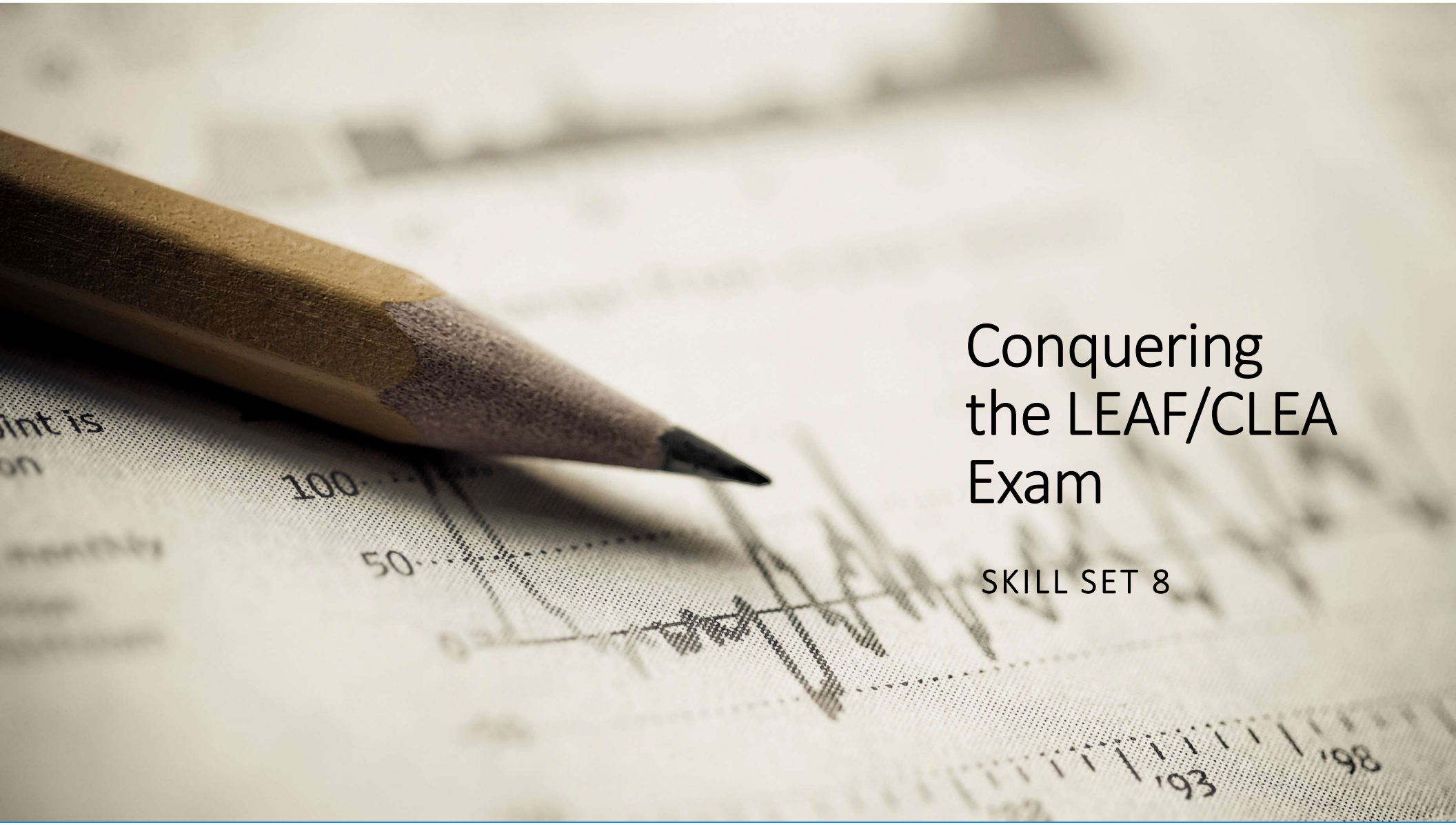


Conquering the LEAF/CLEA Exam

SKILL SET 8

About the Instructor/Course

- Instructor – Jenny Zawitz Jennifer.Zawitz@gmail.com
- CLEA Study Guide: https://iaca.net/wp-content/uploads/2021/06/CLEA-Skill-Sets_Study-Resources-051821.pdf
- LEAF Study Guide: https://iaca.net/wp-content/uploads/2021/06/en_LEAF-Core-Competencies_Study-Resources.pdf
- Exploring Crime Analysis: Readings on Essential Skills (3rd Edition) - IACA
- Each month will cover a different section of the study guide
- Intended as a supplement NOT a substitute for the texts and the Essential Skills classes
  - This course will help you focus your studying, but the courses and text will provide the actual understanding you need to pass the tests

Descriptive Statistics

SKILL SET 8, CHAPTER 8

Statistics General

- Science of collecting and organizing data and then drawing conclusions based on that data
- Three types
- Descriptive – summarize large amounts of information in an efficient and easily understood manner
  - Ex: Measures of central tendency – mean, median, mode
