September, 2019

Instructions for Normal Distribution Workbook -

By John Dunham, send questions or comments to johndunham76@gmail.com

This workbook is designed to work on Microsoft Excel Version 2010 or later; it may not work on older versions. It will certainly work on any version of Excel from a current Office 365 subscription.

Quick-Start Guide: If you only want to view my data and the plots, there is no need for any user input. Open the file named 3_Normal_Distribution_Workbook_Dunham_Sept_2019.xls. This contains porosity data from Slave Field. Just click on the different worksheet tabs to see Porosity data from Facies Types one through six as described in my paper in SEPM Special Publication 109, Mountjoy Symposium 1, on a Devonian Dolostone Reservoir from Alberta Canada. In addition to the Porosity Worksheets, I have included three additional sheets that compile heights and lengths measured in centimeters. These are useful if you have any of your own data that are scaled in numbers rather than percentages. The CM Box tab shows a compilation of these centimeter data sets. Each data set comes from a Normal Distribution, but the Box Plot shows that these data sets obviously come from different populations that show no overlap in their distributions. You are free to use these plots and all data analogous examples of porosity and permeability from a carbonate reservoir that has produced more than 57 million barrels of combined oil and water entirely from non-matrix vuggy porosity.

Guide For Inputting Your Own Data: You are encouraged to use these worksheets to examine your own data. To enter your own data, open the file named 3a_Normal_Distribution_Workbook_Your_Data.xls, and save it with a date added to the file name. If you mess up the sheet, you can always go back to the original to try again. The most important user input is to copy and paste your porosity or integer data into Column A of the sheet, and DON'T FORGET TO SORT THE DATA FROM LOW TO HIGH. Also, make sure that there are no “leftover” values in Column A which might happen if you paste fewer data points than I had into Column A. Check to make sure that Column A only contains the data that you entered . Then, the only other required inputs will be to re-scale the plots to match the range of your data.

I recommend that you do not copy one Worksheet Tab and paste it into a New Worksheet Tab; this is because all the graphs and some of the cell references on the New Tab will still be referenced to cells on the Old Worksheet Tab. It is simpler to re-name the original workbook file and simply enter your new data into the re-named workbook. The instructions below should guide you through the process. Let me know if you have any questions or problems.

User Input: Copy and paste porosity data into column A starting at cell A3. Formulas in the workbook will pull in up to 900 porosity values. If you have more than 900 data points, you will have to modify some of the formulas in the workbook as will be explained below.

User Input: Sort the porosity data from lowest to highest.

Automatic: Column B automatically shows the count order of the sorted porosity data.

Automatic: Column C is the percentile value for each data point in the distribution. The percentile is calculated from the formula i/n+1 where i is the count number from column c and n is the total number of data points. The i/n+1 formula assumes that your porosity samples have not captured the entire population of all possible porosity values.

Automatic: Column D takes the percentile from Column C and automatically calculates a Z score measured in units of Standard Deviations from the Excel Function Norm.S.Inv.

Automatic: Column E takes the Z score from Column D and calculates a Frequency percentile for each of the porosity values in column A based on the assumption that the data are following a normal distribution. If the last term in the function is set to TRUE, the function returns Cumulative Frequency; if set to FALSE, it returns the relative frequency. Column E is set to TRUE, returning cumulative frequency.

Automatic: Column F is set to FALSE and returns the relative frequency, which is the probability of occurrence of each particular porosity value.

Automatic: Cell G3 calculates the Mean porosity based on the Excel Function =AVERAGE(A3:A902). If you have more than 900 porosity values in column A, you will have to modify the formula.

Automatic: Cell H3 calculates the Standard Deviation of porosity based on function =STDEV.S(A3:A902). Modify if you have more than 900 data points.

