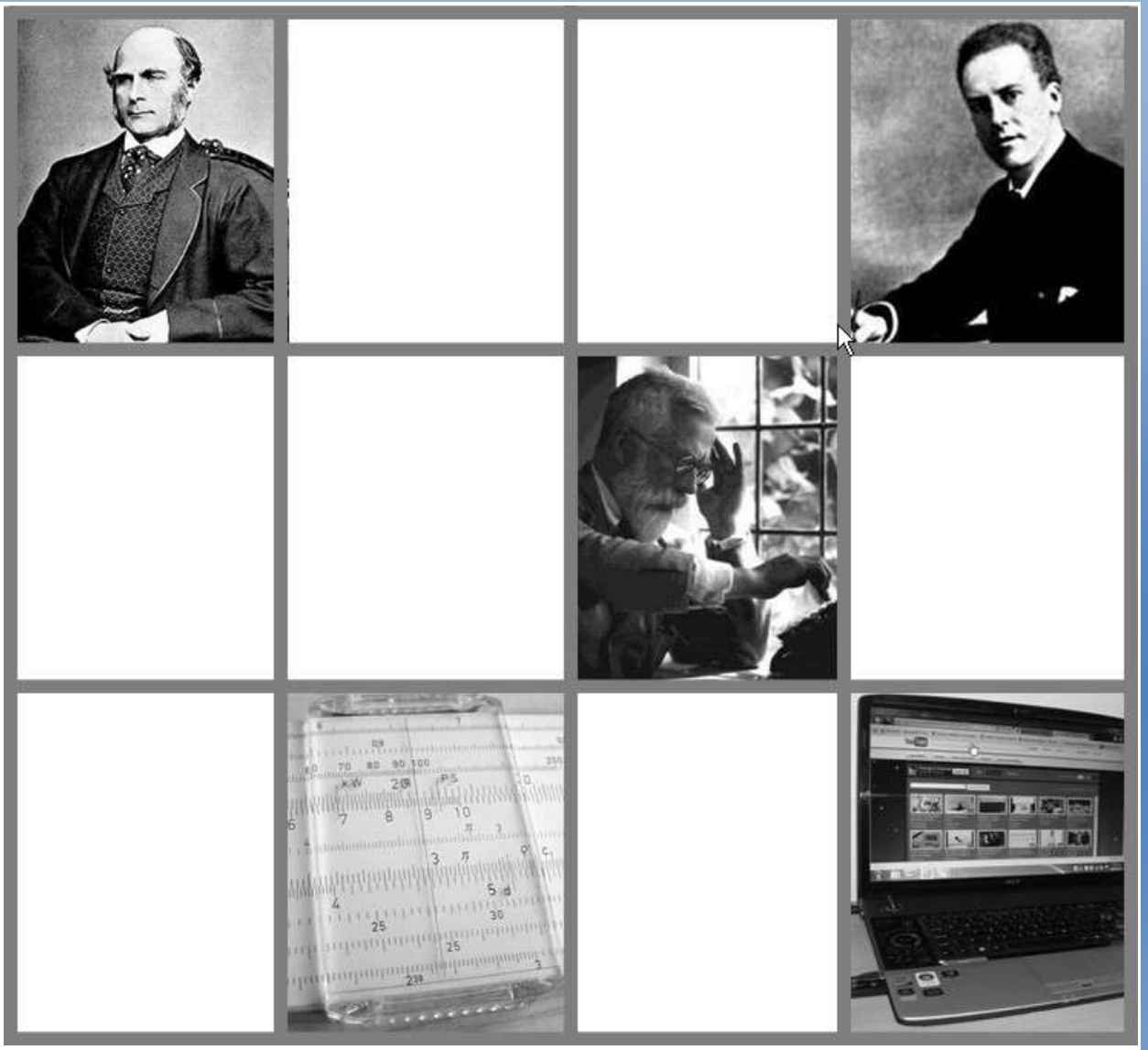


GRAD SCHOOL JUNGLE

STATISTICALLY EXCELLENT

Inferential Statistics using MS Excel 2007 | **MIKE ARIEH MEDINA**

Disclosures: MS Excel 2007 is a copyrighted commercial software application. This eBook is not intended to promote the sale of such product nor use its name for personal monetary gain.

Cover Photo: (From upper left to lower right) Francis Galton, Karl Pearson, Ronald Fisher, a typical slide rule, a modern laptop computer. (Photos taken from: *Wikimedia Commons*)

Please feel free to post this on your blog or email it to whomever you believe would benefit from reading it. **Thank you.**

Statistically Excellent

Inferential Statistics using MS Excel 2007

What's this all about?

In one of the previous post in my blog Grad School Jungle, I have stressed out the importance of a dependable computer set when you are enrolled in a graduate school program. One of the reasons is in data analysis. Following is an excerpt from the post entitled "*Computer: a grad student's best friend*":

> "A decade ago statistical computations for grad researches are done either manually or with the help of a scientific calculator. You'll need pages of scratch paper for this. Nowadays, statistical software such as SPSS, SAS, STATA, and Minitab can compute large amounts of data in seconds.

However, I fail to mention that these are expensive applications in which only institutions can afford and justify the purchase of a licensed version. Individual students therefore can only resort to the trial or student versions of the software. These sometimes have limited features and can only be functional for a short period of time.

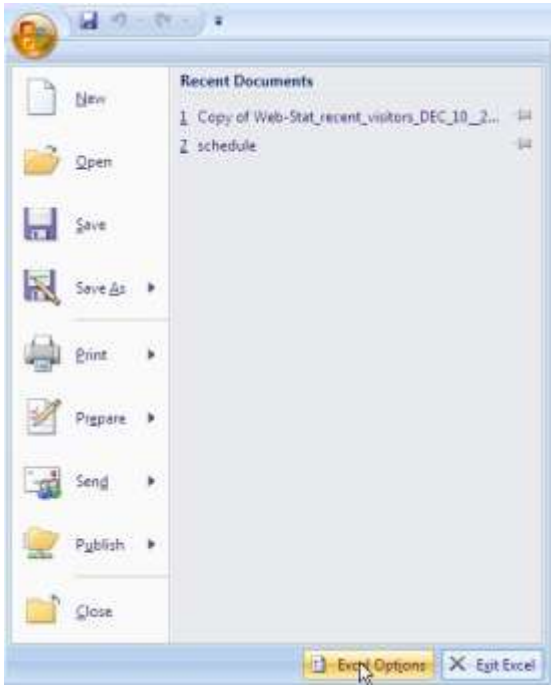
In this regard, I am providing you with this eBook about analyzing statistical data using MS Excel, a Microsoft Office Application for spreadsheets. Although Excel cannot exceed the mainstream statistical software packages in terms of performance, this eBook will try to teach you how to analyze small datasets using inferential statistics and along the way increase your understanding about the statistical analysis minus the hassles of the manual computations.

Mike Arieh Medina

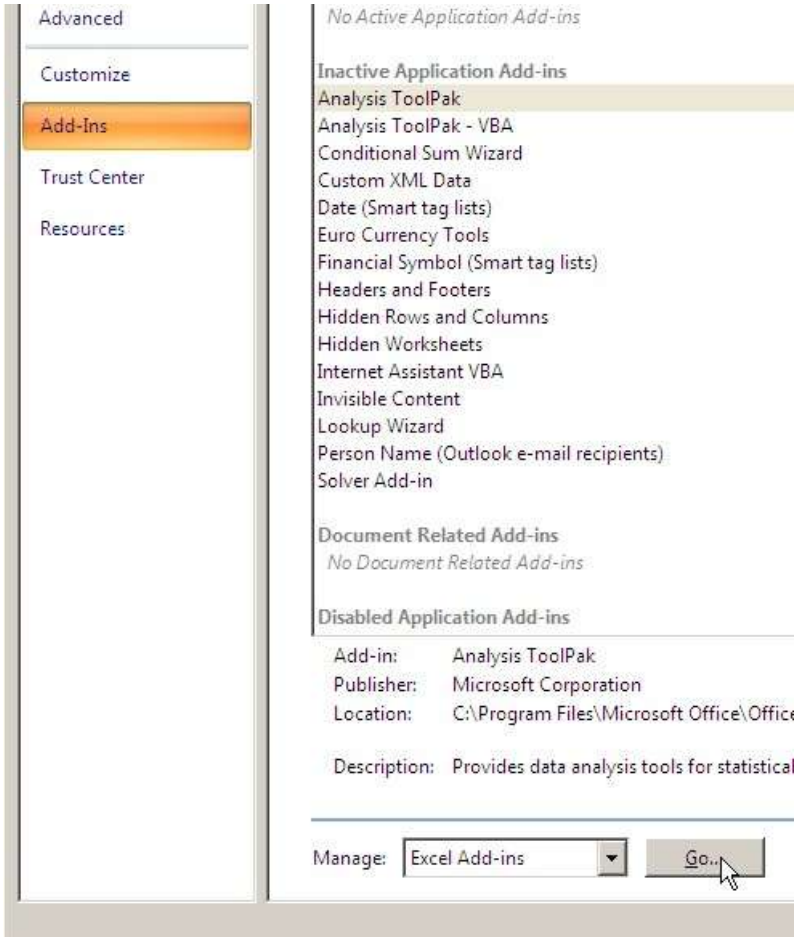
How to load the MS Excel add-in for data analysis

