

The Central Limit Theorem

Objectives:

1. To practice with the normal distribution
2. To explore the Central Limit Theorem

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

Part I. Means for Normals

A central idea in statistics is the notion of *sampling variability*. Specifically, we have noted how a sample estimate (i.e., a *statistic* such as \bar{x} or \hat{p}) of a characteristic of the population (i.e., a *parameter* such as μ or p) varies among different samples taken from that population. Since the value of a statistic varies from sample to sample and its value depends on the outcome of each sample, a statistic is a random variable with its own probability distribution.

Central Limit Theorem for \bar{x} : Suppose X is a random variable such that $E(X) = \mu$ and the standard deviation of X is σ . If we take repeated samples of size n from this population and for each sample calculate the sample mean, \bar{x} , then the probability distribution of \bar{x} (for large n) has the following approximate characteristics:

1. The distribution of \bar{x} is approximately normal. Note that this implies that it is unimodal and symmetric.
2. The mean of \bar{x} is $E(\bar{x}) = \mu$ and its standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

If the individual observations X are normally distributed, then \bar{x} will be normally distributed even for small sample sizes.

Let's first generate some random draws from a normal distribution. From the JMP Starter Window, choose "New Data Table". Double-click on the head 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**. Some new options should appear, and click on the button for **Random Normal**. The box for **Number of rows** should default to 100, which is what we want, and we'll leave the default mean at 0 and the default standard deviation at 1. Click on "OK". You should now have 100 random normals in a single column of your data table. Let's check on their distribution. Go back to the JMP Starter Window and go to **Basic** and "Distribution". Make a histogram (feel free to rotate it) and get summary information on your 100 draws.

Question #1 Does the histogram look approximately normal?

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

Now let's look at what happens when we take the average of four normal draws, and do that 100 times, so we get 100 means of four normals.

Question #3 For the sampling distribution of the mean of four standard normals, what is its theoretical mean and its standard deviation?

We'll need to generate 100 draws in four separate columns and then take the means of the rows. Go back to your data table window (with the one column of 100 normal draws). First, we'll generate three more columns of 100 random normal draws. You can get new columns by going to the red hot spot on the middle left side of the window where it says **Columns** and click on the hot spot and choosing **Add Multiple Columns**. In the dialog window that pops up, the second box is **How many columns to add:**, so change that to 3 and hit "OK" (at the bottom). You should now have four columns, the first one with random draws, and the other three with missing values (shown by dots). The new columns really should be labeled "Column 2", "Column 3", and "Column 4", but if you are using version JMP version 7, you might see "Column 1 2", "Column 2", and "Column 3" instead. If so, you will need to adjust the instructions that follow appropriately.

For each of the new columns, double-click on the header and then put in 100 random standard normal draws (see instructions above if you've forgotten; in this case, the number of rows is already set so you don't need to worry about it here). Now you should have four columns of 100 random normal draws. Let's just make sure we have that by making histograms and getting summary information for each of the columns. Go back to the JMP Starter Window, choose **Basic** and "Distribution", and then in the dialog box that opens up, in the left box (**Select Columns**), click on **Column 1** and then shift-click on the last one (**Column 4**, or if JMP gave you strange column labels, perhaps **Column 3** or something else). This should highlight all four columns. Now click on "Y, Columns" and then on "OK". You should get four histograms and summary information for each of the columns.

Question #4 Do all four columns look approximately normal with approximately the right mean and standard deviation?

If everything looks reasonable, then we'll move forward into computing means in groups of four. You can close any of the "Distribution" windows that are open at this point. Go back to the data table window. We now want to get the mean of each row and put that into a new column. To the right of the header of the last column (either "Column 4" or "Column 3", depending on how JMP labeled them) double-click where a new column header would go, and it should add a new column labeled "Column 5". Right-click on the new "Column 5" header and choose **Formula...** to bring up the formula dialog box. Click on **Column 1** in the upper left box to start the formula (it should appear in the large box). Then click on the "+" sign just to the right of the list of columns, which will add a "+" to the formula. Now click on **Column 2**, then "+", then **Column 3**, then "+", then **Column 4** (or whatever the columns are labeled). Your formula box should now be showing the addition of all four columns, and the red box should be around the last one (e.g., **Column 4**). You

now need to move the red box so that it is around the entire formula. Try to click on the light grey box that goes around the whole formula to turn it red (if you are having trouble, you can try clicking in the empty white space in the formula box to make the red box go away, and then try again to click on the outer grey box). Once you have the red box around the whole thing, then click on “ \div ”, and it should turn your formula into a fraction, with an empty red box on the bottom. Put a 4 into the red box on the bottom (the denominator) and hit enter. Now your formula should be the mean of the first four columns (i.e., add them up and divide by 4). Click “OK” and you should get the means in Column 5. Make a histogram and get the numerical summary information for your sample means. Recall that each sample mean is a mean of four random normal draws, and that you now have a set of 100 sample means.

Question #5 Does the sampling distribution of the means look normal?

↪ **Question #6** What are the mean and standard deviation of the means? Does this match your answer to Question #3?

