FREQUENCIES, CENTRAL TENDENCY, DISPERSION/VARIABILITY, GRAPHING DISTRIBUTIONS

FREQUENCY DISTRIBUTIONS

FREQUENCY DISTRIBUTIONS

A table reporting the number of observations falling into each category of the variable

FREQUENCY DISTRIBUTIONS

Nominal Variables

Identity	Frequency (f)
Native American	947,500
Native American of multiple ancestry	269,700
Native American of Indian descent	5,537,600
Total (N)	6,754,800

FREQUENCY DISTRIBUTIONS

Ordinal Variables

Happiness	Tallies	Freq. (f)	Percentage
Very Happy	IIIIIIIII	9	22.5
Pretty Happy	IIIIIIIIIIIIIIIIIIIIIIIII	25	62.5
Not too Happy	IIIIII	6	15.0
Total (N)		**40**	**100.0**

FREQUENCY DISTRIBUTIONS

Interval-Ratio Variables

Class Interval	*f*	%
95–99	3	4.23
90–94	2	2.82
85–89	4	5.63
80–84	7	9.86
75–79	12	16.90
70–74	17	23.94
65–69	12	16.90
60–64	5	7.04
55–59	5	7.04
50–54	4	5.63
Total	71	100 [a]

FREQUENCY DISTRIBUTIONS: PERCENTAGES & PROPORTIONS

Race/Eth	Frequency (*f*)	Proportion (*P*)	Percentage (%)
Black	1	1/27=.037	3.7%
Native Am.	1	.037	3.7%
Latinx	9	.333	33.3%
As. Am.	4	.148	14.8%
White	12	.444	44.4%
Total (N)	27	1.0	100%

$$\% = \frac{f}{N}(100) = P(100)$$

A relative frequency obtained by dividing the frequency in each category by the total number of cases and multiplying by 100

- *A proportion multiplied by 100*

GRAPHING DISTRIBUTIONS

GRAPHIC PRESENTATION

Pie Charts

Bar Graphs

Histograms

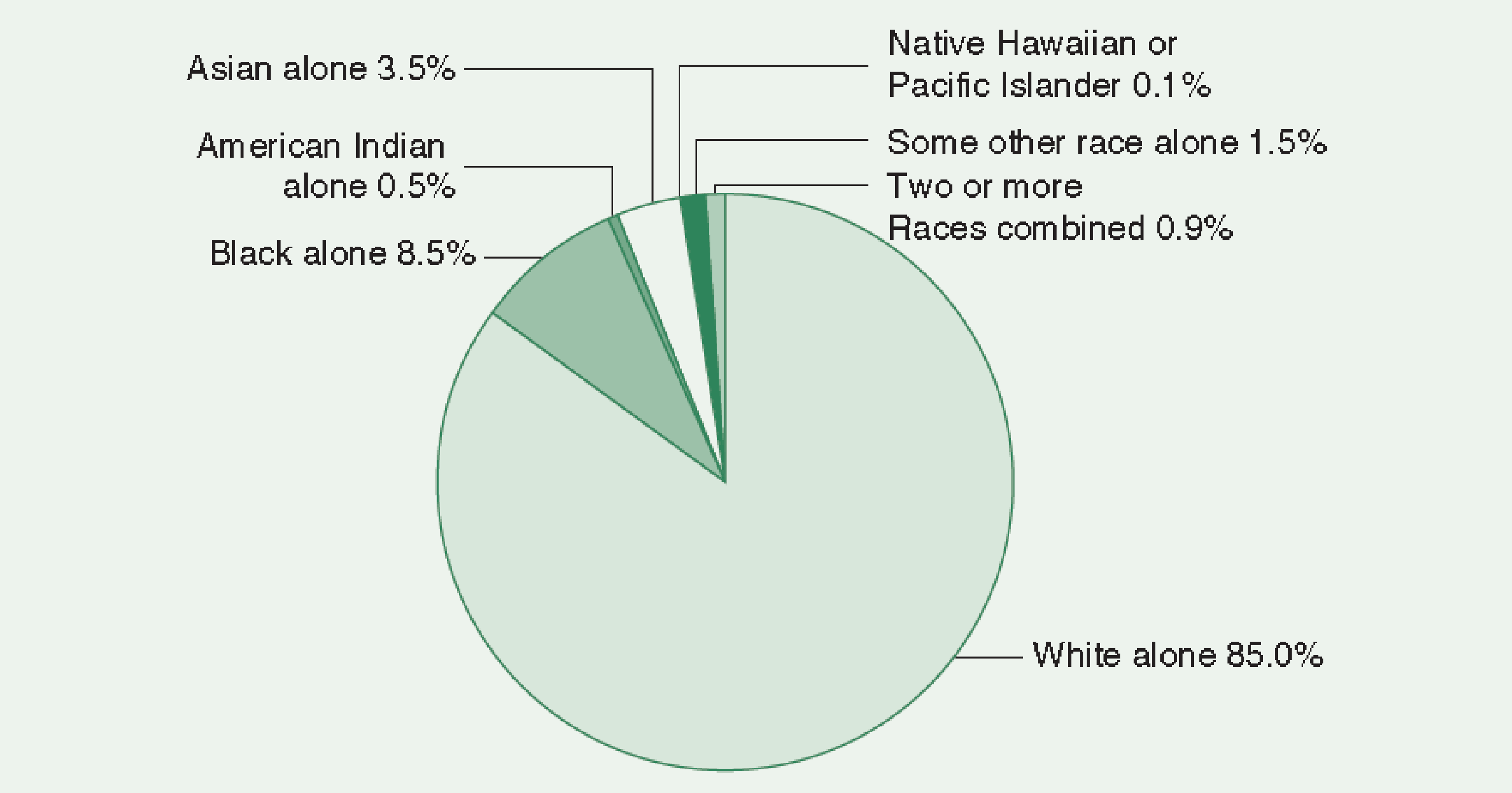
PIE CHART

A graph showing the differences in percentages among categories of a nominal (or ordinal) variable

