Lies, Damned Lies and Statistics

Chapter 22

Cartoon by Jeff Mallett, May 8, 2006

- “If you can’t prove what you want to prove, demonstrate something else and pretend they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anyone will notice the difference” *David Huff*
- “If you can’t dazzle them with brilliance, baffle them with BS” (modern paraphrase)

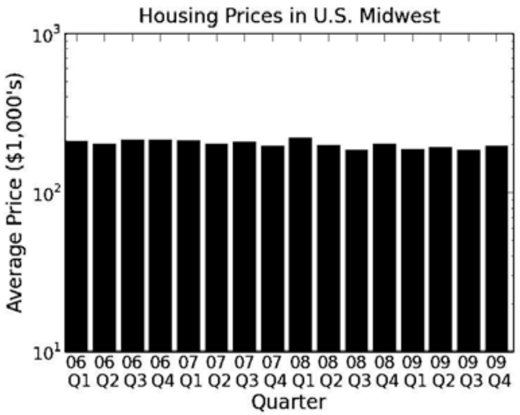
GIGO – Garbage In – Garbage Out

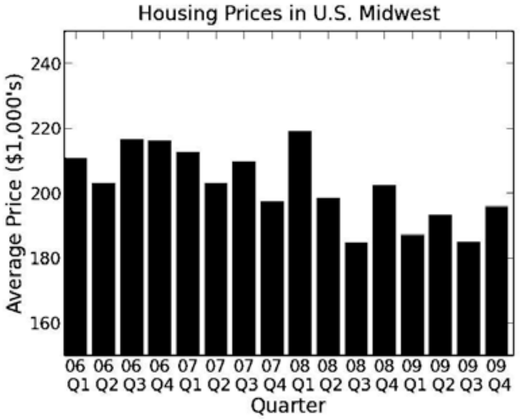
- If your incoming data is flawed, no amount of manipulation can produce a meaningful result
  - Researcher bias
  - Bad data collection
  - Bad sampling techniques
- **Assumption of Independence**
  - Errors will balance each other out

Tests are imperfect

- False positives and negatives
- Design can be wrong
- Samples insufficient
- Measurements inexact and insufficient

