BIOSTATISTICS LECTURE NOTES

1. INTRODUCTION

When we say statistics we know that statistic has more than one meaning. When we say statistics, we understand;

1. Collecting, organizing, and summarizing data,
2. Calculation of descriptive statistics for the observed data,
3. If data is collected for more than one feature, investigation of the relationship between them,
4. Analysing the observed data according to the factors if there are factors taken into consideration in the study,
5. All the methods used for the interpretation, evaluation and generalization of the results obtained are called statistics (methods).

As science, statistics are divided into two categories: theoretical (mathematical) and applied statistics. Mathematical statistics is concerned with the study and development of statistical theory and methods in abstract. Applied statistics deals with the application of statistical methods to solve real problems using the collected data or randomly generated data, and the development of new statistical methodology.

Roughly speaking, statistics is a field of study concerned with the collection of data, organization, analysis of data, interpretation of the results, generalization of the results, and the drawing inferences about a body of data when only a part of data is used.

No matter in which field a researcher works he/she must analyse the data related to the topic statistically and to interpret the results obtained. Therefore, the tools of statistics are employed in many fields, such as agriculture, education, health sciences, etc.

Biostatistics is the branch of applied statistics. That is why it is called biostatistics is because it deals with biological data and consists of methods mainly applied to the biological data. In other word, when the data is obtained from the biological sciences and medicine, the term of "biostatistics" is used.

There are basic concepts that need to be known in order to understand and be successful. Now let's look at these concepts in turn:

2. BASIC CONCEPTS of BIOSTATISTICS

DATA: The raw material of statistics is DATA. Data are the numerical expression of the variables being studied.

In order to obtain reliable results in a given study, data must be collected in accordance with the subject. In practice, data can be collected from different sources:

1. Regularly kept records,
2. From the questionnaires,
3. From published sources,
4. From the experience,
5. From surveys,

VARIABLE: A characteristic being studied that takes different values for different entities is called variable. A variable can be quantitative or qualitative, categorical or numerical, depending on its possible values. To study a variable statistically, there must be variability amongst individual in terms of the variable being studied. For example, the number of eyes of the students cannot be studied statistically because each student has two eyes, except for any health problem. In other words, there is no variability in terms of the number of eyes. The variables such as success grades of any course, weight of students, amount of any vitamin in vitamin syrup, and heights can be examined statistically because having variability amongst entities.

POPULATION: A population is all the objects (individual, entity) of interest. Object in definition can be everything depending on working area, such as people, student, machines, cells, etc. For example, if a researcher is interested in amount of vitamin-B2 in the bottles of vitamin syrup, population includes all the bottles.

POPULATION SIZE (N): The number of individual in the population is called the population size and is denoted by N.

Example:

Let's assume that you are planning to carry on a study related to the students of the University of Ankara. In this case you mean that you are dealing with all the students of the University of Ankara and all of your students constitute your population. Collecting data from all of the students is very difficult and often impossible. Because this will be constrained by the time and financial means of the investigator. In some branches of science, the feature studied will prevent data collection from all individuals. For example, a pharmacist cannot use all of the vitamin syrup bottles to analyse the amount of any vitamin produced in the vitamin syrup. In this case, the researcher should select a random sample to represent the population.

SAMPLE: A sample contains a limited number of individuals that are randomly selected from the population and represents it. On other word, a sample is a part of population, which is representative of it. If the individuals are selected randomly, random sampling means that each individual in the population has equal probability of being selected.

SAMPLE SIZE (n): The number of elements in the sample is called the sample size and is denoted by n.

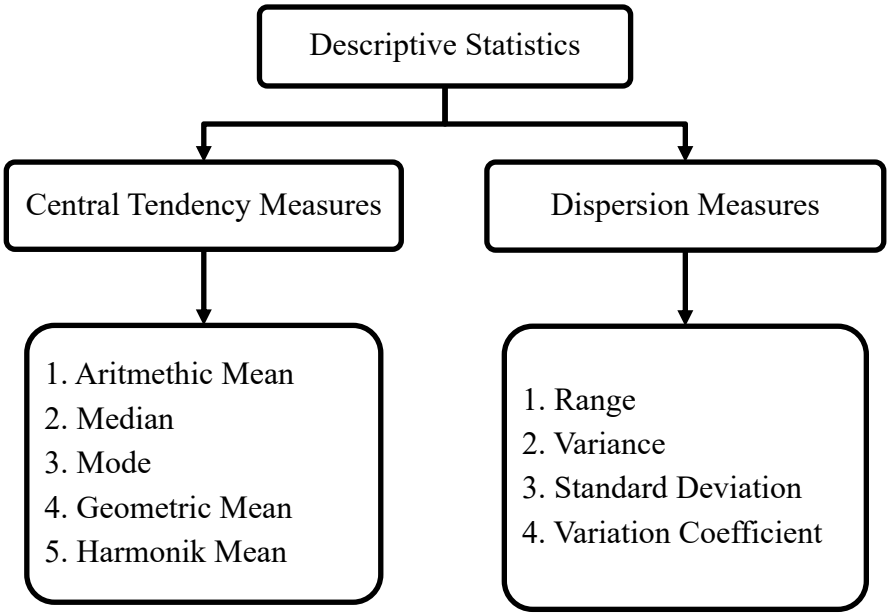
PARAMETER: A parameter is a numerical value that is calculated from the population data and summarizes some aspects of it as a whole. In general, a parameter is known in general but a researcher is interested to know true values of the parameters.

STATISTIC: A statistic is a numerical value computed from the sample data. It summarizes some aspects of the sample. A statistic is the prediction of the parameter. The statistics calculated from the sample must be reliable. In order to ensure your reliability,

1. All the individuals in the sample should be selected by chance from the population.
2. The individuals in the sample should be homogeneous in terms of all the factors that affect the variables being studied.
3. The sample size should be increased as much as possible. As the sample size increases, the reliability of the calculated statistics increases. However, care must be taken to ensure that homogeneity is not impaired while the sample size is increased.

3. DESCRIPTIVE STATISTICS

Experiments are carried to inspect what is going on around us and enable researcher to have information on the variable being studied. A sample is constituted and data on the variable which is the raw material of statistic is collected from the experiments carried out. When the data is collected, the researcher wants to describe the sample better, to learn more and to express it better. In this case, descriptive statistics need to be calculated. Descriptive statistics give better information about the sample studied. Descriptive statistics can be summarized as follows:



3.1. Central Tendency Measures

Measure of central tendency is a single value that is considered to be representative of the data set as a whole. Measures of central tendency provide information on the average value of the data. Data incline to gather around the measures of central tendency. This supply information on the value which all data aggregated (gather up) around.

3.1.1 Arithmetic Mean:

The mean or arithmetic mean is calculated by adding all the values in a sample or population and dividing them by the number of values. In other words, mean is calculated by dividing the sum of all the data by the number of observations. If it is calculated from the population data, it is called parameter and calculated as;

$$\mu_X = \frac{\sum_{i=1}^{N} X_i}{N}$$

When it is calculated from the sample data, it is referred as statistic and calculated as;

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

The mean is the best representative value of the sample and it summarizes the properties of the collected data into a single value. There is only one mean for a sample. The calculation of it is very easy and takes all of the collected data into account. Mean is greatly affected by outliers in the collected data because outliers pull the mean towards itself and give a misleading result.

Example:
Suppose that our collected data from a sample whose size is 5 is: 5, 6, 7, 5, 8. Then sample mean is found as;

$$\overline{X} = \frac{5+6+7+5+8}{5} = 6.2$$

The calculated sample mean gives the information that the collected data shows a tendency to gather around the calculated value. Everybody should keep in mind that the calculated mean cannot be smaller than the smallest value in the collected data, nor greater than the largest value.

3.1.2. Median

The median value is the middle value of the data ordered by their size. The median divides the ordered data into two equal parts. Half of the data are equal to or less than the median value, while the other half is equal to and greater than the median value. If the sample size (n) is odd, ((n + 1) / 2)th value, which is the middle value of the ordered data, is the median value. If the sample size is even number, the mean of $(n / 2)^{th}$ and $((n / 2) + 1)^{th}$ values, which are the two middle values, is the median value. The median is not as severely influenced by extreme values as is the mean.

Suppose that our collected data from a sample whose size is 5 is: 5, 6, 7, 5, 8.

To find the median of the data for the collected data is first ordered. Then the middle value is determined as the median value. So, the median of the sample is;

5 5 6 7 8 after that median equals to 6.

If the data have very small or very large values compared to others, the median value must be calculated for the sample. Because it is affected by outliers and is misleading. For example, if a set consists of values: 125, 120, 130, 140, and 900. Then the value 900 is an outlier. Outliers can make mean value considerably unreliable. For example, the mean of the data is 1415/5=283, which is unreliable and the median is 130. In this case the median value most properly summaries the data, better than the mean.

3.1.3. Mode

The mode of the data is the value which occurs most frequently. In other words, the mode of the sample is the value with highest frequency. For example, the mode of the X= 6, 7, 7, 5, 7, and 8 is 7 since it has the highest frequency.

There might be no mode in the collected data while the data collected on some variable can have more than one mode.

The ordering of these measures in the XY coordinate axis according to their size gives information about the distribution shape of the data. If the arithmetic mean, median value and mode value calculated for a sample are equal to each other, we have the knowledge that the collected data show bell-shaped symmetrical distribution, which is normal distribution (Figure 1) which is normal distribution. We should keep in mind that if all the measures of tendency are approximately equal to each other, distribution of data is symmetric. If the mean is the largest of the central tendency measurements, the distribution of the data is skewed to the right (Figure 3), while if the mean is the smallest in these measurements, it indicates that the distribution is skewed to the left (Figure 2).

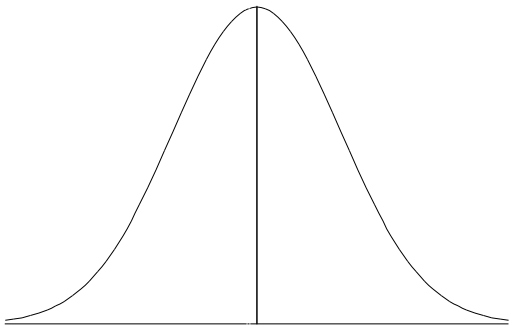


Figure 1. Normal (symmetric) distribution

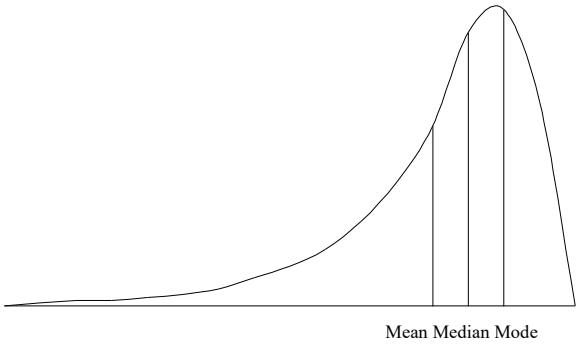


Figure 2. Distribution skewed to the left

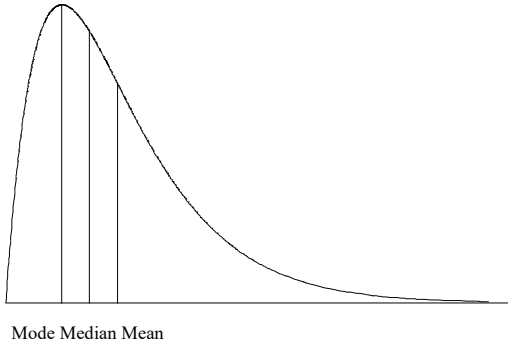


Figure 3. Distribution skewed to the right

3.1.4. Geometric Mean

The geometric mean of n positive values is the n th root of their product and is calculated as follows:

$$G.O = \sqrt[n]{x_1 x_2 x_3 ... x_n}$$

It is calculated for geometrically increasing data.

3.1.5. Harmonic Mean

Harmonic mean is the inverse of the mean of the inverse of the data. And it is calculated as follows:

$$H.O = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$$

$$= \frac{n}{\sum \frac{1}{x_i}}$$

For example, if 3 cars are to go from Ankara to Istanbul, the distance for the 3 cars is the same. If the mean speed of the cars is to be calculated, the harmonic average is calculated.

3.2. Dispersion Measures

In order to study the variable statistically, it should vary. In other words, there should be variability in terms of the variable studied. Therefore, a researcher wants to express the variability amongst individuals numerically. The values calculated for this purpose are called dispersion measures. Let's start to learn dispersion measures in turn:

3.2.1. Range

Range is the difference between the largest and smallest data in the population or sample, that is, it takes into account only two of all the data. Range just gives the difference the between the

largest and smallest data, but not information on variability amongst data. Therefore, it is not commonly used and not a preferable dispersion measure.

