

Confidence Intervals

Objectives:

1. To explore confidence intervals via simulation
2. To practice computing confidence intervals

Getting Started: Log onto your machine and start JMP. There are no datasets to download for today.

Part I. Confidence Intervals for Means

The Central Limit Theorem says that the sample mean is approximately normally distributed for large samples. We can take advantage of this result to make confidence intervals for unknown population means from data. First we'll explore why some intervals are based on the normal (z), and why some are based on the t distribution.

Start by generating nine columns of 1000 random draws from the standard normal distribution. Recall that you can do this with the following sequence: Starting at the JMP Starter Window, choose "New Data Table". Double-click on the header for "Column 1", which will bring up a dialog box. On the second-to-bottom line, you should see "Initial Data Values Missing/Empty Number of rows 0". Click on **Missing/Empty** and choose **Random**, and then click on the button for **Random Normal**. Change the **Number of rows** from 100 to 1000, and then click on "OK". You should now have 1000 random normals in a single column of your data table. Go to the red hot spot on the middle left side of the window where it says **Columns** and click on the hot spot and choose **Add Multiple Columns**. In the dialog window that pops up, tell it to add 8 and hit "OK". You should now have nine columns, the first one with random draws, and the other eight with missing values (shown by dots). For each of the new columns, double-click on the header and then put in 1000 random standard normal draws.

Now you should have nine columns of 1000 random normal draws. What we are going to do is to think of this as 1000 samples each of nine observations. So each sample is in a row, and each column represents one observation in that sample. For example, these might be the standardized weights of nine people riding in a van, and we might be looking at the average weights of people in 1000 different vans. We'll construct confidence intervals for each of these 1000 samples. To do this, we'll need the mean for each row.

Create a new column to the right of the current data by double-clicking where a new header would go. Right-click on the new column header and choose **Formula...** to bring up the formula dialog box. Go to the **Functions (grouped)** menu and scroll down a bit to find **Statistical**. Choose that, and from the submenu, choose **Mean** (note: not **Col Mean**, just the plain **Mean**). Then click on **Column 1** in the upper left box and then shift-click on the bottom column, so that it puts the names of all of the columns in the formula (inside of the mean). Hit "OK" and you should get the row means in your new column. Label your column "Sample Mean" (single click on the header to edit the column label).

Question #1 What are the theoretical expected mean and standard deviation of the sample means in this new column?

Question #2 What are the observed mean and standard deviation? Are they close to what you expect?

Question #3 Describe the shape of the histogram of your sample means.

Since we simulated this dataset, we know the true population standard deviation for the individual observations. Here $\sigma = 1$. And in Question #1, you computed the population standard deviation for the sample mean, $\sigma_{\bar{X}}$. Assuming the population standard deviation is known, the confidence interval uses a z , i.e., it is $\bar{X} \pm E$ where $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

Question #4 For a 95% confidence interval, what is the margin of error, E in this case?

So a 95% confidence interval goes from $\bar{X} - E$ to $\bar{X} + E$. We don't want to look at 1000 intervals by hand. What we'd like to see is how many of them contain the true mean, which in this case is zero, because that's how we simulated the data. Create a new blank column. Then right-click to bring up the **Formula...** dialog box. From **Functions (grouped)**, choose **Comparison** and then choose $a < b \leq c$, which should bring up three boxes in the formula region, with the left one surrounded by the active red box. We want to see if the true mean of zero is in the confidence interval. So from the upper left **Table Columns** menu, click on **Sample Mean** (you may need to scroll down to find it), which should appear in the left box. Then hit "-" and in the red box that appears type in your value for the margin of error, E . Click on the middle box of the formula and enter "0". Then click on the right box and put in **Sample Mean** "+" your value of E . Click on "OK". Your new column should consist of "1" and "0", where it is "1" when your confidence interval includes zero, and "0" when it does not. Take a look at a couple of rows to see that it is "1" when the sample mean is close to zero, and it is "0" only when the sample mean is far enough away from zero (either positive or negative).

Question #5 What fraction of the time do the 95% confidence intervals include the true population mean? (You can get this by going back to the JMP Starter window, choosing **Basic** and "Distribution", clicking on the column name for your new column of "1"'s and "0"'s and looking at the mean of the column, which is equal to the fraction of the rows that are equal to "1".

Question #6 Is this close to what you expect?

Now let's see what would have happened if we had used a t instead of a z . The $t_{.05/2}$ for 8 degrees of freedom is 2.306. Notice that this is larger than the z value of 1.96, so the confidence intervals will be wider. Recompute the margin of error and edit the formula for your last column (right-click on the header to get back to the **Formula...** dialog box, then edit the values for your margin of error and hit "OK"). (If your distributions window is still open, you may get a warning message telling you that the data table has changed. Don't worry about that.)