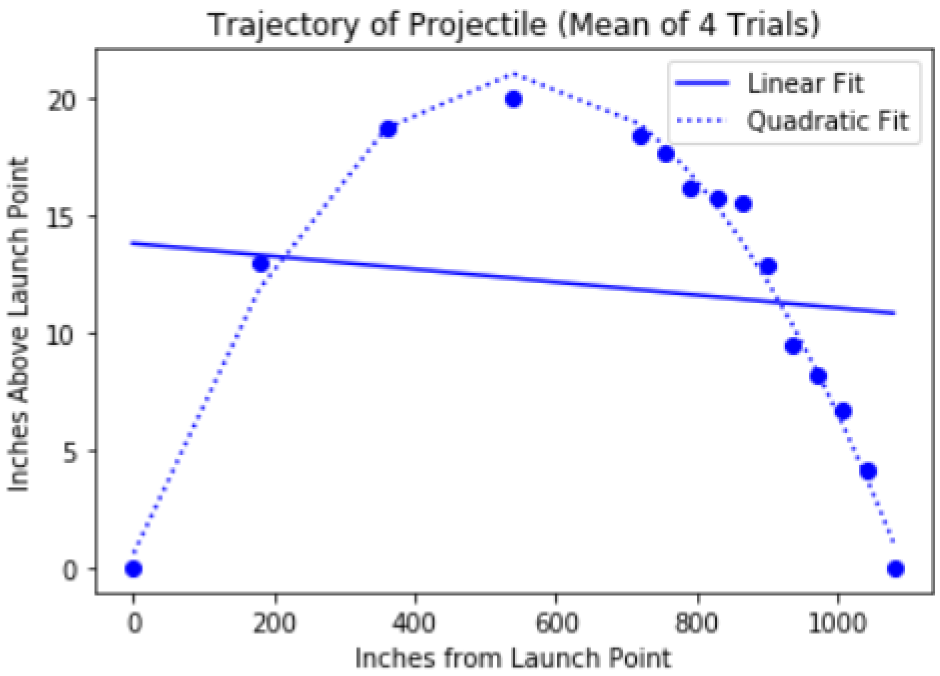
Pictures can be deceiving

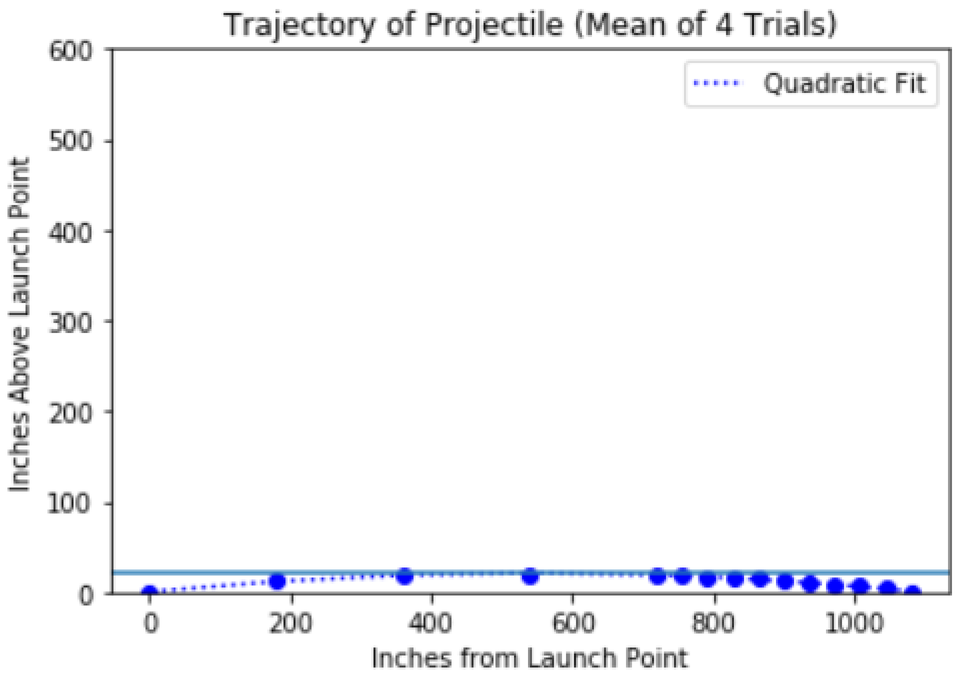




