STATISTICS, MATHEMATICS, AND TEACHING

David S. Moore
Purdue University

In discussing our teaching, we may focus on *content*, what we want our students to learn, or on *pedagogy*, what we do to help them learn. These two topics are of course related. In particular, changes in pedagogy are often driven in part by changing priorities for what kinds of things we want students to learn. It is nonetheless convenient to address content and pedagogy separately. Pedagogy, certainly the less specific of the two, is the topic of my second paper. This paper concerns content, and in particular contains one side of a conversation between a statistician and mathematicians who may find themselves teaching statistics.

How does statistics differ from mathematics? How does this affect the teaching of statistics to beginners? Statisticians are convinced that statistics, while a mathematical science, is not a subfield of mathematics. Like economics and physics, statistics makes heavy and essential use of mathematics, yet has its own territory to explore and its own core concepts to guide the exploration. I will not rehearse the evidence that statistics is not mathematics, which appears in a somewhat polemical form in [17].

Given those convictions, we would naturally prefer that beginning statistics be taught *as statistics*. The American Statistical Association and the MAA have formed a joint committee to discuss the curriculum in elementary statistics. The recommendations of that group reflect the view that statistics instruction should focus on *statistical* ideas. Here are some excerpts (Cobb [8]; a longer discussion appears in [9]):

> Almost any course in statistics can be improved by more emphasis on data and concepts, at the expense of less theory and fewer recipes. To the maximum extent feasible, calculations and graphics should be automated.
>
> Any introductory course should take as its main goal helping students to learn the basics of statistical thinking. [These include] the need for data, the importance of data production, the omnipresence of variability, the quantification and explanation of variability.

The recommendations of the ASA/MAA committee reflect changes in the field of statistics over the past generation. Academic statistics, unlike mathematics, is linked to a larger body of non-academic professional practice. Computing technology has completely changed the practice of statistics. Academic researchers, driven in part by the demands of practice and in part by the capability of new technology, have changed their taste in

research. Bootstrap methods, nonparametric data smoothing, regression diagnostics, and more general classes of models that require iterative fitting are among the recent fruits of renewed attention to analysis of data and scientific inference. Efron and Tibshirani [10] describe some of this work for non-specialists.

Many mathematicians recognize that statistics (unlike probability theory) is a distinct discipline. I am fond of the statement by the eminent probabilist David Aldous [1] that he "is interested in the applications of probability to all scientific fields *except statistics.*" I should add at once that although mathematics can prosper without statistics, the converse fails. Bullock's [3] claim that "Many statisticians now insist that their subject is something quite apart from mathematics, so that statistics courses do not require any preparation in mathematics." draws a clearly false implication. All statistics courses require some preparation in mathematics, and some require a great deal. Elaborate mathematical theories undergird some parts of statistics, and the study of these theories is part of the standard training of statisticians.

Our topic of conversation here is rather the nature of beginning instruction in statistics at any level. Even mathematically sophisticated students would do well not to begin their study of physics with a course in analytical mechanics that revels in the mathematical formalism and assumes some prior acquaintance with the physical phenomena that the mathematics describes. So it is with statistics. The subject matter of statistics is data, and any responsible introduction to statistics should begin by giving students experience with data and a working knowledge of the concepts that organize the statistician's approach to data.

Neither Mathematics Nor Magic

It is helpful to rehearse the reasons why instruction in beginning statistics that is driven by theory (even when that theory is not explicitly taught) is a bad idea. More detail on this appears in [19]. Here is an example of a simple statistical problem. (From [20]; the full study is described by Lyle et al. [16]).

> Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Researchers therefore conducted an experiment.
>
> The subjects in part of the experiment were 21 healthy black men. A randomly chosen group of 10 of the men received a calcium supplement for 12 weeks. The control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic blood pressure for a subject after 12 weeks, in millimeters of mercury.

Take Group 1 to be the calcium group and Group 2 the placebo group. Here are the data for the 10 men in Group 1 (calcium),

$$7 \quad -4 \quad 18 \quad 17 \quad -3 \quad -5 \quad 1 \quad 10 \quad 11 \quad -2$$

and for the 11 men in Group 2 (placebo),

$$-1 \quad 12 \quad -1 \quad -3 \quad 3 \quad -5 \quad 5 \quad 2 \quad -11 \quad -1 \quad -3$$

From the data, calculate the summary statistics:

Group	Treatment	Sample size	Mean	Standard deviation
1	Calcium	10	5.000	8.743
2	Placebo	11	−.273	5.901

The calcium group shows a drop in blood pressure, $\overline{x} = 5.000$, while the placebo group had almost no change, $\overline{y} = -.273$. Is this outcome good evidence that calcium decreases blood pressure in the entire population of healthy black men more than a placebo does?

Standard procedures for analyzing this example assume that the data fit this mathematical model:

$$X_1, X_2, \ldots, X_n \text{ iid } N(\mu_1, \sigma_1)$$
$$Y_1, Y_2, \ldots, Y_m \text{ iid } N(\mu_2, \sigma_2)$$

A routine significance test derived from this model answers the "Is this outcome good evidence?" question. But is the model adequate?

