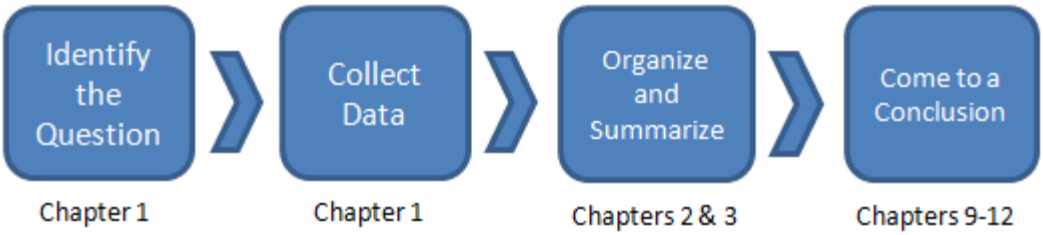
Chapter 2: Organizing and Summarizing Data

Let's review the process of statistics we introduced in Section 1.1:



In Chapter 1, we focused on how to collect data. In this next chapter, we'll talk about how to organize and summarize data using tables in graphs. Section 2.2 will focus on qualitative data, while sections 2.2 and 2.3 will focus on quantitative data. The last section, Section 2.4, talks about various ways that data can be misrepresented.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

Section 2.1: Organizing Qualitative Data

Objectives

By the end of this section, you will be able to...

1. organize qualitative data in tables
2. construct bar graphs
3. construct pie charts

Frequency and Relative Frequency Tables

Let's suppose you give a survey concerning favorite color, and the data you collect looks something like the table below.

blue	red	blue	orange	blue
blue	green	blue	purple	blue
blue	red	pink	green	blue

yellow	green	red	pink	
blue	green	yellow	pink	
yellow	green	blue		

Clearly, we need a better way to summarize the data. The most obvious thing to do would be to make a table with the list of favorite colors and the frequency for each.

favorite color	frequency
blue	10
red	3
orange	1
yellow	3
green	5
pink	3
purple	1

Officially, we call this a **frequency distribution**.

A **frequency distribution** lists each category of data and the number of occurrences for each category.

Sometimes, we really want to know the frequency of a particular category in reference to the total. We can do this just by finding the total, and dividing the frequency for each category by that total.

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

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

A **relative frequency distribution** lists each category of data together with the relative frequency of each category.

favorite color	relative frequency
blue	10/26 ≈ 0.38
red	3/26 ≈ 0.12
orange	1/26 ≈ 0.04
yellow	3/26 ≈ 0.12
green	5/26 ≈ 0.19
pink	3/26 ≈ 0.12
purple	1/26 ≈ 0.04

Technology

Here's a quick overview of how to create frequency and relative frequency tables in StatCrunch.

1. Enter or import the data.
2. Select **Stat** > **Tables** > **Frequency**.
3. Select the column(s) you want to summarize and click **Next**.
4. Add any modifications for an "Other" category and how to order the categories.
5. Click **Calculate** and another window with these numbers calculated will pop up.
6. You can then choose **Options** > **Copy** to copy the output for use elsewhere.

Bar Graphs

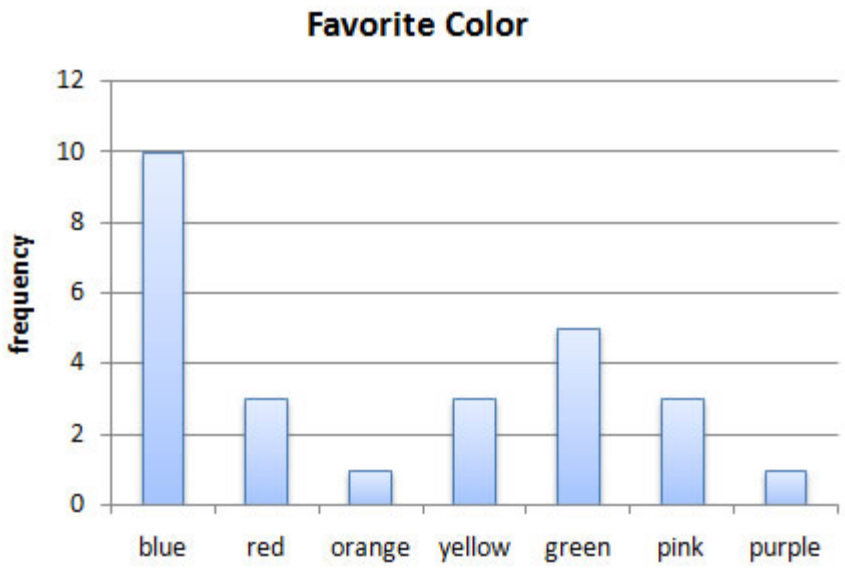
Bar graphs are probably the most commonly used graphs, and one you're already familiar with. I won't mention much more here, except to state a couple keys:

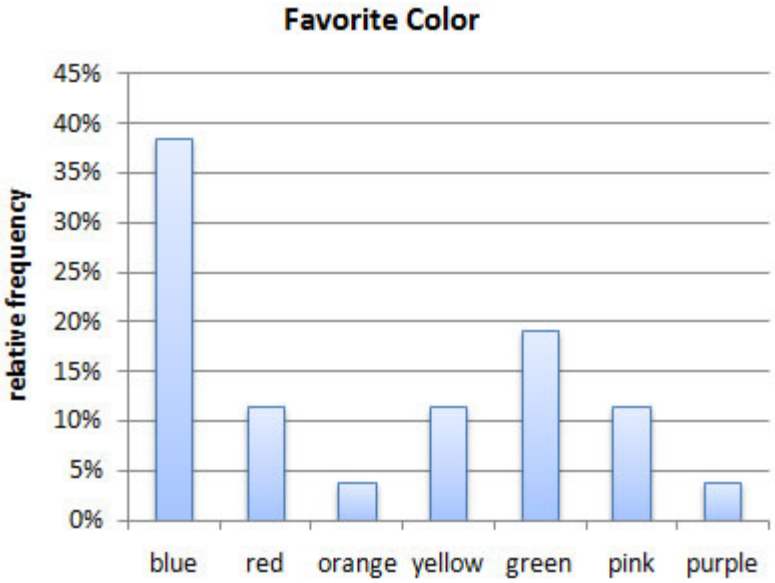
1. heights can be frequency or relative frequency
2. bars must not touch

Using our the data from our previous color example,

favorite color	frequency	relative frequency
blue	10	10/26 ≈ 0.38
red	3	3/26 ≈ 0.12
orange	1	1/26 ≈ 0.04
yellow	3	3/26 ≈ 0.12
green	5	5/26 ≈ 0.19
pink	3	3/26 ≈ 0.12
purple	1	1/26 ≈ 0.04

we could then make both frequency and relative frequency bar graphs.





Technology

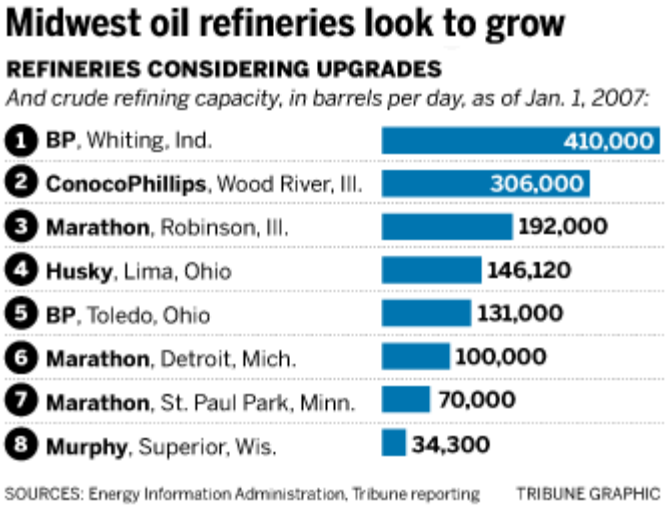
Here's a quick overview of how to create bar graphs in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Bar Graph**, then choose **with data** or **with summary**.
3. If you chose *with data*, select the column(s) you wish to use and click **Next**. If you chose *with summary*, set the columns containing the categories and counts and click **Next**.
4. Choose the type (*Frequency* or *Relative Frequency*) and click **Next**.
5. Enter any modifications and/or color schemes and click **Create Graph!**
6. You can then choose **Options** > **Copy** to copy the box plot for use elsewhere.

Pareto Charts

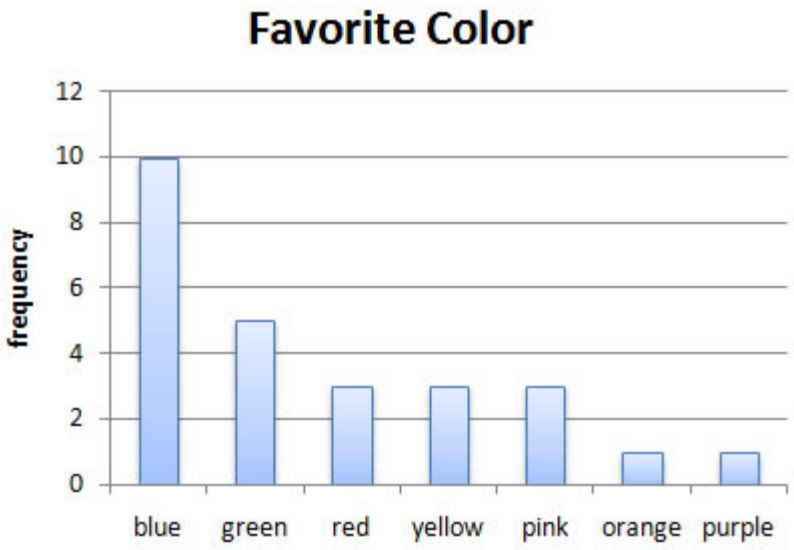
> A Pareto chart is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

You see Pareto charts fairly often in the newspaper, because often the article is trying to show that one particular category is the highest or lowest. The image below, for example, is from the Chicago Tribune. You can see clearly from the graph that it's attempting to show that the local BP refinery in Whiting, Indiana is the highest-capacity refinery that is considering expansion.



If you don't remember the issue, you can read up about BP's plan to expand it's refinery in this article from CBS2 Chicago.

