



Mathematics, Statistics, and Teaching

George W. Cobb; David S. Moore

The American Mathematical Monthly, Vol. 104, No. 9. (Nov., 1997), pp. 801-823.

Stable URL:

<http://links.jstor.org/sici?sici=0002-9890%28199711%29104%3A9%3C801%3AMSAT%3E2.0.CO%3B2-7>

The American Mathematical Monthly is currently published by Mathematical Association of America.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/maa.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

Mathematics, Statistics, and Teaching

George W. Cobb and David S. Moore

How does statistical thinking differ from mathematical thinking? What is the role of mathematics in statistics? If you purge statistics of its mathematical content, what intellectual substance remains?

In what follows, we offer some answers to these questions and relate them to a sequence of examples that provide an overview of current statistical practice. Along the way, and especially toward the end, we point to some implications for the teaching of statistics.

1. INTRODUCTION: AN OVERVIEW OF STATISTICAL THINKING. 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.

1.1. The role of context. The focus on variability naturally gives statistics a particular content that sets it apart from mathematics itself and from other mathematical sciences, but there is more than just content that distinguishes statistical thinking from mathematics. Statistics requires a different *kind* of thinking, because *data are not just numbers, they are numbers with a context*.

Example 1. The mystery of Andover. The finite sequence (3, 5, 23, 37, 6, 8, 20, 22, 1, 3) shows a distinctive pattern when plotted (Figure 1) but the numbers and the pattern have no meaning or interest until we know their context. They are in fact monthly totals of people formally accused of witchcraft in Essex County, Massachusetts, beginning in February, 1692. The plot shows two waves of accusations, separated by a low point in the summer of 1692. The pattern becomes still more meaningful when we know that the first hanging of a convicted witch (Bridget Bishop) took place June 10, 1692: it is not hard to imagine the sobering effect of that first execution in the small community of Salem Village (now Danvers). But why the second wave of accusations? It turns out that the accusations in the first wave were directed against residents of Salem Village, Salem Town, and all but one of the half-dozen immediately adjacent towns; in the second wave the majority of the accusations were directed against residents of the one other adjacent town, Andover. Our sources [3, 4] do not provide much explanation for what happened in Andover, but the pattern, together with what we know of the context, tells at least part of a story and raises some interesting questions.

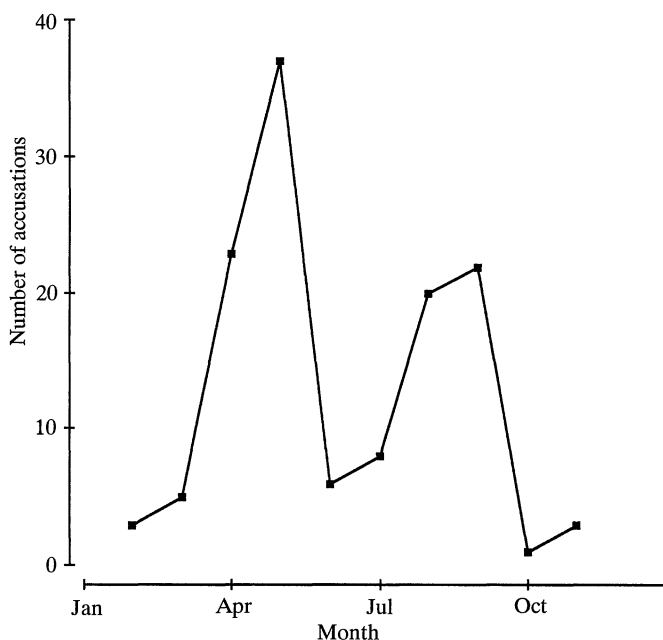


Figure 1. Numbers of people accused of witchcraft in Essex County, MA, 1692.

Although this first example has almost no mathematical content, its interplay between pattern and context is typical of the interpretive part of statistical thinking. For a more familiar example of a very different sort, consider testing that two normal distributions have equal means.

Example 2a. A model for comparing normal means. Consider the standard model involving two sets of independent, identically distributed (iid) random variables:

$$X_1, X_2, \dots, X_n \text{ iid } N(\mu_1, \sigma_1^2) \quad Y_1, Y_2, \dots, Y_m \text{ iid } N(\mu_2, \sigma_2^2)$$

It follows that $\bar{x} = (\sum x_i)/n$ and $s_1^2 = \sum(x_i - \bar{x})^2/(n - 1)$ are sufficient statistics for μ_1 and σ_1^2 , with parallel results for the Y s. Informally, a statistic is *sufficient* for a parameter if it uses all the information about that parameter contained in the sample. More formally, the conditional distribution of the data, given the sufficient statistic, doesn't depend on the parameter. The Rao-Blackwell Theorem guarantees that no unbiased estimator can have a smaller variance than one based on a sufficient statistic. Both \bar{x} and s_1^2 are unbiased: $E(\bar{x}) = \mu_1$ and $E(s_1^2) = \sigma_1^2$. Finally, their joint distribution is known: the sample mean \bar{x} is normal with variance σ_1^2/n , and, independently, $(n - 1)s_1^2/\sigma_1^2$ is chi-square on $(n - 1)$ degrees of freedom. Suppose now we want to test $H_0: \mu_1 = \mu_2$. If $\sigma_1^2 = \sigma_2^2$ then a sufficient and unbiased estimator for the common variance is obtained by pooling:

$$s_p^2 = [(n - 1)s_1^2 + (m - 1)s_2^2]/(n + m - 2)$$

If H_0 is true, then $(\bar{x} - \bar{y})/s\sqrt{(1/n) + (1/m)}$ has a Student's t -distribution on $n + m - 2$ degrees of freedom, and we can use the value of t computed from the data to test the null hypothesis. If t is far enough from 0, we conclude that $\mu_1 \neq \mu_2$.

