



# LEAN SIGMA CORPORATION

Lean Six Sigma Black Belt Training  
Featuring Examples from Minitab 18



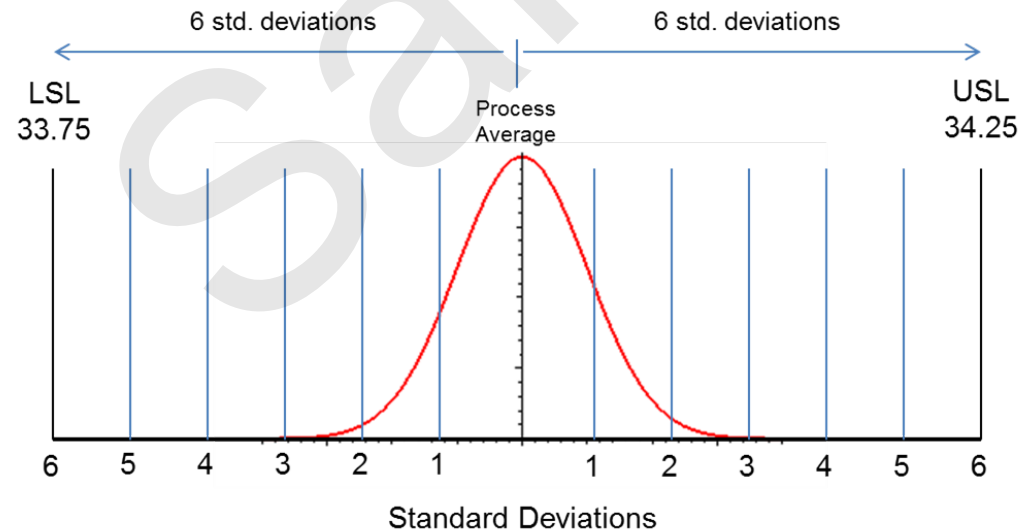
## 1.1.1 What is Six Sigma

Sample



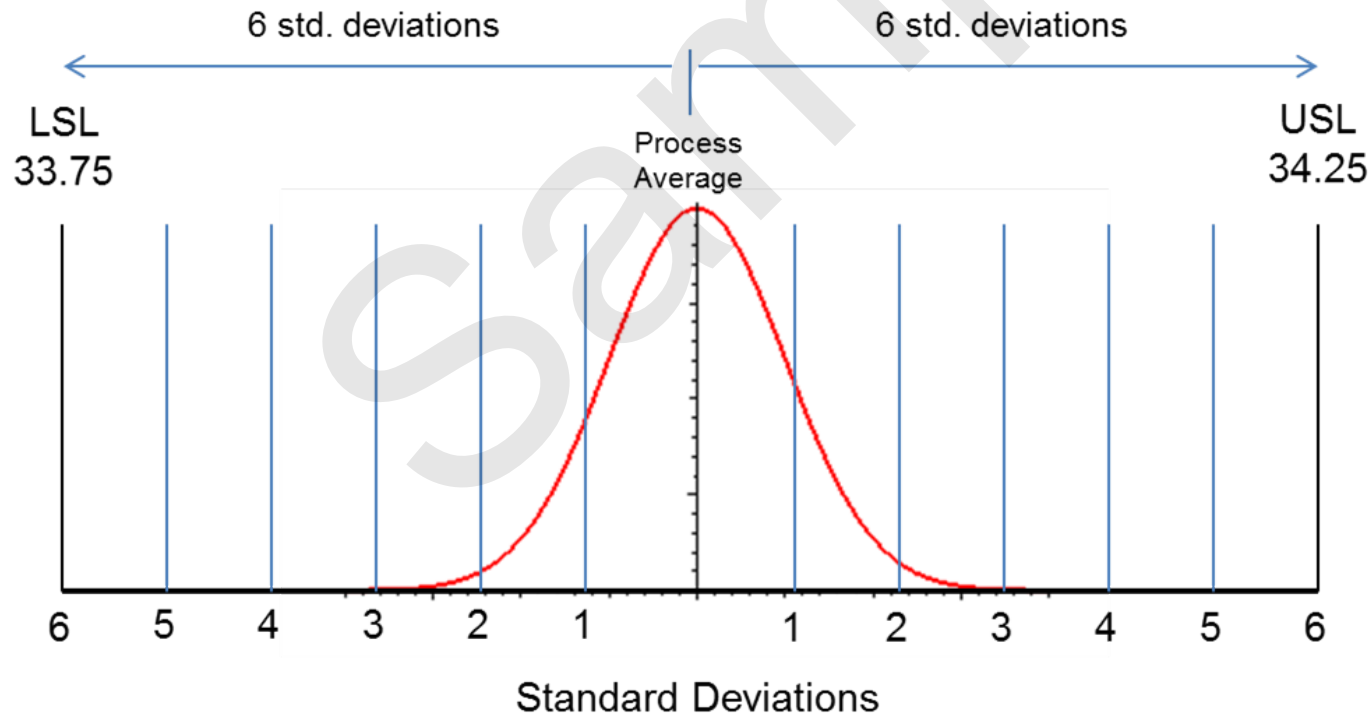
# What is Six Sigma?

- What is “sigma”?
  - In statistics, **sigma** ( $\sigma$ ) refers to “standard deviation,” which is a measure of variation.
  - You will come to learn that variation is the enemy of any quality process. We need to understand, manage, and minimize process variation.
- What is “Six Sigma”?
  - **Six Sigma** is an aspiration or goal of process performance.
  - A Six Sigma “goal” is for a process average to operate approximately  $6\sigma$  away from customer’s high and low specification limits.



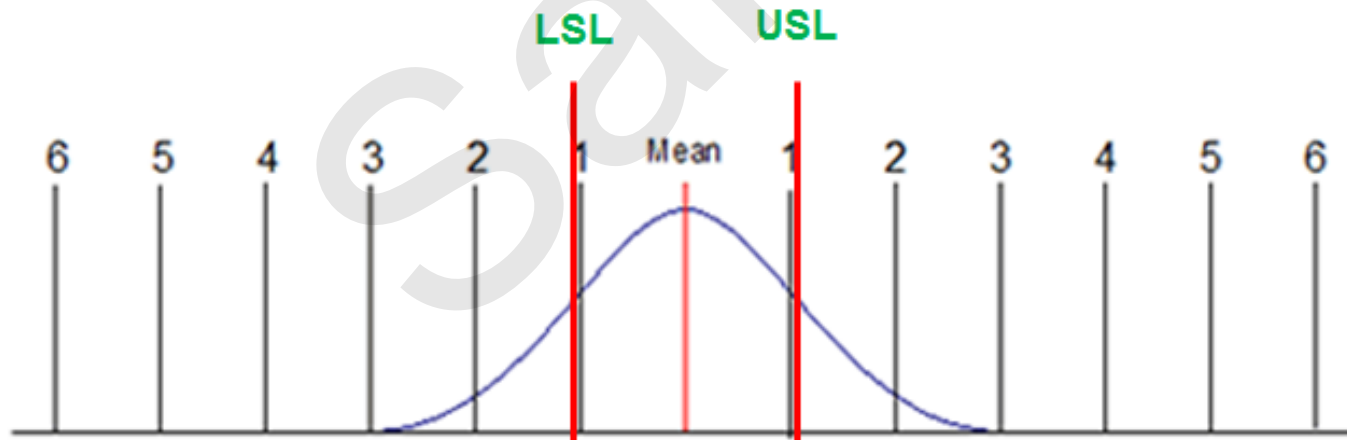
# What is Six Sigma?

- A process whose average is about  $6\sigma$  away from the customer's high and low specification limits has abundant room to "float" before approaching the customer's specification limits.
- A Six Sigma process only yields 3.4 defects for every million opportunities! In other words, 99.9997% of the products are defect-free!



# What is Six Sigma: Sigma Level

- **Sigma level** measures how many “sigma” there are between your process average and the nearest customer specification.
- Let us assume that your customers upper and lower specifications limits (USL & LSL) were narrower than the width of your process spread.
- The USL & LSL below stay about 1 standard deviation away from the process average. Therefore, this process operates at **1 sigma**.



# What is Six Sigma: Sigma Level

---

- A process operating at 1 sigma has a defect rate of approximately 70%.



- This means that the process will generate defect-free products only 30% of the time.
- What about processes with more than 1 sigma level?
- A higher sigma level means a lower defect rate.
- Let us take a look at the defect rates of processes at different sigma levels.



# What is Six Sigma: Sigma Level

- This table shows each sigma level's corresponding defect rate and DPMO (defects per million opportunities).
- The higher the sigma level, the lower the defective rate and DPMO.

Sigma Level	Defect Rate	DPMO
1	69.76%	697612
2	30.87%	308770
3	6.68%	66810
4	0.62%	6209
5	0.023%	232
6	0.00034%	3.4

These Defect Rates Assume a 1.5 sigma shift

- How does this translate into things you might easily relate to?



