How to Handle Non-Normal Data

Russ Aikman

The University of Texas at Arlington

The University of Texas
ARLINGTON™

- Much of this presentation was adapted from **Lean Six Sigma Black Belt Course Material** created by George Group (now a division of Accenture)

- **Other references:**
  - ***Advanced Statistics Demystified*** **by Stephens**
  - ***Practical Nonparametric Statistics*** **by Conover**
  - ***Normality and the Process Behavior Chart*** **by Wheeler**
  - ***The Six Sigma Practitioner's Guide to Data Analysis*** **by Wheeler**

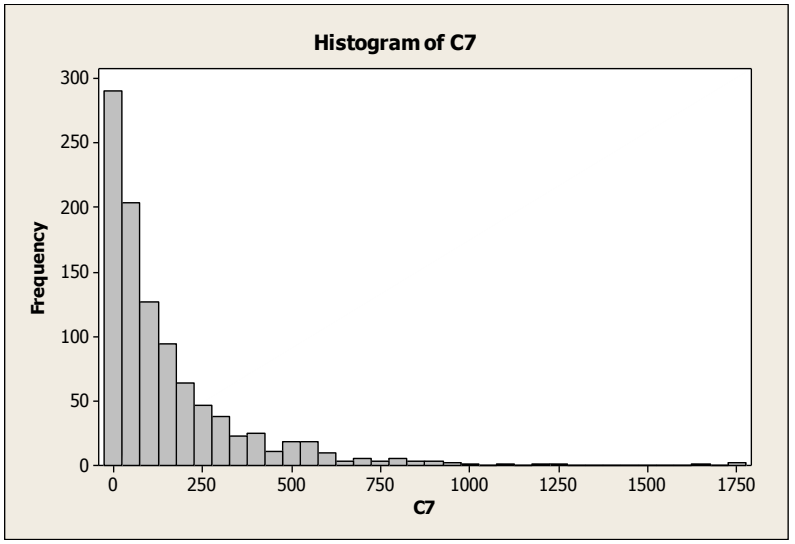
- **Examples are shown using Minitab statistical software**

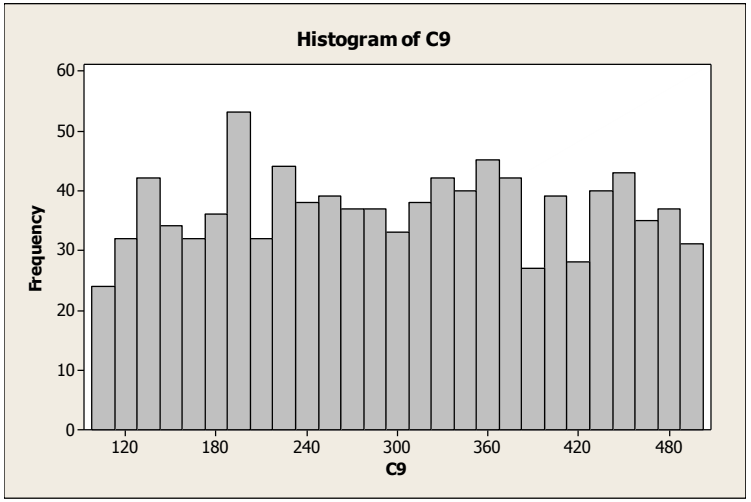
AGENDA

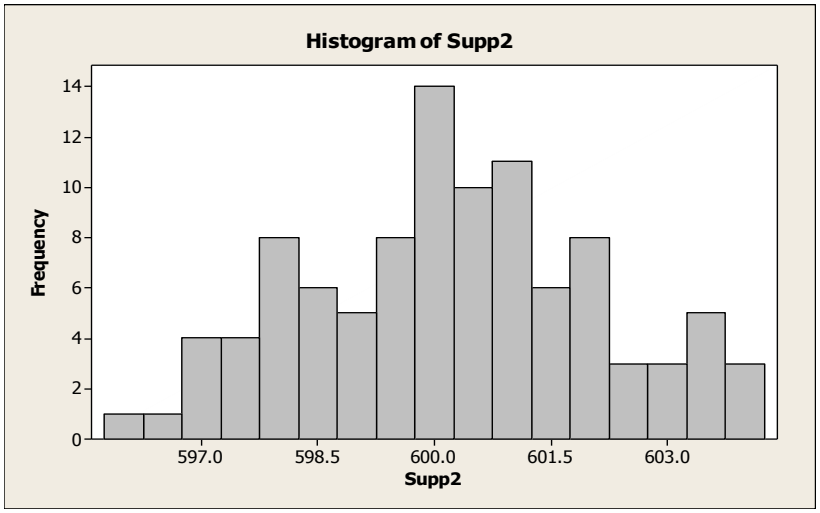
- **Data Distributions**
- **Handling Non-Normal Data: *A General Approach***
- **Validate the Data**
- **Data Stratification & Outliers**
- **Control Chart & Capability Considerations**
- **Non-Parametric Statistics**
- **Transformation of Data**
- **Appendix**
  - **Identifying Distribution Type**

Data Distributions

- Many different types of data distributions exist. When graphed using a histogram, the shape of a distribution can be more easily discerned.







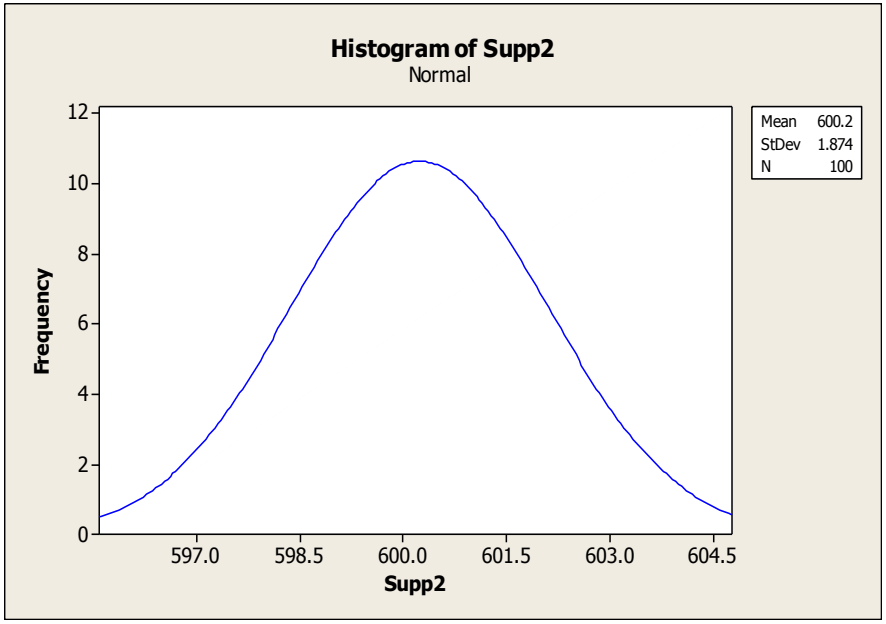
Data Distributions

- Distributions show the frequency and shape of aggregated data values.
- The distribution provides a 'snapshot' of a system based on data collected over a period of time.
- Different types of physical systems tend to have specific data distributions.

The statistical tools or methods that are appropriate for analyzing data are *dependent on the type of distribution*.

Data Distributions

- The most commonly used distribution in statistics is the **Normal Distribution**.



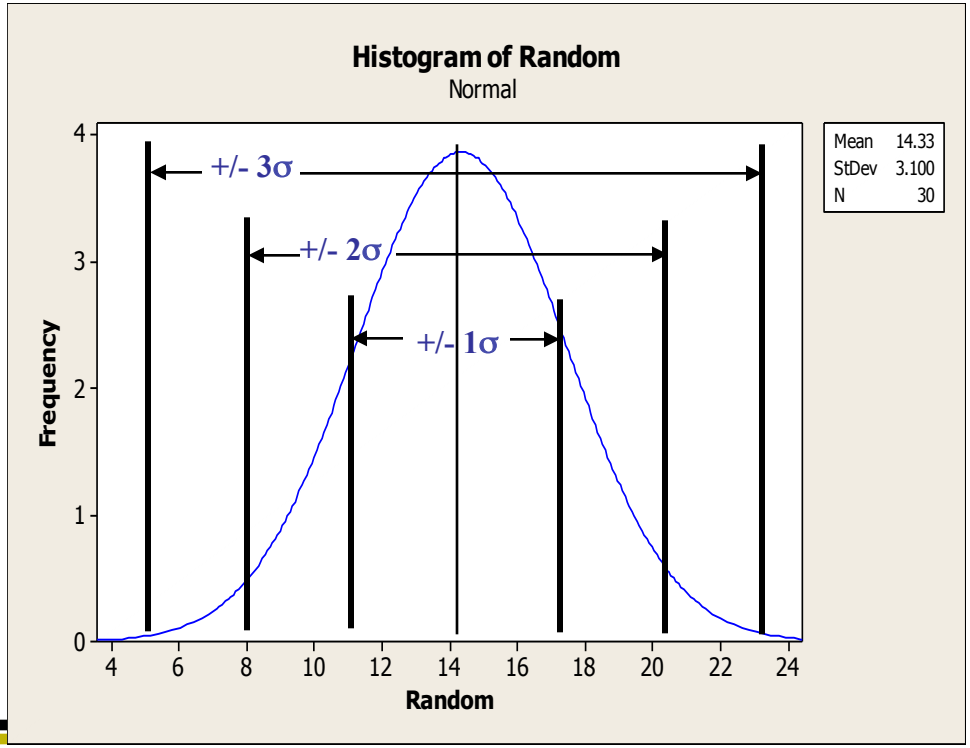
- A dataset must be Normally Distributed in order to use many statistical tools. This common assumption *must* be verified.

The Normal Distribution

- **Characteristics of a Normal Distribution:**
  - Classic, symmetrical bell shaped curve
  - The area under the curve can be used to predict process performance over time

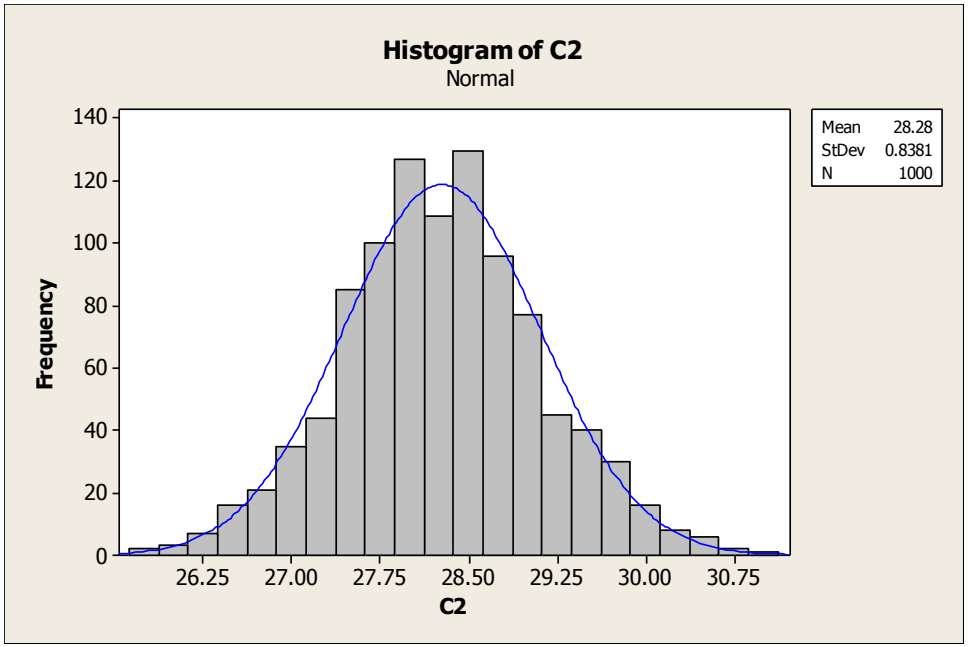
- **Properties:**
  - 68.3% of normally distributed data falls within +/- 1σ of the mean
  - 95.5 % falls within +/- 2σ
  - 99.73% falls within +/- 3σ

Many physical systems are normally distributed



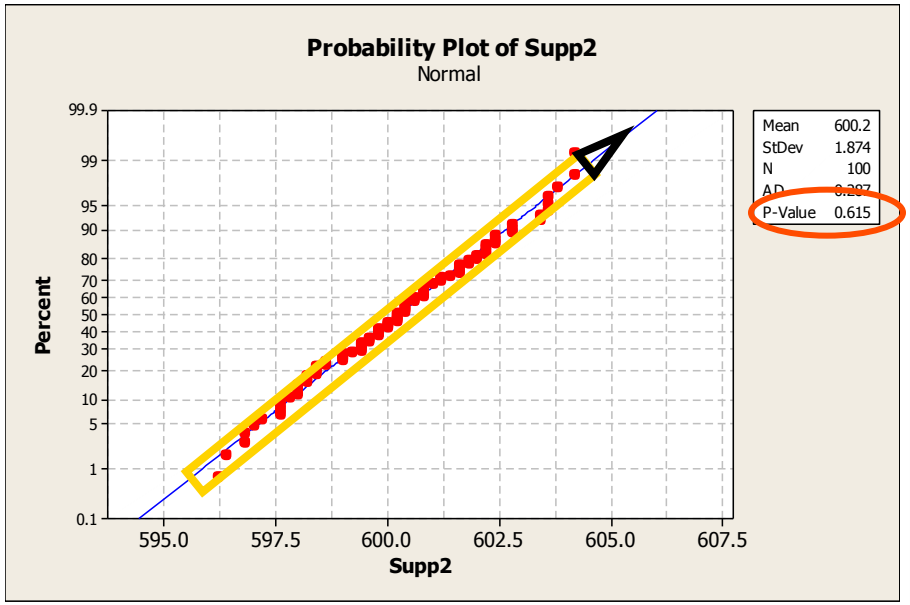
The Normal Distribution

- Examples of Normally Distributed phenomena
  - Diastolic blood pressure
  - Chest size
  - Height & Weight
  - Diameter of trees
  - Length of cut pipe
  - Weight of gear
  - Rainfall amounts
  - Astronomy – location of stars
  - Investment returns - diversified portfolio

