We will use R to investigate the normal distribution.

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 by double clicking on their row: `Boston`, `Cars93`, `USCereal`, `cats`, `mammals` and `michelson`.

We discussed a few new graphics in class: the boxplot and the quantile-quantile plot.

1 Boxplots

Boxplots visualize the five-number summary with a graphic that quickly allows on the see the center (by the median); the spread (by the IQR or range); and the shape (symmetric or skewed). In addition, it allows us to identify the presence of outliers. Thats alot for a graphic that also makes it possible to easily compare these quantities for several different groups of data at once.

The `Lattice explorer` dialog will draw boxplots for us. Find this under the `Plots` menu. Drag the `waiting` variable from the `geyser` data set and drop it on the plot area of the lattice explorer. By default, the density plot is shown. Change this to `bwplot` (for box-and-whisker plot).

This graphic is drawn with a dot marking the median, rather than a line. Otherwise we see the data broken up by its quartiles: the left whisker covers the range of the lowest 25% of the data, the box the range of the middle 50% of the data and the right whisker covers the range of the top 25% of the data. This data set shows no outliers and from the boxplot looks relatively symmetric.

1. What "feature" of the shape is apparent from the densityplot *but not* the boxplot?

Now clear the lattice explorer and drag the `Hwt` variable from the `cats` data set. Make a boxplot. You should see two values marked as outliers.

1. What is the IQR for this data? (Estimate from the graphic)
2. What is the value of Q_3?
3. What is $Q_3 + 1.5 * IQR$?
4. Estimate the two values that are plotted as outliers and compare these to the value of $Q_3 + 1.5 * IQR$? Is this consistent?

The boxplot works well with multiple groups. Clear the lattice explorer then drag over the `Speed` variable from the `michelson` data set. This data comes from a famous experiment done in 1888 to measure the speed of light.

1. Describe the shape of this data set: is it skewed? symmetric?
2. Are there outliers?

This experiment consisted of a series of five experiments. The `Expt` varible records the experiment. Drage this variable over and you should see five boxplots now.

1. Which experiment has the smallest "center?"
2. Which experiment has the greatest "spread?"
3. Explain why more points are marked as outliers than before.
4. Which experiment looks much different than the others. Clarify what you mean by "different."

Now, from the `cats` data set, make boxplots of the `Bwt` (birthweight) variable for each level of the `Sex` variable.

1. Write a sentence or two describing what this graphic says in plain language.

2 Quantile-quantile plots (qqplots)

The quantile-quantile plot plots the quantiles of one distribution against those of another. The `qqmath` plot of the `lattice explorer` uses the normal distribution as the reference distribution for the y axis.

As discussed in class, this graphic will consist of points more or less aligned on a straight line **if** the two distributions have the same shape; otherwise the points will not align on a line and if there is a curve it can often be traced to either a longer tail or a shorter tail in the non-normal distribution.

1. For the same `Bwt` versus `Sex` graph, change to `qqmath` to make the quantile-quantile graph. Do the two different sets of data appear to have a normal distribution? Why or why not.

2. The `temp` variable in the `beav1` data set measures temperature throughout the day for a beaver. Are these values normally distributed? Explain how you decide that.

3. For the `michelson` data, make quantile-quantile plots of the `Speed` variable grouped by the `Expt` variable (drag this over second).

   Which groups, if any, seem normally distributed?

3 Probability calculator

Under the `Plots` menu the teaching demos window can be called up. Do so, then change the demo to "probaiblity calculator" (typo noted, now fixed.) This demo replaces the table. It can be used to find probaibilities, or quantiles.

To find a probability from a z-score, simple leave the values for the mean (`mean`) and standard deviation (`sd`) at 0 and 1 respectively, then in the Result area put in a z-score on the `Value` line. For instance 0.5. Then click on the `update` button. If all goes well, you should see the Result and the graph of the normal should be drawn with shading indicating what probability is found.

1. What is the probability to the right of z when $z = 0, .5, 1, 1.5, 2$?

2. Change the mean to 100 and the standard deviation to 100. Verify that you get the same answers as before for the `Value` being 100, 150, 200, 250, 300.

3. What is more likely: A ACT score of 25 or an SAT score of 600 assuming ACT scores are normal with mean 22 and 1.75 standard deviation and SAT scores are normal with mean 480 and 100 standard deviation.

4. For the ACT scores, what is the area under the curve between 21 and 23?

The quantiles are returned, instead of the probabilities, when the `Find quantile` button is selected. Now the `Value` must be between 0 and 1 (percentiles are between 0 and 100).

1. For a mean of 0 and standard deviation of 1, find the quantile for a value of 0.75?

2. For a mean of 0 and standard deviation of 1, find the quantile for a value of 0.80?

3. What value of z corresponds to 85% of the area between $-z$ and z for the mean 0, sd=1 normal?

4. For the ACT data, what value is the 90th percentile?

5. For the ACT data, between what values are the middle 60% of the scores?