Module 3:
What do you mean?

MANG 434

Spring 2020

Agenda for Module 3

- 3/2/2020 – 3/4/2020
  - Summarizing data (frequency distributions); fitting data (central tendency and shape); interpretation and communication; issues in datasets

Agenda for Module 3

- 3/2/2020 – 3/4/2020
  - Summarizing data (frequency distributions); fitting data (central tendency and shape); interpretation and communication; issues in datasets

- 3/4/2020
  - Exam 1 review session (BE 347 | 5-7pm)

Agenda for Module 3

- 3/2/2020 – 3/4/2020
  - Summarizing data (frequency distributions); fitting data (central tendency and shape); interpretation and communication; issues in datasets

- 3/4/2020
  - Exam 1 review session (BE 347 | 5-7pm)

- 3/6/2020
  - No class – Meet with Jamie day (need to meet with me this week!)
  - Exam 1 posted on eCampus

Agenda for Module 3

- 3/9/2020
    - Excel skills (IF AND statements)

Agenda for Module 3

- 3/9/2020
    - Excel skills (IF AND statements)

- 3/11/2020
    - Excel skills (IF OR statements)

Agenda for Module 3

- 3/9/2020
  - Excel skills (IF AND statements)
- 3/11/2020
  - Excel skills (IF OR statements)
- 3/13/2020
  - Exam day (Exam 1 is due by 11:59PM ET on this day – no exceptions!)
  - Evidence that group project data collection is (near) completion is also due

Agenda for Module 3

- 3/2/2020 – 3/4/2020
  - Summarizing data (frequency distributions); fitting data (central tendency and shape); interpretation and communication; issues in datasets

- 3/4/2020
  - Exam 1 review session (BE 347 | 5-7pm)

- 3/6/2020
  - No class – Meet with Jamie day (need to meet with me this week!)
  - Exam 1 posted on eCampus

Agenda for Module 3

- Let’s get started! ☺

Summarizing Data

- Frequency distribution

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

			Job	Pay
Rating	Stress	WLB	satisfaction	satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

Summarizing Data

- Frequency distribution
    - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

Rating	Stress	WLB	Job satisfaction	Pay satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

CAN BE VISUALIZED IN A BARPLOT

Summarizing Data

- Frequency distribution
  - A table or graph that shows each possible score along with the number of times that score was observed in the data.

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Table 2. Frequency Distribution

Rating	Stress	WLB	Job satisfaction	Pay satisfaction
10	0	0	0	3
9	0	0	0	0
8	0	2	1	0
7	3	0	3	2
6	2	0	0	2
5	2	1	0	0
4	0	1	2	0
3	0	0	1	0
2	0	3	0	0
1	0	0	0	0
0	0	0	0	0
Count	7	7	7	7

CAN BE VISUALIZED IN A BARPLOT

CAN BE USED TO SUMMARIZE ALL TYPES OF DATA (SEE MODULE 2)

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - “What proportion of the respondents gave a rating of 7 for stress?”

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - “What proportion of the respondents gave a rating of 7 for stress?”

$$\text{Relative frequency} = \frac{frequency\ of\ response}{total\ number\ of\ responses}$$

Summarizing Data

- Relative frequency
  - Compared to the (raw) frequency itself, this is a way to make even better sense of observed data
  - Represents how often a response is observe relative to the total number of responses
    - "What proportion of the respondents gave a rating of 7 for stress?"

$$\text{Relative frequency} = \frac{frequency\ of\ response}{total\ number\ of\ responses}$$

$$= \quad \frac{3}{7} \quad = \quad 43\%$$

Summarizing Data

- Cumulative frequency and cumulative percentage
    - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

Summarizing Data

- Cumulative frequency and cumulative percentage
  - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

$$\text{Cumulative frequency}_n = \text{frequency}_n + \textit{cumulative frequency}_{n-1}$$

Table 3. Frequency Distributions for Stress

		Relative	Cumulative	Cumulative
Rating	Frequency	frequency	frequency	percentage
10	0	0 (0%)	7	1.0 (100%)
9	0	0 (0%)	7	1.0 (100%)
8	0	0 (0%)	7	1.0 (100%)
7	3	.43 (43%)	7	1.0 (100%)
6	2	29 (29%)	4	.58 (58%)
5	2	.29 (29%)	2	.29 (29%)
4	0	0 (0%)	0	0 (0%)
3	0	0 (0%)	0	0 (0%)
2	0	0 (0%)	0	0 (0%)
1	0	0 (0%)	0	0 (0%)
0	0	0 (0%)	0	0 (0%)

