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?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(Professor Emeritus of Medical Statistics, University of London)

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.' "

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

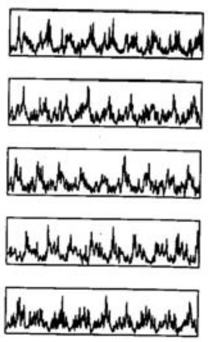
Techniques

- 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*
2. Brad Efron (1979) "Computers and the Theory of Statistics: Thinking the Unthinkable" SIAM Review, 21(4):460-480

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.

Things that surprised me when I started teaching stats

- So many tests! t-, p-, z-, paired-t, chi-squared test, rank-sum test, etc. Don't they have a common underlying logic?
- Emphasis on correlation coefficient, t, not effect size.
- Distinctions that were too subtle for students:
  - one-sided versus two-sided (or is it "tailed"?)
  - equal variance versus unequal variance t-test
  - t distribution with 10 or 20 d.f. versus normal.
  - "confidence" versus "probability"
  - p versus "the probability of the null is true"
- "No causation without experimentation."
- Simple regression, but not multiple regression.
- Significance, but not power.
- Beware of "lurking variables"!
- Bayes Rule isn't central! Support for decisions not central!

To an outsider, some problems are obvious

There's no big picture!

- Archaic and misleading language: e.g., "significance", "deviation," "error," "standard"
- Don't teach algorithms, but how to use the algorithms.
- Need measures of effect size, not just correlation. 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.

Emery Brown and Robert Kass: "What is Statistics?"

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.

[I]n emphasizing the logic of data manipulation, teachers of statistics are instilling excessive cautiousness. 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.

The American Statistician, May 2009, 63(2): 105:110

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

A Religious Attitude

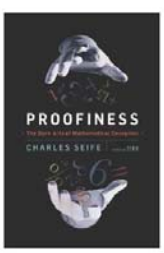
Proofiness
Emphasizing statements that are demonstrably **valid**, that is deducible from stated premises, but allowing the misconception/misinterpretation that the statements are therefore **true**, that is, applicable to the real world.
Antonym: "truthiness."

Truthiness
Truthiness is a "truth" that a person claims to know intuitively "from the gut" without regard to evidence, logic, intellectual examination, or facts. Coined/popularized by Stephen Colbert in 2005.

Quotations Critiquing Proofiness

- "Far better an approximate answer to the **right** question, which is often vague, than an exact answer to the **wrong** question, which can always be made precise." — John Tukey
- "Essentially, all models are wrong, but some are useful." — George E.P. Box
- "As far as the laws of mathematics refer to reality, they are not certain; as far as they are certain, they do not refer to reality." — Albert Einstein
- "Mathematicians do not study objects, but the relations between objects; to them it is a matter of indifference if these objects are replaced by others, provided that the relations do not change. Matter does not engage their attention, they are interested in form alone." — Henri Poincaré

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 Pitt

I see this as a special form of Proofiness: the exploitation of people's willingness to believe arguments framed using numbers, since arithmetic is always **valid**, even if it's not always true.
Synonym: **bullshit**.

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."

Two Different anti-Proofiness Attitudes

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

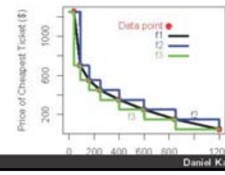
Examples of Encapsulation

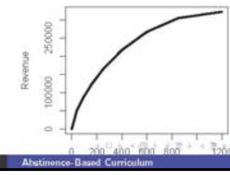
Multiplication on the computer.

```
> 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 = approxfun( demand.points, price.points,method="cons
actual.revenue = antiD( f3 )
```





An Example of Proofiness: The Alternative Hypothesis

A theoretical problem: Should the χ^2-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 "graduation 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 ...

Textbook Coverage of the Alternative

The ***alternative hypothesis*** is denoted by H_a. It states what the researchers suspect or hope to be true 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:

(a) $H_a: \theta < \theta_o$ or (b) $H_a: \theta > \theta_o$ or (c) $H_a: \theta \neq \theta_o$

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

Source: Rossman and Chance, p. 435, *Workshop Statistics: Discovery with Data* 2/e, 2001

Textbook Coverage of the Alternative

- The **alternative hypothesis**, represented by the symbol $\mathbf{H_a}$, is a statement that something is happening. In most situations, this hypothesis is what the researcher hopes to prove. It may be a statement that the assumed status quo is false, or that there is a relationship, or that there is a difference.