This model is in fact incomplete in a most serious way: it does not distinguish between observational data (e.g., from a sample survey) and data from a randomized comparative experiment such as the calcium study. The distinction between observation and experiment is one of the most important in statistics. The researchers want to reach *causal* conclusions: calcium *causes* a reduction in blood pressure. Experiments often allow causal conclusions, while observational studies almost always leave issues of causation unsettled and subject to debate. Yet the mathematical models of statistical theory are identical for observational and experimental data.

The model, like most idealized mathematical models for real phenomena, is also unrealistic. In the words attributed to the statistician George Box, "All models are wrong, but some are useful." The user of inference methods based on this model must carefully explore its adequacy to her setting and her data. Were there flaws in the data production (whether sample or experiment) that render inference meaningless? Are the data, which are certainly not independent observations on a perfectly normal distribution, sufficiently normal to allow use of standard procedures? This question is answered by exploratory examination of the data themselves, combined with knowledge of how "robust" the planned analysis is under deviations from normality.

Theory-based instruction tends to emphasize how methods follow from models, often with only the most general warnings about the realities of practice. Statistics in practice is close to a dialog between models and data. Models for the process that produced our data do indeed play a central role in statistical inference. The mathematical exploration of properties and consequences of models is therefore important (as it is in economics and physics). But the data are also allowed to criticize and even falsify proposed models. We can modify Box's dictum into a practical version of the statement that statistics is not just mathematics: *Mathematical theorems are true; statistical methods are sometimes effective when used with skill.*

The mathematical model provides a basis for *formal statistical inference*, the confidence intervals and significance tests familiar to all students of statistics. Our brief discussion has pointed to the importance of the *design of data production* and of *exploratory analysis of data.* These aspects of statistics are not founded on a mathematical theory and will be neglected in a mathematically-structured treatment. Yet they are fundamentally important in both statistical practice and statistical research. An over-emphasis on probability-based inference is one mark of an overly mathematical introduction to statistics.

The reluctance of mathematically trained teachers to abandon a theory-driven presentation of basic statistics has a respectable basis: to avoid presenting statistics as magic. It is certainly common to teach beginning statistics as magic. The user of statistics is in many ways very like the sorcerer's apprentice. The incantation has an automatic effectiveness, rendering theses acceptable and studies publishable. We are not meant to understand how the incantation works—that is the domain of the sorcerer himself. The incantation must follow the recipe exactly, lest disaster ensue—exploration and flexibility, like understanding, are forbidden to the apprentice. Fortunately, the sorcerer has provided software that automates the exact following of approved incantations.

The danger of statistics-as-magic is real. But the proper defense is not a retreat to a mathematical presentation that is inadequate to the subject and often incomprehensible to students. *Mathematical understanding is not the only kind of understanding.* It is not even the most helpful kind in most disciplines that employ mathematics, where understanding of the target phenomena and core concepts of the discipline take precedence. We should attempt to present an intellectual framework that makes sense of the collection of tools that statisticians use and encourages their flexible application to solve problems. A student understands mathematics when she appreciates the power of abstraction, deduction, and symbolic expression, and can use mathematical tools and strategies flexibly in dealing with varied problems. Reasoning from uncertain empirical data is a similarly powerful and pervasive intellectual method. What follows is an inadequate attempt to describe the intellectual framework of statistics and to comment on implications for teaching.

What is Statistics?

Statistics is a methodological discipline. It exists not for itself but rather to offer to other fields of study a coherent set of ideas and tools for dealing with data. The need for such a discipline arises from *the omnipresence of variability.* Individuals vary. Repeated measurements on the same individual vary. In some circumstances, we want to find unusual individuals in an overwhelming mass of data. In others, the focus is on the variation of measurements. In yet others, we want to detect systematic effects against the background noise of individual variation. Statistics provides means for dealing with data that take into account the omnipresence of variability. It is helpful to organize the subject matter of statistics under three heads:

1. Analyzing and describing data.
2. Producing data.
3. Inference from data.

Data analysis

Data analysis is the contemporary form of "descriptive statistics," powered by more numerous and more elaborate descriptive tools, but especially by a philosophy due in large measure to John Tukey of Bell Labs and Princeton. The philosophy is captured in the now-common name, *exploratory data analysis*, or EDA. The goal of EDA is to see what the data in hand say, on the analogy of an explorer entering unknown lands. We put aside (but not forever) the issue of whether these data represent any larger universe. Here is an elementary summary (from [20]) of the distinctions between EDA and standard inference:

EXPLORATORY DATA ANALYSIS	STATISTICAL INFERENCE
Purpose is unrestricted exploration of the data, searching for interesting patterns.	Purpose is to answer specific questions, posed before the data were produced.
Conclusions apply only to the individuals and circumstances for which we have data in hand.	Conclusions apply to a larger group of individuals or a broader class of circumstances.
Conclusions are informal, based on what we see in the data.	Conclusions are formal, backed by a statement of our confidence in them.