- Moral: Look at Axes, labels and scales

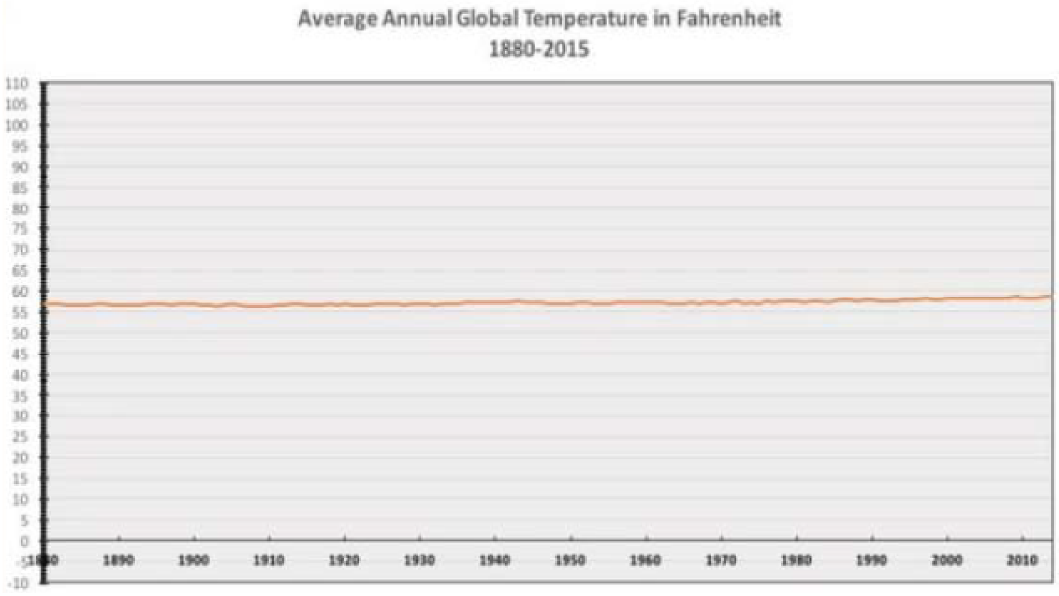
Remember this ...

