# 2 – 1 Organizing Data

## Objective 1: Organize Data Using a Frequency Distribution

When we first collect data, it is \_\_\_\_\_\_\_ data, unorganized and rather “messy.”

### Example 2 -1 Ages of Wealthiest People

Suppose we collect the ages of the top fifty wealthiest people in the United States. ([Forbes Magazine](http://www.forbes.com/billionaires/list/#version:realtime), 08/27/2016) (With this link you can update this list over time.)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 60 | 80 | 85 | 52 | 32 | 72 | 76 | 74 | 80 | 76 |
| 93 | 43 | 43 | 67 | 68 | 71 | 66 | 62 | 77 | 88 |
| 83 | 60 | 51 | 59 | 78 | 98 | 86 | 88 | 76 | 80 |
| 44 | 59 | 65 | 68 | 54 | 51 | 51 | 76 | 59 | 63 |
| 67 | 52 | 77 | 61 | 73 | 50 | 61 | 62 | 50 | 61 |

A researcher wants to organize the dataset into a frequency distribution showing **classes** and **frequencies**. Each data value is in a category called a **class.** The **frequency** of a class is the number of data values in that specific class.

For quantitative values, like ages, the classes must 1) have **equal width**, 2) include values between the highest and lowest values (called the **range**), and 3) have values in, at least, the first and last class. Keep in mind that **classes do not overlap**.

Select the number of classes desired and the beginning value. Since 32 is the least data value, we will begin with 30 as the first **lower class limit**. The **range** is 98-32 = 66, so the class width for 6 classes must be larger than 66/6 = 11. We will use 12 as the class width.

Thus the first class begins at 30, the second begins at 30 + 12 = 42, the third begins at 54, the fourth at 66, the fifth begins at 78, and the sixth begins at 90. Notice that a seventh class would begin at 102, which is larger than any of the data.

| **Class limits** | Tally | Frequency |
| --- | --- | --- |
| 30 - 41 |  |  |
| 42 - 53 |  |  |
| 54 - 65 |  |  |
| 66 - 77 |  |  |
| 78 - 89 |  |  |
| 90 - 101 |  |  |

Classify each data value in a class using tally marks, then count the tally marks to find the frequency for each class.

The result can be shown using a frequency distribution with two columns, leaving out the tally marks.

### Organize Qualitative Data Using Categorical Frequency Distributions

Data that can be placed in categories, such as nominal and ordinal data, can be organized using a frequency distribution.

### Example 2 -2 Blood Type

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

A B B AB O

O O B AB B

B B O A O

A O O O AB

AB A O B A

The data are categorical, so discrete classes, one for each blood group, can be used.

**Step 1** Make a table showing class, frequency and percentage

**Step 2** Count the frequency of each category. (Use tally marks, if needed)

**Step 3** Find the percentage of values in each class by dividing the frequency of each class by the total number of values.

**Step 4** Find the totals for each column.

| **Class: Blood Type** | **Frequency** | **Percent** |
| --- | --- | --- |
| A | 5 | 20% |
| B | 7 | 28% |
| O | 9 | 36% |
| AB | 4 | 16% |
| **Totals** | 25 | 100% |

### Grouped Frequency Distributions

When the \_\_\_\_\_\_\_\_\_\_\_ of quantitative data is large, the data must be grouped into classes.

Decide how many classes to use. There should be between 5 and 20 classes. The classes must not overlap.

The **lower class limit** is the smallest data value that can be included in class: thus, the first lower class limit must be no larger than the smallest value in the data set.

Find the \_\_\_\_\_\_\_\_\_\_\_ class limit for each class. This is the largest value that can be included in each class.

**Class boundaries** are the numbers used to separate **classes**. The size of the gap between **classes** is the difference between the upper **class** limit of one **class** and the lower **class** limit of the next **class**. The class boundary is the midpoint of the upper class limit of one class and the lower class limit of the next class.

If the data values are whole numbers, find the boundaries by subtracting 0.5 from the lower class limit and adding 0.5 to the upper class limit.

Lower class limit – 0.5 = lower class boundary

Upper class limit + 0.5 = upper class boundary

If the data values are in tenths, subtract 0.05 (half of the difference between tenths) from the lower class limit and add 0.05 to the upper class limit.

The **class width** is the difference between the lower class limit of one class and the lower class limit of the next class. (It is preferable that the class width be an odd number.)

