Our Abstinence-Based Curriculum and other Stories from Statistics Education

Daniel Kaplan
Macalester College St. Paul, Minnesota, USA
November 10, 2010

Abstract:
The subtitle might have been, “How religious fervor has shaped the way we teach statistics.” But this would mislead you into thinking about actual religion and contemporary disputes about evolution, sex education, and so on.

No, the talk will be about plain, old, everyday introductory statistics and the familiar topics that are studied: the t-test, one-and two-tailed tests, association vs. causation, etc.

The conventional choice of topics and approaches, it will be claimed, is based largely on honoring the faith of our statistical fathers and the quasi-theological disputes of old, and not on extracting useful information from the sorts of data commonly encountered today. With an eye toward curricular reformation, a few theses will be offered.

Austin Bradford Hill, Presidential Address
Meeting of the Royal Statistical Society, Section of Occupational Medicine, on January 14, 1965

The Environment and Disease: Association or Causation

Hill on Association and Causation
“... in what circumstances can we pass from this observed association to a verdict of causation? Upon what basis should we proceed to do so?” Provides nine (not ten!) “aspects of that association [that we should] especially consider before deciding that the most likely interpretation of it is causation.”

1. Strength — 16 inches
2. Consistency — 20
3. Specificity — 9
4. Temporality — 2
5. Biological gradient — 5
6. Plausibility — 5
7. Coherence — 7
8. Experiment — 2
9. Analogy — 1
The number shows the amount he wrote on each aspect, in column inches.

Hill on Experiment
“Occasionally, it is possible to appeal to experimental, or semi-experimental, evidence. For example, because of an observed association some preventive action is taken. Does it in fact prevent? The dust in the workshop is reduced, lubricating oils are changed, persons stop smoking cigarettes. Is the frequency of the associated events affected? Here the strongest support for the causation hypothesis may be revealed.”

Hill on Tests of Significance
“... the glitter of the t table diverts attention from the inadequacies of the fare. Only a tithe, and an unknown tithe, of the factory personnel volunteer for some procedure or interview. 20% of patients treated in some particular way are lost to sight, 30% of a randomly-drawn sample are never contacted. The sample may, indeed, be akin to that of the man who, according to Swift, ‘had a mind to sell his house and carried a piece of brick in his pocket, which he showed as a pattern to encourage purchasers.’ ”
Our Abstinence-Based Curriculum and other Stories from Statistical Education

Nov 10, 2010

Stranger in a Strange Land

My journey to the land of statistics

- Studied physics, philosophy, engineering: all hostile to statistics.
- Involved in analysis of data in terms of chaos theory. Hypothesized application to cardiology: Is fibrillation chaotic? Can we predict sudden death?
- Textbooks in nonlinear dynamics applied to biology and computer programming for scientists.
- My first statistics course was the one I was assigned to teach in 1997 — using Moore and McCabe.
- Now I’m the director of Macalester’s Applied Math and Statistics major. Recently finished an introductory textbook on statistical modeling.

What I Learned in Grad School

Techniques

- Non-parametric models: locally linear/quadratic, neural networks, radial basis functions.
- Time series analysis: autocorrelation, fourier transform
- Unsupervised clustering, k-means, ...
- “Edgy” chaos theory techniques: dimension estimation, nonlinear prediction.
- NOT ever: t-tests, ANOVA, simple or multiple regression and diagnostics.

And when I needed Stats?

1. Press et al., Numerical Recipes

Things that surprised me when I started teaching stats

- Are the hormone fluctuations chaotic?
- Can you distinguish them from linear dynamics with noisy inputs?
- Work with clustering, neural networks, classifiers, etc, with no statistical background whatsoever.