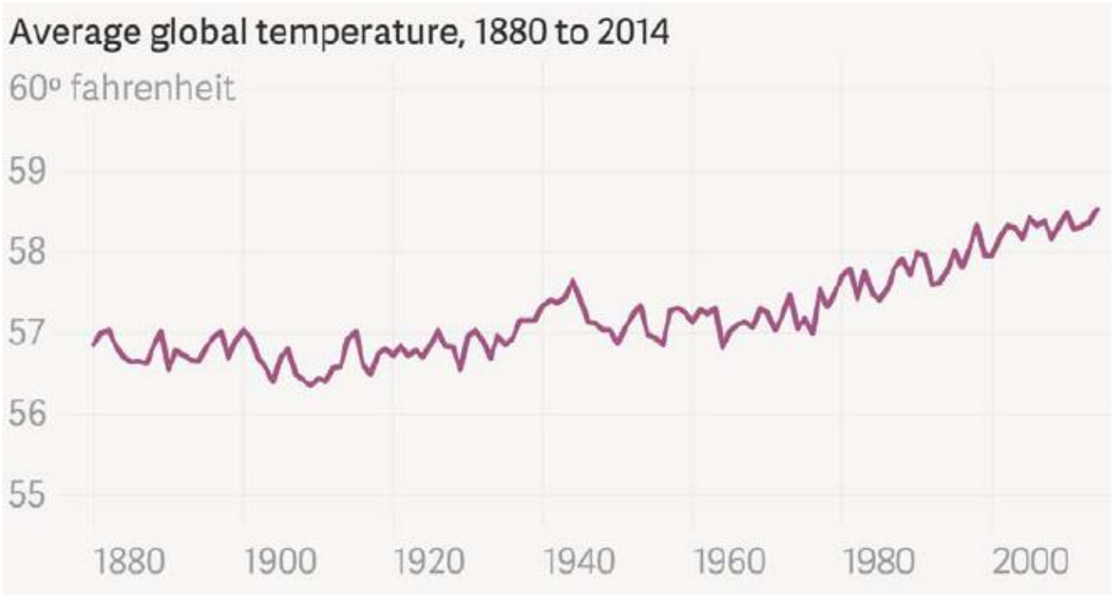




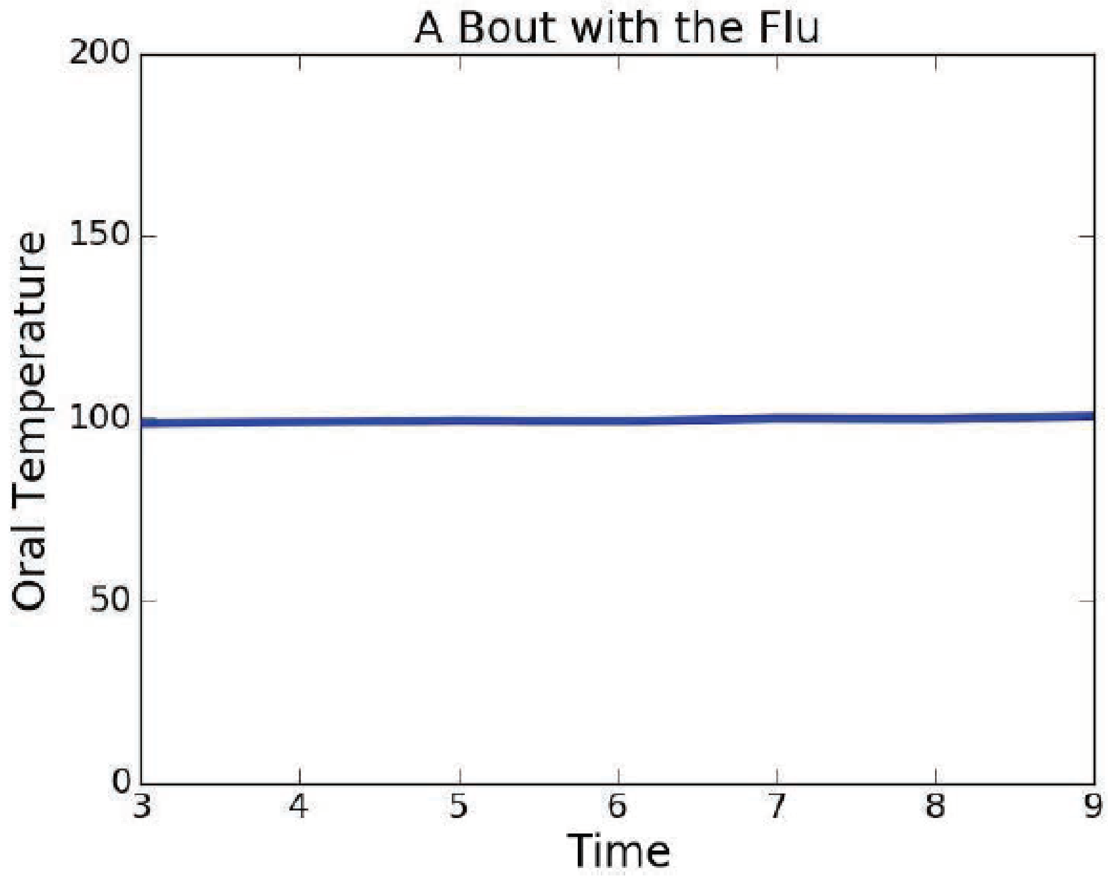

```
21.0485515522 inches= 0.534633209426 meters
Time to impact 0.330204225184 seconds
Horizontal speed =  41 m/sec
Horizontal speed =  149 km/hr
Terminal velocity = 5 m/sec
```