- Multivariate – allow comparisons among factors by isolating the effect of one factor or variable from others that may distort conclusions
  - Ex: Regression, correlation analysis
- Inferential – suggest statements about a population based on a sample drawn from the population (covered in July)
  - Statistical significance, t-test, p-value

Levels of Measurement

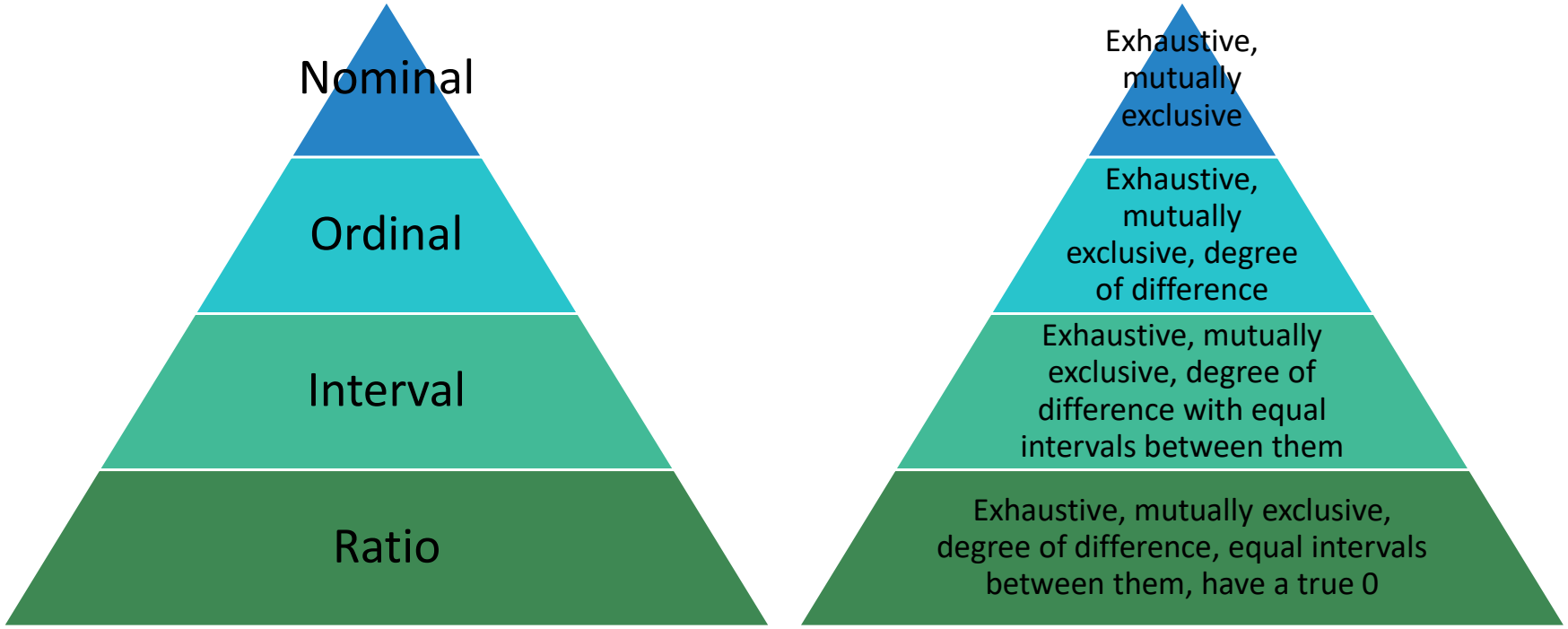
- Used in research and statistics, process of assigning numbers or labels to units of analysis
- Nominal: classify data into categories.
  - Lowest level of measurement.
  - All categories must be mutually exclusive. Only provides labels/names for observations.
  - No way to rank order or perform calculations.
  - Ex: Young vs. Old; Married vs. Unmarried
- Ordinal: Non-meaningful degree of difference
  - Exhaustive and mutually exclusive
  - Also exhibiting a degree of difference between the categories on a scale.
  - Indicates order or ranking between categories but the actual distance between the categories has no meaning.
  - Ex: Under 5, 6-11, 12-21, 21-50, 51-65, 66-85, Over 85

