(15)

03/04/19

③ SUMMARIZING DISTRIBUTIONS

A) CENTRAL TENDENCY:

1) What is Central Tendency?

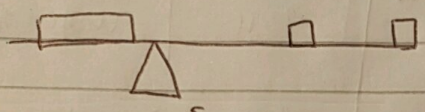
- location of the center of a distribution → comparing individual scores to dist.
- 3 definitions of central tendency:

① Balance scale:

- point at which dist is in balance

*) Abs. diff sim
- is there a formula for this?

*) Sqd Diff Demo
- not really sure about the point of this whatever it



↖ where to place the fulcrum (geometric middle)
asymmetric distribution

② Smallest Absolute Deviation

- number for which the sum of absolute deviations is smallest

③ Smallest Squared Deviation

- target that minimizes the sum of squared deviations

*) Balance scale simulation

- Positive skew: mean > median
  ↖ Balances at mean
- Balanced Unskewed: mean = median
- Negative skew: mean < median
  ↖ Balances on the mean
- Bimodal Dist: Bal Pt = mean

→ if the center of a distribution is defined as the balance pt, then the mean is at the center of the distribution

2) measures of central tendency:

① Arithmetic mean

$$M = \frac{\Sigma X}{N}$$

& geometric mean

$$M = \frac{\Sigma X}{N}$$

② median

- midpoint of the distribution → same # of scores above & below it
- 50th percentile
- mean of 2 middle #'s

③ mode
- most frequently occuring value
- continuous → computed from grouped freq dist → midde of interval

3) median and mean:
- mean = point on wic a dist would balance
- median: value that minimizes the sum of abs deviatns
- mean: value that minimizes sum of squared dev.
- symmetric dist: mean = median
- Bell shaped dist: mean = median = mode

*) mean & median demo:
- depends on the skew of distribution

4) Additional measures of central Tendency

① TRIMEAN
- weighted ave of 25th, 50th, and 75th percentile

$$\text{trimean} = \frac{(P25 + 2P50 + P75)}{4}$$

* why is it over 4 and not 3? → 2 P50

② GEOMETRIC MEAN
- multiply all #'s then take nth root of product

$$\left(\prod X\right)^{\frac{1}{N}}$$

eg. 1, 10, 100

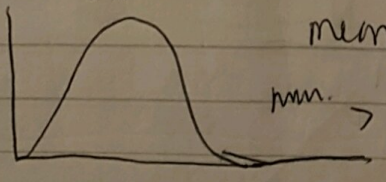
$1 \times 10 \times 100 = 1000$

$\sqrt[3]{1000}$ (3 #'s) → 10

the analog of Arithmetic mean is geometric mean.
- close relationship to logs
- geometric mean only make sense if all the #'s are positive
- good for averaging rates

③ TRIMMED MEAN
- hybrid of mean & median
- remove some of the highest & lowest & compute mean
- 10% → Remove top 5% and Bottom 5%.

5) comparing measures of central Tendency
- symmetric → mean = median = midmean = trimmed mean = mode
  - ↳ except Bimodal
- differences over skew
- Positive skew



mean > median > trimean & trimmed mean > midm > geom mean

03/05/19

B) VARIABILITY

- how much the #'s in a dist. differ from each other

1) What is variability?:

- "spread" OF scores
  - ↳ "variability", "spread", "dispersion"
- measures of variability:

① RANGE

highest score − lowest score = range

② INTERQUARTILE RANGE:

- range of middle 50 % of scores in a dist

IQR = P75 − P25

- "H-spread" → upper hinge − lower hinge
- SEMI-INTERQUARTILE RANGE:
  - → IQR / 2
  - − symmetric dist: SIQR contains 50% scores in dist.

(median)

③ VARIANCE

- "how close the scores are to the middle of dist"

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$ − sample

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$ sample var used to estimate − population var.

