Using EXCEL for Statistical Analysis

Brian W. Sloboda

University of Phoenix

bsloboda@email.phoenix.edu

June 25, 2020

Purpose of this Session

1. First Section
   - The purpose of this presentation is to learn how to use EXCEL to conduct statistical analysis.

Descriptive Statistics

- The first part of this session is to review the procedures to calculate the descriptive statistics using EXCEL.
- (This step only needs to be done once.) Go to TOOLS-ADD INS and select the Analysis Toolpaks and OK. This will add the analysis tools to your EXCEL.
- If for some reason, when you use Data Analysis in the future and it is not there, just download it again.

Descriptive Statistics

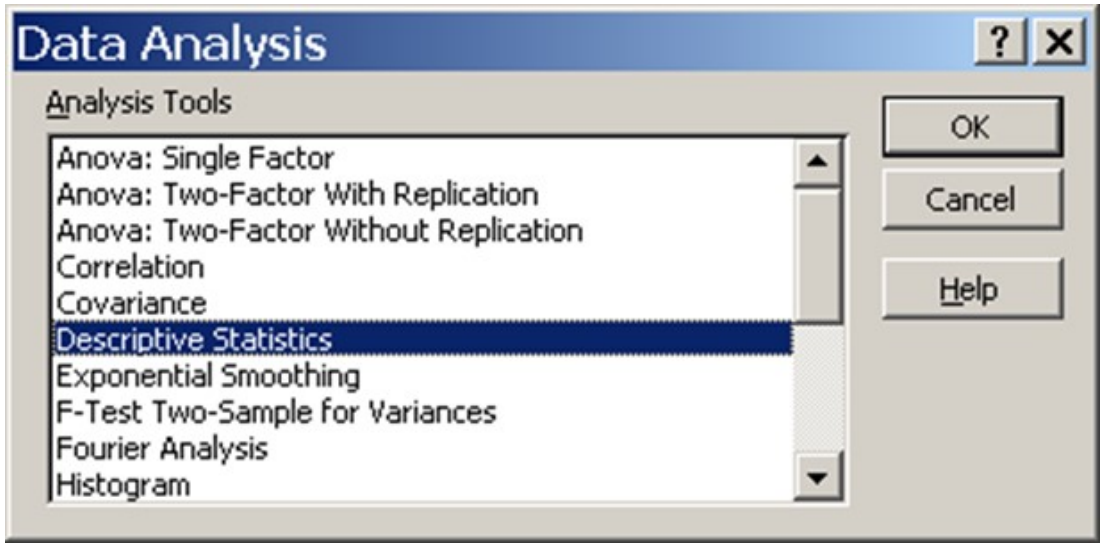
Here is sample data to illustrate descriptive statistics

Observation Number	Variable 1	Variable 2
1	106	121
2	98	104
3	111	104
4	108	115
5	97	118
6	92	96
7	89	85
8	105	106
9	101	111
10	87	100
11	101	111
12	90	105

Table: Sample Data

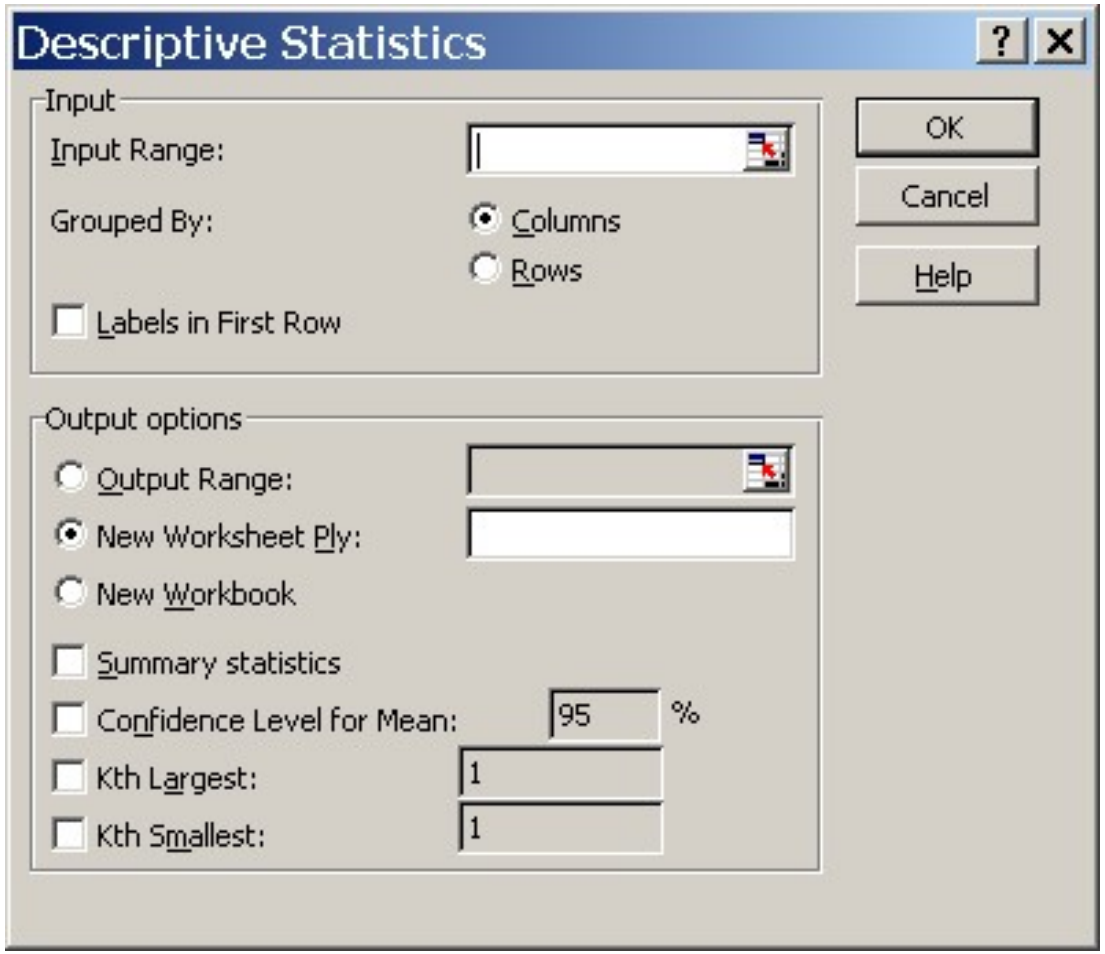
Descriptive Statistics

To run the descriptive statistics on the data, go to TOOLS-DATA ANALYSIS (it should be the last option in the TOOLS menu and will enable once you have loaded it after Step 1). Select DESCRIPTIVE STATISTICS and OK.



Descriptive Statistics

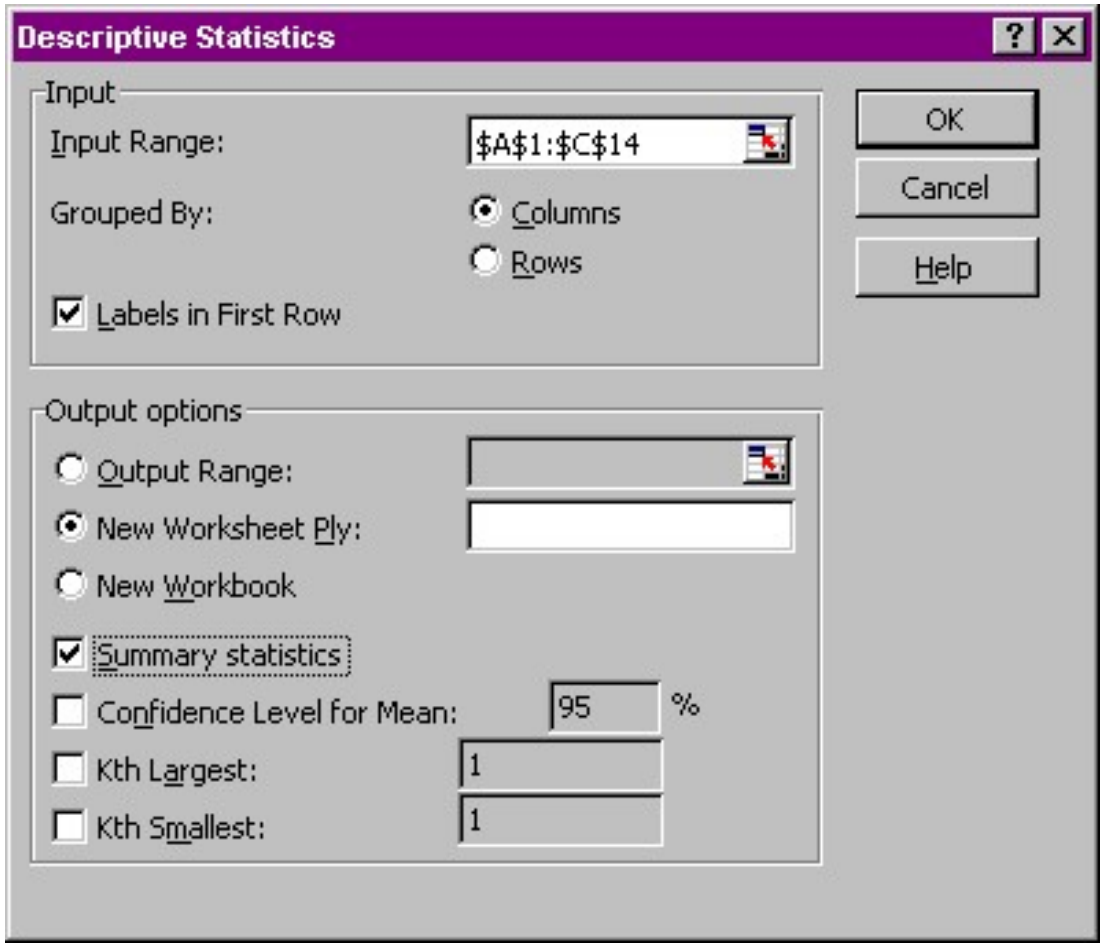
You should now have a table that looks like this.