Levels of Measurement

- Interval: scale of equal units/intervals between them
  - Exhaustive and mutually exclusive
  - Exhibiting a degree of difference between the categories on a scale.
  - Order/ranking between categories but have a scale of equal intervals of measurement between them.
  - Ex: 1-20, 21-40, 41-60, 61-80, 81-100

- Ratio: Scale of equal intervals with a true 0 point
  - Exhaustive and mutually exclusive
  - Exhibiting a degree of difference between the categories on a scale.
  - Order/ranking between categories but have a scale of equal intervals of measurement between them.
  - Contains a true 0 point which allows for the indication of the absence of whatever is measured
  - Ex: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Levels of Measurement



Distributions

- Starting to figure out what kind of intervals we want in our data.
- Frequency distribution: list the number or frequency of scores or labels for each individual case.
- Range: condensing frequency information. Highest score minus the lowest score. Use this to create groupings of information into class intervals.
- Can also add percentages to increase information output when demonstrating frequency distributions.

Frequency Distribution Example

Raw numbers: 1, 2, 2, 3, 1, 4, 5, 3, 2, 6, 7, 9, 9, 10, 2, 1, 3

Raw Numbers	Frequency	Freq x Raw
1	3	3
2	4	8
3	3	9
4	1	4
5	1	5
6	1	6
7	1	7
8	0	0
9	2	18
10	1	10
	N = 17	Sum = 70

Count times data occurs

Multiply raw number and frequency of occurrence

Add frequency together

Sum Frequency by raw

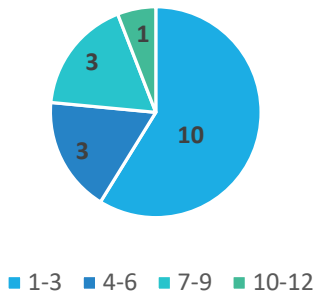
Determining Interval from Frequency

Interval = Range / number of desired intervals

Previous example

Interval = (10-1) / 3 = 3

Can use charts/graphs to portray distribution



Interval	Frequency	Percent	Cumulative %
1-3	10	58.8%	58.8%
4-6	3	17.6%	76.4%
7-9	3	17.6%	94.0%
10-12	1	5.9%	100.0%
	N = 17		

Missing Data?

- Need to deal with missing data or results could be misleading
- Ex: survey respondents didn't answer a question
- Could create a segment labeled "missing cases" and include then in the total
  - Deflate or reduce the other segments with regard to their relative percentages
- Omit missing cases all together
  - Not included in interval or the total case numbers
  - Inflate the relative percentages for each class interval
- Must always indicate if missing cases have been included or excluded.