The Analysis ToolPak is a Microsoft Office Excel add-in. This is a program that is available when you install Microsoft Office or Excel. This add-in provides an indispensable tool for data analysis be it for descriptive or inferential statistics for your data set. After installing MS Office however, you don't see it in the Data tab during your first use. In order to use it in Excel you need to load it first.

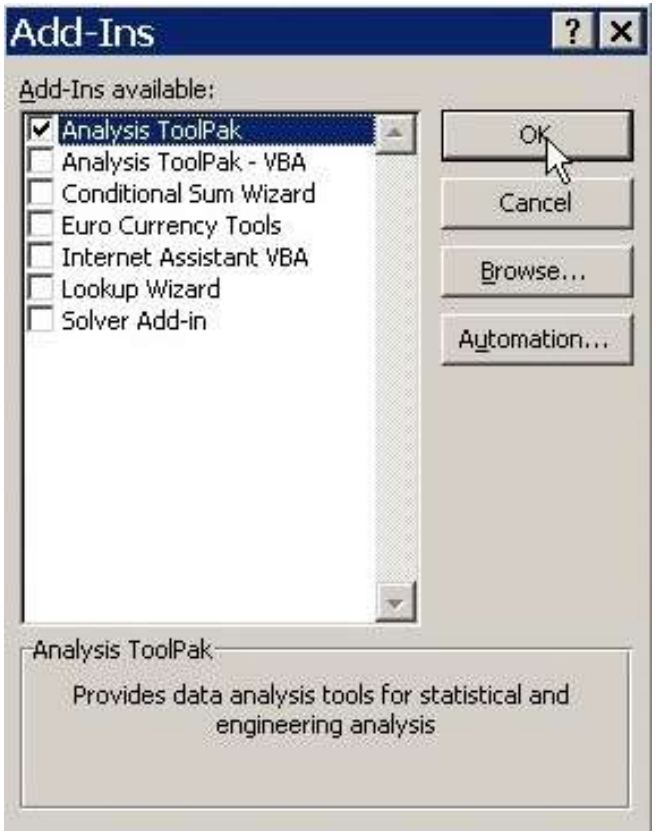
1. Once you have opened MS Excel, click the **Microsoft Office Button**, and then click **Excel Options** at the lower left part of the dialogue box.



2. Click **Add-Ins** found in the left panel of the dialogue box, and then in the **Manage** box found at the lower part, select **Excel Add-ins**, then click **Go.**



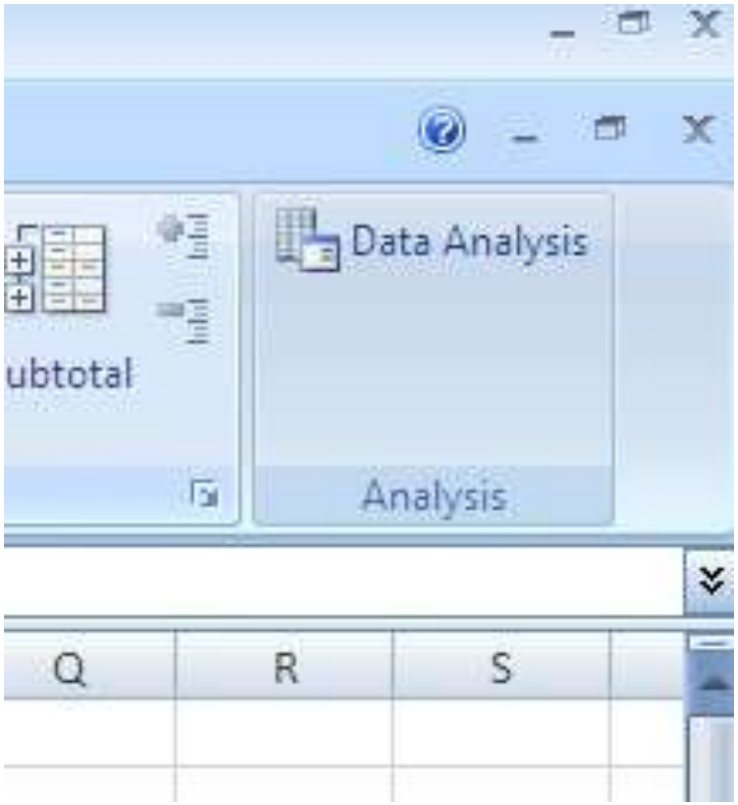
3. In the **Add-Ins available** dialogue box, click the check box for the **Analysis ToolPak**, and then click **OK**.



Note: If **Analysis ToolPak** is not found in the **Add-Ins available** dialogue box, click **Browse** to locate it.

If you get a notification that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

4. After you have loaded the Analysis ToolPak, the **Data Analysis** icon will now be available in the **Analysis** group on the **Data** tab.

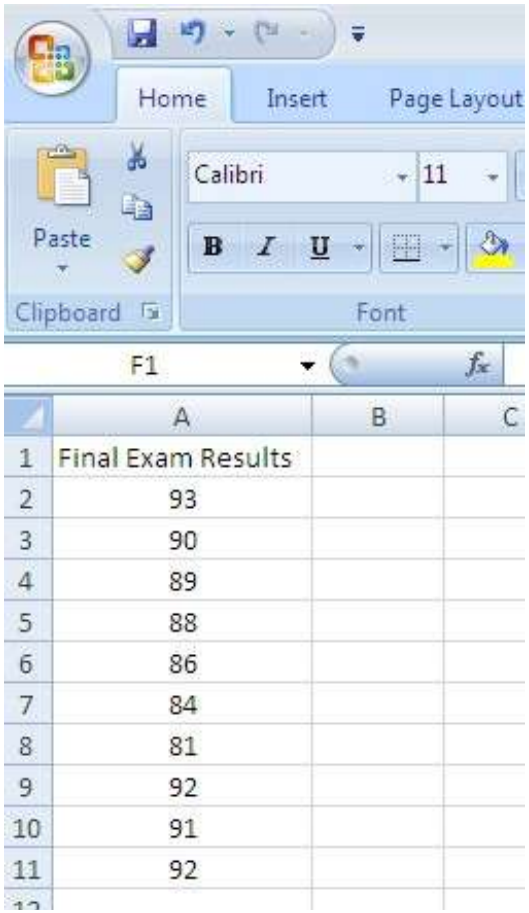


Let's try how this add-in works! Read the next chapter to try some simple descriptive statistics using the **Data Analysis** command.

How to perform descriptive statistics using MS Excel

Here's a way to do some descriptive statistics using the MS Excel Analysis ToolPak add-in.

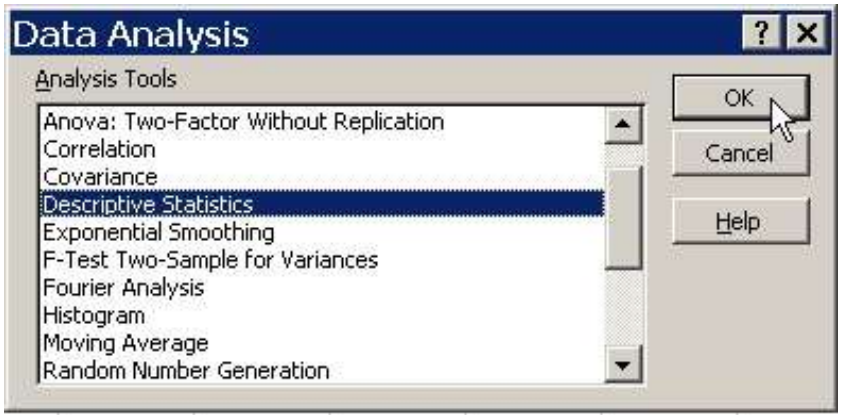
1. Encode your data vertically in a worksheet.



