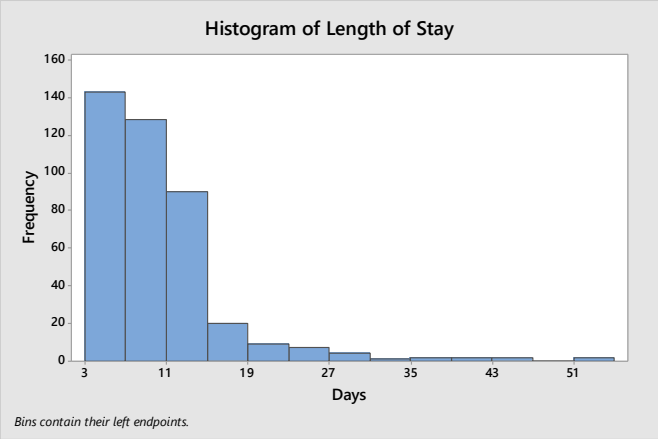
Minitab®

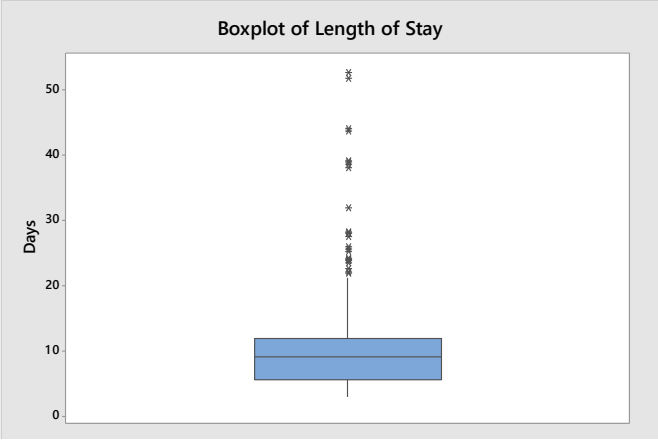
SPC

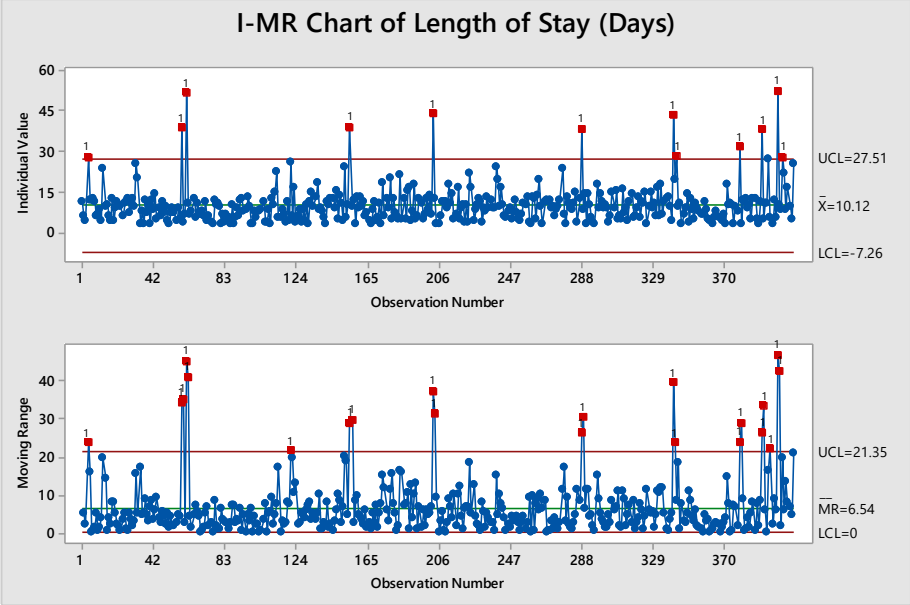
# LESSON: Transforming Non-Normal Data

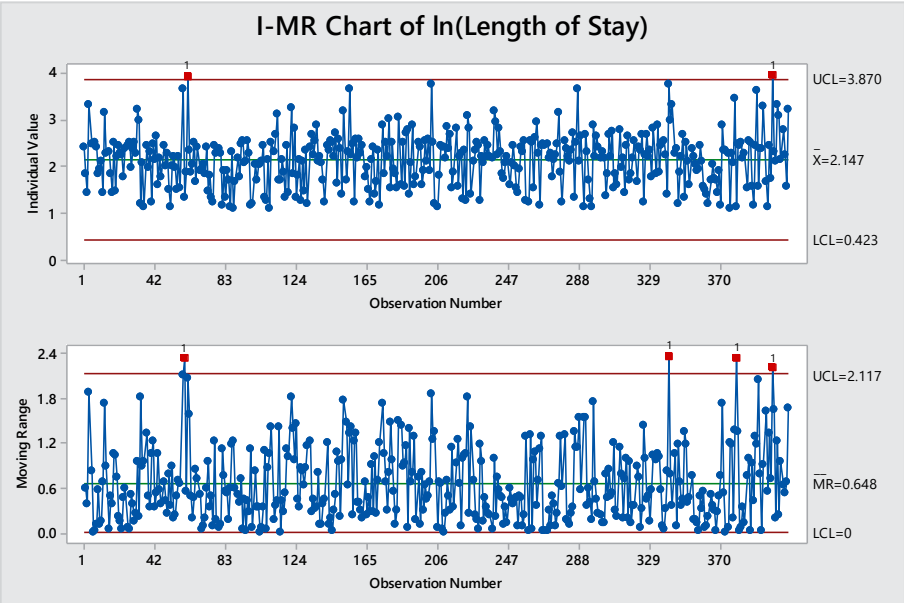
This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

**Example 1:** Length of stay when visiting family over the summer break (in days)






Non-normal data and the Box-Cox Transformation
Many processes do not follow the normal distribution. Some examples of non-normal distributions would include:

- Cycle time data: amount of time to complete a process
- Time between calls to a switchboard
- Customer waiting time
- Time between shots on goal at a hockey game

