

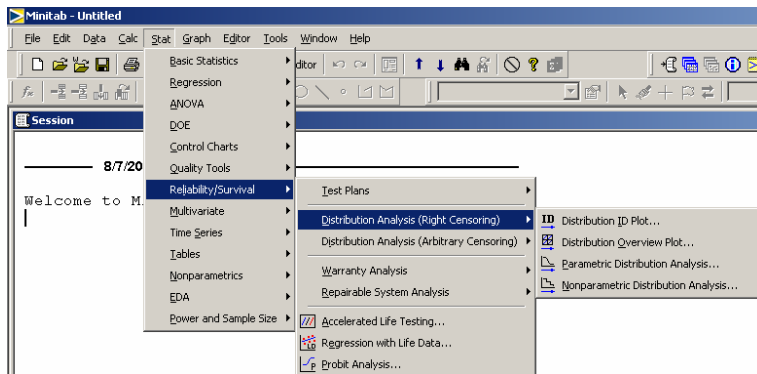
# Censored Data Analysis to Fix “Bad” Data

By: Rick Haynes, Master Black Belt – Smarter Solutions, Inc.

First of all, what is “Bad” data and what is “Censored Data Analysis?”. Let us define “Bad” data as values that are not accurate due to rounding or to data collection errors. Censored data analysis is a method developed to solve reliability issues, which we will discuss using in new ways.

This article will discuss a common method that Black and Green Belts could use that is never covered in a typical course. The analytic methods are very common and included in nearly every full-service statistical package; it is the application of the tool in a non-traditional method that we will cover here.

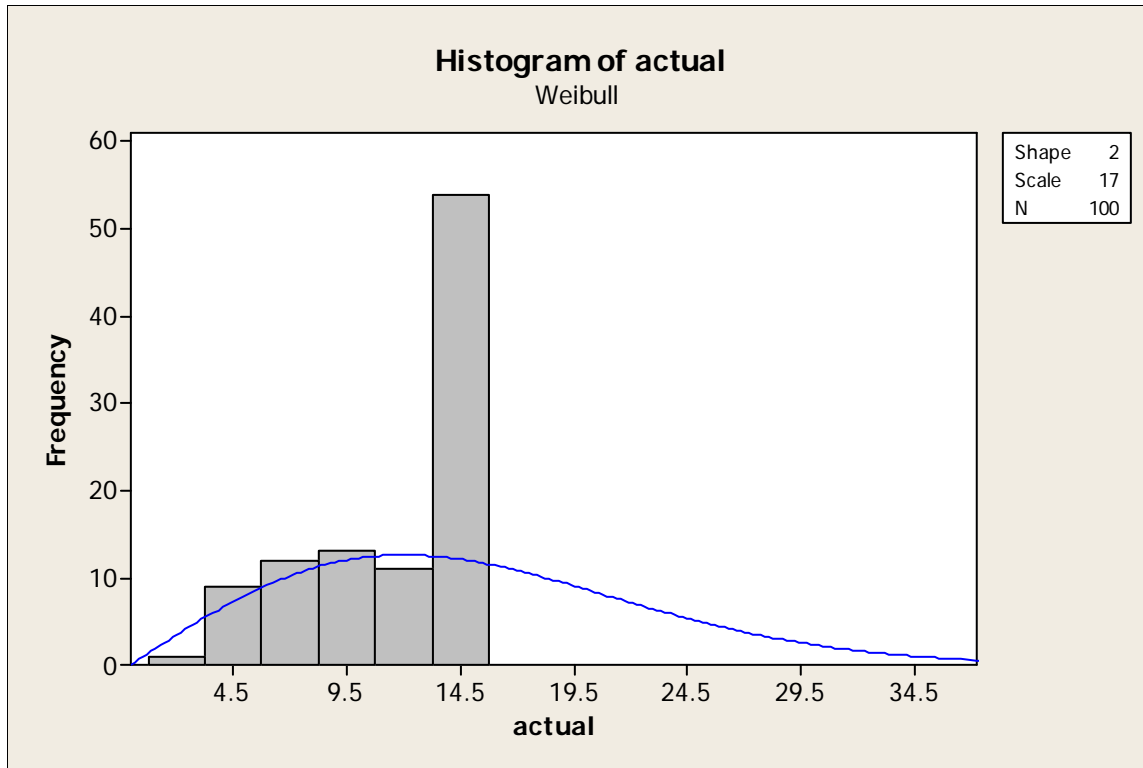
In Minitab, there are a number of reliability and survival methods. It is the 2<sup>nd</sup> and 3<sup>rd</sup> menu items that we will discuss.



