

To understand the study of statistics, student must be aware of the importance of the terms used in statistics, and the following are the most important of those definitions:

A variable: is any characteristic, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables. It is called a variable because the value may vary between data units in a population, and may change in value over time.

For example; 'income' is a variable that can vary between data units in a population (i.e. the people or businesses being studied may not have the same incomes) and can also vary over time for each data unit (i.e. income can go up or down).

Types of variables:

There are different ways variables can be described according to the ways they can be studied, measured, and presented.

Numeric variables

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore, numeric variables are **quantitative variables**.

Numeric variables may be further described as either continuous or discrete:

- **A continuous variable** is a numeric variable. Observations can take any value between a certain set of real numbers. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.
- **A discrete variable** is a numeric variable. Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of which measured as whole units (i.e. 1, 2, 3 cars).

The data collected for a **numeric** variable are **quantitative** data.

Categorical variables

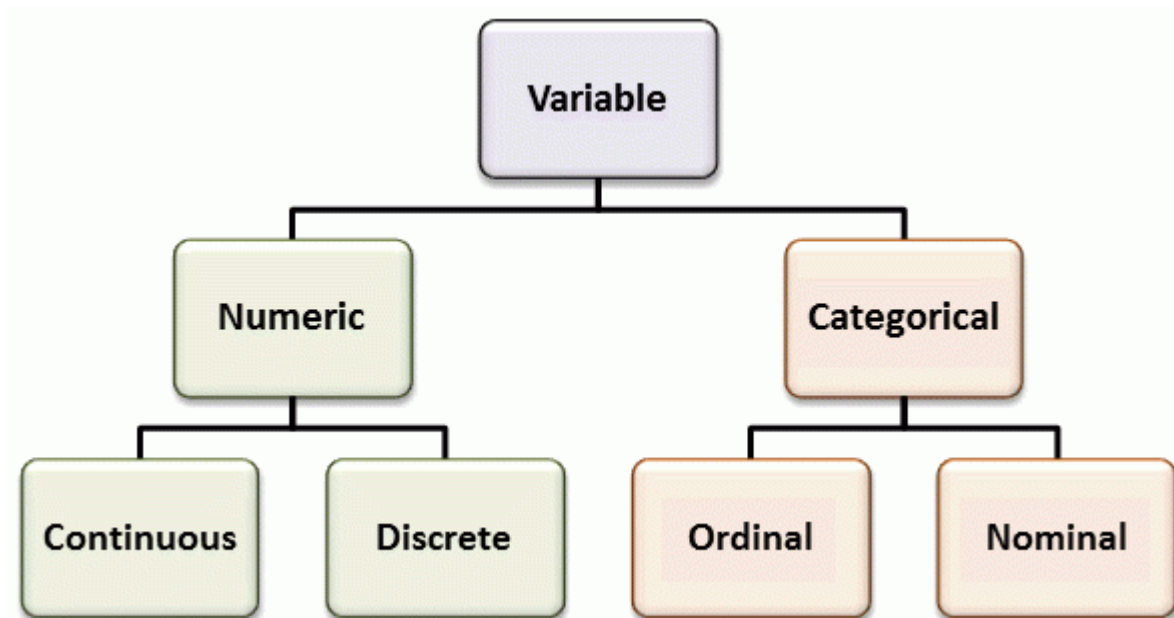
Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables fall into mutually exclusive (in one category or in another) and exhaustive (include all possible options) categories. Therefore,

categorical variables are **qualitative variables** and tend to be represented by a non-numeric value.

Categorical variables may be further described as ordinal or nominal:

- **An ordinal variable** is a categorical variable. Observations can take a value that can be logically ordered or ranked. The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category. Examples of ordinal categorical variables include academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra-large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree).
- **A nominal variable** is a categorical variable. Observations can take a value that is not able to be organized in a logical sequence. Examples of nominal categorical variables include sex, business type, eye color, religion and brand.

The data collected for a **categorical** variable are **qualitative** data.



The other names of two main types of variables are **Parametric** and **Non-parametric** which are equivalent in meaning **Numeric** and **Categorical** respectively.

Population: The whole group of people, items, or element of interest.

Sample: A subset of the population that researchers select and include in their study.

Researchers might want to learn about the characteristics of a population, such as its mean and standard deviation. Unfortunately, they are usually too large and expensive to study in their entirety.

Instead, the researchers draw a sample from the population to learn about it. Collecting data from a subset can be more efficient and cost-effective.

If we had to measure entire populations, we'd never be able to answer our research questions because they tend to be too large and unwieldy. Fortunately, we can use a subset to move forward.

In statistics, the term population doesn't always refer to the number of people; it can be any group, entity, or occurrence we want to study. "A population is a distinct group of individuals, creatures, or objects that can be distinguished for the purposes of data collection and analysis by at least one shared characteristic". The population's quantitative features are mean and standard deviation.

Population comes in many different types. We will concern to two types: **Finite population** and **Infinite population**.

Stages of Statistical process:

1. **Collecting of data**
2. **Descriptive Statistics**
3. **Inferential (Analytical) Statistics**

Some researchers divide the statistics into two types: **Descriptive and inferential**.

Data collection tools

A list of basic data collection tools includes the following:

- Observation method
- Interview method
- Questionnaire method
- Surveys
- Experiments

Errors in Measurement:

Measurement is the foundation for all experimental science. All the great technological development could not have been possible without ever-increasing levels of accuracy of measurements. The measurement of an amount is based on some international standards, which are completely accurate compared with others. Just like your vegetable vendors, measurements are taken by comparing an unknown amount with a known weight. Every measurement carries a level of uncertainty which is known as an error. This error may arise in the process or due to a mistake in the experiment. So 100% accurate measurement is not possible with any method.

An error may be defined as the difference between the measured and actual values. For example, if the two operators use the same device or instrument for measurement. It is not necessary that both operators get similar results. The difference between the measurements is referred to as an ERROR.

There are three types of errors that are classified based on the source they arise from; They are: **Gross Errors, Random Errors** and **Systematic Errors**.

Gross Errors

This category basically takes into account human oversight and other mistakes while reading, recording, and readings. The most common human error in measurement falls under this category of measurement errors. For example, the person taking the reading from the meter of the instrument may read 23 as 28. Gross errors can be avoided by using two suitable measures, and they are written below:

Proper care should be taken in reading, recording the data. Also, the calculation of error should be done accurately. By increasing the number of experimenters, we can reduce the gross errors. If each experimenter takes different readings at different points, then by taking the average of more readings, we can reduce the gross errors

Random Errors

The random errors are those errors, which occur irregularly and hence are random. These can arise due to random and unpredictable fluctuations in experimental conditions (Example: unpredictable fluctuations in temperature, voltage supply, mechanical vibrations of experimental set-ups, etc, errors by the observer taking readings, etc. For example, when the same person repeats the same observation, he may likely get different readings every time.

Systematic Errors:

Systematic errors can be better understood if we divide them into subgroups; They are:

Environmental Errors, Observational Errors and Instrumental Errors.

Environmental Errors: This type of error arises in the measurement due to the effect of the external conditions on the measurement. The external condition includes temperature, pressure, and humidity and can also include an external magnetic field. If you measure your temperature under the armpits and during the measurement, if the electricity goes out and the room gets hot, it will affect your body temperature, affecting the reading.

Observational Errors: These are the errors that arise due to an individual's bias, lack of proper setting of the apparatus, or an individual's carelessness in taking observations. The measurement errors also include wrong readings due to Parallax errors.

Instrumental Errors: These errors arise due to faulty construction and calibration of the measuring instruments. Such errors arise due to the hysteresis of the equipment or due to friction. Lots of the time, the equipment being used is faulty due to misuse or neglect, which changes the reading of the equipment. The zero error is a very common type of error. This error is common in devices like Vernier callipers and screw gauges. The zero error can be either positive or negative. Sometimes the scale readings are worn off, which can also lead to a bad reading.

Instrumental error takes place due to : 1. An inherent constraint of devices 2. Misuse of Apparatus and 3. Effect of Loading.

Keeping an eye on the procedure and following the below listed points can help to reduce the error.

- Make sure the formulas used for measurement are correct.

- Cross check the measured value of a quantity for improved accuracy.
- Use the instrument that has the highest precision.
- It is suggested to pilot test measuring instruments for better accuracy.
- Use multiple measures for the same construct.
- Note the measurements under controlled conditions.

Accuracy

The ability of an instrument to measure the accurate value is known as accuracy. In other words, it is the closeness of the measured value to a standard or true value. Accuracy is obtained by taking small readings. The small reading reduces the error of the calculation. The accuracy of the system is classified into three types as follows:

Point Accuracy

The accuracy of the instrument only at a particular point on its scale is known as point accuracy. It is important to note that this accuracy does not give any information about the general accuracy of the instrument.

Accuracy as Percentage of Scale Range

The uniform scale range determines the accuracy of a measurement. This can be better understood with the help of the following example:

Consider a thermometer having the scale range up to 500°C. The thermometer has an accuracy of ± 0.5 percent of scale range i.e. $0.005 \times 500 = \pm 2.5$ °C. Therefore, the reading will have a maximum error of ± 2.5 °C.

Accuracy as Percentage of True Value

Such type of accuracy of the instruments is determined by identifying the measured value regarding their true value. The accuracy of the instruments is neglected up to ± 0.5 percent from the true value.

Precision

The closeness of two or more measurements to each other is known as the precision of a substance. If you weigh a given substance five times and get 3.2 kg each time, then your measurement is very precise but not necessarily accurate. Precision is independent of accuracy. The below examples will tell you about how you can be precise but not accurate and vice versa. Precision is sometimes separated into:

Repeatability

The variation arising when the conditions are kept identical and repeated measurements are taken during a short time period.

Reproducibility

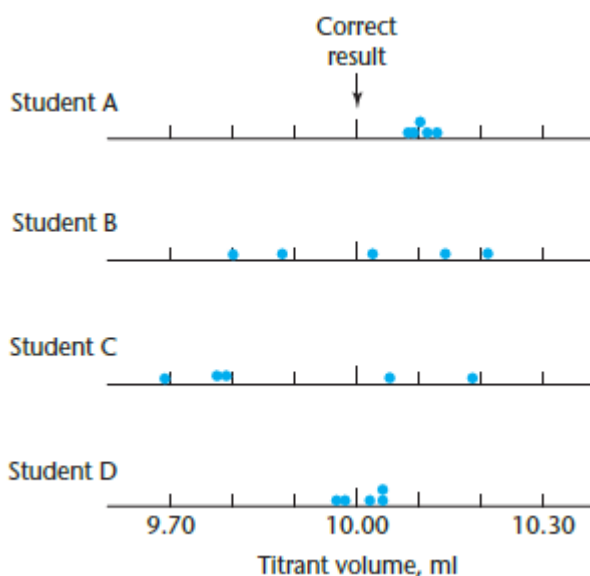
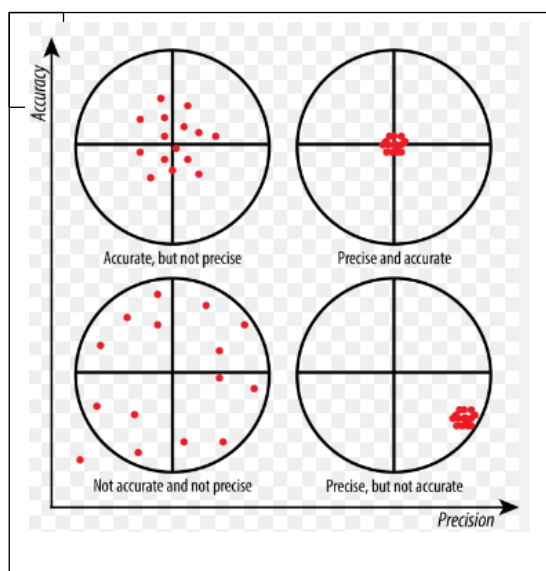
The variation arises using the same measurement process among different instruments and operators, and over longer time periods.

