Department of Statistics and Operations Research

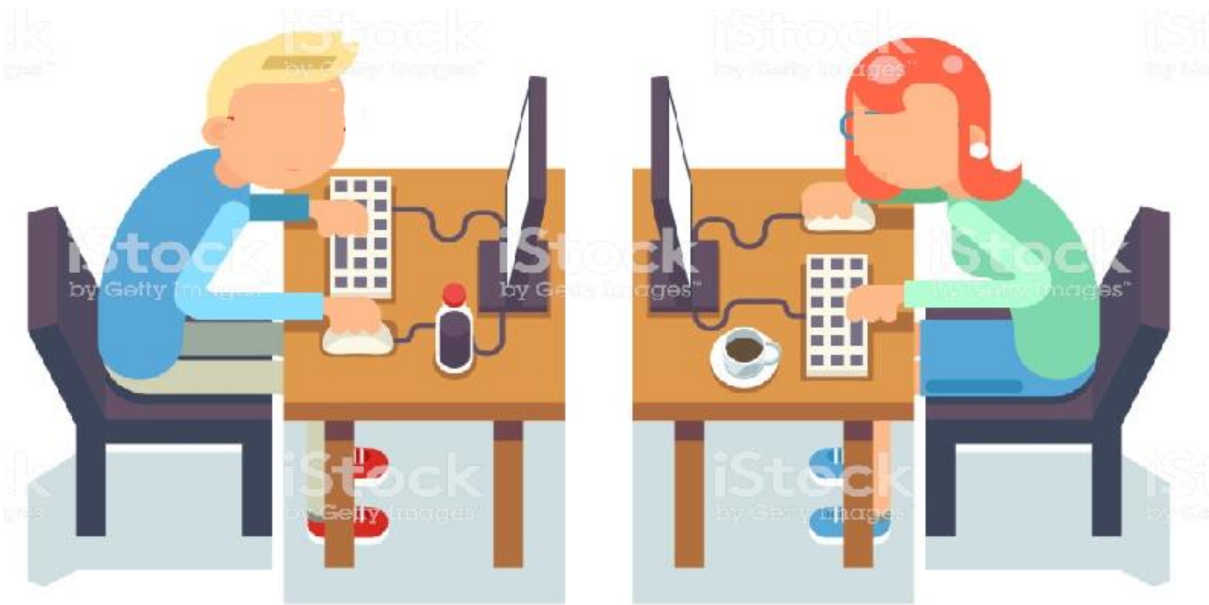
College of Science

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

	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	

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:

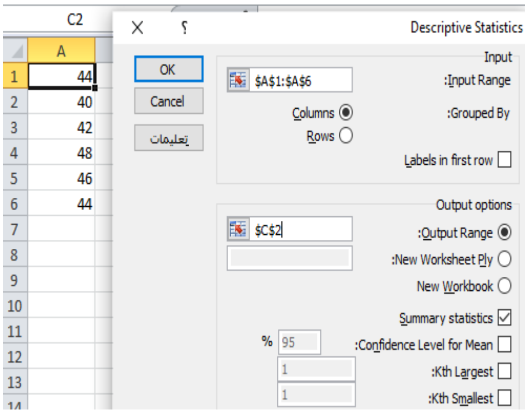
Mean=44	AVERAGE(C2:C7)
Median=44	MEDIAN(C2:C7)
Mode=44	MODE.SNGL(C2:C7)
Sample standard deviation=2.828	STDEV.S(C2:C7)
Sample variance=8	VAR.S(C2:C7)
Kurtosis=-0.3	KURT(C2:C7)
Skewness=4.996E-17	SKEW(C2:C7)
Minimum=40	MIN(C2:C7)
Maximum=48	MAX(C2:C7)
Range=8	MAX(C2:C7)-MIN(C2:C7)
Count=6	COUNT(C2:C7)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

* Range= Maximum-Minimum

** Coefficient of variation= $\frac{\text{Sample standard deviation}}{\text{Mean}} \times 100$

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	
14			
15			
16			
17			
18			
19			
20			

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:

(i) If we call heads a success then this X has a binomial distribution with parameters $n = 6$ and $p = 0.3$.

$$P(X = 2) = \binom{6}{2}(0.3)^2(0.7)^4 = 0.324135$$

(ii)

$$P(X = 3) = \binom{6}{3}(0.3)^3(0.7)^3 = 0.18522.$$

(iii) We need $P(1 < X \leq 5)$

$$\begin{aligned} & P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) \\ = \; & 0.324 + 0.185 + 0.059 + 0.01 \\ = \; & 0.578 \end{aligned}$$

	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 ≥ 2)= 1-p(x< 2) = 1 - p(x ≤ 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 ≥ 2)= 1-p(x< 2) = 1 - p(x ≤ 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
1			
2			
3	A)	-2.144786688	
4			
5	B)	-2.763769458	
6			
7	C)	3.499483297	
8			

Find : A) $t_{0.025}$ $where$ $v = df = 14$

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

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

	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 $P(X = a)$

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

(short)

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

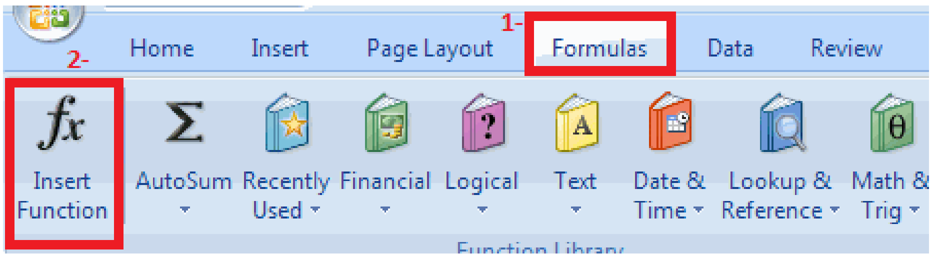
= name of distribution . inv (probability , parameter of distribution)

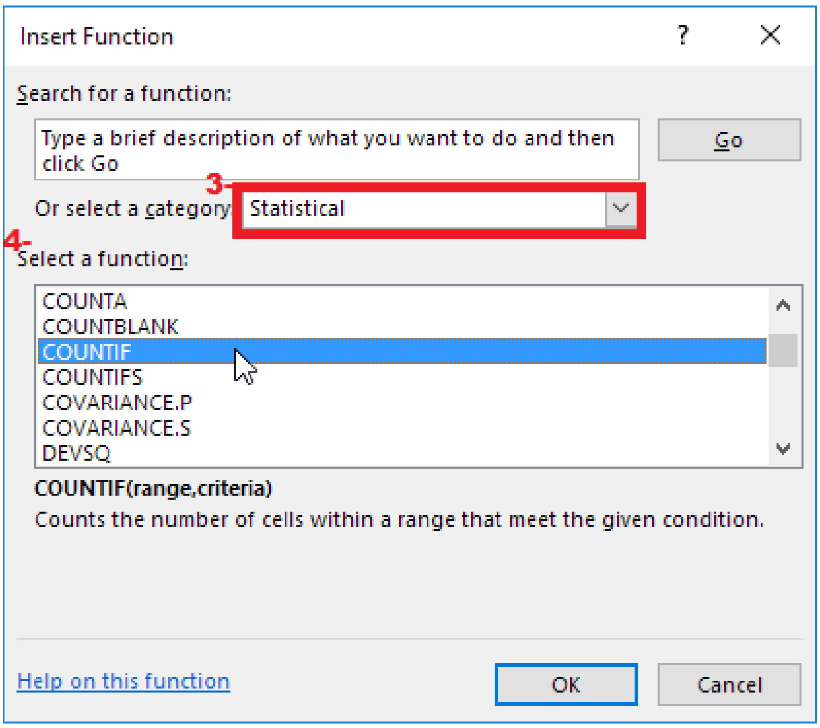
(short)

other ways:

you can find the functions of distribution from:

Formulas – Insert function



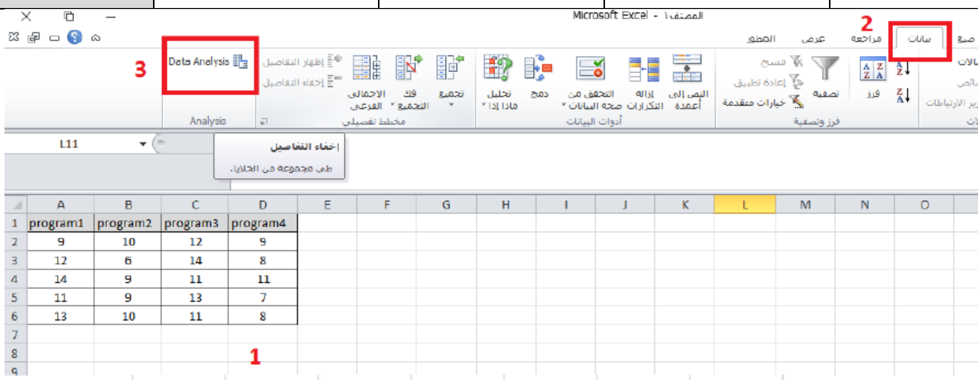


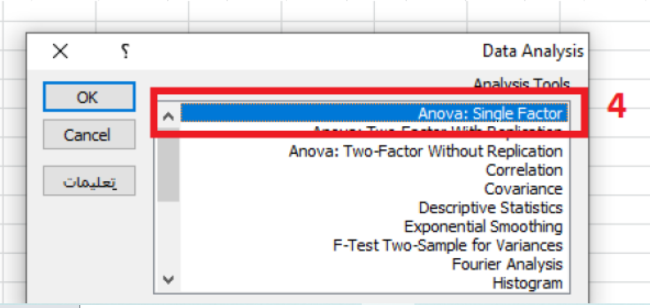
HYPOTHESIS TESTING STATISTICS AND CONFIDENCES INTERVAL

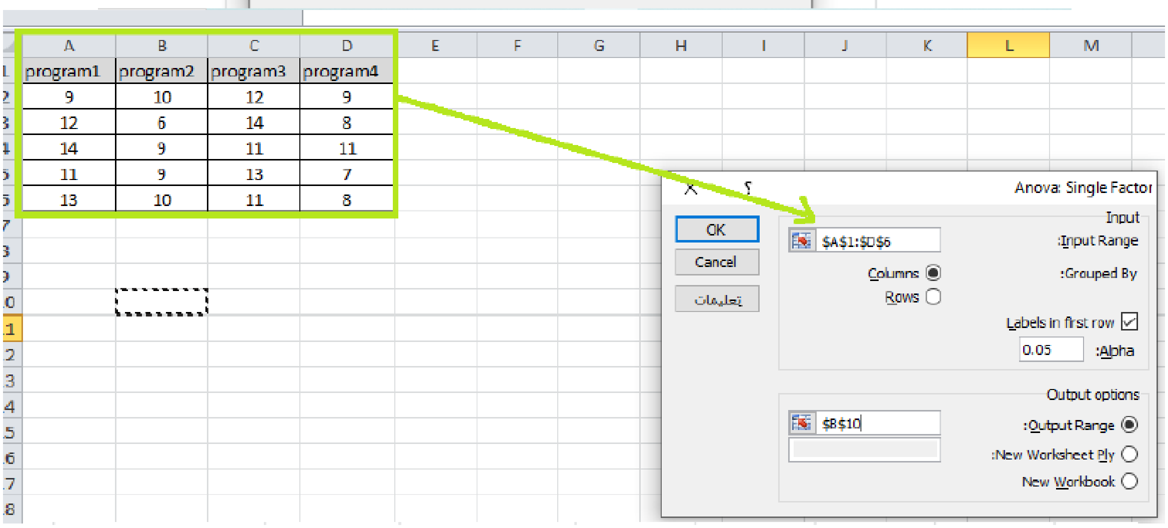
1-One way AVOVA :

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 Variation	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

Two Factor ANOVA with Replication:

2) In a study on fertilizer levels and spacing between plants, plots were assigned to combinations and the yield of potatoes (in kg/plot) was measured

Spacing between plants	Fertilizer level (in tons/ha)	
	1	2
25 cm	16.01 16.78 16.44	15.89 16.23 16.18
33 cm	13.42 13.25 13.32	13.32 13.47 13.26

Make all appropriate tests (α=0.05)

$H_{0\,AB}: \gamma_{ij} = 0 \; for \; all \; i, j$ (there is **no** interaction between factor **A** at level i and factor **B** at level j) $H_{1\,AB}: at\ least\ one\ \gamma_{ij} \neq 0$

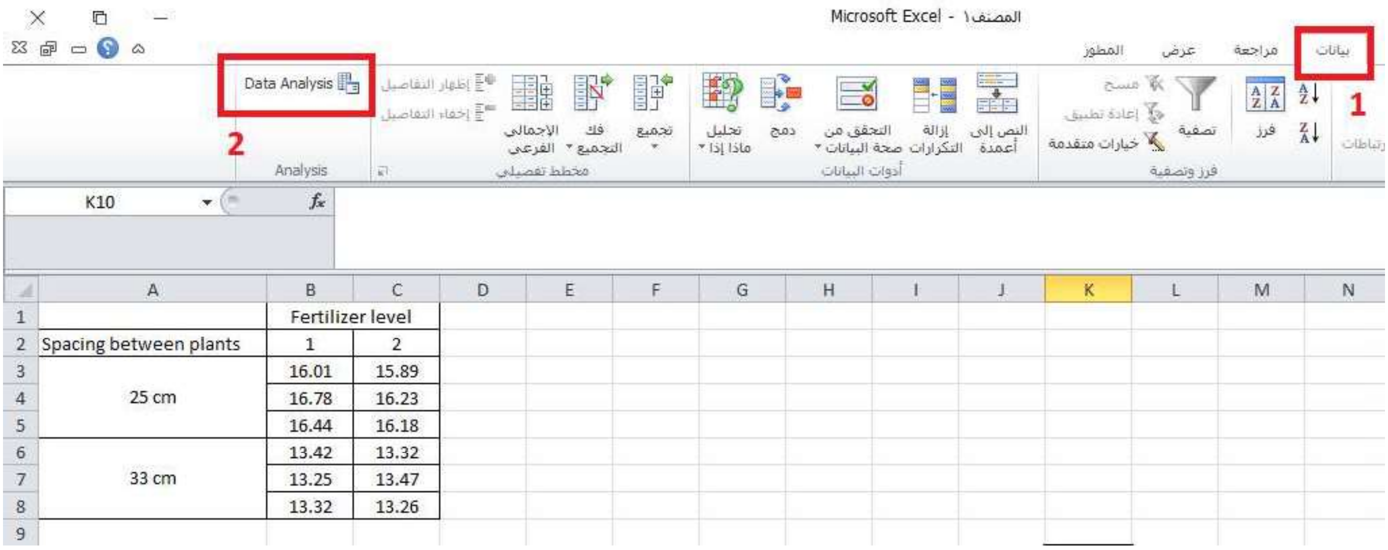
P-value= 0.2387 > 0.05 , we accept H_0, which means there is no interaction between fertilizer and spacing between plants.

