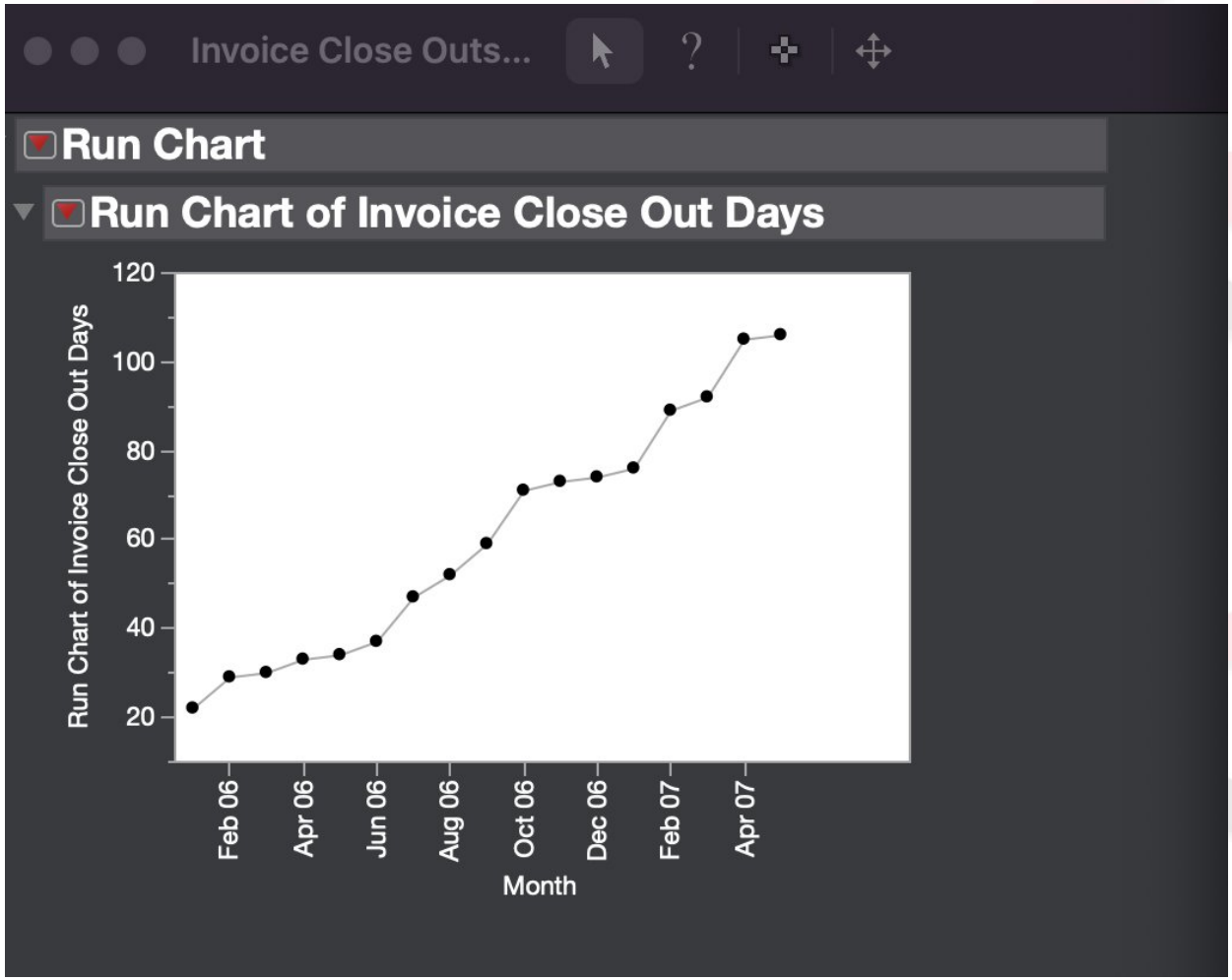
Run Charts

- Run Charts
  1. Open **Invoice Close Outs.jmp**.
  2. Select **Analyze → Quality and Process → Legacy Control Charts → Runs Chart**.
  3. Select **Invoice Close Out Days → Process**.
  4. Select **OK**.

Run Chart


Invoice Close Outs...
Run Chart
Run Chart of Invoice Close Out Days
Run Chart of Invoice Close Out Days
120
100
80
60
40
20
Feb 06
Apr 06
Jun 06
Aug 06
Oct 06
Dec 06
Feb 07
Apr 07
Month

Lean Six Sigma DMAIC Process

DEFINE	MEASURE	ANALYZE	IMPROVE	CONTROL
What problem are we addressing?	What data is needed and what is the current performance?	What are the root causes of the problem?	What is the best solution to remove each root cause?	How can we insure the gains are maintained?
✓Potential projects evaluated and selected ✓Project charter completed ✓VOC collected and analyzed ✓Process mapped	✓Data collection plan created ✓Data collection completed ✓Process baseline established	❑Potential root causes identified ❑Analysis of data completed	❑Brainstorm potential solutions ❑Pilot solutions ❑Optimize process outputs ❑Document solution implementation plan	❑Select appropriate controls ❑Document control plan ❑Deliver project documentation ❑Celebrate completed project

Measure Tollgate

- Determine project critical Ys and Xs
- Create the data collection plan
- Complete a measurement systems analysis
- Collect data
- Calculate process capability and baseline
- Update charter

INTRODUCTION TO ANALYZE

Objectives

- Review the typical flow through the Analyze phase of a project
- Understand the role of process analysis and root cause analysis in determining root causes

Overview of Analyze Phase

Brainstorm to create a list of potential root causes.

Examine data with graphical tools

Review Process Flow Maps

Create value stream maps

Use 5 Whys to List Potential Root Causes

Verify Root Causes

Prioritize Root Causes

Analyze

$$Y = f(X)$$

- During the analyze phase of a project, the team's objective is to determine which of the potential Xs under investigation are critical to the outcome of the process.
- The two phases of analyze are:
  - Process Analysis
  - Data Analysis

Analyze

Data Analysis

- Using data collected to find patterns, trends, and other differences that can suggest, support, or reject theories about the causes of defects.

Process Analysis

- A detailed look at the existing key processes that supply customer requirements in order to identify cycle time, rework, downtime, and other steps that don't add value for the customer.

Process Analysis Tools

- Process flow maps
- Value Matrix and/or Value Stream Map
- FMEA
- 5 Whys

Data Analysis Tools

- Graphical analysis
  - Histograms
  - Run Charts
  - Pareto Plots
  - Box Plots
  - Scatter Plots
- Statistical analysis
  - Hypothesis tests

LEAN SIX SIGMA TE/TTM/TT 533 ONE-SAMPLE T - TESTS

Objectives

- Understand when to use an upper tailed, lower tailed, or two-tailed t-test
- Interpret the results of the one sample t-test in JMP
- Evaluate the assumptions of the one sample t-test

t-test on the Mean

- The t-test on the mean is used to test whether a mean is different from a hypothesized value.
- Examples:
  - Are expense reports being paid in less than 10 days?
  - Did the number of units returned last quarter exceed 100 units?
  - Are we billing significantly more or less than our target of $55,000 per month?

Selecting the Correct Test

- Two-tailed test
  - Used to determine if the mean is significantly greater than or lower than a hypothesized or stated value (first p-value in JMP output)
- Upper-tailed test
  - Used to determine if the mean is significantly greater than a hypothesized or stated value (second p-value in JMP output)
- Lower-tailed test
  - Used to determine if the mean is significantly less than a hypothesized or stated value (third p-value in JMP output)

Two-Tailed Test

Practical Question: Are we invoicing $50,000 per month?

Hypothesis

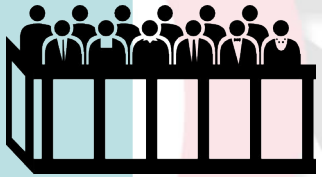
$H_o: \mu = \mu_o$
$H_A: \mu \neq \mu_o$

μ_o **is a known value**

Significance

$\alpha = 0.05$

Based on risk tolerance



Analyze Data

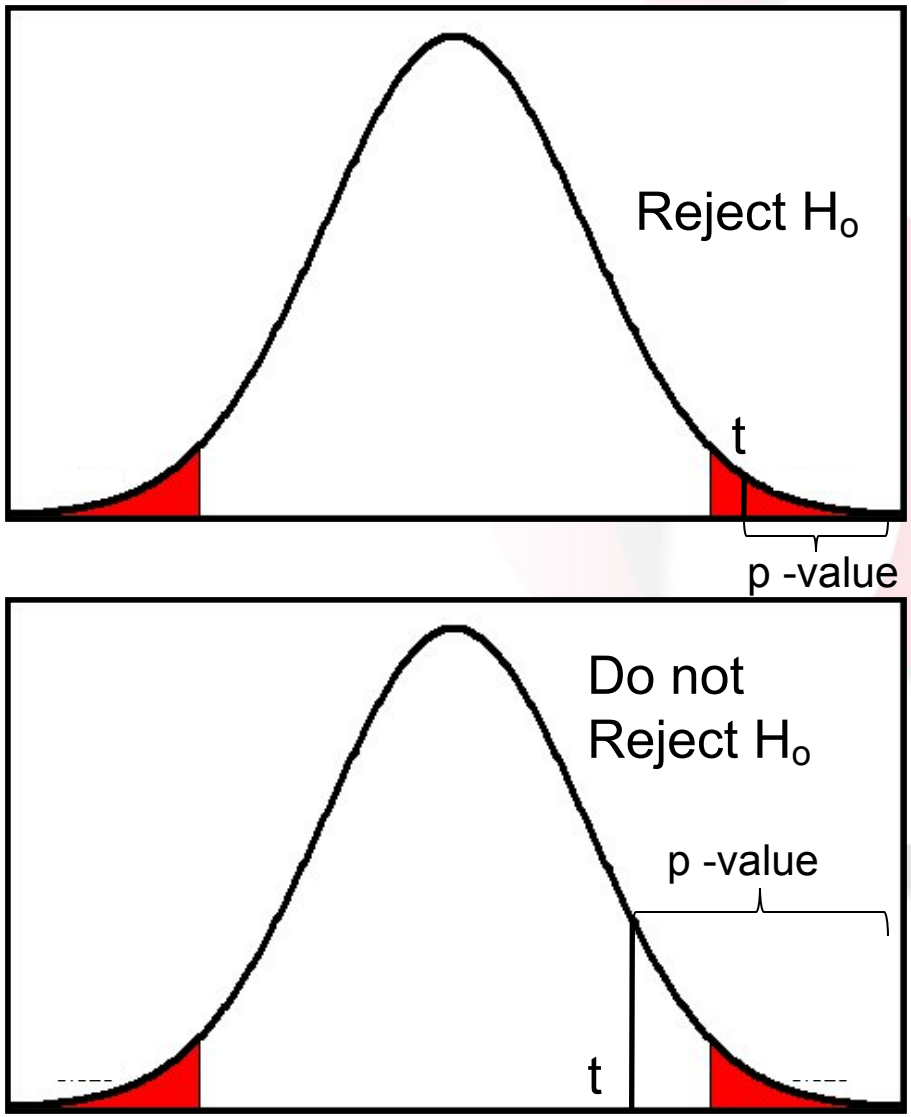
Calculate t statistic

Decision

$p < 0.05$ Reject H_o
$p > 0.05$ Fail to Reject H_o

Two-Tailed Test

- The rejection region is divided evenly in each tail based on the specified alpha value.
- Calculate the t-statistic and determine whether or not it is in one of the rejection regions.
- The p-value is the area under the curve past the test statistic.



Upper-Tailed Test

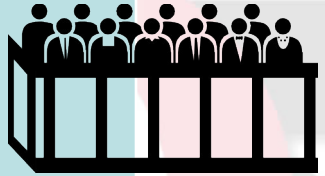
Hypothesis

$H_o: \mu <= \mu_o$
$H_A: \mu > \mu_o$

μ_o is a known value

Significance

$\alpha = 0.05$



Based on risk tolerance

Analyze Data

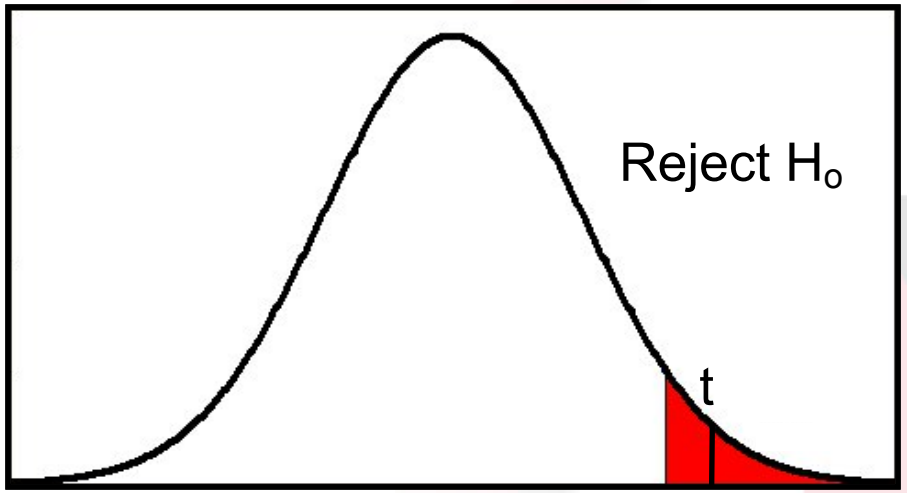
Calculate t statistic

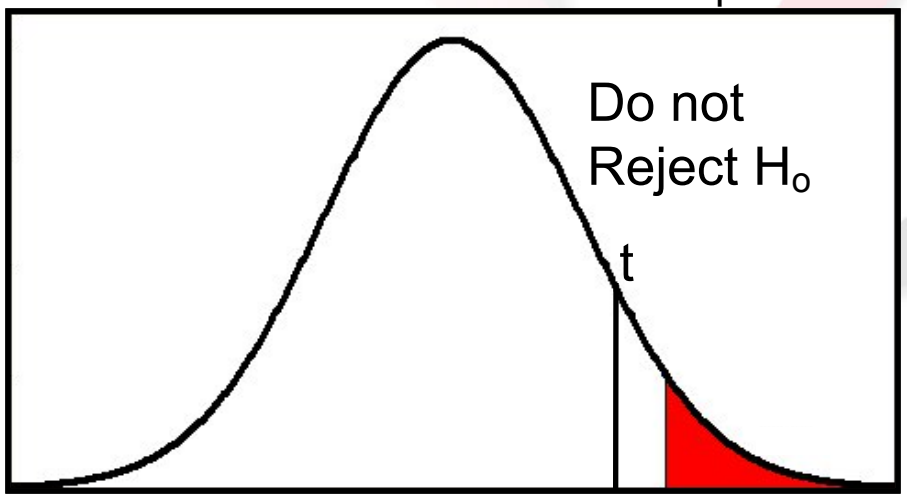
Decision

$p < 0.05$ Reject H_o
$p > 0.05$ Fail to Reject H_o

Upper Tailed Test

- All the rejection region is in the right or upper tail of the normal distribution.
- Calculate the t-statistic and determine whether or not it is in the rejection regions.
- The p-value is the area under the curve past the test statistic.





Lower-Tailed Test

Hypothesis

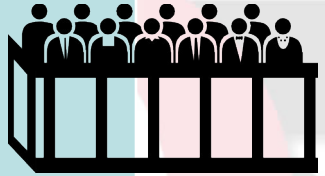
H_o: $\mu >= \mu_o$
H_A: $\mu < \mu_o$

μ_o is a known value

Significance

$\alpha = 0.05$

Based on risk tolerance



Analyze Data

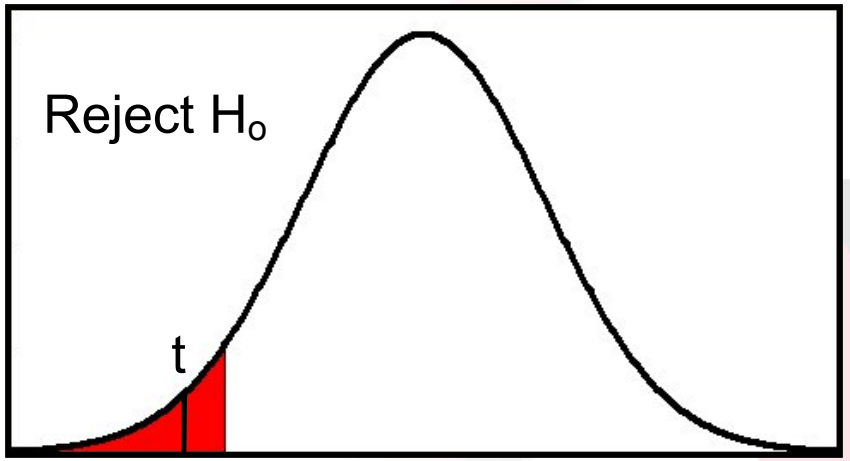
Calculate t statistic

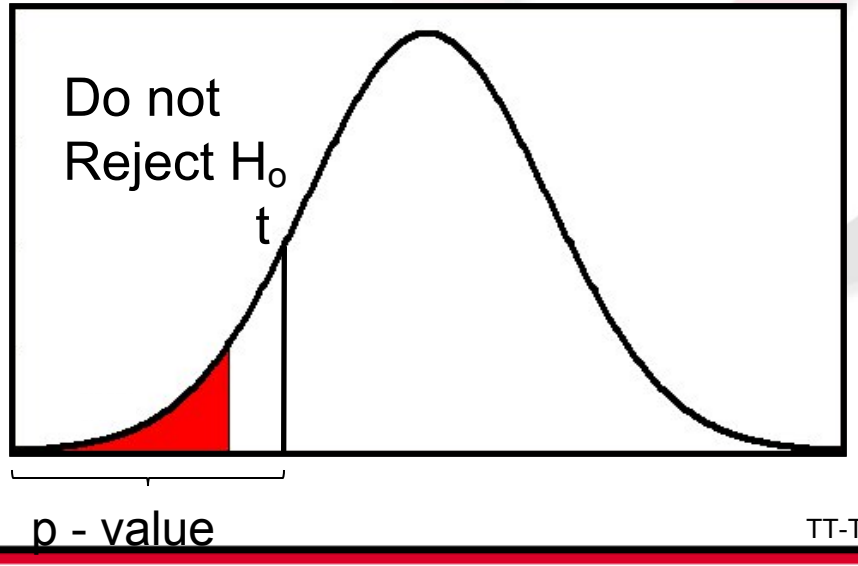
Decision

$p < 0.05$ Reject H_o
$p > 0.05$ Fail to Reject H_o

Lower Tailed Test

- All the rejection region is is in the left or lower tail based on the specified alpha value.
- Calculate the t-statistic and determine whether or not it is in one of the rejection regions.
- The p-value is the area under the curve past the test statistic.





