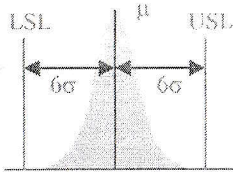
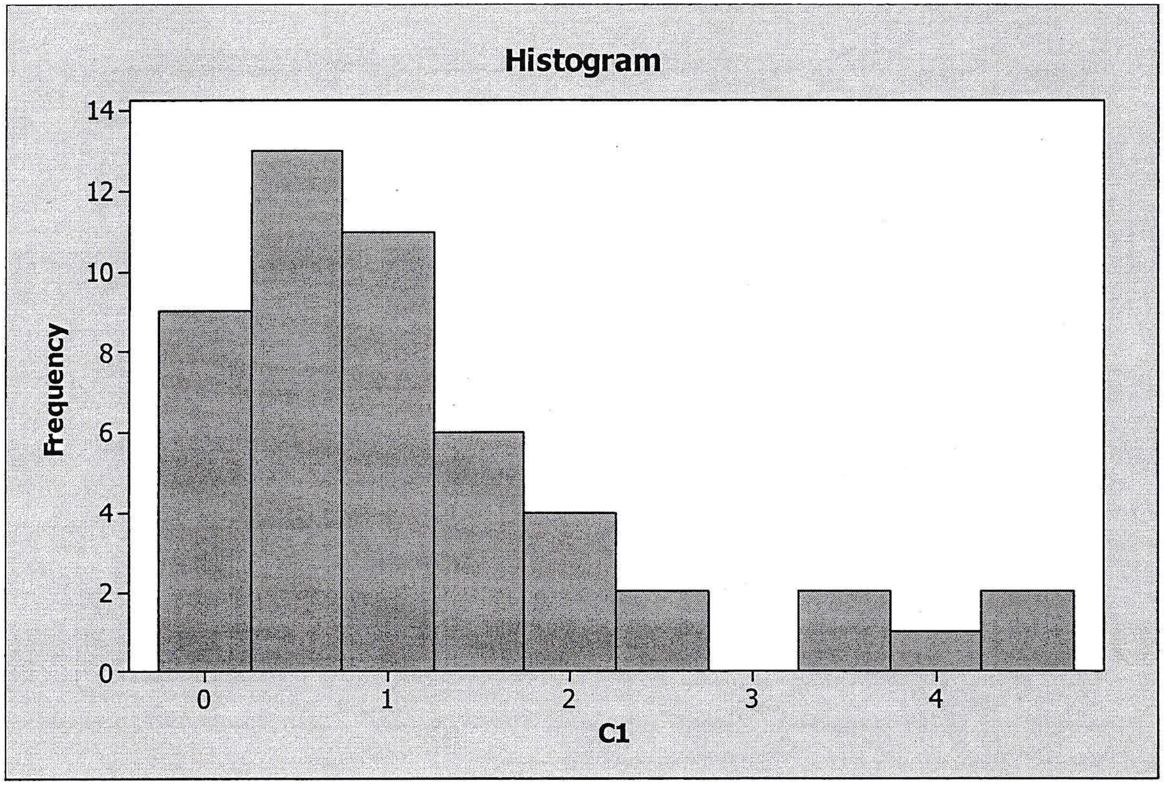


The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

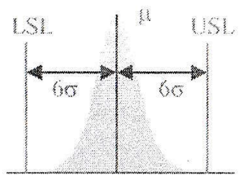
Data Transformation

Not All Data Is Normal

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS



Using "Normal Theory" to answer questions correctly depends on the data actually being Normal and Stable (no special causes).



Non-Normal Data

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

- Not all data is *Normal*, even after the "Special" causes have been removed.

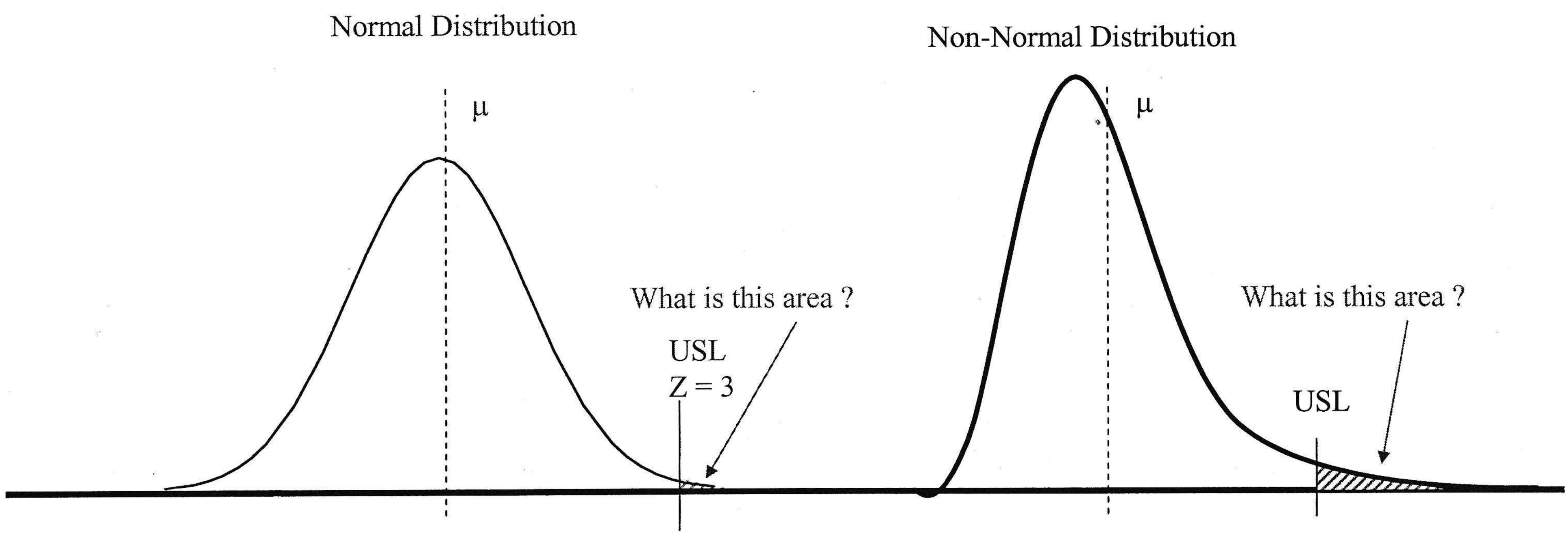
- Nothing is wrong with the data, it just naturally occurs this way.

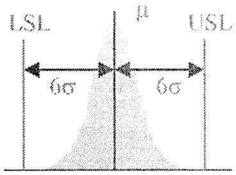
- When data is not normal, many of the statistical tools and methods won't work well…we'll need a method to overcome this problem.

Non-Normal Data Consequences

Consequences when using a statistical method that requires normality.

Example: Calculating Process Sigma or a Z-score.

