Jonathan Go

# STA 371H Notes 1/26/15

Scribing Information: 5% of grade
- -Be VERY detailed with notes

Good Statistical Graphics: appeal with visual evidence for your eye
- -Characteristics:
  - -Vehicles for comparison: fair
  - -Multivariate:
  - -Truthful about magnitude: the relative size of numbers
  - -Not usually small data sets
  - -Not pie charts: eye is very poor at distinguishing relative areas
- -Categorical/Grouping Variables: died vs. survived in Titanic
  - -Use Contingency Table
    - -Cross tabulates all data into 1 of __ amount of options (e.g. 2 by 2 table)
    - -Can normalize by row or column
      - -Each row/column adds up to 100, displayed in percentages
      - -Look for patterns in normalized data
- -Numerical/Quantitative Variables: average temperatures in San Diego
  - -Use Histogram:
    - -To compare properly, x axis, y axis, and bin sizes must be the same
  - -Use standard deviation:
    - -Measures deviation in data
  - -Use coverage interval:
    - -Data from specified percentage, e.g. quantiles (percentiles in fractions)
- -Comparing Numerical vs. Grouping Variables:
  Need to have dispersion, not just average (e.g. college vs average SAT)
  - -Use box plot:
    - -Shows groups, dispersion, and percentages
    - -Both between-group and within-group variation)
    - -e.g. a within-group variation may be much higher than the between-group variation

Bad Statistical Graphics:
- -Truncating the y-axis:
  - e.g. only show 94-100 vs. 1-100 on y axis
- -Percentages that don't add up to 100%
- -Distorting relative sizes:
  - e.g. pyramids=3D figures are interpreted as volume of a solid vs the 2D side of pyramid
- -Low information density:
  - e.g. only 3 years of data for a period for 10 years
- -Pictures/graphics should never be a distraction: let data speak for itself
- -Doesn’t' address lurking variables compared to correlation over time:
  - e.g. data from around 2008 that ignores the recession going on

$Y_i$ = outcome for case i

$Y_i$ = Group Mean + Deviation From Mean = Actual Case = Fitted Value + Residual = $\hat{Y}_i + e_i$

Fitted Case Value gets a hat ($\hat{Y}_i$) because it's an imposter

Jonathan Go

Vocabulary Items

- -Longitudinal Studies:
  - -Same location over a period of time
  - e.g. comparing Austin before and after cell phone ban
  - (make sure to minimize cofounding variables like previously existing trends)
- -Cross Sectional Study:
  - -Multiple units at the same time, trying to minimize confounding variables
  - e,g. comparing both sides of Texarkana (half of city in Texas, half in Arkansas)
- -Natural Experiments:
  - -Experiment that occurs naturally, but behaves like a designed experiment
  - e.g. lottery tickets
- -Case Control Study:
  - -Different than experimental intervention
  - -Matches instead of manipulating variables
  - e.g. "Don't drink when you're pregnant"
    - -Looked down the line at negative vs positive outcomes for babies
    - -Then looked back at the alcohol history
    - -However, the control group actually abused cocaine twice as much as subject group, so the study was incorrect at establishing a causal relationship
    - -Example of matching done horribly wrong

Given a study of aspirin and heart attacks:

- -Endogenous Variables:
  - e.g. find people who have taken aspirin in the last year and ask if they've had a heart attack
  - -confounding variables (e.g. health consciousness)
  - -try to avoid endogenous
  - -"aspirin is confounded"
- -Exogenous Variables:
  - -Controlled from outside the system
  - e.g. set up a controlled experiment to give subjects a controlled, set amount of aspirin every day and measure results
  - -subjects can't decide how much/when to take aspirin, and are forced to take it, therefore aspirin is a exogenous variable
  - -"aspirin is not confounded"

Announcements:
TA Session: Thursday 3-5pm CBA 4.326

*See other set of scribing notes for pictures of the instructor's notes*