Math 122
Introduction to Statistics
Summarizing and Graphing Data

THINGS TO CONSIDER WITH DATA

Distribution: the "shape" of the data. Are the data values clustered around some center value or around many values? Are the data values spread out evenly? Is the data symmetric?

Center: a representative value such as the average, middle, or most common value.

Variation: a measure of how spread out the data is.

Outliers: An outlier is a data value which is significantly different from the other data values.

FREQUENCY DISTRIBUTION

Frequency Distribution: A frequency distribution or frequency table is a table which lists the numbers of data values which fall into specific intervals.

Terminology: Each interval in a frequency distribution is called a *class*. The lowest number in a class is the *lower class limit*. The highest number in a class is the *upper class limit*. The difference between the upper and lower class limits is the *class width.*

Cumulative Frequency: A cumulative frequency distribution displays the total number of data values less than or equal to a value rather than the number of values in an interval.

Relative Frequency Distribution: A relative frequency distribution reports percentages rather than counts.

CONSTRUCTING A FREQUENCY DISTRIBUTION

Range: The range of a data set is the difference between the maximum data value and the minimum data value.

We construct a frequency distribution for this data:

4 5 7 7 7 8 9 10 10 11 11 11 11 11 12 13 13 13 14 18

as we outline the steps invovlved in finding a frequency distribution.

1. Find the range of the data values.
   The range of this data is $18 - 4 = 14$.

2. Decide how many classes you want. This is almost arbitrary. In the next step, we will divide by one less than the desired number of class, so you should keep that in mind.

Whether or not we actually end up with exactly this number of classes depends some on what we pick for our lowest class limit and how we round in the next step.
We will use about 8 classes.

3. Find the class width using width $= \frac{\text{range}}{\text{number of classes-1}}$. If necessary, round for convenience.
Our class width is $\frac{14}{7} = 2$.

4. Select a convenient starting point at or below the minimum data value.
We will use the least value 4 as the starting lower limit.

5. Find the lower class limits by repeatedly adding the class width to the starting point.
By adding 2 to 4 repeatedly, we get these lower limits:

4 6 8 10 12 14 16 18

6. Select upper class limits below next lower class limits.
We will use these upper limits:

5.9 7.9 9.9 11.9 13.9 15.9 17.9 18.9

7. Count.
Counting gives us the numbers in the table on the left below.

Dividing each frequency by 20 (the number of data values) and converting to percentages gives the relative frequency table in the middle. Adding class frequencies gives the cumulative frequency table on the right.

Frequency Table	
Value	Frequency
4-5.9	2
6-7.9	3
8-9.9	2
10-11.9	7
12-13.9	4
14-15.9	1
16-17.9	0
18-19.9	1

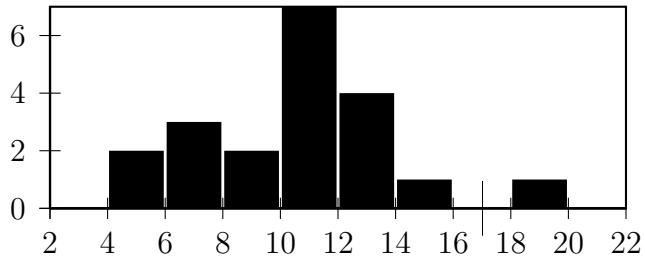
Relative Frequency	
Value	Frequency
4-5.9	10%
6-7.9	15%
8-9.9	10%
10-11.9	35%
12-13.9	20%
14-15.9	5%
16-17.9	0%
18-19.9	5%

Cumulative Frequency	
Value	Frequency
Less than 5.9	2
Less than 7.9	5
Less than 9.9	7
Less than 11.9	14
Less than 13.9	18
Less than 15.9	19
Less than 17.9	19
Less than 19.9	20

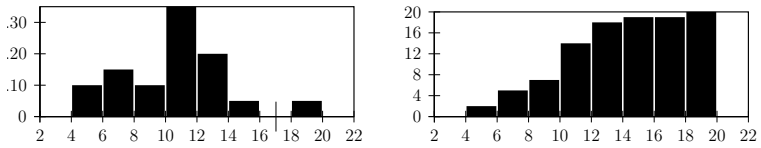
HISTOGRAMS

Histogram: A histogram is a bar graph of a frequency table. The widths of the bars in a histogram are the class widths. The heights of the bars are the frequency counts.

