Sample data is usually too big to easily digest by simply looking at the actual data, so it is necessary to be familiar with methods of making the main features of a dataset understandable. We will review both *numerical* ways to summarize a dataset and *graphical* ways to visualize it. Data comes in several types; for this review these two categories are useful:

(1) *Discrete* (unordered categorical, ordered categorical, or numerical), and

(2) *Numerical* (discrete or continuous).

These categories are not perfect, as it is possible for data to be both discrete and numerical. In this case which type of method to use will depend on the data set; sometimes both types may be helpful.

1 Discrete Variables

Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. The questionnaire also asked for gender information. Data and story from `http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html`. The data looks something like this...

girl	Urban	Sports
girl	Suburban	Grades
boy	Rural	Popular
...		

1.1 A numerical summary: Contigency Tables

A contigency table (or frequency table) simply records the count for each category. It can be done either for variables by themselves or together.

For Goals and Gender, separately:

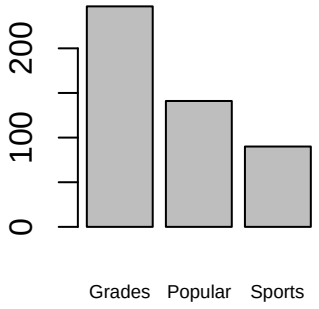
```
Goals
 Grades Popular  Sports
    247     141      90

Gender
 boy girl
 227  251
```

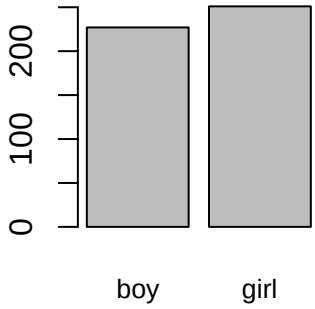
For Goals and Gender, together (a two-way contigency table)

```
         Gender
Goals     boy girl
  Grades  117  130
  Popular  50   91
  Sports   60   30
```

1.2 A graphical summary: Bar Graphs

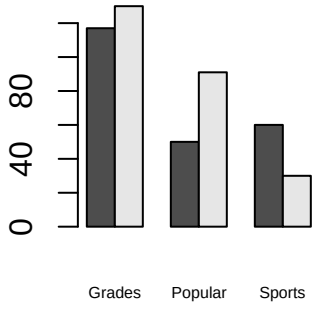
For individual variables, a bar graph plots a bar for each category, with the height equal to the number in that category.

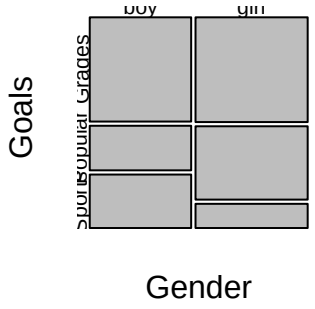




Would more students rather be popular or be good at sports?

For viewing two variables together, there are two common ways; the first is a bar plot, with each bar split into the desired categories; the second is a mosaic plot, where the area corresponds to the frequency.





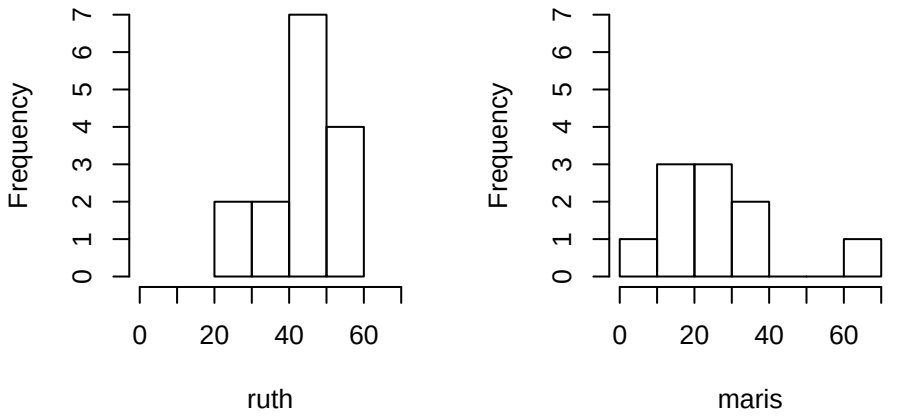
Would more boys rather be popular or good at sports?
Would more girls rather be popular or good at sports?
Do you think it's likely that Gender and Goals are independent? Why or why not?

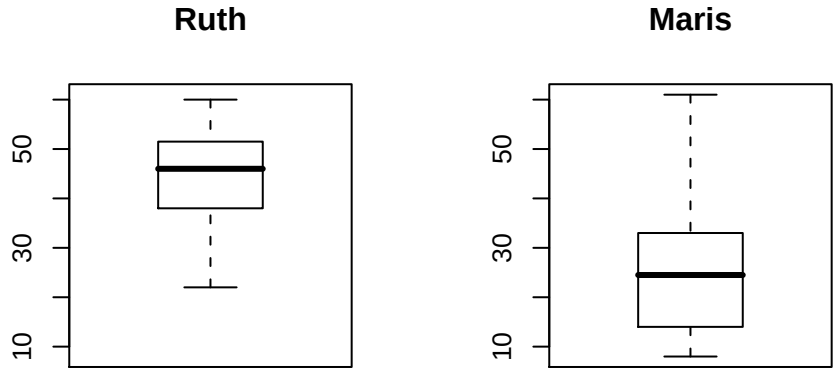
2 Single Numerical Variables

2.1 Graphical Methods

For univariate numerical data, there are three graphical methods we'll look at for showing what the data looks like, stem-leaf plots, histograms, and box plots.

```
> ruth

 [1] 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

> maris

 [1]  8 13 14 16 23 26 28 33 39 61
```





More complex boxplots (with outliers)...

