Department of Statistics and Operations Research

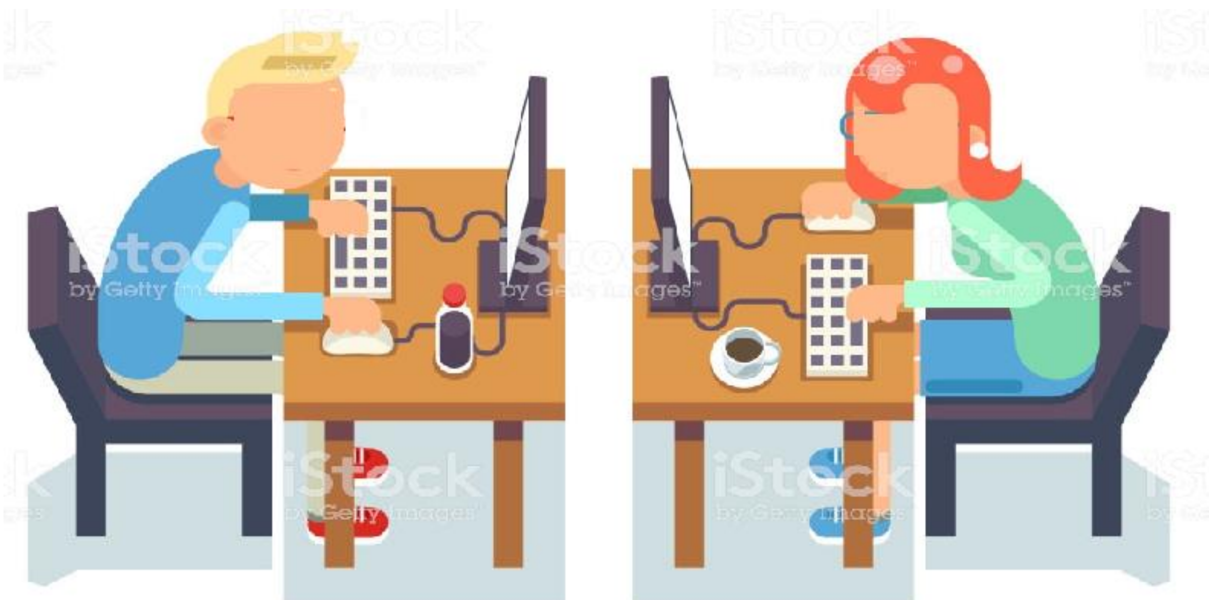
College of Science

King Saud University

Tutorial

STATISTICAL PACKAGES

STAT 328



Editor by : kholoud Basalim

Course outline

STAT 328 (Statistical Packages) 3 credit hours

Course Scope Contents:

Using program code in a statistical software package

(Excel – Minitab – SPSS - R)

to write a program for data and statistical analysis. Topics include creating and managing data files, graphical presentation - and Monte Carlo simulations.

#	Topics Covered
1	Introduction to statistical analysis using excel
2	Some mathematical, statistical and logical functions in excel
3	Descriptive statistics using excel
4	Statistical tests using excel
5	Correlation and regression using excel
6	Introduction to Minitab- Descriptive statistics using Minitab
7	Statistical distributions in Minitab
8	Statistical tests using Minitab
9	Correlation and regression using Minitab
10	Introduction to SPSS
11	Descriptive statistics using SPSS
12	Statistical tests using SPSS
13	Correlation and regression using SPSS
14	Introduction to R
15	Statistical and mathematical functions in R
16	Descriptive statistics using R
17	Statistical distributions in R
18	Statistical tests using R
19	Correlation and regression using R
20	Programming and simulation in R

Excel

MATHEMATICAL FUNCTIONS

Write the commands of the following :

Absolute value	|-4| =4	=abs(-4)	=ABS(-4) 4
Combination	$\binom{10}{6}$	=combin(10 ; 6)	=COMBIN(10;6) 210
Exponential function	$e^{-1.6}$	=exp(-1.6)	=EXP(-1.6) 0.201897
Factorial	$3!$	=fact(3)	=FACT(3) 6
Natural logarithm	$\ln 23$	=ln(23)	=LN(23) 3.135494
Logarithm with respect to any base	$\log_9 4$	=log(4 ; 9)	=LOG(4;9) 0.63093
Logarithm with respect to base 10	$\log 12$	=log10(12)	=LOG10(12) 1.079181
Square root	$\sqrt{4}$	=sqrt(4)	=SQRT(4) 2
Summation	Summation of : 450 ,11, 20 , 5	=sum(450;11;20;5)	=SUM(450;11;20;5) 486
Permutations	$10P6$	=permut(10 ;6)	=PERMUT(10;6) 151200
Product	Product of : 450 , 11 ,20 , 5	=product(450 ; 11; 20; 5)	=PRODUCT(450;11;20;5) 495000
Powers	10^{-4}	=power(10 ; -4)	=POWER(10;-4) 0.0001

CONDITIONAL FUNCTION (IF) AND COUNT CONDITIONAL FUNCTION

We have marks of 14 students :

73 45 32 85 98 78 82 87 60 25 64 72 12 90

1- Print student case being successful if (mark ≥60) and being a failure if (mark< 60).

	A	B	C
1	marks		
2	73	=if(A2>=60;"S";"F")	
3	45		
4	32		
5	85		
6	98		
7	78		
8	82		
9	87		
10	60		
11	25		
12	64		
13	72		
14	12		
15	90		
16			

	A	B	C
1	marks		
2	73	S	
3	45	F	
4	32	F	
5	85	S	
6	98	S	
7	78	S	
8	82	S	
9	87	S	
10	60	S	
11	25	F	
12	64	S	
13	72	S	
14	12	F	
15	90	S	

2- How many successful students ?

	A	B	C
1	marks		
2	73	S	
3	45	F	
4	32	F	
5	85	S	
6	98	S	
7	78	S	
8	82	S	
9	87	S	
10	60	S	
11	25	F	
12	64	S	
13	72	S	
14	12	F	
15	90	S	
16			
17			
18		=countif(B2:B15;"S")	

	A	B	C
1	marks		
2	73	S	
3	45	F	
4	32	F	
5	85	S	
6	98	S	
7	78	S	
8	82	S	
9	87	S	
10	60	S	
11	25	F	
12	64	S	
13	72	S	
14	12	F	
15	90	S	
16			
17			
18		10	
19			

3- How many students whose marks are less than or equal to 80 ?

	A	B	C	D
1	marks			
2	73	S		
3	45	F		
4	32	F		
5	85	S		
6	98	S		
7	78	S		
8	82	S		
9	87	S		
10	60	S		
11	25	F		
12	64	S		
13	72	S		
14	12	F		
15	90	S		
16				
17				
18		10		
19				
20		=countif(A2:A15;"<=80")		

	A	B	C
1	marks		
2	73	S	
3	45	F	
4	32	F	
5	85	S	
6	98	S	
7	78	S	
8	82	S	
9	87	S	
10	60	S	
11	25	F	
12	64	S	
13	72	S	
14	12	F	
15	90	S	
16			
17			
18		10	
19			
20		9	

DESCRIPTIVE STATISTICS

We have students' weights as follows:

44 , 40 , 42 , 48 , 46 , 44.

Find:

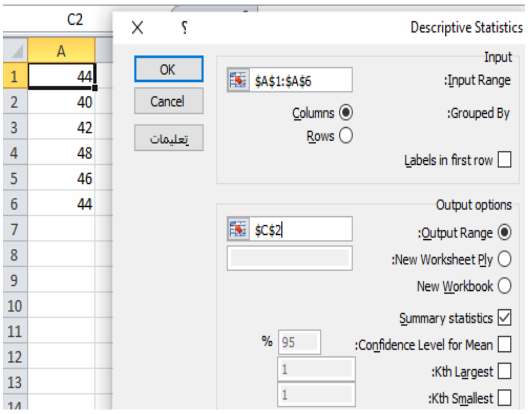
Mean=44	AVERAGE(C2:C7)
Median=44	MEDIAN(C2:C7)
Mode=44	MODE.SNGL(C2:C7)
Sample standard deviation=2.828	STDEV.S(C2:C7)
Sample variance=8	VAR.S(C2:C7)
Kurtosis=-0.3	KURT(C2:C7)
Skewness=4.996E-17	SKEW(C2:C7)
Minimum=40	MIN(C2:C7)
Maximum=48	MAX(C2:C7)
Range=8	MAX(C2:C7)-MIN(C2:C7)
Count=6	COUNT(C2:C7)
Coefficient of variation=6.428%	STDEV.S(C2:C7)/AVERAGE(C2:C7)*100

* Range= Maximum-Minimum

** Coefficient of variation= $\frac{\text{Sample standard deviation}}{\text{Mean}} \times 10$

other ways :

Data - Data Analysis - Descriptive Statistics



PROBABILITY DISTRIBUTION FUNCTIONS

Discrete Distribution :

1-Binomial Distribution :

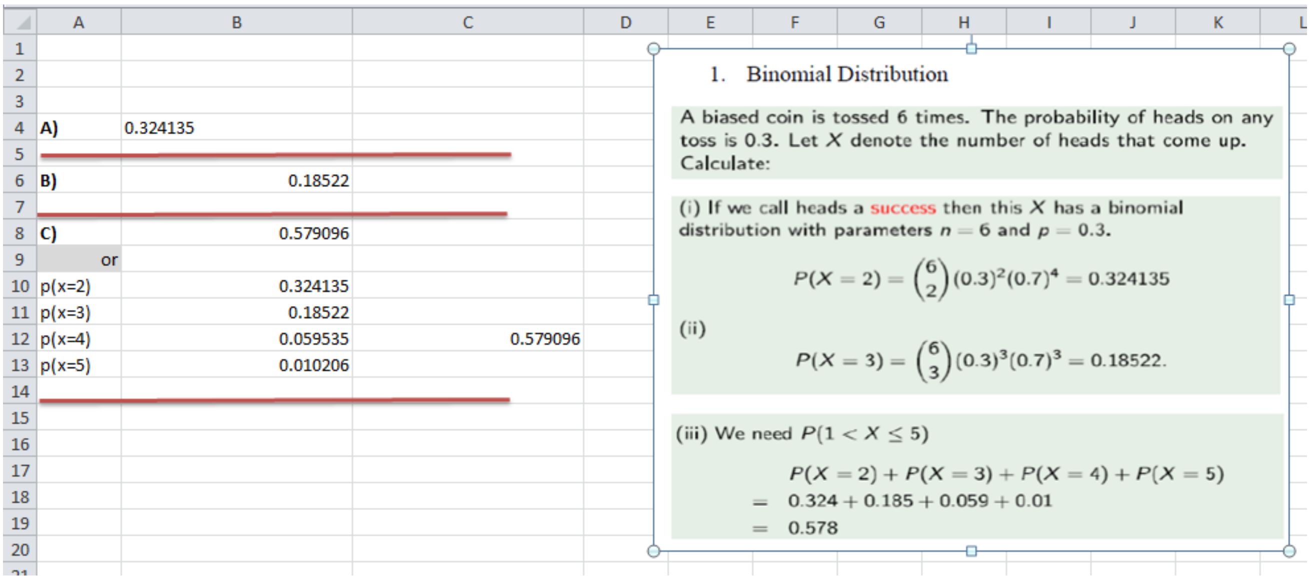
A biased coin is tossed 6 times. The probability of heads on any toss is 0.3. Let X denote the number of heads that come up. Calculate

X~Bin(n=6 , p=0.3)

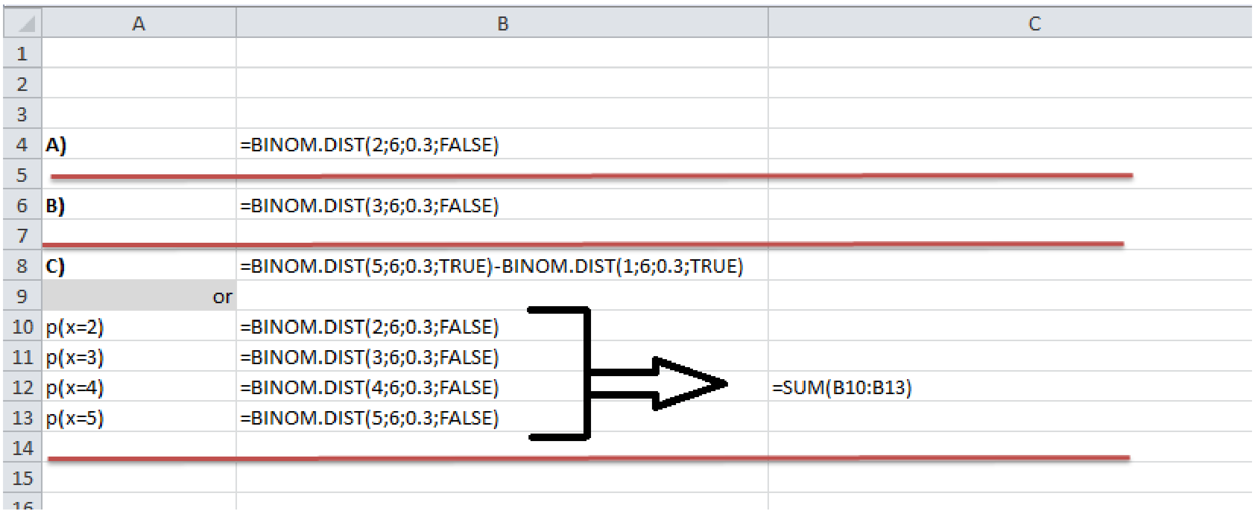
a) P(X = 2)

b) P(X = 3)

c) P(1 < X ≤5).



	A	B	C
1			
2			
3			
4	A)	0.324135	
5			
6	B)	0.18522	
7			
8	C)	0.579096	
9	or		
10	p(x=2)	0.324135	
11	p(x=3)	0.18522	
12	p(x=4)	0.059535	0.579096
13	p(x=5)	0.010206	



	A	B	C
1			
2			
3			
4	A)	=BINOM.DIST(2;6;0.3;FALSE)	
5			
6	B)	=BINOM.DIST(3;6;0.3;FALSE)	
7			
8	C)	=BINOM.DIST(5;6;0.3;TRUE)-BINOM.DIST(1;6;0.3;TRUE)	
9	or		
10	p(x=2)	=BINOM.DIST(2;6;0.3;FALSE)	
11	p(x=3)	=BINOM.DIST(3;6;0.3;FALSE)	
12	p(x=4)	=BINOM.DIST(4;6;0.3;FALSE)	=SUM(B10:B13)
13	p(x=5)	=BINOM.DIST(5;6;0.3;FALSE)	
14			
15			

2.Poisson Distribution :

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

a) What is the probability of observing 4 births in a given hour at the hospital?

b) What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

X~poisson(λ=1.8)

	A	B	C	D
1				
2	A)	0.072302		
3				
4				
5	B)	$p(x\geq 2)=1-p(x<2)=1-p(x\leq 1)$		
6				
7				0.537163

2. Poisson Distribution

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability of observing 4 births in a given hour at the hospital?

Let X = No. of births in a given hour

(i) Events occur randomly
(ii) Mean rate $\lambda = 1.8$ $\Rightarrow X \sim \text{Po}(1.8)$

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X=4) = e^{-1.8}\frac{1.8^4}{4!} = 0.0723$$

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want $P(X \geq 2) = P(X=2) + P(X=3) + \ldots$

i.e. an infinite number of probabilities to calculate

but

$$\begin{aligned} P(X \geq 2) &= P(X=2) + P(X=3) + \ldots \\ &= 1 - P(X<2) \\ &= 1 - (P(X=0) + P(X=1)) \\ &= 1 - (e^{-1.8}\frac{1.8^0}{0!} + e^{-1.8}\frac{1.8^1}{1!}) \\ &= 1 - (0.16529 + 0.29753) \\ &= 0.537 \end{aligned}$$

	A	B	C	D
1				
2	A)	=POISSON.DIST(4;1.8;FALSE)		
3				
4				
5	B)	$p(x\geq 2)=1-p(x<2)=1-p(x\leq 1)$		
6				
7				=1 -POISSON.DIST(1;1.8;TRUE)
8				
9				
10				

Continuous Distribution :

1. Exponential Distribution :

If X~exp(λ=1/10) , Find P(X>7)

	A	B
1		
2		
3	A)	p(x > 7) = 1- p(x < 7)
4		
5		=1-EXPON.DIST(7;(1/10);TRUE)
6		
7		
8		
9		
10		

	A	B	C
1			
2			
3	A)	p(x > 7) = 1- p(x < 7)	
4			
5		0.496585304	
6			
7			

2. Normal Distribution :

If x ∽ N(μ = 20 , σ = 3) . Find :

A) P(X ≤ 25) = P(X < 25)

B) P(X ≤ x_0) =0.55 , x_0 =

	A	B	C	D	E	F	G	H	I
1									
2					If x ∽ N(μ = 20 , σ = 3) . Find :				
3	A)	0.95221							
4					A) P(X ≤ 25) = P(X < 25)				
5									
6	C)	20.37698			C) P(X ≤ x_0) =0.55 , x_0 =				
7									

	A	B
1		
2		
3	A)	=NORM.DIST(25;20;3;TRUE)
4		
5		
6	C)	=NORM.INV(0.55;20;3)
7		

If z ∽ N(μ = 0 , σ = 1) . Find :

A)P(Z ≤ 1.78) = P(Z < 1.78) =

B) P(Z ≤z_0)=0.55 , z_0 =

	A	B	C	D	E	F	G
1							
2							
3	A)	0.96246202					
4							
5							
6	B)	0.125661347					
7							

If z ∽ N(μ = 0 , σ = 1) . Find :

A)P(Z ≤ 1.78) = P(Z < 1.78) =

B) P(Z ≤z_0)=0.55 , z_0 =

	A	B
1		
2		
3	A)	=NORM.S.DIST(1.78;TRUE)
4		
5		
6	B)	=NORM.S.INV(0.55)

3.Student's t-distribution :

Find : A) $t_{0.025}$ where $v = df = 14$ ⟺ P(T<t_0) =0.025 df=14

B) $t_{0.01}$ where $v = df = 10$

C) $t_{0.995}$ where $v = df = 7$

	A	B	C	D	E	F	G	H
1								
2								
3	A)	-2.144786688		Find :	A) $t_{0.025}$	*where*	$v = df = 14$	
4								
5	B)	-2.763769458			B) $t_{0.01}$	*where*	$v = df = 10$	
6								
7	C)	3.499483297			C) $t_{0.995}$	*where*	$v = df = 7$	
8								

	A	B
1		
2		
3	A)	=T.INV(0.025;14)
4		
5	B)	=T.INV(0.01;10)
6		
7	C)	=T.INV(0.995;7)
8		

4- chi-square distribution:

Find : $\chi^2_{0.995}$ **when** $\nu = 19$

	A	B	C	D
1				
2	38.58226		: $\chi^2_{0.995}$ when $\nu = 19$	
3				

	A	B	C
1			
2	=CHISQ.INV(0.995;19)	: $\chi^2_{0.995}$ when $\nu = 19$	
3			

5- F distribution:

Find : $F_{0.995,15,22}$ ⟷ $P(F < f) = 0.995\,,\ df_1 = 15\ ,df_2 = 22$

6	
7	=F.INV(0.995;15;22)
8	

3.359998884	: $F_{0.995,15,22}$

∗ Find the value of a : $P(X \le a)$ or $P(X = a)$

short
= name of distribution . dist (X , parameter of distribution , True :if calculate Cumulative Distribution ≤
False : if calculate Probabilistic Distribution =

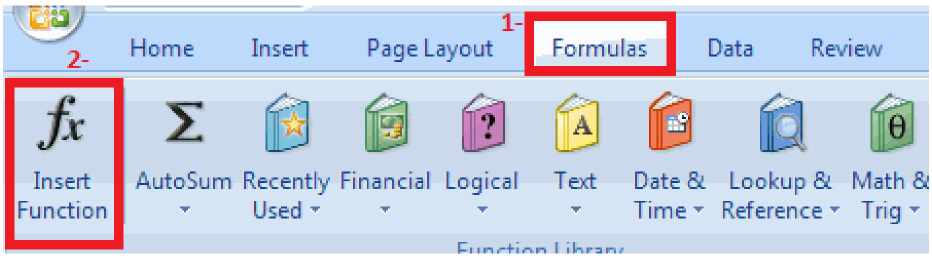
∗ Find the value of k : $P(X \le k) = b$

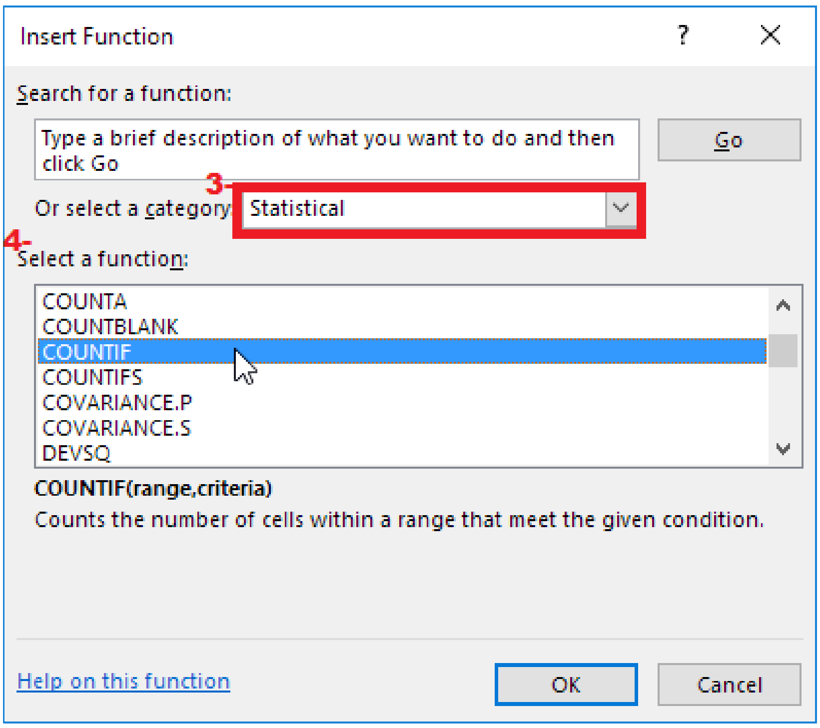
short
= name of distribution . inv (probability , parameter of distribution)

other ways:

you can find the functions of distribution from:

Formulas – Insert function



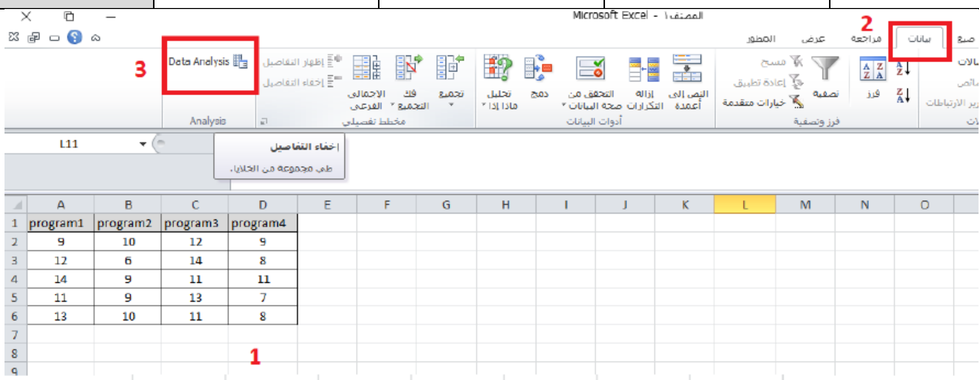


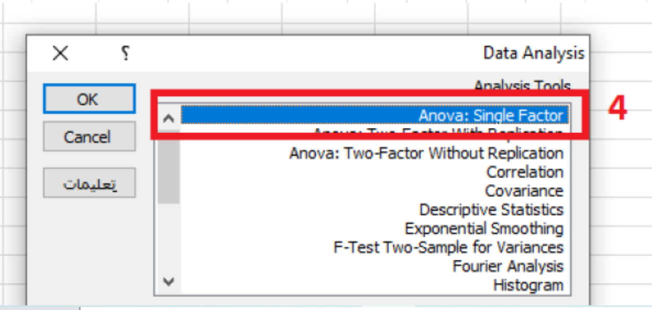
HYPOTHESIS TESTING STATISTICS AND CONFIDENCES INTERVAL

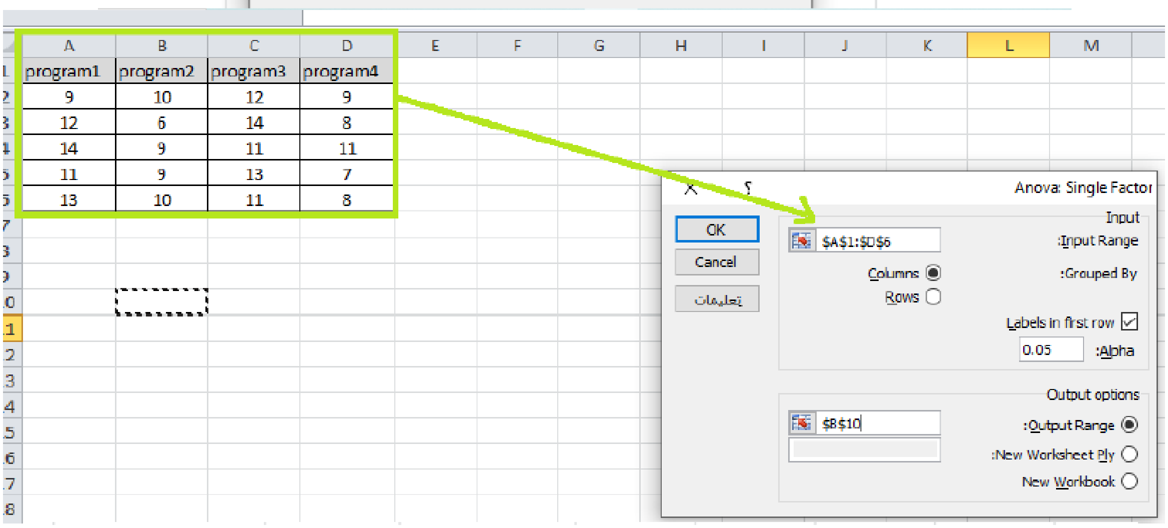
1-One way AVOVA :

A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results:

observation	Program 1	Program 2	Program 3	Program 4
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8







Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
program1	5	59	11.8	3.7		
program2	5	44	8.8	2.7		
program3	5	61	12.2	1.7		
program4	5	43	8.6	2.3		
ANOVA						
Source of Variatior	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	54.95	3	18.31667	7.044872	0.003113	3.238872
Within Groups	41.6	16	2.6			
Total	96.55	19				

1) **Hypotheses testing:**

$$H_0: \mu_2 = \mu_2 = \mu_2 = \mu_2 \qquad VS \qquad H_1: at\ least\ one\ of\ means\ is\ different$$

2) **Test statistic:**

F = 7.04487

3) **Critical region:**

$$F_{crit} = 3.23887$$

4) **P-value = 0.003113** < α

Reject H_0 $if\ F > F_{crit}$

Or

p-value ≤ α

Decision:

we reject the null hypothesis. There are difference in the means

2-Two-Sample t Statistic :

Q: The phosphorus content was measured for independent samples of skim and whole:

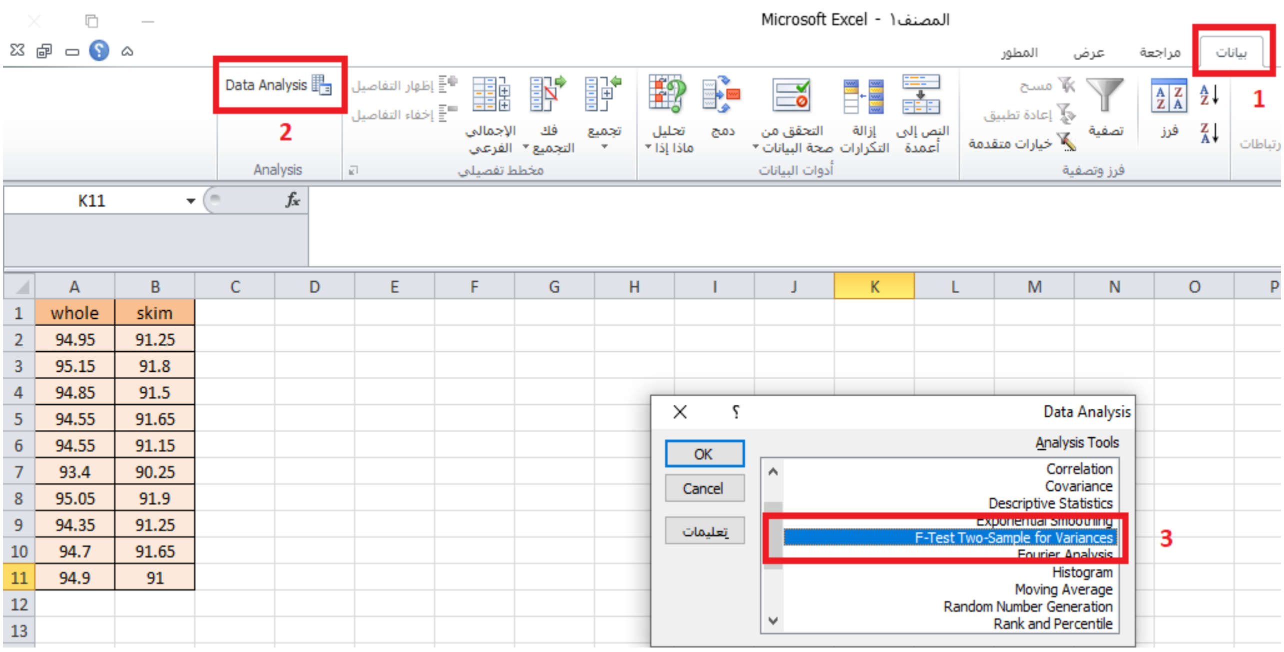
Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

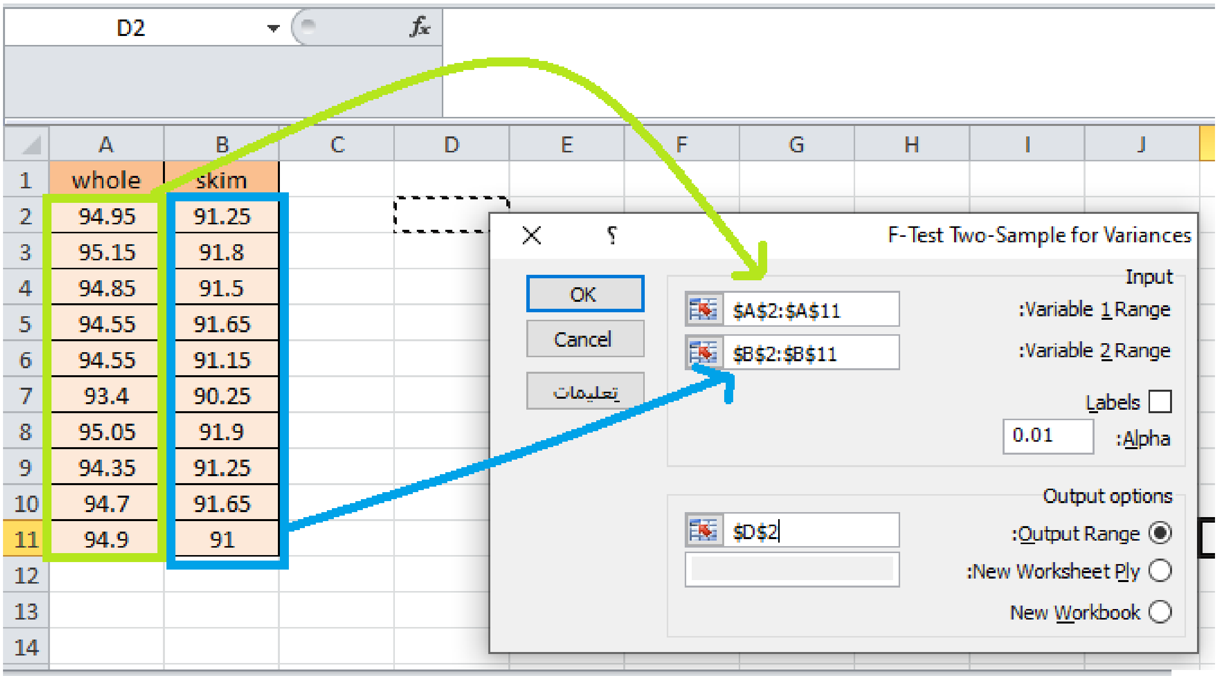
Assuming normal populations .

a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**

1-Test for equality of variance :

Data ⟶ Data Analysis ⟶ F –test two –sample for variance





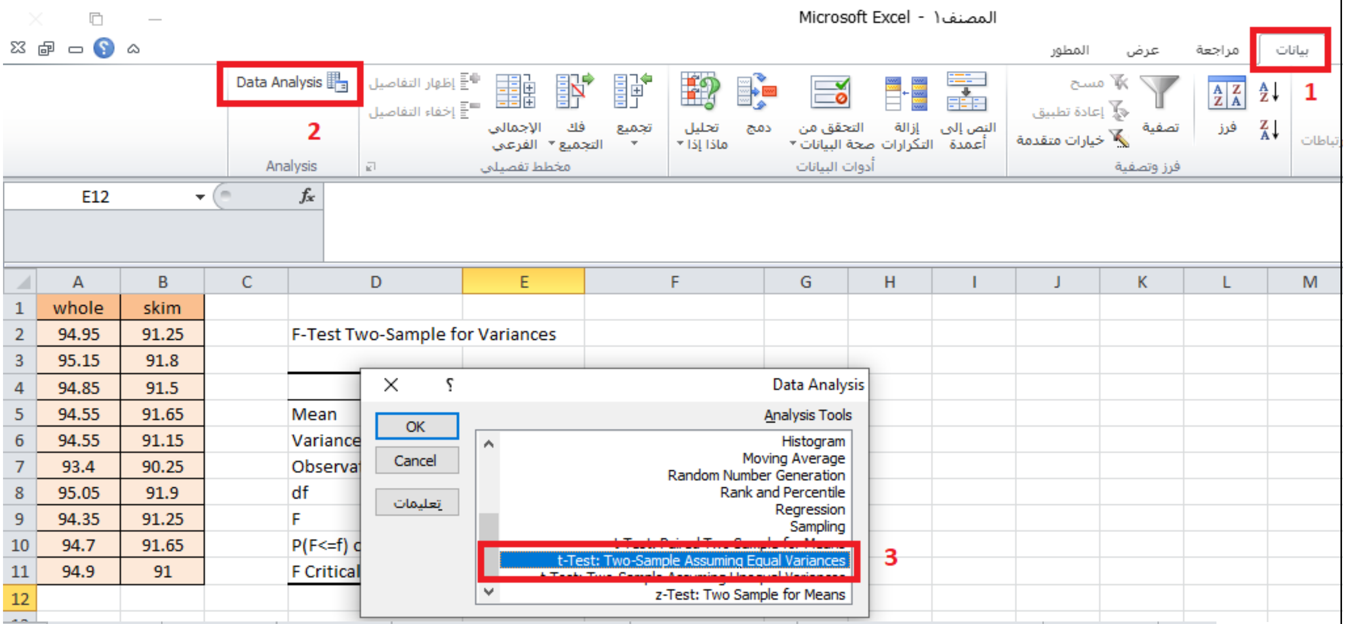
	A	B	C	D	E	F
1	whole	skim				
2	94.95	91.25		F-Test Two-Sample for Variances		
3	95.15	91.8				
4	94.85	91.5			*Variable 1*	*Variable 2*
5	94.55	91.65		Mean	94.645	91.34
6	94.55	91.15		Variance	0.253027778	0.233222222
7	93.4	90.25		Observations	10	10
8	95.05	91.9		df	9	9
9	94.35	91.25		F	1.084921391	
10	94.7	91.65		P(F<=f) one-tail	0.452676255	
11	94.9	91		F Critical one-tail	5.351128861	
12						

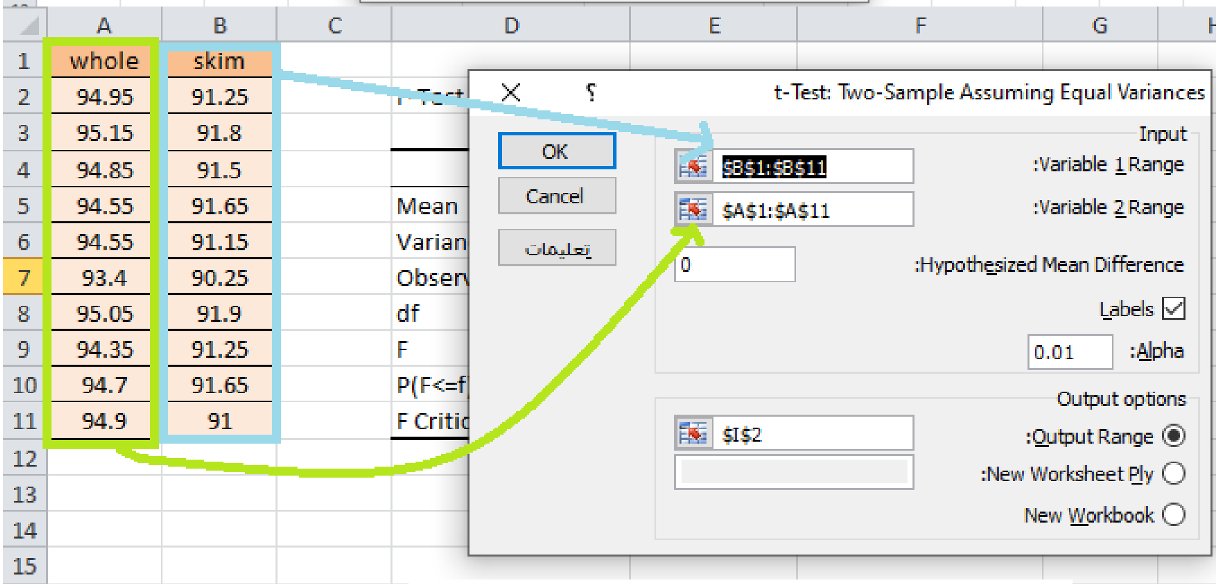
Hypothesis: $H_0: \sigma_1^{\ 2} = \sigma_2^{\ 2}$ $\quad VS \quad$ $H_1: \sigma_1^{\ 2} \neq \sigma_2^{\ 2}$

Conclusion: As F ≯ F Critical one-tail, we fail reject the null hypothesis. This is the case, 1.0849 ≯3.1789. Therefore, we fail to reject the null hypothesis.
<u>The variances of the two populations are equal</u> .

2-T Test two samples for mean assuming Equal Variance :

Data → Data Analysis → T Test: Two -samples Assuming Equal Variance





	A	B
1	whole	skim
2	94.95	91.25
3	95.15	91.8
4	94.85	91.5
5	94.55	91.65
6	94.55	91.15
7	93.4	90.25
8	95.05	91.9
9	94.35	91.25
10	94.7	91.65
11	94.9	91

$$H_0: \mu_{Skim} - \mu_{Whole} \geq 0 \quad VS \quad H_1: \mu_{Skim} - \mu_{Whole} < 0$$

t-Test: Two-Sample Assuming Equal Variances		
	skim	*whole*
Mean	91.34	94.645
Variance	0.233222222	0.253027778
Observations	10	10
Pooled Variance	0.243125	
Hypothesized Mea	0	
df	18	
t Stat	-14.98793002	
P(T<=t) one-tail	6.53252E-12	
t Critical one-tail	2.55237963	
P(T<=t) two-tail	1.3065E-11	
t Critical two-tail	2.878440473	

1-Hypothesis:

$$H_0: \mu_{Skim} \geq \mu_{Whole} \quad VS \quad H_1: \mu_{Skim} < \mu_{Whole}$$

$$H_0: \mu_{Skim} - \mu_{Whole} \geq 0 \quad VS \quad H_1: \mu_{Skim} - \mu_{Whole} < 0$$

2- Test statistic : T= - 14.98

3- T critical value (one tail) = -2.55238

4- Conclusion:

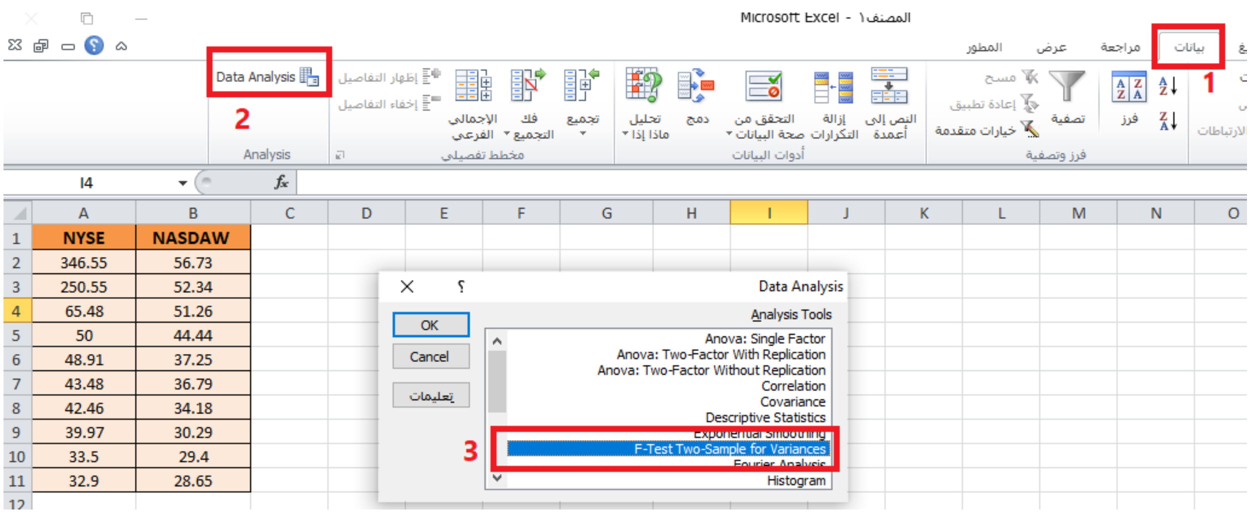
We do a one-tailed test . if t Stat < -t Critical one-tail, we reject the null hypothesis. As -14.9879 < -2.55238 (p-value=0.00000653<α=0.01) . Therefore, we reject the null hypothesis

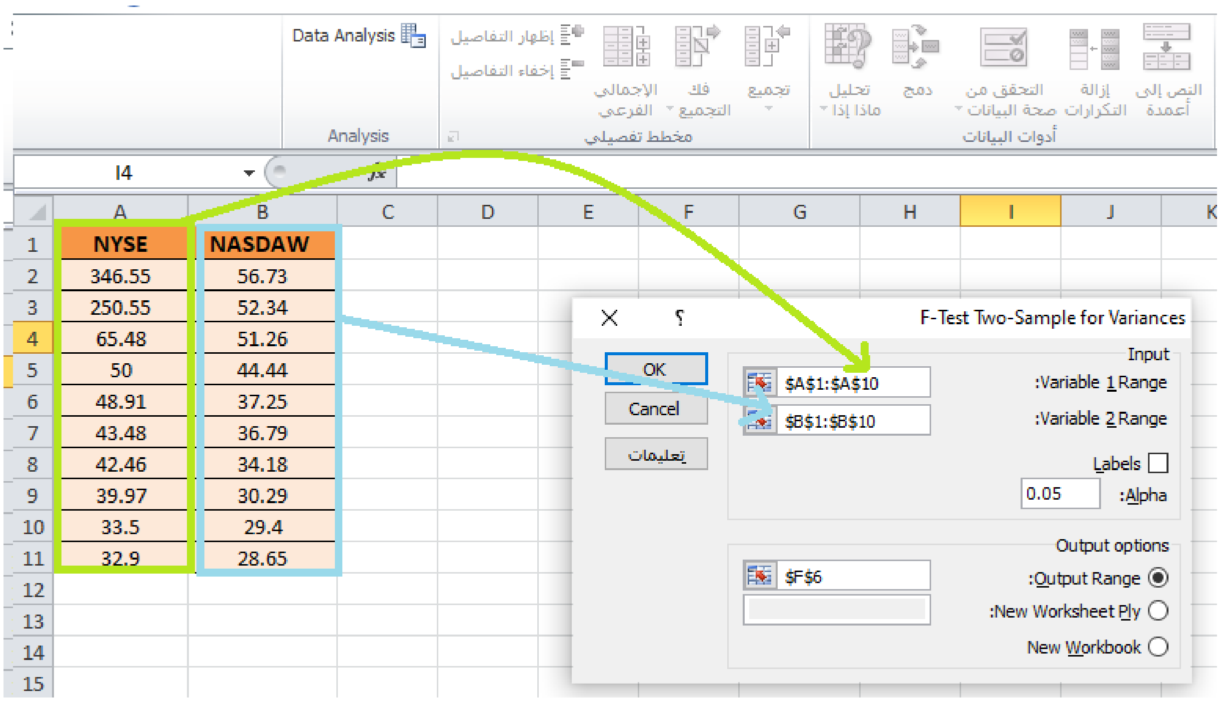
Q : The example below gives the Dividend Yields for the top ten NYSE and NASDAW stocks. Use the t-test tool to determine whether there is any indication of a difference between the means of the two different populations. α =0.05

NYSE	NASDAW
346.55	56.73
250.55	52.34
65.48	51.26
50	44.44
48.91	37.25
43.48	36.79
42.46	34.18
39.97	30.29
33.5	29.4
32.9	28.65

1-Test for equality of variance :

Data → Data Analysis → F –test two –sample for variance





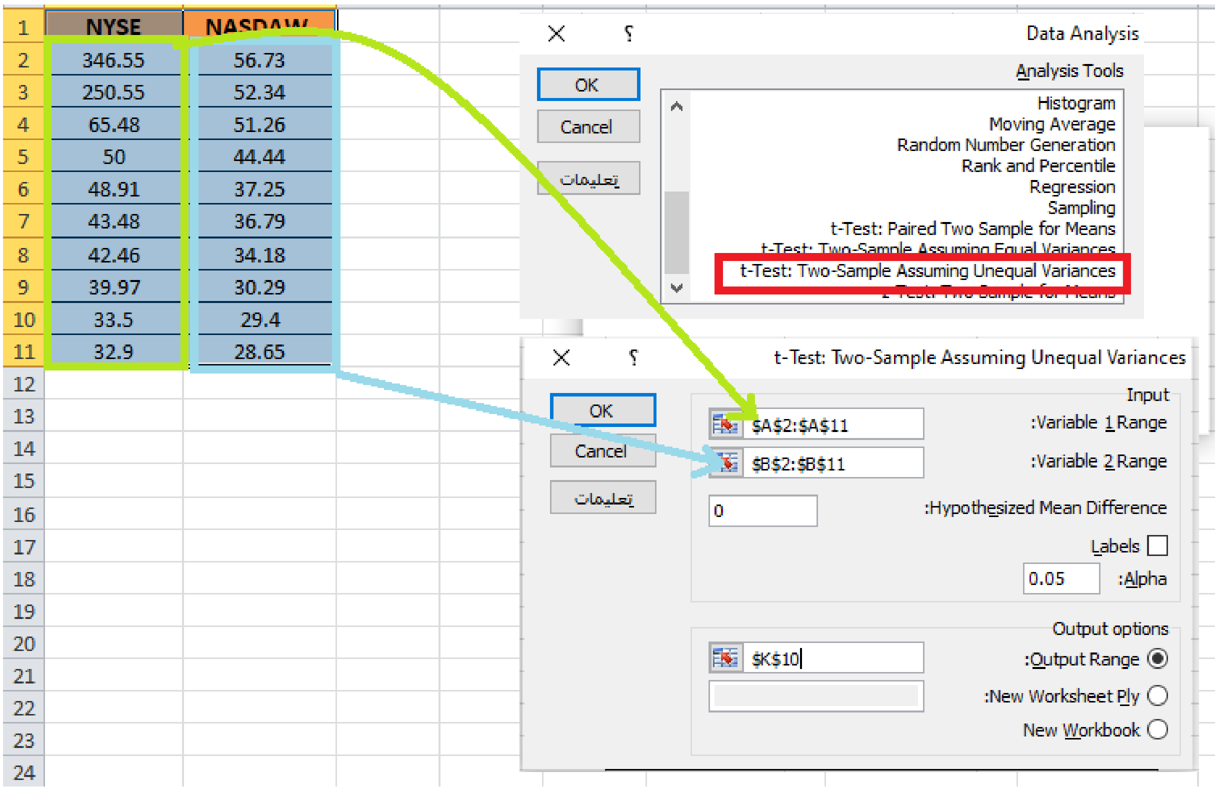
F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	95.38	40.133
Variance	12063.80027	107.2998233
Observations	10	10
df	9	9
F	112.4307561	
P(F<=f) one-tail	3.6621E-08	
F Critical one-tail	3.178893104	

Hypothesis: $H_0: {\sigma_1}^2 = {\sigma_2}^2$ $\quad VS \quad$ $H_1: {\sigma_1}^2 \neq {\sigma_2}^2$

Conclusion: As F > F Critical one-tail, we reject the null hypothesis. Therefore, reject the null hypothesis. The variances of the two populations are unequal .

2-T Test two samples for mean assuming Unequal Variance :

Data → Data Analysis → T Test: Two -samples Assuming Unequal Variance



t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	95.38	40.133
Variance	12063.80027	107.2998233
Observations	10	10
Hypothesized Mean Difference	0	
df	9	
t Stat	1.583593765	
P(T<=t) one-tail	0.073873163	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.147746326	
t Critical two-tail	2.262157163	

1-Hypothesis:

$$H_0: \mu_1 = \mu_2 \quad VS \quad H_1: \mu_1 \neq \mu_2$$

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

2- Test statistic : T= 1.58359

3- T critical value (two tailed) = $\pm$ 2.26215

4- Conclusion:

We do a two-tailed test (inequality). if t Stat < -t Critical two-tail or t Stat > t Critical two-tail, we reject the null hypothesis. This is not the case, -2.26215< 1.58359 < 2.26215. Therefore, we do not reject the null hypothesis (p-value=0.1477$\nless \alpha$=0.05)
there is no significant difference in the means of each sample.

3- paired test :

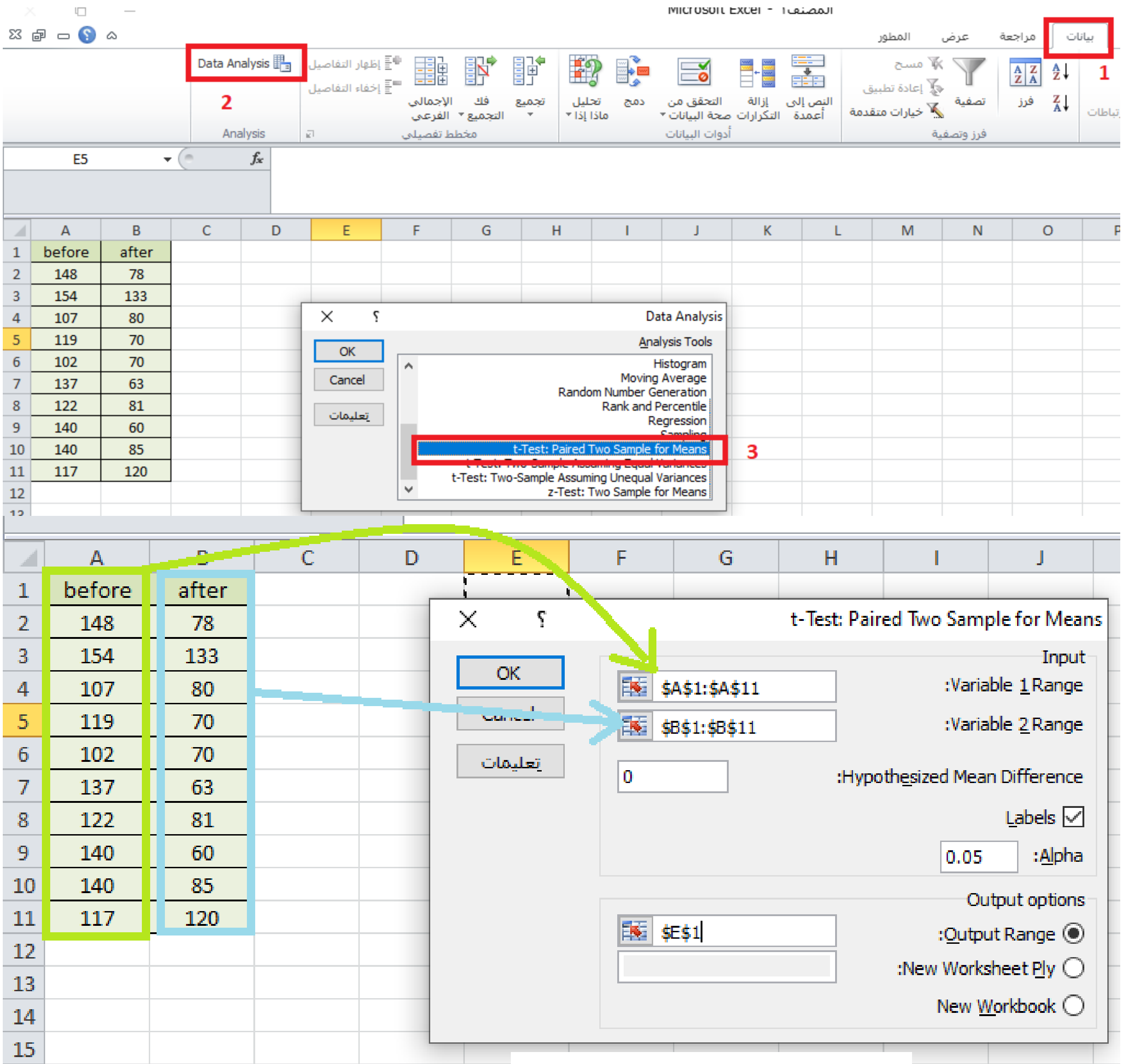
Q : In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

Before surgery (X)	**148**	**154**	**107**	**119**	**102**	**137**	**122**	**140**	**140**	**117**
After surgery (Y)	**78**	**133**	**80**	**70**	**70**	**63**	**81**	**60**	**85**	**120**

We assume that the data comes from normal distribution. Find :

1- **Test whether the data provide sufficient evidence to indicate a difference in the average weight before and after surgery. ($\mu_D = 0$ versus $\mu_D \neq 0$)**

Data → Data Analysis → T Test: Paired Two –sample for Means



$H_0: \ \mu_D = 0$ vs $H_1: \ \mu_D \neq 0$

	E	F	G
	t-Test: Paired Two Sample for Means		
		before	*after*
	Mean	128.6	84
	Variance	310.7111111	574.2222222
	Observations	10	10
	Pearson Correlation	0.232799676	
	Hypothesized Mean Difference	0	
	df	9	
	t Stat	5.375965714	
	P(T<=t) one-tail	0.000223426	
	t Critical one-tail	1.833112933	
	P(T<=t) two-tail	0.000446852	
	t Critical two-tail	2.262157163	

1-Hypothesis:

$$H_0: \mu_1 = \mu_2 \quad VS \quad H_1: \mu_1 \neq \mu_2$$

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

2- Test statistic : T= 5.3759

3- T critical value (two tailed) = ± 2.26215

4- Conclusion:

We do a two-tailed test . if t Stat < -t Critical or t Stat > t Critical two-tail, we reject the null hypothesis. As 5.3759 > 2.26215 (p-value=0.00044 < α=0.05) . Therefore, we reject the null hypothesis

Regression :

Q : Ten Corvettes between 1 and 6 years old were randomly selected from last year's sales records in Virginia Beach, Virginia. The following data were obtained, where x denotes age, in years, and y denotes sales price, in hundreds of dollars.