Question #7 Now what fraction of time do the 95% confidence intervals include the true population mean? (You will need to redo the mean analysis as before, starting from the JMP Starter window.)

Question #8 Which is closer to 95%, the z or the t intervals?

Now suppose that we don't know the true population standard deviation, and instead we have to estimate it from the sample. Thus we will use our point estimate s to estimate σ . First we need to compute the sample standard deviation s for each row. Create a new empty column, then go back to the **Formula...** dialog box, and this time from **Statistical**, choose **Std Dev** (right under **Mean**), and put in all nine of your data columns. Do NOT include your column of sample means or your column of "1"'s and "0"'s. Label this column "Standard Deviation" (or "Std Dev").

Question #9 Because the data are simulated, we know the true standard deviation for each observation is 1. Are the observed sample standard deviations close to 1? (Find the mean from the "Distribution" analysis.)

OK, now we're ready to compute confidence intervals. In this case, the margin of error will be different for each row, because s is different for each row. Again, we'll try this with both a z and a t . First let's do the z . Create a new blank column. Then right-click to bring up the **Formula...** dialog box. From **Functions (grouped)**, choose **Comparison** and then choose $a < b \leq c$. From the upper left **Table Columns** menu, click on **Sample Mean**, then hit "-" and in the red box that appears type in 1.96, then hit "x", then click on **Standard Deviation** from **Table Columns**, then click on "÷" and type in 3 (which is $\sqrt{9}$). You should now have the expression for the value of $\bar{X} - E$ in the left box of the formula. In the middle box, type in "0". In the right box, put in the **Sample Mean** "+" 1.96 "x" **Standard Deviation** "÷" 3. Click on "OK". This should give you a new columns of "1"'s and "0"'s, again with a "1" when the confidence interval contains the true population mean of zero.

Question #10 What fraction of the time do the 95% z confidence intervals contain the true population mean?

Question #11 Now go back and edit the **Formula** to make it a t interval, i.e., replace the 1.96 with 2.306. Now what fraction of the time do the 95% t confidence intervals contain the true population mean?

Question #12 In this case, which is closer to 95%, the z or the t intervals?

↪ **Question #13** Explain why the z intervals should be closer to 95% in the first case, and the t intervals should be closer to 95% in the second case.

Once you're checked off, you can close the data table and distribution windows, leaving just the JMP Starter window.

Part II. Confidence Intervals for Proportions

The Central Limit Theorem also applies for binomial distributions, so we can make confidence intervals for proportions. We'll simulate 10,000 draws from a binomial with $n = 17$ and $p = .5$, e.g., counting the number of heads on 17 flips of a fair coin. Then we'll compute a confidence interval for the population proportion. Since we simulated the data, we can see how often the confidence intervals contain the true value.

Open a **New Data Table** and double-click on the header for **Column 1**. Near the bottom, change the **Number of rows** from 0 to 10000 and hit "OK" to get 10000 rows of missing values. Right-click on the column header and choose **Formula...** From **Functions (grouped)**, scroll down and choose **Random** and **Random Binomial**. For **n** enter in 17, and for **p** enter in .5, and click on "OK". You should get 10000 draws of your binomial (with all values from 0 to 17).

Let's compute the estimated probability, \hat{p} , and put that in the next column. First create a new column by double-clicking the header area for the new column. Label this new column **p-hat**. Then right-click the header and bring up the **Formula** dialog box. From the upper left **Table Columns** menu, click on **Column 1**, then click on " \div " and type in 17 (which is the size of each binomial sample) and hit "OK".

Next we'll compute the margin of error and put it in the third column. Again Create a new column by double-clicking the header area for the new column. Label this new column **E** (for margin

of error). Then right-click the header and bring up the **Formula** dialog box. The margin of error for a binomial is

$$E = 1.96 * \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

So type in **1.96** (which should appear in the red box in the formula area) and hit “×” then hit “ $y\sqrt{x}$ ” and then click on **p-hat** in the upper left to put it in the box inside the square root sign. Then click on “×” and type in **1** and hit “-” and click on **p-hat** again. You should now have $p\text{-hat} * (1 - p\text{-hat})$ inside of the square root sign. Now click on the grey box that goes around that whole expression (everything inside of the square root sign) and then click on “÷” and type in **17** into the denominator. Click on “OK”.

Finally, let’s create a column of “1”’s and “0”’s to see which 95% confidence intervals contain the true population proportion of .5. Create a new column and go to its formula dialog box. From **Functions (grouped)**, choose **Comparison** and $a < b \leq c$. In the left box, put in **p-hat** “-” **E**. In the middle box, type in **.5**. In the right box, put in **p-hat** “+” **E**. Hit “OK”.

Question #14 What fraction of the time do the 95% z confidence intervals contain the true population proportion?

↪ **Question #15** Does the z work for a binomial? Would a wider interval with a t be needed?

Quit JMP and please remember to **Log Off**.