Measures of Central Tendency

- Mean: Average. Distribution of data can only have one mean.
  - Distribution impacted by outliers or extreme scores.
  - Useful with interval and ratio data
- Median: Midpoint or middle score of the distribution.
  - Point where 50% of scores are above the median and 50% are below.
  - Strength in that it is not impacted by extreme scores
  - If odd number of cases, rank order the scores and determine the middle case
    - (n + 1) / 2 with n being the total number of cases
  - If even number of cases, average the two cases at the midpoint.
- Mode: most frequent score
  - Can be unimodal (one mode) or multimodal (more than one mode)
  - Best used with nominal data.

Measures of Central Tendency Examples

Raw numbers: 1, 2, 2, 3, 1, 4, 5, 3, 2, 6, 7, 9, 9, 10, 2, 1, 3

Numbers in rank order: 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 6, 7, 9, 9, 10 (n=17)

Mode: 2 (score occurs 4 times)

Median: 3 (score in the middle).

Mean: (1 + 2 + 2 + 3 + 1 + 4 + 5 + 3 + 2 + 6 + 7 + 9 + 9 + 10 + 2 + 1 + 3) / 17 = 4.12

- Can see how the mean is a little more extreme because of the 10 and two 9s.

Measures of Dispersion or Variability

- Range: create groupings of data into class intervals. Indicates distance between the highest and the lowest values in a distribution.
  - Can be unstable because it only looks at extreme values. Could be a large distance between the two.
- Variance: sum of the squared deviations of each score from the mean, divided by the total number of cases. Summarizes the dispersion of scores around the mean.
  - 1) calculate mean of distribution. 2) calculate the deviation from the mean for each score. 3) square each number in step 2. 4) find the mean of the squared deviations in step 3.
- Standard Deviation: measures the average distance that each data item is away from the mean of all the data in the distribution. Attempt to create a curve from the data. Demonstrates how scores compare with each other and comparison between two distributions.

Measures of Dispersion Examples

Raw Data: 1, 6, 12, 17, 18, 24

Range: 24-1 = 23

Variance: First, determine mean. (1+6+12+17+18+24)/6 = 78/6 = 13

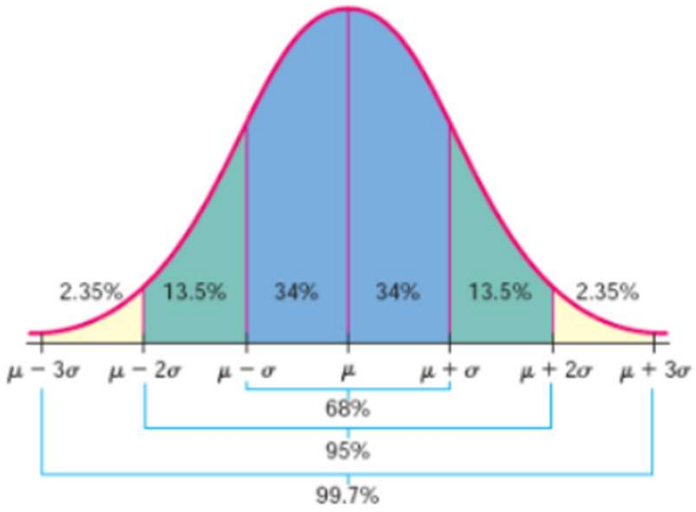
Raw Data	Mean	Deviation	Dev. Squared
1	13	(1-13) = -12	(-12)^2 = 144
6	13	(6-13) = -7	(-7)^2 = 49
12	13	(12-13) = -1	(-1)^2 = 1
17	13	(17-13) = 4	(4)^2 = 16
18	13	(18-13) = 5	(5)^2 = 25
24	13	(24-13) = 11	(11)^2 = 121
		Sum should be 0	356