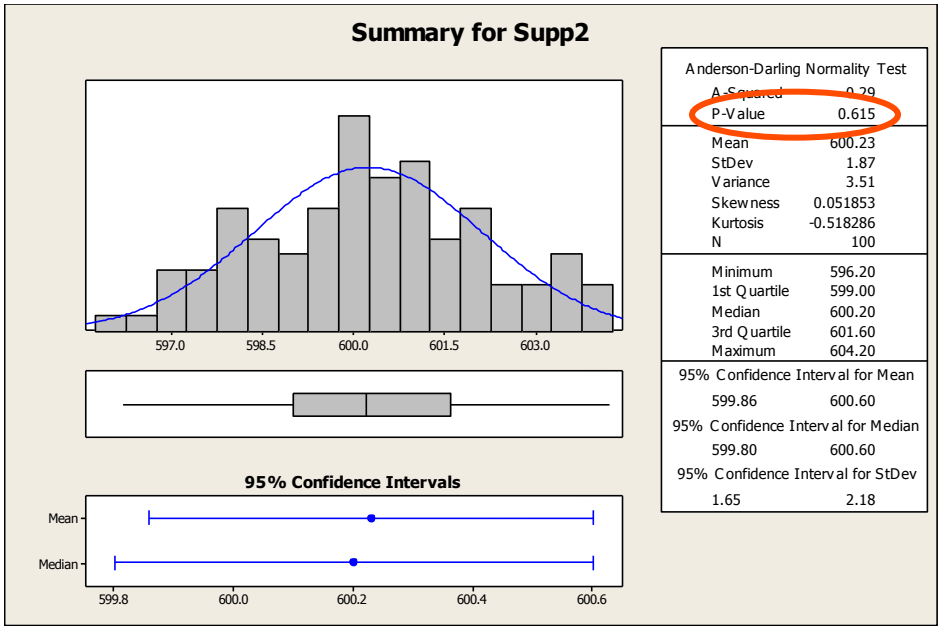


The Normal Distribution

- How to verify if a data set is Normally Distributed?



Stat>Basic Statistics>Normality Test

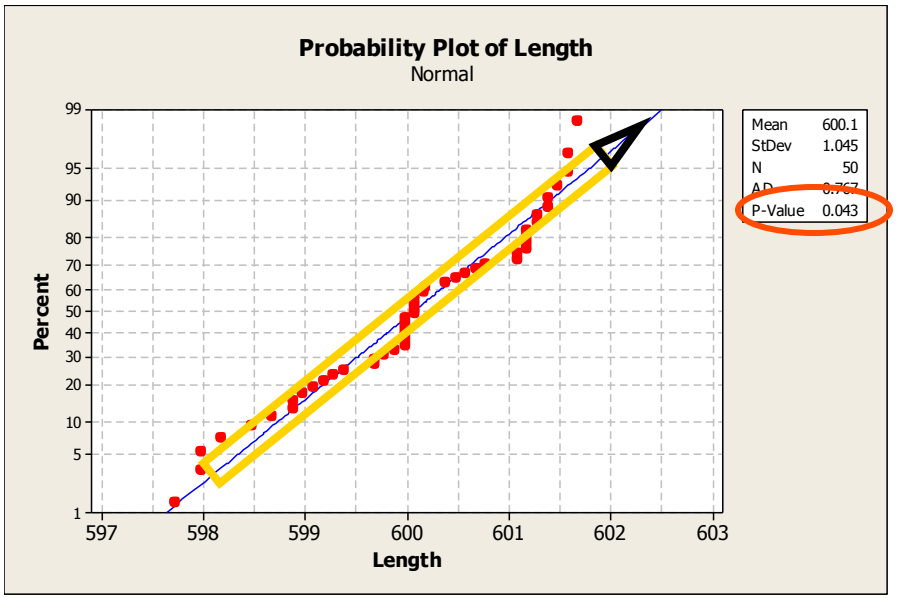


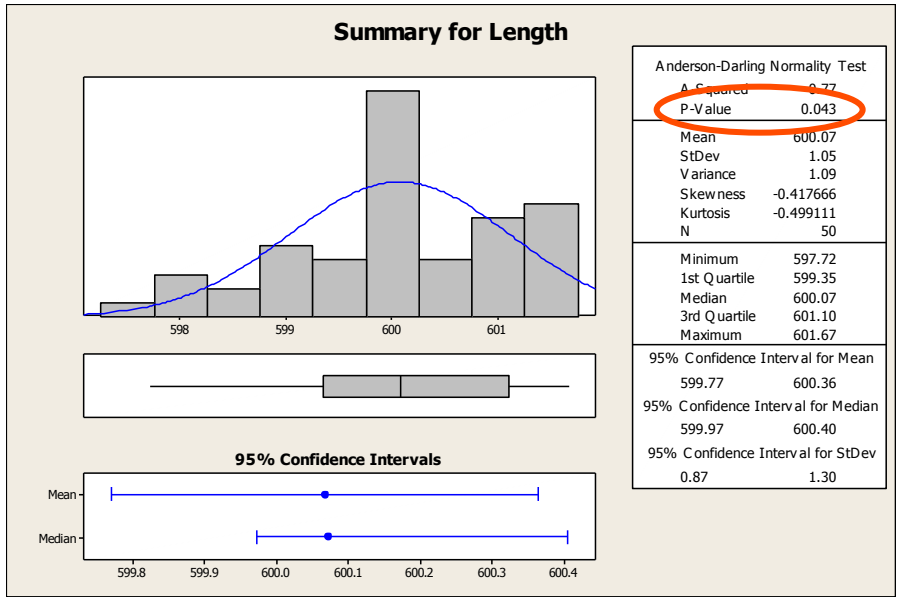
Stat>Basic Statistics>Graphical Summary

- *Make sure p-value is greater than 0.05*
- *'Fat pencil' test*

Non-Normal Distributions

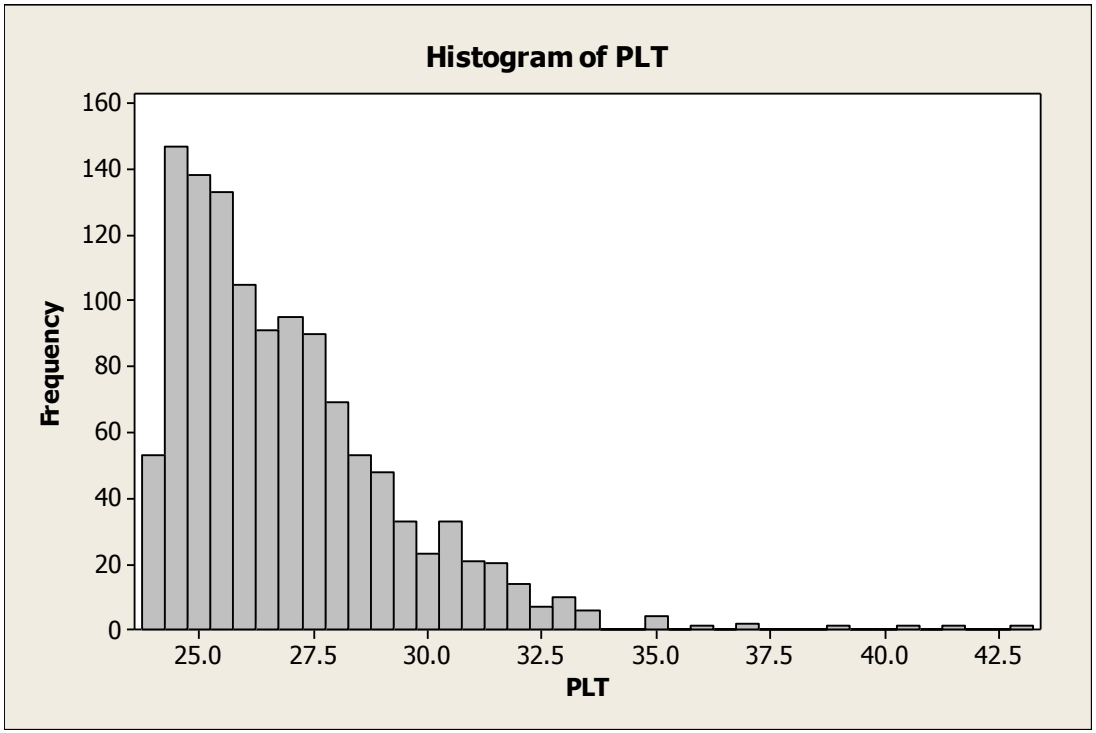
- But what if your data set is NOT Normally Distributed?
- *Example*:





Non-Normal Distributions

- Examples of phenomena not Normally Distributed
  - Process lead time
  - Flatness
  - End play
  - Concentricity
  - Taper
  - Machine efficiency
  - Leak rate
  - Contamination level
  - Life span



An Approach to Non-Normal Data

- **A general approach to analyzing Non-Normal Data:**
  1. **Validate raw data**
  2. **Stratify the data and check for outliers**
  3. **Use appropriate control chart and capability analysis**
     - **Sub group averaging vs. Individual data**
  4. **Use tools that do not require normality**
     - **Non-parametric statistics**
  5. **Transform the data**

Handling Non-Normal Data: *Validate Data*

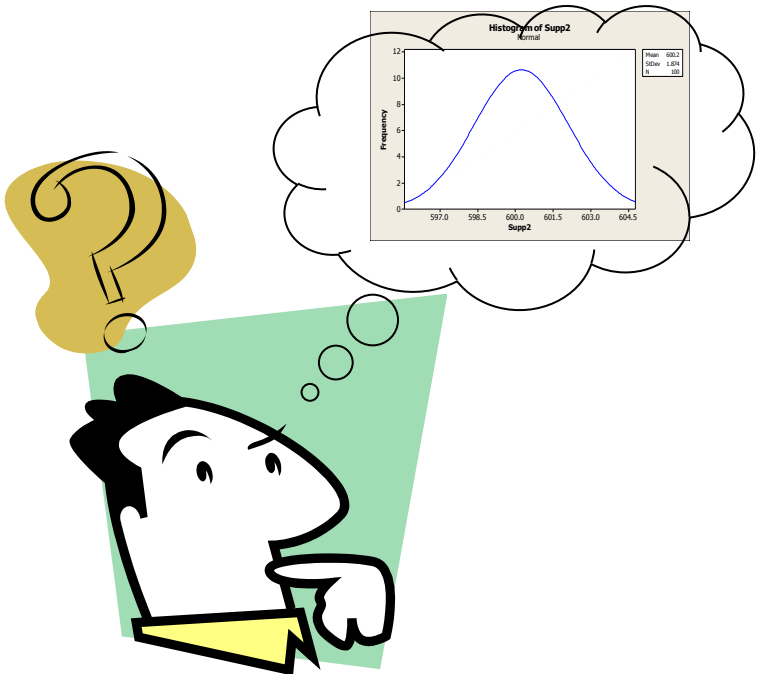
- **Step 1A: Validate the Data:** ***Basics of Measurement***
  - **Were the Operational Definitions consistent?**
  - **Were there errors in data collection OR data entry?**
  - **Has the measurement system been calibrated?**
  - **Is the measurement system valid? Has Measurement System Analysis been performed?**
    - **Gage R&R studies**
      - **Inspector methods**
      - **Tools / measuring devices**

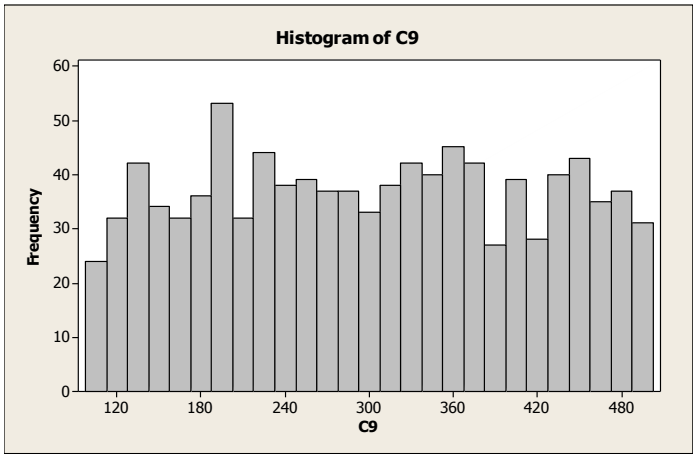
Handling Non-Normal Data: *Validate Data*

- **Step 1B: Validate the Data: *Sample Size***
  - **Is there enough data to properly characterize the distribution?**
    - **Small sample sizes are extremely sensitive to any variation and therefore are more likely to be considered non-normal**
  - **Review sample size calculation**
    - Stat > Power and Sample Size > . . .

