Department of Statistics and Operations Research

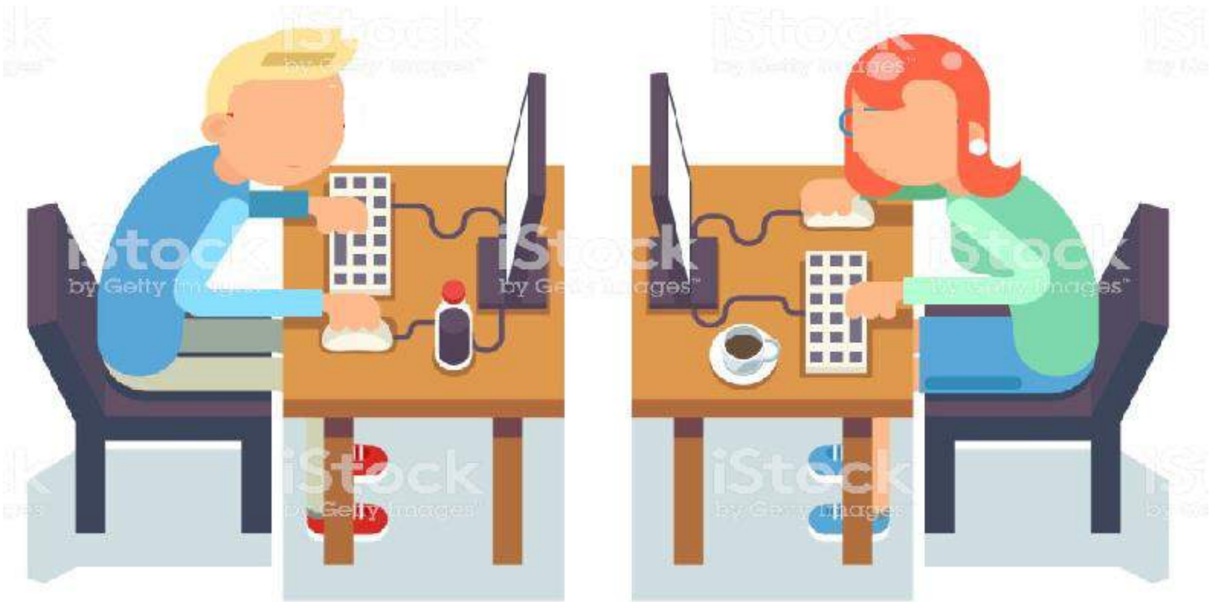
College of Science

King Saud University

Tutorial

STATISTICAL PACKAGES

STAT 328


iStock by Getty Images

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

Department of Statistics and Operations Research

College of Science

King Saud University

Tutorial

STATISTICAL PACKAGES(EXCEL)

STAT 328

MATHEMATICAL FUNCTIONS

Write the commands of the following:

		By Excel **(**using (*fx*)**)**	**By Minitab** calc → calculator
Absolute value	$\lvert -4 \rvert$=4	ABS(-4)	
Combinations	$\binom{10}{6}$=10C6=210	COMBIN(10;6)	
The exponential function	$e^{-1.6}$=0.201897	EXP(-1.6)	
Factorial	110! =1.5882E+178	FACT(110)	
Floor function	$[-3.15]$= -4	INT(-3.15)	
Natural logarithm	ln(23)= 3.135494216	LN(23)	
Logarithm with respect to any base	$\log_9(4)$ = 0.630929754	LOG(4;9)	
Logarithm with respect to base 10	log(12) = 1.079181246	LOG10(12)	
Multinomial Coefficient	$\binom{9}{2\quad 2\quad 5}$= 756	MULTINOMIAL(2;2;5)	
Square root	$\sqrt{85}$= 9.219544457	SQRT(85)	
Summation	Summation of: 450,11,20,5 = 486	SUM(450;11;20;5)	
Permutations	10P6=151200	PERMUT(10;6)	
Product	Product of: 450,11,20,5 = 495000	PRODUCT(450;11;20;5)	
Powers	10^{-4}= 0.0001	POWER(10;-4)	

CONDITIONAL FUNCTION (IF) AND COUNT CONDITIONAL FUNCTION

By Excel

(using (*fx*)**)**

We have grades of 10 students

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

1. Print student case being successful (Mark >=60) and being a failure (Mark <60).
2. How many successful students?
3. How many students whose grades are less than or equal to 80?

DESCRIPTIVE STATISTICS

We have students' weights as follows: 44, 40, 42,48,46,44. Find:

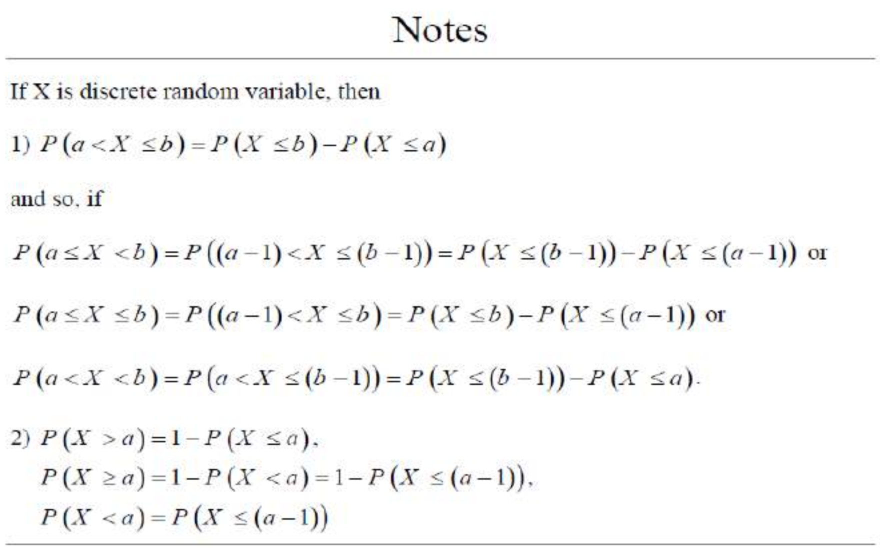
	By Excel (using (*fx*) and (Data Analysis))	**By Minitab** stat → basic statistics → display descriptive statistics + See Appendix -1-
Mean=44	AVERAGE(C2:C7)	
Median=44	MEDIAN(C2:C7)	
Mode=44	MODE.SNGL(C2:C7)	
Sample standard deviation=2.828	STDEV.S(C2:C7)	
Sample variance=8	VAR.S(C2:C7)	
Kurtosis=-0.3	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= $\frac{\text{Sample standard deviation}}{\text{Mean}} \times 10$

PROBABILITY DISTRIBUTION FUNCTIONS

Discrete Distributions



Notes

If X is discrete random variable, then

1) $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$

and so, if

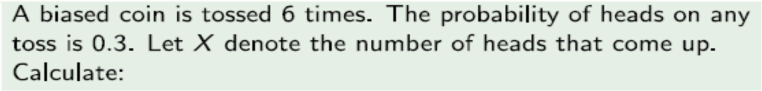
$P(a \leq X < b) = P((a-1) < X \leq (b-1)) = P(X \leq (b-1)) - P(X \leq (a-1))$ or

$P(a \leq X \leq b) = P((a-1) < X \leq b) = P(X \leq b) - P(X \leq (a-1))$ or

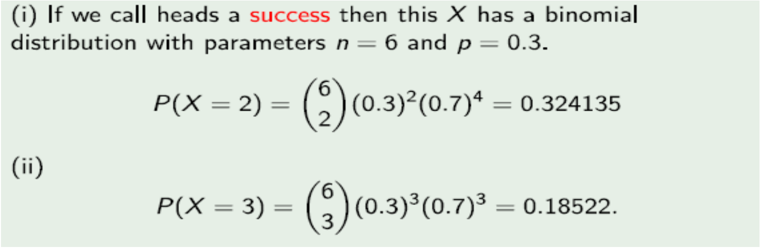
$P(a < X < b) = P(a < X \leq (b-1)) = P(X \leq (b-1)) - P(X \leq a)$.

2) $P(X > a) = 1 - P(X \leq a)$,
$P(X \geq a) = 1 - P(X < a) = 1 - P(X \leq (a-1))$,
$P(X < a) = P(X \leq (a-1))$

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:

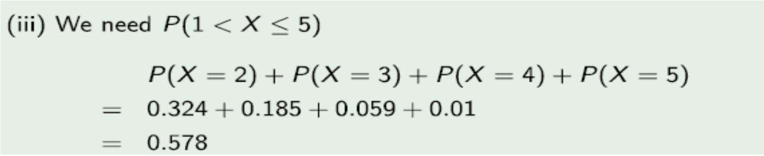


(i) If we call heads a success then this X has a binomial distribution with parameters $n = 6$ and $p = 0.3$.

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

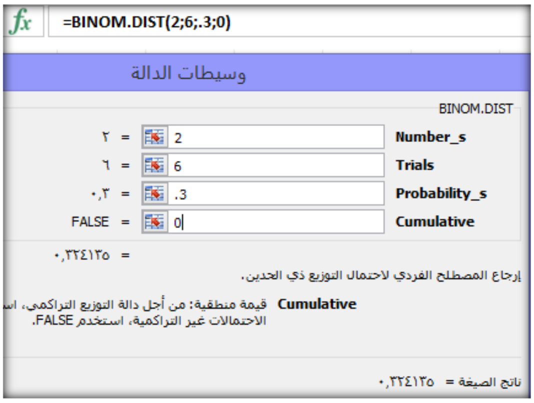
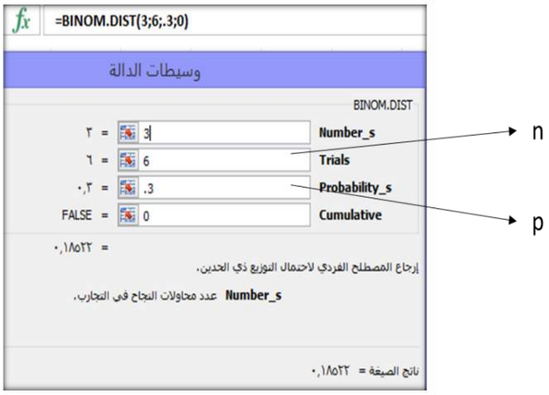
(ii)

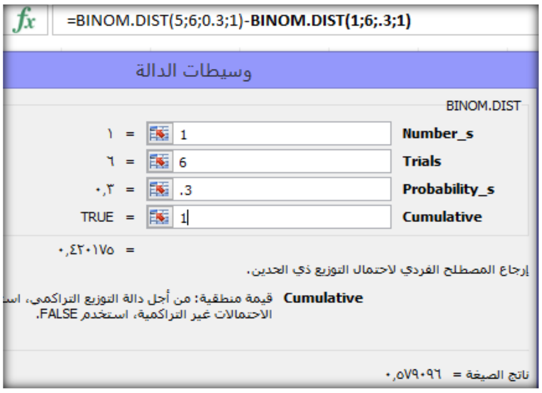
$$P(X = 3) = \binom{6}{3}(0.3)^3(0.7)^3 = 0.18522.$$



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

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

	By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
i	 =BINOM.DIST(2;6;.3;0) وسيطات الدالة BINOM.DIST Number_s 2 = ٢ Trials 6 = ٦ Probability_s .3 = ٠,٣ Cumulative 0 = FALSE = ٠,٣٢٤١٣٥ إرجاع المصطلح الفردي لاحتمال التوزيع ذي الحدين. **Cumulative** قيمة منطقية: من أجل دالة التوزيع التراكمي، اس الاحتمالات غير التراكمية، استخدم FALSE. ناتج الصيغة = ٠,٣٢٤١٣٥	
ii	 =BINOM.DIST(3;6;.3;0) وسيطات الدالة BINOM.DIST Number_s 3 = ٣ Trials 6 = ٦ Probability_s .3 = ٠,٣ Cumulative 0 = FALSE = ٠,١٨٥٢٢ إرجاع المصطلح الفردي لاحتمال التوزيع ذي الحدين. **Number_s** عدد محاولات النجاح في التجارب. ناتج الصيغة = ٠,١٨٥٢٢ 	

iii	 =BINOM.DIST(5;6;0.3;1)-**BINOM.DIST(1;6;.3;1)** وسيطات الدالة BINOM.DIST **Number_s** 1 = ١ **Trials** 6 = ٦ **Probability_s** .3 = ٠,٣ **Cumulative** 1 = TRUE = ٠,٤٢٠١٧٥ إرجاع المصطلح الفردي لاحتمال التوزيع ذي الحدين. **Cumulative** قيمة منطقية: من أجل دالة التوزيع التراكمي، اس الاحتمالات غير التراكمية، استخدم FALSE. ناتج الصيغة = ٠,٥٧٩٠٩٦	

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?

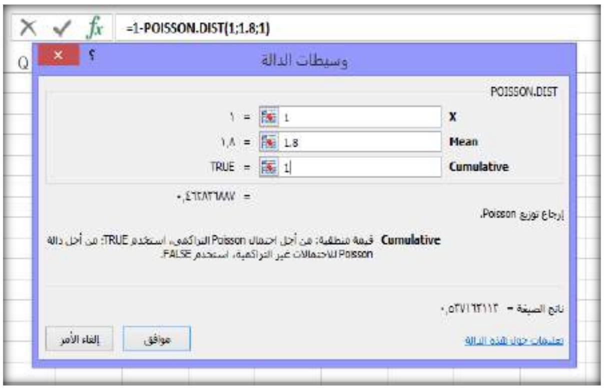
We want $P(X \geq 2) = P(X = 2) + P(X = 3) + \ldots$

i.e. an infinite number of probabilities to calculate

but

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

(using (*fx*))	**By Minitab** calc → probability distribution
 =POISSON.DIST(4;1.8;0) وسيطات الدالة POISSON.DIST X 4 = ٤ Mean 1.8 = ١,٨ Cumulative 0 = FALSE = ٠,٠٧٢٣٠١٧٢٤ إرجاع توزيع Poisson. Cumulative قيمة منطقية: من أجل احتمال Poisson التراكمي، استخدم TRUE؛ من أجل دالة Poisson للاحتمالات غير التراكمية، استخدم FALSE. ناتج الصيغة = ٠,٠٧٢٣٠١٧٢٤ تعليمات حول هذه الدالة 	

	 =1-POISSON.DIST(1;1.8;1) وسيطات الدالة POISSON.DIST X: 1 = ١ Mean: 1.8 = ١,٨ Cumulative: 1 = TRUE = ٠,٤٦٢٨٣٦٨٨٧ إرجاع توزيع Poisson. **Cumulative** قيمة منطقية: من أجل احتمال Poisson التراكمي، استخدم TRUE: من أجل دالة Poisson للاحتمالات غير التراكمية، استخدم FALSE. ناتج الصيغة = ٠,٥٣٧١٦٣١١٣

Continuous Distributions

Notes

If X is continuous symmetric random variable (as Normal distribution and Student's t-distribution), then

1) $P(X \geq x) = 1 - P(X \leq x)$ and $P(X \leq x) = 1 - P(X \geq x)$
2) $P(X \leq x) = 1 - P(X \leq -x)$ and $P(X \geq x) = 1 - P(X \geq -x)$

---

1. Exponential Distribution

On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.

What is the probability that a computer part lasts more than 7 years?

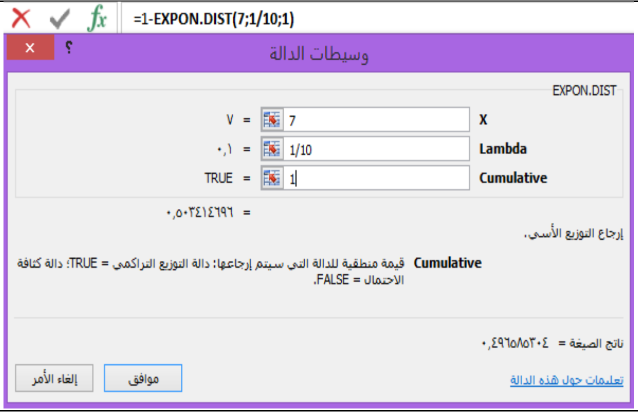
Solution

Let X = the amount of time (in years) a computer part lasts.

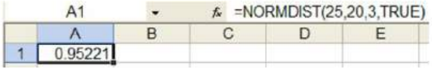
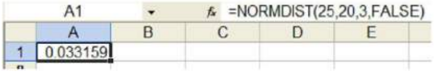
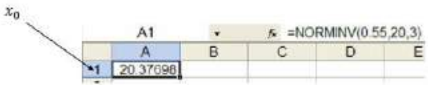
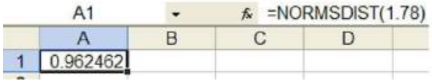
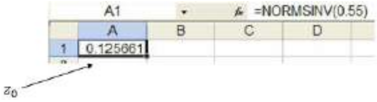
$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

$P(X > 7) = 1 - P(X < 7)$.

$P(X > 7) = e^{-0.1 \cdot 7} = 0.4966$. The probability that a computer part lasts more than 7 years is 0.4966.

By Excel (using (*fx*))	**By Minitab** calc → probability distribution
 =1-EXPON.DIST(7;1/10;1) وسيطات الدالة EXPON.DIST X = 7 = ٧ Lambda = 1/10 = ٠,١ Cumulative = 1 = TRUE = ٠,٥٠٣٤١٤٦٩٦ إرجاع التوزيع الأسي. **Cumulative** قيمة منطقية للدالة التي سيتم إرجاعها: دالة التوزيع التراكمي = TRUE؛ دالة كثافة الاحتمال = FALSE. ناتج الصيغة = ٠,٤٩٦٥٨٥٣٠٤ 	

2. Normal Distribution

	By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
$P(X \leq 25)$ $= P(X < 25)$ at $\mu = 20$ and $\sigma = 3$	 A1 =NORMDIST(25,20,3,TRUE) 1 0.95221	
$f_X(25)$ at $\mu = 20$ and $\sigma = 3$	 A1 =NORMDIST(25,20,3,FALSE) 1 0.033159	
$P(X \leq x_0)$ $= P(X < x_0)$ $= .55$ at $\mu = 20$ and $\sigma = 3$	 A1 =NORMINV(0.55,20,3) 1 20.37698	
$P(Z \leq 1.78)$ $= P(Z < 1.78)$ at $\mu = 0$ and $\sigma = 1$	 A1 =NORMSDIST(1.78) 1 0.962462	
$P(Z \leq z_0) = .55$ at $\mu = 0$ and $\sigma = 1$	 A1 =NORMSINV(0.55) 1 0.125661	

3. Student’s t Distribution

Notes in Excel

1) **=T.DIST(*t*, *v*; 0)** $\leftrightarrow f_{T_v}(t)$
2) **=T.DIST(*t*, *v*; 1)** $\leftrightarrow P(T_v \le t)$
3) **=T.DIST.RT(*t*, *v*)** $\leftrightarrow P(T_v \ge t)$

4) **=T.DIST.2T(*t*, *v*)** | $\leftrightarrow 2P(T_v \ge t)$

5) **=T.INV(*p*, *v*)** $\leftrightarrow P(T_v \le t_0) = p$

6) **=T.INV.2T(*p*, *v*)** $\leftrightarrow 2P(T_v \ge t_0) = p$

Find: $(a)\, t_{0.025}$ *when* $v = 14$

$(b)\, t_{0.01}$ *when* $v = 10$

$(c)\, t_{0.995}$ *when* $v = 7$

	By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
(a) $P(T_{14} \le t) = 0.025$	 =T.INV(.025;14) وسيطات الدالة T.INV Probability .025 = ٠,٠٢٥ Deg_freedom 14 = ١٤ = ٢,١٤٤٧٨٦٦٨٨- إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. Deg_freedom عدد صحيح موجب يشير إلى عدد درجات الحرية التي تميز التوزيع. ناتج الصيغة = ٢,١٤٤٧٨٦٦٨٨- تعليمات حول هذه الدالة موافق إلغاء الأمر	
(b) $P(T_{10} < t) = 0.01$	 وسيطات الدالة T.INV Probability .01 = ٠,٠١ Deg_freedom 10 = ١٠ = ٢,٧٦٣٧٦٩٤٥٨- إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. Deg_freedom عدد صحيح موجب يشير إلى عدد درجات الحرية التي تميز التوزيع. ناتج الصيغة = ٢,٧٦٣٧٦٩٤٥٨- تعليمات حول هذه الدالة موافق إلغاء الأمر	
(c) $P(T_7 < t) = 0.995$	 =T.INV(0.995;7) وسيطات الدالة T.INV Probability 0.995 = ٠,٩٩٥ Deg_freedom 7 = ٧ = ٣,٤٩٩٤٨٣٢٩٧ إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. Deg_freedom عدد صحيح موجب يشير إلى عدد درجات الحرية التي تميز التوزيع. ناتج الصيغة = ٣,٤٩٩٤٨٣٢٩٧ تعليمات حول هذه الدالة موافق إلغاء الأمر	

Given a random sample of size **24** from a normal distribution, find **k** such that:

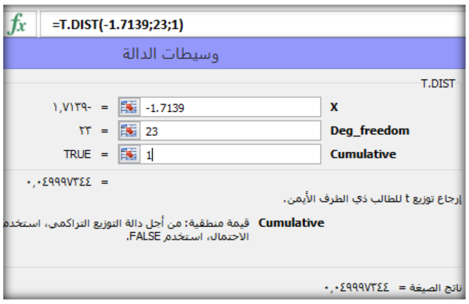
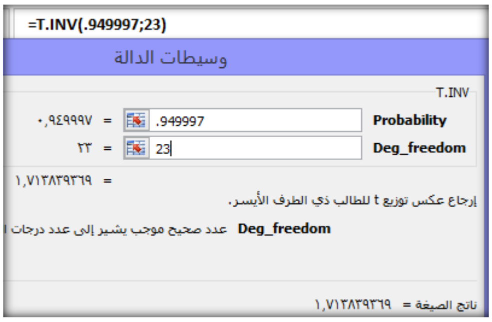
$$(a)\, P(-1.7139 < T < k) = 0.90$$

$$(b)\, P(k < T < 2.807) = 0.95$$

$$(c)\, P(-k < T < k) = 0.90$$

(a)

$P(-1.7139 < T_{23} < k) = 0.9$
$\leftrightarrow P(T_{23} < k) - P(T_{23} < -1.7139) = 0.9$
$\leftrightarrow P(T_{23} < k) = 0.9 + P(T_{23} < -1.7139)$
$\leftrightarrow P(T_{23} < k) = \ 0.949997$

By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
 =T.DIST(-1.7139;23;1) وسيطات الدالة T.DIST X: -1.7139 = ١,٧١٣٩- Deg_freedom: 23 = ٢٣ Cumulative: 1 = TRUE = ٠,٠٤٩٩٩٧٣٤٤ إرجاع توزيع t للطالب ذي الطرف الأيمن. **Cumulative** قيمة منطقية: من أجل دالة التوزيع التراكمي، استخدم الاحتمال، استخدم FALSE. ناتج الصيغة = ٠,٠٤٩٩٩٧٣٤٤  =T.INV(.949997;23) وسيطات الدالة T.INV Probability: .949997 = ٠,٩٤٩٩٩٧ Deg_freedom: 23 = ٢٣ = ١,٧١٣٨٣٩٣٦٩ إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. **Deg_freedom** عدد صحيح موجب يشير إلى عدد درجات ناتج الصيغة = ١,٧١٣٨٣٩٣٦٩ ○*In excel you might make it in one step too* $P(T_{23} < k) = 0.9 + P(T_{23} < -1.7139)$ *so,* $\quad = T.INV(0.9 + T.DIST(-1.7139{,}23{,}1), 23) \ = 1.713839369$	

(b)

$P(k < T_{23} < 2.807) = 0.95$
$\leftrightarrow P(T_{23} < 2.807) - P(T_{23} < k) = 0.95$
$\leftrightarrow P(T_{23} < k) = (T_{23} < 2.807) - 0.95$
$\leftrightarrow P(T_{23} < k) \;= 0.044996$

By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
 =T.DIST(2.807;23;1) وسيطات الدالة T.DIST **X** 2.807 = ٢,٨٠٧ **Deg_freedom** 23 = ٢٣ **Cumulative** 1 = TRUE = ٠,٩٩٤٩٩٦١٣٩ إرجاع توزيع t للطالب ذي الطرف الأيمن. **X** القيمة الرقمية المراد تقييم التوزيع عندها. ناتج الصيغة = ٠,٩٩٤٩٩٦١٣٩  =T.INV(D1;23) وسيطات الدالة T.INV **Probability** D1 = ٠,٠٤٤٩٩٦١٣٩ **Deg_freedom** 23 = ٢٣ = -١,٧٦٩٩٥٢٥٧٦ إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. **Deg_freedom** عدد صحيح موجب يشير إلى عدد درجات الحرية التي تم ناتج الصيغة = -١,٧٦٩٩٥٢٥٧٦ ○*In excel you might make it in one step too* $P(T_{23} < k) = (T_{23} < 2.807) - 0.95$ *so,* $\quad = T.INV(-0.95 + T.DIST(2.807{,}23{,}1), 23) \;= -1.769952576$	

(c)

$(i)\ P(T_{23} < k) - P(T_{23} < -k) = .9$
$\leftrightarrow P(T_{23} < k) - \{1 - P(T_{23} < k)\} = 0.9$
$\leftrightarrow 2P(T_{23} < k) - 1 = 0.9$
$\leftrightarrow 2P(T_{23} < k) = 1.9$
$\leftrightarrow P(T_{23} < k) = 0.95$
$so, \quad = T.inv\ (0.95{,}23) = 1.71387$

$(ii) P(T_{23} < k) - P(T_{23} < -k) = .9$
$\leftrightarrow 1 - P(T_{23} > k) - P(T_{23} < -k) = 0.9$
$\leftrightarrow 1 - P(T_{23} > k) - \{1 - P(T_{23} > -k)\} = 0.9$
$\leftrightarrow 1 - P(T_{23} > k) - \{1 - [1 - P(T_{23} > k)]\} = 0.9$
$\leftrightarrow 1 - P(T_{23} > k) - \{1 - 1 + P(T_{23} > k)\} = 0.9$
$\leftrightarrow 1 - P(T_{23} > k) - P(T_{23} > k) = 0.9$
$\leftrightarrow 1 - 2P(T_{23} > k) = 0.9$
$\leftrightarrow 2P(T23 > k) = 0.1$
$so, \quad = T.inv.2t(0.1{,}23) = 1.71387$

By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
 =T.INV(.95;23) وسيطات الدالة T.INV Probability .95 = ٠,٩٥ Deg_freedom 23 = ٢٣ = ١,٧١٣٨٧١٥٢٨ إرجاع عكس توزيع t للطالب ذي الطرف الأيسر. Probability الاحتمال المقترن بتوزيع t للطالب ثنائي الطرف، رقم بين ناتج الصيغة = ١,٧١٣٨٧١٥٢٨  =T.INV.2T(.1;23) وسيطات الدالة T.INV.2T Probability .1 = ٠,١ Deg_freedom 23 = ٢٣ = ١,٧١٣٨٧١٥٢٨ إرجاع عكس توزيع t للطالب ثنائي الطرف. Deg_freedom عدد صحيح موجب يشير إلى عدد درجات الحرية ناتج الصيغة = ١,٧١٣٨٧١٥٢٨	

4. Chi-Square Distribution

Notes in Excel

1) $= \mathbf{CHISQ.DIST}(x, v, 0) \quad \leftrightarrow \quad f_{\chi_v}(x)$
2) $= \mathbf{CHISQ.DIST}(x, v, 1) \quad \leftrightarrow P(\chi_v \le x)$
3) $= \mathbf{CHISQ.DIST.RT}(x, v, 1) \quad \leftrightarrow P(\chi_v \ge x)$
4) $= \mathbf{CHISQ.INV}(p\,, v) \quad \leftrightarrow P(\chi_v \le x_0) = p$
5) $= \mathbf{CHISQ.INV.RT}(p\,, v) \quad \leftrightarrow P(\chi_v \ge x_0) = p$

By using chi- square distribution ,Find:

$$\chi^2_{0.995} \; when \quad v = 19$$

	By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
P(χ_{19} < x) =0.995	 =CHISQ.INV(.995;19) وسيطات الدالة CHISQ.INV Probability .995 = ٠,٩٩٥ Deg_freedom 19 = ١٩ = ٣٨,٥٨٢٢٥٦٥٥ إرجاع عكس الاحتمال ذي الطرف الأيسر لتوزيع كاي تربيع. Deg_freedom عدد درجات الحرية، رقم بين ١ و١٠^١٠، باستـ ناتج الصيغة = ٣٨,٥٨٢٢٥٦٥٥	

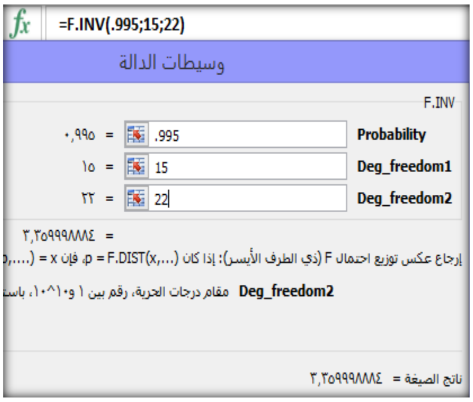
5. F Distribution

Notes in Excel

1) $= \mathbf{F.DIST}(f, v_1, v_2, 0) \quad \leftrightarrow \quad f_{F_{v_1,v_2}}(f)$
2) $= \mathbf{F.DIST}(f, v_1, v_2, 1) \quad \leftrightarrow P(F_{v_1,v_2} \leq f)$
3) $= \mathbf{F.DIST.RT}(f, v_1, v_2, 1) \quad \leftrightarrow P(F_{v_1,v_2} \geq f)$
4) $= \mathbf{F.INV}(p, v_1, v_2) \quad \leftrightarrow P\left(F_{v_1,v_2} \leq f_0\right) = p$
5) $= \mathbf{F.INV.RT}(p, v_1, v_2) \quad \leftrightarrow P\left(F_{v_1,v_2} \geq f_0\right) = p$

From the tables of F- distribution ,Find:

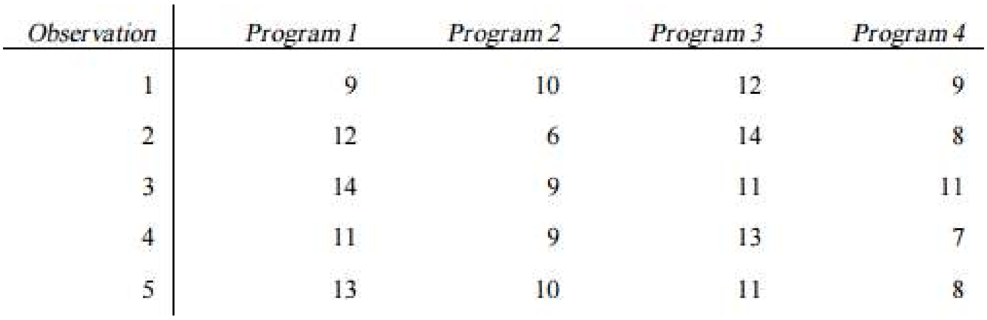
$F_{0.995,15,22}$

	By Excel **(using (*fx*))**	**By Minitab** calc → probability distribution
$P(F_{15,22} < f)$ $=0.995$	 =F.INV(.995;15;22) وسيطات الدالة F.INV Probability .995 = ٠,٩٩٥ Deg_freedom1 15 = ١٥ Deg_freedom2 22 = ٢٢ = ٣,٢٥٩٩٩٨٨٨٤ إرجاع عكس توزيع احتمال F (ذي الطرف الأيسر): إذا كان p = F.DIST(x,...)، فإن x = (...,p Deg_freedom2 مقام درجات الحرية، رقم بين ١ و١٠^١٠، باس ناتج الصيغة = ٣,٢٥٩٩٩٨٨٨٤	

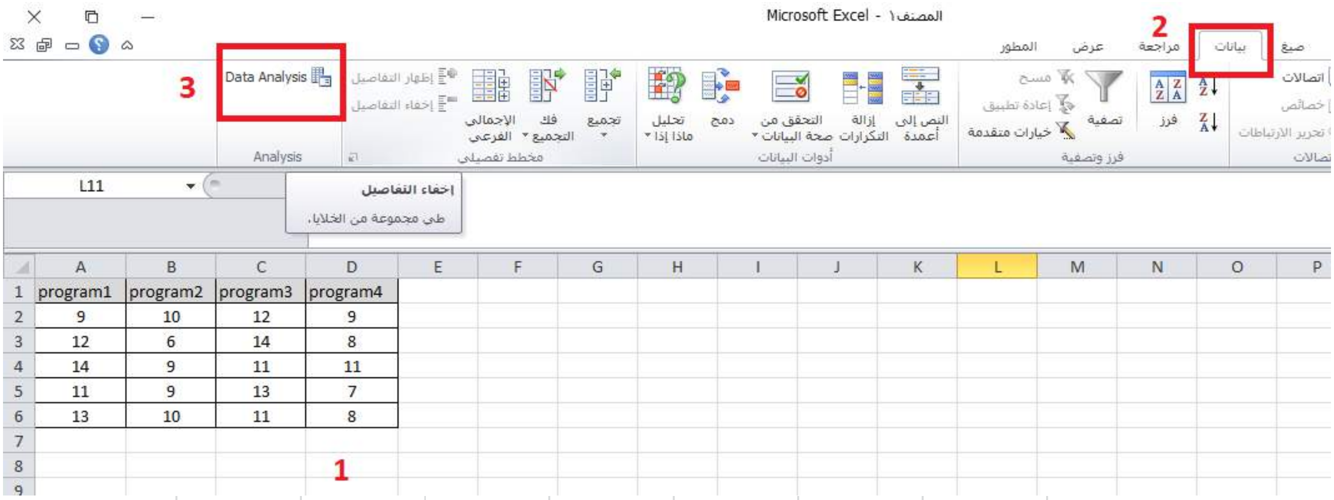
HYPOTHESIS TESTING STATISTICS AND CONFIDENCES INTERVAL

1)

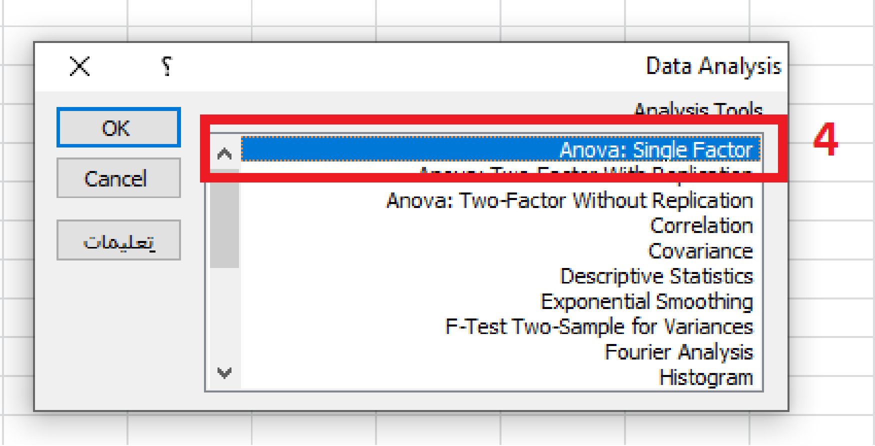
A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number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

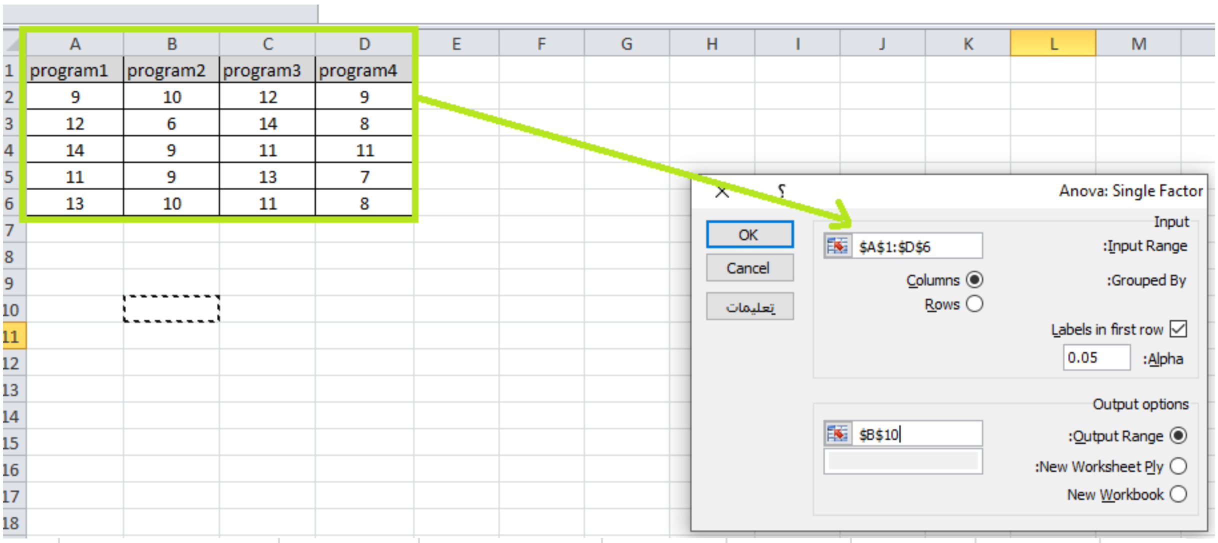


Observation	*Program 1*	*Program 2*	*Program 3*	*Program 4*
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8



	A	B	C	D
1	program1	program2	program3	program4
2	9	10	12	9
3	12	6	14	8
4	14	9	11	11
5	11	9	13	7
6	13	10	11	8





program1	program2	program3	program4
9	10	12	9
12	6	14	8
14	9	11	11
11	9	13	7
13	10	11	8

Anova: Single Factor

Input

:Input Range $A$1:$D$6

:Grouped By Columns ● Rows ○

Labels in first row ☑

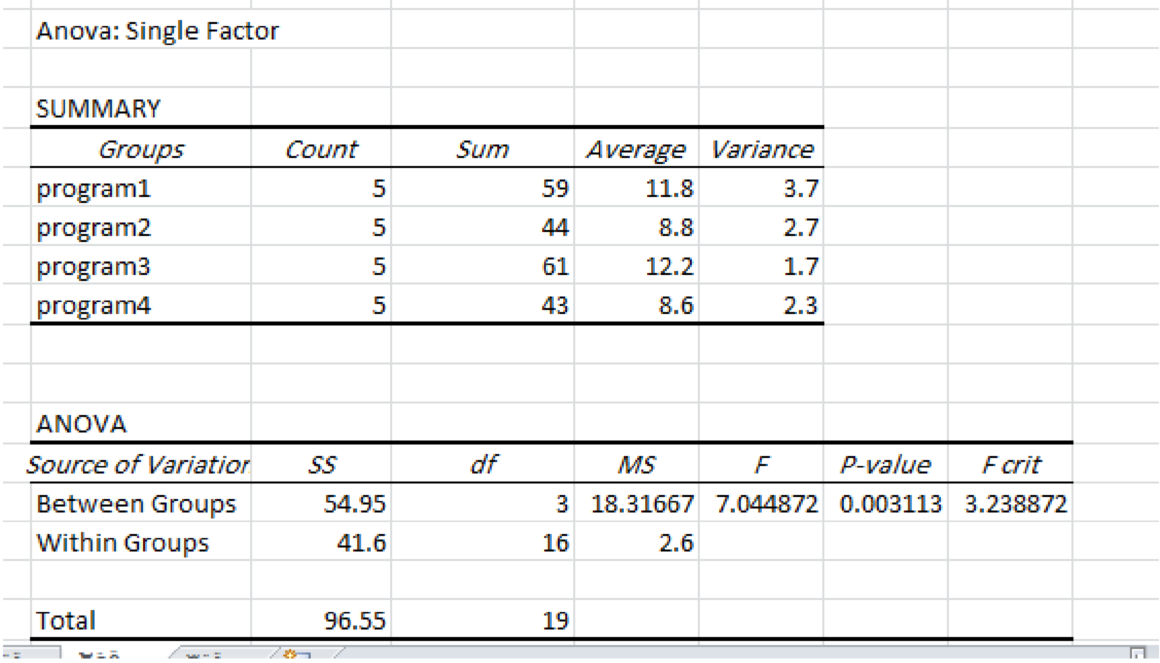
:Alpha 0.05

Output options

:Output Range ● $B$10

:New Worksheet Ply ○

New Workbook ○



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

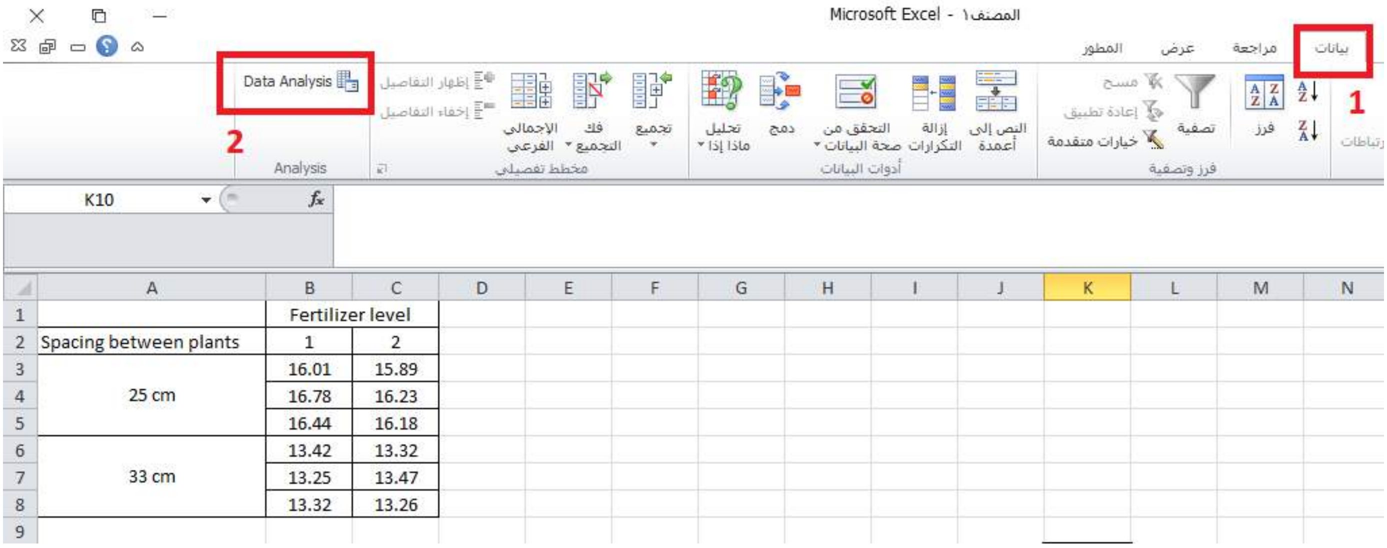
ANOVA

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				

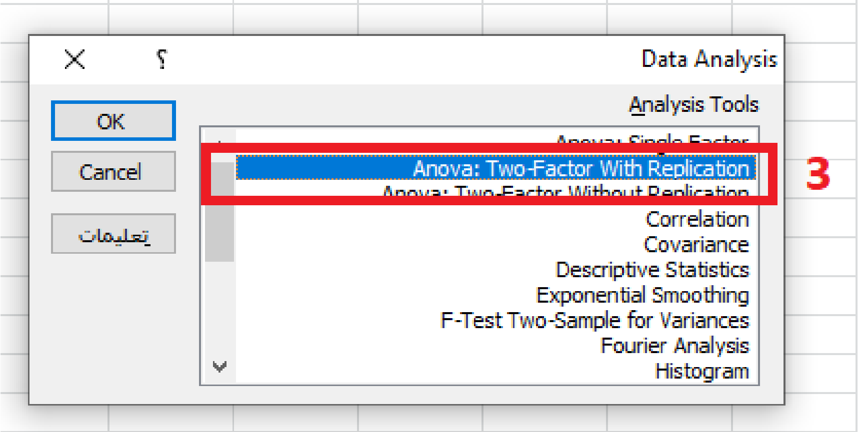
2) In a study on fertilizer levels and spacing between plants, plots were assigned to combinations and the yield of potatoes (in kg/plot) was measured

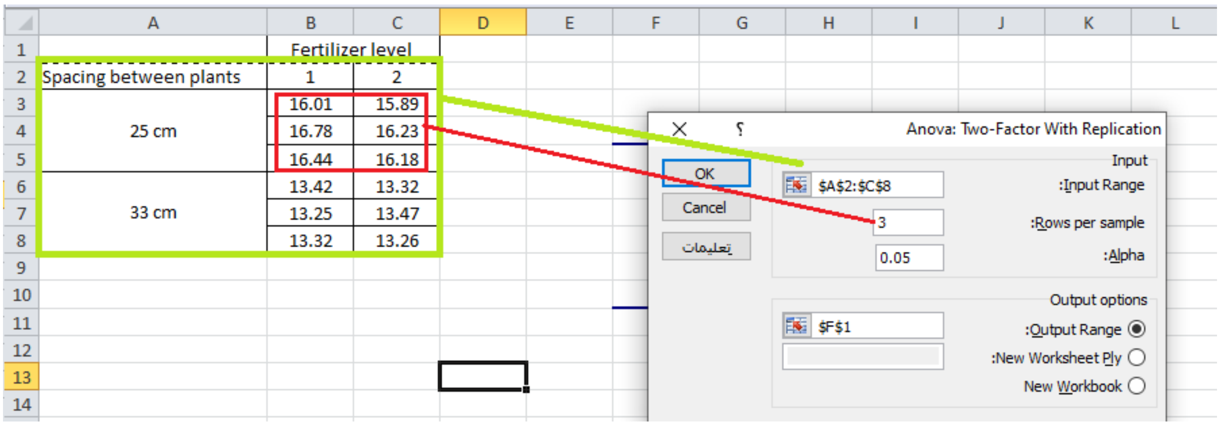
Spacing between plants	Fertilizer level (in tons/ha)	
	1	2
25 cm	16.01	15.89
	16.78	16.23
	16.44	16.18
33 cm	13.42	13.32
	13.25	13.47
	13.32	13.26

Make all appropriate tests ($\alpha=0.05$)



	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




	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

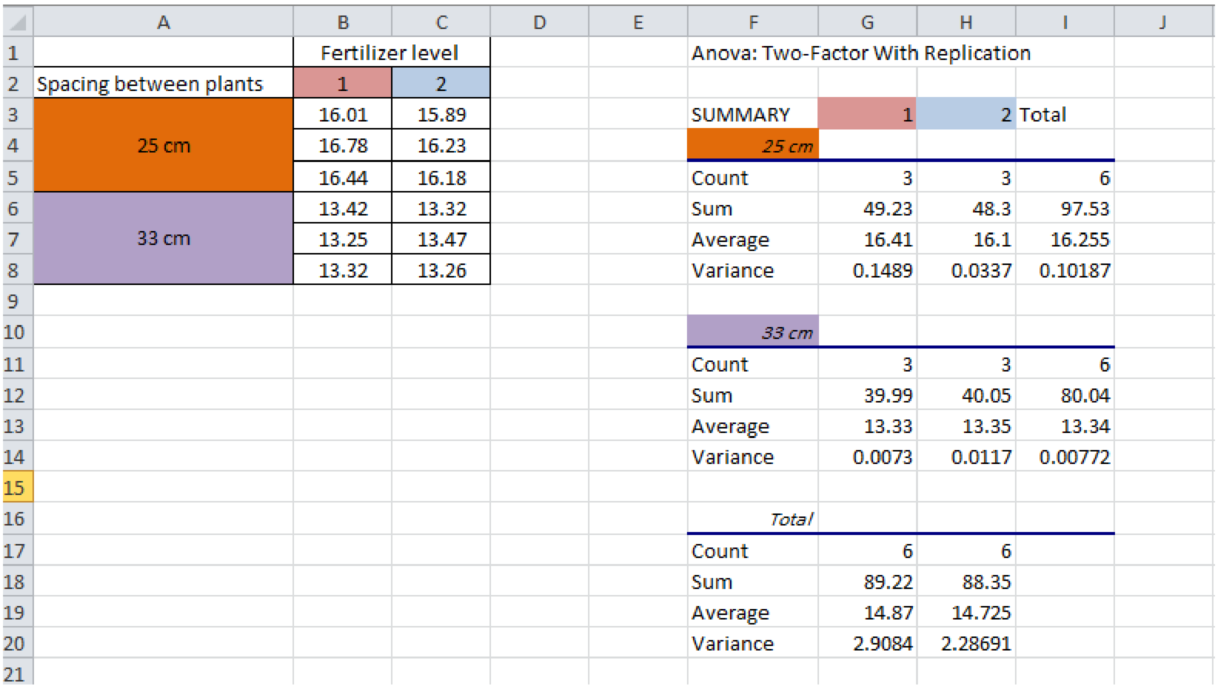
:Alpha 0.05

Output options

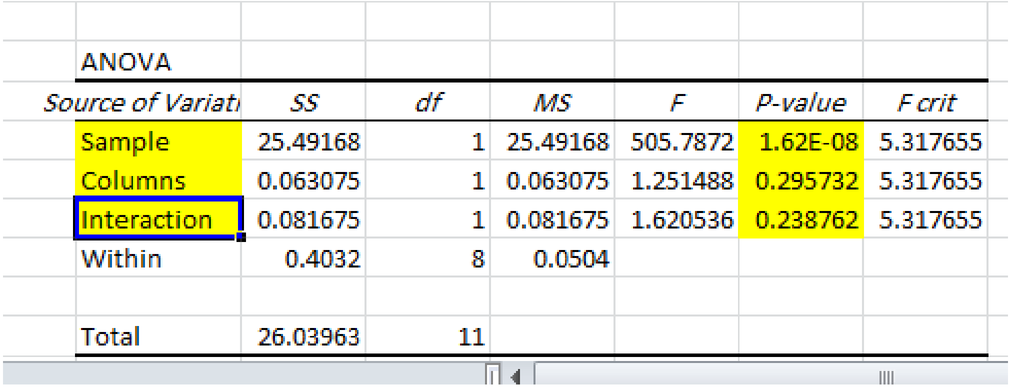
:Output Range ● $F$1

: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										



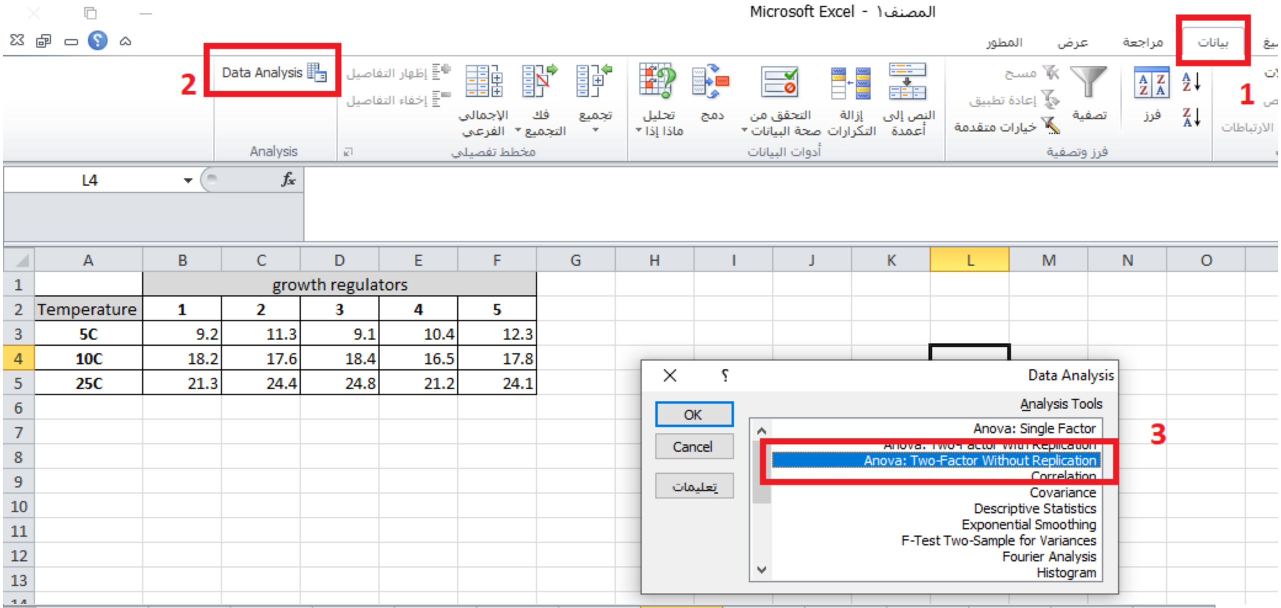
ANOVA

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				

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

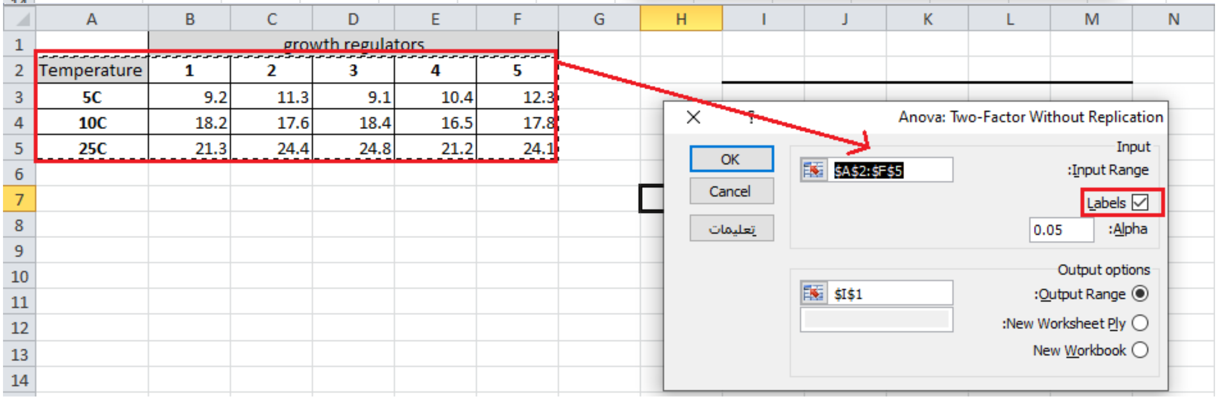
Assuming no interaction, test if there is a difference in the effects of the five growth regulators on the percent weight loss of oranges. Also test if there a difference in the effects of the three storage temperature (α=0.05)



	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 With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram



Anova: Two-Factor Without Replication

Input

:Input Range $A$2:$F$5

Labels ☑

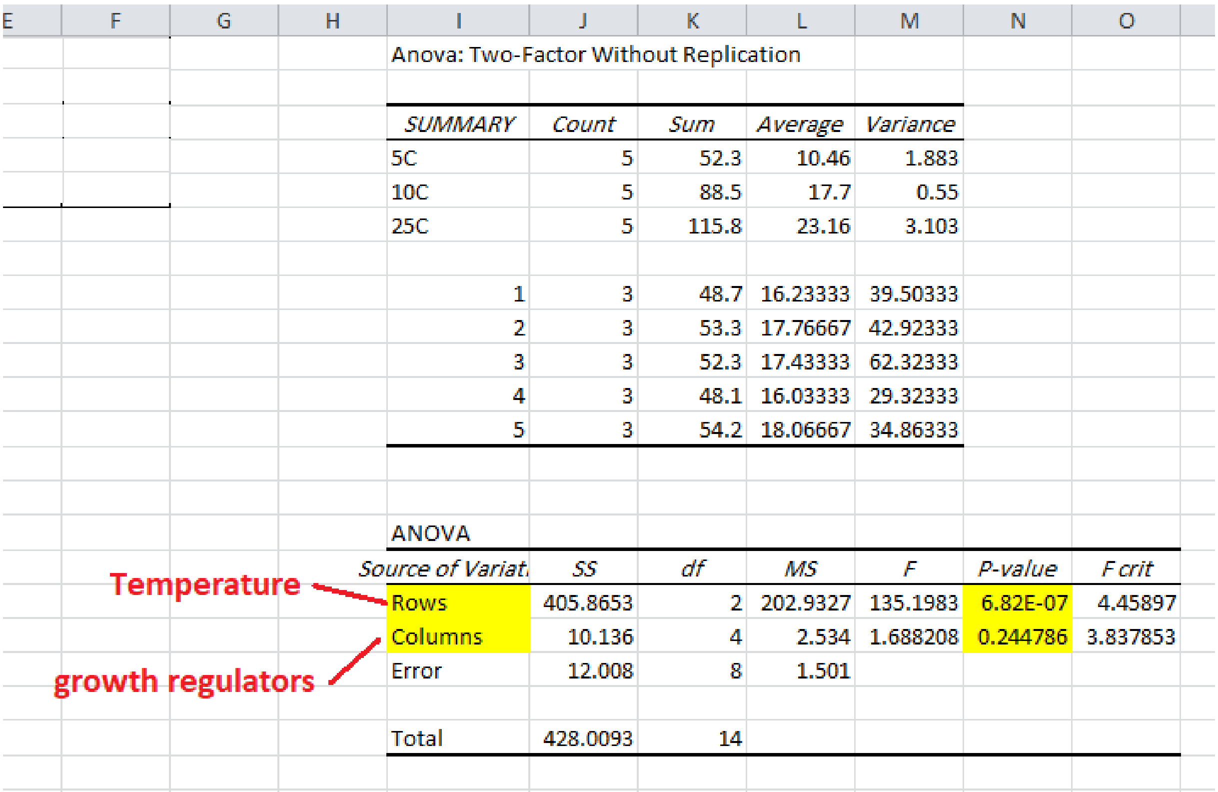
:Alpha 0.05

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	405.8653	2	202.9327	135.1983	6.82E-07	4.45897
Columns	10.136	4	2.534	1.688208	0.244786	3.837853
Error	12.008	8	1.501			
Total	428.0093	14				

Temperature

growth regulators

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

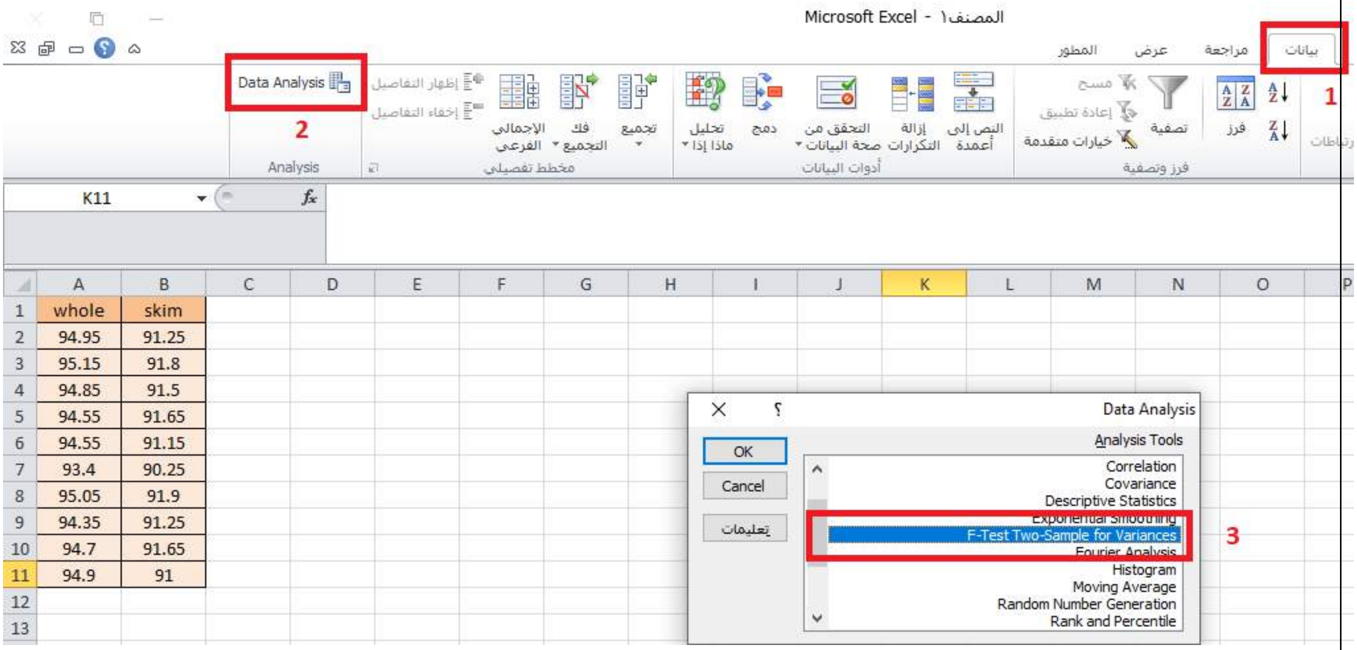
Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

Assuming normal populations with equal variance

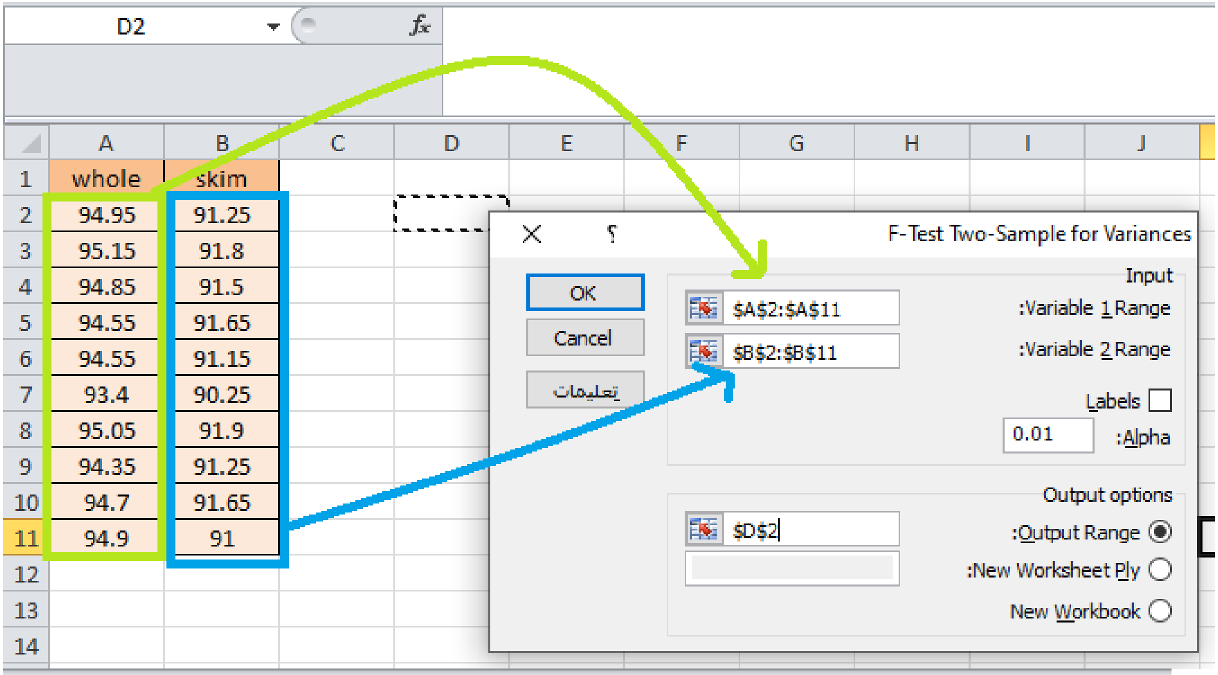
a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**

1-Test for equality of variance :

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



	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

F-Test Two-Sample for Variances

Input

:Variable 1 Range $A$2:$A$11

:Variable 2 Range $B$2:$B$11

Labels ☐

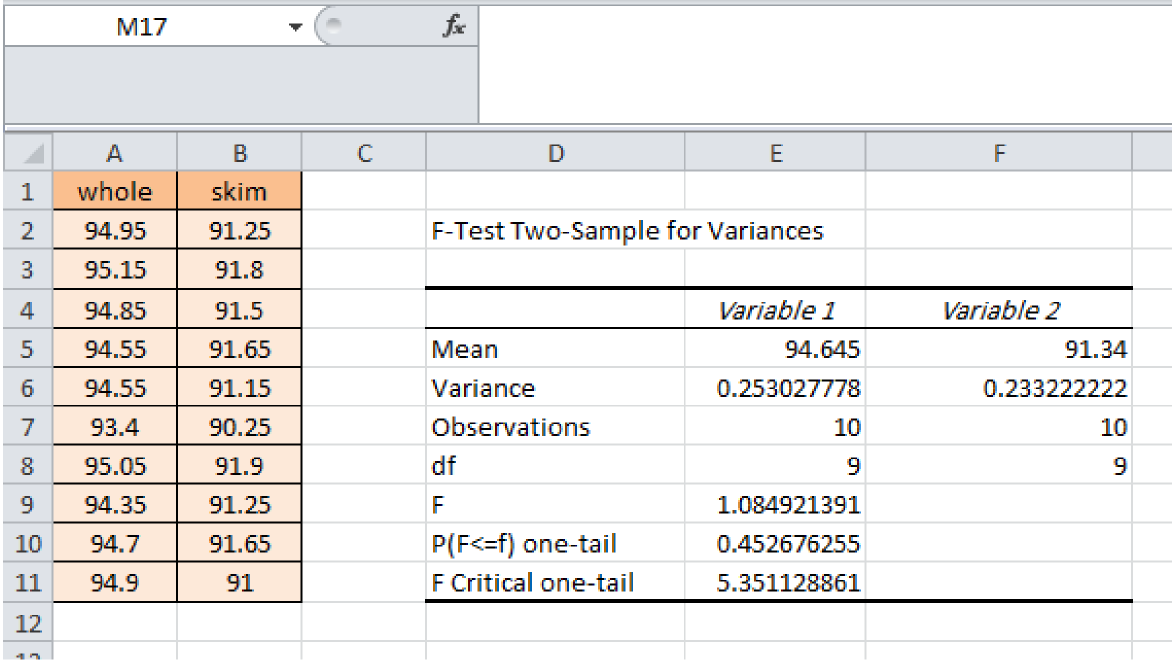
:Alpha 0.01

Output options

:Output Range ● $D$2

:New Worksheet Ply ○

New Workbook ○



	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

F-Test Two-Sample for Variances

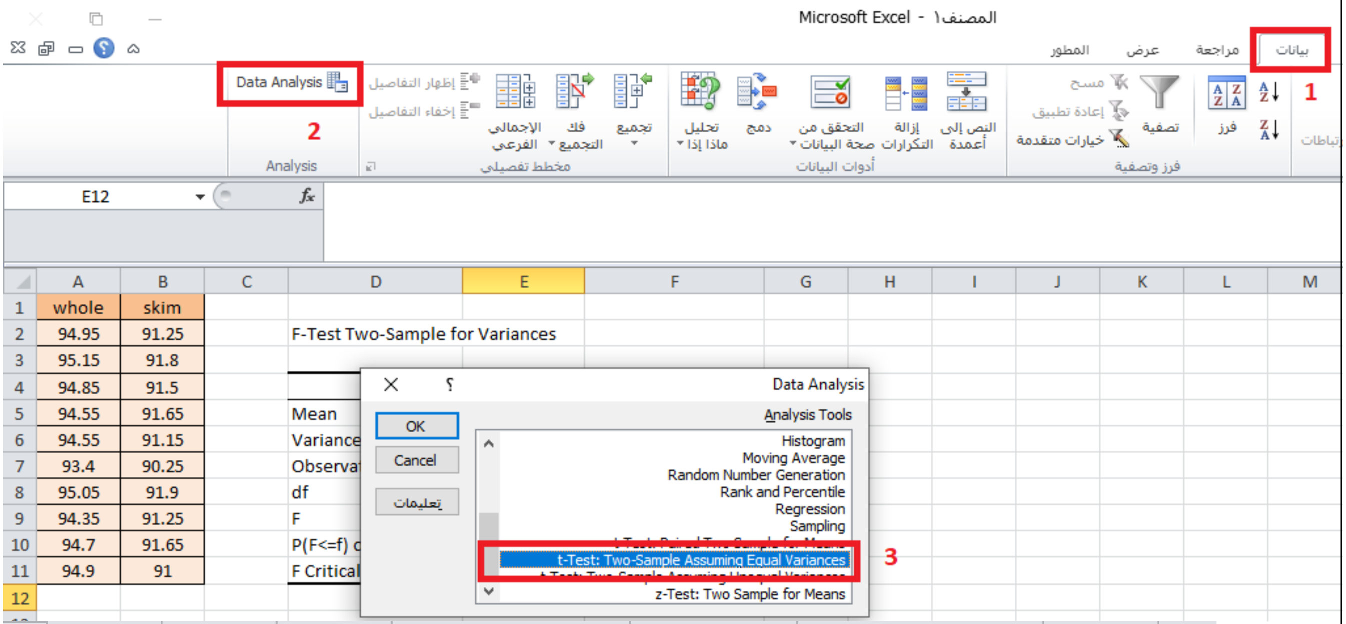
	Variable 1	*Variable 2*
Mean	94.645	91.34
Variance	0.253027778	0.233222222
Observations	10	10
df	9	9
F	1.084921391	
P(F<=f) one-tail	0.452676255	
F Critical one-tail	5.351128861	

Hypothesis: $H_0: \sigma_1 = \sigma_2$ $\quad VS \quad$ $H_1: \sigma_1 \neq \sigma_2$

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. The variances of the two populations are equal (p-value=0.4527$\nless\alpha$=0.01).

