Definitions

Individuals	Students in Survey of Math	States
–**Variables**	–final exam score	–sales tax
	–hair colour	–unemployment rate
	–height	–population
	–final grade	–unemployment rate

A set of n individuals would have the n observation values for a given variable: $x_1, x_2, x_3, \ldots, x_n$.

A distribution of a variable tells us what values the variable takes and how often it takes these values.

A distribution can be displayed by a

- stemplot (good for small numbers of individuals)
- histogram (good for large or small numbers of individuals)
- boxplots (good if the distribution is large and not symmetric).

We are interested in describing individual distributions and comparing multiple distributions.

Constructing Stemplots

Stemplots work very well for small data sets that have values between 0 and 100. So they are useful for "back-of-the-envelope" descriptions of test scores!

1. Separate each observation into a stem consisting of all but the final digit, and a leaf which is the final digit (leaves only have one digit!).
2. Write the stems in a vertical column with smallest at top. Draw a vertical line to the right of the column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Test Score	Stem	Leaf
75	7	5
64	6	4
86	8	6
72	7	2
77	7	7
90	9	0
29	2	9
86	8	6
74	7	4
88	8	8
91	9	1

Original Distribution

0	
1	
2	9
3	
4	
5	
6	4
7	2457
8	668
9	01

Stemplot

2	9
6	4
7	2457
8	668
9	01

Not the Stemplot

On the right, you would lose the immediate recognition of 29 as a data element that is separate from the others. Stemplots work well with small to mid-size data sets, and they contain all the data in the data set.

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 <u>outlier</u> is a value of a variable that falls outside the overall pattern.
- A distribution is <u>symmetric</u> if the left and right sides of the distribution are approximately mirror images of each other.
- A distribution is <u>skewed to the right</u> if the right side (or larger values) of the distribution extends much farther out than the left side (or smaller values).
- A distribution is <u>skewed to the left</u> if the left side (or smaller values) of the distribution extends much farther out than the right side (or larger values).
- The **center of a distribution** can be measured by calculating the <u>mean</u> (sometimes called the average) or the <u>median</u> (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

This section is taken from http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html.

Chambers, Cleveland, Kleiner, and Tukey. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group, Belmont, CA, 351. Original Source: Simpson, Alsen, and Eden. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics* 17, 161-166.

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

NOTE: For reasons that don't concern us, the investigators took the logarithm of the measured data. I have done that for us. The reason this was done is based some more advanced ideas from statistics than we cover in this class.

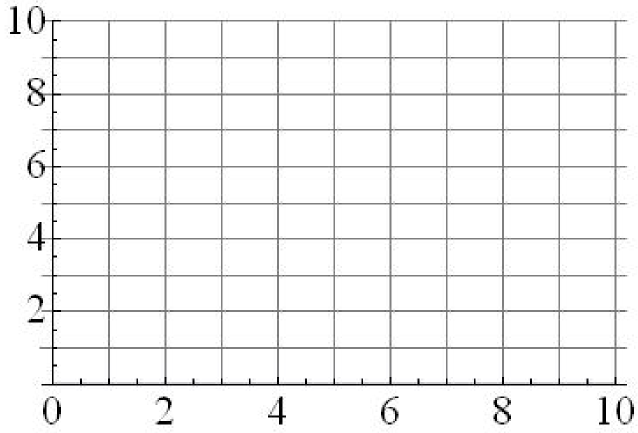
Constructing Histograms

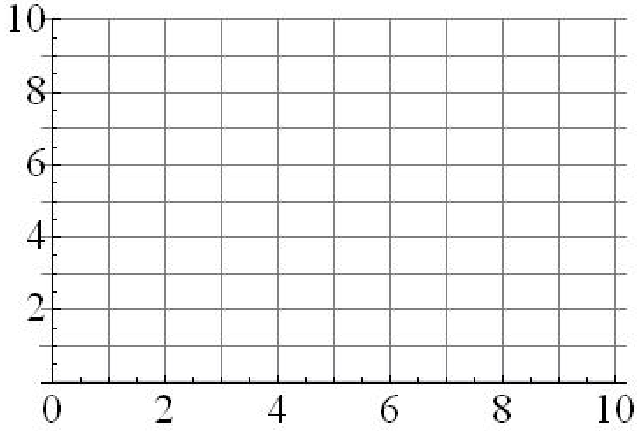
A histogram is a graphical representation of data: the height of the bars in a histogram represent the number of times data was collected in the range at the base of the bar over a sample, or population, or even a number of different samples!

