Statistical Thinking

Teaching Statistics
in Five Unnatural Acts

Paul F. Velleman
Cornell University
pfv2@cornell.edu
“A variant of Gresham’s law applies: things that are cheap to assess, like computational skills, tend to drive out things that are of actual value, like statistical thinking.”

-- George Cobb

…Things that can be learned by rote tend to drive out things that require reasoning.
Statistics is the art of learning about the world from data.

But it is always a challenge to teach an art. There's a way of thinking about the world, about the data, and about how they relate to each other that informs good statistics.

Teaching how to think statistically is as difficult as teaching how to write well or how to draw a draped figure.
We must confront the reasons why it is so hard to teach students to think statistically.
Five Unnatural Habits of Mind Essential for Statistics

1. Think Critically.
2. Be Skeptical.
3. Focus on what we don’t know.
4. Think first about Variation.
5. Think clearly about Conditional and Rare chance events.
Statistics is Subversive

Those who wish to hustle us out of our money, take our votes without our thought, or control us don’t want us to think critically, be skeptical, acknowledge our ignorance (and demand the missing facts) or look too closely at rare or compound chance events. Students who learn to do so make better citizens.

These habits ask for an effort, they challenge us to take responsibility for our choices, and they don’t come without hard thought.
Good News:

Sophomores love being subversive. So we are catching them at the right time: If only we can help them to see how truly dangerous this course can be, they’ll love it.
Think Critically

• Challenge the data’s credentials.
• Look for bias.
• Know what we want to know.
• Look for Lurking variables.
• Check Assumptions and Conditions.

Critical thinking requires creativity. You must think about things that are not in front of you and imagine ways in which things might have gone wrong.
“You can use all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgment. “

--- Alvin Toffler
Be Skeptical

• Be cautious about making claims based on data.
• “Trust every analysis, but plot the residuals.” Skeptical statisticians expect the unexpected, so we go looking for it.
• The three rules of data analysis:
  – I Plot the data
  – II Plot the data, and
  – III Plot the data
The vote in the 2000 Presidential election for Buchanan and the vote for Nader, (the two principal alternatives to Bush and Gore), has a correlation of 0.65 over the counties of Florida.

Ask:
  Linear?
  Any Outliers?
  Subgroups?
Without Palm Beach county and its “butterfly ballot”, the correlation is 0.91.
Hypothesis Testing is Formalized Skepticism

- The null hypothesis is a skeptical claim about the data—that’s actually what makes it the null (not a parameter value of zero.)
Critical Thinking and Skepticism

• Critical thinking is open-ended questioning of the data’s credentials.
  – We wonder whether the data are competent to tell us what we want to know.

• Skepticism questions whether what the data appear to be telling us is the whole truth.
Focus on What We Don’t Know

• We naturally prefer to think about what we do know, and we tend to focus on positive information.

• But that’s not the best way to gather information.
For Example:

A stock analyst claims that *when she says the market will go up, it does*. Pick 2 items to judge her claim.

<table>
<thead>
<tr>
<th>You see</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict Up</td>
<td>Predict Down</td>
<td>Market Up</td>
<td>Market Down</td>
<td></td>
</tr>
<tr>
<td>You can learn</td>
<td>Market result</td>
<td>Market result</td>
<td>Prediction</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Confidence Intervals Focus on What We Don’t Know:

• We don’t say “The mean is 3”.
• We don’t say “The mean is probably 3”
• We don’t say “The mean is close to 3”.
• All we can manage is “The mean is close to three…. Probably
  – (but I may be wrong – and I’m willing to spend the effort to give you a whole interval of plausible values and then to spend extra effort to estimate how likely it is that even that interval is wrong.)”
All Models are Wrong…

Statisticians love models--*because they are wrong.*

When we fit a model, where do we look? Do we admire the model and show it off? No way. We examine the residuals!

To a Statistician the most interesting thing about a model is often what it fails to account for.
Think about Variation

We find it easier to think about an estimated value than to think about the underlying variation.

