

Integrating Statistical Thinking and Ethics into Teaching, Consulting and Daily Life



*Professor Jessica Utts
Department of Statistics
University of California, Irvine
March 15, 2018*



Original Title (AP Statistics talk)

For Advanced Placement Statistics high school teachers and introductory statistics college teachers:

What do Future Senators, Scientists, Social Workers, and Sales Clerks Need to Learn from Your Statistics Class?



Why this talk is for you

- If you are a teacher:
 - Most people will take at most one Statistics class in their lives.
 - If they are your students, you have that one chance to teach them how to make informed decisions!
- If you are a practicing statistician:
 - Many people you interact with will have had at most one statistics course, and need guidance in applying what they learned (or didn't!).
- If you are someone who took at most one statistics class... here is what you need to know!



My Top 10 Important Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Sample size and statistical significance
4. Why many studies fail to replicate
5. Does decreasing risk actually increase risk?
6. Personalized risk
7. Poor intuition about probability and risk
8. Using expected values to make decisions
9. Surveys and polls – good and not so good
10. Confirmation bias

A (Partially True) Story

- Senator Blumberg is interested in children's health issues, and sees this (real) headline:

“Breakfast Cereals Prevent Overweight in Children”



- The article continues:

“Regularly eating cereal for breakfast is tied to healthy weight for kids, according to a new study that endorses making breakfast cereal accessible to low-income kids to help fight childhood obesity.”



Hmm, Senator Blumberg Thinks...

- Maybe I should introduce the Blumberg Cereal Bill to make breakfast cereal available to low-income children throughout the United States! They would all lose weight! I would be a hero!
- But Senator Blumberg remembers some cautions from her statistics class and decides to investigate a bit more.
- What is revealed?



Some Details

- This was an observational study
- 1024 children, only 411 with usable data
 - Mostly low-income Hispanic children in Austin, TX
 - Control group for a larger study on diabetes
- Asked what foods they ate for **3 days**, in each of grades 4, 5, 6 (same children for 3 years)
- Study looked at number of days they ate cereal = **0 to 3** each year (Frosted flakes #1!)



More Details: The analysis

- Multiple regression was used
 - Response variable = BMI percentile each year (BMI = body mass index)
 - Explanatory variable = days of eating cereal in each year (0 to 3), modeled as linear relationship with BMI!
- Did not differentiate between other breakfast or no breakfast (for days without cereal)
- Also included (adjusted for) age, sex, ethnicity and some nutritional variables



Uh-oh, Some Problems!

Topic #1: Confounding variables

- Observational study – no cause/effect.
- Obvious possible confounding variable is general quality of nutrition in the home
 - Unhealthy eating for breakfast (non-cereal breakfast or no breakfast), probably unhealthy for other meals too.
- High metabolism could cause low BMI and the need to eat breakfast. Those with high metabolism require more frequent meals.

Recall what the story said:

“Breakfast Cereals Prevent Overweight in Children”

- The article continues:

“Regularly eating cereal for breakfast is tied to healthy weight for kids, according to a new study that endorses making breakfast cereal accessible to low-income kids to help fight childhood obesity.”

- Notice that the *quote* does not imply cause and effect, but the **headline** does.





Senator Blumberg Knew to Ask:

- Who did the study?
 - Lead author = Vice President of Dairy MAX, a regional dairy council. (Fair disclosure: Study funded by NIH, not Dairy MAX)
- What was the size of the effect?
 - Reduction of just under 2% in BMI percentile for each extra day (up to 3) of consuming cereal (regression coefficient was -1.97)
- So the Cereal Bill died before it left Senator Blumberg's desk!



Who Else Needs to Know How to Evaluate Such Studies?

- Scientists – understand how to conduct study and report results.
- Social workers – if the program had been mandated for low income kids, how important is compliance?
- Sales clerks – does it matter if her/his kids eat cereal for breakfast?
- In other words, everyone!



