THE INTRODUCTORY STATISTICS COURSE:
A SABER TOOTH CURRICULUM?

George W. Cobb Mount Holyoke College

After dinner talk, USCOTS
Columbus, Ohio May 20, 2005

INTRODUCTION

In rural North Carolina, where I grew up, there was a story about an old mountaineer and his one-string cello. Every evening after supper, he'd haul his cello out onto the porch and sit there overlooking his valley, with one finger clamped down on the lone string, while he contentedly bowed back and forth, back and forth, back and forth. One day his wife, a woman of immense patience, offered a tactful suggestion: "I've noticed," she said, "that when other people play the cello, they use all four strings, and they also move their fingers up and down the strings while they play." "I know," the mountaineer replied. "*They* are *looking* for the place. *I* have *found* it."

I'm here this evening to bedevil you with my worry that the curriculum of the introductory statistics course is a bit too much like the mountaineer's cello playing. Those of us who teach the beginning course have settled on a curriculum that we embrace with a remarkable show of consensus. We act too much like *we* have found *the* place. As Dick Scheaffer pointed out a few years back, "With regard to the content of an introductory statistics course, statisticians are in closer agreement today than at any previous time in my career." [1]

I'm quoting Dick here not just because I agree with him about the remarkable consensus, but also as a segue: I want to use this occasion to acknowledge his role in many deep and important changes in statistics education that we have all benefited from. I'm going to talk this evening mainly about curriculum, and Dick has made many contributions in this area, but two of his important contributions to the reform of statistics education are not really about curriculum. One is about pedagogy, about *how* we teach rather than *what* we teach. Dick has led, by his example, in showing all of us that using activities and getting students to experience statistics as something you actually *do* can make a major difference in whether they enjoy the course and in how much they learn and take away from the course. Over the years this lesson from Dick has had a huge impact on what I do in my classes, and a huge impact on how my students react to my teaching.

A second contribution of Dick's is less recognized explicitly, but will be instantly recognizable to anyone who has ever attended one of his workshops or worked with him on a project. Dick's leadership is of a stealth variety: his radicalism tends to fly beneath your radar. He is invariably unassuming and friendly, never preachy. He presents radical ideas in the modest spirit of "Here's something neat that you might want to try." I'm convinced that this style of Dick's has had a lot to do with why it is that so many good

[1] Scheaffer, Richard, discussion of Moore, D.S. (1997). "New pedagogy and new content: the case of statistics," *International Statistical Review*, vol. 65, No. 2, pp. 123 – 165.

and important changes have been made in the way we teach statistics, without any of the divisiveness or rancor that has sometimes been part of the reform of introductory calculus, a rancor that has much too often been part of the attempts to reform the K12 mathematics curriculum. For these reasons, I'd like to take this occasion to thank Dick for all he's taught me, and all he's done for statistics education. Please join me …

Now, before I launch my attack on our consensus curriculum, let's take a closer look at some of what's *good* about it, about some of the wonderful big-picture consequences of our reform efforts. For a concise summary of where we are at this point, I urge you to read the report of Joan Garfield's GAISE project, and Chris Franklin's parallel report for K-12.

In a nutshell, thanks to Dick Scheaffer, David Moore, Ann Watkins, and many others, statistics is now taught as the *science of data*. Statistics is recognized as a subject in its own right – the study of producing and analyzing data. Most mathematicians still regard statistics as a subject they don't *want* to teach, but it is no longer the case that they regard statistics as a subject they *could* teach … with their brains tied behind their backs. When it comes to statistics, of course, "behind their backs" is precisely the place where many mathematicians could be accused of keeping their brains for a very long time. To be fair, the mathematicians' contempt for the old style statistics course was all too often entirely justified. Be that as it may, the change in the mathematicians' attitude toward statistics is, I believe, a mark of the success of the reform of our curriculum. We have ascended, evolutionarily speaking … from taking out the intellectual garbage – anyone *could* do it, but who would ever want to? – all the way up the ladder to automotive repair – there may be a lot of worldly grime under our intellectual fingernails, but what we do is at least acknowledged to be of practical importance, and recognized, also, as a job that takes brains to do well. The statistician is no longer regarded by mathematicians as the deformed little Mr. Hyde that Dr. Jekyll turned into because he thought he wasn't good enough to pass the comprehensive exams in mathematics.

So far, so good. But not far enough, and not good enough. We're at a place where we could easily get stuck in a rut, complacently bowing back and forth on our cello, convinced that although the mathematicians may be still searching, when it comes to the beginning statistics course, *we* have found the *place*. By way of warning, here's another story from my rural upbringing.