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

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

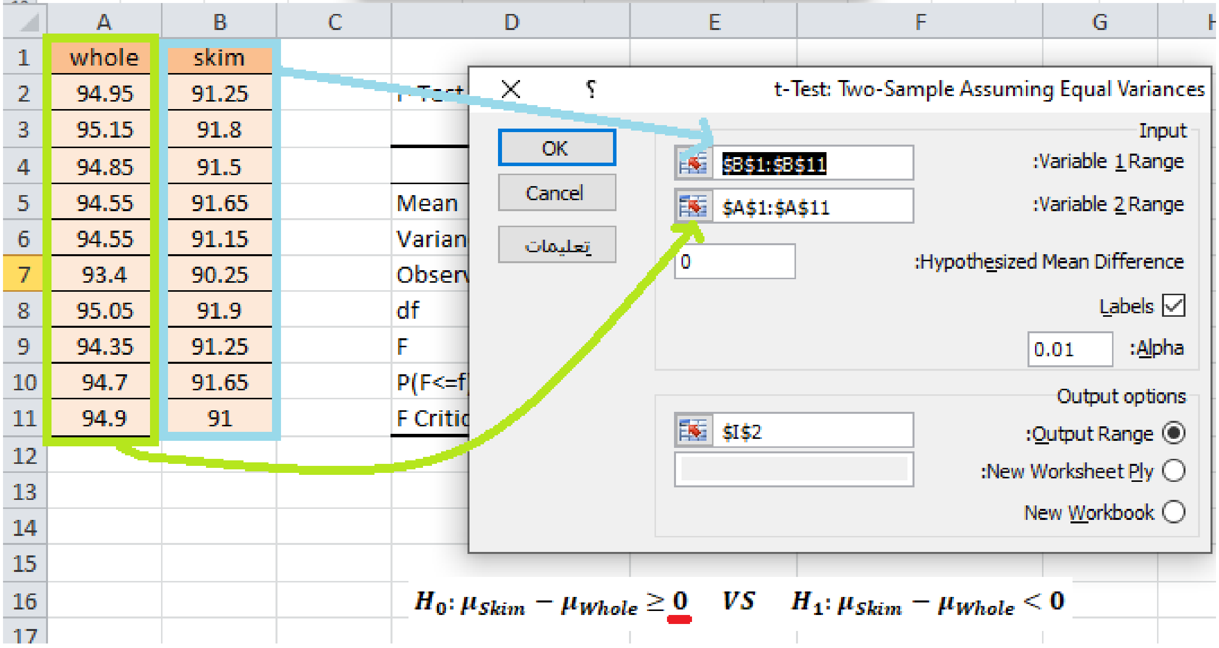


	A	B
1	whole	skim
2	94.95	91.25
3	95.15	91.8
4	94.85	91.5
5	94.55	91.65
6	94.55	91.15
7	93.4	90.25
8	95.05	91.9
9	94.35	91.25
10	94.7	91.65
11	94.9	91

F-Test Two-Sample for Variances

Mean
Variance
Observa
df
F
P(F<=f) o
F Critical

Selected: t-Test: Two-Sample Assuming Equal Variances



Variable 1 Range: $B$1:$B$11

Variable 2 Range: $A$1:$A$11

Hypothesized Mean Difference: 0

Labels ☑

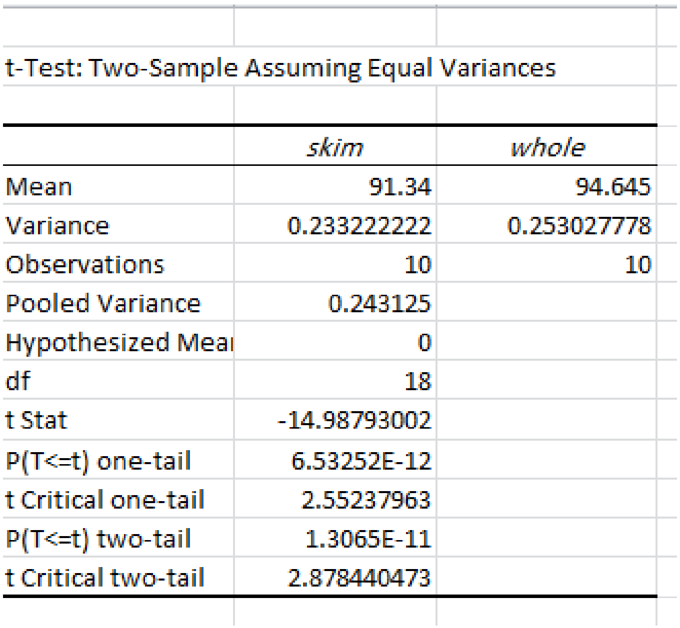
Alpha: 0.01

Output Range: ● $I$2

New Worksheet Ply: ○

New Workbook ○

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



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 Meai	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 one tail= 2.55238

4- Conclusion:

We do a one-tail 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

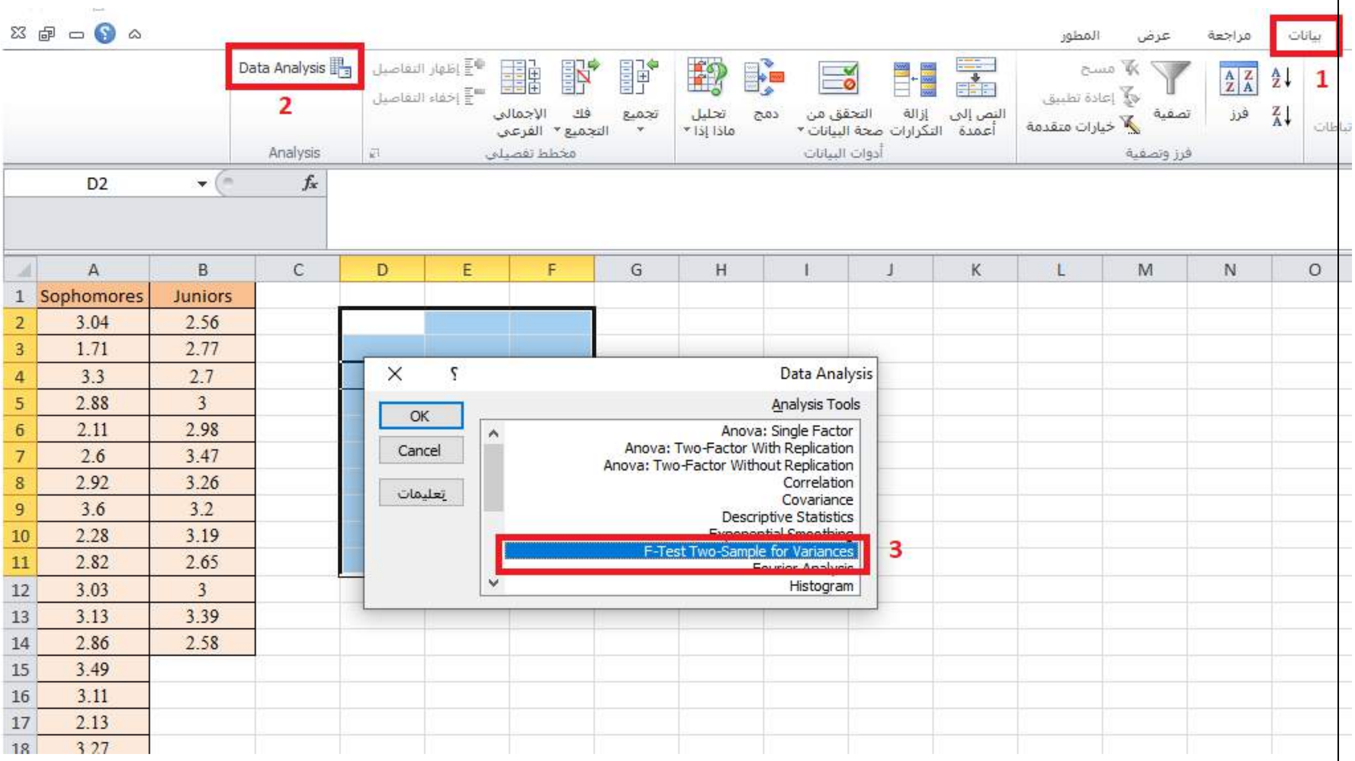
Q : Independent random samples of 17 sophomores and 13 juniors attending a large university yield the following data on grade point averages

sophomores			juniors		
3.04	2.92	2.86	2.56	3.47	2.65
1.71	3.60	3.49	2.77	3.26	3.00
3.30	2.28	3.11	2.70	3.20	3.39
2.88	2.82	2.13	3.00	3.19	2.58
2.11	3.03	3.27	2.98		
2.60	3.13				

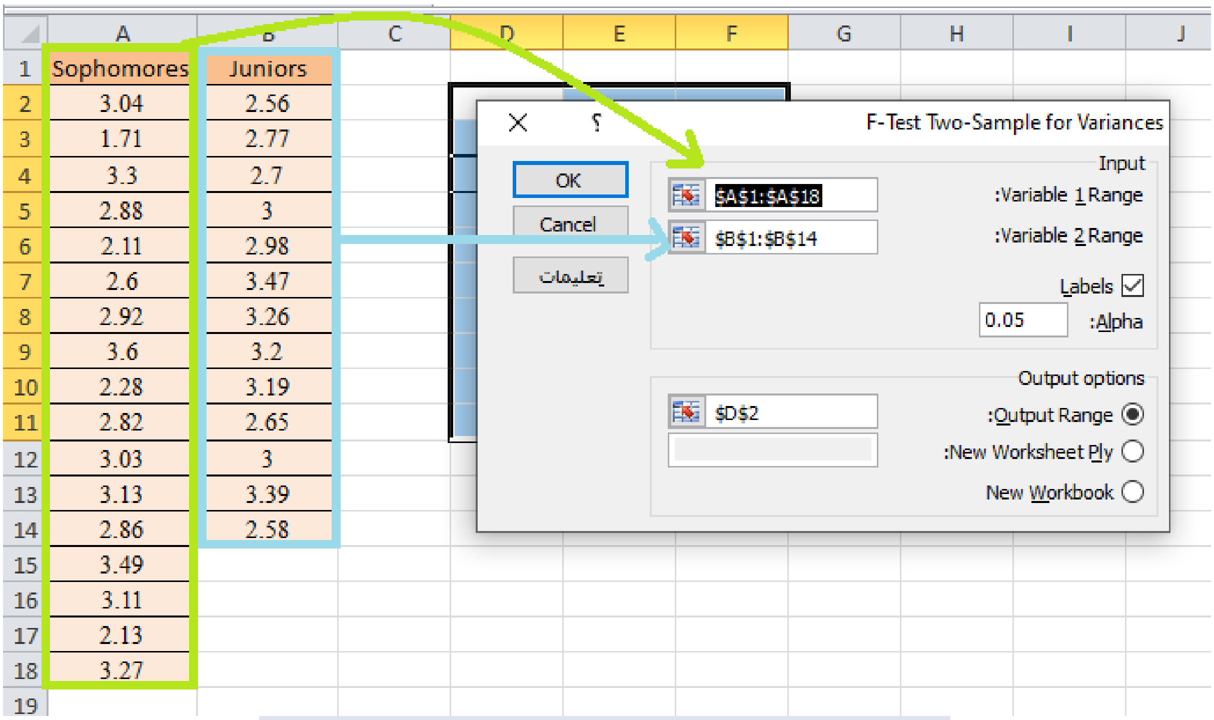
Assuming normal population. At the 5% significance level, do the data provide sufficient evidence to conclude that the mean GPAs of sophomores and juniors at the university different ?

1-Test for equality of variance :

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



	A	B
1	Sophomores	Juniors
2	3.04	2.56
3	1.71	2.77
4	3.3	2.7
5	2.88	3
6	2.11	2.98
7	2.6	3.47
8	2.92	3.26
9	3.6	3.2
10	2.28	3.19
11	2.82	2.65
12	3.03	3
13	3.13	3.39
14	2.86	2.58
15	3.49	
16	3.11	
17	2.13	
18	3.27	



	Sophomores	Juniors
1	Sophomores	Juniors
2	3.04	2.56
3	1.71	2.77
4	3.3	2.7
5	2.88	3
6	2.11	2.98
7	2.6	3.47
8	2.92	3.26
9	3.6	3.2
10	2.28	3.19
11	2.82	2.65
12	3.03	3
13	3.13	3.39
14	2.86	2.58
15	3.49	
16	3.11	
17	2.13	
18	3.27	

F-Test Two-Sample for Variances

Input

:Variable 1 Range $A$1:$A$18

:Variable 2 Range $B$1:$B$14

Labels ☑

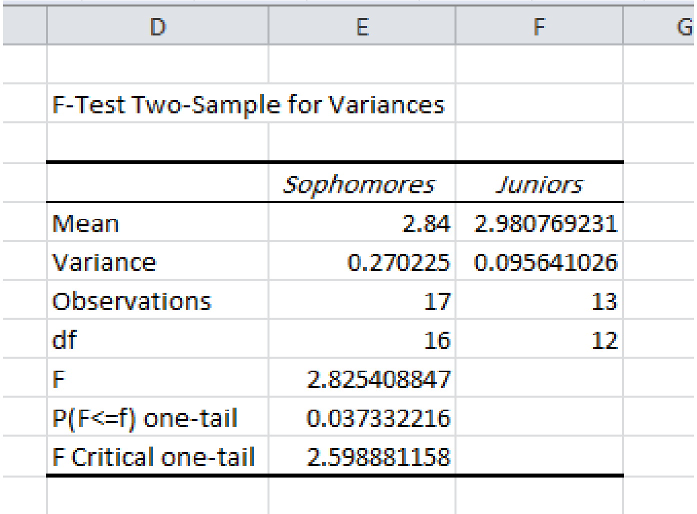
:Alpha 0.05

Output options

:Output Range ● $D$2

:New Worksheet Ply ○

New Workbook ○



F-Test Two-Sample for Variances

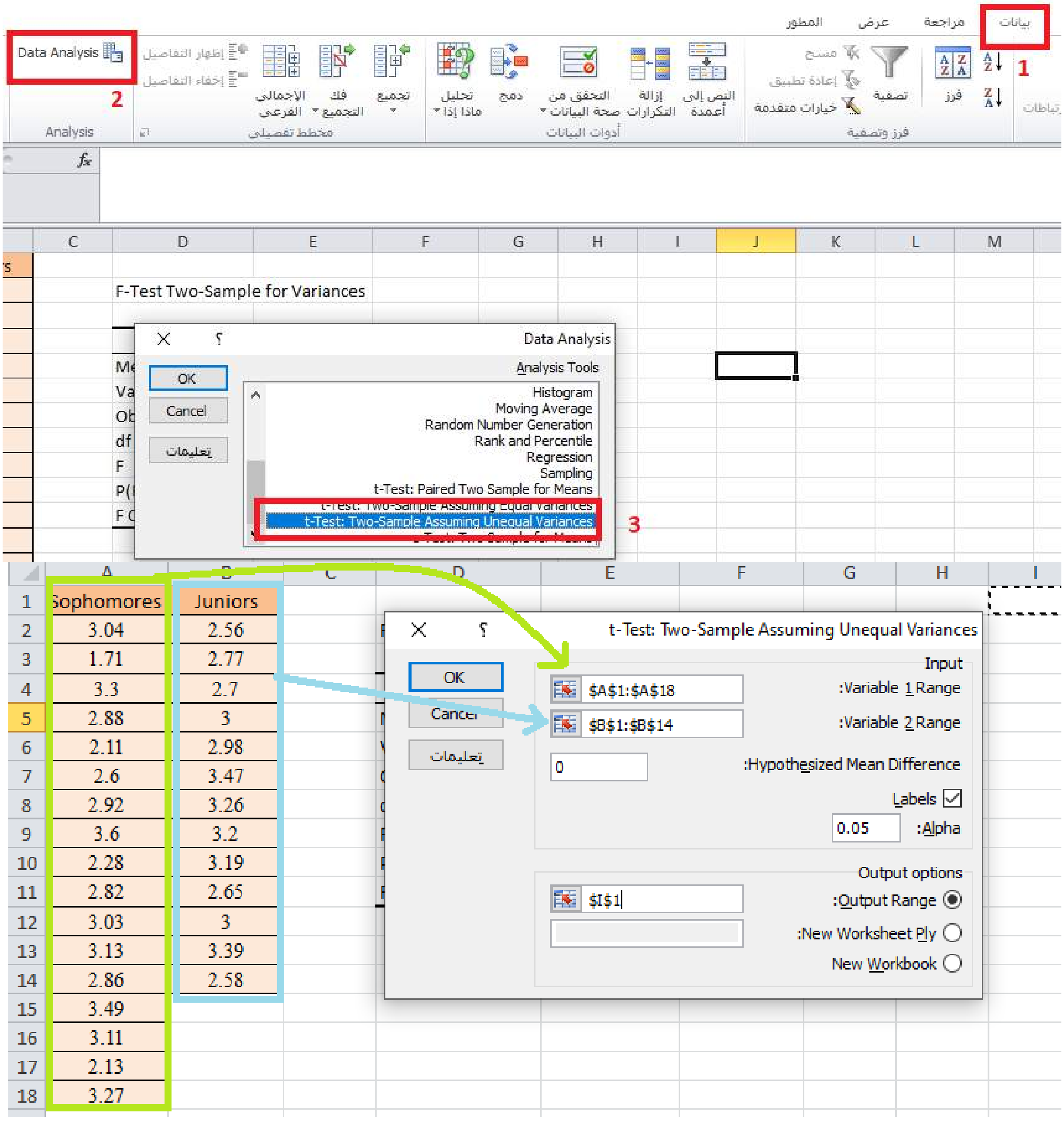
	Sophomores	*Juniors*
Mean	2.84	2.980769231
Variance	0.270225	0.095641026
Observations	17	13
df	16	12
F	2.825408847	
P(F<=f) one-tail	0.037332216	
F Critical one-tail	2.598881158	

Hypothesis: $H_0: \sigma_1 = \sigma_2$ $\quad VS \quad$ $H_1: \sigma_1 \neq \sigma_2$

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 (p-value=0.03733<α=0.05)

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

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



	A	B
1	Sophomores	Juniors
2	3.04	2.56
3	1.71	2.77
4	3.3	2.7
5	2.88	3
6	2.11	2.98
7	2.6	3.47
8	2.92	3.26
9	3.6	3.2
10	2.28	3.19
11	2.82	2.65
12	3.03	3
13	3.13	3.39
14	2.86	2.58
15	3.49	
16	3.11	
17	2.13	
18	3.27	

t-Test: Two-Sample Assuming Unequal Variances

Input

:Variable 1 Range $A$1:$A$18

:Variable 2 Range $B$1:$B$14

:Hypothesized Mean Difference 0

Labels ☑

:Alpha 0.05

Output options

:Output Range ● $I$1

:New Worksheet Ply ○

New Workbook ○

1-<u>Hypothesis:</u>

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

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

2- <u>Test statistic :</u> T= - 0.9231

3- <u>T critical two tail</u>= 2.05183

4- <u>Conclusion:</u>

We do a two-tail test (inequality). If t Stat < -t Critical two-tail or t Stat > t Critical two-tail, we reject the null hypothesis. This is not the case, -2.05183< -0.9231 < 2.05183. Therefore, we do not reject the null hypothesis (p-value=0.3641$\nless \alpha$=0.05)

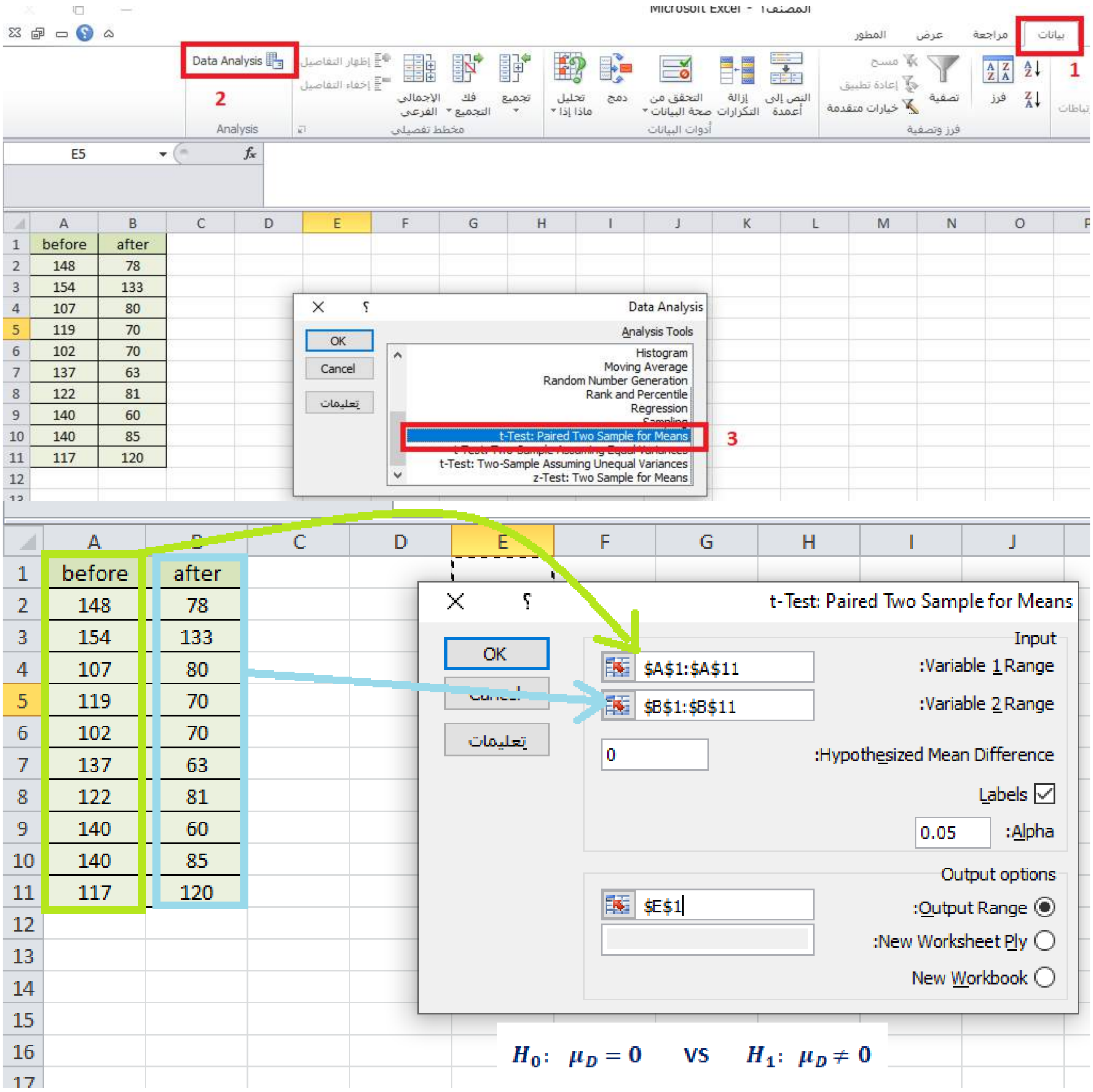
Q : In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

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 <u>data comes from normal distribution</u>. 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



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

t-Test: Paired Two Sample for Means

Input

:Variable 1 Range $A$1:$A$11

:Variable 2 Range $B$1:$B$11

:Hypothesized Mean Difference 0

Labels ☑

:Alpha 0.05

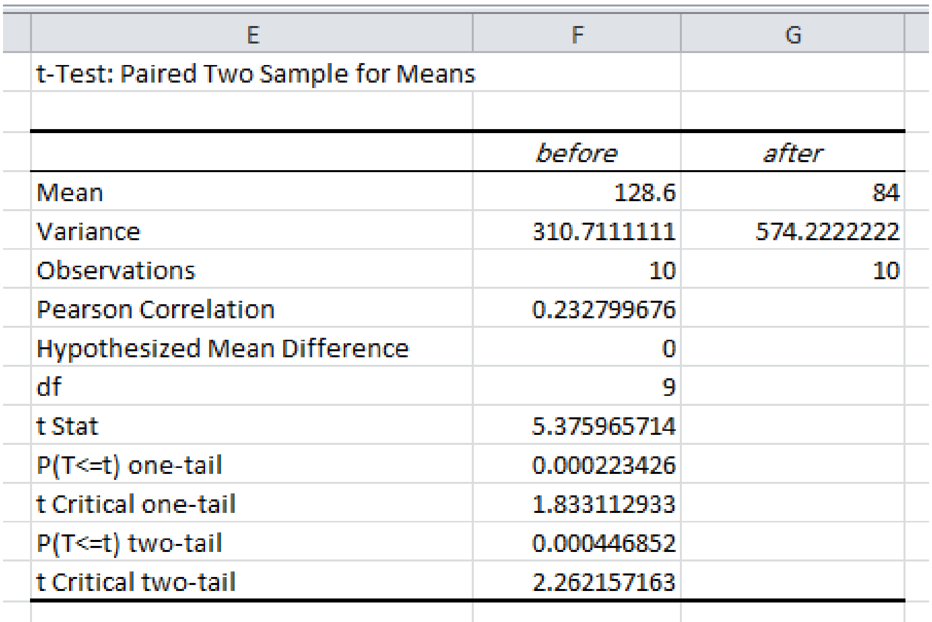
Output options

:Output Range ● $E$1

:New Worksheet Ply ○

New Workbook ○

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



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 two tail= 2.26215

4- Conclusion:

We do a two-tail 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

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.

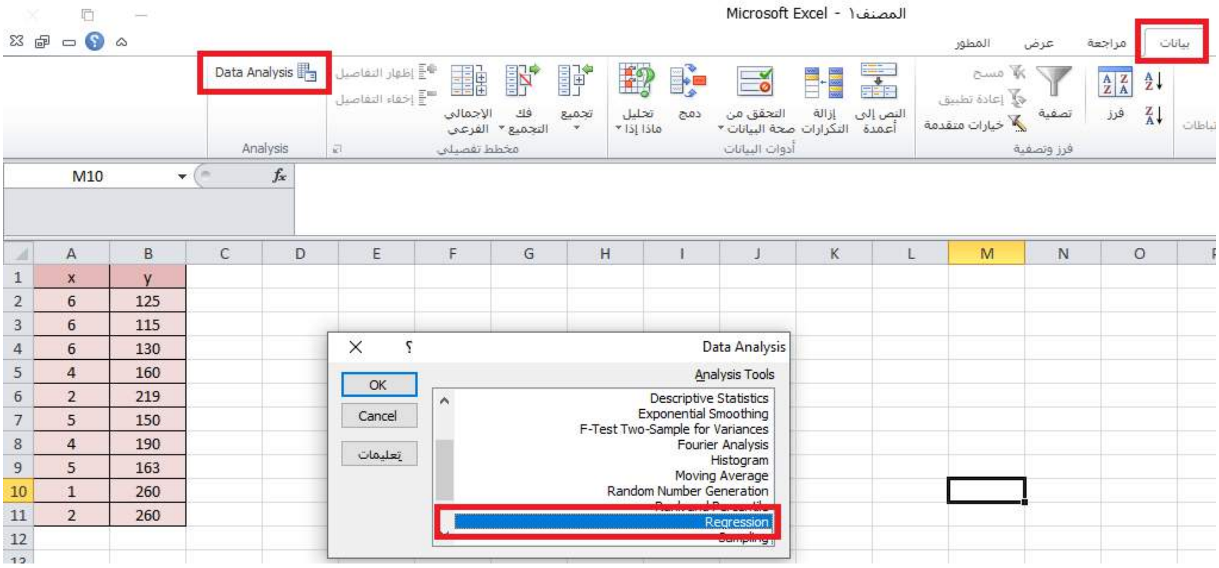
X	6	6	6	4	2	5	4	5	1	2
y	125	115	130	160	219	150	190	163	260	260

a) Determine the regression equation for the data.

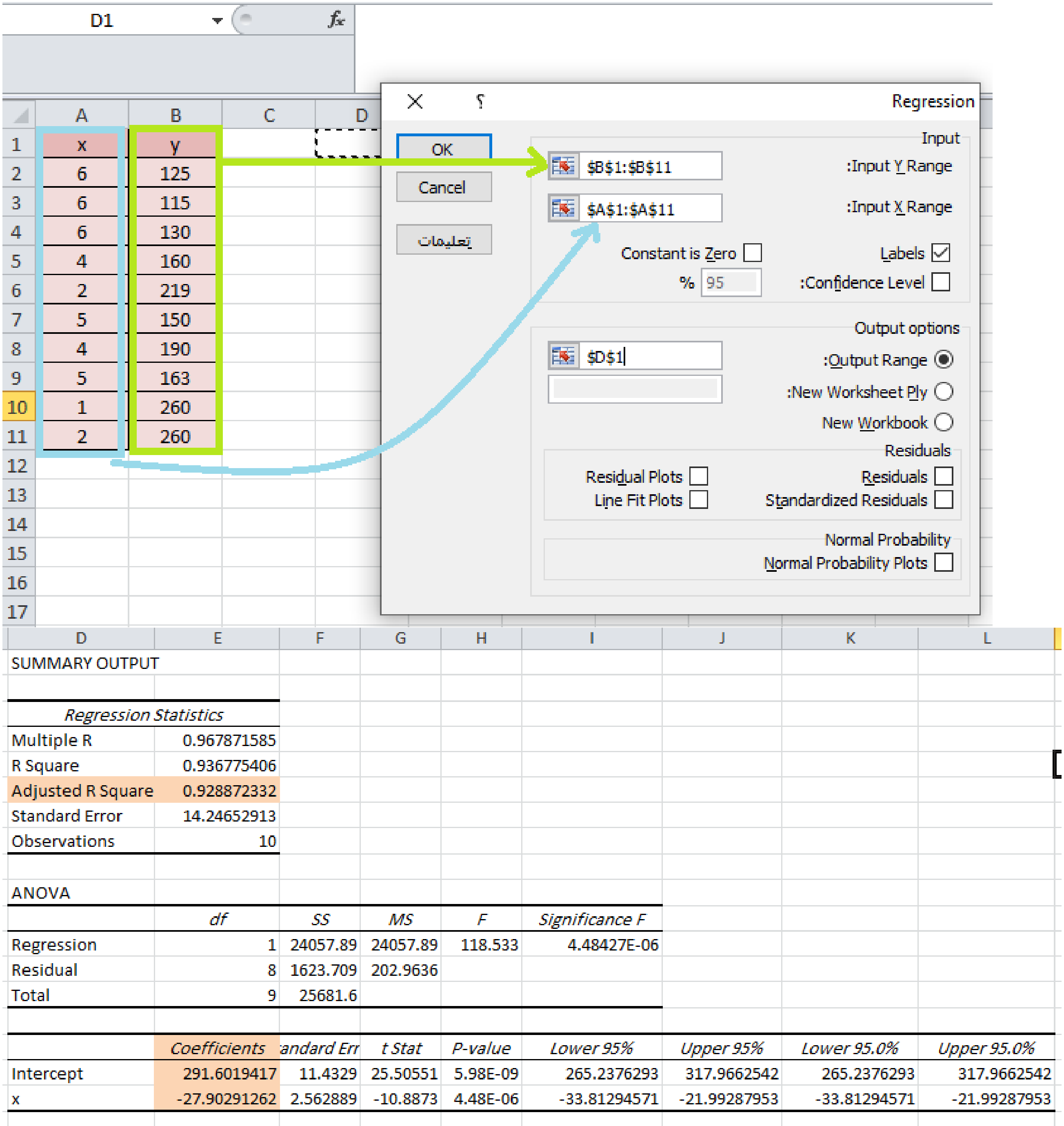
b) Compute and interpret the coefficient of determination, r^2.

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

Data → Data Analysis → Regression



	A	B
1	x	y
2	6	125
3	6	115
4	6	130
5	4	160
6	2	219
7	5	150
8	4	190
9	5	163
10	1	260
11	2	260



	A	B
1	x	y
2	6	125
3	6	115
4	6	130
5	4	160
6	2	219
7	5	150
8	4	190
9	5	163
10	1	260
11	2	260

Input

:Input Y Range $B$1:$B$11

:Input X Range $A$1:$A$11

Labels ☑ Constant is Zero ☐

:Confidence Level ☐ 95 %

Output options

:Output Range ● $D$1

:New Worksheet Ply ○

New Workbook ○

Residuals

Residuals ☐ Residual Plots ☐

Standardized Residuals ☐ Line Fit Plots ☐

Normal Probability

Normal Probability Plots ☐

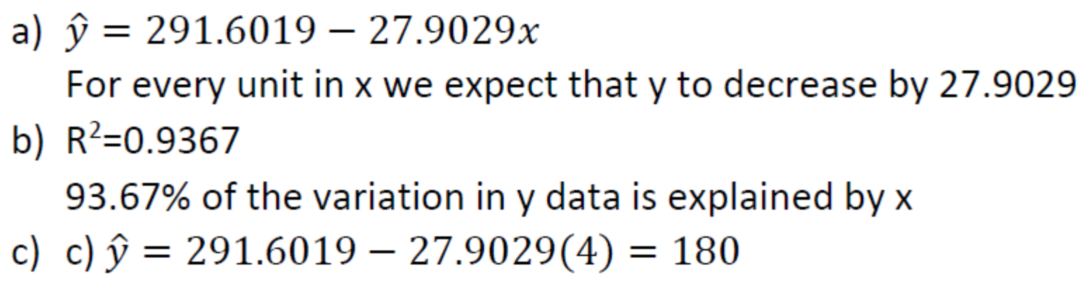
SUMMARY OUTPUT

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	24057.89	24057.89	118.533	4.48427E-06
Residual	8	1623.709	202.9636		
Total	9	25681.6			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	291.6019417	11.4329	25.50551	5.98E-09	265.2376293	317.9662542	265.2376293	317.9662542
x	-27.90291262	2.562889	-10.8873	4.48E-06	-33.81294571	-21.99287953	-33.81294571	-21.99287953



a) $\hat{y} = 291.6019 - 27.9029x$

For every unit in x we expect that y to decrease by 27.9029

b) R^2=0.9367

93.67% of the variation in y data is explained by x

c) c) $\hat{y} = 291.6019 - 27.9029(4) = 180$

Results:

The regression line is: y = sales price = 291.6019 - 27.9029 * age .

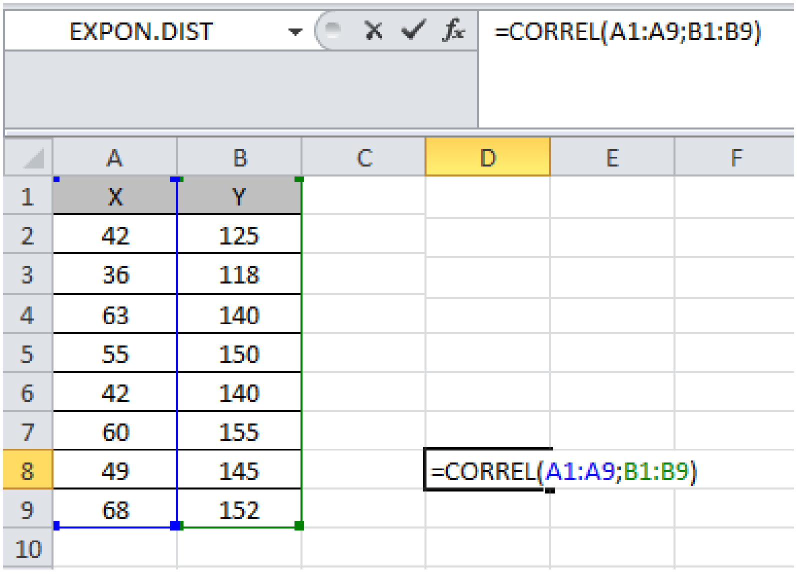
In other words, for increasing the age by one, the sales price decreasing by 27.9029 , while there is 291.6019 minutes does not depend on the age .

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:

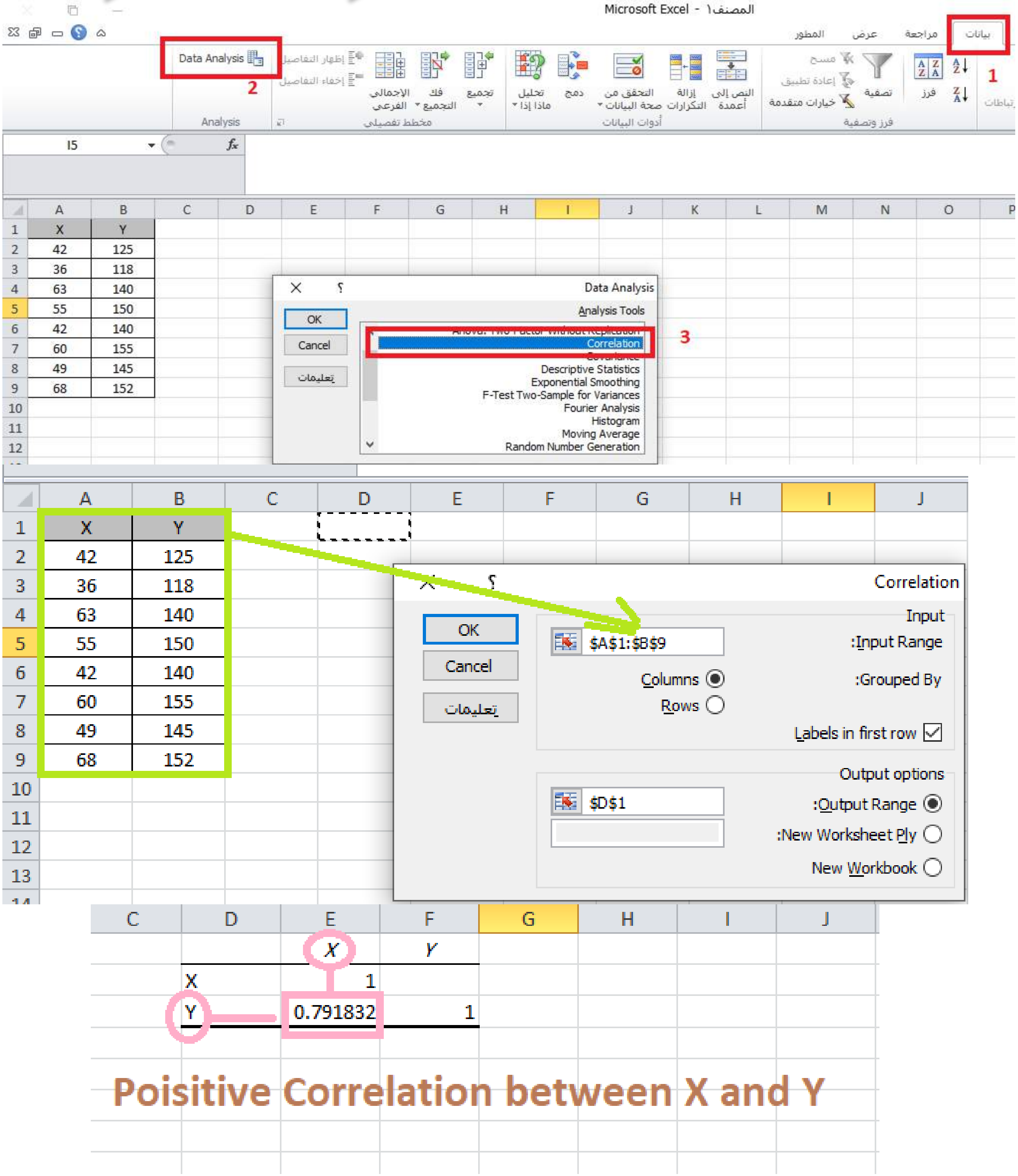
	By Excel (using (*fx*) and (Data Analysis))
Correlation=0.791832	CORREL(M3:M10;N3:N10)



=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						

Data → Data Analysis → Correlation



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

Input Range: $A$1:$B$9

Grouped By: Columns ● Rows ○

Labels in first row ☑

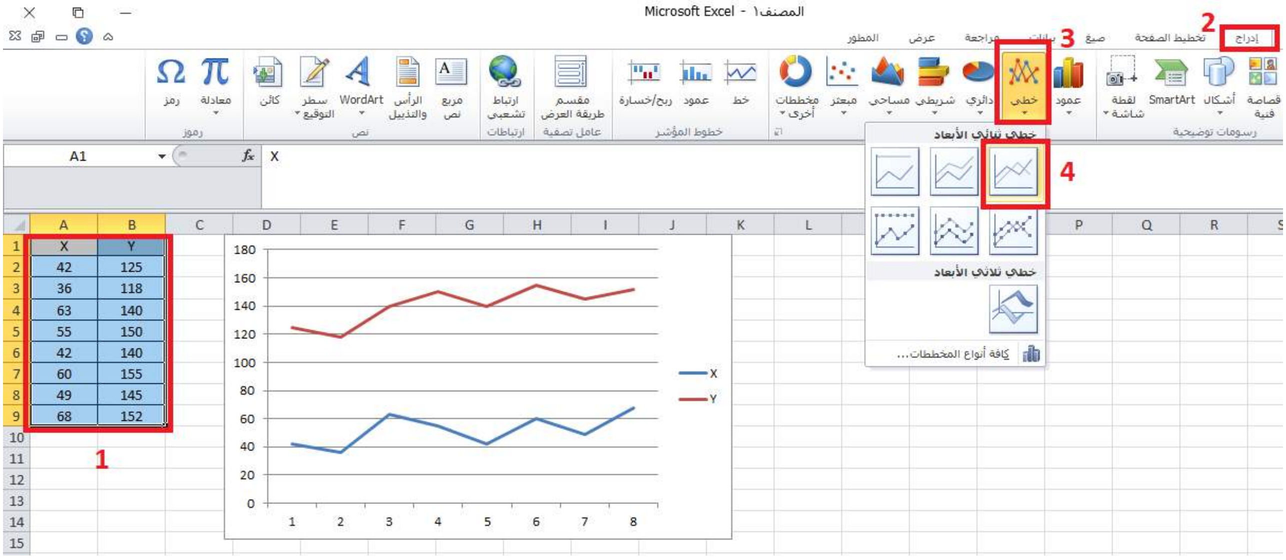
Output Range: ● $D$1

New Worksheet Ply: ○

New Workbook ○

	X	Y
X	1	
Y	0.791832	1

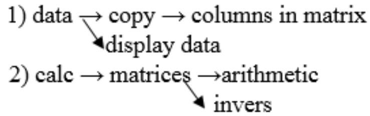
Poisitive Correlation between X and Y



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

MATRICES

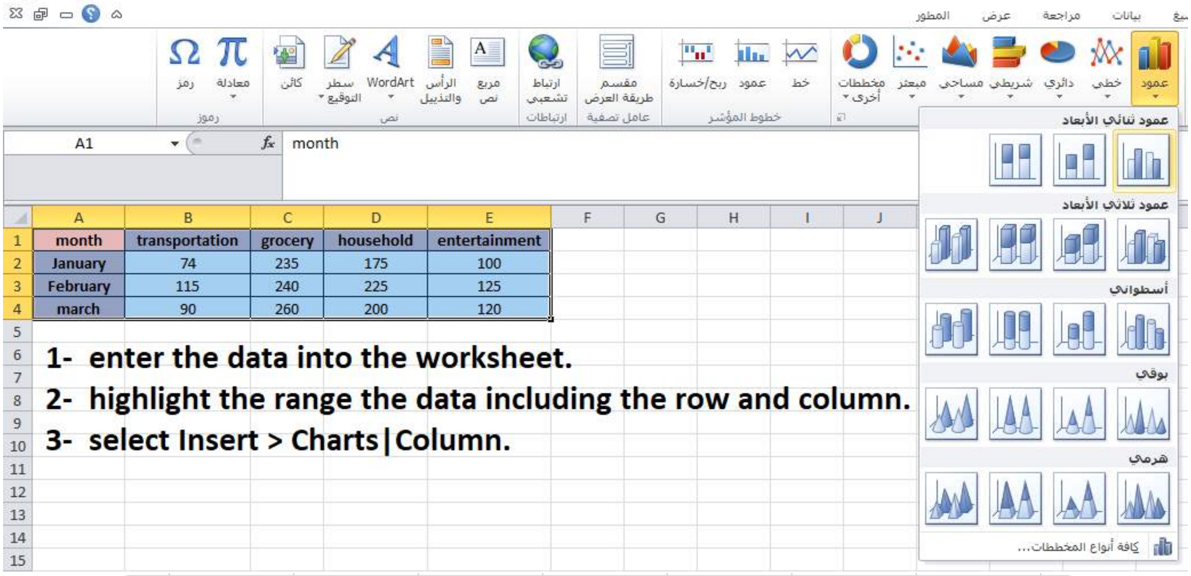
Write the commands of the following:

		By Excel **(**using (*fx*)**)**	**By Minitab** 1) data → copy → columns in matrix ↘ display data 2) calc → matrices → arithmetic ↘ invers •The name of matrices in columns in matrix keeps their names + ✕ Names of matrix containing.... •The name of new matrices in arithmetic and invers is (M#).
Addition of Matrices	$A=\begin{bmatrix}-5&0\\4&1\end{bmatrix},B=\begin{bmatrix}6&-3\\2&3\end{bmatrix}$ $\Rightarrow A+B=\begin{bmatrix}-5+6&0+-3\\4+2&1+3\end{bmatrix}=\begin{bmatrix}1&-3\\6&4\end{bmatrix}$		
Subtract of Matrices	$C=\begin{bmatrix}1&2\\-2&0\\-3&-1\end{bmatrix},D=\begin{bmatrix}1&-1\\1&3\\2&3\end{bmatrix}$ $\Rightarrow C-D=\begin{bmatrix}1-1&2-(-1)\\-2-1&0-3\\-3-2&-1-3\end{bmatrix}=\begin{bmatrix}0&3\\-3&-3\\-5&-4\end{bmatrix}$		
Additive Inverse of Matrix	$A=\begin{bmatrix}1&0&2\\3&-1&5\end{bmatrix}$ $\Rightarrow -A=\begin{bmatrix}-1&0&-2\\-3&1&-5\end{bmatrix}$		
Scalar Multiplication of Matrices	$D=\begin{bmatrix}-3&0\\4&5\end{bmatrix}$ $\Rightarrow 3D=\begin{bmatrix}-9&0\\12&15\end{bmatrix}$		
Matrix Multiplication	$E=\begin{bmatrix}1&4&7\\2&5&8\\3&6&9\end{bmatrix},F=\begin{bmatrix}1&4\\2&5\\3&6\end{bmatrix}$ $\Rightarrow E\times F=\begin{bmatrix}30&66\\36&81\\42&96\end{bmatrix}$		
Determinant and Inverse Matrices	$G=\begin{bmatrix}3&-1\\-5&2\end{bmatrix}$ $\Rightarrow \det(G)=1 \text{ and } G^{-1}=\begin{bmatrix}2&1\\5&3\end{bmatrix}$		

The following data represents the expenses in dollars by month :

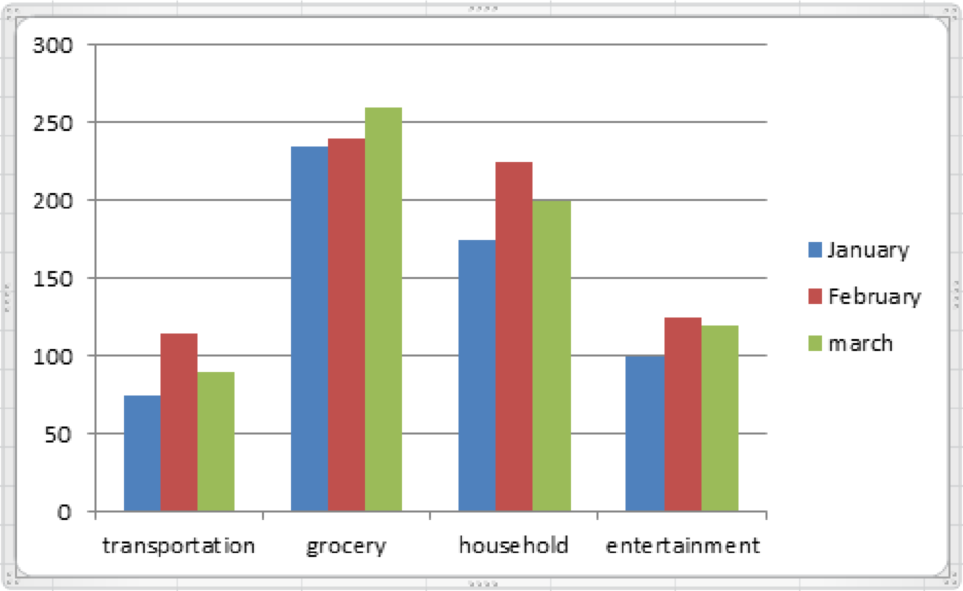
month	transportation	grocery	household	entertainment
January	74	235	175	100
February	115	240	225	125
march	90	260	200	120

a- Bar chart :

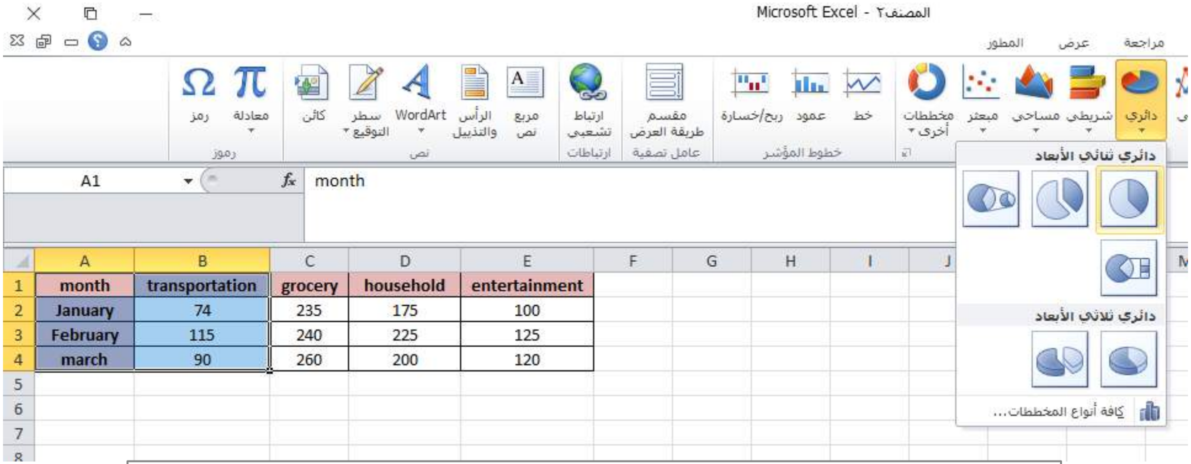


	A	B	C	D	E
1	month	transportation	grocery	household	entertainment
2	January	74	235	175	100
3	February	115	240	225	125
4	march	90	260	200	120

1- enter the data into the worksheet.
2- highlight the range the data including the row and column.
3- select Insert > Charts|Column.

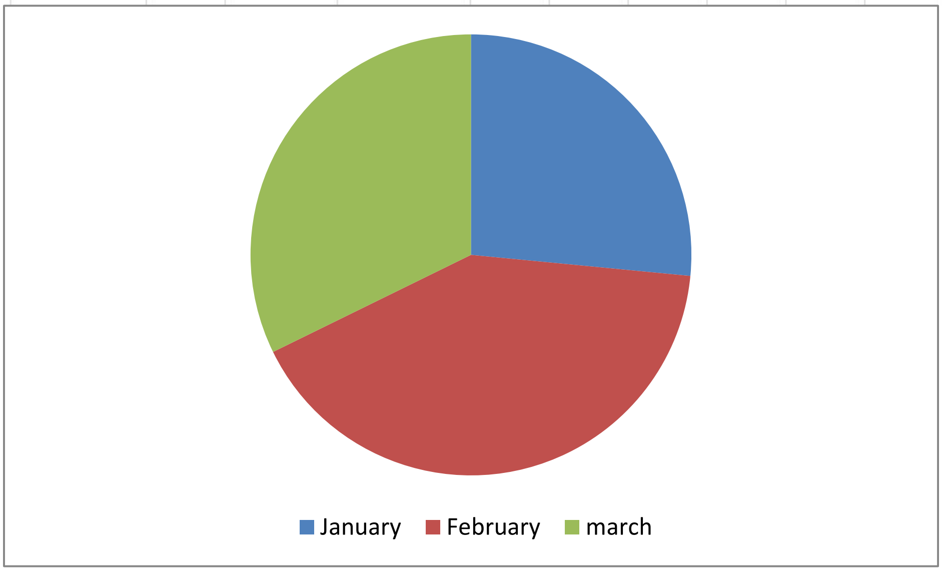


b- Pie chart : (For the previous example draw the pie chart for transportation)

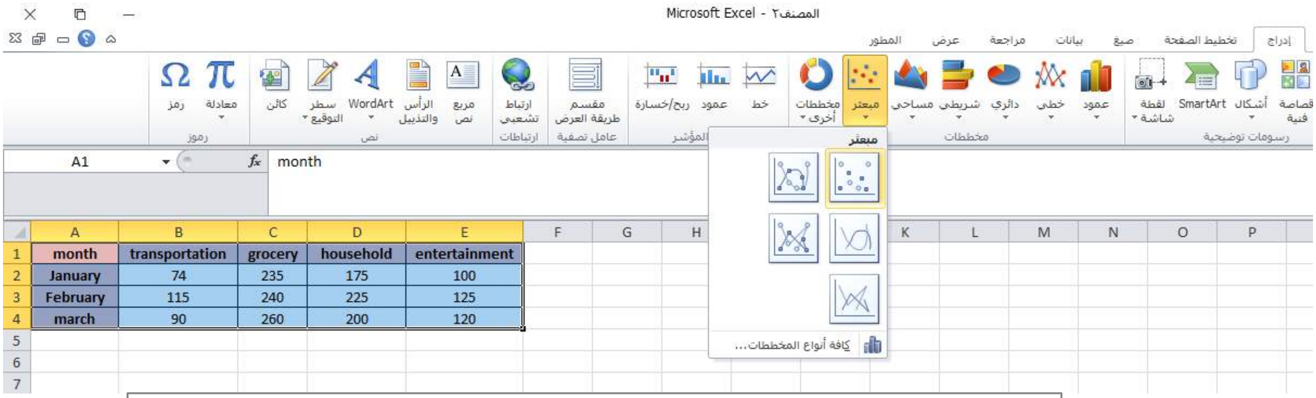


A1 | f_x month

	A	B	C	D	E
1	month	transportation	grocery	household	entertainment
2	January	74	235	175	100
3	February	115	240	225	125
4	march	90	260	200	120

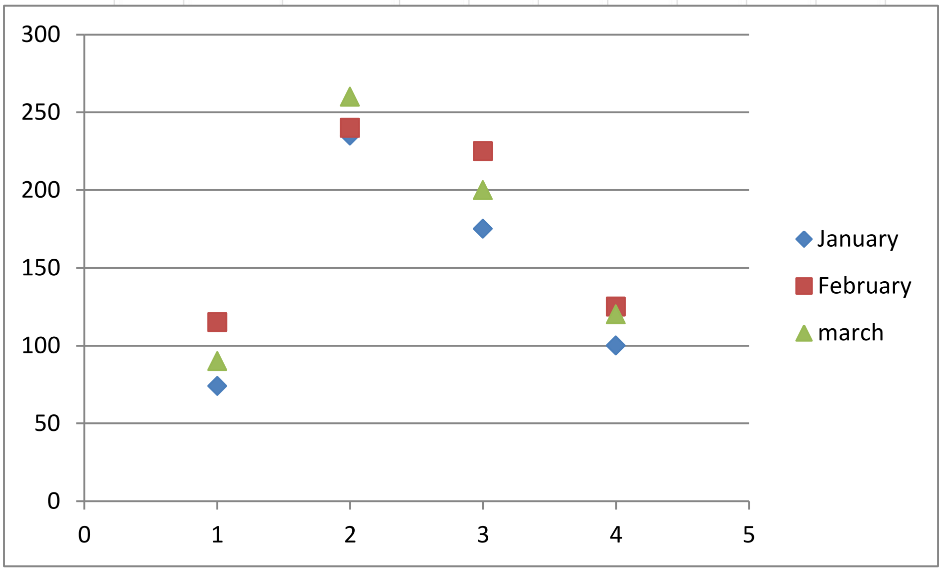


b- Scatter plot :



A1 | month

	A	B	C	D	E
1	month	transportation	grocery	household	entertainment
2	January	74	235	175	100
3	February	115	240	225	125
4	march	90	260	200	120



Histogram :

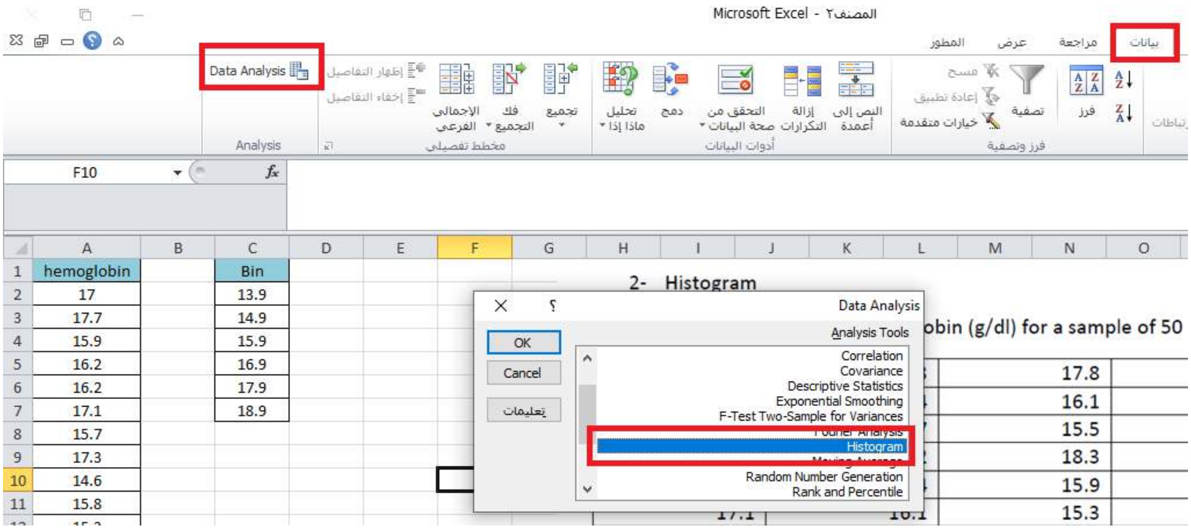
The following data represent hemoglobin (g/dl) for a sample of 50 women :

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

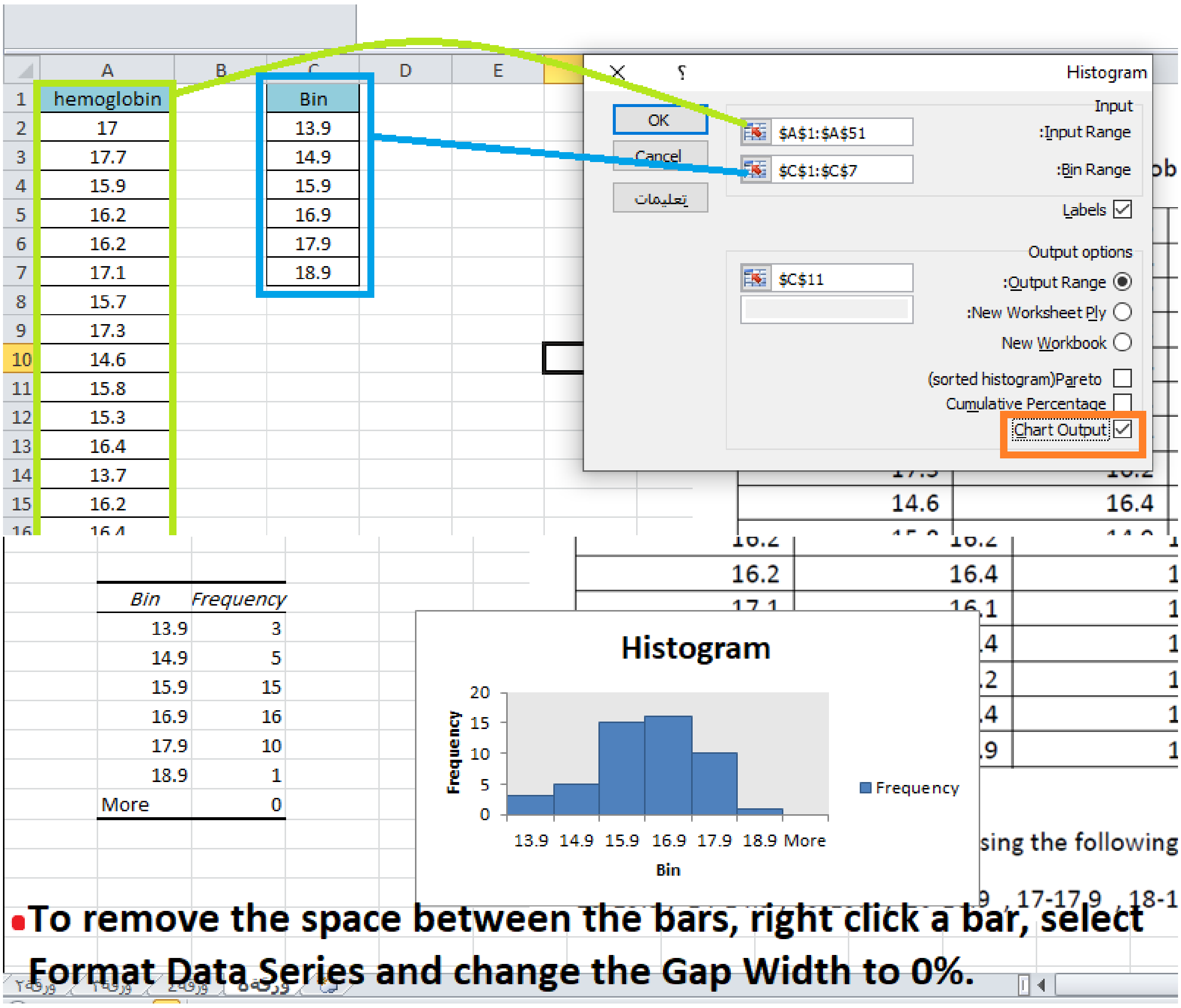
We wish to summarize these data using the following class intervals

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

Data → Data Analysis → Histogram



	A	B	C
1	hemoglobin		Bin
2	17		13.9
3	17.7		14.9
4	15.9		15.9
5	16.2		16.9
6	16.2		17.9
7	17.1		18.9
8	15.7		
9	17.3		
10	14.6		
11	15.8		



	A	B	C
1	hemoglobin		Bin
2	17		13.9
3	17.7		14.9
4	15.9		15.9
5	16.2		16.9
6	16.2		17.9
7	17.1		18.9
8	15.7		
9	17.3		
10	14.6		
11	15.8		
12	15.3		
13	16.4		
14	13.7		
15	16.2		

Input
:Input Range $A$1:$A$51
:Bin Range $C$1:$C$7
Labels ☑

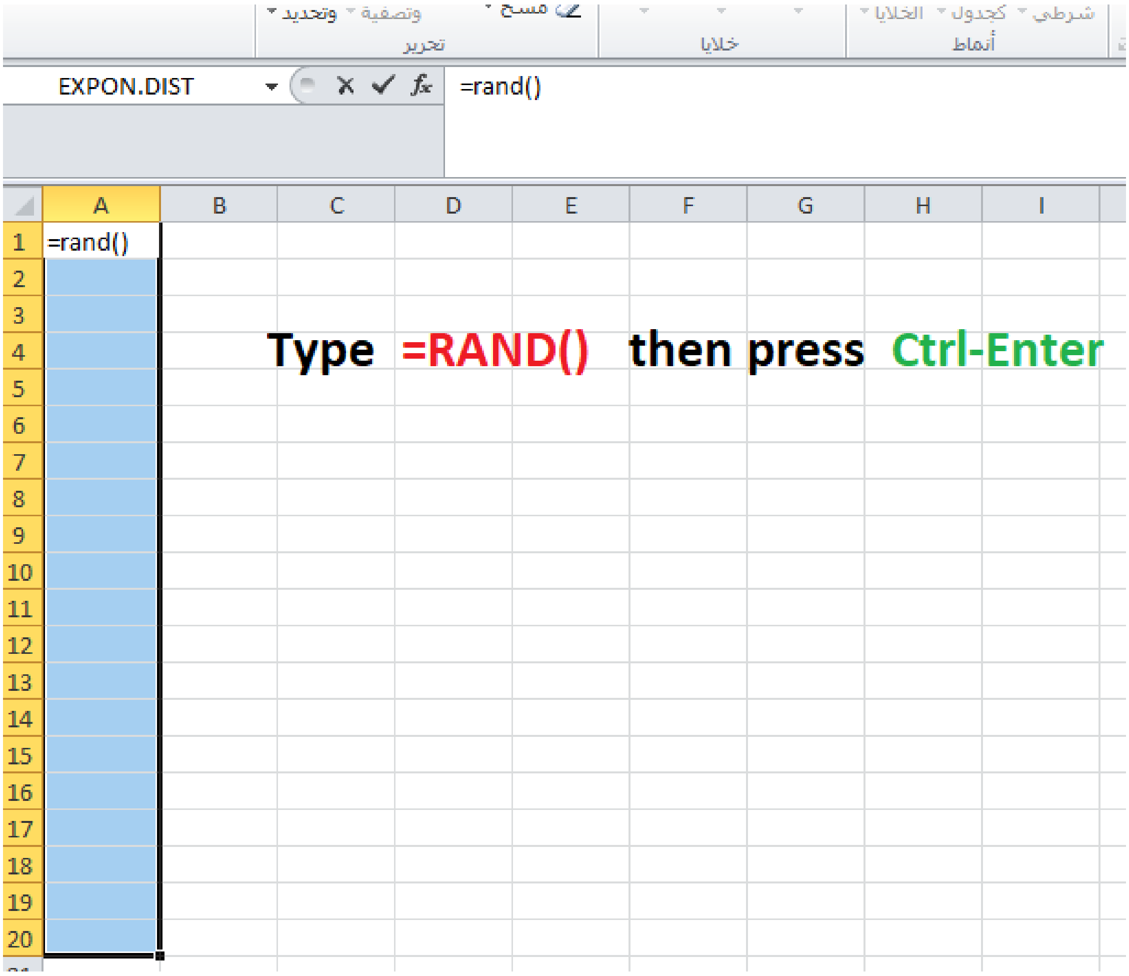
Output options
:Output Range ● $C$11
:New Worksheet Ply ○
New Workbook ○
(sorted histogram)Pareto ☐
Cumulative Percentage ☐
Chart Output ☑

Bin	*Frequency*
13.9	3
14.9	5
15.9	15
16.9	16
17.9	10
18.9	1
More	0

- **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

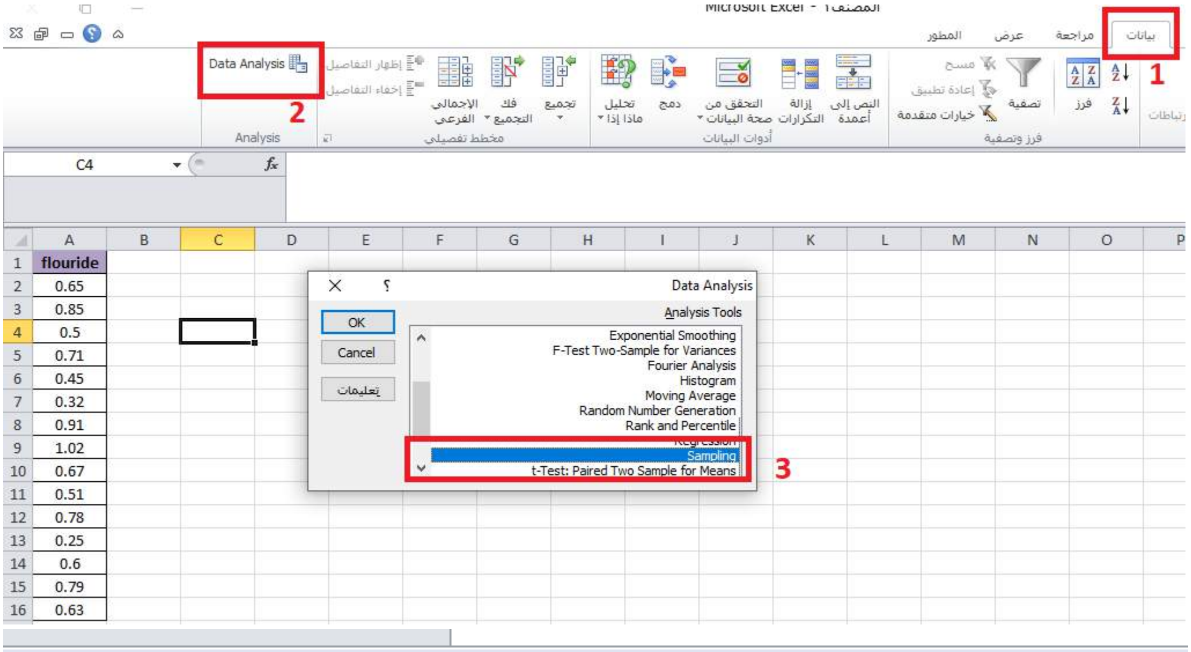


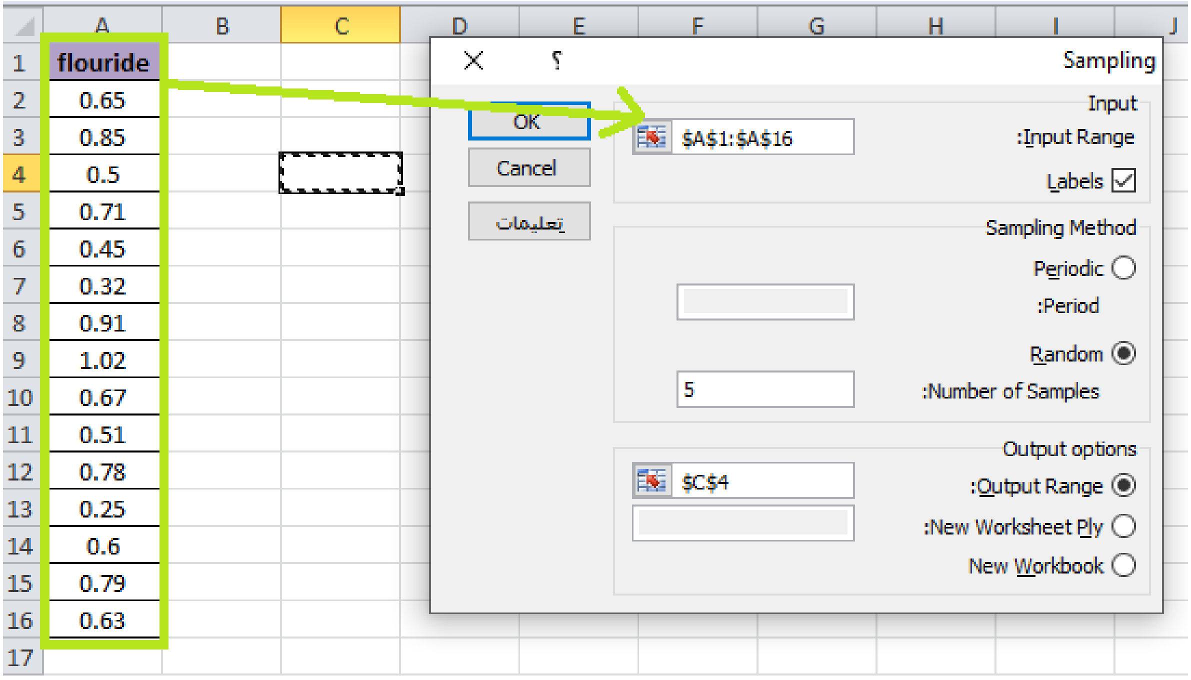
=rand()

Type =RAND() then press Ctrl-Enter

2- Sampling

Data → Data Analysis → sampling





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

Input Range: $A$1:$A$16

Labels ☑

Periodic ○

Period:

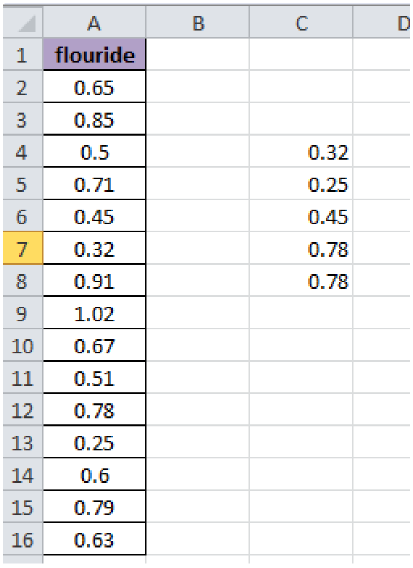
Random ●

Number of Samples: 5

Output Range: ● $C$4

New Worksheet Ply: ○

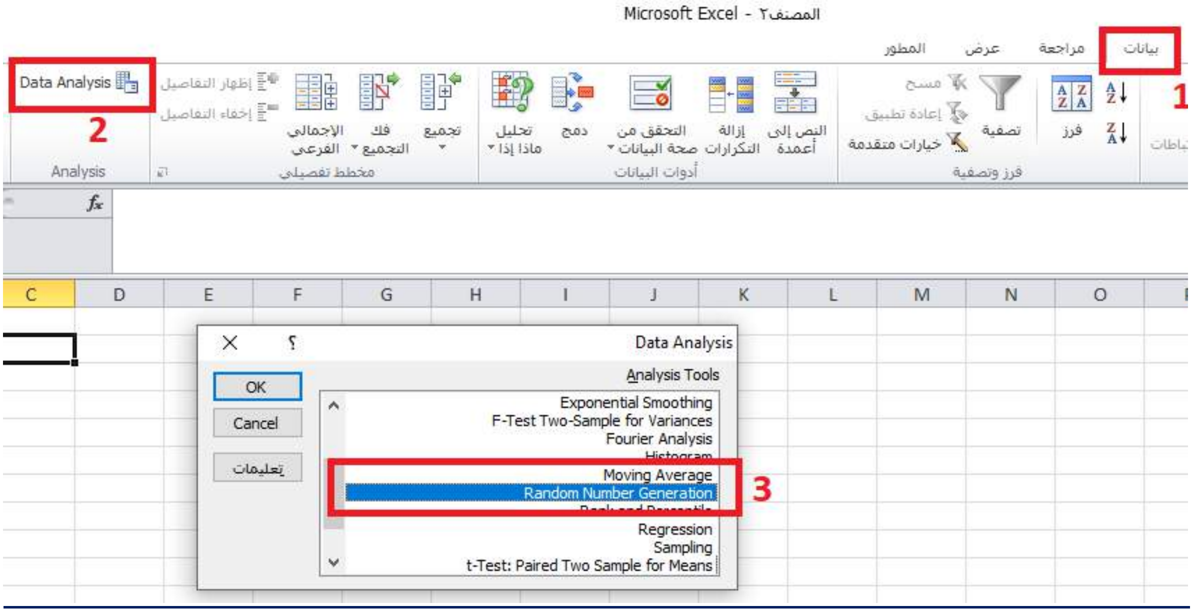
New Workbook ○



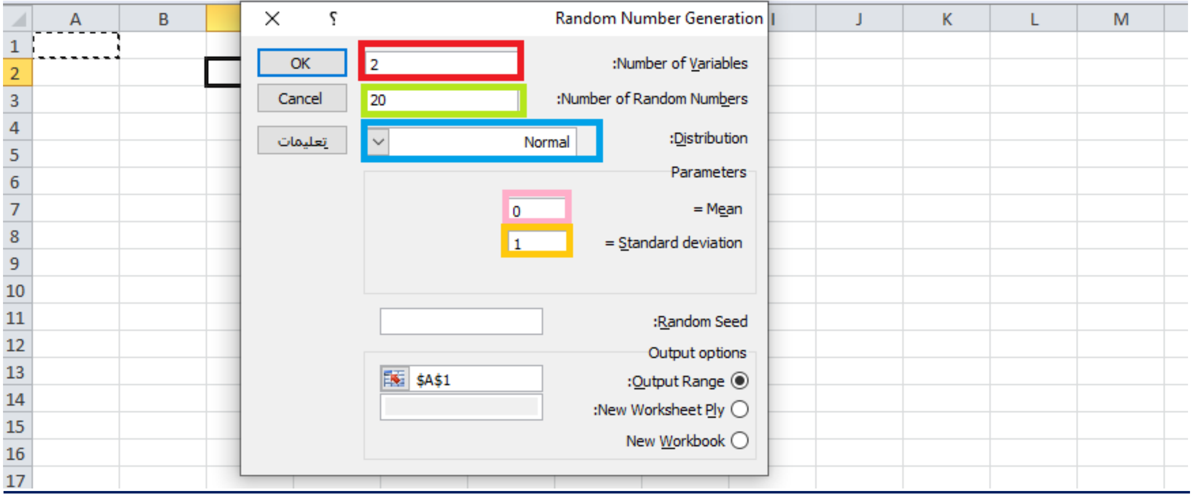
	A	B	C	D
1	**flouride**			
2	0.65			
3	0.85			
4	0.5		0.32	
5	0.71		0.25	
6	0.45		0.45	
7	0.32		0.78	
8	0.91		0.78	
9	1.02			
10	0.67			
11	0.51			
12	0.78			
13	0.25			
14	0.6			
15	0.79			
16	0.63			

3- Random number generation from distributions

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



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



Random Number Generation

Number of Variables: 2

Number of Random Numbers: 20

Distribution: Normal

Parameters

Mean = 0

Standard deviation = 1

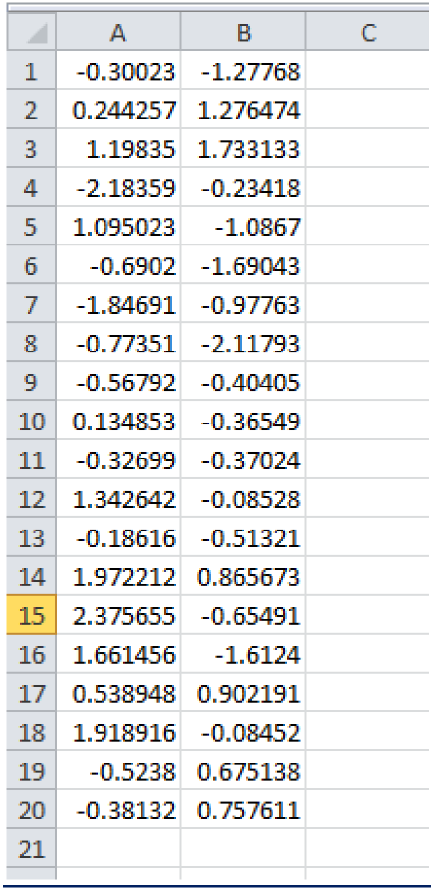
Random Seed:

Output options

● Output Range: $A$1

○ New Worksheet Ply:

○ New Workbook



	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			

Department of Statistics and Operations Research

College of Science

King Saud University

Tutorial

STATISTICAL PACKAGES(SPSS)

STAT 328

SPSS (tutorial 1)

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 :

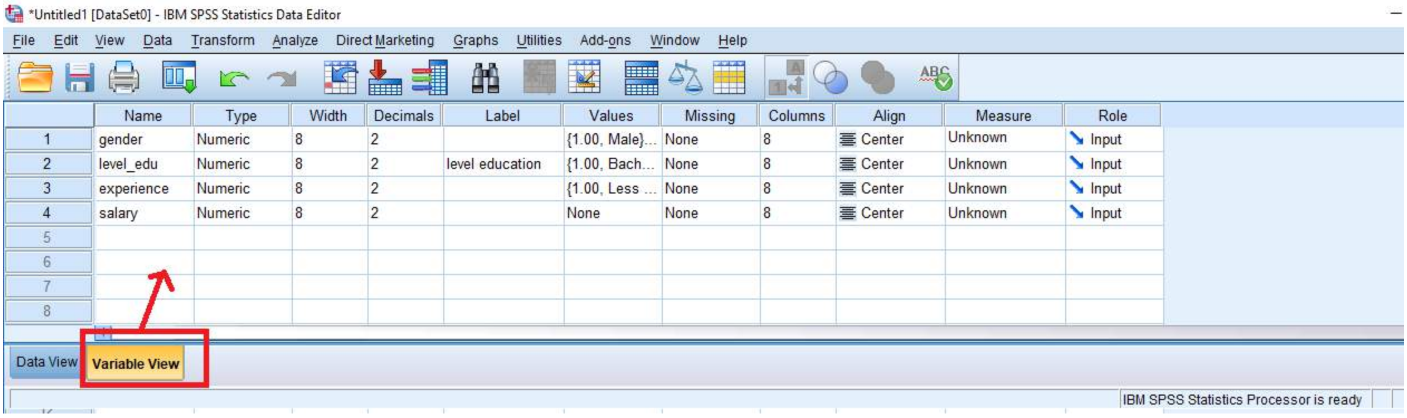
Gender: 1: Male 2: female :

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

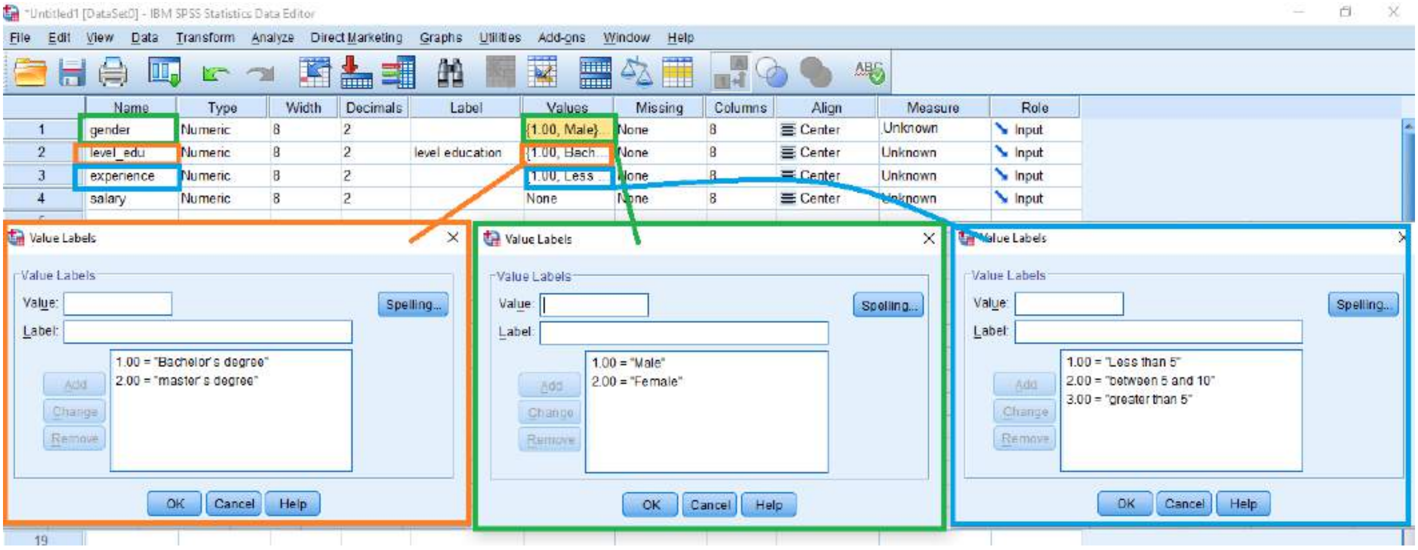
Experience 1: Less than 5 years 2: between 5 and 10 years 3: greater than 5 years

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



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	2		{1.00, Male}...	None	8	Center	Unknown	Input
2	level_edu	Numeric	8	2	level education	{1.00, Bach...	None	8	Center	Unknown	Input
3	experience	Numeric	8	2		{1.00, Less ...	None	8	Center	Unknown	Input
4	salary	Numeric	8	2		None	None	8	Center	Unknown	Input
5											
6											
7											
8											



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gender	Numeric	8	2		{1.00, Male}...	None	8	Center	Unknown	Input
2	level_edu	Numeric	8	2	level education	{1.00, Bach...	None	8	Center	Unknown	Input
3	experience	Numeric	8	2		{1.00, Less...	None	8	Center	Unknown	Input
4	salary	Numeric	8	2		None	None	8	Center	Unknown	Input

Value Labels

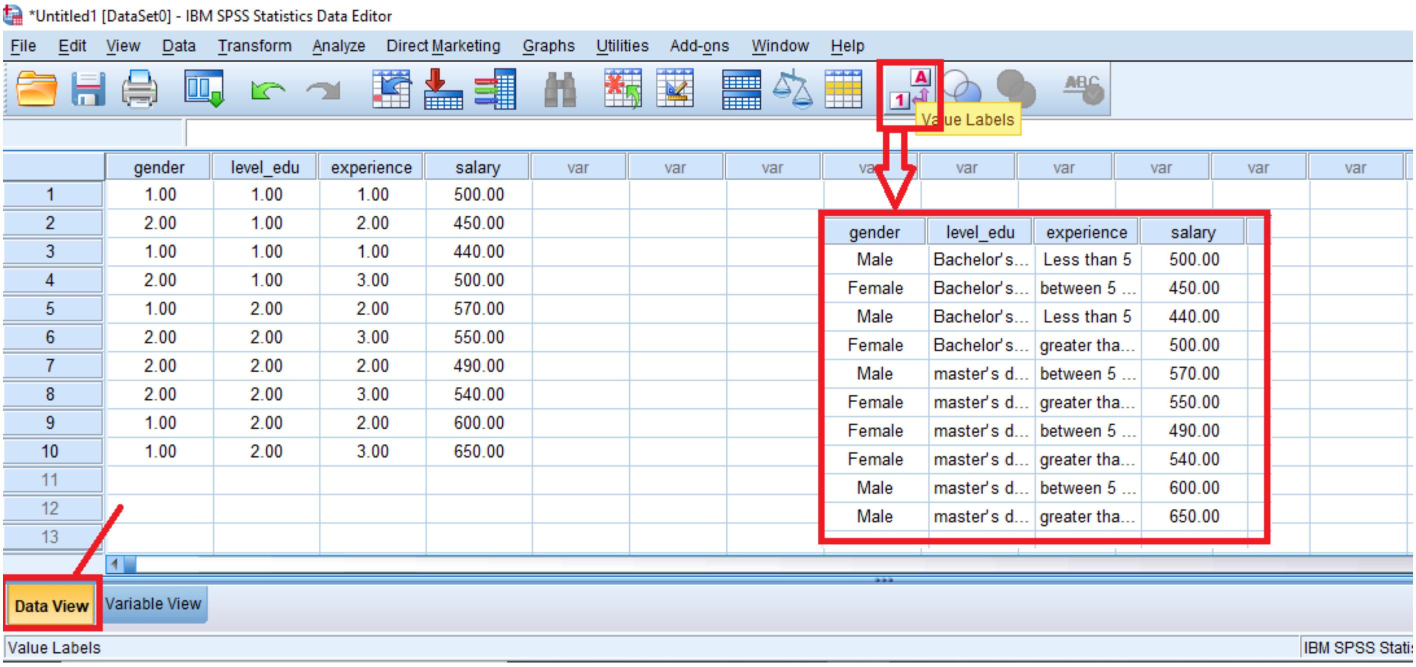
1.00 = "Bachelor's degree"
2.00 = "master's degree"

Value Labels

1.00 = "Male"
2.00 = "Female"

Value Labels

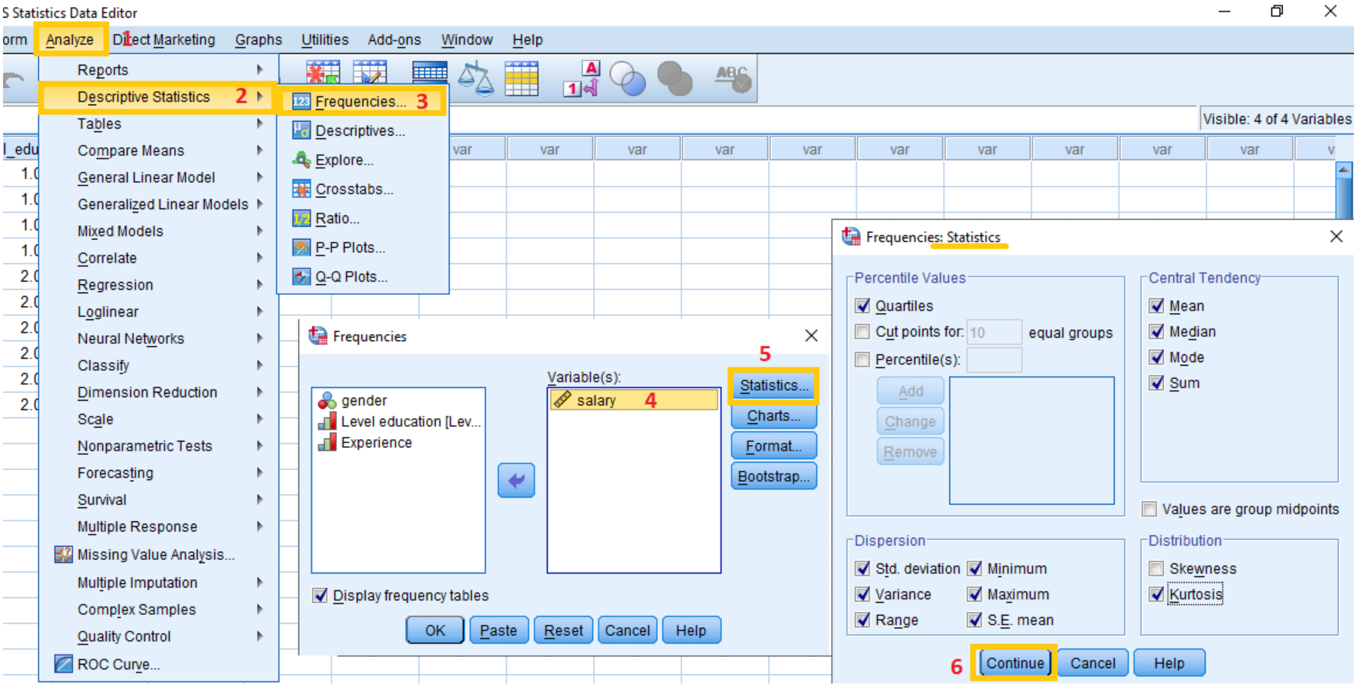
1.00 = "Less than 5"
2.00 = "between 5 and 10"
3.00 = "greater than 5"



	gender	level_edu	experience	salary
1	1.00	1.00	1.00	500.00
2	2.00	1.00	2.00	450.00
3	1.00	1.00	1.00	440.00
4	2.00	1.00	3.00	500.00
5	1.00	2.00	2.00	570.00
6	2.00	2.00	3.00	550.00
7	2.00	2.00	2.00	490.00
8	2.00	2.00	3.00	540.00
9	1.00	2.00	2.00	600.00
10	1.00	2.00	3.00	650.00

gender	level_edu	experience	salary
Male	Bachelor's...	Less than 5	500.00
Female	Bachelor's...	between 5 ...	450.00
Male	Bachelor's...	Less than 5	440.00
Female	Bachelor's...	greater tha...	500.00
Male	master's d...	between 5 ...	570.00
Female	master's d...	greater tha...	550.00
Female	master's d...	between 5 ...	490.00
Female	master's d...	greater tha...	540.00
Male	master's d...	between 5 ...	600.00
Male	master's d...	greater tha...	650.00

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



☑ Display frequency tables

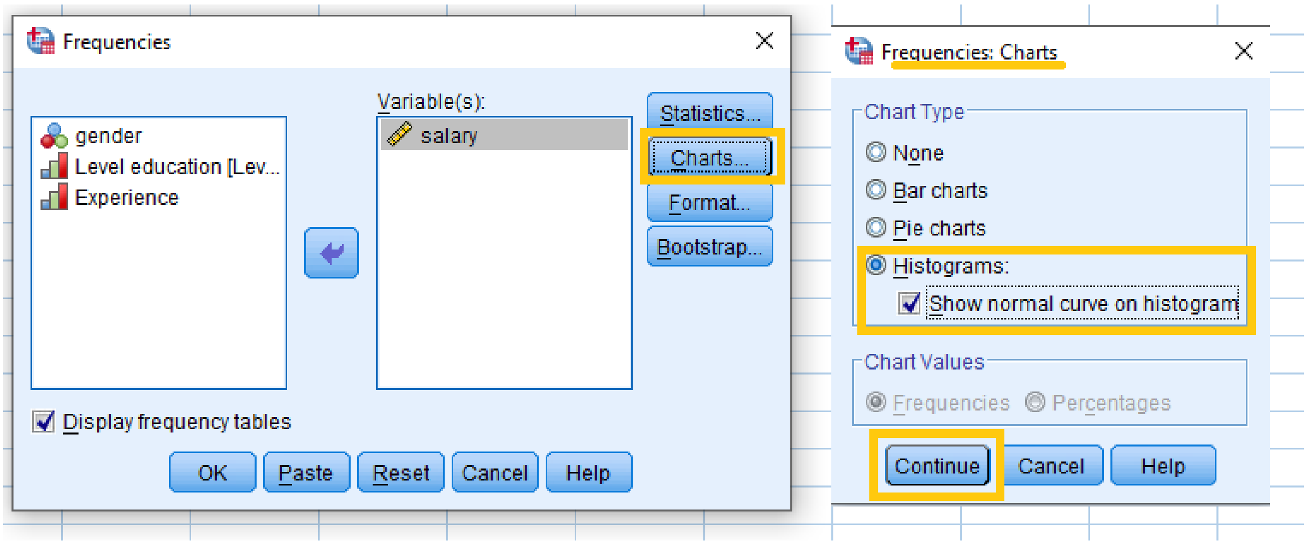
Percentile Values
- ☑ Quartiles
- ☐ Cut points for: 10 equal groups
- ☐ Percentile(s):

Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- ☑ Sum

☐ Values are group midpoints

Dispersion
- ☑ Std. deviation
- ☑ Minimum
- ☑ Variance
- ☑ Maximum
- ☑ Range
- ☑ S.E. mean

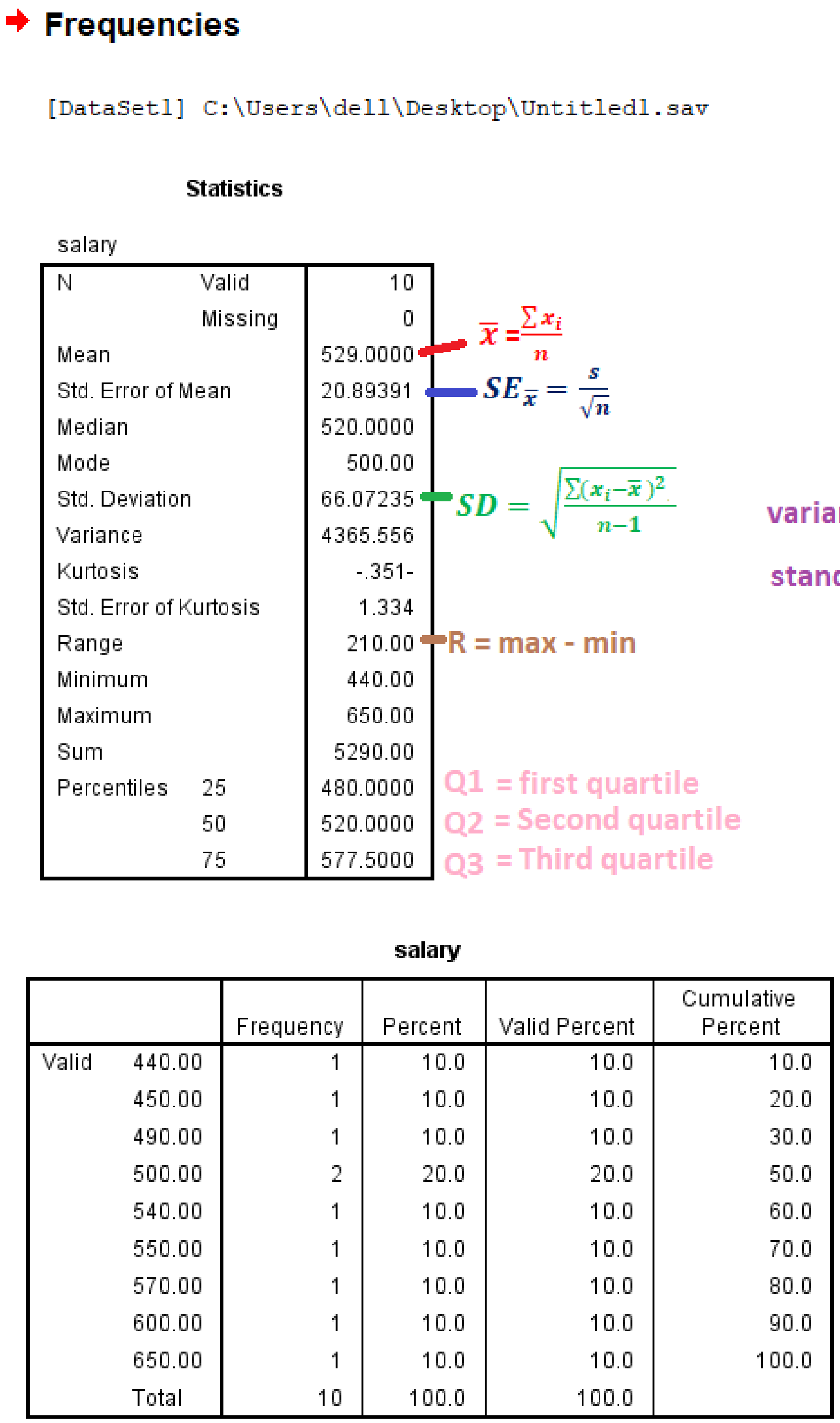
Distribution
- ☐ Skewness
- ☑ Kurtosis



☑ Display frequency tables

Chart Type
- ○ None
- ○ Bar charts
- ○ Pie charts
- ● Histograms:
  - ☑ Show normal curve on histogram

Chart Values
- ● Frequencies ○ Percentages



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

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

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

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

variance = (standard deviation)^2

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

R = max - min

Q1 = first quartile

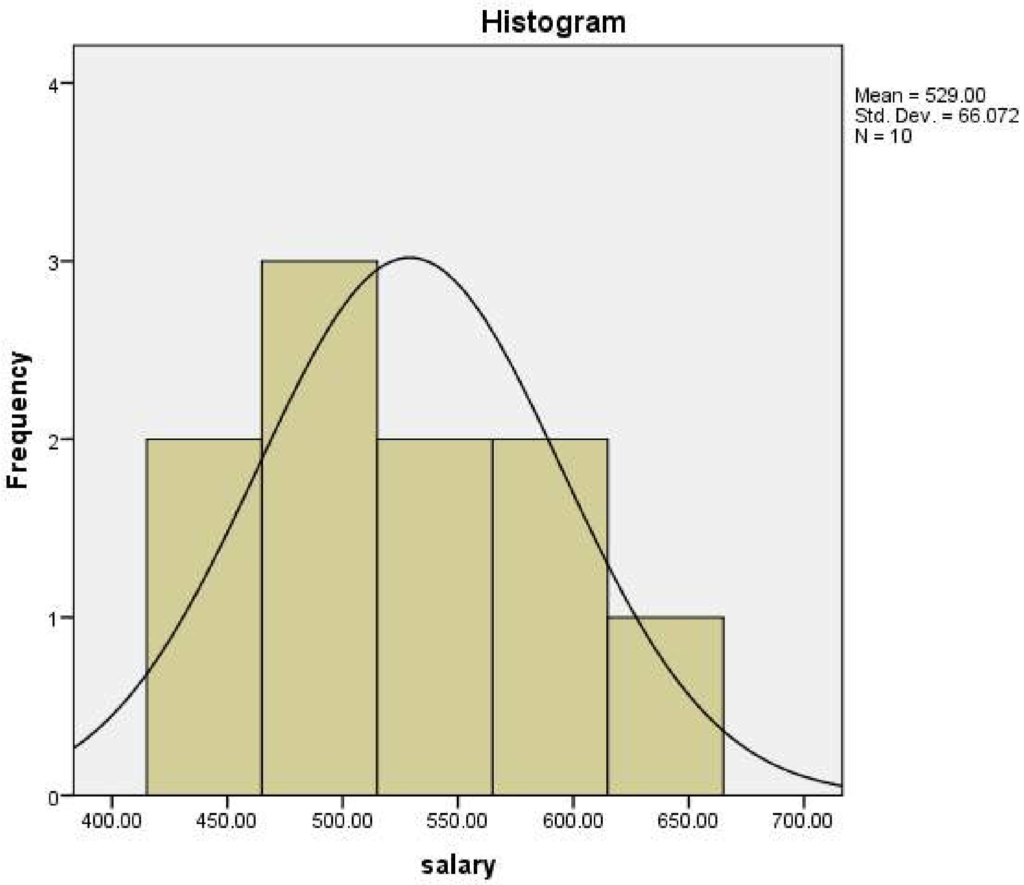
Q2 = Second quartile

Q3 = Third quartile

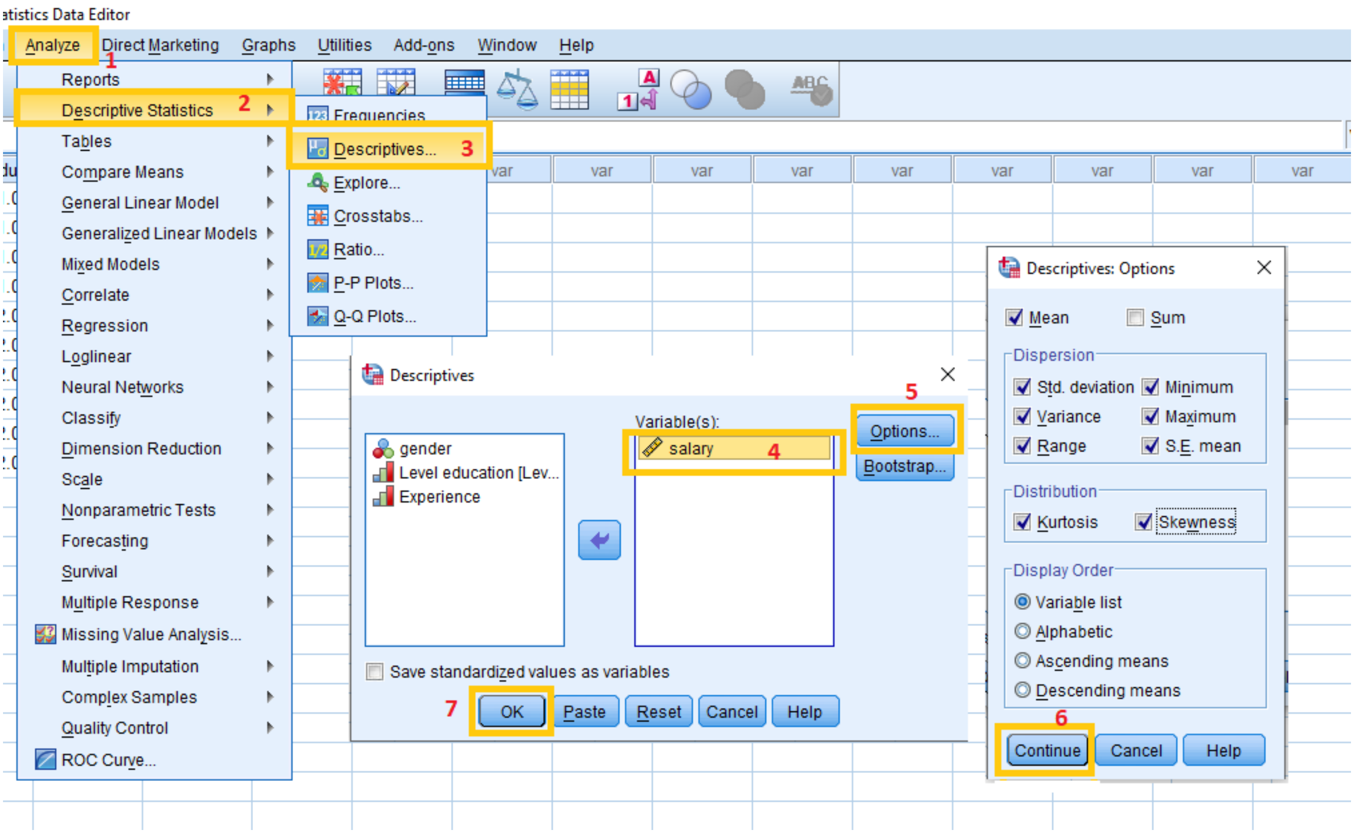
salary

		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	

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


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 :



Variable(s): salary

☐ Save standardized values as variables

☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum

☑ Variance ☑ Maximum

☑ Range ☑ S.E. mean

Distribution

☑ Kurtosis ☑ Skewness

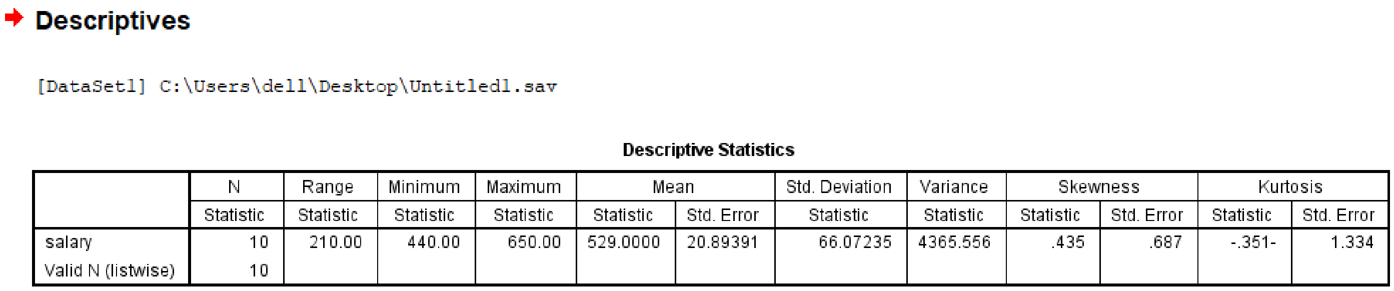
Display Order

● Variable list

○ Alphabetic

○ Ascending means

○ Descending means



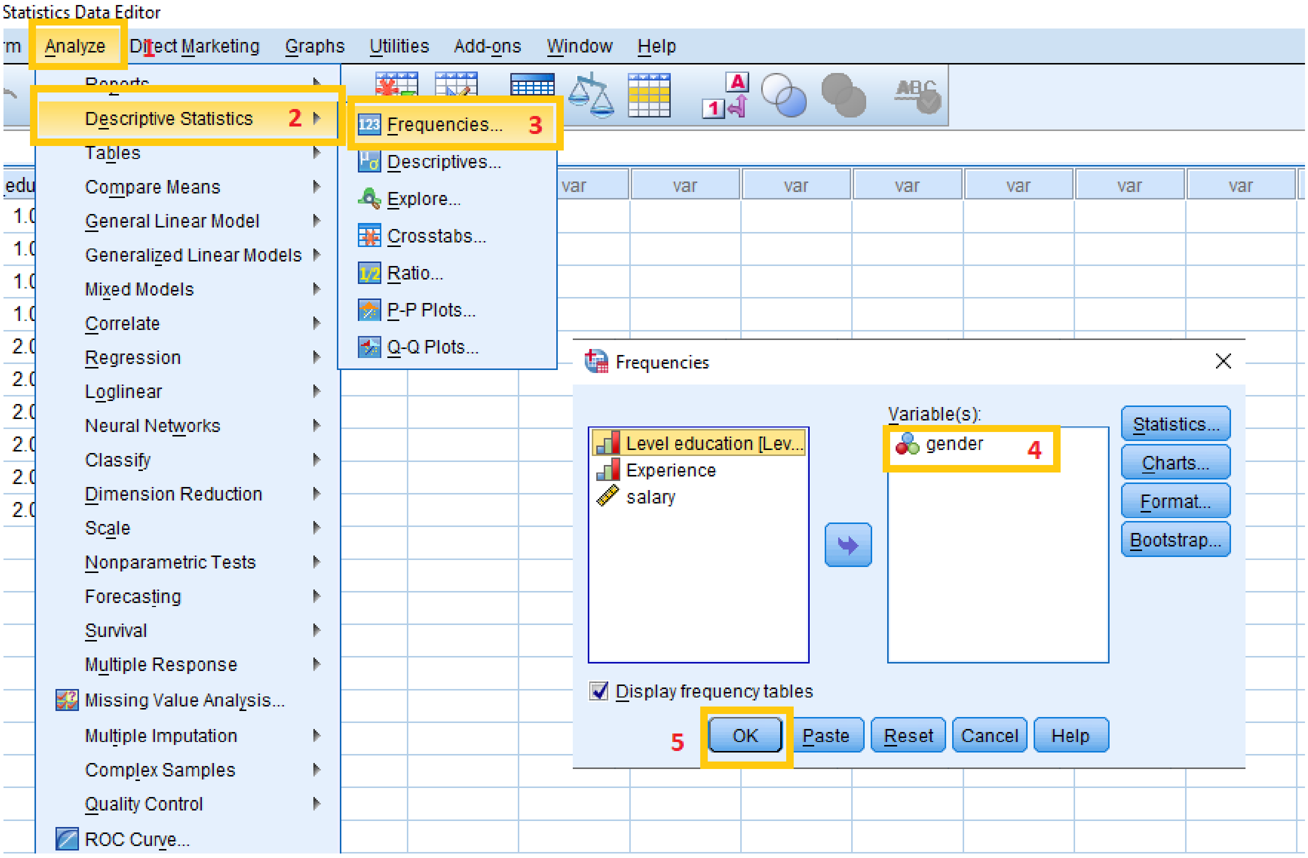
Descriptives

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

Descriptive Statistics

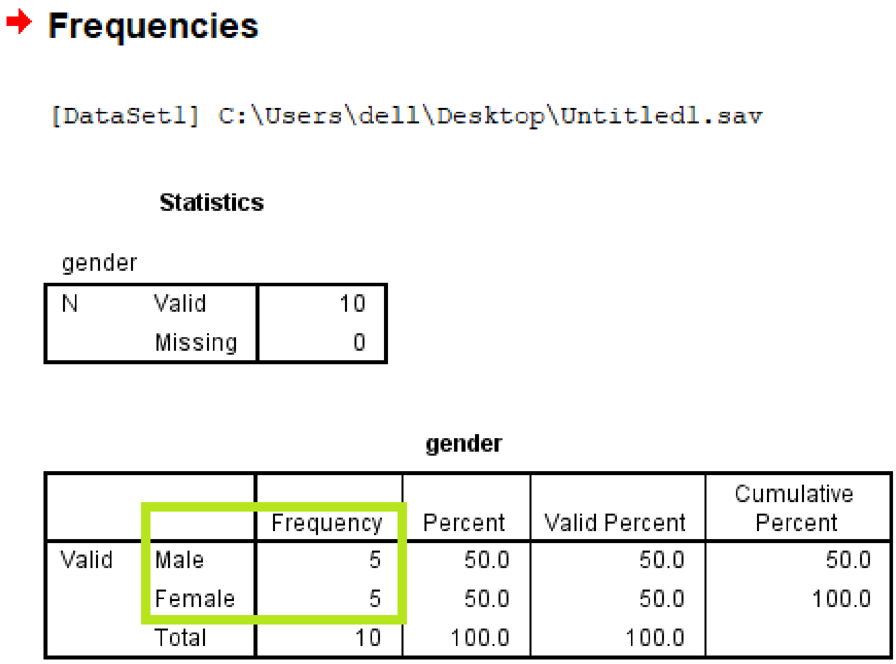
	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
salary	10	210.00	440.00	650.00	529.0000	20.89391	66.07235	4365.556	.435	.687	-.351-	1.334
Valid N (listwise)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

3- How many male and female?



Variable(s): gender

☑ Display frequency tables



Frequencies

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

Statistics

gender

N	Valid	10
	Missing	0

gender

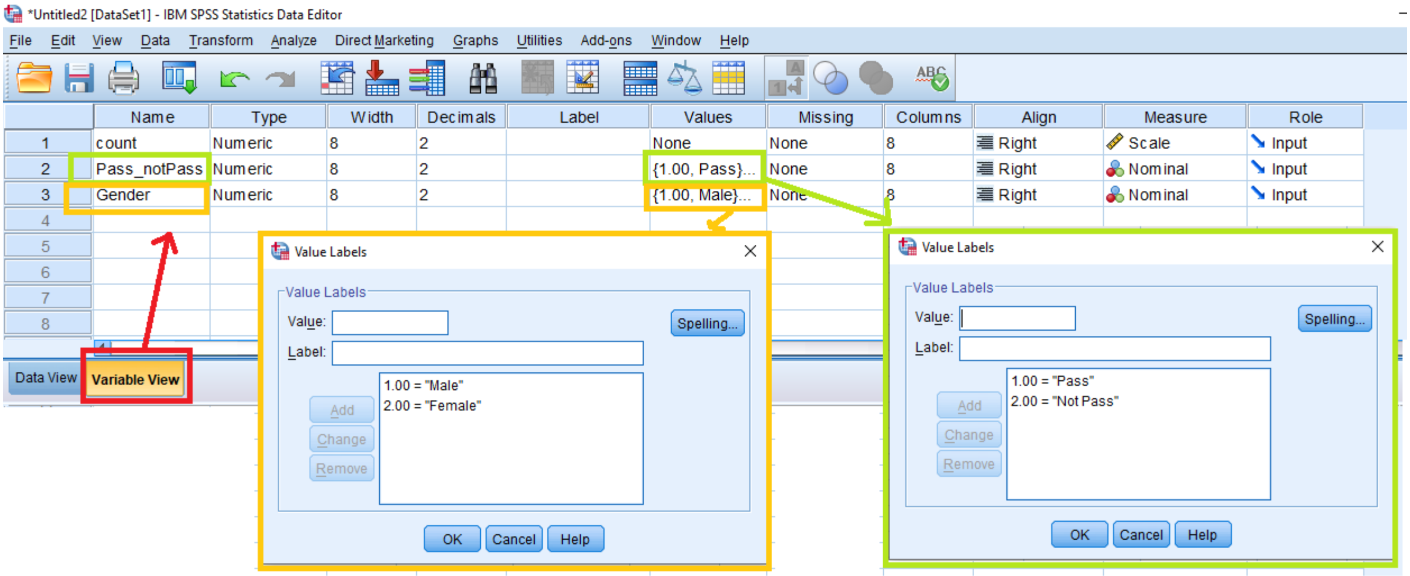
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	5	50.0	50.0	50.0
	Female	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

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**

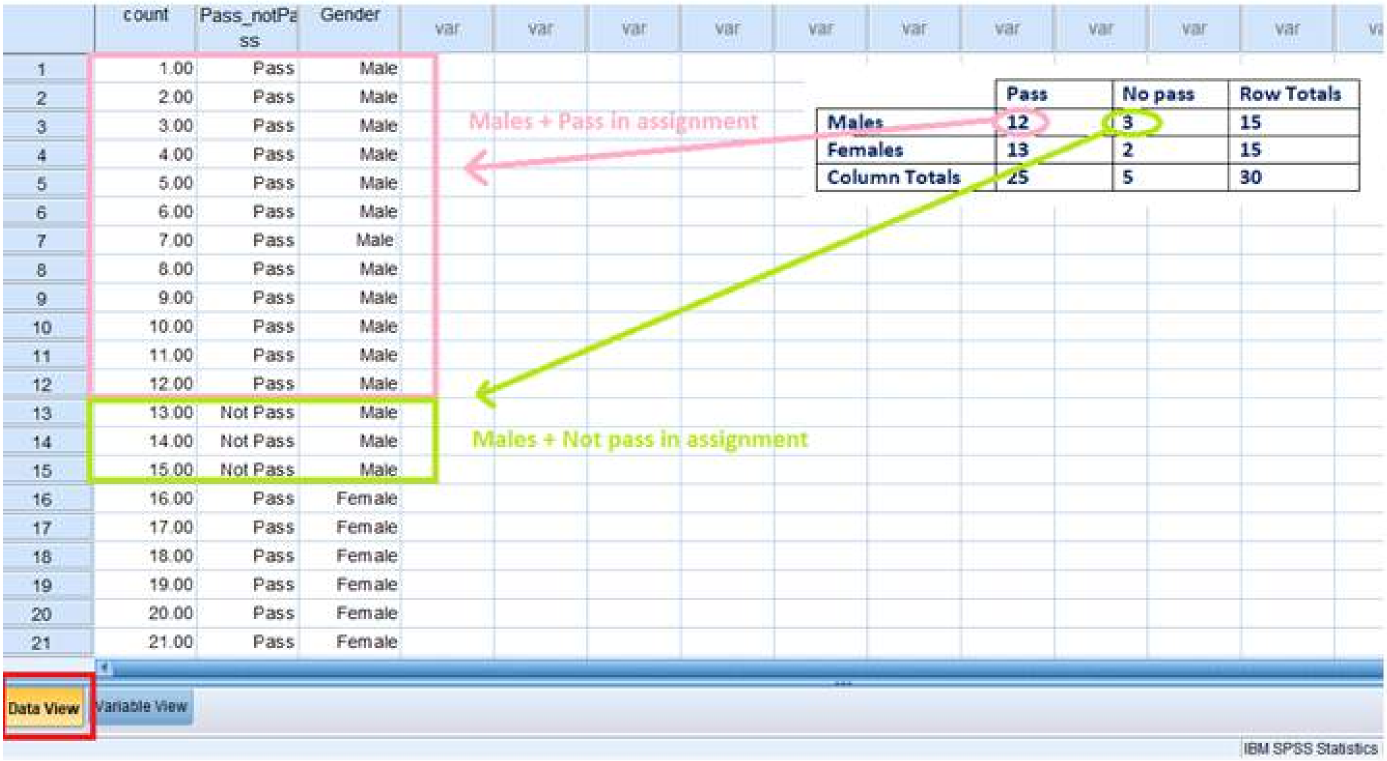
H_1: **the gender of the students is not independent of pass or no pass test grade**



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	count	Numeric	8	2		None	None	8	Right	Scale	Input
2	Pass_notPass	Numeric	8	2		{1.00, Pass}...	None	8	Right	Nominal	Input
3	Gender	Numeric	8	2		{1.00, Male}...	None	8	Right	Nominal	Input
4											
5											
6											
7											
8											

1.00 = "Male"
2.00 = "Female"

1.00 = "Pass"
2.00 = "Not Pass"

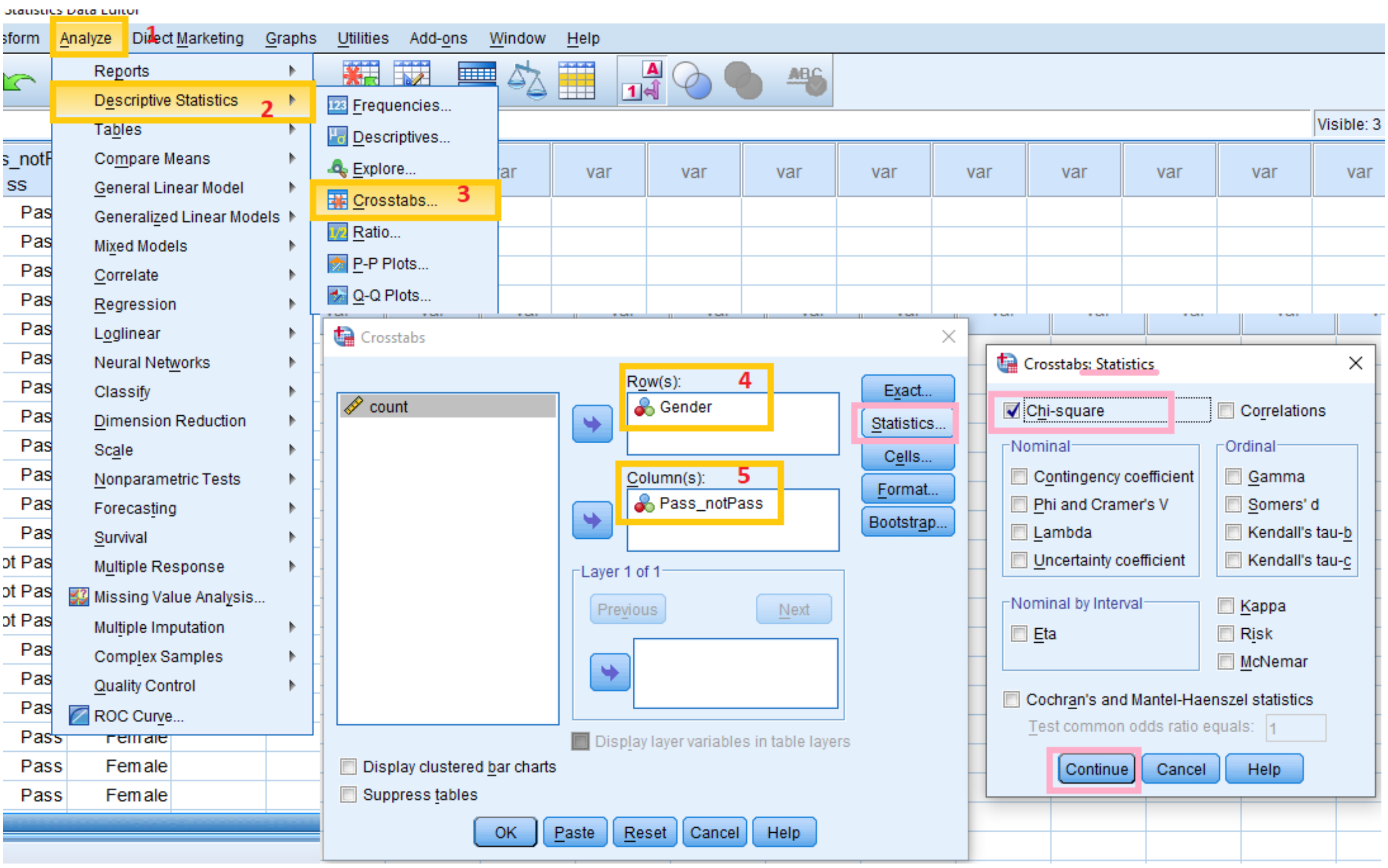


	count	Pass_notPass	Gender
1	1.00	Pass	Male
2	2.00	Pass	Male
3	3.00	Pass	Male
4	4.00	Pass	Male
5	5.00	Pass	Male
6	6.00	Pass	Male
7	7.00	Pass	Male
8	8.00	Pass	Male
9	9.00	Pass	Male
10	10.00	Pass	Male
11	11.00	Pass	Male
12	12.00	Pass	Male
13	13.00	Not Pass	Male
14	14.00	Not Pass	Male
15	15.00	Not Pass	Male
16	16.00	Pass	Female
17	17.00	Pass	Female
18	18.00	Pass	Female
19	19.00	Pass	Female
20	20.00	Pass	Female
21	21.00	Pass	Female

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

Males + Pass in assignment

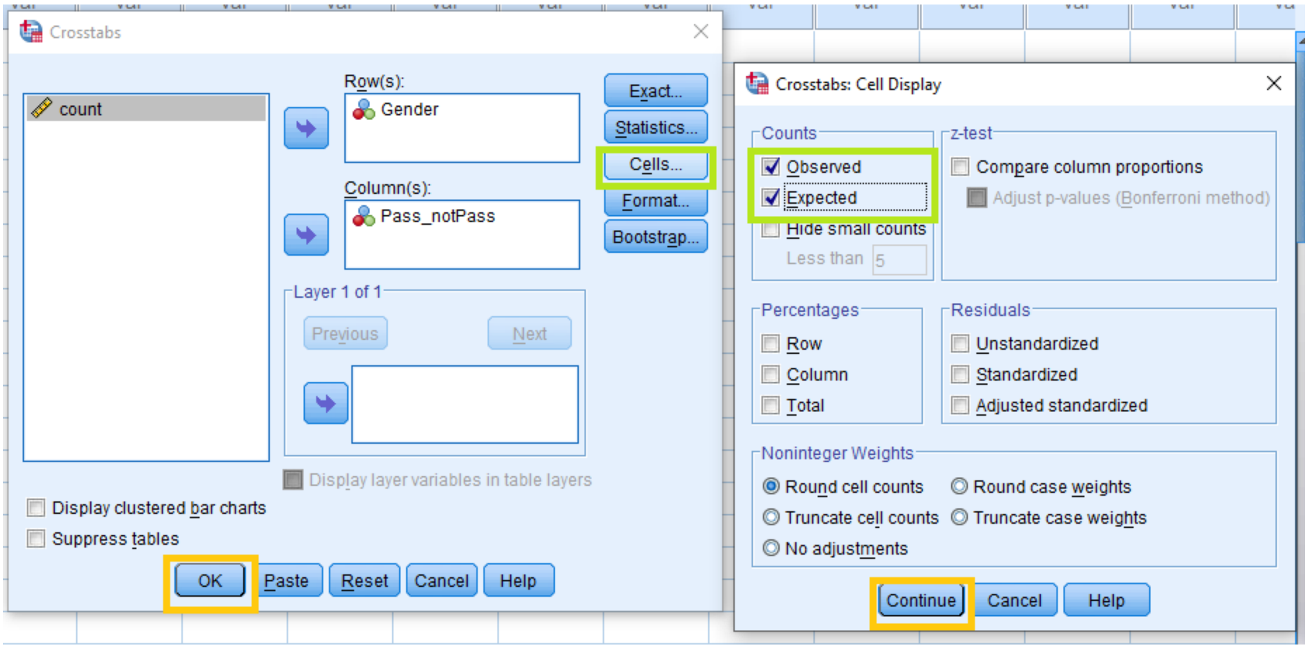
Males + Not pass in assignment

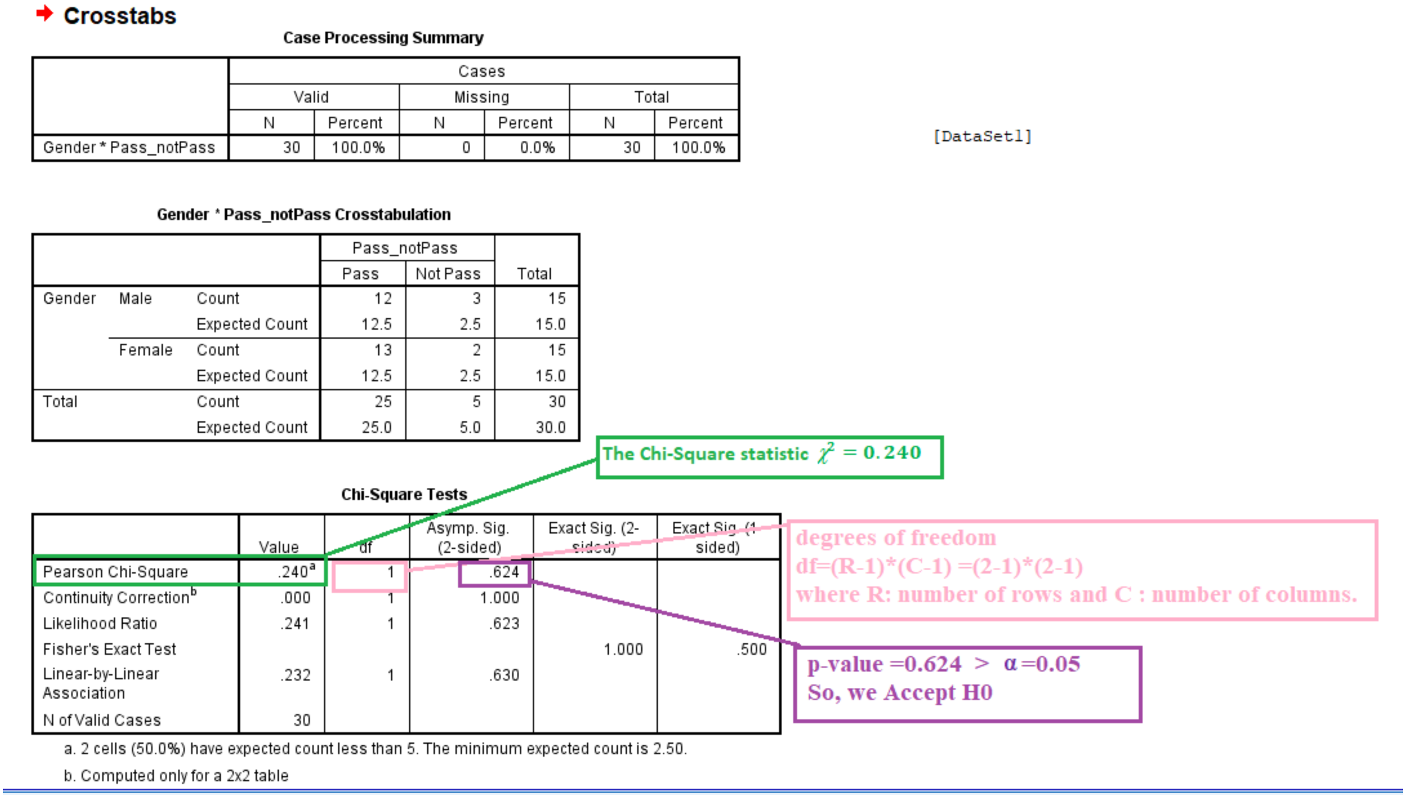


4 Row(s): Gender

5 Column(s): Pass_notPass

☑ Chi-square





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				

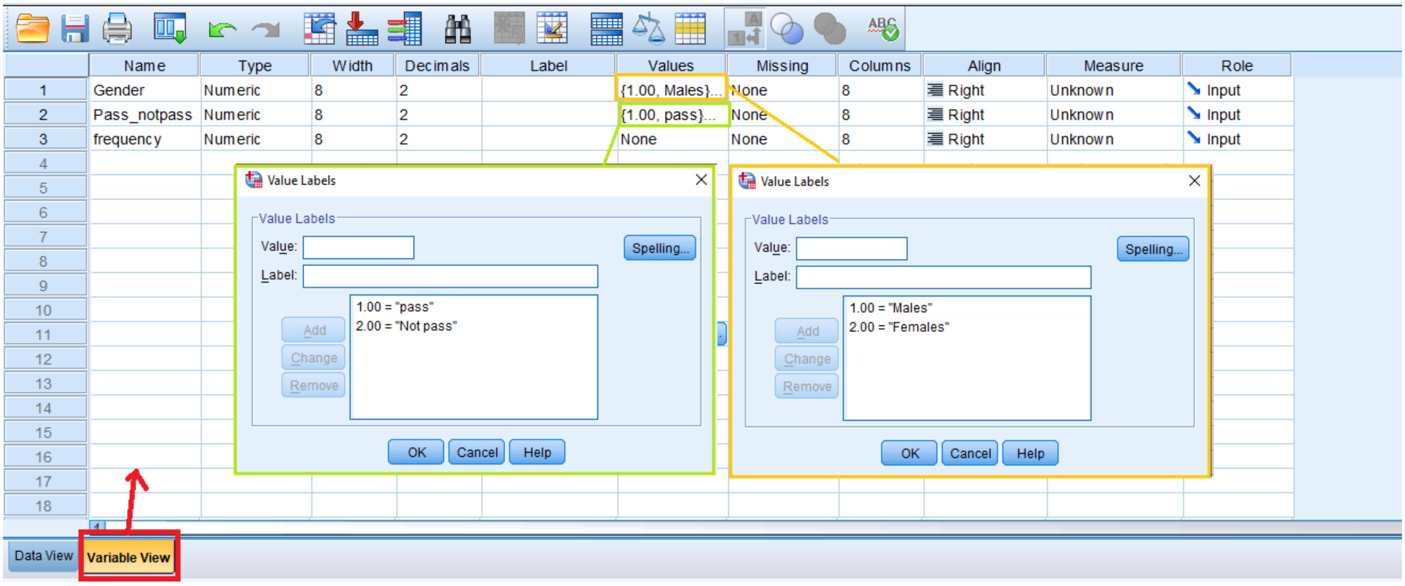
a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.50.

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 :



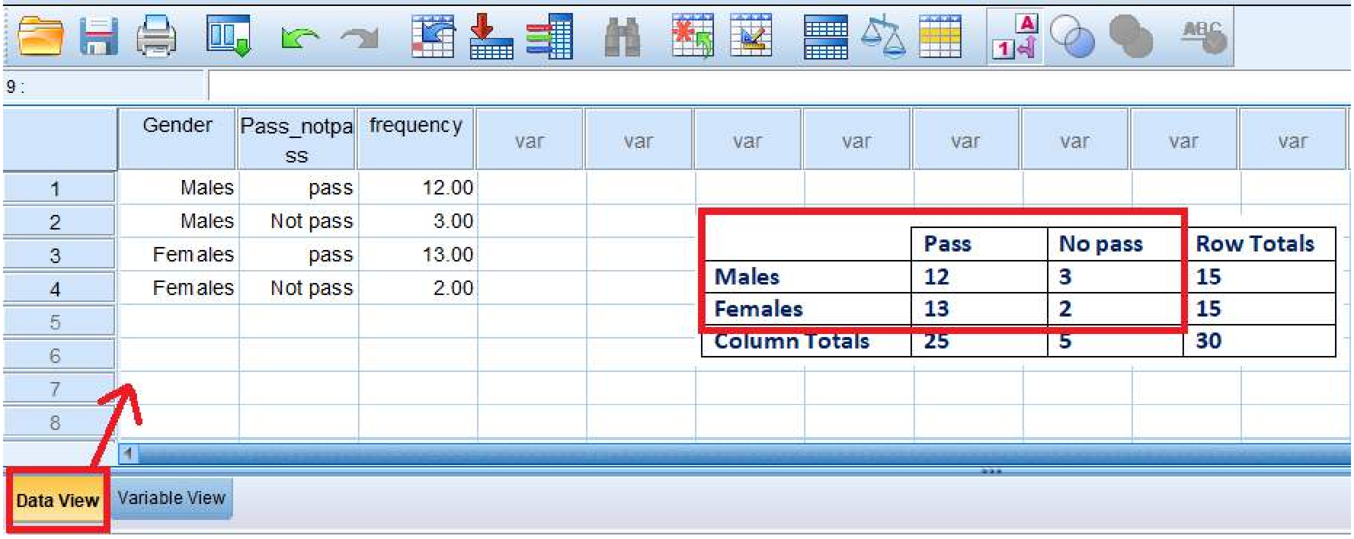
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	2		{1.00, Males}...	None	8	Right	Unknown	Input
2	Pass_notpass	Numeric	8	2		{1.00, pass}...	None	8	Right	Unknown	Input
3	frequency	Numeric	8	2		None	None	8	Right	Unknown	Input

Value Labels

1.00 = "pass"
2.00 = "Not pass"

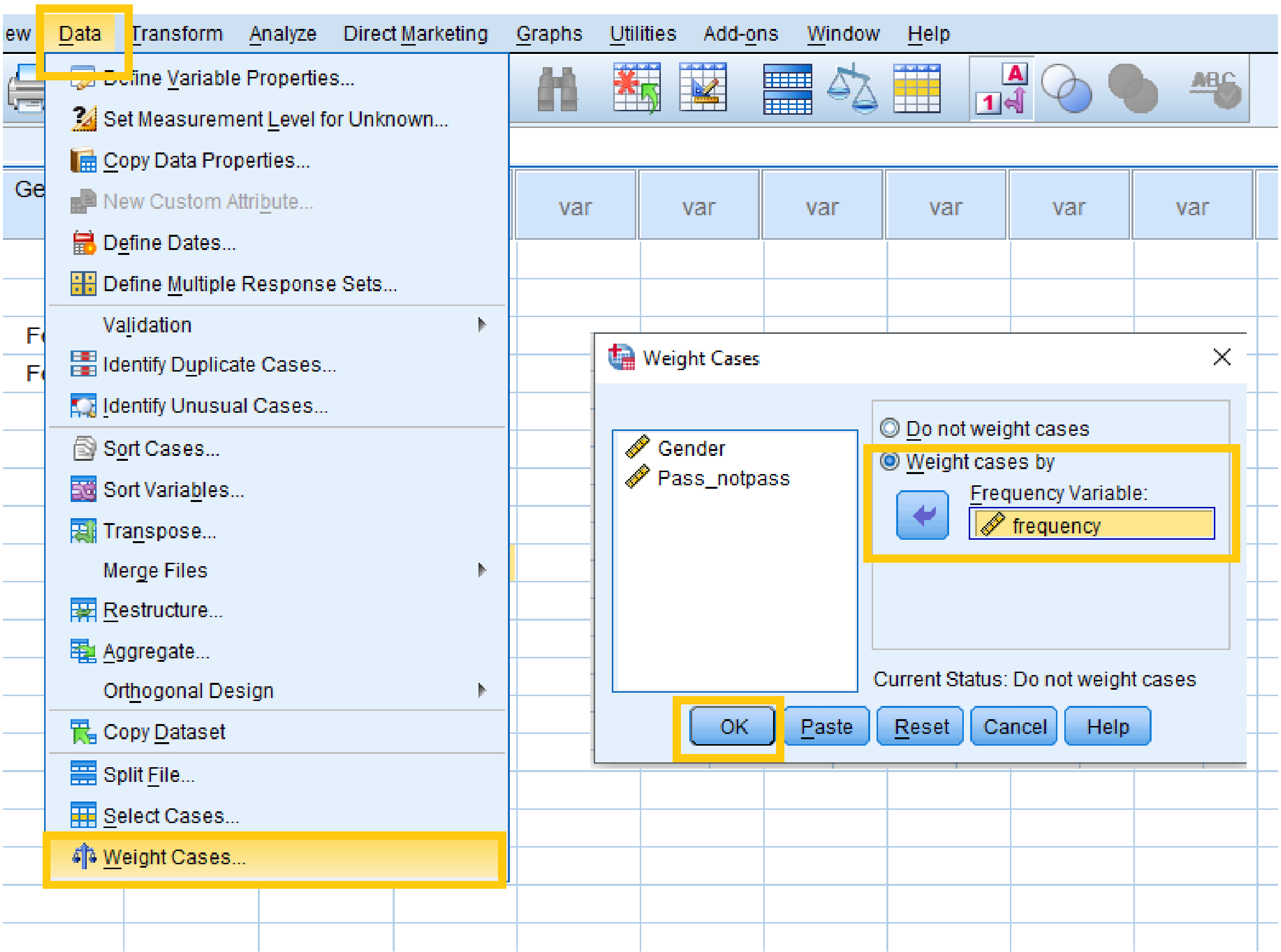
Value Labels

1.00 = "Males"
2.00 = "Females"



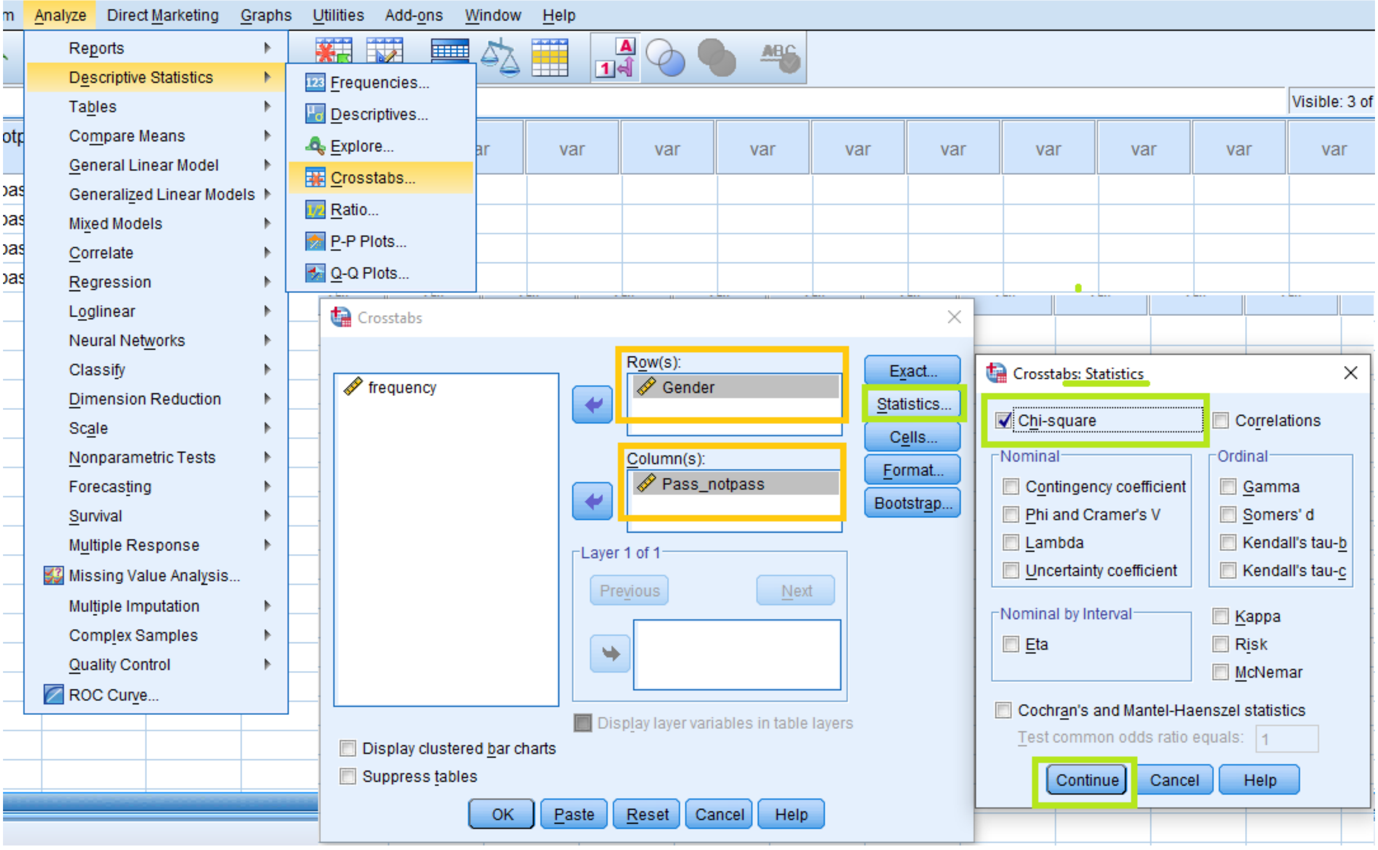
	Gender	Pass_notpass	frequency
1	Males	pass	12.00
2	Males	Not pass	3.00
3	Females	pass	13.00
4	Females	Not pass	2.00

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



○ Do not weight cases
● Weight cases by
Frequency Variable: frequency

Current Status: Do not weight cases

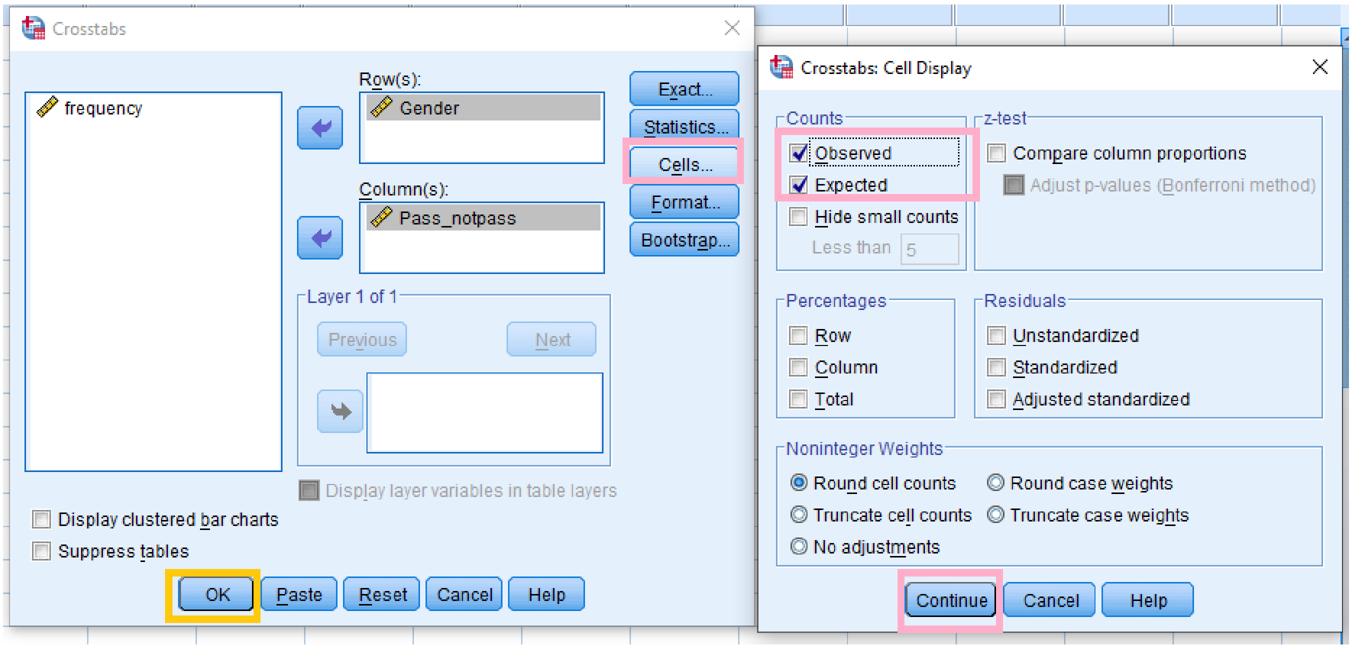


Row(s): Gender

Column(s): Pass_notpass

☐ Display clustered bar charts
☐ Suppress tables

☑ Chi-square
☐ Correlations
☐ Contingency coefficient
☐ Phi and Cramer's V
☐ Lambda
☐ Uncertainty coefficient
☐ Gamma
☐ Somers' d
☐ Kendall's tau-b
☐ Kendall's tau-c
☐ Eta
☐ Kappa
☐ Risk
☐ McNemar
☐ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

