University of California, Santa Cruz
Department of Applied Mathematics and Statistics
Baskin School of Engineering
Fall 2018

AMS 7: Statistical Methods for the Biological, Environmental and Health Sciences

- **Instructor:** David Draper (office BE 357C); telephone 459–1295; email

  `draper@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; please put *AMS 7 student, Fall 2018* in the subject line of any email message to me).

- **Background:** AMS 7 is a 5–unit class with lectures and discussion sections devoted to *statistical methods for the biological, environmental and health sciences*, and AMS 7L is a separate 2–unit computing lab connected to AMS 7, in which you'll get a chance to do hands-on statistical data analyses using a package called `JMP` that's popular in biology, the health sciences and environmental studies. Concurrent enrollment in AMS 7 and 7L is required.

  AMS 7 and 7L are required courses for undergraduates majoring in Ecology and Evolutionary Biology (EEB) and Environmental Studies (ES), and they're strongly recommended courses (a) for the Health Sciences and Environmental Toxicology majors and (b) in a number of degree programs in the Molecular, Cell and Developmental Biology (MCDB) Department; in particular, MCDB students can use AMS 7/7L to satisfy one of their laboratory course requirements.

  Also, one of the General Education requirements is that all undergraduate students must take one or more classes in *Statistical Reasoning* (SR); AMS 7 fulfills this requirement.

  Basically, if you want to know more about (a) how to gather data to decrease uncertainty about something of interest to you and (b) what can validly be concluded from a given body of evidence (data + logic + reasonable assumptions and judgments), then you should seriously consider taking this class.

- **Lectures** for AMS 7: TuTh 11.40am–1.15pm, in Baskin Auditorium (room 101).

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

  `ams007-fall18-01.courses.soe.ucsc.edu`

  It will be regularly updated with announcements and handwritten lecture notes created on the document camera during classes, which supplement the official lecture notes (more on that below).

- **TAs:** There are three TAs for the class:

Name	Email Address	Office
Raquel Barata (Head TA)	`rbarata@ucsc.edu`	BE 356
Zee Hu	`zhu95@ucsc.edu`	BE 354
Sarah Jarvis	`sagjarvi@ucsc.edu`	BE 356

If you have a question about a grade you got on any written work in AMS 7, please speak with the Head TA (Raquel Barata). Note, however, that she has the right both to add points to your score and to subtract points if she sees other problems on your paper that were incorrectly graded in your favor, so bring grading questions to Raquel sparingly.

- **Webcasting:** The lectures in this class will be **webcast**, so that you can watch videos of the lectures after they occur: many students find this helpful in reinforcing the learning that goes on in class. The URL for the site where the videos will be available is

  `webcast.ucsc.edu`

  There is a row called **AMS 7** in the table of available classes; click on `Video List` in that row in the table. You'll be taken to a page headed `Welcome to Opencast`; in the top yellow window type the user name `ams-7-1`, and in the second yellow window type the password `ams-7-01-fall-2018`. You'll then go to a page that lists all of the lectures recorded so far; click on the one you want, and use the arrow keys in the resulting screen to watch, or pause, or rewind.

- **Office Hours** are as follows (RB = Raquel Barata, DD = David Draper, ZH = Zee Hu, SJ = Sarah Jarvis, TBA = to be announced; *Jack's Lounge* the area with white boards on the ground floor of Baskin Engineering, at the opposite end of the building from the coffee kiosk):

Day	Time	Who	Where
Mon	10–11am	RB	TBA
Mon	5–6pm	RB	TBA
Tue	10–11am	SJ	TBA
Tue	1.30–2.30pm	DD	Jack's Lounge
Wed	1.30–2.30pm	ZH	TBA
Thu	1.30–2.30pm	DD	Jack's Lounge
Fri	10–11am	ZH	TBA

  The purpose of the office hours is to help you with the homework assignments and midterm and final exams. The first homework assignment will be handed out on **Tue 2 Oct 2018** in class; we won't hold any office hours this quarter until the week of 1–5 Oct 2018, but from that point on we'll be available to you each weekday of the quarter except for holidays.