Here's another one, using the favorite color data from the last section:

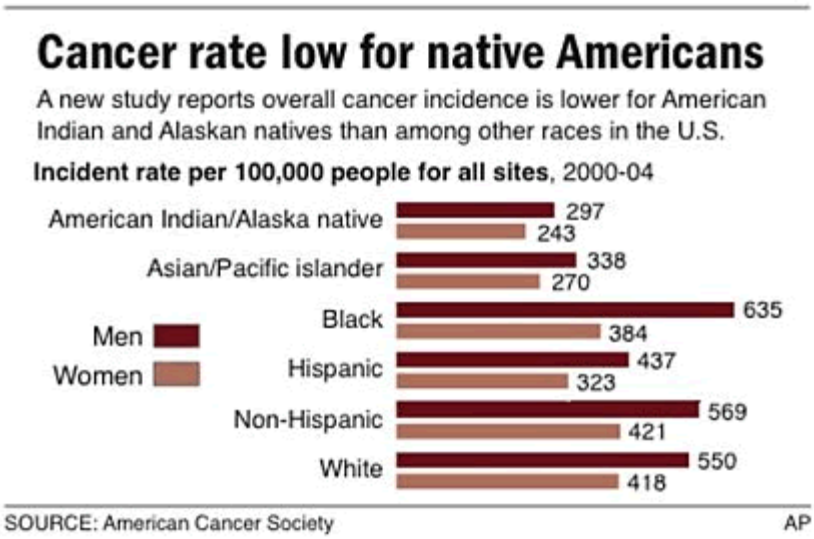


Side-by-Side Bar Graphs

Side-by-side bar graphs are used when you want to compare two different populations. The key with side-by-side bar graphs is that you **must use relative frequencies**. Do you know why?

I think so. But just in case...

Here's a good example of a side-by-side chart, from the Associated Press.

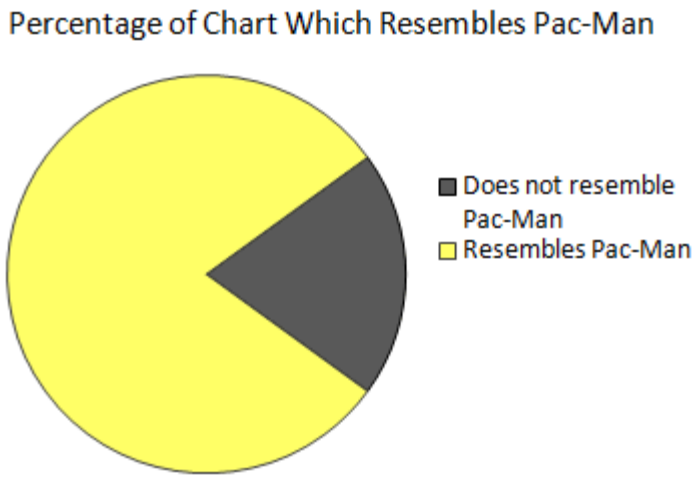


What's shown isn't quite a relative frequency as we've defined it - it's the number per 100,000, where ours as a percent is the number per 100. The reason why the rate per 100,000 is used here is because the percents would all be less than 1% and difficult to read. Still, if frequency was used instead, the "White" category would be the largest, simply because that's the largest segment of the U.S. population.

Technology

Here's a quick overview of how to create side-by-side bar graphs in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Chart** > **Columns**
3. Select the columns you'll be using.
4. Select the location of the lablels (*Row labels in*).
5. If desired, choose an order.
6. Choose the plot type (*vertical bars* for a side-by-side bar graph) and click **Next**.
7. Enter any modifications and/or color schemes and click **Create Graph!**
8. You can then choose **Options** > **Copy** to copy the box plot for use elsewhere.



Pie Charts

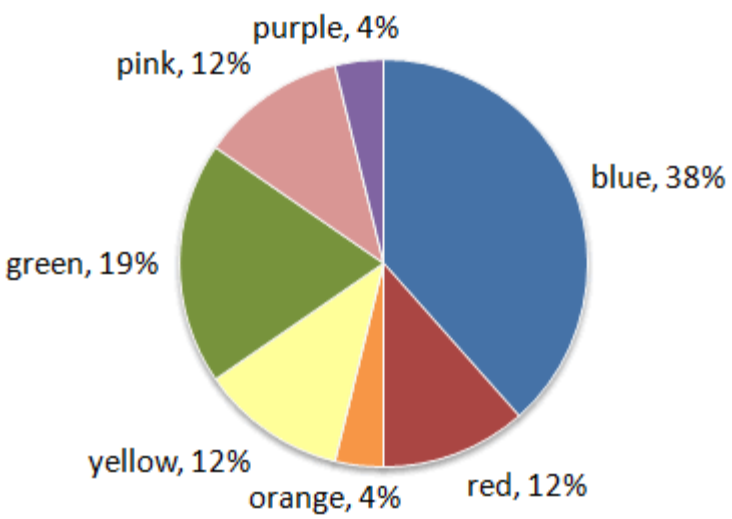
Like bar graphs, pie charts are very common. You're probably already aware of these as well. I'll just include a couple comments:

1. should always include the relative frequency
2. also should include labels, either directly or as a legend

Using our the data from our previous color example,

favorite color	frequency	relative frequency
blue	10	10/26 ≈ 0.38
red	3	3/26 ≈ 0.12
orange	1	1/26 ≈ 0.04
yellow	3	3/26 ≈ 0.12
green	5	5/26 ≈ 0.19
pink	3	3/26 ≈ 0.12
purple	1	1/26 ≈ 0.04

we get this pie chart:.