- Categories are displayed as segments of a circle whose pieces add up to 100% of the total

PIE CHART

Figure 3.1 Three-Year Estimates of the U.S. Population 65 Years and Over by Race, 2009–2011



Source: U.S. Census Bureau, American Fact Finder, 2011, Table S01013

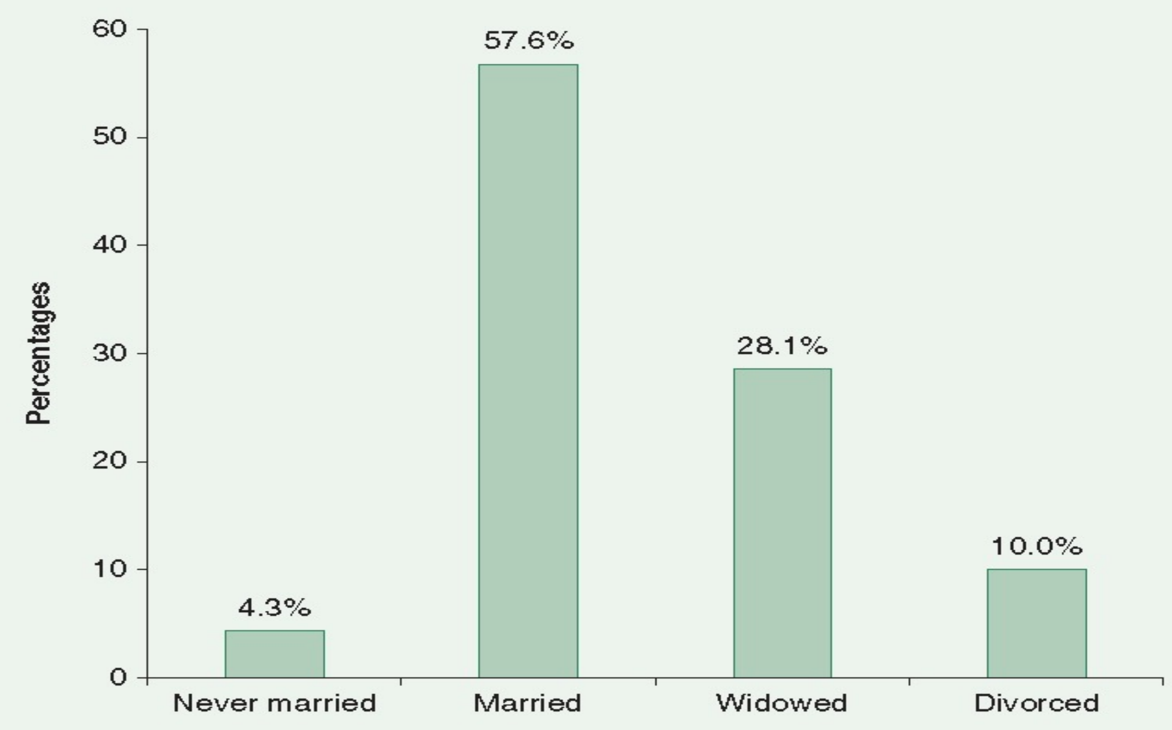
BAR GRAPH

A graph showing the differences in frequencies or percentages among categories of a nominal or ordinal variable

- Categories are displayed as rectangles of equal width with their height proportional to the frequency or percentage of the category
- The taller the bar, the greater the frequency or percentage

BAR GRAPH

Figure 3.3 Marital Status of U.S. Elderly (65 and older), Percentages, 2010



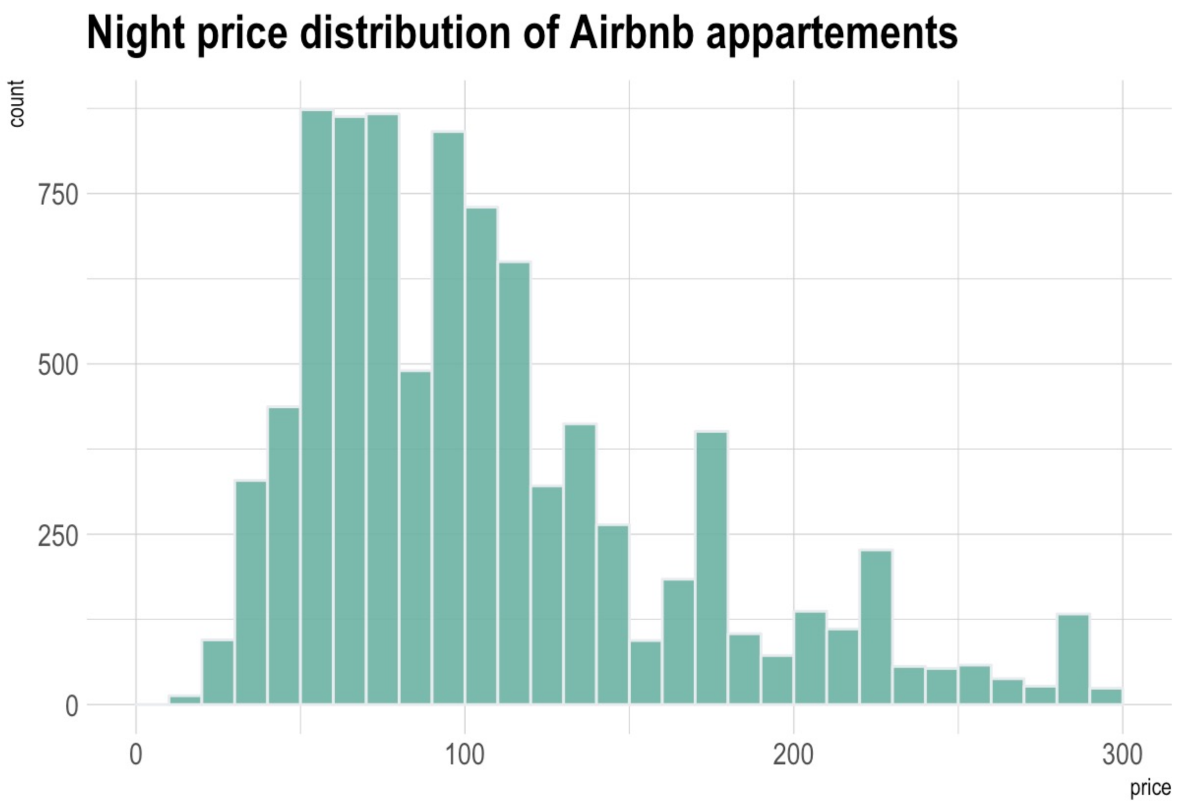
Source: U.S. Census Bureau, *2012 Statistical Abstract of the United States*, Table 34.