  I will hold extra office hours (probably including on weekends) during the periods of time in which you'll be working on the take-home midterm and final (more about these tests below).

- **Enrolling:** We can't go above about 250 people, but up to that limit I'm happy to give permission codes to anybody who needs to take this class this quarter.

- **Discussion Sections:** These have already been arranged, and you're required to enroll in one of them as part of taking the class (E2 = Engineering 2 building):

Section	Day	Time	Place	Instructor
01F	Mo	8–9.05am	E2 194	DD
01A	Mo	noon–1.05pm	Cowell Com 134	RB
01B	Mo	1.20–2.25pm	E2 194	RB
01C	We	2.40–3.45pm	E2 194	SJ
01D	We	4–5.05pm	E2 194	ZH
01E	Fr	8–9.05am	E2 194	ZH

- **Lab Instructors:** There are two lab instructors for AMS 7L:

Name	Section	Email Address	Office
Sharmistha Guha (SG)	AMS 7L–01	`shguha@ucsc.edu`	BE 354
Matt Heiner (MH)	AMS 7L–02	`mheiner@ucsc.edu`	BE 354

- **Lab Sections:** These have also already been arranged, and you're also required to enroll in one of them as part of taking AMS 7L together with AMS 7.

> **AMS 7L is an entirely separate course from AMS 7, taught by a different instructor; for details on enrollment and other matters, please email one or both of the AMS 7L lab instructors listed above.**

- **Structure:** The content of AMS 7 will be presented in three weekly meetings: the TuTh lectures and a 65–minute discussion section. It's your responsibility to attend one of the discussion sections (**quizzes that are a part of the grade for AMS 7 will be given in discussion sections every week**). To keep the class sizes roughly uniform 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.

  Discussion sections will start on Mon 1 Oct 2018 and will continue every week thereafter; new material in the discussion sections will be presented each Mon, and the same material will be covered Mon–Tue–Wed–Thu–Fri; please go to your chosen discussion section next week.

  One holiday to note this quarter: there will be no lecture on **Thu 22 Nov 2018** (Thanksgiving); there will be **no discussion sections for the entire Thanksgiving week (19–23 Nov 2018)**; and there will be **no office hours on Thu and Fri of that week (22–23 Nov 2018)**.

- **Individual tutoring:** A small number of hours of individual tutoring will be available for those who 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 would benefit from individual tutoring, please see me to request this. Tutoring is also available through

  (a) the Multicultural Engineering Program (MEP; see `mep.soe.ucsc.edu` for details), and

  (b) Learning Support Services (LSS, through their *Modified Supplemental Instruction (MSI)* program; see

  `lss.ucsc.edu/programs/modified-supplemental-instruction`

  for more information). **Many students who found themselves struggling a bit in this class in the past received significant help from the MEP tutors.**

Readings

There is only one required set of readings for the course:

- Draper D (2018). *Statistical Methods for the Biological, Environmental and Health Sciences.* Course materials packet (approximately 750 pages) containing draft book manuscripts, lecture notes and reader; available soon at the **Bay Tree Bookstore** at photocopy cost.

I'll also draw some examples and case studies from

- Triola MM, Triola MF (2006). *Biostatistics For the Biological and Health Sciences.* Boston MA: Pearson-Addison Wesley, and
- Zar JH (2009). *Biostatistical Analysis* (fifth edition). Upper Saddle River, NJ: Prentice Hall;

these books are not required (and in fact Zar is not even recommended; it does have some good examples, but it's far too dry and difficult to read, and it contains some colossal mis-statements of fact).

Course Prerequisites and General Education Codes

The formal prerequisites for the class are as follows:

> Score of 300 or higher on the mathematics placement examination (MPE), or course AMS 2 or 3 or 6 or 11A or 15A or Mathematics 3 or 11A or 19A, or by permission of instructor.