Conclusion

Accuracy is the degree of closeness between a measurement and its true value. Precision is the degree to which repeated measurements under the same conditions show the same results.

Accuracy and Precision Examples

A good analogy for understanding accuracy and precision is to imagine a football player shooting at the goal. If the player shoots into the goal, he is said to be accurate. A football player who keeps striking the same goalpost is precise but not accurate. Therefore, a football player can be accurate without being precise if he hits the ball all over the place but still scores. A precise player will hit the ball to the same spot repeatedly, irrespective of whether he scores or not. A precise and accurate football player will not only aim at a single spot but also score the goal.



Student	Results (ml)					Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise, unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

1. Some statistical operations using summation (Σ):

- a. $\Sigma x \dots$ sum all the values of x
- b. $\Sigma x^2 \dots$ square each x and then sum the values of x^2
- c. $\Sigma xy \dots$ multiple each pair of x and y to obtain xy and then sum the values of xy
- d. $(\Sigma x)^2 \dots$ sum all the values of x and then square Σx
- e. $\Sigma x \Sigma y \dots$ sum all the values of x , sum all the values of y , and then multiple the two sums Σx and Σy . Note that $\Sigma xy \neq \Sigma x \Sigma y$
- f. $\Sigma(x + 2) \dots$ add 2 to each x and then sum
- g. $\Sigma(x - y) \dots$ subtract y from each x and then sum. This is the same as doing $\Sigma x - \Sigma y$
- h. $\Sigma(x + y) \dots$ add y to each x and then sum. This is the same as doing $\Sigma x + \Sigma y$
- i. $\Sigma(2x) \dots$ add each x and then multiply the sum by 2. Note that $\Sigma(2x) = 2 \Sigma x$

$$\sum_{i=1}^n y_i^2 = y_1^2 + y_2^2 + y_3^2 + \dots + y_n^2$$

ويرمز لمربع مجموع المشاهدات بالرمز $\left(\sum_{i=1}^n y_i\right)^2$ ويساوي :

$$\left(\sum y_i\right)^2 = (y_1 + y_2 + y_3 + \dots + y_n)^2$$

كما ي رمز لحاصل ضرب متغيرين مثل x و y بالرمز $\sum x_i y_i$ وهو يساوي :

$$\sum x_i y_i = y_1 x_1 + y_2 x_2 + y_3 x_3 + \dots + y_n x_n$$

ويرمز لحاصل ضرب مجموعتين لقيم متغيرين بالرمز $(\sum x_i)(\sum y_i)$ وهو يساوي :

$$(\sum x_i)(\sum y_i) = (x_1 + x_2 + x_3 + \dots + x_n)(y_1 + y_2 + y_3 + \dots + y_n)$$

وفيما يلي بعض القواعد الهامة في عملية الجمع :
قاعدة (1)

$$\sum_{i=1}^n c = nc$$

إذا كانت C أي عدد ثابت فإن :

$$\sum_{i=1}^n c = c_1 + c_2 + c_3 + \dots + c_n = nc$$

قاعدة (2)

إذا كانت C أي عدد ثابت فإن : -

$$\sum cy_i = c \sum y_i$$

البرهان : -

$$\begin{aligned} \sum cy_i &= cy_1 + cy_2 + cy_3 + \dots + cy_n \\ &= c(y_1 + y_2 + y_3 + \dots + y_n) \\ &= c \sum y_i \end{aligned}$$

القاعدة (3)

جميع قيم متغيرين او اكثر هو مجموع جميعهم أي ان :-

$$\sum (x_i + y_i) = \sum x_i + \sum y_i$$

البرهان :-

$$\begin{aligned} \sum (x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \dots + (x_n + y_n) \\ &= (x_1 + x_2 + \dots + x_n) + (y_1 + y_2 + \dots + y_n) \\ &= \sum x_i + \sum y_i \end{aligned}$$

• ويجب التفريق بين بعض الرموز الاحصائية مثل :-

$$-\sum \frac{x_i}{y_i} = \frac{x_1}{y_1} + \frac{x_2}{y_2} + \dots + \frac{x_n}{y_n}$$

وهذا يختلف عن الصورة التالية :

$$-\frac{\sum x_i}{\sum y_i} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{y_1 + y_2 + y_3 + \dots + y_n}$$

وكذلك فان :

$$-\sum (x_i - 3) = \sum x_i - n(3)$$

وهي تختلف عن الصورة :-

$$-\sum x_i - 3$$

سؤال : اذا علمت ان قيم كل من المتغيرين x و y هي كالاتي :

$$\begin{aligned} x_i &= 2, 6, 3, 1 \\ y_i &= 3, 9, 6, 2 \end{aligned}$$

اوجد قيمة كل مما ياتي :-

1-

$$\sum_{i=1}^n y_i$$

9 -

$$\sum x_i y_i^2$$

2 -

$$\sum_{i=2}^3 y_i$$

10 -

$$\sum (y_i - 3)$$

3 -

$$\sum y_i^2$$

11 -

$$\sum y_i - 3$$

4 -

$$(\sum y_i)^2$$

12 -

$$\sum \frac{x_i + 2}{y_i}$$

5 -

$$\sum x_i y_i$$

13 -

$$\sum \frac{\sum (x_i + 2)}{\sum y_i}$$

6 -

$$(\sum x_i)(\sum y_i)$$

14 -

$$\sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

7 -

$$\sum (y_i - x_i)^2$$

15 -

$$\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

8 -

$$\sum (x_i - 3)(y_i - 5)$$

1 -

$$\begin{aligned}\sum yi &= y1 + y2 + y3 + y4 \\ &= 3 + 9 + 6 + 2 = 20\end{aligned}$$

2 -

$$\begin{aligned}\sum_{i=2}^3 yi &= y2 + y3 \\ &= 9 + 6 = 15\end{aligned}$$

3 -

$$\begin{aligned}\sum yi &= y_1^2 + y_2^2 + y_3^2 + y_4^2 \\ &= (3)^2 + (9)^2 + (6)^2 + (2)^2 \\ &= 130\end{aligned}$$

4 -

$$\begin{aligned}(\sum yi)^2 &= (y1 + y2 + y3 + y4)^2 \\ &= (3 + 9 + 6 + 2)^2 = (20)^2 \\ &= 400\end{aligned}$$

5 -

$$\begin{aligned}\sum xi yi &= x1y1 + x2y2 + x3y3 + x4y4 \\ &= (2)(3) + (6)(9) + (3)(6) + (1)(2) \\ &= 80\end{aligned}$$

6 -

$$\begin{aligned}(\sum xi)(\sum yi) &= (x1 + x2 + x3 + x4)(y1 + y2 + y3 + y4) \\ (12)(20) &= 240\end{aligned}$$

7 -

$$\begin{aligned}\sum (yi - xi)^2 &= (y1 - x1)^2 + (y2 - x2)^2 + (y3 - x3)^2 + (y4 - x4)^2 \\ &= (3 - 2)^2 + (9 - 6)^2 + (6 - 3)^2 + (2 - 1)^2 \\ &= 20\end{aligned}$$

8 -

$$\begin{aligned}\sum (xi - 3)(yi - 5) &= (x1 - 3)(y1 - 5) + (x2 - 3)(y2 - 5) + (x3 - 3)(y3 - 5) + (x4 - 3)(y4 - 5) \\ &= (2 - 3)(3 - 5) + (6 - 3)(9 - 5) + (3 - 3)(6 - 5) + (1 - 3)(2 - 5) \\ &= 20\end{aligned}$$

وهنا ايضا يمكن الوصول الى نفس النتيجة بفتح الاقواس ثم التعويض كما يلي :

$$\begin{aligned}\sum (yi - 3)(xi - 5) &= \sum (xi yi - 5xi - 3yi + 15) \\ &= \sum xi yi - 5 \sum xi - 3 \sum yi + (4)(15) \\ &= 80 - 5(12) - 3(20) + 60 = 20\end{aligned}$$

9-

$$\begin{aligned} \sum x_i y_i^2 &= x_1 y_1^2 + x_2 y_2^2 + x_3 y_3^2 + x_4 y_4^2 \\ &= (2)(3)^2 + (6)(9)^2 + (3)(6)^2 + (7)(2)^2 \\ &= 616 \end{aligned}$$

10-

$$\begin{aligned} \sum (y_i - 3) &= \sum y_i - \sum 3 \\ &= \sum y_i - n(3) \\ &= \sum y_i - (4)(3) \\ &= 20 - 12 = 8 \end{aligned}$$

11-

$$\begin{aligned} \sum y_i - 3 \\ &= 20 - 3 \\ &= 17 \end{aligned}$$

12-

$$\begin{aligned} \sum \frac{x_i + 2}{y_i} &= \frac{x_1 + 2}{y_1} + \frac{x_2 + 2}{y_2} + \frac{x_3 + 2}{y_3} + \frac{x_4 + 2}{y_4} \\ &= \frac{2+2}{3} + \frac{6+2}{9} + \frac{3+2}{6} + \frac{1+2}{2} \\ &= \frac{164}{36} \end{aligned}$$

13-

$$\begin{aligned} \sum \frac{\sum (x_i + 2)}{\sum y_i} &= \frac{\sum x_i + (n)(2)}{\sum y_i} \\ \frac{12+8}{20} &= 1 \end{aligned}$$

14-

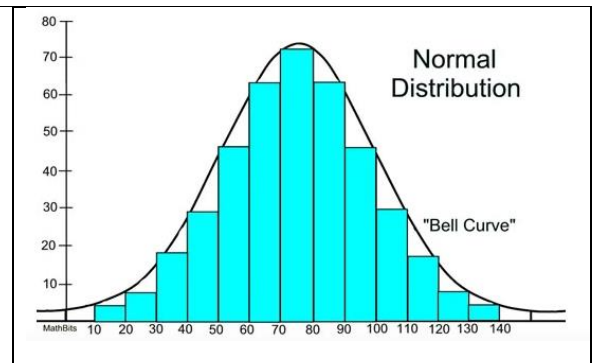
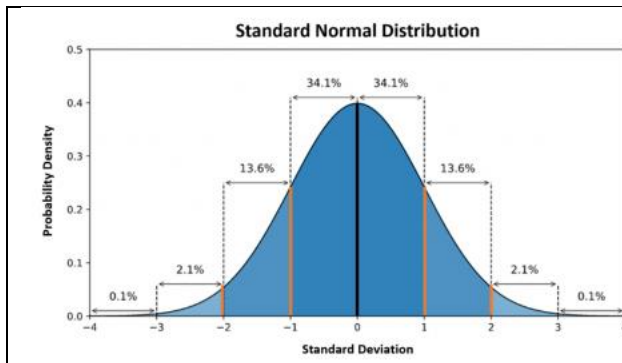
$$\begin{aligned} \sum y_i - \frac{(\sum x_i)^2}{n} \\ &= y_1^2 + y_2^2 + y_3^2 + y_4^2 - \frac{(y_1 + y_2 + y_3 + y_4)^2}{4} \\ &= (3)^2 + (9)^2 + (6)^2 + (2)^2 - \frac{(3+9+6+2)^2}{4} \\ &= 130 - \frac{(20)^2}{4} \\ &= 130 - 100 = 30 \end{aligned}$$