Handling Non-Normal Data: *Data Stratification*

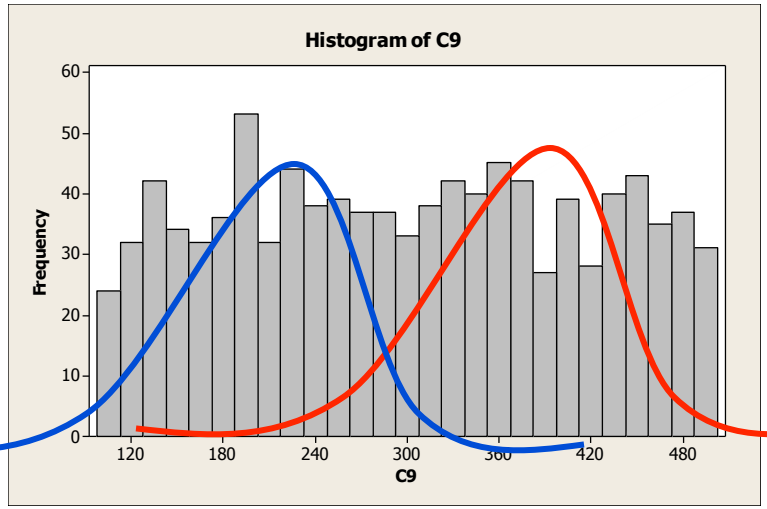
- **Step 2A: Stratify the Data**
- **Sometimes process knowledge indicates that data should be normally distributed when sample data do not appear to be normally distributed**





Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratify the Data**
  - **Datasets can often be segmented into smaller groups by stratification of data**
  - **Each segment of data can then be examined for normality**
  - **Once segmented, non-normal datasets often become normal data sets**



Handling Non-Normal Data: *Data Stratification*

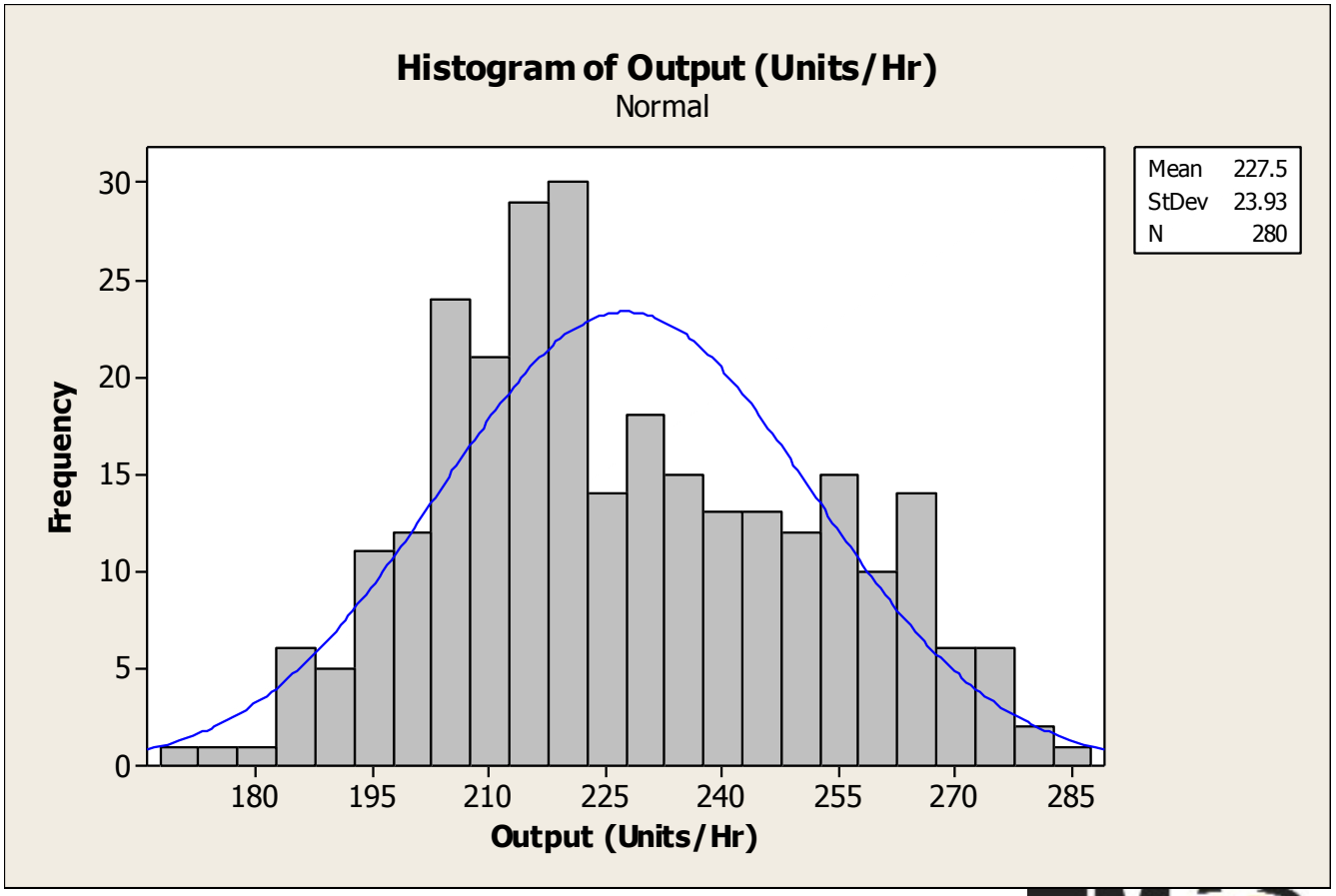
- **Step 2A: Stratify the Data**
  - **Common approaches to stratification**
    - **Time:** Day of Week, Shift, Hour of Day
    - **Location:** City, County, State, Facility
    - **Equipment:** Machine, Server, Engine
    - **Material:** Type, Freshness
    - **Employee:** Experience, Skill, Strength, Knowledge
    - **Product Difficulty/Complexity:** Easy vs. Hard
    - **Customer:** Size, Purchase Criteria

Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratification of Data**
- **Example: Consider the dataset OutputData.XLS**

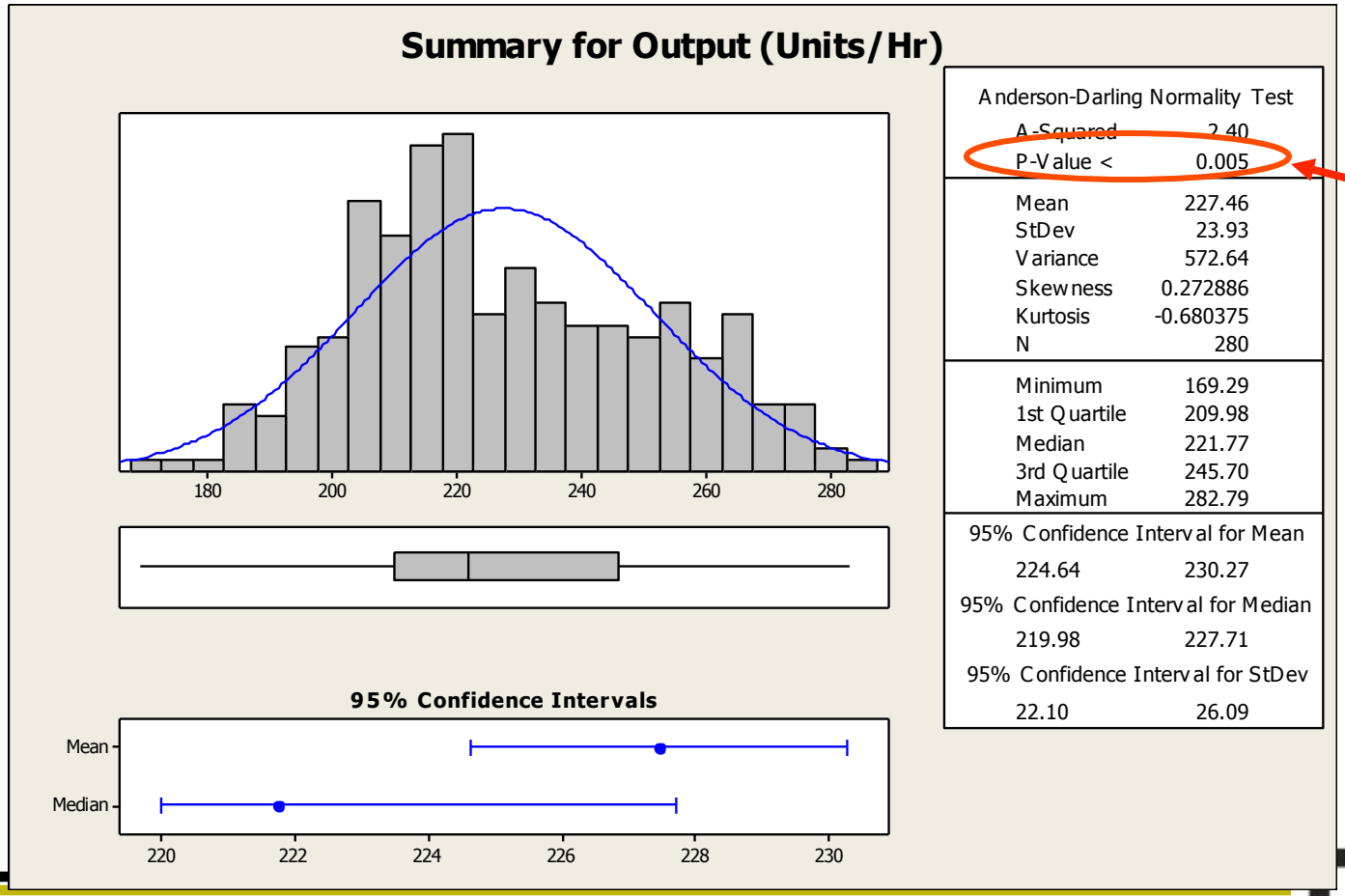
Output (Units/Hr)	Day
217.459	Mon
202.338	Tue
241.054	Wed
205.254	Thu
272.727	Fri
232.235	Sat
201.176	Mon
209.783	Tue
226.847	Wed
236.227	Thu
261.129	Fri
259.317	Sat
224.446	Mon
221.014	Tue
212.858	Wed
183.041	Thu
269.511	Fri
251.532	Sat
257.207	Mon
234.381	Tue
195.119	Wed
214.5	Thu
237.47	Fri
279.564	Sat

• • •



Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratification of Data**
- **Example: Perform a Graphical Summary . . .**

