Lecture 1

Developing your common sense

Good statistics requires a combination of mathematical methods, common sense, and communication skills. This course will emphasize all three aspects with less emphasis on raw mathematical skills. Conducting standard analyses by hand in the age of modern computers is pointless. Many software packages including Excel, free specialized statistical packages such as R, and expensive specialized software packages such as SAS can conduct almost all elementary analyses and many non-elementary ones. However, the user must know which analyses are appropriate in any given situation, must interpret the output of the software appropriately, and must communication results fairly and accurately.

These lectures assume that you have had at least one elementary statistics course previously. In such a course, you would have covered some elementary probability theory, central limit theorem based hypothesis tests and confidence intervals involving a single mean, the difference in two means, a single proportion, and the difference in two proportions, and elementary statistical design. In these lectures, we will only discuss probability as it comes up for the statistics at hand. We will review elementary hypothesis tests and confidence intervals but will emphasize checking assumptions needed for the central limit theorem to be approximately valid and will discuss alternative methods that can be used if those assumptions are not met. First, however, we will review elementary statistical design issues which are closely linked to the development of statistical common sense and the proper interpretation of results.

Example 1 A real estate agency shows black clients housing in black neighborhoods and white clients housing in white neighborhoods. Does this prove that the real estate agency is biased? What would "prove" to you whether the agency was biased or not? ■

Example 2 Do scientific journals discriminate against female authors? Consider the following two research designs for studying this phenomenon: (1) Roughly equal numbers of male and female reviewers for five behavioral journals were chosen. Half the female reviewers and half the male reviewers received the same paper with the names of 2 fictional female authors; the other half of both groups received the same manuscript with the names of 2 fictional male authors. The decisions of each group were summarized and compared. (2) Twenty-four research journals in ecology and evolution were solicited to participate in a study. Seven agreed to participate. The editors of these journals recorded the sex of the first author and whether the publication was ultimately accepted or rejected. The outcomes for male and female first authors were compared.

Which design is better and why?

The first study found female reviewers were biased in favor of female authors. The second found no significant difference. However, after the journals instituted a double-blind review policy, the proportion of papers with female authors published increased. Do you have any comments on these results? ■

Example 3 It is routine for psychology courses to give extra credit to students who agree to participate in research studies conducted by students and faculty in psychology. Suppose a group of students from one such class agreed to participate in a self-esteem study. Half the class is assigned to a treatment group and half to a control group. It is determined that the treatment works and improves grades. Should the self-esteem treatment be used in all University classes? Why or why not?

If you do not believe the results necessarily generalize to all University students, what are the implications for the results from the Framingham Heart Study, where the initial participants were all from the town of Framinghan in Massachusetts? One of the results using the Framingham participants is that low bone mineral density in women is associated with future dementia. What are the implications for the results from the Nurses' Health Study that consuming margarine is associated with an increased risk of heart disease? All initial participants were, you guessed it, nurses. What are the implications for the results from the Physicians' Health Study that taking aspirin daily reduces the risk of a first heart attack? Do you believe these results? Why or why not?

Most published research findings, at least in epidemiology, are false. The use of a sample that is not representative of the population of interest is only one of the reasons for this phenomenon. Published research findings often echo the current prejudices of science and society and are often the result of data fishing (also called data mining) or, similarly, the result of imprecise initial hypotheses and experimental design. Later in this course we will discuss many methods of correcting for multiple test which may be of some use in minimizing the likelihood of publishing false findings. ■

Example 4 A major hospital conducts a randomized, double-blind clinical trial to determine whether the injection of a certain peptide by diabetic patients improves their blood sugar control. Would you believe the results from such a study?

Randomized, double-blind clinical trials are about the best experimental design possible. However, here's a story about such a design. My husband is a Type I diabetic and was in conversation with a participant in a study as described above. The participant wanted to break the blind. My husband suggested the following: take the stuff you are injecting and take a urine-protein test strip and see if the stuff you are injecting is a protein. If it is, then you can't be sure whether it is the specific peptide being studied or not. But if it is not a protein, then you can be sure you are in the control group. The participant followed his advice and found that he was in the control group. Whether this method actually works or not is not even relevant; the patient now believes he is in the control group which causes a kind of anti-placebo effect. Randomized, double-blind clinical trials are not necessarily all they are cracked up to be.