Explanation: Both a Relative Frequency and a Cumulative Frequency plot of porosity data are shown by the graphs anchored at Column T. The red data points are plotted with Porosity Percent on the X axis and Frequency Percent on the Y axis. The black line on each plot shows what a true Normal Distribution would look like based on the Mean and Standard Deviation of porosity in cells G3 and H3. The scatter of red data points away from the black lines shows that the data are not exactly following a normal distribution.

Automatic: Columns I through L are used to calculate the Bell Curve on the Relative Frequency Plot and the Cumulative Curve on the Cumulative Frequency Plot. These lines are based on the mean and standard deviation of porosity observed in this data set. The lines are not calculated or based in any way on the red data points; they come only from the mean and standard deviation values.

Automatic: Columns N through R automatically generate the histogram shown on the Relative Frequency plot. Column N contains the User supplied Bin Boundaries. Column O samples the Porosity data in Column A and automatically fills each bin with the number of samples present in each bin. This formula samples from cells A3 to A903, so if you have more than 900 data points, you will have to modify this formula. The Porosity range goes up to 40%, which should capture most porosity distributions. If you have significant porosity greater than 40% as in a mudstone, you will have to change the array. Just "google" histogram array in Excel. Column P reports the width of each bin; note that the number in column N corresponds to the top of each bin width.

A series of plots is anchored to Column T. If a user wishes to enter new data into this spreadsheet, it may be necessary to re-scale these plots. For example, the Facies_1_Porosity worksheet has plots that are scaled from 0% to 14% porosity, which captures the range of data in this facies type. In contrast the Facies_2_Porosity worksheet has plots that are scaled from 0% to 25% porosity.

User Input, Scaling the Relative Frequency Plot: The Relative Frequency plot has a Bell Curve overlain on a histogram. This is actually a group of two plots overlain on one another. Therefore, if this group is re-scaled, then both the histogram plot and the Bell Curve plot must be re-scaled, as follows. 1) Click anywhere on the Relative Frequency plot. 2) Go to the ribbon at the top of the Excel sheet and click on "Format". 3) On the right side of the ribbon, click on "Selection Pane". The Selection Pane appears on the right edge of the worksheet; the pane shows 6 different plots. It is the Bell_Prob and Bell_Histo plots that are overlain to form the Relative Frequency plot. If you change the scale on one, you will have to change the scale on the other, as follows: a) On the Selection Pane, click on the Bell_Histo line, note the "eyeball" icon to the right of each plot name. b) Click on the eyeball to the right of Bell_Histo plot, the histogram should disappear as the eyeball closes. c) Now click on the X axis of the Bell_Prob plot. d) The Format Axis panel appears on the right of the spreadsheet. e) The Minimum and Maximum Bounds are displayed in decimal format; change the Maximum to the highest porosity value you want to display; be sure to enter 0.25 and not 25 for example. f) Now you need to rescale the histogram. On the selection pane, click on Bell Histo eyeball to turn it back on, and turn off Bell_Prob eyeball. g) Click on any one of the vertical histogram bars; this will light-up blue outlines in Columns N and O that set the range of the histogram plot. To rescale the histogram up to 25% for example, just change the N and O ranges to include the row for 25% in Column N and set Column O to the same row. This will automatically re-scale the histogram. You do not have to do any work on the Format Axis pane.

Turn the Bell_Prob plot eyeball back on. What you should see is the 1% histogram bar centered between the 1% and 2% porosity values on the Bell Curve. This is correct because the Bin-Top of the Histogram is 1%. Similarly, the 2% histogram bar is centered between 1% and 2% on the Bell Curve, because 2% is the Bin-Top of the Histogram such that the 2% bar shows the number of values in the bin from 1.1% to the Bin-Top of 2%.