The **class midpoint** is the average of the \_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ class limits (or the upper and lower class boundaries). The midpoint is the numeric location of the \_\_\_\_\_\_\_ of the class. For a class with a lower limit of 24 and an upper limit of 30, the midpoint is .

### To Construct a Grouped Frequency Distribution, Follow These Rules:

1. There should be between 5 and \_\_\_\_\_ classes.
2. It is preferable, but not necessary, to have a class width that is an\_\_\_\_\_\_ number.
3. Classes do not \_\_\_\_\_\_\_\_\_\_ so no data value can be in two classes.
4. Classes are continuous.
5. There must be enough classes to accommodate \_\_\_\_\_\_ data values.
6. Class widths are \_\_\_\_\_\_\_\_.

### Example 2 – 3 Salaries of College and University Coaches

The data are salaries (in hundred thousands of dollars) of a sample of 30 college and university coaches in the United States. Construct a frequency distribution for the data using 8 classes.

164 225 225 140 188

210 238 146 201 544

550 188 415 261 164

478 684 330 307 435

857 183 381 275 578

450 385 297 390 515

Find the range: Range = Highest value – Lowest value = \_\_\_\_\_\_\_

Determine the class width: Range/8 = \_\_\_\_\_\_, round up to nearest whole number.

Find the class limits.

Find the class boundaries.

Find the midpoint of each class.

Find the frequencies for each class.

Compute cumulative frequencies, showing the frequency of values less than or equal to the upper class boundary of a specific class.

Compute relative frequency, the quotient of the class frequency and the number of data values, of each class. (List the result in decimal form.)

Compute the cumulative relative frequency, the quotient of the cumulative frequency of the specific class and the total number of data values, for each class.

| **Class Limit** | **Class Boundary** | **Class Midpoint** | **Frequency** | **Cumulative Frequency** | **Relative Frequency** | **Cum. Rel. Frequency** |
| --- | --- | --- | --- | --- | --- | --- |
| 140-229 | 139.5-229.5 | 184.5 | 12 | 12 | .375 | .375 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  | 30 |  | 1 |
| **Totals** | | | **30** |  | **1** |  |

# 2-2 Histograms, Frequency Polygons, and Ogives

## Objective 2. Represent Data in Frequency Distributions Graphically, Using Histograms, Frequency Polygons, and Ogives.

When data are organized into a frequency distribution, the summary can be presented in graphical form. The most commonly used graphs are:

Histogram – graph that displays data using contiguous \_\_\_\_\_\_\_\_\_\_\_\_ bars,

unless the class frequency is 0, of heights representing the   
 frequencies of the classes.

Frequency Polygon – graph that displays data using lines that connect points

plotted at heights representing class frequencies at the midpoints of

the classes.

Cumulative Frequency Graph, or \_\_\_\_\_\_\_\_\_\_\_\_\_ - graph that represents

cumulative frequencies for the classes in a frequency distribution.

Construct a histogram, frequency polygon, and ogive for the frequency distribution representing salaries (in hundred thousands of dollars) of a sample of 30 colleges and university coaches in the United States.

| **Class Limit** | **Class Boundary** | **Class Midpoint** | **Frequency** | **Cumulative Frequency** | **Relative Frequency** | **Cum. Rel. Frequency** |
| --- | --- | --- | --- | --- | --- | --- |
| 140 - 229 | 139.5-229.5 | 184.5 | 11 | 11 | .37 | .37 |
| 230 – 319 | 229.5 – 319.5 | 274.5 | 5 | 16 | .17 | .53 |
| 320 – 409 | 319.5 – 409.5 | 364.5 | 4 | 20 | .13 | .67 |
| 410 – 499 | 409.5 – 499.5 | 454.5 | 4 | 24 | .13 | .80 |
| 500 – 589 | 499.5 – 589.5 | 544.5 | 4 | 28 | .13 | .93 |
| 590 – 679 | 589.5 – 679.5 | 634.5 | 0 | 28 | .00 | .93 |
| 680 - 769 | 679.5 – 769.5 | 724.5 | 1 | 29 | .03 | .98 |
| 770 - 859 | 769.5 – 859.5 | 814.5 | 1 | 30 | .03 | 1.00 |
| **Totals** | | | **30** |  | **1** |  |

### Example 2 – 4 Construct a Histogram

**Step 1** Draw the \_\_\_\_\_\_\_. The x axis is the horizontal axis. The y axis is the vertical axis.

**Step 2** \_\_\_\_\_\_\_\_\_\_\_\_ the axes. The x axis is the horizontal axis labeled with lower class boundaries. The y axis is the vertical axis labeled with values for frequencies.