This example differs most strikingly from the first in two ways: mathematical content and the role of context. Example 1, which has essentially no mathematical content, finds its intellectual substance almost entirely in the interplay between pattern and story. Example 2, which has essentially no content apart from mathematics, gets its intellectual substance without any explicit reference to applied context.

Although mathematicians often rely on applied context both for motivation and as a source of problems for research, the ultimate focus in mathematical thinking is on abstract patterns: the context is part of the irrelevant detail that must be boiled off over the flame of abstraction in order to reveal the previously hidden crystal of pure structure. *In mathematics, context obscures structure.* Like mathematicians, data analysts also look for patterns, but ultimately, in data analysis, whether the patterns have meaning, and whether they have any value, depends on how the threads of those patterns interweave with the complementary threads of the story line. *In data analysis, context provides meaning.*

The difference has profound implications for teaching. To teach statistics well, it is not enough to understand the mathematical theory; it is not even enough to understand also the additional, non-mathematical theory of statistics. One must, like a teacher of literature, have a ready supply of real illustrations, and know how to use them to involve students in the development of their critical judgment. In mathematics, where applied context is so much less important, improvised examples often work well, and teachers of mathematics become skillful at inventing examples on the spot. (Need a function to illustrate the chain rule? No problem: just make one up.) In statistics, however, improvised examples don't work, because they don't provide authentic interplay between pattern and context. Much as Bertrand Russell likened mathematics to sculpture for the austerity of its abstraction, one might think of data analysis as like poetry, where pattern and context are inseparable. Imagine yourself teaching a lesson on basic prosody, introducing dactylic hexameter. It is not enough to say "TA ta ta, TA ta ta, TA ta ta, . . . ;" your students need to hear dactyls in a real poem [20]: "This is the forest primeval. The murmuring pines and the hemlocks." In a similar spirit, the teacher of statistics needs to know the data literature. If, for example, when you teach plots for data distributions, you use data on inter-eruption times for Old Faithful [30] and lengths of reigns of English kings and queens [13], your students can learn more than just the methods themselves. The bimodal shape of the inter-eruption times suggests two kinds of eruptions, and the distribution of monarchs' reigns shows the skewness toward high values that is typical of waiting times.

The contrasting roles of context in mathematics and statistics, especially as illustrated in the deliberately extreme first two examples, might seem to lend support to the false implication in Bullock's [5] assertion that "Many statisticians now claim that their subject is something quite apart from mathematics, so that statistics courses do not require any preparation in mathematics." In fact, while we find the evidence that statistics is not mathematics persuasive (see [22], [24]), 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 those theories is part of the training of statisticians. But although statistics cannot prosper without mathematics, the converse fails. That statistics is not a necessary part of a mathematician's training is implicit in the statement by the eminent probabilist David Aldous [1] that he "is interested in the applications of probability to all scientific fields *except statistics.*"

What then, is the role of mathematics in the science of statistics? An answer should begin with a more systematic look at the logic of analyzing data.

1.2. A schematic overview of statistical analysis. An old-style course that wanted to be conscientious about applications might finish off the second example with a little coda of an exercise. The data, although not this invented exercise, are from [25]; the full study is described in [21].

Example 2b. Calcium and blood pressure. Does increasing the amount of calcium in our diet reduce blood pressure? The following numbers give the decrease after 12 weeks in systolic blood pressure for 21 human subjects. The 10 subjects in Group 1 took a calcium supplement for 12 weeks; the 11 in Group 2 took a placebo. Test the hypothesis that the calcium had no effect on blood pressure.

Group 1 (calcium): 7, -4, 18, 17, -3, -5, 1, 10, 11, -2
 Group 2 (placebo): -1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3

This exercise, put so tersely, is a caricature, one that encourages the mistaken view that once the mathematical derivations from a model are completed, applications are largely a matter of routine arithmetic. For a more realistic perspective, consider Figure 2, a diagram of the stages in a statistical analysis. Before considering this crude outline in detail, two cautions are essential.

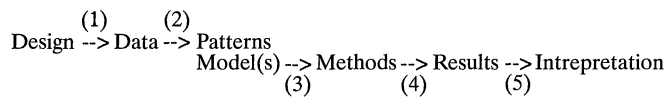


Figure 2. A schematic representation of the phases of data production and analysis.

