AMS 5: Statistics

- **Lectures**: MWF 11am–12.10pm, in the Media Theater M110.

- **Instructor**: David Draper, Baskin Engineering (BE) 135; telephone 459–1295; email draper@ams.ucsc.edu (I’ll do my best, but due to the volume of email I receive, I can’t guarantee quick response to any message you send me).

- **Web Page**: There is a course web page: its URL is

  http://www.soe.ucsc.edu/classes/ams005/Spring06/

- **Instructor Office Hours**: will be announced soon (initially my office hours will be held in BE 135; depending on how many people show up, we’ll often move to a bigger space nearby called Jack’s Lounge).

- **TAs**: The TAs for the class are

<table>
<thead>
<tr>
<th>Name</th>
<th>Email address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethan Arenson</td>
<td><a href="mailto:ethan@ams.ucsc.edu">ethan@ams.ucsc.edu</a></td>
</tr>
<tr>
<td>Vaneet Batish</td>
<td><a href="mailto:vaneet@soe.ucsc.edu">vaneet@soe.ucsc.edu</a></td>
</tr>
<tr>
<td>Ken Hullett</td>
<td><a href="mailto:khullett@soe.ucsc.edu">khullett@soe.ucsc.edu</a></td>
</tr>
</tbody>
</table>

  Their office hours will also be announced in class soon (you can meet the TAs outside BE 142 at the start of their office hours).

- **Discussion Sections**: These have already been arranged, and you’re required to enroll in one of them as part of taking the class.

<table>
<thead>
<tr>
<th>Section</th>
<th>Day</th>
<th>Time</th>
<th>Place</th>
<th>Enrolled</th>
<th>Max</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>01A</td>
<td>Mon</td>
<td>2–3:10pm</td>
<td>BE 372</td>
<td>35</td>
<td>35</td>
<td>closed</td>
</tr>
<tr>
<td>01B</td>
<td>Mon</td>
<td>3:30–4:40pm</td>
<td>BE 372</td>
<td>31</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01C</td>
<td>Mon</td>
<td>5:00–6:10pm</td>
<td>BE 372</td>
<td>29</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01D</td>
<td>Tue</td>
<td>8:30–9:40am</td>
<td>BE 372</td>
<td>9</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01E</td>
<td>Tue</td>
<td>10–11:10am</td>
<td>BE 372</td>
<td>33</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01F</td>
<td>Tue</td>
<td>6–7:10pm</td>
<td>BE 372</td>
<td>26</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01G</td>
<td>Wed</td>
<td>8–9:10am</td>
<td>BE 372</td>
<td>8</td>
<td>0</td>
<td>closed</td>
</tr>
<tr>
<td>01H</td>
<td>Wed</td>
<td>5–6:10pm</td>
<td>BE 372</td>
<td>29</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01I</td>
<td>Wed</td>
<td>6:30–7:40pm</td>
<td>BE 372</td>
<td>11</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01J</td>
<td>Thu</td>
<td>4–5:10pm</td>
<td>BE 372</td>
<td>31</td>
<td>35</td>
<td>open</td>
</tr>
<tr>
<td>01K</td>
<td>Fri</td>
<td>8–9:10am</td>
<td>BE 372</td>
<td>2</td>
<td>0</td>
<td>closed</td>
</tr>
<tr>
<td>01L</td>
<td>Fri</td>
<td>9:30–10:40am</td>
<td>BE 372</td>
<td>25</td>
<td>35</td>
<td>open</td>
</tr>
</tbody>
</table>
The content of the course will be presented in four weekly meetings: the MWF lectures and a 70-minute discussion section. Ten discussion sections will be given each week, and it’s your responsibility to attend one of these (quizzes that are a part of the grade will be given in these discussion sections). To keep the class sizes roughly uniform (at 25–30 people) I ask you to regularly go to the section you’re enrolled in, but from time to time you can go to another one if you need to.