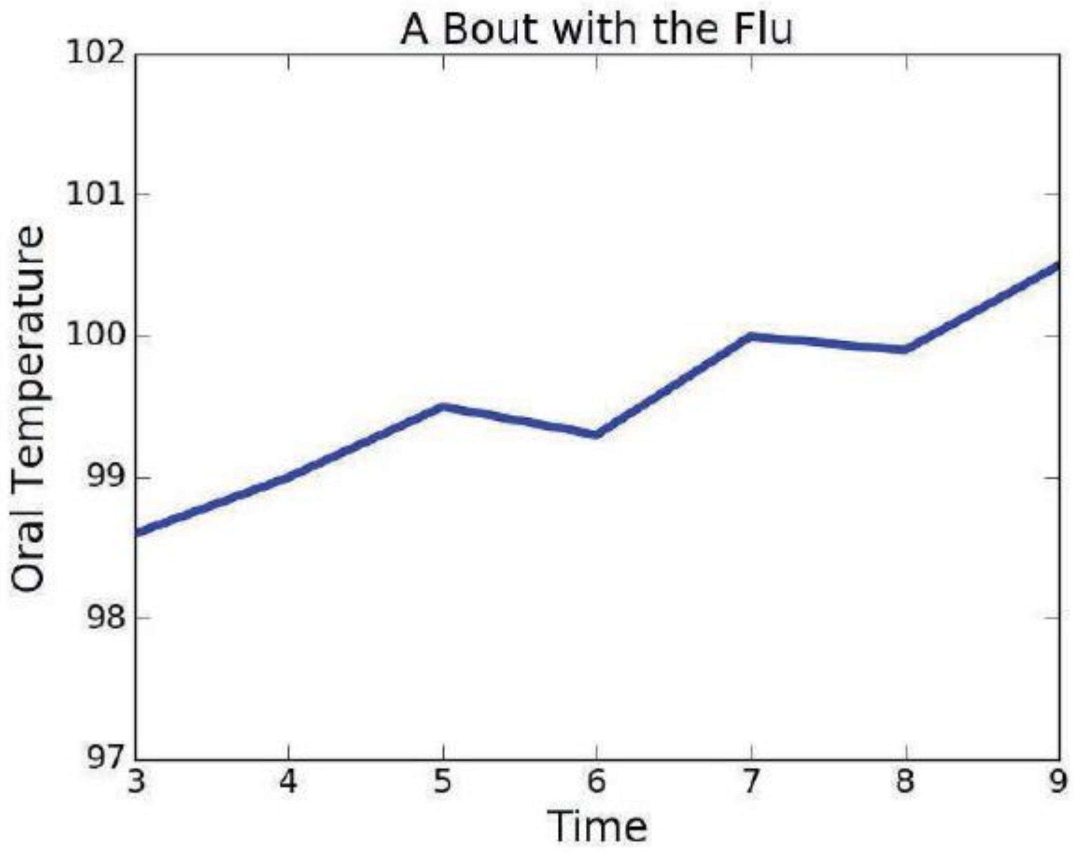
Global warming





Temperature increase with the flu





Some good references for information visualization

- INFO H517 – Visualization Design, Analysis, and Evaluation
- The Visual Display of Quantitative Information, Tufte, Edward R.
  - www.edwardtufte.com
- Storytelling with Data: A Data Visualization Guide for Business Professionals, Knaflic, Cole N.

Con hoc ergo propter hoc

With this, therefore because of this

- Just because data are **correlated** it does not mean that one **caused** the other.
  - Correlation is not causation

  - Flue does not cause school
  - Ice cream does not cause murder
- There can be **lurking** or more commonly called **confounding** variables
- Or sometimes they are not related at all
  - Years that end in 0 do not cause American presidents to die
  - The winner of a sporting event has no effect on elections
    - Redskins Rule

Statistical measures don’t tell the whole story

- Wait, why did we spend all that time studying statistics?
- Because they still matter but we may need to look beyond the “summary statistics” to the data