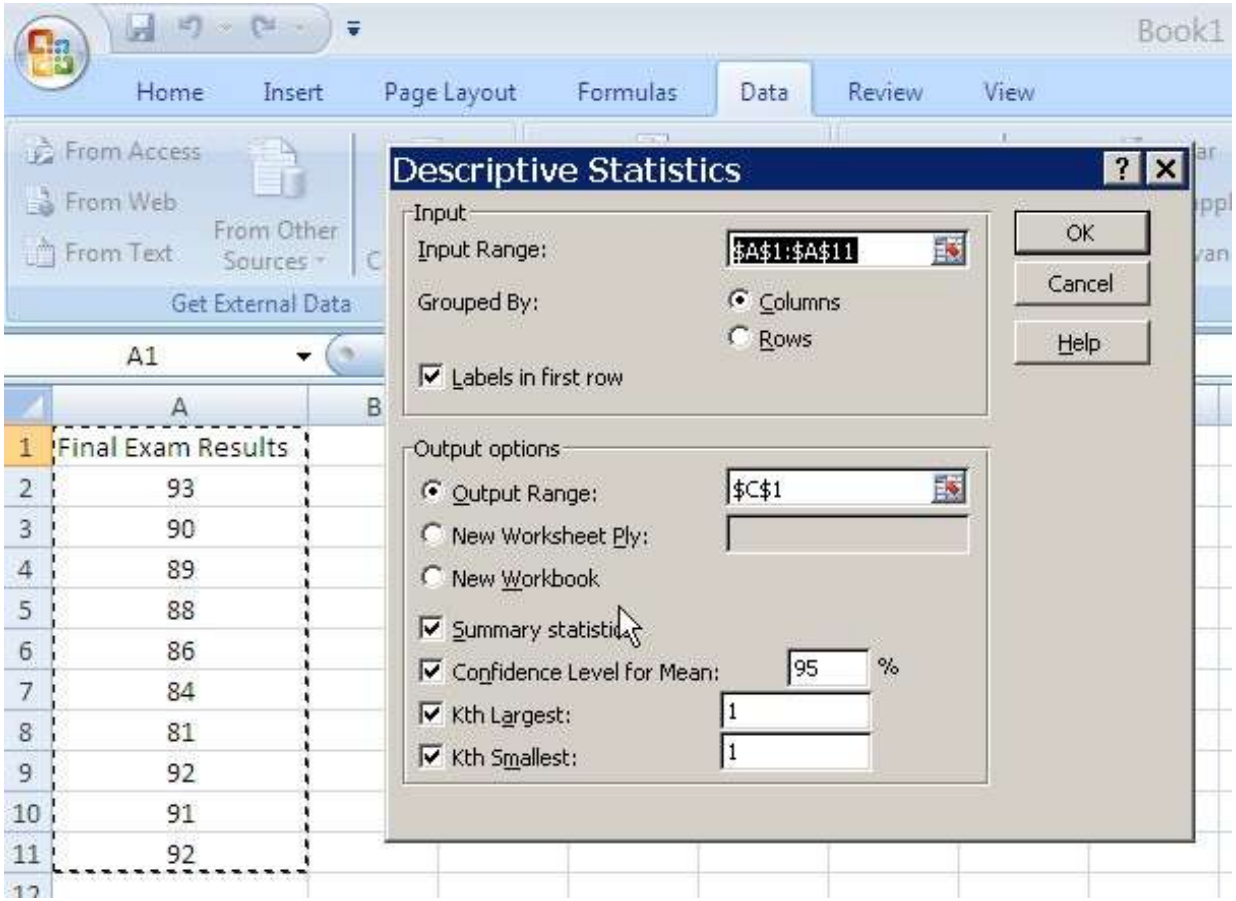
	A	B	C
1	Final Exam Results		
2	93		
3	90		
4	89		
5	88		
6	86		
7	84		
8	81		
9	92		
10	91		
11	92		

2. Click the **Data tab**, and then click the **Data Analysis** command icon.

3. In the Data Analysis dialogue box, select descriptive statistics, and then click Ok.



4. In the Descriptive Analysis dialogue box, enter the input range and the output range values, check the descriptive statistics you wish to be computed (summary statistics, confidence level, etc.) and then click Ok. Note: If you have included the column title in the input range, check Labels in first row.



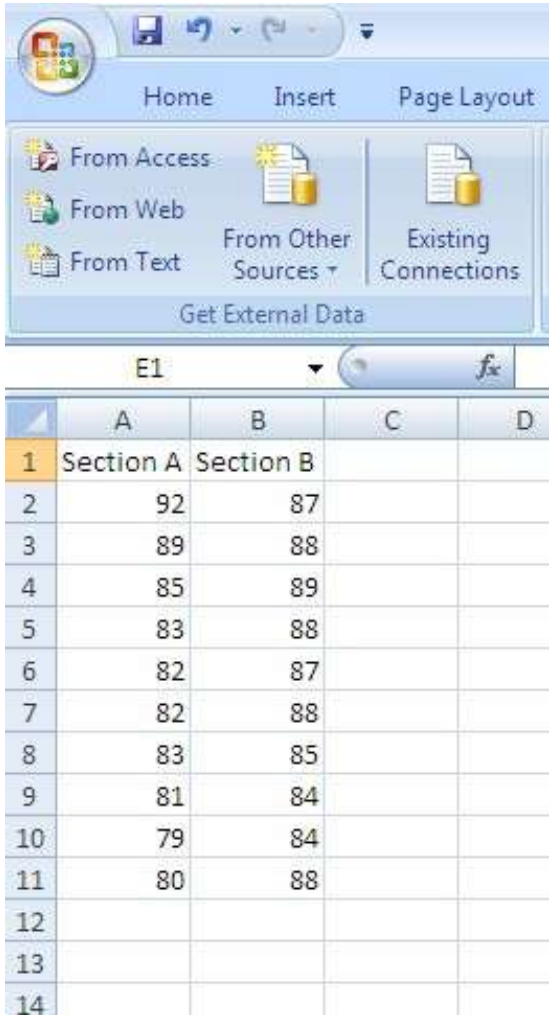
5. The results will be displayed in the output range you have specified.

C	D
Final Exam Results	
Mean	88.6
Standard Error	1.231079021
Median	89.5
Mode	92
Standard Deviation	3.893013686
Sample Variance	15.15555556
Kurtosis	-0.054883982
Skewness	-0.87964931
Range	12
Minimum	81
Maximum	93
Sum	886
Count	10
Largest(1)	93
Smallest(1)	81
Confidence Level(95.0%)	2.784894219

Now, let's try hypothesis testing in the next chapter.