$H_{0\,A}: \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad v.s \quad H_{1\,A}: at\ least\ one\ \alpha_i \neq 0$ (α_i the effect of fertilizer at level i)

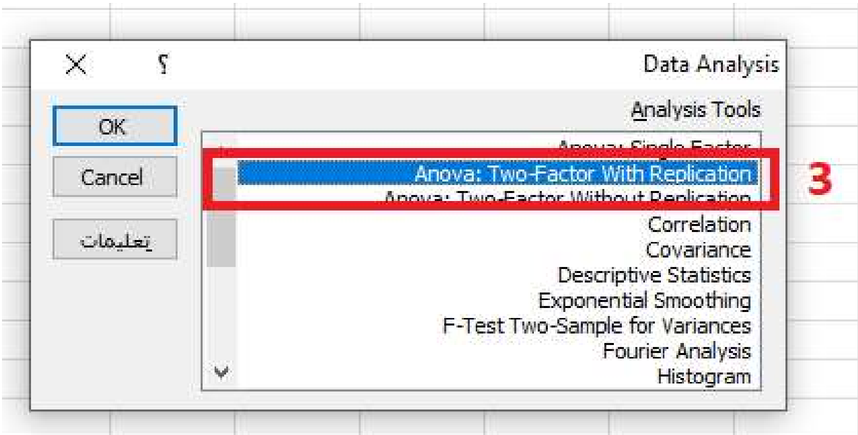
P-value= 0.2957 > 0.05 , we accept H_0, which means there is no effect of fertilizers on potato yield.

$H_{0\,B}: \beta_1 = \beta_2 = \beta_3 = 0 \quad v.s \quad H_{1\,B}: at\ least\ one\ \beta_j \neq 0$ (β_j the effect of spacing between plants at level j)

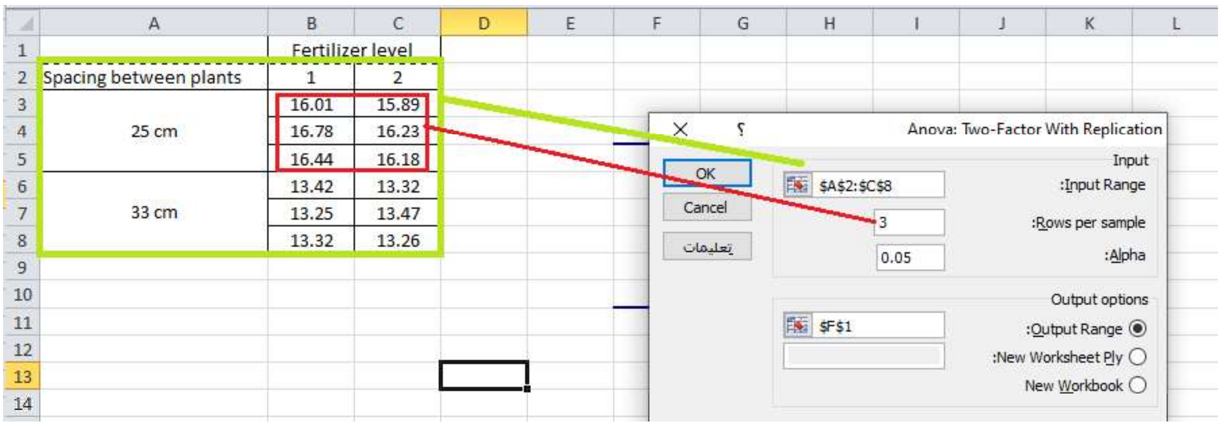
P-value= 0.000 ≤ 0.05 , we cannot accept H_0, which means that there is an effect of distance between plants on potato yield.



Microsoft Excel - المصنف١
بيانات
مراجعة
عرض
المطور
1
Data Analysis
2
Analysis
K10
Fertilizer level
Spacing between plants
1
2
25 cm
16.01
15.89
16.78
16.23
16.44
16.18
33 cm
13.42
13.32
13.25
13.47
13.32
13.26



Data Analysis
Analysis Tools
OK
Cancel
تعليمات
Anova: Two-Factor With Replication
3
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram



Fertilizer level
Spacing between plants
1
2
25 cm
16.01
15.89
16.78
16.23
16.44
16.18
33 cm
13.42
13.32
13.25
13.47
13.32
13.26
Anova: Two-Factor With Replication
Input
:Input Range
$A$2:$C$8
:Rows per sample
3
:Alpha
0.05
OK
Cancel
تعليمات
Output options
:Output Range
$F$1
:New Worksheet Ply
New Workbook

	A	B	C	D	E	F	G	H	I	J
1		Fertilizer level				Anova: Two-Factor With Replication				
2	Spacing between plants	1	2							
3	25 cm	16.01	15.89			SUMMARY	1	2	Total	
4		16.78	16.23			*25 cm*				
5		16.44	16.18			Count	3	3	6	
6	33 cm	13.42	13.32			Sum	49.23	48.3	97.53	
7		13.25	13.47			Average	16.41	16.1	16.255	
8		13.32	13.26			Variance	0.1489	0.0337	0.10187	
9										
10						*33 cm*				
11						Count	3	3	6	
12						Sum	39.99	40.05	80.04	
13						Average	13.33	13.35	13.34	
14						Variance	0.0073	0.0117	0.00772	
15										
16						*Total*				
17						Count	6	6		
18						Sum	89.22	88.35		
19						Average	14.87	14.725		
20						Variance	2.9084	2.28691		
21										

ANOVA

Source of Variati	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	25.49168	1	25.49168	505.7872	1.62E-08	5.317655
Columns	0.063075	1	0.063075	1.251488	0.295732	5.317655
Interaction	0.081675	1	0.081675	1.620536	0.238762	5.317655
Within	0.4032	8	0.0504			
Total	26.03963	11				

Two Factor ANOVA without Replication:

2) Suppose that interest is in 5 growth regulators. Baladi orange trees were randomly sprayed with one of the growth regulators, at harvest, 3 orange from each treatment were randomly assigned to a storage temperature. After a period of storage, the percent weight loss was measured.

Temperature	Growth regulator				
	1	2	3	4	5
5^0C	9.2	11.3	9.1	10.4	12.3
10^0C	18.2	17.6	18.4	16.5	17.8
25^0C	21.3	24.4	24.8	21.2	24.1

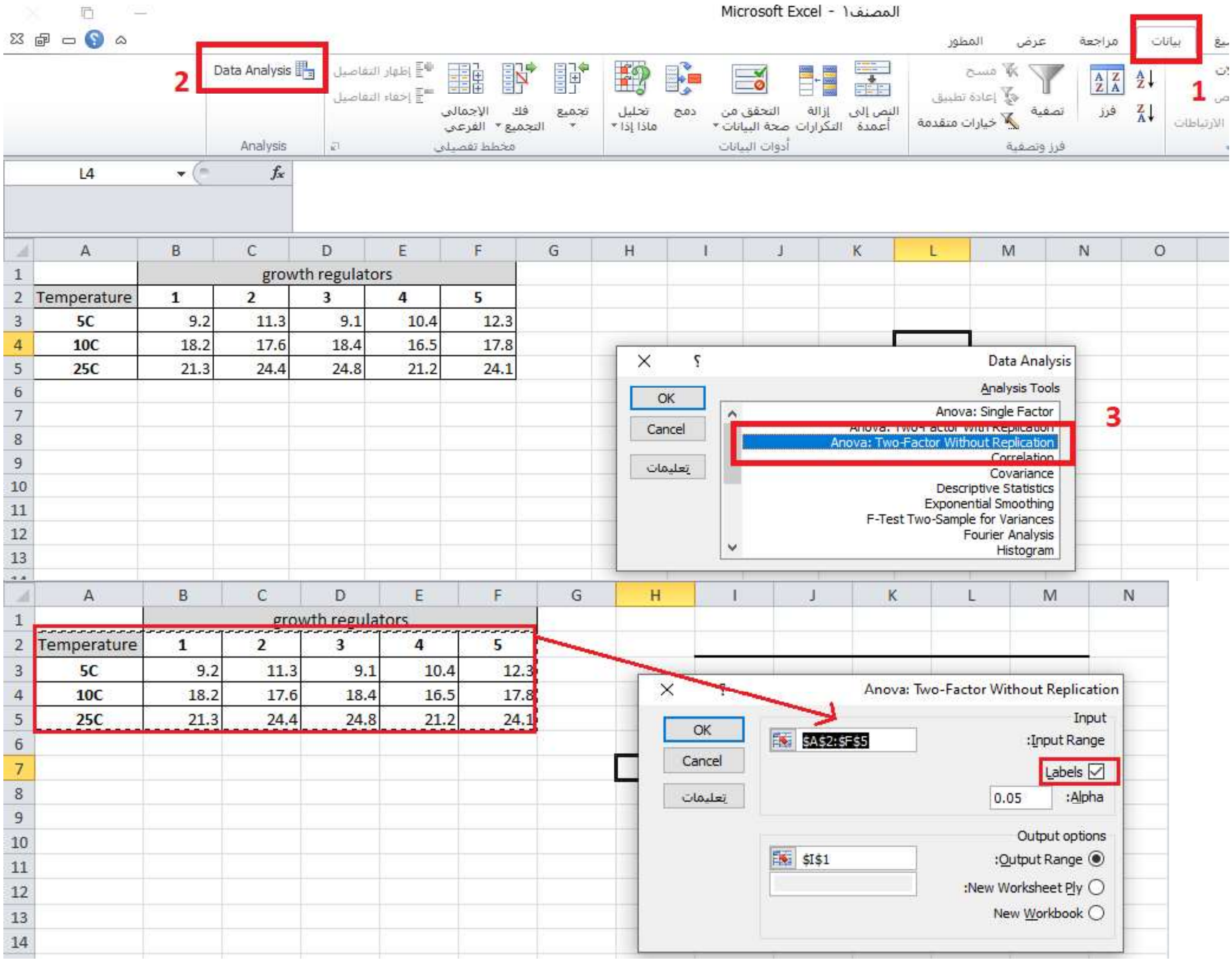
Assuming no interaction, test if there is a difference in the effects of the five growth regulators on the percent weight loss of oranges. Also test if there a difference in the effects of the three storage temperature (α=0.05)

H_{0A}: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ $v.s$ H_{1A}: $at\ least\ one\ \alpha_i \neq 0$ (α_i the effect of temperature at level **i**)

P-value= 0.000 < 0.05 , we reject H_0, which means that there is an effect of temperature on the percentage of weight loss in oranges.

H_{0B}: $\beta_1 = \beta_2 = \cdots = \beta_5 = 0$ $v.s$ H_{1B}: $at\ least\ one\ \beta_j \neq 0$ (β_j the effect of growth regulators at level **j**)

P-value= 0.2447 > 0.05 , we cannot reject H_0, which means that there is no effect of growth regulators on the percentage of weight loss in oranges.