# What is Six Sigma: Sigma Level

---

- Let us take a look at processes operating at 3 sigma.
- 3 sigma processes have a defect rate of approximately 7%. What would happen if processes operated at 3 sigma?
  - Virtually no modern computer would function\*.
  - 10,800,000 health care claims would be mishandled each year.
  - 18,900 US savings bonds would be lost every month.
  - 54,000 checks would be lost each night by a single large bank.
  - 4,050 invoices would be sent out incorrectly each month by a modest-sized telecommunications company.
  - 540,000 erroneous call details would be recorded each day from a regional telecommunications company.
  - 270 million erroneous credit card transactions would be recorded each year in the United States.

(\*<http://www.qualityamerica.com>)





# What is Six Sigma: Sigma Level

---

- What if processes operated with 1% defect rate?
  - 20,000 lost articles of mail per hour\*.
  - Unsafe drinking water almost 15 minutes per day.
  - 5,000 incorrect surgical operations per week.
  - Short or long landings at most major airports each day.
  - 200,000 wrong drug prescriptions each year.
  - No electricity for almost 7 hours per month.
- Even at 1% defect rate, some processes would be unacceptable to you and many others.
- **So what is Six Sigma?**
  - Sigma level is the measure!
  - Six is the goal!

(\* Implementing Six Sigma – Forest W. Breyfogle III)



# What is Six Sigma: The Methodology

---

- Six Sigma itself is the **goal**, not the method.
- In order to achieve Six Sigma, you need to improve your process performance by:
  - Minimizing the process variation so that your process has enough room to fluctuate within customer's spec limits
  - Shifting your process average so that it is centered between your customer's spec limits.
- Accomplishing these two process improvements (*along with stabilization and control*), you can achieve Six Sigma.
- DMAIC is the systematic methodology prescribed to achieve Six Sigma.



# What is Six Sigma: The Methodology

---

- DMAIC is a systematic and rigorous methodology that can be applied to any process in order to achieve Six Sigma.
- It consists of 5 phases of a project:
  - **D**efine
  - **M**easure
  - **A**nalyze
  - **I**mprove
  - **C**ontrol.
- You will be heavily exposed to many concepts, tools, and examples of the DMAIC methodology through this training.
- You will be capable of applying the DMAIC methodology to improve the performance of any process at the completion of the curriculum.



## 1.1.3 Six Sigma Approach

Sample



# Six Sigma Approach: $Y = f(x)$

---

- The Six Sigma approach to problem solving uses a transfer function.
- A **transfer function** is a mathematical expression of the relationship between the inputs and outputs of a system.
- **$Y = f(x)$**  is the relational transfer function that is used by all Six Sigma practitioners.
- It is absolutely critical that you understand and embrace this concept.



# Six Sigma Approach: $Y = f(x)$

---

- “Y” refers to the measure or output of a process.
  - Y is usually your primary metric
  - Y is the measure of process performance that you are trying to improve.
- $f(x)$  means “function of x.”
  - x’s are factors or inputs that affect the Y
- Combined, the  $Y = f(x)$  statement reads “Y is a function of x.”
- In simple terms: “My process performance is dependent on certain x’s.”
- The objective in a Six Sigma project is to identify the critical x’s that have the most influence on the output (Y) and adjust them so that the Y improves.



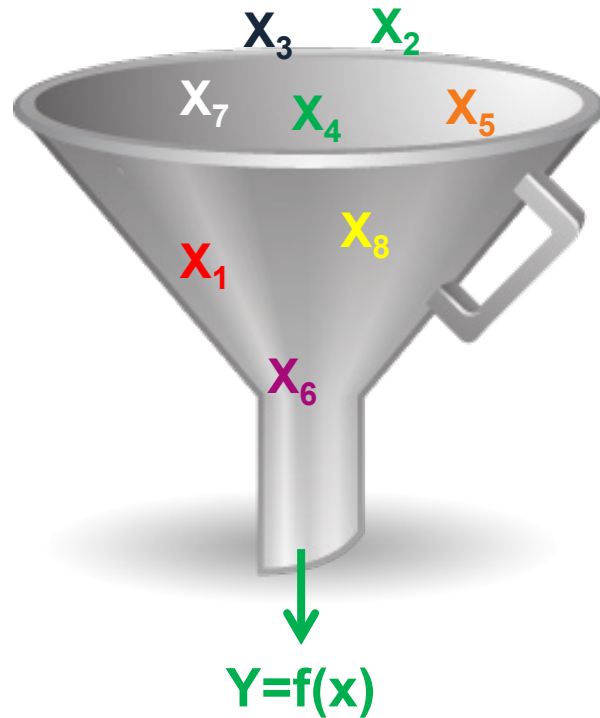
# Six Sigma Approach: $Y = f(x)$

---

- Let us look at a simple example of a pizza delivery company that desires to meet customer expectations of on-time delivery.
  - Measure = on-time pizza deliveries
    - $Y$  = percent of on-time deliveries
  - $f(x)$  would be the  $x$ 's or factors that heavily influence timely deliveries
    - $x_1$ : might be traffic
    - $x_2$ : might be the number of deliveries per driver dispatch
    - $x_3$ : might be the accuracy of directions provided to the driver
    - $x_4$ : might be the reliability of the delivery vehicle
    - etc.
- The statement  $Y = f(x)$  in this example will refer to the proven  $x$ 's determined through the steps of a Six Sigma project.



# Six Sigma Approach: $Y = f(x)$



- With this approach, all potential x's are evaluated throughout the DMAIC methodology.
- The x's should be narrowed down until the vital few x's that significantly influence “on-time pizza deliveries” are identified!





# Six Sigma Approach: $Y = f(x)$

---

- This approach to problem solving will take you through the process of determining all potential  $x$ 's that **might** influence on-time deliveries and then determining through measurements and analysis which  $x$ 's **do** influence on-time deliveries.
- Those significant  $x$ 's become the ones used in the  $Y = f(x)$  equation.
- The  $Y = f(x)$  equation is a very powerful concept and requires the ability to measure your output and quantify your inputs.
- Measuring process inputs and outputs is crucial to effectively determining the significant influences to any process.



## 1.2.5 Pareto Charts and Analysis

Sample



# Pareto Principle

---

- The **Pareto principle** is commonly known as the “law of the vital few” or “80:20 rule.”
- It means that the majority (approximately 80%) of effects come from a few (approximately 20%) of the causes.
- This principle was first introduced in early 1900s and has been applied as a rule of thumb in various areas.
- Example of applying the Pareto principle:
  - 80% of the defects of a process come from 20% of the causes.
  - 80% of sales come from 20% of customers.



# Pareto Principle

---

- The Pareto principle helps us to focus on the vital few items that have the most significant impact.
- In concept, it also helps us to prioritize potential improvement efforts.
- Since this 80:20 rule was originally based upon the works of Wilfried Fritz Pareto (or Vilfredo Pareto), the Pareto principle and references to it should be capitalized because Pareto refers to a person (proper noun).
  - Mr. Pareto is also credited for many works associated with the 80:20, some more loosely than others:
    - Pareto's Law
    - Pareto efficiency
    - Pareto distribution etc.



# Pareto Charts

---

- A **Pareto chart** is a chart of descending bars with an ascending cumulative line on the top.

- **Sum or Count:**

The descending bars on a Pareto chart may be set on a scale that represents the total of all bars or relative to the biggest bucket, depending on the software you are using.

- **Percent to Total:** A Pareto chart shows the percentage to the total for individual bars.
- **Cumulative Percentage:** A Pareto chart also shows the cumulative percentage of each additional bar. The data points of all cumulative percentages are connected into an ascending line on the top of all bars.



# Pareto Charts

---

- Case study time!
  - Next we will use Minitab to run Pareto charts on exactly the same data set.
  - Open the Minitab data file labeled Pareto and follow the instructions over the next few pages to run Pareto charts in Minitab.

Sample



# Create a Pareto Chart in Minitab

---

- Steps to generate a Pareto chart using Minitab:
  1. Open the spreadsheet with the count data for individual categories.
  2. Click on Stat → Quality Tools → Pareto Chart.
  3. A new window with the title “Pareto Chart” pops up.
  4. Select “Category” into the box “Defects or attribute data in” and “Count” into the box “Frequency in.”
  5. Click “OK.”
  6. The Pareto chart is created in a newly-generated window.



# Create a Pareto Chart in Minitab

Pareto Chart

C1	Count
C2	Category

Defects or attribute data in:

Frequencies in:  (optional)

BY variable in:

Default (all on one graph, same ordering of bars)

One group per graph, same ordering of bars

One group per graph, independent ordering of bars

Combine remaining defects into one category after this percent:

Do not combine

Select

Help

Options...