As you can see from the table above, discussion sections are given every day of the week from Monday to Friday. However, as far as the discussion sections in this course are concerned, we’ll be using a different calendar: discussion section week 1 runs from Thu 6 Apr through Wed 12 Apr inclusive, discussion section week 2 runs from Thu 13 Apr through Wed 19 Apr inclusive, and so on, with one exception toward the end of the quarter which I’ll explain later (Mon 29 May is a holiday, so we have to do something a bit different during the ninth week of classes). There are two conclusions to be drawn from this: (a) required attendance in discussion sections begins tomorrow, Thu 6 Apr, and (b) new content in the discussion sections will begin every Thursday this quarter.

- **Individual tutoring:** There will be a relatively small number of hours of individual tutoring available for those that most need it. You should get the great majority of your help in this course by coming to class, discussions sections, and the office hours that the TAs and I will give; it’s best to regard the modest availability of individual tutoring as a last resort after these other resources prove insufficient. If you feel you must have individual tutoring, please see me to request this. Tutoring is also available for qualified students through the Multicultural Engineering Program (MEP); see mep.soe.ucsc.edu for details.

- **Reader:** Given the size of the class and the number of handouts we need to work with for best understanding of the material, I’ve prepared a reader which will include all of the handouts. This will be available as soon as it’s ready (on Friday, probably) in class at xerox cost ($10, in cash).

### General Content

**Statistics** is the **study of uncertainty**: how to measure it, and what to do about it. It comes up in two kinds of things people do:

- **Science** (knowledge for its own sake), and

- **Decision-making** (putting that knowledge to work to make a choice among different possible actions).

Science is mostly about *facts* (“until 1999, when physicists decided to make it a definition instead of something to be measured, the speed of light was thought to be about 299,792,458 meters/second, give or take about 2 meters/second”) and *relationships* (like
Newton’s law relating force \( F \) to mass \( m \) and acceleration \( a \), \( F = ma \), which holds at least to a good approximation. Statistics is helpful with both: coming up with estimates and give-or-takes (measures of uncertainty) about facts, and identifying which relationships are causal (“Smoking causes lung cancer and heart disease”) and which are spurious (“Drinking soda pop causes polio,” or so they thought for awhile back in the 1930s). Along the way we’ll learn some of the most important basic rules of probability, which is the part of mathematics devoted to quantifying uncertainty.

Decision-making is mostly about predicting the future under different sets of conditions (An example from when I gave this class in the fall of 2002: “If we attack Iraq I predict that Saddam Hussein will be killed or captured in about 2 months at the cost of about 100 American military lives and the possibility of a strong terrorist backlash against civilian lives in the U.S. and elsewhere ...”) and choosing your favorite future (“... whereas if we apply diplomatic pressure with our allies and achieve a strong program of detection and destruction of Iraq’s [alleged] weapons of mass destruction I predict that Saddam Hussein’s threat will be neutralized at the cost of close to zero American and other lives. I like the latter better than the former”). Statistics has a lot to say both about how to predict things and how to figure out how accurate your predictions are likely to be. (Note that my predictions under the first scenario listed above weren’t very accurate: it took about 9 months to find Saddam Hussein, and the current casualty toll for American soldiers as of early April 2006 (to say nothing of the Iraqi civilian casualties, which are estimated to be at least 33,814) is 2,332 dead and 17,381 wounded; see icasualties.org/oif for details.)