Example:
X: 7 11 10 12 15 range for X is 15-7=8
Y: 14 7 8 7 15 range for Y is 15-7=8

When we look at the calculated range, we have the same result for the variability between individuals in X and Y samples. Is that right for you? When we look at the examples, the data tells us that the variability in the sample Y is greater. But as we just said, the range does not tell us about the variability between the data, as it is the difference between the greatest and the smallest. This is why we do not use the rage.

Variance is the best and most important dispersion measure giving information about the variability between data.

3.2.2. Variance

Variance is the best and most important dispersion measure giving information about the variability in data. The variance takes account of the distance of data to the mean. When the data is close to the mean, the variance gets smaller, but the data is far from the mean, variance becomes larger. The variance of the data is zero (no variation) when all observations have the same value.

If the variance is calculated from the population data, it is parameter which is denoted by Greek letter and is calculated using the following formula:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(X_i - \mu_x)^2}{N}$$

If the variance is calculated from the sample data, it is statistic which is denoted by Latin letter and is calculated using the following formula:

$$S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum d_x^2}{n-1}$$

Where, $\sum d_x^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$ is sum of squares, (n-1) is degrees of freedom.

Example:
Let's calculate the variances for the samples X and Y used in Range

X: 7 11 10 12 15
Y: 14 7 8 7 15

$$\sum d_x^2 = (7^2 + 11^2 + 10^2 + 12^2 + 15^2) - \frac{(55)^2}{5} = 34$$
$$S_x^2 = \frac{34}{(5-1)} = 8.5$$

$$\sum d_y^2 = (14^2 + 7^2 + 8^2 + 7^2 + 15^2) - \frac{(51)^2}{5} = 62.8$$
$$S_x^2 = \frac{62.8}{(5-1)} = 15.7$$

3.2.3. Standard Deviation

Standard deviation is the square root of variance. Although the unit is not used with variance, the unit of standard deviation is the unit of measurement. The standard deviation gives a more accurate information about the dispersion of values in a population or sample.

If the standard deviation is calculated from the population data, it is parameter which is denoted by Greek letter and is calculated using the following formula:

$$\sigma_x = \sqrt{\sigma_x^2}$$

If the variance is calculated from the sample data, it is statistic which is denoted by Latin letter and is calculated using the following formula:

$$S_x = \sqrt{S_x^2}$$

3.2.4. Variation Coefficient

Generally, as the sample means grows, the standard deviation calculated gets larger. When the coefficient of variation is calculated, the effect of mean over the standard deviation is removed. The coefficient of variation is calculated dividing the standard deviation by mean and multiplying by 100. The variation coefficient expresses the variability observed in the cumulative data as%. If there are zero and / or negative values between the collected data, the coefficient of variation is not calculated. Because the coefficient of variation to be calculated is greater than 100%, which is impractical. The coefficient of variation is calculated using the following formula:

$$CV = \frac{S_x}{\overline{X}}.100$$

If the two sample having different arithmetic means needs to be compared in terms of variability, the coefficient of variation must first be calculated for each sample. In this way, the effect of arithmetic mean over standard deviation is removed. After that, two sample can be compared with each other in terms of variability.

Example:
In a study, it is found that the mean and the standard deviation of body length for anchovy sample are 10 cm and 1.5 cm, respectively. Moreover, the mean and the standard deviation of body length for perch sample are 75 cm and 9.5 cm, respectively. The researcher wonders which sample has greater variability in terms of body length. In such circumstances, if the two samples having different arithmetic means needs to be compared in terms of variability, the coefficient of variation must first be calculated for each sample.

$$\overline{A} = 10 \qquad \overline{P} = 75$$
$$S_A = 1.5 \qquad S_P = 9.5$$

To decide which sample has larger variability in terms of body length, the coefficient of variation for each sample is calculated. Therefore, the coefficients of variation are calculated as CV_A=15% and CV_P=12.7%. The coefficients of variation indicate that variability in terms of body length in anchovy sample is larger than perch sample.

Example:
Two pharmacists record the prescription number of insured patients who come to their pharmacies throughout a week. They calculate mean and standard deviation for number of prescriptions. They are given as follow:

$$\overline{P}_1 = 20 \qquad \overline{P}_2 = 50$$
$$S_{P1} = 8 \qquad S_{P2} = 12$$

If we want to know which pharmacy has greater variability in terms of number of prescription, we calculate coefficient of variation for each pharmacy. Because we know that while sample mean increases, sample variance increases too. Therefore, coefficients of variation are calculated as CV_{p1}=40% and CVp_2=24%. The coefficients of variation indicate that variability in terms of number of prescription in pharmacy 1 is larger than pharmacy 2.

Questions
1. Calculate the mode of this data set: 10, 4, 6, 4
a)7.0 b) 10.0 c) 5.0 d) 4.0

2. Calculate the median of this data set: 10, 3, 2, 4, 1
a) 2.5 b) 3.0 c) 2.0 d) 4.0

3. Calculate the standard deviation of this data set: 5, 5, 5, 5
a) 2.0 b) 4.0 c) 0.0 d) 5.0

4. Outliers tend to;
a) reduce the standard deviation,
b) equal the mean,
c) equal the median,
d) pull the mean away from the median toward the outliers,

5. A large sample standard deviation implies that the sample values are;
a) large b) small c) dispersed d) skewed

6. The median value of a set of data is the;
a) middle value of the ordered data
b) largest value middle value of the ordered data
c) most common value middle value of the ordered data
d) smallest value middle value of the ordered data

7. Which of the following is not a measure of central tendency?
a) mode b) median c) range d) mean

8. The mean of data set is:
a) The sum of the values divided by the number of values
b) Lower than the minimum value in the series
c) Lower than the maximum value in the series
d) An indicator of central tendency for the values of the series

9. Standard deviation:
a) is the square root of variance
b) is measured using the unit of the variable
c) is measured using the squared unit of the variable
d) has values generally comparable with the average value

10. If the mean of a series of values is 10 and their variance is 4, then the coefficient of variation is:
a) 40% b) 20% c) 80% d) 10%

11. The median of a series of numerical values is:
a) A value for which half of the values are higher and half of the values are lower
b) The value located exactly midway between the minimum and maximum of the series
c) The most commonly encountered values among the series
d) A measure of the eccentricity of the series
12. If a series of values consists of 21 numbers, then, for finding the median, we ordered the series;
a) The 11th value in the ordered series

b) The mean between the 10th and 11th values
c) The mean between the 11th and 12th values
d) The 10th value in the ordered series

13. Explain the following concepts with examples:

a. Population	b. Sample	c. Statistic
d. Parameter	e. Variable	d. Sum of square

14. Sample-X consists of 6 observations: 1, 3, 5, 7, 7, 4

a. Calculate the mean of the sample X
b. Calculate the median of the sample X
c. Calculate the mode of the sample X
d. Calculate the variance of the sample X
e. Calculate the standard deviation of the sample X

15. In a study, it is found that the mean and the standard deviation of body length for trout sample are 20 cm and 1.8 cm, respectively. Moreover, the mean and the standard deviation of body weight for trout sample are 250 gr and 15 gr, respectively. Which variable has greater variability?

4. CORRELATION AND REGRESION

4.1. INTRODUCTION

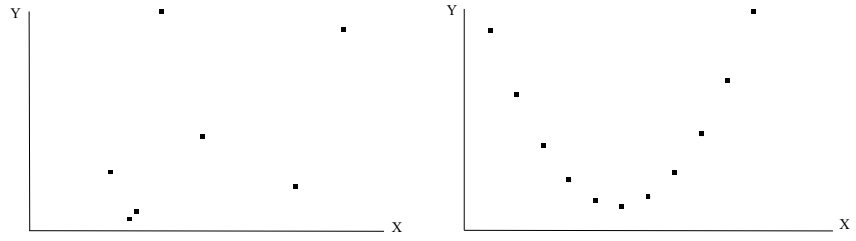
In many cases, when a sample taken from the population, the data on more than one variable are collected. In such cases, it would not be sufficient for the researcher to calculate and know only the descriptive statistics. This is because this would be a loss of information, and it is possible to investigate whether there is a relationship between two variables by using the available data, and how one of the variable changes due to a change of 1 unit in the other one.

In this circumstances, the statistics that need to be calculated are the coefficients of correlation and regression. In this book, while the coefficients are explained, two variables linearly related are taken into account.

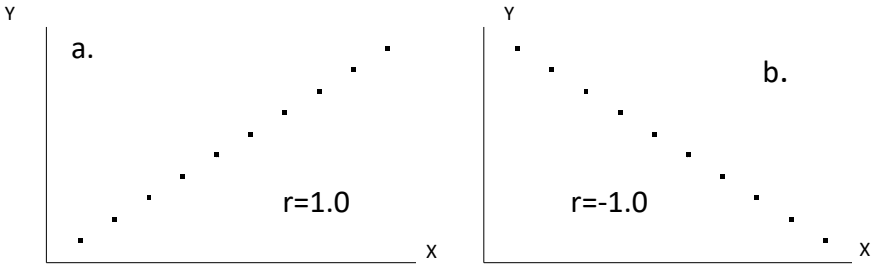
4.2. CORRELATION

When being collected the data on two variables, a scatter plot is used to show the pattern of relationship between them. In other words, if the collected data on two variables are displayed in the coordinate system, the researcher will have a preliminary information about the relationship.

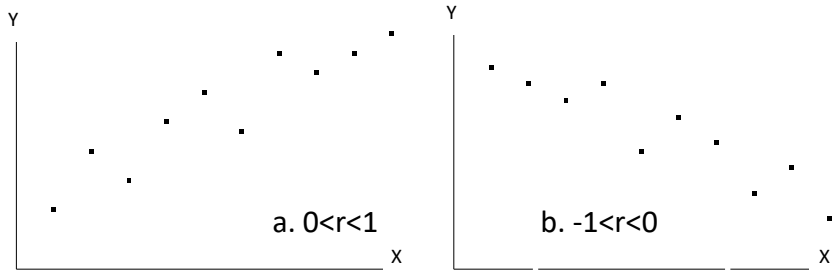
If the graphs as follows are obtained, they mention that there is no relationship or linear relationship between these two properties.



When the data of the two properties are marked in the coordinate system, the points can be sorted on one line as shown in the graphs below. This indicates an exact relationship between the two variables, that is, there is a single Y value that corresponds to each X value. When the correlation is perfect, the correlation coefficient is 1.0. The direction of relationship is indicated by the sign of the correlation coefficient. If the relationship between two variables is an increasing relationship as in the following graph (a), i.e., if Y increases as X increases, the correlation has positive sign, i.e., r = 1.0. However, the relationship between the two variable is negative (opposite, decreasing), as shown in following graph (b), then the sign of correlation coefficient is negative, i.e. r = -1.0.



When the data are displayed in the coordinate system, the graphs can exhibit a relationship seen in the following graphs:



As seen in graph (a), there is no complete relationship between the two variables, that is, the points are not aligned on a straight line. In this case, if there is an incomplete positive relationship between the two variables as in figure (a), the correlation coefficient takes a value between 0 and +1. When there is a negative incomplete relation between the two variables in in the figure (b), the correlation coefficient is between -1 and 0.

As explained above, the correlation coefficient varies between -1 and +1. If the relationship is complete, it takes value 1. The sign of the correlation coefficient indicates the direction of the relationship. If the correlation coefficient is calculated from the population data, it is a parameter and is calculated by using the following formula:

$$\rho = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$

If the correlation coefficient is calculated from the sample data, it is a statistic and is calculated by using the following formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

And it is shown briefly as follows:

$$r = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$$

In equality, $\sum d_x d_y$ sums of products, $\sum d_x^2$ sums of squares of X variables, and $\sum d_y^2$ is the sum of squares of Y variables.