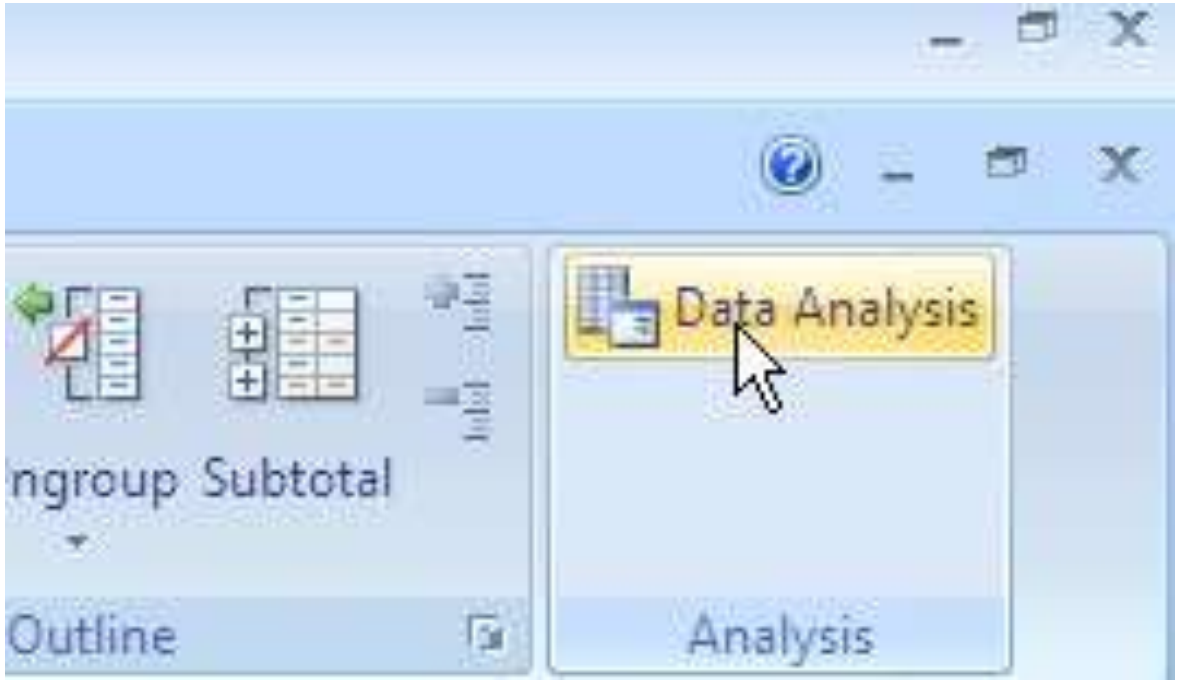
Performing a t-test in Excel

1. Encode your data set in an Excel worksheet.

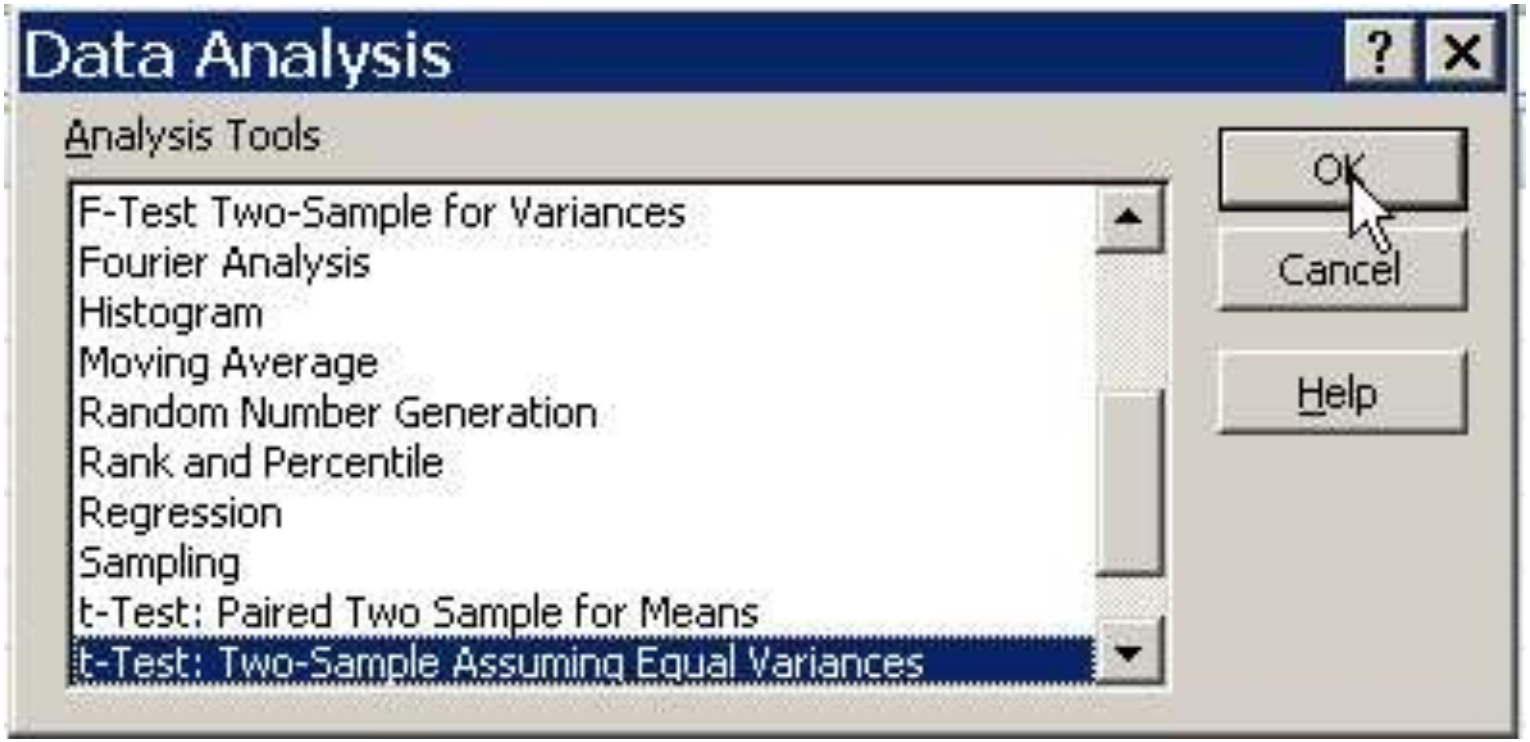


	A	B	C	D
1	Section A	Section B		
2	92	87		
3	89	88		
4	85	89		
5	83	88		
6	82	87		
7	82	88		
8	83	85		
9	81	84		
10	79	84		
11	80	88		
12				
13				
14				

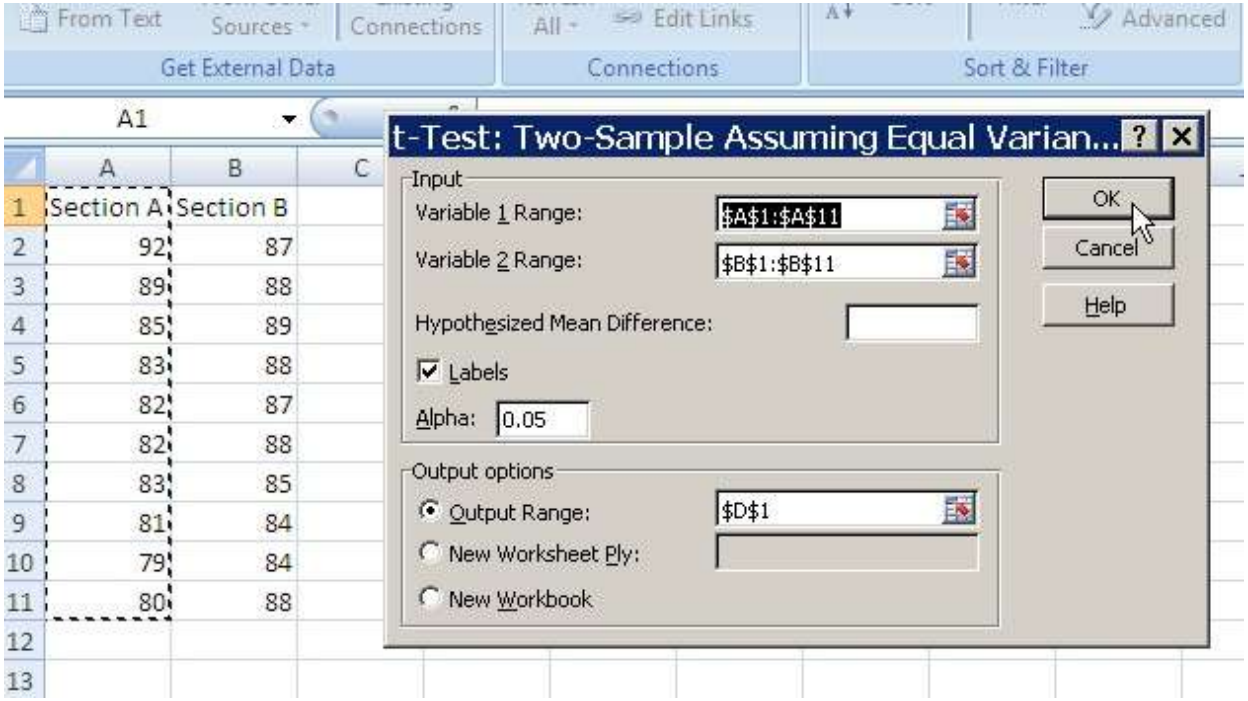
2. From the Data tab click Data Analysis.



