The Introductory Statistics Course: A Ptolemaic Curriculum

George W. Cobb
Mount Holyoke College

1. INTRODUCTION

The founding of this journal recognizes the likelihood that our profession stands at the threshold of a fundamental reshaping of how we do what we do, how we think about what we do, and how we present what we do to students who want to learn about the science of data. For two thousand years, as far back as Archimedes, and continuing until just a few decades ago, the mathematical sciences have had but two strategies for implementing solutions to applied problems: brute force and analytical circumventions. Until recently, brute force has meant arithmetic by human calculators, making that approach so slow as to create a needle's eye that effectively blocked direct solution of most scientific problems of consequence. The greatest minds of generations past were forced to invent analytical end runs that relied on simplifying assumptions and, often, asymptotic approximations. Our generation is the first ever to have the computing power to rely on the most direct approach, leaving the hard work of implementation to obliging little chunks of silicon.

In recent years, almost all teachers of statistics who care about doing well by their students and doing well by their subject have recognized that computers are changing the teaching of our subject. Two major changes were recognized and articulated fifteen years ago by David Moore (1992): we can and should automate calculations; we can and should automate graphics. Automating calculation enables multiple re-analyses; we can use such multiple analyses to evaluate sensitivity to changes in model assumptions. Automating graphics enhances diagnostic assessment of models; we can use these diagnostics to guide and focus our re-analyses.

Both flavors of automation are having major consequences for the practice of statistics, but I think there is a third, as yet largely unrecognized consequence: we can and should rethink the way we present the core ideas of inference to beginning students. In what follows I argue that what we teach has always been shaped by what we can compute. On a surface level, this tyranny of the computable is reflected in the way introductory textbooks have moved away from their former cookbook emphasis on recipes for calculation using artificial data with simple numbers. Computers have freed us to put our emphasis on things that matter more. On a deeper level, I shall argue, the tyranny of the computable has shaped the development of the logic of inference, and forced us to teach a curriculum in which the most important ideas are often made to seem secondary. Just as computers have freed us to analyze real data sets, with more emphasis on interpretation and less on how to crunch numbers, computers have freed us to simplify our curriculum, so that we can put more emphasis on core ideas like randomized data production and the link between randomization and inference, less emphasis

on cogs in the mechanism, such as whether 30 is an adequate sample size for using the normal approximation.

To illustrate my thesis, I offer the following example from Ernst (2004).

Example: Figure 1 shows 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.

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

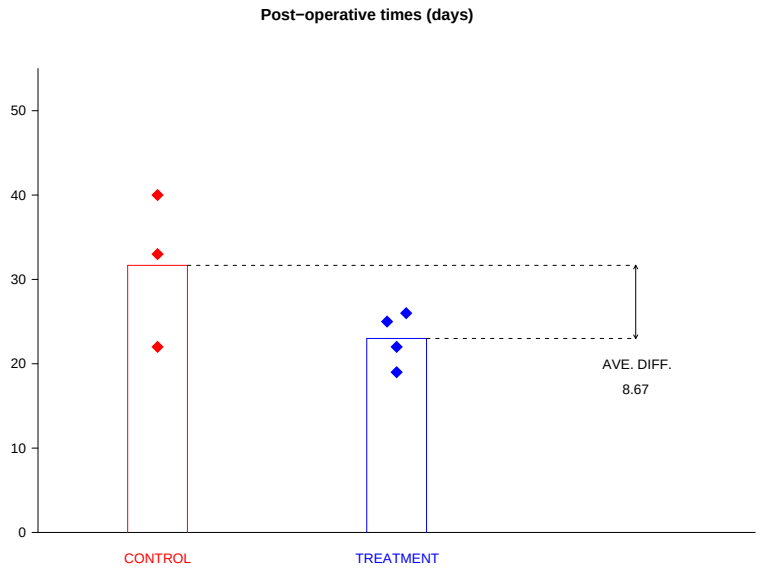

Figure 1: Results of a randomized controlled experiment.

How would your best students, 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. Every statistical method is derived from a model, and one of our goals in teaching statistical thinking should be to encourage our students to think explicitly about the fit between model and reality. Here, that fit is not good.

> Model: Data constitute two independent simple random samples from normal populations.
>
> Reality: Data constitute a single convenience sample, 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, as illustrated in Figure 2 (Ramsey and Shafer, 2002, p.9)?