Summarizing Data

- Cumulative frequency and cumulative percentage
  - An assessment of the total frequency (percentage) of all categories up to and including the category of interest

Cumulative percentage$_n$ = percentage$_n$ + cumulative percentage$_{n-1}$

Table 3. Frequency Distributions for Stress

Rating	Frequency	Relative frequency	Cumulative frequency	Cumulative percentage
10	0	0 (0%)	7	1.0 (100%)
9	0	0 (0%)	7	1.0 (100%)
8	0	0 (0%)	7	1.0 (100%)
7	3	.43 (43%)	7	1.0 (100%)
6	2	29 (29%)	4	.58 (58%)
5	2	.29 (29%)	2	.29 (29%)
4	0	0 (0%)	0	0 (0%)
3	0	0 (0%)	0	0 (0%)
2	0	0 (0%)	0	0 (0%)
1	0	0 (0%)	0	0 (0%)
0	0	0 (0%)	0	0 (0%)

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Step 1: Calculate column mean (average)

Average job satisfaction rating = $\frac{7+7+7+8+3+4+4}{7}$ = 5.71

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of "high" vs. "low" responses observed in a dataset
  - Example: How many people have "high" and "low" levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of “high” vs. “low” responses observed in a dataset
  - Example: How many people have “high” and “low” levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify “high” (i.e., > 5.71) vs. “low” (i.e., < 5.71) scores

Summarizing Data

- Mean (or median) splits
  - A method used to estimate the number of "high" vs. "low" responses observed in a dataset
  - Example: How many people have "high" and "low" levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify "high" (i.e., > 5.71) vs. "low" (i.e., < 5.71) scores

Step 4: Calculate "high" vs. "low" frequencies and percentages

Summarizing Data

4 out 7 ="high" scores

4/7 = .57 (57%)

3 out 7 ="low" scores

3/7 = .43 (43%)

- Mean (or median) splits
  - A method used to estimate the number of "high" vs. "low" responses observed in a dataset
  - Example: How many people have "high" and "low" levels of job satisfaction?

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
6	2	8	6
5	8	7	9
5	8	7	9
6	2	7	9
7	4	4	7
7	5	4	7
7	2	3	6

Step 1: Calculate column mean (average)

$$\text{Average job satisfaction rating} = \frac{7+7+7+8+3+4+4}{7} = 5.71$$

Step 2: Rearrange observed data (largest → smallest)

Step 3: Identify "high" (i.e., > 5.71) vs. "low" (i.e., < 5.71) scores

Step 4: Calculate "high" vs. "low" frequencies and percentages

Central tendency

- Mean, median, mode

Central tendency

- Mean, median, mode
  - Represents a simple statistical model of the center of the distribution of scores.
  - A hypothetical estimate of the "typical" score

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Calculate column mean (average)

Average job satisfaction rating = $\frac{7+7+7+8+3+4+4}{7}$ = 5.71

Central tendency

- Mean, **median**, mode
  - Represents the middle score of a set of ordered observations
  - When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

Central tendency

- Mean, **median**, mode
  - Represents the middle score of a set of ordered observations
  - When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

→

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Ordered from low-to-high

Central tendency

- Mean, **median**, mode
  - Represents the middle score of a set of ordered observations
  - When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

Ordered from low-to-high

Central tendency

- Mean, **median**, mode
  - Represents the middle score of a set of ordered observations
  - When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

Ordered from low-to-high

Calculate column median (mid-point of distribution)

Median job satisfaction rating = 7

Central tendency

- Mean, **median**, mode
  - Represents the middle score of a set of ordered observations
  - When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

Ordered from low-to-high

Calculate column median (mid-point of distribution)

Median job satisfaction rating = 7

Central tendency

- Mean, median, mode
  - Represents the most frequently occurring score in a set of observations
  - Can be bi-modal or even multi-modal

Table 1. Observed Data

		Job	Pay
Stress	WLB	satisfaction	satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Central tendency

