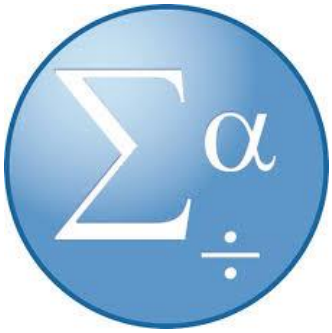
Introductory and Advance Research Analysis

Applied Statistics for Research

School of Graduate Studies and Research Training



Interdisciplinary Research Center
School of Graduate Studies and Research
UNIVERSITY OF MARYLAND EASTERN SHORE

1. Introductory Statistical Analysis in MS Excel

This seminar focuses on how to use excel tools in conducting basic and semi-advance statistical analysis for decision-making. Apart from various steps in how to perform various analysis in excel, the seminar will also highlight why and when to use specific analysis and how to interpret results of the various statistical tests.

Topics to be covered:

1. **Descriptive Statistics in Excel**
2. **Correlation Analysis in Excel**
3. **Hypothesis Testing (F-test Analysis in Excel)**
4. **Hypothesis Testing (T-test Analysis in Excel)**
5. **ANOVA – Analysis of Variance in Excel**
6. **Regression Analysis in Excel**

2. Advance Statistical Analysis using STATA software

This seminar focuses on how to use stata software in conducting basic, semi-advance and advanced statistical analysis for decision-making in Research.

Topics to be covered:

1. **Descriptive Statistics in STATA and how to present results in publishable format**
2. **Correlation Analysis in STATA and how to present results in publishable format**
3. **Hypothesis Testing in STATA**
4. **ANOVA – Analysis of Variance in STATA**
5. **Regression Analysis in STATA and how to present results in publishable format**

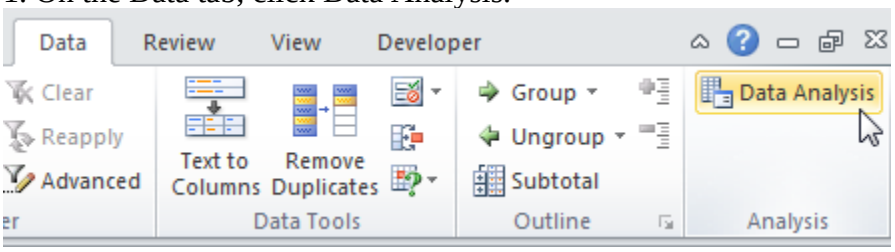
1. Descriptive Statistics

Generating descriptive statistics. For example, you may have the scores of 14 participants for a test.

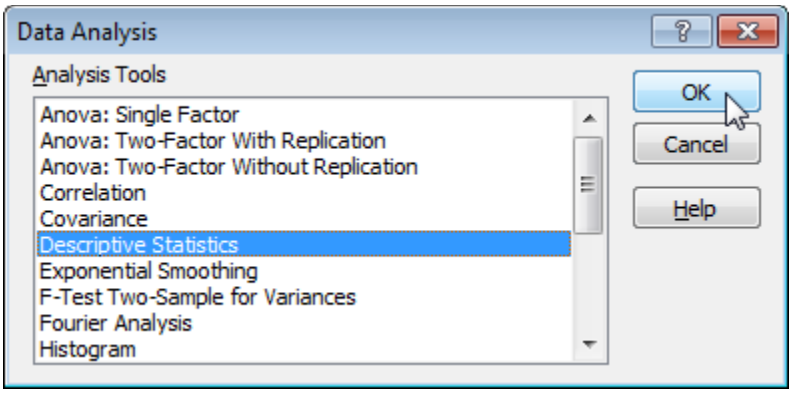
M26

	A	B
1	Scores	
2	82	
3	93	
4	91	
5	69	
6	96	
7	61	
8	88	
9	58	
10	59	
11	100	
12	93	
13	71	
14	78	
15	98	
16		
17		

