General Education Statistics
Class Notes
Organizing Quantitative Data (Section 2.2)

We will see histograms, stem-and-leaf plots, dot plots, and time-series graphs. We will also explore the shapes of these graphs.

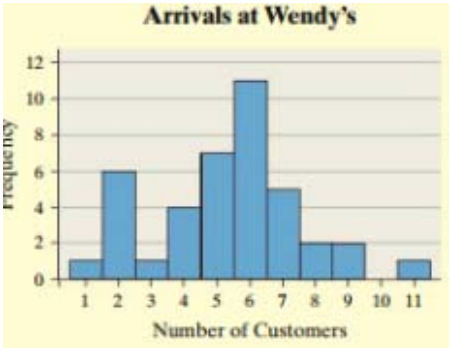
We will see many different kinds of displays here meant to summarize the data. We will use a worksheet to cover many types of displays. We will learn how to read them as well.

Discrete or Continuous Data:
The first step in summarizing quantitative data is to determine whether the data are discrete or continuous.

If the data are discrete and there are relatively few different values of the variable, the categories of data (**classes**) will be the observations (as in qualitative data).

If the data are discrete, but there are many different values of the variables, or if the data are continuous, the categories of data (the **classes**) must be created using intervals of numbers.

Definition: A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and *the rectangles touch each other*.



This data was obtained by counting the number of customers arriving during 40 randomly selected 15-minute intervals of time during lunch at a Wendy's restaurant.

The histogram shows us the **distribution** of a data set -- more on that later.

Definition: A **stem-and-leaf plot** or **stemplot** breaks the numbers up by using some digits to form the **stem.** Other digits (usually the last ones) form the **leaves**.

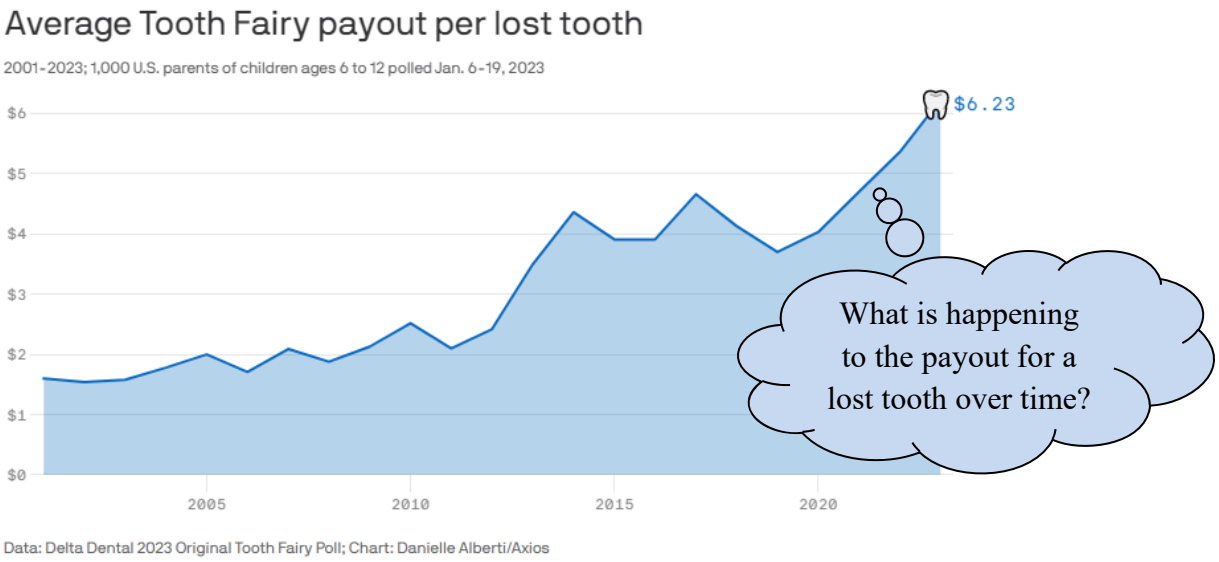
For example, a data value of 147 may have 14 as the stem and 7 as the leaf. Another example using the number 15.6 may use 15 as the stem and 6 as the leaf. Include a legend to explain how the data is read.

This makes little sense until you make and read them, which we will on this section's worksheet.

A stemplot will show the distribution of the data, much like a histogram. However, stemplots do this while recording every single observation rather than grouping them.

Definition: Time series data: If the value of a variable is measured at different points in time, the data are referred to as **time series data**.

