

We will explore tables, bar graphs, pie charts, and more.

Statistics

Class Notes

Chapter 2: Data Displays (Sections 2.1-2.4)

Organizing Qualitative Data (Section 2.1)

If we have raw data but cannot display it accurately and succinctly, it does us no good. We will explore these common ways for displaying qualitative data. Quantitative data will be dealt with later.

Definition: A **frequency distribution** lists each category of data and the number of occurrences for each category. Imagine a bag of M&Ms and you count how many candies of each color.

Well, now that's well and good. But what if we wanted to show what proportion or percentage of the bag was red or blue M&Ms? We could construct a **relative frequency distribution**.

Definition: The **relative frequency** is the proportion (or percent) of observations within a category and is found using the following formula.

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

$$\text{percent} = \frac{\text{part}}{\text{whole}}$$

Definition: A **relative frequency distribution** lists each category of data alongside their relative frequencies.

expl 1: Let's create a relative frequency table. Here is the frequency distribution of a bag of M&Ms. Complete the table. Round to three decimal places.

Color	Frequency	Relative Frequency	Percentage
brown	12		
yellow	10		
red	9		
orange	6		
blue	3		
green	5		
	Total = 45		

What should these relative frequencies add up to? Do they?

Recall: Percentages:

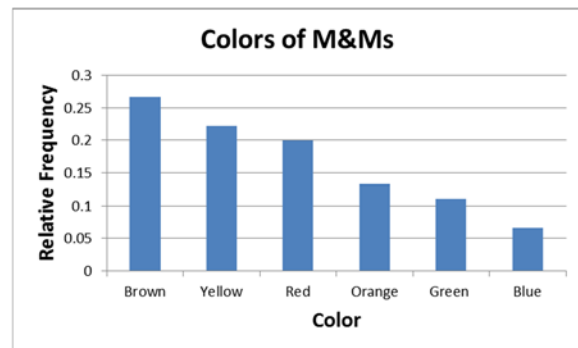
Do you remember how to find percents? How do you convert these numbers to percent form? When you do, remember to include the percent symbol (%).

Definition: A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis, with the frequency (or relative frequency) of the category on the other axis. The height of each rectangle represents the category's frequency (or relative frequency).

Rectangles of equal width are drawn for each category.

Rectangles in bar graphs should *not* touch each other.

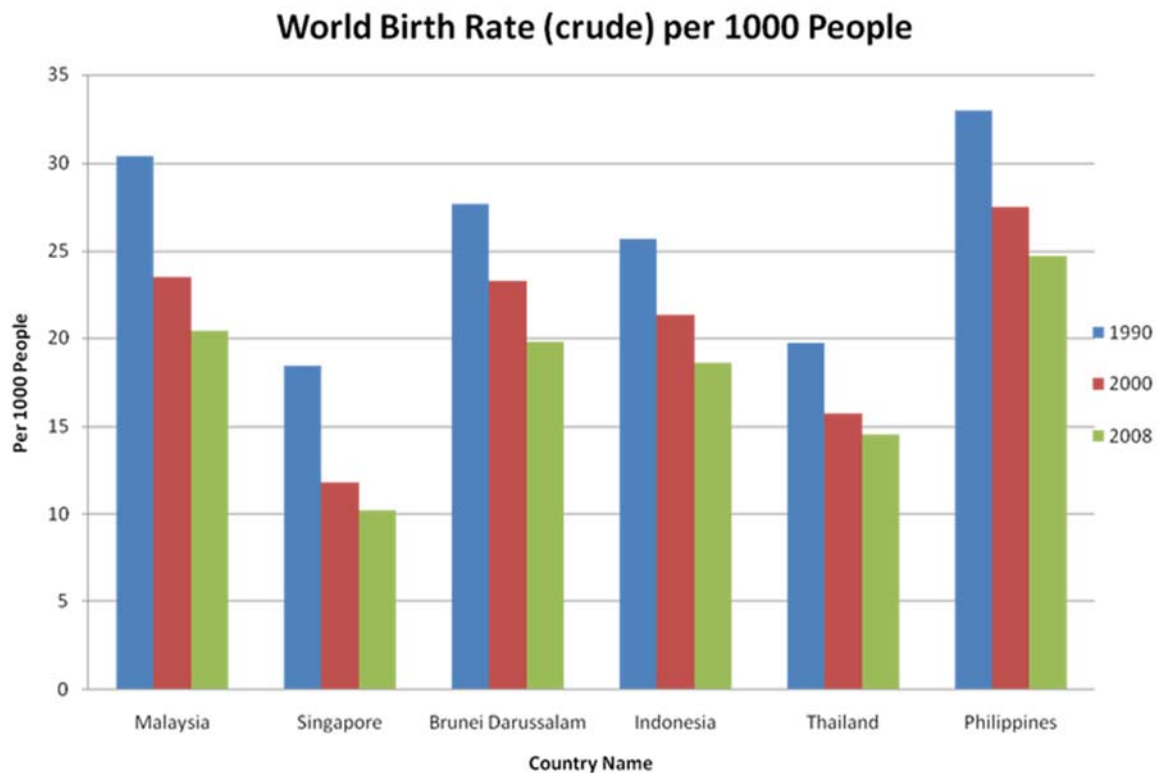
Definition: A **Pareto chart** is a bar graph where the bars are drawn in decreasing order of frequency or relative frequency.