I'm going to try to give a permission code to anybody who needs to take the class this quarter (subject to available seats in the lecture room and discussion sections); basically 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 (but there will be liberal use of formulas involving summation notation, which I'll review soon). If you have any questions about whether you satisfy these prerequisites, please see me.

As noted above, this course satisfies the **SR** (statistical reasoning) General Education requirement.

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 25–35% A, 35–45% B, 20–30% C, 0–10% D/F; anyone who sincerely tries in this class — by turning in every assignment and taking every quiz and exam, and demonstrating a basic level of understanding of the material — will pass the course.) **The material in the course is cumulative, and you'll probably notice that its difficulty level rises slowly each week up through about week 6 and then stays roughly constant thereafter.** The final grade for AMS 7 will have four components: homework, midterm, discussion sections, and final exam.

- **Homework** (25%) will be assigned 4 times during the quarter and due 1–1$\frac{1}{2}$ weeks later. Because of the procedural problems inherent in the grading for a large class, **late homework will only rarely be accepted**, based on reasons such as (severe) illness. 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. For copyright reasons it's not possible to post homework solutions on the course web page. In

the past I put solutions in a glass case in BE for students to look at, but — because of problems with cheating (people took pictures of the solutions with their cell phones and then tried to turn perfect homework papers in late, claiming illness) — I've been instructed by the Engineering School not to post solutions in the glass case any more.

There's 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. So here's the current feedback system: when you get your graded homework, quiz and midterm papers back, if you have any questions about the right answers please bring them to us in office hours; if you think you've been mis-graded please speak with the Head TA, Raquel Barata, who's authorized to correct grading errors (but please note that (as mentioned above) when you bring your paper to her, she has the right to notice two kinds of errors — those that are in your favor, and those that are against you — so don't over-use this opportunity or you may find yourself sometimes with fewer points than when you started asking about the grading).

- **Midterm** (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** (20%). Statistics is something that people 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** (30%). I want to give a take-home open-book open-notes final, but that will depend on you guys: if there's too much cheating on the midterm I'll be forced to give an in-class (open-book, open-notes) final. Either way it will be cumulative, but with emphasis on the material after the midterm.

You'll be submitting your homework and midterm (and probably your take-home final too) in this class at `canvas.ucsc.edu`, as PDF files. If you have a smart phone, you can get any of a number of free apps for taking photos of your written work and creating a PDF file from the photos (a good choice is called `CamScanner`) — once the app converts your photos of each page of your document into a single PDF file, you can email it to yourself and upload it to `canvas`. If you don't have a smart phone, you can go to any photocopier on campus, scan your written work into a single PDF file, and then email the file to yourself and upload it to `canvas` (as in the smart-phone approach).

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, and that also applies to the final if it's take-home**).

In fairness to the many people who do not cheat, instances of plagiarism and other forms of cheating will be dealt with vigorously. For example, the first time (say) three people are caught turning in exactly the same solution to a homework problem and that solution (if it had not been part of an instance of plagiarism) would receive (say) 21 out of the possible 24 points, each of the three people will receive $\frac{21}{3} = 7$ points ($\frac{7}{24} \doteq 29\%$, a failing grade) on that problem; the second and subsequent instances of this kind will be reported to the relevant College Provosts.

If you work in a study group, here's how to avoid plagiarism on the homework: you can talk about the solutions to the problems with the other people in your study group, but then each of you has to go away and write your answers out on your own: your write-up of the homework paper you turn in must be entirely your own effort.

Calculators

Everybody should have available a calculator (with charged batteries or solar power) for use during discussion sections (you'll need this to take the quizzes) and class.

All smart phones have built-in calculators, although you may need to rotate your phone through a 90° angle to get it to calculate square roots, which will come up often in this class. If you don't have a smart phone, cheap calculators with a square root key are readily available for purchase online, or you can type expressions such as `27 / sqrt( 6 )` at `google.com` and thereby gain access to a free online calculator.