15-

$$\begin{aligned} \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} &= x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 \\ &= (2)(3) + (6)(9) + (3)(6) + (1)(2) - \frac{(12)(20)}{4} \\ &= 80 - \frac{(12)(20)}{4} = 20 \end{aligned}$$

2. some types of figures:

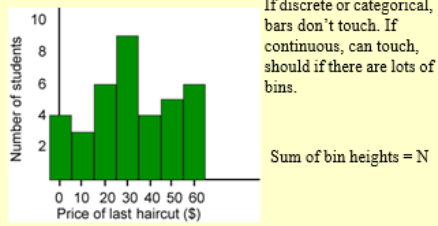
a. bell curve: distribution curve:



b. Histogram

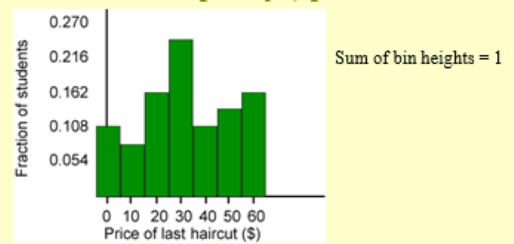
Discrete data

Histogram



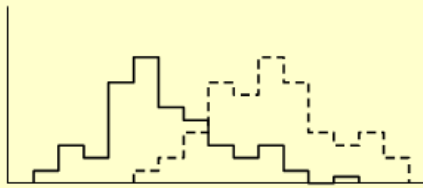
Continuous data

Alternative: density (or "relative frequency") plot



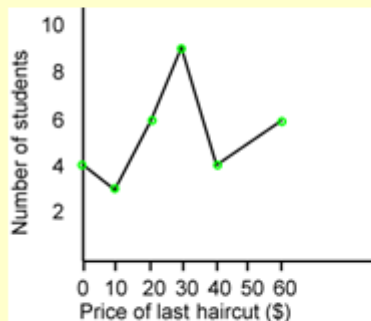
We can plot two histograms in the same plot, to compare them

- E.G. distribution of heights for women (solid) vs. men (dashed)

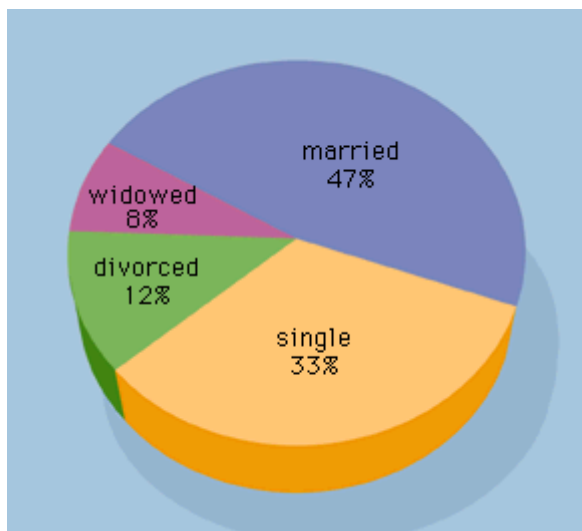


c. Frequency polygon

Alt: for lots of bins, continuous variable, draw histogram as a "frequency polygon"



d. Pie chart



3. Central Tendency:

A score that indicates where the center of the distribution tends to be located.

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

The **mean** (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the **median** and the **mode**.

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, \dots, x_n the sample mean, usually denoted by \bar{x} (pronounced "x bar"), is:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

This formula is usually written in a slightly different manner using the Greek capitol letter, Σ , pronounced "sigma", which means "sum of:"...

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

You may have noticed that the above formula refers to the sample mean.

To calculate mean from frequency table the formula is:

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

As the figure below

Marks scored	Frequency	Mid-point	Frequency \times Mid-point
0 - 9	3	$\frac{0 + 9}{2} = 4.5$	$3 \times 4.5 = 13.5$
10 - 19	5	$\frac{10 + 19}{2} = 14.5$	$5 \times 14.5 = 72.5$
20 - 29	8	$\frac{20 + 29}{2} = 24.5$	$8 \times 24.5 = 196$
30 - 39	4	$\frac{30 + 39}{2} = 34.5$	$4 \times 34.5 = 138$
	n = 20		Total = 420

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is \$30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the \$12k to 18k range.

Central Tendency

Mean, median, and mode are different measures of center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

Mean: The "average" number; found by adding all data points and dividing by the number of data points.

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Mode: The most frequent number—that is, the number that occurs the highest number of times.

The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

Mean:

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero. But the mean has one main disadvantage: it is particularly susceptible to the influence of outliers.

Ungrouped data: $\bar{x} = \frac{\sum xi}{n}$

Grouped data: $\bar{x} = \frac{\sum xifi}{\sum fi}$, where xi represents the center of period, and fi means the frequencies.

To calculate the mean for grouped data you need to construct a table called frequency table which is composed of (Classes, frequencies) and to estimate mean you have to add a column called center of class. The steps of constructing the table are as follows:

1. Estimate the range. (max-min)
2. Calculate the classes number, (should be between 5 to 15 and most times we can choose the number without calculation)

There are two ways to calculate the number of classes: either $2.5 * \sqrt[4]{n}$ or $(1 + 3.32 \log n)$

3. Estimate the class length by dividing the range on number of classes.
4. We choose the smallest reading in the data to be the beginning of the lower limit of the first class and add the length of the class to it, so we get the beginning of the second class.
5. Determine the center of the class using the formula $\frac{\text{lower limit} + \text{upper limit}}{2}$
6. Mark the number of frequencies of each class.
7. Multiply frequency by center of class the summate them.

EXAMPLE: The following data represent number of teaching staff in each laboratory 4, 1, 5, 4, 3, 7, 2, 3, 4, 1

Calculate Mean?

$$\bar{x} = \frac{\sum xi}{n} \quad \bar{x} = \frac{4+1+5+4+3+7+2+3+4+1}{10} = 3.4$$

Exercise 1: Find the mean a. 8, 10, 12, 14, 7, 16, 5, 7, 9, 11 b. 108, 99, 112, 111, 108

c. 64, 66, 65, 61, 67, 61, 57

EXAMPLE: The following data represent weight of 40 students. Find the mean?

40 28 29 31 35 33 32 40 41 29
 45 32 29 44 29 45 39 35 43 42
 45 31 28 43 28 44 38 34 42 41
 41.5 29.5 30.5 32.5 36.5 34.5 33.5 41.5 42.5 30.5

Solution:

Range= 45-28 = 17

No. of classes= $1 + 3.32 \log 40 = 1 + 3.32(1.6) = 6.3$ so it is 6, but probably we need

Class length = range/ No. of classes = $17/6 \approx 3$

Center of class = $(28+30)/2 = 29$

classes	Frequency (f)	Centre of class (xi)	fixi
28-30	8	29	232
31-33	7	32	224
34-36	5	35	175
37-39	4	38	152
40-42	9	41	369
43-45	7	44	308
Summation Σ	40		1460

$$\bar{x} = \frac{\sum xifi}{\sum fi} \quad \bar{x} = \frac{1460}{40} = 36.5$$

EXERCISE: The following data is marks of 50 students in statistics. Find the mean?

66 68 69 74 77 79 65 78 76 67
 70 72 73 78 81 83 69 82 80 71
 75 77 78 83 86 88 74 87 85 76
 81 83 84 89 92 94 80 93 91 82
 87 89 90 95 98 100 86 99 97 88

Median:

The median is the middle point in a dataset—half of the data points are smaller than the median and half of the data points are larger.

To find the median:

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
 If the number of values are even, median is the mean of middle two values.

By formula

When n is odd, Median = Md = $(\frac{n+1}{2})^{th}$ value

- If the number of data points is even, the median is the average of the two middle data points in the list.
 When n is even, Average of $(\frac{n}{2})^{th}$ and $(\frac{n}{2} + 1)^{th}$ values

Example

If the weights of chemicals are 45, 60,48,100,65 gms, calculate the median

Solution

Here $n = 5$

First arrange it in ascending order 45, 48, 60, 65, 100

Median = $(\frac{n+1}{2})^{th} = (\frac{5+1}{2})^{th}$, so Median is 60

Example

If the weight of chemicals are 5,48, 60, 65, 65, 100 gms, calculate the median.

Solution

Here $n = 6$

First arrange it in ascending order 45, 48, 60, 65, 100

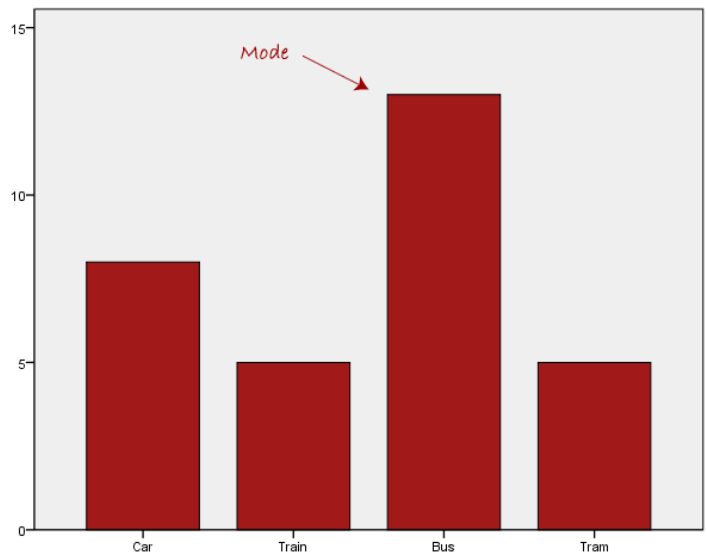
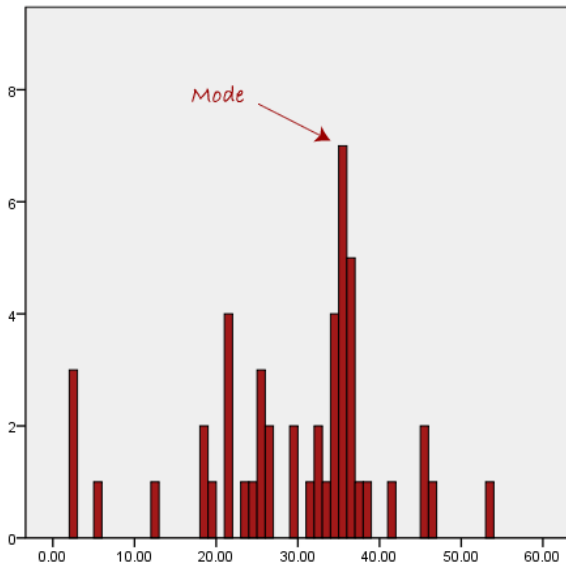
Median = Average of $(\frac{n}{2})^{th}$ and $(\frac{n}{2} + 1)^{th}$ value, that means Average of $(\frac{6}{2})^{th}$ and $(\frac{6}{2} + 1)^{th}$ values

Median = $(\frac{60+65}{2}) = 62.5$

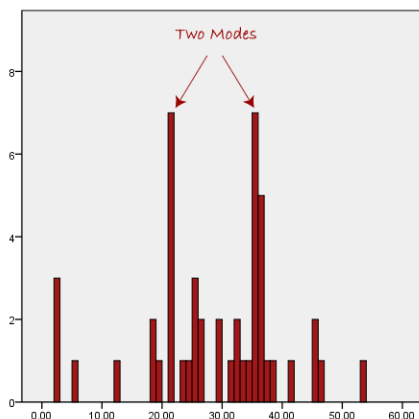
One of the most important features of the median is that it is not affected by extreme values. It can also be found in the case of nonparametric data that can be ordered. The problem of median: It does not take all values into account when calculating it.

