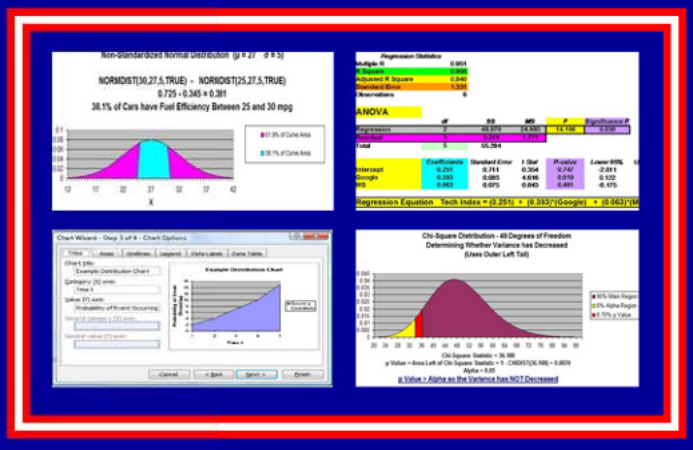
Excel **MASTER** Series

Normality Testing in Excel

The Complete Guide

Excel Statistical Master



Clear and Simple
yet THOROUGH
Statistical Instruction for the
Graduate Student and
Business Manager
with
LOTS of Worked-Out Problems
and Screen Shots

Mark Harmon MBA

Normality Testing in Excel

The Excel Statistical Master

By Mark Harmon

mark@ExcelMasterSeries.com

www.ExcelMasterSeries.com

ISBN: 978-0-9833070-5-1

Nonparametric Tests

Used When Data Is Not Normally Distributed

Statistical procedures are either parametric or nonparametric. Parametric statistical tests require assumptions about the population from which the samples are drawn. For example, many tests such as the t Test, Chi-Square tests, z Tests, and F tests, and many types of hypothesis tests require the underlying population to be normally distributed. Some tests require equal variances of both populations.

Sometimes these assumptions cannot be always be assumed. Examples of this would be if the population is highly skewed or if the underlying distribution or variances were entirely unknown.

Nonparametric tests have no assumptions regarding distribution of underlying populations or variance. Most of theses are very easy to perform but they are not usually as precise as parametric tests and the Null Hypothesis usually requires more evidence to be rejected in a nonparametric test.

Nonparametric tests are often used as shortcut replacements for more complicated parametric tests. You can quite often get a quick answer that requires little calculation by running a nonparametric test.

Nonparametric tests are often used when the data is ranked but cannot be quantified. For example, how would you quantify consumer rankings such as very satisfied, moderately satisfied, just satisfied, less than satisfied, dissatisfied?

Nonparametric tests can be applied when there are a lot of outliers that might skew the results. Nonparametric tests often evaluate medians rather than means and therefore if the data have one or two outliers, the outcome of the analysis is not affected.

They come in especially handy when dealing with non-numeric data, such as having customers rank products or attributes according to preference.

Correctable Reasons Why Your Data Is Not Normally Distributed

In the ideal world, all of your data samples are normally distributed. In this case you can usually apply the well-known parametric statistical tests such as ANOVA, the t Test, and regression to the sampled data.

What can you do if your data does not appear to be normally distributed?

You can either:

- Apply nonparametric tests to the data. Nonparametric tests do not rely on the underlying data to have any specific distribution

- Evaluate whether your “non-normal” data was really normally- distributed before it was affected by one of the seven correctable causes listed below:

The Biggest 7 Correctable Causes of Non-Normality in Data Samples

1) Outliers – Too many outliers can easily skew normally-distributed data. If you can identify and remove outliers that are caused by error in measurement or data entry, you might be able to obtain normally-distributed data from your skewed data set. Outliers should only be removed if a specific cause of their extreme value is identified. The nature of the normal distribution is that some outliers will occur. Outliers should be examined carefully if there are more than would be expected.

2) Data has been affected by more than one process – It is very important to understand all of the factors that can affect data sample measurement. Variations to process inputs might skew what would otherwise be normally-distributed output data. Input variation might be caused by factors such as shift changes, operator changes, or frequent changes in the underlying process. A common symptom that the output is being affected by more than one process is the occurrence of more than one mode (most commonly occurring value) in the output. In such a situation, you must isolate each input variation that is affecting the output. You must then isolate the overall effect which that variation had on the output. Finally, you must remove that input variation's effect from output measurement. You may find that you now have normally-distributed data.

3) Not enough data – A normal process will not look normal at all until enough samples have been collected. It is often stated that 30 is the where a "large" sample starts. If you have collected 50 or fewer samples and do not have a normally-distributed sample, collect at least 100 samples before re-evaluating the normality of the population from which the samples are drawn.

4) Measuring devices that have poor resolution – Devices with poor resolution may round off incorrectly or make continuous data appear discrete. You can, of course, use a more accurate measuring device. A simpler solution is to use a much larger sample size to smooth out sharp edges.

5) A different distribution describes the data – Some forms of data inherently follow different distributions. For example, radioactive decay is described by the exponential distribution. The Poisson distribution describes events event that tend to occur at predictable intervals over time, such as calls over a switchboard, number of defects, or demand for services. The lengths of time between occurrences of Poisson-distributed processes are described by the exponential distribution. The uniform distribution describes events that have an equal probability of occurring. Application of the Gamma distribution often based on intervals between Poisson-distributed events, such as queuing models and the flow of items through a manufacturing process. The Beta distribution is often used for modeling planning and control systems such are PERT and CPM. The Weibull distribution is used extensively to model time between failure of manufactured items,

finance, and climatology. It is important to become familiar with the applications of other distributions. If you know that the data is described by a different distribution than the normal distribution, you will have to apply the techniques of that distribution or use nonparametric analysis techniques.

6) Data approaching zero or a natural limit – If the data has a large number of value than are near zero or a natural limit, the data may appear to be skewed. In this case, you may have to adjust all data by adding a specific value to all data being analyzed. You need to make sure that all data being analyzed is "raised" to the same extent.

7) Only a subset of process' output is being analyzed – If you are sampling only a specific subset of the total output of a process, you are likely not collecting a representative sample from the process and therefore will not have normally distributed samples. For example, if you are evaluating manufacturing samples that occur between 4 and 6AM and not an entire shift, you might not obtain the normally-distributed sample that a whole shift would provide. It is important to ensure that your sample is representative of an entire process.

If you are unable to obtain a normally-distributed data sample, you can usually apply non-parametric tests to the data.

The Normality Test

Simple and Done in Excel

The normality test is used to determine whether a data set resembles the normal distribution. If the data set can be modeled by the normal distribution, then statistical tests involving the normal distribution and t distribution such as **Z test**, **t tests**, **F tests**, and **Chi-Square tests** can performed on the data set. In this chapter we will discuss two very simple tests of normality that can easily performed in Excel.

The Histogram - The Simplest Normality Test

Probably the easiest normality test is to plot the data in an Excel histogram and then compare the histogram to a normal curve. This method works much better with larger data sets. It is extremely simple to perform in Excel. Here is an example of how a Histogram is used in Excel as the most basic Normality test:

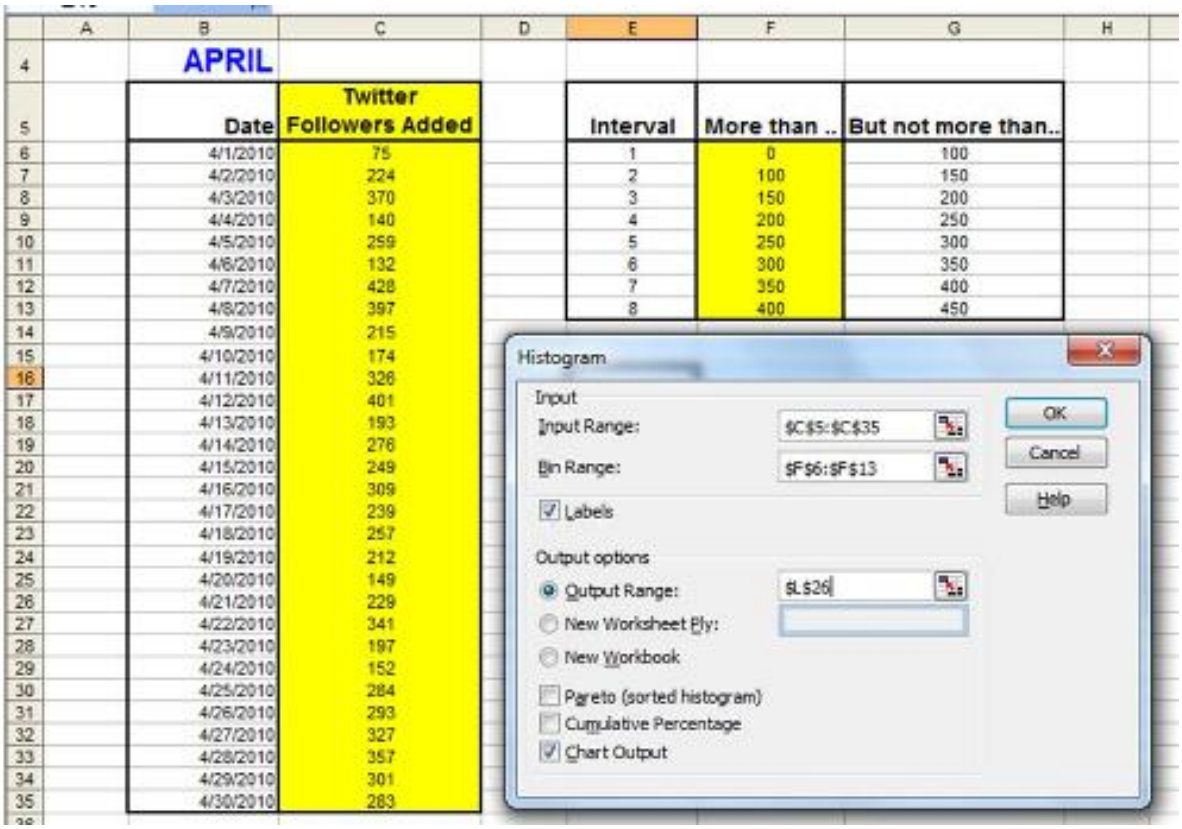
We are going to evaluate the following data for Normality using a Histogram:

APRIL	
Date	Twitter Followers Added
4/1/2010	75
4/2/2010	224
4/3/2010	370
4/4/2010	140
4/5/2010	259
4/6/2010	132
4/7/2010	428
4/8/2010	397
4/9/2010	215
4/10/2010	174
4/11/2010	326
4/12/2010	401
4/13/2010	193
4/14/2010	276
4/15/2010	249
4/16/2010	309
4/17/2010	239
4/18/2010	257
4/19/2010	212
4/20/2010	149
4/21/2010	229
4/22/2010	341
4/23/2010	197
4/24/2010	152
4/25/2010	284
4/26/2010	293
4/27/2010	327
4/28/2010	357
4/29/2010	301
4/30/2010	283

After the input data is arranged as above, we need to determine how we want the data to be grouped when it is broken down into a Histogram. Excel calls the groups "bins." We need to determine the upper and lower range of each bin. When the data is inserted into Excel, we need only to provide the lower boundary of each bin. Here is how I have arbitrarily set up lower boundaries for each bin:

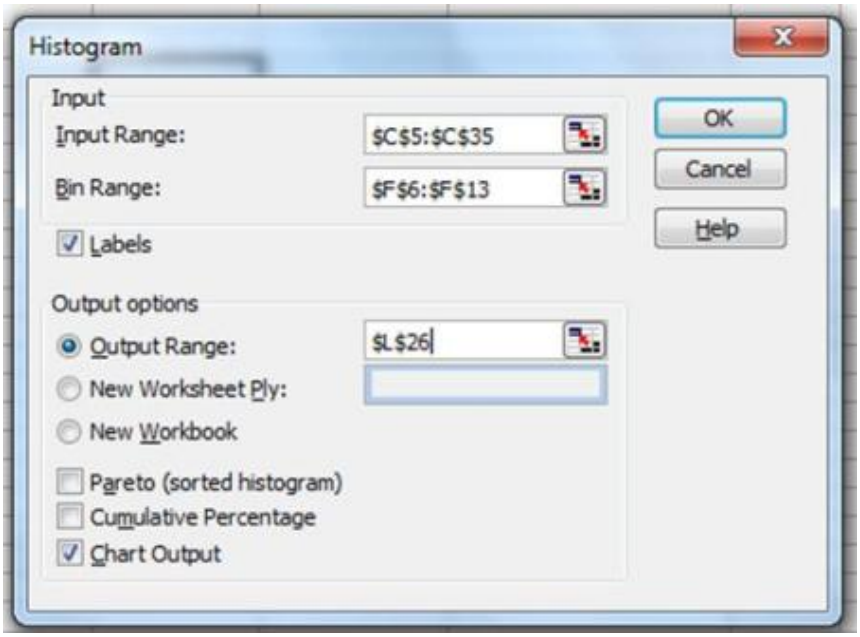
Interval	More than ..	But not more than..
1	0	100
2	100	150
3	150	200
4	200	250
5	250	300
6	300	350
7	350	400
8	400	450

Now we are ready to create a Histogram with Excel. Access the Excel Histogram in Excel 2003 from: **Tools / Data Analysis / Histogram**. A dialogue box will appear. The following dialogue box is shown completed. Highlight the input data and bin range data by selecting yellow-colored data cells as is shown above. Your dialogue box will look like this one when you are ready to create the Histogram:

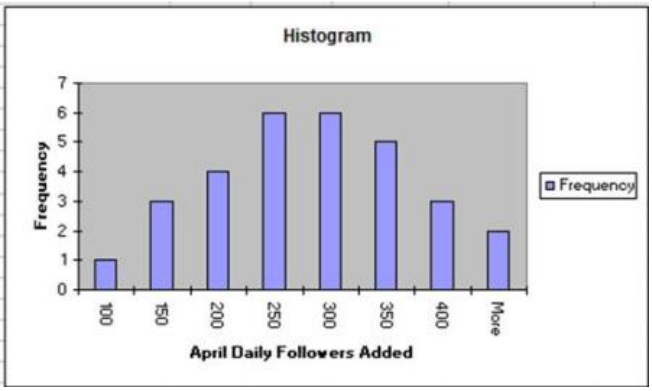


	A	B	C	D	E	F	G	H
4		APRIL						
5		Date	Twitter Followers Added		Interval	More than ..	But not more than..	
6		4/1/2010	75		1	0	100	
7		4/2/2010	224		2	100	150	
8		4/3/2010	370		3	150	200	
9		4/4/2010	140		4	200	250	
10		4/5/2010	259		5	250	300	
11		4/6/2010	132		6	300	350	
12		4/7/2010	428		7	350	400	
13		4/8/2010	397		8	400	450	
14		4/9/2010	215					
15		4/10/2010	174					
16		4/11/2010	326					
17		4/12/2010	401					
18		4/13/2010	193					
19		4/14/2010	276					
20		4/15/2010	249					
21		4/16/2010	309					
22		4/17/2010	239					
23		4/18/2010	257					
24		4/19/2010	212					
25		4/20/2010	149					
26		4/21/2010	229					
27		4/22/2010	341					
28		4/23/2010	197					
29		4/24/2010	152					
30		4/25/2010	284					
31		4/26/2010	293					
32		4/27/2010	327					
33		4/28/2010	357					
34		4/29/2010	301					
35		4/30/2010	283					

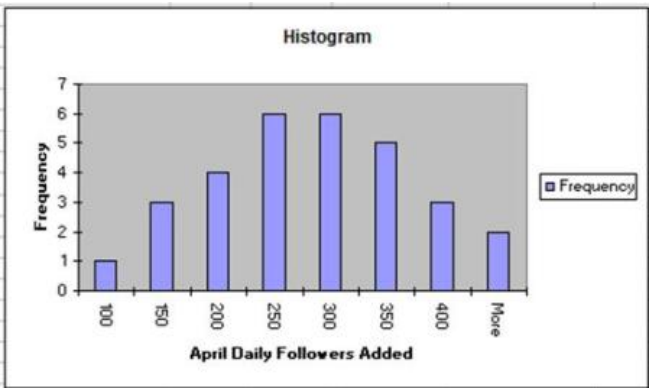
Here is a close-up of the dialogue box that was just shown:

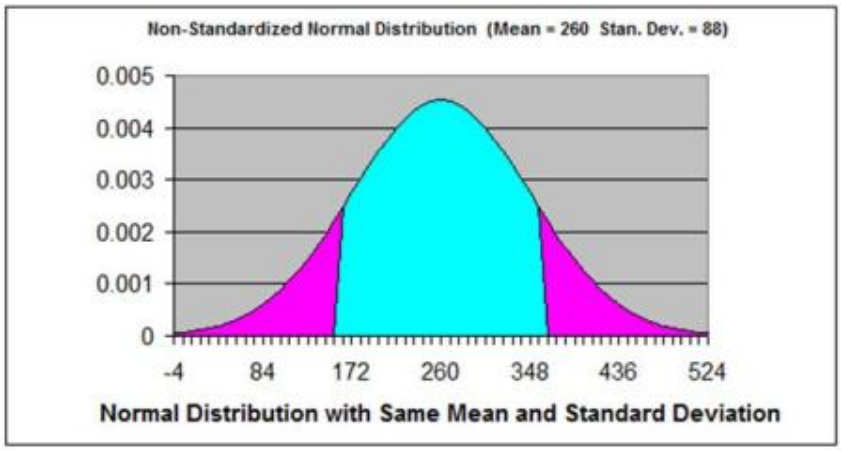


Hitting the OK button will give a completed Histogram that will look like this:



Compare this to a Normal curve with the same mean and standard deviation. The data is Normally distributed.





The Normal Probability Plot
A Simple, Quick Normality Test for Excel

Another normality test that is very easy to implement in Excel is called the Normal Probability Plot. There are 2 ways to create the Normal Probability Plot. They both create the same output. I use the 1st method because it is accompanied with an explanation of why the method works. I personally have difficulty with applying a method that I don't understand. Here are both methods, starting with my preferred choice:

Creating the Normal Probability Plot - Method 1

One characteristic that defines the Normal distribution is that Normally-distributed data will have the same amount of area of Normal curve between each point. For example, if there were 7 sampled points total that were perfectly Normally-distributed, The area under the Normal curve between each point would contain 1/7 of the total area under the Normal curve.

The area under the Normal curve between 2 points can be determined by using the CDF (Cumulative Distribution Function) as follows:

Using the CDF To Calculate Area Between 2 Points On the Normal Curve

We can obtain the normal curve area between two sample points (on the X-axis) by using the **Cumulative Distribution Function (CDF).** The CDF at any point on the x-axis is the total area under the curve to the left of that point. We can obtain the percentage of area in normal curve for each regionby subtracting the CDF at the x-Value of region's lower boundary from the CDF at the x-Value of the region's upper boundary.

The normal distribution that we are trying to fit data has as its two and only parameters the sample's mean and standard deviation.

