Chapter 2 - Summarizing and Graphing Data

Characteristics of Data:

1. Center: A representative or average value that indicates where the middle of the data set is located.

2. Variation: A measure of the amount that the data values vary.

3. Distribution: The nature or shape of the spread of the data over the range of values (such as bell-shaped, uniform, or skewed).

4. Outliers: Sample values that lie very far away from the vast majority of the other sample values.

5. Time: Changing characteristics of the data over time.

The Mnemonic is "Computer Viruses Destroy Or Terminate."

Frequency Distribution: (A list of Values vs Corresponding frequencies.) A frequency distribution (or frequency table) shows how a data set is partitioned among several categories (or classes) by listing all categories along with the number of data values in each category.

Pulse Rates (beats per minute) of Females and Males:

Females:

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64

Males:

68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

Lower class limits are the smallest numbers that can belong to the different classes. (has lower class limits of 60, 70, 80, 90, 100, 110, and 120.)

Upper-class limits are the largest numbers that can belong to the different classes. (has upper-class limits of 69, 79, 89, 99, 109, 119, 129.)

Class boundaries are the numbers used to separate the classes but without the gaps created by class limits. Figure 2-2 shows the gaps created by the class limits from Table 2-2. In Figure 2-2 we see that the values of 69.5, 79.5, . . . , and 119.5 are in the centres of those gaps. These are the class boundaries. Following the pattern established, we see that the lowest class boundary is 59.5, and the highest class boundary is 129.5. So, the complete list of class boundaries is 59.5, 69.5, 79.5, . . . , 119.5, 129.5.

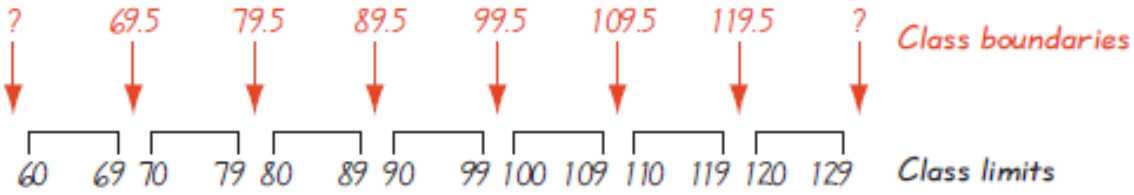


Figure 2-2 Finding Class Boundaries

Class midpoints are the values in the middle of the classes. (Table 2-2 has class midpoints of 64.5, 74.5, 84.5, 94.5, 104.5, 114.5, and 124.5.) Each class midpoint is found by adding the lower class limit to the upper-class limit and dividing the sum by 2.

The Class width is the difference between two consecutive lower class limits or two consecutive lower class boundaries in a frequency distribution. (Table 2-2 uses a class width of 10.)

Procedure for Constructing a Frequency Distribution:

We construct frequency distributions so that (1) large data sets can be summarized, (2) we can analyze the nature of data, and (3) we have a basis for constructing graphs (such as histograms, introduced in the next section). Although technology allows us to automatically generate frequency distributions, the steps for manually constructing them are as follows:

1. Determine the number of classes. The number of classes should be between 5 and 20, and the number you select might be affected by the convenience of using round numbers.

2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

Round this result to get a convenient number. (We usually round up.) If necessary, change the number of classes so that they use convenient values.

3. Choose either the minimum data value or a convenient value below the minimum data value as the first lower class limit.

4. Using the first lower class limit and the class width, list the other lower class limits. (Add the class width to the first lower class limit to get the second lower class limit. Add the class width to the second lower class limit to get the third lower class limit, and so on.)

5. List the lower-class limits in a vertical column and then enter the upper-class limits.

6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

When constructing a frequency distribution, be sure the classes do not overlap. Each of the original values must belong to exactly one class. Include all classes, even those with a frequency of zero. Try to use the same width for all classes, although it is sometimes impossible to avoid open-ended intervals, such as "65 years or older."

Relative Frequency Distribution: A variation of the basic frequency distribution is a relative frequency distribution. In a relative frequency distribution, the frequency of a class is replaced with a relative frequency (a proportion) or a percentage frequency (a per cent).