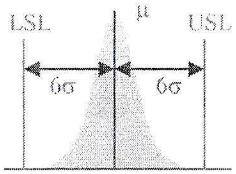




Non-Normal Data Consequences

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Method	Consequence of Non-Normality
Process Sigma Calculation	Incorrect process sigma value
Control Chart	False detection of some special causes, missed signal of others
Hypothesis Testing (t-test, ANOVA, etc)	Incorrect conclusion about differences between groups
Regression	Misidentification of important factors. Poor predictive abilities



Normality Exercise

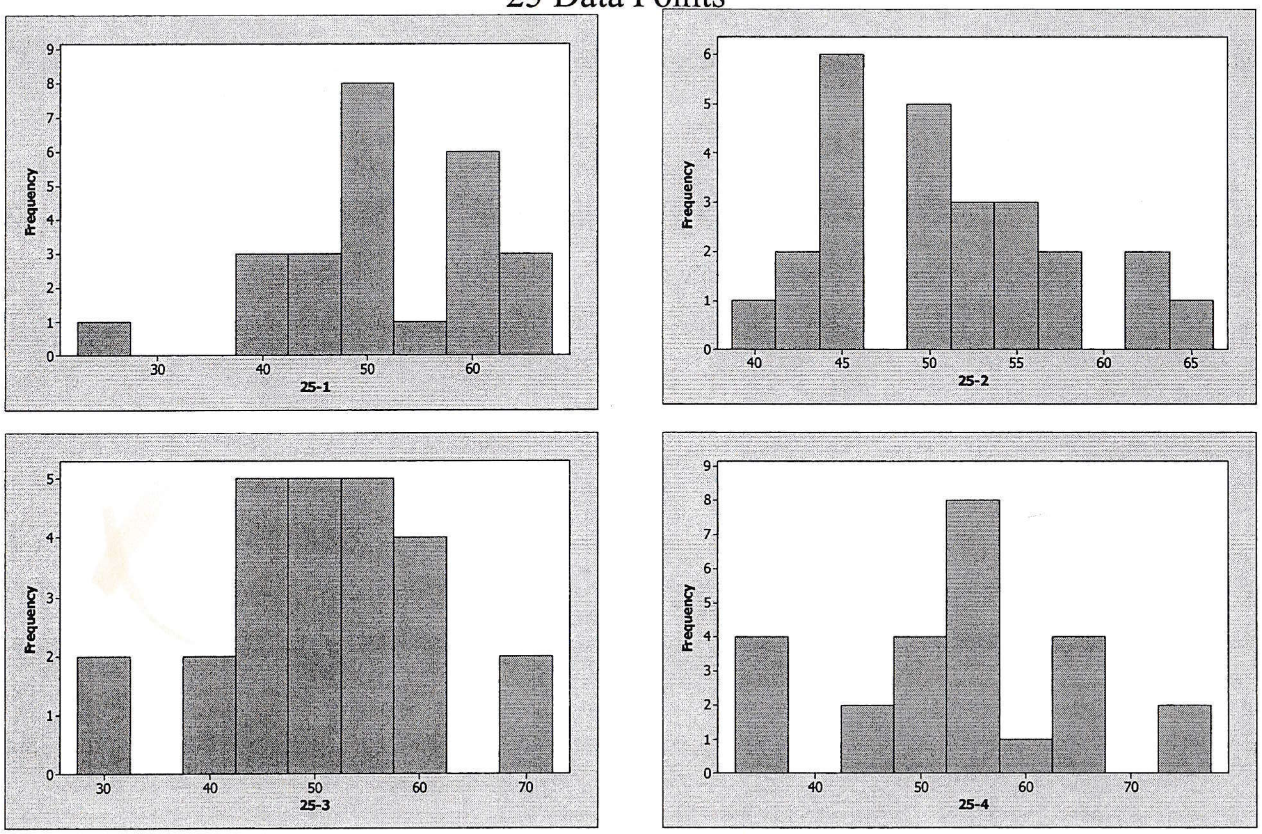
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Directions:

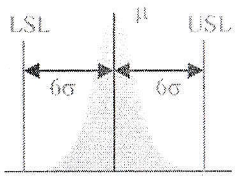
1. Look at each histogram and assess which data sets come from a normal distribution.
2. Circle or mark the ones you think are normal.