Many of us have used the Stat>Reliability/Survival>Distribution Analysis (Right Censoring) > Distribution ID Plot.. function to examine a set of numbers to determine the best-fitting distributions. Since reliability statistics is not a common Belt topic, students may never have known how these functions work. Knowing how they work will open a few doors to non-traditional uses that can be quite helpful.

We will first examine right censoring, and work towards arbitrary censoring, which is the goal of this article. Censoring is just the term for data about which we know something but not everything. Let us consider testing light bulbs to determine their average life. How do you think the average life is calculated? Do we turn on 100 bulbs and time each one until it burns out and then take an average? We could, but it could take a long long time. Generally, the test could involve say 100 light bulbs. We would turn them all on and time them until a significant number had burned out, say one half by day 14. The test is now stopped. What do we know? WE know the exact failure time of 50 bulbs, and we know that 50 have not yet burned out. We actually know that the 50 working bulbs did not burn out by the time the test stopped, at 14 days.

We would consider the 50 failed bulbs as actual data, with the 50 working bulbs to be censored data, censored in that we know only that they will fail at greater than 14 days. A histogram would look like this.



This histogram has the distribution actually used to generate the random data drawn on the curve to show you what the truth should be. Now why is it right censored? Because the true failure value is to the right of the reported time of 14 on all the bulbs that have not failed. The average of the current data is 11.4, which is less than the true average time to fail if we ran all 100 bulbs to failure. To find the estimated average failure time, we need to use the reliability tools in Minitab.

**Distribution ID Plot - Right Censoring**

Variables:  
actual

Frequency columns (optional):

By variable:

Use all distributions

Specify:

Distribution 1: Weibull

Distribution 2: Lognormal

Distribution 3: Exponential

Distribution 4: Normal

**Distribution ID Plot - Censor**

Censoring Options

Use censoring columns:  
censored

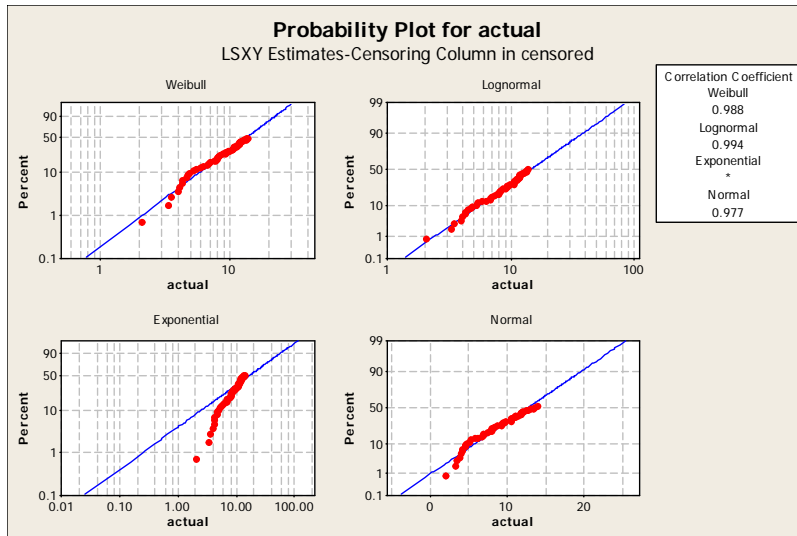
Censoring value: 0

Time censor at:

Failure censor at:

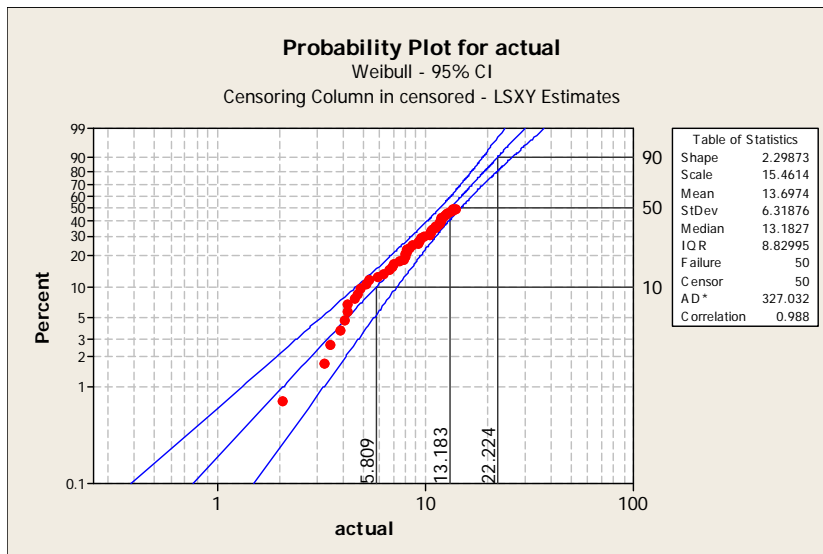
	C2	C3
	censored	actual
1	1	9.5554
2	0	14.0000
3	1	2.5905
4	1	5.0072
5	0	14.0000
6	0	14.0000
7	0	14.0000
8	0	14.0000
9	0	14.0000
10	0	14.0000
11	0	14.0000
12	0	14.0000
13	0	14.0000
14	0	14.0000
15	0	14.0000
16	0	14.0000
17	0	14.0000
18	0	14.0000
19	0	14.0000
20	0	14.0000