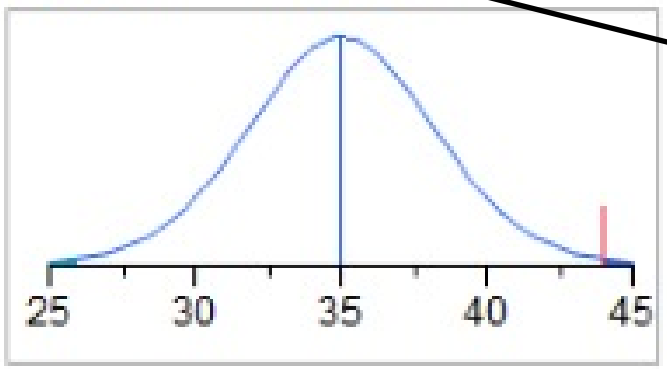
Examining JMP Output

Test Mean

Hypothesized Value	35
Actual Estimate	44.0417
DF	23
Std Dev	15.5437

t Test

Test Statistic	2.8497
Prob > |t|	0.0091*
Prob > t	0.0045*
Prob < t	0.9955



Two – tailed (| t |)

Upper – tailed (>)

Lower – tailed (<)

Must Know Formula

$$t = \frac{\bar{y} - m_o}{s_{\bar{y}}}$$

m_o is the hypothesized mean

$\bar{y}$ is the sample mean

$s_{\bar{y}}$ is the estimated standard error of the mean

One Sample t – test Assumptions

- The means are normally distributed
  - Histogram
  - Normal Quantile Plot
  - Box and Whisker Plot
  - Shapiro – Wilk test
- Observations are independent

If the assumptions are not validated, the conclusions drawn from the results may not be valid.

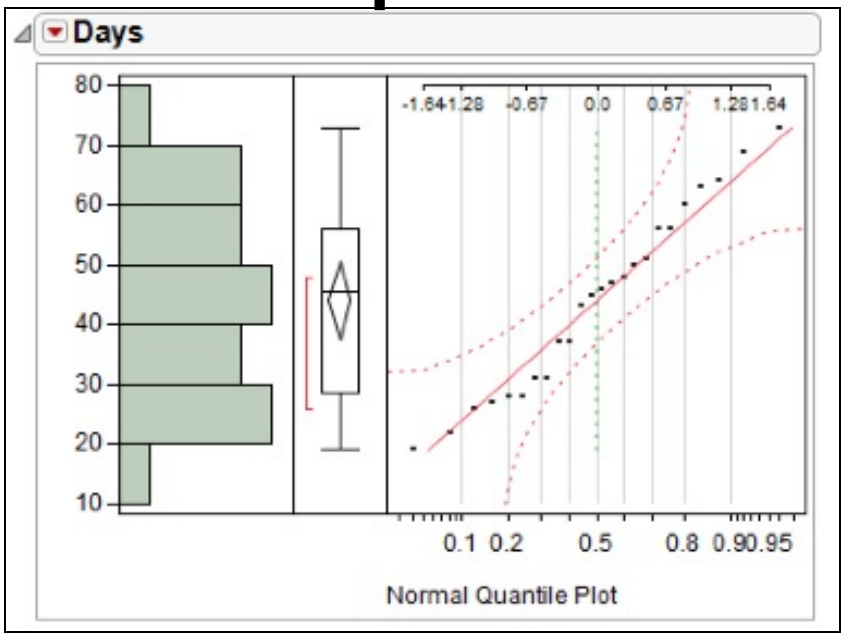
Example

An insurance company is concerned that claims are not being processed in 35 days. They randomly selected 24 claims that were processed during the last quarter and recorded the number of days it took to process each claim.

$$H_o: \mu = 35 \quad \text{versus} \quad H_a: \mu \neq 35$$

1. Open **Processing Days.jmp**.
2. Select → **Analyze** → **Distribution**.
3. Select → **Days** → **Y, Columns**.
4. Select **OK**.

Example – Validating Assumptions



5. Select the red triangle next to Days and select **Normal Quantile Plot**.

- Does the histogram look bell shaped?
- How are the mean and median related?
- Is there an obvious curvature in the normal quantile plot?

Example – Validating Assumptions

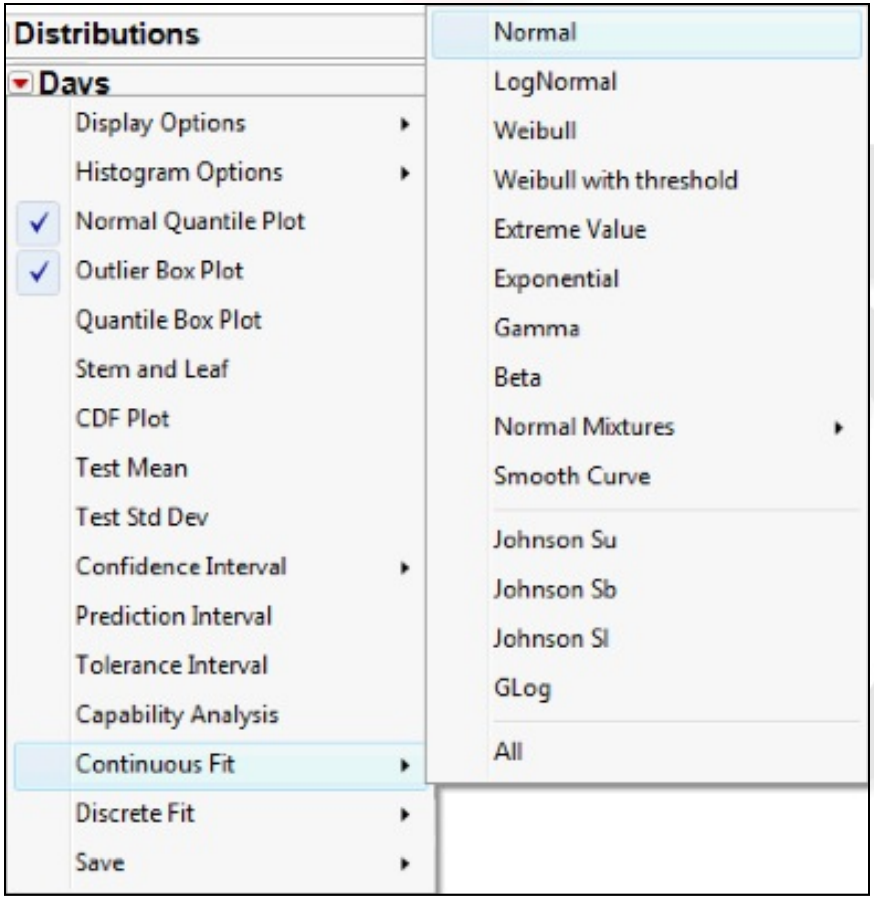
In addition to assessing normality graphically, the Shapiro Wilk test may be used to test the following hypothesis:

H_o: The data is normally distributed

H_a: The data is not normally distributed

Example – Validating Assumptions

1. Click the red triangle next to Days.
2. Select **Continuous Fit→ Normal**.
3. Click the red triangle next to **Fitted Normal** and select **Goodness of Fit**.



Example – Validating Assumptions

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	44.041667	3.1728463	37.823002	50.260331
Dispersion	σ	15.543709	0.5877208	14.433445	16.838811

Measures	
-2*LogLikelihood	198.80454
AICc	203.37597
BIC	205.16064

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9622983	0.4863

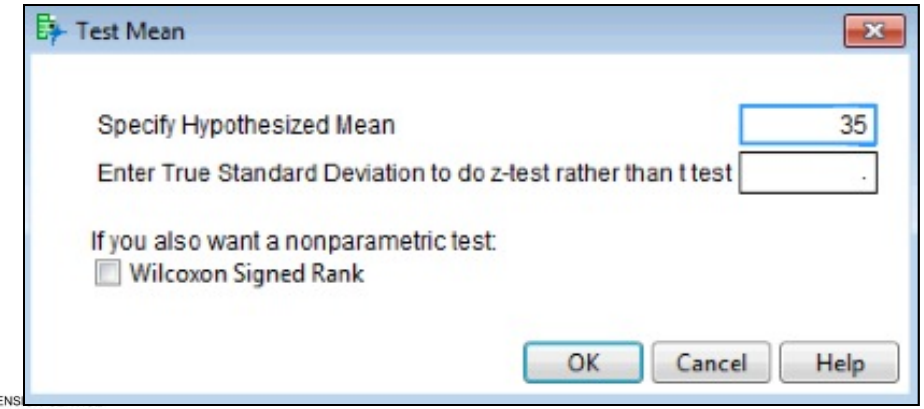
	A2	Simulated p-Value
Anderson-Darling	0.3072583	0.5416

If α = 0.05, what conclusion should be made based on the Shapiro – Wilk Test?

Example – Testing the Mean

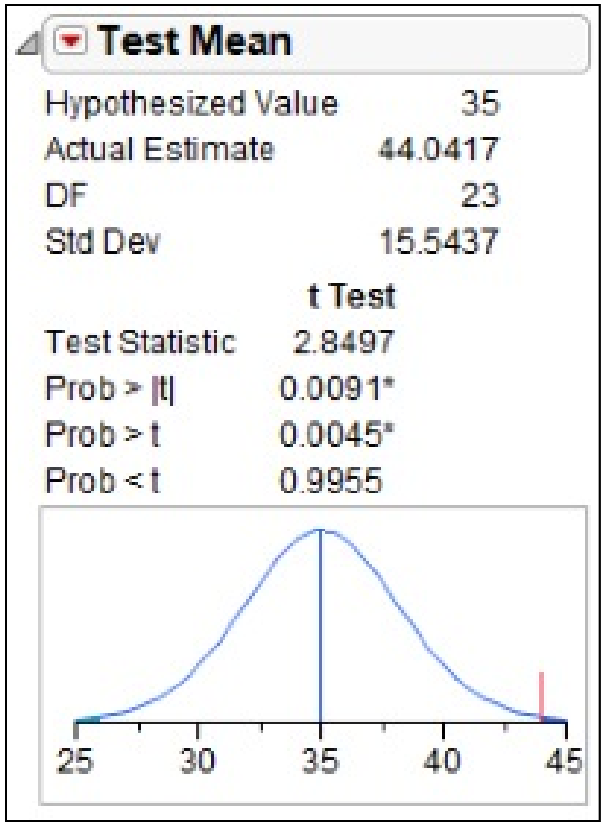
Now that the assumptions have been validated and it is concluded that the one sample t – test will be valid, the hypothesis can be tested.

1. Click the red triangle next to **Days** and select **Test Mean**.
2. Enter **35** as the hypothesized mean.



Example – Testing the Mean

3. Select OK.



- Which p – value should be evaluated?
- Why?
- What are your conclusions?
- State your conclusions in practical terms.

NC STATE UNIVERSITY

Lean Six Sigma TE/TTM/TT

533

Two Sample Analysis

Two Sample Analysis

Objective: To investigate the effect of independent factors on a variable response factor.

Deliverables: Updated list of input variables.

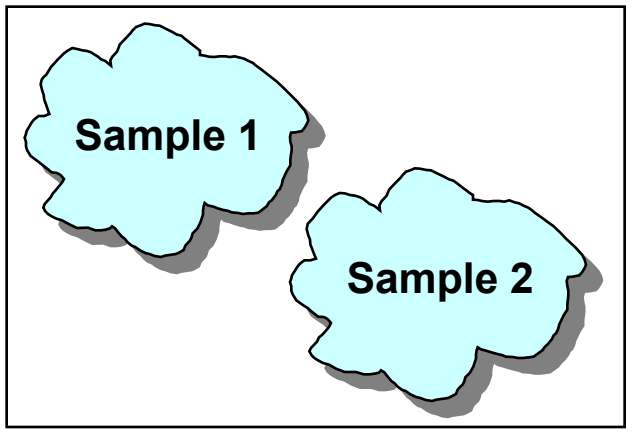
Two Sample Analysis

Hypothesis Testing Situations

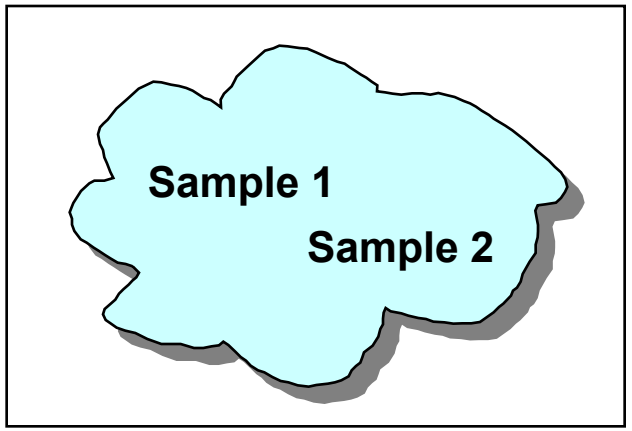
1. Given two samples: Compare the Standard Deviations between the two samples. [Ho: $s_{population1} = s_{population1}$]
2. Given two samples: Compare the Means between the two samples. [Ho: $m_{population1} = m_{population2}$]
3. Given two "Paired" Samples: Compare the average paired difference, d, to a target mean = 0. [Ho: $d_{paired\ dif} = 0$]

Why Test Population Parameters Against Each Other?

- We never know the true population parameters. Thus, we use statistics from samples to determine how likely it is that these two samples came from one population.
- These comparisons allow us to compare the performance of two process m's or s's.
- Then, we can conclude with some degree of statistical confidence that these parameters came from one population or two.

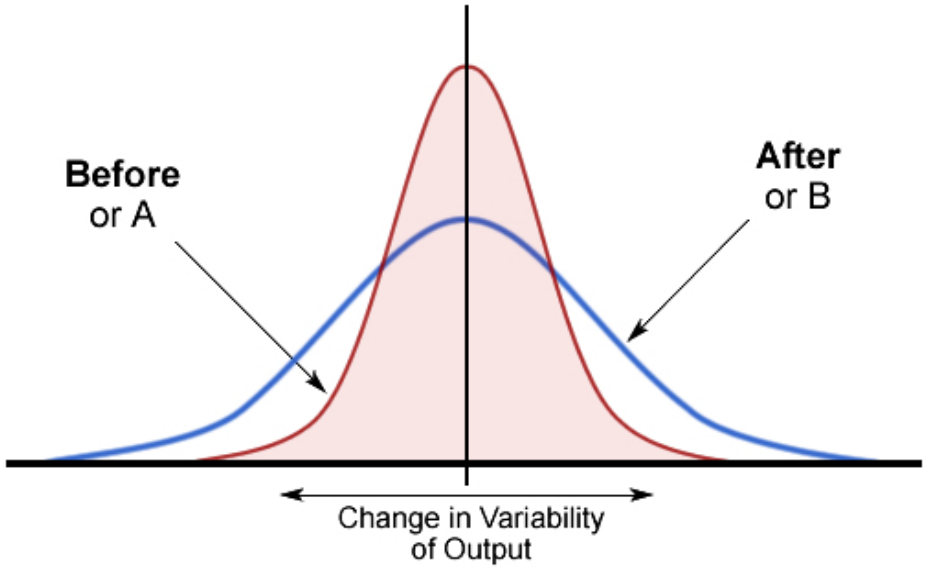


OR



Comparing 2 Variances

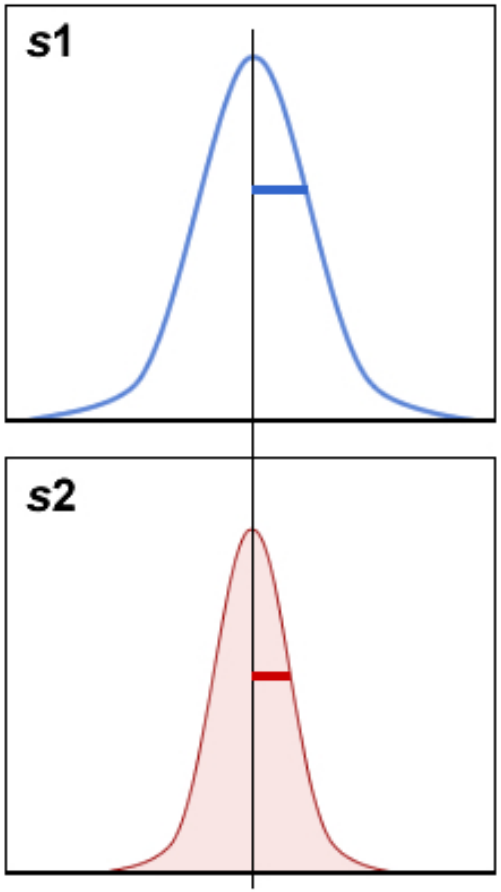
- Given two samples: Compare the Standard Deviations between the two samples. [Ho: $s_{population1} = s_{population1}$]
- Is there a significant difference in the VARIABILITY of Outputs?



Comparing 2 Variances

- When comparing two population means using variables data, first decide if a statistical difference exists in the variances (Unequal Variances). This test is important since it affects the formula used to perform the test on the means.
- We also need to know if the distributions are normally distributed since this determines the type of variance test used.
- The results of the normality and variance tests will determine underlying assumptions needed for the analysis of the population central tendencies (means versus medians, with or without equal variances).