25 Data Points

Bi



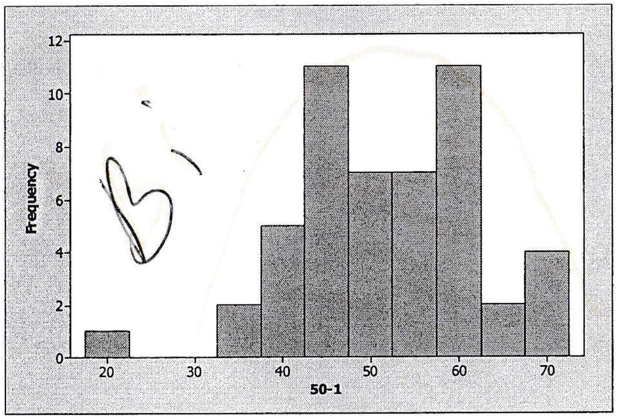
skew



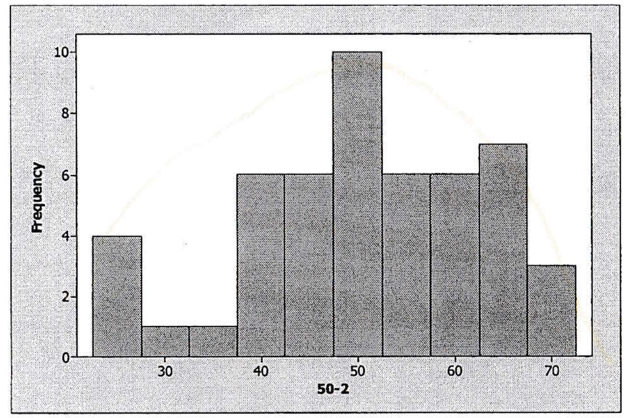
Normality Exercise

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

50 Data Points

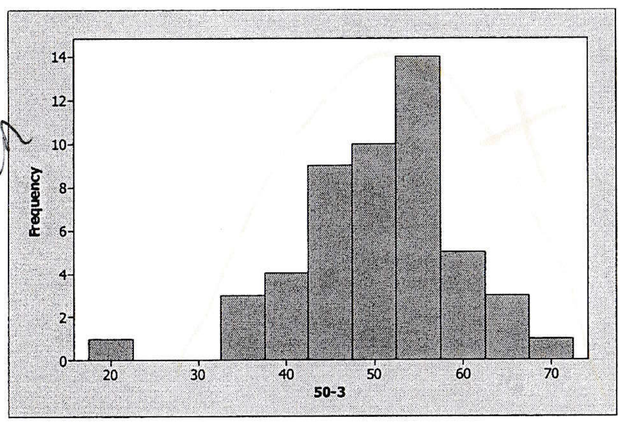


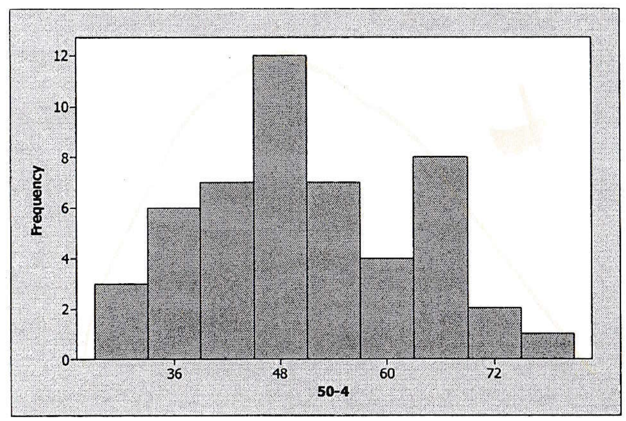
Bi



No

Normal
outlier





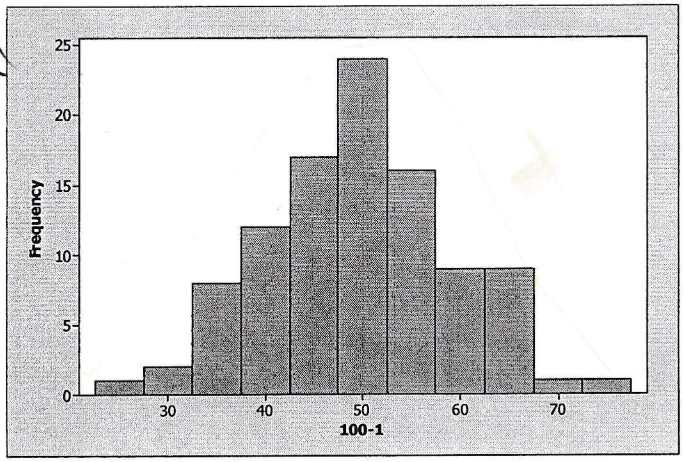
Bi

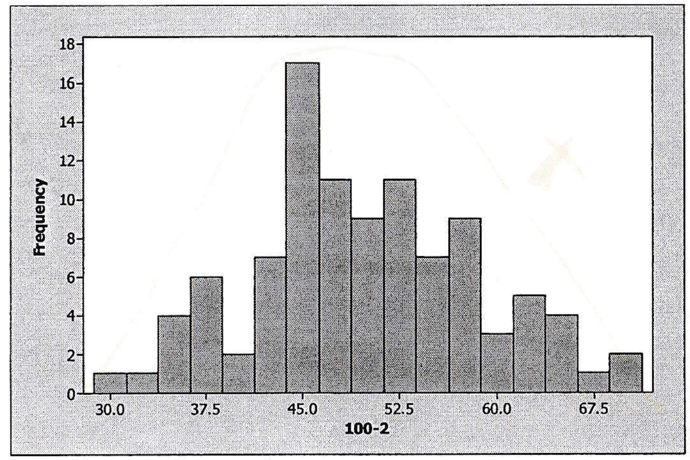
Normality Exercise

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

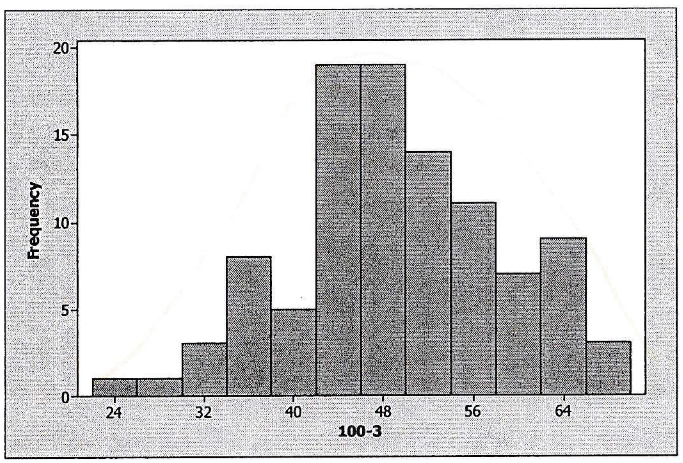
100 Data Points

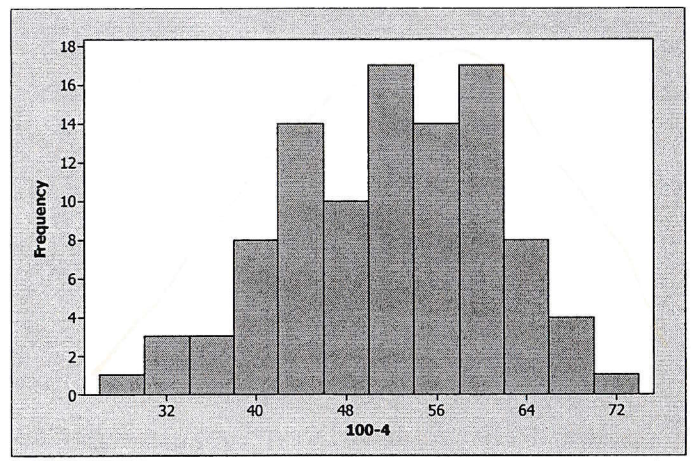
Normal



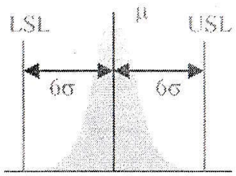


Normal





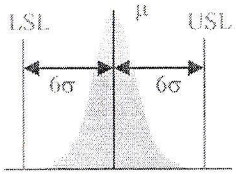
Bi



Normality Exercise: Answers

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

- Looking at histograms can be deceptive.
  - Each of the data sets on the previous pages were randomly generated in Minitab using the "Normal Distribution" function (Mean = 50, Std. Dev. = 10).
  - **They are all NORMAL Distributions.**
- As you can see, it is difficult to tell if data is Normal by looking at histograms of n=25, n=50 and even for n=100.