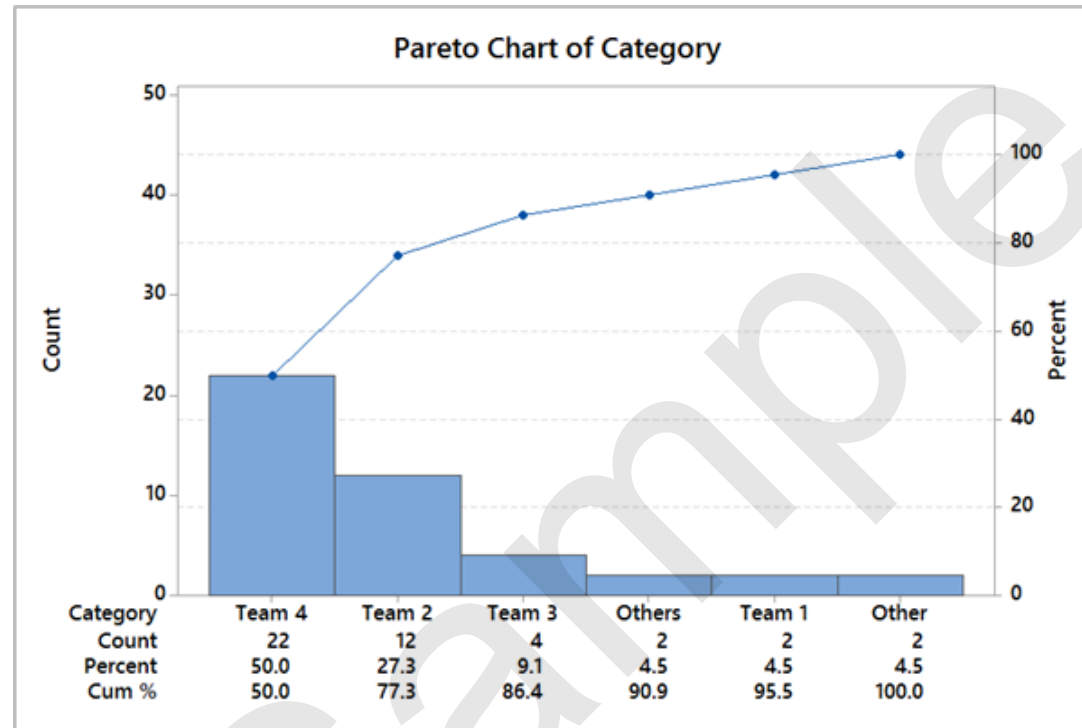
OK

Cancel





# Create a Pareto Chart in Minitab



- The above Pareto chart generated in Minitab presents the count of defective products by team.
- The bars are descending on a scale with the peak at 50, which is approximately the total count of all defective products for all teams.
- The table below the chart contains counts, individual percentages, and cumulative percentages.
- The cumulative percentages are the red data points driving the red line that spans across the graphic.



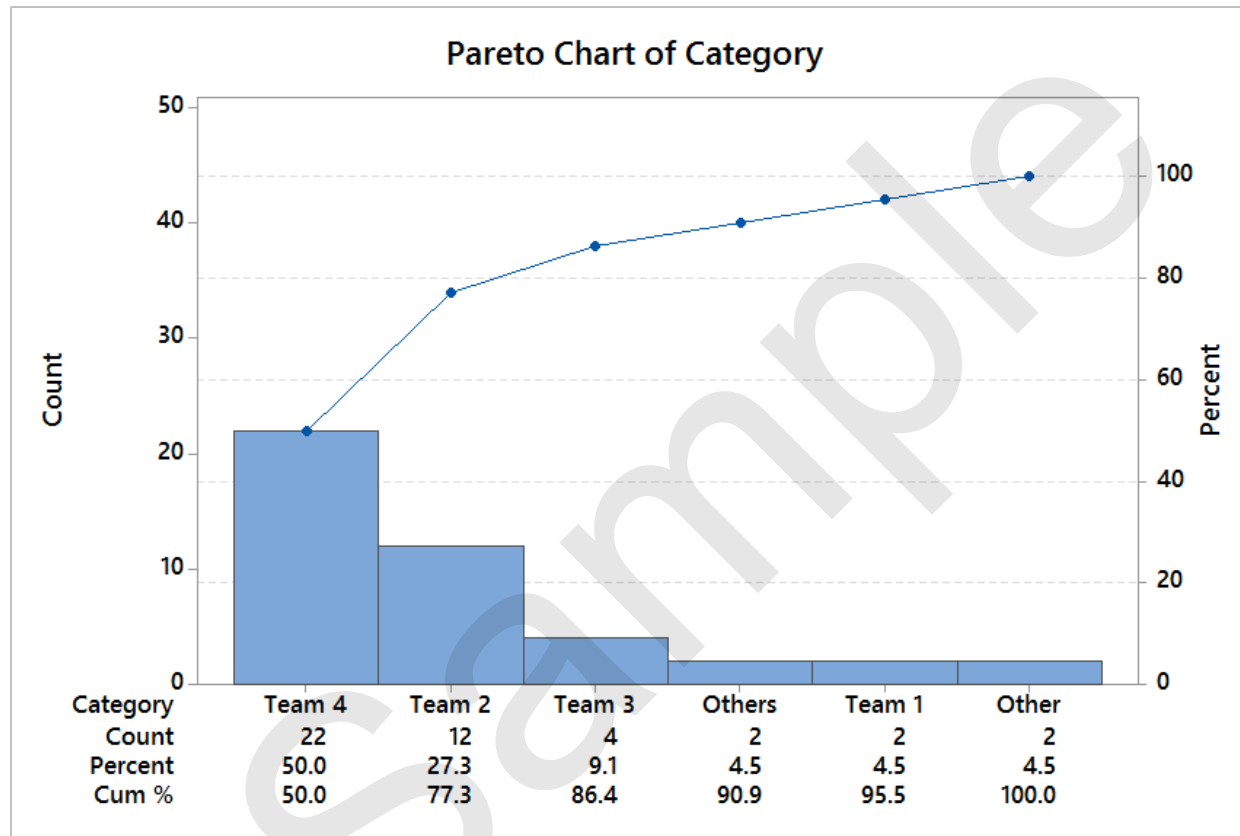
# Pareto Analysis

---

- The **Pareto analysis** is used to identify the root causes by using multiple Pareto charts.
- In Pareto analysis, we drill down into the bigger buckets of defects and identify the root causes of defects that contribute heavily to total defects.
- This "drill down" approach effectively solves a significant portion of the problem.
- Next you will see an example of three-level Pareto analysis.
  - The second-level Pareto is a Pareto chart that is a subset of the tallest bar on the first Pareto.
  - The third-level Pareto is a subset of the tallest bar of the second-level Pareto.



