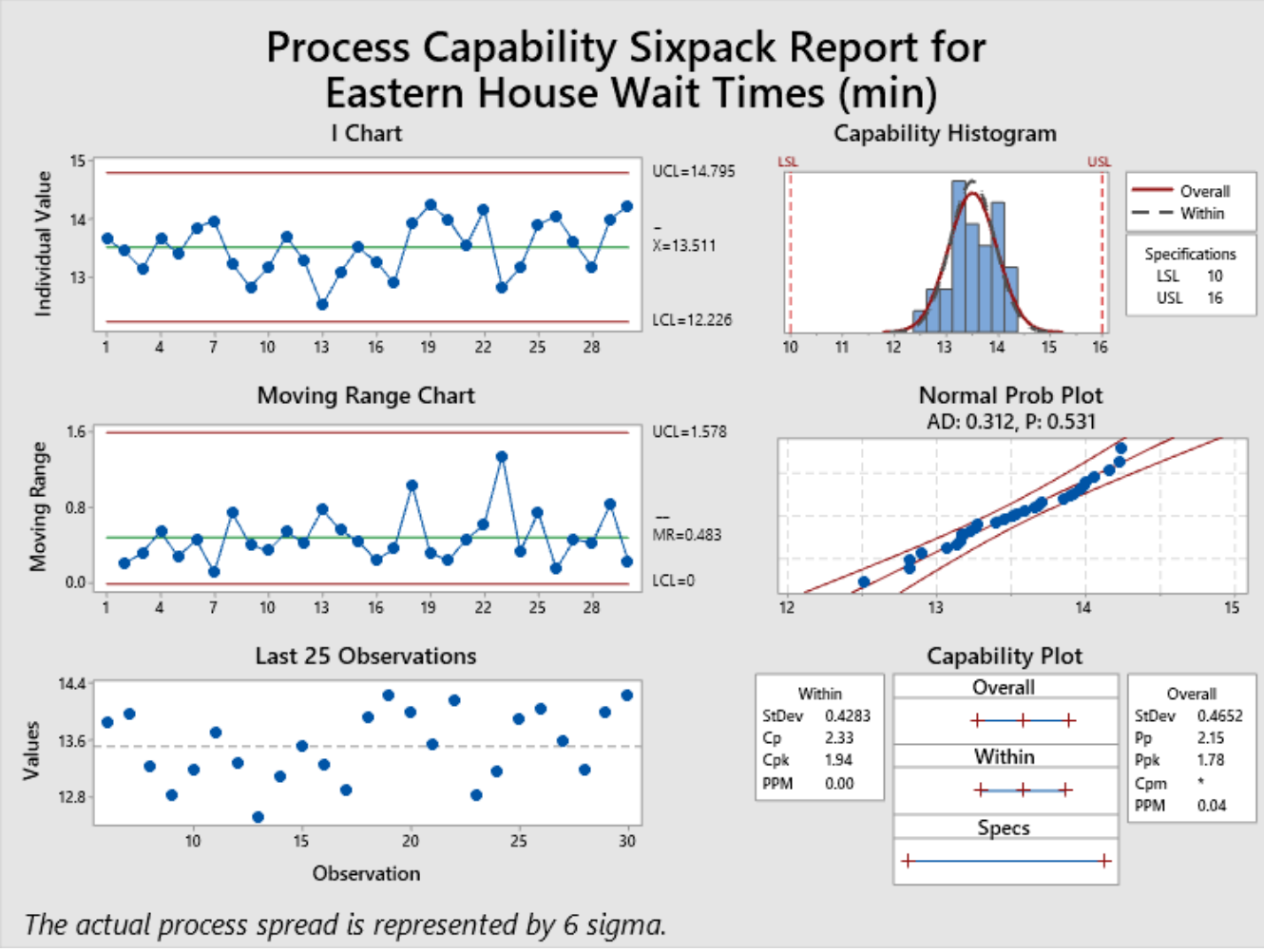
Minitab®

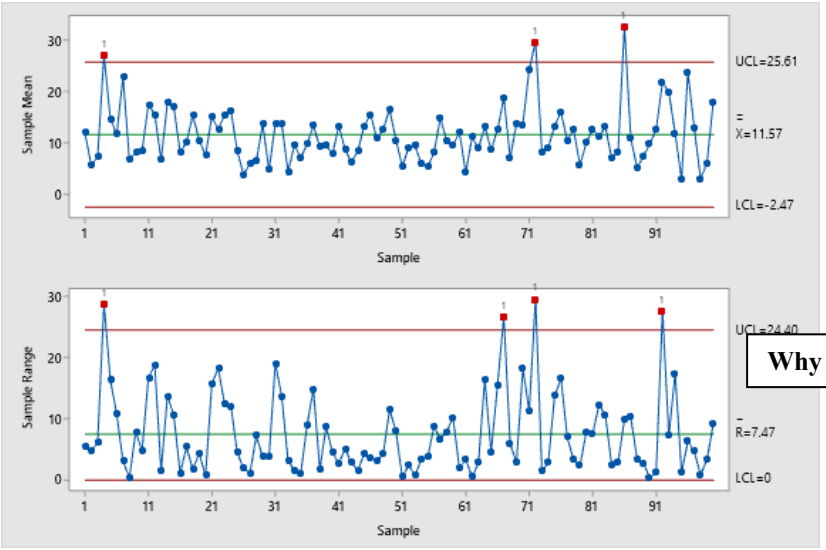
SPC

LESSON: Capability Analysis with *Non-Normal Data*

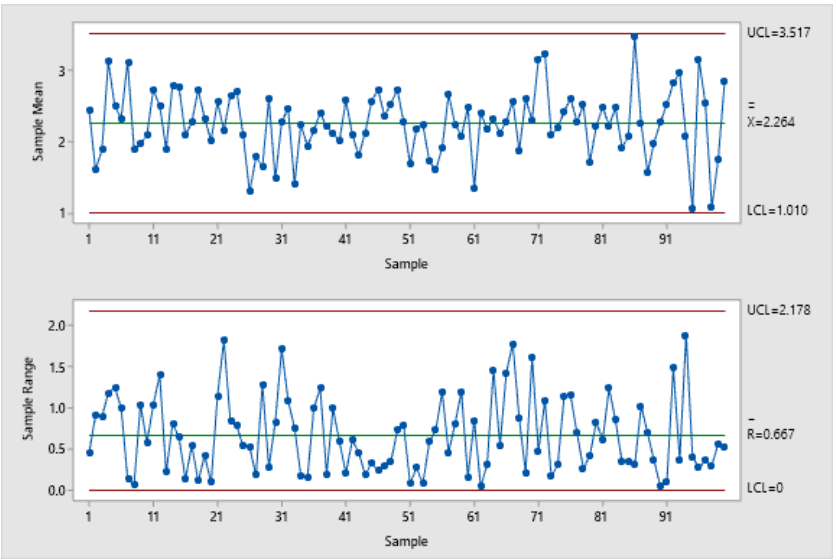
Today's Topic: Why do we need to transform non-normal data? Below are lengths of stay at a hospital for a particular procedure.



Note: $\bar{X}$ **and R charts** with non-normal data (top); $\bar{X}$* **and R* charts** with transformed (normal) data (bottom)



Why so many out of control points on R chart?



Capability Analysis ASSUMPTIONS

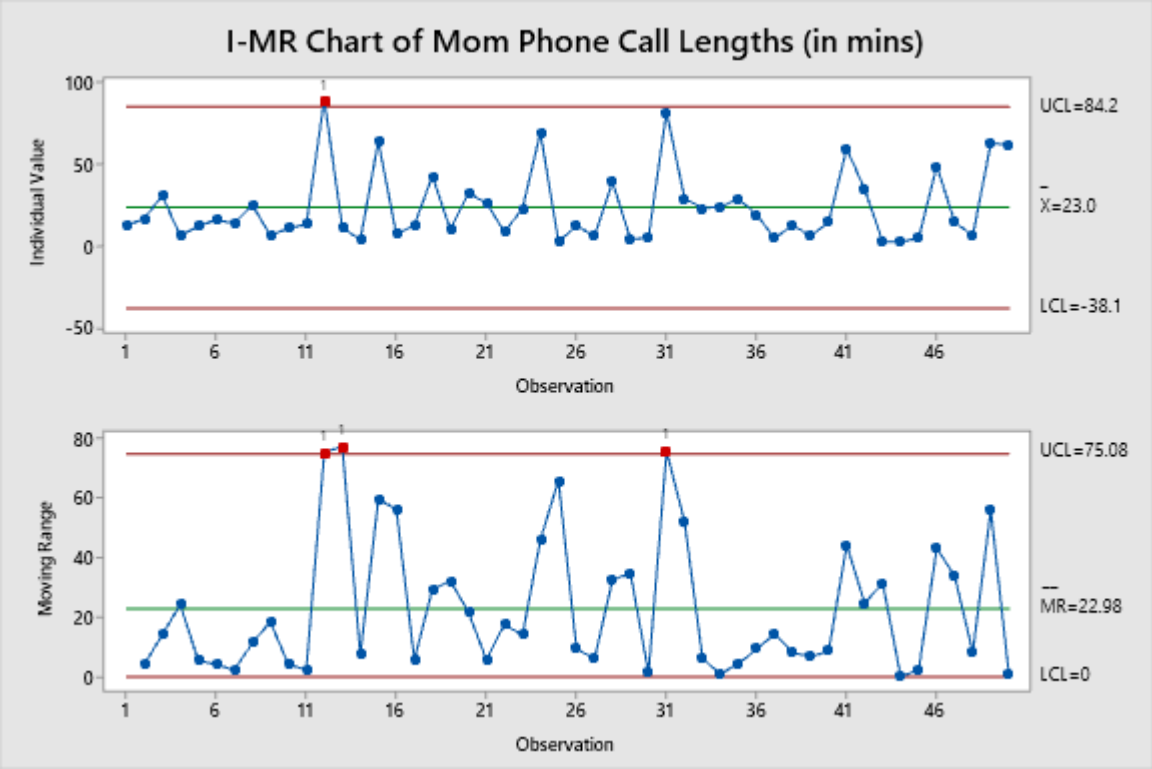
1. What must be true about our process before we can perform a capability analysis?
2. What other assumption must we check that I've clearly alluded to in all of my graphics for capability?
3. Another assumption we *should* check.