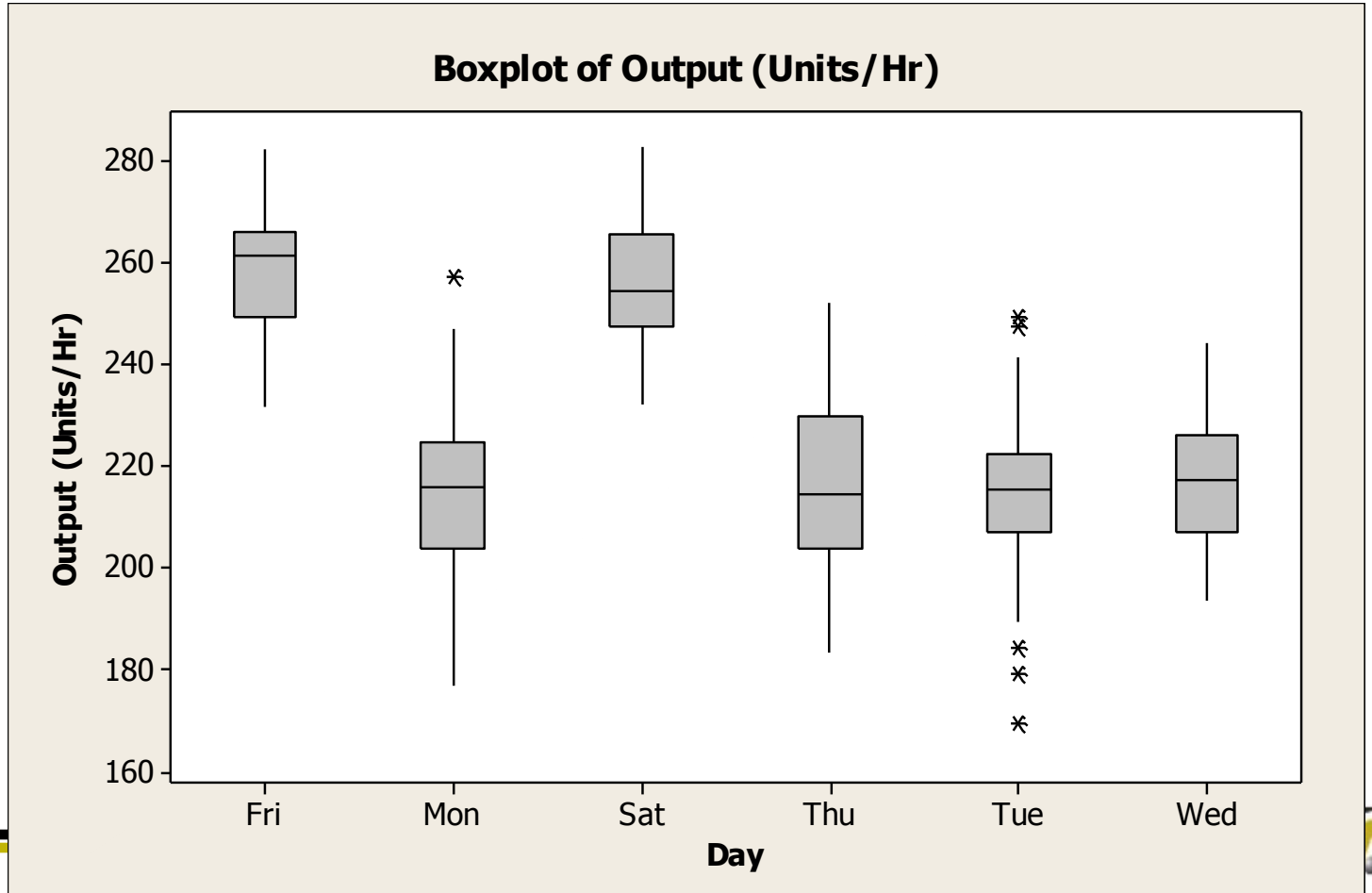


Anderson-Darling Normality Test	
A-Squared	2.40
P-Value <	0.005
Mean	227.46
StDev	23.93
Variance	572.64
Skewness	0.272886
Kurtosis	-0.680375
N	280
Minimum	169.29
1st Quartile	209.98
Median	221.77
3rd Quartile	245.70
Maximum	282.79
95% Confidence Interval for Mean	
224.64	230.27
95% Confidence Interval for Median	
219.98	227.71
95% Confidence Interval for StDev	
22.10	26.09

$p < .05$ indicating the data is not normal

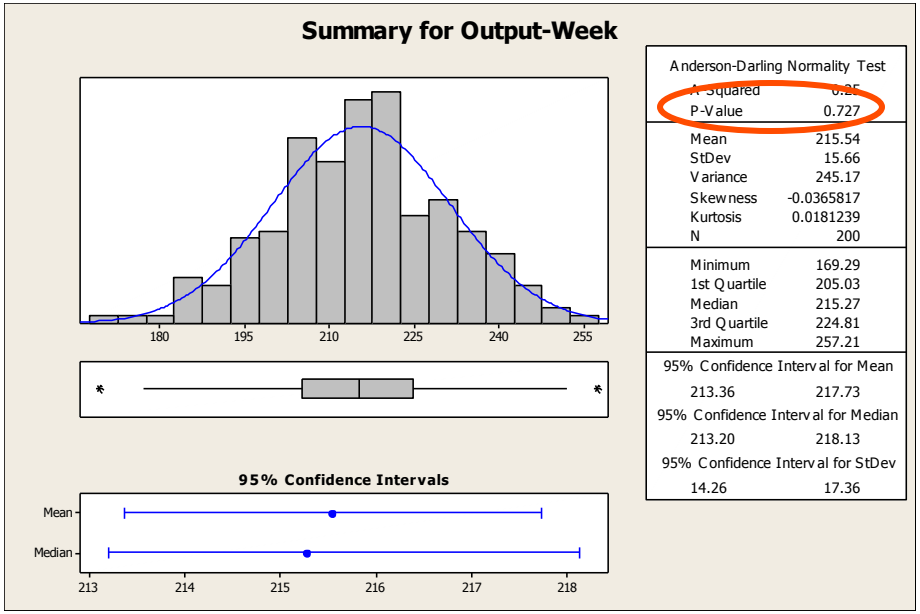
Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratification of Data**
- **<u>Example</u>: Next data is stratified graphically by Day using boxplots**

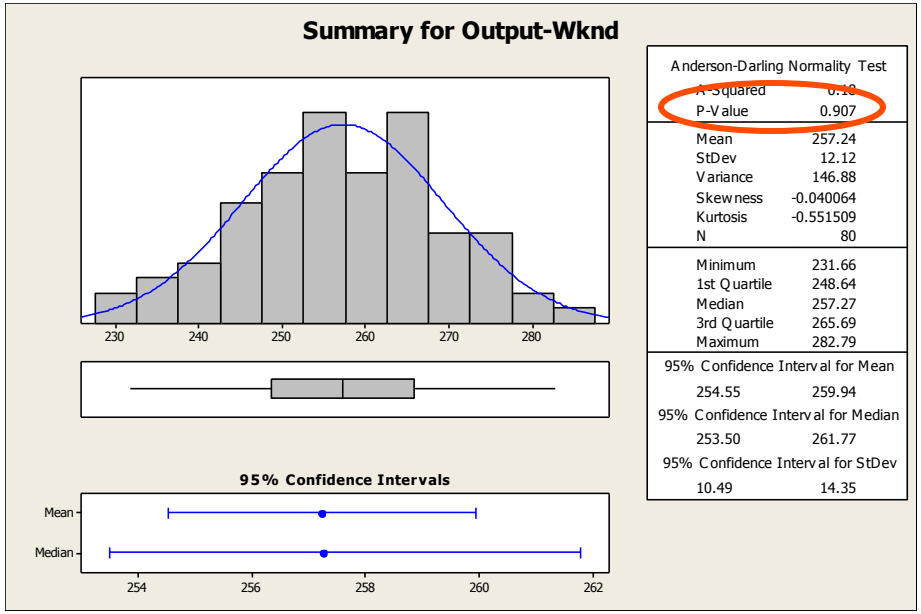


Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratification of Data**
- **<u>Example</u>: Data is then grouped into Mon-Thu and Fri-Sat**



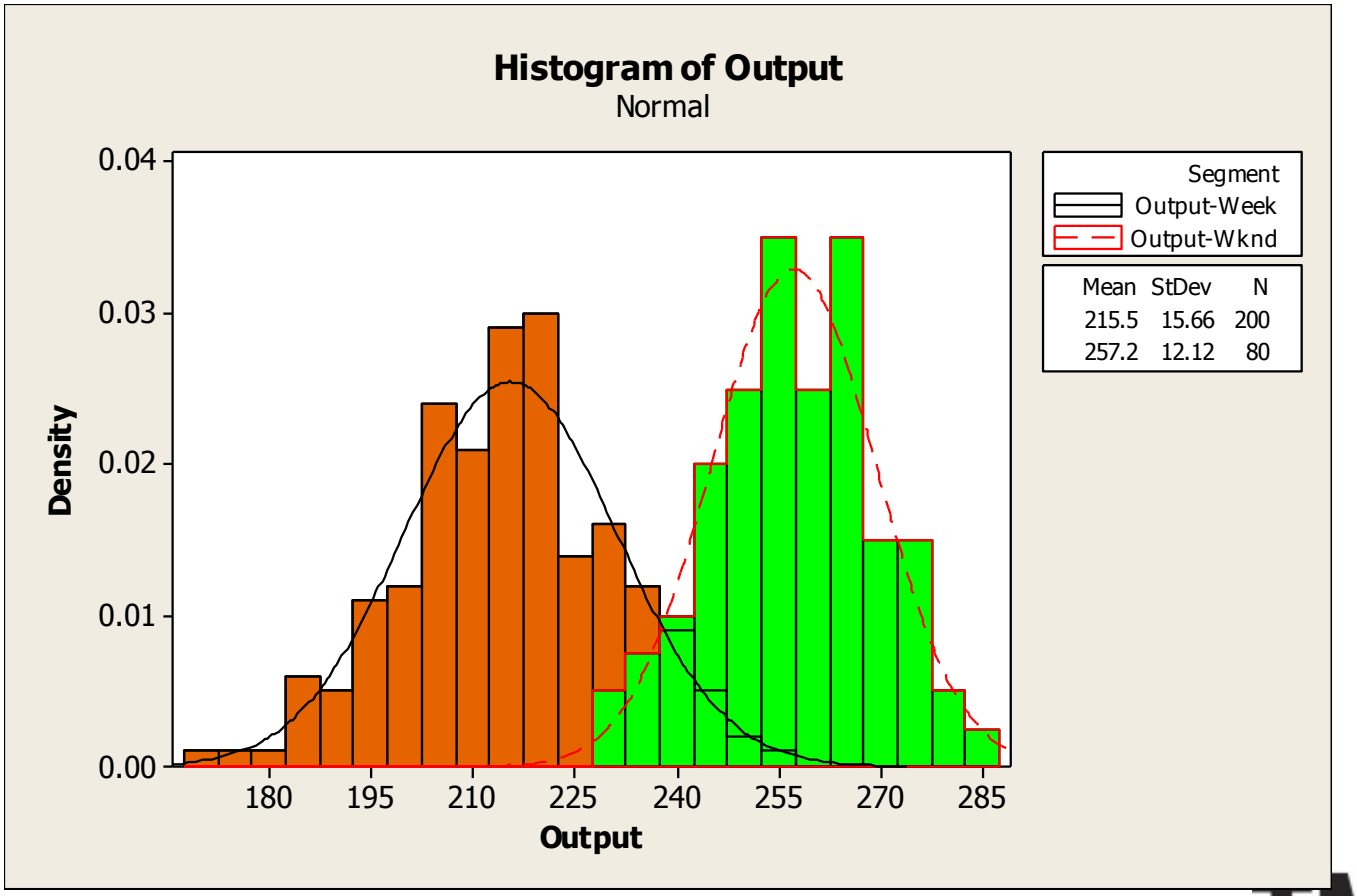
- **Data for Mon through Thu**
- **p = 0.727**



- **Data for Fri and Sat**
- **p = 0.907**

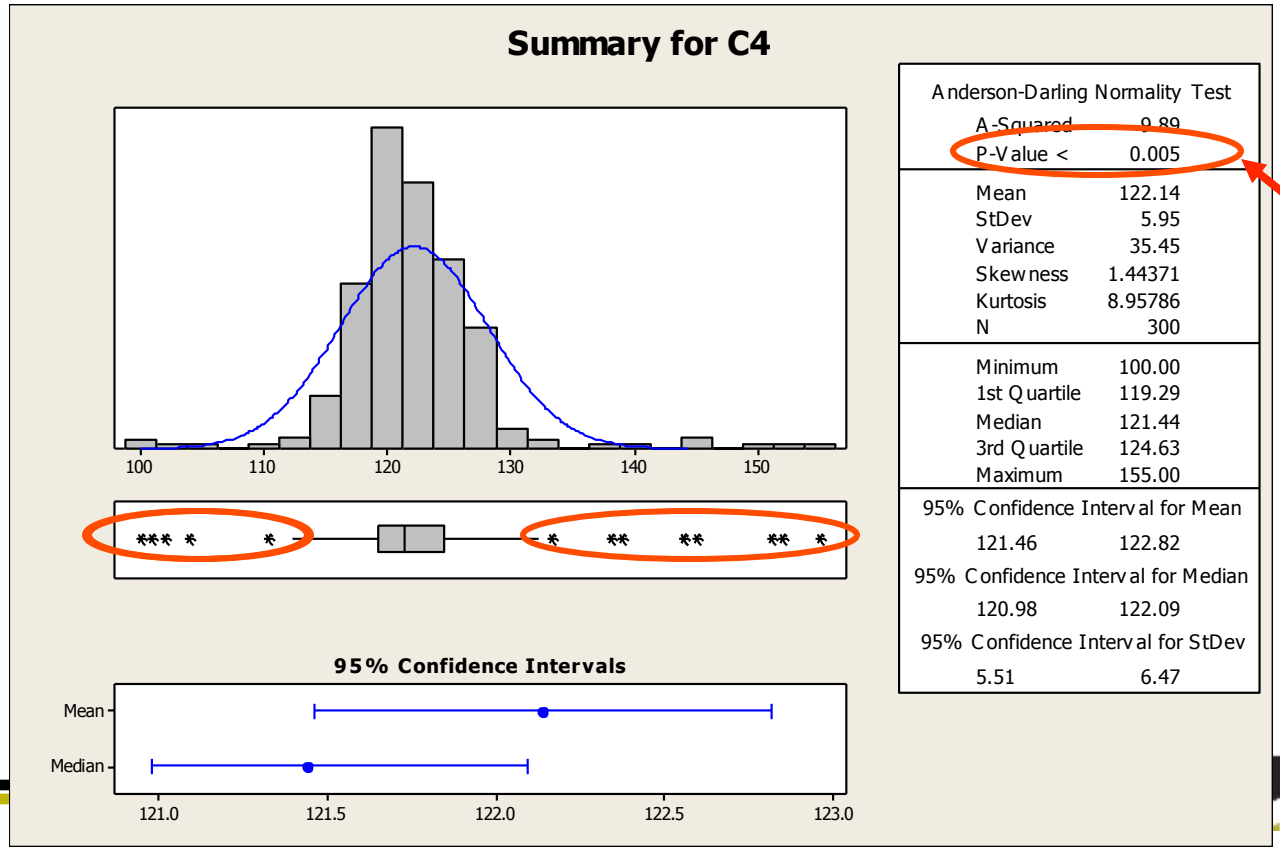
Handling Non-Normal Data: *Data Stratification*