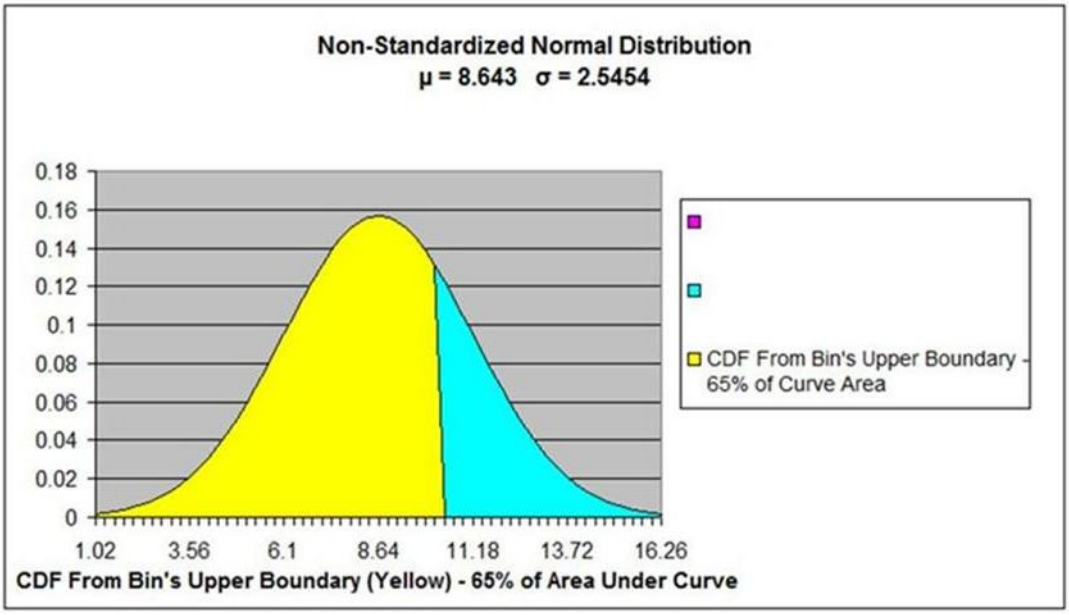
The CDF of this normal distribution at any point on the x-Axis can be determined by the following Excel formula:

CDF = NORMDIST (x Value, Sample Mean, Sample Standard Deviation, TRUE)

Once again, this formula calculate the CDF at that x Value, which is the area under the normal curve to the left of the x Value. That normal curve has as its parameters the sample's mean and standard deviation.

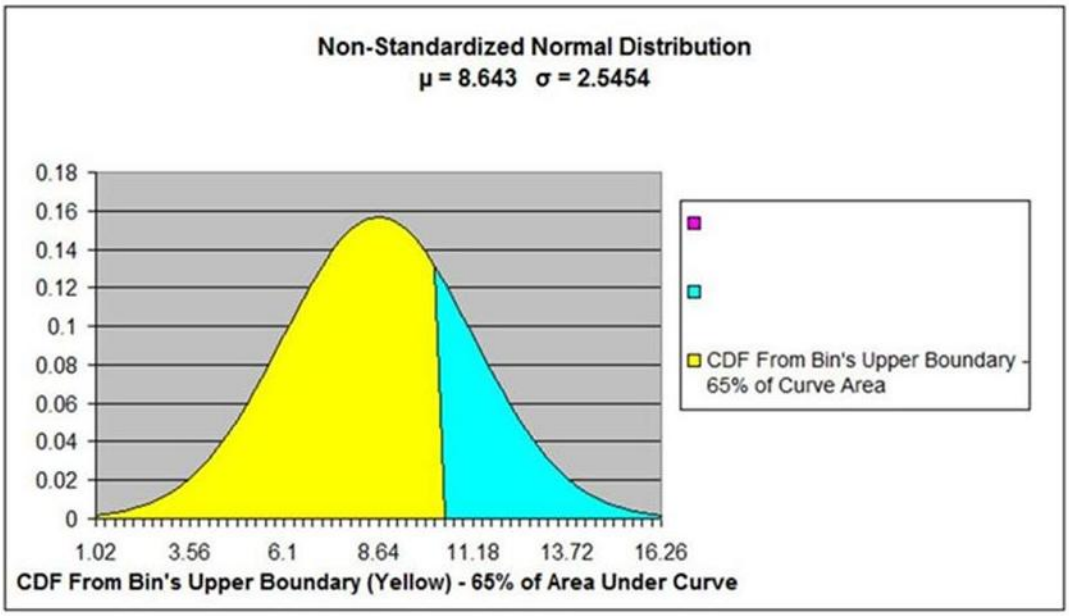
Graphical Interpretation of the CDF

CDF (65% of Curve Area From Upper Boundary of Region)



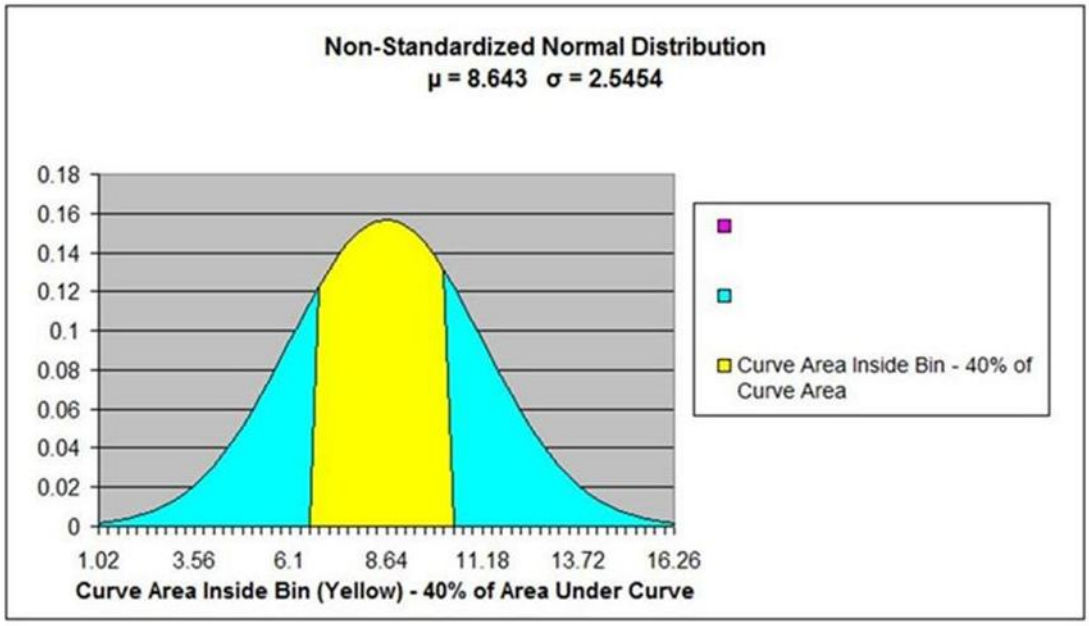
MINUS

CDF (25% of Curve Area From Lower Boundary of Region)



EQUALS

25% of Curve's Total Area Is Inside Region



Given the above, here are the Steps to creating a Normal Probability Plot to evaluate the Normality of sampled data.

Here is a set of 7 sampled points that we are going test for Normality using the Normal Probability Plot:

Actual Sample Values
-4.0
-3.0
0.8
1.8
3.9
6.2
6.5

From these samples, we need to calculate sample size (count - number of samples), sample mean, and sample standard deviation. Here are those calculations:

Sample Mean =	1.743
Sample Stan. Dev. =	4.154
n =	7

Given the above sample size, mean, and standard deviation, if the sample were perfectly Normally-distributed, the sample would have been as follows:

Expected Sample Values	CDF At Each Sample Point For n Samples If Sample Is Normally Distributed	Z Score At Each Sample Point If Sample Is Normally Distributed
-4.343	1/14	-1.465
-1.545	3/14	-0.792
0.222	5/14	-0.366
1.743	1/2	0.000
3.264	9/14	0.366
5.031	11/14	0.792
7.829	13/14	1.465

If there are 7 sampled data points that were perfectly Normally distributed, there would be 1/7 of the total Normal curve area between each sampled point.

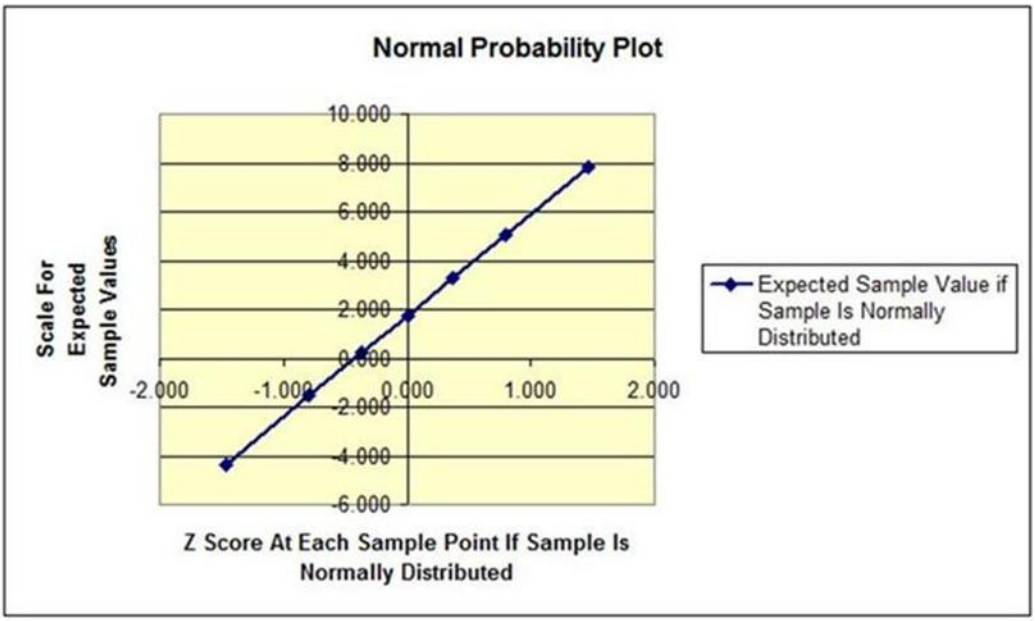
The Z Score at each sampled point are found with the following Excel formula:

NORMSINV (CDF at each Sample Point)

The Expected Sample Values are found by the following Excel formula:

NORMINV (CDF at Sample Point, Sample Mean, Sample Stan. Dev.)