If these assumptions are not met, we can still perform a capability study, but the indices will be meaningless.

Example 1. I have been collecting the length of phone calls that I've had with my Mom for the last 50 calls. The data are in **Lesson16DATA_CapabilityAnalysis_Nonnormal**.

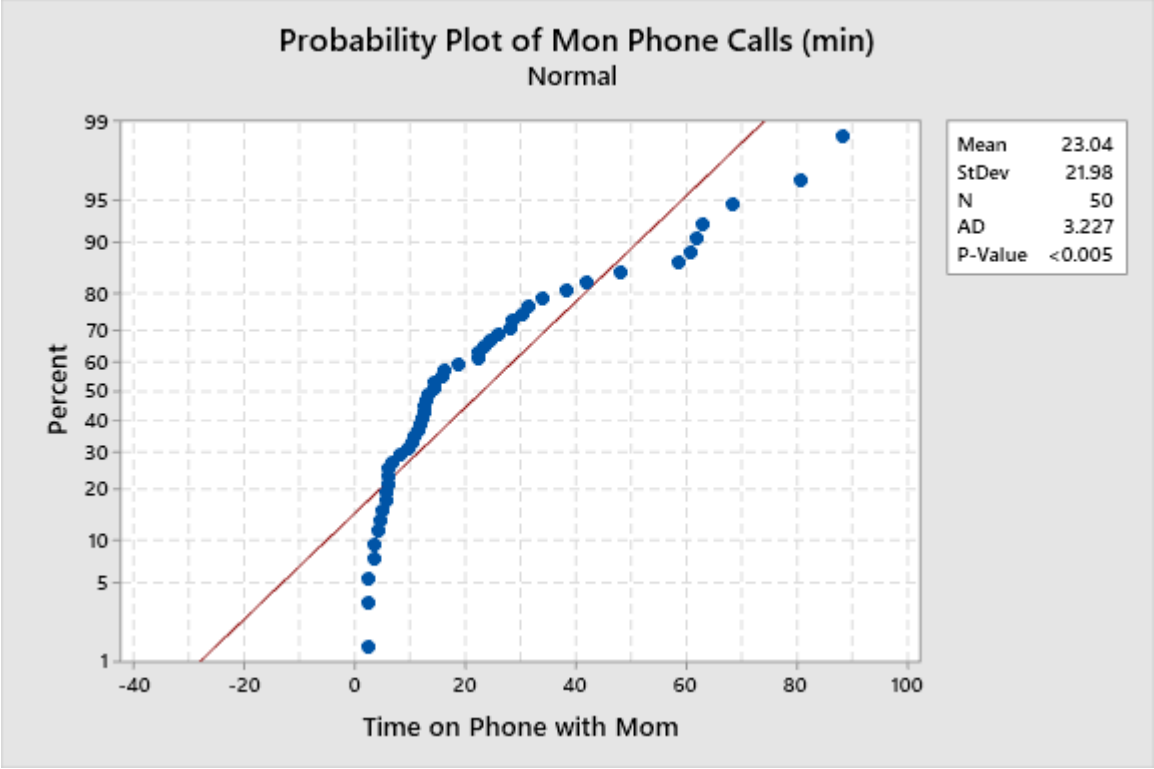
(a) First, let's make sure the length of these calls is a "stable" process.

Stat > Control Charts > Variable Charts for Individuals > I-MR



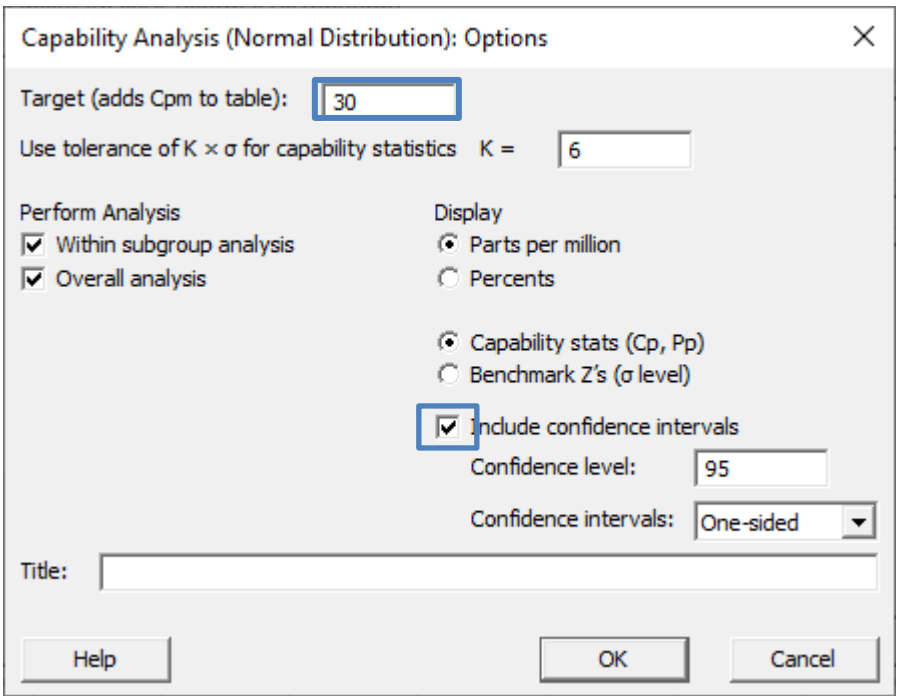
(b) Is the data normally distributed?

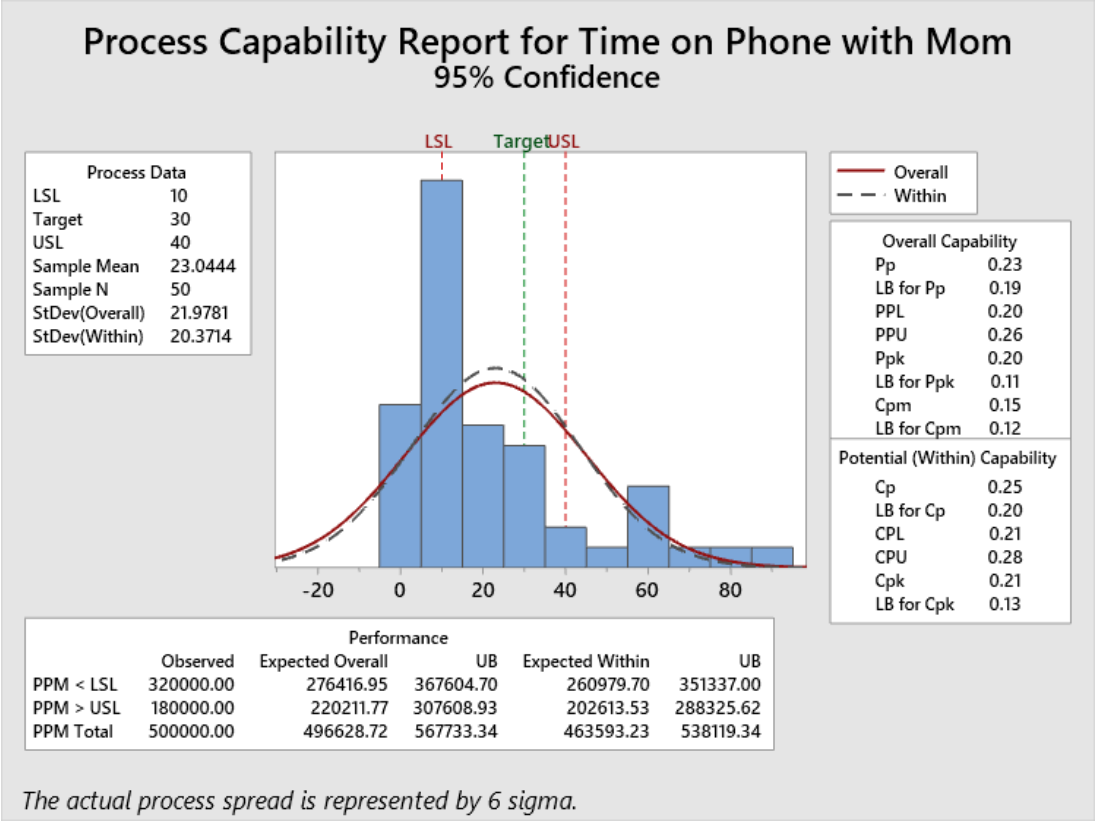
You are performing the following hypothesis test:
H_0: Data is from a normally distributed process
H_a: Data is not from a normally distributed process



Setting specification limits: I am going to establish the lower specification of 10 minutes since if we talk less than 10 minutes then my Mom may feel sad that I'm trying to get off the phone quickly. I get "antsy" being on the phone with anyone more than 40 minutes. So, I'll set the lower specification of 10 minutes and the upper specification of 40 minutes. I'd say my ideal target time is 30 minutes.