Technology

Here's a quick overview of how to create pie charts in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Pie Chart**, then choose **with data** or **with summary**.
3. If you chose *with data*, select the column(s) you wish to use and click **Next**. If you chose *with summary*, set the columns containing the categories and counts and click **Next**.
4. Enter any modifications (labels, title, color scheme, etc) and click **Create Graph!**
5. You can then choose **Options** > **Copy** to copy the box plot for use elsewhere.

Section 2.2: Organizing Quantitative Data: The Popular Displays

Objectives

By the end of this section, you will be able to...

1. organize quantitative data into tables
2. construct histograms for discrete and continuous data
3. draw stem-and-leaf plots
4. draw dot plots
5. identify the shape of a distribution

Like qualitative data in the last section, quantitative data can (and should) be organized into tables. We'll break this page up into two parts - discrete and continuous.

Organizing Discrete Data into Tables

If you recall from Section 1.2,

> A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of values. (*Countable* means that the values result from counting - 0, 1, 2, 3, ...)

Since we can list all the possible values (that's essentially what *countable* means), one way to make a table is just to list the values along with their corresponding frequency.

Example 1

Here's some data I collected from a previous students Mth120 course. It refers to the number of children in their family (including themselves).

2	2	2	4	5	3	3	3	3
2	1	2	3	5	3	4	3	1
2	3	5	3	2	1	3	2	

An easy way to compile the data would then be to make a frequency or relative frequency table as we did before.

children	frequency	relative frequency
1	3	3/26 ≈ 0.12
2	8	8/26 ≈ 0.31
3	10	10/26 ≈ 0.38
4	2	2/26 ≈ 0.08
5	3	3/26 ≈ 0.12

Sometimes, however, we have too many values to make a row for each one. In that case, we'll need to group several values together.

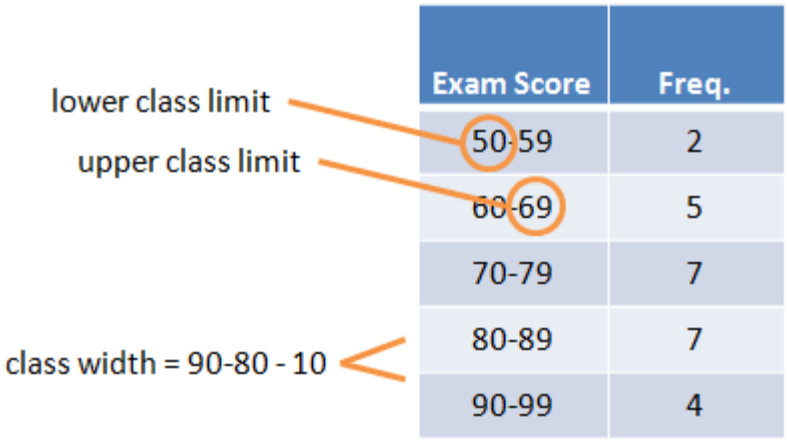
Example 2

A good example might be the scores on an exam, ranging from 1-100. Here are some data from a past Mth120 class.

62	87	67	58	95	94	91	69	52
76	82	85	91	60	77	72	83	79
63	88	79	88	70	75	87		

In this case, we'll have to set up intervals of numbers called **classes**. Each class has a **lower class limit** and an **upper class limit**, along with a **class width**. The class width is the difference between successive lower class limits.

To be consistent, the class width should be same for each class. One good option might look something like this:



Exam Score	Freq.
50-59	2
60-69	5
70-79	7
80-89	7
90-99	4

Organizing Continuous Data into Tables

Organizing continuous data is similar to organizing multi-valued discrete data. We have to form classes which don't overlap. I usually try to design a class width that's either logical (i.e. 10 points for grades above) or so that I have 5-8 classes when complete.

Example 3

For this example, let's consider the average commute for each of the 50 states. The data below show the average daily commute of a random sample of 15 states.

23.1	18.3	23.2	19.9	26.6
24.8	23.1	23.2	22.7	29.4
22.3	30.0	25.8	21.9	16.7

Source: US Census

Do you know why this is a continuous random variable and not discrete? (Hint: It's *not* because of the decimal.)

I think I know!

To make a frequency or relative frequency for continuous data, we use the same strategy we'd use for multi-valued discrete data.

average commute	frequency	relative frequency
16-17.9	1	1/15 ≈ 0.07
18-19.9	2	2/15 ≈ 0.13
20-21.9	1	1/15 ≈ 0.07
22-23.9	6	6/15 = 0.40
24-25.9	2	2/15 ≈ 0.13
26-27.9	1	1/15 ≈ 0.07
28-29.9	1	1/15 ≈ 0.07
30-31.9	1	1/15 ≈ 0.07

Once we have these tables, we'll need to learn how to create some charts to display the information, which is what the next few page are about.

Technology

Here's a quick overview of how to create frequency and relative frequency tables for quantitative data in StatCrunch.

Discrete Data

1. Enter or import the data.
2. Select **Stat** > **Tables** > **Frequency**.
3. Select the column(s) you want to summarize and click **Next**.
4. Add any modifications for an "Other" category and how to order the categories, and click **Calculate**.