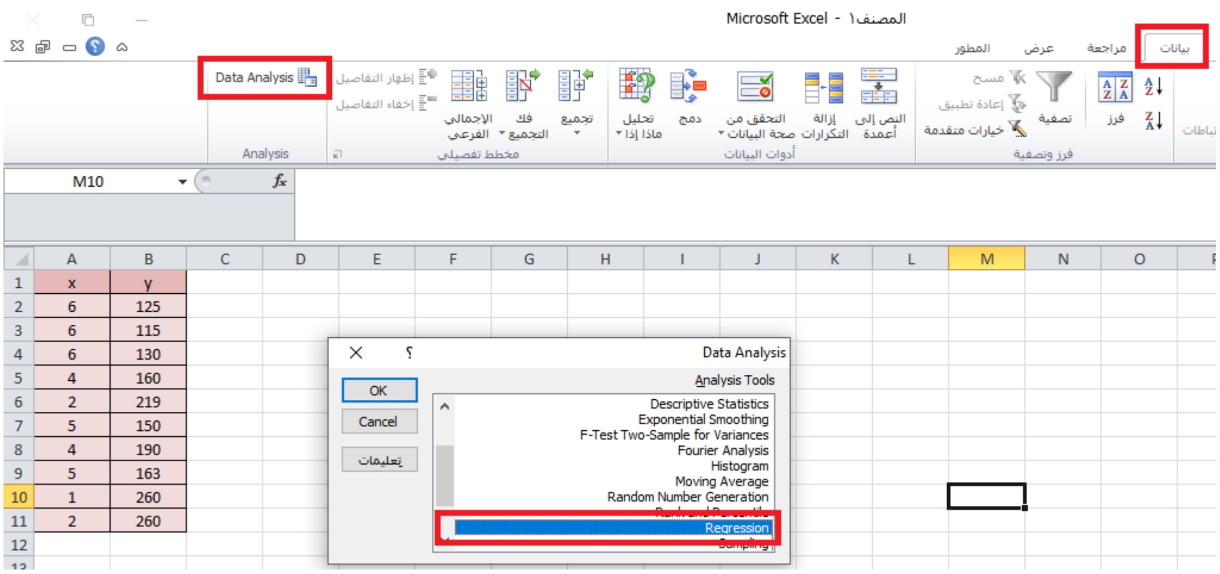
X	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

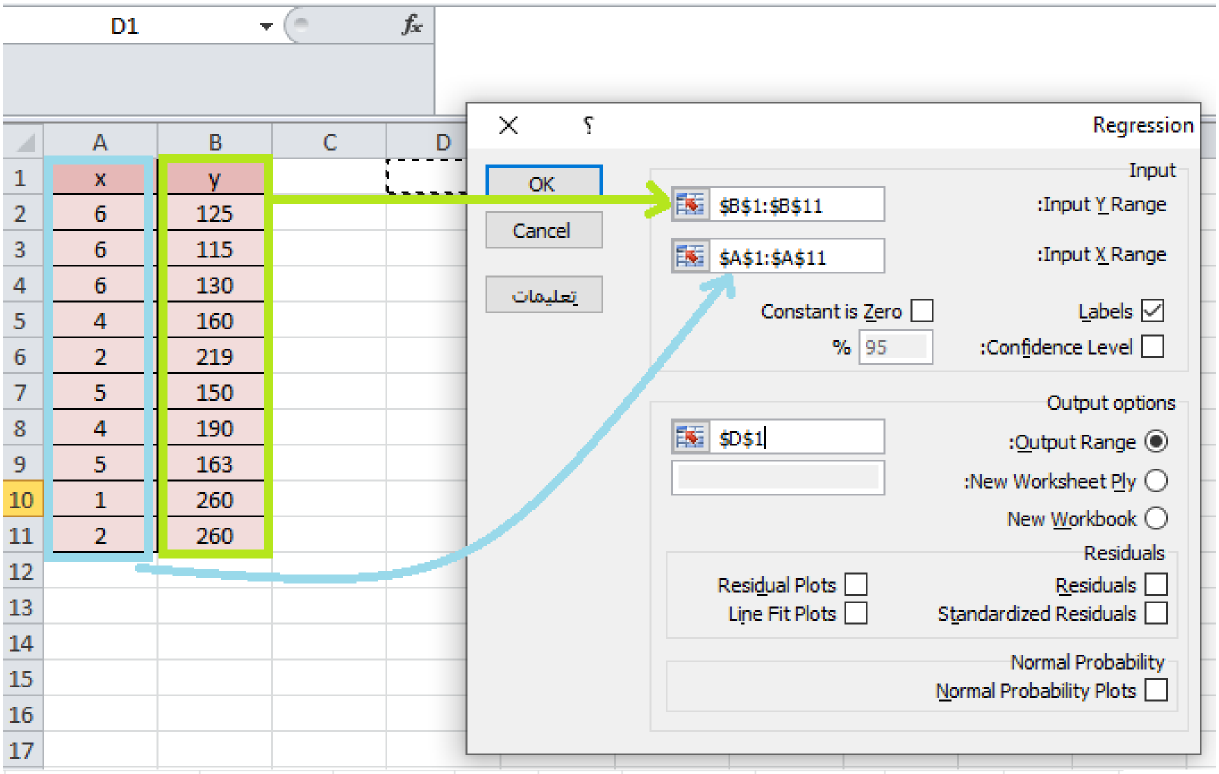
a) Determine the regression equation for the data.

b) Compute and interpret the coefficient of determination, r^2.

c) Find the predicted sales price of 4-year-old Corvette.

Data → Data Analysis → Regression





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.967871585
R Square	0.936775406
Adjusted R Square	0.928872332
Standard Error	14.24652913
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	24057.89126	24057.89	118.533	4.48427E-06
Residual	8	1623.708738	202.9636		
Total	9	25681.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	291.6019417	11.43289905	25.50551	5.98E-09	265.2376293	317.9662542	265.2376293	317.9662542
X Variable 1	-27.90291262	2.562889198	-10.8873	4.48E-06	-33.81294571	-21.99287953	-33.81294571	-21.99287953

Results:

a) **The regression equation : $\hat{y}$ = sales price = 291.6019 - 27.9029 * age .**
In other words, for increasing the age by one, the sales price decreasing by 27.9029 , while there is 291.6019 of Y does not depend on the age .

b) $r^2 = 0.9367$
The coefficient of determination is 0.9368; therefore, about 93.68% of the variation in the price data is explained by age. The regression equation appears to be very useful for making predictions since the value of r^2 is close to 1.

c) The predicted sales price is 17999.0291 dollars ($17,999.0291).

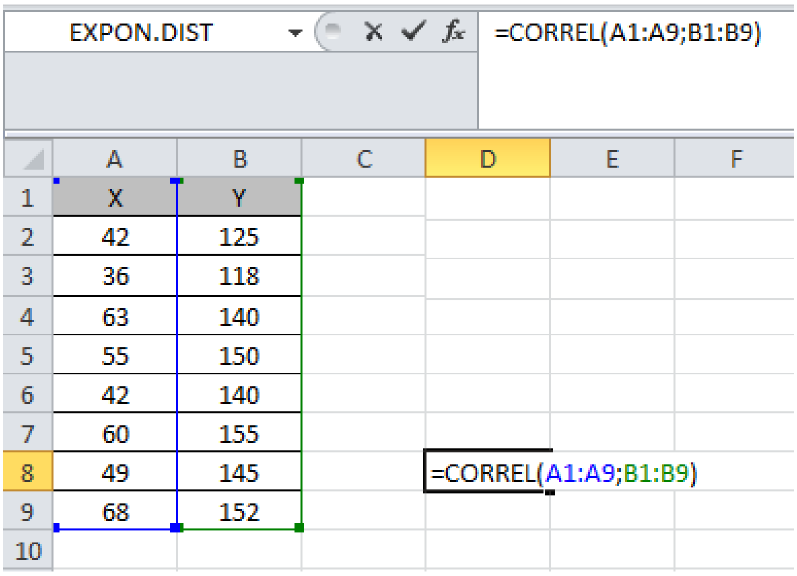
Correlation :

Q : We have the table illustrates the age X and blood pressure Y for eight female.

X	42	36	63	55	42	60	49	68
Y	125	118	140	150	140	155	145	152

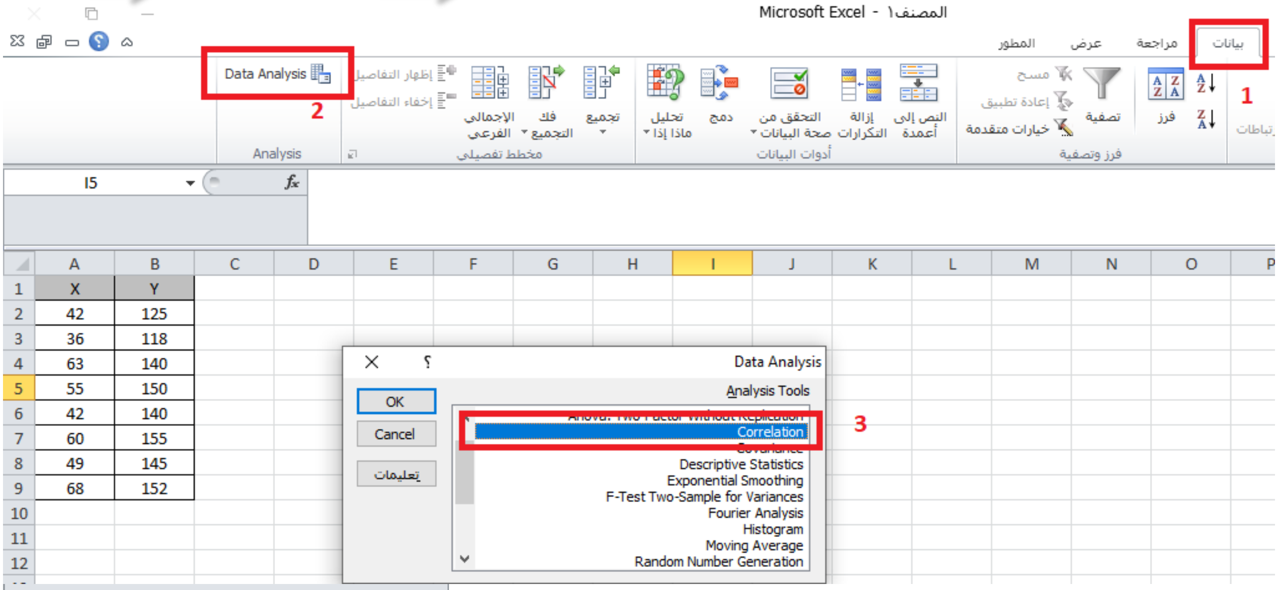
Find:

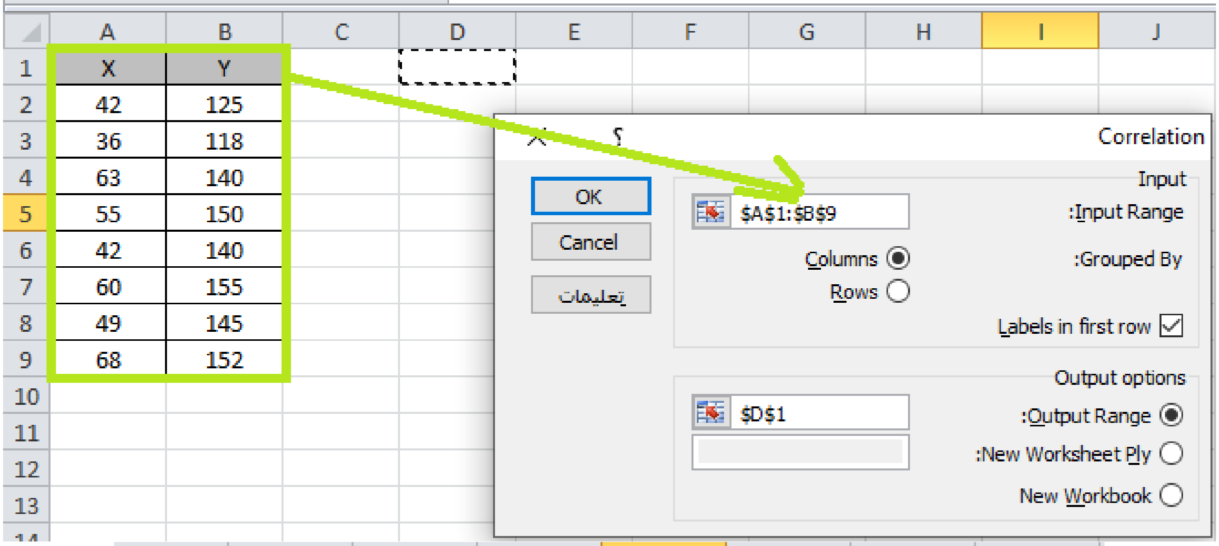
	By Excel (using (fx) and (Data Analysis))
Correlation=0.791832	CORREL(M3:M10;N3:N10)

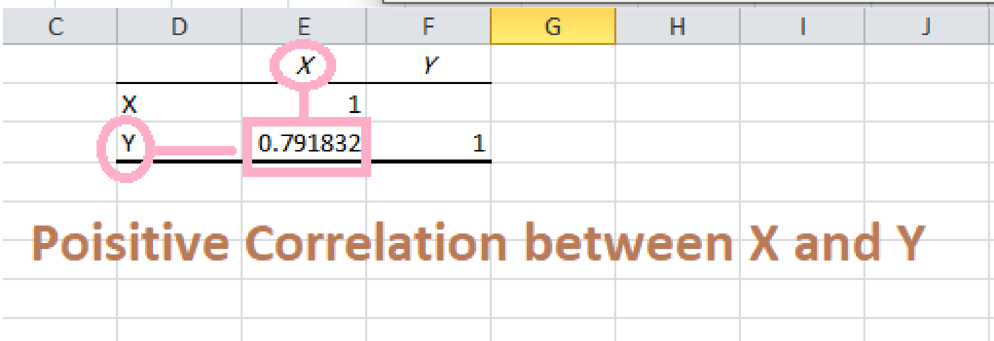


OR:

Data → Data Analysis → Correlation

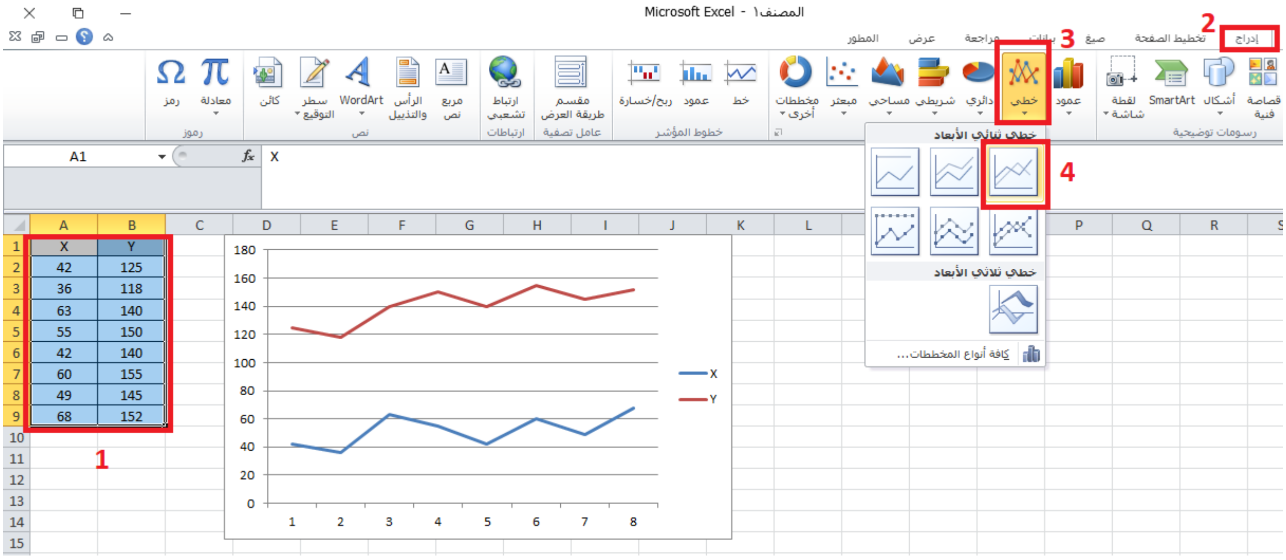






Poisitive Correlation between X and Y

The Graph showing correlation between two variables:



MATRICES

Write the commands of the following:

Addition of Matrices:

$A = \begin{bmatrix} -5 & 0 \\ 4 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 6 & -3 \\ 2 & 3 \end{bmatrix}$

A+B=

Addition of Matrices					
A=	-5	0	B=	6	-3
	4	1		2	3
A+B=	1	-3			
	6	4			

Subtract of Matrices

$C = \begin{bmatrix} 1 & 2 \\ -2 & 0 \\ -3 & -1 \end{bmatrix}$, $D = \begin{bmatrix} 1 & -1 \\ 1 & 3 \\ 2 & 3 \end{bmatrix}$,

C-D =

Subtract of Matrices					
C=	1	2	D=	1	-1
	-2	0		1	3
	-3	-1		2	3
C-D=	0	3			
	-3	-3			
	-5	-4			

Additive Inverse of Matrix

$A = \begin{matrix} 1 & 0 & 2 \\ 3 & -1 & 5 \end{matrix}$, -A=

Additive Inverse of Matrix			
A=	1	0	2
	3	-1	5
− A=	-1	0	-2
	-3	1	-5

Scalar Multiplication of Matrices

$D = \begin{bmatrix} -3 & 0 \\ 4 & 5 \end{bmatrix}$, 3D=

Scalar Multiplication of Matrices

D=	-3	0
	4	5
3D=	-9	0
	12	15

Matrix Multiplication

$$E = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \quad , \quad F = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

ExF=

Matrix Multiplication

E=	1	4	7	F=	1	4
	2	5	8		2	5
	3	6	9		3	6

E X F = =mmult(N3:P5;R3:S5)

MMULT(array1; **array2**)

Ctrl +Shift +Enter

E X F=	30	66
	36	81
	42	96

transpose of (G)

$$G = \begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$$

$G^T =$

transpose

G=	3	-1
	-5	2

=TRANSPOSE(N12:O13

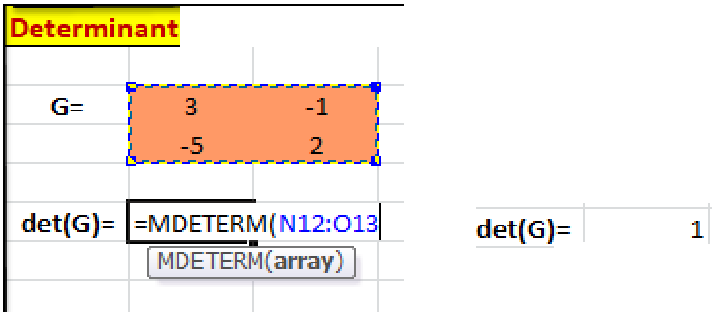
TRANSPOSE(**array**)

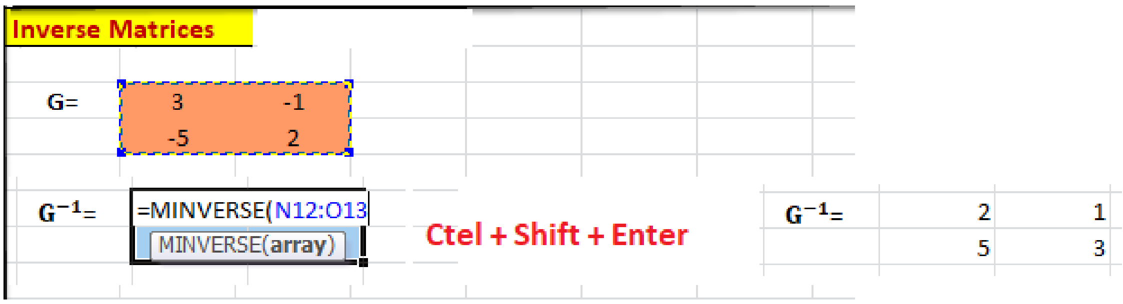
Ctel + Shift + Enter

$G^T =$	3	-5
	-1	2

Determinant and Inverse Matrices

$$G = \begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$$



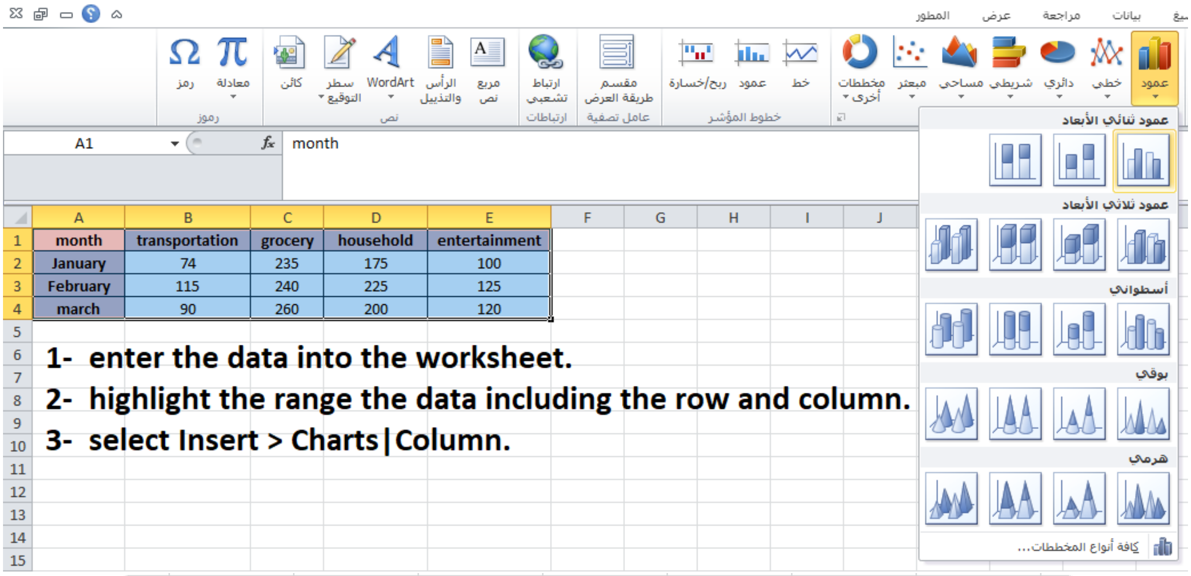


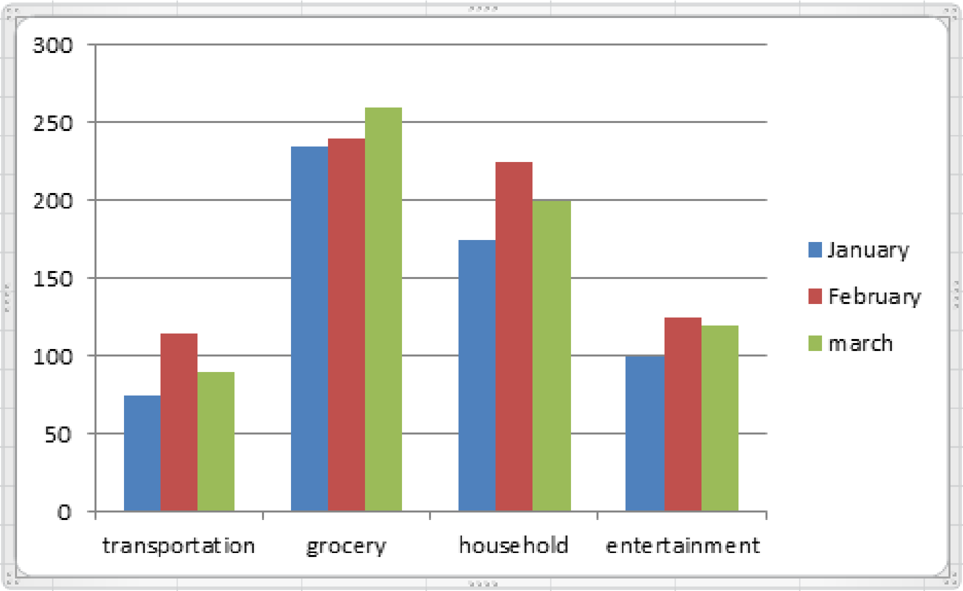
Some Statistical Charts

The following data represents the expenses in dollars by month :

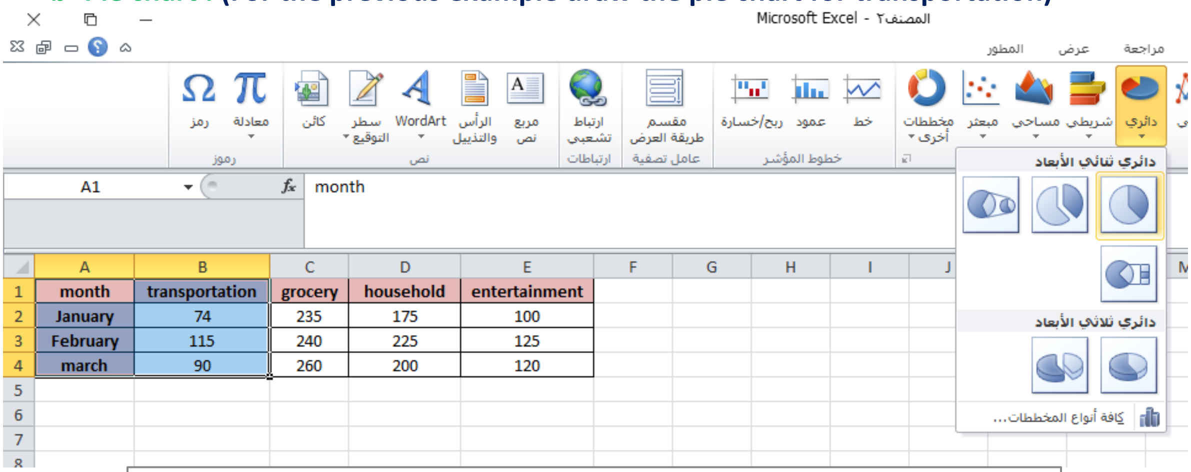
month	transportation	grocery	household	entertainment
January	74	235	175	100
February	115	240	225	125
march	90	260	200	120

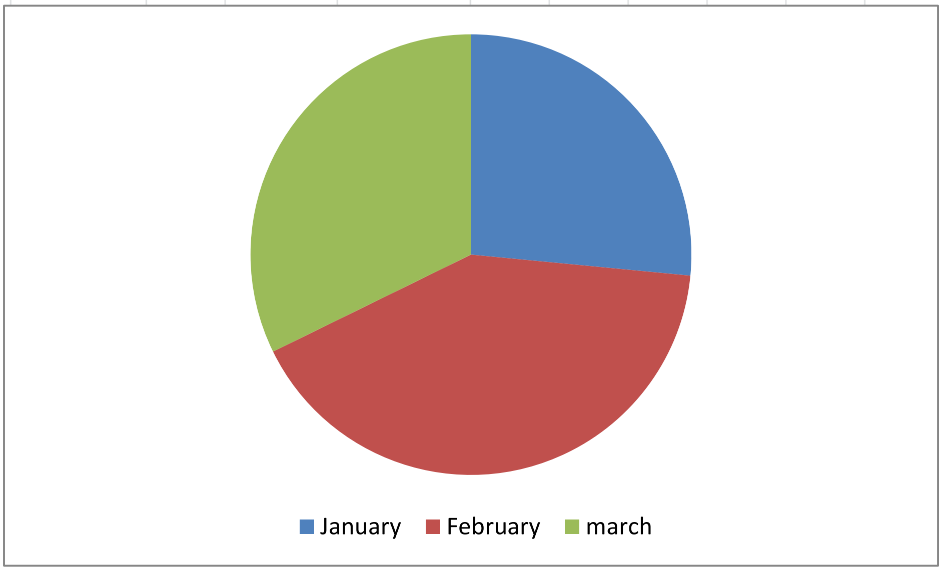
a- Bar chart :



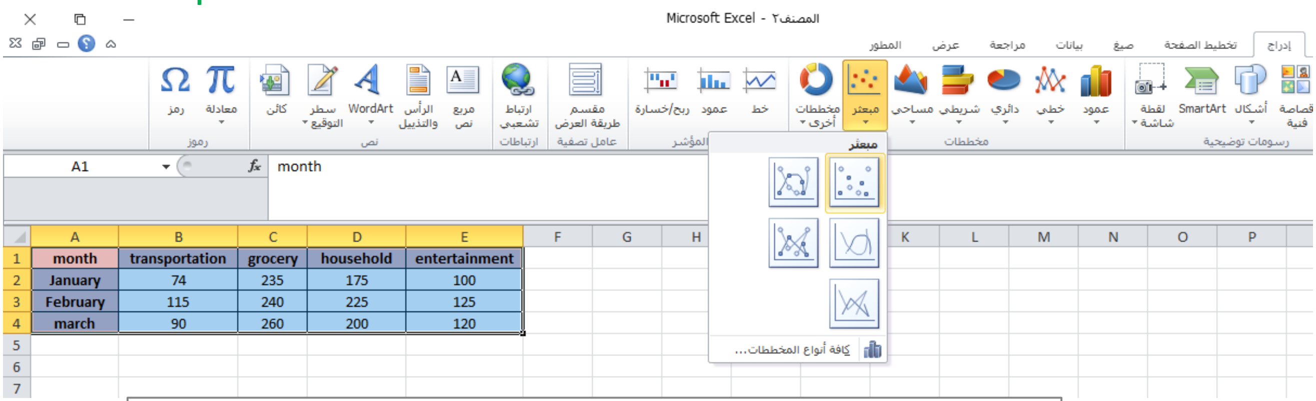


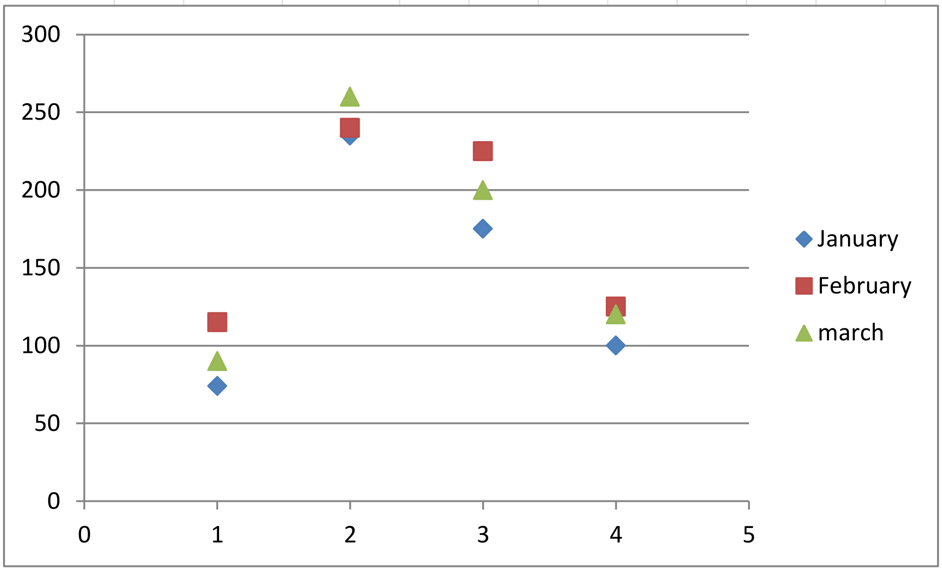
b- Pie chart : (For the previous example draw the pie chart for transportation)





c- Scatter plot :





d- Histogram :

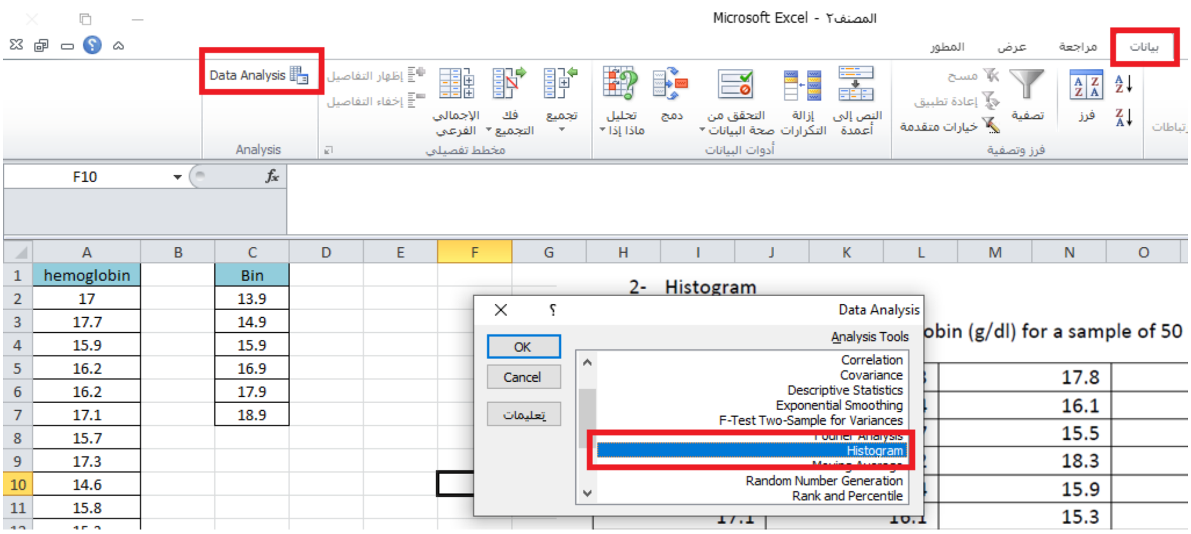
The following data represent hemoglobin (g/dl) for a sample of 50 women :

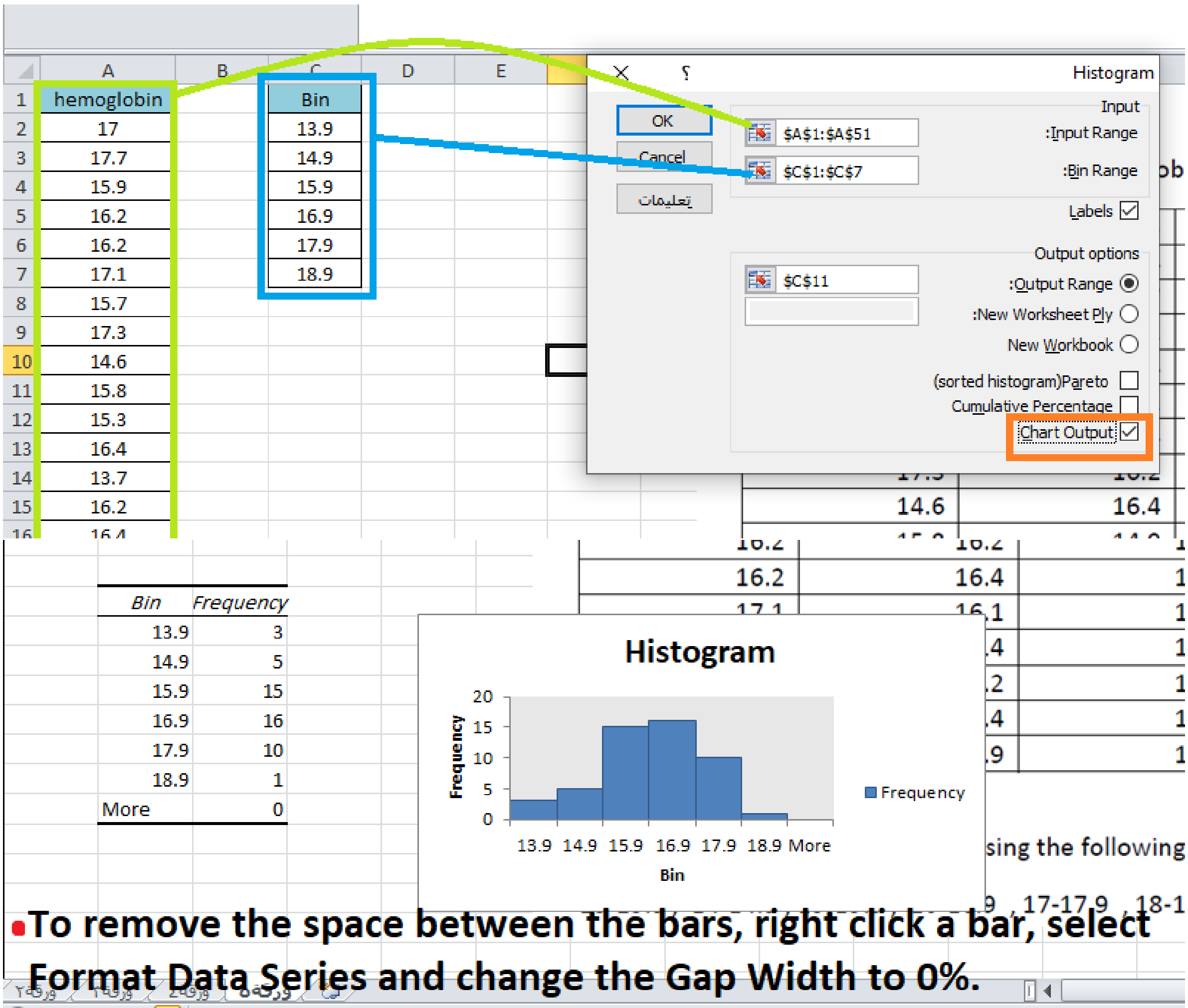
17	15.3	17.8	17.4	16.3
17.7	16.4	16.1	15	15.9
15.9	13.7	15.5	14.2	16.7
16.2	16.2	18.3	16.1	15.1
16.2	16.4	15.9	15.7	15.8
17.1	16.1	15.3	15.1	13.5
15.7	14	13.9	17.4	17
17.3	16.2	16.8	16.5	15.8
14.6	16.4	15.9	14.4	17.5
15.8	14.9	16.3	16.3	17.3

We wish to summarize these data using the following class intervals

13-13.9 , 14-14.9 , 15-15.9 , 16-16.9 , 17-17.9 , 18-18.9

Data → Data Analysis → Histogram

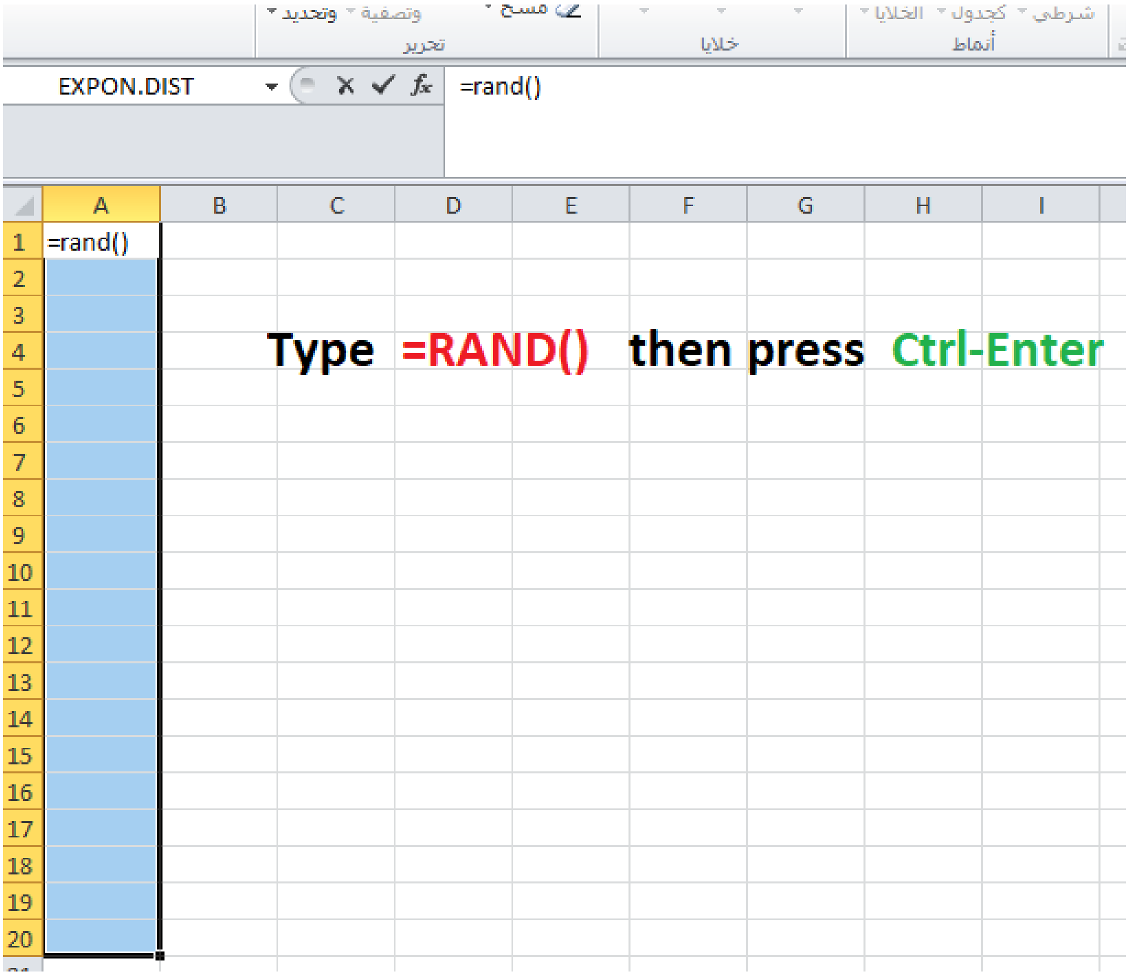




To remove the space between the bars, right click a bar, select Format Data Series and change the Gap Width to 0%.

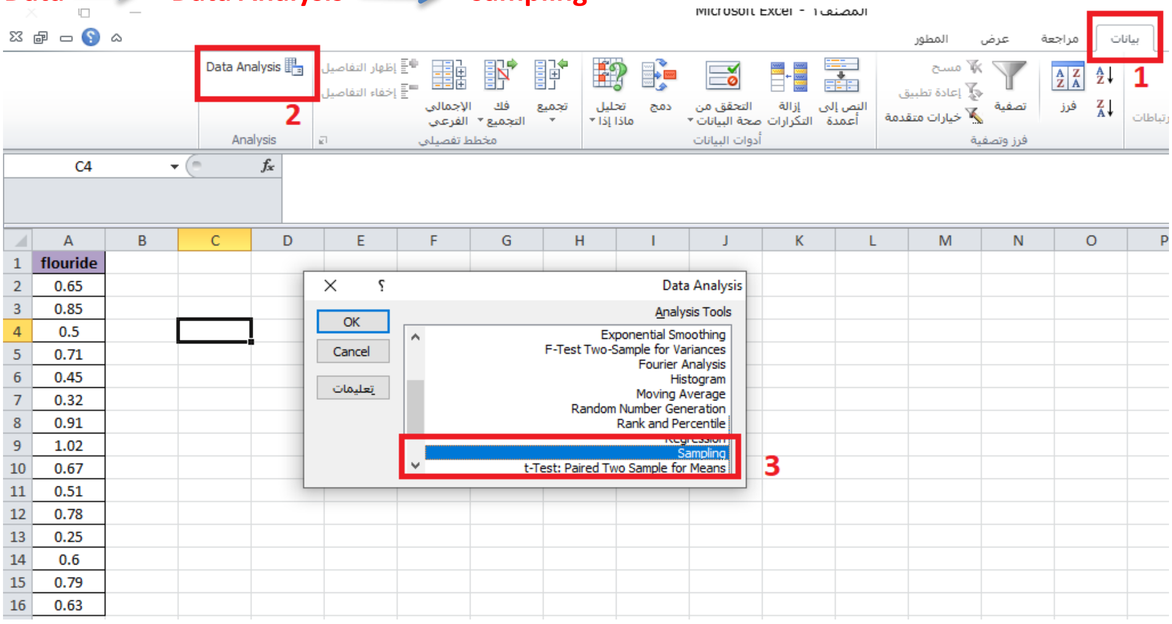
Generation Random samples :

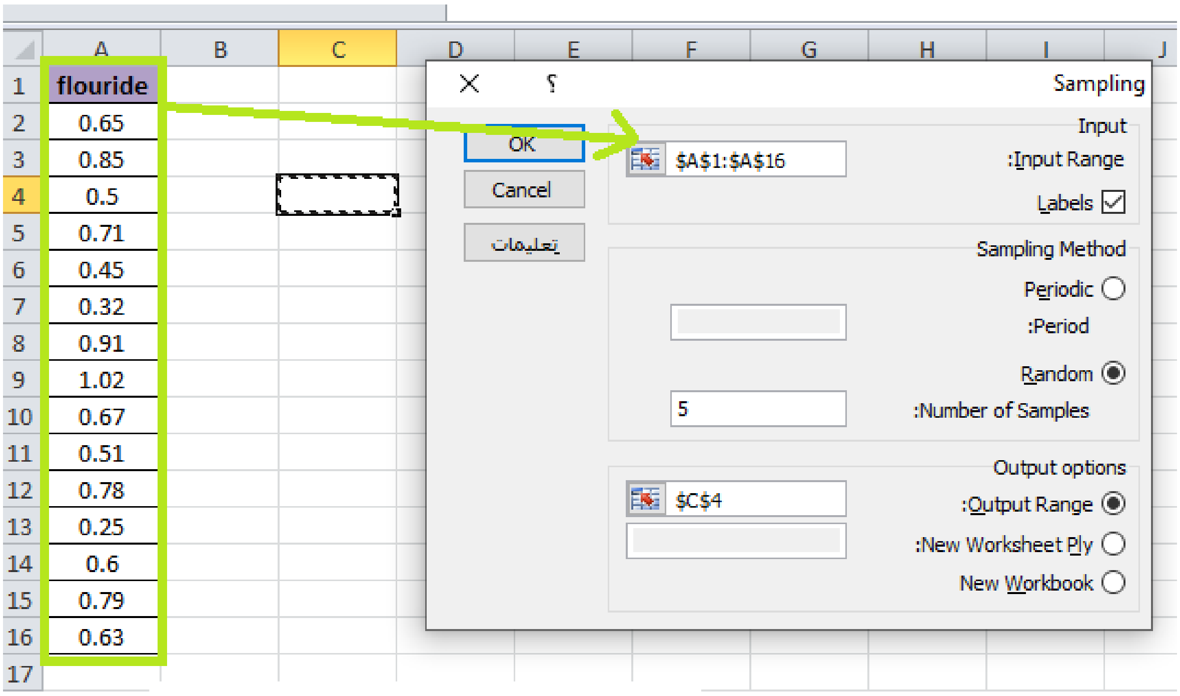
1- generate a random sample of size 20 between 0 and 1



2- Sampling

Data → Data Analysis → sampling

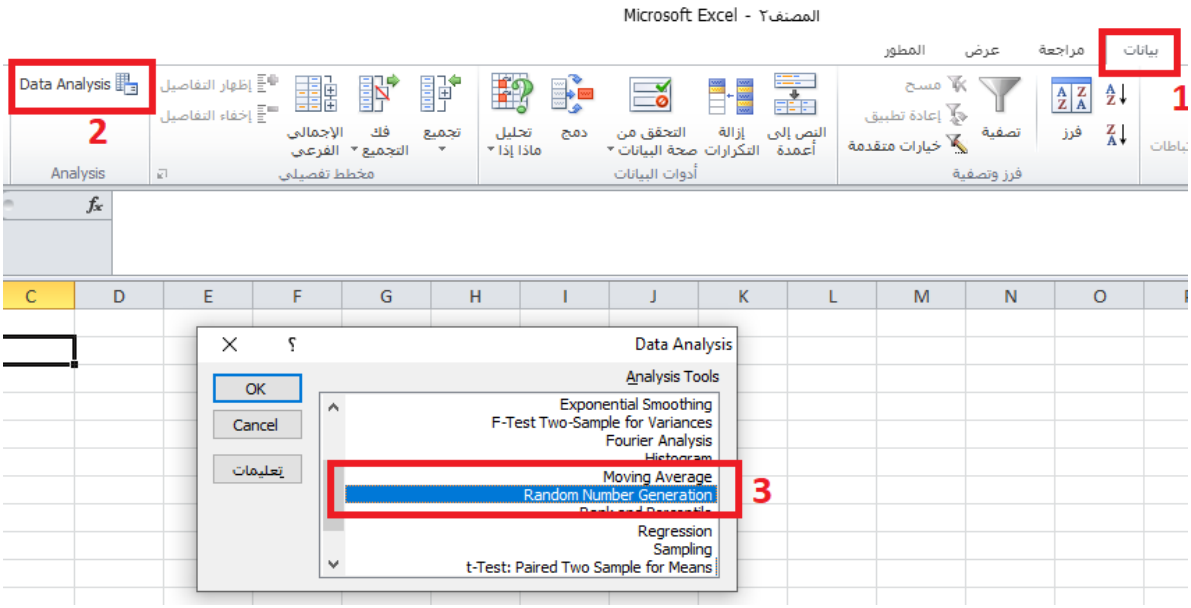




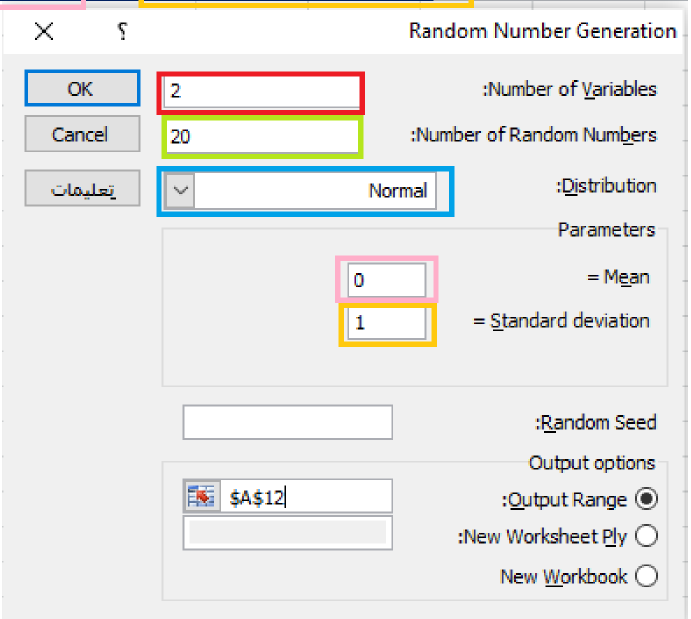
	A	B	C	D
1	flouride			
2	0.65			
3	0.85			
4	0.5		0.32	
5	0.71		0.25	
6	0.45		0.45	
7	0.32		0.78	
8	0.91		0.78	
9	1.02			
10	0.67			
11	0.51			
12	0.78			
13	0.25			
14	0.6			
15	0.79			
16	0.63			

3- Random number generation from distributions

To generate two random sample of size 20 from normal distribution with mean 0 and standard deviation 1



To generate two random sample of size 20 from normal distribution with mean 0 and standard deviation 1



	A	B	C
1	-0.30023	-1.27768	
2	0.244257	1.276474	
3	1.19835	1.733133	
4	-2.18359	-0.23418	
5	1.095023	-1.0867	
6	-0.6902	-1.69043	
7	-1.84691	-0.97763	
8	-0.77351	-2.11793	
9	-0.56792	-0.40405	
10	0.134853	-0.36549	
11	-0.32699	-0.37024	
12	1.342642	-0.08528	
13	-0.18616	-0.51321	
14	1.972212	0.865673	
15	2.375655	-0.65491	
16	1.661456	-1.6124	
17	0.538948	0.902191	
18	1.918916	-0.08452	
19	-0.5238	0.675138	
20	-0.38132	0.757611	
21			

Note: RANDOM SAMPLES

Use the Insert Function button on the standard tool bar or type directly.
=RAND( ) returns a random number between 0 and 1.
=RANDBETWEEN(bottom,top) returns a random number between the designated values.

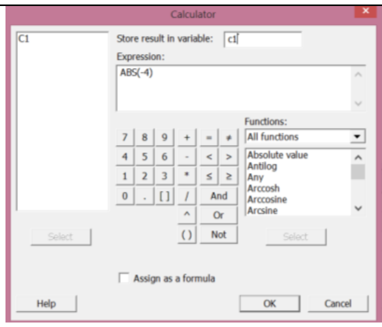
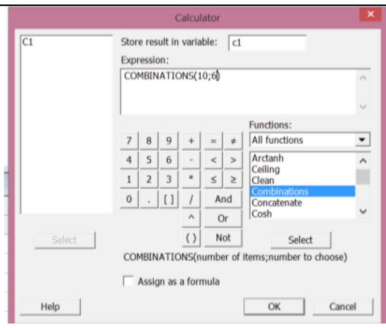
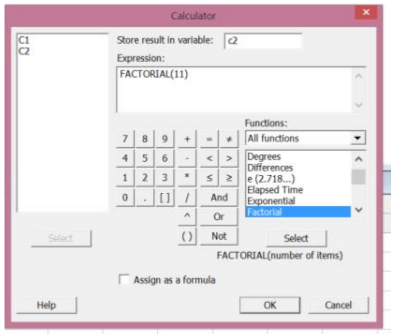
Use the menu selection >Data>Data Analysis to access the dialog box.
Sampling returns a random sample from a designated cell range.

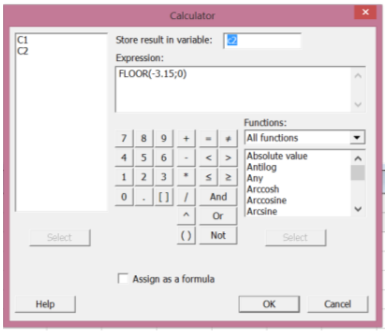
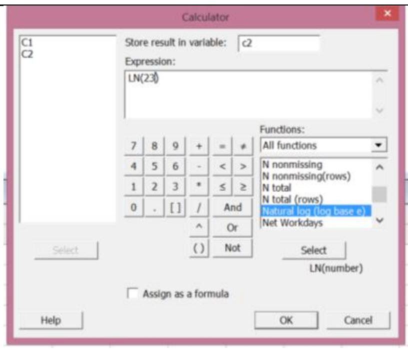
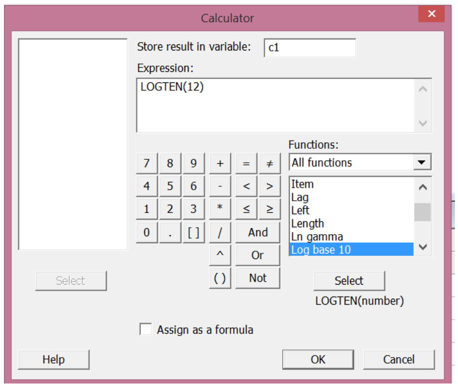
Random Number Generator returns a random sample from a designated distribution (uniform, normal, binomial, poisson).

