How to Lie with Statistics

Supplementary Material for CFB3333/PHY3333 Professors John Cotton and Stephen Sekula March 23, 2012 Based on the following information on the web: http://www.physics.smu.edu/pseudo/LieStat

Resources

- Huff, Darrell. "How to Lie with Statistics"
 - first published in 1954
 - some of the examples show their age, but they still very effectively communicate the tricks and traps of statistics
- Statistics what is it?
 - very simply: it is the study of the collection, organization, and interpretation of data
 - used correctly, it's a powerful tool in interpreting the results of an experiment
 - used incorrectly, or misunderstood, it's a powerful tool for manipulating people to get them to agree with you

Digression about Elections

- There is no perfect vote counting system
 - as a result, every vote counting system MUST have an inherent uncertainty (e.g. statistical or systematic, where "systematic" errors are errors of measurement)
- In 2000, President George W. Bush and Vice President Al Gore ended their bids for the Presidency in Florida
 - With other states too close to call, Florida's 25 electoral votes were the "prize to win" to seal victory
 - Bush's lead over Gore was less than 2000 votes, and in one recount narrowed to as little as 300 votes
- This is the first election in U.S. history where the margin of victory for electoral votes was essentially within some measure of uncertainty on the actual vote count.

Confusion over Palm Beach County ballot



Sun-Sentinel graphic/Daniel Niblock



"Proving" a Coin is Biased

- We did this on Monday
 - You "know" that the probability of flipping a coin and getting heads is 50/50
 - But that means that in a large (e.g. infinite) number of coin flips, the number of heads will equal 50% of the total flips
 - In a small set of trials, the chance of getting heads 7,8,9 times is not small and can happen
 - Seeing "biased coins" in a small sample of trials is an example of "cherry picking" data to suit your opinion or ideology. In a small enough number of trials, you can find all kinds of data that appears to support your notions.

Distributions

- You are dealing with a population of data
 - e.g. pilot salaries, or factory worker salaries, incomes in a neighborhood, etc.
- You are asked to summarize the data in some way
- The "Average" is a very common way to do this
 - but . . . which average? There are 3 kinds!
 - Mean, Median, and Mode are all "averages," but can all have different meanings depending on the data

Averages

- Mean: the "arithmetic mean" is when you add up all the numbers in the population and DIVIDE the sum by the total number of data points
- Median: the value such that half of the numbers in the population lie below, and half above, that value ("the middle")
- Mode: the number that appears MOST FREQUENTLY in the population

Example			
Salary	Mean	Median	Mode
\$8,000	\$37,727	\$14,000	\$23,000
\$10,000			
\$11,000			
\$12,000			
\$12,000			
\$14,000			
\$23,000			
\$23,000			
\$23,000			
\$23,000			
\$256,000			

When does it matter?

- When data are distributed according to THE NORMAL DISTRIBUTION (also known as "the bell curve") then it DOESN'T MATTER whether you quote mean, median, or mode as "the average" they are all basically the same number.
- Otherwise, you need to know which average is being used. Skewed distributions, like those salaries, can be interpreted VERY differently depending on whether we use mean, median, or mode.





Extrapolation

- This is when you use past behavior of a data sample to infer future behavior
- "I've seen this pattern before, and it's going to happen again."
 - a very common stock broker philosophy
 - it's also usually dead wrong
- Except when well-defined laws are at work in the control of the data outcomes, even if they are probabilistic, extrapolation can be a dangerous and/or deceptive technique.



Shown are times (in seconds) measured for the fastest mile runners (y-axis) plotted against the days since Dec. 30, 1899. They appear to decrease linearly, so I fit a trend line to them (a straight line). Extrapolation of the data would suggest that by around the year 2500, humans will be able to run a mile in ZERO SECONDS.











Foam impact experiment, at speeds estimated from video of strike on actual shuttle. Resulting damage. Piece hitting Columbia was 400 times bigger than any previous observed strike – outside experience of foam strike models.

Post-hoc Thinking

- Post Hoc Ergo Propter Hoc Latin for, "After this, therefore because of this."
- Data are collected after some event; the event is assumed to cause the outcomes in the data
- Darrell Huff uses 1950s college statistics on men and women:
 - 93% of middle-aged Cornell male graduates were married
 - 65% of middle-aged Cornell female graduates were married
 - Conclusion: college is bad for a woman's chance of marrying!
 - is there an alternative explanation of the data?

College Makes You Less Religious?!

• Senator Rick Santorum cited this statistic recently:

He claimed that "62 percent of kids who go into college with a faith commitment leave without it," but declined to cite a source for the figure. [CBS News. Political Hotsheet Blog. Feb. 23, 2012.]

• Any thoughts on this? Anybody know what is wrong with this kind of post hoc thinking?

What the study actually says

 The study in question was written by Mark Regnerus and Jeremy Uecker, and published on Feb. 5, 2007 in the journal "Social Forces." http://sf.oxfordjournals.org/content/85/4/1667.short

• It finds that:

- If you attended college and get a bachelors degree, your odds ratio of disaffiliating from a religious institution is about 1.3 meaning there is a 1.3 x 50% = 65% chance that you stop affiliating with a religious institution.
- However, the study finds that if you DID NOT attend college, your odds ratio is 1.6! That means a 1.6 x 50% = 80% chance of disaffiliation!