For AP Exam... NOT in ORDER of MR. MICEK!!

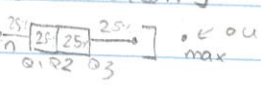
AP Stats Unit 1

Categorical data
- Bar graphs (put # inside not out) / Pareto (tall→small)
- Pie charts
- Frequency tables

Quantitative data
- Pictographs
- Line
- Ogive (cumulative) (Technically quantitative) → shows the % of data at or below a point
- density curves, N/A to below

Ana: High, Low, Trend, Why

- 2 types: discrete (limited, countable), continuous (infinite, decimals included)
- Histogram/frequency table: tally marks, has bins. chose a number of bins (5-7), and to find width, do $\frac{\text{max} - \text{min}}{\#\text{ of bins}}$ (round up). Start with smallest data value, count the width into the class limits (eg, if w=3, limit is 1-3). x-axis label is the start of each class limit.
- dot plots: good w/ mdm amt of data w/o high spread
- Stem & leaf plots: find max+min, draw a line splitting data into places, record, then organize (note, can do 2 lines per stem or back to back)
  - include a key!
- box + whisker: get a 5 # summary (min, max, Q1/Q2 (median)/Q3), IQR is Q3 − Q1, outlier fence is Q3 + 1.5(IQR) and Q1 − 1.5(IQR). Wider box/line means more spread out data



Ana
- Center (id which)
- Shape (eg, unimodal, bimodal, symmetric, etc)
- Spread (id which + a word)
- Outliers (under 5%, large gap, or IQR method)

The Averages

- Mean: $\frac{\Sigma x}{n}$, easily influenced by outliers. $\bar{x}$ for sample, μ for popln
- Median: midpoint. $\frac{n+1}{2}$ is location. resistant to outliers
  - if mean = median, symmetric
  - if mean < median, skewed left (tail left)
  - if mean > median, skewed right (tail right)
- Mode: most common occuring value
- Trimmed mean: eliminates pull of outliers. Eg, if 5%, cut off 5% of the amt of data points (.05(n), round up) from both sides

Measures of Spread

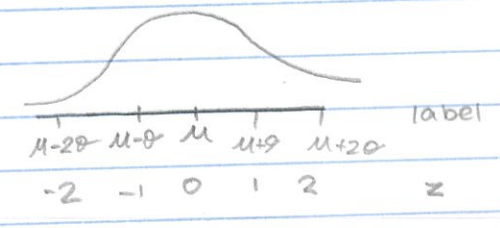
- Range: max - min, influenced by outliers
- IQR (Q3 - Q1)
- Standard deviation: $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = s$ (or θ in popln)
- Coefficient of variation is $s/\bar{x}$ or θ/μ
  - 0-10% clustered
  - 13-33% low
  - 35-67% moderate
  - 70-99% high
  - 100%+ data gone wild!

Transformation of data:

- Adding / Subtracting (a constant to every data value) changes the center but not spread
- Multiplication / division (by a constant to every data value) affects all measures: center + spread (by same constant)
- Adding in or taking away a data point changes mean + spread if its an outlier

Normal distribution:

- continuous quantitative variables
- empirical rule: 68%, 95%, 99.7%
- unimodal, symmetric, mound shaped
- Z-scores: $\frac{x-\mu}{\sigma}$



AP Stats Unit 2 (Two-variable data)

one set of data with 2 variables! are they associated?

Categorical data:

- use a two-way table (proportions, not just counts)
  - marginal relative frequencies: grand totals (÷ by total data)
  - joint relative frequencies: the word "AND". (÷ by total data)
  - conditional relative frequencies: changes denominator, instead of grand total = of data as before, the denom is based on the condition
  - segmented bargraph: based on conditional relative frequency
  - to tell if there is association between variables, look at marginal relative frequencies verus conditional relative frequencies for each
    - if the rf are same, no association = independent
    - if the rf are different, shows an association/dependence

Quantitative data:

- scatter plot
  - x-axis is explanatory variable, y-axis is response variable
  - identify direction, form, strength, unusual features
    - direction: positive or negative
    - form: linear, curve, parabolic
    - strength: how closely the individual points fit the pattern
    - unusual features: outliers, clusters, etc
  - Ana: +/- association
    trends (sentence)
    influentials/outliers
    correlation(r) + "word"
    influential follows pattern, just appear they affect the slope
  - correlation coefficient: how strong relationship is. (r) (linear only)
    negative means direction is −
    positive means direction is +
    - between -1 and 1
    (perfect straight line is ±1)
- least squares regression line:
  - $\hat{y} = a + bx$ (a is y-int, b is slope) (a sometimes doesn't make sense in context)
  - do not extrapolate / make predictions for values outside the data range
- residual
  - $y - \hat{y}$ → difference between actual value & predicted value
  - all points have their own residual value, all should add to 0
  - sd of residuals tells us how far off the model is in making predictions

- residual plot: plot of all residuals for all points
  - linear is a good fit when the plot is random, centered at 0, w/ no clear pattern
  - linear is a bad fit if there is a clear pattern like a curve

- coefficient of determination (r^2)
  - always positive, between 0 and 1
  - measures % of variation in response variable (y) that is explained by the variation in the explanatory variable (x)... how well the variables are connected! how reliable the LSRL is, higher means more reliable
  - must say, r^2= ___ % of the variation in the y (___) is explained by the least squares regression line and the variation in x (___).

- outliers/influentials:
  - outliers
    - occur in one direction: x or y, can have a normal x value but odd y, or vice versa.
    - outliers don't fit the pattern, large residuals. weaken correlation
  - influentials: fit the pattern, just far off in both direction have small residuals. if removed, change slope/correlation/y-int... change model in general. high-leverage points (have lg/sm x-values) is an example

Formulas:

$\hat{y} = bx + a$

$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$a = \bar{y} - b\bar{x}$

$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

AP Stats Unit 3

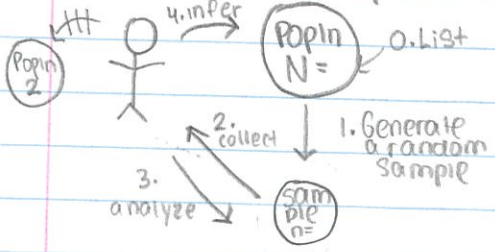
- everything starts w/ a research question regarding a variable or a relationship between 2 variables
- popln = set of all subjects we are interested in (N), stated as "quantitative variable of all ___"
- sample = subset of popln, used to make generalizations ab the larger popln, stated as "quantitative variable of ___ ..." (eg, avg # of pages read by 123 students in P4)

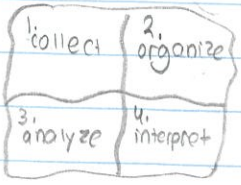
Types of sampling:

- cluster: ÷ popln into clusters (← heterogenous, no shared trait) + randomly chose (← SAS) certain clusters + census them
- stratified: ÷ popln into categories (strata / ← homogenous / shared trait, eg color) + chose random sample from each strata
- systematic: chose a number + chose sample of every # (eg, every 5th object)
- simple random sampling: (only one that) allows for equal prob of an object and samples to be chosen.
  - one way is using a RNT. 1) number each thing from min to max. all things should have the same # of digits (eg 001-678), 2) chose any random spot on table, 3) go in any direction but stick in the same direction. keep anything within limits + throw out anything beyond the entire deal, 4) stop when you get the desired # of samples
  - another way is using calc. 1) # each thing, 2) press math, 3) go to prob and randInt No Rep, 4) put lower + upper, n= amt of samples wanted, 5) enter + take the ones used

(+ census: whole pop)

- inferential statistics pic:





Levels of Measurement:

1) Nominal: categorical/names/ order not important
2) ordinal: categorical, order important (like ratings)
3) interval: quantitative, arranged in order, but no natural zero
4) ratio: quantitative zero exists, can divide

types of bias:

- nonresponse: individuals refuse to participate/cant → undercoverage
- voluntary response: those w/ strong feelings are more likely to respond → skews. (not representative) (self-selection bias)
- convenience: asking those around, not representative of popln, as diff location = diff types of ppl
- response bias: lying or faulty recall
- availability bias: relying on info that comes easier to mind
- recall bias: giving false response be hard to remember
- under-coverage: part of popln is excluded
- confirmation bias

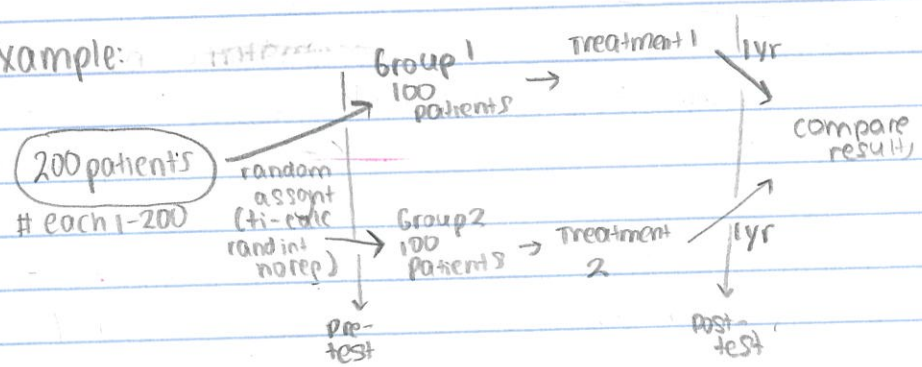
Experimental Design:

- 4 important principles: comparison, random assignment of treatments, direct control of confounding variables, and replication
- experimental unit: subjects
- explanatory variable: manipulated / independent variable, also called factors
  - usually one factor w/ 2+ treatments, if more, has combos of both as treatment
- response variable: dependent variable, what you're measuring
- control group: group that doesn't receive treatment, could get placebo
  - active control group receive existing treatment
- confounding: two factors cannot be distinguished in their effects on the response variable. control for this!
- single-blind: subjects don't know what group they are in / dk what they're getting
- double-blind: subjects nor the research team know the assignments of subjects → done to prevent any bias

3 experimental designs:

1. completely randomized: want groups to be as similar as possible, only difference is the treatment
2. randomized block design: before randomly assigning to groups, block the sample by any possible confounding variables (ex, gender, age, grade, etc). within each block, continue like normal
3. matched-pairs design: pair subjects together, one gets treatment, other in pair gets control, chosen randomly. the pairs themselves have a lot in common (eg, both men 75 w severe alzheimer)

Example:



AP Stats Unit 4

- outcome: result of a random event (rolling a 5)
- event is a collection of outcomes (rolling an even number)
- Long run relative frequency: after a large # of repetitions, #/n, give # and word
  - law of large numbers: simulated probabilities approach true probabilities as the number of trials increases
- Simulation:
  1) state the problem / what you're looking for
  2) state any assumptions (pre-req probabilities)
  3) assign digits
  4) many reps (20+), note that a rep is not the # of #s you look at
  5) state conclusion (avg out all probs for the reps)
- sample space: all outcomes
- complement of an event is the prob of the event not happening
  - $P(A^c) = 1 - P(A)$
- "And" probability: both happen at the same time
  - $P(A \text{ and } B) \;/\; P(A \cap B) = \boxed{P(A) \cdot P(B|A)}$
  - if they can't happen at the same time, called mutually exclusive. $P(A \text{ and } B) = 0$
- "Or" probability: combination of 3 probabilities: P(A) only, P(B) only, P(A and B)
  - $P(A \text{ or } B) \;/\; P(A \cup B)$
  - ↳ Conditional probability: P(A) given that B has occured
    - $P(A|B)$
    - formula: $\boxed{P(A|B) = \frac{P(A \text{ and } B)}{P(B)}}$
  - $\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$
    - if not mutually exclusive / overlap, must subtract the overlap (A and B). if they are mutualy exclusive, P(A and B) = 0 so that part is negligible
- Independence of events
  - A and B are independent iff event A doesn't change prob of B
  - if $P(A) = P(A|B)$ or $P(B) = P(B|A)$, are independent
- know how to interpret a 2-way table!
- use tree diagrams to simplify
- Odds is a ratio of favorable: unfavorable

- random variables:
  1) continuous (unit 5)
  2) discrete
     - can take on a countable # of values, each w/ own probabilty
     - can make a probability distribution (graph / table)
     - written as X and P(x) in a table/graph
- analysis of probability distr: CSSO
  - mean/expected value: what we expect the outcome to be in long run
    $\mu = \Sigma(x \cdot p(x))$ sum each outcome x its probability
  - standard deviation: precision of the expected value
    $\sigma_x = \sqrt{\Sigma(x-\mu_x)^2 \cdot P(x)}$
  - on calc: input all x and P(x) into L1 and L2, then 1-var-stats, freq list is L2.
- Transforming and combining random variables
  - transforming is converting from one unit to another
    ↳ multiplying a rv by a constant multiplies mean + sd by the same constant
    ↳ adding / subtracting of a constant only affects mean only by adding / subtracting that constant. sd stays same
  - combining is adding rv tgt or repeating 1 variable many times
    1) each outcome must be independent of others
    2) means can be added or subtracted simply
    3) cannot add/sub standard deviations ★
       you are allowed to combine variance.
       ↳ only add variance never subtract. even when difference
       ↳ so, convert sds to variance, then add variance, the sq root to get back to sd.

AP Stats Unit 4 cont...

• Binomial Distribution:

- outcomes are countable, numerical values
- only 2 outcomes: success or fail
- rules:
  1) success is clearly defined, called p
  2) probability is same for each trial
  3) $q = 1-p$ (failure)
  4) each trial is independent
  5) a set number of trials
- $P(X=x) = {}_nC_x \cdot p^x \cdot q^{n-x}$
  - ${}_nC_x = \frac{n!}{(n-x)!\,x!}$
- on calc: 2nd vars → binompdf ! binomcdf is cumulative of all before it (at most x)
- mean: $np = \mu$ (expected value)
- Standard deviation: $\sqrt{\mu q} = \sigma$

Geometric Distribution

- outcomes also countable, numeric values
- X = number of the trial where the 1st success occurs
- rules:
  1) success is clearly defined, called p
  2) probability same for each trial
  3) $q = 1-p$ (failure)
  4) each trial independent

  } same as binomial
  5) no n, meaning no set # of trials
- $P(X=x) = q^{x-1} \cdot p$
- if looking for a success to happen after a certain point, $P(x > x) = q^x$
- mean: $\mu = \frac{1}{p}$
- Standard deviation: $\frac{\sqrt{q}}{p}$
- on calc, geompdf

AP STATISTICS: AGAINST ALL ODDS
Video 14 Worksheet

Name Dasha Heydari

SAMPLES AND SURVEYS

1. What is an estimate based on a sample? _____

2. What is a true value that describes an entire population? Parameter

3. What is the process of dividing a population into similar units? _____

4. What example of stratification is used in the video? _____

   How many strata are used? _____

5. In 1936, the Literary Digest predicted Alf Landon would win the presidential election. How many readers did the magazine poll? 2.3 million How many people did Gallup poll? 50,000

   Who did Gallup predict as the winner? Roosavelt

   What was the problem with the magazine's poll? Only covered the wealthy and undercovered the rest.