Minitab Statistical Software

MATHEMATICAL FUNCTIONS

Write the commands of the following:

		By Minitab: **calc → calculator**
Absolute value	$\|-4\|=4$	
Combinations	$\binom{10}{6}$=10C6=210	
The exponential function	$e^{-1.6}$=0.201897	H.W
Factorial	11! =39916800	

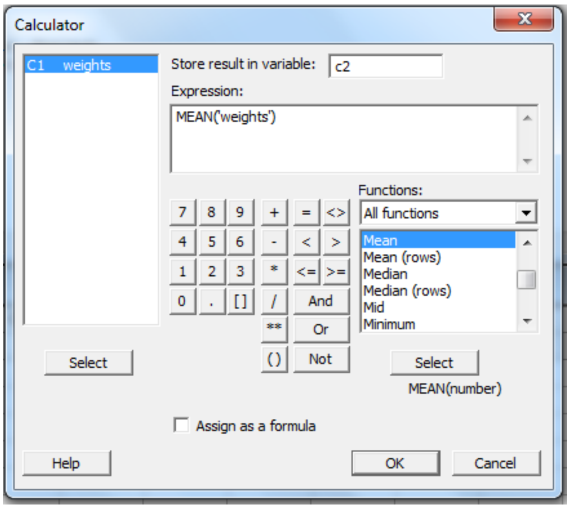
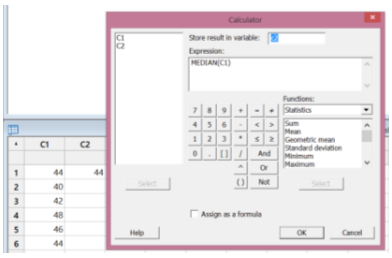
Floor function	$[-3.15]$= -4	
Natural logarithm	**ln(23)= 3.135494216**	
Logarithm with respect to any base	$\log_9(4) = 0.630929754$	H.W
Logarithm with respect to base 10	**log(12) = 1.079181246**	
Square root	$\sqrt{85}$= 9.219544457	HW
Summation	**Summation of: 450,11,20,5 = 486**	H.W
Permutations	**10P6=151200**	H.W
Powers	10^{-4}= 0.0001	H.W

DESCRIPTIVE STATISTICS

We have students' weights as follows:

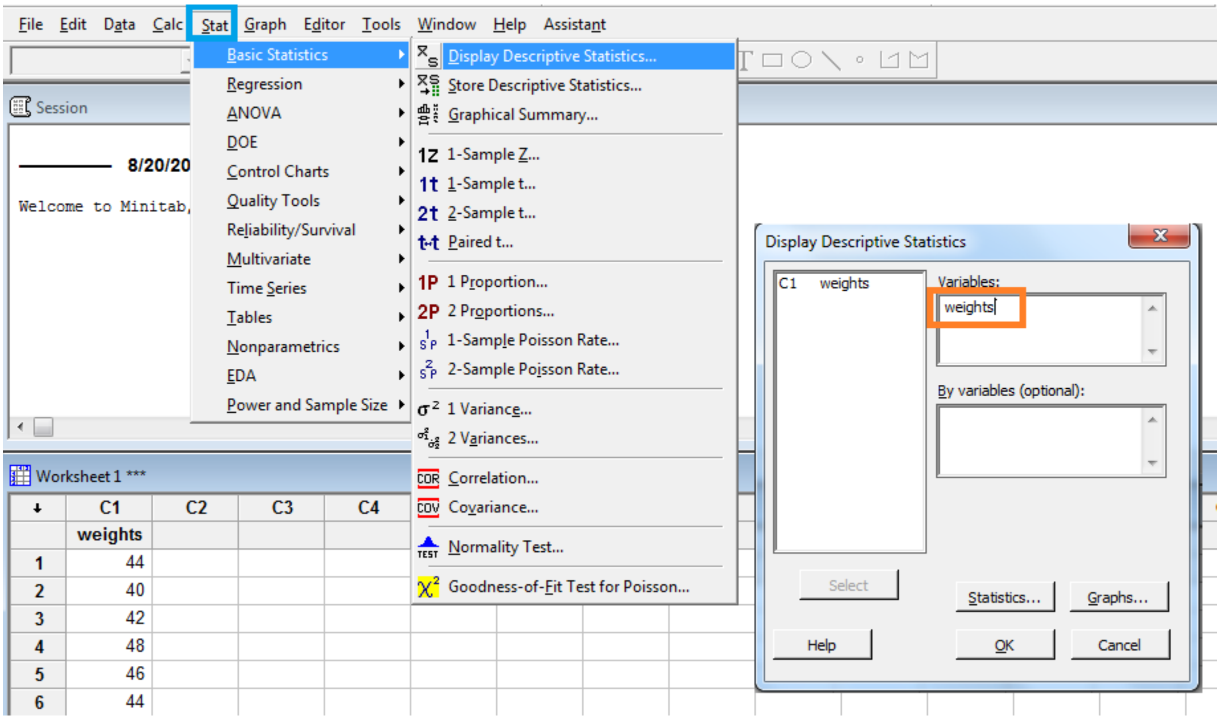
44 , 40 , 42 , 48 , 46 , 44

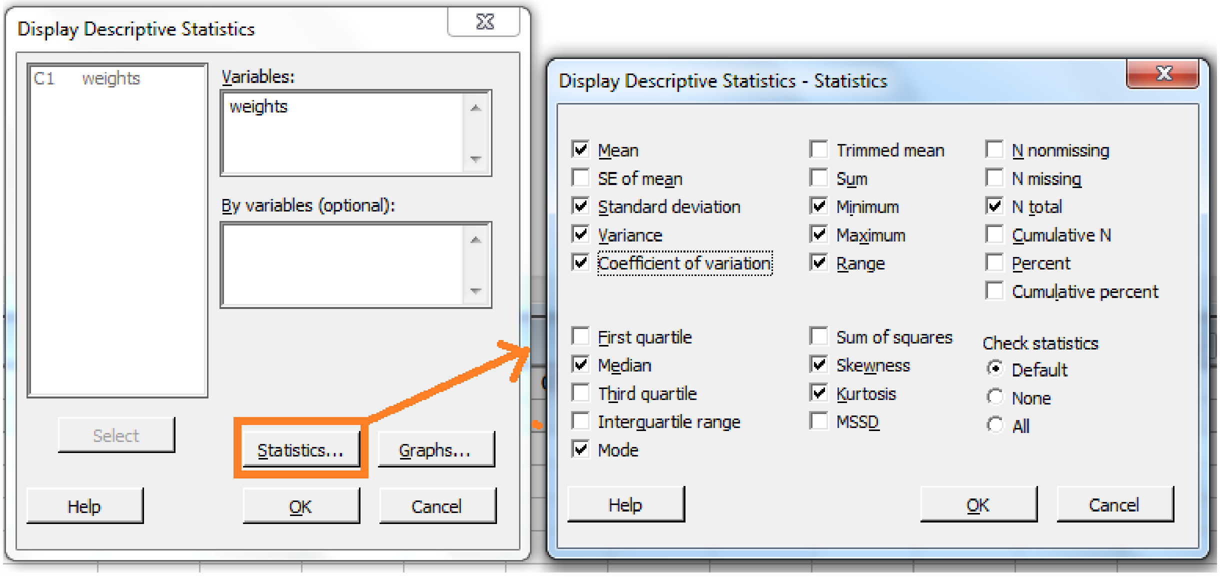
Find:

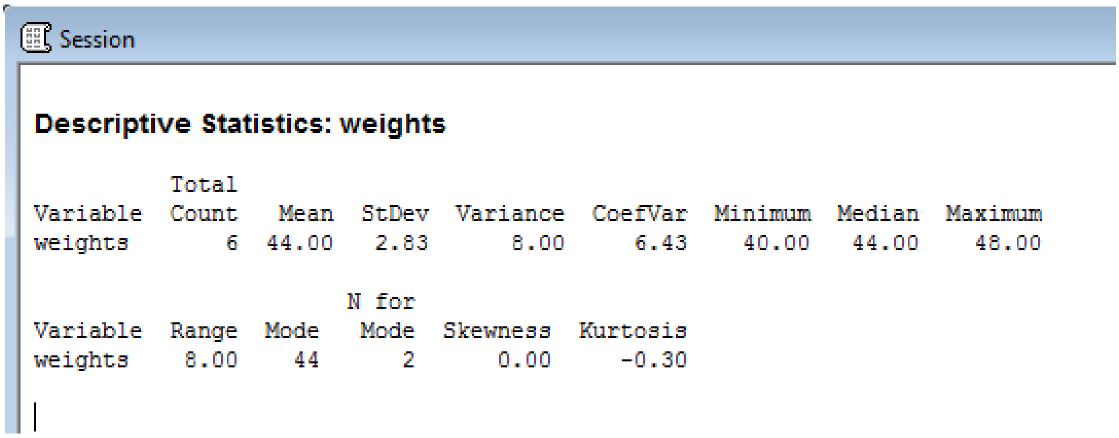
	By Minitab **calc → calculator**
Mean=44	
Median=44	

Mode=44	
Sample standard deviation=2.828	
Sample variance=8	
Kurtosis=-0.3	
Skewness=4.996E-17	
Minimum=40	
Maximum=48	
Range=8	
Count=6	
Coefficient of variation=6.428%	

OR







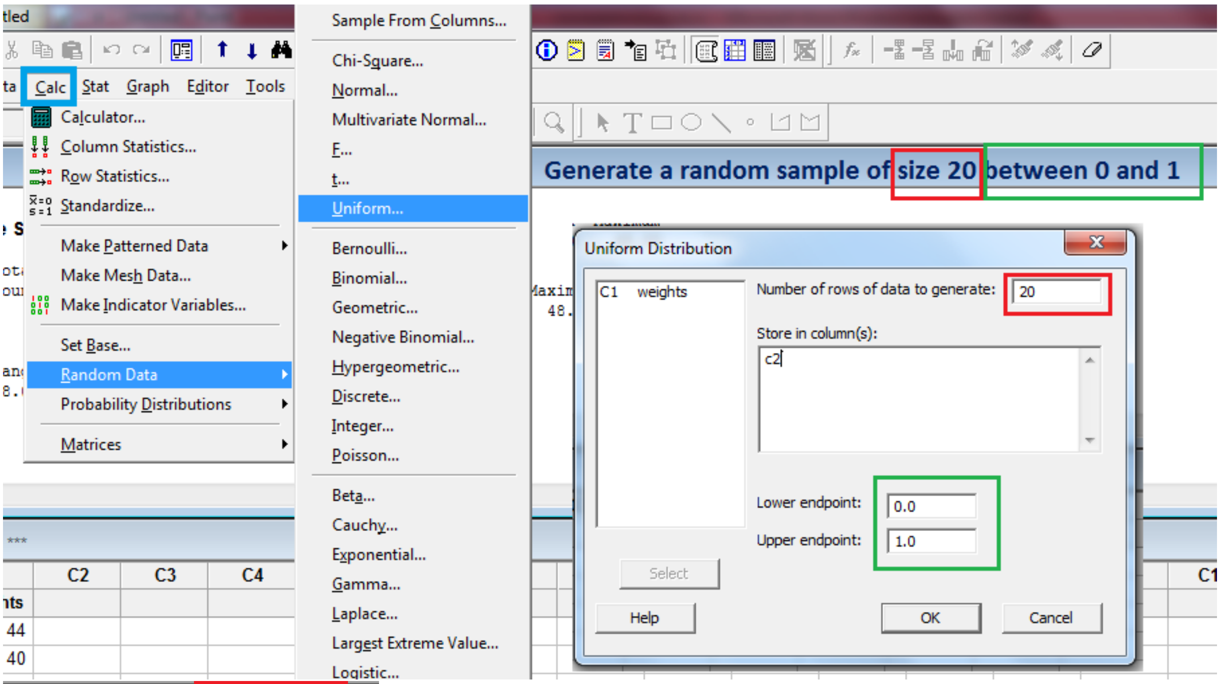
Descriptive Statistics: weights

Variable	Total Count	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum
weights	6	44.00	2.83	8.00	6.43	40.00	44.00	48.00

Variable	Range	Mode	N for Mode	Skewness	Kurtosis
weights	8.00	44	2	0.00	-0.30

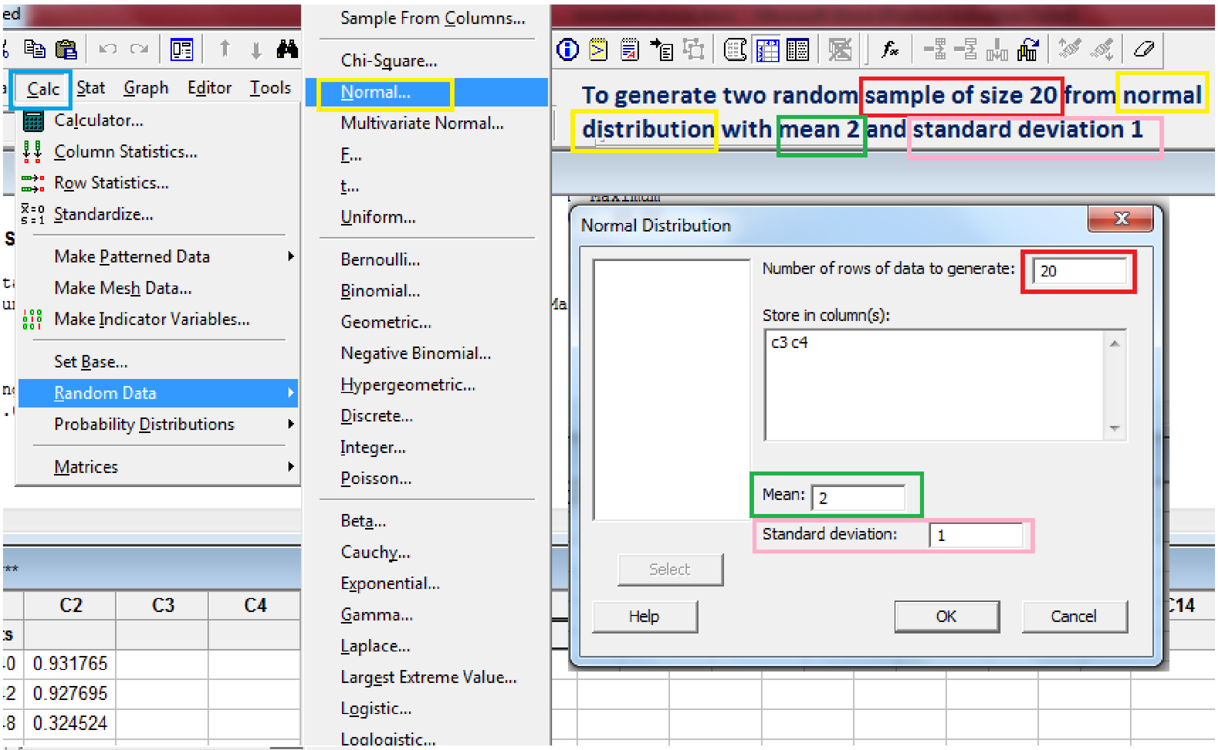
Generation Random samples

Generate a random sample of size 20 between 0 and 1



↓	C1	C2
	weights	
2	40	0.931765
3	42	0.927695
4	48	0.324524
5	46	0.778341
6	44	0.489656
7		0.322434
8		0.941024
9		0.659563
10		0.732617
11		0.989092
12		0.135542
13		0.926717
14		0.345093
15		0.412306
16		0.673885
17		0.852221
18		0.536819
19		0.664379
20		0.529992
21		

To generate two random sample of size 20 from normal distribution with mean 2 and standard deviation 1



↓	C3	C4
2	1.90538	3.24232
3	1.53281	0.79785
4	2.26516	1.61199
5	2.02053	1.55126
6	2.77391	3.18251
7	0.82860	2.13605
8	1.62206	2.28298
9	1.60335	2.53755
10	1.78842	3.14198
11	1.52002	1.91811
12	2.39124	1.91966
13	0.58042	1.06965
14	0.63477	1.07440
15	1.72190	1.46281
16	2.92055	2.64073
17	2.23790	0.28269
18	3.49190	2.45408
19	2.09695	0.77887
20	2.31778	0.22604

PROBABILITY DISTRIBUTION FUNCTIONS

Discrete Distribution :

1-Binomial Distribution :

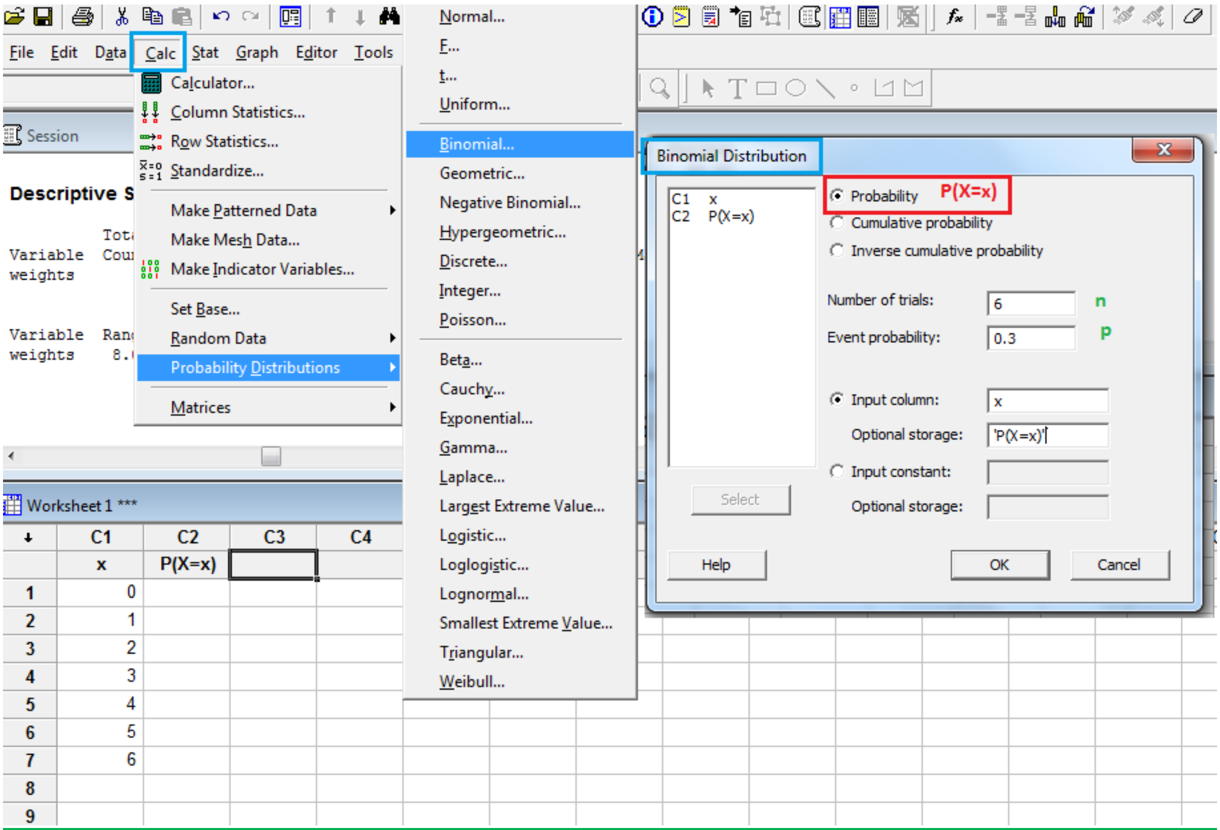
A biased coin is tossed 6 times. The probability of heads on any toss is 0.3. Let X denote the number of heads that come up. Calculate

X~Bin(n=6 , p=0.3)

a) $P(X = 2)$

b) $P(X = 3)$

c) $P(1 < X \leq 5)$.



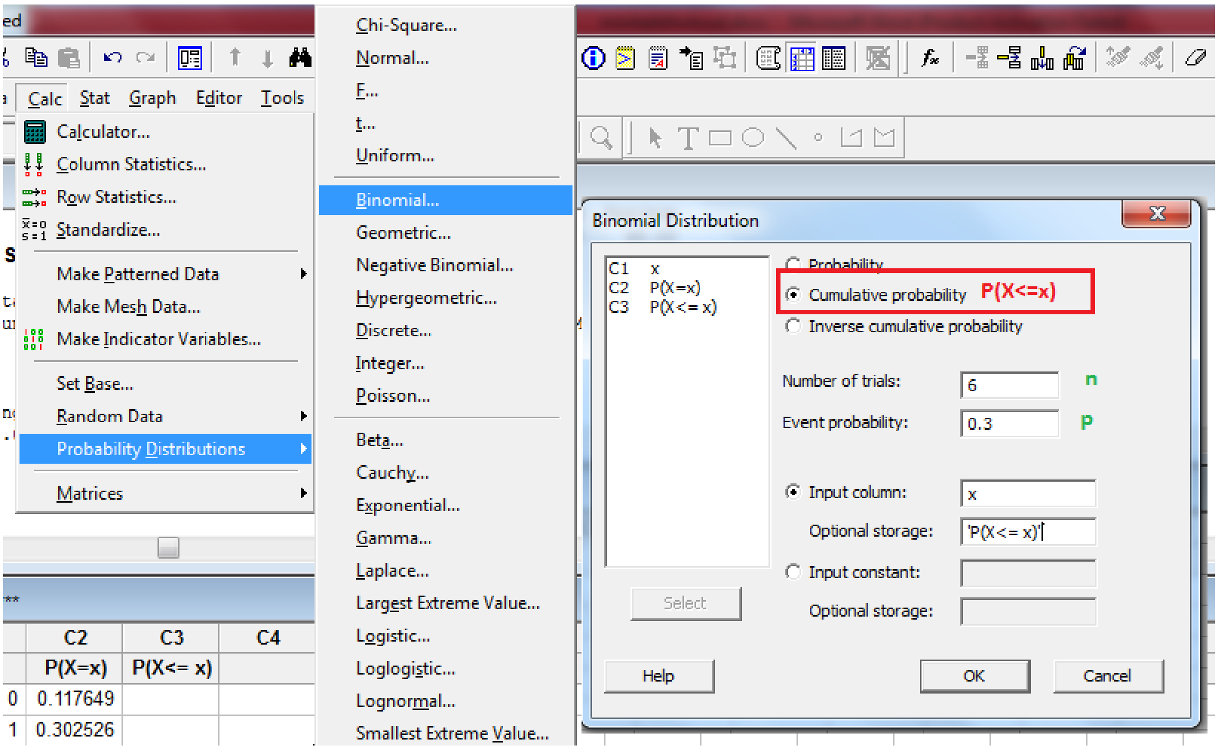
Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9
	x	P(X=x)							
1	0	0.117649	P(X=0)						
2	1	0.302526	P(X=1)						
3	2	0.324135	P(X=2)	(i)					
4	3	0.185220	P(X=3)	(ii)					
5	4	0.059535	P(X=4)						
6	5	0.010206	P(X=5)						
7	6	0.000729	P(X=6)						

(iii)

0.324145+0.185220+0.059535+0.010206 =0.578

OR



Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C1
	x	P(X=x)	P(X<= x)									
1	0	0.117649	0.11765									
2	1	0.302526	0.42017									
3	2	0.324135	0.74431									
4	3	0.185220	0.92953									
5	4	0.059535	0.98906									
6	5	0.010206	0.99927									
7	6	0.000729	1.00000									

$$P(1 < X \leq 5) = P(X \leq 5) - P(X \leq 1) = 0.99927 - 0.42017 = 0.578$$

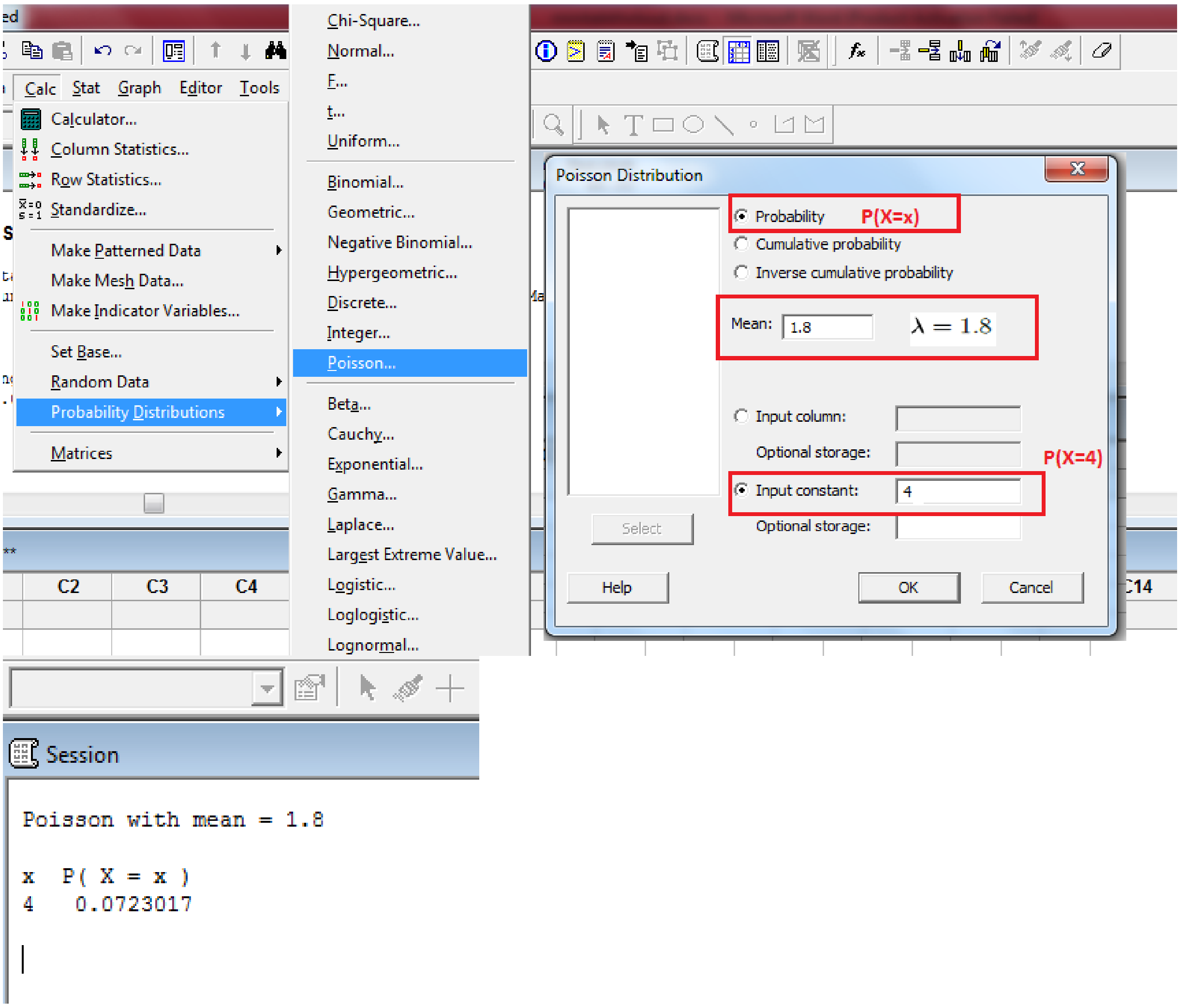
2.Poisson Distribution :

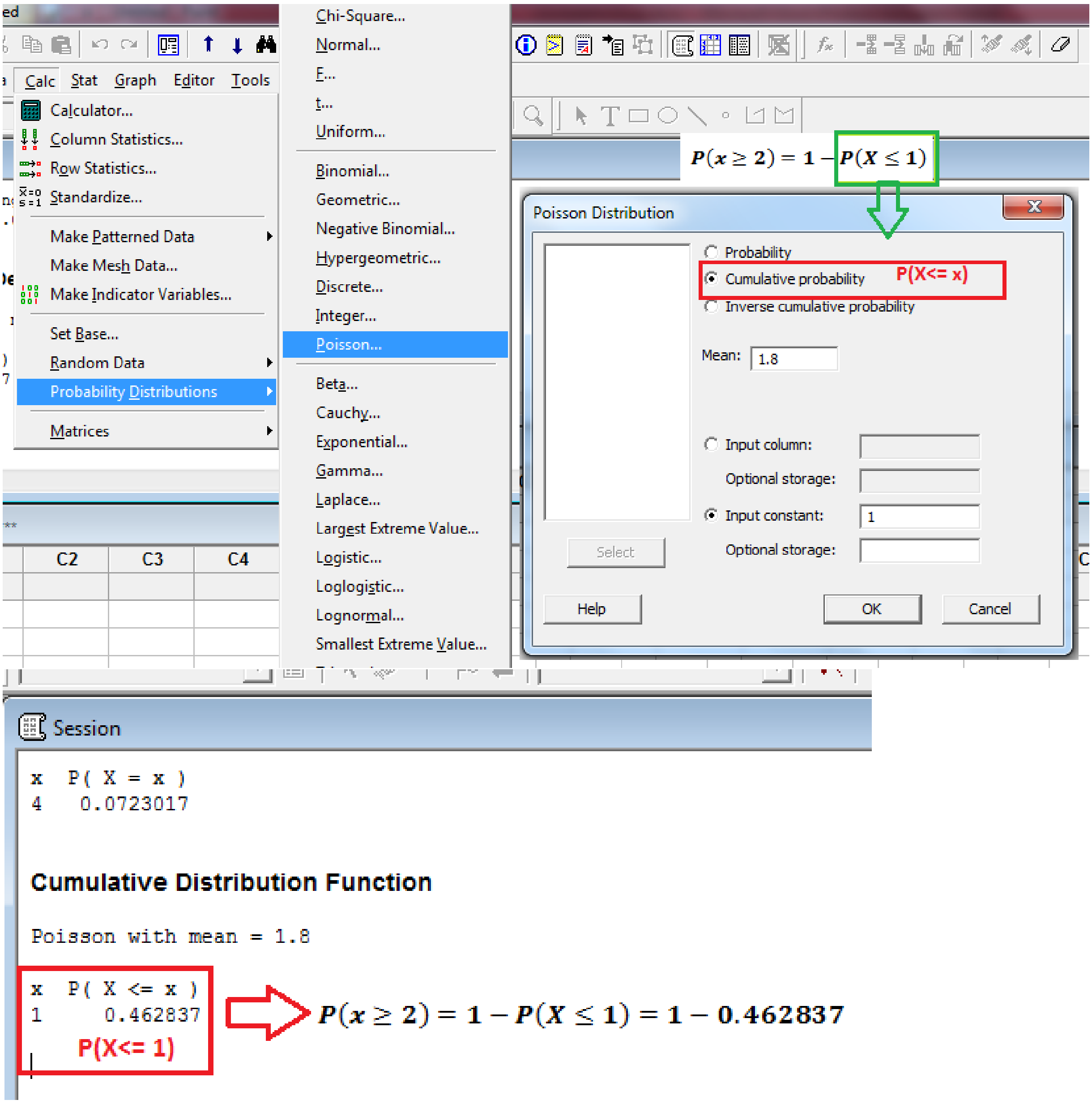
Births in a hospital occur randomly at an average rate of 1.8 births per hour.

a) What is the probability of observing 4 births in a given hour at the hospital?

b) What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

X~poisson(λ=1.8)

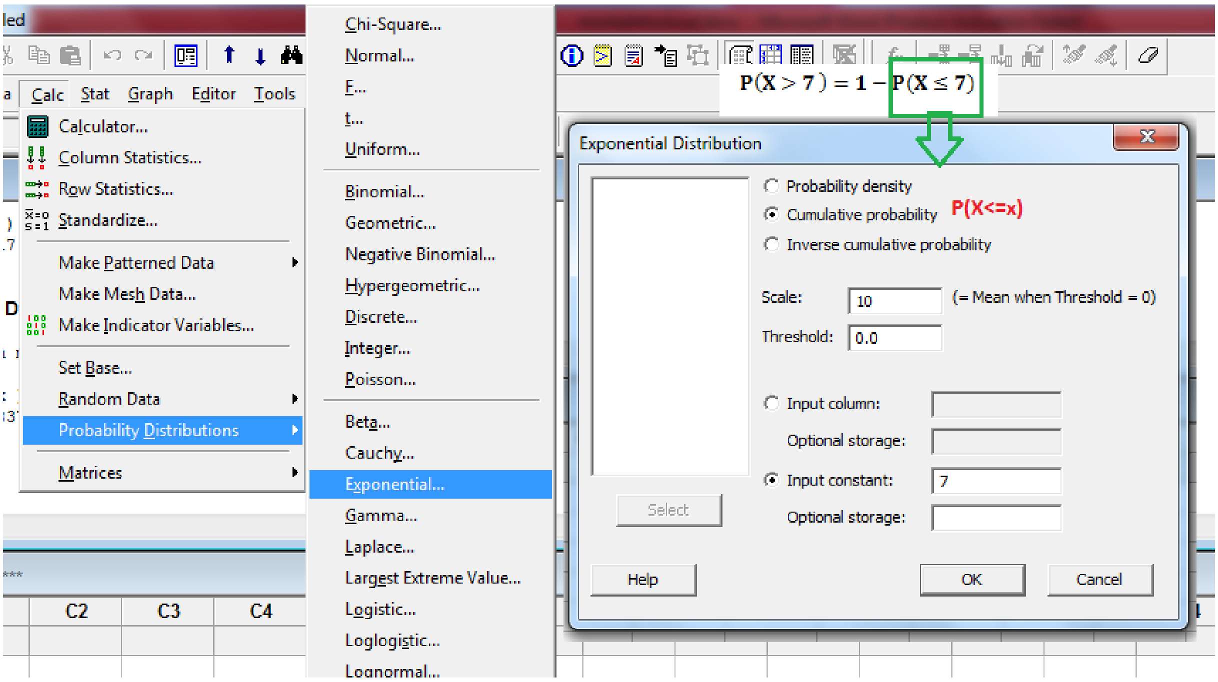


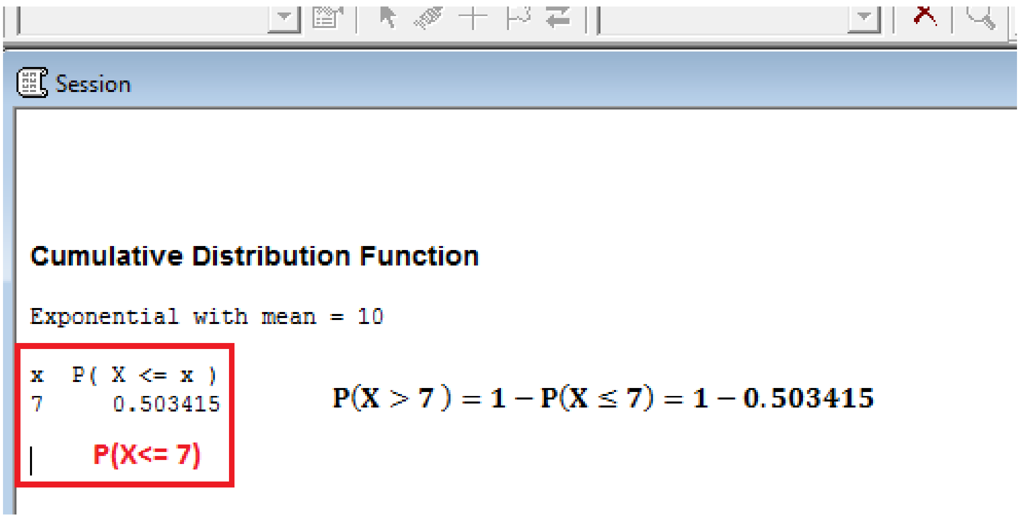

Calc Stat Graph Editor Tools
Calculator...
Column Statistics...
Row Statistics...
Standardize...
Make Patterned Data
Make Mesh Data...
Make Indicator Variables...
Set Base...
Random Data
Probability Distributions
Matrices
Chi-Square...
Normal...
F...
t...
Uniform...
Binomial...
Geometric...
Negative Binomial...
Hypergeometric...
Discrete...
Integer...
Poisson...
Beta...
Cauchy...
Exponential...
Gamma...
Laplace...
Largest Extreme Value...
Logistic...
Loglogistic...
Lognormal...
Smallest Extreme Value...
C2 C3 C4
$P(x \geq 2) = 1 - P(X \leq 1)$
Poisson Distribution
Probability
Cumulative probability
P(X<= x)
Inverse cumulative probability
Mean: 1.8
Input column:
Optional storage:
Input constant: 1
Optional storage:
Select
Help
OK
Cancel
Session
x P(X = x)
4 0.0723017
Cumulative Distribution Function
Poisson with mean = 1.8
x P(X <= x)
1 0.462837
P(X<= 1)
$P(x \geq 2) = 1 - P(X \leq 1) = 1 - 0.462837$

Continuous Distribution :

1. Exponential Distribution :

If X~exp(λ=1/10) , Find P(X>7)

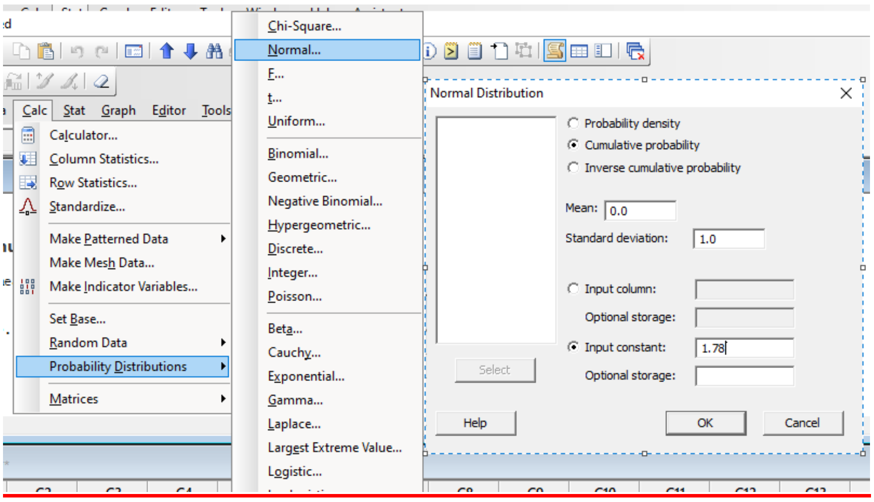




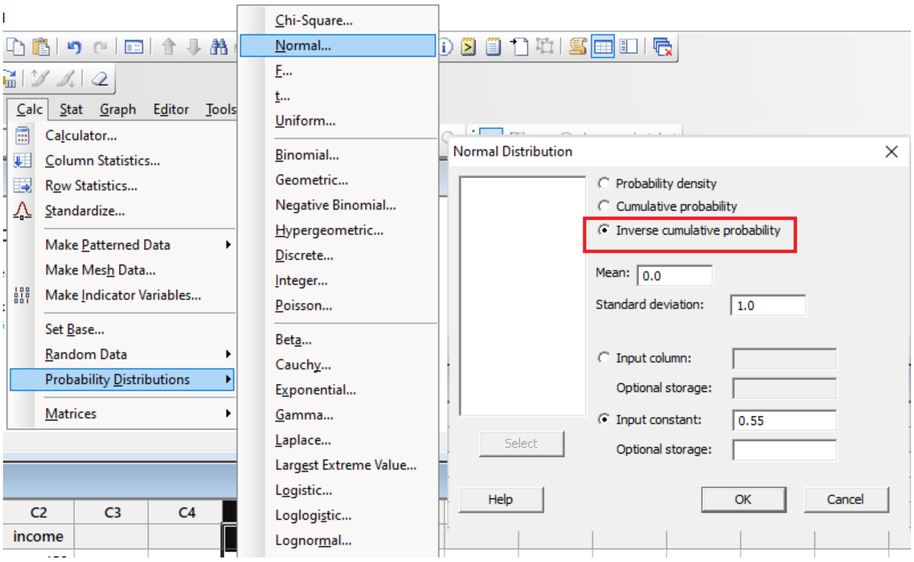
2. Normal Distribution :

If z ∽ N(μ = 0 , σ = 1) . Find :

A)P(Z ≤ 1.78) = P(Z < 1.78) =



B) P(Z ≤z_0)=0.55 , z_0 =



Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1

   x  P( X ≤ x )
1.78    0.962462
```

A)P(Z ≤ 1.78) = P(Z < 1.78) =

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1

P( X ≤ x )         x
      0.55  0.125661
```

B) P(Z ≤z_0)=0.55 , z_0 =

MATRICES

Addition of Matrices	$A=\begin{bmatrix}-5 & 0\\ 4 & 1\end{bmatrix}, B=\begin{bmatrix}6 & -3\\ 2 & 3\end{bmatrix}$ $\Rightarrow A+B=\begin{bmatrix}-5+6 & 0+-3\\ 4+2 & 1+3\end{bmatrix}=\begin{bmatrix}1 & -3\\ 6 & 4\end{bmatrix}$	**MTB > copy c1-c2 m1** **MTB > copy c3-c4 m2** **MTB > add m1 m2 m3** **MTB > print m3**
Subtract of Matrices	$C=\begin{bmatrix}1 & 2\\ -2 & 0\\ -3 & -1\end{bmatrix}, D=\begin{bmatrix}1 & -1\\ 1 & 3\\ 2 & 3\end{bmatrix}$ $\Rightarrow C-D=\begin{bmatrix}1-1 & 2-(-1)\\ -2-1 & 0-3\\ -3-2 & -1-3\end{bmatrix}=\begin{bmatrix}0 & 3\\ -3 & -3\\ -5 & -4\end{bmatrix}$	**MTB > copy c3-c4 m4** **MTB > copy c5-c6 m5** **MTB > subt m5 m4 m6** **MTB > print m6**
Additive Inverse of Matrix	$A=\begin{bmatrix}1 & 0 & 2\\ 3 & -1 & 5\end{bmatrix}$ $\Rightarrow -A=\begin{bmatrix}-1 & 0 & -2\\ -3 & 1 & -5\end{bmatrix}$	**MTB > copy c7-c9 m7** **MTB > mult -1 m7 m8** **MTB > print m8**
Scalar Multiplication of Matrices	$D=\begin{bmatrix}-3 & 0\\ 4 & 5\end{bmatrix}$ $\Rightarrow 3D=\begin{bmatrix}-9 & 0\\ 12 & 15\end{bmatrix}$	**MTB > copy c10-c11 m9** **MTB > mult 3 m9 m10** **MTB > print m10**
Matrix Multiplication	$E=\begin{bmatrix}1 & 4 & 7\\ 2 & 5 & 8\\ 3 & 6 & 9\end{bmatrix}, F=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$ $\Rightarrow E\times F=\begin{bmatrix}30 & 66\\ 36 & 81\\ 42 & 96\end{bmatrix}$	**MTB > copy c11-c13 m11** **MTB > copy c14-c15 m12** **MTB > mult m11 m12 m13** **MTB > print m13**
Inverse Matrices	$G=\begin{bmatrix}3 & -1\\ -5 & 2\end{bmatrix}$ $\Rightarrow \det(G)=1 \text{ and } G^{-1}=\begin{bmatrix}2 & 1\\ 5 & 3\end{bmatrix}$	**MTB > copy c16-c17 m14** **MTB > inver m14 m15** **MTB > print m15**

```
MTB > copy c1 c2 m1
MTB > print m1
```

input matrix A and denoted by m1

Data Display

```
 Matrix M1

-5   0
 4   1

MTB > copy c3 c4 m2
MTB > print m2
```

input matrix B and denoted by m2

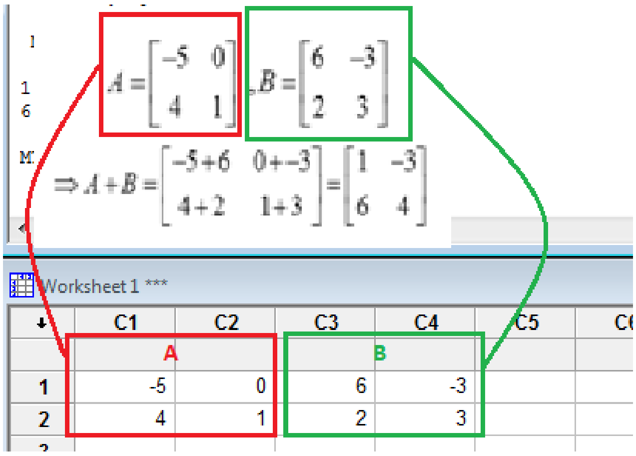
Data Display

```
 Matrix M2

6   -3
2    3

MTB > add m1 m2 m3
MTB > print m3
```

m1 + m2 and denoted by m3

Data Display

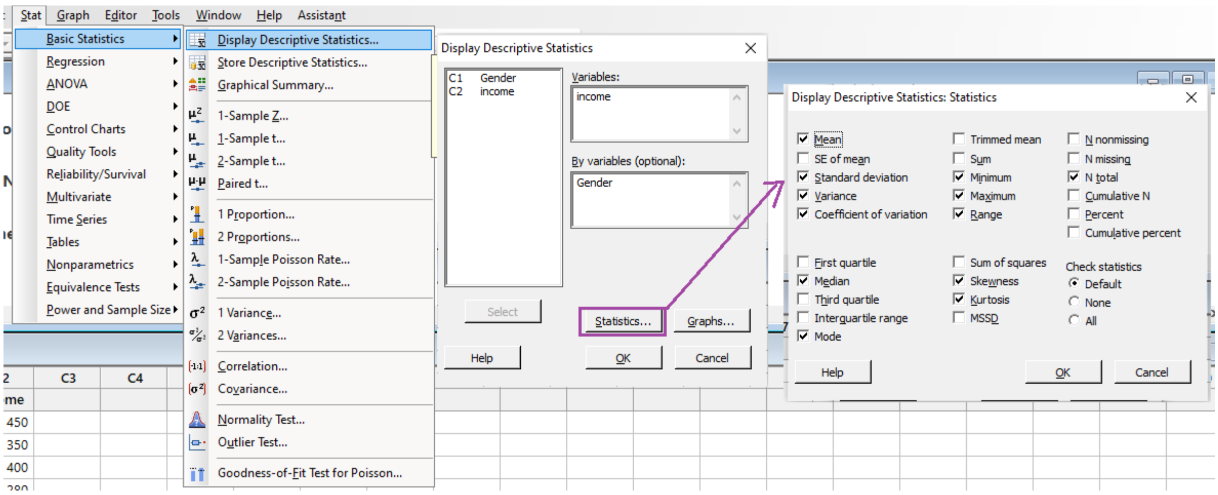
```
 Matrix M3

1   -3
6    4

MTB >
```

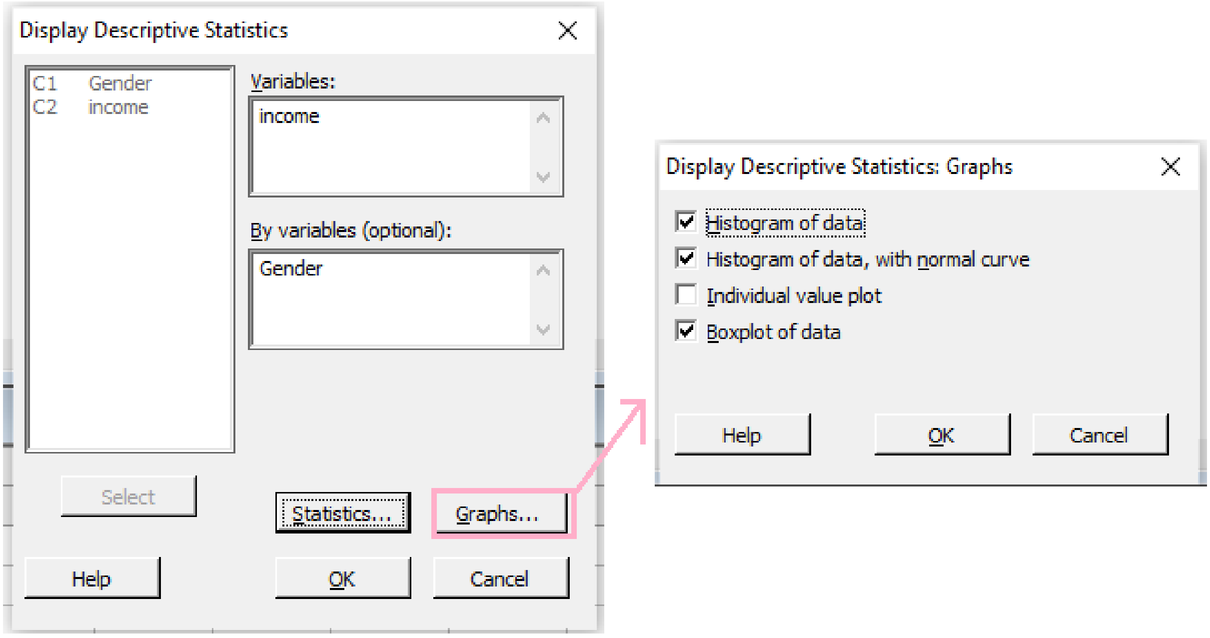


The following table gives the monthly income for sample of employees; analyze the data based on gender (M: Male , F: Female)

Gender	M	F	F	F	M	M	F	M	F	F	M	M
income	450	350	400	280	500	480	300	300	260	240	520	400

↓	C1-T	C2
	Gender	income
1	M	450
2	F	350
3	F	400
4	F	280
5	M	500
6	M	480
7	F	300
8	M	300
9	F	260
10	F	240
11	M	520
12	M	400

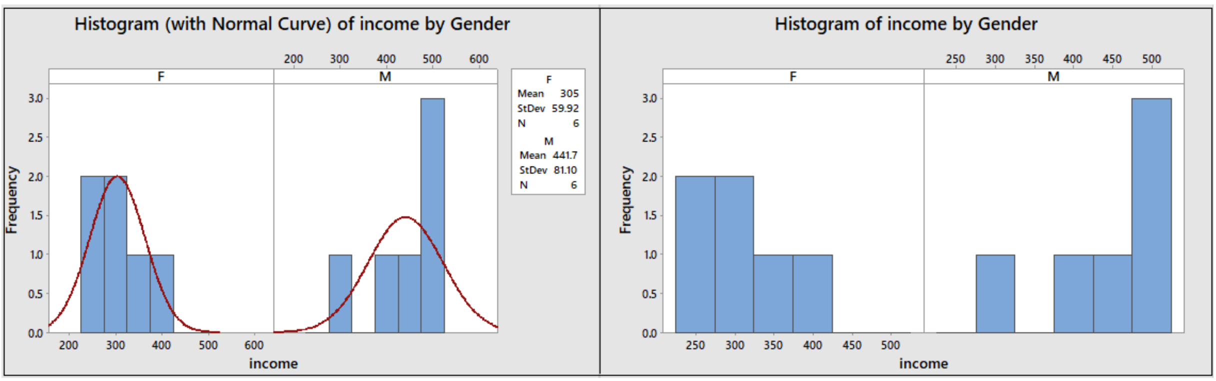


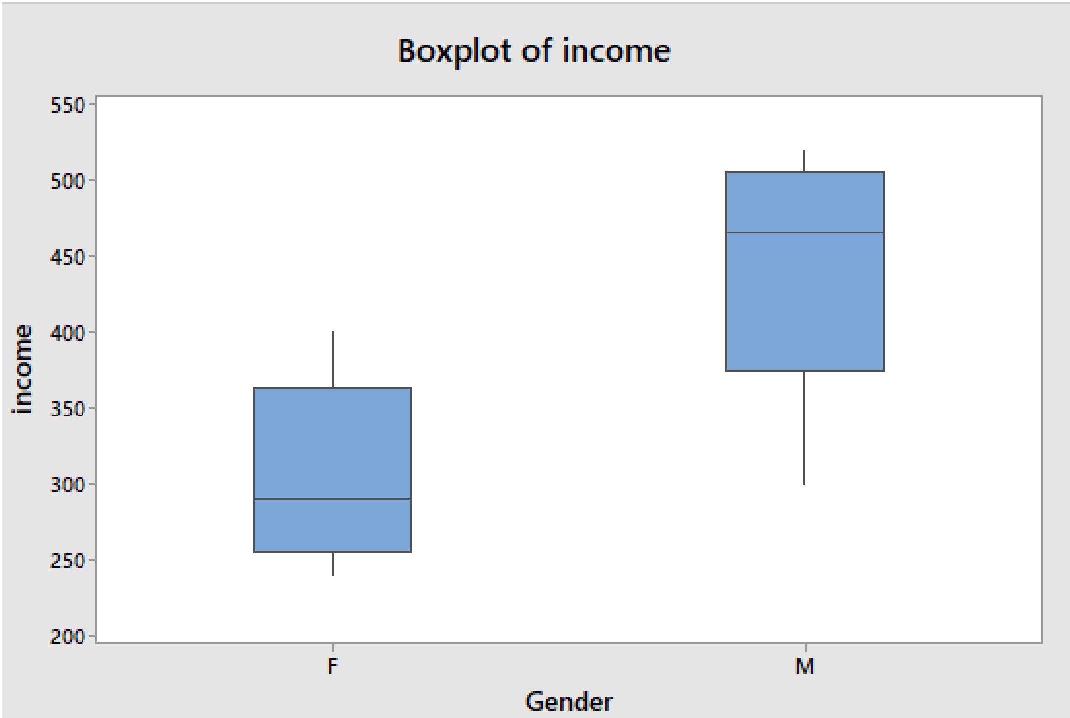


Session

Variable	Gender	Total Count	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum	Range
income	F	6	305.0	59.9	3590.0	19.64	240.0	290.0	400.0	160.0
	M	6	441.7	81.1	6576.7	18.36	300.0	465.0	520.0	220.0

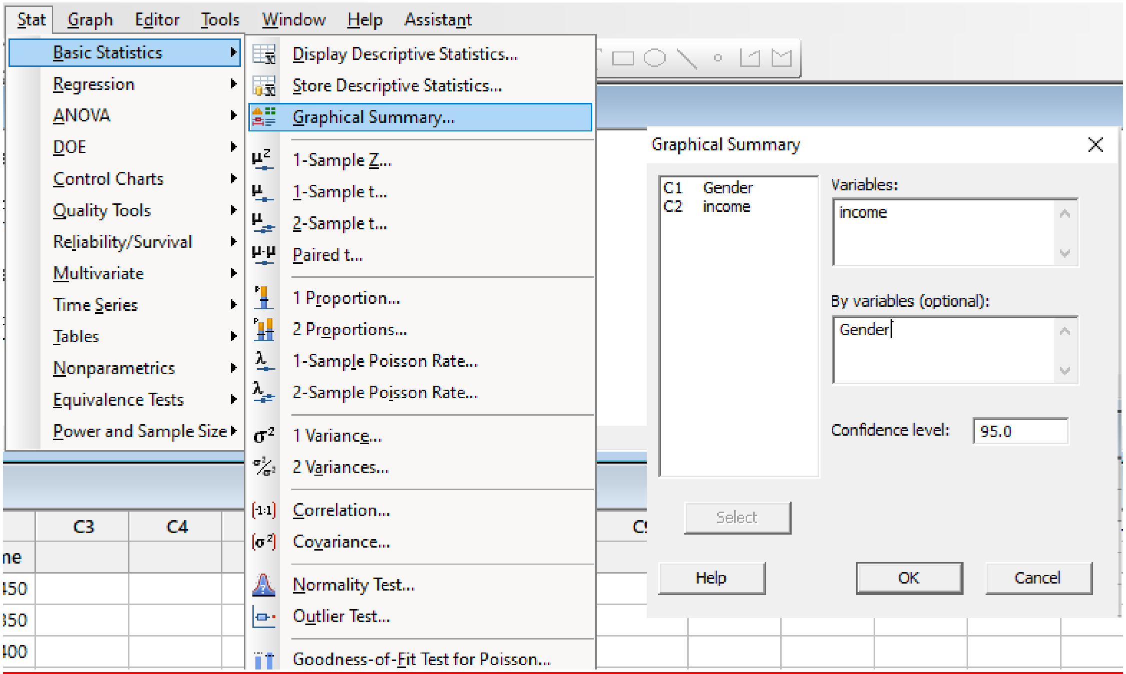
Variable	Gender	Mode	N for Mode	Skewness	Kurtosis
income	F	*	0	0.79	-0.39
	M	*	0	-1.23	1.15