A graph of Expected Sample Values vs. Z Score will be a straight line, as follows:

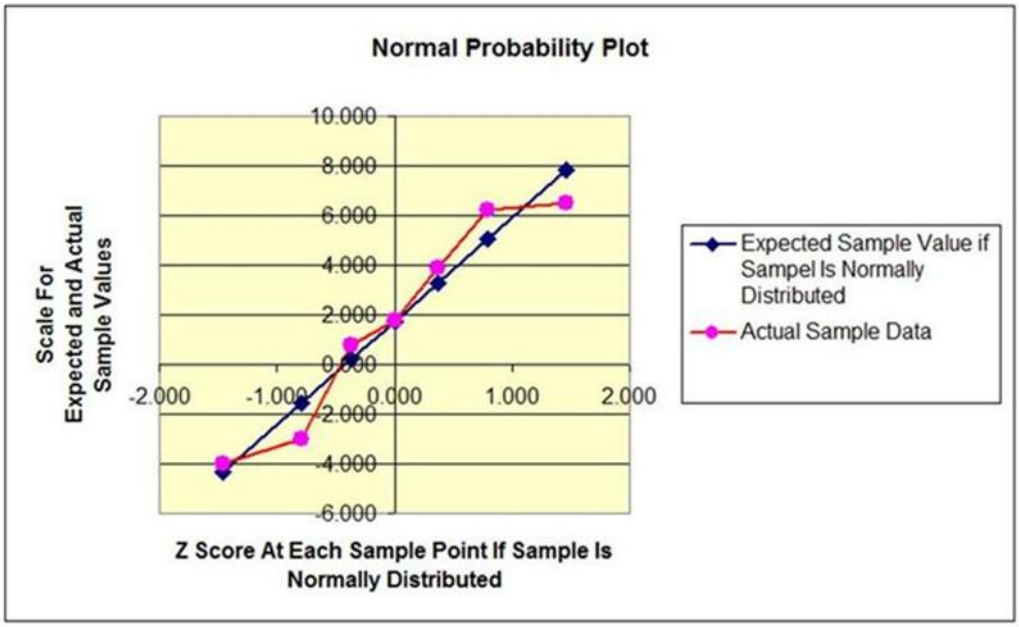


We now observe the actual data samples compared to the Expected Data Samples for Normally-distributed data having the same mean and standard deviation:

Actual Sample Values
-4.0
-3.0
0.8
1.8
3.9
6.2
6.5

Actual Sample Values	Expected Sample Values	CDF At Each Sample Point For n Samples If Sample Is Normally Distributed	Z Score At Each Sample Point If Sample Is Normally Distributed
-4.0	-4.343	1/14	-1.465
-3.0	-1.545	3/14	-0.792
0.8	0.222	5/14	-0.366
1.8	1.743	1/2	0.000
3.9	3.264	9/14	0.366
6.2	5.031	11/14	0.792
6.5	7.829	13/14	1.465

We now wish to see how close the Actual Sample Values graph to the straight line of the Expected Sample Values, as follows:



We can see that the Actual Sample Data (in purple) maps closely to the Expected Sample Values (in dark blue) so we conclude that the data appears to be derived from a Normally-distributed population. **One caution**: **A larger sample size (at least 50) should be used to obtain valid results.** The small sample size (7) was used here for simplicity.

Creating the Normal Probability Plot - Method 2

The data set is ranked in order and then plotted on a graph. Each point in the data set represents a y value of a plotted point. The x values of the points are Normal Order Statistic Medians. The closer than the graph is to a straight line, the more closely the data set resembles the normal distribution. Correlation analysis can also be performed the data set (called the Order Responses) and the Normal Order Statistic Medians. The closer the correlation coefficient is to 1, the more the data set resembles the normal distribution.

An Example

An example is the best way to illustrate the Normal Probability Plot. Evaluate the following data set of 6 points for normality:

{66, 76, 17, 23, 44, 41}

The rank of each data point is:

5, 6, 1, 2, 4, 3

The data in ranked order is:

{17, 23, 41, 44, 66, 76}

Now we have to calculate the **Normal Order Statistic Medians**. We know that we have 6 points so n = 6. The **Normal Order Statistic Medians** are given by the following formula:

N(i) = G(U(i))

U(i) are the Uniform Order Statistic Medians defined by this formula:

m(i) = 1 - m(n) for i = 1

m(i) = (i - 0.3175)/(n + 0.365) for i = 2, 3, ..., n-1

m(i) = 0.5(1/n) for i = n

G is called the Percent Point of the Normal Distribution. It is the inverse of the cumulative distribution function. In Excel, it would be the NORMSINV(x) function. It tells you the probability the x has a value of m(i) or less. Variable x is normally distributed on a standard normal curve (μ = 0 and σ = 1).

Given the above information, here is how the **Normal Order Statistic Medians** are calculated:

n = 6

Now calculate **U(i) – the Uniform Order Statistic Medians**.

U(i) are the **Uniform Order Statistic Medians** defined by this formula:

m(i) = 1 - m(n) for i = 1

m(i) = (i - 0.3175)/(n + 0.365) for i = 2, 3, ..., n-1

m(i) = 0.5(1/n) for i = n

i = 1 -->
m(1) = 1 – m(n) = 1 – m(6) = 1 – 0.8909 = 0.1091

i = 2 -->
m(2) = (i - 0.3175)/(n + 0.365) = (2 – 0.3175) / (6 + 0.365) = 0.2643

i = 3 -->
m(3) = (i - 0.3175)/(n + 0.365) = (3 – 0.3175) / (6 + 0.365) = 0.4214

i = 4 -->
m(4) = (i - 0.3175)/(n + 0.365) = (4 – 0.3175) / (6 + 0.365) = 0.5786

i = 5 -->
m(5) = (i - 0.3175)/(n + 0.365) = (5 – 0.3175) / (6 + 0.365) = 0.7357

i = 6 -->
m(6) = m(i) = 0.5(1/n) for i = n = m(i) = 0.5(1/6) = 0.8909

So,

U(1) = 0.1091
U(2) = 0.2643
U(3) = 0.4214
U(4) = 0.5786
U(5) = 0.7357
U(6) = 0.8909

The **Normal Order Statistic Medians** are given by the following formula: N(i) = G(U(i)) --> G(U(i)) is the inverse of the cumulative distribution function. It tells the x value that corresponds to the probability U(i) that a random sample taken from a standardized normally distributed population will have a value of x or less.

This is found in Excel by the following formula:

N(i) = G(U(i)) = NORMSINV(U(i))

So, the **Normal Order Statistic Medians** are given by G(U(i)) = NORMSINV(U(i))

N(1) = NORMSINV(U(1)) = NORMSINV(0.1091) = -1.23
N(2) = NORMSINV(U(2)) = NORMSINV(0.2643) = - 0.63
N(3) = NORMSINV(U(3)) = NORMSINV(0.4214) = - 0.20
N(4) = NORMSINV(U(4)) = NORMSINV(0.5786) = 0.20
N(5) = NORMSINV(U(5)) = NORMSINV(0.7357) = 0.63
N(6) = NORMSINV(U(6)) = NORMSINV(0.8908) = 1.23

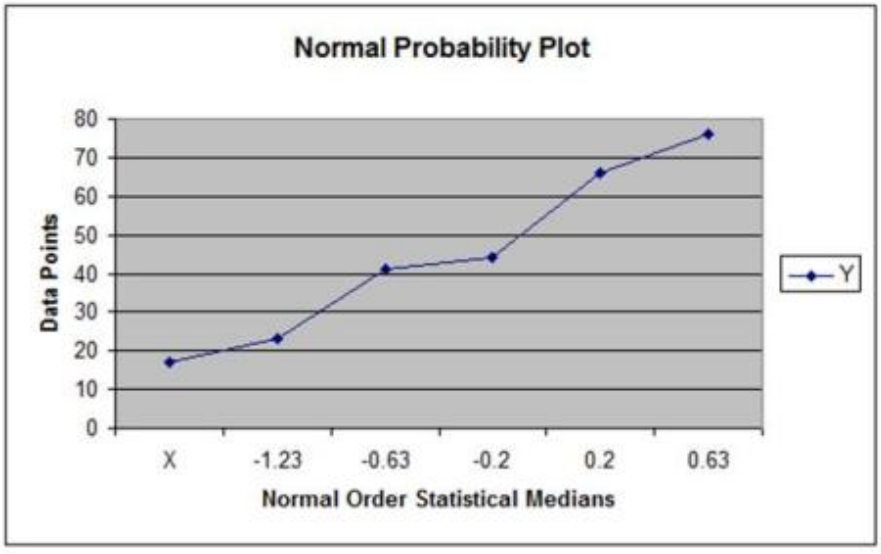
The above are the X values of the data points whose Y values are the ranked point in the data set. The ranked data set is:

{17, 23, 41, 44 66, 76}

So, the following points can be plotted:

(-1.23, 17) (-0.63, 23) (-0.20, 41) (0.20, 44) (0.63, 66) (1.23, 76)

The final graph of these plotted point will resemble this chart:



The closer that the plotted resembles a straight line, the closer the data set resembles the normal distribution. You can also run correlation analysis between the data set of Ordered Responses and the Normal Order Statistic Medians. The closer the correlation coefficient is to 1, the more closely the data set resembles the normal distribution.

There are other well-known Normality tests such as the **Chi-Square Goodness-of-Fit Test.** This is shown in the next section in detail.

If you are going to perform any statistical analysis that uses the normal distribution or t distribution such as Z test, t tests, F tests, and chi-square tests, you should first test your data set for normality. The Normal Probability Plot described in this article is probably the easiest and quickest way to do it in Excel.