1. The summary oversimplifies by suggesting a strict left-to-right progression. In reality, the process of data analysis is neither linear nor unidirectional. Several transitions involve a dialog of sorts, sometimes between adjacent elements, but sometimes among more than just two. Thus, for example, the choice of design for data production determines the structure of the resulting data, but knowledge based on data already in hand can help shape the design, as when knowing the size of variation from one subject to another helps decide how many subjects will be needed. Similarly, the data may suggest a model, but the model leads to methods that send us back to the data to check for possible violations of the model's assumptions. Perhaps most important of all, as we shall see, the final stage, interpretation of the results, depends in a crucial way on the first stage, the kind of design used for producing the data.
2. The rough and qualified ordering of stages here is not meant to suggest that we think the topics taught in an introductory statistics course should follow the same order. For reasons presented later, we recommend beginning with methods for exploring and describing data, then going "back" to data production, and from there to formal inference.

With these cautions assumed, the flowchart can provide a useful framework for examining the role of mathematics in statistics and summarizing elements of the

non-mathematical substance of the subject. Here are four quick observations:

1. Design, exploration, and interpretation are core elements of statistical thinking. All three elements are heavily dependent on context, but at the *introductory* level they involve very little mathematics. The (largely non-mathematical) theory of experimental design is decades old and well developed; the theory of exploration is newer, and at present still primitive, although computer-based tools for exploration have become quite sophisticated; the theory of interpretation is fragmentary at best.
2. The classical course in mathematical statistics corresponds so neatly to transition (3) that “from models to methods” might almost serve as a course title. Context is largely irrelevant here, because models are presented abstractly, as in Example 2a, and a typical derivation simply applies one optimality principle or another (least squares, maximum likelihood) to deduce the method *de jour*.
3. Transition (4), from methods to results, is the focus of the old-style cookbook course, in which each method is summarized by a set of formulas. Context is irrelevant here also, in that you can learn computational algorithms, and in fact learn them more efficiently, if you resist any temptation to encumber your brain with concern about what the methods are good for. All the same, some courses have tried to make the throat-clogging bolus of rote easier to get down by sugar-coating it with a thin glaze of ersatz context. Fortunately, the computer is fast sweeping courses like these into the dustbin of curricular history.
4. It is perhaps ironic that transitions (3) and (4), the two that have most often been the focus of courses at the introductory level, are precisely the two that are intellectually most automatic (given our current limited understanding and less developed theory of the other transitions) and so offer the least room for judgment and creativity.

To develop these points in more detail, we return to the example of calcium and blood pressure. In what follows, we combine the stages of Figure 2 under three broader headings: data production, data analysis, and inference.

2. THE CONTENT OF STATISTICS

2.1. Data production. The standard model of Example 2a is 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. This distinction, between observation and experiment, is one of the most important in statistics. Researchers often 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 calcium study was in fact an experiment:

Example 2c. The design of the calcium study [21]. 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.

Can we conclude that calcium has caused a reduction in blood pressure? Such an inference, that an observed difference may be taken at face value, stands on three legs. Two of the three are grounded in data production:

- (1) an argument—automatic only for random samples and randomized experiments—that a probability model applies to the data;
- (2) an argument—probability-based, and comparatively straightforward—that the observed difference is “real,” i.e., too big to be plausibly explained as due just to chance variation; and
- (3) an argument—often thorny and fraught with pitfalls, except in the case of randomized experiments—that the observed difference is not due to some confounding influence distinct from the factor of interest.

The *t*-test of Example 2a, like all statistical tests and confidence intervals, deals only with the second argument: “If we assume that a particular chance model applies, how likely is it to get an observed difference this big?” The other two arguments depend on the design.

The clinical trial on the effect of calcium on blood pressure was a *randomized comparative experiment*. Figure 3 presents the design in outline form. The great virtue of assigning the subjects at random is that it makes arguments (1) and (3) automatic, and so reduces the problem of inferring cause to checking the fit of a model, and then, given adequate fit, carrying out a straightforward calculation. 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.

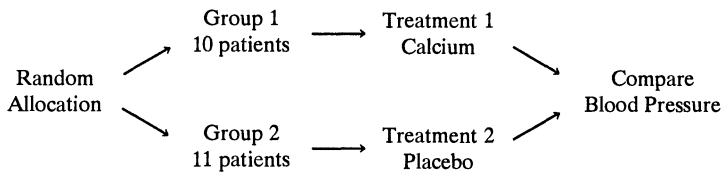


Figure 3. The simplest randomized comparative experiment.

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. However, unlike randomized experiments, observational studies do not lend themselves in so straightforward a way to an inference of causation, as the following example shows. The original study by Best and Walker appears as an example in [12]; our presentation here follows [26].

Example 3. Smoking and health. One of the early observational studies of smoking and health compared mortality rates for three groups of men. The rates, in deaths per year per 1000 men, were:

Non-smokers 20.2, Cigarette smokers 20.5, Cigar and pipe smokers 35.5.