M: mean of sample taken from pop'l mean μ

- other formulas:

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

④ STANDARD DEVIATION

- square root of variance
- useful when dist is normal or approx normal bec. prop. of dist w/in given # of SDs from μ can be calculated
- σ = population SD

  s = sample SD

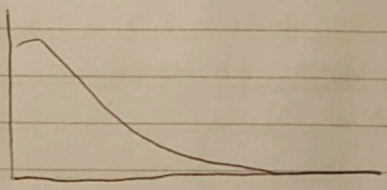
Ⓐ VARIABILITY SIMULATION

Ⓑ VARIANCE ESTIMATION SIMULATION

2) SHAPES OF DISTRIBUTIONS:

- numerical indexes for measure of shape

① SKEW



- large positive skew
- mean > median

- Pearson numerical index of skew:

$$\frac{3(\text{mean} - \text{median})}{\sigma}$$

- third moment about the mean:

$$\sum \frac{(x-\mu)^3}{\sigma^3}$$

② KURTOSIS

$$\sum \frac{(x-\mu)^4}{\sigma^4} - 3$$

Ⓐ normal distribution → to define no kurtosis

Ⓐ) COMPARING DISTRIBUTIONS SIMULATION

3) EFFECTS OF LINEAR TRANSFORMATIONS:

- if a var X has mean μ, SD of σ and var σ^2, then new var Y created using linear transformation

$$Y = bX + A$$

will have mean $b\mu + A$

SD $b\sigma$

var $b^2\sigma^2$

4) variance sum law I

$$\sigma^2_{sum} = \sigma^2_M + \sigma^2_F$$

↑ var(males) ↑ var(females)

$$\sigma^2_{difference} = \sigma^2_M + \sigma^2_F$$

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$ → VARIANCE SUM LAW

- only applies when vars are independent

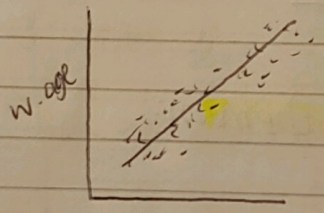
5) Statistical literacy:

④ DESCRIBING BIVARIATE DATA:

- Dataset w 2 vars = "bivariate"

A) INTRODUCTION TO BIVARIATE DATA:

- summarizing data similar to/analogous to univariate
- eg. age
  - Histogram, μ, σ
  - scatter plot



"positive association" - relationship

- cluster along else "linear"

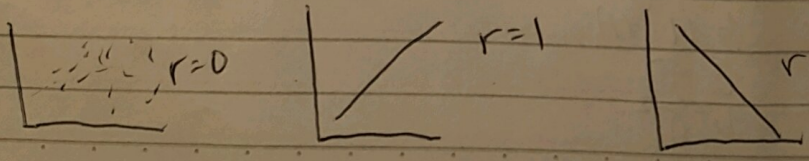
B) VALUES OF THE PEARSON CORRELATION

- "Pearson product-moment correlation coefficient"
- measure of strength of the linear relnshp b/w 2 vars.
  - ↳ if not linear, doesn't adequately report relnshp
- ρ = population
  r = sample
- [-1, 1] range

1 = perfect positive relnship

-1 = perfect negative relationship

0 = no linear relnship



C) PROPERTIES OF PEARSON'S r

- symmetric: "the corr of x with y = corr of y with x"
- unaffected by linear transformations
- range from -1 to 1

@ guessing correlation demo

D) COMPUTING PEARSON'S r

x, y → deviation scores from the mean (means of x, y = 0)
X, Y → variables for wc to get corr of
xy → high value for sum of Σxy means + corr, and low val for Σxy means negative corr

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \longleftarrow \sqrt{\Sigma x^2 \Sigma y^2}$$

"r should be the same despite linear transformations"

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

} alternative formula.

~~) Range restriction demonstration~~

E) VARIANCE SUM LAW II

- when X and Y are not assumed to be independent — aka. when X and Y are correlated.

POPULATION
$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\rho \sigma_X \sigma_Y$$

↑ corr b/tn x and y in pop.

SAMPLE
$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2r s_X s_Y$$

F) STATISTICAL LITERACY