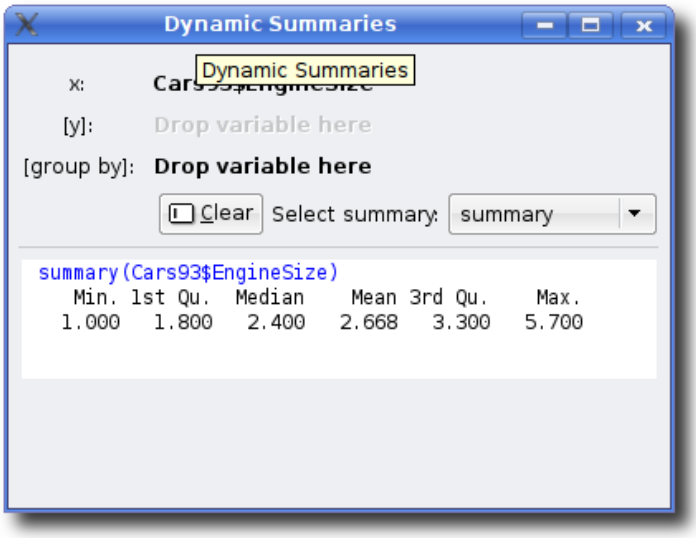

Figure 1: `Dynamic summaries` dialog showing a summary of the `EngineSize` variable in the `Cars93` data set.

We will use R to investigate the big 3 concepts of exploratory data analysis: the center, the spread and shape.

As before we a) double click to load R b) type `library(pmg)` to load `pmg` c) minimize the main R window and d) wait.

Once `pmg` is loaded we use the `Data::Load data sets...` dialog to open some data sets. from this dialog open the following data sets from the `MASS` package: `Animals`, `Boston`, `Cars93`, `Insurance`, `cats`, and `galaxies`.

1 Measures of center

There are various ways to define the "center" of a set of data. The average, or mean; the median, or middle point; or the mode, the most common value. Each has its value.

We can compute measures of the center very quickly with the `Dynamic Summaries` dialog under the `Data` menu. Open this up and change the summary popup to `summary`. Then drag a variable over to the `x:` line and drop it where it says **Drop variable here**. The five number summary and the mean are shown.

For example, if you drag the `EngineSize` variable from the `Cars93` dataset you will get something like Figure 1. From this, we can see that the mean is slightly more than the median.

1. Find the mean and median for the following data sets:

   (a) The `black` variable from the `Boston` data set.

   (b) The `dis` variable from the `Boston` data set.

   (c) The `Length` variable from the `Cars93` data set.

   (d) The `claims` variable from the `Insurance` data set.

   Write down the values, and note which is greater, the mean or the median

The term "dynamic summaries" is used as they can update dynamically when a value from the `Data` page is updated. Click on the `Data` tab, then the `Open` button, and then click on the `Add` button. This opens a new sheet to add values to.

Type in the values 1,2,1,4 under the `X1` variable. Now drag the `X1` column head and drag this onto the Dynamic Summaries area. You should get the summary of the 4 values. Now when you type in new values or edit old ones the summary will update automatically.

1. Starting with 1,2,1,4 type in a 5th value so that the mean is 5, What value do you use? (If you don't like to compute, you can use trial and error.)

2. Now try to type in a data set with just 5 values which has mean more than 10 and median less than 2. What do you get?

3. Now try to type in a data set with just 5 values which has median equal to 5 and mean equal to 10. What is the data set?

2 Measures of spread

The spread of a data set gives us a sense of how variable the values are. We discussed three measures: the range – a natural sense of spread, but sensitive to just one large or small values; the IQR – the range of the middle 50%; and the standard deviation (or variance) which basically computes how far on average are the values from their center.

We can again use the `Dynamic summaries` dialog to compute values for us.

1. For the `galaxies` data set, which is more the standard deviation or IQR?
2. For the `black` variable in the `Boston` data set which is more the standard deviation or the IQR?
3. For bell shaped data (normal data) the IQR is about 2/3rds the standard deviation. For the `MPG.highway` variable in the `Cars93` data set, is the IQR more or less than 2/3rds the standard deviation?

3 Interpretation

From a densityplot, histogram or dotplot we can visually assess how large the mean, median and IQR will be:

1. The mean is the balance point
2. The median is the 50% point
3. The IQR covers the middle 50% of the data

As for the standard deviation, *when the data is bell shaped (Normal)* then between one standard deviation to the left of the mean and one to the right will sit approximately 68% of the data.

To look at densityplots easily, we can use the dialog found under the `Data::Univariate Summaries::quantiles` menu item. To use this, simple drag a variable to the bold area **Add variable here**. The density plot is drawn.

To see the median

By changing the `probs:` value to `c(.25, .5, .75)`, the quartiles will be shaded.

1. Use this dialog to make a density plot of the `horsepower` variable in the `Cars93` dataset. Guess which will be more the mean or median, based on the above interpretations. Check your guess.
2. Repeat with the `Claims` variable in the `Insurance` data set.

4 Shape

There are some basic terms to describe the general shape of the distribution of a data set. A densityplot shows the basic shape, as does a histogram.

The number of peaks or modes of a dataset is an important feature. If the densityplot shows one prominent peak then the data set is termed *unimodal*; if two peaks then *bimodal*; and if two or more peaks *multimodal*. Now that's easy? Except the number of peaks depends upon how the density plot is made. Just like the histogram where the number of bins can decide how rough or smooth the graphic is, the same is true for a density plot. Basically if a peak just shows up, it may or may not be real. You should focus on dominant peaks.

For unimodal data sets, the central peak serves as some notion of center. Relative to this a data set is *symmetric* if it looks more or less the same on both sides. It is *skewed right* if it has a longer tail on the right; and *skewed left* if it has a longer tail on the left.

1. For each of these data sets, describe them as uni/bi/multi modal
   (a) `Max.Price` in the `Cars93` data set
   (b) `length` in the `Cars93` data set
   (c) The `galaxies` data set
2. Characterize these variables as symmetric, skewed right or skewed left
   (a) The `black` variable from the `Boston` data set.
   (b) The `dis` variable from the `Boston` data set.
   (c) The `Length` variable from the `Cars93` data set.
   (d) The `claims` variable from the `Insurance` data set.
3. Now compare your last answers with the first question, where you were asked to find the mean and median for these data sets. Try fill in the blanks with *more than*, *less then*, or *about equal to*:

(a) For a symmetric data set the mean is to the median

(b) For a skewed right data set the mean is to the median

(c) For a skewed left data set the mean is to the median

4. The last question on the `body` variable shows a really skewed data set. Many times these can be *transformed* to symmetric data sets by applying a function such as square root, squared, or a logarithm.

   To apply such a function, click on the variable and you should see that it can be edited. Type in of the following *then hit enter* to see the effect of the transformation; then characterize it in terms of skew or symmetry

   (a) `Animals$body^2`

   (b) `Animals$body^(1/2)`

   (c) `Animals$body^(1/3)`

   (d) `log(Animals$body)`