Mode:

The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:



However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:



Graphic Methods:

The first question to ask when considering how best to display data is whether a graphical method is needed at all. It's true that in some circumstances a picture may be worth a thousand words, but at other times, frequency tables do a better job than graphs at presenting information. This is particularly true when the actual values of the numbers in different categories, rather than the general pattern among the categories, are of primary interest. Frequency tables are often an efficient way to present large quantities of data and represent a middle ground between text (paragraphs describing the data values) and pure graphics (such as a histogram).

One of the statistics they collect is the Body Mass Index (BMI), calculated as weight in kilograms divided by squared height in meters. The BMI is not an infallible measure. For instance, athletes often measure as either underweight (distance runners, gymnasts) or overweight or obese (football players, weight throwers), but it's an easily calculated measurement that is a reliable indicator of a healthy or unhealthy body weight for many people.

The BMI is a continuous measure, but it is often interpreted in terms of categories, using commonly accepted ranges. The ranges for the BMI shown below, established by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), are generally accepted as useful and valid.

BMI range	Category
< 18.5	Underweight
18.5–24.9	Normal weight
25.0–29.9	Overweight
30.0 and above	Obese

If our students numbers in College of Applied Sciences in 2022 were as shown for each category:

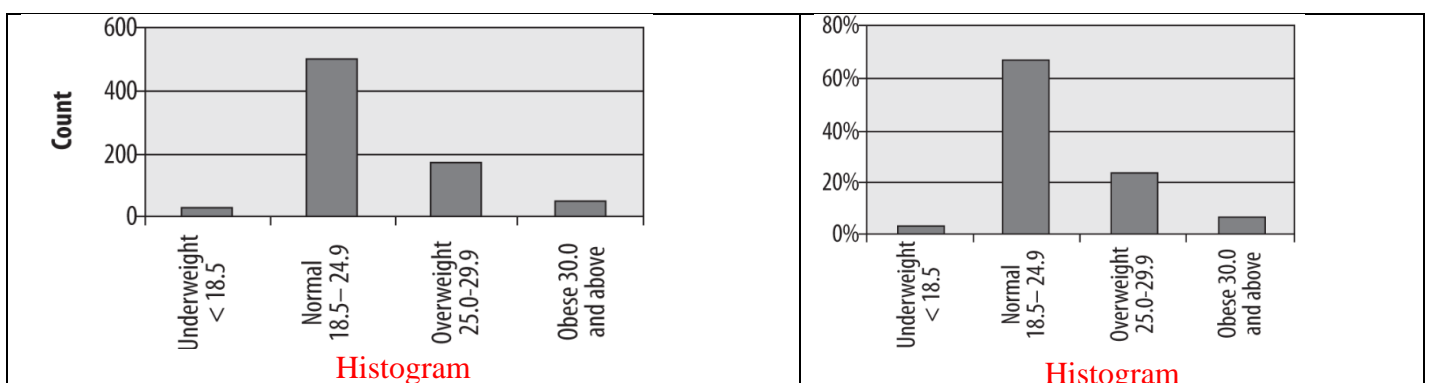
BMI range	Number
< 18.5	25
18.5–24.9	500
25.0–29.9	175
30.0 and above	50

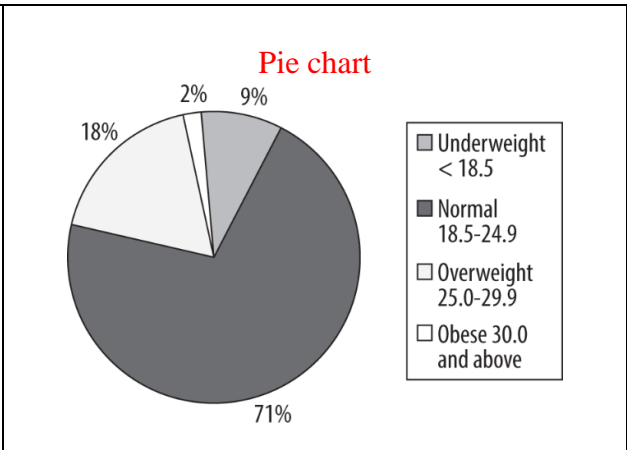
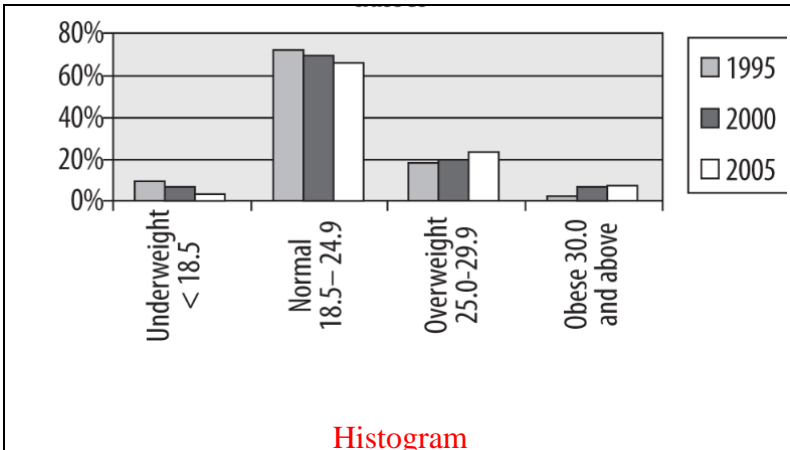
frequency table

BMI range	Number	Relative frequency
< 18.5	25	3.3%
18.5–24.9	500	66.7%
25.0–29.9	175	23.3%
30.0 and above	50	6.7%

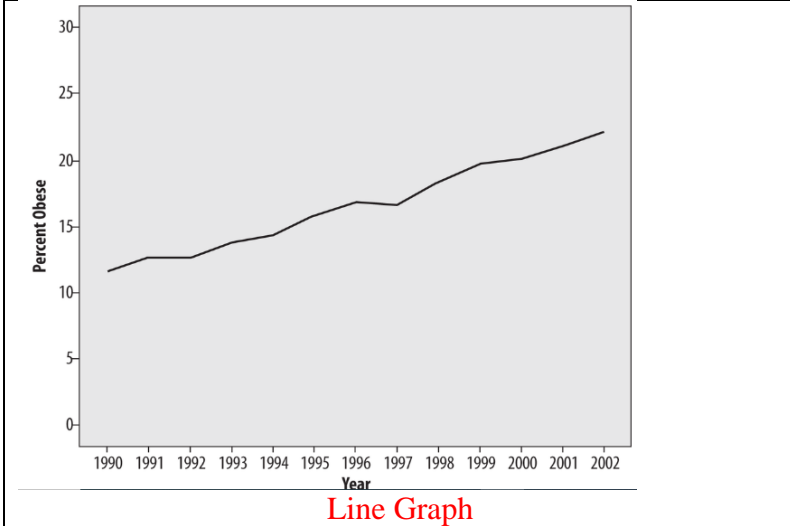
Relative frequency table

So we can use different figures to explain the frequency table.

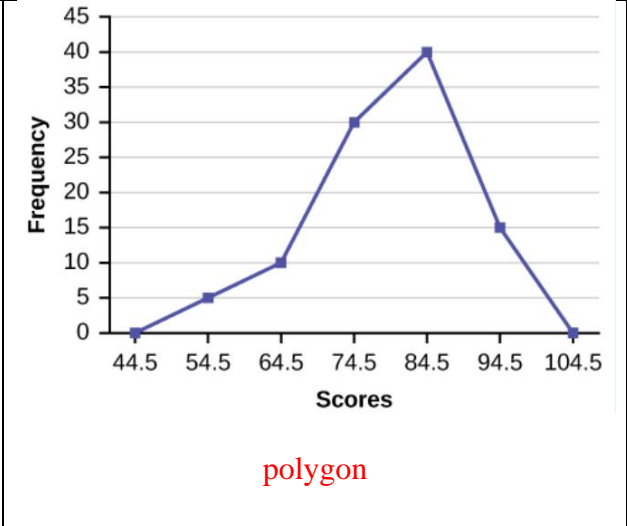




Histogram



Line Graph



polygon

Application of Descriptive Statistics Using Microsoft Excel

The screenshot shows a Microsoft Excel worksheet with the following data table:

40	28	29	31	35	33	32	40	41	29
60	48	49	51	55	53	52	60	61	49
33	34	36	40	38	37	45	46	34	55
71	75	73	72	80	81	69	90	35	35

The formula bar shows the function `=max` being applied to a range of cells. The dropdown menu shows the following options:

- MAX
- MAXA
- MAXIFS
- DMAX

The status bar at the bottom right indicates the date is 2024/03/19 and the time is 05:31.

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به

مشاركة

بحث فرز وتصفية وتحديد

جمع تلقائي تصفية مسح

تنسيق حذف إدراج خلايا

تنسيق الخلية أنماط شرطية

0.00 0.00 %

إرجاع أصغر قيمة موجودة في مجموعة من القيم. يتم تجاهل القيم والنصوص المنطقية

MIN MINA MINIFS MINUTE MINVERSE DMIN NOMINAL NORMINV

1																		
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

1 ورقة 2 ورقة

100%

إدخال

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به

مشاركة

بحث فرز وتصفية وتحديد

جمع تلقائي تصفية مسح

تنسيق حذف إدراج خلايا

تنسيق الخلية أنماط شرطية

0.00 0.00 %

إرجاع أصغر قيمة موجودة في مجموعة من القيم. يتم تجاهل القيم والنصوص المنطقية

MIN MINA MINIFS MINUTE MINVERSE DMIN NOMINAL NORMINV

1																		
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

1 ورقة 2 ورقة

100%

إدخال

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بها تزيد القيام به

تحرير: =med

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90		28											
10					48.875													
11					=med													
12					MEDIAN													
13					إرجاع الوسيط أو الرقم الموجود في منتصف مجموعة من الأرقام المحددة.													
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 1 ورقة 2

100%

Type here to search

05:36 ص 2024/03/19

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بها تزيد القيام به

تحرير: =MEDIAN(D4:M7)

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90		28											
10					48.875													
11					=MEDIAN(D4:M7)													
12					MEDIAN(number1; [number2]; ...)													
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 1 ورقة 2

100%

Type here to search

05:36 ص 2024/03/19

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أخرجني بما تريد القيام به

تحرير

عامة

نسخ

الخط

E12

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90	28												
10					48.875													
11					45.5													
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 2 ورقة 1

100%

05:37 ص 2024/03/19

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أخرجني بما تريد القيام به

تحرير

عامة

نسخ

الخط

SUM

=mode

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90	28												
10					48.875													
11					45.5													
12					=mode													
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 2 ورقة 1

100%

05:37 ص 2024/03/19

MODE.MULT
MODE.SNGL
MODE

هذه الدالة متوفرة للتوافق مع Excel 2007 والإصدارات السابقة.
إرجاع القيمة الأكثر تكراراً أو الأكثر ظهوراً في صفيف أو في نطاق من البيانات

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به

مشاركة

بحث فرز تصفية وتحديد

تحرير

خلايا

أنماط

عام

رقم

محاذاة

خط

الحافظة

الشريط الرئيسي

ملف

قص

لصق

نسخ

نسخ التنسيق

الحافظة

D4 =MODE(D4:M7)

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90	28												
10					48.875													
11					45.5													
12					=MODE(D4:M7)													
13					MODE(number1; [number2]; ...)													
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 2

100%

نقطة

Type here to search

05:38 ص 2024/03/19

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به

مشاركة

بحث فرز تصفية وتحديد

تحرير

خلايا

أنماط

عام

رقم

محاذاة

خط

الحافظة

الشريط الرئيسي

ملف

قص

لصق

نسخ

نسخ التنسيق

الحافظة

E13

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2																		
3																		
4				40	28	29	31	35	33	32	40	41	29					
5				60	48	49	51	55	53	52	60	61	49					
6				33	34	36	40	38	37	45	46	34	55					
7				71	75	73	72	80	81	69	90	35	35					
8																		
9					90	28												
10					48.875													
11					45.5													
12					40													
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		

ورقة 2

100%

نقطة

Type here to search

05:38 ص 2024/03/19

Word - Third Lecture Statistics ...