The sum of the products is the sum of the multiplication of the deviations of the X and Y values from their own mean, and is shown as $\sum d_x d_y = \sum (x_i - \bar{x})(y_i - \bar{y})$. In practice, the sum of products is calculated using the following formula:

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

In summary:

1. Correlation coefficient measures the strength of the association between two variables.
2. If the correlation coefficient is calculated from the population data it is denoted by "ρ". If the correlation coefficient is calculated from the sample data, it is denoted by "r".
3. The sign of the correlation coefficient describes the direction of the relationship between two variables.
4. The absolute value of the correlation coefficient describes the magnitude of the relationship between two variables.
5. The correlation coefficient always takes value between -1.0 and +1.0. It cannot be smaller than -1.0 and larger than +1.0, that is $-1.0 \leq r \leq +1.0$.
6. The greater the absolute values of correlation coefficient, the stronger the linear relationship.
7. The strongest linear relationship is indicated by a correlation coefficient of -1.0 or +1.0, which indicates that all the data points would lie along a straight line.
8. The correlation coefficient value close to (or equal to) zero indicates no relationship or linear relationship between two variables.
9. A positive correlation coefficient means that if one variable gets bigger, the other variable tends to get bigger too. In other words, a positive correlation indicates a positive relationship, that is, increasing values in one variable correspond to increasing values in the other variables.
10. A negative correlation coefficient means that if one variable gets bigger, the other variable tends to get smaller. In other words, a negative correlation indicates a negative relationship, that is, increasing values in one variable correspond to decreasing values in the other variables.
11. The correlation coefficient has no unit.

4.3. REGRESSION COEFFICIENT

One of the two variables studied can be treated as the function of the other. Not only this may be the nature of the features, but also it could be more practical in terms of researcher's study field. The population of a country increases as the years go by. So there is a positive relationship. Here the population of the country is a function of years (time). If the population Y is defined by time X, this relationship is briefly denoted as y = f (x). The trail (in meters) that any vehicle leaves at the end of the brake is a function of its speed (km / h). In fact, the

relationship brake distance = f (speed). However, in everyday life it is desirable to know the speed when an accident occurred. This is very important for the parties involved in the accident. However, the accident is over and there is evidence of a brake left on the road as evidence. Here, in order to predict the speed, in contrast to the current situation, it is taken into account as speed = f (brake distance).

When one of the variables is defined as the function of the other, the variable on the left side of the equation, which is also called the dependent variable, depends on the independent variables on the right side of the equation. The dependent variable can also be treated as a function of more than one variable, such as y = f (x1, x2, x3).

In this section, the relationship between a dependent and an independent variable will be discussed.

Y = f (x); X is the independent variable, and Y is the dependent variable. If the pairs X and Y are pointed at the coordinate axis and the decision is made that the relationship can be assumed to be linear, the equation of the relationship is written as:

$$Y = a + bX + e.$$

In the equation, b is the coefficient of the regression which indicates the amount of change in dependent variable due to a change of one unit of independent variable. The correlation coefficient gives the degree of relationship between the two variables and it has no unit. However, the unit of the regression coefficient is the amount by which the dependent variable will change on average in terms of its unit.

If there are two variables, that is X and Y, two regression coefficients can be calculated: b_{yx} and $b_{xy.}$

b_{yx} = variable Y regressed on variable X
b_{xy} = variable X regressed on variable Y

In general terms, the following equation is used for the calculation of the regression coefficient:

$$b = \frac{\text{Sum of products}}{\text{Sum of squares of independent variable}}$$

In this case, b_{yx} coefficient is calculated as;

$$b_{yx} = \frac{\sum d_x d_y}{\sum d_x^2}$$

b_{xy} coefficient is calculated as;

$$b_{xy} = \frac{\sum d_x d_y}{\sum d_y^2}$$

When the regression coefficient is calculated using the equations given above, the sign of the coefficient is positive or negative depending on the direction of the relationship between the two variables. If there is an inverse relationship between the two variable, the sign of coefficient is negative. If there is an increasing relationship, the coefficient is found as positive. The sign of the regression coefficient and the sign of the correlation coefficient can never be different because the sign of both is determined by the sign of the sum of products.

4.4. REGRESSION EQUATION

If the marked points were on a straight line, the equation could be easily constructed. If the points are distributed on the right side, rather than on an exact line, a straight line can be formed that passes nearest to all of the points corresponding to the X and Y values. This straight line to be formed is called "Regression Line" and this equation is called "Regression Equation" or "Pre-Estimation Equation", and is given as:

$$\hat{Y} = a + b_{yx}X.$$

Where, "$\hat{Y}$" is the predicted value of the dependent variable Y, "a" is a constant, the point at which the line crosses the Y axis when X=0, "b_{yx}" is the slope of the regression line, i.e., the regression coefficient, and "X" is the observed value of the independent variable.

4.5. DETERMINATION COEFFICIENT

The accuracy of the estimates made by the regression equation expressed by determination coefficient and is denoted by r^2." r^2" equals the square of the correlation coefficient calculated. As the absolute value of the correlation coefficient approaches 1, it means that the accuracy of the prediction by regression equation increases.

In summary:

1. b_{yx}, is the regression coefficient of variable Y regressed on variable X. Here variable Y is dependent variable and variable X is the independent variable. It is the amount of change in variable Y due to a change of one unit of variable X, on average.
2. b_{xy}, is the regression coefficient of variable X regressed on variable Y. Here variable X is dependent variable and variable Y is the independent variable. It is the amount of change in variable X due to a change of one unit of variable Y, on average.
3. The unit of the regression coefficient is the measurement unit of the dependent variable.
4. A positive regression coefficient indicates the amount of increment in variable Y due to a change of one unit of variable X, on average.
5. A negative regression coefficient indicates the amount of reduction in variable Y due to a change of one unit of variable X, on average.

EXAMPLE 1:
In a study conducted to examine the change in the amount of vitamin C in any vitamin syrup kept at room temperature after opening cover, the amount of vitamin C per unit volume for weekly periods was recorded as follows.

Storage time	Vitamin C in the unit volume
0	100
2	90
4	70
6	40
8	30

Here, the storage time is independent variable (X) and the dependent variable is the amount of vitamin C. The amount of vitamin is taken as a function of time, that is, Y = f (X).

The following table is prepared for calculating the coefficients of this model as described above.

	X	Y	X^2	Y^2	XY
	0	100	0	10000	0
	2	90	4	8100	180
	4	70	16	4900	280
	6	40	36	1600	240
	8	30	64	900	240
Sum:	20	330	120	25500	940

$$\sum X = 20,\ \sum Y = 330,\ \sum XY = 940,\ \sum X^2 = 120,\ \sum Y^2 = 25500$$

Mean of independent variable: $\overline{X} = \frac{20}{5} = 4$

Mean of dependent variable: $\overline{Y} = \frac{330}{5} = 66$

Sum of squares of independent variable: $\sum d_X^2 = 120 - \frac{(20)^2}{5} = 40$

Sum of squares of dependent variable: $\sum d_Y^2 = 25500 - \frac{(330)^2}{5} = 3720$

Sum of products: $\sum d_X d_Y = 940 - \frac{(20)(330)}{5} = -380$

The regression coefficient of the amount of vitamin C in the unit volume over time is calculated as follows:

$$b_{yx} = \frac{-380}{40} = -9.5$$

The meaning of the coefficient -9.5 is "when the independent variable (weekly) increases by one unit (one week, the dependent variable (vitamin) decreases by 9.5 mg, on average".

According to these results, regression equality,

$$a = \overline{Y} - b_{yx}\overline{X}$$

$$a = 66 - (-9.5)(4) = 104$$

Regression equation is: $\hat{Y} = 104 - 9.5X$

The correlation coefficient between the time and the unit of vitamin C in the unit volume is calculated as follows:

$$r_{xy} = \frac{-380}{\sqrt{(40)(3720)}} = -0.985$$

The calculated correlation coefficient showed a linear decreasing association of 98.5% between the storage time and the amount of vitamin C. In other words, the correlation coefficient of -0.985 indicates a strong negative correlation between the storage time and the amount of vitamin C. For this example, the determination coefficient is $(0.985)^2$=0.97. This means that 97% of variability in the dependent variable can be accounted for by the independent variable.

EXAMPLE 2:
To study the relationship between drug doses (X) and the number of heartbeat, 6 rats are with different drug doses and the heartbeats are recorded as follows:

X(drug doses)	3.0	3.1	3.2	2.9	3.5	3.4
Y(heartbeat)	85	90	92	93	86	88

The following values are calculated using the observed heartbeats:

$\sum X = 19.1$ $\quad\quad$ $\sum Y = 534$

$\sum d_X^2 = 0.268$ $\quad\quad$ $\sum d_Y^2 = 52$

$\sum d_X d_Y = -1.6$

a. What is the strength of association between drug doses and heartbeats?
b. What is the amount of change in variable Y on average due to an increase of 1 unit in X?
c. How much of the variation observed in the Y variable can be explained by X variable?

The correlation coefficient is calculated as follows:

$$r_{xy} = \frac{-1.6}{\sqrt{(0.268)(52)}} = -0.429$$

This indicates that there is a decreasing linear relationship of 42.9% between drug doses and heartbeat.

To have information on The amount of change in number of heartbeat on average due to an increase of 1 unit in dose regression coefficient is calculated as follows:

$$b_{yx} = \frac{-1.6}{0.268} = -5.97$$

This points to a decline of 5.97 in heartbeat number on average due to an increase of 1 unit in dose.

To estimate the heartbeat number corresponding to the doses, the regression equation is calculated. First, "a" which is the point at which the regression line crosses Y-axis is calculates as:

$$a = \overline{Y} - b_{yx}\overline{X}$$

$$a = 89 - (-5.97)(3.18) \cong 108$$

Then, regression equation is written as:

$$\hat{\mathbf{Y}} = \mathbf{108} - \mathbf{5.97\ X}$$

For this example, the determination coefficient is $(-0.429)^2=0.183$. This means that 18.3% of variability in the heartbeats can be accounted for by the drug doses

QUESTIONS:

1.Some data are given as:

X	Y
1	16
2	23
4	35
3	28
5	44
6	40
3	22
8	61
9	82

d. Compute the correlation coefficient and explain what it means.

b. Compute the coefficients of regression b_{yx}, and explain what it means

c. What is the estimated value, for X = 7?

d. What is the estimated value, for X = 17?

2) The correlation coefficient between study time (h/day) and success grade is computed as 0.90. Which of the following expression is true?

A) As the study time increases, the success grade decreases.

B) %90 of the hard working students gets high marks.

C) %81 of the variation in success grade can be explained by study time.

D) %90 of the variation in success grade can be explained by study time.

E) None of them.

3) The correlation of linear relationship between X and Y variables is - 0.75. Which of the following expressions are absolutely true for regression coefficient?

A) $-\infty<b<\infty$ B) $-\infty<b<0$ C) $0<b<\infty$ D) $0<b<1$ E) $-1<b<1$

4) If the sign of the regression coefficient is (-), which of the following is absolutely true for correlation coefficient?

A) $-\infty<r<0$ B) $-1<r<1$ C) $-1<r<0$

D) $r = -1$ E) $r=0$

(5th-7th). Answer the questions according to the following information.

The following values were calculated for the data collected from 10 students:

X: Attendance to biometrics course (week)

Y: Final grades

Sum of squares for X = 134

Sum of squares for Y = 6000

Sum of products of X and Y = 730

5) What is the correlation coefficient between X and Y variables?

A) 0.814 B) 5.448 C) 0.663 D) 0.122

6) What is the amount of change in variable Y due to an increase of 1 unit in X, on average?
A) 0.814 B) 5.448 C) 0.663 D) 0.122

7) How much of the variation observed in the Y variable can be explained by X variable?
A) 0.814 B) 5.448 C) 0.663 D) 0.122

(8th-11th). Answer the questions according to the following information.
The following values were calculated for 15 individuals in terms of X (weight, kg) and Y (height, cm). (NOTE: Weights range from 45kg to 80kg, height ranges from 150cm to 175cm.)

$\Sigma X = 855$ $\Sigma Y = 2430$ $\Sigma X^2 = 50738$ $\Sigma Y^2 = 398800$ $\Sigma XY = 140681$

8) What is the he degree of linear relationship between the X and Y variables?
A) 1.084 B) 0.422 C) 0.677 D) 0.458 E) 0.178

9) Which of the following is the regression coefficient of variable Y regressed on variable X?
A) 1.084 B) 0.422 C) 0.677 D) 0.458 E) 0.178

10) When variable Y increases by 1 cm, what is the amount of change in variable X, on average?
A) 1.084 B) 0.422 C) 0.677 D) 100.21 E) 0.178

11) Estimate the weight of an individual whose height is 145 cm?
A) 49.826 kg B) 257.392 kg
C) 11.364 kg D) 100.212 kg
E) cannot be predicted

1. The lithium and sodium measurements in water samples taken from six different water source are given on the right side.