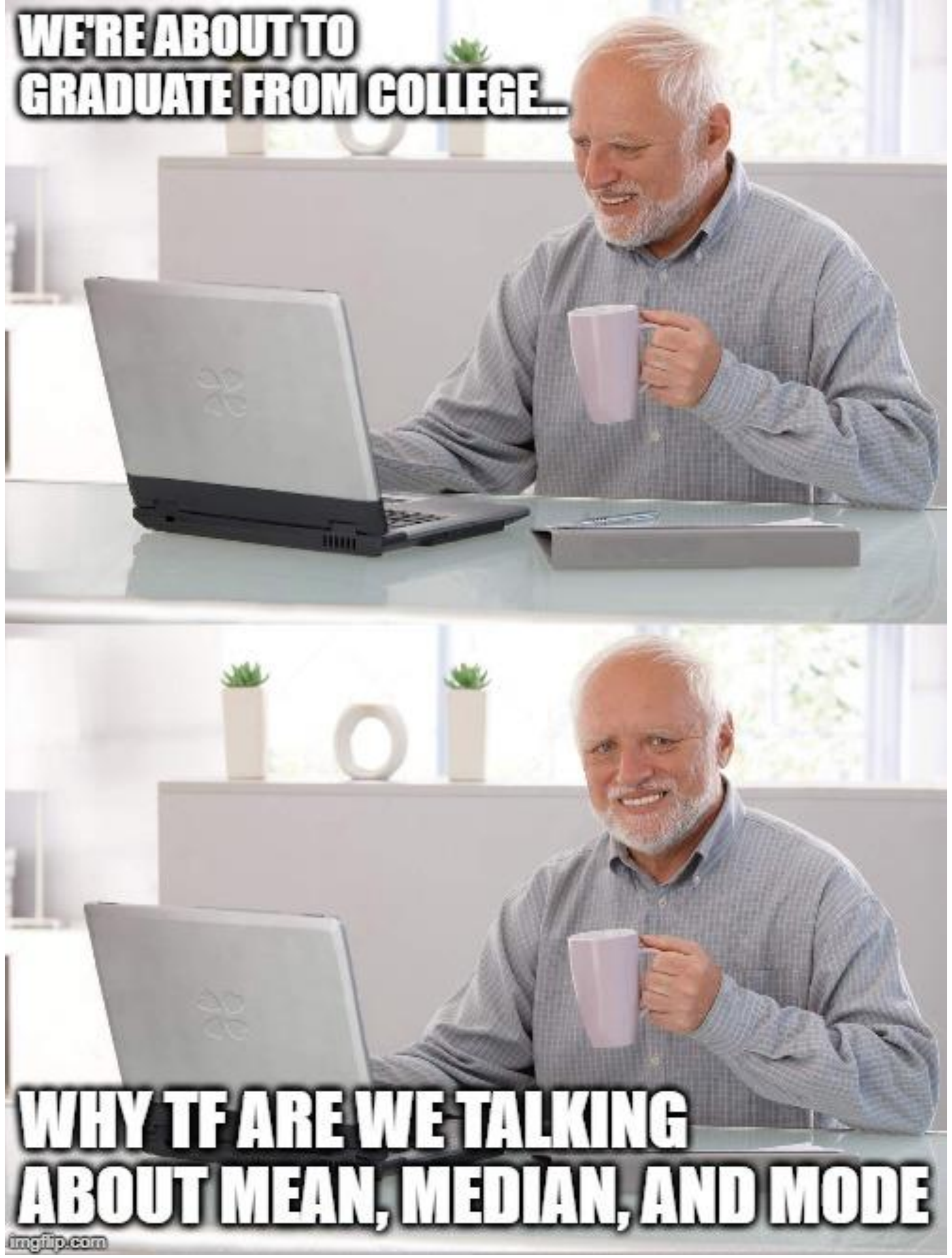
- Mean, median, mode
  - Represents the most frequently occurring score in a set of observations
  - Can be bi-modal or even multi-modal