To test whether the observed differences might be due to chance, we could use a model similar to the one in Example 2a. The sample sizes were so large that we can easily rule out chance variation as an explanation for the observed differences, leaving us with the apparent conclusion that cigarettes pose little risk but pipes or cigars or both are quite dangerous. Indeed, that conclusion would be valid *if* these data had come from a randomized, controlled double-blind experiment like the calcium study. However, the premise is clearly untenable. Because this is an observational study, we need to ask about other factors, linked to smoking habits, that might be responsible for the observed difference. Here, age is the main such factor: pipe and cigar smokers tend to be older than cigarette smokers, and the risk of death increases with age. In this study, the average ages for the three groups were:

Non-smokers 54.9 years, Cigarette smokers 50.5 years,
Cigar and pipe smokers 65.9 years.

Only after adjusting the death rates for the differences in age do we get numbers more in line with what we have come to expect:

Non-smokers 20.3, Cigarette smokers 28.3, Cigar and pipe smokers 21.2.

Taken together, the last two examples offer what we consider two of the most important lessons for mathematicians who teach statistics: one, the conclusions from a study depend crucially on how the data were produced, and two, the standard mathematical models ignore 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.

2.2. Data analysis: exploration and description. 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.

Table 1 presents an elementary summary [25] of the distinctions between EDA and standard inference:

TABLE 1. EXPLORATORY DATA ANALYSIS VS. FORMAL PROBABILITY-BASED INFERENCE

Exploratory Data Analysis	Statistical Inference
Purpose is unrestricted exploration of the data, searching for interesting patterns.	Purpose is to answer specific questions, posted 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. This is one reason why 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.

As we have already seen, the model of Example 2a, because it does not distinguish between observation and experiment, is incomplete. It is also, like most idealized mathematical models for real phenomena, 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 the setting and the 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 the assumptions of the model.

Example 2d. Preliminary exploration of the calcium data. An analysis might start from a simple outline: plot, shape, center, spread.

Plot. A stemplot splits each data value into a stem and leaf, then sorts leaves onto shared stems. Figure 4 shows a back-to-back stemplot useful for comparing two groups:

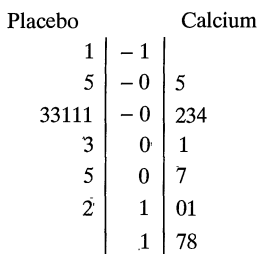


Figure 4. Parallel stemplot of reduction in systolic blood pressure for two groups of men.

Shape. The distribution for the placebo group is unimodal and symmetric. The treatment group, however, contains a faint suggestion of bimodality, which raises the possibility of two kinds of subjects. Might there be some who respond to calcium, and others who do not? There is no way to tell from these data, but the possibility is worth noting.

Center and spread. A useful plot for comparing centers, spreads and symmetries is the boxplot (Figure 5). Each box locates the quartiles and median of a distribution; the “whiskers” extend from the quartile to the most extreme points within 1.5 interquartile ranges of the nearest quartile, and points at a greater distance from the median are shown separately. Here we find a difference in medians, but also a pronounced difference in spreads, one that should raise suspicions about the assumption of equal variances used to justify a pooled estimate in Example 2a.

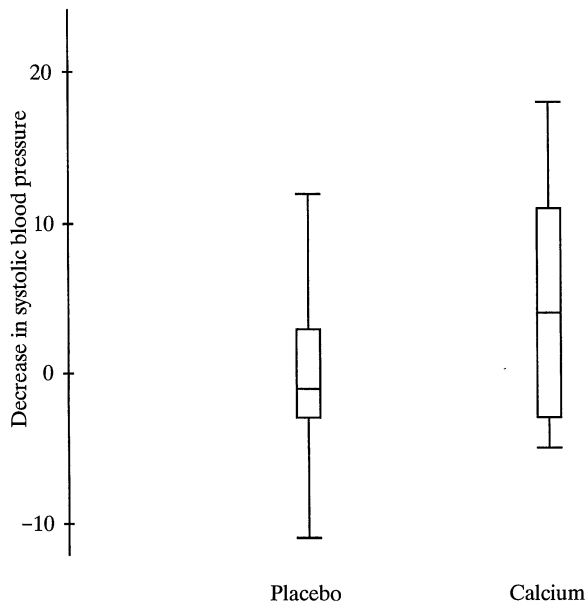


Figure 5. Parallel boxplots of reduction in systolic blood pressure for two groups of men.

Normal quantile plot. Looking ahead to a t -test to compare means, it is prudent to ask whether the data give us reason to question the normal model of Example 2a. Here we subtract the group mean from each observation to get residuals, then plot the ordered residuals against the corresponding quantiles of a normal distribution; see Figure 6. Our ordinates are the 21 ordered residuals, which divide the real line into 22 sub-intervals. The corresponding abscissas are the 21 values that divide the real line into 22 segments that are equiprobable under the normal model. If the data come from a single normal distribution, we can expect the points to fall near a line.