HISTOGRAM (SPECIAL BAR GRAPH)

A bar graph showing the differences in frequencies or percentages among categories of an interval-ratio variable

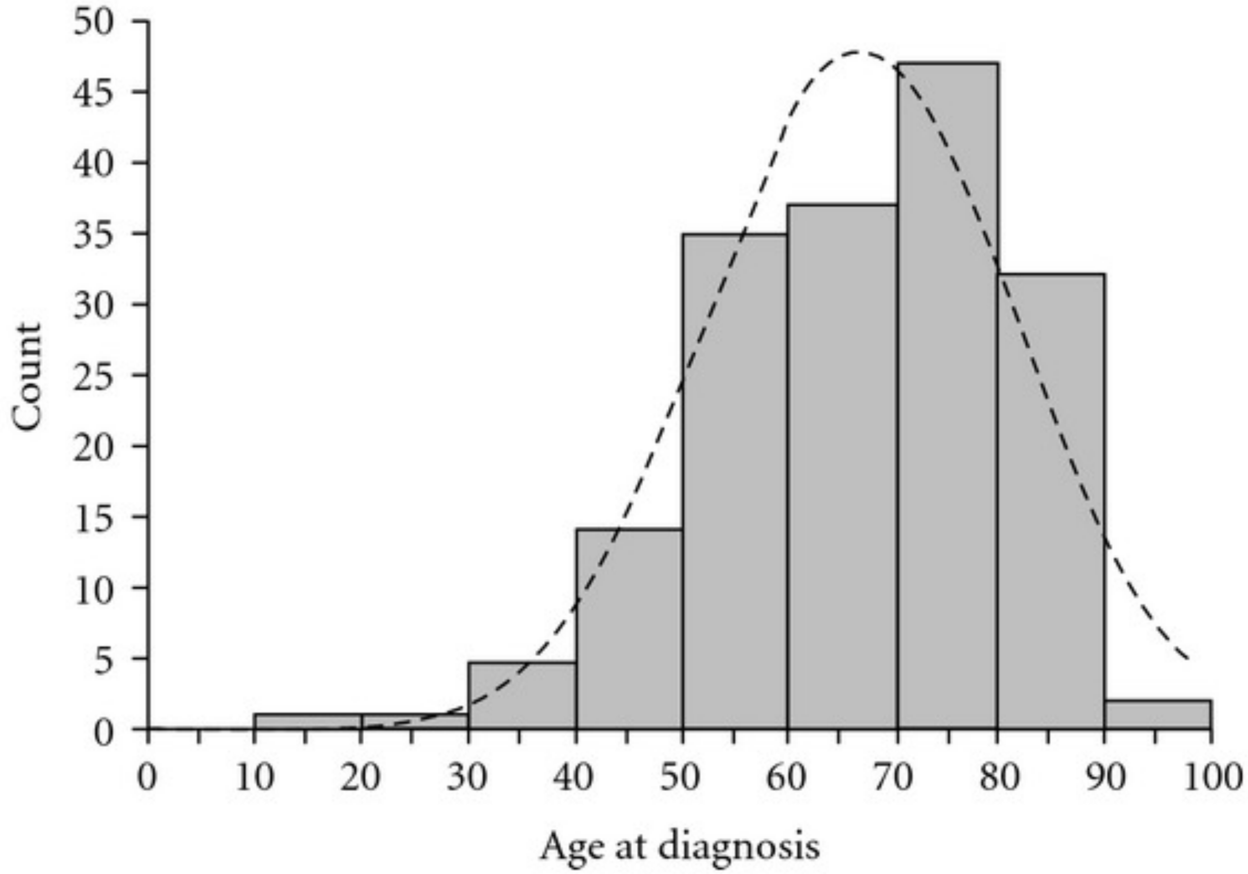
- Categories displayed as contiguous bars, with width proportional to the width of the category and height proportional to the frequency or percentage of that category

HISTOGRAM



Night price distribution of Airbnb appartements
count
750
500
250
0
0
100
200
300
price

HISTOGRAM (W/ NORMAL CURVE)



MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY

Measures that describe what is “average” or “typical” of the distribution of numbers

MEASURES OF CENTRAL TENDENCY

Mode

Median

Mean

THE MODE

The category with the largest frequency (percentage) in the distribution

Can be calculated for following variable levels:

- Nominal
- Ordinal
- Interval-ratio

THE MEDIAN

The score that divides the distribution into two equal parts, so that half the cases are above it and half below it (the 50^{th} percentile)

Can be calculated for following variable levels:

- Ordinal
- Interval-ratio

To do:

- Put data in ascending or descending order
- Position of median is : $\frac{N+1}{2}$

THE MEAN

Distribution's average

Obtained by adding up all the scores and dividing by the total number of scores

Center of gravity for distribution

- So, sensitive to extremes

Can be calculated for following variable levels:

- Interval-ratio

THE MEAN

Calculating the mean

$$\bar{X} = \frac{\sum X}{N}$$

$\bar{X}$ = mean of X

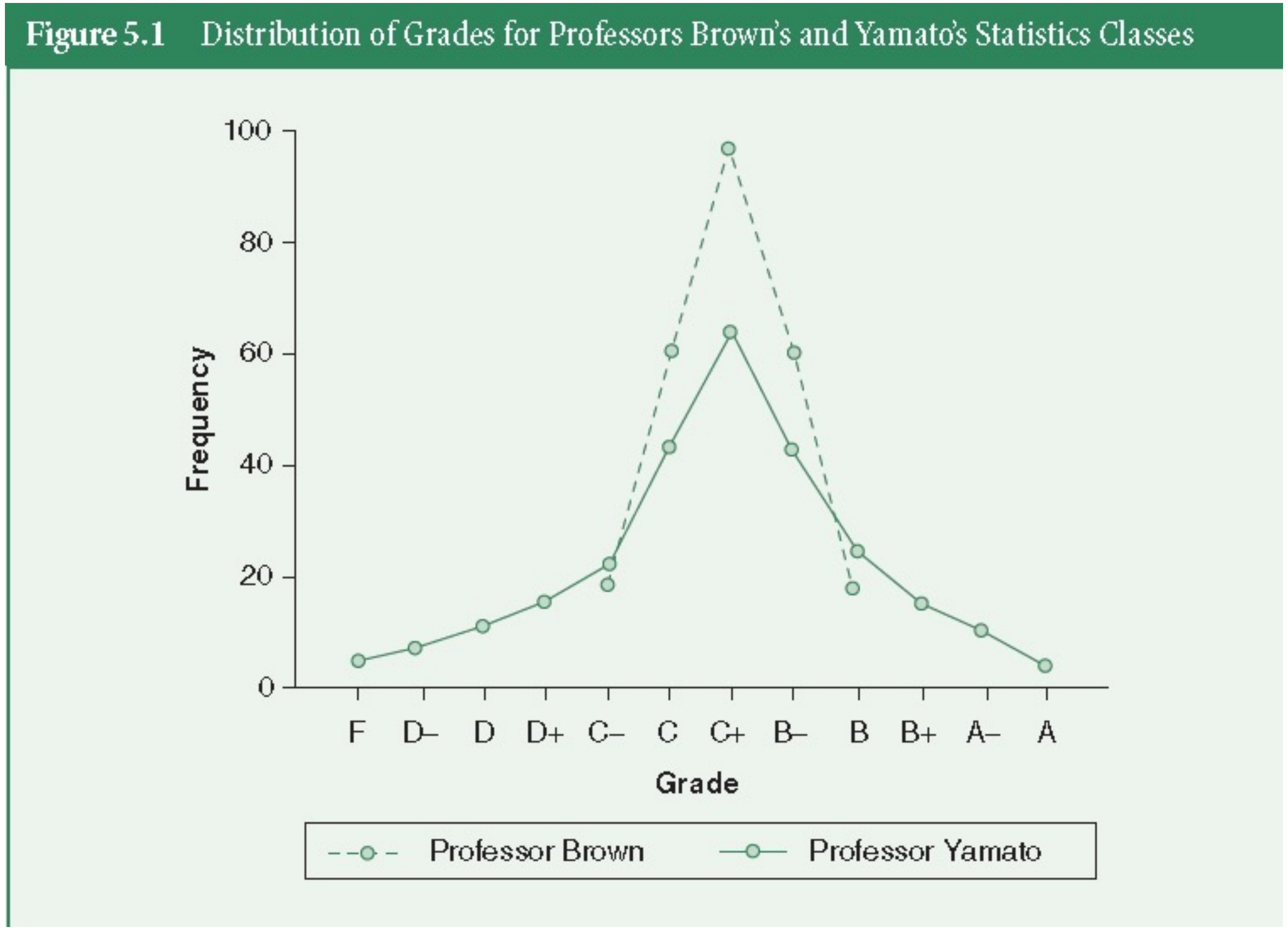
X = the raw scores of the variable

ΣX = the sum of all the X (raw) scores

N = the number of observations or cases

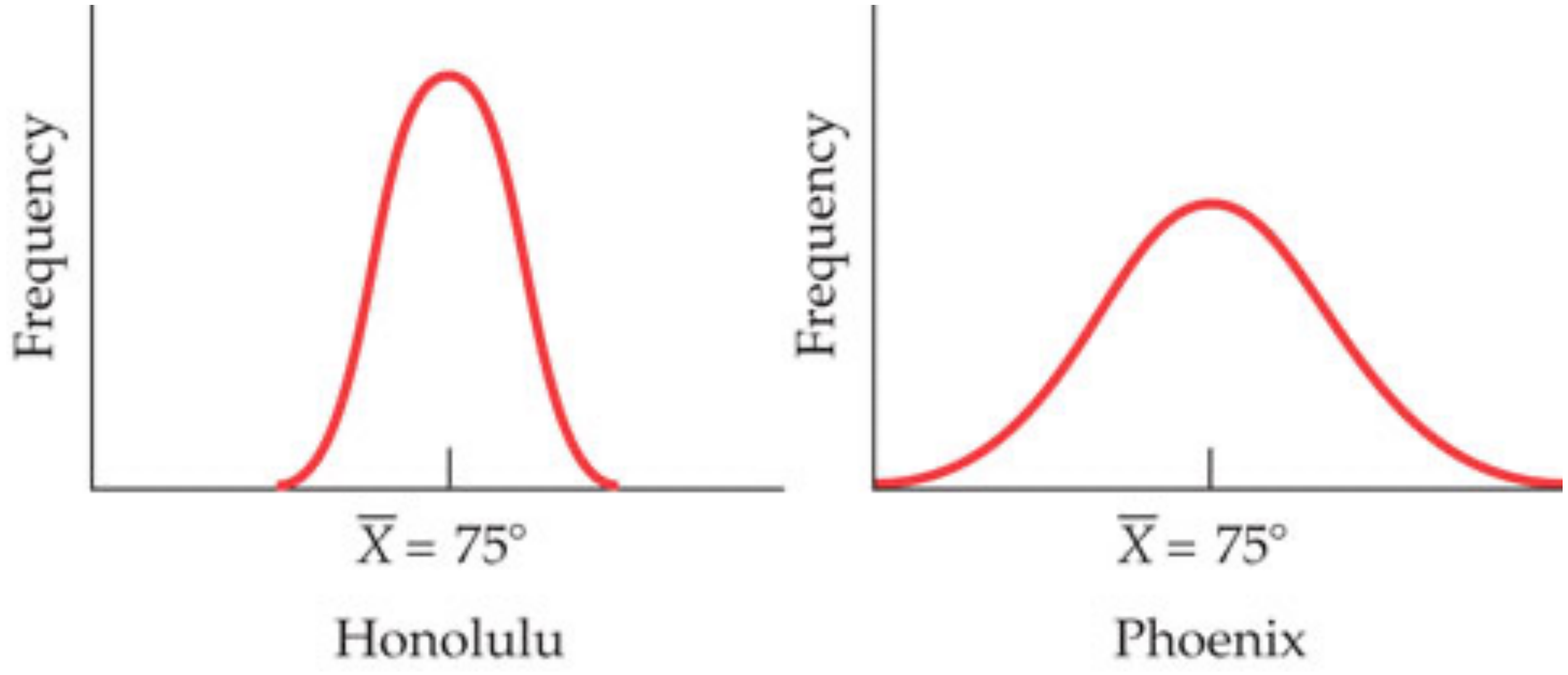
MEASURES OF VARIABILITY

WHY NOT JUST CENTRAL TENDENCY?



Notice that both distributions have the same mean, yet they are shaped differently

WHY NOT JUST CENTRAL TENDENCY?



WHY NOT JUST CENTRAL TENDENCY?

Measures of central tendency can be misleading

Need a measure of the spread/clustering of data

MEASURING VARIABILITY

Once we have the mean (central tendency) we want to know how much variation is around the mean