Side-by-Side Bar Graphs:

Consider the information this graph provides. Here we are able to compare each country with one another, as well as how each country changed over time.

These data sets should be compared by using relative frequencies, because different sample or population sizes make comparisons using frequencies difficult or misleading.



(Source: <https://cloudfour808.wordpress.com/2011/03/29/bar-graph-aisyah-zahidah-107168/>)

expl 2: Use the above graph to answer these questions.

- What country and year do we see the highest birth rate given on the graph? What is that birth rate?
- Describe the trend over time for the country of Malaysia.
- Who has the higher birth rate in 2008, Indonesia or Malaysia?

expl 3: (refers to graph on previous page) Let's delve deeper into the question asked in example 2c. The populations and total counts of births for Indonesia and Malaysia in 2008 are given below in the table. Fill in the birth rates, estimated from the graph, in the table below. Explain why we use birth rates (per 1000 people) and *not* the number of births to compare these countries.

Country	2008 Birth Rates (per 1000 people, read from graph)	Approximate number of births	Population (2008)
Indonesia		4,400,512	234,693,997
Malaysia		506,354	24,821,286

Definition: A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

expl 4: Consider the frequency table for the M&Ms example. Let's draw a pie chart for this data using StatCrunch.

Color	Frequency
brown	12
yellow	10
red	9
orange	6
blue	3
green	5
	Total = 45

Instructions for how to make pie charts in StatCrunch are given on the next page.

Instructions: Select **StatCrunch** from the menu in **MyStatsLab**. From there, select **Visit StatCrunch website**. Once there, select **Open StatCrunch** from the tabs at the top.

Label the column marked “var1” by clicking on “var1” and renaming it “Color”. Do the same for “var2”, renaming it “Count”. Enter the data into the columns.

Select **Graph > Pie Chart > With Summary**. You will need to tell it which columns have the categories and counts. Also, title your graph under **Graph Properties > Title**. There are other options but you can leave them on default. Then click “Compute”. It produces the pie chart, with counts and percentages in place.

The textbook has good instructions on this and other displays as well as different software.

Recall: Inferential versus Descriptive Statistics:

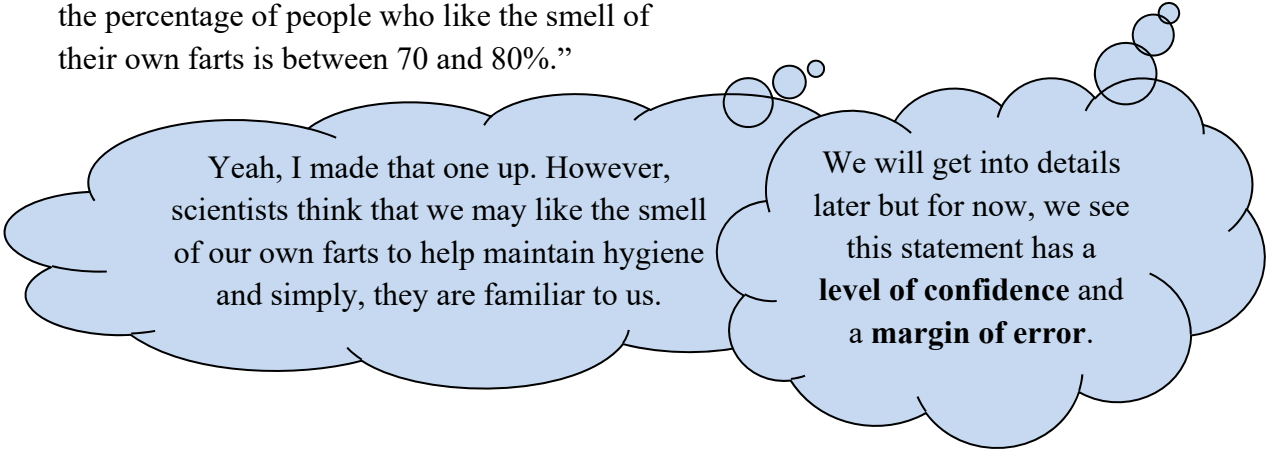
Recall: Definitions: Descriptive Statistics versus Inferential Statistics:

A **statistic** is a numerical summary of a *sample*. **Descriptive statistics** consist of organizing and summarizing data. It uses numerical summaries, tables, and graphs.

Inferential statistics uses methods that take the result from a sample and extend it to the population. It also provides a measure of the reliability of the result. A **parameter** is a numerical summary of the *population*. It is this information that statistics can help us estimate by looking at the sample.

If you are told about the result of a sample in word or graph form, consider this to be **descriptive statistics**.

If you are told information about a whole population (i.e. for all adults, all cars, all states, etc.), consider this to be **inferential statistics**. The statisticians took a sample and then *inferred* from it something about the larger population. They are said to be making a **prediction** using the limited information from a sample. Often, such things will be worded like, “We are 95% confident that the percentage of people who like the smell of their own farts is between 70 and 80%.”



Yeah, I made that one up. However, scientists think that we may like the smell of our own farts to help maintain hygiene and simply, they are familiar to us.

We will get into details later but for now, we see this statement has a **level of confidence** and a **margin of error**.

Organizing Quantitative Data (Section 2.2)

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:

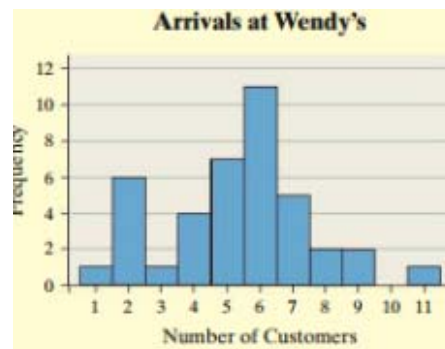
The first step in summarizing quantitative data is to determine whether the data are discrete or continuous.

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*.

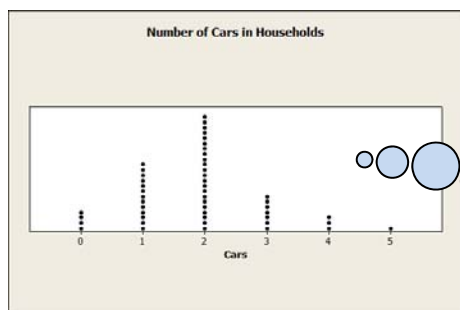
This data was obtained by counting the number of customers arriving during 40 randomly selected 15-minute intervals of time during lunch at a Wendy's restaurant.



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

Definition: A **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.



Here, we simply put a dot for each observation above that response.

