Confidence Intervals

I. Confidence Intervals for \( \mu \)
   For continuous data, we can easily obtain confidence intervals for the population mean, \( \mu \). Create a histogram for the variable of interest by following the steps “Analyze” then “Distribution”. In the distribution window, click on the red triangle next to the variable, and choose “Confidence Interval”. You are now given options on different \( \alpha \)-levels. Choose the level you prefer, and the confidence interval will appear in the distribution window.

   If your data contains two different groups (as is the case with our “Big Class” data set, which contains boys and girls), you might want to create two confidence intervals, one for each group. Create separate histograms by specifying a “by” variable in the distribution dialogue window (in the “Big Class” data set, this would be the variable “sex”). When that is done, simply create a confidence interval for each group. If you want to compare the two intervals, you have to remember to choose the same confidence level!

   The procedure of making confidence intervals is based on the assumption that the data is a random sample from a population with a normal population frequency curve. We will discuss different ways to check if this is true.

II. Confidence Intervals for Population Proportion
   A CI for \( p \) can be obtained in JMP-IN by defining a categorical variable with two levels, corresponding to the success and failure categories, and a frequency column that identifies the number of successes and failures in the sample.

   Get the frequencies for success and failure by the steps “Analyze” then “Distribution”. The confidence interval is created as above. Click on the red triangle next to the variable, choose “Confidence Intervals” and specify the confidence level. The confidence interval appears in the distribution window.

   To make confidence intervals for population proportions, the sample size needs to be large enough. A simple rule of thumb is to check that \( np > 5 \) and \( n(1 - p) > 5 \) for the method to be suitable.

   If the population proportion \( p \) is unknown, you should use the sample estimate in these formulas to check the suitability of the CI.

III. Interpreting a Confidence Interval
   Before the data is collected, the interval is unknown and is viewed as random since it will depend on the actual sample selected. The confidence interval is determined once the data is collected and the confidence coefficient is specified. Hence, different samples give different intervals.
The interpretation of a 95% confidence interval is as follows. If intervals are constructed accordingly, over and over again, 95% of them will contain the true value for the parameter of interest. Note that the interval constructed will either cover the parameter or will not.

**Testing for Normality**

Are your data from a normal distribution? There are several ways we can determine if the distribution can be considered normal.

**I. Visual Representations**

First, construct a stem-and-leaf representation of the data. Does the data seem unimodal? Symmetrical? If so, then it may be normal. If not, then it probably is not normal.

Second, create a boxplot. If the distribution is normal, then the arms on the sides should be roughly equal in length. If there are extreme outliers, then there should be roughly the same number of such outliers on each end.

Third, create a histogram. Check for modality, symmetry, and outliers again. Also look for the “bell” shape of the normal distribution. Keep in mind that the “bell” could be stretched out or very peaked.

Using JMP, we can overlay a possible bell-curve over top of the histogram to check for normality. Once you have the histogram created, click the red triangle above the graph and choose “Fit Distribution”, then “Normal”. A bell-shaped curve should appear on your histogram. To see if your histogram roughly follows this curve, click the white hand button on the upper button bar and modify the thickness of the bars as we have done in class before.

This bell-shaped curve is created from the mean and the standard deviation of your data. JMP doesn’t know any more than you about the true nature of your data’s distribution, so don’t let this curve fool you into believing that this is what your distribution “should” look like. It is only a guide to determine normality.

**II. Better Tests for Normality.**

a) **Normal Quantile Plot**

First, create a histogram in the normal manner. Second, click the red triangle above the histogram. Choose “Normal Quantile Plot”. A new graph is created that shows a bunch of dots plotted roughly along a straight red line between two other curved red lines. This graph is also called a Normal Probability Plot or a Normal Scores Plot.

If your data are truly from a normal distribution, the arrangement of dots should be linear and should follow the middle red line. They should not be above the upper curved red line, nor should they be below the lower curved red line.

It is important to remember that even if we sample from a perfectly normal
population, there will be some variability between the sample we obtain and the theoretical normal scores. What we are looking for is the general trend in the plot.

We do not want to see a curve in our normal plots, since curves indicate that the population is skewed. A normal plot concave down indicates data skewed to the left, while a normal plot concave up indicates data skewed to the right.

It is not uncommon that we see normal curves that are s-shaped; steep at the ends and flat in the middle. This happens when the distribution we are sampling from has long left- and right-hand tails.

b) Shapiro-Wilk Test (or Wilk-Shapiro Test)

There is a statistical test to determine how the probability that a certain distribution is normal. It is the Wilk-Shapiro Test, and JMP will do it for you. Follow the steps above to draw the bell curve on top of your histogram. In doing this, new information will appear in your window below (or to the right of) the Moments heading. The heading for this information is “Fitted Normal”, and it will show you some information that you already know which it used to create the bell curve. Click the red triangle next to “Fitted Normal” and choose “Goodness of Fit”. What you will see is something similar to this:

<table>
<thead>
<tr>
<th>Goodness-of-Fit Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk Test</td>
</tr>
<tr>
<td>W</td>
</tr>
<tr>
<td>Prob&lt;W</td>
</tr>
<tr>
<td>0.989210</td>
</tr>
</tbody>
</table>

Don’t worry about the number under the W. The other number (under “Prob<W”) is called a p-value, which is something we’ll be learning a lot about soon. In this case, the p-value is the probability of seeing this arrangement of data if the underlying distribution is indeed normal.

In short, the smaller the p-value, the less chance that what we are looking at is a normal distribution. The cut-off is usually 0.05, so if this number is less than 0.05, we are confident that the distribution is not normal. If it is greater than 0.05, we say that we cannot reject our assumption of normality.