Microsoft Excel - المصنف١
بيانات
مراجعة
عرض
المطور
1
2
Data Analysis
Analysis
فرز وتصفية
أدوات البيانات
مخطط تفصيلي
L4
growth regulators
Temperature
1
2
3
4
5
5C
9.2
11.3
9.1
10.4
12.3
10C
18.2
17.6
18.4
16.5
17.8
25C
21.3
24.4
24.8
21.2
24.1
Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
تعليمات
3
Anova: Two-Factor Without Replication
Input
Input Range:
$A$2:$F$5
Labels
Alpha:
0.05
Output options
Output Range:
$I$1
New Worksheet Ply:
New Workbook

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
5C	5	52.3	10.46	1.883
10C	5	88.5	17.7	0.55
25C	5	115.8	23.16	3.103
1	3	48.7	16.23333	39.50333
2	3	53.3	17.76667	42.92333
3	3	52.3	17.43333	62.32333
4	3	48.1	16.03333	29.32333
5	3	54.2	18.06667	34.86333

ANOVA

Source of Variat.	SS	df	MS	F	P-value	F crit
Rows (Temperature)	405.8653	2	202.9327	135.1983	6.82E-07	4.45897
Columns (growth regulators)	10.136	4	2.534	1.688208	0.244786	3.837853
Error	12.008	8	1.501			
Total	428.0093	14				

2-Two-Sample t Statistic :

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

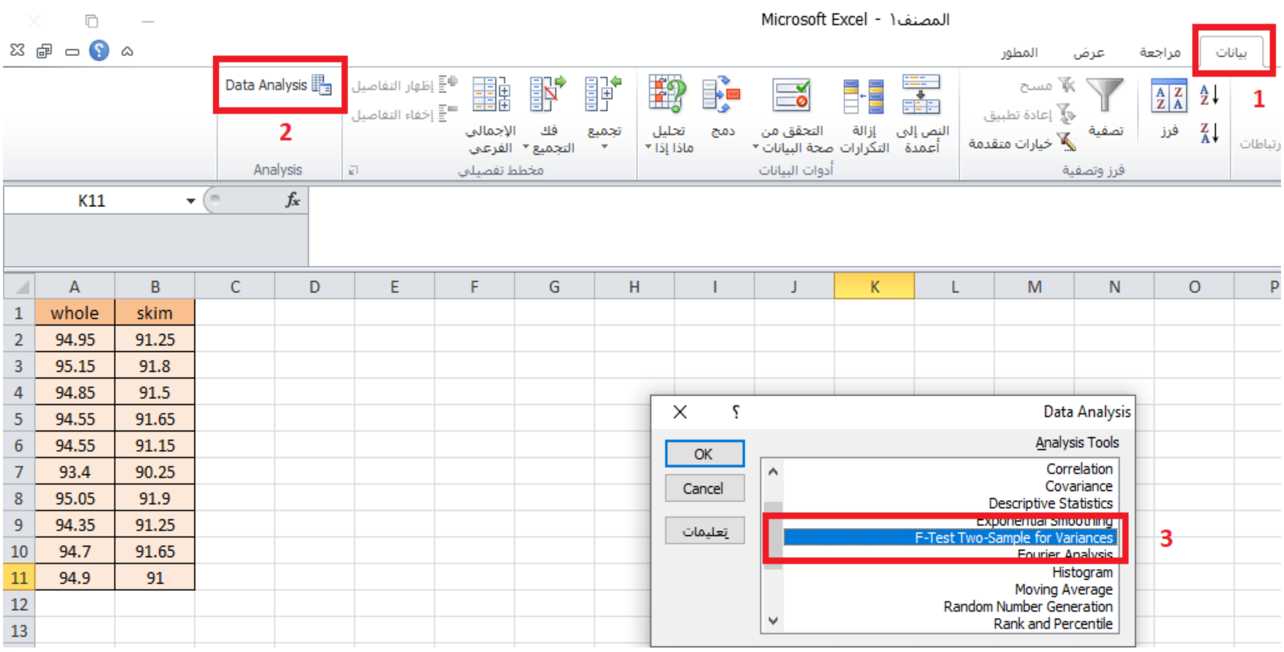
Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

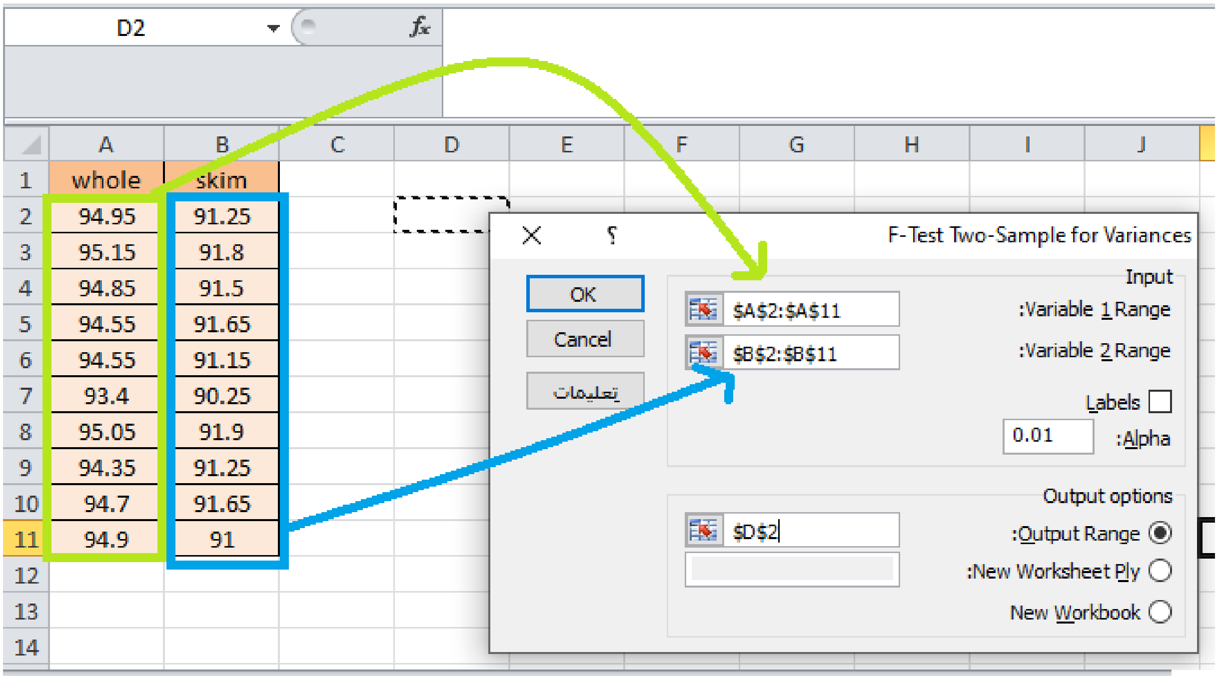
Assuming normal populations .

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





M17

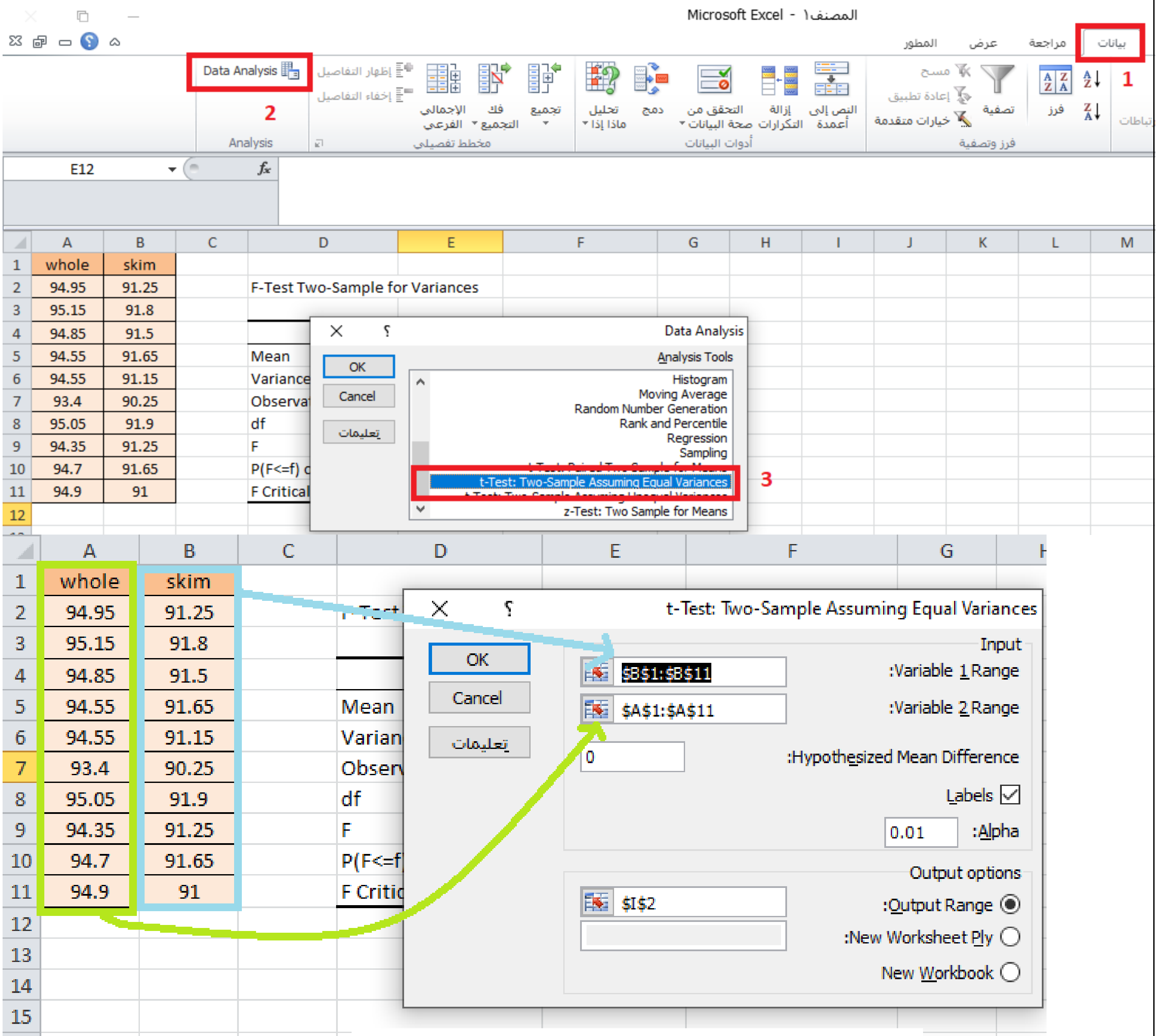
	A	B	C	D	E	F
1	whole	skim				
2	94.95	91.25		F-Test Two-Sample for Variances		
3	95.15	91.8				
4	94.85	91.5			*Variable 1*	*Variable 2*
5	94.55	91.65		Mean	94.645	91.34
6	94.55	91.15		Variance	0.253027778	0.233222222
7	93.4	90.25		Observations	10	10
8	95.05	91.9		df	9	9
9	94.35	91.25		F	1.084921391	
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 ≯ 5.351 . Therefore, we fail to reject the null hypothesis.
The variances of the two populations are equal .

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

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



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

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:

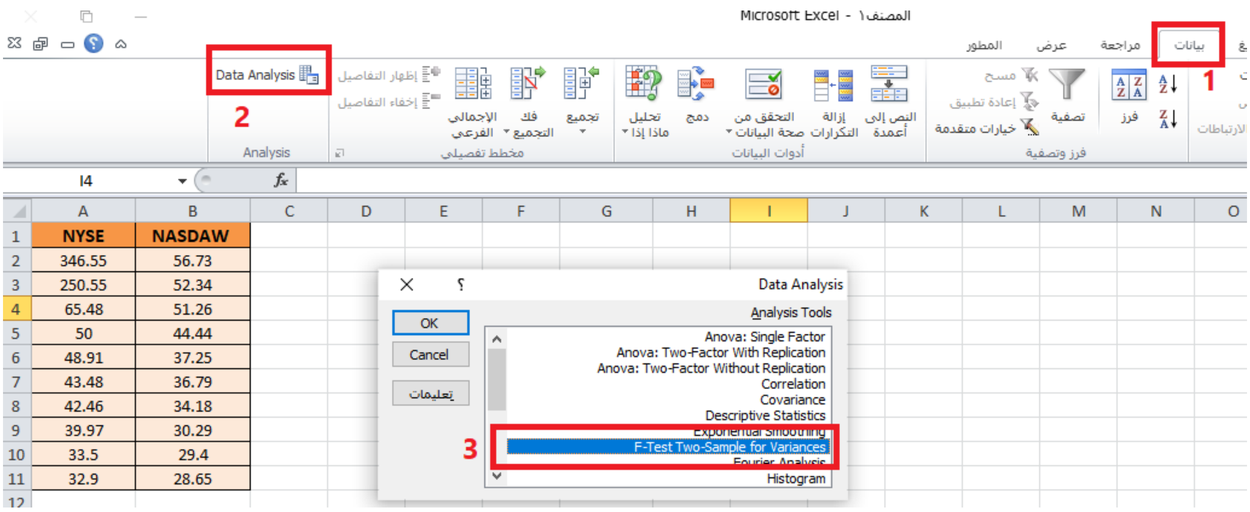
We do a one-tailed test . if t Stat < -t Critical one-tail, we reject the null hypothesis. As -14.9879 < -2.55238 (p-value=0.00000653<α=0.01) . Therefore, we reject the null hypothesis

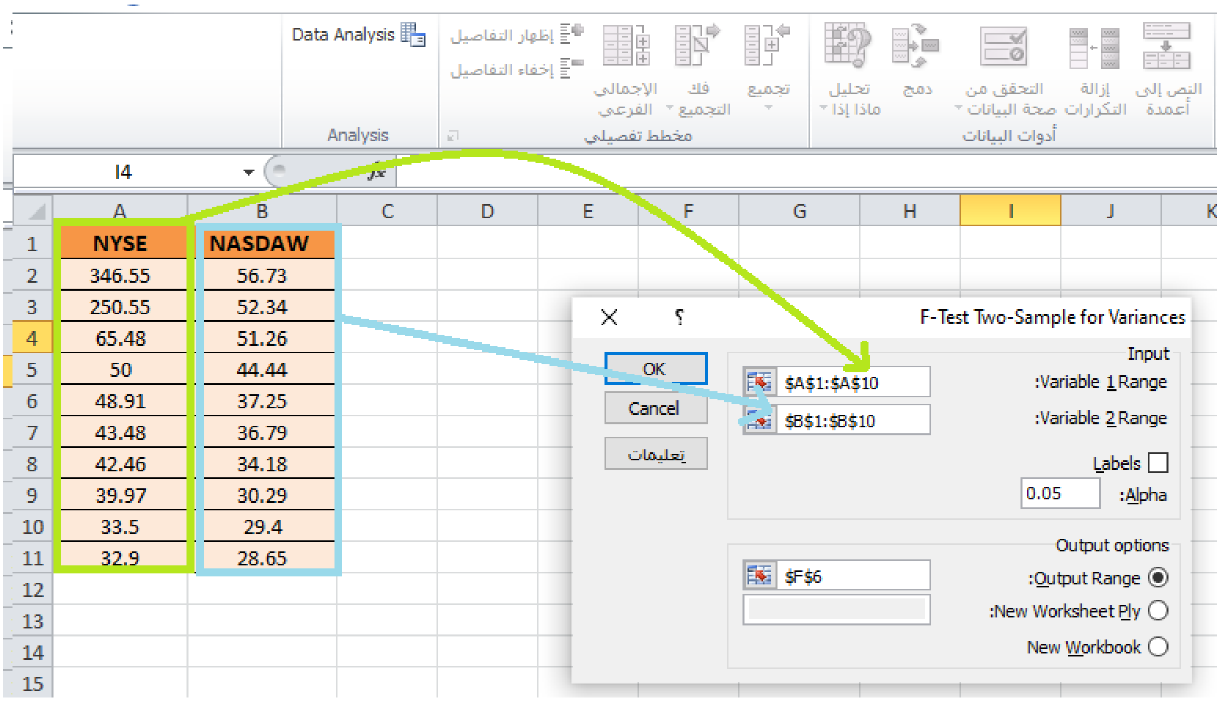
Q : The example below gives the Dividend Yields for the top ten NYSE and NASDAW stocks. Use the t-test tool to determine whether there is any indication of a difference between the means of the two different populations. α =0.05

