Properties of Distribution

The distribution has certain properties that we can use to describe the look of the distribution. Usually, these properties can be easily described in very general terms from a histogram, boxplot, or stemplot.

The properties are typically:

- groupings
  Are the values of the variables closely grouped together?
  Is there one or more values of the variable that is far away from the other values of the variables?
- overall shape
  How many peaks are there?
  Is the distribution periodic (does it repeat itself)?
  Is the distribution symmetric (is one side like the other)?
- center
  Is there essentially one peak in the distribution?
  Where does this peak occur?
  We can describe the center of the distribution in terms of the average value, or the middle value.
- spread
  How wide is the distribution?

When describing a distribution:

- Look for outliers and try to explain why they are outliers.
- Concentrate on main feature, ignoring the outliers if you have been able to explain their existence to your satisfaction.
- Look for rough features, not precise variations.

These general descriptions can be converted to more formal descriptions using the following terms:

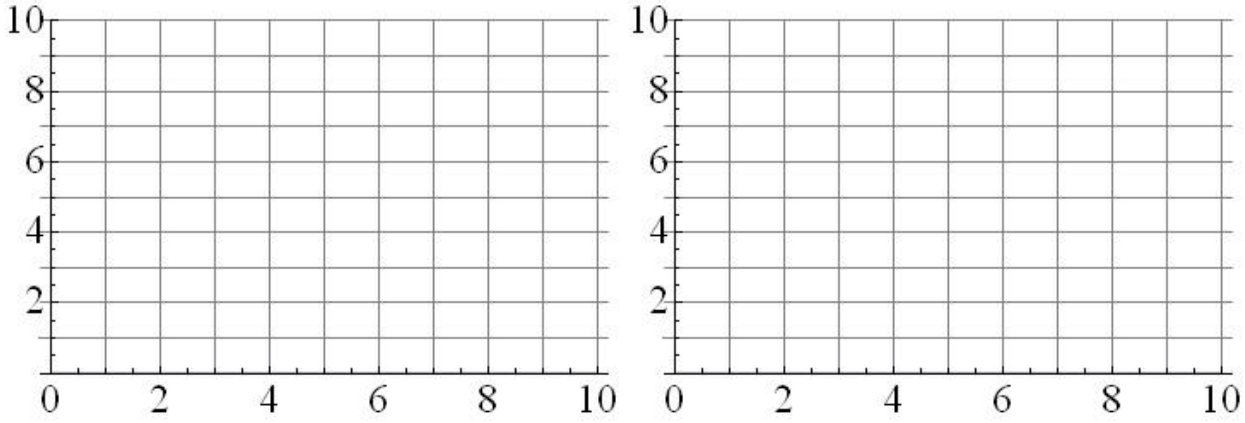
- An outlier is a value of a variable that falls outside the overall pattern.
- A distribution is symmetric if the left and right sides of the distribution are approximately mirror images of each other.
- A distribution is skewed to the right if the right side (or larger values) of the distribution extends much farther out than the left side (or smaller values). The tail extends further to the right.
- A distribution is skewed to the left if the left side (or smaller values) of the distribution extends much farther out than the right side (or larger values). The tail extends further to the left.
- The **center of a distribution** can be measured by calculating the mean (sometimes called the average) or the median (sometimes called the middle) of the distribution.
  Mean $= \bar{x} = \dfrac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$.
  Median: The number such that half the observations are smaller and half are larger.
  1. Arrange the observations from smallest to largest.
  2. If n is odd, the median is the observation in the middle of the list.
  3. If n is even, the median is the average of the two center observations in the list.
- The **spread of a distribution** is described by quartiles and standard deviation.

Example: Seeding Clouds from http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html.

Clouds were randomly seeded or not with silver nitrate. Rainfall amounts were recorded from the clouds. The purpose of the experiment was to determine if cloud seeding increases rainfall (notice that this gives us two distributions that we can compare).

Unseeded Clouds	Seeded Clouds
7.09	7.92
6.72	7.44
5.92	7.41
5.84	6.89
5.77	6.56
5.50	6.19
5.09	6.06
5.00	5.81
4.55	5.71
4.47	5.62
4.40	5.62
4.23	5.54
3.86	5.49
3.72	5.30
3.60	5.29
3.37	4.86
3.35	4.78
3.27	4.77
3.26	4.75
3.19	4.53
3.08	3.70
2.85	3.49
2.44	3.45
1.59	2.86
1.59	2.04
0	1.41