1. On the Data tab, click Data Analysis.

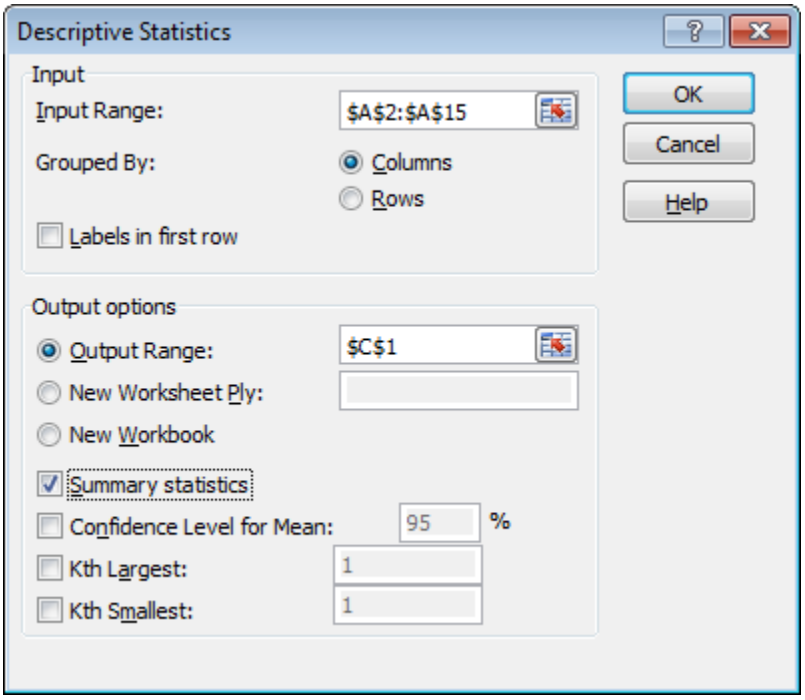


2. Select Descriptive Statistics and click OK.



3. Select the range A2:A15 as the Input Range.

4. Make sure Summary statistics is checked.



6. Click OK. Result:

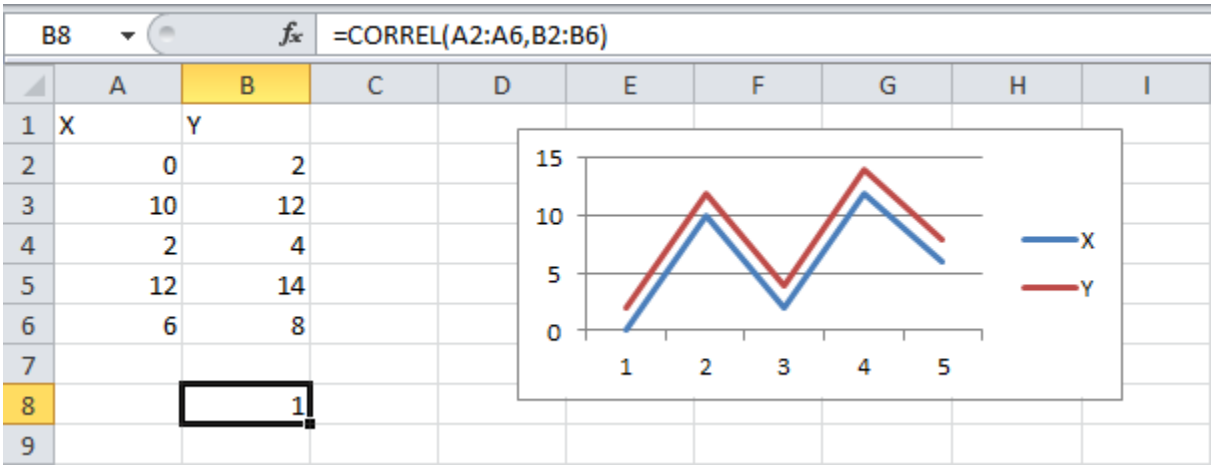
L28

	A	B	C	D	E
1	Scores		*Column1*		
2	82				
3	93		Mean	81.21428571	
4	91		Standard Error	4.045318243	
5	69		Median	85	
6	96		Mode	93	
7	61		Standard Deviation	15.13619489	
8	88		Sample Variance	229.1043956	
9	58		Kurtosis	-1.426053506	
10	59		Skewness	-0.402108004	
11	100		Range	42	
12	93		Minimum	58	
13	71		Maximum	100	
14	78		Sum	1137	
15	98		Count	14	
16					
17					