Table 1. Observed Data

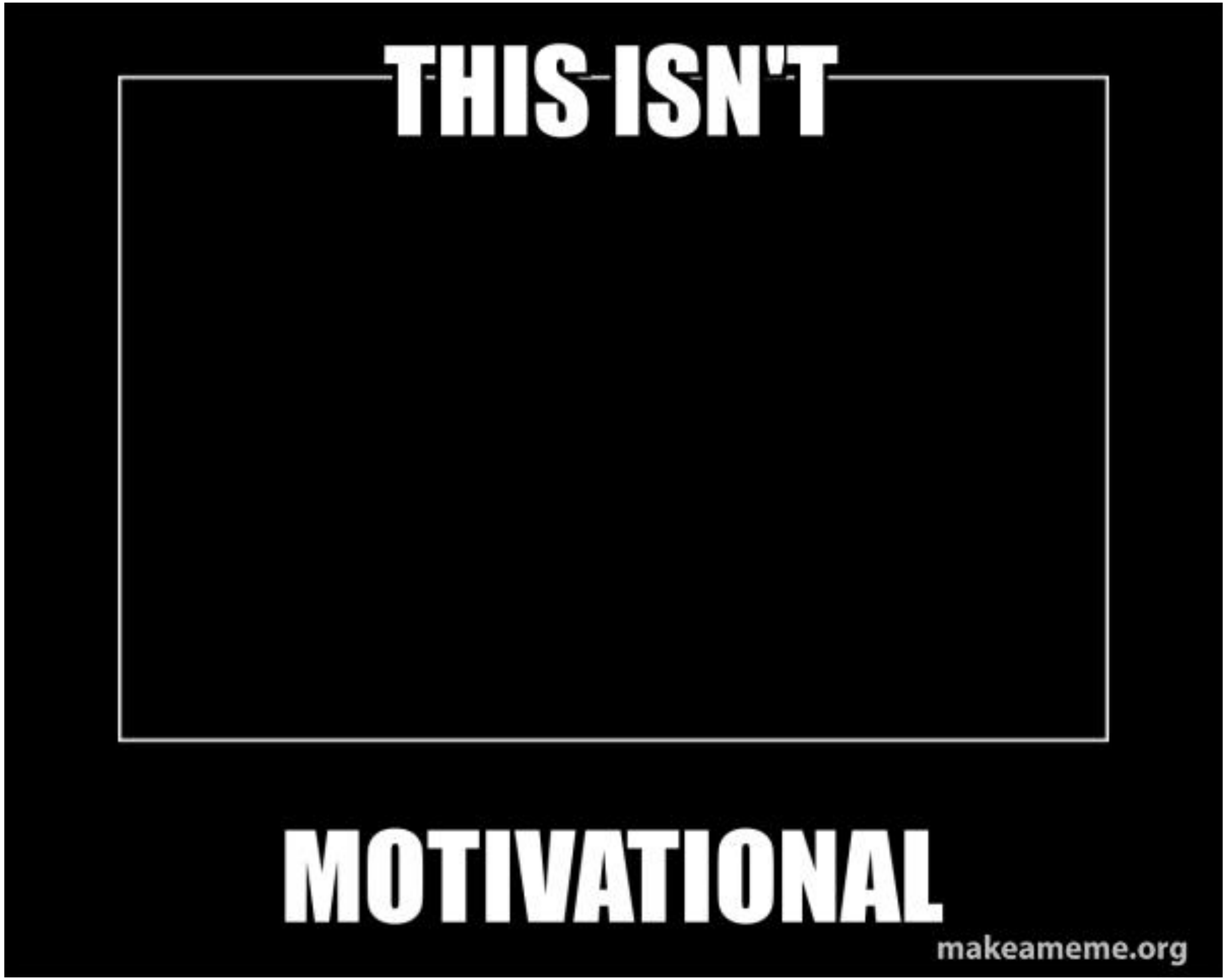
Stress	WLB	Job satisfaction	Pay satisfaction
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6
7	2	3	6
7	4	4	7
7	5	4	7

Calculate column mode

Modal job satisfaction rating = 7


WE'RE ABOUT TO GRADUATE FROM COLLEGE...
WHY TF ARE WE TALKING ABOUT MEAN, MEDIAN, AND MODE
imgflip.com

Because mean, median, and mode are important
CHANGE MY MIND
imgflip.com



THIS ISN'T
MOTIVATIONAL
makeameme.org

ME WAITING ON THE POINT TO BE MADE
makeameme.org

The point is…

- Although we know about these measures of central tendency, we may not be using them to their full potential

- Many of the descriptive statistics that we aware of (e.g., mean) are meaningless if they are not reported in tandem with other important information