Statistics is good both for telling you how much (or little) you know about something and for figuring out how to design experiments or sample surveys to get new information to reduce your uncertainty. There’s a lot of emphasis on good graphics: drawing pictures of your data that provide insight not readily found just by looking at the numbers (example: scatterplot of polio deaths against soft drink consumption). Statistics includes both descriptive methods to summarize facts (“The death rate within 30 days of admission for patients aged 65 and over with a principal diagnosis of heart attack at these 10 hospitals from January through July 2005 was 17%”) and methods to draw inference about counterfactuals (“I’m pretty sure that I would have gotten there faster if I had taken Soquel instead of the freeway”). Statistics uses math, mainly probability, but common sense and good judgment are at least as important as math in most good statistical work. This class will remind you more of a logic course—figuring out what’s valid to conclude from a given set of evidence, and what’s not valid—than a math course. A long time ago (in the late 1700s) a really good mathematician, Pierre Simon de Laplace, put it best:

\[
\text{Statistics is common sense reduced to calculation.}
\]

If any of this sounds relevant to your interests, maybe this course is for you.
General Style

The course will be based on a series of case studies drawn from my own consulting work and that of people whose work I’m familiar with. These case studies will mainly come from the natural and social sciences and medicine, but there will also (for example) be decision-theory examples from business and other fields. The case studies typically have four components:

1. In the first step we fully examine the real-world problem and make the central question(s) clear.

2. Then we “invent” one or more methods to solve the problem in step (1).

3. Next we apply the methods from step (2) to completely solve the problem and understand the real-world implications of the solution.

4. Finally, we stand back and examine the general properties of the methods “invented” in step (2): what other kinds of problems can they help to solve? Under what conditions do they work best, and what does it take to make them fail?

I like to help people learn in an interactive fashion, with questions and answers going back and forth between you and me on a regular basis during the “lectures.” In this manner we will trace the discovery process that led to the original development of the methods we study. The idea is for some real learning to occur in class, in addition to note-taking.

Texts

- Required: Draper, David (2006), *Thinking About Uncertainty: An Introduction to Probability and Statistics*. The first 11 chapters of a book (DD for short) I’m writing for ideas/methods courses like this one; the first five weeks of the class will be drawn pretty directly from this. Available at xerox cost in class ($20, in cash).

- Required: Freedman, David; Pisani, Robert; and Purves, Roger (1998), *Statistics*, Third Edition. New York: Norton. An excellent book (FPP) for an ideas course, with many good examples (some of which I’ll use in class). This is available now at Slug Books: 224 Cardiff Place, Santa Cruz (phone 469 7584). They have some used copies of FPP in addition to the new ones, and if you buy a new one you’ll be able to sell it back to them or to the UCSC bookstore when the course is over if you want; you can also get used copies at half.com and other discount book-sellers on the web. You’ll get the most out of reading from the third edition of this book; earlier editions are cheaper but will only give you about 50–75% of what you’ll get from the latest edition.
Course Prerequisites and General Education Codes

There are no formal prerequisites, but you should be comfortable with high school mathematics at roughly the level of college algebra; in particular, no calculus will be used in this class. If you have any questions about whether you satisfy these informal prerequisites, please see me.

This course satisfies the following General Education Codes: IN and Q.

Course Requirements and Grades

My basic approach to grades is to try to get everybody to work hard to absorb as much of the material as they can in one quarter and then give the best grades I can, more or less consistent with past grading standards for the course. (The grade distribution is usually approximately 25% A, 40% B, 25% C, 5% D, 5% F, give or take 5 or 10 percentage points in each of the first three categories). You’ll probably notice that the material in the course is cumulative, and that its difficulty level rises slowly each week to a peak at around week 7 or 8. The final grade will have four components: homework, midterm, discussion sections, and final exam.

- **Homework** (about 25%) will be assigned about 6 times during the quarter and due at the beginning of class 1–1\frac{1}{2} weeks later. In order that solutions be available promptly, and because of the procedural problems inherent in the grading for such a large class, LATE HOMEWORK WILL NOT BE ACCEPTED FOR ANY REASON. To compensate for emergencies or bad luck, your lowest homework score will be dropped from the grade computation (each homework will have about the same weight). Note that none of the homework assignments is optional.

  One possible strategy in view of the dropping of the lowest score is of course to neglect to turn in an assignment, but people who have done this in the past have noticed that they are unprepared on the corresponding material at exam time.