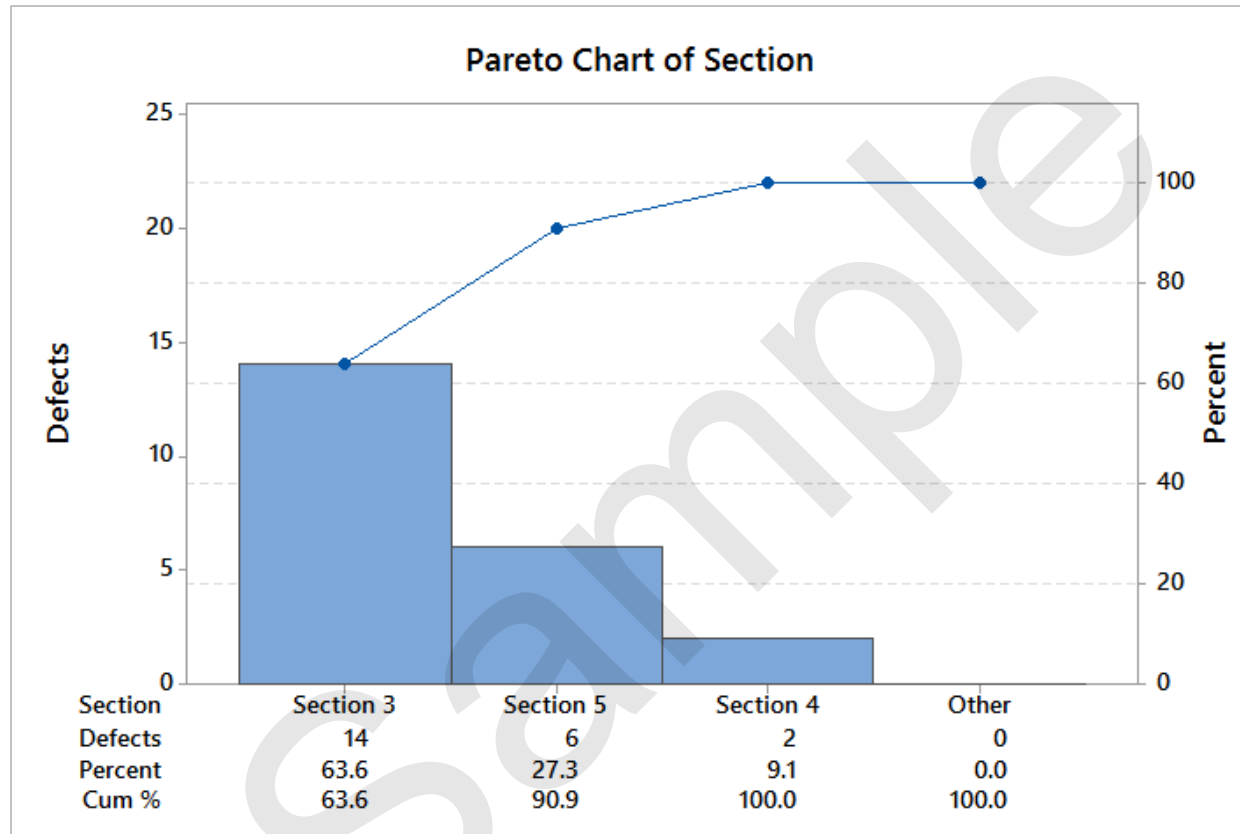
# Pareto Analysis: First Level



- First-level Pareto
- Shows the count of defective items by team
- Next level will only show the defective items of team 4



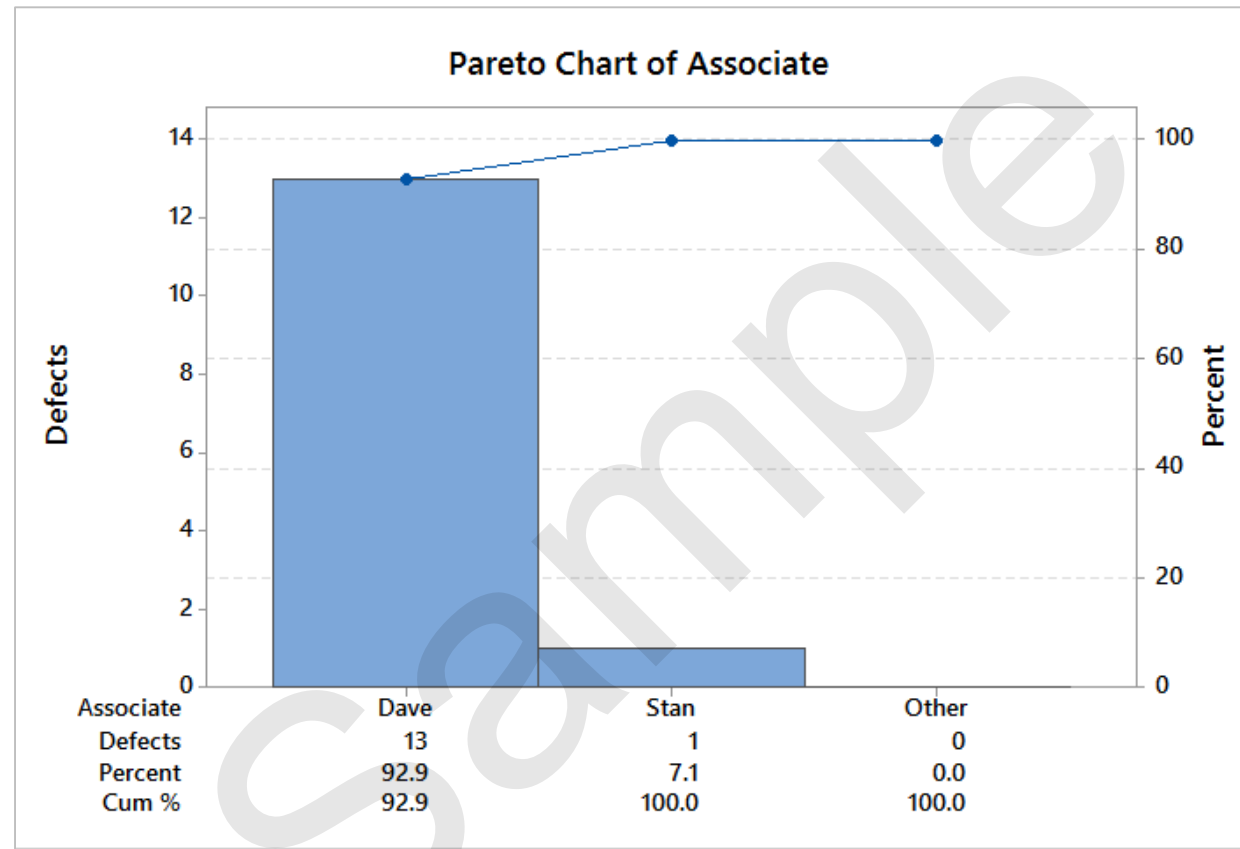
# Pareto Analysis: Second Level



- Second-level Pareto
- Shows the count of the defective items by section for only team 4
- Next level will only show the defective items of section 3



# Pareto Analysis: Third Level

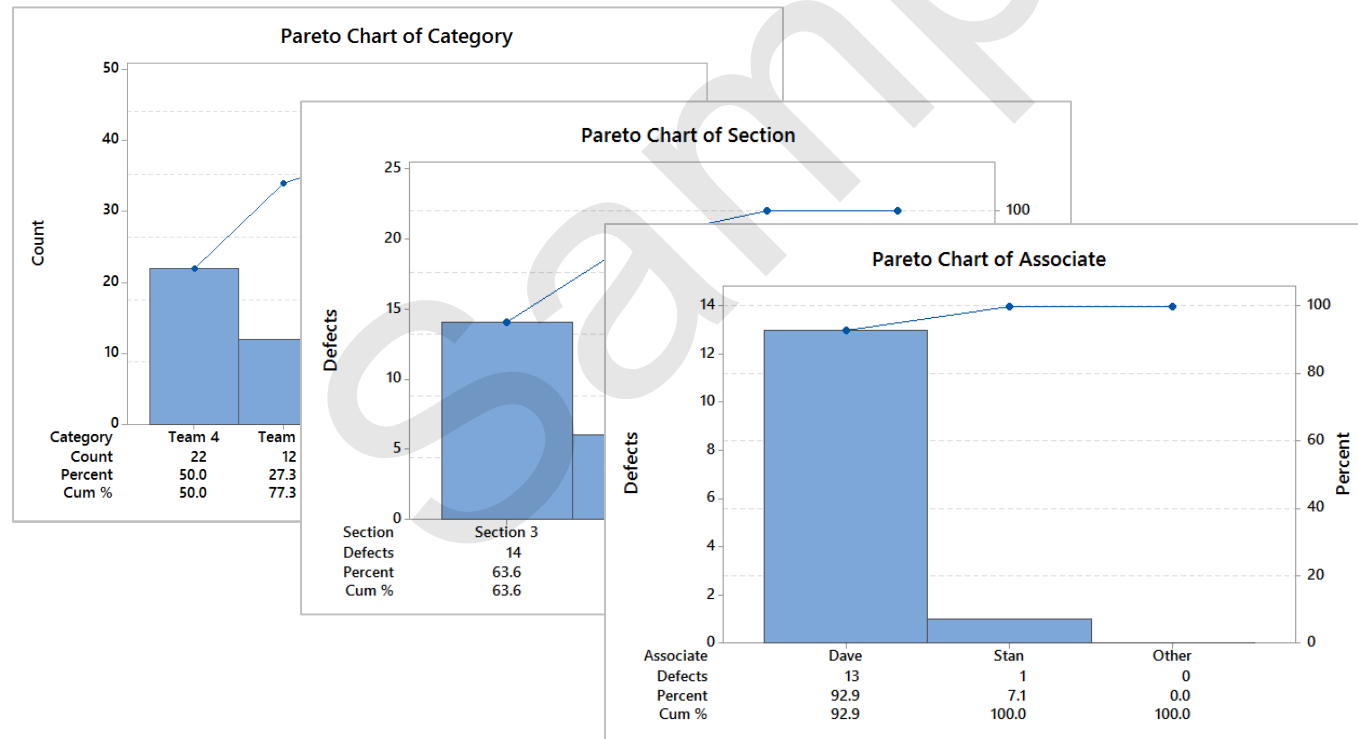


- Third-level Pareto
- Shows the count of defective items by associate for only section 3 of team 4
- Next level will only show the defective items of Dave



# Pareto Analysis: Conclusion

- After drilling down three levels we find that most of the defective products are from Dave who is in Section 3 of Team 4.
- Determining what Dave might be doing differently and solving that problem can potentially fix about 30% of the entire defective products (13/44).