To an outsider, some problems are obvious

- Archaic and misleading language: e.g., “significance”, “deviation,” “error,” “standard”
- Don’t teach algorithms, but to use the algorithms.
- Need measures of effect size, not just correlations.
- Units are important in science!
- To be interesting, statistics has to deal with situations of interesting complexity. Can’t be about means versus medians.
- There’s lots of judgment in statistics, to give students settings that require judgment, e.g. Which model is best? What’s a normal value in this setting?
- Students don’t understand even simple formulas, e.g. $s/\sqrt{n}$
- Archaic technology, e.g., tables, and failure to adopt computing.
- Adopting computing means teaching about computing, because nobody else is doing it.
- Using teaching software instead of software principles.

How I reacted

- Almost all those tests are based on regression, so teach that.
- Techniques that give 1.5 digits of the p-value are good enough to start with.
- Students don’t understand even simple formulas, so teach them the consequences of sampling by drawing actual samples.
- Descriptive statistics should be rich: multiple regression and modeling, so that you can describe a complicated world.
- Teach basics of conditional probability so that p-values make sense.
- Create the worlds of the Null and Alternative hypothesis: draw data from them.
- Variables are not lurking, they are there to be dealt with appropriately: measuring covariates, adjustment, randomization.
- Teach Bonferroni, and worry about the details later.

Common Sense or Hypothesis Testing

- Are the hormone fluctuations chaotic?
- Can you distinguish them from linear dynamics with noisy inputs?
- Work with clustering, neural networks, classifiers, etc, with no statistical background whatsoever.

How I reacted

- Almost all those tests are based on regression, so teach that.
- Techniques that give 1.5 digits of the p-value are good enough to start with.
- Students don’t understand even simple formulas, so teach them the consequences of sampling by drawing actual samples.
- Descriptive statistics should be rich: multiple regression and modeling, so that you can describe a complicated world.
- Teach basics of conditional probability so that p-values make sense.
- Create the worlds of the Null and Alternative hypothesis: draw data from them.
- Variables are not lurking, they are there to be dealt with appropriately: measuring covariates, adjustment, randomization.
- Teach Bonferroni, and worry about the details later.
Physicists and engineers (and likewise computer scientists) are ambitious; when faced with problems, they tend to attack, sweeping aside impediments stemming from limited knowledge about the procedures that they apply.

In emphasizing the logic of data manipulation, teachers of statistics are instilling excessive caution. Students seem to develop extreme risk aversion, apparently fearing that the inevitable flaws in their analysis will be discovered and pounced upon by statistically trained colleagues.


Looking for a Generalization …

The Unhappy ABCs of Statistics Education
1. Abstinence-Based Curriculum
2. Association Beats Causation
3. Algebra Better than Computing
4. Algebra But not Calculus
5. Not Adjustment Based Conclusions

Proofiness

Emphasizing statements that are demonstrably valid, that is, deductible from stated premises, but allowing the misconception/misinterpretation that the statements are therefore true, that is, applicable to the real world.

Antonym: “truthiness.”

Truthiness

Truthiness is a “truth” that a person claims to know intuitively “from the gut” with regard to evidence, logic, intellectual examination, or facts. Coined/popularized by Stephen Colbert in 2005.

Proofiness: The Book

Charles Seife, Proofiness: The Dark Arts of Mathematical Deception

“it all comes down to numbers, the author argues, and the ways they can be used to make people believe things that are not true.” — David Piet

Bullshit

Philosopher Harry G. Frankfurt, in On Bullshit (2005), writes about the bullshitter’s complete disregard for whether what he’s saying corresponds to facts in the physical world. He “does not reject the authority of the truth, as the liar does, and oppose himself to it. He pays no attention to it at all. By virtue of this, bullshit is a greater enemy of the truth than lies are.”
Approximation
Our answers are only rough. They need to be good enough to guide decisions. We also need an operational way to see how rough our answers are and whether they are good enough to guide our decisions.