Back in the first half of the last century, after rural electrification had come to the Tennessee valley, local officials got a call from a man whose house had been recently wired. He wanted to know how to keep from burning his fingers. It seems that before electrification, he had used a candle for light, and he would take his single candle with him as he went from one room to the next. After his house went on the new grid, he had bought only a single light bulb, and was using it the way he had used his candle, patiently unscrewing it and taking it with him from one room to the next. No wonder he burnt his fingers!

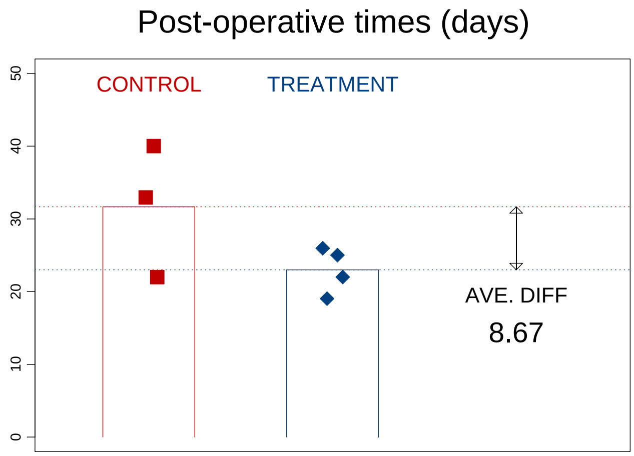
There's a similar story about the farmer who bought his first tractor. He hated it, because it took him so much longer to get the plowing done. Back when his team of mules had to pull only the plow, he could get a whole field done in one day, but now that his mules had to pull the *tractor*, it took three times as long.

I don't suggest that either story is literally true, but I want to regard them as allegories with lessons about the introductory statistics curriculum. Computers have brought us opportunities for change as potentially revolutionary as the opportunities brought by rural electrification and the invention of the internal combustion engine, but to a greater extent than we realize, our curriculum is still mulishly pulling a tractor behind it, and our students are still going from room to room with a single light bulb.

Ready to burn your fingers? I'm going to describe a small data set. As I describe it, think about how the students in your introductory statistics course would analyze the data.

Example: Here are post-surgery recovery times in days, for seven patients who were randomly divided into a control group of three that received standard care, and a treatment group of four that received a new kind of care.[2]

	Times (days)				Mean	SD
Control (standard):	22	33	40		31.67	3.0
Treatment (new):	19	22	25	26	23.00	9.1



How would the students from your introductory course, and mine, analyze this data set? A two-sample *t*-test? Assuming unequal SDs, and so using the Welch-adjusted *t*-statistic? Let's examine that answer.

[2] Ernst, Michael D. (2004). "Permutation methods: A Basis for Exact Inference," *Statistical Science*, vol. 19, pp.676-685.

In the old language, what *assumptions* are we making in choosing this analysis? In the newer and better language that we owe to Jeff Witmer of Oberlin, what *conditions* must be satisfied in order to justify this analysis?

Conditions:

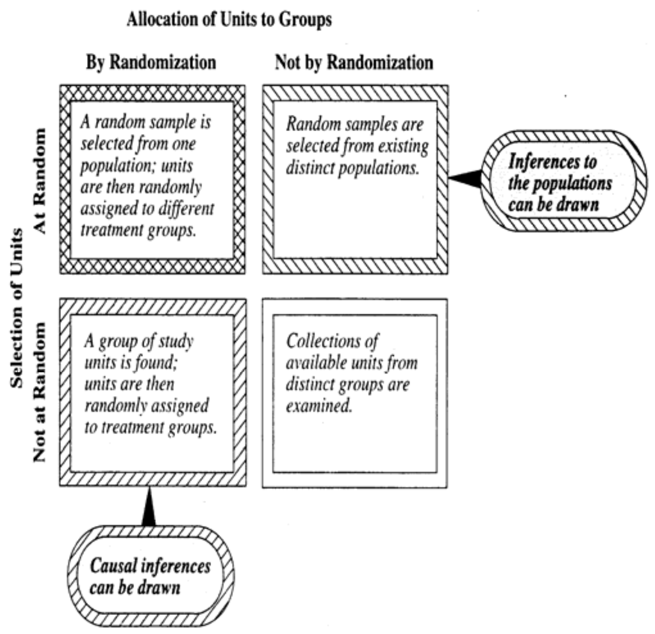
Data constitute two independent simple random samples from normal populations

Reality:

Data constitute a single convenience sample, not necessarily normal, randomly divided into two groups.