6. List three mistakes that can occur in polling.
   a. Word choice (and loaded questions) can influence people's feelings, thoughts, and perceptions.
   b. Polling people around others, which can change their responses if they are conscious
   c. Interviewer/surveyor can also subconsciously influence responses (based on appearance, gender, etc)

7. How many personal interviews are conducted each year as the core of the GSS? _____

8. What is the histogram of the sampling process called? _____

9. What pattern does this distribution follow? _____

10. What is the peak of the distribution? _____

11. What happens to the distribution when the sample size is increased? _____

12. What determines precision? _____

Stats August 10, 2023

1.1 5-8, 11, 12

Dasha Heydari
Period 4

10

(a)

(5) Nominal, because there is no difference between data and no criteria to actually put the data from smallest to largest. There is no numerical value.
(b) Ordinal, because the choices can be ordered, but there is no meaningful numerical difference between the options.

(6) No, Lucy's observations do not apply to all adults as this is not a good sample. Her friends will most likely be more similar to each other than the actual population of adults: it is not representative. Based on this description alone, we cannot draw those conclusions, other than that they will probably have similarities to Lucy and her statuses.

(7) (a) The variable is the meal (breakfast, lunch, dinner) ordered.
(b) The variable is qualitative.
(c) Data from all U.S. adult fast food customers in the U.S.

(8) (a) The variable is average miles per gallon.
(b) Quantitative.
(c) Data from all new hybrid small cars.

(11) (a) Ratio
(b) Interval
(c) Nominal
(d) Ordinal
(e) Ratio
(f) Ratio

(12) (a) Ordinal
(b) Ratio
(c) Nominal
(d) Interval
(e) Ratio
(f) Nominal

Dasha Heydari
Period 4

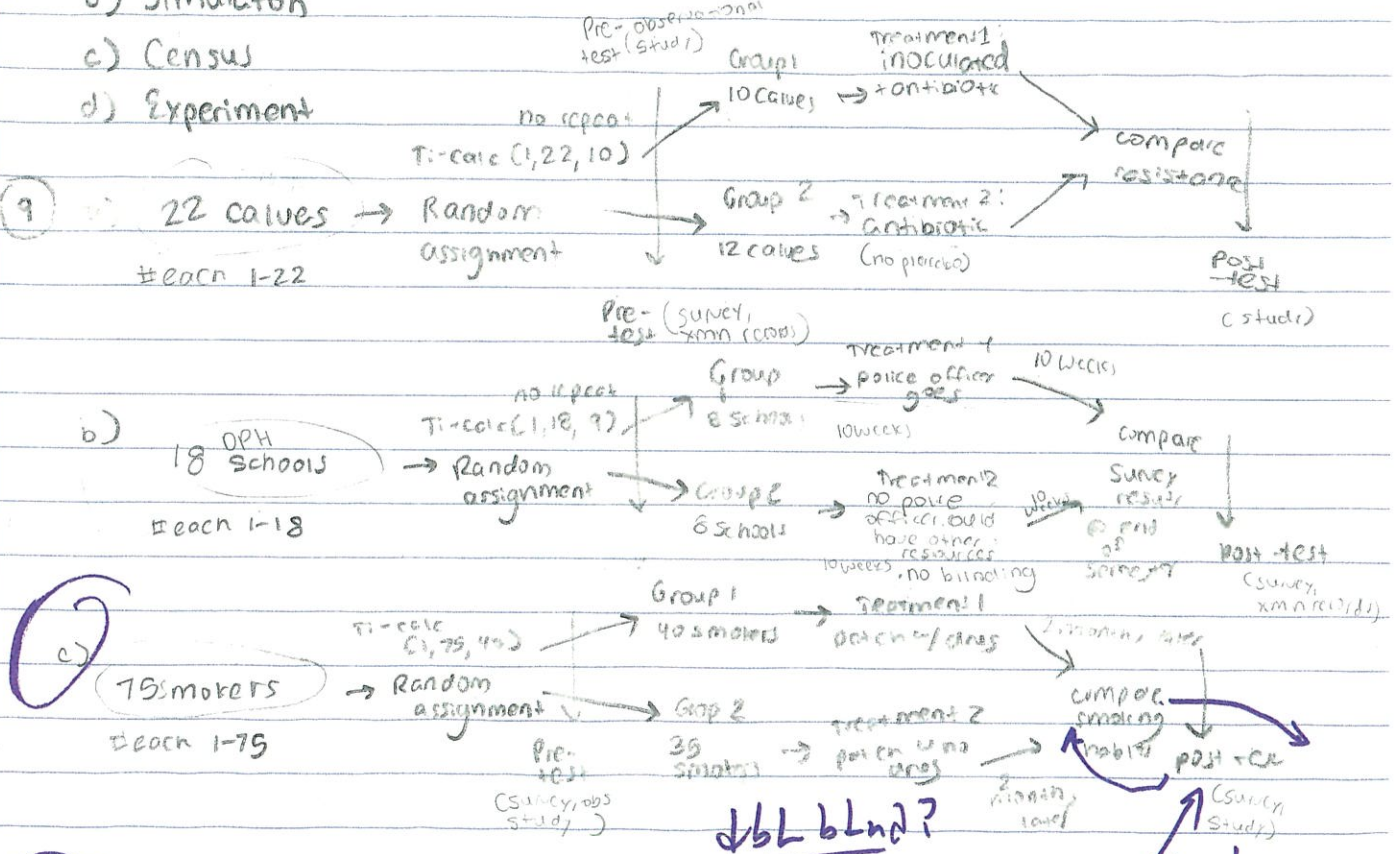
9+

1.3 #7-10 (9 outline)

(7) a) observational study
b) experiment
c) experiment
d) observational study

(8) a) Sampling
b) Simulation
c) Census
d) Experiment

(9)



(10) a) This is very subjective. Chosing a speafic amount of years, such as 2 or 3 is much more precise and eliminates the vagueness.

b) It would be different, because it causes the individual to feel introspective and guilty if they answered "yes" to the previous question, altering the way they answer to the one above them.

c) The answers would predominantly be "no" if it was between yes or no, because people don't like to associate to

extremes, especially with the word choice of "too much."
With options of "sometimes", "rarely," and "frequently,"
it would be more spread out widely, and
likely more truthful, and give more insight

book

Chapter One

1.1

Stats:
- collect
- organize
- analyze
- interpret

- Individuals: people/objects included in the study
- Variable: characteristic of individual to be measured/observed
  - Categorical variable = qualitative variable
- Population data: every individual (complete)
- Sample data: only some individuals (not complete)
- Population parameter: numerical measure that describes an aspect of a population. Sample statistic is same thing for a sample

Levels of Measurement

- Nominal (lowest)
  - names, labels, categories & no order (smallest → largest)
- Ordinal
  - arranged in order, but differences between data are meaningless (same as distance)
- Interval
  - arranged in order, differences are meaningful
- Ratio (highest)
  - arranged in order, differences & ratios are meaningful

Nominal	• can be categorized
Ordinal	• smallest → largest, or [illegible] data can be compared, but not numerical
Interval	• order like above, and take in differences between data
Ratio	• order, take in differences, and find ratio (age are data type)

1.2

Simple random Sample: subset of the population selected in such a manner that every sample has an equal chance of being selected

Other sampling:

- Stratified: ÷ pop into subgroups called strata, draw random samples from each stratum
- Systematic: # all members of a pop, then select every nth member
- Cluster: ÷ into pre-existing clusters (often geographic), and chose random sample of clusters
- multistage: create successively smaller groups at each stage
- convenience: data from pop members readily available

Sample frame: list of individuals in sample

Undercoverage: doesn't match population

Sampling error: doesn't represent

Nonsampling error: poor sample design, sloppy collection, etc

1.3

Planning a Study

1) Identify area of interest
2) Specify variables / protocols
3) Determine: Sample or pop → entire population = census, part = sample
4) Ethics / privacy
5) Collect data
6) Make decisions
7) Note concerns

Simulation!!