In practice, exploratory analysis is a prerequisite to formal inference. Most real data contain surprises, some of which can invalidate or force modification of the inference that was planned. Running data through a sophisticated (and therefore automated) inference procedure before exploring them carefully is the mark of a statistical novice. The dialog

between data and models continues with more advanced diagnostic tools that allow data to criticize specific models. These tools combine the EDA spirit with the results of mathematical analysis of the consequences of the models.

Wide availability of cheap computing, especially graphics, has combined with the desire to "let the data speak" to generate an abundance of new tools: stemplots, boxplots, model-free scatterplot smoothers, resistant regression algorithms, clever ideas for display of high-dimensional data on two-dimensional screens, and many more advanced diagnostic tools for specific situations. Standard statistical software implements much of this. The books [5] and [7], by Bell Labs scientists influenced by Tukey, present much of the basic graphical material. The software packages S and S-PLUS, which originated at Bell Labs, implement more of the new graphics and also implement several new classes of models. See [6] for detailed discussion of the latter.

At the level of beginning instruction, it is easy to view data analysis as a collection of clever tools (stemplots, five-number summaries, ...). We should attempt to also offer our students an overview, to help them grasp the strategies that organize the examining of data:

1. Proceed from simple to complex: first examine each variable individually, then look at relationships among them.

2. Use a hierarchy of tools: first plot the data, then choose appropriate numerical descriptions of specific aspects of the data, then if warranted select a compact mathematical model for the overall pattern of the data.

3. Look at both the overall pattern and at any striking deviations from that pattern.

We reinforce these principles by filling in the specifics in each of several settings. Given data on a single quantitative variable, we may expect students to display the distribution by a stemplot, note that it reasonably symmetric, calculate the mean and standard deviation as numerical summaries, and use a normal quantile plot to see whether a normal distribution is a suitable compact model for the overall pattern. Given two quantitative variables, we draw a scatterplot, measure the direction and strength of linear association by the correlation, and, if warranted, use a fitted straight line as a model for the overall pattern.

I expect students to write coherent descriptions of data. To help them, I provide outlines for implementing the second and third points above in various settings. Figure 1, for example, is the outline for describing a single quantitative variable. Following this outline requires both knowledge of the tools mentioned and judgment to choose among them and interpret the results. Judgment is formed by experience with data. Students cannot at first "read" graphs any more than they can read words or equations. Here is an example of a basic one-variable data analysis. Describing relations among several variables requires more elaborate tools and finer judgment.

Figure 1: Outline for describing data on a single quantitative variable

A. Describe the data

- number of observations
- nature of the variable
- how it was measured
- units of measurement

B. Plot the data; choose from

- dotplot
- stemplot
- histogram

C. Describe the overall pattern

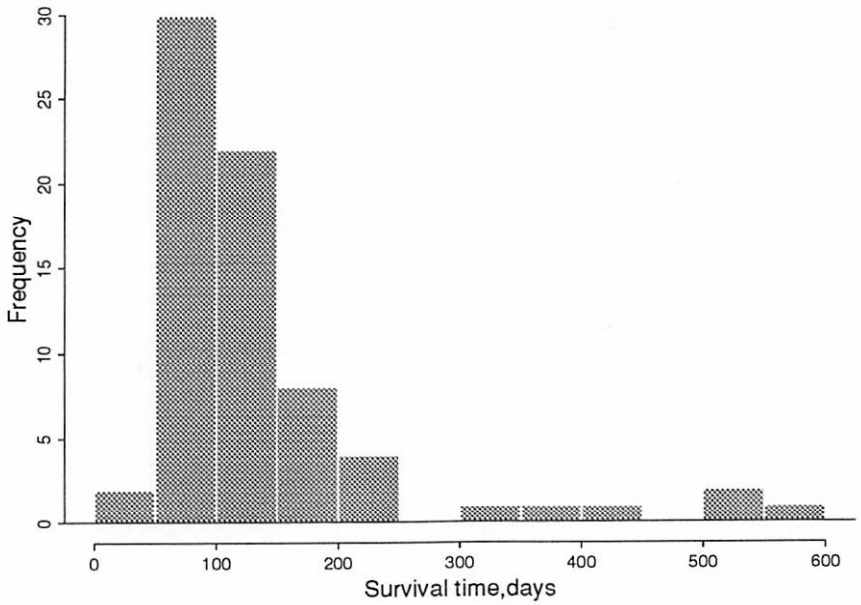
- shape
  - no clear shape?
  - skew or symmetric?
  - single or multiple peaks?
- center and spread; choose from
  - five-number summary
  - mean and standard deviation
- is normality an adequate model (normal quantile plot)?

D. Look for striking deviations from the overall pattern

- outliers
- gaps or clusters

E. Interpret your findings in C and D in the language of the problem setting. Suggest plausible explanations for your findings.

