
Data

Basic Tools:

- **average/mean:** sum of the values divided by the number of values
- **median**: the middle value when the data is sorted
  - If the number of values is even, the median is the average of the two middle values
- **mode**: value that occurs most often. Can have multiple modes.

Example of Basic Tools

http://demonstrations.wolfram.com/MeanMedianMode/

More Tools!

- **standard deviation**: how much variation or dispersion from the average exists (example).
- **range:** the difference between the largest and smallest values
- **frequency:** the rate at which something occurs or is repeated over a particular period of time or in a given sample.

Ways to Display Data

http://vimeo.com/29862153

Definitions:

- **scatter plots:** The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis
- **best fit lines:** A line on a graph showing the general direction that a group of points seem to be heading.
- **histogram:** A histogram is a bar graph that shows how frequently data occur within certain ranges or intervals. The height of each bar gives the frequency in the respective interval.

Visualization Tools

- Excel
  - Simple Excel Guide
- Google Fusion Tables
  - demo
- Snap! Data Lab
  - lab exercise

Demos!

- scatter plot demo
- Histogram Demo: visual example of Irwin-Hall Distribution
- Best Fit Line Demo: If every 100 meter freestyle olympic raced against each other


Pitfalls in Data

Overgeneralization

- *Definition*: Given a few samples, generalize to entire populations.
- Example
- Anyone else have an example?

Cause VS. Correlation

Cause: X causes Y ≠ Y causes X

Correlation:
X and Y are correlated = Y and X are correlated

Therefore:
X correlates with Y ≠ X causes Y
X correlates with Y ≠ Y causes X
Y correlates with X ≠ X causes Y
Y correlates with X ≠ Y causes X

Cause VS. Correlation

Example

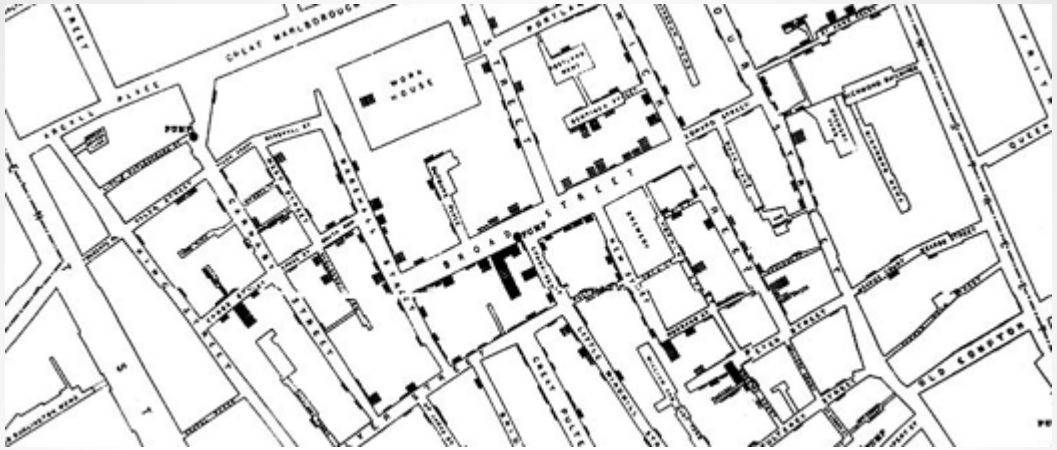
How to Discuss Data

Which John Snow?

John Snow

- discovered source of Cholera outbreak **using data in 1854!**
- used a dot map to illustrate the cluster of cholera cases. Used statistics to illustrate the connection between the quality of the water source and cholera cases

John Snow's Original Scatterplot



Modern Version Using Google Maps

John Snow

John Snow's study regarded as the founding event of the science of **epidemiology**: the study of the patterns, causes, and effects of health and disease conditions in defined populations.

What Questions Can You Ask Data?

How do 9 States affect the outcome of an election?

How can you track the rights of women over the world since 1892?

How many people have been killed by guns since Sandy Hook Elementary shooting?

How do you know what drink you want?

How do you remember all the callback jokes in Arrested Development?

Now it's your turn

To make a change!

(with big data)