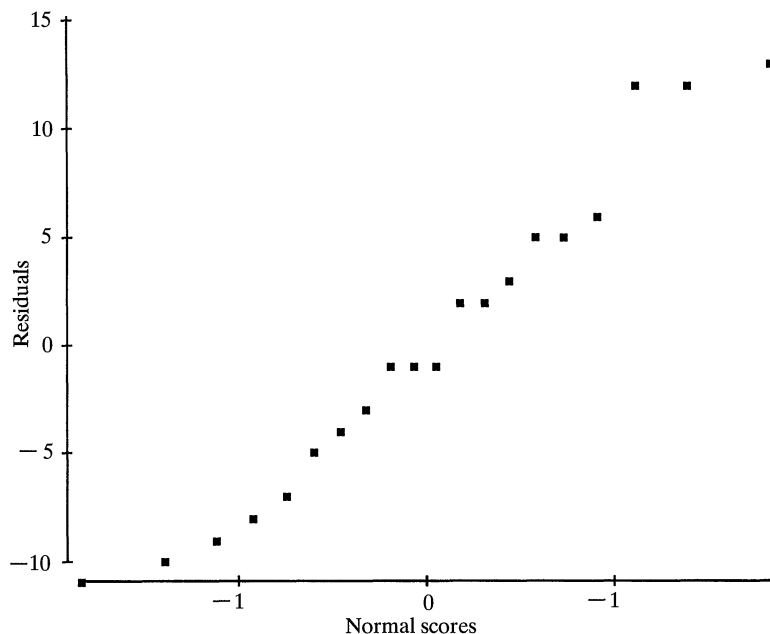


Figure 6. Normal quantile plot for the blood pressure data.

For the calcium data, the pattern is reasonably linear, although the vertical jump before the three right-most points shows observed residuals that are larger than predicted by the normal model, a pattern consistent with the unequal spreads in the boxplots.

Mathematically structured instruction, which tends to emphasize how methods follow from models, often provides only the most general warnings about the realities of practice. Statistics in practice resembles 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. In the calcium examples, the exploratory analysis warns us not to rely heavily on the assumption of equal variances, and to use a modified t -test that estimates separate variances for the two groups. 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.*

Wide availability of cheap computing, especially graphics, has combined with the desire to "let the data speak" to generate an abundance of new tools: at the low end we have the stemplots and boxplots of Example 2c; but there are also model-free scatterplot smoothers, resistant regression algorithms, clever ideas for display of high-dimensional data on two-dimensional screens, and many still more advanced diagnostic tools for specific situations. Standard statistical software implements much of this. The books [7] and [9], 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 [8] for detailed discussion of the latter.

Although it may be tempting for the neophyte to view data analysis as merely a collection of clever tools, the value of these tools comes from using them in a systematic way, according to 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.

It is part of the unifying (but non-mathematical) theory of EDA that these principles apply in each of several settings. Given data on a single quantitative variable, we might 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. Thus the univariate “*Plot, shape, center, spread,*” returns in the context of bivariate data as “*Plot, shape, direction, strength.*”

Here, as elsewhere, an analysis is not just a search for patterns, but a search for *meaningful* patterns. The *best* fit is not necessarily the *most useful*, as the following example illustrates.

Example 3. Dormitories and cities. Each point in Figure 7 represents one of the 50 U.S. states with horizontal coordinate equal to the state’s urban population, and

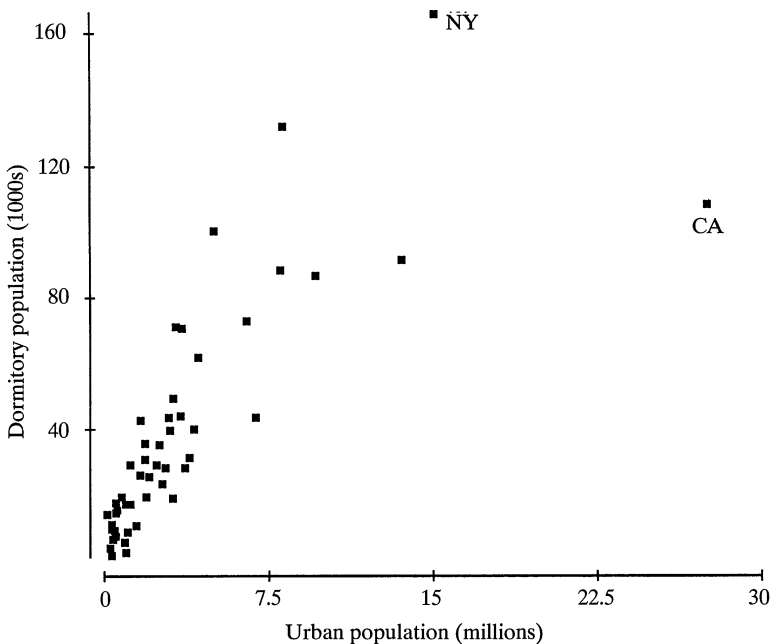


Figure 7. Scatterplot of dormitory population versus urban population for the 50 U.S. states.

vertical coordinate equal to the number of the state’s college students housed in dormitories. Several features of the plot’s shape stand out. For example, the plot is fan shaped, with many points bunched in the lower left: most states have relatively small urban populations (a couple of million or so) and relatively small dormitory populations as well (under 50,000); only a few states have very large urban populations or very large dormitory populations, and the variability from state to state is larger (more space between points) for the states with larger values. The pattern of association between the two variables is positive and strong: smaller urban populations go with smaller dormitory populations, larger urban populations with larger dormitory populations and, for all but a few of the states, knowing the size of a state’s urban population allows us to predict its dormitory population to within a fairly narrow range.