(c) Capability Analysis? Can we perform a process capability study assuming normality and a stable process? Note that I added a "target value" to the analysis (30 minutes). Click **Options**.





Non-normal data and the Box-Cox Transformation

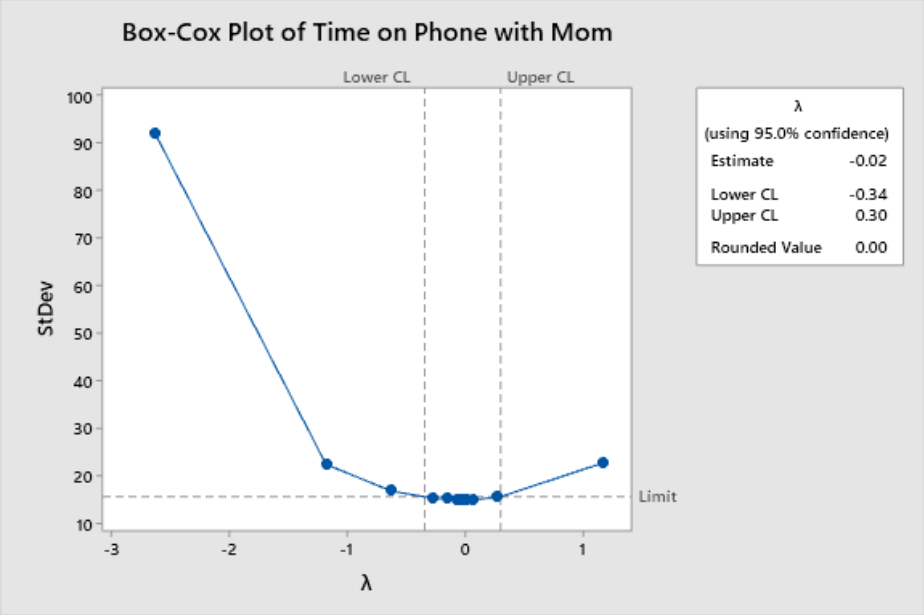
- Non-normality is a way of life; no characteristic will have exactly a normal distribution.
- One strategy to make non-normal data resemble normal data is by using a **transformation**.
- Which transformation to select for the situation at hand? The choice is usually not obvious.
- In Minitab, the **Box-Cox Transformation** estimates lambda values, as shown below, which minimize the standard deviation of a standardized transformed variable. The resulting transformation is **Y^{λ} when $\lambda \neq 0$** and **$\log_e Y$ or $\ln(Y)$ when $\lambda = 0$.**

The Minitab method searches for a value of lambda from λ = -5 to 5 that makes the transformed data "most" normal.
Here are some common transformations where *Y'* is the transform of the data *Y*:

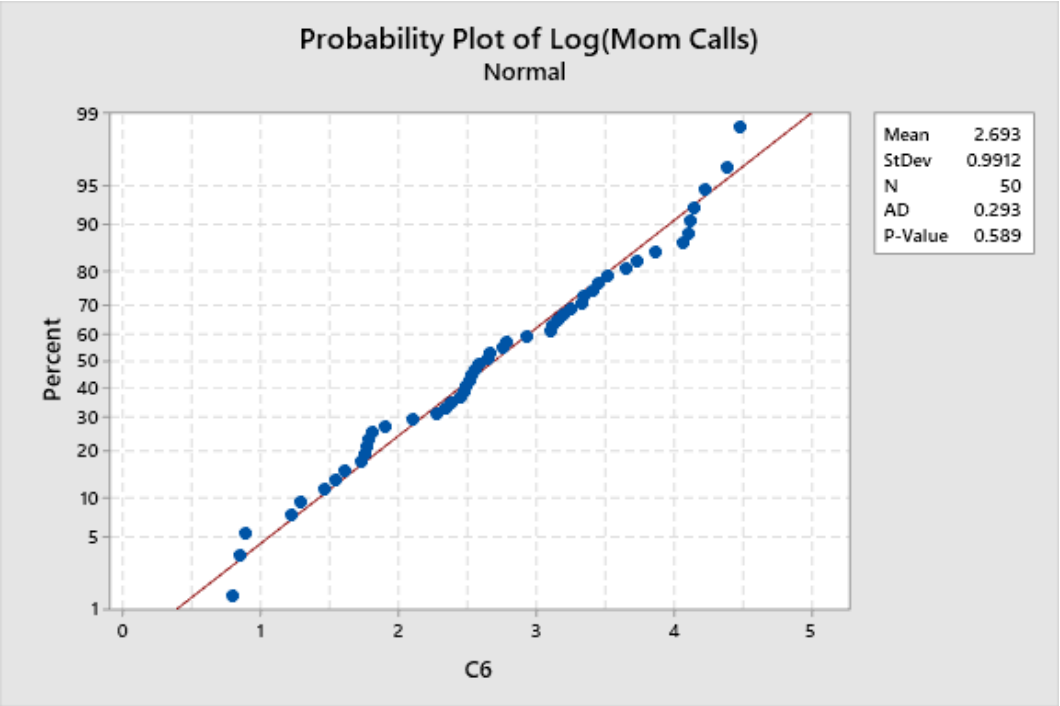
Lambda (λ) value	Transformation
λ=2	$Y' = Y^2$
λ = 0.5	$Y' = \sqrt{Y}$
λ = 0	$Y' = \log_e Y$, i.e., $\ln(Y)$
λ = -0.5	$Y' = 1 / (\sqrt{Y})$
λ = -1	$Y' = 1 / Y$

(d) What is the appropriate transformation for the "Mom Talk Length" data? What transformation (e.g., ln(*X*), X^2, $\sqrt{X}$, etc.) are you applying to the data?

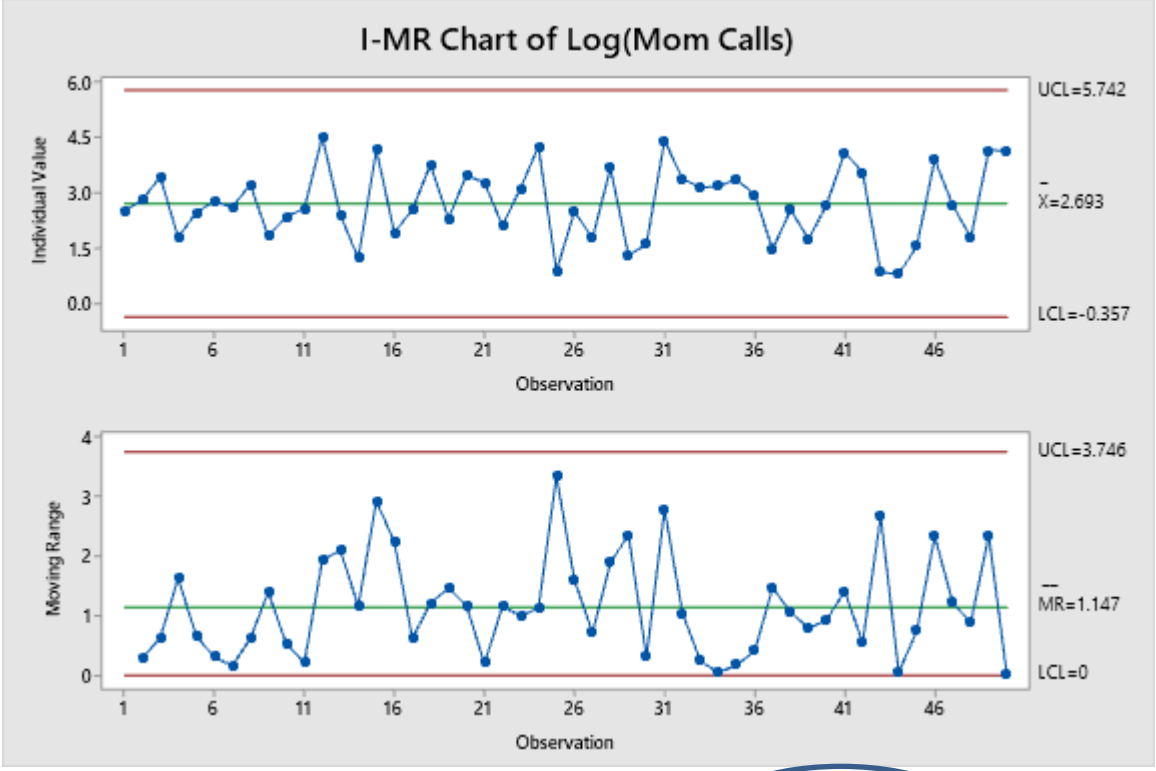
Stat > Control Charts > Box Cox Transformation



Either you can transform the data **Y' as Y' = ln(Y)** (if you are using Minitab's rounded value) or **$Y' = Y^{-0.02}$** (if you are using Minitab's estimated value). The rounded value for λ is just the more "simple" and understandable value for λ for the casual user. If the value of λ is close to 0, most users are more comfortable with the transformation ln(Y) than $Y^{-0.02}$.



(e) Capability Analysis? Can we perform a capability analysis with the transformed data assuming normality and a stable process?

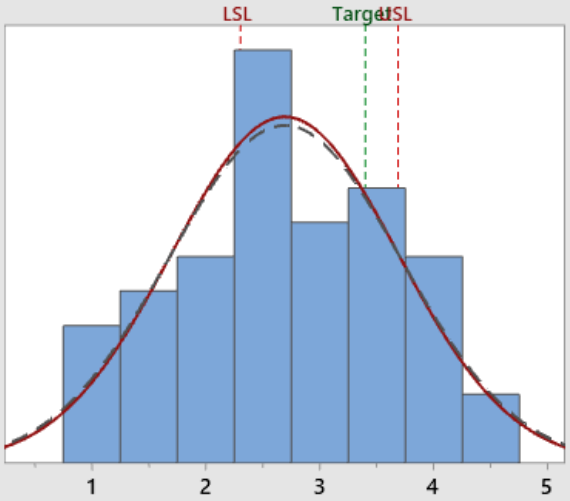


Where did I get these numbers?

