

**Guest Editorial in the
Journal of Undergraduate Math and Its Applications:**

A New Approach to Introductory Statistics

Robert L. Wardrop
Department of Statistics
University of Wisconsin–Madison
wardrop@stat.wisc.edu

Introduction

The standard organization of topics in introductory statistics texts begins with descriptive statistics, follows with probability theory (including distributions and sampling theory), and finally proceeds to inference—but inference only for random samples from populations.

To be sure, texts differ in the emphasis given to each of these “course segments,” and some authors are better statisticians or writers than others, but few books deviate far from the norm. George Cobb writes [1993]:

... [I]f one could superimpose maps of the routes taken by all elementary books, the resulting picture would look much like a time-lapse night photograph of car taillights all moving along the same busy highway.

This editorial provides a brief description of my very different approach to introductory statistics.

What Students Know

Students enter a statistics course familiar with several methods of learning. For example, one can *memorize* a poem or a speech; a physical skill or proper pronunciation of a foreign language can be learned by *repetition*; or something can be learned by going to a proper source and *looking it up*. In addition, students know that mathematicians learn by applying the rules of logic to a set of assumptions or facts.

From Frustration, a New Goal

Students enter the class knowing that scientists learn about the natural, social, and political worlds by experimentation and observation; but they are woefully inexperienced at, and very uncomfortable with, doing science themselves. Few students view themselves as scientists, even in a broad sense, and all are unclear about or unaware of the importance of statistical principles and methods in the work of scientists. I became disillusioned with the standard approach to introductory statistics when I realized that after finishing my course, even my “A” students had nary a clue about the usefulness of statistics in science.

As a result of my frustration with the standard approach, I decided to create a course that would have as its fundamental goal:

To enable students to discover that statistics can be an important tool in daily life.

This goal would be achieved by having the students realize that daily life can be enriched by being a scientist, and that a good understanding of statistics can make a person a better scientist. The standard approach, which teaches students a collection of methods, and hopes they can figure out how to use them, did not work in my classes. I decided that I would focus on making students do science, and in the process of doing science they

would learn the usefulness of statistics. I will present the flavor of this approach by describing the activities of three of my students during the first three weeks of my course.

The First Three Weeks

The course begins with several examples of comparative studies. A comparative study requires a protocol for obtaining information—called an experimental design, and methods for learning from the information obtained. The experimental design of a comparative study has four components: subjects, response, treatments, and the method of assigning subjects to treatments. If subjects are assigned to treatments by randomization, the design is called a completely randomized design (CRD). The first three weeks of the course restrict attention to CRDs with two treatments and a dichotomous response. No assumptions are made about how the subjects are selected for inclusion in the study. It is helpful to note that there are two broad classes of subjects: distinct individuals and trials.

Example of a Student Project

Examples in my course are drawn from the published scientific literature, of course, but also include over 80 small projects performed by my former students. These projects put into action the material covered in the course. The work of Therese Nyswonger provides an example of a good project. Therese performed a CRD related to marital infidelity. More precisely, she wanted to know whether the gender of the cheater would influence a person's decision to tell the wronged spouse. Therese's subjects were 20 female co-workers, and her treatments were two versions of a question. The first version of the question was,

You are friends with a married couple. You are equally fond of the man and the woman. You discover that the husband is having an affair. The wife suspects that something is going on. She asks you if you know anything about her husband's having an affair. Do you tell?

The second version reversed the roles of the husband and wife, but otherwise was identical to the first version. The response was the subject's answer—yes or no.

Therese found that of the ten subjects who read the cheating-husband version, seven said they would tell the wife, and of the ten subjects who read the cheating-wife version, only four said they would tell the husband. Clearly, the version read did not *determine* the response, but there is evidence that the version read *influenced* the response. The meaning of the evidence, however, is not clear.

The Skeptic's Argument

After performing her study, and examining and summarizing her data, Therese turned to the consideration of what she had learned. (To paraphrase Professor Jessica Utts of the University of California at Davis, author of the excellent new book, *Seeing Through Statistics* [1995], Therese needed to decide whether the results of her study were meaningful enough to encourage her to change her lifestyle, attitudes, or beliefs.)

I emphasize to the class the importance of making any conclusion as precise as possible. To this end, I introduce the *Skeptic's Argument*, which states that the pattern in Therese's data is completely due to chance, or, in other words, that the version read is totally irrelevant. It is merely the case, the Skeptic claims, that 11 of Therese's friends were "tellers," and nine were not. Confronted with either version of Therese's question, the tellers would say that they would tell and the others would say that they would not tell. (Therese believed that it would be a mistake to give each subject both versions because the knowledge gained by reading both versions could affect the responses.)