Figure 2: Histogram of guinea pig survival times

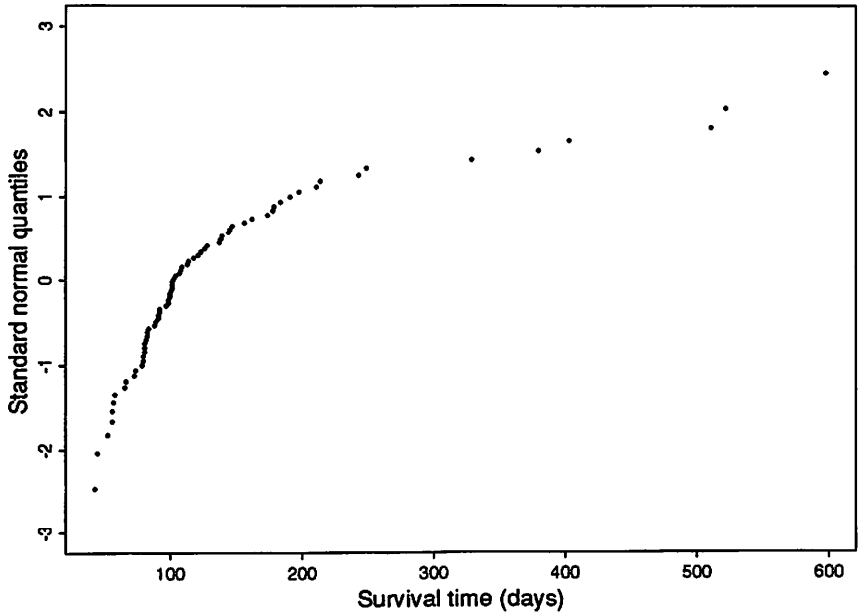


In a study of resistance to infection [2], researchers injected 72 guinea pigs with tubercle bacilli and measured their survival time in days after infection. Both a histogram (Figure 2) and a normal quantile plot (Figure 3) show that the distribution of survival times is strongly skewed to the right. There are no outliers—although some individuals survived far longer than the average, this appears to be a characteristic of the overall distribution rather than pointing to, for example, errors in measuring or recording these individuals.

The strong skewness suggests that the five number summary (min = 43 days, first quartile = 82.5 days, median = 102.5 days, third quartile = 151.5 days, max = 598 days) is a better numerical summary than the mean and standard deviation ($\overline{x}$ = 141.8 days, s = 109.2 days). There is very large variation in survival times among the individuals—for example, the third quartile is almost 150% of the median and the largest 6 observations are more than double the median. Without more information, we cannot accurately predict the survival time of an infected individual. Moreover, standard t procedures should not be used for inference about survival time. Inference could employ a non-normal distribution as a model or seek a transformation to a scale that is more nearly normal.

I have tried to suggest that there is a coherent (though not mathematical) set of ideas and associated tools for exploring data. This material is core statistics. Moreover, students like it and find that they can do it, a substantial bonus when teaching a subject feared by many. Finally, exploration of data raises issues that prepare the way for

Figure 3: Normal quantile plot for guinea pig survival times



inference. A week of "descriptive statistics" at the beginning of a course isn't an adequate introduction to data analysis.

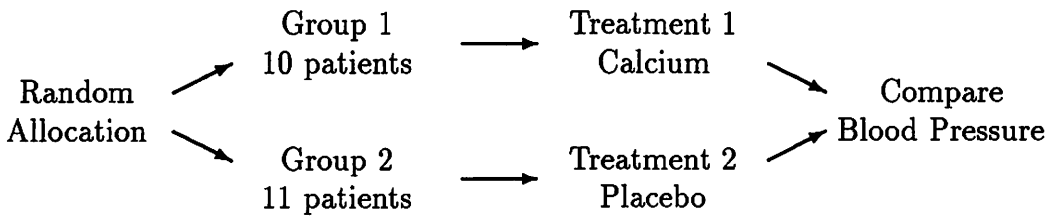
Data production

Statistical ideas for producing data to answer specific questions are the most influential contributions of statistics to human knowledge. Badly designed data production is the most common serious flaw in statistical studies. Well designed data production allows us to apply standard methods of analysis and reach clear conclusions. Professional statisticians are paid for their expertise in designing studies; if the study is well designed (and no unanticipated disaster occurred), you don't need a professional to do the analysis. In other words, the design of data production is *really* important. If you just say "Suppose X_1 to X_n are iid observations," you aren't teaching statistics.

The clinical trial on the effect of calcium on blood pressure was a *randomized comparative experiment.* Figure 4 presents the design in outline form.

Randomized comparative experiments are intended to produce good evidence that the experimental treatments actually *cause* changes in the response. The random assignment of subjects eliminates bias in forming the treatment groups and produces groups that differ only through chance variation before we apply the treatments. The comparative design reminds us that all subjects are treated exactly alike except for the contents of the pills they take. Thus if we observe differences in the mean reduction in blood pressure greater than could be expected to arise by chance, we can be confident that the calcium brought about the effect we see.

Figure 4: The simplest randomized comparative experiment



Students should understand why randomized comparative experiments are the gold standard for evidence of causation. Only then can they understand, for example, the arguments against making available to patients with AIDS or other fatal disease any treatment that has any promise of helping: we would then never learn which of these treatments actually help, which have no effect, and which are on balance harmful. The book [4] edited by the physicians Bunker and Barnes and the statistician Mosteller contains striking examples of medical treatments that became standard in the days before medicine adopted randomized comparative experiments, and were found to be worthless when subjected to proper testing. Some AIDS activists come close to demanding a return to the time when unproved and perhaps harmful treatments could avoid the scrutiny of statistically designed trials.