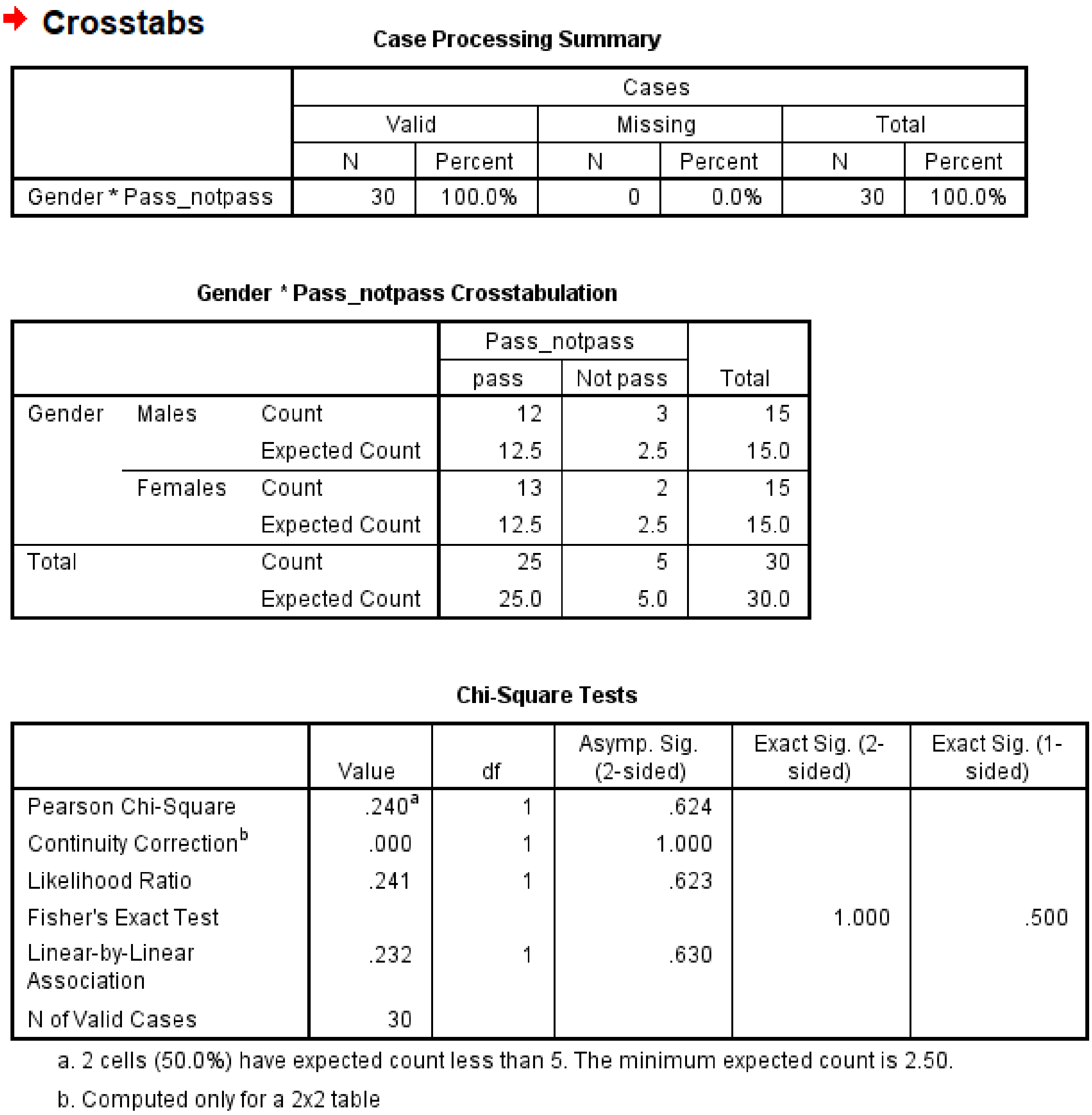


Row(s): Gender

Column(s): Pass_notpass

☐ Display clustered bar charts
☐ Suppress tables

☑ Observed
☑ Expected
☐ Hide small counts
Less than 5
☐ Compare column proportions
☐ Adjust p-values (Bonferroni method)
☐ Row
☐ Column
☐ Total
☐ Unstandardized
☐ Standardized
☐ Adjusted standardized
● Round cell counts
○ Round case weights
○ Truncate cell counts
○ Truncate case weights
○ No adjustments



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				

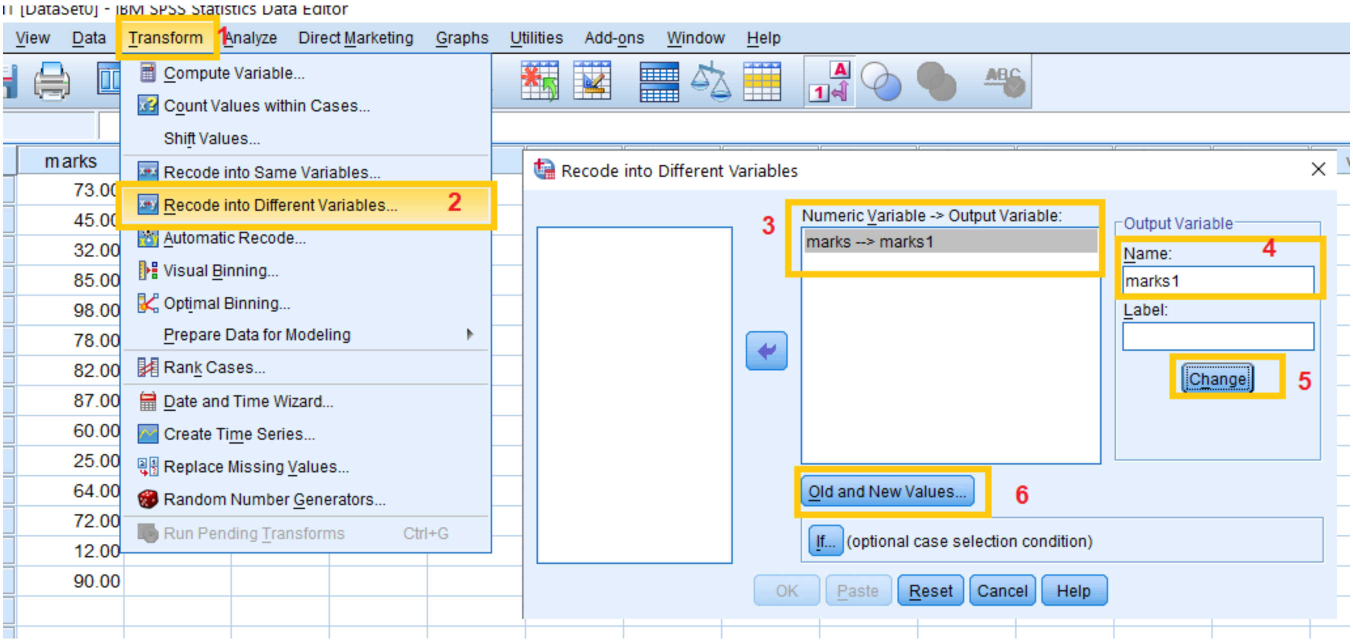
a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.50.

b. Computed only for a 2x2 table

Q3: We have marks of 14 students

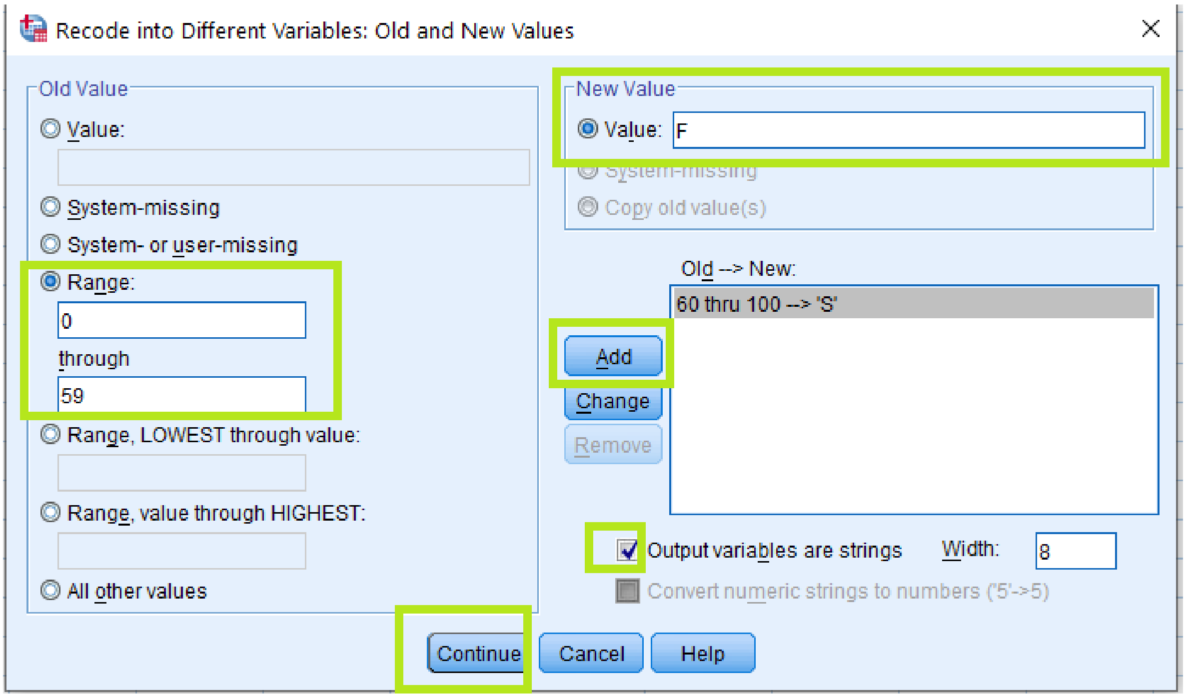
73 45 32 85 98 78 82 87 60 25 64 72 12 90

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



Numeric Variable -> Output Variable: marks --> marks1

Name: marks1

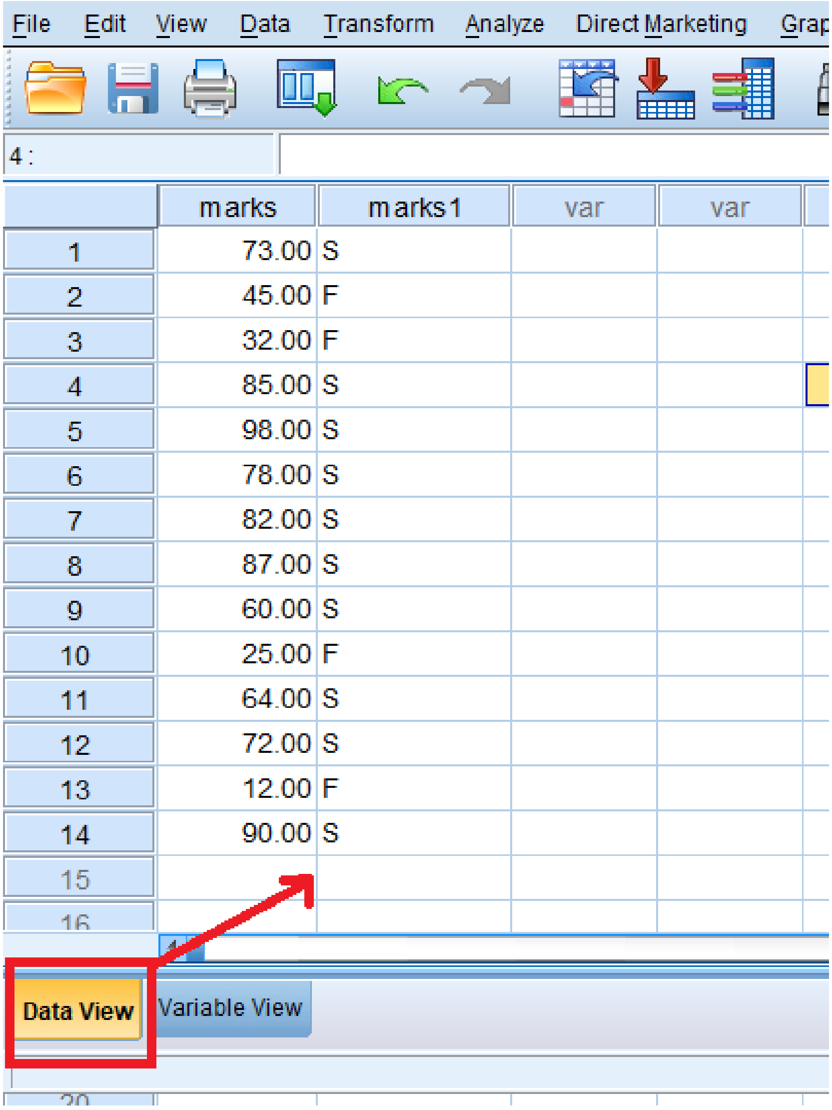


● Range: 0 through 59

● Value: F

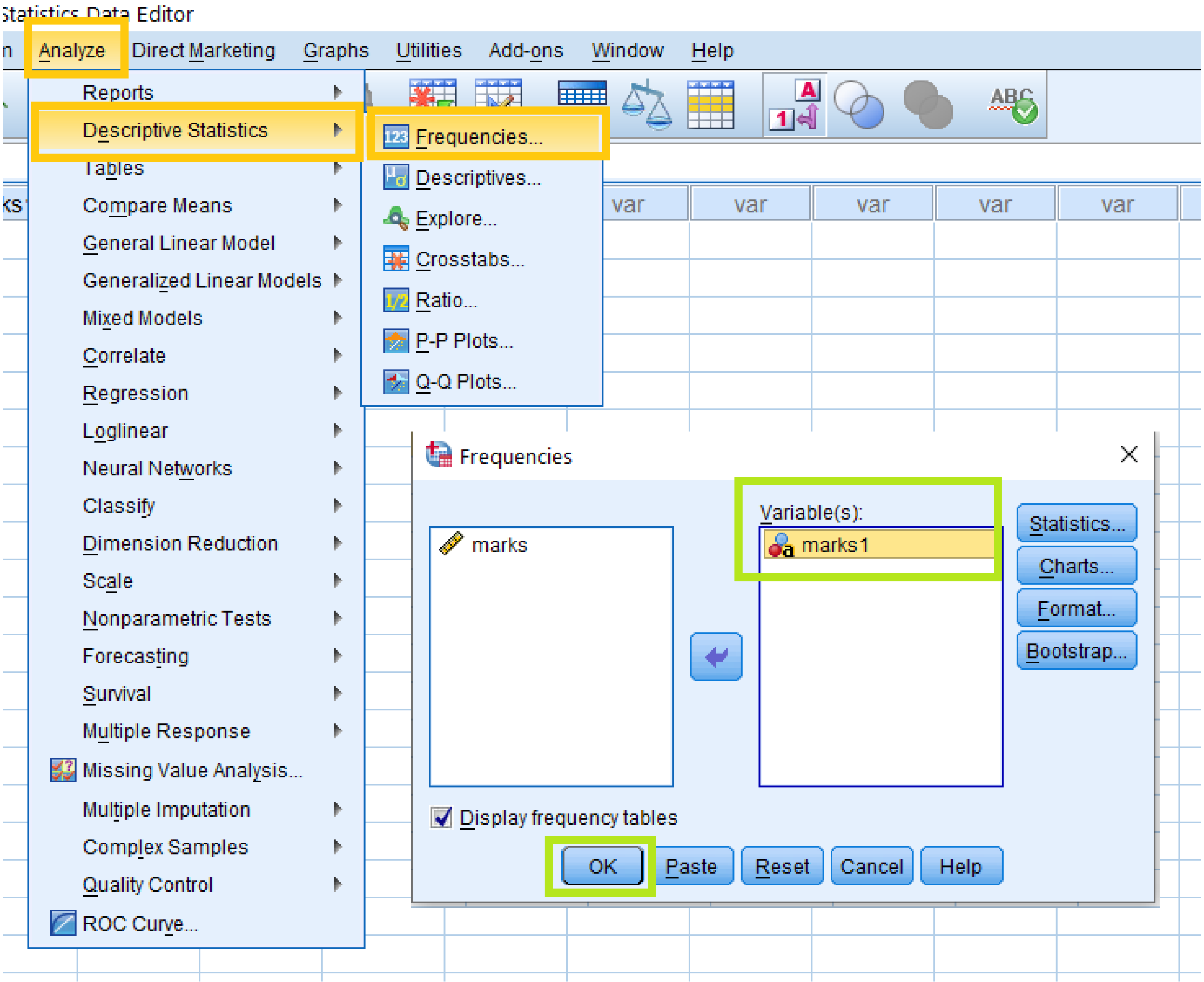
Old --> New: 60 thru 100 --> 'S'

☑ Output variables are strings Width: 8



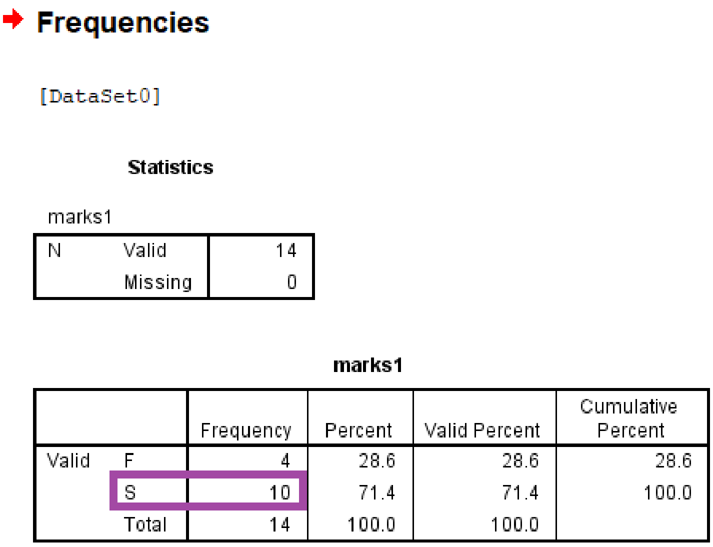
	marks	marks1	var	var
1	73.00	S		
2	45.00	F		
3	32.00	F		
4	85.00	S		
5	98.00	S		
6	78.00	S		
7	82.00	S		
8	87.00	S		
9	60.00	S		
10	25.00	F		
11	64.00	S		
12	72.00	S		
13	12.00	F		
14	90.00	S		
15				

2. How many successful students?



Variable(s): marks1

☑ Display frequency tables



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 <u>approximately normally distributed</u>, find and interpret a <u>90% confidence interval</u> for the average fruit shape.

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

$$H_0\text{: the population follows a normal distribution}$$

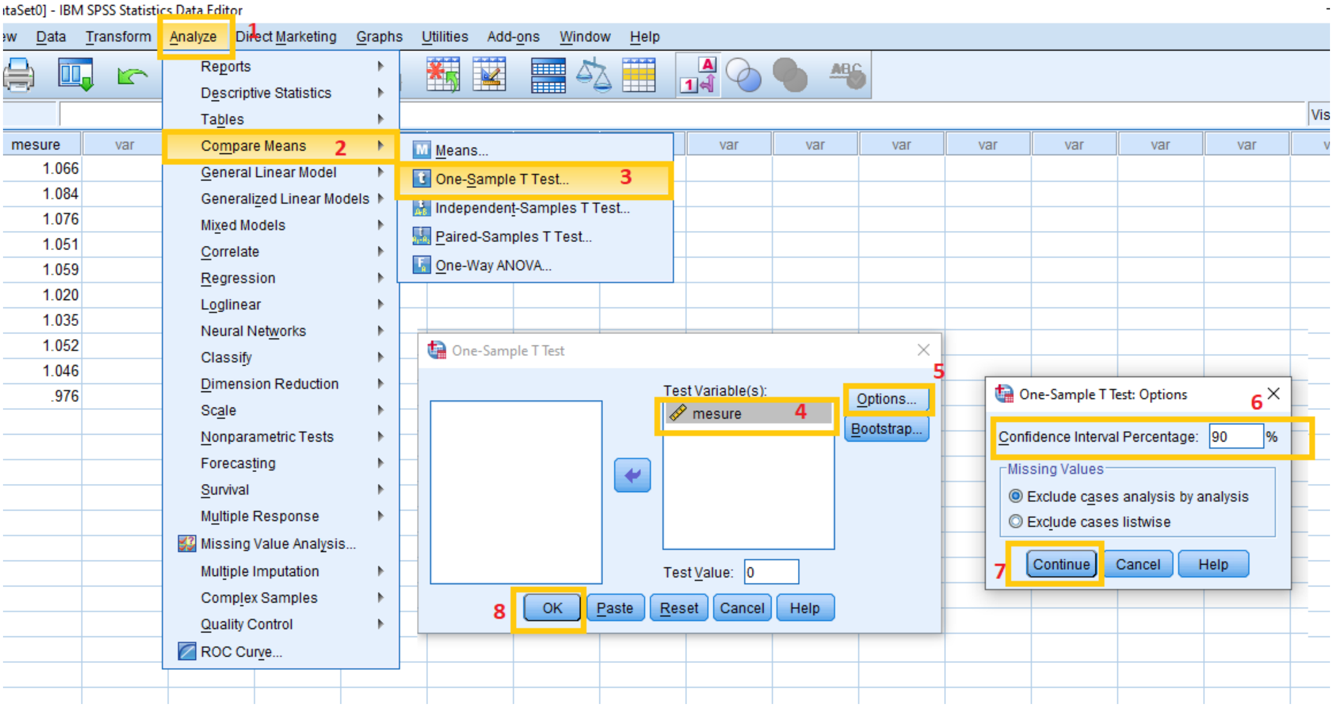
Vs

$$H_1\text{: the population does not follow a normal distribution}$$

we find the question he 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:



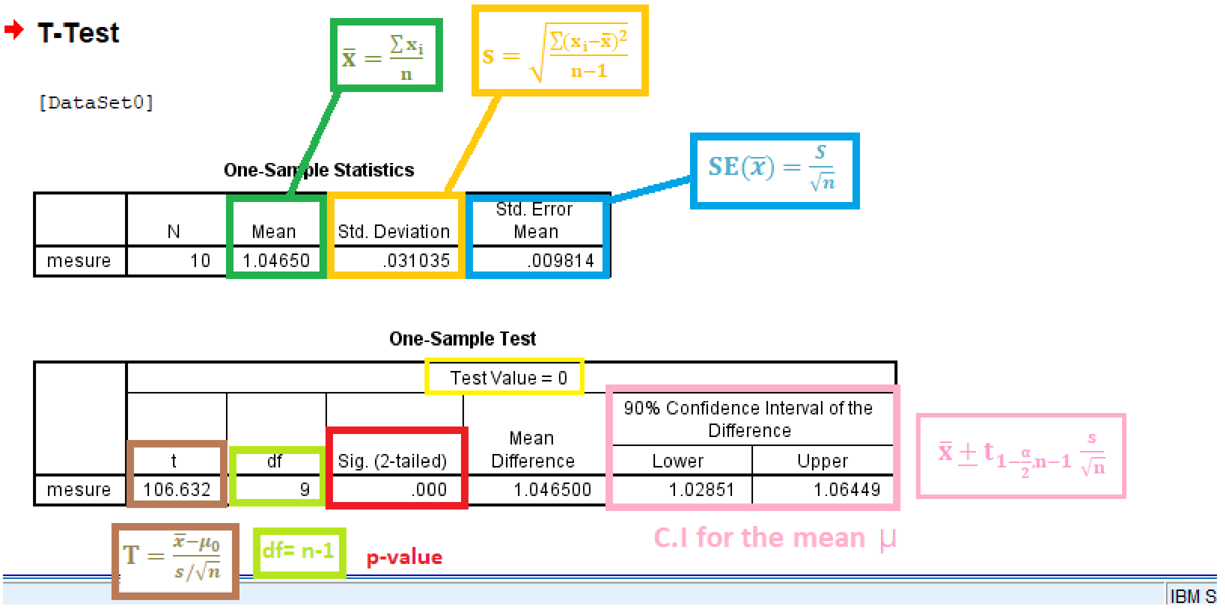
Test Variable(s): mesure

Test Value: 0

Confidence Interval Percentage: 90 %

● Exclude cases analysis by analysis

○ Exclude cases listwise



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

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

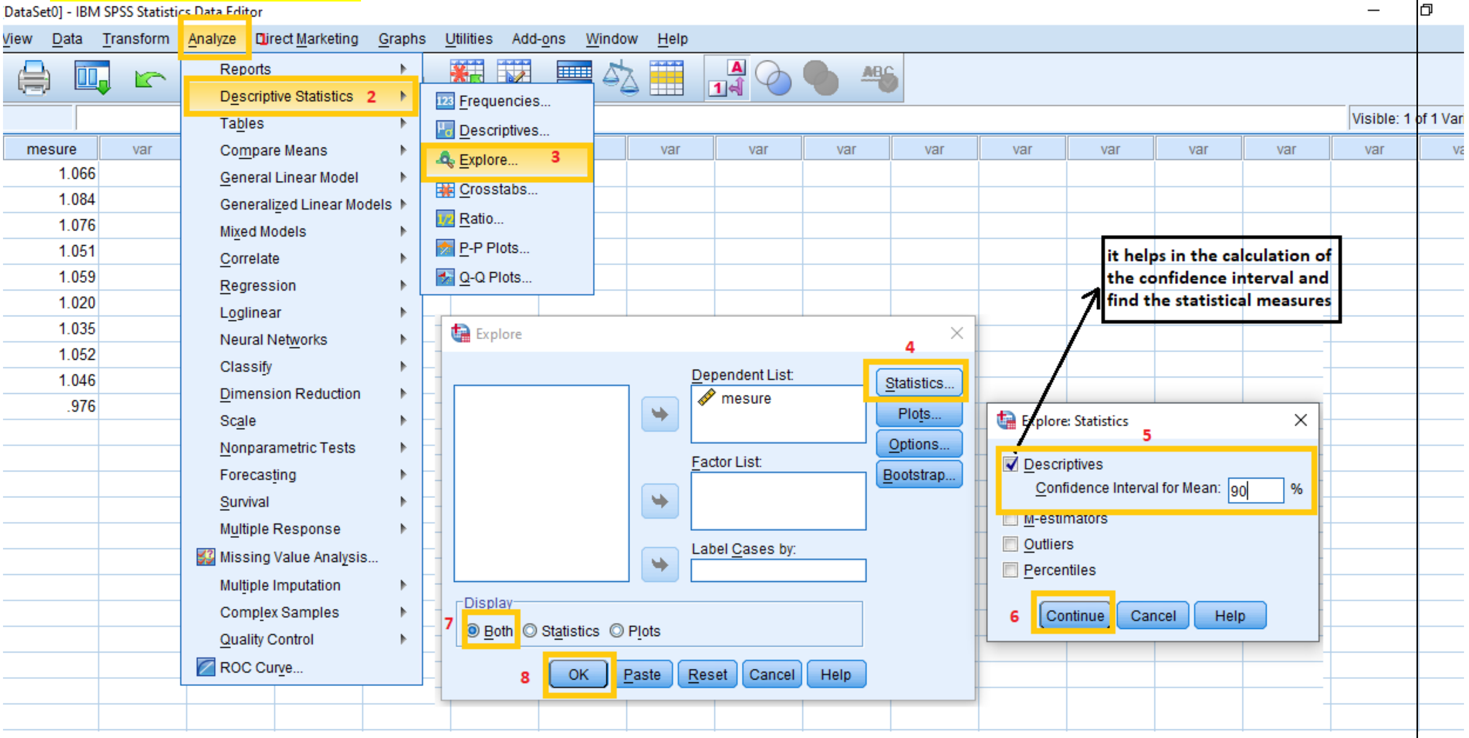
$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

df= n-1

p-value

C.I for the mean μ

2) The second method:



DataSet0] - IBM SPSS Statistics Data Editor

mesure
1.066
1.084
1.076
1.051
1.059
1.020
1.035
1.052
1.046
.976

2 Descriptive Statistics

3 Explore...

Explore

Dependent List: mesure

Factor List:

Label Cases by:

4 Statistics...

Explore: Statistics

5 ☑ Descriptives

Confidence Interval for Mean: 90 %

☐ M-estimators

☐ Outliers

☐ Percentiles

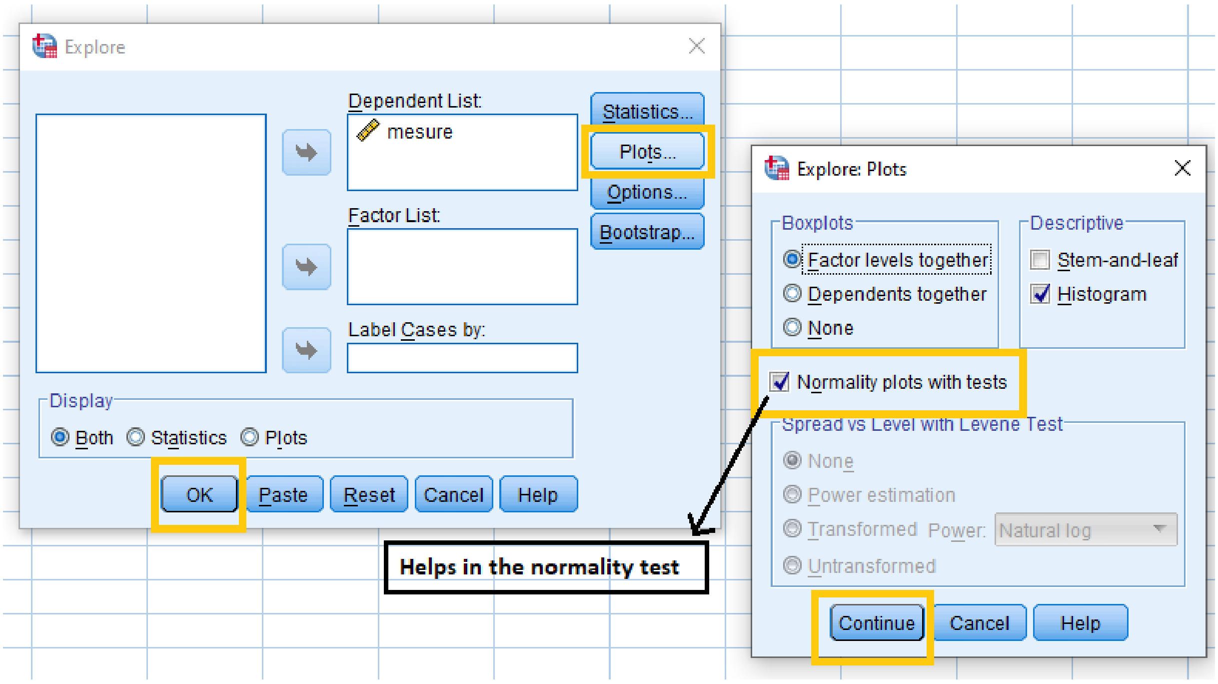
6 Continue

Display

7 ● Both ○ Statistics ○ Plots

8 OK

it helps in the calculation of the confidence interval and find the statistical measures



Dependent List: mesure

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

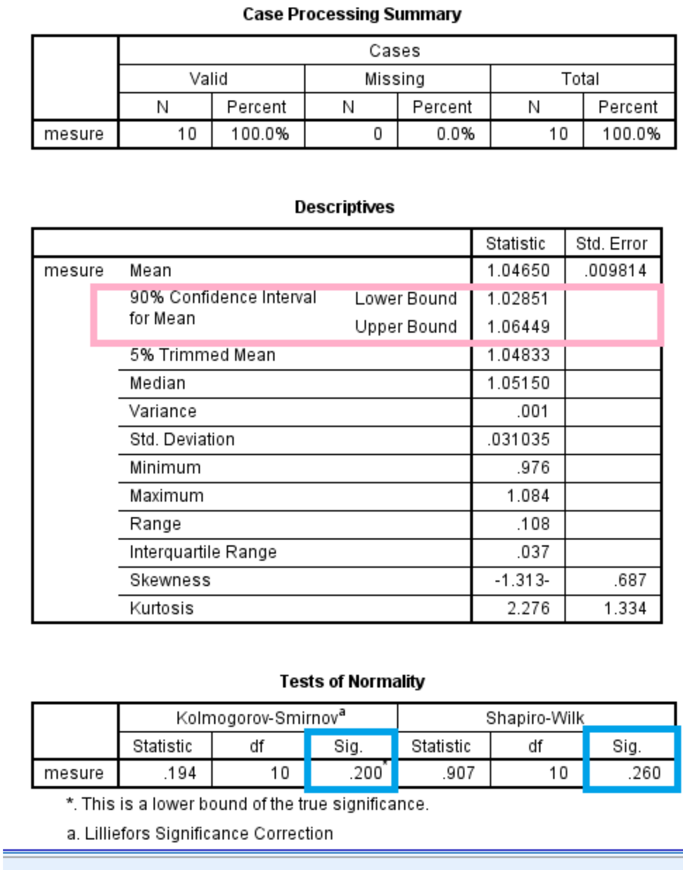
Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☐ Stem-and-leaf ☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None

Helps in the normality test



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 <u>skim milk is less than the average phosphorus content of whole milk</u> . 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.

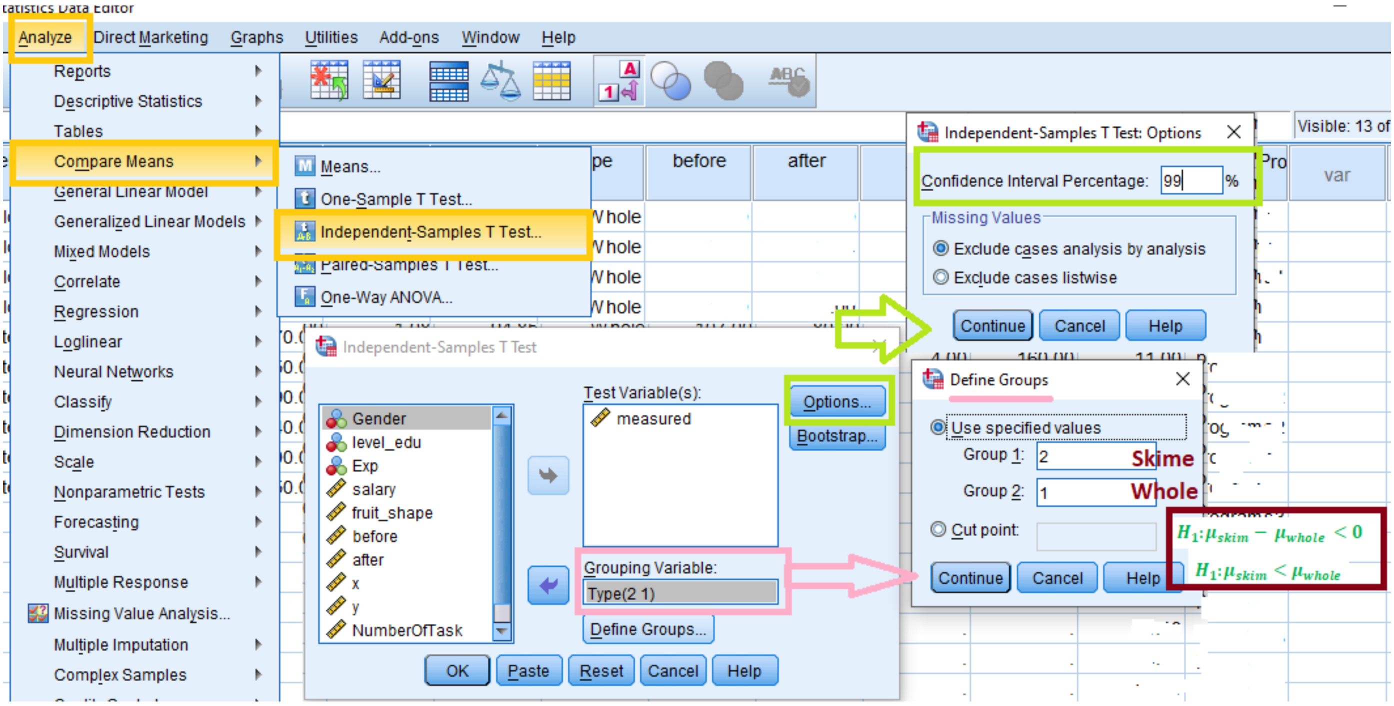
2) The populations follow a normal distribution

$$H_0\text{: the population follows a normal distribution}$$
$$\textbf{Vs}\ \ H_1\text{: the population does not follow a normal distribution}$$

However, we find the question he 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) **90 % Confidence interval of** $\mu_{skim} - \mu_{whole}$



Test Variable(s): measured

Grouping Variable: Type(2 1)

Confidence Interval Percentage: 99 %

Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

● Use specified values

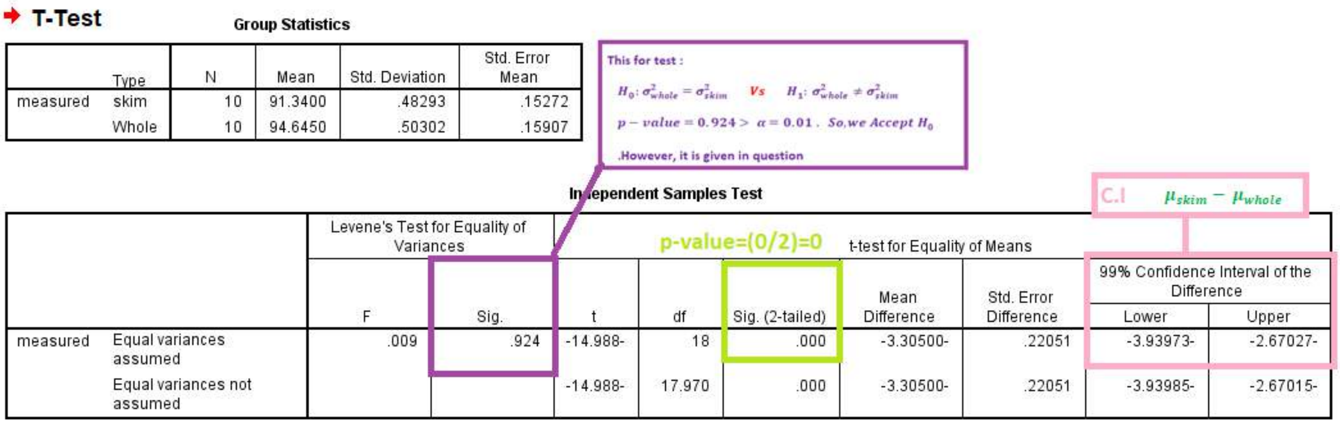
Group 1: 2 **Skime**

Group 2: 1 **Whole**

○ Cut point:

$H_1: \mu_{skim} - \mu_{whole} < 0$

$H_1: \mu_{skim} < \mu_{whole}$



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

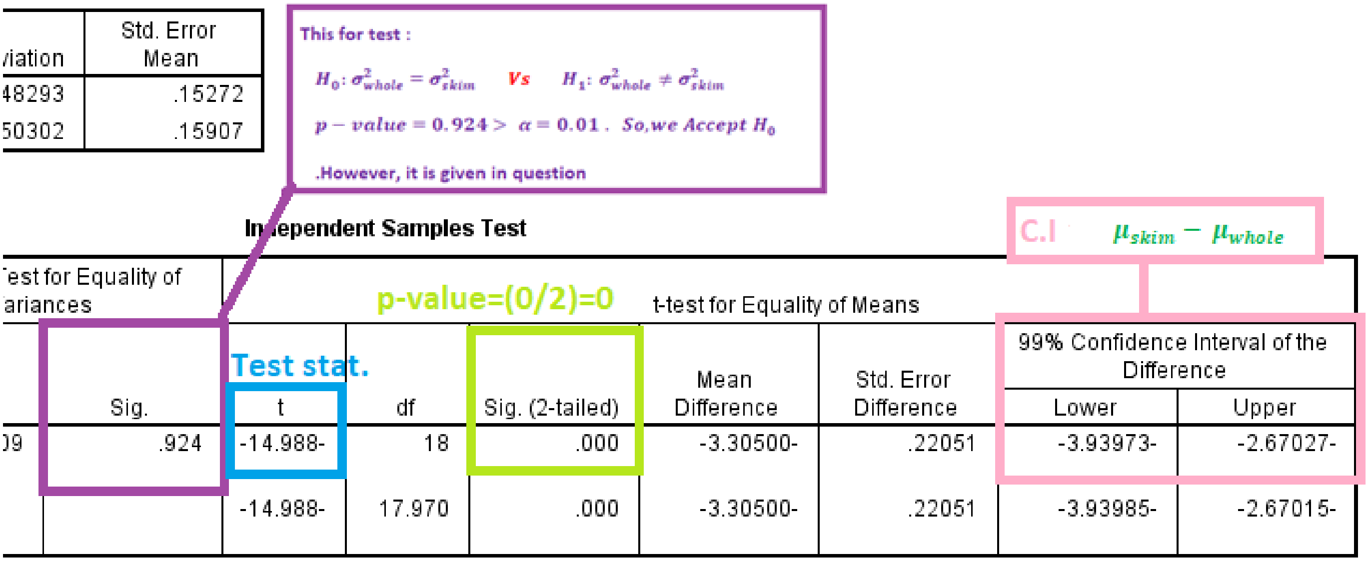
,However, it is given in question

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-



iation	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

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

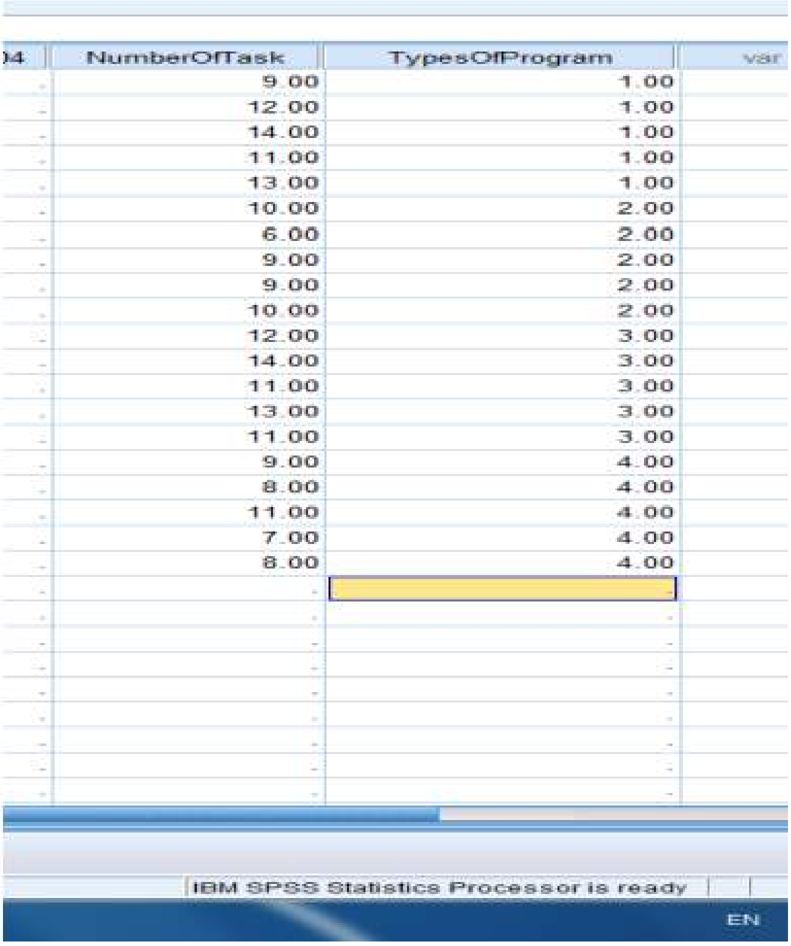
to use the one way ANOVA- test, we need to make sure that :

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

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

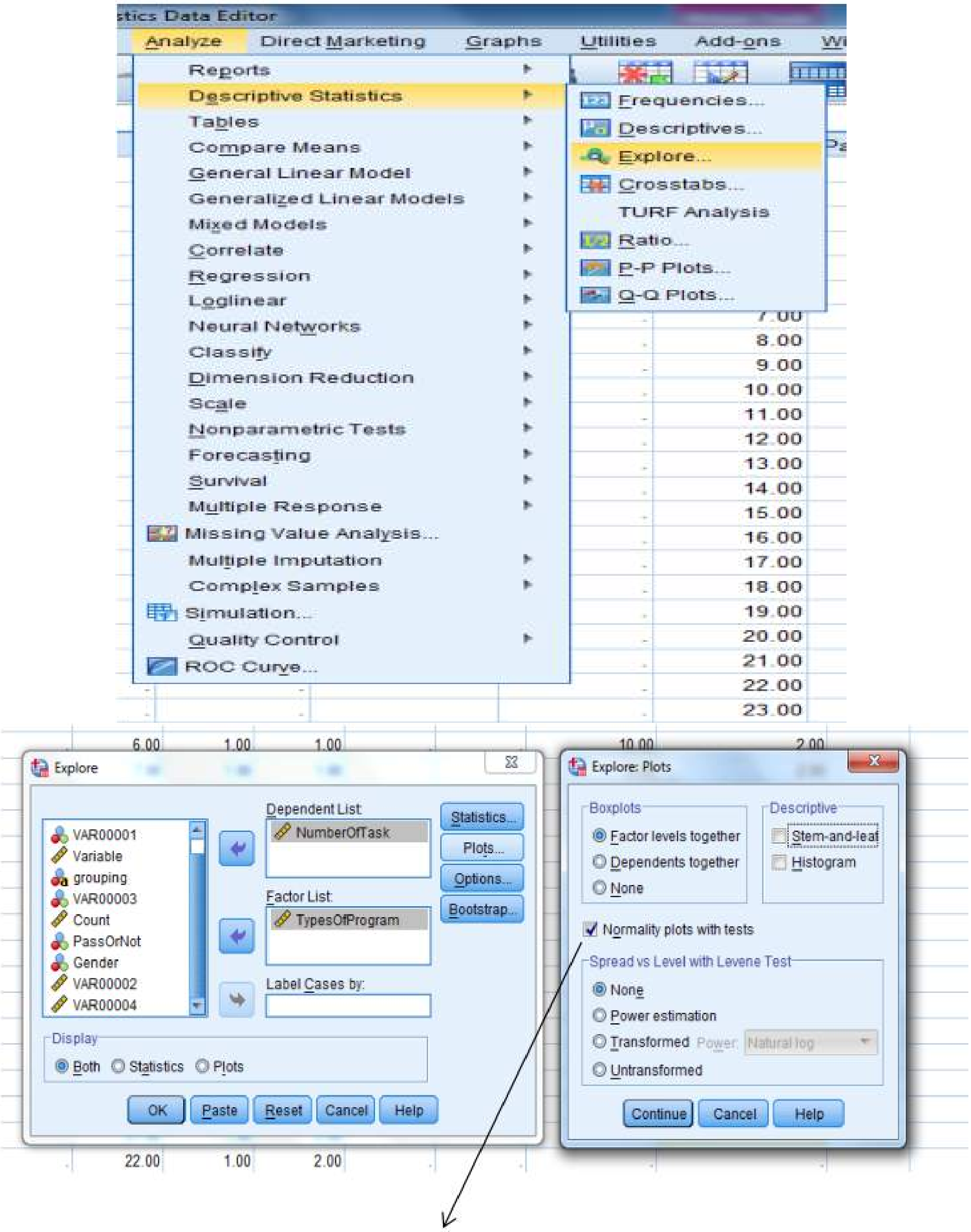
2) **The populations follow a normal distribution :**

$$H_0\text{: the four population follows a normal distribution}$$
$$\textbf{Vs}\ \ H_1\text{: the four population does not follow a normal distribution}$$



NumberOfTask	TypesOfProgram	var
9.00	1.00	
12.00	1.00	
14.00	1.00	
11.00	1.00	
13.00	1.00	
10.00	2.00	
6.00	2.00	
9.00	2.00	
9.00	2.00	
10.00	2.00	
12.00	3.00	
14.00	3.00	
11.00	3.00	
13.00	3.00	
11.00	3.00	
9.00	4.00	
8.00	4.00	
11.00	4.00	
7.00	4.00	
8.00	4.00	
.	.	
.	.	
.	.	
.	.	
.	.	
.	.	
.	.	
.	.	
.	.	

IBM SPSS Statistics Processor is ready



Dependent List: NumberOfTask

Factor List: TypesOfProgram

Display: ● Both ○ Statistics ○ Plots

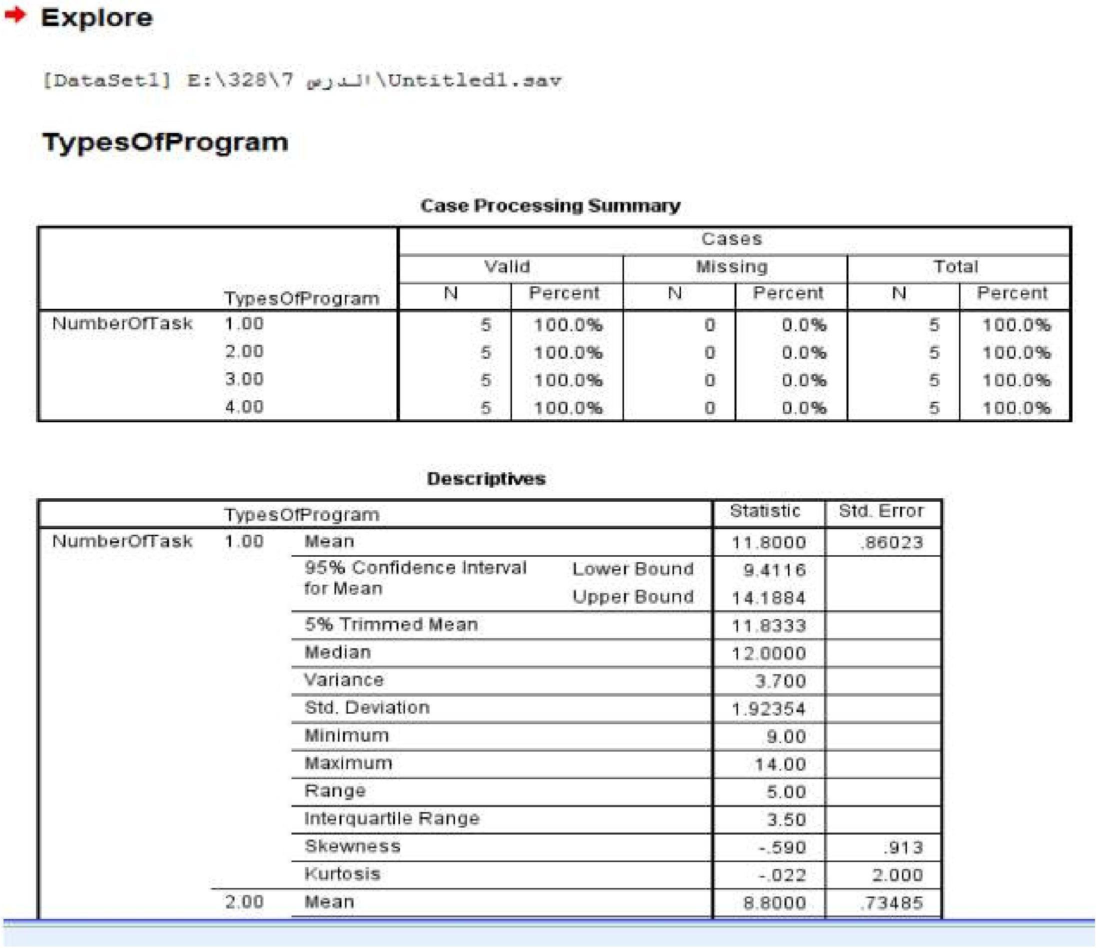
Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☐ Stem-and-leaf ☐ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

Helps in the normality test



Explore

[DataSet1] E:\328\7 الدرس\Untitled1.sav

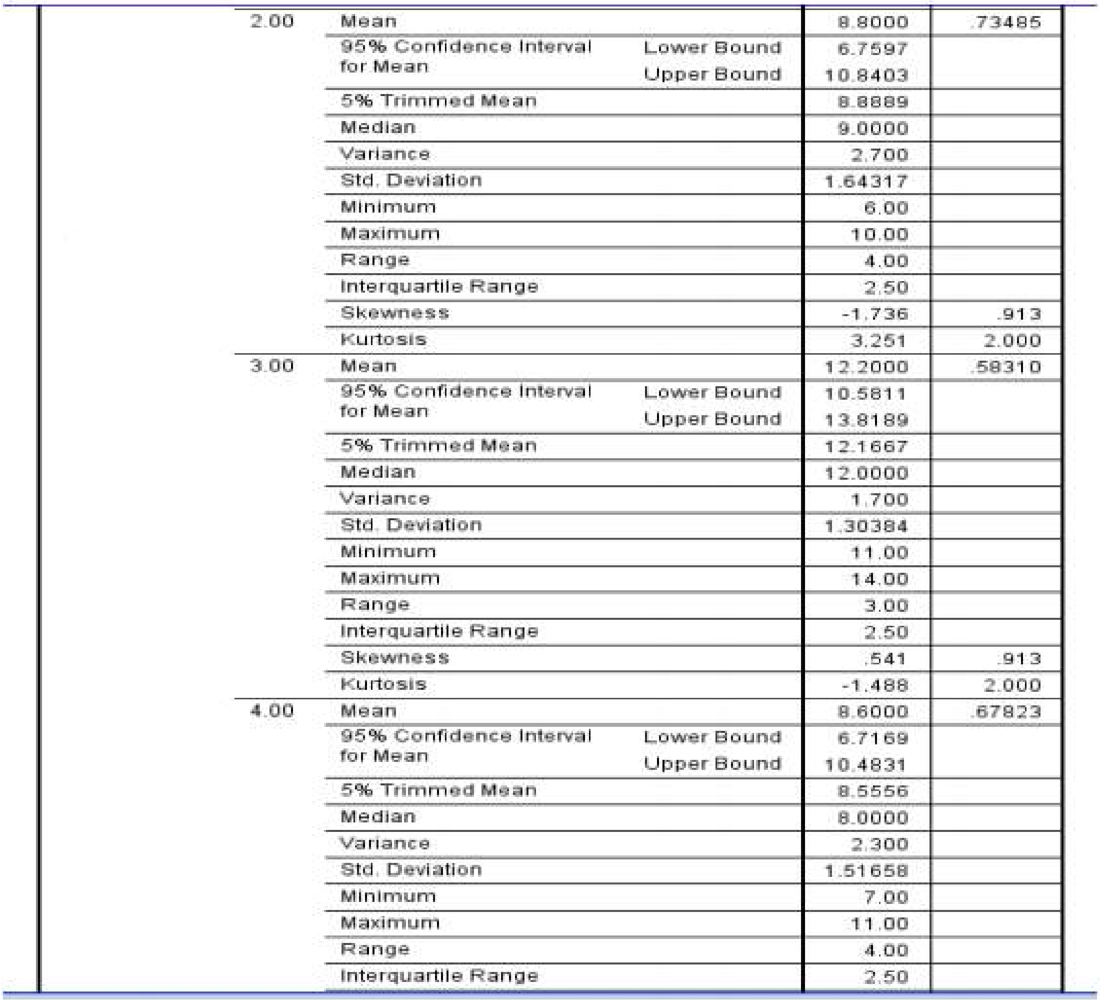
TypesOfProgram

Case Processing Summary

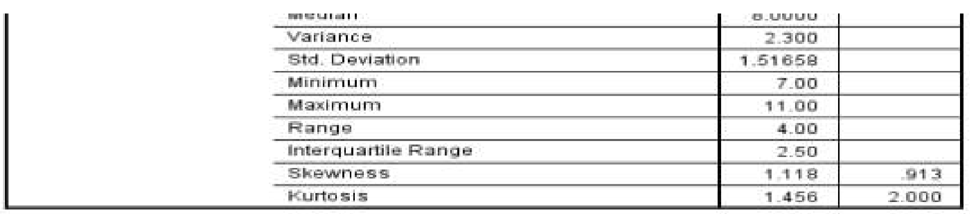
		Cases					
		Valid		Missing		Total	
	TypesOfProgram	N	Percent	N	Percent	N	Percent
NumberOfTask	1.00	5	100.0%	0	0.0%	5	100.0%
	2.00	5	100.0%	0	0.0%	5	100.0%
	3.00	5	100.0%	0	0.0%	5	100.0%
	4.00	5	100.0%	0	0.0%	5	100.0%

Descriptives

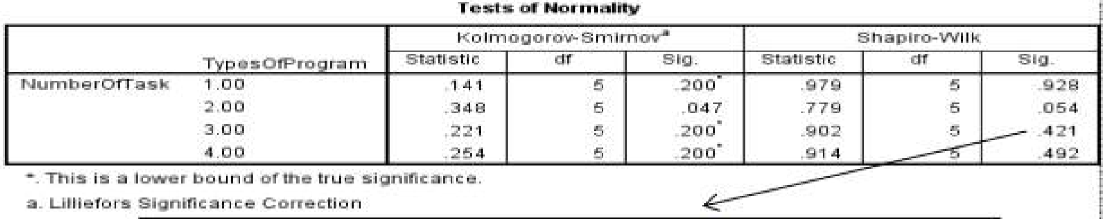
	TypesOfProgram			Statistic	Std. Error
NumberOfTask	1.00	Mean		11.8000	.86023
		95% Confidence Interval for Mean	Lower Bound	9.4116	
			Upper Bound	14.1884	
		5% Trimmed Mean		11.8333	
		Median		12.0000	
		Variance		3.700	
		Std. Deviation		1.92354	
		Minimum		9.00	
		Maximum		14.00	
		Range		5.00	
		Interquartile Range		3.50	
		Skewness		-.590	.913
		Kurtosis		-.022	2.000
	2.00	Mean		8.8000	.73485



2.00	Mean		8.8000	.73485
	95% Confidence Interval for Mean	Lower Bound	6.7597	
		Upper Bound	10.8403	
	5% Trimmed Mean		8.8889	
	Median		9.0000	
	Variance		2.700	
	Std. Deviation		1.64317	
	Minimum		6.00	
	Maximum		10.00	
	Range		4.00	
	Interquartile Range		2.50	
	Skewness		-1.736	.913
	Kurtosis		3.251	2.000
3.00	Mean		12.2000	.58310
	95% Confidence Interval for Mean	Lower Bound	10.5811	
		Upper Bound	13.8189	
	5% Trimmed Mean		12.1667	
	Median		12.0000	
	Variance		1.700	
	Std. Deviation		1.30384	
	Minimum		11.00	
	Maximum		14.00	
	Range		3.00	
	Interquartile Range		2.50	
	Skewness		.541	.913
	Kurtosis		-1.488	2.000
4.00	Mean		8.6000	.67823
	95% Confidence Interval for Mean	Lower Bound	6.7169	
		Upper Bound	10.4831	
	5% Trimmed Mean		8.5556	
	Median		8.0000	
	Variance		2.300	
	Std. Deviation		1.51658	
	Minimum		7.00	
	Maximum		11.00	
	Range		4.00	
	Interquartile Range		2.50	



Median	8.0000	
Variance	2.300	
Std. Deviation	1.51658	
Minimum	7.00	
Maximum	11.00	
Range	4.00	
Interquartile Range	2.50	
Skewness	1.118	.913
Kurtosis	1.456	2.000



Tests of Normality

		Kolmogorov-Smirnov^a			Shapiro-Wilk		
	TypesOfProgram	Statistic	df	Sig.	Statistic	df	Sig.
NumberOfTask	1.00	.141	5	.200*	.979	5	.928
	2.00	.348	5	.047	.779	5	.054
	3.00	.221	5	.200*	.902	5	.421
	4.00	.254	5	.200*	.914	5	.492

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

a. Lilliefors Significance Correction

As P – value > .05 for the four populations.

So, we except H_0: *the four populations follow* a normal distribution

3) Homogeneity of Variance (to get a test of the assumption of homogeneity of variance) i.e.

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

i.e. the variances of each sample are equal

$$Vs$$

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

This will be clear later.

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

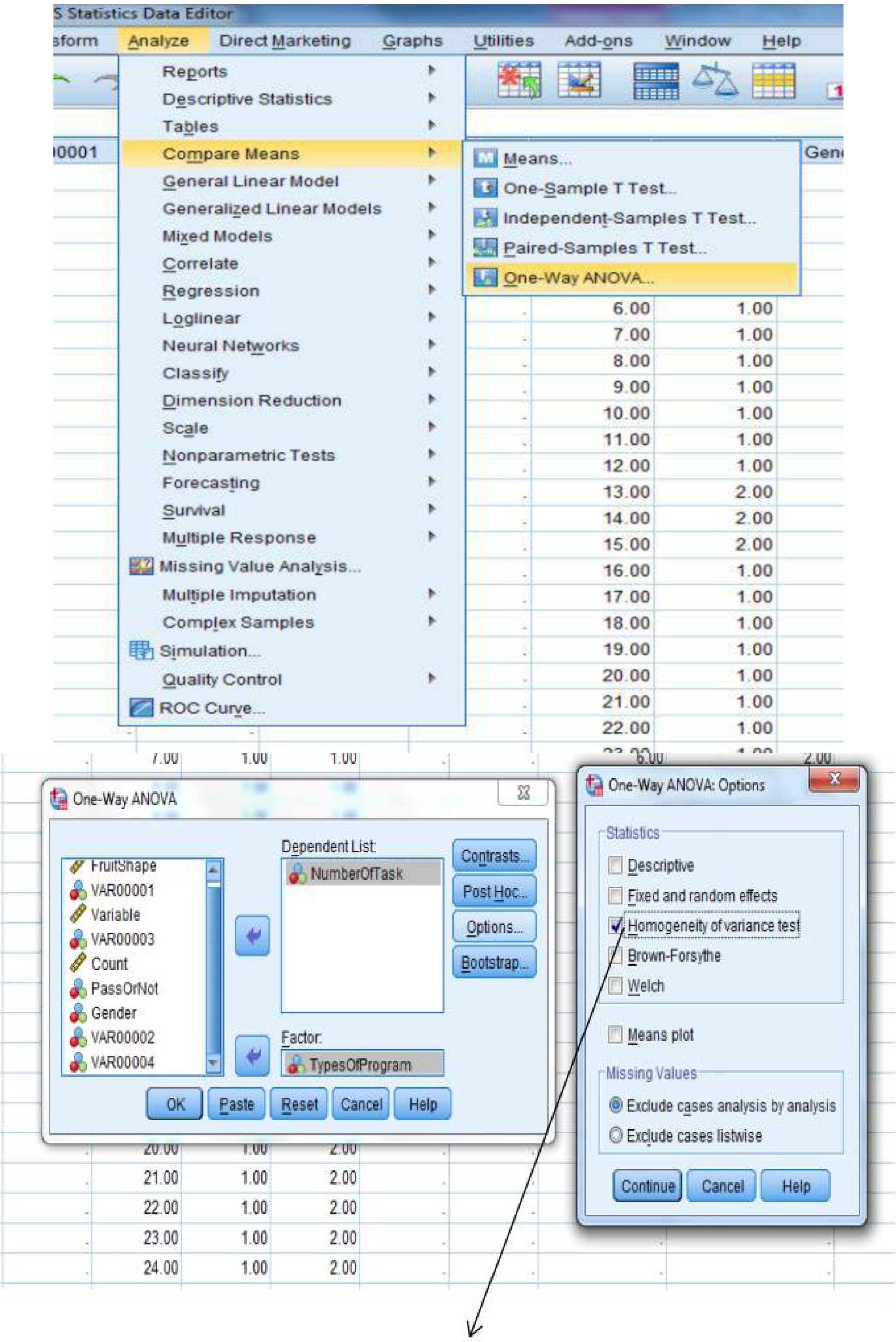
$$H_0: \mu_{program\ 1} = \mu_{program\ 2} = \mu_{program\ 3} = \mu_{program\ 4}$$

i.e. treatments are equally effective

$$Vs$$

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

$$at\ \alpha = .05$$



Dependent List: NumberOfTask

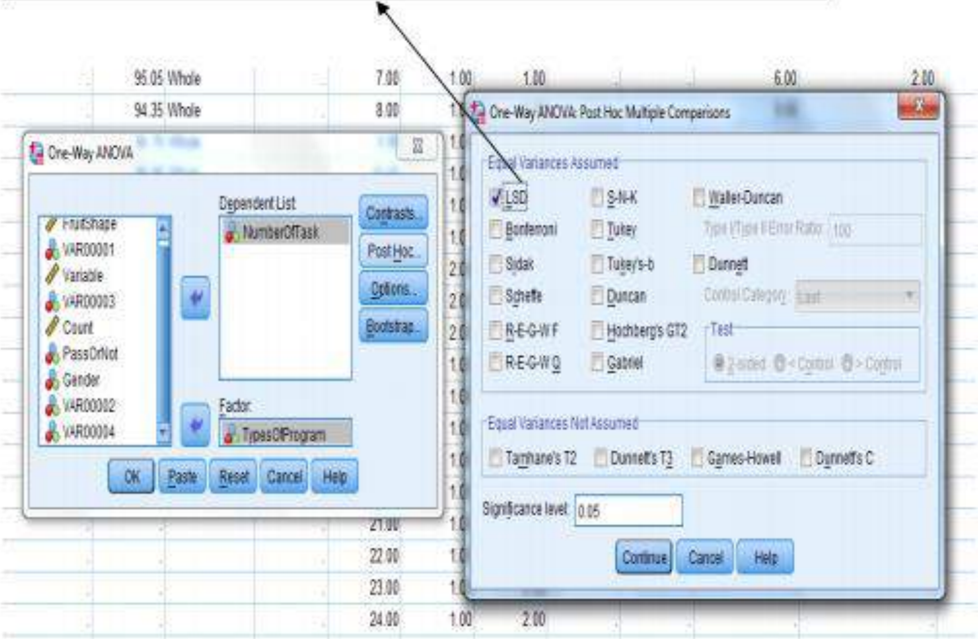
Factor: TypesOfProgram

☑ Homogeneity of variance test

● Exclude cases analysis by analysis

Helps in the homogeneity of variance test

If we reject H₀ in Analysis of Variance (ANOVA one way-test) we need to look at the multiple comparisons output by use the appropriate post hoc procedure (LSD) to determine whether unique pairwise comparisons are significant.



Dependent List: NumberOfTask

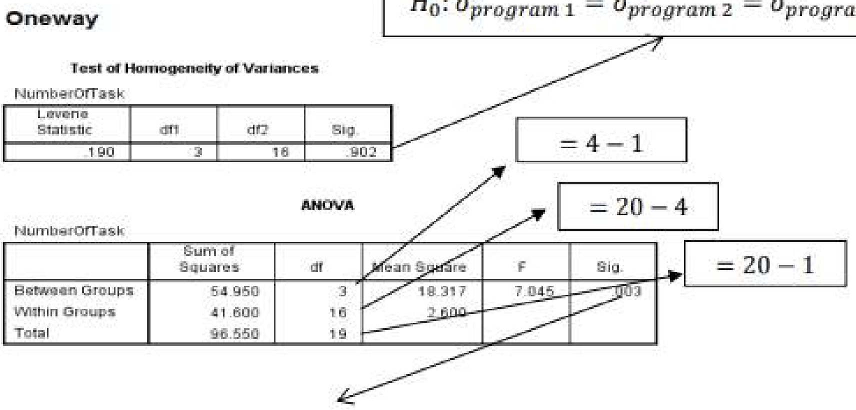
Factor: TypesOfProgram

☑ LSD

Significance level: 0.05

As P – value > .05 .So, we except

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



Oneway

Test of Homogeneity of Variances

NumberOfTask

Levene Statistic	df1	df2	Sig.
.190	3	16	.902

$= 4 - 1$

$= 20 - 4$

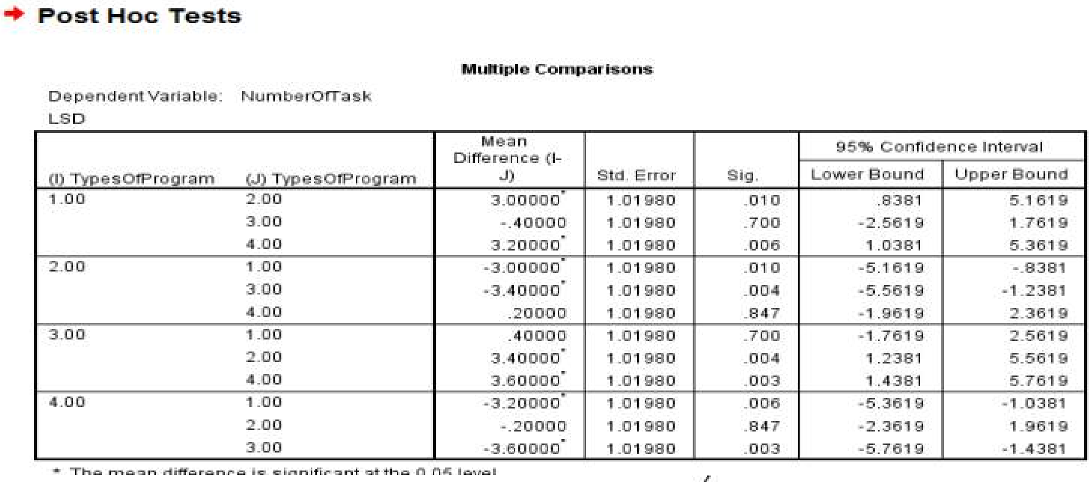
ANOVA

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			

$= 20 - 1$

as $P - value < .05$,then we reject $H_0: \mu_{program\,1} = \mu_{program\,2} = \mu_{program\,3} = \mu_{program\,4}$.