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

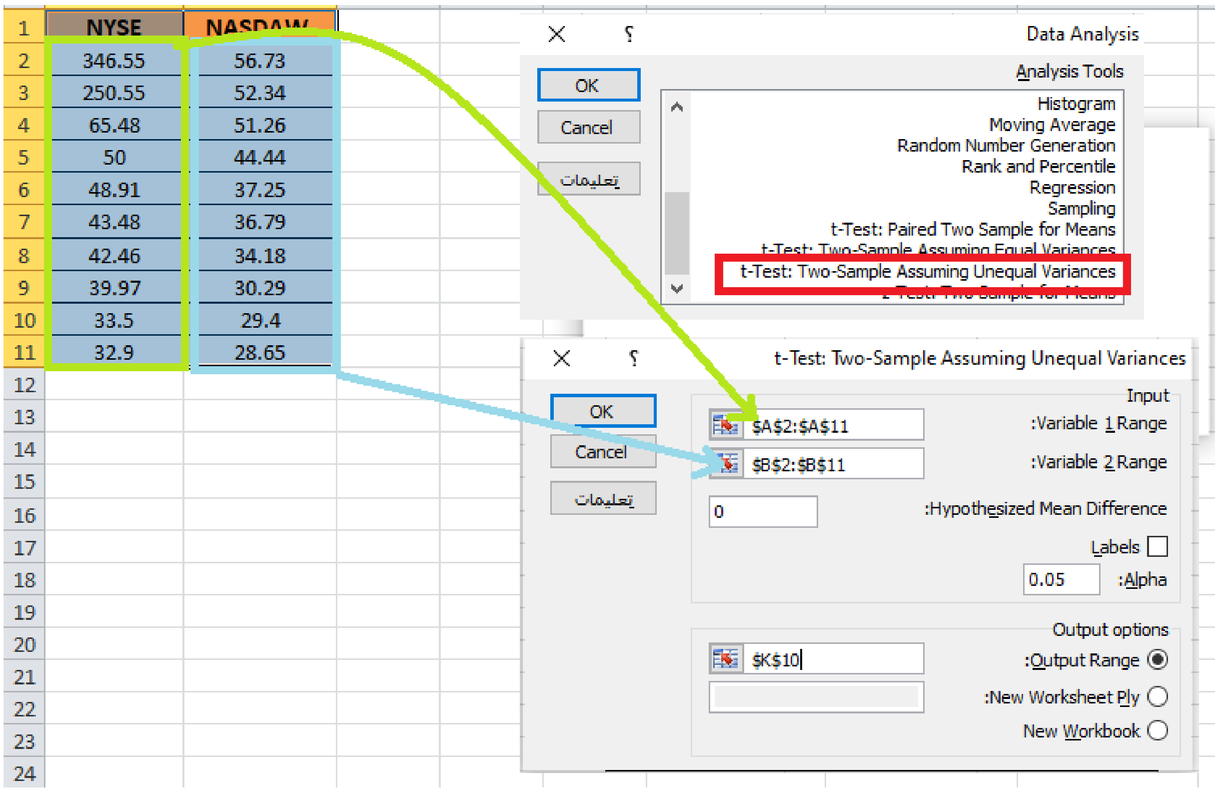
	Variable 1	Variable 2
Mean	95.38	40.133
Variance	12063.80027	107.2998233
Observations	10	10
df	9	9
F	112.4307561	
P(F<=f) one-tail	3.6621E-08	
F Critical one-tail	3.178893104	

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 :

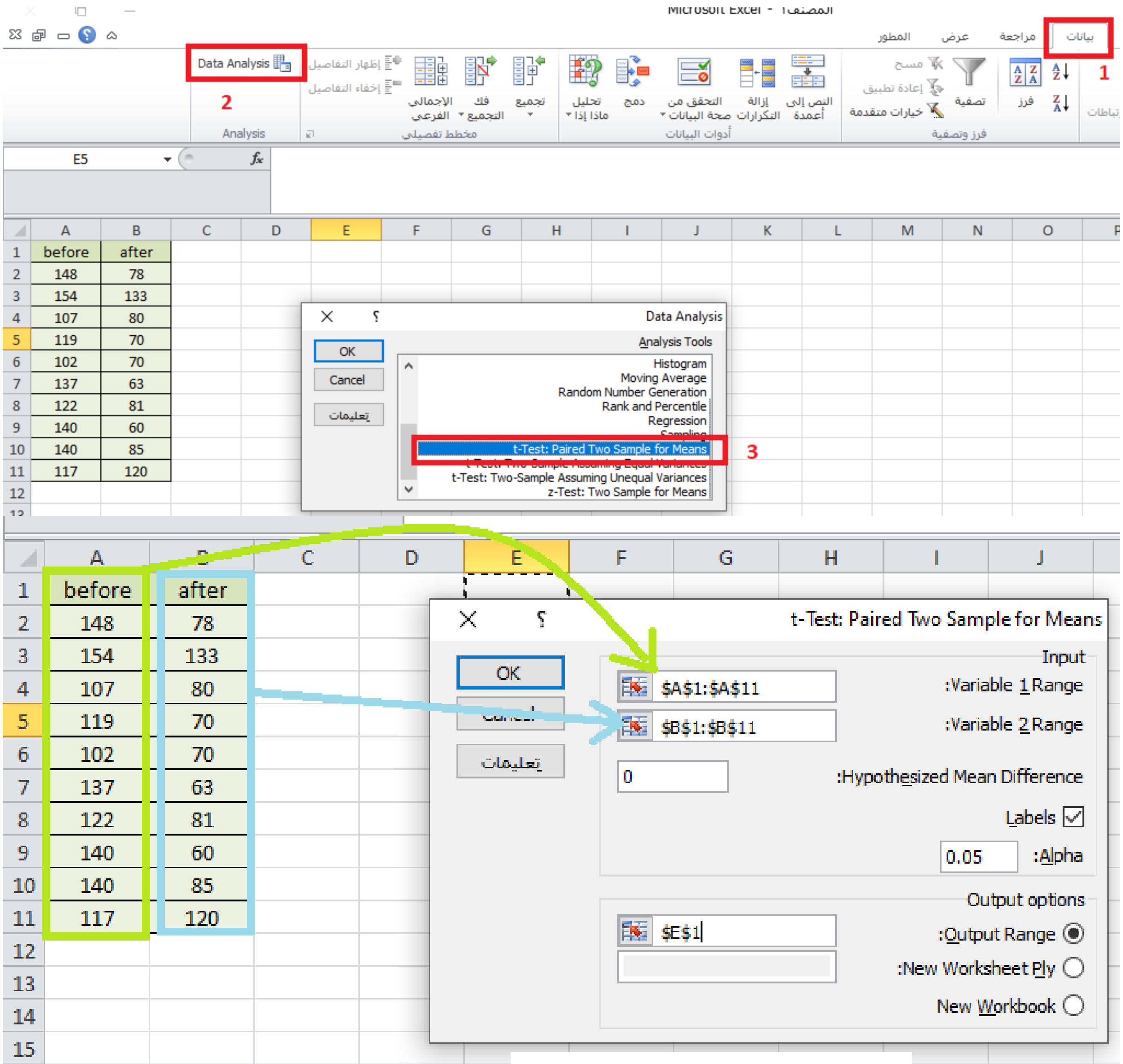
Q : In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

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



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

	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) = $\pm$ **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.

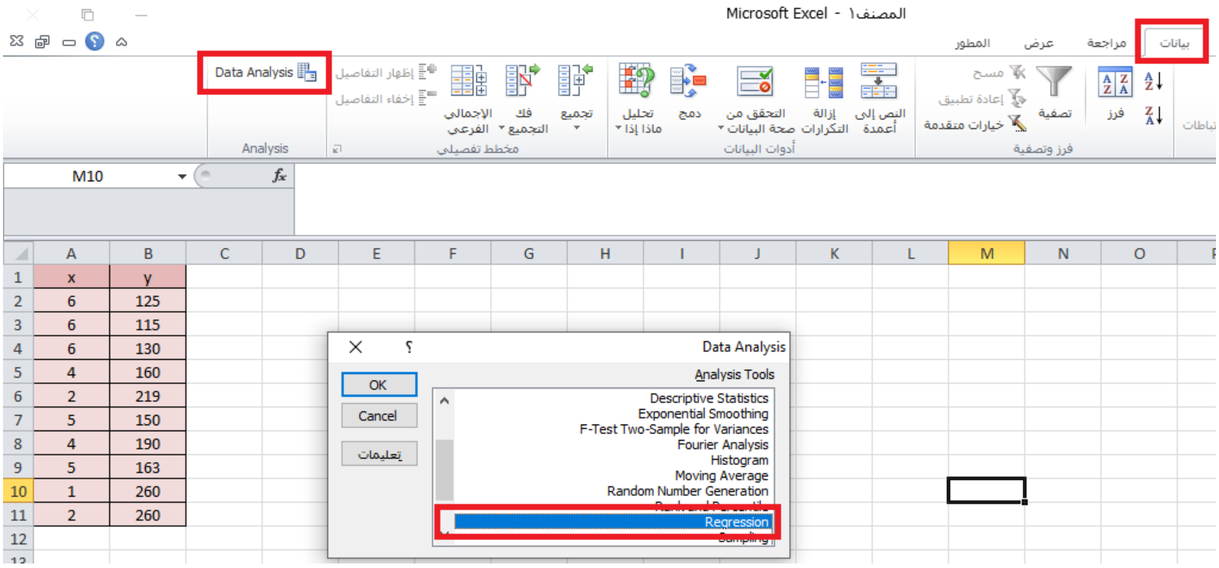
X	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

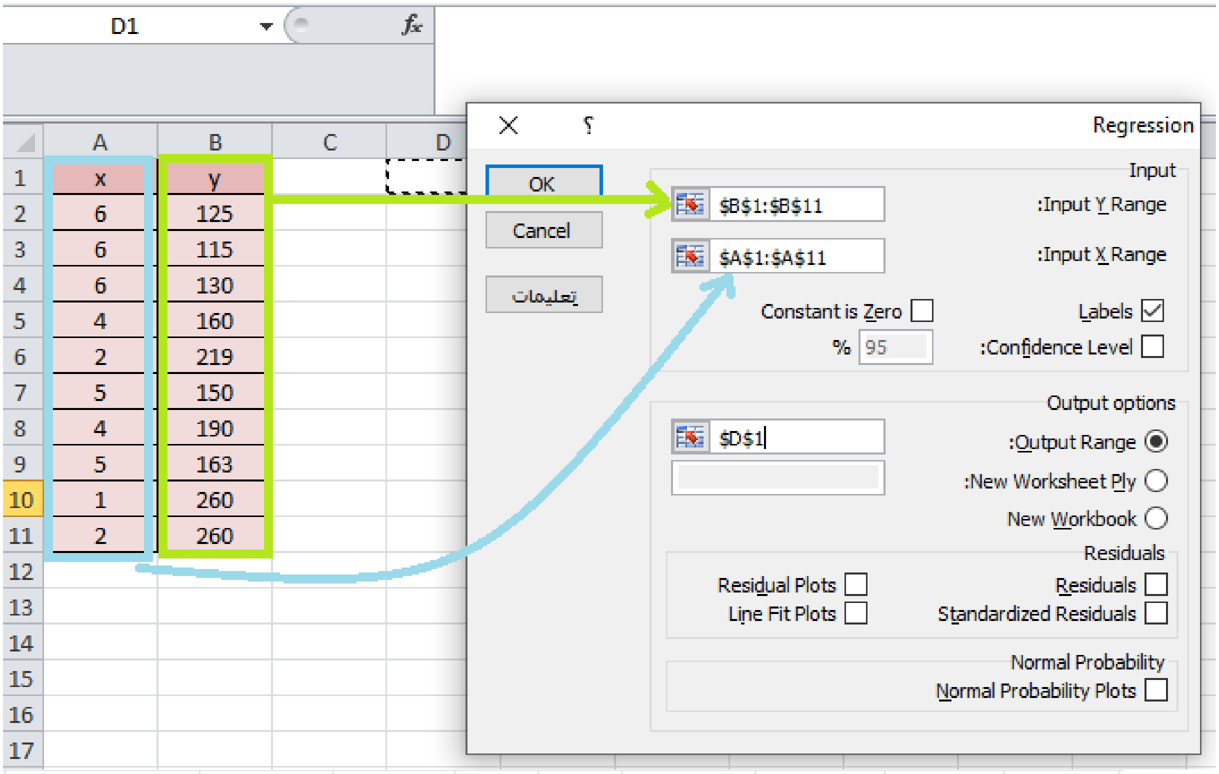
a) Determine the regression equation for the data.

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) $\mathbf{r^2 = 0.9367}$
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).

