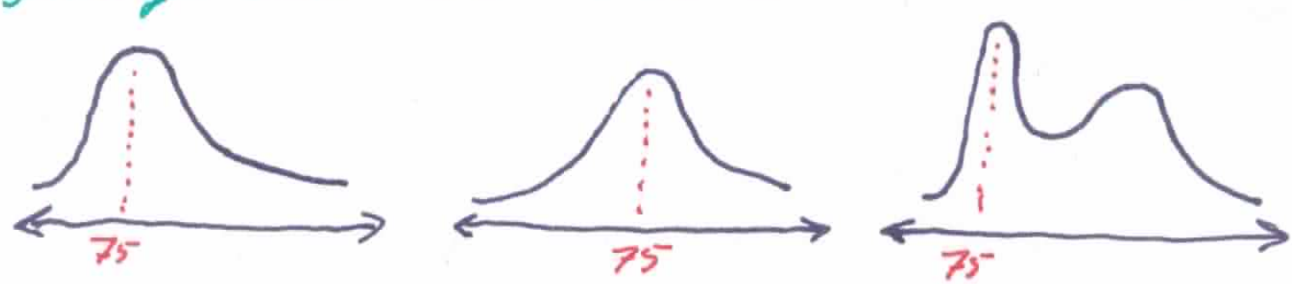
Lecture 2: Measures of Central Tendency and Variability

central tendency (averages)

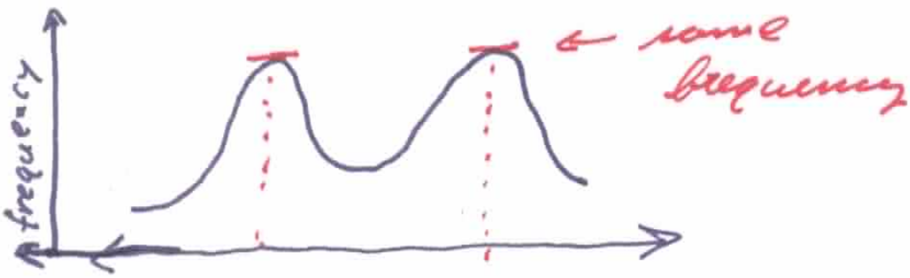
- mean
- median
- mode ← good for categorical data

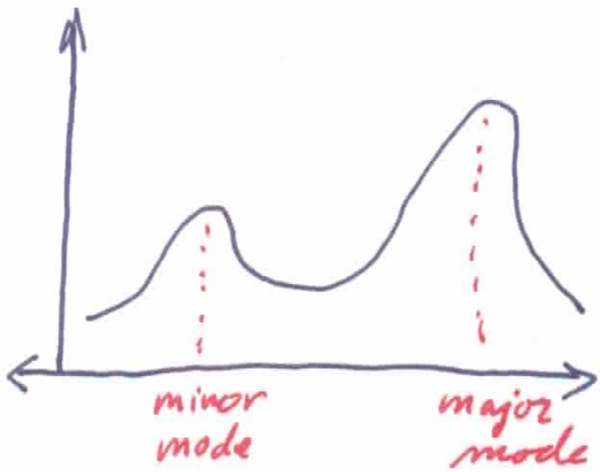
Mode

- ask how many students are freshman, sophomores, juniors & seniors
- the mode is the most frequently occurring **category** (nominal-level measure)
- problem w/ using mode in numerical data is it does not incorporate enough information about the distribution



- but when working with nominal-level measures (e.g. freshman, sophomore, etc.), the mode is the only choice *bimodal distribution





median

- the value in the middle
- ~~[illegible]~~ cannot be used w/ nominal-level measurements, but can be used w/ order-level measure
- the median depends on the order among values (from high to low or vice versa)
- when even number, we take the average of the middle two values

- advantage of median is that it is not affected by a few extreme values (outliers)

- median often used when distribution is skewed (e.g. Census Bureau reports median household income)