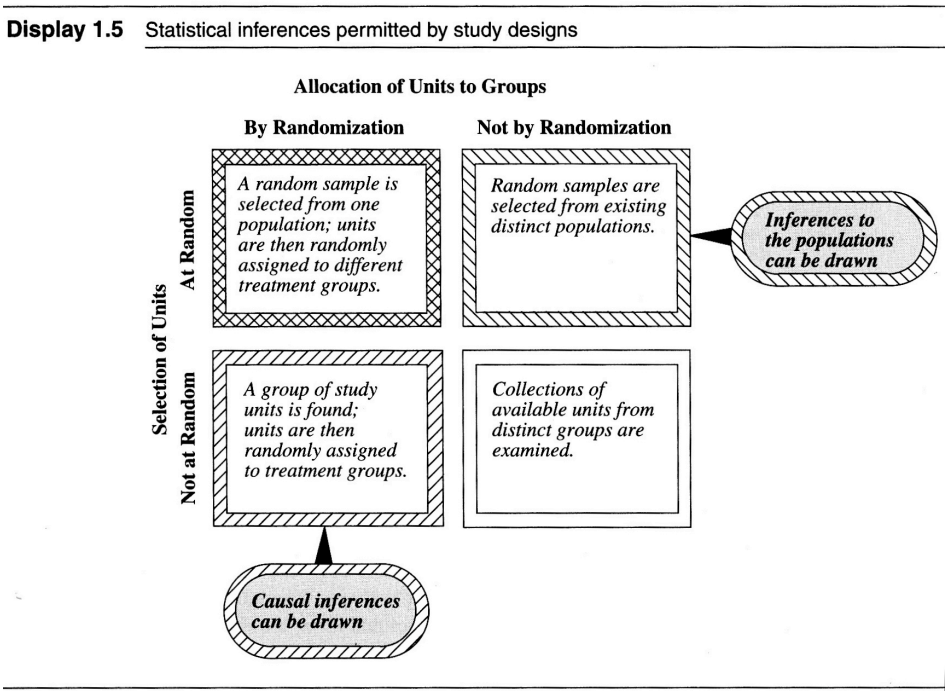

Figure 2: A chart from Ramsey and Shafter (2002).

An alternative to the t-test for this data set is the permutation test: Start with the null hypothesis that the treatment has no effect. Then the 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 is that both the content and the structure of our introductory curriculum are shaped by old history. 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 needlessly complicated, because he put the earth at the center of his system, instead of putting the sun at the center. Our curriculum is needlessly 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:

1. *What* we teach: our Ptolemaic curriculum
2. *What is wrong* with what we teach: three reasons
3. *Why* we teach it anyway: the tyranny of the computable
4. *What we should* teach instead: putting inference at the center
5. *Why we should* teach it: an unabashed sales pitch.

2. WHAT WE TEACH: OUR PTOLEMAIC CURRICULUM

I offer Ptolemy's description of the universe as a metaphor for our current introductory curriculum, an instructive 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 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. 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. Eventually Ptolemy ended up with epicycles on epicycles, at least eighty in all.

Now consider parallels with 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 tend to be roughly normal. 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." Only now can students begin to construct confidence intervals or hypothesis tests, and only for the unrealistically simple case of a single normal sample with known standard deviation. Notice that none of the probability scaffolding is essential: students can learn randomization-based hypothesis testing without any of it. (For a visionary textbook that introduces inference by basing it directly on randomized data production, I recommend Robert Wardop's (1994).)

No sooner have we reached the central result about $\bar{x}$ than 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: s for σ; 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 samples 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 Agresti and Coull's (1998) artificial pairs of successes and failures to the actual data in order to get a better approximation. We may not have a total of eighty epicycles, but each adjustment we ask our students to learn takes their attention away from more basic ideas such as the fit between model and reality.

In the standard approach, there are so many variations that it is hard for students to recognize and appreciate the unifying theme. We ourselves may understand that there is a single paradigm based on the use of (observed – expected)/(standard error), but students encountering this material for the first time must master the details of a daunting array of at least seven variations:

Measured data	**Counted data**
z (known standard deviation)	one sample, normal approximation
one-sample t	two-sample normal approximation
two-sample t, pooled SDs	Agresti/Coull adjustment
two-sample t, two SDs	

It is no wonder that beginners rarely manage to get past the details to the main ideas.