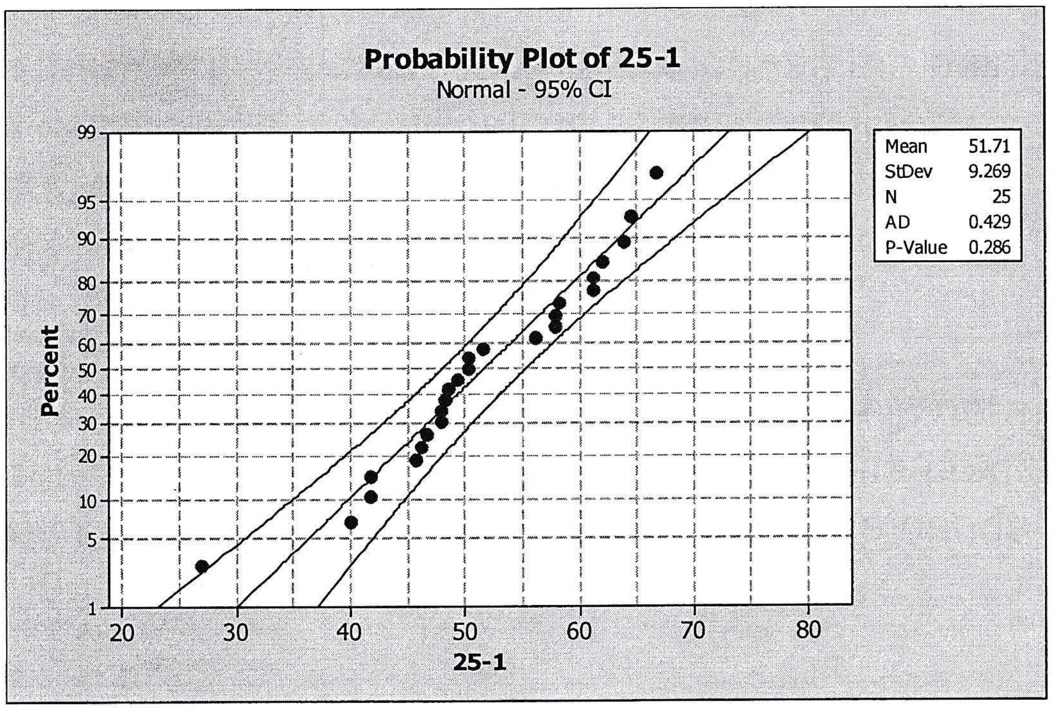


Normal Probability Plots

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

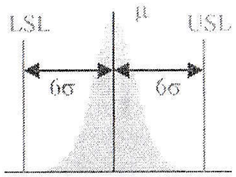
Normal probability plots offer a better way to asses Normality.

Listed below is the normal probability plot for the 1st data set used in the last exercise (n=25)



- If the data is Normal, the points will fall on a "straight" line.
- "Straight" means within the 95% confidence bands.

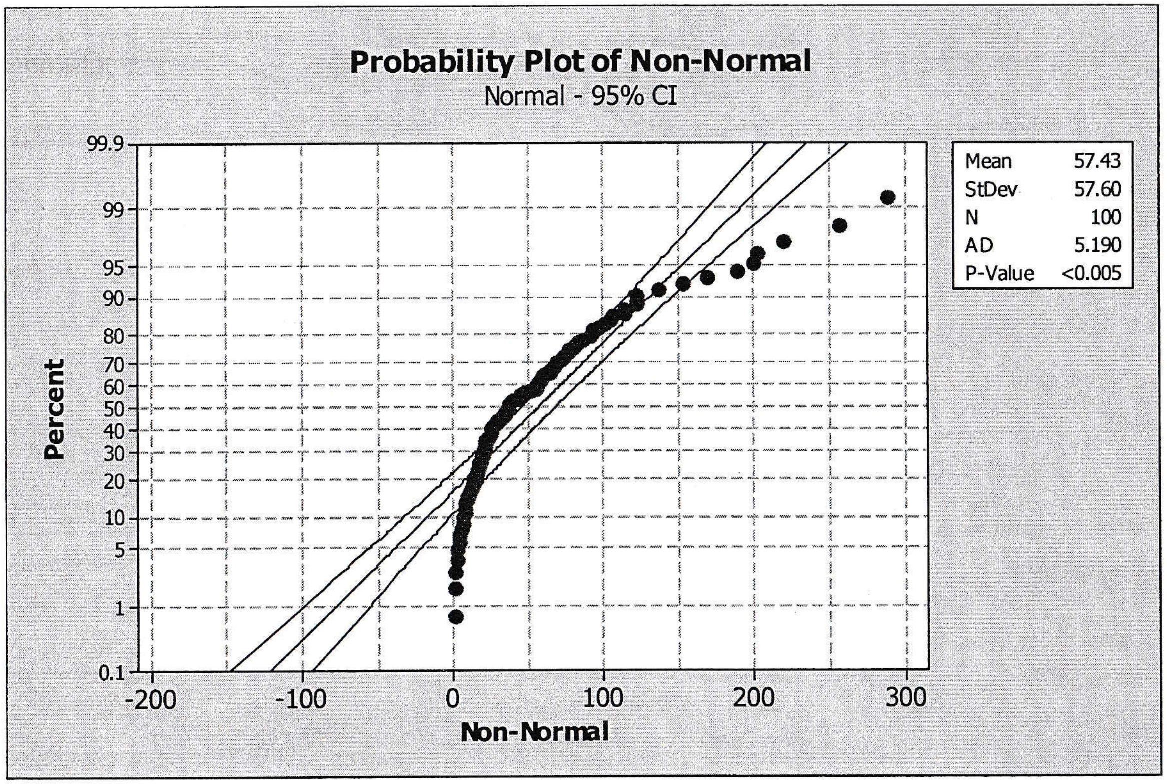
Or you can say the data is normal if approx. 95% of the data points fall within the confidence bands

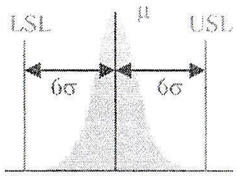


Normal Probability Plot

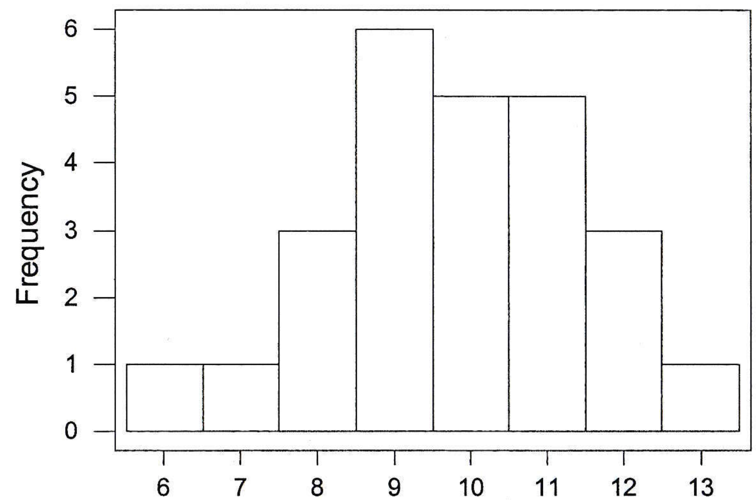
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