observational study

experiment

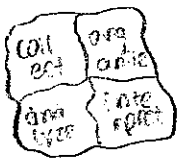
- placebo group + treatment group
  - randomized
- block = individuals w/ common features (gender, age)
  - do treatment to each block + compare results within the block

Survey

- pitfalls:
  - nonresponse
  - truthfulness
  - faulty recall
  - hidden bias
  - vague wording
  - interviewer influence
  - voluntary response

Stats Day 1

code: ufacpft



Statistical Data
- found everywhere in variety of forms
- assists in making informed decisions when faced w/ uncertainty and w/o statistical bias

Stats → collection, organization, analysis and interpretation of numerical info

1. Collect → **Collecting:** (eg observation, survey)
   - samples
   - experiments
   - census (all/everyone)
   - simulations
2. Organize
   - frequency tables
   - charts
   - graphs
3. Analyze
   - center/shape/spread/outliers, gaps
   - trends, as x ↑, y does?
   - measures of central tendency
   - measures of dispersion/spread
   - confidence intervals
   - hypothesis tests
4. Interpret
   - assumptions
   - p-value + Alpha-level
   - confidence level
   - z score or T score or Chi sq score
   - written summary/ conclude in context of problem

Pop Quiz:
i) observational
ii) survey
iii) experiment
iv) anecdote (bad)

"The #s have no way to speak for themselves we speak for them"

Caution ⚠
- it is easy to deceive/lie w/ our statistics
- work diligently to use statistics to show the entire picture + the truth

Stats day 2

Population → N = —

- Measrmt of entire group of interest
- Data from all...
- Quantative variable of all...

must use this wording when asked, what is the population?

Example: Studying all stat students in CA...
Answer: studying the avg # of hrs studied by all stat students in CA

Example 2: Avg # of pages of USBB read of all Students in period 4

Sample → n = —

- part of the popltn
- smaller section of those being measr
- Data from...
- Quantitative variable of...
- random and representative is desired
- (response bias: lying when asked a question in sample)
- (non-response bias: refusing to answer)
- 3x5 cards in box:
- random # table
- calculator chose random
- computer chose random

must use this wording when asked, what is the sample?

yes! for choosing sample)

Example: Avg Number of hours studied by 123 stat Students in CA → same as top, just omit "all"

Example 2: Avg # of pages of USBB read of 12 students in period 4

example

HW problem:

(8) Wt is avg miles per gallon for all new hybrid small cars? Using CR, a random sample of such a vehicles gave an avg of 35.7 mpg.

a) Identify the variable: Avg mpg
b) Quantitative vs qualitative: Quantative
c) What is the implied population? Data from all new hybrid small cars

must say data from...

SRS → simple random sample

- calculator
- computer
- RNT

SRS

Random Number Table in case of inspecting out of 678 cars

1) Number each car 001-678
   - every # has to have same # of digits as the rest
2) Drop it like its hot on a random spot
3) Go any direction but stick in one direction (eg, right)
   - take the first 3 digits that appear in that direction
   - keep anything between 001-678, but toss 678 +. (it is not applicable)
   - toss the entire deal (all 3 digits, not just one digit)
4) Inspect and look at those 7 (or however many samples you have) cars

SRS

Calculator

1) # each student 1-33
2) press math
3) go to prob and then randInt NoRep((amt of winners)
4) put lower as 1, upper as 33, n=4
5) four winners selected! (: woot :)

Dr. Nik video notes

- Sample has to be unbiased + representative
- there will always be variation + sampling error

1. SRS: ideal. list all membas + chose numbers that are all equally likely to be chosen
2. Convenience: bad. chose ppl nearby/close/what is easy. often biased.
3. Systematic: chose a certain number and chose every object apart (eg every 5th object)
4. Cluster: pop ÷ into clusts that are then some clust. chosen by random. Census those clusters chosen
5. Stratified: categorized by characteristics and chose random # from each Strata

Stats Day Three

Inferential Statistics Pic

Process listed below:

Step 0) get a List of everyone

Step 1) Random Sample generated

Step 2) Collect data from that sample

*change pop statement to sample statement (goes from all to whatever number)

Step 3) Analyze

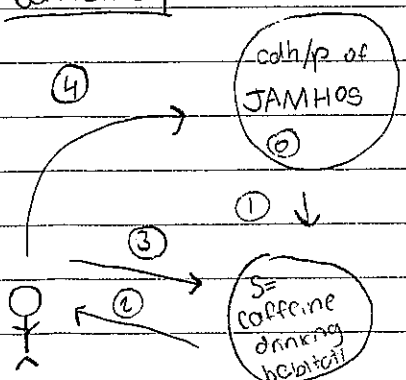
Step 4) Infer (about the whole population from that sample)

*only look at population you drew from, not any other population

↑n = ↑c { • to get higher confidence level, increase sample size!!

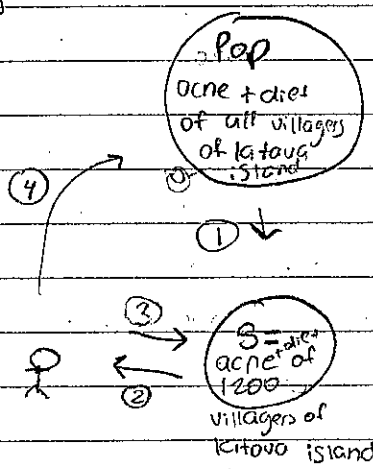
sample size ↓ confidence level ↓

Caffeine:



and personsons on Japanese American Males who are older/live in Hawaii

Acne



Limits

1) Ethnicity (could be genetic)
2) Gender (only on men)
3) Location
4) Age (not old enough)
5) Observation Study (not an experiment)

Bias

Nonresponse: individuals either cannot be contacted or refuse to participate → leads to undercoverage

Voluntary response: those with strong feelings are more likely to respond, skews in one direction / not representative → self selection bias

Convenience sample: asking whoever is around, limits to those present and not representative of population
diff
locations have diff types of people w/ diff opinions

Response bias: lies, whether or not it is on purpose or not, it can be faulty recall or just refusing to say truth

Availability bias: rely on info that comes readily to mind, memories that are easier
recall bias is that it's hard to remember what happened or when, so you may give false response

} thinks that if something is readily/easily available in memory, it must be more common; frequency in memory = frequency in real life

Under-coverage bias: part of a population is excluded from sample

Confirmation bias search for info that reinforces our pre-existing beliefs

Types of Bias

- Non-response bias
- under coverage bias
- Response bias
- Convenience sample/selection bias
- Voluntary response sample
- Confirmation bias
- Availability bias

Stats Day Five

Levels of Measurement

Nominal

- qualitative / categorical
- names, labels, listings
- order not important
- hair color, eye color, car models, students in a class, etc.

Ordinal

- categorical data / qualitative
- can be arranged in order, but differences have no # meaning
- rankings / ratings
  - excellent (4) / good (3) / fair (2) / poor (1), low / medium / high
  - even tho surveys may have numbers they are position values, not actual numerical value

Interval

- quantitative data
- arranged in order AND can subtract; numerical differences between values are meaningful
- no natural zero / starting point / where nothing exists
- eg: olympic years, temperatures, bc 0°C (and 0°F) doesn't = 0 heat
- years because 2021 is 21 years after 2000, but there is no year 0 (0 AD doesn't mean absence of time)

Ratio

- highest level of quantitative data
- order, subtract difference, division makes sense, and a zero exists
- Distance you live from ORHS (miles)
- Length of class before you get bored (minutes)
- Time it takes to read (hours)
- in a system built on a zero
- Ages in years

Ages of Students: Ratio

IQs of Students: Interval (100 iQ)

Rankings of Student Effort: Ordinal

vid notes:

1) Critical mindset of all data, put it thru the test

- dont believe everything you see
- keep emotions in check (don't react abruptly)

2) Do some digging

- Spelling + weird urls
- fact checking website

3) See who else is reporting the same story

4) Dont take image @ face value

- photoshop

5) Check w/ your common sense, does it "sound right"

1) Fake

2) Fake

3) Fake

4) True

5) T

Darya Heydari

10

1.2 #1-4, 9-12, 16, 20

① A stratified sample seperates the population into groups (called strata) that share a common characteristic. Then, a random simple sample of a certain size is drawn from each strata. A cluster also seperates the population into groups, but they are based on demographic area. Then, a certain number of clusters is selected, and every member in those clusters is included in the sample, whereas in stratified sampling, a simple random sample is chosen, not a census, and all groups are included partially.