**Non-normality** is a way of life; no characteristic will have exactly a normal distribution.
One strategy to make non-normal data resemble normal data is by using a **transformation**.
Which transformation to select for the situation at hand? The choice is usually not obvious. The most popular 3 transformations are $X^2$, *ln(X), and* $\sqrt{X}$

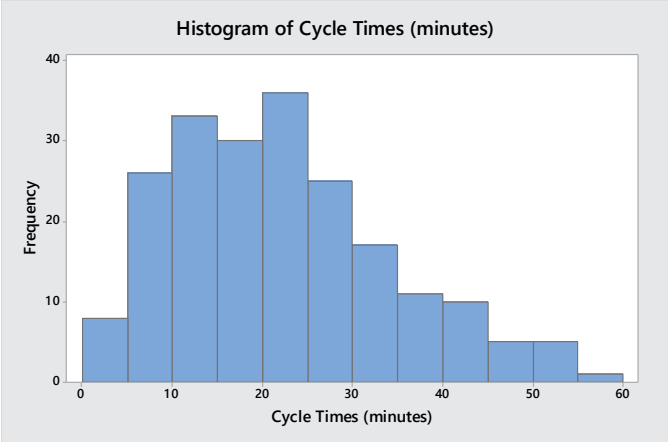
In Minitab, the **Box-Cox Transformation** estimates lambda values which minimize the standard deviation of a standardized transformed variable. The resulting transformation is $Y^{\lambda}$ when $\lambda \neq 0$ and *ln(Y)* when $\lambda = 0$.

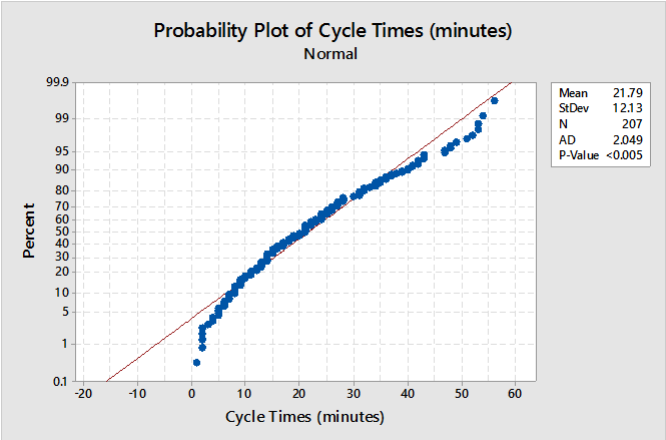
The Minitab method searches for a value of lambda from −5 to 5 that makes the transformed data "most" normal.