Dispersion tendency in statistics:**Variance:**

Variance in statistics is **a measurement of the spread between numbers in a data set**. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set, so Variance defined as the average of the squared differences from the mean.

Variance measures how far a data set is spread out:

$$\text{Sample Variance, } s^2 : \\ s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{Population Variance, } \sigma^2 : \\ \sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

Where: x_i = Value of each data point, \bar{x} = Mean, μ =mean (population), n = Number of data points

Variance cannot be negative. A zero value means that all of the values within a data set are identical.

If the variance is low that's mean the data collect near average, while If the variance is high the data will spread from the average.

Problem 1:

The heights (in cm) of students of a class is given to be 163, 158, 167, 174, 148. Find the variance.

Solution:

To find the variance, we need to find the mean of the given data and total members in the data set.

Total number of elements, $n = 5$

$$\bar{X} = (163+158+167+174+148)/5 = 162$$

$$s^2 = \frac{(163 - 162)^2 + (158 - 162)^2 + (167 - 162)^2 + (174 - 162)^2 + (148 - 162)^2}{5 - 1}$$

$$s^2 = \frac{(1)^2 + (-4)^2 + (5)^2 + (12)^2 + (-14)^2}{5 - 1}$$

$$\therefore s^2 = 95.5$$

Properties of Variance:

1. All values are used in the calculation.
2. It is not extremely influenced by outliers (non-robust).
3. The units of variance are awkward: the square of the original units.
4. Therefore, standard deviation is more natural since it recovers the original units.

For grouped data, variance can be estimated from the formula:

$$S^2 = \frac{\sum_{i=1}^n f_i(x_i - \bar{x})^2}{f_i - 1}$$

Where: f_i = frequencies and x_i = mid-point

Problem 2: calculate the variance of the following data from given table:

Classes	Frequencies
30-34	4
35-39	5
40-44	2
45-49	9
summation	20

Solution:

Classes	Frequencies (f_i)	Mid class	Fixi	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
30-34	4	32	128	41	32-41	$(-9)^2$	$4*(81) = 324$
35-39	5	37	185	41	37-41	$(-4)^2$	$5*(16) = 80$
40-44	2	42	84	41	42-41	$(1)^2$	$2*(1) = 1$
45-49	9	47	423	41	47-41	$(6)^2$	$9*(36) = 324$
summation	20		820				730

$$\begin{aligned} \text{Mean } \bar{x} &= \frac{\sum f_i \cdot x_i}{\sum f_i} \\ &= \frac{820}{20} = \mathbf{41} \end{aligned}$$

$$S^2 = \frac{\sum_{i=1}^n f_i(x_i - \bar{x})^2}{f_i - 1}$$

$$S^2 = \frac{730}{20 - 1}$$

$$S^2 = 38.4$$

Standard deviation: S or SD

Standard deviation is a **measure of dispersion in statistics**. "Dispersion" tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the mean or average.

It is the most robust and widely used measure of dispersion since, unlike the range, it takes into account every variable in the dataset.

For example, are all your scores close to the average? Or are lots of scores way above (or way below) the average score?

When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large. **The standard deviation is usually presented in conjunction with the mean and is measured in the same units.**

Sample Standard Deviation, s :

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

Population Standard Deviation, σ :

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

S or SD = $\sqrt{S^2}$

Now for problem 2 the standard deviation is the square root of 38.4 equal 6.198

It is very easy now to describe the dataset by knowing mean and standard deviation

So, mean \pm SD = **41 \pm 6.198**

Coefficient of Variation: (CV)

Also known as the relative standard deviation (RSD), The co-efficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The co-efficient of variation represents the **ratio of the standard deviation to the mean**, and it is a useful statistic for **comparing the degree of variation from one data series to another**, even if the means are drastically different from one another.

$$CV = \frac{SD}{\bar{x}} * 100$$

Problem 3: calculate CV of the following data, (23, 35, 56, 35, 77)

Solution: it must calculate both mean and Standard deviation to calculate CV

$$\text{Mean } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$= (23+35+56+35+77)/5$$

$$= 45.2$$

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

$$SD = \sqrt{\frac{(23 - 45.2)^2 + (35 - 45.2)^2 + (56 - 45.2)^2 + (35 - 45.2)^2 + (77 - 45.2)^2}{5 - 1}}$$

$$SD = 21.38$$

$$CV = \frac{SD}{\bar{x}} * 100$$

$$CV = \frac{21.38}{45.2} * 100$$

$$CV = 47.3$$

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به تسجيل الدخول

39.5814
=STDEV.S(
STDEV.S(number1; [number2]; ...)

Age (year)	43	54	54	44	55	49	45	48	56	49	44
	56	59	43	53	63	52	48	56	55	59	48
	46	47	54	59	46	45	48	51	55	48	57
	38	56	58	57	48	55	54	62	49	43	35
Urea (mg/dL)	15.7	16	26.4	18	19.2	26.3	25	40.6	13.5	25.9	23.9
	25.8	23.1	33	21.1	36	27	27.2	28.2	26.1	19.6	36
	39	21	30	32	31	26	29	27	21	27	29
	33	32	38	34	29	25	27	29	28	34	37

39.5814
40.35493

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به تسجيل الدخول

B2 =STDEV.S(B2:L5)

Age (year)	43	54	54	44	55	49	45	48	56	49	44
	56	59	43	53	63	52	48	56	55	59	48
	46	47	54	59	46	45	48	51	55	48	57
	38	56	58	57	48	55	54	62	49	43	35
Urea (mg/dL)	15.7	16	26.4	18	19.2	26.3	25	40.6	13.5	25.9	23.9
	25.8	23.1	33	21.1	36	27	27.2	28.2	26.1	19.6	36
	39	21	30	32	31	26	29	27	21	27	29
	33	32	38	34	29	25	27	29	28	34	37

39.5814
=STDEV.S(B2:L5)
STDEV.S(number1; [number2]; ...)

40.35493

New Microsoft Excel Worksheet - Excel

ملف الشريط الرئيسي إدراج تخطيط الصفحة الصغ بيانات مراجعة عرض تعليمات أجزئي بما تريد القيام به تسجيل الدخول

B2 =STDEV.S(B2:L5)

Age (year)	43	54	54	44	55	49	45	48	56	49	44
	56	59	43	53	63	52	48	56	55	59	48
	46	47	54	59	46	45	48	51	55	48	57
	38	56	58	57	48	55	54	62	49	43	35
Urea (mg/dL)	15.7	16	26.4	18	19.2	26.3	25	40.6	13.5	25.9	23.9
	25.8	23.1	33	21.1	36	27	27.2	28.2	26.1	19.6	36
	39	21	30	32	31	26	29	27	21	27	29
	33	32	38	34	29	25	27	29	28	34	37

39.5814
=STDEV.S(B2:L5)
STDEV.S(number1; [number2]; ...)

40.35493

Correlation and Dependence

Correlation is a statistical relationship between two variables or data sets that measures the degree to which they move in relation to each other. It is a measure of the strength and direction of the linear relationship between two variables. If two variables are positively correlated, they move in the same direction, meaning that when one variable increases, the other variable also tends to increase. If two variables are negatively correlated, they move in opposite directions, meaning that when one variable increases, the other variable tends to decrease. Correlation can be calculated using various methods, such as the Pearson correlation coefficient, which measures the linear relationship between two variables, or the Spearman rank correlation coefficient, which is more robust and sensitive to nonlinear relationships. Correlation is a useful tool in statistics because it can indicate a predictive relationship that can be exploited in practice. However, correlation does not imply causation, meaning that a correlation between two variables does not necessarily mean that one variable causes the other to change.

Dependence, on the other hand, signifies a variable whose value depends on the value assigned to another variable (independent variable). While correlation focuses on linear relationships, dependence encompasses a wider range of statistical relationships beyond linearity. Various statistical measures, such as Spearman's rank correlation and Kendall's coefficient, have been developed to capture different aspects of dependence between variables, especially when dealing with nonlinear relationships.

The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It can range from 0 to 1 or -1 to 0, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no linear correlation. The correlation coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations.

The Pearson correlation coefficient (R) or r: is the most commonly used type of correlation coefficient and measures the linear relationship between two variables.

It can be calculated as follows:

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

There is another formula to calculate r but the above is easier.

To evaluate the correlation coefficient after calculation:

r=1: Perfect positive correlation. This means that as one variable increases, the other variable also increases in a perfectly linear fashion.

r=0.8 to r=0.99: Strong positive correlation. This indicates a strong linear relationship between the variables.

r=0.5 to r=0.79: Moderate positive correlation. The variables are moderately related in a linear manner.

r=0: No correlation. There is no linear relationship between the variables.

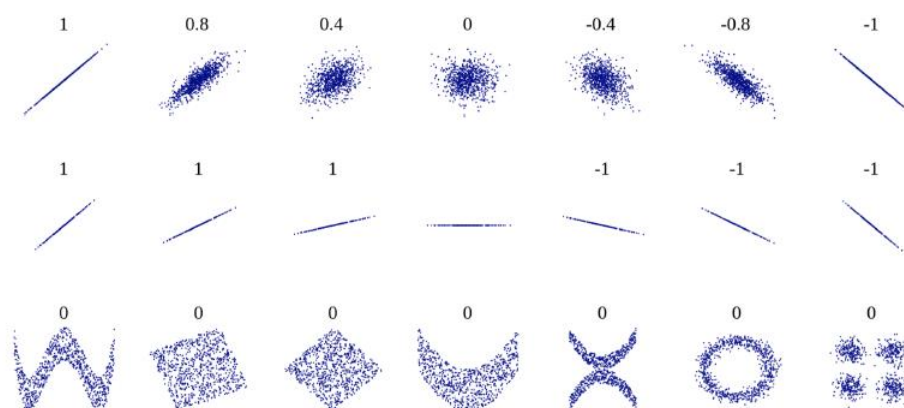
r=-0.5 to r=-0.79: Moderate negative correlation. The variables are moderately related in a negative linear manner.

r=-0.8 to r=-0.99: Strong negative correlation. This indicates a strong linear relationship in the negative direction.

r=-1: Perfect negative correlation. As one variable increases, the other variable decreases in a perfectly linear fashion.

The following figures indicate the types of correlation

Pearson Correlation



However, correlation does not imply causation, meaning that a correlation between two variables does not necessarily mean that one variable causes the other to change.

The Pearson correlation coefficient is most effective when used with data that is jointly normally distributed.

Spearman's rank correlation coefficient is a nonparametric measure of rank correlation, which assesses the statistical dependence between the rankings of two variables. It is named after Charles Spearman and often denoted by the Greek letter ρ . Spearman's correlation coefficient measures how well the relationship between two variables can be described using a monotonic function, whether linear or not. It is appropriate for both continuous and discrete ordinal variables.

It can be calculated from the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

Spearman's rank correlation coefficient is suitable for:

1. non-parametric measure of rank correlation that assesses the strength and direction of the relationship between two variables, whether linear or not.
2. It is a good option when the data is not normally distributed, or when the relationship between the variables is not linear, and it is not affected by the range of the data or the presence of ties.
3. It is a good option for assessing test-retest reliability for ordinal data.

Example: check if there is a correlation or not between the reactants S quantity and yielding products H_2SO_4 from the following data:

Reactants (grams)	20	22	24	26	28	30	32
Product Yield (grams)	14	17	19	19.5	21	24	25

	x	y	xy	x ²	y ²
	20	14	280	400	196
	22	17	374	484	289
	24	19	456	576	361
	26	19.5	507	676	380.25
	28	21	588	784	441
	30	24	720	900	576
	32	25	800	1024	625
Σ	182	139.5	3725	4844	2868.25
	x ² =33124	y ² =19460.25			

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


$$r = \frac{3725 - \frac{182 * 139.5}{7}}{\sqrt{(4844 - \frac{33124}{7})(2868.25 - \frac{19460.25}{7})}}$$

$$r = \frac{98}{\sqrt{112 * 88.21}}$$

$$r = 0.986$$

x	y					
20	14					
22	17					
24	19					
26	19.5					
28	21					
30	24					
32	25					

=corr

 CORREL

ارجاع معامل الارتباط بين مجموعتين من البيانات

x	y					
20	14					
22	17					
24	19					
26	19.5					
28	21					
30	24					
32	25					

=CORREL(

CORREL(array1; array2)

x	y
20	14
22	17
24	19
26	19.5
28	21
30	24
32	25

=CORREL(A2:A8
CORREL(array1; array2)

x	y
20	14
22	17
24	19
26	19.5
28	21
30	24
32	25

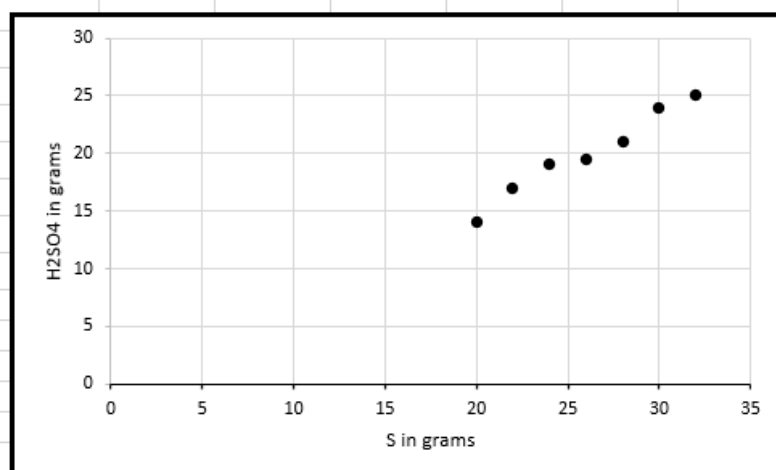
=CORREL(A2:A8;B2:B8)
CORREL(array1; array2)

fx =CORREL(A2:A8;B2:B8)

A	B	C	D	E	F
x	y				
20	14				
22	17				
24	19		0.985933		
26	19.5				
28	21				
30	24				
32	25				

The results of this example indicate there is an excellent positive correlation between the independent variable (quantity of S in grams, and the yielding product (H₂SO₄ in grams))

x	y
20	14
22	17
24	19
26	19.5
28	21
30	24
32	25



Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship between the variables, which can be used to make predictions about the dependent variable based on the independent variables.

An equation of linear relationship can be used to describe many variants but in our lecture we will deal with a simple linear regression that means a first order equation connect between y and x only.

The equation for a simple linear regression model is often written as:

$$y = \alpha x + \beta$$

Y is the dependent variable (the variable we want to predict or explain).

X is the independent variable (the variable we use to make predictions or explain variations in(y)).

α is slope (the change in Y for a one-unit change in X).

β is the intercept (the value of Y when X is 0).

Linear regression is a powerful statistical tool for modeling the relationship between a dependent variable and one or more independent variables. It allows us to make predictions about the dependent variable based on the independent variables, and to make inferences about the population parameters based on the sample data.

The coefficient of determination, also known as R-squared or R^2 , is a measure of the goodness of fit of a regression model. The R^2 value ranges from 0 to 1, with higher values indicating a better fit of the regression line to the data. An R^2 value of 1 means that all the variation in the dependent variable is predictable from the independent variable(s), while an R^2 value of 0 means that none of the variation in the dependent variable is predictable from the independent variable(s).

Here some examples of equation between y versus x

1. In chemistry, linear regression is commonly used to model the relationship between variables in various analytical methods. One of the best chemical equations that exemplify linear regression is the Beer-Lambert Law, which relates the concentration of a substance in a solution to its absorbance. The equation is: $A = \epsilon \cdot b \cdot C$

Where:

A is the absorbance of the solution,

ϵ is the molar absorptivity of the substance,

b is the path length of the cuvette (usually in cm),

c is the concentration of the substance in the solution (usually in mol/L).

concentration mol/L	Absorbance
0	0
3	0.05
6	0.09
9	0.18
12	0.22
15	0.34
18	0.46
21	0.5
24	0.68

New Microsoft Excel Worksheet - Excel

إدراج | تخطيط الصفحة | المصغ | بيانات | مراجعة | عرض | تعليمات | أيقوني بما تريد القيام به

الشريط الرئيسي | إدراج | تخطيط الصفحة | المصغ | بيانات | مراجعة | عرض | تعليمات | أيقوني بما تريد القيام به

الحصول على الوظائف الإضافية | الوظائف الإضافية الخاصة بي | رسومات توضيحية | جدول PivotTables | جدول

الوظائف الإضافية | الوصفي بها | الموصي بها | المخططات | المخططات | الموصي بها

مستتر

خريطة ثلاثية الأبعاد | خرائط | PivotChart | خرائط ثلاثية الأبعاد

خط عمود ربح / خسارة | خطوط المؤشر | خطوط المؤشر

مقسم الخطوط | طريقة العرض الزمني | عوامل تصفية

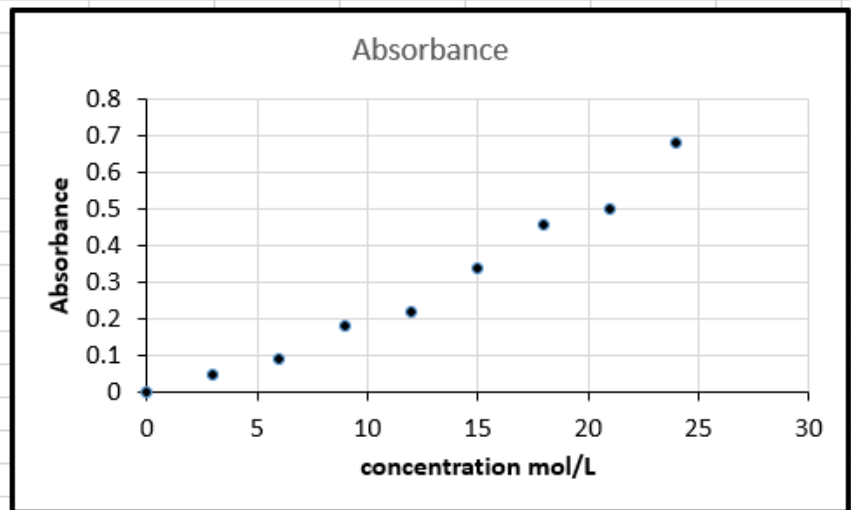
ارتباطات

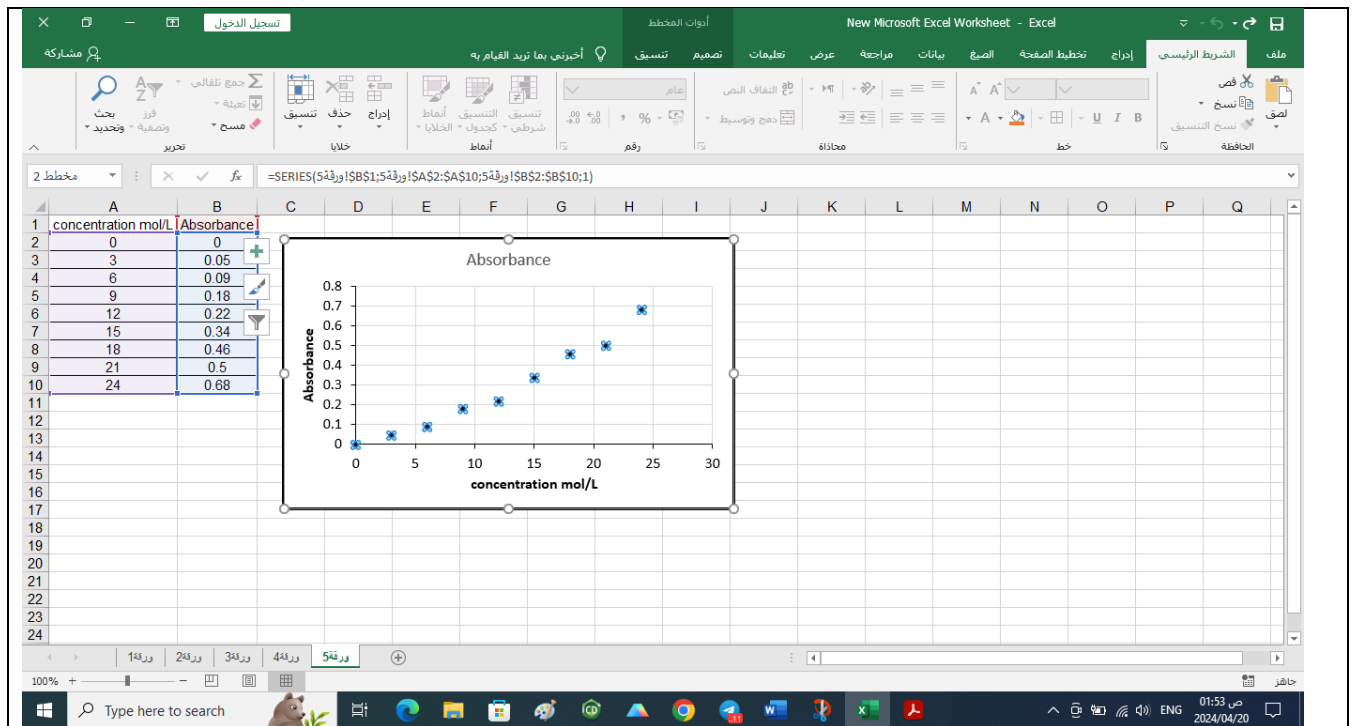
رموز | نصي | ارتباطات

concentration mol/L

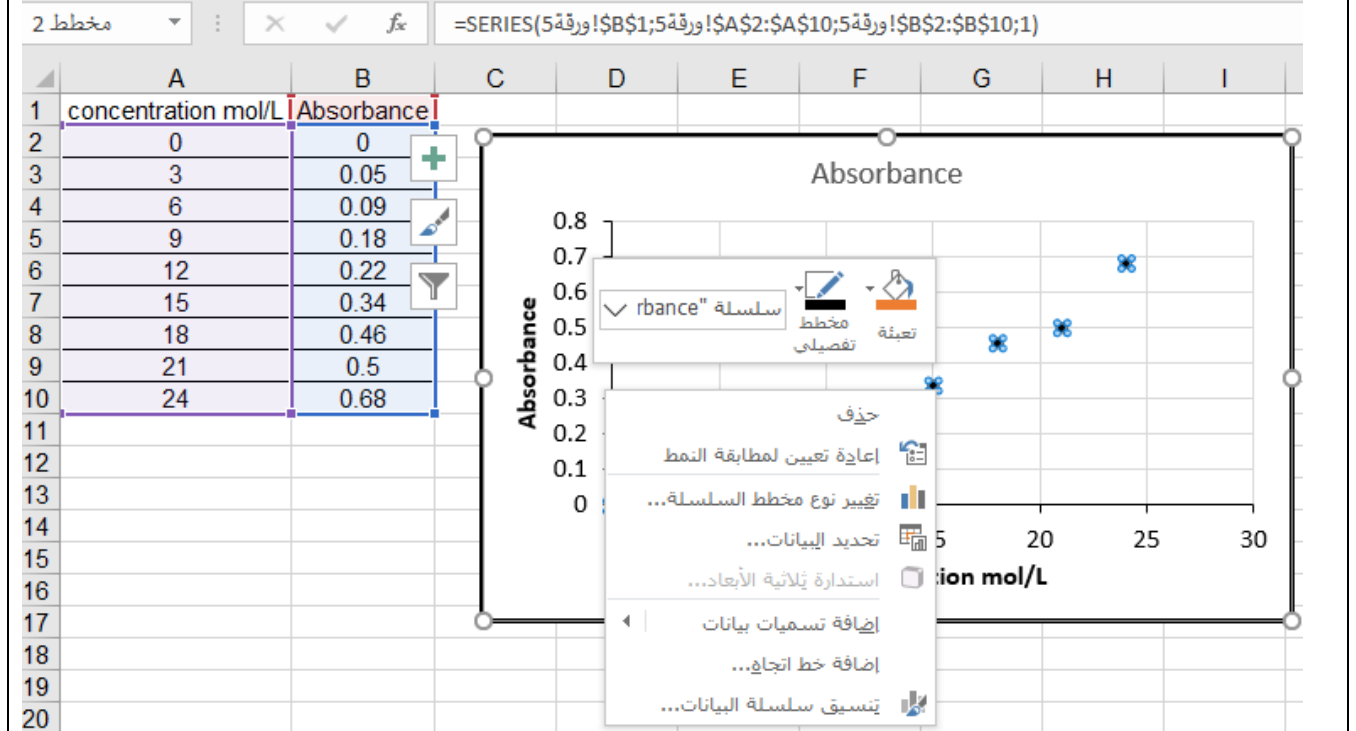
A	B	C	D	E	F	G
concentration mol/L	Absorbance					
0	0					
3	0.05					
6	0.09					
9	0.18					
12	0.22					
15	0.34					
18	0.46					
21	0.5					
24	0.68					

concentration mol/L	Absorbance
0	0
3	0.05
6	0.09
9	0.18
12	0.22
15	0.34
18	0.46
21	0.5
24	0.68

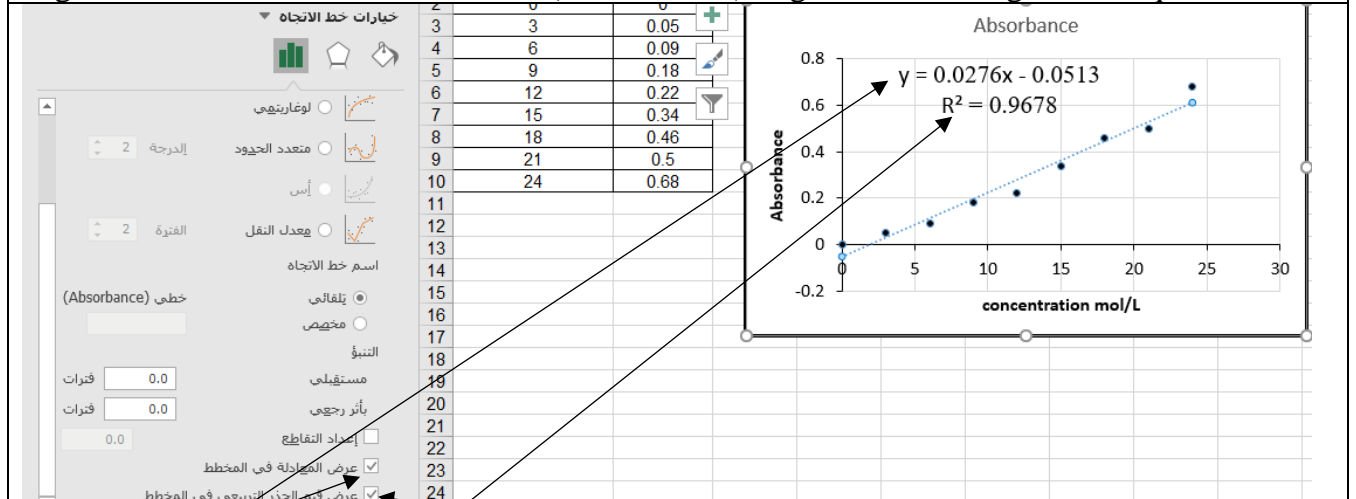




Put pointer on the data points and



Right click then choose add trendline (إضافة خط اتجاه) to get the linear regression equation



Show equation show R^2 is a measure of the goodness of fit of a regression model

2. Gas Laws: Gas laws describe the behavior of gases under different conditions. For example, the ideal gas law is often written as: $PV=nRT$

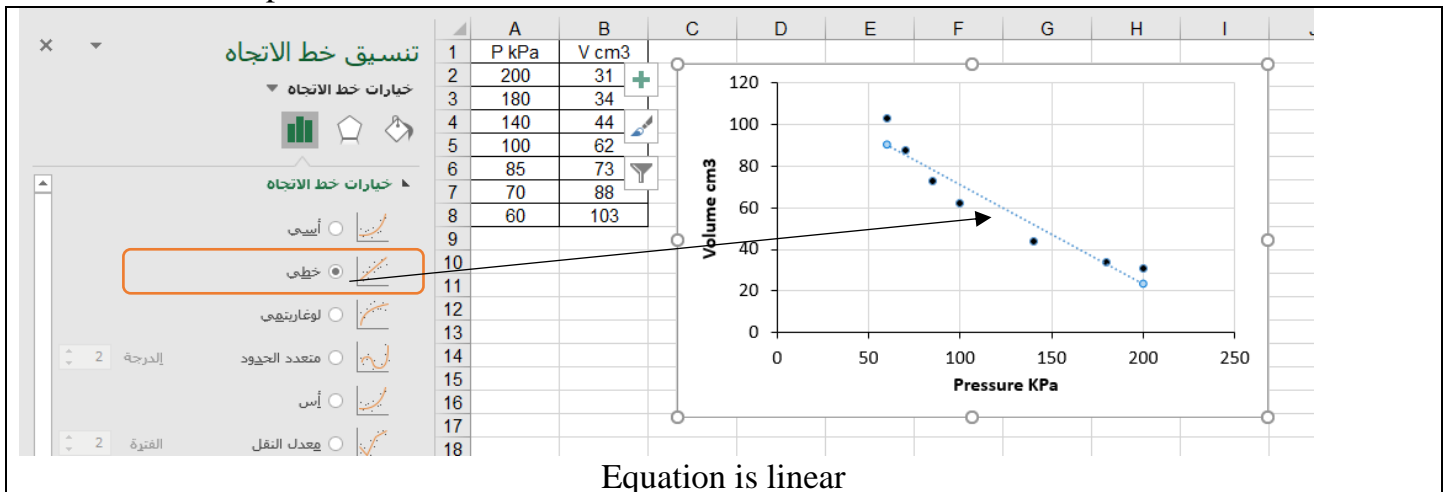
P is the pressure of the gas.

V is the volume of the gas.

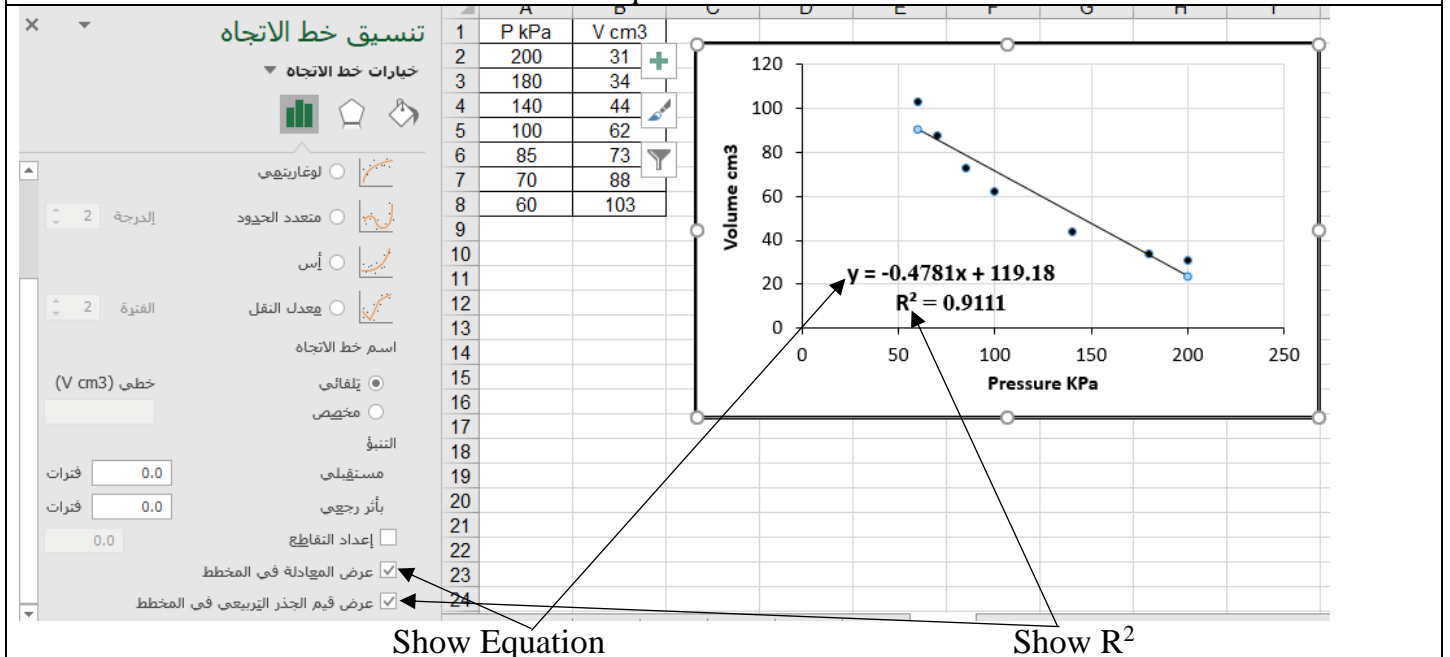
n is the number of moles of gas.