- Non-Normal Data





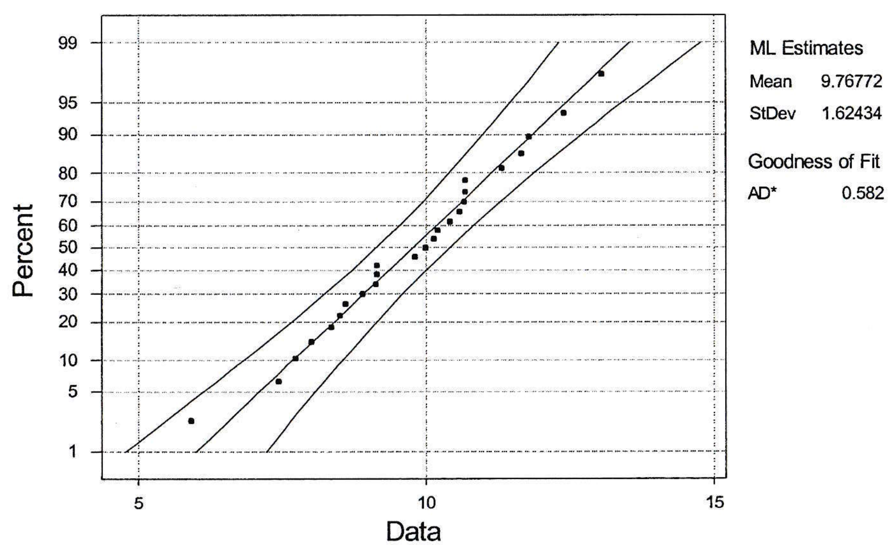
Normality Test in Minitab



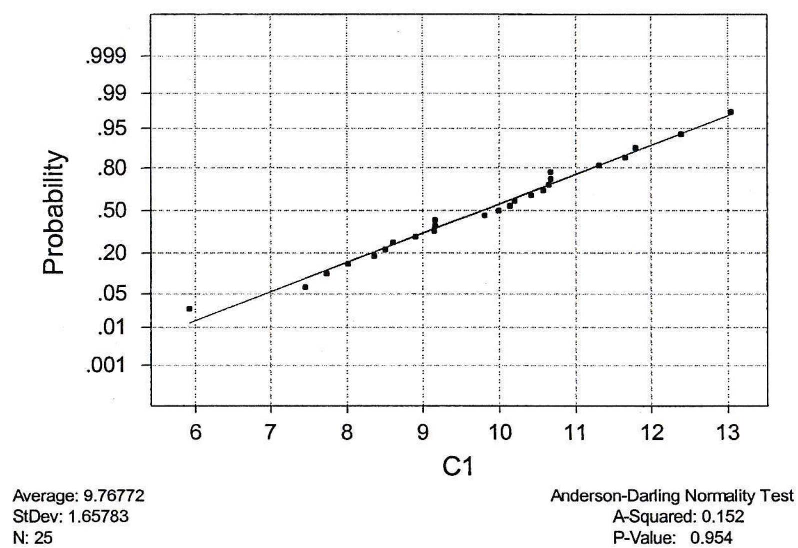
Results:

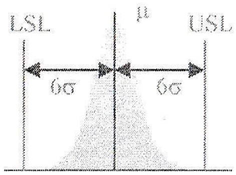
Your results will be different

Normal Probability Plot for C1
ML Estimates - 95% CI



Normal Probability Plot



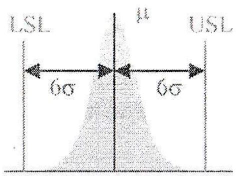


What do I do if my data is not Normal?

Caution! Sometimes “Special” causes make normal data appear skewed.

- Plot data in time series and look for shift in the process averages or in the variability.

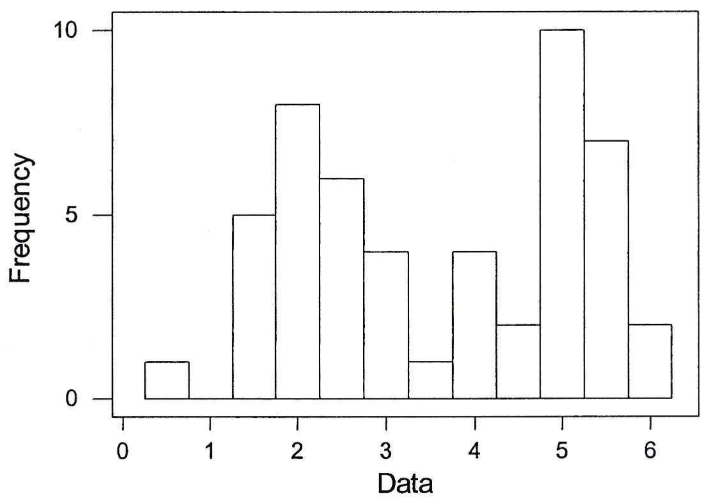
- Graph the data (Dot plots, Box Plots, etc) using rational sub-grouping to determine if there is a “group” that is causing the skewed distribution.

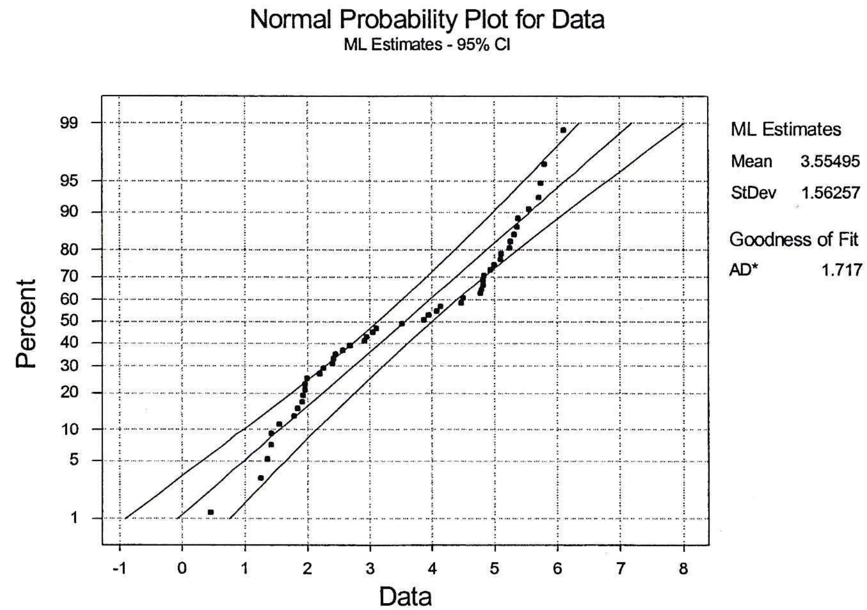


Example

A BB is assessing the time required to complete a transactional process. Data is collected for 50 transactions and they would like to determine what % of the transactions take more than 6 hours to complete.

