# Section 2.5 Measures of the Center of the Data

In addition to the median which measures the center of the data, there is another measurement called the mean or average. Unlike the median, the mean considers all data values.

n = the number of data values

This can be a good measurement of the center of a dataset, if all the values are close together. However, if we have values that are extreme (too low or too high) compared to the data values. For example, in a list of salaries we have one person who earns 40 times the salary as the others, the mean as the center of the data may not be the best measurement.

60000, 55000, 65000, 58000, 58000, 59750, 62000, 63000, 2500000

When you have extreme values or outliers then the median is the best measure of center of the dataset.

Remember the reason we compute the median or mean/average is to locate a measurement that tells us about the population of interest.

# **Law of Large Numbers**

The Law of Large Numbers indicates that if the sample size gets larger and larger the sample mean, , approached the population mean, .

# Calculating the Mean of Grouped Data



# **Section 2.6 Skewness and the Mean, Median, and Mode**

The mode is defined to be the value that occurs most often. You can multiple modes in a data set.

5, 6, 7, 7, 8, 8, 9, 10

The modes are 7 and 8.

If a distribution is symmetric, the mean, median, and mode are the equal.

5, 6, 6, 7, 7, 7, 8, 8, 9

# Left skewed distributions

If the distribution is left skewed, then extreme values appears on the left.

0, 1, 7, 7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11

# Right Skewed Distribution

7, 7, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11, 26

# **Section 2.7 Measures of the Spread of Data**

The range is a measure of spread of the data. The range is defined as the maximum value – minimum value.

Range = Maximum - Minimum

It can tell you the distance between the two extreme values; but it does not take into account all the data values.

The next, measure of spread is the interquartile range.

IQR = Q3 – Q1

It measures the distance where 50% of the data lies. The interquartile range is used to determine outliers. But it does not use all the data in computing the range.

A new measure of spread is called the standard deviation. It considers all the data values. It measures the difference of each data value from the mean and squares the distance. Then it totals the squared deviations and divides by n is we are looking at all values in the population. Or it divides by n-1 if we are looking at the sample distribution. The value is called variance

Then we take the square root of the formula for standard deviation.

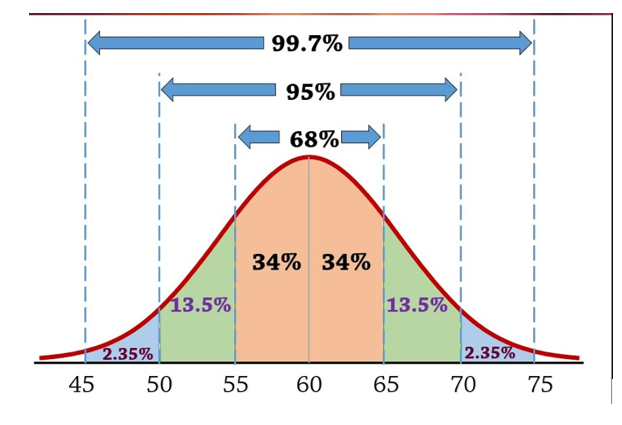


# **Empirical Rule**

When the data is symmetric, we know the approximate percentage of data with standard deviations.

* Between
* Between
* Between

The chart below shows the Empirical Rule when the



60 – 5 = 55 and 60 + 5 = 65 68% of the data is between 55 and 65

**60 – 2\*5 = 50 and 60 + 2\*5 = 70 95% of the data is between 50 and 70**

60 – 3\*5 = 45 and 60 + 3\*5 = 75 99.7% of the data is between 45 and 75

# **Outliers**

We say that values outside the 95% interval is unusual. To find extreme values, use the following formula.

the values are outliers.

Not all curves are symmetric, when they are skewed, we can Tchebysheff’s Theorem to determine the percent of data within standard deviations from the mean.

# **Tchebysheff’s Theorem**

When the data cannot use the Empirical rule to describe it. For example, the data set can be skewed (left or right), flat, or some shape other than a bell-shaped curve. There is a more conservative rule to describe the data.

𝑥̅± 𝑠 at least 0%

𝑥̅± 2𝑠 at least 75%

𝑥̅± 3𝑠 at least 89%

𝑥̅± 𝑘𝑠 at least 100%

Comparing the Variability of two data sets

Suppose you wanted to compare two blood pressure drugs. The measures below are the amount of time in hours the medicine will control the blood pressure in a normal range.



To find the mean of the data set copy the data to an Excel spreadsheet and enter the following formulas.

To find the Mean enter =AVERAGE(highlight the data)

To find the Standard deviation enter =STDEV.S(highlight the data)

To find the Coefficient of Variation enter =100\*cell of Standard deviation/cell of Average

Click on the table above to see the Excel Spreadsheet.

Drug A seems to be a bit better medication to use looking only at the mean. However, when you look at the spread of the values in each data set, the Drug A is more spread out, the standard deviation is larger. Therefore, the range of values within 2 standard deviations from the mean is 27.3 – 2\*12.68 and 27.3 + 2\* 12.68(1.94 to 52.66 hours).

Drug B is less spread out. That means when you take the drug it is more reliable since the range of values is 26-2\*1.70 and 26 + 2\*1.70 ( 22.6 to 29.4 hours).

Another way to measure variability is the **Coefficient of Variation.** It is defined to be

Whichever, population or sample has the smallest percentage is the less variable.