

“Most real-life statistical problems have one or more nonstandard features. There are no routine statistical questions; only questionable statistical routines.” – D.R. Cox

“... much of what statistics teaches seems surprising, counterintuitive, and even unscientific.”
– R. Royall

“It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place.” – J. Tukey

E538—Statistics for Environmental Science—Spring 2003 (A Second Course in Data Analysis and Statistics)

School of Public and Environmental Affairs—Indiana University

Prerequisite: An introductory course in statistics¹

Web site: www.indiana.edu/~e538stat

Professor: David Parkhurst

Office: SPEA 355

Phone: 855-4556

Section 8828

9:30–10:45 a.m. TR, SPEA 277

Labs, Friday in SPEA 277:

§8829, 9:05–11:05 am; §8830, 1–3 pm

E-mail: parkhurs@indiana.edu

Office Hours: Tuesday 11:00–12:00; Wednesday 11:00–12:00, and by appointment other times.

You are welcome to stop in at other times if my office door is open—I'll be happy to talk with you if I'm free at the time.

Associate instructor: Denise Wessels (dwessels@indiana.edu)

Office location and office hours to be announced

Purpose of statistics

The primary purpose of statistics is to allow us to use data to improve our knowledge, and, in applied situations, to use that knowledge effectively in making decisions about how to act. As illustrated in the figure on the next page, those processes require knowledge both of the subject being studied, and of statistical and data-analysis methods. The latter are the focus of this course.

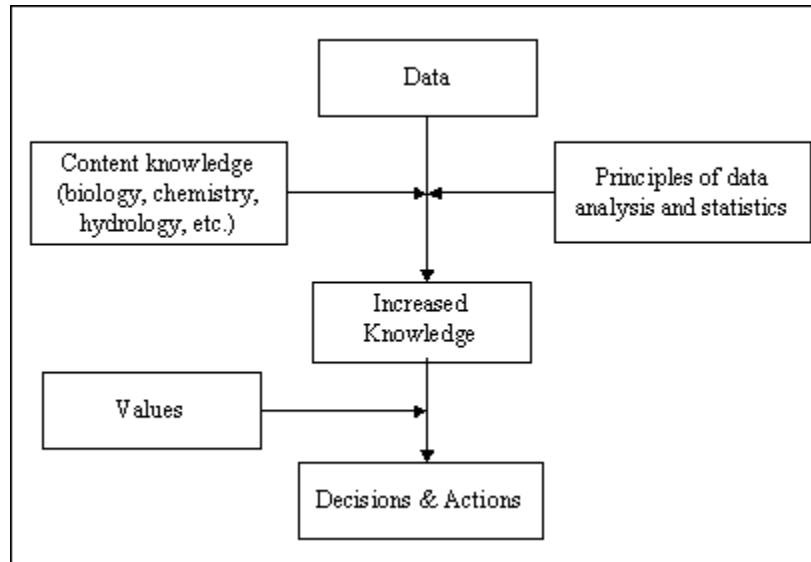
Course description:

The popular view of “statistics” seems to correspond with the title of the infamous book, *How to Lie with Statistics* (D. Huff. 1954. Penguin), but a worthy alternate goal for environmental scientists is to use statistical methods for help in learning what is true, and acting appropriately as a result. We need tools for analyzing data—with their associated variability and uncertainty—to improve our understanding (basic science) and to guide our management decisions and actions (applied science).

To analyze data effectively, environmental scientists need to know some basic statistical “recipes,” but also need to know when those recipes should or should not be used, and how to interpret the results they produce. Thus the concepts that underlie the recipes are important as well. Beyond this, statistics is currently undergoing several major revolutions that demonstrate both real problems with the conventional statistical methods that most scientists now use, and

¹ If you have not previously taken an introductory course in stats, please see Prof. Parkhurst right away.

relatively recent changes that overcome some of those problems. In addition to conventional statistics, therefore, we will consider three of these revolutions at least briefly, *i.e.*:



- Bayesian methods, which are based on so-called subjective or personal probabilities, and which often give answers much closer to what scientists really want to know, compared with most answers available from standard (frequentist) statistics,
- Elimination of, or greatly reduced emphasis on, null hypothesis significance testing, which has many logical problems and tends to hide adverse environmental effects, and
- Resampling methods, which reduce the need to make sometimes dubious assumptions about data.

We will also consider what these alternative views of data analysis teach us about the shortcomings of conventional methods.

Our overall approach should help you to be effective environmental scientists and managers, not only in the short run but also as your career develops in the longer term.

Course objectives:

The three major goals of this course are that you should:

1. Learn how to separate the "signal" (or information, or pattern) from the "noise" (or random variation) in data, and then to use the information obtained to increase your knowledge and when appropriate, to help your decision making.
2. Learn how to obtain data in efficient ways, to provide maximum information for the time and money spent.
3. Learn how to recognize illogical uses and interpretations of statistical methods by others, to aid in advancing scientific knowledge and for effective environmental management.