Variance = 356/(6-1) = 356/5 = 71.2

Standard deviation = $\sqrt{71.2}$ = 8.44

Normal Distribution and Skewness

- Unimodal distribution can be normal, positive, or negative
- Skewness: shows the spread of scores weighted to one side of the mean (positive or negative)
  - The farther apart the measures, the more skewed the distribution
- Normal distribution (bell shaped curve) means that all scores are evenly distributed throughout the distribution. No skewness or extreme scores.
  - Key is that mean, median, and mode will be approximately equivalent.
  - Curve is symmetrical about the mean and it never touches the x-axis.
  - Approximately 68% of all cases are within 1 SD, 95% are within 2 SD, and 99.7 are within 3 SD.



Skewness

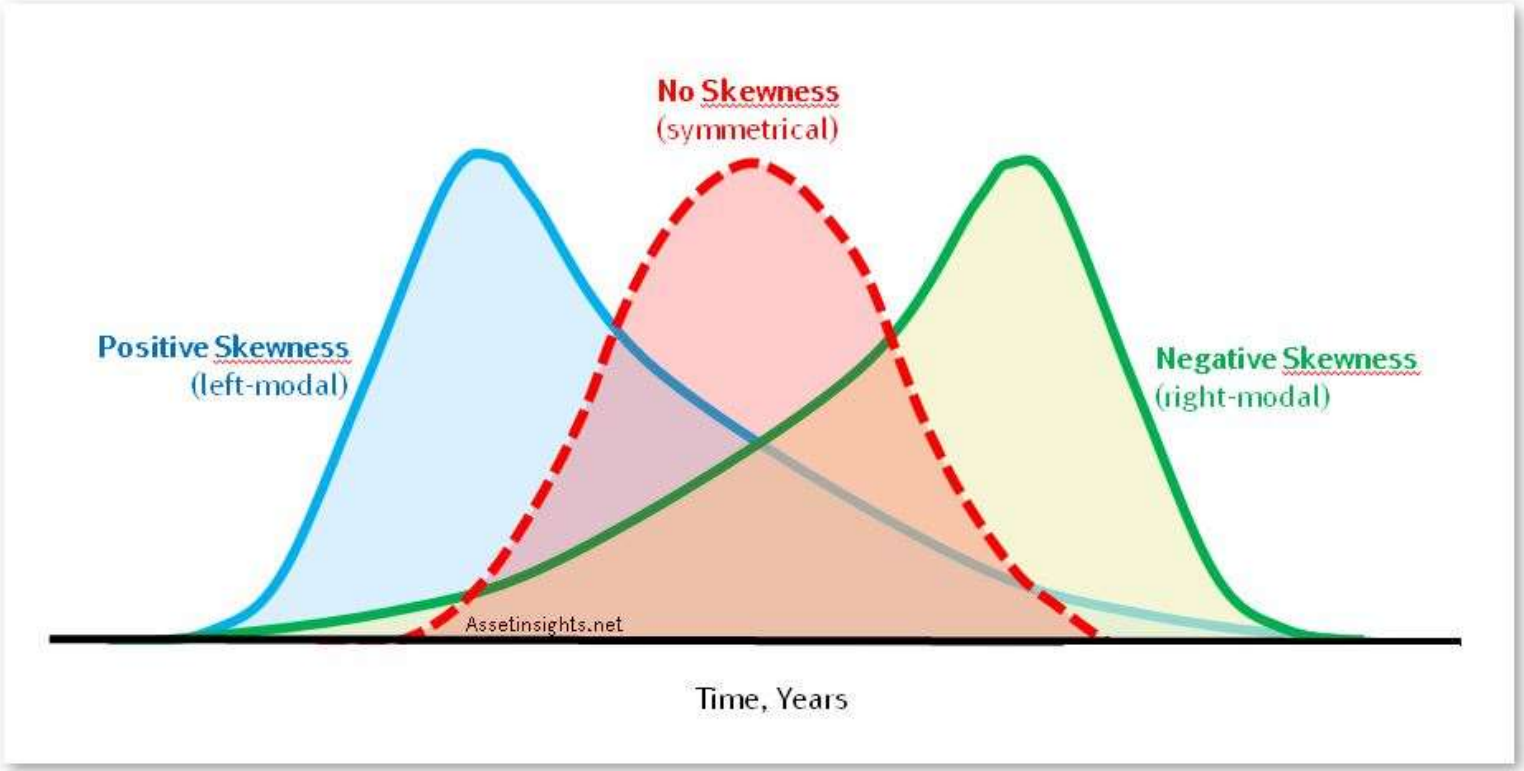
POSITIVELY SKEWED

- Scores weighted to the left (hump located to the left and the tail moves right)
- Right tail is longer because a few scores have values much higher than the rest
- Mode is smallest value followed by median then mean

NEGATIVELY SKEWED

- Scores weighted to the right (hump located to the right and the tail moves left)
- Left tail is longer because a few scores have values much lower than the rest
- Generally the mean is smaller than the median and the median is smaller than the mode

Skewness



Rates

- Standardize some measure for comparative purposes. Allow for comparison of numbers of different sizes.
- Calculate by dividing raw numbers by a comparable denominator
- Ex: Crime rate. Compare instances of crime or a specific crime by population.
- Number of crimes/population X 100,000
- Determine crime rate per 100,000 with the 100,000 being the number that standardizes the rates.

Proportion and Percent

- Proportion is a dichotomous variable which means that it only has two categories.
- Proportion has a range between 0 and 1
- Divide the value by the whole.
- Ex: What is the proportion of crimes that are Part 1 Crimes in a city in which there are 25 Part 1 crimes out of 100 total crimes each month?