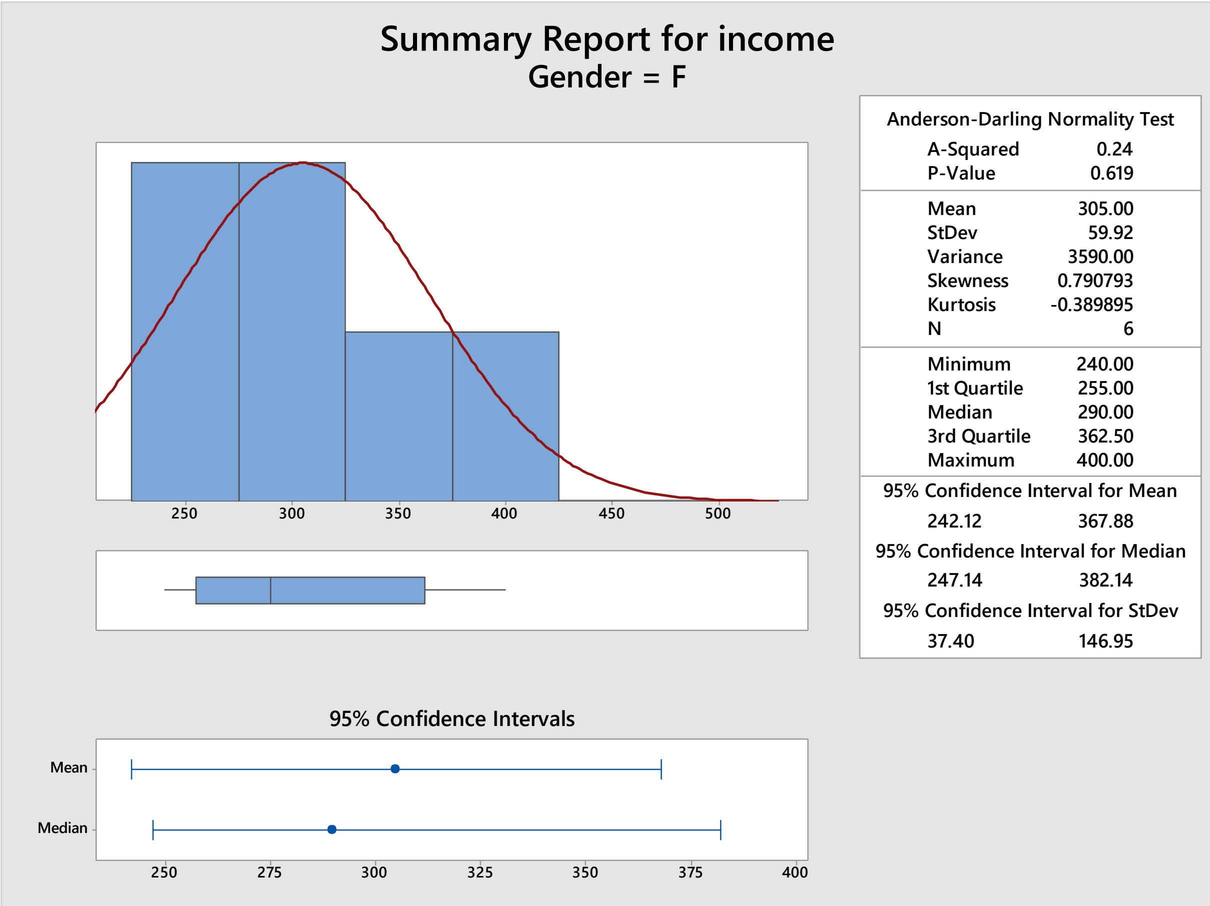


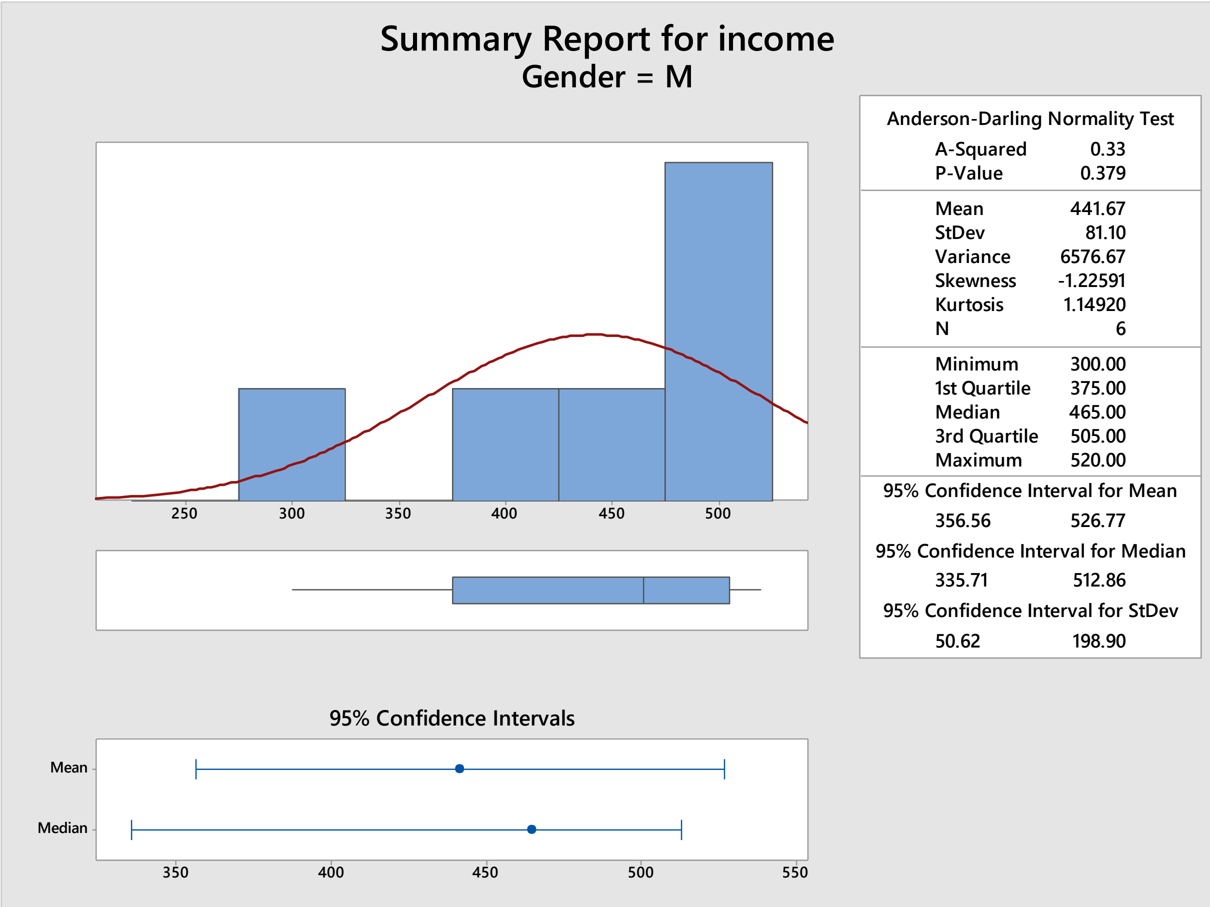


- **Graphical Summary**

The graphical summary can be also introduced for the income of both male and female as follows




Summary Report for income
Gender = F
Anderson-Darling Normality Test
A-Squared 0.24
P-Value 0.619
Mean 305.00
StDev 59.92
Variance 3590.00
Skewness 0.790793
Kurtosis -0.389895
N 6
Minimum 240.00
1st Quartile 255.00
Median 290.00
3rd Quartile 362.50
Maximum 400.00
95% Confidence Interval for Mean
242.12 367.88
95% Confidence Interval for Median
247.14 382.14
95% Confidence Interval for StDev
37.40 146.95
250 300 350 400 450 500
95% Confidence Intervals
Mean
Median
250 275 300 325 350 375 400

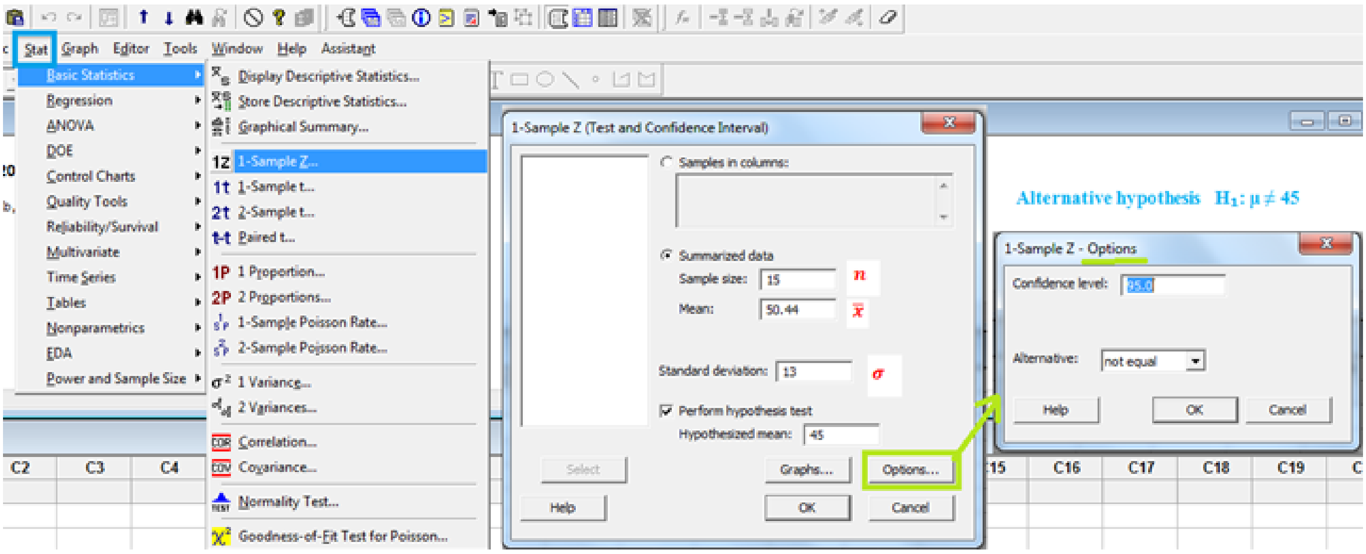

Summary Report for income
Gender = M
Anderson-Darling Normality Test
A-Squared 0.33
P-Value 0.379
Mean 441.67
StDev 81.10
Variance 6576.67
Skewness -1.22591
Kurtosis 1.14920
N 6
Minimum 300.00
1st Quartile 375.00
Median 465.00
3rd Quartile 505.00
Maximum 520.00
95% Confidence Interval for Mean
356.56 526.77
95% Confidence Interval for Median
335.71 512.86
95% Confidence Interval for StDev
50.62 198.90
250 300 350 400 450 500
95% Confidence Intervals
Mean
Median
350 400 450 500 550

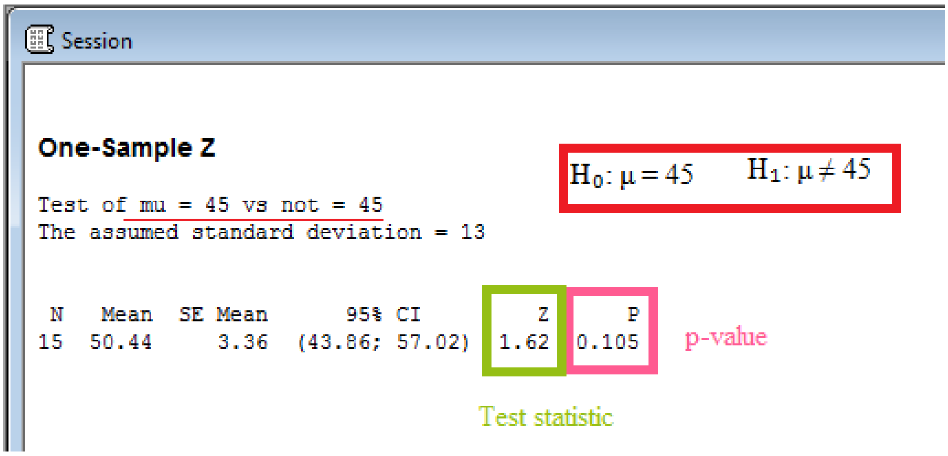
One-sample z-test

Q: In a study on samples fruit grown in central Saudi Arabia, 15 samples of ripe fruit were analyzed for Vitamin C content obtaining a mean of 50.44 mg/100g . Assume that Vitamin C contents are normally distributed with a standard deviation of 13. At α=0.05,

a) **Test whether the true mean vitamin C content is different from 45 mg/100g**
b) **Find a 95% confidence interval for the average vitamin C content .**

> *** Use the 1-sample Z-test to estimate the mean of a population and compare it to a target or reference value when you know the standard deviation of the population**
> **Using this test, you can:**
> **- Determine whether the mean of a group differs from a specified value.**
> **- Calculate a range of values that is likely to include the population mean.**





1- **Hypothesis:**

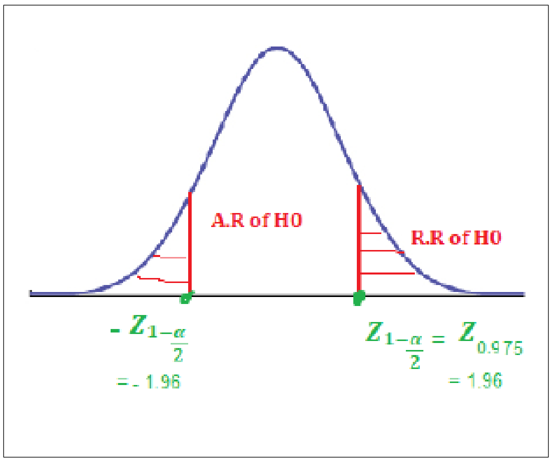
Null hypothesis H_0: $\mu = 45$ VS Alternative hypothesis H_1: $\mu \neq 45$

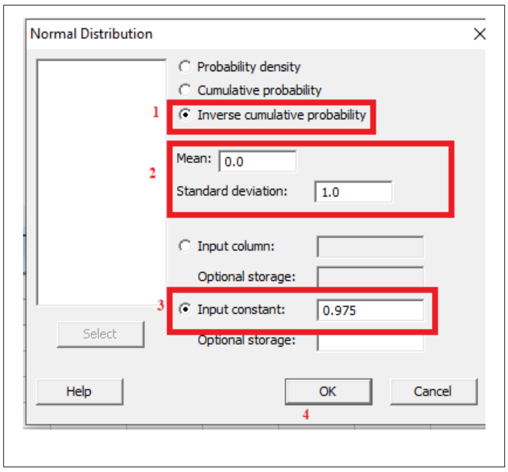
2- **Test statistic :**

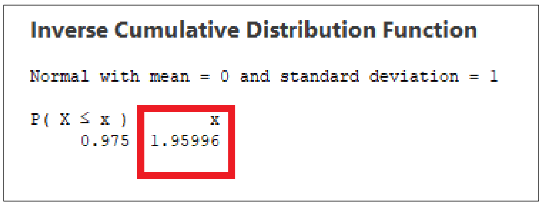
Z=1.62

3- **The critical region(s)**

Calc>> probability distributions>>Normal







4- **Decision:**

Since p-value =0.105 > α= 0.05 . we can not reject H_0

The 95% CI for the mean μ : (43.86 , 57.02)

One-sample t-test

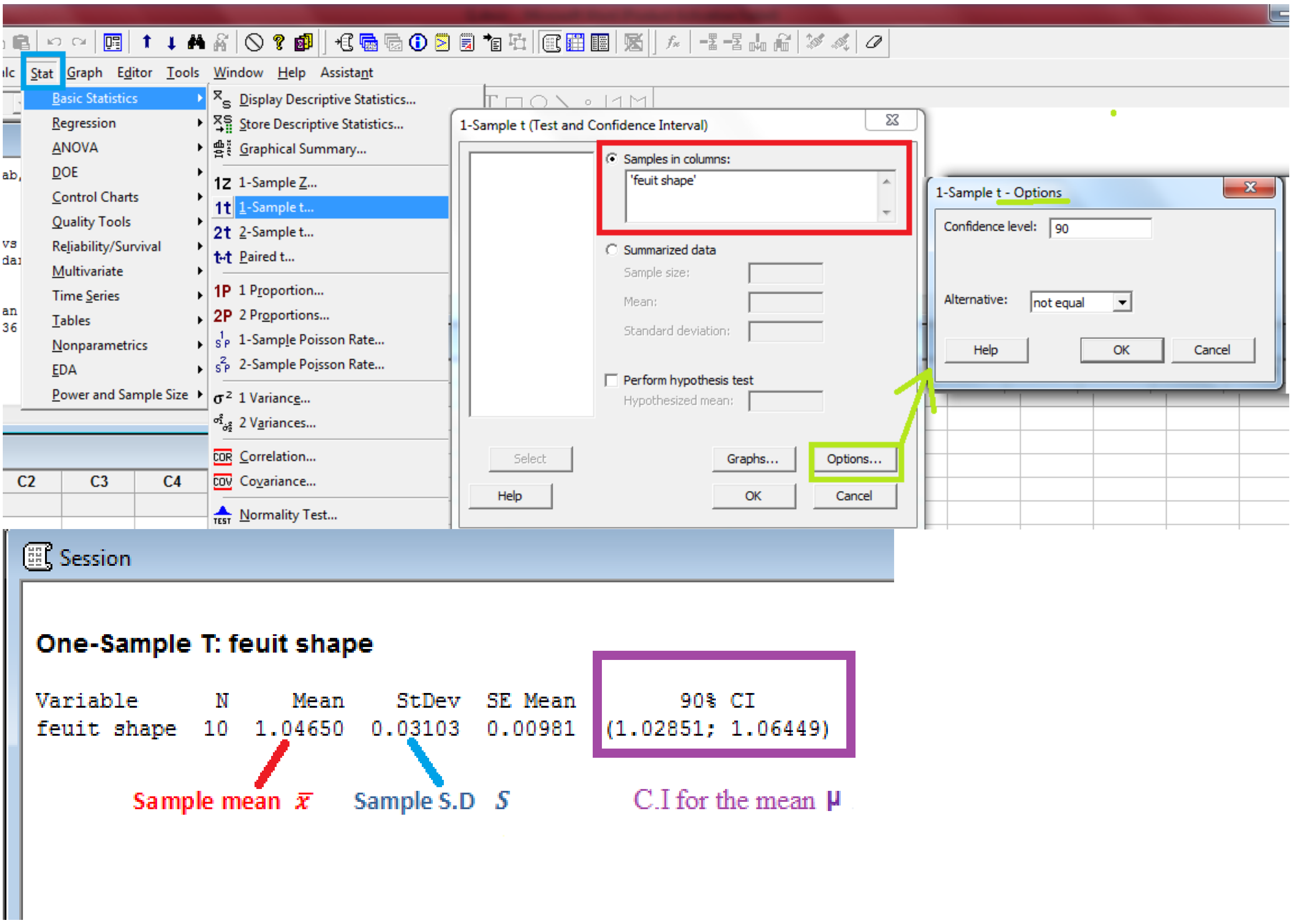
Q: For a sample of 10 fruits from thirteen-year-old acidless orange trees, the fruit shape (determined as a diameter divided by height) was measured [Shaheen and Hamouda (8419b)]:

1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

Assuming that fruit shapes are approximately normally distributed, find and interpret a 90% confidence interval for the average fruit shape.

*Use the 1-sample t-test to estimate the mean of a population and compare it to a target or reference value **when you do not know the standard deviation of the population**.
Using this test, you can:
- Determine whether the mean of a group differs from a specified value.
- Calculate a range of values that is likely to include the population mean.

↓	C1
	feuit shape
1	1.066
2	1.084
3	1.076
4	1.051
5	1.059
6	1.020
7	1.035
8	1.052
9	1.046
10	0.976
11	



The 90% CI for the mean μ: (1.02851 , 1.06449)

Two-sample t-test

Q: The phosphorus content was measured for independent samples of skim and whole:

Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

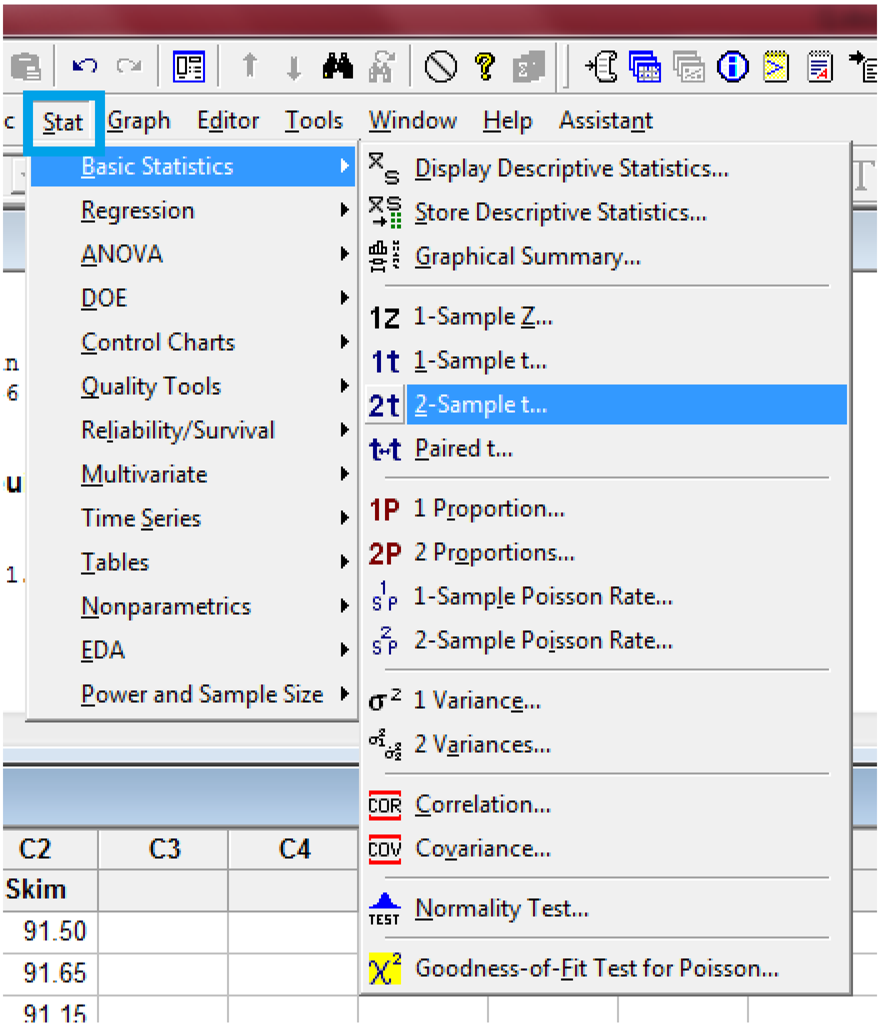
Assuming normal populations with equal variance

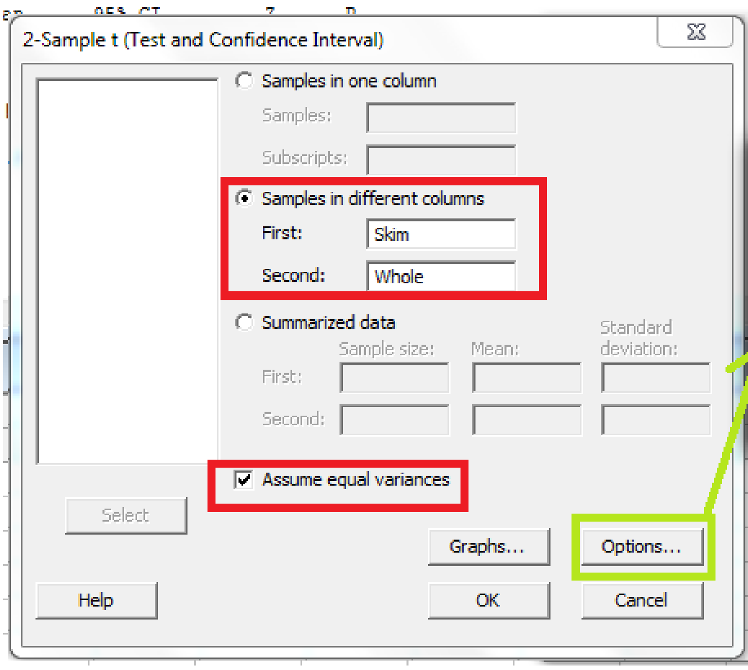
a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**

b) **Find and interpret a 99% confidence interval for the difference in average phosphorus contents of whole and skim milk .**

*Use the 2-sample t-test to **two compare between two population means**, **when the variances are unknowns**

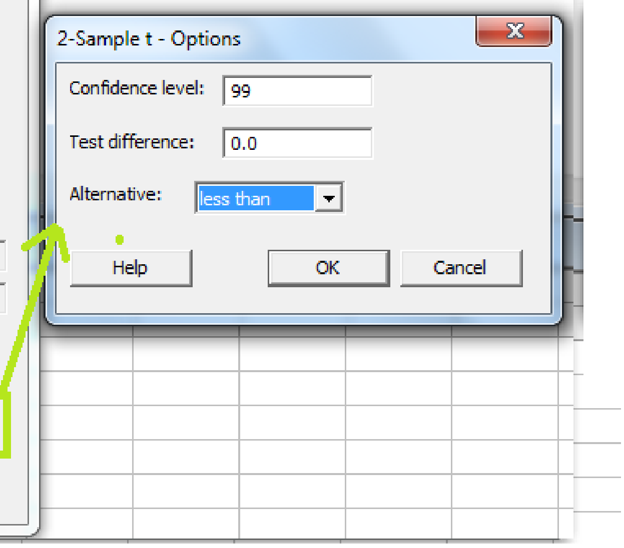
↓	C1	C2
	Whole	Skim
3	94.85	91.50
4	94.55	91.65
5	94.55	91.15
6	93.40	90.25
7	95.05	91.90
8	94.35	91.25
9	94.70	91.65
10	94.90	91.00
11		





$H_0: \mu_{Skim} \geq \mu_{Whole} \quad vs \quad H_1: \mu_{Skim} < \mu_{Whole}$

$H_0: \mu_{Skim} - \mu_{Whole} \geq 0 \quad vs \quad H_1: \mu_{Skim} - \mu_{Whole} < 0$



Two-Sample T-Test and CI: Skim; Whole

```
Two-sample T for Skim vs Whole

       N    Mean  StDev  SE Mean
Skim   10  91.340  0.483     0.15
Whole  10  94.645  0.503     0.16


Difference = mu (Skim) - mu (Whole)
Estimate for difference:  -3.305
99% upper bound for difference:  -2.742
T-Test of difference = 0 (vs <): T-Value = -14.99  P-Value = 0.000  DF = 18
Both use Pooled StDev = 0.4931
```

$H_1: \mu_{Skim} - \mu_{Whole} < 0$

T= -14.99   p-value = 0.00   Degree of freedom=18

a)

1- **Hypothesis :**

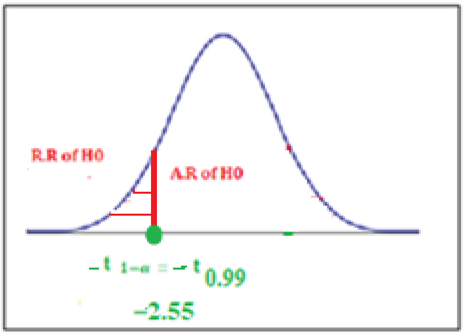
$H_0: \mu_{Skim} \geq \mu_{Whole} \quad vs \quad H_1: \mu_{Skim} < \mu_{Whole}$

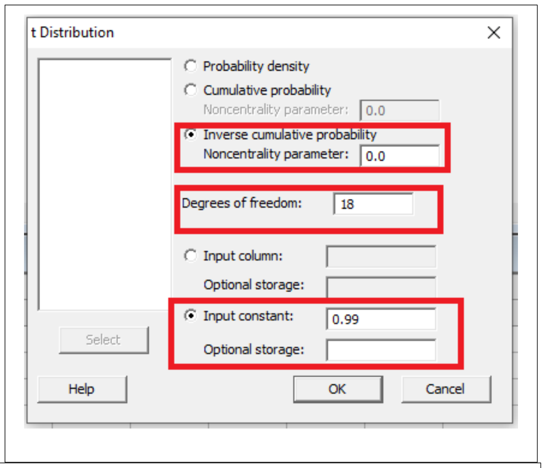
$H_0: \mu_{Skim} - \mu_{Whole} \geq 0 \quad vs \quad H_1: \mu_{Skim} - \mu_{Whole} < 0$

2- **Test statistic : T= -14.99**

3- **The critical region(s):**

Calc>> probability distributions>> t





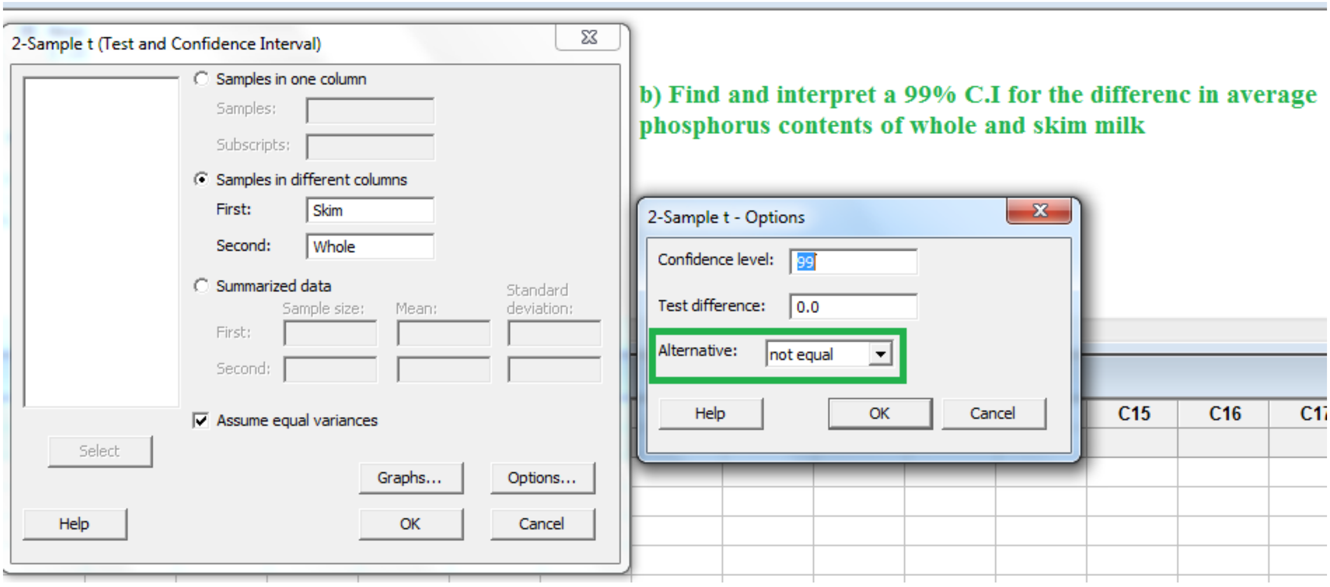
Inverse Cumulative Distribution Function

```
Student's t distribution with 18 DF

P( X ≤ x )        x
      0.99  2.55238
```

4- **Decision:**

Since p-value =0.00 < α= 0.01 . we reject H_0

b)



Session

```
        N    Mean  StDev  SE Mean
Skim   10  91.340  0.483     0.15
Whole  10  94.645  0.503     0.16


Difference = mu (Skim) - mu (Whole)
Estimate for difference:  -3.305
99% CI for difference:  (-3.940; -2.670)
T-Test of difference = 0 (vs not =): T-Value = -14.99  P-Value = 0.000  DF = 18
Both use Pooled StDev = 0.4931
```

$\mu_{Skim} - \mu_{Whole} \in (-3.940\ , -2.670\ )$

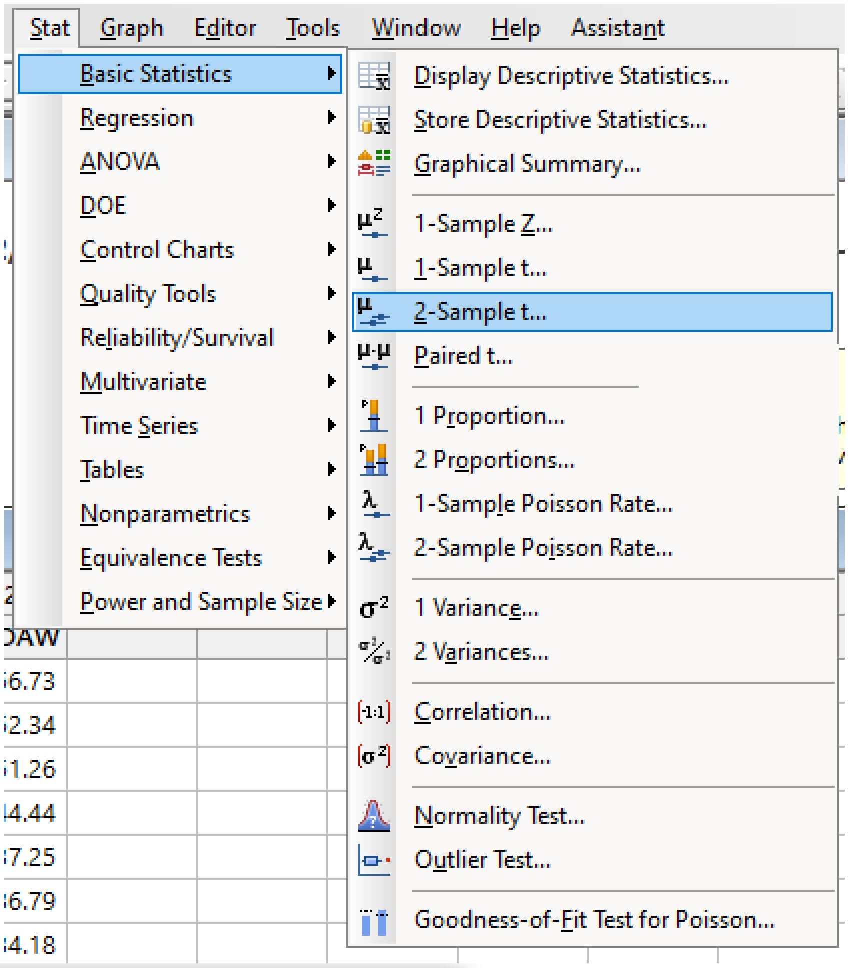
Two-sample t-test

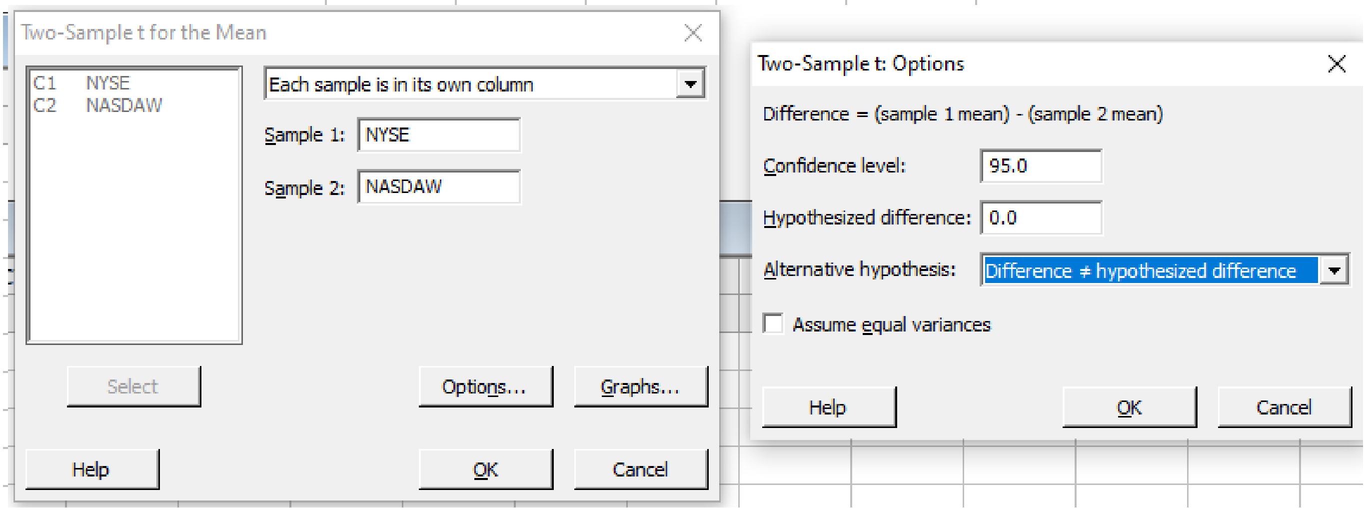
Q : The example below gives the Dividend Yields for the top ten NYSE and NASDAW stocks. Use the t-test tool to determine whether there is any indication of a difference between the means of the two different populations. α= 0.05

NYSE	NASDAW
346.55	56.73
250.55	52.34
65.48	51.26
50	44.44
48.91	37.25
43.48	36.79
42.46	34.18
39.97	30.29
33.5	29.4
32.9	28.65

Worksheet 1 ***

↓	C1	C2
	NYSE	NASDAW
1	346.55	56.73
2	250.55	52.34
3	65.48	51.26
4	50.00	44.44
5	48.91	37.25
6	43.48	36.79
7	42.46	34.18
8	39.97	30.29
9	33.50	29.40
10	32.90	28.65
11		


Stat
Graph
Editor
Tools
Window
Help
Assistant
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Time Series
Tables
Nonparametrics
Equivalence Tests
Power and Sample Size
Display Descriptive Statistics...
Store Descriptive Statistics...
Graphical Summary...
1-Sample Z...
1-Sample t...
2-Sample t...
Paired t...
1 Proportion...
2 Proportions...
1-Sample Poisson Rate...
2-Sample Poisson Rate...
1 Variance...
2 Variances...
Correlation...
Covariance...
Normality Test...
Outlier Test...
Goodness-of-Fit Test for Poisson...


Two-Sample t for the Mean
C1 NYSE
C2 NASDAW
Each sample is in its own column
Sample 1: NYSE
Sample 2: NASDAW
Select
Options...
Graphs...
Help
OK
Cancel
Two-Sample t: Options
Difference = (sample 1 mean) - (sample 2 mean)
Confidence level: 95.0
Hypothesized difference: 0.0
Alternative hypothesis: Difference ≠ hypothesized difference
Assume equal variances
Help
OK
Cancel

Two-Sample T-Test and CI: NYSE; NASDAW

```
Two-sample T for NYSE vs NASDAW

        N  Mean  StDev  SE Mean
NYSE    10    95    110       35
NASDAW  10  40.1   10.4      3.3


Difference = μ (NYSE) - μ (NASDAW)
Estimate for difference:  55.2
95% CI for difference:  (-23.7; 134.2)
T-Test of difference = 0 (vs ≠): T-Value = 1.58  P-Value = 0.148  DF = 9
```

T-test   p-value

$H_0: \mu_1 = \mu_2 \quad vs \quad H_1: \mu_1 \neq \mu_2$

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

a)

1- **Hypothesis :**

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_1: \mu_1 \neq \mu_2$$

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

2- **Test statistic :** T= 1.58

3- **Decision:**

Since p-value =0.148 > α= 0.05 . we can not reject H_0 that there is no significant difference in the means of each sample.

Paired-sample t-test

Q : In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

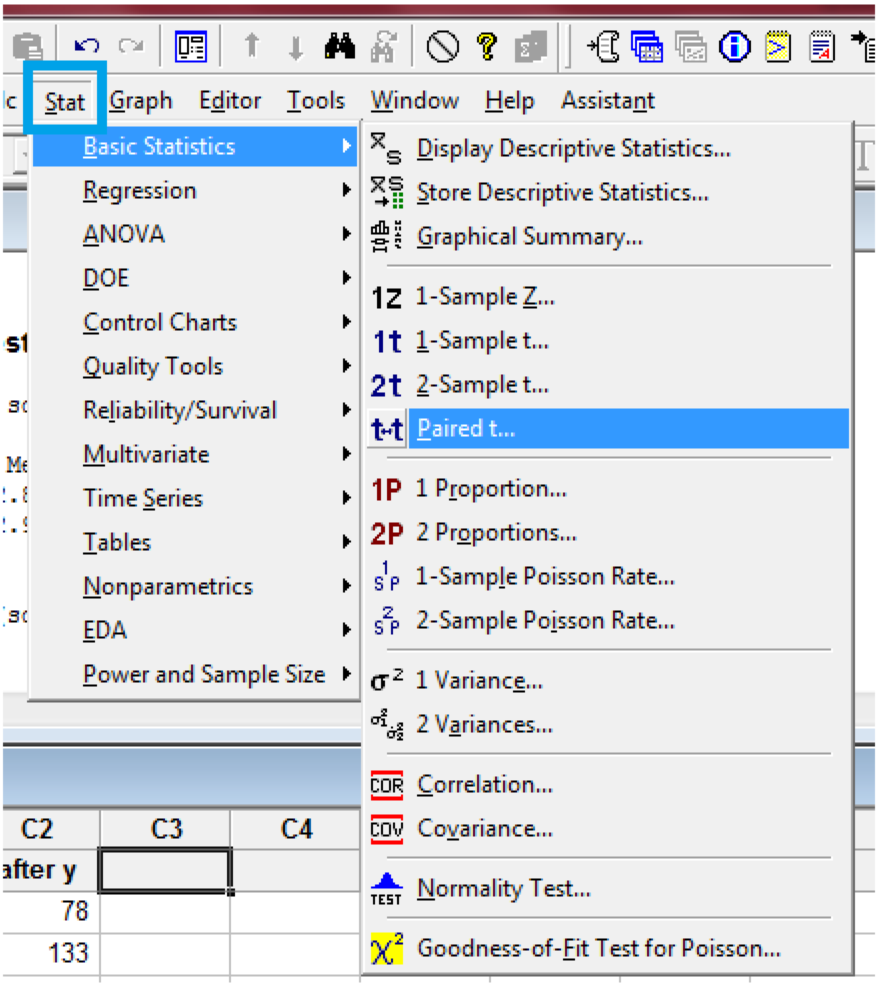
Before surgery (X)	**148**	**154**	**107**	**119**	**102**	**137**	**122**	**140**	**140**	**117**
After surgery (Y)	**78**	**133**	**80**	**70**	**70**	**63**	**81**	**60**	**85**	**120**

We assume that the data comes from normal distribution. Find :

a) **Test whether the data provide sufficient evidence to indicate a difference in the average weight before and after surgery. ($\mu_D = 0$ versus $\mu_D \neq 0$)**
b) **Find 95% confidence interval for μ_D, where μ_D is the difference in the average weight before and after surgery.**

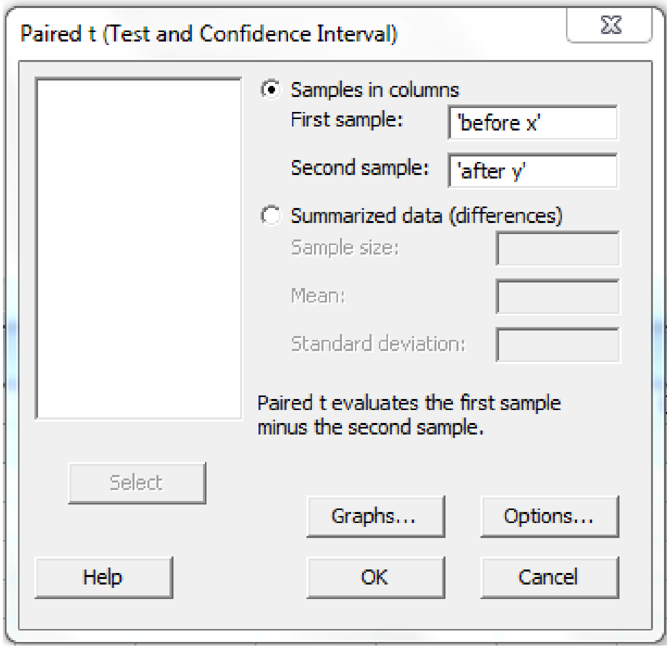
*Use the Paired-sample t-test to compare between the means of paired observations taken from the same population. This can be very useful to see the effectiveness of a treatment on some objects.

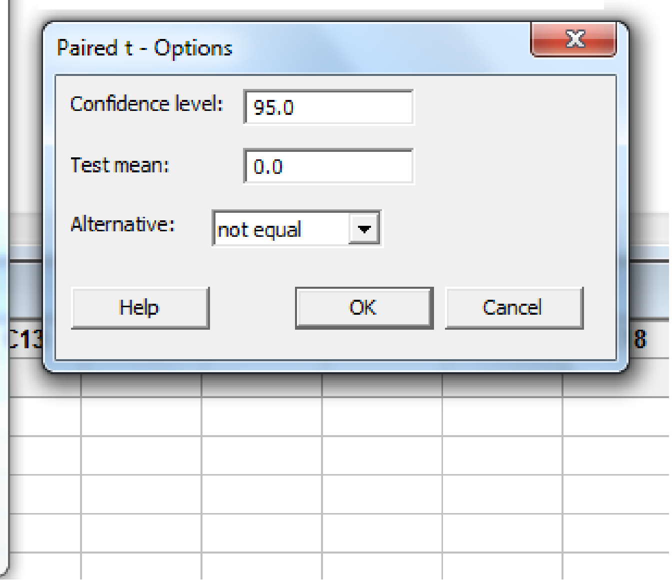
↓	C1	C2
	before x	after y
1	148	78
2	154	133
3	107	80
4	119	70
5	102	70
6	137	63
7	122	81
8	140	60
9	140	85
10	117	120
11		



$$H_0: \mu_D = 0 \quad vs \quad H_1: \mu_{D1} \neq 0$$

$$H_0: \mu_x - \mu_y = 0 \quad vs \quad H_1: \mu_x - \mu_y \neq 0$$





Paired T-Test and CI: before x; after y

```
Paired T for before x - after y

             N    Mean  StDev  SE Mean
before x    10  128.60  17.63     5.57
after y     10   84.00  23.96     7.58
Difference  10   44.60  26.23     8.30

95% CI for mean difference: (25.83; 63.37)
T-Test of mean difference = 0 (vs not = 0): T-Value = 5.38  P-Value = 0.000
```

$H_0: \mu_D = 0 \quad vs \quad H_1: \mu_{D1} \neq 0$

$H_0: \mu_x - \mu_y = 0 \quad vs \quad H_1: \mu_x - \mu_y \neq 0$

a)

1- **Hypothesis:**

$\mu_D = 0 \quad \text{vs} \quad \mu_D \neq 0$

2- **Test Statistic :**

T= 5.38

3- **Decision:**

Since p-value =0.00 < α= 0.05 . we reject H_0

b)

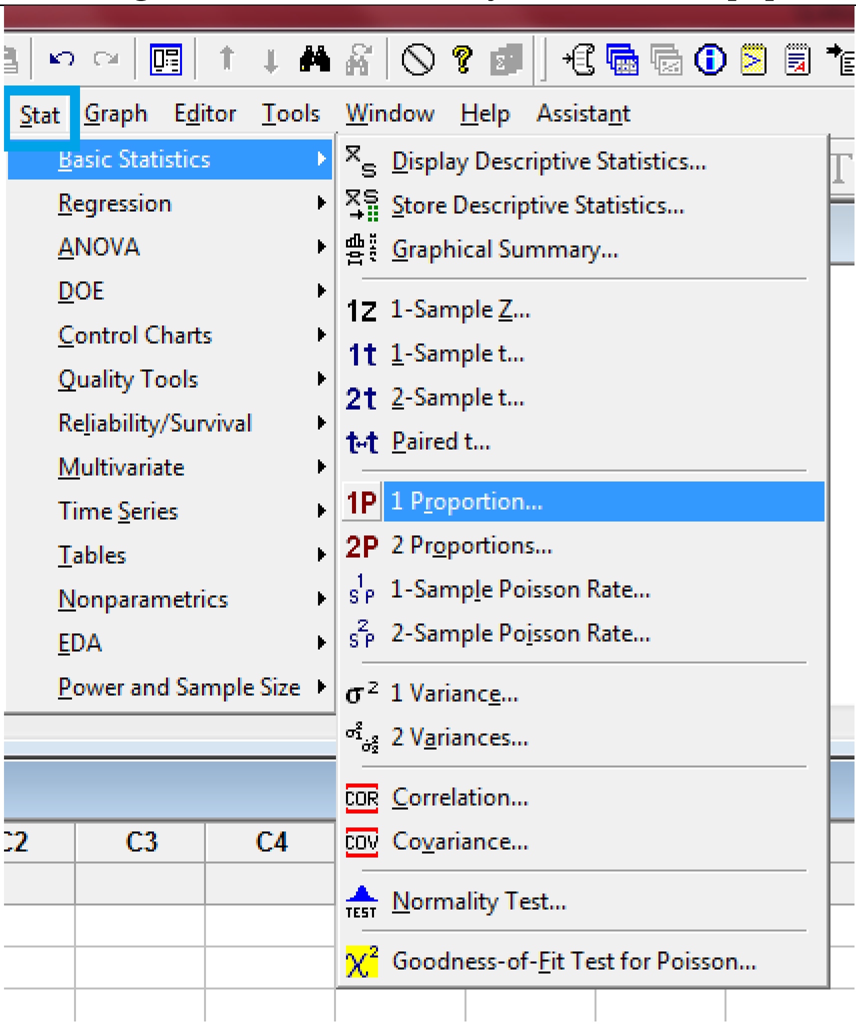
$\mu_D \in (25.83 \ , \ 63.37 \ )$

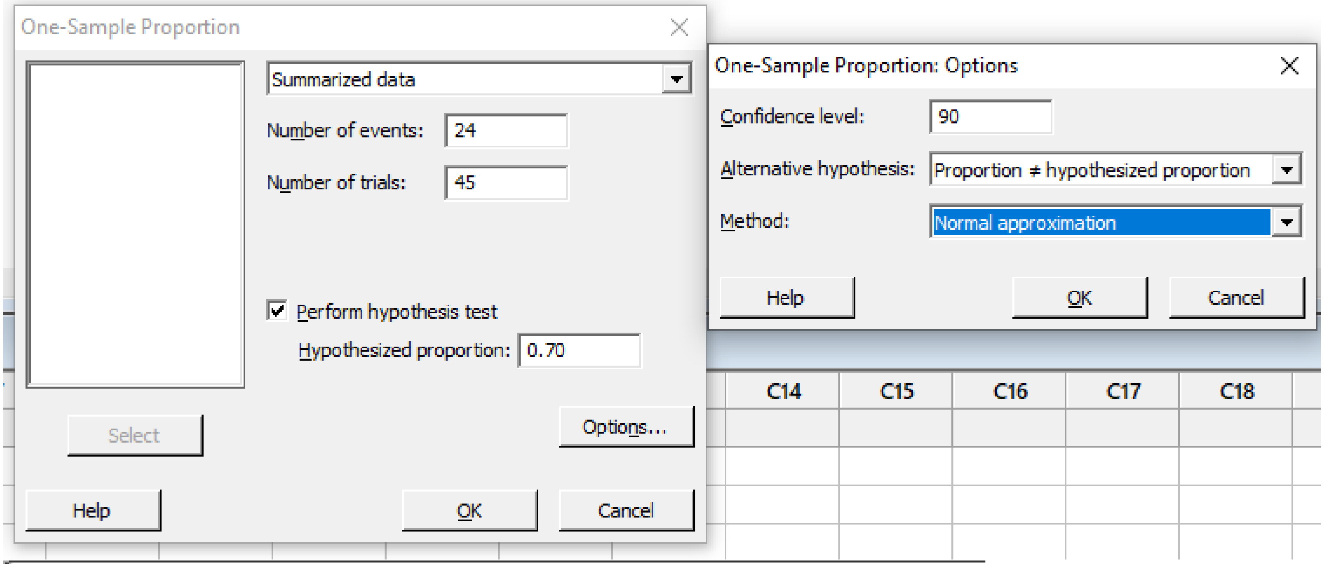
One sample proportion

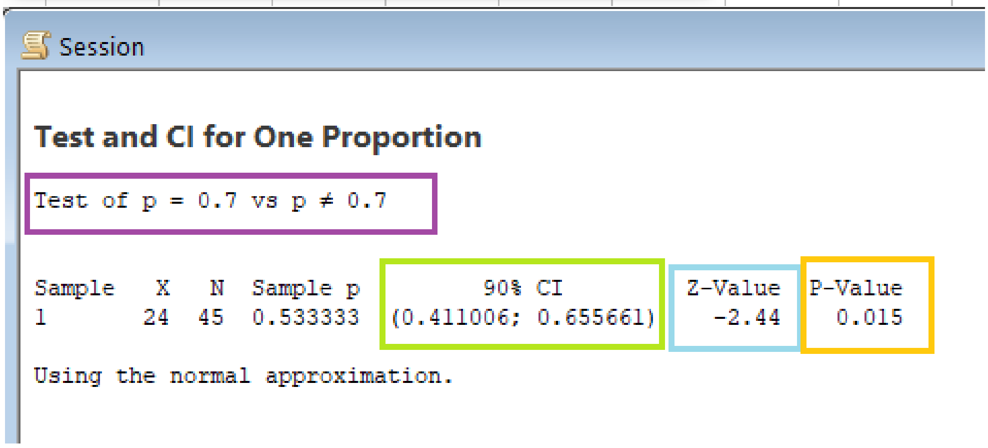
Q: A researcher was interested in the proportion of females in the population of all patients visiting a certain clinic. The researcher claims that 70% of all patients in this population are females.

a) **Would you agree with this claim if a random survey shows that 24 out of 45 patients are females? α=0.1**
b) **Find a 90% confidence interval for the true proportion of females**

Use the 1 proportion test to **estimate the proportion of a population** and **compare it to a target or reference value.**
Using this test, you can:
Determine whether the proportion for a group differs from a specified value.
Calculate a range of values that is likely to include the population proportion.







p: event proportion

Normal approximation method is used for this analysis.

a)

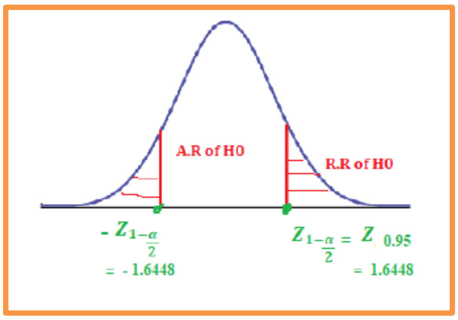
1- **Hypothesis:**

$$H_0: P = 0.70 \qquad vs \qquad H_1: P \neq 0.70$$

2- **Test statistic :**

Z= -2.44

3- **z critical** = 1.645



4- **conclusion is:**

Since p-value =0.015 < α= 0.1 . we reject the null hypothesis H_0

We do not agree with the claim stating that 70% of the population are females.