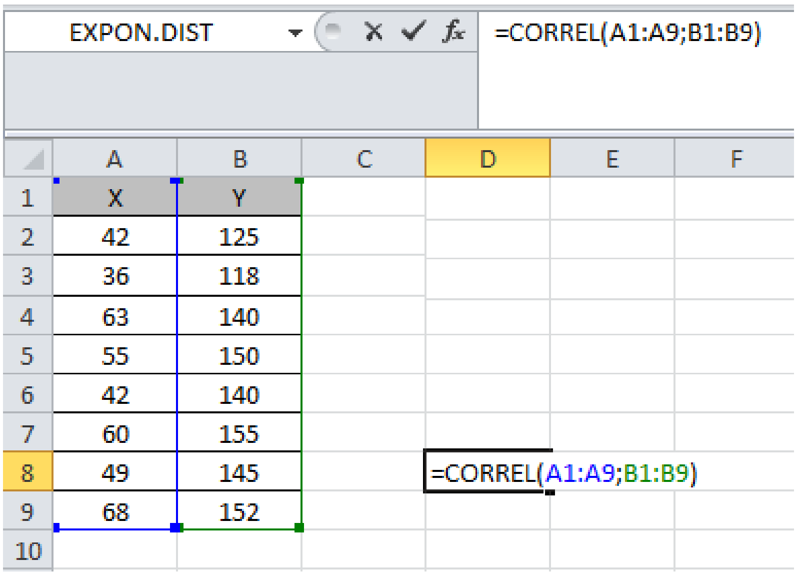
Correlation :

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)

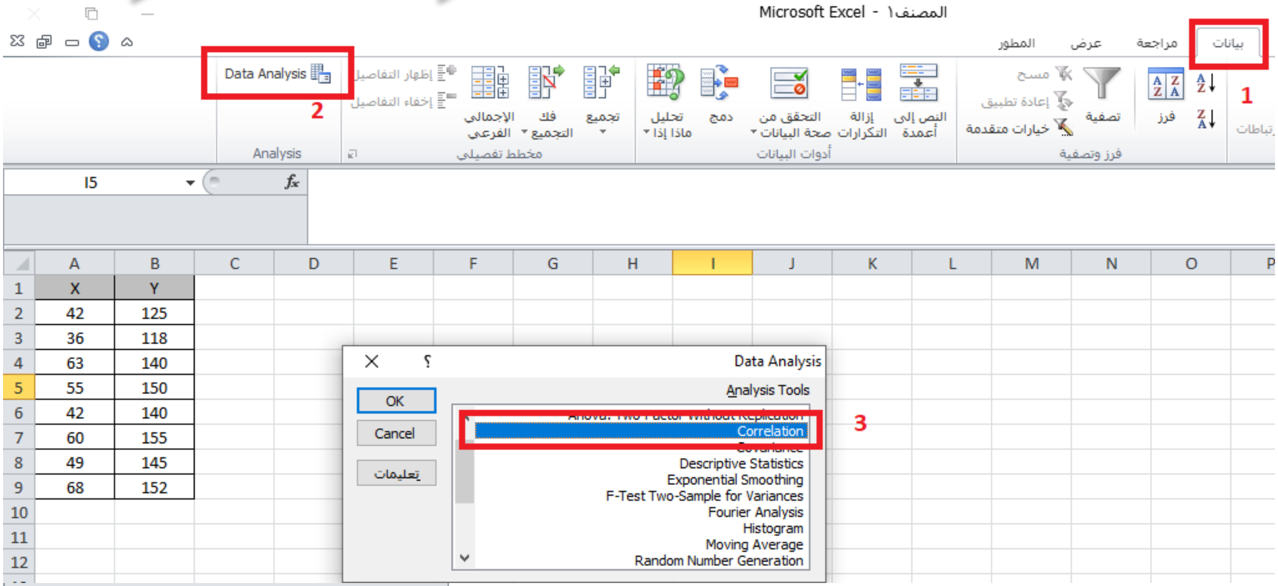


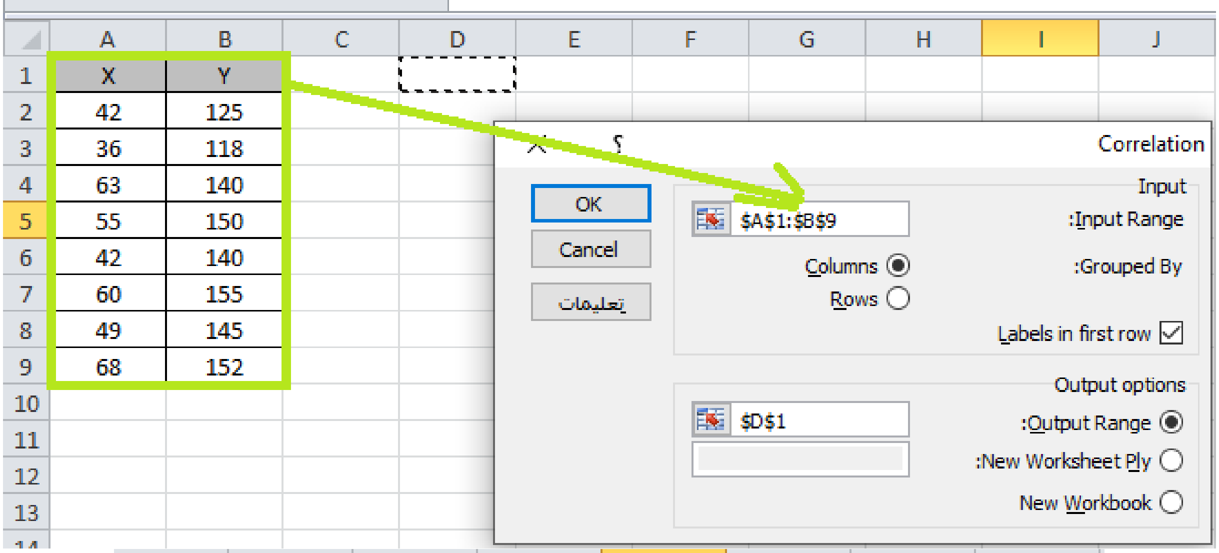
EXPON.DIST | =CORREL(A1:A9;B1:B9)

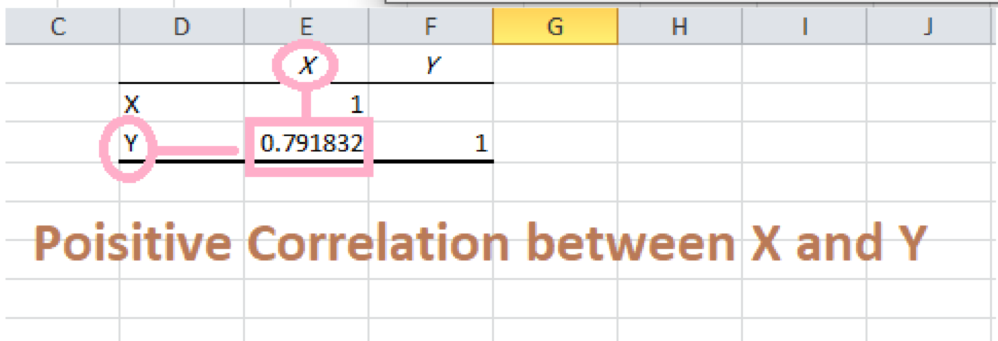
	A	B	C	D	E	F
1	X	Y				
2	42	125				
3	36	118				
4	63	140				
5	55	150				
6	42	140				
7	60	155				
8	49	145		=CORREL(A1:A9;B1:B9)		
9	68	152				
10						

OR:

Data → Data Analysis → Correlation

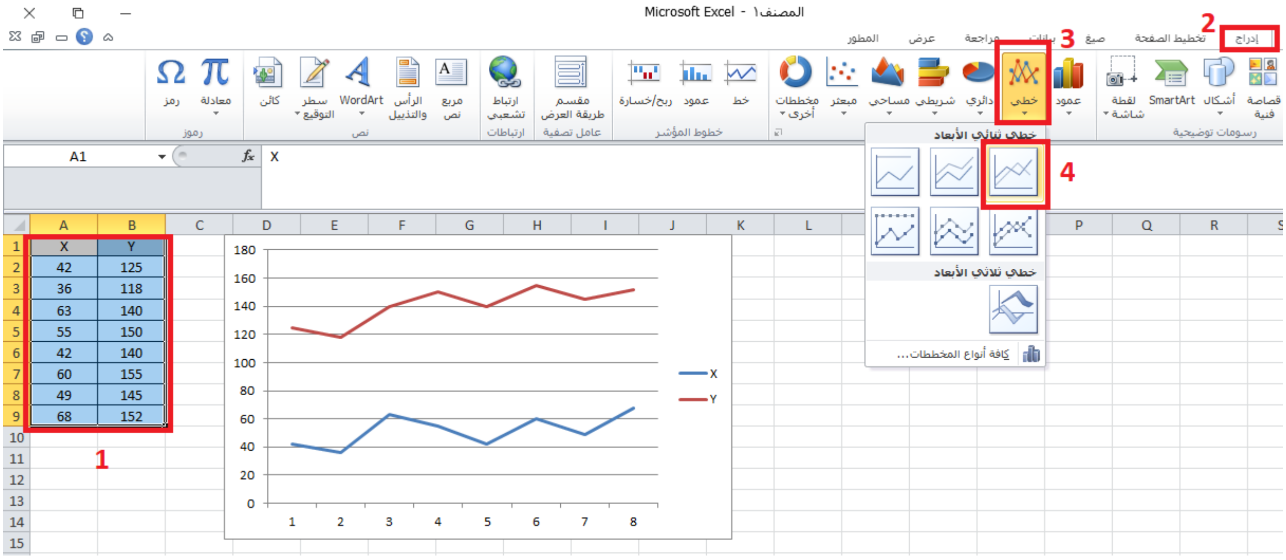






Poisitive Correlation between X and Y

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

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

Determinant

G=	3	-1
	-5	2

det(G)= =MDETERM(N12:O13

MDETERM(**array**)

det(G)= 1

Inverse Matrices

G=	3	-1
	-5	2

G^{-1}= =MINVERSE(N12:O13

MINVERSE(**array**)

Ctel + Shift + Enter

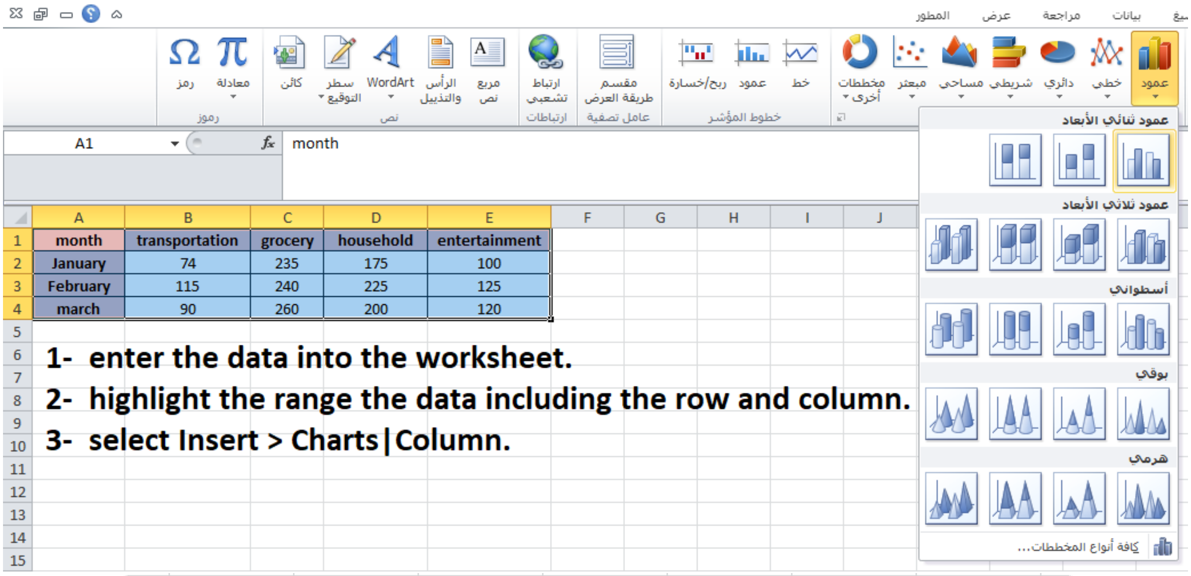
G^{-1}=	2	1
	5	3

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

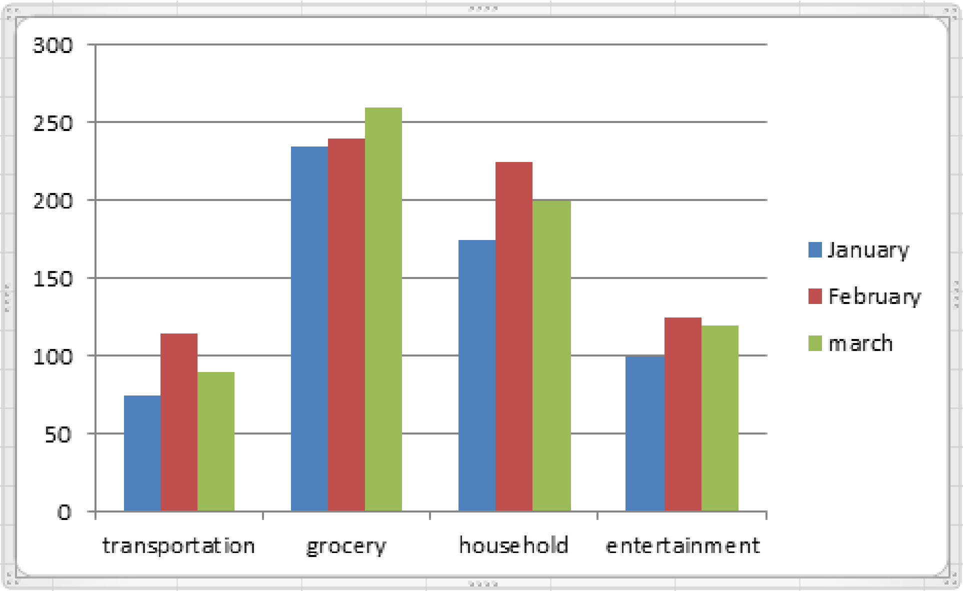
a- Bar chart :



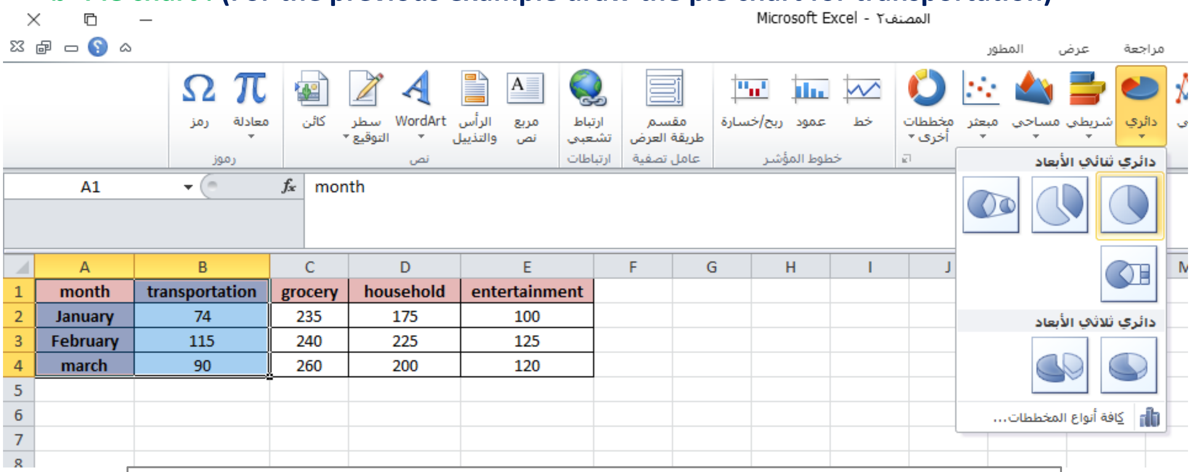
1- enter the data into the worksheet.