More of my Favorite Headlines

- “6 cups a day? Coffee lovers less likely to die, study finds”
- “Oranges, grapefruits lower women's stroke risk”
- “Yogurt Reduces High Blood Pressure, says a New Study”
- “Walk faster and you just might live longer”
 - “Researchers find that walking speed can help predict longevity”
 - “The numbers were especially accurate for those older than 75”



Assessing possible causation

Some features that make causation *plausible* even with observational studies:

- There is a reasonable explanation for how the cause and effect would work.
- The association is consistent across a variety of studies, with varying conditions.
- Potential confounding variables are measured and ruled out as explanations.
- There is a “dose-response” relationship.



Another Story (also partially true)

- Mr. Buckley has a daughter.
- He would like to have a son.
- So he asks his wife if she would please eat cereal for breakfast. Not because she's fat...
- But because he saw a news story in a reputable outlet, from a reputable journal



More about Cereal: Does it Produce Boys?

- Headline in *New Scientist*: “Breakfast cereal boosts chances of conceiving boys” Numerous other media stories of this study.
- Study in *Proc. of Royal Soc. B* showed of pregnant women who ate cereal, 59% had boys, of women who didn't, 43% had boys.
- Problem #1 revisited:
Headline implies eating cereal *causes* change in probability, but this was an observational study. (Confounding variables???)



Topic #2: Multiple Testing

- The study investigated 132 foods the women ate, at 2 time periods for each food = 264 possible tests! (Stan Young pointed this out in a published criticism.)
- By chance alone, *some* food would show a difference in birth rates for boys and girls.
- Main issue: Selective reporting of results when many relationships are examined, not adjusted for multiple testing. Quite likely that there are “false positive” results.



Common Multiple Testing Situations

- *Genomics*: Looking for genes related to specific disease, testing many thousands.
- *Diet and disease*: For instance, ask patients and controls about many dietary habits.
- *Interventions (e.g. Abecedarian Project)*:
 - Randomized study gave low-income kids (infant to kindergarten) educational program (or not).
 - Kids in program were almost 4 times as likely to graduate from college. (Many other differences; too many to all be multiple testing.)



Multiple Testing: What to do?

- There are statistical methods for handling multiple testing. See if the research report mentions that they were used.
- See if you can figure out how many different relationships were examined.
- If *many* significant findings are reported (relative to those studied), it's *less likely* that the significant findings are false positives.



Yet Another Story

- There is a planet similar to earth, Planet PV
- On that planet, babies are only allowed to be born in the spring.
- No one knows about the beneficial effects of taking aspirin to prevent heart attacks.
- Lots of other false notions from statistical studies (even more than here!).
- Why? Because on Planet PV everything is decided by p -values!



On Planet PV, They Read This Headline

Spring Birthday Confers Height Advantage

- Austrian study, heights of 507,125 military recruits.
- Test of difference in mean heights for men born in spring versus fall found tiny p-value
- Men born in spring were, on average, about 0.6 cm taller than men born in fall, i.e. about 1/4 inch (Weber et al., *Nature*, 1998, 391:754–755).
- Sample size so large that even a very small difference was *highly statistically significant*.



Does Aspirin Prevent Heart Attacks?

Physicians' Health Study

5-year randomized experiment

22,071 male physicians (40 to 84 years old)

$\chi^2 = 25.4$, p -value ≈ 0

| Condition | Heart Attack | No Heart Attack | Attacks per 1000 |
|-----------|--------------|-----------------|------------------|
| Aspirin | 104 | 10,933 | 9.42 |
| Placebo | 189 | 10,845 | 17.13 |

But on Planet PV, $n = 2207$ instead, same rates

So $\chi^2 = 2.54$, p -value = .111, not significant!



Topic #3: Role of sample size in statistical significance

- The p -value does *not* provide information about the *magnitude/importance* of the effect.
- If sample size **large** enough, almost **any null hypothesis can be rejected**.
- If the sample size is **too small** it is very hard to achieve statistical significance (low power)
- Don't equate statistical significance with whether or not there is a real, important effect.
- If possible, get a confidence interval.