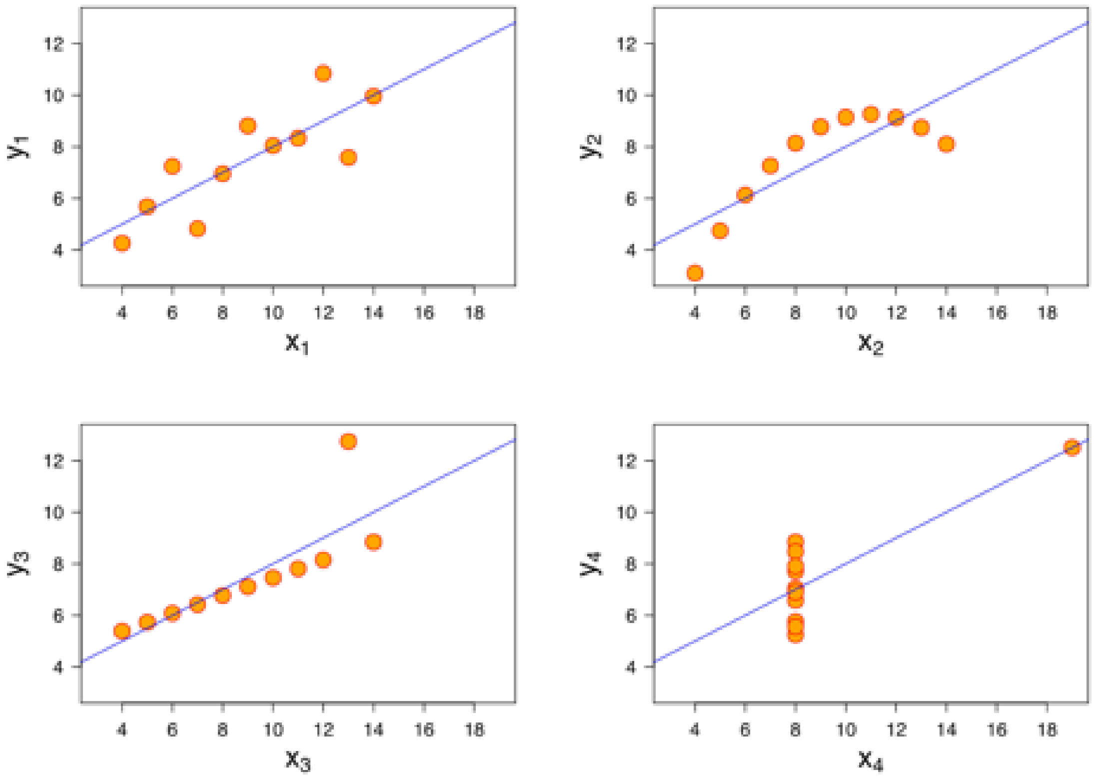
Anscombe’s quartet

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- Summary statistics for groups are identical
  - Mean x = 9.0
  - Mean y = 7.5
  - Variance of x = 10.0
  - Variance of y = 3.75
  - Linear regression model: y = 0.5x + 3
- Are the four data sets really similar?

Anscombe's quartet displayed



- **Moral: Statistics about the data is not the same as the data**
- **Moral: Use visualization tools to look at the data itself**