The INPUT RANGE is the data that will analyzed. Either select the red box and highlight the range, or enter the cell ranges of the data. The cells to be analyzed are A1 to C14. When the data are highlighted, hit ENTER

Descriptive Statistics

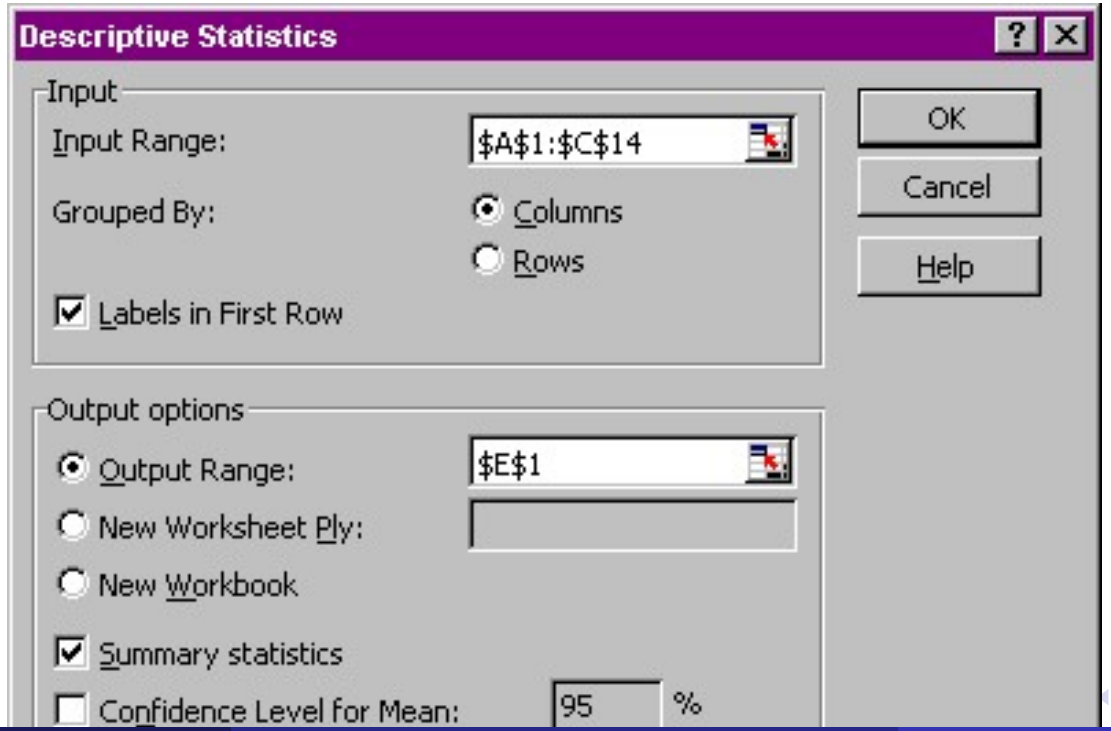
Next, select the Summary Statistics box (which will do a summary statistics table).



Descriptive Statistics

Finally, decide where the output should be produced. If you would like the summary table on the same sheet as the data, select the first option (Output Range). If you would like the table in the same Excel workbook, but on a different sheet, select the second option (New Worksheet Ply). Finally, the output can be generated to a whole new Excel file (the third option).

Let's select the second option.



Descriptive Statistics

Select OK and a summary table will be displayed that should look like this:

Var 1		*Var 2*	
Mean	98.75	Mean	107.1667
Standard Error	2.296654	Standard Error	2.383127
Median	99.5	Median	105.5
Mode	101	Mode	104
Standard Deviation	7.955844	Standard Deviation	8.255393
Sample Variance	63.29545	Sample Variance	68.15152
Kurtosis	-1.27125	Kurtosis	-0.81325
Skewness	-0.05516	Skewness	0.169683
Range	24	Range	26
Minimum	87	Minimum	95
Maximum	111	Maximum	121
Sum	1185	Sum	1286
Count	12	Count	12

Descriptive Statistics

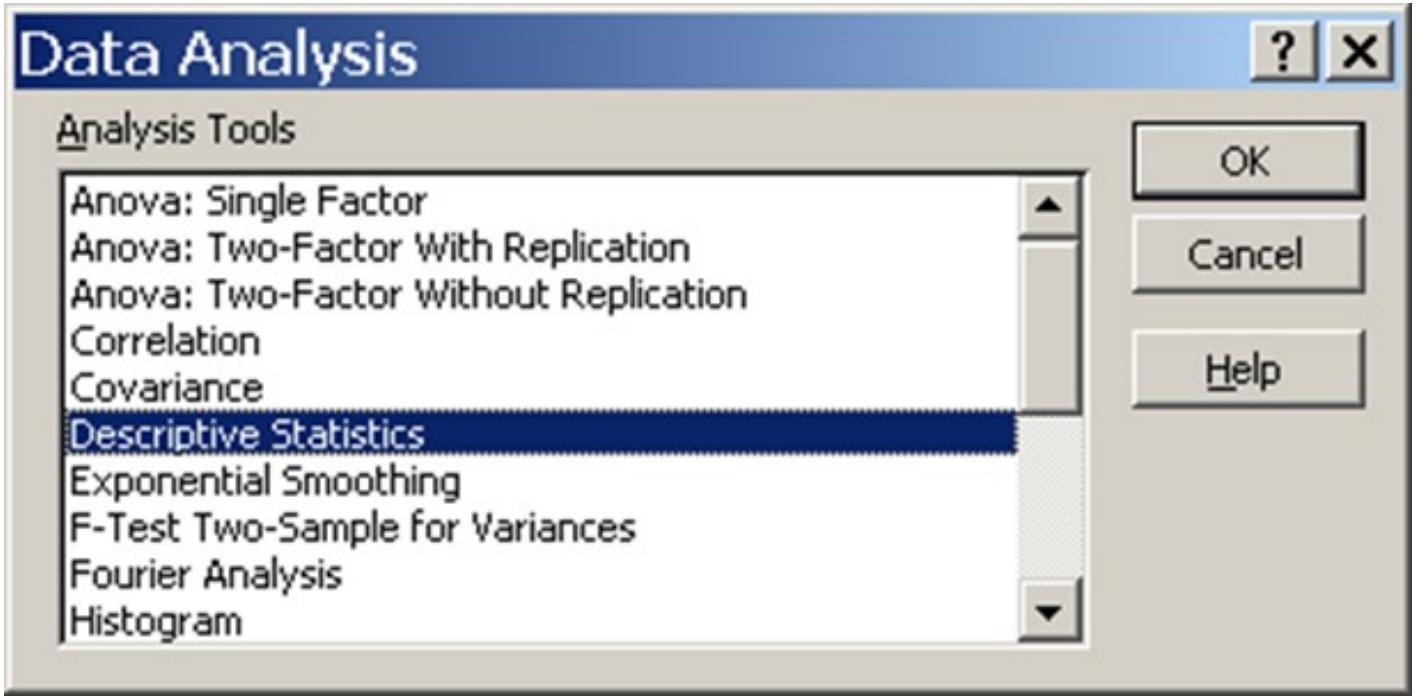
Here is an explanation of what each of the descriptive statistics is describing:

Descriptive statistics *describe* the data; it does not establish relationships or provide cause-and-effect results. Here are what each of the items from the EXCEL output means:

Mean – the average of the data (sum of data / number of data points)
Standard Error – standard deviation of a sampling distribution
Median – middle value when the values are rank-ordered
Mode – most frequent value
Standard deviation – square root of the variance
Variance – describes the distribution of the data around the mean. Is the average of the squared deviations of the data values from the mean.
Kurtosis – describes the peakedness of the normal curve
Skewness – describes the symmetry of the normal curve
Range – distance between the highest and lowest value
Minimum – smallest value
Maximum – largest value

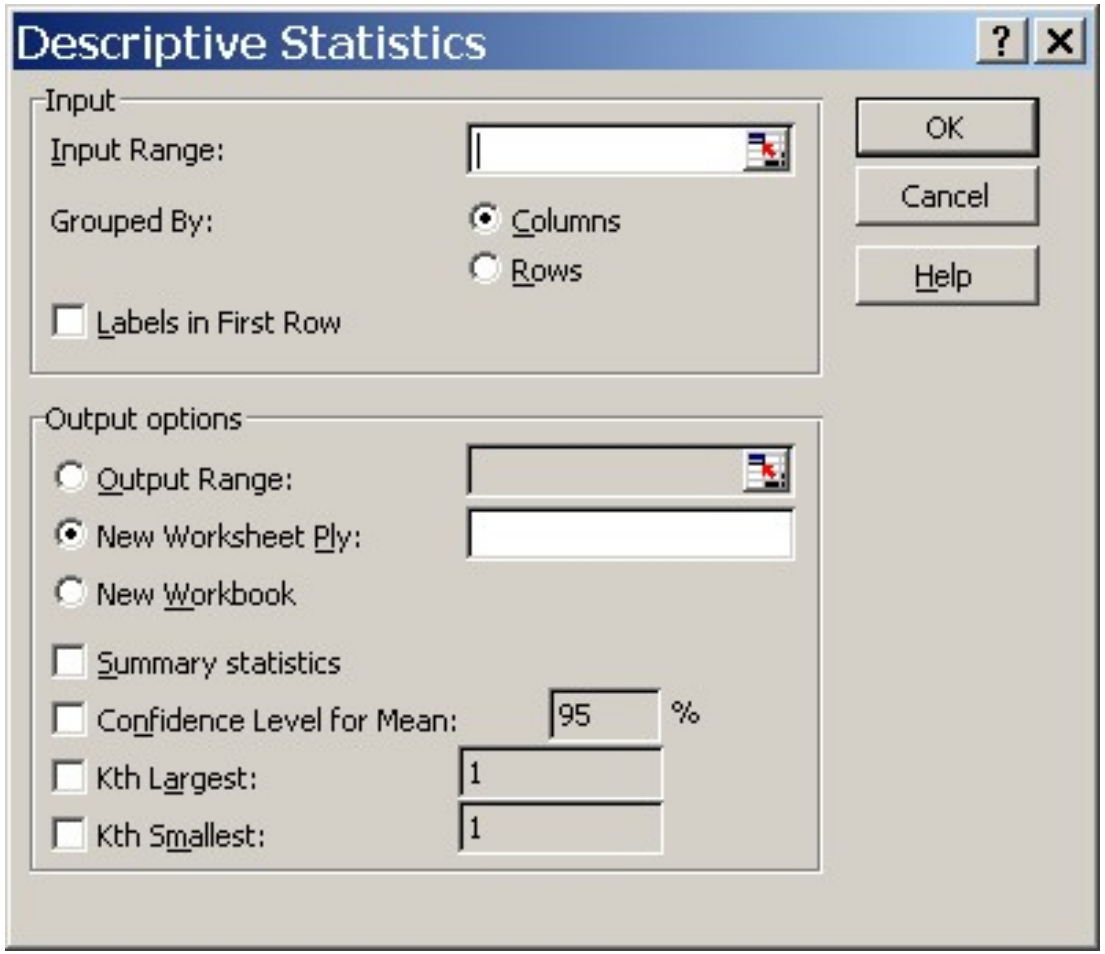
Confidence Intervals

Now we will shift gears from descriptive statistics and start some statistical inference, namely confidence intervals. So we need to go to Descriptive Statistics box



Confidence Intervals

Then, in the following box select, Confidence Interval for Means. The default in EXCEL is 95 percent



Confidence Intervals

The confidence interval is not calculated directly by EXCEL, so you will need to take the formula for a confidence interval and translate into EXCEL language. Recall the formula for the confidence interval for a large sample: μ, 95% confidence interval $(\hat{\mu}_L, \hat{\mu}_U)$ of μ is an interval that satisfies

$$P(\hat{\mu}_L \leq \mu \leq \hat{\mu}_U) = 0.95.$$

We usually make the interval centered so that

$$P(\hat{\mu}_L \leq \mu) = P(\mu \leq \hat{\mu}_U) = 0.025.$$

Table of Data for Confidence Intervals

Obs. Number	MZ1	MZ2
1	88	111
2	115	106
3	88	89
4	92	76
5	115	111
6	48	64
7	83	112
8	77	91
9	99	89
10	77	62
11	109	111
12	95	75

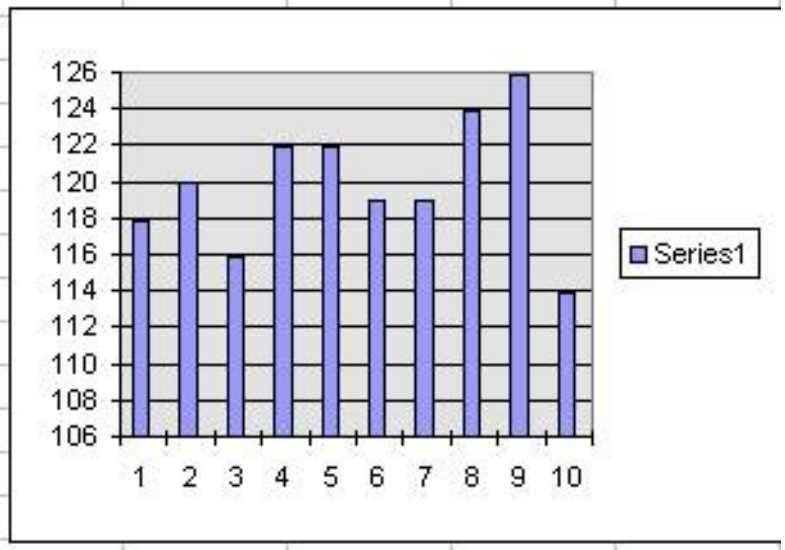
Table: Sample Data- Not all is given

Confidence Intervals

Here is the output from the EXCEL

117.9		*Column1*	
119.9			
115.9		Mean	119.9
121.9		Standard Error	1.145038
121.9		Median	119.4
118.9		Mode	121.9
118.9		Standard Deviation	3.620927
123.9		Sample Variance	13.11111
125.9		Kurtosis	-0.33482
113.9		Skewness	0.017553
		Range	12
		Minimum	113.9
		Maximum	125.9
		Sum	1199
		Count	10
		Confidence Level(95.0%	2.590257

119.9



You will see the confidence interval information in the last row.