Despite the nice fit between picture and story, the analysis so far has overlooked a most important feature. If we take at face value the pattern that states with large urban populations also have large dormitory populations, we might be tempted to conclude that cities must attract colleges. Although plenty of confirming instances come to mind, this naive interpretation is wrong: both our variables are indirect measures of the size of the states’ populations, so it is hardly surprising that the two measures show a strong positive association. To uncover a more meaningful relationship, we have to “adjust for the lurking variable:” divide urban population by total population to get percent urban, divide dormitory population by total population to get percent living in dormitories, and plot the result (Figure 8).

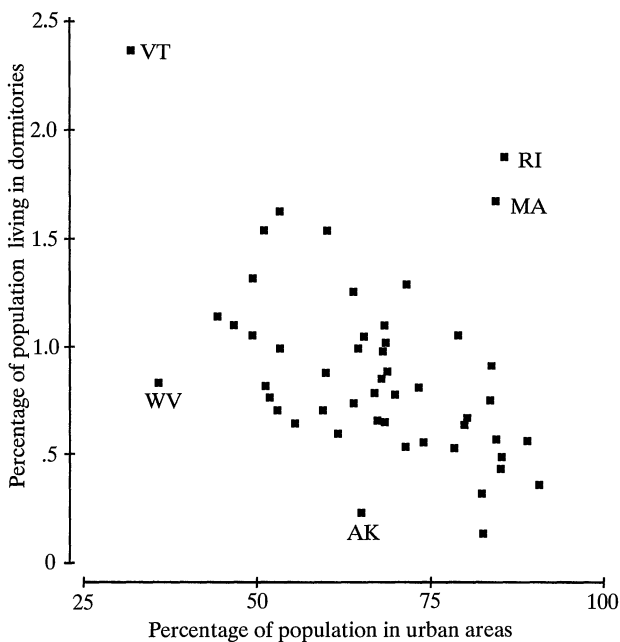


Figure 8. Scatterplot of the dorms-and-cities data after adjusting for the “lurking variable” population.

Now the relationship is weaker, but what it tells us is more interesting. The direction is reversed: rural states—those with a lower percentage of their residents living in metropolitan areas—have a higher percentage of their residents living

in college dormitories. On reflection, this makes sense. Think about Pullman, Washington, or Ames, Iowa; about Norman, Oklahoma, or Lawrence, Kansas. Rural states may have fewer colleges and universities in absolute numbers, but their students make up a higher percentage of the total population of the state, and are more likely to live in dormitories.

2.3. Formal inference: the argument against chance. 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 out 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 “estimate \pm 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 rarely occur if the effect we seek were absent is good evidence that the effect is in fact present. Figure 9 illustrates this reasoning in our medical example. The normal curves in that figure represent the sampling distribution of the difference $\bar{x} - \bar{y}$ 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,

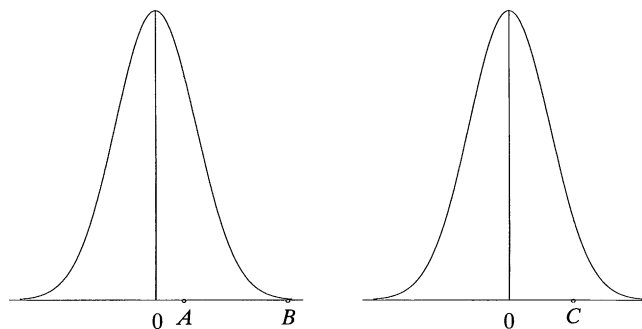


Figure 9. The idea of statistical significance: is this observation surprising?

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 $\bar{x} - \bar{y} = 5.273$. 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: $p = 0.072$.

3. TEACHING. 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. This section, in keeping with the rest of this article, concerns content, and in particular contains one side of a conversation between statisticians and mathematicians who may find themselves teaching statistics.

3.1. Statistics should be taught as statistics. 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. 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 [10]; a longer discussion appears in [11]:

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 [14] describe some of this work for non-specialists.

3.2. Neither Mathematics Nor Magic. An over-emphasis on probability-based inference is one mark of an overly mathematical introduction to statistics, and yet 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. Students understand mathematics when they appreciate 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. How can we best lead our students to understand, appreciate, and begin to assimilate this intellectual method?

3.3. Begin with exploratory data analysis. Although the implied chronology of Figure 2 suggests starting with data production, experience says otherwise. For one thing, exploratory data analysis makes a better beginning because it is more concrete. There is no need to distinguish population and sample, and no need to discuss the features of randomization that protect against bias. Basic methods are conceptually and algorithmically simple, and the data are in hand—actual numbers on a page, as opposed to mere ghosts of data-in-the-future, the way they are in designing an experiment. Moreover, providing motivation is not a problem. Students like exploratory analysis and find that they can do it, a substantial bonus when teaching a subject feared by many. Engaging them early on in the interpretation of results, before the harder ideas come along to claim their attention, can

help establish good habits that pay dividends when you get to inference. Finally, starting with data analysis prepares for design and for inference. Experience with data distributions introduces students to the omnipresence of variability, and to the potential for bias, the two main reasons we need careful design. If you teach design before data analysis, it is harder for students to understand why design matters. Experience with data distributions is also the best way to get ready to tackle the difficult idea of a sampling distribution.