For this example, we have a value of 1 if the bulb failed, a 0 if it was censored.



Now I know this is Weibull distributed data because I created it, but Minitab recommends the lognormal distribution as a best guess. Since I know this is wear-out failure data, I will still choose the Weibull, which is the typical distribution of this type of data. But choosing lognormal would still give a reasonable answer.

Using Stat>Reliability/Survival>Distribution Analysis (Right Censoring)>Parametric Distribution Analysis.. we get the following output;



The analysis estimates the mean failure time at 13.7 days, which is closer to the average of the original data of 15. Upon inspection of the probability plot, we should see that only 50 points are plotted, in that most points are below the 50% point. Minitab estimates the probability of getting the actual failure points and accounts for the probability of being in the censored region and plots the data.

If this is your project, you would assume that the failure times follow a distribution of a Weibull(shape = 2.298, scale=15.46). You could use the probability plot to estimate the percent failing above any value by adding percentile lines to the chart.

If you are still following the concept, we will consider the more general case of “Arbitrarily Censored”. This method includes right-censored, left-censored, and interval-censored data. Interval censored means that we know the event occurred between two values or within an interval, but we do not know the exact event time. To use this function, we generally need to create two columns of data. The format for these columns is best found in the Minitab help screen, data option for the arbitrary censored analysis. It is shown below.

**Distribution Analysis (Arbitrarily Censored Data)**

When your data consist of exact failures and a varied censoring scheme, including right-, left- and interval-censored data, your data is arbitrarily-censored. For general information on life data and censoring, see [Distribution Analysis Data](#).

You can enter up to 50 samples per analysis. Minitab estimates the functions independently for each sample, unless you assume a common shape ([Weibull](#)) or scale (other distributions). All the samples display on a single plot, with different colors and symbols, which helps you compare the various functions between samples.

Minitab analyzes systems with one cause of failure or multiple causes of failure. For systems that have more than one cause of failure, see [Multiple Failure Modes \(Arbitrarily Censored Data\)](#).

Enter your data in table form, using a Start column and End column:

For this observation...	Enter in the Start Column...	Enter in the End Column...
Exact failure time	Failure time	Failure time
Right censored	Time that the failure occurred after	Missing value symbol "*"
Left censored	Missing value symbol "*"	Time before which the failure occurred
Interval censored	Time at start of interval during which the failure occurred	Time at end of interval during which the failure occurred

This data set illustrates tabled data. For observations with corresponding columns of frequency, see [Using frequency columns](#).

Start	End	
*	10000	Left censored at 10000 hours.
10000	20000	
20000	30000	
30000	30000	Exact failures at 30000 hours.
40000	50000	
50000	50000	
50000	60000	Interval censored between 50000 and 60000 hours.
60000	70000	
70000	80000	
80000	90000	
90000	*	Right censored at 90000 hours.

When you have more than one sample, you can use separate columns for each sample. Alternatively, you can stack all the samples in one column, then set up a column of grouping indicators, which can be numbers or text. For an illustration, see [Stacked vs. Unstacked data](#).

There are four different methods to enter a single observation.

Left censored: start = \*, end = time where failure noticed

Exact failure time known: Start = End = time of failure.

Right censored data: start = time of operation without failure, End = \*

Interval censored: start = last time it was known to work, End: time it was known to be failed.