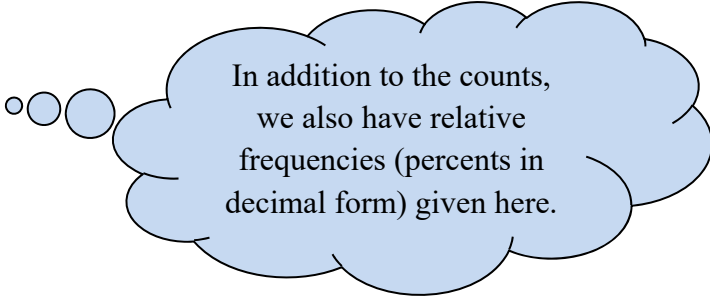
# of Cars	Frequency
0	4
1	13
2	22
3	7
4	3
5	1

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.

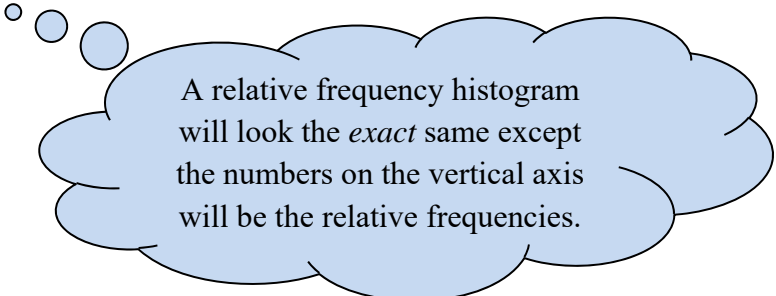
# of Cars	Frequency	Relative Frequency
0	4	$\frac{4}{50} = 0.08$
1	13	$\frac{13}{50} = 0.26$
2	22	0.44
3	7	0.14
4	3	0.06
5	1	0.02



In addition to the counts, we also have relative frequencies (percents in decimal form) given here.

expl 5: 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.



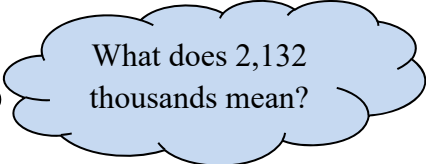
A relative frequency histogram will look the *exact* same except the numbers on the vertical axis will be the relative frequencies.

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



What does 2,132 thousands mean?

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.

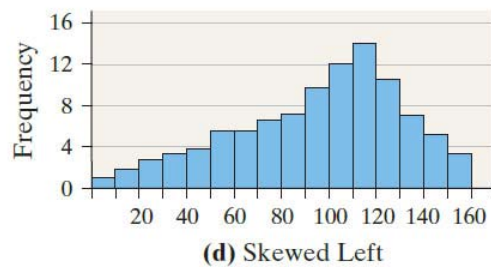
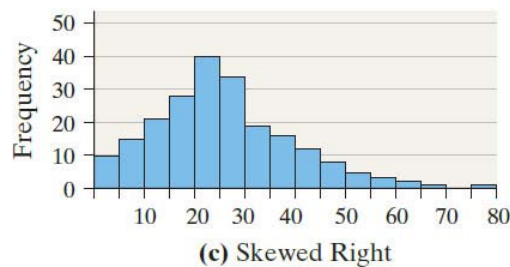
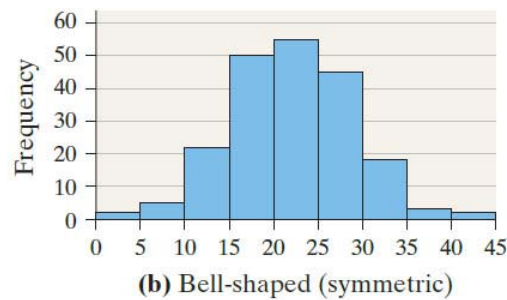
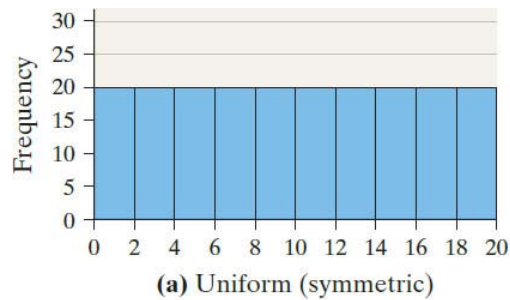
Definition: Data table: A **data table** is an arrangement of data into rows and columns as you see above. A title, row and/or column labels, and qualifiers like “numbers are in thousands” are important.