ملاحظه: في حالة فترات
الثقة يكون اختيار الفرض
الاحصائي لا يساوي

b)

$$P \in (0.411006 \ , \ 0.655661)$$

Two sample proportion

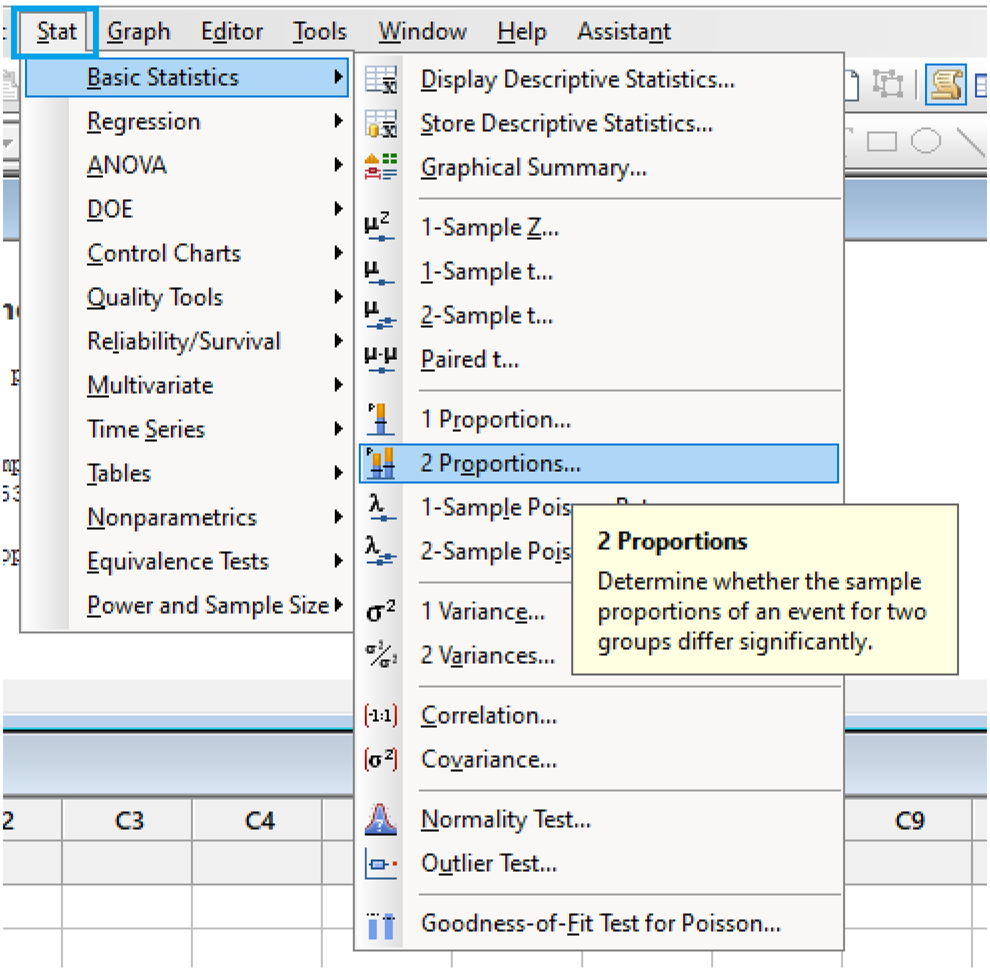
In a study about the obesity (overweight), a researcher was interested in comparing the proportion of obesity between males and females. The researcher has obtained a random sample of 150 males and another independent random sample of 200 females. The following results were obtained from this study

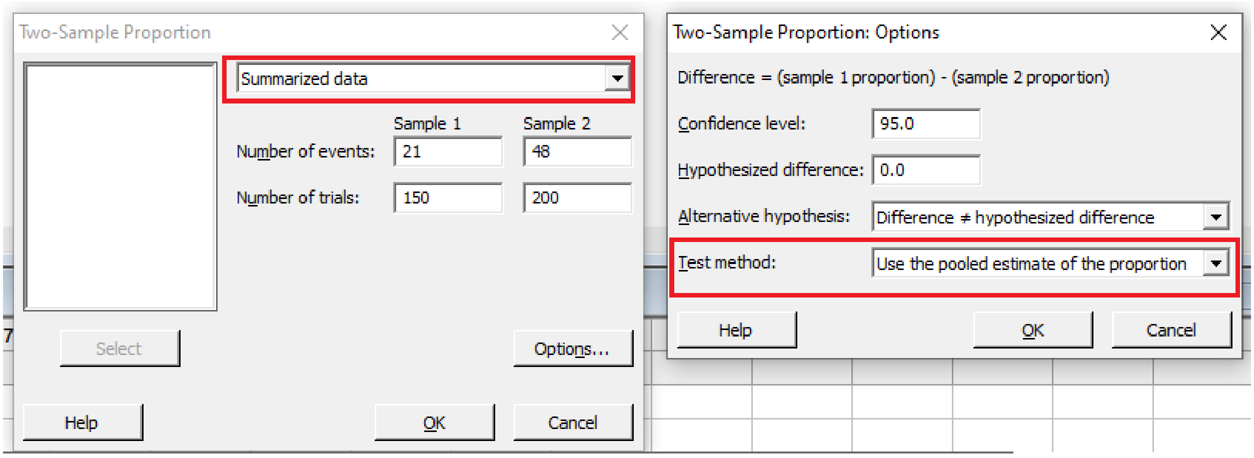
	n	Number of obese people
Males	150	21
Females	200	48

a) Can we conclude from these data that there is a difference between the proportion of obese males and proportion of obese females? Use α = 0.05.

b) Find a 95% confidence interval for the difference between the two proportions.

- Determine whether the proportions of two groups differ
- Calculate a range of values that is likely to include the difference between the population proportions





Test and CI for Two Proportions

```
Sample    X    N  Sample p
1        21  150  0.140000
2        48  200  0.240000


Difference = p (1) - p (2)
Estimate for difference:  -0.1
95% CI for difference:  (-0.181159; -0.0188408)
Test for difference = 0 (vs ≠ 0):  Z = -2.33  P-Value = 0.020

Fisher's exact test: P-Value = 0.021
```

p_1: proportion where Sample 1 = Event

p_2: proportion where Sample 2 = Event

Difference: **p_1 - p_2**

a)

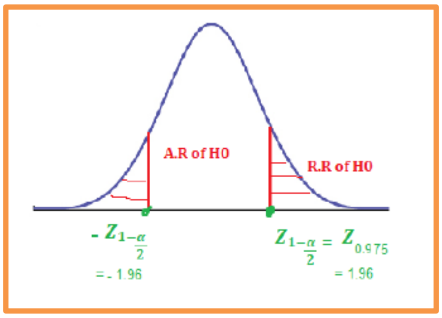
1- **Hypothesis:**

$$H_0: P_1 = P_2 \quad vs \quad H_1: P_1 \neq P_2$$

2- **Test statistic :**

Z= -2.33

3- **z critical = 1.96**



4- **conclusion is:**

Since p-value = 0.020 < α= 0.05 . we reject H_0

We conclude that there is a difference between the proportion of obese males and proportion of obese females .

b)

$$P_1 - P_2 \in (-0.181159 \;,\; -0.018841 \;)$$

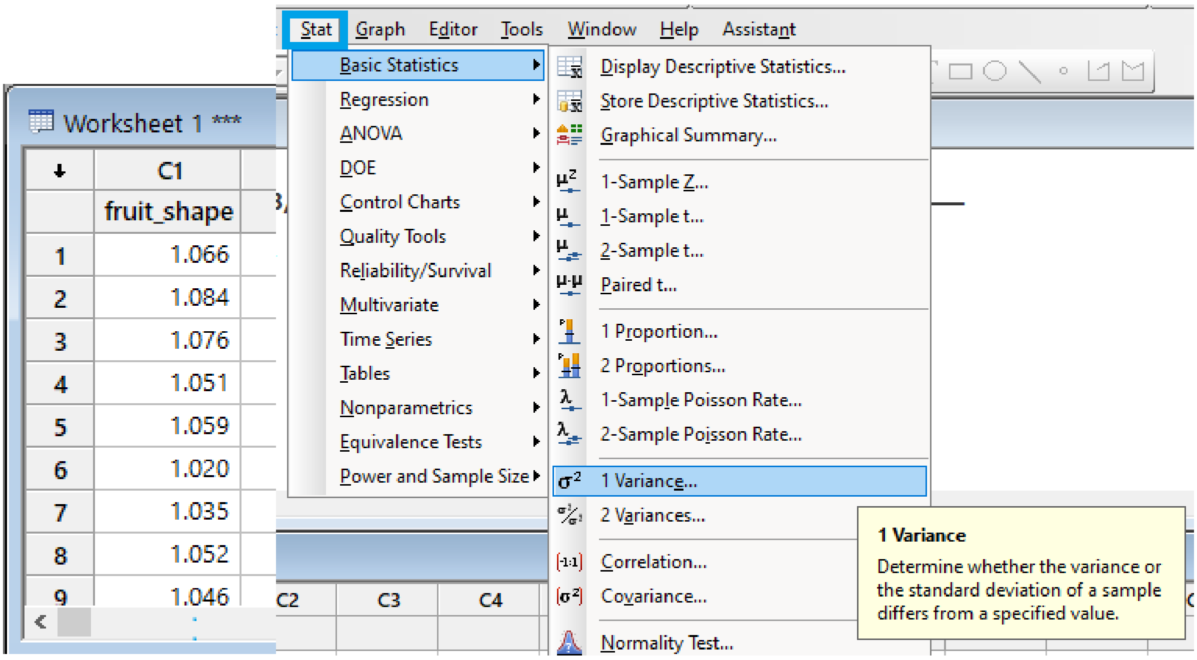
ملاحظه: في حالة فترات الثقة يكون اختيار الفرض الاحصائي لا يساوي

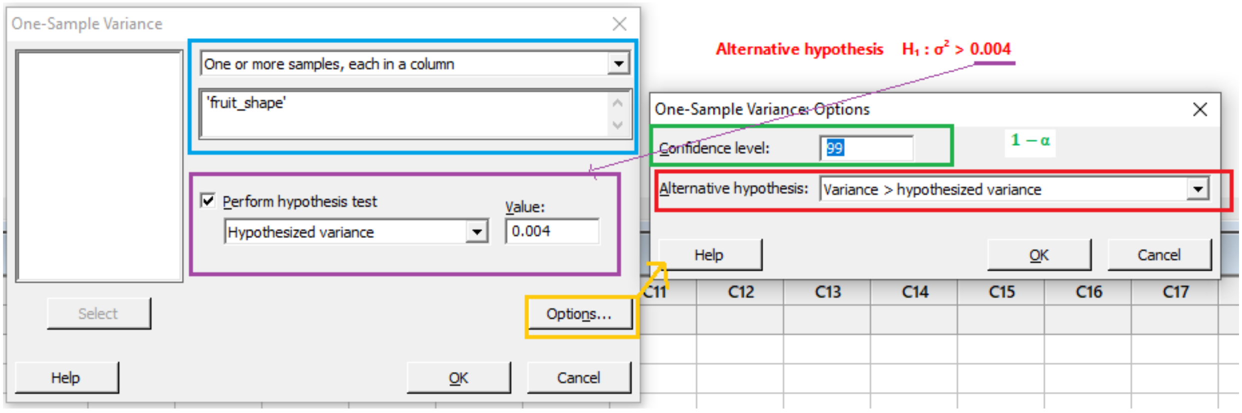
one sample variance

Q: For a sample of 10 fruits from thirteen-year-old acidless orange trees, the fruit shape (determined as a diameter divided by height) was measured [Shaheen and Hamouda (8419b)]:

1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

Assuming that fruit shapes are approximately normally distributed, Test whether the variance of fruit shape is more than 0.004, use α=0.01





Test and CI for One Variance: fruit_shape

```
Method

Null hypothesis          σ-squared = 0.004
Alternative hypothesis   σ-squared > 0.004

The chi-square method is only for the normal distribution.
The Bonett method is for any continuous distribution.


Statistics

Variable      N   StDev  Variance
fruit_shape  10  0.0310  0.000963


99% One-Sided Confidence Intervals

                         Lower Bound   Lower Bound
Variable     Method        for StDev  for Variance
fruit_shape  Chi-Square       0.0200      0.000400
             Bonett           0.0129      0.000168


Tests

                              Test
Variable     Method      Statistic  DF  P-Value
fruit_shape  Chi-Square       2.17   9    0.989
             Bonett              —   —    1.000
```

Null hypothesis $H_0: \sigma^2 = 0.004$

Alternative hypothesis $H_1: \sigma^2 > 0.004$

p-value= 0.989

$\chi^2 = 2.17$

df= n-1 =10 -1

1- **The hypothesis:**

$H_0: \sigma^2 = 0.004$ **vs** $H_1: \sigma^2 > 0.004$

2- **p-value= 0.989 > α=0.01 , we can not reject H0**

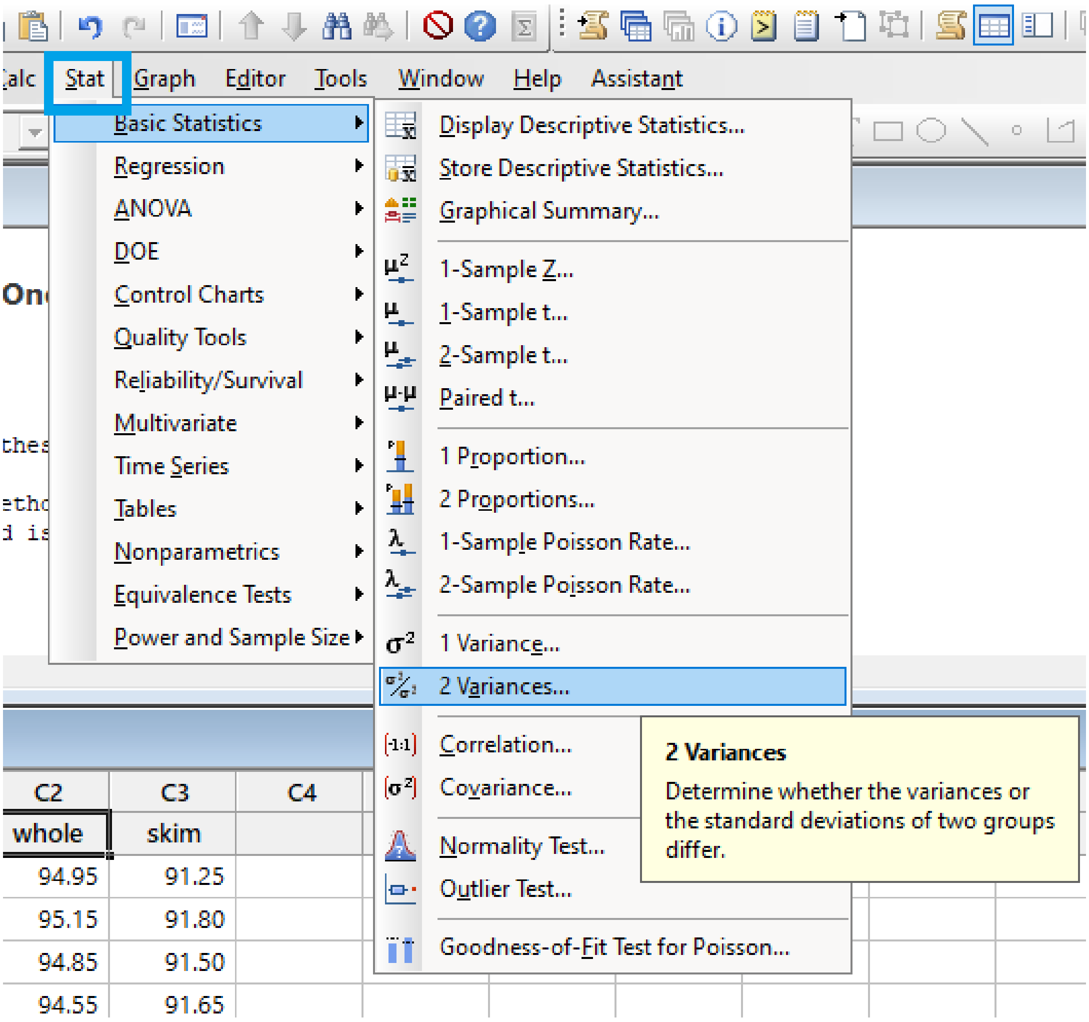
Two sample variance

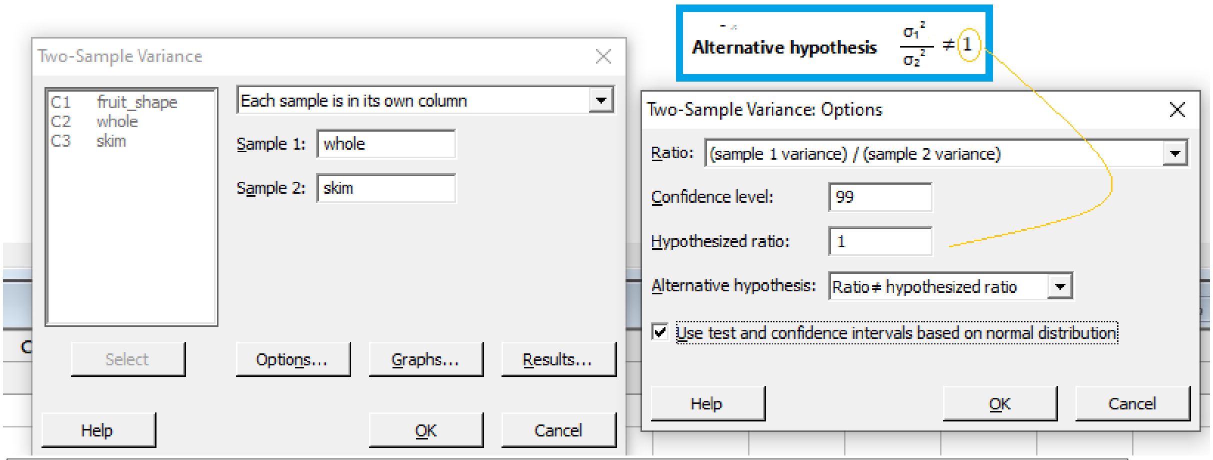
Q: The phosphorus content was measured for independent samples of skim and whole:

Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

Assuming normal populations . Test whether the variance of phosphorus content is different for whole and skim milk.

That is test whether the assumption of equal variances is valid. Use α=0.01





Test and CI for Two Variances: whole; skim

```
Method

Null hypothesis          Variance(whole) / Variance(skim) = 1
Alternative hypothesis   Variance(whole) / Variance(skim) ≠ 1
Significance level       α = 0.01
```

$H_0: \frac{\sigma_1^2}{\sigma_2^2} \neq$ vs $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

```
F method was used. This method is accurate for normal data only.


Statistics

                                        99% CI for
Variable    N  StDev  Variance          Variances
whole      10  0.503     0.253    (0.097; 1.313)
skim       10  0.483     0.233    (0.089; 1.210)

Ratio of standard deviations = 1.042
Ratio of variances = 1.085


99% Confidence Intervals

                                   CI for
          CI for StDev           Variance
Method       Ratio                 Ratio
F       (0.407; 2.664)     (0.166; 7.097)


Tests

                        Test
Method  DF1  DF2   Statistic  P-Value
F         9    9        1.08    0.905
```

p-value =0.905

F=1.08

df 1 = n1 - 1 =9

df 2 = n2 - 1 = 9

Current Worksheet: Worksheet 1

1- **Hypothesis :**

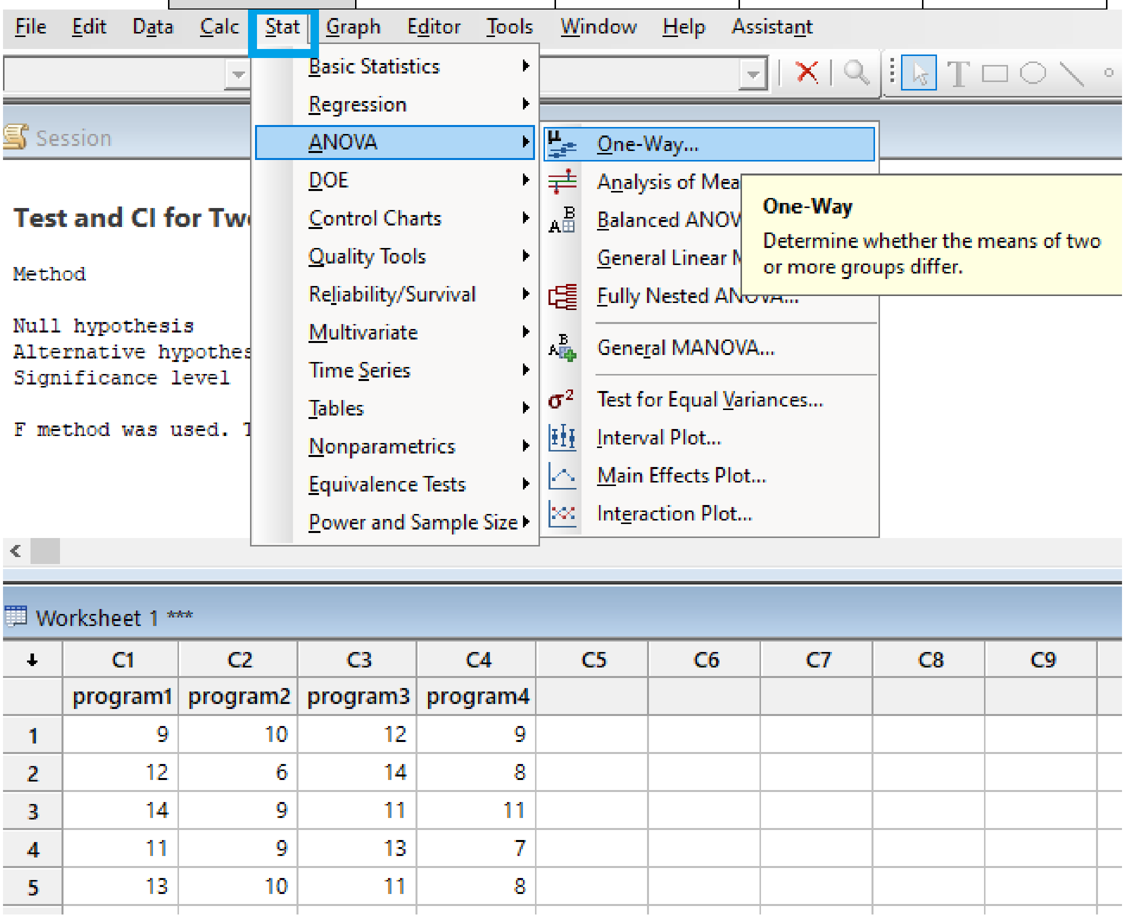
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \neq \quad vs \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

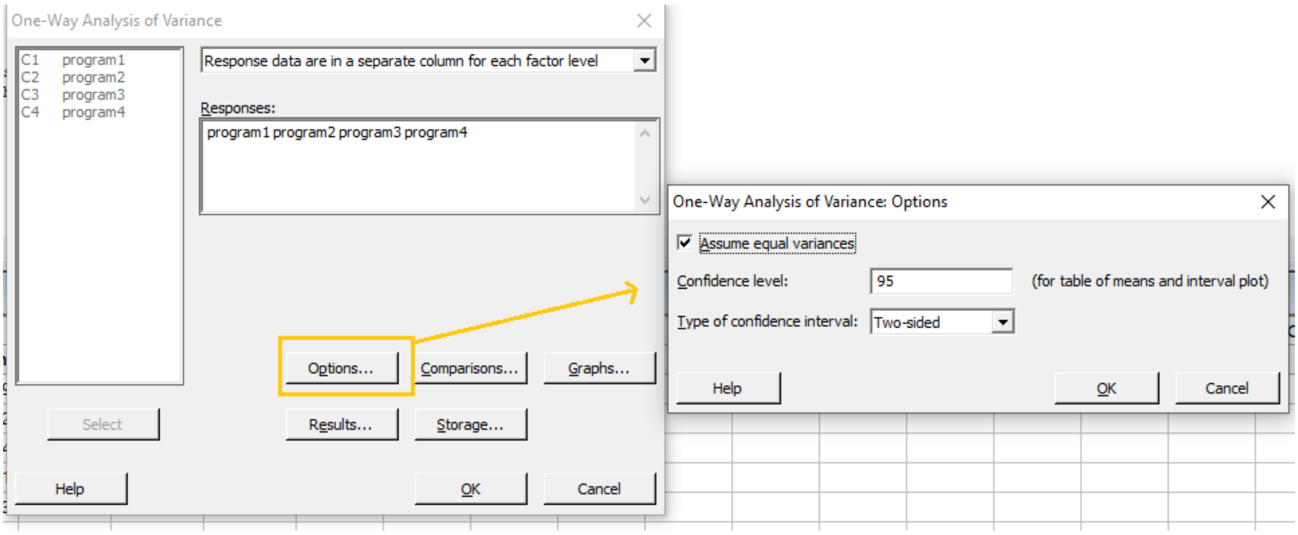
2- **P-value : 0.905 > α=0.01 , we cannot reject H0 , The variances of the two populations are equal**

ANOVA

Q: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results

Observation	Programs 1	Programs 2	Programs 3	Programs 4
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8





One-way ANOVA: program1; program2; program3; program4

Method

Null hypothesis	All means are equal
Alternative hypothesis	At least one mean is different
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Factor	4	program1; program2; program3; program4

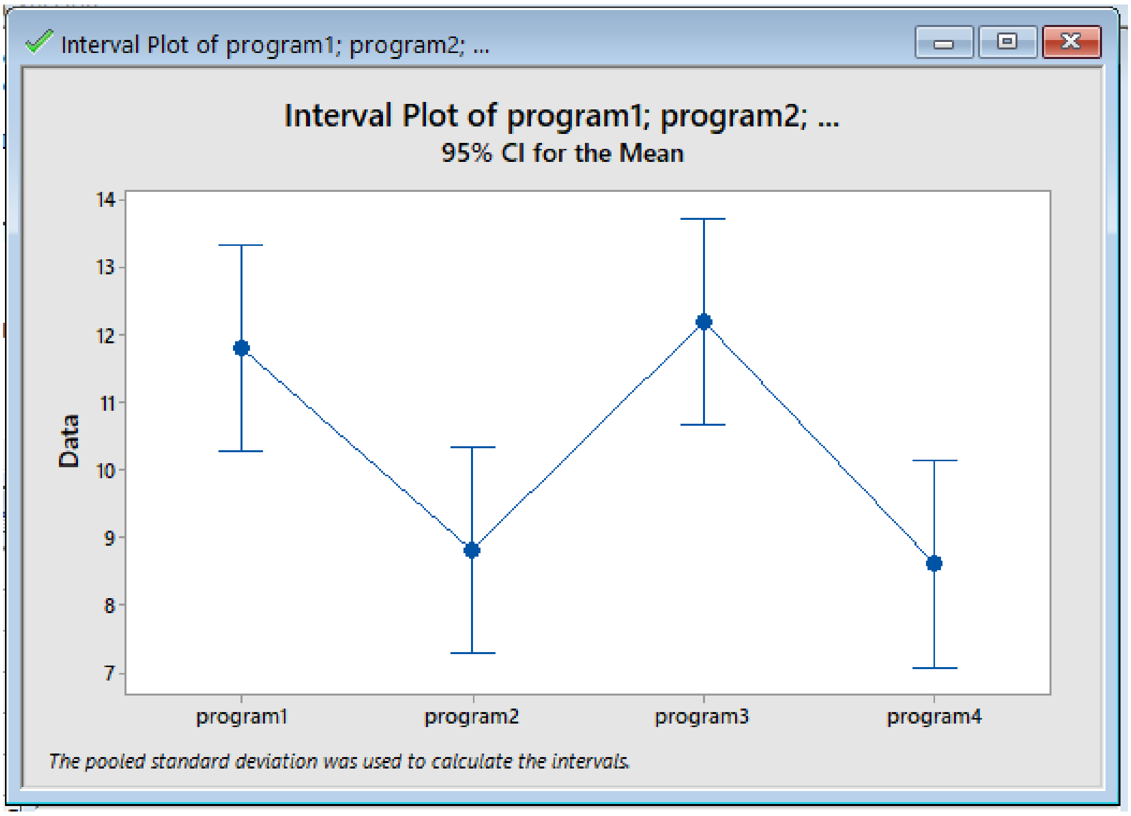
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	54.95	18.317	7.04	0.003
Error	16	41.60	2.600		
Total	19	96.55			

p-value = 0.003 < α = 0.05

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.61245	56.91%	48.83%	32.68%



1-Hypothesis :

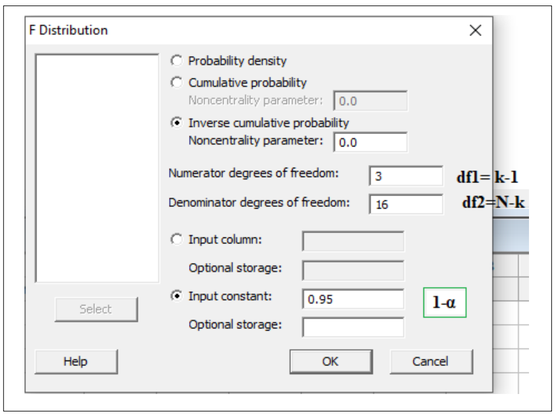
$$H_0: \mu_{program1} = \mu_{program2} = \mu_{program3} = \mu_{program4}$$

$$H_1: at\ least\ one\ mean\ is\ diffrenet$$

2- Test statistic :

F = 7.04

3- p-value = 0.003 < α=0.05 , Reject $H_0: \mu_{program1} = \mu_{program2} = \mu_{program3} = \mu_{program4}$



Calc>> probability distributions>>F

$$F_{critical} = F_{1-\alpha,\ df1=k-1,\ df2=N-k}$$

$$= F_{0.95,\ 3,16} = 3.288$$

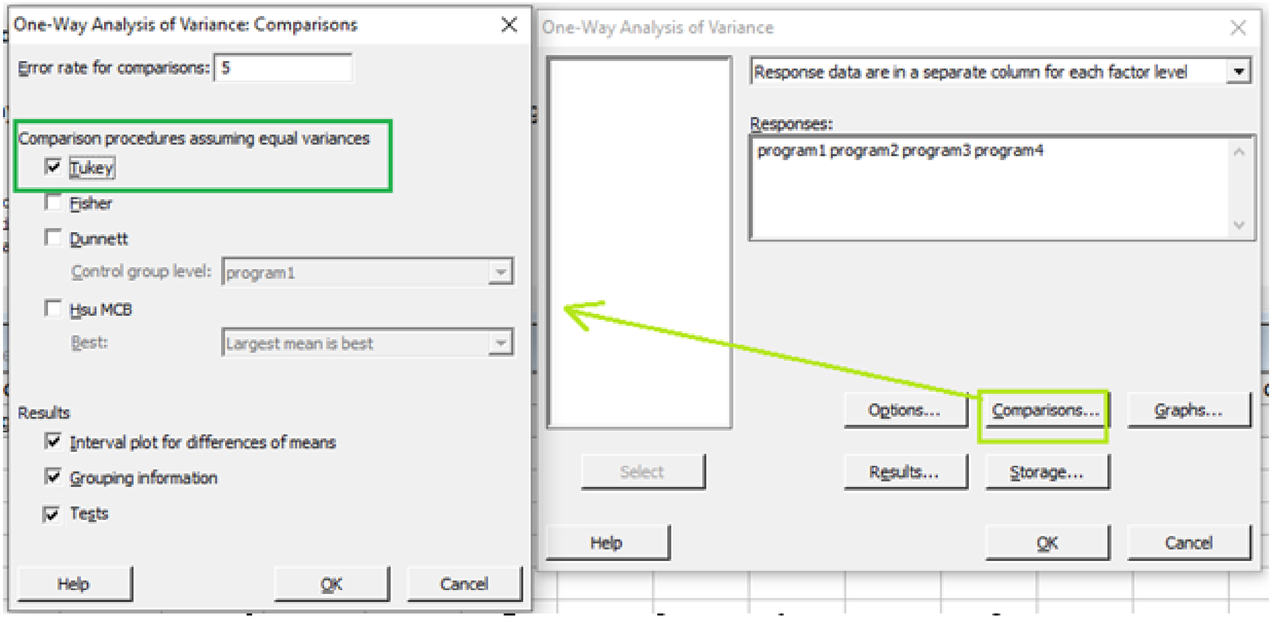
Inverse Cumulative Distribution Function

F distribution with 3 DF in numerator and 16 DF in denominator

P(X ≤ x)	x
0.95	3.23887

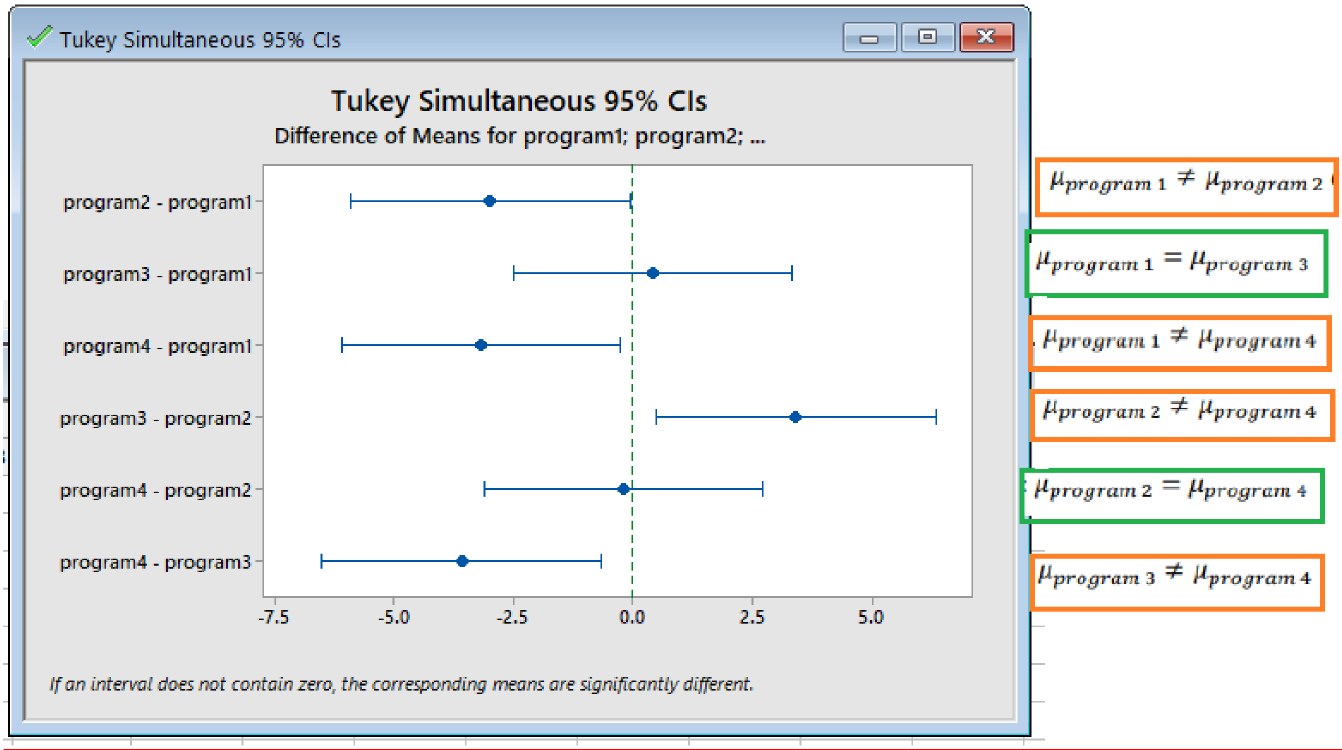
now we use Tukey test to determine which means different

Stat > ANOVA > One-Way



Tukey Pairwise Comparisons

```
Grouping Information Using the Tukey Method and 95% Confidence

Factor    N    Mean  Grouping
program3  5  12.200  A
program1  5  11.800  A
program2  5   8.800    B
program4  5   8.600    B

Means that do not share a letter are significantly different.


Tukey Simultaneous Tests for Differences of Means

                      Difference      SE of                                 Adjusted
Difference of Levels    of Means  Difference       95% CI        T-Value    P-Value
program2 - program1        -3.00        1.02  (-5.92; -0.08)      -2.94      0.043
program3 - program1         0.40        1.02  (-2.52;  3.32)       0.39      0.979
program4 - program1        -3.20        1.02  (-6.12; -0.28)      -3.14      0.029
program3 - program2         3.40        1.02  ( 0.48;  6.32)       3.33      0.020
program4 - program2        -0.20        1.02  (-3.12;  2.72)      -0.20      0.997
program4 - program3        -3.60        1.02  (-6.52; -0.68)      -3.53      0.013

Individual confidence level = 98.87%
```



Chi-square

Q2: What is the relationship between the gender of the students and the assignment of a Pass or No Pass test grade? (Pass = score 70 or above)

	Pass	No pass	Row Totals
Males	12	3	15
Females	13	2	15
Column Totals	25	5	30

1-Hypothesis :

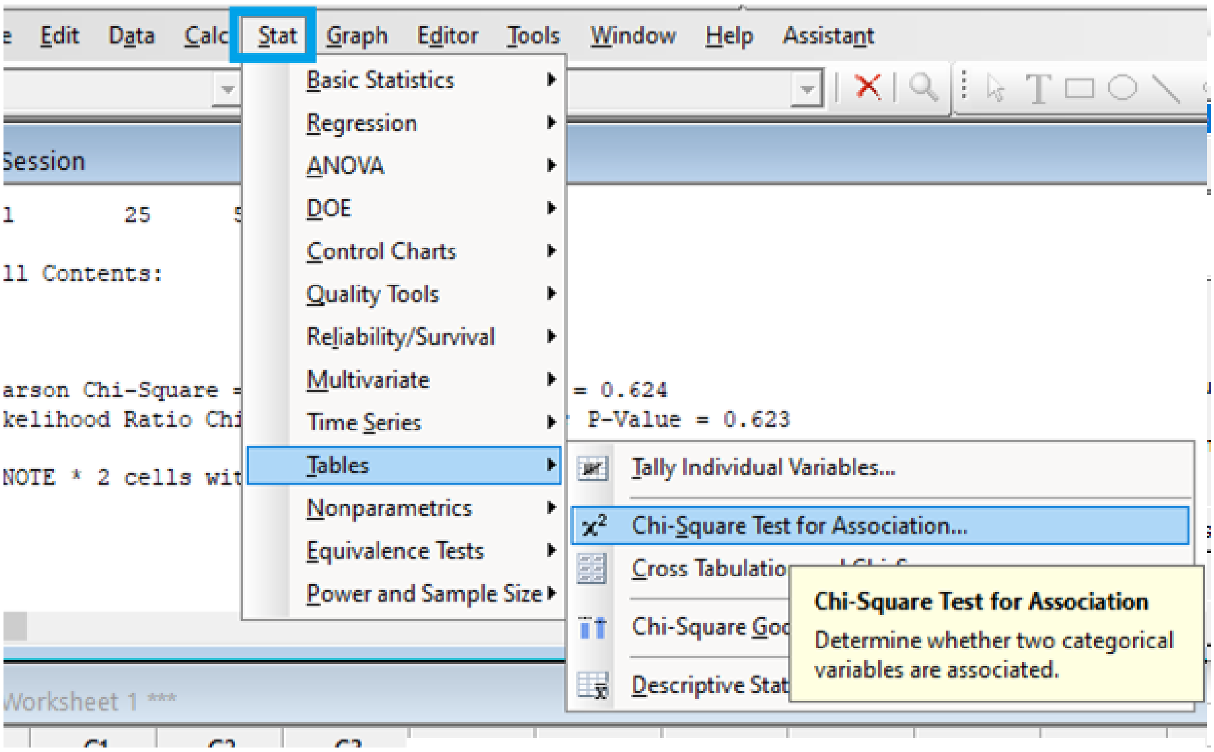
H_0: **the gender of the students is independent of pass or no pass test grade**

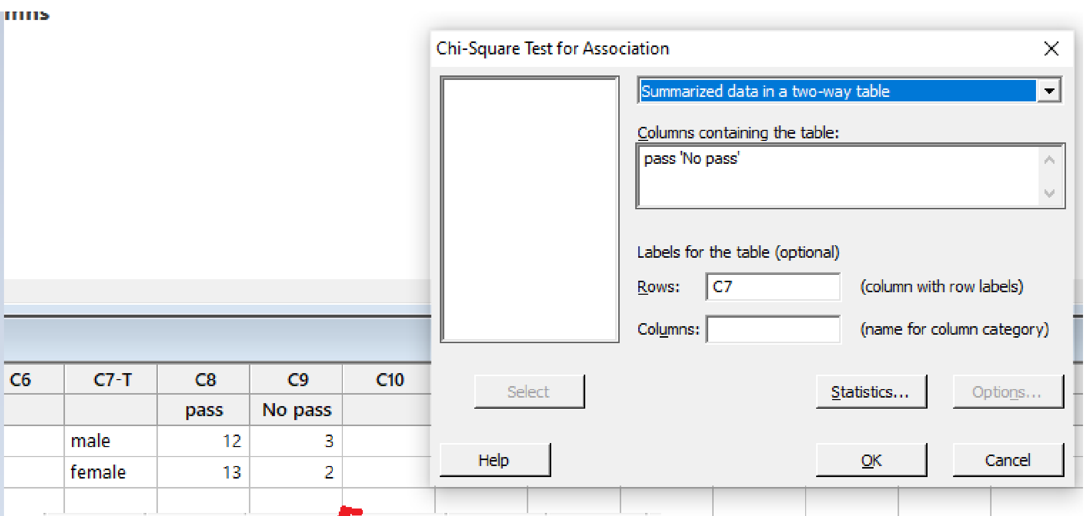
H_1: **the gender of the students is not independent of pass or no pass test grade**

2- **Test statistic :** $\chi^2 = 0.240$

3- **p-value =0.624 >** α**=0.05 , we Accept H0**

4-





	Pass	No pass	Row Totals
Males	12	3	15
Females	13	2	15
Column Totals	25	5	30

Chi-Square Test for Association: C7; Worksheet columns

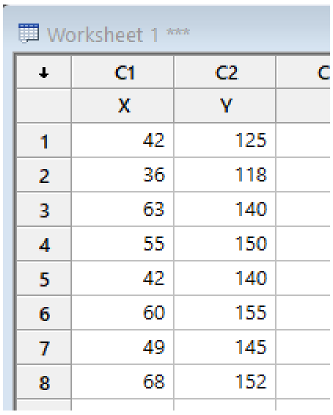
```
Rows: C7   Columns: Worksheet columns

            pass  No pass  All

male          12        3   15
          12.500    2.500

female        13        2   15
          12.500    2.500

All           25        5   30

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 0.240; DF = 1; P-Value = 0.624
Likelihood Ratio Chi-Square = 0.241; DF = 1; P-Value = 0.623

* NOTE * 2 cells with expected counts less than 5
```

Correlation

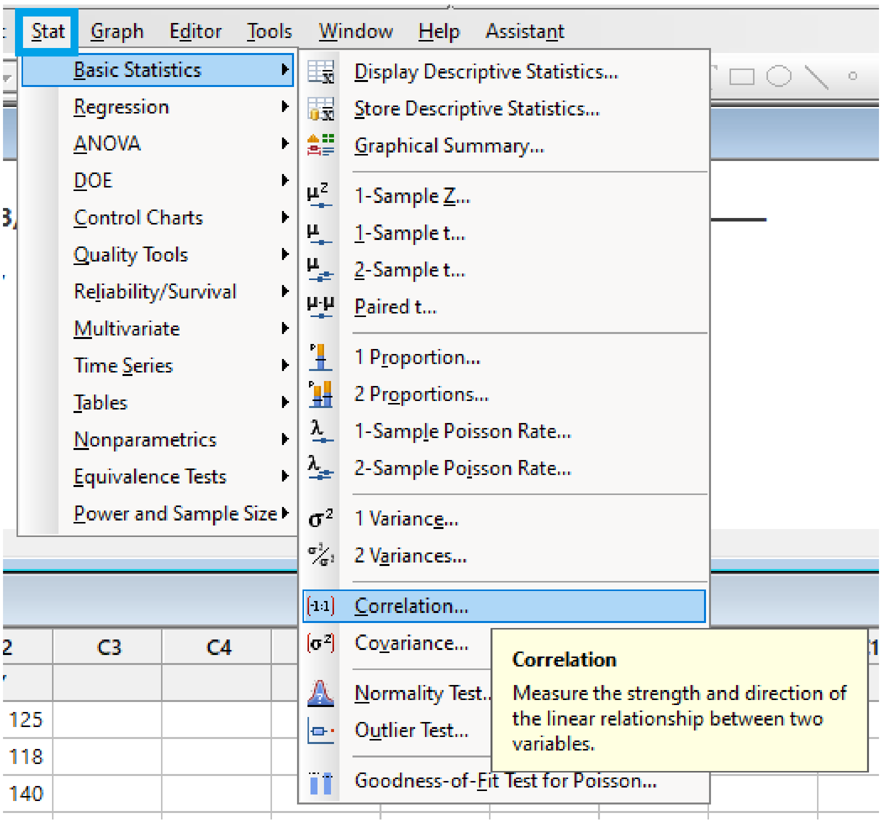
We have the table illustrates the age X and blood pressure Y for eight female.

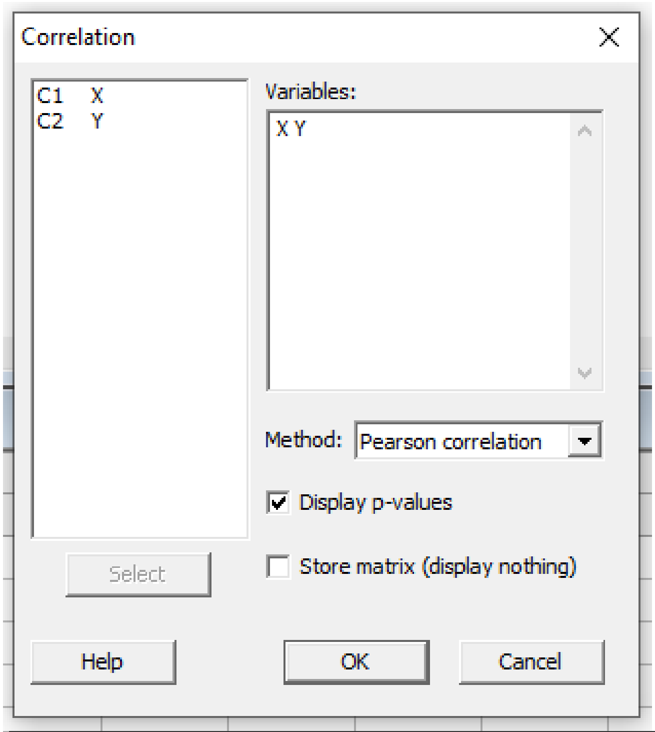
X	42	36	63	55	42	60	49	68
Y	125	118	140	150	140	155	145	152

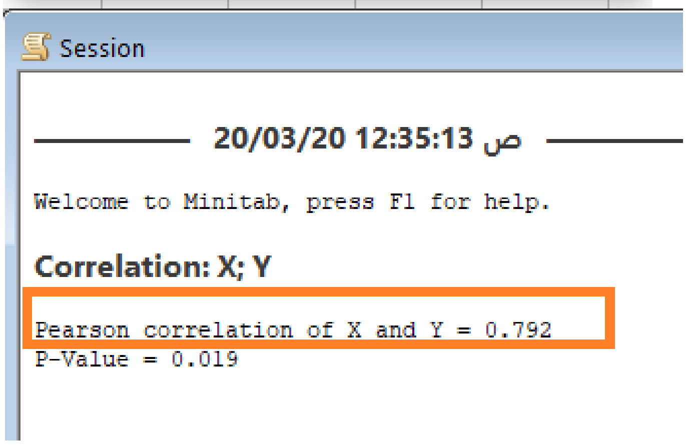
Find the correlation coefficient between x and y



↓	C1	C2
	X	Y
1	42	125
2	36	118
3	63	140
4	55	150
5	42	140
6	60	155
7	49	145
8	68	152







r = 0.792 positive correlation

Regression

Ten Corvettes between 1 and 6 years old were randomly selected from last year's sales records in Virginia Beach, Virginia. The following data were obtained, where x denotes age, in years, and y denotes sales price, in hundreds of dollars.

X	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

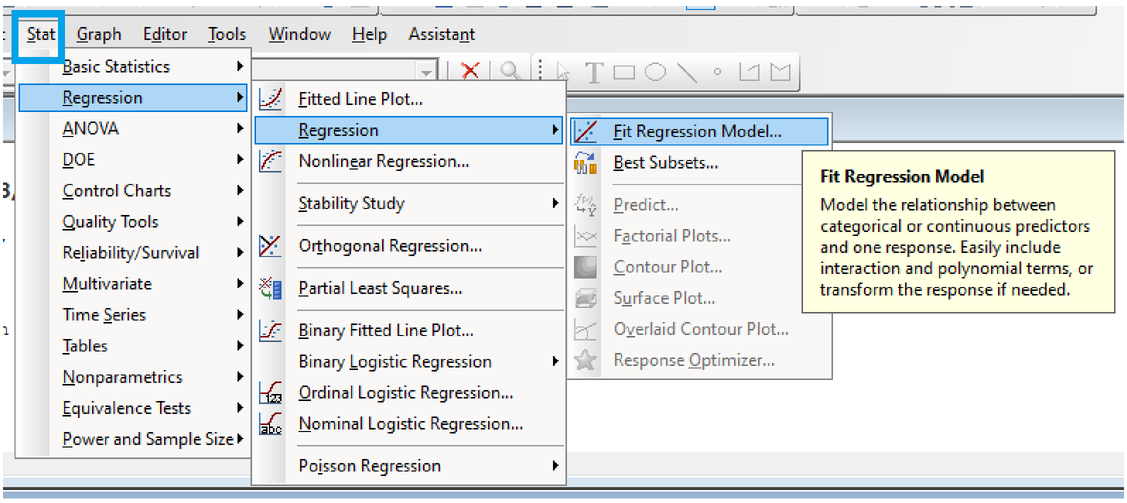
a) Determine the regression equation for the data.

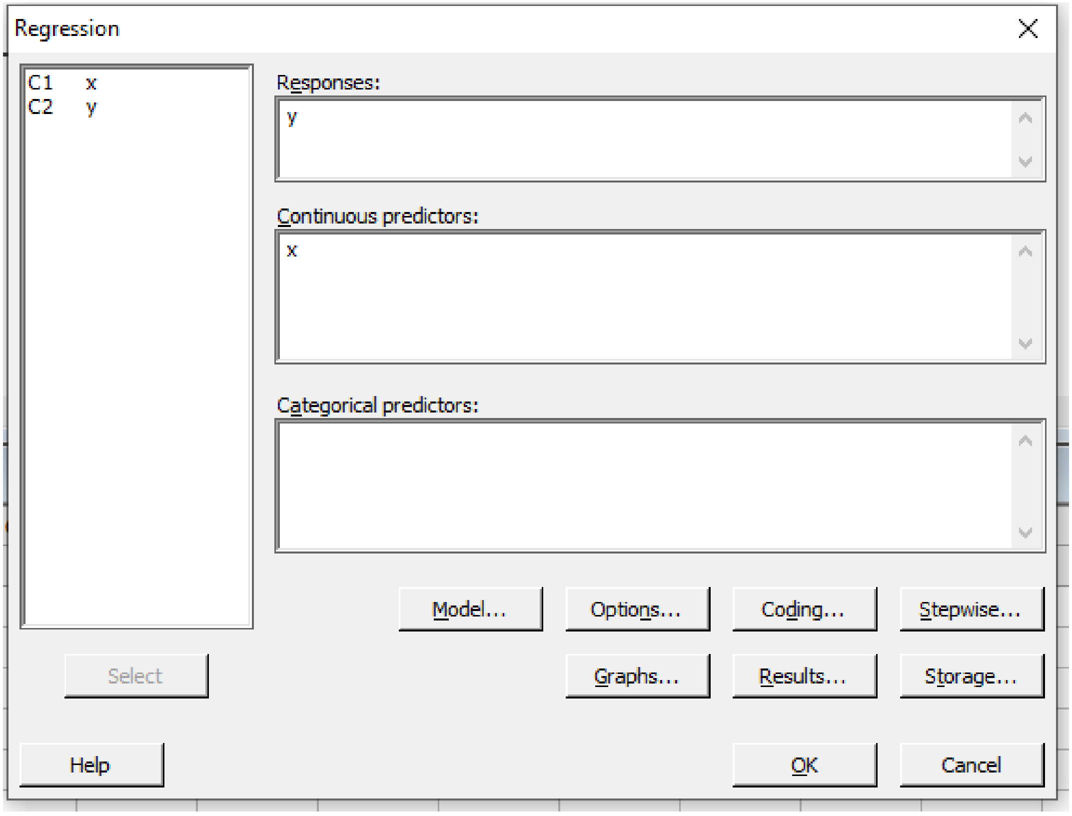
b) Compute and interpret the coefficient of determination, r^2.

c) Obtain a point estimate for the mean sales price of all 4-year-old Corvettes.

(Ans b) $R^2 = 0.9368 \rightarrow 93.68\ \%$ *of the variation in y data is explained by x*)

(Ans c) $\hat{y} = 291.6 - 27.90(4) = 180$)




Regression
C1 x
C2 y
Responses:
y
Continuous predictors:
x
Categorical predictors:
Model...
Options...
Coding...
Stepwise...
Select
Graphs...
Results...
Storage...
Help
OK
Cancel

Regression Analysis: y versus x

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	24057.9	24057.9	118.53	0.000
x	1	24057.9	24057.9	118.53	0.000
Error	8	1623.7	203.0		
Lack-of-Fit	3	132.0	44.0	0.15	0.927
Pure Error	5	1491.7	298.3		
Total	9	25681.6			

Model Summary **b)**

S	R-sq	R-sq(adj)	R-sq(pred)
14.2465	93.68%	92.89%	90.16%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	291.6	11.4	25.51	0.000	
x	-27.90	2.56	-10.89	0.000	1.00

a)

Regression Equation

y = 291.6 - 27.90 x

Results:

a) **The regression equation : $\hat{y}$ = sales price = 291.6 - 27.90 * age .**
In other words, for increasing the age by one, the sales price decreasing by 27.90 , while there is 291.60of Y does not depend on the age .

a) $r^2 = 0.9367$
The coefficient of determination is 0.9368; therefore, about 93.68% of the variation in the price data is explained by age. The regression equation appears to be very useful for making predictions since the value of r^2 is close to 1.

b) **The predicted sales price is 18000.000 dollars ($18,000.000).**

SPSS

"Statistical Package for the Social Sciences"

Q1 :In the following example, ten women and men employees in a company were asked about the educational Level, the number of years of experience, and the current salary.