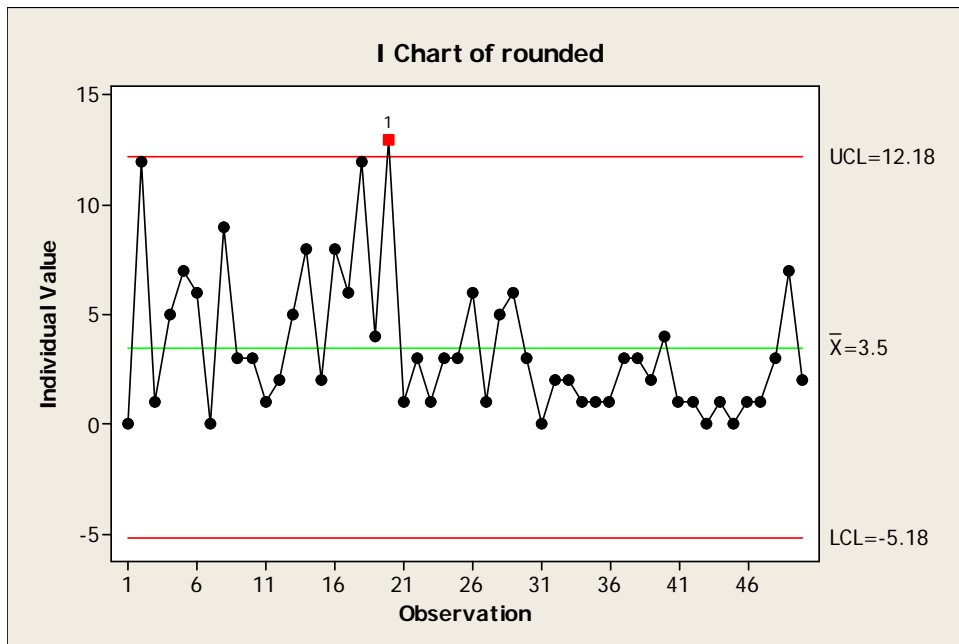
Consider the last option, interval censored. This is the option we will use to our advantage in some improvement projects.

## Rounded Data Analysis

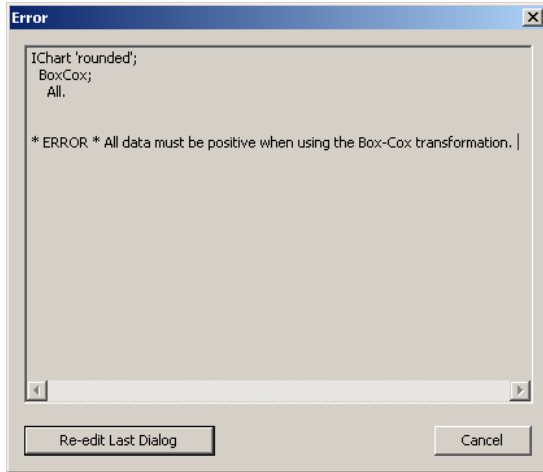
Consider you have a project to reduce the time it takes to process a transaction. It typically takes between 1 to 5 days. When you are provided the data, it looks like this;

0	1	1	0	1
12	2	3	2	1
1	5	1	2	0
5	8	3	1	1
7	2	3	1	0
6	8	6	1	1
0	6	1	3	1
9	12	5	3	3
3	4	6	2	7
3	13	3	4	2

So you check for predictability with an I-chart.



Well, the chart has an out-of-control point and an apparent natural limit at zero. Since it is time data, values less than zero are not possible. This appears to be non-normal data, so that you check it with a probability plot. You then try the Box-Cox transform option on the I-chart and receive the following message;



Now when you examine the data, there are zeros included in the data. What does zero time mean? When

you ask the process owner, you find that the computer that tracks the elapsed time subtracts the start from the end day to calculate time, so that a zero means it was completed on the same day it started. Since there are zeros in the data, you cannot use the Box-Cox transformation function or the

distribution analysis functions to solve for the most common distributions for time data, lognormal and Weibull. OK, what do you do now? Use arbitrary censored data analysis!