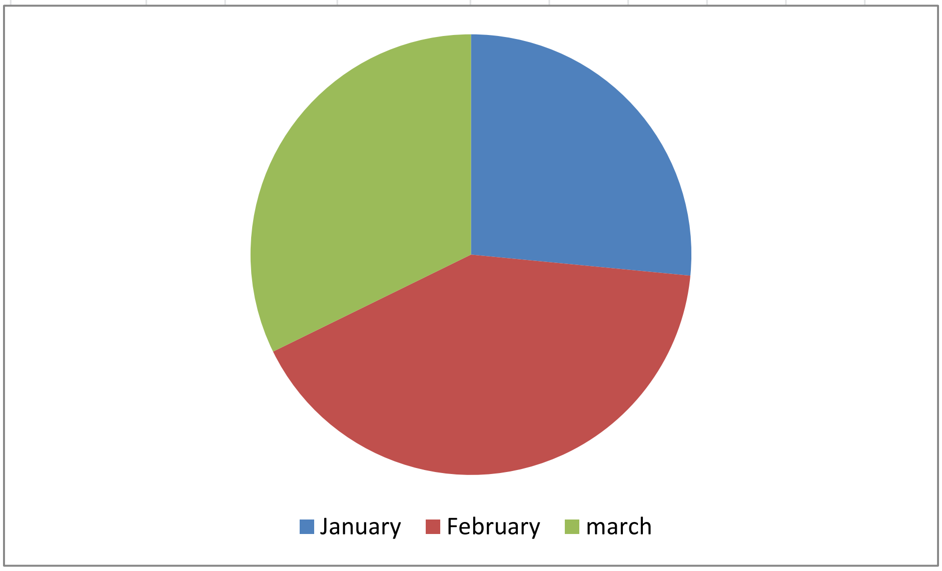
2- highlight the range the data including the row and column.

3- select Insert > Charts|Column.

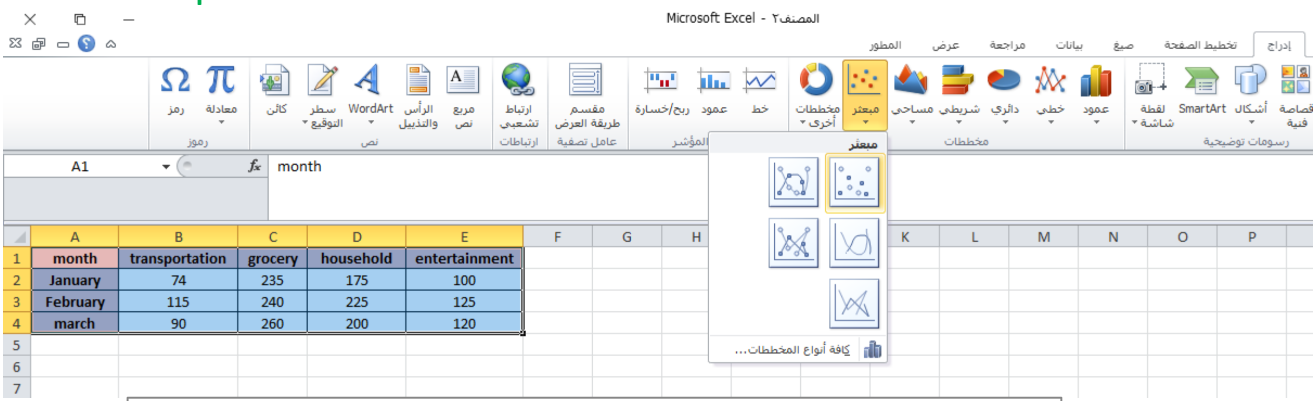


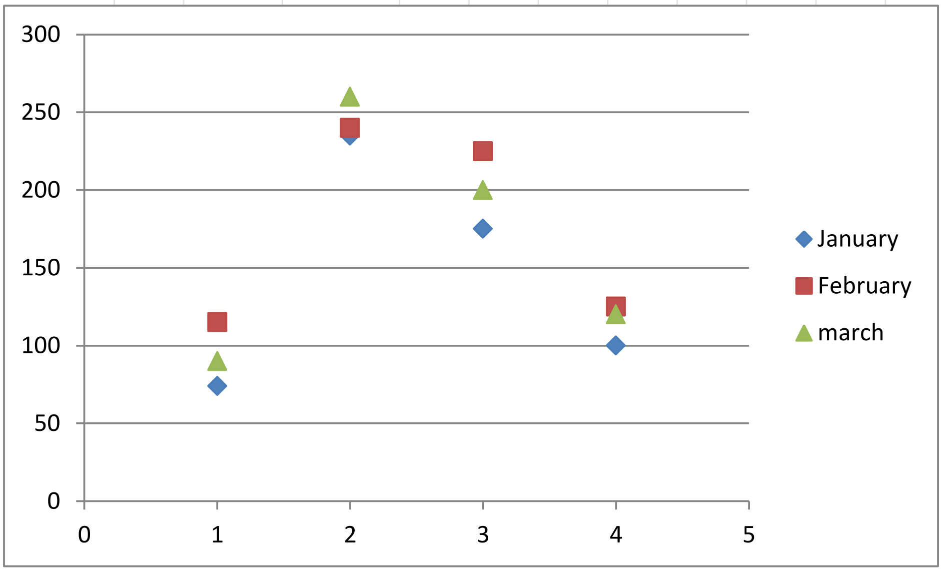
b- Pie chart : (For the previous example draw the pie chart for transportation)





c- Scatter plot :





d- Histogram :

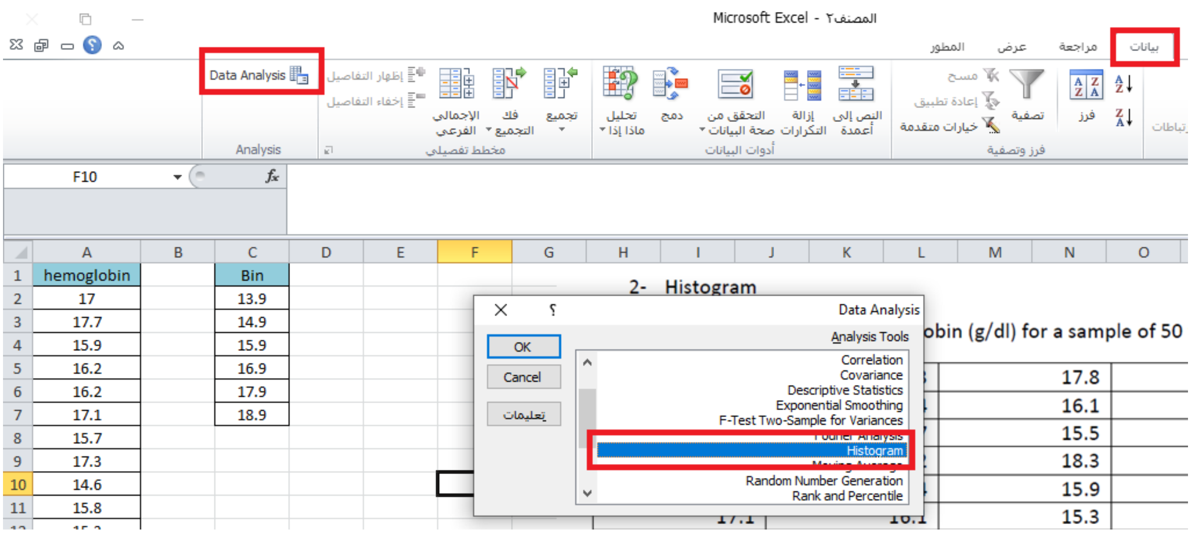
The following data represent hemoglobin (g/dl) for a sample of 50 women :

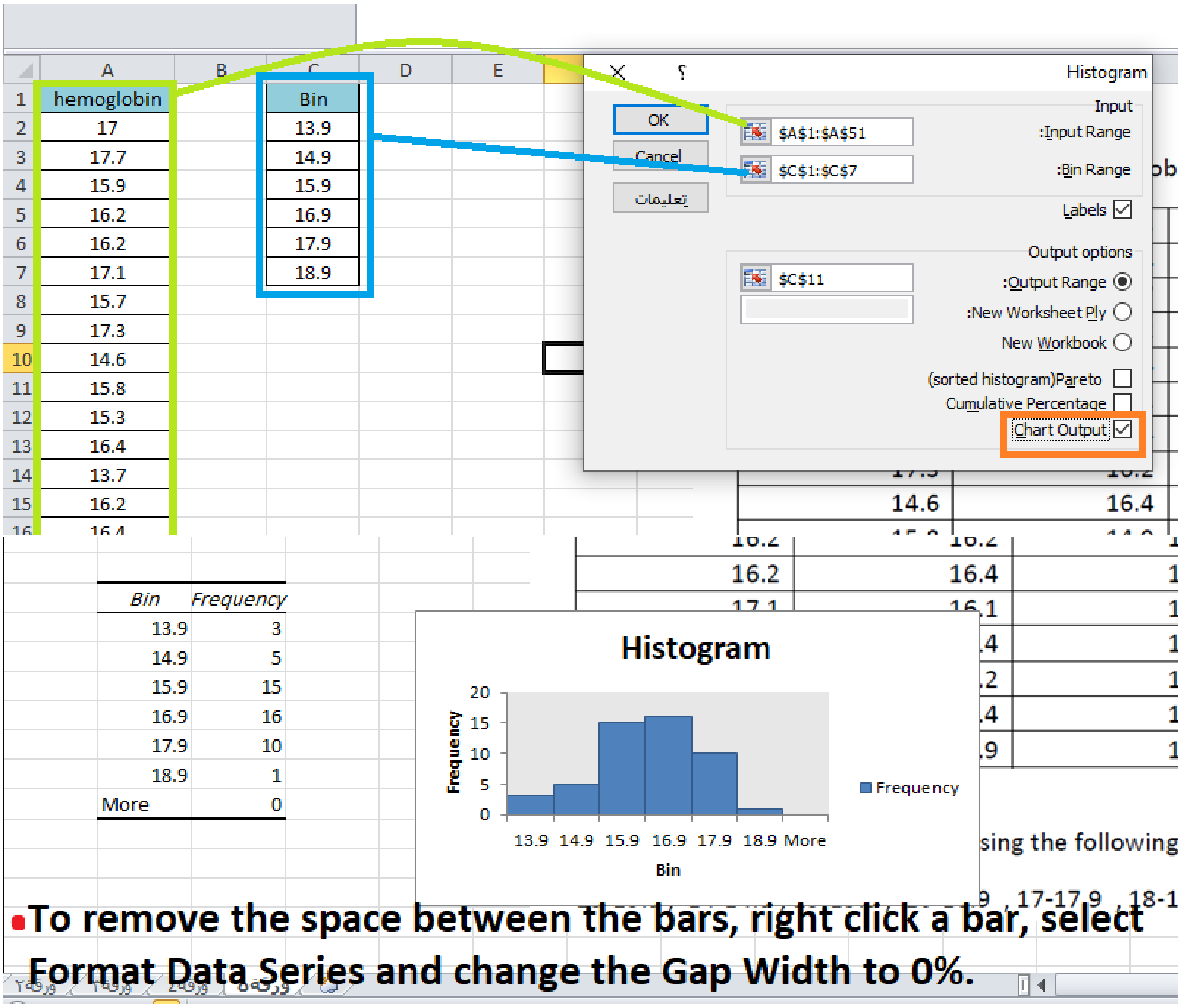
17	15.3	17.8	17.4	16.3
17.7	16.4	16.1	15	15.9
15.9	13.7	15.5	14.2	16.7
16.2	16.2	18.3	16.1	15.1
16.2	16.4	15.9	15.7	15.8
17.1	16.1	15.3	15.1	13.5
15.7	14	13.9	17.4	17
17.3	16.2	16.8	16.5	15.8
14.6	16.4	15.9	14.4	17.5
15.8	14.9	16.3	16.3	17.3

