

# Reasoning with Data Notes

saffron\_\_

## Contents

<b>Course Summary</b>	<b>2</b>
<b>Exploratory Data Analysis</b>	<b>2</b>
<b>Study Design</b>	<b>5</b>
<b>Elementary Probability</b>	<b>6</b>
<b>Probability Models of Data</b>	<b>7</b>
<b>Sampling Distribution of a Statistic: The Probability Model of Sample Mean and the CENTRAL LIMIT THEOREM</b>	<b>8</b>

## Course Summary

- see `lecture1written.pdf`

## 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 giving 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 ( $\bar{x}$ )
    - \* 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 quantitative 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  $\rightarrow$  quantitative: side by side boxplots
    - \* more difference if boxes are further apart on “y-axis”
    - \* summaries: numerical summaries of response for each category
  - categorical  $\rightarrow$  categorical: contingency table
    - \* summaries: conditional percents of responses, conditioned on explanatory (ie what percent of people from Cali are stat majors?)
  - quantitative  $\rightarrow$  quantitative: scatterplots
    - \* describe: direction, form, outliers
    - \* summaries (only if reasonably linear): correlation coefficient ( $R$ ), least squares regression line ( $\hat{y} = b_0 + b_1X$ )
  - quantitative  $\rightarrow$  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

- probability: measure of variation due to a random phenomenon
- random: produced so that probability applies. satisfies:
  - equal likelihood ( $A$ 's selection is as likely as  $B$ 's)
  - independence ( $A$ 's selection doesn't influence  $B$ 's)
- coin:  $P(\text{heads}) = 0.5$
- relative frequency: if selections are random (eq likelihood and independent),

$$P(\text{outcome}) = \frac{\# \text{ of ways outcome occurs}}{\# \text{ of possible outcomes}}$$

- long-run approx. can be as precise as desired: Law of Large Numbers
- not replacing violates independence
  - \* consider a bag with a blue and red ball. selecting blue and not putting it back makes  $P(\text{red}) = 1$ .
  - \* but this is common—a pollster doesn't survey two houses twice
- not replacing is fine if population  $\geq 10$  or 20 times the sample
  - \* consider 2000 blue, 2000 red. selecting blue makes  $P(\text{red}) \approx 0.5$
- probability experiment: an outcome's success/failure varies
  - eg flipping a coin: outcome (like get heads) varies
  - trials: different “runs” of a probability experiment
- disjoint/mutually exclusive:  $A \cap B$  is empty
- elementary outcomes: irreducible for assigning probabilities. mutually disjoint.
- sample space:  $S$ , the set of all elementary outcomes in a probability experiment
- complement:  $A^C$ , “opposite” of a probability.  $A + A^C = 1$ .
- intersection:  $\cap$ , “and” of two probabilities. middle of venn diagram.
- union:  $\cup$ , “or” of two probabilities. all circles' overlap area.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- joint probability tables
  - each inner box is an elementary outcome
  - rows sum to row marginal probabilities, same with columns
  - 2x2 tables can represent venn diagrams

- conditional probability ( $A$ , if we know that  $B$ ):
  - $P(A \text{ given } B) = P(A | B) = P(A \cap B) \div P(B)$
  - for a d20,  $P(4 | \text{even}) = 0.1$
  - take from box and margin of a joint probability table
- statistically independent: unassociated events. if you know the outcome of one event, it doesn't change the probability of the other.
  - with variables: association b/t  $A$  and  $B$  is nearly what is expected from chance
  - with sampling: any person similarly likely to be selected, regardless of previous selection
  - $A$  and  $B$  are independent
    - \* iff  $P(B | A) = P(B | A^C) = P(B)$
    - \* iff  $P(A \cap B) = P(A) \cdot P(B)$

## Probability Models of Data

- random variable: outcome of randomness that takes on numerical values
- discrete rv: takes on “separated” outcomes. modeled with table or rel-freq histogram
- continuous rv: takes on an interval of outcomes. modeled with density curve
- rv example: cars occupancy
  1. define prob. experiment: randomly select car on road
  2. define rv: let  $X$  = number of people in car
  3. assign probabilities to  $X$ . denote prob. of  $X = 1$  as  $P(X = 1)$
- Binomial Counts “Experiment”/Study/Trial
  - conditions:
    - (a) fixed sample size  $n$  for every run of study
    - (b) two categories (success/failure) for each observation
    - (c) observations must succeed or fail independently
    - (d) fixed prob  $p$  of success for each observation
  - conditions c and d are guaranteed if selections are random
  - the count of successes in  $n$  trials is a discrete rv  $X$ , and follows the Binomial Distribution (ie,  $X$  is binomial. see by graphing the histogram)
  - $P(X = x) = {}_n C_x \cdot p^x \cdot (1 - p)^{n-x}$
  - binomial coefficient  ${}_n C_k = C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$  (recall  $0! = 1$ )
  - eg: 8 rand selected lab animals get vaccine w/ prevention rate of 40% and then are infected. Then  $X$  = surviving animals out of  $n = 8$  with  $p = 0.4$ .

- density curve
  - graph a histogram, imagine a idealized curve fitting it
    - \* idealized curve means true populations, so variables are greek letters
    - \* mean =  $\mu$  (mu)
    - \* stddev =  $\sigma$  (sigma)
  - area = proportion of observations = relative freq = probability
  - area = 1
  - exponential distribution
    - \* let  $X$  be the time until the next event, with events occuring randomly with a mean waiting time  $\lambda$ .
    - \* denoted  $X \sim \text{Exp}(\lambda)$
    - \* fun fact: this is graphed  $f(x) = \lambda e^{-\lambda x}$
    - \* eg: time until next earthquake, time until a phone call ends
  - uniform distribution
    - \* a process produces values  $\in [a, b]$  with equal likelihood.
    - \* denoted  $X \sim \text{Uniform}(a, b)$
    - \* fun fact: this is graphed  $f(x) = \frac{1}{b-a}$  for  $x \in [a, b]$ ,  $f(x) = 0$  elsewhere
    - \* eg: result of dice roll
  - normal/gaussian distribution
    - \* models result of mixing many independant factors and Central Lim Thrm
    - \* denoted  $X \sim N(\mu = u, \sigma = s)$
    - \* fun fact: these are variations on the graph  $f(x) = e^{-x^2}$
    - \* eg: human height
    - \* points of inflection are 1 stddev away from mean
    - \* 68/95/99.7 percent of population falls within 1/2/3 stddev of mean
  - standard normal distribution
    - \* for discussing probability on normal distributions: on all normal curves, portions with the same  $Z$  score have the same area.
    - \* standardized/ $Z$  score: stddevs from mean for a number
      - $Z = \frac{\text{observation} - \mu}{\sigma}$
    - \*  $N(\mu = 0, \sigma = 1)$  is already standardized. we may denote  $Z \sim N(0, 1)$
    - \* eg: consider adult male weight  $X$ .
      - we know  $X \sim N(\mu = 165, \sigma = 30)$
      - if P weighs 202 pounds, how unusual are they?
      - $Z = \frac{202-165}{30} = 1.23$  and  $P(X < 202) = P(Z < 1.23) = 0.8907$ .
    - \*  $Z$  score tables and such exist for these calculations
    - \* be able to calculate:  $Z$  score to prob, prob to  $Z$  score, stdizing to find prob, unstdizing to find score (see Lecture 14 notes)



# Sampling Distribution of a Statistic: The Probability Model of Sample Mean and the CENTRAL LIMIT THEOREM

TODO