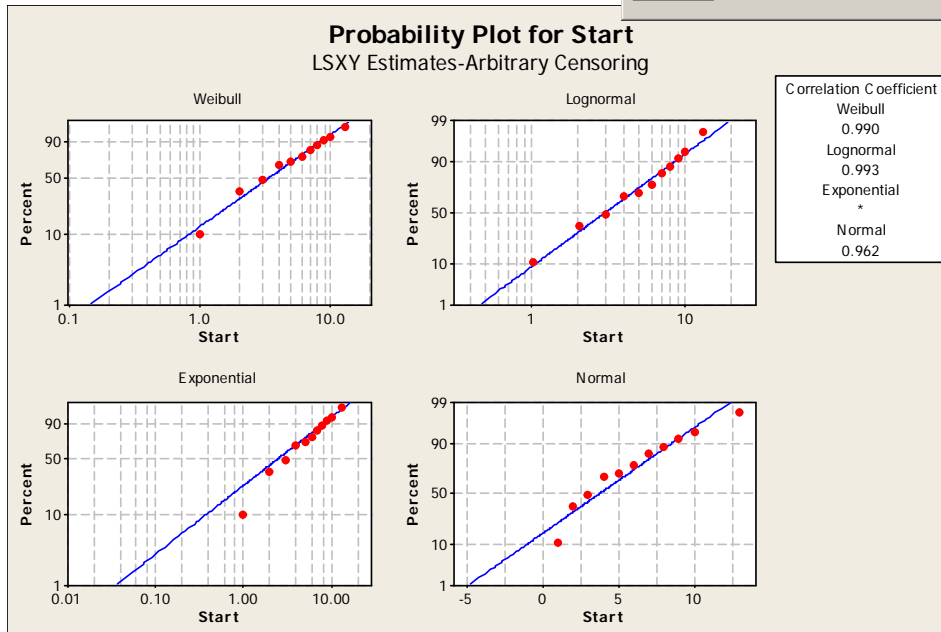
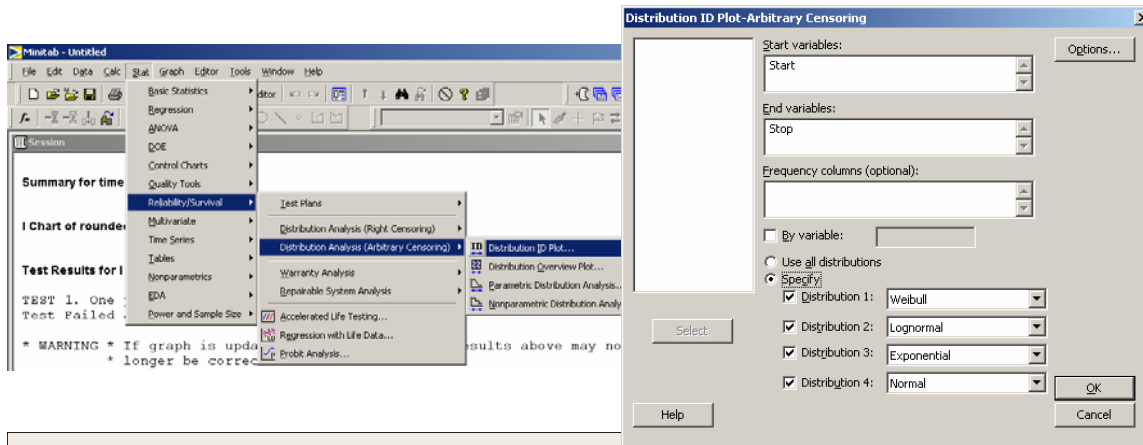
What do we know about the data is that the recorded value is the start and stop date subtracted, so that a zero means it took between 0 and 1 day. A 1 means it took between 1 and 2 days. Now we need to create the two columns for the data analysis, because we do not know the true completion time, but we know the interval that it fell within.

For the start column we will put the reported time. For the stop column we will put the reported time +1.

	C1	C2	C3	C4	C5
	time	rounded		Start	Stop
1	0.6514	0		0	1
2	12.9043	12		12	13
3	1.7144	1		1	2
4	5.9731	5		5	6
5	7.1447	7		7	8
6	6.2815	6		6	7
7	0.4183	0		0	1
8	9.0882	9		9	10
9	3.8928	3		3	4
10	3.1981	3		3	4
11	1.2965	1		1	2
12	2.9394	2		2	3
13	5.7617	5		5	6
14	8.4090	8		8	9
15	2.8089	2		2	3
16	8.3162	8		8	9
17	6.7441	6		6	7
18	12.3418	12		12	13

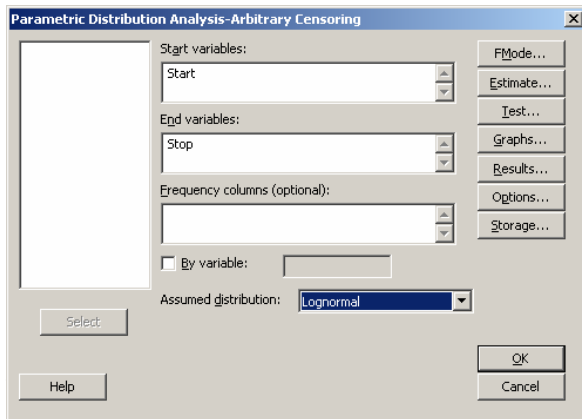
$Start = rounded$
$End = rounded + 1$

Now solve for the distribution of the data.

