Outlier

Section Summary | Introduction to Statistics
Descriptive Statistics Part I

Graphs for Visualizing Data

1. Graphing and visualizing data
    a. Visualizing data will help you determine what statistical tests to run
    b. Types of graphs:
        i. Frequency table
            1. Lists the number of data points that have a certain characteristic
            2. Good for qualitative data
            3. Can be visualized using a bar graph
        ii. Bar graph
            1. Good for qualitative data
            2. The horizontal axis displays the possible values of the variable
            3. The height of a bar corresponds to the frequency that the value appears in the data
            4. Make sure the bars are all the same width because information is only contained in the height of the bars
        iii. Dot plot
            1. Good for quantitative data
            2. The horizontal axis represents the range of values for the variable
            3. Each data point is plotted above the correct value on the horizontal axis
            4. Repeat points are stacked on top of each other
            5. Can be used to visualize the shape of the data, such as the center and spread
            6. Can also be used to compare multiple datasets
            7. Dot plots are similar to bar graphs in that the height of the dots corresponds to the frequency of that value
        iv. Time series graph

1. Time series data is quantitative data collected at various points in time
2. Good for seeing change in a variable over time
3. Horizontal axis is time
4. The scale of the horizontal axis depends on the frequency that data was collected (every year, every month, every minute, etc)
5. The vertical scale depends on the data
6. Connect the data points in your graph to visualize the trend

2. Manipulating Data
    a. Truncation: starting the axis at a value greater than 0 to exaggerate differences
    b. Scale Manipulation: changing the range of an axis
        i. Can alter the shape of the graph
        ii. Be sure to choose a scale that fits the context of the data to avoid ethics complains
    c. Cherry picking: choosing to exclude certain data points

Histograms

1. Frequency charts for quantitative data
    a. Just like with qualitative data, we can make frequency charts for quantitative data following these extra steps:
        1. Define classes
            - Your classes must cover the full range of data
            - Break the range of the data into smaller classes of equal width
            - The number of classes depends on context, but it's usually 5-20
            - If you choose too few classes, all your data will clump up into a couple blocks
            - If you choose too many classes, your data will be very spread out
            - Neither of these scenarios will be informative, so try to choose an appropriate number of classes based on the size of your data
            - Use round numbers to define your classes to make analysis easy
            - Lower class limit = lowest value of each class
            - Upper class limit = largest value in each class
            - Class midpoint = value exactly in the middle of each class, calculated as (lower limit + upper limit)/2
            - Class width = difference between two consecutive lower class limits (not the difference between upper and lower class limits)
        2. Calculate frequencies
            - Tally up all the data points in each class
            - The number of data points in a class is that class's frequency

- Relative frequency = frequency divided by the total number of data points
- Relative frequency tells how large a frequency is compared to the size of the data, so it adds context to a frequency

b. Also like frequency tables for qualitative data, we have methods of visualizing frequency tables for quantitative data

2. Histograms
    a. A histogram is a graph of bars with equal width where the height of each bar represents the frequency for that class
    b. These are like bar graphs, but the horizontal axis is now a continuous scale of values
    c. Histograms help us see the shape of the data (center and spread)
    d. Outliers - extreme values that lie far from the center
    e. Histograms help us identify outliers
    f. They can also help visualize the difference between two datasets
    g. You can make relative frequency histograms as well
        i. A relative frequency histogram will have the same shape as an absolute frequency histogram
        ii. The only difference is the scale of the vertical axis
        iii. An absolute frequency histogram will have counts as the vertical axis, while a relative frequency histogram will have percentages for the vertical axis
    b. Frequency polygon
        i. Mark the top-center of each bar with a dot and connect the dots with a line
        ii. When you take away the bars, you will leave a frequency polygon that shows the shape of the histogram
        iii. These are used to simplify the graph so we can plot multiple polygons on the same graph for comparison
        iv. It also becomes easier to view the changes from class to class
3. Common distribution shapes
    a. Uniform distribution

   i. All bars are about the same height, so each class has the same frequency
   ii. You have an equal probability of falling into each class
b. Normal distribution
   i. Very common and very important
   ii. Symmetrical about its center
   iii. Shaped like a bell
   iv. A lot of real-world data is normally distributed
c. Right- and left-skewed distributions
   i. These look like normal distributions, but they have tails extending to the right or left
   ii. When the tail extends to the right, it is called right-skewed
   iii. When the tail extends to the left, it is called left-skewed
   iv. These graphs are named by the location of the elongated tail, not the location of the majority of the data (a common mistake)

Paired Data

1. Scatter Plots
    a. Paired data is two datasets with a 1 to 1 relationship
    b. We can use a scatter plot to represent paired data
    c. 1 variable goes on the horizontal axis and the other variable goes on the vertical axis
    d. Choose axis scales that suit the data
    e. Plot data points as ordered pairs where the first coordinate is that data point's value for the horizontal axis variable and the second coordinate is that data point's value for the vertical axis variable
    f. Scatter plots are good to show correlation between the two variables
2. Types of correlations
    a. Linear correlation: the relationship between the two variables can be plotted as a line
    b. A line can be drawn through the center of all the data points to represent the linear relationship is known as the regression line, line of best fit, or least squares line
    c. Nonlinear correlations occur when the variables exhibit some patterned relationship that is not a straight line
    d. Positive correlation: as 1 variable changes, the other variable changes in the same direction (think positive slope)
    e. Negative correlation: as 1 variable changes, the other variable changes in the opposite direction (think negative slope)
    f. Strong positive correlation: the correlation between the 2 variables is positive, but there is a lot of variance in the data, making the points group more loosely

g. Strong negative correlation: the correlation between the 2 variables is negative, but there is a lot of variance in the data, making the points group more loosely
h. No correlation: there is no pattern in the data, and the data points seem to be laid out randomly

1.