Example: Here is a histogram for the data above:

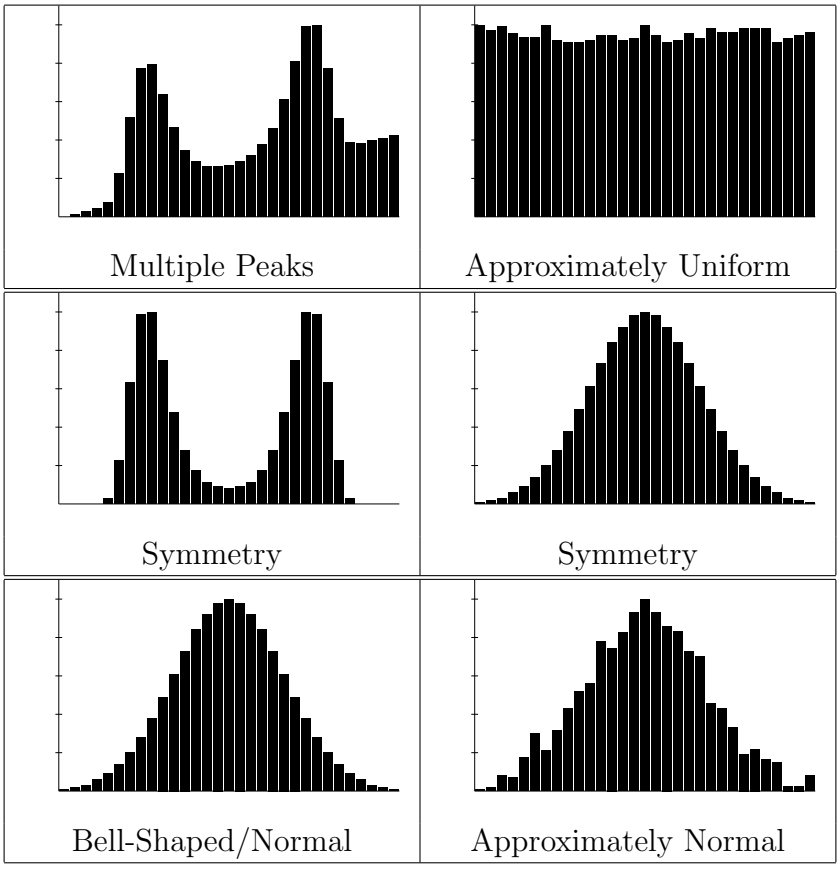


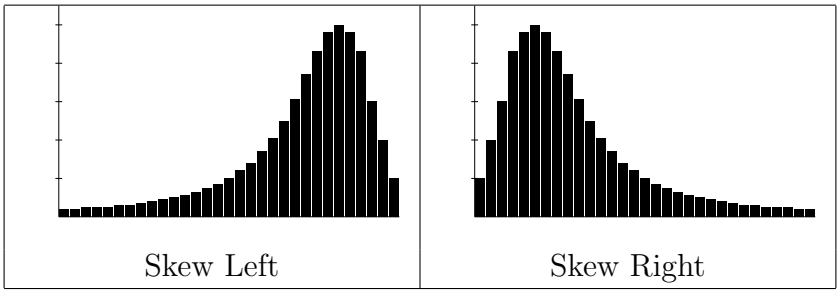
A histogram of the *relative frequency distribution* (on the left below) looks almost identical.