Na (g/L) (X)	Li (mg/L) (Y)
0.6	0.8
0.9	1.6
1.5	1.1
3.7	1.8
0.7	1.0
4.0	3.0

a. Calculate the regression coefficient of Li on Na, and explain the meaning of regression coefficient.
b. Calculate the regression equation
c. Calculate correlation coefficient and explain its meaning.
d. If the regression equation is used to predict Li-values for Na-values, what is the accuracy of predictions? Calculate, and explain its meaning.

$\left(\sum d_x^2 = 11.94 \ \sum d_y^2 = 3.24 \ \sum d_x d_y = 5.26\right)$

a. Calculate the regression coefficient of Li on Na, and explain the meaning of regression coefficient.
b. Calculate the regression equation
c. Calculate correlation coefficient and explain its meaning.
d. If the regression equation is used to predict Li-values for Na-values, what is the accuracy of predictions? Calculate, and explain its meaning.

5.STATISTICAL DISTRIBUTIONS

The studied variables represent populations whose distribution functions are known. Binomial and Normal distributions whose probability density functions are known are most commonly encountered distributions.

5.1. Binomial Distribution

A population in which each trial has two possible outcomes which are named as success and failure distributes in accordance with binomial distribution which is a discrete distribution. Researcher deals with the outcome which is named as success. In binomial distribution, sequential trials should be independent of each other and the probability of each two outcomes is assumed to be constant. To keep the probability of outcomes constant throughout the experiment, the sequential trials should be done with replacement. If the trials are not done with replacement, the probability of the outcomes cannot remain constant.

The probability of a success is equal to $\boldsymbol{\pi}$ if we are working on population and is equal to **p** if we are working on sample. The value of p remains constant throughout the experiment.

In binomial distribution the number of success out of "n" trials are denoted by "r". The probability of r successes out of n trials is calculated by using the following probability density function of binomial distribution:

$$\mathbf{P(r)= C(n,r)\ p^r\ (1-p)^{n-r}}$$

Where, n is the number of trials, p is the probability of success, (1-p) is the probability of failure, r is the number of success out of n trials, P(r) is the probability of r successes out of n trials, nCr is the number of combinations that can be made regardless of order by taking r of n different things.

If the number of successes distributes in accordance with the binomial distribution whose parameters are n and p, the mean of distribution equals to $\mu(r) = np$ and variance of distribution equals to $\sigma^2(r) = np(1-p)$.

EXAMPLE:

After an exam, it was found that 60% of the students were successful. If 5 students were chosen randomly from the students who took this course;

a. What are the probability of 5 being successful?
b. What is the probability of 4 being successful out of 5 students?
c. What is the probability of 3 being successful out of 5 students?
D. What is the probability of 2 being successful out of 5 students?
to. What is the probability of 1 being successful out of 5 students?
f. What is the probability that they all fail?
g. What are the probability of at least 2 being successful?
h. What are the probability of 4 of them being successful?

a. The probability of 5 being successful is:

$$P(5) = C(5,5)(\frac{3}{5})^5(\frac{2}{5})^{5-5} = \frac{5!}{(5-5)!5!}(\frac{3}{5})^5(\frac{2}{5})^0$$

$$= (\frac{3}{5})^5 = \frac{243}{3125} = 0.07776$$

that is, %7.776.

b. The probability of 4 being successful out of 5 students randomly selected is:

$$P(4) = C(5,4)(\frac{3}{5})^4(\frac{2}{5})^{5-4} = \frac{5!}{(5-4)!4!}(\frac{3}{5})^4(\frac{2}{5})^1$$

$$= 5(\frac{3}{5})^4(\frac{2}{5}) = \frac{810}{3125} = 0.2592$$

That is, %25.92.

c. The probability of 3 being successful out of 5 students randomly selected is;

$$P(3) = C(5,3)(\frac{3}{5})^3(\frac{2}{5})^{5-3} = \frac{5!}{(5-3)!3!}(\frac{3}{5})^3(\frac{2}{5})^2$$

$$= 10(\frac{3}{5})^3(\frac{2}{5})^2 = \frac{1080}{3125} = 0.3456$$

that is, %34.56.

d. The probability of 2 being successful out of 5 students randomly selected is;

$$P(2) = C(5,2)(\frac{3}{5})^2(\frac{2}{5})^{5-2} = \frac{5!}{(5-2)!2!}(\frac{3}{5})^2(\frac{2}{5})^3$$

$$= 10(\frac{3}{5})^2(\frac{2}{5})^3 = \frac{720}{3125} = 0.2304$$

that is, %23.04.

e. The probability of 1 being successful out of 5 students randomly selected is;

$$P(1) = C(5,1)(\frac{3}{5})^1(\frac{2}{5})^{5-1} = \frac{5!}{(5-1)!1!}(\frac{3}{5})^1(\frac{2}{5})^4$$

$$= 5(\frac{3}{5})(\frac{2}{5})^4 = \frac{240}{3125} = 0.0768$$

that is, %7.68.

f. The probability of 0 being successful out of 5 students randomly selected is;

$$P(0) = C(5,0)(\frac{3}{5})^0(\frac{2}{5})^{5-0} = \frac{5!}{(5-0)!0!}(\frac{3}{5})^0(\frac{2}{5})^5$$

$$= (\frac{2}{5})^5 = \frac{32}{3125} = 0.01024$$

that is, %1.024.

Notice that, the sum of the probabilities of all possible cases is equal to 1.

$$P(5) + P(4) + P(3) + P(2) + P(1) + P(0) = 1.0$$

$$\frac{243}{3125} + \frac{810}{3125} + \frac{1080}{3125} + \frac{720}{3125} + \frac{240}{3125} + \frac{32}{3125} = \frac{3125}{3125} = 1.0$$

g. Success of at least 2 students means 2, 3, 4 or 5 students are successful. In this case, the probability to be calculated is shown in P (r≥2) and P (r ≥2) = P (2) + P (3) + P (4) + P (5), that is:

$$P(r\geq 2)= \frac{720}{3125} + \frac{1080}{3125} + \frac{810}{3125} + \frac{243}{3125} = \frac{2853}{3125} = 0.91296$$

or;

$$P(r\geq 2)=1-[P(0)+P(1)]$$

$$P(r\geq 2)=1-\frac{240}{3125} + \frac{32}{3125} = 1 - \frac{272}{3125} = \frac{2853}{3125} = 0.91296$$

h. The probability of at most 3 successful out of 5 students is calculated as the sum of

$$P(r\leq 3)=P(0)+P(1)+P(2)+P(3)$$

$$P(r\leq 3)= \frac{32}{3125} + \frac{240}{3125} + \frac{720}{3125} + \frac{1080}{3125} = \frac{2072}{3125} = 0.66304$$

or;

$$P(r\leq 3)=1-[P(4)+P(5)]$$

$$P(r\leq 3)=1-\frac{243}{3125} + \frac{810}{3125} = 1 - \frac{1053}{3125} = \frac{2072}{3125} = 0.66304$$

Question: The probability that a person suffering from migraine headache will obtain relief with a particular drug is 0.9. Three randomly selected sufferers from migraine headache are given the drug. Find the probability that the number obtaining relief will be:

a. Exactly zero b. More than one c. Exactly two d. Two or fewer

Question: It is known that in a faculty, the probability of smoking is 25%. If we select 7 students randomly, what is the probability of 4 students who are smokers out of 7 students

5.2. Normal Distribution

The normal distribution is a continuous distribution unlike binomial distribution. If the differences amongst the data on a continuous variable result from the uncontrollable factors (which is referred as accidental differences), it distributes in accordance with normal distribution. For example, the differences in terms of weight amongst students who are in a certain age group, same gender, and under the same conditions in terms of weights occurs because of the small effect of uncontrollable factors on weight. Properties occurring under such effects show normal distribution.

Normal distribution is a continuous distribution and is the most important distribution in statistics. This distribution is also known Gaussion distribution. The normal distribution is defined by the following probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

Normal distribution probability function has two parameters: μ (mean) and σ (standard deviation). In the function, e (2.718) and π (3.1416) are mathematical constants.

The probability density function is used to calculate the probability being in a specified interval. For example, assume that the student weights show a normal distribution with a mean of 60 kg and with standard deviation of 5 kg. If the researcher intends to know the probability of students whose weights are between 65 and 69 kg, he should take integral of the probability density function between 65 and 69, as following:

$$\int_{65}^{69} f(x)\,dx = \int_{65}^{69} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{X-60}{5})^2} dx$$

This integration gives the probability of students whose weights are in the interval of 65 and 69 kg. In other words, it gives percent of students whose weights are in the interval of 65 and 69 kg.

5.2. Characteristics of Normal Distribution

1. As seen in the following figure, normal distribution is a bell-shaped distribution, which means that the majority of the data is in the middle of the distribution.

2. Normal distribution has two parameters. These distributions differentiate from each other with the difference in their parameters.

3. Normal distribution is a symmetric distribution. It means that if the distribution curve is cut from the mean line, the left and right sides are the same.

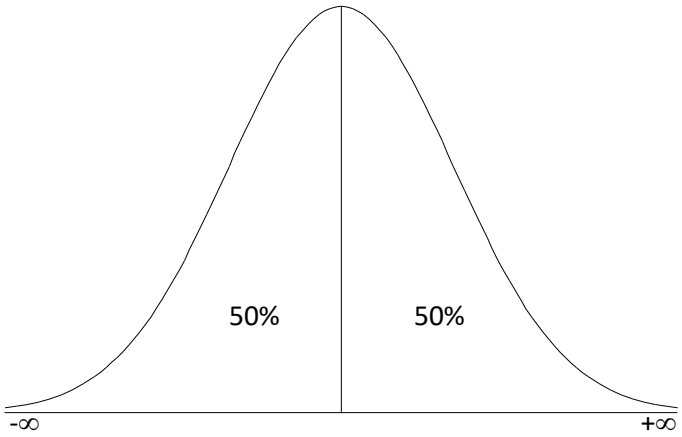
4. In normal distribution, mean, median, and mode are equal to each other.

5. In normal distribution, 50% of observations is smaller than mean and 50% of observations is greater than mean.

6. The total area under a normal distribution curve is equal to 1 or 100%.

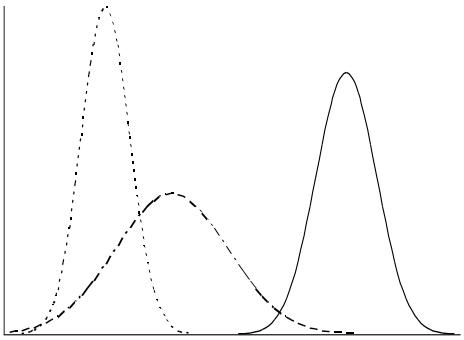
$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} dx = 1.0$$

7. The normal distribution curve is asymptotic to the horizontal axis.
8. Normal distribution includes all the values from -∞ to +∞.

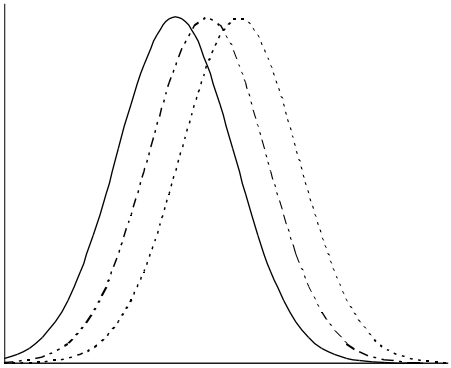


9. Normal distributions differentiate from each other with difference in their parameter(s).

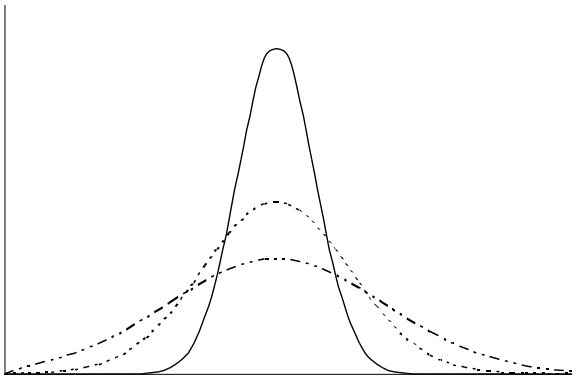
As seen in the following figure, normal distributions are different from each other with the difference in both parameters.



As seen in the following figure, normal distributions are different from each other with the difference in their mean.



As seen in the following figure, normal distributions are different from each other with the difference in their standard deviation.