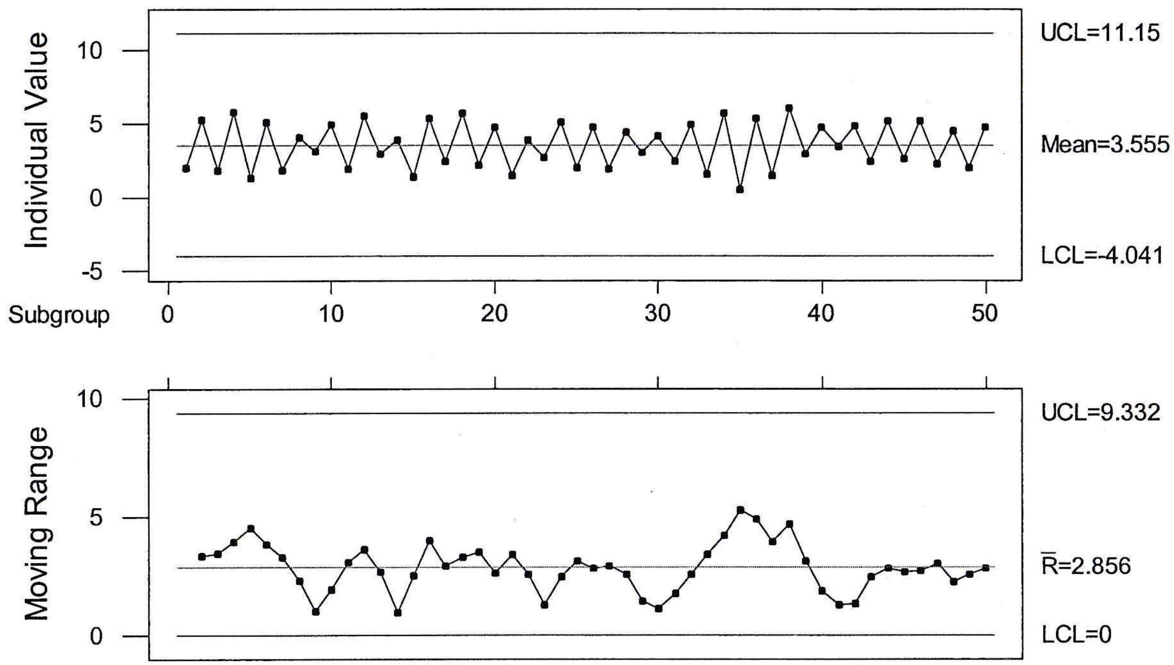
The normality assessment tools indicate that the data is not Normal





Time Series Graph

I and MR Chart for Data

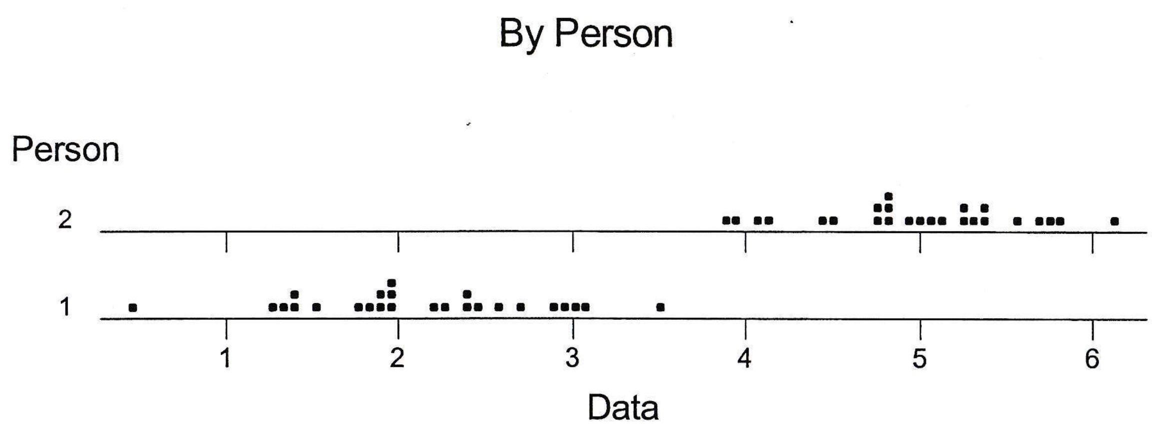


Pattern in the data ?

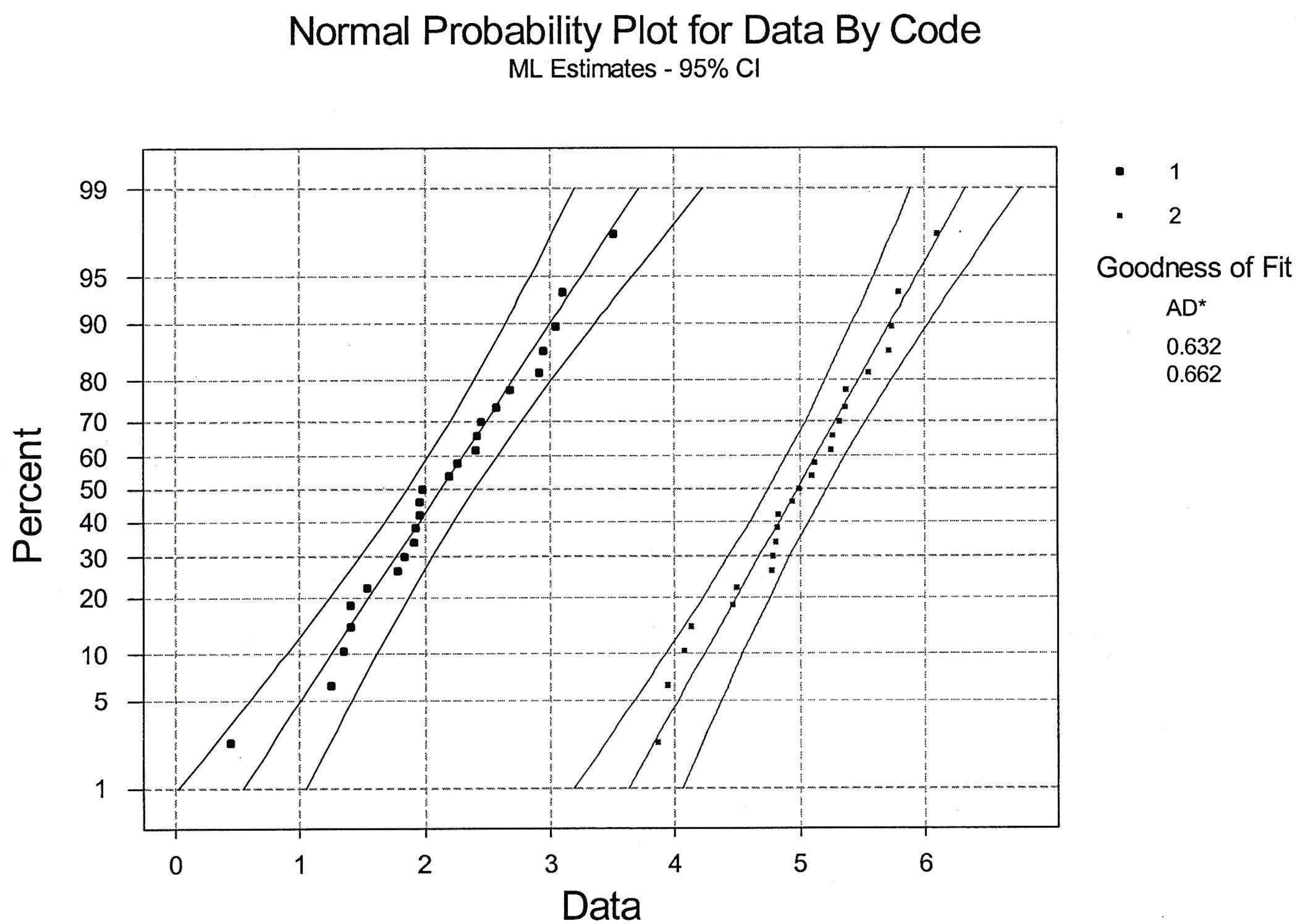
Graph by Rational Grouping

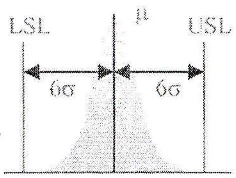
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

The BB reviews the PMAP and determines that there are two individuals that perform this type of transaction. The data is graphed using this information.