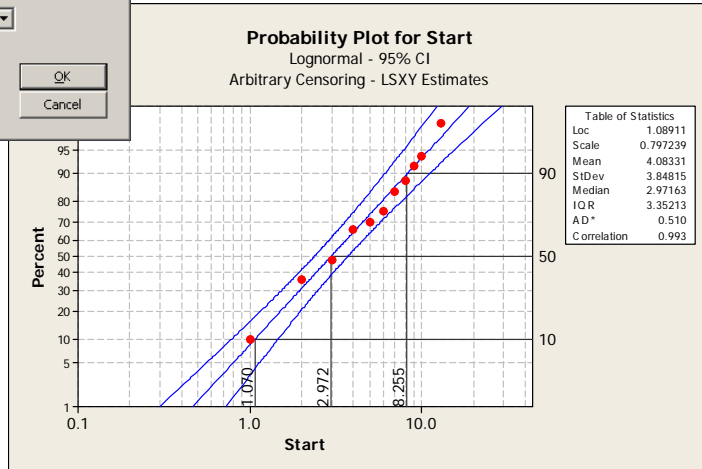


From this analysis, it looks like the lognormal is the best fit, which feels good since it is the most common distribution of cycle time data. Upon inspection, there are only 11 points on the probability plot, although there are 50 points in the data. This is because this function places only one point for each repeated value in the data. There were only 11 unique numbers with the set of 50 transactions. The software accounts for the value and the number of values occurring for each one when it estimates the distribution.

Next we will use the parametric distribution analysis function

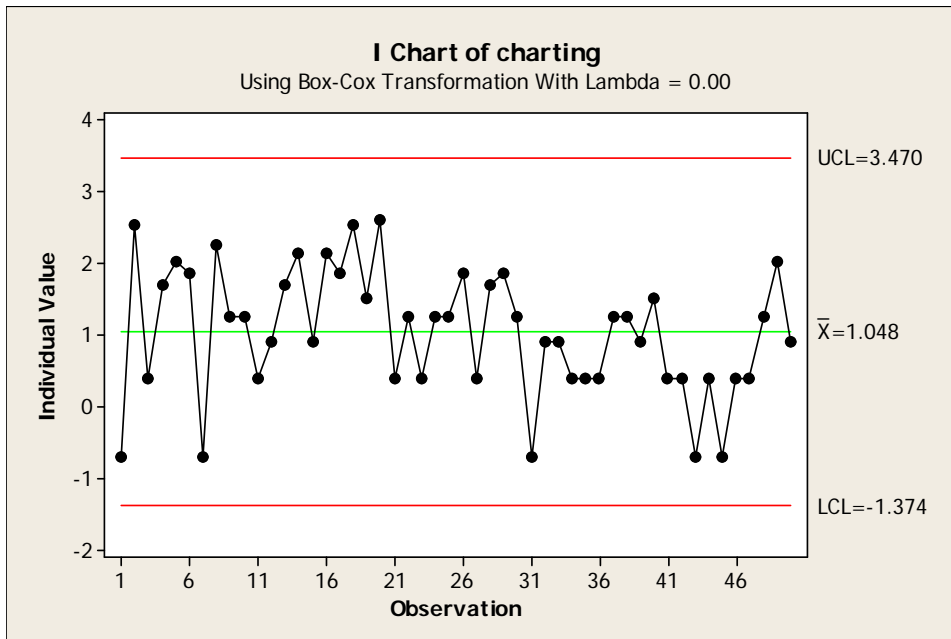


distribution has a location parameter of around 1.089 and scale of around 0.797. The mean to complete a transaction is estimated at 4.1 days, with most (80%) of the transactions completed between 1 and 8.3 days. Now we have a baseline capability to begin our improvement.



