Computationally intensive methods warrant reconsideration of pedagogy in statistics

GORDON BEAR

Ramapo College, Mahwah, New Jersey

Computationally intensive methods of statistical inference do not fit the current canon of pedagogy in statistics. To accommodate these methods and the logic underlying them, I propose seven pedagogical principles: (1) Define inferential statistics as techniques for reckoning with chance. (2) Distinguish three types of research: sample surveys, in which statistics affords generalization from the cases studied; experiments, in which statistics detects systematic differences among the batches of data obtained in the several conditions; and correlational studies, in which statistics detects systematic associations between variables. (3) Teach random-sampling theory in the context of sample surveys, augmenting the conventional treatment with bootstrapping. Regarding experimentation, (4) note that random assignment fosters internal but not external validity, (5) explain the general logic for testing a null model, and (6) teach randomization tests as well as t , F , and χ^2 . (7) Regarding correlational studies, acknowledge the problems of applying inferential statistics in the absence of deliberately introduced randomness.

The textbooks from which students of psychology learn statistics have become quite uniform in the topics that they cover, the order in which they present those topics, and the way in which they conceptualize inferential techniques. Prominent in the canon are three doctrines: (1) The observations to be analyzed derive from a population through random sampling; (2) the goal of statistical inference is to determine the parameters of populations; and (3) the methods of choice for statistical inference rely on the t , F , and χ^2 distributions. Alternative methods are presented, to be sure, but still as devices for generalizing from random samples to their parent populations.

These three doctrines, which date from early in this century, fail to accommodate inferential methods recently made practicable by high-speed computing machinery. To incorporate modern methods—and the reasoning underlying them—I propose seven pedagogical principles for the introductory course in applied statistics for students of psychology.

1. Define Inferential Statistics as Techniques for Reckoning With Chance

In distinguishing between descriptive and inferential techniques, the latter should be defined as devices that account for chance in drawing conclusions from observations. Some texts conceive inferential statistics too narrowly—solely in terms of sampling, as devices for

generalizing from a sample to a population—but this conception does not accommodate randomization tests such as Fisher's Exact Test and others described below.

2. Clarify the Functions of Statistics by Distinguishing Types of Research

Exactly what descriptive and inferential techniques do depends on the kind of research in which they are employed. I urge that statistics students learn to distinguish three kinds of research: sample surveys, experiments, and correlational studies (cf. Bohling, Forsyth, & May, 1990). Table 1 shows the identifying features of each and the purposes for which they employ statistics.

3. Teach Random-Sampling Theory and the Confidence Interval in the Context of Sample Surveys

After basic techniques for describing a collection of observations (e.g., measures of central tendency and measures of variability) are introduced, the sample survey and the confidence interval should be explained. These comprise the methodology and the inferential device most familiar to beginning students and easiest for them to understand.

With appropriate definitions and examples, the teacher should convey the following points: (1) A sample survey is a method of investigating a population of cases; the goal is a conclusion about the population (e.g., a guess about the percentage of all the cases that possess a certain attribute or about their mean score on some variable). (2) The method requires selecting a sample of cases, making observations on the sample, and generalizing from the sample to the population. (3) If and only if the cases were selected by means of probability sampling (one variety of which is random sampling), one

This article is based on a paper presented at the November, 1994 meeting of the Society for Computing in Psychology, St. Louis. For inspiration and edification, I am grateful to Richard May, Michael Hunter, and Michael Masson of the University of Victoria. Correspondence concerning this article should be addressed to G. Bear, School of Science, Ramapo College, Mahwah, NJ 07430-1680 (e-mail: gbear@ultrix.ramapo.edu).

Table 1
Characteristics of Three Methods of Research

Method	Characteristics		
	Overall Purpose of Statistics	Function of Descriptive Statistics	Function of Inferential Statistics
Sample Survey Infer parameters of a population from study of a sample	Generalization from sample to population	Characterize the sample	Estimate the parameter(s) of the population
Experiment Test effect of independent variable(s) on dependent variable(s) by varying the former and measuring the latter	Comparison of the batches of data obtained in the several conditions	Characterize the difference among the several batches of data	Decide whether the difference is systematic (find a pattern)
Correlational Study Discover natural association between two or more variables by measuring them in a batch of cases	Assessment of the association between the variables	Characterize the association across the cases on hand	Decide whether the association is systematic (find a pattern)

can generalize by means of a mathematical model that yields a confidence interval for the unknown parameter.

