Stats Bootcamp

sema barlas 02/28/19

SESSION 1

- Assignment: (10 pts) (03/06) → 9/10
  ch 1-4 - odd # questions

Outline
- Distribution of data
- Graphs
- measures of central tendency and dispersion
- correlation

① INTRODUCTION

A) what are statistics? → "Facts & Figures"
- Field of study concerned w/ summarizing data, interpreting data, & making decisions based on data
* Any calculated in sample to estimate population
- importance of interpretation!
? statistics includes:
  - numerical calculations
  - graphs
  - interpretations & decisions based on #'s and graphs
? ice cream consumption & drowning } correlation ≠ causation
  - third variable involved - summer

B) Importance of statistics:
- credibility
- question the statistics you encounter
? Certain references?
  - need to know more
? Evaluating statistical claims:
  - statistics presented
  - sources of statistical findings
  - procedures used to generate the claims

C) Descriptive statistics
- #s used to summarize & describe data
- data: info. collected from an expt., survey, historical record, etc.
- don't involve generalizing
  ↳ inferential statistics

☐ Ave salary for occupations in 1999
☐ # of unmarried men / 100 unmarried women in US metro areas 1990

②

☐ winning Olympic times for men & women (since 1896)
- can compute desc stats from the data

? 2 basic divisions of statistics
- inferential
- descriptive

? Desc. stat
- summarize & describe data

? Desc. stats
- mean of age
- # of ppl
- median
- height

1) Inferential statistics
- statistics: rely on sample to draw inferences from population
- sample selection challenged
  - over-representation
  - non-random sampling
  - non-proportional
- SIMPLE RANDOM SAMPLING
  - every member of population has to have an equal chance of being selected into the sample
  - selection of 1 member must be independent of every other member
  - "chooses a sample by pure chance"
  * Bias
- SAMPLE SIZE MATTERS
  - sampling procedure, rather than results define 'SRS'
  - small sample size may not be rep. of the pop.
- MORE COMPLEX SAMPLING
  - when SRS not feasible

(1) Random Assignment
- populations often hypothetical
- random division of sample into 2 groups
- critical for expt validity

Ef - Failure to randomize → invalidates experimental findings
- non-random sample → restrict generalizability

(2) Stratified sampling
- can be used if population has # of distinct strata / groups
  1. identify members of sample in each group

2. randomly sample from subgroups of proportionalities

? our data come from ~~samples~~, but we really care most about populations.
? A random sample:
- stratified sampling more rep. of pop than random sampling
? participants assigned for RSch study put into treatment groups on basis of chance
- Random Assignment
? uncertainty re. conclusions about popularm. can be eliminated if
- obtain data from all members of the pop.
? using a Random sample:
- is to accept uncertainty abt conclusions
- enables you to calculate statistic
- is to risk drawing wrong conclusions about population
? Random sample
- equal prob of selection
? SRS of student body:
- ID #'s picked from a table of random #'s
? A Biased sample:
- will likely have groups from the population over-represented or under-represented due to systematic sampling factors.

◁ sampling simulations (Concise book)
- SRS vs stratified SRS

E) VARIABLES

- **independent & dependent variables:**
  - var: property or char that takes on value
  - independent var = manipulated by an experimenter (x)
    - determine effect of independent var on dependent var (y)
  - Levels of independent variables
    - experimental & control → 2 levels
    - 5 diets → 5 levels
    - # of levels = # of experimental conditions
- **Qualitative & Quantitative var:**
  - do not imply numerical ordering → QUALITATIVE
  - Qualitative Attribute (eg gender) →
  - measured numerically (eg. height) → QUANTITATIVE
- **DISCRETE & CONTINUOUS VARS.**
  - DISCRETE - possible values are distinct points on a scale (eg. # of children)
  - CONTINUOUS - continuous scale (eg. response time)