Process Capability Report for Log(Mom Calls)

95% Confidence

Process Data	
LSL	2.302
Target	3.401
USL	3.689
Sample Mean	2.69263
Sample N	50
StDev(Overall)	0.99119
StDev(Within)	1.01643



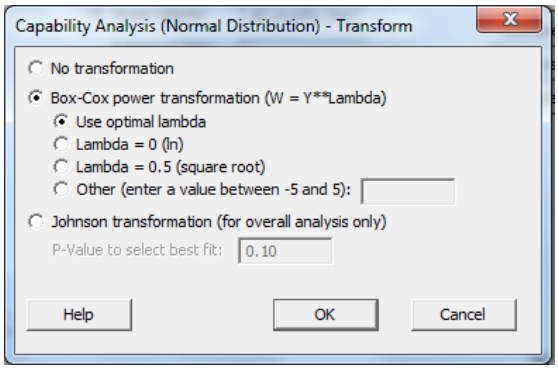
Overall
Within

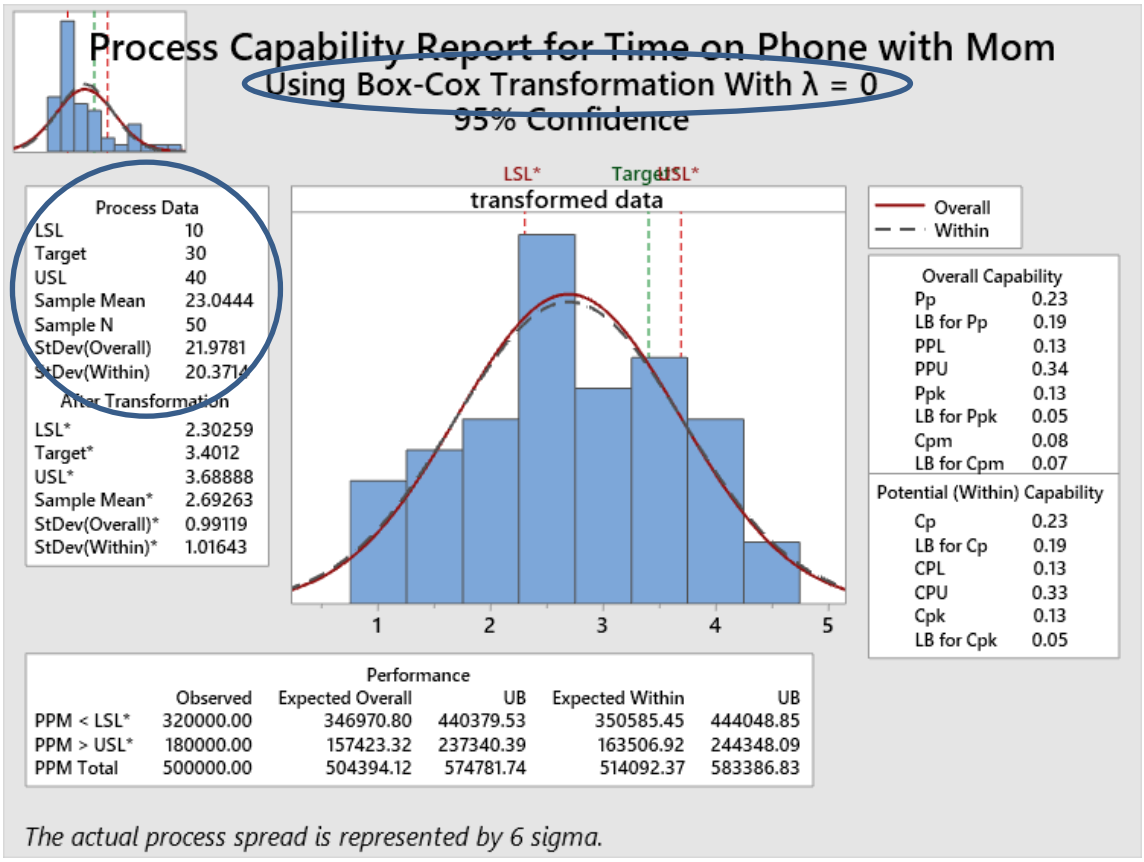
Overall Capability	
Pp	0.23
LB for Pp	0.19
PPL	0.13
PPU	0.34
Ppk	0.13
LB for Ppk	0.05
Cpm	0.08
LB for Cpm	0.07

Potential (Within) Capability	
Cp	0.23
LB for Cp	0.19
CPL	0.13
CPU	0.33
Cpk	0.13
LB for Cpk	0.05

Performance					
	Observed	Expected Overall	UB	Expected Within	UB
PPM < LSL	320000.00	346752.87	440158.14	350372.13	443832.4
PPM > USL	180000.00	157394.04	237306.54	163477.66	244314.50
PPM Total	500000.00	504146.92	574561.14	513849.79	583170.95

Easier way? **Stat > Quality Tools > Capability Analysis > Normal**, click **Transform** box.



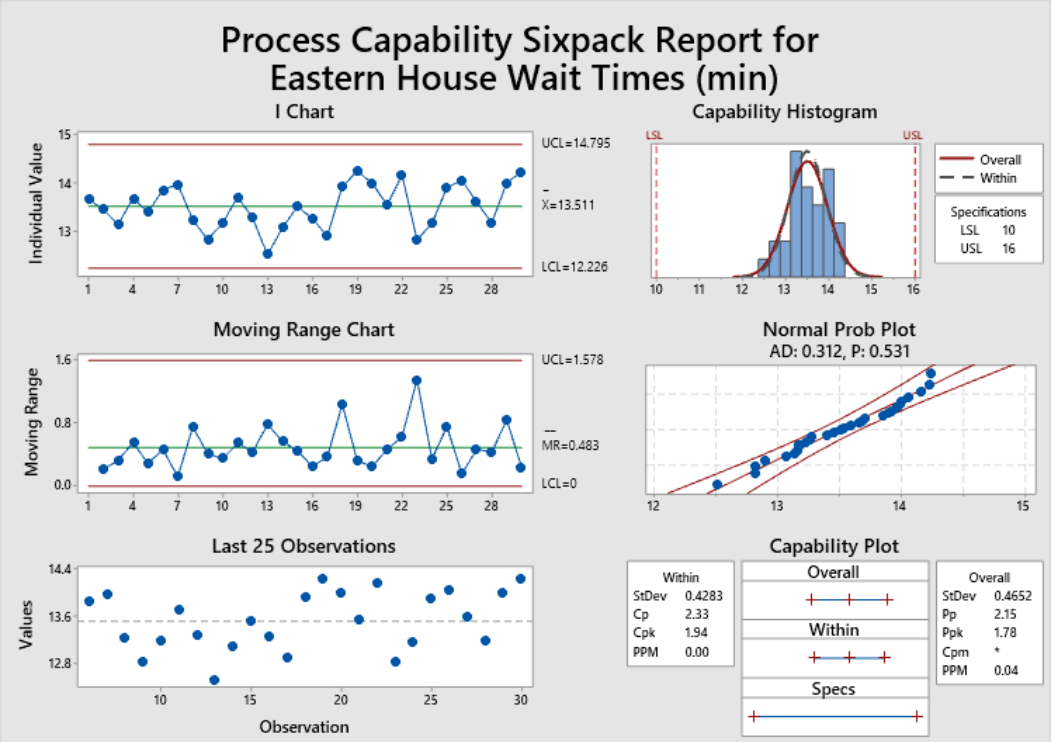


The Difference Between Cpk and Ppk

Example 2. I have been able to collect data for **waiting times** from the Eastern House restaurant for mealtimes starting at 7p.m. The manager wanted to improve customer satisfaction related to customer waiting for their meals. The **lower specification of 10 minutes** was established to increase beverage sales. The manager found if the wait time was less than 10 minutes people wouldn't order a second drink. The **upper specification was 16 minutes** because manager started to receive complaints when wait time went over 16 minutes. If the manager could decrease the variation in wait time, then she could better predict how long people would have to wait for tables and how many customers she could serve in one evening.