If the data set given represents choices from a list, then it may be more appropriate to create a histogram with the number of times each number is listed. In that case, we would label the base of the histogram bars with the number it represents.

Histograms work well for large data sets, but you usually have to take some care to choose the bar width (sometimes called bin width) wisely to adequately convey the information in your data set to your audience. Check out histogram for Old Faithful on the course webpage to see this!

Example Create the Histogram for the Unseeded and Seeded distributions.





The Five Number Summary & Quartiles

We use quartiles to measure the spread of a distribution. They are similar to the median which is used to measure center.

The first quartile Q_1 is the number such that 25% of the observations are below it and 75% of the observations are above it.

The Median is the second quartile, the number such 50% of the observations are below it and 50% of the observations are above it.

The third quartile Q_3 is the number such that 75% of the observations are below it and 25% of the observations are above it.

The quartiles break the distribution up into quarters (hence the name).

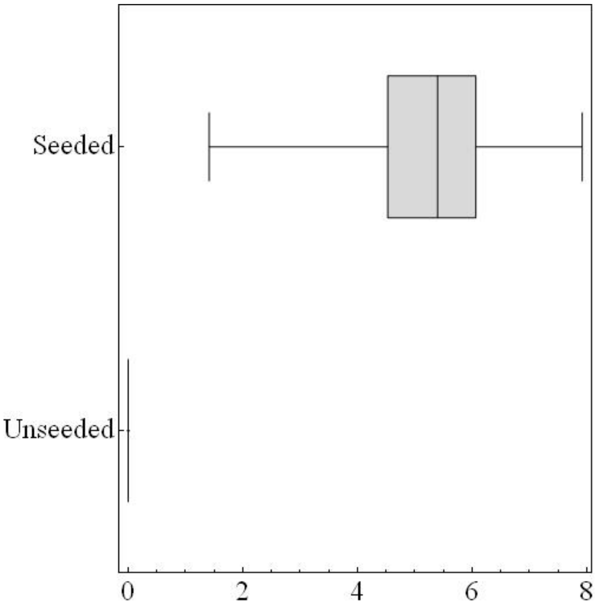
We can combine these numbers with the minimum and maximum of the distribution to create the five-number summary of the distribution, which describes the distribution in some detail (center, spread).

$$\text{Minimum} \quad Q_1 \quad \text{Median} \quad Q_3 \quad \text{Maximum}$$

Example Find the five-number summary for the seeded and unseeded clouds distribution.

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

We can construct a box plot which visually represents the five number summary.



The box represents the spread of the middle half of the distribution; notice the median is NOT in the middle of the box.

In this representation, the horizontal distances are important, and the vertical distances meaningless.

Variance and Standard Deviation

There are other ways of describing the center and spread of a data set, which are more often reported. They involve the mean and standard deviation.

The standard deviation is a measure of how far the observations are from their mean. The standard deviation is the square of the variance.

$$\text{Mean } = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Variance } = s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

$$\text{Standard Deviation } = s = \sqrt{s^2}$$

You can use a calculator or Excel to determine the standard deviation if you need it, so don't worry about memorizing this formula.

Note that the standard deviation has the same units as the mean. Since we are squaring in the standard deviation, information about skewness is completely lost–so the mean and standard deviation should only be used to describe symmetric distributions.

Example Using the formulas, we find for the

- 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 say:

- 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. The unseeded is skewed the the right. 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.

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.