- Are scores closely clustered around the mean or spread out?

Numbers that describe diversity or variability in the distribution

MEASURES OF VARIABILITY

Range

Variance

Standard Deviation

THE RANGE

A measure of variation in interval-ratio variables.

It is the difference between the highest (maximum) and the lowest (minimum) scores in the distribution

- Range = Highest score – Lowest score

PROBLEMS AND FIXES

Range sensitive to extremes… looks at

Instead, want measure of clustering around the mean

RAW SCORE DEVIATION

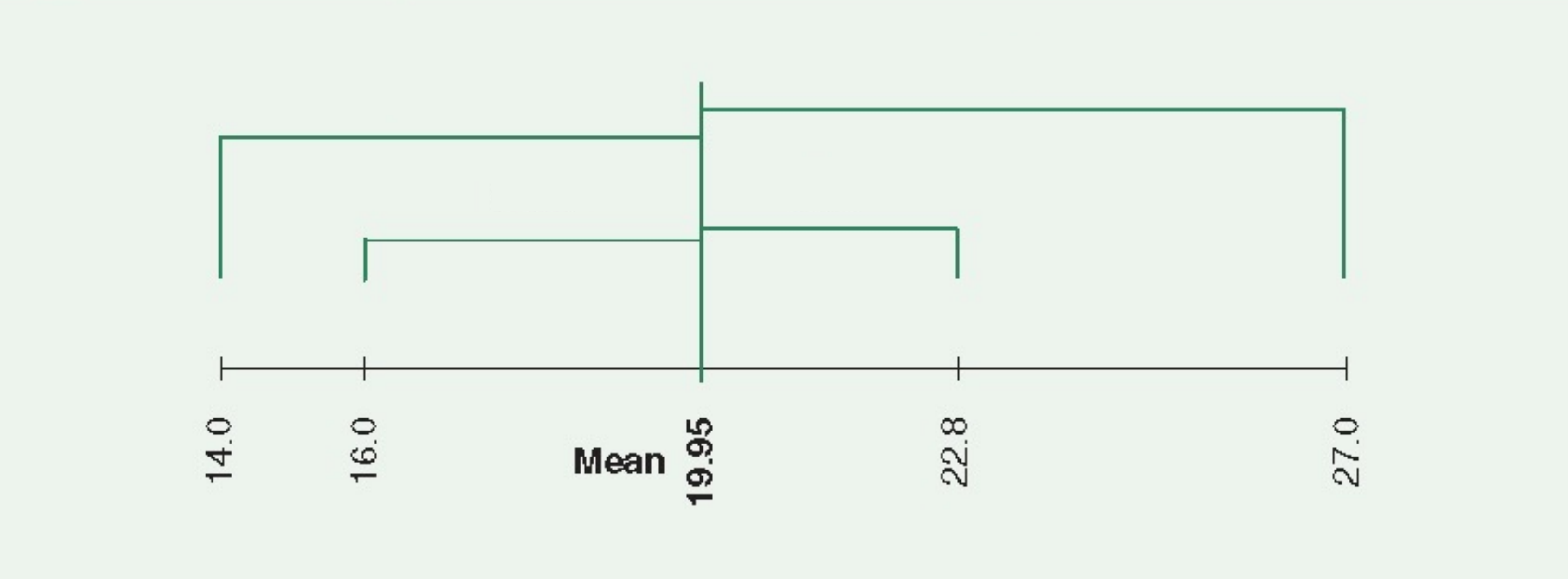
A value for how **close** or **far** a raw score is from the mean