How many of us devote much time in our class to the difference between model and reality here? One of the most important things our students should take away from an introductory course is the habit of always asking, “Where was the randomization, and what inferences does it support?” How many of us ask our students, with each new data set, to distinguish between random sampling and random assignment, and between the corresponding differences in the kinds of inferences supported – from sample to population if samples have been randomized, and from association to causation if assignment to treatment or control has been randomized? (See display.)[3]

Display 1.5 Statistical inferences permitted by study designs


Allocation of Units to Groups
By Randomization
Not by Randomization
Selection of Units
At Random
Not at Random
A random sample is selected from one population; units are then randomly assigned to different treatment groups.
Random samples are selected from existing distinct populations.
Inferences to the populations can be drawn
A group of study units is found; units are then randomly assigned to treatment groups.
Collections of available units from distinct groups are examined.
Causal inferences can be drawn

An alternative to the *t*-test for this data set is the **permutation test**. Here’s how it works for the example. Start with the null hypothesis that the treatment has no effect. Then the

[3] The chart is from Fred L. Ramsey and Daniel W. Schafer (2002). *The Statistical Sleuth*. Pacific Grove, CA: Duxbury, p. 9.

seven numerical values have nothing to do with treatment or control. The only thing that determines which values get assigned to the treatment group is the randomization. The beauty of the randomization is that we can repeat it, over and over again, to see what sort of results are typical, and what should be considered unusual. This allows us to ask, “Is the observed mean difference of 8.67 too extreme to have occurred just by chance?” To answer this question, put each of the seven observed values on a card. Shuffle the seven cards, and deal out three at random to represent a simulated control group, leaving the other four as the simulated treatment group. Compute the difference of the means, and record whether it is at least 8.67. Reshuffle, redeal, and recompute. Do this over and over, and find the proportion of the times that you get an average of 8.67 or more. This turns out to be 8.6% of the time: unusual, but not quite unusual enough to call significant.

Notice that because the set of observed values is taken as given; there is no need for any assumption about the distribution that generated them. Normality doesn’t matter. We don’t need to worry about whether SDs are equal either; in fact, we don’t need SDs at all. Nor do we need a *t*-distribution, with or without artificial degrees of freedom. The model simply specifies that the observed values were randomly divided into two groups. Thus there is a very close match between the model and what actually happened to produce the data. It is easy for students to follow the sequence of links, from data production, to model, to inference.

> Question: Why, then, is the *t*-test the centerpiece of the introductory statistics curriculum?
> Answer: The *t*-test is what scientists and social scientists use most often.
>
> Question: Why does everyone use the *t*-test?
> Answer: Because it’s the centerpiece of the introductory statistics curriculum.

So why *does* our curriculum teach students to do a *t*-test? What are the costs of doing things this way? What could we be doing instead?

My thesis this evening is that both the content and the structure of our introductory curriculum are shaped by old history, or, one might even say, by old Europe. What we teach was developed a little at a time, for reasons that had a lot to do with the need to use available theory to handle problems that were essentially computational. Almost one hundred years after Student published his 1908 paper on the *t*-test, we are still using 19th century analytical methods to solve what is essentially a technical problem – computing a *p*-value or a 95% margin of error. Intellectually, we are asking our students to do the equivalent of working with one of those old 30-pound Burroughs electric calculators with the rows of little wheels that clicked and spun as they churned out sums of squares.

Now that we have computers, I think a large chunk of what is in the introductory statistics course should go. In what follows, I’ll first review the content of a typical introductory course, offering Ptolemy’s geocentric view of the universe as a parallel. Ptolemy’s cosmology was way too complicated, because he put the earth at the center of his system, instead of putting the sun at the center. Our curriculum is way too complicated because we put the normal distribution, as an approximate sampling distribution for the mean, at the center of our curriculum, instead of putting the core logic of inference at the center.

After reviewing our consensus curriculum, I'll give three reasons why this curriculum hurts our students and our profession: it's confusing, it takes up time that could be better spent on other things, and it even carries a whiff of the fraudulent. If it's all that bad, you may ask, how did we ever paint ourselves into the corner of teaching it? Until recently, I suggest, we had little choice, because our options were limited by what we could compute. I'll offer a handful of historically based speculations regarding the tyranny of the computable.

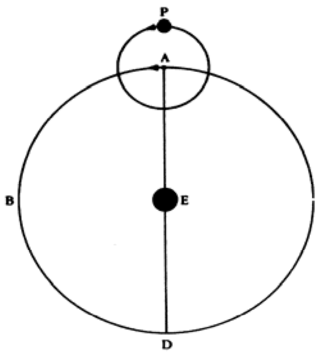
After that, I'll sketch an alternative approach to the beginning curriculum, and conclude with an unabashed sales pitch. Here, then, is the plan:

A. WHAT we teach: our Ptolemaic curriculum
B. WHAT'S WRONG with what we teach: three reasons
C. WHY we teach it anyway: the tyranny of the computable
D. WHAT SHOULD we teach instead: putting inference at the center
E. WHY SHOULD we teach it: an unabashed sales pitch