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
1.00	2.00	3.00000*	1.01980	.010	.8381	5.1619
	3.00	-.40000	1.01980	.700	-2.5619	1.7619
	4.00	3.20000*	1.01980	.006	1.0381	5.3619
2.00	1.00	-3.00000*	1.01980	.010	-5.1619	-.8381
	3.00	-3.40000*	1.01980	.004	-5.5619	-1.2381
	4.00	.20000	1.01980	.847	-1.9619	2.3619
3.00	1.00	.40000	1.01980	.700	-1.7619	2.5619
	2.00	3.40000*	1.01980	.004	1.2381	5.5619
	4.00	3.60000*	1.01980	.003	1.4381	5.7619
4.00	1.00	-3.20000*	1.01980	.006	-5.3619	-1.0381
	2.00	-.20000	1.01980	.847	-2.3619	1.9619
	3.00	-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 $\mathrm{P} - \mathrm{value} = .7 > .05$, 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 $\mathrm{P} - \mathrm{value} = .847 > .05$, 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.

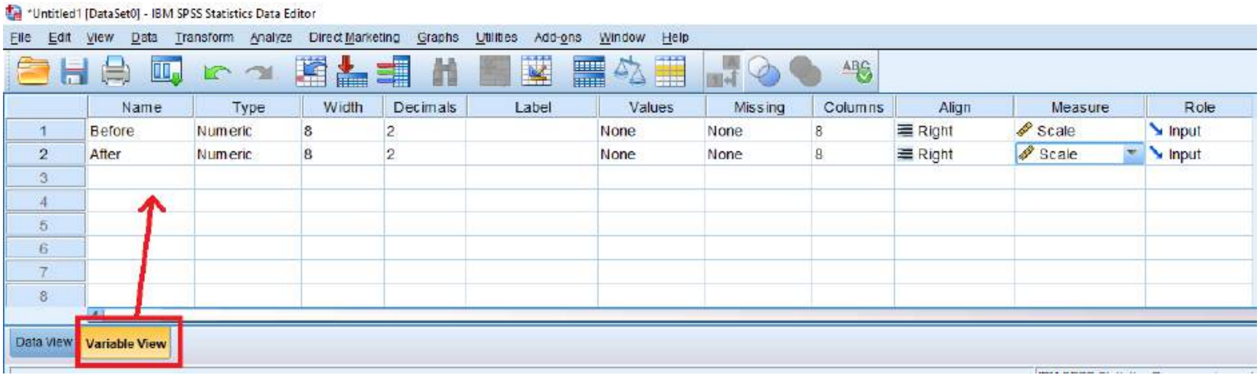
Q4: In a study of a surgical procedure used to decrease the amount of food that person can eat. A sample of 10 persons measures their weights before and after one year of the surgery, we obtain the following data:

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**

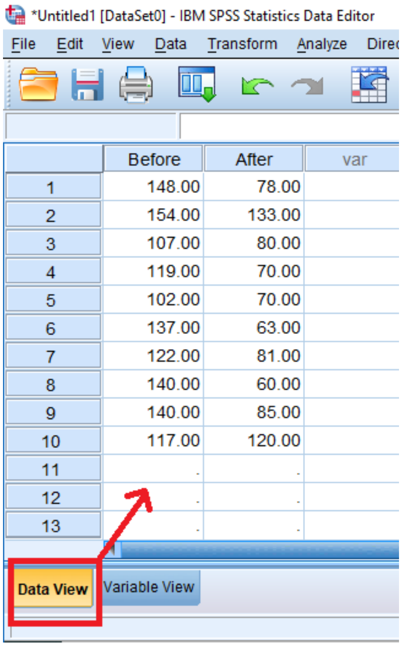
We assume that the data comes from normal distribution. Find :

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 $\mu_D \neq 0$)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Before	Numeric	8	2		None	None	8	Right	Scale	Input
2	After	Numeric	8	2		None	None	8	Right	Scale	Input
3											
4											
5											
6											
7											
8											



	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	.	.	

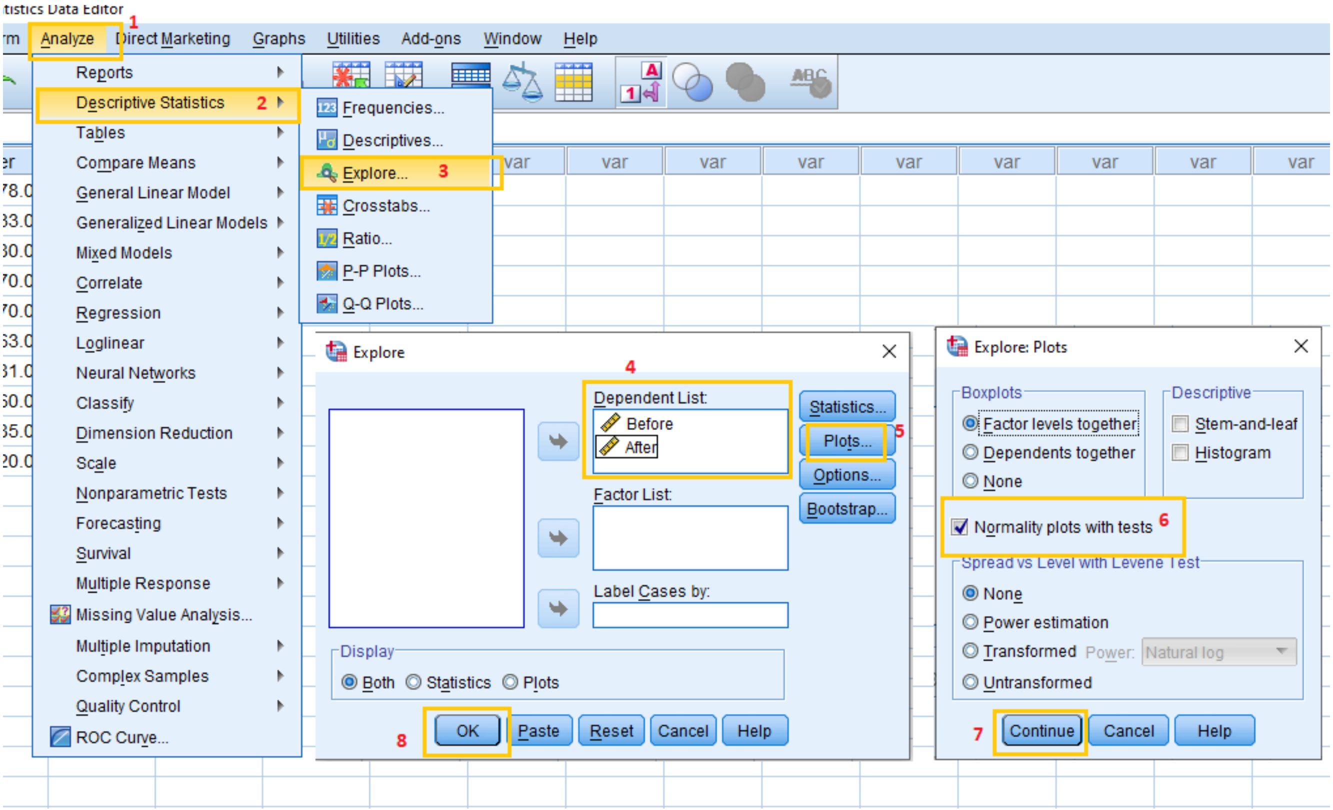
To use the **Paired-Samples T-Test**, we need to make sure that the population follows a normal distribution:

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

Vs

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

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



Dependent List: Before, After

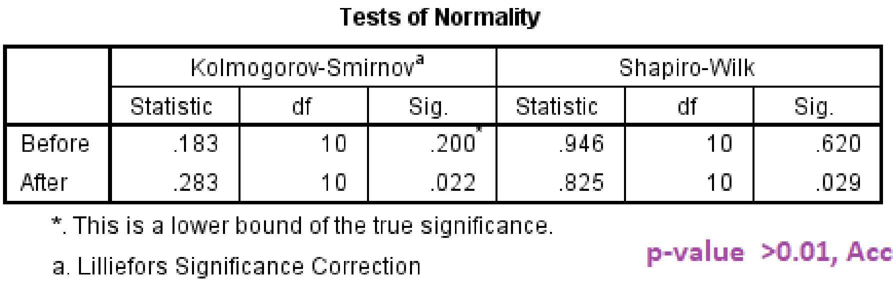
Display: ● Both ○ Statistics ○ Plots

Boxplots: ● Factor levels together ○ Dependents together ○ None

Descriptive: ☐ Stem-and-leaf ☐ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed ○ Untransformed



Tests of Normality

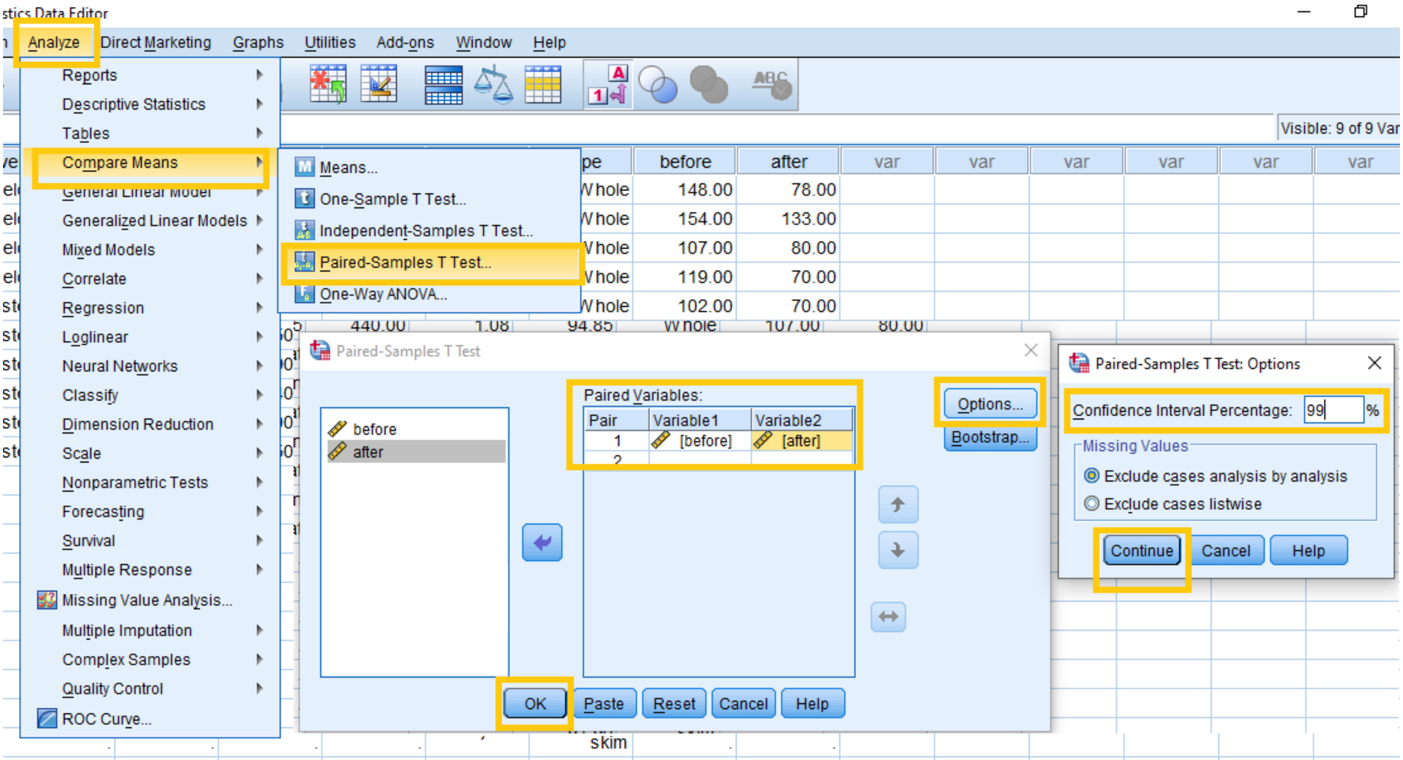
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Before	.183	10	.200*	.946	10	.620
After	.283	10	.022	.825	10	.029

*. 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 $\mu_D \neq 0$:



Paired Variables:

Pair	Variable1	Variable2
1	[before]	[after]
2		

Confidence Interval Percentage: 99 %

Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

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

		Paired Differences							
D=before –after		$\bar{D}$	SD_D	Std. Error Mean	99% Confidence Interval of the Difference				
		Mean	Std. Deviation		Lower	Upper	t	df	Sig. (2-tailed)
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

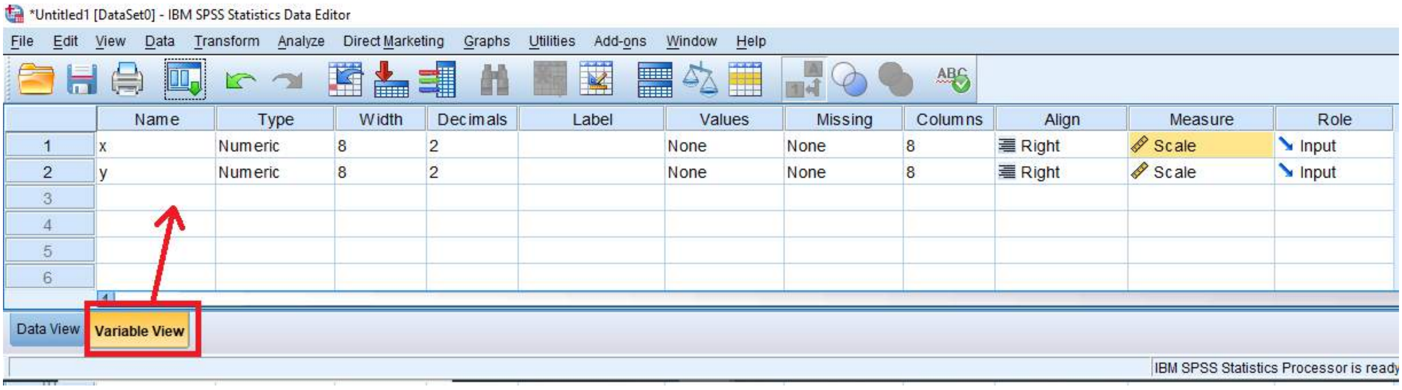
a) Compute and interpret the linear correlation coefficient, r.

b) Determine the regression equation for the data.

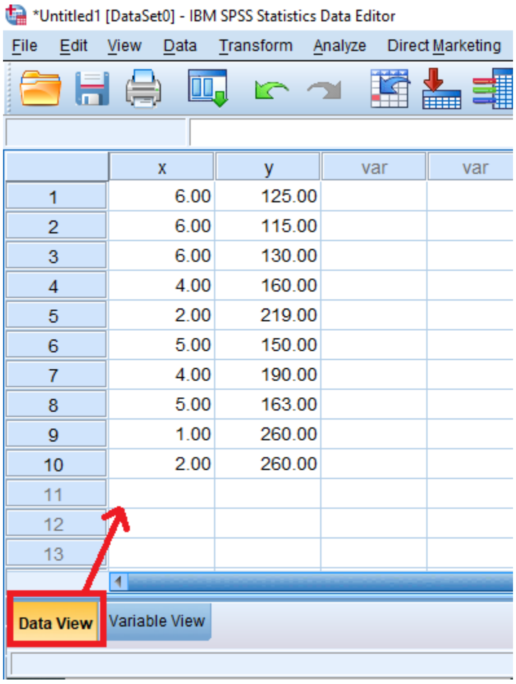
c) Compute and interpret the coefficient of determination, r^2.

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).

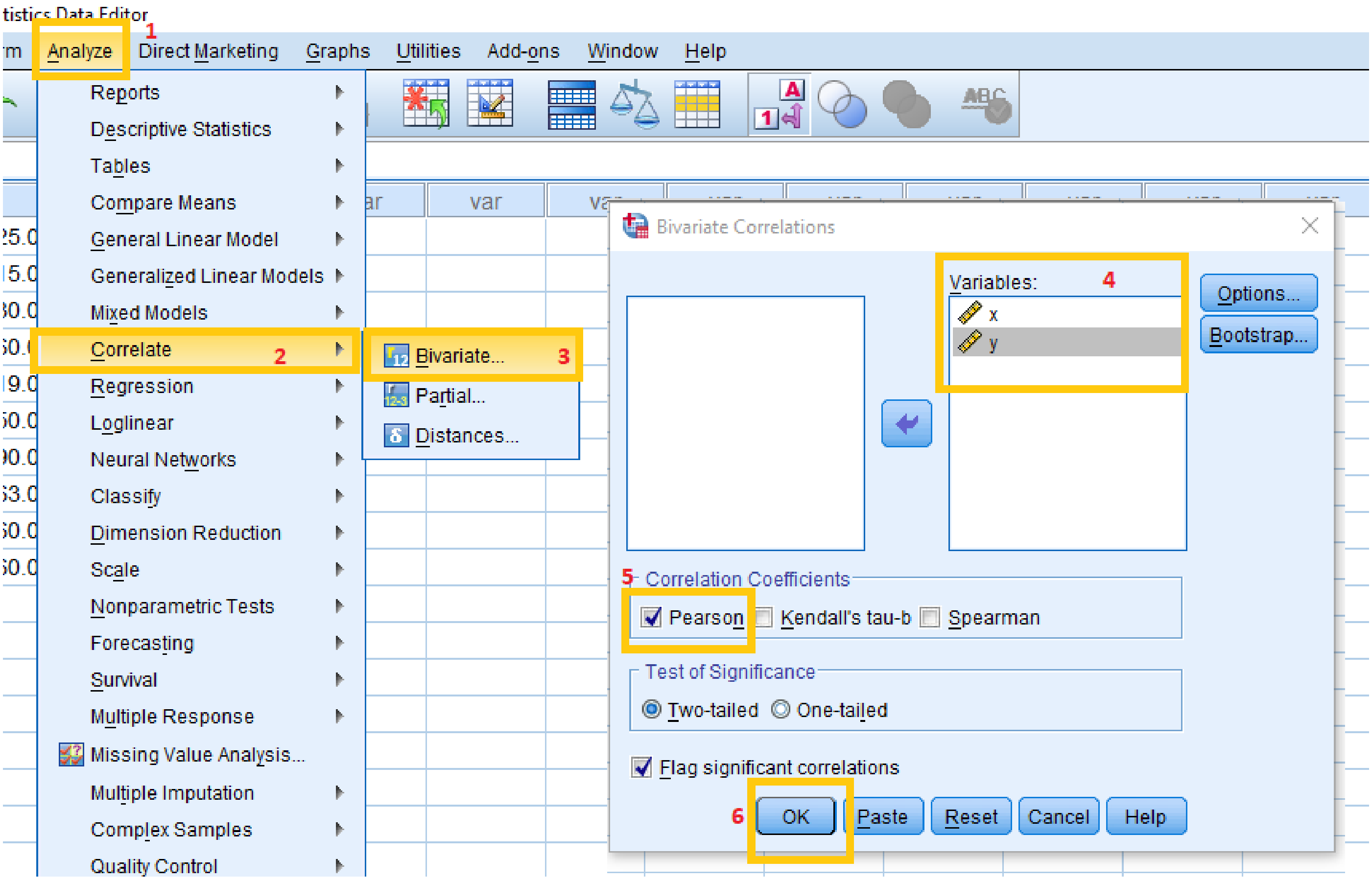


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	x	Numeric	8	2		None	None	8	Right	Scale	Input
2	y	Numeric	8	2		None	None	8	Right	Scale	Input
3											
4											
5											
6											



	x	y
1	6.00	125.00
2	6.00	115.00
3	6.00	130.00
4	4.00	160.00
5	2.00	219.00
6	5.00	150.00
7	4.00	190.00
8	5.00	163.00
9	1.00	260.00
10	2.00	260.00
11		
12		
13		

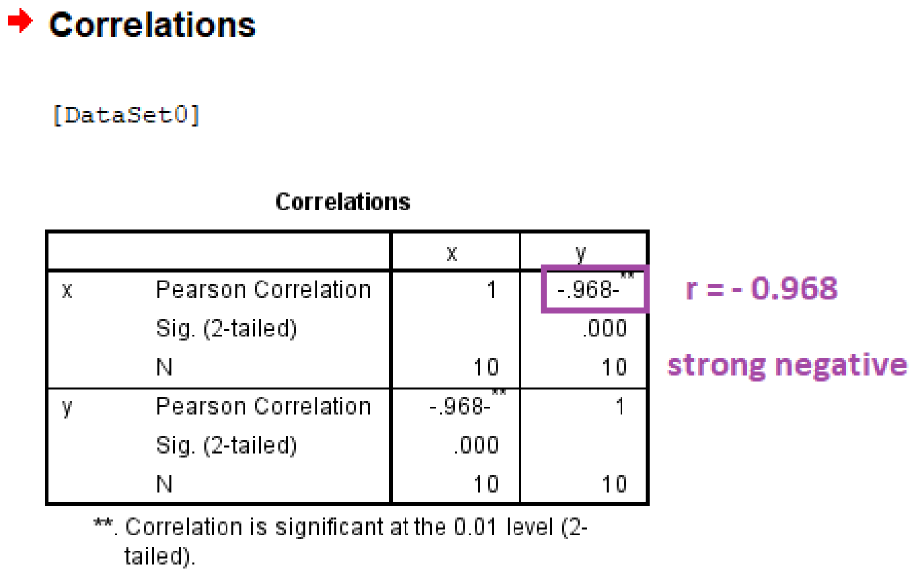
a) Select Analyze → Correlate → Bivariate... (see figure, below).



☑ Pearson ☐ Kendall's tau-b ☐ Spearman

● Two-tailed ○ One-tailed

☑ Flag significant correlations



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

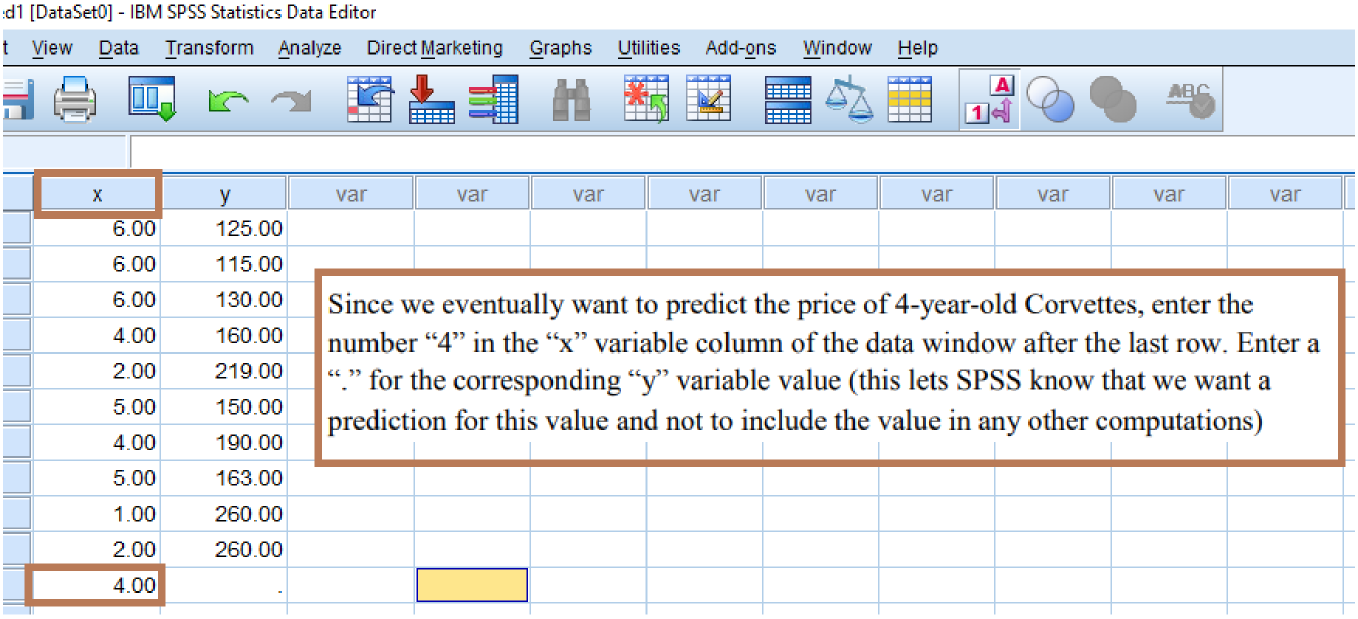
**. Correlation is significant at the 0.01 level (2-tailed).

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.

b, c and d)

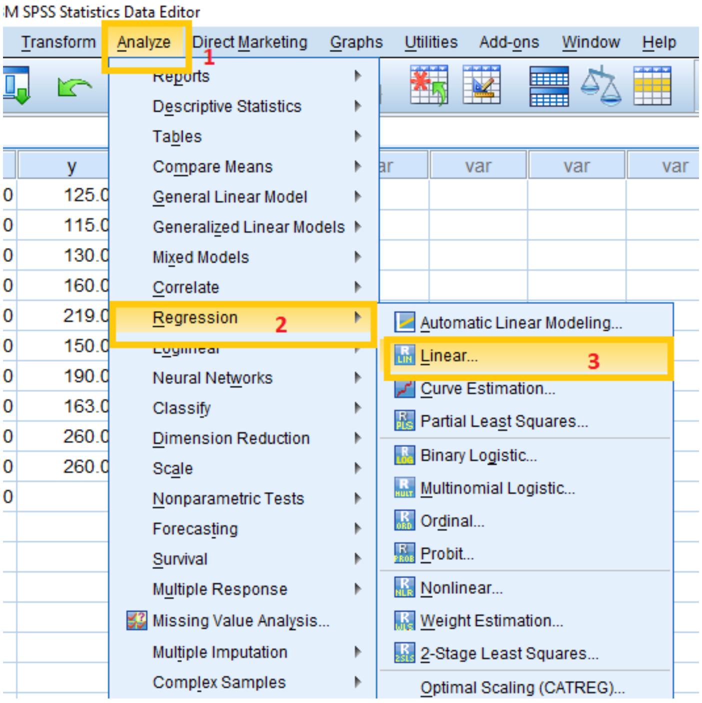


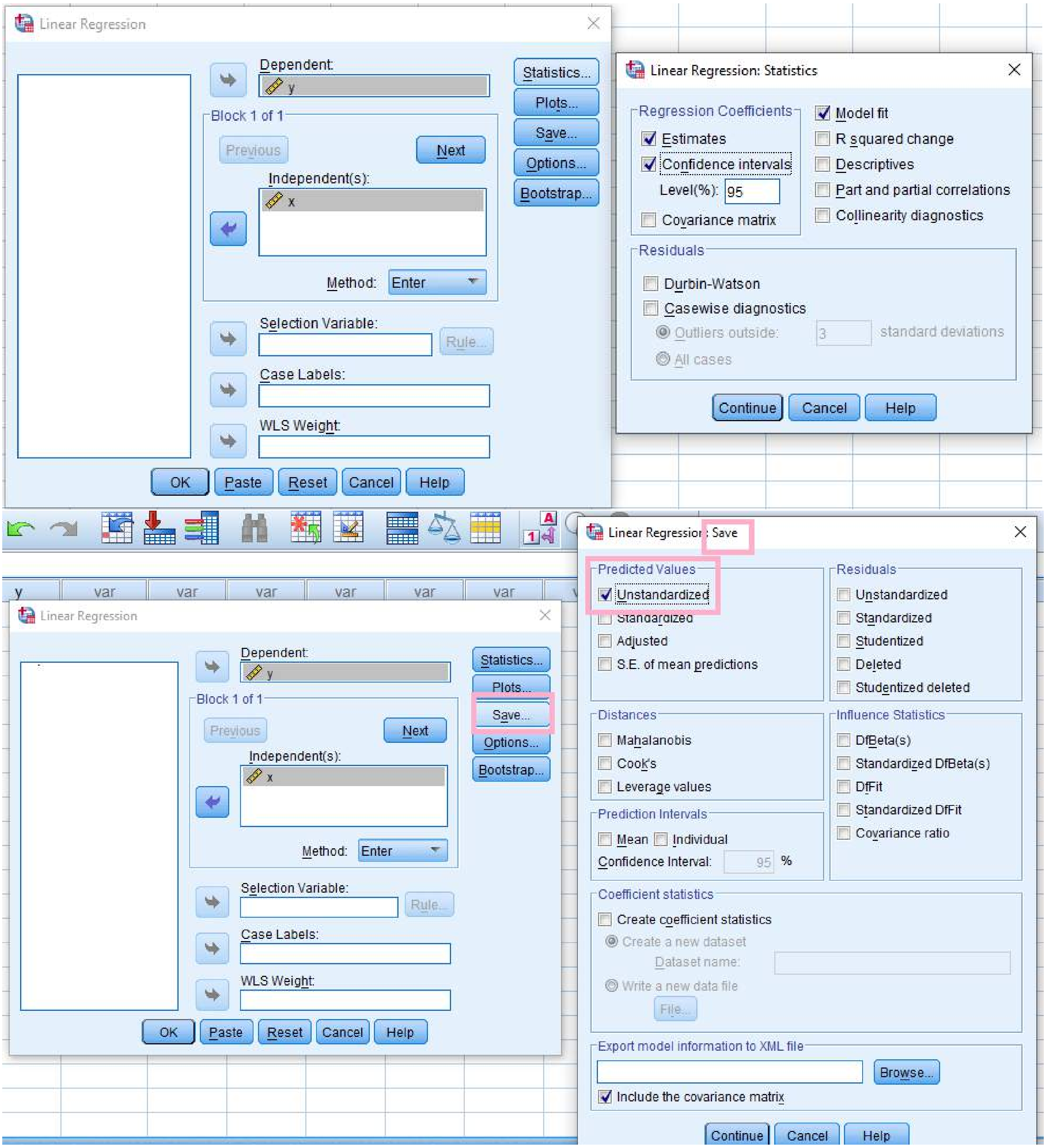
x	y
6.00	125.00
6.00	115.00
6.00	130.00
4.00	160.00
2.00	219.00
5.00	150.00
4.00	190.00
5.00	163.00
1.00	260.00
2.00	260.00
4.00	.

Since we eventually want to predict the price of 4-year-old Corvettes, enter the number “4” in the “x” variable column of the data window after the last row. Enter a “.” for the corresponding “y” variable value (this lets SPSS know that we want a prediction for this value and not to include the value in any other computations)

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".





Linear Regression

Dependent: y

Independent(s): x

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

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

Linear Regression: Save

Predicted Values
- ☑ Unstandardized
- ☐ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

Residuals
- ☐ Unstandardized
- ☐ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

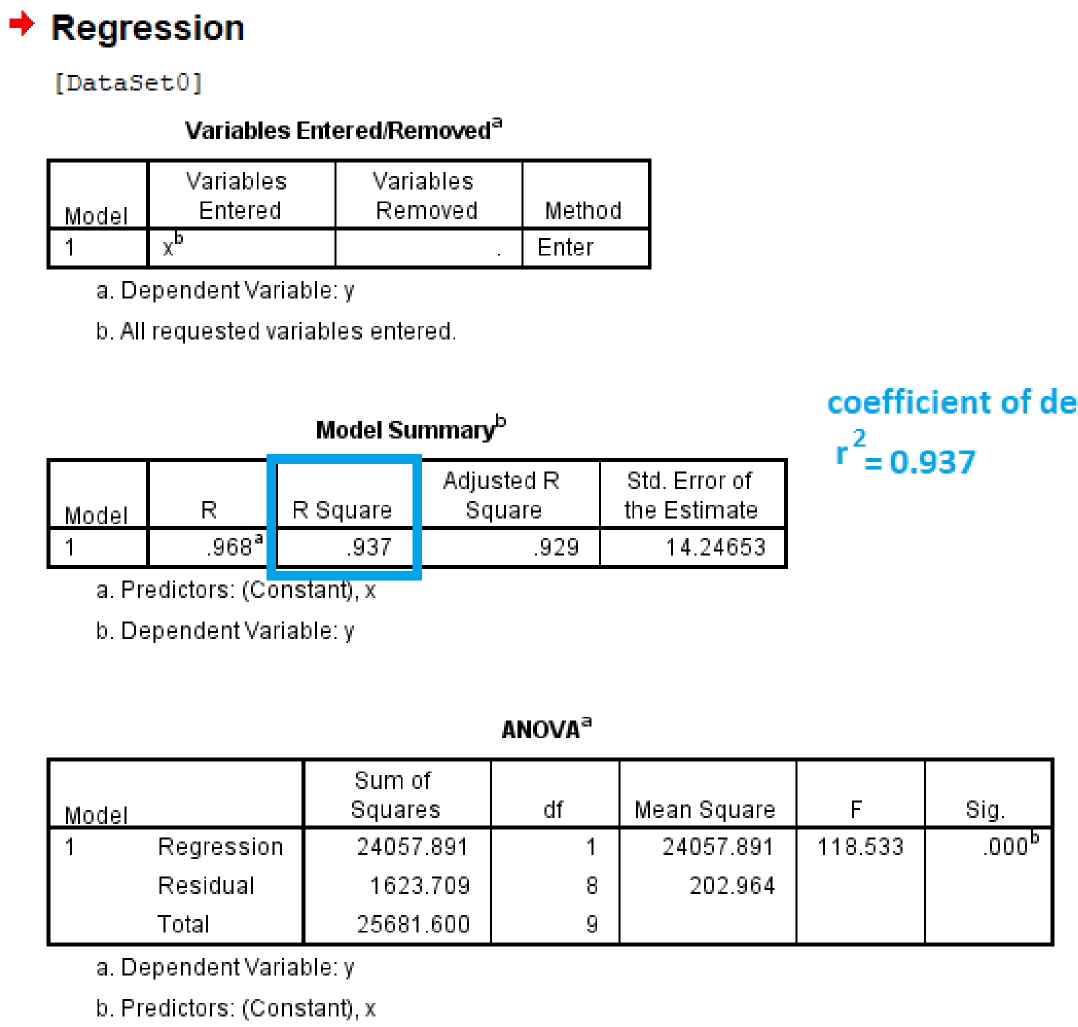
Distances
- ☐ Mahalanobis
- ☐ Cook's
- ☐ Leverage values

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



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 :

$r^2 = 0.937$

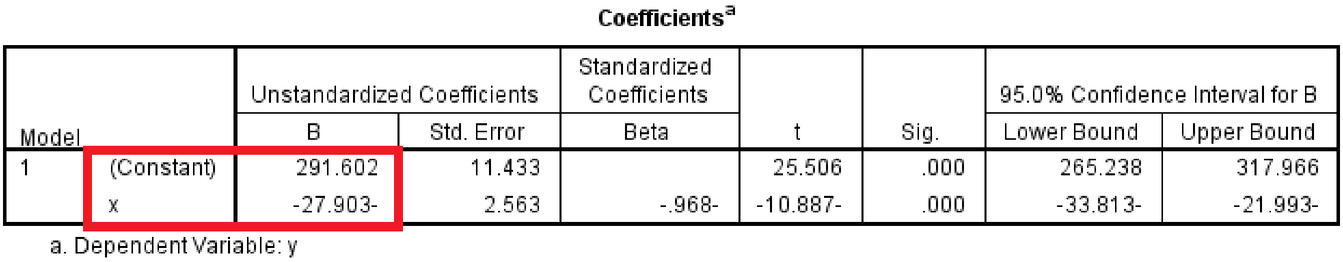
ANOVA[a]

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]

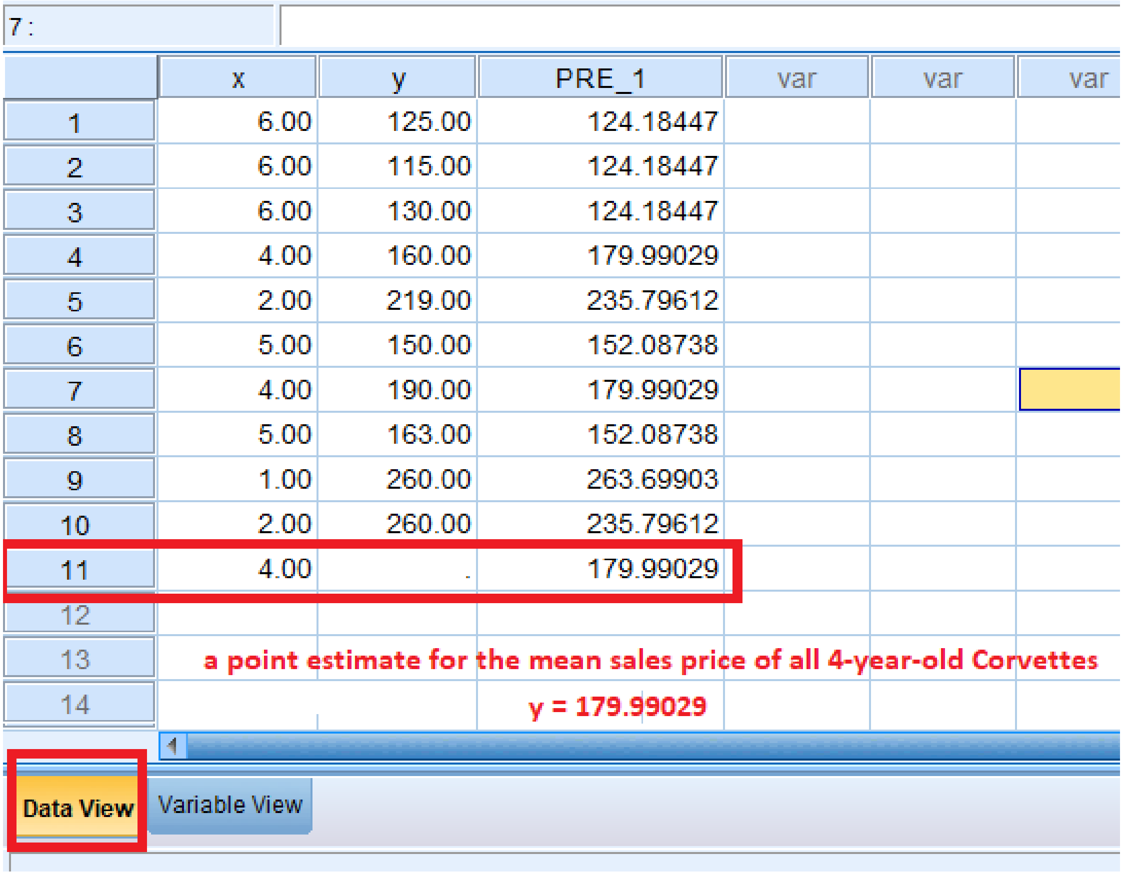


Coefficients[a]

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 : $y = 291.602 - 27.903\, x$



	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

- **From above, the regression equation is: y = 29160.1942 – (2790.2913)(x).**

- **The coefficient of determination is 0.9368; therefore, about 93.68% of the variation in y data is explained by x.**

Department of Statistics and Operations Research

College of Science

King Saud University

Tutorial

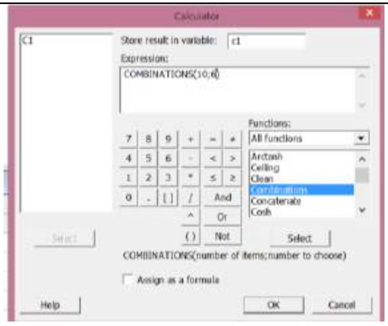
STATISTICAL PACKAGES(Minitab)

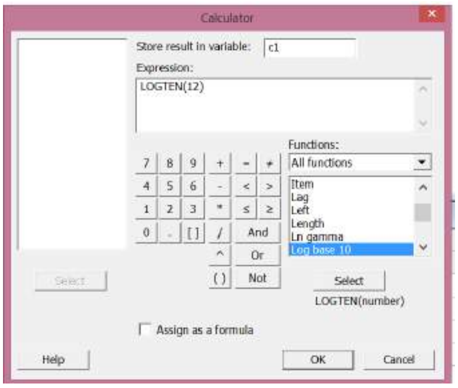
STAT 328

Minitab Statistical
Software

MATHEMATICAL FUNCTIONS

Write the commands of the following:

		By Minitab: **calc → calculator**
Absolute value	**$\|-4\|=4$**	 Store result in variable: c1 Expression: ABS(-4)
Combinations	**$\binom{10}{6}$=10C6=210**	 Store result in variable: c1 Expression: COMBINATIONS(10;6) COMBINATIONS(number of items;number to choose)
The exponential function	**$e^{-1.6}$=0.201897**	H.W
Factorial	**11! =39916800**	 Store result in variable: c2 Expression: FACTORIAL(11) FACTORIAL(number of items)

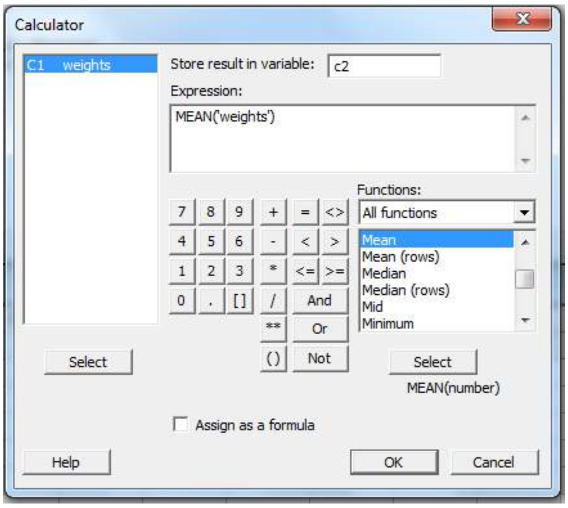
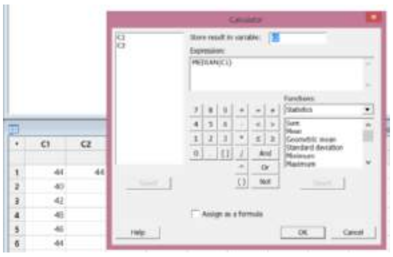
Floor function	$[-3.15]$= -4	 Store result in variable: c2 Expression: FLOOR(-3.15;0) Functions: All functions
Natural logarithm	ln(23)= 3.135494216	 Store result in variable: c2 Expression: LN(23) Functions: All functions Natural log (log base e) LN(number)
Logarithm with respect to any base	$\log_9(4)$ = 0.630929754	H.W
Logarithm with respect to base 10	log(12) = 1.079181246	 Store result in variable: c1 Expression: LOGTEN(12) Functions: All functions Log base 10 LOGTEN(number)
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:

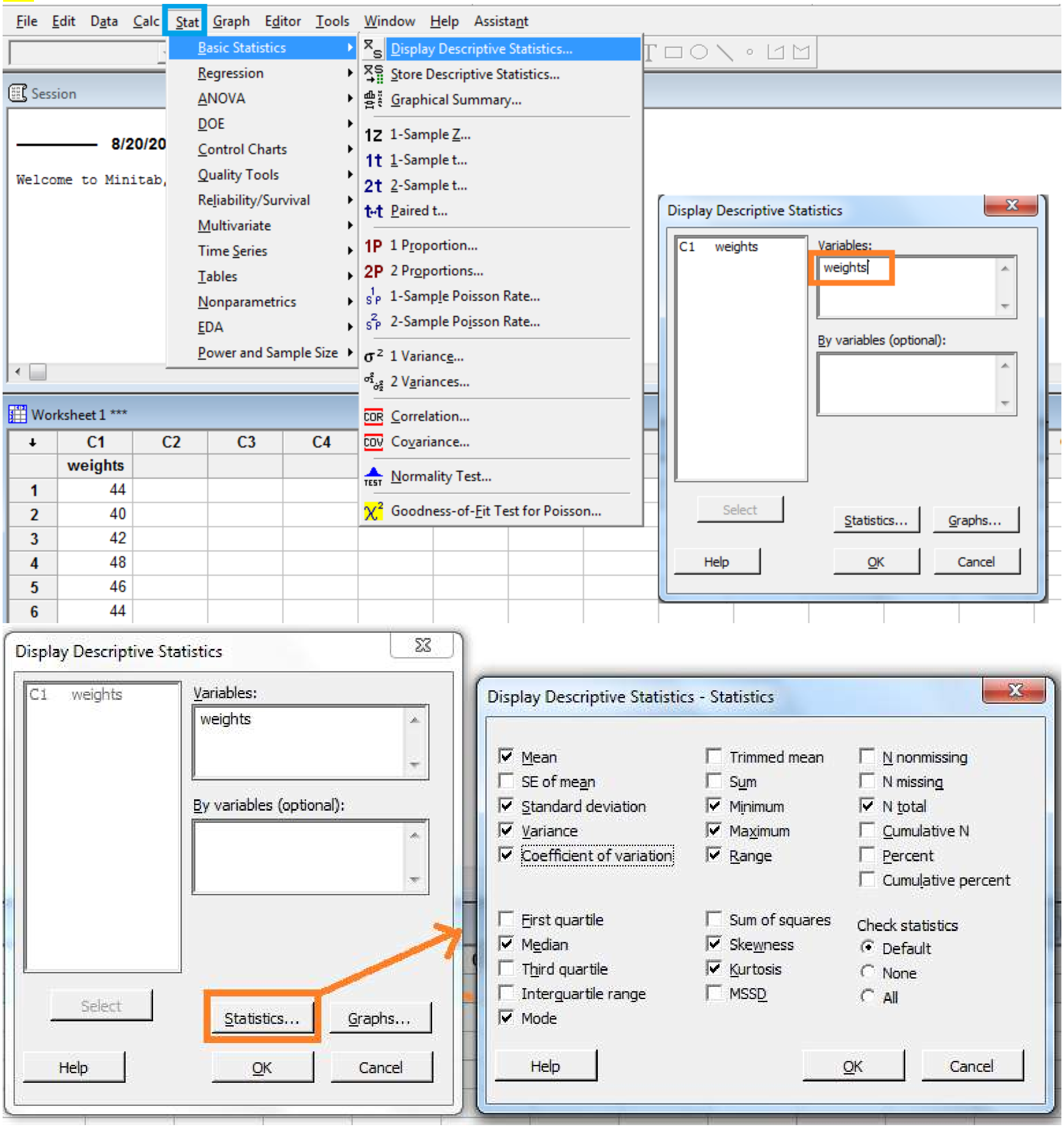
44 , 40 , 42 , 48 , 46 , 44

Find:

	By Minitab **calc → calculator**
Mean=44	 Store result in variable: c2 Expression: MEAN('weights') Functions: All functions
Median=44	 Expression: MEDIAN(C1)

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



	C1	C2	C3	C4
	weights			
1	44			
2	40			
3	42			
4	48			
5	46			
6	44			

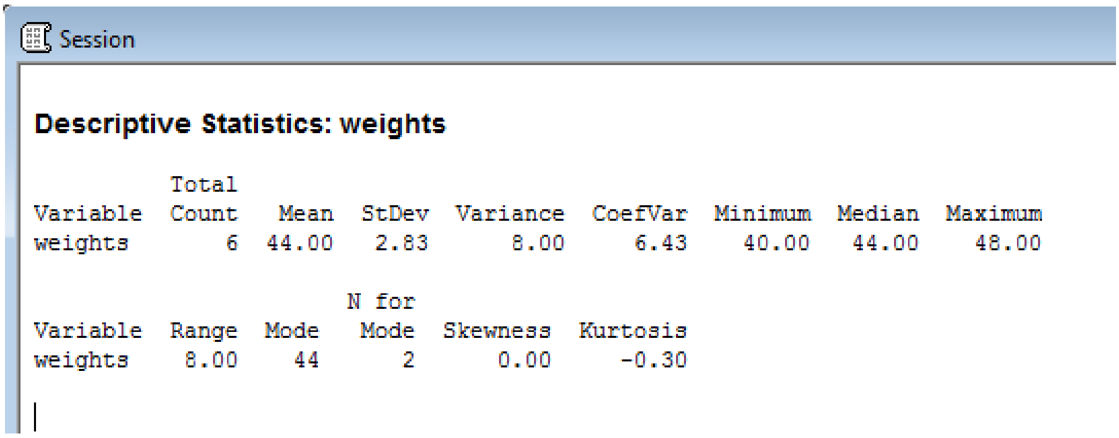
C1 weights

Variables: weights

By variables (optional):

☑ Mean
☐ SE of mean
☑ Standard deviation
☑ Variance
☑ Coefficient of variation
☐ First quartile
☑ Median
☐ Third quartile
☐ Interquartile range
☑ Mode
☐ Trimmed mean
☐ Sum
☑ Minimum
☑ Maximum
☑ Range
☐ Sum of squares
☑ Skewness
☑ Kurtosis
☐ MSSD
☐ N nonmissing
☐ N missing
☑ N total
☐ Cumulative N
☐ Percent
☐ Cumulative percent

Check statistics
● Default
○ None
○ All

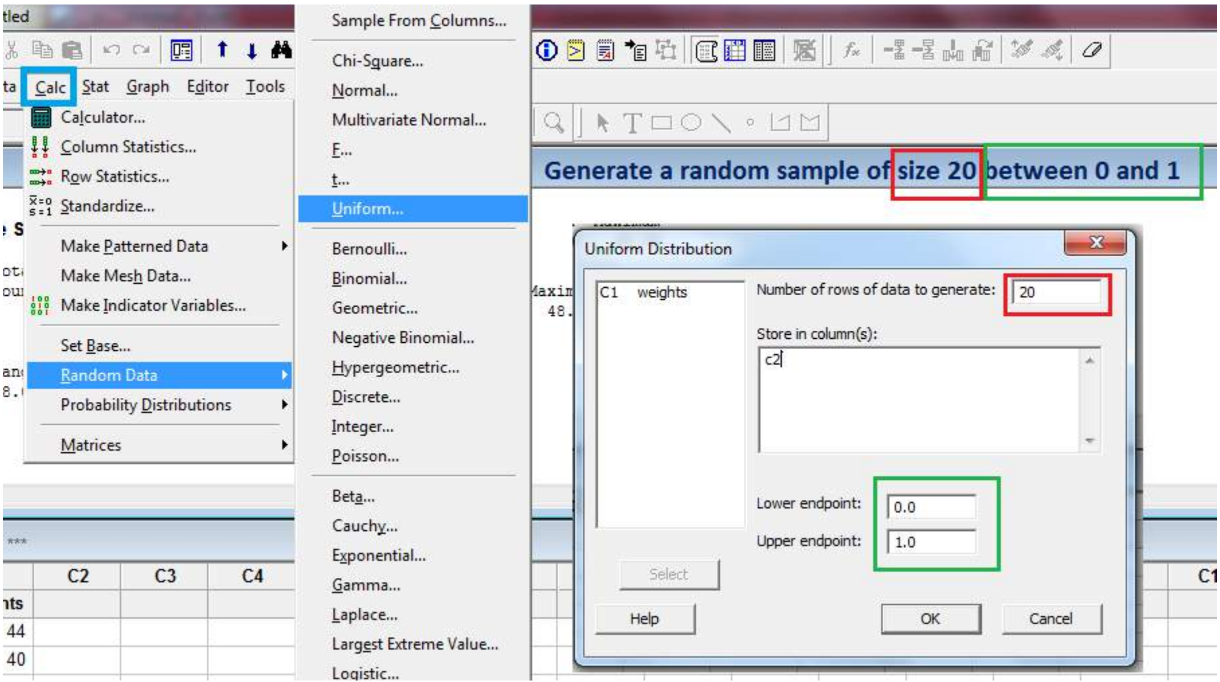


Descriptive Statistics: weights

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

                       N for
Variable   Range  Mode   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



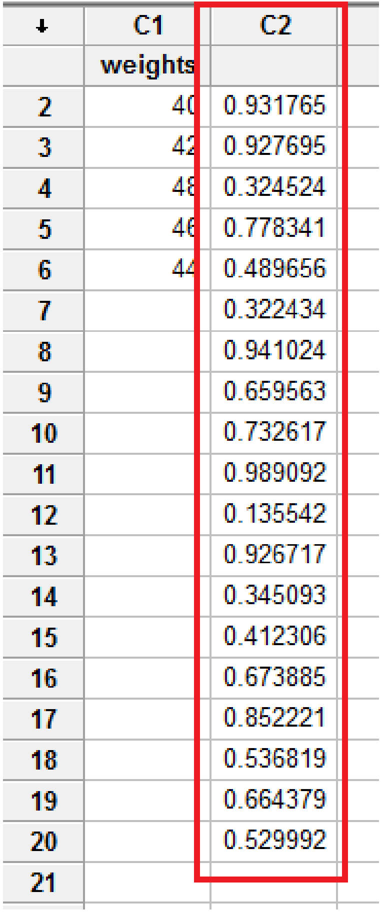
Generate a random sample of size 20 between 0 and 1

Number of rows of data to generate: 20

Store in column(s): c2

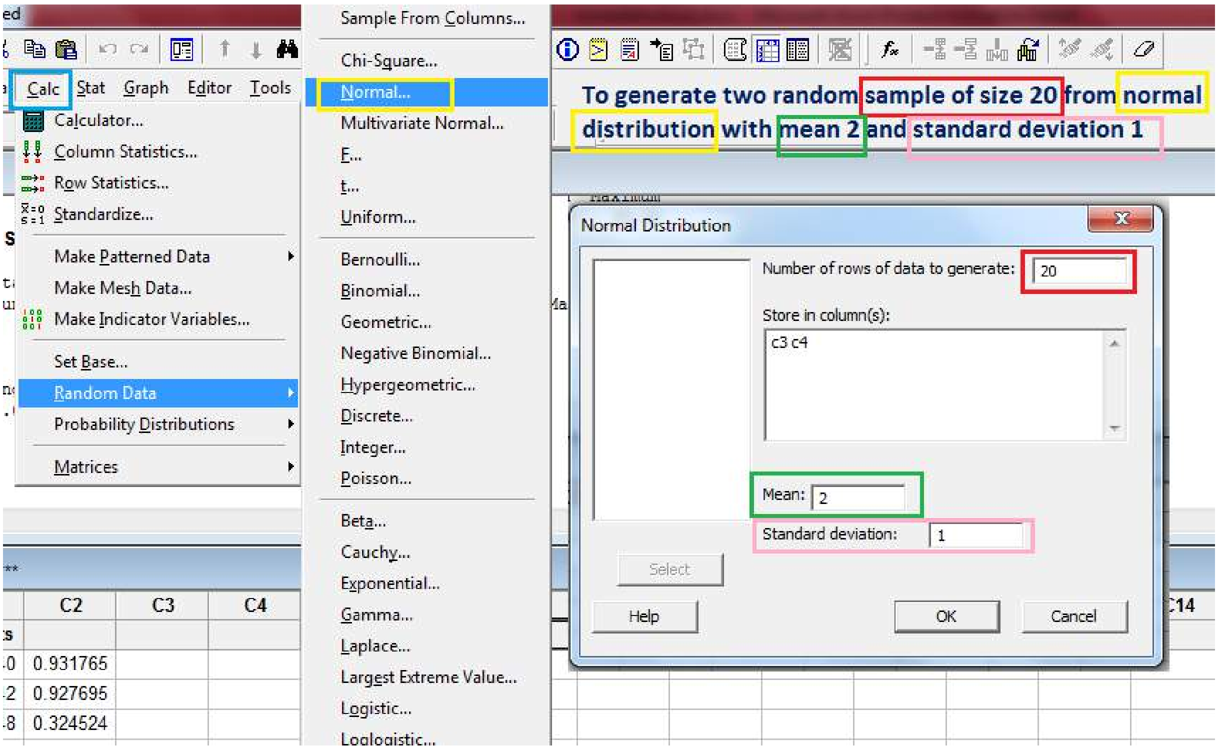
Lower endpoint: 0.0

Upper endpoint: 1.0



↓	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



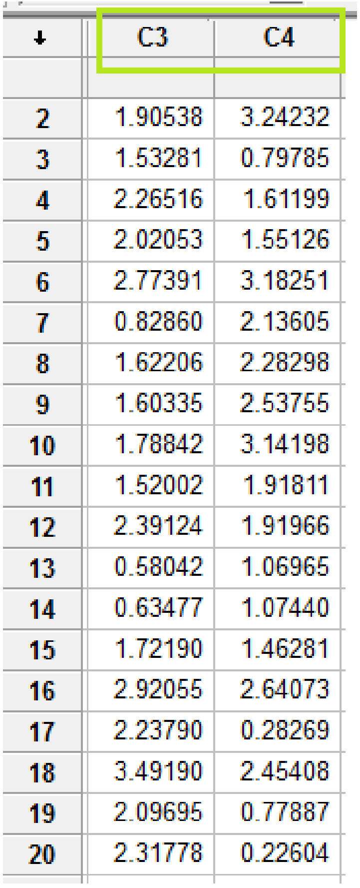
To generate two random sample of size 20 from normal distribution with mean 2 and standard deviation 1

Number of rows of data to generate: 20

Store in column(s): c3 c4

Mean: 2

Standard deviation: 1

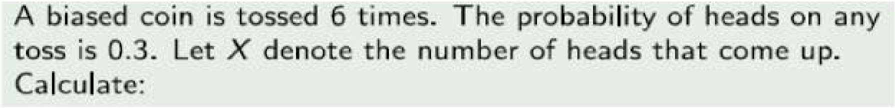


↓	C3	C4
2	1.90538	3.24232
3	1.53281	0.79785
4	2.26516	1.61199
5	2.02053	1.55126
6	2.77391	3.18251
7	0.82860	2.13605
8	1.62206	2.28298
9	1.60335	2.53755
10	1.78842	3.14198
11	1.52002	1.91811
12	2.39124	1.91966
13	0.58042	1.06965
14	0.63477	1.07440
15	1.72190	1.46281
16	2.92055	2.64073
17	2.23790	0.28269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

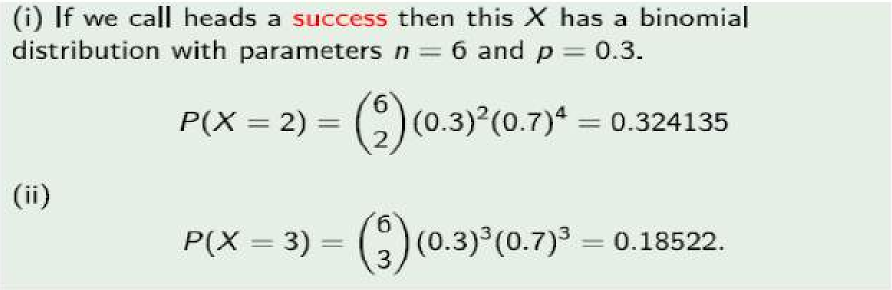
PROBABILITY DISTRIBUTION FUNCTIONS

Discrete Distributions

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:

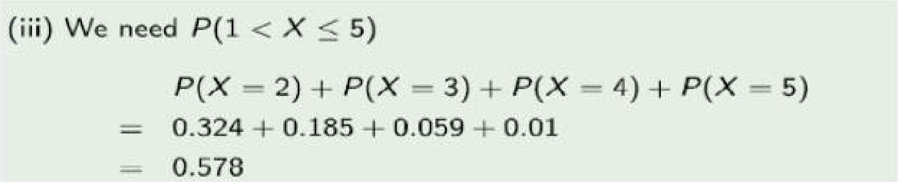


(i) If we call heads a success then this X has a binomial distribution with parameters $n = 6$ and $p = 0.3$.

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

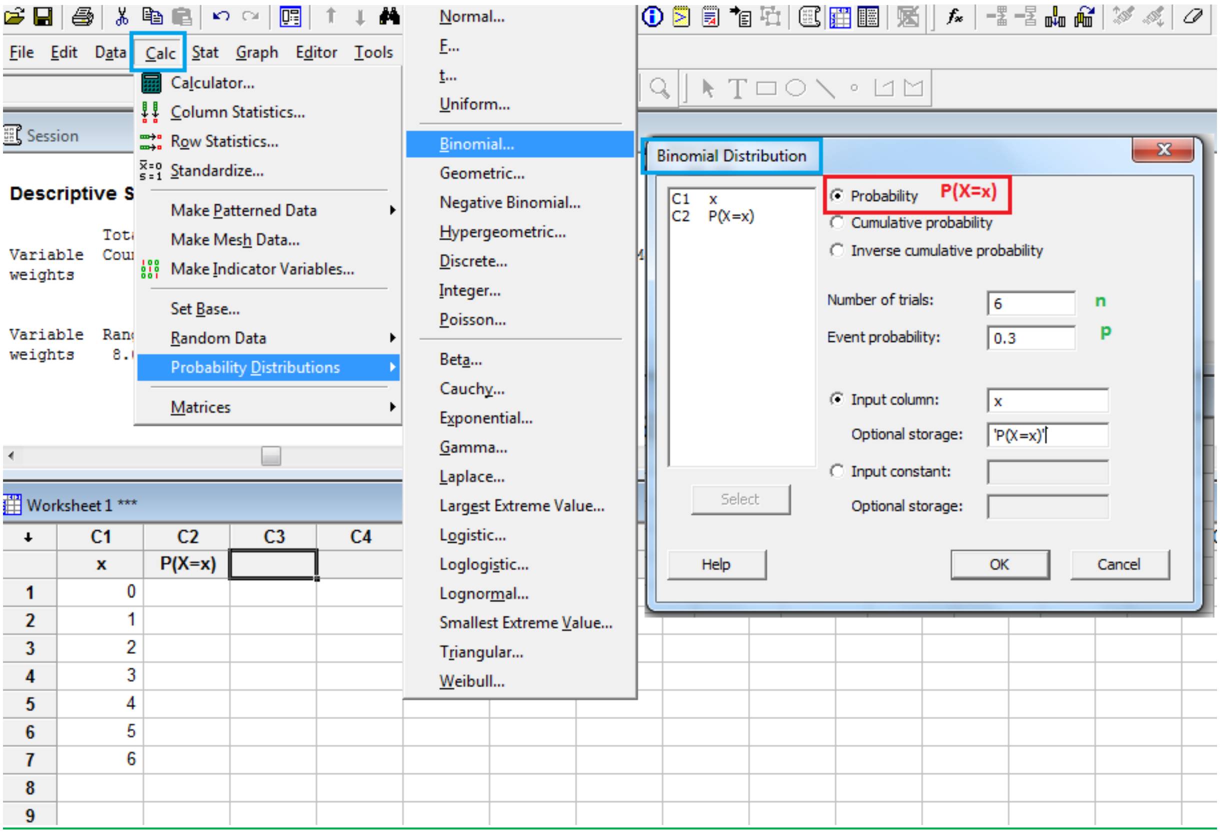
(ii)

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



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

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



● Probability P(X=x)
○ Cumulative probability
○ Inverse cumulative probability

Number of trials: 6 n

Event probability: 0.3 p

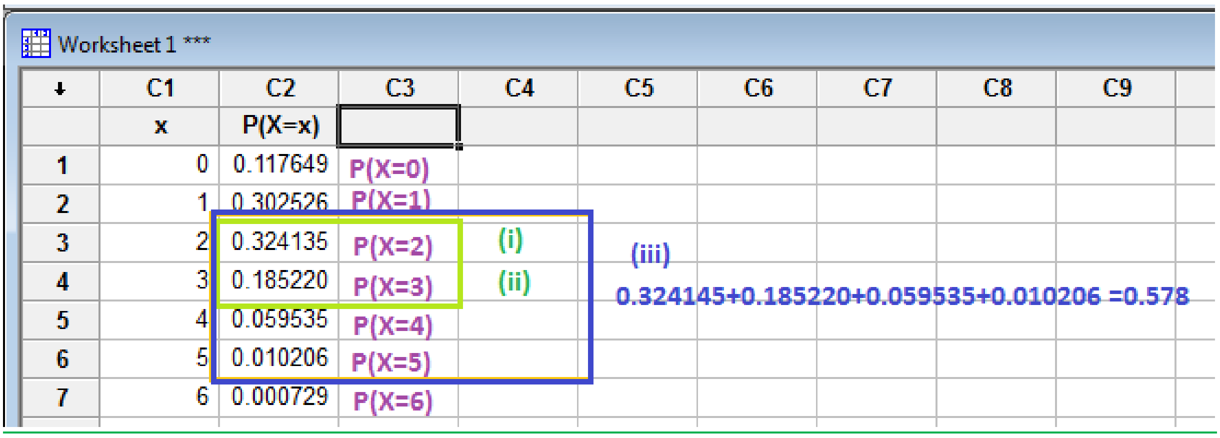
● Input column: x

Optional storage: 'P(X=x)'

○ Input constant:

Optional storage:

↓	C1	C2	C3	C4
	x	P(X=x)		
1	0			
2	1			
3	2			
4	3			
5	4			
6	5			
7	6			
8				
9				



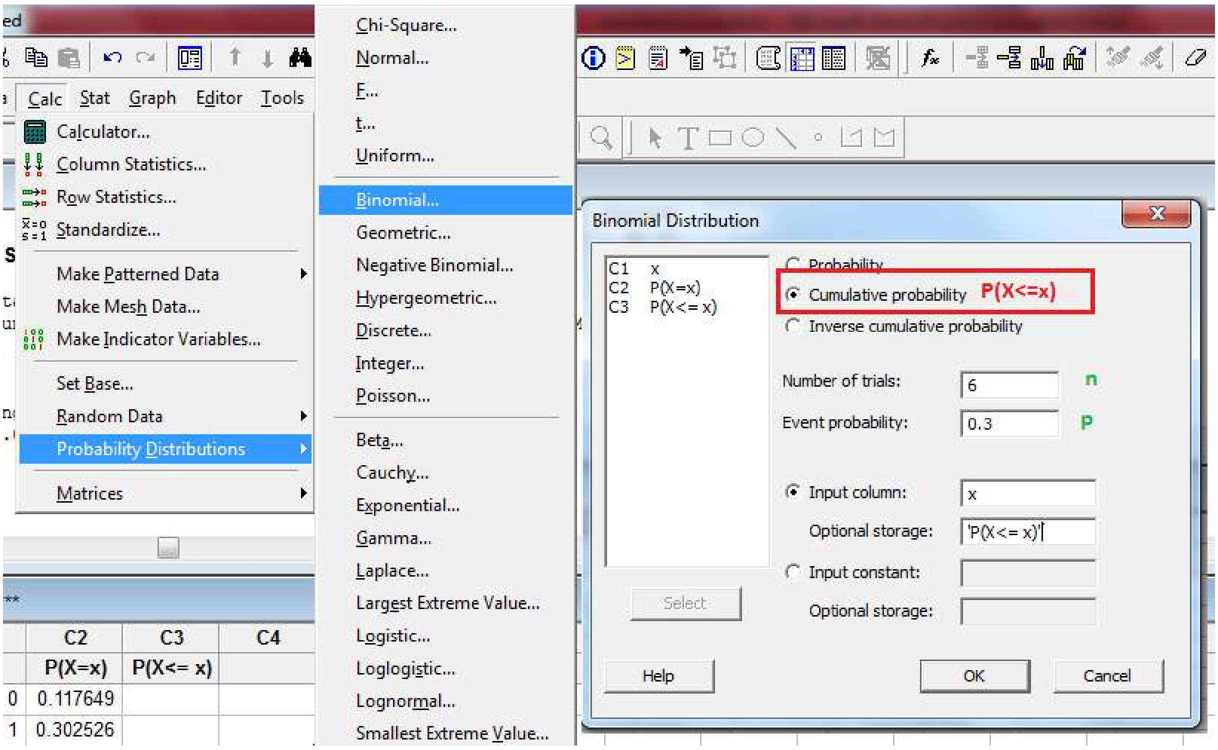
Worksheet 1 ***

↓	C1	C2	C3	C4
	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



Binomial Distribution

C1 x
C2 P(X=x)
C3 P(X<= x)

○ Probability

● Cumulative probability P(X<=x)

○ Inverse cumulative probability

Number of trials: 6 n

Event probability: 0.3 p

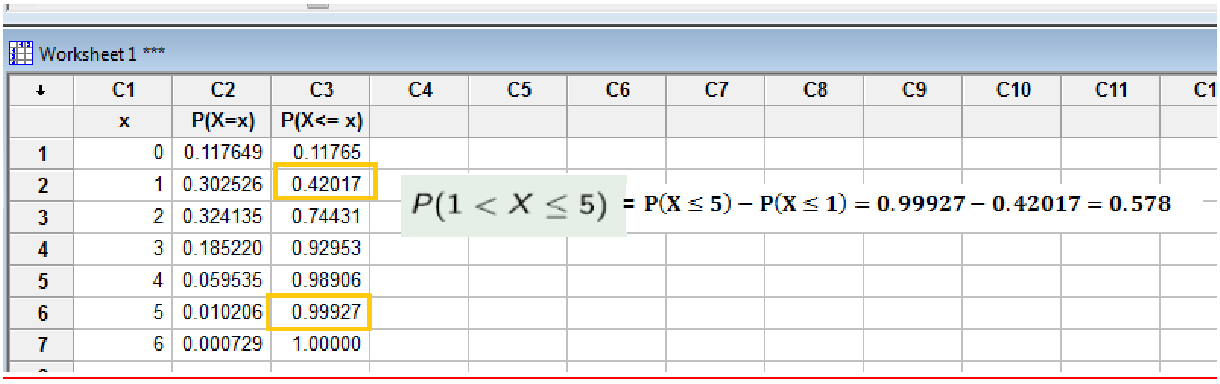
● Input column: x

Optional storage: 'P(X<= x)'

○ Input constant:

Optional storage:

	C2	C3	C4
	P(X=x)	P(X<= x)	
0	0.117649		
1	0.302526		



Worksheet 1 ***

↓	C1	C2	C3	C4
	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.

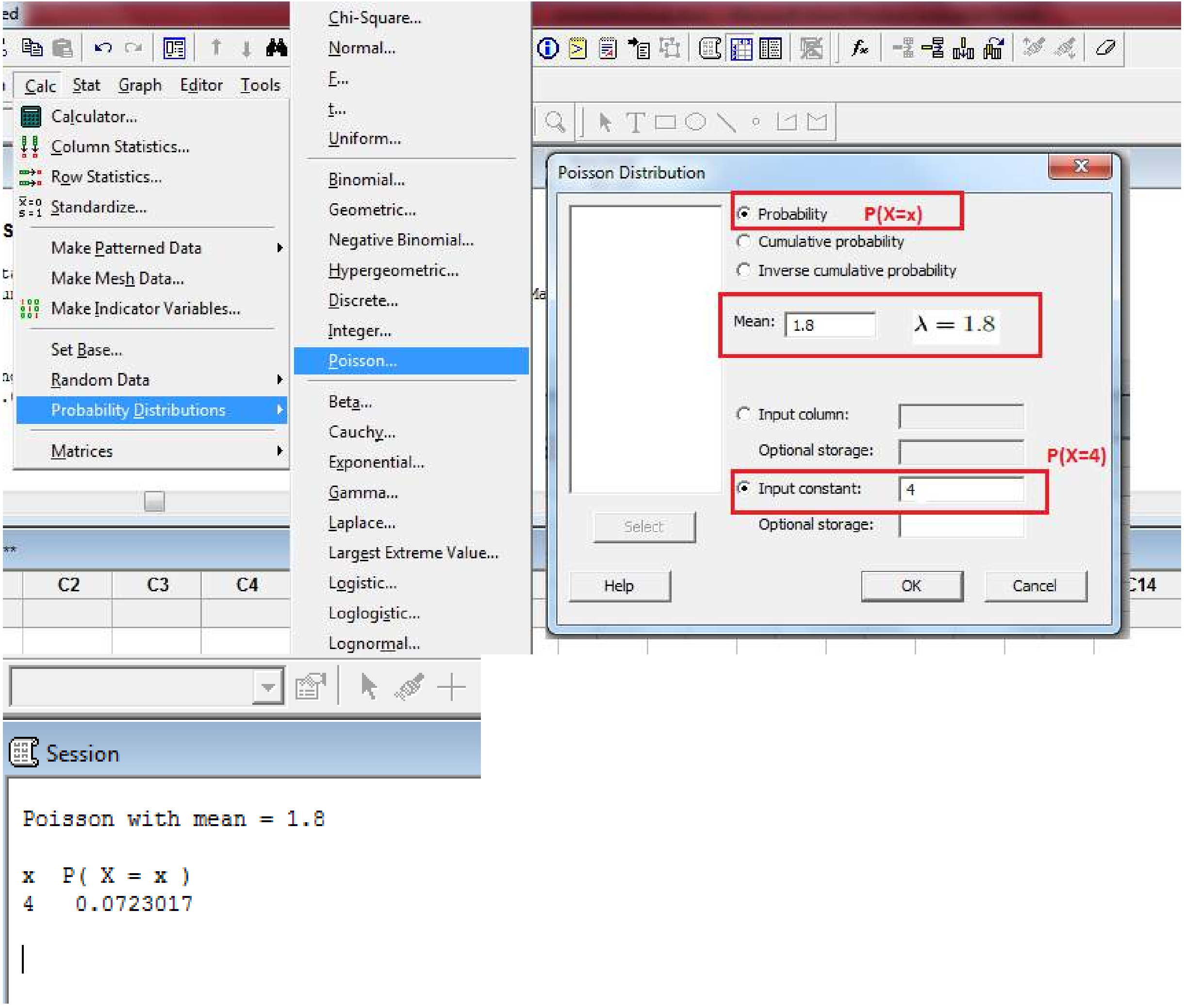
What is the probability of observing 4 births in a given hour at the hospital?