Normal Probability Plot





Transforming Data

Transformation is a re-expression of the data.

Examples:

- Transform US Dollars to Euros
- Transform seconds into minutes
- Transform Fahrenheit to Celsius

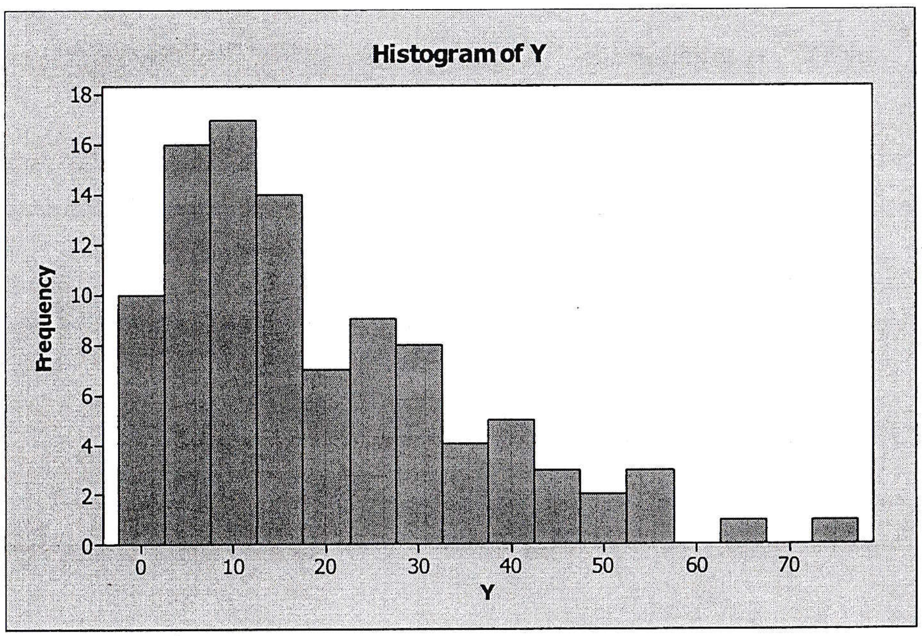
Or Transform Non-Normal Data to Normal Data

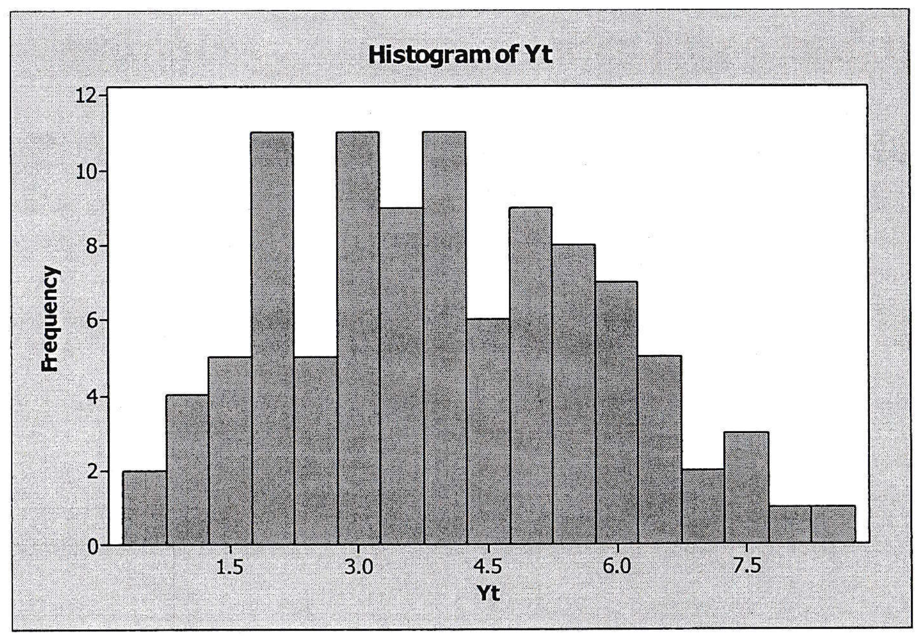
Shape Changing Transformations

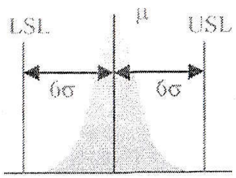
Any transformation that changes the shape (distribution) of the data.

Example: Square Root Transformation $$Y_T = \sqrt{Y}$$

Note: Y_T denotes the transformed Y







Types of Transformations

- Square Root to Normal $\quad Y_T = \sqrt{Y}$
- Log to Normal $\quad Y_T = Log(Y)$
- Inverse to Normal $\quad Y_T = \dfrac{1}{Y}$
- Inverse Square Root to Normal $\quad Y_T = \dfrac{1}{\sqrt{Y}}$
- Quarter Root to Normal $\quad Y_T = \sqrt[4]{Y}$
- Etc…

Choosing a Transformation

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

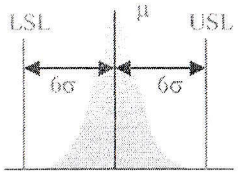
How do I pick the correct transformation for my data ?

1. Trial & Error

Try all of the transformations and pick the one that produces the most Normal data.

2. Use the Box-Cox power transformation method.

- A power transformation raises the Y value to a power denoted by lambda (λ): $Y_T = Y^{\lambda}$
- λ typically falls between –2 and 2
- The Box-Cox method finds the λ that minimizes the departure from Normality.