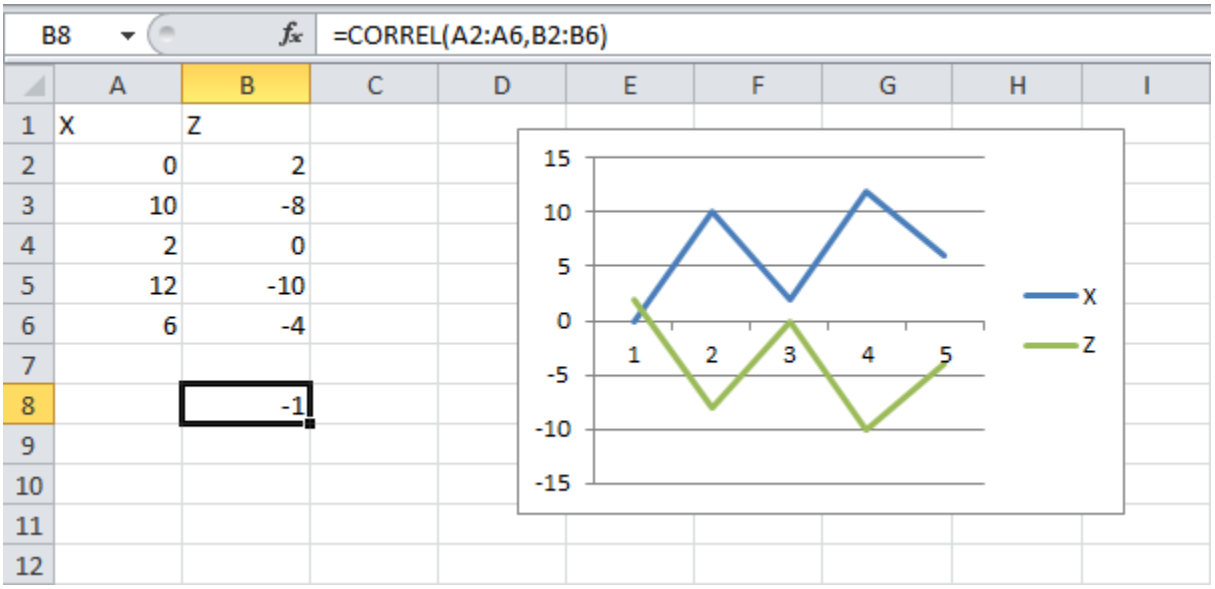
2. Correlation Analysis

The correlation coefficient (a value between -1 and +1) tells you how strongly two variables are related to each other. We can use the CORREL function or the Analysis Toolpak add-in in Excel to find the correlation coefficient between two variables.

- A correlation coefficient of +1 indicates a perfect positive correlation. As variable X increases, variable Y increases. As variable X decreases, variable Y decreases.



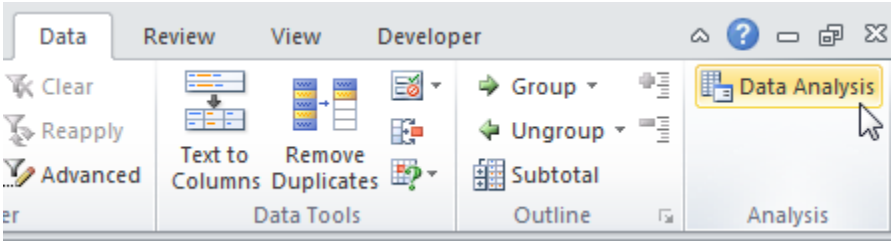
- A correlation coefficient of -1 indicates a perfect negative correlation. As variable X increases, variable Z decreases. As variable X decreases, variable Z increases.



- A correlation coefficient near 0 indicates no correlation.

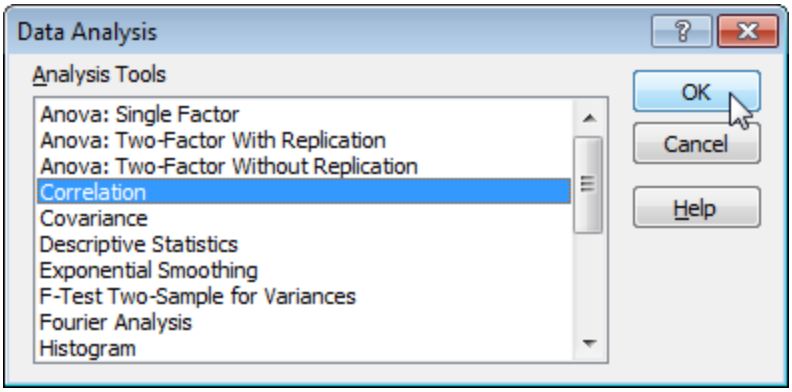
To use the Analysis Toolpak add-in in Excel to quickly generate correlation coefficients between multiple variables, execute the following steps.