0.01 = 0.004769
0.10 = 0.50

Encapsulation/Information Hiding
Encapsulation is "the process of compartmentalizing the elements of an abstraction that constitute its structure and behavior; encapsulation serves to separate the contractual interface of an abstraction and its implementation." — Grady Booch

An Example of Proofiness: The Alternative Hypothesis
A theoretical problem: Should the χ²-test be one- or two-sided.
- Fisher: Two-sided. Too good a fit is a sign of a problem. "If P is between .1 and .9, there is no reason to suspect the hypothesis being tested." Fisher 1925, Statistical Methods for Research Workers, p. 71.
- Pearson: One-sided. We should not reject a "gradation curve" [a model of a distribution] because it is too close to the data. It's not clear to me who is right, since it depends on the purpose of the test. Looking for fraud, for signs of too many parameters, or flaws in the error model?

A theoretical resolution
The Neyman-Pearson lemma. A one-tailed test gives greater power at any specified significance. This doesn't really address the issue of the purpose of the test, but ...

Examples of Encapsulation
Multiplication on the computer.
> 1138/1000 == 0.001*1138
[1] FALSE

Interpolation and anti-differentiation
demand.points = c(0,40,90,160,250,400,600,850,1200)
price.points = c(1250,1250,700,550,450,350,250,150,50)
f3 = approx(demand.points, price.points, method="const"
actual.revenue = antiD( f3 )

Textbook Coverage of the Alternative
The alternative hypothesis is denoted by H₁. It states what the researchers suspect or hope to learn about the parameter of interest. It depends on the purpose of the study and must be specified before the data are examined. The alternative hypothesis can take one of three forms:

1. H₁: μ > μ₀ (one-sided)
2. H₁: μ ≠ μ₀ (two-sided)
3. H₁: μ < μ₀ (one-sided)

The first two forms are called one-sided alternatives, while the last is a two-sided alternative.

Textbook Coverage of the Alternative
The alternative hypothesis is denoted by H₁. It states what the researchers suspect or hope to learn about the parameter of interest. It depends on the purpose of the study and must be specified before the data are examined. The alternative hypothesis can take one of three forms:

1. H₁: μ > μ₀ (one-sided)
2. H₁: μ ≠ μ₀ (two-sided)
3. H₁: μ < μ₀ (one-sided)

The possible null and alternative hypotheses are one of these three choices, depending on the research question:

1. H₀: μ ≥ μ₀ versus H₁: μ < μ₀ (one-sided)
2. H₀: μ ≥ μ₀ versus H₁: μ = μ₀ (two-sided)
3. H₀: μ = μ₀ versus H₁: μ > μ₀ (one-sided)

Often the null hypothesis for a one-tailed test is written as H₀: μ = μ₀ instead.

Source: Utts and Heckard, Mind on Statistics 1/e, 2000
Textbook Coverage of the Alternative

“Stating $H_1$ is often the more difficult task.”

Source: Moore and McCabe, p. 450 “Introduction to the Practice of Statistics” 2/e, 1993

The "Anything but" Hypothesis

The alternative is generally introduced as an “anything but” hypothesis — anything but the null.

Problems with this ...

- It encourages students to think too abstractly. If we believe that applied statistics requires some knowledge of the field of application, why encourage a statement of the alternative that has absolutely no contact with the field of application.
- We're starting students off with a mathematically complicated form of a hypothesis: a "compound hypotheses," not a simple statement of what the world is like.
- A consequence of the mathematical complexity of the "anything but" hypothesis is that it becomes difficult to estimate power.

The "Anything but" Hypothesis

- Benefit: Increases the power (Neyman-Pearson Lemma) You can prove this!
- Cost (for teaching): Adds complexity.
- Cost (for teaching): Disconnects statistics from science. You don't have to talk about the specifics of the system when presenting examples.
- Cost (for teaching): Lost opportunity to teach about things that affect the results substantially, e.g., covariates and adjustment.
- Cost (for research): Encourage fraud — ex post facto trimming of the p-value. Statistical circumcision. $p < 0.10$ is satisfactory, so long as you can justify a one-sided test.