Confidence Intervals

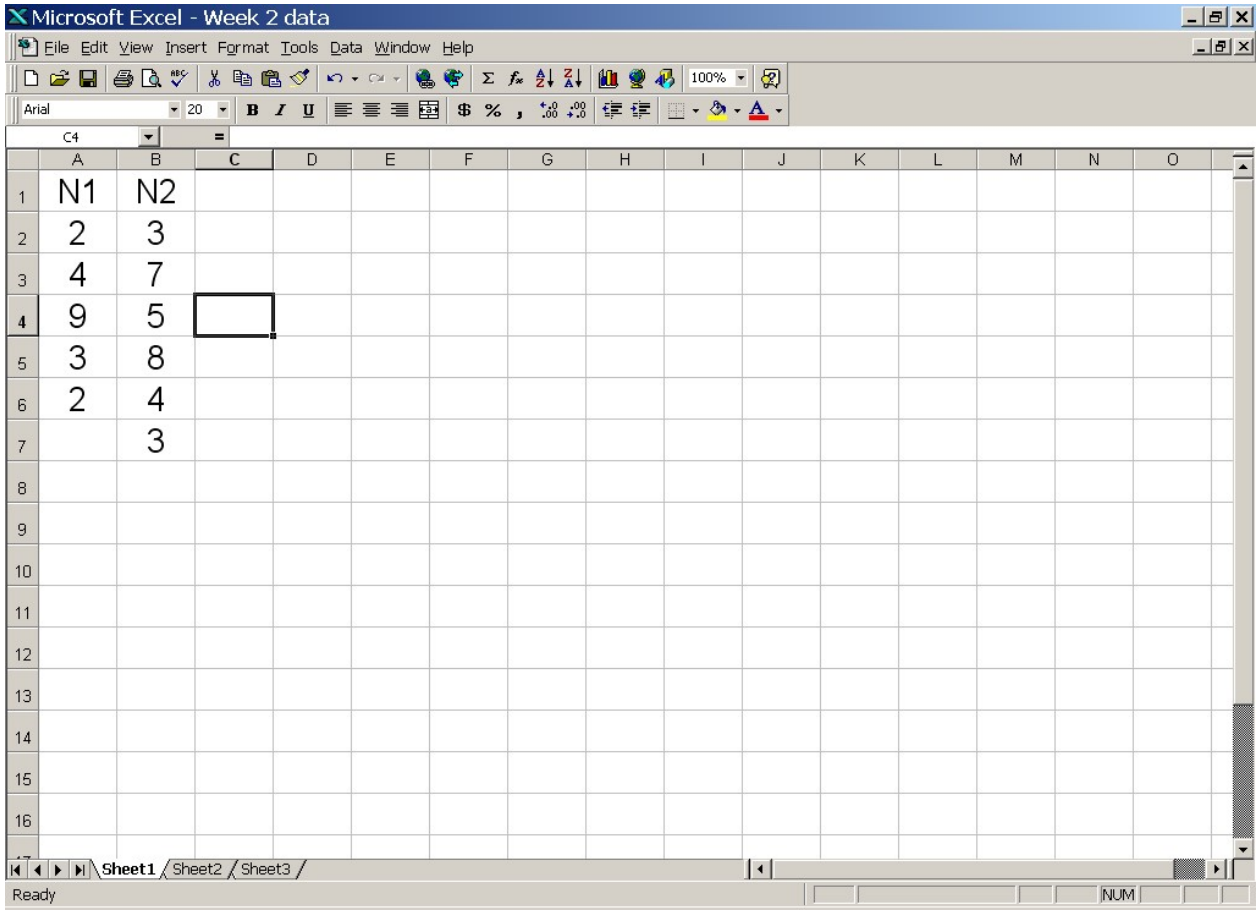
A 95% confidence interval for the population mean is the sample mean plus or minus the "confidence level" reported by EXCEL.
Here this yields 119.90-2.59 and 119.90+2.59 = (117.31, 122.49).
Interpretation: This means that the data are consistent 95% of the time with a data generating process with population mean of mu in the range 117.31 to 122.49.

Inferential Statistics for Test of Means of Two Samples

As long as you have the size of the sample, mean, and standard deviation, a t-test will work on small sample comparison, even if the total sample is not provided. But the t-test is not limited to small sample research designs and can also be used for large samples and can be a fairly robust. There are actually a variety of different designs that compare mean differences.

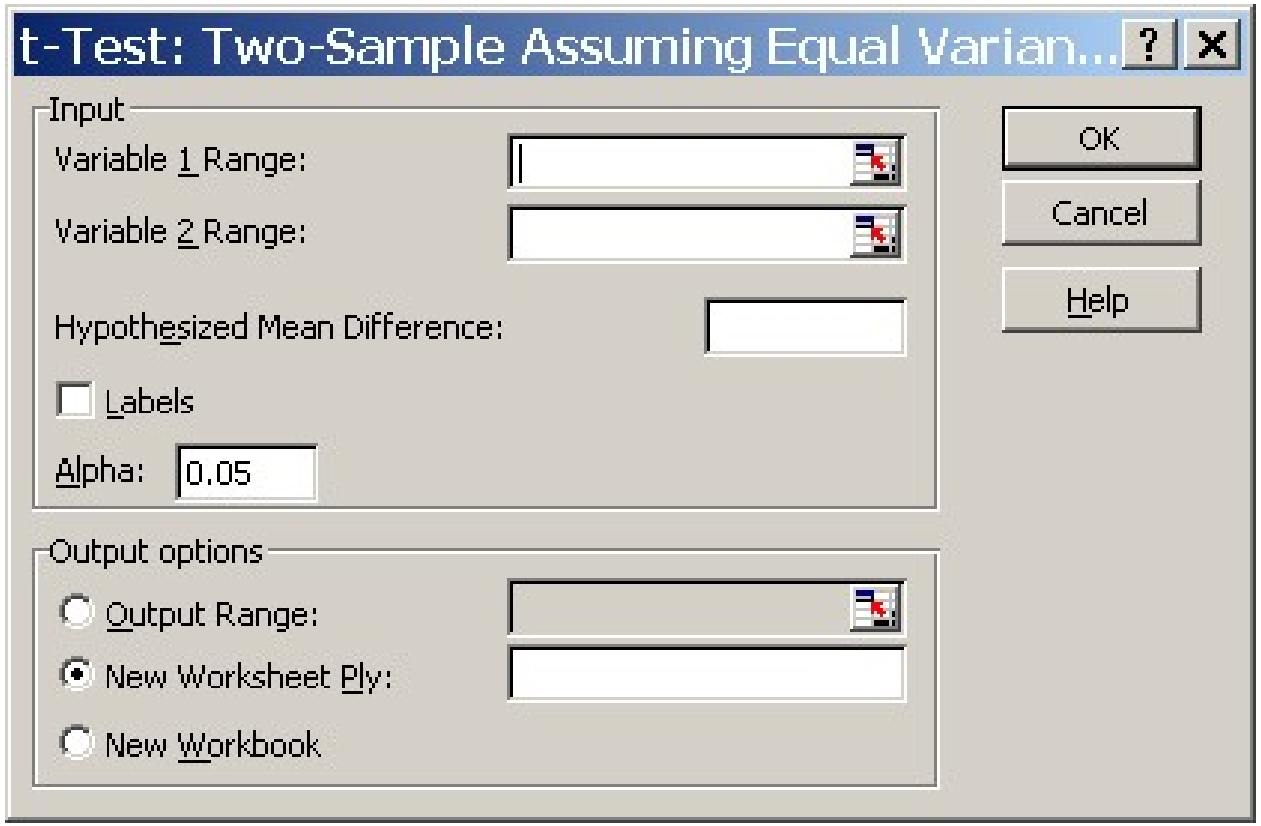
Inferential Statistics for Tests of Means of Two Samples

Open an Excel spreadsheet and enter the values from the following example. Go to Tools-Data Analysis-"t-test: Two Sample Assuming Equal Variances."



Inferential Statistics for Tests of Means of Two Samples

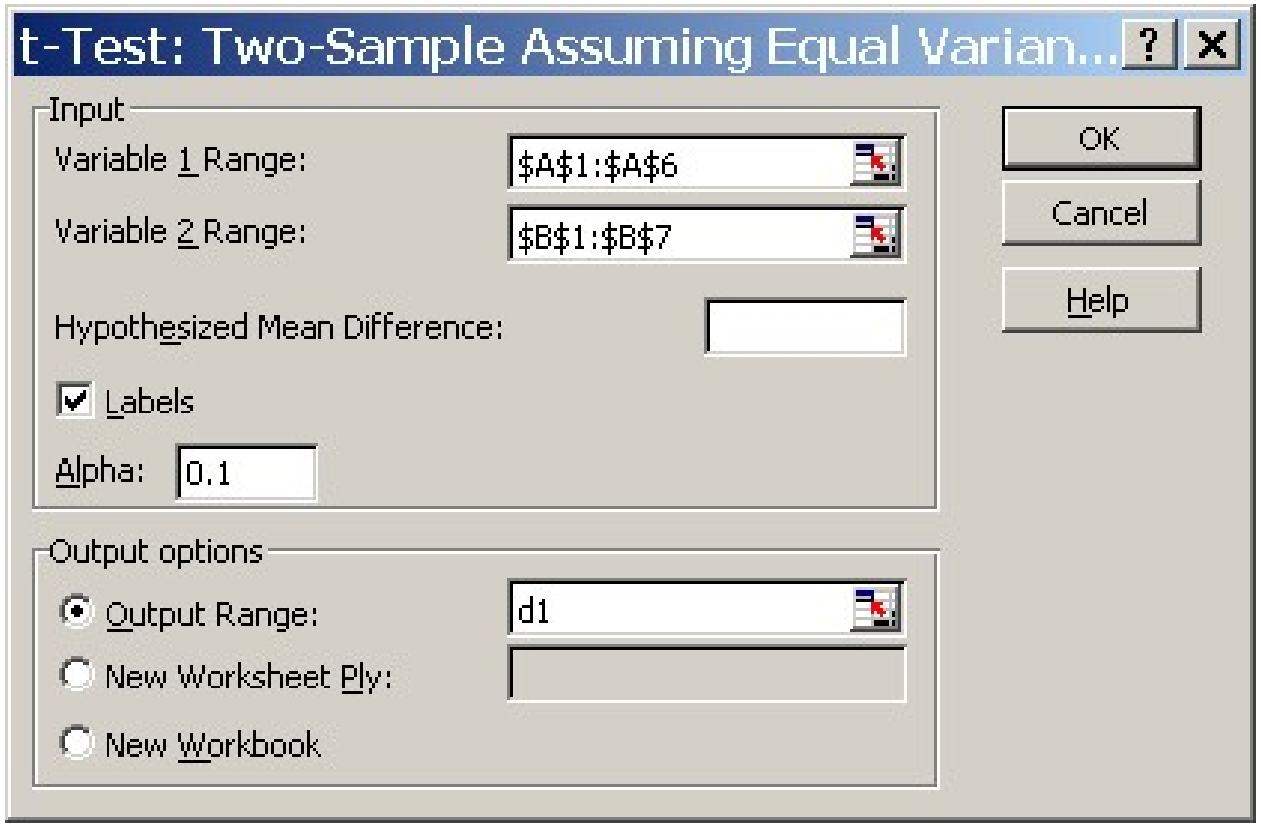
Go to Tools-Data Analysis-"t-test: Two Sample Assuming Equal Variances."