We have tried to suggest that there is a coherent (though not mathematical) set of ideas and associated tools for exploring data. Students need to practice these ideas and tools by writing coherent descriptions of data. To help them, we provide both outlines for what to write, and examples that can serve as models. Figure 10, for example, is the outline for describing 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 10. Outline for describing data on a single quantitative variable.

Following this outline requires both knowledge of the tools it mentions 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.

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 11) and a normal quantile plot (Figure 12) 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.

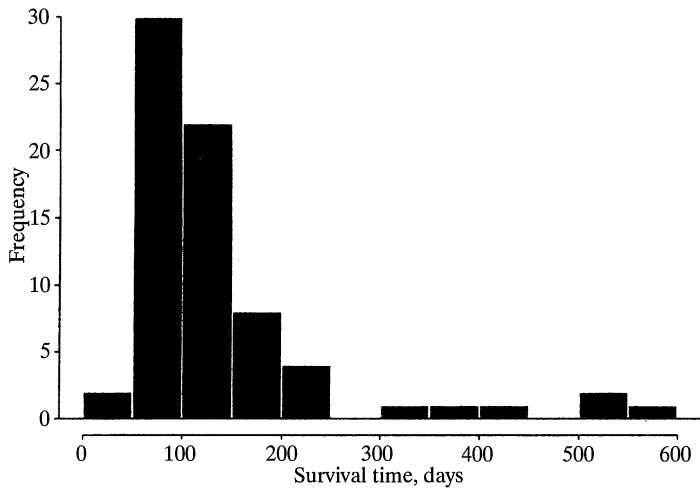


Figure 11. Histogram of guinea pig survival times.

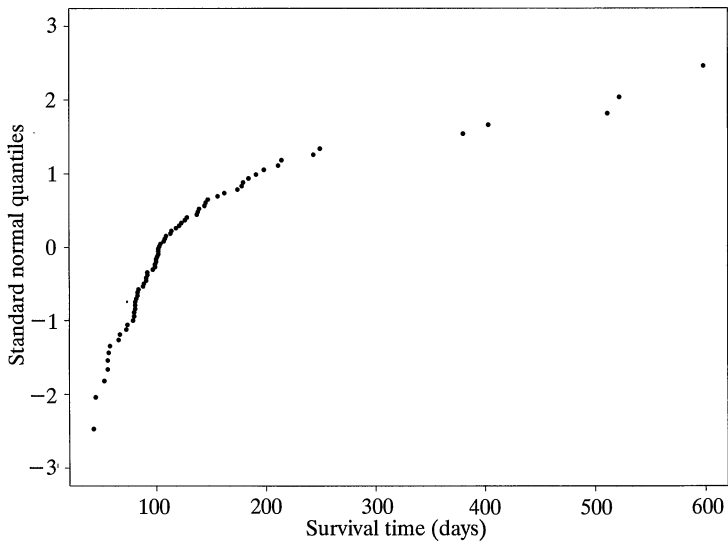


Figure 12. Normal quantile plot for guinea pig survival times.

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 ($\bar{x} = 141.8$ 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.

Although many students come to a first statistics course expecting empty ritual, EDA offers them the pleasant surprise that the methods exist to serve

the search for meaning. This surprise is so welcome that it carries a danger of pushing the pendulum too far the other way. Some students may drift into a complacent conviction that any story about the data that fits the patterns with coherence and plausibility must be true. The timing is right for a dose of design and skepticism.

3.4. Teach design as the bridge between data analysis and inference. An introduction to design for data production fits naturally between exploratory analysis and inference: sound design is what makes inference possible. Waiting to introduce probability distributions until after the basics of design has a number of advantages. For one thing, this order helps make clear that the justification for probability models must come from the randomness in the data production process, and so provides some protection against unthinking adoption of probability models. For another, learning about data production introduces students to essential concepts like population and sample, parameter and statistic, before they encounter the sampling distribution, which is conceptually difficult all by itself.

The single most important point for students to understand is why randomized comparative experiments are the gold standard for evidence of causation. A rich source of true-life cautionary tales is the book [6], edited by the physicians Bunker and Barnes and the statistician Mosteller, which 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.

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 your judgment of the needs of your audience.

3.5. Inference: two barriers to understanding. Section 2.3 has described briefly how inference works. Because the details are in practice automated, we would like students to put most of their effort into grasping the 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. We’d like most of our students to understand the idea of a sampling distribution; we know that quite a few won’t understand the reasoning of significance tests. Our 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 are 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 the 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, \dots, X_n iid $N(\mu_1, \sigma_1)$, Y_1, Y_2, \dots, Y_m iid $N(\mu_2, \sigma_2)$. 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. We 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 this paradigm. This is a (strongly held) minority position. Deep water ahead.

3.6. 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. Here, however, we are discussing introductory statistics rather than mathematics.