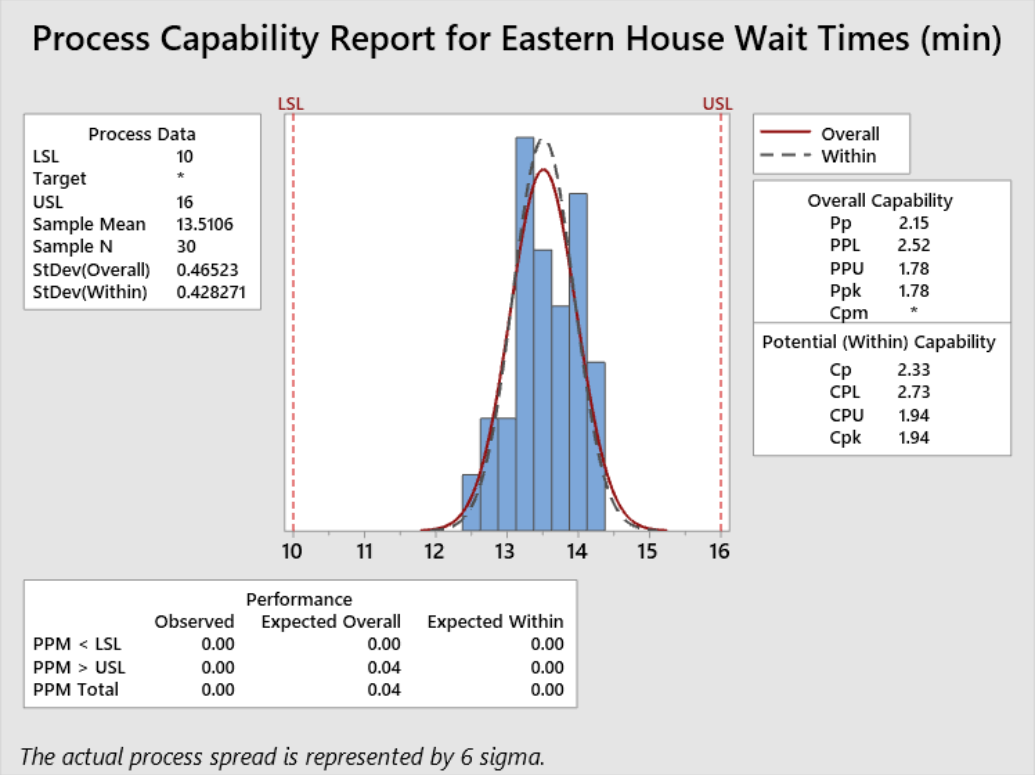
Stat > Quality Tools > Capability Sixpack > Normal

Are the necessary conditions satisfied for performing a capability analysis assuming normal data?



The actual process spread is represented by 6 sigma.

Stat > Quality Tools > Capability Analysis > Normal



The actual process spread is represented by 6 sigma.

AIAG (Automotive Industry Action Group) is considered the main group in developing and determining the "standard" definition of the following Indices:

Cp = Capability Index
Pp = Performance Index
Cpk = Capability Index which accounts for process centering
Ppk = Performance Index which accounts for process centering

Formulas

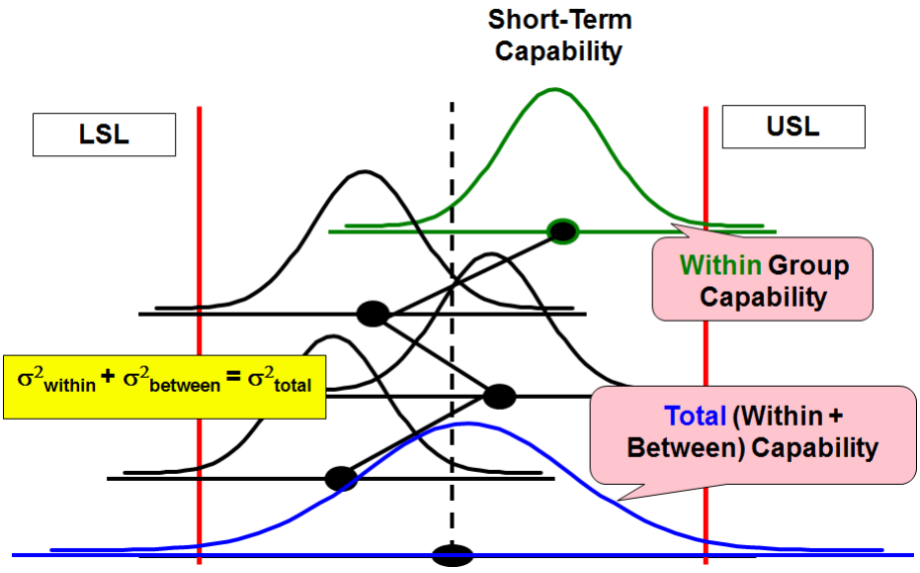
$$Cp = \frac{(USL - LSL)}{6 * \hat{\sigma}_{\bar{R}/d_2}} \qquad Pp = \frac{(USL - LSL)}{6 * \hat{\sigma}_S}$$

$$Cpl = \frac{(Mean - LSL)}{3 * \hat{\sigma}_{\bar{R}/d_2}} \qquad Ppl = \frac{(Mean - LSL)}{3 * \hat{\sigma}_S}$$

$$Cpu = \frac{(USL - Mean)}{3 * \hat{\sigma}_{\bar{R}/d_2}} \qquad Ppu = \frac{(USL - Mean)}{3 * \hat{\sigma}_S}$$

$$Cpk = Min(Cpl, Cpu) \qquad Ppk = Min(Ppl, Ppu)$$

Note the only difference is the estimator of Sigma



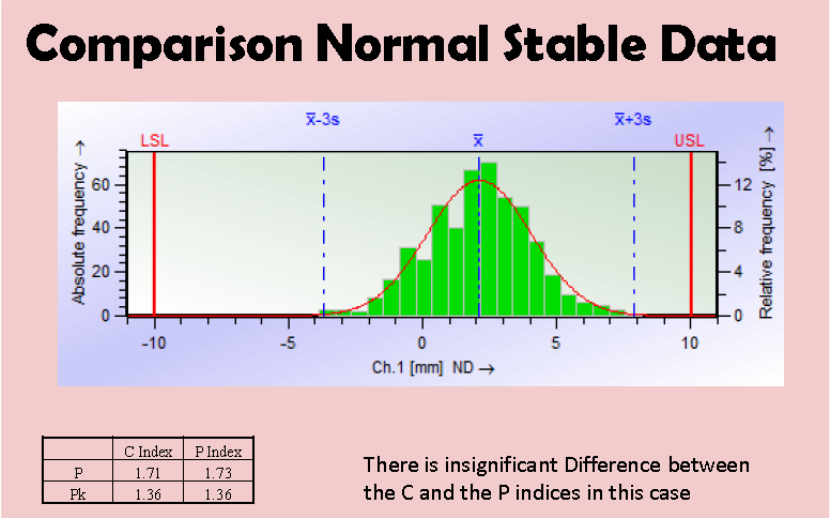
Definitions of $\hat{\sigma}$: σ is the estimate of the true process standard deviation; it can be computed in several ways:

1. $\hat{\sigma} = \frac{\bar{R}}{d_2}$, where $\bar{R}$ **is the mean of the ranges** for each subgroup and d_2 **is an unbiasing constant** dependent on the sample size n, or similarly $\hat{\sigma} = \frac{\bar{s}}{c_4}$, where $\bar{s}$ **is the mean of the standard deviations** for each subgroup and c_4 is an unbiasing constant dependent on the sample size n.

2. $\hat{\sigma} = \frac{\overline{MR}}{d_2}$, where $\overline{MR}$ **is the mean of the moving ranges** for the k independent trials and d_2 **is an unbiasing constant**

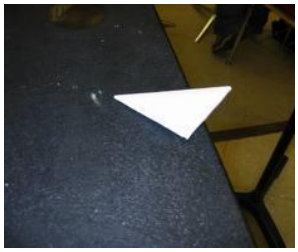
3. $\hat{\sigma} = s$ is the "typical" definition of the sample standard deviation s that you learn in your first stats class, specifically, $s = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m-1}}$.

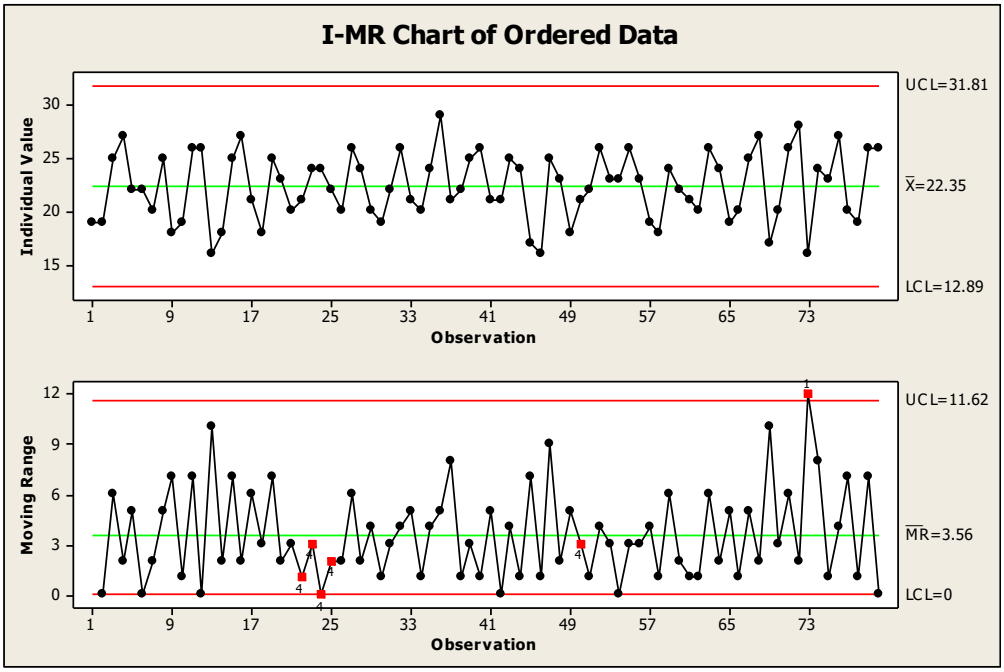
When a process is "stable" or "in-control," then there will be little difference between Cp and Pp or Cpk and Ppk
Why?



Example 3. Greg, Sara, Rick, and I are practicing for an upcoming Paper Football Championship against the Math Department. In order to hone our skills, we all sit down at a long table in a classroom and take consecutive turns trying to slide the paper football a given distance of 20 inches. I recorded the distances for us in a Minitab spreadsheet.