5.2. Standard Normal Distribution

Standard normal distribution is a normal distribution with a mean of zero and a standard deviation of one. Standard normal distribution is defined by the following probability density function:

$$f(z) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{z^2}{2}}$$

If parameters of normal distribution are known, each normal distribution can be transformed to the standard normal distribution by transforming any X-value into Z-value using the following formula:

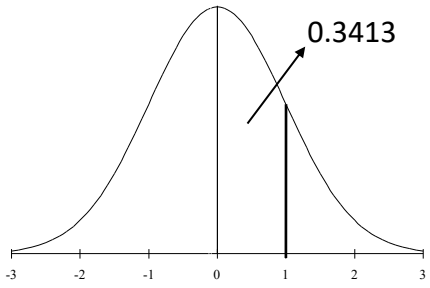
$$Z = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} = \frac{X - \mu_x}{\sigma_x}$$

Z-formula standardizes any normal distribution. Z value is the number of standard deviations, which shows the distance of any value from mean. If each value of a normally distributed variable is transformed into Z-value, the result will be standard normal distribution. Z-value can be negative, which implies that Z-value is below mean. However, area is always positive.

After the integral of the probability density function is taken from zero to the determined Z-value, the probabilities of being between zero and determined Z-values are organized as table. The standard normal distribution table is used to calculate area under the standard normal curve.

EXAMPLE 1:

What is the probability of Z values between 0 and 1.0 values?



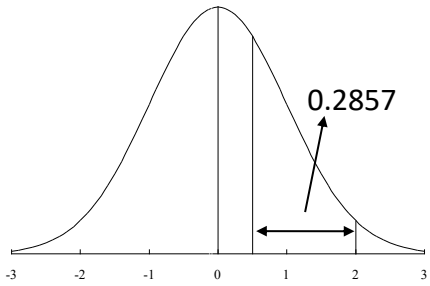
From Standard normal distribution table, the area between 0 and +1 under standard normal distribution curve is found as 0.3413. In other words, 34.13% of all the Z-values are between 0 and +1. In other words; the probability of any Z value between zero and 1 in the standard normal distribution is 34.13%.

It can be said that in normal populations, 34.13% of the observation values are in the area between the mean and 1 standard deviation from this.

EXAMPLE 2:

In standard normal distribution, what is the probability of Z values between 0.5 and 2.0 values. This probability is calculated as follows:

$$P(0.5<Z<2.0)=P(0<Z<2.0)-P(0<Z<0.5)$$



To calculate this probability, the probability of Z-values to be between 0 and 0.5 is subtracted from the probability of Z-values to be between 0 and 2.0. From standard normal distribution table, it is found that; $P(0<Z<2.0) = 0.4772$ and

$P(0<Z<0.5) = 0.1915$

In this case. $P(0.5<Z<2.0)$ equals to

$$0.4772-0.1915=0.2857.$$

EXAMPLE 3:

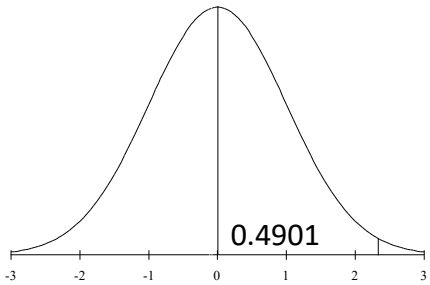
It is known that the weight of newborn babies has a normal distribution with a mean weight of 3.2 kg and a standard deviation of 0.3 kg.

a. What is the probability of newborn babies whose weights are between 3.2 kg and 3.9 kg?

Here, the probability of newborn babies whose weights between 3.2 kg and 3.9 kg is asked. To calculate this, first, the values of normal distribution, which are 3.2 kg and 3.9 kg, are transformed into standard normal distribution by using the following formula;

$$P(3.2 < X < 3.9) = P(\frac{3.2 - 3.2}{0.3} < Z < \frac{3.9 - 3.2}{0.3})$$
$$= P(0 < Z < 2.33)$$

so, the z-values corresponding to 3.2 and 3.9 kg are calculated as 0 and 2.33, respectively.



From standard normal distribution table, the area between 0 and 2.33 under standard normal distribution curve is found as 0.4901, which gives the probability of Z-values between 0 and 2.33, which is the desired probability. This results clarifies that the probability of newborn babies whose weights are between 3.2 kg and 3.9 kg is 49.01%.

EXAMPLE 4:

The Vitamin-C content of a particular brand of vitamin supplement pills is normally distributed with mean 490 mg and standard deviation 12 mg. Then;

a. What is the probability that a randomly selected pill contains at least 500 mg of Vitamin C?

b. What is the probability that a randomly selected pill contains Vitamin C between 500 and 510 mg?

c. What is the probability that a randomly selected pill contains less than 500 mg of Vitamin C?

d. What is the lowest amount of vitamin C in 2.5% of vitamin pills containing the highest vitamin C?

Question 1. Describe the normal distribution.
Question 2. Describe the standard normal distribution.
Question 3. Explain how the standard normal distribution is used in statistics.
Question 4.

a. Find the area between Z=0 and Z=1, P(0<Z<1)
b. P(-1<Z<0) c. P(-1<Z<1) d. P(0<Z<1.5)
e. P(0<Z<1.64) f. P(-1.0<Z<1.5) g. P(Z<1.5)
h. P(Z>1.5) i. P(Z>-1.5)

Question 5. In a faculty, it is known that the student weights show a normal distribution with a mean of 60 kg and with standard deviation of 5 kg. Find the probability of students;

a. More than 70 kg c. More than 50 kg
b. Between 55 and 72 kg d. less than 62 kg

Question 6: The cholesterol content of large chicken eggs is normally distributed with a mean of 200 milligrams and standard deviation 15 milligrams.

a. What proportion of these eggs has cholesterol content above 205 milligrams?
a) 0.2004 b) 0.6293 c) 0.7250 d) 0.3300 e) 0.3707

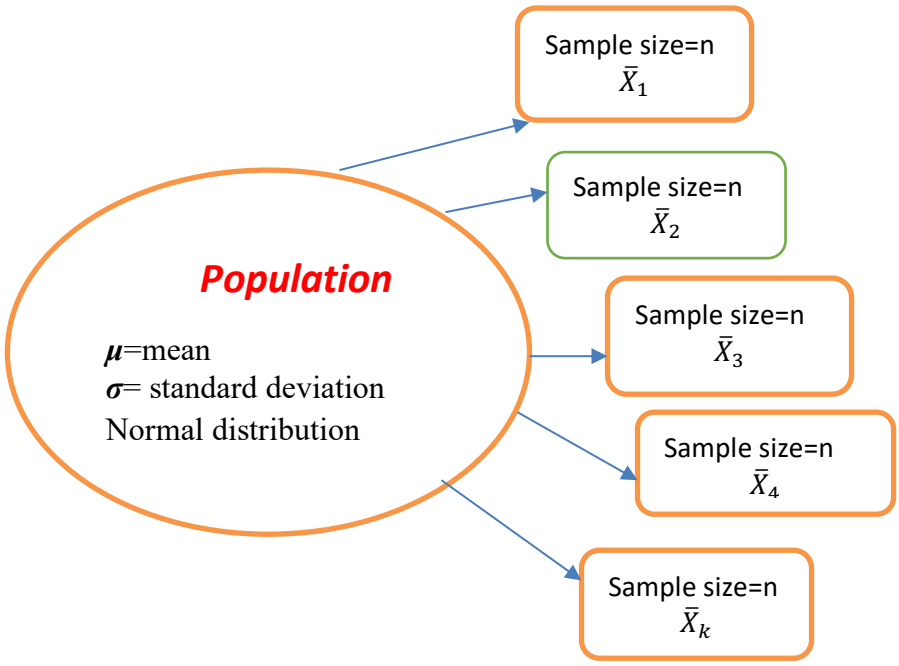
b. In sixty-seven percent of the eggs, the cholesterol content is less than a certain value "C". Find the value of "C".
a) 0.33 b) 206.6 c) 210 d) 0.44 e) 193.4

6. Sampling distribution

Sampling distributions are theoretical distributions. A sampling distribution is a distribution of statistics produced by repeated sampling with replacement from the same population. Some of the calculated statistics will be equal to population parameter, some of them are less than population parameter and some of them are greater than population parameter. Therefore, they will show a distribution which is called **sampling distribution**.

6.1. Sampling Distribution of the Mean

When samples are drawn from one population with equal sample size repeatedly and means are calculated, the calculated means show a distribution which is called **Sampling distribution of the mean.**



The calculated mean from each sample is a prediction of population mean. Some of the calculated means will be equal to population mean, some of them are less than population mean and some of them are greater than population mean. Therefore, they will show a distribution which is called **sampling distribution of the mean.**

If the samples were randomly taken from a population showing normal distribution, the sampling distribution of the mean shows also normal distribution. It has two parameters: $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

The mean of the sampling distribution is: $\mu_{\bar{x}} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3+\cdots\bar{X}_k}{k} = \mu_x$

The standard deviation of the sampling distribution is: $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

Example 1:
It is known that weight of students in a faculty show normal distribution with mean of 60 kg and standard deviation of 5 kg.

If repeated samples of 4 students with replacement are selected from this faculty and the means of weight are calculated, the calculated weight means show normal distribution. This is because the weight of students shows normal distribution. The distribution is called sampling distribution of mean.

The sampling distribution of mean has two parameters: the mean of the sampling distribution,

$$\mu_{\bar{x}} = \mu_x = 60kg$$

and the standard deviation of the sampling distribution.

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{5kg}{\sqrt{4}} = 2.5kg$$

When the sample means are calculated, the researcher can calculate the probability of means between 58 and 63 kg, $P(58 < \bar{X} < 63)$?

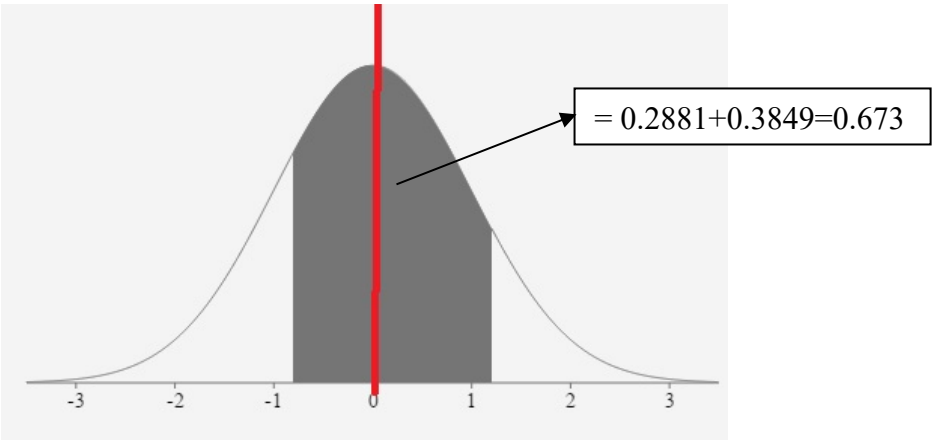
As we said before, all the calculated sample means are not equal each other. Therefore, there is variation in this sample means and they show normal distribution. The probability can be calculated by using standard normal distribution. The weights of 58 and 63 can be transformed into Z-value by using the following formula:

$$Z_1 = \frac{\overline{X}_1 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

Two Z-values are calculated

$$Z_1 = \frac{\overline{X}_1 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{58-60}{2.5} = \frac{-2}{2.5} = -0.8 \text{ and } Z_2 = \frac{\overline{X}_2 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{63-60}{2.5} = \frac{-2}{2.5} = 1.2$$

Now, we are interested in the probability of Z-values between -0.8 and 1.2:



This means that 67.3% of the means are between 58 and 63 kg.

When the sample means are calculated, the researcher can calculate the probability of means between 61 and 65 kg, $P(61 < \bar{X} < 65)$?

First, Z-values are calculated as:

$$Z_1 = \frac{\bar{X}_1 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{61-60}{2.5} = \frac{1}{2.5} = 0.4$$

$$Z_1 = \frac{\bar{X}_1 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{65-60}{2.5} = \frac{5}{2.5} = 2$$

Then, the probability of Z-values between 0.4 and 2 is calculated using the standard normal distribution as 0.3218, which means that 32.18% of weight means are between 61 and 65 kg.