② A simple random sample makes sure that every individual has an equal chance of being selected, and to get a simple random sample, one would use a calculator, computer, or random number table (or index cards.) In a systematic sample, the population is ordered in a natural sequence. There is no ordering in a simple random sample. Also, in a systematic sample, people are chosen by using a certain number and choosing the person/thing that comes in that order, and so on. There is no counting of a systematic way in a simple random sample.

③ Although Maria did make "some samping error", the advice discredits the fact than samping error can be detrimental to the entire study. Sening size can matter in terms of price, so not considering this can make the data inaccurate and not an "even playing field." Therefore, she should go back and redo it.

④ The sample frame does not include all students enrolled in the college, only a sample of those who use the recreation center. This being the entire sample for a study would be bad because it is undercoverage: it does not

represent the rest of the students at the college who don't use the recreation study, who might have much different opinions that could drastically change the entire study's results. Asking only those at the recreation center will skew it in one direction.

(9) You could get a random sample in many different ways, such as a random number table to get four students with corresponding number

(a) The first few students tend to have similarities such as wanting to get to class early to be prepared and tend to like class more than the rest. This means that it is not random.

(b) Similarly to above, they have certain similarities that the rest don't share, so it might not be random. They could all be in a same different class or have a longer commute. "conv" → UCB

(c) Those in the back row may not be as attentive or have a passion for learning like the others, meaning they might have a similar mindset which is not random.

(d) The tallest students could all be male which undercovers the women in the class and isn't a random sample.

(10) @ The sample frame could omit those who are taking a class at the same day and are in different fields / have different interests.

(b) Homeschooled students

(11) 16, 41, 02, 16, 07, 40, 42, 01, 39, 90
I selected a random part and went right, selecting two digits at a time.

Dana Heydon

(12) 273, 865, 000, 405, 398, 940, 312, 928
I chose a random spot and went down, saving 3 digits at a time.

(16) 254, 227, 152, 182, 050, 121, 145, 168, 001, 101, 028, 161, 281, 131, 325, 062, 008, 135, 118, 082, 081, 100, 178, 144, 309, 153, 139, 192, 032, 227
. Two "people" had the same birthday. I would expect results to be different since it is all random but there is a chance it will also have a shared b-day on there too.

(20) (a) Stratified sample
(b) Simple random sample
(c) Cluster sample
(d) Systematic sample
(e) Convenience sample.

CONTENT OVERVIEW

A **census** is an attempt to gather information about every member of some group, called the **population**. This unit introduces the U.S. Census and its problems in collecting data on the entire U.S. population. One of the most serious problems is undercounting certain segments of the population. Unfortunately not all groups of people are undercounted at the same rates. For example, undercount rates for minority groups are higher than for whites and undercounted rates for renters are higher than for homeowners. Moreover, undercount rates for those living in poverty are higher than for the affluent. The U.S. government uses sampling to estimate undercount rates for various groups. However, it never changes the official headcount number based on the results from sampling.

A **sample** allows the researcher to gather information from only a part of the population. Sampling – collecting data from a portion of the population – is the general means of gathering information about a population when it is not possible to get information from each individual in the population. Sampling saves both time and money. In some cases, such as for Frito-Lay potato chips, both the whole potatoes in a sample and the chips in a sample are destroyed as part of the data collection process. In such cases a census would be out of the question or there would be no product left to sell.

In order for a sample to provide good information about a population, the sample needs to be representative of the population. A **simple random sample**, a sample in which each member of the population is equally likely to wind up in the sample, is one means of ensuring that the sample is representative of the population and not biased. A simple random sample can be selected from the population in the same way that a subgroup is randomly selected from a larger group to receive a certain treatment. Hence, you should refer to Unit 15, Designing Experiments, for directions on selecting a random sample.

Sampling bias occurs when a sample is collected in such a way that some members of the population are less likely to be included than others. A voluntary television poll is an example of a biased sample. Since it is voluntary, only those with strong views are likely to call or text in to vote. Furthermore, only those watching the particular station at the time the poll is given will participate. In this case, the entire segment of the population who do not watch that particular station will be left out of the sample.

KEY TERMS

The entire group of objects or individuals about which information is wanted is called the **population**.

A **census** is an attempt to gather information about every individual in a population.

A **sample** is a part of the population that is actually examined in order to represent the whole. A **simple random sample** of size *n* consists of *n* individuals from the population chosen in such a way that every set of n individuals has an equal chance of being the selected sample.

Sampling bias occurs when a sample is collected in such a way that some individuals in the population are less likely to be included in the sample than others. Because of this, information gathered from the sample will be slanted toward those who are more likely to be part of the sample.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. Are recent U.S. Censuses more or less accurate than early Censuses?

Accuracy has been increasing since early Censuses, but they are still not perfect.

2. Why is the U.S. Census undercount, which is quite small as a percent of the population, so important?

Statistics are important in a democracy and the underrepresented lose representation in congress, and their fair share of funds is lost. (hospitals, schools, social services, etc.)

3. What is a simple random sample?

A sample chosen in a way that each individual has an equal chance in being selected. This helps prevent any sampling bias that could undercover parts of the population

4. How many uses of sampling can you spot in the account of Frito-Lay potato chips?

Cluster Sampling to select which potatos to examine

Systematic sampling to inspect each chip for quality.

CONTENT OVERVIEW

This unit describes methods related to conducting surveys. Particularly when populations are large, geographically-dispersed human populations, it would be nearly impossible to include everyone in a survey. So, one aspect of conducting a good survey is the **sampling design** – the method used to choose a **sample** that is representative of the **population**. Equally important are the design of the questions and interviewer training.

Convenience sampling and **voluntary sampling** are two methods for choosing a sample that may not produce a **representative sample**. In convenience sampling, a sample is chosen in a way that makes it easy to obtain. For example, the pollster could stand outside a grocery store on some weekday morning and interview people as they enter the store. That would be an easy way to get a sample, but the sample probably won't be representative of the opinions of the population -- for one thing, most likely there will be more women in the sample than men, and the sample won't contain people who work weekdays 9 to 5. So the sample will be **biased** toward the views of women who are not working weekday mornings. Voluntary sampling is equally hazardous. A television show might ask people to call or text in their responses. Generally people who feel strongly about a topic are more likely to volunteer.

Using random sampling techniques as part of the sampling plan produces samples that are more likely to be representative of the population. In a **simple random sample**, every person in the population has an equal chance of being chosen for the sample. However, for large populations, a simple random sample can be difficult to conduct. Here are two new concepts of sample design: **multistage samples** and **stratified samples**. For a two-stage sampling process, a sample of clusters is first selected and then random samples within each cluster are chosen. For a stratified sampling process, two or more strata are defined and then random samples are taken from each stratum.

Questionnaire design concerns the wording of questions and the overall order and length of the questionnaire. In terms of wording, consider the following:

- Don't use long words when a shorter word would mean the same thing.
- Stay clear of words that might be unfamiliar to respondents.
- Be sure that questions are neutral and do not lead the respondent in a particular direction.
- Keep sentences relatively short and simple.