Let X = No. of births in a given hour

$$\left.\begin{array}{l}\text{(i) Events occur randomly}\\ \text{(ii) Mean rate } \lambda = 1.8\end{array}\right. \Rightarrow X \sim \mathrm{Po}(1.8)$$

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

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



● Probability

Mean: 1.8

● Input constant: 4

```
Poisson with mean = 1.8

x  P( X = x )
4    0.0723017
```

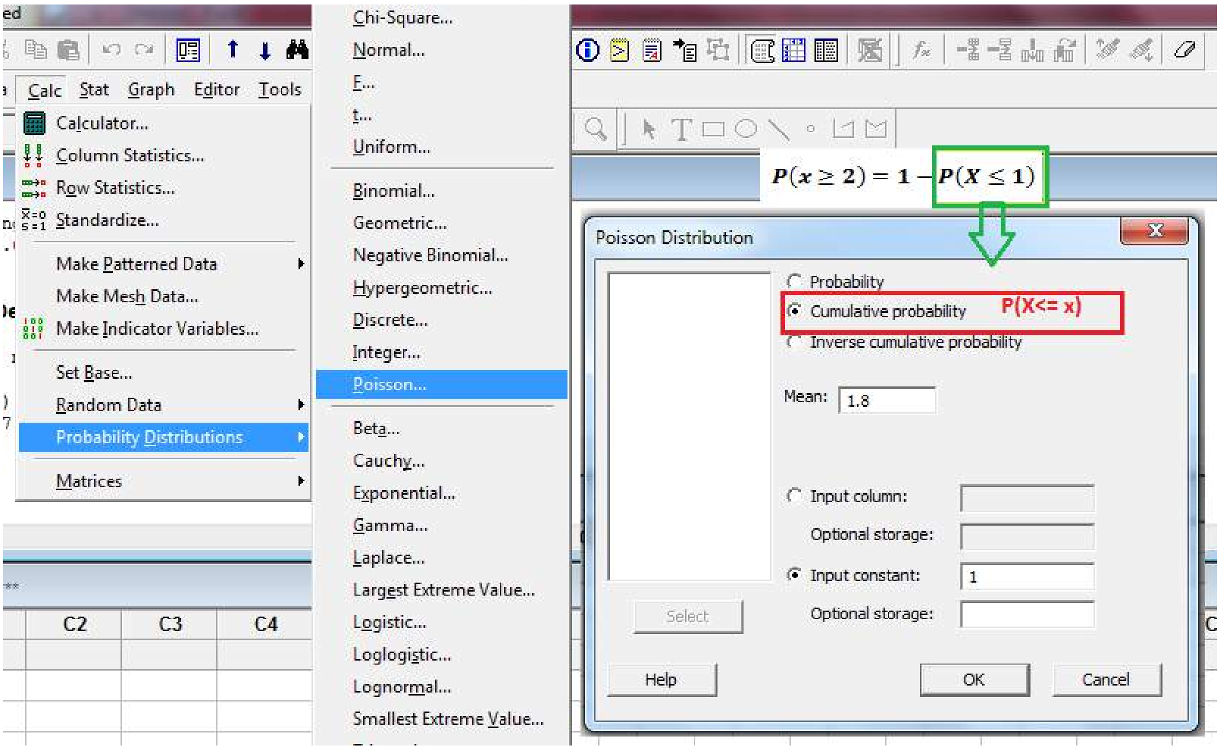
What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

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

i.e. an infinite number of probabilities to calculate

but

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



$P(x \geq 2) = 1 - \boxed{P(X \leq 1)}$

Poisson Distribution

○ Probability

● Cumulative probability P(X<= x)

○ Inverse cumulative probability

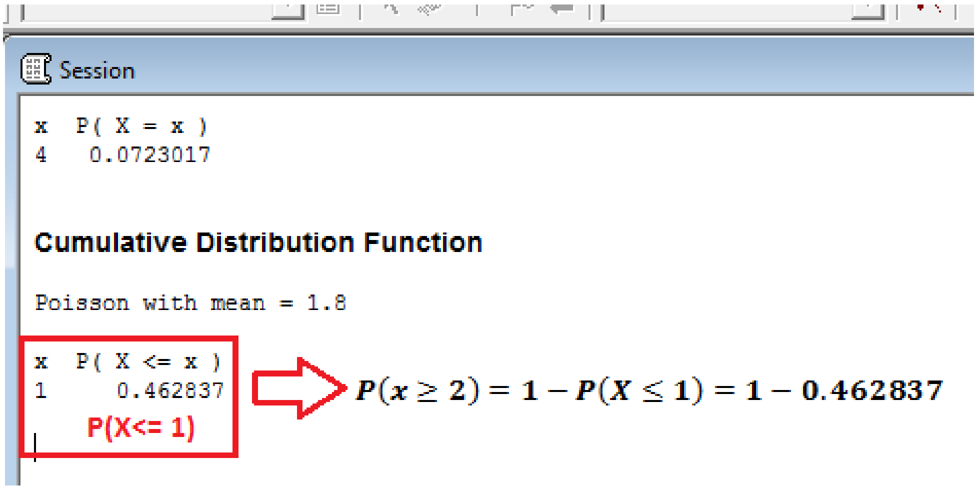
Mean: 1.8

○ Input column:

Optional storage:

● Input constant: 1

Optional storage:



```
x  P( X = x )
4   0.0723017
```

Cumulative Distribution Function

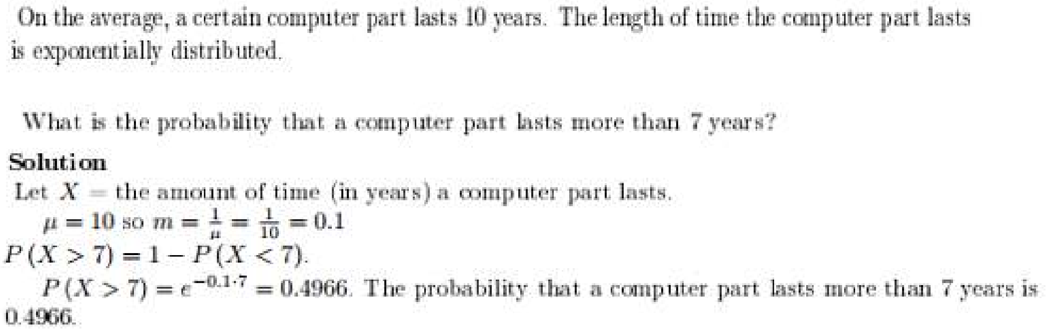
```
Poisson with mean = 1.8

x  P( X <= x )
1     0.462837
```

P(X<= 1)

$$P(x \geq 2) = 1 - P(X \leq 1) = 1 - 0.462837$$

Continuous Distributions

1. Exponential Distribution



On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.

What is the probability that a computer part lasts more than 7 years?

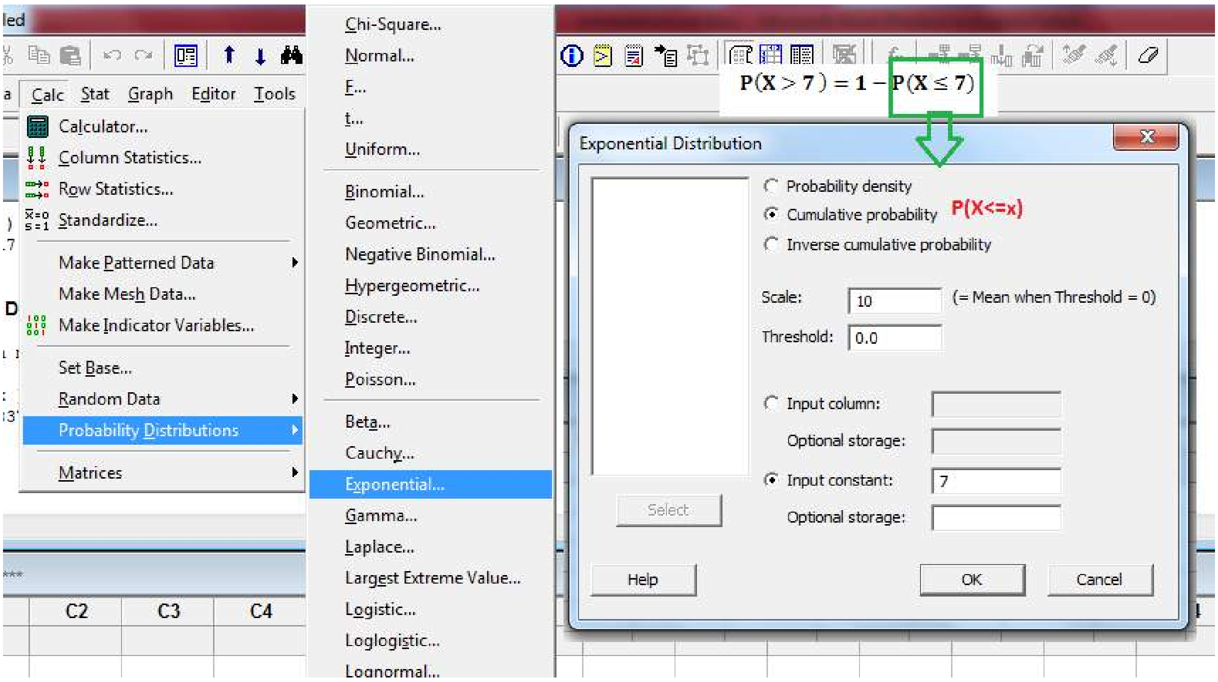
Solution

Let X = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

$P(X > 7) = 1 - P(X < 7)$.

$P(X > 7) = e^{-0.1 \cdot 7} = 0.4966$. The probability that a computer part lasts more than 7 years is 0.4966.



$P(X > 7) = 1 - P(X \le 7)$

Exponential Distribution

○ Probability density

● Cumulative probability P(X<=x)

○ Inverse cumulative probability

Scale: 10 (= Mean when Threshold = 0)

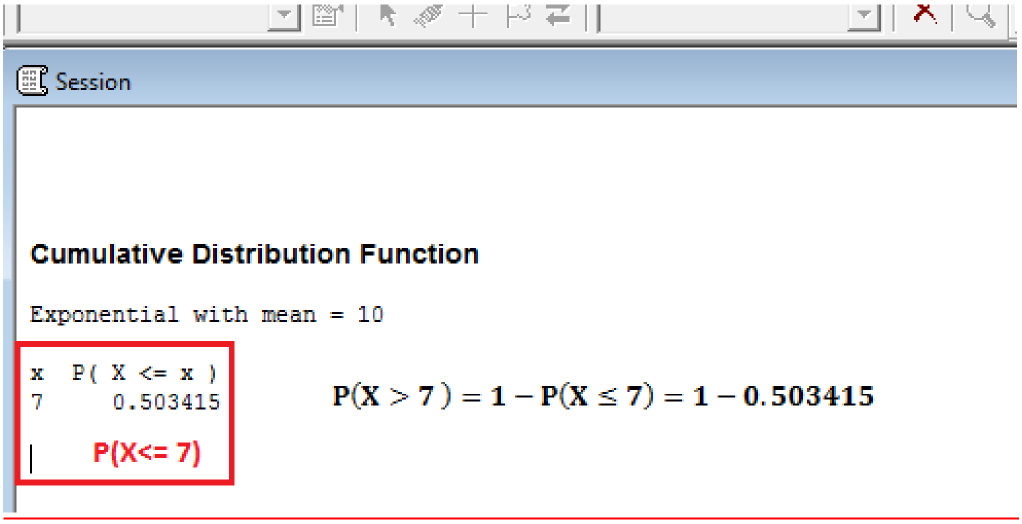
Threshold: 0.0

○ Input column:

Optional storage:

● Input constant: 7

Optional storage:



Cumulative Distribution Function

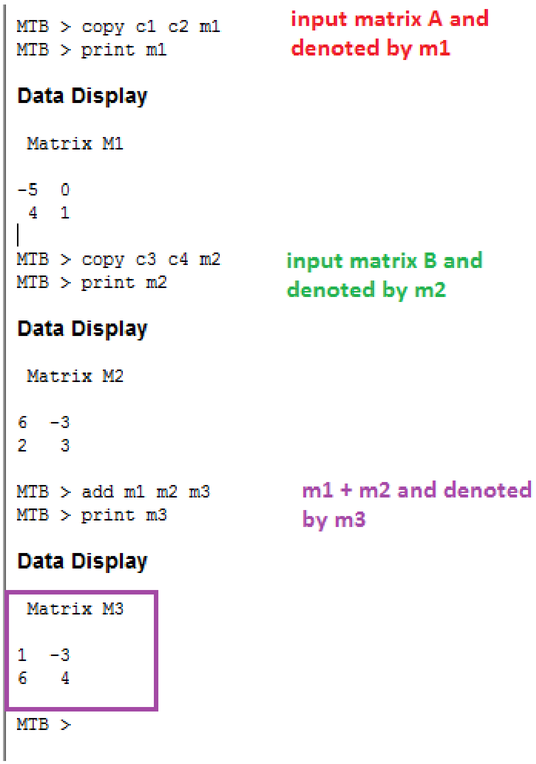
```
Exponential with mean = 10

x  P( X <= x )
7     0.503415
```

P(X<= 7)

$$P(X > 7) = 1 - P(X \le 7) = 1 - 0.503415$$

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 > 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

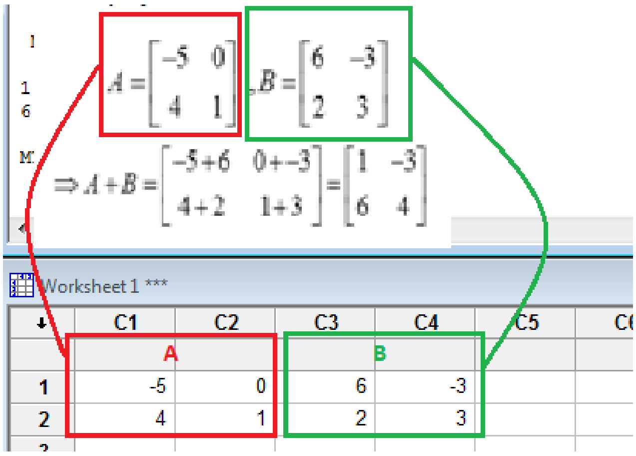
Data Display

```
 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 >
```



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

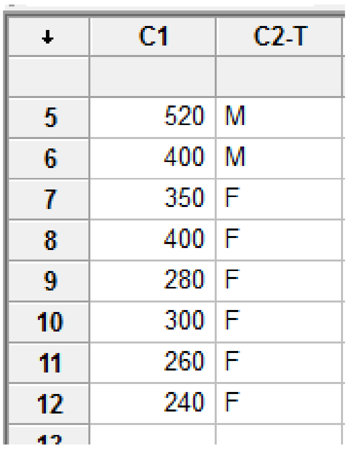
$$\Rightarrow A+B=\begin{bmatrix}-5+6 & 0+-3\\ 4+2 & 1+3\end{bmatrix}=\begin{bmatrix}1 & -3\\ 6 & 4\end{bmatrix}$$

↓	C1	C2	C3	C4	C5	C6
	A		B			
1	-5	0	6	-3		
2	4	1	2	3		

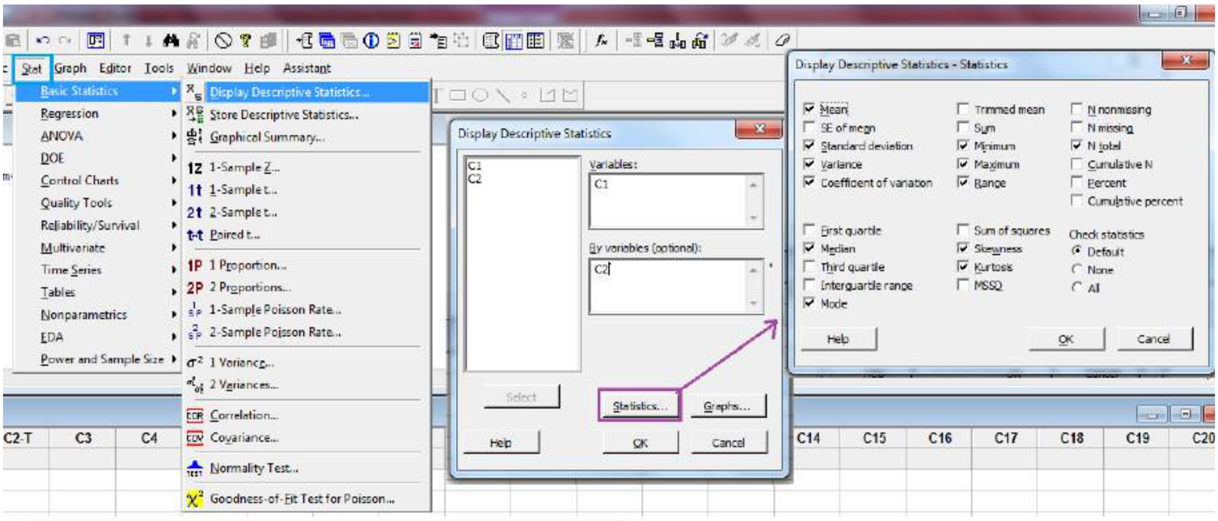
The following table gives the monthly income for sample of employees; analyze the data based on gender

Male	**450**	**500**	**480**	**300**	**520**	**400**
Female	**350**	**400**	**280**	**300**	**260**	**240**

Analyze the data based on the gender



↓	C1	C2-T
5	520	M
6	400	M
7	350	F
8	400	F
9	280	F
10	300	F
11	260	F
12	240	F



Display Descriptive Statistics

C1
C2

Variables:
C1

By variables (optional):
C2

Display Descriptive Statistics - Statistics

☑ Mean
☐ SE of mean
☑ Standard deviation
☑ Variance
☑ Coefficent of variation

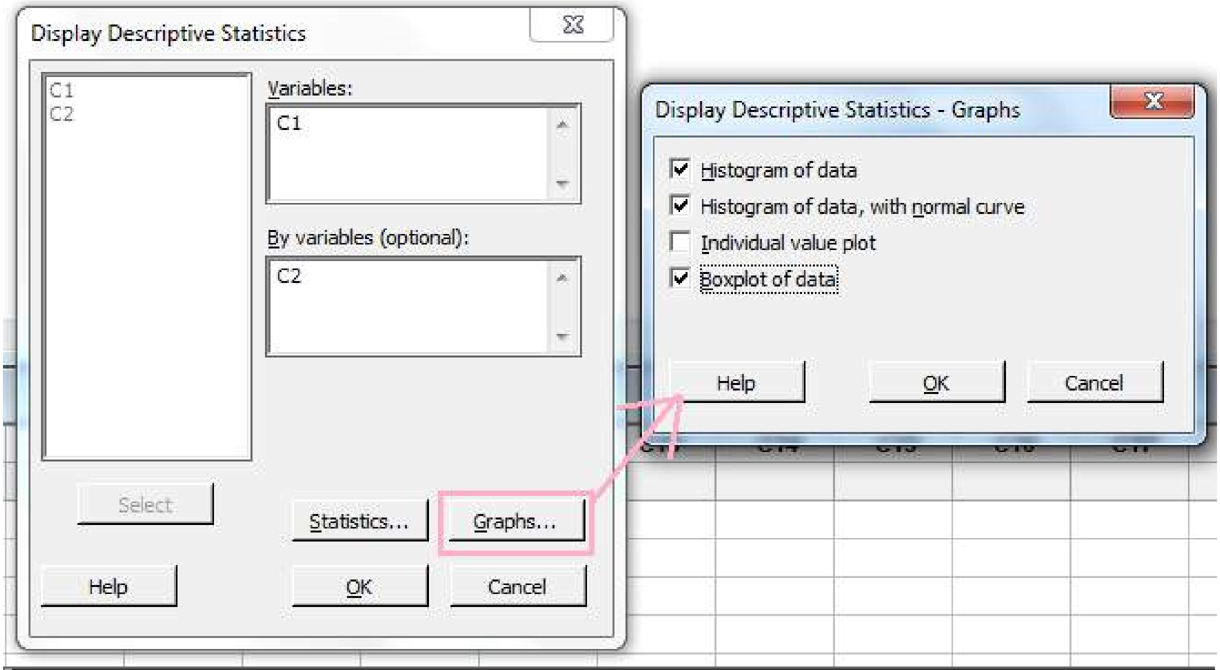
☐ First quartile
☑ Median
☐ Third quartile
☐ Interquartile range
☑ Mode

☐ Trimmed mean
☐ Sum
☑ Minimum
☑ Maximum
☑ Range

☐ Sum of squares
☑ Skewness
☑ Kurtosis
☐ MSSD

☐ N nonmissing
☐ N missing
☑ N total
☐ Cumulative N
☐ Percent
☐ Cumulative percent

Check statistics
● Default
○ None
○ All


Display Descriptive Statistics

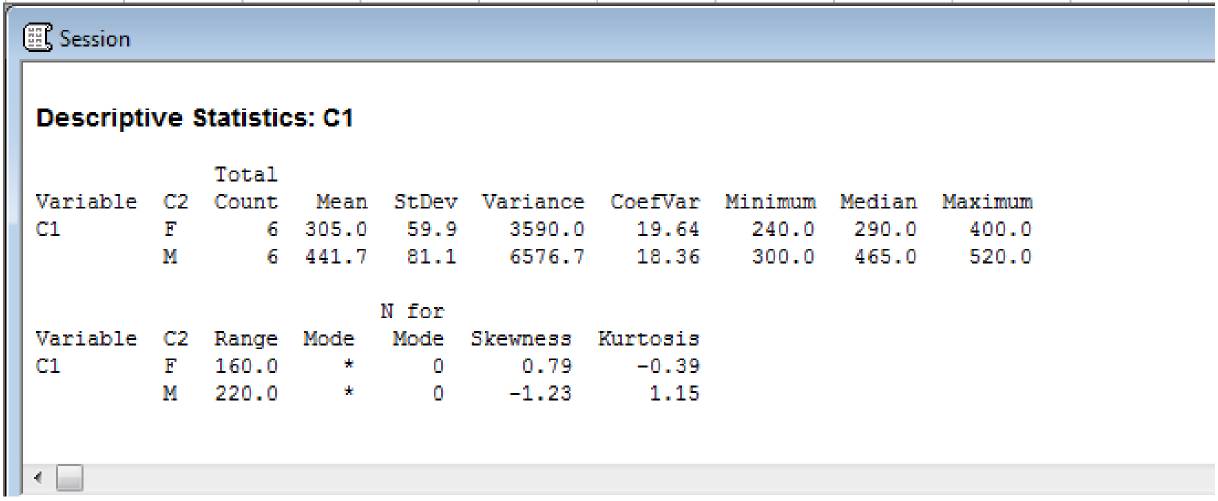
C1
C2

Variables:
C1

By variables (optional):
C2

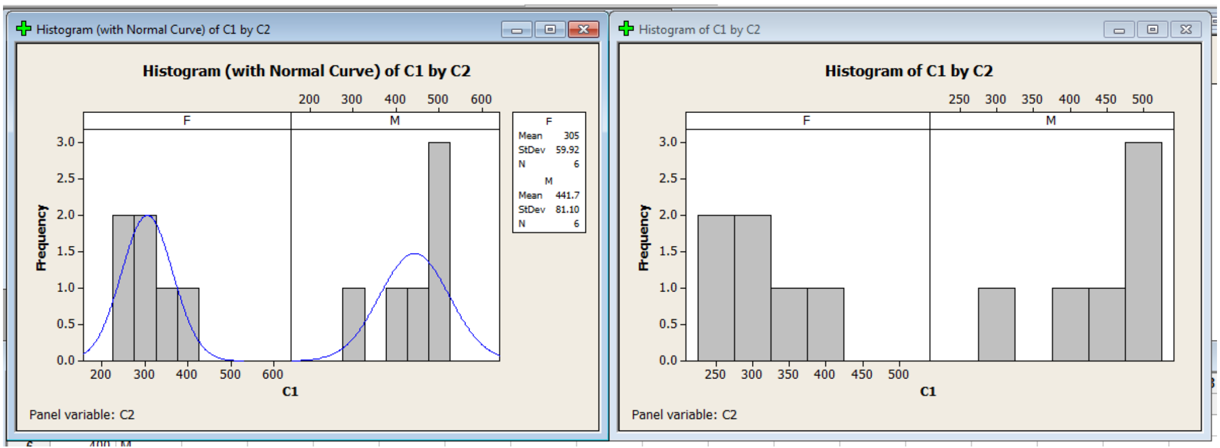
Display Descriptive Statistics - Graphs

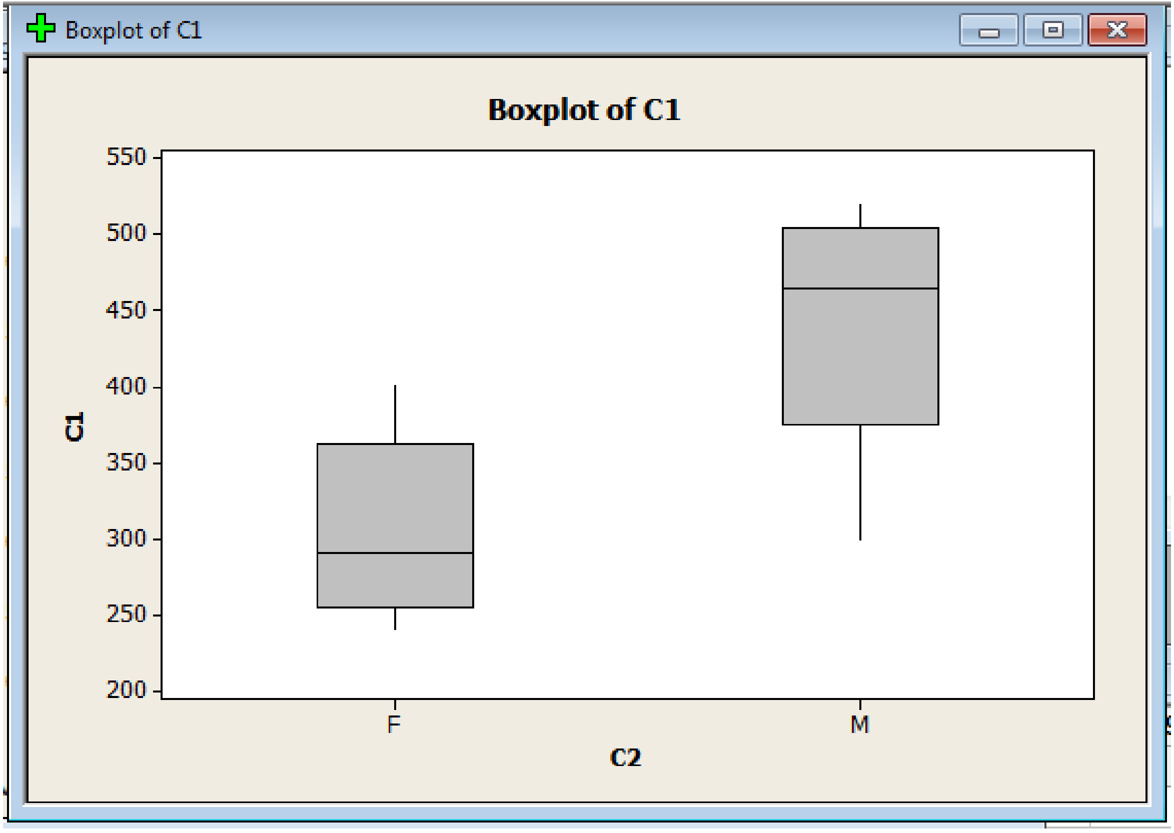
☑ Histogram of data
☑ Histogram of data, with normal curve
☐ Individual value plot
☑ Boxplot of data



Descriptive Statistics: C1

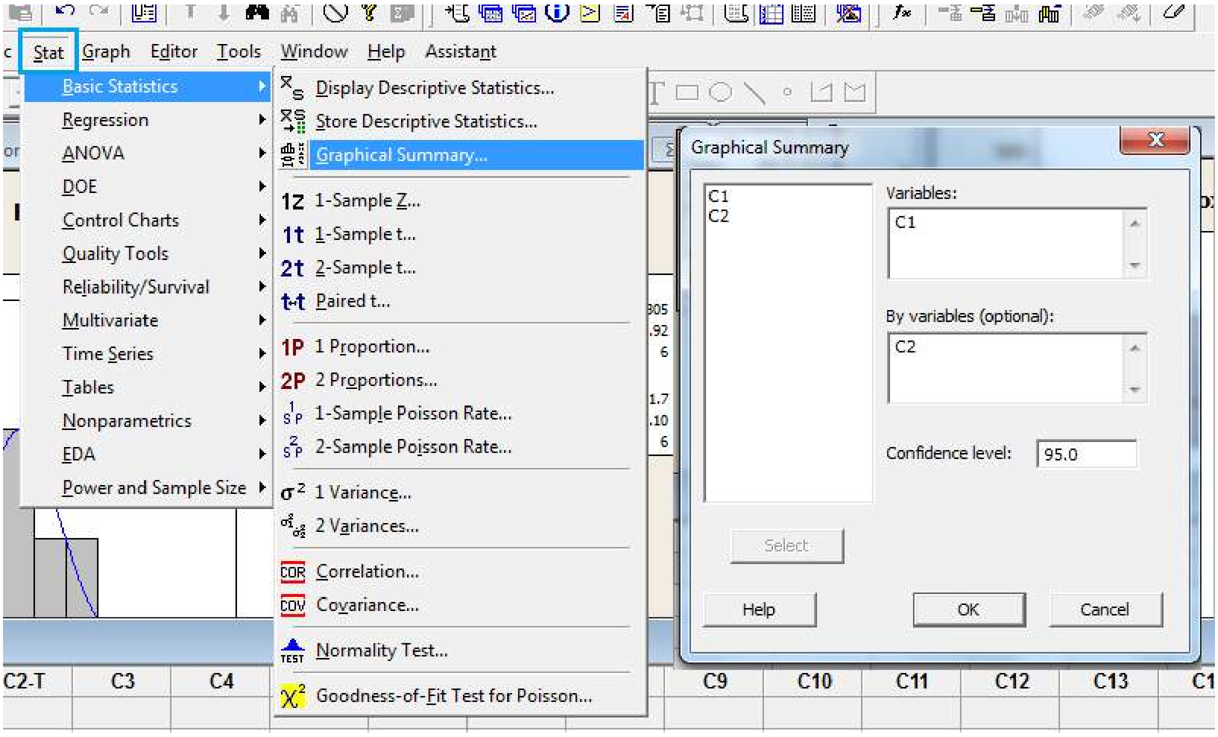
```
                Total
Variable  C2    Count   Mean  StDev  Variance  CoefVar  Minimum  Median  Maximum
C1        F         6  305.0   59.9    3590.0    19.64    240.0   290.0    400.0
          M         6  441.7   81.1    6576.7    18.36    300.0   465.0    520.0


                             N for
Variable  C2    Range  Mode   Mode  Skewness  Kurtosis
C1        F     160.0     *      0      0.79     -0.39
          M     220.0     *      0     -1.23      1.15
```





- **Graphical Summary**

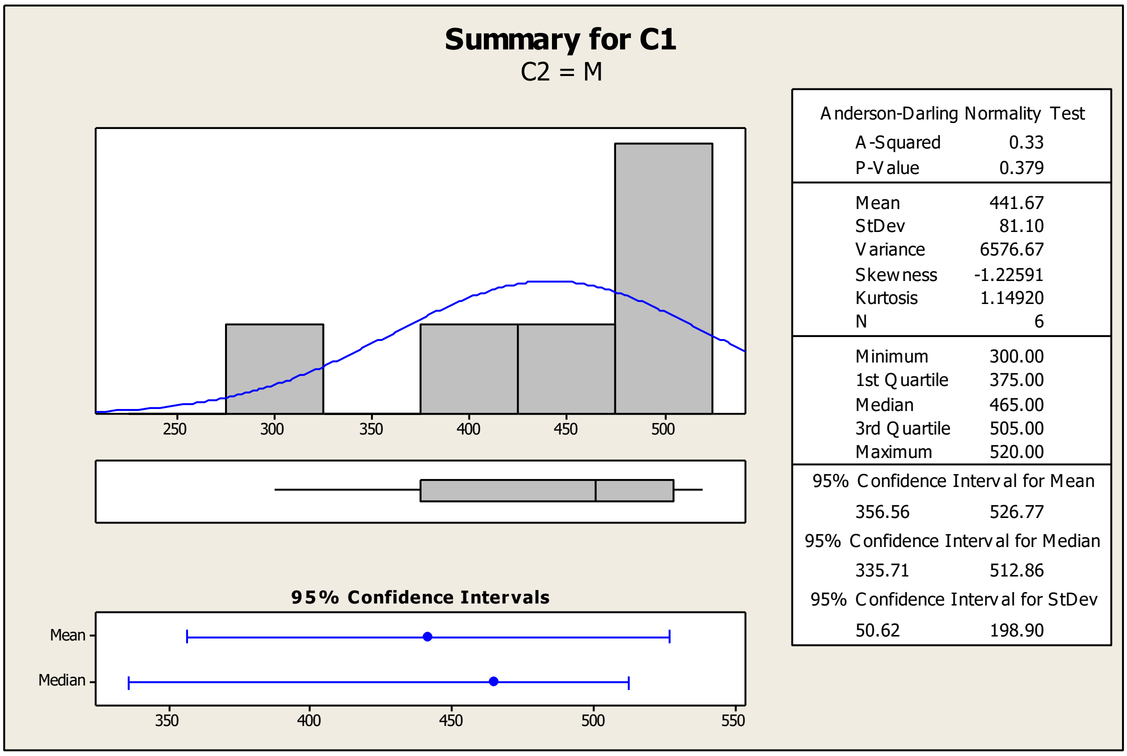
The graphical summary can be also introduced for the income of both male and female as follows



Variables: C1

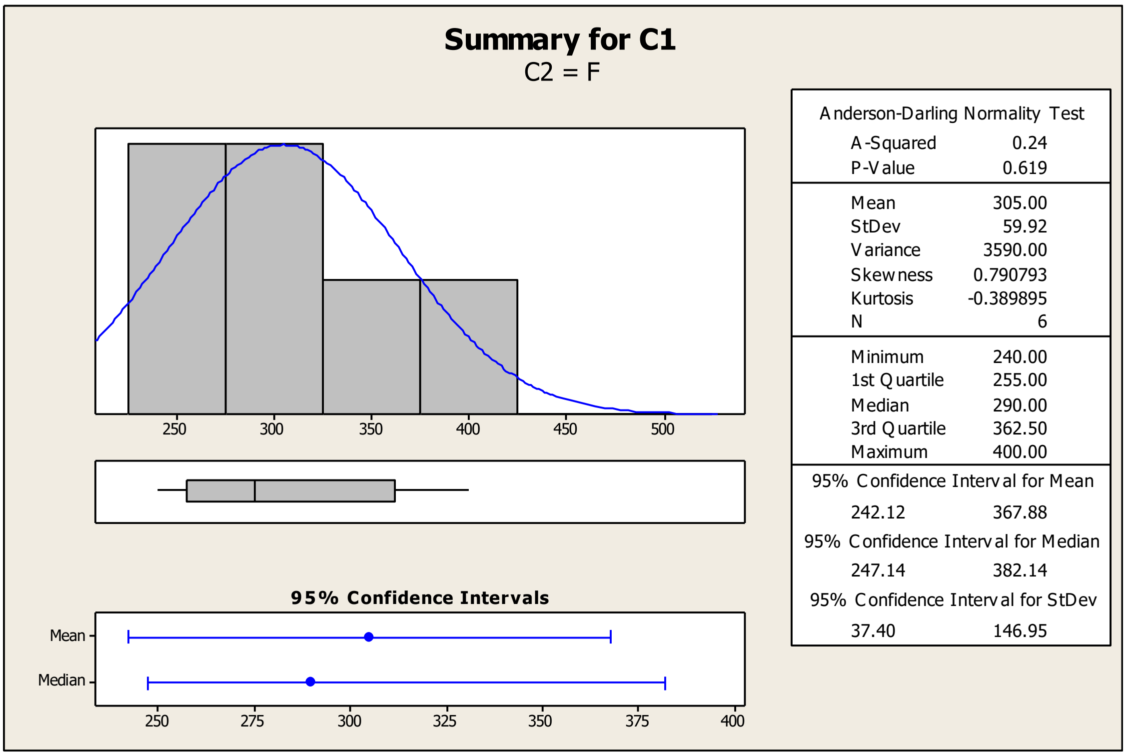
By variables (optional): C2

Confidence level: 95.0



Summary for C1
C2 = 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
```



```
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
```

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,

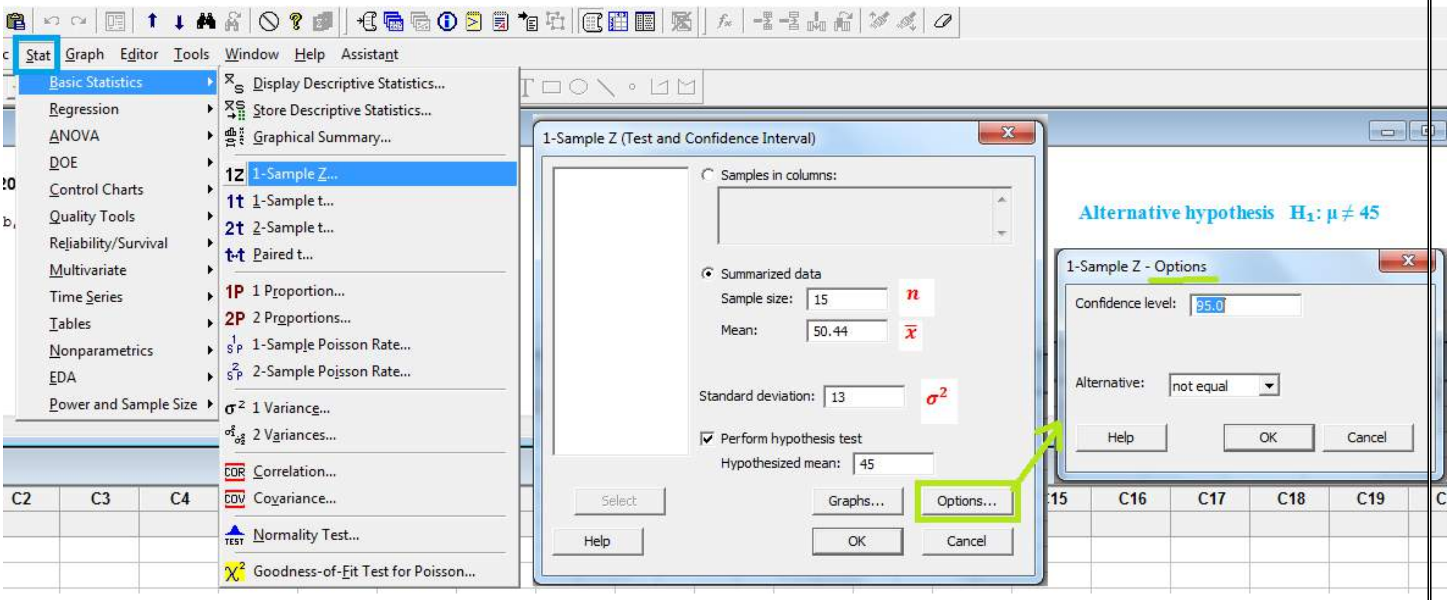
a) **Test whether the true mean vitamin C content is different from 45 mg/100g**
b) **Find a 95% confidence interval for the average vitamin C content .**

*** 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.

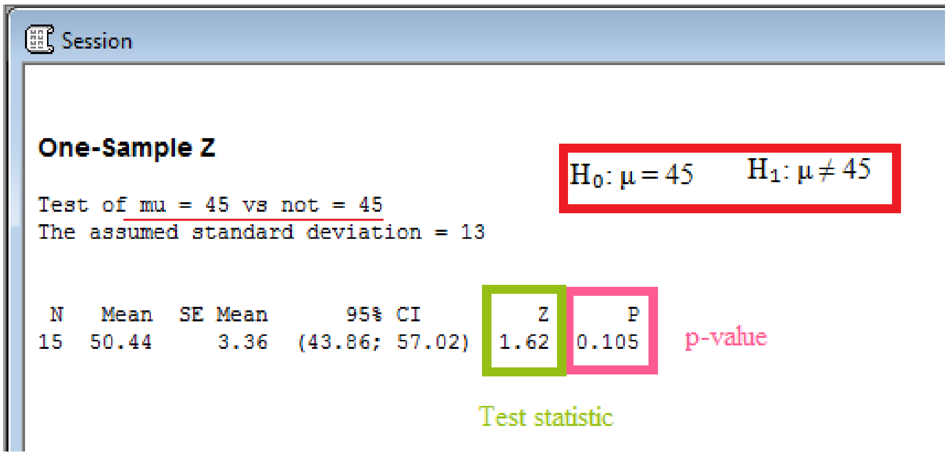
- Calculate a range of values that is likely to include the population mean.



Alternative hypothesis $H_1: \mu \neq 45$

○ Samples in columns:
● Summarized data
Sample size: 15
Mean: 50.44
Standard deviation: 13
☑ Perform hypothesis test
Hypothesized mean: 45

Confidence level: 95.0
Alternative: not equal



```
One-Sample Z

Test of mu = 45 vs not = 45
The assumed standard deviation = 13


 N   Mean  SE Mean       95% CI          Z      P
15  50.44     3.36  (43.86; 57.02)    1.62  0.105
```

1- **Hypothesis:**

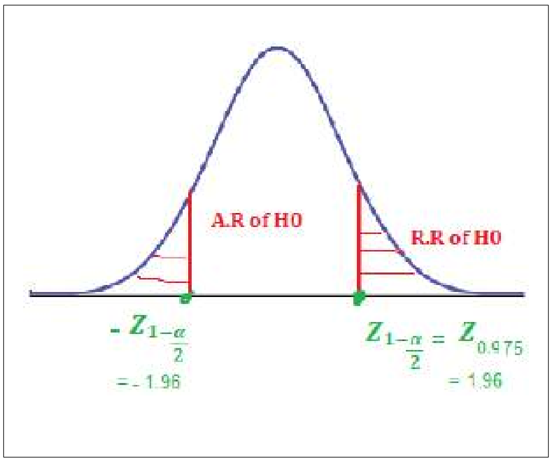
Null hypothesis H_0: $\mu = 45$ VS Alternative hypothesis H_1: $\mu \neq 45$

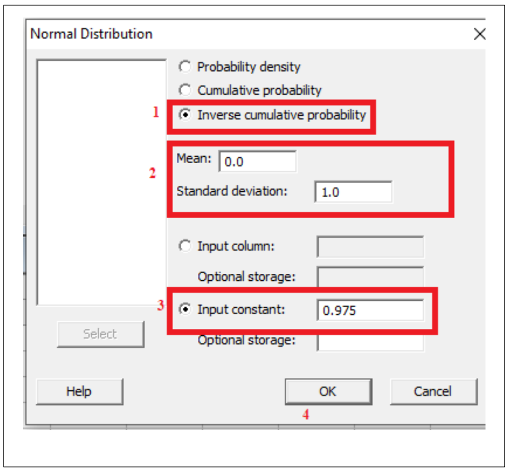
2- **Test statistic :**

Z=1.62

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

Calc>> probability distributions>>Normal



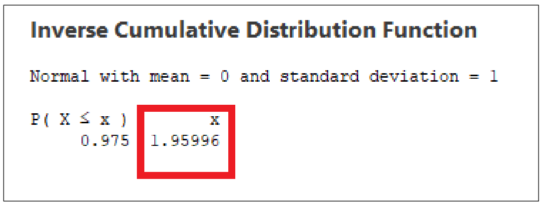


1 ● Inverse cumulative probability

2 Mean: 0.0
Standard deviation: 1.0

3 ● Input constant: 0.975

4 OK



Inverse Cumulative Distribution Function

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

P( X ≤ x )        x
     0.975  1.95996
```

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)]:

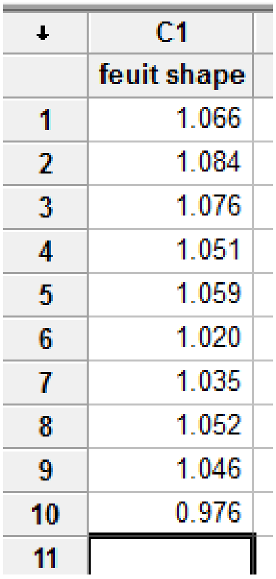
1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

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

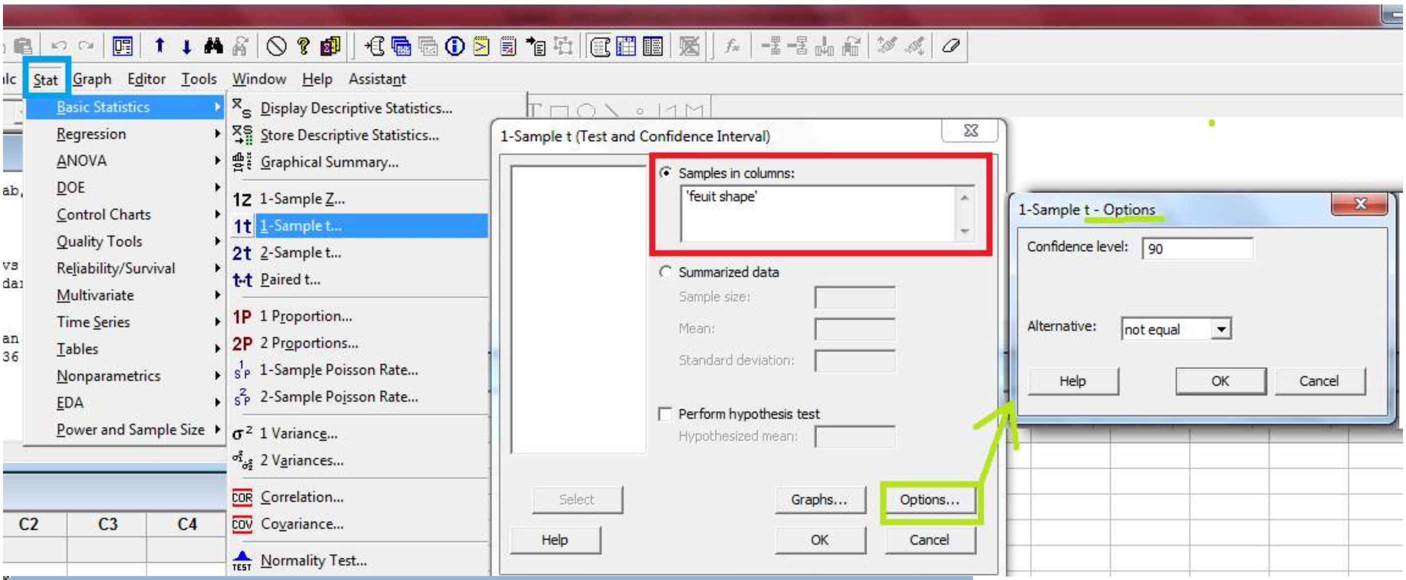
*Use the 1-sample t-test to estimate the mean of a population and compare it to a target or reference value **when you do not know the standard deviation of the population**.

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.



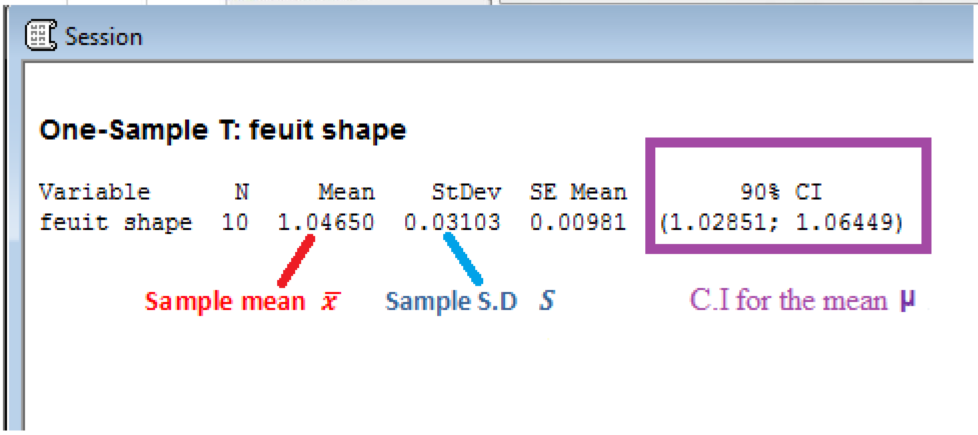
↓	C1
	feuit shape
1	1.066
2	1.084
3	1.076
4	1.051
5	1.059
6	1.020
7	1.035
8	1.052
9	1.046
10	0.976
11	



● Samples in columns: 'feuit shape'

Confidence level: 90

Alternative: not equal



```
One-Sample T: feuit shape

Variable      N     Mean    StDev  SE Mean         90% CI
feuit shape  10  1.04650  0.03103  0.00981  (1.02851; 1.06449)
```

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:

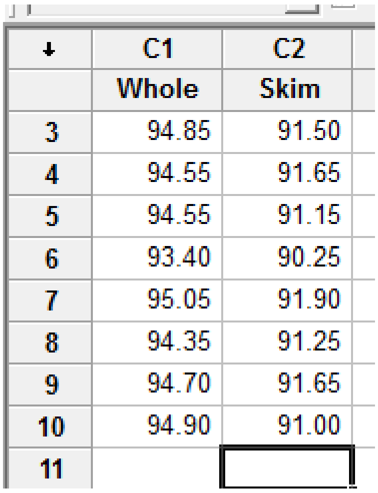
Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

Assuming normal populations with equal variance

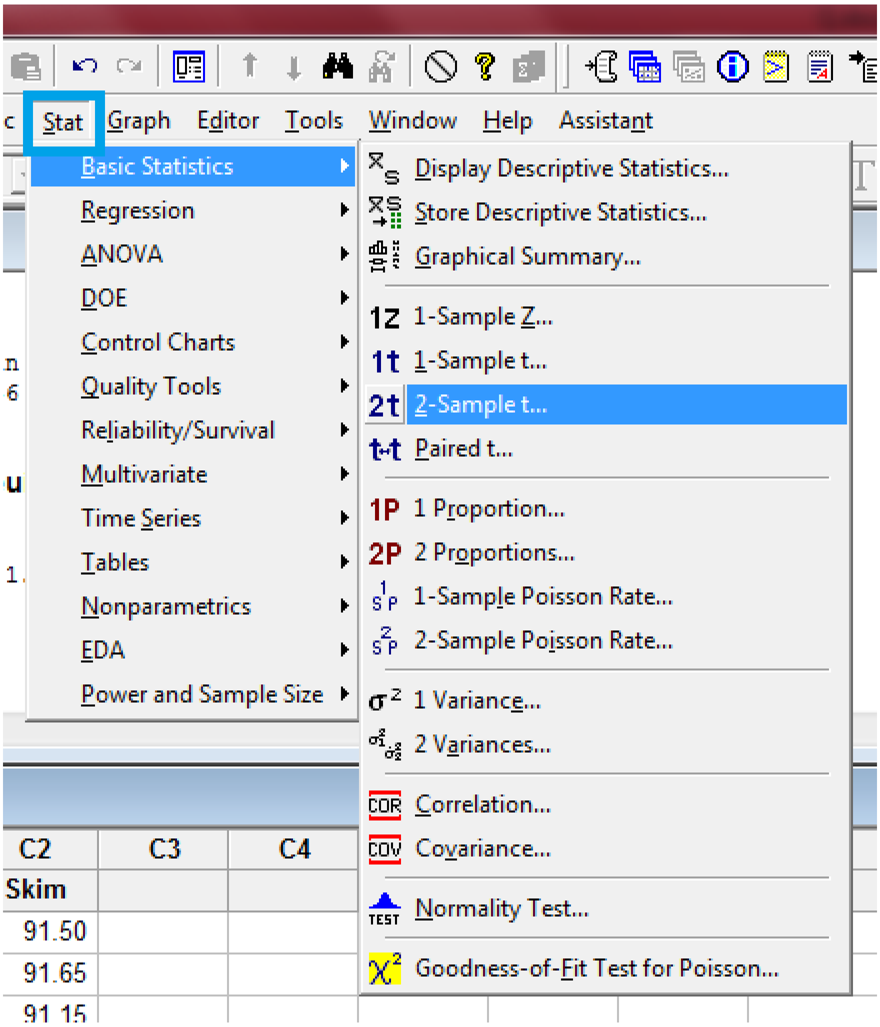
a) **Test whether the average phosphorus content of skim milk is less than the average phosphorus content of whole milk . Use α=0.01**

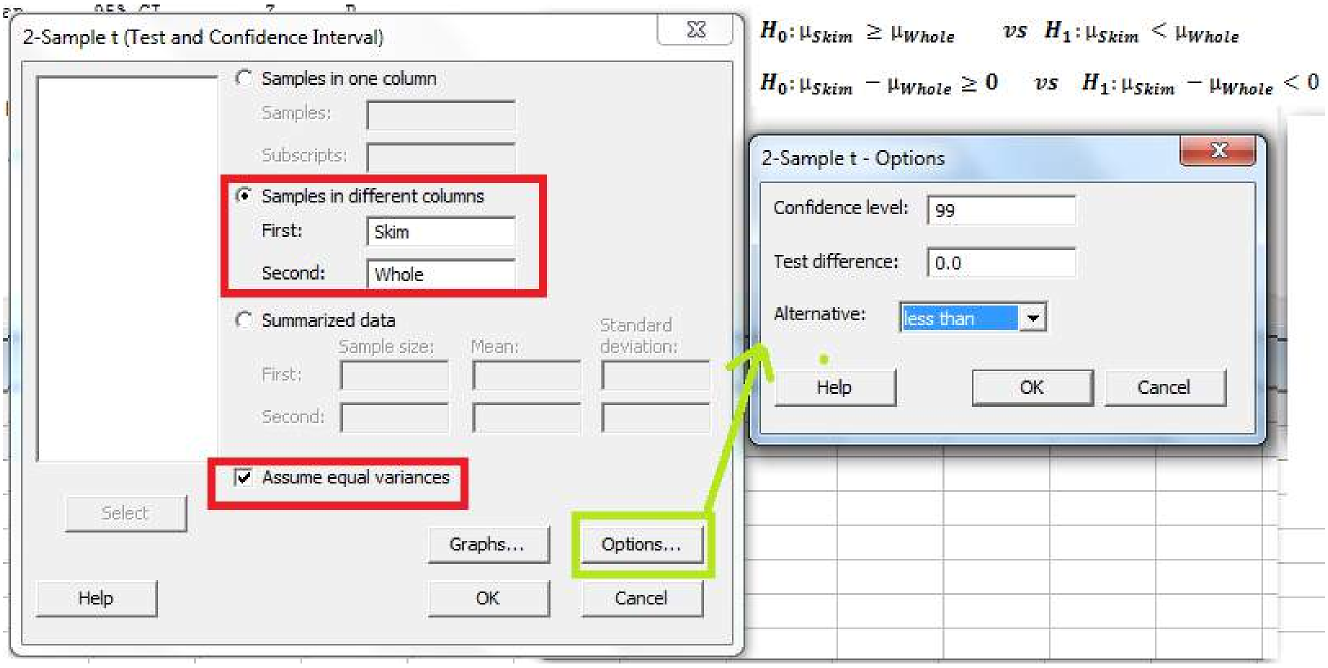
b) **Find and interpret a 99% confidence interval for the difference in average phosphorus contents of whole and skim milk .**

*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		




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

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

2-Sample t (Test and Confidence Interval)

○ Samples in one column

● Samples in different columns

First: Skim

Second: Whole

○ Summarized data

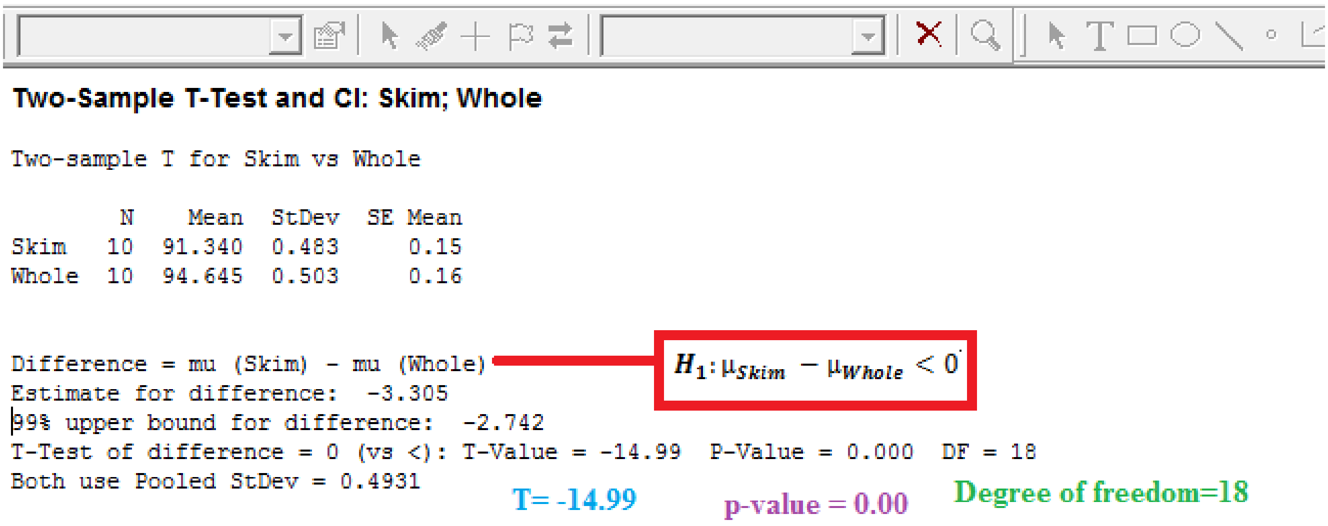
☑ Assume equal variances

2-Sample t - Options

Confidence level: 99

Test difference: 0.0

Alternative: less than



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 :

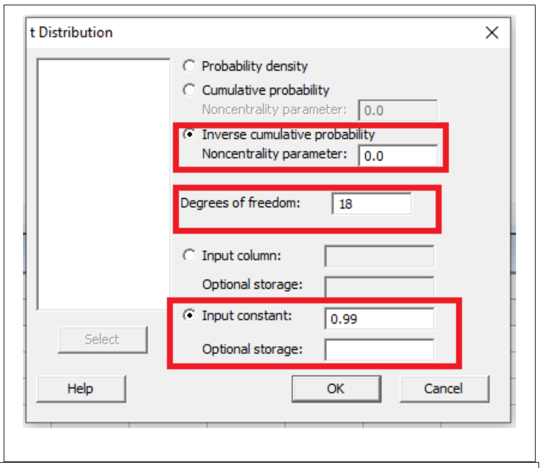
$$H_0: \mu_{Skim} \geq \mu_{Whole} \quad vs \quad H_1: \mu_{Skim} < \mu_{Whole}$$

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

2- Test statistic : T= -14.99

3- The critical region(s):

Calc>> probability distributions>> t

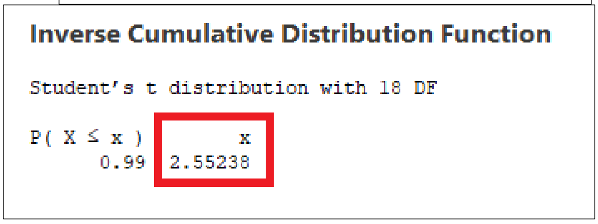


t Distribution

○ Probability density
○ Cumulative probability
● Inverse cumulative probability
Noncentrality parameter: 0.0

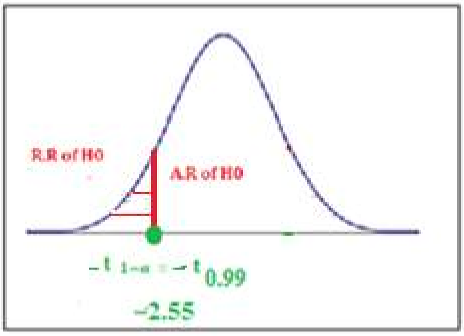
Degrees of freedom: 18

○ Input column:
Optional storage:
● Input constant: 0.99
Optional storage:



Inverse Cumulative Distribution Function

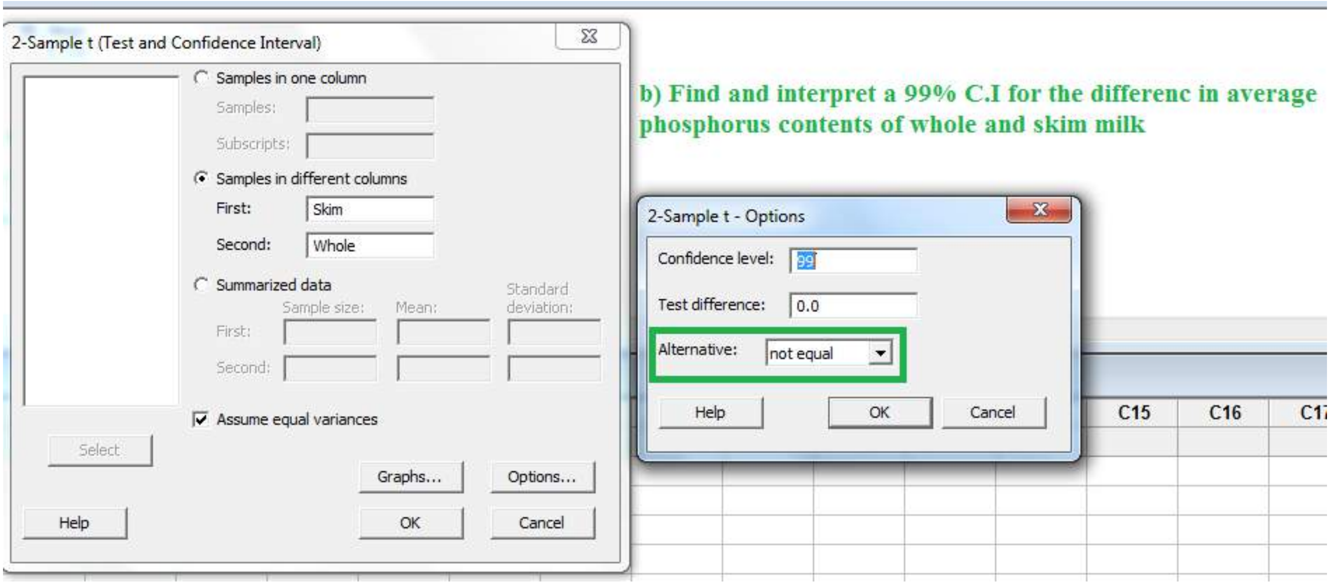
```
Student's t distribution with 18 DF

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



4- Decision:

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

b)



b) Find and interpret a 99% C.I for the differenc in average phosphorus contents of whole and skim milk

○ Samples in one column

● Samples in different columns

First: Skim

Second: Whole

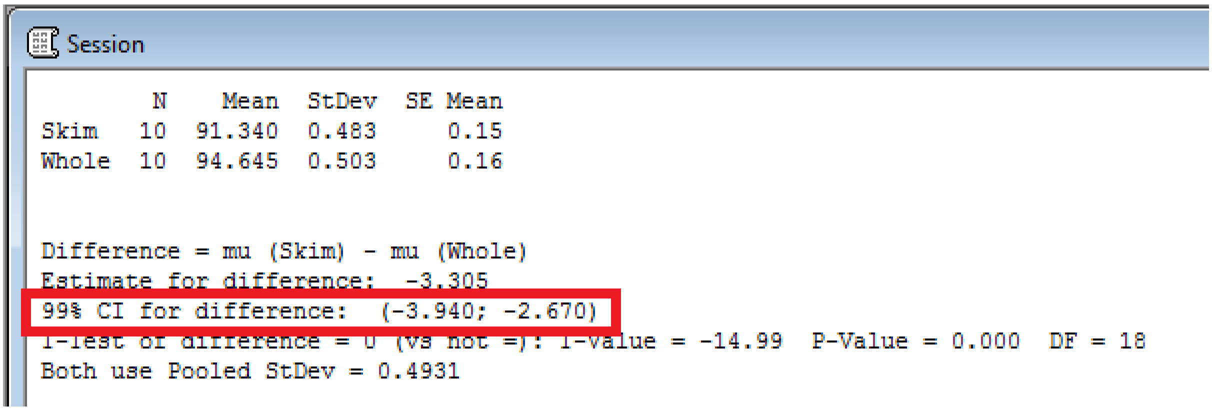
○ Summarized data

☑ Assume equal variances

Confidence level: 99

Test difference: 0.0

Alternative: not equal



```
        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
```

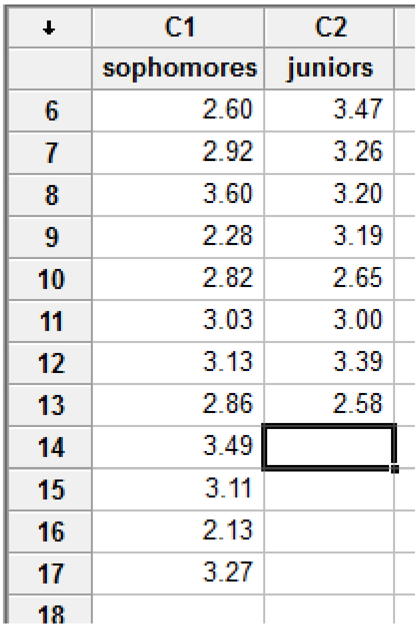
$$\mu_{Skim} - \mu_{Whole} \in (-3.940 \ , -2.670 \ )$$

Two-sample t-test

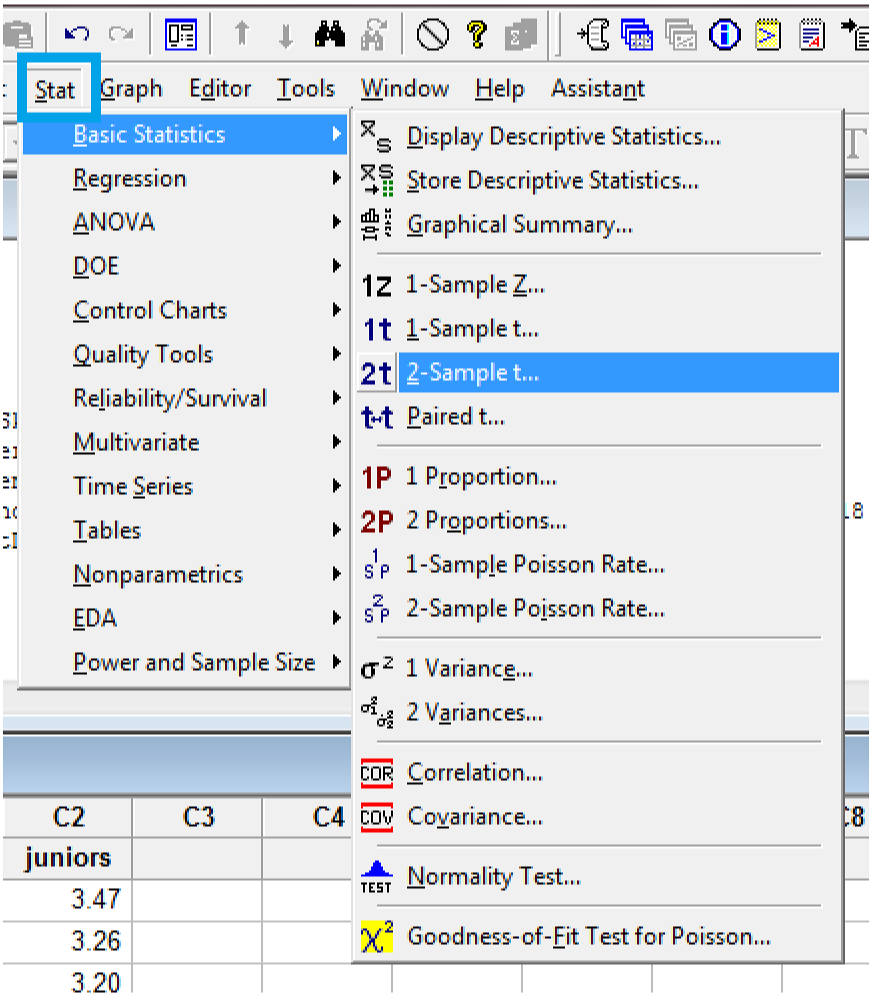
Q : Independent random samples of 17 sophomores and 13 juniors attending a large university yield the following data on grade point averages

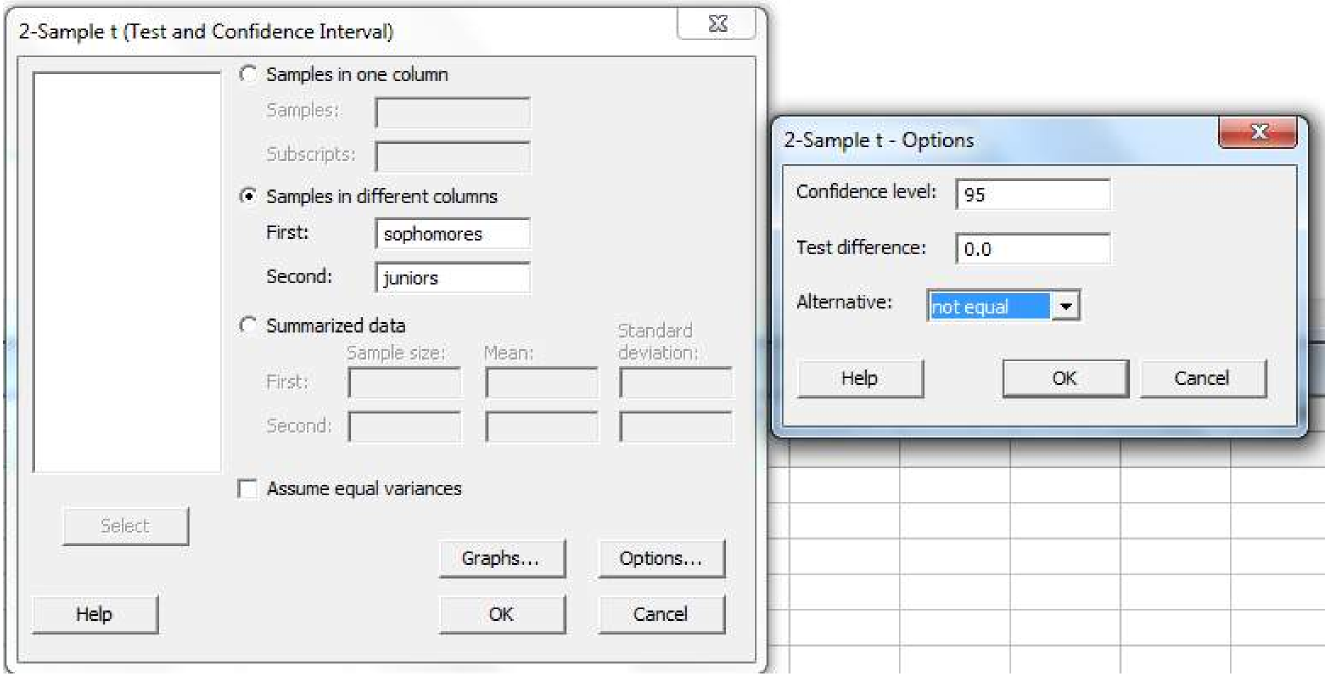
sophomores			juniors		
3.04	**2.92**	**2.86**	**2.56**	**3.47**	**2.65**
1.71	**3.60**	**3.49**	**2.77**	**3.26**	**3.00**
3.30	**2.28**	**3.11**	**2.70**	**3.20**	**3.39**
2.88	**2.82**	**2.13**	**3.00**	**3.19**	**2.58**
2.11	**3.03**	**3.27**	**2.98**		
2.60	**3.13**				

Assuming normal population. At the 5% significance level, do the data provide sufficient evidence to conclude that the mean GPAs of sophomores and juniors at the university different ?



↓	C1	C2
	sophomores	juniors
6	2.60	3.47
7	2.92	3.26
8	3.60	3.20
9	2.28	3.19
10	2.82	2.65
11	3.03	3.00
12	3.13	3.39
13	2.86	2.58
14	3.49	
15	3.11	
16	2.13	
17	3.27	
18		

Stat
Graph
Editor
Tools
Window
Help
Assistant
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Time Series
Tables
Nonparametrics
EDA
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...
Goodness-of-Fit Test for Poisson...
C2
C3
C4
juniors
3.47
3.26
3.20



2-Sample t (Test and Confidence Interval)

○ Samples in one column

Samples:

Subscripts:

● Samples in different columns

First: sophomores

Second: juniors

○ Summarized data

Sample size: Mean: Standard deviation:

First:

Second:

☐ Assume equal variances

2-Sample t - Options

Confidence level: 95

Test difference: 0.0

Alternative: not equal

Interval Plot of program1; program2; ...

Two-Sample T-Test and CI: Sophomores; Juniors

