Outlier

Section Summary | Introduction to Statistics
Descriptive Statistics Part II

Measures of Center

1. Resistance: a measure of center is resistant if the presence of outliers does not change the center by much
2. Mean
    a. Average of the data
    b. Calculated by adding up all the values and dividing by the total number of values
    c. Sample Mean:
        i. $\bar{x} = \frac{\Sigma x}{n}$
        ii. $\bar{x}$ = "x bar" = sample mean
        iii. Σ = "sigma" = sum up
        iv. x = individual data value
        v. Σx = sum up all the individual data values
        vi. n = sample size
        vii. $\bar{x}$ is a sample statistic
    d. Population mean
        i. $\mu = \frac{\Sigma x}{N}$
        ii. μ = "mu" = population mean
        iii. Σx = sum up all the individual data values
        iv. N = population size
        v. μ is a population parameter
    e. Mean is not resistant because extreme values (outliers) get pulled into the calculation
3. Median
    a. The data point in the middle of the dataset when the values are arranged in ascending order
    b. When your dataset has an odd number of values, there is only 1 value directly in the middle - that value is your median
    c. When your dataset has an even number of values, there are 2 values in the middle - the median is the sum of those 2 values divided by 2

   d. Median is resistant because we don't directly use all data points in finding the median - outliers, by definition, will never fall in the center
   e. When your data is symmetrical like the normal distribution, the mean is approximately equal to the median
4. Mode
   a. The value with the greatest frequency
   b. No calculation is necessary - just look at your data and find the value that occurs the most
   c. Unlike the mean and median, the mode can be used for both quantitative and qualitative data
   d. When you have qualitative data, the only measure of center you have is the mode
   e. You can have 0, 1, or multiple modes depending on how many values occur the most
   f. A dataset with 2 modes is called bimodal
   g. A dataset with more than 2 modes is called multimodal
   h. Mode is resistant because outliers, by definition, will not be the most frequent value

Measures of Spread

1. Why do we need measures of spread?
    a. Measures of center don't tell the whole story
    b. 2 distributions can have the same center but vastly different shapes
    c. Spread gives more information on how a distribution is shaped
2. Range
    a. The difference between the largest value and smallest value in your dataset
    b. Range = max - min
    c. Range is not resistant because outliers will be those max/min values
    d. Easy to compute, but not the most informative metric for spread
3. Standard deviation
    a. Measure of how much individual data points deviate from the mean
    b. Sample standard deviation:
        i. $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$
        ii. s = sample standard deviation
        iii. Σ = sum
        iv. x = individual data values
        v. $\bar{x}$ = sample mean
        vi. $\Sigma(x - \bar{x})^2$ = sum of the squared deviations, i.e. take each value, subtract the mean, square that difference, then sum up all the squared differences
        vii. n = sample size
    c. Population standard deviation:

i. $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

ii. σ= "sigma" = population standard deviation

iii. μ = "mu" = population mean

iv. N = population size

d. The formula for sample standard deviation uses n-1 instead of n because this gives a better estimate for the population standard deviation (it also has to do with the reduced number of degrees of freedom, which you'll get into later)

e. Standard deviation can never be negative, and can only be 0 if all the data points are the same

f. Larger standard deviation means more spread

g. Standard deviation is not resistant because every value is included in the calculation

h. Standard deviation can be difficult to calculate, but you can use Desmos

i. Type your dataset into Desmos as x = [x_1, x_2, x_3,...,x_n] where x_i are your data points and they are separated by commas

ii. Calculate your sample standard deviation with the function stdev(x)

iii. Calculate your population standard deviation with the function stdevp(x)

i. Standard deviation can be used to identify extreme values

j. In general, points that are more than 2 standard deviations away from the mean can be considered extreme

4. Variance

a. Variance is simply the square of the standard deviation

b. Sample variance: $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

c. Population variance: $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$

d. Same properties as the standard deviation

i. Never negative

ii. Only 0 if all the values are equal
iii. Not resistant
iv. Larger variance means more spread

e. Some statistical measures require either standard deviation or variance specifically, so we define them separately
f. Given the standard deviation, find the variance by squaring it
g. Given the variance, find the standard deviation by taking the square root of it

Measures of Location + Boxplots

1. Percentile
   a. Percentiles divide the data into 100 groups with about 1% of the data in each group
   b. Given a data point, x, the percentile of x = $\frac{number\ of\ values\ less\ than\ x}{total\ number\ of\ values} \times 100$
   c. This tells the percentage of data points beneath x
   d. Given a percentile, we can locate the specific data point(s) corresponding to that percentile
      i. L = $(\frac{k}{100})n$
      ii. L = locator that tells the position of the desired data point in an ordered (ascending) list of the data points
      iii. n = total number of data points
      iv. k = percentile
      v. If L is a whole number, $P_k = \frac{L^{th}\ value + (L+1)^{th}\ value}{2}$ where P_k is the desired data point
      vi. If L is not a whole number, round L up to the next whole number, and P_k is the L^{th} value after L is rounded up
   e. Many programs have several different methods of calculating percentiles, but each method gives very similar results
2. Quartiles
   a. Quartiles divide the data into 4 groups with about 25% of the data in each group
   b. Quartiles are really just special names for the 25^{th}, 50^{th}, and 75^{th} percentiles
   c. Q1 = 25^{th} percentile

d. Q2 = 50th percentile
e. Q3 = 75th percentile
f. Along with the maximum and minimum data points, these 5 values gives us the "5 value summary"

3. Box Plots
    a. You can use this 5 value summary to make a box plot
    b. Draw 3 horizontal lines that mark Q1, Q2, and Q3
    c. Draw a box around these 3 lines
    d. Extend vertical tails to the maximum and minimum values to complete the plot
    e. Interquartile Range (IQR): the difference between Q1 and Q3 (Q3 - Q1)
    f. Outlier: any data point that is above Q3 + $(1.5 \times IQR)$ or below Q1 - $(1.5 \times IQR)$
    g. The IQR defines a "normal" range of values
    h. You can show outliers in a modified box plot
        i. Contract the vertical lines to only extend to the maximum and minimum values that are not outliers
        ii. Mark outliers with stars
        iii. This will make your vertical lines more proportional to the box