Reasoning with Data Notes

saffron_

Contents

Course Summary

- see `lecture1written.pdf`

Course

Exploratory Data Analysis

EDA for 1-variable categorical data

- population: complete set on interest (eg all US workers). can't be measured perfectly
- sample: subset of pop that can actually be obtained
- parameter: summary of population (eg average weight). also can't be measured
- statistic: estimation of parameter using sample.
- inference: specifying estimate of a parameter (ie gving estimate and measure of how far off it is)
- individuals/units/cases: objects described in dataset
- variable: column of spreadsheet, something measured
- quantitative: numerical (eg age)
- categorical/qualitative: not numerical (eg ethnicity)
- frequency table: show how often each *categorical* variable shows up
  - relative freq. table: how percent often they show up, adding up to 1 or 100%: best summary of a categorical variable
- frequency/percent/relative (add to 1) frequency bar graphs: to visualize categorical data. v. similar to their freq table variants.
- pie graph: to visualize a *single* categorical variable, with each slice being a category.

EDA for 1-variable quantitative data

- grouped frequency/grouped relative frequency table: for one quantitative var. boring version of a histogram.
- histogram: for one quantitative variable. visualizes a grouped (rel.) freq table.
- distribution of one quantitative var (anal. doesn't really hold up for qualitative)
  - modality: how many 'clusters'? (eg unimodal/bimodal/etc)
  - symmetric/skewness? (skew right means tail is to right)
  - center: mean ($\overline{x}$)
    - median if heavy outliers. using mode is cursed but you do you
  - spread: standard deviation (S) (usually, $\frac{2}{3}$ of values are 1 stddev away from mean)
    - interquartile range (IQR) if heavy outliers.
  - outliers?
- fun facts: variance: square of std dev. easier in formulas because it removes the square root, but in real world, std dev is preferred because it has same units as the data
- box plot: other way to graph quantiative data (using min, first quartile, median, third quartile, max)
  - interquartile range (IQR): difference between quartiles, a measure of spread. is rough, not as useful as stddev.
  - shows less data than histogram, so best for concise comparison of *multiple* distributions (see section on 2-variable EDA)

EDA for 2-variable data

- relationship/association: one variable can tell you about another
- explanatory variable: "input" variable, the x-axis
- response variable: "output" variable, the y-axis
- if we need analysis from explanatory → response
  - categorical → quantiative: side by side boxplots
    - more difference if boxes are futher apart on "y-axis"
    - summaries: numerical summaries of response for each category
  - categorical → categorical: contingency table
    - summaries: conditional percents of responses, conditioned on explanatory (ie what percent of people from Cali are stat majors?)
  - quantitative → quantiative: scatterplots
    - describe: direction, form, outliers
    - summaries (only if reasonably linear): correlation coefficient (R), least squares regression line ($\hat{y} = b_0 + b_1 X$)
  - quantiative → categorical: outside scope of this course

Study Design

- no matter what, we want to consider and get…
  - reliability, statistical significance: low random variation/error
    - use a large sample size
    - can be measured with stddev
  - validity: trustworthy estimates and predictions
    - consider and declare outliers. remove them if needed, but with caution!
    - beware extrapolation outside range of data that produced the model
    - validate model (eg check for linearity if you use linear regression)
  - generalizability: no bias aka systemic error
    - called instrument bias if instrument is set wrong (can be social instrument such as misleading survey)
      - remove via resetting instrument (reset scale, rewrite survey)
    - called sampling bias if sample is systematically not representative
      - remove via random sampling. such as simple random sampling (SRS): every individual has same chance to be chosen as any other; every pair same chance as other pair; etc
- and if we have 2 or more variable relationships…
  - causality: no lurking/confounding variables; if we want to find causation and not just correlation
    - use randomized assignment of explanatory variable
    - experimental study: study where lurking/confounding variables are removed
    - observational study: not experimental. subjects decided which treatments to get (eg survey vitamin c usage vs flu symptoms)
    - placebo control group: 'non-active' treatment to avoid placebo effect
    - 'double blind': neither the researchers nor the subjects know which treatment they are receiving

Elementary Probability

- TODO lecture 9 onwards