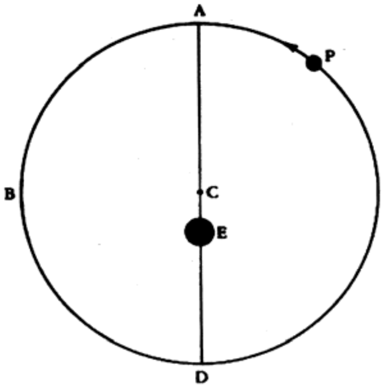
A. WHAT WE TEACH: OUR PTOLEMAIC CURRICULUM

I actually know very little about the history of cosmology, so you should regard what I'm about to tell you as more of a metaphor than a history lesson. Ptolemy's description of the universe is a useful example of a model that started from a very simple beginning, but which didn't quite fit the facts, and so got extended and modified and patched and tweaked until it had become quite complicated. Ptolemy's system began with two very simple ideas due to Aristotle: one, that the sun and planets travel at constant speeds in circular orbits, and two, that these orbits all have the same center, namely, the earth. Unfortunately, this model didn't quite fit. In particular, it could not account for retrograde motion, the apparent backward motion of certain planets at those times when they are closest to the earth. So Ptolemy added *epicycles*, little circles that revolved around the original big circles:[4]

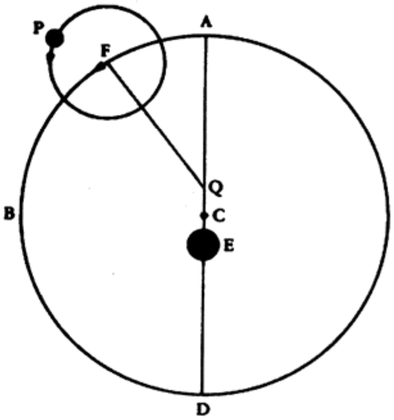
[4] Figures are from The Galileo Project, http://galileo.rice.edu/sci/theories/ptolemaic_system.html.



Now each planet was assumed to travel around its little circle as that little circle traveled around the big one. This new model could account for the retrograde motion, but, unfortunately, it still didn't quite fit all the observable data. Planets didn't move at constant speeds relative to the earth. To explain non-uniform motion, Ptolemy introduced the notion of the *eccentric*. The earth was no longer at the center of the big circle, but merely nearby. The center of the circle was at a different point.



And then came other adjustments:



There's actually a lot more to the model than I've told you, because the modifications I've described turned out not to be adequate, either, and so Ptolemy ended up with epicycles on epicycles, at least eighty epicycles in all, but I think you have enough to get the idea. With this as background, let’s now take a look at the core of the typical introductory curriculum:

Ptolemy started with an idealized construct, the circle, and a simple idea, that the planets and sun had circular orbits centered at the earth. The inference part of our curriculum starts with an idealized construct, the normal curve, and a simple idea, that most distributions we care about can be related to the normal, at least approximately. Just as circles can have different centers and different spreads, normal distributions also have different centers, and different spreads, and so our students have to learn about expected values and standard errors for the sample mean and sample proportion. They also learn that, according to the central limit theorem, the sampling distribution of the sample mean is roughly normal. Thus the center of our curricular model is that “*x*-bar is roughly normal, with mean μ and standard error σ over the square root of *n*.” *Note that this is a probability theorem, not a statistical idea*. In our current curriculum, this is where we have to get to before students can construct tests and confidence intervals for the mean.

But immediately, we encounter a problem. We have to *estimate* the standard deviation from the sample, and once we substitute *s* for σ, we no longer have a normal distribution. So Student and Fisher added an epicycle, and brought us the *t*-distribution: put *s* for σ; put *t* for 2 . This gives intervals and tests based on the *t*-statistic and the *t*-distribution. But we really want to be able to compare *two* groups, *two* means. So we need the expected value of the difference of two sample means, and the standard deviation of the difference as well. So maybe we now go to the two-sample *t* with pooled estimate of a common variance. At least that *t*-statistic still has a *t*-distribution. But of course it’s not what we really want our students to be using, because it can give the wrong answer when the populations have different standard deviations. So we introduce the variant of the *t*-statistic that has an un-pooled estimate of standard error. But this *t*-statistic no longer has a *t*-distribution, so we have to use an approximation based on a *t*-distribution with different degrees of freedom.

There’s more, of course. When we work with proportions instead of means, we estimate the standard deviation, but we continue to use *z* instead of *t* for confidence intervals. However, some of us now add Alan Agresti’s artificial pair of successes and artificial pair of failures to the actual data in order to get a better approximation. We may not have a total of eighty epicycles, but I think you get the general idea.