Each of the following statements is an example of a null hypothesis:

- There is no extrasensory perception.
- There is no difference between the mean pulse rates of men and women.
- There is no relationship between exercise intensity and the resulting aerobic benefit.

Some examples of alternative hypotheses are

- There is extrasensory perception.
- Men have lower mean pulse rates than women do.
- Increasing exercise intensity increases the resulting aerobic benefit.

Source: Utts and Heckard, *Mind on Statistics* 1/e, 2000

Textbook Coverage of the Alternative

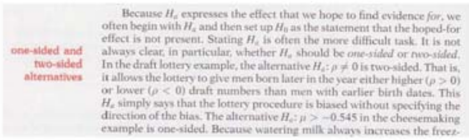
The possible null and alternative hypotheses are one of these three choices, depending on the research question:

1. $H_0: p = p_0$ versus $H_a: p \neq p_0$ (two-sided)
2. $H_0: p \geq p_0$ versus $H_a: p < p_0$ (one-sided)
3. $H_0: p \leq p_0$ versus $H_a: p > p_0$ (one-sided)

Often the null hypothesis for a one-sided test is written as $H_0: p = p_0$ instead. We will write it both ways in this chapter, to give you practice with the two ways you may encounter the null hypothesis in journal articles. Remember that a p-value is computed by assuming the null hypothesis is true, and the specific null value p_0 is what is assumed to be the truth about the population for that computation.

Source: Utts and Heckard, *Mind on Statistics* 1/e, 2000

Textbook Coverage of the Alternative


one-sided and two-sided alternatives

Because H_a expresses the effect that we hope to find evidence *for*, we often begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present. Stating H_a is often the more difficult task. It is not always clear, in particular, whether H_a should be *one-sided* or *two-sided*. In the draft lottery example, the alternative $H_a: \rho \neq 0$ is two-sided. That is, it allows the lottery to give men born later in the year either higher ($\rho > 0$) or lower ($\rho < 0$) draft numbers than men with earlier birth dates. This H_a simply says that the lottery procedure is biased without specifying the direction of the bias. The alternative $H_a: \mu > -0.545$ in the cheesemaking example is one-sided. Because watering milk always increases the freez-

"Stating H_a 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**.

Costs and Benefits to One-Sided Tests

- 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): Encourages 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.

What to Teach about Power?

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.

But What Should the Alternative Be?

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.

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 use 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."

Forming Specific Alternative Hypotheses (III)

3. The population is like the sample from my preliminary study.

- Often we collect a small amount of preliminary data to guide our future study.
- In addition to telling us about means (or coefficients), the preliminary data tells us something about standard errors.

Design your study to make your standard error adequately small.

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.

Label Type	Measurement type	Test
2 levels	quantitative	t-test
$>$ 2 levels	quantitative	ANOVA
2 levels	2 levels	p-test
$\geq$ 2 levels	$\geq$ 2 levels	χ^2 test
quantitative	quantitative	F-test

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 there are often benefits to making such inferences. When we do this,

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

The New York Times **Search**

Your Search "after controlling for" Search Advanced Search »

Today | Past 7 Days | **Past 30 Days** | Past 12 Months | All Results Since 1851

All Result Types | Articles | Multimedia 1-6 of 6

Thinking Like an Economist
Caplan and Miller found that "the estimated effect of education sharply falls **after controlling for** intelligence. ...

Grassroots Schooling in India - Room for Debate
Children in these low-cost private schools outperform children in government schools, even **after controlling for** parental wealth and ...

How Doctors Can Motivate Patients to Diet
The difference persisted even **after controlling for** age, race, sex, economic status, educational level, general health and other factors. ...

Confronting Income Inequality
For example, even **after controlling for** other factors, these counties had the largest increases in bankruptcy filings. ...

Vital Signs - 'Chemo Brain' May Be From Cancer, Not the Treatment ...
After controlling for differences between the groups, like age, education and overall health, researchers concluded that people with a ...

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 top SAT scores. Only one of 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.