The combination of probability sampling and confidence interval affords a generalization (an interval estimate) with three virtues: (1) It is objective (given the same data, anyone with sufficient knowledge will arrive at the same estimate); (2) it is precise (stated in terms of numbers); and (3) it warrants a known degree of confidence (which the researcher chooses—usually 95%). In the absence of probability sampling, one can, of course, still calculate a confidence interval, but one cannot know to what population, if any, the interval applies or what degree of confidence it deserves. More generally, one can use statistics to generalize beyond the cases on hand only if those cases were chosen by means of probability sampling (a point that psychologists and their students often fail to appreciate, as May & Hunter [1988] have demonstrated).

An even more general point worth making here is that a mathematical model (such as that of simple independent random sampling) applies to the real world most clearly when people make the world fit the model (by truly choosing a sample at random).

In learning the theory underlying interval estimation, students encounter important concepts: sampling variation, sampling error, random-sampling distributions, standard error, the normal-curve model, and the central limit theorem. They achieve closure in this part of the course when they comprehend the conduct of a sample survey and the calculation of an interval estimate.

An elementary understanding of probability will suffice for students here, and there is no need to burden them yet with the logic of hypothesis testing. If the instructor considers only the large samples that are common in survey research (e.g., in public-opinion polling), the complexities of the *t* distribution may also be postponed. Students should not, however, be shielded from Micceri's (1989) discovery that many scores of interest in psychology and education are not distributed normally.

The arithmetic required for interval estimation is simple enough for ordinary calculators. If a computer is available, appropriate software will permit helpful demonstrations of random-sampling distributions.

Also feasible with a computer—indeed, virtually impossible without one—is bootstrapping (Efron & Tibshirani, 1993; Simon, 1992), a method for empirically approximating the random-sampling distribution of any statistic. Given a statistic calculated from a sample, the method hopes to capture its random-sampling distribution by repeatedly sampling with replacement from the sample and calculating the statistic for each pseudosample so obtained. (In effect, the method treats the actual sample as a population.) Whether or not students have access to the technology of bootstrapping, they should know of the method and appreciate the excitement and controversy that it has sparked.

4. Regarding Experimentation, Note that Random Assignment Fosters Internal, but not External, Validity

The professor should introduce experimentation with words and examples to this effect:

The goal of an experiment is a conclusion about the effect of one variable on another. Probability sampling is not required, and is, in fact, rare, but random assignment is. However they obtain their subjects, experimenters (in a simple between-subjects design) manipulate the first variable, assigning subjects at random to one variation of it or another, and then measure the second variable, thereby obtaining some number of observations under each variation of the independent variable.

Here, the student should be apprised of the purpose of random assignment (also called randomization) and of the distinction between internal and external validity: An experiment possesses internal validity if it permits a valid conclusion about whether the independent variable affected the dependent variable in those subjects and

under those circumstances actually studied; it possesses external validity if the conclusion about the effect of the independent variable on the dependent variable can be generalized beyond the subjects and circumstances of the experiment (Campbell & Stanley, 1963; for elaborations of this basic distinction, see Pedhazur & Schmelkin, 1991). Internal validity is the primary concern in experimentation (Berkowitz & Donnerstein, 1982; Mook, 1983), and it is fostered by random assignment, but random assignment cannot create external validity—an other principle that psychologists and their students often fail to appreciate (May & Hunter, 1988). To honor the distinction between internal and external validity, and to acknowledge that the subjects of an experiment were acquired through a process other than probability sampling, I suggest, in this context, replacing the word *sample* with *batch*, the term employed in exploratory data analysis (Hoaglin, Mosteller, & Tukey, 1983; Velleman & Hoaglin, 1981).

5. Regarding Experimentation, Teach the General Logic for Testing a Null Model

Before introducing any particular statistical test appropriate for experimental data, the professor should foster insight into the purpose of statistics here. Simple descriptive techniques suffice to characterize the variation among the batches. For example, one calculates the difference between the mean of the scores obtained in the experimental condition and the mean of those obtained in the control condition. Given such a finding, which is an index of the variation among the batches, students should learn to ask a critical question: Couldn't it have arisen through a process of pure chance? They should also welcome inferential techniques as devices for answering this question.