I will emphasize conceptual understanding with the hopes that in the future:

- You will know when standard statistical methods are being properly applied and when they are not.

- You can modify and extend standard (and non-standard) methods to fit new situations.
- You will know to consult a statistical expert for complex problems that go beyond your statistical knowledge.

For each major topic, I will try to follow a common pattern by discussing:

- What the concept is, and why it is useful.
- How to do any necessary calculations (sometimes by hand and sometimes with software).
- How and why the concept works (simple theory).
- Examples from environmental science, from the viewpoint of both basic and applied science.

Course format and resources

We deal with complex but important ideas in E538, and many students will have to invest considerable effort in the course. We will help you as much as we can through:

- Lectures. Most days, about an hour of the class will be in lecture mode. You should feel free to interrupt and ask questions if you “lose the thread.”
- In-class problems. I will frequently pose problems for you to work, preferably in groups of two or three. We’ll discuss these after you’ve had some time to attack them.
- Discussion. Occasionally, lecture time may be replaced with small group discussion, to help you grapple with important philosophical issues.
- Textbooks and notes. The Moore and McCabe book is an introductory but substantial text that is appropriate for a strong undergraduate or introductory graduate course in statistics. I have chosen it because it gives a solid basis for the more advanced consideration we give to certain topics that practicing environmental scientists ought to know. Parts of this book ought to be review for you, but some of the material will be new for most students. In either case, you may see a few exam problems based on material from the book that is not repeated in lecture. Advanced topics are covered in lectures and in the supplementary lecture notes that you should pick up at a local copy shop—listen for an announcement in lecture. One such topic is Bayesian statistical methods; the Iverson book provides a readable entry into that subject.
- Labs. These Friday sessions allow time, among other things, for sampling exercises in which each student helps generate part of a large dataset to illustrate some critical point. In addition, the lab periods encourage you to work together on problems. The teaching assistant will be there to provide examples and to help with the assigned problems.
- Exercises. You will be provided with sets of exercises each week to show how the principles under study apply to environmental science and management. Many of these problems are taken from past exams, so they give you practice solving the kinds of problems you can expect to find on this year’s exams.

Study hint: We will provide detailed answers for some problems, but we **strongly** recommend that you not look at the details until you have worked the problem yourself. You will learn a thousand times more (that’s a rough estimate) by working through a problem yourself than by looking at a prepared solution and convincing yourself that you understand it.

Remember that in your future work you will have to solve problems yourself, not just understand someone else's solutions, and the same will be true in exams.

If you help other students with a problem, you can do them the most good not by working it for them, but rather by either (a) working some similar problem for them, or (b) finding out where they are “stuck,” and asking leading questions that will get them on the right track. If you go to the effort to help others in these ways, you will likely learn a lot yourself in the process.

- Exams. As a student, I learned a great deal from the process of wrestling with ideas while taking exams, and I try to write questions that will give you that same benefit. This means that most questions will test your knowledge of underlying concepts, and not just your ability to “spit back” standard recipes. Questions will be similar in general form and difficulty to the assigned exercises and to those on past exams, but will not be identical to them with just the numbers changed.
- E-mail. Each day at the end of class, I will ask part or all of the class to write out a brief note about (a) what you found most interesting that day, (b) what you understood least that day, and (c) any specific questions you might have.. I try within a day or two to provide answers to questions, via e-mail that I send to everyone in the class. PLEASE READ THESE MESSAGES—I view the information they provide as supplements to lecture, and will assume you have that information when I write exam questions. Sometimes these messages may include announcements about changes in schedule, modified office hours, etc. as well.
- Office hours: Both Professor Parkhurst and the associate instructor hold regular office hours, and will be happy to help you when you can't work a problem or don't understand some concept. See Dr. Parkhurst after class or send e-mail to make appointments with him for other times. Please take advantage of our availability whenever there is something you don't understand!
- Homework: A few homework assignments spread through the semester will give you practice in analyzing data and interpreting statistical results.

Some or all of the homework will be assigned as group assignments, for two reasons. First, education research shows that most students can learn better by working with their peers than by just listening to a professor and then working alone. Second, agencies and firms that hire our students often tell us that because people generally work together in the “real world,” they encourage us to give students experience in working with others.

For homework assigned to groups, you must participate fairly in your group's work—if you do not, you will receive a zero for that assignment. (I will be asking each person to rate the participation of group members for each group assignment.)

Required homework must be handed in by the deadline to receive credit. Because computer systems and printers occasionally fail, you should plan to complete your work before the due date (a day early, for example), to allow for such problems.