Here are some common transformations where $Y^*$ is the transform of the data $Y$:

| Lambda (λ) value | Transformation |
|---|---|
| $\lambda = 2$ | $Y^* = Y^2$ |
| $\lambda = 0.5$ | $Y^* = \sqrt{Y}$ |
| $\lambda = 0$ (it doesn't make sense why this is log, it's just Minitab's notation) | $Y^* = \ln(Y)$ |
| $\lambda = -0.5$ | $Y^* = 1 / (\sqrt{Y})$ |
| $\lambda = -1$ | $Y^* = 1 / Y$ |

**Example 1:** Cycle time data (in minutes)



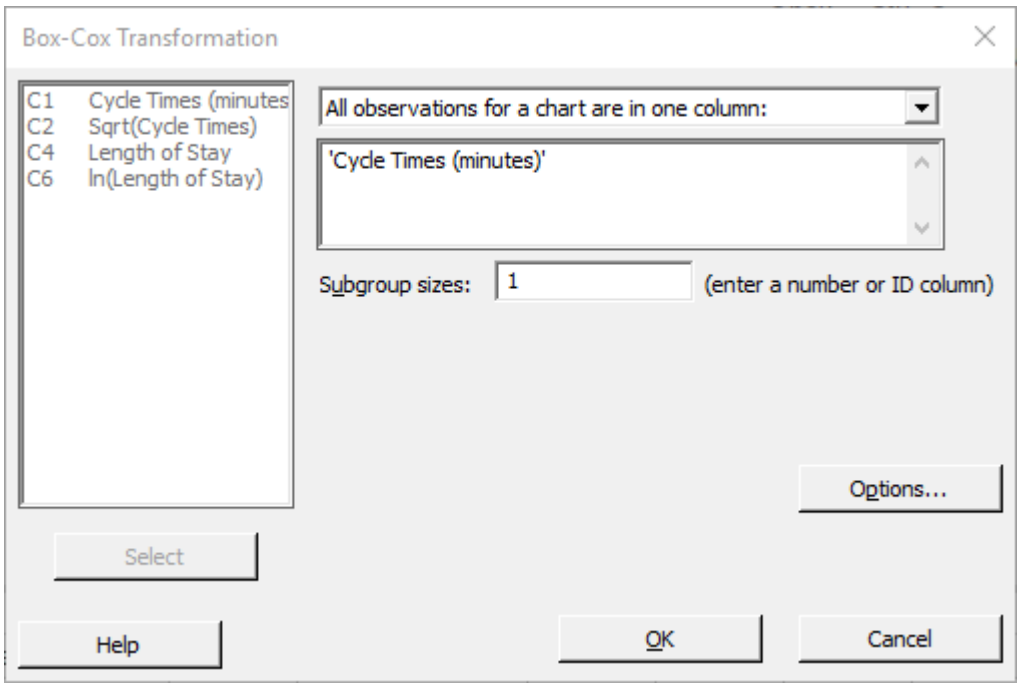




The fact that the process data is **bounded below by zero** is an important point to consider when you decide to model with a normal distribution.

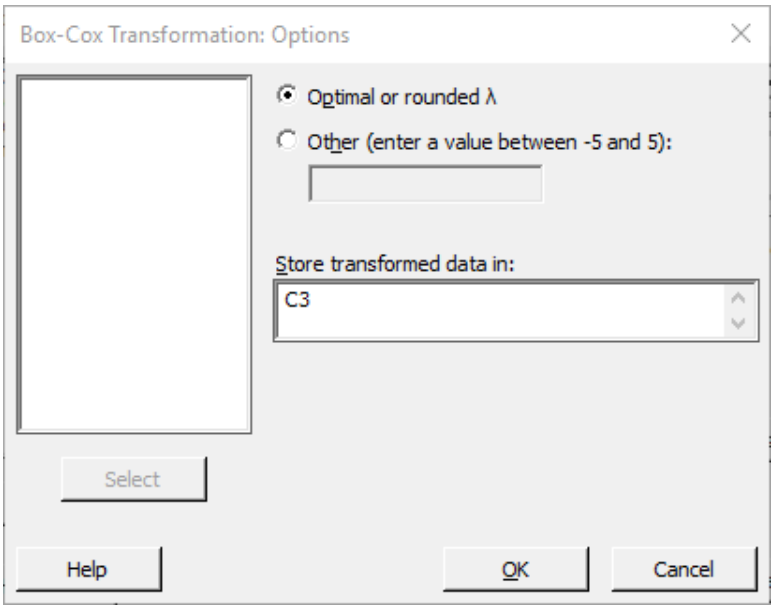
The **normal probability** plot above indicates that the data is not from a normal distribution.