3. From the Data Analysis dialogue box, scroll down until you see t-Test: Two-Sample Assuming Equal Variances; click it, and then click on OK.



4. Click on the box for “Variable 1 Range”, highlight the cells containing the data for the first group. After they have been selected, click on the box labeled "Variable 2 Range". Then, highlight the cells containing the data for the other sample.



If you have included highlighting the label for the groups of data, check on Labels.

In the "Output Options", click on the button beside the Output Range; then click in the box beside it. Select a cell on your worksheet into which you would like the results to be placed. Click OK.

5. The results of the T-Test will appear in the designated range. There is a significant difference between the two sample means if the absolute value of the t statistic is greater than the t critical value or P is less than the alpha level you have set.

D	E	F
t-Test: Two-Sample Assuming Equal Variances		
	Section A	*Section B*
Mean	83.6	86.8
Variance	16.48888889	3.28888889
Observations	10	10
Pooled Variance	9.888888889	
Hypothesized Mean	0	
df	18	
t Stat	-2.27541822	
P(T<=t) one-tail	0.017670139	
t Critical one-tail	1.734063592	
P(T<=t) two-tail	0.035340278	
t Critical two-tail	2.100922037	

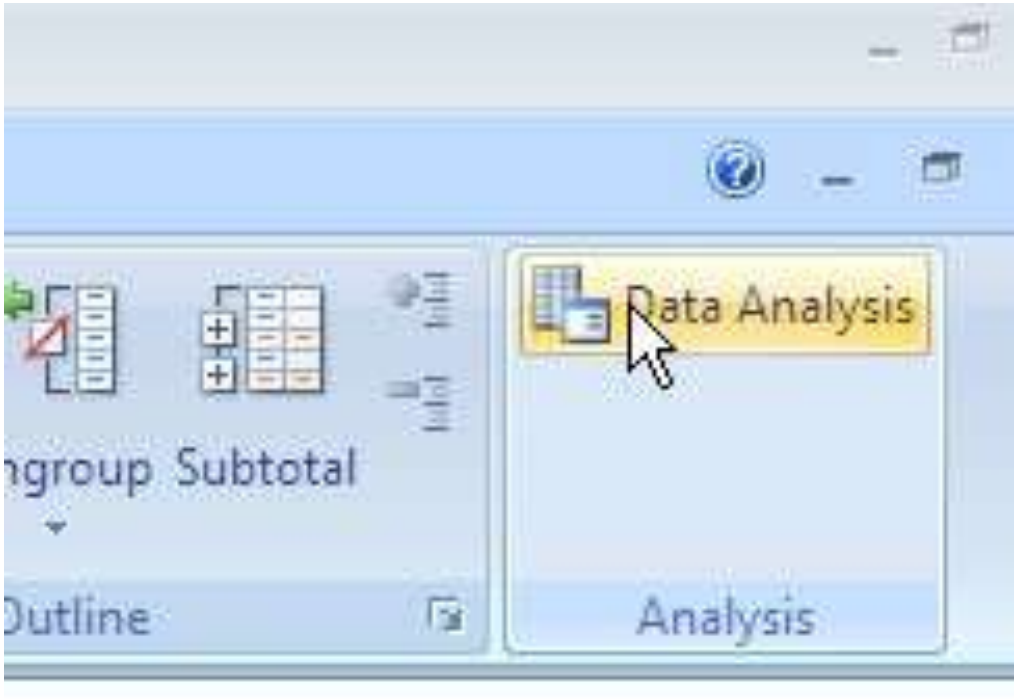
Now, in the next chapter, will try comparing 3 groups!

The ANOVA using Excel

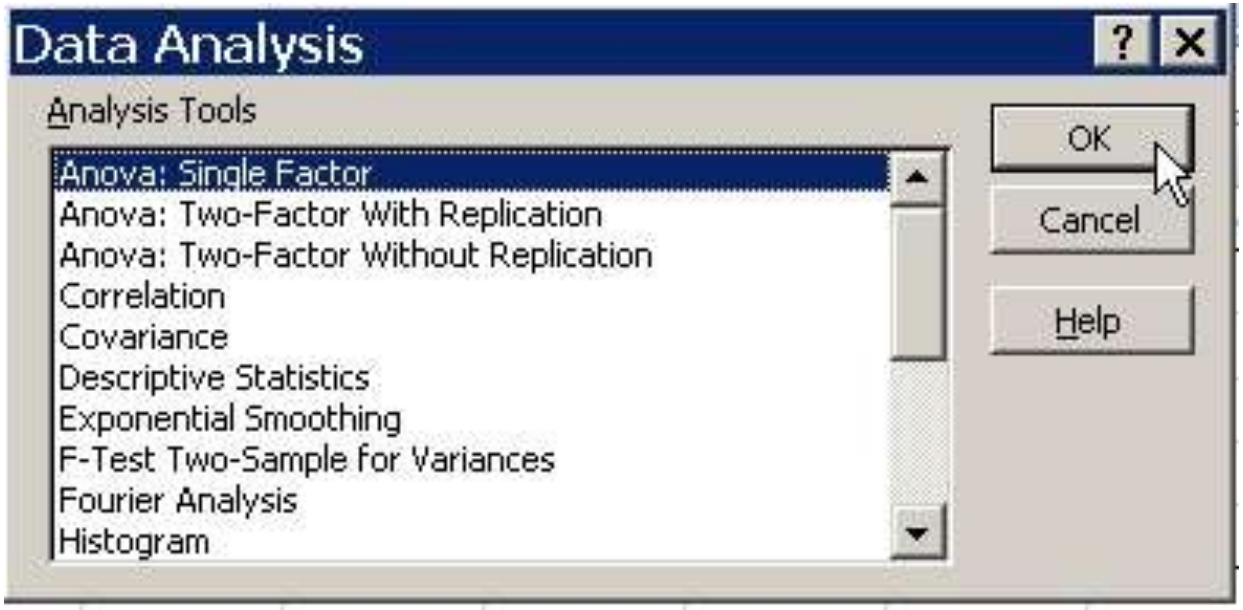
1. Encode your data in Excel.

G	H	I
Julia	**Joey**	**Mike**
90	88	84
89	86	85
87	85	87
88	87	81
86	83	80

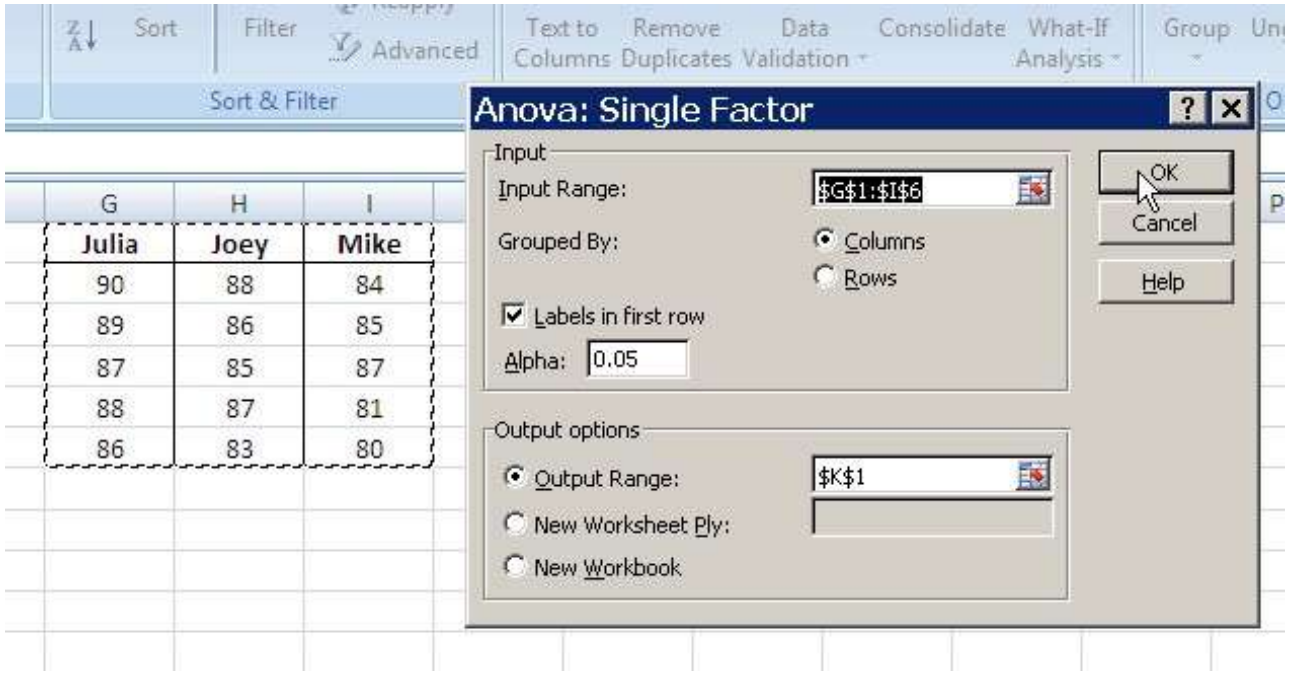
2. From the Data tab click Data Analysis.



