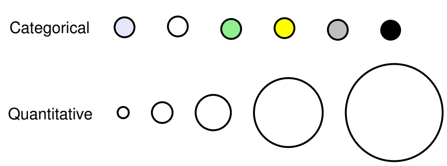
Math 324 Lecture 3
Dr Bolstad
http://math324sfsu.bmbolstad.com

Variables

- **variable** a characteristic which may change between different objects in a population.

  Categorical Variable a variable that places each individual into one of a specified number of categories. This may or may not be ordered. eg high or low, eye color, gender

  Quantitative Variable a variable which takes numerical values. Something for which we could add or average. eg heights, volumes, weights



Two types of statistical analysis

- **Descriptive Statistics** sometimes called *exploratory data analysis* is about investigating the measurements of the variables in your dataset. This could involve each variable by itself or relationships between variables.

- **Inferential Statistics** sometimes called *confirmatory data analysis* use information from our sample to draw conclusions about the population from which it was drawn.

What will we do in this class?

Mostly inferential statistics. But before we really get to that we will spend considerable time learning about probability. Why? Because probability is the branch of mathematics which deals with uncertainties. And we must first learn about the uncertainties that might arise when taking a sample from a population. But, I'm getting ahead of myself we will start looking at probability in lectures 3 or 4 lectures from now. Instead, we will spend a few lectures on some topics in descriptive statistics.

Stem and Leaf plot

The stem and leaf plot is useful for getting a quick picture of the shape of the distribution while also including the actual numerical values into the graph. It is typically most useful when you have a small number of observations. Each observation is separated into two parts, a *stem* and a *leaf*. Stems consist of all but the least significant (most right) digit. The leafs contain only a single digit. The stem and leaf plot is created by first writing the stems in a vertical column and then the leafs in increasing order to the right of the stems. Usually a vertical line is drawn to separate the stems from the leafs.

An example

Suppose we have the following data:

```
 -0.9 -2.0  0.1  0.1  0.6  0.4  0.4  1.1  0.4 -0.5  0.6
  1.0 -2.6  0.7 -0.6  0.7  1.6  0.7  1.4  0.9  2.0  1.6
  1.1  2.3  1.8
```

then stem and leaf plot will be

```
  -2 | 60
  -1 |
  -0 | 965
   0 | 11444667779
   1 | 0114668
   2 | 03
```

The stems can also be split, This is particularly useful with large datasets. For our example

```
-2 | 6
-2 | 0
-1 |
-1 |
-0 | 965
-0 |
 0 | 11444
 0 | 667779
 1 | 0114
 1 | 668
 2 | 03
```

where we split the stems so that the leaves 0, 1, 2, 3, 4 fall on one stem and 5, 6, 7, 8, 9 fall on the other stem. With observed variables with many digits you might find it easier to round the numbers before making the stem and leaf plot.

Histograms

Histograms are more versatile than stem and leaf plots. A histogram breaks the range of values of a variable into intervals and displays only counts of observations that fall into each interval. This makes them much easier to use when there are a large number of observations. The intervals for a histogram should be of equal length.

Steps to drawing a histogram

1. Divide the range of the data into equal length classes

2. Count the number of individuals/units/observations in each class. A table of these values is called a frequency table.

3. Draw the histogram. The height of each bar represents the number of observations in the class. The width represents the length of the class.

An Example

The data

```
0.75 2.04 1.99 1.59 0.84 1.38 0.68 1.22 1.31 0.43
1.55 1.97 1.76 1.05 1.19 0.78 0.21 1.25 1.03 1.23
0.68 2.39 1.81 0.33 1.05
```

A frequency table

Class	Freq
0.00 < x <= 0.50	1
0.25 < x <= 0.50	2
0.50 < x <= 0.75	3
0.75 < x <= 1.00	2
1.00 < x <= 1.25	6
1.25 < x <= 1.50	3
1.50 < x <= 1.75	2
1.75 < x <= 2.00	4
2.00 < x <= 2.25	1
2.25 < x <= 2.50	1

The resultant histogram