1. On the Data tab, click Data Analysis.

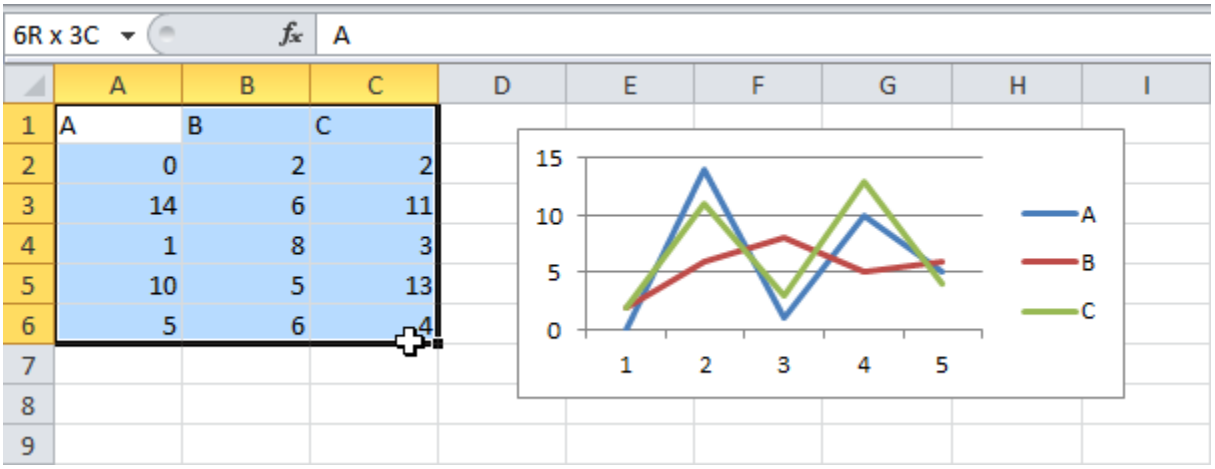


Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

2. Select Correlation and click OK.



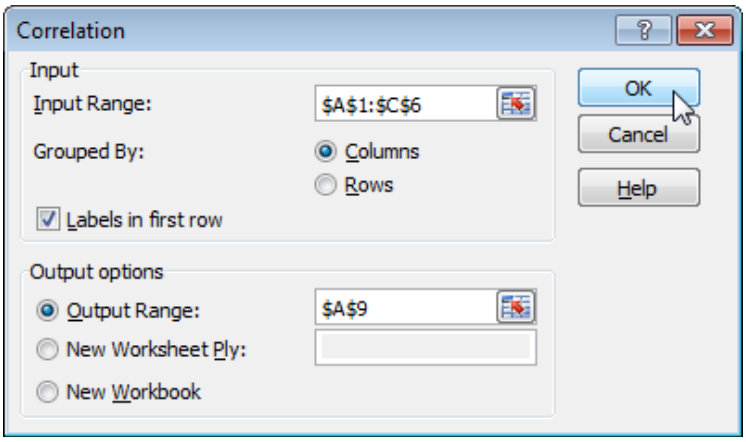
3. For example, select the range A1:C6 as the Input Range.



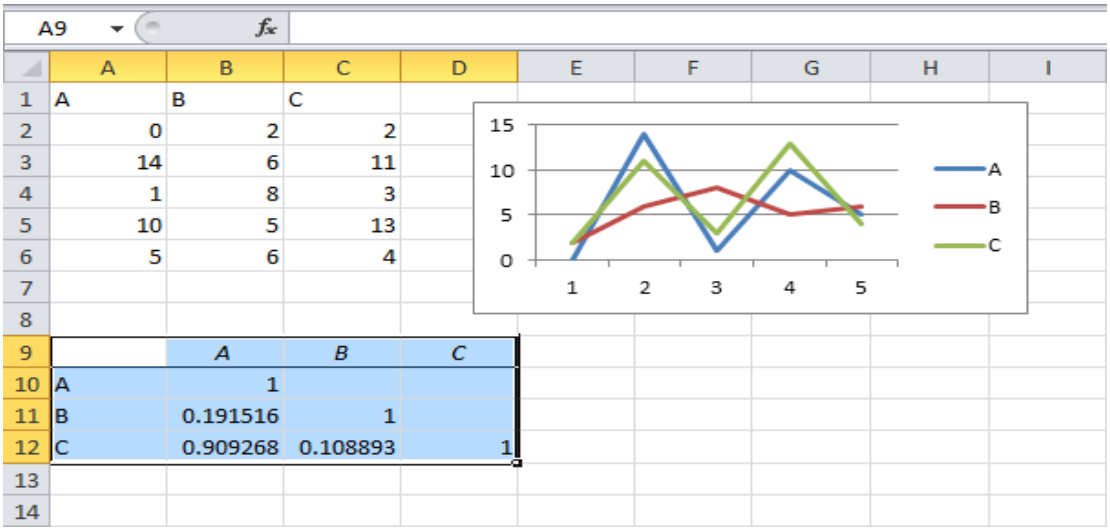
4. Check Labels in first row.

5. Select cell A9 as the Output Range.

6. Click OK.



Result.



Conclusion: variables A and C are positively correlated (0.91). Variables A and B are not correlated (0.19). Variables B and C are also not correlated (0.11). You can verify these conclusions by looking at the graph.

3. ANOVA

This example teaches you how to perform a single factor ANOVA (analysis of variance) in Excel. A single factor or one-way ANOVA is used to test the null hypothesis that the means of several populations are all equal.

Below you can find the salaries of people who have a degree in economics, medicine or history.

H_0: $\mu_1 = \mu_2 = \mu_3$

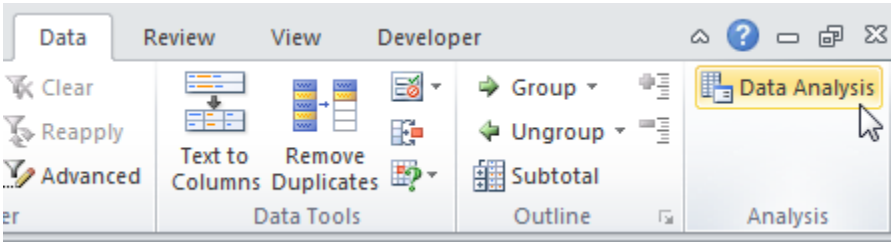
H_1: at least one of the means is different.