This is pretty simple for the Porosity Plots that start at 0% on both Bell and Histogram, but it is a little trickier on the Centimeter Length and Height plots. Go to the Worksheet Tab labeled Adult Height_cm. The Bell Curve plot is scaled from 150 to 200 cm which brackets all the data and has enough room on either side to show the tails of the Bell Curve. Click on the Bell Curve Plot and go to Format > Selection Pane, and click on the eyeball to turn off the Bell_Prob Plot. Look at the histogram and you'll see that there are no values on the X axis. This is because they have been set to White Text Fill in order to keep them from overposting the X axis of the probability plot. We will turn them back on now. Click just below the X axis of the histogram and an outline box will appear. On the upper right of the ribbon, in the Font Area, you'll see a box on the right side that has an A with a white bar under it. This sets the Font Color. Click on the Drop Down arrow and select Black for the font color; the X axis values have appeared. Next, click on any of the histogram bars. Then, scroll the worksheet down so that you can see Row 153 of the worksheet and you will see the outlines of two blue boxes in Columns N and O. These columns specify the scale of the Histogram. The important thing to note is that the histogram starts at the value 151 in Row N. This is because the Bell Curve is scaled starting at 150 cm, and so logically, the first bar of the histogram will contain all values between 150 and 151 cm, and this bar will be centered between 150 and 151 on the bell curve. Turn the Bell Curve back on by clicking the eyeball, and you should see the histogram bars centered between the proper values on the Bell Curve. You would think that if the Bell curve starts at 150, then the histogram should start at 150, but that is not the case. Always start histograms so that they contain the first value AFTER the first value on the Bell Curve.

Now that you've turned the Bell Curve back on, you can click it on and off a few times to confirm that the bars are centered in the correct places, but you will also notice that you have two scales on the X axis that are overprinting one another. Go back to the histogram and set the X axis fonts back to White and turn the Bell_Prob plot back on.

The plots should still be perfectly aligned, but if you did happen to move one or the other during this process, you can re-align them as follows. On the selection pane, click on Bell_Prob. Then, hold the Ctrl key down and click on Bell_Histo; both names should now be highlighted on the selection pane. Then on the Ribbon, click on Format; just to the left of the height and width area is a box marked Align. Click on Align to open a drop-down menu. First select Align-Top, then select Align-Left. Both plots are now perfectly aligned.

Note on the histogram that there are very small labels under each vertical bar; these labels show that your two plots are properly aligned. The 1% bar on the histogram corresponds to the bin that ranges from 0 to 1% and the 2% bar ranges from 1.1 to 2%. Each bar corresponds to the Bin-Top, and each vertical bar is centered between its corresponding porosity range on the Bell_Prob plot. For example, the vertical bar at 1% shows 6 samples in that bin. On the Bell_Prob plot, note that there are 6 red data points between 0 and 1%. The Black Line is a plot of a theoretical Normal Distribution that is calculated from the Mean and Standard Deviation of the Porosity values from Column A. The histogram is useful for indicating whether the distribution has more than one Mode, and whether it is strongly skewed. The Bell Plot is useful for detecting the scatter of points away from the theoretical Normal Distribution Plot.

User Input, Scaling the Cumulative Frequency Plot: This is easer because it is just one plot. Click on the X axis and change the upper porosity number to your desired value; enter it as decimal 0.25 and not 25 for example. As in the case of the Bell Curve plot, the black line corresponds to a theoretical normal distribution calculated from observed porosity mean and standard deviation. The points do scatter away from the black line, but generally seem to track along it. The plot is useful for extracting porosity values that correspond to cumulative frequency probability percent's. For example, if the data did come from an ideal Normal Distribution, then 10% of the values would be less than 1.2% porosity, 25% of the values would be less than 2.7% porosity, and so on. These numbers are automatically displayed on the plot from an array of cells from AK53 to AT57. This is all automatic and requires no user input.