**Step 3** Using the frequencies as heights, \_\_\_\_\_\_\_\_\_\_ vertical bars for each class.

The histogram corresponds to the frequency chart of coaches salaries.  Each bar is labeled with the class midpoint on the horizontal axis and the class frequency on the vertical axis.

### Example 2 – 5 Construct a Frequency Polygon

**Step 1** Find the midpoints of each class.

**Step 2** \_\_\_\_\_\_\_\_ the x and y axes. The x axis is labeled with class \_\_\_\_\_\_\_\_\_\_.   
 The y axis is t labeled with class \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Step 3** At each class midpoint, plot points at the height representing \_\_\_\_\_\_\_\_\_\_\_\_\_\_ for each class .

This is a line graph of the freqency for each class.  To make the polygon, the class before the data begins and the class following the data must be included with zero frequencies.

### Example 2 – 6 Construct an Ogive

**Step 1** Find the cumulative \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_of each class.

|  | **Cumulative Frequency** |
| --- | --- |
| Less than 229.5 |  |
| Less than 319.5 |  |
| Less than 409.5 |  |
| Less than 499.5 |  |
| Less than 589.5 |  |
| Less than 679.5 |  |
| Less than 769.5 |  |
| Less than 859.5 |  |

**Step 2** \_\_\_\_\_\_\_\_ the x and y axes. The x axis is labeled with class \_\_\_\_\_\_\_\_\_\_\_\_.   
The y axis is labeled with a scale appropriate for the cumulative \_\_\_\_\_\_\_\_\_\_\_\_\_.

**Step 3** At each upper class boundary, plot the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ frequency points.

**Step 4** Starting with the first upper class boundary, connect adjacent points with line segments.

A **relative frequency graph** uses the relative frequencies, or appropriate scale, instead of the frequencies. The graphs look the same, only the scale on the y-axis is different.

### Using Histograms to Describe the Shape of Distributions

Although distributions are not often perfectly shaped, identify an overall pattern:

| **Bell Shaped** |  | **Uniform** |
| --- | --- | --- |
| Graph shows a symmetric histogram with a bell shape. |  | Graph shows a symmetric histogram with a uniform shape. |
|  |  |  |

| **J-Shaped** |  | **Reverse J-Shaped** |
| --- | --- | --- |
| Graph shows a symmetric histogram with a J shape. |  | Graph shows a symmetric histogram with a reverse J shape. |
|  |  |  |
| **Right-Skewed** |  | **Left-Skewed** |
| Graph shows a symmetric histogram with a right skewed shape. |  | Graph shows a symmetric histogram with a left skewed shape. |
|  |  |  |
| **Bimodal** |  | **U-Shaped** |
| Graph shows a symmetric histogram with a bi-modal shape. |  | Graph shows a symmetric histogram with a U  shape. |

## 2-3 Other Types of Graphs

## Objective 3. Represent Data Using Bar Graphs, Pareto Charts, Time Series Graphs, Pie Graphs and Dotplots.

### Bar Graph

Agraph that represents data by using \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ or \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ bars with heights or lengths representing \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

1. Draw and label the x and y axes so that one represents the group or class and the other represents the frequency.
2. Draw bars corresponding to the frequencies.

### Example 2 – 7 Sales for Fast-Food Franchises

The data gives the worldwide sales, in billions of dollars, for fast-food franchises for a specific year.

Sales in a specific year for each of f9ve fast food franchises.Wendy''s ($8.7), KFC ($14.2), Pizza Hut ($9.3), Burger King ($12.7), and Subway ($10.0).

Bar graph representing the frequency distribution for a categorical variable with vertical bars, labeled with category names. or Bar graph representing the frequency distribution for a categorical variable with hoeizontal bars, labeled with category names.

### Pareto Chart

A special version of a bar graph that represents the frequency distribution for a

categorical variable. The frequencies are displayed by heights of vertical bars, arranged

from \_\_\_\_\_\_\_\_\_\_\_\_\_ to \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Example 2-8 Construct a Pareto Chart

To create a Pareto chart, use vertical bars and put the categories in order from the highest frequency to the lowest frequency.