- Each raw score has an associated deviation

- Because mean ($\bar{X}$) is the "center of gravity" of all raw scores (X), each raw score will be some distance from the mean

RAW SCORE DEVIATION

Figure 5.6 Illustrating Deviations From the Mean



RAW SCORE DEVIATION

Subtract each raw score from the mean

- Each raw score (observation) will have a deviation score

Raw score deviations can be negative or positive

- Scores below mean have negative deviation and scores above mean have positive deviation

Shows how close or far they are from the mean (how much each score differs)

RAW SCORE DEVIATION

Calculating the raw score deviation

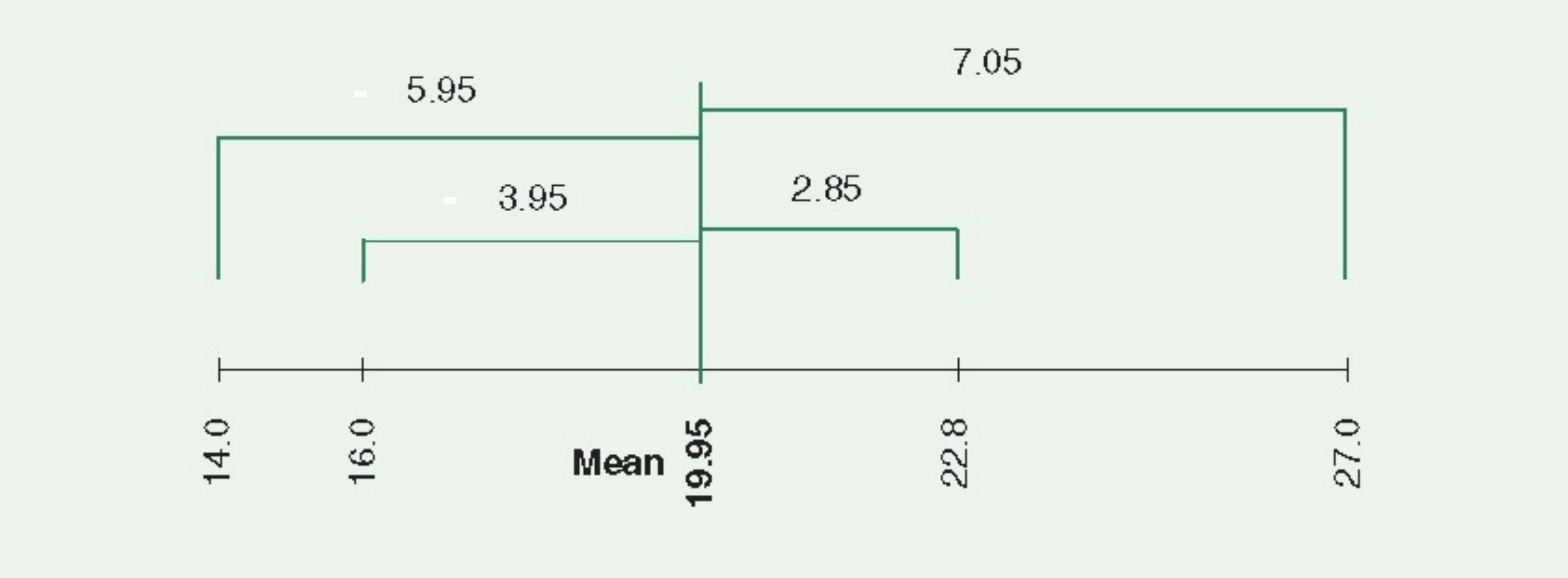
$$Raw\ Dev = (X - \bar{X})$$

$\bar{X}$ = mean of X

X = the raw scores of the variable

RAW SCORE DEVIATION

Figure 5.6 Illustrating Deviations From the Mean



RAW SCORE DEVIATION

Instead of reporting deviation for <u>every</u> raw score, we want to come up with <u>**one**</u> overall <u>summary score</u>

We can do this by <u>summing</u> (adding up) <u>every raw score deviation</u> scores

RAW SCORE DEVIATION

Summary Score: How much *all scores* vary from the mean?