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

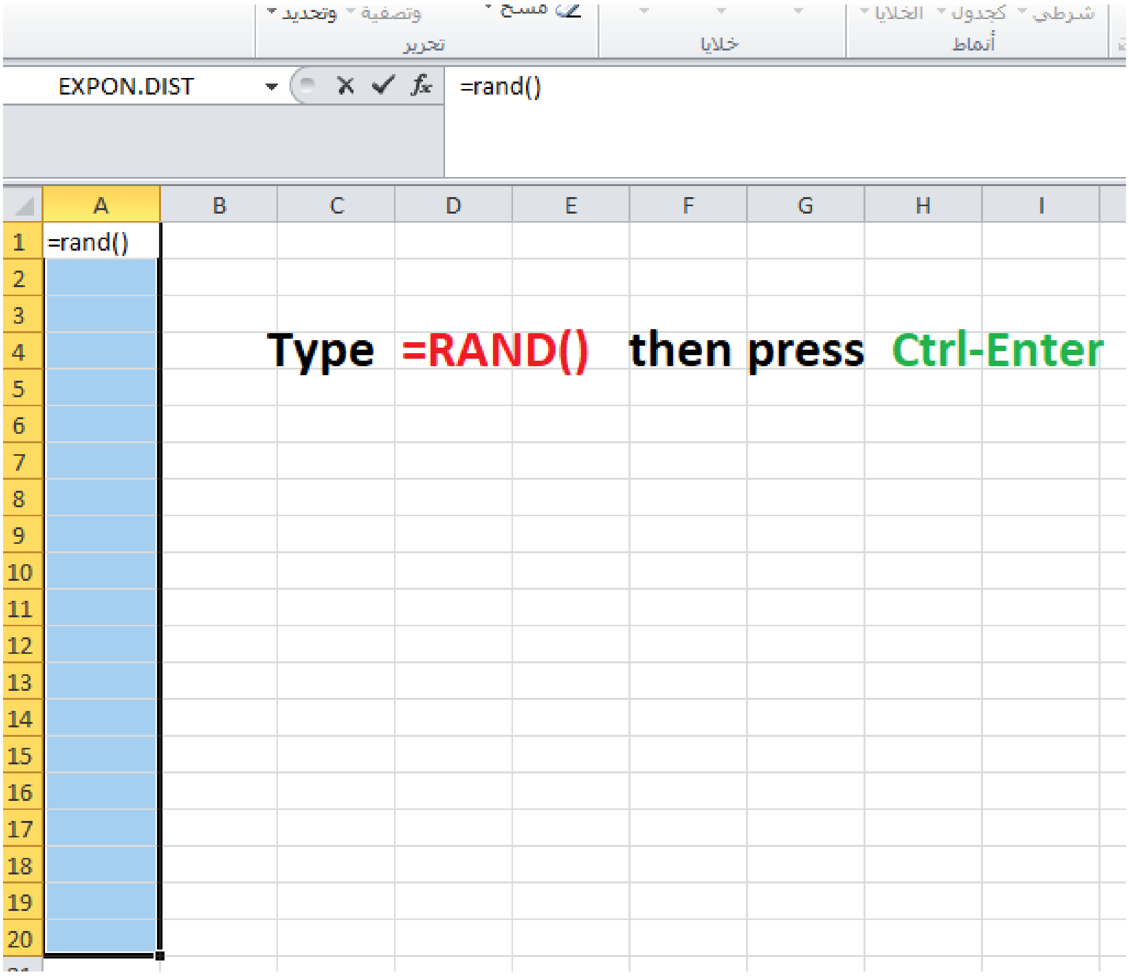
Data → Data Analysis → Histogram




hemoglobin
17
17.7
15.9
16.2
16.2
17.1
15.7
17.3
14.6
15.8
15.3
16.4
13.7
16.2
Bin
13.9
14.9
15.9
16.9
17.9
18.9
Histogram
Input
Input Range: $A$1:$A$51
Bin Range: $C$1:$C$7
Labels
OK
Cancel
تعليمات
Output options
Output Range: $C$11
New Worksheet Ply:
New Workbook
Pareto (sorted histogram)
Cumulative Percentage
Chart Output
Bin Frequency
13.9 3
14.9 5
15.9 15
16.9 16
17.9 10
18.9 1
More 0
Histogram
Frequency
Bin
13.9 14.9 15.9 16.9 17.9 18.9 More
20 15 10 5 0
To remove the space between the bars, right click a bar, select Format Data Series and change the Gap Width to 0%.

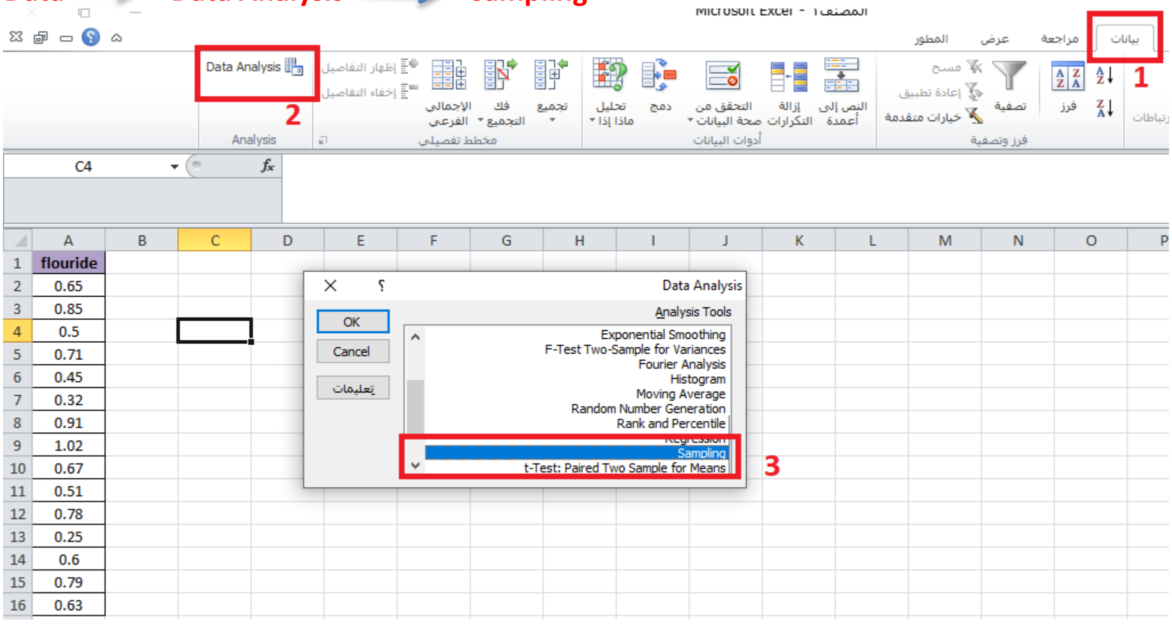
Generation Random samples :

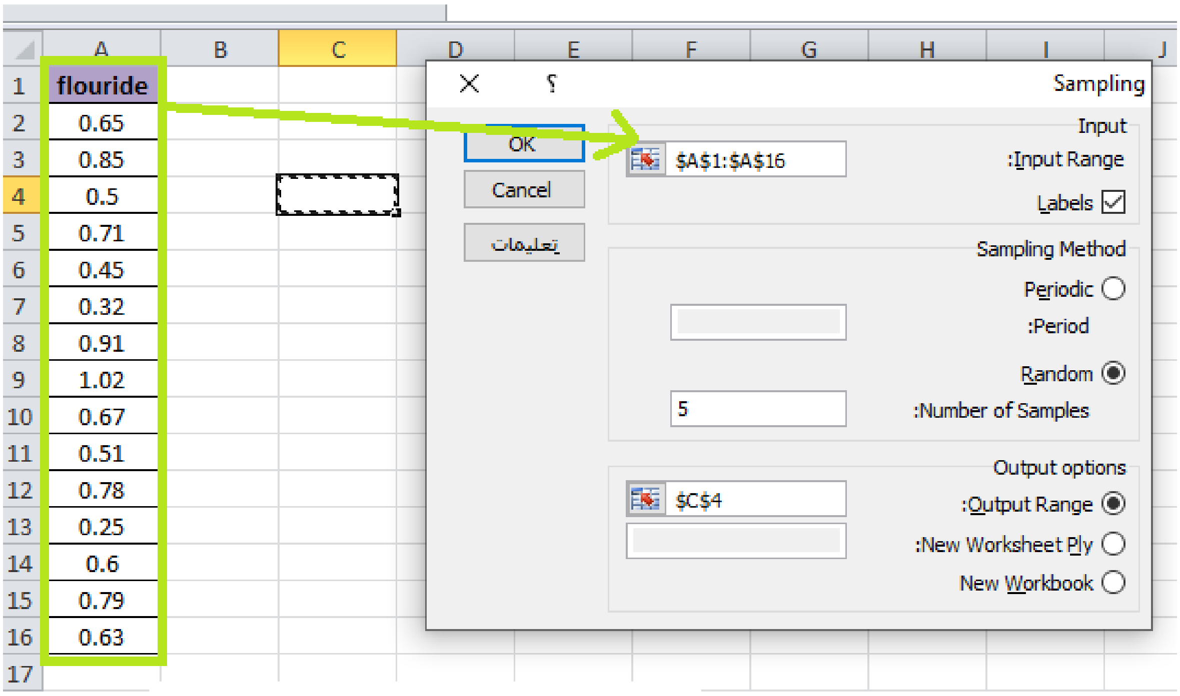
1- generate a random sample of size 20 between 0 and 1

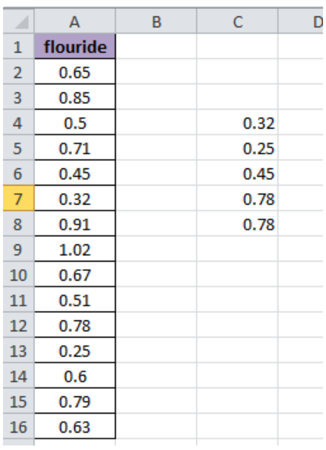


2- Sampling

Data → Data Analysis → sampling



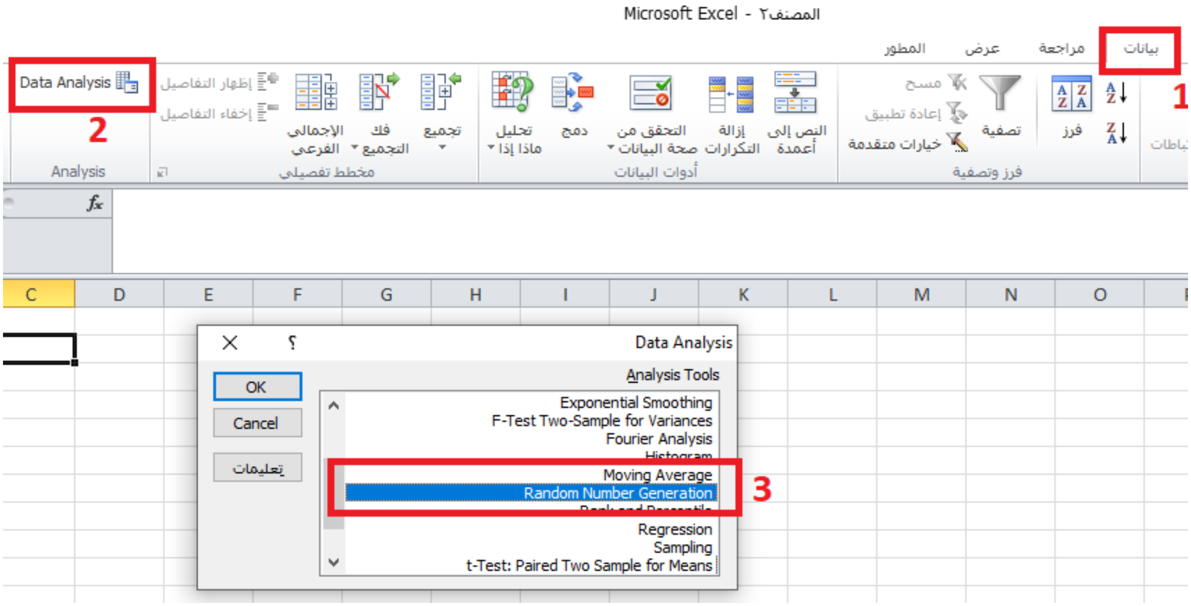




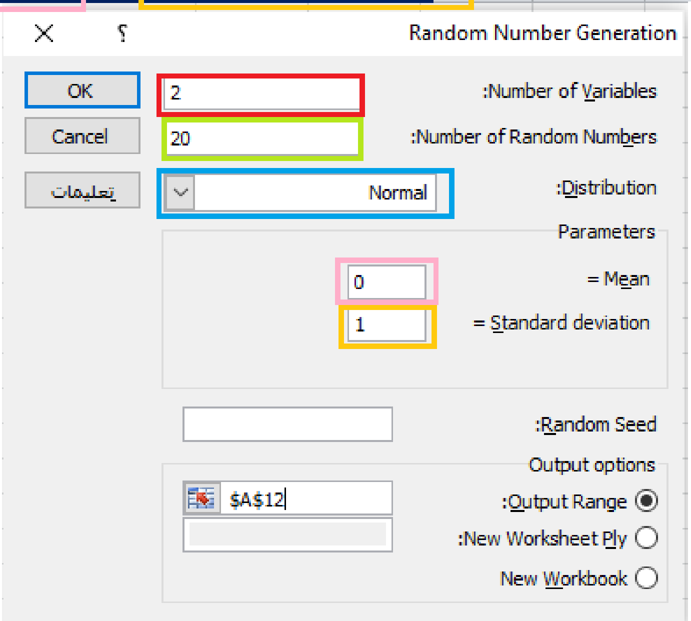
	A	B	C
1	flouride		
2	0.65		
3	0.85		
4	0.5		0.32
5	0.71		0.25
6	0.45		0.45
7	0.32		0.78
8	0.91		0.78
9	1.02		
10	0.67		
11	0.51		
12	0.78		
13	0.25		
14	0.6		
15	0.79		
16	0.63		

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.