The other major means of producing data are *sample surveys* that choose and examine a sample in order to produce information about a larger population. Interesting examples abound—opinion polls sound and unsound, government collection of economic and social data, academic data sources such as the National Opinion Research Center at the University of Chicago. Statistical designs for sampling begin by insisting that impersonal chance should choose the sample. The central idea of statistical designs for producing data, through either sampling or experimentation, is the deliberate use of chance. Explicit use of chance mechanisms eliminates some major sources of bias. It also ensures that quite simple probability models describe our data production processes, and therefore that standard inference methods apply. Designs for data production offer the most secure basis for statistical inference in practice, and also provide a natural transition to inference in teaching.

There is of course more to the statistical side of designing experiments and sample surveys than "randomize." The designs used in practice are often quite complex, and must balance efficiency with the need for information of varying precision about many factors and their interactions. Simple designs—randomized experiments comparing two or several treatments, simple random samples from one or several populations—illustrate the most important ideas and support the inference taught in a first statistics course. You must talk about these designs, but need not go farther. Some other important material, for example, procedures for developing and testing survey questions and for training and

supervising interviewers, is not usually presented in statistics courses. Statistics students should be aware that these practical skills do matter, and that data production can go awry even when we start with a sound statistical design. How much time to spend here is a matter of a teacher's judgment of the needs of her audience.

Inference

Statistical inference provides methods for drawing conclusions from data about the population or process from which the data were drawn. It now becomes essential (as it was not in data analysis) to distinguish sample *statistics* from population *parameters*. The true values of the parameters are unknown to us. We have the statistics in hand, but they would take different values if we repeated our data production. Inference must take this sample variability into account.

Probability describes one kind of variability, the chance variability in random phenomena. When a chance mechanism is explicitly used to produce data, probability therefore describes the variation we expect to see in repeated samples from the same population or repeated experiments in the same setting. That is, probability answers the question, "What would happen if we did this many times?" Standard statistical inference is based on probability. It offers conclusions from data *along with* an indication of how confident we are in the conclusions. The statement of confidence is based on asking "What would happen if I used this inference method many times?" That is exactly the kind of question probability can answer (which is why we ask it). The indication of our confidence in our methods, expressed in the language of probability, is what distinguishes formal inference from informal conclusions based on e.g., an exploratory analysis of data.

Any particular inference procedure starts with a statistic (perhaps several statistics) calculated from the sample data. The *sampling distribution* is the probability distribution that describes how this statistic would vary if we drew many samples from the same population. In elementary statistics we present two types of inference procedures, confidence intervals and significance tests. A confidence interval estimates an unknown parameter. A significance test assesses the evidence that some sought-after effect is present in the population.

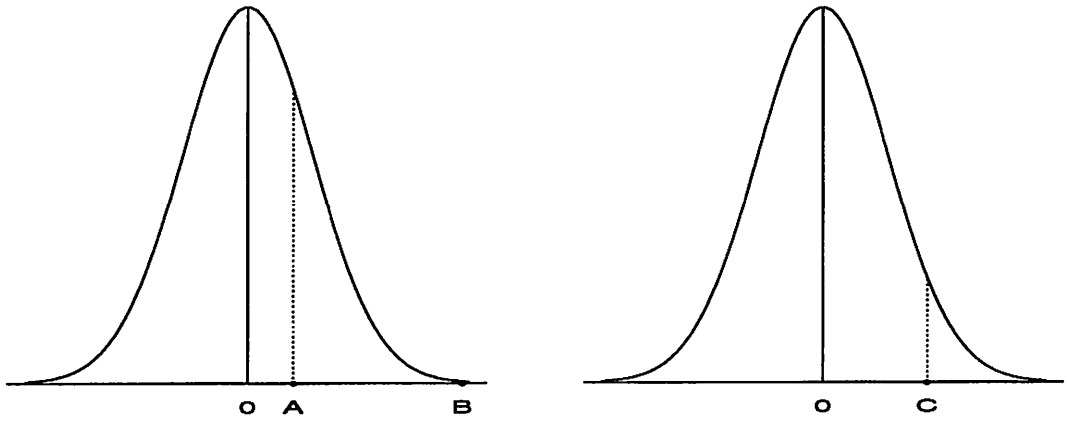
A *confidence interval* consists of a recipe for estimating an unknown parameter from sample data, usually of the form

$$\text{estimate} \pm \text{margin of error}$$

and a confidence level, which is the probability that the recipe actually produces an interval that contains the true value of the parameter. That is, the confidence level answers the question, "If I used this method many times, how often would it give a correct answer?"