To go back to the light bulb metaphor, we ask our students to learn the one-sample *z*, then unscrew that light bulb, and take it with them, pulling a John Deere behind them, of course, to the next room, where they learn the one-sample *t*. Then they unscrew the bulb again and go to a new room for the two-sample *t* with pooled SD, then on into the next room for the two-sample *t* with unequal variances. Then it’s upstairs for counted data: one proportion, two proportions, the plus-four adjustment, and maybe chi-square.

B. WHY IT'S WRONG: obfuscation, opportunity cost, and fraud.

Notice how little of any of this deals directly with the core ideas of inference! Randomized data production? That was *chapters* ago. The connection between randomized data production and a sampling distribution? Or even just the idea of what a sampling distribution *is* -- in David Moore's words, the answer to the question, "What will happen if I repeat this many times?" That, too, has gotten buried under an avalanche of technical details. In a typical book, at least a third of the pages will be devoted to this stuff. This distribution-centered approach to statistics dates from a time when direct simulation was too slow to be practical. (The distribution theory based on the central limit theorem is also, I suspect, one of the few parts of our curriculum that actually *appeals* to mathematicians.)

What are the costs of doing things the way we do? I see three groups of reasons: obfuscation, opportunity cost, and fraud.

First, consider **obfuscation**. A huge chunk of the introductory course, at least a third, and often much more, is devoted to teaching the students sampling distributions, building up to the sampling distributions for the difference of two means and the difference of two proportions. Why is this bad? It all depends on your goal. The sampling distribution is an important idea, as is the fact that the distribution of the mean converges to a normal Both have their place in the curriculum. But if your goal is to teach the logic of inference in a first statistics course, the current treatment of these topics is an intellectual albatross. Sampling distributions are *conceptually difficult*; sampling distributions are *technically complicated*; and sampling distributions are *remote from the logic of inference*. The idea of a sampling distribution is inherently hard for students, in the same way that the idea of a derivative is hard. Both require turning a *process* into a mathematical *object*. Students can find the slope of a tangent line, and understand the process of taking a limit at a single point, but the transition from there to the derivative as a *function*, each of whose values comes from the limiting process, is a hard transition to make. Similarly, students can understand the *process* of drawing a single random sample and computing a summary number like a mean. But the transition from there to the sampling distribution as the probability distribution each of whose outcomes corresponds to the process of taking-a-sample-and-computing-a-summary-number is again a hard transition to make. Just at the point where we try to introduce this very hard idea to students, we also burden them with the most technically complicated part of the course as well. Many brains get left behind, because by now the tractor has become too heavy to pull any more. A course that may have started well with exploratory data analysis and continued well with ideas of sampling and experimental design has now lost not only momentum but also a sense of coherence. There's a vital logical connection between randomized data production and inference, but it gets smothered by the heavy, sopping wet blanket of normal-based approximations to sampling distributions.

Now consider the **opportunity cost.** Does anyone recognize what this is?

According to my former brother-in-law, who lives in Austin, that's a picture of a Texan after he's had the … (I'll use a polite circumlocution) … pompous vacuity kicked out of him. I offer it as a metaphor for what our course might look like if we purge it of all the unnecessary distribution theory. What's missing represents the part of our course time that would become available for us to reclaim and use for other purposes if we were to re-center our curriculum.

Enrollments in introductory statistics courses have skyrocketed in recent years, as more and more students recognize that they need to know how to deal with data. It's easy for us to be optimistic about the future of our subject, and to keep on bowing complacently at our cellos. But consider. Just as one example, I've been working currently with a molecular immunologist and her honors student doing research with microarrays. The methods they are using are at a cutting edge of our subject, developed by people like Brad Efron and Terry Speed and Rob Tibshirani and their colleagues. The methods are computer-intensive, but they have proved to be conceptually accessible – just barely -- to the student of my immunologist colleague. That student would have liked to take a statistics course to help her understand the papers by Efron and Speed and Tibshirani, but *none of the traditional courses would have been of much value.* Another example: Many of our computer science students are studying robotics, or image processing, or other applications that use Bayesian methods. They, too, would like a statistics course that would help them understand these topics, but the standard introductory curriculum would not offer much of a connection to the statistical topics they want to learn.

As statisticians we are used to going into a new area and learning quickly enough of what we need to learn in order to work successfully in a particular applied area. That's what I'm trying to do in connection with the microarray data. I checked my college catalog, and here's what I would nominally have to take just to get within range of the content that I'm teaching myself:

- Chem 101 – General chemistry I
- Chem 201 – General chemistry II
- Chem 202 – Organic chemistry I
- Biol 150 – Intro Biol I: form & function
- Biol 200 – Intro Biol II: org. development
- Biol. 210 – Genetics & molecular biology
- Biol 340 – Eukaryotic molecular genetics

We don't expect ourselves to endure this long, slow slog, and we shouldn't expect the same of our students. More and more, they'll be learning their statistics in an ecology course, or a genetics course, or a data mining course taught in a computer science department, or as part of a lab project, unless we change our courses to make them more responsive to student needs.