H18

	A	B	C	D
1	economics	medicine	history	
2	42	69	35	
3	53	54	40	
4	49	58	53	
5	53	64	42	
6	43	64	50	
7	44	55	39	
8	45	56	55	
9	52		39	
10	54		40	
11				
12				

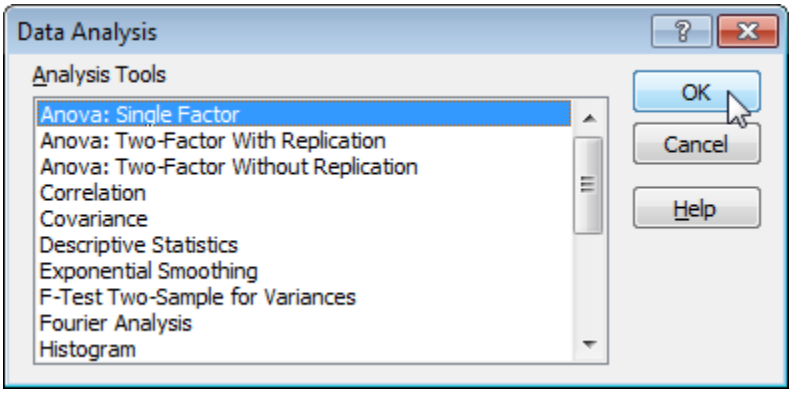
To perform a single factor ANOVA, execute the following steps.

1. On the Data tab, click Data Analysis.



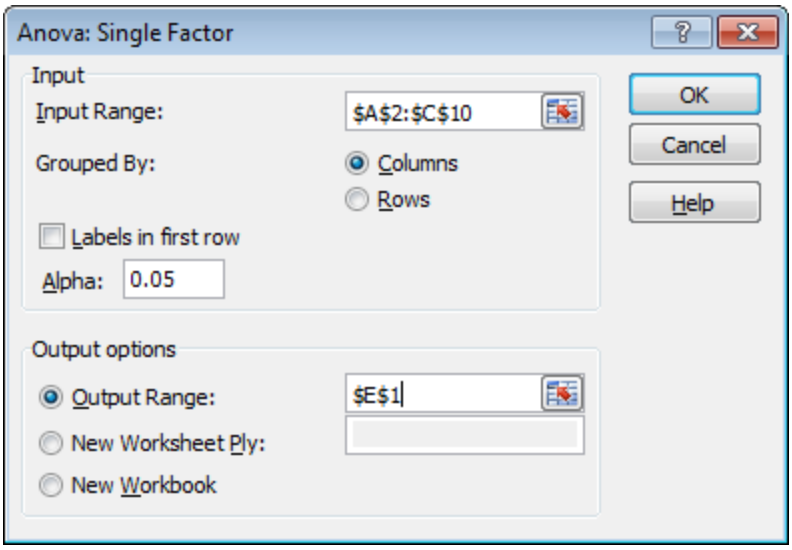
Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

2. Select Anova: Single Factor and click OK.



3. Click in the Input Range box and select the range A2:C10.

4. Click in the Output Range box and select cell E1.



5. Click OK.

Result:

E	F	G	H	I	J	K
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	9	435	48.33333	23.5		
Column 2	7	420	60	32.33333		
Column 3	9	393	43.66667	50.5		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1085.84	2	542.92	15.19623	7.16E-05	3.443357
Within Groups	786	22	35.72727			
Total	1871.84	24				

Conclusion: if F > F crit, we reject the null hypothesis. This is the case, 15.196 > 3.443. Therefore, we reject the null hypothesis. The means of the three populations are not all equal. At least one of the means is different. However, the ANOVA does not tell you where the difference lies. You need a t-Test to test each pair of means.

4. F-Test Analysis

This example teaches you how to perform an F-Test in Excel. The F-Test is used to test the null hypothesis that the variances of two populations are equal.
Below you can find the study hours of 6 female students and 5 male students.