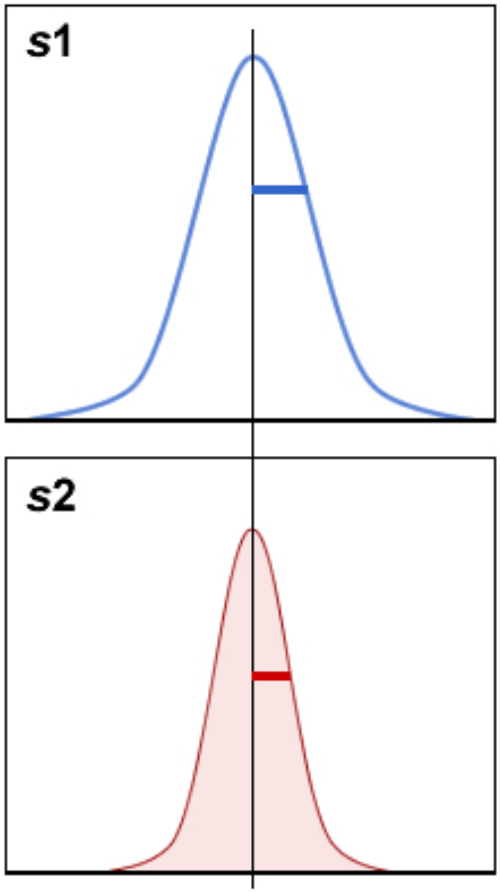
Ho: $\sigma_{pop1} = \sigma_{pop2}$
Ha: $\sigma_{pop1} > \sigma_{pop2}$



Comparing 2 Variances

- "Is there a real difference between s_{pop1} and s_{pop2}?"
- This test requires the use of the **F test** statistic
- **Bartlett's test** statistic can be used for a comparison of 2 or more variances. This test assumes the data are **Normally distributed**
- **Levene's test** statistic is also used to compare 2 or more variances and is appropriate for continuous data which are **not Normally distributed**

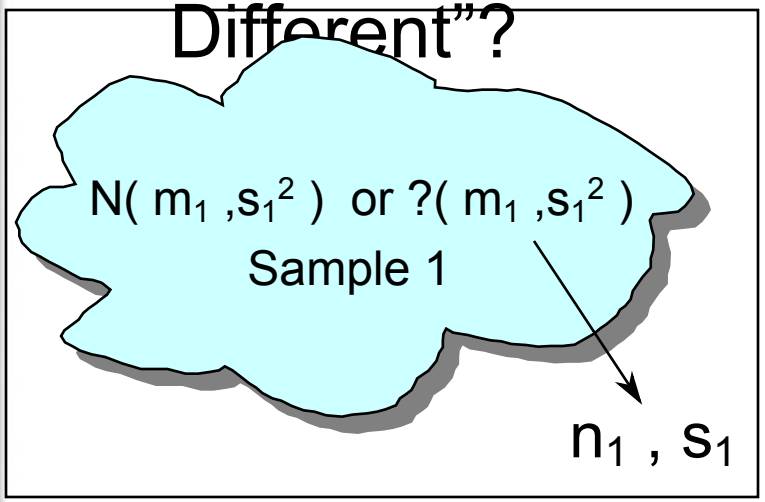
Ho: $\sigma_{pop1} = \sigma_{pop2}$
Ha: $\sigma_{pop1} > \sigma_{pop2}$

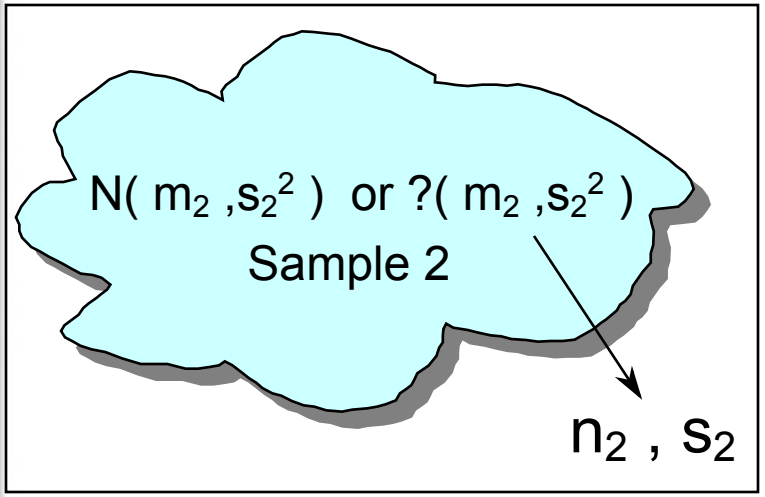


Test for Equal Variance

Ho: $\sigma_{pop1} = \sigma_{pop2}$
Ha: $\sigma_{pop1} > \sigma_{pop2}$

Are Variances the "Same or Different"?





Calculate F_{calc} such that $F_{calc} > 1$

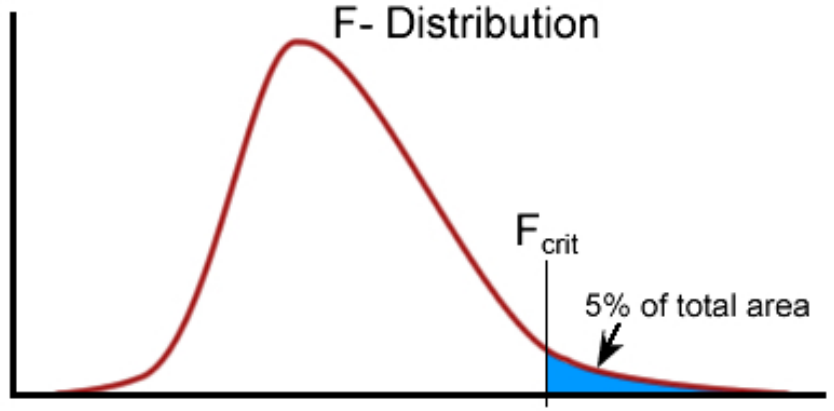
$F_{calc} = s_1^2 / s_2^2$

Select s_1, as the largest, $s_1 > s_2$

Compare to F_{crit} (for a = 0.05)

If $F_{calc} > F_{crit}$, then REJECT Ho.

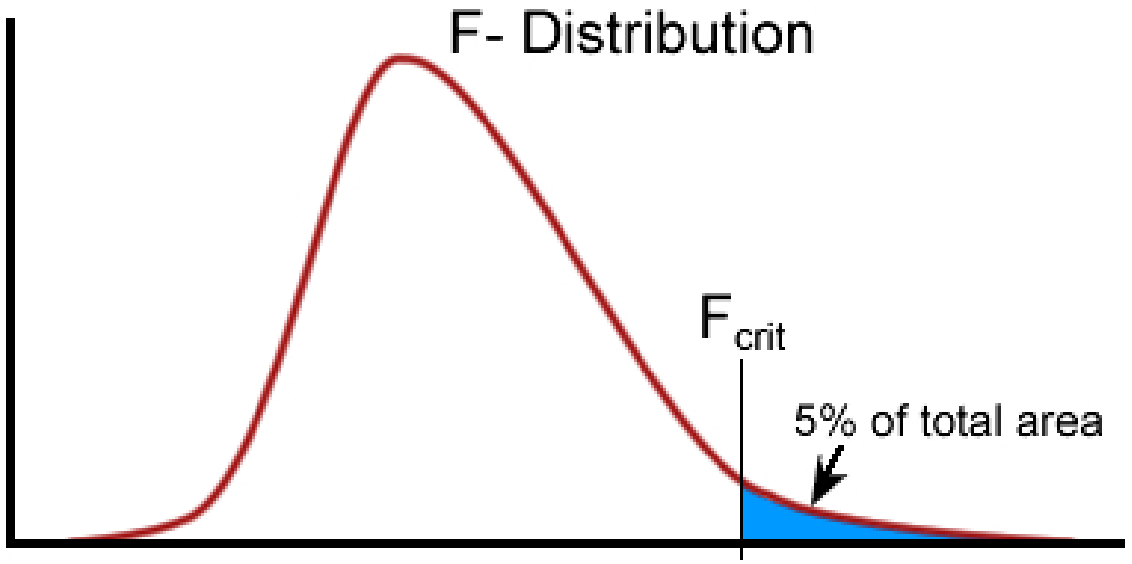
(Ho: variances s_1^2 and s_2^2 are the same)



What is $F_{critical}$?

- The critical value can be looked up in a table or calculated by JMP. The value depends upon the alpha level and the degrees of freedom for both factors (denominator and numerator).
- When the calculated F value exceeds the critical F (at a = 0.05), the p-value will be less than 0.05. A high calculated F and a low p-value indicate that the two variances are different.

Ho: $\sigma_{pop1} = \sigma_{pop2}$
Ha: $\sigma_{pop1} > \sigma_{pop2}$



Test for Unequal Variances

- If the normal probability plot indicates we are dealing with normally distributed data, then we use **Bartlett's Test Statistic** in JMP (or if there are only two variances, we use the F test)
- If the data are not normally distributed, then we use **Levene's Test Statistic** in JMP
- **Analyze>Fit Y by X>Unequal Variances**

$$H_o: \sigma_{pop1} = \sigma_{pop2}$$
$$H_a: \sigma_{pop1} \neq \sigma_{pop2}$$

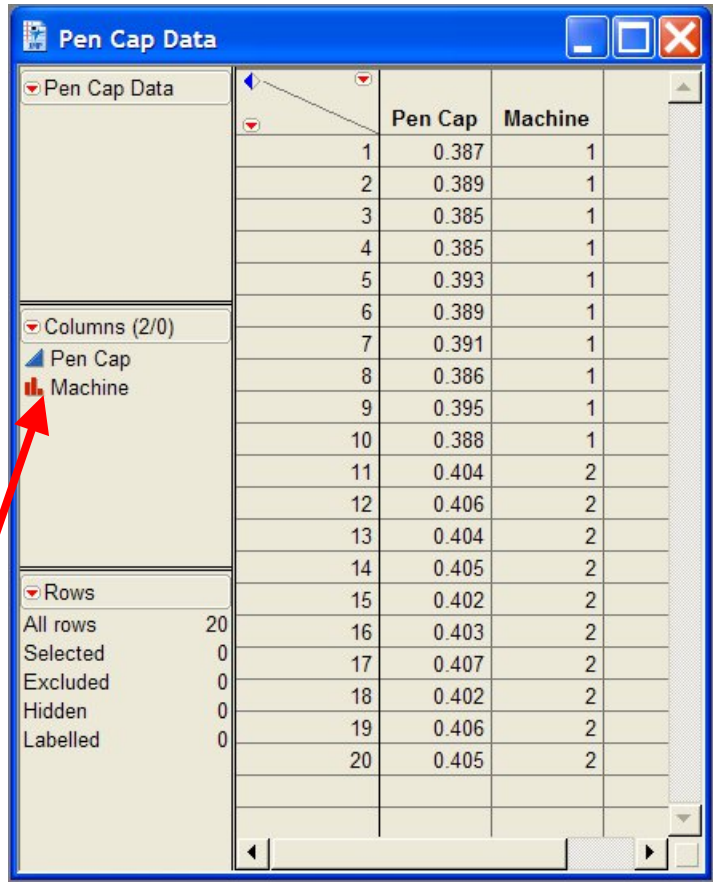
Pen Cap Machine Example

Open JMP with Pen Cap Data (Diameter Machine Data)

Machine 1	Machine 2
0.387	0.404
0.389	0.406
0.385	0.404
0.385	0.405
0.393	0.402
0.389	0.403
0.391	0.407
0.386	0.402
0.395	0.406
0.388	0.405

Remember that you have to stack the data when entering it into JMP

Note: The Machine column is Nominal Data



Pen Cap Machine Example *(Cont.)*

Step 1: Practical Problem – Modifications have been made to one of the machines. We want to see if we have "significantly" improved the cap diameter of the machine with these modifications before we spend significant time and resources modifying the other machine. After we get cap diameter samples from each machine, how do we determine if there is a "real" difference between the two machines?

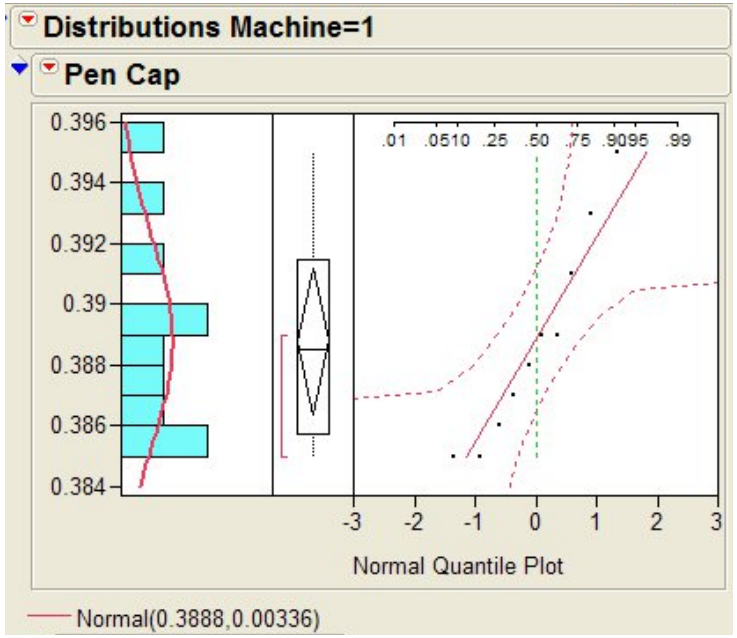
Pen Cap Machine Example *(Cont.)*

Step 2: Determine if the machine samples are normally distributed.

- **JMP>Analyze>Distribution**
  - Select *Pen Cap* as the **Y, Columns**
  - Select *Machine* as the **By, Columns**
  - Click **OK**
- From the **Red Triangle** under **Machine =1**, **Pen Cap**, select **Normal Quantile Plot** and then **Continuous Fit > Fit Normal**
- From the **Fitted Normal Red Triangle**, select **Goodness of Fit**
- Repeat the steps for **Machine =2**

Pen Cap Machine Example *(Cont.)*

- Normality Test for Machine #1



Moments

Mean	0.3888
Std Dev	0.0033599
Std Err Mean	0.0010625
upper 95% Mean	0.3912035
lower 95% Mean	0.3863965
N	10

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	0.3888	0.3863965	0.3912035
Dispersion	σ	0.0033599	0.0023111	0.0061339

Goodness-of-Fit Test

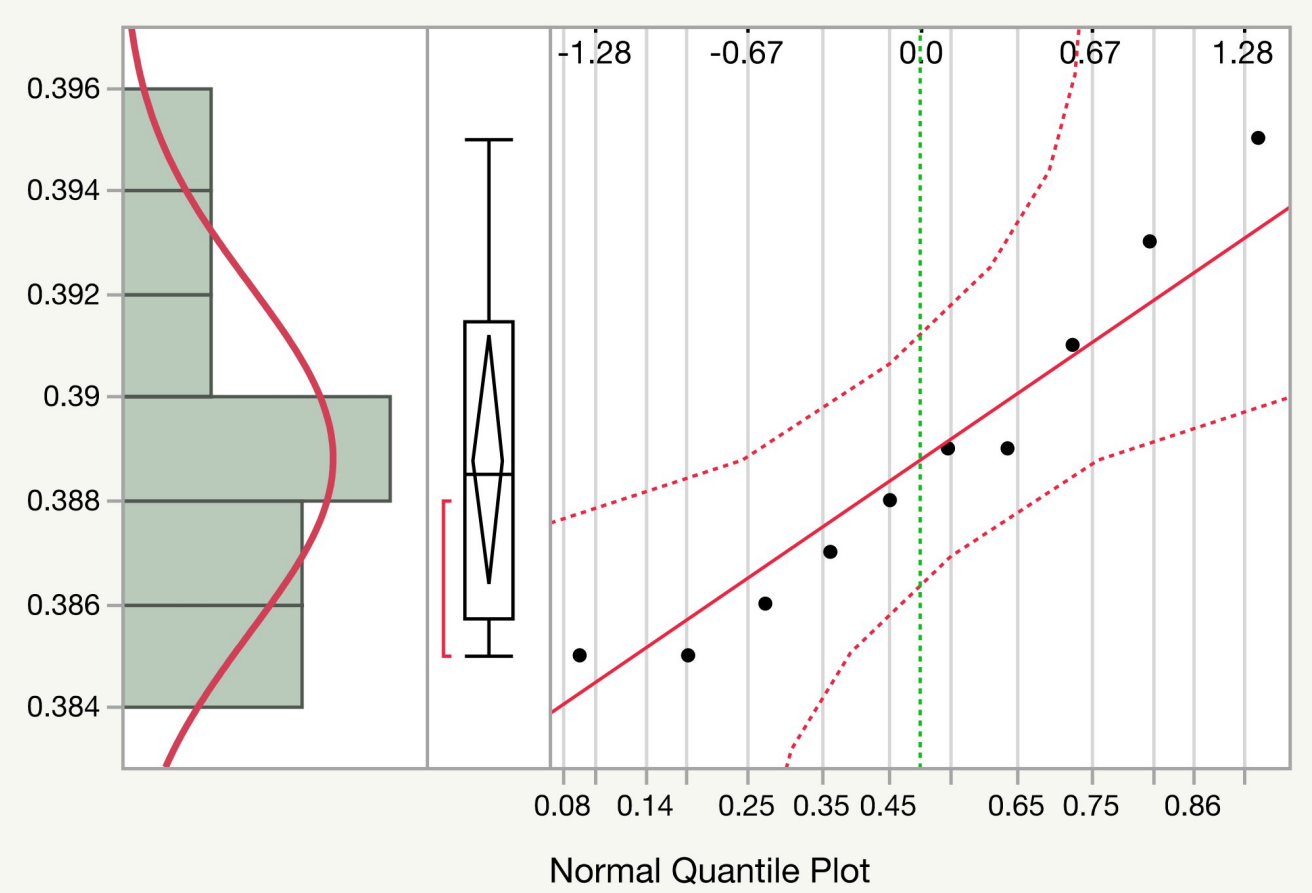
Shapiro-Wilk W Test

W	Prob<W
0.931116	0.4590

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

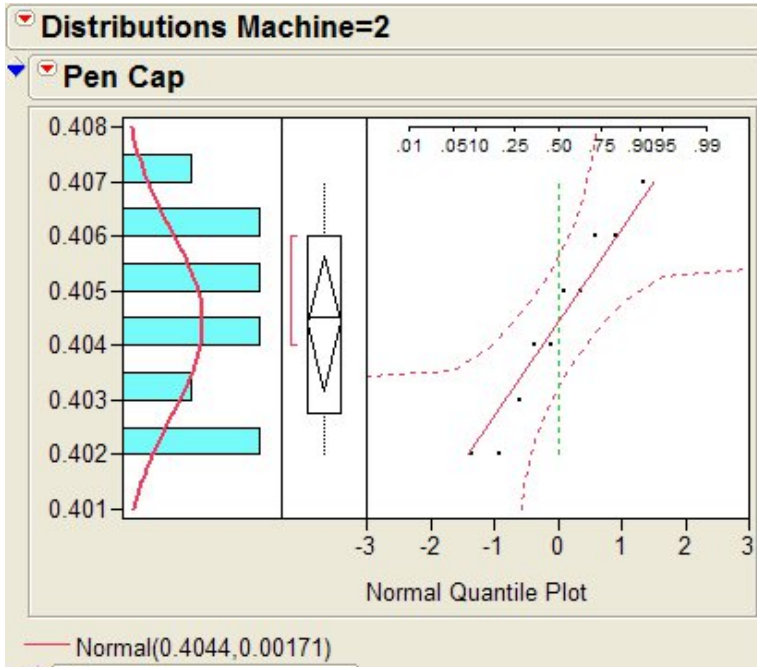
Pen Cap Machine Example *(Cont.)*

- Normality Test for Machine #1



Pen Cap Machine Example *(Cont.)*

- Normality Test for Machine #2



Moments

Mean	0.4044
Std Dev	0.0017127
Std Err Mean	0.0005416
upper 95% Mean	0.4056252
lower 95% Mean	0.4031748
N	10