Third, consider **fraud**. It's a strong word, and I admit to using it mainly to get your attention, but consider the consequences of teaching the usual sampling model. *Either you limit your examples and applications to data that come from simple random samples, or you fudge things.* I find some kinds of fudge harder to swallow than others. If the random samples were stratified, for example, I don't have a problem with pretending that we can ignore the stratification, because the key link between data production and inference is preserved. Samples were random, so generalization to a population is supported. What I have trouble getting down my throat is the use of the sampling model for data from randomized experiments. Do we want students to think that as long as there's any sort of randomization, that's all that matters? Do we want them to think that because the assignment of treatments was randomized, then they are entitled to regard the seven patients in the example as representative of all patients, regardless of age or whether their operation was a tonsillectomy or a liver transplant? Do we want students to leave their brains behind and pretend, as we ourselves apparently pretend, that choosing at random from two independent normal populations is a good model for randomly assigning treatments to a single convenience sample?

If the beginning course is even a tiny part as bad as I'm suggesting that it is, how did we ever come to teach it? I suggest that until fairly recently it was in fact quite a good course. It's not that the course has gotten worse; rather, it's that the opportunities to do a lot better have grown much faster than the course content has.

C. WHY WE TEACH IT: THE TYRANNY OF THE COMPUTABLE

J. Abner Pediwell, in his book *The Saber Tooth Curriculum,* wrote that "The important thing is to recognize the principle, not to do obeisance before one of the cogs of its mechanism." As a general principle, it's hard to argue with, but unfortunately, history shows that cogs and mechanisms have more to do with our choices than we might like. I've become convinced that a huge chunk of statistical theory was developed in order to compute things, or approximate things, that were otherwise out of reach. Until very recently, we had no choice but to rely on analytic methods. The computer has offered to free us and our students from that, but our curriculum is still mulishly insisting that our students drag the tractor behind them.

For historical perspective, and as a kind of proof-of-concept argument, I now invite you to join me in some speculation. To set the stage, I ask you to think about two questions with similar answers -- answers that I think contain an important lesson for those of us who teach introductory statistics.

Archimedes and Bayes.

The first of the two questions deals with the history of calculus. More than two thousand years ago, Archimedes knew a version of integral calculus, and showed how to use limits to compute areas under curves. *The Question*: If Archimedes knew about limits and how to use them to compute areas, back around 350 BCE, why did we have to wait another two thousand years for Newton and Leibniz to give us the modern version of calculus?

Question 2 has a similar structure. Thomas Bayes did his work on what we now call Bayesian inference around 1760. Laplace did a lot with Bayesian methods in the 1770s. Yet roughly 200 years later, in the 1950s, 60s, and 70s, hardly any statisticians were doing Bayesian data analysis. Several influential statisticians (De Finetti, Good, Lindley, Savage) wrote many widely read papers on the logical foundations of statistics, papers containing rigorous proofs to the effect that you had to be mentally deficient not to be a Bayesian in your orientation. Nevertheless, these impeccable arguments by influential statisticians won few converts. Most of us read the proofs, nodded in agreement, and continued to practice our deficiencies. Three more decades passed. Then, just in the last 15 years or so, our world has experienced a Bayesian Renaissance. Why?

I suggest that the answers to these two questions are similar. Consider first the calculus question. The work of Archimedes, like all of Greek mathematics at the time, was grounded in geometry. The geometric approach had two major limitations: it didn't lend itself easily to generalization -- finding the area under a parabola doesn't lead easily to finding the area under an arbitrary curve – and it didn't lead easily to a solution of the inverse problem – finding the slope of a tangent line. For two millennia, the geometric calculus of Archimedes remained largely dormant, a sleeping beauty, waiting for the magic awakening that was to begin in the watershed year of 1637. During the intervening two thousand years of dormancy, Arabic numerals made their way from the Madrassa in Fez, Morocco across the Mediterranean to the Vatican in Rome, brought by Pope Leo IX, and algebra made its way across North Africa to Gibraltar to Renaissance Italy. Finally, in 1637, Fermat and Descartes made geometry *computable* via the coordinate system of analytic geometry, and after that computational innovation it took a mere three short decades before Newton and Leibniz gave us the modern derivative. The core *idea* of calculus – taking a limit – was known to Archimedes two millennia earlier. What had held things up was not a missing *idea* so much as a missing *engine*, a missing crank to turn. The sleeping beauty was awakened not by a magic kiss, but by a cog in the mechanism.

