Reasoning with Data Notes

saffron_

Contents

Course Summary

- see `lecture1written.pdf`

Course

EDA for one 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 numereical (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 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.

- distribution: representation of data that shows what values the variable takes and how often

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/median? using mode is cursed but you do you
  - spread: std deviation (usually, $\frac{2}{3}$ of values are 1 stddev away from mean)
  - 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 $\rightarrow$ response

  - categorical $\rightarrow$ quantiative: side by side boxplots
  - categorical $\rightarrow$ categorical: contingency table
  - quantitative $\rightarrow$ quantiative: scatterplots
  - quantiative $\rightarrow$ categorical: outside scope of this course

  TODO lec 5-6-7 for summarizing above

Study Design

- no matter what, we want to consider and get. . .

  - sensitivity: low random variation (use a large sample size)
  - 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 (use random sampling)

- and if we have 2 or more variable relationships. . .

      – causality: no lurking/confounding variables; if we want to find causation and not just correlation (randomized assignment of explanatory variable)

- bias/systemic error: consistent tendency toward inaccurate outcomes

      – instrument bias if instrument is set wrong (can be social instrument such as misleading survey)

      – sampling bias if sample is systematically not representative

      – remove sampling bias via random sampling

- simple random sampling (SRS): every individual has same chance to be chosen as any oher; every pair same chance as other pair; etc

Elemetary Probability

- TODO lecture 9 onwards