Box-Cox Transformation procedure in Minitab:
Performs a Box-Cox procedure for process data used in control charts. To use Box-Cox procedure, *the data must be positive*.

Minitab:
**Choose Stat > Control Charts > Box-Cox Transformation.**
Complete the dialog box as shown below.

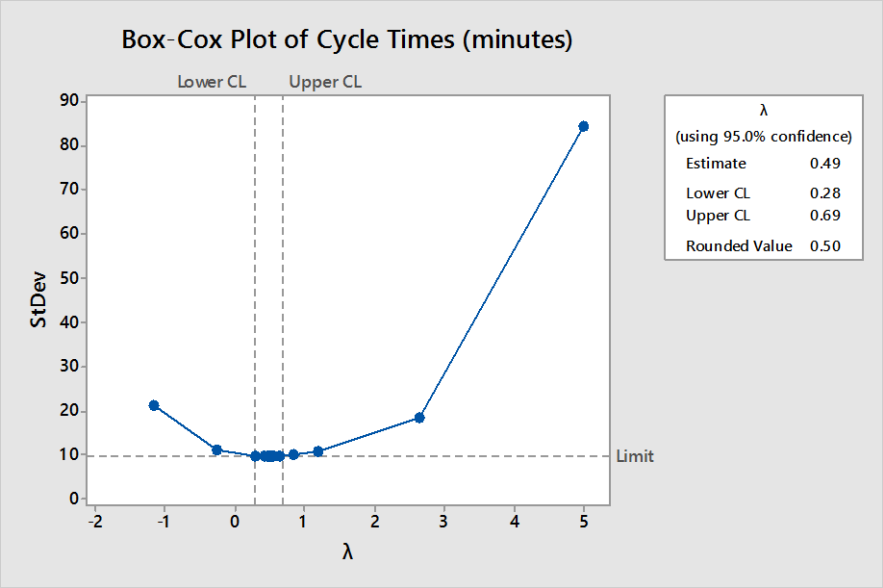



Click **Options**.




Click **OK** in each dialog box.

Minitab displays the following and indicates:

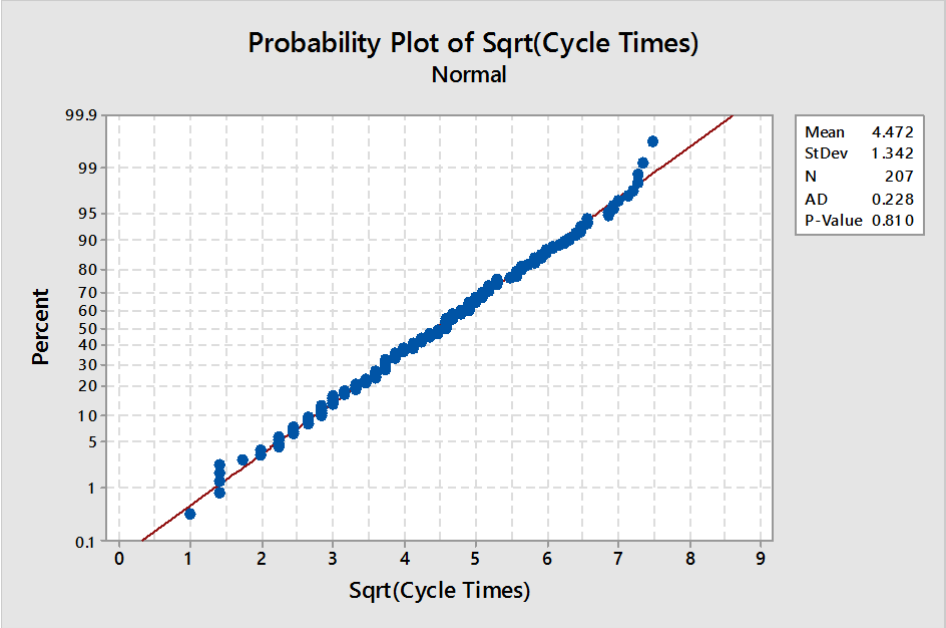



- The **best estimate of λ** for the transformation (above – it's λ = 0.49)
- A **95% confidence interval** for **λ** (above – it's [0.28, 0.69])
- A rounded value for λ (above – it's λ = 0.5)