But

*Statistics is about Variation*
For Example:

A town has two hospitals. In the larger one, about 45 babies are born each day. In the smaller hospital, about 15 babies are born each day. Over a period of a year, would one hospital, the other, or neither have more days on which the babies born there were more than 60% male?
The Standard Deviation is the Statistician’s Ruler

• Most of the inference seen in the introductory course compares a statistic to its standard deviation to see whether it is “big”.
• This idea carries into advanced methods as well.
Thinking about Conditional Events

• This is just plain hard.
• It is easy to show that we don’t naturally think clearly about conditional probabilities.
• But we must for rational decision making.
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and she participated in antinuclear demonstrations.
Rank order the following in terms of the likelihood that they are true of Linda:

a) Linda is a teacher in an elementary school
b) Linda works in a bookstore and takes yoga classes.
c) Linda is active in the feminist movement.
d) Linda is a psychiatric social worker
e) Linda is a member of the League of Women Voters.
f) Linda is a bank teller.
g) Linda is an insurance salesperson.
h) Linda is a bank teller who is active in the feminist movement.
Which is More Likely?

• A massive flood somewhere in North America in 2004, in which 1,000 people drown.

Or

• An earthquake in California during 2004, causing a flood in which more than 1,000 people drown.
Rare events and Salience

Worldwide, in 2001
Fewer airline passengers died in accidents than during an average year

-- including the deaths of air passengers on 9/11.
Is Statistical Thinking Unnatural?

We haven’t evolved to be Statisticians. Our students who think Statistics is an unnatural subject are right. This isn’t how humans think naturally.

But it is how humans think rationally. And it is how scientists think. This is the way we must think if we are to make progress in understanding how the world works and, for that matter, how we ourselves work.
Teaching Critical Thinking

**Think**

- For all data, know the W’s (*Who, What, When, Where, Why, how*) If the Who and What are not clear. Stop; you can't analyze these data.

- Think about how the data were collected. Was the survey random? Was the experiment randomized, blinded, and placebo-controlled? Are there likely lurking variables?
  - Be especially wary of voluntary response, such as internet-based surveys. They are worthless.

- Think about Assumptions and Conditions.
Teaching Skepticism

Show

- Skepticism is a natural trait in students. We only need to focus and direct it.
- The basic tool is graphics. Plot the data, plot the data, plot the data. *Expect the unexpected.*
- Enumerate what kinds of patterns to look for and what kinds of deviations from them are interesting. These patterns form the bases for the inference methods we’ll teach later in the course.
Teaching a Focus on Ignorance

Tell

• When we point out this aspect of confidence intervals, it often helps students to see how they work.

• The task’s not done until we write a sentence or two that summarizes and explains our conclusions in terms of the measurements (and their units).

• Discussions of confidence intervals must carefully assign the uncertainty to ourselves, not to the parameter we hope to trap in the (random) interval.
Teaching about Variation

Simulate

• Technology makes this a lab course.
• W.S. Gosset (Student) had to simulate by actually drawing slips of paper from a hat. Our students can reproduce his study instantly.
• Students can discover the Law of Large Numbers, the Central Limit Theorem, the reasoning of Hypothesis testing, and the construction of confidence intervals for themselves. And when they do that, they understand and remember.
Sample Distribution
\( \hat{p} = \text{Proportion Red} \)
Teaching about Chance

(We may get Beyond the Formulas, but we can’t forget them)

• Students need to know:
  – A (frequentist) definition of probability.
  – The basic rules for combining probabilities
  – Conditional probabilities and Independence

• General multiplication rule, the definition of independence, the general addition rule, and definition of disjoint events.

• Tree diagrams can help as well.
44% of college students engage in binge drinking, 37% drink moderately, and 19% abstain entirely.

Another study finds that among binge drinkers aged 21-34, 17% have been involved in an alcohol-related automobile accident, while among non-bingers of the same age, only 9% have been involved in such accidents.

What’s the probability that a student who has had an alcohol-related accident is a binge drinker?

Almost 70%
Taking Responsibility

There is an overall theme to these unnatural acts. Each asks us to take responsibility, to think “outside the box,” to be willing to say “this is what I think might be the case, but I may be wrong.”