- Statistical software: We will use a statistical software package known as R that is available at all of IU's windows-based computer clusters. Early in the semester, we'll meet in one of the clusters for a startup session—stay tuned for time and place. To install R (for free) on your own computer, follow the instructions I sent by email in the week before classes.

Grading:

Many exam and homework problems involve sequences of calculations and logical steps. When we find an error in your work, we try to count off once for that error, and then to grade the rest of the problem starting from that point. That is, we try not to let a small error at the start of the problem cause you to lose too many points overall.

After each exam, I will give you the conversion from scores to grades, so you know right away the contribution of that exam to your course grade. I decide on the conversions by the following sequence of reasoning. First, I treat each score as a percentage of the points available, and assign grades on a scale where 90% = lowest A-, 80% = lowest B-, 70% = lowest C-, etc. My exams can be challenging, and if that process leads to too many low grades, I will boost the distribution, usually with a conversion of the form $\text{Grade} = a + b \times \text{Score}$ where a and b are appropriate constants.

Texts:

- D.S. Moore & G.P. McCabe. 1998. *Introduction to the Practice of Statistics* (4th ed.). W.H. Freeman & Co.
- Class notes. To be purchased at a local copy shop.
- G.R. Iverson. 1984. *Bayesian Statistical Inference*. Sage Publications. Newbury Park, CA. I'll try to place a copy of this on reserve for those who don't want to purchase it.
- Some journal papers, to be placed on reserve in the library, will be required reading. Watch for announcements.
- You will need a pocket calculator with $\log x$, e^x , Σ , means, standard deviations, and regression functions. Calculators with several internal memories help you avoid writing down intermediate results. A single storage register is acceptable, however. Try to find the instruction manual for your calculator. If you can't locate one, some calculator companies make them available via the web—a search engine like Google may help you to locate the appropriate one.
- This may be useful for some: Gonick, L. and Smith, W. 1993. *The Cartoon Guide to Statistics*. New York, Harper Perennial. This is a well written book that might be helpful to you. You can probably find it in local bookstores—I have not ordered copies for the course, however.
- Please bring graph paper to every lab. For consistency, I recommend paper with one-cm squares, and lighter one-mm squares.

Prerequisites:

This course has a prerequisite of an introductory course in statistics, and that background **will be assumed**. It is possible you may have learned the necessary concepts in some science class, but if you haven't specifically had a stats course, please see me right away. You should also have a firm grasp of freshman level algebra and know the general ideas of both derivatives and integrals from calculus.

We will encourage you to perform (or sometimes to check) your homework calculations using statistical software like R (or perhaps other packages).

Grading:

The two mid-term exams and the final will contribute 21%, 23%, and 26%, respectively, to your course grade. The homework will contribute the remaining 30%. Exams will be cumulative, but each with greatest emphasis on the newest material. The rising weights are designed to allow you to learn from the first exam whether you are studying enough and in the right ways, and to recover later if need be.

Course ethics:

Many of my colleagues share with me the frustration that a few students seem increasingly not to know the difference between intellectual honesty and dishonesty. Cheating on tests is obvious enough, but issues involving out-of-class assignments (like the homework for this class) seem less clear. These difficulties are increased by my hope that you will work together—**to a point**—but that you (or your group, for group work) will turn in work that is your own. Computers add another source of ambiguity. The following discussion is intended to help clarify my expectations about the homework. Please ask if you want further guidance.

Here is a statement (modified from one provided by Prof. D. Willard) that I use in my undergraduate classes:

ACADEMIC MISCONDUCT. Academic dishonesty is not common, but **is** serious and intolerable. Please familiarize yourself with the Student Handbook guidelines. I assume that you all know and understand what plagiarism and cheating are; if you don't, find out. The rules are simple. Do your own work. Don't copy or even seem to copy from others. Allowing someone else to use your work as if it were their own is as serious as using someone else's work without full written acknowledgment in whatever you turn in. If you make legitimate use of work done by others, always document your sources.

Please be fully aware—If I determine that a student has cheated in this course², then that student will be dropped from the course and will receive a failing course grade. As required, I will also report such infractions to SPEA's Graduate Program Director and to the Dean of Students.

For E538 homework, I hope that you **will** work together in the following acceptable ways:

- Verbally discuss the purposes of an assignment. What is the point of an analysis like this? Why might we want to know the results being asked for? What assumptions are required for a particular kind of analysis? Why are these methods useful? In what ways is the exercise like something you have done in the past, or might have to do in the future? What would the results mean if they came out in various ways?
- You may also discuss the general ways of dealing with the types of calculation required for the exercise. If you know how to do the calculation, and want to help a classmate from a different group who hasn't figured it out yet, I would be pleased if you did that. Indeed, probably the best way to learn well how to do something is to show someone else how to do

² I don't expect to see this in a graduate course, but it does seem to happen on average about once a semester in medium to large undergraduate courses. I have occasionally had indications of cheating even in graduate classes, but that has been rare, and I hope it stays that way!

it— such cooperation should help both you and the person you are helping. However, do this by making up a similar set of data and showing your classmate how to analyze the made-up data. Then, **leave them to analyze the real homework exercise themselves**. Of course, for group assignments you may work with others in your group on the actual exercise.

