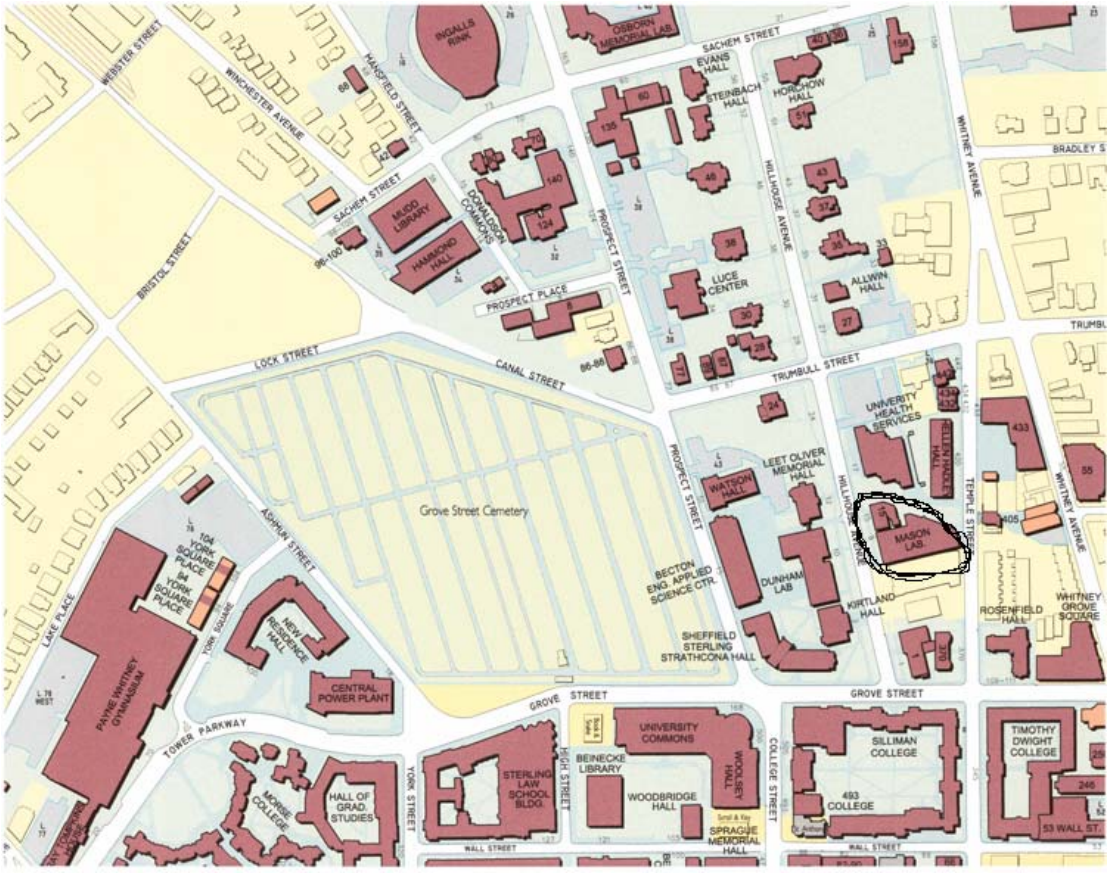
Jay Emerson
Assistant Professor of Statistics
Yale University
john.emerson@yale.edu

TIME: Monday, Wednesday, and Friday 10:30-11:20 (with a special note regarding the Friday classes).

CLASSROOM: Mason 211. Note the change from the original location, OML 203, which is not a lecture hall. We apologize for the confusion.

TEACHIING FELLOWS (as of 1/9):

Anna Kochetkova **anna.kochetkova@yale.edu**	**Qi Yan** **qi.yan@yale.edu**



This course offers a foundation for statistical reasoning, emphasizing how statistics can help us to better 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 how to apply statistical concepts to data, analyzing and reaching conclusions to real-world problems. There are no prerequisites, and the course is intended for non-majors.

I will assume only that students are familiar with elementary algebraic notation at the high school level. Regular problem sets (often daily, very short) will involve a heavy dose of common sense, traditional mathematical exercises, data analysis, and basic simulation techniques. Two short midterm exams (in addition to the final exam) will help students synthesize the material. Upon completion of this course, students should be able to think critically about data, to use graphical and numerical summaries, to apply standard statistical inference procedures, and to draw conclusions from such analyses.

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.

Schedule

Please see

http://www.stat.yale.edu/~jay/100/schedule.pdf

for the most recent version of the posted class schedule including homework assignments. The book chapters are short and easy (and fun) to read. Homework assignments should be similarly short and not overly taxing if you stay on top of the material.

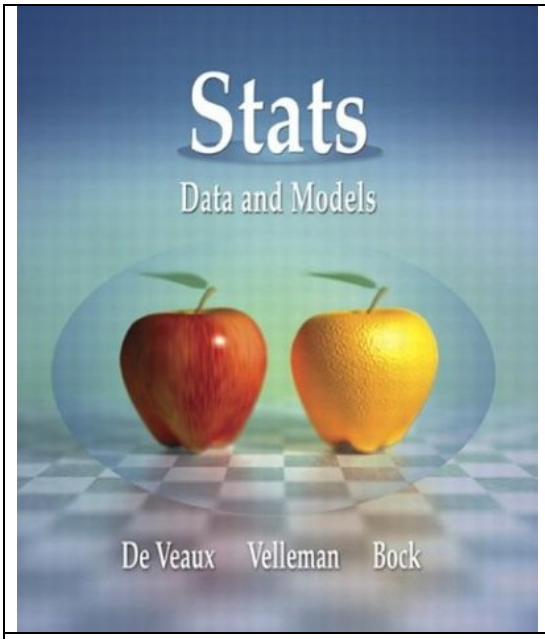
Topics

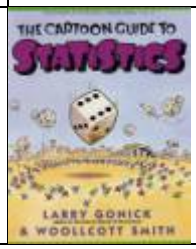
The topics to be covered largely coincide with those of STAT 101-106a. Here, I provide a list of general topics (with a selection of specific topics and examples included) that reflects my philosophical approach to introductory statistics: statistics can be fun, useful, and accessible to all students.

