

## Vocabulary for Statistical Sleuths

1. Hypothesis - *The American Heritage College Dictionary, Third Edition* defines statistics as a tentative explanation that accounts for a set of facts and can be tested by further investigation; a theory. [NOTE: HOW TO FORMULATE A HYPOTHESIS - First students observe an occurrence and explore why it happens. Next, students write down what happened in their own words. To further explore what happened, students decide what they want to find out about the occurrence. Using observations and by designing logical how and why questions, students write a statement that describes what they want to do. (This defines the purpose of the experiment.) Based on the questions identified in the preliminary investigation, students make a list of **answers** to each of the questions. These answers should be worded in statements that describe the how or why of what happened in the occurrence. These statements become the hypotheses. These statements need to have logical, measurable outcomes. One occurrence can have many hypotheses.]
2. Experiment - A process of determining if a hypothesis is true or false. This process is usually scientific in nature and provides answers for measurable questions.
3. Population - An entire group to be examined (including every member of the group).
4. Statistical survey - A method designed to collect statistical data about a specific population. Surveys can be designed to answer who, what, when, where, why, or how about a population.
5. Statistics - *The American Heritage College Dictionary, Third Edition* defines statistics as the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. (Boston, New York: Houghton Mifflin Company, 2000) 1328.
6. Sample - A specific portion of a population (usually consisting of people or objects).
7. Random sample - Each member of the population is selected entirely by chance and has an equal chance of being selected.
8. Systematic sample - One member of the population on a random basis is selected, and each additional member is selected at evenly spaced intervals until the desired number for the sample space has been collected. NOTE: AN EXAMPLE - If every fifth, tenth, fifteenth, etc. member of the population is selected.
9. Stratified sample - The entire population is divided into meaningful subgroups (strata) and then each group is randomly sampled (as described by the random sample above).
10. Sample bias - When the method used to acquire a sample results in a sample that is systematically different from the population, it is called a biased sample. Note: Examples of common misuses of probability and statistics include inadequate sample size; incomplete or incorrect graphs; over-generalized results; over-interpretation of numerical data; use of raw data, percents, or statistics (range, median, mean, or mode) to misrepresent the data collected; misrepresentation of the likelihood and significance of a result. (See the FCAT specifications provided

## Vocabulary for Statistical Sleuths Continued

by the Florida Department of Education at  
<http://www.firn.edu/doe/sas/fcat/fcatis01.htm> .)

11. Data displays - An organized collection of data. NOTE: Several forms of data displays: tables, line graphs, charts, bar graphs, histograms, and box-and-whisker graphs. Each type of data display has a specific function. A line graph is appropriate to show data that ***changes over time***. A bar graph helps to show a comparison between two or more things. The circle graph does a better job of showing percentages and relationships of the part to the whole. A table and a chart are primarily used to organize data in some sort of structure, but do not usually go the extra step to show a relationship between the data. (See the definitions for histogram (#12) and box-and-whisker graph (#24) in this document.
12. Frequency - *The American Heritage College Dictionary, Third Edition* defines frequency as the number times a specified measurement occurs in a sample. (Boston, New York: Houghton Mifflin Company, 2000) 545.
13. Histogram - *The American Heritage College Dictionary, Third Edition* defines a histogram as a bar graph of a frequency distribution in which the widths of the bars are proportional to the classes into which the variable has been divided and the heights of the bars are proportional to the class frequencies. (Boston, New York: Houghton Mifflin Company, 2000) 644. NOTE: Another way to visualize a histogram is a collection of data that is arranged on a coordinate plane with a vertical and a horizontal axis. A histogram shows frequency for a variable within equal intervals. The data on the horizontal axis usually represents the appropriate intervals for the kind of data it represents. (In other words, if the data represented the age of humans and there were intervals 150-160+, then there is a problem because no human has ever lived to that ripe old age.) The vertical axis usually shows the frequency of the data. A histogram looks like a bar graph; however, there are some differences between the two. In a histogram, BOTH axes represent numerical values. However, in a bar graph, either axis can be any variable (examples - dog, Lisa, TV, etc.) and have no numerical value. On a histogram, there is no space between the bars, but in a bar graph there is some separation between the bars. The horizontal axis is usually represented with equal intervals (example - # 0 - 9, 10 - 19, 20 - 29, etc.); however, in a bar graph there is only a single variable. A histogram can be misleading if there are too many or not enough bars used in the data set.
14. Box-and-whisker graph - A specific data illustration including spread of the data (minimum and maximum values), the median (Q2), lower quartile (Q1) (also known as the median of the lower half of all the values in the set), the upper quartile (Q3), (also known as the median of the upper half of all the values in the set), and the interquartile range (IQR = Q3 - Q1).
15. Minimum value - The lowest value in a data set.
16. Maximum value - The highest value in a data set.

## Vocabulary for Statistical Sleuths Continued

17. Mean - The average value of the data set. NOTE: HOW TO FIND THE MEAN: add the total of all of the values in set of data and divide by the number of values in the list (also known as the addends).
18. Median - The middle value in a data set arranged from least to greatest in value. NOTE: HOW TO FIND THE MEDIAN Given a set of data, arrange the data from least to greatest in value; find the middle number in the range (if there are an even number of data in the set, take both values and average them).
19. Mode - The value or values that occur the most often in a set of data. NOTE: It is possible to have several modes in a data set.
20. Range - Given a set of data, this value tells us the difference between the highest and the lowest value (extreme).
21. Measures of central tendency - Three measures of central tendency: mean, median, and mode. These terms are used to describe where the CENTER of the data TENDS to fall.
22. Quartiles - Divides a set of data into three sections: Q2 - the *median* of the entire data set; Q1 - the median of the lower half of the data; Q3 - the median of the upper half of the data. NOTE: HOW TO DETERMINE EACH QUARTILE Q2 (See #18-how to find the median for description); Q1 - identified by selecting the minimum value and the median of the entire data set and finding the median of that section; Q3 - identified by selecting the maximum value and the median of the entire data set and finding the median of that section.
23. Interquartile range (IQR) - The difference between the upper half of the data Q3 and the lower half of the data Q1:  $(IQR=Q3-Q1)$
24. Outlier - A value that does not appear to fit with the rest of the data set. An example would be if you were collecting data about the average height of men and in the data that you collected, you had only one individual who was 8 feet tall and the closest height to that was 7 feet 4 inches, then 8 feet would be considered an outlier because it is an example of an unusual case and is not representative of the data.