2. Click OK to the first choice, ANOVA: Single Factor.



4. Click and drag your mouse from Julia's name to the last score in Mike's column. This will automatically complete the Input Range for you. Click the box labeled "Labels in First Row." Click Output Range. Then either type in an empty cell location, or mouse click an empty cell as shown. Click OK.



5. Interpret the probability results by evaluating the *F* ratio. If the *F* ratio is larger than the *F* critical value, *F crit*, there is a statistically significant difference. If it is smaller than the *F crit* value, the score differences are best explained by chance.

K	L	M	N	O	P	Q
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Julia	5	440	88	2.5		
Joey	5	429	85.8	3.7		
Mike	5	417	83.4	8.3		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	52.93333	2	26.46667	5.475862	0.020427	3.885294
Within Groups	58	12	4.833333			
Total	110.9333	14				

The *F* ratio 5.47 is larger than the *F crit* value 3.88. Julia is a better student. The difference between her and the other two students is statistically significant.

We have finished with comparison statistics now we shall try Excel in computing for statistics of relationships!

Linear Regression with Excel

A linear regression is a statistical tool employed in order to determine whether or not two (or more) variables are linearly related.

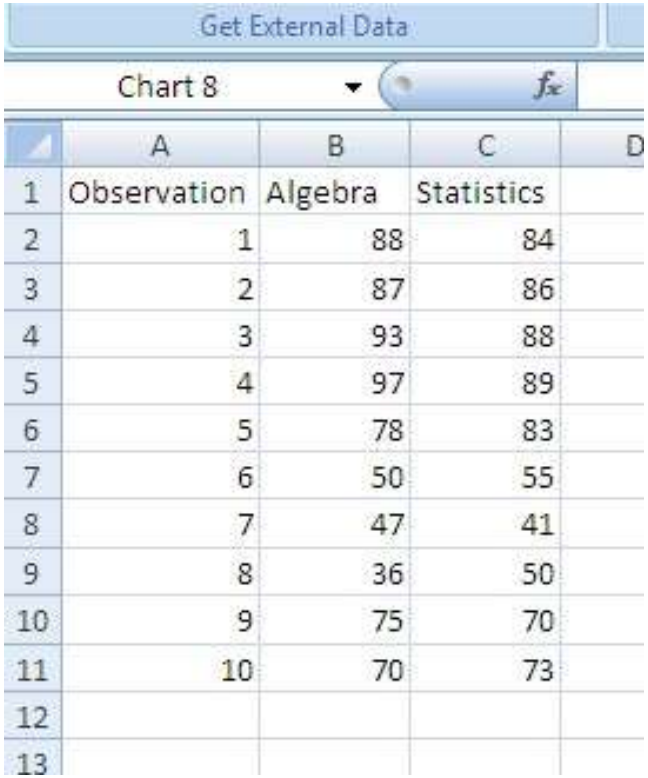
Suppose you want to determine whether a student's grade in statistics is a function of his or her grade in algebra. The general form of the relationship is:

$$Y_i = a + bX_i$$

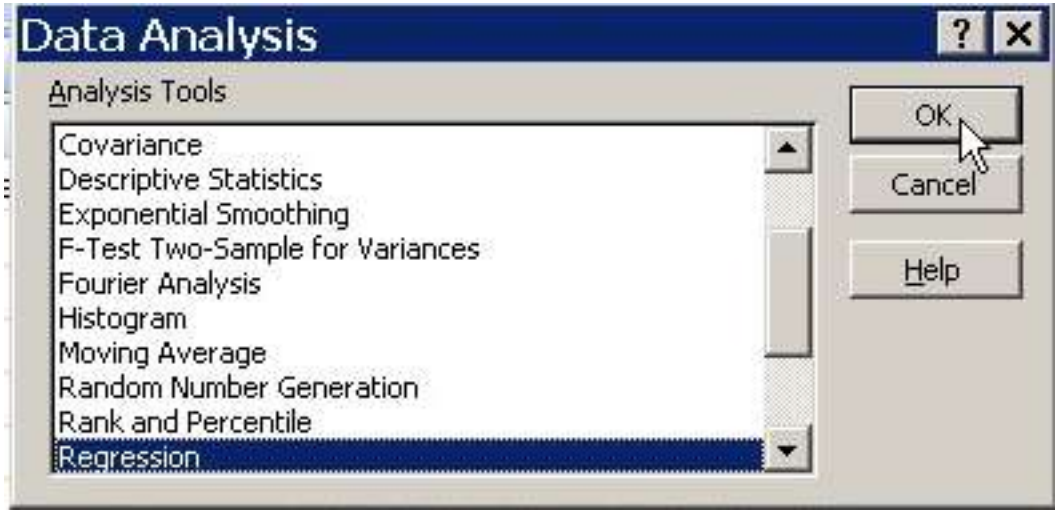
where:

- Y_i = value of Y (statistics grade) for observation i
- a = average value of Y (statistics grade) when X (algebra grade) is zero
- b = average change in Y (statistics grade) given a one unit increase in X (algebra grade), i.e. the average increase in statistics grade for each additional grade point in algebra
- X_i = value of X (algebra grade) for observation i

You start by collecting a random sample of observations, and recording them in your spreadsheet. For ease of computation, it helps to put the dependent variable (Y) in the left column and the independent variable (X) in the right column.