Another story involving my husband involves his own participation in a clinical trial at Washington University in St. Louis. The nurse told him to manipulate his C-peptide levels by eating candy shortly before taking the tests needed to enter the study. That sort of manipulation is unlikely to be reported in any journal article relaying the results but may very well affect the results of the trial. ■

Example 5 Experiments where one can randomize treatments and record continuous outcomes provide one type of statistics. Another situation is polling, where one selects a random sample of individuals and records responses to questions, often yes/no questions. The reported results from most political polls seem to fall in this category, but, actually, most polls use a more sophisticated sampling design than simple random sampling.

All polls, regardless of sampling design, are subject to all kinds of problems. One is suppose to be able to reach the random target sample and get their honest answers to questions. But pollsters are restricted in the hours they can call you. How does this affect the people they can reach? How does calling on land-line telephones affect the people pollsters can reach? In the Dewey vs. Truman election the effect was that the pollsters could not reach people in poorer communities. The effect is different today given the wide use of cell phones.

Would you answer questions about your salary, your sex life, and your drug use honestly? Delicate questions such as these are often left to the end of the survey so that the pollster can develop a relationship with the person he is questioning. There are other methods for getting honest responses to delicate questions such as providing a biased spinner. The result is similar to saying if you get a 1 or a 2 on a die roll answer the sensitive question but if you get a 3 or 4 or 5 or 6 on the die roll answer an innocuous question for which we know the results for the population (for instance, do you own your own home?) We know the proportion of yes's expected to this question. The point is that the pollster is NOT told the result of the spinner/die so he doesn't know which question you are answering when you say "yes." ■

The kind of analysis you conduct on data depends on its nature. The first and major part of this course is devoted to analyzing continuous response data. The results from polls tend to be binary (yes/no) response data. The second part of this course will be devoted to models to analyze that kind of data.

References and Readings

[1] Anonymous. Doctors confirm benefits of aspirin. *New York Times*, page B6, July 20 1989.

[2] Tamsin L. Braisher, Matthew R.E. Symonds, and Neil J. Gemmell. Publication success in *Nature* and *Science* is not gender dependent. *BioEssays*, 27:858–859, 2005.

[3] Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julie Koricheva, Rooda Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution*, 23:4–6, 2008.

[4] Marian Burros. Study links heart disease to margarine. *New York Times*, page A10, March 5 1993.

[5] Editorial. Working double-blind: Should there be author anonymity in peer review? *Nature*, 451:605–606, February 2008.

[6] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:e124, 2005.

[7] Margaret E. Lloyd. Gender factors in review recomendations for manuscript publication. *Journal of Applied Behavior Analysis*, 23:539–543, 1990.

[8] John Rice. *Mathematical Statistics and Data Analysis*. Brooks/Cole, 3^{rd} edition, 2006.

[9] Beena Sood, Virginia Delaney-Black, Chandice Covington, Beth Nordstrom-Klee, Joel Ager, Thomas Templin, James Janisse, Susan Martier, and Robert J. Sokol. Prenatal alcohol exposure and childhood behavior at age 6 to 7 years: I. Dose-response effect. *Pediatrics*, 108:e34, 2001.

[10] Tom Tregenza. Gender bias in the refereeing process? *Trends in Ecology and Evolution*, 17:349–350, 2002.

[11] Michael Winerip. Lines that divide towns and the races. *New York Times*, page E7, Dec 22 1985.

Exercises for Lecture 1

1. Recent studies suggest that alcohol consumption during pregnancy can lead to long-term behavioral problems in the resulting children. See, for instance, [9]. Are such studies based on experiments or retrospective observations and why might one be skeptical of the conclusions?

2. An article in BioEssays cited the following statistics as evidence that the journals *Science* and *Nature* do not discriminate against authors on the basis of sex: out of 136 female researchers and 305 male researchers in Britain and Australia over the period 1999-2004, the average number of publications men had in the two journals was 0.64, the average number the women has was 0.33 and this was a significant difference. However, the difference might be due to the fact that women published fewer articles on average over this period (9.1 vs. 12.4). The mean percentage of output for the two sexes published in the two journals was not significantly different.

   Assess the experimental design these statistics are based on.