time  
have  
able

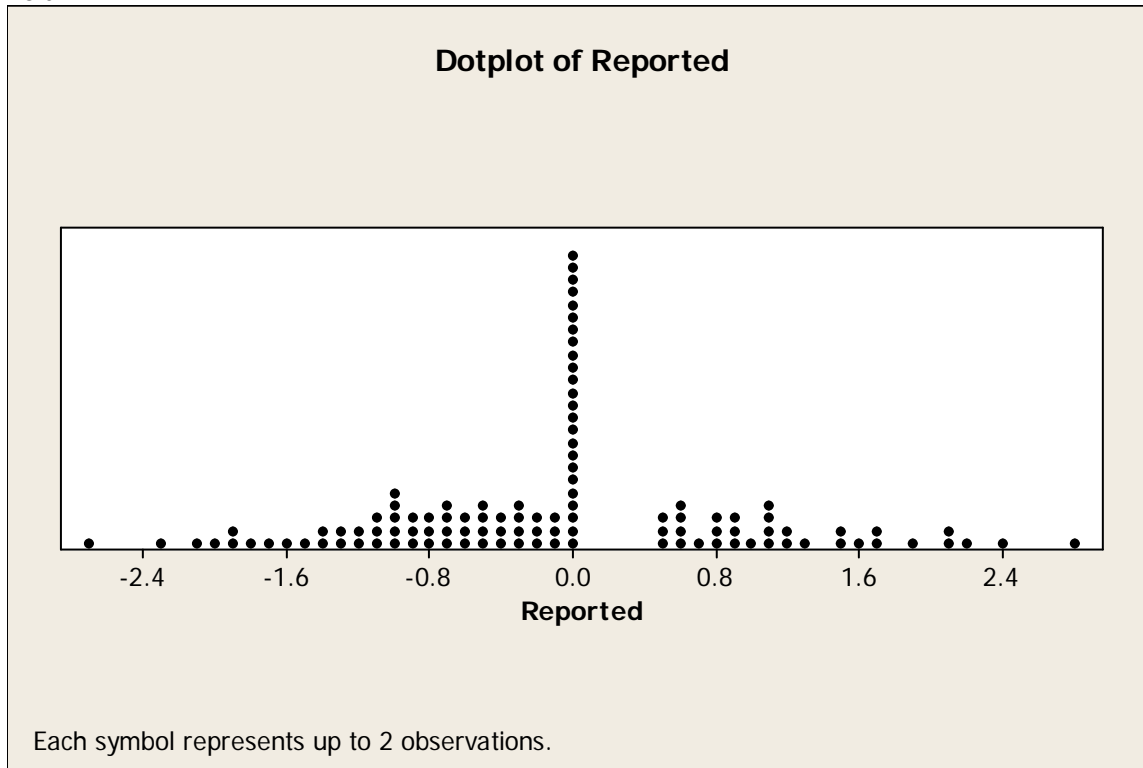
Sounds good, but we have not been to assess the predictability. This is a much harder task since we do not have any fancy tools to work with the data. There are a few options, but the simplest is to recognize that the data is really in an interval of 1.0 width. We will add ½ the interval width to each data point and plot an I-chart. Again we should use the Box-Cox option to choose the best lambda for the data transformation.



Now the process looks predictable with a log transformation (lambda=0).

Using an understanding that the recorded time data was rounded down to an integer day did not stop us from making a supportable assessment of the process predictability and capability/performance.

One other case study that came from a student is applicable to the censored data analysis. In this case, the measurement was of a component fit into a recessed area. The component was required to be centered to a specific tolerance and the belt wanted to assess the process capability to meet the process requirement. The measurements were taken with a “feeler gauge” measuring the space between the target and the component. A value of zero meant that it was centered, negative values indicate it is low, positive values indicate it is high. When he received the data, it looked wrong. Here is what shape it held.

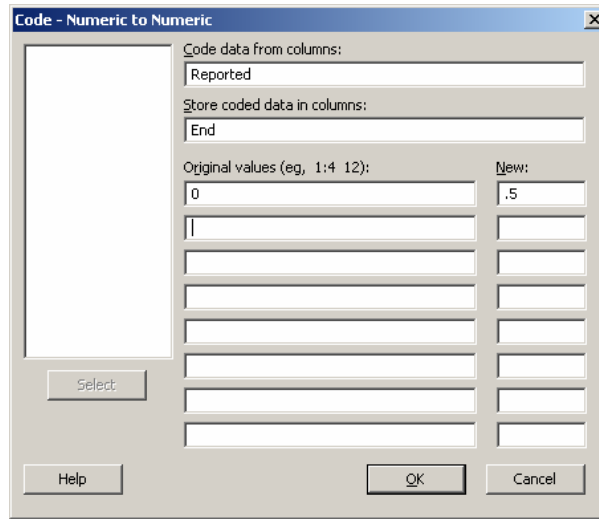
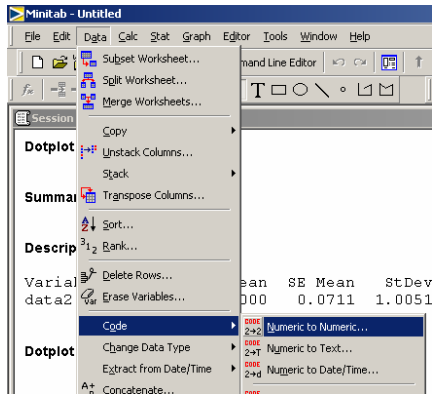


There were no points between 0 and 0.5. When he questioned the inspectors, they said that is what they measured. Without understanding why this could happen, they went back to re-measure the parts. What they found is that there was another component that interfered with the measurements as they went positive. When alignment value reached a bit higher than +0.5 there was a clearance around the other component and the inspector could get the measurement tool to record a value. What the student learned was that components with a value between 0 and 0.5 were all recorded as a zero. OK, what can be done with the existing data? Did he have to toss the data out and start over again with an improved measurement system?

No. Upon realizing that all of the zero values really are values in an interval between zero and 0.5, he could use the arbitrary censoring, parametric distribution analysis to estimate the mean and standard deviation of the process, and its true capability.