The only difference here is that now the vertical axis is labeled with fractions or percentages rather than counts. We can also draw a histogram of the cumulative distribution (on the right).

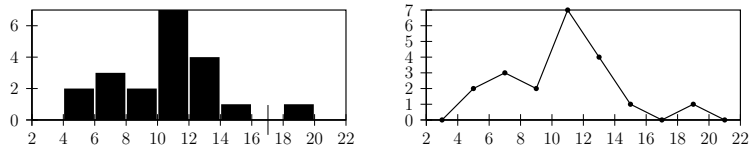
THINGS TO LOOK FOR IN HISTOGRAMS



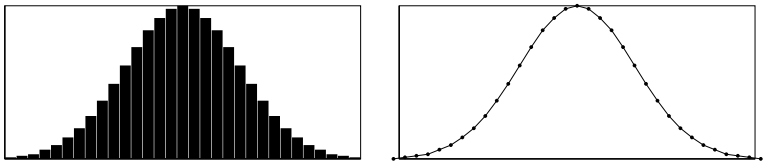


OTHER STATISTICAL GRAPHS

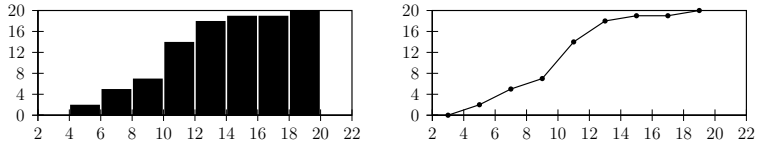
Frequency Polygon: A frequency polygon is a graph of data from a frequency distribution which uses a series of line segments connecting points rather than vertical bars like a histogram. The points which the line segments connect are the center points of the tops of the histogram bars. Here are the histogram and frequency polygon for the example data above:



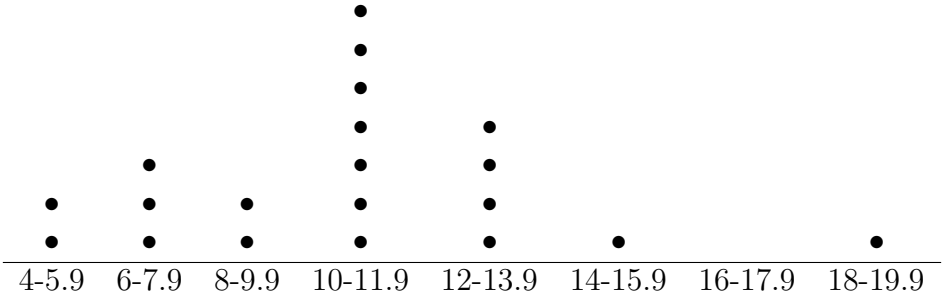
Here are the histogram and frequency polygon of some bell shaped data. Note how with more classes the frequency polygon begins to look like a smooth curve.



Ogive: A line graph for a cumulative frequency histogram is called an ogive. Here is a cumulative frequency histogram accompanied by the corresponding ogive:



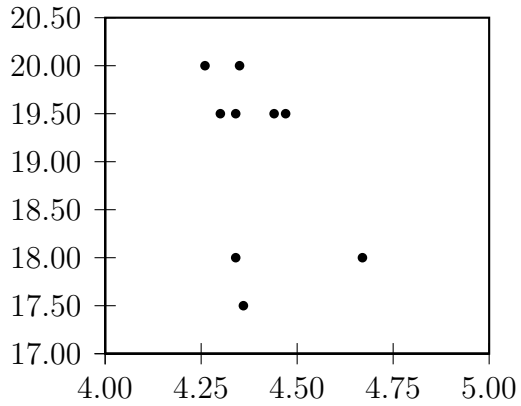
Dot Plot: Dot plots are easy to construct by hand. Here is a dot plot of the data above.