- 25 (value)/ 100 (whole) = .25
- Rarely report proportion but instead report percent. Percent is a dichotomous variable with values between 0 and 100. Multiply the proportion by 100 for the percent.
- Ex: What is the percent of crimes that are Part 1 Crimes in a city in which there are 25 Part 1 crimes out of 100 total crimes each month?
- (25/ 100) * 100 = .25 *100 = 25%

Percent Change

- ((New – Old) / Old) * 100
- Ex: Percent change where there were 25 Part 1 crimes this month and 15 Part 1 crimes last month.
- ((25-15) / 15) * 100 = (10/15) * 100 = 66.7 or a 66.7% increase
- Can be positive or negative (positive means increase, negative means decrease)
- Ex: Percent change where there were 25 Part 1 crimes this month and 50 Part 1 crimes last month.
- ((25-50) / 50) * 100 = (-25/50) * 100 = -50 or a 50% decrease
- Be careful comparing two time periods exclusively, especially where a long time has passed
- Always report percentages and percent changes with their raw numbers
- Note: small changes in small numbers can lead to big percent changes.

Multivariate Statistics

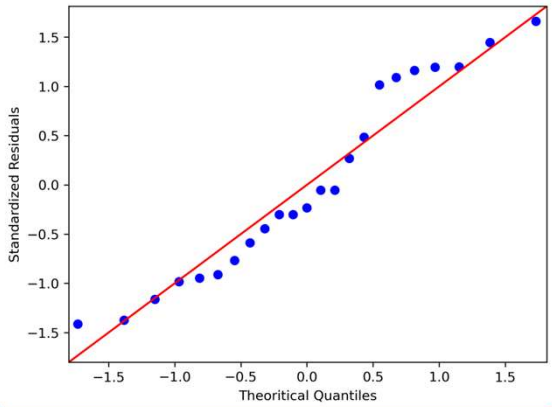
- Analysis of two or more variables looking for an association between/among the variables.
- If you subject a dataset to univariate or multivariate analysis, you are assuming that they relate to each other in some way.
- Before running your analysis, consider if the variable concepts logically relate to each other and consider the data type. You also want to make sure that they are the same standard of measure.
- Stats are multivariate when two or more variables (typically ratio or interval type) are examined together to see if there are observable relationships between them.
- Ex: measure of a variable over a number of intervals of time or a time series analysis.
- Best to put data into an Excel spreadsheet to visualize it, usually with a scatter diagram graph. Looking for a trendline to see if we can find a pattern (increase/decrease in crime)