Hypothesis testing paradox:

- Researcher conducts test, $n = 100$, finds $t = 2.50$, p -value = 0.014, reject null hypothesis ($t = \frac{\bar{x} - 0}{s/\sqrt{100}}$)
- Just to be sure, repeats with $n = 25$
- Uh-oh, finds $t = 1.25$, p -value = 0.22, cannot reject null! The effect has disappeared!
- To salvage, decides to combine data, so $n = 125$. Finds $t = 2.795$, p -value = 0.006!
- Paradox: The 2nd study *alone* did not replicate finding, but when *combined* with 1st study, the effect seems even stronger than 1st study!

What's going on?

- Both studies have the same effect size! $e.s. = \bar{x}/s$
- Combined data also has that effect size
- The value of the test statistic and p -value depend on the sample size through \sqrt{n} .
- Effect size is t / \sqrt{n} and $t = \sqrt{n}$ (effect size)

| Study | n | Effect size | Test stat | P -value |
|----------|-----|-------------|-----------|------------|
| 1 | 100 | 0.25 | 2.50 | 0.014 |
| 2 | 25 | 0.25 | 1.25 | 0.22 |
| Combined | 125 | 0.25 | 2.795 | 0.006 |



Why Effect Sizes are Important

- Unlike p -values, they don't depend on sample size (but accuracy of estimating them does).
- They are a measure of the true effect or difference in the population.
- They can be compared even when different units or different tests are used.
- Replication should be defined as getting approximately the same effect size, *not* as getting approximately the same p -value!



Topic #4: Conflicting Results of Studies

Ioannidis (2005) looked at replication:

- 45 high-impact medical studies in which treatments were found to be effective
 - Each published in top medical journal, and had been cited more than 1000 times
 - Studies were repeated with same or larger size, and same or better controls for 34 of them.
- How many do you think replicated original result of effective treatment? All? Most?



Conflicting results, continued

- The 45 studies included 6 observational studies and 39 randomized controlled trials.
- Replication results:
 - Only 20 of the 45 attempted replications were successful (i.e. found the same or better effect)
 - Of the 6 observational studies, 5 found smaller or reversed effects (83%).
 - Of the 39 randomized experiments, 9 found smaller or reversed effects (23%).



Possible explanations

Ioannidis suggests these explanations:

- Confounding variables in observational studies
- Multiple testing problems in the original studies
- Multiple researchers looking for a positive finding; by chance alone, someone will find one



Other possible reasons

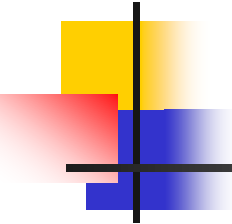
- Different conditions or participants
 - It's difficult and not very interesting to conduct an exact replication.
- Original study was surprising, replications are not. No incentive to publish successful replications.
 - Instead, surprising “non-replications” are published.
 - So there is a replication “file drawer” effect!



Topic #5:

Avoiding Risk May Put You in Danger

- In 1995, UK Committee on Safety of Medicines issued warning that new oral contraceptive pills “increased the risk of potentially life-threatening blood clots in the legs or lungs by twofold – that is, by 100%” over the old pills
- Letters to 190,000 medical practitioners; emergency announcement to the media
- Many women stopped taking pills.



Clearly there is increased risk, so what's the problem with women stopping pills?

Probable consequences:

- Increase of 13,000 abortions the following year
- Similar increase in births, especially large for teens
- Additional \$70 million cost to National Health Service for abortions alone
- Additional deaths and complications probably *far exceeded* pill risk.



Actual Risk versus Relative Risk

- “Twofold” risk of blood clots:
 - 1/7000 to 2/7000, not a big change in absolute risk, and still a small risk.
- *Absolute* risk is what is important:
 - 2/7000 likely to have a blood clot
 - Compare to other risks of pregnancy
- But *Relative* risk (2 in this case) is what makes news!



