12.1 Sampling, Frequency Distributions, and Graphs

**Statistics** is a method for collecting, organizing, analyzing, and interpreting data, as well as drawing conclusions based on the data. **Descriptive statistics** involves collecting, organizing, summarizing, and presenting data. **Inferential statistics** is concerned with making generalizations about, and drawing conclusions from, the data collected.

# Objective 1: Describe the population whose properties are to be analyzed

A **population** is the set containing all the people or objects whose properties are to be described and analyzed by the data collector. Since it is often not feasible to observe or question every element of the population, a sample is needed.

# Objective 2: Select an appropriate sampling technique

A **sample** is a subset or a subgroup of the population. A **random sample** is a sample obtained in such a way that every element in the population has an equal chance of being selected for the sample.

**RANDOM SAMPLING TECHNIQUE**

1. Identify each element in the population.
2. Assign numbers to each element in the population.
3. Randomly select numbers.
4. Assign the elements in the population who have those numbers to the sample set.

Surveys and polls involve data from a sample of some population. Regardless of the sampling technique used, the sample should exhibit characteristics typical of those possessed by the target population. This type of sample is called a **representative sample**.

# Objective 3: Organize and present data

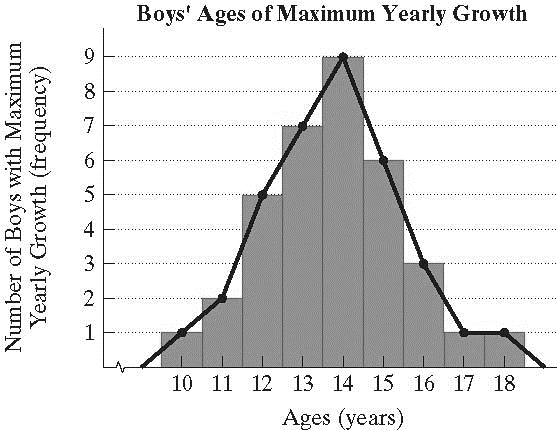
Collected data can be presented using a **frequency distribution**. A frequency distribution is a table with two columns. The data values are listed in one column and are generally listed from smallest to largest. The adjacent column is labeled **frequency** and indicates the number of times each value occurs.

When data values are arranged or grouped into **classes**, a **grouped frequency distribution** can be constructed. For example, test scores can be grouped into 10-point intervals. The leftmost number in each class of a grouped frequency distribution is called the **lower class limit.** The rightmost number in each class is called the **upper class limit**. The difference between any two consecutive lower (or upper) class limits is called the **class width**. With the possible exception of the first and last classes, all classes in a grouped frequency distribution must have the same width.

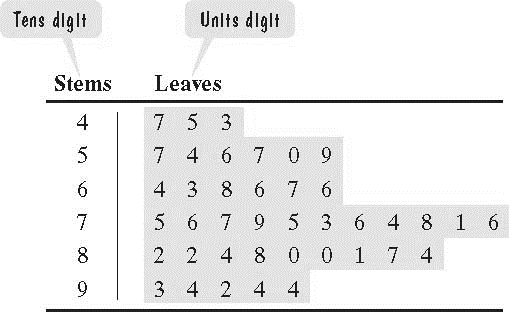
A **histogram** is a visual representation of a frequency chart. Each vertical bar represents a data value or class with the height of the bar indicating its frequency. The bars of a histogram touch each other because the values or classes on the *x*-axis are quantitative and consecutive. This is in contrast to bar graphs which often have categorical *x* values. A **frequency polygon** is a line graph formed by connecting the midpoints of the top of each bar of a histogram. The endpoints of the frequency polygon are always on the *x*-axis.

On the left below is a frequency distribution for a boy’s age of maximum yearly growth. The graph below on the right shows a histogram and a frequency polygon on the same axes for this frequency distribution. Given that children grow at different rates, the data represents the age at which the boys in the sample have their ‘growth spurt,’ or maximum growth over the year.

| **Age of**  **Maximum Growth** | **Number of Boys (Frequency)** |
| --- | --- |
| 10 | 1 |
| 11 | 2 |
| 12 | 5 |
| 13 | 7 |
| 14 | 9 |
| 15 | 6 |
| 16 | 3 |
| 17 | 1 |
| 18 | 1 |



A **stem-and-leaf plot** is another visual representation of a frequency distribution. It is constructed by separating each data item into a **leaf** which is usually the last digit of the number and a **stem** which is the first digit of a two-digit number. If the data values are greater than 99, the stems may have two or more digits. The stem-and-leaf plot gives a visual view of the data similar to that of a histogram. The plot below represents the statistics test scores for a group of 40 students.



# Objective 4: Identify deceptions in visual displays of data

Graphs can be used to clarify data but they can also create false impressions. These may be inadvertent or deliberate attempts by the presenter to lead or mislead.

**CONSIDERATIONS FOR USE OF VISUAL DISPLAYS OF DATA**

* The wording of titles and labels and the use of pictures within the graph design may distract from the actual data or bias its appearance.
* The scale of the vertical axis and the choice of intervals on the horizontal access may change the apparent significance of trends.
* The data may not be representative and the presenter may be biased. Consider the source of data and whether it represents an entire population or a sample. If it is a sample, ask how the sample is selected.

Notice the different impressions given by the graphs below.

Two line graphs for percentage of people in the United States living below the poverty level, 2000 through 2004. The x-axis is year and y-axis is poverty rate. In the first graph, the y-axis increases on a scale of 4 tenths of 1 percent.  Stretching the scale on the vertical axis creates an impression of a rapidly increasing poverty rate. In the second graph, the y-axis increases on a scale of 2 percent. Compressing the scale on the vertical axis creates an impression of a slowly increasing poverty rate.