Similarly, I suggest, with Bayesian inference. Bayes gave us the *idea*, posthumously, in 1763. Laplace showed how to apply the same logic much more broadly. Two centuries later, Lindley and Savage and the other Foundational Evangelists made clear that if we failed to convert to Bayes, we risked the damnation of eternal incoherence. Their evangelism was all to no avail: the tent of Bayesian Revival remained largely empty. No one answered the alter call. Then along came Markov chain Monte Carlo, and suddenly Bayes was everywhere. What had held things up was not a missing *idea* so much as a missing *engine*, a missing crank to turn – a cog in the mechanism.

To me, an important lesson in all this is that, historically, we have always tended to underestimate the extent to which what we are *able* to do shapes what we think we *ought* to do. Almost surely you know the story of the drunk who was looking for his keys under a street light. “Where’d you lose them?” he was asked. “Back down the block.” “Shouldn’t you look for them back there?” “No, it’s too dark there. I couldn’t see what I was doing.” Much as we’d like to think that our sense of what’s appropriate drives our sense of what choices we have, reality is much less logical. The set of choices we have available to us constrains the range of decisions we evaluate as possible options. The drunk looks under the light not because it’s appropriate but because that’s where he can see. We statisticians are not much more sophisticated. In the 1960s we did *not* shift our center of intellectual gravity toward Bayes, because computationally, we couldn’t see how to do it. In the 1990s we *did* shift toward Bayes, because by then we could. Intellectually, we were not much better than the drunk. [5]

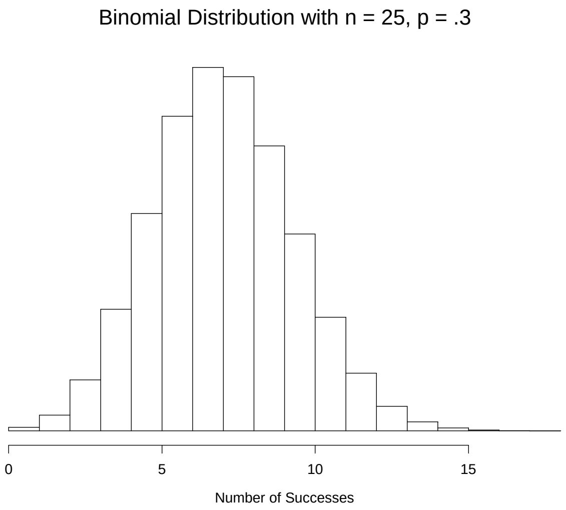
Bernoulli and De Moivre

In much the same way, two of the major results of introductory statistics – the law of large numbers and the central limit theorem -- were driven by computation. Turn back the clock to the early 1690s, to Jacob Bernoulli and the Weak Law of Large Numbers (aka Law of Averages). What Bernoulli really wanted in his research was the equivalent of a confidence interval for a binomial probability: He wanted to know, “How many trials do you have to observe in order to be able to estimate *p* with a given precision?” He wasn’t able to solve this problem, and had to settle for the limit theorem that bears his name. Why couldn’t he solve it? Because *he couldn’t compute the tail probabilities*:

$$\binom{n}{r} p^r (1-p)^{n-r} + \binom{n}{r+1} p^{r+1}(1-p)^{n-(r+1)} + \ldots + \binom{n}{n} p^n (1-p)^0$$

---

[5] There’s a crude joke about a dog that explains what he *does* do in terms of what he *can* do. If you don’t know the joke, don’t hold your breath: I’m not about to tell you. What matters is that the joke captures an important fact about life: the dog does what he does not because it’s right but because he can. Our profession’s conversion to Bayesian data analysis was similar. In the 1960s, it was shown that to be Bayesian was right, but because Bayesian methods were out of reach at the time, few converted. They couldn’t, so they didn’t. Many of those who became Bayesians in the 1990s were merely taking advantage of the fact that, computationally speaking, their posteriors had become accessible.



Bernoulli ended up using a geometric series (and a huge pile of algebra) to bound the tail probabilities, because he *did* know how to sum a geometric series. Then he took a limit to show that as the sample size increased, the tail probability went to zero.[6] Thus one of our major theorems, the Law of Averages, arose *as an end run around a computing impasse*.

Now fast forward 30 years to the 1720s and De Moivre's version of the Central Limit Theorem. What problem was De Moivre working on? The same one that Bernoulli had been unable to solve: the problem of how to compute binomial tail probabilities. He found the normal distribution as a way to approximate those tail probabilities. In this sense, the normal distribution and the Central Limit Theorem arose *as a by-product of a 30-year struggle with a computing impasse.*

To me, the lesson here is clear. In statistics, our vision has always been blinkered by what we can compute. *No algorithm, no option*. We are always at risk of remaining intellectually imprisoned in the labyrinth of computability. If we are to soar like Daedalus above the maze, we must cut away the constraining thorns of old analytical paradigms, and ask where we really want to go, rather than limit ourselves and our curriculum short-sightedly to the next available turn in the hedge.

