

Reasoning with Data Notes

saffron__

Contents

Course Summary	2
Exploratory Data Analysis	2
Study Design	5
Elementary Probability	6
Probability Models of Data	7
Sampling Distribution of a Statistic	8
Formal Inference (Confidence Intervals, Hypo. Testing)	10

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

good luck on exam 1 <3

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 ≥ 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 (and a joint probability like $A \cap B$)
 - 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 $P(A | B) = P(A)$
 - * iff $P(A | B) = P(A | B^C)$
 - * 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 $p = 0.4$.

- distributions (density curves)
 - graph a histogram, imagine a idealized curve fitting it
 - * idealized curve means true populations, so variables are greek letters
 - * mean = μ (mu)
 - * stddev = σ (sigma)
 - area = proportion of observations = relative freq = probability
 - total area = 1
 - exponential distribution
 - * let X be the time until the next event, with events occurring randomly with a mean waiting time λ .
 - * 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

- population distribution: eg how long it takes for literally every lightbulb to burn out
- (sample) data distribution: eg how long it takes for this random sample of $n = 1000$ lightbulbs to burn out
- sampling distribution of a statistic: the probability distribution of all possible values of a statistic (from taking many samples of same sample size from same population)
- sampling distribution of \bar{X} , the mean

- center: if sampling is random, then mean of sample is the population mean. ie,

$$\mu_{\bar{x}} = \mu \text{ if sampling is random}$$

- spread: if sampling is random and outcomes are statistically independent (eg sampling with replacement or infinitely large pop), then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- shape: if sample is large ($n \geq 30$), then by the Central Limit Theorem, the sampling distribution of \bar{X} approaches the normal distribution. (resembles population otherwise)
- lightbulb example problem in Lecture 16. Note AFSOC for inference.

- sampling distribution of \hat{P} , proportion

- consider a binomial study (in particular, random and independent) with sample size n and success probability p . then,

- center:

$$\mu_{\hat{p}} = p$$

- spread:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- shape: if np and $n(1-p)$ both ≥ 10 , then by the Central Limit Theorem, \hat{P} approaches the normal distribution. (resembles population otherwise?? maybe?)
- coinflip example problem in Lecture 18.

good luck on exam 2 <3

Formal Inference (Confidence Intervals, Hypo. Testing)

- we study two branches of formal inference:
 - confidence intervals: what are most likely values of a parameter?
 - hypothesis/significance testing: what is the likelihood of a parameter to be a value?
- Confidence Intervals

- Theory: \bar{x} is “probably” somewhat “close” to μ .
 - * how probable? confidence level
 - * how close? margin of error
- if random sampling, stat. independent, normal, then the confidence interval is

$$\bar{X} \pm Z_{\text{critical}} \frac{\sigma}{\sqrt{n}}$$

- * where n = sample size, \bar{X} = sample mean, σ = assumed population stddev, $Z_c = 1.645$ for $C = 90\%$, $Z_c = 2$ for $C = 95\%$, $Z_c = 3$ for $C = 99.7\%$, etc
- * eg: $\approx 95\%$ of \bar{x} 's are within $\pm 2\sigma_{\bar{x}}$
- similarly, if binomial (in particular, random sampling and independent) and normal, the confidence interval is

$$\hat{p} \pm Z_{\text{critical}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- Hypothesis Testing
 1. State Hypotheses
 - null (H_0): $\mu = \text{value}$
 - alt (H_a): $\mu <, >, \text{ or } \neq \text{value}$
 2. choose α (usually 0.01 or 0.05)
 3. perform study, get sample statistic, get p value of that sample statistic
 4. conclusion: if $p < \alpha$, this is very unlikely. we reject the null hypothesis
 5. state in context: what does parameter represent? what is the likely value/why?
- Note: ‘ \neq ’ confidence test with significance α is equiv to making a confidence interval with level $1 - \alpha$ ie α^C
 - ie: Hypothesis test says do not reject $H_0 \Leftrightarrow H_0$ value is inside confidence interval
- TODO formal template for writing this. an example on lec24 p5 says “since the p value is not less than 0.05, we do not reject the null hypo. [explain what this means]”

- T-Test for mean when sigma unknown
 - we instead use sample stddev, S .
 - like how we call $\frac{\sigma}{\sqrt{n}}$ $\sigma_{\bar{x}}$, we call $\frac{S}{\sqrt{n}}$ standard error
 - notice how similar our equations are
 - hypothesis test for a mean
 - * test stat = $t = \frac{\bar{X} - \mu_{null}}{S/\sqrt{n}}$ (“stderrs away from mean” instead of stddevs away from mean)
 - confidence interval for a mean
 - * $\bar{X} \pm t_{\text{critical}} \cdot \frac{S}{\sqrt{n}}$
 - * note: t_{critical} is a bit over 2 for 95% because the t distribution is fatter at tails than the normal
 - instead of $n \geq 30$ for normality, you may inspect the distribution. since t-distribution is robust, vauge normality is even fine as samples sizes increase (as long as no severe outliers). ie, if $n \geq 40$ or so, t dist is valid regardless of shape
 - t approaches norm distribution as n increases.
- what the FUCK are lect 25-6 cooking? about tests for two independent means and proportions. do you just get the t or z scores out of thin air? aka from isle.stat.cmu.edu?