R is the ideal gas constant.

T is the temperature in Kelvin.

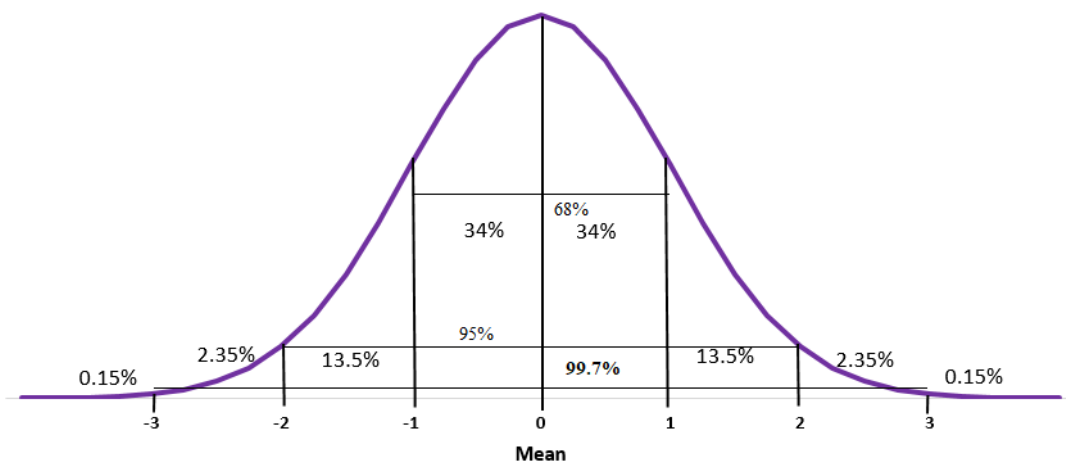


Equation is linear



Normal Distribution

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.



Null hypothesis and Alternative hypothesis

The null hypothesis and alternative hypothesis are statistical concepts used in hypothesis testing to determine whether a given phenomenon or relationship exists in a population based on a sample of data.

Null hypothesis in correlation means there is no correlation between two variable series of study, while if there is a correlation between them that means the null hypothesis is rejected and we should follow the alternative hypothesis. In case of comparison between two groups, null hypothesis means; there is no significant differences.

significant differences: the difference comes by effect of some reason, not occurred by chance.

Student t test

A statistical test used to determine if there is a significant difference between the means of two groups.

The t-test is a statistical technique used to determine whether there is a significant difference between the means of two groups. It is commonly used in hypothesis testing to compare the means of two samples, typically to determine if there is a statistically significant difference between the means of two groups.

It is known a Student's t-test also. To do the test, there are some conditions:

1. The variables are quantitative.
2. The data must be small.
3. The data obey normal distribution.

In case of other conditions, we cannot apply t test. Instead, we almost should apply ANOVA.

To do the comparison and apply t test, we will calculate the value of t ($t_{\text{calculated}}$), then compared it with the value of t from tables ($t_{\text{tabulated}}$). If $t_{\text{calculated}}$ is larger than $t_{\text{tabulated}}$, the difference between groups is significant.

Now let us explain how to estimate the two values.

Firstly, to calculate the value of t calculated, we should use the formula:

$$t_{\text{calculated}} = \frac{|\bar{x}_1 - \bar{x}_2|}{S_{\text{pooled}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

\bar{x}_1 : the mean of group 1,

\bar{x}_2 : mean of group 2,

n_1 : number of group 1 observations.

n_2 : number of group 2 observations

S_{pooled} is the pooled standard deviation, that is a critical component used to estimate the standard deviation of the difference between the means of two independent samples.

The formula for the pooled standard deviation (S_{pooled}) is as follows:

$$S_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Secondly, we should estimate $t_{\text{tabulated}}$ from given tables. The above row represents the confidence level, while the first column represents the degree of freedom, as shown below.

Before learning how to calculate $t_{\text{tabulated}}$, we should understand the meaning of significant differences.

The significance is an important criterion depends on confidence level. Confidence limits quantify the uncertainty in estimating a population parameter from a sample. They allow researchers to make inferences about the population while accounting for sampling variability. The confidence level (e.g., 95%) indicates the probability that the confidence interval contains the true population parameter. A 95% confidence level means there is a 95% chance the true parameter falls within the confidence limits. A 95% confidence level means the probability is equal or less than 5% (0.05). **The probability cannot be avoided, and cannot be treated but it can be estimated. So, confidence level of 99% means propability is ≤ 0.01 (1%).**

Now to estimate $t_{\text{tabulated}}$, using a table of t values, we calculate the tabular t value at certain confidence level (i.e. probability level), and then by the intersection of the confidence level with the degrees of freedom, as shown below:

To calculate the value of $t_{\text{tabulated}}$ at 95% confidence level and the degree of freedom equal 10; the value from table is: **2.228**.

Degree of freedom: is the maximum number of logically independent values, which may vary in a data sample.

In two Sample t-test: The degrees of freedom are calculated as the sum of the degrees of freedom for each sample, which is the sample size minus one for each sample, i.e., $df = n_1 + n_2 - 2$.

Degree of freedom (n-1)	Confidence level 90%	Confidence level 95%	Confidence level 99%	Confidence level 99.9%
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.500	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
15	1.753	2.131	2.947	4.073
20	1.725	2.086	2.845	3.850

Example 1: A group of 5 patients treated with medicine A is of weight 42,39,38,60 &41 kgs. Second

group of 7 patients from the same hospital treated with medicine B is of weight 38, 42, 56, 64, 68, 69, & 62 kgs. Find whether there is a significant difference between medicines at probability level of $p \leq 0.05$?

Solution:

From data we can summarize the estimated data

n1	\bar{x}_1	S_1^2	SD ₁	n2	\bar{x}_2	S_2^2	SD ₂
5	44	82.5	9.08	7	57	154.33	12.42

Data size is small. Now let us examine the normal distribution: $2SD < \bar{x}$, $2(9.08) < 44$ and $2(12.42) < 57$

\therefore both groups are normally distributed.

$$t_{calculated} = \frac{|\bar{x}_1 - \bar{x}_2|}{S_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad S_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

$$S_{pooled} = \sqrt{\frac{82.5(5-1) + 154.33(7-1)}{5+7-2}} = 11.21$$

$$t_{calculated} = \frac{|44-57|}{11.21} \sqrt{\frac{5*7}{5+7}} = 1.832$$

From table the intersection of the confidence level (95%) with the degrees of freedom 10: t tabulated is 2.228

\therefore t calculated < t tabulated

∴ the difference is not significant

Example 2: Measurements of uric acid (mg/dL) were collected for a group of 7 diabetic patients and a group of 7 healthy people as well, and they were as follows in the table:

SD	mean	
0.66	5.71	Healthy people
0.86	6.81	Diapetic patients

Explain whether there are significant differences between the two groups through the t test, on the confidence level of 95% and 99%, taking advantage of the attached table.

Degree of Freedom

$$n_1+n_2-2=12$$

Degrees of Freedom (n-1)	$\alpha = 0.20$	0.10	0.05	0.02	0.01	0.002
1	3.078	6.314	12.706	31.821	63.657	318.300
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	3.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.305	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.746	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733

Solution:

∴ the data obey normal distribution (2SD<mean for both group) and size of data is small

$$t_{calculated} = \frac{|\bar{x}_1 - \bar{x}_2|}{S_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t_{calculated} = \frac{|5.71 - 6.81|}{0.766551} \sqrt{\frac{7 * 7}{7 + 7}}$$

= 2.685

$$S_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

$$S_{pooled} = \sqrt{\frac{(0.66)^2(7-1) + (0.86)^2(7-1)}{7+7-2}}$$

= 0.766551

95% ($\alpha=0.05$)

t calculated	t tabulated
2.685	2.179
∴ t calculated (2.685) > t tabulated (2.179)	
∴ there is a significant difference	

99% ($\alpha=0.01$)

t calculated	t tabulated
2.685	3.055
∴ t calculated (2.685) < t tabulated (3.055)	
∴ there is no significant difference	

ANOVA "Analysis of variance"

Statistical method for analyzing differences among means. **ANOVA** can be defined also as a collection of statistical models and associated estimation procedures used to analyze the differences among group means. It is developed by Ronald Fisher.

In inferential statistics, if the study requires a comparison between two small-sized groups and the values are normally distributed, we use the t test, but if the comparison is among more than two groups, it is not suitable to use the t test between each of the two groups, it is inappropriate in terms of complexity of calculations, and the error rate will be high.

Therefore, the scientist Ronald Fisher developed a statistical analysis method that fits these requirements for comparing averages and is based on the variance among groups. That is, the analysis is performed directly and compares among the study groups, as well as this method is convenient to compare between two groups if the t test conditions are not met.

Dependent variable: This is the item being measured that is theorized to be affected by the independent variables.

Independent variable/s: These are the items being measured that may have an effect on the dependent variable.

A null hypothesis (H₀): This is when there is no difference between the groups or means. Depending on the result of the ANOVA test, the null hypothesis will either be accepted or rejected.

An alternative hypothesis (H₁): When it is theorized that there is a difference between groups and means.

Factors and levels: In ANOVA terminology, an independent variable is called a factor which affects the dependent variable. Level denotes the different values of the independent variable that are used in an experiment.

Types of ANOVA

One-way ANOVA: Compares the means of three or more independent groups.

Two-way ANOVA: Analyzes the effects of two independent variables on a dependent variable.

MANOVA (Multivariate ANOVA): Extends ANOVA to situations with multiple dependent variables.

One-way ANOVA

The one-way analysis of variance is also known as single-factor ANOVA or simple ANOVA. As the name suggests, the one-way ANOVA is suitable for experiments with only one independent variable (factor) with two or more levels. For instance, a dependent variable may be what month of the year there are more flowers in the garden. There will be twelve levels. A one-way ANOVA assumes:

Just as the t test depends on the calculations on the S pooled and the degrees of freedom, so the ANOVA test depends on the sum of the squares SS from which the variance is calculated and the degrees of freedom.

Example 1:

To get complete comparison in marks of students in English Language in all stages of our college, we cannot use t test because we have more than two groups. So, ANOVA is the suitable analysis.

The result of comparing using ANOVA will show whether there are significant differences or not among the groups depending on a probability level, almost $p \leq 0.05$.

Example 2:

In the laboratory if we are looking for differences among groups of students, the ANOVA is the convenient analysis.

Example 3:

Comparison between two groups but the data is not small, the ANOVA is the convenient analysis.

Example 4:

Comparison between two groups but the data is not under normal distribution, the ANOVA is the convenient analysis.

To know which group is the reason of significant difference, there are other tests beyond ANOVA must be done. These analyses are called Post-hoc Tests.

If the ANOVA rejects the null hypothesis, post-hoc tests can be used to determine which specific group means differ.

Examples of post-hoc tests include Tukey's HSD (Honestly Significant Difference), least significant difference (LSD), Duncan, and Bonferroni correction.

In summary, ANOVA is a powerful statistical technique used to compare the means of multiple groups and assess the effects of independent variables on a dependent variable, while accounting for the variability within and between groups