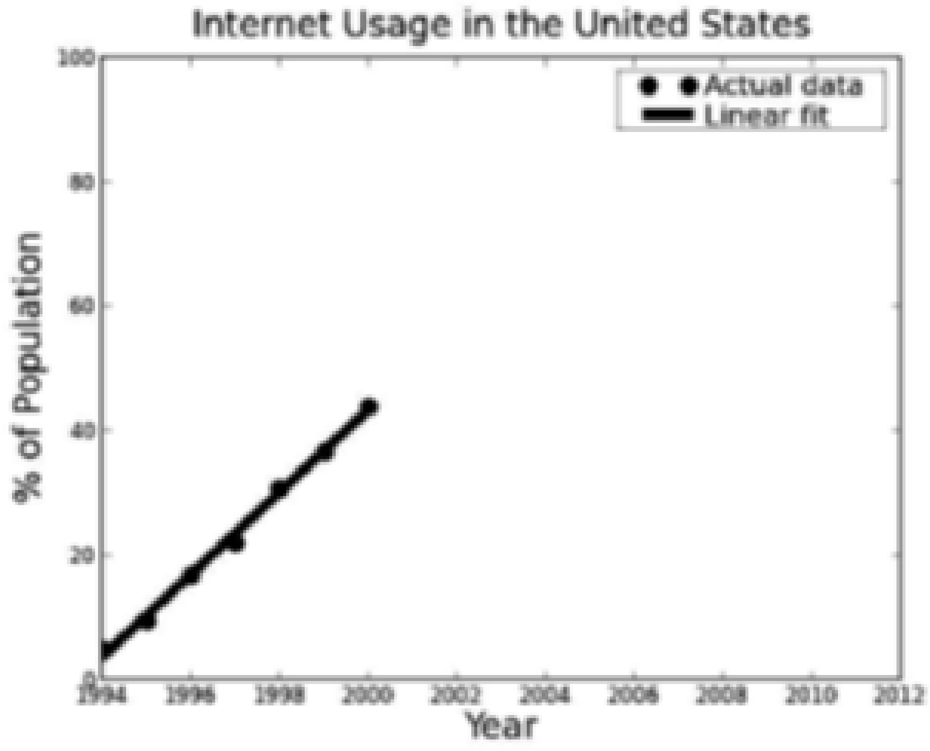
Sampling bias

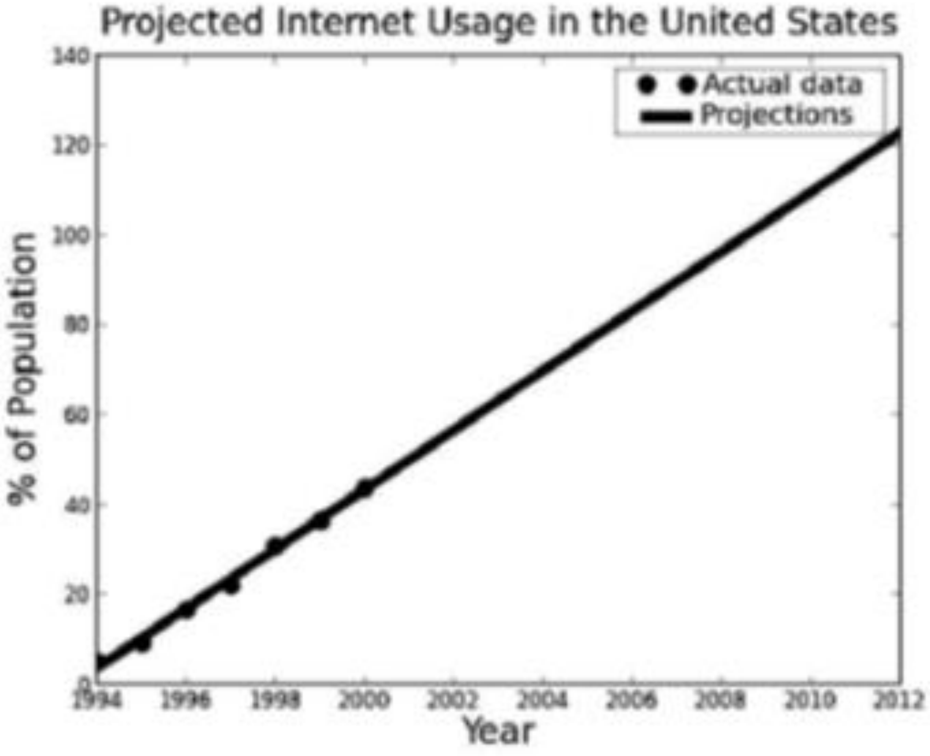
- All statistical techniques are based upon the assumption that by sampling a subset of a population we can infer things about the population as a whole.
- As we have seen, if random sampling is used, one can make meaningful mathematical statements about the expected relation of the sample to the entire population.
- Easy to get random samples in simulations.
- Not so easy in the field, where some examples are more convenient to acquire than others.