Confidence level, 360–361
Confidentiality, 22, 227–228, 232
Confounding, 156, 177, 191, 398
Continuity correction, 326–327
Contrast, 643, 655–660
Control group, 181, 191, 397, 400
Correlation, 101–105
and regression, 115, 121
and scatterplots, 103–104
between random variables, 281–282, 286
cautions about, 125–137
multiple, 613–614. *See* Coefficient of determination
nonsense, 133
inference for, 591–593
population, 590
squared, 117–119, 581
test for, 591

for means, 455, 460–461, 489
paired-*t*, 516
for predicted values, 573–574, 581
for proportions, 379–380, 403, 404, 480
regression slope for, 568, 580
Student's *t* and, 455–457
Confounding, 263–264, 266
Context, 7, 11
Contingency tables
examination of, 22, 297
explanation of, 18–19, 26, 534, 539
Continuous random variables, 314–315, 324, 338
Control groups, 259, 266
Control treatments, 259
Convenience samples, 242, 247
Copernicus, 311

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, Scheaffer, & 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

Confounding-Related Terms in the Index Number of indexed pages in each text for confounding-related words.

Index Term	Confounding	Lurking Variable	Case-Control	Control Variable	**Covariate**	Causation	Simpson's Paradox
A	13	2	4	0	0	14	2
B	6	10	1	0	0	13	2
C	5	12	5	1	0	6	4
D	4	4	0	0	0	4	0
E	4	3	0	0	0	1	2
F	0	0	0	0	0	6	0

Other words that don't appear: Adjustment, Standardization.

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, Frerichs, Kaelin, Nezami, Teitelbaum, 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 skip the statistics education.

Displaying Confounding

Dominant modes of presenting confounding: stories, two-way tables, scatter plots.

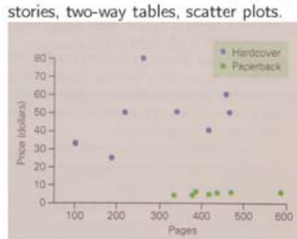


FIGURE 3.1 Price versus pages for two types of books

Utts & Heckard, p. 59

Displaying Confounding

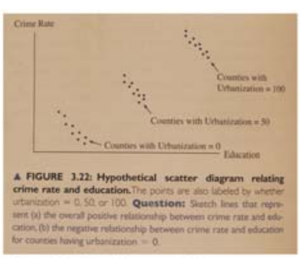
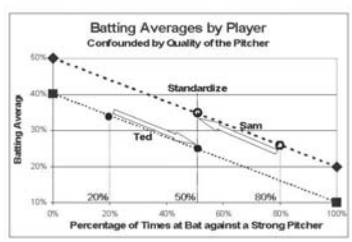


FIGURE 3.22: Hypothetical scatter diagram relating crime rate and education.

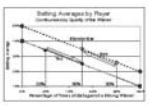
Agresti & Franklin, p. 130

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.

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.

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.

Three Components of an Approach

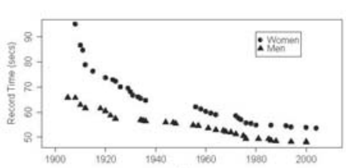
Models, not Means
Rather than calculating means of groups, use regression. Show how to interpret the coefficients as groupwise means in simple cases.

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.

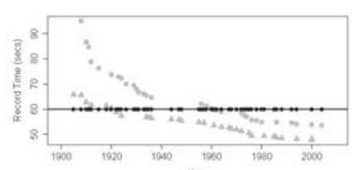
Graphics and Geometry
Use graphics that display the relationships among explanatory variables and make it easy to see why confounding occurs and how to arrange things to mitigate its impact. [Example: BK-Plot]

Models, not Means

Illustrative example: World record times in the 100m freestyle swim race. Two explanatory variables: year and sex.

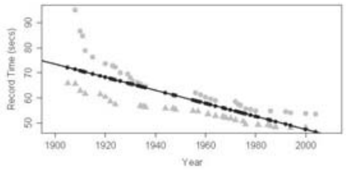


The Mean as a Model



Time ~ 1
Intercept
59.92

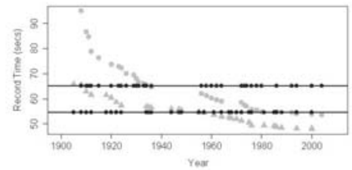
Simple Regression



Time ~ 1 + year

Intercept	year
567.24	−0.26

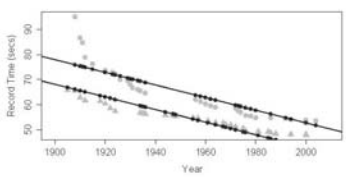
Groupwise Means



Time ~ 1 + sex

Intercept	sexM
65.19	−10.54

Multiple Regression



Time ~ 1 + sex + year

Intercept	sexM	year
555.72	−9.80	−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.