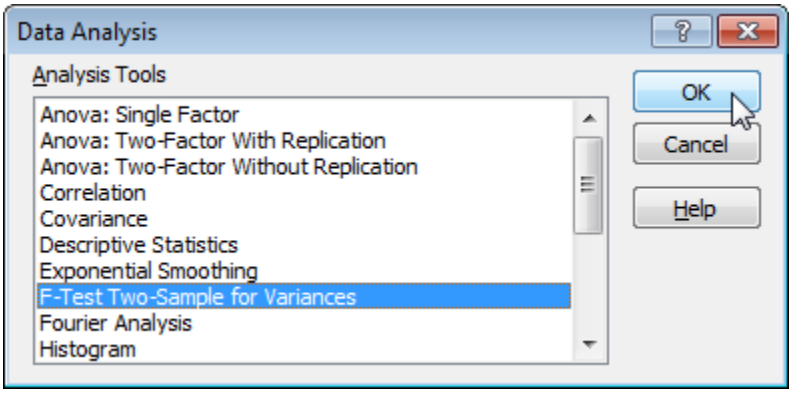
H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

H15 f_x

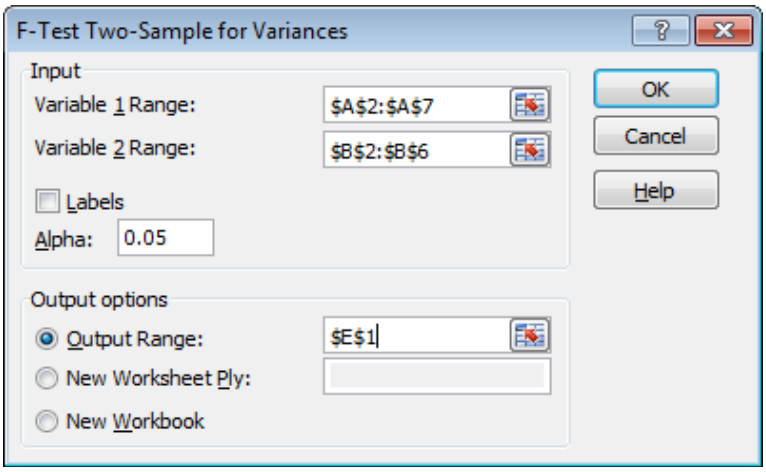
	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			
9			

To perform an F-Test, execute the following steps.

1. On the Data tab, click Data Analysis.



3. Click in the Variable 1 Range box and select the range A2:A7.
4. Click in the Variable 2 Range box and select the range B2:B6.
5. Click in the Output Range box and select cell E1.



6. Click OK.

E	F	G
F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	33	24.8
Variance	160	21.7
Observations	6	5
df	5	4
F	7.373271889	
P(F<=f) one-tail	0.037888376	
F Critical one-tail	6.256056502	

Conclusion: if F > F Critical one-tail, we reject the null hypothesis. This is the case, 7.373 > 6.256. Therefore, we reject the null hypothesis. The variances of the two populations are unequal.

5. T-Test Analysis

This example teaches you how to perform a t-Test in Excel. The t-Test is used to test the null hypothesis that the means of two populations are equal.
Below you can find the study hours of 6 female students and 5 male students.

H_0: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

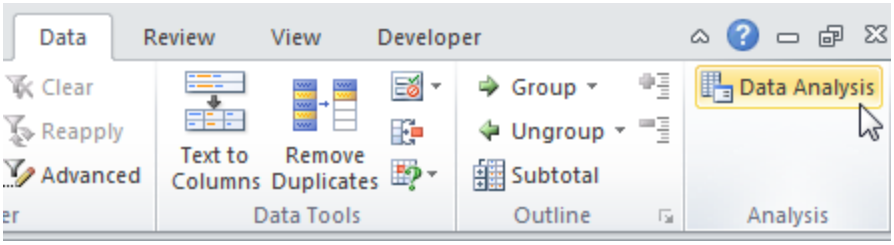
H15

	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			
9			

To perform a t-Test, execute the following steps.

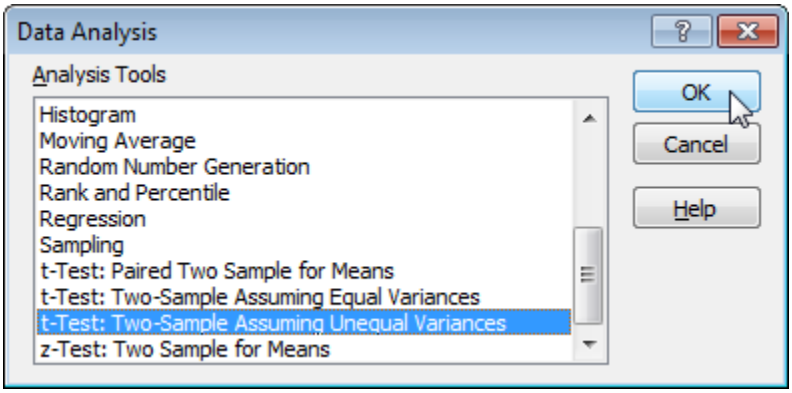
1. First, perform an F-Test to determine if the variances of the two populations are equal. This is not the case.