Inferential Statistics for Tests of Means of Two Samples

Enter the data from N1 into Variable 1 Range (A1:A6) and the data from N2 into Variable 2 Range (B1:B7). Don't forget to check the labels box! Notice that the default alpha is .05. What is the alpha of this data? You will need to change the level of significance to .10. Let's make the output on the same page beginning in cell D1. Select OK.

Here is the boxes that need to be filled out."



Inferential Statistics for Tests of Means of Two Samples

Here are the results from EXCEL

The output should look like this:

t-Test: Two-Sample Assuming Equal Variances

	N1	N2
Mean	4	5
Variance	8.5	4
Observations	5	6
Pooled Variance	6.2222222	
Hypothesized Mean Difference	0	
Df	9	
t Stat	-0.662051	
P(T<=t) one-tail	0.2622632	
t Critical one-tail	1.3830288	
P(T<=t) two-tail	**0.5245264**	
t Critical two-tail	1.8331139	

Analysis of Variance (ANOVA)

- An ANOVA (Analysis of Variance), sometimes called an F test, is closely related to the t test. The major difference is that, where the t test measures the difference between the means of two groups, an ANOVA tests the difference between the means of two or more groups
- A one-way ANOVA, or single factor ANOVA as mentioned in EXCEL, tests differences between groups that are only classified on one independent variable which is called the treatments.

Analysis of Variance (ANOVA)

- Example (Using One Way ANOVA) Comparing Hotel Prices Some professional associations are reluctant to hold meetings in New York because of high hotel prices and taxes. Are hotels in New York more expensive than hotels in other major cities?
- We have a random sample of eight hotels and their prices that were taken from the 1992 Mobil Travel Guide to Major Cities.

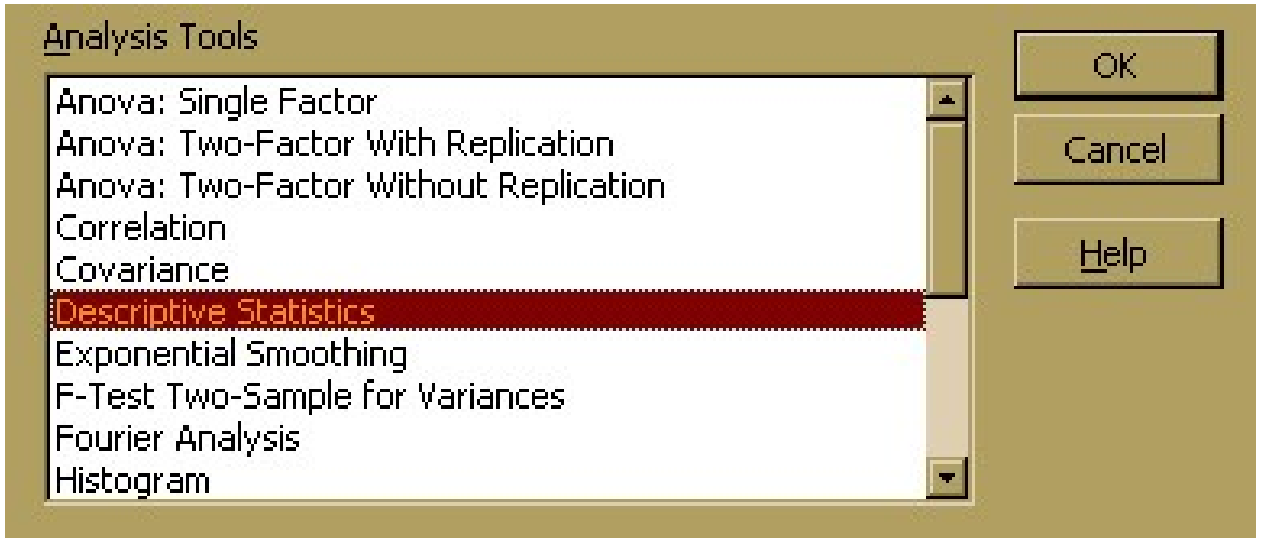
Analysis of Variance (ANOVA)

LA	SF	DC	NY
119	99	115	170
150	185	185	135
110	265	186	185
79	109	189	250
145	169	125	250
140	99	64	170
165	175	120	210
175	110	119	215

Table: Hotel Data

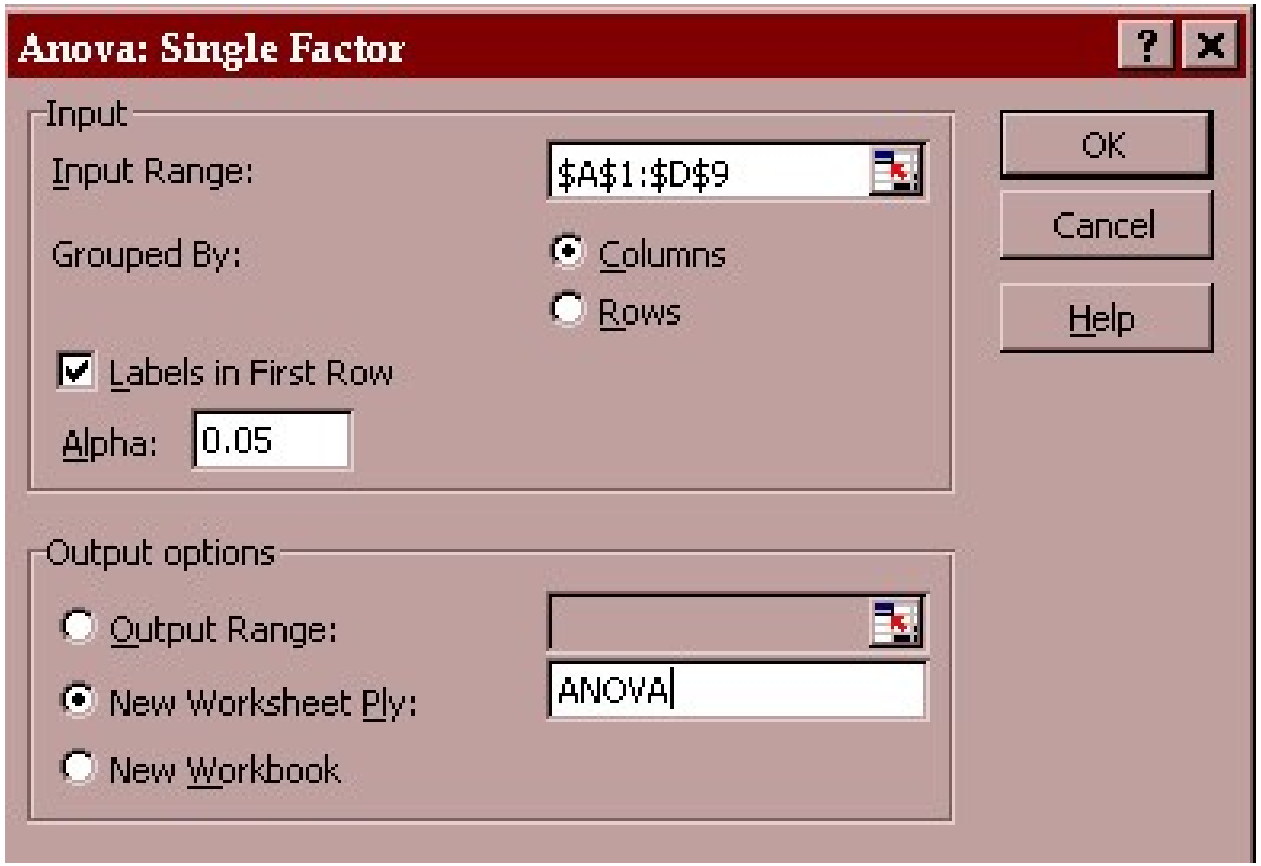
Analysis of Variance (ANOVA)

Click on Tools, then on Data Analysis. When you do this, you will see the following screen.