A *significance test* starts by supposing that the sought-after effect is *not* present in the population. It asks "In that case, is the sample result surprising or not?" A probability (the P-value) says how surprising the sample result is. A result that would

Figure 5: Is this observation surprising?



rarely occur if the effect we seek were absent is good evidence that the effect is in fact present. Figure 5 illustrates this reasoning in our medical example. The normal curves in that figure represent the sampling distribution of the difference $\overline{x} - \overline{y}$ between the mean blood pressure decreases in the calcium and placebo groups, for the case of no difference between the two population means. This distribution, which shows the variability due to chance alone, has mean 0. Outcomes greater than 0 come from experiments in which calcium reduces blood pressure more than the placebo. If we observe result A, we are not surprised; an outcome this far above 0 would often occur by chance. It provides no credible evidence that calcium beats the placebo. If we observe result B, on the other hand, the experiment has produced an effect so strong that it would almost never occur simply by chance. We then have strong evidence that the calcium mean does exceed the placebo mean. The P-value (the right tail probability) is 0.24 for point A and 0.0005 for point B. These probabilities quantify just how surprising an observation this large is when there is no effect in the population. What about the actual data? Point C shows the observed value $\overline{x} - \overline{y} = 5.273$. The corresponding P-value is 0.055. Calcium would beat the placebo by at least this much in 5.5% of many experiments just by chance variation. The experiment gives some evidence that calcium is effective, but not extremely strong evidence. (A note for those who worry about details: These P-value calculations took the variability of the sample means to be known. In practice, we must estimate standard deviations from the data. The resulting test has a larger P-value, $P = 0.072$.)

Those five paragraphs describe briefly how inference works. Because the details are in practice automated, we would like students to grasp these ideas. They are not easy to grasp. The first barrier is the notion of a sampling distribution. Choose a simple setting, such as using the proportion $\hat{p}$ of a sample of workers who are unemployed to estimate the proportion p of unemployed workers in an entire population. Physical examples (sampling beads from a box), computer simulations, and encouraging thought experiments all help

convey the idea of many samples with many values of $\hat{p}$. Keep asking, "What would happen if I did this many times?" That question is the key to the logic of standard statistical inference.

Once the idea of a sampling distribution begins to settle, the tools of data analysis help us take the next steps. Faced with any distribution, we ask about shape, center, and spread. The shape of the sampling distribution of $\hat{p}$ is approximately normal. The mean is equal to the unknown population proportion p. This says that $\hat{p}$ as an estimator of p has no bias, or systematic error. The precision of the estimator is described by the spread of the sampling distribution, which (thanks to normality) we measure by its standard deviation. We are now only details away from confidence intervals.

The second major barrier is the reasoning of significance tests. Although the basic idea ("Is this outcome surprising?") is not recondite, the details are daunting. There's no escape from null and alternative hypotheses and one- versus two-sided tests. The logic of testing, which starts out "Suppose for the sake of argument that the effect we seek is not present ..." isn't straightforward. I'd like most of my students to understand the idea of a sampling distribution; I know that quite a few won't understand the reasoning of significance tests. My fallback position is to insist that they be able to verbalize the meaning of P-values produced by software or reported in a journal. This is part of insisting that students write succinct summaries of statistical findings. "The study compared two methods of teaching reading to third-grade students. A two-sample t test comparing the mean scores of the two treatment groups on a standard reading test had P-value $P = 0.019$. That is, the study observed an effect so large that it would occur just by chance only about 2% of the time. This is quite strong evidence that the new method does result in a higher mean score than the standard method."

Two concluding remarks about inference. First, a conceptual grasp of the ideas is almost pictorial, based on picturing the sampling distribution and following the tactics learned in data analysis. No amount of formal mathematics can replace this pictorial vision, and no amount of mathematical derivation will help most of our students see the vision. The mathematics is essential to our knowing the facts, but this does not imply that we should impose the mathematics on our students.

Second, we want our students to know a good deal more than the big picture and several recipes that implement it in specific settings. Here are some further points, both practical and conceptual, roughly in order of importance. How far down the list you should go depends on your audience.

- Study of specific inference procedures reveals behaviors that are common and that all students should understand. To get higher confidence from the same data, you must pay with a larger margin of error. Even effects so small as to be practically unimportant will be highly significant in the statistical sense if we base a significance test on a very large sample.

- Lots of things can go wrong that make inference of dubious value. Comparing subjects who *choose* to take calcium against others who don't tells little about the

effects of calcium, because those who choose to take calcium may be very health-conscious in general. One extreme outlier could pull the conclusion of our medical experiment in either direction, again invalidating making the results of inference. Examine the data production. Plot the data. Then, perhaps, go on to inference.

- Inference procedures themselves don't tell us that something went wrong. The margin of error in a confidence interval, for example, includes *only* the chance variation in random sampling. As the *New York Times* says in the box that accompanies its opinion poll results, "In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll."

- Common inference procedures really are based on mathematical models like the one that appears in our medical example,

$$X_1, X_2, \ldots, X_n \quad \text{iid} \quad N(\mu_1, \sigma_1)$$

$$Y_1, Y_2, \ldots, Y_m \quad \text{iid} \quad N(\mu_2, \sigma_2)$$

  This model isn't exactly true; is it useful? In fact, the two-sample t procedures that follow from this model when we want to compare μ_1 and μ_2 are quite robust against non-normality. So the model does lead to practically useful procedures. But the variance ratio F statistic for comparing σ_1 and σ_2 is extremely sensitive to non-normality, so much so that it is of little practical value. Even beginners need to be aware of such issues.