2. On the Data tab, click Data Analysis.

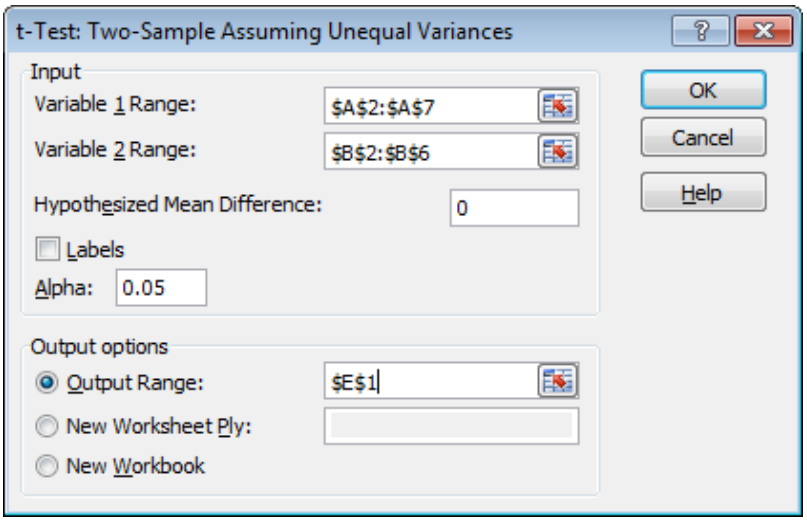


Note: can't find the Data Analysis button? Click here to load the Analysis ToolPak add-in.

3. Select t-Test: Two-Sample Assuming Unequal Variances and click OK.



4. Click in the Variable 1 Range box and select the range A2:A7.
5. Click in the Variable 2 Range box and select the range B2:B6.
6. Click in the Hypothesized Mean Difference box and type 0 (H_0: $\mu_1 - \mu_2 = 0$).
7. Click in the Output Range box and select cell E1.



8. Click OK. Result:

E	F	G
t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	33	24.8
Variance	160	21.7
Observations	6	5
Hypothesized Mean Difference	0	
df	7	
t Stat	1.47260514	
P(T<=t) one-tail	0.092170202	
t Critical one-tail	1.894578605	
P(T<=t) two-tail	0.184340405	
t Critical two-tail	2.364624252	

Conclusion: We do a two-tail test (inequality). lf t Stat < -t Critical two-tail or t Stat > t Critical two-tail, we reject the null hypothesis. This is not the case, -2.365 < 1.473 < 2.365. Therefore, we do not reject the null hypothesis. The observed difference between the sample means (33 - 24.8) is not convincing enough to say that the average number of study hours between female and male students differ significantly.

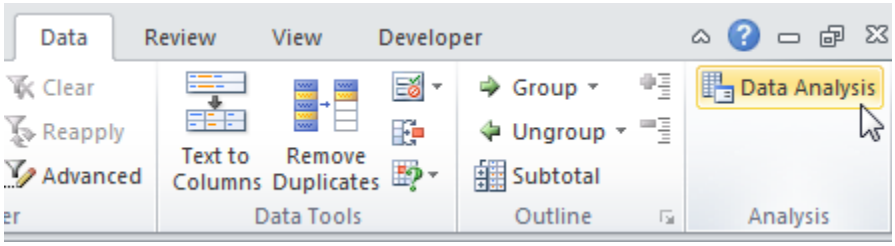
6. Multiple Regression Analysis

This example teaches you how to perform a regression analysis in Excel and how to interpret the Summary Output.

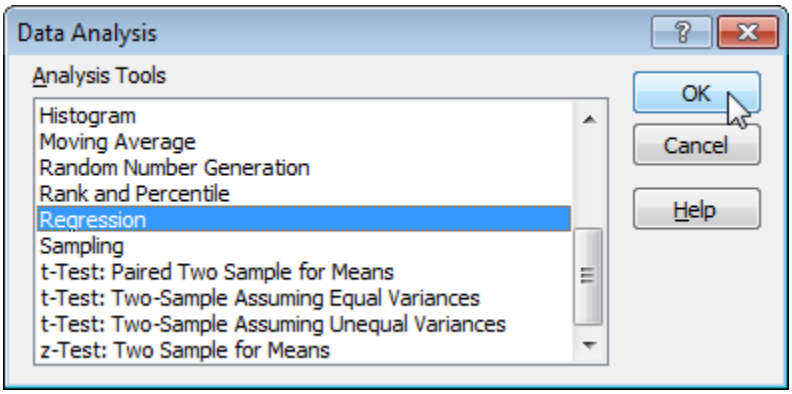
Below you can find our data. The big question is: is there a relation between Quantity Sold (Output) and Price and Advertising (Input). In other words: can we predict Quantity Sold if we know Price and Advertising?

