We have learned the following vocabulary, a part of Exploratory Data Analysis:

Center The center as measured with the mean or median

Spread the variation in the data measured by the standard deviation, the IQR and the range

Shape The general shape of the data as indicated by one of several graphics

Position A sense of how large a data value is in a data set. We've defined the **percentile rank** and the **z-score** to measure this.

This project shows how to compute the percentile rank and the z-score.

Before beginning we load some data sets. Using the "File::Load package..." dialog load the MASS package. Then load the data sets `Aids2`, `Animals`, `Cars93`, `leuk` and `michelson`.

Loading the MASS package is equivalent to typing this command

```
> library(MASS)
```

1 Using the "Commands" area

In this project we will make use of the "Commands" portion of PMG. This is the first tab on the left just to the right of where the variables appear. Click in here, you should see something like Figure 1. The command window has two states: one to enter in commands, and one to show how they are evaluated. A simple command would look like `mean(x)` which would find the mean of the variable `x`. For variables in data sets, the names are slightly more complicated. For instance `Cars93$MPG.highway` refers to a variable in the `Cars93` data set. We don't need to type these if the data sets are in the variable window – just drag it to Commands area and drop the text.

2 Finding the length of a variable

In R the length of a variable is what we call n, the number of data points collected for the variable. To find it, we use the `length()` function. To illustrate do the following:

- If needed, click the "edit" button to allow editing in the command area. Click "clear" to clear out any old commands.
- Now type `length(` and then drag the `MPG.highway` variable to drop it after your typing.
- Now match the parenthesis for `length` by typing `)`. This is a command. If you click "evaluate" you should get output like

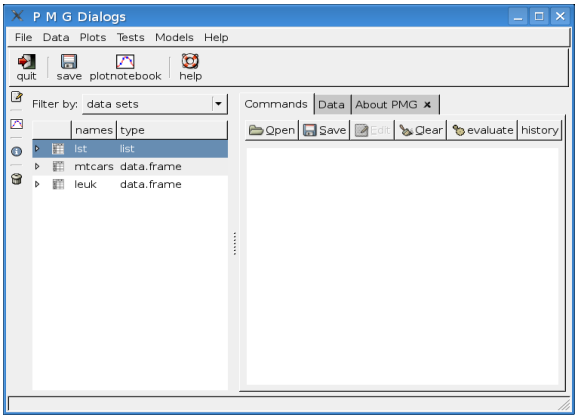

Figure 1: A screenshot showing the Commands tab ready for input. The window has two states: one where you can edit the commands, and another where the commands are evaluated. Previous commands can be recalled using the history button. Dragging a variable from the left will add the name of that variable to the command window.

```
> length(Cars93$MPG.highway)

[1] 93
```

The 93 indicates that there are 93 values.

2.1 Problems

1. Find the length of the `age` variable in the `Aids2` data set.
2. Find the length of the `time` variable in the `leuk` data set.

3 The percentile rank

The percentile rank is defined in class by $(B + (1/2)E)/n \times 100\%$. We know how to find n, but what about B and E? For this we use a nifty feature of R. To illustrate, we can find the percentile rank of 25 for the `MPG.highway` variable with these commands.

```
> x = Cars93$MPG.highway
> n = length(x)
> B = sum(x < 25)
> E = sum(x == 25)
> (B + 1/2 * E)/n * 100

[1] 19.35484
```

You only type the commands after the prompt ">". To be precise type just:

```
x = Cars93$MPG.highway
n = length(x)
B = sum(x < 25)
E = sum(x == 25)
(B + 1/2*E)/n * 100
```

In the above the data is assigned to the variable *x* so that we don't have to continually type the longer name for the data. Then each piece of the formula is found. Notice, for instance, how *B* is found using `sum` to add up all values less than 25. *E* is found by summing up all the values equal to 25. (Equals is two equals signs.) The answer of 19.35 indicates that roughly 19 percent of the data is less than 25.

3.1 Problems

1. Find the percentile rank of 35 for the `MPG.highway` variable. (Edit the appropriate place, don't start anew.)
2. For the `age` variable in the `Aids2` data set find the percentile rank of 25. (Replace the definition of *x*.)
3. For the `death` variable *minus* the `diag` variable in `Aids2` find the percentile rank of 365 (the percentage who died within a year of diagnosis).

4 The *z*-score

Finding the *z* score of a number is done by finding the mean and standard deviation and then dividing. The mean is found with `mean()` and the standard deviation with the `sd()` function. To find the *z* score of 45 in the `MPG.highway` data we have these commands

```
> x = Cars93$MPG.highway
> (45 - mean(x))/sd(x)

[1] 2.984770
```

[Again, you only type the values after the prompt]

The *z*-score is almost 3. If the data is bell shaped, then this is unusally large.

4.1 Problems

1. Find the *z*-score of 30 for the `MPG.highway` data set. Is this large?
2. Find the *z*-score of 40 for the `time` variable in the `leuk` data set. Is this large?

5 The relation between the z-score and the percentile rank

If data is bell shaped, then the z-score can tell us information in the same manner as the percentile rank. If the data is *skewed*, then it is usually better to use the median, IQR, and percentile rank to describe it, as the mean and standard deviation may be strongly influenced by a few large (or small) values.

Let's assume we have symmetric data (check it with a boxplot or a density plot). Then we have these rough guides

- About 68% of z-scores are in $[-1, 1]$
- About 95% of z-scores are in $[-2, 2]$
- About 99.8% of z-scores are in $[-3, 3]$.

For a given data set, how can we easily check how many z scores are in a range? The `scale` function will find the z-scores for an entire data set. From there we just need to have R count, in a manner similar to how we found the percentile rank.

To illustrate, these commands will find how many values have z-scores in $[-1, 1]$ for the `MPG.highway` variable.

```
> x = scale(Cars93$MPG.highway)
> n = length(x)
> sum(-1 < x & x < 1)/n

[1] 0.7741935
```

Again, type only the commands after the prompt. The `sum()` function uses the ampersand to combine expressions. The `-1 < x & x < 1` is read "x is greater than -1 *and* less then 1.

We get about 77% of the data.

5.1 Problems

1. What percent of the data in `MPG.highway` has a z-score in $[-2, 2]$?
2. For the `age` variable in the `Aids2` data set, what percent of the data has a z-score in $[-2, 2]$. Make a graph of this data, is the data set symmetric?
3. Let x be the variable `Aids2$death - Aids2$diag`. (Thetime an aids patient lived.) data set, what percent of the data has a z-score in $[-2, 2]$. Make a graph of this data, is the data set symmetric?