Non-representative bias

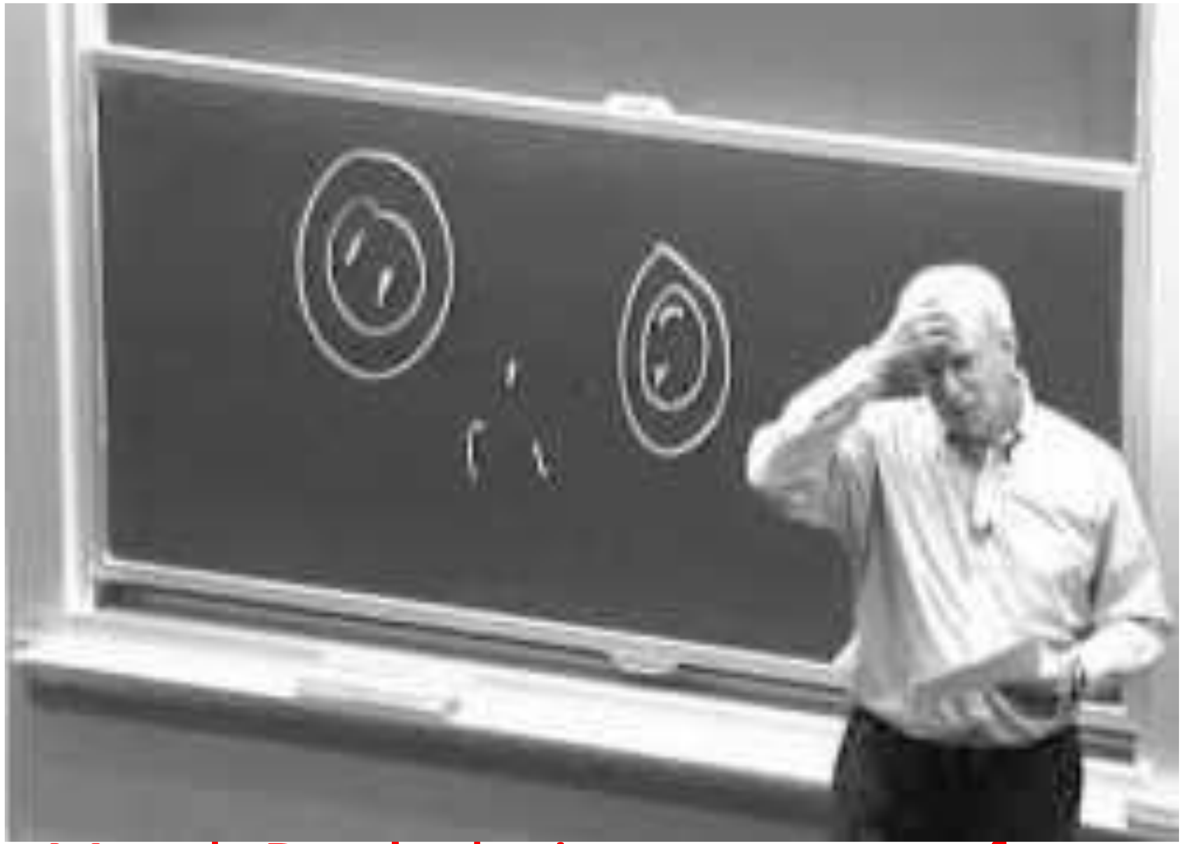
- “Convenience sampling” not usually random, e.g.,
  - Reviews are usually good or bad, not many with an ‘okay’ experience take time to submit a rview
- Non-response bias, e.g., opinion polls conducted by mail or online
- When samples not random and independent, we can still do things like compute means and standard deviations, but we should not draw conclusions from them using things like the empirical rule and central limit theorem.
- **Moral: Understand how data was collected, and whether assumptions used in the analysis are satisfied. If not, be wary.**

Beware of extrapolation





Texas sharpshooter fallacy



- Moral: Don’t design your test after you collect the data

Percentages can confuse

- Always know the basis of the percentage
  - A 16% gain after a 15% loss is not a 1% gain
  - 100 – 15% = 100 – 15 = 85
  - 85 + 16% = 85 + (85*0.16) = 85 + 13.6 = 98.6
- When stores offer multiple discounts find out what discounts what

“Statistically significant” differences may not be significant

- Differences may lie within standard error
- Sample sizes may not be large enough to show true nature of a relation

Regression fallacy

- **Regression to the mean** After an extreme event the next event is likely to be close to the mean (chapter 15)
    - We may tend to ignore the fact that extreme events occur and count on our "lucky" (pencil, socks, seat, song, etc…)

A final word or three

- Skepticism is warranted when drawing inferences from data
  - Not the same as denial
- Present the data – don’t explain it
  - Do explain the collection – don’t hide bad input
- Be sure appropriate statistical tests are applied and the data is appropriate for the test
- When comparing data compare their units
- Does the data reflect reality
- Can you present the analysis such that it has explanatory power