	A	B	C	D
1	Quantity Sold	Price	Advertising	
2	8500	$2	$2,800	
3	4700	$5	$200	
4	5800	$3	$400	
5	7400	$2	$500	
6	6200	$5	$3,200	
7	7300	$3	$1,800	
8	5600	$4	$900	
9				
10				

1. On the Data tab, click Data Analysis.

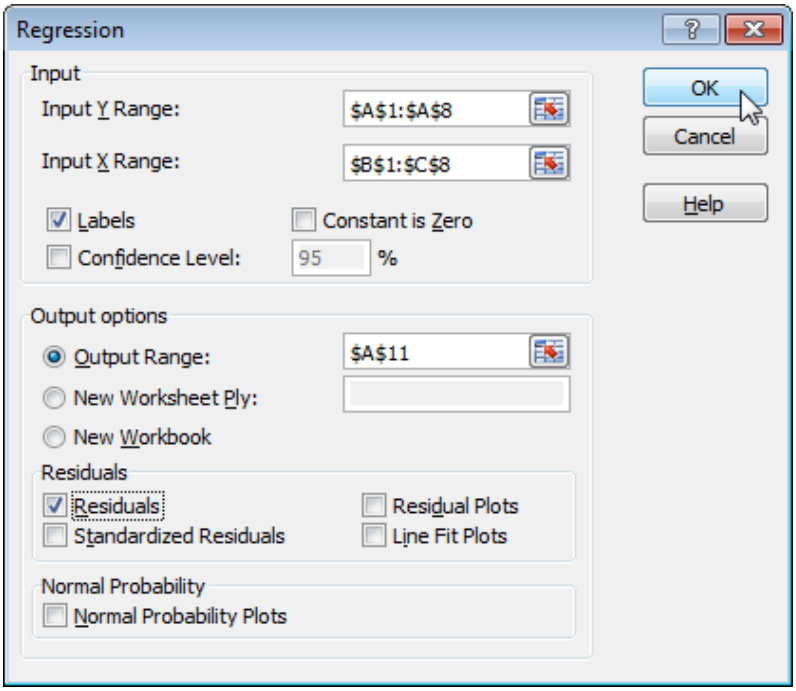


2. Select Regression and click OK.



3. Select the Y Range (A1:A8). This is the predictor variable (also called dependent variable).
4. Select the X Range(B1:C8). These are the explanatory variables (also called independent variables). These columns must be adjacent to each other.
5. Check Labels.
6. Select an Output Range.
7. Check Residuals.

8. Click OK.



Excel produces the following Summary Output (rounded to 3 decimal places).

R Square

R Square equals 0.962, which is a very good fit. 96% of the variation in Quantity Sold is explained by the independent variables Price and Advertising. The closer to 1, the better the regression line (read on) fits the data.

11	SUMMARY OUTPUT	
12		
13	*Regression Statistics*	
14	Multiple R	0.981
15	R Square	0.962
16	Adjusted R Square	0.943
17	Standard Error	310.524
18	Observations	7
19		

Significance F and P-values

To check if your results are reliable (statistically significant), look at Significance F (0.001). If this value is less than 0.05, you're OK. If Significance F is greater than 0.05, it's probably better to stop using this set of independent variables. Most or all P-values should be below below 0.05. In our example this is the case. (0.000, 0.001 and 0.005).

20	ANOVA						
21		*df*	*SS*	*MS*	*F*	*Significance F*	
22	Regression	2	9694299.568	4847149.784	50.269	0.001	
23	Residual	4	385700.432	96425.108			
24	Total	6	10080000.000				
25							
26		*Coefficients*	*Std Error*	*t Stat*	*P-values*	*Lower 95%*	*Upper 95%*
27	Intercept	8536.214	386.912	22.062	0.000	7461.975	9610.453
28	Price	-835.722	99.653	-8.386	0.001	-1112.404	-559.041
29	Advertising	0.592	0.104	5.676	0.005	0.303	0.882
30							

Coefficients

The regression line is: y = Quantity Sold = 8536.214 -835.722 * Price + 0.592 * Advertising. In other words, for each unit increase in price, Quantity Sold decreases with 835.722 units. For each unit increase in Advertising, Quantity Sold increases with 0.592 units. This is valuable information.

You can also use these coefficients to do a forecast. For example, if price equals $4 and Advertising equals $3000, you might be able to achieve a Quantity Sold of 8536.214 -835.722 * 4 + 0.592 * 3000 = 6970.

Residuals

The residuals show you how far away the actual data points are from the predicted data points (using the equation). For example, the first data point equals 8500. Using the equation, the predicted data point equals 8536.214 -835.722 * 2 + 0.592 * 2800 = 8523.009, giving a residual of 8500 - 8523.009 = -23.009.

33	RESIDUAL OUTPUT		
34			
35	*Observation*	*Predicted Quantity Sold*	*Residuals*
36	1	8523.009	-23.009
37	2	4476.048	223.952
38	3	6265.938	-465.938
39	4	7160.883	239.117
40	5	6252.733	-52.733
41	6	7095.058	204.942
42	7	5726.330	-126.330
43			