Classify the data using the following variables and enter it to SPSS program :

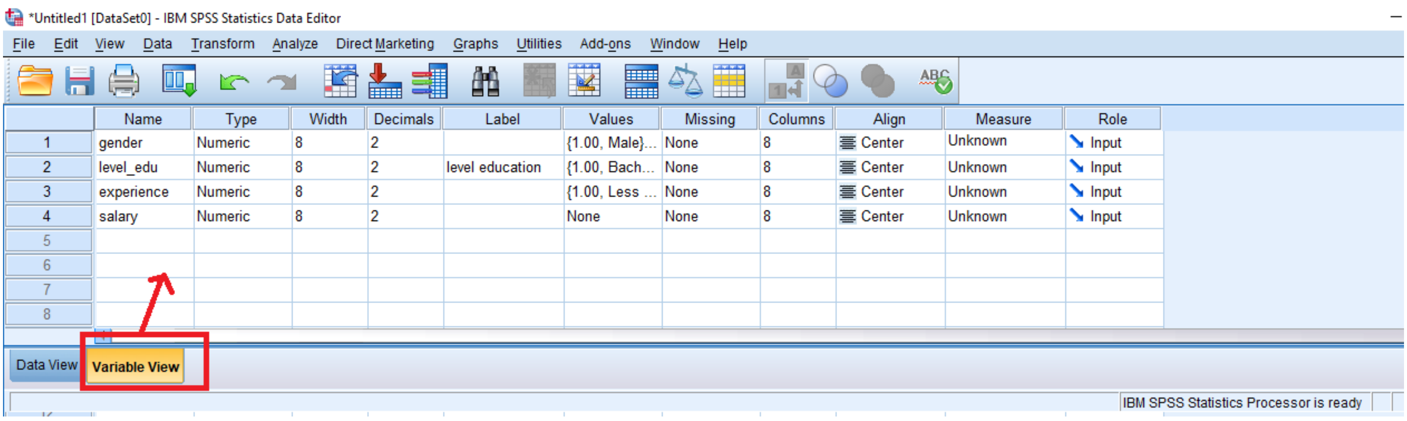
Gender: 1: Male 2: female :

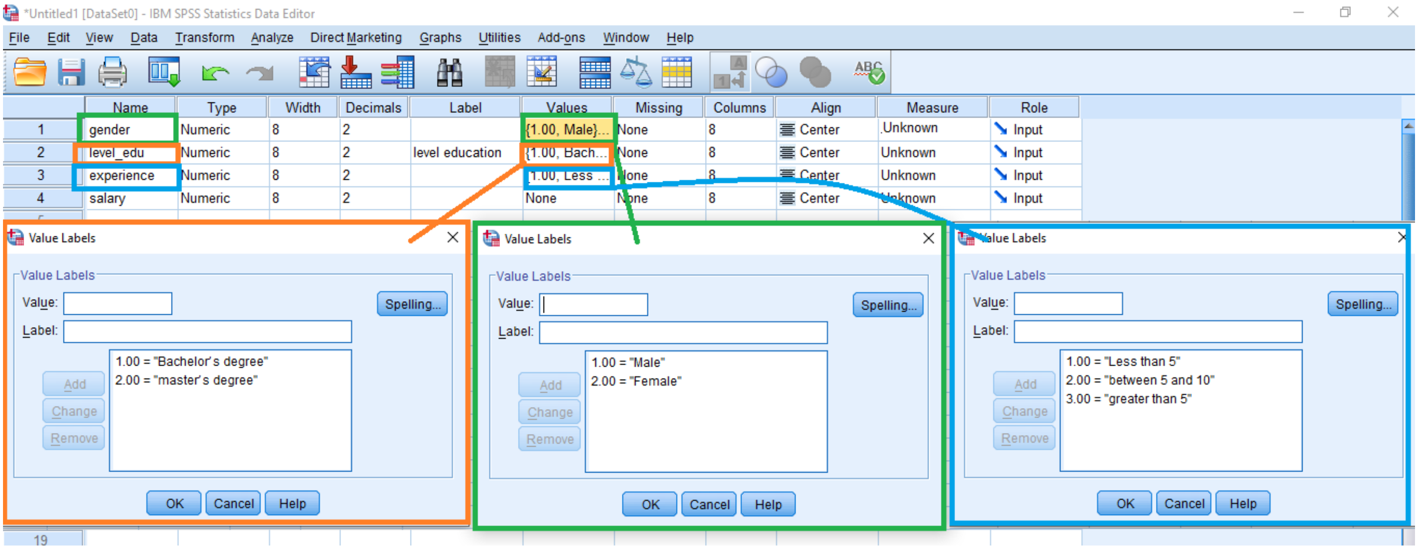
Level education 1: Bachelor's degree:1 2: master's degree

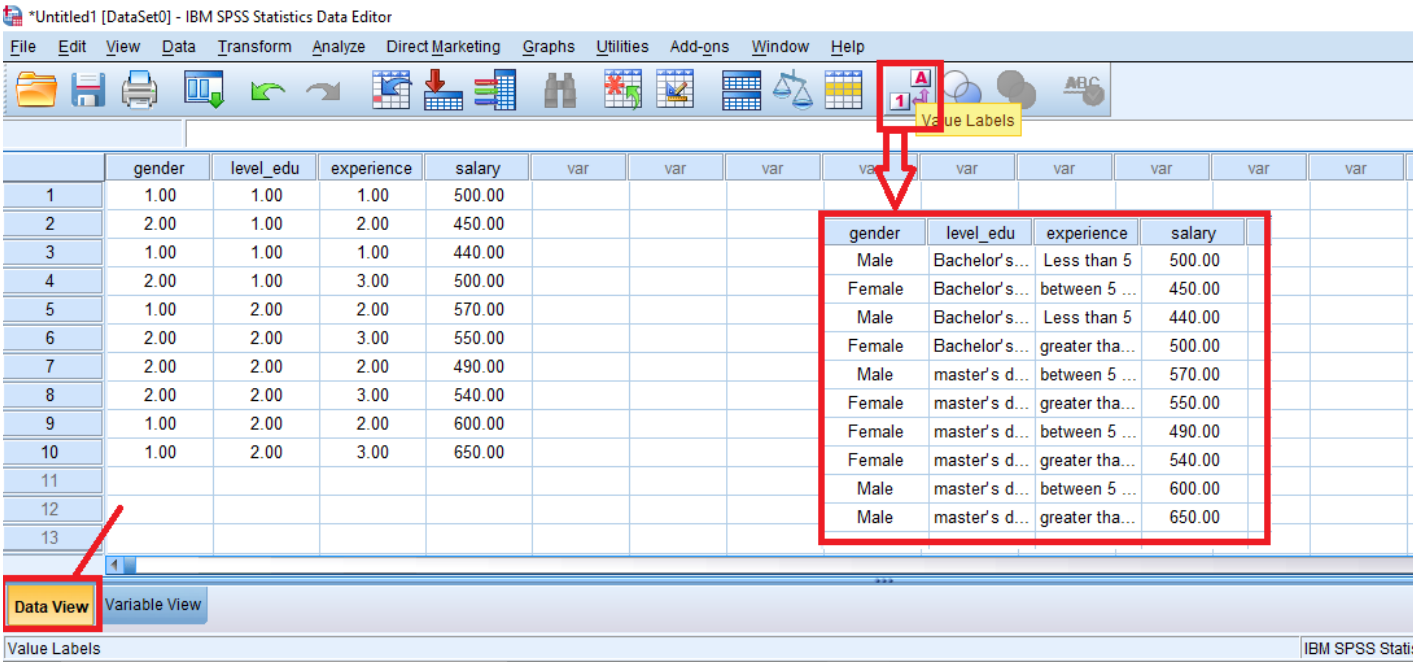
Experience 1: Less than 5 years 2: between 5 and 10 years 3: greater than 5 years

Salary

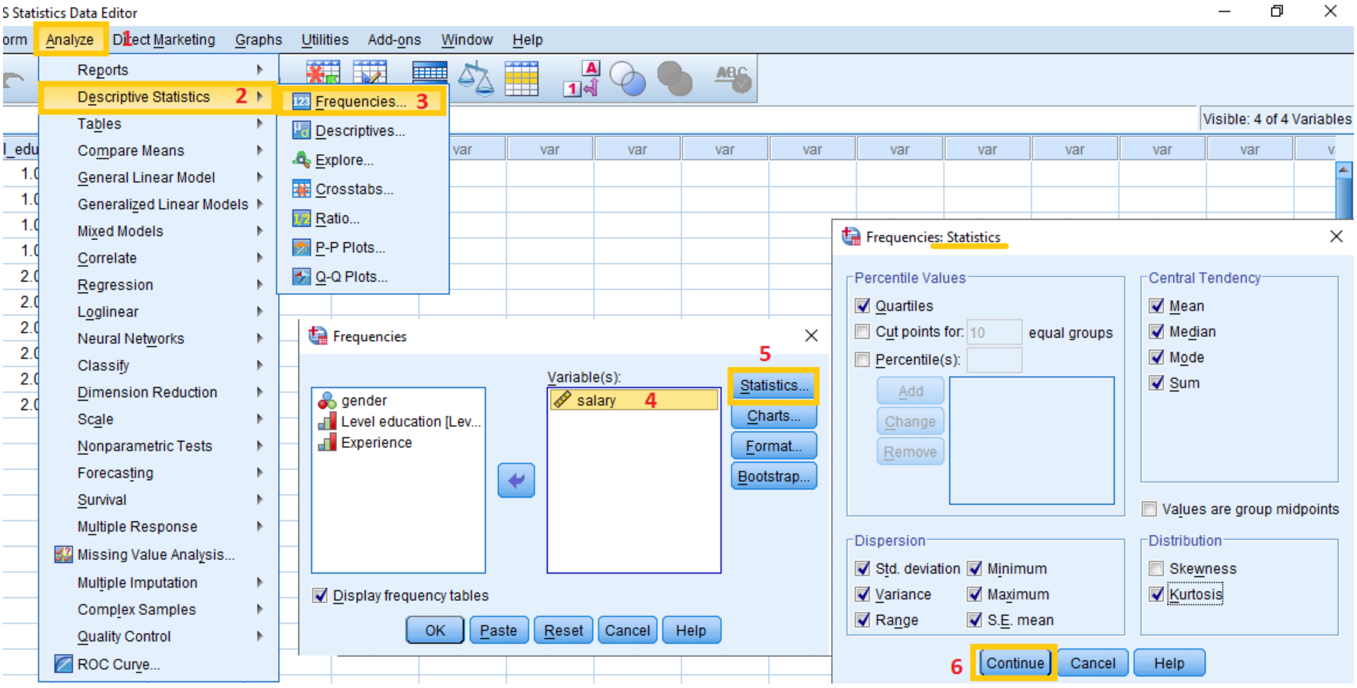
Gender	Level education	Experience	Salary
Male	Bachelor's degree	Less than 5	500.00
Female	Bachelor's degree	between 5 and 10	450.00
Male	Bachelor's degree	Less than 5	440.00
Female	Bachelor's degree	greater than 5	500.00
Male	master's degree	between 5 and 10	570.00
Female	master's degree	greater than 5	550.00
Female	master's degree	between 5 and 10	490.00
Female	master's degree	greater than 5	540.00
Male	master's degree	between 5 and 10	600.00
Male	master's degree	greater than 5	650.00

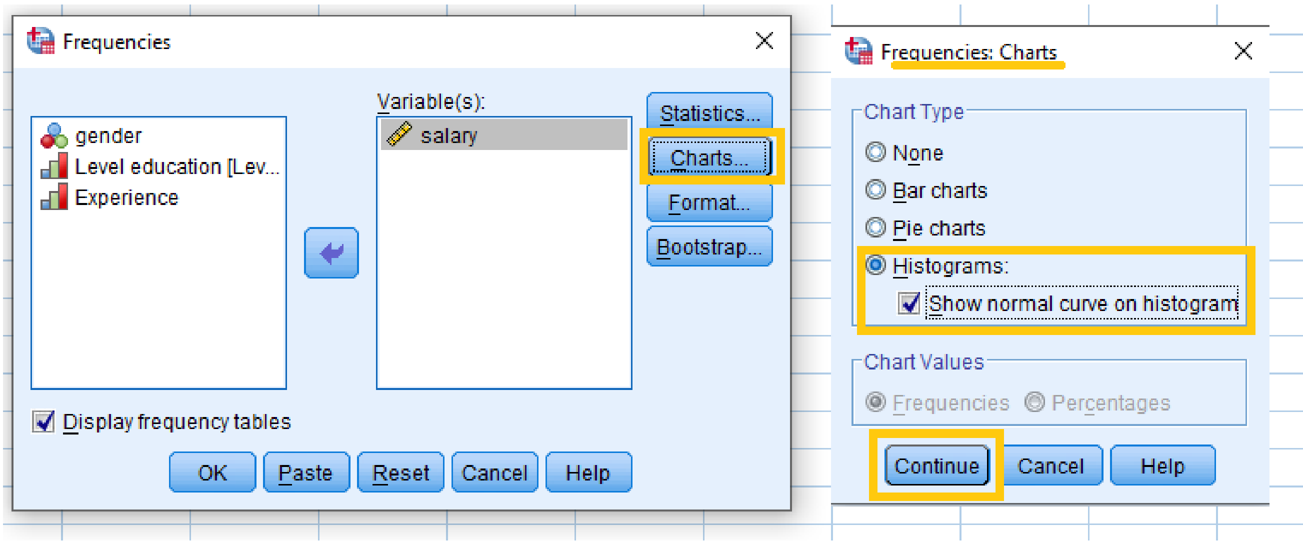



*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
Name
Type
Width
Decimals
Label
Values
Missing
Columns
Align
Measure
Role
gender
level_edu
experience
salary
level education
{1.00, Male}...
{1.00, Bach...
{1.00, Less ...
Value Labels
1.00 = "Bachelor's degree"
2.00 = "master's degree"
1.00 = "Male"
2.00 = "Female"
1.00 = "Less than 5"
2.00 = "between 5 and 10"
3.00 = "greater than 5"


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
Value Labels
gender
level_edu
experience
salary
Male
Female
Bachelor's...
master's d...
Less than 5
between 5 ...
greater tha...
Data View
Variable View

1- Use the **Frequencies** option for calculating statistical measures and frequency table for salaries :





Frequencies

[DataSet1] C:\Users\dell\Desktop\Untitled1.sav

Statistics

salary

N	Valid	10
	Missing	0
Mean		529.0000
Std. Error of Mean		20.89391
Median		520.0000
Mode		500.00
Std. Deviation		66.07235
Variance		4365.556
Kurtosis		-.351-
Std. Error of Kurtosis		1.334
Range		210.00
Minimum		440.00
Maximum		650.00
Sum		5290.00
Percentiles	25	480.0000
	50	520.0000
	75	577.5000

$\bar{x} = \frac{\sum x_i}{n}$

$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

R = max - min

Q1 = first quartile

Q2 = Second quartile

Q3 = Third quartile

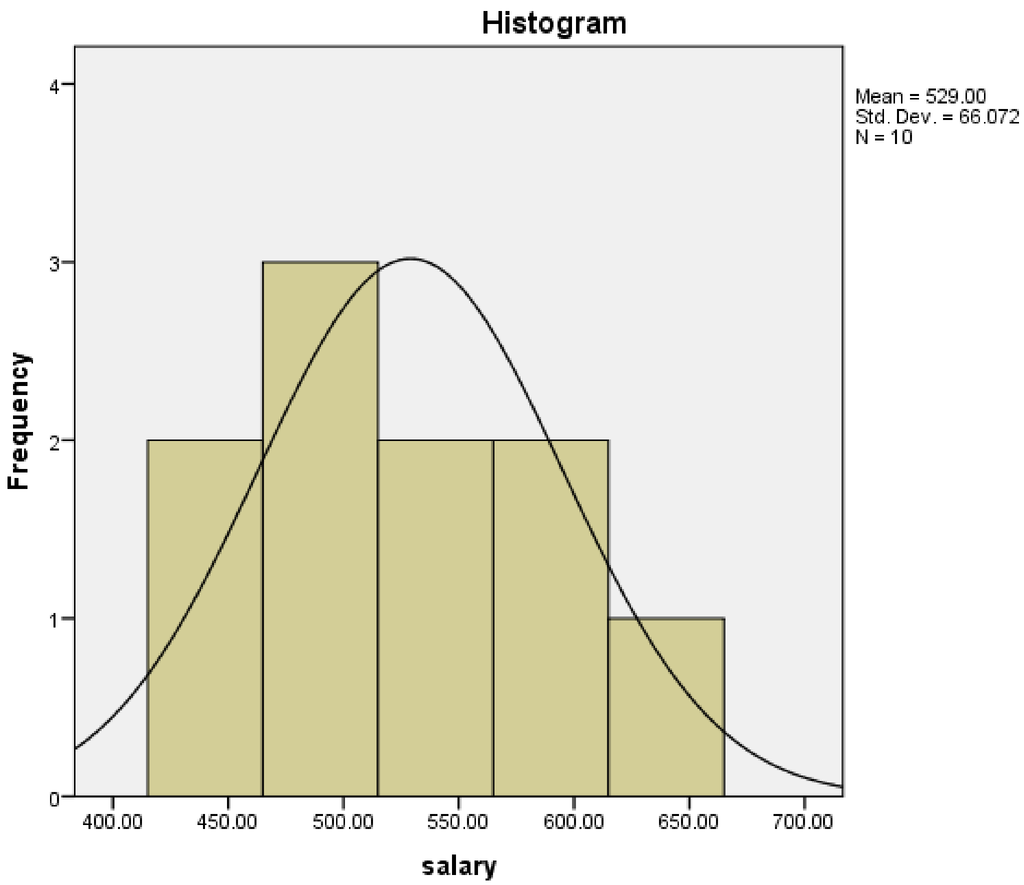
variance = $(\text{standard deviation})^2$

standard deviation = $\sqrt{\text{variance}}$

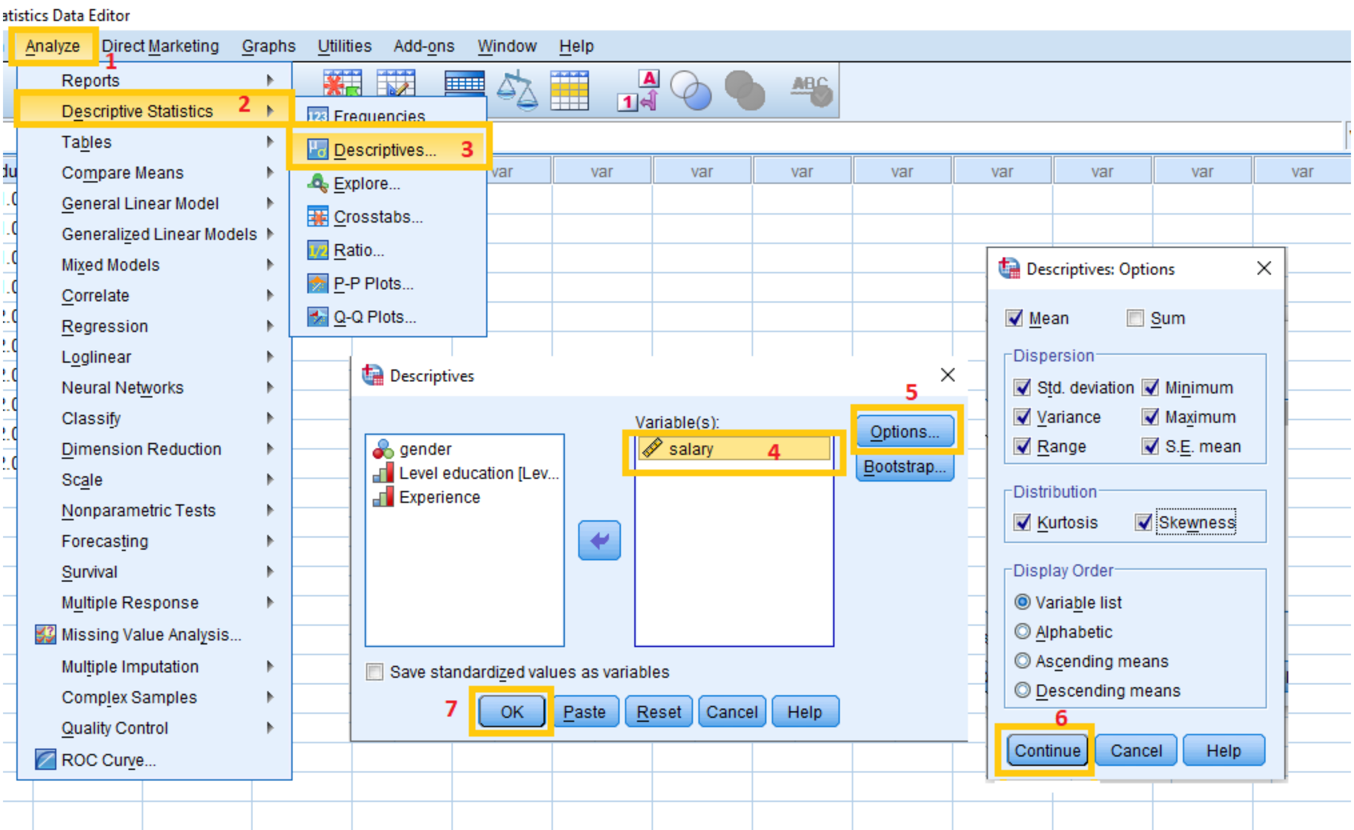
The first quartile, Q1, is the 25th percentile. The second quartile, Q2, is the 50th percentile. The third quartile, Q3, is the 75th percentile

salary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	440.00	1	10.0	10.0	10.0
	450.00	1	10.0	10.0	20.0
	490.00	1	10.0	10.0	30.0
	500.00	2	20.0	20.0	50.0
	540.00	1	10.0	10.0	60.0
	550.00	1	10.0	10.0	70.0
	570.00	1	10.0	10.0	80.0
	600.00	1	10.0	10.0	90.0
	650.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	


Histogram
Mean = 529.00
Std. Dev. = 66.072
N = 10
Frequency
0
1
2
3
4
400.00
450.00
500.00
550.00
600.00
650.00
700.00
salary

2- Use the descriptive option for calculating statistical measures for salaries :



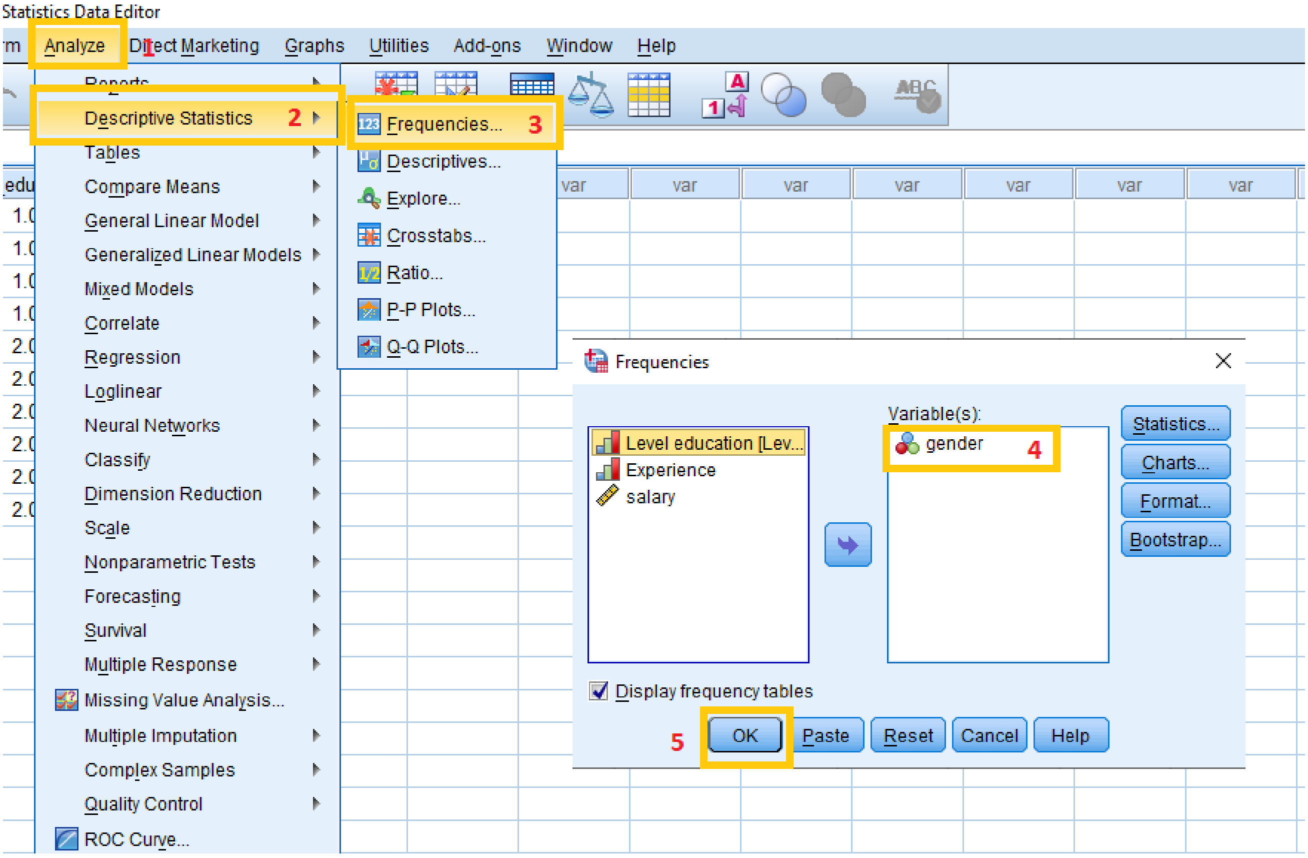
Descriptives

```
[DataSet1] C:\Users\dell\Desktop\Untitled1.sav
```

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
salary	10	210.00	440.00	650.00	529.0000	20.89391	66.07235	4365.556	.435	.687	-.351-	1.334
Valid N (listwise)	10											

3- How many male and female?



Frequencies

[DataSet1] C:\Users\dell\Desktop\Untitled1.sav

Statistics

gender

N	Valid	10
	Missing	0

gender

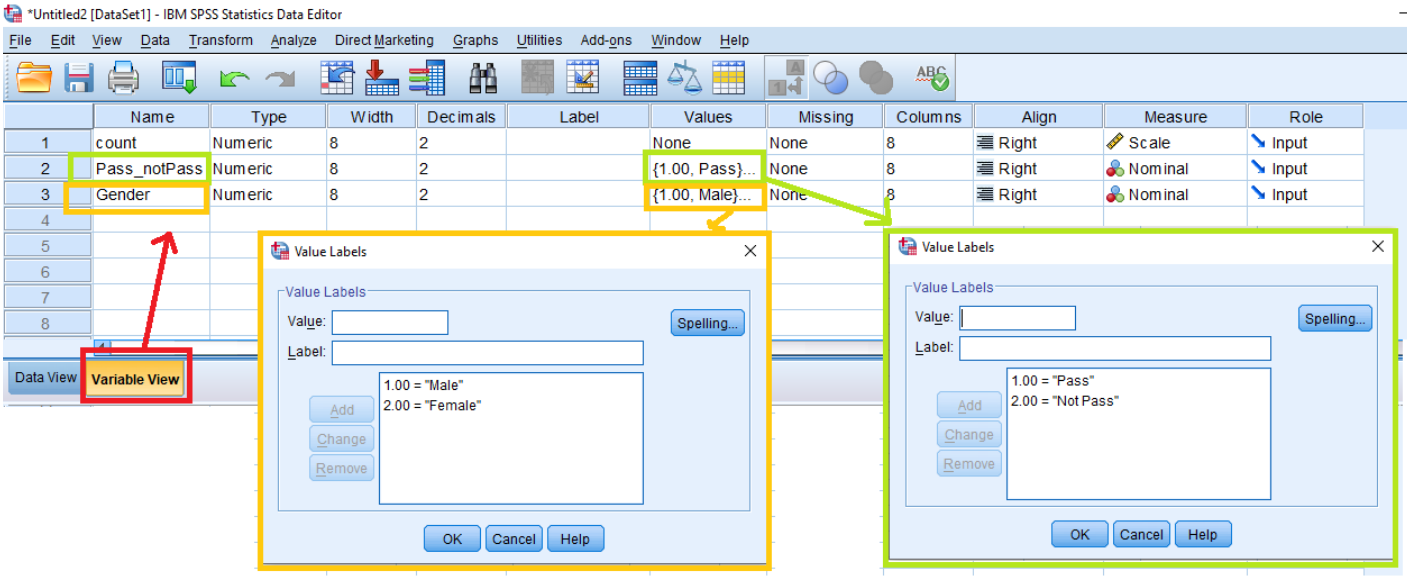
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	5	50.0	50.0	50.0
	Female	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

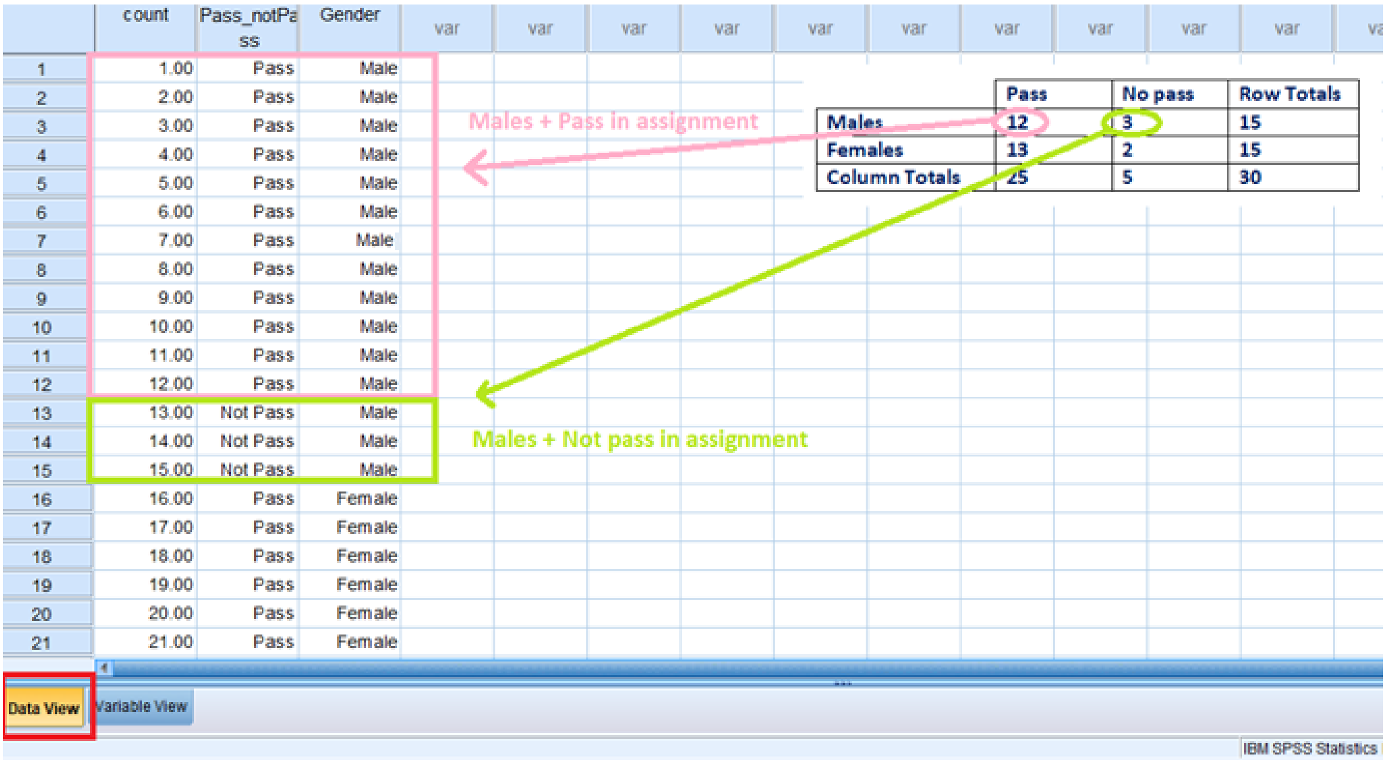
Q2: What is the relationship between the gender of the students and the assignment of a Pass or No Pass test grade? (Pass = score 70 or above)

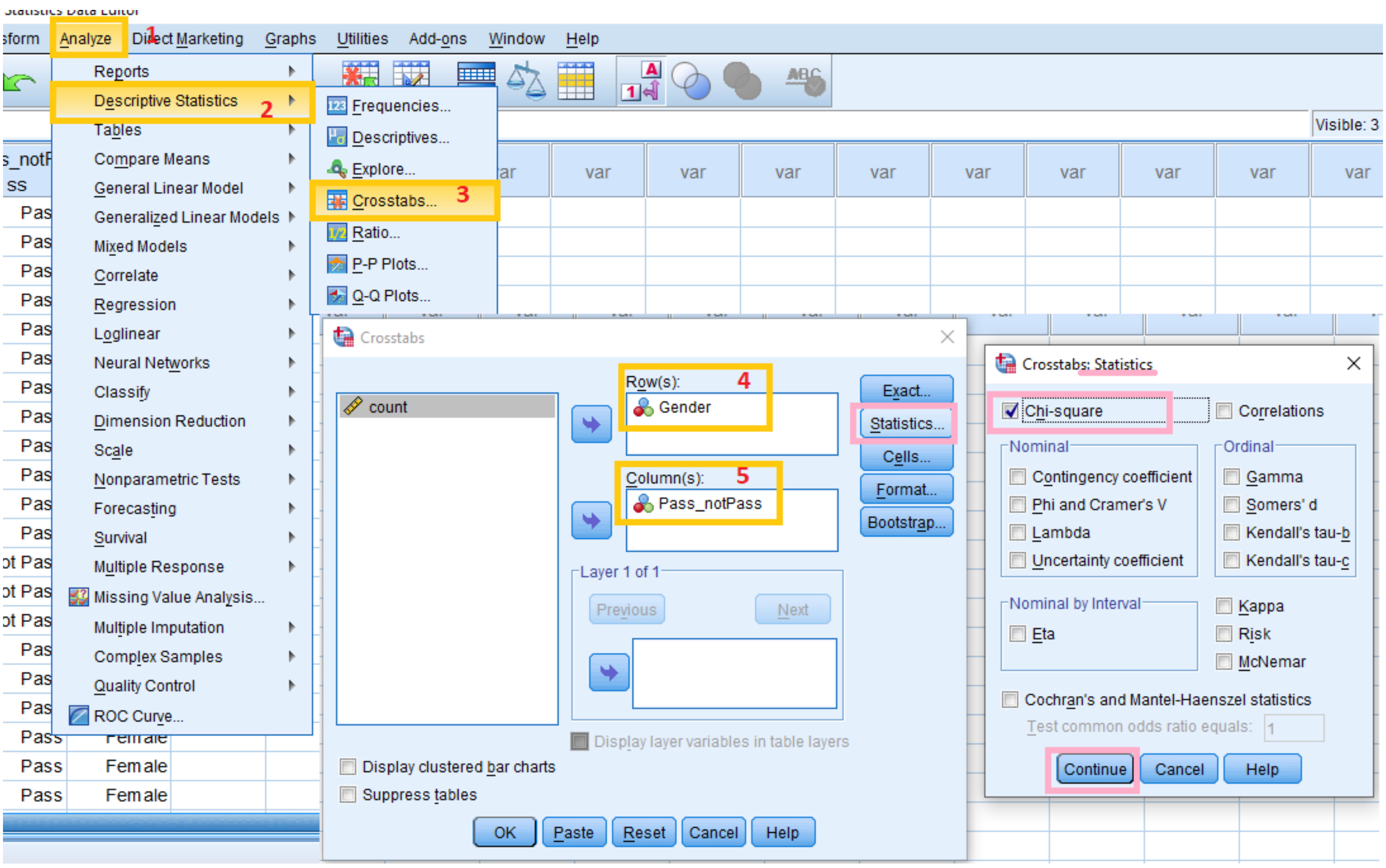
	Pass	No pass	Row Totals
Males	12	3	15
Females	13	2	15
Column Totals	25	5	30

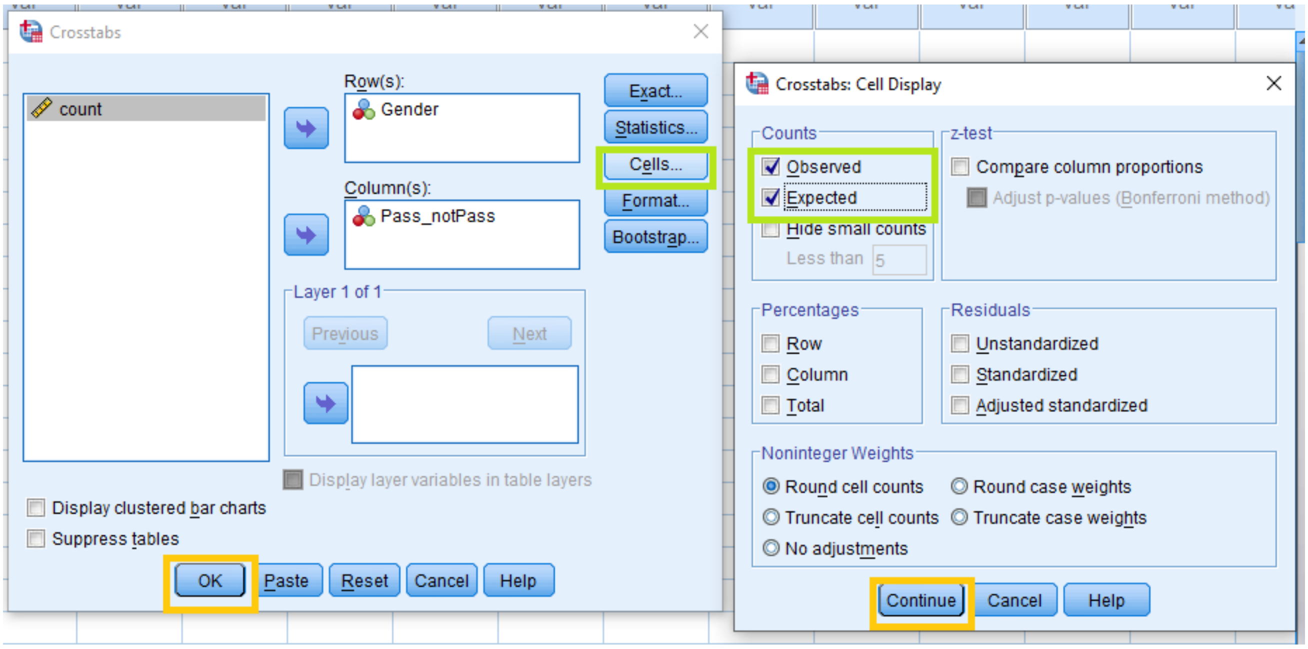
H_0: **the gender of the students is independent of pass or no pass test grade**

H_1: **the gender of the students is not independent of pass or no pass test grade**




Males + Pass in assignment
Males + Not pass in assignment
Pass
No pass
Row Totals
Males
12
3
15
Females
13
2
15
Column Totals
25
5
30
Data View
Variable View
IBM SPSS Statistics


Analyze
Descriptive Statistics
Frequencies...
Descriptives...
Explore...
Crosstabs...
Crosstabs
Row(s):
Gender
Column(s):
Pass_notPass
Statistics...
Crosstabs: Statistics
Chi-square
Continue
1
2
3
4
5



Crosstabs

[DataSet1]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Pass_notPass	30	100.0%	0	0.0%	30	100.0%

Gender * Pass_notPass Crosstabulation

			Pass_notPass		Total
			Pass	Not Pass	
Gender	Male	Count	12	3	15
		Expected Count	12.5	2.5	15.0
	Female	Count	13	2	15
		Expected Count	12.5	2.5	15.0
Total		Count	25	5	30
		Expected Count	25.0	5.0	30.0

The Chi-Square statistic $\chi^2 = 0.240$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.240[a]	1	.624		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.241	1	.623		
Fisher's Exact Test				1.000	.500
Linear-by-Linear Association	.232	1	.630		
N of Valid Cases	30				

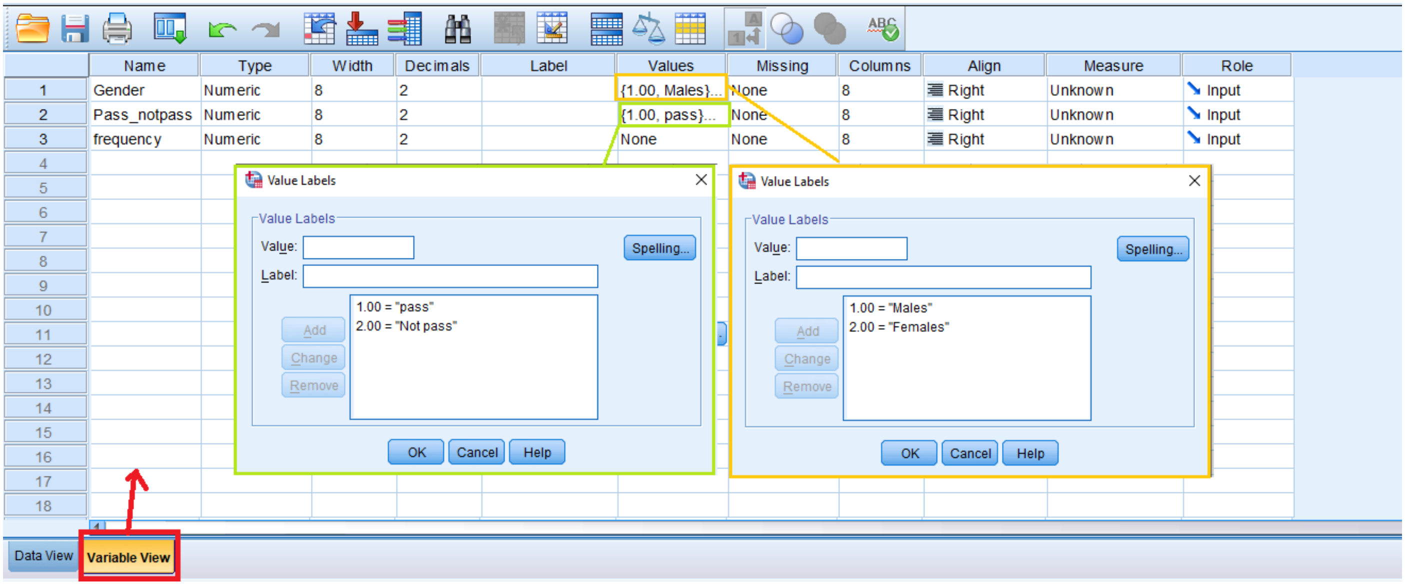
a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.50.

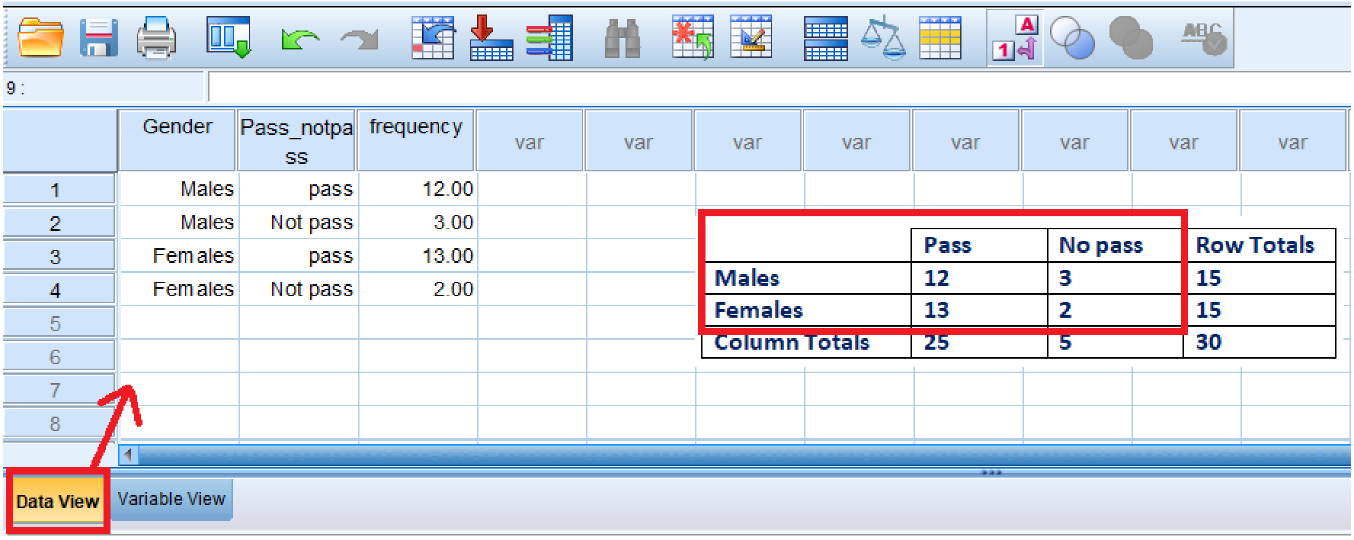
b. Computed only for a 2x2 table

degrees of freedom
df=(R-1)*(C-1) =(2-1)*(2-1)
where R: number of rows and C : number of columns.

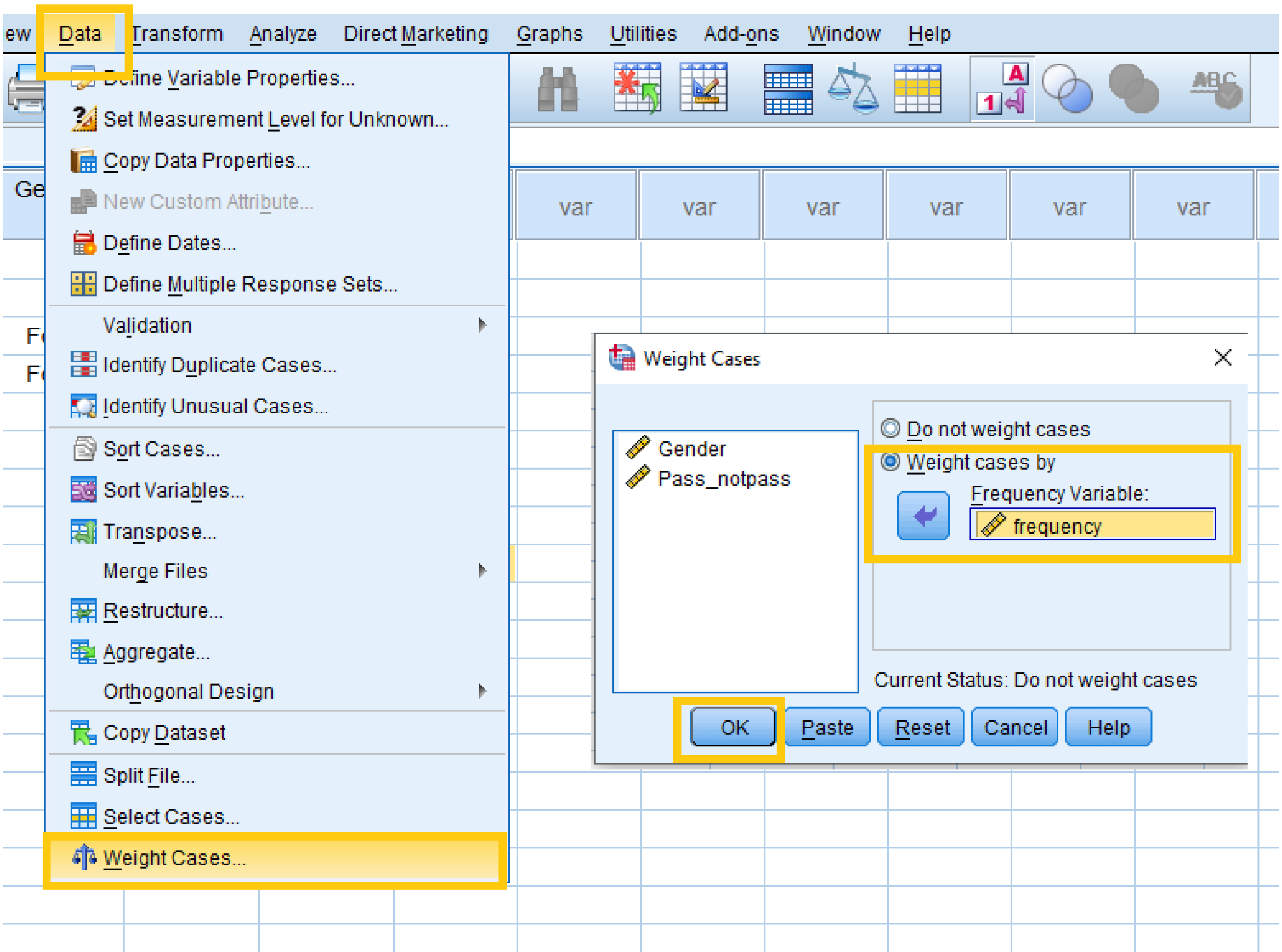
p-value =0.624 > α =0.05
So, we Accept H0

Other method :

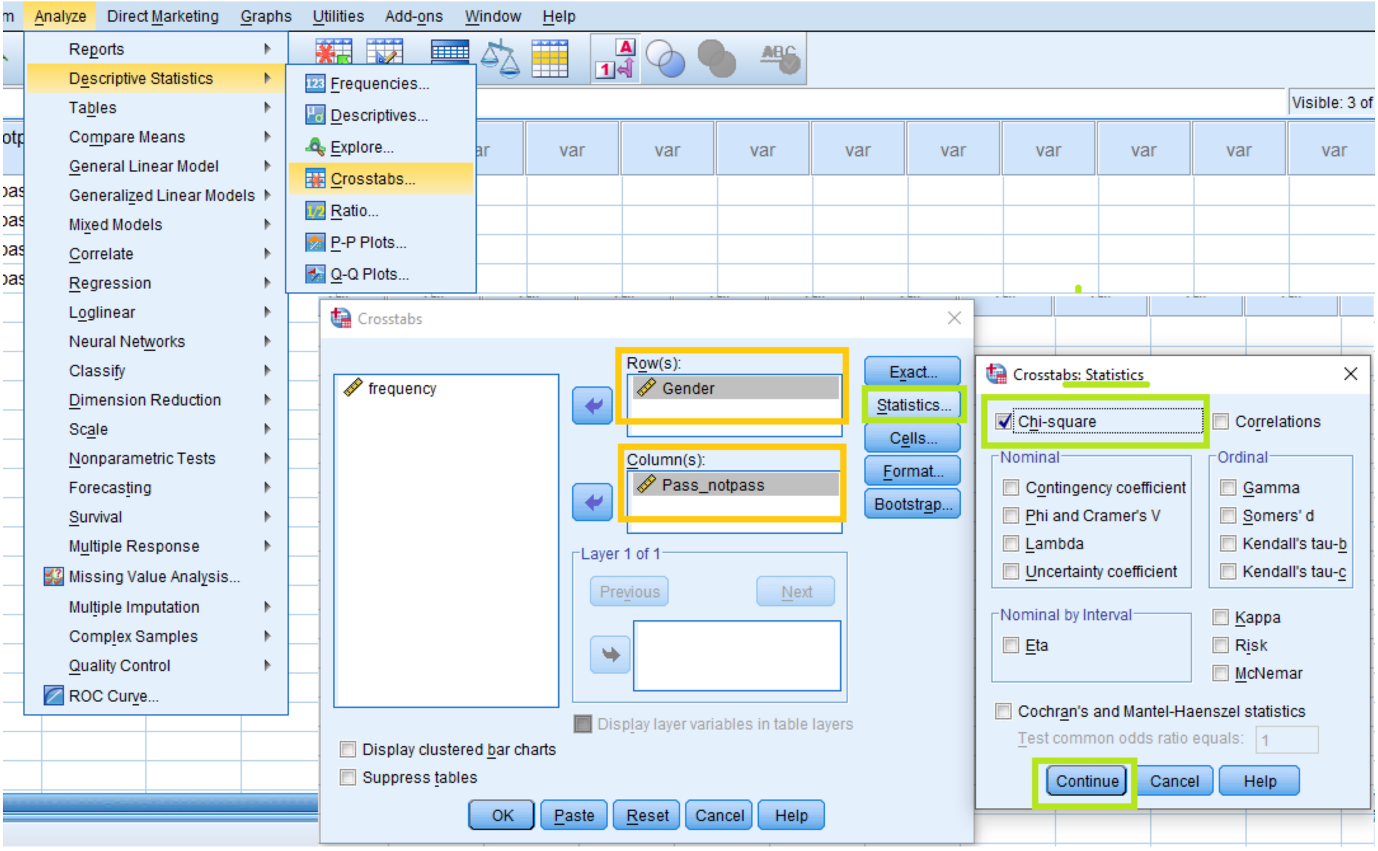


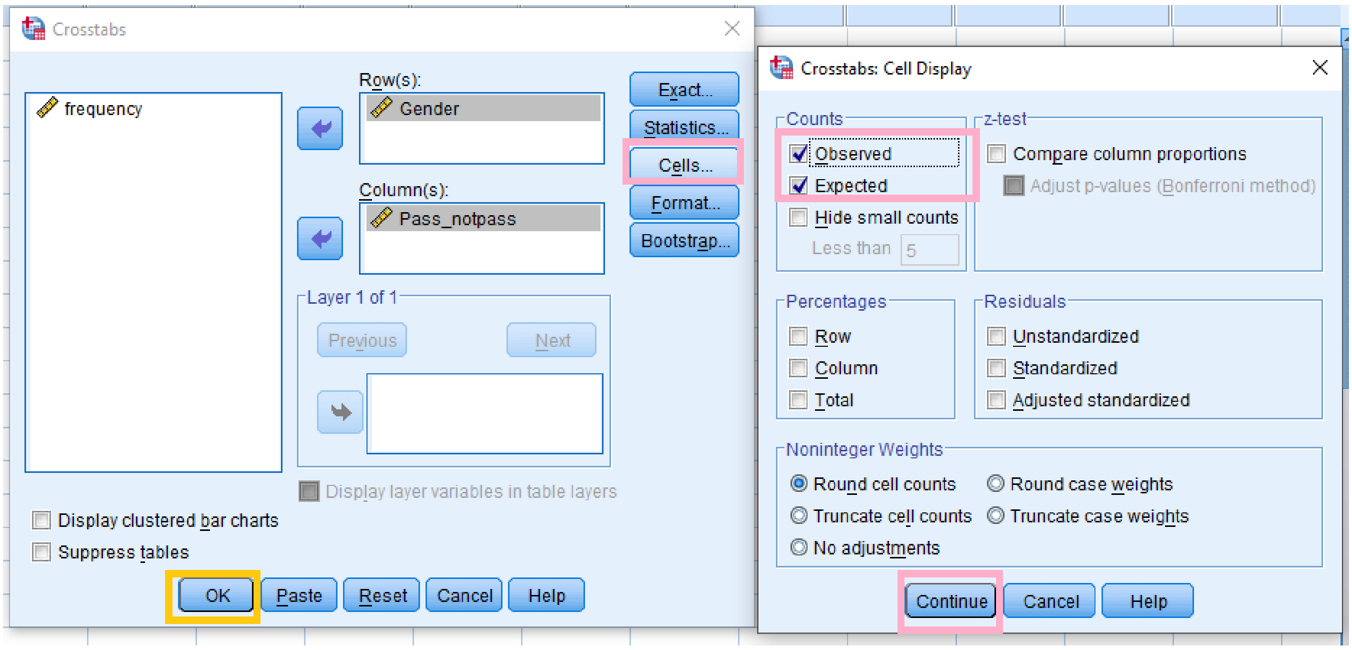


	Pass	No pass	Row Totals
Males	12	3	15
Females	13	2	15
Column Totals	25	5	30



ew
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Define Variable Properties...
Set Measurement Level for Unknown...
Copy Data Properties...
New Custom Attribute...
Define Dates...
Define Multiple Response Sets...
Validation
Identify Duplicate Cases...
Identify Unusual Cases...
Sort Cases...
Sort Variables...
Transpose...
Merge Files
Restructure...
Aggregate...
Orthogonal Design
Copy Dataset
Split File...
Select Cases...
Weight Cases...
var
Weight Cases
Gender
Pass_notpass
Do not weight cases
Weight cases by
Frequency Variable:
frequency
Current Status: Do not weight cases
OK
Paste
Reset
Cancel
Help


Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
Visible: 3 of
var
Crosstabs
frequency
Row(s):
Gender
Column(s):
Pass_notpass
Exact...
Statistics...
Cells...
Format...
Bootstrap...
Layer 1 of 1
Previous
Next
Display layer variables in table layers
Display clustered bar charts
Suppress tables
OK
Paste
Reset
Cancel
Help
Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:
1
Continue
Cancel
Help


Crosstabs
frequency
Row(s):
Gender
Column(s):
Pass_notpass
Exact...
Statistics...
Cells...
Format...
Bootstrap...
Layer 1 of 1
Previous
Next
Display layer variables in table layers
Display clustered bar charts
Suppress tables
OK
Paste
Reset
Cancel
Help
Crosstabs: Cell Display
Counts
Observed
Expected
Hide small counts
Less than
5
z-test
Compare column proportions
Adjust p-values (Bonferroni method)
Percentages
Row
Column
Total
Residuals
Unstandardized
Standardized
Adjusted standardized
Noninteger Weights
Round cell counts
Round case weights
Truncate cell counts
Truncate case weights
No adjustments
Continue
Cancel
Help

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Pass_notpass	30	100.0%	0	0.0%	30	100.0%

Gender * Pass_notpass Crosstabulation

			Pass_notpass		Total
			pass	Not pass	
Gender	Males	Count	12	3	15
		Expected Count	12.5	2.5	15.0
	Females	Count	13	2	15
		Expected Count	12.5	2.5	15.0
Total		Count	25	5	30
		Expected Count	25.0	5.0	30.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.240[a]	1	.624		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.241	1	.623		
Fisher's Exact Test				1.000	.500
Linear-by-Linear Association	.232	1	.630		
N of Valid Cases	30				

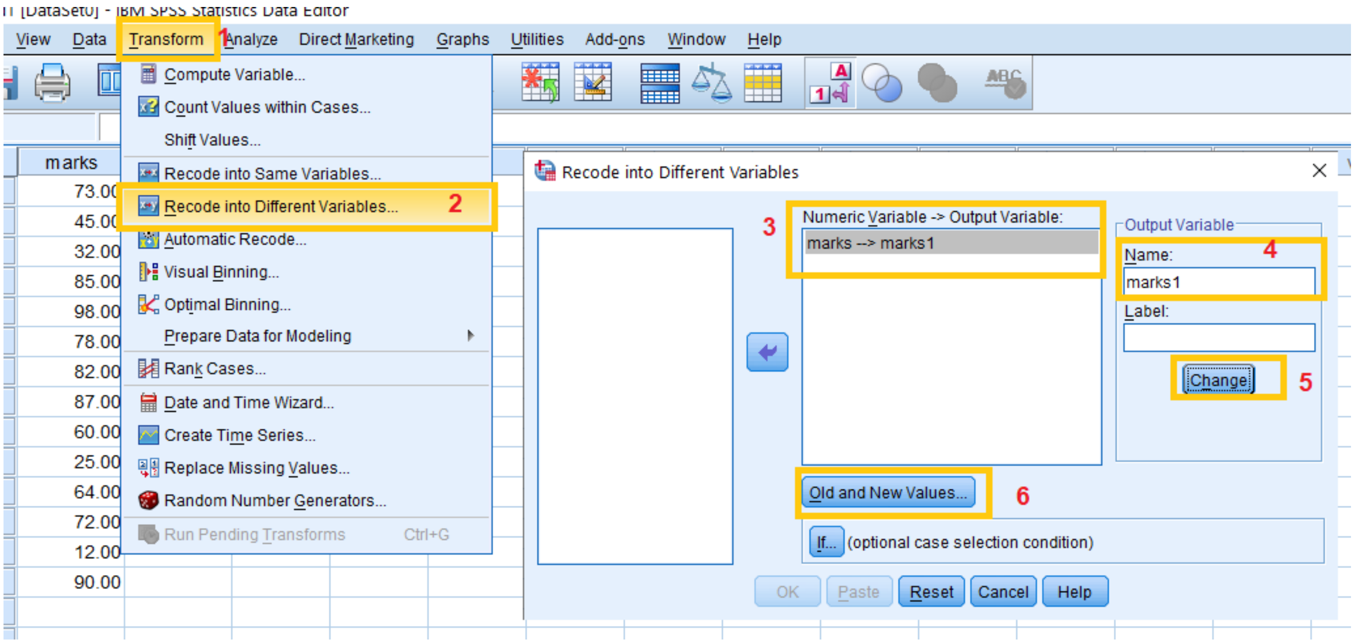
a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.50.