- **Step 2A: Stratification of Data**
- **<u>Example</u>: Data is then grouped with Mon-Thu and Fri-Sat**



Handling Non-Normal Data: *Outliers*

- **Step 2B: Removal of Outliers**
- **Outliers may affect normality of a dataset**
- **Example:**



p < .05 indicating the data is not normal

Handling Non-Normal Data: *Outliers*

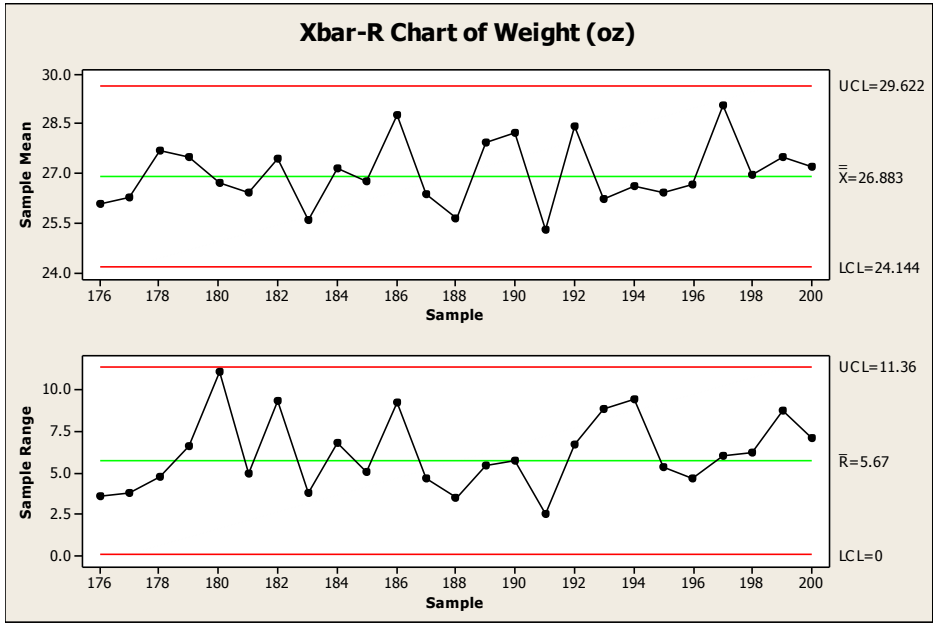
- **Step 2B: Removal of Outliers**
- **Some guidelines on outliers**
  - **All outliers are signals of a lack of homogeneity**
  - **Simply removing outliers from computations in order to improve statistical computations will not fix the underlying reality that the process is being operated unpredictably**
  - **Data manipulation may result in clearer inferences, but it divorces those inferences from the process generating the data**
    - ***The Six Sigma Practitioner's Guide to Data Analysis* by Wheeler**

Handling Non-Normal Data: *Outliers*

- **Step 2B: Removal of Outliers**

  - **Outliers are indicators of Special Cause Variation, and often result from external influences.**
  - **They can be safely removed from the dataset IF**
    1. **The cause is clearly understood,**
    2. **It is not likely to repeat itself, AND**
    3. **The number of outliers is two or less**
  - **If you have many outliers, your process may be in a "state of chaos". In this case, outliers are a reoccurring part of the existing process and may not be removed from the dataset.**
    - ***Adapted from LSS Black Belt by George Group***

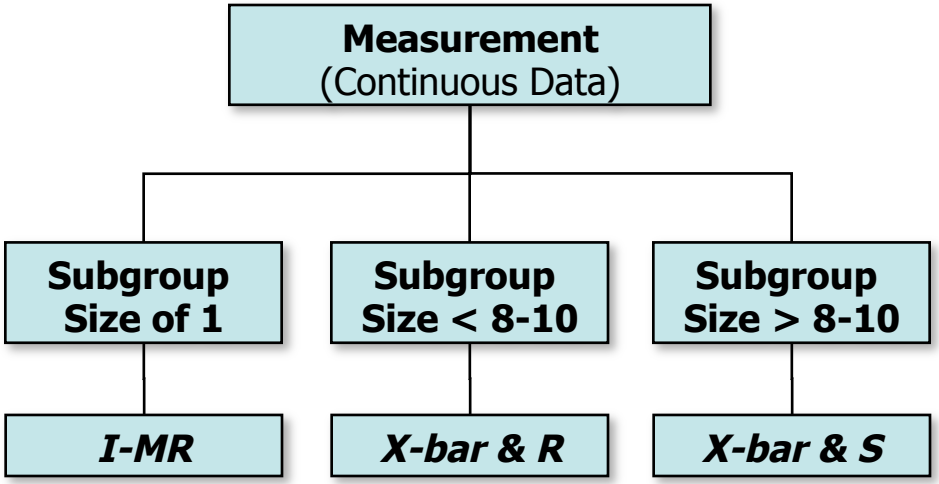
Handling Non-Normal Data: *Control Chart Considerations*



- **Step 3A: Use Appropriate Control Chart**
  - **Control Charts are used to monitor many types of processes**
  - **The theory behind control charts for continuous data is generally thought to be based on an assumption of normally distributed data**
  - **Many statisticians emphasize the importance of confirming that the underlying distribution is normal before using such charts**

Which Control Chart to Use?

- **Step 3A: Use Appropriate Control Chart**



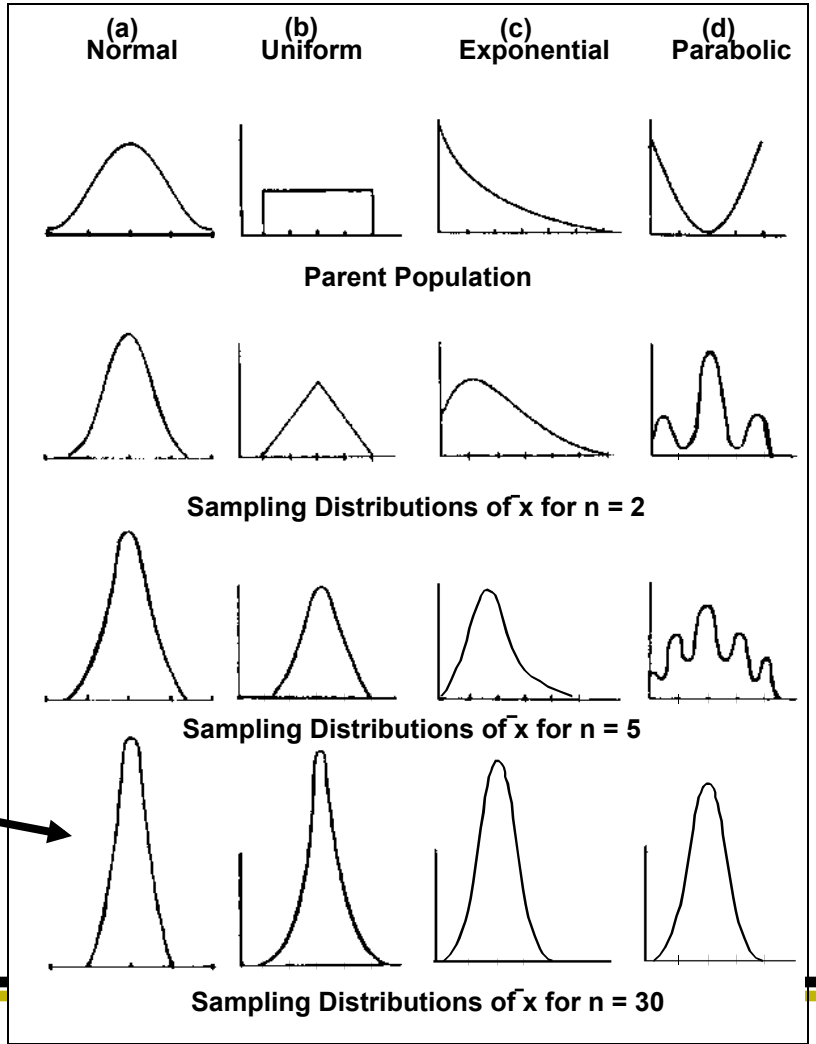
- **For continuous data, the recommended chart is dependent on size of the subgroup**
  - **An I-MR chart (also called X-MR) is often used when there isn't a natural subgroup, or when beginning to monitor a new process**
  - **A common subgroup size for X-bar & R charts is 5**

Handling Non-Normal Data: *Control Chart Considerations*

- **Step 3A: Use Appropriate Control Chart**
- **But what if the data isn't normal?**
  - **With X-Bar R or X-Bar S, there is Subgroup Averaging of Data**
  - **The distribution of subgroup means usually is normal, *even if parent distribution is NOT normal***
  - **Based on Central Limit Theorem**
  - **Remember: The more skewed the data, the larger the sample needed**
    - ***NOTE: The underlying (i.e., parent) population is STILL not normal***

Handling Non-Normal Data: *Control Chart Considerations*

- **Step 3: Use Appropriate Control Chart**



Central Limit Theorem at work

Increasing Sample Size

Handling Non-Normal Data: *Control Chart Considerations*

- **Step 3A: Use Appropriate Control Chart**
- **What about an I-mR chart?**
  - **With individual data, the requirement of normality would seem to be more important because there is no subgroup averaging**
- **But is normality really required?**

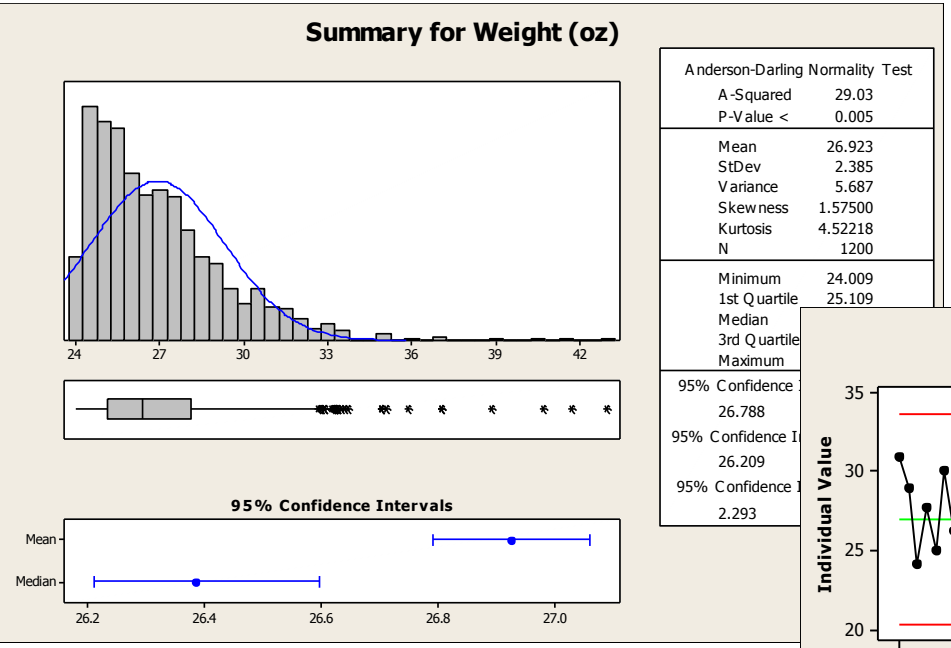
Three-sigma limits are not tied to any particular probability. . . Regardless of the shape of the distribution, they filter out virtually all of the probable noise so that points outside the limits are very likely to be signals of exceptional variation . . .

So while normal distributions happen, the normality of the data is neither a prerequisite for a process behavior chart, nor is it an inevitable consequence of a predictable process.

