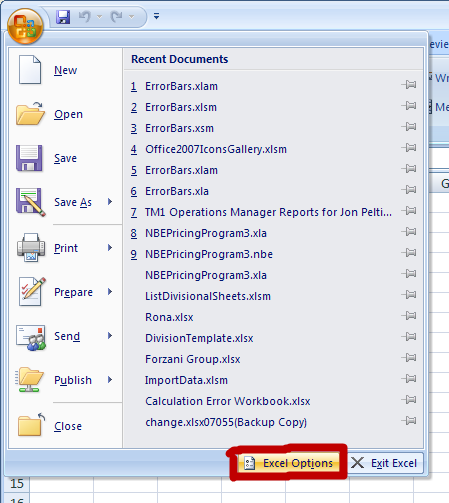
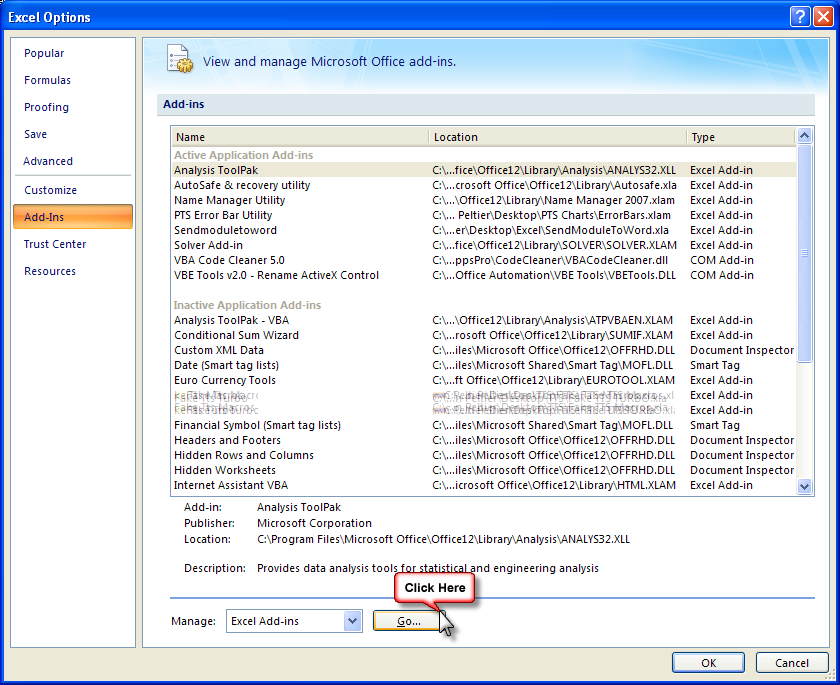
**Statistics using Excel**

1. **How to install Analysis ToolPack Add-In**

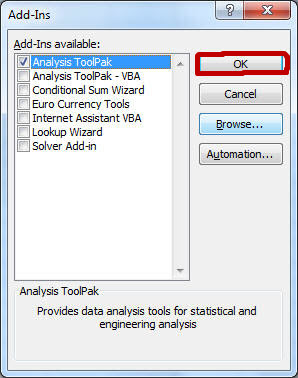
Step 1: Click on the **Excel Options** located at the bottom of the windows menu.



Step2: Select **Add–Ins** option and then Click on the Go button.



Step3: In the dialog box appeared on the screen Select **Analysis ToolPack** and then press Ok button.



1. **Descriptive Statistics**

It answers the following questions:

* What is the value that best describes the data set?
* How much a data set speads from its average value?
* What is the smallest and largest number in a data set?

It provides information on summary statistics that includes Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count.

In other words, it consists of measures of **central tendency** and **variability.**

**Measures of central tendency** are used to find the single value that best describes about the entire distribution. There are three main measures of central tendency: **Mean, Median and Mode.**

**Measures of Variability** refers to the spread or dispersion of scores**.** There are four main measures of variability: **Range, Inter quartile range and Standard deviation and Variance.**

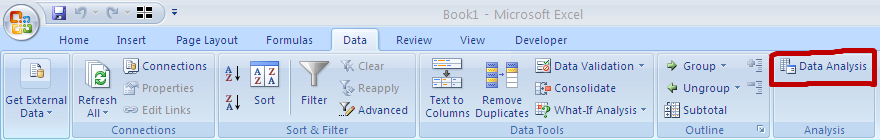
**Example 1:** Suppose you are asked to calculate the average asset value of top stock funds and check whether there is any variability in the assets of these stock funds. You would answer this question with a **measure of central tendency and variability.**