To understand these techniques, students must know that they work by assessing the plausibility of a null model, which is a portrait of how chance could have been the sole reason for the variation among the batches (Hunter & May, 1995). Any null model implies a null distribution, which is the probability distribution of the possible findings of the experiment predicated on the assumption that the model is true. This is the distribution that one divides into a region of rejection and a region of retention, and it is this distribution to which one refers the actual finding of the experiment. (The procedures that researchers take to be tests of a null hypothesis are really tests of the entire null model in which it is embedded.)

To enhance understanding, the professor can explain that hypothesis testing in statistics is just a variant of the procedure by which one tests any kind of hypothesis about any kind of unobservable phenomenon (Popper, 1935/1959). To build on students' prior knowledge, a lesson should show how hypothesis testing resembles the procedure by which a physician tests a diagnosis.

6. Regarding Experimentation, Teach Randomization Tests as well as Methods that Assume Random Sampling

In the canon of statistical pedagogy, the null model for an experiment supposes that each batch of observations is a sample selected at random from a population, but when the subjects were not acquired through such sampling, these populations are not easily identified (Oakes, 1986, pp. 4–5, 15, 120, 122, 154–157). Some textbooks completely evade this issue, even failing to cite Fisher's (1925) classic solution to the problem, which imagines the data that would result from replicating the experiment an infinite number of times and supposes them to be normally distributed. It is Fisher's controversial conceptualization that underlies the t , F , and χ^2 tests that are currently standard in psychology.

Before learning the standard methods and the debates that they have sparked, students might be taught a simpler, more straightforward method for determining whether a difference observed in an experiment is systematic or plausibly ascribed to chance: randomization testing. This class of techniques has an honorable history dating back to Fisher himself (1935; see also Edgington, 1987; Pitman, 1937a, 1937b), but Fisher's Exact Test for the 2×2 contingency table is the only such device familiar to most psychologists. Mathematical statisticians and research scientists alike are currently devoting extensive attention to the development of techniques for randomization testing (e.g., Hooper, 1989; Hunter & May, 1993; May & Hunter, 1993; Romano, 1989; van den Brink & van den Brink, 1989; Welch, 1990).

Randomization testing affords a direct answer to the critical question about the finding of a randomized experiment: Was the observed difference between the batches simply induced by the randomization of the subjects—in which case it was present from the beginning—or did it arise during the course of the experiment as an effect of the variation in the conditions?

Applied to the typical between-subjects design, a randomization test assesses the plausibility of a null model supposing that (1) each observation was a fixed property of the subject whom it characterizes, and (2) any difference among the batches derived solely from the randomization and was thus not systematic. This claim is tantamount to the assertion that the observations were independent of the conditions of the experiment, and it is that assertion that becomes the null hypothesis (May, Masson, & Hunter, 1990). The alternative hypothesis claims that the observations did depend on the condition, which is tantamount to the assertion that the condition of the experiment influenced the observed variable. To assess the plausibility of this null model, given how the experiment turned out, one must calculate the outcome measure (e.g., the difference between the means in a two-condition experiment) for each possible way of randomizing the observations, thereby obtaining the null distribution.

Randomization tests are easy to teach and easy to grasp, but except for the tiniest of batches and the simplest of designs, they require calculation so extensive as to be impractical without a computer. Software is now available that makes them feasible and accessible for beginning students. Choices include *NPSTAT* (May, Hunter, & Masson, 1993; May, Masson, & Hunter, 1989); *Resampling Stats*, which also does bootstrapping (Bruce, 1993); and *StatXact* (1995).

In learning the standard tests based on the t , F , and χ^2 distributions, students should understand that these methods rest on a null model that is controversial and that easily engenders confusion: The tests look like devices for generalizing beyond the observations, but mathematical generalization is impossible in the absence of probability sampling, and in fact the methods function as devices for pattern finding. A rationale for employing them in the absence of probability sampling is that they often yield p -values approximating those derived from randomization methods.