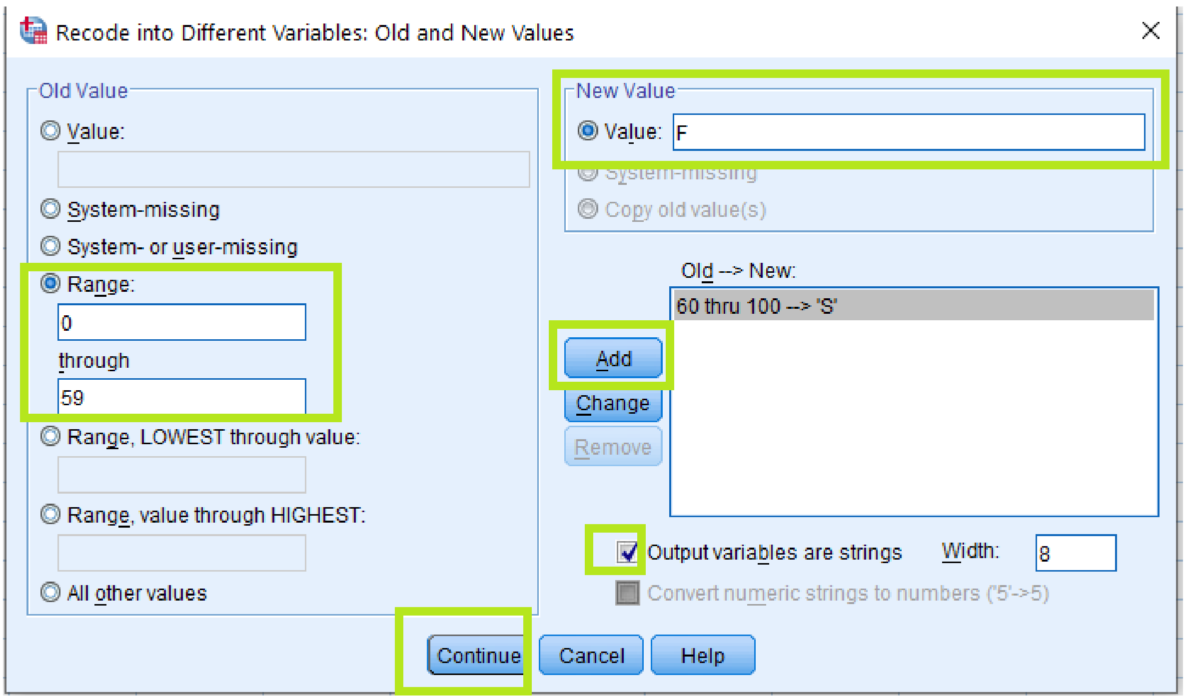
b. Computed only for a 2x2 table

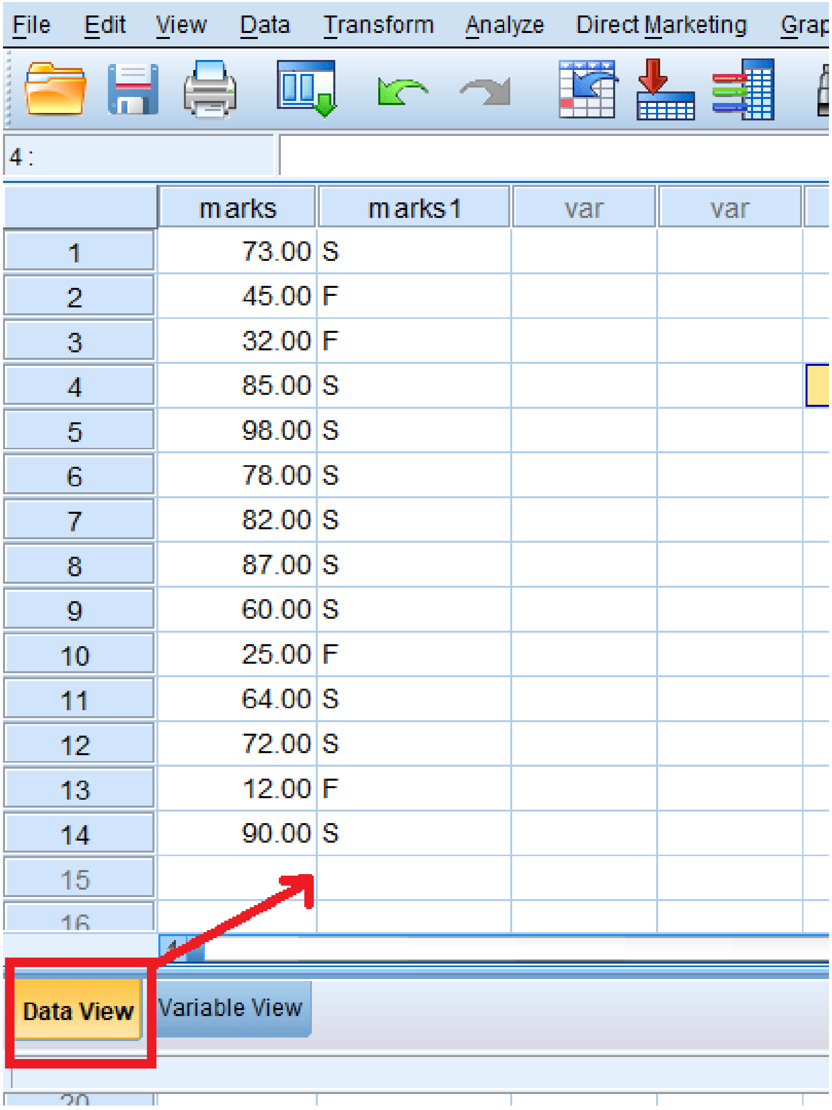
Q3: We have marks of 14 students

73 45 32 85 98 78 82 87 60 25 64 72 12 90

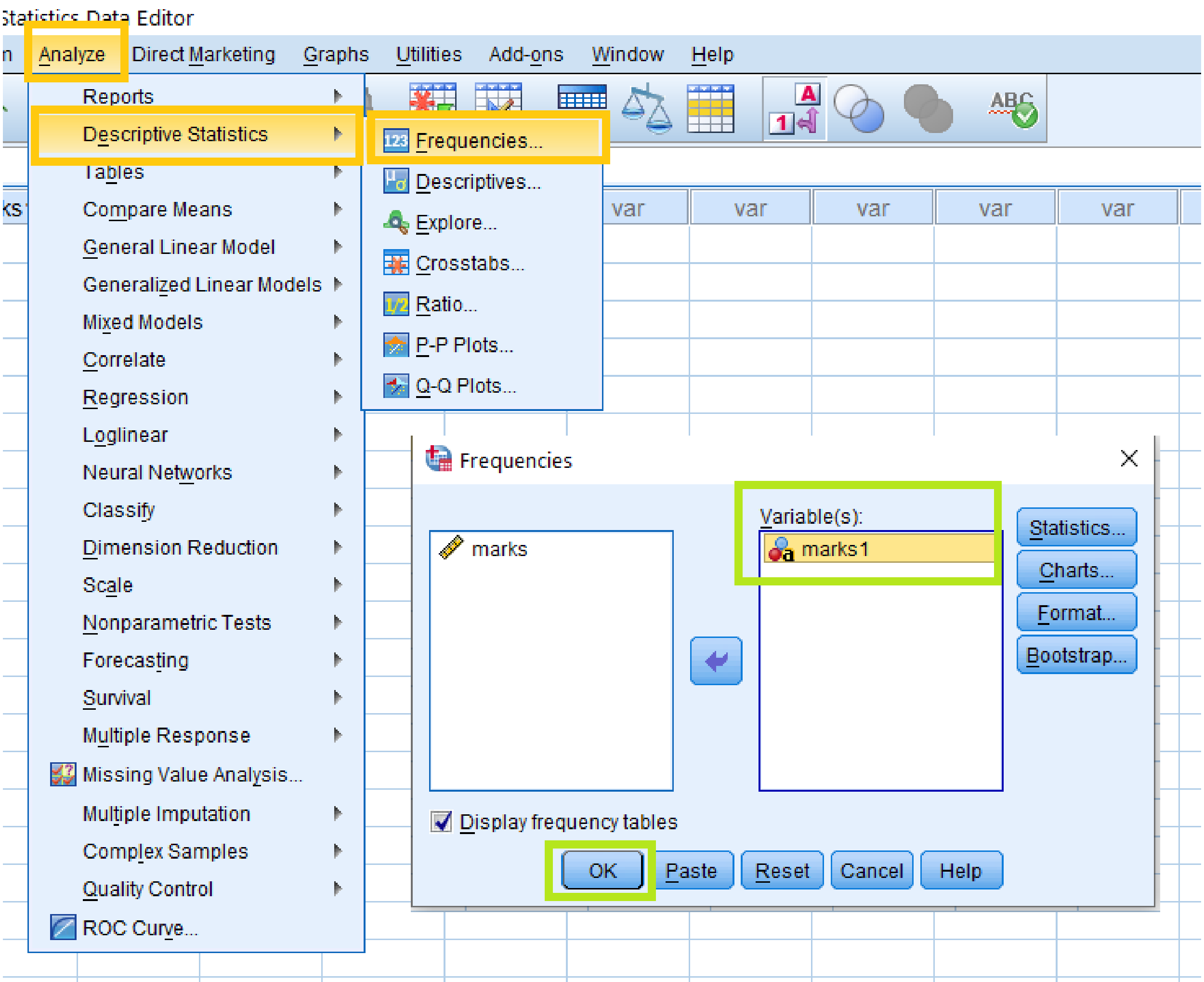
1. **Recode the students' marks to be successful (if the mark is> = 60) and be a failure (if Mark < 60)?**






File Edit View Data Transform Analyze Direct Marketing Grap
4 :
marks
marks1
var
var
1 73.00 S
2 45.00 F
3 32.00 F
4 85.00 S
5 98.00 S
6 78.00 S
7 82.00 S
8 87.00 S
9 60.00 S
10 25.00 F
11 64.00 S
12 72.00 S
13 12.00 F
14 90.00 S
15
Data View
Variable View

2. How many successful students?



Frequencies

[DataSet0]

Statistics

marks1

N	Valid	14
	Missing	0

marks1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	4	28.6	28.6	28.6
	S	10	71.4	71.4	100.0
	Total	14	100.0	100.0	

Q1: For a sample of 10 fruits from thirteen-year-old acidless orange trees, the fruit shape (determined as a diameter divided by height) was measured [Shaheen and Hamouda (8419b)]:

1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

Assuming that fruit shapes are approximately normally distributed, find and interpret a 90% confidence interval for the average fruit shape.

To use the **T- test** , we need to make sure that the population follows a normal distribution:

H_0: **the population follows a normal distribution**

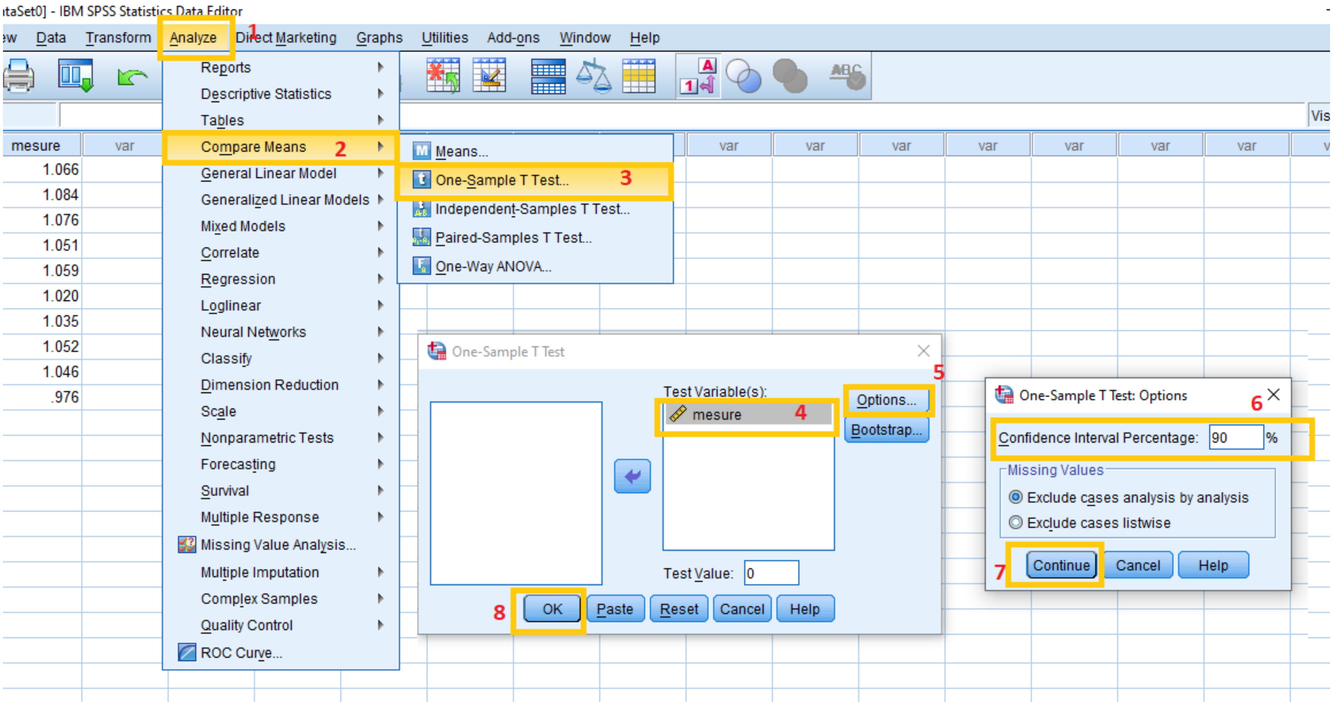
Vs

H_1: **the population does not follow a normal distribution**

we find the question said that the population follows a normal distribution, so is not necessary to make this test.

Now, 90% Confidence interval of the mean can be found in two ways :

1) The first method:



T-Test

[DataSet0]

$\bar{x} = \frac{\sum x_i}{n}$ $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ $SE(\bar{x}) = \frac{s}{\sqrt{n}}$

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
mesure	10	1.04650	.031035	.009814

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	90% Confidence Interval of the Difference Lower	90% Confidence Interval of the Difference Upper
mesure	106.632	9	.000	1.046500	1.02851	1.06449

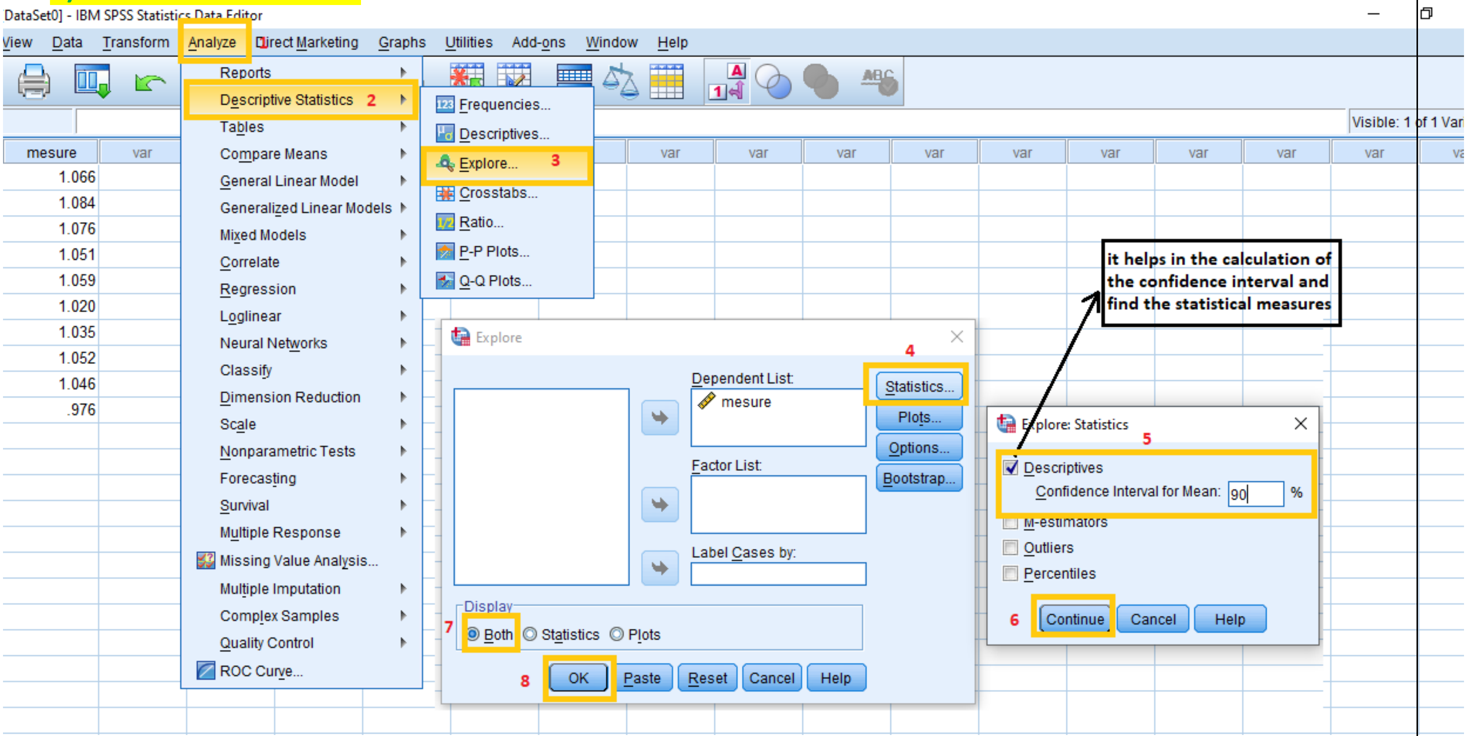
$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ df= n-1 p-value

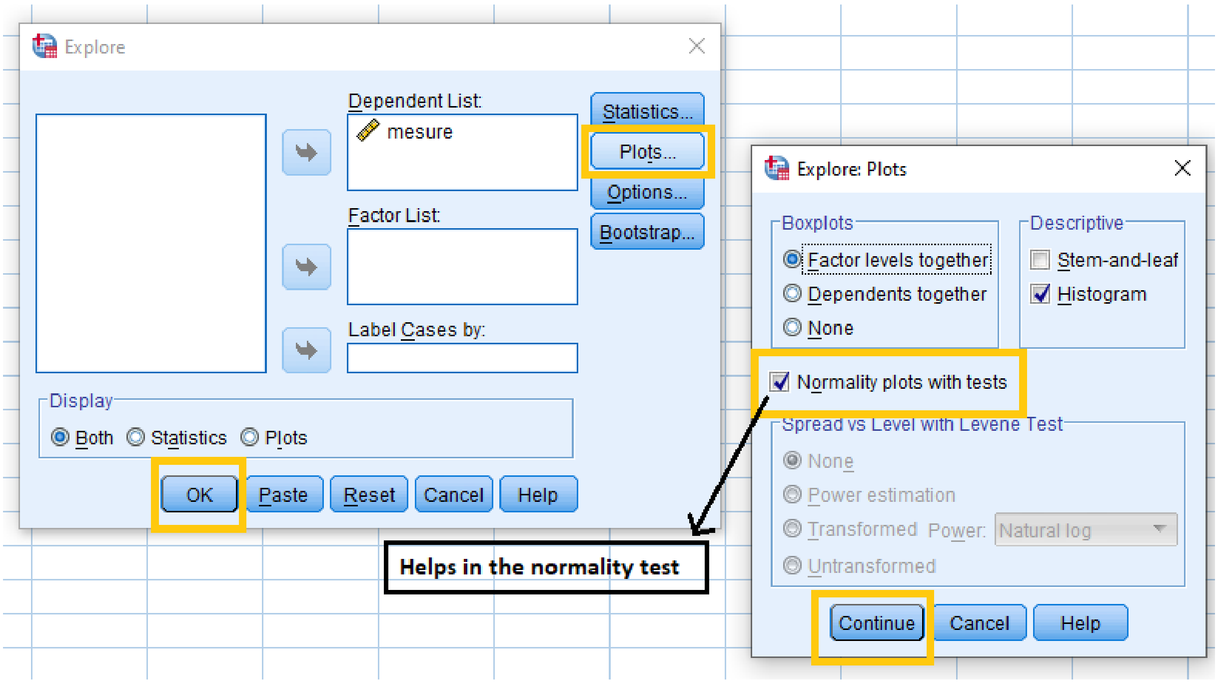
$\bar{x} \pm t_{1-\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$

C.I for the mean μ

IBM S

2) The second method:





Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
mesure	10	100.0%	0	0.0%	10	100.0%

Descriptives

			Statistic	Std. Error
mesure	Mean		1.04650	.009814
	90% Confidence Interval for Mean	Lower Bound	1.02851	
		Upper Bound	1.06449	
	5% Trimmed Mean		1.04833	
	Median		1.05150	
	Variance		.001	
	Std. Deviation		.031035	
	Minimum		.976	
	Maximum		1.084	
	Range		.108	
	Interquartile Range		.037	
	Skewness		-1.313-	.687
	Kurtosis		2.276	1.334

C.I for the mean

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
mesure	.194	10	.200*	.907	10	.260

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

p-value > 0.1 = α

so , we accept H0: the population follows a normal distribution

Q2: The phosphorus content was measured for independent samples of skim and whole:

Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

Assuming normal populations with equal variance

a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**
b) **Find and interpret a 99% confidence interval for the difference in average phosphorus contents of whole and skim milk .**

to use the T- test for two sample, we need to make sure that

1) **The independence of the two samples:**

It is very clear that there is no correlation between the values of the two samples.

2) **The populations follow a normal distribution**

To use the T- test for two sample, we need to make sure that :

1) The independence of the two samples:

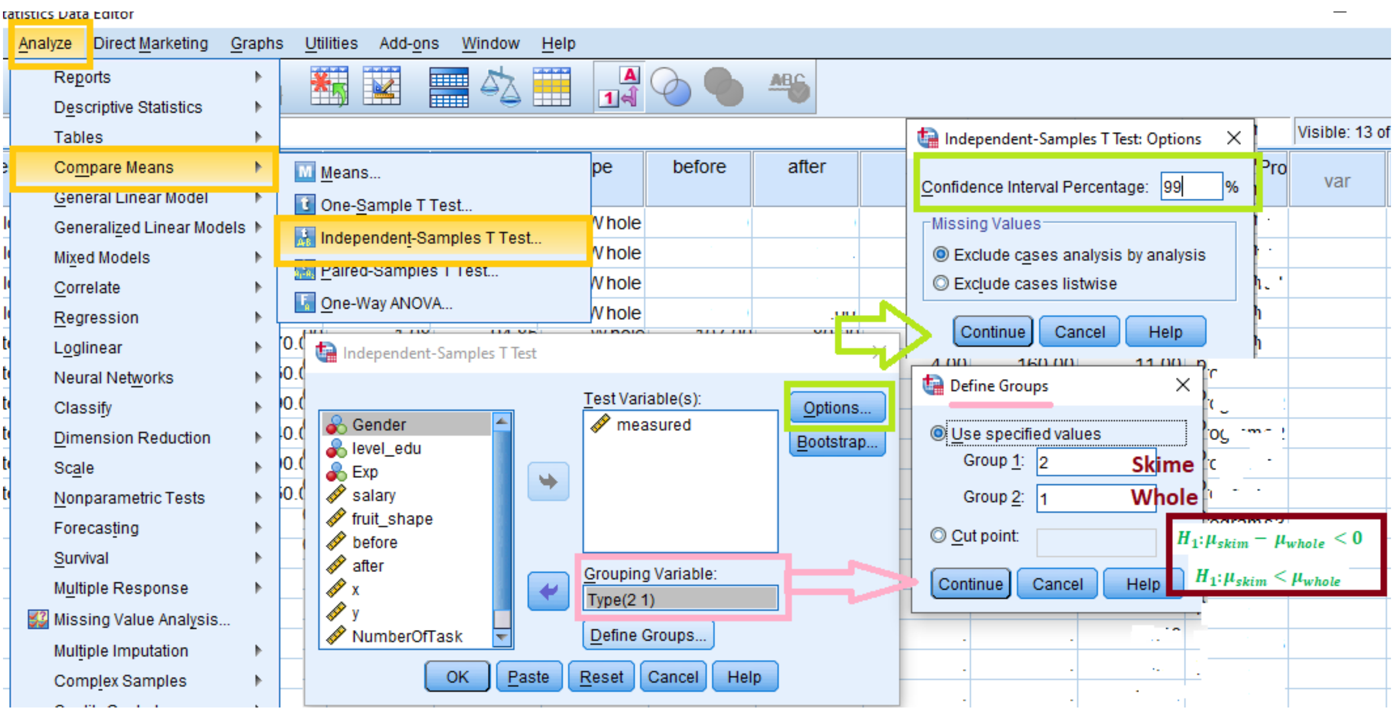
It is very clear that there is no correlation between the values of the two samples.

2) The populations follow a normal distribution

H_0: the population follows a normal distribution
Vs H_1: the population does not follow a normal distribution

However, we find the question said that the populations follows a normal distribution, so is not necessary to make this test.

a) $H_0: \mu_{skim} - \mu_{whole} = 0 \quad Vs \quad H_1: \mu_{skim} - \mu_{whole} < 0 \qquad at\ \alpha = 0.01$
b) **99 % Confidence interval of** $\mu_{skim} - \mu_{whole}$



T-Test

Group Statistics

	Type	N	Mean	Std. Deviation	Std. Error Mean
measured	skim	10	91.3400	.48293	.15272
	Whole	10	94.6450	.50302	.15907

This for test :

$H_0: \sigma^2_{whole} = \sigma^2_{skim}$ Vs $H_1: \sigma^2_{whole} \neq \sigma^2_{skim}$

$p-value = 0.924 > \alpha = 0.01$. So, we Accept H_0

.However, it is given in question

Independent Samples Test

C.I $\mu_{skim} - \mu_{whole}$

p-value=(0/2)=0

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									99% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
measured	Equal variances assumed	.009	.924	-14.988-	18	.000	-3.30500-	.22051	-3.93973-	-2.67027-
	Equal variances not assumed			-14.988-	17.970	.000	-3.30500-	.22051	-3.93985-	-2.67015-

...viation	Std. Error Mean
...48293	.15272
...50302	.15907

This for test :

$H_0: \sigma^2_{whole} = \sigma^2_{skim}$ Vs $H_1: \sigma^2_{whole} \neq \sigma^2_{skim}$

$p - value = 0.924 > \alpha = 0.01$. So,we Accept H_0

.However, it is given in question

Independent Samples Test

C.I $\mu_{skim} - \mu_{whole}$

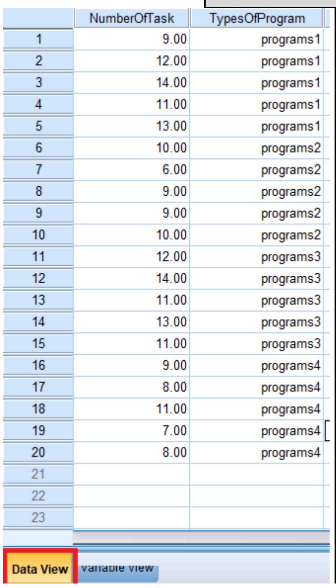
p-value=(0/2)=0

Test stat.

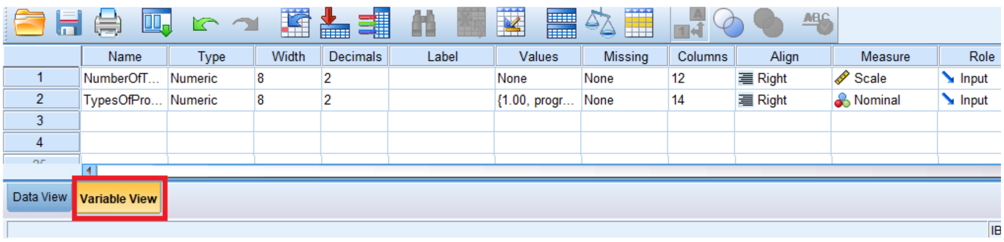
...est for Equality of ...ariances		t-test for Equality of Means						
							99% Confidence Interval of the Difference	
	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
...09	.924	-14.988-	18	.000	-3.30500-	.22051	-3.93973-	-2.67027-
		-14.988-	17.970	.000	-3.30500-	.22051	-3.93985-	-2.67015-

Q3: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results (α =0.05)

Observation	Programs 1	Programs 2	Programs 3	Programs 4
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8



	NumberOfTask	TypesOfProgram
1	9.00	programs1
2	12.00	programs1
3	14.00	programs1
4	11.00	programs1
5	13.00	programs1
6	10.00	programs2
7	6.00	programs2
8	9.00	programs2
9	9.00	programs2
10	10.00	programs2
11	12.00	programs3
12	14.00	programs3
13	11.00	programs3
14	13.00	programs3
15	11.00	programs3
16	9.00	programs4
17	8.00	programs4
18	11.00	programs4
19	7.00	programs4
20	8.00	programs4



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	NumberOfT...	Numeric	8	2		None	None	12	Right	Scale	Input
2	TypesOfPro...	Numeric	8	2		{1.00, progr...	None	14	Right	Nominal	Input

to use the one way ANOVA- test, we need to make sure that :

1) **The independence of the four samples:**

It is very clear that there is no correlation between the values of the four samples .

2) **The populations follow a normal distribution :**

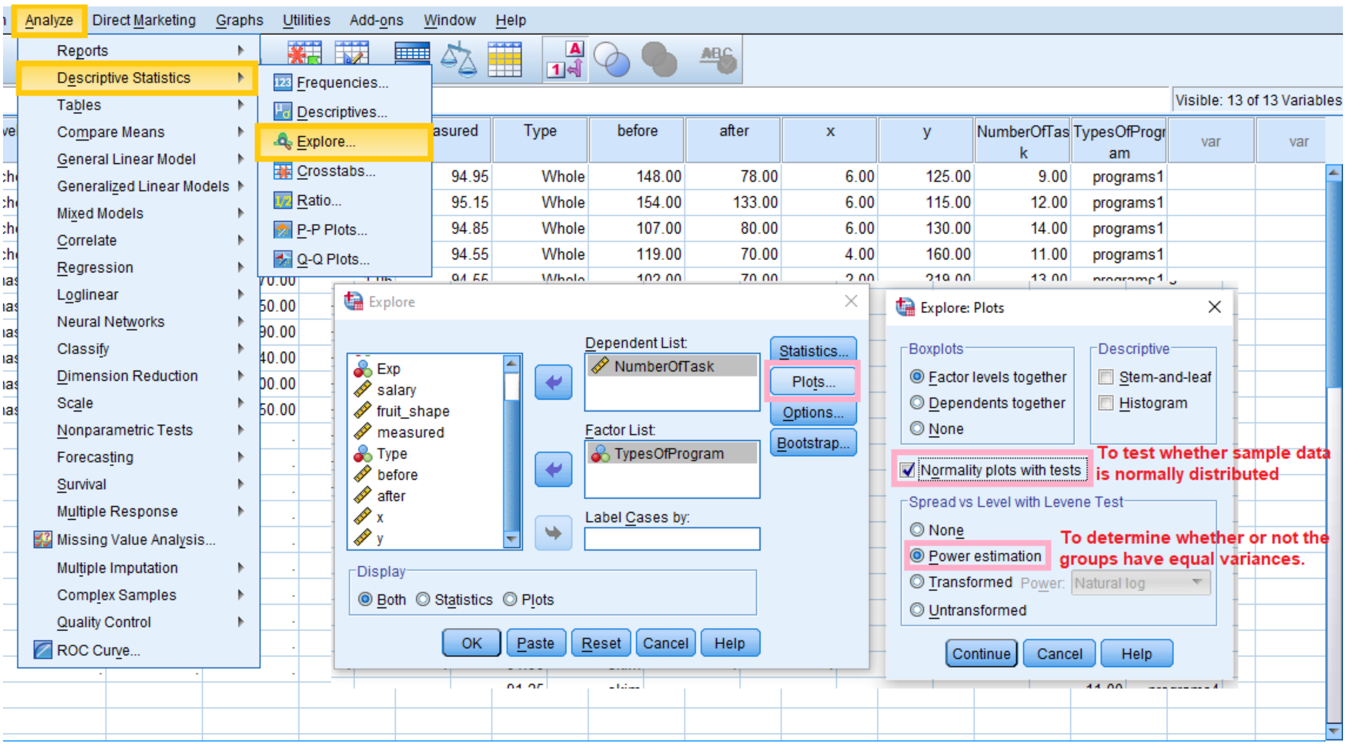
H_0: The data is normally distributed **Vs** H_1: The data does not normally distributed

3) **homogeneity variance :** using Levene's Test for Equality of Variances.

$$H_0: \sigma^2_{\text{program 1}} = \sigma^2_{\text{program 2}} = \sigma^2_{\text{program 3}} = \sigma^2_{\text{program 4}}$$

i.e. the variance of each sample are equal

$Vs\ H_1$: The variances are not all equal



Tests of Normality

NumberOfTask	TypesOfProgram	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
NumberOfTask	programs1	.141	5	.200*	.979	5	.928
	programs2	.348	5	.047	.779	5	.054
	programs3	.221	5	.200*	.902	5	.421
	programs4	.254	5	.200*	.914	5	.492

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Hypotheses :

H_0: The data is normally distributed

Vs

H_1: The data does not normally distributed

Decision :

p-value > 0.05

We **not reject** the null hypothesis the data is normally distributed

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
NumberOfTask	Based on Mean	.190	3	16	.902
	Based on Median	.167	3	16	.917
	Based on Median and with adjusted df	.167	3	14.311	.917
	Based on trimmed mean	.191	3	16	.901

Hypotheses :

$H_0: \sigma^2_{\text{program 1}} = \sigma^2_{\text{program 2}} = \sigma^2_{\text{program 3}} = \sigma^2_{\text{program 4}}$

i.e. the variance of each sample are equal

Vs H_1: The variances are not all equal

Decision :

p-value > 0.05

we fail to reject the null hypothesis.

NumberOfTask

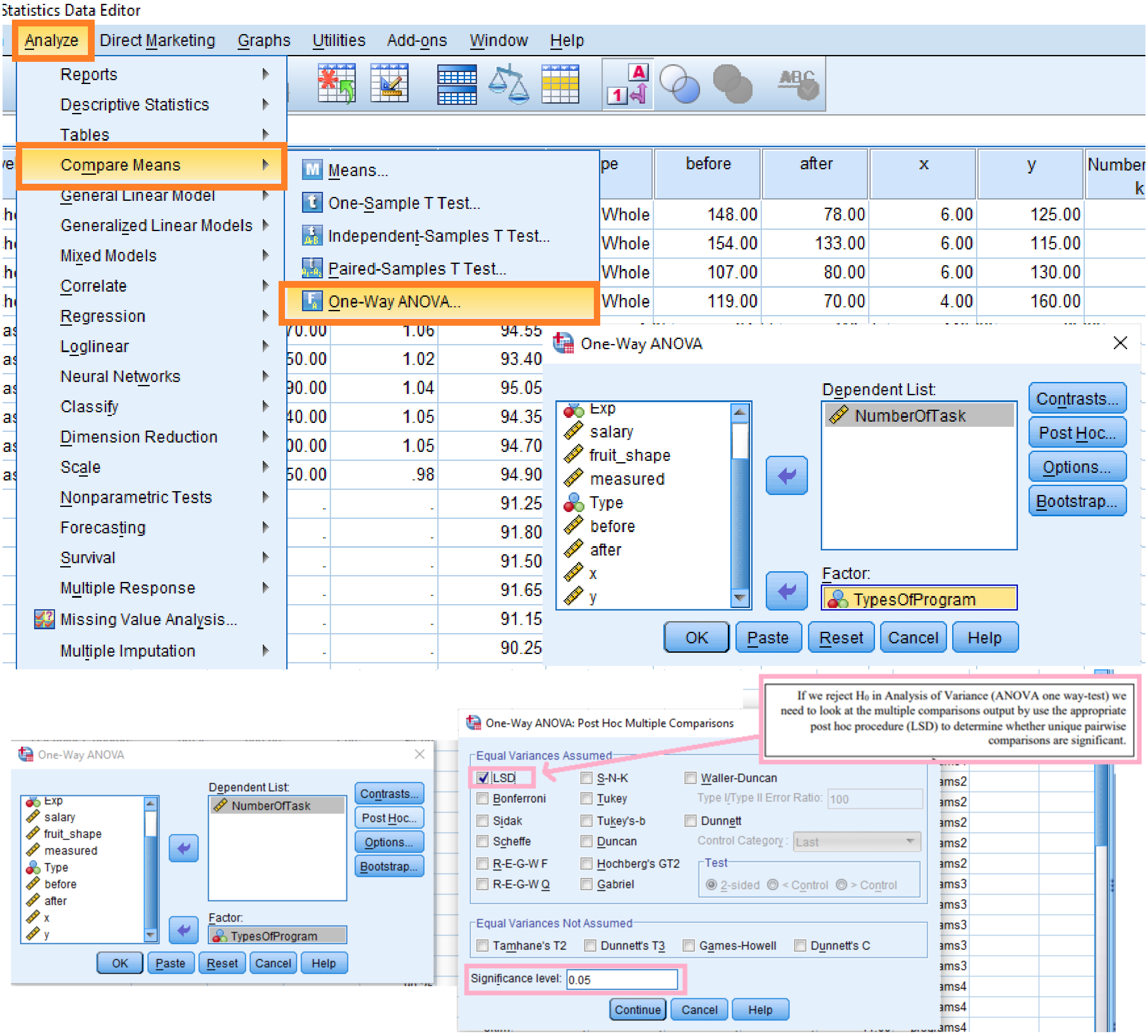
Now, the **goal** of the question:

$$H_0: \mu_{program\ 1} = \mu_{program\ 2} = \mu_{program\ 3} = \mu_{program\ 4}$$

i.e. treatments are equally effective

Vs

$$H_1: The\ means\ are\ not\ all\ equal$$



ANOVA

NumberOfTask

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	54.950	3	18.317	7.045	.003
Within Groups	41.600	16	2.600		
Total	96.550	19			

p-value < 0.05 , than we reject H0

$$H_0: \mu_{program\ 1} = \mu_{program\ 2} = \mu_{program\ 3} = \mu_{program\ 4}$$

Post Hoc Tests

Multiple Comparisons

Dependent Variable: NumberOfTask
LSD

(I) TypesOfProgram	(J) TypesOfProgram	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
programs1	programs2	3.00000*	1.01980	.010	.8381	5.1619
	programs3	-.40000-	1.01980	.700	-2.5619-	1.7619
	programs4	3.20000*	1.01980	.006	1.0381	5.3619
programs2	programs1	-3.00000-*	1.01980	.010	-5.1619-	-.8381-
	programs3	-3.40000-*	1.01980	.004	-5.5619-	-1.2381-
	programs4	.20000	1.01980	.847	-1.9619-	2.3619
programs3	programs1	.40000	1.01980	.700	-1.7619-	2.5619
	programs2	3.40000*	1.01980	.004	1.2381	5.5619
	programs4	3.60000*	1.01980	.003	1.4381	5.7619
programs4	programs1	-3.20000-*	1.01980	.006	-5.3619-	-1.0381-
	programs2	-.20000-	1.01980	.847	-2.3619-	1.9619
	programs3	-3.60000-*	1.01980	.003	-5.7619-	-1.4381-

*. The mean difference is significant at the 0.05 level.

1) $H_0: \mu_{program\,1} = \mu_{program\,2}$ vs $H_1: \mu_{program\,1} \neq \mu_{program\,2}$ at $\alpha = .05$

as $P-value = .01 < .05$, then we reject H_0.

2) $H_0: \mu_{program\,1} = \mu_{program\,3}$ vs $H_1: \mu_{program\,1} \neq \mu_{program\,3}$ at $\alpha = .05$

as $P-value = .7 > .05$, then we except H_0.

3) $H_0: \mu_{program\,1} = \mu_{program\,4}$ vs $H_1: \mu_{program\,1} \neq \mu_{program\,4}$ at $\alpha = .05$

as $P-value = .006 < .05$, then we reject H_0.

4) $H_0: \mu_{program\,2} = \mu_{program\,3}$ vs $H_1: \mu_{program\,2} \neq \mu_{program\,3}$ at $\alpha = .05$

as $P-value = .004 < .05$, then we reject H_0.

5) $H_0: \mu_{program\,2} = \mu_{program\,4}$ vs $H_1: \mu_{program\,2} \neq \mu_{program\,4}$ at $\alpha = .05$

as $P-value = .847 > .05$, then we except H_0.

6) $H_0: \mu_{program\,3} = \mu_{program\,4}$ vs $H_1: \mu_{program\,3} \neq \mu_{program\,4}$ at $\alpha = .05$

as $P-value = .003 < .05$, then we reject H_0.

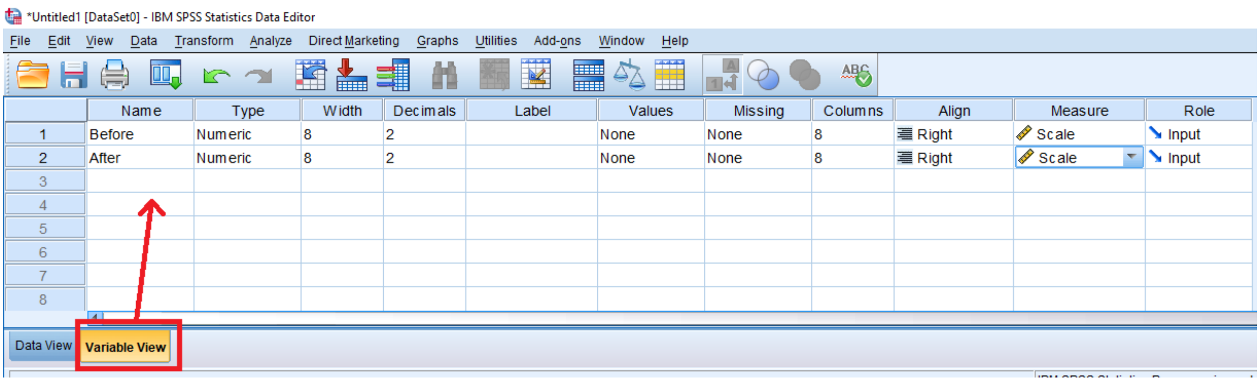
Q4: In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

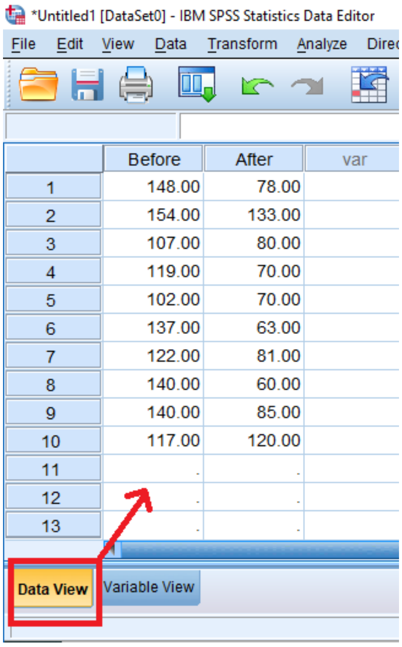
Before surgery (X)	**148**	**154**	**107**	**119**	**102**	**137**	**122**	**140**	**140**	**117**
After surgery (Y)	**78**	**133**	**80**	**70**	**70**	**63**	**81**	**60**	**85**	**120**

We assume that the <u>data comes from normal distribution</u>. Find :

1- 99% confidence interval for µD, where µD is the difference in the average weight before and after surgery.

2- Test whether the data provide sufficient evidence to indicate a difference in the average weight before and after surgery. ($\mu_D = 0$ versus $\mu_D \neq 0$) (α =0.01)




*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direc

	Before	After	var
1	148.00	78.00	
2	154.00	133.00	
3	107.00	80.00	
4	119.00	70.00	
5	102.00	70.00	
6	137.00	63.00	
7	122.00	81.00	
8	140.00	60.00	
9	140.00	85.00	
10	117.00	120.00	
11	.	.	
12	.	.	
13	.	.	

Data View Variable View

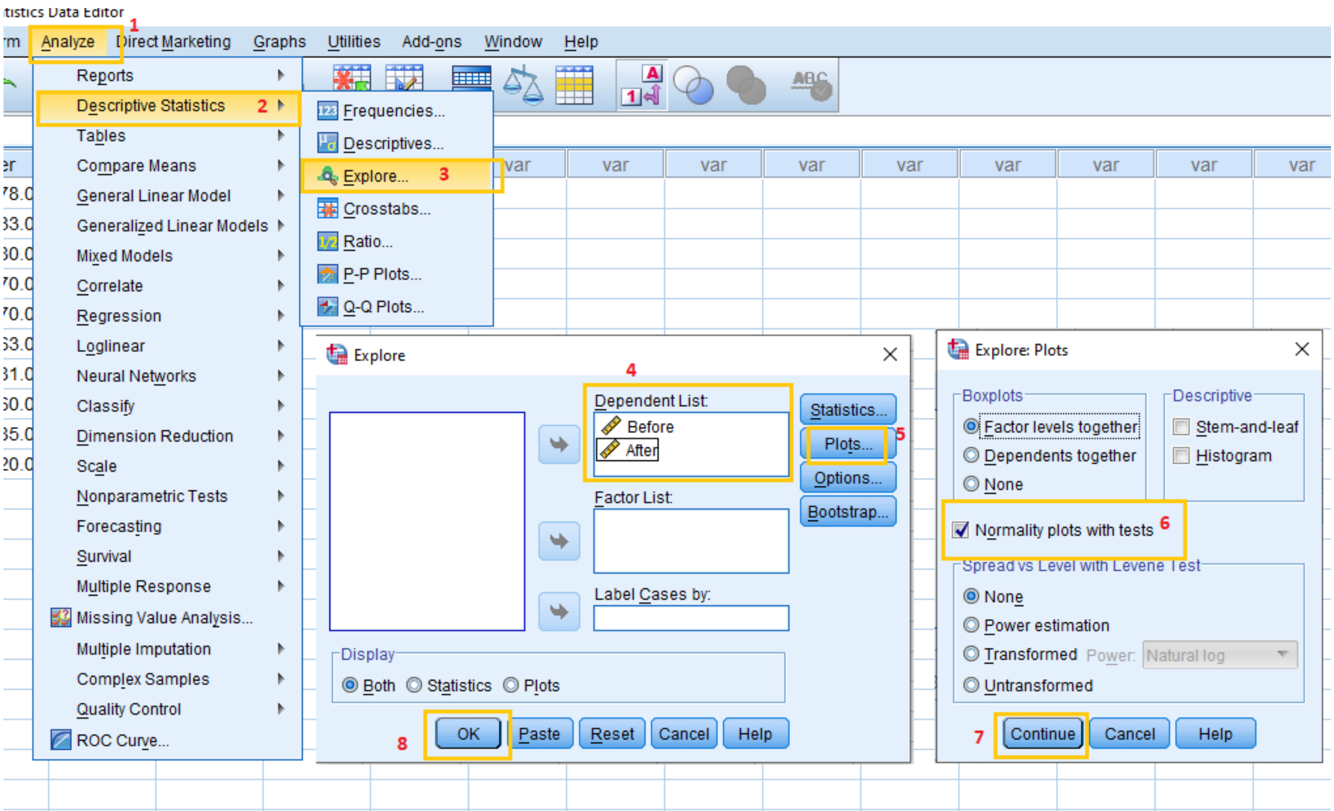
To use the **Paired-Samples T-Test**, we need to make sure that the population follows a normal distribution:

H_0: **the population follows a normal distribution**

Vs

H_1: **the population does not follow a normal distribution**

However, we find the question said that the population follows a normal distribution, so is not necessary to make this test



Tests of Normality

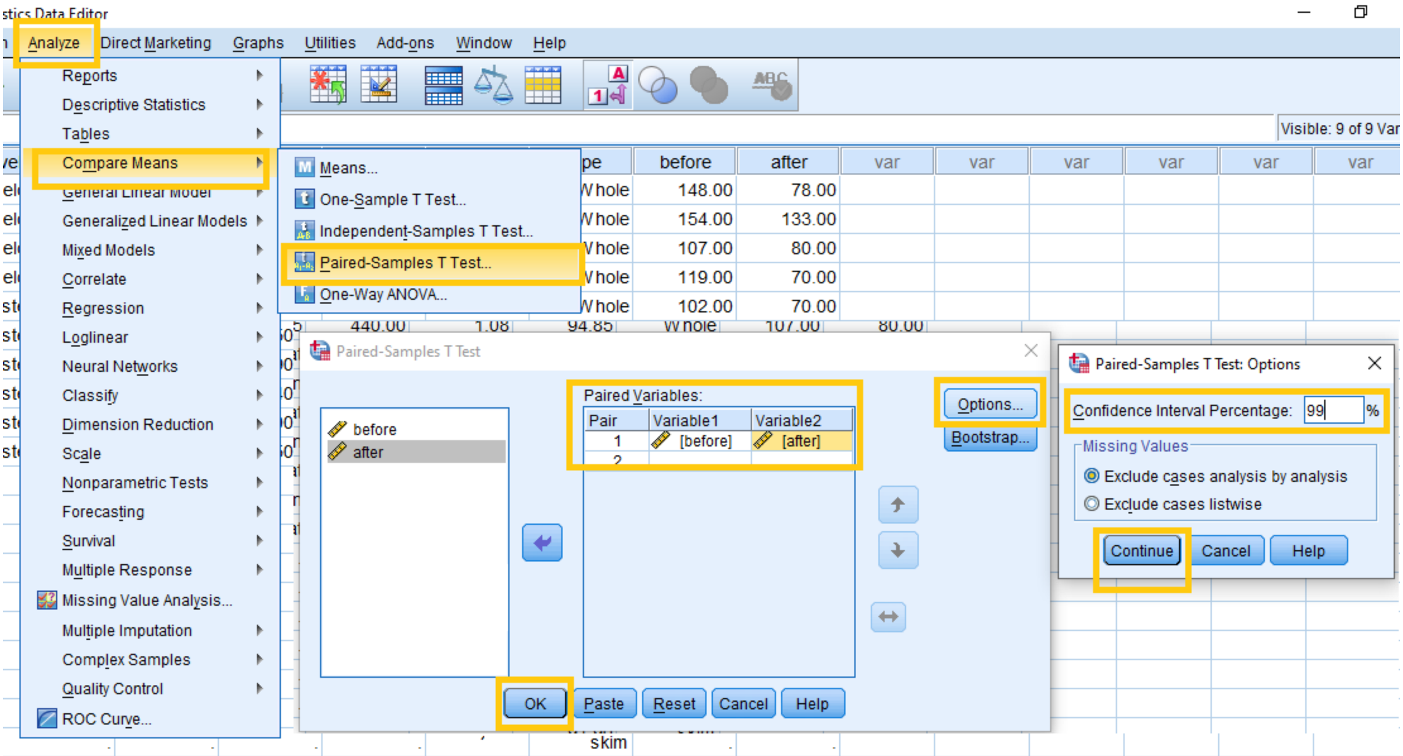
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Before	.183	10	.200*	.946	10	.620
After	.283	10	.022	.825	10	.029

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

p-value >0.01, Accept H0

Now, 99 % Confidence interval for μD and test $\mu_D = 0$ versus $\mu_D \neq 0$:



T-Test

[DataSet0]

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	before	128.6000	10	17.62700	5.57415
	after	84.0000	10	23.96293	7.57775

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	before & after	10	.233	.517

Test statistic : $T = \frac{\bar{D}}{S_D \div \sqrt{n}}$

degree of freedom : n-1 =9

Paired Samples Test

D=before –after		Paired Differences					t	df	Sig. (2-tailed)
		$\bar{D}$ Mean	SD_D Std. Deviation	Std. Error Mean	99% Confidence Interval of the Difference Lower	Upper			
Pair 1	before - after	44.60000	26.23484	8.29618	17.63877	71.56123	5.376	9	.000

$p - value = 0 \leq \alpha = 0.01$
So ,we Reject H_0

Q5: Ten Corvettes between 1 and 6 years old were randomly selected from last year's sales records in Virginia Beach, Virginia. The following data were obtained, where x denotes age, in years, and y denotes sales price, in hundreds of dollars.