**Example 2:** Suppose you are asked to provide a figure that best describes the annual salary offered to students in ABC College you would answer with a **measure of central tendency and variability.**

Using Excel:

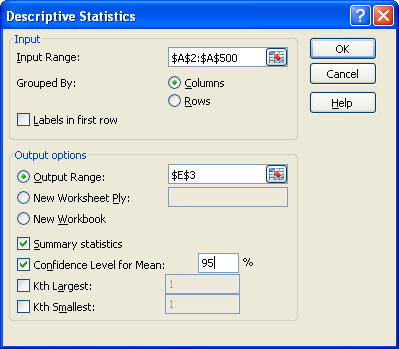
1. If you haven’t already installed the **Analysis ToolPack ,** Click the **Microsoft Office Button**, then click on the **Excel Options** , and then select **Add-Ins** , Click Go, check the **Analysis ToolPack** box, and click Ok.
2. Select **Data** tab, then click on the **Data Analysis** option, then selects **Descriptive Statistics** from the list and Click **Ok**.

(Data tab >> Data Analysis >> Descriptive Statistics)

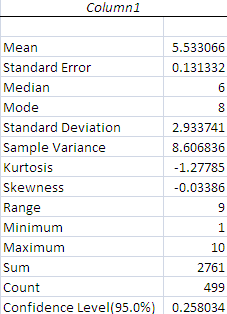




1. In the **Input Range** we select the data, and then select **Output Range** where you want the output to be stored. If you don’t specify the output range it will throw output in the new worksheet.
2. Check **Summary Statistics and Confidence Level for Mean** options. By default the confidence level is 95%. You can change the level as per the hypothesis standard of study.



1. When you click Ok, you will see the result in the selected output range.



1. **Normal Distribution**

**Problem 1**: Suppose you are asked to calculate the chances that income of a randomly selected person will be less than 3000 dollars. It is known that the income levels are normally distributed with a mean of 3500 and standard deviation of 500.

Using Excel:

1. Go to **Formulas** tab , then Press **Insert Function** button and Choose **Statistical** then **NORMDIST** from the function name , Cick **OK**
2. **NORMDIST** function returns the normal distribution for the specified mean and standard deviation

**SYNTAX**:

*= NORMDIST (value for which you want the distribution, mean, standard deviation,* ***cumulative distribution or probability mass function****)*

Note: If fourth part of the above syntax is **TRUE,** NORMDIST returns the **cumulative distribution function**; if it is **FALSE**, it returns the **probability mass function**.

**=NORMDIST (3000, 3250, 500, TRUE)**

**Problem 2:** Suppose you are asked to calculate the income level which could be exceeded by only top 25 percent income bracket individuals. This means to calculate the income of an individual whose percentile is 75.

Using Excel:

1. Go to **Formulas** tab , then Press **Insert Function** button and Choose **Statistical** then **NORMINV**  from the function name , Cick **OK**
2. **NORMINV** function returns the inverse of the normal distribution for the specified mean and standard deviation

**SYNTAX:**

*= NORMINV (probability, mean, standard deviation)*

**=NORMINV (0.75, 3250, 500)**

**4. Correlation**

**Correlation** is used to measure strength of the relationship between two variables. It can be positive, negative or zero.

In a **positive relationship** both variables tend to move in the same direction: If one variable increases, the other tends to increase. If one decreases, the other tends to decrease.

In a **negative relationship** the variables tend to move in the opposite directions: If one variable increases, the other tends to decrease, and vice-versa.

Note: The correlation coefficient may take on any value between +1 and -1.

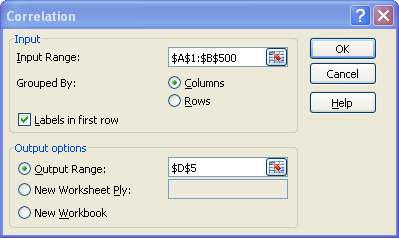
Example 1: How are sales of ABC Company and GDP related?

Example 2: How are annual returns on Treasury Bills and Bonds related?

Problem1: Suppose you would like to know whether there is a relationship between grades and number of hours studied.

Using Excel:

1. If you haven’t already installed the **Analysis ToolPack ,** Click the **Microsoft Office Button**, then click on the **Excel Options** , and then select **Add-Ins** , Click Go, check the **Analysis ToolPack** box, and click Ok.
2. Go to **Data** tab and click on **Data Analysis**
3. When **Data Analysis** dialog appears, Click on **Correlation**.