  The purpose of the homework is to develop facility in statistical thinking through regular practice, and to provide early and regular feedback on your performance in the course. Solutions to the homework will be posted shortly after the due date in a glass case in BE. You should consult these solutions and compare them to your own, and talk with the TAs and/or with me to resolve any confusions that remain. There is an enormous volume of homework that the graders must examine in a short time, and it’s impossible for them to make detailed comments on each paper and still return them quickly enough to be useful to you. For this reason you should examine the posted solutions carefully, even if you receive a high score on your paper, since there may be some ideas you’ve not fully understood.

- **Midterm** (about 25%). This will be a take-home open-book open-notes exam given out around the end of the fifth week and due a week later. This will not come early enough for you to use it in any decision you might need to make about dropping the course, but you should have enough feedback from the homework and quizzes by then to make that decision.
• **Discussion sections** (about 20%). Statistics is something you learn by doing, so it’s important to work a lot of problems, both by yourself and by talking with other people. You’ve already enrolled into a discussion section; attendance at these sections is required. The idea is to have sessions in which the TAs lead the discussion on how to solve some problems, chosen to illustrate in practice the topics being considered in lecture at that time. There will typically be one problem like the ones solved in the discussion section or like what’s going on in class at the time; you’ll be asked to solve this problem (open-book, open-notes) and turn your solution in for credit as a kind of small quiz.

• **Final exam** (about 30%). The final will be an in-class, open-book, open-notes exam. It will be cumulative, but with emphasis on the material after the midterm.

Two final notes about grades:

• Incompletes will be given only in clear cases of emergency.

• Anybody who is a senior and who needs to pass this course by the end of this year to graduate should start working today—waiting til nearly the last minute to take the course does not guarantee a passing grade.

**Collaboration, Plagiarism, and Cheating**

You’re encouraged to form study groups for the purpose of discussing the homework problems, but **all of the written work you turn in for this class must be your own efforts**. Even though the volume of homework the graders will be evaluating is large, it’s surprisingly easy to spot instances where someone has simply copied someone else’s solution, and this will be even easier to identify with the take-home midterm (unlike the homework, **you’re not allowed to discuss the take-home midterm problems with anybody else**). In fairness to the many people who do not cheat, instances of plagiarism and other forms of cheating will be dealt with as harshly as possible in the UCSC system.

**Calculators**

Everybody should have available a **calculator** (with charged batteries or solar power) for use during exams and discussion sections. It’s important that this machine have a square root key in addition to the usual arithmetic operations, and it’s helpful to have at least one memory. You can find calculators like this for $5–10 (or even less) these days.
Lectures, Discussion Sections, and Readings

You’re responsible for everything that goes on in class and discussion sections and for obtaining any written material that’s distributed. The TAs and I will often refer back in lectures and discussion sections to handouts originally covered in previous classes, so I recommend that once the reader is ready you should get one and bring it to all lectures and discussion sections.

You should do the assigned readings before coming to class or discussion section. Ordinarily, the lecture will discuss aspects of the readings in detail or will present additional material not contained in the texts. Neither the lectures nor the readings can be substituted for one another. The discussion sections will sometimes introduce new material and will involve turning in some written work at their conclusion, so regular nonattendance will clearly hurt your chances of performing well. There is a strong causal relationship in this class between \{taking all of the homeworks and quizzes seriously\} and \{getting a good grade\}.

Preparing Homework

Here are some guidelines for getting your homework ready to turn in; please follow them. The graders have an amazingly small amount of time to look at your paper and pass judgment on it—anything you can do to improve its form, by making it relatively neat and easy to follow, will maximize your chance of a good grade on the homework.

• Submit homework on 8.5 by 11 or 8.5 by 14 paper only, and make sure that your name is clearly printed on all pages of anything you turn in.

• Use staples or paper clips to hold together submissions of more than one page.

• Write legibly and coherently. Manuscripts that are unintelligible in either content or handwriting are not likely to be looked on favorably.