Regression Analysis

- Procedure for pattern recognition.
- If a trendline can be found and it represents enough of the points, it may have predictive qualities.
- Regression and correlation analysis go hand in hand. Regression analysis locates the best line that runs through the data points and shows you what it looks like. Correlation analysis tells you how strong the association is and to what extent you can draw inferences from it.
- Remember correlation does not equal causation.
- If your regression has a curve, the trendline will be considered non-linear. An example of this is the seasonal effect of crime – moves like a wave with a rise in the summer typically.
- Linear functions generally show up as a straight line in your data points and can be examined with algebraic functions. (think slope)

Linear Regression

- Object of linear regression is to find the best straight line through points. Points may not be in a line, so you may have points surrounding your line.
- Formula for a straight line is y=mX + b where
  - b is the y intercept aka the location on the y axis where the line intercepts
  - m is coefficient of slope (constant and used as multiplier of x)
- Regression line takes the place of the mean in univariate analysis
- Like standard deviation in univariate analysis, standard error is the square root of the variance in regression.
- Larger standard error = less accuracy of regression and less reliability of projections.



Independent and Dependent Variables

- Want to not only determine the trend line for ratio or interval variables but also the degree of association or correlation between them. How much does one variable impact the other.
- Correlation does not equal causation as many factors can also be influencing any relationship the variables may have.
- Start with a hypothesis that is explaining what relationship you believe the variables will have. Hypothesis must be testable.
- Independent variable is the thing you are manipulating while dependent variable is being measured.
- Dependent variable depends on the independent variable. IV is what you change and the DV changes because of that. The score you get on your test (DV) DEPENDS on how much you study (IV).

Correlation Analysis

- Measures the “fit” or degree of association between two variables. Provides a measure of quality of the regression line.
- Pearson’s product-moment coefficient of correlations – represented by “r”
- Ranges from -1.0 to 1.0
- -1.0 is a perfect negative correlation while 1.0 is a perfect positive correlation. 0 is no correlation (very unlikely).
- Negative correlation means as one variable goes up the other goes down.
- Positive correlation means as one variable increases, so does the other OR as one variable decreases so does the other.

Correlation Analysis

- To determine how much of the variance is accounted for by the change in variables, use the coefficient of determination or r-squared (r^2)
- Basically this determines how much of the change in one variable is explained by the change in the other.
- Assume the correlation between two variables is 0.8 (r = 0.8). The coefficient of determination is 0.8 squared or 0.64 (r^2 = (0.8)^2 = 0.64). This indicates that 64% of the variance is accounted for by the changes in the independent variable. The remaining percentage is explained by other factors.

Summary Multivariate Stats

- Regression analysis involves pattern recognition and a placement of a trendline (straight or curved) with a mathematical function.
- Correlation analysis tells us how strong the association is and to what extent we can assume it is meaningful, useful, and reliable.
- Nonlinear relationships are more common.
- Logistic curve – gradual takeoff, period of exponential increase, then slowing or deceleration of the rate of growth. True of population, crime rate, etc. Nothing can go up forever.

Conclusions

- Read the books and take the classes to strengthen understanding.
- Check out Statistics for People Who Hate Statistics.
- Try to apply the things learned to your every day work to “make them stick”.
- Use the study guides.
  - https://iaca.net/about-clea/ (links for program outline and study guides here)
  - https://iaca.net/about-leaf/ (links for program outline and study guides here)
- Next month: Inferential Statistics (Skill Set 9)

Any questions?