- What other important information should accompany the mean…

Variance

- Standard deviation
  - SD is an estimate of the average variability (spread) of a set of observations around the mean
  - Importantly, SD is expressed in the same units of measurement as the raw scores
  - It is the square root of the variance (sqrt[sum of squares/number of values])

Variance

- Range
  - The range of scores is the value of the smallest score subtracted from the highest score

Table 1. Observed Data

Stress	WLB	Job satisfaction	Pay satisfaction
7	2	3	6
7	4	4	7
7	5	4	7
5	8	7	9
5	8	7	9
6	2	7	9
6	2	8	6

$$
\begin{aligned}
\text{Range} &= \mathit{Highest\ score} - \mathit{lowest\ score} \\
&= 8 - 3 \\
&= 5
\end{aligned}
$$

Shape

- Skewness

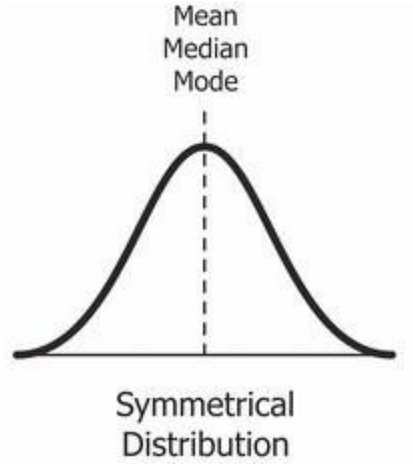
- Kurtosis

Shape

- Skewness → a measure of the symmetry of a *frequency distribution*
- Kurtosis

Shape

- Skewness → a measure of the symmetry of a *frequency distribution*
- Kurtosis

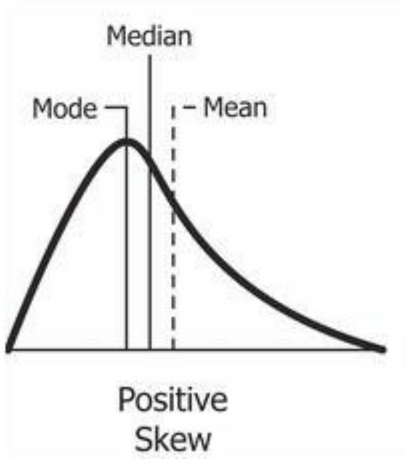


Symmetrical distributions have a skew of 0

Shape

- Skewness → a measure of the symmetry of a *frequency distribution*

- Kurtosis

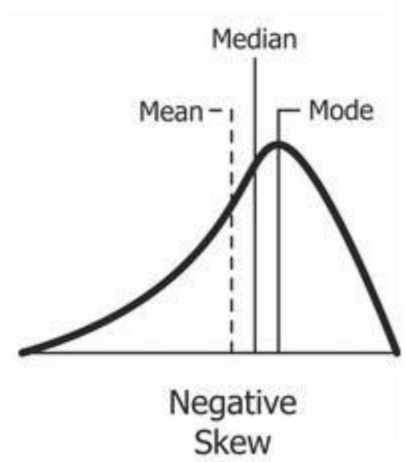


Symmetrical distributions have a skew of 0

When the frequent scores are clustered at the lower end of the distribution and the tail points to the higher (more positive) scores, the value of skew is positive