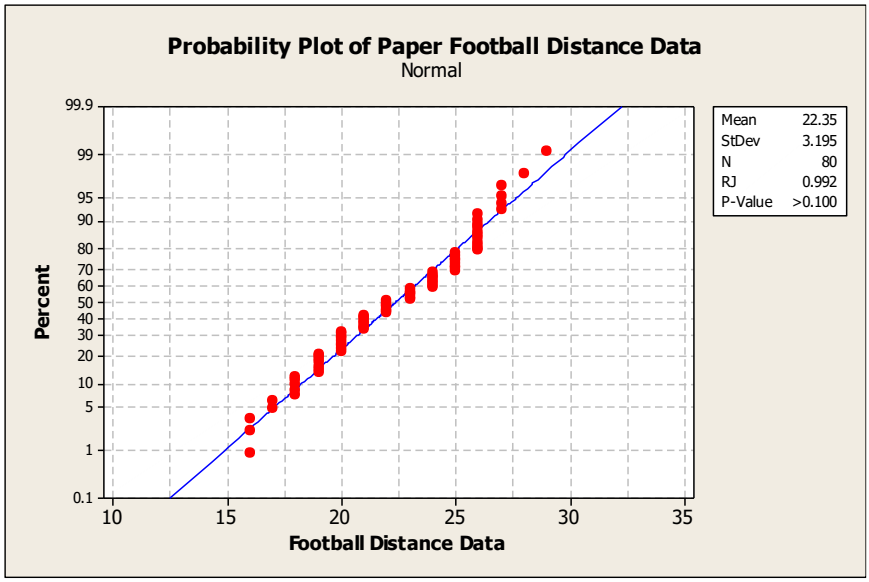
In order to determine if the process is stable, I'll first construct an I-MR chart. The distances are ordered in the following manner: Greg, Sara, Rick, Diane



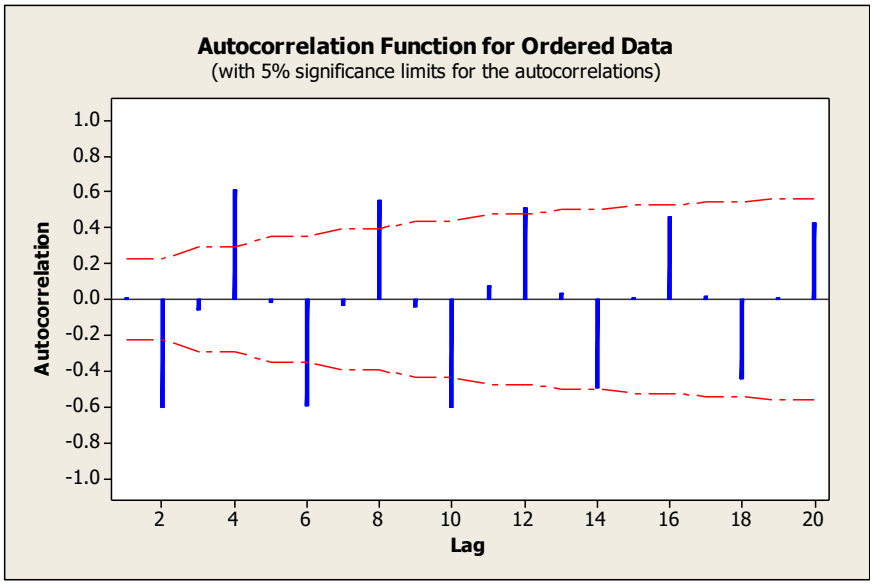


TEST 4. 14 points in a row alternating up and down. Test Failed at points: 22, 23, 24, 25, 50

The process seems somewhat stable, so let's move ahead with a normality test. Can we assume normality?

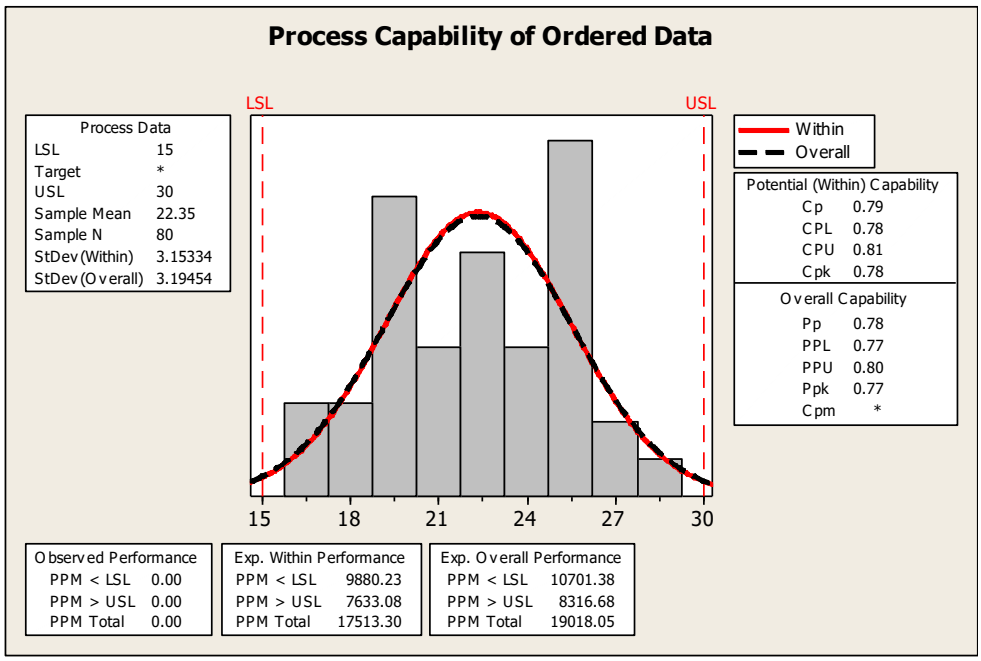


While we're checking, let's make sure there is no dependency in the data. What is this autocorrelation plot telling us about the process? What is happening with the football distances?



For now, we'll be the typical industry and ignore dependency … since most places would fail to even check for it.

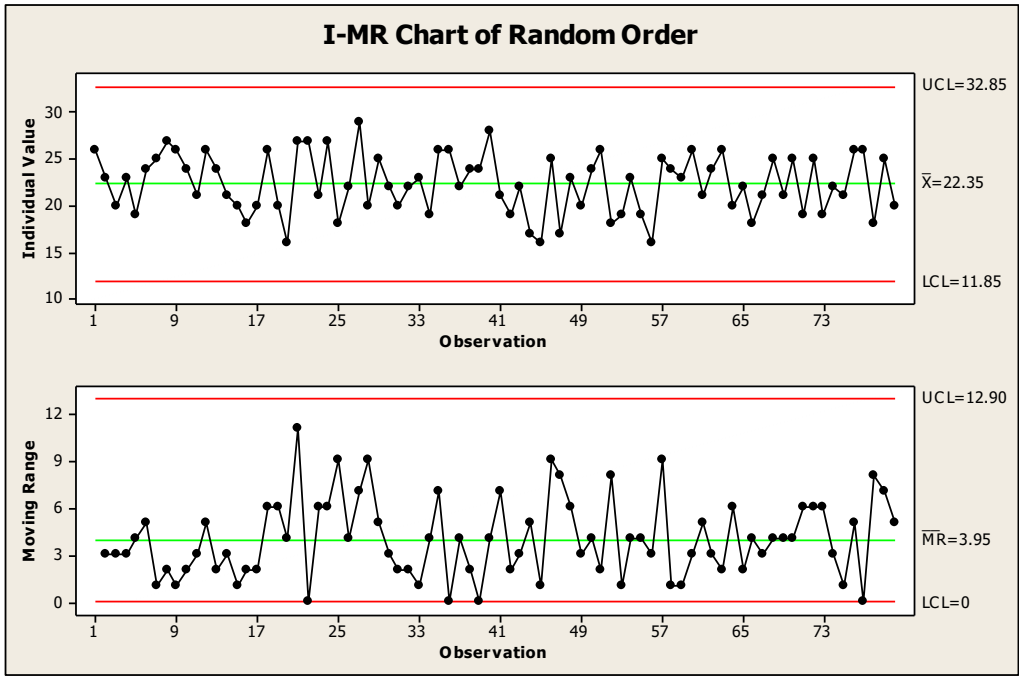
We'll run a capability analysis. Note the values for Cpk and Ppk.



Let's suppose that I randomized the order of our trials – so, now I'm just mixing up the trials by Greg, Sara, Rick, and Diane. Did the order of our shots really matter anyway? Let's check out the new I-MR chart.