User Input, Scaling the Probability Plot: A Probability Plot is just the same as a Cumulative Frequency Plot, except that the Cumulative Frequency Axis has been "stretched" at the tails and "compressed" in the center in order to get the "Curved" line on the Cumulative Plot to correspond to a Straight line on the Probability Plot. This was done in 'the old days' to let people plot data points in pencil on "Probability Paper" and then use a ruler to draw an 'eyeball best fit' of a straight line to the data points. Once the line was drawn, then the user could extract P10, P50, and P90 values for use in Resource Volume calculations. Compare the probability plot to the cumulative plot above it; you can see that the distribution and scatter of points is almost the same relative to the black lines. However, there is one difference. The Black Line on the Cumulative plot doesn't really "see" the data points; it comes only from the Mean and Standard Deviation. On the other hand, the Black Line on the Probability Plot is coming from the slope and intercept of the straight line that is fit to the actual data. If you click on any single red data point on the Probability Plot, you'll see that the points come from Column A (porosity) and Column D (normsinv). NORM.S.INV is the Excel Function that calculates the Z value for a Normal Distribution that corresponds to the Percentile value shown in Column C. The Z value 0 for the 0.5 percentile of the distribution and extends in the negative direction for percentiles less than 0.5 and in the positive direction for percentiles greater than 0.5; the Z value is calculated based on the assumption that the data come from a Normal Distribution. But, the Y axis on the Probability Plot is scaled in Percent, and not in Z values! Where do these percent numbers come from??

The answer is that the Percent Numbers on the Y axis are Data Labels that correspond to an array extending from cells BO1 to BQ18. Column BO is a list of Cumulative Frequency Percent's that I want to display on my Probability Plot. Column BP then computes the Z value that would correspond to these percentiles using the Excel Function NORMSINV. Now I have Z values that correspond to Percent values. Column BQ lists the X axis "anchor" that is used to get all the Percentile Labels to plot on the Y axis where it crosses the X axis at 0%. If you change the porosity range on your plots so it starts at a higher porosity value than 0%, you will need to change the X axis anchor to whatever you used to set your minimum porosity. If the labels disappear from the Y axis, this is the likely reason. The placement of the Y axis percent labels is automatic and does not require user input.

People have complained that Excel does not have a built-in probability scale, but this is how you make one. Click on the Y axis to select the "Probability%" series that is posted on the Y axis from cells BQ2:BQ28 and BP2:BP28, which are the Z values and the X anchor values. With that series selected, go to the ribbon and click on Design, and on the left side of the Design ribbon is a box labeled Add Chart Element, click to open a drop-down menu and select Data Labels. A drop down menu appears, select Left. Now, go back to the upper left of the ribbon and click on the down arrow in the white box that probably says "Chart Area", a long list of names appears on the drop down menu, search for the line that says Series "Probability %" Data Labels, and click on that line. Next, click on the Format Selection button that is just below the white box with the selected "Probability %" name. When you click on Format Selection, a Format Data Labels pane appears on the right side of the worksheet. Click on the Check Box that says Value from Cells, a box appears that says Select Data Label Range, click on the up arrow at the right of the box and scroll the worksheet over so you can select cells BO2:BO18, and these cells should now be in the white box. Click the down arrow and then click OK. Now, **be sure** to uncheck any other box such as Y value or Show Leader Lines or anything else; make sure that Values from Cells is the only box selected. Hit the enter key and look at the Y axis on the Probability plot. The percent labels should be there. Of course, you don't want to have the Z value labels posted on the Y axis. They won't be there on this worksheet because their font-color has already been set to White. If you are using this method to make your own probability plots, just click on any one of the Z values posted on the Y axis, and use the Font box on the ribbon to change the font color to White; the Z values will disappear. To format the percent data labels that you created, just click on them on the Y axis and use the Font box on the ribbon to set the font, size, and color of your labels.

Automatic: P10, P50, and P90 values are automatically posted on the straight-line fit of the Probability Plot, along with P25 (first quartile) and P75 (third quartile) values. Note that the numbers from the Probability Plot may not be exactly the same as the numbers on the Cumulative Plot. This is because the Probability Plot is fitting a line to real data while the Cumulative Plot is not fitted to the real data. If the data are closely following a true Normal Distribution, then the numbers on both plots should be very close, but if there is significant departure from a Normal Distribution, then the numbers will be different. The amount of difference gives an idea of the degree of uncertainty in the true frequency distribution of porosity in this data set.