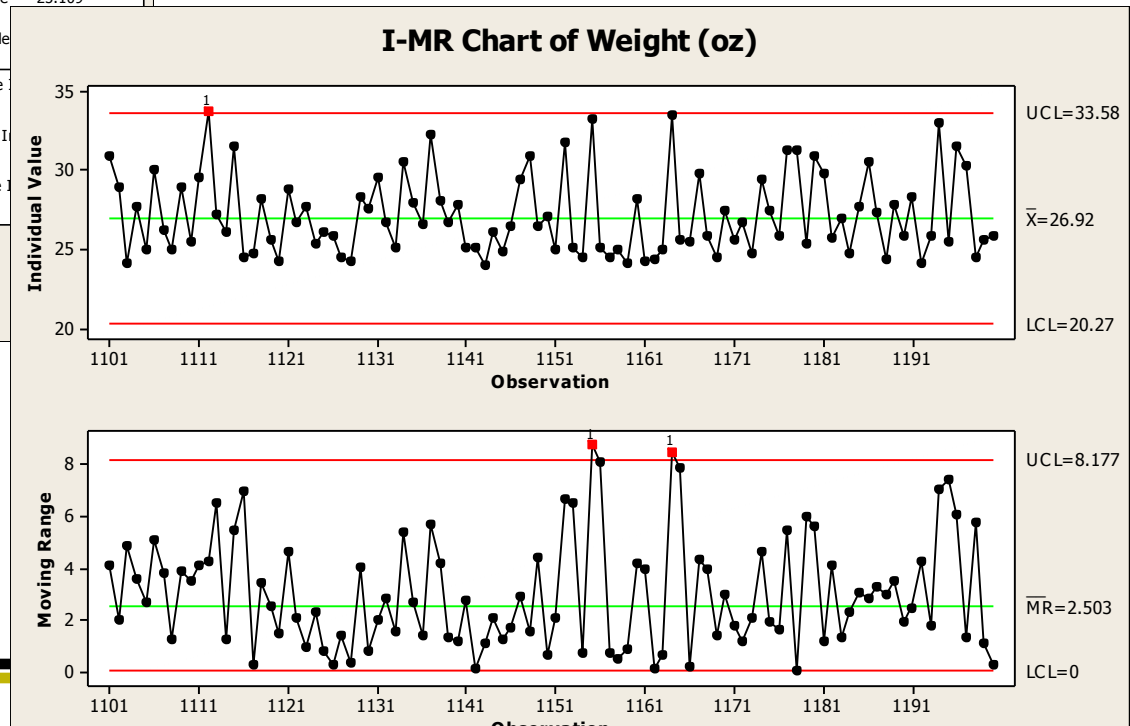
- *Normality and the Process Behavior Chart* by Wheeler

Handling Non-Normal Data: *Control Chart Considerations*

- **Step 3A: Use Appropriate Control Chart**
- **Example: I-mR chart with non-normal data**

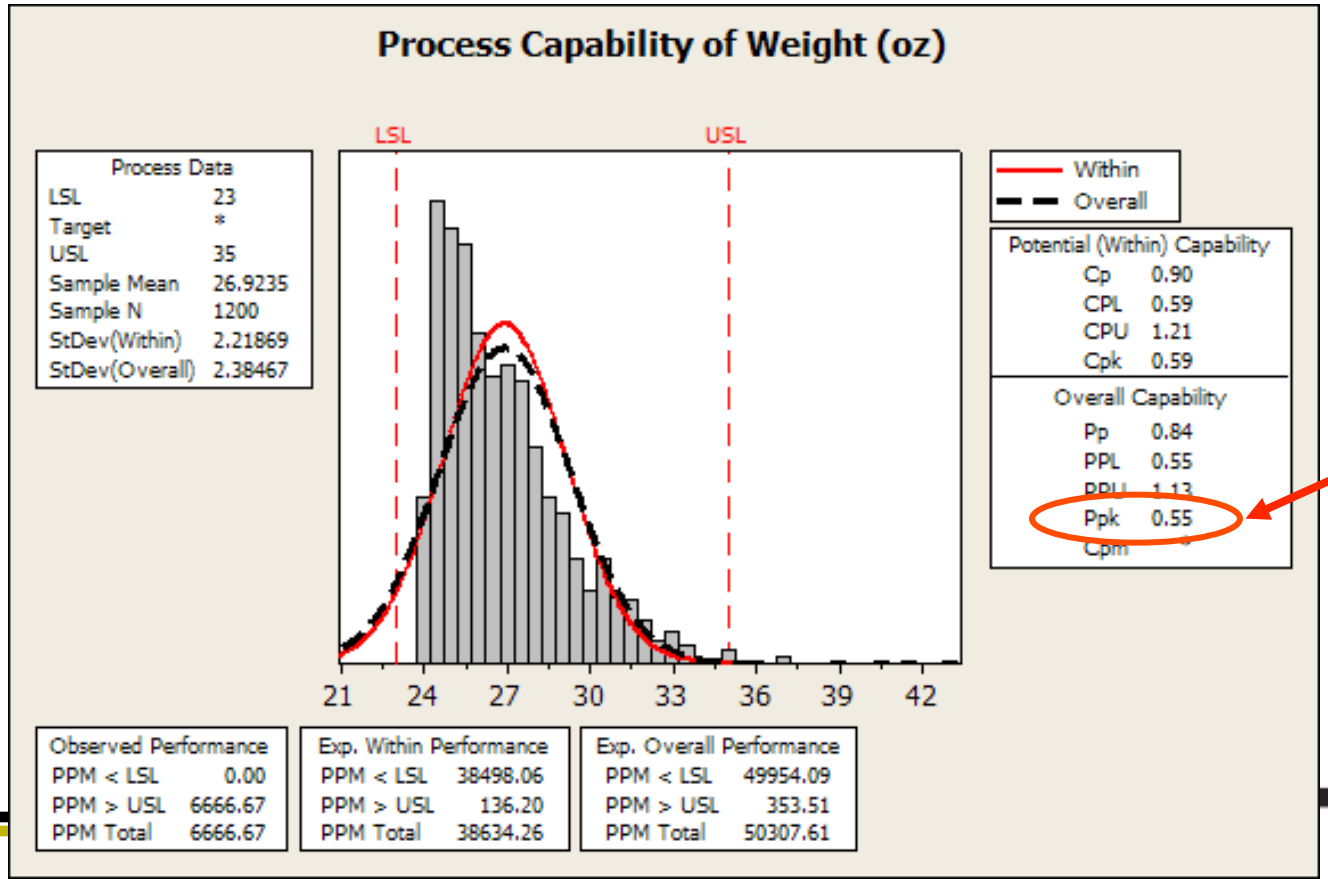


Anderson-Darling Normality Test	
A-Squared	29.03
P-Value <	0.005
Mean	26.923
StDev	2.385
Variance	5.687
Skewness	1.57500
Kurtosis	4.52218
N	1200
Minimum	24.009
1st Quartile	25.109
Median	
3rd Quartile	
Maximum	
95% Confidence	
26.788	
95% Confidence I	
26.209	
95% Confidence I	
2.293	



Handling Non-Normal Data: *Capability Considerations*

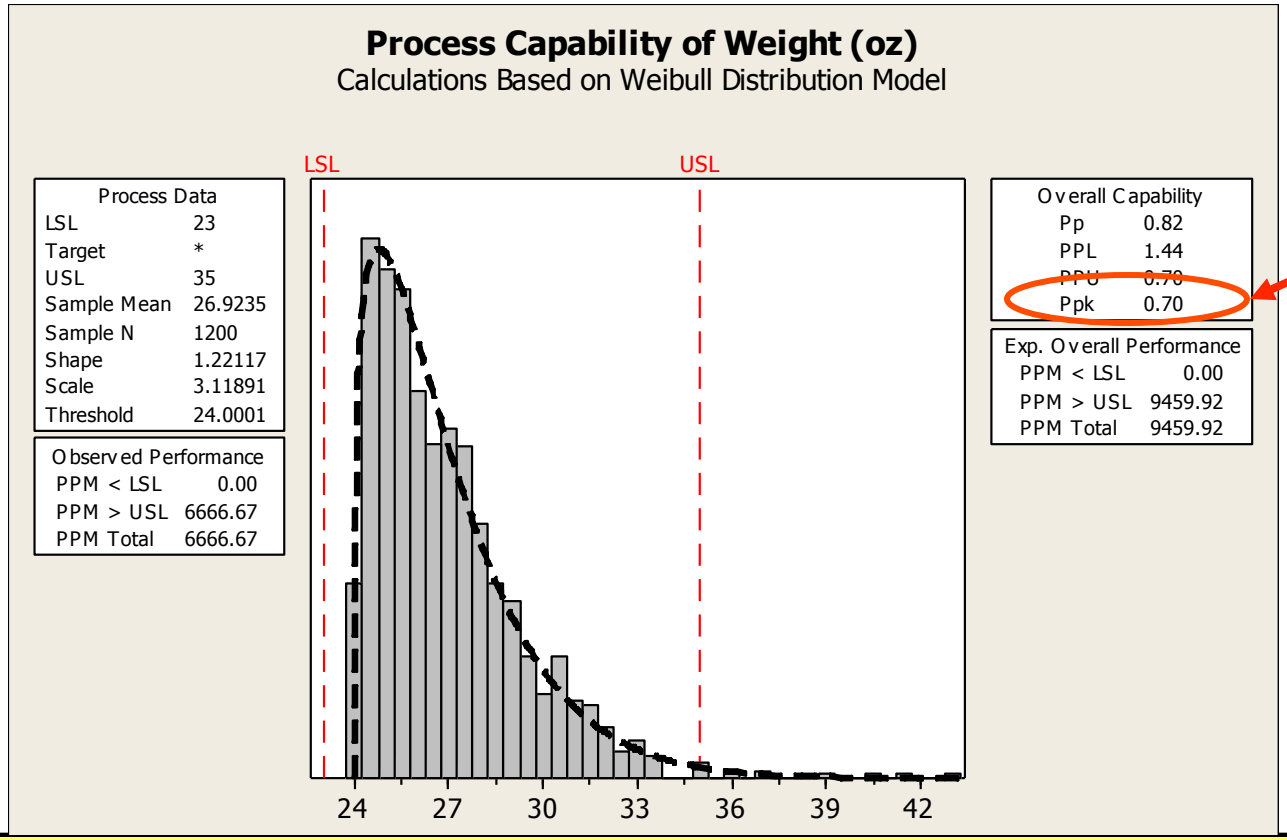
- **Step 3B: Use Appropriate Capability Chart**
- **Minitab can be used to calculate the process capability of a set of data assuming the data *is* normally distributed**
  - Stat > Quality Tools > Capability Analysis > Normal



Process Capability assuming data are normal

Handling Non-Normal Data: *Capability Considerations*

- **Step 3B: Use Appropriate Capability Chart**
- **Capability of the same data set as calculated using a non-normal distribution (in this case, the Weibull Distribution)**
  - Stat > Quality Tools > Capability Analysis > NonNormal



Process Capability assuming data are *not* normal

NOTE: There are a total of 13 different distributions that can be chosen for the non-normal capability analysis

Handling Non-Normal Data: *Nonparametric Statistics*

- **Step 4: Use nonparametric statistics**
  - **These statistical tools do *not* require normality**
  - **The specific statistical tool used depends on the question to be answered, type of data, etc.**
  - **Each tool has one or more assumptions that must be confirmed**
  - **Comparison of Medians instead of Means**
  - **Minitab Help provides detailed information on requirements of each tool**
    - **Stat > Nonparametrics**
  - **For in-depth understanding:**
    - ***Practical Nonparametric Statistics* by Conover**

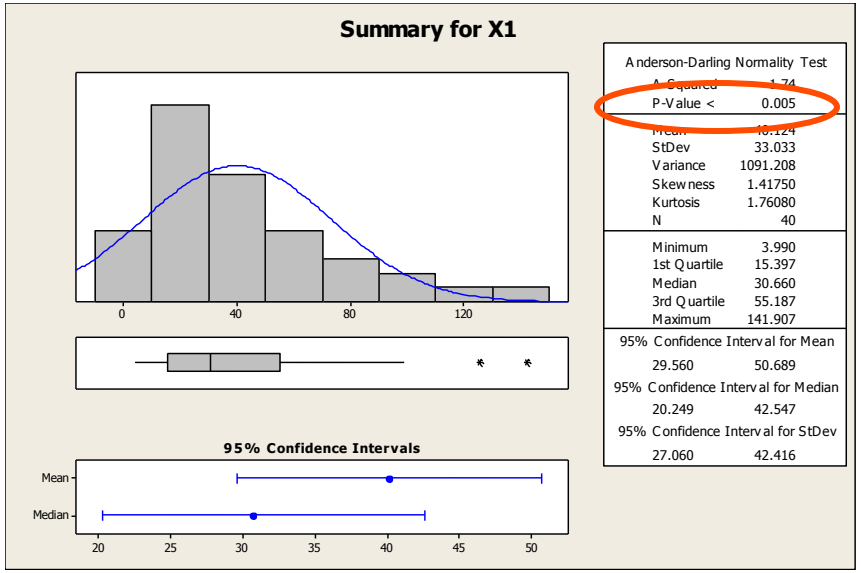
Handling Non-Normal Data: *NonParametric Tools*

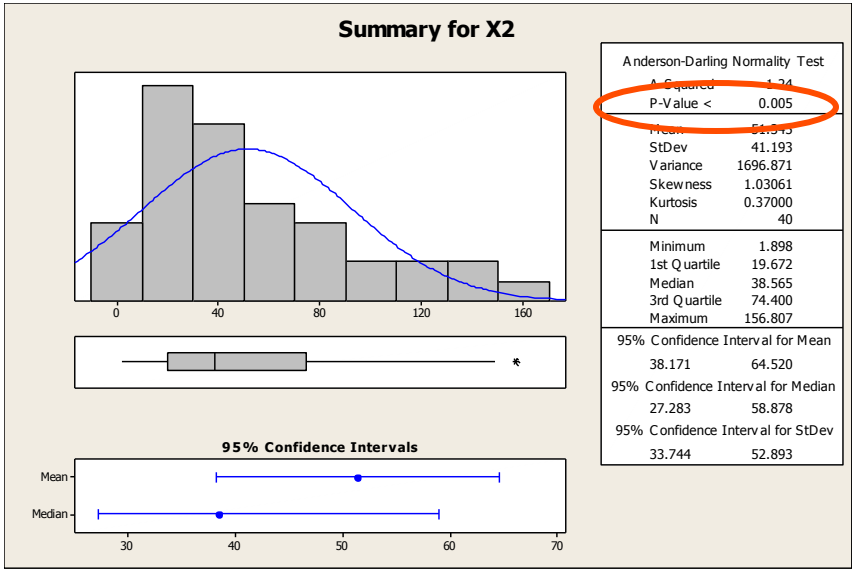
Common statistical tests for normal and nonparametric data