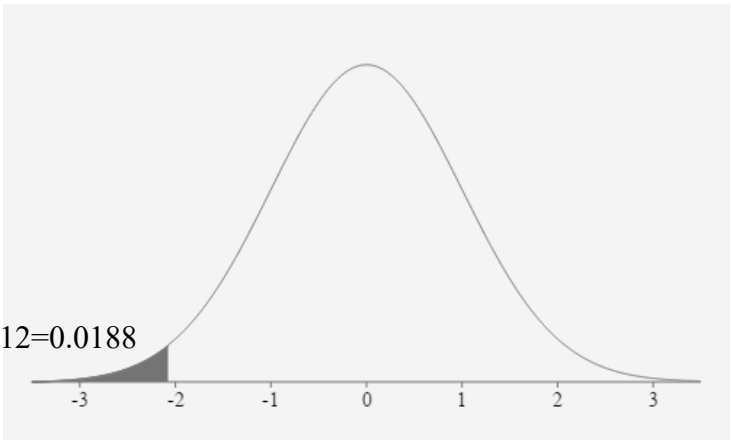
Example 2:

A factory informs that the amount of B_1-vitamin in vitamin pills produced at the factory shows a normal distribution with mean of 25 mg and standard deviation of 1.2 mg. It is claimed that a sample of 25 vitamin pills were randomly taken from this factory and the mean of B_1-vitamin was found to be 24.5 mg. What is the probability of this sample randomly taken from this factory?

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.2}{\sqrt{25}} = \frac{1.2}{5} = 0.24$$

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{24.5-25}{0.24} = -2.083$$

0.5-0.4812=0.0188



So, the probability of our 25 sample being taken from the factory is 1.88%

7. Hypothesis Testing

In general, it is very difficult to collect data from all the individuals in the population being studied, even it is impossible. Therefore, researchers have to take a random sample from the population and work on it. Depending on the sample data, information on population is obtained and the parameters are predicted.

Hypothesis testing is a procedure to determine whether the hypothesis stated is true or not. The steps of hypothesis testing procedure can be ordered as follows:

STEP 1: Stating Hypothesis

Hypothesis is a statement about value of parameter of the population being interested in. There are two hypotheses in hypothesis testing.

Null hypothesis: The null hypothesis is denoted by $\mathbf{H_0}$. This is a statement about the value of population parameter. This hypothesis claims that the difference between population parameter and sample statistic occurs only because of chance. It is stated as:

$$H_0: \mu_{\bar{x}} = \mu_x \text{ (hypothesised value)}$$

Alternative hypothesis: The alternative hypothesis is denoted by $\mathbf{H_1}$. This hypothesis is accepted if the null hypothesis is false and rejected. Alternative hypothesis determines whether hypothesis testing is two-tail or one-tail hypothesis control. If alternative hypothesis is stated as $H_1: \mu_{\bar{x}} \neq \mu_x$ (hypothesised value), this shows two-tailed hypothesis control. If it is stated as $H_1: \mu_{\bar{x}} > \mu_x$ (hypothesised value) or $H_1: \mu_{\bar{x}} < \mu_x$ (hypothesised value), this shows one-tailed hypothesis control. In null and alternative hypotheses, the same value of population parameter must be used.

STEP 2: Test Statistic and Test distribution

After stating the hypotheses, the appropriate test statistics and test distribution is determined. In hypothesis control of mean, if the standard deviation of population is known, the appropriate test statistics is Z-value and calculated as follows:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

The standard normal distribution (Z-distribution) is used as test distribution to make decision.

STEP 3: Significance Level

In hypothesis control, there are two types of error:

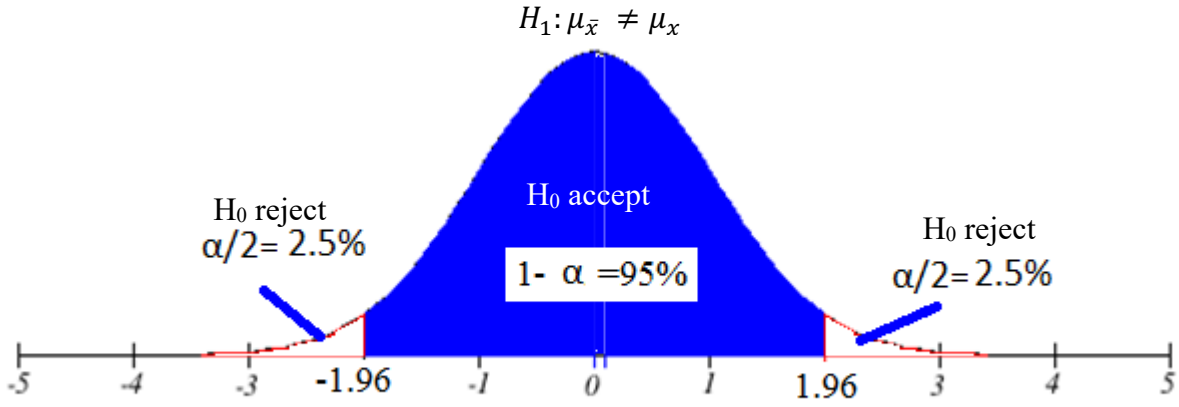
Type I Error: Type I error is made when the null hypothesis is rejected even if it is true in reality. The probability of type I error is called significance level and is denoted by α. In general, α is determined as 1% or 5% in applied and health sciences. When the probability of type I error decreases, reliability of hypothesis testing increases.

Type II Error: Type I error is made when the null hypothesis is not rejected even if it is false in reality.

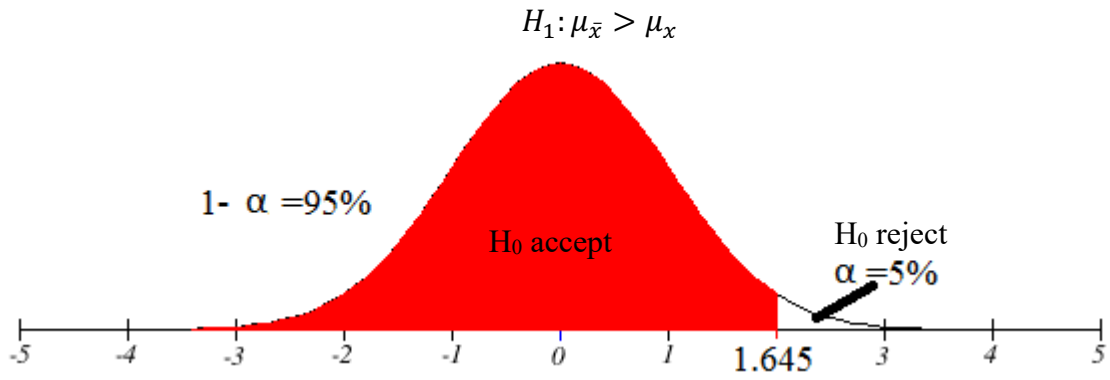
STEP 4: Making Decision

The probability of type I error divides test distribution into two regions: H_0 accept and H_0 rejection. These regions on graphs is shown as follows:

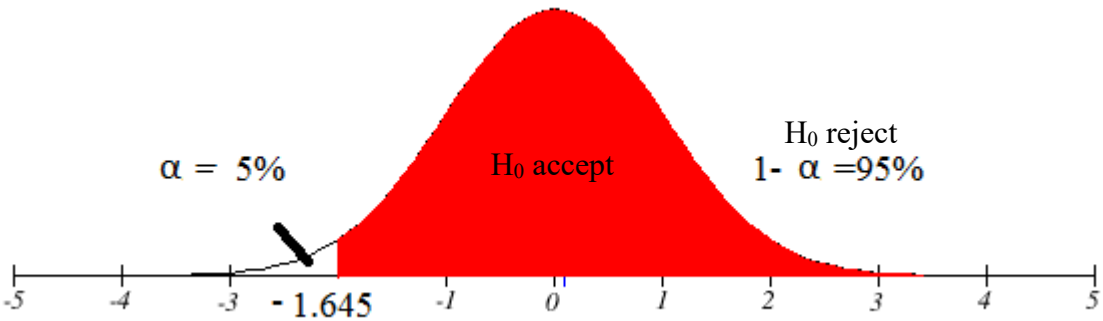
Two tailed:



One tailed:



or $H_1: \mu_{\bar{x}} < \mu_x$



Example 1: A factory informs that the amount of B-vitamin in the vitamin pills shows a normal distribution with a mean of 25 mg and a standard deviation of 1.2 mg. 25 vitamin pills are taken from this factory and the mean B-vitamin is calculated as 24.5 mg. Can we conclude that the mean of B-vitamin is 25 mg?

Step 1:

H_0: There is no statistically significant difference in B-vitamin between the sample and the population means. In another word, this sample is drawn from the population with mean of 25 mg, that is $H_0: \mu_x = 25$

H_1: There is statistically significant difference between the sample and the population means. In another word, this sample is not taken from the population with mean of 25 mg, that is $H_1: \mu_x \neq 25$

Step 2:

Since the population standard deviation is given, the test statistics is z-statistics and the test distribution is z-distribution.

Z-statistics is calculated as $Z = \dfrac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$

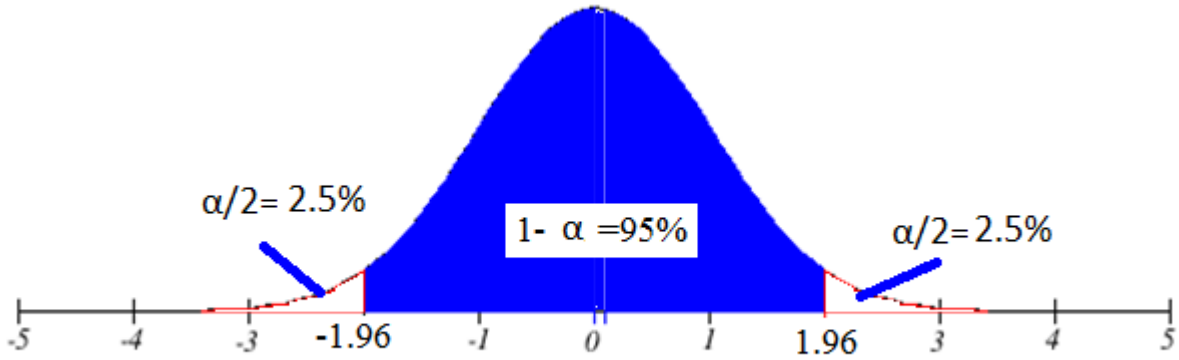
Given → n=25 , $\overline{X} = 24.5$, $\mu_{\overline{X}} = 25$, $\sigma_X = 1.2$

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.2}{\sqrt{25}} = \frac{1.2}{5} = 0.24$$

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{24.5 - 25}{0.24} = -2.083$$

Step 3:

For this example, assume that α is determined as 5%.



Step 4: Decision

The calculated Z-value of -2.083 falls in the critical (H_0 rejection) region, so the null hypothesis is rejected which means that there is statistically significant difference between the population mean (25mg) and the sample mean (24.5). In another word, this sample is not drawn from the population with mean of 25mg and standard deviation of 1.2mg

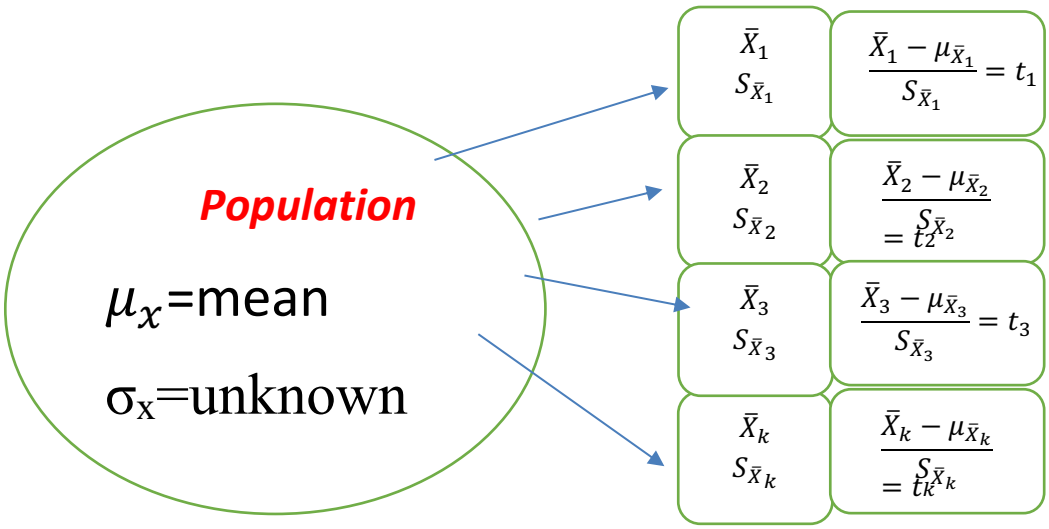
8. t-distribution (student's t-distribution)

When samples of n sample size are repeatedly drawn from a normal population with unknown standard deviation, the standard deviation of population is predicted from the sample data. Therefore, the standard deviation of sampling distribution of mean is predicted from the sample data and is denoted by $S_{\bar{x}}$ and is called standard error of mean. The sample mean can be standardized by using the formula $\frac{\bar{X}-\mu_{\bar{X}}}{S_{\bar{X}}}$. This value is called t-value. For each sample t statistics is calculated as follows, these calculated t-statistics show a distribution with a degree of freedom (n-1). This distribution is called t-distribution.

t-statistics is calculated

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{S_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{S_X/\sqrt{n}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sqrt{S^2{}_X/n}}$$

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$



Properties of t-distribution

- There are infinite t-distributions depending on degrees of freedom,
- T-distribution is a bell-shaped and symmetric distribution,
- T-distribution includes all the t-value from $-\infty\ to\ +\infty$,
- The mean of t-distribution is zero and its variance is $\frac{n-1}{n-3} = \frac{df}{df-2}$,
- When degree of freedom increases variance decreases,
- When degree of freedom increases t-distribution approaches to z-distribution,
- When degree of freedom is equal to infinity t-distribution overlaps to normal distribution.