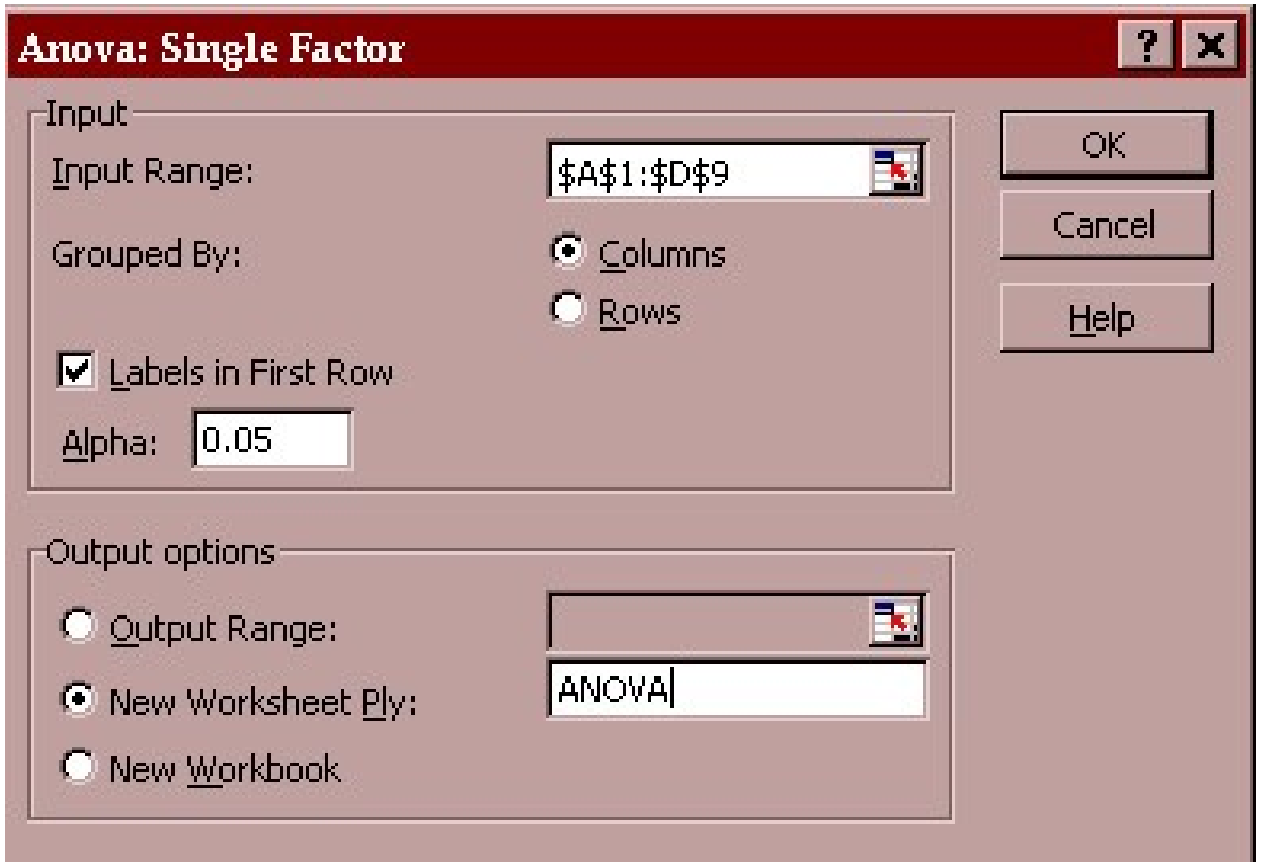
Analysis of Variance (ANOVA)

ANOVA: Single Factor is the first tool on the list. Click on it, then click OK. You will see the following.



Analysis of Variance (ANOVA)

Choose New Worksheet Ply and type the name, ANOVA, to the right. After you have entered all of these values, the screen should look like the following.



Analysis of Variance (ANOVA)

After you have entered all the data, click OK. Excel will calculate the ANOVA table. You should see a table like the following.

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	LA	8	1083	135.375	980.8393		
6	SF	8	1211	151.375	3414.839		
7	DC	8	1083	135.375	1771.125		
8	NY	8	1585	198.125	1649.554		
9							
10							
11	ANOVA						
12	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
13	Between Groups	21145.38	3	7048.458	3.60703	0.02549	2.946685
14	Within Groups	54714.5	28	1954.089			
15							
16	Total	75859.88	31				
17							
18							

Analysis of Variance (ANOVA)

- Basically, the ANOVA compares two variances: the between-city variance vs. the within-city variance. If the between-city variance is much higher than the within-city variation, the cities are significantly different.
- The p-value is 0.025 (cell F13), which is less than the alpha value we specified so we should reject the hypothesis of equal city means at the 5% level.
- Therefore, we could say that, at the 5% significance level, it appears that the expected price of a hotel room is not the same in the four cities.

ANOVA: Two-Factor Without Replication Analysis

- This analysis tool performs a two-factor ANOVA that does not include more than one sampling per group, testing the hypothesis that means from two or more samples are equal (drawn from populations with the same mean).
- For example, if the experimenter in the above Cola example had tested only one can of soda for each of the eight trials instead of using a new can for each trial, we would use the two factor without replication.

ANOVA: Two-Factor Without Replication Analysis

Example: As a production manager, you want to see if 3 filling machines have different mean filling times when used with 5 types of boxes. At the .05 level, is there a difference in machines, in boxes? The data can be given as follows:

Box	Machine 1	Machine 2	Machine 3
1	25.40	21.80	20.00
2	26.31	23.40	22.20
3	24.10	22.75	19.75
4	23.74	23.50	20.60
5	25.10	21.60	20.40

Table: Production Data

ANOVA: Two-Factor Without Replication Analysis

The results from this ANOVA are given as follows:

ANOVA: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
25.4	2	43.4	21.7	5.78
26.31	2	44	22	0.08
24.1	2	43.25	21.625	7.03125
23.74	2	43.35	21.675	2.31125
25.1	2	42	21	0.72
machine 2	5	113.05	22.61	0.778
machine 3	5	102.95	20.59	0.9205

ANOVA

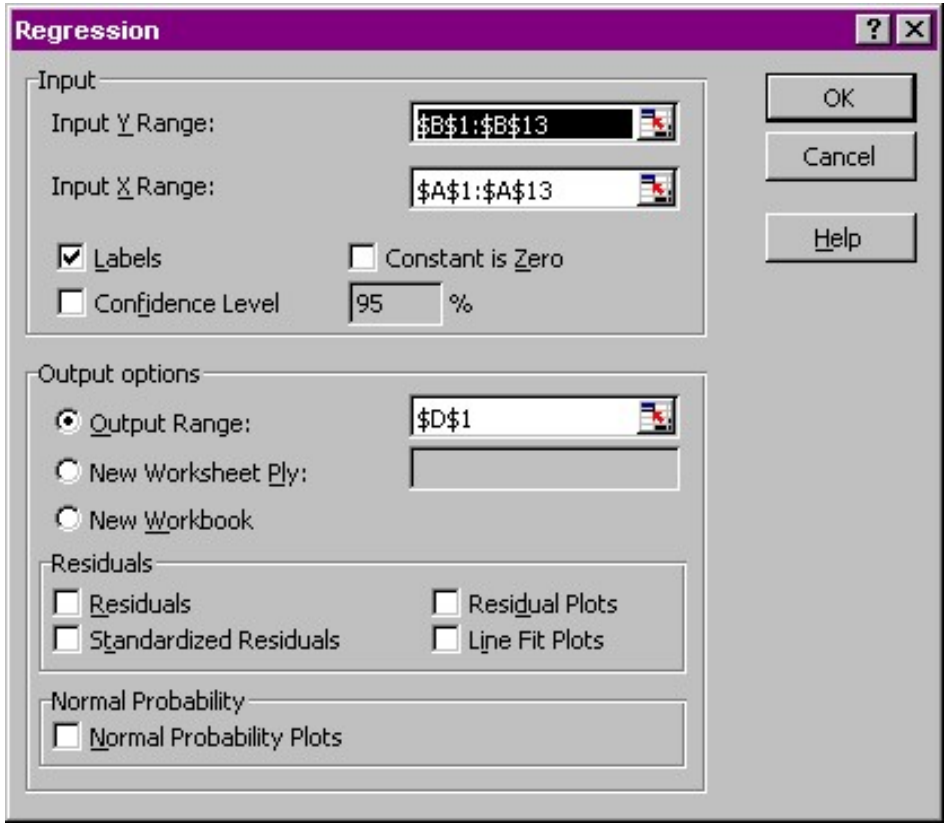
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.0725	4	0.268125	0.187451	0.933108	6.388233
Columns	10.201	1	10.201	7.131696	0.055776	7.708647
Error	5.7215	4	1.430375			
Total	16.995	9				