Continuous or Multi-valued Discrete Data:

1. Enter or import the data.
2. Select **Data** > **Bin Column**.
3. Select the column containing the data, select "Use fixed width bins", and set the lowest class limit (*Start bins at:*) and class (*bin*) width.
4. Click **Calculate.**
5. Select **Stat** > **Tables** > **Frequency**.
6. Select the newly created bin column and click **Calculate**.*

* Note that these classes *seem* to overlap, but that the class "0-k" does not include *Mk.*

Stem-and-Leaf Plots

Stem-and-leaf plots are another way to represent quantitative data. They give more detail because they show the actual data. The idea is to split each data value into two parts - a **stem** and a **leaf**. The **stem** is everything of the right-most digit, and the **leaf** is that right-most digit. Here's an example, using the data from earlier this section regarding exam scores from a previous Mth120 class.

Example 6

62 87 67 58 95 94 91 69 52

76 82 85 91 60 77 72 83 79
63 88 79 88 70 75 87

With these data, the stems are the first digits - 5, 6, 7, 8, and 9. The leafs are all the second digits, 0, 1, ... , 9. The full **stem-and-leaf plot** lists the stems down the left side, a vertical bar between, and then lists the leafs in order to the right. Something like this:

5	2 8
6	2 7 9 0 3
7	6 7 2 9 9 0 5
8	7 2 5 3 8 8 7
9	5 4 1 1

It's interesting that this plot looks very similar to a histogram, only it gives us the actual data. Take a look at this animation to see the relationship:

There are some limitations to stem-and-leaf plots. In particular, we're limited to small data sets - can you imagine the leaves if we had 1,000 test scores? Also, the range in the data needs to be fairly small.

By that, I mean if the data values range from 1-100, our stems can be 0, 10, 20, ... , 90, as they were in this example. On the other hand, if the values range from 1-10,000, the stems would have to be 0, 10, 20, ... , 9,980, 9,990. That's a lot of rows!

Technology

Here's a quick overview of how to create stem-and-leaf plots in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Stem and Leaf**
3. Select the column you wish to use and click **Create Graph!**

Dot Plots

Dot pots are similar to single-valued histograms, but rather than placing rectangles above each particular value, a dot plot just places the required number of dots above each value. Looking at our example again with the number of children, the plot would look something like this:

Technology

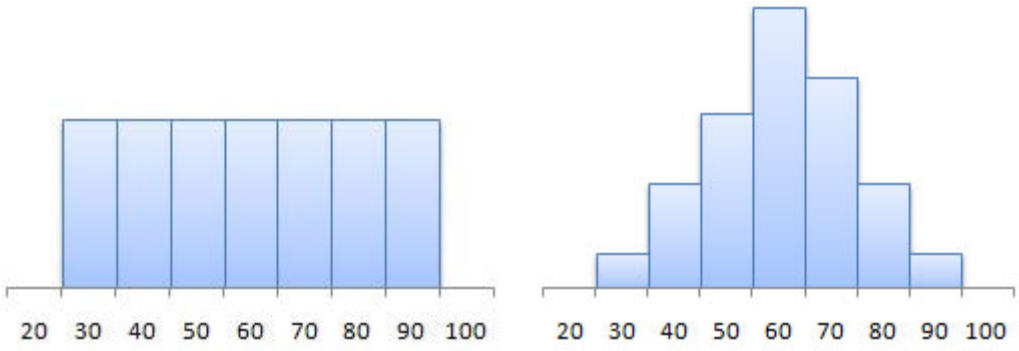
Here's a quick overview of how to create dot plots in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Dotplot**.
3. Select the column you wish to use and click **Next**.
4. Set any options and click **Create Graph!**

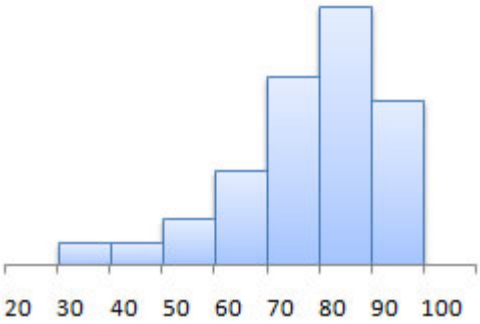
Distribution Shape

A good way to describe a distribution is its shape. In general, we describe a distribution's shape in one of four ways (though there are others):

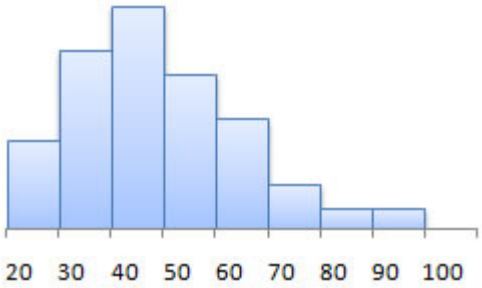
1. **uniform** - frequencies are evenly spread out among all values of the variable
2. **symmetric (bell-shaped)** - highest value is in the middle, with values tailing off to the right and left
3. **left-skewed** - highest value is on the right, with a longer left "tail"
4. **right-skewed** - highest values is on the left, with a longer right "tail"



uniform symmetric (bell-shaped)



left-skewed



right-skewed

<< previous section | next section >>

1 2 3 4 5 6 7 8 9 10 11 12 13

Section 2.3: Additional Displays of Quantitative Data

Objectives

By the end of this section, you will be able to...

1. construct frequency polygons*
2. create cumulative frequency and relative frequency tables
3. construct ogives*
4. draw time-series graphs

* You will not be tested on these objectives.