- another advantage is that it can be computed in the presence of missing values (assuming that you know if they're in upper or lower end)

mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

mean vs. median vs. mode

$X = \{1, 1, 1, 1, 6\}$

median + mode are both 1

but mean is 2

→ mean does not match any of the values

→ mean affected by the outlier

subscripts, summation, notation

$X_i \equiv$ the i-th value of set X

$\overline{X} \equiv$ the mean of X

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \equiv$ the sum of the values of set X divided by the number of observations

Σ is "Sigma" ("sum" begins with an "s")

$i=1$ and N are the lower + upper limits of the summation

choosing a measure of central tendency

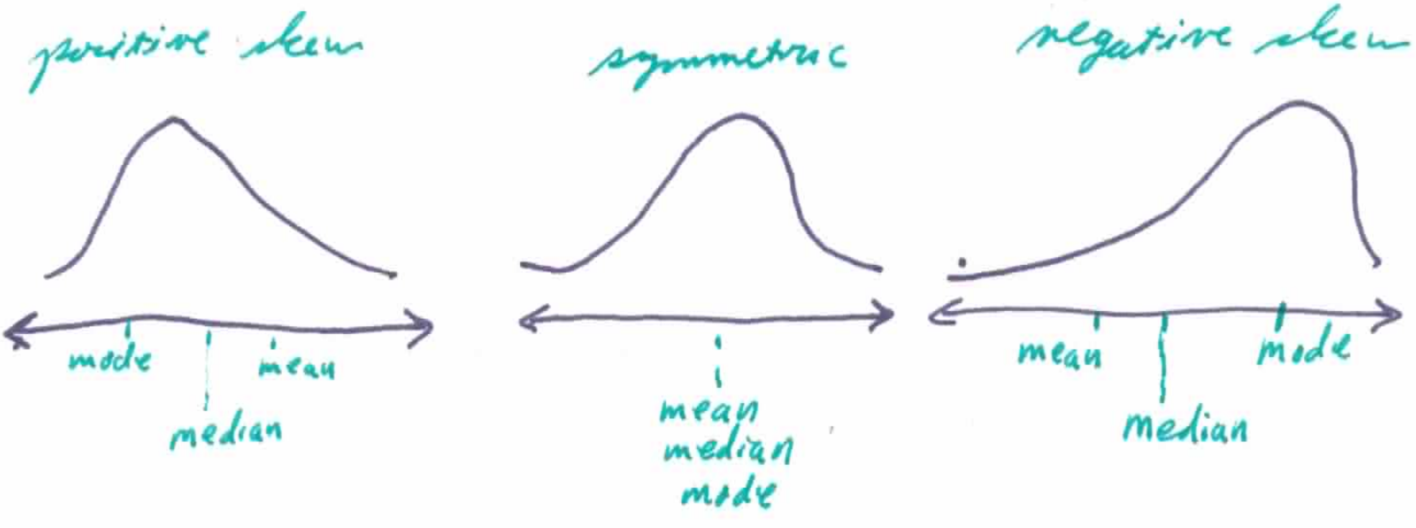
- mode – only choice w/ nominal level data
  (3 Fords, 2 Pontiacs, 4 Toyotas, 1 BMW)
  Toyota is the mode
  - also appropriate when distribution has two or more modes + you want to describe members of each group as typical
  - <u>not</u> appropriate with small samples or ungrouped continuous data

- median
  - only appropriate w/ ordinal-level data or higher
    - CEO, manager, worker
  - also appropriate w/ skewed distributions

- mean
  - requires interval-level data or higher
  - ~~reflects the influence of every~~
  - gives equal weight to each data point
  - **not** appropriate in the presence of extreme outliers

using mean, median + mode to detect skewed distributions

- empirical frequency distributions are noisy (i.e. jagged)



Measures of Variability

→ your textbook has a silly but GOOD example of the limitations of the mean

weights of defensive linemen

team A	team B
215	145
180	165
200	190
205	300

→ the mean of both is 200, but [illegible] ~~that~~ 300 lb lineman is not going to matter, all running plays ~~we~~ will go through the 145 & 165 lb linemen

→ the variability (i.e. dispersion) of the weights is important

→ variability measured by:

- range
- interquartile range
- mean deviation
- variance (and standard deviation)