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

Part II. General Case

The Central Limit Theorem is particularly powerful because it applies (at least approximately) regardless of the distribution of the individual observations (as long as they have the same mean and same standard deviation as each other). Let’s try this with a discrete distribution, in particular, a discrete uniform on $1, \dots, 10$, i.e., the whole numbers from 1 to 10 are each equally likely to be picked. This distribution arises in a number of circumstances, including spinners in board games (like “Life”), or the last digit of many types of numbers (like social security numbers, or entries in tax returns). (As a side note, one of the ways they can catch people cheating on their tax returns is because the entries don’t show the right types of variability — it turns out that people are really bad at making up supposedly random numbers.)

Open a **New Data Table** and double-click on the header for **Column 1**. For **Initial Data Values**, replace **Missing/Empty** with **Random**, and then when the new options appear, leave the button on **Random Integer**. To the right, it should have boxes for **between 1 and 100**. Change the 100 to 10, and then click “OK”. You should get 100 random whole numbers from 1 to 10 in the first column.

Question #7 Make a histogram of these random numbers. Do they look approximately uniform?

Question #8 What is the theoretical expected value (mean) for these random draws? How close is your observed sample mean?

Question #9 The theoretical standard deviation is 2.872. How close is your observed sample standard deviation?

Next add nine more columns and put 100 random whole numbers from 1 to 10 in each of the new columns, so that you have a total of 10 columns (each with 100 rows). (You are welcome to add even more columns if you want, bringing the total to 15 or 20, which will likely improve the results of the simulation. Just replace the 10's below with whatever number of columns you have created.) Now 10 isn't that large of a sample, and our original observations don't look much like the normal distribution, so the Central Limit Theorem won't give that good of an approximation if just took the average across rows like last time. We need a larger sample for the Central Limit Theorem to be helpful. Instead, this time we will take the average of the columns, so that we will have 10 averages of samples of 100 each. To do this, from the data table window, find the hot spot in the upper left next to **Untitled**. Click there and choose **Tables** and then choose **Summary**. In the leftmost box labeled **Select Columns**, click on **Column 1** and then shift-click on the last one (**Column 10**, or whatever JMP has labeled it, such as **Column 9**) to select all 10 columns. Next click on **Statistics** and choose **Mean** from the drop-down menu. It should fill in the box with the means of all of the columns. Click on "OK". You should get a new table with a single row, a column titled **N Rows**, and a column with the mean of each of the previous columns of 100 draws. We'd really like them all in the same column, so we can analyze them. First get rid of the column called **N Rows** by right-clicking on the **N Rows** header and choosing **Delete Columns**. Now we'll flip the table over. Go to the hot spot in the upper left of the window (by **Summary of Untitled**) and choose **Tables** and then **Transpose**. In the dialog box that appears, click on the top one (**Mean(Column 1)**) and then shift-click on the bottom one (e.g., **Mean(Column 10)**) to select all of the rows, then click on "Transpose Columns" and hit "OK". You should get yet a new data table window, this one with all ten of the sample means in a column labeled **Row 1**.

Question #10 What are the smallest and largest observed sample means? Are they all close to the theoretical expected value?

Question #11 Make a histogram of the sample means (from **Basic**, “Distribution”, then click on **Row 1** and “Y, Columns” and “OK”). Ten draws from the sample mean may not be enough for it to look approximately normal (but it might). It should at least not look uniform. Describe what you see in the histogram (don’t close this window until you’ve been checked off for this section). You might find it helpful to adjust the number of bins by using the grabber tool.

Question #12 What is the observed mean of these ten sample means? Is it close to the expected value?

Question #13 Compute the theoretical value of the standard deviation of the sampling distribution for the mean. Is it close to the sample standard error (the observed standard deviation of the sample means)?

↪ **Question #14** How accurate has the Central Limit Theorem been in this exercise?

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

Part III. Normal Approximation to the Binomial and Poisson

Make a **New Data Table**. We’ll put 1000 draws from a binomial with $n = 24$ and $p = .25$. For example, this could be guessing randomly on a multiple choice test with four choices for each question. Recall that to get random draws from the binomial distribution, first tell JMP how many draws you want by double-clicking on the **Column 1** header and putting 1000 in the box to the right of **Number of rows**, and hitting “OK”. Next right-click on the **Column 1** header and choose **Formula . . .**, giving you the formula box. From the **Functions (grouped)** menu, scroll down and choose **Random** and **Random Binomial**. Fill in $n = 24$ and $p = .25$ and hit “OK”.

Question #15 For a draw from a binomial with $n = 24$ and $p = .25$, what are the theoretical expected value and standard deviation?

Question #16 What are your observed sample mean and standard deviation? Are they close to the theoretical values?

Question #17 Does the histogram of your sample look approximately normal?

As long as both $np \geq 5$ and $np(1 - p) \geq 5$, the normal is a good approximation to the binomial distribution. Similarly, the normal is a good approximation to the Poisson distribution for Poissons with large means. Generate 1000 random draws from a Poisson distribution with mean 30.

Question #18 For a draw from a Poisson with mean 30, what are the theoretical mean and standard deviation?

Question #19 What are your observed sample mean and standard deviation? Are they close to the theoretical values?

Question #20 Does the histogram of your sample look approximately normal?

↪ **Question #21** How good of an approximation does the Central Limit Theorem provide here?

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