Costs and Benefits to One-Sided Tests

- Benefits: Increases the power (Neyman-Pearson Lemma) You can prove this!
- Cost (for teaching): Adds complexity.
- Cost (for teaching): Disconnects statistics from science. You don’t have to talk about the specifics of the system when presenting examples.
- Cost (for teaching): Lost opportunity to teach about things that affect the results substantially, e.g., covariates and adjustment.
- Cost (for research): Encourage fraud — ex post facto trimming of the p-value. Statistical circumcision. $p < 0.10$ is satisfactory, so long as you can justify a one-sided test.

How much does it increase the power? Is it worthwhile?

- Calculations indicate that for a study with 80% power, giving up the added power of the one-sided test is compensated by an approximately 20% increase in sample size.
- That’s not nothing. But if we are going to be worrying about factors of 20% in sample size, we had better have already covered the matters that lead to order of magnitude estimates of sample size. This we typically have not done.
- Indeed, many statistics instructors that I talk to say that they don’t cover power, and so how can a student understand the benefit of a one-sided test anyways?

Pedagogical Triage

It’s a truism that our curricula are overwhelmed. There is too much to teach.

Triage

A process of determining the priority of patients' treatments based on the severity of each patient's condition. Used when resources are insufficient for all patients to be treated immediately.

Triage in Statistics Education

First introduce the concept of hypothesis testing and give them the tools to get the first digit right, then worry about calculations on the order of 20%.

Proposal

Make it a habit to develop every hypothesis testing example with a specific alternative hypothesis. Not just the “direction” of the effect but the magnitude.

Benefits

- Helps to develop students' concepts of "significance" and how it differs from "substantial."
- Allows a discussion of sample size.
- Highlights the disconnect between the alternative and the p-value. (“But where did we use the alternative hypothesis?”)
- Provides a connection to the practice of science. It’s not usually the null that you’re interested in.
The motivation for the one- vs two-sided distinction is power. What things affect power? In rough order of importance.

- Sample size.
- The hypothesized value for the alternative hypothesis. (If the effect is small ... don’t bother.)
- Covariates.
- Appropriate experimental design, e.g. cross-over
- Precision of measurements.
- Multiple tests.
- One- vs two-sided.

We’re teaching the last item on this list and largely ignoring the preceding items.

What to Teach about Power?

Objection I hear from instructors:
• I’m teaching statistics, not public health (or biology, or economics, or .... )
• I don’t have the time to talk about the details of the examples.
These answers are about proofiness.

Real data
- GAISE emphasizes importance of using “real data.”
- What makes it real is the context it which it was collected.
- The context is therefore important.

We should be modeling for our students the idea that you should know something about the field of application if you are going to work in it.

But What Should the Alternative Be?

Three General Forms for a Specific Alternative

If you don’t know what the alternative should be, you have an opportunity to show your students how to figure it out.

Proofiness doesn’t apply: you need to know about the world to frame a meaningful, specific alternative.

You need never be without a specific alternative.
1. Your best guess.
2. The smallest interesting effect.
3. The population is like your preliminary study.

There’s no such thing as the “right” alternative. It’s a hypothesis!

Isn’t any Effect Interesting?

Examples:
• ESP: If you can do even some, that tells us something.
• Political polls: It takes only 1 vote to win!

But the sampling/experimental methodology isn’t infinitely reliable. You need to have some sense of the size of possible bias, e.g., sampling of non-voters, cues to the ESP guesser.

Austin Bradford Hill
“The glitter of the t table diverts attention from the inadequacies of the fare.”

A Proofiness Question

Suppose we have a single population that has been assigned meaningless labels randomly. We take a sample and measure some quantity. What can we say about how the labels relate to the measured quantity?

You can prove statements about this situation: the Null Hypothesis.