The Chi-Square Goodness-Of-Fit Test

The Easiest and Most Robust

Normality Test In Excel

As a marketer, anytime that you are running a t Test, and regression, a correlation, or ANOVA, you should make sure you're working with normally distributed data, or your test results might not be valid . The quick-and-dirty Excel test is simply to throw the data into an Excel histogram and eyeball the shape of the graph. If there is a still a question, the next (and easiest) normality test is the Chi-Square Goodness-Of-Fit test.

-

The Chi-Square Goodness-Of-Fit test is less well known than some other normality test such as the Kolmogorov-Smirnov test, the Anderson-Darling test, or the Shapiro-Wilk test. The Chi-Square Goodness-Of-Fit test is, however, a lot less complicated, every bit as robust, and a whole lot easier to implement in Excel (by far) than any of the more well-known normality tests. Let's run through an example:

The 1st Step of Normality Testing

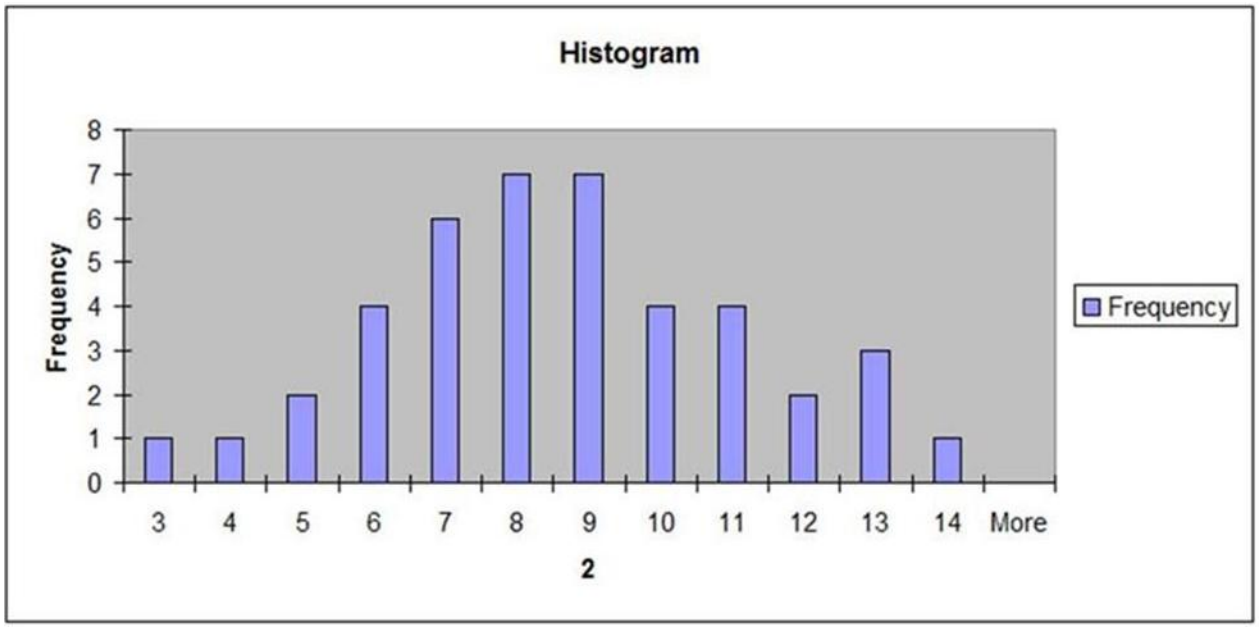
Graph the Data in an Excel Histogram

Here is the data we wish to test to determine if it is normally distributed:

Tuesday Sales			
3	7	8	5
4	8	9	6
5	9	10	7
6	10	11	8
7	7	6	9
8	8	7	13
9	9	8	11
10	10	9	12
11	8	13	6
12	9	11	7
13	14		

The 1st step in the data analysis is to create an Excel Histogram. The resulting histogram for the above data is as follows:

	Frequency
3	1
4	1
5	2
6	4
7	6
8	7
9	7
10	4
11	4
12	2
13	3
14	1
More	0



The histogram above somewhat resembles a normal distribution, but we should still apply a more robust test to it to be sure. The Chi-Square Goodness-of-Fit test in Excel is both robust and easy to perform, understand, and explain to others. Here is how to perform this test on the above data.

The 1st Step of the Chi-Square

Goodness-Of-Fit Test in Excel

We need to know the **mean**, **standard deviation**, and **sample size** of the data that we are about to test for normality. Use the Descriptive Statistics Excel tool to obtain this information. In Excel 2003, this tool can be found at **Tools / Data Analysis / Descriptive Statistics**. The resulting output for this test is as follows:

Tuesday Sales	
Mean	**8.643**
Standard Error	0.392765
Median	8.5
Mode	8
Standard Deviation	**2.5454**
Sample Variance	6.479094
Kurtosis	-0.29534
Skewness	0.095549
Range	11
Minimum	3
Maximum	14
Sum	363
Count	**42**

How the Chi-Square Goodness-Of-Fit Test Works

New that we have the sample mean, standard deviation, and sample size, we are ready to perform the Chi-Square Goodness-Of-Fit test on the data in Excel.

The Chi-Square Goodness-Of-Fit test is a **hypothesis test**. The Null and Alternative Hypotheses being tested are:

H0 = The data follows the normal distribution.

H1 = The data does not follow the normal distribution.

A quick summary of the test is as follows:

We divide the observed samples into groups that have the same boundaries as the bins that were established when the Histogram was created in Excel. In this case, the observed samples fell into the following bins:

3 to 4 - 1 sample had a value in this range
4 to 5 - 1 sample had a value in this range
5 to 6 - 2 samples had a value in this range
6 to 7 - 4 samples had a value in this range
7 to 8 - 6 samples had a value in this range
8 to 9 - 7 samples had a value in this range
9 to 10 - 7 samples had a value in this range
10 to 11 - 4 samples had a value in this range
11 to 12 - 4 samples had a value in this range
12 to 13 - 3 samples had a value in this range
13 to 14 - 1 sample had a value in this range

	Frequency
3	1
4	1
5	2
6	4
7	6
8	7
9	7
10	4
11	4
12	2
13	3
14	1
More	0

The figures above represent the observed number of samples in each bin range. We now need to calculate how many sample we would expect to occur in each bin if the sample was normally distributed with the same mean and standard deviation as the sample taken (mean = 8.634 and standard deviation = 2.5454).

The expected number of sample in each bin is calculated by the following formula:

(Area of the normal curve bounded by the bin's upper and lower boundaries) x (Total number of samples taken)

For example, if there were only 2 bins that meet at the mean, then the corresponding normal curve would have 2 regions with a boundary at the mean of the normal curve. Each of the two regions of the normal curve would contain 50% of the area under the entire normal curve. We would therefore expect 50% of the total number of samples taken to fall in each bin. If, for example, 42 samples were taken, we would expect 21 samples to occur in each bin if the samples were normally distributed.

Given the bin ranges we have established for the Excel Histogram and the number of observed samples in each bin, we now need to calculate the number of samples we would expect to find in each bin. We assume that the samples are normally distributed with the same mean and standard deviation as measured from the actual

sample. Given these assumptions, we use the method described above to calculate how many samples would be expected to occur in each bin.

Once we know the observed and expected number of samples in each bin, we calculate the Chi-Square Statistic.

A **Chi-Square Statistic** is created from the data using this formula:

Chi-Square Statistic = Σ [[(Expected num. - Observed num.)^2] / (Expected num)]

A **p Value** is calculated in Excel from this Excel formula:

p Value = CHIDIST (Chi-Square Statistic, Degrees of Freedom)

We take all of the samples and divide them up into groups. These groups are called bins. We will use the same bins as was used when creating the Histogram in Excel. The bins are as follows:

The size of the p Value determines whether or not we go with the assumption that the samples are normally distributed.

The Decision Rule

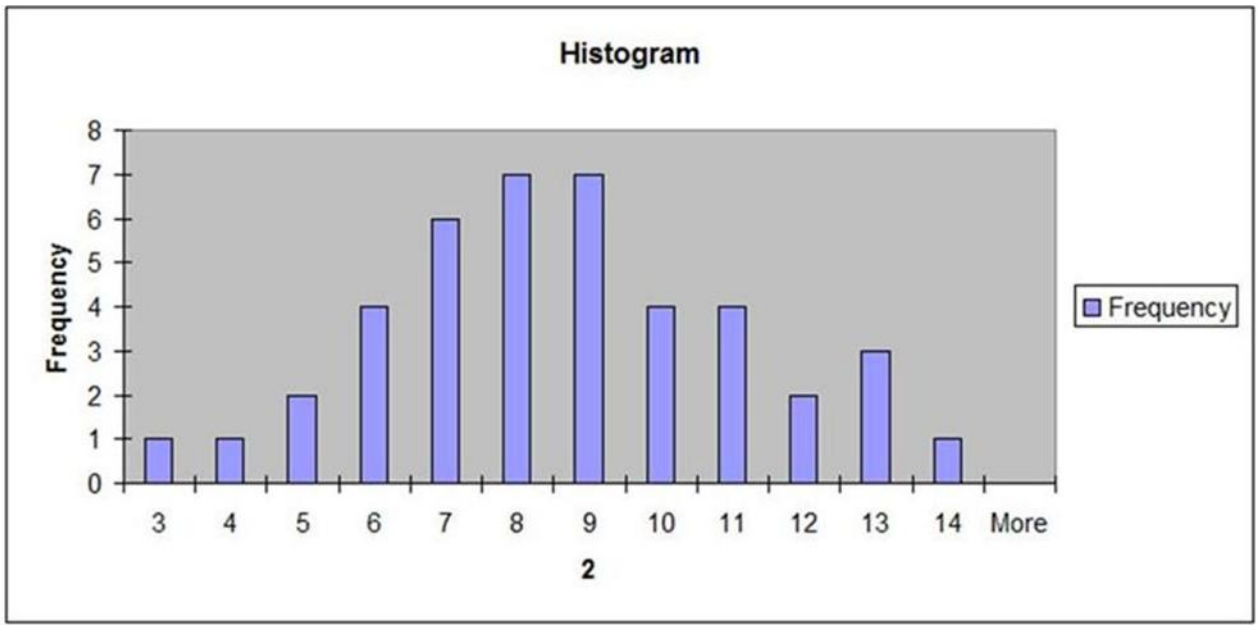
If the resulting p Value is less than the **Level of Significance**, we reject the Null Hypothesis and state that we cannot state within the required Degree of Certainty that the data is normally distributed. In other words, if we would like to state within 95% certainty that the data can be described by the normal distribution, the Level of Significance is 5%. The Level of Significance = 1 - Required Degree of Certainty. If the resulting p Value is greater than 0.05, we can state with at least 95% certainty that the data is normally distributed.