Range

→ largest value minus smallest value

→ football example

A: 215 − 180 = 35
B: 300 − 145 = 155

→ weakness: based only on max & min so extreme values affect the measure

interquartile range

→ the value at the 75th percentile minus the value at the 25th percentile

→ similar in spirit to the concept of the median

→ example ~~[illegible]~~ using income distribution

1967	75th percentile	\$12,150
	50th percentile	\$8,000 (median)
	25th percentile	\$4,900

interquartile range = \$12,150 − \$4,900 = \$7,250

→ interquartile range is very useful with skewed distributions where a few extreme values may distort the std. deviation

→ it also gives us a good idea of what a "typical" value is

mean absolute deviation

→ unlike the range or interquartile range mean absolute deviation (and variance) ~~[illegible]~~ ~~[illegible]~~ incorporates information from every data point

→ mean absolute deviation $= \frac{1}{N} \sum_{i=1}^{N} |X_i - \bar{X}|$

→ team A: $\{215, 180, 200, 205\}$

$$\frac{1}{4}\left(|215-200| + |180-200| + |200-200| + |205-200|\right)$$

$$\frac{1}{4}\left(|15| + |-20| + |0| + |5|\right)$$

$$\frac{1}{4}\left(15 + 20 + 0 + 5\right) = \frac{40}{4} = 10$$

Variance

note: that's $(N-1)$

$$\text{sample variance} = \frac{1}{(N-1)} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

→ squared deviation around the mean

→ unlike mean ~~[illegible]~~ absolute deviation it's not on the same "scale" as the original data, so we usually take the square root of variance to obtain the ~~[illegible]~~ standard deviation

choosing a measure of variability

→ level of measurement

- with nominal-level measures only the mode can be ~~[illegible]~~ computed

→ which measure of central tendency are you using? If using median, then you should use ~~mean absolute deviation~~ the interquartile range (because both based on percentiles) If using mean, then use mean absolute deviation or variance/ std error

→ with highly skewed data (where extreme values can create a misleading picture), you may want to use median + interquartile range

Your textbook presents you with "alternative ways to compute the variance and std deviation" ← IGNORE THEM

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{1}{n}\sum x_i^2 - \frac{2}{n}\bar{x}\sum x_i + \bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2$$

$$= \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

So an equivalent expression appears to present a simpler way to calculate the variance, but it ignores the precision with which a computer stores a number, so you should NOT use the "computing formula" EVER!

EXAMPLE

$X = \{45, 50, 55\}$

$\bar{X} = 50$

$$(45-50)^2 = (-5)^2 = 25$$
$$(50-50)^2 = 0^2 = 0$$
$$(55-50)^2 = 5^2 = 25$$
$$\overline{50}$$

$$s^2 = \frac{50}{3}$$

computing formula also works

$$45^2 = 2025$$
$$50^2 = 2500$$
$$55^2 = 3025$$
$$\overline{7550}$$

$$\frac{7550}{3} - 2500 = \frac{50}{3}$$

but watch what happens when we add 10,000 to each of the numbers
(the dispersion about the mean has remained constant, so the variance should remain the same but it doesn't!)

$$10{,}045^2 = 1{,}0090203 \times 10^8$$
$$10{,}050^2 = 1{,}0100250 \times 10^8$$
$$10{,}055^2 = 1{,}0110303 \times 10^8$$
$$\overline{3{,}0300755 \times 10^8}$$

$$\frac{3{,}0300755 \times 10^8}{3} - 1{,}0100250 \times 10^8$$

$$1{,}0100252 \times 10^8 - 1{,}0100250 \times 10^8$$

$$= 17 \neq \frac{50}{3}$$

Standardized Values (z-score)

quantifies proximity to the mean in multiples of the standard deviation

$$z_i = \frac{x_i - \bar{x}}{s}$$

Textbook example:

	mid	final
John	90	65
Mary	65	90

	mid	final
mean	70	70
sd	20	5

so their raw total scores are the same, but Mary did much better than her classmates on the final exam

in z-scores

	mid	final
John	+1,0	−1,0
Mary	−0,25	+4,0

advantages of z-scores

- standardizes the difference in magnitudes of possible scores
- enables you to compare scores across variables