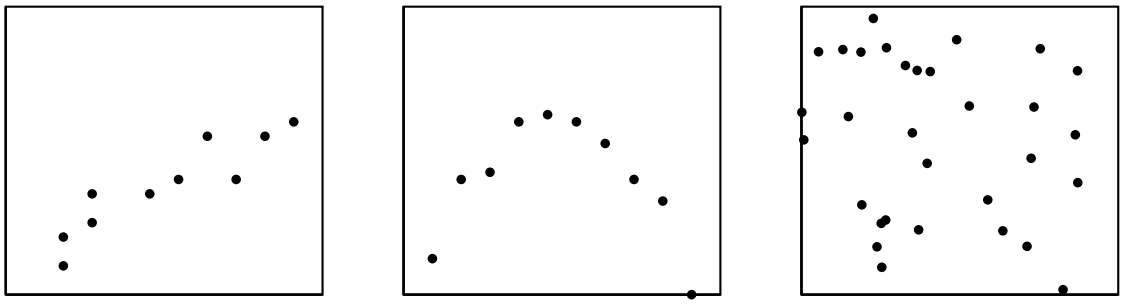
Scatter Plot: A scatter plot of data which comes is pairs graphs each pair of data as a point on a plane. For example here are the 30m times (in seconds) and vertical jump measurements (in inches) of several female athletes:

30m	4.35	4.75	4.47	4.34	4.89	4.36	4.30	4.44	4.67	4.26	4.34
Vertical	20.0	14.5	19.5	18.0	12.5	17.5	19.5	19.5	18.0	20.0	19.5

Here is scatter plot of the same data. The horizontal axis is 30m time, and the vertical axis is vertical jump:



We can look for patterns in scatter plots. In the plot below on the left, the dots seem to almost follow a line up and to the right. This scatter plot depicts a *linear correlation*. In the middle scatter plot, there is a pattern, but it is not linear. The plot on the left exhibits no apparent patter.

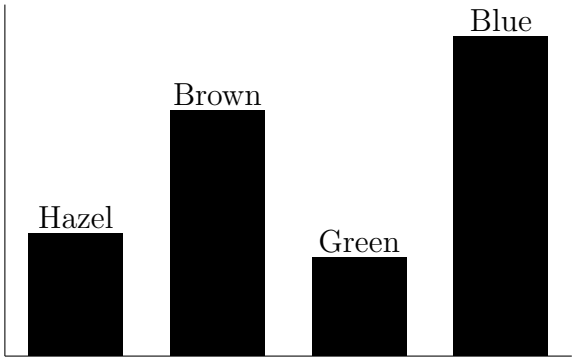


CATEGORICAL DATA

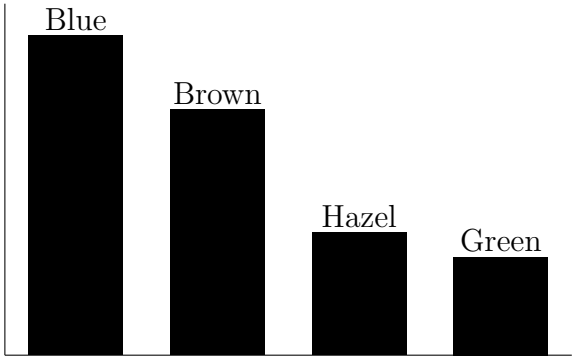
Pareto Chart: We can draw bar charts or histograms of categorical data also. For example, a survey of 64 college students about eye color resulted in this data:

Color	Frequency
Hazel	10
Brown	20
Green	8
Blue	26

Here is a bar chart of this data:



A natural question is what order the bars should be in. To avoid bias in choosing bar order, we usually order the bars from tallest to lowest. This is called a Pareto Chart.



Pie Chart: We can also draw a pie chart of categorical data. Each class or category corresponds to a "slice" of the pie. The size of the slice indicates the size of that category compared to the other categories.