Lying with statistics

Chau Tran & Chun-Che Wang
Why lying?

- People (consciously) lie all the time
  - You should learn to detect lies
- People make unconscious lie, too
  - Making good arguments using statistics is hard
- Learn tricks to make your results more interesting
  - Only works if your work is inherently interesting
Outline

1. Lying with numbers
2. Lying with visualization
Lying with numbers

“The typical Brown graduate earns an average mid-career salary of $119,000”

Lying with numbers

“The typical Brown graduate earns an average mid-career salary of $119,000”

In how many ways can this be a lie?

Sampling bias

Is there any bias in the sample set?
“The typical Brown graduate earns an average mid-career salary of $119,000”
Sampling bias 1

● More successful graduates are more likely to respond to surveys
  ○ They feel good about their earnings
  ○ Surveys are only sent to big companies

● How big is their sample size?
  ○ 1% shouldn’t be called ‘typical’
“The typical Brown graduate earns an average mid-career salary of $119,000”
Sampling bias 2

- **Tendency to exaggerate**
  - Brag
  - School spirit
- **Tendency to minimize**
  - No one likes tax
- **Do they cancel out each other?**
  - No one knows!
The well chosen “average”

What statistics are they actually talking about?
The well chosen “average”

“The typical Brown graduate earns an average mid-career salary of $119,000”
The term “average”

- Average can be mean, median, or mode
  - They can be totally different
- Mean: Evenly distributes the total among individuals
  - Can be unrepresentative when measurements are highly skewed
- Median: Value dividing distribution into two equal parts (50th percentile)
- Mode: Most frequently observed outcome (rarely reported with numeric data)

Slide by Larry Winner, “Overview of How to Lie with Statistics”
Imagine a school with 4 alumni

- A: $1 million/year
- B: $120k/year
- C: $100k/year
- D: $80k/year
- E: $80k/year

- Mean = $276k/year
- Mode = $80k/year
- Median = $100k/year
“The typical Brown graduate earns an average mid-career salary of $119,000”
Correlation vs Causation

What conclusions can you make from this data?

- Does going to Brown make you rich?
Correlation doesn’t imply causation

- Hypothesis 1: Brown only admits smart students, who will make a lot of money no matter what
- Hypothesis 2: Brown only admits rich students, who will inherit a lot of money no matter what
- ...
- Hypothesis n
Lying with numbers - Summary

Three questions you should ask after you read any paper:

1. Is there any bias in the sample set?
   a. Look for unconscious bias
   b. Look for conscious bias

2. What statistics are they actually talking about?

3. What conclusions can we make from their findings?
Lying with visualization

HOW CAN THIS BE TRUE?

Below the four parts are moved around

The partitions are exactly the same, as those used above

From where comes this "hole"?

Lying with visualization

"don't believe everything you see."
Lying with Line Chart

Zero line at the bottom
Chop off the bottom
Change the portion of y-axis
Lying with Diagram

- Say that in a pond, there were
  - 13 Adult frogs in May
  - 39 Adult frogs in September
- Represented in a “stacked-frog” plot
Lying with Diagram

- or we can represent in this way...

Number of Adult Frogs in South Pond

http://www.physics.csbsju.edu/stats/display.html
Lying with Bar charts

V.S

Frog Population

May  Sept

May  Sept

40
30
20
10
0
Some more example

- People at “A” get twice pay than people at “B”
Lying with Maps

- Choropleth map
  - provides an easy way to visualize measurement varies across geographic area
  - Darker-means-more rule

http://upload.wikimedia.org/wikipedia/commons/1/17/World_population_density_map.png
Lying with Choropleth map

- US poverty map

Pic from Guardian data blog
Lying with Choropleth map

- Poverty data range from 6.6% to 22.7%
  - Unequally distributed

If we are measuring inequality, perhaps we should at least use equally distributed classes

http://vis4.net/blog/posts/choropleth-maps/
Lying with Choropleth map

- Choice of colors

http://vis4.net/blog/posts/choropleth-maps/
Lying with Choropleth map

- With equally distributed classes and equidistant colors from a HSV gradient

http://vis4.net/blog/posts/choropleth-maps/
Lying with Choropleth map

● When look at any choropleth map, be aware of
  ○ How they categorize the classes
  ○ How they choose the colors

● Choropleth map classification based on
  ○ Equal-intervals
  ○ quantile classing
    ■ each class has equal number quantity
  ○ Iterative algorithm to find “natural breaks”
Today’s Example (Ad from Verizon)

http://www.techdirt.com/articles/20091103/1818276787.shtml
Today’s Example (Ad from Verizon)

- AT&T actual coverage map (2009)

[Map image]

Today’s Example (Apple WWDC)

http://www.theguardian.com/technology/blog/2008/jan/21/liesdamnliesandstevejobs
Today’s Example (Apple WWDC)

U.S. Smartphone Marketshare

- RIM: 39.0%
- Apple: 21.2%
- Palm: 9.8%
- Motorola: 7.4%
- Nokia: 19.5%
- Other: 3.1%

http://www.theguardian.com/technology/blog/2008/jan/21/liesdamnliesandstevejobs
Lying with Visualization - Summary

1. Choice of ranges on graphs can have huge impact on interpretation
2. Choice of proportion of y-axis and x-axis can distort as well
3. Can distort bar charts by having them start at positive values
4. Pictorial graphs should have areas proportional to values

Slide by Larry Winner, “Overview of How to Lie with Statistics”
Question?