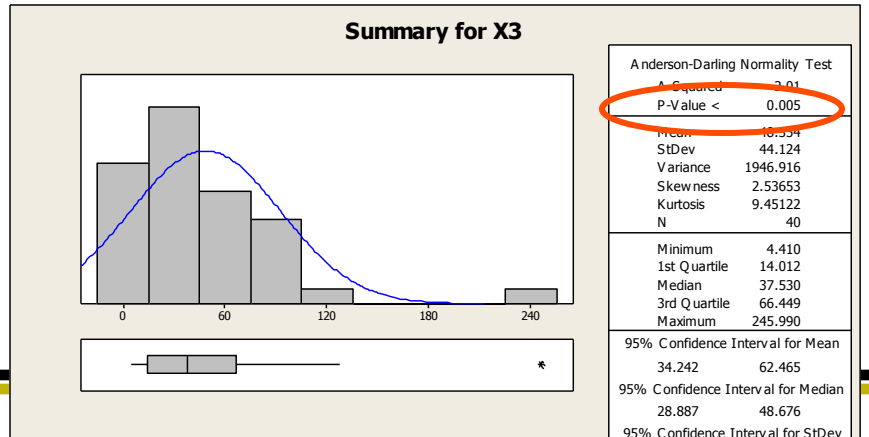
Hyp. Test Assumes Normality	Normality *not* a Required Assumption	Required Assumptions
One sample Z-test	One sample Sign test	Ordinal data
One sample t-test	One sample Wilcoxon test	Continuous & symmetric
Two sample t-test	Mann – Whitney test	Continuous or ordinal, same shape, equal variances
One-way ANOVA	Kruskal – Wallis OR Mood's Median test	Same shape for each subgroup
Randomized Block (2-Way ANOVA)	The Friedman Test	See Minitab Help

Handling Non-Normal Data: *NonParametric Tools*

- **Example: Nonparametrics**
  - **Consider a process with three machines in parallel: X1, X2, and X3**



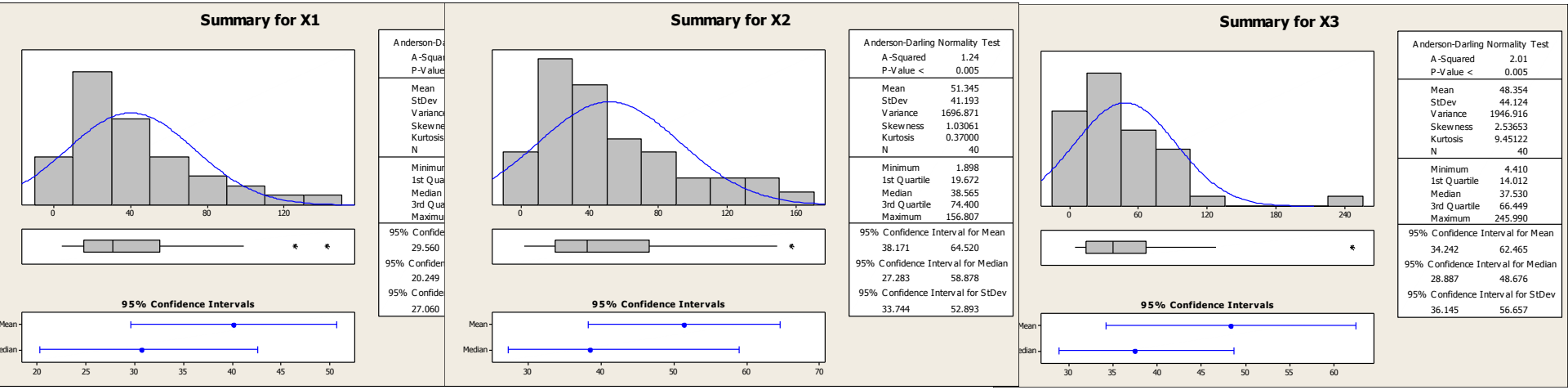




The data are not normally distributed as indicated by the p value of less than .05 for each machine

Handling Non-Normal Data: *NonParametric Tools*

- **Example: Nonparametrics**
  - **The question for a Black Belt: *Are these three machines the same?***
  - **If the data were normally distributed, ANOVA could be used to answer this question**
    - **Null Hypothesis:** $H_0 : \mu_{X1} = \mu_{X2} = \mu_{X3}$
    - **Alternative Hypothesis:** H_a **: At least one** $\mu_i \neq \mu_j$



  - **The Mood's Median Test may be used if the data are continuous, and the datasets being compared have the same shape**

Handling Non-Normal Data: *NonParametric Tools*

- **Example: Nonparametrics – Run the Mood Median Test**
  - **A p value > .05 indicates the null hypothesis cannot be rejected (i.e., that the *median* of the three datasets is the same)**

Mood Median Test: Measure versus Machine

```
Mood median test for Measure

Chi-Square = 1.40    DF = 2    P = 0.497

                                        Individual 95.0% CIs
Machine  N<=  N>  Median  Q3-Q1  ---+---------+---------+---------+---
X1         23  17    30.7   39.8  (--------*--------)
X2         18  22    38.6   54.7         (--------*----------------)
X3         19  21    37.5   52.4          (------*---------)
                                  ---+---------+---------+---------+---
                                    24        36        48        60

Overall median = 35.4
```

Handling Non-Normal Data: *Transforming Data*

- **Step 5: Transforming Data**
- **A transformation of raw data occurs when a mathematical function is applied to the raw data**
  - **Transformations based on: x^2, x^3, $x^{1/2}$, log(x), ln(x), etc.**
  - **Minitab includes Box-Cox & Johnson transformations for continuous data**
    - **Stat>Control Chart>Box-Cox Transformation**
    - **Stat>Quality Tools>Johnson Transformation**
- **The goal of transformation is to make the transformed data normally distributed, which will allow the use of statistical tools requiring normality**
  - ***NOTE: Transformations do not always work***

Recommendation: Only use transformations as a last resort.

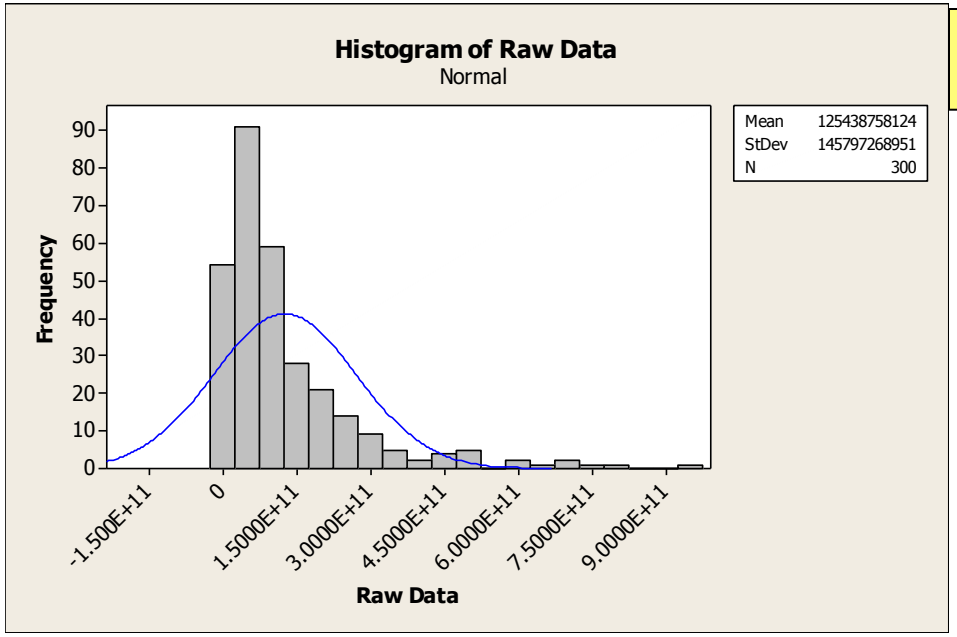
Rules for Transformation of Raw Data

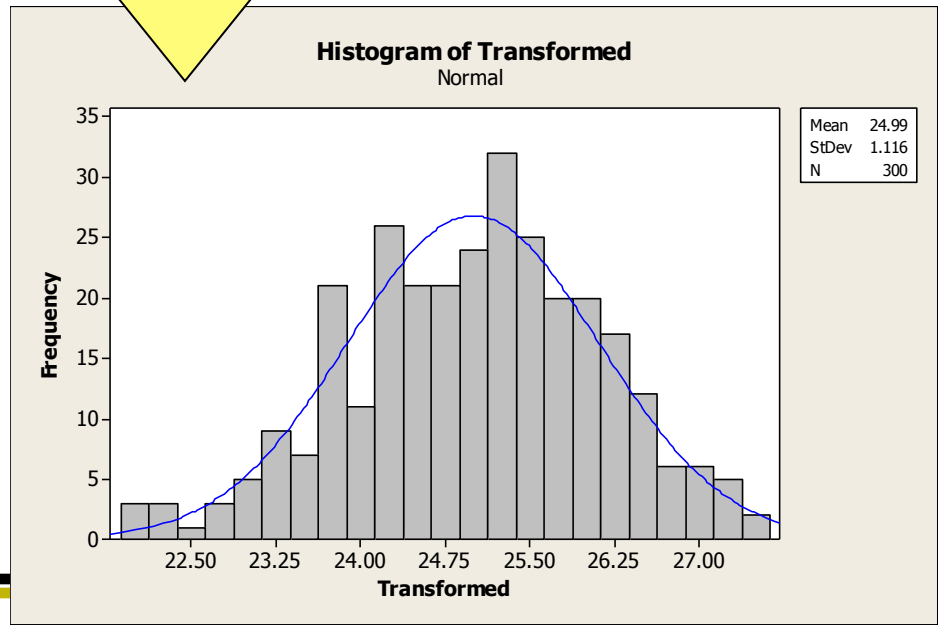
- **General Transform Rules**
  1. **The transform must preserve the order of the data.**
  2. **The transform must be a smooth and continuous function.**
  3. **The transform should, if possible, yield interpretable results.**
  4. **The transform is often useful when the ratio of the largest to smallest value is greater than two.**
  5. **All external reference points (for example: spec limits) should get the exact same transform as the data.**

Handling Non-Normal Data: *Transforming Data*

- **Step 5: Transforming Data**

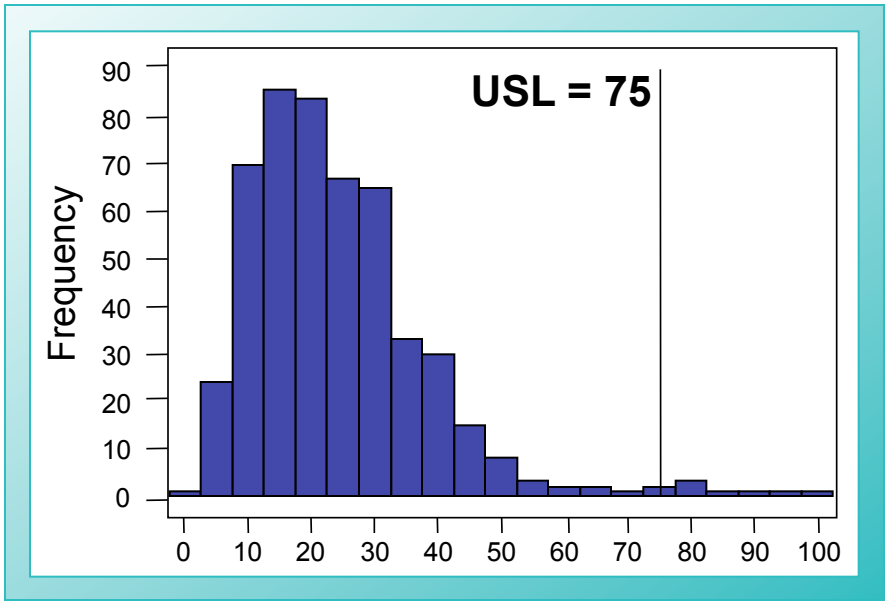
Example



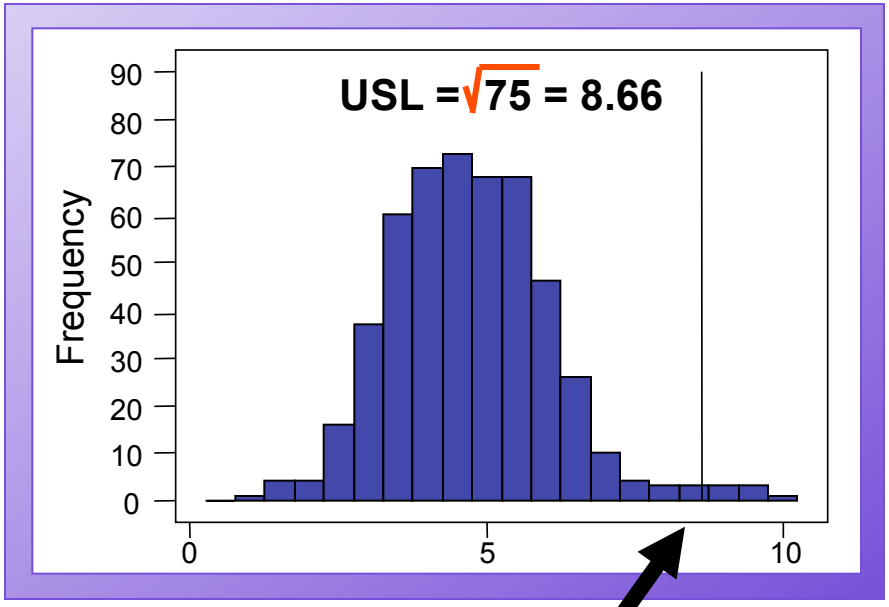


Handling Non-Normal Data: *Transforming Data*

Raw Data (Before Transform)