Topic #6: Reported Risk versus Your Risk

“Older cars stolen more often than new ones”

Davis (CA) Enterprise, 15 April 1994, p. C3

- Of the 20 most popular auto models stolen in California the previous year, 17 were at least 10 years old.
- Many factors determine which cars stolen:
 - Type of neighborhood.
 - Locked garages.
 - Cars not locked and/or don't have alarms.



Topic #6: Reported Risk versus Your Risk

- The real question of interest is:
If I were to buy a new car, would my risk of having it stolen increase or decrease over my old car?
- Article gives no information about that question.



Considerations about Risk

- Changing a behavior based on relative risk may *increase* overall risk of a problem. Trade-offs!
- Find out what the *absolute* risk is, and consider relative risk in terms of additional *number* at risk
- Suppose a behavior doubles risk of cancer
 - Brain tumor: About 7 in 100,000 new cases per year, so adds about 7 cases per 100,000.
 - Lung cancer: About 75 in 100,000 new cases per year, so adds 75 per 100,000, more than 10 times as many!
- Does the reported risk apply to you?
- Over what time period? (Per year? Per lifetime?)



Topic #7: Poor intuition about probability and risk

- William James was first to suggest that we have an *intuitive* mind and an *analytical* mind, and that they process information differently.
- Example: People feel safer driving than flying, when probability suggests otherwise.
- Psychologists have studied many ways in which we have poor intuition about probability assessments.
 - Recommended reading: *Thinking, Fast and Slow* by Daniel Kahneman



Example: Confusion of the Inverse

Gigerenzer gave 160 gynecologists this scenario:

- About 1% of the women who come to you for mammograms have breast cancer (bc)
- If a woman has bc, 90% chance of positive test
- If she does not have bc, there is only a 9% chance of positive test (false positive)

A woman tests positive. What should you tell her about the chances that she has breast cancer?



Answer choices: Which is best?

- The probability that she has breast cancer is about 81%.
- Out of 10 women with a positive mammogram, about 9 have breast cancer.
- Out of 10 women with a positive mammogram, about 1 has breast cancer.
- The probability that she has breast cancer is about 1%.



Percent who chose each answer

- The probability that she has breast cancer is about 81%." 13% chose this
- Out of 10 women with a positive mammogram, about 9 have breast cancer. [i.e. 90% have it] 47% chose this
- Out of 10 women with a positive mammogram, about 1 has breast cancer. [i.e. 10% have it] 21% chose this
- The probability that she has breast cancer is about 1%. 19% chose this



What is the Correct Answer?

Let's look at a hypothetical 100,000 women.
Only 1% have cancer, 99% do not.

| | Test positive | Test negative | Total |
|-----------|---------------|---------------|------------|
| Cancer | | | 1,000 (1%) |
| No cancer | | | 99,000 |
| Total | | | 100,000 |



Let's see how many test positive

90% who have cancer test positive.

9% of those who don't have it test positive.

| | Test positive | Test negative | Total |
|-----------|---------------|---------------|---------|
| Cancer | 900 (90%) | | 1,000 |
| No cancer | 8910 (9%) | | 99,000 |
| Total | 9810 | | 100,000 |

Complete the table for 100,000 women:

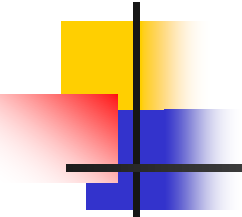
| | Test positive | Test negative | Total |
|-----------|---------------|---------------|---------|
| Cancer | 900 | 100 | 1,000 |
| No cancer | 8910 | 90,090 | 99,000 |
| Total | 9810 | 90,190 | 100,000 |

Correct answer is $900/9810$, just under 10%!