Much of current education has emphasized self-esteem more than personal responsibility, so this may be a shock to students. Nevertheless, personal responsibility is the foundation of wisdom.
Carving the Introductory Statistics Course

Paul F. Velleman
Cornell University
Data Description, Inc.
pfv2@cornell.edu

I saw the angel in the marble and carved until I set it free.

-- Michaelangelo
It is far too easy to add new topics to the Introductory Statistics Course. There’s always one more method or technique that someone thinks should be in the course. But that threatens to bury the essence of the course in trivia.
Advanced classes are defined by what they include.
Introductory classes are defined by what they exclude.

There is an interesting, subtle, beautiful introductory statistics course, but it is often obscured by topics, notation, terminology, and tedium that don’t belong.
The Program for Today

I propose to sculpt the introductory statistics course by carving away the parts that are not needed and then polishing what’s left behind into a simple presentation order so that we can set free the angel.
Audiences

• **College students** taking a first statistics class. Even if they know some calculus, they don’t really want us to use it. And many know no more than high school algebra.

• **High School AP statistics students**. They may be a bit more enthusiastic about the class – it is an elective rather than a required course. But then, they’re high school seniors…
Cutting away the big blocks:
Get out the sledgehammer

• Probability first foundations. Emphasize data rather than theory.
• Hand calculation (use technology), and along with it formulas for calculation – choose formulas for understanding.
Understanding or Computation?

\[
b = \frac{n(\sum x_i y_i) - (\sum x_i \sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \]

\[
b = r \frac{S_y}{S_x}
\]
Chipping away other things not needed:

- Data “types” (Nominal, ordinal, interval, ratio).
- Subscripts that count from 1 to $n$.
- $z$ for means (and those rules for when to switch from $t$ to $z$).
- Critical value approach to testing.
- “Population formulas” for standard deviation and other stats. Divide by $n-1$. 
The Angelic Parts

- **Models** are fundamental from the very beginning. Models are wrong so that they can be useful.
Models...

Require Assumptions
Because they are idealized, they are only really true under idealized assumptions.

Are described by Parameters
Parameters refer to models of populations, not to the populations themselves.
The Angelic Parts

• Models are fundamental from the very beginning. Models are wrong so that they can be useful.
• Differentiate Assumptions from Conditions, and check the latter (often graphically)
Conditions to Check

- Random? - How were the data gathered?
- Independent? Just have to think about it
- (Nearly) Normal? - make a picture
- Linear? - make a picture
- Constant Variance? - make a picture
The Angelic Parts

• **Models** are fundamental from the very beginning. Models are wrong so that they can be useful.

• Differentiate **Assumptions** from **Conditions**, and check the latter (often graphically)

• The **standard deviation** as a ruler.
Standard Deviation

- Often we speak of variation as if it is a problem. We speak of making decisions *in spite of variation*.
- But Statistics doesn’t just see past variation; we *employ* it. The Standard Deviation becomes our principal tool for understanding the magnitude of differences and effects.
The Angelic Parts

• **Models** are fundamental from the very beginning. Models are wrong so that they can be useful.

• Differentiate **Assumptions** from **Conditions**, and check the latter (often graphically)

• The **standard deviation** as a ruler.

• The central value of **randomizing**.
Randomizing

• To many, “random” feels like haphazard; out of control; chaotic.

• As with the standard deviation, Statistics takes what could be a challenge and harnesses it as a workhorse.

• The *deliberate application of randomness* is one of the central insights of Statistics.
The Central Role of Randomization

- Random sampling
- Randomization in experiments
- Simulation as a tool for Understanding
  - Statistics as an experimental science.

- *Without Randomization, your inferences don’t have a chance*
  
  » George Cobb
Soft Curves

Statistics deals with the real world, not MathWorld.

In MathWorld the bicyclist rides in a straight line along a level road at a constant speed without encountering any air resistance past a tree that stands perpendicular to the ground.

No wonder there is a “right answer”.

In Statistics we liberate ourselves from the absurdity of MathWorld, but that liberty comes at a price. Uncertainty and caution and approximation will accompany everything we do. The real world is much more complicated – and much more interesting – than MathWorld.