Square Root Transformed Data



Transformed specification limit

Takeaways

- There are many approaches to non-normal data
- Always check the data to verify that it is truly non-normal
- Approaches to handle non-normal data include:
  - Basics of Data Collection
  - Stratification of data and removal of outliers
  - Control Charts & Nonnormal Capability Analysis
  - Use of nonparametric statistics
  - Transformation using math functions into normal distributions

For More Information

www.texasleansixsigma.com

www.tmac.org

Russ Aikman (817) 307-0400
***email*: raikman@uta.edu**

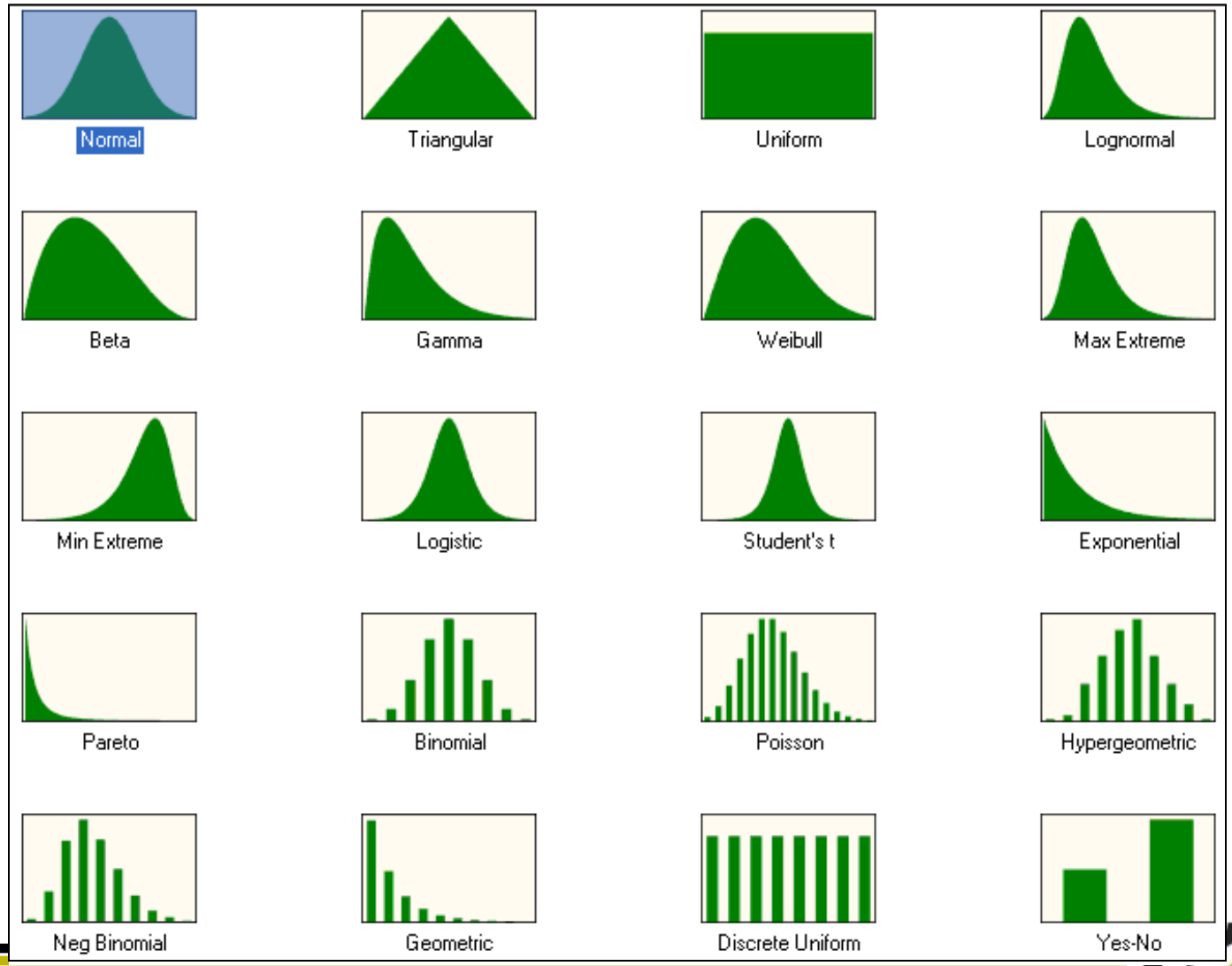
Mark Sessumes (817) 312-5853
***email*: sessumes@arri.uta.edu**

Alberto Yanez (817) 602-4005
***email:* albertoy@arri.uta.edu**

A P P E N D I X

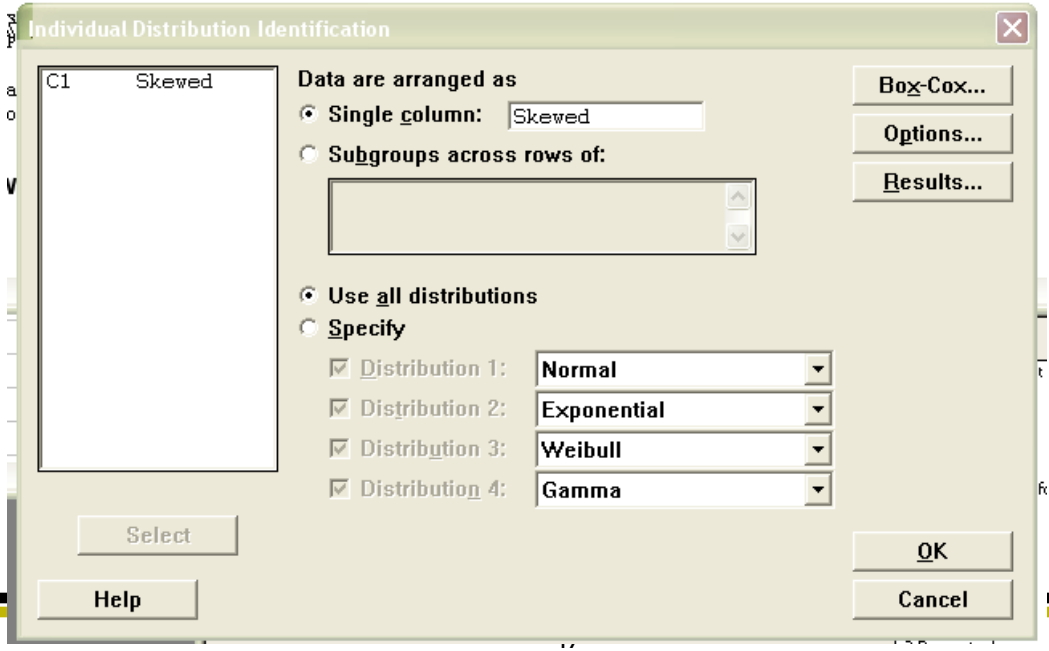
- **Determining the distribution type**

Some Non-Normal Distributions



Handling Non-Normal Data: *Determining the Distribution*

- **A common question from new Black Belts:**
  - ***If the data is not from a normal distribution, what is the appropriate distribution?***
  - **Minitab has an Individual Distribution Identification tool which can be used to determine the dataset's distribution type**
    - Stat > Quality Tools > Individual Distribution Identification



Handling Non-Normal Data: *Determining the Distribution*

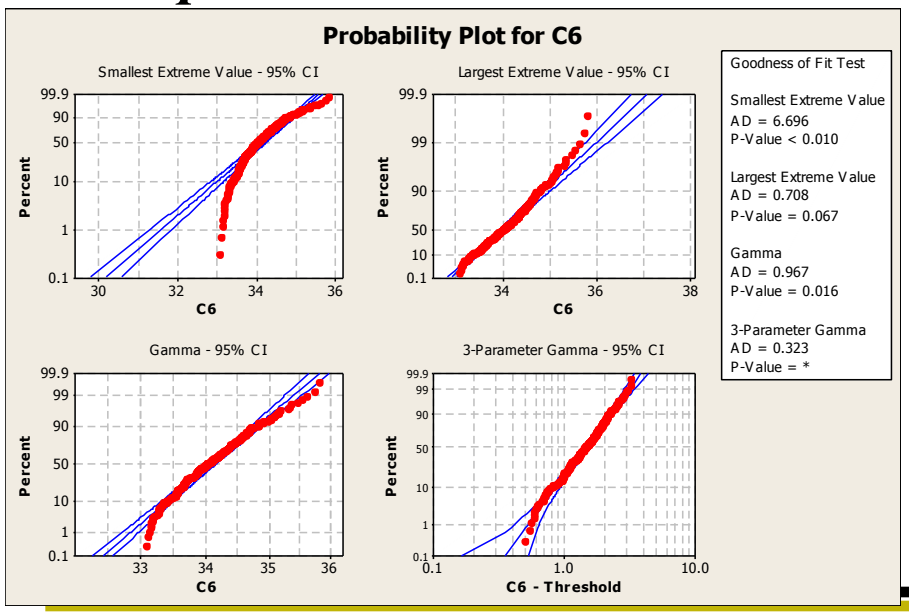
- **The Individual Distribution Identification tool fits different distributions to a set of data in an effort to find the optimal distribution for that data, based on:**
  - **Probability plots and**
  - **Goodness-of-fit tests**
- **The 13 Distributions Analyzed by the IDI:**

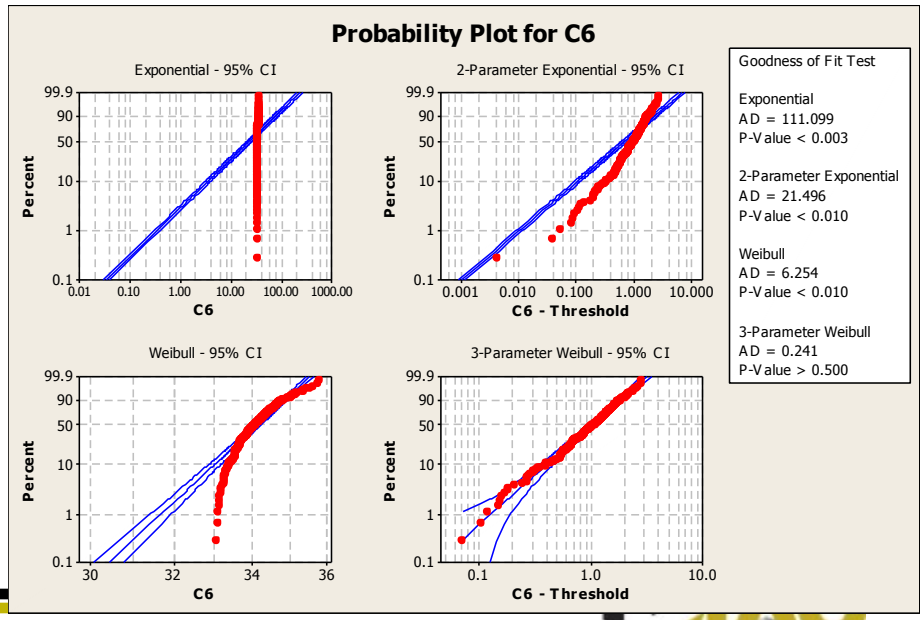
· normal	· largest extreme value
· lognormal	· smallest extreme value
· 3-parameter lognormal	· gamma
· exponential	· 3-parameter gamma
· 2-parameter exponential	· logistic
· Weibull	· loglogistic
· 3-parameter Weibull	· 3-parameter loglogistic

Handling Non-Normal Data: *Determining the Distribution*

- **Q: How to know which distribution best fits the dataset?**
- **A: After using the Individual Distribution Identification tool, a number of probability plots are created.**
  - ***Pick the distribution with the highest p-value***

Example





Handling Non-Normal Data: *Determining the Distribution*

Example

Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	1.050	0.009	
Box-Cox Transformation	0.469	0.246	
Lognormal	0.922	0.019	
3-Parameter Lognormal	0.317	*	0.000
Exponential	111.099	<0.003	
2-Parameter Exponential	21.496	<0.010	0.000
Weibull	6.254	<0.010	
3-Parameter Weibull	0.241	>0.500	0.000
Smallest Extreme Value	6.696	<0.010	
Largest Extreme Value	0.708	0.067	
Gamma	0.967	0.016	
3-Parameter Gamma	0.323	*	0.000
Logistic	0.982	0.006	
Loglogistic	0.919	0.009	
3-Parameter Loglogistic	0.631	*	0.004
Johnson Transformation	0.204	0.875	