3. 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 (1992) 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 kinds: 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 taking-a-sample-and-computing-a-summary-number is again a hard transition to make. (I owe this important idea, that it is hard for students to make the transition from a one-time process like taking a limit at a point, or taking a single sample and computing the mean, to an abstract entity like a derivative or sampling distribution, created by applying the process repeatedly, to Mogen Niss, Roskilde Universitetcenter, via personal communication from David S. Moore.) 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. 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**. 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 is easy for us to be optimistic about the future of our subject, 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 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. We may be living in the early twenty-first century, but our curriculum is still preparing students for applied work typical of the first half of the twentieth century.

As statisticians we are used to going into a new area and learning quickly enough of what

we need 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 in order to get to 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 is 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 persnickety 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 if you choose a random sample but don't randomize the assignment of treatments, it is still valid to conclude that the treatment caused the observed differences? 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 a large normal population is a good model for randomly assigning treatments?

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.

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

A sage once advised "The important thing is to recognize the principle, not to do obeisance before one of the cogs of its mechanism." As a general directive, 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 at best in the early stages of accepting the offer.

To appreciate the extent to which our thinking is kept on a tight leash by what we can compute, consider a pair of questions from the history of mathematics and statistics. 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?

The second question 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, including Birnbaum (1962), De Finetti (1972), Good (1950), Lindley (1965),and Savage (1954), wrote many widely read papers and books addressing the logical foundations of statistics and containing 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, calculus 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 Al-Kaourine Madrassa in Fes, Morocco across the Mediterranean to the Vatican in Rome, brought by Pope Sylvester II (Landau 1958), 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 (1812) 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. 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.

Now turn back the clock to the early 1690s, to Jacob Bernoulli (1693) 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 (see Figure 3):

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

Bernoulli ended up using a geometric series (and prodigious 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. (A good source is Uspensky, J.V. (1937). Introduction to Mathematical Probability, New York: McGraw-Hill, chapter VI.) 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 (1733). 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

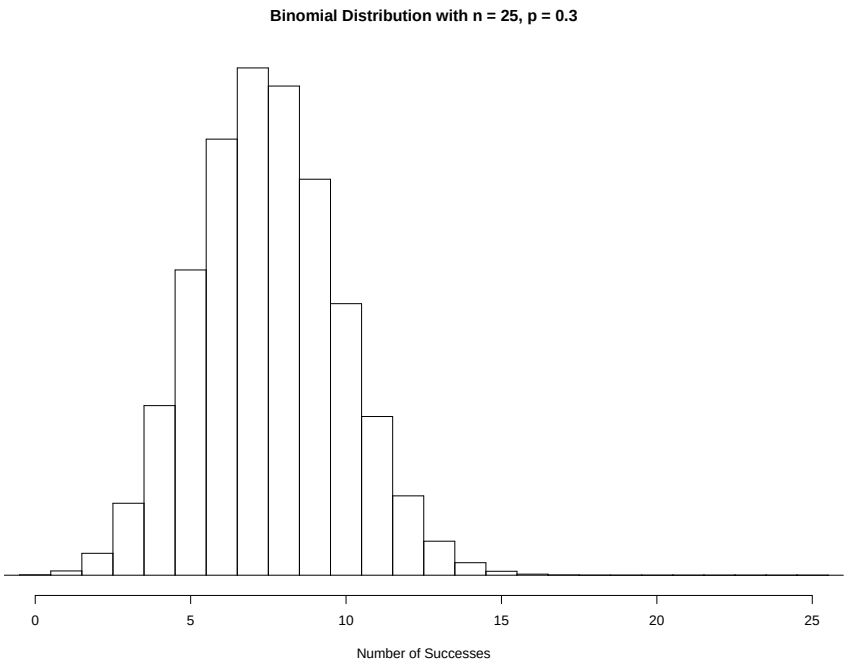

Figure 3: Binomial tail probabilities of the sort Bernoulli wanted to compute.

imprisoned in the labyrinth of the number crunching Minataur. 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. 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!