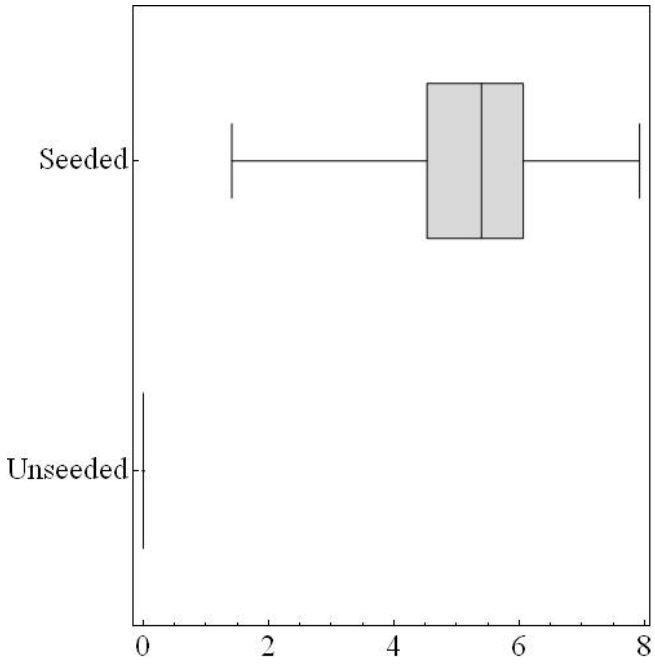
Histograms



The Five Number Summary

Unseeded: Minimum =	Q_1 =	Median =	Q_3 =	Maximum =
Seeded: Minimum = 1.41	$Q_1 = 4.53$	Median = 5.395	$Q_3 = 6.06$	Maximum = 7.92

Boxplot



Variance and Standard Deviation

- unseeded clouds mean = 3.99, standard deviation = 1.64
- seeded clouds mean = 5.13, standard deviation = 1.60

Using either the histograms or boxplots we created we can describe the distributions as follows:

- Seeded generated more rainfall than unseeded (from boxplot, the middle 50% of measurements is higher for seeded).
- The seeded looks more symmetric than the unseeded, however both are skewed to the left. The unseeded would NOT be well described by mean and standard deviation.
- It is possible that the 0 is an outlier in the unseeded distribution. If we exclude 0, then the mean would be 4.15 and standard deviation 1.45 (mean and standard deviation are sensitive to outliers). The quartiles change, but to a lesser extent.

Normal Distribution

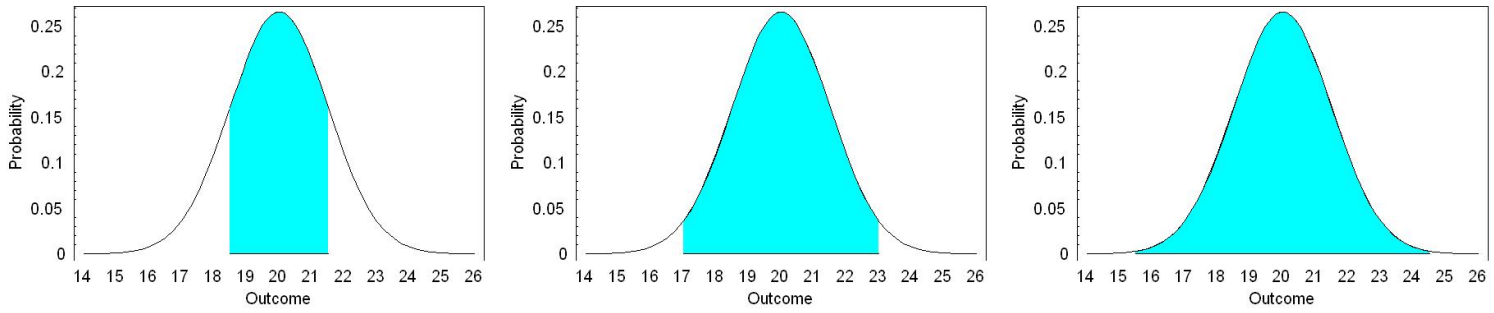
An important distribution is a normal distribution (sometimes called the bell curve), which has the properties:

- The tails of the curve fall off rapidly.
- The distribution is symmetric.
- The mean lies at the center of symmetry.
- The mean is the same as the median (true for any symmetric distribution).
- Since it is symmetric, it is well described by the mean and standard deviation.
- In fact, a normal distribution is completely described by the mean and standard deviation.
- The spread is entirely measured by the standard deviation.
- Where the curve changes from curving down to curving up is one standard deviation away from the mean.
- It has no outliers.

The 68-95-99.7 Rule

All normal distributions behave in certain regular ways. For example,

- the probability that a measurement falls within 1 standard deviation of the mean is 0.68.
- the probability that a measurement falls within 2 standard deviations of the mean is 0.95.
- the probability that a measurement falls within 3 standard deviations of the mean is 0.997.



The above graphs show the 68-95-99.7 rule for a normal distribution with mean 20 and standard deviation of 1.5.

Summary

- graphical description of distributions include:
  - histograms
  - stemplots
  - boxplots
- two methods of numerically describing center and spread of a distribution:
  - five number summary (min, first quartile, median, third quartile, max), leads to boxplots
  - mean and standard deviation, usually included with a histogram
- the mean and standard deviation are greatly affected by outliers.
- the mean and standard deviation do not display the skewness of the distribution.
- the mean and standard deviation are best used for symmetric distributions without outliers.
- skewed distributions are best described by the five number summary, since the boxplot easily displays information about the skewness of the distribution.