Shape of the Distribution:

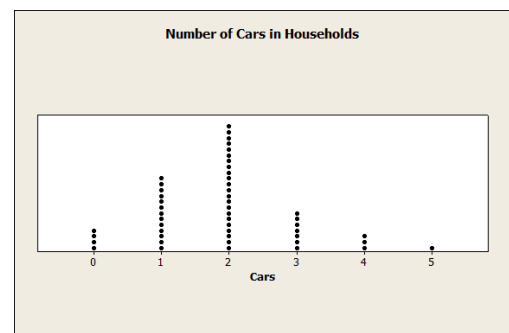
There are four main types of shapes we will discuss. They are

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.

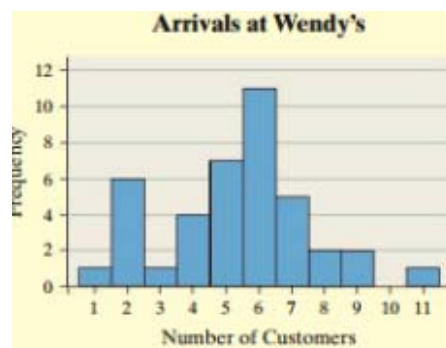
To understand this further, consider the histograms of four different distributions.



expl 6a: Consider the dot plot shown here. Which shape do you think it has?



expl 6b: Consider the histogram shown here. Which shape do you think it has?



Additional Displays of Quantitative Data (Section 2.3)

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.

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.

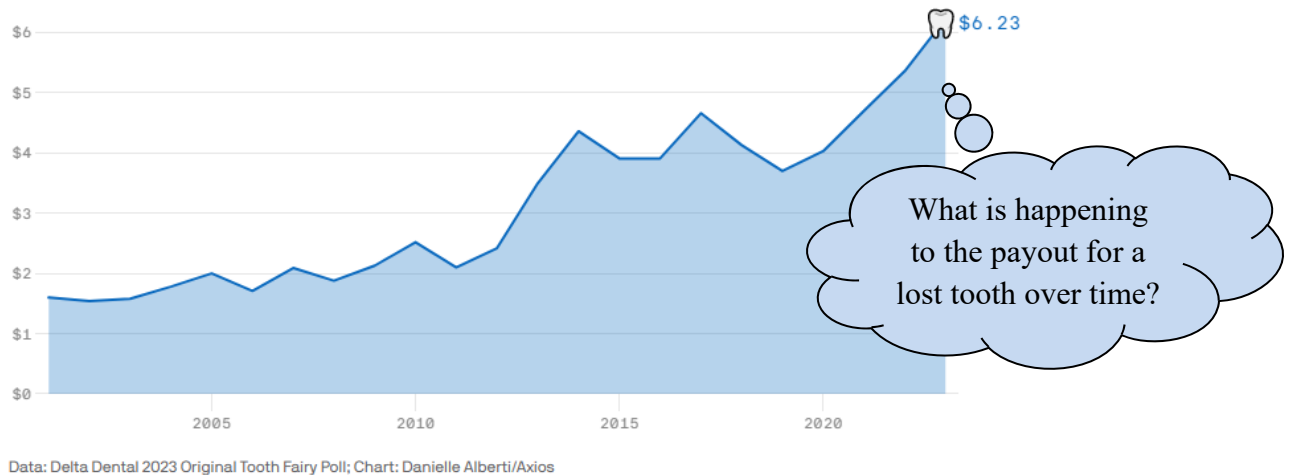
You may decide to **split the stems** on a stemplot. That simply means you use two lines for each leaf rather than a single line. You would do this if the leaves were very long. We will *not* cover this but you can look it up elsewhere.