## 3.4.1 One & Two Sample T-Tests



# What is a T-Test?

---

- In statistics, a **t-test** is a hypothesis test in which the test statistic follows a *Student t* distribution if the null hypothesis is true.
- We apply a t-test when the population variance ( $\sigma$ ) is unknown and we use the sample standard deviation ( $s$ ) instead.

Sample





# What is One Sample T-Test?

---

- **One sample t-test** is a hypothesis test to study whether there is a statistically significant difference between a population mean and a specified value.
  - Null Hypothesis ( $H_0$ ):  $\mu = \mu_0$
  - Alternative Hypothesis ( $H_a$ ):  $\mu \neq \mu_0$

where  $\mu$  is the mean of a population of our interest and  $\mu_0$  is the specific value we want to compare against.



# Assumptions of One Sample T-Test

---

- The sample data drawn from the population of interest are unbiased and representative.
- The data of the population are continuous.
- The data of the population are normally distributed.
- The variance of the population of our interest is unknown.
- One sample t-test is more robust than the z-test when the sample size is small ( $< 30$ ).



# Normality Test

---

- To check whether the population of our interest is normally distributed, we need to run normality test.
  - Null Hypothesis ( $H_0$ ): The data are normally distributed.
  - Alternative Hypothesis ( $H_a$ ): The data are not normally distributed.
- There are a lot of normality tests available:
  - Anderson-Darling
  - Sharpiro-Wilk
  - Jarque-Bera etc.



# Test Statistic and Critical Value of One Sample T-Test

- Test Statistic

$$t_{calc} = \frac{\bar{Y}}{s / \sqrt{n}}, \text{ where}$$

$\bar{Y}$  is the sample mean,  $n$  is the sample size, and  $s$  is the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

- Critical Value

- $t_{crit}$  is the t-value in a Student t distribution with the predetermined significance level  $\alpha$  and degrees of freedom  $(n - 1)$ .
- $t_{crit}$  values for a two-sided and a one-sided hypothesis test with the same significance level  $\alpha$  and degrees of freedom  $(n - 1)$  are different.



# Decision Rules of One Sample T-Test

---

- Based on the sample data, we calculated the test statistic  $t_{\text{calc}}$ , which is compared against  $t_{\text{crit}}$  to make a decision of whether to reject the null.
  - Null Hypothesis ( $H_0$ ):  $\mu = \mu_0$
  - Alternative Hypothesis ( $H_a$ ):  $\mu \neq \mu_0$
- If  $|t_{\text{calc}}| > t_{\text{crit}}$ , we reject the null and claim there is a statistically significant difference between the population mean  $\mu$  and the specified value  $\mu_0$ .
- If  $|t_{\text{calc}}| < t_{\text{crit}}$ , we fail to reject the null and claim there is not any statistically significant difference between the population mean  $\mu$  and the specified value  $\mu_0$ .



# Use Minitab to Run a One-Sample T-Test

---

- *Case study:* We want to compare the average height of basketball players against 7 feet.
  - Data File: “One Sample T-Test” tab in “Sample Data.xlsx”
  - Null Hypothesis ( $H_0$ ):  $\mu = 7$
  - Alternative Hypothesis ( $H_a$ ):  $\mu \neq 7$



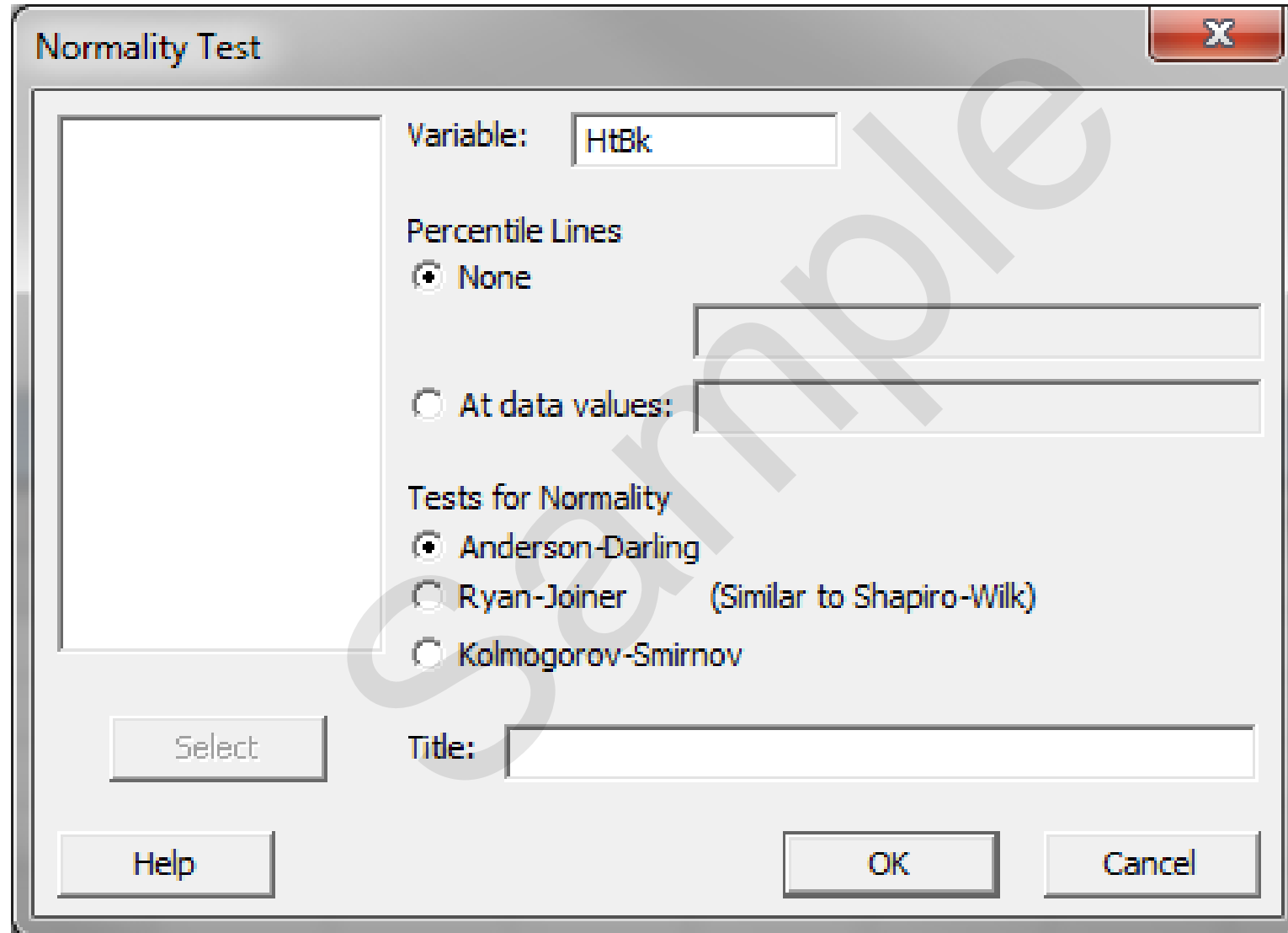
# Use Minitab to Run a One-Sample T-Test

---

- Step 1: Test whether the data are normally distributed
  - 1) Click Stat → Basic Statistics → Normality Test.
  - 2) A new window named “Normality Test” pops up.
  - 3) Select “HtBk” as the variable.
  - 4) Click “OK.”
  - 5) A new window named “Probability Plot of HtBk” appears, which covers the results of the normality test.



# Use Minitab to Run a One-Sample T-Test



The image shows a screenshot of the 'Normality Test' dialog box in Minitab. The dialog box has a title bar with the text 'Normality Test' and a close button (X) in the top right corner. The main area contains the following options:

- Variable:** A text box containing 'HtBk'.
- Percentile Lines:** Three radio button options:
  - None
  - At data values:
- Tests for Normality:** Three radio button options:
  - Anderson-Darling
  - Ryan-Joiner (Similar to Shapiro-Wilk)
  - Kolmogorov-Smirnov