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Cumulative Frequency Distribution: The cumulative frequency for a class is the sum of the frequencies for that class and all previous classes.

Finding Frequency, Relative Frequency, Cumulative Frequency:

Females:

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64

Pulse Rate (Class Width)	Frequency	Relative Frequency	Cumulative Frequency
60 - 69	12	30%	12
70 - 79	14	35%	12 + 14 = 26
80 - 89	11	27.5%	26 + 11 = 37
90 - 99	1	2.5%	37 + 1 = 38
100 - 109	1	2.5%	38 + 1 = 39
110 - 119	0	0	39 + 0 = 39
120 - 129	1	2.5%	39 + 1 = 40

	Σ f = 40		

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Class Boundaries = 59.5, 69.5, 79.5, 89.5, 99.5, 109.5, 119.5, 129.5

Class MidPoints = 64.5, 74.5, 84.5, 94.5, 104.5, 114.5, and 124.5

Normal Distribution:

1. The frequencies start low, then increase to one or two high frequencies, then decrease to a low frequency.
2. The distribution is approximately symmetric, with frequencies preceding the maximum roughly a mirror image of those following the maximum.

Histograms: A histogram is a graph consisting of bars of equal width drawn adjacent to each other (without gaps). The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.

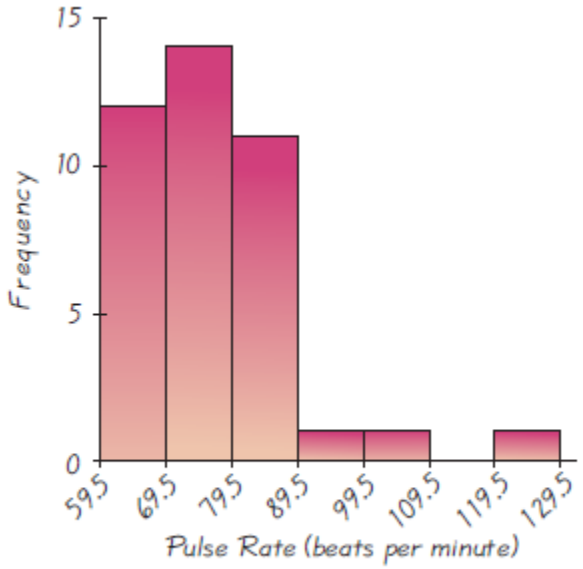


Figure 2-3 Histogram

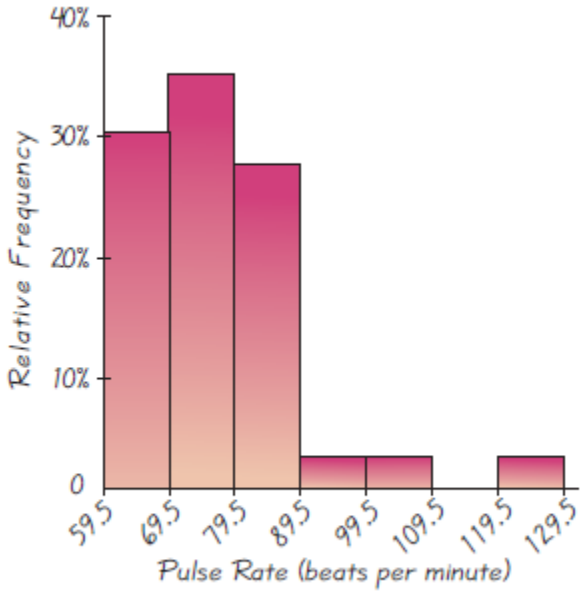


Figure 2-4 Relative Frequency Histogram

Males:

68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

Pulse Rate (Class Width)	Frequency	Relative Frequency	Cumulative Frequency
50 - 59	6	15%	6
60 - 69	17	42.5%	6 + 17 = 23
70 - 79	8	20%	23 + 8 = 31
80 - 89	8	20%	31 + 8 = 39
90 - 99	1	2.5%	39 + 1 = 40
100 - 109	0	0%	40 + 0 = 40
110 - 119	0	0%	40 + 0 = 40
120 - 129	0	0%	40 + 0 = 40

	Σ f = 40		

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Statistical Graphics:

Frequency Polygon: One type of statistical graph involves the class midpoints. A frequency polygon uses line segments connected to points located directly above class midpoint values. We construct a frequency polygon from a frequency distribution as shown in Example 1.