Example 1:

It is known that the mean recovery time of patients used an old antidepressant is 8 months and also known recovery time is normally distributed. Newly developed antidepressant is given to 25 patients and the mean recovery time is calculated to be 7.5 months and variance 0.81. Can we conclude that the new antidepressant reduces recovery time?

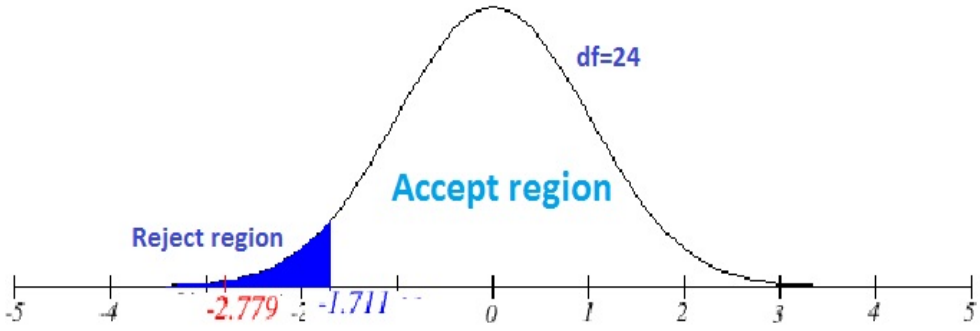
Step1: Construct the two hypotheses and specify the hypothesis type

H_0: The new antidepressant doesn't reduce recovery time, $\mu_{\bar{x}} = 8\ months$
H_1: The new antidepressant reduces recovery time, $\mu_{\bar{x}} < 8\ months$

Step2: Specify and calculate test statistics, calculate degrees of freedom
Test statistics= t-statistics
Test distribution =t- distribution
df= n-1= 25-1=24
Variance, $S_x^2 = 0.81$ and standard deviation; $S_x = 0.9$
Standard error of the mean: $S_{\bar{x}} = \frac{S_x}{\sqrt{n}} = \frac{0.9}{\sqrt{25}} = 0.18$

Then, $t = \frac{\bar{X} - \mu_{\bar{X}}}{S_{\bar{X}}} = \frac{7.5-8}{0.18} = 2.778$

Step3: Specify type I error level (usually α=5%)
Step4: Decision: to make a decision, plot a t-distribution with given degrees of freedom



The calculated t – statistics falls in reject region, therefore control hypothesis is rejected, that is, since the null hypothesis is rejected we can conclude that the new antidepressant reduces recovery time.

Example 2:

In health individuals hemoglobin in blood is measured to be 15mg/100ml and normally distributed

Randomly selected 16 individuals' hemoglobin mean and standard error of mean are found to be $15.6 \pm 0.4\ mg/100ml$. Based on hemoglobin amount in blood, can we conclude that these 16 individuals are healthy?

Step1: Construct the two hypotheses and specify hypothesis type

H_0: based on hemoglobin amount in blood, these 16 individuals are healthy.

$$H_0: \mu_{\bar{x}} = 15mg/100ml$$

H_1: based on hemoglobin amount in blood these, 16 individuals are not healthy.

$$H_1: \mu_{\bar{x}} \neq 15mg/100ml$$

Step2: Specify and calculate test statistics, calculate degree of freedom

Test statistics= t-statistics

Test distribution =t- distribution

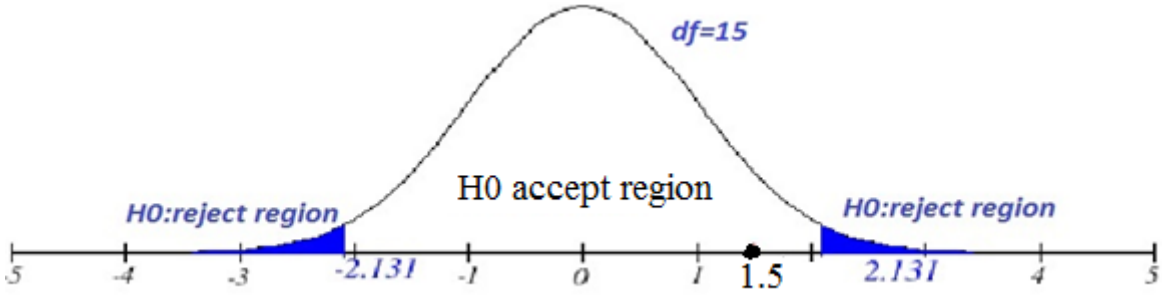
df= n-1= 16-1=15

Standard error of the mean = 0.4mg/100ml

Then, $t = \frac{\bar{X}-\mu_{\bar{X}}}{S_{\bar{X}}} = \frac{15.6-15}{0.4} = 1.5$

Step3: Specify type I error level (usually α=5%)

Step4. Decision



Calculated t-statistics falls in H_0 hypothesis accept region, therefore control hypothesis is accepted. Since the control hypothesis is accepted we can conclude that; based on hemoglobin amount in blood these 16 individuals are healthy.

9. Two sample t-test (comparison of two independent sample means)

Sometimes there might be a need to compare means of two samples where population is distributed normally and we do not know standard deviation of the population. This kind of analysis compare whether the difference between this two means is come out of chance or not.

Taking two samples from normally distributed population repeatedly; calculate means and calculate differences between these two means repeatedly shows a distribution. This distribution is called **sampling distribution of difference between means.** Sampling distribution of difference between means is normally distributed and has two parameters.

If the difference between two sample means that are drawn from the same population is really due to chance, their difference is accepted as zero.

If these two samples are taken from the population with unknown variance/standard deviation, the variance of the population is estimated from the calculated sample variances. This estimated variance is called pooled variance or weighted variance. Pooled variance is the best estimator of population variance and is calculated as follow:

$$S_x^2 == \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{(n_A - 1) + (n_B - 1)}$$

Testing the difference, between two means groups whether it is significant or not, is done using hypothesis testing. In the hypothesis testing, since population variance is estimated from sample variances, the test distribution is t-distribution and the test statistics is t-statistics. The test statistics for comparison of two means is calculated as follows:

$$t = \frac{(\overline{A} - \overline{B}) - \mu_D}{S_D} = \frac{(\overline{A} - \overline{B})}{S_D}$$

Where, μ_D is mean of sampling distribution difference between means, $\mu_D = 0$, and S_D is estimated standard deviation of Sampling distribution difference between means. Shortly, standard error of the differences between means. Standard error of the differences between means is calculated as follows:

$$S_D = \sqrt{\frac{\sum d_A^2 + \sum d_B^2}{(n_A - 1) + (n_B - 1)} * \frac{(n_A + n_B)}{(n_A n_B)}} = \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{(n_A - 1) + (n_B - 1)} * \frac{(n_A + n_B)}{(n_A n_B)}} \text{ if } n_1 \neq n_2,$$

If $n_1 = n_2 = n$, the expression above is simplified as follows;

$$S_D = \sqrt{S_{\overline{A}}^2 + S_{\overline{B}}^2}$$

The test statistics, t-value, shows t-distribution whit degrees of freedom of **$(n_A - 1) + (n_B - 1)$.**

Example 1. Final exam grades of 5 females, and 5 male students are givens as follows:

Female (A)	Male (B)
65	77
60	90
70	65
81	58
80	80

Based on the data
a) What is the best estimator of population variance, calculate.
b) Based on the data, is there statistically significant difference in exam grades between female and male students?

a)

$$S_x^2 \frac{\sum d_A^2 + \sum d_B^2}{(n_A - 1) + (n_B - 1)} = \frac{338.8 + 638}{(5-1)+(5-1)} = 122.1$$

b)

H_0: there is no statistically significant difference in exam grades between female and male students, that is, $H_0 : \mu_{\bar{A}} = \mu_{\bar{B}}$ or $\mu_{\bar{A}} - \mu_{\bar{B}} = 0$

H_1: there is statistically significant difference in exam grades between female and male exam score, that is, $H_1 : \mu_{\bar{A}} \neq \mu_{\bar{B}}$ or $: \mu_{\bar{A}} - \mu_{\bar{B}} \neq 0$

Standard error of the differences between means

Way1.

$$S_D = \sqrt{\frac{\sum d_A^2 + d_B^2}{(n_A - 1) + (n_B - 1)} * \frac{(n_A + n_B)}{(n_A n_B)}} = \sqrt{122.1 * \frac{(5+5)}{(5*5)}} \cong 6.99$$

Way2.

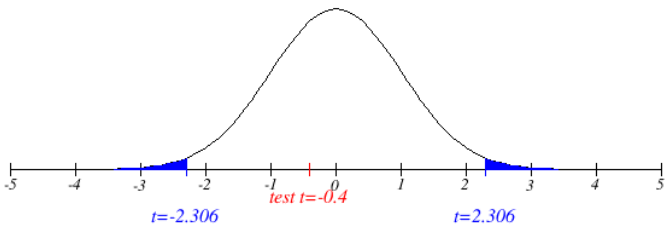
$$S_{\bar{A}}^2 = \left(\frac{S_A}{\sqrt{n_A}} \right)^2 = \frac{\sum \mathrm{d}_\mathrm{A}^2 / (n_A - 1)}{n_A} = \frac{338.8/4}{5} = \frac{84.7}{5} = 16.94$$

$$S_{\bar{B}}^2 = \left(\frac{S_B}{\sqrt{n_B}} \right)^2 = \frac{\sum \mathrm{d}_\mathrm{B}^2 / (n_B - 1)}{n_B} = \frac{638/4}{5} = \frac{159.5}{5} = 31.9$$

$$S_D = \sqrt{S_{\bar{A}}^2 + S_{\bar{B}}^2} = \sqrt{16.94 + 31.9} = 6.99$$

Test statistics is:

$$t = \frac{(\bar{A} - \bar{B}) - \mu_D}{S_D} = \frac{(71.2 - 74)}{6.99} = -0.4$$ with degrees of freedom of 8.



Since the calculated t-statistics falls in H_0 accept region, the control hypothesis is accepted. Therefore, there is no significant difference in exam grades between females and males final.

Example 2:

Out of 25 complaints on insomnia Administered in a hospital, 10 of them are treated with Drug-A and the rest of them are treated with Drug-B. After the treatment, means and standard errors of means of their sleeping time is reported as $\bar{A} \pm S_{\bar{A}} = 4.5 \pm 0.70 hrs$ and $\bar{B} \pm S_{\bar{B}} = 3.5 \pm 0.90 hrs$ respectively for the two drugs. Based on these results, can we conclude Drug-A is more effective?

The objective of this research (base on correcting sleeping time of these insomnia patients) is to examine whether the effect of Drug-A is better than Drug-B. To conclude the effect of Drug-A is better than Drug-B, sleeping time of patients treated with Drug-A must be longer than those who were treated with Drug-B. This implies that the hypothesis, which is going to be tested, is one-tailed.

In this case the two hypothesis are as follows:

H_0: The difference in sleeping time between patients who were treated with Drug-A and Drug-B is not statistically significant. Based on improving sleeping time we cannot conclude Drug-A is more effective than Drug-B.

$H_0 : \mu_{\bar{A}} = \mu_{\bar{B}}$ or $\mu_{\bar{A}} - \mu_{\bar{B}} = 0$

H_1: The difference in sleeping time between patients who were treated with Drug-A and Drug-B is statistically significant. Based on improving sleeping time we can conclude Drug-A is more effective than Drug-B.