Breaking the Normal Curve into Regions

The Chi-Square Goodness-Of-Fit test requires that the normal distribution be broken into sections. In each section we count how many occur. This is our Observed # for each section. The Excel Histogram function has already done this for us. Once again, here is the Excel Histogram output:

	Frequency
3	1
4	1
5	2
6	4
7	6
8	7
9	7
10	4
11	4
12	2
13	3
14	1
More	0

When we created the Excel Histogram from the data, we had to specify how many "bins" the samples would be divided into. Excel counted the number of observed samples in each bin and then plotted the results in the following histogram:



Since Excel has already counted how many observed samples are in each bin, we wil also use the bins as our sections for the Chi-Square Goodness-Of-Fit test. We know how many actual samples have been observed in each bin. We now need to calculate how many samples would have been expected to occur in each bin.

Calculating the Expected Number of Samples in Each Bin

The size of each bin determines how many samples would have been expected to occur in that bin. Each bin represents a percentage of the total area under the distribution curve that we are evaluating. That percentage of the total area that is associated with a bin represents the probability that each observed sample will be drawn from that bin.

Here is a simple example that will hopefully clarify the above paragraph. If we were evaluating a data set for normality, we would be trying to determine whether the data fits the normal curve. We have to determine what the bins ranges that we will divide the data into. The simplest bin arrangement would be to place all the data into only two bins on either side of the sample's mean. If the data were normally distributed, we would expect half of the samples to occur in each bin.

In other words, if the bins were placed along the x-axis relative to the sample's mean so each bin would be directly under 50% of a normal curve with the same mean, then we would expect 50% of the samples to occur in each bin. If there were 60 total samples taken, we would expect 30 samples to occur in each bin.

The expected number of samples for a single bin = **Exp**.

Exp. = (Area under the normal curve over the top of the bin) x (Total number of samples)

Calculating the CDF

We can obtain the normal curve area over each bin by using the **Cumulative Distribution Function (CDF).** The CDF at any point on the x-axis is the total area under the curve to the left of that point. We can obtain the percentage of area in normal curve for each bin by subtracting the CDF at the x-Value of bin's lower boundary from the CDF at the x-Value of the bin's upper boundary.

The normal distribution that we are trying to fit data has as its two and only parameters the sample's mean and standard deviation.

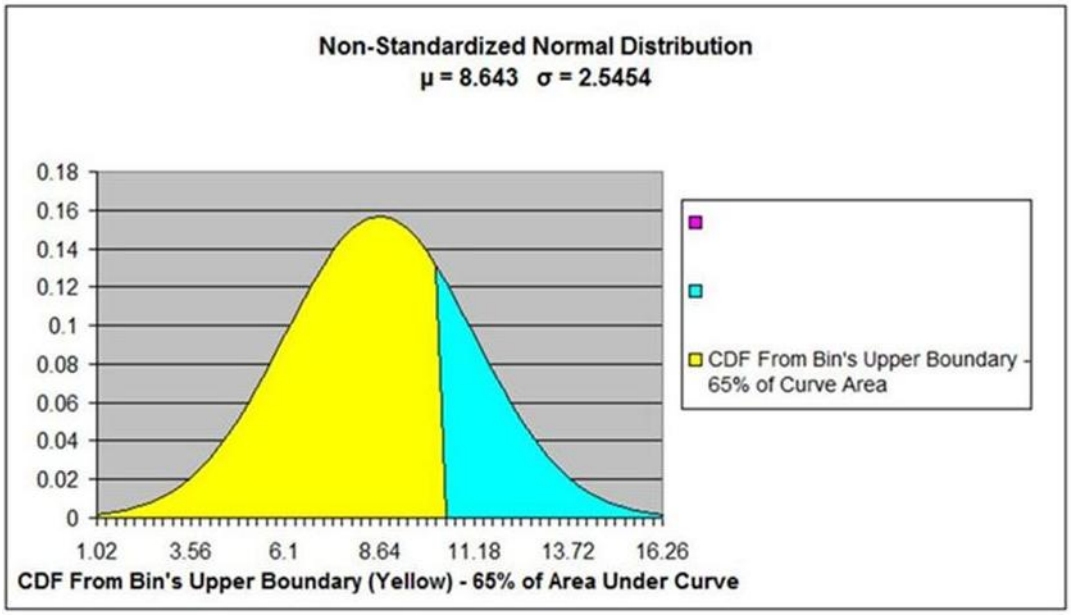
The CDF of this normal distribution at any point on the x-Axis can be determined by the following Excel formula:

CDF = NORMDIST (x Value, Sample Mean, Sample Standard Deviation, TRUE)

Once again, this formula calculate the CDF at that x Value, which is the area under the normal curve to the left of the x Value. That normal curve has as its parameters the sample's mean and standard deviation.

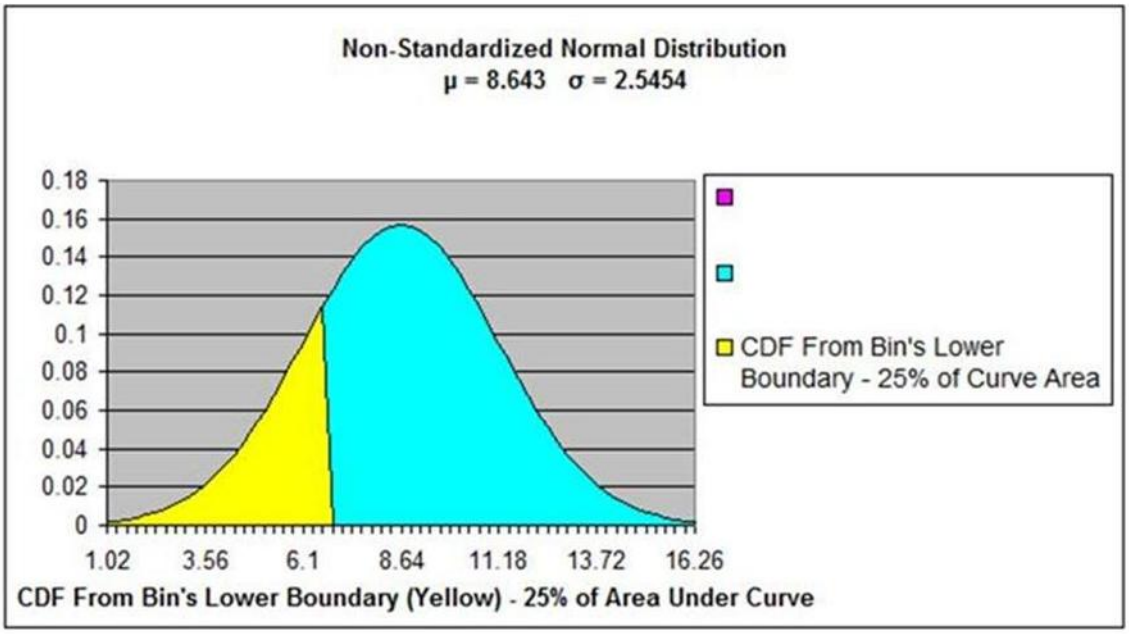
Graphical Interpretation of the CDF

CDF (65% of Curve Area From Upper Boundary of Bin)



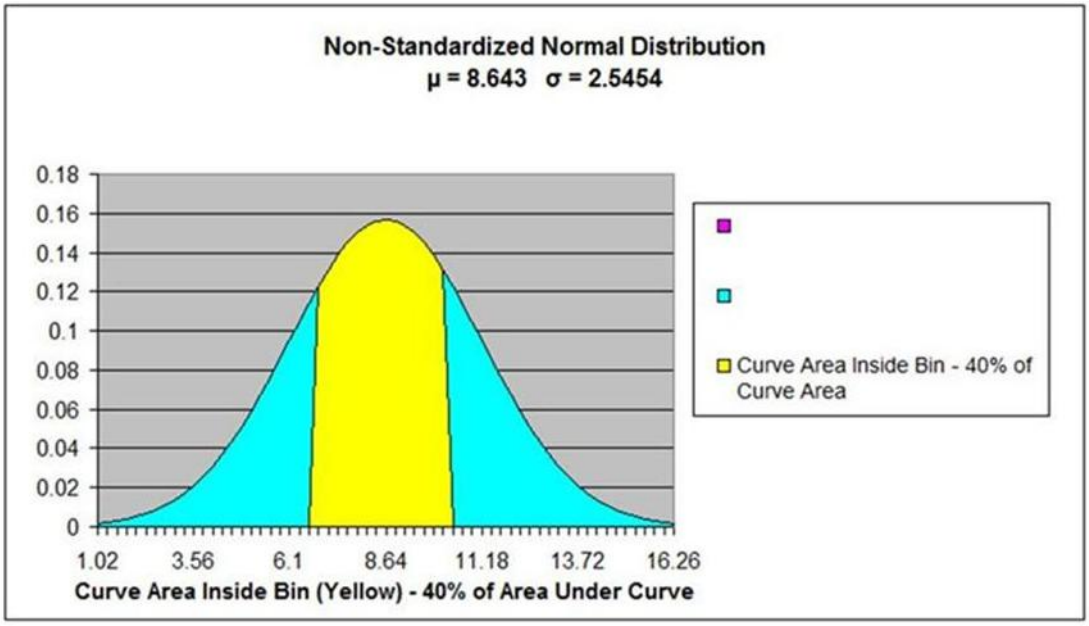
MINUS

CDF (25% of Curve Area From Lower Boundary of Bin)