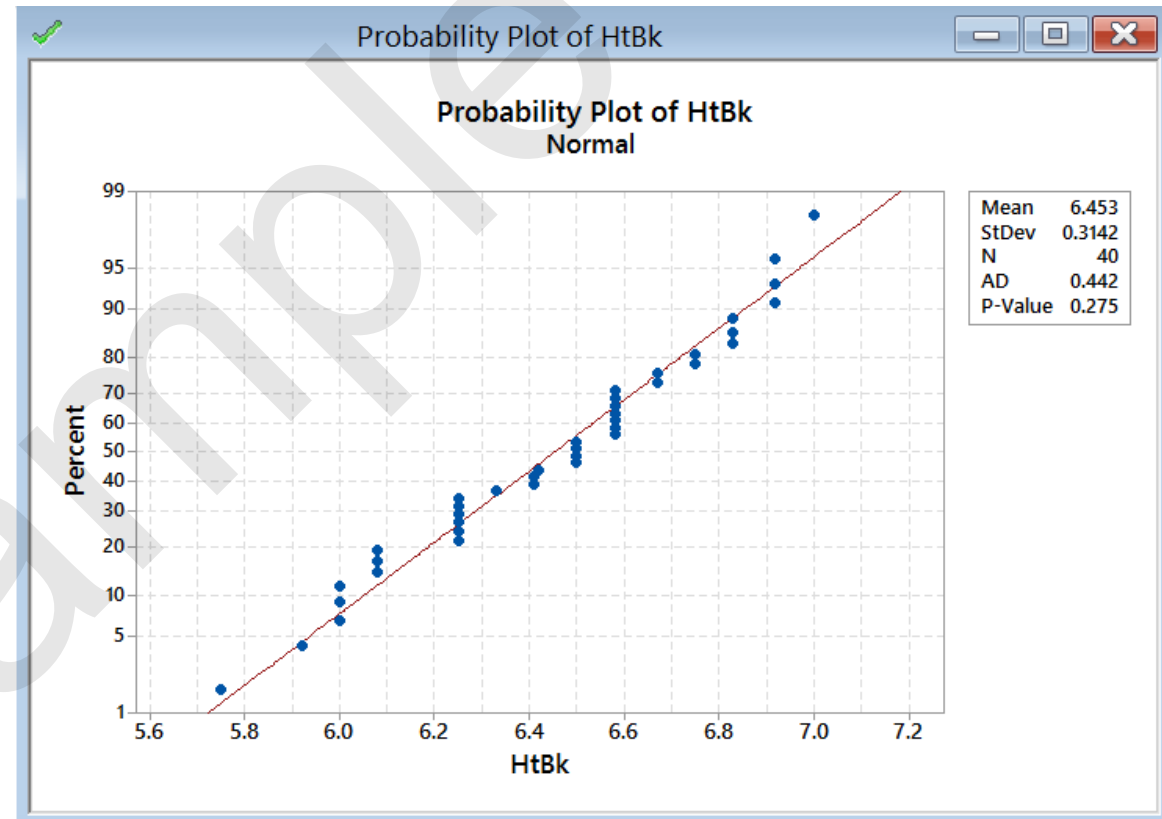
At the bottom of the dialog box, there are four buttons: 'Select', 'Help', 'OK', and 'Cancel'. The 'Select' button is positioned to the left of the 'Title:' text box, which is currently empty.





# Use Minitab to Run a One-Sample T-Test

- Null Hypothesis ( $H_0$ ): The data are normally distributed.
- Alternative Hypothesis ( $H_a$ ): The data are not normally distributed.
- Since the p-value of the normality is 0.275, which is greater than alpha level (0.05), we fail to reject the null and claim that the data are normally distributed.
- If the data are not normally distributed, you need to use hypothesis tests other than the one sample t-test.



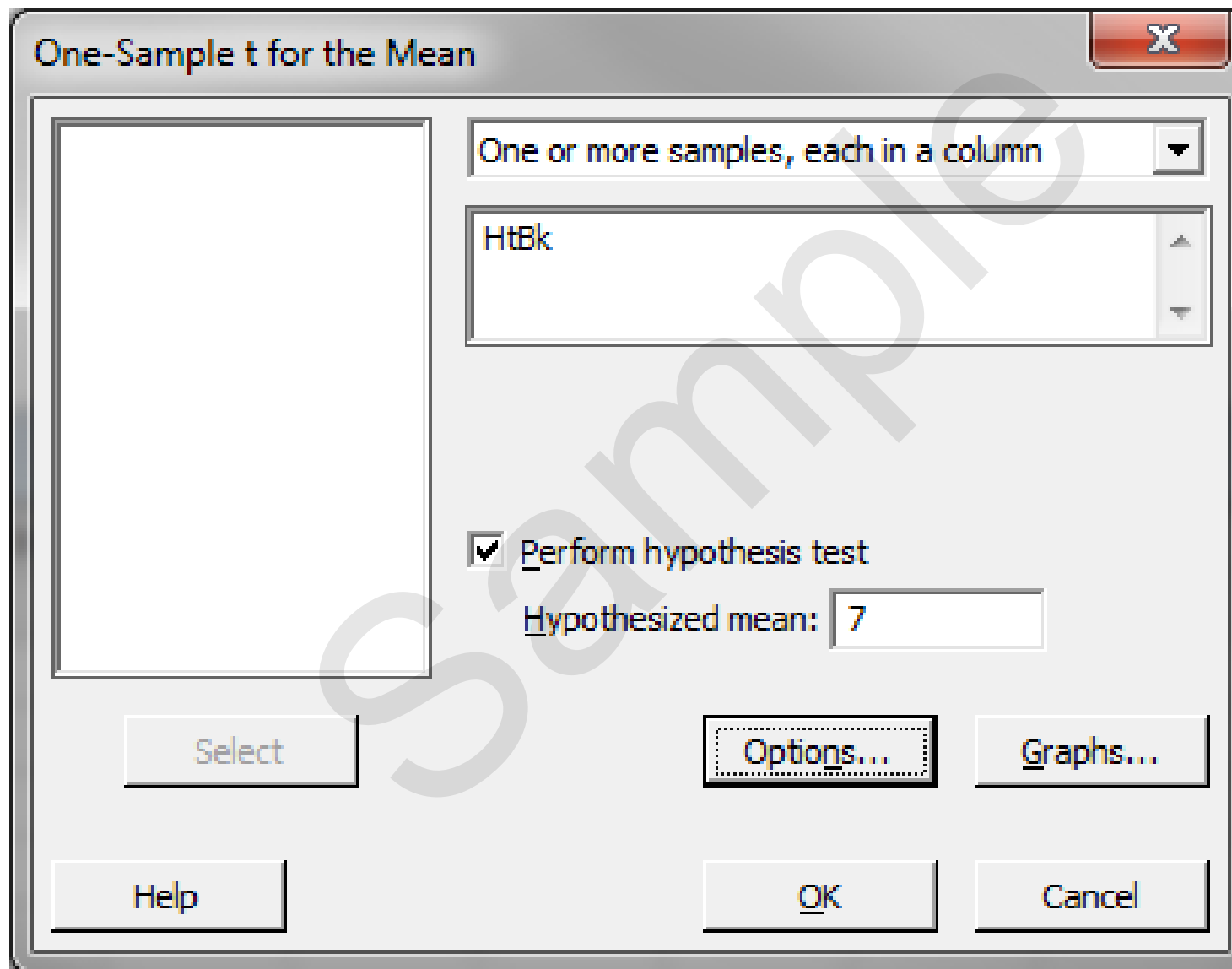
# Use Minitab to Run a One-Sample T-Test

---

- Step 2: Run the one-sample t-test
  - 1) Click Stat → Basic Statistics → 1 Sample t.
  - 2) A new window named “1 Sample t (Test and Confidence Interval)” pops up.
  - 3) Click in the blank box under “Samples in columns” and “HtBk” appears in the list box on the left.
  - 4) Select “HtBk” as the “Samples in columns.”
  - 5) Check the box of “Perform hypothesis test.”
  - 6) Enter the hypothesized value “7” into the box next to “Perform hypothesis test.”
  - 7) Click “OK.”
  - 8) The one-sample t-test result appears automatically in the session window.



# Use Minitab to Run a One-Sample T-Test



# Use Minitab to Run a One-Sample T-Test

- Null Hypothesis ( $H_0$ ):  $\mu = 7$
- Alternative Hypothesis ( $H_a$ ):  $\mu \neq 7$
- Since the p-value is smaller than alpha level (0.05), we reject the null hypothesis and claim that average of basketball players is statistically different from 7 feet.

## One-Sample T: HtBk

### Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for $\mu$
40	6.4532	0.3142	0.0497	(6.3528, 6.5537)

$\mu$ : mean of HtBk

### Test

Null hypothesis  $H_0: \mu = 7$   
Alternative hypothesis  $H_1: \mu \neq 7$

T-Value	P-Value
-11.00	0.000



## 4.2.2 Multiple Linear Regression

Sample



# What is Multiple Linear Regression?

---

- **Multiple linear regression** is a statistical technique to model the relationship between one dependent variable and two or more independent variables by fitting the data set into a linear equation.
- The difference between simple linear regression and multiple linear regression:
  - Simple linear regression only has one predictor.
  - Multiple linear regression has two or more predictors.



# Multiple Linear Regression Equation

---

$$Y = \alpha_1 * X_1 + \alpha_2 * X_2 + \dots + \alpha_p * X_p + \beta + e$$

- $Y$  is the dependent variable (response).
- $X_1, X_2, \dots, X_p$  are the independent variables (predictors). There are  $p$  predictors in total.
- Both dependent and independent variables are continuous.
- $\beta$  is the intercept indicating the  $Y$  value when all the predictors are zeros.
- $\alpha_1, \alpha_2, \dots, \alpha_p$  are the coefficients of predictors. They reflect the contribution of each independent variable in predicting the dependent variable.
- $e$  is the residual term indicating the difference between the actual and the fitted response value.