But we must deal with Vague Concepts
Vague Concepts

- Representative (sample)
- Bias
- Symmetric/Skewed
- Unimodal/Bimodal/Multimodal
- Center
- Spread
- Association
- Outlier
At What Point are Data Paired?

• Equal number of 20-year old men and women.
• One man and one woman surveyed on each day of the study
• Both from the same state
• Well, actually from the same town
• Well, now that you ask, from the same college
• … same statistics class
• And the question was about their opinion of the value to them of the material
• … in the lecture they just attended…
• …And they left the class holding hands
• … and heard each other’s responses
Clean Lines: Don’t Say…

• **Measure of central tendency.** Without the LLN, there is no tending. And even with it, that’s not what we’re measuring.

• **Empirical rule.** David Moore’s proposal to call it the **68-95-99.7 Rule** puts the important facts in the name where they might be remembered.
...and Don’t Say...

- **Independent/Dependent Variables.** Confusing terms that are used elsewhere to mean other things.
- **Best** model (e.g. Best fitting line…) “The best is the enemy of the good.” We should be willing to entertain alternative models. (After all, they are all wrong.)
Say, But not Too Soon:

• Don’t talk about the population at the start. Indeed, don’t even mention it until much later. Inference is not the (only) goal of Statistics.

• Don’t talk about samples early. Data are interesting in their own right.

• Don’t say probability, say relative frequency until we’re well into the course.

• And, of course, nothing about drawing inferences until we know what to look for in data.
The Pedestal: Why Statistics?

We often do a poor job of answering the question that our students have that supports the entire course:
“Why am I taking this course?”

“It’s required (shut up and sit down)”
“Why am I taking this course?”

“To learn tools and techniques that will enable you to make **inferences** from **samples** to **populations** as long as you have a sufficiently large **simple random sample** or a properly randomized **experiment**, with **independent, identically distributed errors**, and a clear idea of what **parameters** you are estimating…”
“Why am I taking this course?”

“Because I’m a sadist: You’re about to be confused, made to feel stupid, and bored for an entire term. You haven’t a prayer of getting the right answers on the exams without a lot of memorizing (shut up and sit down.)”
"Why am I taking this course?"

"Data are interesting. And they are interesting because they help us to understand the world –
You’re taking the course; to learn the **Art of Understanding the World through Data**
(Stand up and tell us what you see.)"
Order

• Now that we have a suitable sculpture, we must help our students construct their own models of it. This is more like building with clay than like carving marble.

• To be learned, each new concept must attach to other concepts already in place.

• That requires careful attention to the order in which concepts are introduced.
Foundations

• Graphing and what to look for
• Summary statistics, especially the standard deviation (as a ruler)
• Standardizing, 68-95-99.7 rule, Normal model
• Association, correlation, regression models
• Randomness and simulation
• Gathering data with randomization
• Basic probability, LLN
• CLT
<table>
<thead>
<tr>
<th>Topic</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. One Proportion</td>
<td>CI concept</td>
</tr>
<tr>
<td>2. One Proportion</td>
<td>HT concept</td>
</tr>
<tr>
<td>3. 2 Proportions</td>
<td>Variances add</td>
</tr>
<tr>
<td>4. One Mean</td>
<td>$t$ - models</td>
</tr>
<tr>
<td>5. Difference of Means</td>
<td>2-sample $t$</td>
</tr>
<tr>
<td>6. Mean of Differences</td>
<td>Matching</td>
</tr>
<tr>
<td>7. Dist’n of Categorical Data</td>
<td>$\chi^2$ Goodness of fit</td>
</tr>
<tr>
<td>8. Association (Categorical)</td>
<td>$\chi^2$ Independence</td>
</tr>
<tr>
<td>9. Association (Quantitative)</td>
<td>Regression</td>
</tr>
<tr>
<td>10. Association (mixed)</td>
<td>ANOVA</td>
</tr>
</tbody>
</table>
Step back and View the Whole

The introductory statistics course should be a course in how to think. Students should learn to think critically about data. They should be vigilant in their skepticism about what the data may be saying. They should understand the central importance of variation.
Taking it Home

- Students should take home a complete piece, not a pile of chips.
- That understanding is far more important than any particular analysis skill they may learn.