EQUALS

25% of Curve's Total Area Is Inside Bin



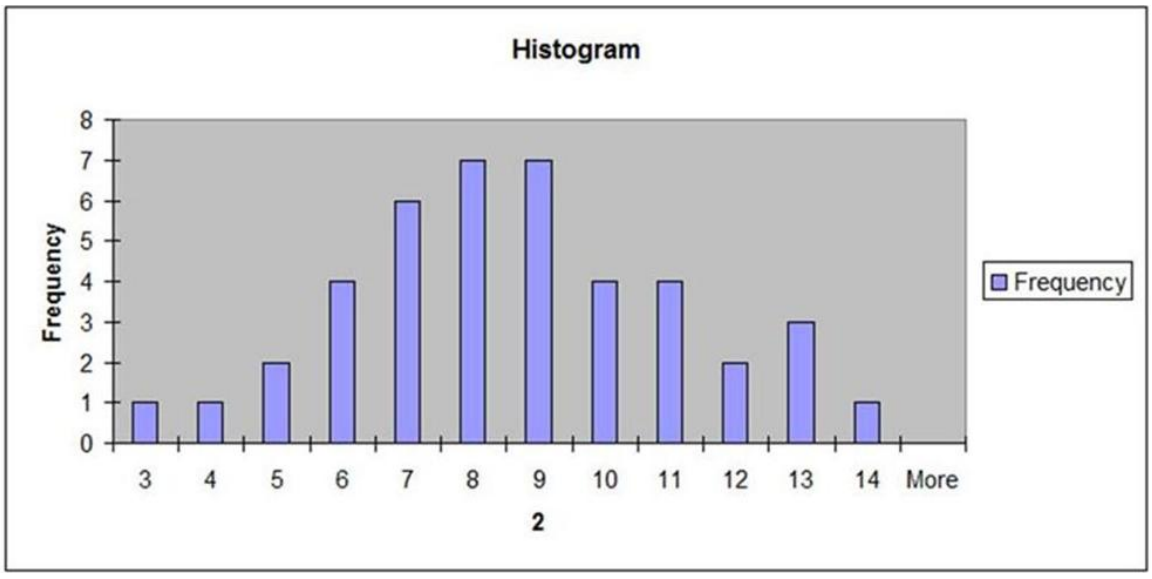
Once again, here is the original data and the calculations performed in Excel using the Histogram bin ranges and a sample mean of 8.643 and standard deviation of 2.5454:

Tuesday Sales			
3	7	8	5
4	8	9	6
5	9	10	7
6	10	11	8
7	7	6	9
8	8	7	13
9	9	8	11
10	10	9	12
11	8	13	6
12	9	11	7
13	14		

Tuesday Sales	
Mean	**8.643**
Standard Error	0.392765
Median	8.5
Mode	8
Standard Deviation	**2.5454**
Sample Variance	6.479094
Kurtosis	-0.29534
Skewness	0.095549
Range	11
Minimum	3
Maximum	14
Sum	363
Count	**42**

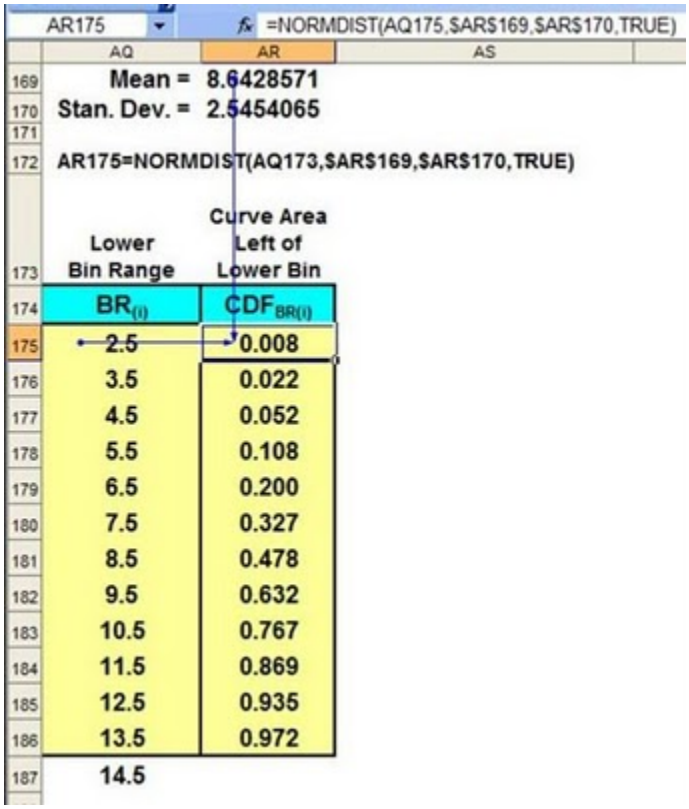
Histogram Output

	Frequency
3	1
4	1
5	2
6	4
7	6
8	7
9	7
10	4
11	4
12	2
13	3
14	1
More	0

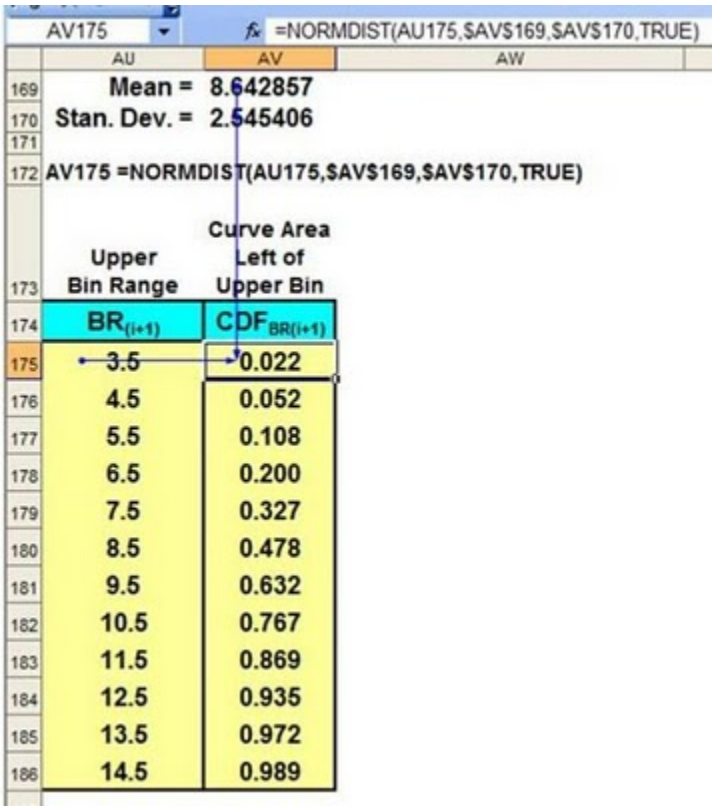


Lower Bin Range	Upper Bin Range	Actual Number of Observations In Each Bin
$BR_{(i)}$	$BR_{(i+1)}$	Obs.
2.5	3.5	1
3.5	4.5	1
4.5	5.5	2
5.5	6.5	4
6.5	7.5	6
7.5	8.5	7
8.5	9.5	7
9.5	10.5	4
10.5	11.5	4
11.5	12.5	2
12.5	13.5	3
13.5	14.5	1
14.5		

Next we calculate the **curve area to the left of the lower edge of each bin**:



Next we calculate the **curve area to the left of the upper edge of each bin**:



Finally we calculate the **percentage of total curve area contained within each bin** by subtracting the area to the left of the bin's lower edge from the curve area to the left of the curve's upper edge in Excel as follows:

Curve Area Left of Lower Bin	Curve Area Left of Upper Bin	Area in Bin =
$CDF_{BR(I)}$	$CDF_{BR(I+1)}$	$CDF_{BR(I+1)} - CDF_{BR(I)}$
0.008	0.022	0.014
0.022	0.052	0.030
0.052	0.108	0.057
0.108	0.200	0.091
0.200	0.327	0.127
0.327	0.478	0.151
0.478	0.632	0.154
0.632	0.767	0.135
0.767	0.869	0.102
0.869	0.935	0.066
0.935	0.972	0.037
0.972	0.989	0.017

We can now **calculate the Expected number of samples in each bin** by the following formula:

Exp. number of samples in each bin =

(Percentage of Curve Area in that Bin) x Total number of samples

This calculation for each bin is completed in the 1st column below. There are 42 total samples taken for this exercise.

Area in Bin = $CDF_{BR(i+1)} - CDF_{BR(i)}$		Expected Number of Observations = Bin Area * n (42) Exp.
0.014	x 42 =	0.578
0.030	x 42 =	1.266
0.057	x 42 =	2.380
0.091	x 42 =	3.842
0.127	x 42 =	5.325
0.151	x 42 =	6.338
0.154	x 42 =	6.477
0.135	x 42 =	5.684
0.102	x 42 =	4.283
0.066	x 42 =	2.771
0.037	x 42 =	1.540
0.017	x 42 =	0.735

Expected Number of Observations = Bin Area * n (42)	Actual Number of Observations	
Exp.	Obs.	Exp - Obs.
0.578	1	-0.422
1.266	1	0.266
2.380	2	0.380
3.842	4	-0.158
5.325	6	-0.675
6.338	7	-0.662
6.477	7	-0.523
5.684	4	1.684
4.283	4	0.283
2.771	2	0.771
1.540	3	-1.460
0.735	1	-0.265

Exp - Obs.	Exp.	$(Exp - Obs)^2 / (Exp)$
-0.422	0.578	0.308
0.266	1.266	0.056
0.380	2.380	0.061
-0.158	3.842	0.007
-0.675	5.325	0.086
-0.662	6.338	0.069
-0.523	6.477	0.042
1.684	5.684	0.499
0.283	4.283	0.019
0.771	2.771	0.215
-1.460	1.540	1.385
-0.265	0.735	0.096

The end result of the Excel calculations is the final column below of (Exp. - Obs.)^2 / Exp. for each bin. These figures are then summed as follows to give us **the overall Chi-Square Statistic for the sample data**. In this case, the sample data's Chi-Square Statistics is 2.841.