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 you put the course materials packet in two ring binders — one for the draft books, the other for the lecture notes and reader — and bring the lecture notes/reader binder 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 readings. 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 for credit at their conclusion, so regular non-attendance will clearly hurt your chances of performing well. **It has been amply demonstrated in the past that there is a strong cause-and-effect relationship in this class between {taking all of the homeworks and quizzes seriously} and {getting a good grade}**.

Preparing Written Work for Submission

Here are some guidelines for getting your homework, midterm and final 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 written work in this class.

- Make sure that your **name** is **clearly printed** on all pages of anything you turn in.
- Write **legibly** and **coherently**. Manuscripts that are unintelligible in either content or handwriting are not likely to be looked on favorably.
- **Important:** Before you upload your homework, midterm and final papers to `canvas`, preview the PDF file on your laptop or desktop: if it's too faint to be easily read, you could end up with a bad score even if your answers are right, because the grader can't read your work. **Two hints:** (1) write up the answers you submit in black ink (not pencil), and (2) learn how to adjust the `contrast` setting (either in the app you're using or on the campus photocopiers) to make your pages dark enough to read easily.

General Content

Statistics is the **study of uncertainty**: how to measure it well, and how to make good choices in the face of it. **Uncertainty** is a state of incomplete or imperfect information about something of interest to you, for example

(a) the percentage p of the deer who lived on the UCSC campus as of Aug 2018 who have chronic wasting disease, or

(b) the pollution status of Monterey Bay in 2022 if a law regulating the dumping of refuse from ships into the Bay comes into effect in 2019, or

(c) the survival rate two years from diagnosis for patients with advanced liver cancer who take the relatively new drug *regorafenib*, a *multikinase inhibitor* whose use "may result in the [blocking] of cellular division/proliferation and the induction of apoptosis [death] in tumor cells" `[National Cancer Institute web site]`.

Statistics comes up mainly in two kinds of things people do:

- *Science* (acquiring 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* (for example, the percentage p mentioned in (a) above might be about 1.8%) and *relationships* (for instance, how the wing length of recently-born sage sparrows relates to their age). Statistics is helpful with both: coming up with *estimates* (intelligent guesses) and give-or-takes (measures of uncertainty) about facts (for example, on the basis of some data I have I might estimate p to be 1.8%, give or take 0.6%), and identifying which relationships are *causal* ("Smoking causes lung cancer and heart disease in humans") and which are just *associations* ("Drinking soda pop causes polio," or so they thought for awhile back in the 1930s; it does turn out that soft drink consumption and polio incidence were associated with each other, but as it happens neither was causing the other). 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 and choosing your favorite future; for example, policy-makers might need to choose between enacting

or not enacting the law regulating the dumping of refuse from ships into Monterey Bay mentioned in (b) above, and until they gathered some data and figured out how to analyze it they would be uncertain about the two possible futures {amount of pollution in the Bay in 2022 if the law *were not* enacted} and {amount of pollution if the law *were* enacted}. 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.

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 (*data*) to reduce your uncertainty; an example would be designing a *randomized controlled trial* to estimate the efficacy of the liver cancer drug regorafenib mentioned above. 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 (for example, a *scatterplot* of polio deaths against soft drink consumption). Statistics includes both **descriptive** methods to summarize *factuals* ("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 Jun through Aug 2018 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"). Along the way we'll talk about *sample size calculations* (methods for figuring out how much data you should gather in any given situation: it should make good intuitive sense that it's possible to have too little data, but surprisingly it's also possible to have *too much data*), and methods for quantifying the strength of the relationship between two variables (*correlation, regression*, the *analysis of variance*, and the *analysis of categorical data*).

Statistics uses math, mainly probability, but common sense, logic and good judgment are at least as important as math in most good statistical work. A long time ago (in the late 1700s) the great French mathematician Laplace put it best:

Statistics is common sense reduced to calculation.

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 (including a variety of examples from my drafts of the book for this class and other biological/environmental/health statistics textbooks, and also from journal articles in the biological, environmental and health sciences). 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'll trace the discovery process that led to the original development of the methods we study (I'll tell you a bit about the *history* of probability and statistics along the way). The idea is for some real learning to occur in class, not just note-taking.