A variation of the basic frequency polygon is the relative frequency polygon, which uses relative frequencies (proportions or percentages) for the vertical scale. When trying to compare two data sets, it is often very helpful to graph two relative frequency polygons on the same axes.

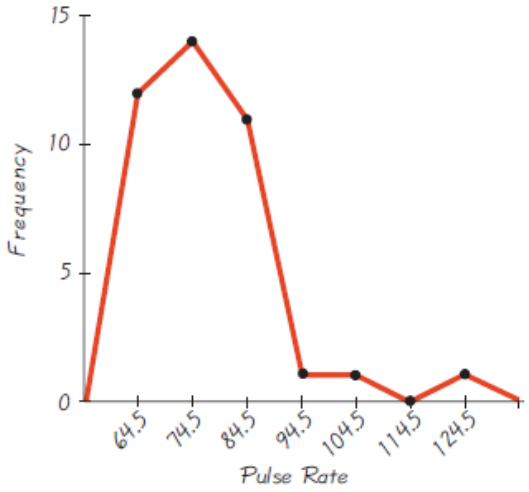


Figure 2-5 Frequency Polygon: Pulse Rates of Women

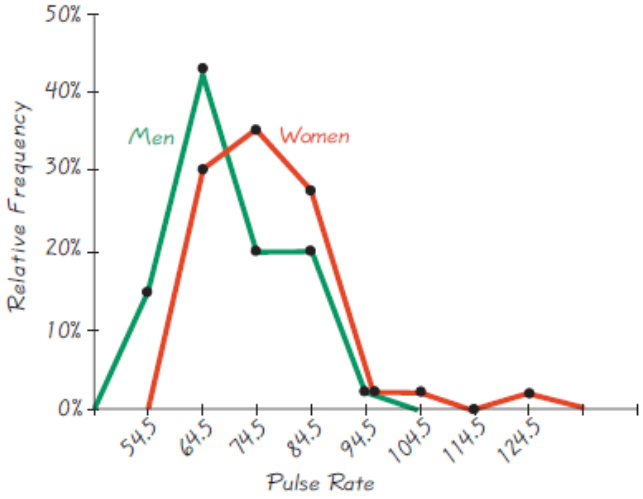


Figure 2-6 Relative Frequency Polygons: Pulse Rates of Women and Men

Ogive: Another type of statistical graph called an ogive (pronounced "oh-jive") involves cumulative frequencies. Ogives are useful for determining the number of values below some particular value, as illustrated in Example 3. An ogive is a line graph that depicts cumulative frequencies. An ogive uses class boundaries along the horizontal scale and cumulative frequencies along the vertical scale.