Shape

- Skewness → a measure of the symmetry of a *frequency distribution*
- Kurtosis



Symmetrical distributions have a skew of 0

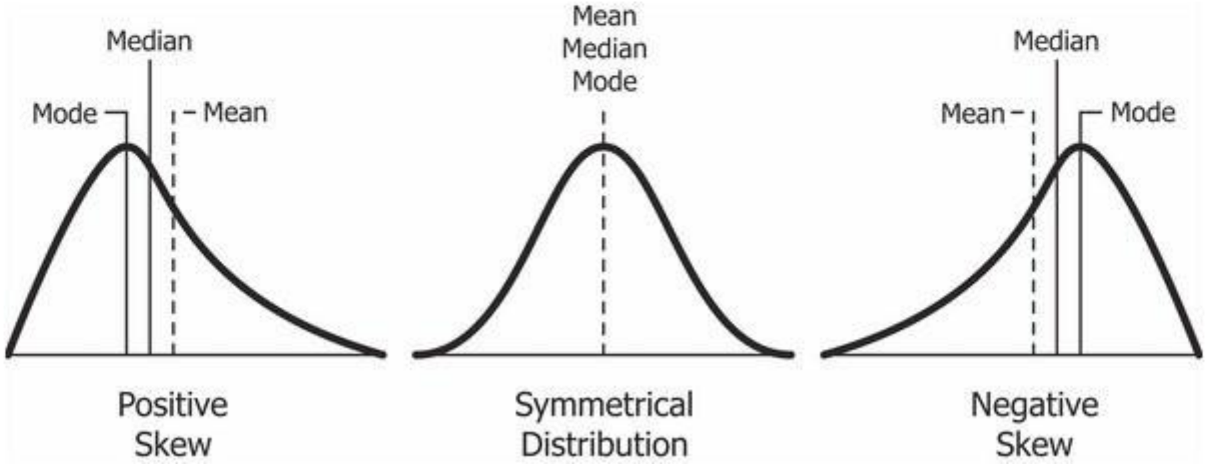
When the frequent scores are clustered at the lower end of the distribution and the tail points to the higher (more positive) scores, the value of skew is positive

When the frequent scores are clustered at the higher end of the distribution and the tail points to the lower (more negative) scores, the value of skew is negative

Shape

- Skewness → a measure of the symmetry of a *frequency distribution*

- Kurtosis



Symmetrical distributions have a skew of 0

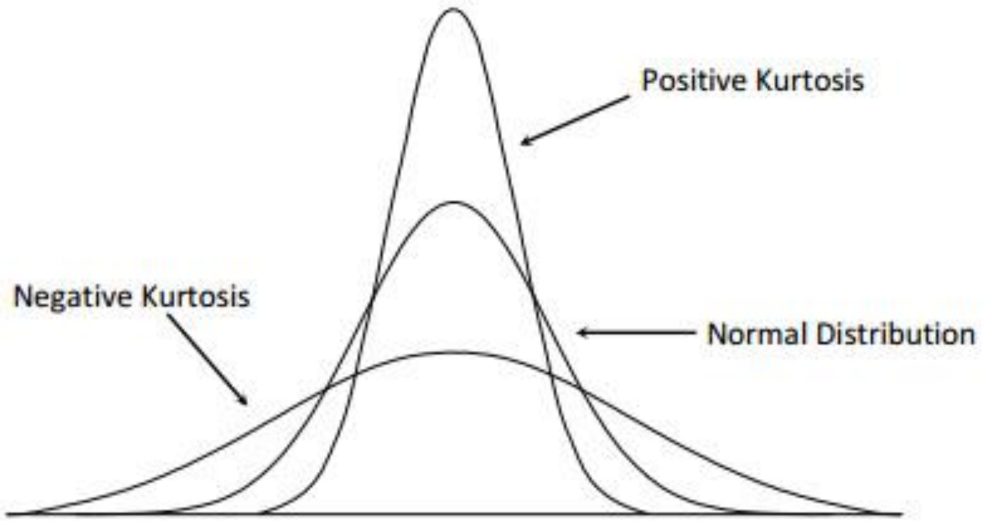
When the frequent scores are clustered at the lower end of the distribution and the tail points to the higher (more positive) scores, the value of skew is positive

When the frequent scores are clustered at the higher end of the distribution and the tail points to the lower (more negative) scores, the value of skew is negative