Note: Xbar will still be the same value

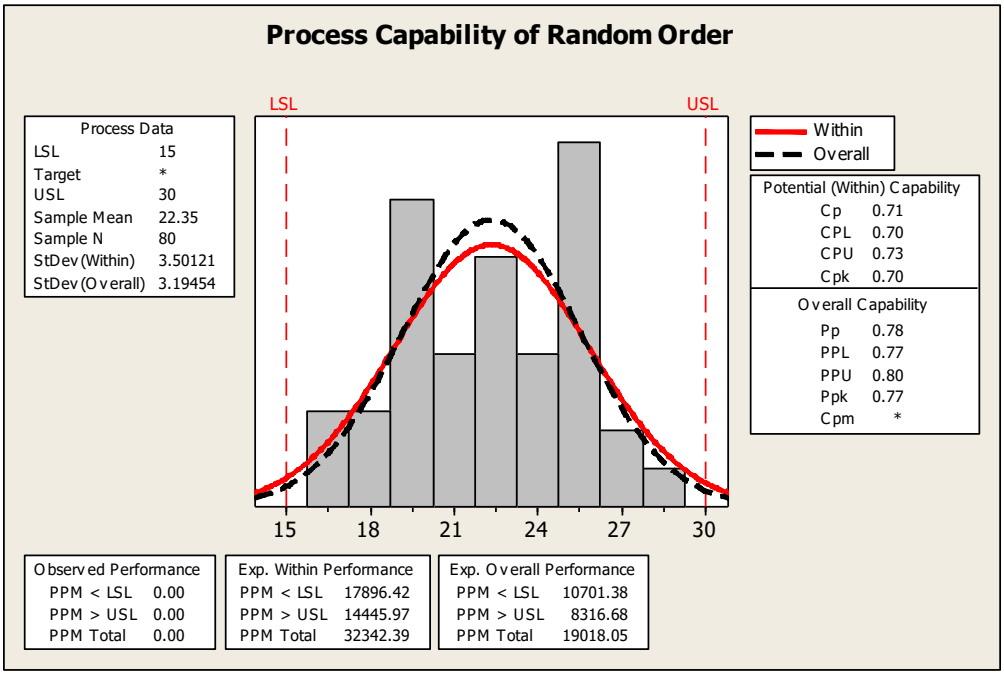
MRbar will now be a different value since the MR's are now different because of the change in data order



Will the data still be normally distributed? Do I need to recheck it for normality?

Will there still be dependency in the data?

Now, a capability analysis is definitely in order. Why has Cpk changed, while Ppk has remained the same?



Moral of the story?

Example 4. A rational subgrouping plan for collecting Easter Peep data follows:

- Select the first subgroup to consist of four samples from machine A, the second subgroup to consist of four samples from machine B, the third subgroup to consist of four samples from machine C, and the fourth subgroup to consist of four samples from machine D.
- Repeat this sequence over time. A diagram of this plan is displayed in Figure 1.

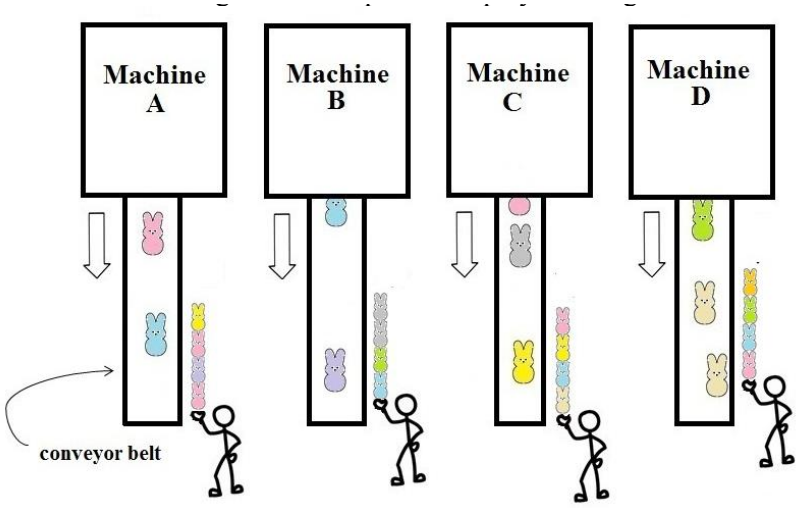
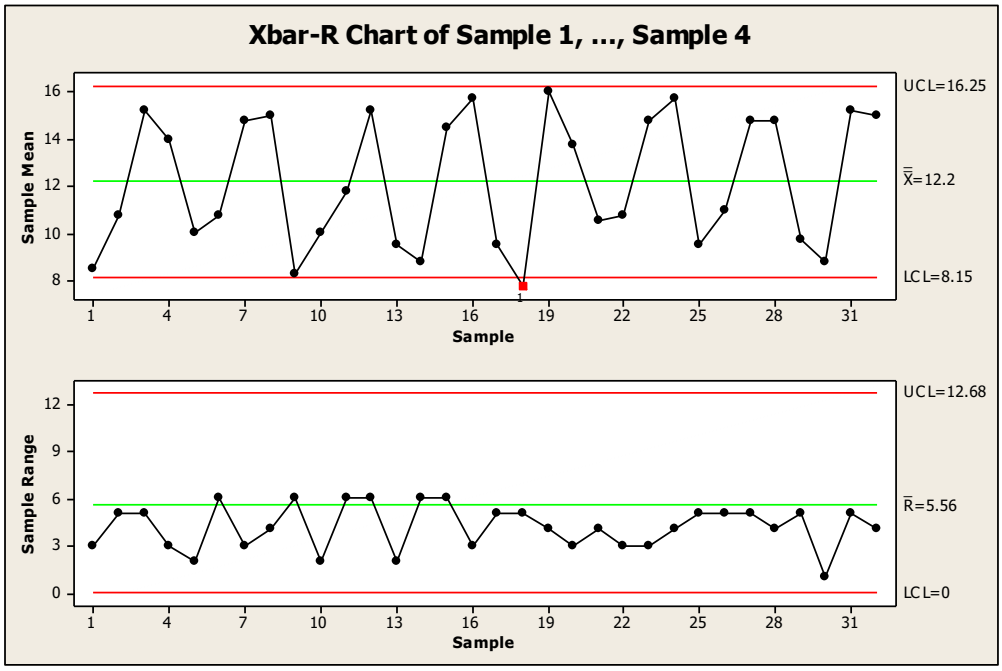
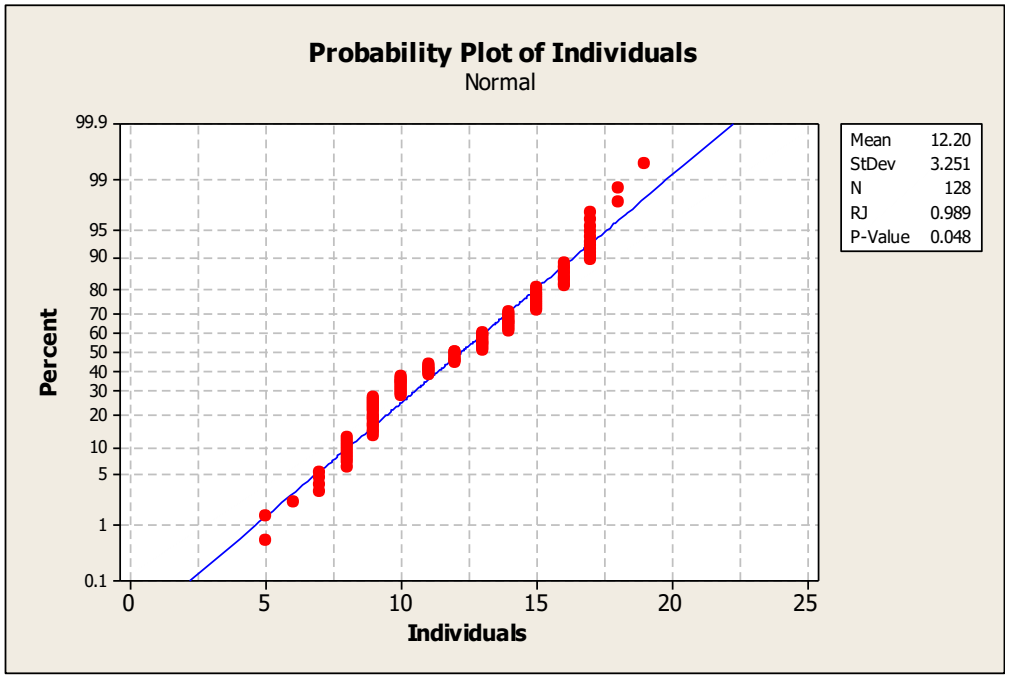


Figure 1. Peeps are produced by machines A, B, C, and D. At the end of the line, workers sample 4 peeps per machine and average the 4 peeps' sponginess. This plan subgroups or samples by machine.

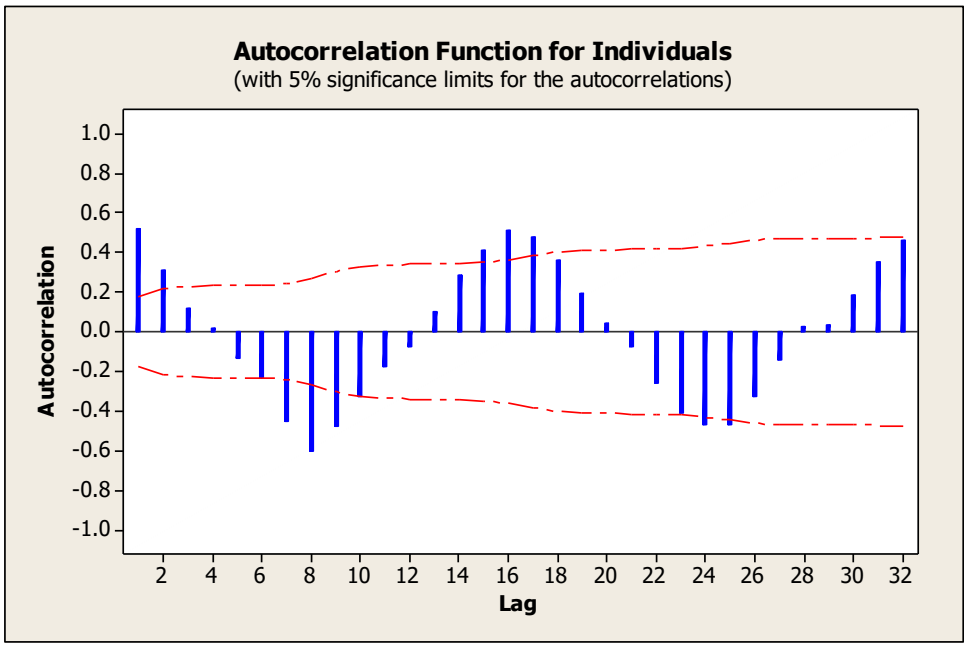
I'm going to make an Xbar-R chart for this data because it's variable (measurement data) and 4 samples are collected from each machine and their sponginess values are averaged. Unlike previous charts, **I set the process mean and process standard deviation.**