Physicians confused two probabilities:

$P(\text{positive test} \mid \text{cancer}) = .9$ or 90%

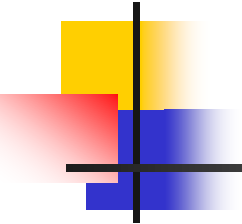
$P(\text{cancer} \mid \text{positive test}) = 900/9810 = .092$ or 9.2%



Confusion of the inverse: Other examples

Cell phones and driving (2001 study):

- Given that someone was in an accident:
 - $P(\text{Using cell phone}) = .015$ (1.5% on cell phone)
 - $P(\text{Distracted by another occupant}) = .109$
(10.9% gave this reason)
 - Does this mean other occupants should be banned while driving but cell phones are okay??
- $P(\text{Using cell phone}|\text{accident}) = .015$
- But what we really want is
 - $P(\text{Accident}|\text{cell phone})$,
 - Much harder to find; need $P(\text{Cell phone})$



Confusion of the inverse: DNA Example

- Dan is accused of crime because his DNA matches DNA at a crime scene (found through database of DNA). Only **1 in a million** people have this specific DNA.
- Suppose there are **6 million** people in the local area, so about **6 have this DNA**. Only one is guilty!
- Is Dan almost surely guilty??



DNA Example continued

- Remember, only 6 people with this DNA out of 6 million people
- $P(\text{DNA match} \mid \text{Dan is innocent})$
 ≈ 5 out of 6 million, extremely low!
 - *Prosecutor would emphasize this*
- But... $P(\text{Dan is innocent} \mid \text{DNA match})$
 ≈ 5 out of 6, fairly high!
 - *Defense lawyer should emphasize this*
- Jury needs to understand this difference!

The Conjunction Fallacy: Survey Question

Plous (1993) presented readers with this test:

Place a check mark beside the alternative that seems most likely to occur within the next 10 years:

- An all-out nuclear war between the United States and Russia
- An all-out nuclear war between the United States and Russia in which neither country intends to use nuclear weapons, but both sides are drawn into the conflict by the actions of a country such as Iraq, Libya, Israel, or Pakistan.

Survey in my class: Using your intuition, pick the more likely event at that time.

44/138 = **32%** chose first option – CORRECT!

94/138 = **68%** chose second option – Incorrect!

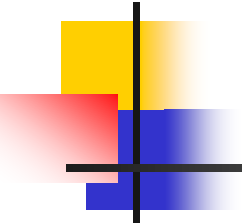
The Representativeness Heuristic and the Conjunction Fallacy

- **Representativeness heuristic:** People assign higher probabilities than warranted to scenarios that are *representative* of how they *imagine* things would happen.
- This leads to the **conjunction fallacy** ... when detailed scenarios involving the conjunction of events are given, people assign *higher* probability assessments to the *combined event* than to statements of one of the simple events alone.
- Remember that $P(A \text{ and } B) = \text{can't } \underline{\text{exceed}} P(A)$



Other Probability Distortions

- Coincidences have higher probability than people think, because there are so many of us and so many ways they can occur.
 - UCI Statistics Department story of 13s
 - The one in a million event
- Low risk, scary events in the news are perceived to have *higher* probability than they have (readily brought to mind).
- High risk events where we think we have control are perceived to have *lower* probability than they have.



Topic 8: Expected Values: A (partially) true story

Ashley S. lives outside of Washington DC and is going to hear a presentation of the WSS in DC. If the weather is bad, she will stay at a hotel. She looks at hotels and finds a room with the following:

- Pay \$170 now, nonrefundable *OR*
- Pay \$200 when she arrives, but only if she needs the hotel
- What should she do? What additional information would help her decide?



Expected value for her decision

- Define p = probability she needs the hotel.
- Expected costs for each decision:
 - If she pays advance purchase, $E(\text{Cost}) = \$170$
 - If she doesn't pay in advance
$$E(\text{Cost}) = \$200(p) + \$0(1 - p) = \$200p$$
- Which is lower?
 $\$200p < \170 when $p < (170/200) = 0.85$.
- Decision: Pay advance purchase if $p > 0.85$, but not otherwise.