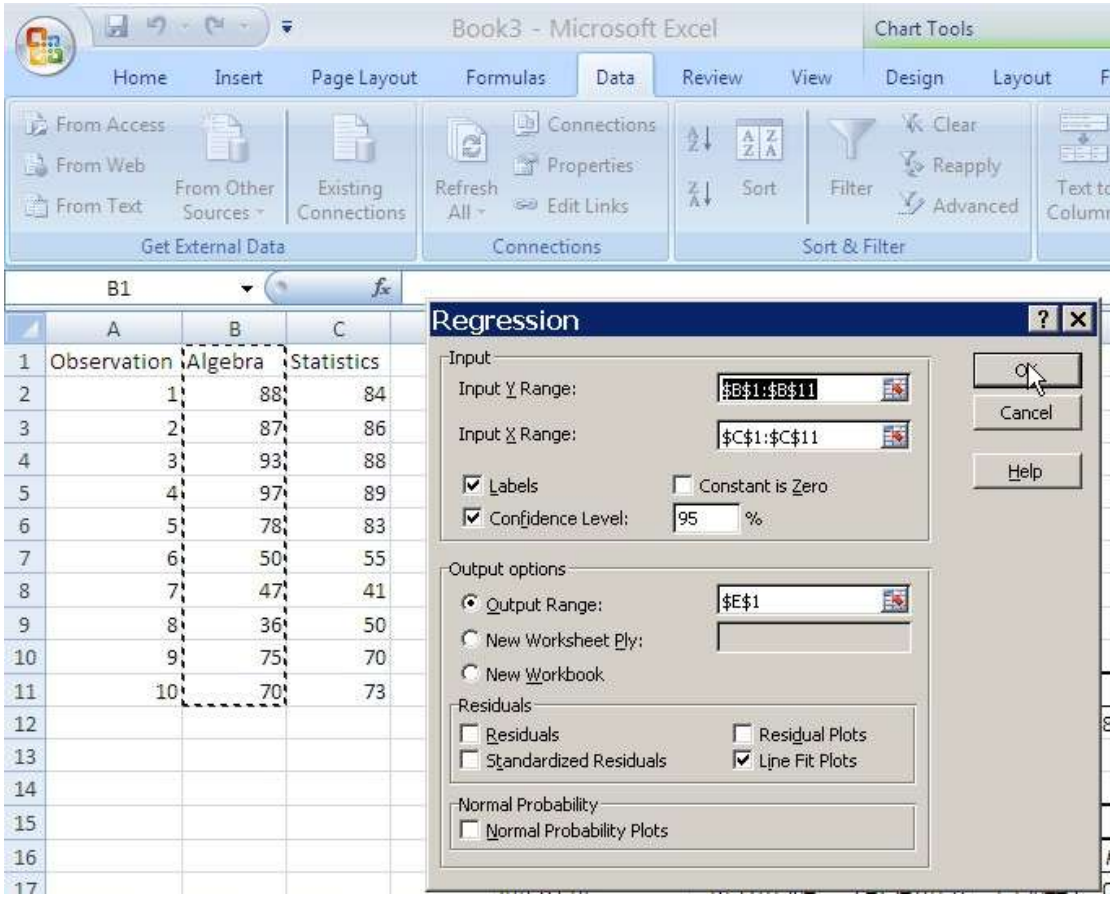


From the Data tab, Click on Data Analysis. A dialogue box for Data Analysis will appear.



Scroll down until you see the "Regression" tool, click on it, then click OK.

You will see a box like this:

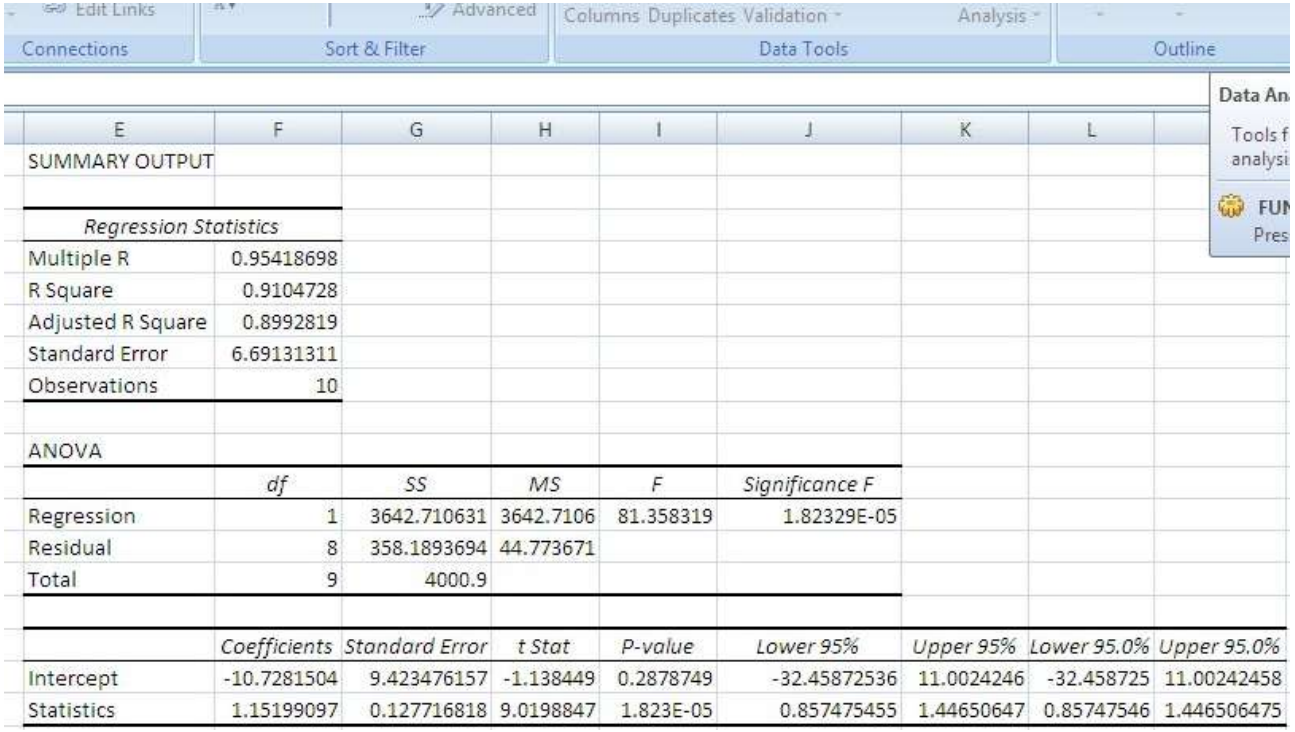


Click inside the box labeled "Input Y Range:" and then click on cell B1 and hold the left mouse button down and highlight cells B1 through B11

Next, click inside the box labeled "Input X Range:" and then click on cell C1 and hold the left mouse button down and highlight cells C1 throught C11

Since we have labels at the top of each data column (and included their cells in the ranges above) click the "Labels" checkbox and then click the "Line Fit Plots" checkbox.

Now, click OK and Excel will perform the linear regression, and the output will be shown on a range you have specified in the “Output Range”.



	E	F	G	H	I	J	K	L	
SUMMARY OUTPUT									
	Regression Statistics								
	Multiple R	0.95418698							
	R Square	0.9104728							
	Adjusted R Square	0.8992819							
	Standard Error	6.69131311							
	Observations	10							
	ANOVA								
		df	*SS*	*MS*	*F*	*Significance F*			
	Regression	1	3642.710631	3642.7106	81.358319	1.82329E-05			
	Residual	8	358.1893694	44.773671					
	Total	9	4000.9						
		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
	Intercept	-10.7281504	9.423476157	-1.138449	0.2878749	-32.45872536	11.0024246	-32.458725	11.00242458
	Statistics	1.15199097	0.127716818	9.0198847	1.823E-05	0.857475455	1.44650647	0.85747546	1.446506475

Values to note:

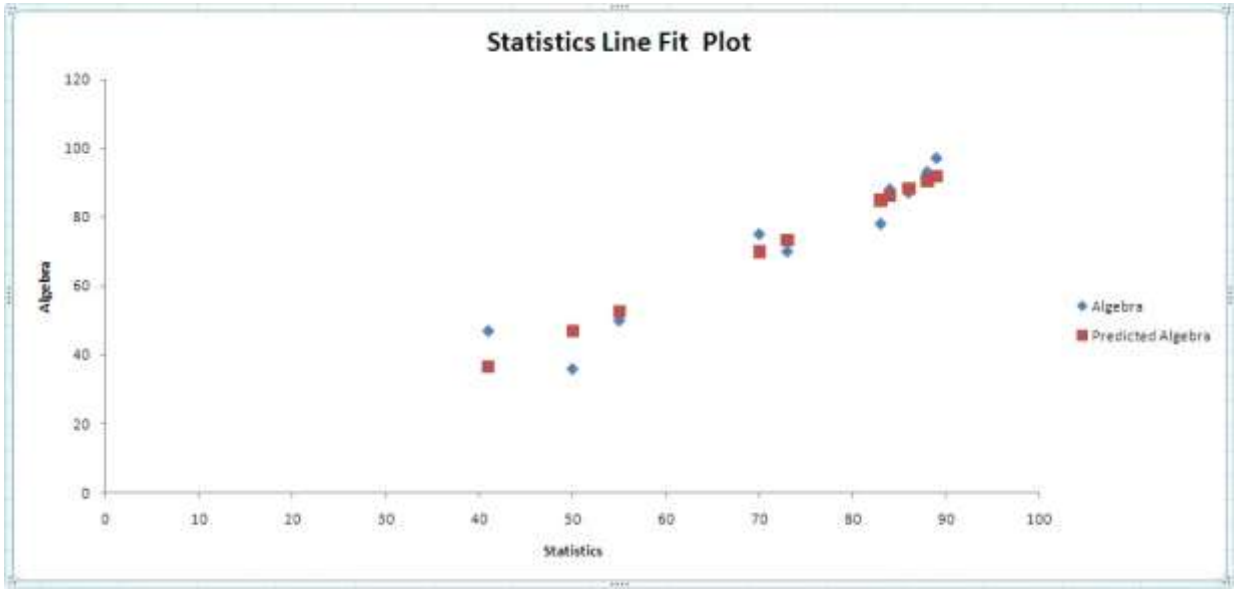
The "intercept" is the value of "a" in the equation: $Y_i = a + bX_i$

The “slope” is the value of "b" in the equation: $Y_i = a + bX_i$

So, our regression equation is: statistics grade = -10.728 + 1.512(algebra grade)
We interpret it as: the statistics grade of a student increases by 1.512 with every increase of the algebra grade by one.