```
Two-sample T for Sophomores vs Juniors

             N   Mean  StDev  SE Mean
Sophomores  17  2.840  0.520     0.13
Juniors     13  2.981  0.309    0.086


Difference = μ (Sophomores) - μ (Juniors)
Estimate for difference:  -0.141
95% CI for difference:  (-0.454; 0.173)
T-Test of difference = 0 (vs ≠): T-Value = -0.92  P-Value = 0.364  DF = 26
```

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

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

T test p-value

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= -0.92

3- **Decision:**

Since p-value =0.364 > α= 0.05 . we can not reject H_0

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:

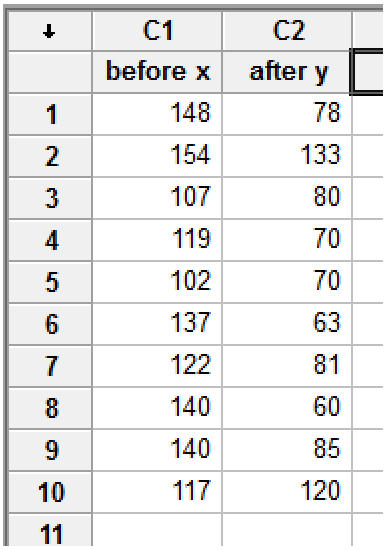
Before surgery (X)	**148**	**154**	**107**	**119**	**102**	**137**	**122**	**140**	**140**	**117**
After surgery (Y)	**78**	**133**	**80**	**70**	**70**	**63**	**81**	**60**	**85**	**120**

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

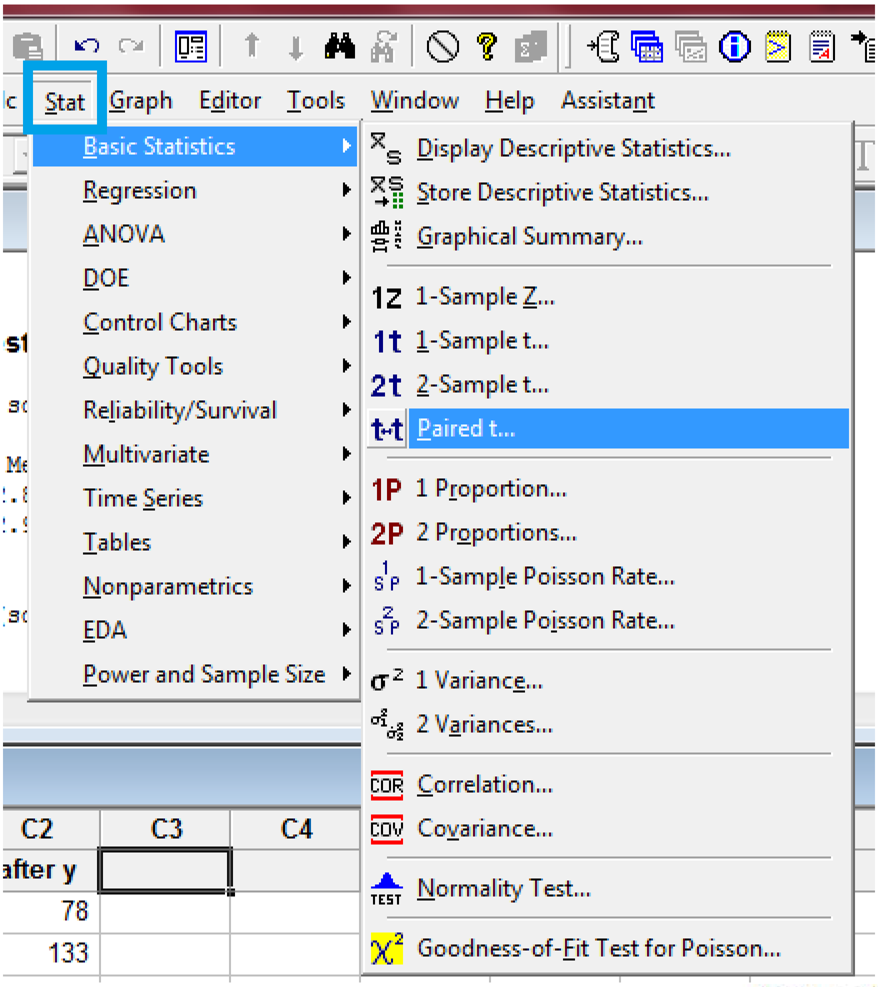
a) **Test whether the data provide sufficient evidence to indicate a difference in the average weight before and after surgery. ($\mu_D = 0$ versus $\mu_D \neq 0$)**

b) **Find 90% 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.

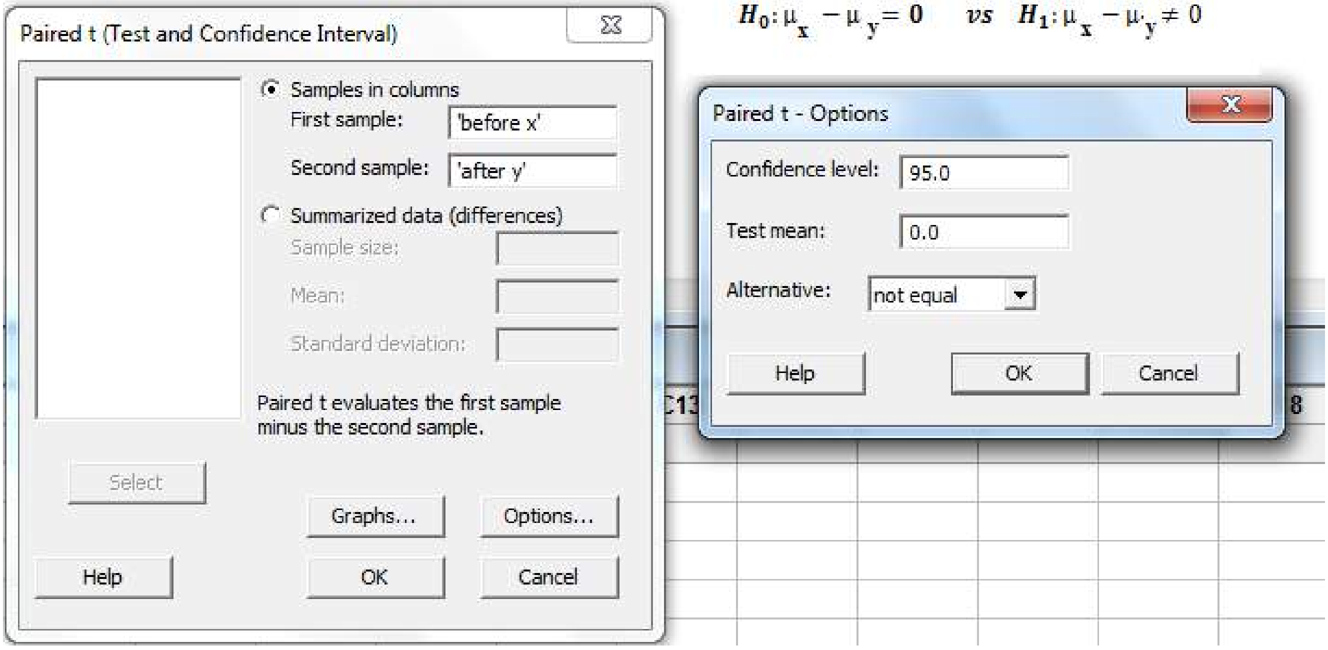


↓	C1	C2
	before x	after y
1	148	78
2	154	133
3	107	80
4	119	70
5	102	70
6	137	63
7	122	81
8	140	60
9	140	85
10	117	120
11		



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

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


Paired t (Test and Confidence Interval)

● Samples in columns

First sample: 'before x'

Second sample: 'after y'

○ Summarized data (differences)

Sample size:

Mean:

Standard deviation:

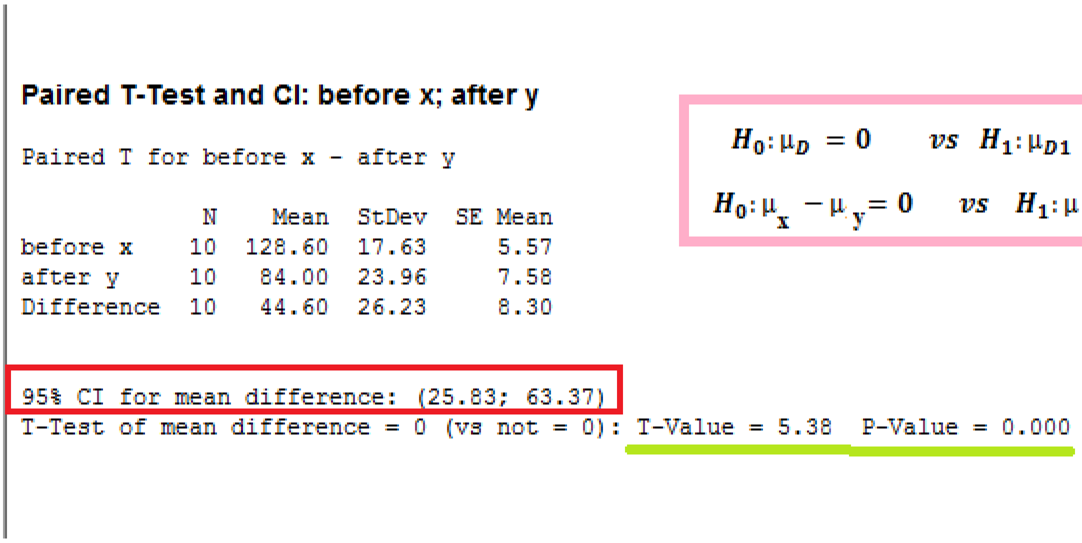
Paired t evaluates the first sample minus the second sample.

Paired t - Options

Confidence level: 95.0

Test mean: 0.0

Alternative: not equal



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
```

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

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

a)

1- **Hypothesis:**

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

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.

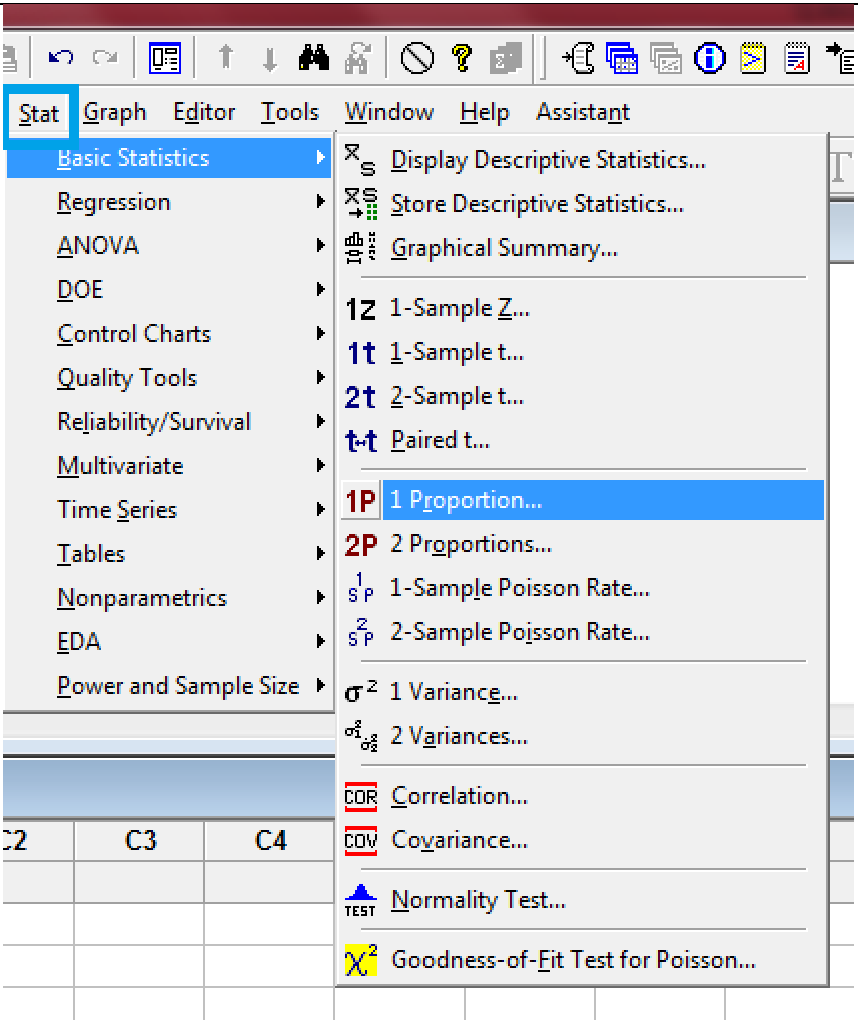
a) **Would you agree with this claim if a random survey shows that 24 out of 45 patients are females? α=0.1**
b) **Find a 90% confidence interval for the true proportion of females**

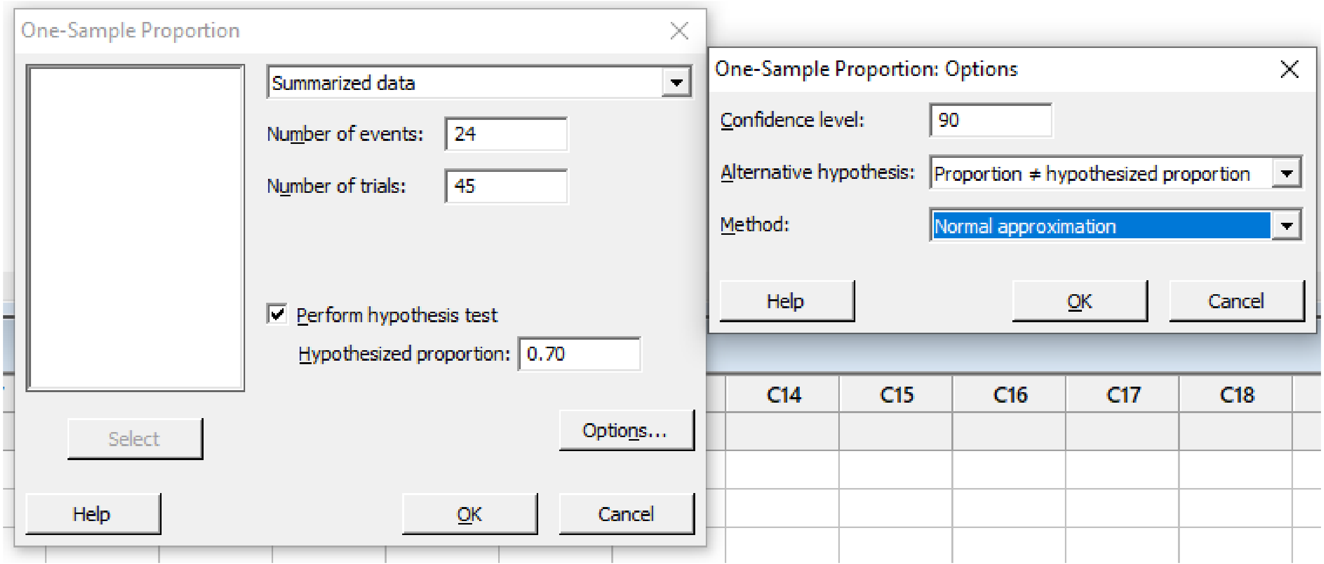
Use the 1 proportion test to **estimate the proportion of a population** and **compare it to a target or reference value.**

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.





One-Sample Proportion

Summarized data

Number of events: 24

Number of trials: 45

☑ Perform hypothesis test

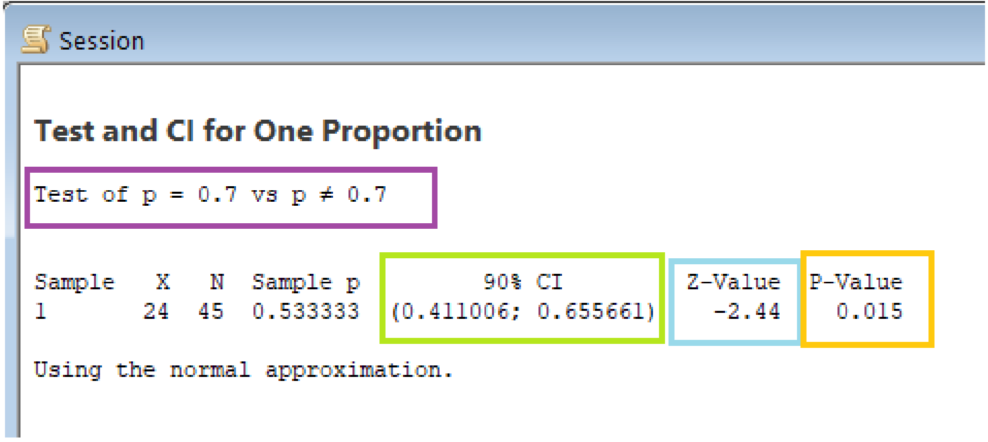
Hypothesized proportion: 0.70

One-Sample Proportion: Options

Confidence level: 90

Alternative hypothesis: Proportion ≠ hypothesized proportion

Method: Normal approximation



```
Test and CI for One Proportion

Test of p = 0.7 vs p ≠ 0.7

Sample   X   N  Sample p          90% CI          Z-Value  P-Value
1       24  45  0.533333  (0.411006; 0.655661)    -2.44    0.015

Using the normal approximation.
```

p: event proportion

Normal approximation method is used for this analysis.

a)

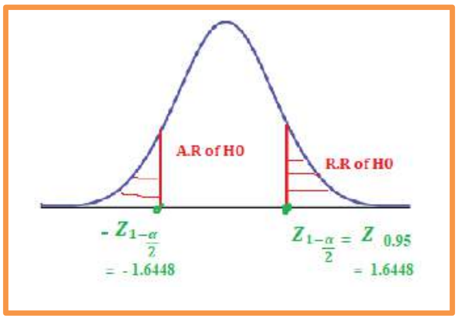
1- **Hypothesis:**

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

2- **Test statistic :**

Z= -2.44

3- **z critical** = 1.645



4- **conclusion is:**

Since p-value =0.015 < α= 0.05 . 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

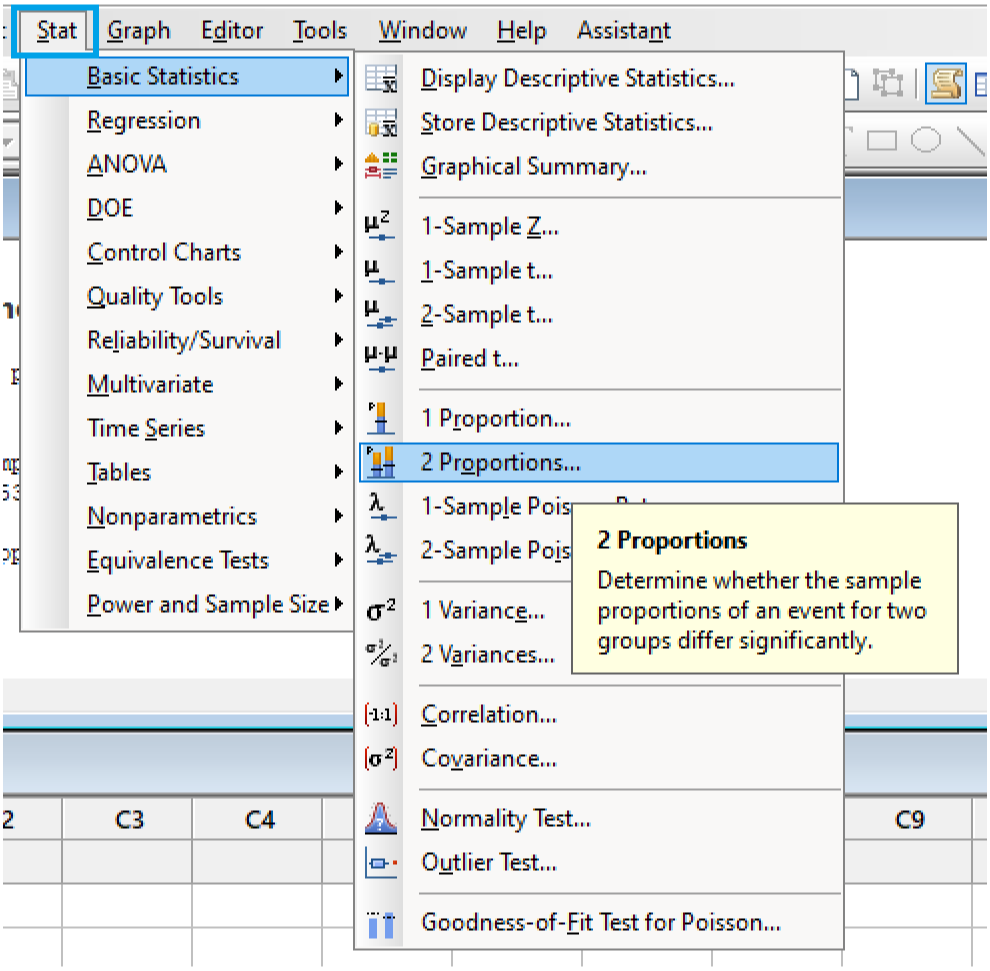
In a study about the obesity (overweight), a researcher was interested in comparing the proportion of obesity between males and females. The researcher has obtained a random sample of 150 males and another independent random sample of 200 females. The following results were obtained from this study

	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.**

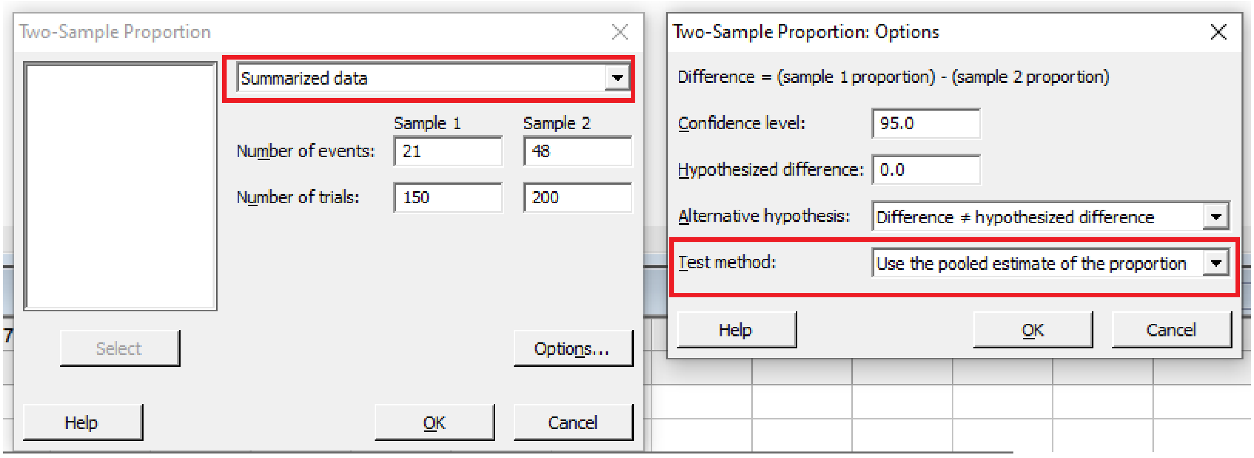
b) **Find a 95% confidence interval for the difference between the two proportions.**

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



2 Proportions

Determine whether the sample proportions of an event for two groups differ significantly.



Two-Sample Proportion

Summarized data

	Sample 1	Sample 2
Number of events:	21	48
Number of trials:	150	200

Two-Sample Proportion: Options

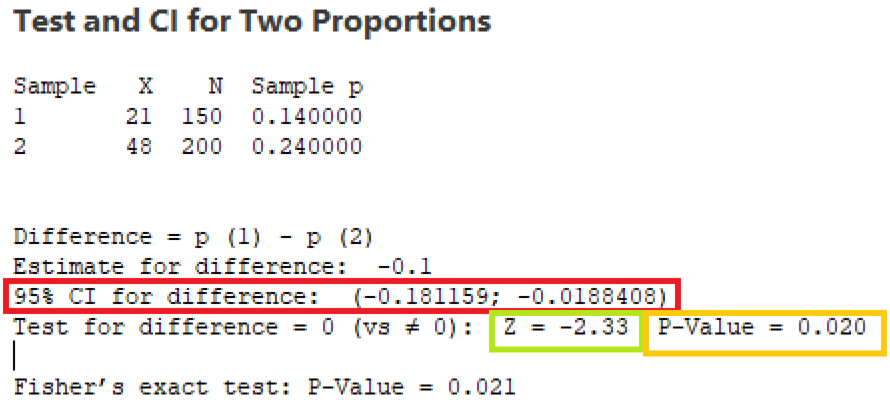
Difference = (sample 1 proportion) - (sample 2 proportion)

Confidence level: 95.0

Hypothesized difference: 0.0

Alternative hypothesis: Difference ≠ hypothesized difference

Test method: Use the pooled estimate of the proportion



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)

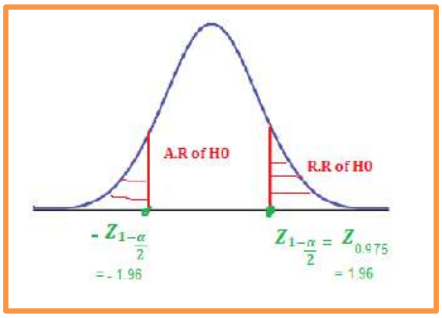
1- **Hypothesis:**

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

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)

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

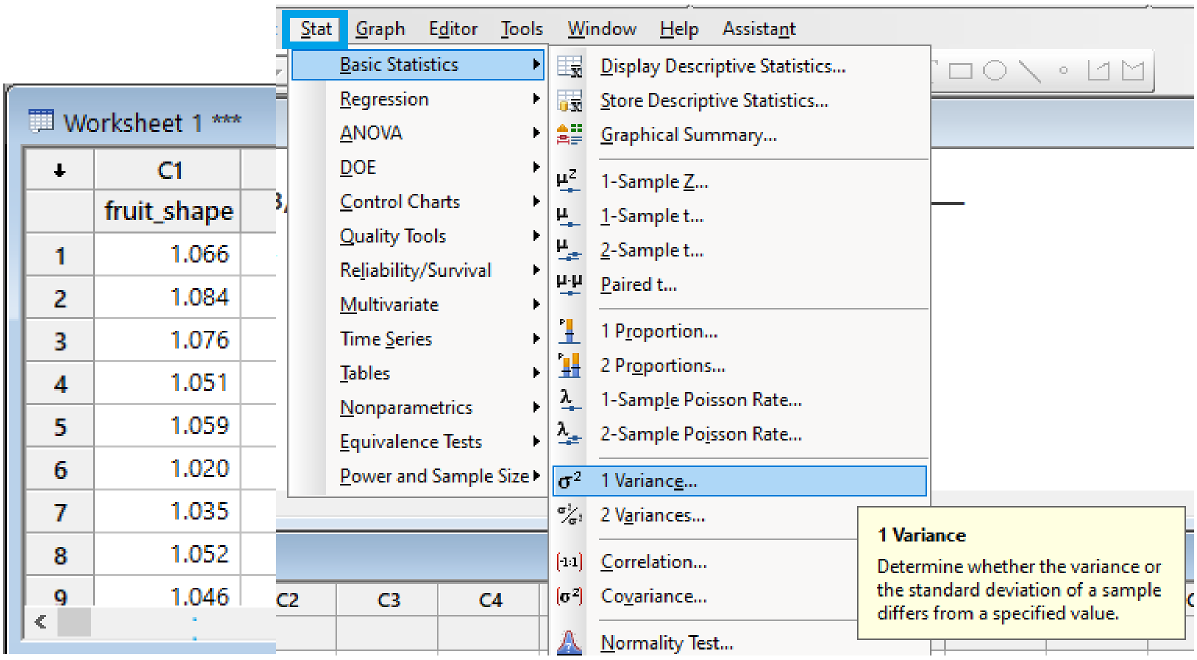
ملاحظه: في حالة فترات الثقة يكون اختيار الفرض الاحصائي لا يساوي

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)]:

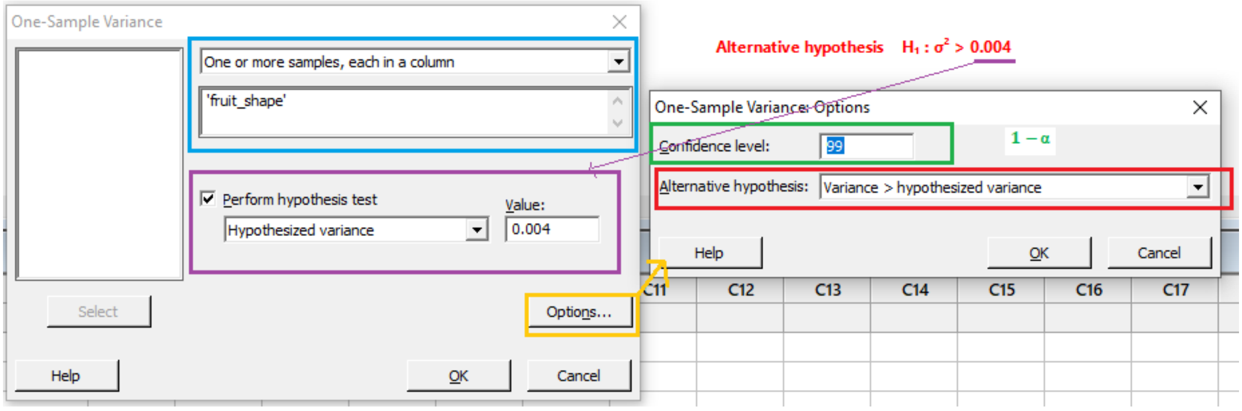
1.066 1.084 1.076 1.051 1.059 1.020 1.035 1.052 1.046 0.976

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



	C1
	fruit_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

1 Variance
Determine whether the variance or the standard deviation of a sample differs from a specified value.



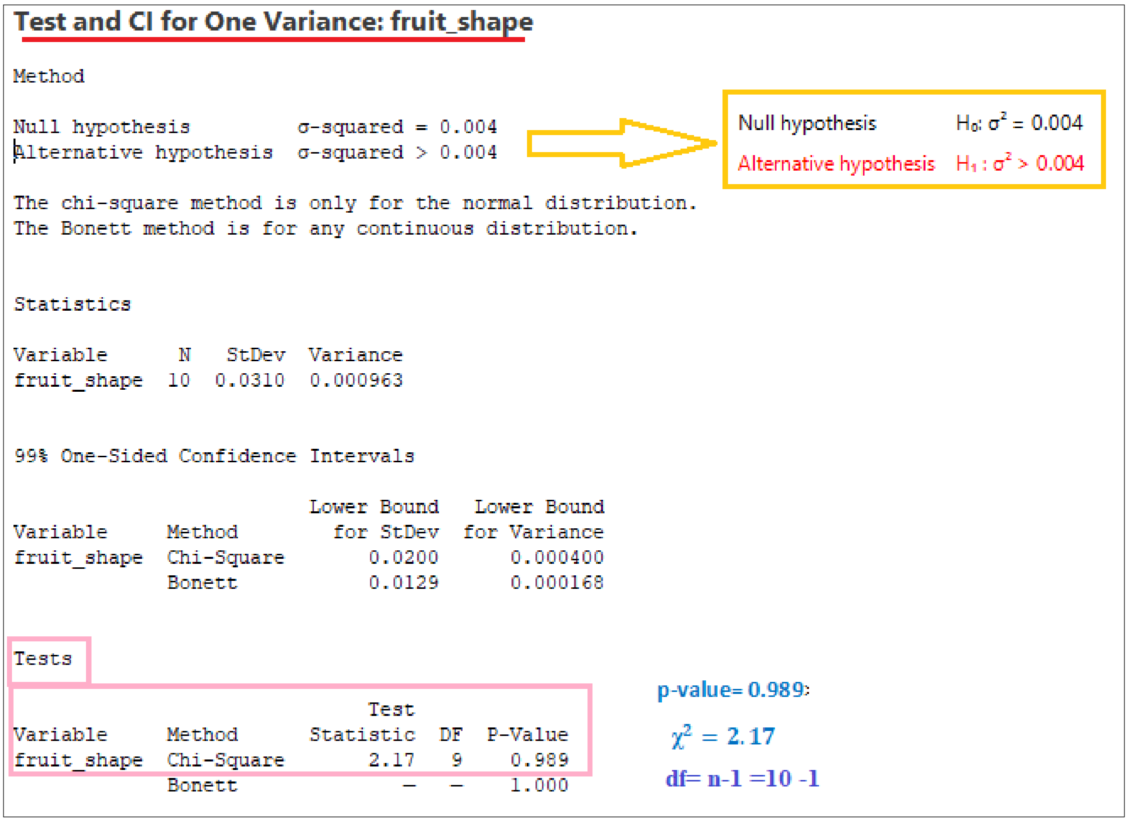
One or more samples, each in a column
'fruit_shape'

☑ Perform hypothesis test
Hypothesized variance
Value: 0.004

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

Confidence level: 99 $1-\alpha$

Alternative hypothesis: Variance > hypothesized variance



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

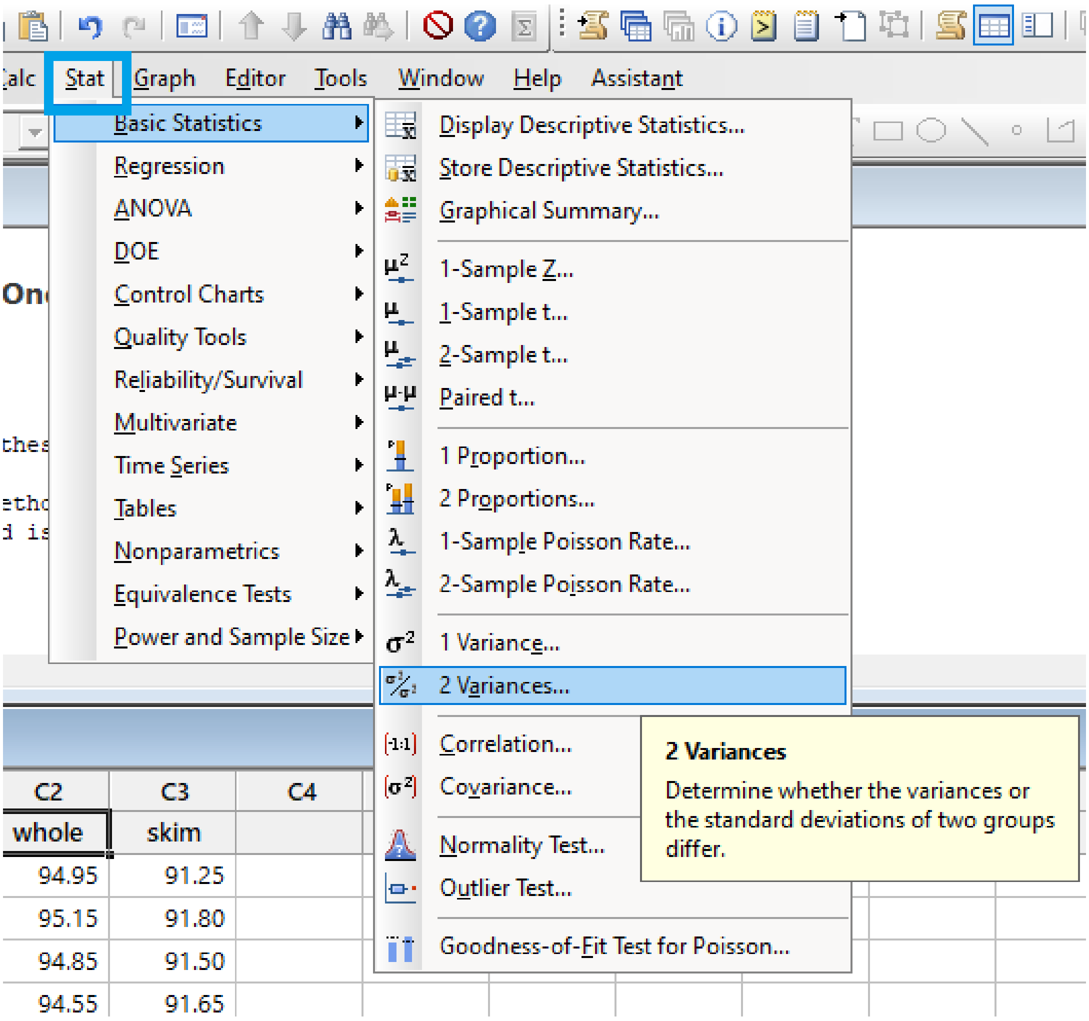
Two sample variance

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

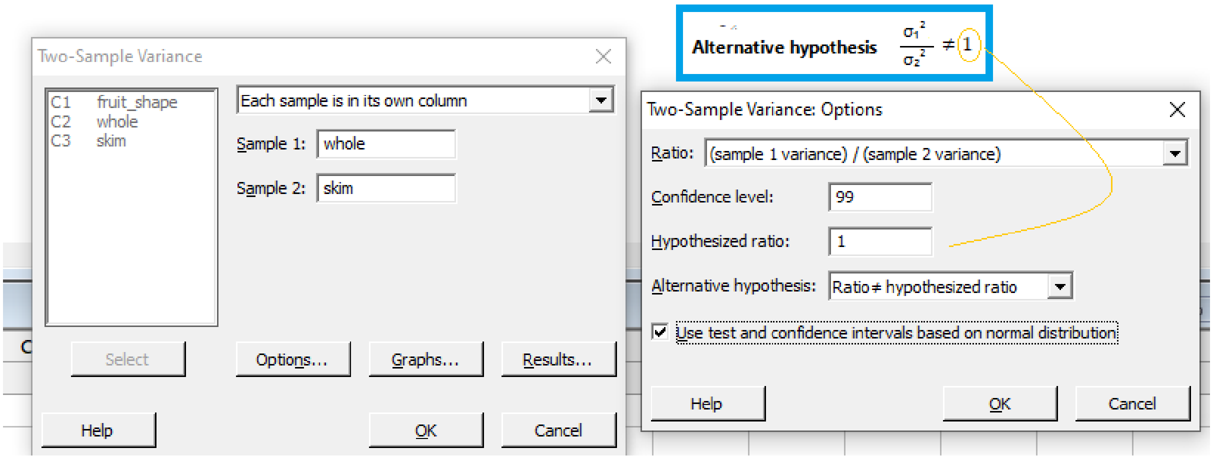
Whole	**94.95**	**95.15**	**94.85**	**94.55**	**94.55**	**93.40**	**95.05**	**94.35**	**94.70**	**94.90**
Skim	**91.25**	**91.80**	**91.50**	**91.65**	**91.15**	**90.25**	**91.90**	**91.25**	**91.65**	**91.00**

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



C2	C3	C4
whole	skim	
94.95	91.25	
95.15	91.80	
94.85	91.50	
94.55	91.65	



Alternative hypothesis $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

Two-Sample Variance

C1 fruit_shape
C2 whole
C3 skim

Each sample is in its own column

Sample 1: whole

Sample 2: skim

Two-Sample Variance: Options

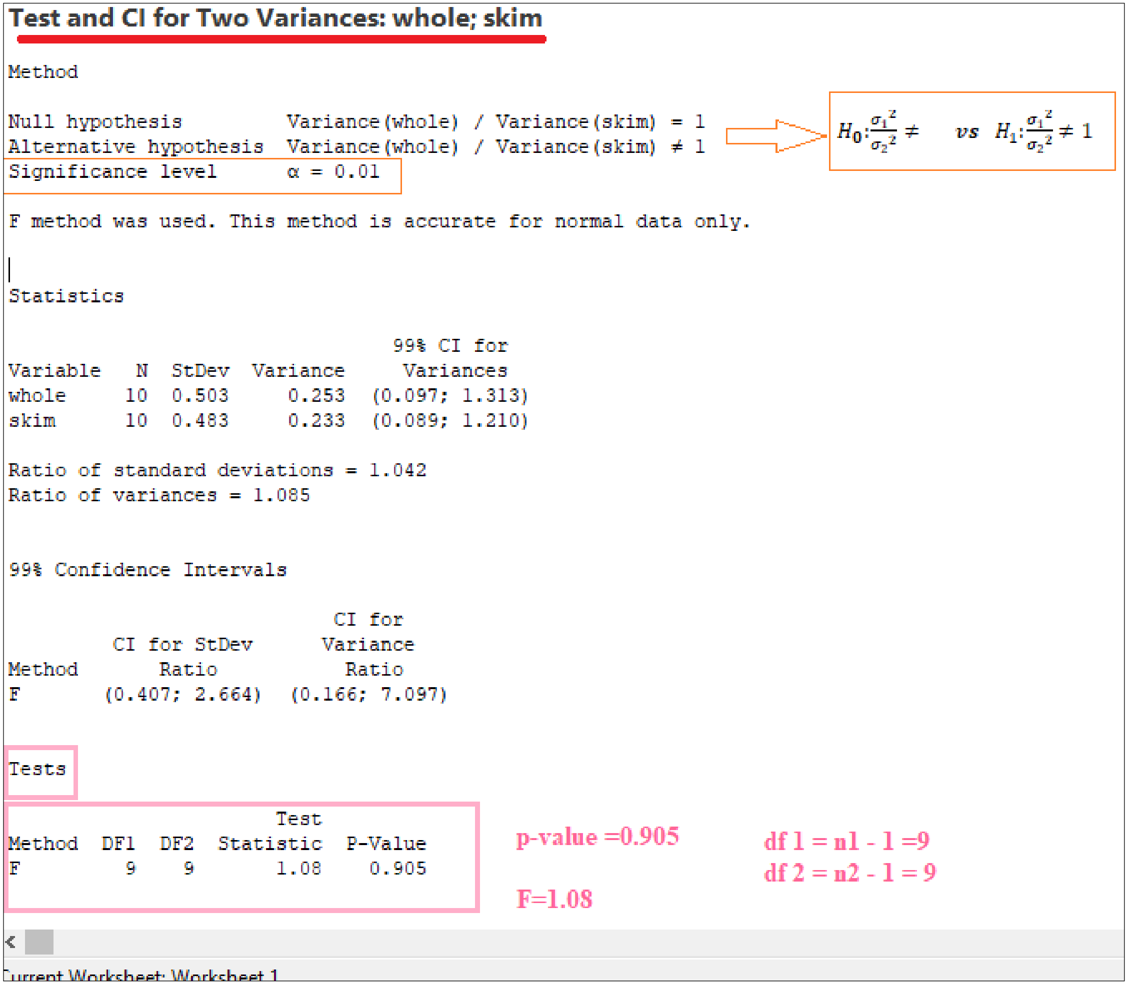
Ratio: (sample 1 variance) / (sample 2 variance)

Confidence level: 99

Hypothesized ratio: 1

Alternative hypothesis: Ratio ≠ hypothesized ratio

☑ Use test and confidence intervals based on normal distribution



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

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
```

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

p-value =0.905

F=1.08

df 1 = n1 - 1 =9

df 2 = n2 - 1 = 9

Current Worksheet: Worksheet 1

1- **Hypothesis :**

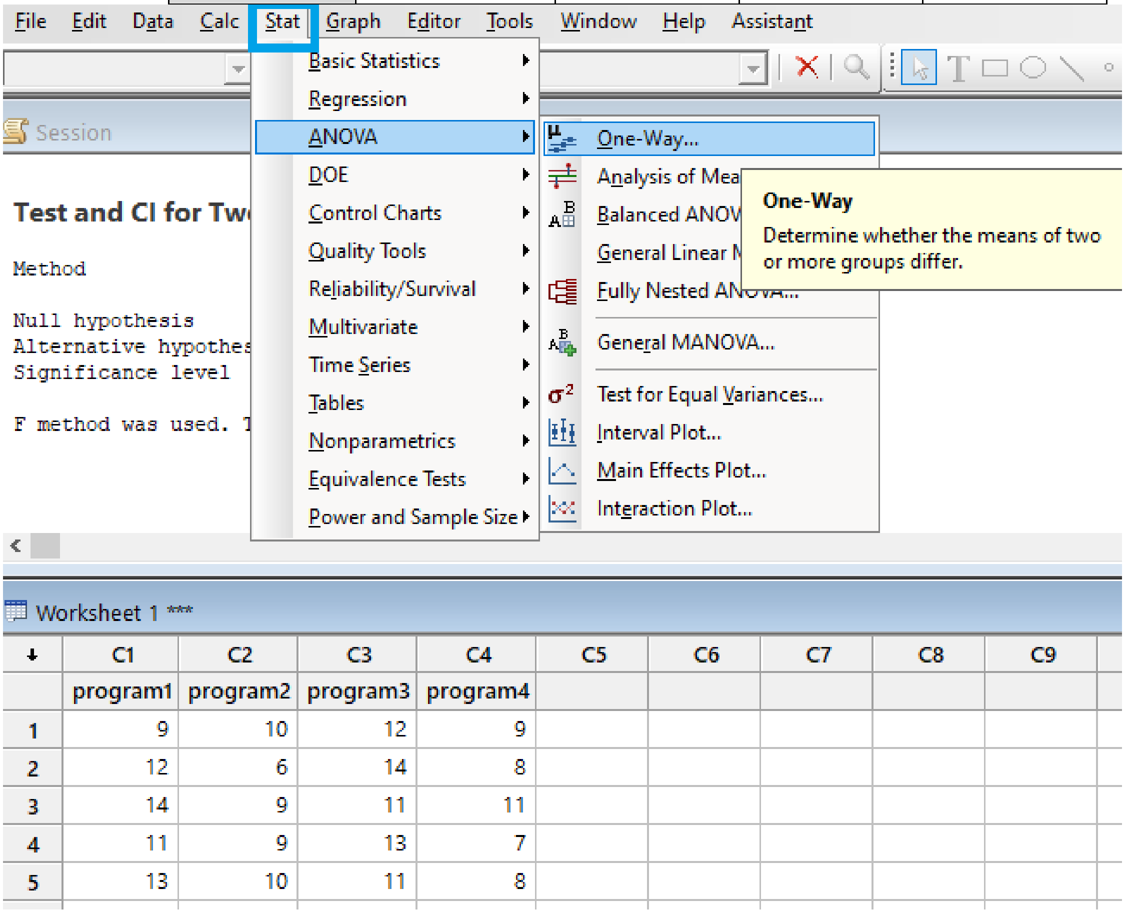
$$H_0: \frac{\sigma_1^2}{\sigma_2^2} \neq \quad \text{vs} \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

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

ANOVA

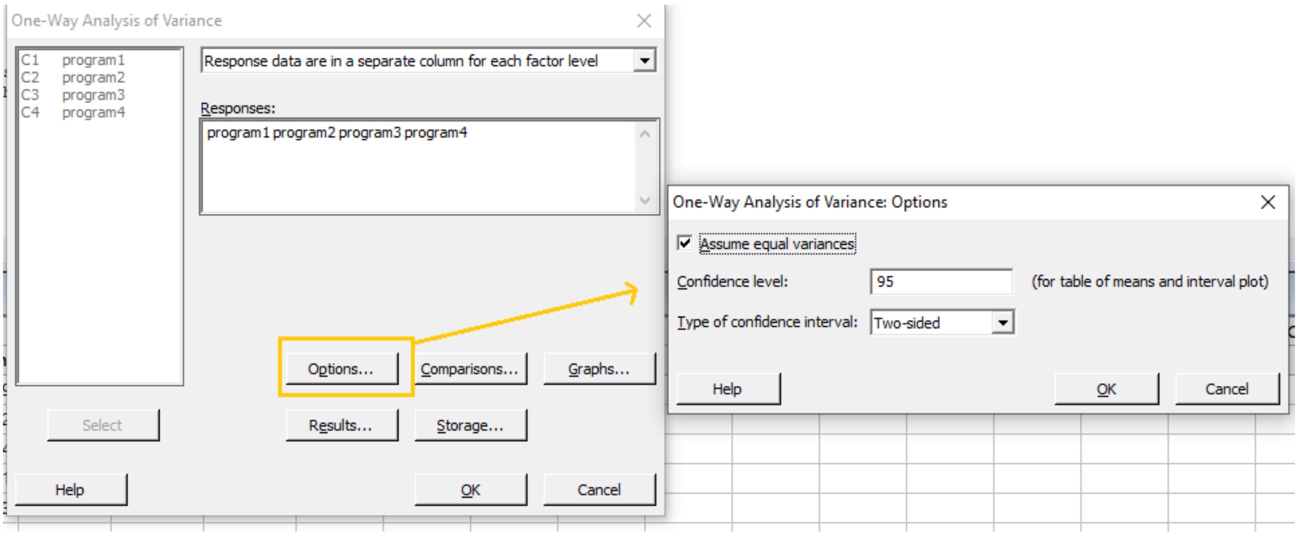
Q: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number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



```
Test and CI for Tw

Method

Null hypothesis
Alternative hypothes
Significance level

F method was used. T
```

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9
	program1	program2	program3	program4					
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					



Response data are in a separate column for each factor level

Responses: program1 program2 program3 program4

☑ Assume equal variances

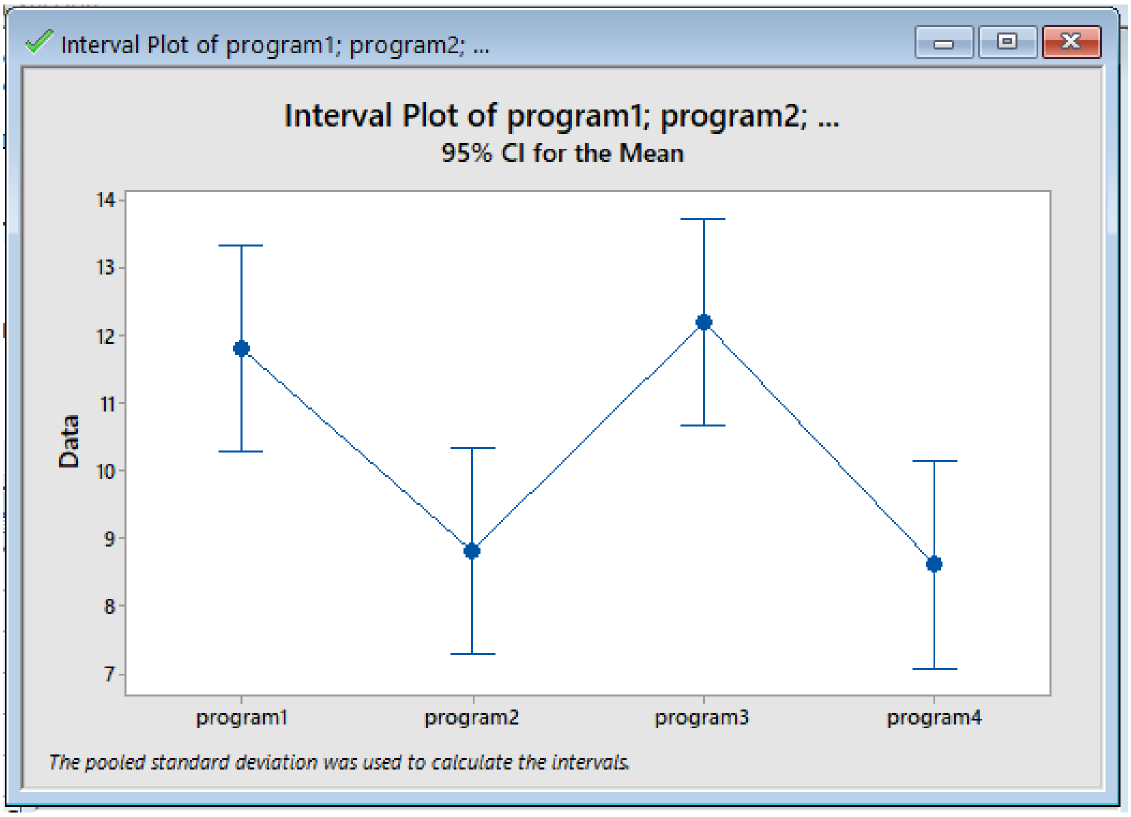
Confidence level: 95 (for table of means and interval plot)

Type of confidence interval: Two-sided

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

Model Summary

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

p-value = 0.003 < α= 0.05



1-Hypothesis :

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

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

2- Test statistic :

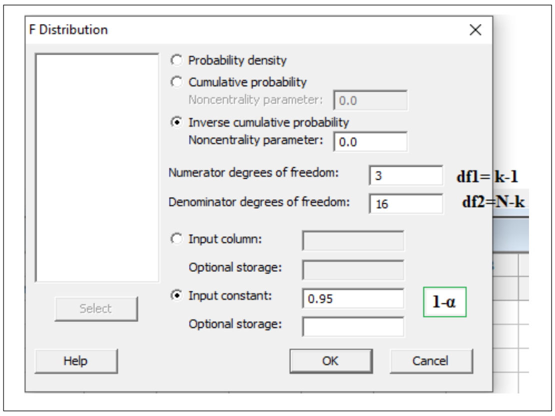
F = 7.04

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

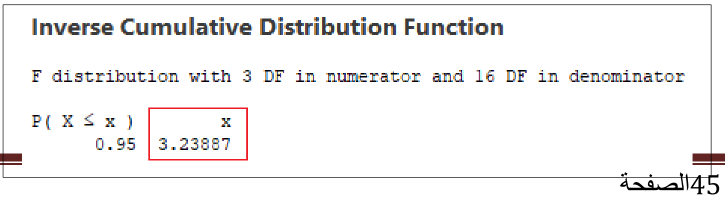
Calc>> probability distributions>>F

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

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



○ Probability density
○ Cumulative probability
Noncentrality parameter: 0.0
● Inverse cumulative probability
Noncentrality parameter: 0.0
Numerator degrees of freedom: 3 **df1=k-1**
Denominator degrees of freedom: 16 **df2=N-k**
○ Input column:
Optional storage:
● Input constant: 0.95 **1-α**
Optional storage:

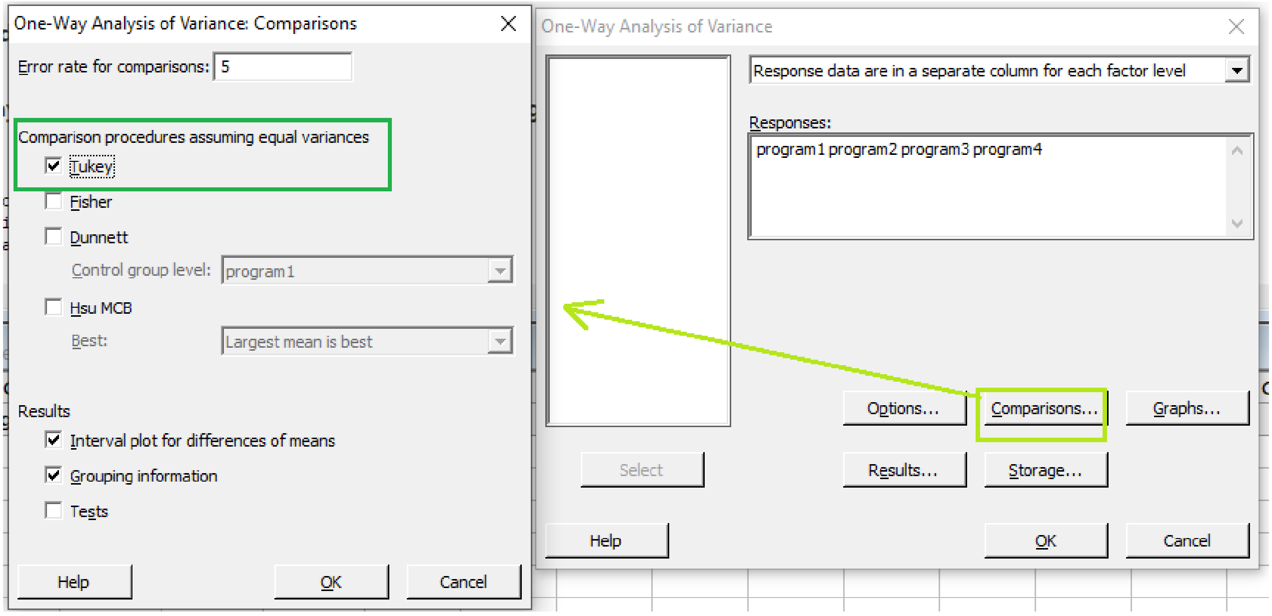


Inverse Cumulative Distribution Function

```
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



One-Way Analysis of Variance: Comparisons

Error rate for comparisons: 5

Comparison procedures assuming equal variances

☑ Tukey

☐ Fisher

☐ Dunnett

Control group level: program1

☐ Hsu MCB

Best: Largest mean is best

Results

☑ Interval plot for differences of means

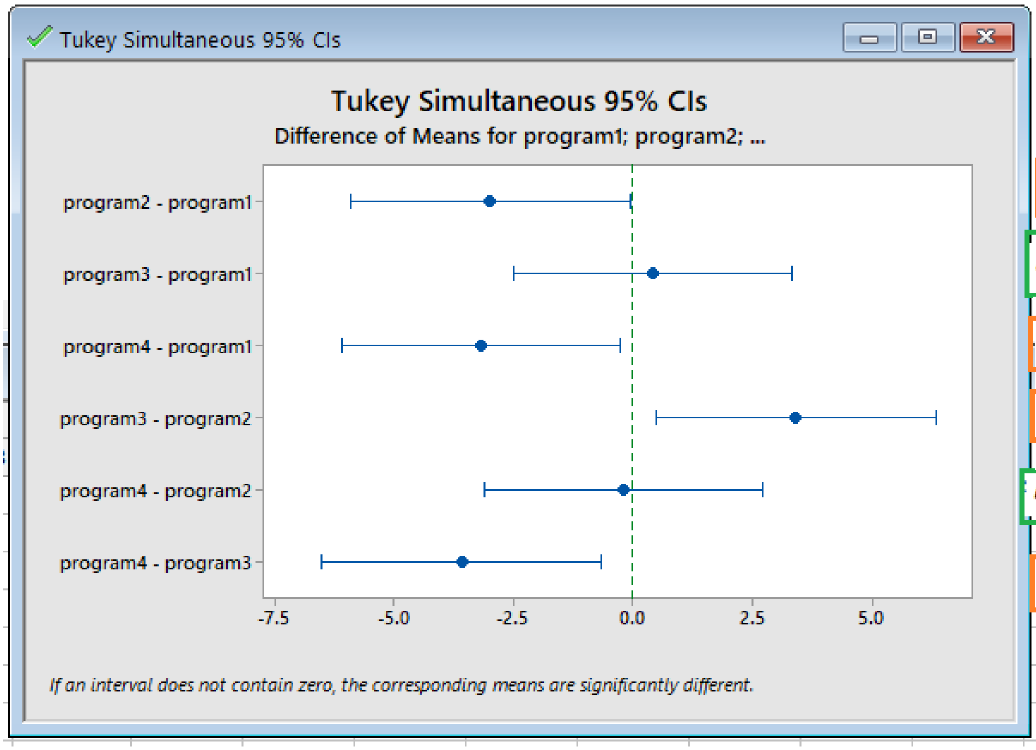
☑ Grouping information

☐ Tests

One-Way Analysis of Variance

Response data are in a separate column for each factor level

Responses: program1 program2 program3 program4



If an interval does not contain zero, the corresponding means are significantly different.