Insurance, lottery, extended warranty

Should you buy an extended warranty?

What about insurance? (e.g. earthquake?)

- *On average* the company wins
- But *some* consumers will be winners, and some will be losers.
- You can use knowledge of your own circumstances to assess which is likely for you.



Understanding Expected Value: Survey Question (my class)

Which one would you choose in each set?
(Choose either A or B and either C or D.)

- A.** A gift of \$240, guaranteed
- B.** A 25% chance to win \$1000 and a 75% chance of getting nothing.
- C.** A sure loss of \$740
- D.** A 75% chance to lose \$1000 and a 25% chance to lose nothing



Survey Question Results

Which one would you choose in each set?
(Choose either A or B and either C or D.)

85%

A. A gift of \$240, guaranteed

15%

B. A 25% chance to win \$1000 and a 75% chance of getting nothing.

30%

C. A sure loss of \$740

70%

D. A 75% chance to lose \$1000 and a 25% chance to lose nothing



The Amount Makes a Big Difference

Which one would you choose in each set?

A. A gift of \$5, guaranteed

B. A 1/1000 chance to win \$4000

Now 75% chose B.

This is like buying lottery tickets.

C. A sure loss of \$5

D. A 1/1000 chance of losing \$4000

Now 80% chose C.

Like buying insurance or extended warranty.



Probability, Intuition, Expected Value

Examples of Consequences in daily life:

- Assessing probability when on a jury
Lawyers provide detailed scenarios – people give higher probabilities, even though *less* likely.
- Extended warranties and other insurance
“Expected value” favors the seller
- Gambling and lotteries
Again, average “gain” per ticket is negative
- Poor decisions (e.g. driving versus flying)



Topic #9: Surveys and polls

Most of you probably know about common problems, such as:

- Biased wording posing as objective surveys
- Confusing wording and/or possible responses
- Problems with getting a representative sample, and getting people to respond
- Responses given with desire to please or give socially acceptable answers

Let's look at some subtle examples...

Wording is Important and Difficult to Get Right!

Small change of words can lead to big change in answers.

Example: How Fast Were They Going?

Students asked questions after shown film of car accident.

- About how fast were the cars going when they contacted each other?
Average response = 31.8 mph
- About how fast were the cars going when they collided with each other?
Average response = 40.8 mph

Ref: Loftus & Palmer, *Journal of Verbal Learning and Verbal Behavior*

Ordering of Questions

The order in which questions are presented can change the results.

Example:

1. How happy are you with life in general?
2. How often do you normally go out on a date?
About _____ times a month.

Almost no correlation in answers. When order was *reversed*, there was a strong correlation! Respondents seem to think the happiness question was now, “Given what you just said about going out on dates, how happy are you?”

Ref: Clark and Schober, *Questions about Questions*, J. Tanur, Ed.



Topic #10: Confirmation bias – ethical issue, or argument for Bayes?

- People tend to give more credence to statistical results and data that support their beliefs
- Effect is stronger for emotionally charged and deeply held beliefs
- I have seen this when I present data showing strong statistical evidence for psychic abilities
- Ethical issue: Should all data be treated equal?
- Or let's just admit that we are all Bayesians, and we combine data with prior beliefs even if we think we are objective statisticians!



Again: My Top 10 Important Topics

1. Observational studies, confounding, causation
2. The problem of multiple testing
3. Sample size and statistical significance
4. Why many studies fail to replicate
5. Does decreasing risk actually increase risk?
6. Personalized risk
7. Poor intuition about probability and risk
8. Using expected values to make decisions
9. Surveys and polls – good and not so good
10. Confirmation bias



QUESTIONS?

Contact info:

jutts@uci.edu

<http://www.ics.uci.edu/~jutts>

UCIrvine

UNIVERSITY OF CALIFORNIA, IRVINE