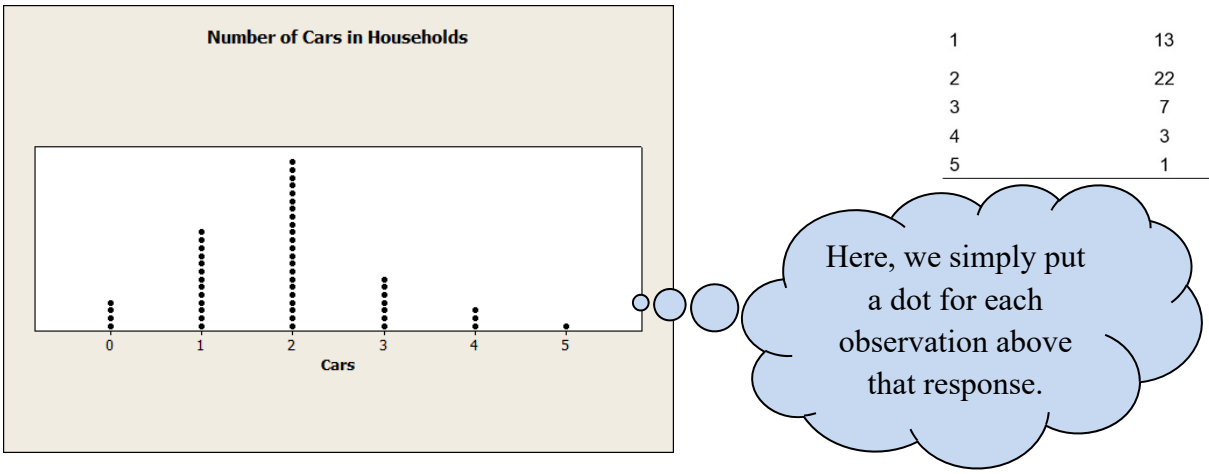
A **time-series graph** is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis. Line segments are then drawn connecting the points. The term **line graph** is also used to describe this graph.



Definition: A **dot plot** is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed. It will look similar to a bar graph.

Suppose we sampled 50 households to ask how many cars they had and obtained the data here. Below is a dot plot of this sample data.

# of Cars	Frequency
0	4
1	13
2	22
3	7
4	3
5	1

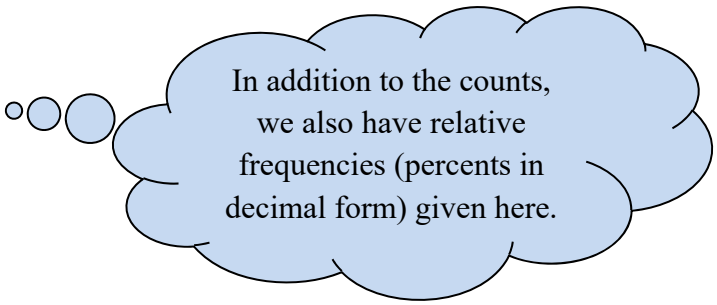


This seems like a pain in the neck (to read) but it does give us an idea of the shape of the **distribution** of data…

EXAMPLE Drawing a Histogram for Discrete Data

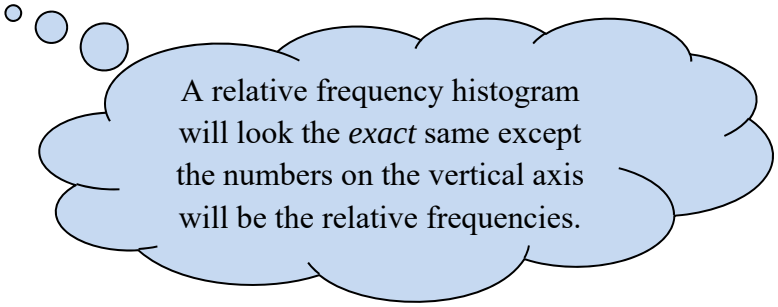
Again, suppose we asked 50 households how many cars they had. We will draw a frequency histogram for this "number of cars per household" data.

# of Cars	Frequency	Relative Frequency
0	4	$\frac{4}{50} = 0.08$
1	13	$\frac{13}{50} = 0.26$
2	22	0.44
3	7	0.14
4	3	0.06
5	1	0.02



expl 1: Histograms for discrete data are like bar graphs but the bars *will* touch each other. Also, we will draw each bar so that it is centered above the value on the axis. Because of this, we place 0 cars to the right of the vertical axis, *not* where you would expect it on the Cartesian plane.

Let's draw this histogram. Start with two labeled axes and a title. Your horizontal axis should be "Number of cars" with "Frequency" on the vertical axis.

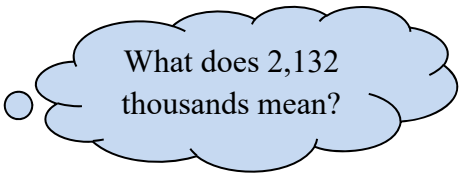


Displaying Continuous Data:

Definitions: The **lower class limit** of a class is the smallest value within the class while the **upper class limit** of a class is the largest value within the class. For the example below, the lower class limit of the first class is 25. The lower class limit of the second class is 35. The upper class limit of the first class is 34. Notice the classes do *not* overlap but encompass all possible values.