④

03/03/19

F) PERCENTILES

- percent of scores lower than yours (no universally accepted def)

$$\text{Rank} = \frac{P}{100} \times (N+1)$$

P = desired percentile

N = # of entries

R = integer then R, otherwise integer portion, fractional portion: IR, IR+1
↳ interpolate by multiplying difference

- DEFINITIONS

1) "the lowest score that's greater than 65% of the scores"
2) "smallest score greater than or equal to 65% of scores"
3) Rank:

① Find R of percentile desired

R=

② integer: then Pth percentile is # at Rank R

Percentile = 9 ; Rank = 2

R=2.25

② R ≠ integer, find Pth percentile by interpolation

③ IR = 2

④ FR = 0.25

⑤ Find scores at Rank IR & Rank IR + 1
= 2, 3 (Rank) (not scores)

⑥ (.25)(7-5) + 5 = 5.5

Query? multiply diff btn scores by FR then add to lower IR

G) LEVELS OF MEASUREMENT

- Types of scales:
  - measure dependent var depending on type

① NOMINAL SCALES:
- > names / categorized responses
- > no ordering
- > lowest level of measurement

② ORDINAL SCALES:
- > ordered
- > allow comparisons of degree wtc 2 subjects passed dpt var
- > BUT: diff btn 2 levels can't be assumed to be same
  (diff btn adjacent scale may not rpst eql intervals on underlying scale)
  ↳ even if numbers are used

③ INTERVAL SCALES:
- > numerical scales w same interpretation throughout
- > eg. Farenheit
- > BUT: no true zero point → no ratio comparisons.

- measurement demonstration
  - interval vs ordinal

④ RATIO SCALES:
> most informative scale
> interval with zero point (all 3 scales into 1)
> labels, ordered, diff scale = same meaning, ratio
> eg. Kelvin, amt of $

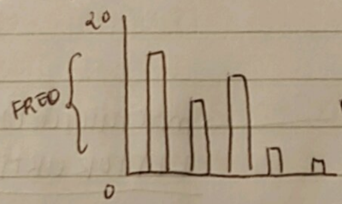
⊛ psychological var:
> 5-7 pt scale —> ordinal (pain)
> memory = ratio, but early u hard
} generally "inappropriate"

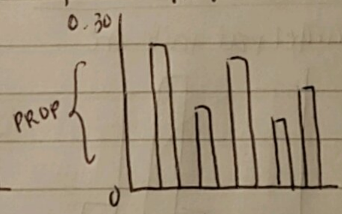
- CONSEQUENCES OF LEVEL OF MEASUREMENT:
  - helps in statistical analysis

H) DISTRIBUTIONS

- DISTRIBUTIONS OF DISCRETE VARIABLES:
  - Frequency table, freq dist., Also prop. —> probability dist

COLOR	FREQ



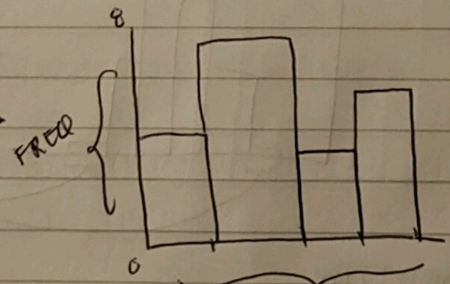


- CONTINUOUS VARIABLES:
  - eg. time to respond —> grouped freq dist since entries are unique (ranges) = tabulated
  - histogram

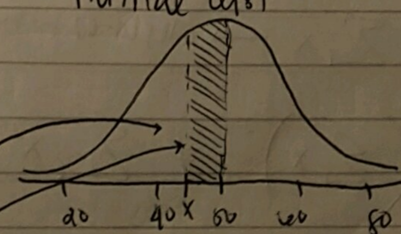
grouped freq dist

Range	freq

—>



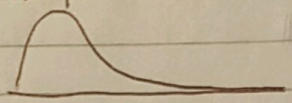
- Probability Densities:
  - represent all potential times at once —> plot dist of cont var. (cont dist)
    - "PROBABILITY DENSITY"
    - normal dist