2.2 Numerical Summary Statistics

The most common methods of numerically summarizing a data set are the sample mean, sample standard deviation, and the five number summary, which is the minimum, first quartile, median, third quartile, and maximum.

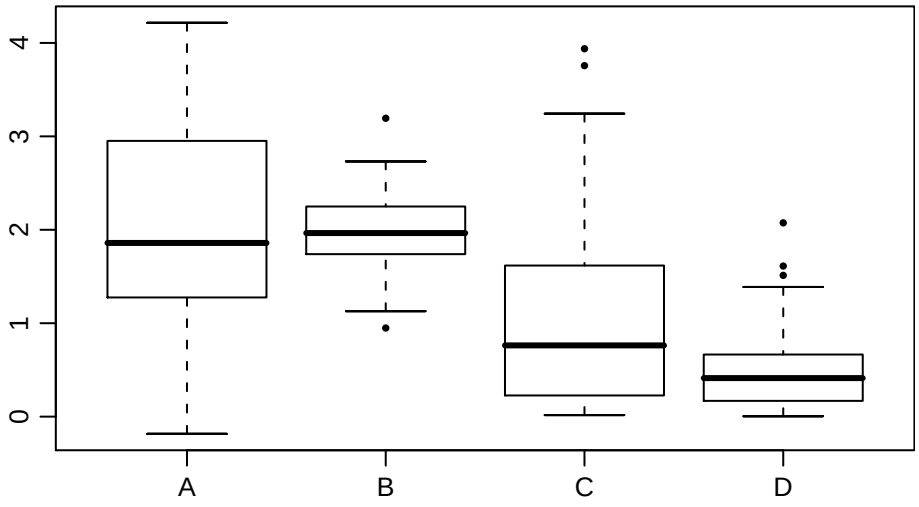
The sample mean is a simple average, and is usually denoted with a bar over the variable:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The sample standard deviation is computed similarly to the population standard deviation, except by using the actual data, though the divisor is $n-1$, not n, as you might expect. The usual notation is s, where

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Compare the sample mean and standard deviation with the boxplots for these made-up datasets:



for A, mean=2.06, sd=1.07
for B, mean=2, sd=0.41
for C, mean=1.04, sd=0.99
for D, mean=0.53, sd=0.46

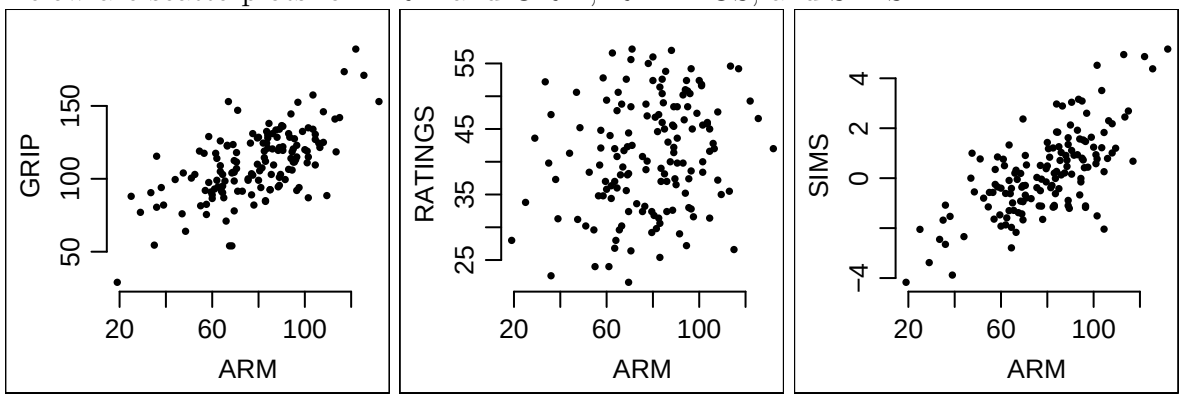
3 Two or More Numerical Variables

Does physical strength matter for physically demanding jobs? To answer this, 147 individuals working as electricians, construction workers, auto mechanics, and other physically demanding jobs were measured in both strength and job performance. Their strength was measured with two standard strength tests, one for arm strength (ARM), and one for grip strength (GRIP). For job performance, they were first evaluated by their employer and given a score (RATINGS), and also evaluated by simulating using a wrench (SIMS). Data and story from
`http://www.ruf.rice.edu/~lane/case_studies/physical_strength/index.html`.

3.1 Graphical: Scatterplot

We can look at each variable individually by looking at stem-leaf plots, histograms, and boxplots. To look at two variables together, we need a new tool, called the scatterplot. It simply puts the variables on the two axes, and plots each data point. This helps us to determine how the two variables are related. Are they linearly related, or is there a more complex pattern? If linear, is the relationship positive or negative? How strong is the relationship?

Below are scatterplots for ARM and GRIP, RATINGS, and SIMS.



Is arm strength related to grip strength? To the rating of the employer? To their score on the simulation? How, and how strongly?

3.2 Numerical: Correlation

A numerical way to measure the degree of linear relationship between two predictors is the sample correlation, often denoted with r. It is calculated in a similar way to the population version of the correlation, except the sample data is used. That is,

$$r(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

Like the population correlation, it is always between -1 and 1. Values closer to 0 mean the relationship is weaker, and the sign corresponds to the direction of the relationship.

The correlation for the three relationships plotted above are 0.63, 0.22, and 0.69, respectively. Do these numbers agree with what the plots seemed to say?

It is important to remember that correlation only measures *linear* relationships.
Compare the following scatterplots with their sample correlations.