
# SIX SIGMA techniques – part 1

**Prof. Alberto Portioli Staudacher**
**Dipartimento Ing. Gestionale**
**Politecnico di Milano**
**Dep. Management, Economics and Industrial Engineering**
**Alberto.portioli@polimi.it**

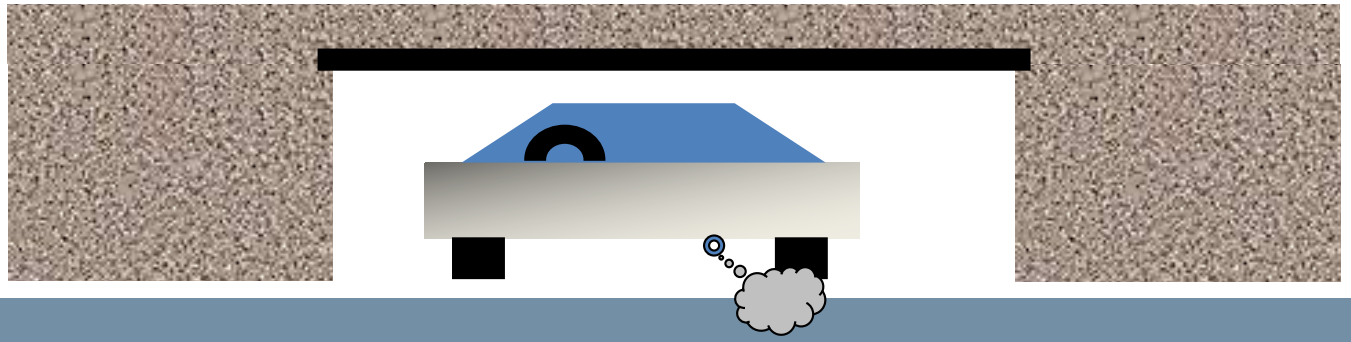
# Process Capability

Quantifiable comparison of Voice of Customer
(spec limits) to Voice of the Process (control limits).

Most measures have some target value and acceptable limits of variation around the target: for example, Viscosity, Laminate flatness, weight, etc.

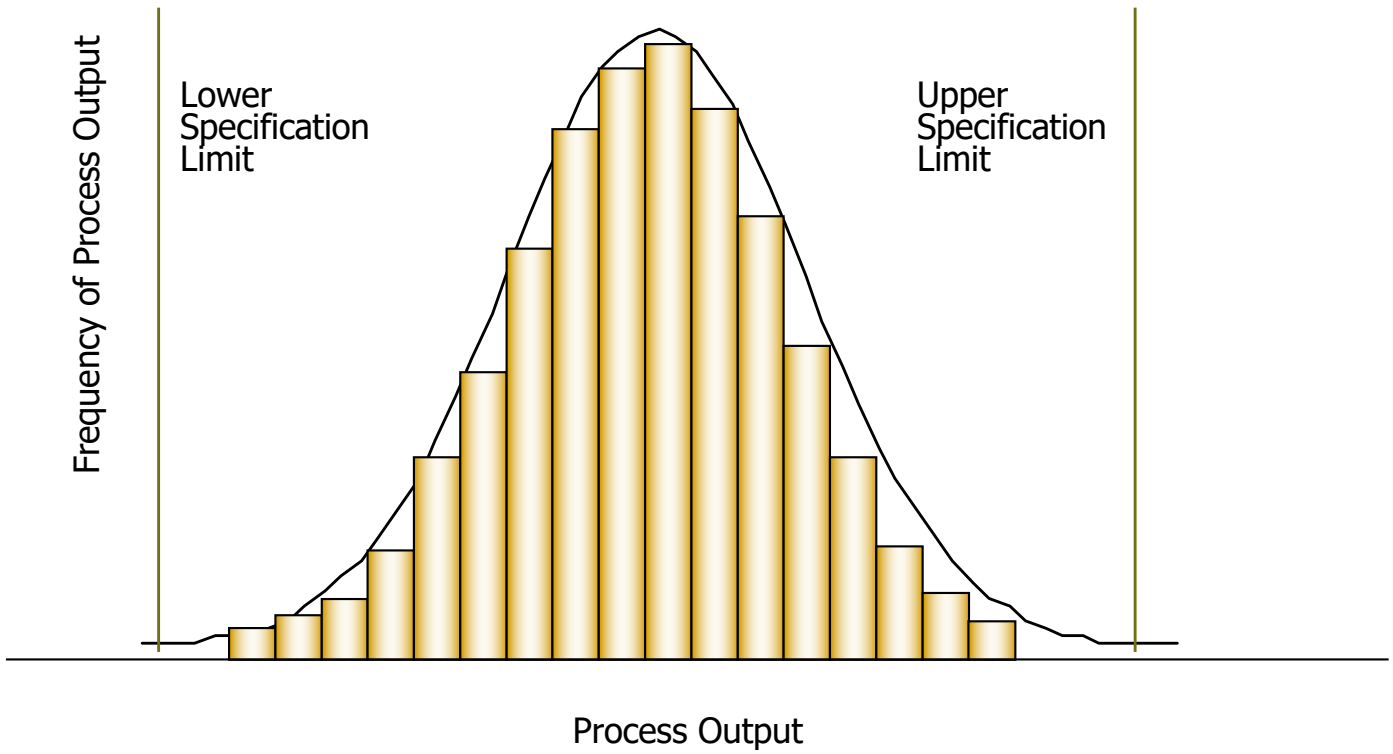
Once a process is in statistical control, that is producing consistently, you probably then want to determine if it is capable, that is meeting specification limits and producing “good” parts

You determine capability by comparing the width of the process variation with the width of the specification limits. The process needs to be in control before you assess its capability; if it is not, then you will get incorrect estimates of process capability



# Process Capability

After having build the control chart (Voice of your process) you can assess process capability graphically by drawing  the histogram against the specification limits.




These graph help you assess the distribution of your data.

You can also calculate capability indices, which are ratios of the specification limits to the natural process variation (common causes of variation).

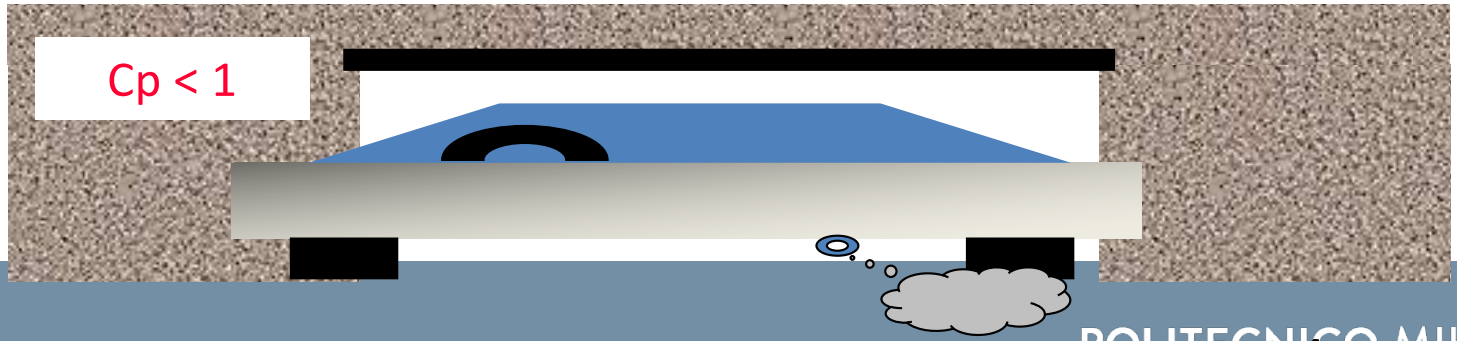
# Process Capability Ratio – Cp

Ratio of total variation allowed by the specification to the total actual variation of the process

Use $C_p$ when the mean can easily be adjusted (*i.e.*, plating, grinding, polishing, machining operations, and many transactional processes where resources can easily be added with no/minor impact on quality) AND the mean is monitored (so operator will know when adjustment is necessary – using control charts is one way of monitoring).

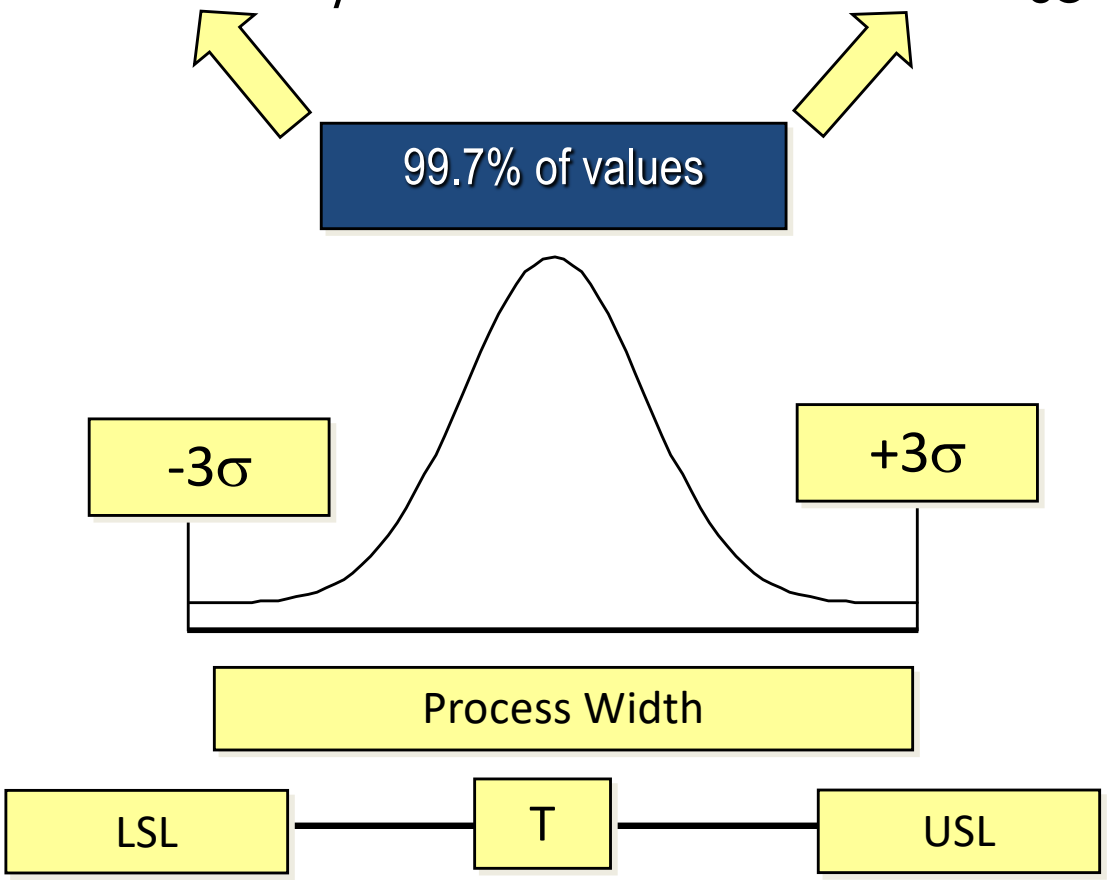
Typical goals for $C_p$ are greater than 1.33 (or 1.67 for safety items)

If $C_p < 1$ then the variability of the process is greater than the specification limits.



# Process Capability Ratio – Cp

$$C_p = \frac{\textit{Allowed variation (spec.)}}{\textit{Normal variation of the process}} \quad \text{or} \quad C_p = \frac{|USL - LSL|}{6s}$$



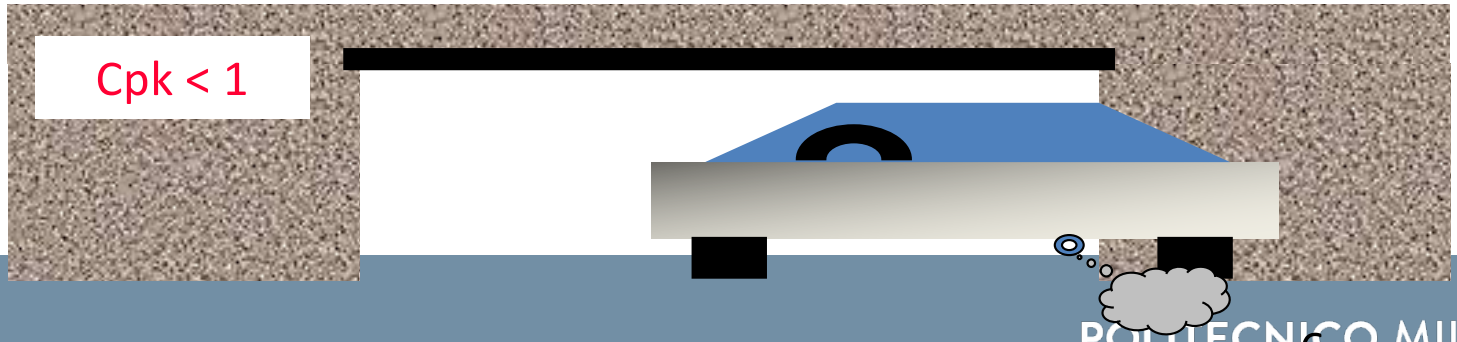
# Process Capability Ratio – Cpk

Ratio of 1/2 total variation allowed by spec. to the actual variation, with only the portion closest to a spec. limit being counted.

This index accounts for the dynamic mean shift in the process – the amount that the process is off target.

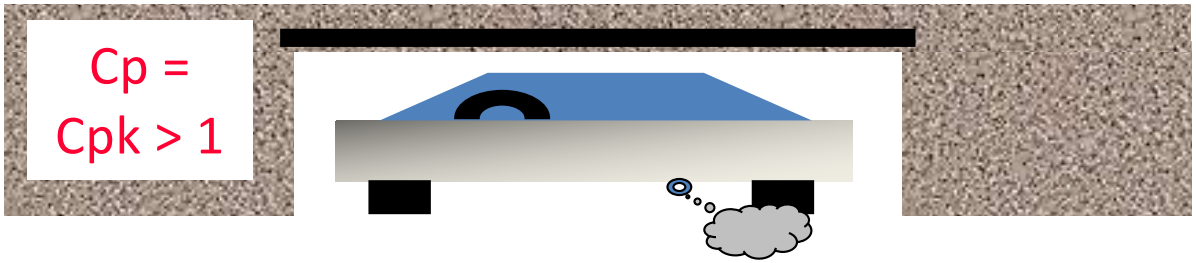
Typical goals for $C_{pk}$ are greater than 1.33 (or 1.67 if safety related)

$$C_{pk} = Min\left[\frac{USL - \bar{x}}{3s} \text{ or } \frac{\bar{x} - LSL}{3s}\right]$$

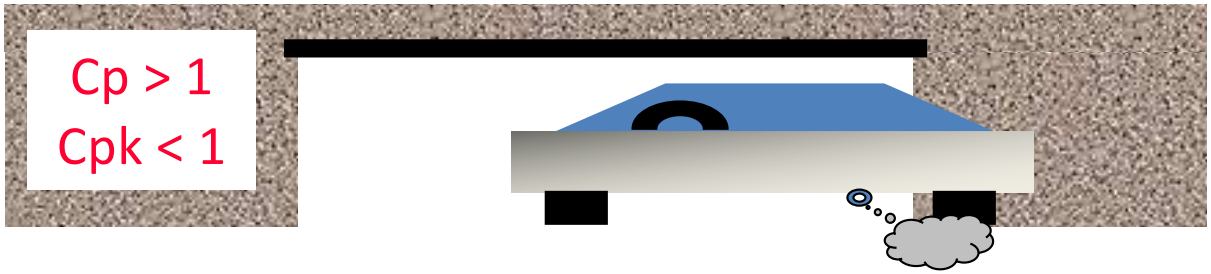


# Relationship between Cp & Cpk

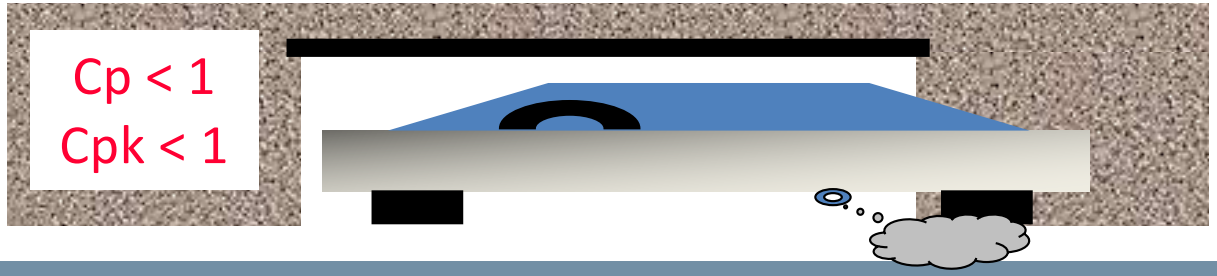
Cp = Cpk > 1, The process is centred:




Cp > 1, Cpk < 1, The process is NOT centred, but is still potentially good




Cp < 1, Cpk < 1, The process is NOT centred, and is NOT potentially good



# Uses of Capability Analysis

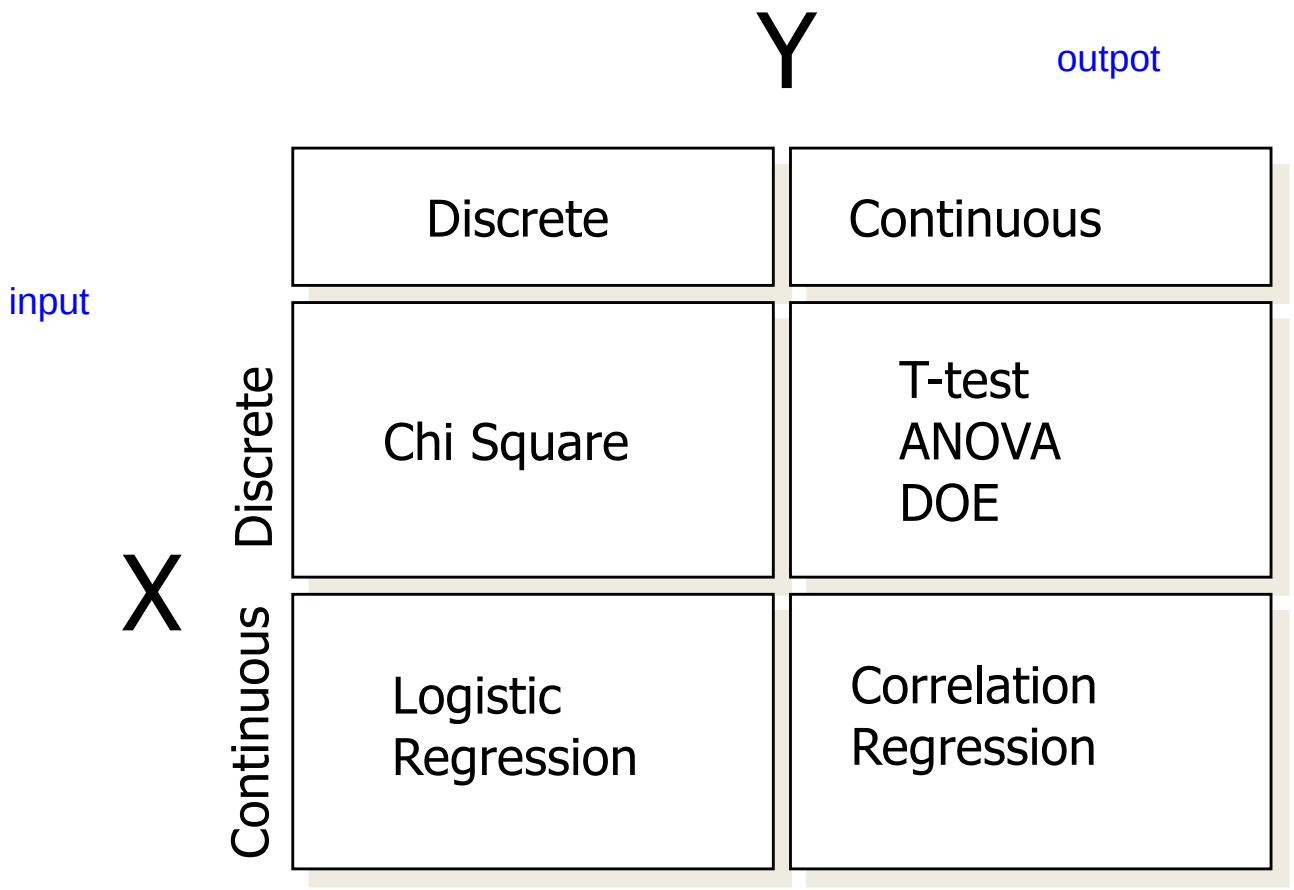
Performed on new equipment as part of the qualification and approval process.

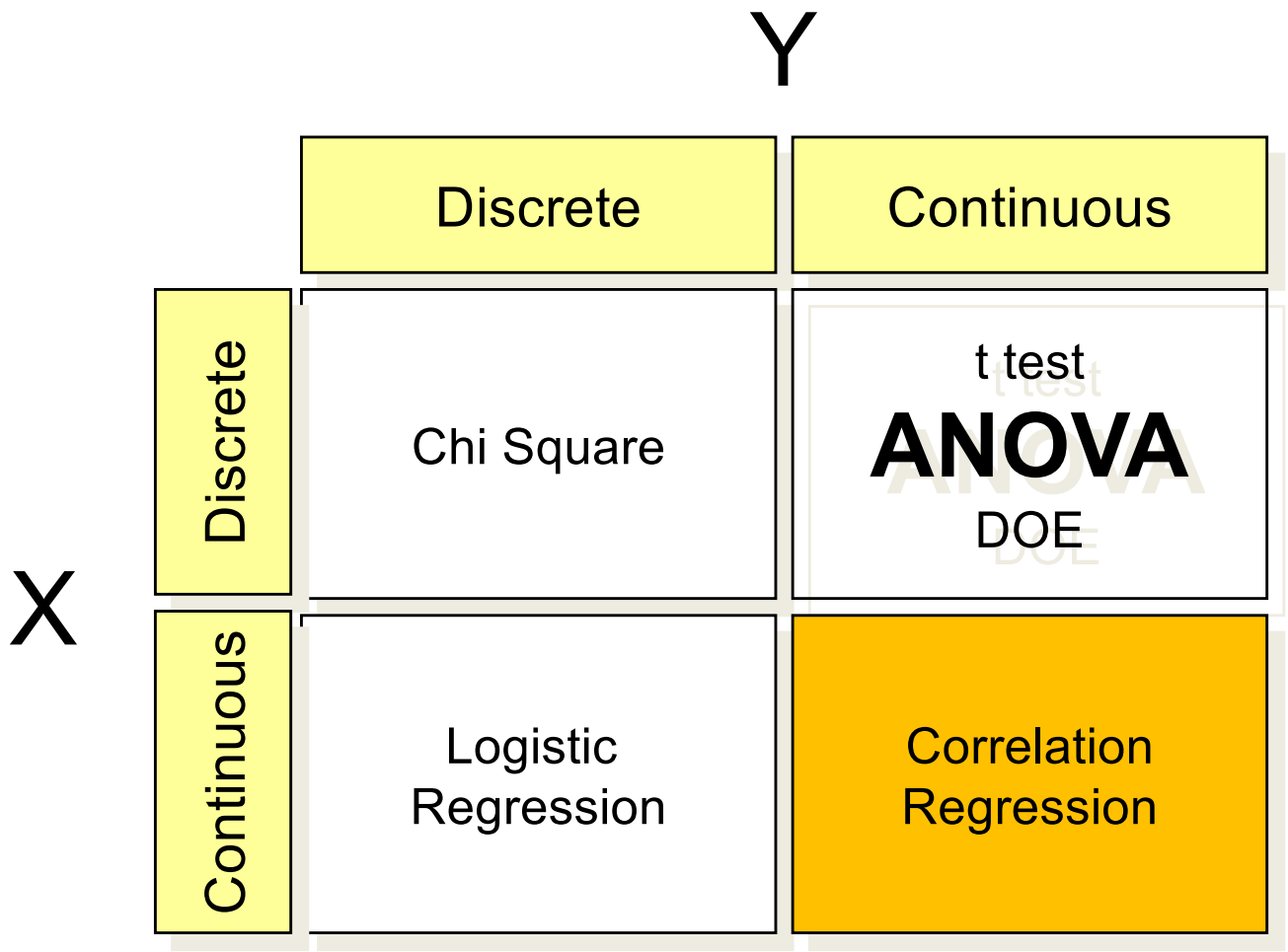
Performed on existing processes as a means of establishing a baseline of current operations, and an indicator to measure improvement.

When done periodically, is a mean of monitoring wear and tear on equipment, and deterioration of a process for whatever reason (material, personnel, environment, etc.).

Can be done on any process that has a spec. established, manufacturing or transactional (spec. is needed for the values in numerator), and has a capable measuring system (needed for valid values in denominator).

# Statistical Tools for Validating Root Causes




Y
Discrete
Continuous
X
Discrete
Continuous
Chi Square
t test
ANOVA
DOE
Logistic
Regression
Correlation
Regression

# Scatter Diagrams (Plots)

Scatter Diagrams are a graphical representation of the **relationship** between pairs of variables (factors). This relationship may be demonstrated between:

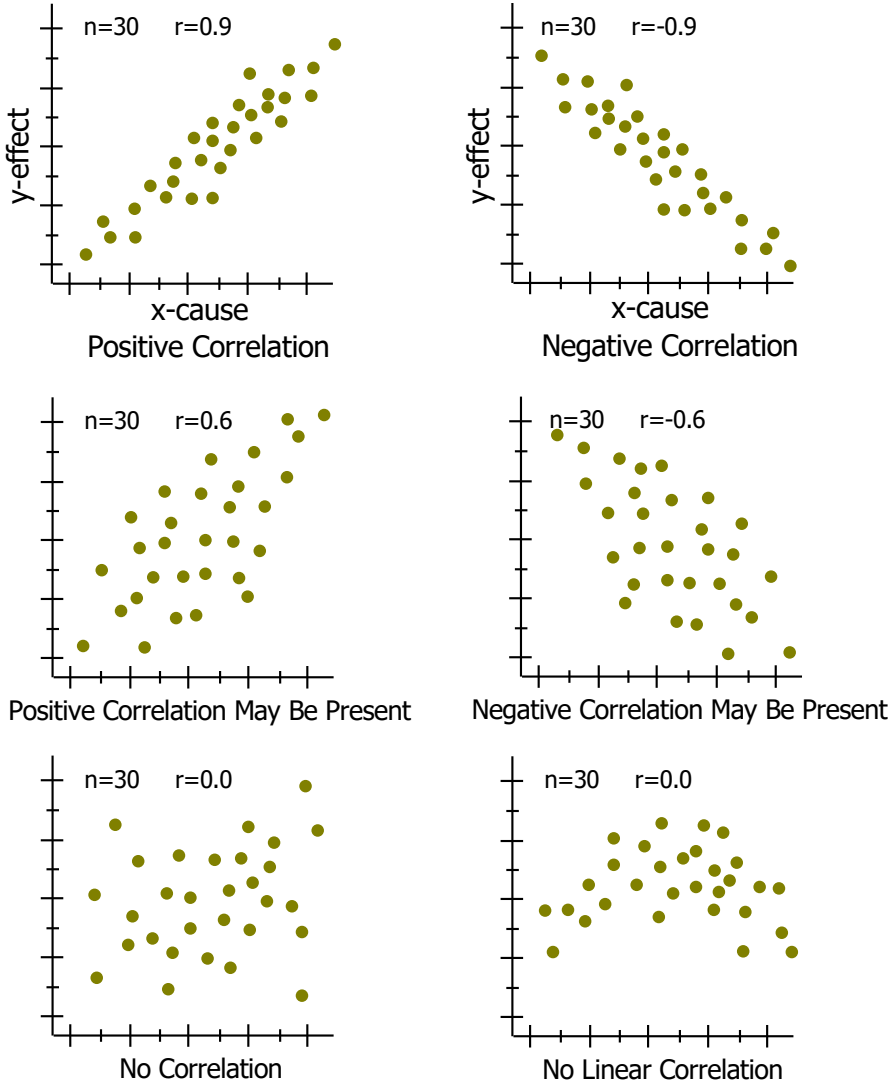
- A potential root cause (X) and the problem or output (Y)
- Two potential causes, inputs, or factors, (X1 and X2 )

For Example: To what extent does increasing advertising investments increase sales?

Scatter diagrams have the same basic format:

The horizontal axis is always an input (X). The vertical axis may be another input, or it may be the output (Y). The points are paired (X,Y) data.

# Scatter Diagram Interpretation



# Correlation Analysis

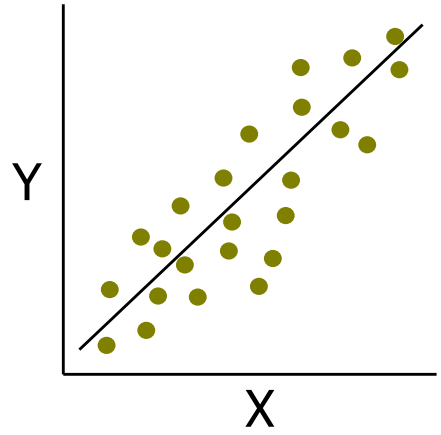
Scatter diagrams or plots provides a graphical representation of the relationship.

Correlation Analysis places a magnitude on that relationship.

- Range of correlation: -1 to +1

  - Perfect positive relationship +1
  - No relationship 0
  - Perfect negative relationship -1

What is linear regression analysis?




If we were to fit a straight line carefully between the points in the above graph, we could describe the equation for that line as follows:

$$Y=b_0+b_1X_1$$

where

$b_0$ = the predicted value of Y when X=0

$b_1$ = slope of line

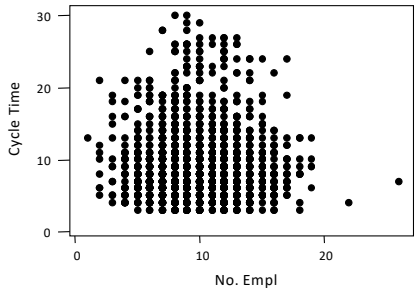
Using this equation, we could predict the value of Y given a certain value of X.

# Coefficient $R^2$

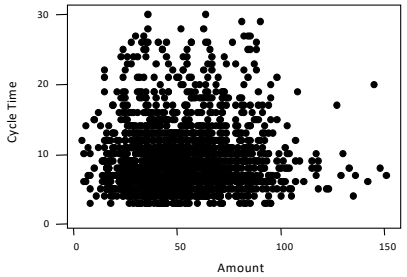
In linear regression, with one input, we used $R^2$ as a measure of what proportion of the variability in the output was explained by the input.

Obviously, the higher $R^2$, the more the response Y is explained by the input X.

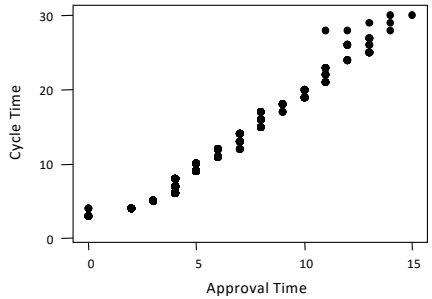
# Correlation Analysis




R=-0.023




R=0.018




R=0.978

# Multiple Linear Regression

**Multiple Regression** will sort out many potential root causes and verify or validate those that have a significant impact on the Critical To Quality /Critical Customer's Requirements or response variable.

$$Y = f(X) = X_1 + X_2 + X_3 \ldots$$

Root cause analysis leads to the underlying source of the defect, so the team can design solutions and change the process to permanently eliminate the defects.

# Multiple Regression

In Simple Linear Regression, we have:

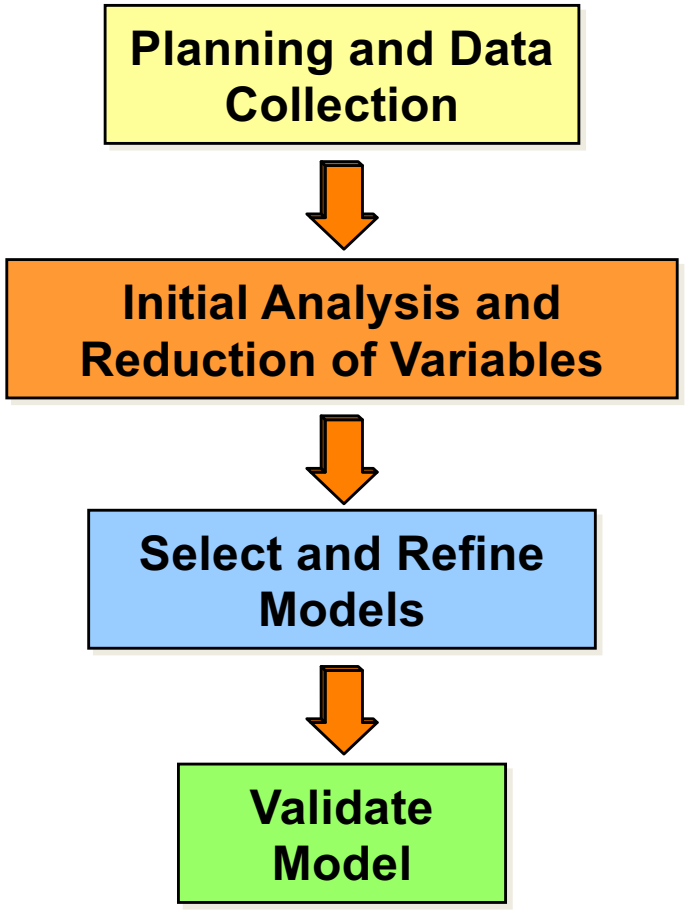
- Only one independent (predictor) variable (x)
- Example: $Y = B_0 + B_1x$

In **Multiple Linear Regression**, we have:

- More than one independent (predictor) variables ($x_1$, $x_2$, …, $x_k$)
- Example: $Y = B_0 + B_1x_1 + B_2x_2 + B_3x_3$

We'd like to identify which, if any, of the predictor variables are useful in predicting Y

# Review: General Strategy for Regression Modeling




- What variables?
- How will I get the data?
- How much data do I need?

- What input variables have the biggest effect on the response?
- What are some candidate prediction models?

- Check model assumptions
- Outliers, Influential points?
- What is the best model?

- How well does the model work predicting new observations?
- How much variation is there in the predictions?

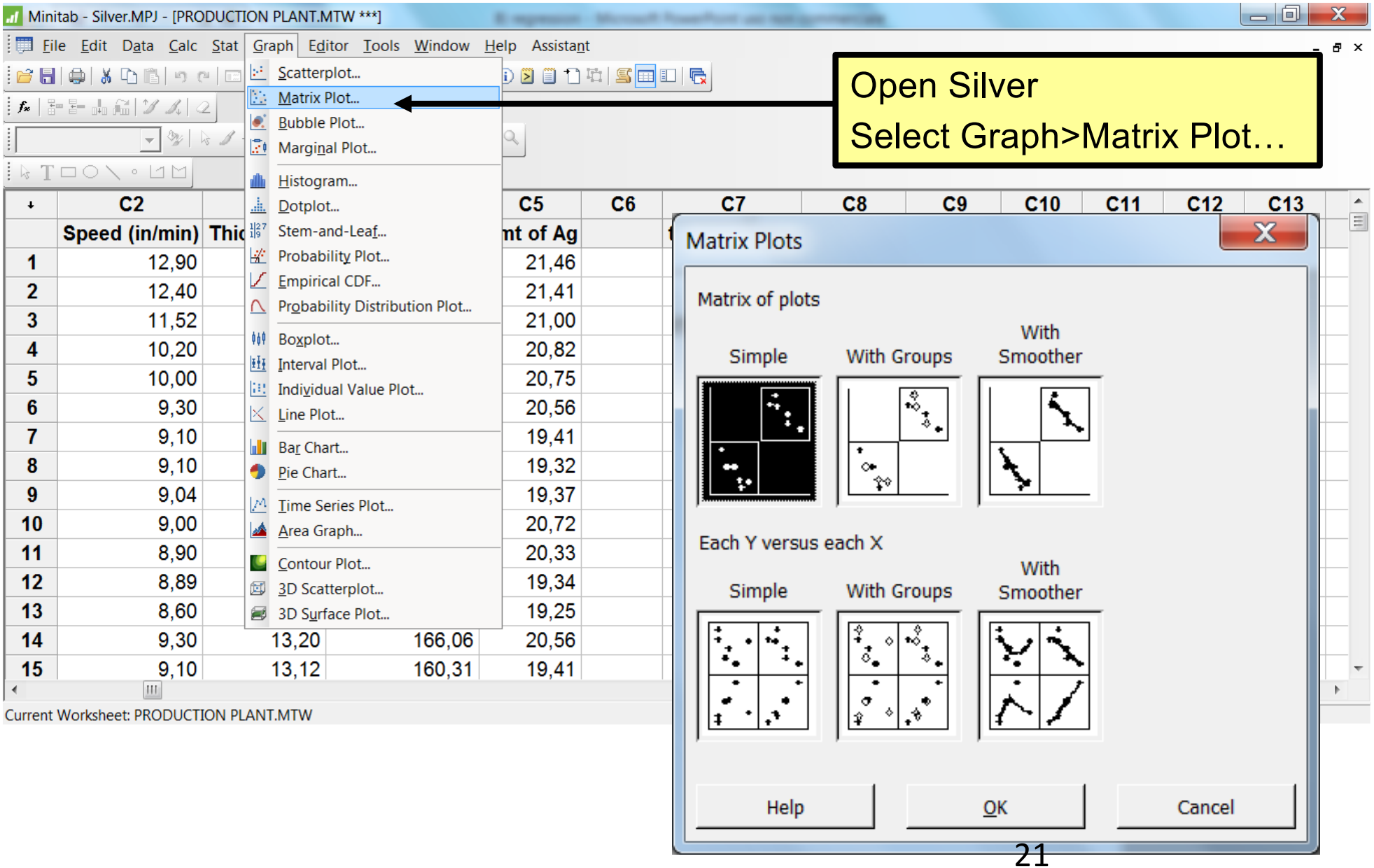
# Example: Production Plant

A chemical engineer is investigating the amount of silver required in the high volume production of contact switches. Although only a small amount of silver is deposited on the switches, a larger amount is wasted through a multiple step process. He has collected data and would like to develop a prediction model.
A Black Belt suggests that multiple regression might be useful.

The variables are given below

- $x_1$ = Average temperature of rinse bath (degrees C)
- $x_2$ = Speed of reel that feed the switches through the line (inches/min)
- $x_3$ = Thickness of silver deposit (angstroms)
- $x_4$ = Water consumed (gallons per day)
- Y = Amount of silver consumed (pounds/day)

Source: *Applied Regression Analysis*, Draper and Smith


Minitab - Silver.MPJ - [PRODUCTION PLANT.MTW ***]
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Scatterplot...
Matrix Plot...
Bubble Plot...
Marginal Plot...
Histogram...
Dotplot...
Stem-and-Leaf...
Probability Plot...
Empirical CDF...
Probability Distribution Plot...
Boxplot...
Interval Plot...
Individual Value Plot...
Line Plot...
Bar Chart...
Pie Chart...
Time Series Plot...
Area Graph...
Contour Plot...
3D Scatterplot...
3D Surface Plot...
Open Silver
Select Graph>Matrix Plot…
C2 C5 C6 C7 C8 C9 C10 C11 C12 C13
Speed (in/min) nt of Ag
1 12,90 21,46
2 12,40 21,41
3 11,52 21,00
4 10,20 20,82
5 10,00 20,75
6 9,30 20,56
7 9,10 19,41
8 9,10 19,32
9 9,04 19,37
10 9,00 20,72
11 8,90 20,33
12 8,89 19,34
13 8,60 19,25
14 9,30 13,20 166,06 20,56
15 9,10 13,12 160,31 19,41
Current Worksheet: PRODUCTION PLANT.MTW
Matrix Plots
Matrix of plots
Simple
With Groups
With Smoother
Each Y versus each X
Simple
With Groups
With Smoother
Help
OK
Cancel

# Potential Problems with Several Predictor Variables

Sometimes the x's are correlated (dependent). This condition is known as ***Multicollinearity***

**Multicollinearity can cause problems** (sometimes severe)

- Estimates of the coefficients are affected (unstable, inflated variances)
- Difficulty isolating the effects of each x
- Coefficients depend on which x's are included in the model

# Coefficient of Multiple Determination, R2

In linear regression, with one input, we used $R^2$ as a measure of what proportion of the variability in the output was explained by the input. In multiple linear regression, the measure is similar, but is known as the “coefficient of multiple determination”, $R^2$.

$R^2$ is a measure of the proportion of the variability in the output that is explained by all of the inputs taken together, not any one input individually.

# Some Cautions About the Coefficients

Relative importance of predictors cannot be determined from the size of their coefficients

Coefficients depend on the scale of the input variables and may not be directly comparable

- Suppose we measured production time in hours rather than days
- You may consider standardizing the data

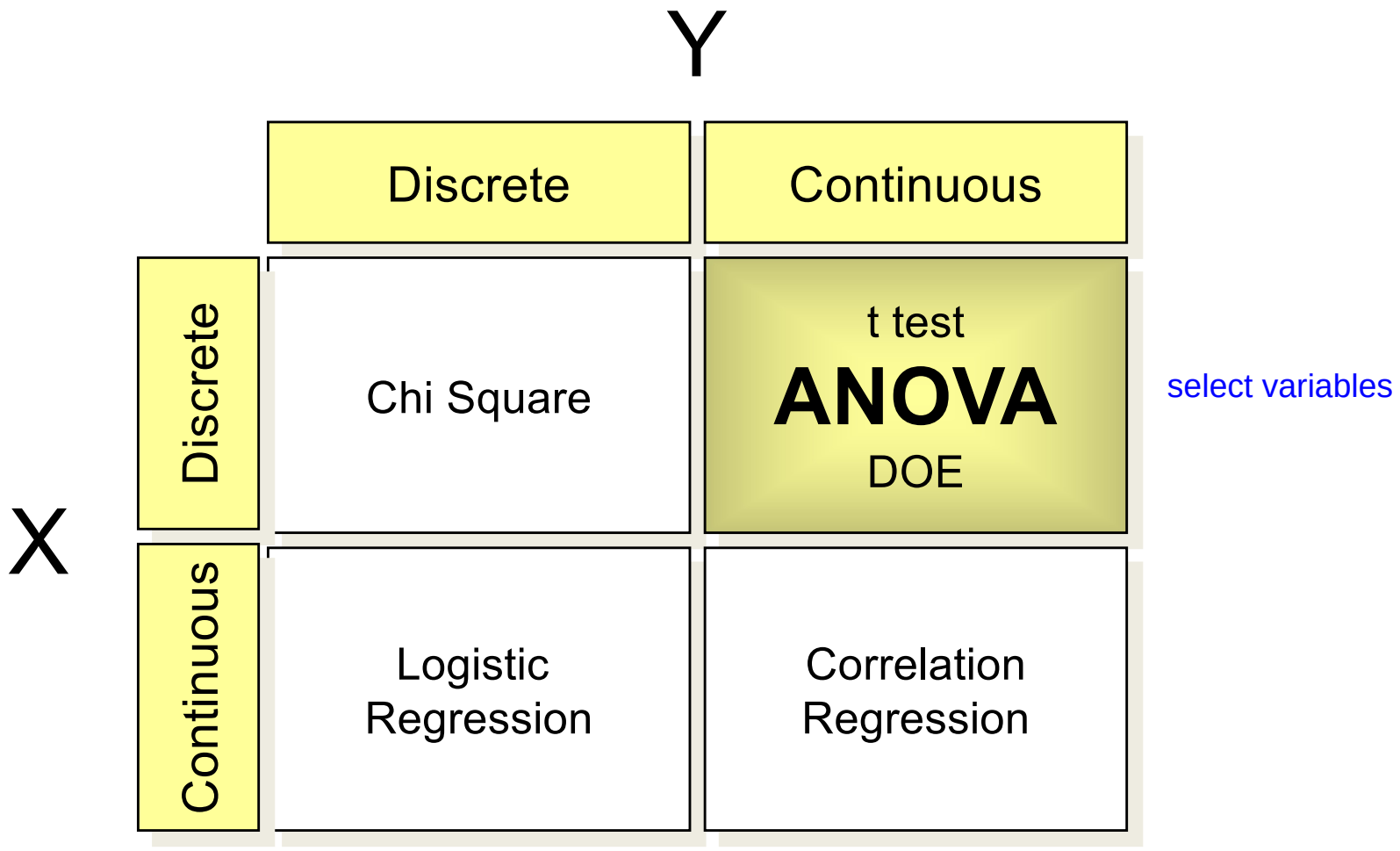
So, what do we do? How can we identify the important variables?

## Cautions on Using Regression Analysis

This has been a general introduction to analyzing potential root causes using linear regression. Before using a regression equation to predict the actual value of Y, considerable more analysis must be undertaken.

One general rule should always be followed when doing this kind of analysis.

- **"Never carry out regression analysis without first drawing the scatter diagram."**


Y
Discrete
Continuous
X
Discrete
Continuous
Chi Square
t test
ANOVA
DOE
Logistic Regression
Correlation Regression
select variables

# Examples

**Manufacturing** – An engineer is trying to determine if there is any relationship between the day of the week and the productivity level of the fish processing and packaging plant

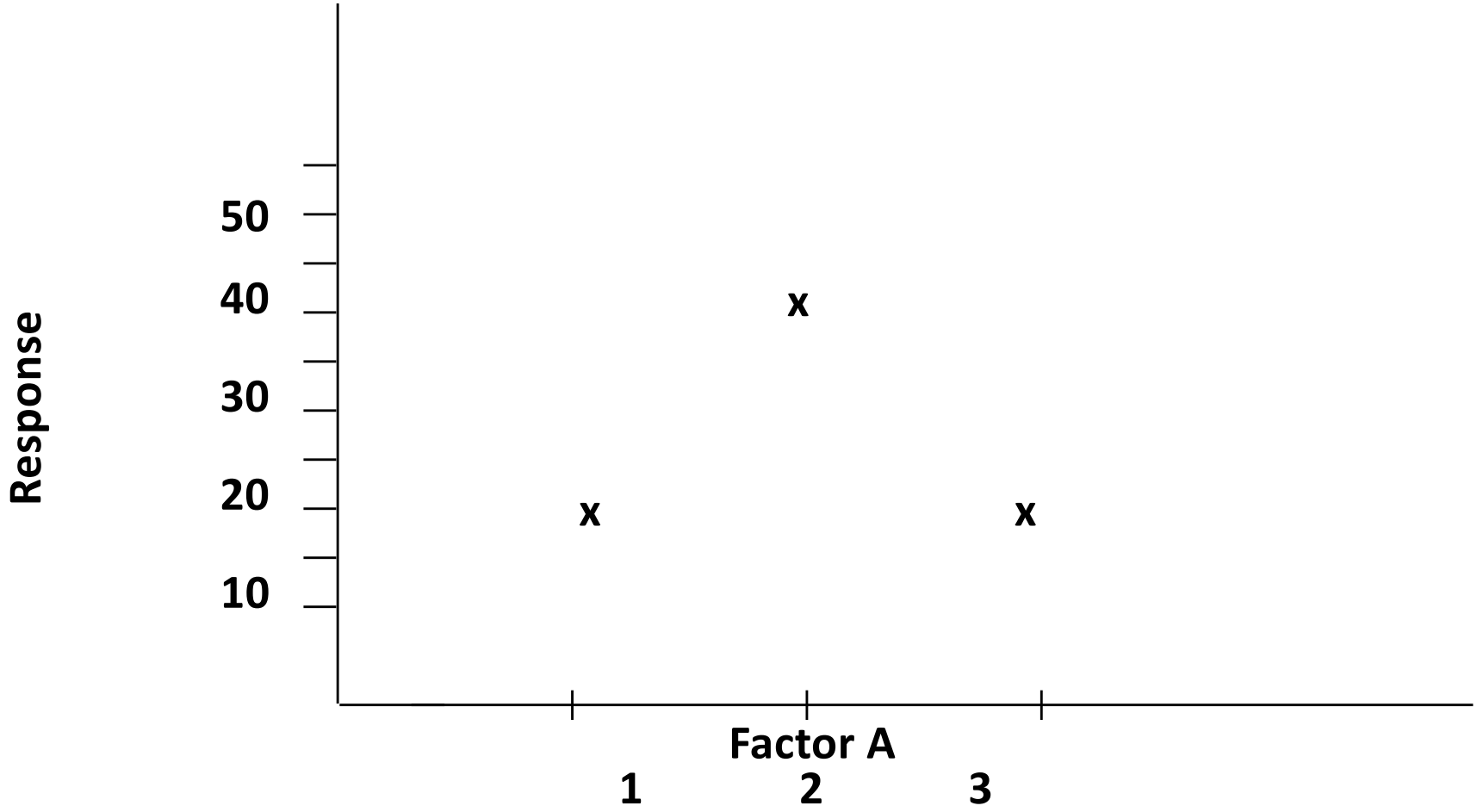
**Transactional** – A manager wants to understand how different attendance policies (in presence, on-line synchronous, on-line asynchronous) may affect productivity

**Design** – A design engineer needs to gather data on a proposed design change and its possible interactions with existing system parameters

**Marketing** – A marketing responsible wants to know if the type of packaging and shelf placement has an impact on sales.

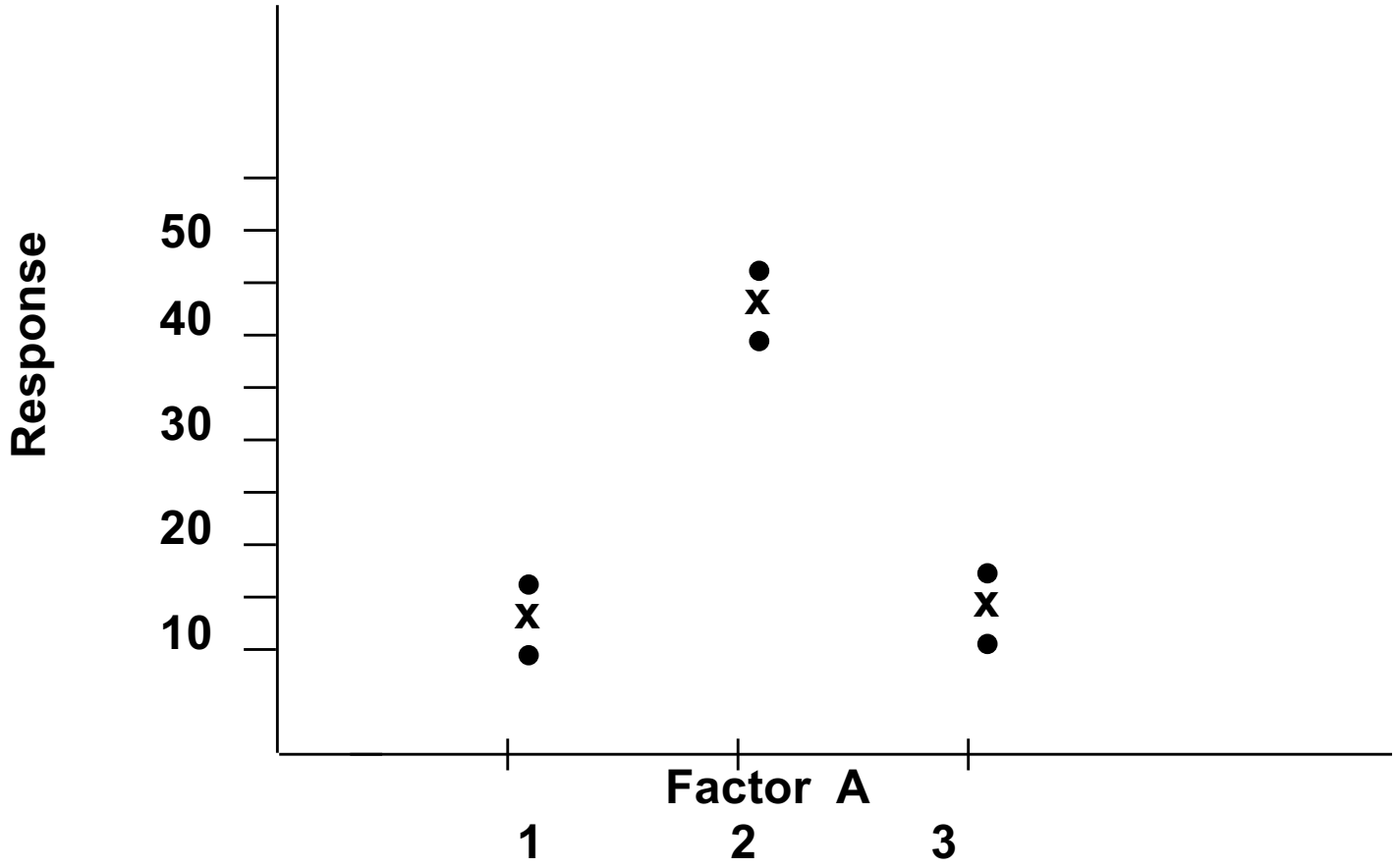
# Is there a difference?

Does factor A make a difference? Why or why not?



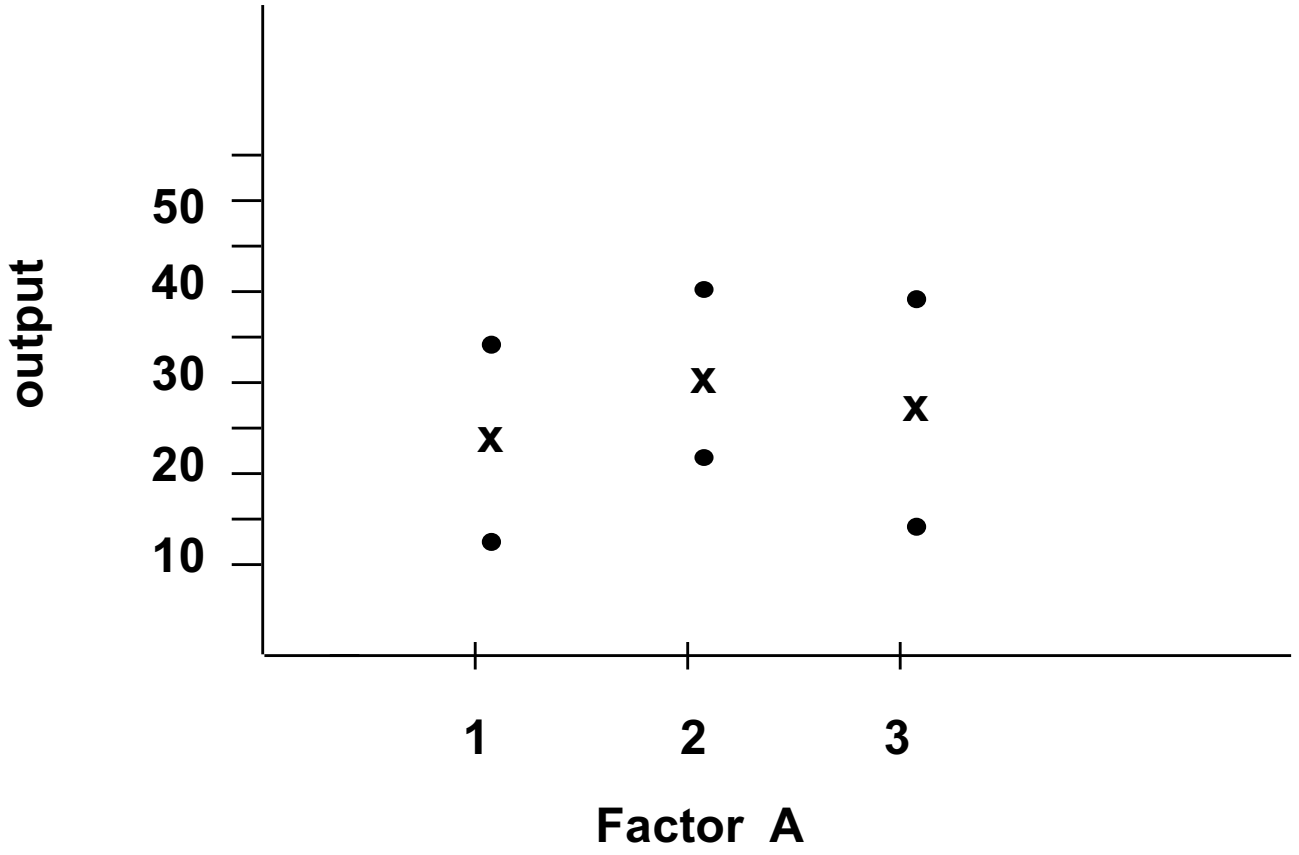
# Conceptual ANOVA

Does factor A make a difference? Why or why not?



# Conceptual ANOVA

Does factor A make a difference? Why or why not?



# One Factor Experiments

One-way analysis of variance (ANOVA) is a statistical method for comparing more than two sample means of the same factor.

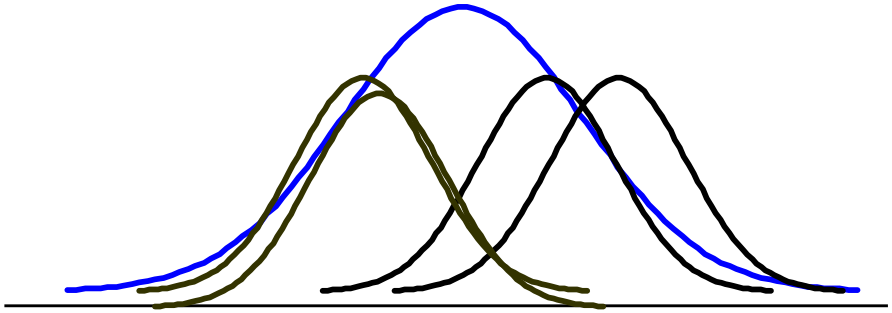
The hypothesis tested is:

$$\text{Ho}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$$
$$\text{Ha}: \text{ At least one } \mu_k \text{ is different}$$

Simply speaking, an ANOVA tests whether any of the population means are different. ANOVA does not tell us which ones are different. We'll supplement ANOVA with multiple comparison procedures for that.

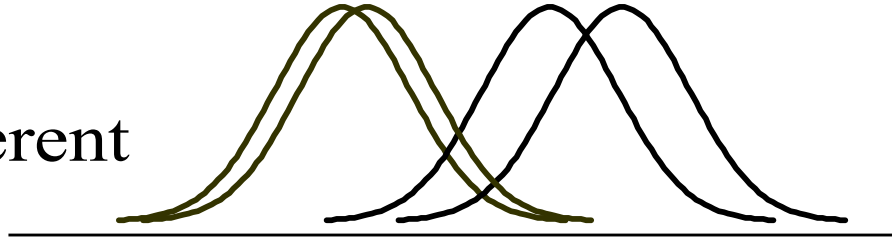
# Questions Asked by ANOVA

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$$



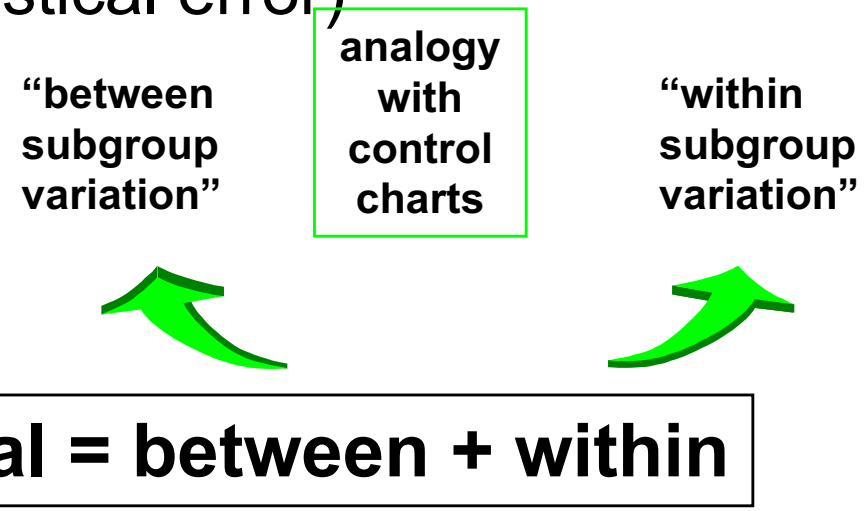
**Are some of the 4 population means different?**

$$H_a : \text{ At least one } \mu_k \text{ is different}$$

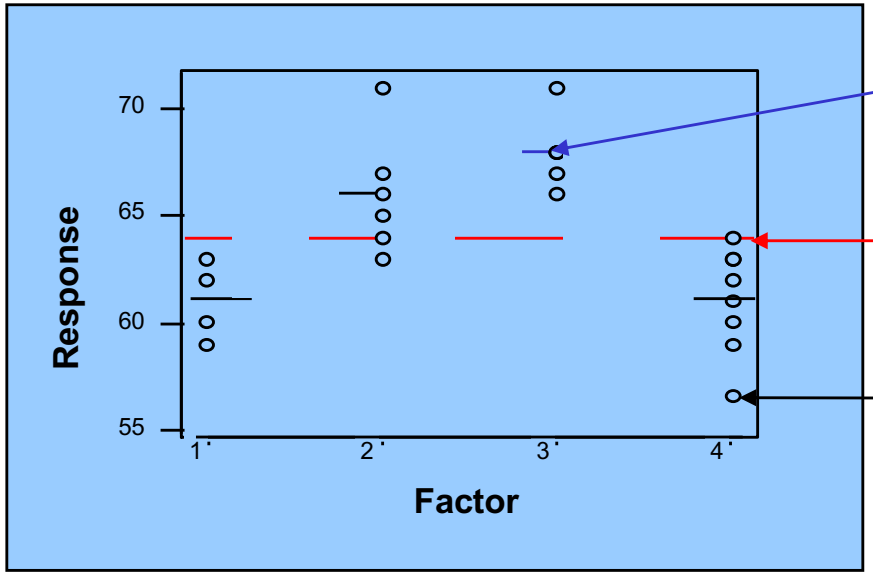


ANOVA looks at three sources of variability:

- Total = total variability among all observations
- Between = variation between group means (factor)
- Within = random (chance) variation within each group (noise, or statistical error)



# Understanding the Fundamentals – Sums of Squares




$\bar{y}_j$ - *Mean of Group*

$\bar{\bar{y}}$ - Grand Mean of the experiment

$y_{ij}$ - individual measurement

i = represents a data point within the $j^{th}$ group

j = represents the $j^{th}$ group

g = total # of groups

# Sums of Squares – Formula

$$\sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij}-\bar{\bar{y}})^2 = \sum_{j=1}^{g} n_j(\bar{y}_j-\bar{\bar{y}})^2 + \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$$

SS(Total) = SS(Factor) + SS(Error)

**SS(Tot) = Total Sum of Squares of the Experiment (individuals - Grand Mean)**

**SS(Factor) = Sum of Squares of the Factor (Group Mean - Grand Mean)**

**SS(Error) = Sum of Squares within the Group (individuals - Group Mean)**

By comparing sums of squared differences, we can tell if the observed difference is due to a true difference or random chance.

# ANOVA Sum of Squares

We can separate the total sum of squares into two components (within and between).

If the factor we are interested in has little or no effect on the average response, then these two estimates (within and between) should be fairly equal and we will conclude all subgroups could have come from one larger population.

As these two estimates (within and between) become significantly different, we will attribute this difference as originating from a difference in subgroup means.

# ANOVA – Null and Alternate Hypothesis

Ho : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha : At least one $\mu_k$ is different

To determine whether we can accept or reject the null hypothesis, we must calculate the Test Statistic (F-ratio) using the Analysis of Variance table as described on the following slide.

| SOURCE | SS | df | MS (=SS/df) | F {=MS(Factor)/MS(Error)} |
|---|---|---|---|---|
| **BETWEEN** | SS(Factor) | g - 1 | SS (factor)/df factor | MS(Factor) / MS(Error) |
| **WITHIN** | SS(Error) | $\sum_{j=1}^{g}\left(n_j - 1\right)$ | SS(Error) / df error | |
| **TOTAL** | SS(Total) | $\left(\sum_{j=1}^{g} n_j\right) - 1$ | | |

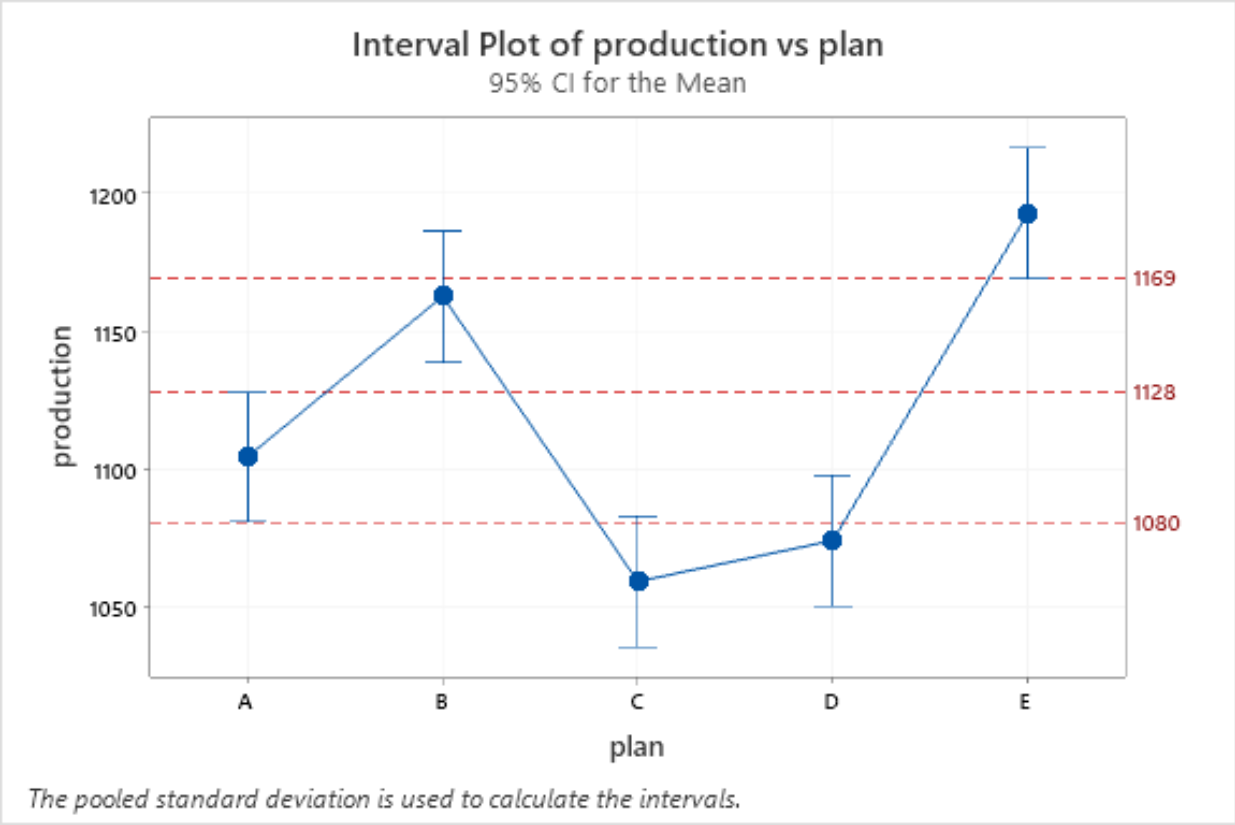
**Why is Source "Within" called the Error or Noise?**

**In practical terms what is the F-ratio telling us?**

**What do you think large F-ratios mean?**

# Tukey Analysis

The Tukey Analysis or pairwise comparisons, compare each group with all other, that is, the first is compared with the second, third, fourth and fifth; the second with the third, the fourth and the fifth (the comparison with the first had already been done, etc.)




This graph represents:

- Averages
- Confidence interval for each average

Overlapping between two confidence intervals means that the two situations are included into the same cluster

# Two-Ways ANOVA

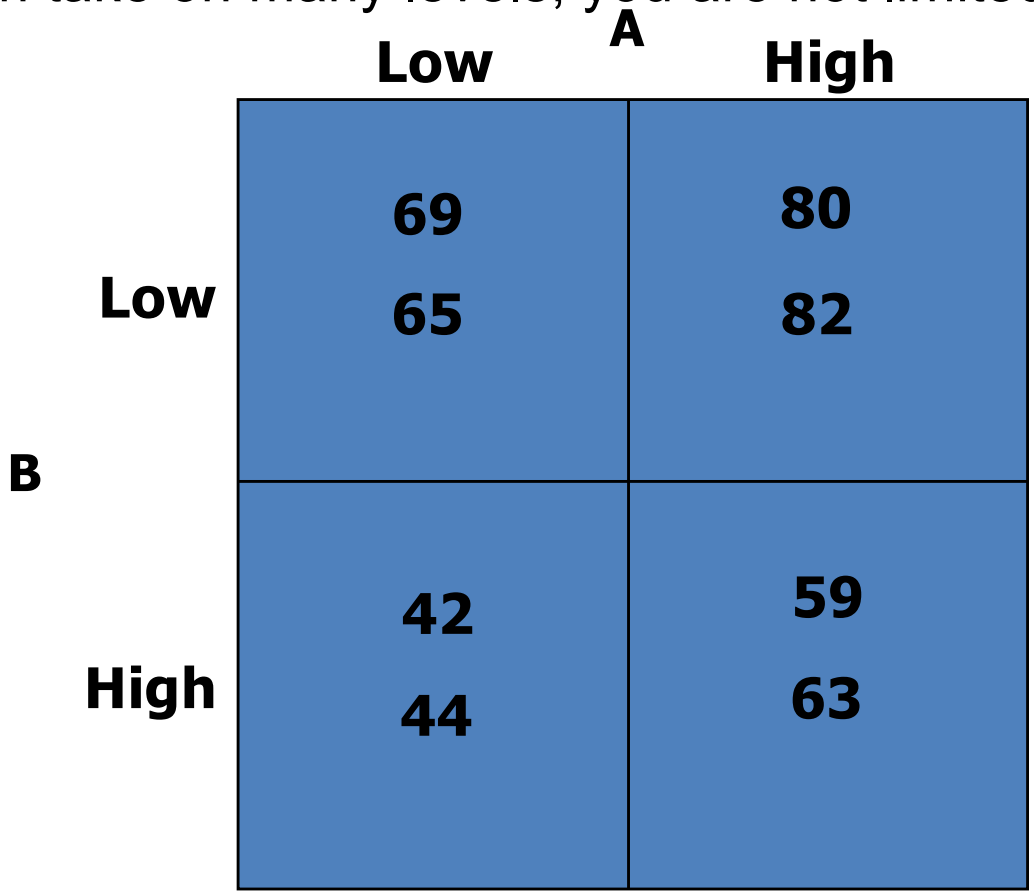
In One-Way ANOVA, we looked at how different levels of a single factor impacted a response variable.

In Two-Way ANOVA, we will examine how different levels of two factors and their interaction impact a response variable.

# Two-Ways ANOVA → Two Factors

At a high level, a two-ways ANOVA (two factor) can be viewed as a two-factor experiment.

The factors can take on many levels; you are not limited to two levels for each.



# Two-Ways ANOVA

Experiments often involve the study of more than one factor.

Factorial designs are very efficient methods where the combinations of the levels of factors are investigated.

These designs evaluate the effect on the response caused by different levels of factors and their interaction.

As in the case of One-Way ANOVA, we will be building a model and verifying some assumptions.

# Two Ways ANOVA

Just as in the one-factor analysis of variance, the total variability can be segmented into its component sum of squares:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_e$$

- Given:
- SST is the total sum of squares,
- SSA is the sum of squares from factor A,
- SSB is the sum of squares from factor B,
- SSAB is the sum of squares due to the interaction of A with B, and
- SSe is the sum of squares from error.

# Two Ways ANOVA: braking distance at 100 km/h

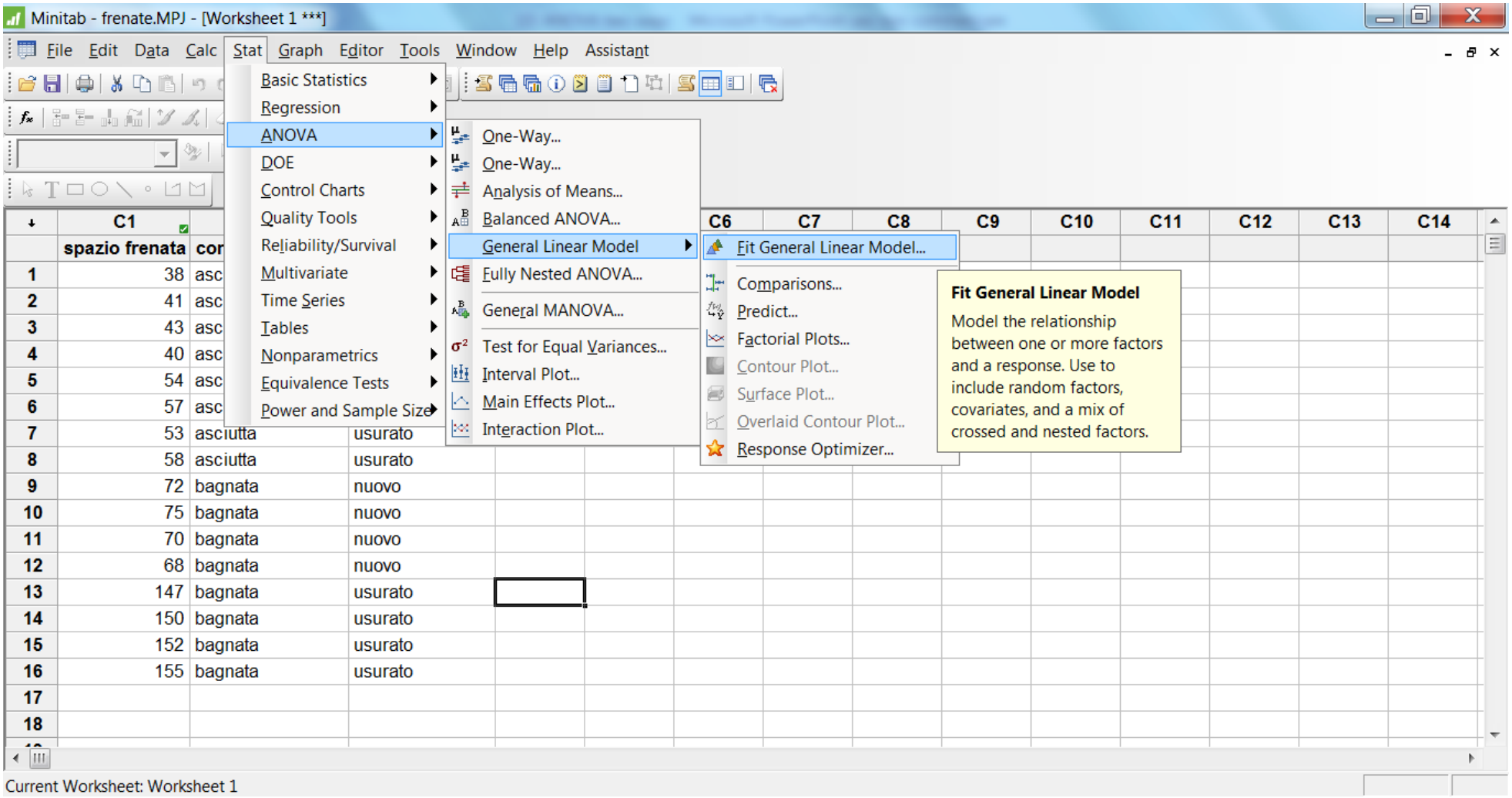
A classic test for assessing the quality of a tire is the distance to come to a halt from 100 km / hour.

In this example we have tested 4 conditions:

- New tire / dry
- Worn tire / dry
- New tire / wet
- Worn tire / wet

For each condition were made 4 tests

# Two Ways ANOVA: braking distance at 100 km/h



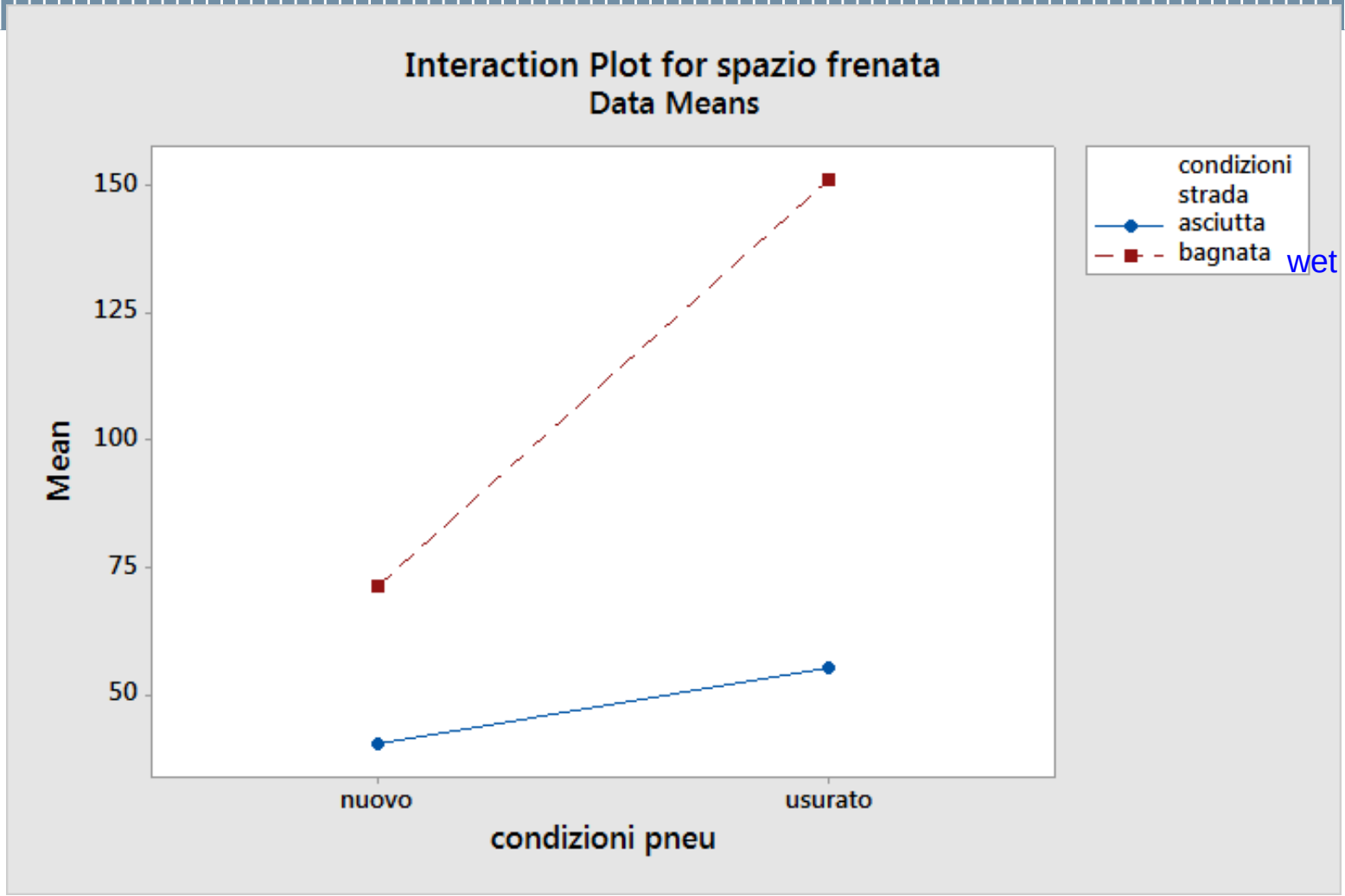

| | C1 | C2 | C3 |
|---|---|---|---|
| | spazio frenata | cor | |
| 1 | 38 | asc | |
| 2 | 41 | asc | |
| 3 | 43 | asc | |
| 4 | 40 | asc | |
| 5 | 54 | asc | |
| 6 | 57 | asc | |
| 7 | 53 | asciutta | usurato |
| 8 | 58 | asciutta | usurato |
| 9 | 72 | bagnata | nuovo |
| 10 | 75 | bagnata | nuovo |
| 11 | 70 | bagnata | nuovo |
| 12 | 68 | bagnata | nuovo |
| 13 | 147 | bagnata | usurato |
| 14 | 150 | bagnata | usurato |
| 15 | 152 | bagnata | usurato |
| 16 | 155 | bagnata | usurato |
| 17 | | | |
| 18 | | | |

# Two Ways ANOVA: braking distance at 100 km/h

The two parameters and interaction are significant in terms of influence on the braking distance.

The model explains more than 99% of the overall variability and the error is therefore negligible

# Two Ways ANOVA: interaction plot



POLITECNICO
MILANO 1863