<table>
<thead>
<tr>
<th>Label Type</th>
<th>Measurement type</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 levels</td>
<td>quantitative</td>
<td>t-test</td>
</tr>
<tr>
<td>&gt; 2 levels</td>
<td>quantitative</td>
<td>ANOVA</td>
</tr>
<tr>
<td>2 levels</td>
<td>2 levels</td>
<td>p-test</td>
</tr>
<tr>
<td>&gt; 2 levels</td>
<td>&gt; 2 levels</td>
<td>( \chi^2 ) test</td>
</tr>
<tr>
<td>quantitative</td>
<td>quantitative</td>
<td>F-test</td>
</tr>
</tbody>
</table>

But how does this relate to the real world?
A Worldly Dialog

Investigator: I think that trait A influences trait B. I recognize that there might be other traits — C, D, E, etc. — that could also be connected to A and B. How do I measure the extent to which A influences B?

Statistician: Do an Experiment. Assign A in a way that’s independent of C, D, E, etc. Then measure the association between A and B.

Investigator: Nice! But, actually, I can’t assign A. (Or, I can influence A, but C, D, E, will continue to play a role.) What do I do?

Statistician: Sorry. You’re out of luck.

Overhearing the exchange, Econometrician/Political Scientist/Epidemiologist: May I help you?

Our Abstinence Based Curriculum

"Correlation is Not Causation"
This is like saying, “sex is not love.”

"No Causation without Experimentation"
This is like saying, “No sex before marriage.” It might be a good rule, but ...

As a rule, we warn students away from making causal inferences, despite the fact that they are often beneficial to making such inferences. When we do this,

- We signal that what we are teaching is irrelevant to what students and researchers are done to.
- Failure to provide students with the tools and concepts to evaluate and criticize causal claims.

SAT Scores and School Spending

[T]he 10 states with the lowest per pupil spending included four — North Dakota, South Dakota, Tennessee, Utah — among the 10 states with the highest per pupil expenditures — Wisconsin — was among the 10 states with the highest SAT scores. New Jersey has the highest per pupil expenditures, an astonishing $10,561, which teachers’ unions elsewhere try to use as a negotiating benchmark. New Jersey’s rank regarding SAT scores? Thirty-ninth... The fact that the quality of schools... [fails to correlate] with education appropriations will have no effect on the teacher unions’ insistence that money is the crucial variable. George F. Will, (September 12, 1993), “Meaningless Money Factor,” The Washington Post, C7.

How would you deal with this example using the topics found in introductory statistics?

[Drawn from Deborah Lynn Guber (1999) JSE 7(2)]

Coverage of Confounding in Textbooks

How to measure coverage?

Simplistic approach — Count the number of pages on the topic listed in the index.

- Activities 7, 8, 3, 4, 8, 9, 10, 11, 12
- Pooled data, 3, 5, 7, 96
- For particular studies, 7, 8, 10, 11
- For proportions, 10, 11, 12
- For regression, 10, 11, 12, 13
- For contingency, 10, 11, 12
- For significance, 10, 11, 12
- For effect size, 10, 11, 12

Books Surveyed

A. Utts & Heckard, Mind on Statistics
B. Moore & McCabe, Introduction to the Practice of Statistics 4/e
C. Agresti & Franklin, Statistics: The Art and Science of Learning from Data
D. Watkins, Schofield, & Cobb, Statistics in Action: Understanding a World of Data
E. De Veaux, Velleman, & Bock, Stats Data and Models
F. McClave, Benson, & Sincich, Statistics for Business and Economics 9/e
Our Abstinence-Based Curriculum and other Stories from Statistical Education

Why Not Teach about Causation?

The standard approach in statistics education is to Abstain from Causation without Experimentation. Reasons to abstain:

- Pragmatic: It's too hard to teach the techniques: multivariate modeling, model specification, adjustment, instrumental variables, matched sampling, ...
- Proofiness: We don't want to. We can't make a provably correct statement about causation without experimentation.
- We'll teach about the things we can prove.