Thus, according to the Skeptic's Argument, the pattern in Therese's data (70 percent versus 40 percent telling), is simply the result of the chance assignment of seven tellers to the cheating-husband version, and four tellers to the cheating-wife version. Debating the Skeptic is the Advocate, who acknowledges that the Skeptic could be correct, but argues that the pattern is unlikely to be due to chance. Instead, claims the Advocate, the pattern in the data reflects a preference *among Therese's subjects* to tell on a husband rather than a wife.

Quantifying the Debate

At this point in the class, I ask students which argument they find more convincing for Therese's data—the Skeptic's or the Advocate's. Typically, students give approximately the same support to each argument.

Hypothesis testing, I announce, is a technique that allows a researcher to *quantify* the debate between the Skeptic and the Advocate. Probability is defined as a measure of uncertainty, with special emphasis on probability induced by the chance mechanism of randomization. In particular, all possible assignments of subjects to treatments are equally likely. Probability theory is a huge subject, but attention is restricted to those few and minor aspects of probability that were needed by Therese to analyze her data. Her null hypothesis stated that the Skeptic is correct, and the P-value measures the (relative) strength of the evidence in support of the alternative hypothesis. For Therese's project the P-value is too large to discard the Skeptic's argument. The difference between 70% and 40% is substantial, but it is obtained from too few subjects to be convincing.

Of course, an exact P-value can be computationally inaccessible or tedious. This difficulty is addressed by presenting two methods for obtaining an approximate P-value: simulation and the standard normal curve. First, a computer simulation experiment is presented, and it is shown that 10,000 runs yield an excellent approximation to the sampling distribution of the test statistic. This is a pedagogically important topic because

- A simulation experiment is very intuitive; no fancy mathematical arguments are needed to convince a student that looking at *some* assignments of subjects to treatments is a reasonable approximation to looking at *all* assignments of subjects to treatments.
- Simulation is a powerful tool for the consideration of robustness and power later in the course.

I show students that if the probability histograms of the sampling distribution of the test statistic are drawn for each of several studies, the most striking difference between these pictures is in their amount of spread. If these test statistics are standardized, the resulting probability histograms are similar to each other and to the standard normal curve. Thus, a P-value can be approximated by using the standard normal curve. This approach has two noteworthy features:

- The standard normal curve is introduced as an approximation device without any reference to the abstract notion of a continuous random variable.
- The student learns that the standard deviation is a reasonable way to measure spread because it works: standardizing by dividing by the standard deviation yields probability histograms that are similar.

Another Project Rejects the Skeptic's Argument

Therese's subjects were distinct individuals—people. Two other students, Teresa Chervenka and Lan Nguyen, performed a study in which each subject was a trial—a shot of an arrow by Teresa's boyfriend John. A trial was a success if John hit a 4-inch diameter bull's-eye from 20 yards, and was a failure otherwise. The treatments were using wet plastic vanes or wet feather vanes on the arrow. (According to Teresa and Lan, most archers agree that feather vanes fly better than plastic vanes in dry weather.) The study consisted of 100 trials, with 50 trials assigned to each treatment by randomization. John obtained 48 successes with the plastic vanes, but only 36 successes with the feather vanes.

The P-value computed by Teresa and Lan led them to conclude that the pattern in their data was too strong to reasonably be attributed to chance—they rejected the argument of the Skeptic.

What Students Learn in the First Three Weeks

Let us review what has been accomplished during the first three weeks of the course. The students have been introduced to a large number of important statistical principles and methods, all in the context of a scientific investigation. The students have been able to work through a complete statistical analysis—the selection of a scientific issue for study; framing that issue, if possible, as a CRD with two treatments and a dichotomous response; collecting, presenting, summarizing and interpreting the data; and performing a hypothesis test to further understand the data's message.

Later in the Course

This new approach naturally stresses the limitations of the conclusions of a study. Therese learned about her 20 co-workers; whether her findings reflect the attitudes of the general population is not known. John learned that the difference in his performance using wet plastic vanes versus wet feather vanes was too strong to reasonably be attributed to chance *on the occasion of his experiment*. The questions of whether the pattern of the study would persist on another occasion, or if John were hunting, or for a different archer remain unanswered. It has been my experience that the standard approach to introductory statistics encourages students to make unwarranted generalizations because it teaches inference only for random samples.