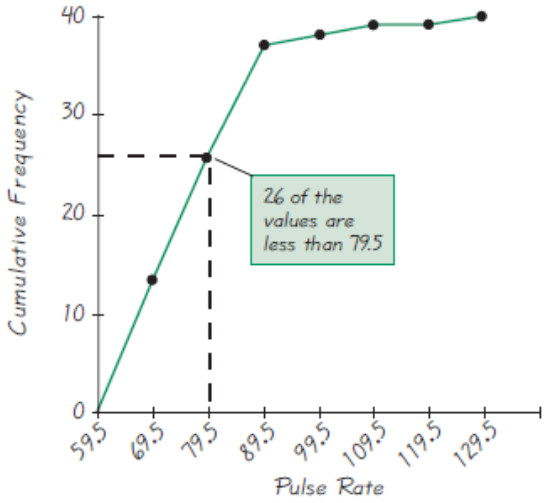
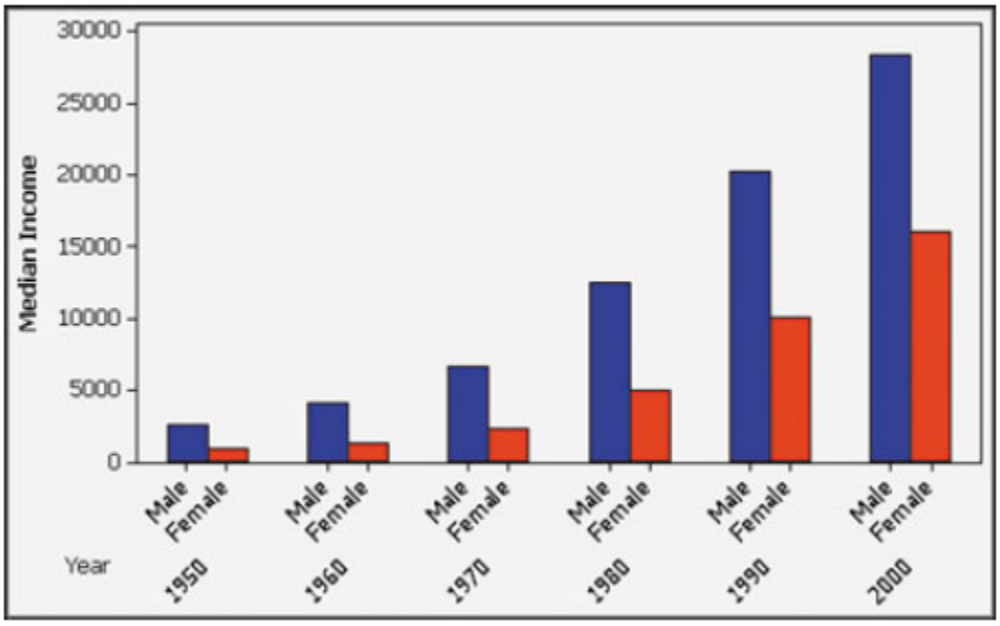
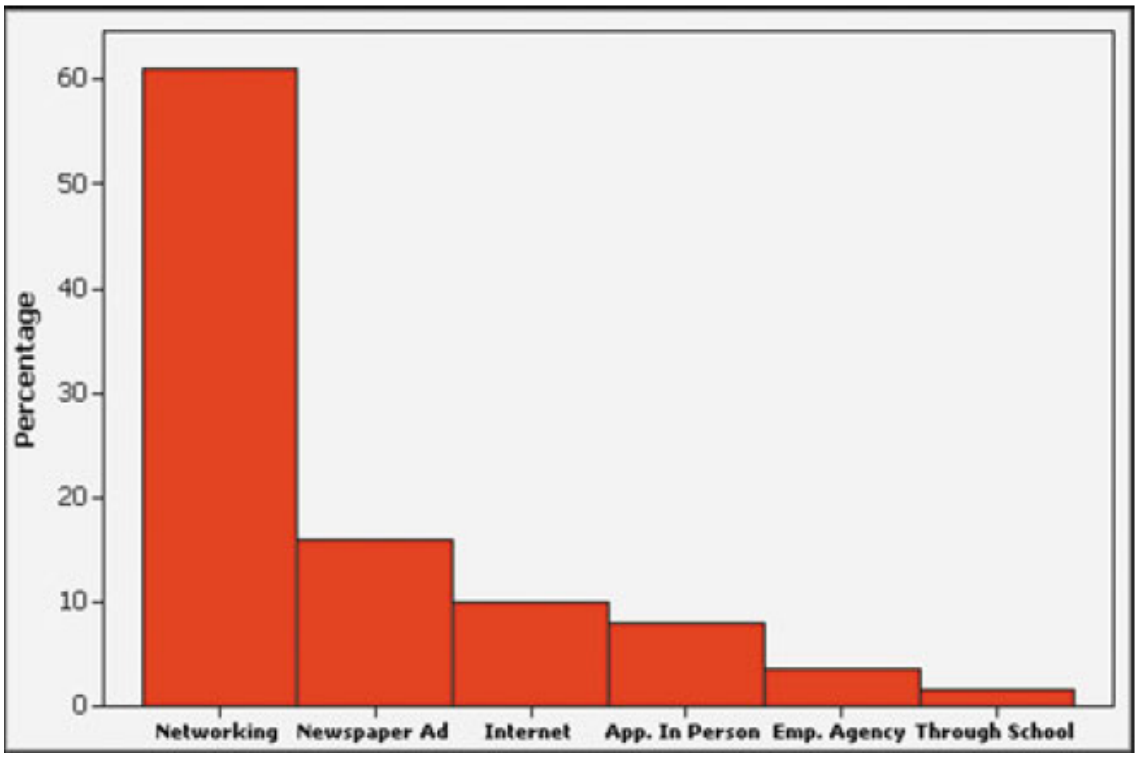