- Avoid asking two questions in one – for example, the question "Have you argued with your friends or parents this month?" is really two questions in one.
- Be specific and avoid terms that are vague. For example, words such as "often" or "sometimes" should be replaced by specific terminology such as "every day" or "once a week."
- Finally, interviewers need to be trained not to show their own opinions and not to suggest answers, but to encourage people to respond. In addition, the gender or race of an interviewer needs to be taken into account. For example, people may give different answers about racial issues depending on the race of the interviewer.

KEY TERMS

The **population** is the entire group of individuals about which information is desired. A **sample** is a subset of the population from which information will be extracted. A **representative sample** is one that accurately reflects the members of the entire population. A **biased sample** is one in which some individuals or groups from the population are less likely to be selected than others due to some attribute.

A **sampling design** describes how to select the sample from the population. There are many sampling designs, including the following:

- **Simple random sampling** is a sampling design that chooses a sample of size *n* using a method in which all possible samples of size *n* are equally likely to be selected.
- **Convenience sampling** is a sampling design in which the pollster selects a sample that is easy to obtain, such as friends, family, co-workers, and so forth.
- **Voluntary sampling** or **self-selecting sampling** is a sampling design in which the sample consists of people who respond to a request for participation in the survey.
- **Multistage sampling** is a sampling design that begins by dividing the population into clusters. In stage one, the pollster choses a (random) sample of clusters. In subsequent stages, random samples are chosen from each of the selected clusters.
- **Stratified sampling** is used to ensure that specific non-overlapping groups of the population are represented in the sample. The non-overlapping groups are called **strata**. In a **stratified random sample**, the sample is obtained by taking random samples from each of the strata.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. Why was the *Literary Digest* poll so far wrong in predicting the outcome of the 1936 presidential election?

   Literary Digest drew their sample from a list of car and telephone owners, items that were expensive and indicative of wealth, which omitted the large Pro-Roosevelt poor population.

2. Why would a simple random sample of counties in a state give results that might not represent the entire state?

   It could leave out groups of interest, in this case counties, and undercover different, important parts of a state with differing opinions.

3. In sampling, what are strata?

   Groups with similar characteristics

4. You are an interviewer for an opinion poll. How should you react to answers that seem anti-social or immoral?

   You should respond in a diplomatic and respectful way.

KEY TERMS

In an **observational study** researchers observe subjects and measure variables of interest. However, the researchers do not try to influence the responses. The purpose is to *describe* groups of subjects under different situations. In an **experimental study**, researchers deliberately apply some treatment to the subjects in order to observe their responses. The purpose is to study whether the treatment *causes* a change in the response.

In a **double-blind** experiment neither the subjects nor the individuals measuring the response know which subjects are assigned to which treatment. In a **single-blind** experiment the subjects do not know which treatment they are receiving but the individuals measuring the response do know which subjects were assigned to which treatments.

A **placebo** is something that is identical in appearance to the treatment received by the treatment group but has no effect.

A **control group** is an experimental group that does not receive the treatment under study. The control group could receive a placebo to hide the fact that no treatment is being given. In an **active control group**, the subjects receive what might be considered the existing standard treatment.

The explanatory variables in either an observational study or experiment are called **factors**. A **treatment** is any specific condition applied to the subjects in an experiment. If an experiment has more than one factor, then a treatment is a combination of specific values for each factor.

Two factors (explanatory variables) are **confounded** when their effects on a response variable are intertwined and cannot be distinguished from each other.

Darya Heydari

THE VIDEO

1) Random Assignment
2) Comparison
3) Double Blindness
4) Replication
5) Sample size

10

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. Why is the study of the effect of humans on the coral reefs *not* an experiment?

They only studied pre-existing coral reefs and humans influence, they did not try to influence it, only take a look and collect data. No influence on humans or the reefs; or interference / imposition

2. Who were the subjects in the Glucosamine/Chondroitin study? What did researchers want to find out?

Patients with osteoarthritis in their knees. They wanted to see if osteoarthritis pain would go away. (the response variable is reported reduction in knee pain.)

3. Why were subjects randomly assigned to the treatments?

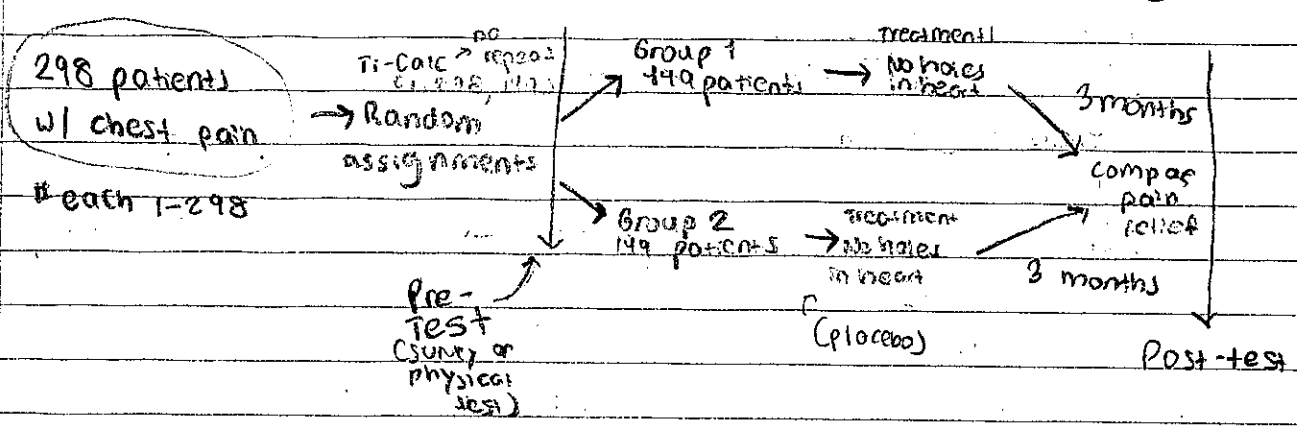
They were randomly assigned to the treatment to eliminate bias and produce accurate results.

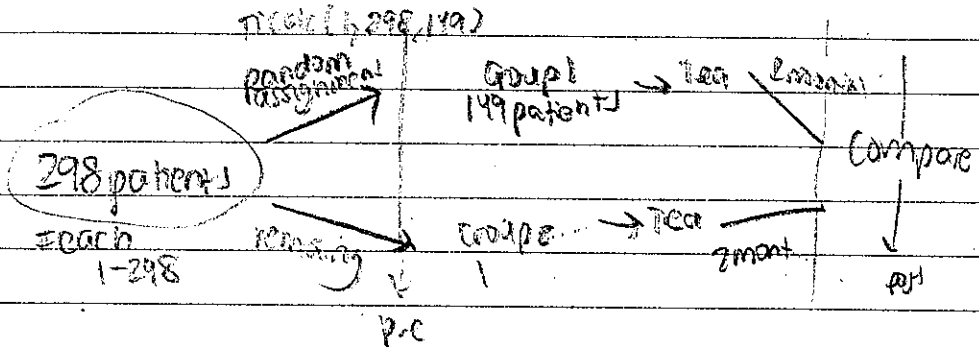
4. Dr. Confound conducted a very badly designed experiment on mood-altering medication. List some of the problems with his experiment.