Power Transformations

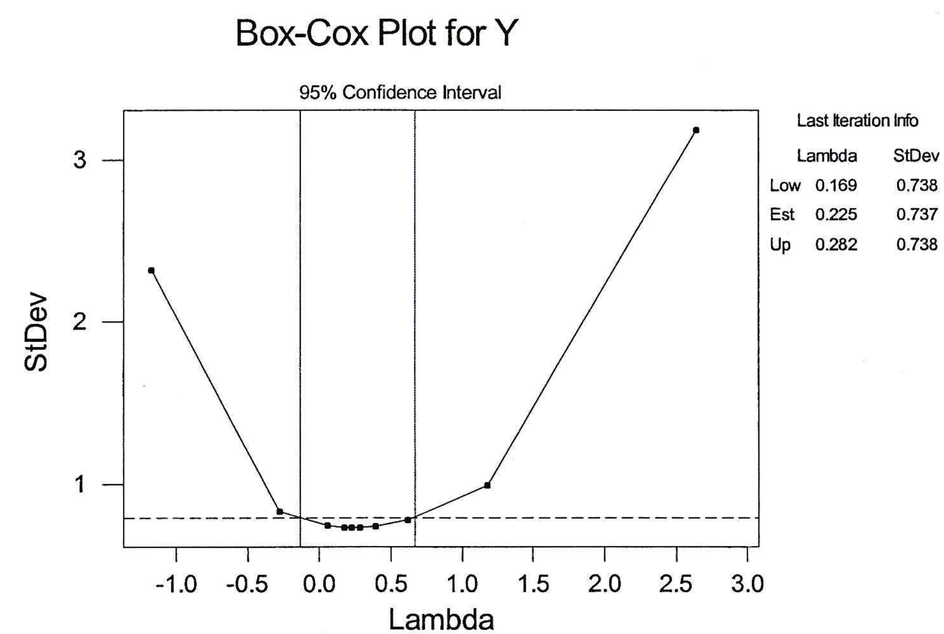
The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

λ (Power)	Y^{λ}	Common Names
-2	$\frac{1}{Y^2}$	Reciprocal squared
-1	$\frac{1}{Y}$	Reciprocal
-0.5	$\frac{1}{\sqrt{Y}}$	Reciprocal square root
0	$Log(Y)$	Log*
0.5	$\sqrt{Y}$	Square root
1	---	No transformation
2	Y^2	Squared

* $Y^0 = 1$, but Y^{λ} as λ approaches zero = Log (Y)

Selecting a λ Value

- Choose a lambda value from within the recommended 95% confidence interval that corresponds to a simple transformation, if possible.



For the data set above, Box-Cox recommends a λ anywhere from –0.3 to 0.7

λ=0.25 is within this range (1/4 root) $Y_T = \sqrt[4]{Y}$

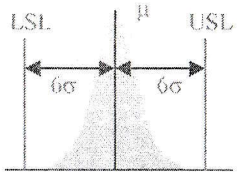
Transformation Exercise

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Background:

A customer service center measured the time to respond to and close customer complaints. The time for the last 100 complaints is listed in the File: “Complaint.mtw”

1. Use one of the methods to determine if the data is “Normal”.
2. If the data is NOT Normal, use the Box-Cox method to determine a lambda value and transform the data.
3. Test the transformed data for normality



Transforming Back to Natural Units

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

Once data has been transformed and analyzed, sometimes it is necessary to transform the data back to the Natural units.

Example: To determine the control limits for a control chart.

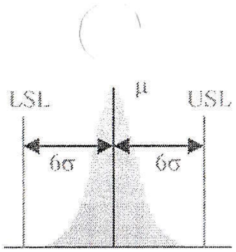
<u>Log Transformations:</u>

The back transformation for the Log is called the **antilog**

Antilog = 10^x

If $\log_{10}(Y) = Y_t$, then $10^{Yt} = Y$

$Log_{10}(100) = 2$, then $10^2=100$



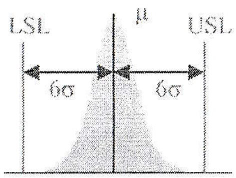
Transforming Back to Natural Units

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS

- Other "Back" transformations

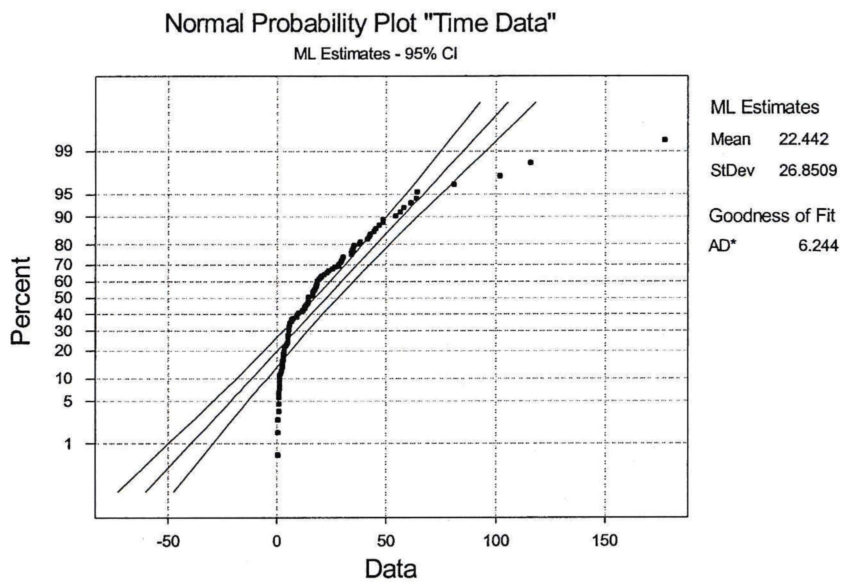
Transformation	Back Transformation
Square	Square Root
Square Root	Square
Reciprocals	Multiply by the function in the denominator*

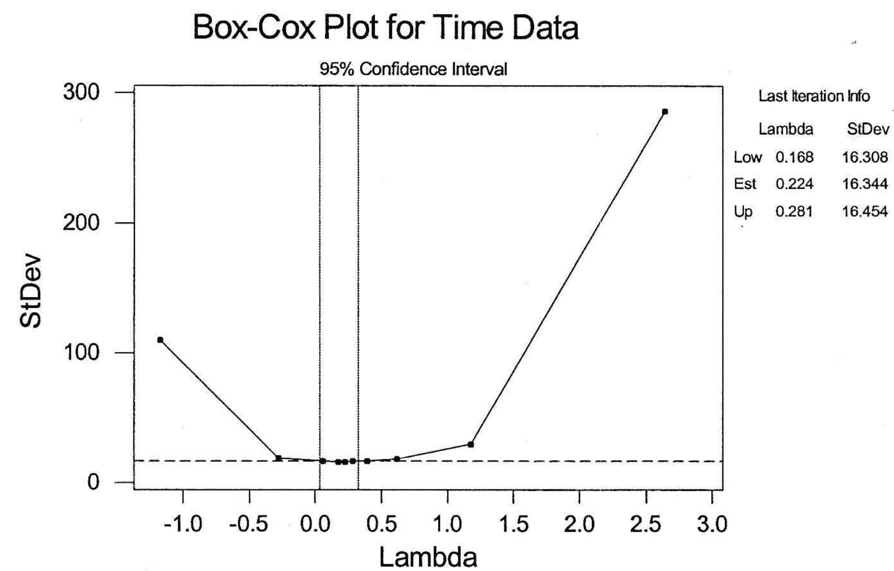
* Caution: Back transforming the reciprocal switches the UCL and the LCL and should only be used if Y is positive.



Answer to Transformation Exercise (Pg 29)

The Tech Group
WE PROVIDE INNOVATIVE SOLUTIONS





Transformed Data

$\lambda = 0.25$