$$\sum(X - \bar{X})$$

RAW SCORE DEVIATION

Cannot sum all the raw score deviations…

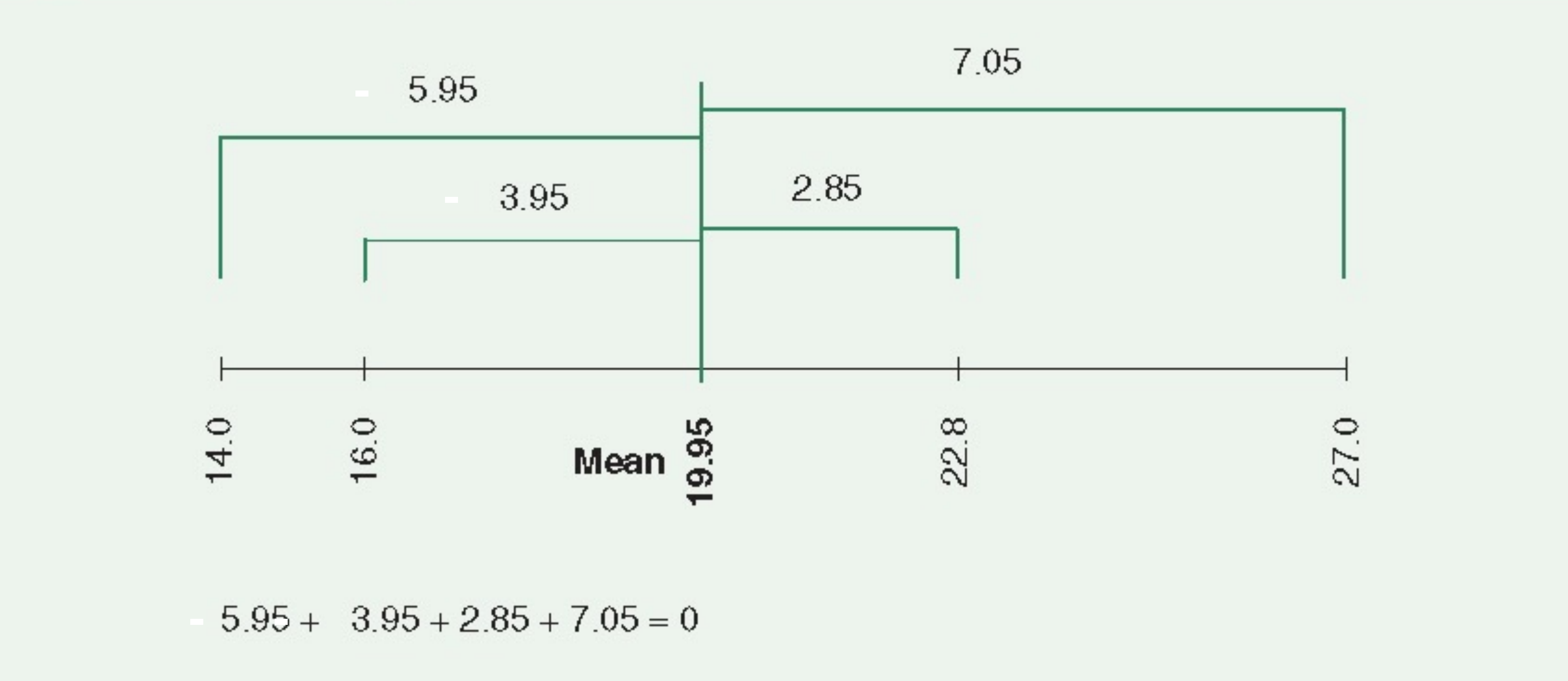
- *Negative* and *positive* **cancel** each other out and **sum to zero**

Because mean is ”center of gravity” …

$$\sum(X - \bar{X}) = 0$$

RAW SCORE DEVIATION

Figure 5.6 Illustrating Deviations From the Mean



FIXING RAW SCORE DEVIATION

Instead, we can square the sum of the deviations to make them all positive

Sum of the Squared Deviations from the Mean (Sum of Squares)

$$\sum(X - \bar{X})^2$$

FIXING RAW SCORE DEVIATION

We create an average-summed-deviation by dividing by the number of cases involved

- (*number of cases minus 1, to correct for error*)

Variance

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

VARIANCE

A measure of variation for interval-ratio variables

- How much *all scores* vary from the mean, on average?
- *how closely* the scores cluster around the mean, on average

Average of the squared-deviations from the mean

FIXING VARIANCE

In “squared units”

- We square the deviations because, when we add deviations, they sum to zero
- When we square numbers, they become larger (in absolute value) and positive, so you won’t have zeros
- However, squared numbers aren’t in the original units as the variable X

FIXING VARIANCE

Must take the square root of variance to get it back to the original units (not squared units)

Standard Deviation

$$SD = \sqrt{var} = \sqrt{\frac{\sum(X - \bar{X})^2}{(N-1)}}$$

STANDARD DEVIATION

A measure of variation for interval-ratio variables

Average overall deviation from the mean

- Equal to the square root of the variance

VARIABILITY AND DISTRIBUTIONS

STANDARD DEVIATION IN RELATION TO DISTRIBUTIONS

Real measure for average variability in a distribution

Larger scores indicate more variability/spread around the mean

Can be used to interpret position of a raw score, relative to all other scores within a distribution

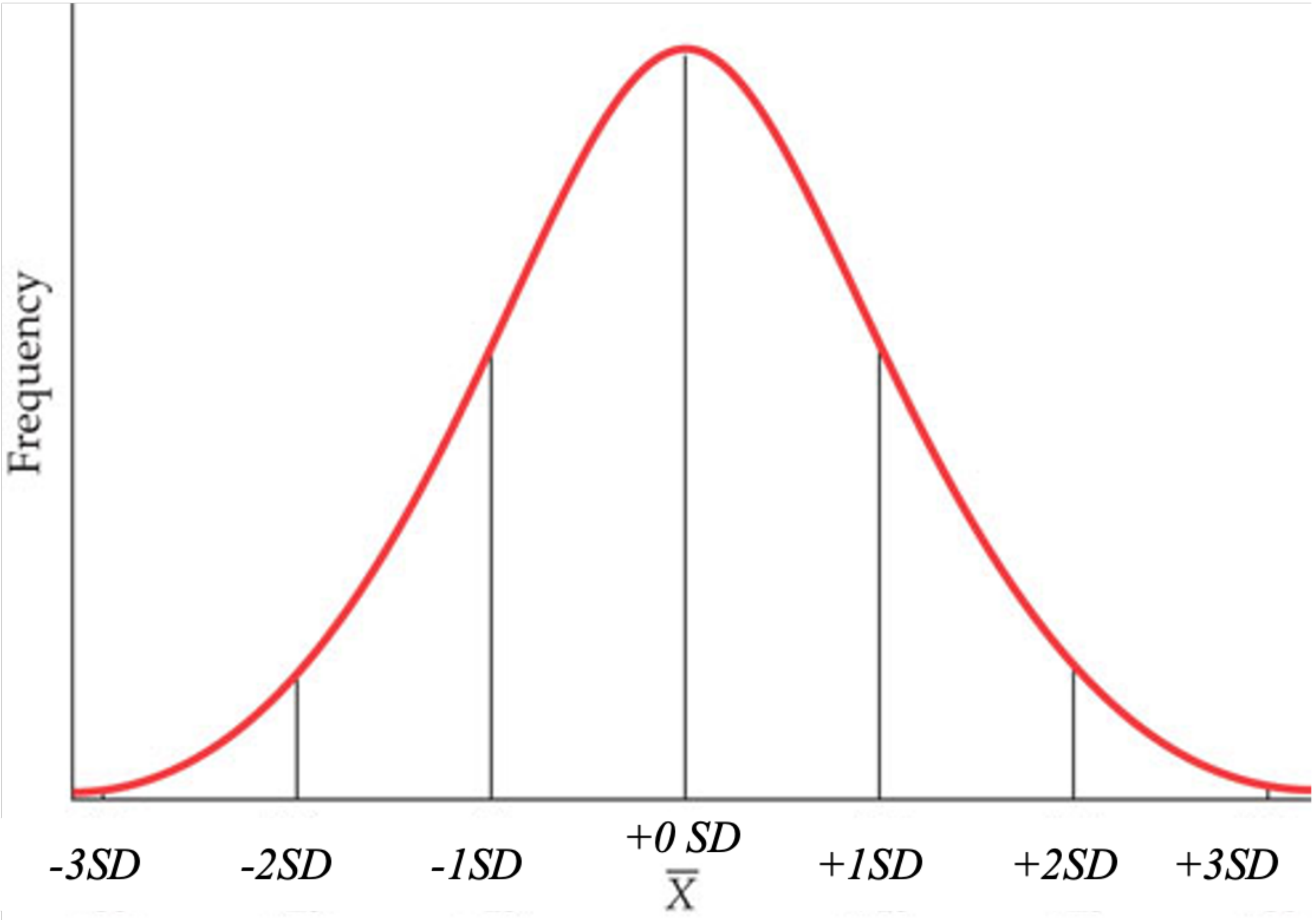
Standardized (e.g. SD), so can be used to compare variability in one distribution to variability in other distributions