In addition to histograms, stem-and-leaf plots, and dot plots, there are some other, section common plots. We'll introduce a couple in this section. The first type, **frequency polygons**, are not a type of plot that will be expected of you on exams, though you will be asked questions about them on homework.

Frequency Polygons

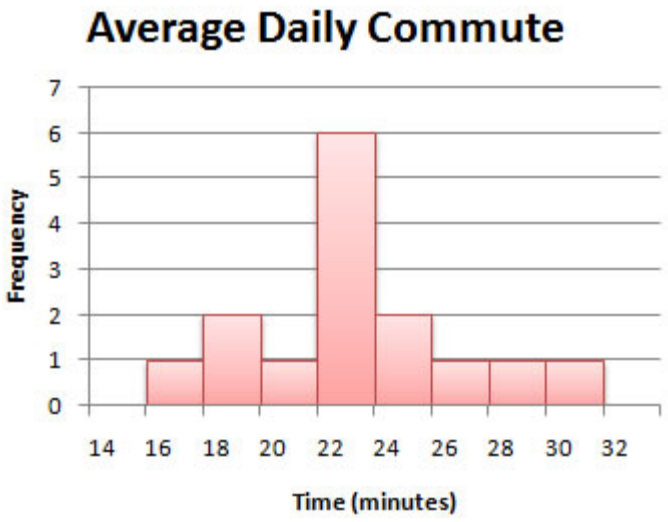
A **frequency polygon** is drawn by plotting a point above each class midpoint and connecting the points with a straight line. (**Class midpoints** are found by average successive lower class limits.)

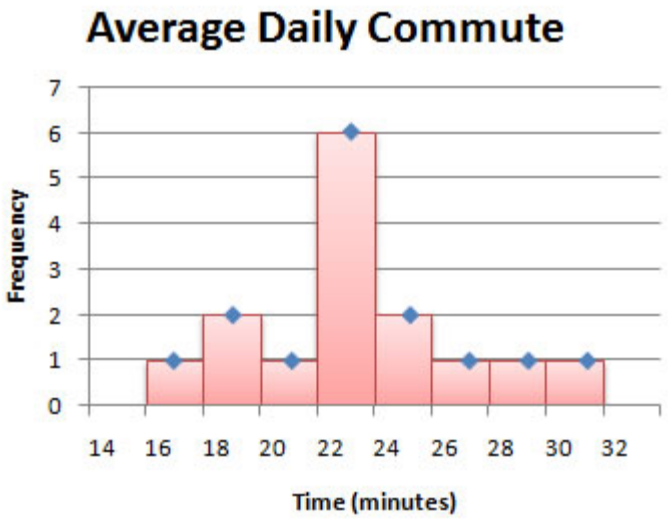
Example 1

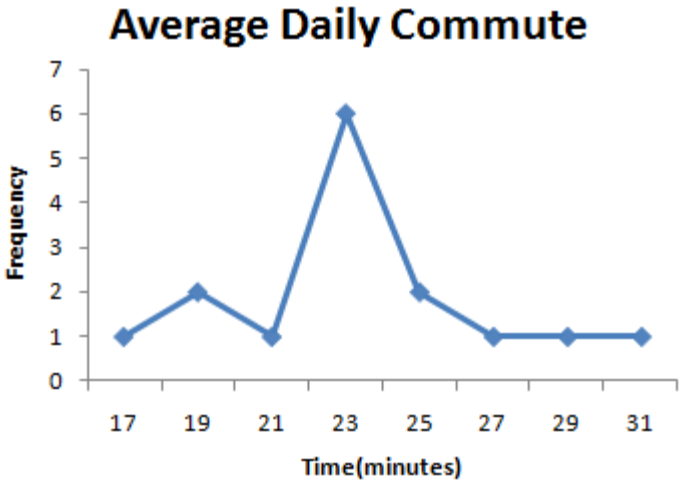
To illustrate the idea, let's look at the average commute data from the last section.

average commute	midpoint	frequency	relative frequency
16-17.9	17	1	1/15 ≈ 0.07
18-19.9	19	2	2/15 ≈ 0.13
20-21.9	21	1	1/15 ≈ 0.07
22-23.9	23	6	6/15 = 0.40
24-25.9	25	2	2/15 ≈ 0.13
26-27.9	27	1	1/15 ≈ 0.07
28-29.9	29	1	1/15 ≈ 0.07
30-31.9	31	1	1/15 ≈ 0.07

The three images below show the relationship between the histogram and the frequency polygon.







Note: No technology section this time, since you won't be asked to do this for exams.

Cumulative Tables

Cumulative tables are just what they imply - they show the sum of values up to and including that particular

category. As with regular tables, we can have both cumulative frequency and relative frequency.

Example 2

To illustrate the idea, let's look at the average commute data from the last section.

average commute	frequency	cumulative frequency
16-17.9	1	1
18-19.9	2	3
20-21.9	1	4
22-23.9	6	10
24-25.9	2	12
26-27.9	1	13
28-29.9	1	14
30-31.9	1	15

average commute	relative frequency	cumulative relative frequency
16-17.9	1/15 ≈ 0.07	1/15 ≈ 0.07
18-19.9	2/15 ≈ 0.13	3/15 ≈ 0.20
20-21.9	1/15 ≈ 0.07	4/15 ≈ 0.27
22-23.9	6/15 = 0.40	10/15 ≈ 0.67
24-25.9	2/15 ≈ 0.13	12/15 = 0.80
26-27.9	1/15 ≈ 0.07	13/15 ≈ 0.87
28-29.9	1/15 ≈ 0.07	14/15 ≈ 0.93
30-31.9	1/15 ≈ 0.07	15/15 = 1.00

Technology

Unfortunately, there is no easy way to create cumulative tables in StatCrunch. The best method is to create a regular frequency or relative frequency table and compute the cumulative values by hand.