Some Fields Do Teach Causation

Report and Recommendations from the Working Group on Epidemiology 101. Riegelman, Fraser, French, Kaelin, Neuman, Tetelbaum, Winston, Woodin. They describe a no-prerequisite course. "Epidemiology 101." Among their learning goals:

- Explain basic statistical and epidemiologic concepts of estimation, inference, and adjustment to establish association.
- Explain how to use evidence of an association to make a judgment about whether an association is causal.
- Describe the basic epidemiologic study designs that are used to test hypotheses, identify associations, and establish causation.

Emphasizing that this should be done at a non-technical level, they say: "Epidemiology 101 should be taught without need for prerequisite courses such as statistics. Required statistical principles should be integrated into the course."

That is, let's give the statistics education.

Displaying Confounding

Unconventional Display of Confounding

From: Milo Schield (2006) "Presenting Confounding and Standardization Graphically"
Why is this Unconventional?

It's an explanation rather than a presentation of data.
- It has a lot of content that is NOT the data.
- It relies on models constructed from things other than the data.
- It's not based on one of the standard display modalities, e.g. scatter plot or two-way table.

CRAZY! Multiple regression? As an introductory topic?

Understandable reaction: It’s hard enough to teach the t-test.
I ask you to suspend disbelief for a moment and consider how you would design an introductory statistics course if the stated goal was to bring students to an understanding of confounding and the ways to deal with it (and their limitations).
- NOT, how to add multiple regression into an existing course.
- INSTEAD, how to design a course from scratch, if necessary leaving out some of the existing canon.

Building an Introductory Statistics Course around Confounding

Rather than treating confounding as peripheral and mysterious (“lurking” variables), let’s make it central. This means emphasizing techniques that can illuminate confounding rather than obscuring it.
- The essence of confounding is the existence of multiple explanatory variables.
- The t-test doesn’t qualify, since it has at most a single explanatory variable. We need multiple regression.

Three Components of an Approach

Models, not Means
Rather than calculating means of groups, use regression. Show how to interpret the coefficients as groupwise means in simple cases.

Exploration & Simulation
Give students data and let them see how results change. Give them simulations where the answer is known, and let them explore how they can recover the known result. [Example: campaign spending]

Graphics and Geometry
Use graphics that display the relationships among explanatory variables and make it easy to see why confounding occurs and how to arrange things to mitigate its impact. [Example: BK-Pbo]

Models, not Means
Illustrative example: World record times in the 100m freestyle swim race. Two explanatory variables: year and sex.

The Mean as a Model
Illustration: Regression
- Time \( y \) vs. Year
- Intercept \( 50.02 \)
Our Abstinence-Based Curriculum and other Stories from Statistical Education

Simple Regression

\[ \text{Time} \sim 1 + \text{year} \]
\[ \text{Intercept} \approx 567.24 \text{ year} \approx -0.26 \]

Groupwise Means

\[ \text{Time} \sim 1 + \text{sex} \]
\[ \text{Intercept} \approx 68.19 \text{ sexM} \approx -10.54 \]

Multiple Regression

\[ \text{Time} \sim 1 + \text{sex} + \text{year} \]
\[ \text{Intercept} \approx 555.72 \text{ sexM year} \approx -9.80 \text{ -0.25} \]

Extensions

Of course you can go on to elaborate, e.g., introducing an interaction term, etc.

Summary

I believe that statistics has a huge amount to offer students across the curriculum. But statistics education...

- Has failed to embrace its natural allies: Computation and Approximation
- The dominance of proofiness
  - Pushes us to teach formal approaches to problems that are not of general interest.
  - Prevents us from prioritizing techniques. (Why do we spend time on one- and two-sided tests? Why teach equal and unequal variance tests?)

Let’s use computation (encapsulation, packaging that makes daunting operations simple) and approximation (simulation, modeling) to allow us to teach meaningfully about causation.