STANDARD DEVIATION IN RELATION TO DISTRIBUTIONS

Can be used to divide up a normal curve (AKA normal distribution)

- About 99.9% of cases in a distribution fall between -3 SDs and +3 SDs of the mean
- Just over 2/3 of cases in a distribution fall between -1 SDs and +1 SDs of the mean

STANDARD DEVIATION IN RELATION TO DISTRIBUTIONS



A NOTE ON THE SHAPE OF THE DISTRIBUTION

Skewness (symmetry)

- measure of how close or far a distribution is from symmetry (the normal curve). It measures the clustering of scores along the X-axis, with regard to the position of the mode, median, and mean.

Kurtosis (peakedness)

- measure of the peakedness of the distribution - how high or low the distribution is on the Y-axis, and how different it is from mesokurtic (e.g. middle kurtosis, or the normal curve).

SKEWNESS AND KURTOSIS

Skewness (symmetry)

- **Symmetrical** (e.g. Normal Curve)
  - Mean, median, mode all at center
- **Negatively Skewed** (e.g. Final Grades)
  - left skew, tail to the left
  - mean < median < mode
- **Positively Skewed** (e.g. STIs)
  - right skew, tail to the right
  - mode < median < mean

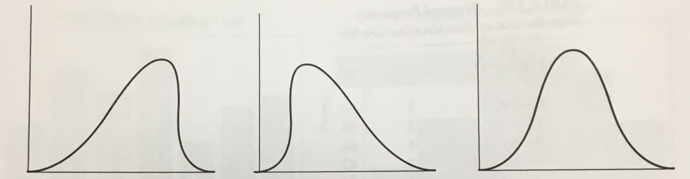
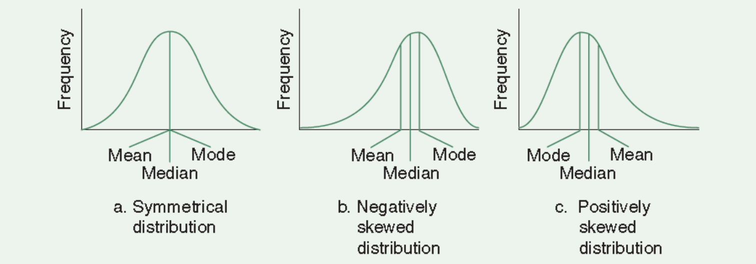


FIGURE 2.7 *Three Distributions Representing Direction of Skewness*

Figure 4.6 Types of Frequency Distributions



Kurtosis (peakedness)

- **Mesokurtic**
  - Normal Curve
- **Leptokurtic**
  - Very peaked
- **Platykurtic**
  - Very flat

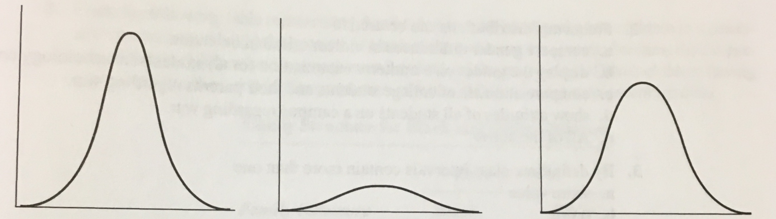


FIGURE 2.8 *Some Variation in Kurtosis among Symmetrical Distributions*

SKEWNESS AND KURTOSIS (RANGES AND CUTOFFS)

Skewness (symmetry)

- Range: $-\infty$ to ∞.
- Signs: $-$ is negative skewness, $+$ is positive skewness, and 0 is no skew… (AKA symmetry, AKA the normal curve).
- Cutoffs:
  - High Skewness: $x \geq |1|$
  - Moderate Skewness: $|1| \geq x \geq |.5|$
  - Low Skewness: $|.5| \geq x \geq |0|$

Kurtosis (peakedness)

- Range: -2 to $+\infty$.
- Signs: $-$ is platykurtic, $+$ is leptokurtic ($+$ but close to 0 is mesokurtic)
- Cutoffs:
  - Platykurtic: $0 \geq x \geq -2$
  - Mesokurtic: $1 \geq x \geq 0$
  - Leptokurtic: $+\infty \geq x \geq 1$