

## Chapter 1

**The  $1.5 \times IQR$  Rule for Outliers** Call an observation an outlier if it falls more than  $1.5 \times IQR$  above the third quartile or below the first quartile.

**Association** Occurs between two variables if specific values of one variable tend to occur in common with specific values of the other.

**Back-to-back stemplot** (*also called a back-to-back stem-and-leaf plot*) Used to compare the distribution of a quantitative variable for two groups. Each observation in both groups is separated into a **stem**, consisting of all but the final digit, and a **leaf**, the final digit. The stems are arranged in a vertical column with the smallest at the top. The values from one group are plotted on the left side of the stem and the values from the other group are plotted on the right side of the stem. Each leaf is written in the row next to its stem, with the leaves arranged in increasing order out from the stem.

**Bar graph** Used to display the distribution of a categorical variable or to compare the sizes of different quantities. The horizontal axis of a bar graph identifies the categories or quantities being compared. Drawn with blank spaces between the bars to separate the items being compared.

**Bimodal** Describes a graph of quantitative data with two clear peaks.

**Boxplot** A graph of the five-number summary. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.

**Categorical Variable** Places an individual into one of several groups or categories.

**Conditional distribution** Describes the values of one variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.

**Data analysis** A process of describing data using graphs and numerical summaries.

**Dotplot** A simple graph that shows each data value as a dot above its location on a number line.

**Distribution** Tells what values a variable takes and how often it takes these values.

**First quartile  $Q_1$**  If the observations in a data set are ordered from lowest to highest, the first quartile  $Q_1$  is the median of the observations whose position is to the left of the median.

*The Practice of Statistics for AP\*, 4<sup>th</sup> Edition Glossary*

**The Five-Number Summary** Consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum  $Q_1$   $M$   $Q_3$  Maximum

**Frequency table** Displays the count (*frequency*) of observations in each category or class.

**Histogram** Displays the distribution of a quantitative variable. The horizontal axis is marked in the units of measurement for the variable. The vertical axis contains the scale of counts or percents. Each bar in the graph represents an equal-width class. The base of the bar covers the class, and the bar height is the class frequency or relative frequency

**Individuals** Objects described by a set of data. Individuals may be people, animals, or things.

**Inference** Drawing conclusions that go beyond the data at hand.

**Interquartile range**  $IQR = Q_3 - Q_1$

**Marginal distribution** The marginal distribution of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

**Mean** The arithmetic average. To find the mean  $\bar{x}$  of a set of observations, add their values and divide by the number of observations.

**Median  $M$**  The midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution: 1. Arrange all observations in order of size, from smallest to largest. 2. If the number of observations  $n$  is odd, the median  $M$  is the center observation in the ordered list. 3. If the number of observations  $n$  is even, the median  $M$  is the average of the two center observations in the ordered list.

**Mode** The value or class in a statistical distribution having the greatest frequency.

**Multimodal** Describes a graph of quantitative data with more than two clear peaks.

**Outlier** An individual value that falls outside the overall pattern of a distribution.

**Overall pattern** In any graph of data, look for the overall pattern and for striking departures from that pattern. *Shape, center, and spread* describe the overall pattern of the distribution of a quantitative variable.

## *The Practice of Statistics for AP\*, 4<sup>th</sup> Edition Glossary*

**Pie chart** Shows the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percents for the categories. A pie chart must include all the categories that make up a whole.

**Quantitative Variable** Takes numerical values for which it makes sense to find an average.

**Range** The range of a set of quantitative data is the maximum value minus the minimum value.

**Relative frequency table** Shows the percents (*relative frequencies*) of observations in each category or class.

**Resistant measure** A statistic that is not affected very much by extreme observations.

**Roundoff error** The difference between the calculated approximation of a number and its exact mathematical value.

**Segmented bar graph** Used to compare the distribution of a categorical variable in each of several groups. For each group, there is a single bar with “segments” that correspond to the different values of the categorical variable. The height of each segment is determined by the percent of individuals in the group with that value. Each bar has a total height of 100%.

**Side by side bar graph** Used to compare the distribution of a categorical variable in each of several groups. For each value of the categorical variable, there is a bar corresponding to each group. The height of each bar is determined by the count or percent of individuals in the group with that value.

**Simpson’s paradox** An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined.

**Skewness** A distribution is **skewed to the right** if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side. It is **skewed to the left** if the left side of the graph is much longer than the right side.

**Splitting stems** A method for spreading out a stemplot that has too few stems.

**Standard deviation**  $s_x$  Measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. In symbols,  $s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$ .

**Stemplot** (*also called a stem-and-leaf plot*) A simple graphical display for fairly small data sets that gives a quick picture of the shape of a distribution while including the

*The Practice of Statistics for AP\*, 4<sup>th</sup> Edition Glossary*

actual numerical values in the graph. Each observation is separated into a **stem**, consisting of all but the final digit, and a **leaf**, the final digit. The stems are arranged in a vertical column with the smallest at the top. Each leaf is written in the row to the right of its stem, with the leaves arranged in increasing order out from the stem.

**Symmetry** If the right and left sides of a graph are approximately mirror images of each other.

**Third quartile  $Q_3$**  If the observations in a data set are ordered from lowest to highest, the third quartile  $Q_3$  is the median of the observations whose position is to the right of the median.

**Two-way table** A two-way table of counts organizes data about two categorical variables.

**Unimodal** Describes a graph of quantitative data with a single peak.

**Variables** Any characteristic of an individual. A variable can take different values for different individuals.

**Variance  $s_x^2$**  The average squared distance of the observations in a data set from their

mean. In symbols,  $s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$ .