The **graph** can be used to assess the **appropriateness of the transformation**.

- If the **95% confidence interval** for λ is "close" to **1**, this indicates that a transformation should not be done. In the case that the optimal λ is close to 1, you would gain very little by performing the transformation.
- If the optimal λ is "close" to 0.5, you could simply take the square root of the data, since this transformation is simple and understandable as compared to λ = 0.49; management can understand $\sqrt{Y}$, but maybe not $Y^{0.49}$.

Below is a normality plot of the **transformed** cycle times. The applied transformation is $\sqrt{Y}$.




**Question:** What if the **Box-Cox algorithm** does **not find a suitable transformation?**

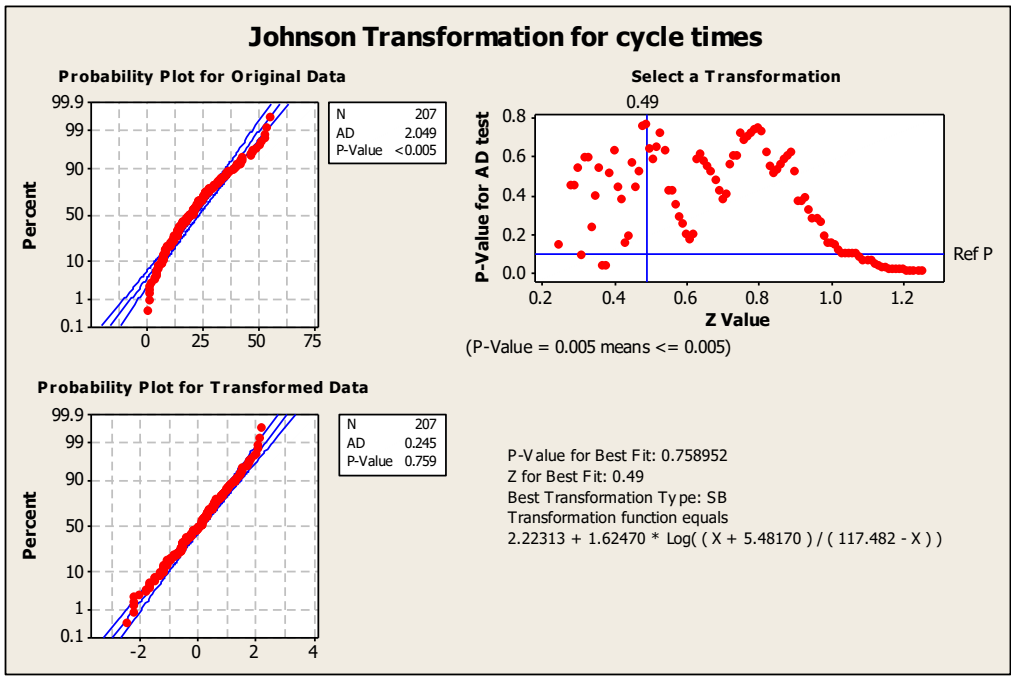
**Answer 1:** Try a **Johnson Transformation**. The Johnson Transformation uses a different algorithm than the Box-Cox Transformation. While the J**ohnson Transformation** is based on **three potential functions in the Johnson transformation family**, the **Box-Cox Transformation** simply finds a **power transformation: $Y^{\lambda}$**

How to determine the appropriate Johnson Transformation in Minitab?

**Stat > Quality Tools > Johnson Transformation**

The Johnson transformation optimally selects a function from three families of distributions of a variable, which are easily transformed into a standard normal distribution.
These distributions are labeled as *SB*, *SL*, and *SU*, where *B*, *L*, and *U* refer to the variable being bounded, lognormal, and unbounded.
Minitab displays normal probability plots for original and transformed data and their *p*-values for comparison.
You can also store the transformed data in another column for further analysis.