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	0.4044	0.4031748	0.4056252
Dispersion	σ	0.0017127	0.0011781	0.0031267

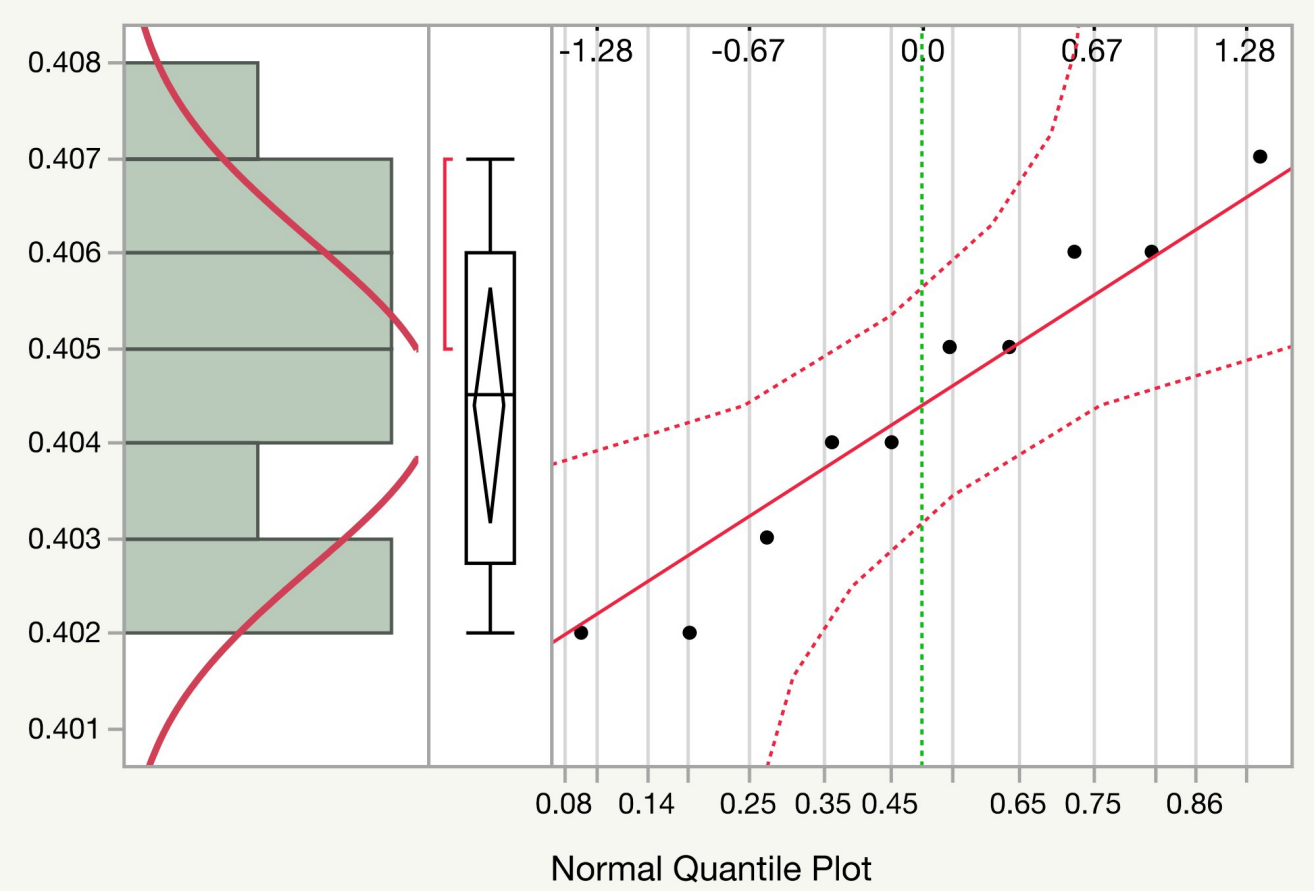
Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.943329	0.5906

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Pen Cap Machine Example *(Cont.)*



Pen Cap Machine Example *(Cont.)*

Step 2 (cont.):

Ho: Machine sample is Normal

Ha: Machine sample is NOT Normal

Fail to reject Ho.

Machine samples are normal.

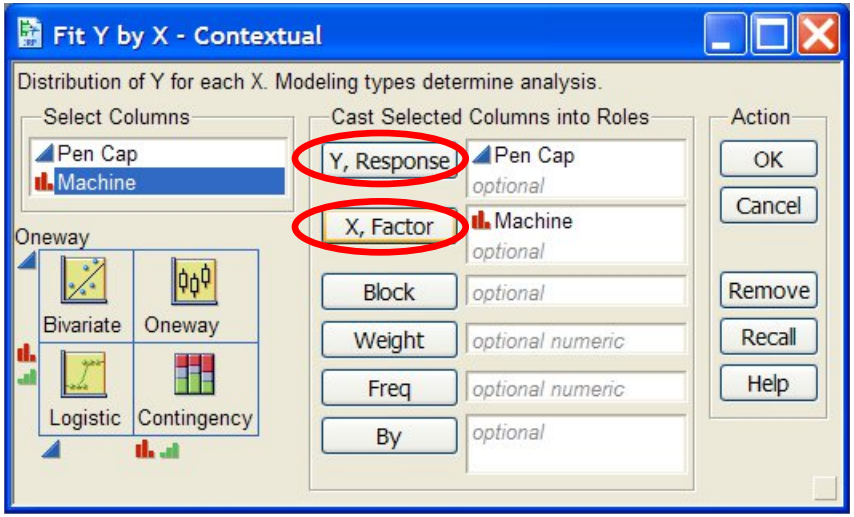
Step 3: In order to determine if the machine yield has improved or not, we must first check for equality of variances. State the Null and the Alternative Hypotheses:

For s: Ho: $s_{machine1} = s_{machine2}$

Ha: $s_{machine1}{}^{1} \neq s_{machine2}$

Pen Cap Machine Example *(Cont.)*

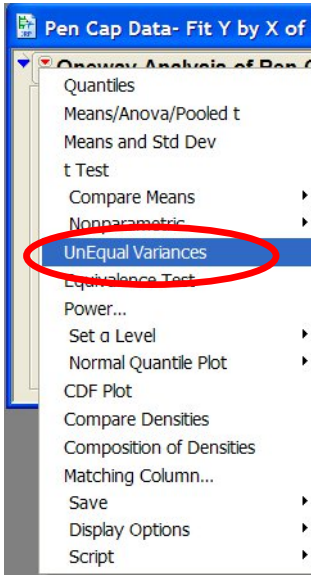
- Return to the Pen Cap data table
- **Analyze>Fit Y by X** select Y, Response, *Pen Cap*, select X, Factor, *Machine*)
- Click **OK**

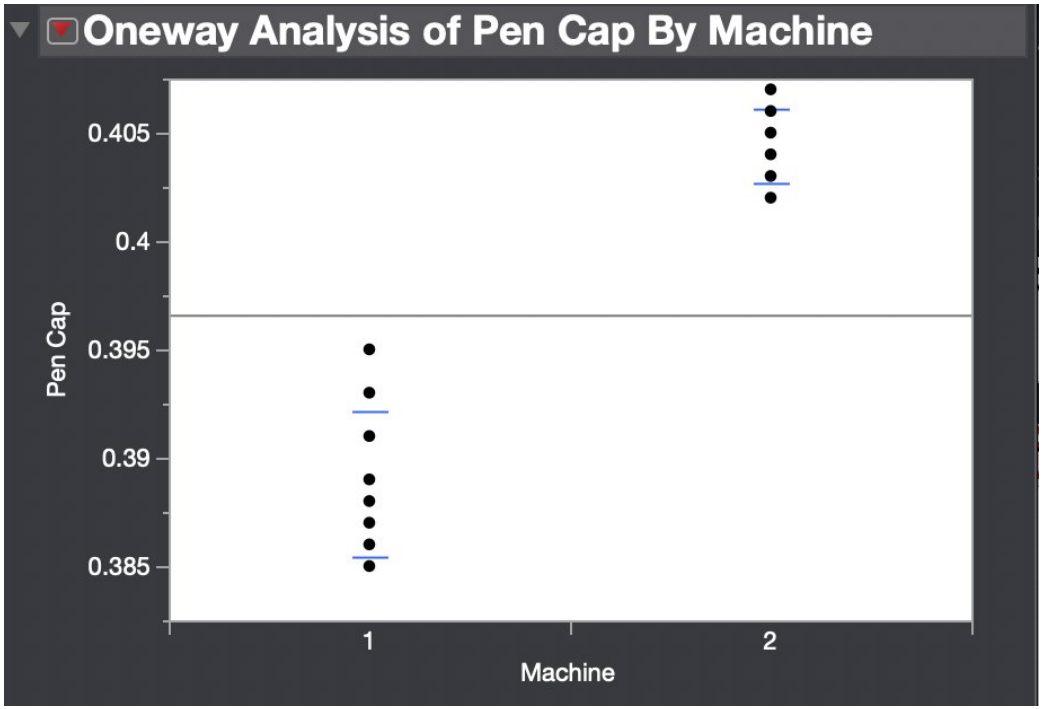


Six Sigma – *Two Sample Analysis*

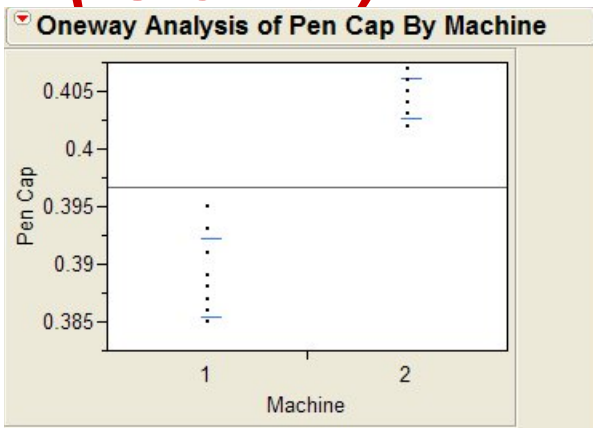
Pen Cap Machine Example *(Cont.)*

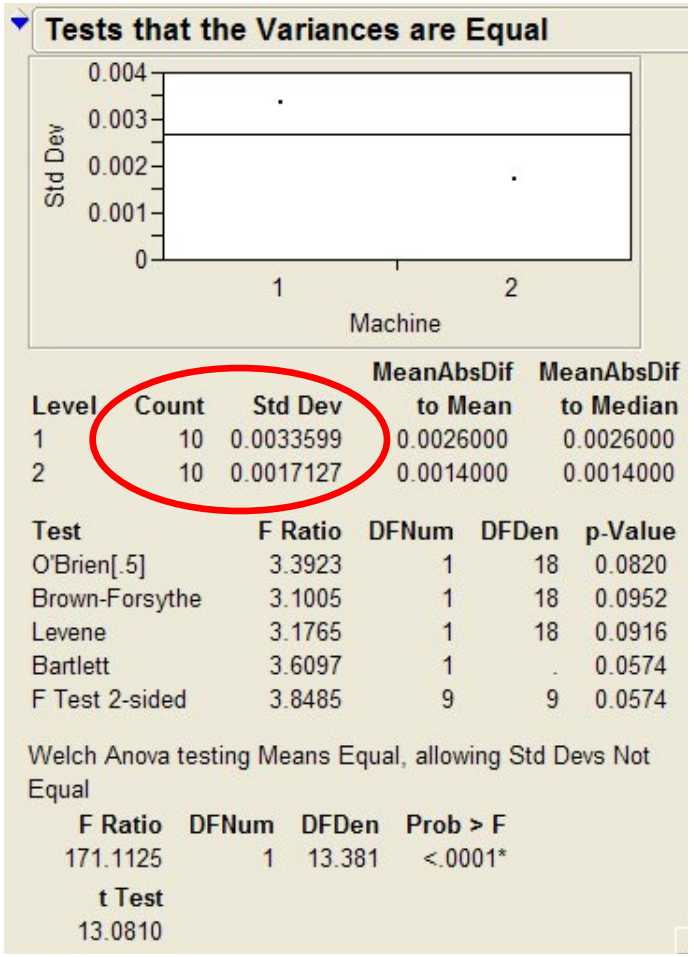
- From the Oneway Analysis Pen Cap Red Triangle, select **Unequal Variances**





Pen Cap Machine Example *(Cont.)*





Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	10	0.0033599	0.0026000	0.0026000
2	10	0.0017127	0.0014000	0.0014000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	3.3923	1	18	0.0820
Brown-Forsythe	3.1005	1	18	0.0952
Levene	3.1765	1	18	0.0916
Bartlett	3.6097	1	.	0.0574
F Test 2-sided	3.8485	9	9	0.0574

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
171.1125	1	13.381	<.0001*

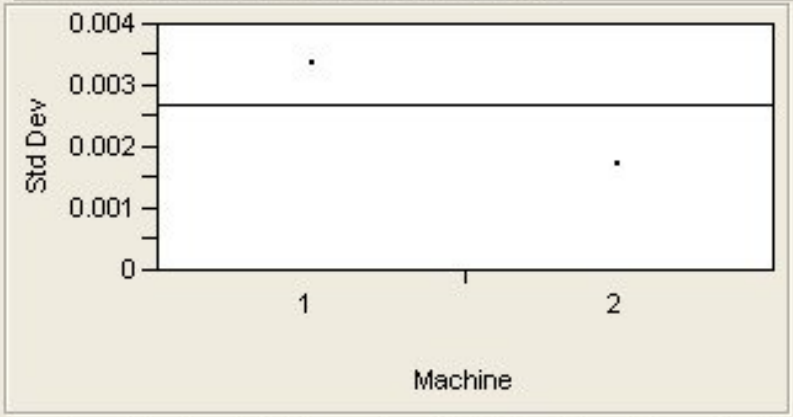
t Test
13.0810

Ho: smachine1 = smachine2
Ha: smachine1 ≠ smachine2

Is the Std Dev of Machine 1 statistically different from the Std Dev of Machine 2?

Pen Cap Machine Example (Cont.)

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	10	0.0033599	0.0026000	0.0026000
2	10	0.0017127	0.0014000	0.0014000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	3.3923	1	18	0.0820
Brown-Forsythe	3.1005	1	18	0.0952
Levene	3.1765	1	18	0.0916
Bartlett	3.6097	1	.	0.0574
F Test 2-sided	3.8485	9	9	0.0574

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
171.1125	1	13.381	<.0001*

t Test
13.0810

Ho: smachine1 = smachine2
Ha: smachine1 ≠ smachine2

Note: smachine1 is larger than smachine2

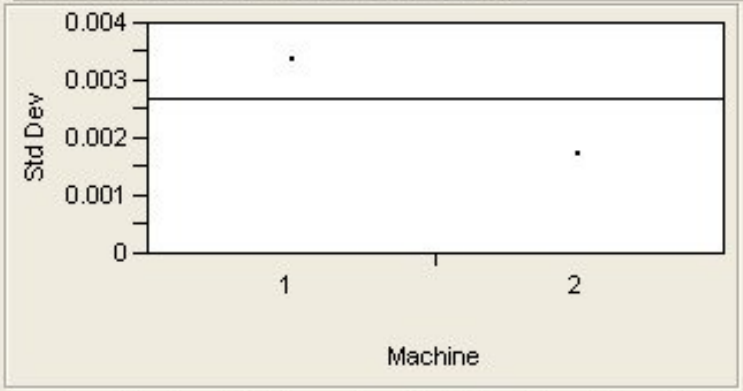
Use Levene if Data Sets are **NOT** normally distributed

Use Bartlett if Data Sets are normally distributed

If the Std Dev's are statistically different, use Welch to test means

Pen Cap Machine Example (Cont.)

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
1	10	0.0033599	0.0026000	0.0026000
2	10	0.0017127	0.0014000	0.0014000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	3.3923	1	18	0.0820
Brown-Forsythe	3.1005	1	18	0.0952
Levene	3.1765	1	18	0.0916
Bartlett	3.6097	1		0.0574
F Test 2-sided	3.8485	9	9	0.0574

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
171.1125	1	13.381	<.0001*

t Test
13.0810

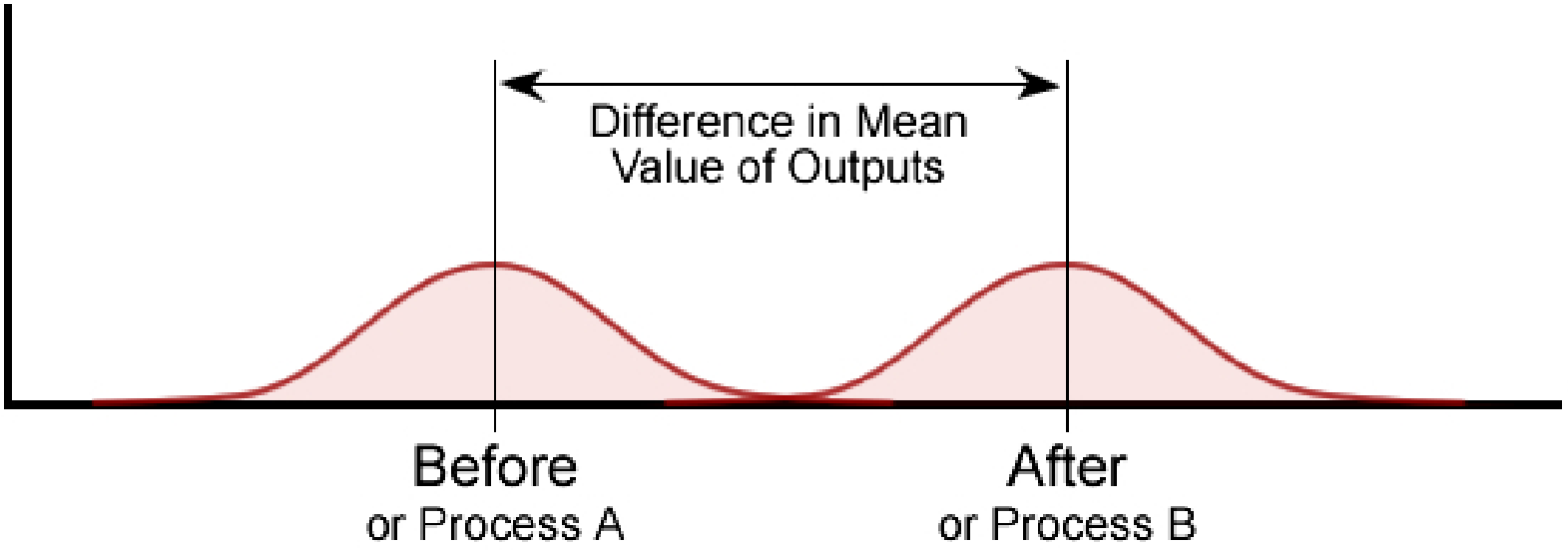
Ho: smachine1 = smachine2
Ha: smachine1 ≠ smachine2

Note: smachine1 is larger than smachine2

Data sets are normally distributed, use Bartlett test p-value 0.0574
Fail to reject Ho (Std Devs may be the same)

Comparing Two Means

- Given two samples: Compare the Means between the two samples. [Ho: $m_{population1} = m_{population2}$]
- Is there a significant difference in the MEAN VALUE of Outputs?