7. Regarding Correlational Research, Acknowledge the Problems of Applying Inferential Statistics in the Absence of Deliberately Introduced Randomness

Having progressed this far in their introduction to statistics, students can appreciate a vexing issue in applying inferential statistics to correlational research: What is gained by testing null models of chance? A correlational study requires neither probability sampling nor manipulation with randomization, merely the acquisition of subjects and the measurement of two or more variables, and simple techniques of description such as Pearson's r may suffice to detect and summarize the associations between the variables across the subjects on hand. If chance was not deliberately introduced in a known way—by means of random sampling or random assignment—it is unclear what is gained by applying inferential devices for assessing the role of chance (Oakes, 1986, Section 7.1).

The issue is especially interesting because in the underlying mathematics, techniques for detecting systematic differences in sample surveys and randomized experiments are identical to techniques for detecting systematic associations in correlational studies. The formulas do not know how the data were collected.

Students should be warned that although regression techniques look like devices for generalizing from the data on hand, they cannot be used for that purpose absent probability sampling. Students should also know that as alternatives to the conventional methods, randomization tests exploiting the power of the computer are under development.

REFERENCES

BERKOWITZ, L., & DONNERSTEIN, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory research. *American Psychologist*, 37, 245-257.

- BOHLING, P. H., FORSYTH, G. A., & MAY, R. B. (1990). Research methodology taxonomy and interpreting research studies. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry, & M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 38-43). Washington: American Psychological Association.
- BRUCE, P. C. (1993). *Resampling Stats: The "new statistics"* [Software manual]. Arlington, VA: Resampling Stats.
- CAMPBELL, D. T., & STANLEY, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- EDGINGTON, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.
- EFRON, B., & TIBSHIRANI, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- FISHER, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.
- FISHER, R. A. (1935). *The design of experiments*. Edinburgh, U.K.: Oliver and Boyd.
- HOAGLIN, D. C., MOSTELLER, F., & TUKEY, J. W. (Eds.) (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- HOOPER, P. M. (1989). Experimental randomization and the validity of normal-theory inference. *Journal of the American Statistical Association*, 84, 576-586.
- HUNTER, M. A., & MAY, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology*, 34, 384-389.
- HUNTER, M. A., & MAY, R. B. (1995). *Statistical testing and null distributions*. Manuscript submitted for publication.
- MAY, R. B., & HUNTER, M. A. (1988). Interpreting students' interpretations of research. *Teaching of Psychology*, 15, 156-158.
- MAY, R. B., & HUNTER, M. A. (1993). Some advantages of permutation tests. *Canadian Psychology*, 34, 401-406.
- MAY, R. B., HUNTER, M. [A.], & MASSON, M. (1993). *NPSTAT 3.7* [Computer software]. University of Victoria, Canada: Authors.
- MAY, R. B., MASSON, M. E. J., & HUNTER, M. A. (1989). Randomization tests: Viable alternatives to normal curve tests. *Behavior Research Methods, Instruments, & Computers*, 21, 482-483.
- MAY, R. B., MASSON, M. E. J., & HUNTER, M. A. (1990). *Application of statistics in behavioral research*. New York: Harper & Row.
- MICCERI, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- MOOK, D. G. (1983). In defense of external invalidity. *American Psychologist*, 38, 379-387.
- OAKES, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- PEDHAZUR, E. J., & SCHMELKIN, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- PITMAN, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society (Series B)*, 4, 119-130.
- PITMAN, E. J. G. (1937b). Significance tests which may be applied to samples from any population: II. *Journal of the Royal Statistical Society (Series B)*, 4, 225-232.
- POPPER, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books. (Original work published 1935)
- ROMANO, J. P. (1989). Bootstrap and randomization tests of some non-parametric hypotheses. *Annals of Statistics*, 17, 141-159.
- SIMON, J. L. (1992). *Resampling: The new statistics*. Arlington, VA: Resampling Stats.
- StatXact* [Computer software] (1995). Cambridge, MA: Cytel Software.
- VAN DEN BRINK, W. P., & VAN DEN BRINK, S. G. J. (1989). A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *Journal of Mathematical & Statistical Psychology*, 42, 183-189.
- VELLEMAN, P. F., & HOAGLIN, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury.
- WELCH, W. J. (1990). Construction of permutation tests. *Journal of the American Statistical Association*, 85, 693-698.

(Manuscript received February 24, 1995;
accepted for publication February 27, 1995.)