- Similarly, if you are helping someone learn to perform statistical analyses (for example, in R), show them the general ideas, but do not leave a computer file in a state such that they can enter the homework data into a ready-made template that you have prepared. (Again, if working in an assigned group, you may exchange computer files **with others in your specific group**.)

Here are some examples of what is **NOT** permissible, for individual assignments or with others outside your group for group assignments:

- Copying from another student's written answers, even if you make substantial changes to the wording. Also, as noted above, it is just as unacceptable to provide your answers to another student as it is for them to use your answers.
- Analyzing the data in a computer file, then allowing someone else to use or modify that file (or a copy of it) in any way.
- Giving a copy of **any** computer file related to the homework assignment to any other student in the class. The one exception to this is that if two or more students are to analyze the same data, one person may type in the data and pass a file **containing the raw data only** to others.

If you are uncertain about what is fair and what is not, please err on the careful side, then ask for clarification later.

Policy on “Incompletes”:

The University policy on grades of “incomplete” includes the following statement:

CIRCUMSTANCES PERMITTING INCOMPLETES

The grade of Incomplete used on the final grade reports indicates that the work is satisfactory as of the end of the semester but has not been completed. The grade of Incomplete may be given **ONLY** when the completed portion of a student's work in the course is of passing quality. Instructors may award the grade of Incomplete upon a showing of such hardship to a student as would render it unjust to hold the student to the time limits previously fixed for the completion of his or her work.

After discussing this statement with my colleagues, I believe that the "hardship" referred to does not include poor preparation or planning, an overloaded schedule, or similar factors. Rather, it refers to substantial illness, family emergencies, and the like. Any incompletes granted in E538 will be based on this University policy.

Policy on exams:

You must take mid-term exams at the times they are scheduled for the lab in which you are enrolled. (I may, however, attempt to arrange with you for the exams for both sections to occur at a common time.) If that creates a conflict with some other course, you may ask to take the exam for the other lab section and I'll most likely give you permission to do so, depending on availability of space. Such requests will be considered only if made at least seven days prior to the exam time.

You must take the final exam at the University's published time. **It is your responsibility to arrange employment requirements, job interviews, airplane flights, and the like so they do not conflict with any of the scheduled exams.** The final exam is officially scheduled for Tuesday, May 6, 8–11 in the morning. However, I hope to work with you to find a three-hour slot some other time during finals week.

E538—Spring 2003—Tentative Schedule:

Week of	Topic	Reading³
Jan 14	Introduction; Distributions, display, and description of data	Preface, Introduction 1
Jan 21 Jan 23	R introduction, Student Bldg. 221 Relationships among two or more variables	2
Jan 28	Experimental Design	3
Jan 35 (Feb4)	Probability	4; Iv7–9
Feb 11	Probability, conditional probability	4.4, 4.5; Iv12–17
Feb 18	Binomial and Poisson distributions; Properties of sample means; Introduction to estimation	5.1, Notes 5.2, 6.1
Feb 21	First mid-term exam, in lab	
Feb 25	Decision making under uncertainty---I Decision trees, probabilities, and values	Notes
Mar 4	Decision making under uncertainty---II Randomization hypothesis tests Types I and II error, α , β , and power	Notes 6.2–6.4 Anderson, et al. (2000) ⁴
Mar 11	Decision making under uncertainty---III t tests	Notes 7.1, 7.2
Mar 18	Spring break!	
Mar 25	t tests, power, equivalence and reverse tests	Parkhurst (2001) ⁵
Apr 1	Inference about variability Bayesian inference for proportions	7.3 Iv9–11, Iv18–39
April 4	Second mid-term exam, in lab	
Apr 8	Bayesian inference for means	Iv70–77; (Review prev. Iv)
Apr 15	Transformations One-way anova	Notes 12; 13
Apr 22	Two-way anova Regression statistics	Notes 10
Apr 29	Logistic regression	15 (on M&M's CD)
May 6	Final exam—8–10 am, or TBA	

³ Most references are to chapters or sections of Moore & McCabe. Numbers preceded by “Iv” are page numbers in the Iverson book. “Notes” refer to readings from the class notes.

⁴ Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 64: 912--923. On reserve.

⁵ Parkhurst, D. F. 2001. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience* 51: 1051-1057. On reserve in SPEA library.