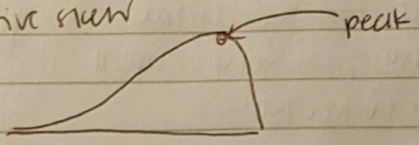
① Area = 1
② prob of any exact value of X = 0
③ prob that rand # will fall btn 2 pts

-shapes of distributions:

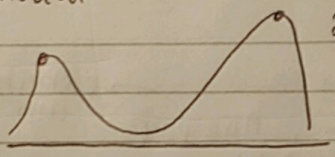
- symmetric
- tails → longer + tail = positive skew (skewed to the right
  - large / extreme



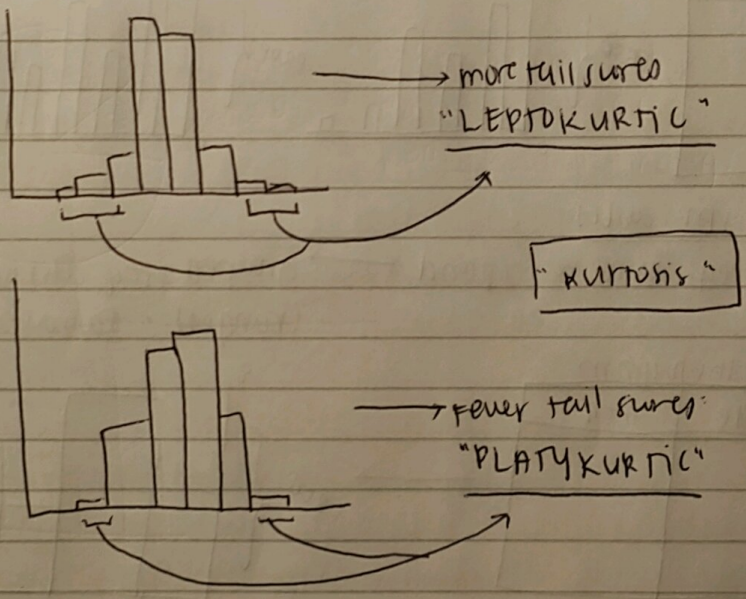
- Negative skew



- Bimodal dist



- How large/fat tails are



I) SUMMATION NOTATION:

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

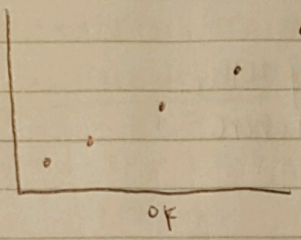
↑ index var

$\sum x$ → all scores to be summed

$$\left(\sum x\right)^2 \neq \sum x^2$$

J) LINEAR TRANSFORMATIONS

- Transform data from one measurement scale to another
- Form a straight line = "linear"



K) LOGARITHMS

- log transformation reduces positive skew
- BASICS OF LOGS:
  - opposite of exponents
    $10^2 = 100$
    $\log_{10}(100) = 2$
    $\log_{10}(1000) = 3$
  - $e = 2.718$ → natural log → $\ln(x)$ $\log_e(x)$
  - changing the base of a log changes the result by a multiplicative constant:
    $\log_{10}$ to $\ln$ → $\times 2.303$
    $\ln$ to $\log_{10}$ → $\div 2.302$
  - "antilog" undoes logging → raise base to N
- LOGS and proportional change:
  - proportional raw changes are equal in log units
  - Arithmetic ops:
    $\log(AB) = \log(A) + \log(B)$
    $\log(A/B) = \log(A) - \log(B)$

L) STATISTICAL LITERACY

- case study

⑧

03/03/19

② GRAPHING DISTRIBUTIONS

A) QUALITATIVE VARIABLES:

- graphing: first step in data analysis
- William Playfair 18th c., John Tukey 20th c.
- iMAC computer expansion of Apple's market share
  - 500 iMAC users interviewed
- no ordering pre-established

① Frequency Tables:

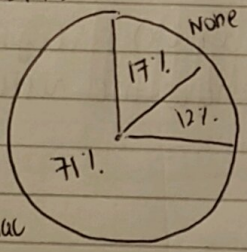
Result	Freq	Rel. Freq
None	x	z
window		
mac		
TOTAL	y	

Response categ.

proportion of responses in each category

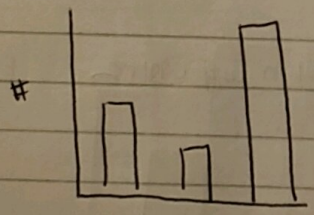
$z = \frac{x}{y}$

② Pie charts:



- size proportional to % responses
- relative freq × 100
- good for small # categ.

③ Bar charts:



- comparing distributions:
  - Bar charts ⟶ diff btn dist (oriented horiz)
- Graphical mistakes to Avoid:
  - 3D Bar charts
  - lie Factor = size of effect shown in a graph : size shown in data (Edward Tufte) ⟶ >1.05 or <0.95 = Bad
  - setting baseline ≠ 0
  - using line graph for Qual vars

B) QUANTITATIVE VARIABLES

- measured on numeric scale

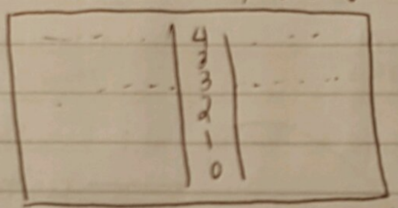
1) Stem & Leaf Displays

- Best for small/moderate amts of data
- # of touchdown passes

3	2 3 3 7
2	0 0 0 1 4 2 3
1	8 7 6 5
0	6 9

↑ tens — ↑ ones

- to clarify shape of dist.
- can also split
- "Back-to-Back stem & leaf display"



→ football rating

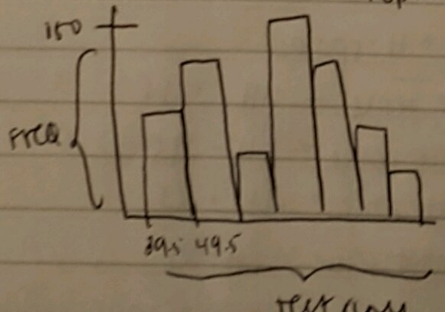
1998 data — 2000 data

Accuracy
- effects of timing
- rounded to be easy to do...
- useful up to N = 200

} 2 place accuracy

2) Histograms

- shape of distribution → large # of observations
- steps
    1. create freq. table (intervals) — class intervals
       ↳ place intervals midway 0.5 b/n 2 #'s
    & class freqs represented by bars



# of from 5 to 10 (bin widths) - Sturges Rule - Rice Rule

→ discrete

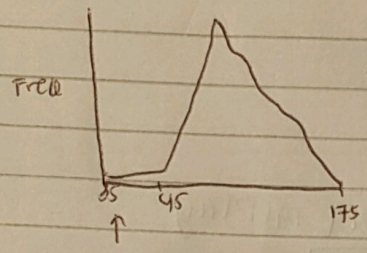
- for continuous, can bucket into whole #s

- can also be based on relative frequencies rather than actual frequencies
    ↳ proportion rather than #

3) Frequency Polygons → comparing distributions (overlaying)

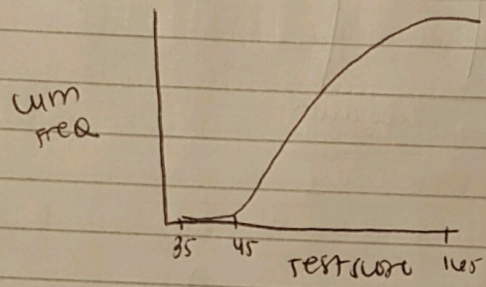
- shape of distributions
- similar to histograms but for sets of data
- cumulative freq dist

steps:
1. choose class interval
2. x-axis = score values
3. middle of int w/ bar, label w/ mid value
4. y-axis = freq of class
5. connect the pts



represents interval from 29.5 to 39.5

← Frequency Polygon



← cumulative frequency polygon

4) Box plots:

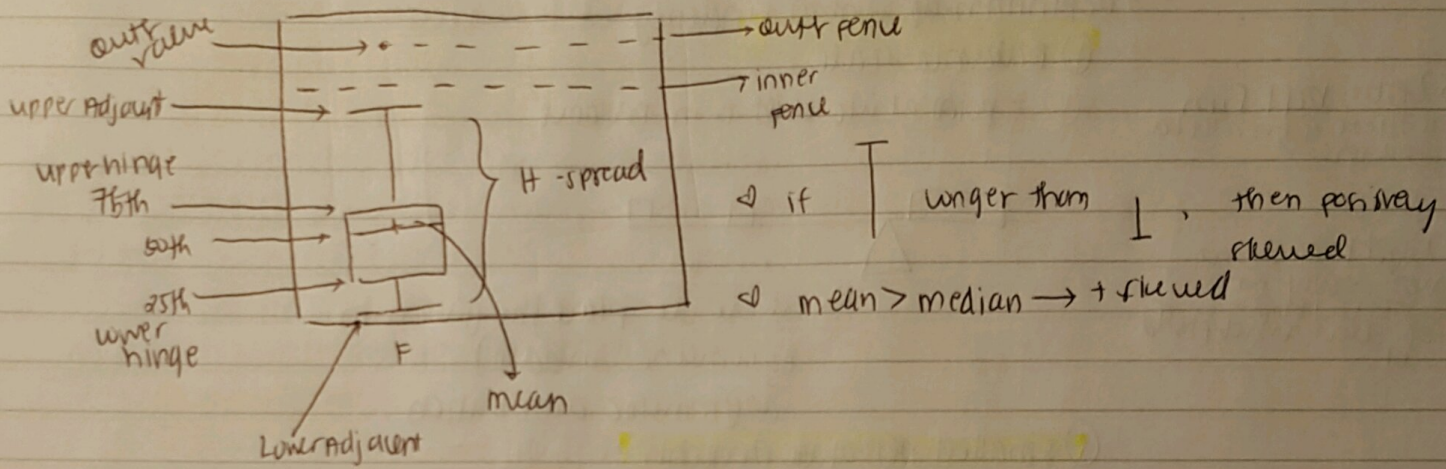
- identifying outliers & comparing distributions

* steps:

1. 25th, 50th, 75th percentiles
2. check p. 94 for terminology

terminology:
- upper hinge (75th), lower hinge (25th)
- H-spread = UH − LH
- step = 1.5 × H-spread
- upper inner fence = UH + 1S
- lower inner fence = LH − 1S
- upper outer fence = UH + 2S
- lower outer fence = LH − 2S
- upper adjacent – largest value below UIF
- lower adjacent – smallest " above LIF
- outside value – beyond IF but not OF
- far out value = value beyond OF

- whiskers → UH & LH to UA & LA.
- mean = +
- OV = o
- FO V = *



- variation on Boxplots

4) Bar charts:

- frequency counts
- % increases
- change over time
- compare means of diff experimental conditions = Box plot

5) Line Graphs:

- Bar graph w top of Bar represented by pt joined by line (rest of the bar is supressed)
- appropriate when both x and y axis display ordered (rather than qualitative) variables
- changes over time

6) DOT Plots:

- various information

C) STATISTICAL LITERACY:

Eye color	Freq	Rel Freq
Blue	10	10/26
Brown	11	11/26
Grey	1	1/26
Green	4	4/26
Total	26	26

→ Relative Freq table