From the point of view of deductive logic that has shaped so much of statistical teaching in the past, probability is more basic than statistics: probability provides the chance models that describe the variability in observed data. From the point of view of the development of understanding, however, we believe that statistics is more basic than probability: whereas variability in data can be perceived directly, chance models can be perceived only after we have constructed them in our own minds. In the ideal Platonic world of mathematics, we can start with a probabilistic chicken and use deductive logic to lay a statistical egg, but in the messier world of empirical science, we must start with the egg as observed data and construct a prior probabilistic chicken as an inference. In an introductory *statistics* course, the chicken's only value is to explain where eggs come from. It seems a bit unfair, in that context, at least, to ask beginning students to learn about egg-generators before they've become familiar with eggs—less extreme, but in the same spirit as starting the study of chemistry with quantum mechanics.

What then, should be the place of probability in beginning instruction in statistics? Our 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*. Although the theoretical structure of standard statistical inference is based on probability, the role of probability is limited to answering the question “What would happen if we used this method very many times?” The answer is given by 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 $P(A \cup B)$ 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 [16] and [27]) 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., [28] and the collection [19]. What is worse, mathematics educators have found no effective way to correct our defective intuition. Garfield and Ahlgren [15] 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.” We 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 ours 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.

3.7. What About Mathematics Majors? Mathematics majors traditionally meet statistics as the second course in a year-long sequence devoted to probability and statistical theory. We hope it is clear that we 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 do?

Our 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 would ordinarily and most easily 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 we are 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 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 [29]. 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 \bar{x} for predicting y from x . Derive it. Then explore, perhaps using software, analogs of the other measures mentioned in A.

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, *Amer. Math. Monthly* **101**, 223–233. The remark appears in the biographical note accompanying the paper.
 2. Bjerkedal, T. (1960), Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, *American Journal of Hygiene* **72**, 130–148.
 3. Boyer, Paul and Stephen Nissenbaum (1972). *Salem Village Witchcraft*. Belmont, CA: Wadsworth Publishing Co.
 4. Boyer, Paul and Stephen Nissenbaum (1974). *Salem Possessed*. Cambridge, MA: Harvard University Press.

5. Bullock, James O. (1994), Literacy in the language of mathematics, *Amer. Math. Monthly* **101**, 735–743.
6. Bunker, John P., Benjamin A. Barnes, and Frederick Mosteller (eds.) (1977), *Costs, Risks and Benefits of Surgery*. New York: Oxford University Press.
7. Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey (1983), *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.
8. Chambers, John M. and Trevor J. Hastie (1992), *Statistical Model in S*. Pacific Grove, CA: Wadsworth.
9. Cleveland, William S. and Mary E. McGill (eds.) (1988), *Dynamic Graphics for Statistics*. Belmont, CA: Wadsworth.
10. Cobb, George W. (1991), Teaching statistics: more data, less lecturing, *Amstat News*, December 1991, pp. 1, 4.
11. Cobb, George W. (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.
12. Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics* **24**, 205–213.
13. Crystal, David (ed.) (1994), *The Cambridge Factfinder*. Cambridge: Cambridge University Press, pp. 174–175.
14. Efron, Bradley and Rob Tibshirani (1991), Statistical data analysis in the computer age, *Science* **253**, 390–395.
15. 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**, 44–63.
16. Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krüger (1989) *The Empire of Chance*. Cambridge: Cambridge University Press.
17. Hoaglin, D. C. (1992), Diagnostics, in D. C. Hoaglin and D. S. Moore (eds.), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America, pp. 123–144.
18. Hoaglin, David C. and David S. Moore (eds.) (1992), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America.
19. Kapadia, R. and M. Borovcnik (eds.) (1991), *Chance Encounters: Probability in Education*. Dordrecht: Kluwer.
20. Longfellow, Henry Wadsworth (1847), *Evangeline*, Introduction, 1.1.
21. 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**, 1772–1776. Dr. Lyle provided the data in the example.
22. Moore, David S. (1988), Should mathematicians teach statistics (with discussion), *College Math. Journal* **19**, 3–7.
23. Moore, David S. (1992), What is statistics? in David C. Hoaglin and David S. Moore (eds.), *Perspectives on Contemporary Statistics*, MAA Notes 21. Washington, DC: Mathematical Association of America, pp. 1–18.
24. 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.
25. Moore, David S. (1995), *The Basic Practice of Statistics*. New York: W. H. Freeman.
26. Rosenbaum, Paul R. (1995), *Observational Studies*. New York: Springer-Verlag, p. 60.
27. Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass: Belknap.
28. Tversky, Amos and Daniel Kahneman (1983), Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment, *Psychological Review* **90**, 293–315.
29. Watson, G. S. (1994), letter to the editor, *The American Statistician* **48**, p. 269. This is the last in a sequence of comments on this inequality, and contains references to the earlier contributions.
30. Weisberg, Sanford (1985). *Applied Linear Regression*, 2nd edition. New York: John Wiley and Sons, p. 230.

Department of Mathematics, Statistics
and Computer Science
Mount Holyoke College
South Hadley, MA 01075
gcobb@mtholyoke.edu

Department of Statistics
Purdue University
West Lafayette, IN 47907
dsm@stat.purdue.edu