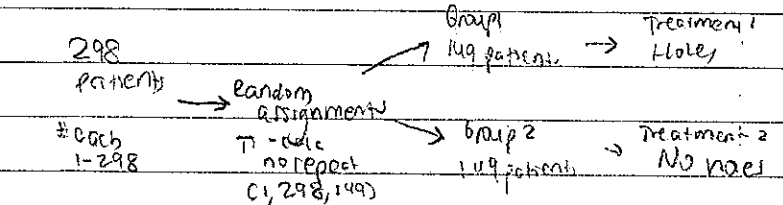
- He has the subject state the mood in front of other people which can influence other people's responses.
- The experimenter bias: he is being kind which can alter the mood without the actual medication
- Confounding variable: one sits and one stands which can affect their physical well-being. The one standing will be annoyed standing for a while, but sitting can be relaxing
- Didn't randomly assign the pills, it is obvious which one got the placebo: lack of blindness
- Manipulating the numbers: bumped up the number to fair when told "between bad and fair"
- Sample too small

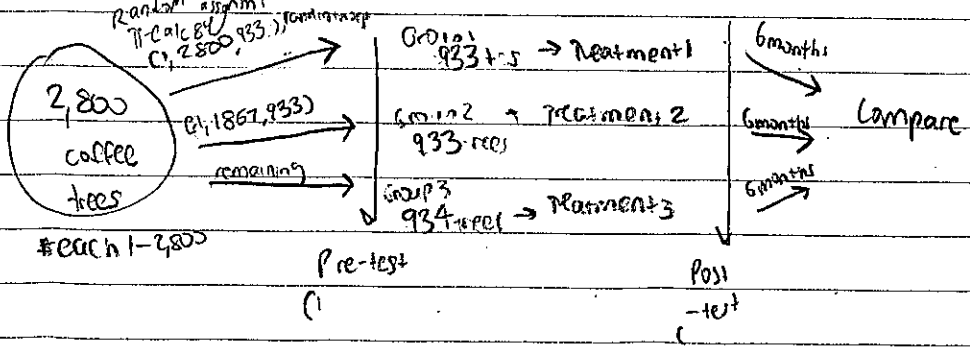
1.3 - Experiments

Experimental outline Design

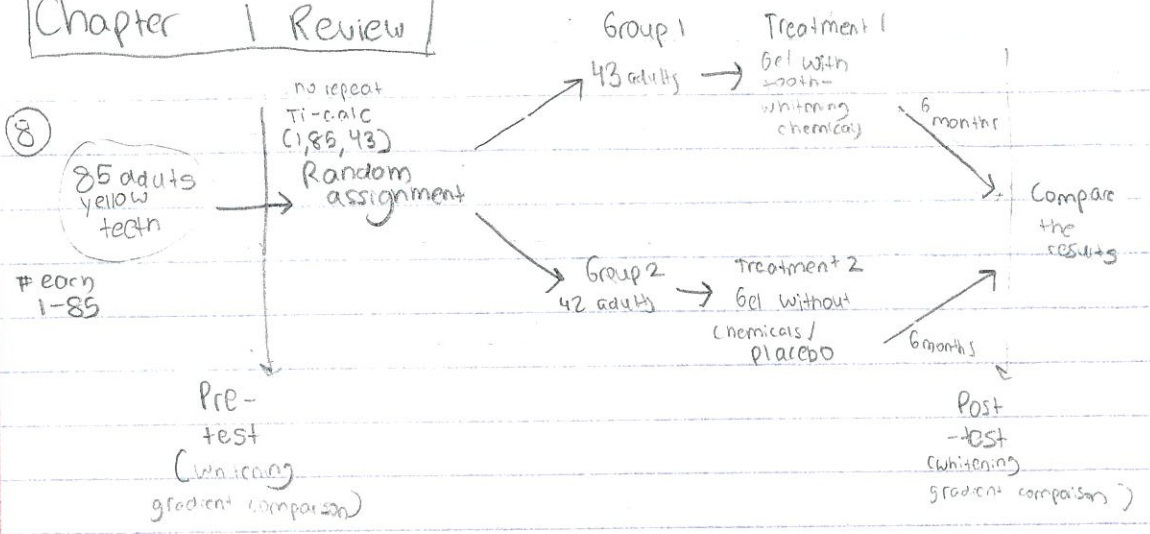




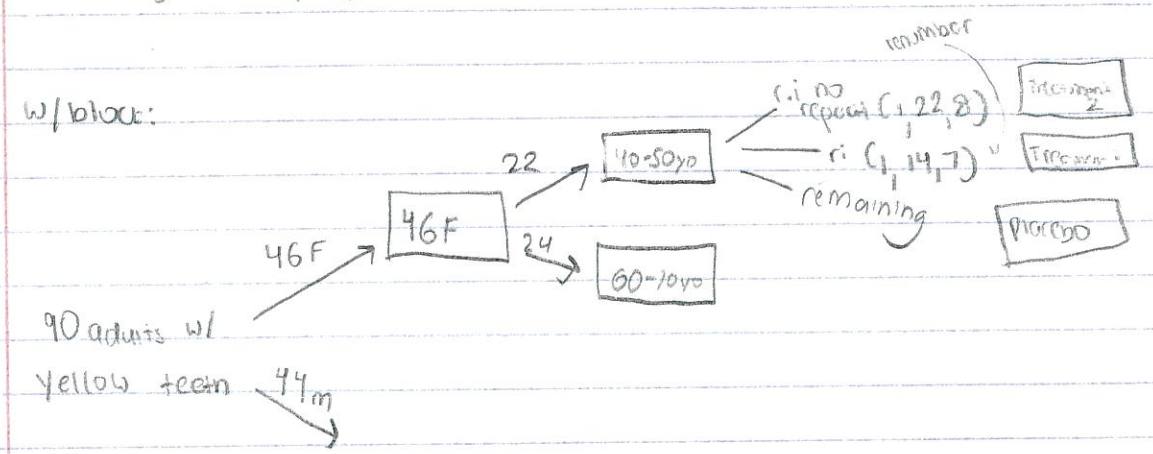




Chapter 1 Review



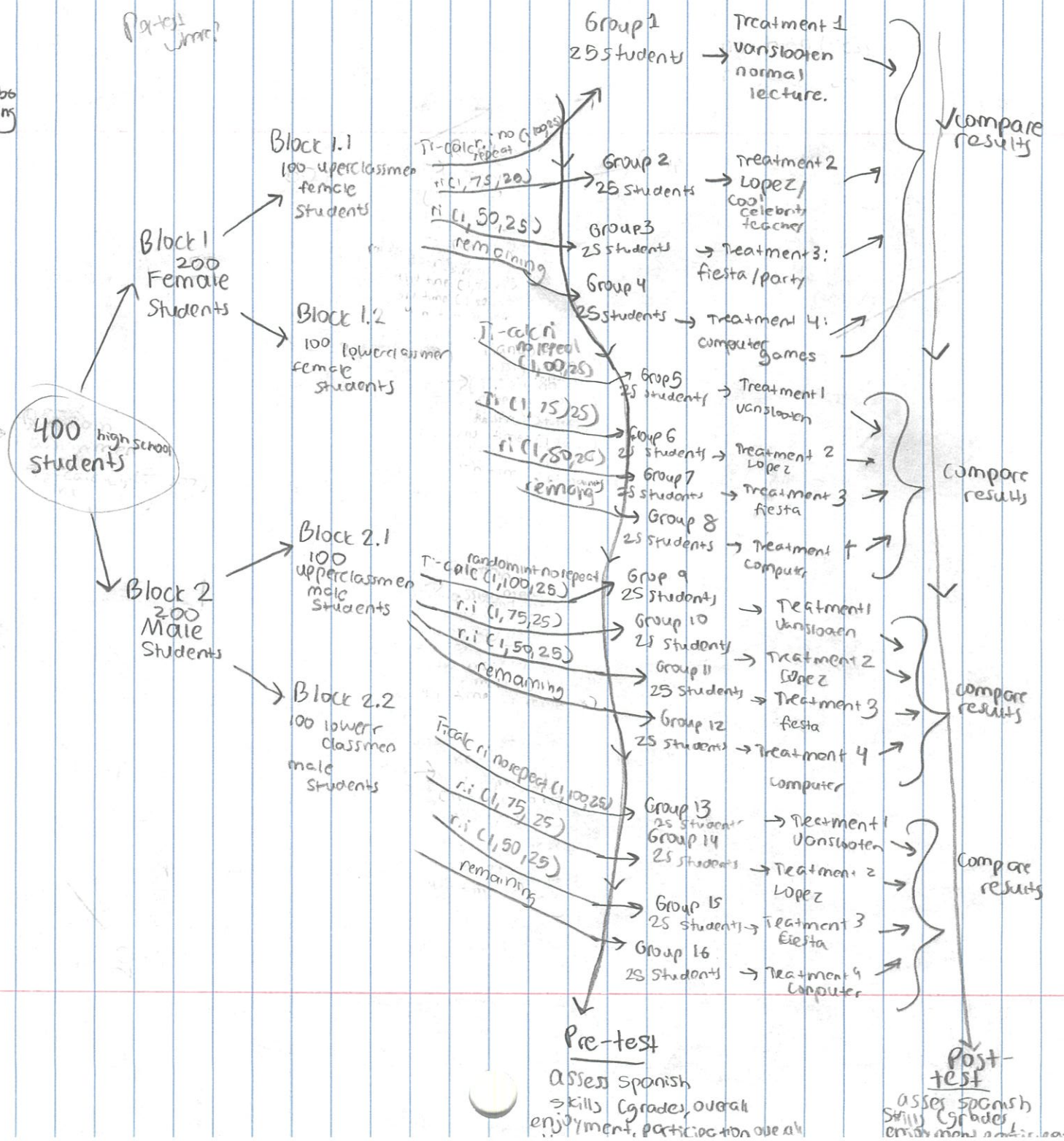
w/ block:



Dasha Heydari
Period 4

Stats Spanish Learning Extra Credit

no placebo blinding



Sim Notes Lecture

Simulation: Imitation of chance behavior, based on a model that accurately reflects the experiment under consideration

Steps

1) State the problem (what are we talking about?)
2) Assumptions
3) Assign digits (similar to real outcomes) *
4) Many reps (do it a lot!) 25+
5) State your conclusion (sim & reality)

Example: coin flipping simulation

1) Hot streak with coins (3+ of the same)
2) Heads & Tails, equally likely
   Each toss independent of the others
3) 1 digit = 1 coin flip (on RNT)
   odd # = hed, even # = tails
   10 #10 coin flips = 1 Rep!!
4) Repeat many times 25+
5) State concl

5.56

1) State prob.:
? 10 of students ab evening exams
2) assumptions
Y ≠ N, 80%, 20%,
not in a group, independent
3) 0-7 yes, 8-9 NO
4) 10 #s = 10 students = 1 Rep
5) Conclude: 3/25
times all 10 students said
Yes. There is a 12%
chance (approximately)

Stats Survey

Tokyo v Rio Olympics

raising: scary for ppl to say honest opinion, people are looking around and can be self-conscious + not express truthfulness (group setting)

- verbal question: dont do them verbal anymore
- leading/suggestive words
- city names instead of year
- x vs x-4, 2016 is less recent than 2021, don't remember Rio as much

Did you enjoy watching Tokyo Olympics

- assumes that we've seen it, there is no way to say "didn't watch"
- good bc isn't in a group setting, more individual
- good bc written and non verbal
- "Did you enjoy" is leading/suggestive
- City: Tokyo (evokes images of no direction to olympics)
- no distinction between strong yes and meek yes
- not quantitative

Please rate the 2020 olympics:

- good bc an even number of choices (1-4), if 5 keeps them on the fence
- no leading/suggestive words, bland basic words is good
- no longer "Tokyo" just 2020
- more choices than yes/no, can express levels of satisfaction
- still only just have ordinal data (ratio is ideal)

Please log hours spent watching Olympic games (by date)

- Ratio
- right now survey, not looking back to the past

Clear, concise?, help feel that its anonymous, assist them in truthful answers (calm anxiety, fear, fears), motivate them to respond, (w incentives)

unforgettable!

Bush Survey

5-60

a 1) Simulate batting average up to bat a lot tracking hits

2) Assumptions: actually batting .320 right now, it is not late in the season, no injury, pitching is normal/average, not in Colorado, normal bat

3) 00-31 = hit
32-99 = miss

40 digits = 20 number = 20 swing = 1 rep

4) 19 22 39 50 34 08 75 62
hit hit no no no hit no no

87 13 96 40 91 25 31 72 54
no hit no no hit no hit no no

48 28 53
no hit no

$\frac{7}{20}$ = 35% → Rep 1

73 67 64 71 50 99 40 00
no no no no no no no hit

10 23 27 75 44 26 42 82 42 53 62 90
hit hit hit no no hit no no no no no no

Rep 2 → $\frac{6}{20}$ = 30%

46 46 77 17 09 77 55 80 00 95 32 85 85 94
hit hit hit

85 82 22 09 09 56
hit hit

Rep 3 → $^5/_{25}$ = 20%

52 71 13 88 89 93 07 46 02 23 40 01
hit hit hit hit hit hit

18 58 45 45 76 76 25 73
hit hit

Rep 4 → $^6/_{25}$ = 24%

Rep 5 = $^6/_{25}$ = 24%. Rep 10 = $\frac{4}{25}$ = 16%. Rep 15 = $\frac{7}{25}$ = 31%.
Rep 6 = $\frac{6}{25}$ = 24%. Rep 11 = $\frac{5}{25}$ = 20% Rep 16 = $\frac{7}{25}$ = 31%.
Rep 7 = $\frac{4}{25}$ = 16%. Rep 12 = $\frac{6}{25}$ = 24%. Rep 17 = 4/25 = 16%.
Rep 8 = $^7/_{25}$ = 31%. Rep 13 = $\frac{9}{25}$ = 36%. Rep 18 = $\frac{6}{25}$ = 24%.
Rep 9 = $\frac{5}{25}$ = 20%. Rep 14 = $\frac{8}{25}$ = 32%. Rep 19 = $\frac{3}{25}$ = 12%.
Rep 20 = $\frac{5}{25}$ = 20%.

5) The average frequency of at-bats in which the player gets a hit is 23.8%.

c) This is a lower number than the player's actual average (.235 vs .320)

5-61

2) ~~Assumptions: There are no twins in the group, Students~~ in class are randomized

For 41 Students:

I decided to do 25 reps of a simulation where I put in my calculator (rand int(1,365,41)) so that the calculator would chose 41 people and see if any had the same birthday (days were #d 1-365)

Rep 1: 3 had same bday (98), 2 had same bday on 128, 2 had same bday on 131, 2 had same bday on 209 (YES) Yes

Rep 2: Yes
Rep 3: Yes
Rep 4: Yes
Rep 5: Yes
R6: NO
R7: Yes
R8: Yes
R9: Yes
R10: Yes
R11: NO
R12: Yes
R13: Yes
R14: Yes
R15: NO
R16 Yes
R17: Yes
R18: YES
R19: Yes
R20: Yes
R21: YES
R22: Yes
R23: Yes
R24: Yes
R25: Yes

3) 1-365

4) 25 reps, 41 # = 1 Rep

5.) Conclusion

22 out of 25 times, at least 2 people share a birthday, which is 88%. This matches up and is close to 90%.

For 23 students:

R1: NO	R10: YES		
R2: Yes	R11: Yes	R20: YES	Conclusion:
R3: Yes	R12: NO	R21: Yes	$\frac{12}{25}$ times
R4: NO	R13: YES	R22: Yes	two people or
R5: Yes	R14: NO	R23 Yes	more shared
R6: NO	R15: Yes	R24: NO	the same
R7: NO	R16: NO	R25: NO	birthday, or
R7: Yes	R17: NO		48% of the
R8: NO	R18: NO		time which is close
R9: NO	R19: NO		to 50%.