5. WHAT WE SHOULD TEACH: THE THREE Rs 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 by simulation to see what's typical and what's not; 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

To fill out my proposed curriculum a bit more, what I have in mind is that we would teach exploratory data analysis more or less as we do now, and then teach randomized data production, both sampling schemes and experimental design. Then, we would introduce inference by way of the permutation test for randomized experiments. We could introduce confidence intervals as the set of values not rejected by the corresponding test, and analyze the sampling model using the fact that under the null hypothesis, 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 article by Michael Ernst in the August 2004 issue of *Statistical Science*. We could still teach the t-test, but it would appear 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, but its days are numbered, just like the rotary phone and the 8-track tape player."

6. WHY WE SHOULD TEACH IT: A DOZEN REASONS

If we teach the permutation test as the central paradigm for inference, then

1. the model matches the production process, and so it easy to emphasize 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 defined 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; and,

12. finally, we should do it because Fisher (1936) 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."

7. CONCLUSION

Much of what we currently teach to beginning students of statistics – a curriculum shaped by its once-necessary but now-anachronistic reliance on the normal as an approximate sampling distribution – is technically much more demanding, and substantively much more peripheral, than the simpler and more fundamental ideas that now, thanks to computers, we could and should be teaching. Before computers, there was no alternative. Now, there is no excuse. Unfortunately, as suggested both by a reviewer and by a premise of this essay, we teachers of statistics will probably be slow to change. Just as practicing statisticians resisted abstract arguments and were slow to adopt Bayesian methods of data analysis, until, first, MCMC made the computations more nearly automatic, and then, opinion leaders in the profession began to publicize their own Bayesian analyses, I predict that teachers of statistics will initially resist abstract arguments like the one I have made here. Where will we find the keys to change? Right now, the dominant textbooks illuminate a portion of the curricular landscape that is down the block from where I think the keys can be found, and most of us, even if fully sober, choose to hunt where the light of curricular fashion shines most brightly. What we need is a generation of adventurous authors who will choose to beam their light on different ground, thereby making the new curriculum easier to implement, and a new generation of adventurous teachers, willing to lead their students down roads less trammeled.

8. REFERENCES

Agresti A and Coull BA. (1998). "Approximate is better than Ôexact' for interval estimation of binomial proportions. *The American Statistician*, 52:119-126.

Bayes, Thomas (1763). "Essay Towards solving a problem in the doctrine of chances", *Phil. Trans. Roy. Soc.*, 53, 370-418. (Reprinted: Biometrika, 1958, 45, 293-315.)

Bernoulli, J. (1713). *Ars Conjectandi*, Basel: Thurnisorium.

Birnbaum, Allan (1962). "On the Foundations of Statistical Inference," *J. Am. Stat. Assoc.*, 57, 269-306.

De Finetti, Bruno (1972). *Probability, Induction, and Statistics*, New York: John Wiley and Sons.

De Moivre, Abraham (1733). *The Doctrine of Chances*, London: Woodfall.

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

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.

Good, I.J. (1950).*Probability and the Weighing of Evidence.* New York: Hafner Publishing Company.

Hald, Anders (1990). *A History of Probability and Statistics and their Applications before 1750*, New York: John Wiley and Sons.

Hald, Anders (1998). *A History of Mathematical Statistics from 1750 to 1930*, New York: John Wiley and Sons.

Landau, Rom (1958). "The karaouine at Fes," it The Muslim World,vol. 48, pp. 104-12.

Laplace, P.S. (1812). *ThŐorie Analytique des ProbabilitŐs*, Paris: Courcier.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Cambridge University Press.

Moore, David S. (1992). "Teaching Statistics as a Respectable Suibject," in Florence Gordon and Sheldon Gordon, eds., *Statistics for the Twenty-First Century, MAA Notes*, No 26, Mathematical Association of America.

Pitman, EJG (1937). "Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4, 119-130.

Ramsey, Fred L. and Daniel W. Schafer (2002). *The Statistical Sleuth.* Pacific Grove, CA:

Duxbury

Savage, L.J. (1954). *The Foundations of Statistics*, New York: Wiley.

Uspensky, J.V. (1937). *Introduction to Mathematical Probability*, New York: McGraw-Hill.

Wardop, Robert (1994) *Statistics: Learning in the Presence of Variation.* Dubuque, IA: William C. Brown.