ANOVA: Two-Factor Without Replication Analysis

- By looking at the p values we can determine the results. Looking at the columns (the machines also called the treatments), the p value is .055 which is greater than the level of significance of .05. So there are no differences between the means.
- For the rows, which represents the boxes. Its p-value is .933 which is greater than the level of significance of .05. So there are no differences between the block means.
- **Remark:** Though the means differ, we cannot say there is a difference because this is based on causal observation which is scientific as was just done here.

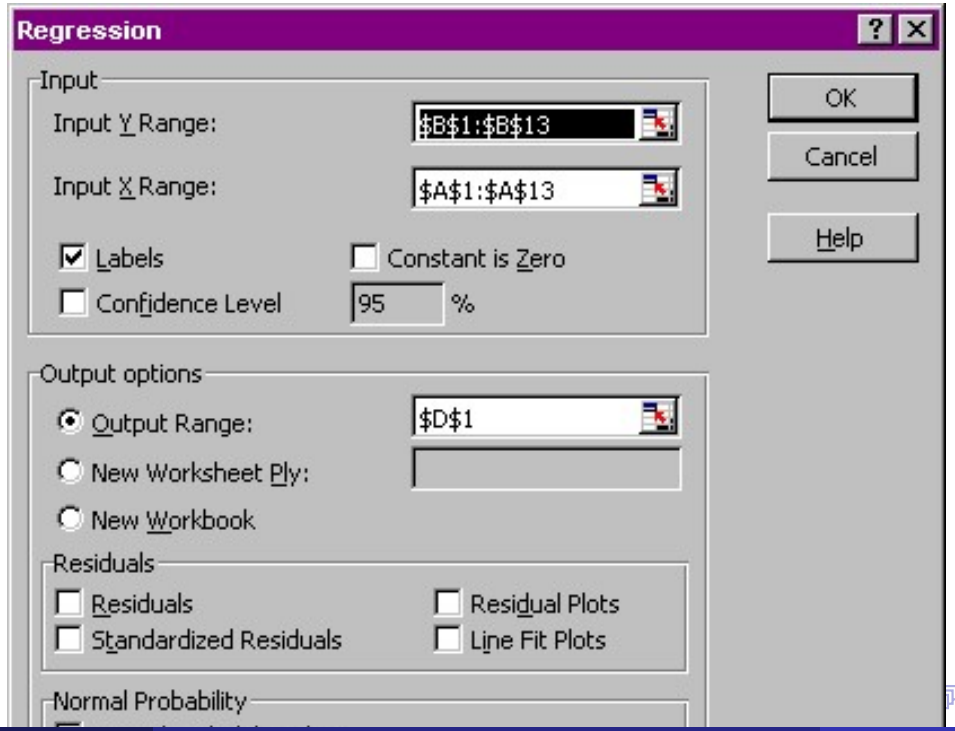
Regression Analysis

It is standard convention to list the x variable before the y variable in a table. You should notice that in Excel, the y-variable is listed first. If the variables are not entered properly—you will end up with the wrong results (GIGO).



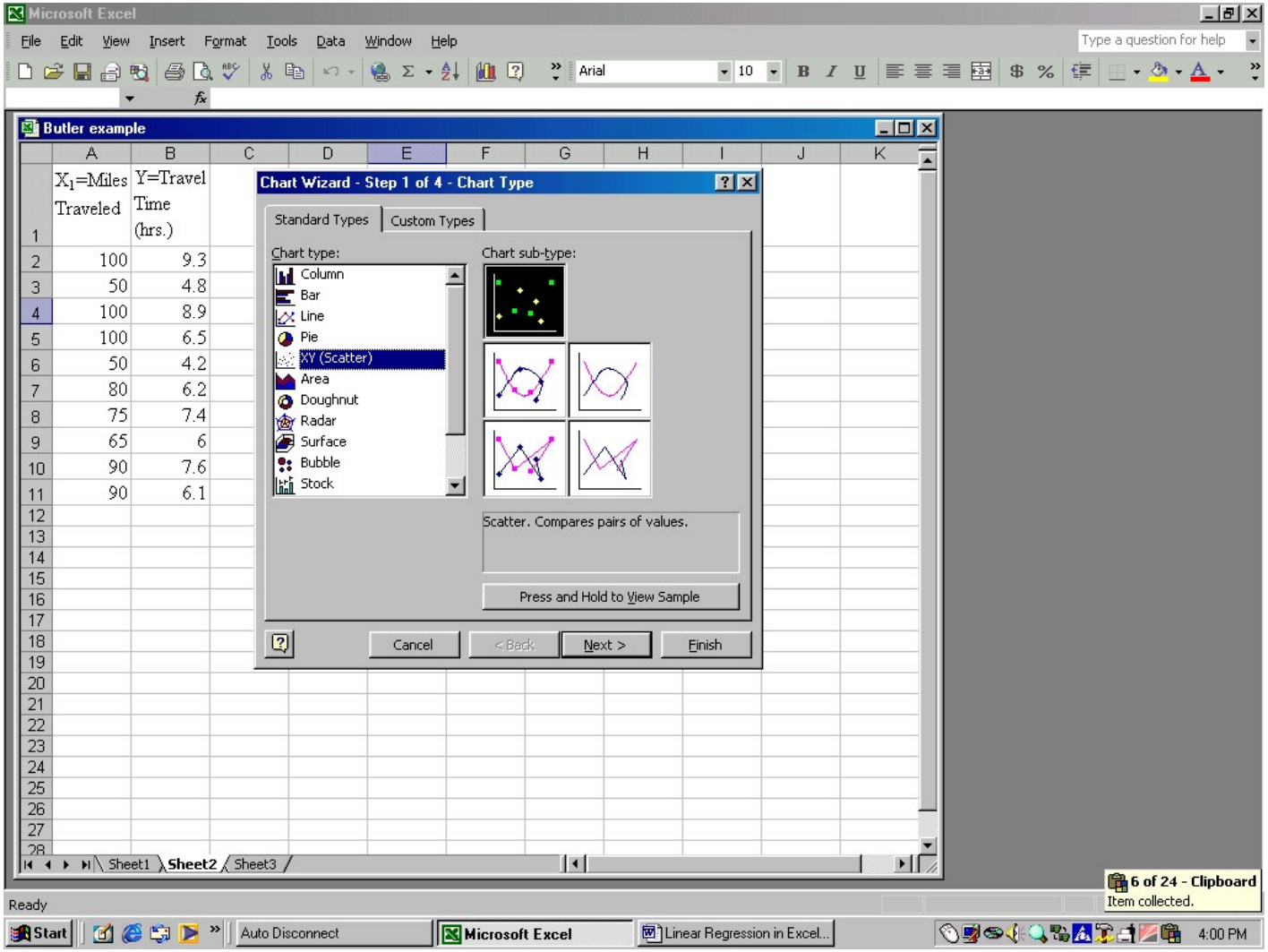
Regression Analysis

Butler's Trucking Company is an independent trucking Company in southern California. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to estimate the total daily travel time for their drivers.
Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries. A simple random sample of 10 driving assignments is provided



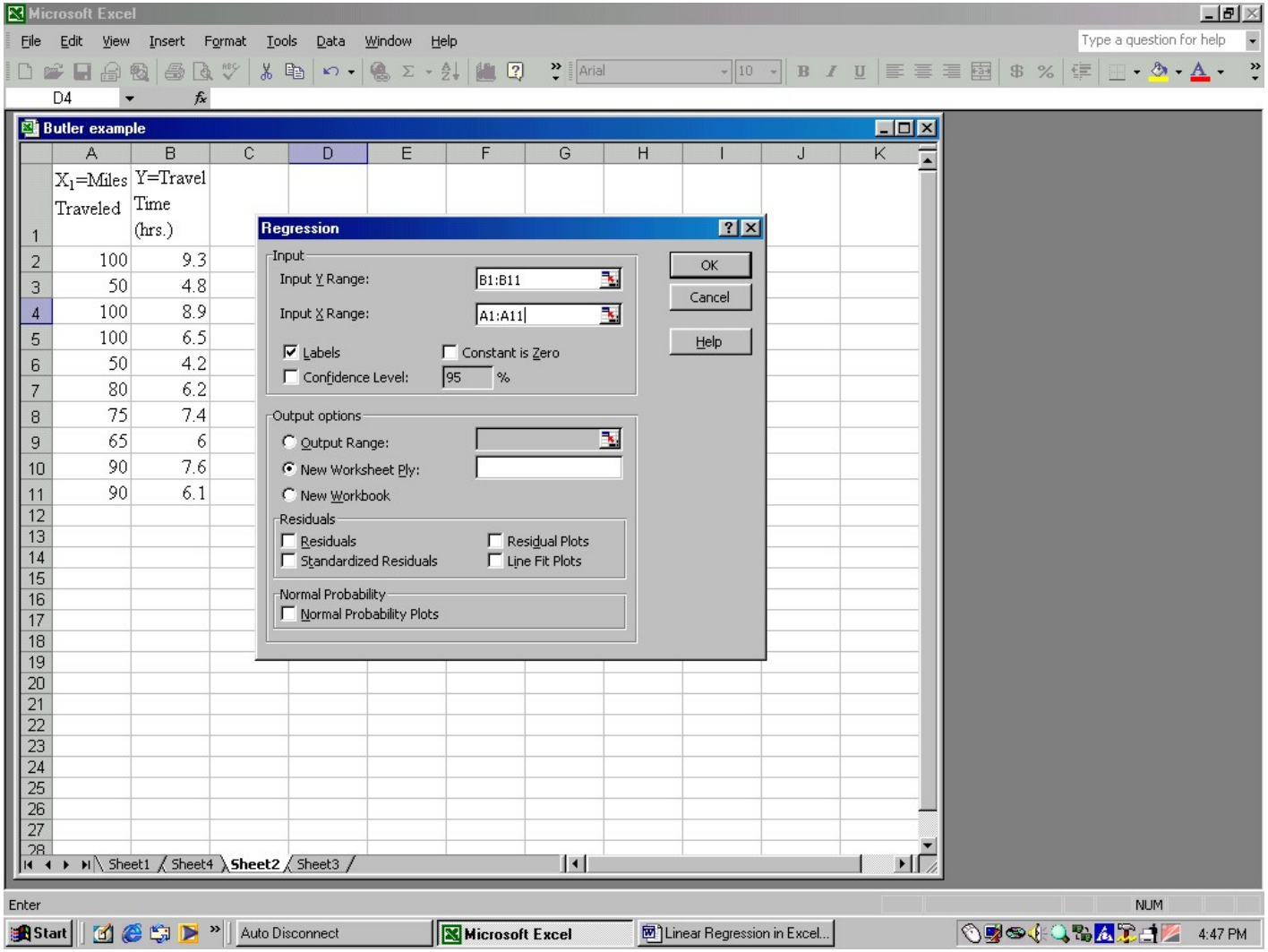
Regression Analysis

Here is the random sample of 10 observations for this example



Regression Analysis

Here is the entering the data into the EXCEL regression.



Regression Analysis

The results from EXCEL

SUMMARY OUTPUT

Correlation Coefficient

Coefficient of Determination

Regression Statistics	
Multiple R	0.8149057
R Square	0.6640713
Adjusted R Square	0.6220802
Standard Error	1.0017919
Observations	10

P value ANOVA

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15.87130435	15.871304	15.814578	0.004080177
Residual	8	8.028695652	1.003587		
Total	9	23.9			

b_0

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.273913	1.400744525	0.9094542	0.3896874	-1.95621171	4.5040378
X1=Miles Traveled	0.0678261	0.017055637	3.9767547	0.0040802	0.028495691	0.1071565

b_1

P value for t test for X_1

Regression Analysis

- In the first summary table, you will find the Coefficient of Determination, R2. Interpretation: 66.4% of the variation in travel time is explained by miles traveled. So 23.6 percent is not explained by the regression.
- For the multiple R, this is the correlation coefficient which .81. Interpretation: It is a strong positive correlation between the miles traveled and travel times.
- The ANOVA table gives the F statistic for testing the claim that there is no significant relationship between your independent and dependent variables. The sig. value is your p value. Interpretation: Since the .004 is less than .05, the model as a whole is good.
- The Columns below the Coefficients box gives the b0 and b1 values for the regression equation. The intercept value is always b0. The b1value is next to your independent variable, x. The regression equation is Travel Miles= 1.23+ .067*MILES TRAVELED

Regression Analysis

- Interpretation: Now we can interpret the slope. The slope is .067 in this simple regression. If there is one additional mile traveled, then travel miles would increase by .067.
- In the last P-value column of the coefficient output data, the p values for individual t tests for our independent variable is given (in the same row as your independent variable). Recall that this t test tests the claim that there is no relationship between the independent variable and your dependent variable. Thus you should reject the claim that there is no significant relationship between your independent variable and dependent variable if p¡. Interpretation: Since the p-value is .004 and it is less than the level of significance of .05, we would reject the null hypothesis and conclude that there is a significant relationship. (Do not need to interpret the constant or the y intercept term)

Table

This section shows the regression analysis which means more than one x variable. Expanding on the data from the earlier regression.

Miles Traveled	Deliveries	Travel Time
100	4	9.3
50	3	4.8
100	4	8.9
100	2	6.5
50	2	4.2
80	2	6.2
75	3	7.4
65	4	6.0
65	3	7.6
90	2	6.1

Table: Delivery Data

Multiple Regression Analysis

The results from EXCEL

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950678166
R Square	0.903788975
Adjusted R Square	0.876300111
Standard Error	0.573142152
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	21.60055651	10.80028	32.87836743	0.00027624
Residual	7	2.299443486	0.328492		
Total	9	23.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.868701467	0.951547725	-0.91294	0.391634304	-3.118752683	1.38134975
X1=Miles Traveled	0.061134599	0.009888495	6.182397	0.000452961	0.037752041	0.084517156
X2=Number of Deliveries	0.923425367	0.221113461	4.176251	0.004156622	0.400575489	1.446275244

Multiple Regression Analysis

- In the first summary table, you will find the Coefficient of Determination, R squared. Interpretation: 90.3% of the variation in travel time is explained by miles traveled and number of delivers. So 9.7% is not explained by the regression.
- Next look at the adjusted R-squared. This is the penalty for adding more independent variables to the regression equation. Interpretation: 87.7 percent of the variance travel time is explained by miles traveled and number of delivers. (Note: this measure is not interpreted in simple regression)
- The ANOVA table gives the F statistic for testing the claim that there is no significant relationship between your independent and dependent variables. The sig. value is your p value. Interpretation: Since the .00027 is less than .05, the model as a whole is good.

Multiple Regression Analysis

- The Columns below the Coefficients box gives the b0 and b1 values for the regression equation. The intercept value is always b0. The b1value is next to your independent variable, x. The regression equation is
  Travel Miles= .868+ .061*MILES TRAVELED+.923*NUMBER DELIVERIES
  Interpretation: Now we can interpret the slope. The slope of MILES TRAVELED is .061. If there is one additional mile traveled, then travel miles would increase by .061. The slope of NUMBER DELIVERIES is .923. If there is one additional delivery, then the travel miles would increase by .923.
- In the last P-value column of the coefficient output data, the p values for individual t tests for our independent variable is given (in the same row as your independent variable). Recall that this t test tests the claim that there is no relationship between the independent variable and your dependent variable.

Multiple Regression Analysis

- You should reject the claim that there is no significant relationship between your independent variable and dependent variable if p¡alpha. Interpretation: Since the p-value is .004 and it is less than the level of significance of .05, we would reject the null hypothesis and conclude that there is a significant relationship. For the NUMBER OF DELIVERIES, the p value is .0004 which is also less than the level of significance so it is significant.

The End