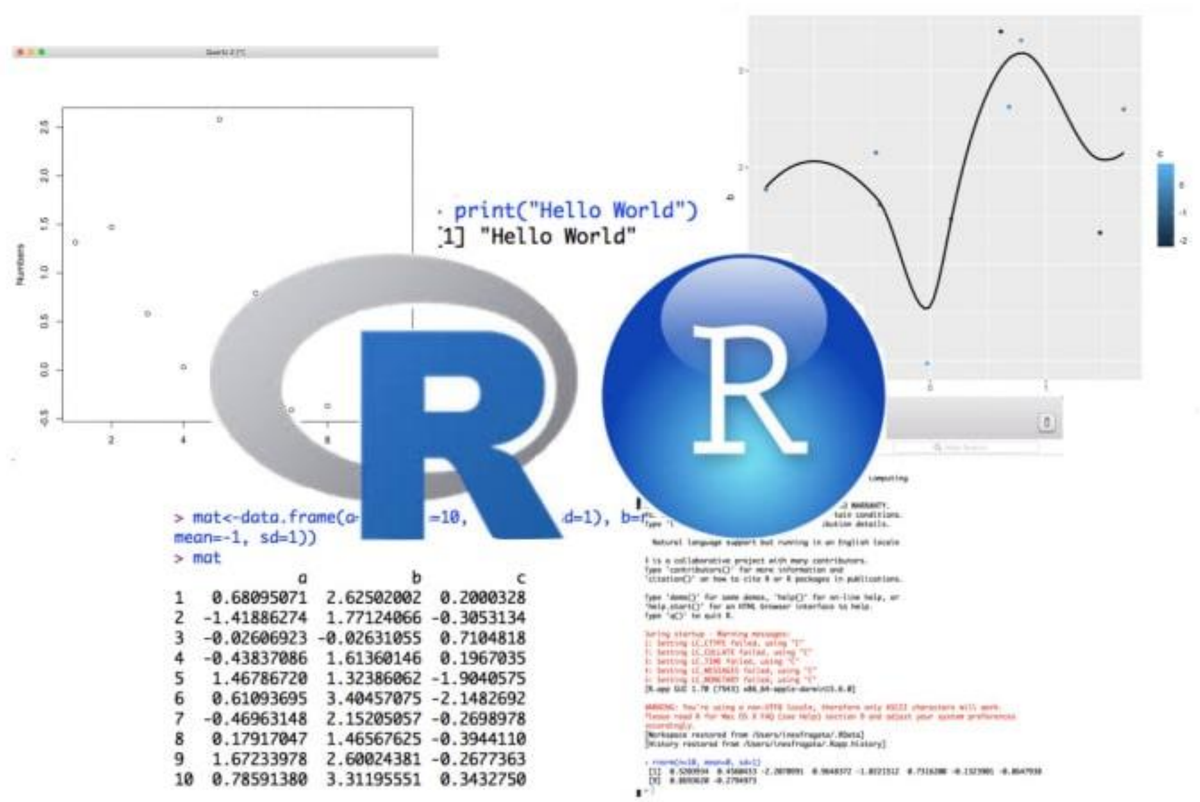
=RANDBETWEEN(bottom,top) returns a random number between the designated values.

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).

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 :

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

```
> 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 determent of matrix and its inverse

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

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.

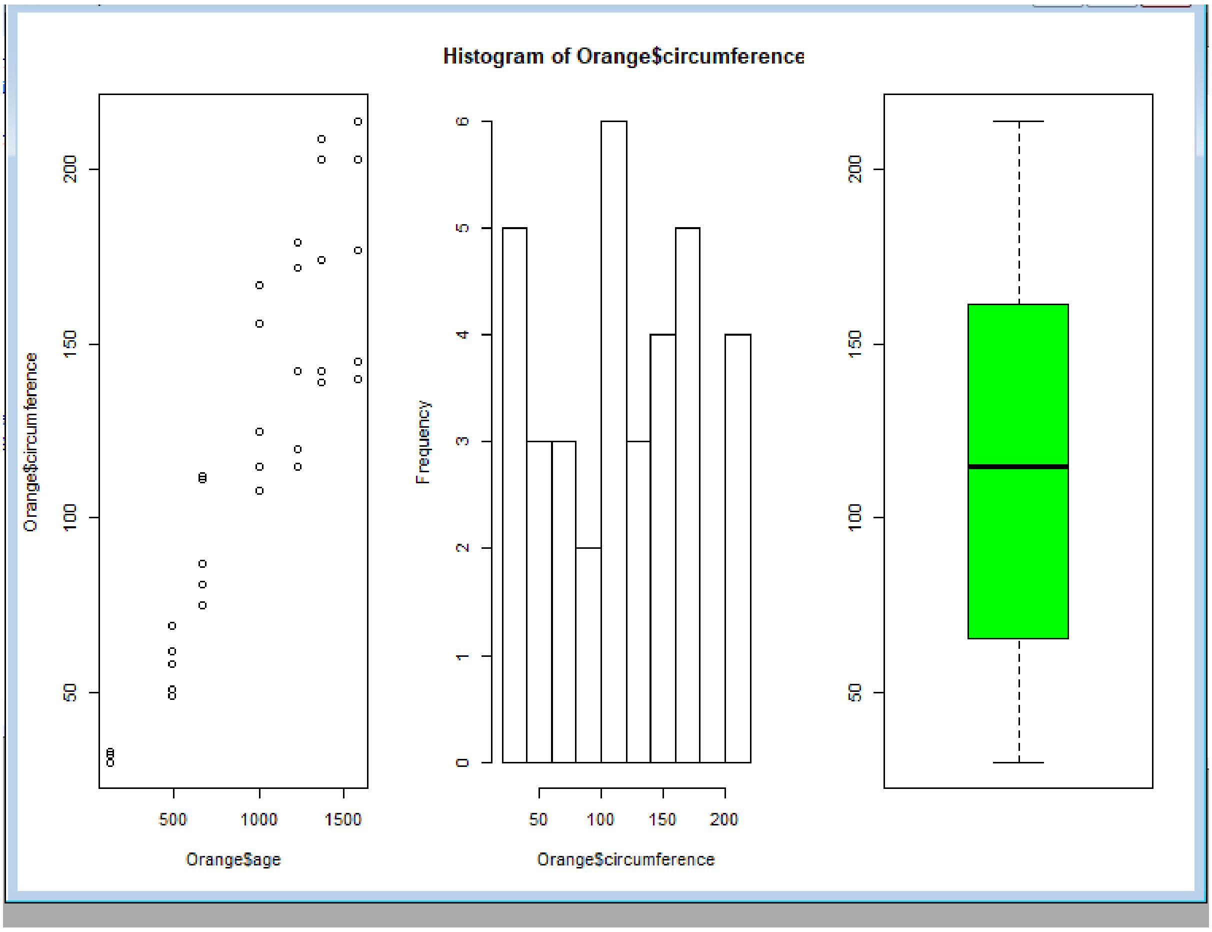
```
> Orange
   Tree  age circumference
1     1  118            30
2     1  484            58
3     1  664            87
.     .  .              .
.     .  .              .
.     .  .              .
30    5  484            49
31    5  664            81
32    5 1004           125
33    5 1231        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

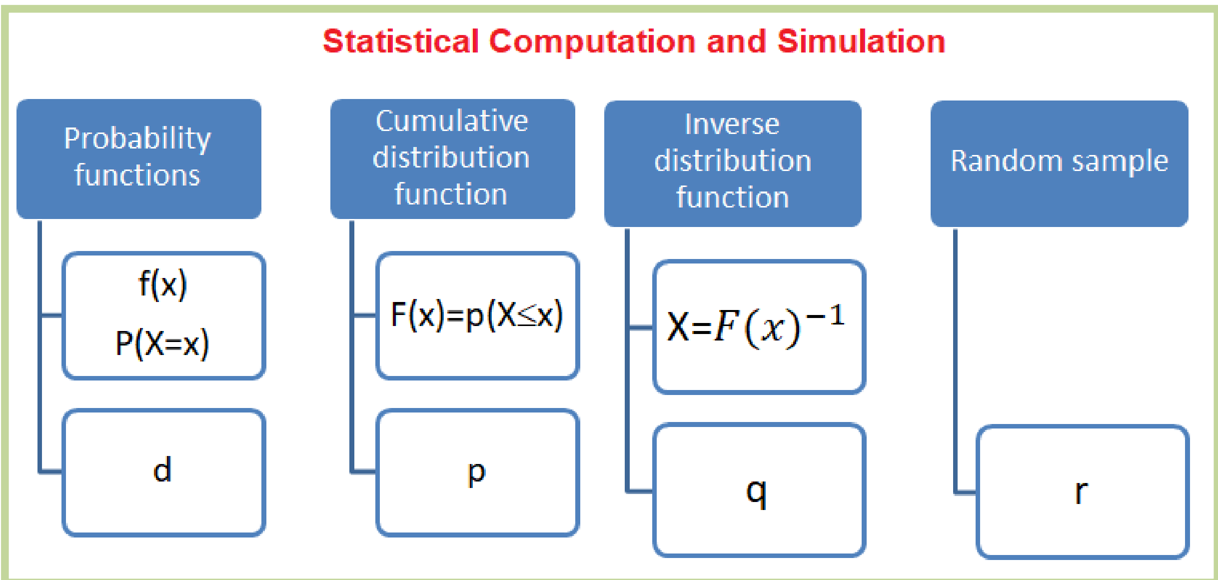
```
> attach(Orange)
> mean(age)
[1] 922.1429
> summary(circumference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   30.0    65.5   115.0   115.9   161.5   214.0 
```

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
50
100
150
200
500
1000
1500
0
1
2
3
4
5
6


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-$\mathbf{F(2.5) = P(X \leq 2.5)}$

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

3- Generate a random sample with size 10 from $\mathbf{N(2\ , 0.25^2)}$ **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 ≤ 5) = P(X≤ 5)- P(X≤ 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**

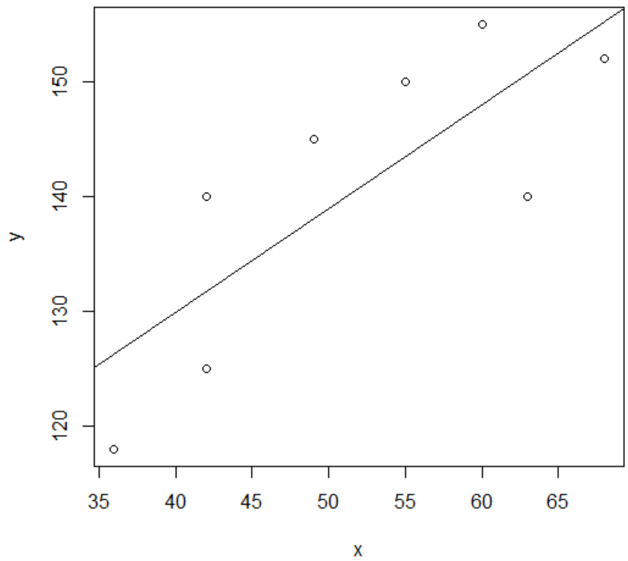
```
> # 2)correlation of X and Y:
> cor(x,y)
[1] 0.7918318
> cor.test(x,y)

        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 <u>mean of fruit shape greater than 1.02 .</u> Use α=0.05

1-Hypothesis :

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

2-Test statistics :

T=2.7002

3- Decision:

$$p-value = 0.01219 < \ \alpha = 0.05$$

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

```
> x <- c(1.066,1.084,1.076,1.051,1.059,1.020,1.035,1.052,1.046,0.976)
> t.test(x,mu=1.02,alternative ="greater",conf.level = 0.95)

        One Sample t-test

data:  x
t = 2.7002, df = 9, p-value = 0.01219
alternative hypothesis: true mean is greater than 1.02
95 percent confidence interval:
 1.02851      Inf
sample estimates:
mean of x
   1.0465
```

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= -14.988

3- **Decision:**

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

```
> Whole <- c(94.95,95.15,94.85,94.55,94.55,93.40,95.05,94.35,94.70,94.90)
> Skim <- c(91.25,91.80,91.50,91.65,91.15,90.25,91.90,91.25,91.65,91.00)
> t.test(Skim,Whole,alternative = "less",conf.level = 0.99,var.equal = TRUE)

        Two Sample t-test

data:  Skim and Whole
t = -14.988, df = 18, p-value = 6.533e-12
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
      -Inf -2.742173
sample estimates:
mean of x mean of y 
   91.340    94.645 
```

For confidence interval we change alternative to not equal

b) $\mu_{Skim} - \mu_{Whole} \in (-3.94\ , -2.67\ )$

```
> t.test(Skim,Whole,alternative = "two.sided",conf.level = 0.99)

        Welch Two Sample t-test

data:  Skim and Whole
t = -14.988, df = 17.97, p-value = 1.341e-11
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -3.939849 -2.670151
sample estimates:
mean of x mean of y 
   91.340    94.645 
```

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:

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

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)
```

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

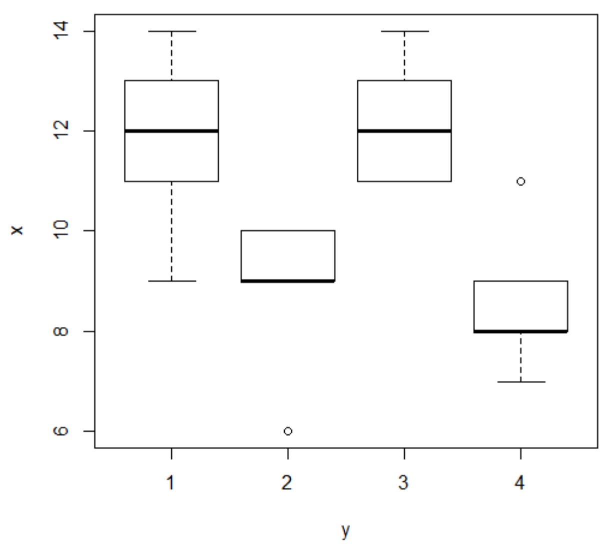
$\mu_{program\,1} = \mu_{program\,3}$

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

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

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

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



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

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

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

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

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

⇨ $B(6\,,5)$