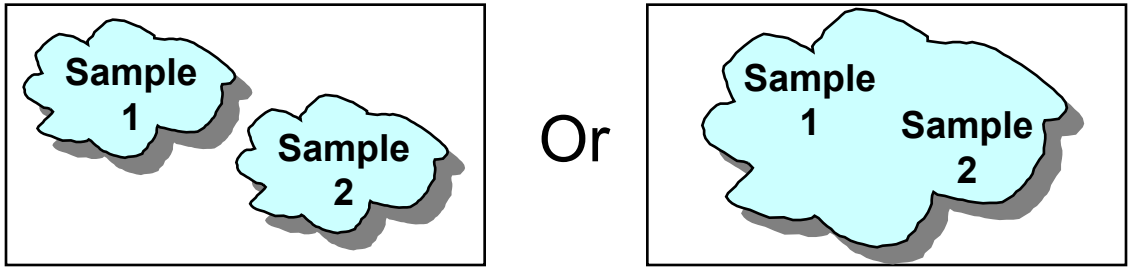
Population 1 Mean vs. Population 2 Mean

- Requires the t-distribution and variable data
- This statistic can be used to test if the samples taken are from one population or two distinct populations
- In other words:

  "Is there a real difference between m_{pop1} and m_{pop2}?"
- The formula for the "2 Sample T-Test" statistic is:
- Equal Variances (degrees of freedom = n1+ n2 - 2):

  $t\ (calc) = (\bar{X}_{pop1} - \bar{X}_{pop2}) / sqrt\{s^2_{pooled} * (1/n_1 + 1/n_2)\}$
- Unequal Variances (degrees of freedom < n1+ n2 - 2):

  $t\ (calc) = (\bar{X}_{pop1} - \bar{X}_{pop2}) / sqrt\{s^2_1/n_1 + s^2_2/n_2\}$



Do Two Samples Come From the "Same" or "Different" Populations?

Pen Cap Machine Example *(Cont.)*

Step 1: Practical Problem – Modifications have been made to one of the machines. We want to see if we have "significantly" improved the cap diameter of the machine with these modifications before we spend significant time and resources modifying the other machine. After we get cap diameter samples from each machine, how do we determine if there is a "real" difference between the two machines?

Step 2: Recall: The data sets are normal and variances are equal.

Step 3: State the Null and Alternate Hypotheses for improved Pen Cap Machine 2:

For m:			
	Ho: $m_{machine1} = m_{machine2}$	Ho: $m_{machine1} = m_{machine2}$	Ho: $m_{machine1} = m_{machine2}$
	Ha: $m_{machine1} \neq m_{machine2}$	Ha: $m_{machine1} > m_{machine2}$	Ha: $m_{machine1} < m_{machine2}$

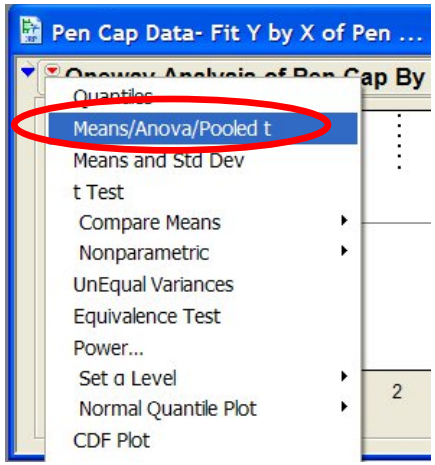
Pen Cap Machine Example *(Cont.)*

Step 4: Compare p-Value to Alpha and reject or fail to reject Ho.

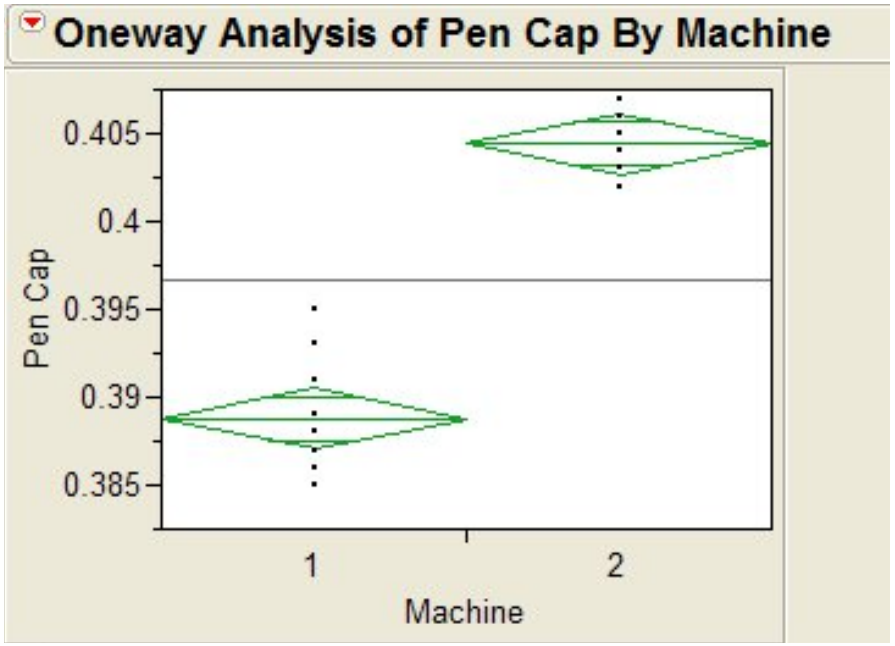
Pen Cap Machine Example *(Cont.)*

Two Sample-t Test

- **JMP>Analyze>Fit Y by X**
  - For **Y,Response** select *Pen Cap* and for **X,Factor** select *Machine,* click **OK**
- From the **Oneway Analysis of Pen Cap Diameter By Machine # Red Triangle,** select **Means/Anova/Pooled t**



Pen Cap Machine Example *(Cont.)*



Six Sigma – *Two Sample Analysis*

Pen Cap Machine Example *(Cont.)*

Difference = Machine2 – Machine1
Difference = 0.404399 - 0.388800
= 0.01560

Oneway Anova

Summary of Fit

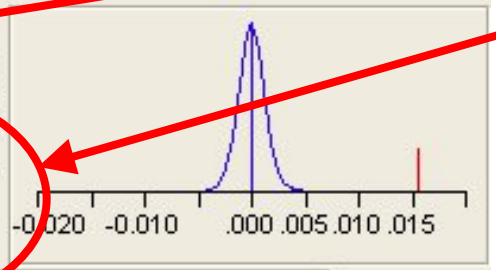
Rsquare	0.904819
Adj Rsquare	0.899531
Root Mean Square Error	0.002667
Mean of Response	0.3966
Observations (or Sum Wgts)	20

t Test

2-1

Assuming equal variances

Difference	0.015600	t Ratio	13.081
Std Err Dif	0.001193	DF	18
Upper CL Dif	0.018105	Prob > |t|	<.0001*
Lower CL Dif	0.013095	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Machine	1	0.00121680	0.001217	171.1125	<.0001*
Error	18	0.00012800	7.111e-6		
C. Total	19	0.00134480			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	10	0.388800	0.00084	0.38703	0.39057
2	10	0.404400	0.00084	0.40263	0.40617

Std Error uses a pooled estimate of error variance

Prob>|t| <.0001 means that the means are different!
Prob>t <.0001 means that dif = 0.0156 is greater than 0.
Prob<t 1.000 means that dif = 0.0156 is NOT less than 0.

Pen Cap Machine Example *(Cont.)*

Step 5: "There is a real difference between $m_{machine1}$ and $m_{machine2}$." We can conclude that our expensive modifications have improved pen cap diameter (based on these two samples of data).

Average of Machine 2 (0.4044) is statistically greater than the average of Machine 1 (0.3888)

Paired Comparisons (T-tests)

Given two "Paired" Samples: Compare the average paired difference, d, to a target mean = 0.

Ho: $d_{\text{paired dif}} = 0$

Ha: $d_{\text{paired dif}} \neq 0$

Paired Comparisons (T-tests)

- Definition of a paired comparison:
  - In some cases, the samples are not independent. Rather, each observation in one sample is paired with an observation in the other.
  - Each row of data has a mate. There are measurements of the same thing twice (the pairing is the multiple measurement).
- "Blocking" is used to block out excess variability caused by the dependency of the samples.

  Block what you can, randomize what you cannot.
- With two samples, the raw data is not used in the analysis. Rather the difference between the two sets of data is compared to zero. Thus if the population curve (of the differences) has a mean of 0, then we know there is no difference between the samples.

Paired Comparison: Shoes Example

- **Problem**: Ten random boys are selected to test two types of shoe material. Each boy wears one shoe made from each material. The shoes are randomly assigned to the right or the left foot. Since both materials cost the same amount, the quality group wants to produce shoes with the material that will wear the best.
- Let's look at the data (**higher values mean more wear**).
- What is the blocking variable in this problem?
- What will blocking do for our analysis?
  - By blocking, we have effectively removed the between boy variability from our test.

Paired Comparison: Shoes Example

Shoes paired t-1.jmp
- Distribution
- Matched Pairs

Columns (4/0)
- Boy
- Mat A
- Mat B
- Diff B-A

	Boy	Mat A	Mat B	Diff B-A
1	1	13.2	14	0.8
2	2	8.2	8.8	0.6
3	3	10.9	11.2	0.3
4	4	14.3	14.2	-0.1
5	5	10.7	11.8	1.1
6	6	6.6	6.4	-0.2
7	7	9.5	9.8	0.3
8	8	10.8	11.3	0.5
9	9	8.8	9.3	0.5
10	10	13.3	13.6	0.3

- Open the file Shoes Paired t.jmp

Paired Comparison: Shoes Example

Step 1: Practical Problem: Which material should the company use to make shoes?

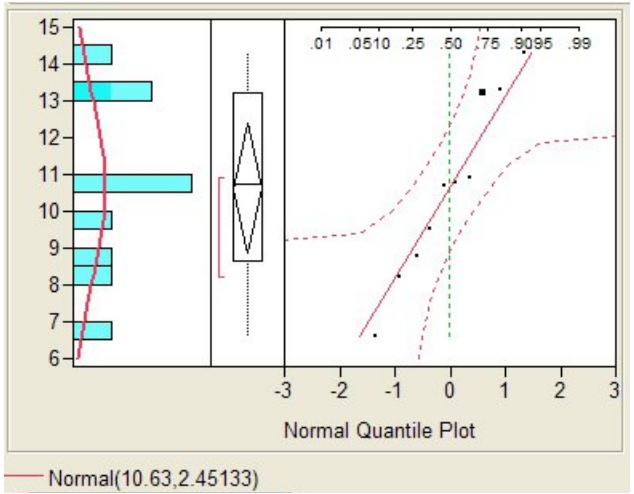
Step 2: Are the material samples normally distributed?

Ho: Material sample is Normal

Ha: Material sample is NOT Normal

Paired Comparison: Shoes Example

Material A

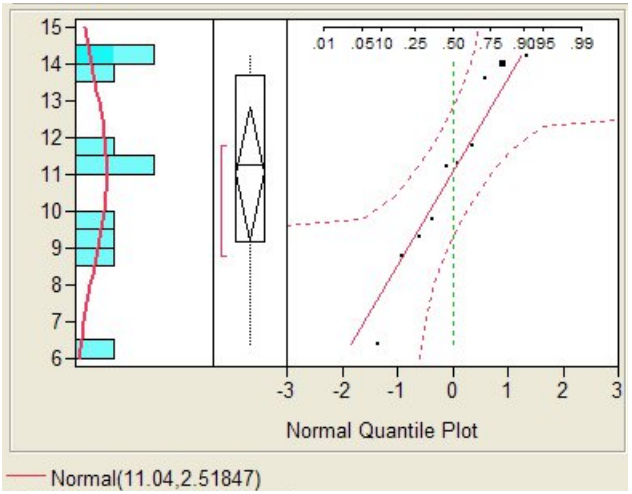


Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.962401	0.8129

Material B



Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.948149	0.6467

Fail to reject Ho.
Material samples appear normal.

Paired Comparison: Shoes Example

Step 3: State the Hypotheses

For d (m2-m1): Ho: d = 0

Ha: d ¹ 0

This becomes a 1-Sample T Test where the target is zero!

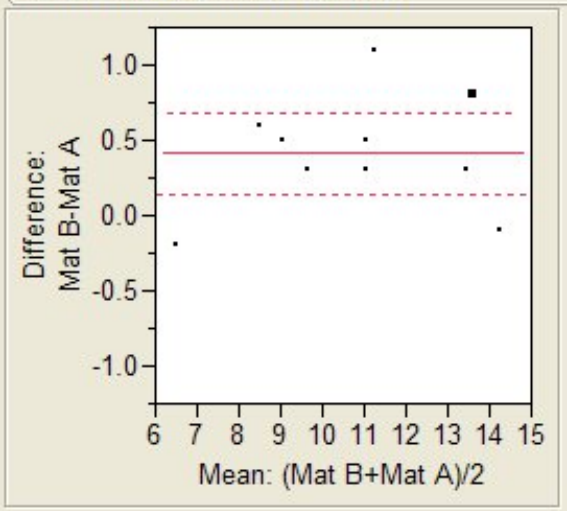
Step 4: Determine the test statistic, determine the critical value, and reject or accept the null hypothesis. (Use JMP).

- **JMP>Analyze>Specialized Modeling>Matched Pairs**
- For **Y, Paired Response** select *material A* and *material B*
- Click **OK**

Paired Comparison: Shoes Example

Matched Pairs

Difference: Mat B-Mat A



Mat B	11.04	t-Ratio	3.348877
Mat A	10.63	DF	9
Mean Difference	0.41	Prob > |t|	0.0085*
Std Error	0.12243	Prob > t	0.0043*
Upper95%	0.68695	Prob < t	0.9957
Lower95%	0.13305		
N	10		
Correlation	0.98823		

Reject Ho – the mean difference between the paired groups is significantly different from 0.

p<0.05
Significant Difference

p<0.05
Material B is significantly greater than Material A

Paired Comparison: Shoes Example

Step 5: Material A wears significantly better than material B. How can we tell?

Note: The boys were homogeneous within groups but different between. Some boys were more active than others.

NC STATE UNIVERSITY

Lean Six Sigma TE/TTM/TT

533

Correlation & Regression

Correlation & Regression

Objective: To use correlation & regression tools to narrow the list of continuous input variables.

Deliverables: Correlation/Regression analysis; updated input list

Correlation & Regression

Assumptions :

Input – continuous variables
Output – continuous variables

Definitions

- **Correlation**: A technique to "quantify" the strength of association between a variable output and a variable input via the *correlation coefficient* = *r*.
- **Regression Equation**: A prediction equation, not necessarily linear, which allows the values of inputs to be used to predict a corresponding output.
- **Coefficient of Determination**: r^2, represents the adequacy of the regression model or the amount of variation explained by the regression equation.

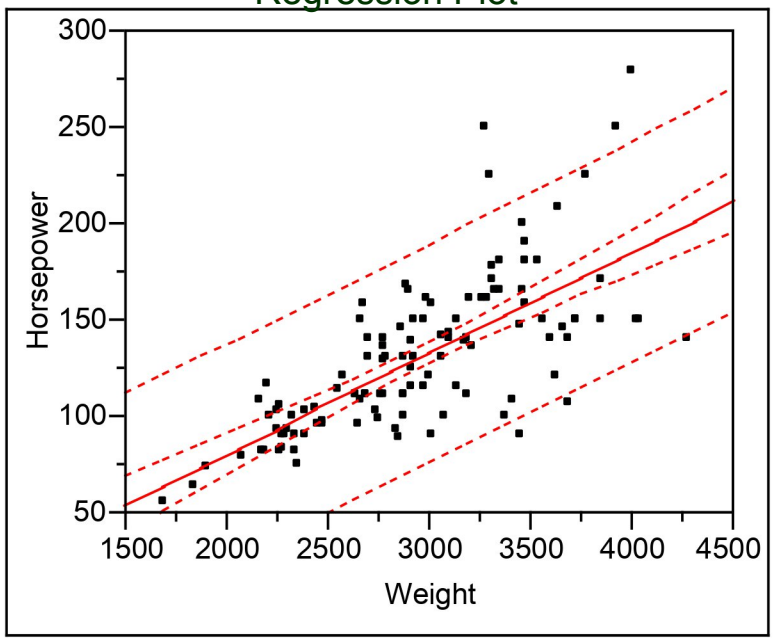
Juran 44.95

Why Do We Use These Tools?

- Gather meaningful data about a process without interruption.
- This type of study leaves the process in a natural operating state (without DOE) versus artificially inducing variation (with DOE).
- Correlation provides a graphical analysis and quantifier of the strength between independent and dependent variables.
- Regression can be used as a means of obtaining prediction equations so independent variables can be controlled.