In this context, we need to remember that Pitman's seminal work on the permutation test was published in 1937, at a time when today's computing power was not even imagined. We've seen what happened to Bayesian methods once statisticians were able to compute posterior distributions. We should think hard about the permutation test now that it is so easy to implement. [7]

[6] A good source is Uspensky, J.V. (1937). *Introduction to Mathematical Probability,* New York: McGraw-Hill, Chapter VI.

[7] I'm inclined to take things a step beyond that, and suggest that we may even be resisting the permutation test in part *because* the theory is so analytically shallow. The computer is the only possibility, which may make things entirely too simple for some people!

D. WHAT WE SHOULD TEACH: THE THREE R's OF INFERENCE

We need a new curriculum, centered not on the normal distribution, but on the logic of inference. When Copernicus threw away the old notion that the earth was at the center of the universe, and replaced it with a system that put the sun at the center, his revolution brought to power a much simpler intellectual regime. We need to throw away the old notion that the normal approximation to a sampling distribution belongs at the center of our curriculum, and create a new curriculum whose center is the core logic of inference.

What is that core logic? I like to think of it as three Rs: randomize, repeat, reject. Randomize data production; repeat to see what's typical; reject any model that puts your data in its tail.

The three Rs of inference: RANDOMIZE, REPEAT, REJECT

1. RANDOMIZE data production
    - To protect against bias
    - To provide a basis for inference
        - random samples let you generalize to populations
        - random assignment supports conclusions about cause and effect
2. REPEAT by simulation to see what's typical
    - Randomized data production lets you re-randomize, over and over, to see which outcomes are typical, which are not.
3. REJECT any model that puts your data in its tail

As it stands, the third R, Reject any model that puts your data in its tail, describes the logic of Fisher's approach to hypothesis testing, but the same logic gives us confidence intervals also, as the set of parameter values not rejected by the corresponding hypothesis test. (Years ago, Jim Swift, Jim Landwehr, and Ann Watkins showed an effective way to do this at the elementary level.) There's even a rejection-based algorithm for finding posterior distributions, which I described two years ago in a talk at the JSM. After describing the randomization distribution for experiments, we could handle the sampling model using the very same distribution, using the fact that under the null hypothesis, and conditional on the observed values, the probability model is exactly the same as for the randomized experiment. To see all this spelled out, I highly recommend the lovely article by Michael Ernst in the August 2004 issue of *Statistical Science*.[8] We could still teach the *t*-test, but it would make only a cameo appearance, almost as an afterthought: "This is what people had to do in the old days, as an approximate method, before computers made it possible to solve the problem directly." For the time being, we could also add, "Lots of people still use this old approximation, out of nostalgia, but its days are numbered, just like the rotary phone and the 8-track tape player."

E. WHY WE SHOULD TEACH IT: a dozen reasons

I promised to conclude with shameless shilling, and so here goes:
If we teach the permutation test as the central paradigm for inference, then

[8] Ernst, *op cit*.

1. the model matches the production process, and so it allows us to stress the connection between data production and inference;
2. the model is simple and easily grasped;
3. the distribution is easy to derive for simple cases (small *n*) by explicitly listing outcomes;
4. the distribution is easy to obtain by physical simulation for simple situations;
5. the distribution is easy to obtain by a computer simulation whose algorithm is an exact copy of the algorithm for physical simulation;
6. expected value and standard deviation can be understood concretely by regarding the simulated distribution as data;
7. the normal approximation is empirical rather than "theory-by-fiat;"
8. the entire paradigm generalizes easily to other designs (e.g., block designs), other test statistics, and other data structures (e.g., Fisher's exact test);
9. it is easy and natural to teach two distinct randomization schemes, random sampling and random assignment, with two kinds of inferences;
10. it offers a natural way to introduce students to computer-intensive and simulation-based methods, and so offers a natural lead-in to such topics as the bootstrap;
11. it frees up curricular space for other modern topics;[9] and,
12. finally, we should do it because Fisher told us to. Actually, he said essentially that we should do it, except that we can't, and so we have been forced to rely on approximations:

> "the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method."[10]

Statistics teachers of the world, unite! You have nothing to lose but your tractor and your cello.

[9] See the introduction to statistics for mathematically inclined students by Allan Rossman and Beth Chance, at California Polytechnic State University, San Luis Obispo, and the introductory curriculum being developed by Daniel Kaplan at Macalester College.

[10] Fisher, R.A. (1936), "The coefficient of racial likeness and the future of craniometry" *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, vol. 66, pp. 57-63, quoted in Ernst (2004), *op cit*.