The first three weeks of my course represent a dramatic departure from the standard approach to introductory statistics. Later in my course, the material becomes more familiar: inference for one or more populations is presented, techniques learned for dichotomous responses motivate methods for numerical responses, and regression is studied. The emphasis on statistics as a tool for scientists, however, leads to important—though perhaps not dramatic—changes from the standard approach, both in presentation and emphasis. Space limitations prevent a cataloging of these changes, but one deserves mentioning:

By learning inference based on randomization before inference for random samples, students are better equipped to understand the limitations in the interpretation of observational studies, including such counterintuitive results as Simpson's Paradox.

A Different Ordering of Topics

Perhaps the most controversial aspect of my approach is the decision to have a somewhat complete presentation of statistical methods *for a dichotomous response* before any consideration of numerical responses. Below are four reasons I prefer this ordering.

1. I definitely would not change the first three weeks of my course. I want to present a complete statistical analysis of a scientific issue as early in the course as possible. If numerical responses are studied at the same time as, or instead of, dichotomous responses, much more time will be required to discuss data presentation and summary. As a result, the complete analysis would be delayed substantially.
2. In my course, the student's first exposure to important statistical ideas is for the computationally and conceptually simple dichotomous response. For example, there is no question of how to summarize dichotomous data—one uses the proportion of successes—but for numerical data the issue of summary is much more complicated. The simplicity of a dichotomous response seems to make it easier for students to understand the new statistical ideas.
3. For a dichotomous response, a population is a single number, the proportion of successes (or probability of a success for Bernoulli trials). For a numerical response, however, a population is a picture (of a probability distribution). Thus, while it is conceptually easy to compare dichotomous populations—you are just comparing two numbers—it is unclear how two numerical response populations should be compared. Related to this, robustness is a nonissue for a study of a dichotomous population or populations, but considerations of robustness are a critically important component of any intellectually complete analysis of a numerical population or populations.

It is my experience that students are better equipped to deal with the complications of a numerical response *after* a thorough exposure to a dichotomous response. A student who already understands confidence intervals and hypothesis testing can better deal with issues of robustness and the selection of methods (e.g. studying means or medians).

4. I am a strong believer in the benefits of repetition in learning. In my course, students receive an early exposure to inference, and the ideas are repeated throughout the course. While the students may not have a good understanding of hypothesis testing after their initial exposure in Week 2, by the end of the semester nearly all of them have a solid understanding of it.

What about All Those Projects?

Finally, you may be thinking, “I am not sure what I think of this new approach, except that I do not have enough time to grade student projects!” Most of the persons who have taught with this approach (see below) have not assigned student projects. Their experience has coincided with mine: the use of student projects in the text is very attractive to students. I have two possible explanations for this phenomenon:

- Students know better than professors—well, this professor anyway!—what interests them. Hence, they find projects performed by other students to be more interesting than many of the examples selected by the instructor.
- Students seem to experience a general sense of validation at seeing the work of their peers displayed.

Acknowledgments

At the University of Wisconsin–Madison, Michael Kosorok, Robert B. Miller, Gary Schroeder, and Xiaodong Zheng (now at Utah State University) have used the approach described in this editorial in their introductory courses. Ronald L. Wasserstein of Washburn University and David Madigan of the University of Washington–Seattle also have taught courses following this new approach. I am indebted to these persons for their helpful comments and suggestions.

I am indebted to the editor of UMAP for his many helpful suggestions that improved this editorial.

References

- Cobb, George W., Reconsidering Statistics Education: Report of a National Science Foundation Conference, *Journal of Statistics Education*, 1 (1) (July 1993): para. 53.
- Utts, Jessica. 1995. *Seeing Through Statistics*, Belmont, CA: Duxbury.
- Wardrop, Robert L. 1995. *Statistics: Learning in the Presence of Variation*. Dubuque, IA: Wm. C. Brown.

About the Author

Robert L. Wardrop has a B.S. in mathematics from Oakland University in Rochester, Michigan, and a Ph.D. in Statistics from the University of Michigan. Since 1974, he has been a member of the faculty of the Department of Statistics at the University of Wisconsin—Madison and currently holds the rank of Associate Professor. He has embodied his philosophy of teaching statistics in the new textbook *Statistics: Learning in the Presence of Variation*, published by William C. Brown.