Shape

- Skewness

- Kurtosis → a measure of the degree

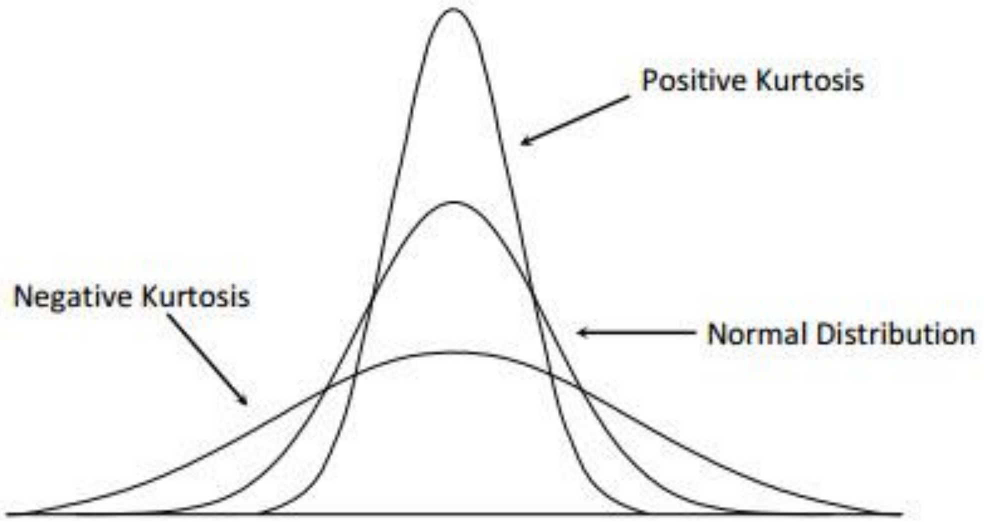


Normal kurtosis = 3

Shape

- Skewness

- Kurtosis → a measure of the degree



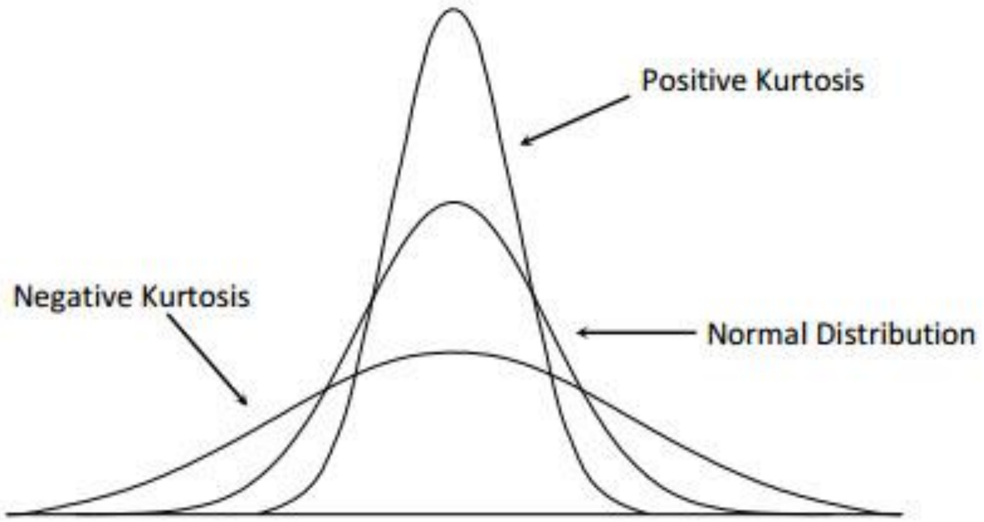
Normal kurtosis = 3

Kurtosis < 3 → Platykurtic
(the distribution produces fewer and less extreme values [e.g., outliers] than does the normal distribution)

Shape

- Skewness

- Kurtosis → a measure of the degree



Normal kurtosis = 3

Kurtosis < 3 → Platykurtic
(the distribution produces fewer and less extreme values [e.g., outliers] than does the normal distribution)

Kurtosis > 3 → Leptokurtic
(this distribution produces more extreme values [e.g., outliers] than the normal distribution)

Threats to descriptive statistics

- Missing data

- Outliers

- Range restriction

Threats to descriptive statistics

- Missing data
- Outliers
- Range restriction

1. Missing Completely at Random (MCAR)

2. Missing at Random (MAR)

3. Missing Not at Random (MNAR; this type of missingness cannot be ignored)

See https://www.theanalysisfactor.com/missing-data-mechanism/ for an explanation of each type of missing data.

Interpreting descriptive statistics

- As previously mentioned, descriptive statistics should be reported in tandem with other descriptive statistics

Interpreting descriptive statistics

- As previously mentioned, descriptive statistics should be reported in tandem with other descriptive statistics
  - The mean is not informative without reporting the corresponding SD
  - The raw frequency is not informative without reporting the corresponding relative frequency
  - Etc.

Interpreting descriptive statistics

- As previously mentioned, descriptive statistics should be reported in tandem with other descriptive statistics
  - The mean is not informative without reporting the corresponding SD
  - The raw frequency is not informative without reporting the corresponding relative frequency
  - Etc.

- Descriptive statistics are the gateway to more sophisticated, in-depth analyses
  - Imagine that you observe low levels of job satisfaction among female employees. The next question that might need to be addressed is, “*Why* are females experiencing low levels of job satisfaction?”