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.

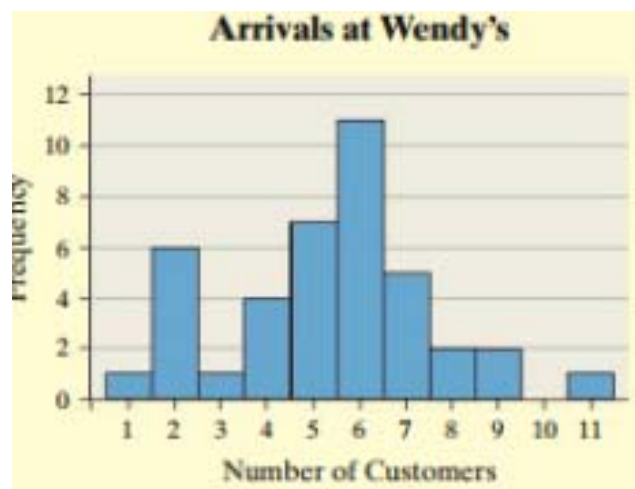
Average Tooth Fairy payout per lost tooth

2001-2023; 1,000 U.S. parents of children ages 6 to 12 polled Jan. 6-19, 2023



Definition: A frequency polygon is a graph that uses points, connected by line segments, to represent the frequencies for the classes. You will plot a point directly above each class' midpoint at that class' frequency. Connect these points and then connect each end to the horizontal axis.

expl 7: Draw a frequency polygon on top of the histogram here.

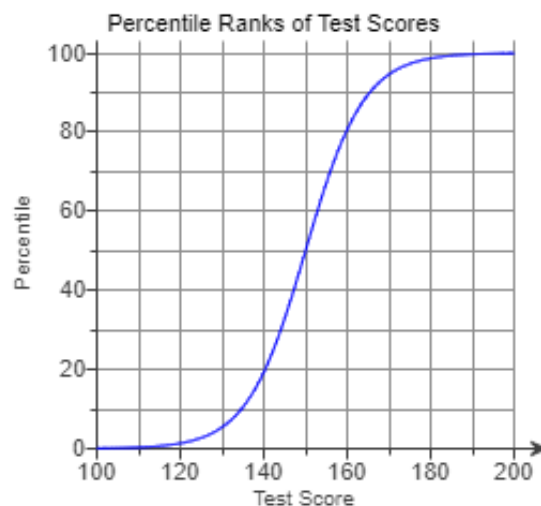
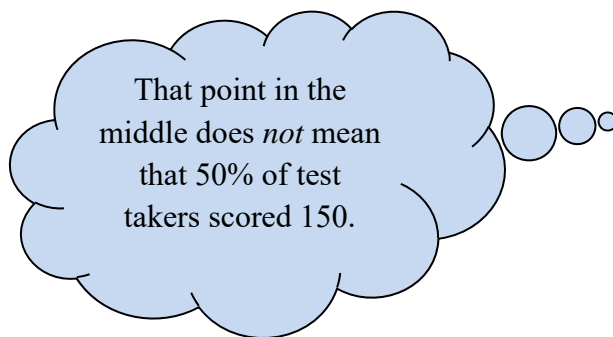


Definition: If needed, the **class midpoint** can be found by adding consecutive lower class limits and dividing by two.

Definition: A **cumulative frequency distribution** displays the *aggregate* frequency of each category. For discrete data, we will add the observations that are less than or equal to the observation being plotted. In other words, we add them as we go to form the **aggregate data**. This is done for continuous data as well, adding those observations that are less than or equal to the upper limit of each class.

If you do this with percentages, and *not* counts, we call this the **cumulative relative frequency distribution**.

Definition: Ogive Graphs: The vertical axis in an ogive (pronounced oh-jive) is the **cumulative relative frequency**. Here is an example which displays the percentiles of a fictional standardized test's scores. (A **percentile**, as we see later, is a score such that a certain percent of scores are less than or equal to it.)