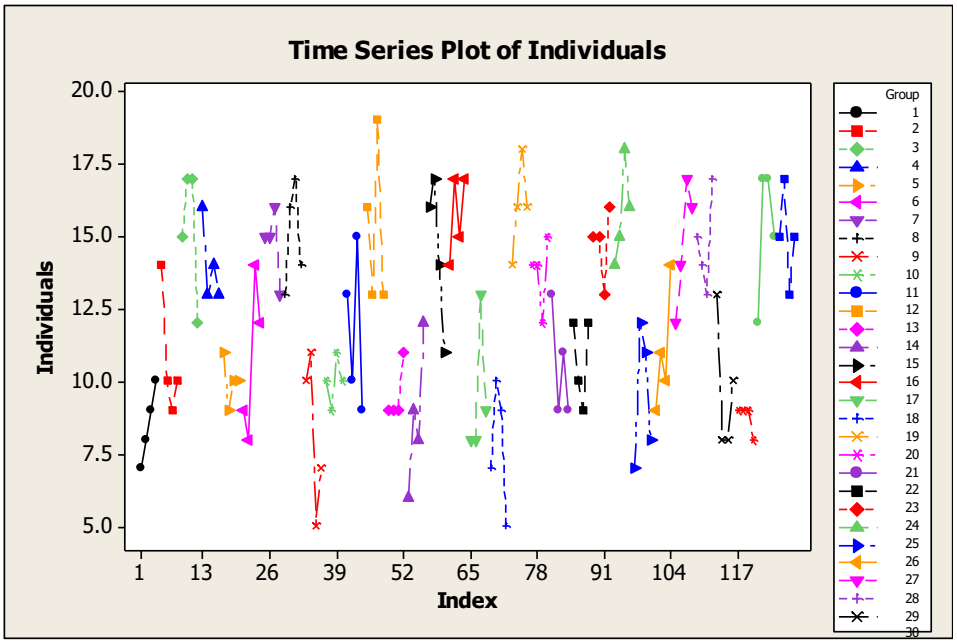
Ryan Joiner normality test:

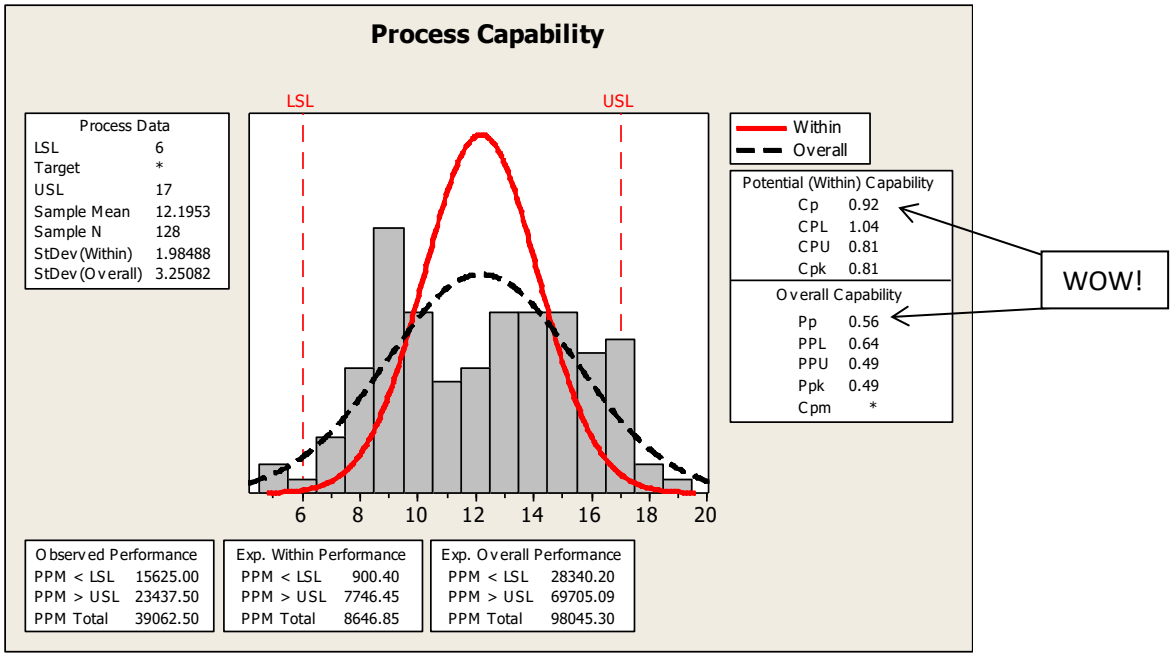


Autocorrelation plot:



Time Series plot for each trial, consisting of a subgroup of size $n = 4$:





One more Capability Index: **Cpm:** It is more concerned with the process being centered around a given **Target Cpm and formula for Cpm:**

- Cpm incorporates the target when calculating the standard deviation
- Like the sigma of the individuals formula, $\hat{\sigma}Cpm$ compares each observation to a reference value
- Instead of comparing the data to the mean $\hat{\mu}$, the data is compared to a given target T. These differences are squared.
- Thus any observation that is different from the target observation T will increase the $\hat{\sigma}Cpm$ standard deviation.

$$C_{pm} = \frac{(USL - LSL)}{6 * \hat{\sigma}_{C_{pm}}} \qquad \hat{\sigma}_{C_{pm}} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - T)^2}{n-1}}$$

- As this difference increases, so does $\hat{\sigma}Cpm$. And as $\hat{\sigma}Cpm$ becomes larger, the Cpm gets smaller.
- If the difference between the data and the target is small, so too is $\hat{\sigma}Cpm$. And as $\hat{\sigma}Cpm$ gets smaller, the Cpm index becomes larger. The higher the Cpm index, the better the process.

Capability Analysis with Non-Normal Data (continued)

If the process data is not normal and we want to do a capability analysis on the data, here are 3 options:

1. Transform by a function of the form **Y^{λ} or ln(Y)** by using the "Box-Cox Transformation" Minitab function to determine the appropriate transformation. Perform the capability analysis on the transformed data. Make sure the specifications have been transformed as well.

2. Perform a **Johnson Transformation**, also done in Minitab. Perform the capability analysis on the transformed data. Make sure the specifications have been transformed as well.

Stat > Quality Tools > Johnson Transformation

3. Determine the distribution that actually fits the original data. We'll do this with Minitab's "Distribution Identification" ability. Perform a capability analysis with a non-normal distribution using the original specification limits.

Example 1. Percentage of recyclables thrown in trash for 15 trash areas over 4 days -> 60 total data points. We set the specification limits at LSL = 0 and USL = 0.3. The data are in **Lesson16DATA_CapabilityAnalysis_Nonnormal.**

0.11765, 0.15385, 0.11765, 0.15385, 0.26190, 0.29032, 0.19298, 0.47273, 0.66279, 0.30137, 0.31429, 0.48333, 0.37143, 0.30000, 0.45000, 0.29730, 0.16384, 0.23684, 0.52174, 0.45000, 0.46774, 0.48000, 0.66667, 0.33333, 0.50000, 0.66667, 0.40909, 0.18182, 0.27358, 0.15652, 0.30303, 0.17333, 0.50000, 0.18182, 0.50000, 0.23529, 0.30952, 0.38462, 0.36585, 0.50000, 0.15909, 0.16176, 0.23529, 0.05405, 0.55000, 0.90000, 0.88889, 0.46154, 0.27500, 0.25926, 0.21429, 0.23529, 0.37736, 0.21569, 0.20313, 0.22059, 1.00000, 0.39394, 0.81818, 0.56250

Problem: The best transformation is ln(Y). But, with LSL = 0, we can't transform the LSL with ln(Y). Better option: fit a distribution to the data.

Minitab will try to fit the data against 8+ distributions: Exponential, Weibull, Lognormal, Loglogistic, etc. Important Minitab commands for future use in engineering courses: Individual Distribution Identification

Stat > Quality Tools > Individual Distribution Identification

Give Minitab your column of data, subgroup size, and it will try various fits

H_0: Data from distribution A versus H_a: data not from distribution A

Low p-value -> distribution is NOT a good fit

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	1.544	<0.005	
Box-Cox Transformation	0.277	0.643	
Lognormal	0.277	0.643	
3-Parameter Lognormal	0.288	*	0.379
Exponential	6.144	<0.003	
2-Parameter Exponential	3.849	<0.010	0.000
Weibull	0.648	0.088	
3-Parameter Weibull	0.428	0.334	0.091
Smallest Extreme Value	3.678	<0.010	
Largest Extreme Value	0.476	0.237	
Gamma	0.367	>0.250	
3-Parameter Gamma	0.542	*	1.000
Logistic	1.106	<0.005	
Loglogistic	0.360	>0.250	
3-Parameter Loglogistic	0.362	*	0.921
Johnson Transformation	0.270	0.666	

Lognormal is a good fit – not surprising! The transformation suggested by Minitab to make the recycling %'s normal is ln(data).

Stat > Quality Tools > Capability Analysis > Non-Normal; select the distribution of your choice