# Use Minitab to Run a Multiple Linear Regression

---

- *Case study:*
  - We want to see whether the scores in exam one, two, and three have any statistically significant relationship with the score in final exam. If so, how are they related to final exam score? Can we use the scores in exam one, two, and three to predict the score in final exam?
  - Data File: “Multiple Regression Analysis” tab in “Sample Data.xlsx.”
- Step 1: Determine the dependent and independent variables. All should be continuous.
  - Y (dependent variable) is the score of final exam.
  - $X_1$ ,  $X_2$ , and  $X_3$  (independent variables) are the scores of exam one, two, and three respectively.
  - All the variables are continuous.





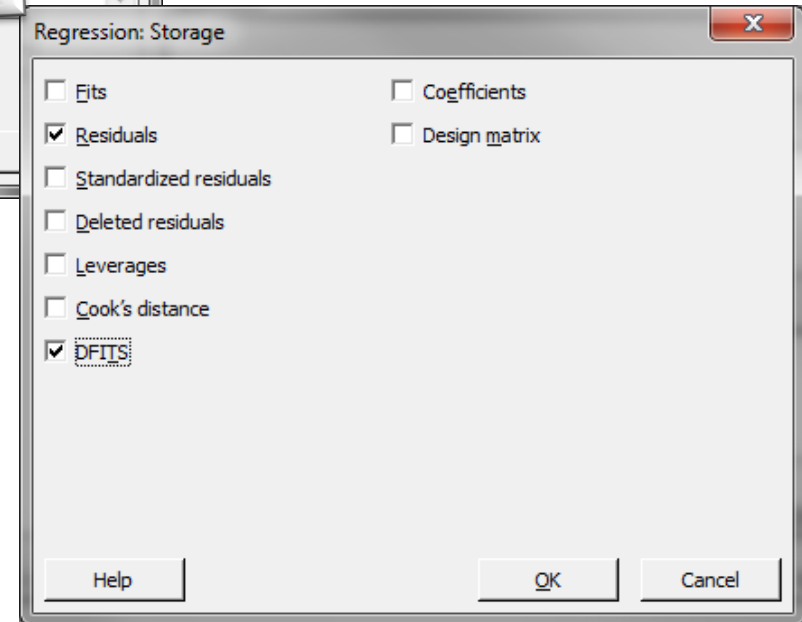
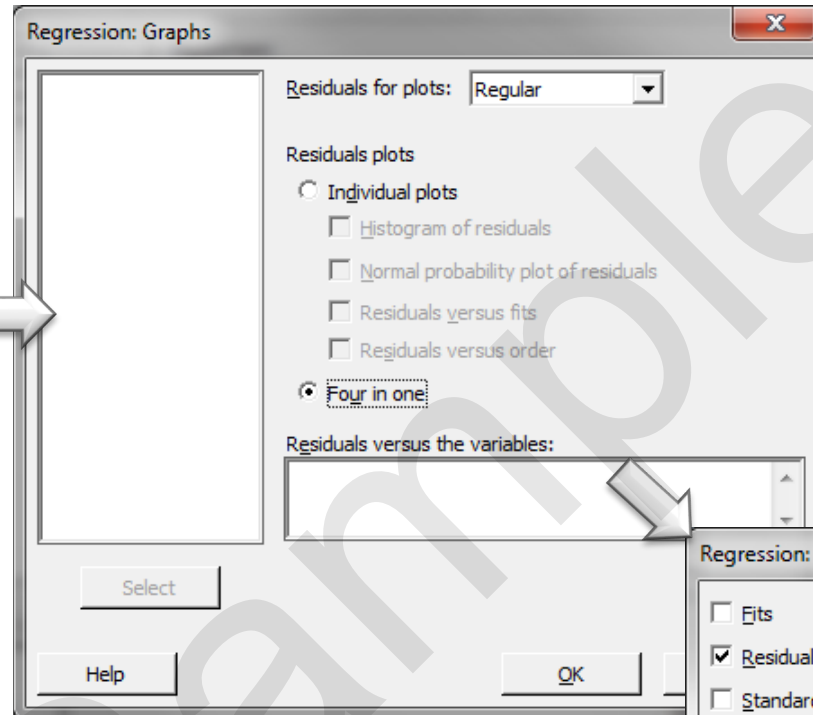
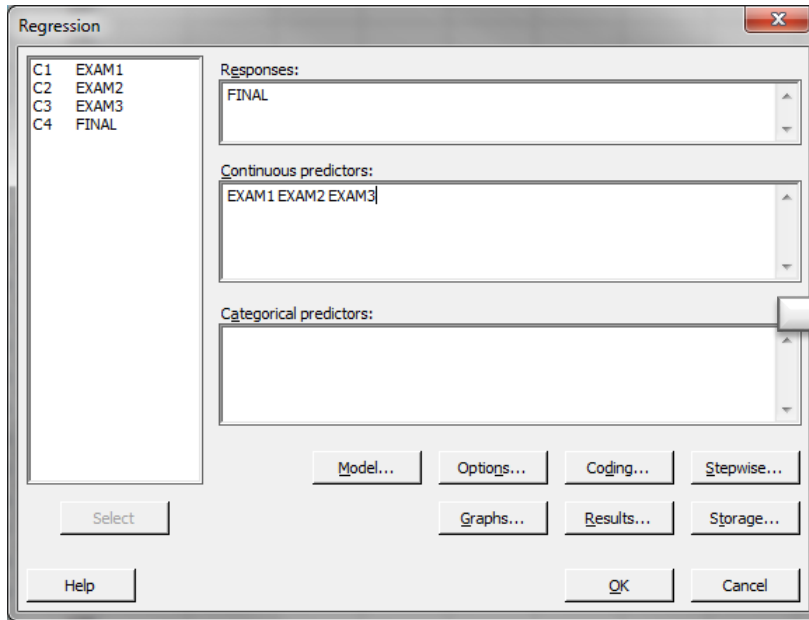
# Use Minitab to Run a Multiple Linear Regression

---

- Step 2: Start building the multiple linear regression model
  - 1) Click Stat → Regression → Regression → Fit Regression Model
  - 2) A new window named “Regression” pops up.
  - 3) Select “FINAL” as “Response” and “EXAM1”, “EXAM2” and “EXAM3” as “Continuous Predictors.”
  - 4) Click the “Graph” button, select the radio button “Four in one” and click “OK.”
  - 5) Click the “Storage” button, check the boxes of “Residuals” and “DFITS” and click “OK.”
  - 6) Click “OK” in the window named “Regression.”
  - 7) The regression analysis results appear in the session window and the four residual plots appear in another window named “Residual Plots for FINAL.”



# Use Minitab to Run a Multiple Linear Regression



# Use Minitab to Run a Multiple Linear Regression

- Step 3: Check whether the whole model is statistically significant. If not, we need to re-examine the predictors or look for new predictors before continuing.
  - $H_0$ : The model is not statistically significant (i.e., all the parameters of predictors are not significantly different from zeros).
  - $H_1$ : The model is statistically significant (i.e., at least one predictor parameter is significantly different from zero).
- In this example, p-value is much smaller than alpha level (0.05), hence we reject the null hypothesis; the model is statistically significant.

## Regression Analysis: FINAL versus EXAM1, EXAM2, EXAM3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	13731.5	4577.17	670.09	0.000
EXAM1	1	58.7	58.73	8.60	0.008
EXAM2	1	197.7	197.67	28.94	0.000
EXAM3	1	877.3	877.30	128.43	0.000
Error	21	143.4	6.83		
Total	24	13875.0			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.61357	98.97%	98.82%	98.51%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.34	3.76	-1.15	0.262	
EXAM1	0.356	0.121	2.93	0.008	7.81
EXAM2	0.543	0.101	5.38	0.000	5.59
EXAM3	1.167	0.103	11.33	0.000	5.16

### Regression Equation

$$\text{FINAL} = -4.34 + 0.356 \text{ EXAM1} + 0.543 \text{ EXAM2} + 1.167 \text{ EXAM3}$$

### Fits and Diagnostics for Unusual Observations

Obs	FINAL	Fit	Resid	Std Resid
23	175.00	167.69	7.31	2.93 R



# Use Minitab to Run a Multiple Linear Regression

- Step 4: Check whether multicollinearity exists in the model.
- The VIF information is automatically generated in the table Coefficients.