We should also look at the "Adjusted R Square" value in the Regression statistic in order to determine how strong the relationship between statistics grade and algebra grade is. In this case, its value is 0.90, which indicates that about 90% of statistics grade is determined by the algebra grade (so about 10% is determined by other factors).

We can also look at the line fit plot to get a visual feel for how "linear" the relationship is:



The pink squares show the "predicted" relationship, i.e. a perfectly straight line from the equation:

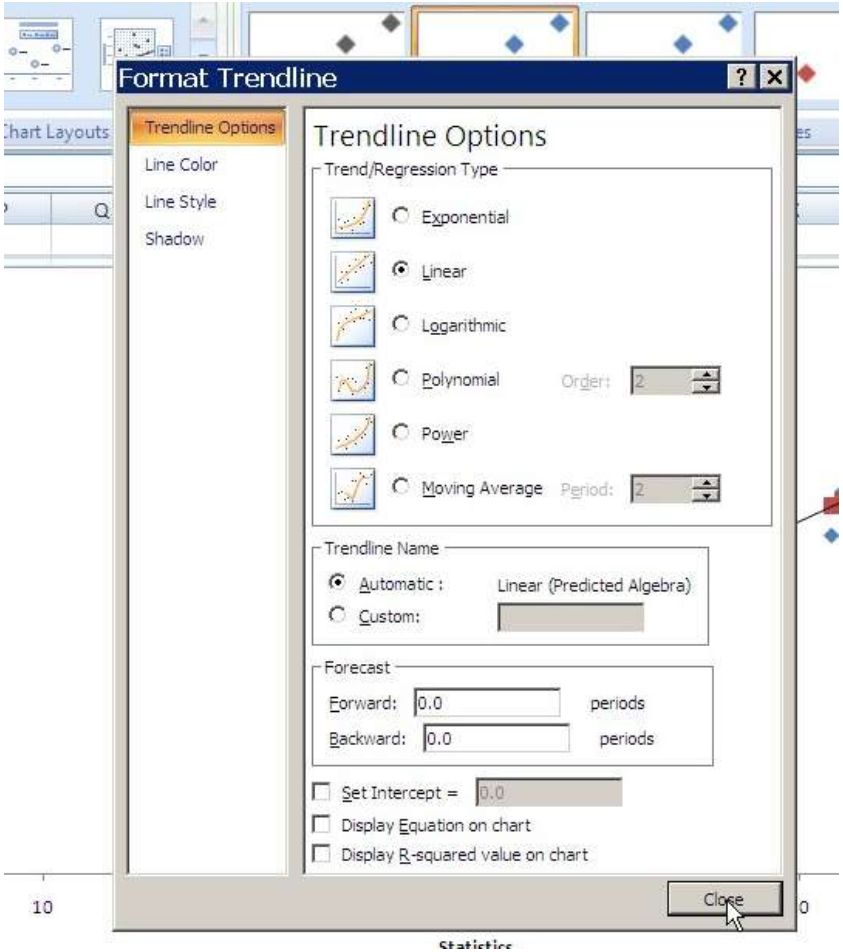
$Y_i = a + bX_i$

The dark blue diamonds show the actual relationship:

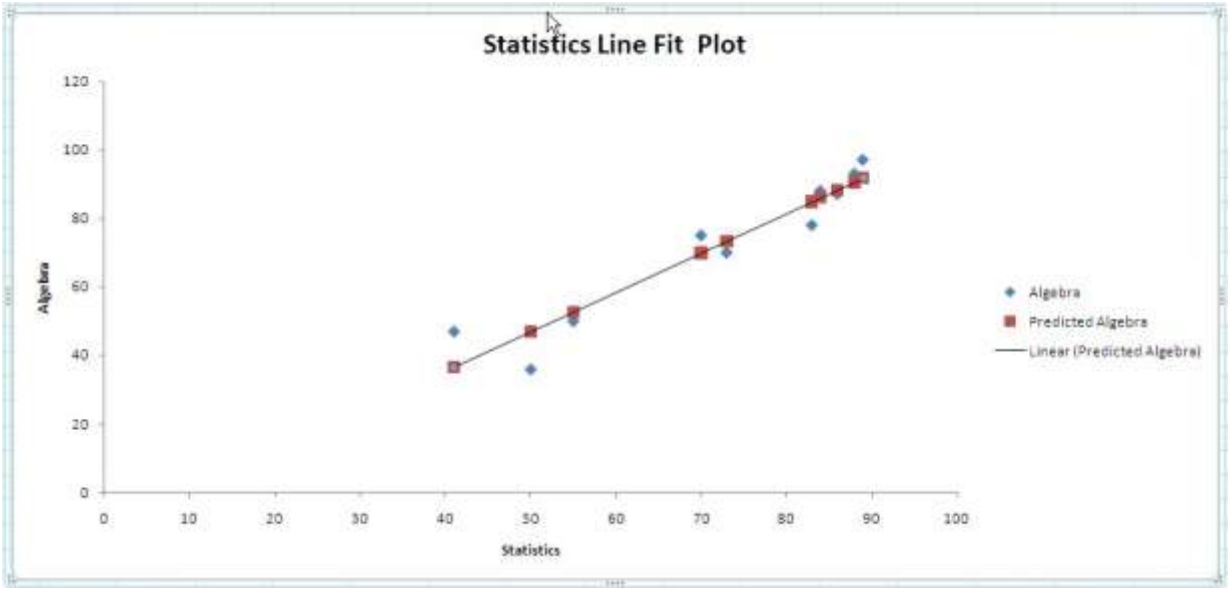
$Y_i = a + bX_i + error_i$

The error term accounts for the fact that the part of salary is due to other factors not included in our model.

We can also connect the pink squares with a straight line also called a regression line by right clicking any of the pink squares. Then click button for “Linear” then click “Close”.


Format Trendline
Trendline Options
Line Color
Line Style
Shadow
Trendline Options
Trend/Regression Type
Exponential
Linear
Logarithmic
Polynomial
Order:
2
Power
Moving Average
Period:
2
Trendline Name
Automatic :
Linear (Predicted Algebra)
Custom:
Forecast
Forward:
0.0
periods
Backward:
0.0
periods
Set Intercept =
0.0
Display Equation on chart
Display R-squared value on chart
Close
10

The graph will then look like this.



About the Author

Mike Arieh Medina is an evaluation and research consultant for development programs by profession. He was formerly a professor in environmental science and teaches research and statistics courses.

Mike is also an eLearning specialist and advocates the use of digital and web technology in education.

He is based in Mati, Davao Oriental but travels around Mindanao, Philippines for his consultancy work.

He is currently writing his dissertation for his PhD in Development Research Administration at the University of Southeastern Philippines in Davao City.

For speaking engagements or consultancy services, you can contact him through email at earth_initiative@yahoo.com

Check out his blog, **Grad School Jungle** a site about life and survival in graduate school at http://gradschooljungle.blogspot.com

Grad School Jungle

Grad School Jungle is an academic blog started in March 2008 by **Mike Arieh Medina**. It was originally meant to cover environmental and development topics for his PhD academic papers. Later on, in September 2010, Mike felt the need of graduate students to have certain resources that would provide them with ways on how to focus on their studies by acquiring information that would motivate them in finishing their graduate degree. Thus, Grad School Jungle now included posts on news, articles, tips and other motivational materials for grad students in general. Furthermore, college students and teachers were also able to benefit from the information posted in the blog. This eBook, **Statistically Excellent: Inferential Statistics using MS Excel 2007** is a compilation of a series of blog posts which was published in Grad School Jungle in January 2011.