Mathematics for Management Online Course:
Statistics Section: Concept Summary

Statistics

Histograms
To summarize data with a histogram, begin by dividing the possible data values into 5 to 10 bin ranges and determine how many data points fall into each bin. The histogram option of the Excel Data Analysis Toolpak makes it easy to create a histogram.

Measures of Central Location
The mean, median, and mode are the primary measures used to summarize the typical value or central location for a data set.

- The mean is just the average of the numbers in a data set.
- Roughly speaking, the median is the point in a data set where half the observations are less than the median and half the observations are more than the median. If the data set consists of an even number of data points, the median is the average of the two middle observations. If the data set consists of an odd number of data points, the median is the middle observation.
- The mode is the most frequently occurring value in a data set.
- Excel's AVERAGE, MEDIAN, and MODE functions can be used to compute measures of central location.

Skewness
If a data set exhibits a lot of skewness, then the median is a better measure of central location than the mean. Otherwise, the mean is a better measure of central location.

Measures of Variability
The sample variance and sample standard deviation are measures of a data set's spread about the mean or variability. Given data points $x_1, x_2, \ldots, x_n$, the sample variance of the data set (written as $S^2$) is defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$  

The sample standard deviation is the square root of the sample variance. Excel's VAR function finds the sample variance, and Excel's STDEV function finds the sample standard deviation.
Rule of Thumb and Outliers
When a data set has a fairly symmetric histogram, approximately 68% of the data is within one standard deviation of the mean and approximately 95% of the data is within two standard deviations of the mean. Any data point that is more than 2 standard deviations away from the mean is called an outlier.

Covariance and Correlation
Covariance and correlation are measures of linear association between two sets of data $X$ and $Y$ consisting of the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The sample covariance between data sets $X$ and $Y$ is given by

$$\text{ Covariance } = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The sample correlation between two data sets $X$ and $Y$ is given by

$$\text{ Correlation } = \frac{\text{ Covariance}}{S_X S_Y},$$

where $S_X$ and $S_Y$ are the standard deviations of the data sets. Correlation is a unit-free measure of linear association, so it is used more often than covariance.