1. Select the data range (both independent and dependent variable) in the **Input Range** box.
2. Check Labels in first row and enter range in the output range box and Click on Ok.

|  |  |  |
| --- | --- | --- |
|  | *Grade* | *Number of Hours Studied* |
| Grade | 1 |  |
| Number of Hours Studied | 0.050575041 | 1 |

**Simple Linear Regression**

**Regression** is used to find a model that predicts one variable in terms of the other variable.Regression analysis fits a straight line to the relationship between the two variables.

The general equation for a straight line is:

Y = a + b X

a : Intercept: The intercept identifies the value of Y when Xi’s is zero.

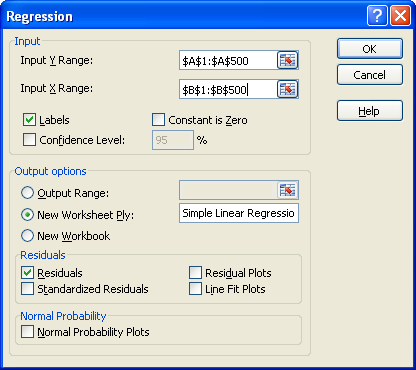
b: The slope specifies how much the variable Y will change when X changes by one unit.

Problem 1: Suppose you want to predict your high grade score using the number of hours you studied.

High Grade Score = Constant + Slope\* (Number of hours studied)

**Using Excel:**

1. If you haven’t already installed the **Analysis ToolPack ,** Click the **Microsoft Office Button**, then click on the **Excel Options** , and then select **Add-Ins** , Click Go, check the **Analysis ToolPack** box, and click Ok.
2. Go to **Data** tab and click on **Data Analysis**
3. When **Data Analysis** dialog appears, Click on **Regression**.
4. When Regression dialog box appears, enter dependent variable data in **Input Y Range and independent variable in Input X Range.**



1. Specify range in the Output Range box. By default it throws output in the new worksheet. You can name the new worksheet.
2. Checking the Residuals box list commands excel to list Residual Output that includes Predicted Grade and Residuals ( Observed Grade – Predicted Cost )
3. To calculate the best predicted grade, you have to find coefficient from the summary output.

The relationship between the number of number of hours studied and the high grade score is

**Y = 80.68 +0.1110 \* (No. of hours studied)**

Run the simple linear regression analysis using **LINEST** function.

*= LINEST (dependent variable data range, independent variable(s) data range, Constant required or not? , additional regression statistics required or not ? )*

Note:

1. If value of **third** part of the above formula is **TRUE, Intercept** is calculated normally.
2. If value of **third** part of the above formula is **FALSE, Intercept** is set equal to **0**
3. If value of **fourth** part of the above formula is **TRUE,** excel returns **additonal regression statistics** such as R square, F ratio, standard error etc. ( Press F2 and Hit **Ctrl Shift Enter** to set this as an array formula)
4. If value of **fourth** part of the above formula is **FALSE,** excel returns only **slope and intercept** of the line.

**Problem**: To predict grade score for the 16 number of hours studied.

=SUM (LINEST (Correlation!A2:A500, Correlation!B2:B500)\*{16,1})

Note: *SUM({m,b}\*{x,1}) equals mx + b, the estimated y-value for a given x-value*

**Multiple Regression**

**Regression** is used to find a model that predicts one variable in terms of the other variable.Regression analysis fits a straight line to the relationship between the two variables.

The general equation for a straight line is:

Y = a + b1 X1 +b2 X2 + ℮i

a : Intercept: The intercept identifies the value of Y when Xi’s are zero.

b1: Slope: Slope of Y with variable X1 holding X2 constant

b2: Slope: Slope of Y with variable X2 holding X1 constant

The slope specifies how much the variable Y will change when X changes by one unit.

Example 1: Suppose you want to predict monthly sales of ABC Company using the size of workforce, amount of advertising expenditures, price.