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

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

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

$\mu_{program\ 2} \neq \mu_{program\ 4}$

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

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

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

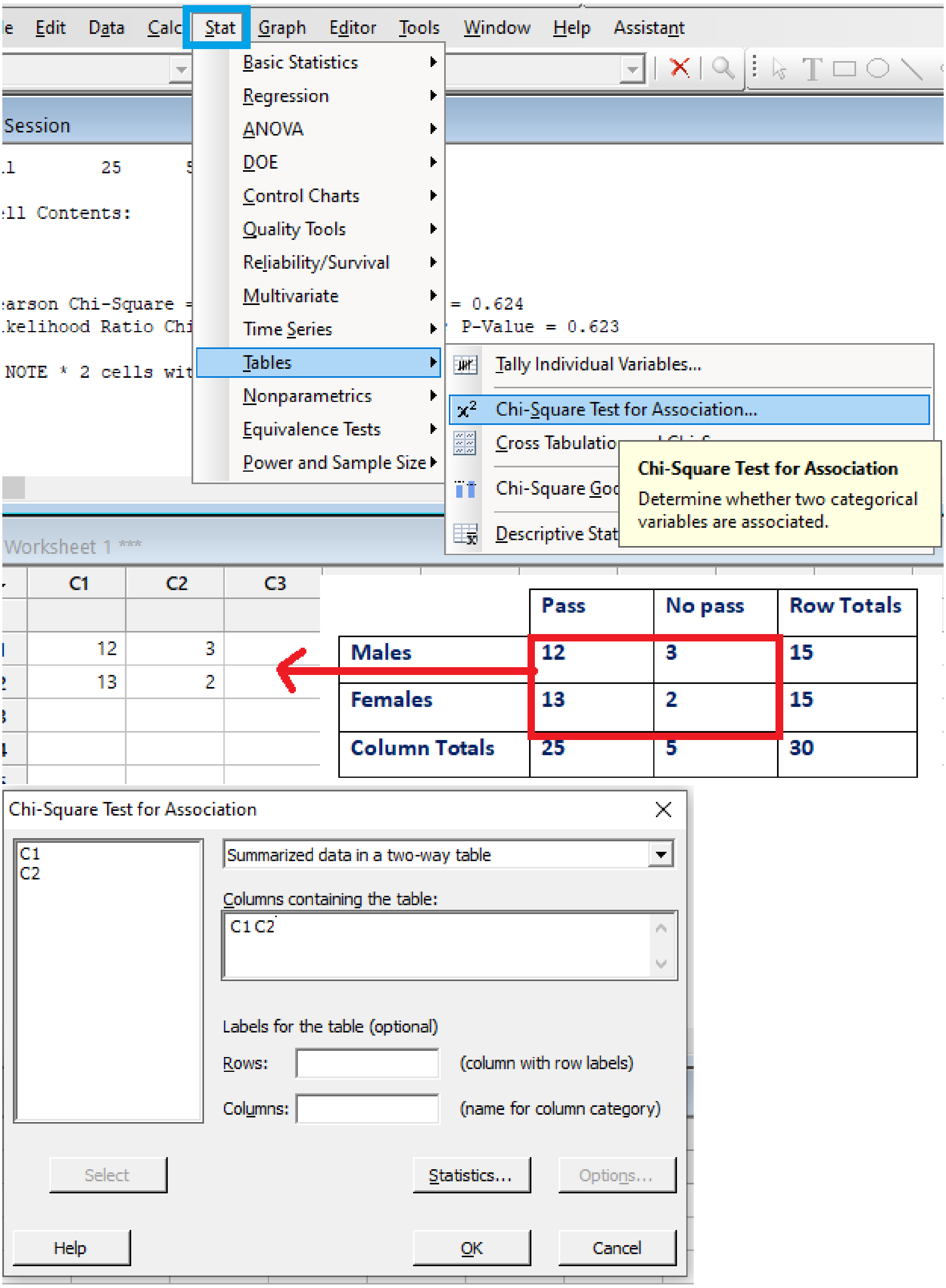
1-Hypothesis :

$\mathbf{H_0}$: **the gender of the students is independent of pass or no pass test grade**

$\mathbf{H_1}$: **the gender of the students is not independent of pass or no pass test grade**

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

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



```
.1          25      5

.ll Contents:


.arson Chi-Square =                    = 0.624
.kelihood Ratio Chi                    P-Value = 0.623

NOTE * 2 cells wit
```

	C1	C2	C3
1	12	3	
2	13	2	
3			
4			

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

Chi-Square Test for Association

C1
C2

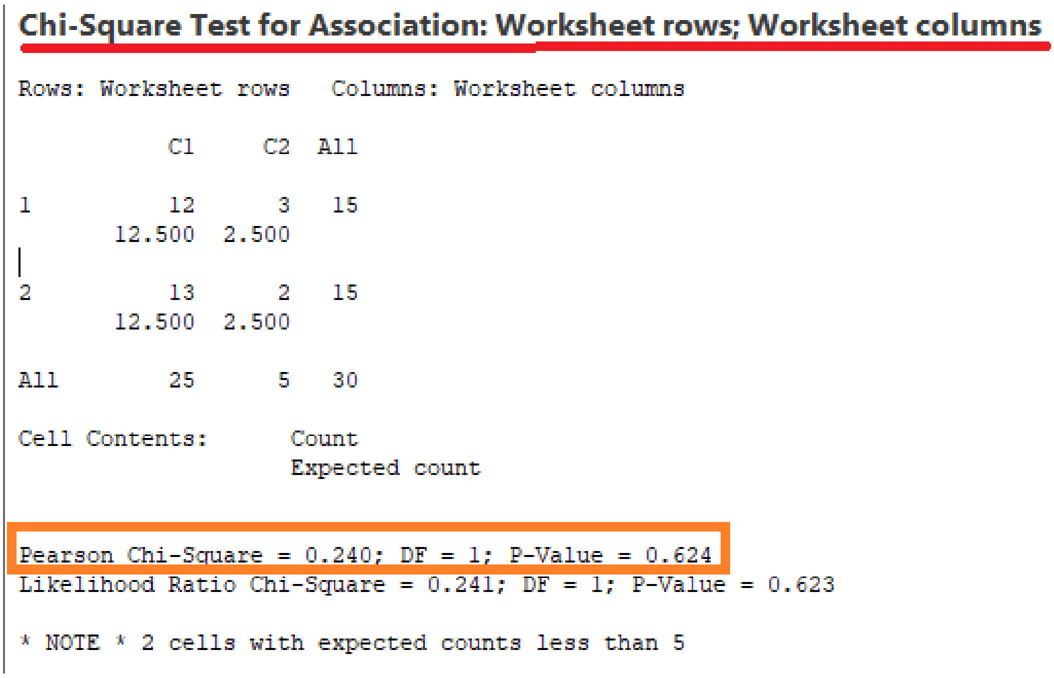
Summarized data in a two-way table

Columns containing the table: C1 C2

Labels for the table (optional)

Rows: (column with row labels)

Columns: (name for column category)



Chi-Square Test for Association: Worksheet rows; Worksheet columns

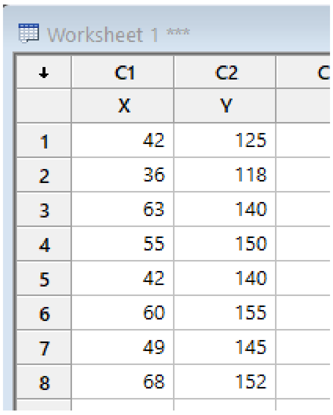
```
Rows: Worksheet rows   Columns: Worksheet columns

           C1      C2  All

1          12       3   15
       12.500   2.500

2          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.

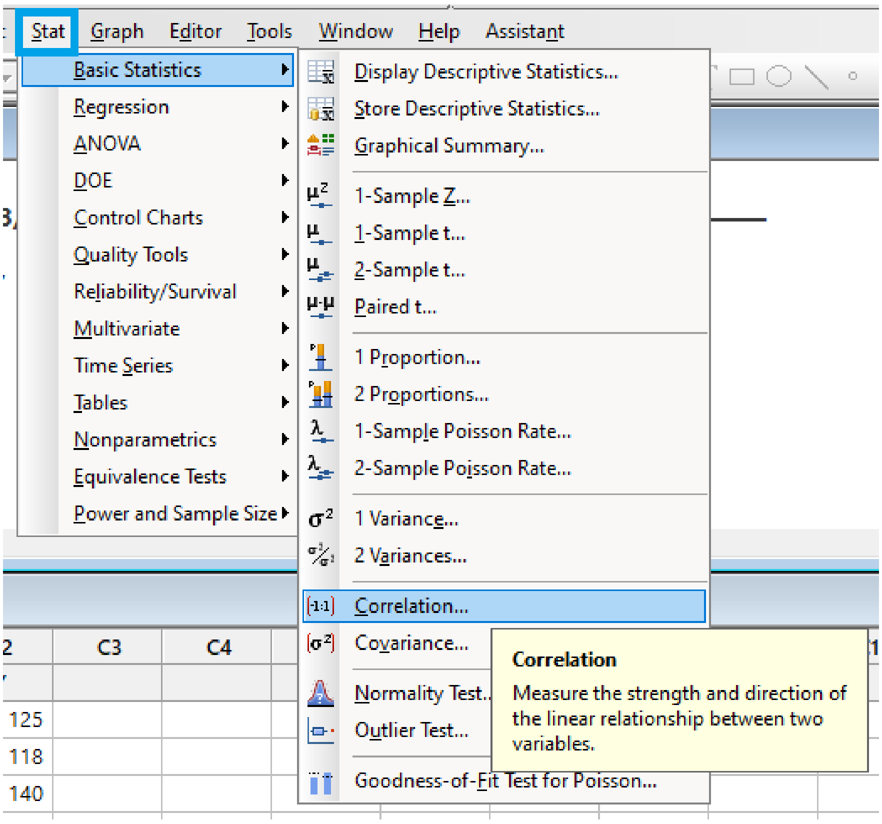
X	**42**	**36**	**63**	**55**	**42**	**60**	**49**	**68**
Y	**125**	**118**	**140**	**150**	**140**	**155**	**145**	**152**

Find the correlation coefficient between x and y

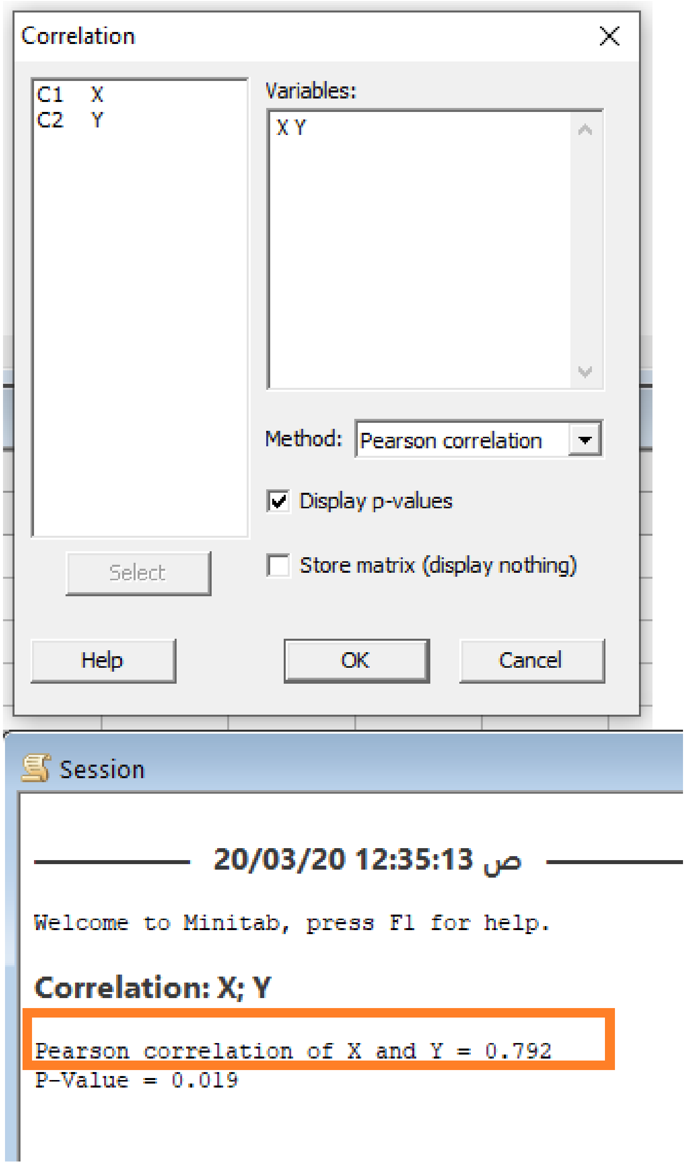


Worksheet 1 ***

↓	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



Correlation
Measure the strength and direction of the linear relationship between two variables.



Variables: X Y

Method: Pearson correlation

☑ Display p-values

☐ Store matrix (display nothing)

```
———— 20/03/20 12:35:13 ص ————

Welcome to Minitab, press F1 for help.
```

Correlation: X; Y

```
Pearson correlation of X and Y = 0.792
P-Value = 0.019
```

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

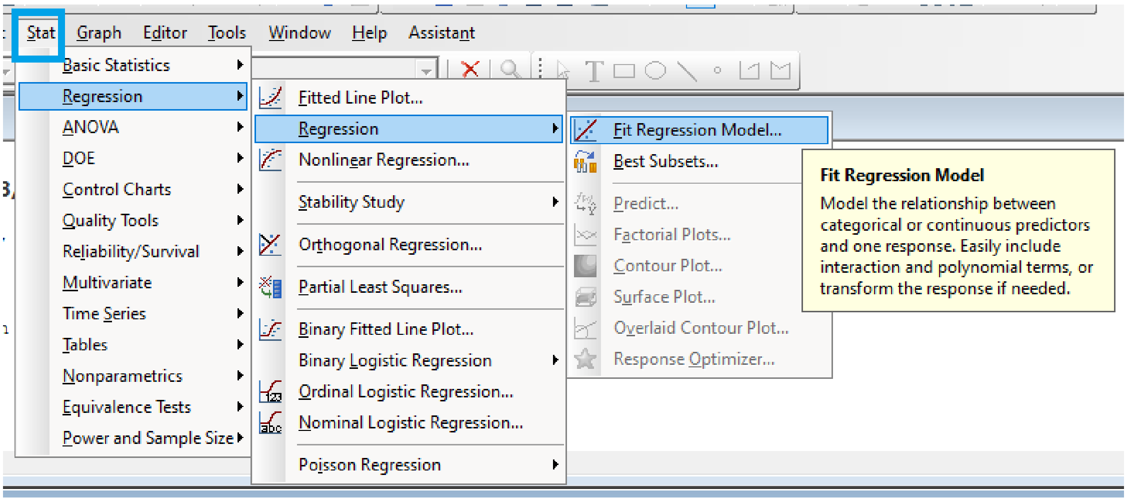
a) Determine the regression equation for the data.

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.

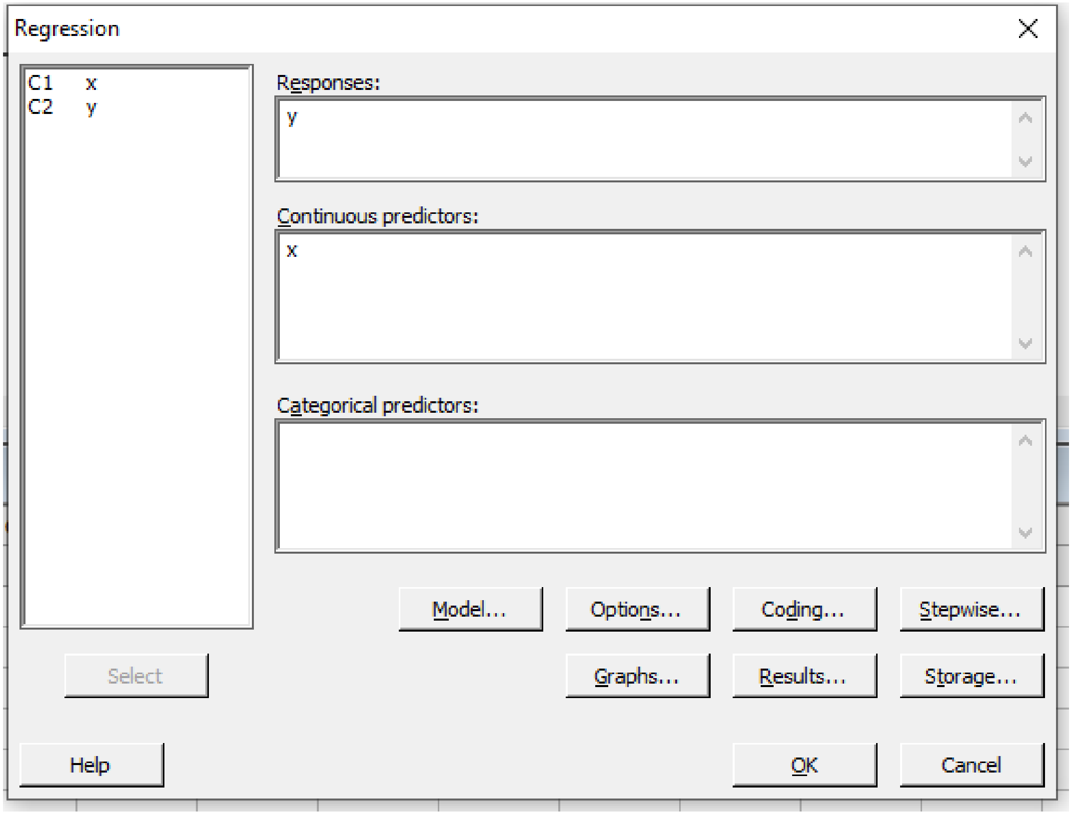
(Ans **b)** $R^2 = 0.9368 \rightarrow 93.68\,\%$ *of the variation in y data is explained by x*)

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



Fit Regression Model

Model the relationship between categorical or continuous predictors and one response. Easily include interaction and polynomial terms, or transform the response if needed.



C1 x
C2 y

Responses: y

Continuous predictors: x

Categorical predictors:

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

      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


Regression Equation

y = 291.6 - 27.90 x
```

b)

a)

Department of Statistics and Operations Research

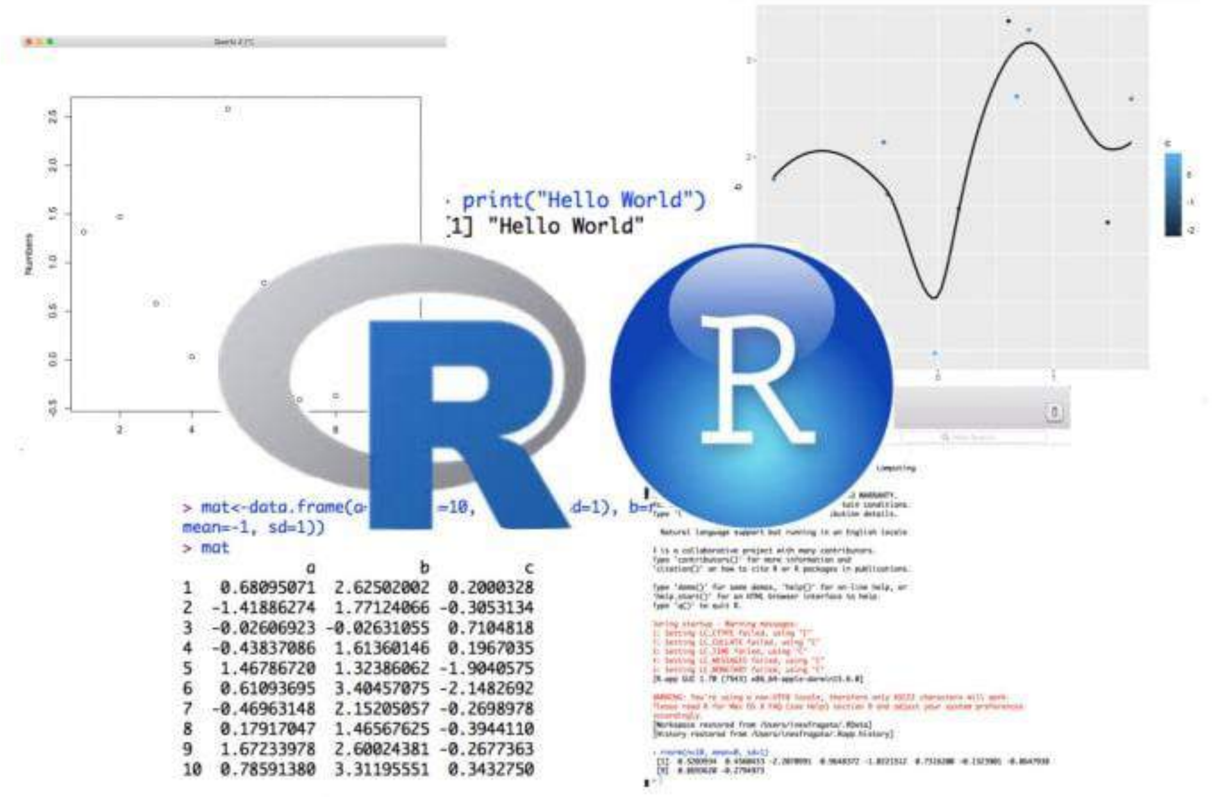
College of Science

King Saud University

Tutorial

STATISTICAL PACKAGES(R)

STAT 328



```
· print("Hello World")
[1] "Hello World"
```

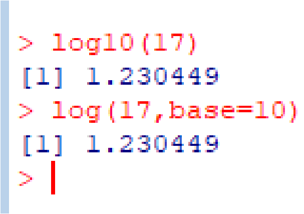
```
> mat<-data.frame(a-          =10,          d=1), b=r
mean=-1, sd=1))
> mat
            a           b          c
1   0.68095071  2.62502002  0.2000328
2  -1.41886274  1.77124066 -0.3053134
3  -0.02606923 -0.02631055  0.7104818
4  -0.43837086  1.61360146  0.1967035
5   1.46786720  1.32386062 -1.9040575
6   0.61093695  3.40457075 -2.1482692
7  -0.46963148  2.15205057 -0.2698978
8   0.17917047  1.46567625 -0.3944110
9   1.67233978  2.60024381 -0.2677363
10  0.78591380  3.31195551  0.3432750
```

R-Part 1

#Mathematical functions :

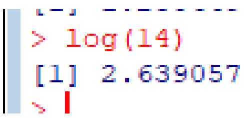
Q1: Write the command and the result to calculate the following :

Log(17)=



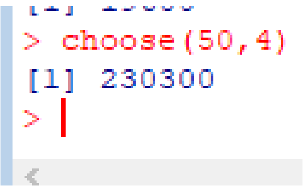
```
> log10(17)
[1] 1.230449
> log(17,base=10)
[1] 1.230449
>
```

Ln(14)=



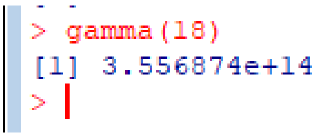
```
> log(14)
[1] 2.639057
>
```

$\binom{50}{4} =$



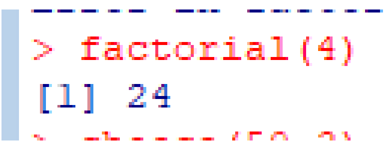
```
> choose(50,4)
[1] 230300
>
```

$\Gamma(18)$,



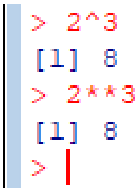
```
> gamma(18)
[1] 3.556874e+14
>
```

4!=



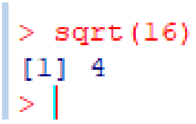
```
> factorial(4)
[1] 24
```

$2^3 =$



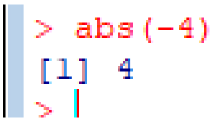
```
> 2^3
[1] 8
> 2**3
[1] 8
>
```

$\sqrt{16} =$



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

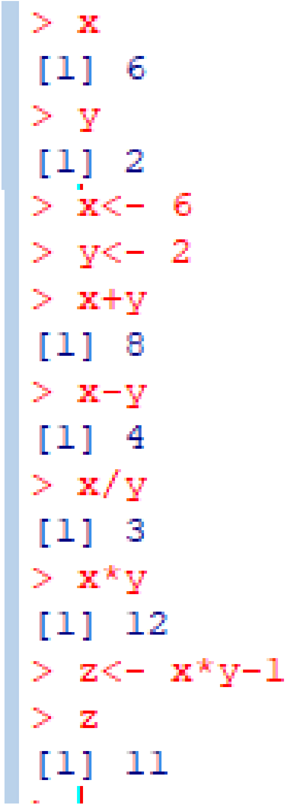
|-4| =



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

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

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

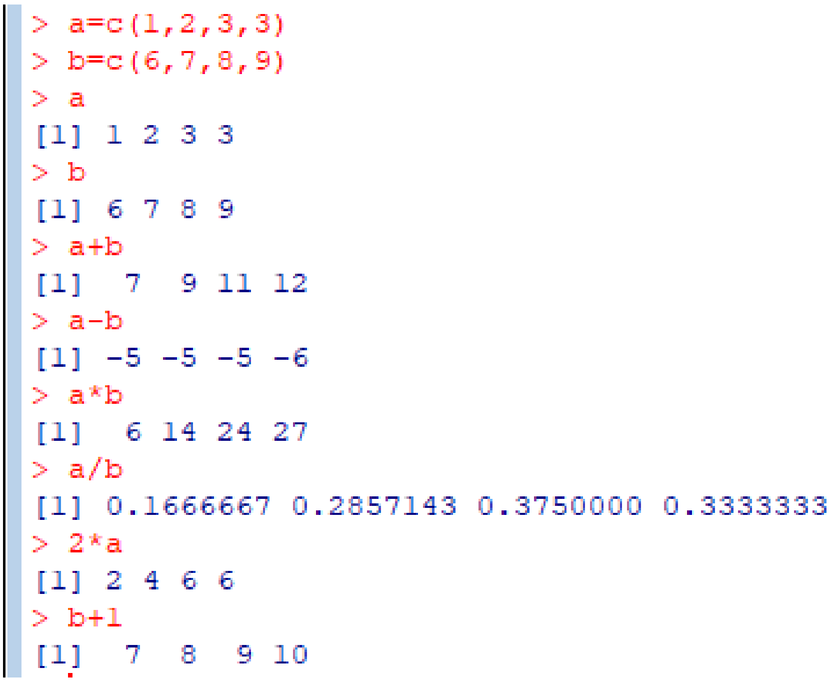


```
> x
[1] 6
> y
[1] 2
> x<- 6
> y<- 2
> x+y
[1] 8
> x-y
[1] 4
> x/y
[1] 3
> x*y
[1] 12
> z<- x*y-1
> z
[1] 11
>
```

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 , $a \div b$, $2\mathrm{a}$, $\mathrm{b} + 1$



```
> 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.

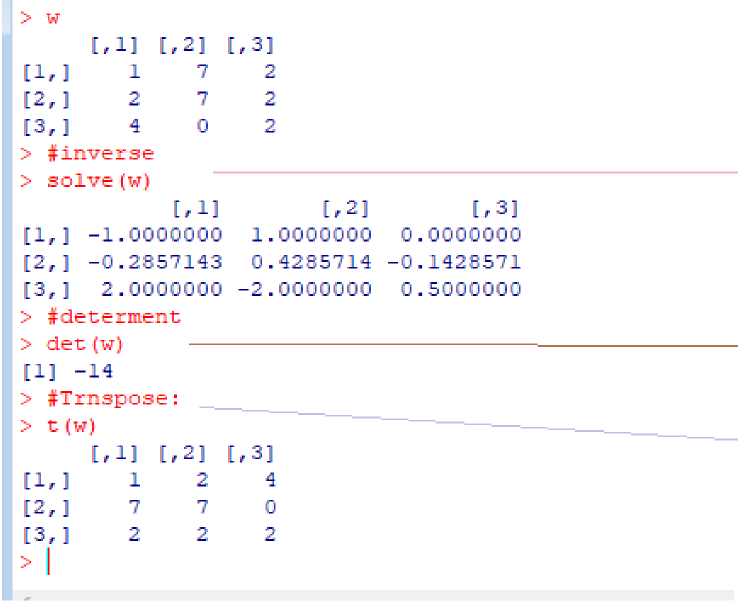
rm() deletes (removes) a variable from a workspace.

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
> 
```

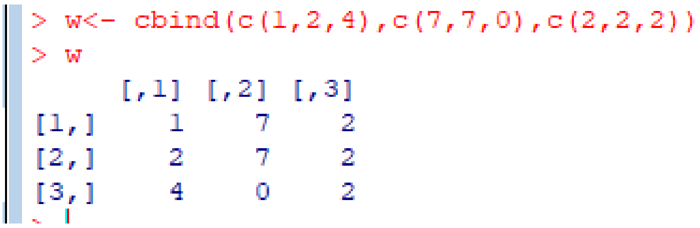
لكتابة المصفوفة نستخدم الامر matrix

لإيجاد المعكوس نستخدم الامر solve

لايجاد محدد المصفوفة نستخدم det

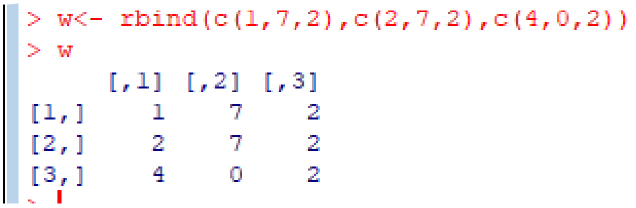
لايجاد منقول المصفوفة نستخدم الامر t

OR



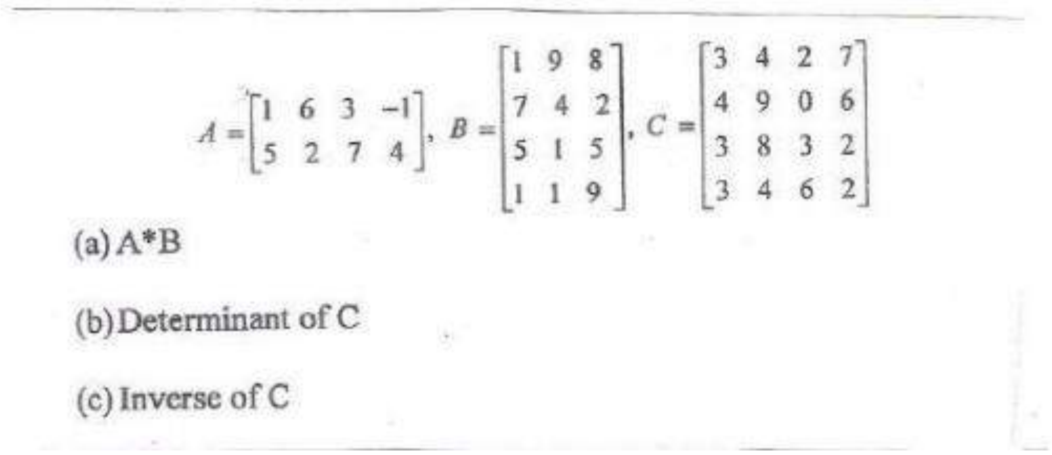
```
> w<- cbind(c(1,2,4),c(7,7,0),c(2,2,2))
> w
     [,1] [,2] [,3]
[1,]    1    7    2
[2,]    2    7    2
[3,]    4    0    2
```

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:

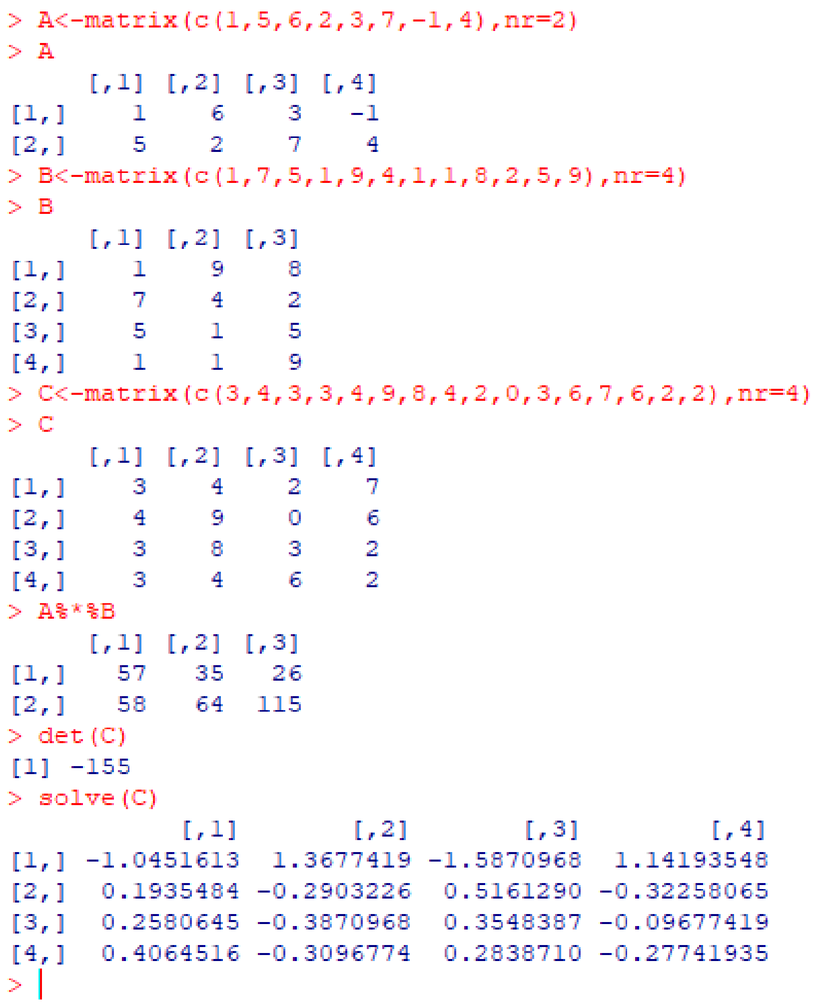


$$A=\begin{bmatrix}1 & 6 & 3 & -1\\ 5 & 2 & 7 & 4\end{bmatrix},\ B=\begin{bmatrix}1 & 9 & 8\\ 7 & 4 & 2\\ 5 & 1 & 5\\ 1 & 1 & 9\end{bmatrix},\ C=\begin{bmatrix}3 & 4 & 2 & 7\\ 4 & 9 & 0 & 6\\ 3 & 8 & 3 & 2\\ 3 & 4 & 6 & 2\end{bmatrix}$$

(a) A*B

(b) Determinant of C

(c) Inverse of C

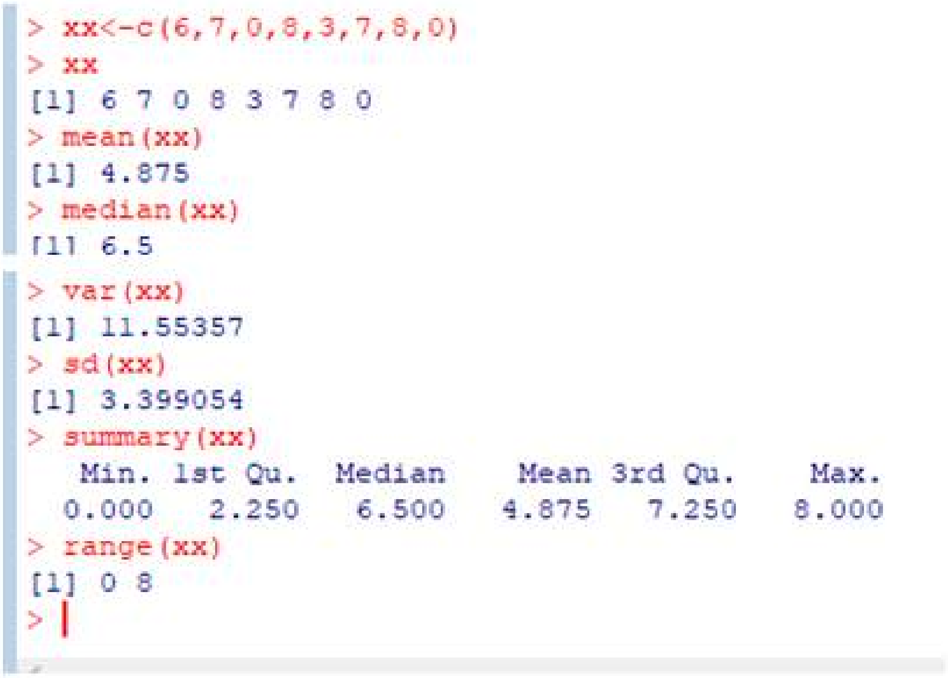


```
> A<-matrix(c(1,5,6,2,3,7,-1,4),nr=2)
> A
     [,1] [,2] [,3] [,4]
[1,]    1    6    3   -1
[2,]    5    2    7    4
> B<-matrix(c(1,7,5,1,9,4,1,1,8,2,5,9),nr=4)
> B
     [,1] [,2] [,3]
[1,]    1    9    8
[2,]    7    4    2
[3,]    5    1    5
[4,]    1    1    9
> C<-matrix(c(3,4,3,3,4,9,8,4,2,0,3,6,7,6,2,2),nr=4)
> 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
> 
```

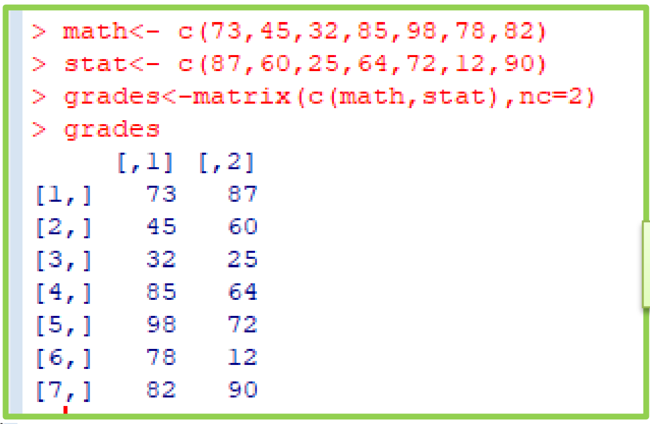
R-Part 2

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

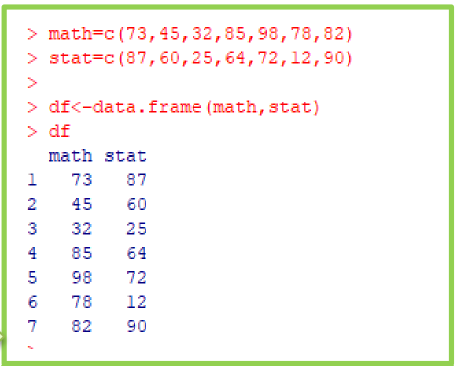
Find

1) summary of math and stat grades

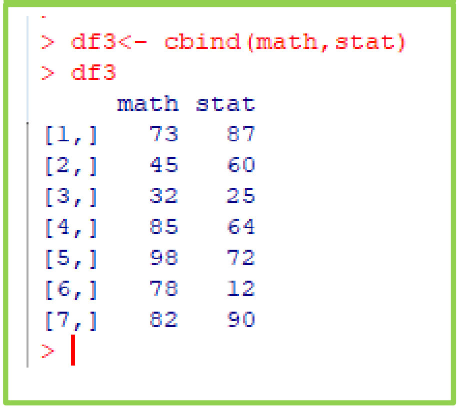


```
> math<- c(73,45,32,85,98,78,82)
> stat<- c(87,60,25,64,72,12,90)
> grades<-matrix(c(math,stat),nc=2)
> grades
     [,1] [,2]
[1,]   73   87
[2,]   45   60
[3,]   32   25
[4,]   85   64
[5,]   98   72
[6,]   78   12
[7,]   82   90
```

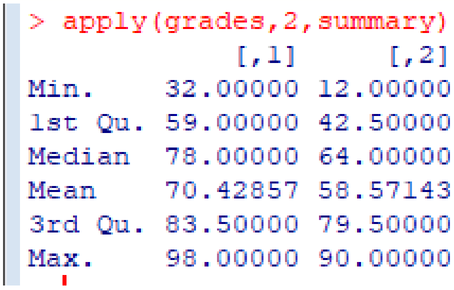
OR



```
> 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
```

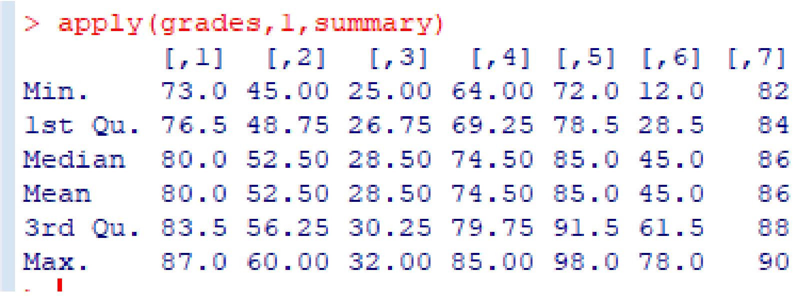


```
> df3<- cbind(math,stat)
> df3
     math stat
[1,]   73   87
[2,]   45   60
[3,]   32   25
[4,]   85   64
[5,]   98   72
[6,]   78   12
[7,]   82   90
>
```



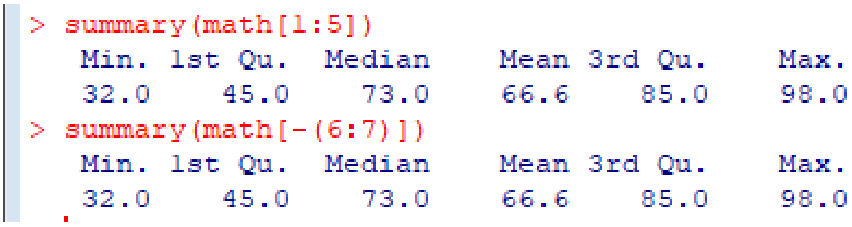
```
> apply(grades,2,summary)
            [,1]     [,2]
Min.    32.00000 12.00000
1st Qu. 59.00000 42.50000
Median  78.00000 64.00000
Mean    70.42857 58.57143
3rd Qu. 83.50000 79.50000
Max.    98.00000 90.00000
```

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



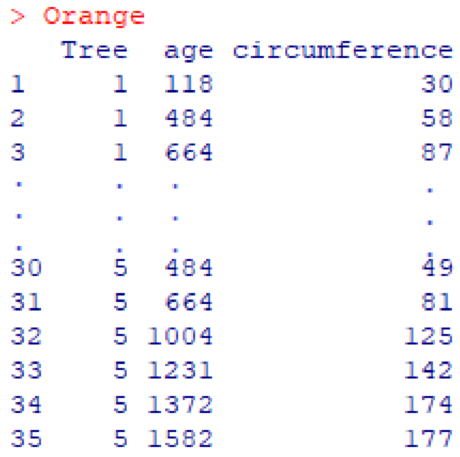
```
> summary(math[1:5])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   32.0    45.0    73.0    66.6    85.0    98.0 
> summary(math[-(6:7)])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   32.0    45.0    73.0    66.6    85.0    98.0 
```

استدعاء بيانات من R و حساب بعض الاحصاءات

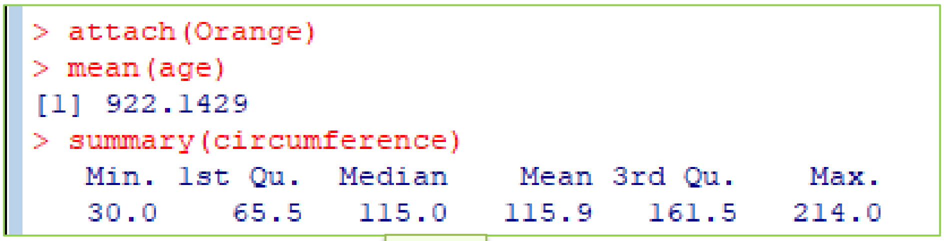
Q2: Growth of Orange Trees

Description

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

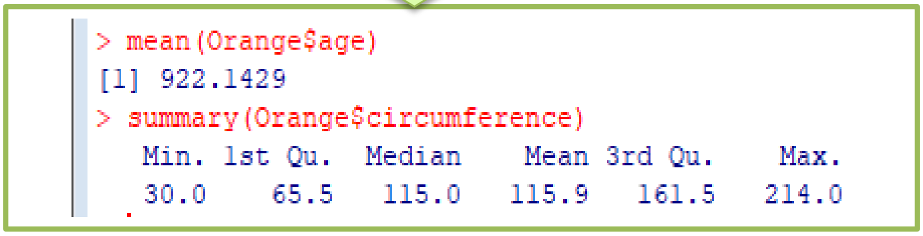


```
> Orange
   Tree  age circumference
1     1  118            30
2     1  484            58
3     1  664            87
.     .   .              .
.     .   .              .
.     .   .              .
30    5  484            49
31    5  664            81
32    5 1004           125
33    5 1231     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

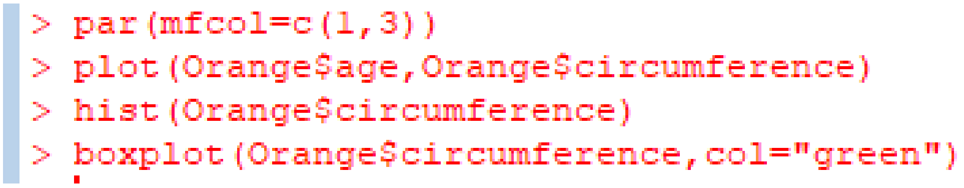


```
> attach(Orange)
> mean(age)
[1] 922.1429
> summary(circumference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   30.0    65.5   115.0   115.9   161.5   214.0 
```

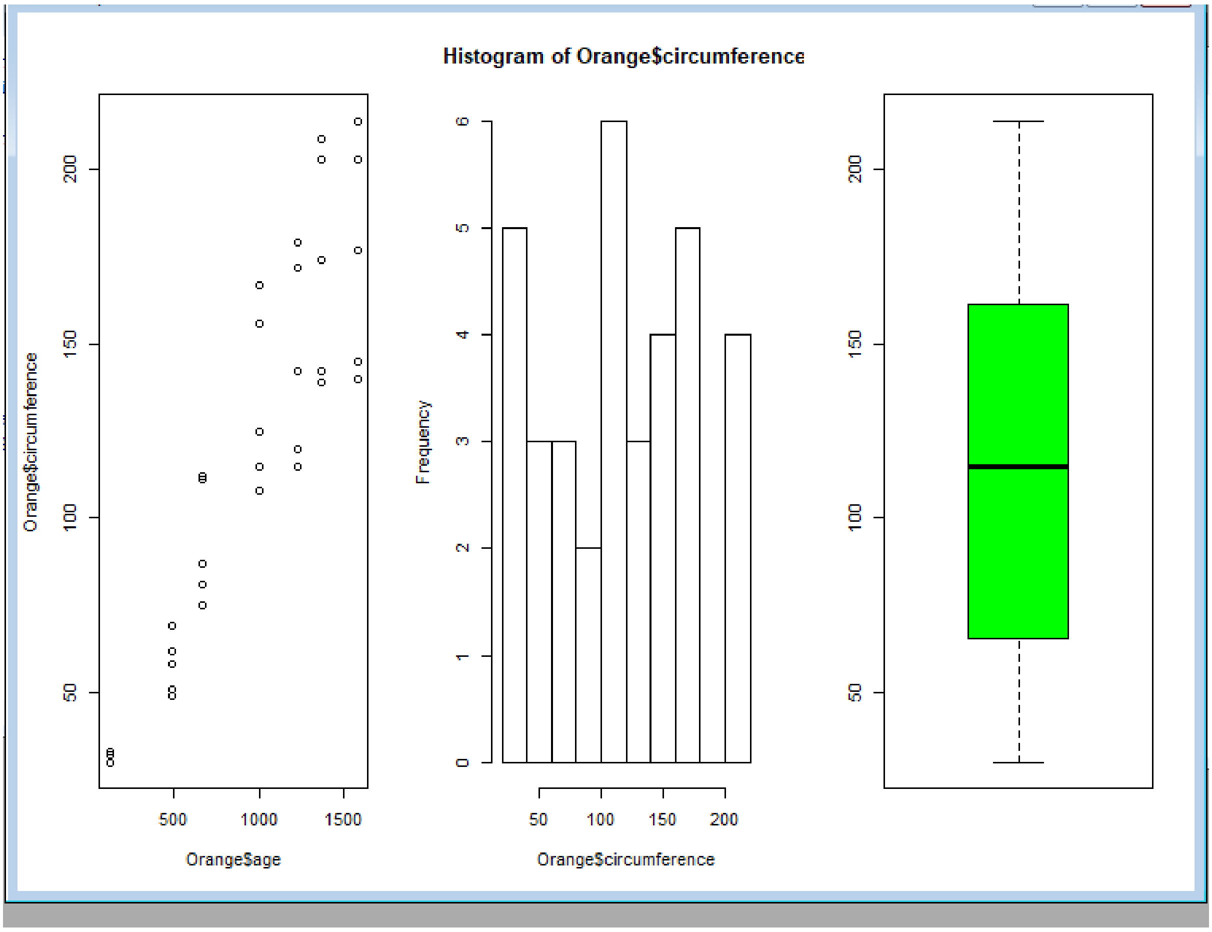
OR

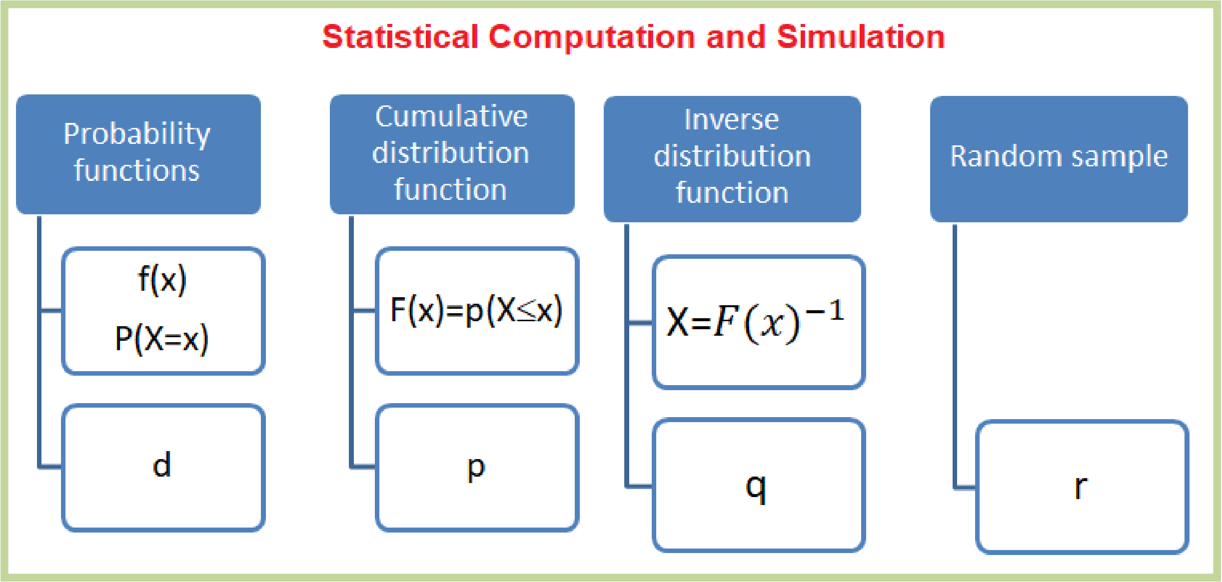


```
> mean(Orange$age)
[1] 922.1429
> summary(Orange$circumference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   30.0    65.5   115.0   115.9   161.5   214.0 
```



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





Statistical Computation and Simulation

Probability functions	Cumulative distribution function	Inverse distribution function	Random sample
f(x) P(X=x)	F(x)=p(X≤x)	X=$F(x)^{-1}$	
d	p	q	r

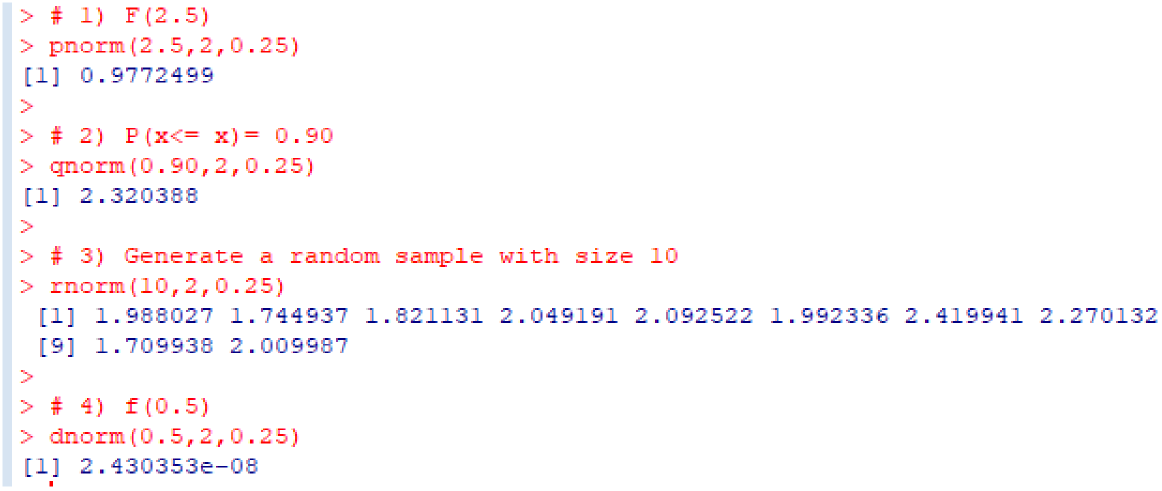
Q3: Suppose X is Normal with mean 2 and standard deviation 0.25 . Find:

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 ?**

4- f(0.5)

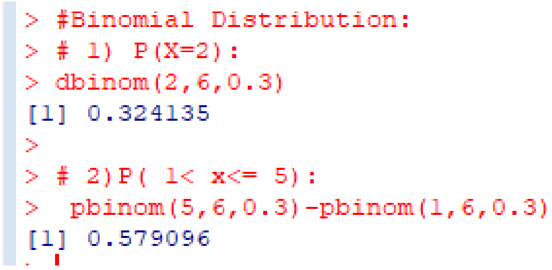


```
> # 1) F(2.5)
> pnorm(2.5,2,0.25)
[1] 0.9772499
>
> # 2) P(x<= x)= 0.90
> qnorm(0.90,2,0.25)
[1] 2.320388
>
> # 3) Generate a random sample with size 10
> rnorm(10,2,0.25)
 [1] 1.988027 1.744937 1.821131 2.049191 2.092522 1.992336 2.419941 2.270132
 [9] 1.709938 2.009987
>
> # 4) f(0.5)
> dnorm(0.5,2,0.25)
[1] 2.430353e-08
```

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)

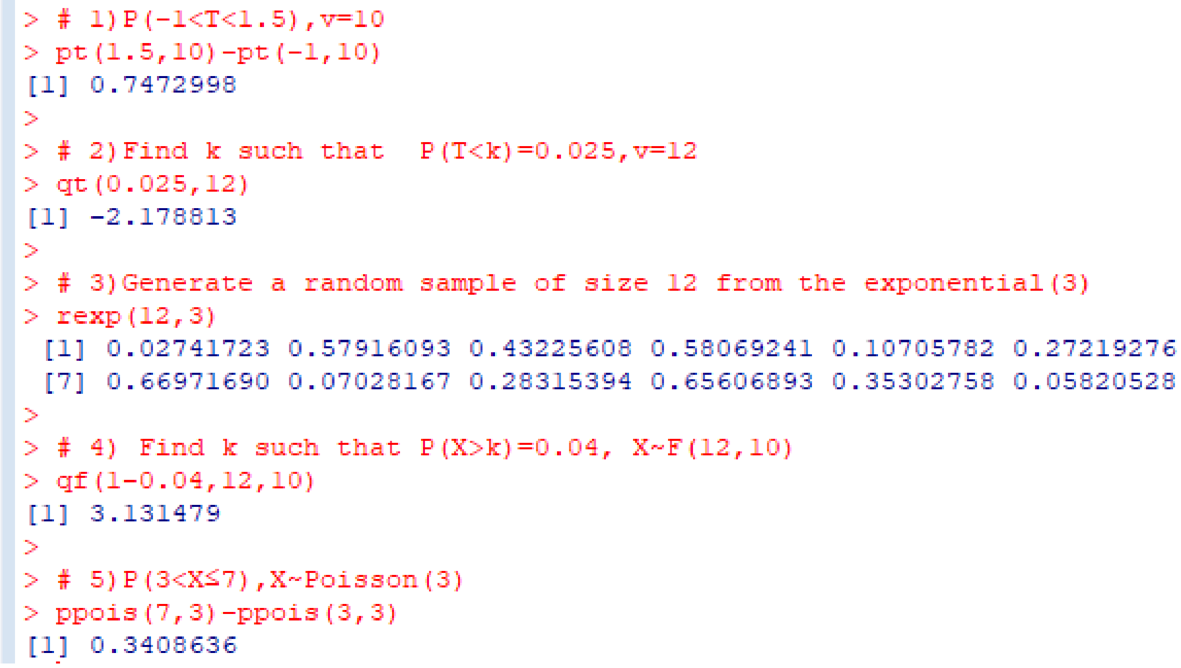
2- $P(1< X \leq 5)$ = $P(X \leq 5) - P(X \leq 1)$



```
> #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

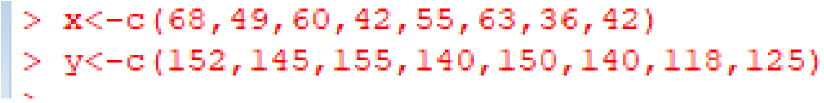
1. $P(-1 < T < 1.5),\ v = 10$
2. Find k such that $P(T < k) = 0.025,\ v = 12$
3. Generate a random sample of size 12 from the exponential(3)
4. Find k such that $P(X > k) = 0.04,\quad X \sim F(12, 10)$
5. $P(3 < X \leq 7),\ X \sim \text{Poisson}(3)$



```
> # 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
```

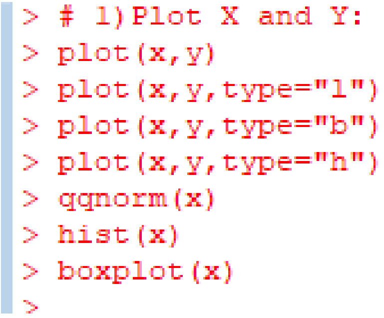
Q6: We have the following table show age X and blood pressure Y of 8 women

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



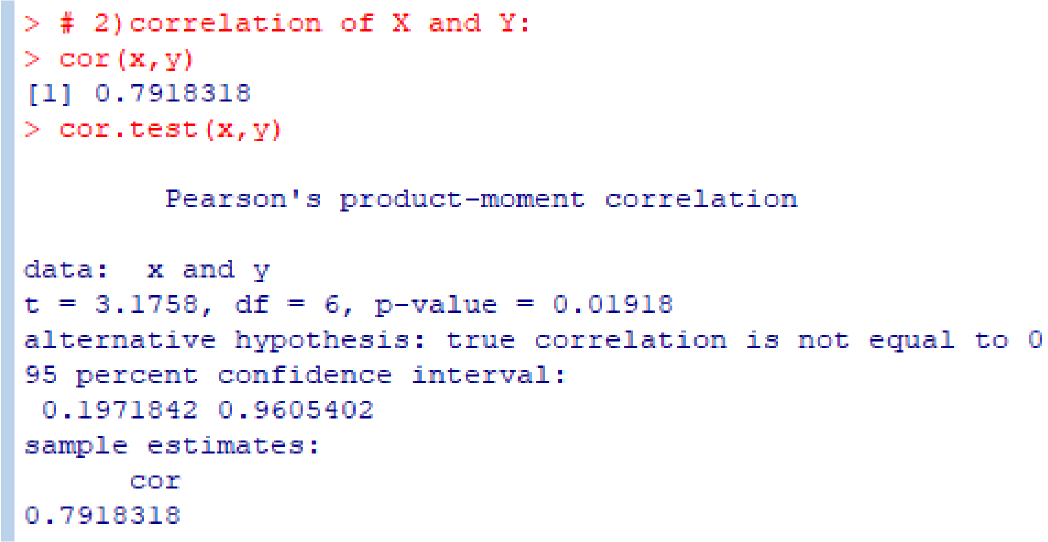
```
> x<-c(68,49,60,42,55,63,36,42)
> y<-c(152,145,155,140,150,140,118,125)
```

1. **Plot X and Y**



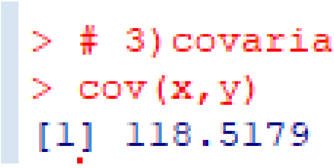
```
> # 1)Plot X and Y:
> plot(x,y)
> plot(x,y,type="l")
> plot(x,y,type="b")
> plot(x,y,type="h")
> qqnorm(x)
> hist(x)
> boxplot(x)
>
```

2. **correlation of X and Y**



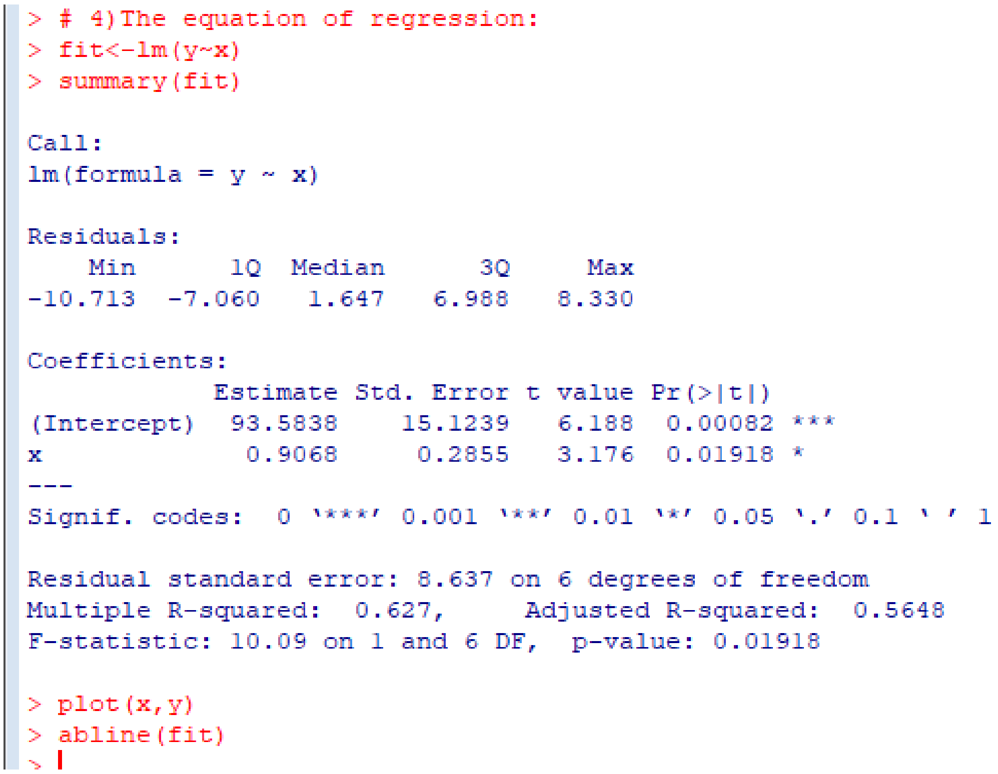
```
> # 2)correlation of X and Y:
> cor(x,y)
[1] 0.7918318
> cor.test(x,y)

        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**

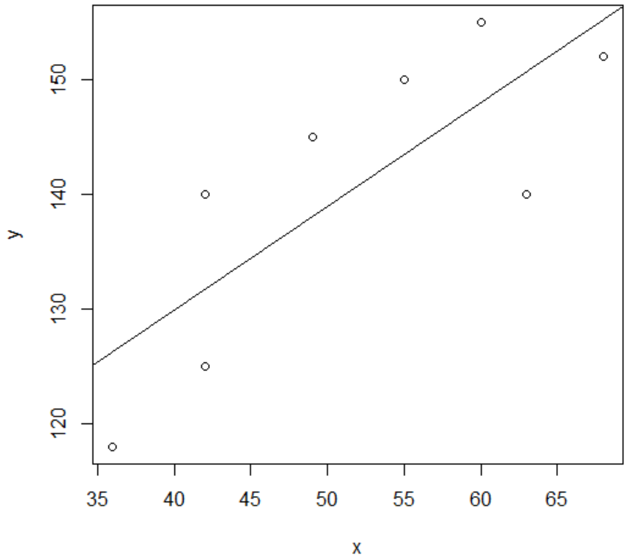


```
> # 3)covariance:
> cov(x,y)
[1] 118.5179
```

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 :

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

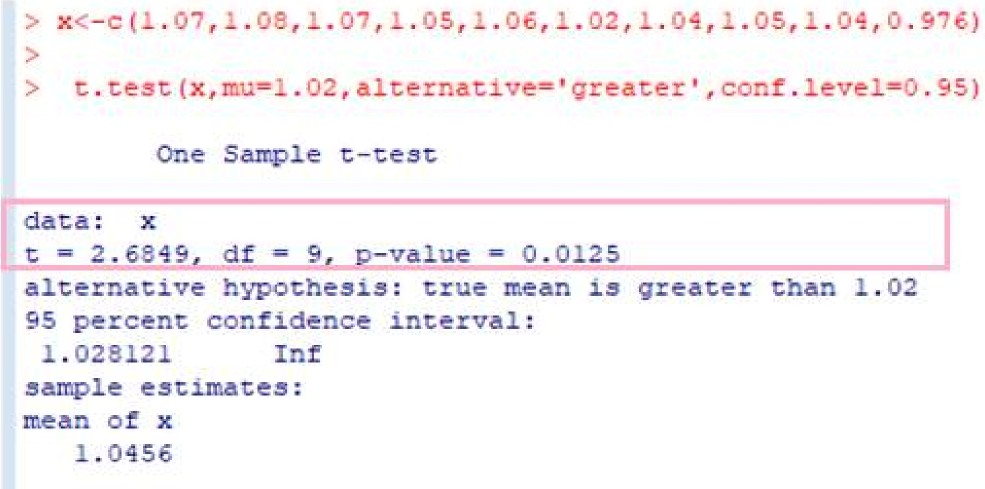
2-Test statistics :

T=2.6849

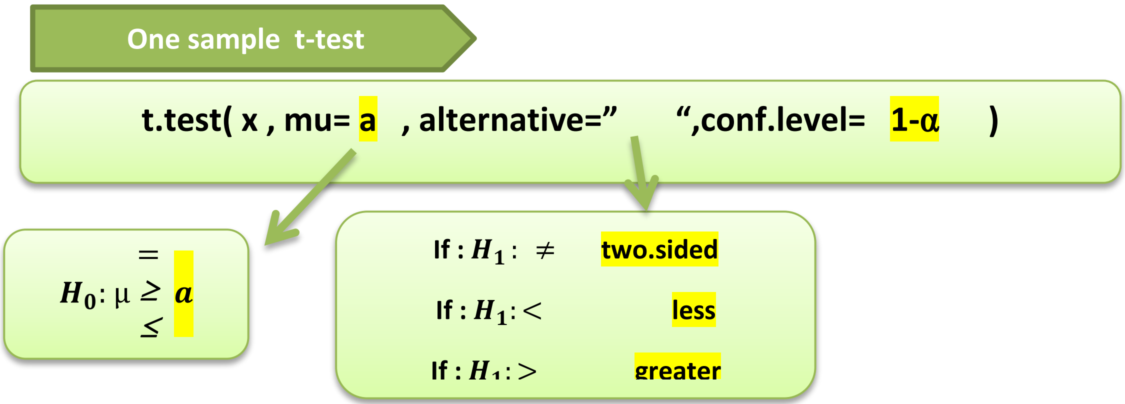
3- Decision:

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

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)

$H_0: \mu \begin{matrix} = \\ \geq \\ \leq \end{matrix} a$

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 :**

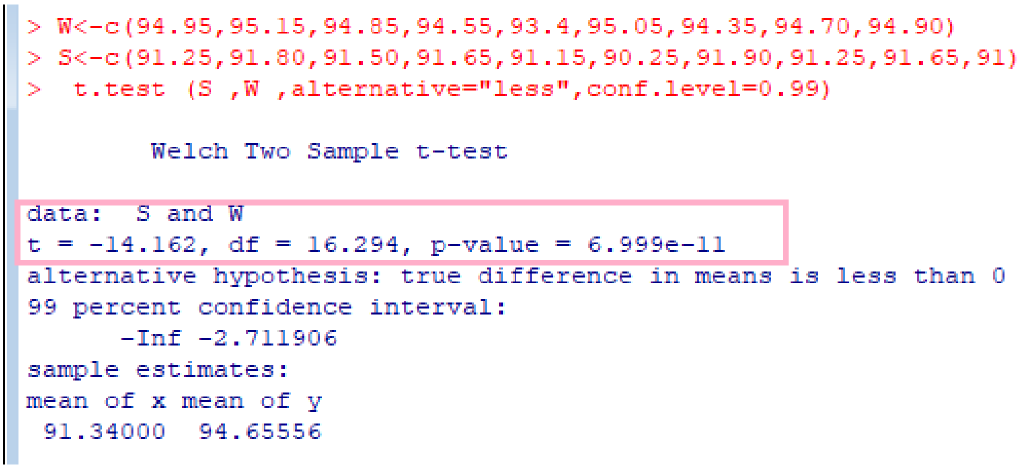
$$H_0: \mu_{Skim} \geq \mu_{Whole} \quad vs \quad H_1: \mu_{Skim} < \mu_{Whole}$$

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

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

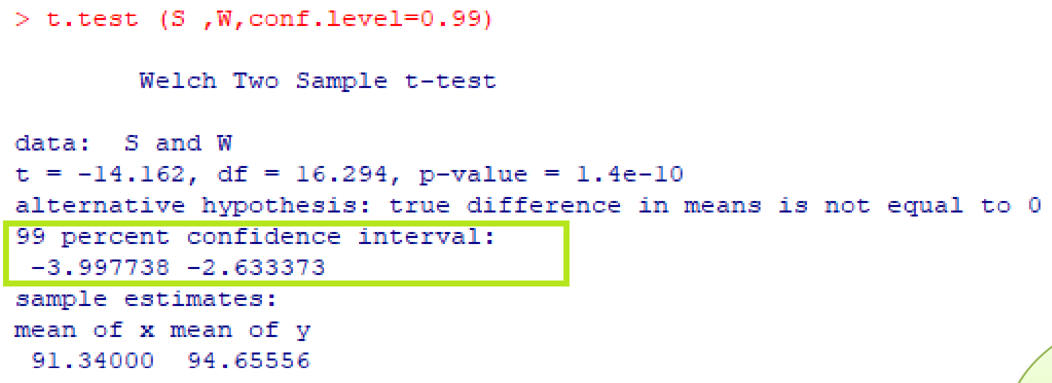
3- **Decision:**

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

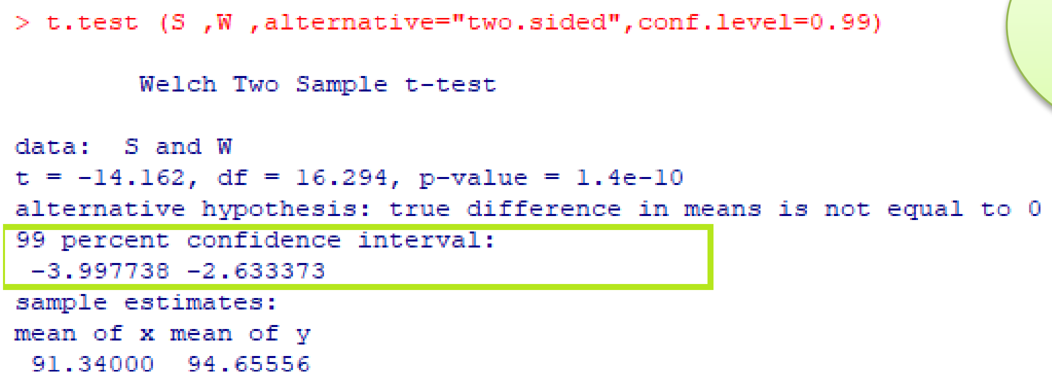


```
> W<-c(94.95,95.15,94.85,94.55,93.4,95.05,94.35,94.70,94.90)
> S<-c(91.25,91.80,91.50,91.65,91.15,90.25,91.90,91.25,91.65,91)
>  t.test (S ,W ,alternative="less",conf.level=0.99)

        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)

$H_0: \mu_x - \mu_y \overset{=}{\underset{\le}{\ge}} a$

If : $H_1: \neq$ **two.sided**

If : $H_1: <$ **less**

If : $H_1: >$ **greater**

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 $\mu_D \neq 0$

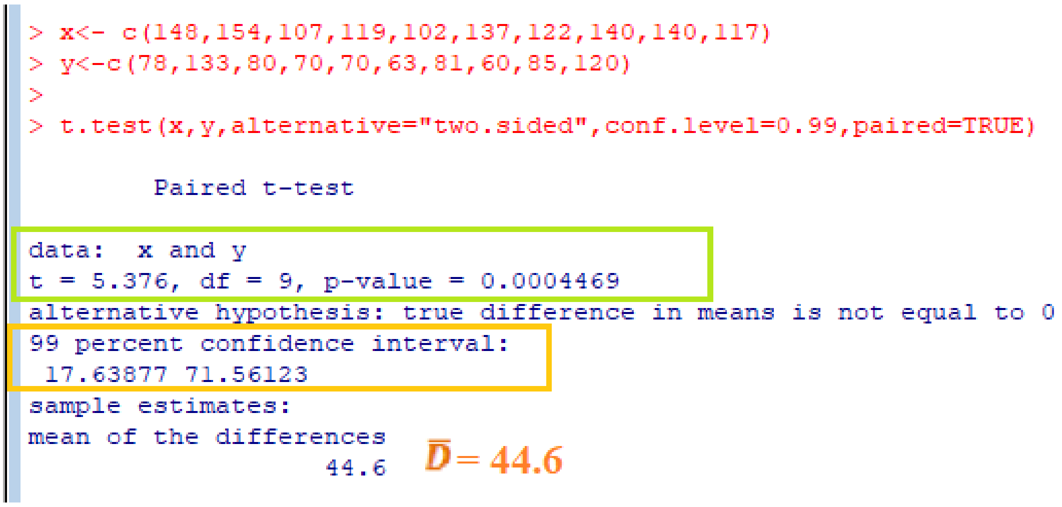
2- Test Statistic :

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 )`

$H_0: \mu_D \begin{matrix} = \\ \geq \\ \leq \end{matrix} a$

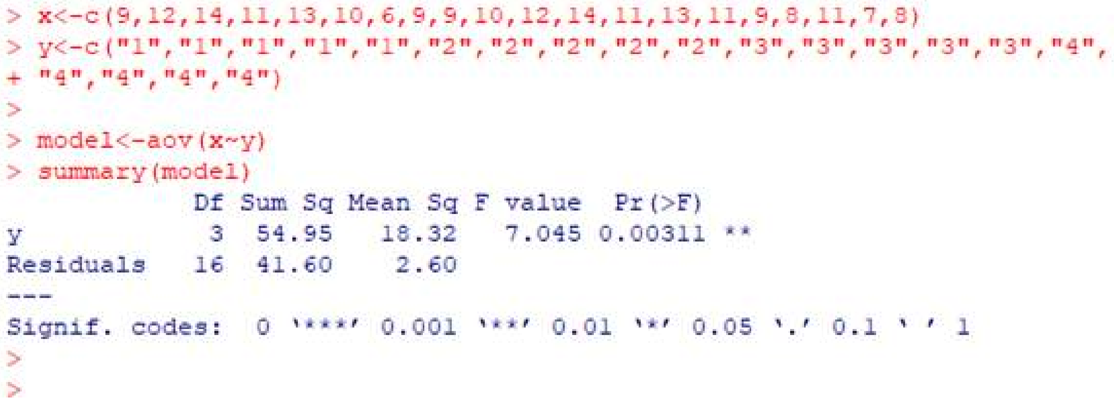
If : $H_1: \neq$ two.sided

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 :

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

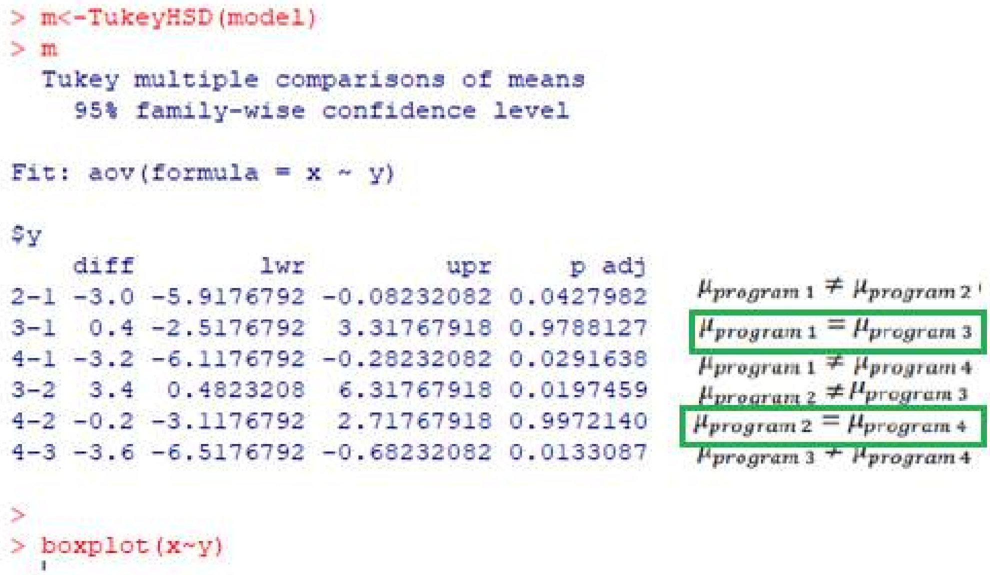
$$H_1: at\ least\ one\ mean\ is\ diffrenet$$

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)
```

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

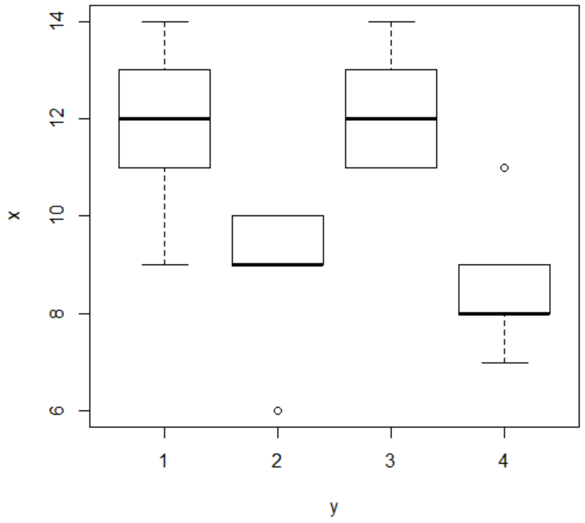
$\boxed{\mu_{program\,1} = \mu_{program\,3}}$

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

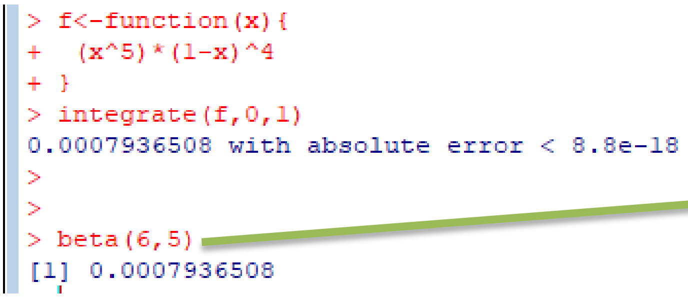
$\mu_{program\,2} \neq \mu_{program\,3}$

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

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



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

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

α-1 = 5 and β-1 = 4

α = 6 | β = 5

$\Rightarrow$ B(6 , 5)