Ogives

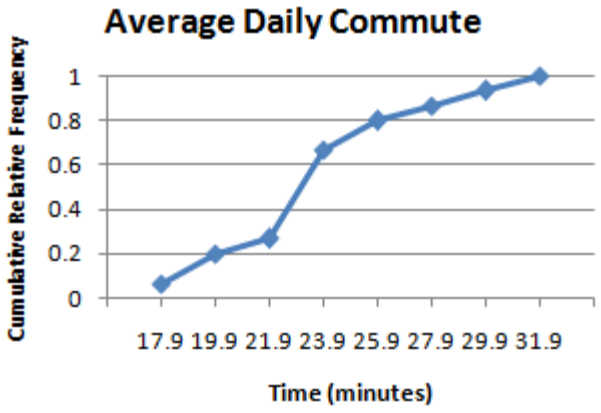
Ogives are pretty funky graphs, and rarely used except in specific areas. We'll just give a quick example here, but like frequency polygons, you won't be expected to create these on an exam. (Though it may come up in homework.)

An **ogive** (read as "oh jive") is a graph that represents the cumulative frequency or cumulative relative frequency for the class. It is constructed by plotting points - the x-coordinates are the upper class limits and the y-coordinate is the corresponding cumulative frequency or cumulative relative frequency.

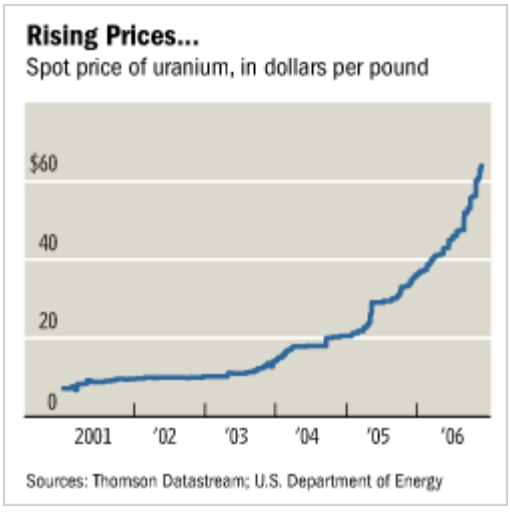
Example 3

To illustrate the idea, let's again use the average commute data from the last section.

average commute	relative frequency	cumulative relative frequency
16-17.9	1/15 ≈ 0.07	1/15 ≈ 0.07
18-19.9	2/15 ≈ 0.13	3/15 ≈ 0.20
20-21.9	1/15 ≈ 0.07	4/15 ≈ 0.27
22-23.9	6/15 = 0.40	10/15 ≈ 0.67
24-25.9	2/15 ≈ 0.13	12/15 = 0.80
26-27.9	1/15 ≈ 0.07	13/15 ≈ 0.87
28-29.9	1/15 ≈ 0.07	14/15 ≈ 0.93
30-31.9	1/15 ≈ 0.07	15/15 = 1.00



Note: No technology section this time, since you won't be asked to do this for exams.



Time-Series Graphs

Time series graphs are much more common than the last couple times we've looked at. It's common to see stock

prices and daily temperature graphs in the news - both are time series plots.

A **time series plot** 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.

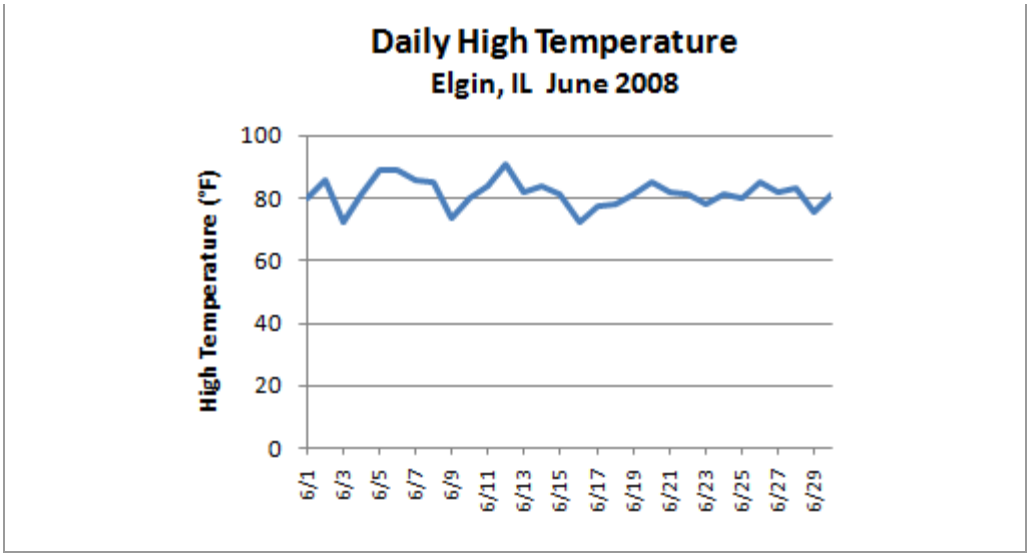
The example above is from the Chicago Tribune and reflects the price of uranium from 2001-2006.

Example 4

Here's another example, using the daily high temperature in Elgin, IL, for the month of June, 2008.

date	daily high temperature
6/1	80
6/2	86
6/3	72
6/4	81
6/5	89
6/6	89
6/7	86
6/8	85
6/9	73
6/10	80
6/11	84
6/12	91
6/13	82
6/14	84
6/15	81
6/16	72
6/17	77
6/18	78
6/19	81
6/20	85
6/21	82
6/22	81
6/23	78
6/24	81
6/25	80
6/26	85
6/27	82
6/28	83
6/29	75
6/30	81

And the time series plot would look something like this:



Technology

Here's a quick overview of how to create a time series plot in StatCrunch.

1. Enter or import the data.
2. Select **Graphics** > **Index Plot**
3. Select the column(s) you want to plot and click **Next**.
4. Set any desired options and click **Create Graph!**

<< previous section | next section >>

1 2 3 4 5 6 7 8 9 10 11 12 13

Section 2.4: Graphical Misrepresentations of Data

Objectives

By the end of this section, you will be able to...

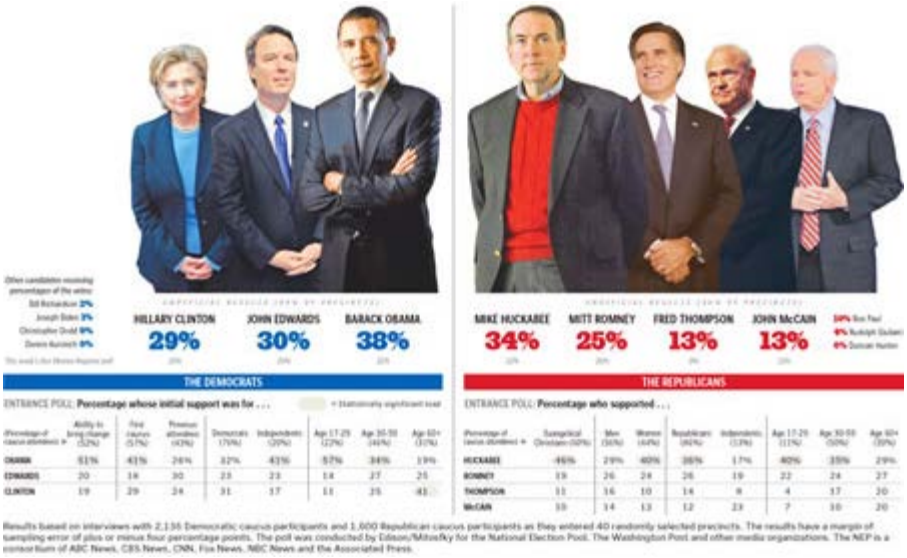
1. describe what can make a graph misleading or deceptive

Misleading and Deceptive Graphs

The author of your text makes an interesting distinction between "misleading" and "deceptive" graphs. It's an important point, so read through that paragraph before continuing on to the examples. (Page 104)

Example 1

This first one was from the Washington Post after the Iowa caucuses in January, 2008. Look carefully at the graphic and try to determine what was misleading about it.



OK, I have an idea.

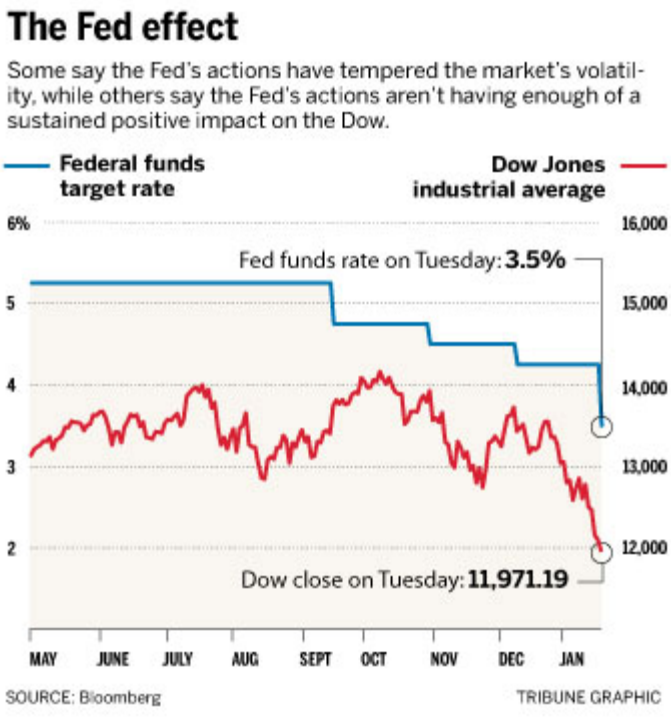
Example 2

This next graphic is attempting to relate the purchasing power of the Canadian dollar (also known as the "Loonie" - I love that!) in relation to the U.S. dollar. This is a bit more subtle. Can you see what's misleading about this?

OK, I'm ready.

Example 3

Here's a classic graphic from the Chicago Tribune. This is very typical of graphics representing the stock market. Can you see what's wrong?



I think so. Let me see if I'm right.

Look for this error next time whenever you read an article that's trying to show how quickly something is increasing or decreasing.

<< previous section | next section >>