(source: MyStatLab, Fundamentals of Statistics, Sullivan)

expl 8a: Use the above ogive to find the percent of test takers who scored 140 or less.



expl 8b: Use the above ogive to find the score that 80% of the test takers scored less than or equal to.

Graphical Misrepresentations of Data (Section 2.4)

“There are three kinds of lies: lies, damned lies, and statistics.” (British Prime Minister Benjamin Disraeli, 1804-1881) This saying was popularized by Mark Twain.

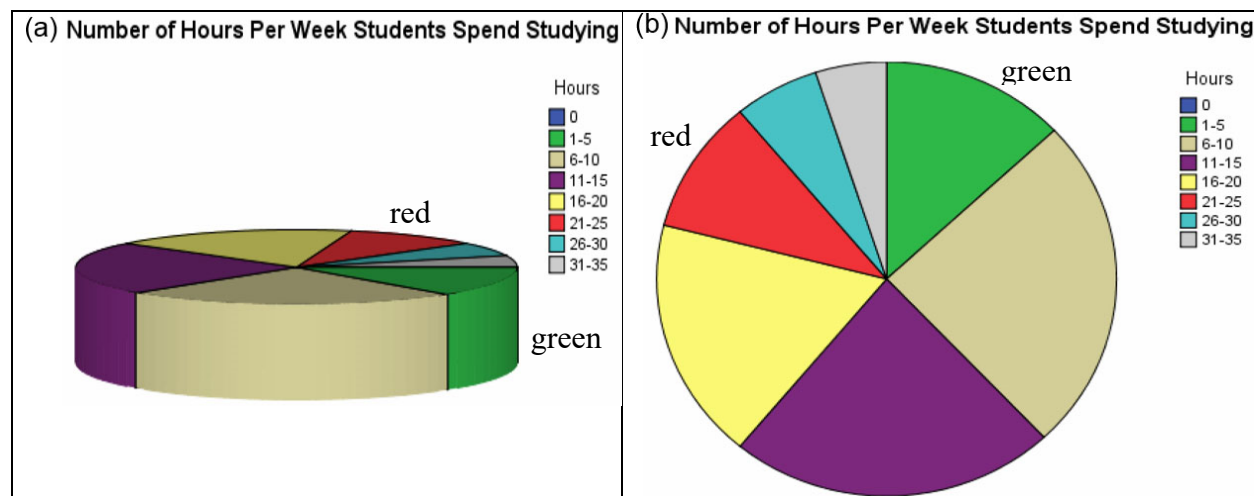
As producers of data displays and statistics, we try hard to provide accurate depictions. As consumers, we must be on the lookout for misleading (intentional or not) displays. We will look at a few ways this can be done.

The book labels it **misleading** if it is unintentional. If it is done on purpose, the book calls that **deceptive**.

Three-dimensional Charts and Graphs:

Consider the two pie charts that purport to tell us the same story.

The National Survey of Student Engagement is a survey that (among other things) asked first year students at liberal arts colleges how much time they spend preparing for class each week. The results from the 2007 survey are summarized in these dramatically different pie charts.



We, in trying to make our graphics more dynamic and interesting, do some pretty irresponsible things. Three-dimensional graphs are often hard to read properly. Which graph shows the true relationship between the “1-5 hours” (green) group and the “21-25 hours” (red) group more accurately?

You should never use three-dimensional graphics. Labeling each sector with its percentage would help. Yet it is still hard for our minds to disavow what our eyes incorrectly see.