Figure 2-7 Ogive

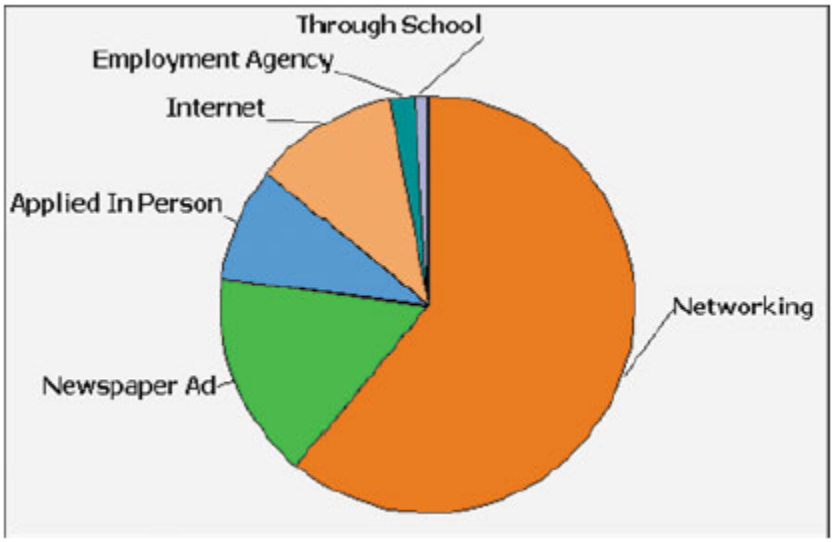
Bar Graphs: **A bar graph uses bars of equal width to show frequencies of categories of qualitative data. The vertical scale represents frequencies or relative frequencies. The horizontal scale identifies the different categories of qualitative data. The bars may or may not be separated by small gaps.**



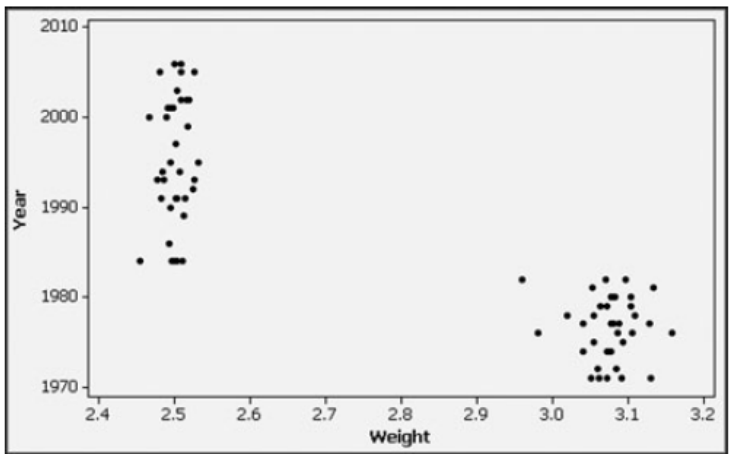
Pareto Charts: **When we want to draw attention to the more important categories, we can use a Pareto chart. A Pareto chart is a bar graph for qualitative data, with the added stipulation that the bars are arranged in descending order according to frequencies. The vertical scale in a Pareto chart represents frequencies or relative frequencies. The horizontal scale identifies the different categories of qualitative data. The bars decrease in height from left to right.**



Pie Charts: **A pie chart is a graph that depicts qualitative data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category.**



Scatterplots: **A scatterplot (or scatter diagram) is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first (x) variable, and the vertical axis is used for the second variable. The pattern of the plotted points is often helpful in determining whether there is a relationship between the two variables.**



Time-Series Graph: **A time-series graph is a graph of time-series data, which are quantitative data that have been collected at different points in time.**