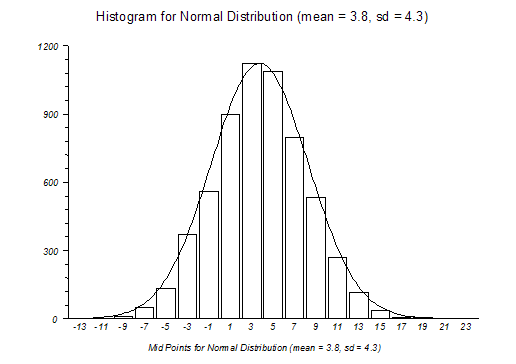
**Topic 7**

**Distributions**

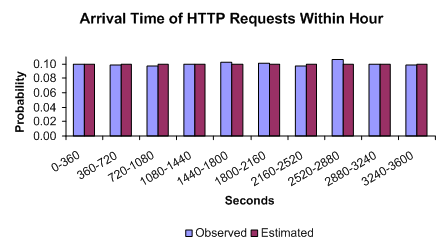
When events occur, they occur with some predictability. Not so surprising, this predictability has ‘shape,’ and we call the shape **distribution** of the data.

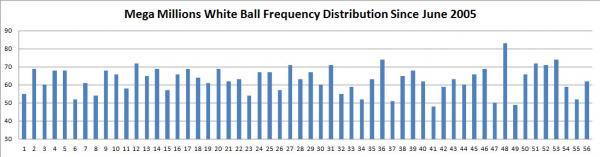
What is occurring is that for most physical systems, some events happen more than others, so in a frequency graph, such as a bar graph or histogram, these events that happen more often are taller than events that happen less often. This is what gives distributions their shape. The most common example of this is the bell curve (or Normal distribution, or Gaussian distribution)



Note here that the numbers between 3 and 5 are the most common events.

Sometimes all events have equal probability. This is called a Uniform Distribution. A simple example of this might be the roll of a die. All numbers, 1 through 6, have equal probability of occurring. It is a fairly common distribution.





Certain types of distributions **occur over, and over, and over again in nature**. Some are more common than others. The Normal curve shown first is by far the most common. The Uniform distribution shown above is another. What others are there?

First, lets talk about some *features of distributions*:

Discrete or Continuous

* Discrete distributions involve variables that can take on discrete values.
  + Bournoulli, Binomial, Poisson, Geometric, Hypergeometric, Skellam, …
* Continuous distributions involve variables that take on a range of values.
  + Normal, Lognormal, Exponential, Uniform, Pareto, Weibull, …

Finite or infinite domain (support)

* Finite domain start and end at finite values
* Infinite domain start or end (or both) at infinity

These factors are important when either using a distribution to analyze your data.

* Select a distribution based on knowledge of variables
* Fit a distribution to the data

**Very often we decide what a distribution is by looking at a visual construction based on the frequencies of the data values.**

When looking at a data distribution, we look for **4 things.**

1. **Center** estimation
2. The **spread**, or variability
3. **3 Shape** aspects
   1. *Symmetric*
   2. *Skewed* (to right or to left)
   3. *Peaked* (kurtosis)
4. **Multimodality**
   1. *Multi-modal* distributions – more than one distribution combined in the graph
   2. *Ouliers* – data points from another distribution, noticeably far from the center

When viewing distributions, we have **two more graphing tools** (and one modification).

1. **Stemplot** (aka stem and leaf plot)
   1. Ordinal or continuous data
   2. Small data sets
   3. Leaves (or stems in book) represent a 10th power (e.g. 50, for 51, 52, 54, …)
   4. Modify to achieve 5-15 lines (stem in book, leaves elsewhere)
   5. The shape o f the plot estimates the distribution
2. **Histogram** 
   1. Data is ordinal or continuous data
   2. Large data sets
   3. Bars represent a range
   4. The height of a bar is the count for that range
   5. The shape o f the plot estimates the distribution
3. Distribution Comparison plots
   1. Side-by-side stemplot
   2. Population plot