$H_1 : \mu_{\bar{A}} > \mu_{\bar{B}}$ or $: \mu_{\bar{A}} - \mu_{\bar{B}} > 0$

$n_A = 10 : \bar{A} \pm S_{\bar{A}} = 4.5 \pm 0.70 hrs$

$n_B = 15 : \bar{B} \pm S_{\bar{B}} = 3.5 \pm 0.90 hrs$

Based on this calculation

$$S_{\bar{A}} = \frac{S_A}{\sqrt{n_A}}$$

$$\Rightarrow S_{\bar{A}}^2 = \left(\frac{S_A}{\sqrt{n_A}} \right)^2 \Rightarrow S_A^2 = S_{\bar{A}}^2 n_A$$

$$S_A^2 = \frac{\sum_{i=1}^{n} \left(A_i - \bar{A} \right)^2}{\left(n_A - 1 \right)} = \frac{\sum d_A^2}{\left(n_A - 1 \right)}$$

$$\sum d_A^2 = S_A^2 n_A \left(n_A - 1 \right) = \left(0.7 \right)^2 * 10 * \left(10 - 1 \right) = 44.1$$

$$S_{\bar{B}} = \frac{S_B}{\sqrt{n_B}}$$

$$\Rightarrow S_{\bar{B}}^2 = \left(\frac{S_B}{\sqrt{n_B}}\right)^2 \Rightarrow S_B^2 = S_{\bar{B}}^2 n_B$$

$$S_B^2 = \frac{\sum_{i=1}^{n}\left(B_i - \bar{B}\right)^2}{\left(n_B - 1\right)} = \frac{\sum d_B^2}{\left(n_B - 1\right)}$$

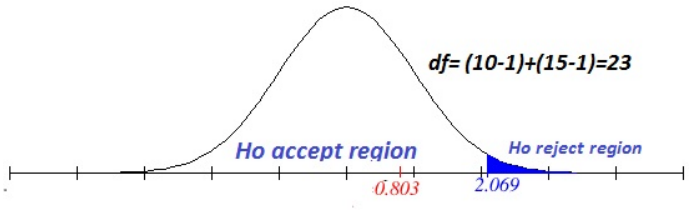
$$\sum d_B^2 = S_B^2 n_B\left(n_B - 1\right) = \left(0.9\right)^2 * 15 * (15 - 1) = 170.1$$

$$S_D = \sqrt{\frac{\sum d_A^2 + d_B^2}{\left(n_A - 1\right) + \left(n_B - 1\right)} * \frac{\left(n_A + n_B\right)}{\left(n_A n_B\right)}} = \sqrt{\frac{44.1 + 170.1}{\left(10 - 1\right) + \left(15 - 1\right)} * \frac{\left(10 + 15\right)}{\left(10,*15\right)}} \cong 1.246$$

$$t = \frac{\left(\bar{A} - \bar{B}\right) - \mu_D}{S_D} = \frac{\left(4.5 - 3.5\right)}{1.246} = 0.803$$

Specify type I. error α=0.05, and degree of freedom

Decision:



Conclusion: since the calculated t-statistics falls in H_0 accept region, the control hypothesis is accepted. Therefore, based on improving sleeping time, there is no significant difference between Drug-A and Drug-B.

10. Comparison of Two Dependent Groups

In some experiments, the groups to be compared may be dependent, which means that measurements in two groups are taken from the same individuals. For instance, measurements of blood pressure before and after treatment in same individual, weight loose before and after 6-week exercise, etc. The measurements taken from the same individual in groups is called pair.

Test used to compare two dependent groups is called paired t-test. Paired t-statistics is calculated as follows

$$t = \frac{\bar{D} - \mu_{\bar{D}}}{S_{\bar{D}}}$$

Where

$\bar{D}$: mean of the difference between paired observations (Before-After or After-Before)

$\mu_{\bar{D}}$: Population mean of the difference between paired observations, it is accepted as zero.

$S_{\bar{D}}$: The standard error of mean of differences. It is calculated as follows;

$$S_{\bar{D}} = \frac{S_D}{\sqrt{n}} = \sqrt{\frac{S_D^2}{n}}$$

S_D: standard deviation of differences

S_D^2: variance of differences

**Calculated paired t-test shows t-distribution with (n-1) degrees of freedom.

If the hypothesis is two-tailed, the sign of calculated t-statistics depends on "before-after" or "after- before" subtraction. If the hypothesis is one-tailed, when alternative hypothesis is constructed the differences are taken into consideration.

Example 1;

The measurement of 6 hypertension patients' blood pressure before and after treatment is measured as follows. Can we conclude that the treatment is effective in decreasing blood pressure of patients?

Before treatment	150	155	140	165	155	165
After treatment	110	135	145	125	155	120

H_0 : The mean of differences after treatment and before treatment can be accepted as zero. Means, the treatment is not effective in decreasing blood pressure of patients. Shortly,

$H_0 : \mu_{\bar{D}} = 0$

H_1: the mean of differences after treatment and before treatment cannot be accepted as zero. Means, the treatment is effective in decreasing blood pressure of patients. Shortly,

The alternative hypothesis is constructed in two ways

First way; If we calculate differences by subtracting before from after

$H_0 : \mu_{\bar{D}} < 0$

Second way; If we calculate differences by subtracting after from before

$H_0 : \mu_{\bar{D}} > 0$

Let's calculate differences by subtracting before treatment from after treatment;

Before	150	155	140	165	155	165
After	120	135	145	125	155	150
Differences (after-before)	-30	-20	5	-40	0	-15

$$\sum D = (-30) + ... + (-15) = -100$$

$$\bar{D} = \frac{-100}{6} = -16.67$$

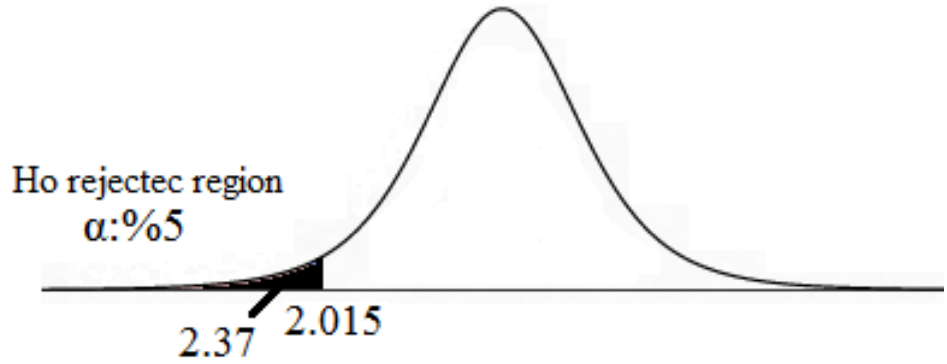
$$\sum d_D^2 = (-30)^2 + ... + (-15)^2 - \frac{(-100)^2}{6} = 1483.33$$

$$S_{\bar{D}}^2 = \frac{\sum d_D^2}{n-1} = \frac{1483.33}{6-1} = 296.67$$

$$S_{\bar{D}} = \sqrt{\frac{S_{\bar{D}}^2}{n}} = \sqrt{\frac{296.67}{6}} = 7.03$$

$$t = \frac{\bar{D} - \mu_{\bar{D}}}{S_{\bar{D}}} = \frac{-16.67 - 0}{7.03} = -2.37$$

Degree of freedom=(n-1) = (6-1) =5. Critical table value for 5 degree of freedom for one - tailed hypothesis control is 2.015.



H_0 hypothesis is rejected, which means that the treatment is effective in decreasing blood pressure of patients.

Example 2;

The measurement of 8 students' body temperature before and after an exam is measured as follows. Can we conclude that the exam is changed body temperature?

Before exam	37.5	36.8	37.1	36.8	37.4	36.1	36	37
After exam	37.2	37	37.1	37	37.3	36.3	36.1	36.9

H_0 : The mean of differences in body temperature after exam and before exam can be accepted as zero, which means that the exam does not change body temperature. Shortly;

$H_0 : \mu_{\bar{D}} = 0$

H_1: The mean of differences body temperature after exam and before exam cannot be accepted as zero, which means that the exam changes body temperature. Shortly;

$H_0 : \mu_{\bar{D}} \neq 0$

Let's calculate differences by subtracting after exam from before exam;

Before exam	37.5	36.8	37.1	36.8	37.4	36.1	36	37
After exam	37.2	37	37.1	37	37.3	36.3	36.1	36.9
Differences (before-after)	0.3	-0.2	0	-0.2	0.1	-0.2	-0.1	0.1

$$\sum D = 0.3 + ... + 0.1 = -0.2$$

$$\bar{D} = \frac{-0.2}{8} = -0.025$$

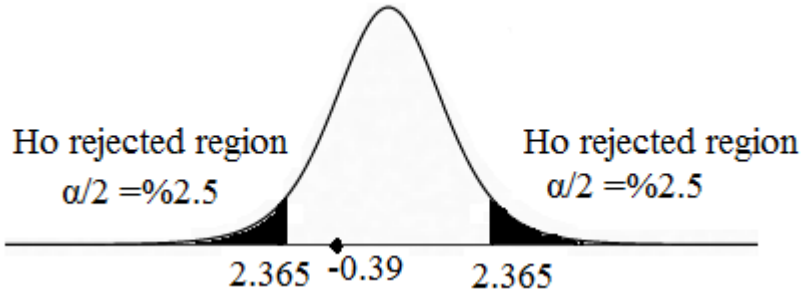
$$\sum d_D^2 = 0.3^2 + ... + 0.1^2 - \frac{(-0.2)^2}{8} = 0.235$$

$$S_{\bar{D}}^2 = \frac{\sum d_D^2}{n-1} = \frac{0.235}{8-1} = 0.0336$$

$$S_{\bar{D}} = \sqrt{\frac{S_{\bar{D}}^2}{n}} = \sqrt{\frac{0.0336}{8}} = 0.0648$$

$$t = \frac{\bar{D} - \mu_{\bar{D}}}{S_{\bar{D}}} = \frac{-0.025 - 0}{0.0648} = -0.39$$

Degree of freedom=(n-1)=(8-1)=7. Critical table value for 7 degree of freedom for two-tailed hypothesis control is 2.365.



Decision: H_0 hypothesis is accepted, which means that the exam does not change body temperature.

REFERENCES:

1. **Biostatistics, Basic Concepts and Methodology for the Health Sciences, By Wayne W. Daniel, Prepared By: Sana A. Abunasrah, Revision By: Saba M. Alwan**
2. Lectures of Stat -106 (Biostatistics) Text book Biostatistics Basic Concepts and Methodology for the Health Sciences By Wayne W. Daniel
3. **Biostatistics,** Frank H. Osborne, Ph. D. Professor, https://slideplayer.com/slide/5151033/
4. **Basic Biostatistics,** Presenter: Lesego Gabaitiri, PhD (UB)
5. Lectures of Bio733Applied Biostatistics Biostatistics Basic Concepts and Methodology for the Health Sciences By Wayne W. Daniel
   vulms.vu.edu.pk › Biostatistics-Handouts
6. King Saud University College of Science Department of Statistics & OR STAT – 145 BIOSTATISTICS Summer Semester 1431/1432 Lectures' Notes Prof. Abdullah Al-Shiha

http://fac.ksu.edu.sa/sites/default/files/Note-145%20biostat%20Prof.%20Abdullah%20Al-Shiha.pdf

7. **Biostatistics,** Frank H. Osborne, Ph. D. Professor, https://image.slideserve.com/1088723/biostatistics-l.jpg
8. DANIEL, W. W. (1995). Biostatistics, John Wiley & Sons. Inc. USA
9. SNEDECOR, W. and COCHRAN W. G. 1980. Statistical Methods. Seventh Edition. The Iowa state University Press, Ames, Iowa, USA.
10. STAT 6200 Introduction to Biostatistics Lecture Notes

https://faculty.franklin.uga.edu/dhall/sites/faculty.franklin.uga.edu.dhall/files/lec1.pdf

11. Introductory Statistics , Saylor URL: http://www.saylor.org/books
12. **Basic Definitions and Concepts**

https://saylordotorg.github.io/text_introductory-statistics/s05-01-basic-definitions-and-concepts.html

13. Normal Distribution, https://web.stanford.edu › hrp259 › lecture6
14. Properties of Normal Distribution
    https://faculty.elgin.edu/dkernler/statistics/ch07/7-1.html
15. The Normal Distribution

https://edisciplinas.usp.br/pluginfile.php/1153982/mod_resource/content/1/Sharpe_vellemman_cap9.pdf