- 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 Supplementay Material

<table>
<tr>
<td> 
</td>
<td>I will use a new book by Richard D. De Veaux, Paul Velleman, and David Bock, Stats: Data and Models. I will provides lecture notes for supplentary 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.

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).</td>
</tr>
<tr>
<td>The Cartoon Guide to Statistics, by Gonick and Smith, is a nice, fun, light, inexpensive optional text.</td>
<td> 
</td>
</tr>
</table>

Special Format (tentatively, based on course enrollment)

The class is scheduled to meet MWF 10:30-11:20. Depending on the course enrollment, I might try to split the Friday class into smaller sections. On Fridays, you may have the option* of attending class at times other than 10:30-11:20, although I would continue to hold class meetings in during this regular time slot. This could serve several purposes. First, it would allow computational labs to be held in the StatLab facility (which would be appropriate for groups of 20-30 students). Second, it would allow a higher degree of student participation, which I have found to be effective in the past. I will teach these sections with the help of the Teaching Fellows. (* or, a time shift may be required if the enrollment is extremely large, sorry about that!)

Alternatively (and I'm still exploring this option as of 1/6), I might lecture in Mason Lab 211, with an A/V connection to the Statlab and the Garage (or is it the Jungle?) in Dunham, where students can follow the lecture and work along with me on computation examples and exercises. Stay tuned for more information!

Grading (tentative)

Homework	30%
Midterm Exams (during class hours)	20% each
Final Exam (5/6/05, 2:00?)	30%

Review Sessions / Discussion Sections

I will hold occasional (or even weekly) optional review sections as announced.

The Course Website

I will not be using classes.yale.edu. Rather, I will be helping test the next generation, called Sakai (available easily as http://sakai.yale.edu). However, I ask that you add STAT 100b to your classes.yale.edu list in addition to logging into Sakai. Early in the course, I may make use of the classes.yale.edu email lists until I'm sure everything is working with Sakai.

Computing

I plan to use a free program called **R** in the course. It is quite similar to the commercial software package **S-Plus**. Think of it as a calculator on steroids. I have several reasons for using **R** instead of **Minitab**, **Stata, Data Desk**, or **Excel**, for example. I'll spend some time talking about this in class. 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 want to use another software package, great! In fact, I encourage it. My only word of caution is that it may be difficult to get help (at least, quickly) if you run into troubles.
- Don't forget to add this to your resume or CV! It is much more impressive than knowing **Excel**.

If you want to get started, you can download and install **R** directly from the web, at

http://www.r-project.org/

Follow the "Download > CRAN" link on the left side of the page. In the list of "mirror sites" choose something in the USA close to home, maybe the site in Boston,

http://www.bioconductor.org/CRAN/

Note that there are versions for Windows, the Mac, and Linux. Under "Precompiled Binary Versions", most of you will choose "Windows (95 and later)". You want the "base" package, and all you really need is the setup program, "rw2001.exe". I may even be able to give you the link right here:

http://bioconductor.org/CRAN/bin/windows/base/rw2001.exe

Save it to your desktop. Double-click it to start the installation. There are some very silly Bill Gates-ish questions along the way. Yes, of course you accept the licensing agreement! In fact, none of the questions require any guidance from me: do the obvious things and let it go. If you have any questions, of course, come to us for help!

There are two references I might suggest for **R**, but please not that these are not required. They may be helpful, but strictly will not be necessary. One, however, is free, from the R Project (and I'll make it available on the web):

http://www.stat.yale.edu/~jay/100/R-intro.pdf

The other reference is a book by Peter Dalgaard, *Introductory Statistics with R*. It is available from Amazon.com (for example, though I'm sure other places online have it as well):

http://www.amazon.com/exec/obidos/tg/detail/-/0387954759/qid=1105303537/sr=8-1/ref=sr_8_xs_ap_i1_xgl14/102-1907237-0337701?v=glance&s=books&n=507846

More Computing

The textbook comes with a CD-ROM containing ActivStats and and Excel plug-in called Data Desk (DDXL). I've looked a bit at ActivStats, and it has some interesting features that students might find helpful. I may also try out **DDXL** now and then, and (as noted in the previous section) students are welcome to use this as an alternative to **R**.

Expectations, hints, and other information and comments.

1. Daily reading and regular homework assignments (some written, some computational) are required of everyone. Homework must be turned in at the beginning of class; late homeworks receive at most half credit; the lowest homework grade will be dropped. The homework is an essential part of the course; they are the primary way of learning the material.
2. Attendance at all classes, including the Friday lab classes, is required.
3. Here is a reasonable guideline: spend at least 2 hours on your own for every class/lab 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.
4. 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).
5. I am confident that the material learned will be useful in future courses.
6. Many students use statistical research as part of their senior essays.
7. Knowledge of statistics can be invaluable in summer internships.
8. 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 Foundataion 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. We are both offering similar courses this semester (using the same textbook), and I have adopted some of his ideas, organization, and a few of the hints, above, in my syllabus.