- We often want to do inference when our data do not come from a random sample or randomized comparative experiment. Think, for example, of measurements on successive parts flowing from an assembly line. Inference is justified by a probability model for the process that produced our data, and the correctness of the model can to some extent be assessed from the data themselves. Randomized data production is the paradigm and the most secure setting for inference, but it is not the only allowable setting.

- Inductive inference from data is conceptually complex. It's not surprising that there are alternative ways of thinking about it. Standard statistical theory tends to think of inference as if its purpose were to make decisions. A test must decide between the null and alternative hypotheses, for example. This leads at once to Type I and Type II errors and so on. The decision-making approach fits uneasily with the "Is this outcome surprising?" logic expressed by P-values. I think that assessing the strength of evidence is a much more common goal than making a decision, but not everyone agrees. The Bayesian school of thought goes farther, by introducing an explicit description of the available prior information into any statistical setting and combining prior information with data to reach a decision. Almost all statisticians think this is sometimes a good idea. Bayesians think *all*

statistical problems can be made to fit their paradigm. This is a (strongly held) minority position. Deep water ahead.

What About Probability?

Probability is an essential part of any mathematical education. It is an elegant and powerful field of mathematics that enriches the subject as a whole by its interactions with other fields of mathematics. Probability is also essential to serious study of applied mathematics and mathematical modeling. The domain of determinism in natural and social phenomena is limited, so that the mathematical description of random behavior must play a large role in describing the world. Whether our mathematical tastes run to purity or modeling, probability helps to satisfy them.

We are, however, discussing introductory statistics rather than mathematics. Probability is the branch of mathematics most heavily applied in statistics. In particular, probability provides the theoretical structure of standard statistical inference, which is based on asking "What would happen if we used this method very many times?" What should be the place of probability in beginning instruction in statistics? My position is not standard, though it is gaining adherents: first courses in statistics should contain essentially no formal probability theory.

Why? First, because *informal probability is sufficient for a conceptual grasp of inference.* The "what would happen" question of standard inference is answered by referring to the sampling distribution of a statistic, which records the pattern of variation of the outcomes of, for example, many random samples from the same population. If we agree that actually deriving these distributions is better left to more advanced study, they can be understood as distributions using the tools of data analysis, without the apparatus of formal probability. Rules for $P(A \bigcup B)$ add very little to a statistics course.

The second reason to avoid formal probability is that *probability is conceptually the hardest subject in elementary mathematics.* The history of probabilistic ideas (see for example [12] and [21]) is fascinating but a bit frightening. Better minds than ours long found the subject confusing in the extreme. Psychologists, beginning with Tversky and his collaborators, have demonstrated that confusion persists, even among those who can recite the axioms of formal probability and who can do textbook exercises. Our intuition of random behavior is gravely and systematically defective. See e.g. Tversky and Kahneman [22] and the collection by Kapadia and Borovcnik [15]. What is worse, mathematics educators have found no effective way to correct our defective intuition. Garfield and Ahlgren [11] conclude a review of research by stating that "teaching a conceptual grasp of probability still appears to be a very difficult task, fraught with ambiguity and illusion." They suggest study of "how useful ideas of statistical inference can be taught independently of technically correct probability." I believe that concentrating on the idea of a sampling distribution allows this, at least at the depth appropriate for beginners.

The concepts of statistical inference, starting with sampling distributions, are of course also quite tough. We ought to concentrate our attention, and our students' limited

patience with hard ideas, on the essential ideas of statistics. We faculty imagine that formal probability illumines those ideas. That's simply not true for almost all of our students.

What About Mathematics Majors?

Mathematics majors traditionally meet statistics as the second course in a year-long sequence devoted to probability and statistical theory. I hope it is clear that I don't regard a tour of sufficient statistics, unbiasedness, maximum likelihood estimators, and the Neyman-Pearson theorem as a promising way to help students understand the core ideas of statistics. On the other hand, mathematics majors should certainly see some of the mathematical structure of statistical inference. What ought we to do?

My preference is to precede the study of theory by a thorough data-oriented introduction to statistical ideas and methods and their applications. That is, mathematics students are not necessarily an exception to the principle that a first introduction to statistics should not be based on formal probability. If the students have strong quantitative backgrounds, a data-oriented course can move quickly enough to present genuinely useful statistics and serious applications. The need for theory can be made clear as we face issues of practice, and the theory makes much more sense when its setting in practice is clear. In many institutions, however, constraints or faculty hesitation make this path difficult. In others, there is little coordination between the "applied" and theoretical courses, so that the latter does not in fact build on the former.

We ought therefore to reconsider what a one-semester introduction to statistics for mathematics majors and other quantitatively strong students should look like. This course will naturally follow a course in probability. Here we encounter another barrier: we can't in good conscience retool both semesters of the standard probability-statistics sequence to optimize the introduction to statistics. Probability is important in its own right, not just as preparation for statistical theory. The more emphasis a department places on applications and modeling in its major curriculum, the more the probability course must play an essential role in this emphasis. An introduction to probability that emphasizes modeling and includes simulation and numerical calculation certainly sets the stage for statistics, but I am hesitant to move any strictly statistical ideas into the probability semester. The reform of probability and the reform of statistics are distinct issues.