Automatic: Calculation of the Anderson-Darling Test to determine quality of fit of data points to a true Normal Distribution. In the old days, likelihood that the data really are coming from a Normal Distribution was qualitatively judged by looking at how many of the data points plotted on the line or how far they scattered from the line. The Anderson-Darling test is calculated in the array of cells ranging from Columns Bi through Columns BQ. The need for User Input will occur if you are replacing the data in this worksheet with new data from your own project. Go to Column BI and make sure that all the new data that you loaded into Column A is now displayed in Column BI.

The spreadsheet for the Anderson Darling test comes from spcforexcel.com, and the detailed reference is posted in Cell BW1. The question is whether a data set comes from a normal distribution, or not. The results of the Anderson Darling Test are posted next to the Probability Plot in cells AO60 to AP69. If the p Value in Cell AP66 is greater than 0.05, then there is a 95% probability that the data are from a Normal Distribution; if cell AP66 is less than 0.05, then the hypothesis of a Normal Distribution is rejected. Compare cell AL 66 on worksheet tab Facies_1_Porosity to cell AL66 on worksheet tab Facies_2_Porosity. The normal distribution is accepted for Facies 1 but rejected for Facies 2. Compare the two probability plots, note the goodness if fit for Facies 1 and the much higher scatter for Facies 2.

Automatic: A Horizontal Box Plot is automatically calculated from the porosity data and is posted at Column T below Row 78, just below the Probability Plot. The array of cells that produces the box plot are in Columns AY through Column BF. There is also a Vertical Box Plot anchored in Column AS. The Box Plots are explained as follows. The First (P25), Second (P50 = Median), and Third (P75) Quartiles are calculated from the sorted porosity data; the First and Third Quartiles form the outline of the Box. The “whiskers” extending away from the Box correspond to 1.5 times the range between Q1 and Q3. Any data points that plot beyond the whiskers are plotted as Outliers. The Median is shown as a line located near the middle of the Box, between Q1 and Q3. The Median line is placed within a notch. The notch corresponds to the 95% confidence interval for the location of the Median as described by Chambers et al., 1983. The formula for the 95% confidence interval is Median +/- 1.57*(Q3-Q1)/(number of samples^0.5). The Sample Mean is posted as a point near the center of the Box Plot, with error bars that correspond to the 95% confidence interval for location of the Mean. The Box Plot is useful for instantly allowing visualization of the shape of the data distribution. For Facies 1, the Median to Q3 is a little broader than the Median to Q1, indicating a slight positive skew. The Median and the Mean are close to one another and lie within the 95% confidence intervals of both parameters. The Vertical Box Plot in Column AY is useful for comparing data distributions among several different facies types. Go to Worksheet Tab Por Box Plots to see a comparison of porosity distributions by facies type. Facies 2 stands out as significantly “good” while Facies 4 is significantly bad. Similarly go to the CM Box Worksheet tab and note that these different data sets clearly come from different populations.

User Input: Note the Legend in the Box Plots; the name for the Box Outline is set to n =, which is the number of data points in the data set. If you load your own data, be sure to click on the outline of the box, which will open the Series Name dialog on the formula bar, and set the n= to the number of data points in your data set.

User Input: The final plot in Column T below Row 78 is another probability plot, but with axes reversed to show Cumulative Frequency on the X axis and Porosity Percent on the Y axis. Some companies prefer this arrangement of axes, as I found out when I worked for one of them. If you do not see all your data points or the straight-fit line, then you will have to manually adjust the input cells for the “Data Points” data series. Make sure that Column D and Column A are set to all the rows of data in your data set. The plot won’t work right if there are more rows or fewer rows selected for input, than exist in your data set.

That completes the explanation for how to use these Normal Distribution spreadsheets to make your own plots with your own data. Send questions or comments to johndunham76@gmail.com.