Use vertical bars

Make the bars the same \_\_\_\_\_\_\_\_\_\_\_\_.

Arrange the data from the largest to the smallest, by frequency.

Units used for the frequency scale must be equal in size.

Construct the Pareto chart for Worldwide Sales in Example 2-7. Include category labels and the scale for the frequencies:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

### Time Series Graph

A graph that represents data that occur over a specific period of time.

**Step 1** Draw and label the x and y axes.

**Step 2** Label the x axis for years and the y axis for the variable over time.

**Step 3** Plot each point according to the table.

**Step 4** Draw the line segments (not curves) connecting adjacent points.

### Example 2-9. Percent of Cigarette Smokers

The data show the percentage of U.S. adults who smoke. Draw and analyze a time series graph for the data.

| **Year** | 1970 | 1980 | 1990 | 2000 | 2010 |
| --- | --- | --- | --- | --- | --- |
| **Percent** | 37 | 33 | 25 | 23 | 19 |

There has been a decrease in the percentage of U.S. adults who smoke over a 40-year period.

### Example 2-10. Amount Spent on Valentine’s Day

The data show the average amount of money spent by consumers on Valentine’s Day

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
| **Amount** | $120 | $123 | $103 | $103 | $110 | $126 |

Source: National Retail Foundation

Draw a time series graph for the data and comment on the trend.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

### Pie Charts

A circle that is divided into sections or wedges according to the \_\_\_\_\_\_\_\_\_\_ of frequencies in each category of the distribution.

**Step 1** Divide the frequency for each class to a proportional part of the circle.

Degrees = 0, *f* = frequency, *n* = sum of the frequencies

**Step 2** Convert each frequency to a percentage

% =

**Step 3** Using a protractor and compass, draw the graph with angles measuring the degree values found in Step 1 and label each section with the class name and percentages.

### Example 2 – 11. Reasons We Travel

Use the data to determine the degrees and percentages needed to construct a pie graph for Reasons We Travel data. Construct the pie chart.

**Reasons We Travel**

| **Purpose** | **Frequency** | **Degrees** | **Percentage** |
| --- | --- | --- | --- |
| *Personal business* | 146 |  |  |
| *Visit friends or relatives* | 130 |  |  |
| *Work-related* | 225 |  |  |
| *Leisure* | 299 |  |  |
| Source: *USA TODAY*. | | | |

### Dotplots

A \_\_\_\_\_\_\_\_\_\_\_\_ is a statistical graph in which each data value is plotted as a point (dot) above the horizontal axis.

**Step 1** Find the lowest and highest data values, and decide what scale to use on the horizontal axis.

**Step 2** Draw a horizontal line, and draw the scale on the line.

**Step 3** Plot each data value above the line. If the value occurs more than once, plot the other point(s) above the first point.

### Example 2-12. Teacher Strikes

In Pennsylvania the numbers of teacher strikes for the last 14 years are shown. Construct a dotplot for the data.

9 13 15 7 7 14 9 10 14 18 7

8 8 3

*Solution:*

**Step 1** Find the lowest and highest data values, and decide what scale to use on the horizontal axis.

The lowest value is 3 and the highest value is 18.

**Step 2** Draw a horizontal line, and draw the scale on the line.

The scale will be ones from 0 to 20.

**Step 3** Plot each data value above the line. If the value occurs more than once, plot the other point(s) above the first point.

The dot plot shows a horizontal axis numbered from 0 through 20.  Above 3 there is one dot. Three dots are above 7. Two dots are above 8. Two dots are above 9. One dot is above 10.  One dot is above 13.  Two dots are above 14.  One dot is above 15. One dot is above 18.  

## Stem and Leaf Plot

## Objective 4. Draw and Interpret a Stem and Leaf Plot.

A data plot that uses part of the data value as the stem (the leftmost digit or digits) and part of the data as the leaf (usually the last digit) to form groups or classes. The plot shows the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the data while preserving the detail.

**Step 1** Arrange the data in order.

**Step 2** Separate the data according to the first digit.

**Step 3** Construct the display by using the leading digit(s) as the stem and the trailing digit as the leaf.

### Example 2 – 13. Out-Patient Cardiograms

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25 31 20 32 13 14 43 02 57 23

36 32 33 32 44 32 52 44 51 45

*Solution:*

Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32, 32, 33, 36, 43, 44, 44, 45, 51, 52, 57

*Note:* Arranging the data in order is not essential and can be cumbersome when the data set is large; however, it is helpful in constructing a stem and leaf plot. The leaves in the final stem and leaf plot should be arranged in order.

