For *univariate*, *numeric* data sets we learn the following terminology for describing the data:

Center Roughly (we'll be precise soon) the center of the data refers to the "middle" value. This can be defined by where a graph of a data set visually balances or where a graph splits into equal area pieces.

Why is understanding the center important? Because it gives us the best single number summary of a data set. (How many times have you said "on average..." to refer to some set of numbers.)

Spread The spread of a data set refers roughly to how wide is the data relative to its center. The *range* is the difference between the maximum value and the minimum value in the data, or simply those two numbers. Although the range measures spread, it is generally not used for this, as a single observation can throw off the interpretation. Rather, we use something like the range of the middle 50% of the data or the upcoming standard deviation, which is harder to visualize from a graph.

Why is spread important? Because the center is often not enough to make wise decisions: Do you think anybody would play the lottery if they knew that on average they would make 1 cent back for every dollar (that's a center)? I don't think so, I think people think they will make somewhere in the range which could be 50,000,000 dollars.

Shape A "bell-shaped" distribution will be a common assumption about data that we can make statistical inference about. As such, we need to know when data is bell shaped, and if it isn't what kind of shape is it.

We have the following descriptions. From a densityplot we have the *modes*, which are the peaks in the density (pronounced ones). A data set can be *unimodal*, *bimodal*, or *multimodal*. A bell shaped data set is unimodal. Peaks are important, because typically that is where most of the area is, and hence most of the data (area = likelihood that data is found there).

For unimodal data sets we have bell-shaped ones which are roughly symmetric. If a data set is non-symmetric then it is called skewed. A skewed right data set looks like someone pulled a bell-shaped density to the right. A skewed-left data set is pulled the other way. See Figure 1 for a bell shaped data set, a skewed left one and a skewed right one.

In this project, you will do the following repeadetedly:

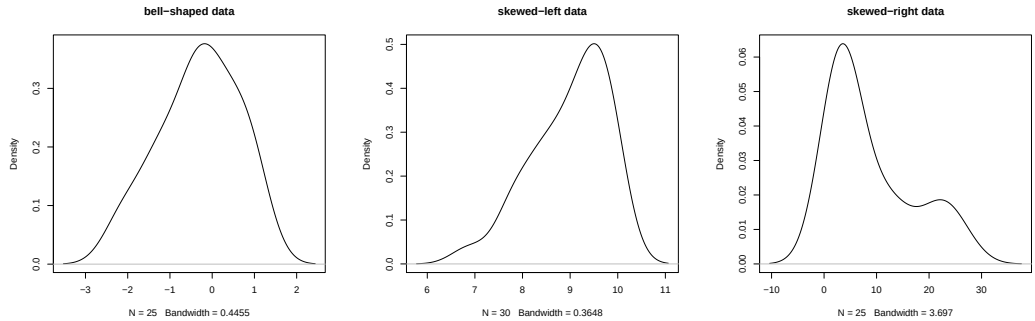

Figure 1: A bell shaped data set, a skewed left data set and a skewed right data set.

1. Load a data set
2. Look at a graphic of the data set
3. Estimate from the graphic the center and spread, then describe the shape using the vocabulary above.

If you forgot how to use R to load a data set here are some commando instructions. Refer to the first week's handout for more detail.

1. Start R by double clicking its icon on the desktop
2. From the R command line, issue the command `library(pmg)`.
3. Minimize the main R window so that only the pmg window shows.
4. Under the "Plots" menu open Lattice Explorer.
5. To load a package, such as MASS, under the File menu open "load package...". When the dialog opens, double click (once on the package name. Its status should change to TRUE.)
6. To load a data set, under the Data menu open the "load data set..." dialog. Double click on any data set you want to load.

Before we begin, let's load the MASS package. Do this using the dialog or by issuing the command

```
> library(MASS)
```

Okay, here is an example of what I want for this assignment.

For the data set `women` describe the `weight` variable.

First, we can load the data through the dialog or with the command

```
> data(women)
```

Now `women` is a data frame, click on the "+" sign to see that it has two variables: `height` and `weight`. Drag the `weight` one to the lattice explorer and make a histogram or a densityplot. The density plot looks something like Figure 2.

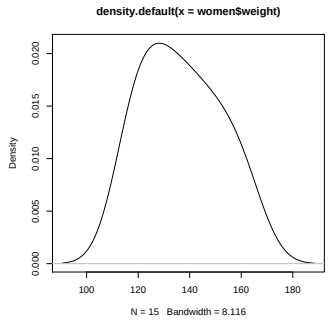

Figure 2: Density plot of `weight` variable.

From this graph we see the following: The center is around 135; the spread is about 125 to 155, or 30; and the data is unimodal and more or less bell shaped.

Remember to "clear" the graph before adding a new one, otherwise a multi-variate graph is drawn. (The "Clear" button is buggy!)

1. Load the `mtcars` data set. Describe the `mpg` variable.

2. From the `mtcars` data sets describe the `wt` data set.

3. Load the `Cars93` data set from the MASS pacakge. Describe the `MPG.highway` variable. This is a similar data set as the previous `mpg` data, is your description the same?

4. Clear the graph. Then from the `Cars93` data set first drag the `MPG.highway` variable, and then the `Cylinders` variable. A graphic with the MPG data broken up by the number of cylinders the car has is produced. There are now 6 plots. Describe the similarities and differences between them.

5. Load the `Animals` data set from the MASS package. Describe the variable `brain`.

6. Load the `Sitka` data set from the MASS package. Describe its `size` variable.

7. Load the `UScereal` data set from the MASS package. Describe the `calories` variable. Now drag the `mfr` variable to break the data out by manufacture (The help page has: 'mfr' Manufacturer, represented by its first initial: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.) Now describe the similarities and differences in the number of calories for different manufacturers.

8. Load the `Aids2` data set from the MASS package. This is data on Aids patients in Australia. The `age` variable records the age of the patient at diagnosis. Describe the age variable.

   The `state` variable gives the "state" of origin. Break the age variable up using the state variable to see if there are differences for the age between the states. Describe any differences or similarities