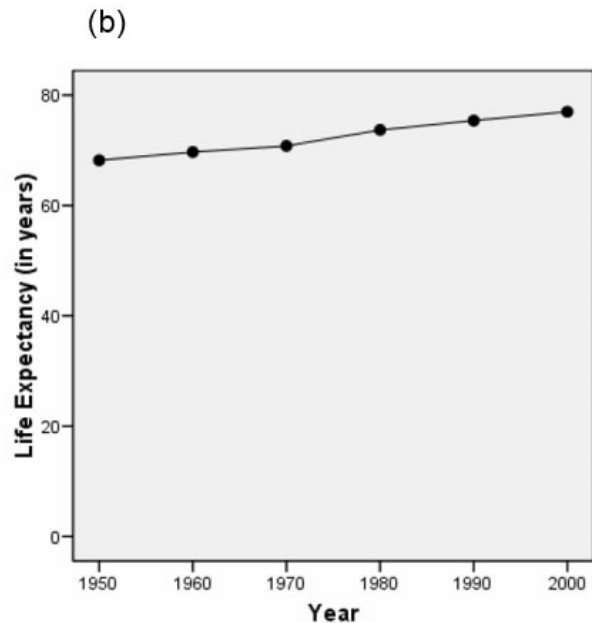
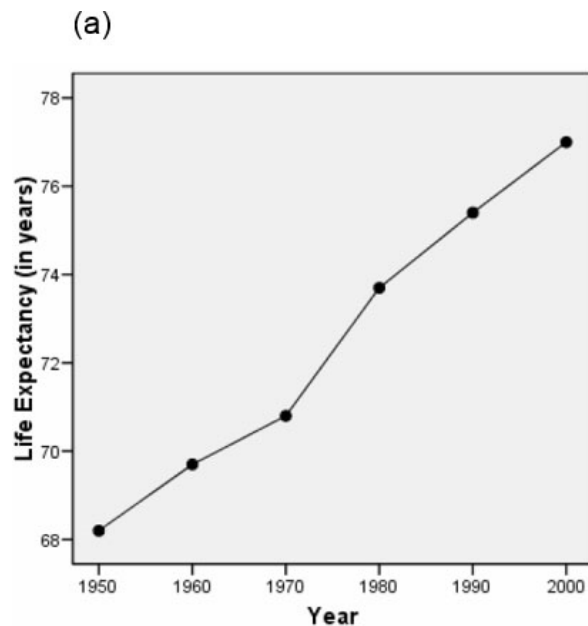
Adjusting the Scale of Graphs:

expl 9: Consider the data for the life expectancies (in years) of residents of the United States through the last half-century. (Source: National Center for Health Statistics)

Year	Life Expectancy (in years)
1950	68.2
1960	69.7
1970	70.8
1980	73.7
1990	75.4
2000	77.0

Looking at the table,
how would you
describe the data?

The table lists out the data in a
straightforward way. How could a
graph possibly change that?



Comment on how these graphs make you feel about the trend in life expectancy in the US. What makes the difference? Which do you interpret as more accurate? Why?

Pictographs:

Pictographs use pictures to make the graphics look fancier. Consider these examples.

expl 10: Here we see a pictograph. The numbers show that the 1980 purchasing power of the Canadian dollar (in terms of the American dollar) is twice that for the year 1995. However, the area of the coins (which is what our minds really compare) tell a different story. Why is that?



Purchasing power:
the value of money
in terms of how
much it can buy

(Source: https://www.statcan.gc.ca/edu/power-pouvoir/ch9/img/5214825_02-eng.jpg)

Not always bad:

Pictographs, if carefully constructed, can make good displays.

This pictograph shows the same information. However, you will notice that the relative sizes are *not* skewed. Only one dimension was reduced to show the decrease in purchasing power.

Purchasing Power of the Canadian Dollar, 1980 to 2000



Guidelines for Constructing Good Graphics:

- Title and label the graphic axes clearly, providing explanations, if needed. Include units of measurement and a data source when appropriate.
- Avoid distortion. Never lie about the data.
- Minimize the amount of white space in the graph. Use the available space to let the data stand out. If scales are truncated, be sure to clearly indicate this to the reader.
- Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader.
- Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion of the graph to draw attention to that area. Don't try to force the reader to any specific part of the graph. Let the data speak for themselves.
- Avoid relative graphs that are devoid of data or scales.

Optional Worksheet: Data Displays 2: This worksheet reviews stem-and-leaf plots and pie charts. We will also see how the scale of a line graph affects its meaning.

Optional Worksheet: Do the Numbers Make Sense?: This worksheet has several examples concerned with miscalculations and misinterpretations of statistics, and not their graphical displays.