The chart shows the leading or tens digit  in the "stem" column and the trailing digit or ones digit in the "leaf" column.  The data are in order.  The plot retains the detail of the data set while also showing the shape of the data.  

### Example 2 – 14. Calories per 1 Ounce of Salad Dressing

A listing of calories per 1 ounce of selected salad dressings (not fat-free) is given below. Construct a stem and leaf plot for the data.

100 130 130 130 110 110 120 130 140 100 140 170 160 130 160 120 150 100 145 145 145 115 120 100 120 160 140 120 180 100 160 120 140 150 180 160

**Step 1** Arrange the data in order.

**Step 2** Determine the stems (the leading two digits of each three digit number) and leaves (the last digit of each number).

The stems are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

**Step 3** Construct the stem and leaf plot.

| **Stems** | **Leaves** |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Look for peaks and gaps in the distribution. Is the distribution symmetric? Skewed? What does the graph show about variability?

**Back to Back Stem and Leaf Plots –** Two stem and leaf plots used to compare \_\_\_\_\_ data sets, so that the leaves for one set of data are on one side of the stem and leaves for the other are on the other side. Each set of leaves are arranged in numerical order from the stem.

### Example 2 – 15 Number of Stories in Tall Buildings

The number of stores in two selected samples of tall buildings in Atlanta and Philadelphia is shown. Construct a back-to-back stem and leaf plot, and compare the distributions.

**Atlanta**

55 70 44 36 40 63 40 44 34 38

47 52 32 32 50 53 32 28 31 52

34 32 50 26 29 60 32

**Philadelphia**

61 40 38 32 30 58 40 40 25 30

54 40 36 30 30 53 39 36 34 33

50 38 36 39 32

*Solution*:

The range of the data from Atlanta is from 26 through 70. The range of the data from Philadelphia is from 25 through 61. The stems needed are 2, 3, 4, 5, 6, and 7. Once the data is ordered, arrange each data set in order from smallest to largest from the stem out. In this case, the Atlanta data is on the left, so the smaller values are right justified next to the stems and the Philadelphia data, which is on the right, is left justified next to the stems.

This stem and leaf plot shows two data sets using the same stems.  The data from Atlanta is on the left, arranged smallest to largest beside the appropriate stem from right to left.  The data from Philadelphia is on the right, arranged from smallest to largest beside the anppropirate stem from right to left.

### Misleading Graphs

Graphs are a visual \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ enabling readers to analyze and interpret data more easily than they could by looking at the numbers.

Inappropriately drawn graphs can misrepresent the data and guide readers to incorrect conclusions.

The scale of the vertical axis should not be truncated unless the reader is made aware of the change.

### Example 2–16 Changing the Scale on the Vertical Axis to Magnify the Differences

This bar graph is misleading because the differences between the categories are magnified by using a vertical scale from 95% to 100%. This bar graph shows that the differences are not as large as it appears in the misleading graph by using the scale of 0% to 100% on the vertical scale.

Using scale from 95% to 100% Using scale from 0% to 100%

### Example 2–17 Changing the Scale on the Vertical Axis to Show a Small Difference

The first graph in this series is scaled correctly, but the differences are so small that teh differences barely show up. The second graph in this series is scaled from 20% to 25%, but it allows the differences to show.  The importance is that the scale is noted so that the reader is aware of the change in scale.

Using scale from 0% to 100% Using scale from 20% to 25%

Graphs should always be labeled on both the x and y axes so the magnitude of the differences is apparent to the reader.

One-dimensional increases or decreases should be illustrated using one dimensional means instead of exaggerating the differences by using two-dimensional pictures.

### Example 2–18 Differences Shown With a One Dimensional Graph and a Two Dimensional Graph

For a one dimensional graph, difference o fa one dimensional variable is shown correctly if the scale begins at zero.  In a two dimensional graph, the area makes the differences appear larger.  The value of the square of the smaller value of the variable is dwarfed by the square of the larger value of the variable.

The source for the information presented should be included so the reader can analyze or determine the reliability of the organization presenting the data.

### Example 2 – 19 Comparison without Vertical Scale

### .

Four economic factors for four areas are compared, but no vertical scale is given, so the magnitude of the differences is not apparent.