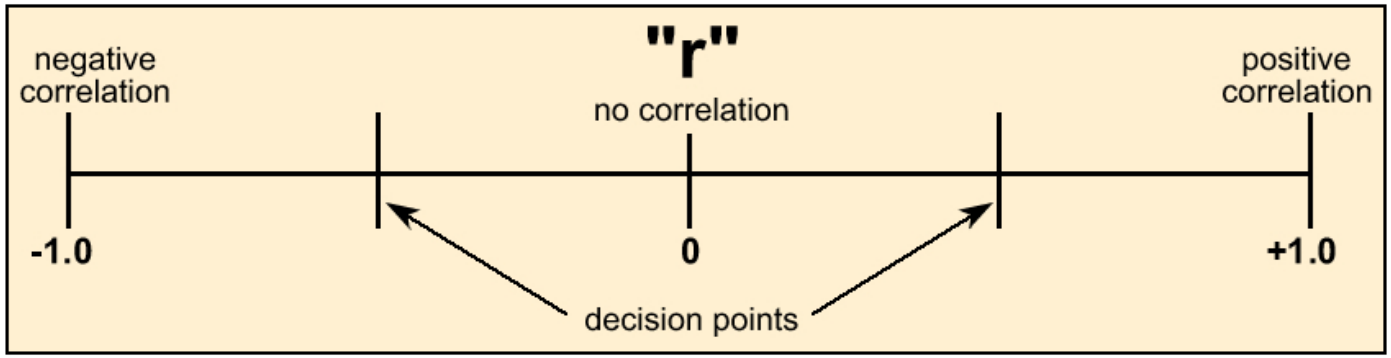
All Possible X's

Probable X's



Correlation

- Correlation is a measure of strength of association between two quantitative variables (e.g., pressure and yield) which aids in establishing Y = F (X).
- Correlation measures the strength of the relationship between two variables using the correlation coefficient, r.
- The correlation coefficient, r, will always be between -1 and +1.
- Guideline (usually based on sample size):
  - If | r | > 0.80, then relationship is significant
  - If | r | < 0.20, then relationship is not significant



*NOTE: JMP uses the Pearson's formula

Correlation

- Table for decision points in determining the correlation (positive or negative) for different sample sizes of n:

n	Decision point	n	Decision point
5	0.878	18	0.468
6	0.811	19	0.456
7	0.754	20	0.444
8	0.707	22	0.423
9	0.666	24	0.404
10	0.632	26	0.388
11	0.602	28	0.374
12	0.576	30	0.361
13	0.553	40	0.312
14	0.532	50	0.279
15	0.514	60	0.254
16	0.497	80	0.22
17	0.482	100	0.196

Data Requirements

- To conduct a correlation study, one must have:
- **Bivariate Data**: data from two variables from the same object/person
- Bivariate data is made up of ordered pairs

(Factor)	**(Response)**
X (input)	**Y (output)**
x1	y1
x2	y2
x3	y3
x4	y4
x5	y5
x_nth	y_nth

Correlation

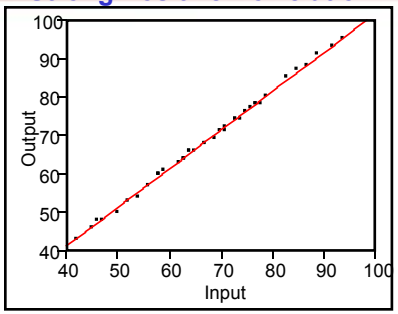
The correlation coefficient (r):

- Always falls between -1 and +1
- Is a positive value – as the value of one variable increases, so does the other.
- Is a negative value – as the value of one variable increases, the other decreases.

The Correlation Formula:

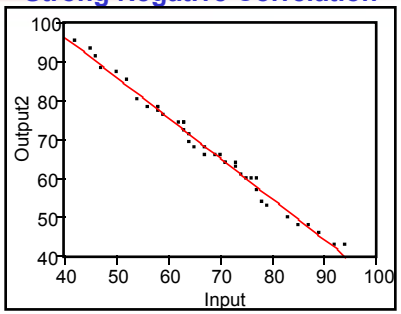
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Strong Positive Correlation



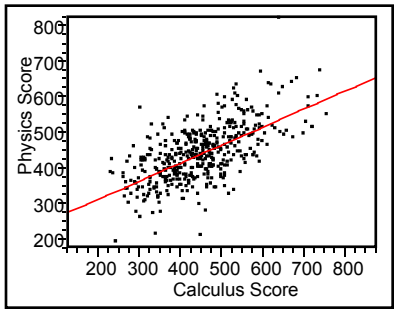
RSquare 0.997691

Strong Negative Correlation



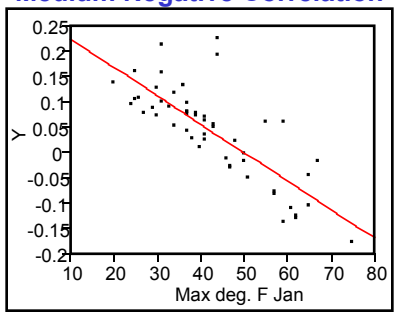
RSquare 0.988842

Medium Positive Correlation



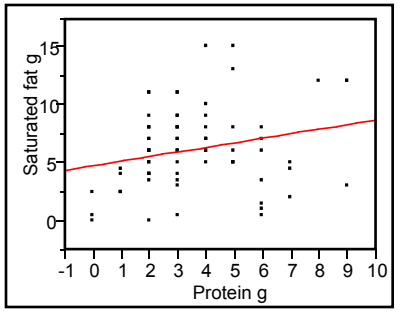
RSquare 0.392685

Medium Negative Correlation



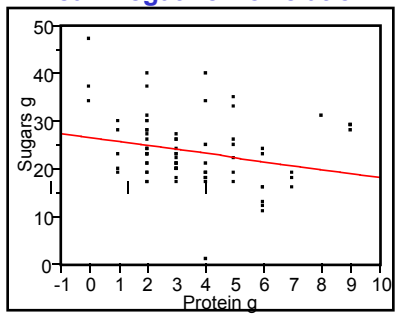
RSquare 0.631387

Weak Positive Correlation



RSquare 0.058543

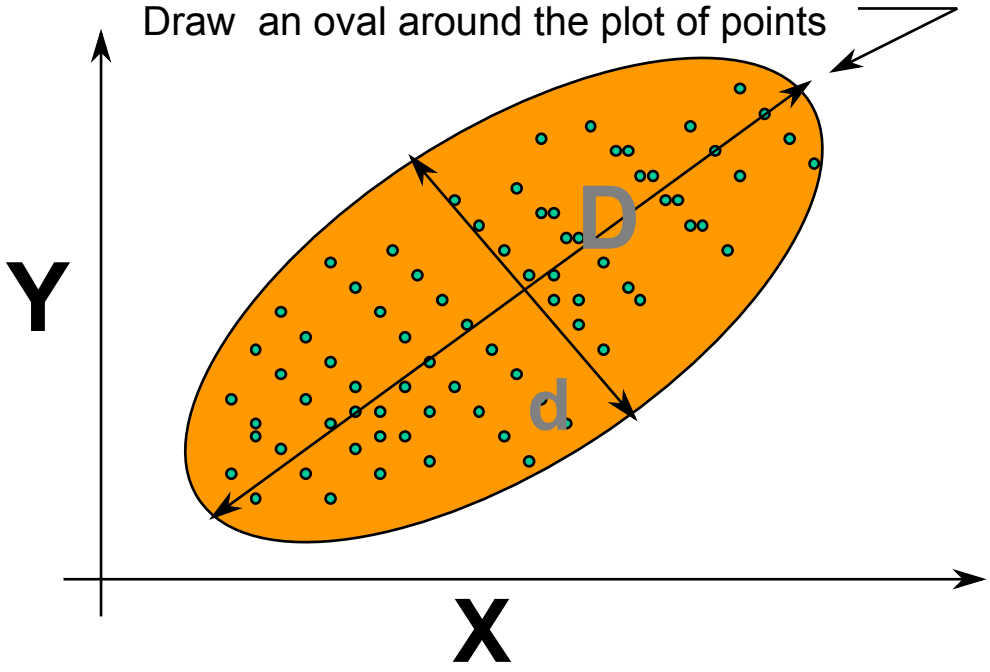
Weak Negative Correlation



RSquare 0.058026

Estimating the Correlation Coefficient

1. Measure the length of the MAX diameter (D) of the oval with a scale
2. Measure the length of the MIN diameter (d) of the oval with a scale
3. Estimate the value of " r " by calculating: ± (1- [d/D])
4. Attach the sign in the direction of the slope of D



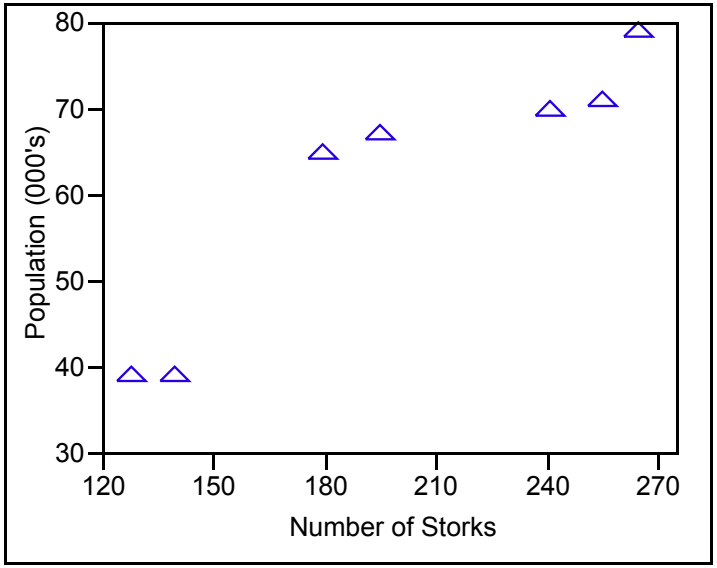
Abuse and Misuse of Correlation

- If we establish a correlation between y and x1, that does NOT necessarily mean variation in x1 caused variation in y.
- A third variable may be 'lurking' that causes both x1 and y to vary.
- In conclusion, an association between two variables does NOT mean there is a cause-and-effect relationship.

Correlation does NOT determine causation!

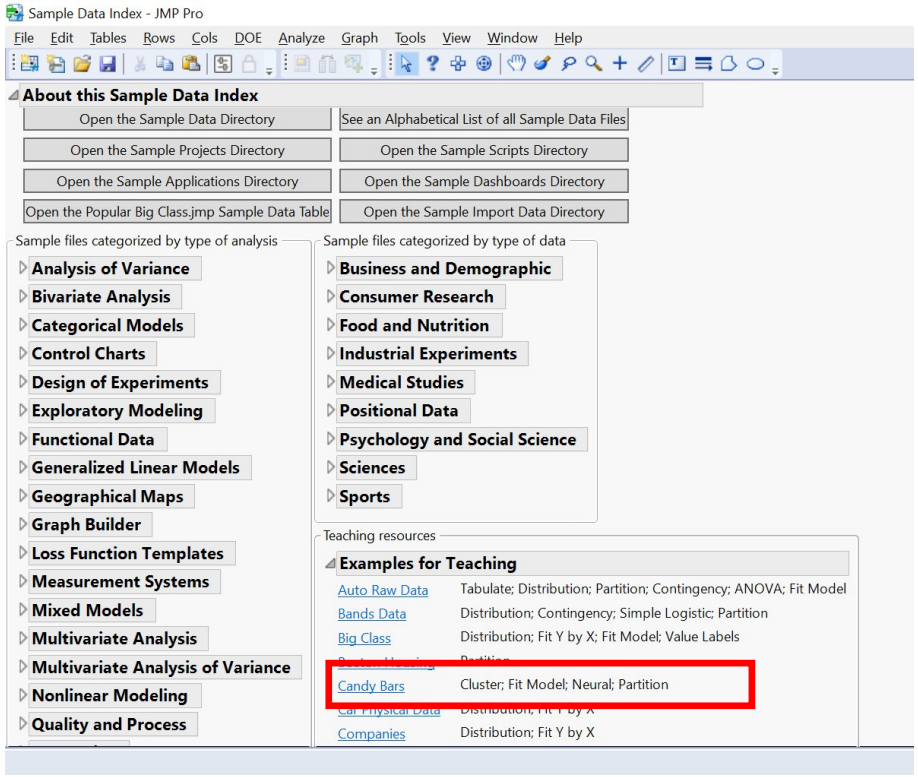
Stork Example

- **Correlation does not imply causation.**
- Look at the graph of population and storks below.
- Question: Would killing storks be an adequate method of birth control?
- We may identify a relationship by observing a process; two variables tend to increase together and decrease together. However, this does not necessarily mean that we can adjust one variable by manipulating the other variable.



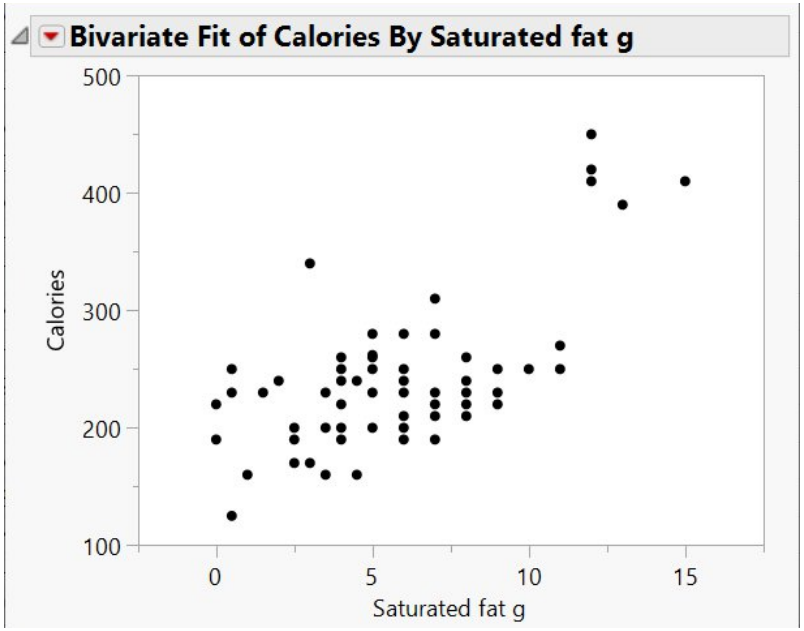
Correlation Example 1

- Open the data table **JMP>Help>Sample Index> Examples for Teaching> Candy Bars.jmp**



Correlation Example 1

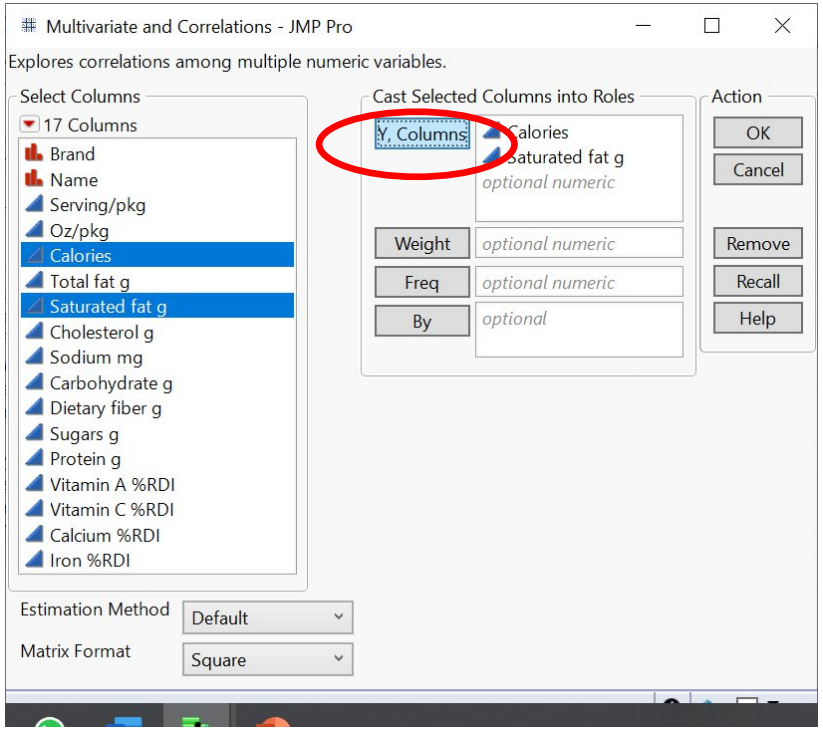
- Plot the data first
  - **JMP>Analyze>Fit Y By X**
  - For **Y, Response** select *Calories*
  - For **X, Factor** select *Saturated fat g*
  - Click **OK**



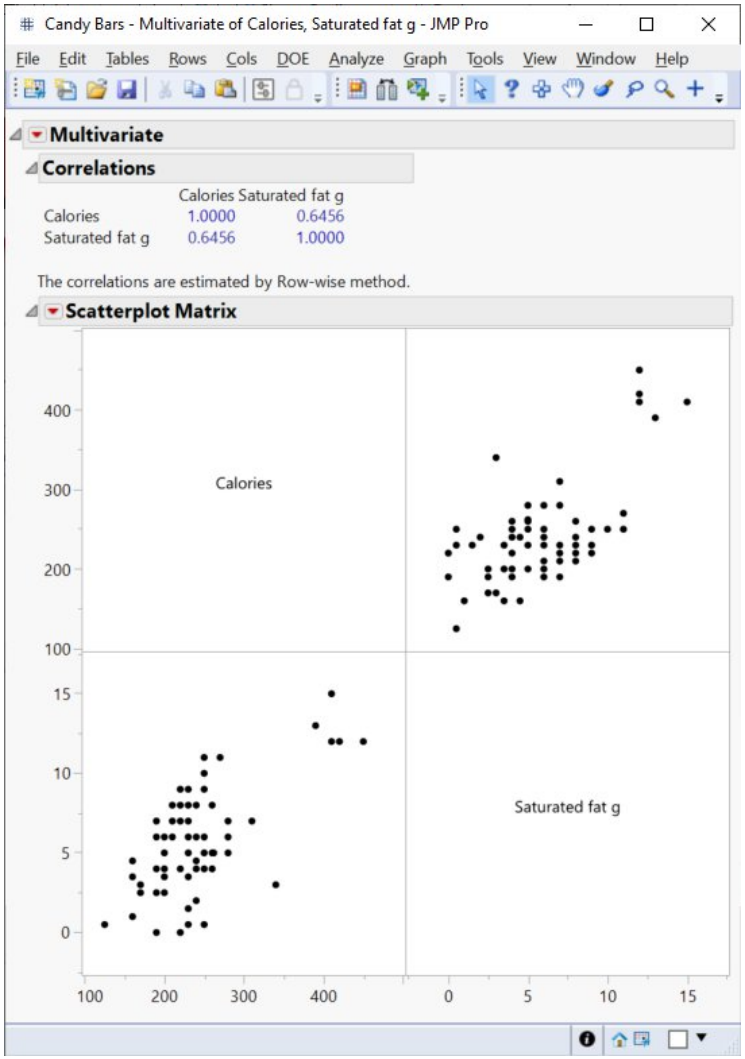
Correlation Example 1

Run the Correlation Analysis

- **Analyze>Multivariate Methods>Multivariate**
- For **Y, Columns**, choose *Calories* and then *Saturated fat g*
- Click **OK**

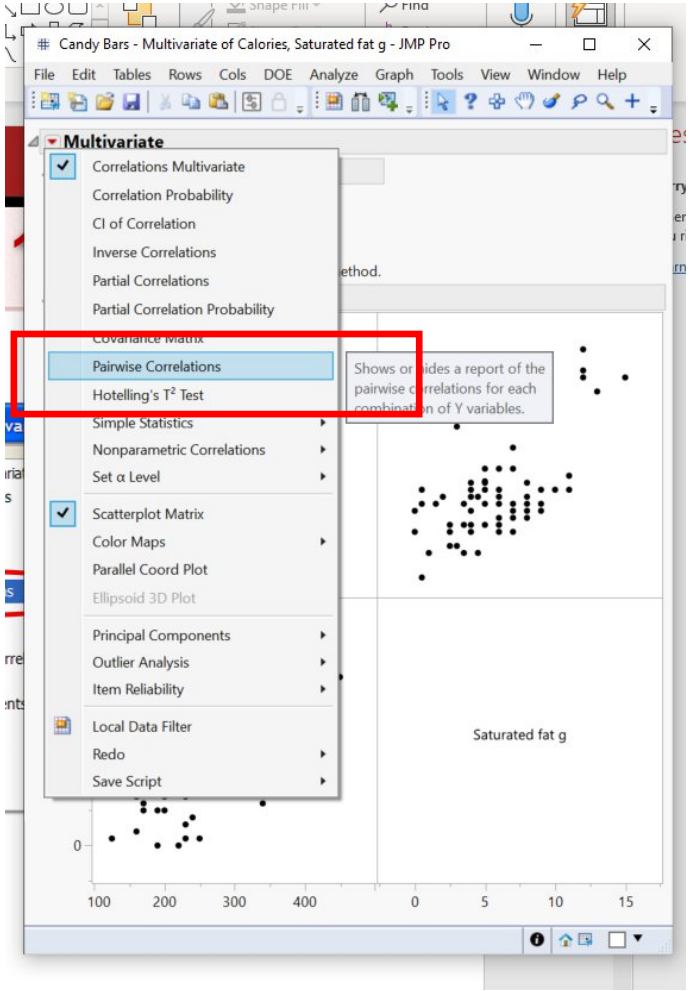


Correlation Example 1



Correlation Example 1

- Click-on the **Multivariate Red Triangle** and select **Pairwise Correlations** to see the p-values.



Correlation Example 1

Pairwise Correlations

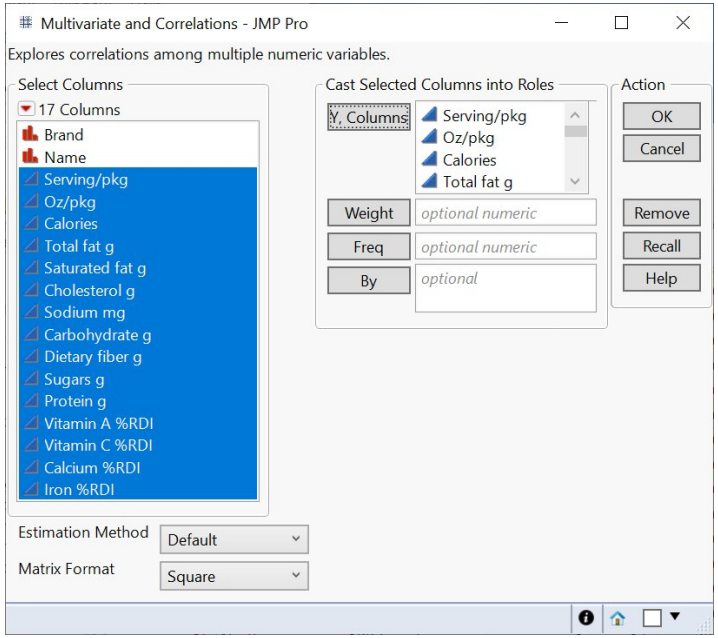
Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
Saturated fat g	Calories	0.6456	75	0.4906	0.7611	<.0001*	

Graphical picture of Correlation (r = 0.6456)

- Are the two variables related? What is r?
- What are your conclusions?

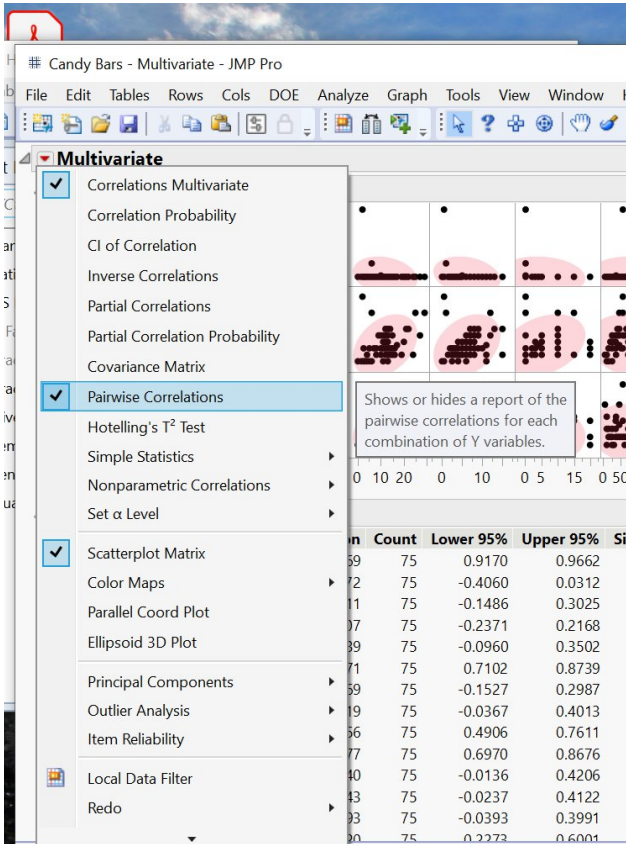
Correlation Example 1

- **Analyze>Multivariate Methods>Multivariate**
- For **Y,Columns**, Choose *Servings/pkg* through *Iron %RDI*
- Click **OK**
- All the different combinations of the correlation graphs for the selected variables are generated in the output window.



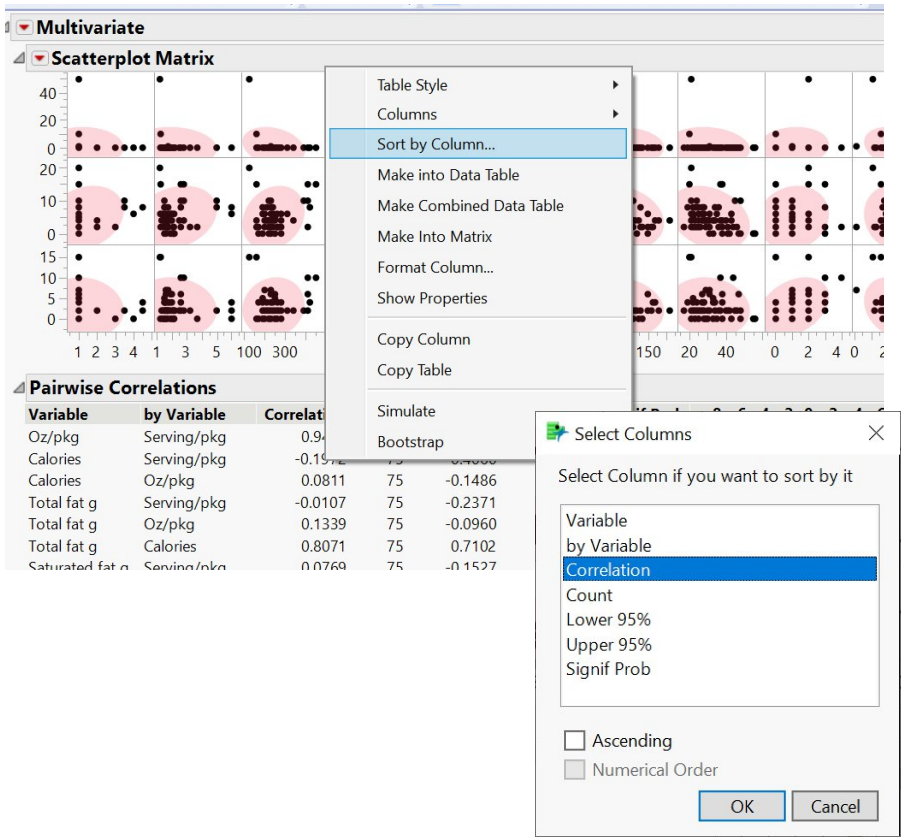
Correlation Example 1

- **Click on red diamond in Multivariate, chose**
- **Multivariate> Pairwise Correlations**
- Analyze all combinations of Scatterplots and the Correlation Coefficients



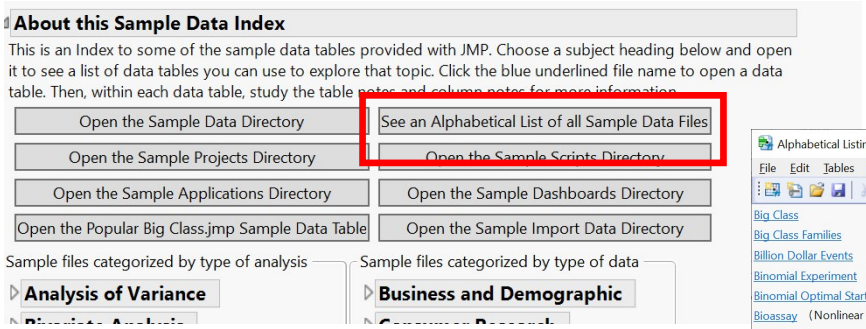
Correlation Example 1

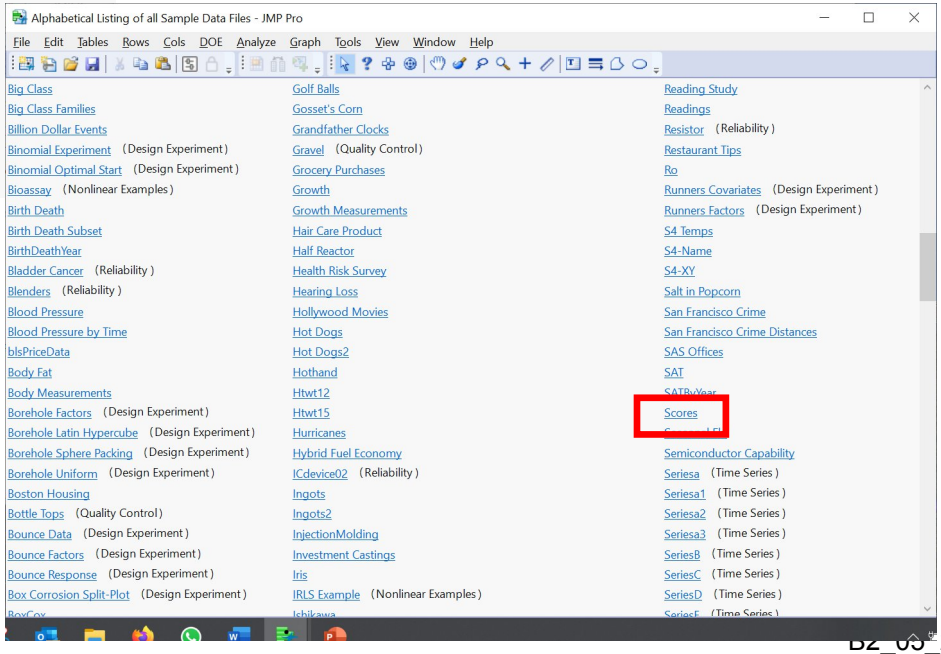
- TIP: You can order the correlation values by right-clicking on the Pairwise Correlations table and selecting **Sort by Column**. Then select the "Correlation" column
- The table is rank ordered by correlation coefficient from +1 to -1.



Correlation Example 2

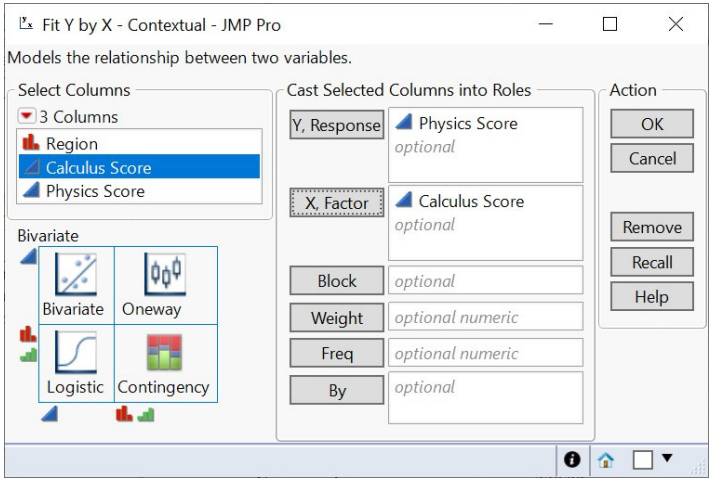
- Open the menu **JMP>Help>Sample Index**
- Click on See an **Alphabetical List of all Sample Data Files**
- Scroll & Select **Scores.jmp**

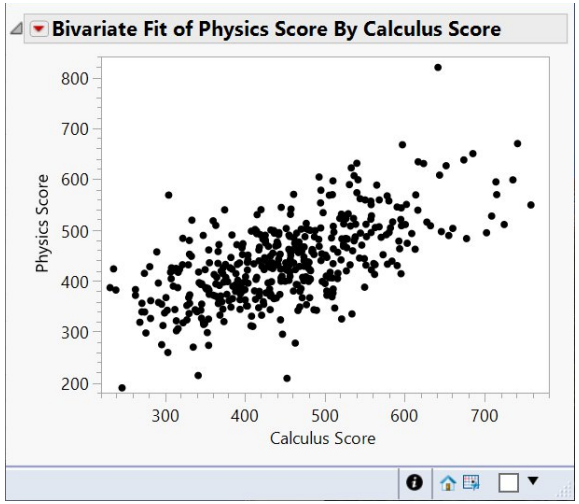




Correlation Example 2

- Always plot the data first
  - **JMP>Analyze>Fit Y by X**
  - For **Y, Response** select *Physics Score*
  - For **X, Factor** select *Calculus Score*
  - Click **OK**

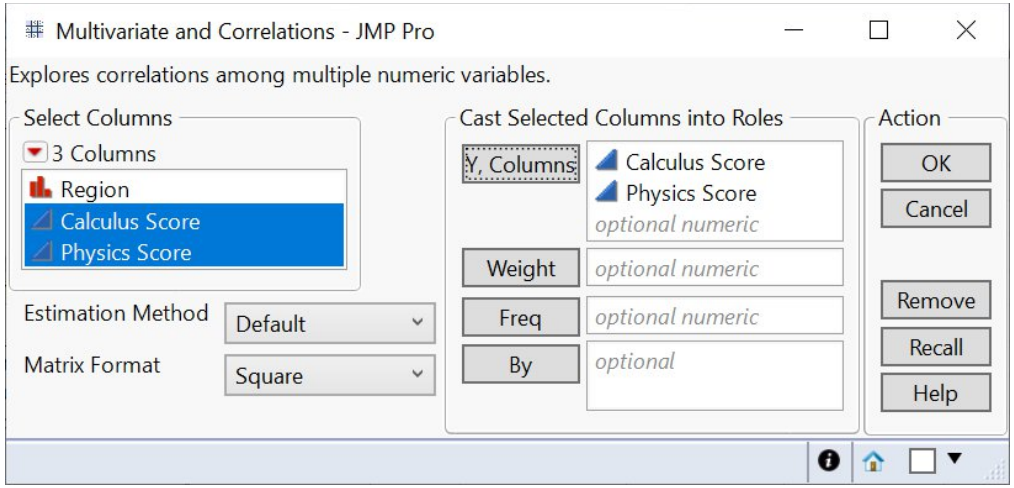




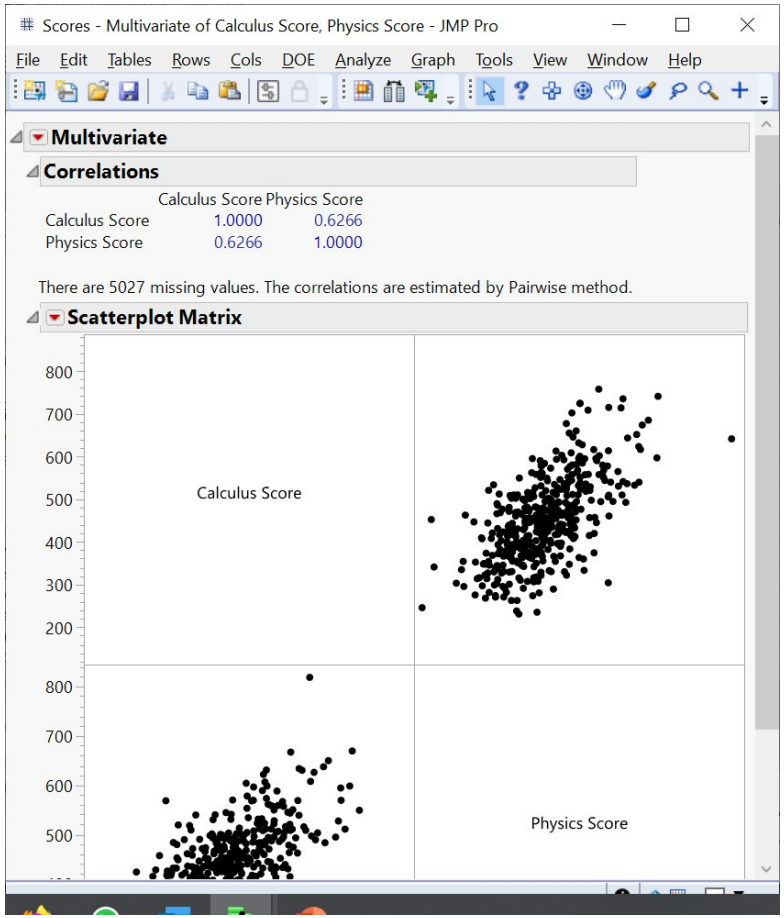
Correlation Example 2

Run the Correlation Analysis

- **Analyze>Multivariate Methods>Multivariate**
- For **Y, Columns**, choose *Calculus Score* and then *Physics Score*
- Click **OK**

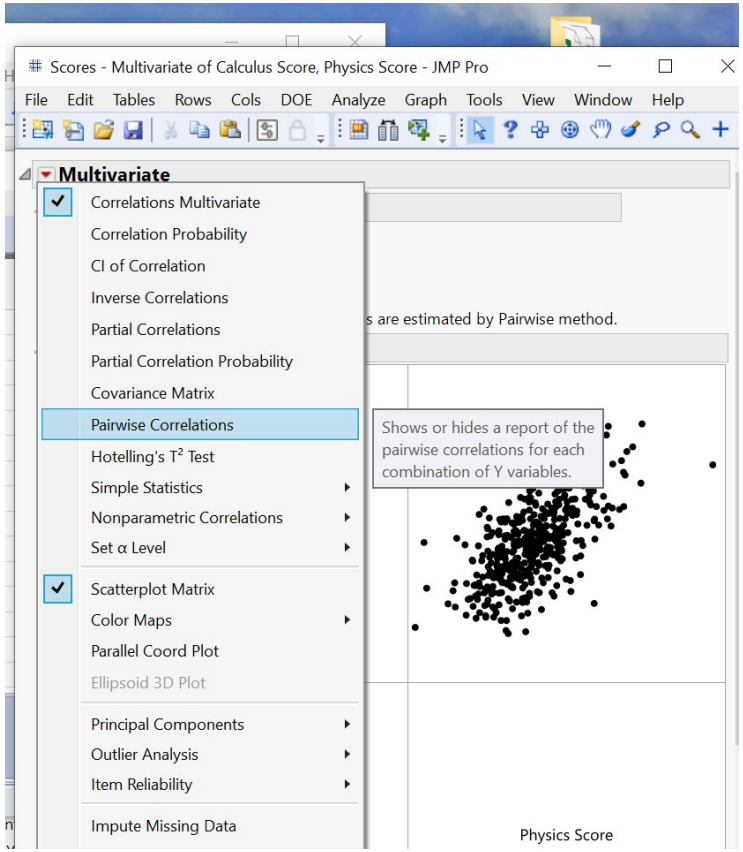


Correlation Example 2



Correlation Example 2

- Click-on the **Multivariate Red Triangle** and select **Pairwise Correlations** to see the p-values.