$(Exp - Obs)^2 / (Exp)$
0.308
0.056
0.061
0.007
0.086
0.069
0.042
0.499
0.019
0.215
1.385
0.096

Chi-Square Statistic =

χ^2 = = Sum of Above

χ^2 = 2.841

The Degrees of Freedom

The Chi-Square-Goodness-Of-Fit test requires the number of Degrees of Freedom be calculated for the specific test being run. The formula for this is as follows:

Degrees of Freedom = df = (number of filled bins) - 1 - (number of parameters calculated from the sample)

The number of filled bins = 12

We calculated the mean and standard deviation from the sample. This is 2 parameters.

df = 12 - 1 - 2 = 9

P Value Calculation

We can now **calculate the p Value** from Chi-Square Statistics and the Degrees of Freedom as follows:

Chi-Square Statistic =

χ^2 = = Sum of Above

χ^2 = 2.841

df = (# bins) - 1 - (# of parameters)

df = 12 - 1 - 2

df = 9

p Value = Prob. Of Normal Distribution = CHIDIST (x , df)

= CHIDIST (2.841 , 9)

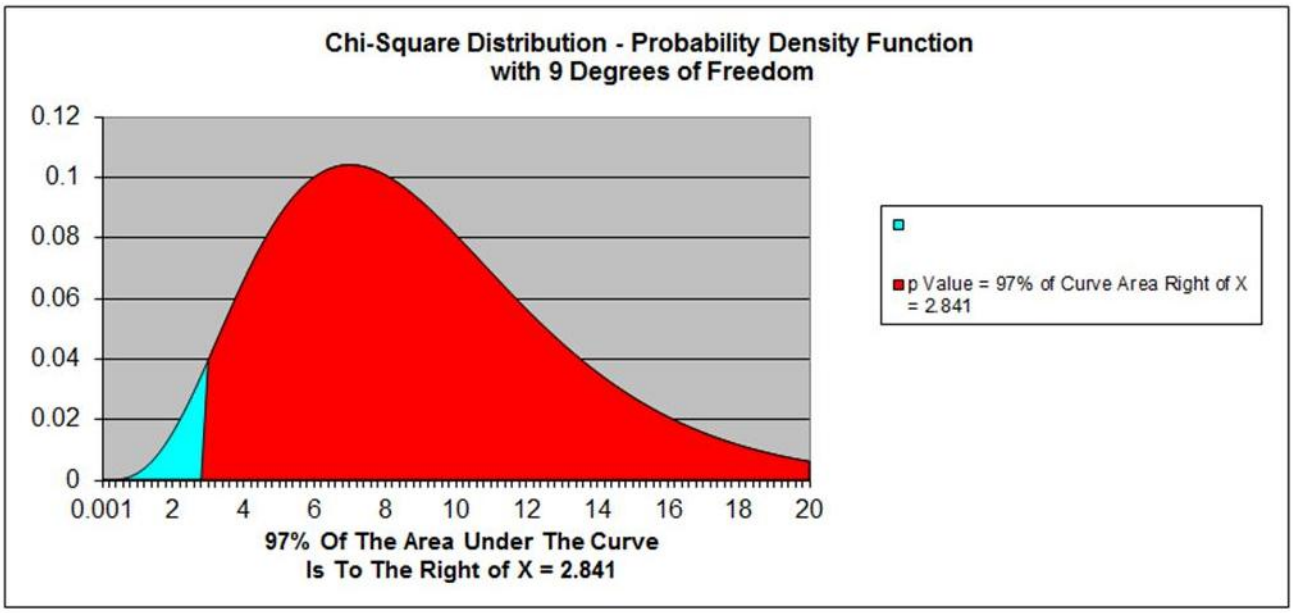
= 0.970263312

= 97.03%

The p Value's Graphical Interpretation

The p Value's graphical interpretation is shown below. The p Value represents the percentage of area (in red) to the right of X = 2.841 under a Chi-Square distribution with 9 Degrees of Freedom. If the p Value (.9703) is greater than the Level of Significance (0.05), we do not reject the Null Hypothesis.

In this case, we state that we do not reject the Null Hypothesis and do not have sufficient evidence that the data is not normally distributed.



Comparing Solving Chi-Square Problems With the p Value vs. Using the Critical Value

Chi-Square problems, such as the Goodness-Of-Fitness Test shown here or the Chi-Square Independence Test, can be solved using either of two equivalent ways. They can be solved by comparing the p Value with alpha as we did here, or they can be solved by comparing the Chi-Square Value with the Chi-Square Critical Value. Let's briefly look at both methods:

1) Comparing the p Value With Alpha

We used this approach here. We found the p Value (0.9703) to be greater than the Level of Significance (Alpha, 0.05) so we do not reject the Null Hypothesis that states that data is not different than we would expect it to be if it were normally distributed.

The p Value is the area under the curve to the right of the Chi-Square Value on the X-axis. In the problem above we calculated the Chi-Square Value to be 2.841. The p Value equals the percentage of area under the curve to the right of X = 2.841 (in red on the graph). In this case the p Value = 0.9703 or 97%. This is greater than Alpha, which equals 0.05 or 5%).

2) Comparing the Chi-Square Value With the Chi-Square Critical Value

The p Value is the area under the curve to the right of the Chi-Square Value on the X-axis. In the problem above we calculated the Chi-Square Value to be 2.841. The p Value equals the percentage of area under the curve to the right of X = 2.841 (in red on the graph). In this case the p Value = 0.9703 or 97%. This is greater than Alpha, which equals 0.05 or 5%).

Equivalently, we could compare Chi-Square Value (2.841) to the Critical Chi-Square Value. The Critical Chi-Square Value is the point on the X-axis that Alpha (0.05 or 5% of the area under the curve is to the right of). If the Critical Chi-Square Value is greater than the Chi-Square Value, we do not reject the Null Hypothesis.

Critical Chi-Square Value in Excel = CHIINV(Alpha,Degrees of Freedom)

=CHIINV(0.05,9) = 16.91

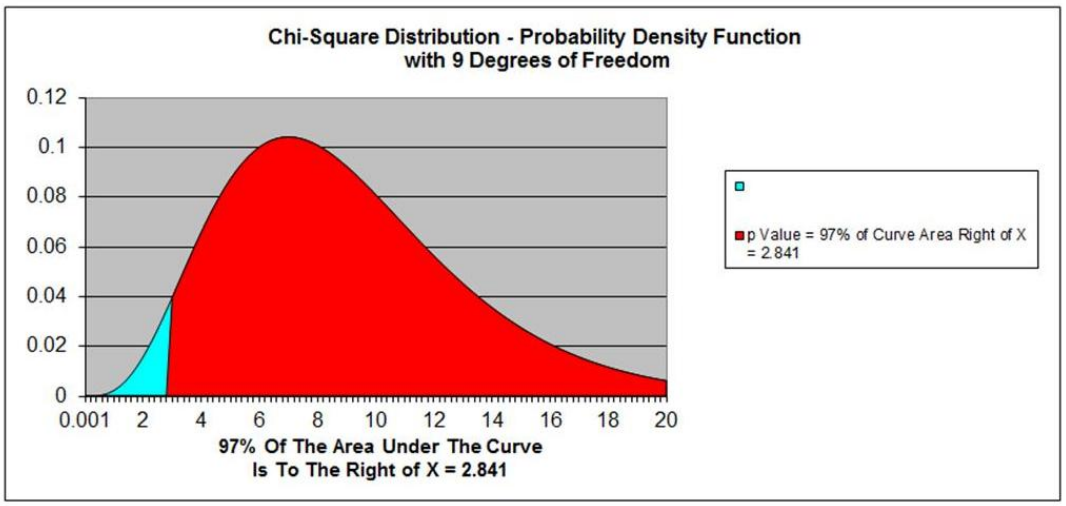
We can see on the above graph that Alpha or 5% of the area under the curve would be to the right of the Chi-Square Critical Value of X = 16.91.

We can also see on the graph that the p Value of 97% of the area under the curve is to the right of the calculated Chi-Square value of X = 2.841.

From this we can see that comparing the p Value (.97) to Alpha (0.05) is equivalent to comparing the Chi-Square Value (2.842) to the Critical Chi-Square Value (16.91) to determine whether to reject the Null Hypothesis, which states that the sample data fits the distribution to which we are comparing the sample data to.

We do not reject the Null Hypothesis if the Chi-Square Value (2.841) is less than the Critical Chi-Square Value (16.91) or, equivalently, the p Value (0.97) is greater than Alpha (0.05). The Chi-Square Value and Critical Chi-Square Value are points on the X-axis. The p Value and Alpha are areas under to curve to the right of those points on the X-axis.

Here is the graphical representation of the answer for reference again:



Meet the Author

Mark Harmon is a master number cruncher. Creating overloaded Excel spreadsheets loaded with complicated statistical analysis is his idea of a good time. His profession as an Internet marketing manager provides him with the opportunity and the need to perform plenty of meaningful statistical analysis at his job.

Mark Harmon is also a natural teacher. As an adjunct professor, he spent five years teaching more than thirty semester-long courses in marketing and finance at the Anglo-American College in Prague and the International University in Vienna, Austria. During that five-year time period, he also worked as an independent marketing consultant in the Czech Republic and performed long-term assignments for more than one hundred clients. His years of teaching and consulting have honed his ability to present difficult subject matter in an easy-to-understand way.

Harmon received a degree in electrical engineering from Villanova University and MBA in marketing from the Wharton School.