General Assembly

STUDY GUIDE

DISTRIBUTIONS AND DESCRIPTIVE STATISTICS

Key Terms

Exploratory Data Analysis: An approach to analysis that uncovers characteristics and trends within a data set using basic statistics and simple visualizations.

Probability Distribution: A table or graph that shows a value's observed frequency of occurrence — in other words, how often a value or range of values occur in the data set.

Descriptive Statistics: Statistics that summarize a sample data set. Descriptive statistics tell us information about the data without making any predictions or assumptions about a population.

Inferential Statistics: Statistics that extrapolate conclusions about a population based on a sample. Inferential statistics make assumptions based on patterns established by the sample.

Mean/Average: The arithmetic mean of a column or row of numbers. It's calculated by finding the sum of values and dividing that by the number of values included in the sum.

Median: The middle value of a column or row of numerical values sorted from smallest to largest.

Mode: The value that appears most frequently in a row or column of data.

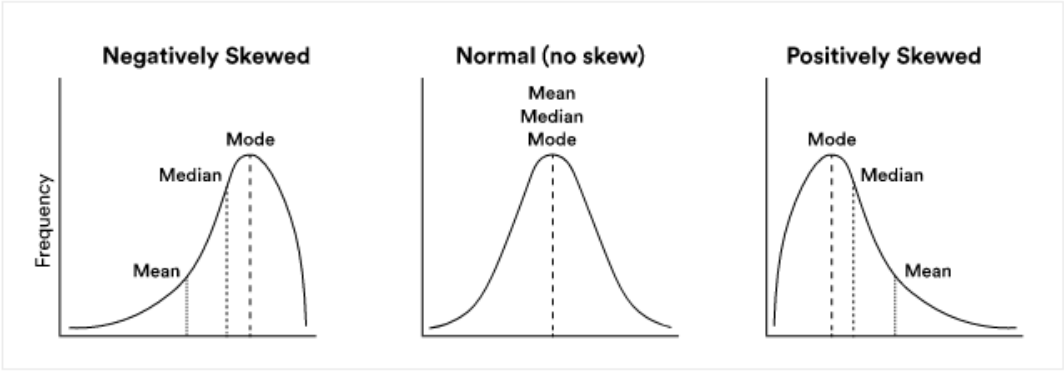
Standard Deviation: A measure of how greatly a data set varies from its mean. The more spread apart the data, the higher the deviation. The STDEV.P formula in Excel is used for calculating the standard deviation of a population, while the STDEV.S formula is used for a sample.

Histogram: A chart that displays the frequency distribution of a set of quantitative data. These are not the same as bar charts. Use the Data Analysis ToolPak to create one.

Cheat Sheet

1. Understanding Distributions
    - » Skewness: The degree to which outliers have affected the balance of a data set, as measured by the difference between the median and the mean.
    - » If the mean is larger than the median, the data has a positive (right) skew.

» If the median is larger than the mean, the data has a negative (left) skew.



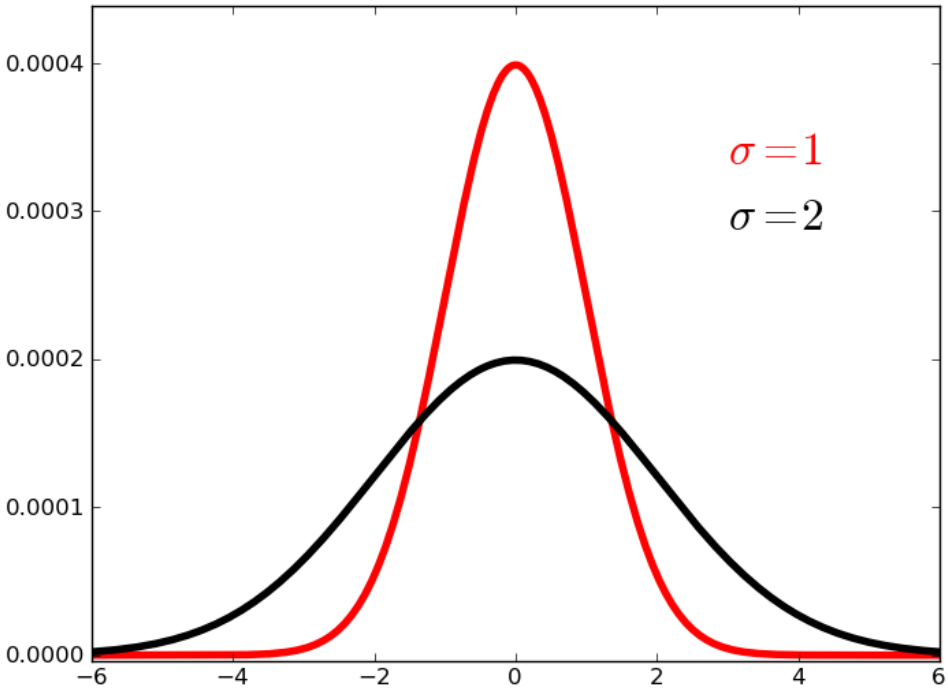
2. Spread: How far the data are from the center of the data set.

» The average of 0 and 100 is 50. The average of 49 and 51 is 50. But 0 and 100 are much further from 50 than 49 and 51. Therefore, 0 and 100 would have a larger spread.

» For normally distributed data, we talk about spread in terms of standard deviation.

» For skewed data, we talk about spread in terms of IQR.

» In the chart below, the red distribution has a smaller standard deviation than the black distribution. These curves show the general shape of the underlying histogram.



3. Percentiles

» A measure indicating the value below which a given percentage of observations in a group of observations fall.

» Are based on the number of items in the data set, not the values of the items.

» Use the PERCENTILE.EXC function in Excel to get the value of a particular percentile.

» Use the PERCENTRANK.EXC function in Excel to get the percentile for a particular value.