Correlation Example 2

Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
Physics Score	Calculus Score	0.6266	436	0.5660	0.6805	<.0001*	

- Are the two variables related? What is r?
- What are your conclusions?

Regression Analysis

- Correlation tells us the strength of a relationship, not the exact numerical relationship.
- The next step for analyzing continuous data is the determination of the regression equation.
- Regression analysis calculates a "prediction equation" which can mathematically predict Y for any given X.
- The primary objective of regression analysis is to make **PREDICTIONS**.
- The regression equation is simply the one that **BEST FITS** the plotted data.
- Examples of prediction equations:

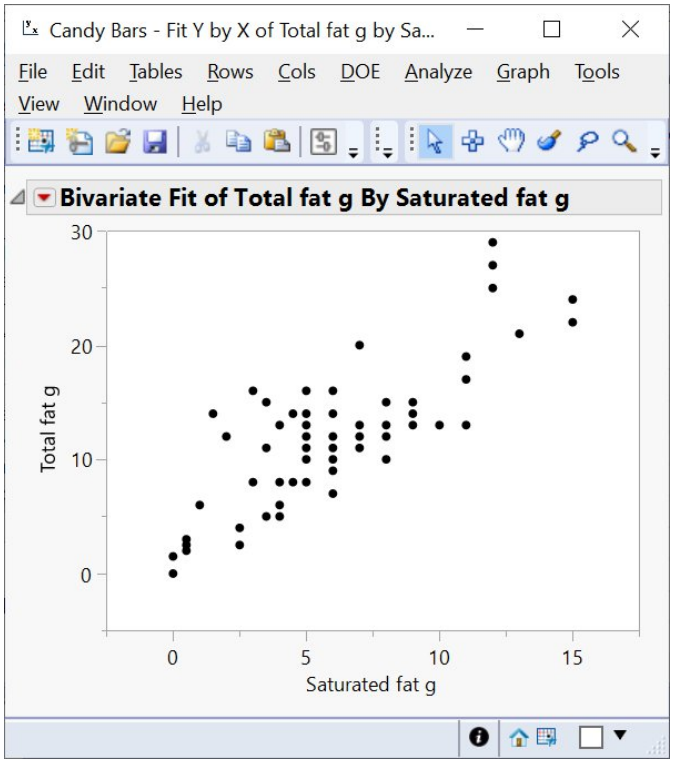
$Y = a + b\,x$	(linear model)
$Y = a + b\,x + c\,x^2$	(with quadratic term)
$Y = a + b\,x + c\,x^2 + d\,x^3$	(with cubic term)
$Y = a\,(b^x)$	(exponential model)

Coefficient of Determination, R-Squared

- The output from the fitted line plot contains an equation which relates the predictor (input variable) to the response (output variable).
- The **R-sq.** value is the square of the correlation coefficient. It is also the **fraction of the variation in the output (response) variable that is explained by the equation**.
- What is a good value? It depends on the process and the industry. For example, a chemist may require an R-sq of 0.99. However, the fact that one input variable may account for 65% of the variation in your final product may be phenomenal too!

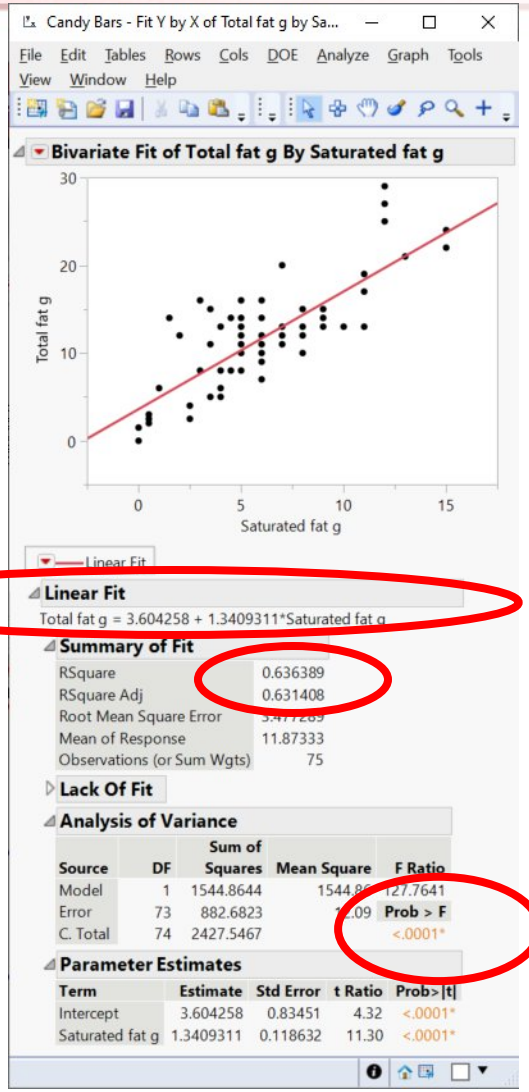
Regression Example (Fitted Line Plot)

- Open the data table **JMP>Help>Sample Index>Examples for Teaching>Candy Bars.jmp**
- Plot the data first
  - **JMP>Analyze>Fit Y by X**
  - For **Y, Response** select *Total Fat g*
  - For **X, Factor** select *Saturated Fat g*
  - Click **OK**



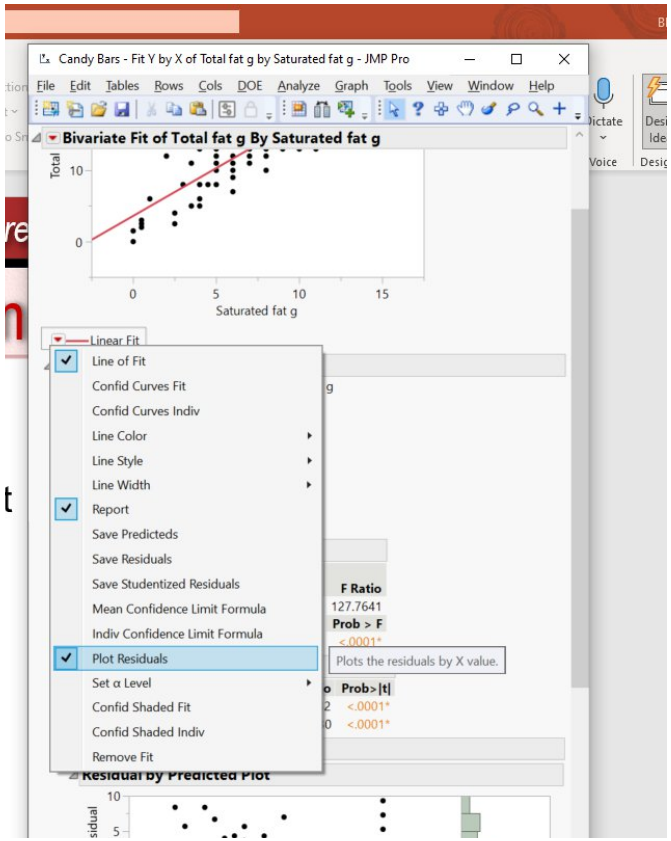
Regression Example (Fitted Line Plot)

- Click-on the **Bivariate Fit of Total Fat by Saturated Fat** and select **Fit Line**.
- Note the Prediction Equation and the RSquare value.
- 63.6% of the variation in the data can be explained by the model.



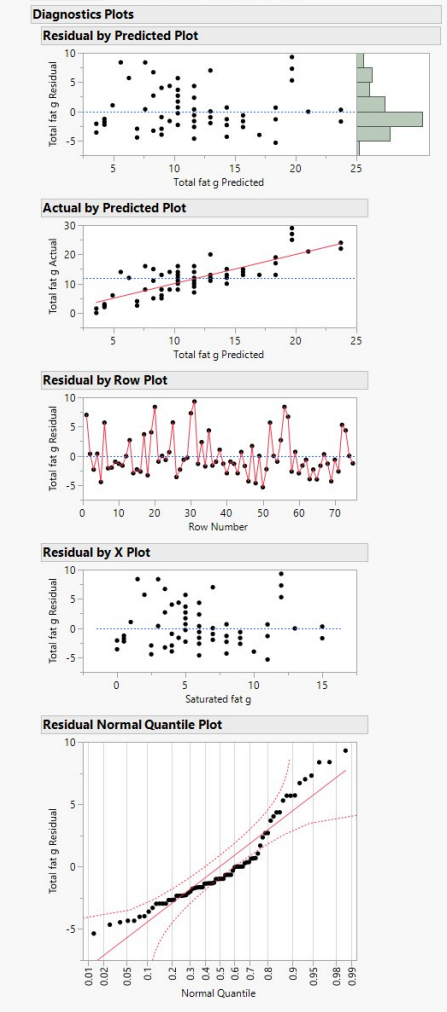
Regression Example (Fitted Line Plot)

- Examine the Residuals
- Click-on the **Linear Fit Red Triangle** and select **Plot Residuals**

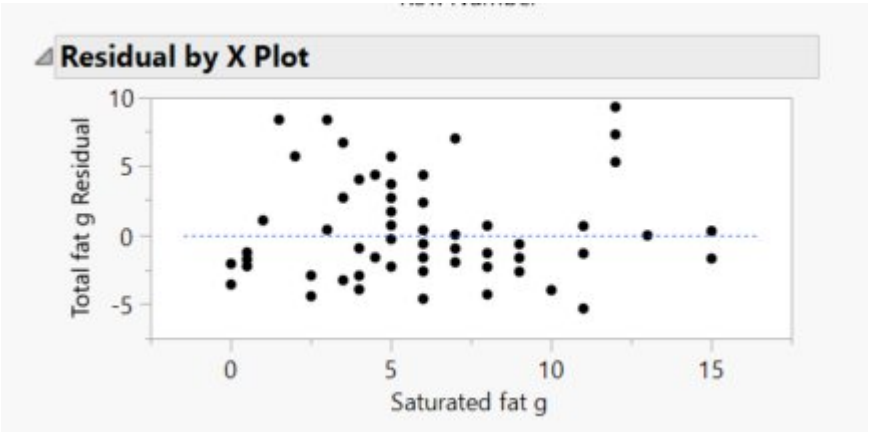


Regression Example (Residuals)

- What do residuals show you?
- What does residual by predicted plot mean?
- What does the Normal Quantile Plot tell you?



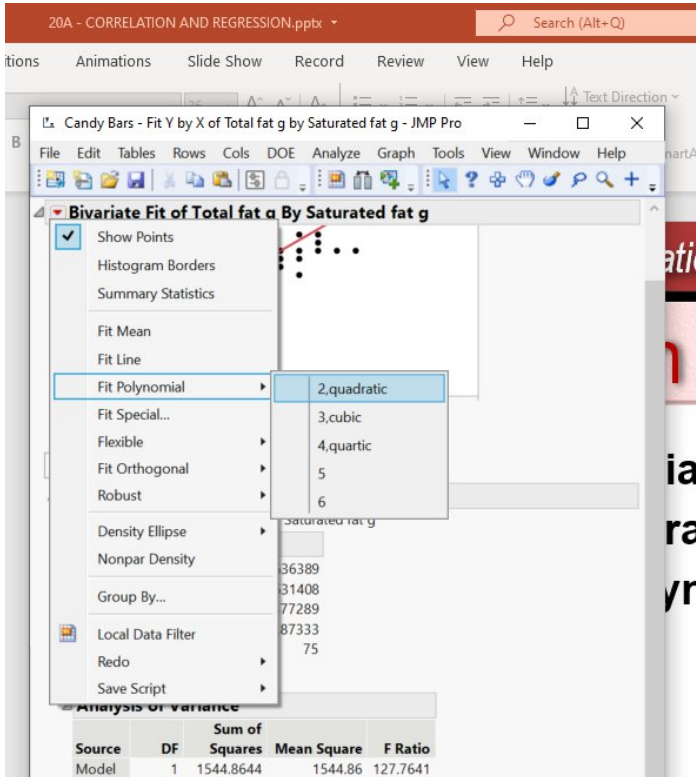
Regression Example (Fitted Line Plot)



- Obtain a prediction equation:
  - Do you need a higher order model?
  - Investigate the residuals plot.

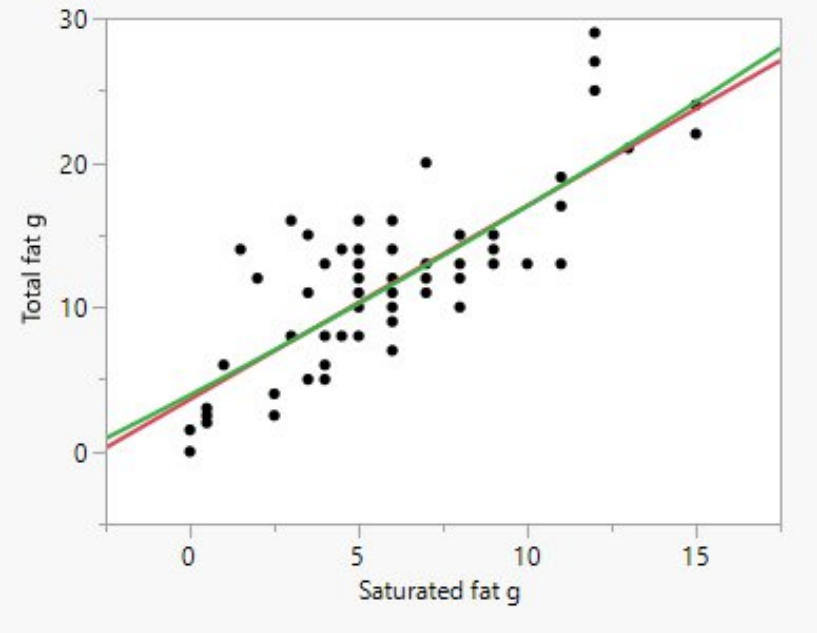
Regression Example (Fitted Line Plot)

- Click-on the **Bivariate Fit of Total Fat by Saturated Fat** and select **Fit Polynomial> 2,quadratic**.



Regression Example (Fitted Line Plot)

- Note that the Quadratic Model is not an improvement over the linear model.



Regression Example (Fitted Line Plot)

- The RSquare value for the quadratic model is 0.636941.
- The squared term in the model does not improve the model, and it is not significant as shown by the p>value of .7417

Polynomial Fit Degree=2

Total fat g = 3.5810782 + 1.3285343*Saturated fat g + 0.0086968*(Saturated fat g-6.16667)^2

Summary of Fit

RSquare	0.636941
RSquare Adj	0.626856
Root Mean Square Error	3.498695
Mean of Response	11.87333
Observations (or Sum Wgts)	75

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1546.2044	773.102	63.1575
Error	72	881.3422	12.241	Prob > F
C. Total	74	2427.5467		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.5810782	0.842565	4.25	<.0001*
Saturated fat g	1.3285343	0.125105	10.62	<.0001*
(Saturated fat g-6.16667)^2	0.0086968	0.026285	0.33	0.7417

Class Examples

Use correlation/regression to analyze the Candy Bars.jmp file. Does the Total Fat, X, predict the Calories, Y, in the candy bars?

- Are the two variables related?
- What is "r"? What is R-Sq?
- What are your conclusions for each analysis?
- Try a polynomial of degree = 2.
- Try a polynomial of degree = 3.
- Do either improve the R-Sq?

Summary

- Correlation is a very useful tool in the process industry.
- Correlation is the measure of the **relationship** between two quantitative variables.
- Correlation does NOT determine causation!
- Regression analysis seeks to find a relationship between the variables in the form of a prediction equation which may or may not be linear.
- In regression, the equation may be the desired answer, or it may be the means to the desired prediction.

TRUE or FALSE Question:

When the calculated value of r is between the table value of r and 1 (positive or negative), have we established a cause- and-effect relationship between the two variables?

NC STATE UNIVERSITY

Lean Six Sigma TE/TTM/TT

533

Correlation & Regression

Correlation & Regression

Objective: To use correlation & regression tools to narrow the list of continuous input variables.

Deliverables: Correlation/Regression analysis; updated input list

Correlation & Regression

Assumptions :

Input – continuous variables
Output – continuous variables

Definitions

- **Correlation**: A technique to "quantify" the strength of association between a variable output and a variable input via the *correlation coefficient* = *r*.
- **Regression Equation**: A prediction equation, not necessarily linear, which allows the values of inputs to be used to predict a corresponding output.
- **Coefficient of Determination**: r^2, represents the adequacy of the regression model or the amount of variation explained by the regression equation.

Juran 44.95