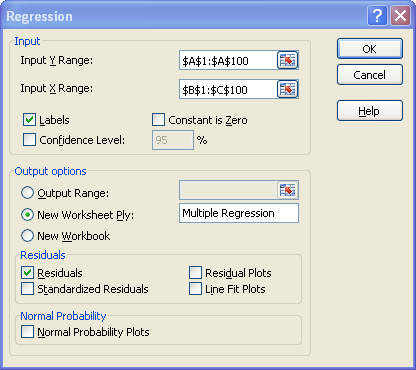
Example 2: Suppose you want to predict your monthly salary using number of years of services, years of education etc

**Problem1**: To predict cost of running auto plant using number of cars of Model A and B produced.

Cost = Constant + Slope1\* (Units of Model A produced) + Slope2\* (Units of Model B produced)

**Using Excel:**

1. If you haven’t already installed the **Analysis ToolPack ,** Click the **Microsoft Office Button**, then click on the **Excel Options** , and then select **Add-Ins** , Click Go, check the **Analysis ToolPack** box, and click Ok.
2. Go to **Data** tab and click on **Data Analysis**
3. When **Data Analysis** dialog appears, Click on **Regression**.
4. When Regression dialog box appears, enter dependent variable data in **Input Y Range and independent variable(s) in Input X Range.**



1. Specify range in the Output Range box. By default it throws output in the new worksheet. You can name the new worksheet.
2. Checking the Residuals box list commands excel to list Residual Output that includes Predicted Cost and Residuals ( Observed Cost – Predicted Cost )
3. To calculate the best predicted cost, you have to find coefficients from the summary output.

The relationship between the number of units Model A and B produced and the cost of running auto plant is  **Y = 116230.5044 -24.62791049 \* (Model A) -3.739179584 \* (Model B)**

Run the multiple regression analysis using **LINEST** function.

*= LINEST (dependent variable data range, independent variable(s) data range, Constant required or not? , additional regression statistics required or not ? )*

Note:

1. If value of **third** part of the above formula is **TRUE, Intercept** is calculated normally.
2. If value of **third** part of the above formula is **FALSE, Intercept** is set equal to **0**
3. If value of **fourth** part of the above formula is **TRUE,** excel returns **additonal regression statistics** such as R square, F ratio, standard error etc. ( Press F2 and Hit **Ctrl Shift Enter** to set this as an array formula)
4. If value of **fourth** part of the above formula is **FALSE,** excel returns only **slope and intercept** of the line.

**=LINEST ('Raw Data'! A2:A100,'Raw Data'! B2:C100, TRUE, TRUE)**

How to use this formula:

* First select four row and 3 column i.e. 4R X 3C
* Put the above formula
* Press **F2** and then **Hit Ctrl Shift Enter**

**Confidence Interval**

For a 95 % confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95 % of these intervals would contain the true mean. It does not mean that there is a 95 % probability that the interval contains the true mean.

**Theoretical Rule:**

1. For Sample Size (n) > 30 we use Z test
2. For Sample Size (n) <=30 we use T test

In real world we use T test as t test will approximate the Z statistic at sample sizes approaching infinity.

**For n >= 30 (Z test)**

The general formula for developing a confidence interval for a population means is:

http://home.ubalt.edu/ntsbarsh/Business-stat/excel/formula1.gif

In this formula http://home.ubalt.edu/ntsbarsh/Business-stat/excel/symbol1.gif is the mean of the sample; Z is the interval coefficient, which can be found from the normal distribution table (for example the interval coefficient for a 95% confidence level is 1.96). S is the standard deviation of the sample and n is the sample size.

**For n< 30 (T test)**

The general formula for developing confidence intervals for the population mean based on small a sample is:

http://home.ubalt.edu/ntsbarsh/Business-stat/excel/formula4.GIF

In this formula http://home.ubalt.edu/ntsbarsh/Business-stat/excel/symbol1.gif is the mean of the sample. http://home.ubalt.edu/ntsbarsh/Business-stat/excel/symbol2.gif is the interval coefficient providing an area of http://home.ubalt.edu/ntsbarsh/Business-stat/excel/symbol3.gifin the upper tail of a t distribution with n-1 degrees of freedom which can be found from a t distribution table (for example the interval coefficient for a 90% confidence level is 1.833 if the sample is 10). S is the standard deviation of the sample and n is the sample size.

**Note:** Excel calculates confidence interval of a population mean based on sample size lesser than 30 the same way it does for sample size greater than 30

**Using Excel:**

1. If you haven’t already installed the **Analysis ToolPack ,** Click the **Microsoft Office Button**, then click on the **Excel Options** , and then select **Add-Ins** , Click Go, check the **Analysis ToolPack** box, and click Ok.
2. Go to **Data** tab and click on **Data Analysis**
3. When **Data Analysis** dialog appears, Click on **Descriptive Statistics**
4. In the **Input Range** we select the data, and then select **Output Range** where you want the output to be stored. If you don’t specify the output range it will throw output in the new worksheet.
5. Check **Summary Statistics and Confidence Level for Mean** options. By default the confidence level is 95%. You can change the level as per the hypothesis standard of study.

The confidence interval is Mean +/- Confidence level