X	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

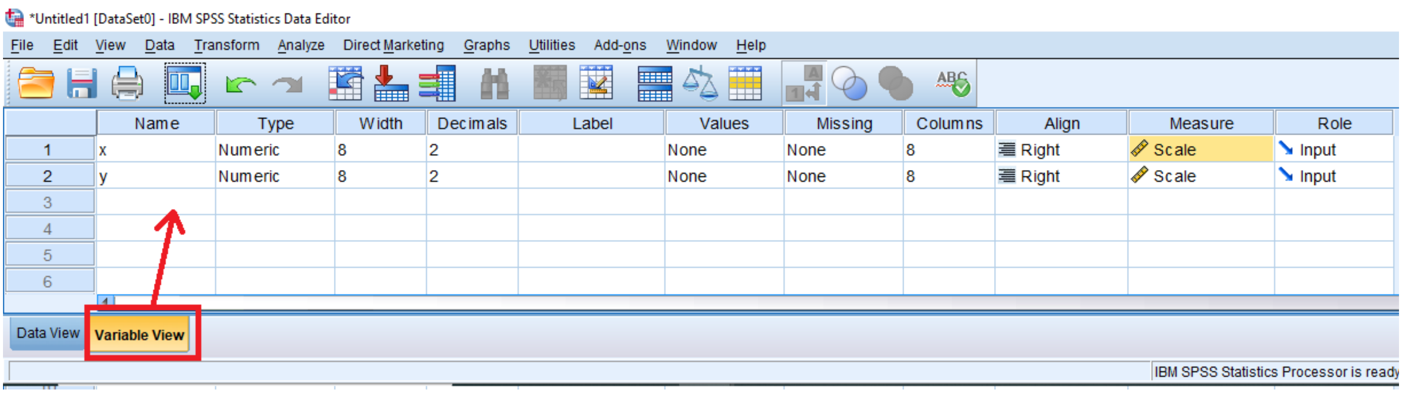
a) Compute and interpret the linear correlation coefficient, r.

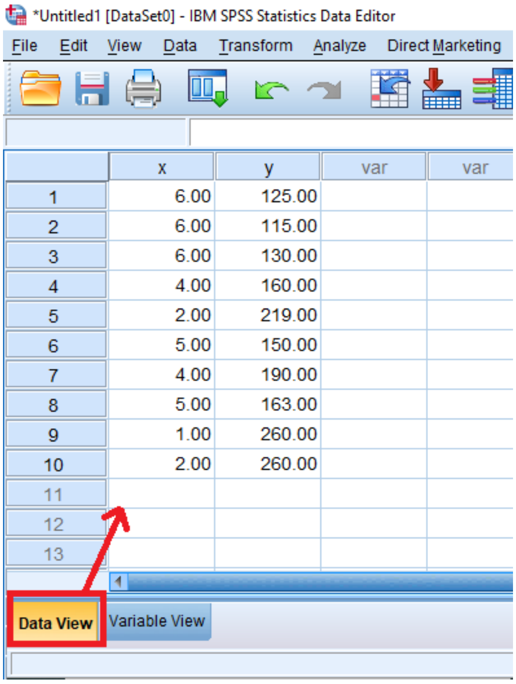
b) Determine the regression equation for the data.

c) Compute and interpret the coefficient of determination, r^2.

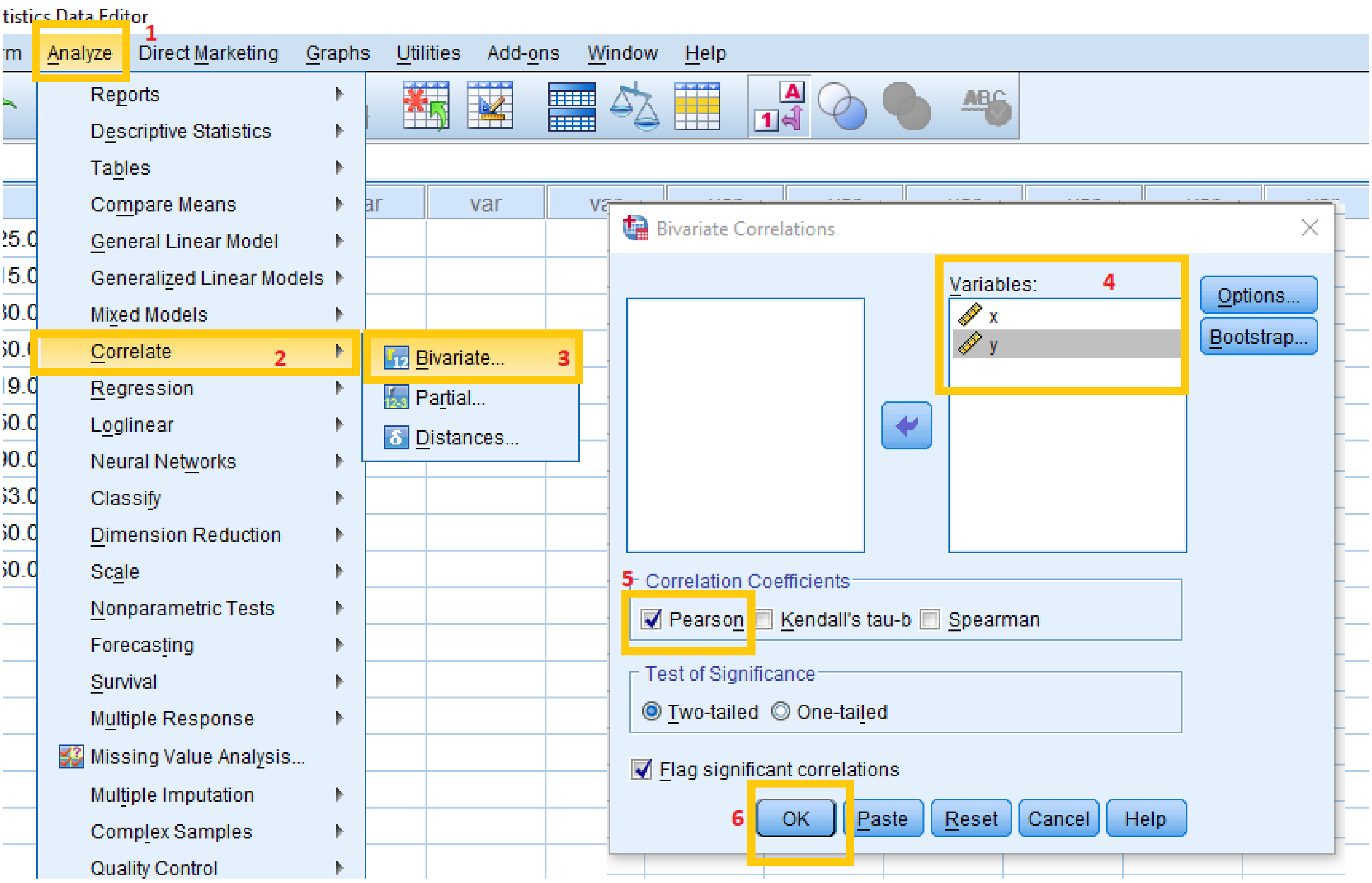
d) Obtain a point estimate for the mean sales price of all 4-year-old Corvettes.

Enter the age values (x) into one variable and the corresponding sales price values (y) into another variable (see figure, below).





a) Select Analyze → Correlate → Bivariate… (see figure, below).



Correlations

[DataSet0]

Correlations

		x	y
x	Pearson Correlation	1	-.968-**
	Sig. (2-tailed)		.000
	N	10	10
y	Pearson Correlation	-.968-**	1
	Sig. (2-tailed)	.000	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

r = - 0.968

strong negative

The correlation coefficient is –0.968 which we can see that the relationship between x and y are negative and strong.

Testing the Significance of the Correlation Coefficient :

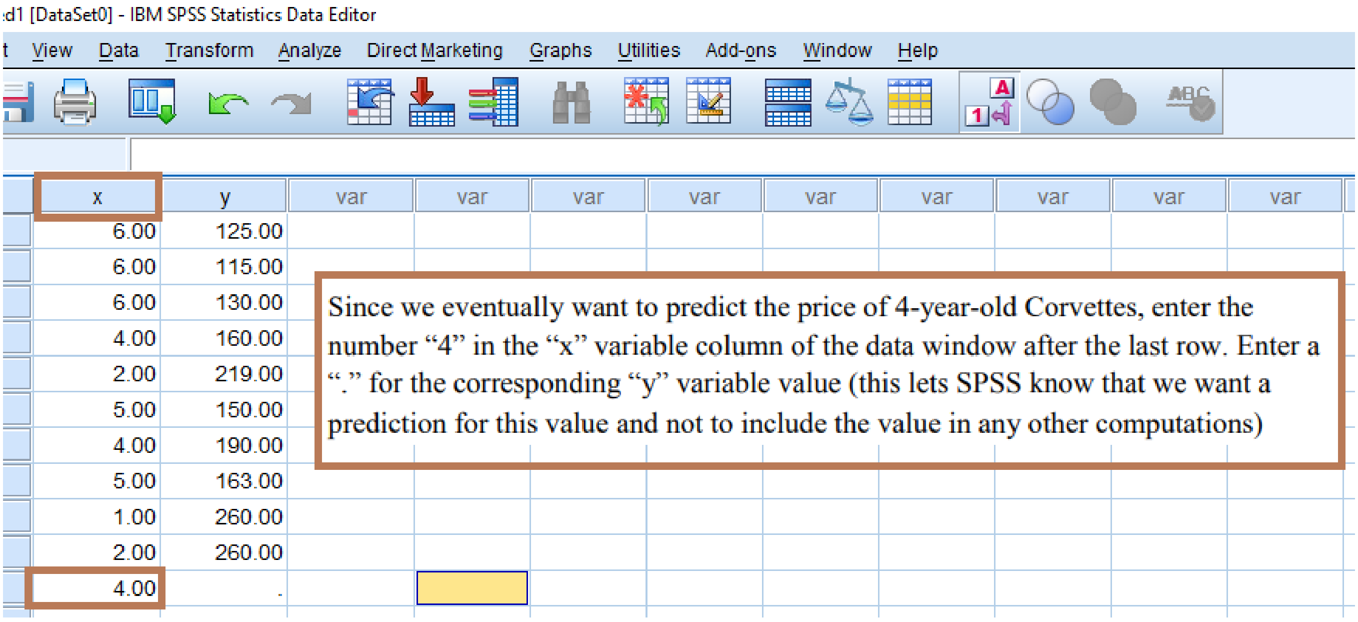
Null Hypothesis: H_0: $\rho = 0$ Vs Alternate Hypothesis: H_a: $\rho \neq 0$

Null Hypothesis H_0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X and Y in the population.

Alternate Hypothesis H_a: The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X and Y_2 in the population.

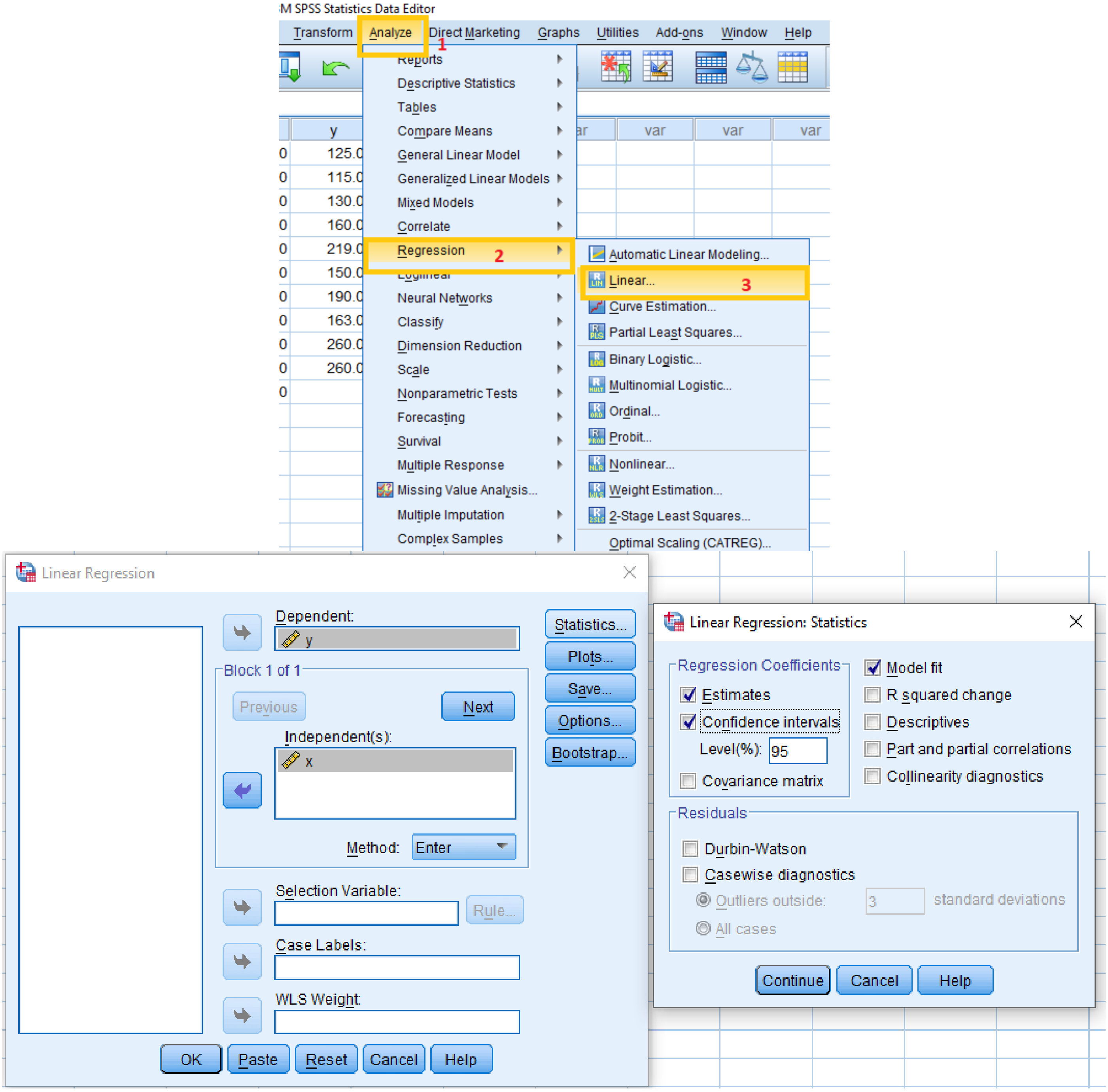
p-value = 0.00 < α = 0.05 . so , we reject H_0: $\rho = 0$.

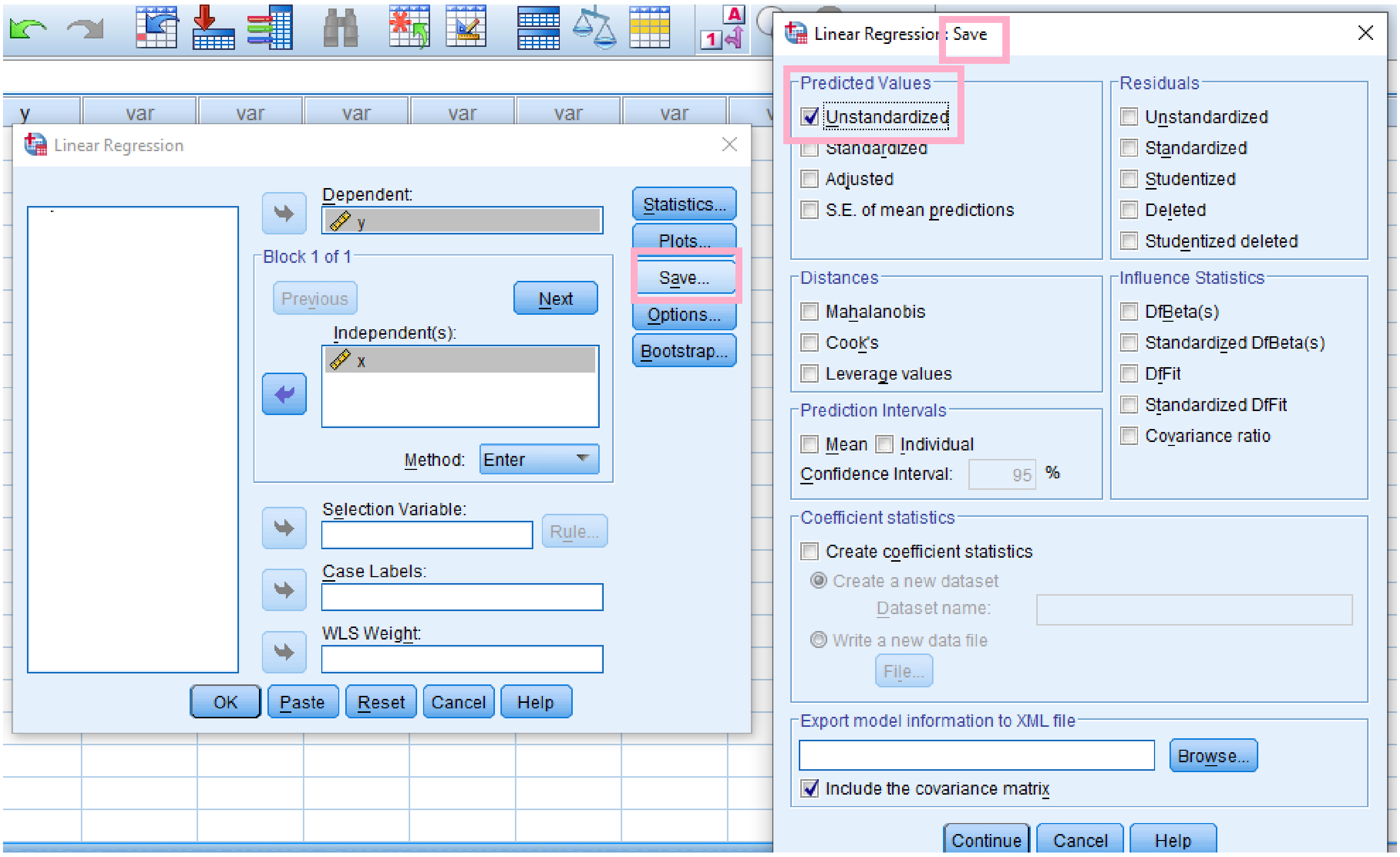
b, c and d)



Select Analyze → Regression → Linear… (see figure).

Select "y" as the dependent variable and "x" as the independent variable. Click "Statistics", select "Estimates" and "Confidence Intervals" for the regression coefficients, select "Model fit" to obtain r^2, and click "Continue". Click "Save…", select "Unstandardized" predicted values and click "Continue". Click "OK".


M SPSS Statistics Data Editor
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
1
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
2
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Automatic Linear Modeling...
Linear...
3
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
y
var
var
var
125.0
115.0
130.0
160.0
219.0
150.0
190.0
163.0
260.0
260.0
Linear Regression
Dependent:
y
Block 1 of 1
Previous
Next
Independent(s):
x
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Level(%):
95
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside:
3
standard deviations
All cases
Continue
Cancel
Help


Linear Regression
Dependent:
y
Block 1 of 1
Previous
Next
Independent(s):
x
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help

Regression

[DataSet0]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	x[b]	.	Enter

a. Dependent Variable: y

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.968[a]	.937	.929	14.24653

a. Predictors: (Constant), x

b. Dependent Variable: y

coefficient of determination : $r^2 = 0.937$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24057.891	1	24057.891	118.533	.000[b]
	Residual	1623.709	8	202.964		
	Total	25681.600	9			

a. Dependent Variable: y

b. Predictors: (Constant), x

Coefficients[a]

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	291.602	11.433		25.506	.000	265.238	317.966
	x	-27.903-	2.563	-.968-	-10.887-	.000	-33.813-	-21.993-

a. Dependent Variable: y

Regression equation : $y = 291.602 - 27.903\, x$

7 :

	x	y	PRE_1	var	var	var
1	6.00	125.00	124.18447			
2	6.00	115.00	124.18447			
3	6.00	130.00	124.18447			
4	4.00	160.00	179.99029			
5	2.00	219.00	235.79612			
6	5.00	150.00	152.08738			
7	4.00	190.00	179.99029			
8	5.00	163.00	152.08738			
9	1.00	260.00	263.69903			
10	2.00	260.00	235.79612			
11	4.00	.	179.99029			
12						
13						
14						

a point estimate for the mean sales price of all 4-year-old Corvettes

y = 179.99029

Data View | Variable View

Results:

b) The regression equation : $\hat{y}$ **= sales price = 291.6019 - 27.9029 * age .**

In other words, for increasing the age by one, the sales price <u>decreasing</u> by 27.9029 , while there is 291.6019 of Y does not depend on the age .

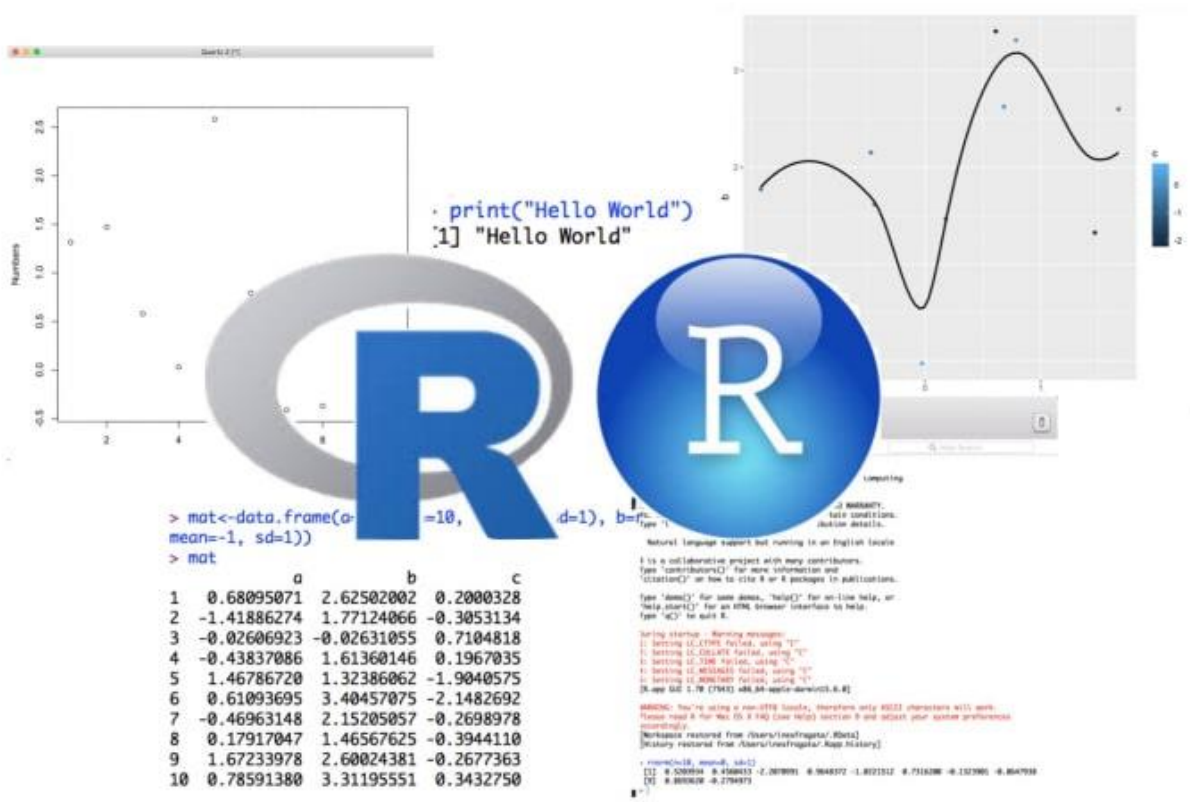
c) $r^2 = 0.9367$

The coefficient of determination is 0.9368; therefore, about 93.68% of the variation in the price data is explained by age. The regression equation appears to be very useful for making predictions since the value of r^2 is close to 1.

d) The predicted sales price is 17999.0291 dollars ($17,999.029).

-

R Programming



R-Part 1

#Mathematical functions :

Q1: Write the command and the result to calculate the following :

Log(17)=

```
> log10(17)
[1] 1.230449
> log(17,base=10)
[1] 1.230449
>
```

Ln(14)=

```
> log(14)
[1] 2.639057
>
```

$\binom{50}{4} =$

```
> choose(50,4)
[1] 230300
>
```

$\Gamma(18)$,

```
> gamma(18)
[1] 3.556874e+14
>
```

4!=

```
> factorial(4)
[1] 24
```

$2^3 =$

```
> 2^3
[1] 8
> 2**3
[1] 8
>
```

$\sqrt{16} =$

```
> sqrt(16)
[1] 4
>
```

|-4| =

```
> abs(-4)
[1] 4
>
```

Q2: Let x=6 and y=2 find:

$x + y$, $x - y$, $x \div y$, xy , z= x y -1

```
> x
[1] 6
> y
[1] 2
> x<- 6
> y<- 2
> x+y
[1] 8
> x-y
[1] 4
> x/y
[1] 3
> x*y
[1] 12
> z<- x*y-1
> z
[1] 11
```

Vector :

Q3: If $a = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$ **, b=** $\begin{bmatrix} 6 \\ 7 \\ 8 \\ 9 \end{bmatrix}$ **. find :**

$$a + b \quad , \quad a - b \quad , \quad ab \quad , a \div b \; , \; 2a \quad , \quad b + 1$$

```
> a=c(1,2,3,3)
> b=c(6,7,8,9)
> a
[1] 1 2 3 3
> b
[1] 6 7 8 9
> a+b
[1]  7  9 11 12
> a-b
[1] -5 -5 -5 -6
> a*b
[1]  6 14 24 27
> a/b
[1] 0.1666667 0.2857143 0.3750000 0.3333333
> 2*a
[1] 2 4 6 6
> b+1
[1]  7  8  9 10
```

ls() is a function in **R** that lists all the object in the working environment.

rm() deletes (removes) a variable from a workspace.

Matrices:

Q3: write the commends and results to find the <u>determent</u> of matrix and its <u>inverse</u>

$$w = \begin{bmatrix} 1 & 7 & 2 \\ 2 & 7 & 2 \\ 4 & 0 & 2 \end{bmatrix}$$

عدد الصفوف

```
> w<-matrix(c(1,2,4,7,7,0,2,2,2),nr=3)
> w
     [,1] [,2] [,3]
[1,]    1    7    2
[2,]    2    7    2
[3,]    4    0    2
> #inverse
> solve(w)
           [,1]       [,2]       [,3]
[1,] -1.0000000  1.0000000  0.0000000
[2,] -0.2857143  0.4285714 -0.1428571
[3,]  2.0000000 -2.0000000  0.5000000
> #determent
> det(w)
[1] -14
> #Trnspose:
> t(w)
     [,1] [,2] [,3]
[1,]    1    2    4
[2,]    7    7    0
[3,]    2    2    2
>
```

لكتابة المصفوفة نستخدم الامر matrix

لايجاد المعكوس نستخدم الامر solve

لايجاد محدد المصفوفة نستخدم det

لايجاد منقول المصفوفة نستخدم الامر t

OR

```
> w<- cbind(c(1,2,4),c(7,7,0),c(2,2,2))
> w
     [,1] [,2] [,3]
[1,]    1    7    2
[2,]    2    7    2
[3,]    4    0    2
```

OR

```
> w<- rbind(c(1,7,2),c(2,7,2),c(4,0,2))
> w
     [,1] [,2] [,3]
[1,]    1    7    2
[2,]    2    7    2
[3,]    4    0    2
```

Q4:

$$A = \begin{bmatrix} 1 & 6 & 3 & -1 \\ 5 & 2 & 7 & 4 \end{bmatrix},\ B = \begin{bmatrix} 1 & 9 & 8 \\ 7 & 4 & 2 \\ 5 & 1 & 5 \\ 1 & 1 & 9 \end{bmatrix},\ C = \begin{bmatrix} 3 & 4 & 2 & 7 \\ 4 & 9 & 0 & 6 \\ 3 & 8 & 3 & 2 \\ 3 & 4 & 6 & 2 \end{bmatrix}$$

(a) A*B

(b) Determinant of C

(c) Inverse of C

```
> A<-matrix(c(1,5,6,2,3,7,-1,4),nr=2)
> A
     [,1] [,2] [,3] [,4]
[1,]    1    6    3   -1
[2,]    5    2    7    4
> B<-matrix(c(1,7,5,1,9,4,1,1,8,2,5,9),nr=4)
> B
     [,1] [,2] [,3]
[1,]    1    9    8
[2,]    7    4    2
[3,]    5    1    5
[4,]    1    1    9
> C<-matrix(c(3,4,3,3,4,9,8,4,2,0,3,6,7,6,2,2),nr=4)
> C
     [,1] [,2] [,3] [,4]
[1,]    3    4    2    7
[2,]    4    9    0    6
[3,]    3    8    3    2
[4,]    3    4    6    2
> A%*%B
     [,1] [,2] [,3]
[1,]   57   35   26
[2,]   58   64  115
> det(C)
[1] -155
> solve(C)
           [,1]       [,2]       [,3]        [,4]
[1,] -1.0451613  1.3677419 -1.5870968  1.14193548
[2,]  0.1935484 -0.2903226  0.5161290 -0.32258065
[3,]  0.2580645 -0.3870968  0.3548387 -0.09677419
[4,]  0.4064516 -0.3096774  0.2838710 -0.27741935
> 
```

Q5: A sample of families were selected and the number of children in each family was considered as follows:

6, 7, 0, 8, 3, 7, 8, 0

Find mean , median , range , variance , standard deviation?

```
> xx<-c(6,7,0,8,3,7,8,0)
> xx
[1] 6 7 0 8 3 7 8 0
> mean(xx)
[1] 4.875
> median(xx)
[1] 6.5
> var(xx)
[1] 11.55357
> sd(xx)
[1] 3.399054
> summary(xx)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.250   6.500   4.875   7.250   8.000
> range(xx)
[1] 0 8
>
```

R-Part 2

Q1: We have grades of 7 students in the following table

math	73	45	32	85	98	78	82
stat	87	60	25	64	72	12	90

Find

1) summary of math and stat grades

```
> math<- c(73,45,32,85,98,78,82)
> stat<- c(87,60,25,64,72,12,90)
> grades<-matrix(c(math,stat),nc=2)
> grades
     [,1] [,2]
[1,]   73   87
[2,]   45   60
[3,]   32   25
[4,]   85   64
[5,]   98   72
[6,]   78   12
[7,]   82   90
```

OR

```
> math=c(73,45,32,85,98,78,82)
> stat=c(87,60,25,64,72,12,90)
>
> df<-data.frame(math,stat)
> df
  math stat
1   73   87
2   45   60
3   32   25
4   85   64
5   98   72
6   78   12
7   82   90
```

```
> df3<- cbind(math,stat)
> df3
     math stat
[1,]   73   87
[2,]   45   60
[3,]   32   25
[4,]   85   64
[5,]   98   72
[6,]   78   12
[7,]   82   90
>
```

```
> apply(grades,2,summary)
            [,1]     [,2]
Min.    32.00000 12.00000
1st Qu. 59.00000 42.50000
Median  78.00000 64.00000
Mean    70.42857 58.57143
3rd Qu. 83.50000 79.50000
Max.    98.00000 90.00000
```

2) Summary of each student grade

```
> apply(grades,1,summary)
        [,1]  [,2]  [,3]  [,4] [,5] [,6] [,7]
Min.    73.0 45.00 25.00 64.00 72.0 12.0   82
1st Qu. 76.5 48.75 26.75 69.25 78.5 28.5   84
Median  80.0 52.50 28.50 74.50 85.0 45.0   86
Mean    80.0 52.50 28.50 74.50 85.0 45.0   86
3rd Qu. 83.5 56.25 30.25 79.75 91.5 61.5   88
Max.    87.0 60.00 32.00 85.00 98.0 78.0   90
```

3) Summary of first five student grades in math

```
> summary(math[1:5])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   32.0    45.0    73.0    66.6    85.0    98.0
> summary(math[-(6:7)])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   32.0    45.0    73.0    66.6    85.0    98.0
```

استدعاء بيانات من R و حساب بعض الاحصاءات

Q2: Growth of Orange Trees

Description

The **Orange** data frame has 35 rows and 3 columns of records of the growth of orange trees.

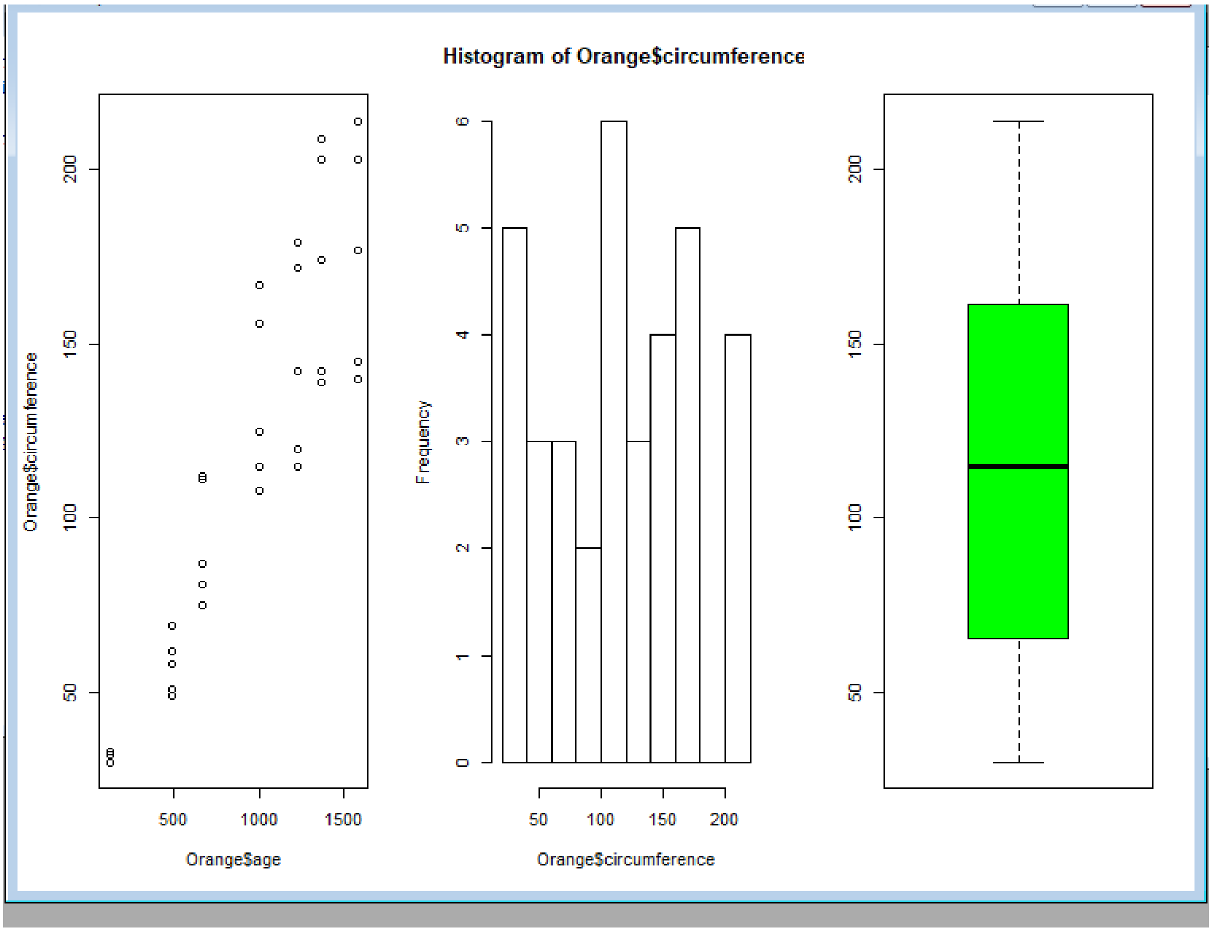
```
> Orange
   Tree  age circumference
1     1  118            30
2     1  484            58
3     1  664            87
.     .  .              .
.     .  .              .
.     .  .              .
30    5  484            49
31    5  664            81
32    5 1004           125
33    5 1231           142
34    5 1372           174
35    5 1582           177
```

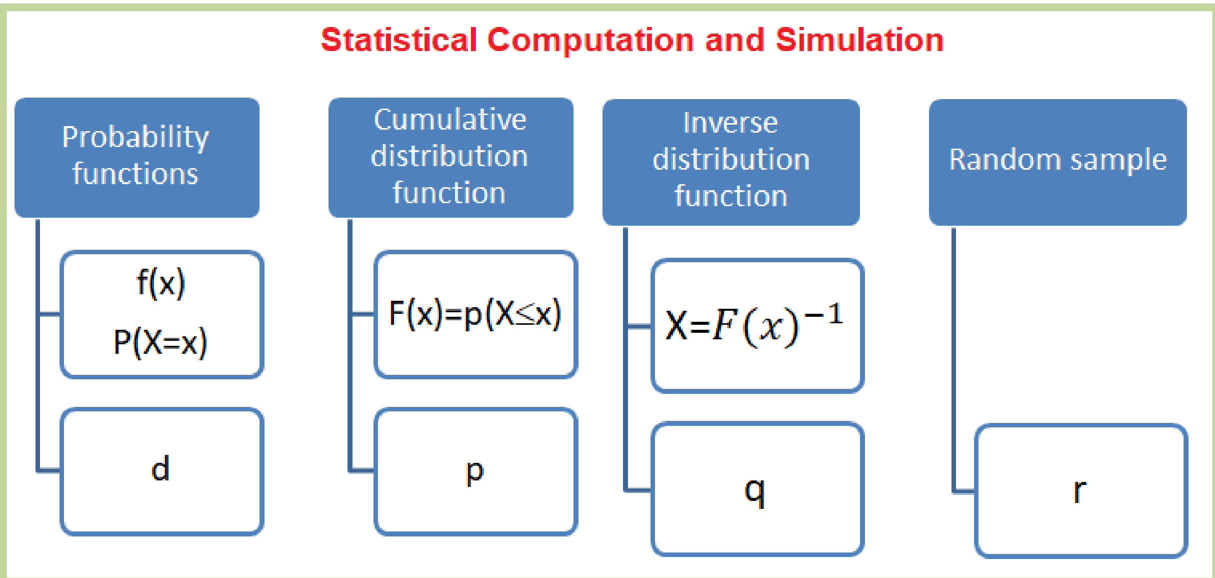
```
> attach(Orange)
> mean(age)
[1] 922.1429
> summary(circumference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   30.0    65.5   115.0   115.9   161.5   214.0 
```

OR

```
> mean(Orange$age)
[1] 922.1429
> summary(Orange$circumference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   30.0    65.5   115.0   115.9   161.5   214.0 
```

```
> par(mfcol=c(1,3))
> plot(Orange$age,Orange$circumference)
> hist(Orange$circumference)
> boxplot(Orange$circumference,col="green")
```


Histogram of Orange$circumference
Orange$circumference
Orange$age
Frequency
Orange$circumference


Statistical Computation and Simulation
Probability functions
f(x)
P(X=x)
d
Cumulative distribution function
F(x)=p(X≤x)
p
Inverse distribution function
X=F(x)^−1
q
Random sample
r

Q3: Suppose X is Normal with mean 2 and standard deviation 0.25 . Find:

1-$F(2.5) = P(X \leq 2.5)$

2-$F^{-1}(0.90)$ or $P(X \leq x) = 0.90$

3- Generate a random sample with size 10 from $N(2\ ,0.25^2)$ distribution ?

```
> # 1) F(2.5)
> pnorm(2.5,2,0.25)
[1] 0.9772499
>
> # 2) P(x<= x)= 0.90
> qnorm(0.90,2,0.25)
[1] 2.320388
>
> # 3) Generate a random sample with size 10
> rnorm(10,2,0.25)
 [1] 1.988027 1.744937 1.821131 2.049191 2.092522 1.992336 2.419941 2.270132
 [9] 1.709938 2.009987
```

Q4: A biased coin is tossed 6 times . The probability of heads on any toss is 0.3 . Let X denote the number of heads that come up. Find :

1-P(x=2)

2- $P(1< X \leq 5)$ = $P(X\leq 5)$- $P(X\leq 1)$

```
> #Binomial Distribution:
> # 1) P(X=2):
> dbinom(2,6,0.3)
[1] 0.324135
>
> # 2)P( 1< x<= 5):
>  pbinom(5,6,0.3)-pbinom(1,6,0.3)
[1] 0.579096
```

Q5: write the commends and results to calculate the following

1. $P(-1 < T < 1.5),\ v = 10$
2. Find k such that $P(T < k) = 0.025,\ v = 12$
3. Generate a random sample of size 12 from the exponential(3)
4. Find k such that $P(X > k) = 0.04,\ X \sim F(12, 10)$
5. $P(3 < X \leq 7),\ X \sim Poisson(3)$

```
> # 1)P(-1<T<1.5),v=10
> pt(1.5,10)-pt(-1,10)
[1] 0.7472998
>
> # 2)Find k such that  P(T<k)=0.025,v=12
> qt(0.025,12)
[1] -2.178813
>
> # 3)Generate a random sample of size 12 from the exponential(3)
> rexp(12,3)
 [1] 0.02741723 0.57916093 0.43225608 0.58069241 0.10705782 0.27219276
 [7] 0.66971690 0.07028167 0.28315394 0.65606893 0.35302758 0.05820528
>
> # 4) Find k such that P(X>k)=0.04, X~F(12,10)
> qf(1-0.04,12,10)
[1] 3.131479
>
> # 5)P(3<X≤7),X~Poisson(3)
> ppois(7,3)-ppois(3,3)
[1] 0.3408636
```

Q6: We have the following table show age X and blood pressure Y of 8 women

X	68	49	60	42	55	63	36	42
Y	152	145	155	140	150	140	118	125

```
> x<-c(68,49,60,42,55,63,36,42)
> y<-c(152,145,155,140,150,140,118,125)
```

1. **Plot X and Y**

```
> # 1)Plot X and Y:
> plot(x,y)
> plot(x,y,type="l")
> plot(x,y,type="b")
> plot(x,y,type="h")
> qqnorm(x)
> hist(x)
> boxplot(x)
>
```

2. **correlation of X and Y**

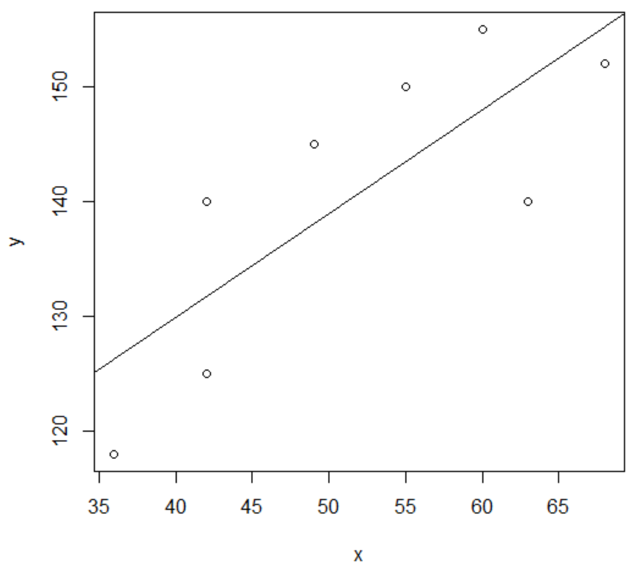
```
> # 2)correlation of X and Y:
> cor(x,y)
[1] 0.7918318
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 3.1758, df = 6, p-value = 0.01918
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1971842 0.9605402
sample estimates:
      cor
0.7918318
```

3. **covariance**

```
> # 3)covariance:
> cov(x,y)
[1] 118.5179
```

4. The equation of regression

```
> # 4)The equation of regression:
> fit<-lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-10.713  -7.060   1.647   6.988   8.330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.5838    15.1239   6.188  0.00082 ***
x             0.9068     0.2855   3.176  0.01918 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.637 on 6 degrees of freedom
Multiple R-squared:  0.627,     Adjusted R-squared:  0.5648
F-statistic: 10.09 on 1 and 6 DF,  p-value: 0.01918

> plot(x,y)
> abline(fit)
>
```



Regression Equation:

$Y = 93.5838 + 0.9068\ x$

R-Part 3

Q1: For a sample of 10 fruits from thirteen-year-old acidless orange trees, the fruit shape (determined as a diameter divided by height) was measured [Shaheen and Hamouda (8419b)]:

1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

Assuming that fruit shapes are approximately normally distributed, Test whether the mean of fruit shape greater than 1.02 . Use α=0.05

1-Hypothesis :

$$H_{0:}\ \mu \leq 1.02 \quad vs \quad H_{1:}\ \mu > 1.02$$

2-Test statistics :

T=2.6849

3- Decision:

$$p - value = 0.0125 < \alpha = 0.05$$

So, we reject $H_{0:}\ \mu \leq 1.02$

```
> x<-c(1.07,1.08,1.07,1.05,1.06,1.02,1.04,1.05,1.04,0.976)
>
>  t.test(x,mu=1.02,alternative='greater',conf.level=0.95)

        One Sample t-test

data:  x
t = 2.6849, df = 9, p-value = 0.0125
alternative hypothesis: true mean is greater than 1.02
95 percent confidence interval:
 1.028121      Inf
sample estimates:
mean of x
   1.0456
```

One sample t-test

t.test(x , mu= a , alternative=" ",conf.level= 1-α)

$H_0: \mu \ \{=, \geq, \leq\}\ a$

If : $H_1: \neq$ two.sided

If : $H_1: <$ less

If : $H_1: >$ greater

Q2: The phosphorus content was measured for independent samples of skim and whole:

Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

Assuming normal populations with equal variance

a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**

b) **Find and interpret a 99% confidence interval for the difference in average phosphorus contents of whole and skim milk .**

a)

1- **Hypothesis :**

$$H_0: \mu_{Skim} \geq \mu_{Whole} \quad vs \quad H_1: \mu_{Skim} < \mu_{Whole}$$

$$H_0: \mu_{Skim} - \mu_{Whole} \geq 0 \quad vs \quad H_1: \mu_{Skim} - \mu_{Whole} < 0$$

2- **Test statistic :** T= -14.162

3- **Decision:**

Since p-value =0.00 < α= 0.01 . we reject H_0

```
> W<-c(94.95,95.15,94.85,94.55,93.4,95.05,94.35,94.70,94.90)
> S<-c(91.25,91.80,91.50,91.65,91.15,90.25,91.90,91.25,91.65,91)
>  t.test (S ,W ,alternative="less",conf.level=0.99)

        Welch Two Sample t-test

data:  S and W
t = -14.162, df = 16.294, p-value = 6.999e-11
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
      -Inf -2.711906
sample estimates:
mean of x mean of y 
 91.34000  94.65556 
```

b) $\mu_{Skim} - \mu_{Whole} \in (-3.99\ ,-2.63\ )$

```
> t.test (S ,W,conf.level=0.99)

        Welch Two Sample t-test

data:  S and W
t = -14.162, df = 16.294, p-value = 1.4e-10
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -3.997738 -2.633373
sample estimates:
mean of x mean of y
 91.34000  94.65556

> t.test (S ,W ,alternative="two.sided",conf.level=0.99)

        Welch Two Sample t-test

data:  S and W
t = -14.162, df = 16.294, p-value = 1.4e-10
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -3.997738 -2.633373
sample estimates:
mean of x mean of y
 91.34000  94.65556
```

For confidence interval we change alternative to not equal

Two independent sample t-test

t.test(x,y , mu= a , alternative=" ",conf.level= 1-α , var.equal =)

$H_0: \mu_x - \mu_y \; (=, \geq, \leq) \; a$

If : $H_1: \neq$ two.sided

If : $H_1: <$ less

If : $H_1: >$ greater

TRUE

FALSE

Q3: In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

Before surgery (X)	**148**	**154**	**107**	**119**	**102**	**137**	**122**	**140**	**140**	**117**
After surgery (Y)	**78**	**133**	**80**	**70**	**70**	**63**	**81**	**60**	**85**	**120**

We assume that the data comes from normal distribution. Find :

1- 99% confidence interval for μD, where μD is the difference in the average weight before and after surgery.

2- Test whether the data provide sufficient evidence to indicate a difference in the average weight before and after surgery. ($\mu_D = 0$ versus $\mu_D \neq 0$)

a) 1- Hypothesis:

$\mu_D = 0$ vs $\mu_D \neq 0$

2- Test Statistic :

T= 5.376

3- Decision:

Since p-value =0.00 < α= 0.05 . we reject H_0

b) 99% $C.I$ $\mu_D \in (17.638 \ , \ 71.56)$

```
> x<- c(148,154,107,119,102,137,122,140,140,117)
> y<-c(78,133,80,70,70,63,81,60,85,120)
>
> t.test(x,y,alternative="two.sided",conf.level=0.99,paired=TRUE)

        Paired t-test

data:  x and y
t = 5.376, df = 9, p-value = 0.0004469
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 17.63877 71.56123
sample estimates:
mean of the differences
                   44.6
```

$\bar{D} = 44.6$

Paired t-test

t.test(x,y , mu= a , alternative=" ",conf.level= 1-α ,paired=T)

$H_0: \mu_D \; = \; \geq \; \leq \; a$

If : $H_1: \neq$ two.sided

If : $H_1: <$ less

If : $H_1: >$ greater

Q4: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results

Observation	Programs 1	Programs 2	Programs 3	Programs 4
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8

```
> x<-c(9,12,14,11,13,10,6,9,9,10,12,14,11,13,11,9,8,11,7,8)
> y<-c("1","1","1","1","1","2","2","2","2","2","3","3","3","3","3","4",
+ "4","4","4","4")
>
> model<-aov(x~y)
> summary(model)
            Df Sum Sq Mean Sq F value  Pr(>F)
y            3  54.95   18.32   7.045 0.00311 **
Residuals   16  41.60    2.60
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
```

1-Hypothesis :

$$H_0: \mu_{program1} = \mu_{program2} = \mu_{program3} = \mu_{program4}$$

$$H_1: at\ least\ one\ mean\ is\ diffrenet$$

2- **Test statistic :**

F = 7.045

3- **p-value = 0.00311 < α=0.05 , Reject** $H_0: \mu_{program1} = \mu_{program2} = \mu_{program3} = \mu_{program4}$

We use Tukey test to determine which means different:

```
> m<-TukeyHSD(model)
> m
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = x ~ y)

$y
    diff        lwr         upr     p adj
2-1 -3.0 -5.9176792 -0.08232082 0.0427982
3-1  0.4 -2.5176792  3.31767918 0.9788127
4-1 -3.2 -6.1176792 -0.28232082 0.0291638
3-2  3.4  0.4823208  6.31767918 0.0197459
4-2 -0.2 -3.1176792  2.71767918 0.9972140
4-3 -3.6 -6.5176792 -0.68232082 0.0133087

>
> boxplot(x~y)
```

$\mu_{program\,1} \neq \mu_{program\,2}$

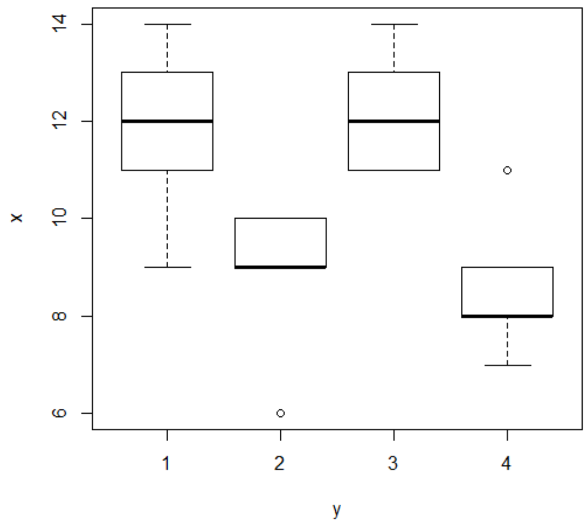
$\mu_{program\,1} = \mu_{program\,3}$

$\mu_{program\,1} \neq \mu_{program\,4}$

$\mu_{program\,2} \neq \mu_{program\,3}$

$\mu_{program\,2} = \mu_{program\,4}$

$\mu_{program\,3} \neq \mu_{program\,4}$



1- $\int_0^1 x^5(1-x)^4dx$

```
> f<-function(x){
+   (x^5)*(1-x)^4
+ }
> integrate(f,0,1)
0.0007936508 with absolute error < 8.8e-18
>
>
> beta(6,5)
[1] 0.0007936508
```

2- $\int_0^1 x^5(1-x)^4dx$

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

$\alpha-1=5$ and $\beta-1=4$

$\alpha=6$ | $\beta=5$

⇨ B(6 , 5)