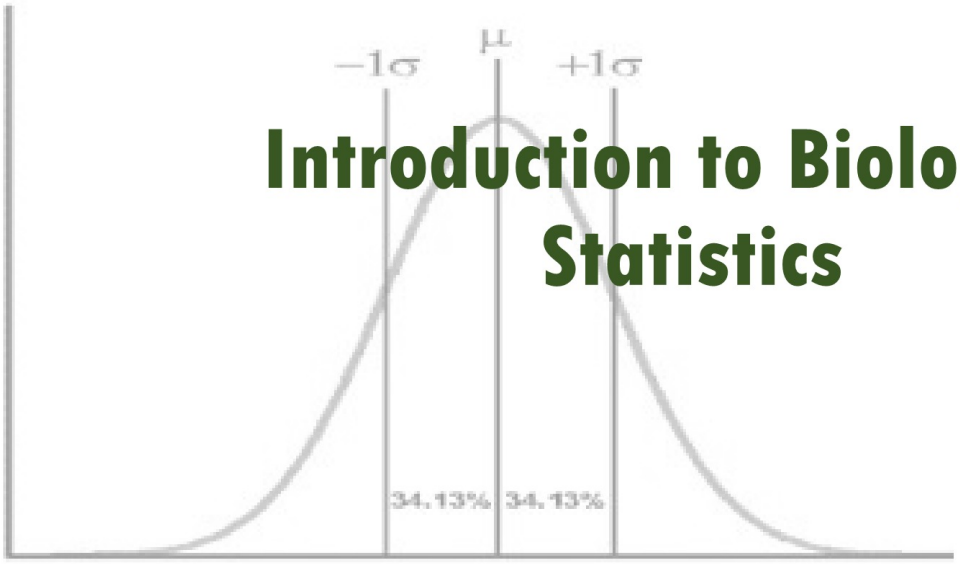


Introduction to Biological Statistics

Lecture Approaches

- Definitions.
- Normal distribution.
- Test of the hypothesis.
- Dogma of biological statistics.
- Choosing the best statistical test.

Definitions

Statistics is:

A branch of mathematics that provides techniques to analyze whether or not your data is significant (meaningful).

A branch of mathematics dealing with the analysis and interpretation of masses of numerical data.

The field of study that involves the collection and analysis of numerical facts or data of any kind.

The study of how information should be employed to reflect on, and give guidance for action, in a practical situation involving uncertainty.

Biostatistics: Application of statistical methods to biological, medicine and health sciences including (Collection and presentation of data, analysis and interpretation of the results and making results on the basis of such analysis).

Definitions

Population: **is any complete group with at least one characteristic in common.**

A population may be studied using one of two approaches: taking a census, or selecting a sample.

Census: **is a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.**

Sample: **is a selective group of the population** (is a subset of units in a population, selected to represent all units in a population of interest).

❑**Statistics use samples as estimators of the corresponding population.**

Definitions

Data: are measurements or observations that are collected as a source of information.

data unit: is one entity (such as a person or business) in the **population** being studied, about which data are collected.

data item: is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income.

Observation: is an occurrence of a specific data item that is recorded about a data unit.

Dataset: is a complete collection of all observations.

Definitions

	age (years)	sex	income ($)
Person 1 (John Smith)	18	m	50000
Person 2 (Joe Bloggs)	18	m	40000
Person 3 (Sally Jones)	20	f	55000
Person 4 (Linda Lee)	22	f	50000
Person 5 (Harry James)	19	m	35000

Data Items

Data Unit - Person 2.

Numeric observation of the data item 'income'

Non-numeric (categorical) observation of the data item 'sex'

Quantitative data: are measures of values or counts and are expressed as numbers. Quantitative data are data about **numeric variables**.

Qualitative data: are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about **categorical variables**.

Definitions

Variable: is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. The value of the variable can "vary" from one entity to another.

Types of variables:

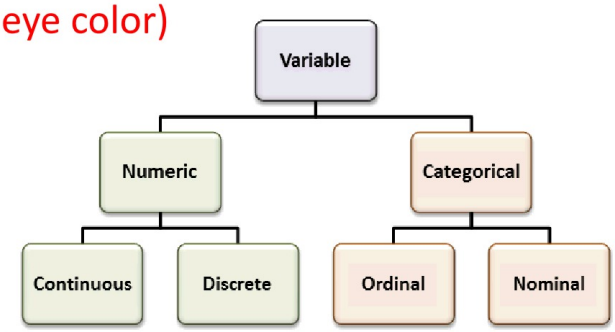
Numeric variables have values that describe a measurable quantity as a number.

- **Continuous variable** is a numeric variable. Observations can take any value between a certain set of real numbers. (*i.e.* 1.5m, 48.2˚C)
- **Discrete variable** is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. (*i.e.* 1, 2, 3 cars)

Definitions

Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit.

- **Ordinal variable** is a categorical variable. Observations can take a value that can be logically ordered or ranked. (*i.e.* small, medium, A, B)
- **Nominal variable** is a categorical variable. Observations can take a value that is not able to be organized in a logical sequence. (*i.e.* sex, business type, eye color)



Definitions

Statistical hypothesis: is any claim about population and this hypothesis could be true or false.

Null hypothesis H_0: Nothing new or interesting happening here! (And anything "interesting" observed is due to chance alone.).

Alternative hypothesis H_a: is the rejection of the null hypothesis (usually be considered the researcher's hypothesis).

A test of hypotheses: is a method for using sample data to decide whether the null hypothesis should be rejected.

Errors in Hypothesis Testing:

A type I error consists of rejecting the null hypothesis H_0 when it was true.

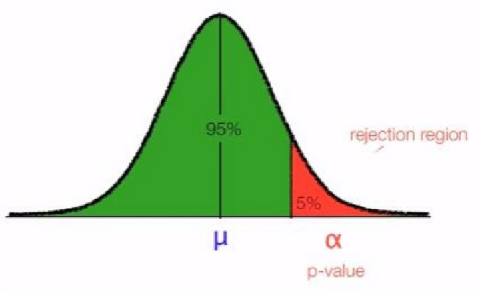
A type II error consists of not rejecting H_0 when H_0 is false.

Definitions

Level α Test: is significant level of experiment (mostly 0.05 in biological experiments).

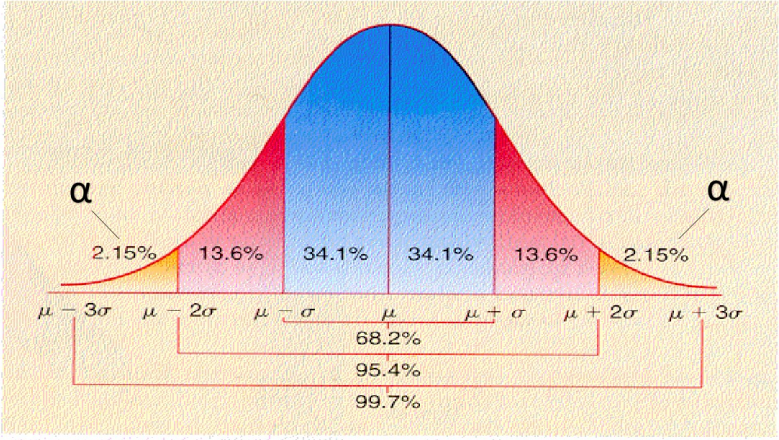
P-value: The P-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given data set.

P-value: is the probability that, if null hypothesis was true.

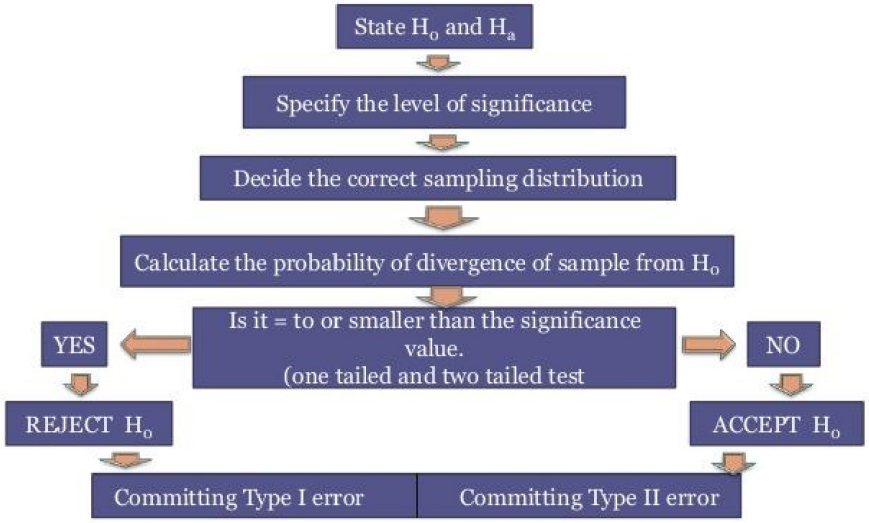


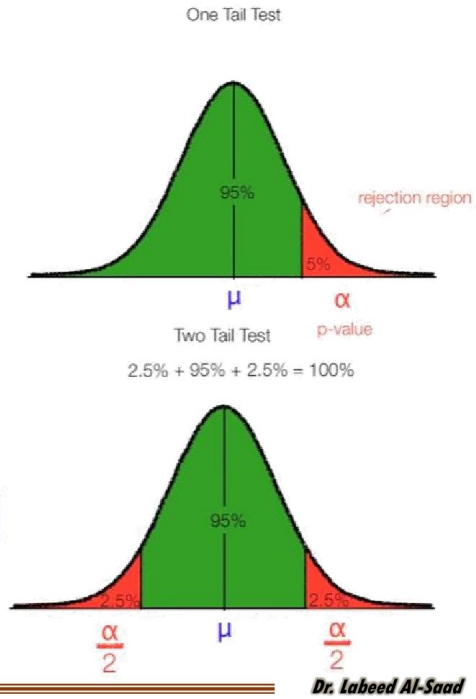
Normal distribution

Normal distribution: A theoretical frequency distribution for a random variable, characterized by a bell-shaped curve symmetrical about its mean. Also called Gaussian distribution



Test of the hypothesis





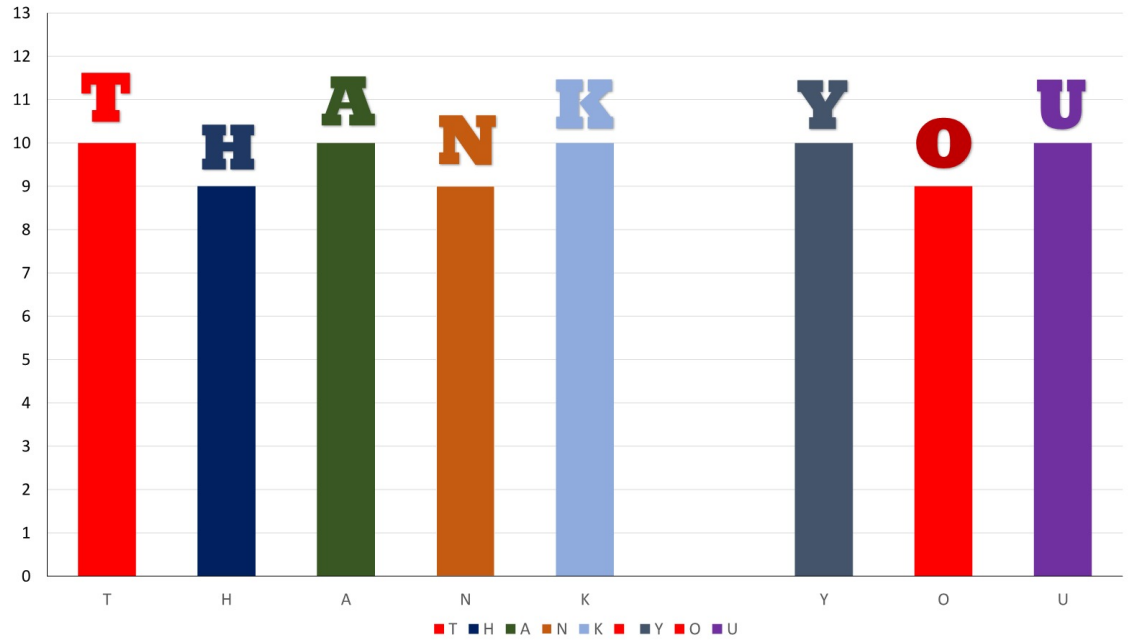
Dogma of biological statistics

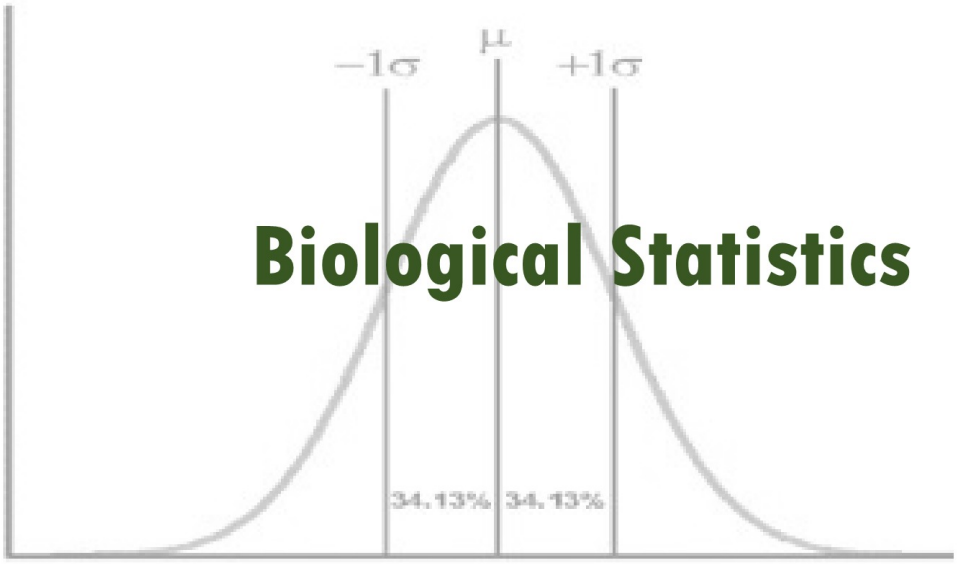
1. **Hypothesis.** (H_0 and H_a)
2. **Significance.** Level of significance α (0.05)
3. **Sample.** Take a sample from population to provide the statistics you need.
4. **p-Value.** Calculate the P-value (This is always done by computer package like SPSS).
5. **Decide.** Use calculated P-Value to decide which hypothesis is true. If the P-value is less than significant level you have to reject null hypothesis

Choosing the best statistical test

- Q1: What type of data do you have?
- Q2: How many samples do you have?
- Q3: What is the test supposed to do?

Q3 → / Q2 ↓ / Q1 →	Compare the Data		Seek Relationships
	Categorical Data	Quantitative Data	
One Sample	1 sample proportion	1 sample t	
Two Samples	2 sample proportions	2 sample t	
Two Samples Special		2 sample t Paired t	Correlation/ Regression
Three or more samples		One-Way ANOVA	


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Tabular and graphical presentation of data ?

A statistical method involves arranging, summarizing, and presenting a set of data in such a way that useful information is produced.

Example: If we need to know the number of males or females in a sample or grouping them according to specific age classes. In this situation we have to build a table divided to specific age classes or into two groups (males and females), so that we'll get an idea (information) about the frequency of each group in the sample

These tables could be simple or composite depending on the researcher or the study requirements. Below an example of this type of tables.

Tabular and graphical presentation of data ?

1. Simple table: This table is based on the representation of one characteristic that is divided into classes whose width is determined by the researcher, and the frequency of each class was counted, and the class mid-point was calculated.

Example: Distribution of students' scores for a biostatistics exam in a particular class containing 29 students.

Students score classes	frequency	Class mid-point
1 - 25	1	13
26 -50	3	38
51- 75	20	63
76 - 100	5	88
sum	29	

Note: that the 29 different grades of students were presented in a frequency table and they became understandable, and we had a clear vision about the levels of students in this class.

The class mid-point represent the average of the class

Based on the above, a new statistical terms were appeared:

- Class: a class is a grouping of values by which data is binned for computation of a frequency distribution. In our past example we grouped score variable in to (0-100) in to four classes.
- Class limits: the upper and lower limits of the class, for example the class (1-25), the lower limit is 1, while the upper limit is 25.
- Class interval: Class interval refers to the numerical width of any class in a particular distribution. Mathematically it is defined as the difference between the upper and the lower class limits.

Based on the above, a new statistical terms were appeared:

- Class Mid-point: is a specific point in the center of the class *i,e.* it is the average of the upper and lower class limits.
- Class frequency: Class frequency refers to the number of observations in each class. In our previous example we had 20 student in class (50-75), which means the frequency is 20.

Note: the number of classes of any sample can be determined based on researcher experience or by one of the statistical methods like Sturges method:

$$No.of\ classes = 1 + (3.3 \times \log n)$$

Based on the above formula the No. of classes of our past example should be 6.

Tabular and graphical presentation of data

1. Composite table: this type of tables based on presenting two or more categories by grouping them in to classes.

Example: grouping students based on length and weight in composite table

Note: we grouped a 220 values in 2 directional classes (length& weight) producing a comprehensible table

Student length / Student weight	50-59 kg	60-69 kg	70-79 kg	Sum
150-160	20	15	5	40
161-170	15	12	8	35
171-180	6	15	14	35
Sum	41	42	27	110

Tabular and graphical presentation of data

What if the data were categorical ?

Example: grouping of students based on stage and gender in composite table

Stage / Gender	Male	Female	Sum
First stage	18	7	25
Second stage	10	8	18
Third stage	21	5	26
Fourth stage	19	6	25
Sum	68	26	94

How to create frequency table

1. Calculate the diference between the maximum and minimum values of the data sample (arrange the data in ascending or descending order to facilitate this).
2. Determining the number of categories either through the researcher's experience or by one of the statistical methods.
3. Calculate the class interval by dividing the product of (step 1) by product of (step 2) and rounding the result to the nearest integer value.
4. Determine the lower and upper limits (the first lower limit being the minimum value in the sample or a little smaller, then the lower limit for the second class is greater than it by the class width and so on, after that we have to determine the upper limits for all classes starting from the first class whose upper limit is less than a the lower limit of the next class by a degree, the next upper limit should increased by the class width, and so on).

How to create frequency table

Example: Create a frequency table for RBC sample of 32 individuals collected randomly?

RBC	3.2	2.4	2.5	4.1	2.5	2.1	3.1	2.6	3.7	2.6	2.4
	3.6	3.7	4.11	4.2	3.2	2.9	2.92	2.66	1.8	1.95	1.8
	3.2	3.3	3.1	2.5	2.44	3.6	3.11	2.88	4.1	5.2	

Solution:

1. Determine the maximum and minimum values and calculate the divergence between them: 5.2 - 1.8 = 3.4
2. Define the number of classes according to the statistical formula: $No.of\ classes = 1 + (3.3 \times \log n)$ which is approximately 6.
3. Determine the class width by dividing the first product by the second product: 3.4 / 6 = 0.56, approximately 0.6

So the table should be as follows:

How to create frequency table

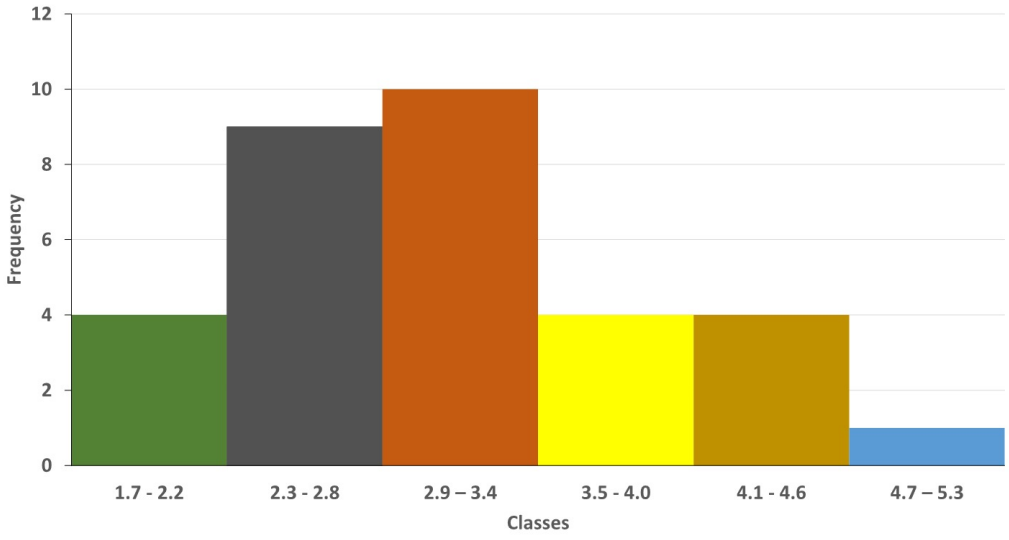
Classed	Frequency	Class mid-point
1.7-2.2	4	1.95
2.3-2.8	9	2.55
2.9-3.4	10	3.15
3.5-4.0	4	3.75
4.1-4.6	4	4.35
4.7-5.3	1	5
Sum	32	

Create the frequency histogram, polygon, and curve

To draw a histogram, based on the frequency table that previously created:

1. Draw the x and y axes.
2. Put the classes on the x-axis.
3. Put the frequencies on the y-axis.
4. Represent each class by drawing a rectangle from the beginning of the lower limit of the class up to the value of the frequency then transversely to the lower limit of the next class, we may leave a space between the rectangles, especially in the case of catagorical classes.

The Histogram

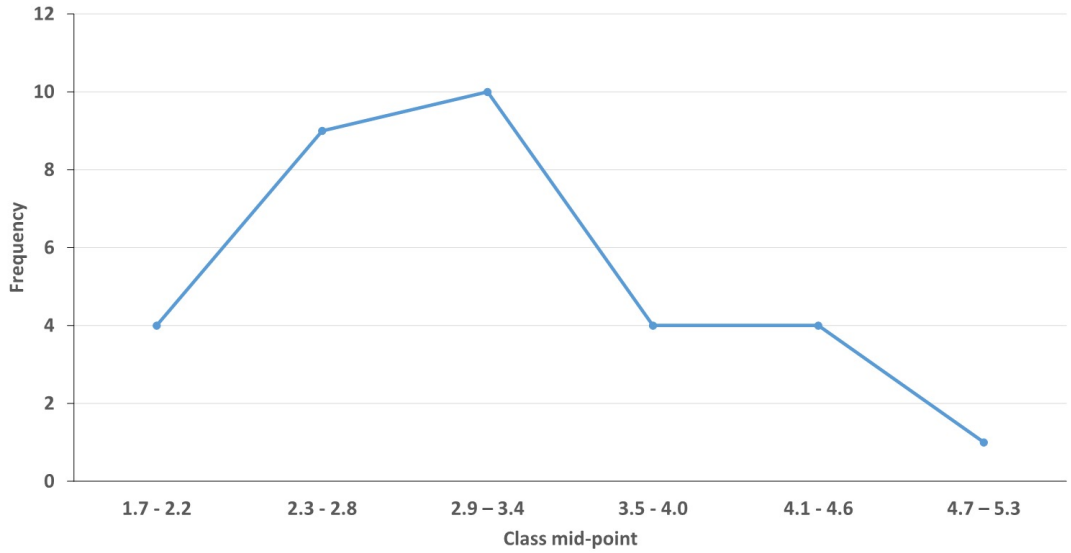


Create the frequency histogram polygon and curve

To draw a frequency polygon, based on the frequency table that previously created:

1. Draw the x and y axes.
2. Put the classes on the x-axis.
3. Put the frequencies on the y-axis.
4. Represent each class Mid-point by plotting a dot indicating the frequency of that class.
5. Connect the dots by a straight lines.

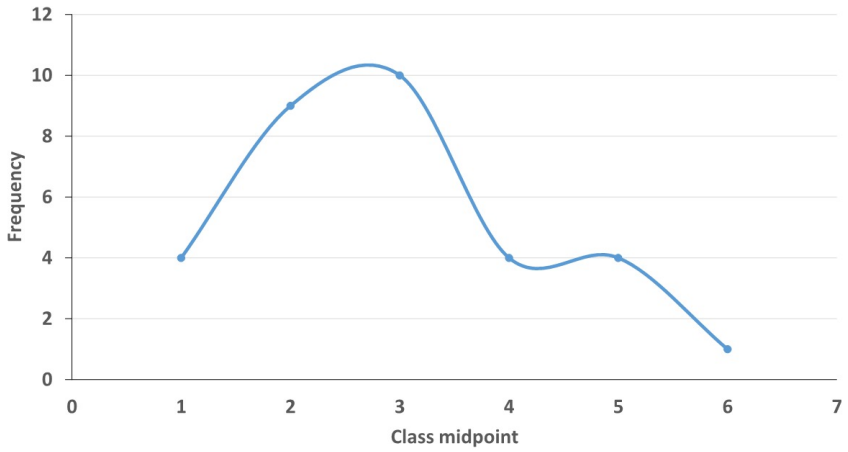
Frequency polygon

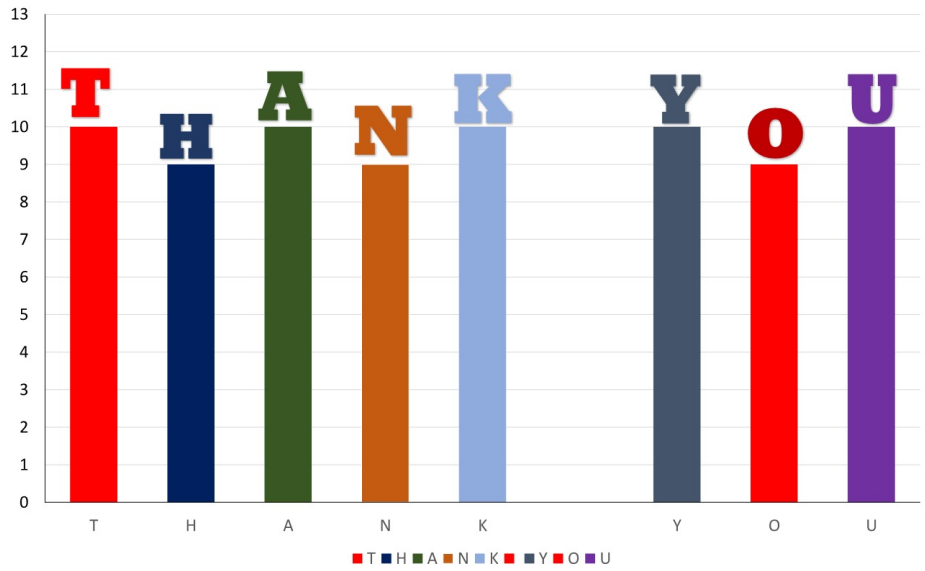


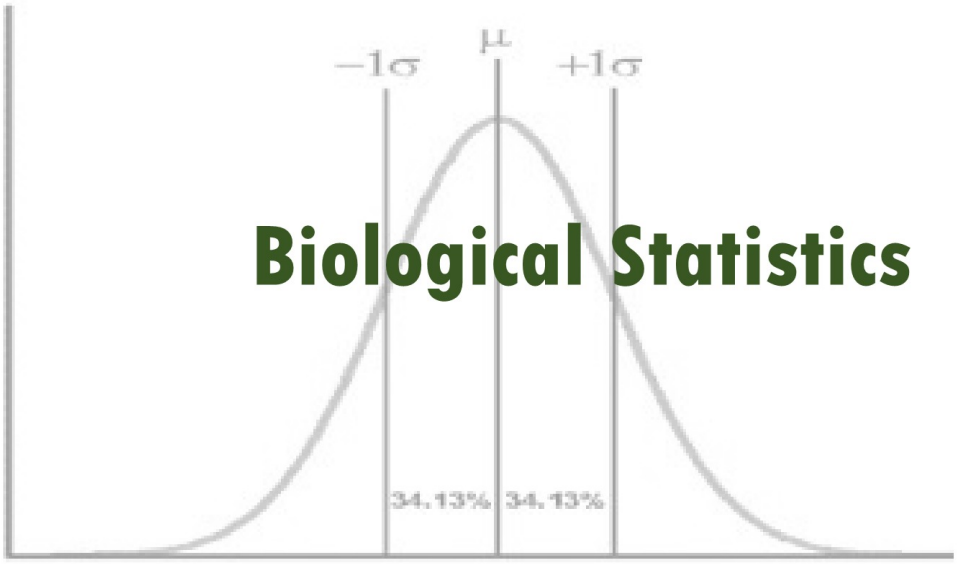
The frequency curve

To draw the Frequency curve

Follow the same polygon protocol, but connect the dots with curved line




13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

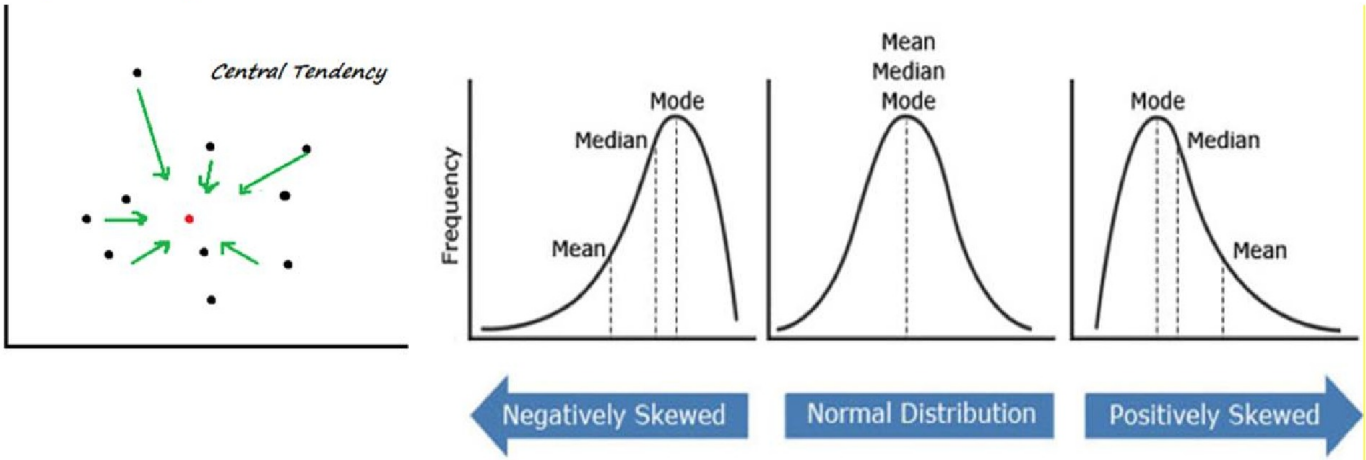
Important statistical sympols

$\sum$	Sum of or summation	n!	N Factorial
$\neq$	Not Equal	α	0.05 statistical level
$>$	Greater than	β	0.01 statistical level
$<$	Less than	$\lvert \quad \rvert$	Absolute value
μ	Population mean	$\bar{x}$	Sample mean
σ^2 σ	Variance and standard deviation of the population	S^2 S	Variance and standard deviation of the sample
$\hat{Y}$	Y hat: indicate an estimate rather than true	X or Y or Z	A variable
Xi	Individual observation	$\sum Xi$	Total sum of servation
n	Sample size	$(\sum xi)^2$	Square total sum
$\sum Xi^2$	Total sum of square	C.V.	Coefficient of ariation
d	Difference (x1 –x2)	*b*	Sample regression coefficient
r	Sample correlation	df	Degrees of freedom
SS	Sum of square	MS	Mean square
ns	Not significant	(*) and (**)	Significant and highly significant
H_o	Null hypothesis	H1	Alternative hypothesis

These symbols will be used in statistical expressions, so it is important to understand their meanings

Central tendency measurements

These are the measures around which the values are centered, the most important of which are the arithmetic mean (the average), the median (the value that lies in the middle), and the mode (the most frequent value).



The Arithmetic Mean

It is the average of the sample observations = summation / number of observations

$$\bar{X} = \frac{\sum xi}{n} = \frac{x1 + x2 + x3 + \cdots + xn}{n}$$

Were: ($\bar{X}$ = the mean), ($\sum xi$ = **Sum**) **and** (n = **The number**)

Example: Find the mean of the following values:

1.4 1.23 1.33 1.36 1.55 1.66 1.85

Solution: The number of values (n)= 7

$$\bar{X} = \frac{\sum xi}{n} = \frac{1.85 + 1.66 + 1.55 + \cdots + 1.4}{7} = 1.4966$$

The properties of the arithmetic mean

1. The sum of the deviations of the values from their arithmetic mean is zero.

$\bar{X} = 7$

Xi	12	3	5	8	7
$\bar{X}$-xi	7-12 = -5	7-3 = 4	7-5 = 2	7-8 = -1	7-7 = 0

$\sum(\bar{X} - xi)$ = -5 + 4 + 2 -1 +0 = 0

2. The sum of the deviation squares is the least possible (that is, if we replace the mean with any sample value and measure the squares of the deviations from it, it would be greater than the value of the sum of the squares of the deviations).

Try replacing the mean value with any of the five sample values and you will notice that the sum of the deviation squares is greater than 46

Xi	12	3	5	8	7
$\bar{X}$-xi	7-12 = -5	7-3 = 4	7-5 = 2	7-8 = -1	7-7 = 0
$(\bar{X} - xi)^2$	25	16	4	1	0

$\bar{X} = 7$

$\sum(\bar{X} - xi)^2 = 25 + 16 + 4 + 1 + 0 = 46$

The properties of the arithmetic mean

3. If a fixed number is added to or subtracted from all values, the value of the mean changes by the amount of the fixed added or subtracted value.

4. If all values are multiplied by a constant number, then the new mean = the old mean × the constant number.

5. The mean of the sum of the values of two variables = the sum of the two means.

	xi	yi	xi+yi
	10	13.16	23.16
	12	4	16
	3	12	15
	5	7	12
Mean	7.5	9.04	16.54

The mean of sum = the sum of means

The Weighted Mean

It is calculated by dividing (the sum of the observations × their weights) by the sum of the weights.

$$\overline{WX} = \frac{\sum wixi}{\sum wi} = \frac{(w1 * x1) + (w2 * x2) + (w3 * x3) + \cdots + (wn * xn)}{w1 + w2 + w3 + \cdots + wn}$$

were is: $\overline{WX}$ = weighted mean ; $\sum wixi = sum\ of\ values \times their\ weights$; wi = value weight

What does weights means ????

The weight is an expression of the importance of the value. For example, if you have 10 Iraqi dinars and 10 dollars, although the numerical value is 10, the weighted value of the 10 dollars is 10 x 1300 = 13,000 dinars. The 1300 is the weight of each dollar, while the weight of each dinar is only 1. Another example, the degrees of the first stage in the college weigh 10% in the final average, while the degrees of the fourth stage weigh 40%, meaning that the value of the degree in the fourth stage = 0.4 of the degree of the final average, while the degree in the first stage = 0.1

The Weighted Mean

When we use Weighted mean ?????

1. When the data represented as a frequency of occurrence.
2. When some data factors are given a greater weight than the other factors

The Median

The median is the value that lies in the middle after arranging the sample values in ascending or descending order.

1. If the number of values is odd then the order of the median = number of values + 1 divided by 2.
2. If the number of values is even, then the value of the median = the number of values divided by 2, then we take this value and the one after it and extract their average, which represents the value of the median.

If the number of values id even

10 ،4 ،6 ،3 ،19 ،13

19 ،13 ،10 ،6 ،4 ،3

Median= (6+10)/2= 8

If the number of values is odd

9 ،10 ،5 ،7 ،4 ،2 ،3

10 ،9 ،7 ،5 ،4 ،3 ،2

Median= (7+1)/2= 4 *i.e.* The fourth value (5)

The Mode المنوال

mode

- It is the most frequently occurring value in the sample.
- Useful in the case of data containing outliers.
- Some data does not contain a mode.
- Not commonly used as medium and medium

How to calculate the mode of data set

- Sort the data in ascending order
- Find the frequently occurred values.

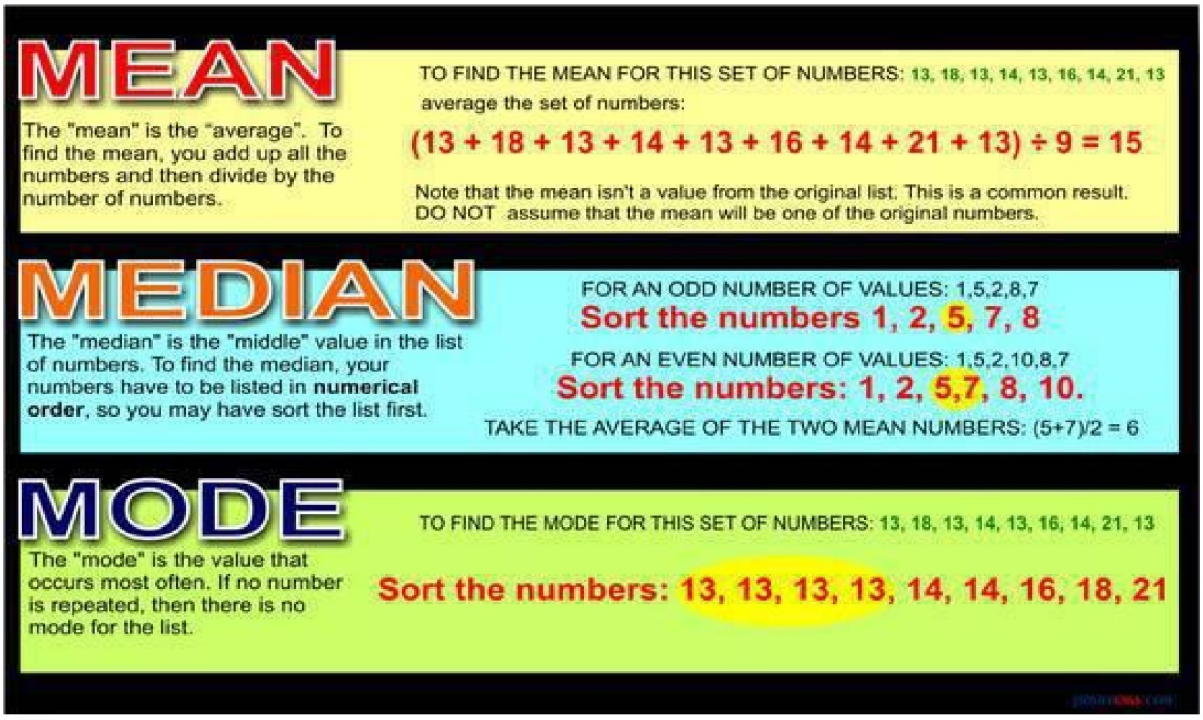
3, 4, 5, 5, 6, 6, 6, 7, 8, 8, 99 mode = 6

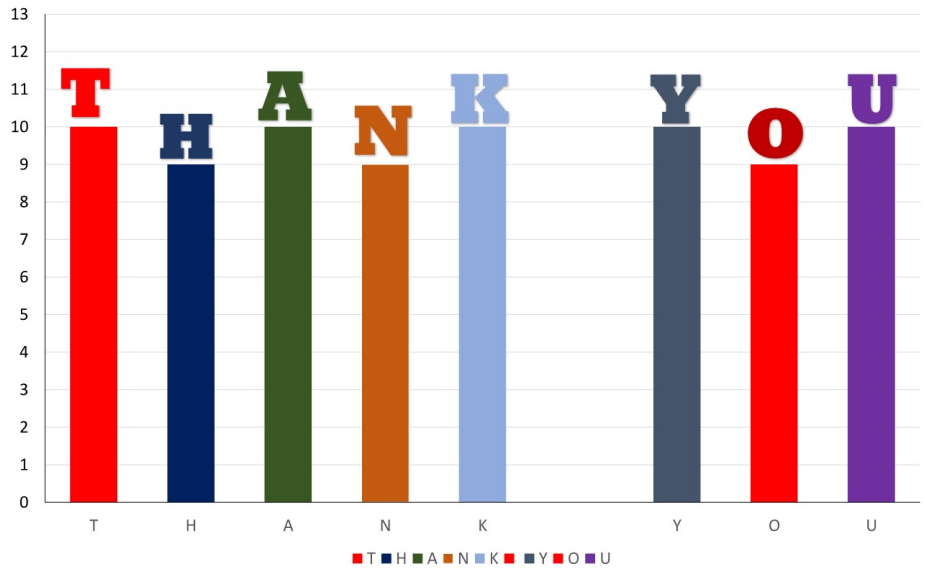
3, 4, 5, 5, 5, 6, 6, 6, 8, 8, 99 modes = 5 and 6

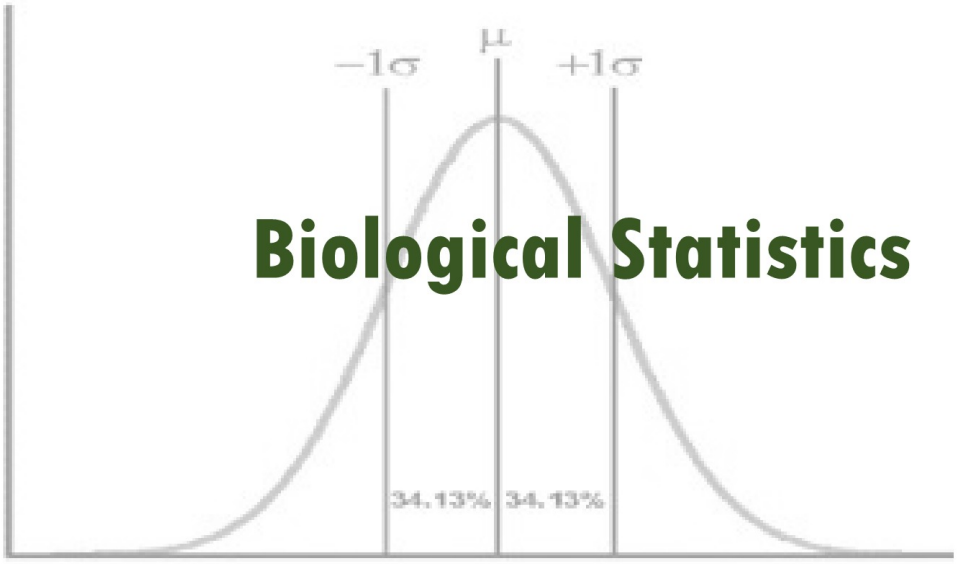
1, 2, 3, 4, 5, 6, 7, 8, 9,10,11 no mode

one mode ~ unimodal, two modes ~ bimodal, more ~ multimodal

Conclusion


MEAN
The "mean" is the "average". To find the mean, you add up all the numbers and then divide by the number of numbers.
TO FIND THE MEAN FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13
average the set of numbers:
(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15
Note that the mean isn't a value from the original list. This is a common result.
DO NOT assume that the mean will be one of the original numbers.
MEDIAN
The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order, so you may have sort the list first.
FOR AN ODD NUMBER OF VALUES: 1,5,2,8,7
Sort the numbers 1, 2, 5, 7, 8
FOR AN EVEN NUMBER OF VALUES: 1,5,2,10,8,7
Sort the numbers: 1, 2, 5,7, 8, 10.
TAKE THE AVERAGE OF THE TWO MEAN NUMBERS: (5+7)/2 = 6
MODE
The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.
TO FIND THE MODE FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13
Sort the numbers: 13, 13, 13, 13, 14, 14, 16, 18, 21


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Dispersion of Variation measurements

Measures of dispersion describe the spread of the data *i.e.* the divergence of the observations from their arithmetic mean that reflects the homogeneity of the values. These measures includes:

1. Range.
2. Variance.
3. Standard deviation.
4. Standard error

Explanation
If we have two sets

X = 17 16 15 18 19 17
Y= 7 13 50 20 11 5

The mean of X = 17
Also the mean of Y= 17 too
But, the values of set Y are more variant than X

Range (R)

It is the difference between the maximum and minimum values of a gevin data set.

How to calculate it ??

1. Arrange data in ascending or descending order.
2. Define the maximum and minimum values the apply the formula below:

$$R = Max. - Min.$$

The range is an inaccurate criterion because it depends on the largest and lowest value. In our example, the two samples have the same range, but if we calculated the variance, the second set will show more variant than the first.

Ex: find the range of the data sets below:

X = 2 5 7 6 10 6 7

Y = 8 13 9 16 8 10 9

Solution: arrange data ascendingly

X = 2 5 6 6 7 7 10 R= 10 – 2 = 8

Y = 8 8 9 9 10 13 16 R= 16 – 8 = 8

Variance (S^2 or σ^2)

One of the most widely used dispersion measures, it can be defined as the mean of the squares of deviations of values from their arithmetic mean.

How to calculate it??

There are two methods:

Squaring values method (preferred)

$$S^2 = \frac{\sum X_i^2 - \frac{\sum (X_i)^2}{n}}{n-1}$$

The deviations method

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Standard deviation (S)

It is the square root of the variance Also, it is commonly used measures, actually more than the variance itself, BECAUSE if we square the deviations in the variance, we must square the units, for example: cm becomes cm^2, it'll be an area unite not a length unit. When you root the value of the square of the deviation, the units will back to the normal status. Therefore, the standard deviation is:

Squaring values method (preferred)

$$S = \sqrt{\frac{\sum {X_i}^2 - \frac{\sum (X_i)^2}{n}}{n-1}}$$

The deviations method

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Standard Error (SE)

It is a measure that tells us how accurate is the arithmetic mean of the sample, *i.e*, is this mean a real estimation of the population mean or not. The lower the value of the standard error, the greater the accuracy of estimating the arithmetic mean, It is also, gave us an impression about the sample homogeneity.

How to calculate it ??

Simply by dividing standard deviation by square root of N:

$$SE = \frac{\delta}{\sqrt{n}}$$

δ= standard deviation
n= number of samples

Note:
S is the same as σ, except that S is used for the sample and σ is used for the population. In terms of calculation, it is the same thing.

Calculating the variance, standard deviation, and standard error of the grouped data (containing classes and frequencies)

These data mean the data that are in the form of frequency tables, meaning that the values are not the real values, but are grouped in classes, each class represented by a frequency, for example: (10-20) There are 5 observations, but indeed we don't know exactly value of each observation because it is vague and we just know that it is within the limits of the class. Anyway, the formula variance is:

The variance formula of grouped data

$$S^2 = \sum FiX_i{}^2 - \frac{\sum(FiX_i)^2}{\sum Fi}$$

Where:

X_i: Class mid-point= (Upper limit + Lower limit) / 2

Fi: Frequency

How to calculate the standard deviation and standard error ?

Calculating the variance, standard deviation, and standard error of the grouped data (containing classes and frequencies)

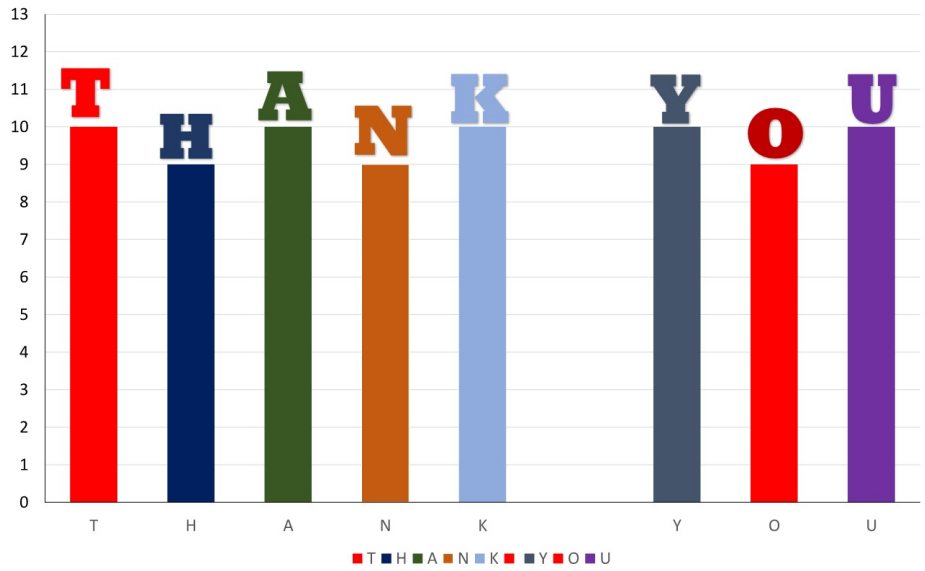
Ex: calculate the variance, standard deviation, and standard error for the following grouped data

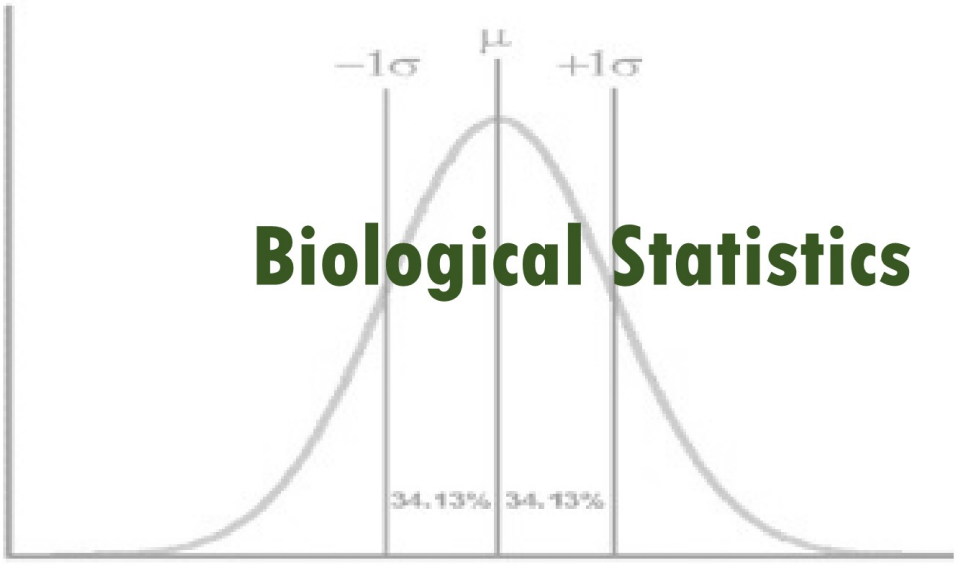
Input	Steps of solution

Classes	Fi	X_i	FiX_i	X_i^2	FiX_i^2
60-62	5	(60+62)/2= 61	(5 ×61) = 305	3721	(5 × 3721) = 18605
63-65	18	(63+65)/2= 64	(18 ×64) = 1152	4096	(18 ×4096) = 73728
66-68	42	(66+68)/2= 67	(42 ×67) = 2814	4489	(42 × 2814) = 188538
69-71	27	(69+71)/2= 70	(27 ×70) = 1890	4900	(27 × 4900) = 132300
72-74	8	(72+74)/2= 73	(8 ×73) = 584	5329	(8 × 5329) = 42632
	$\sum Fi = 100$		$\sum F_iX_i = 6745$		$\sum FiX_i^2 = 455803$

$$S^2 = \sum FiX_i^2 - \frac{\sum(FiX_i)^2}{\sum Fi} \Rightarrow S^2 = 455803 - \frac{(6745)^2}{100} = 852.75 \Rightarrow S = \sqrt{852.75} = 29.2$$

$$\Rightarrow SE = \frac{29.2}{\sqrt{100}} = 2.92$$


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

العينة Sample

What is sample?

A sample is a set of individual represents their population.

What are the good sample proprities?

- Randomness: The sample should be unbiased.
- Normally distributed: the number of correct predictions is approximately equal to the number of false predictions.
- Each observation should be independent.

Sample size

Why we need to calculate idol sample size?

In order to be representative of the population, so that the difference between the mean of the sample and the mean of the real population is as little as possible (not significant), thus all the next statistical analysis will be accurate give a clear vision about the actual population, therefore any considered decisions will be correct.

How to calculate sample size

Online sample size calculator

http://www.raosoft.com/samplesize.html

The formula requirement:

1. Confidence level, which is mostly =%0.95
2. The probable existence of interested trait ($\hat{p}$), which = 0.5 when the population being normally distributed.
3. The error margin (e^2): for 95% confidence level *i.e.* the error percent = 5%, the error margin will be little less = 0.04
4. The population size: for our example= 5000
5. α level divided by 2 = 0.025
6. Z value: can be calculated by MS Excel function: NORM.S.INV(1-0.025) = 1.96
7. Calculate the formula.

$$Sample\ size = \frac{\frac{Z^2 \times \hat{p} \times (1 - \hat{p})}{e^2}}{1 + (\frac{Z^2 \times \hat{p} \times (1 - \hat{p})}{e^2 \times N})}$$

$$Sample\ size = \frac{\frac{1.96^2 \times 0.5 \times 0.5}{0.04^2}}{1 + (\frac{1.96^2 \times 0.5 \times 0.5}{0.04^2 \times 5000})} = 535.91$$

Note: Z . value can be obtained using special statistical tables called a Z value tables, which are available on the Internet, or through Z calculators on the Internet as well.

Confidence intervals

First of all, we have to understand the confidence level concept

It is a percentage (95% for example), which means that if you repeat the same experiment or field survey over and over, you will get 95% results that match the population readings, in other words, your stats will be completely correct and reflect what is actually happening in the population.

What does confidence intervals tell us?

It is a measure of the reliability of the sampling method. As it determines the upper and lower values that sample observations must fall between themed in a particular community within the confidence level specified for the sample, regardless of the repeat sampling.

Confidence intervals

Simple example

If I want to collect a sample of students to study the average height at a confidence level of 95% and the confidence limits were (150-180) and this sample conformed to the conditions of the correct sample, then if I re-sampled this population over and over again, the height values that I'd measure are 95% fall within confidence interval. If not, the sample is not representive, and that any statistical analysis will be incorrect, so any decisions based on the obtained results will be completely wrong and unreliable.

Confidence intervals

When we use it?

When we don't know the population mean (μ). In this case, we cannot compare how close the sample mean from the population mean, therefore we cannot know whether our sample is representative of the community or not?

How to calculate confidence intervals?

Can be done using the following formula: $CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$

CI= Confidence interval.
$\bar{x}$ = Sample mean.
z = Confidence level value.
s= Sample standard deviation
n= Sample size

$z \frac{s}{\sqrt{n}}$

This called the error margin value, when we add or subtract it to/from the sample mean, we'll get the upper and lower intervals of confidence.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n-1}}$$

This is the standard deviation

Z value = t at degrees of freedom = n-1 and significance level equals 1- confidence level

How to calculate confidence intervals

Example: A sample of the lengths of 10 students was collected to determine the average height of the students population in a school. Identify the confidence intervals within the 95% confidence level to confirm the reliability of the collected sample .

Student length: 150 155 160 175 180 176 155 166 179 162

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

Solution:

Xi	150	155	160	175	180	176	155	166	179	162
$(Xi - \bar{X})^2$	249.64	116.64	33.64	84.64	201.64	104.04	116.64	0.04	174.24	14.44

$$\bar{X} = \frac{\sum Xi}{n} = \frac{1658}{10} = 165.8$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n-1}} = \sqrt{\frac{1095.6}{10-1}} = 11.033$$

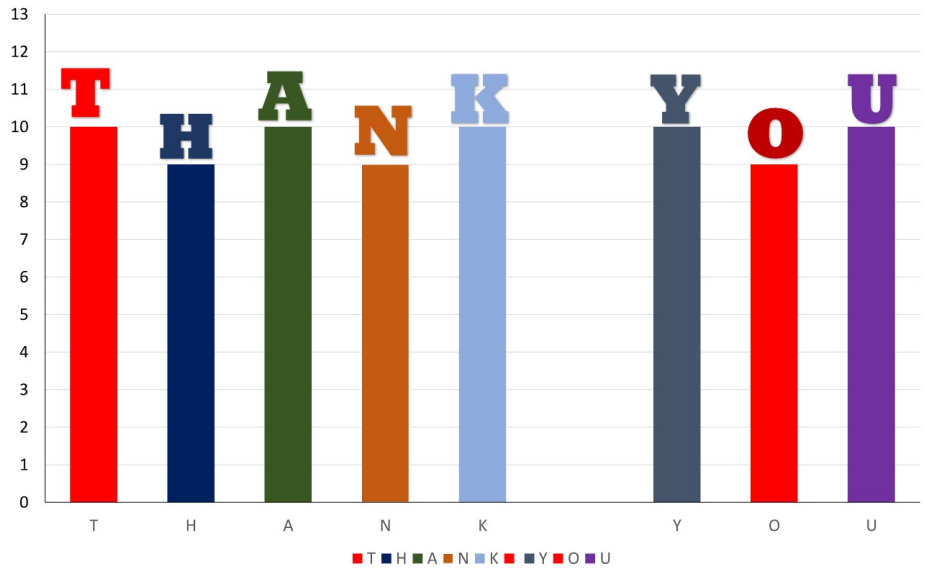
$$z = Norm.s.inv(1 - 0.025) = 1.9599$$

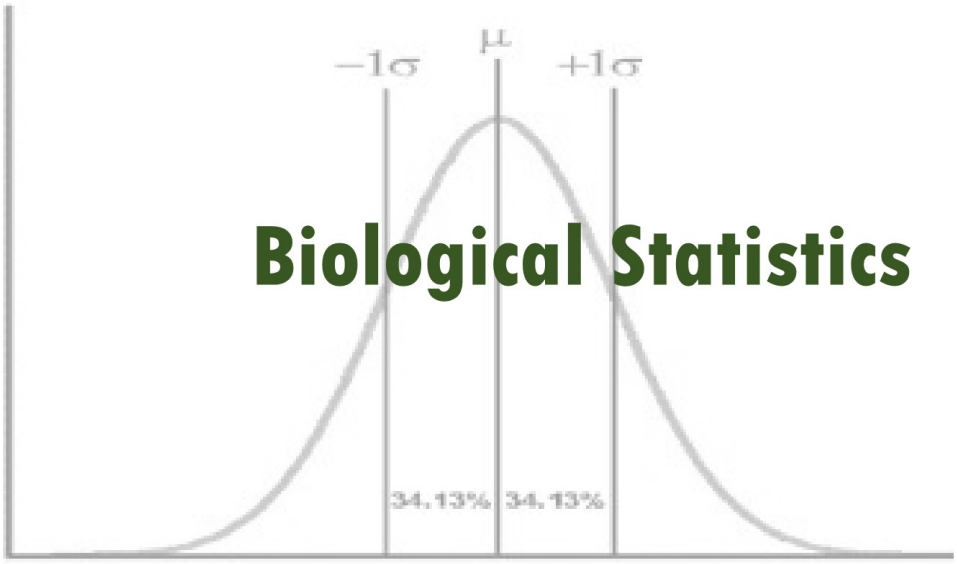
هذه أحد دوال اكسل

$$CI = 165.8 \pm 11.033 \times \frac{1.96}{\sqrt{10}}$$

$$CI = 165.8 \pm 6.83$$

$$CI = (158.97 - 172.63)$$


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Coefficient of variance (CV)

- It is a measure of relative variance, which is the ratio of the standard deviation to the arithmetic mean.
- This scale is usually used to compare the variance of two different variables in size or units. It is calculated using the following equation:

$$CV = \frac{\sigma}{\mu}$$ For population

OR

$$CV = \frac{S}{\bar{X}}$$ For sample

Where is:

σ or S: Standard deviation

μ or $\bar{X}$: Mean.

Coefficient of variance (C.V.)

Example: We have two samples of students for two different classes X and Y, how to know which of sample has the high variance in the students' levels, where the students' grades are as follows:

$$\bar{X} = \frac{543}{9} = 60.33$$

$$\bar{Y} = \frac{636}{9} = 70.67$$

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Xi	50	70	88	30	90	25	65	90	35
$(X_i -$	106.71	93.51	765.63	919.91	880.31	1248.21	21.81	880.31	641.61
Yi	65	70	75	67	85	82	77	55	60
$(Y_i -$	4994.25	4994.25	4994.25	4994.25	4994.25	4994.25	4994.25	4994.25	4994.25

$$S = \sqrt{\frac{5558}{8}} = 26.36$$

$$CV = \frac{S}{\bar{X}} = \frac{26.36}{60.33} = 0.44$$

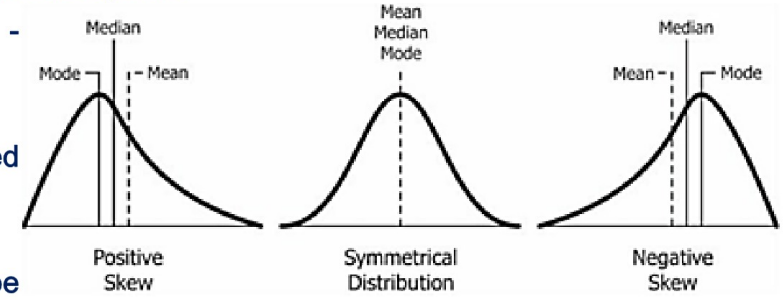
$$S = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}} \quad S = \sqrt{\frac{798}{8}} = 9.99$$

$$CV = \frac{S}{\bar{Y}} = \frac{9.99}{70.67} = 0.14$$

Since the CV of sample X is greater than of sample Y, then we conclude that the variance in the levels of students in sample X is higher than that of sample Y.

Skewness and Kurtosis measurements

- Skewness: It is a measure of the extent to which both sides of the sample distribution curve are symmetrical, i.e. how closely they match the normal distribution curve (the bell), precisely, it is the determination of the extent to which the shape of the sample distribution curve differs from the normal distribution curve.
- The skewness of a normal distribution curve = zero.
- When the skew value is between 0.5 and -0.5 it is acceptable.
- When the skewness value lay between -0.5 and -1, called negative skewness.
- when it is between 0.5 and 1, called positive skewness.
- If the value is less than -1 then it will be highly negative skewness, and if it is greater than 1 then it will be highly positive skewness.



https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

How to calculate skewness??

It can be calculated according the following equation :

$$Skewness = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^3}{n \times \sigma^3}$$

https://www.wallstreetmojo.com/skewness/

Where is:

Xi: the observation value

$\bar{X}$: The sample mean

σ: standard deviation

Skewness and Kurtosis measurements

Example: Suppose that we have the following observations of a random sample collected from a known population, How to certify that it is normally distributed?

Xi	$(X_i - \bar{X})^2$	$(Xi - \bar{X})^3$
12	400	-8000
13	361	-6859
54	484	10648
56	576	13824
25	49	-343
160	$\sum(X_i - \bar{X})^2$= 1870	$(Xi - \bar{X})^3$= 9270

$$\bar{X} = \sum \frac{Xi}{n} = \frac{160}{5} = 32$$

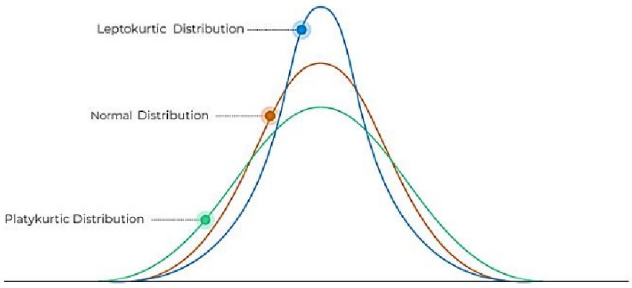
$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{1870}{4}} = 21.62$$

$$Skewness = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^3}{n \times \sigma^3} = \frac{9270}{5 \times 10108.17} = 0.183$$

Since the skewness value is greater than zero, then the distribution curve for this sample is positive skewness and because it lies between 0.5 and -0.5, it'll be acceptable and the data can be considered as normally distributed.

Skewness and Kurtosis measurements

- Kurtosis: This parameter is actually concerned with the ends distribution curve (how wide it is), as it measures the extent to which extreme values (much higher or lower than the mean) are present in the sample because these values make the mean value unrepresentative of the sample. That is, it is a measure of how extreme the sample values are.



- When kurtosis value = 3 it is called Mesokurtic which is the standard case of normal distribution curve.
- When kurtosis being greater than 3, it is called leptokurtic, meaning that the extreme values in the sample are less than in the normally distributed sample.
- When kurtosis being less than 3, it is called platykurtic, which means that the sample contains extreme values compared to the normally distributed sample.
- When Kurtosis lays between -2 and +2 it considered acceptable according to George & Mallery (2010)

George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson.

How to calculate kurtosis

It can be calculated according the following equation:

$$Kurtosis = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^4}{n \times \sigma^4}$$

https://analystprep.com/cfa-level-1-exam/quantitative-methods/kurtosis-and-skewness-types-of-distributions/

Where is:

Xi: the observation value

$\bar{X}$: The sample mean

σ: standard deviation

Since the kurtosis of the normal distribution curve is 3, 3 is subtracted from the product of the above equation to get the actual value of kurtosis when the normal curve is zero

Skewness and Kurtosis measurements

Example : If we calculate the kurtosis for our previous example

$$\bar{X} = \sum \frac{Xi}{n} = \frac{160}{5} = 32$$

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{1870}{4}} = 21.62$$

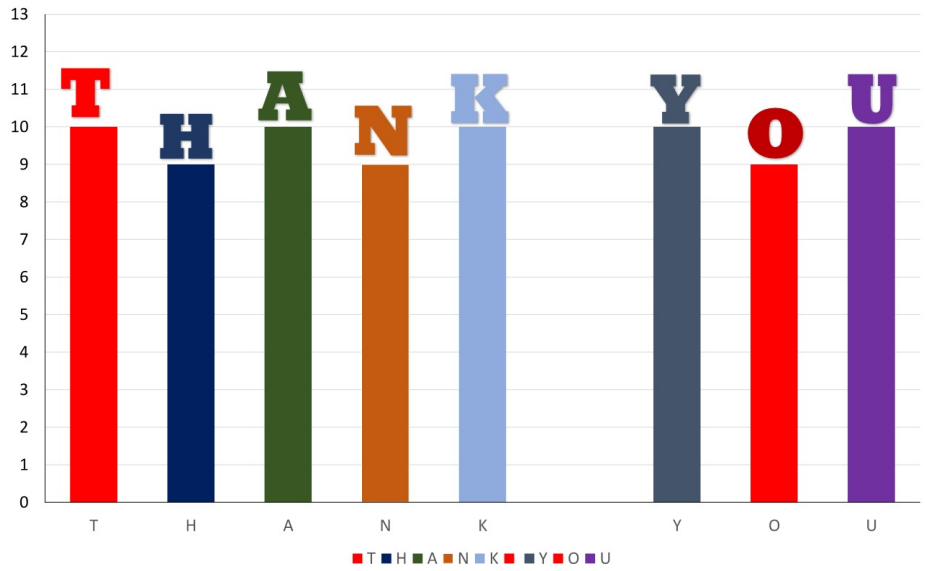
Xi	$(X_i - \bar{X})^2$	$(Xi - \bar{X})^4$
12	400	160000
13	361	130321
54	484	234256
56	576	331776
25	49	2401
160	$\sum(X_i - \bar{X})^2$= 1870	$(Xi - \bar{X})^4$= 858754

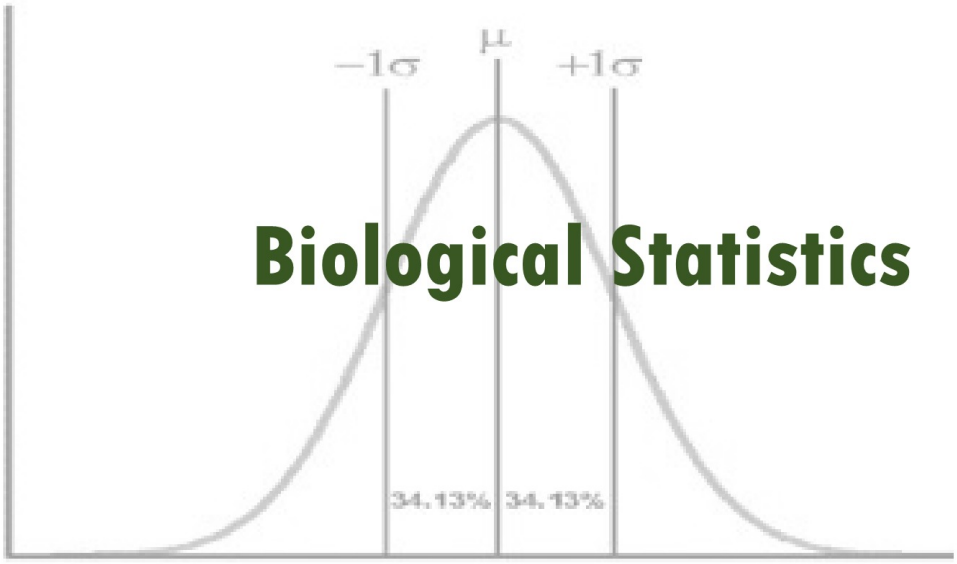
$$Kurtosis = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^4}{n \times \sigma^4} = \frac{858754}{5 \times 218556.3} = 0.785$$

$$Excess\ kurtosis = 0.785 - 3 = -2.213$$

Then, we subtract the standard value of the kurtosis of the normal distribution curve from the value above to get the excess kurtosis or the actual kurtosis :

Since the flatness value is negative, then this kurtosis type is Platykutrosis. Since the value is less than -2, the data is considered as abnormally distributed.


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Simplified review

In order to clarify any confusion about any the methods and/or statistical tests that we learned previously and what we'll learn later we'll summarize here a simplified scheme describes the logical steps to analyze any raw data to extract results and interpret them to make our final decision depending on a correct scientific basis guaranteed solving the problem under study.

All that we have learned up to this lecture is related to descriptive statistics of collected data, which includes reviewing and examining our data to preparing it for statistical tests, and this in fact represents almost half of the statistical analysis. the second half, will include studying relationships or making comparisons and determining whether they are statistically significant or not?

Simplified statistical analysis protocol – 1st slide

Determine the studied problem: Ex.
(Efficiency of a drug to control a disease)

↓

Determining the population of the community under study
(Ex: 4000,00 0 individuals)

↓

Calculate the sample size
(for our example, will be 385)
This means, that statistical analysis of this sample will give us a representative results, as if we analyzed a data of 4 million people

→

Collect data
(Make sure the sample data is random and unbiased)

↓

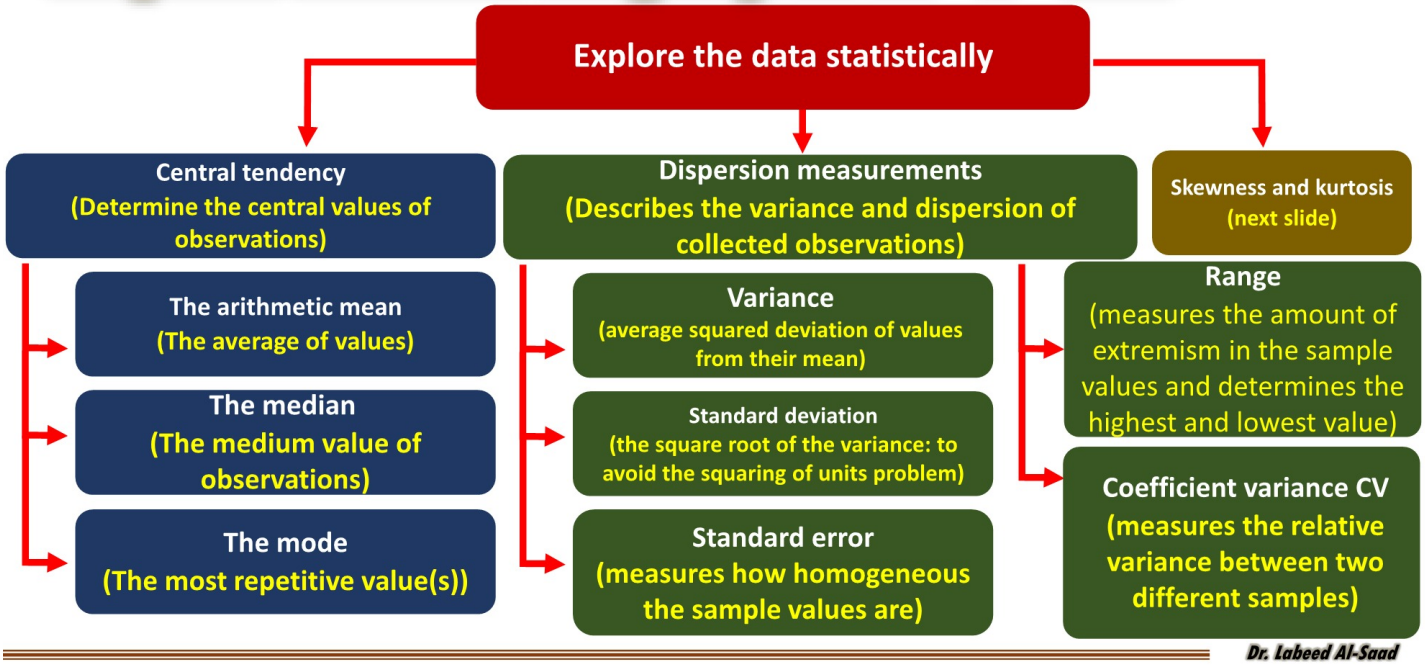
Organize data into tabular tables
(converted from incomprehensible data to information)

↓

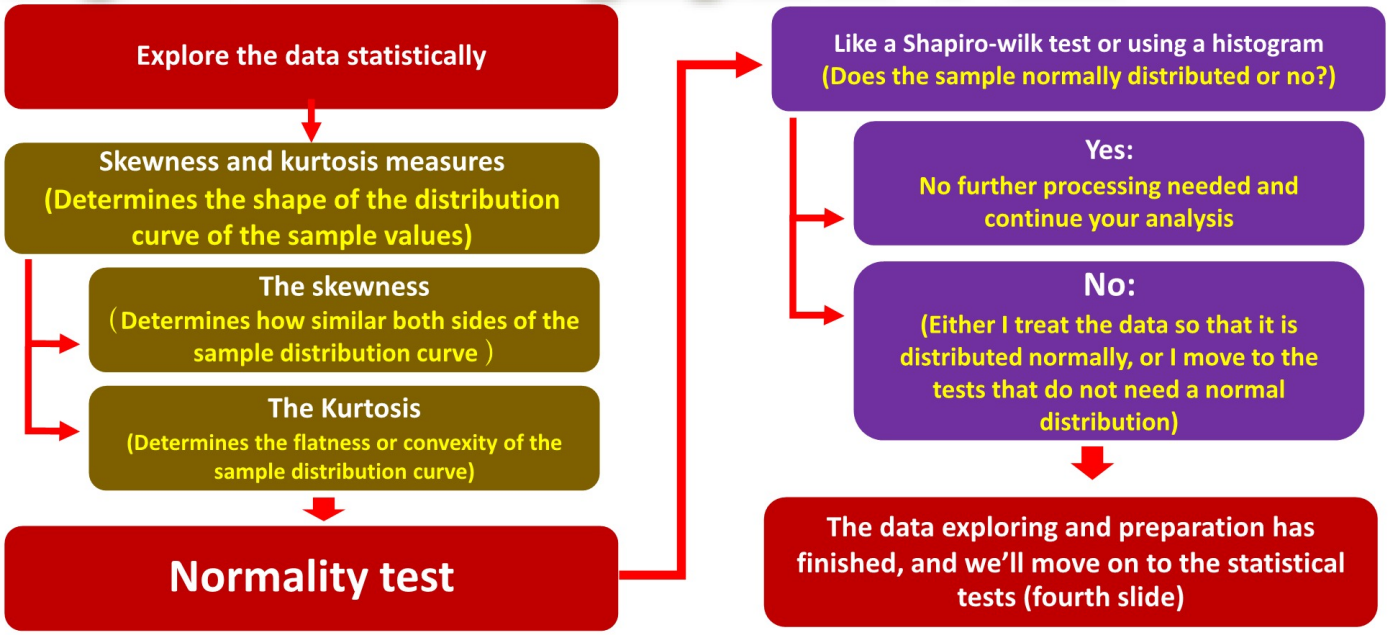
Statistical explore of sample data
(Central tendency, dipersion, slewness and kurtosis)

↓

Simplified statistical analysis protocol – 2nd slide

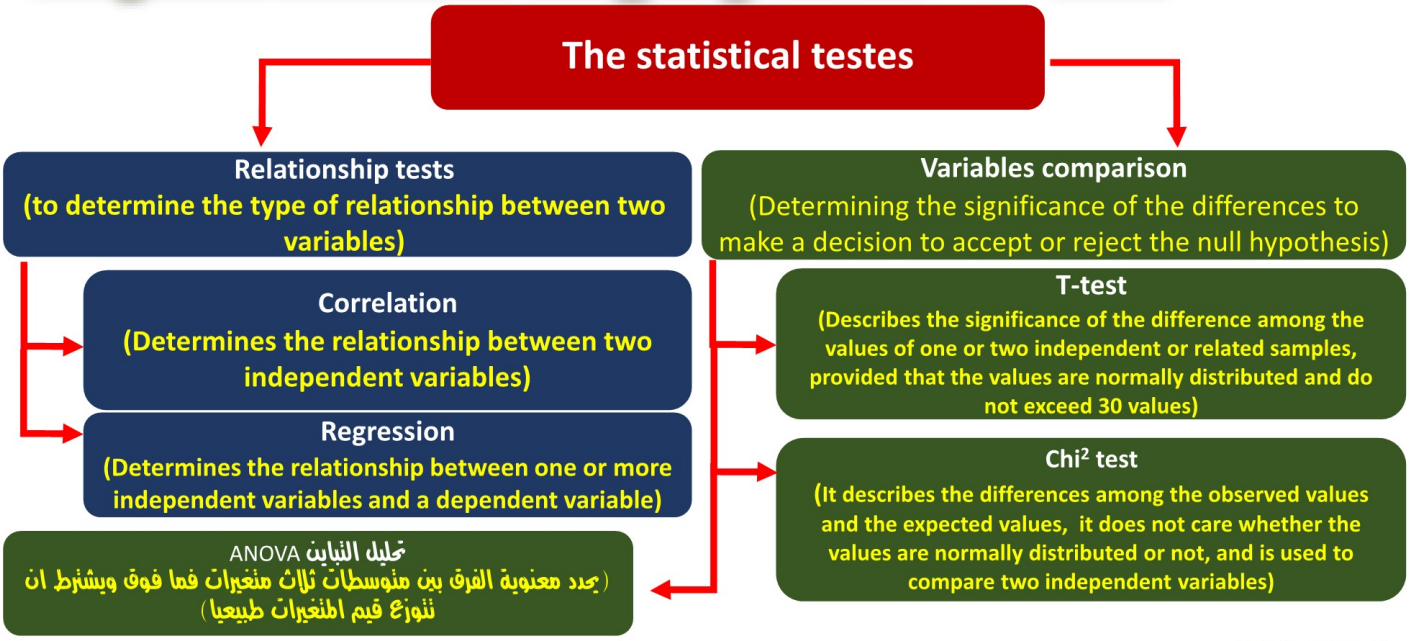


Simplified statistical analysis protocol – 3rd slide



Biological statistics

Simplified statistical analysis protocol – 4th slide



Simplified statistical analysis protocol – 5th slide

Simple table helps to select the suitable statistical test

Choosing the best statistical test

- **Q1: What type of data do you have?**
- **Q2: How many samples do you have?**
- **Q3: What is the test supposed to do?**

Q3 → / Q1 →	Compare the Data		Seek Relationships
Q2 ↓	Categorical Data	Quantitative Data	
One Sample	1 sample proportion	1 sample t	
Two Samples	2 sample proportions	2 sample t	
Two Samples Special		2 sample t Paired t	Correlation/ Regression
Three or more samples		One-Way ANOVA	

The study of relationship among variables - Correlation

The correlation coefficient is used to estimate the linear relationship between two variables and the direction of this relationship. The value of correlation coefficient arranging between +1 and -1 where the sign indicates the direction of relationship.

Types of Correlation coefficient

There are three main types of correlation coefficient:

1. Pearson correlation coefficient: It is a parametric measure that is used with quantitative data (data in the form of real numbers).

The study of relationship among variables - Correlation

Types of Correlation coefficient

2. Spearman's rho correlation coefficient: It is a non-parametric measure and is used with data that is in the form of ranks (that is, it is concerned with the rank of the values not their amount, for example, instead of care about students' scores, we care about who is the first and who is the second .. etc. regardless of the value of the degree).

3. Kendall's tau correlation coefficient: It is also used with non-parametric measures

How to calculate Pearson correlation coefficient

We can perform that using one of the following formulas:

$$r = \frac{\sum_{i=1}^{n}[(xi - \bar{x}) \times (yi - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2 \times \sum_{i=1}^{n}(yi - \bar{y})^2}}$$

OR

$$r = \frac{\sum_{i=1}^{n} xi \times yi - \frac{\sum_{i=}^{n} xi \times \sum_{i=1}^{n} yi}{n}}{\sqrt{\left(\sum_{i=1}^{n} xi^2 - \frac{(\sum_{i=1}^{n} xi)^2}{n}\right) \times \left(\sum_{i=1}^{n} yi^2 - \frac{(\sum_{i=1}^{n} yi)^2}{n}\right)}}$$

Which means simply:

$$r = \frac{Variance\ xy}{Variance\ x\ * Variance\ y} \qquad r = \frac{S^2 x * y}{S^2 x\ \times S^2 y}$$

$$|t| = r \times \sqrt{\frac{n-2}{1-r^2}}$$

The T-test is used to identify the significance of the correlation. If the calculated T is greater than the tabulated T at degrees of freedom df = n-2, then the correlation is significant.

Where:

r: Correlation coefficient

xi : X observations

yi : Y observations

$\bar{x}$: X mean

$\bar{y}$: Y mean

n: observations count.

Ex: The following data represent the observation of two independent random variables. Indicate whether or not the two variables are likely to be related?

X-value	1.24	1.34	1.39	1.41	1.64	1.44	1.48	1.51	1.54	1.54	1.54	1.62
Y-value	1.30	1.50	1.70	1.50	1.44	1.47	1.60	1.60	1.80	1.50	1.70	1.90

Solution:

xi	yi	$(xi-\bar{x})$	$(xi-\bar{x})^2$	$(yi-\bar{y})$	$(yi-\bar{y})^2$	$(xi-\bar{x})\times(yi-\bar{y})$
1.24	1.30	-0.23	0.0529	-0.28	0.08	0.06
1.34	1.50	-0.13	0.02	-0.08	0.01	0.01
1.39	1.70	-0.08	0.01	0.12	0.01	-0.01
1.41	1.50	-0.06	0.00	-0.08	0.01	0.00
1.64	1.44	0.17	0.03	-0.14	0.02	-0.02
1.44	1.47	-0.03	0.00	-0.11	0.01	0.00
1.48	1.60	0.01	0.00	0.02	0.00	0.00
1.51	1.60	0.04	0.00	0.02	0.00	0.00
1.54	1.80	0.07	0.00	0.22	0.05	0.02
1.54	1.50	0.07	0.00	-0.08	0.01	-0.01
1.54	1.70	0.07	0.00	0.12	0.01	0.01
1.62	1.90	0.15	0.02	0.32	0.10	0.05
Sum: 17.69	19.01	0.0500	0.1485	0.050	0.309700	0.116700

$$r = \frac{\sum_{i=1}^{n}[\,(xi - \bar{x}) \times (yi - \bar{y})\,]}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2 \times \sum_{i=1}^{n}(yi - \bar{y})^2}}$$

$$r = \frac{0.116700}{\sqrt{0.148500 \times 0.309700}} = 0.544$$

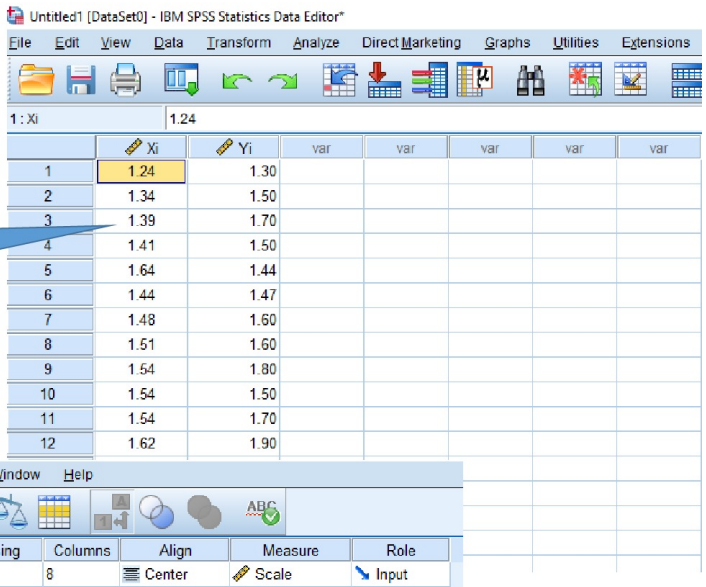
$$|t| = r \times \sqrt{\frac{n-2}{1-r^2}} = 0.544 \times \sqrt{\frac{12-2}{1-0.544^2}} = 2.050$$

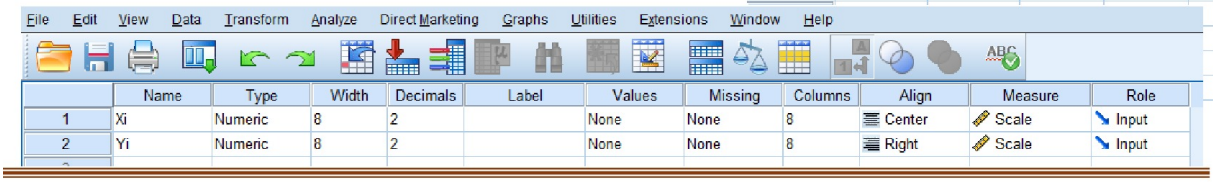
Then we extract the tabular t-value at 10 degrees of freedom and 0.05 alpha level, which equals 2.228

Note that the two variables are positively related, but not significant, *i.e.* there is a direct relationship between them, but it is not significant.

The solution in SPSS

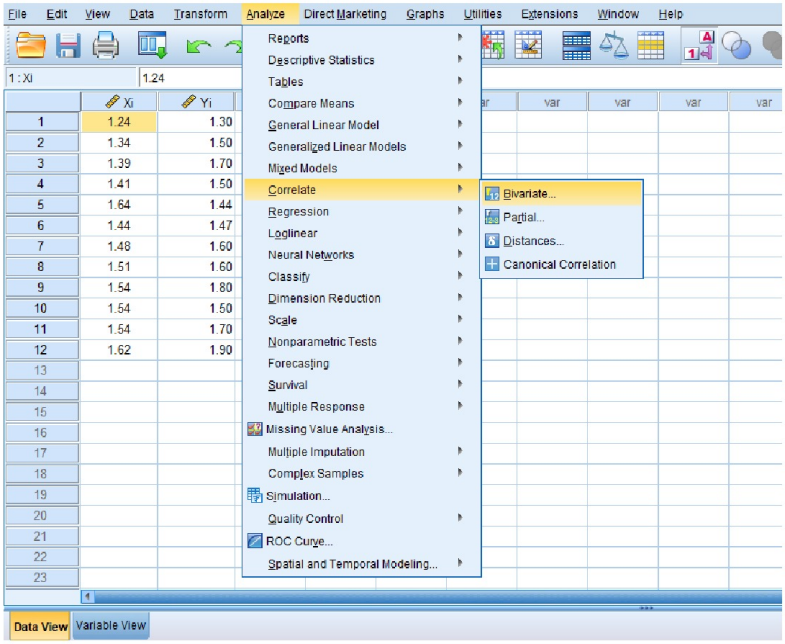
1. Identify variables and input data





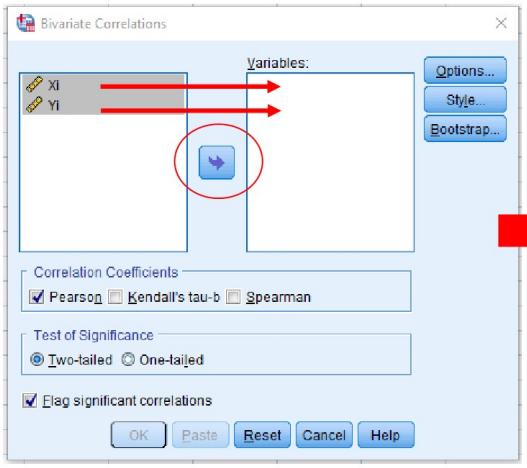
The solution in SPSS

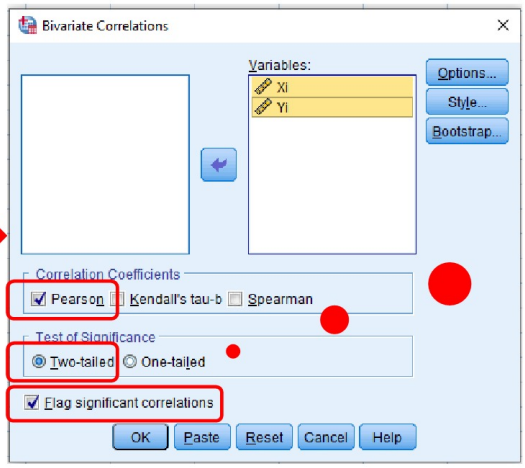
2. Go to Analyze >> Correlate >> Bivariate



The solution in SPSS

3. Transfer variables to the variables list using bottom

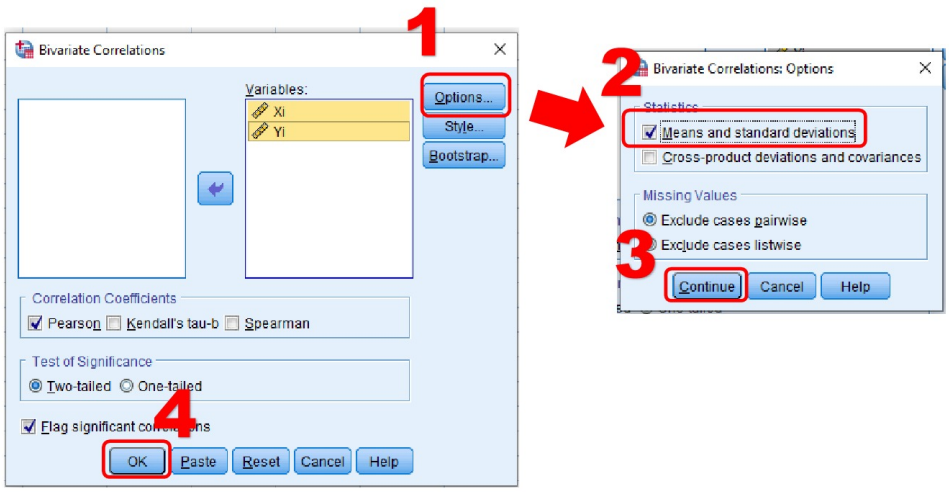




Be sure that Pearson box, Two tailed and Flag significant correlations are selected

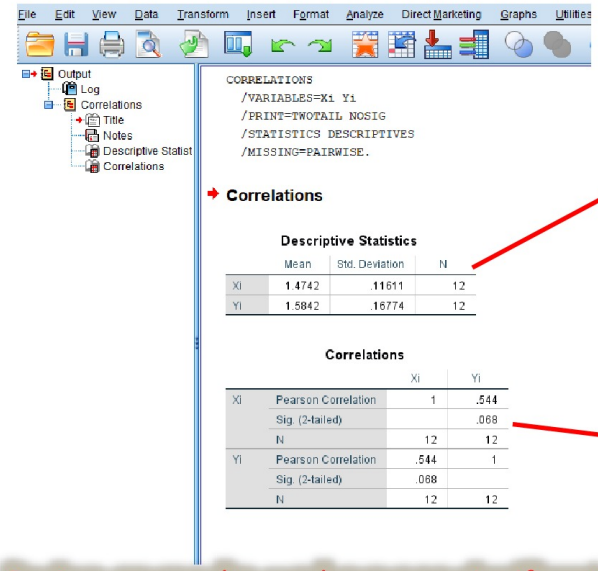
The solution in SPSS

4. Optionally click Options >> select "Means and standard deviations" >> OK



The solution in SPSS

5. The results will appear in output window



Descriptive Statistics

	Mean	Std. Deviation	N
Xi	1.4742	.11611	12
Yi	1.5842	.16774	12

Correlations

		Xi	Yi
Xi	Pearson Correlation	1	.544
	Sig. (2-tailed)		.068
	N	12	12
Yi	Pearson Correlation	.544	1
	Sig. (2-tailed)	.068	
	N	12	12

Correlation value

P- value

Basing on results, we have non significant positive correlation.

Calculating Spearman's rho correlation coefficient

It can be calculated using the following formula

$$r_s = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)}$$

Where:

r_s: Spearman rho correlation coefficient.

d^2: The square difference of two opposite ranks *(rank xi – rank yi)*

n : observation count.

xi	Xi rank	yi	Yi rank	di	di^2
25	3	80	2	1	1
15	4	77	3	1	1
30	2	35	4	-2	4
50	1	90	1	0	0
					$\sum d^2 = 6$

$$r_s = 1 - \frac{6 \times 6}{4 \times 15} = 0.4$$

Calculating Kendall's tau correlation coefficient

It can be calculated using the following formula

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

C means that y of the right pair is greater than y of the left pair and x of the right pair is greater than x of the left pair. The same thing in the case of the smallest means the same direction. If they differ, it becomes D.

xi	yi	Xi rank	Yi rank	Pair 1		Pair 2		C/D
15	77	4	3	4	3	3	2	C
25	80	3	2	4	3	2	4	D
30	35	2	4	4	3	1	1	C
50	90	1	1	3	2	2	4	D
				3	2	1	1	C
				2	4	1	1	C

$$\tau = \frac{4-2}{6} = 0.33$$

whereas:

τ : Kendall Tau correlation coefficient

n_c: the number of concordant pairs (C) i.e. the x-value of the second is greater/less than the x-value of the first and the y-value of the second is greater/less than the y-value of the first.

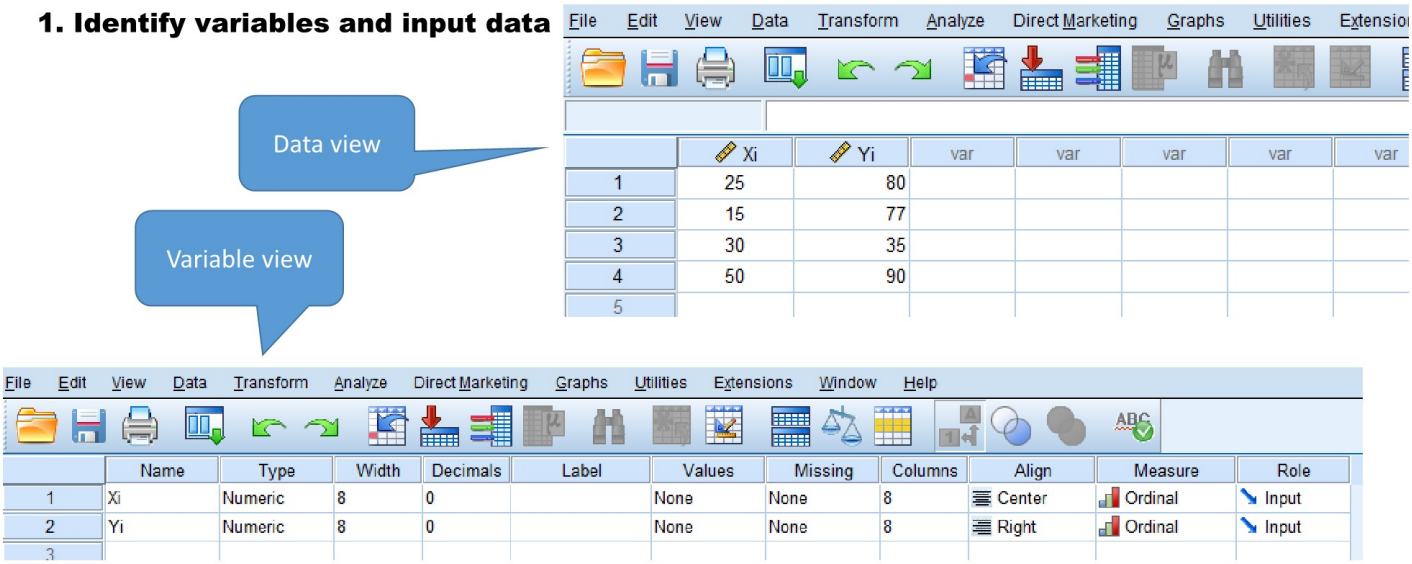
n_d: The number of pairs that are not compatible (D), meaning that the above condition does not apply

n: the number of values.

- The values of X are arranged in ascending order and their rank extracted.
- Y ranks are extracted and each rank is placed in front of its corresponding X.
- Multiple comparisons are made between each pair and the rest of the pairs to determine the number of pairs that are concordant (n_c) and discordant (n_d)

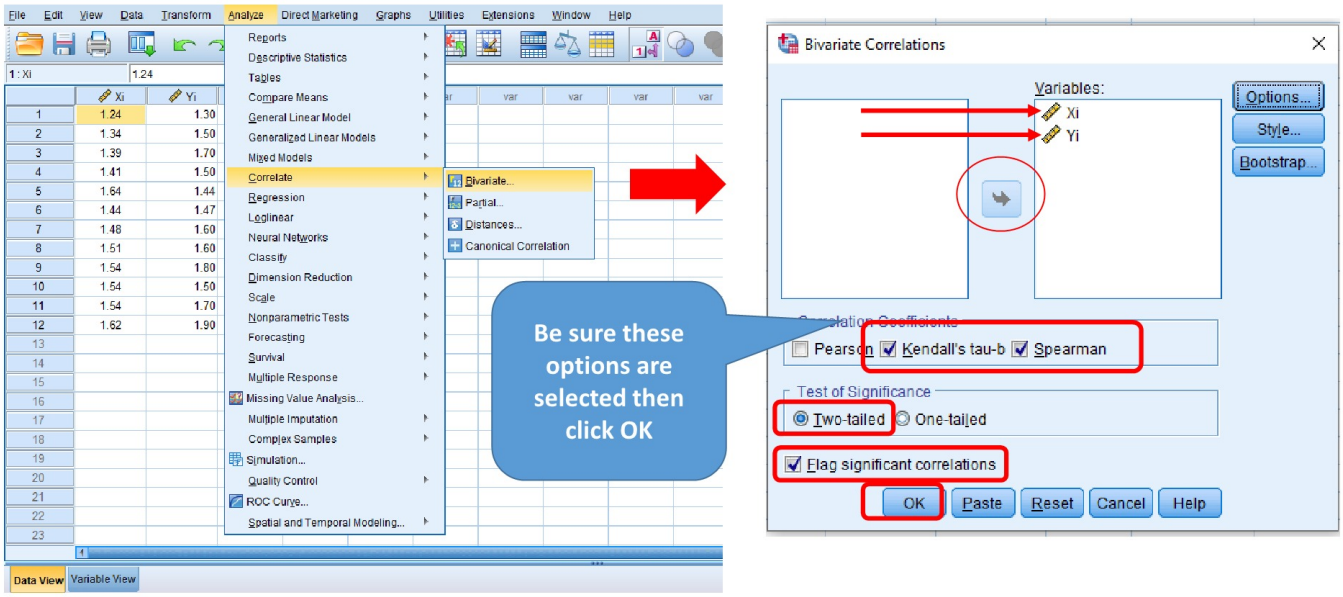
The solution in SPSS

1. Identify variables and input data



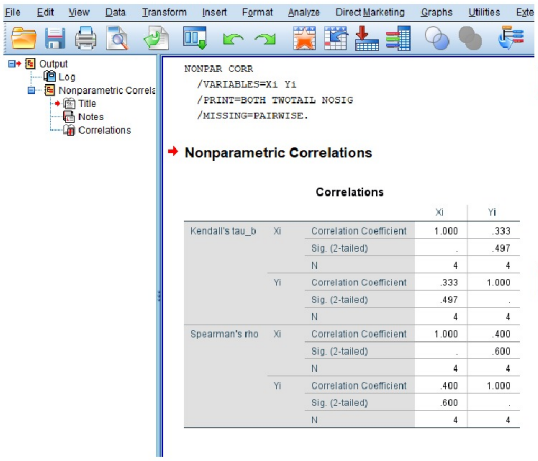
The solution in SPSS

2. Go to Analyze >> Correlate >> Bivariate >> transfer variables to the variables list using bottom



The solution in SPSS

3. The results will appear in output window



Correlations

			Xi	Yi
Kendall's tau_b	Xi	Correlation Coefficient	1.000	.333
		Sig. (2-tailed)	.	.497
		N	4	4
	Yi	Correlation Coefficient	.333	1.000
		Sig. (2-tailed)	.497	.
		N	4	4
Spearman's rho	Xi	Correlation Coefficient	1.000	.400
		Sig. (2-tailed)	.	.600
		N	4	4
	Yi	Correlation Coefficient	.400	1.000
		Sig. (2-tailed)	.600	.
		N	4	4

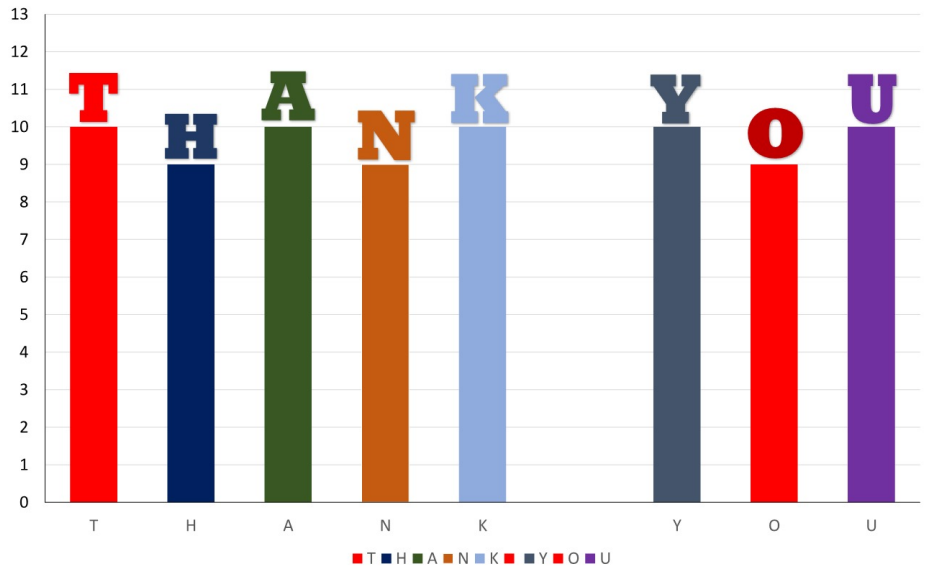
Correlation value

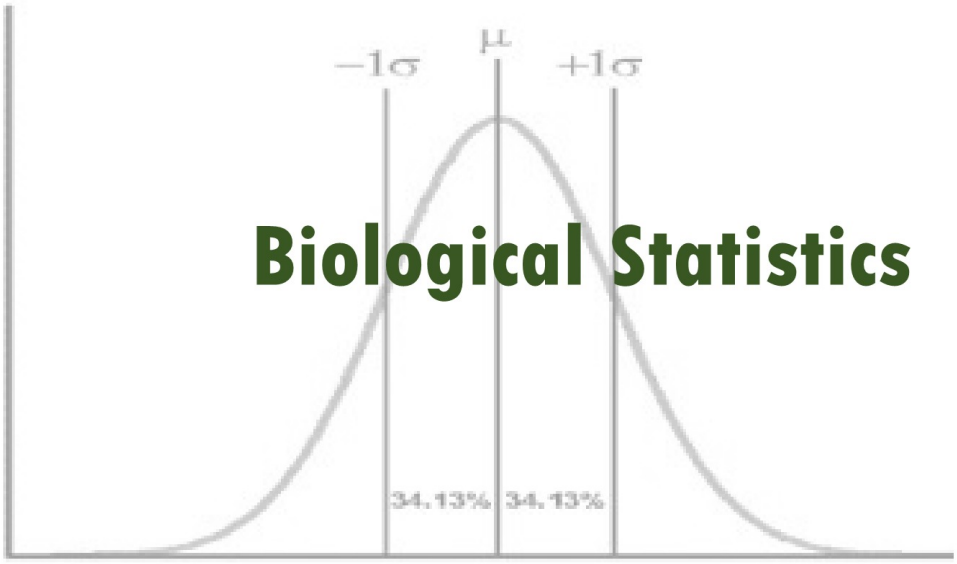
P- value

Correlation value

P- value

Basing on results of both methods, we have non significant positive correlation .


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

T-test to compare between two means

It is a parametric test usually used to compare two means to determine the significance deference between them *i.e.* does the difference return to accident or no. This test use when the sample size being relatively small (30 observations or less)

The test roles

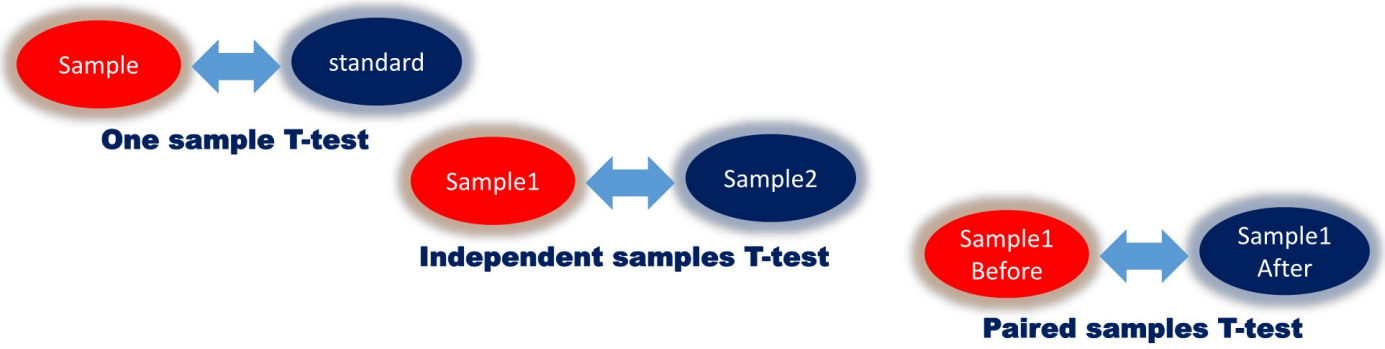
- The sample should be normally distributed.
- The sample should be randomly collected (unbiased).

Types of T-test

There are three types of T-test:

1. One sample T-test: It is used to compare the mean of a sample with a known standard mean and to determine the significance of the difference between them, such as comparing the heart pulses/ minute for a sample of students with the normal rate (72 pulses / minute) to measure the extent to which the sample rate differs from the standard rate and is this difference significant or not? The standard mean here is 72.
2. Independent sample T-test: used to compare between to independent samples mean such as comparing the heart pulses/ minute between two groups of students (A & B)

3. Paired samples T-test: used to compare between the means of two correlated samples such as comparing the heart pulses / minute of a student sample before and after performing a sport exercise.



How are T tests performed?

One sample T test:

This test performed by calculating T value and comparing it with the tabulated T-value at a degrees of freedom = n-1 and a level of significance= 0.05.

$$|t| = \frac{\bar{X} - \mu_0}{S_x/\sqrt{n}}$$

Whereas:

$|t|$ the calculated T value.

$\bar{X}$: The sample mean.

μ_0 : The population mean.

$S_x/\sqrt{n}$: The standard error.

:

Example : Determine wither the heart pulses/ minute of the sample below is differ significantly than the standard heart pulses rate (72 pulses/ minute) or no?

Xi	70	77	90	100	85	80	72	65	90	88	76	101	95	87

Solution : Firstly, we have to ensure that the sample is normally distributed or no using Shapiro-Wilk?

$$W = \frac{\left(\sum_{i=1}^{n} aiX_{(i)}\right)^2}{\sum_{i=1}^{n}(Xi-\bar{X})^2}$$

$$W = \frac{39^2}{1594} = 0.954$$

Since, the tabular W = 0.953, which is slightly smaller than the calculated one, but the value of P = 0.7 was greater than 0.05, which is considered more accurate in estimating the significance, if the sample is normally distributed. So we can move on to the T . test

Xi	*sorted xi*	*(Xi-mean)^2*	*ai*	*X(i)*	*ai*x(i)*
70	65	361	0.5251	36	18.9036
77	70	196	0.3316	30	9.948
90	72	144	0.246	23	5.658
100	76	64	0.1802	14	2.5228
85	77	49	0.124	13	1.612
80	80	16	0.0727	8	0.5816
72	85	1	0.024	2	0.048
65	87	9			
90	88	16			
88	90	36			
76	90	36			
101	95	121			
95	100	256			
87	101	289			
Sum		1594			39.274

Xi	70	77	90	100	85	80	72	65	90	88	76	101	95	87	1176
Xi^2	4900	5929	8100	10000	7225	6400	5184	4225	8100	7744	5776	10201	9025	7569	100378

$$S^2 = \frac{\sum X_i{}^2 - \frac{\sum (X_i)^2}{n}}{n-1} = \frac{100378 - \frac{(1176)^2}{14}}{14-1} = 122.615$$

$$S = \sqrt{S^2} = \sqrt{122.615} = 11.073$$

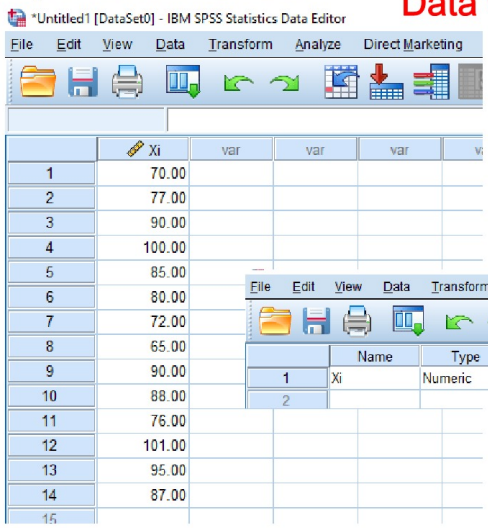
$$|t| = \frac{\bar{X} - \mu_0}{S_x/\sqrt{n}} = \frac{84 - 72}{11.073/\sqrt{14}} = 4.054$$

Since the calculated T value= 4.054, which it larger than tabular one = 2.144, then the average number of heart pulses of people in the sample (84) differs significantly from the standard mean (72 pulses/ min).

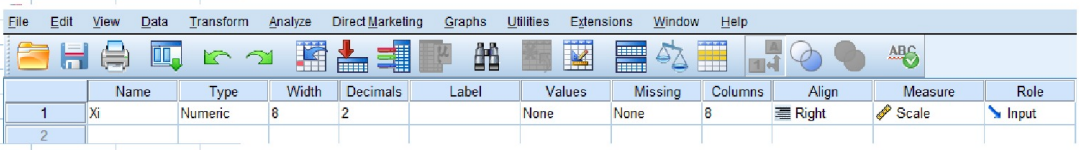
Performing solution in SPSS :

1. Input data in SPSS:

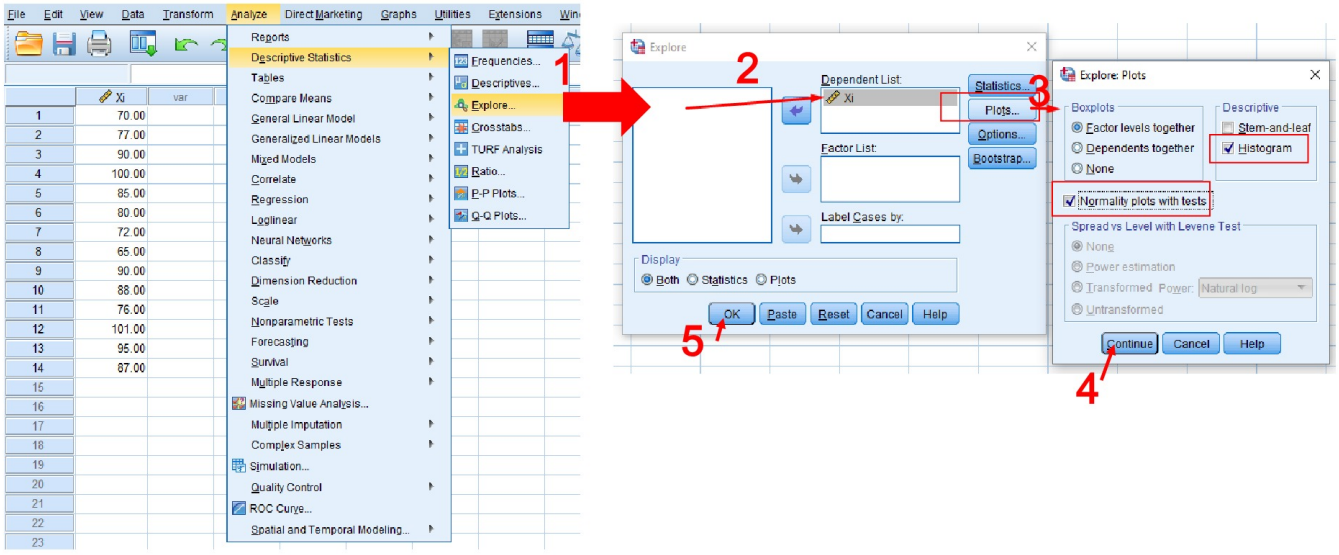
Data view



Variable view



2. Examine normality of sample:



3. The results will appear:

Variance	122.615	
Std. Deviation	11.07318	
Minimum	65.00	
Maximum	101.00	
Range	36.00	
Interquartile Range	16.25	
Skewness	-.092	.597
Kurtosis	-.887	1.154

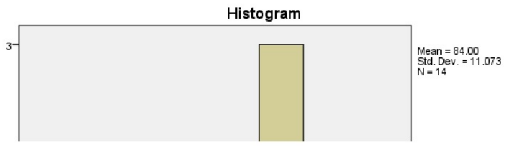
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Xi	.107	14	.200*	.968	14	.847

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

a. Lilliefors Significance Correction

Xi



Tests of Normality P-value

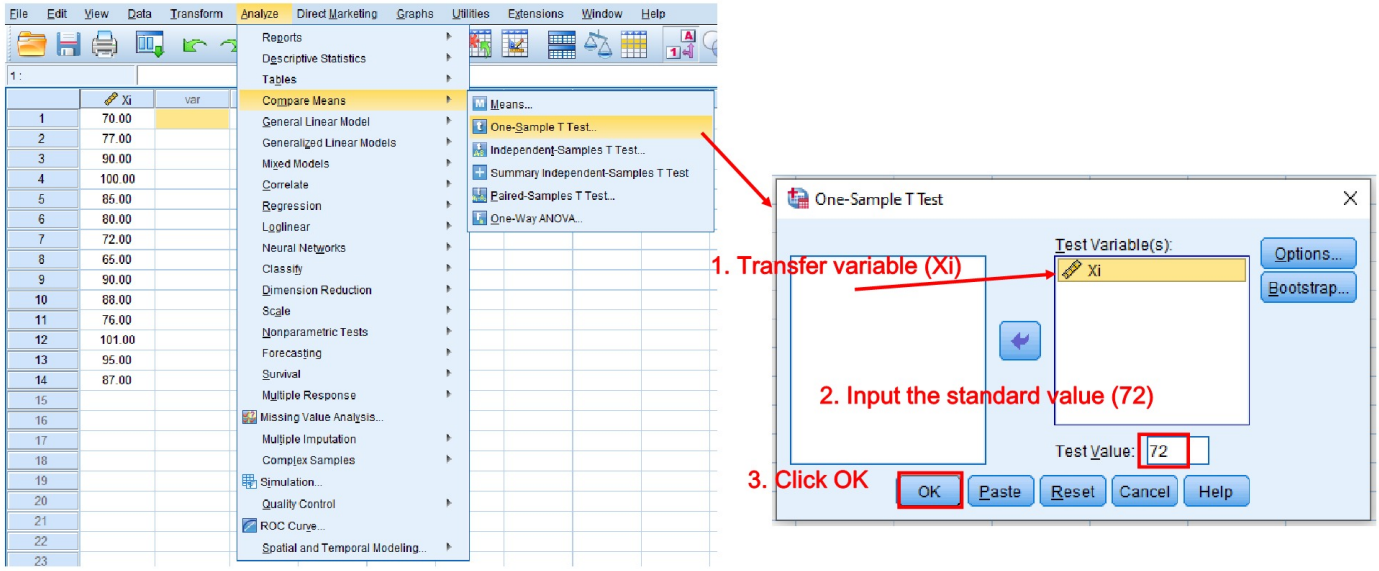
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Xi	.107	14	.200*	.968	14	.847

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

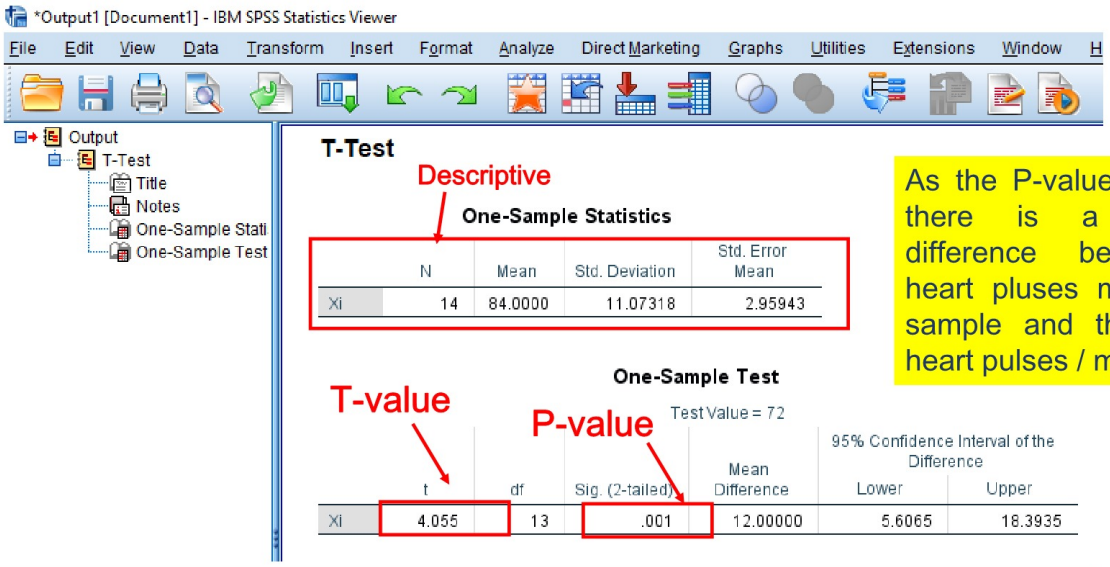
a. Lilliefors Significance Correction

As P-value (0.847) > 0.05, so there is no significant difference between the sample mean and the population mean *i.e.* the sample is normally distributed

4. Go to analyze >> Compare means >> One-sample T test



5. The results of T-test will appear



As the P-value < 0.01, so there is a significant difference between the heart pluses mean of the sample and the standard heart pulses / min.

How are T tests performed?

Independent samples T test:

This test is performed by calculating T value and comparing it with the tabulated one at degrees of freedom = n1 + n2 -2 and a level of significance= 0.05. :

$$|t| = \frac{\bar{X} - \bar{Y}}{\sqrt{(S_x^2/n_1) + (S_y^2/n_2)}}$$

Whereas:

$\bar{X}$: The mean of sample X.

$\bar{Y}$: The mean of population.

S_x^2 : Variance of X.

S_y^2 : Variance of Y.

n_1 : Number of X observations.

n_2 : Number of Y observations.

Example: Examine the significance of differences between means of heart pulses of the groups X and Y that their data listed below considering both of them as normally distributed.

Xi	70.00	77.00	90.00	100.00	85.00	80.00	72.00	65.00	90.00	88.00	76.00	101.00	95.00	87.00
Yi	82.00	80.00	100.00	102.00	77.00	85.00	72.00	70.00	85.00	72.00	77.00	75.00	80.00	75.00

Solution :

	Xi	*Yi*			
	70.00	82.00	3.14	8.76	6.72
	77.00	80.00			
	90.00	100.00			
	100.00	102.00			
	85.00	77.00			
	80.00	85.00			
	72.00	72.00			
	65.00	70.00			
	90.00	85.00			
	88.00	72.00			
	76.00	77.00			
	101.00	75.00			
	95.00	80.00			
	87.00	75.00			
Mean	84.00	80.86			

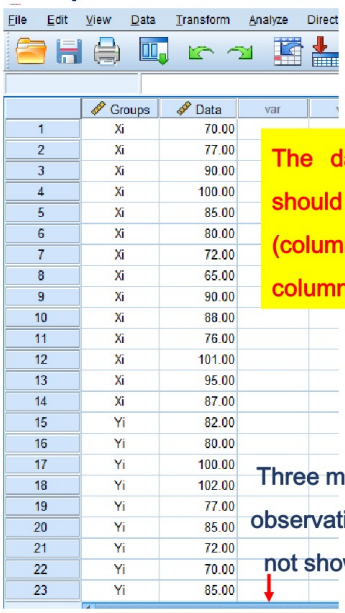
$$|t| = \frac{\bar{X} - \bar{Y}}{\sqrt{(S_x^2/n_1) + (S_y^2/n_2)}} = \frac{84 - 80.86}{\sqrt{8.76 + 6.72}} = 0.798$$

$$df = n_1 + n_2 - 2 = 14 + 14 - 2 = 26$$

Since the tabulated T= 2.06 lager then calculated T= 0.798, so the differences between $\bar{X}$ and $\bar{Y}$ is not significant *i.e.* accepting H_0.

Performing solution in SPSS :

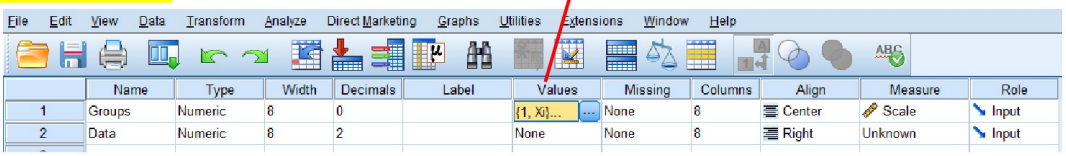
1. Input data: Data view

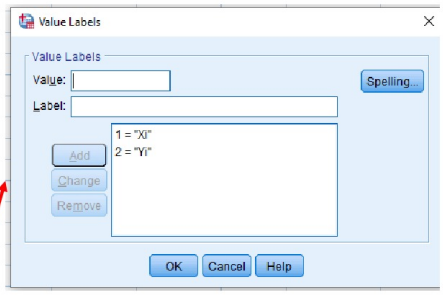


The data input here should be in this style (column for data and column for groups)

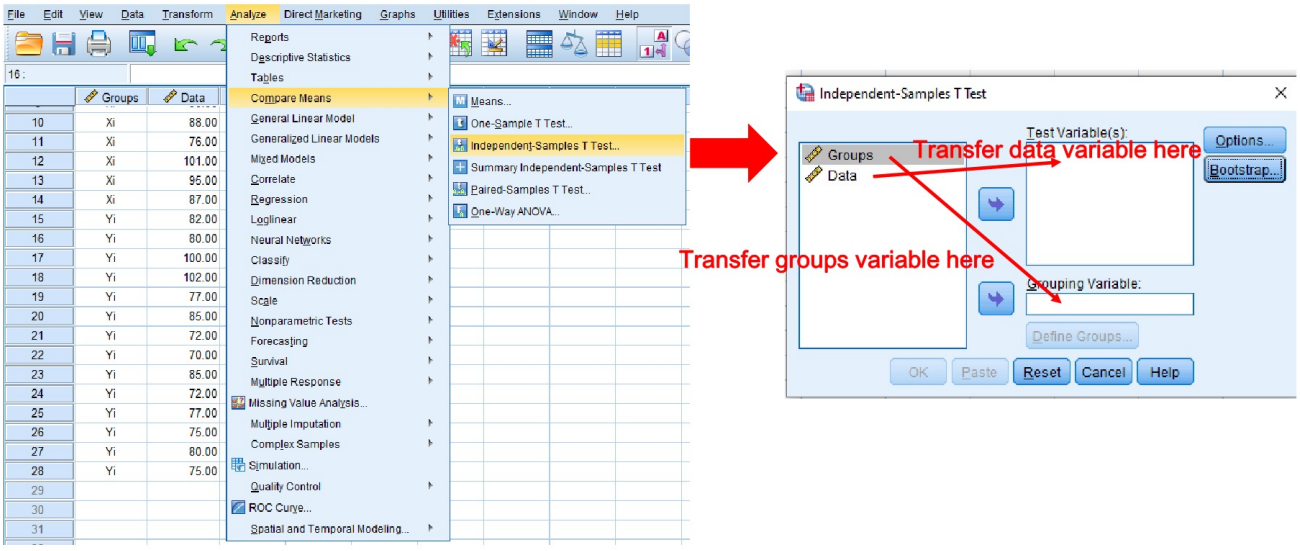
Three more observations not shown

Variable view

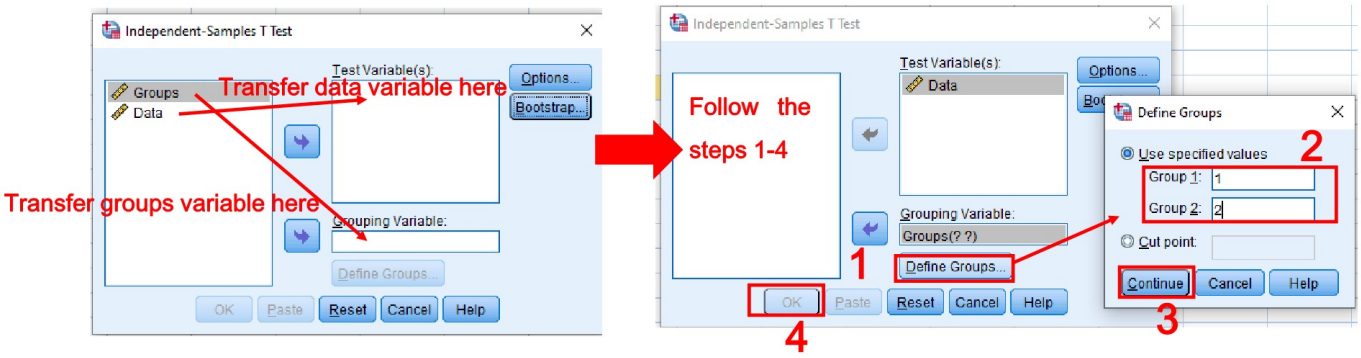




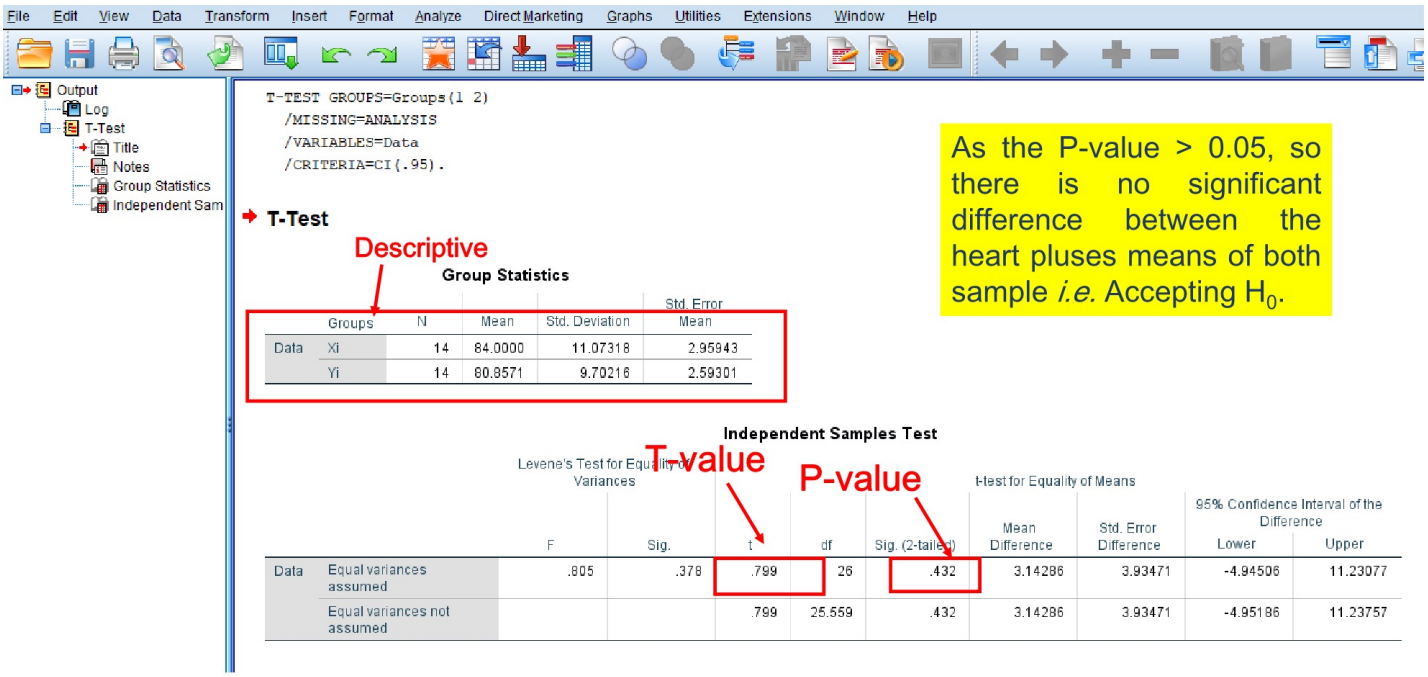
2. Go to Analyze >> Compare means >> Independent - Samples T test



2. The data should be transferred to the test variables box and the groups should be transferred to the grouping variables box then being identified correctly as 1 and 2



3. After clicking OK the results will appear in the output window


File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Extensions Window Help
Output
Log
T-Test
Title
Notes
Group Statistics
Independent Sam
T-TEST GROUPS=Groups(1 2)
/MISSING=ANALYSIS
/VARIABLES=Data
/CRITERIA=CI(.95).
T-Test
Descriptive
Group Statistics
Groups N Mean Std. Deviation Std. Error Mean
Data Xi 14 84.0000 11.07318 2.95943
Yi 14 80.8571 9.70216 2.59301
As the P-value > 0.05, so there is no significant difference between the heart pluses means of both sample i.e. Accepting H0.
Independent Samples Test
T-value
P-value
Levene's Test for Equality of Variances
t-test for Equality of Means
95% Confidence Interval of the Difference
F Sig. t df Sig. (2-tailed) Mean Difference Std. Error Difference Lower Upper
Data Equal variances assumed .805 .378 .799 26 .432 3.14286 3.93471 -4.94506 11.23077
Equal variances not assumed .799 25.559 .432 3.14286 3.93471 -4.95186 11.23757

How are T tests performed?

Paired samples T test :

This test is also performed by calculating T value and comparing it with the tabular T at a degrees of freedom = n-1 and a level of significance= 0.05. :

$$|t| = \frac{\bar{d}}{S_d/\sqrt{n}}$$

Whereas:

$\bar{d}$: It is a difference mean between both paired reads

S_d : standard deviation of differences.

Example : If we assume that we have a group of athletes and we want to test the effect of jogging for a 15 min on the mean of heart pulses, where X1 represents the number of heart pulses before exercise and X2 is the number of heart pulses after exercise. Assuming that the data is normally distributed.

X1i	70.00	77.00	90.00	100.00	85.00	80.00	72.00	65.00	90.00	88.00	76.00	101.00	95.00	87.00
X2i	100.00	102.00	100.00	102.00	110.00	105.00	101.00	95.00	105.00	103.00	100.00	106.00	100.00	105.00

الحل :

X1i	70.00	77.00	90.00	100.00	85.00	80.00	72.00	65.00	90.00	88.00	76.00	101.00	95.00	87.00
X2i	100.00	102.00	100.00	102.00	110.00	105.00	101.00	95.00	105.00	103.00	100.00	106.00	100.00	105.00
di=|X1-X2|	30.00	25.00	10.00	2.00	25.00	25.00	29.00	30.00	15.00	15.00	24.00	5.00	5.00	18.00

$$\bar{d} = \frac{\sum di}{n} = \frac{258}{14} = 18.43$$

$$S^2 = \frac{\sum {d_i}^2 - \frac{\sum (d_i)^2}{n}}{n-1} = 9.866$$

$$|t| = \frac{\bar{d}}{S_d/\sqrt{n}} = \frac{18.43}{9.866/\sqrt{14}} = 6.988$$

Since the tabulated T= 2.160 is less than calculated T=6.988, which means there is a significant difference between No. of pulses before and after exercise, so we'll reject H_0 and accept H_1.

Performing solution in SPSS :

1. Input data:

Data view

File Edit View Data Transform Analyze

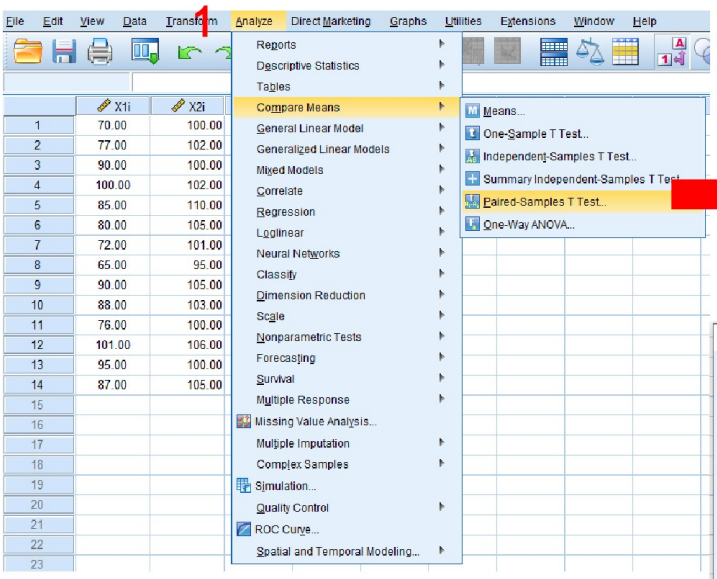
	X1i	X2i	var
1	70.00	100.00	
2	77.00	102.00	
3	90.00	100.00	
4	100.00	102.00	
5	85.00	110.00	
6	80.00	105.00	
7	72.00	101.00	
8	65.00	95.00	
9	90.00	105.00	
10	88.00	103.00	
11	76.00	100.00	
12	101.00	106.00	
13	95.00	100.00	
14	87.00	105.00	

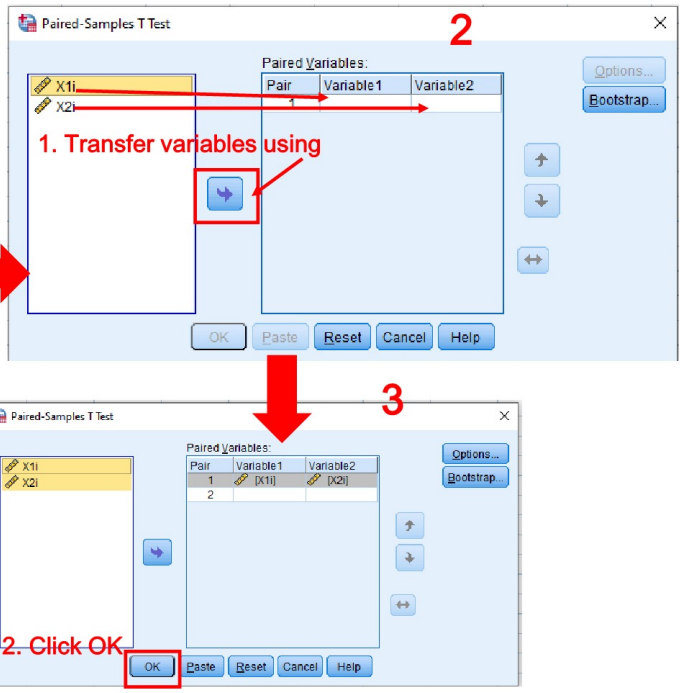
Variable view

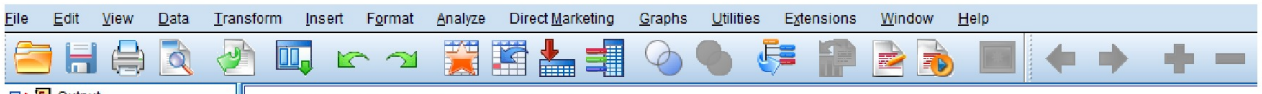
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help

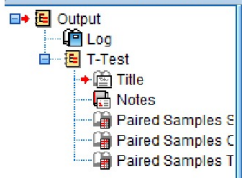
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	X1i	Numeric	8	2		None	None	8	Center	Scale	Input
2	X2i	Numeric	8	2		None	None	8	Right	Scale	Input

1. Go to analyze >> Compare means >> Paired










```
T-TEST PAIRS=X1i WITH X2i (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

T-Test

Descriptive

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	X1i	84.0000	14	11.07318	2.95943
	X2i	102.4286	14	3.63137	.97052

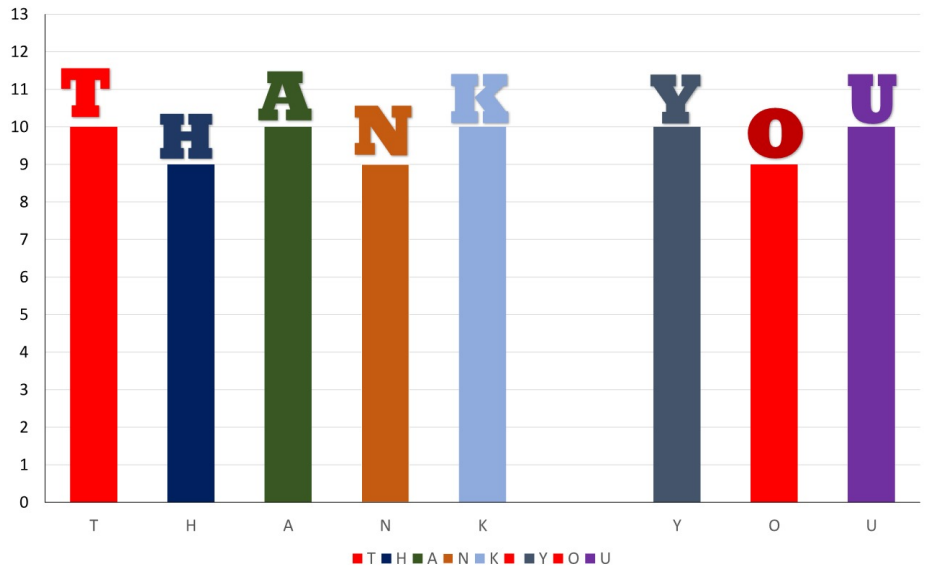
Paired Samples Correlations

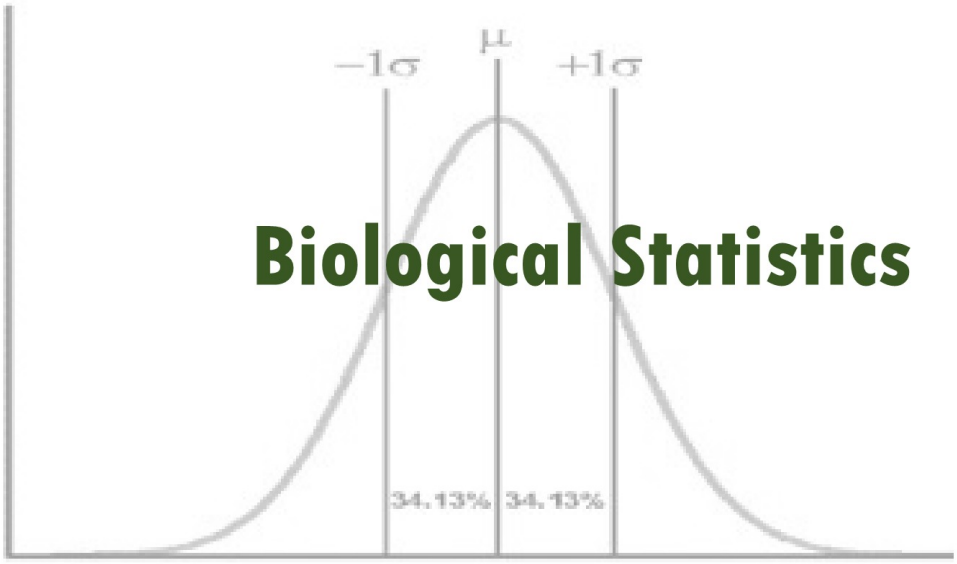
		N	Correlation	Sig.
Pair 1	X1i & X2i	14	.478	.084

Paired Samples Test

		Paired Differences					T-value	df	P-value
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			Sig. (2-tailed)
Pair 1	X1i - X2i	-18.42857	9.86614	2.63684	-24.12511	-12.73203	-6.989	13	.000

As the P-value < 0.01, so there is a significant difference between the heart pluses means before and after exercise *i.e.* rejecting H_0.


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Chi square or X^2 Test

It is one of the efficient tests of non-parametric data, it is suitable for dealing with discrete values such as integers and percentages, also it is used with parametric data. its idea based on a comparison between the observed values (the real values collected from the experiment) and the expected values (the values that the researcher expects to obtain if he applied the same experiment). This test does not require that the data be normally distributed.

When we need to use this test

The best use of this test is with data represented in the form of frequencies, where the comparison is made between the actual frequencies and the expected frequencies.

The cases of application

1. Goodness of fit test (for one sample).
2. Chi-square test of independence: to test for independence, to find the extent of the significance of the difference between two independent samples, and if this difference significant or not?

How to calculate X^2 ?

It is a sum of (the square of the differences between the observed and expected values divided by the expected values). The main point of this test is how to calculate the expected values.

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Whereas:
O: The observed values.
E: The expected values.

Chi square for goodness of fit

There are two cases of estimating the expected values then calculating the chi-square value:

1. Goodness of fit with a particular distribution.
2. Goodness of fit with standard ratio.

Example on goodness of fit with particular distribution

An examination was conducted for a group of students (50 students) to evaluate their level of in a particular subject. The number of students within failed class was 5 students, and the students in the class (satisfy - medium) 15 students, the number of students within the class (good - very good) was 20 students, and the excellence grade students were ten, does the distribution of students' levels in this sample follow a normal distribution or not?

The solution :

Classes	O	$E = \frac{\sum O}{n}$	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
Excellent	10	12.5	- 2.5	6.25	0.5
Good – Very good	20	12.5	7.5	56.25	4.5
Satisfied - Medium	15	12.5	2.5	6.25	0.5
Failed	5	12.5	- 7.5	56.25	4.5
Sum	50		0		$X^2 = 10$

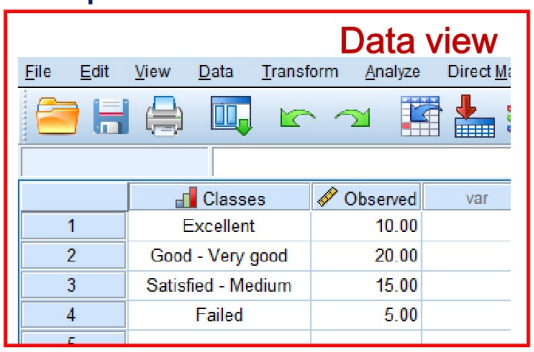
Degrees of freedom = No. of Classes -1 >> 4 - 1 = 3

Based on that, the tabular X^2 within significance level = 0.05 and *df* = 3 will be 7.814. Since the calculated value of chi is greater than the tabular one, this indicates that there are significant differences between the sample distribution and the normal distribution of the community, and that the levels of students' were not within the normal limits.

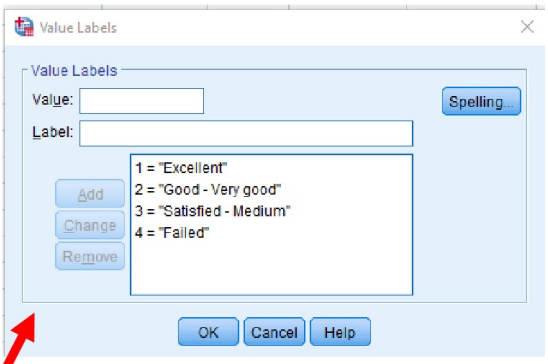
The solution in SPSS

1. Input data

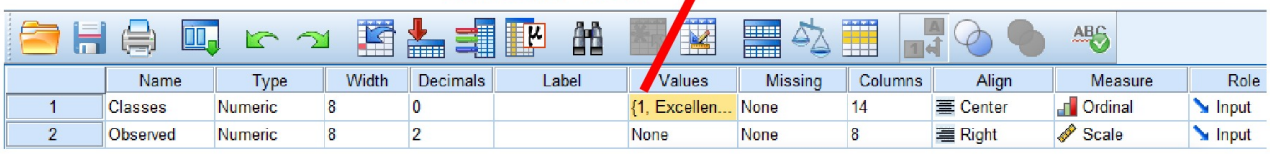
Data view



	Classes	Observed	var
1	Excellent	10.00	
2	Good - Very good	20.00	
3	Satisfied - Medium	15.00	
4	Failed	5.00	



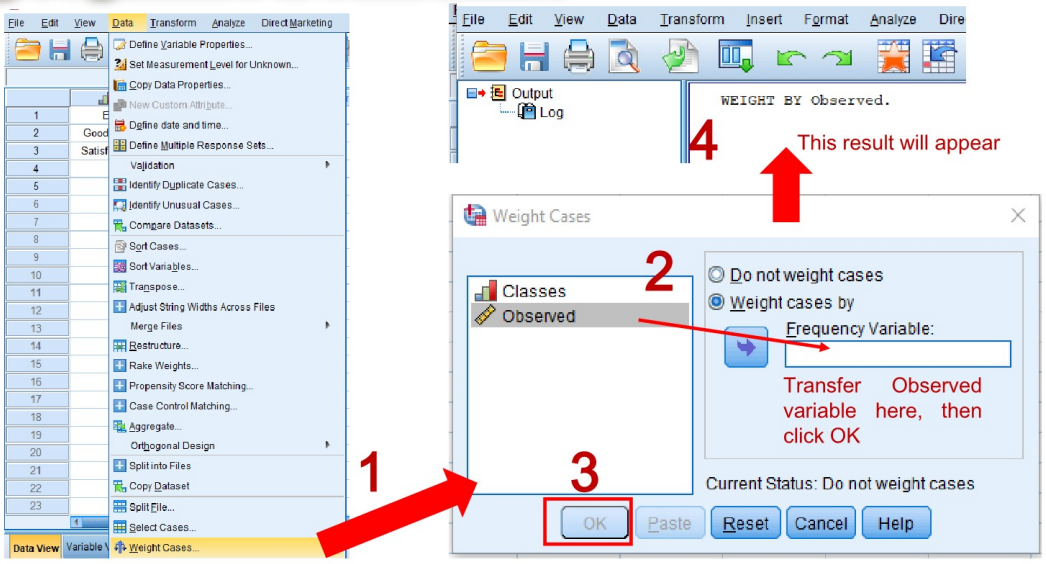
Variable view



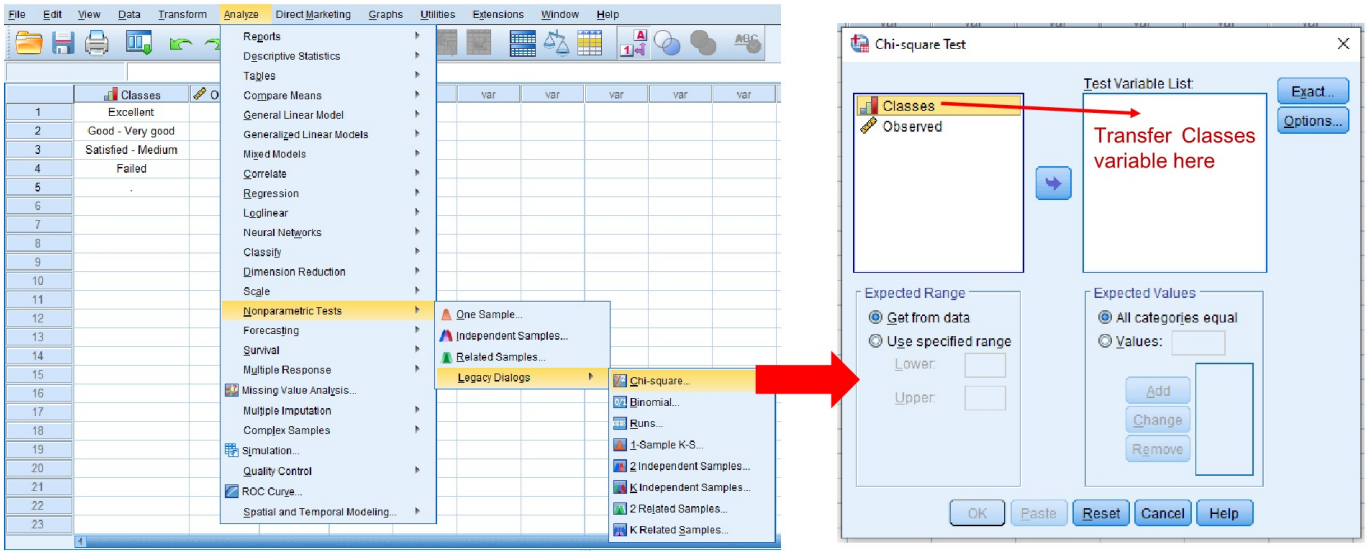
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Classes	Numeric	8	0		{1, Excellen...	None	14	Center	Ordinal	Input
2	Observed	Numeric	8	2		None	None	8	Right	Scale	Input

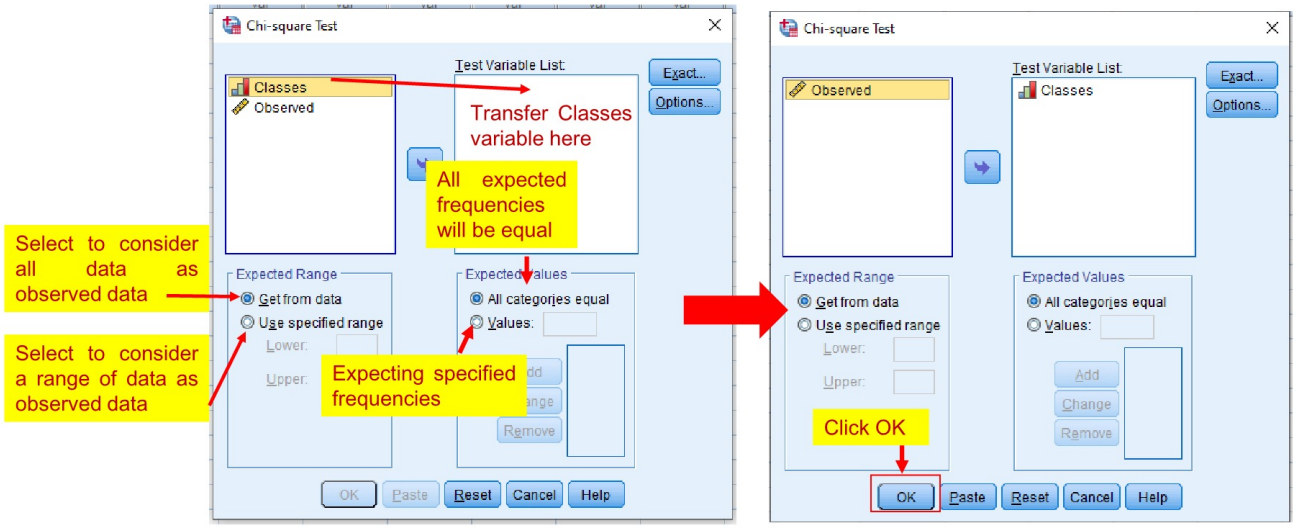
Important note : When the data being in the form of frequencies, we should weight the classes by the frequencies to tell SPSS how many frequencies of each class .

2. Go to Data >> Weight classes.



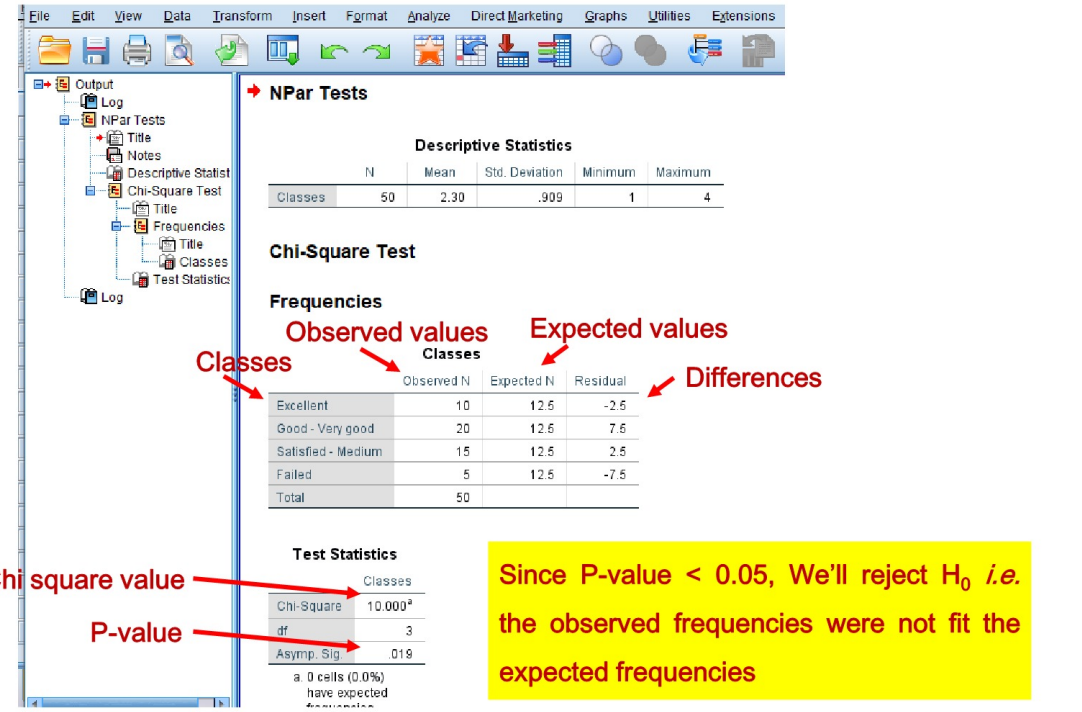
3. Go to Analyze >> Nonparametric Tests >> Legacy dialogs >> Chi-square





In our example we expecting all classes to being in equal frequencies and considering all observations set as observed data

4. The results will appear in output window



Since P-value < 0.05, We'll reject H_0 *i.e.* the observed frequencies were not fit the expected frequencies

Example about goodness of fit with a particular ratio

A particular college includes 230 students and 20 teachers, How to determine whether the student-teacher ratio is within the standard ratio (10:1) or not?

The solution:

The total number of professors and students in the college is 230 + 20 = 250

The Total of the standard ratio is 1 + 10 = 11

That is, the ratio of teachers is 1 / 11 and the ratio of students is 10 / 11

The expected number of teachers is $250 \times \frac{1}{11} = 23$

The expected number of students is $250 \times \frac{10}{11} = 227$

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(230-227)^2}{227} + \frac{(20-23)^2}{23} = 0.43$$

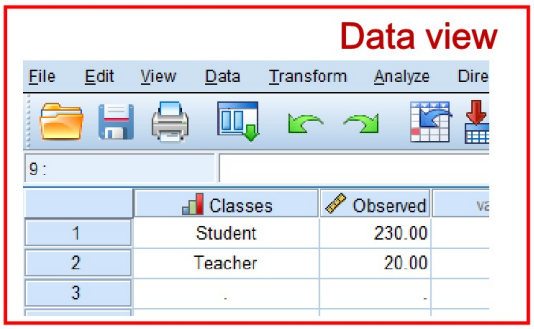
$$df = n - 1 = 2 - 1 = 1$$

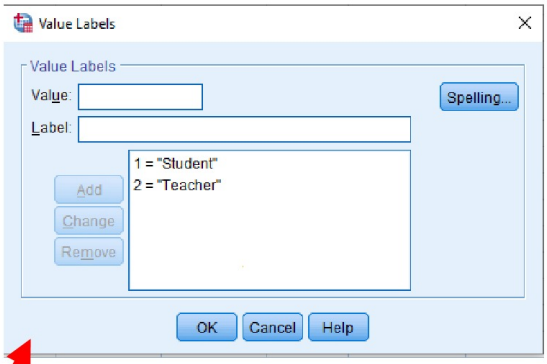
X^2 at 0.05 alpha level 3.841. since the calculated value of X2 (0.43) is less than the tabular value, we conclude that there are no significant differences between the ratio of students to teachers in this college and the standard ratio.

Performing solution using SPSS

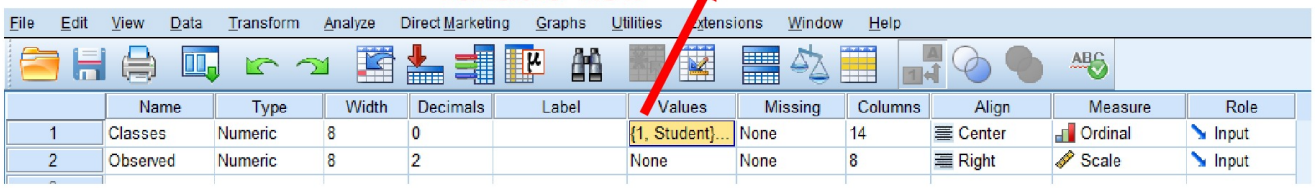
1. Input data

Data view

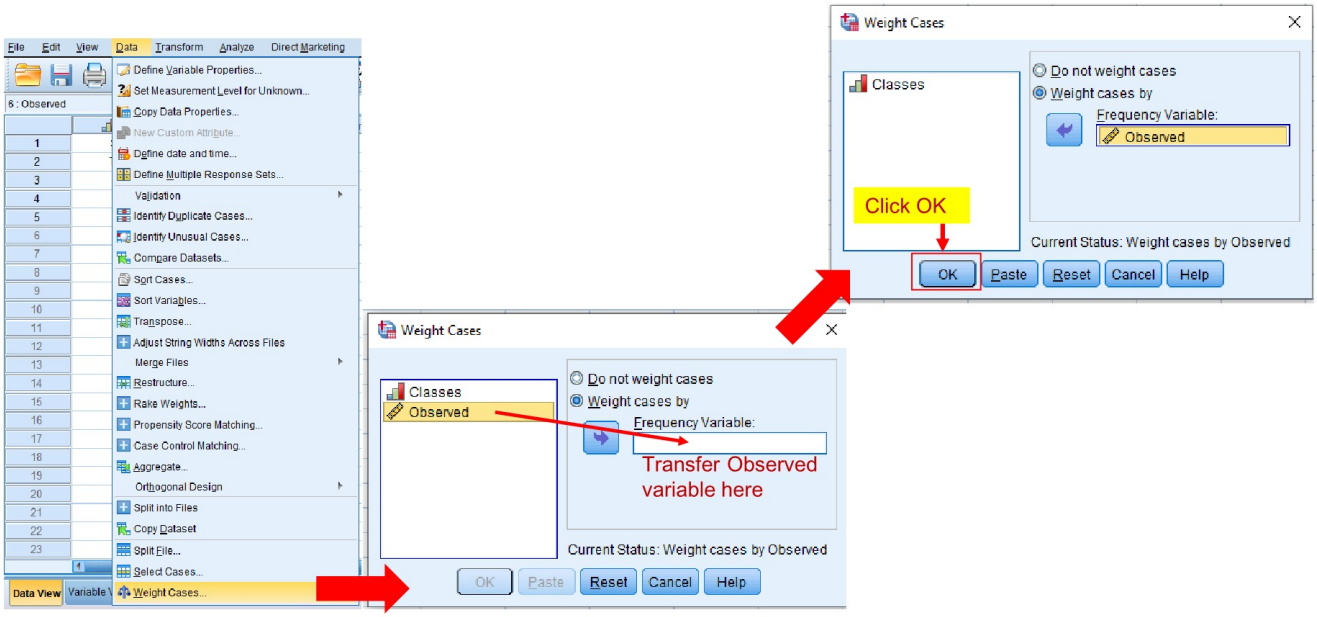




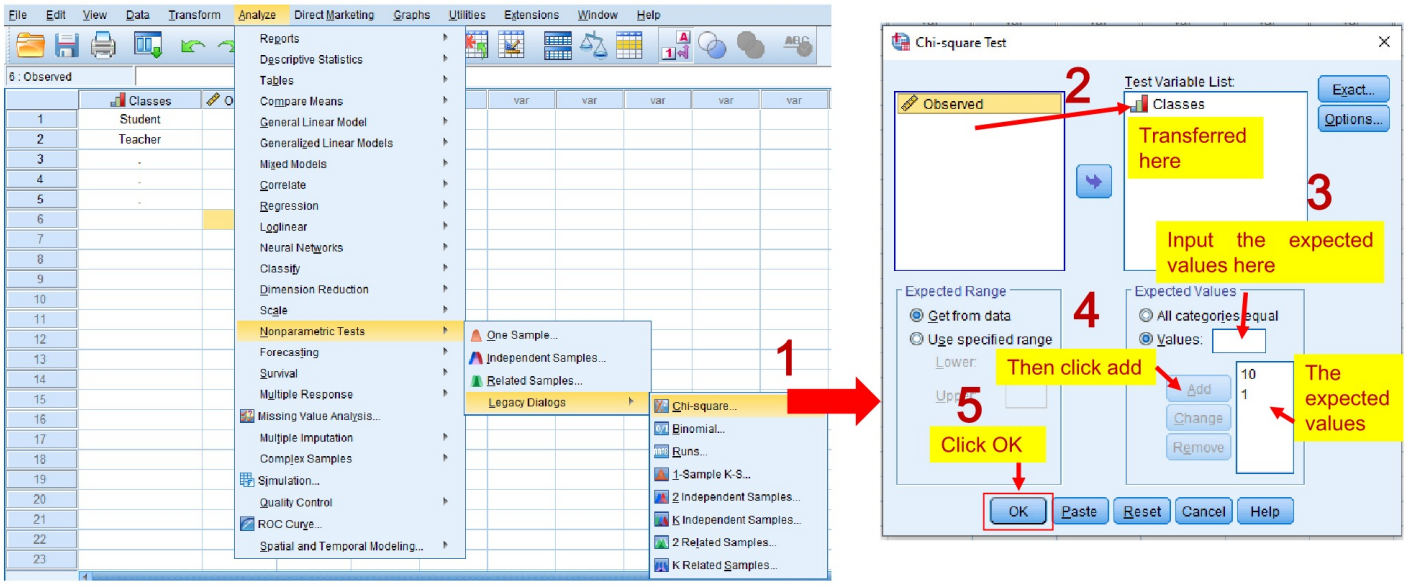
Variable view



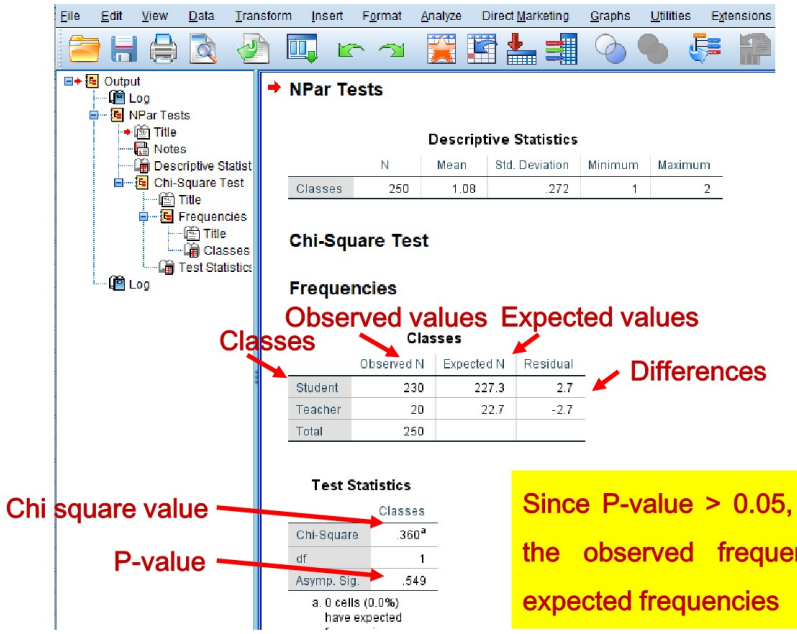
2. Weight classes by going to Data >> Weight cases



3. Go to Analyze >> Nonparametric tests >> Legacy dialogs >> Chi-Sequare



4. The results will appear in output window



Since P-value > 0.05, We'll accept H_0 *i.e.* the observed frequencies were fit the expected frequencies

Other example about goodness of fit with particular distribution

A study was conducted to determine the sensitivity of a sample of 400 people from a community of a particular city to infection with seasonal influenza, the results showed that 75% of them are sensitive to infection, 15% of them are resistant to infection, while 10% were hypersensitive. To what extent do these percentages match the standard percentages of the population, which are 80% sensitive, 12% resistant, and 8% hypersensitive?

The solution : Calculating the observed and expected frequencies:

The observed frequency values (Sample)

$$O\ 75\% = \frac{75}{100} \times 400 = 300$$
$$O\ 15\% = \frac{15}{100} \times 400 = 60$$
$$O\ 10\% = \frac{10}{100} \times 400 = 40$$

The expected frequency values (Population)

$$E\ 80\% = \frac{80}{100} \times 400 = 320$$
$$E\ 12\% = \frac{12}{100} \times 400 = 48$$
$$E\ 8\% = \frac{8}{100} \times 400 = 32$$

Classes	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Sensitive	300	320	-20	400	1.25
Resistant	60	12	48	144	3
hypersensitive	40	8	32	64	2
					$X^2 = 6.25$

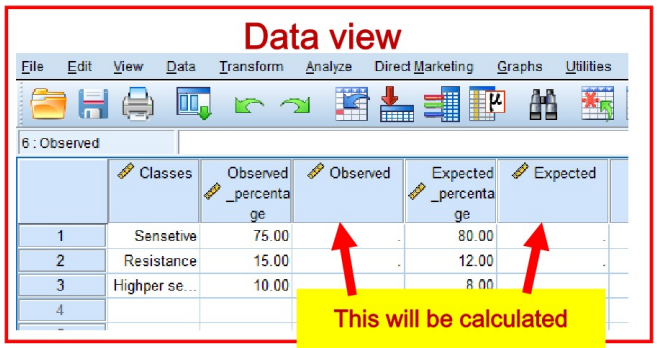
$$df = n - 1 = 3 - 1 = 2$$

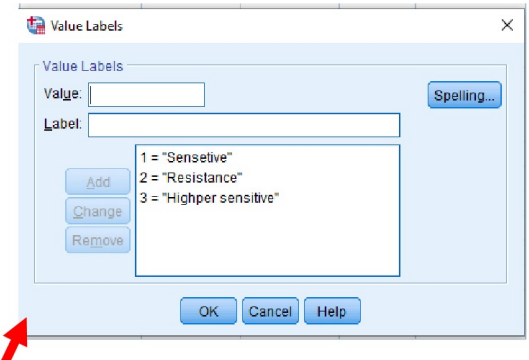
We'll extract the tabular X^2 at degrees of freedom = 2 and a level of significance = 0.05, which equals 5.99. Since the tabular value of X^2 is less than the calculated one, the infection sensitivity of the sample members are significantly different from the standard ratios, and thus we reject the null hypothesis and accept the alternative.

Performing solution using SPSS

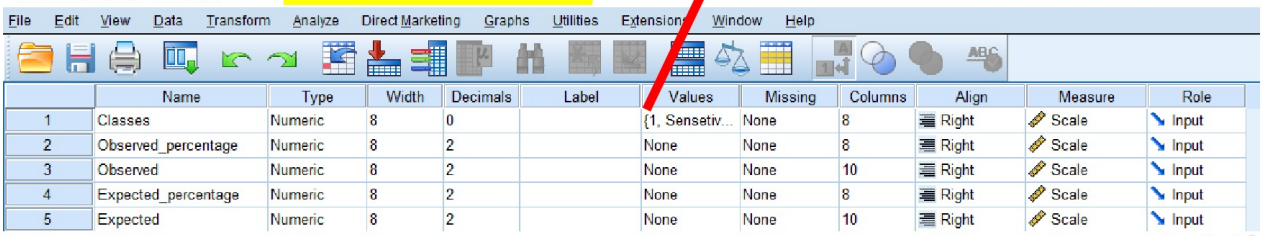
1. Input data

Data view



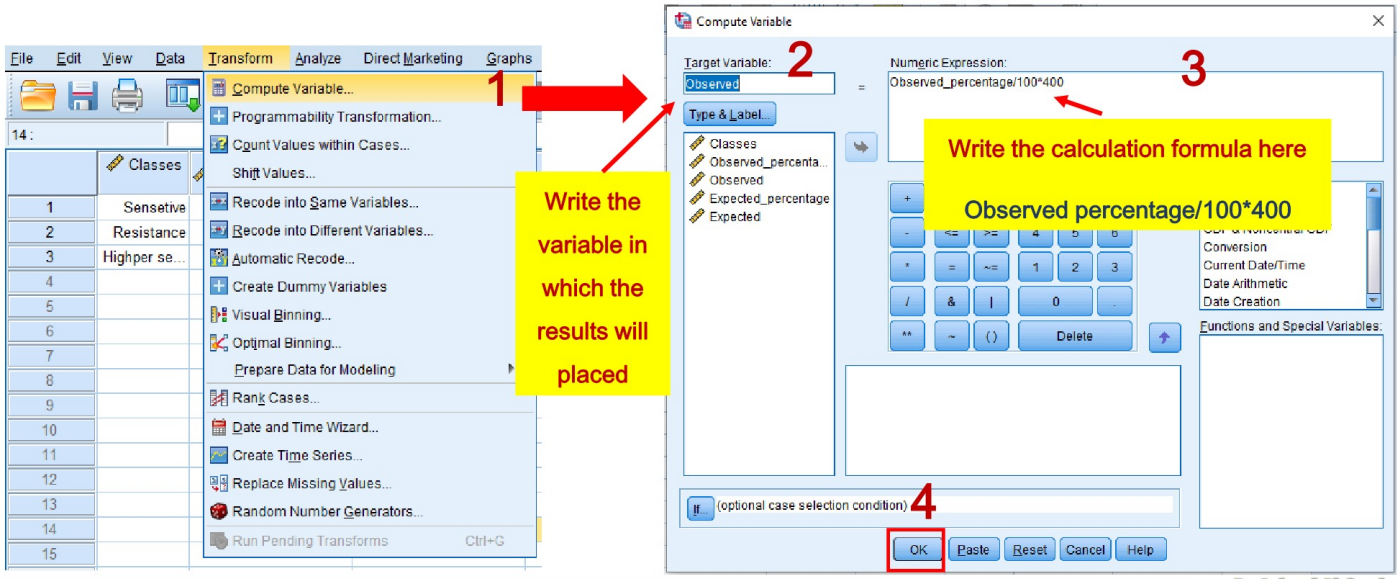


Variable view

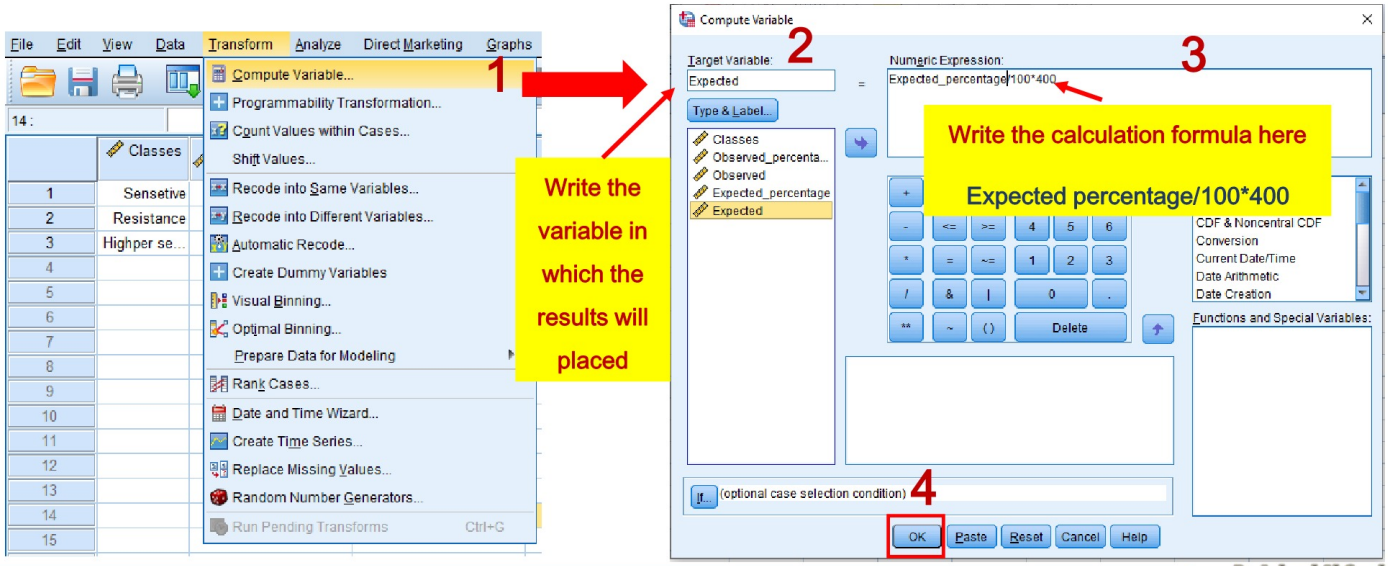


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Classes	Numeric	8	0		{1, Sensetiv...	None	8	Right	Scale	Input
2	Observed_percentage	Numeric	8	2		None	None	8	Right	Scale	Input
3	Observed	Numeric	8	2		None	None	10	Right	Scale	Input
4	Expected_percentage	Numeric	8	2		None	None	8	Right	Scale	Input
5	Expected	Numeric	8	2		None	None	10	Right	Scale	Input

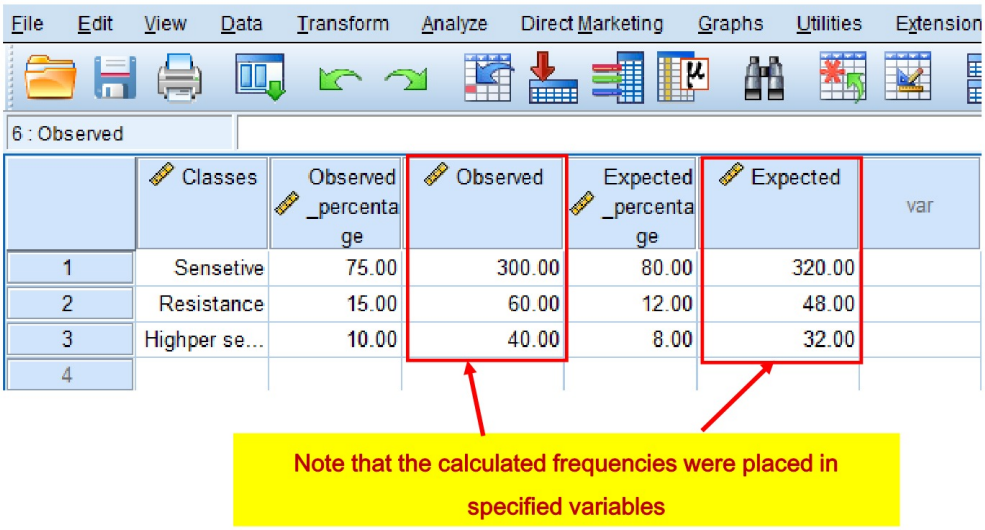
2. Convert percentages to a real frequencies of observed and expected values: Go to Transform >> Compute Variables:



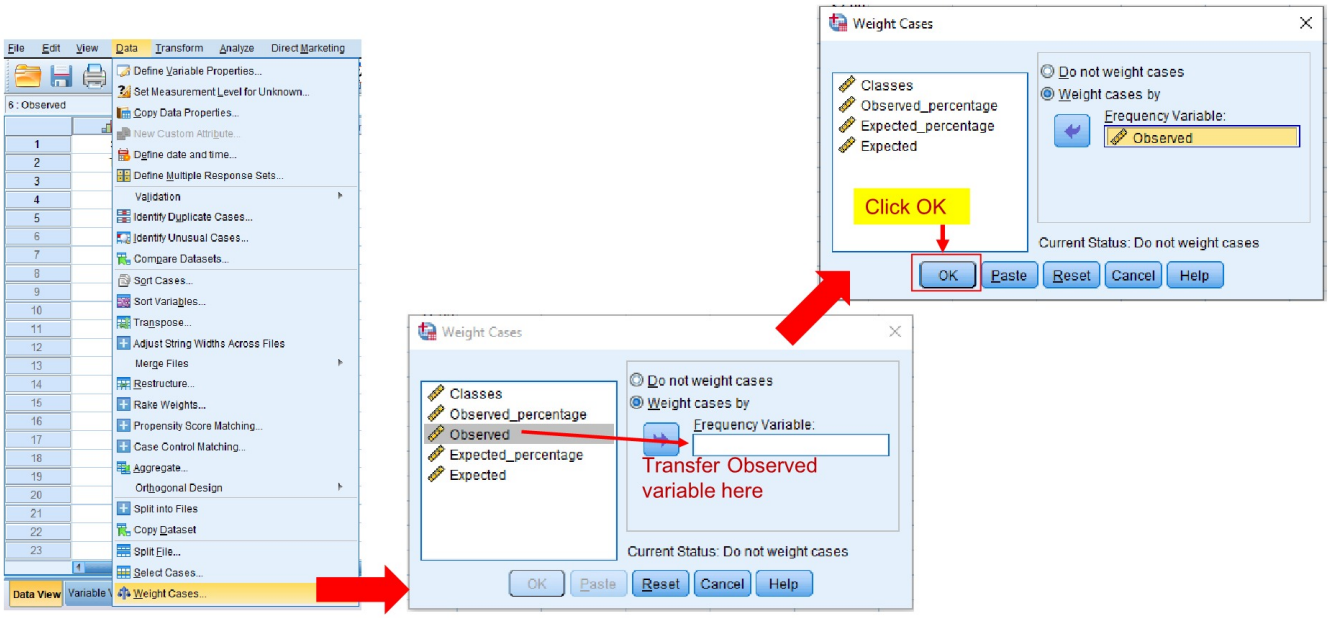
2. Convert percentages to a real frequencies of observed and expected values: Go to Transform >> Compute Variables:



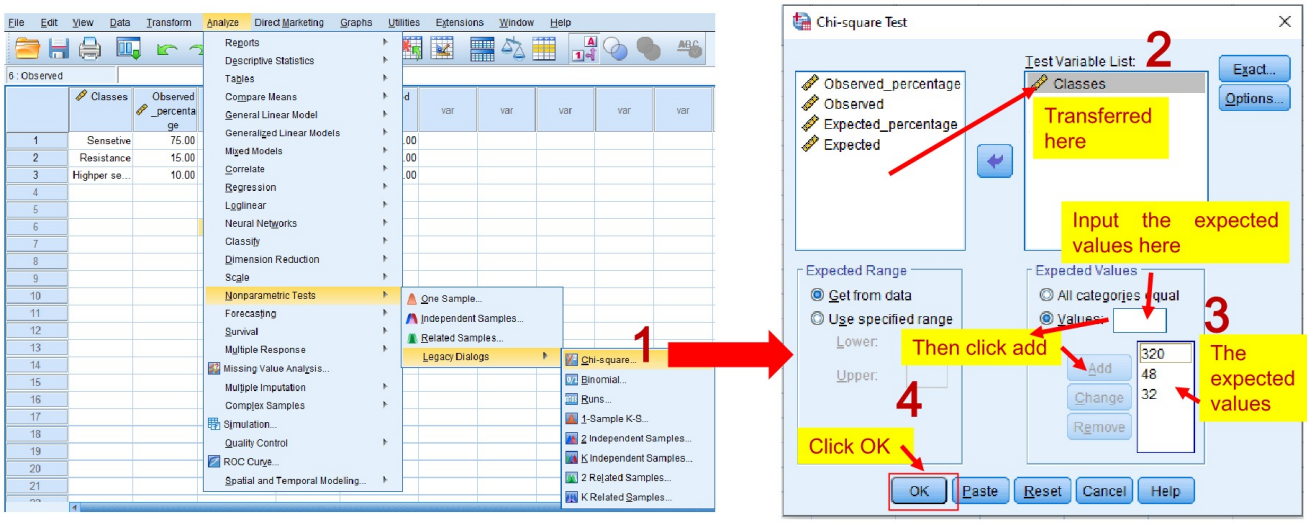
3. The values were calculated and the results were placed in the specified variables



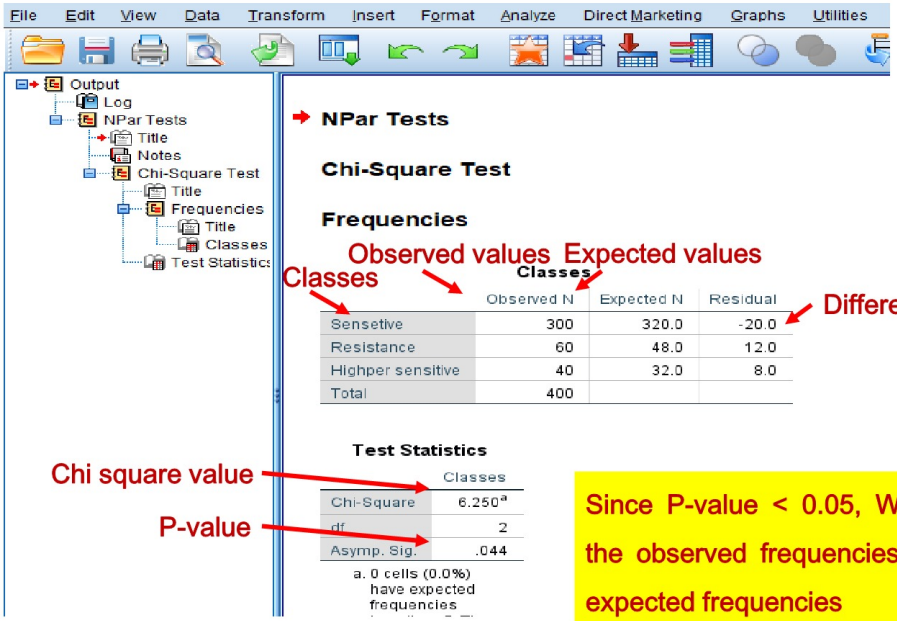
4. Weight classes by going to Data >> Weight cases



5. Go to Analyze >> Nonparametric Tests >> Legacy dialogs >> Chi-square



6. The results will appear in output window



Since P-value < 0.05, We'll reject H_0 *i.e.* the observed frequencies were not fit the expected frequencies

Chi - square test of independence

This test is used to determine whether the variables are independent or not (*i.e.*, is there a correlation between them) basing on a certain factor?

Example :

A researchers used four types of antibiotics, in three doses for each antibiotic, and recorded the percentage of mortality to find out the extent of their independence in the effect depending on the dose, *i.e.*, are there significant differences between them depending on the dose?

Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %
250 mg	50	55	43	30
500 mg	77	80	65	60
750 mg	90	85	72	75

The solution

Firstly we'll calculate the probabilistic position P for each antibiotic and each dose separately by the following equation: $P = \frac{Sum}{Total}$

Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %	Sum	P
250 mg	50	55	43	30	178	0.23
500 mg	77	80	65	60	282	0.36
750 mg	90	85	72	75	322	0.41
Sum	217.00	220.00	180.00	165.00	Total=782	
P	0.28	0.28	0.23	0.21		

$P = \frac{178}{782}$

The second step is to calculate the internal probabilistic position by multiplying the probability value of each drug by the probability value of each dose:

For example, the probabilistic position of the first drug at the first dose is:

0.28 x 0.23 = 0.06

Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %
250 mg	0.06	0.06	0.05	0.05
500 mg	0.10	0.10	0.08	0.08
750 mg	0.11	0.12	0.09	0.09

The third step is to calculate the expected values of mortality percentages by multiplying the value of the internal probabilistic position by the grand total:

For example, the expected value of the first drug is: 0.6 x 782 = 49.39

Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %
250 mg	49.39	50.08	40.97	37.56
500 mg	78.25	79.34	64.91	59.50
750 mg	89.35	90.59	74.12	67.94

Dose	Drug1 %		Drug2 %		Drug3 %		Drug4 %	
	O	E	O	E	O	E	O	E
250 mg	50	49.39	55	50.08	43	40.97	30	37.56
500 mg	77	78.25	80	79.34	65	64.91	60	59.50
750 mg	90	89.35	85	90.59	72	74.12	75	67.94

$(O-E)^2$

Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %
250 mg	0.37	24.21	4.12	57.15
500 mg	1.56	0.44	0.01	0.25
750 mg	0.42	31.25	4.49	49.84

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$= \sum \frac{0.37}{49.39} + \dots\dots\dots\dots + \frac{49.84}{67.94} = 3.29$$

$df = (Columns - 1) \times (Rows - 1) \rightarrow = (4-1) \times (3-1) = 6$

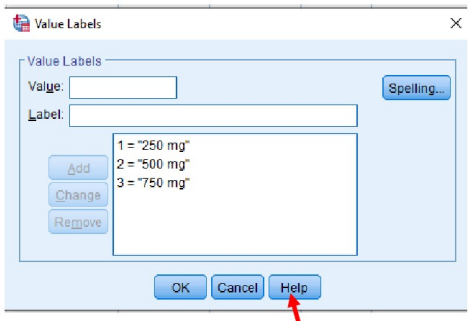
Since the tabular value of X^2 at degrees of freedom = 6 and the significance level =0.05 is 12.59 greater than the calculated one, this means that there are no significant differences between the drugs depending on the dose, meaning that the variables are independent.

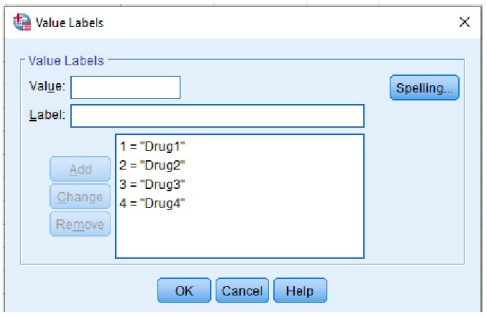
Performing solution using SPSS

1. Input data

Data view

	Drug	Dose	Data
1	Drug1	250 mg	50.00
2	Drug1	500 mg	77.00
3	Drug1	750 mg	90.00
4	Drug2	250 mg	55.00
5	Drug2	500 mg	80.00
6	Drug2	750 mg	85.00
7	Drug3	250 mg	43.00
8	Drug3	500 mg	65.00
9	Drug3	750 mg	72.00
10	Drug4	250 mg	30.00
11	Drug4	500 mg	60.00
12	Drug4	750 mg	75.00

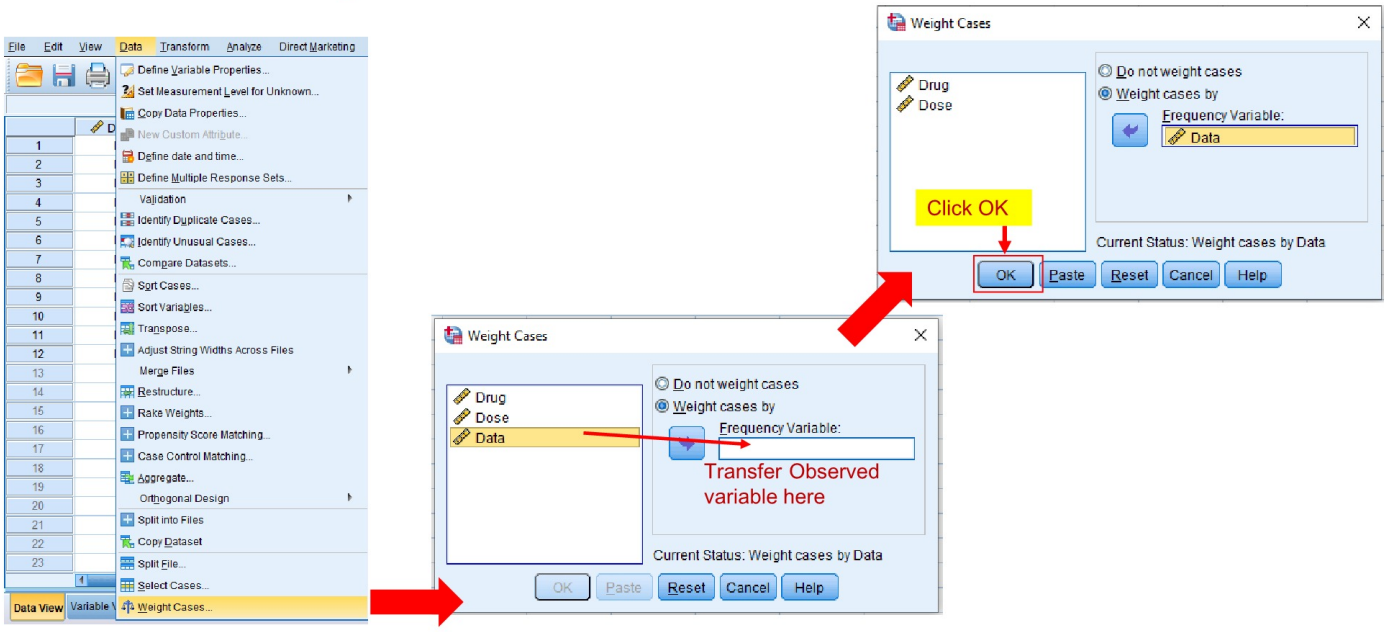




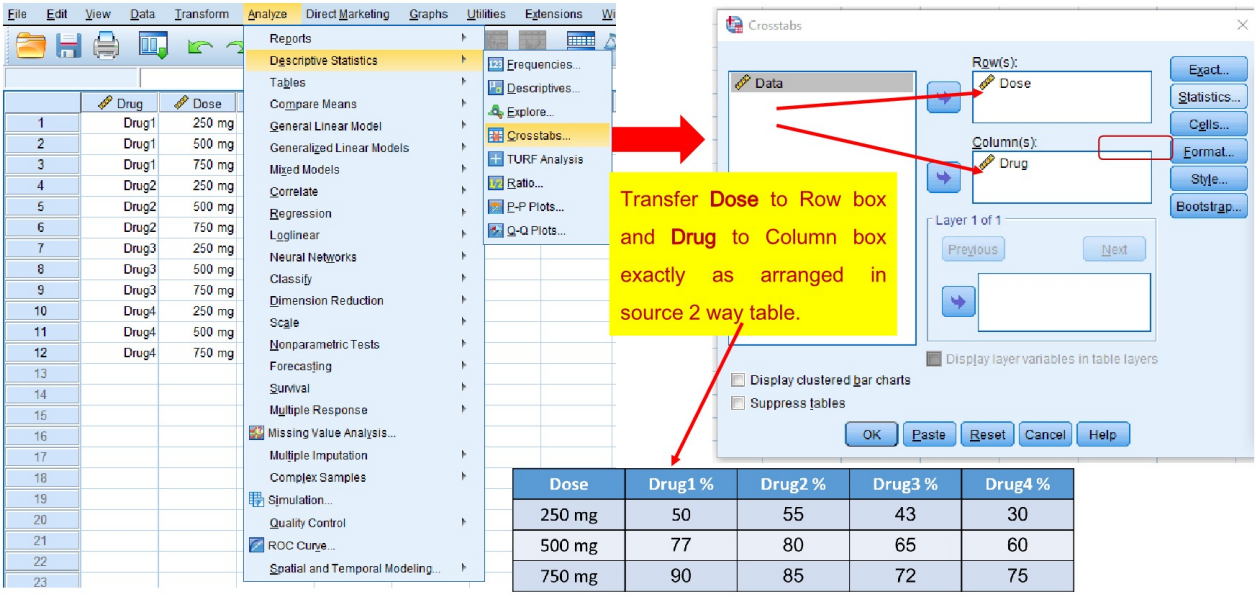
Variable view

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Drug	Numeric	8	0		{1, Drug1}...	None	8	Right	Scale	Input
2	Dose	Numeric	8	0		{1, 250 mg}...	None	8	Right	Scale	Input
3	Data	Numeric	8	2		None	None	8	Right	Scale	Input

2. Go to Data >> Weight cases

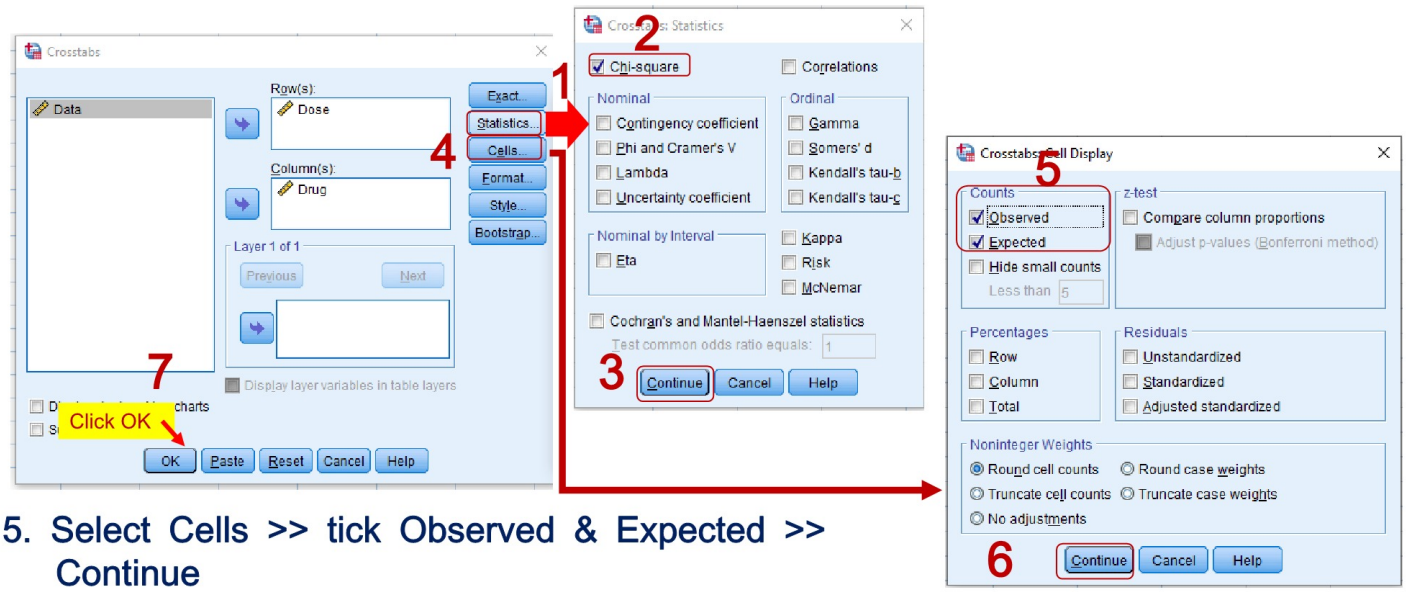


3. Go to Analyze >> Descriptive statistics >> Crosstabs



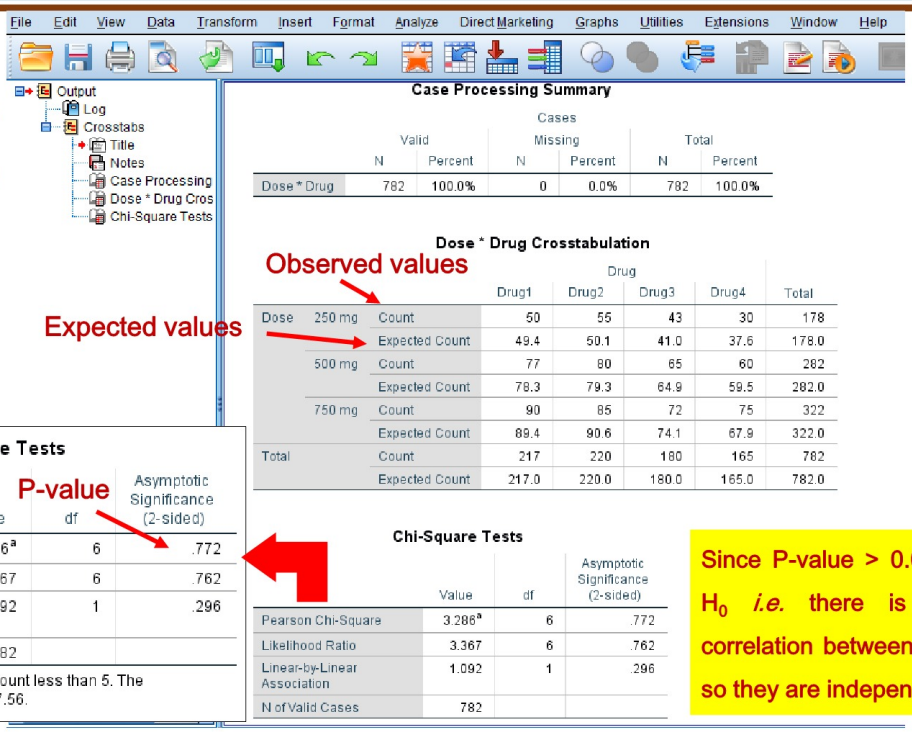
Dose	Drug1 %	Drug2 %	Drug3 %	Drug4 %
250 mg	50	55	43	30
500 mg	77	80	65	60
750 mg	90	85	72	75

4. Select Statistics >> Chi-square >> Continue



5. Select Cells >> tick Observed & Expected >> Continue

5. The results will appear in output window



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Dose * Drug	782	100.0%	0	0.0%	782	100.0%

Dose * Drug Crosstabulation

Observed values

Expected values

			Drug				
			Drug1	Drug2	Drug3	Drug4	Total
Dose	250 mg	Count	50	55	43	30	178
		Expected Count	49.4	50.1	41.0	37.6	178.0
	500 mg	Count	77	80	65	60	282
		Expected Count	78.3	79.3	64.9	59.5	282.0
	750 mg	Count	90	85	72	75	322
		Expected Count	89.4	90.6	74.1	67.9	322.0
Total		Count	217	220	180	165	782
		Expected Count	217.0	220.0	180.0	165.0	782.0

Chi-Square Tests

Chi square value

P-value

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.286[a]	6	.772
Likelihood Ratio	3.367	6	.762
Linear-by-Linear Association	1.092	1	.296
N of Valid Cases	782		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 37.56.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.286[a]	6	.772
Likelihood Ratio	3.367	6	.762
Linear-by-Linear Association	1.092	1	.296
N of Valid Cases	782		

Since P-value > 0.05, We'll accept H_0 *i.e.* there is no significant correlation between drug and dose, so they are independent variables

Other case of Chi – square test of independence

This test is also used in experiments where the results are in the form of a yes or no answer, male and female, win and lose, agree or disagree, so that we have a two-way matrix (two rows representing the two samples and two columns or several columns representing the traits).

Example :

A researchers wanted to test the effect of a person's gender on his academic achievement, that is, does a person's academic achievement depend on his gender, in other words, are the two variables independent (*i.e.*, a person's gender and academic achievement are independent or correlated with each other). The researcher used a sample of 395 males and females and conducted a questionnaire to find out their academic achievement. The result of the questionnaire was as follows:

	Secondary school	Bachelor	Master	PhD
Male	60	54	46	41
Female	40	44	53	57

The solution :

To simplify calculations we'll codding the cells

	Secondary school	Bachelor	Master	PhD	FiH
Male	60 A	54 B	46 C	41 D	201
Female	40 E	44 F	53 G	57 H	194
FiV	100	98	99	98	Fi= 395

The values we denoted from A to H are the observed values (O) and we have to calculate the expected values E for each one of them, using the following equation:

$$E = \frac{\sum FiV \times \sum FiH}{\sum Fi}$$

$$E_A = \frac{201 \times 100}{395} = 50.88$$

$$E_B = \frac{201 \times 98}{395} = 49.86$$

Do the same for other values

Whereas:

$\sum FiH$: The sum of row observation.

$\sum FiV$: The sum of column observation.

$\sum Fi$: The total sum.

We'll get the following expecting values

	Secondary school	Bachelor	Master	PhD
ذكور	50.89	49.87	50.38	49.87
أناث	49.11	48.13	48.62	48.13

After that we'll calculate the X^2 value:

	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	60	50.89	9.11	82.99	1.63
B	40	49.11	-9.11	82.99	1.69
C	54	49.87	4.13	17.06	0.34
D	44	48.13	-4.13	17.06	0.35
E	46	50.38	-4.38	19.18	0.38
F	53	48.62	4.38	19.18	0.39
G	41	49.87	-8.87	78.68	1.58
H	57	48.13	8.87	78.68	1.63
					X^2= **8**

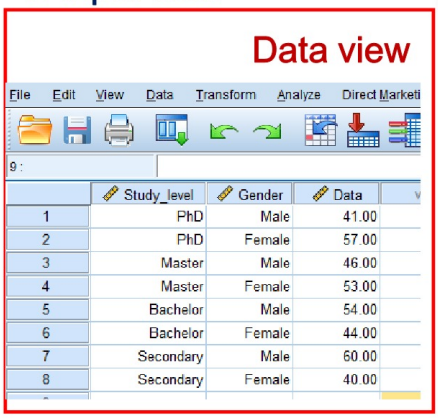
$$df = (4-1) \times (2-1) = 3$$

Then extract the tabular X^2 at degrees of freedom = 3 and a significance level of 0.05, which is equal to 7.81, and since it is less than the calculated one, *i.e.* we'll reject the H_0 and accept the alternative, which means, there is an effect of the individuals' gender on their academic achievement.

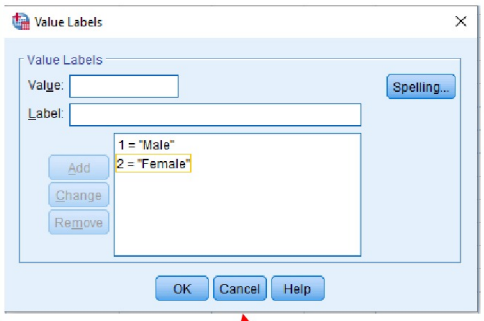
1. Input data

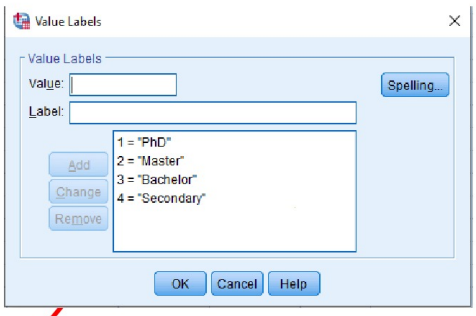
Performing solution using SPSS

Data view

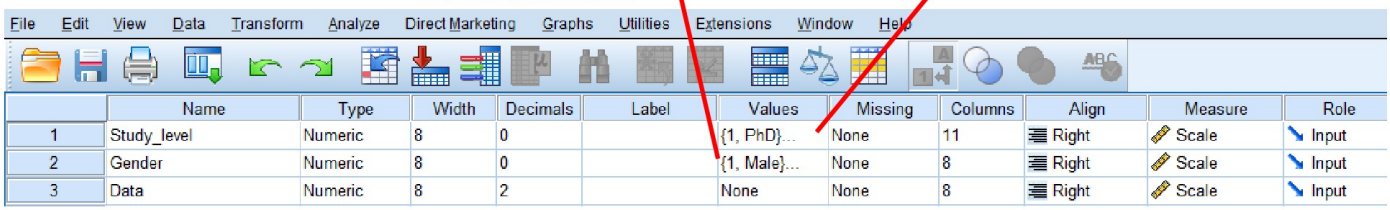


	Study_level	Gender	Data
1	PhD	Male	41.00
2	PhD	Female	57.00
3	Master	Male	46.00
4	Master	Female	53.00
5	Bachelor	Male	54.00
6	Bachelor	Female	44.00
7	Secondary	Male	60.00
8	Secondary	Female	40.00



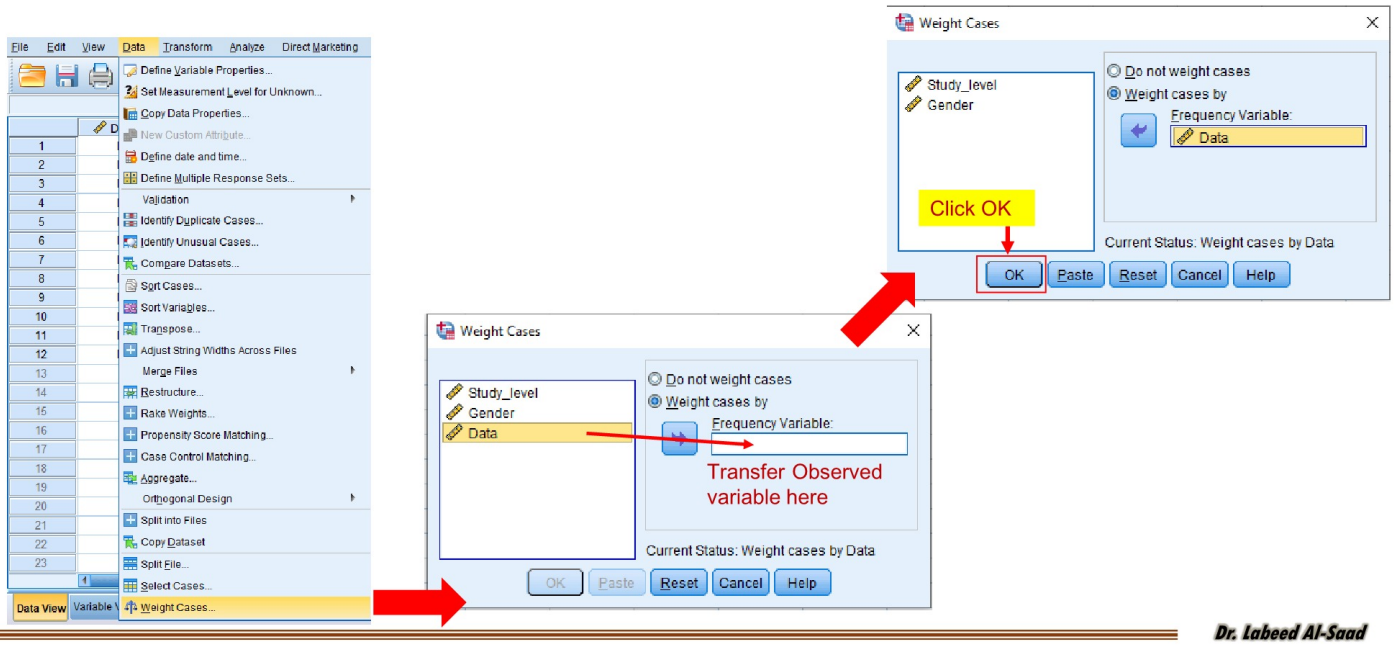


Variable view

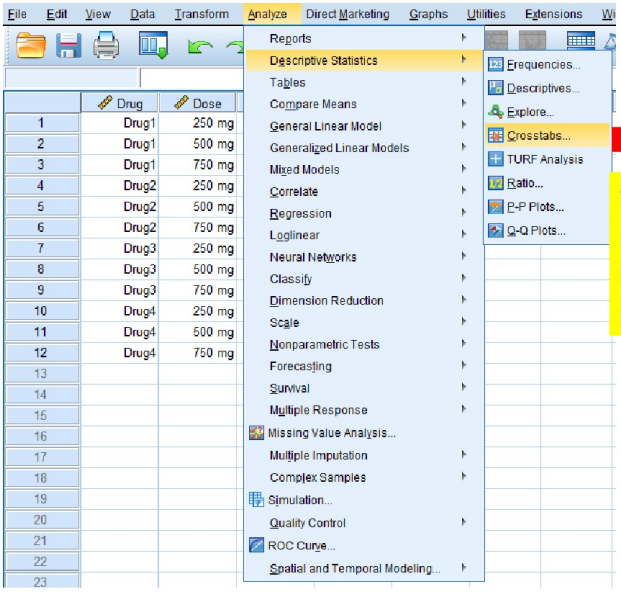


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Study_level	Numeric	8	0		{1, PhD}...	None	11	Right	Scale	Input
2	Gender	Numeric	8	0		{1, Male}...	None	8	Right	Scale	Input
3	Data	Numeric	8	2		None	None	8	Right	Scale	Input

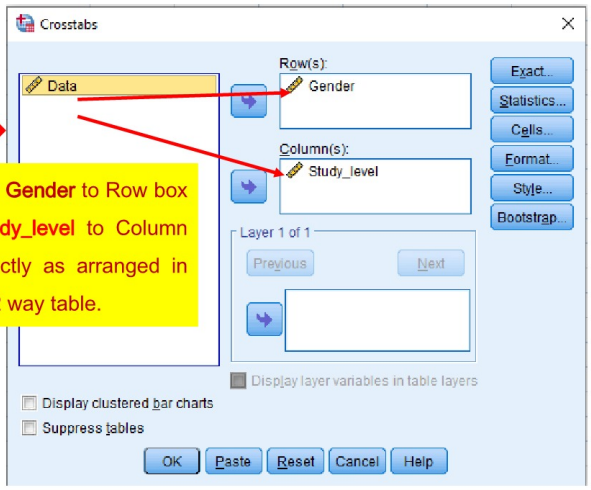
2. Go to Data >> Weight cases



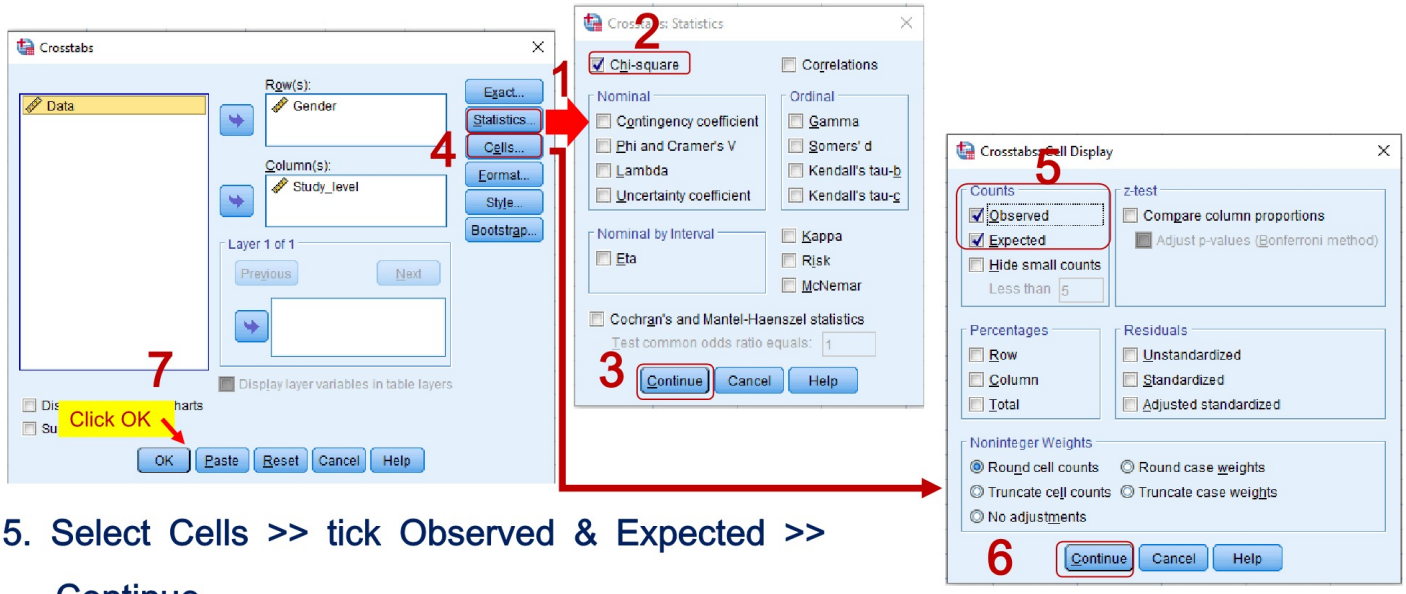
3. Go to Analyze >> Descriptive statistics >> Crosstabs



Transfer **Gender** to Row box and **Study_level** to Column box exactly as arranged in source 2 way table.

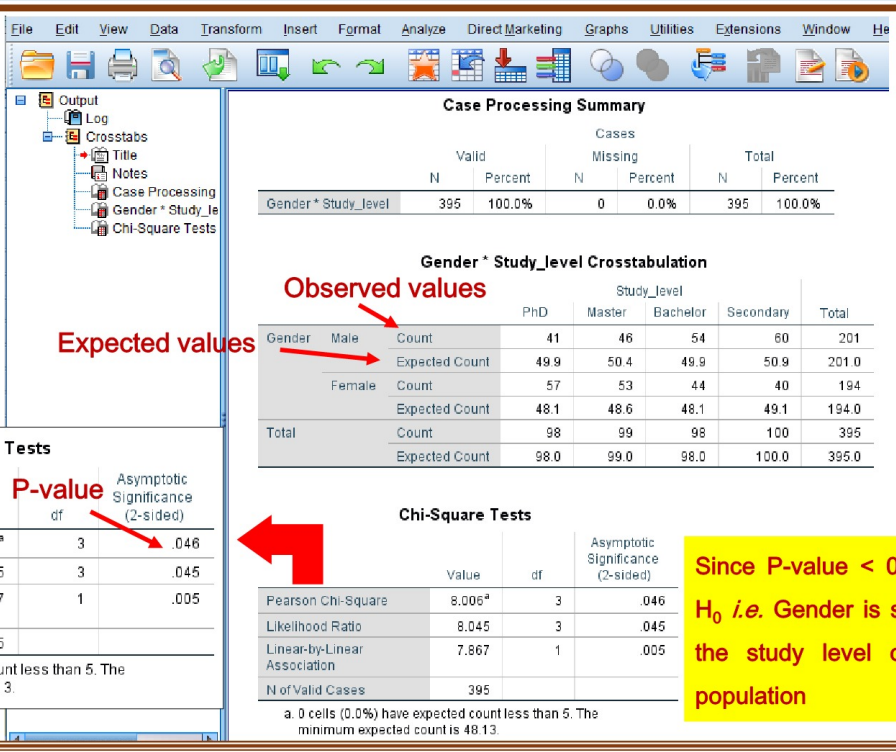


4. Select Statistics >> Chi-square >> Continue



5. Select Cells >> tick Observed & Expected >> Continue

5. The results will appear in output window



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Study_level	395	100.0%	0	0.0%	395	100.0%

Gender * Study_level Crosstabulation

Observed values

Expected values

			Study_level				
			PhD	Master	Bachelor	Secondary	Total
Gender	Male	Count	41	46	54	60	201
		Expected Count	49.9	50.4	49.9	50.9	201.0
	Female	Count	57	53	44	40	194
		Expected Count	48.1	48.6	48.1	49.1	194.0
Total		Count	98	99	98	100	395
		Expected Count	98.0	99.0	98.0	100.0	395.0

Chi-Square Tests

Chi square value

P-value

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.006[a]	3	.046
Likelihood Ratio	8.045	3	.045
Linear-by-Linear Association	7.867	1	.005
N of Valid Cases	395		

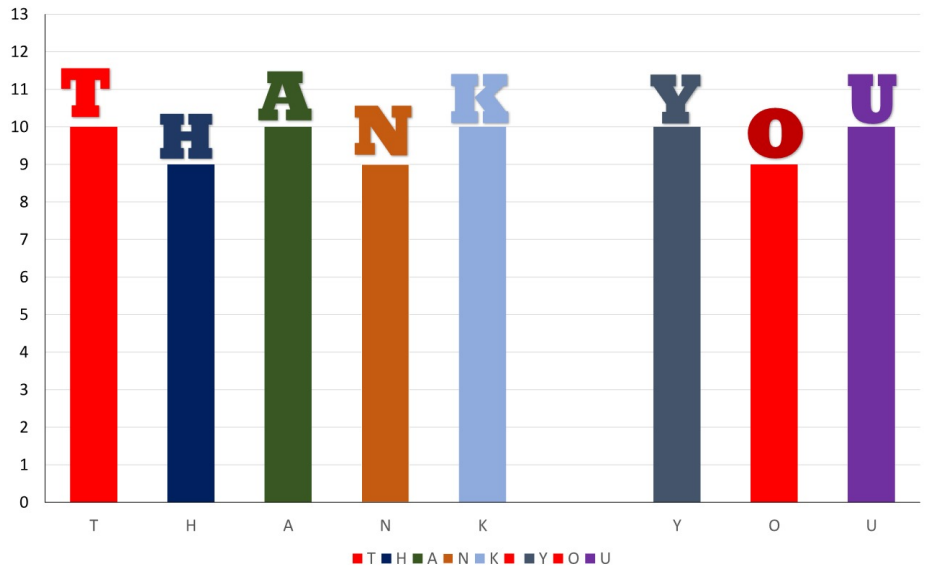
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 48.13.

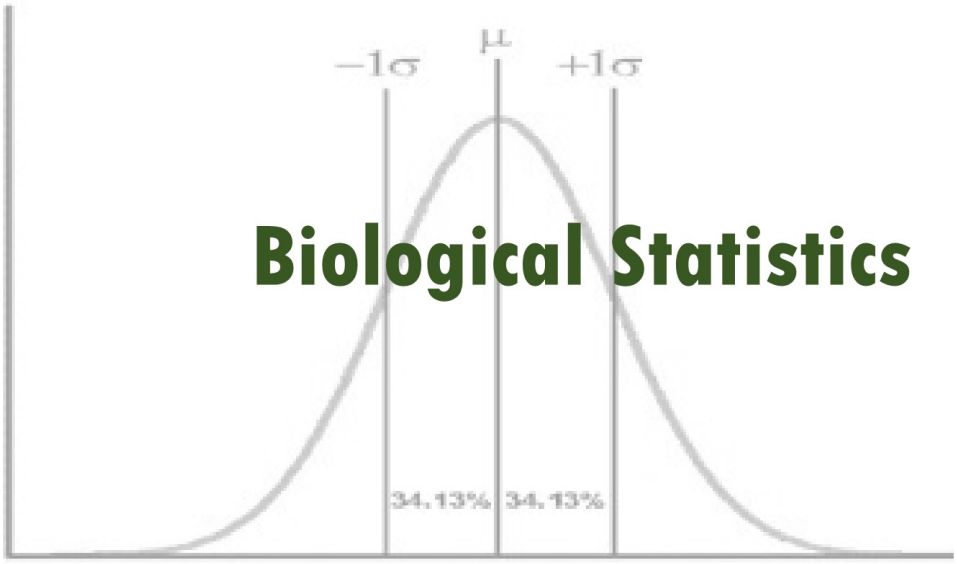
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.006[a]	3	.046
Likelihood Ratio	8.045	3	.045
Linear-by-Linear Association	7.867	1	.005
N of Valid Cases	395		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 48.13.

Since P-value < 0.05, We'll Reject H_0 *i.e.* Gender is significantly affect the study level of the examined population


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U



Biological Statistics

Regression Analysis

- It is a method to study the relationship between two variables, one of them is the dependent variable (the variable that we want to study), while the other is the independent (the factor that affect the studied variable).
- This relationship analysis could be between / among one dependent variable and one / multiple independent variable(s).
- This analysis provide us with an equation describes the relationship between/among the studied variables (dependent and independent(s)).

What we can get from regression Analysis

- Understand the nature of the effect of the independent variable / variables on the dependent variable, in other words, which of these variables is more influential than the others, and which one is unimportant to be excluded in the case of insignificance.
- Prediction or expectation: Through the mathematical model equation that we'll get from the regression analysis, we can assume any value for the independent variable(s) and obtain the expected value of the dependent variable without the need for other additional experiments to find out these values.

What is the difference between regression and correlation ?

- Correlation measures the strength of the relationship between two variables, regardless of their reality (for example, the relationship between population growth in Iraq and rainfall rates in Morocco). Also, the correlation does not care whether any of examined variables are independent and which are dependent.
- In the case of regression, the relationship must be logical and that it measures the effect of the independent variable or variables on the dependent variable (that is, there must be dependent and independent variables), and in the case of more than one independent variable, the regression can determine which of these variables is more influential and represents this relationship with a mathematical equation also enables us to predict, and whenever this equation is accurate, the prediction will be accurate. But if it is approximate, the prediction will have a margin of error that can be estimated and controlled.

Types of Regression

- Generally, the regression is either linear, which is the most common, or non-linear, where the linear relationship can be represented by the equation of a straight line, and the non-linear relationship can be represented by a non-linear equation (curve equation).
- The type of relationship can be identified simply by drawing a Scatter Plot between the dependent variable (on the y-axis) and the independent variable (on the x-axis). Through the spread and direction of the points, it can be determined weather the relationship is a linear (represented by a straight line), or non-linearly (represented by a curve).

Types of Regression

In general, we will focus on linear regression, and this type of regression, in turn, is divided into two types:

- Simple linear regression: In this type, the linear relationship between only two variables is studied, one of them is dependent while the second is independent, and it is a special case of multiple regression so that we eventually find a straight line equation that is in the following form:

$$y = b_0 + b_1 x$$

Whereas:
b_0: regression constant is the point of intersection of the regression line with the y-axis and represents the value of y when the influence of x = 0 .
b_1: Regression coefficient (Slop).
x: The independent variable that we want to examine its effect on dependent variable.

$$b_0 = \frac{\sum y - b \sum x}{n}$$

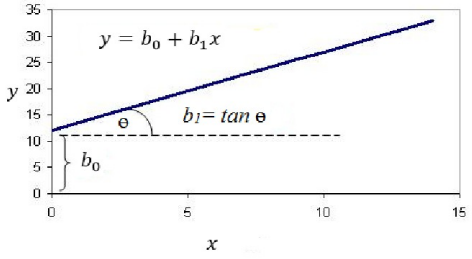
$$b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Types of Regression

$$y = b_0 + b_1 x$$

It is noted from the regression equation above, that after calculating the values of the regression constant and the regression coefficient, you only need to substitute the values of the independent variable x to obtain the corresponding values for the dependent variable y.

- The simple regression relationship can be represented by the following diagram:



Where the regression line is a straight line inclined at an angle of ɵ representing the coefficient of regression b_1 represents the tangent of this angle, and the ordered pairs of (x,y) represent the points that draw this line.

Example :

The table below represents a study of water consumption in ton (Y) for a particular city and the effect of air temperature (X) on it. We try, through simple regression analysis, to find the significant effect of temperature on the rate of water consumption and we'll try to represents this relationship with a mathematical equation to help predict the expected quantities of water that this city needs it in case of high or low temperatures in the future.

Y	212	220	230	235	255	270	283	295	310	330	360	350	225	235	235	242	248	257	279	283	285	300	224	236
X	20	21	23	24	27	29	30	32	33	34	35	35	21	22	23	24	25	26	27	28	29	30	22	24

Solution :

1. Input data

Data view

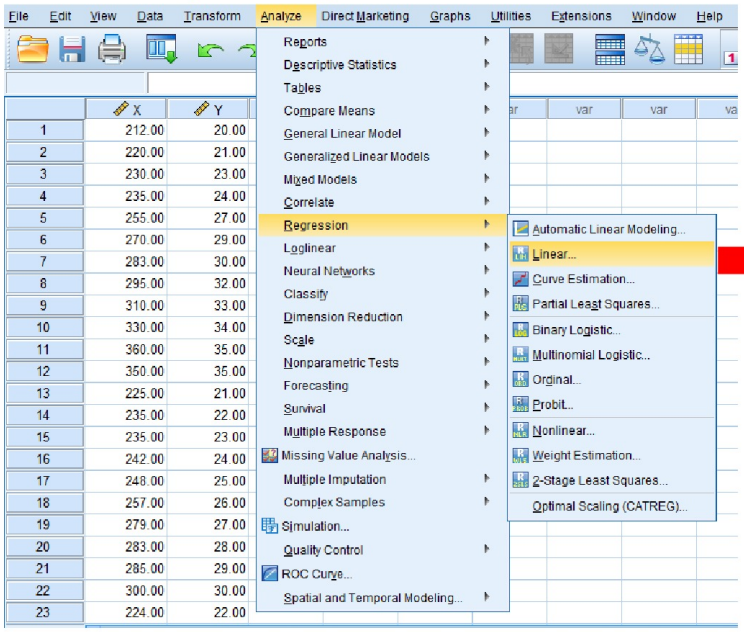
	X	Y
1	212.00	20.00
2	220.00	21.00
3	230.00	23.00
4	235.00	24.00
5	255.00	27.00
6	270.00	29.00
7	283.00	30.00
8	295.00	32.00
9	310.00	33.00
10	330.00	34.00
11	360.00	35.00
12	350.00	35.00
13	225.00	21.00
14	235.00	22.00
15	235.00	23.00
16	242.00	24.00
17	248.00	25.00
18	257.00	26.00
19	279.00	27.00
20	283.00	28.00
21	285.00	29.00
22	300.00	30.00
23	224.00	22.00

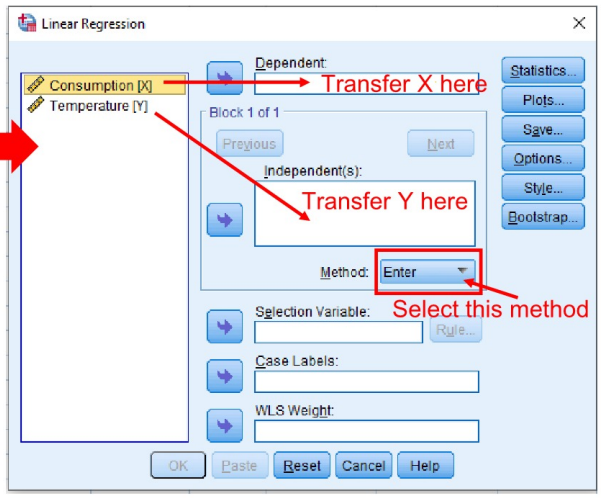
Variable view

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	X	Numeric	8	2	consumption	None	None	8	Right	Scale	Input
2	Y	Numeric	8	2	Temperature	None	None	8	Right	Scale	Input
3											

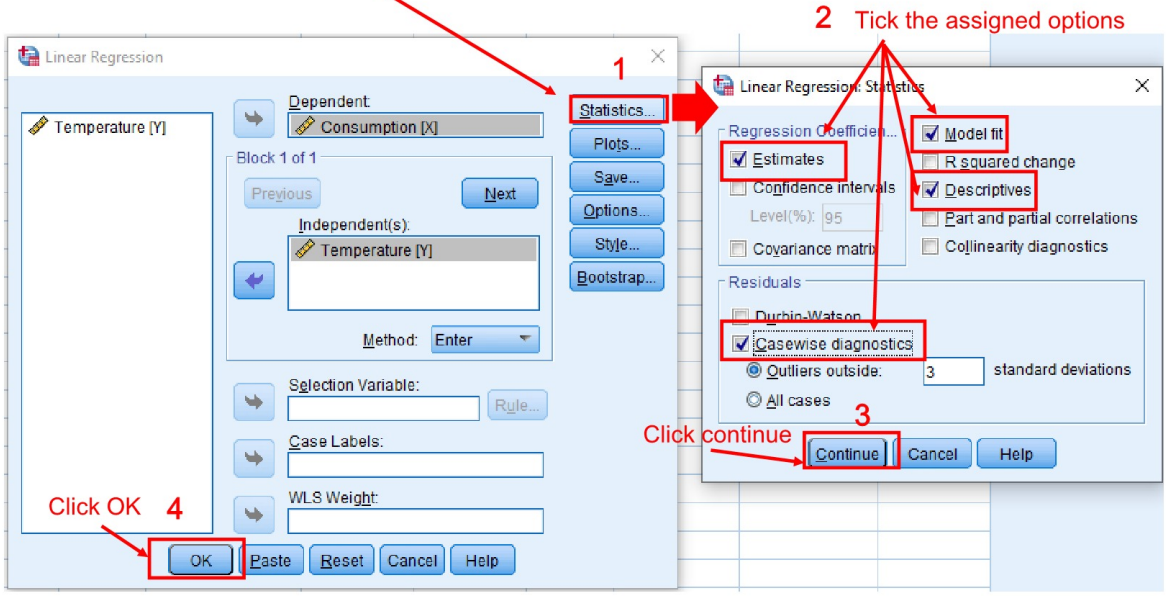
These labels will appear in results instead of variables name (X & Y)

2. Go to Analyze >> Regression >> Linear





2. Select Statistics bottom



3. The results will appear in output window

Descriptive Statistics

The means

	Mean	Std. Deviation	N
Consumption	266.6250	41.54287	24
Temperature	26.8333	4.65941	24

The Correlation is 0.973 which is positive and significant P< 0.01

Correlations

Correlation

		Consumption	Temperature
Pearson Correlation	Consumption	1.000	.973
	Temperature	.973	1.000
Sig. (1-tailed)	Consumption	.	.000
	Temperature	.000	
N	Consumption	24	24
	Temperature	24	24

P value

Variables Entered/Removed[a]

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

a. Dependent Variable: Consumption

b. All requested variables entered.

ANOVA table

ANOVA[a]

The regression is significant P< 0.01

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37591.234	1	37591.234	393.365	.000[b]
	Residual	2102.391	22	95.563		
	Total	39693.625	23			

P value

a. Dependent Variable: Consumption

b. Predictors: (Constant), Temperature

This means, when X (temperature) increases one unit the consumption of water will increase 8.677 folds

Coefficients[a]

This table used to write regression equation

P value

Model		b_1 / b_0 Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	33.804	11.907		2.839	.010
	Temperature	8.677	.437	.973	19.833	.000

a. Dependent Variable: Consumption

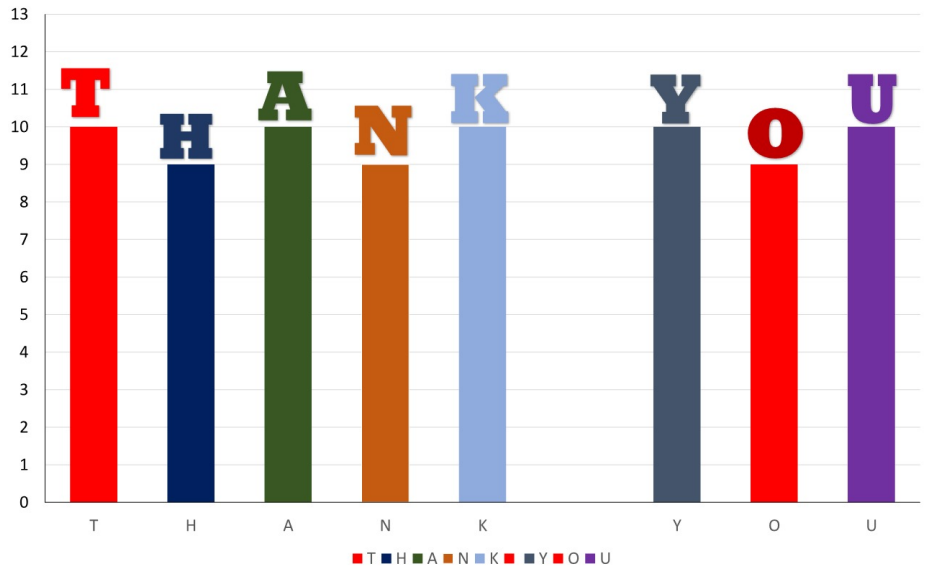
$Y = 33.804 + 8.677\ X$

This table show the residual analysis results, which are the difference between predicted value and actual value

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	207.3351	337.4836	266.6250	40.42773	24
Residual	-16.45394	22.51636	.00000	9.56077	24
Std. Predicted Value	-1.467	1.753	.000	1.000	24
Std. Residual	-1.683	2.303	.000	.978	24

a. Dependent Variable: Consumption


13
12
11
10
9
8
7
6
5
4
3
2
1
0
T
H
A
N
K
Y
O
U
T H A N K Y O U