Our goal, I think should be an integrated statistics course that moves through data analysis, data production, and inference in turn, emphasizing the organizing principles of each. We should certainly take advantage of and strengthen the student's mathematical capacities. Although data analysis and data production have no unifying theory, mathematical analysis can illumine even data analysis. Here are a few examples.

A. Consider the optimality properties of measures of center for n observations. The mean minimizes the mean squared error; the median minimizes

the mean absolute error (and need not be unique); the midrange minimizes the maximum absolute (or squared) error; try minimizing the *median* absolute error for $n = 3$ and examine the unpleasant behavior of the resulting measure.

B. Students met the Chebychev inequality while studying probability. Now they may meet the interesting inequality $|\mu - m| \leq \sigma$ linking the mean, median, and standard deviation of any distribution (see Watson [23]). Describe one-sample data by the empirical distribution (probability $1/n$ on each observed point) to draw conclusions about how far apart the sample mean and median may be.

C. The least-squares regression line is the analog of the mean $\overline{x}$ for predicting y from x. Derive it. Then explore, perhaps using software, analogs of the other measures in A above.

Data production lends itself to probability calculations that illustrate how likely it is that random assignments will be unbalanced in specific ways; the advantages of large samples soon become clear.

Very nice. We can give our students a balanced introduction to statistics that makes use of their knowledge of mathematics. The inevitable consequence is that we spend less time on inference. We must decide what to preserve and what to cut. There is as yet no consensus, because despite much grumbling, the reform of the math major sequence has not yet begun. Imagining such a reform is a good place to end a discussion of statistics, mathematics, and teaching. This is your take-home exam: design a better one-semester statistics course for mathematics majors.

REFERENCES

1. Aldous, David (1994), "Triangulating the circle, at random," *American Mathematical Monthly*, **101**, pp. 223–233. The remark appears in the biographical note accompanying the paper.

2. T. Bjerkedal (1960), "Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli," *American Journal of Hygiene*, **72**, pp. 130–148.

3. Bullock, James O. (1994), "Literacy in the language of mathematics," *American Mathematical Monthly*, **101**, pp. 735–743.

4. Bunker, John. P., Benjamin A. Barnes, and Frederick Mosteller (eds.) (1977), *Costs, Risks and Benefits of Surgery.* New York: Oxford University Press.

5. Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey (1983), *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth.

6. Chambers, John M. and Trevor J. Hastie (1992), *Statistical Models in S.* Pacific Grove, CA: Wadsworth.

7. Cleveland, William S. and Mary E. McGill (eds.) (1988), *Dynamic Graphics for Statistics.* Belmont, CA: Wadsworth.

8. Cobb, George (1991), "Teaching statistics: more data, less lecturing," *Amstat News*, December 1991.

9. Cobb, George (1992), "Teaching statistics," in L. A. Steen (ed.) *Heeding the Call for Change: Suggestions for Curricular Action*, MAA Notes 22. Washington, DC: Mathematical Association of America.

10. Efron, Bradley and Rob Tibshirani (1991), "Statistical data analysis in the computer age," *Science* **253**, pp. 390–395.

11. Garfield, Joan and Andrew Ahlgren (1988), "Difficulties in learning basic concepts in probability and statistics: implications for research," *Journal for Research in Mathematics Education*, **19**, pp. 44–63.

12. Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krüger (1989) *The Empire of Chance.* Cambridge: Cambridge University Press.

13. Hoaglin, D. C. (1992), "Diagnostics," in D. C. Hoaglin and D. S. Moore (eds.), *Perspectives on Contemporary Statistics.*

14. Hoaglin, David C. and David S. Moore (eds.) (1992), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America.

15. Kapadia, R. and M. Borovcnik (eds.) (1991), *Chance Encounters: Probability in Education.* Dordrecht: Kluwer.

16. Lyle, Roseann M. et al. (1987), "Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men," *Journal of the American Medical Association*, **257**, pp. 1772–1776. Dr. Lyle provided the data in the example.

17. Moore, David S. (1988), "Should mathematicians teach statistics" (with discussion), *College Mathematics Journal*, **19**, pp. 3–7.

18. Moore, David S. (1992), "What is statistics," in David C. Hoaglin and David S. Moore (eds.), *Perspectives on Contemporary Statistics.*

19. Moore, David S. (1992), "Teaching statistics as a respectable subject," in Florence Gordon and Sheldon Gordon (eds.), *Statistics for the Twenty-First Century*, MAA Notes 26. Washington, DC: Mathematical Association of America.

20. Moore, David S. (1995), *The Basic Practice of Statistics.* New York: W. H. Freeman.

21. Stigler, S. M. (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge, Mass: Belknap.

22. Tversky, Amos and Daniel Kahneman, D. (1983), "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment," *Psychological Review* **90**, pp. 293–315.

23. Watson, G. S. (1994), letter to the editor, *The American Statistician* **48**, p. 269. This is the last in a series of comments on this inequality, and contains references to the earlier contributions.