**Note:** A Johnson transformation does not always find an optimal function to transform your data.

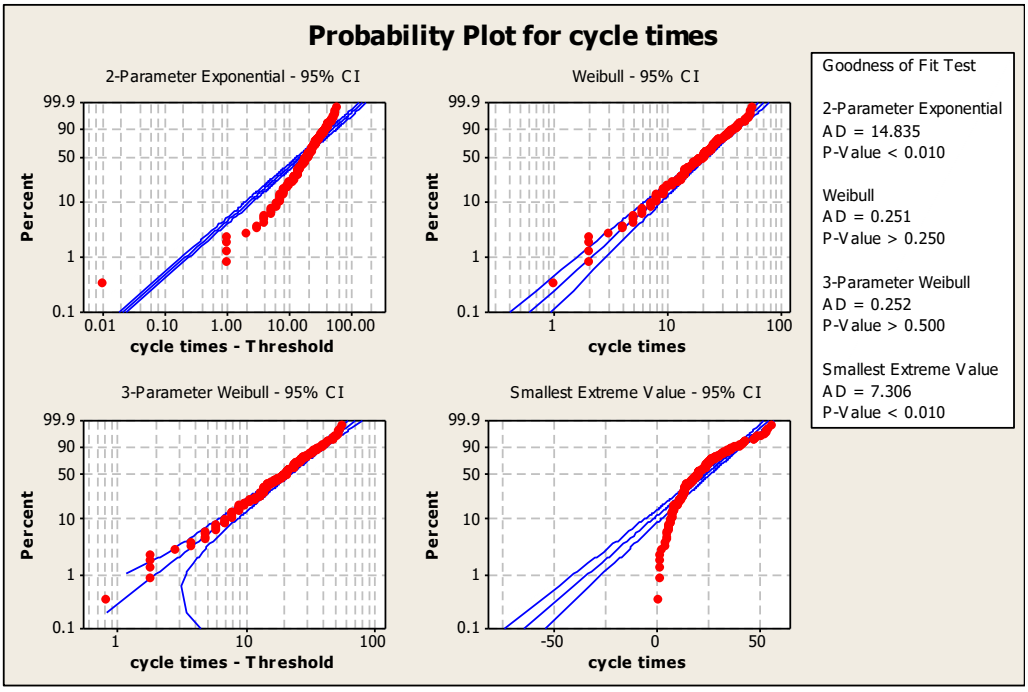



**Question:** What if the **Box-Cox algorithm** does **not find a suitable transformation?**

**Answer 2: Fit the data** with a **"reasonable" distribution** for modeling it (e.g., Exponential, Weibull, Lognormal, etc.) instead of transforming the data to normal

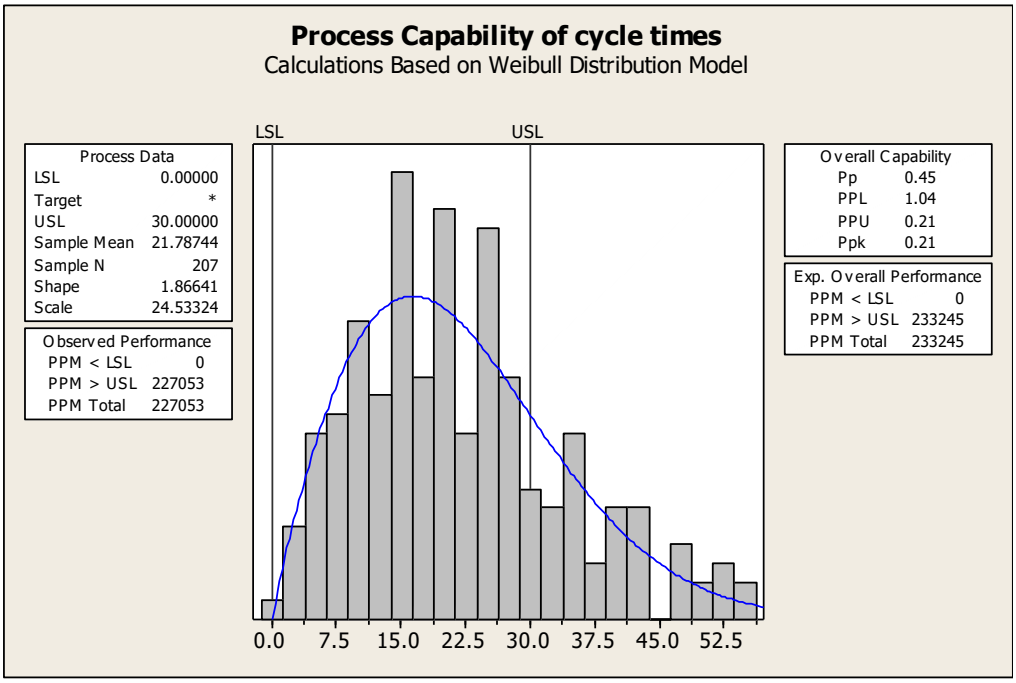
We will use this option later in the course when determining the **capability of a process**

