3 – Organizing Data

A frequency distribution is a *table* that organizes quantitative data into classes. The number of values in each class is the frequency.

Class limits are rough criteria to define classes.

Class boundaries are specific numerical values marking endpoints of the classes.

Class width is the difference between boundaries, or the width of the class.

Class midpoints are the values in the center of the class boundaries

There are some conventional (and logical) rules for frequency distributions:

1. Between 5 and 20 classes is best.
2. Classes may not overlap.
3. Classes with no data are shown.
4. All values in the data set must be included.
5. Classes must be of equal widths
   (exceptions to Rule 5 may be made in some cases for the first and/or last class)

The cumulative frequency distribution for a dataset follows the same idea, except that classes all begin with the minimum value or zero. The frequencies thus become additive, so that each class's cumulative frequency is equal to the sum of its own frequency and all the preceding frequencies. This is computed in the example below.

Relative frequency is computed by dividing frequency by the total number of data in the orginal data set. This gives the *portion* of the data that are within each class. The following properties will apply to all relative frequencies:

- Relative frequencies are between 0 and 1
- The sum of the relative frequencies in a distribution are 1.

Relative frequency is also computed in the example below.

Example: Consider this dataset of 32 responses to the question, "How long would it take to drive home from where you are right now (in minutes)?" from the spring 2012 Math 175 Student Survey.

10	20	35	90
10	20	40	110
10	20	45	150
10	25	45	240
15	25	45	240
15	30	45	270
15	30	60	1500
20	30	60	3960

If we ignore the two highest values (1500 & 3960; one of the students was from Texas, and another from Central America), then we can put the numbers between 10 and 270 into classes. Using 280 because it divides evenly by 20, we will have 15 classes (14 for the data between 10 and 270, and a 15th for the large numbers). That will give us "nice" classes with the following class are: 0 – 20, 20 – 40, . . . , 260 – 280, and 280 +.

(Note the exception to rule 5 for the last class)

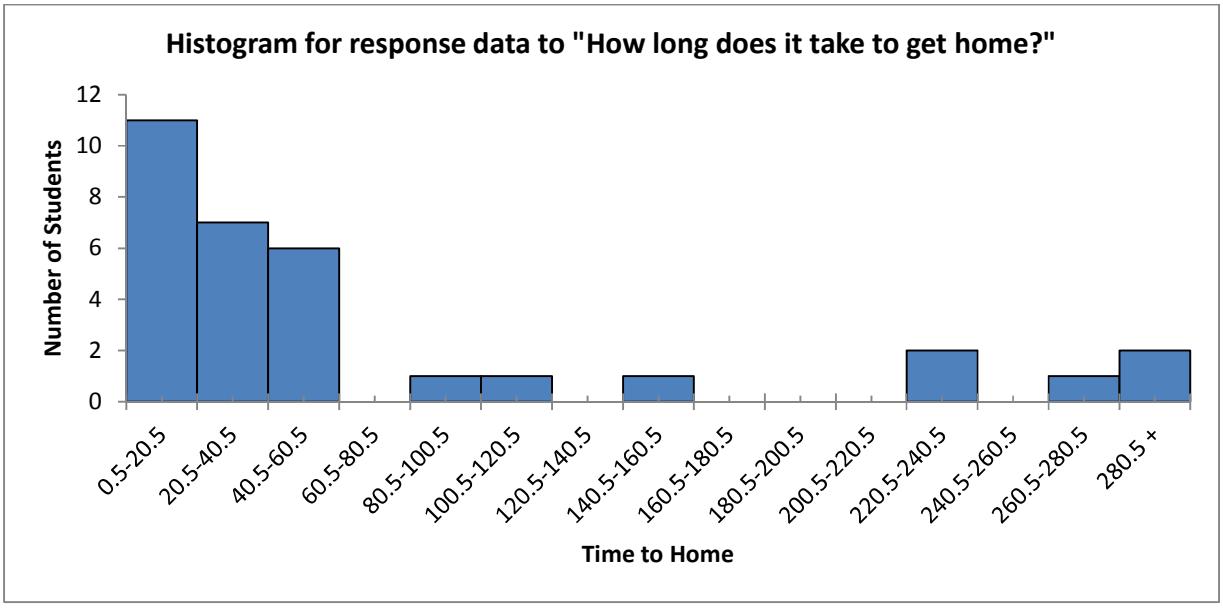
Many of these values are the same as the limits (e.g., 20, 40, etc.) and so we must set up class boundaries in order to follow rule 2. Those boundaries will be: 0.5 – 20.5, 20.5 – 40.5, . . ., which keeps the class width at 20 units.

Lastly, the class midpoints are computed by adding the boundaries of each class and dividing by 2 (aka, their average), which gives us 10.5, 30.5, . . .

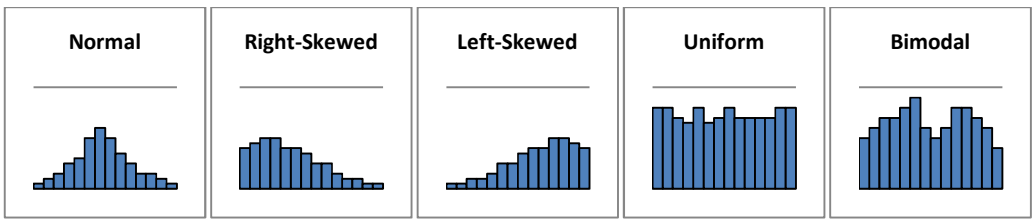
This table shows the distributions for frequency, cumulative frequency, and relative frequency:

Class	*Limits*		*Boundaries*		*Midpoint*	*Frequency*	*Cumulative Frequency*	*Relative Frequency*
1	0	20	0.5	20.5	10.5	11	11	0.344
2	20	40	20.5	40.5	30.5	7	18	0.219
3	40	60	40.5	60.5	50.5	6	24	0.188
4	60	80	60.5	80.5	70.5	0	24	0.000
5	80	100	80.5	100.5	90.5	1	25	0.031
6	100	120	100.5	120.5	110.5	1	26	0.031
7	120	140	120.5	140.5	130.5	0	26	0.000
8	140	160	140.5	160.5	150.5	1	27	0.031
9	160	180	160.5	180.5	170.5	0	27	0.000
10	180	200	180.5	200.5	190.5	0	27	0.000
11	200	220	200.5	220.5	210.5	0	27	0.000
12	220	240	220.5	240.5	230.5	2	29	0.063
13	240	260	240.5	260.5	250.5	0	29	0.000
14	260	280	260.5	280.5	270.5	1	30	0.031
15	280 +		280.5	<>	<>	2	32	0.063

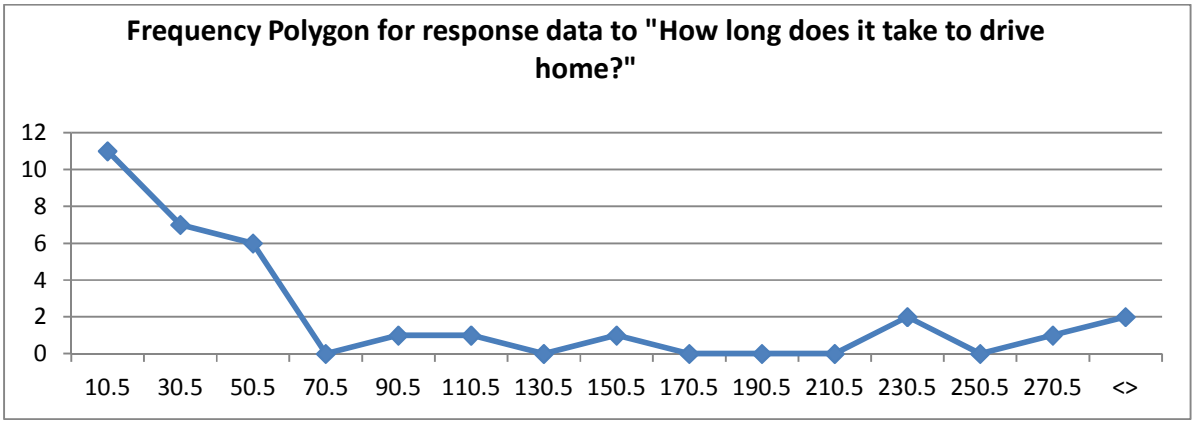
A histogram is a vertical bar graph for frequency data. The histogram for the example dataset is shown below.



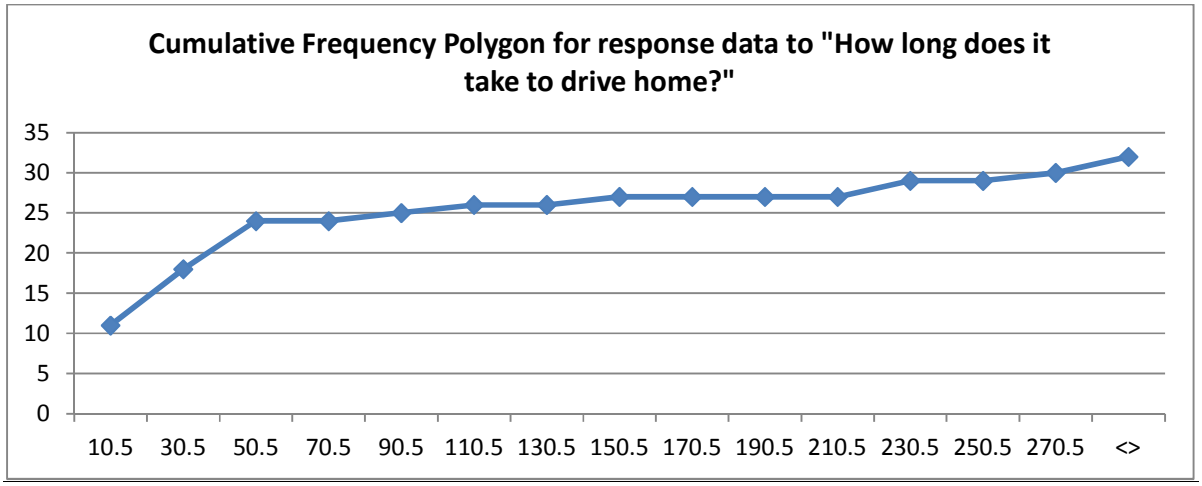
The shape of a distribution is seen in its histogram. The shapes you should know appear here:



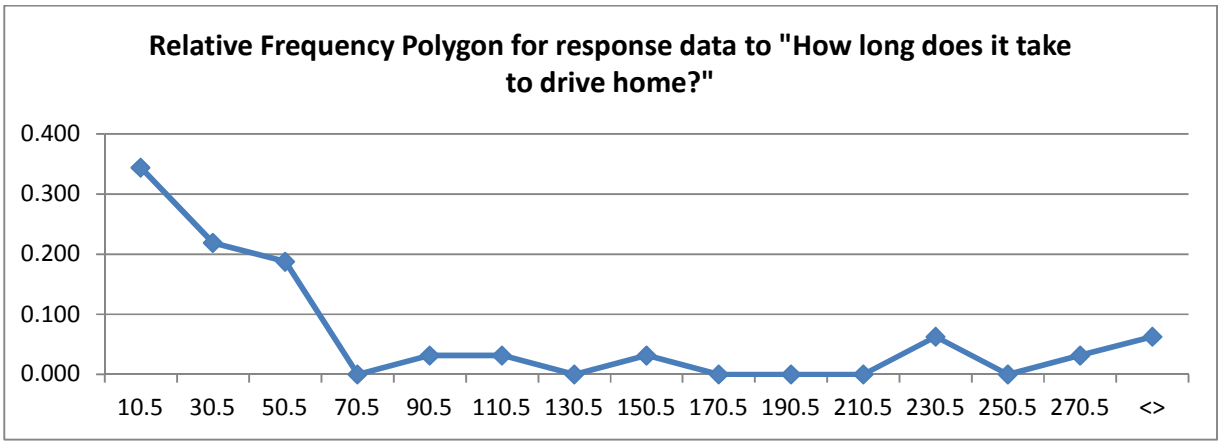
A polygon is a line graph for frequency data. Marks are plotted for each frequency, and they are centered over the midpoint of each class, and line segments are drawn to connect the marks. The frequency, cumulative frequency, and relative frequency polygons for the example dataset are shown below:



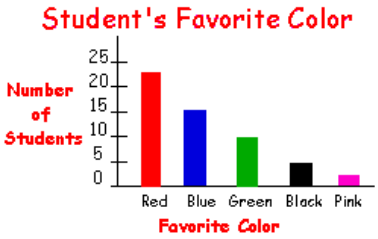
Because of its additive nature, the elevation of a *cumulative* frequency polygon will *never* decrease:

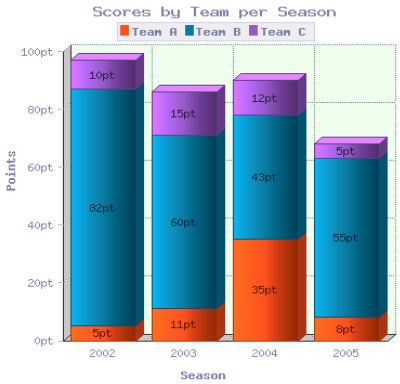


Relative frequency is computed directly from frequency, and so the shapes of those polygons are scaled versions of each other:

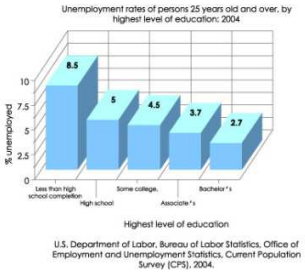


Producing visual displays for qualitative data is typically done with bar graphs. *Bar Graphs* should be straight-forward, 2-dimensional and with non-truncated bars of uniform width in order to avoid misinterpretations. For *nominal* data, bars should be arranged in either descending or ascending order. Here are some examples of some bar graphs done well:

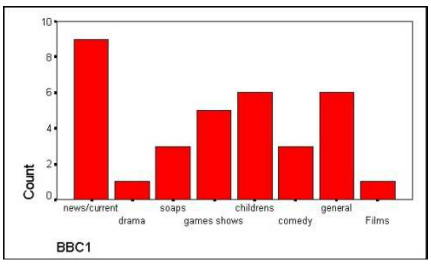




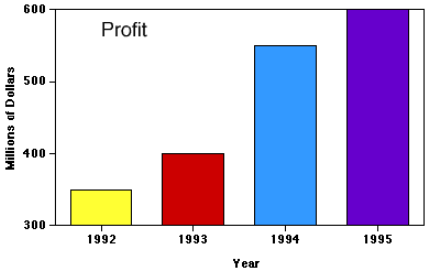
And several that were done poorly:



(3-D enhancement makes the first bar appear larger than it should)



(Bars should be ordered from highest to lowest or lowest to highest)

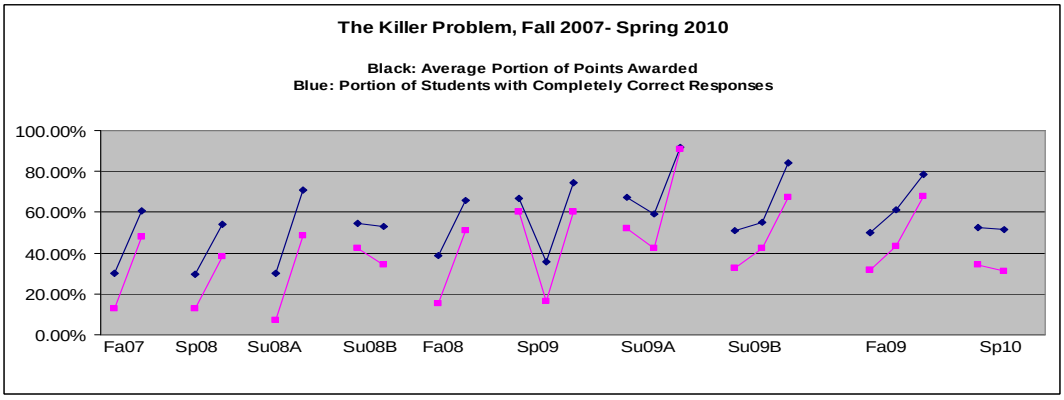


(Truncated bars artificially magnify the differences between bar height)

Note: Pie Charts are generally not recommended because the areas of the wedge-shaped pieces are disproportionate to the actual values.

A time-series display shows data with a chronological sequence. Most often, time-series data is displayed with a dot-plot or polygon. Time will appear as the horizontal axis in order to show how a statistic changes over time.

Here is an example of a time-series display for two distributions:



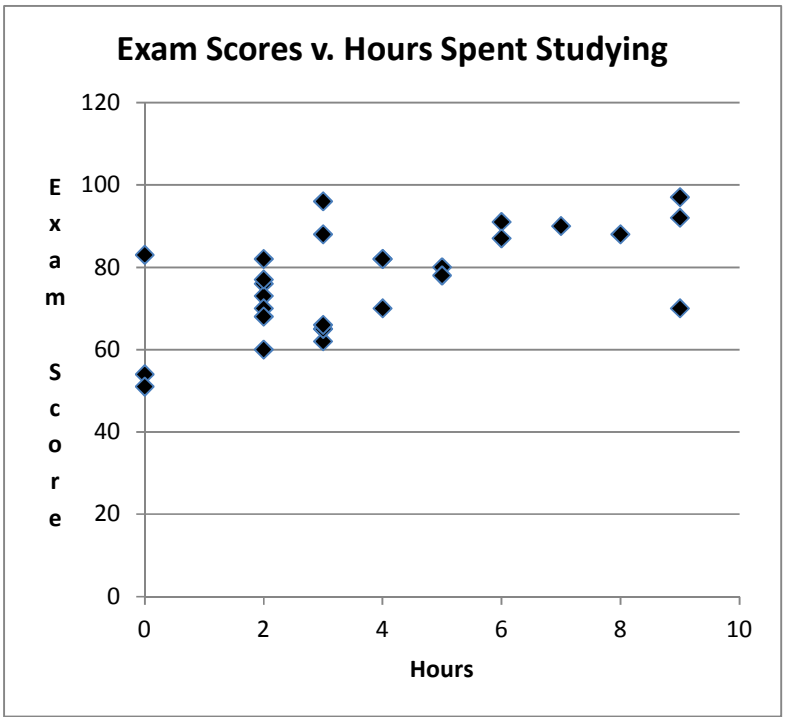
Paired data (or 2-D) are data for which two corresponding values are paired for each data point.

Example: Before an exam, students in a statistics course were asked how many hours they spent studying for the exam. The responses and the exam grades were recorded in the data set shown here

Hours	Exam
0	54
0	51
0	83
2	60
2	76
2	73
2	70
2	77
2	82
2	68
3	96
3	65
3	88
3	62
3	65
3	66
4	70
4	82
4	82
5	80
5	78
6	91
6	87
7	90
8	88
9	97
9	92
9	70

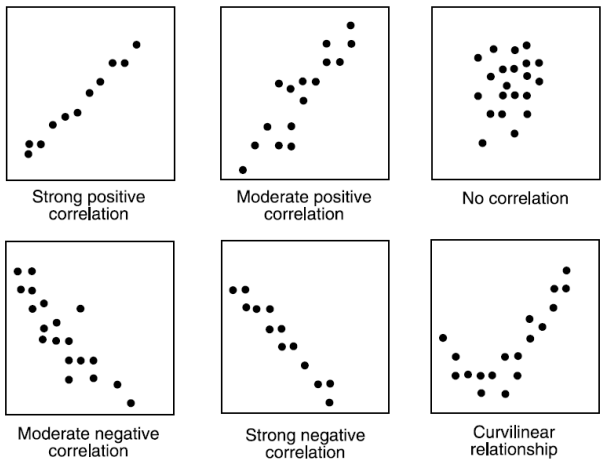
Scatterplots are displays of paired, or 2-dimensional data. The horizontal and vertical axes should be labeled and scaled for each of the variables.

The example data set is shown in a scatterplot below:



Correlation is the term for the type of relationship between two variables in paired data. It will be quantified in a later lecture.

For now, you need to know the difference between positive and negative relationships, between strong, moderate, and weak relationships, and between linear and non-linear (aka, curvilinear) relationships. They are illustrated here:



The last display of data to learn is called a stem-and-leaf plot. This unique display is a virtual bar graph for frequency data that retains the raw data. In histograms and other displays of frequency data, the actual values within the dataset are not apparent in the image. Stem-and-leaf plots use the original data as units for horizontal bars.

First, the values in a dataset being used for a stem-and-leaf plot must all be rounded to the same number of digits and ranked. We will examine the dataset on the right to explain the procedure.

Rounded Age of Math 175 Students, Fall 2012

18.5	20.0	20.6	21.0	22.0	24.0
19.0	20.0	20.8	21.2	22.0	27.0
19.0	20.0	20.8	21.2	22.2	27.0
19.0	20.0	20.8	21.3	22.3	28.1
19.2	20.0	20.8	21.3	23.0	31.7
19.5	20.0	20.9	21.6	23.0	32.0
19.6	20.0	20.9	21.6	23.0	34.0
19.7	20.2	21.0	21.7	23.7	37.8
19.9	20.5	21.0	21.8	23.7	47.9

The values range from 18.5 to 47.9 and must be put into classes of width 1, 2, 5, or 10. It is done in the table below with width 2.

Class			Freq
18.0	-	19.9	9
20.0	-	21.9	27
22.0	-	23.9	9
24.0	-	25.9	1
26.0	-	27.9	2
28.0	-	29.9	1
30.0	-	31.9	1
32.0	-	33.9	1
34.0	-	35.9	1
36.0	-	37.9	1
38.0	-	39.9	0
40.0	-	41.9	0
42.0	-	43.9	0
44.0	-	45.9	0
46.0	-	47.9	1

From here, rather than drawing a histogram, we list the items in each class in such a way as to create the look of horizontal bars. This is done by writing the first digit of each number (1, 2,3, or 4, in our example) in a column. The other columns will contain the remaining digits, and it is imperative that the column width is uniform. When the data are entered, the values create horizontal bars.

The reason that class widths must be 1, 2, 5, or 10 because of our base-ten number system. When done appropriately, the horizontal bar graph appears and the original ranked data is still fully intact.

The finished stem-and-leaf plot is below.

1	8.5	9.0	9.0	9.0	9.2	9.5	9.6	9.7	9.9																			
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.6	0.8	0.8	0.8	0.8	0.9	0.9	1.0	1.0	1.0	1.2	1.2	1.3	1.3	1.6	1.6	1.6	1.8	
2	2.0	2.0	2.2	2.3	3.0	3.0	3.0	3.7	3.7																			
2	4.0																											
2	7.0	7.0																										
2	8.1																											
3	1.7																											
3	2.0																											
3	4.0																											
3	7.8																											
3																												
4																												
4																												
4																												
4	7.9																											

This value, 0.9, really represents 20.9 because it is in the row proceeded by a 2.