The **class width** is the difference between consecutive lower class limits. The class width of the data is 35 − 25 = 10.

Number of Persons Aged 25-64 Who Are Currently Work-disabled	
Age	**Number (in thousands)**
25-34	2,132
35-44	3,928
45-54	4,532
55-64	5,108



You may have to determine the classes for your data. Here are some guidelines.

1. Determine how many classes you want. This amounts to how many bars your histogram will have. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer classes you should have.

2. Determine the class width by computing the following and then rounding up.

$$\text{Class width} = \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}$$

3. Choose the Lower Class Limit of the First Class. You should choose the smallest observation in the data set or a convenient number slightly lower than the smallest observation.

4. Define the Lower Class Limits of the Subsequent Classes starting with the Lower Class Limit of the First Class and adding the Class Width. Do this repeatedly until you have the Lower Class Limits for all Classes.

Definition: Data table: A **data table** is an arrangement of data into rows and columns as you see above. A title, row and/or column labels, and qualifiers like “numbers are in thousands” are important.

Worksheet: Data Displays:
This worksheet will review pie charts and bar graphs. It will then lead us through line graphs (also called time-series graphs), histograms, stem-and-leaf plots, and data tables. We will also explore adjusting the scale of a line graph to change the meaning of it. We will briefly look at the concepts of outlier and the distribution of the data.

Shape of the Distribution:

There are four main types of shapes we will discuss. They are

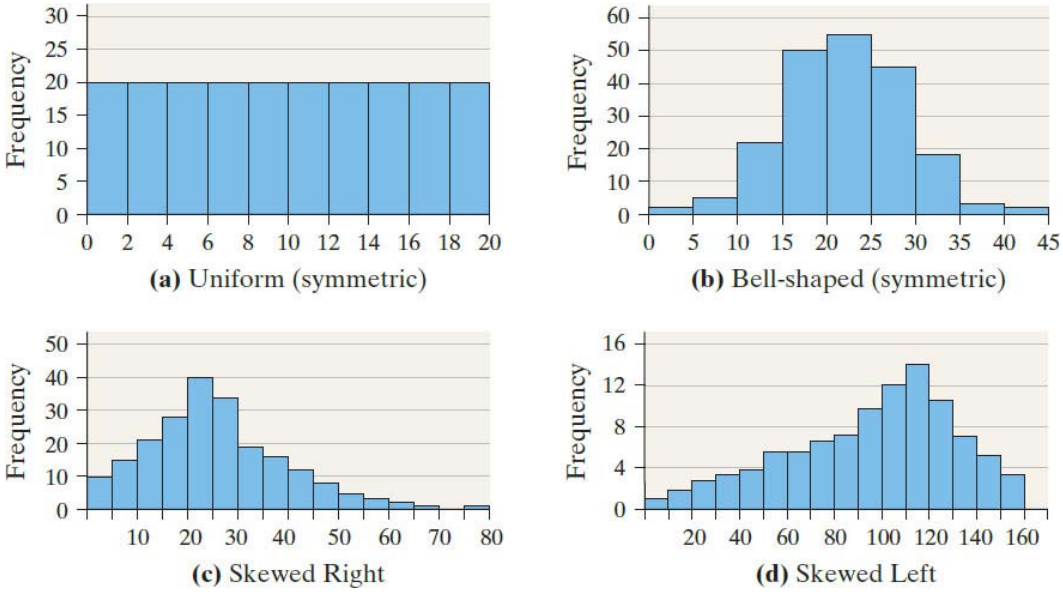
1. **Uniform distribution:** the frequency of each value of the variable is evenly spread out across the values of the variable,

2. **Bell-shaped (symmetric) distribution:** the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle,

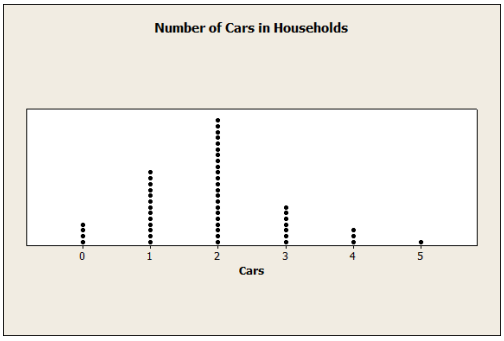
3. **Skewed right:** the tail to the right of the peak is longer than the tail to the left of the peak,

4. **Skewed left:** the tail to the left of the peak is longer than the tail to the right of the peak.

To understand this further, consider the histograms of four different distributions.



expl 2a: Consider the dot plot shown here. Which shape do you think it has?



expl 2b: Consider the histogram shown here. Which shape do you think it has?