**Stat > Quality Tools > Individual Distribution Identification;** look for "good" p-values (> 0.05 at least)



We can see that the **Weibull distribution** is the best fit for the actual (not transformed) data.

**Stat > Control Charts > Quality Tools > Capability Analysis > Non-Normal**, Choose Weibull distribution




## Determining which distribution fits a set of data

**Example:** Cycle time data from Box-Cox Lesson, Example 1.

**Stat > Quality Tools > Individual Distribution Identification;** look for "good" p-values (> 0.05 at least)

$H_0$: Data is from _____ distribution

$H_a$: Data is not from _____ distribution

**Goodness of Fit Test hypothesis test: small *p*-value - reject the fit of the given distribution**

| **Distribution** | **AD** | **P** | **LRT P** | |
|---|---|---|---|---|
| Normal | 2.049 | <0.005 | | |
| Box-Cox Transformation | 0.228 | 0.810 | | # Box-Cox is not a distribution |
| Lognormal | 2.812 | <0.005 | | |
| 3-Parameter Lognormal | 0.308 | * | 0.000 | |
| Exponential | 17.189 | <0.003 | | |
| 2-Parameter Exponential | 14.808 | <0.010 | 0.000 | |
| **Weibull** | **0.251** | **>0.250** | | |
| 3-Parameter Weibull | 0.252 | >0.500 | 0.848 | |
| Smallest Extreme Value | 7.306 | <0.010 | | |
| **Largest Extreme Value** | **0.343** | **>0.250** | | |
| **Gamma** | **0.603** | **0.133** | | |
| 3-Parameter Gamma | 0.267 | * | 0.068 | |
| Logistic | 1.568 | <0.005 | | |
| Loglogistic | 1.510 | <0.005 | | |
| 3-Parameter Loglogistic | 0.536 | * | 0.000 | |
| Johnson Transformation | 0.245 | 0.759 | | # Johnson is not a distribution |

**ML Estimates of Distribution Parameters**

| Distribution | Location | Shape | Scale | Threshold |
|---|---|---|---|---|
| Normal* | 21.78744 | | 12.13489 | |
| Box-Cox Transformation* | 4.47155 | | 1.34216 | |
| Lognormal* | 2.88932 | | 0.69322 | |
| 3-Parameter Lognormal | 3.55392 | | 0.32800 | -15.07332 |
| Exponential | | | 21.78744 | |
| 2-Parameter Exponential | | | 20.88835 | 0.89909 |
| **Weibull** | | **1.86641** | **24.53324** | |
| 3-Parameter Weibull | | 1.84652 | 24.32771 | 0.16530 |
| Smallest Extreme Value | 28.18949 | | 13.35720 | |
| **Largest Extreme Value** | **16.13630** | | **9.84622** | |
| **Gamma** | | **2.75928** | **7.89605** | |
| 3-Parameter Gamma | | 4.18274 | 6.06654 | -3.58732 |
| Logistic | 20.88098 | | 6.88111 | |
| Loglogistic | 2.95179 | | 0.37315 | |
| 3-Parameter Loglogistic | 3.43790 | | 0.21531 | -11.29200 |
| Johnson Transformation* | 0.02833 | | 0.97677 | |

* Scale: Adjusted ML estimate

## Using that distribution to determine probabilities: