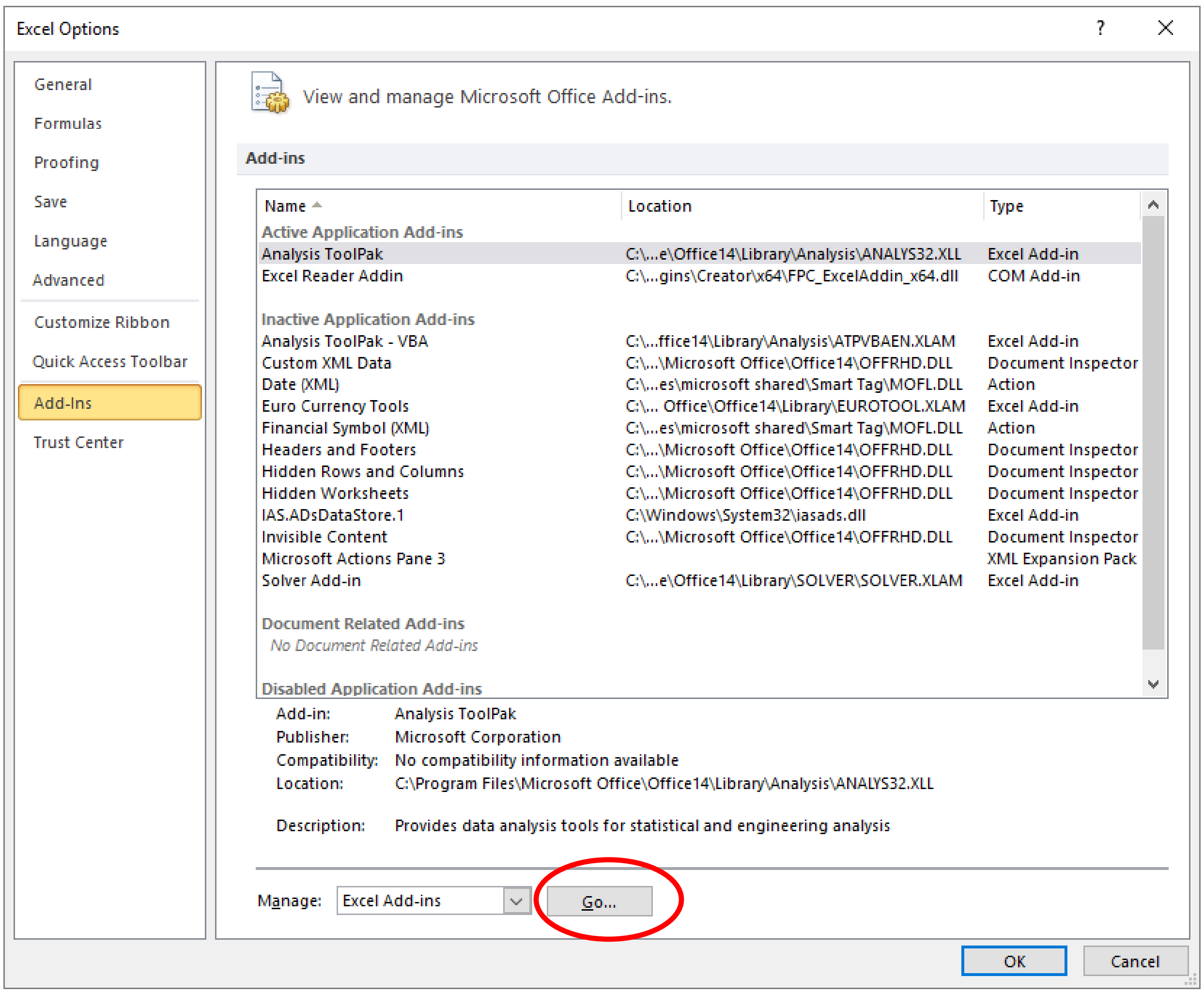
Lab 4: Data Analysis in Excel

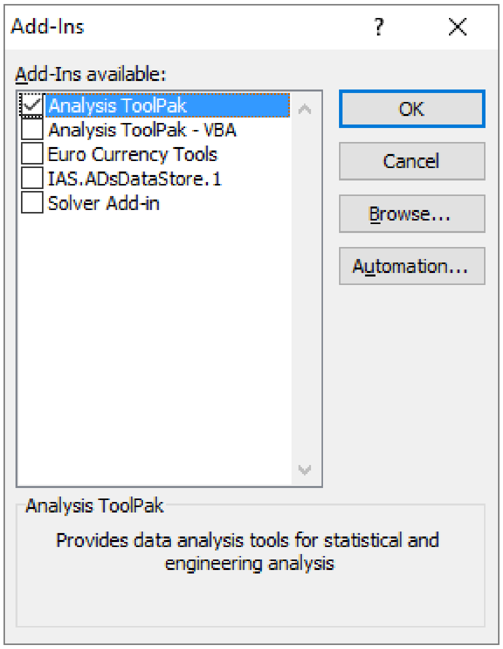
Analysis ToolPak

1. Load ToolPak add-in

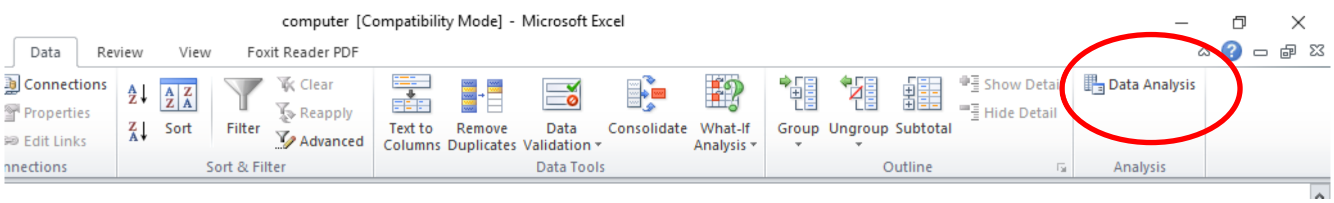
1. Click on the green “File” tab → select “Option”→ An excel option dialog appears, select “Add-ins”.



2. Click Go button,The Add-Ins dialog appears, Check “Analysis ToolPak” and Ok.



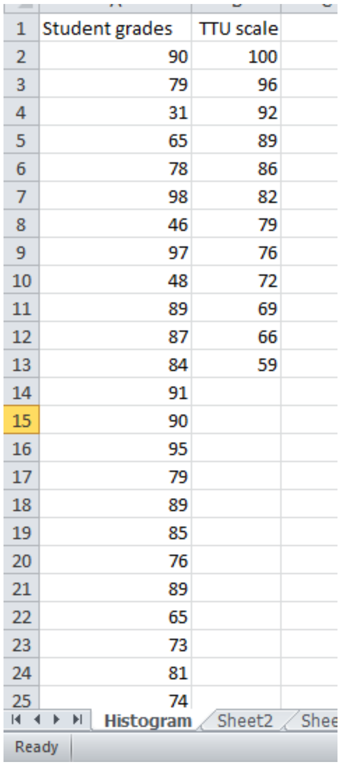
The Data Analysis tool bar now appears under the Data tab.



2. Histogram

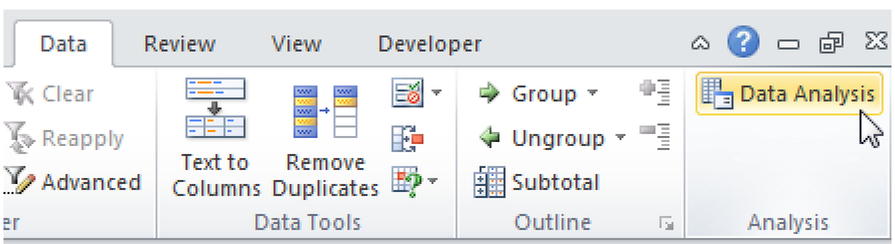
A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable).

1. First, enter the bin numbers (upper levels), for example TTU's scale (A2:A13)

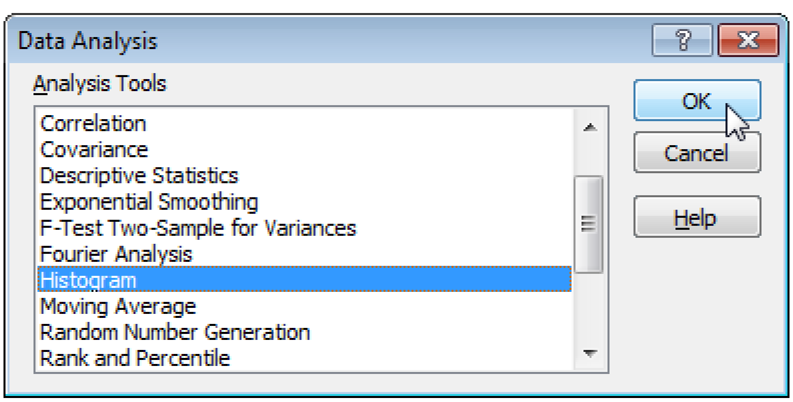


	Student grades	TTU scale
1	Student grades	TTU scale
2	90	100
3	79	96
4	31	92
5	65	89
6	78	86
7	98	82
8	46	79
9	97	76
10	48	72
11	89	69
12	87	66
13	84	59
14	91	
15	90	
16	95	
17	79	
18	89	
19	85	
20	76	
21	89	
22	65	
23	73	
24	81	
25	74	

2. On the Data tab, click Data Analysis.



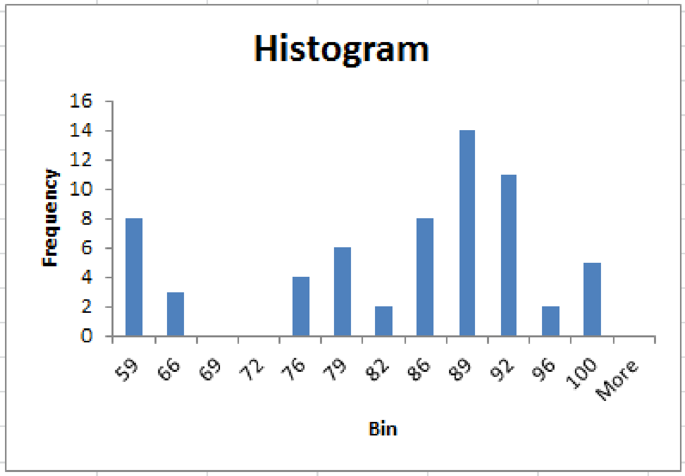
Select Histogram and click OK.



3. . Select the range (student grade column)
4. Click in the Bin Range box and select the range (column TTU scale).

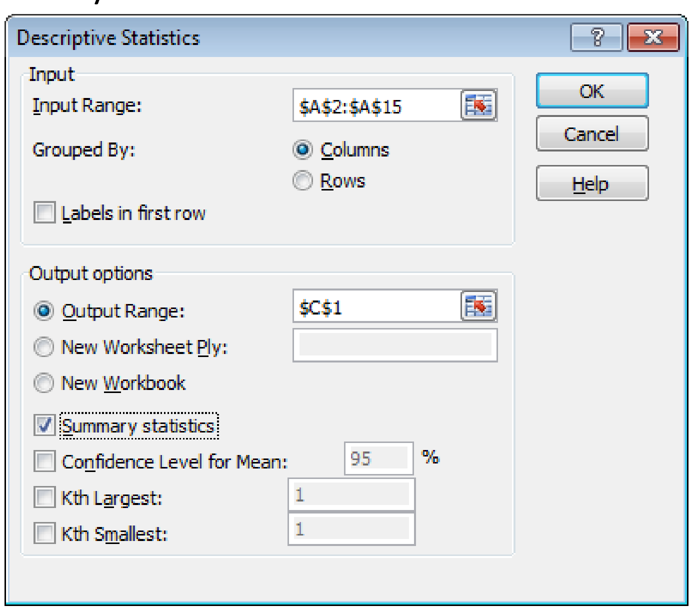
5. Click the Output Range option button, click in the Output Range box and select any cell.
6. Check Chart Output.

Bin	*Frequency*
59	8
66	3
69	0
72	0
76	4
79	6
82	2
86	8
89	14
92	11
96	2
100	5
More	0



3. Descriptive Statistics

1. On the Data tab, click Data Analysis.
2. Select Descriptive Statistics and click OK.
3. Select the range (Student grade) as the Input Range.
4. Select any cell as the Output Range.
5. Make sure Summary statistics is checked.



Result

	Column1
Mean	80.20634921
Standard Error	2.04945995
Median	87
Mode	89
Standard Deviation	16.26708405
Sample Variance	264.6180236
Kurtosis	1.910532125
Skewness	-1.618207824
Range	67
Minimum	31
Maximum	98
Sum	5053
Count	63

4. Analysis of variance (ANOVA)

a. A single factor

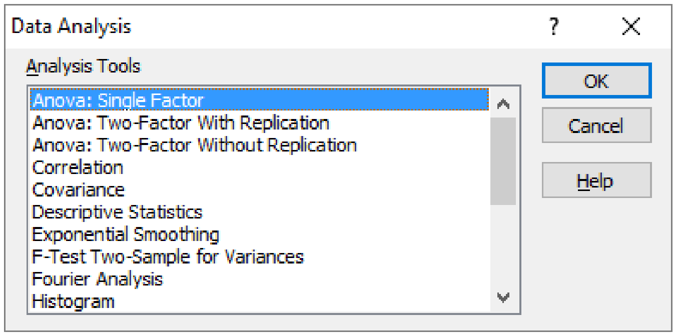
A single factor or one-way ANOVA is used to test the null hypothesis that the means of several groups are identical.

- Suppose that the benefit of a company is separated by different regions as follows.

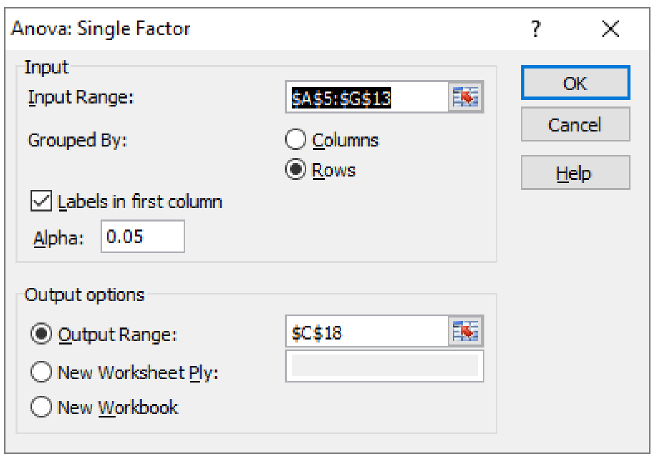
	Region	2009	2010	2011	2012	2013	2014
2	G1	23454	24872	19070	21308	21676	22938
3	G1	21376	21876	21736	21467	23731	22685
4	G3	24501	20231	18033	23457	20436	22215
5	G4	20316	23075	20018	19543	20813	19066
6	G5	20134	23886	23220	20962	18206	20855
7	G6	19219	19715	22526	20019	24654	23925
8	G6	21795	24508	18344	22379	21641	20948
9	G8	24206	19001	21664	18530	21827	23389
10	G9	19962	23148	23695	18376	22462	18786
11							

- Now we consider that there is a difference of benefit among regions (i.e., single factor)?
- Hypothesis
  - $H_0 = M_{G1} \approx M_{G2} \approx M_{G3} \approx M_{G4} \approx M_{G5} \approx M_{G6} \approx M_{G7} \approx M_{G8} \approx M_{G9}$
  - H_1 : there is a difference among regions
- Check these hypothesis using Excel

1. Select Data tab → click Data Analysis → select Anova: single factor



2. Select Input range, choose group by Rows (enter alpha value, not important), select output range (any cell)



3. Result

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
G1	6	133318	22219.66667	4024312.667
G2	6	132871	22145.16667	819536.5667
G3	6	128873	21478.83333	5621386.567
G4	6	122831	20471.83333	1992443.767
G5	6	127263	21210.5	4314079.1
G6	6	130058	21676.33333	5455400.667
G7	6	129615	21602.5	4025922.7
G8	6	128617	21436.16667	5213206.967
G9	6	126429	21071.5	5369663.1

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	13805553	8	1725694.125	0.421632841	0.901935	2.152133
Within Groups	184179760.5	45	4092883.567			
Total	197985313.5	53				

4. Conclusion: $F<F_{critical}$, we accept the hypothesis H_0. If H_0 is rejected, means that at least one of the means is different. However, the ANOVA does not tell you where the difference lies. You need a T-Test (later) to test each pair of means.

b. Two-way ANOVA without replication

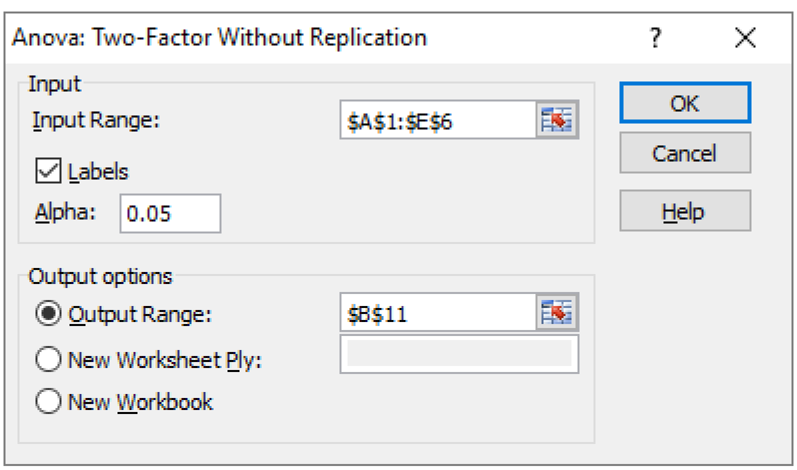
Similar as One-Way Anova, but we consider the influence of two factors on a dependent variable. For example: there are 4 experts to predict the development rate of 5 companies as follows.

Company	Expert 1	Expert 2	Expert 2	Expert 4
C1	8	12	8.5	13
C2	14	10	9	11
C3	11	9	12	10
C4	9	13	10	13
C5	12	10	10	10

Question: is there a difference of mean of development rate among 5 companies and experts?

Using Excel to check this hypothesis as follows.

1. Select Data tab → click Data Analysis → select Anova: two factor without replication → select input range and output range. (change alpha if needed)



2. Result

Anova: Two-Factor Without Replication							
SUMMARY	*Count*	*Sum*	*Average*	*Variance*			
C1	4	41.5	10.375	6.229166667			
C2	4	44	11	4.666666667			
C3	4	42	10.5	1.666666667			
C4	4	45	11.25	4.25			
C5	4	42	10.5	1			
Expert 1	5	54	10.8	5.7			
Expert 2	5	54	10.8	2.7			
Expert 2	5	49.5	9.9	1.8			
Expert 4	5	57	11.4	2.3			
ANOVA							
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
Rows	2.3	4	0.575	0.144654088	0.961924359	3.25916673	
Columns	5.7375	3	1.9125	0.481132075	0.701475354	3.49029482	
Error	47.7	12	3.975				
Total	55.7375	19					

3. Conclusion:
   - F_Rows < F_Rows $_{Crit}$ and F_columns < F_columns $_{Crit}$, we accept the null hypothesis.

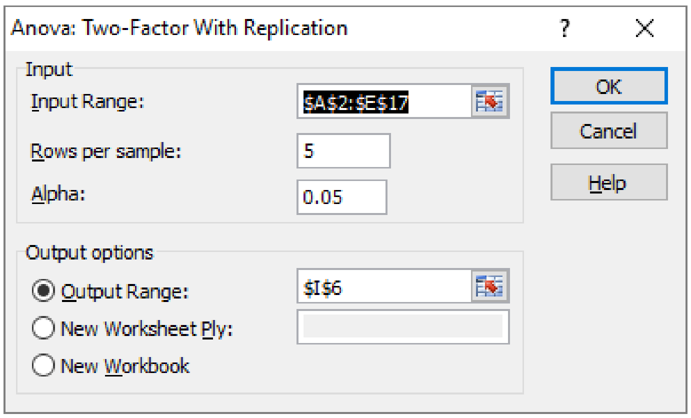
c. Two-way ANOVA with replication

Similar as two-way ANOVA without replication, however a sample has some rows of data as following table.

Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175

	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

1. Select Data tab → click Data Analysis → select **Anova: two factor with replication** → select input range and output range. (change alpha if needed) → Indicate rows per sample, here is 5 (i.e., there are 5 data rows per fertilizer).



2. Result

Anova: Two-Factor With Replication

SUMMARY	Wheat	Corn	Soy	Rice	Total
Blend X					
Count	5	5	5	5	20
Sum	659	677	879	706	2921
Average	131.8	135.4	175.8	141.2	146.05
Variance	844.2	707.8	278.7	354.2	782.3658
Blend Y					
Count	5	5	5	5	20
Sum	716	798	701	827	3042
Average	143.2	159.6	140.2	165.4	152.1
Variance	498.7	978.3	165.7	217.3	511.0421
Blend Z					
Count	5	5	5	5	20
Sum	822	868	932	862	3484
Average	164.4	173.6	186.4	172.4	174.2

Variance	443.3	428.8	212.3	175.8	330.6947	
Total						
Count	15	15	15	15		
Sum	2197	2343	2512	2395		
Average	146.4667	156.2	167.4667	159.6667		
Variance	705.8381	871.0286	605.981	404.9524		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	8782.9	2	4391.45	9.933347	0.000245	3.190727
Columns	3411.65	3	1137.217	2.572355	0.064944	2.798061
Interaction	6225.9	6	1037.65	2.347138	0.045555	2.294601
Within	21220.4	48	442.0917			
Total	39640.85	59				

3. Conclusion:
    - F_sample > F_sample_{Crit} ➔ there is a difference according to fertilizer
    - F_columns < $F\_columns_{Crit}$ ➔ accept the null hypothesis (there is no difference among type of crop)
    - F_interaction > $F\_interaction_{Crit}$ ➔ reject the null hypothesis (there is a difference)

5. F-Test

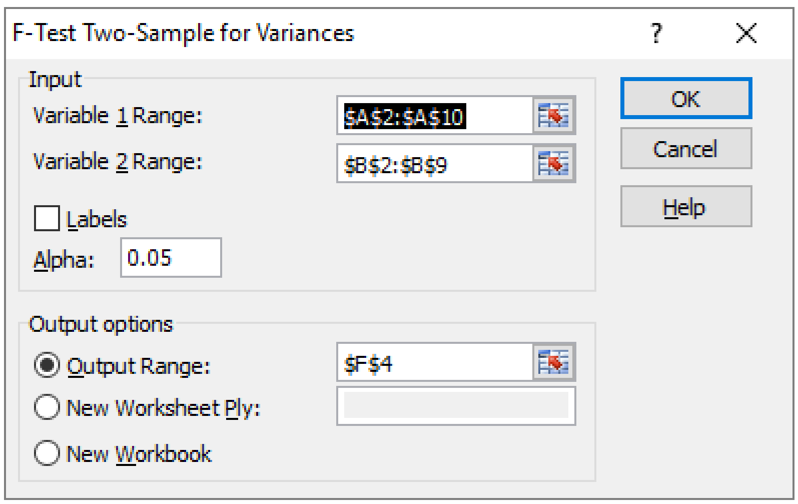
The F-Test is used to test the null hypothesis that the variances of two populations are equal.

Example:

- The study hours of 9 female students and 8 male students as follows.
- There is a difference of variance of two samples?

Female	Male
26	19
25	25
43	23
34	30
18	18
52	25
45	23
26	28
29	

1. Select Data tab → click Data Analysis → select F-Test Two-Sample for Variances and click OK. → select variance 1 and variance 2 range. Choose output range (any cell).



2. Result

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	33.11111111	23.875
Variance	126.1111111	16.69642857
Observations	9	8
df	8	7
F	7.553178847	
P(F<=f) one-tail	0.007518426	
F Critical one-tail	3.725725317	

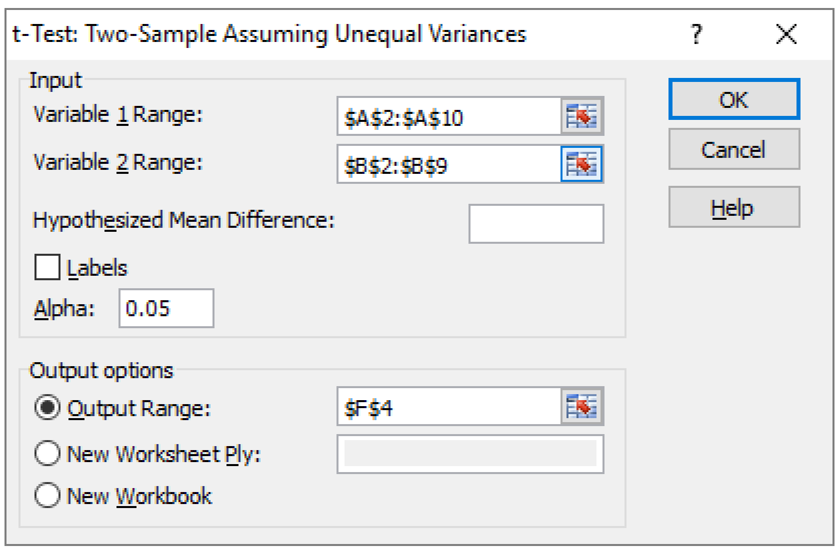
Conclusion: F > F_Crit → reject null hypothesis (there is a difference)

NOTE: be sure that the variance of Variable 1 is higher than the variance of Variable 2. If not, swap your data.

6. T-Test

The t-Test is used to test the null hypothesis that the means of two populations are equal. For example: use the same data as F-Test, we compare the means of study hour of female and male. *Null hypothesis is equal.*

1. Select Data tab → click Data Analysis → select t-Test: Two Sample Assuming Unequal Variance (we choose this because we have knew the variance of female and variance of male are different).



2. Result

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	33.11111111	23.875
Variance	126.1111111	16.69642857
Observations	9	8
Hypothesized Mean Difference	0	
df	10	
t Stat	2.301888661	
P(T<=t) one-tail	0.022056149	
t Critical one-tail	1.812461123	

P(T<=t) two-tail	0.044112298
t Critical two-tail	2.228138852

Conclusion: We do a two-tail test (inequality). If t Stat < -t Critical two-tail or t Stat > t Critical two-tail, we reject the null hypothesis.