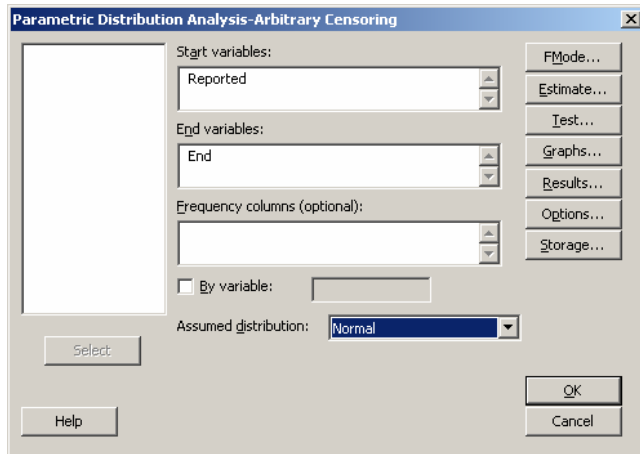
To do this we need to create a second column of data. We know that all the values are exact except for the zero values. Therefore, the second column will have the same data except that all of the zeros will be replaced with a 0.5.

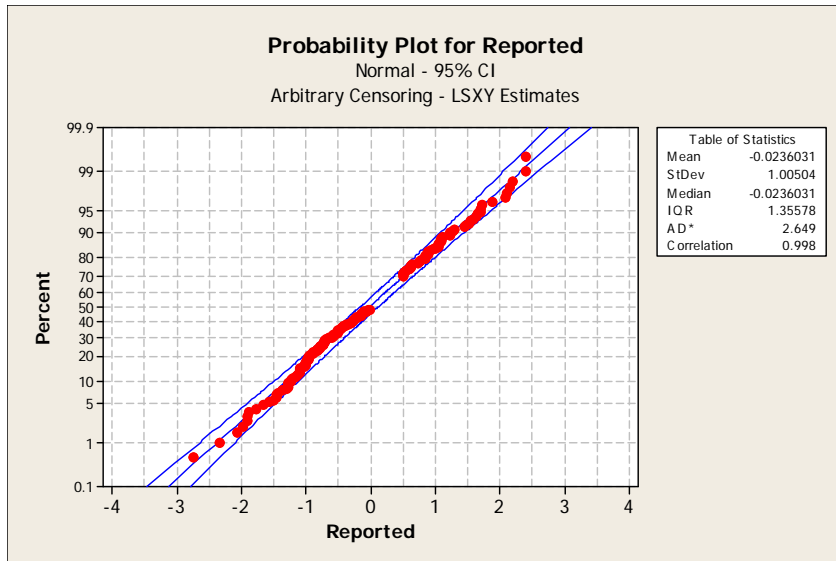
We can easily do this with the code function;



	C1	C2	C3	C4
	Data	Censored	Reported	End
1	-0.99309	1	-0.99309	-0.99309
2	0.37973	0	0.00000	0.50000
3	1.12714	1	1.12714	1.12714
4	-0.31245	1	-0.31245	-0.31245
5	1.05373	1	1.05373	1.05373
6	0.62411	1	0.62411	0.62411
7	-1.46314	1	-1.46314	-1.46314
8	-0.26408	1	-0.26408	-0.26408
9	-0.72233	1	-0.72233	-0.72233
10	0.62988	1	0.62988	0.62988
11	-0.59435	1	-0.59435	-0.59435
12	-0.24482	1	-0.24482	-0.24482
13	-0.47059	1	-0.47059	-0.47059

Now we do the arbitrary censored analysis (Stat>Reliability/Survival>Distribution Analysis (arbitrary Censored)>Parametric Distribution analysis) We could go right to this because we knew the data would fit a normal distribution, both in theory and by looking at the data.





As we saw in the previous use of this tool, the censored data points do not show on the plot as a dot, but Minitab accounts for the data probability on the chart. In this case, it is the space in the probability plot data.

This method estimated a mean of -0.024 and a standard deviation of 1. This is quite close to the uncensored data set which had a mean of zero and standard deviation of 1. Now what were the reported data statistics before we accounted for the censoring of values between 0 and 0.5, the process has a mean of -0.56 and a standard deviation of 0.99.

To sum up: Covered in this article was the use of the arbitrary censoring analysis in Minitab to process “Bad” data, i.e., when we know that the data is either severely rounded or that there was a known error/bias in reporting a portion of the data. This tool can get a Belt through the measure phase steps using supportable Minitab tools rather than just giving up and estimating the baseline performance. It does not provide a perfect answer, but it meets the needs of most improvement projects.