## Regression Analysis: FINAL versus EXAM1, EXAM2, EXAM3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	13731.5	4577.17	670.09	0.000
EXAM1	1	58.7	58.73	8.60	0.008
EXAM2	1	197.7	197.67	28.94	0.000
EXAM3	1	877.3	877.30	128.43	0.000
Error	21	143.4	6.83		
Total	24	13875.0			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.61357	98.97%	98.82%	98.51%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.34	3.76	-1.15	0.262	
EXAM1	0.356	0.121	2.93	0.008	7.81
EXAM2	0.543	0.101	5.38	0.000	5.59
EXAM3	1.167	0.103	11.33	0.000	5.16

### Regression Equation

$$\text{FINAL} = -4.34 + 0.356 \text{ EXAM1} + 0.543 \text{ EXAM2} + 1.167 \text{ EXAM3}$$

### Fits and Diagnostics for Unusual Observations

Obs	FINAL	Fit	Resid	Std Resid
23	175.00	167.69	7.31	2.93 R



# Multicollinearity

---

- **Multicollinearity** is the situation when two or more independent variables in a multiple regression model are correlated with each other.
- Although multicollinearity does not necessarily reduce the predictability for the model as a whole, it may mislead the calculation for individual independent variables.
- To detect multicollinearity, we use VIF (Variance Inflation Factor) to quantify its severity in the model.



# Variance Inflation Factor

---

- VIF quantifies the degree of multicollinearity for each individual independent variable in the model.
- VIF calculation:
  - Assume we are building a multiple linear regression model using  $p$  predictors.

$$Y = \alpha_1 \times X_1 + \alpha_2 \times X_2 + \dots + \alpha_p \times X_p + \beta$$

- Two steps are needed to calculate VIF for  $X_1$ .
    - Step 1: Build a multiple linear regression model for  $X_1$  by using  $X_2, X_3, \dots, X_p$  as predictors.
- $$X_1 = a_2 \times X_2 + a_3 \times X_3 + \dots + a_p \times X_p + b$$
- Step 2: Use the  $R^2$  generated by the linear model in step 1 to calculate the VIF for  $X_1$ .
  - Apply the same methods to obtain the VIFs for other  $X$ s. The VIF value ranges from one to positive infinity.

$$VIF = \frac{1}{1 - R^2}$$



# Variance Inflation Factor

---

- Rules of thumb to analyze variance inflation factor (VIF):
  - If  $VIF = 1$ , there is no multicollinearity.
  - If  $1 < VIF < 5$ , there is small multicollinearity.
  - If  $VIF \geq 5$ , there is medium multicollinearity.
  - If  $VIF \geq 10$ , there is large multicollinearity.



# How to Deal With Multicollinearity

---

- Increase the sample size.
- Collect samples with a broader range for some predictors.
- Remove the variable with high multicollinearity and high p-value.
- Remove variables that are included more than once.
- Combine correlated variables to create a new one.
- In this section, we will focus on removing variables with high VIF and high p-value.





# Use Minitab to Run a Multiple Linear Regression

---

- Step 5: Deal with multicollinearity:
  - Step 5.1: Identify a list of independent variables with VIF higher than 5. If no variable has VIF higher than 5, go to Step 6 directly.
  - Step 5.2: Among variables identified in Step 5.1, remove the one with the highest p-value.
  - Step 5.3: Run the model again, check the VIFs and repeat Step 5.1.
  - Note: we only remove one independent variable at a time.



# Use Minitab to Run a Multiple Linear Regression

---

- Step 6: Identify the statistically insignificant predictors. Remove one insignificant predictor at a time and run the model again. Repeat this step until all the predictors in the model are statistically significant.
  - Insignificant predictors are the ones with p-value higher than alpha level (0.05). When  $p > \alpha$  level, we fail to reject the null hypothesis; the predictor is not significant.
    - $H_0$ : The predictor is not statistically significant.
    - $H_1$ : The predictor is statistically significant.
  - As long as the p-value is greater than 0.05, remove the insignificant variables one at a time in the order of the highest p-value.
  - Once one insignificant variable is eliminated from the model, we need to run the model again to obtain new p-values for other predictors left in the new model.



# Use Minitab to Run a Multiple Linear Regression

- In this example, all three predictors have VIF higher than 5. Among them, EXAM1 has the highest p-value.
- We will remove EXAM1 from the equation and run the model again.

## Regression Analysis: FINAL versus EXAM1, EXAM2, EXAM3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	13731.5	4577.17	670.09	0.000
EXAM1	1	58.7	58.73	8.60	0.008
EXAM2	1	197.7	197.67	28.94	0.000
EXAM3	1	877.3	877.30	128.43	0.000
Error	21	143.4	6.83		
Total	24	13875.0			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.61357	98.97%	98.82%	98.51%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.34	3.76	-1.15	0.262	
EXAM1	0.356	0.121	2.93	0.008	7.81
EXAM2	0.543	0.101	5.38	0.000	5.59
EXAM3	1.167	0.103	11.33	0.000	5.16

### Regression Equation

$$\text{FINAL} = -4.34 + 0.356 \text{ EXAM1} + 0.543 \text{ EXAM2} + 1.167 \text{ EXAM3}$$

### Fits and Diagnostics for Unusual Observations

Obs	FINAL	Fit	Resid	Std Resid	R
23	175.00	167.69	7.31	2.93	R



# Use Minitab to Run a Multiple Linear Regression

---

- Run the new multiple linear regression with only two predictors (i.e., EXAM2 and EXAM3).
- Check the VIFs of EXAM2 AND EXAM3. They are both smaller than 5; hence, there is little multicollinearity existing in the model.

Sample



# Use Minitab to Run a Multiple Linear Regression

- In this example, both predictors' p-values are smaller than alpha level (0.05).
- As a result, we do not need to eliminate any variables from the model.

## Regression Analysis: FINAL versus EXAM2, EXAM3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	13672.8	6836.39	743.91	0.000
EXAM2	1	555.1	555.11	60.41	0.000
EXAM3	1	1686.0	1685.97	183.46	0.000
Error	22	202.2	9.19		
Total	24	13875.0			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.03146	98.54%	98.41%	98.18%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.34	4.37	-0.99	0.331	
EXAM2	0.7222	0.0929	7.77	0.000	3.53
EXAM3	1.3375	0.0987	13.54	0.000	3.53

### Regression Equation

$$\text{FINAL} = -4.34 + 0.7222 \text{ EXAM2} + 1.3375 \text{ EXAM3}$$

### Fits and Diagnostics for Unusual Observations

Obs	FINAL	Fit	Resid	Std Resid
17	147.00	140.78	6.22	2.23 R



# Use Minitab to Run a Multiple Linear Regression

- Step 7: Interpret the regression equation
  - The multiple linear regression equation appears automatically at the top of the session window.
  - “Coefficients” section provides the estimates of parameters in the linear regression equation.

## Regression Analysis: FINAL versus EXAM2, EXAM3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	13672.8	6836.39	743.91	0.000
EXAM2	1	555.1	555.11	60.41	0.000
EXAM3	1	1686.0	1685.97	183.46	0.000
Error	22	202.2	9.19		
Total	24	13875.0			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.03146	98.54%	98.41%	98.18%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.34	4.37	-0.99	0.331	
EXAM2	0.7222	0.0929	7.77	0.000	3.53
EXAM3	1.3375	0.0987	13.54	0.000	3.53

### Regression Equation

FINAL = -4.34 + 0.7222 EXAM2 + 1.3375 EXAM3

### Fits and Diagnostics for Unusual Observations

Obs	FINAL	Fit	Resid	Std Resid
17	147.00	140.78	6.22	2.23 R



# Interpreting the Results

---

- $R^2_{\text{Adj}} = 98.4\%$ 
  - 98% of the variation in FINAL can be explained by the predictor variables EXAM2 & EXAM3.
- P-value of the F-test = 0.000
  - We have a statistically significant model.
- Variables p-value:
  - Both are significant (less than 0.05).
- VIF
  - EXAM2 & EXAM3 are both below 5; we're in good shape!
- Equation:  $-4.34 + 0.722 \cdot \text{EXAM2} + 1.34 \cdot \text{EXAM3}$ 
  - -4.34 is the Y intercept, all equations will start with -4.34.
  - 0.722 is the EXAM2 coefficient; multiply it by EXAM2 score.
  - 1.34 is the EXAM3 coefficient; multiply it by EXAM3 score.



# Interpreting the Results

---



- Let us say you are the professor again, and this time you want to use your prediction equation to estimate what one of your students might get on their final exam.

- Assume the following:
  - Exam 2 results were: 84
  - Exam 3 results were: 102.
- Use your equation:  $-4.34 + 0.722*EXAM2 + 1.34*EXAM3$
- Predict your student's final exam score:
  - $-4.34 + (0.722*84) + (1.34*102) = -4.34 + 60.648 + 136.68 = \mathbf{192.988}$



Nice work again! Now you can use your “magic” as the smart and efficient professor and allocate your time to other students because this one projects to perform much better than the average score of 162.







# LEAN SIGMA CORPORATION

Lean Six Sigma Black Belt Training  
Featuring Examples from Minitab 18

