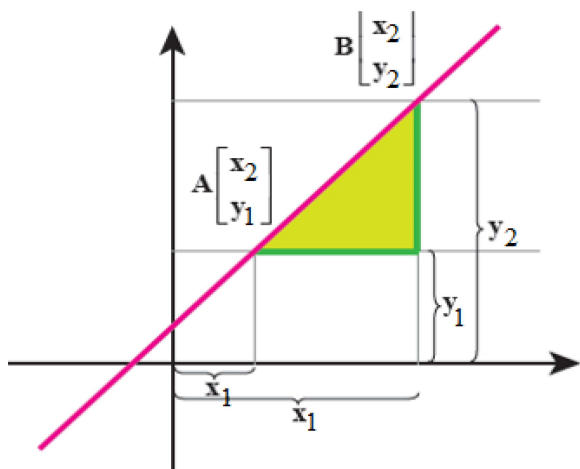
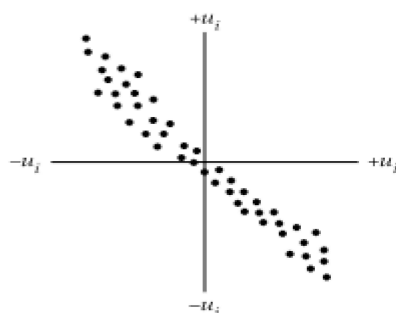
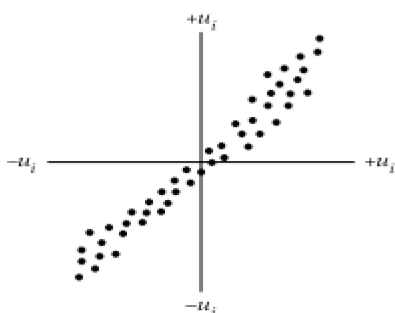


Statistical Methods and Probabilities in Agricultural Science



$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{S_x}{\sqrt{n}} < \mu_x < \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{S_x}{\sqrt{n}}$$



Written by:
Dr. Mohsen MIRZAPOUR
Dr. Saeid HEYDARZADEH
Dr. Harun GİTARİ

STATISTICAL METHODS AND PROBABILITIES IN AGRICULTURAL SCIENCE

Written by:

Dr. Mohsen MIRZAPOUR

Dr. Saeid HEYDARZADEH

Dr. Harun GİTARİ



Copyright © 2022 by iksad publishing house

All rights reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means, including photocopying, recording or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law.

Institution of Economic Development and Social Researches Publications®

(The Licence Number of Publicator: 2014/31220)

TURKEY TR: +90 342 606 06 75

USA: +1 631 685 0 853

E mail: iksadyayinevi@gmail.com

www.iksadyayinevi.com

It is responsibility of the author to abide by the publishing ethics rules. The first degree responsibility of the works in the book belongs to the authors.

Iksad Publications – 2022©

ISBN: 978-625-8246-62-9

Cover Design:

November / 2022

Ankara / Türkiye

Size = 16x24 cm

CONTENTS

Preface	1
CHAPTER 1	
DEFINITIONS AND SOME TERMS OF STATISTICAL DISTRIBUTION	
Assist. Prof. Dr. Mohsen MIRZAPOUR	3
CHAPTER 2	
CENTRAL TENDENCY AND DISPERSION INDICATORS	
Dr. Saeid HEYDARZADEH	45
CHAPTER 3	
POSSIBILITIES	
Dr. Harun GĪTARĪ	121
CHAPTER 4	
RANDOM VARIABLE, MATHEMATIC EXPECTATION AND BINOMIAL DISTRIBUTIONS	
Assist. Prof. Dr. Mohsen MIRZAPOUR	
Dr. Saeid HEYDARZADEH	167
CHAPTER 5	
STATISTICAL ASSUMPTION TEST	
Assist. Prof. Dr. Mohsen MIRZAPOUR	
Dr. Harun GĪTARĪ	243

CHAPTER 6

NORMAL DISTRIBUTIONS, T, CHI-SQUARE, F AND ANALYSIS OF VARIANCE

Dr. Saeid HEYDARZADEH

Dr. Harun GĪTARĪ283

CHAPTER 7

THE RELATIONSHIP BETWEEN VARIABLES

Assist. Prof. Dr. Mohsen MIRZAPOUR

Dr. Saeid HEYDARZADEH

Dr. Harun GĪTARĪ353

CHAPTER 8

ANSWERS TO THE EXERCISES

Dr. Saeid HEYDARZADEH437

REFERENCES479

APPENDIX TABLES496

Preface

Science is a collection of facts that are gradually discovered by humans and classified into different branches. The goal of science is to understand the concepts of nature and to achieve greater prosperity and happiness for humans. Science can be compared to a big building whose lower floors are complete but the upper floors are being built and different people are working harmoniously in different parts of this big structure. The building of science is rising day by day to bring prosperity and perfection to man. Therefore, the most important properties of collective science are its continuity and modifiability.

Research is a process and a path that leads to the discovery of science. Investigating (on weight of activation) means to seek the truth, and literally means to discover the truth and deal with the truth of a matter. In other words, research entails a continuous effort and follow-up to uncover the unknown.

Today, we see that scientists have explained and presented amazing scientific laws, principles, and theories in the field of various sciences, and all these achievements are a result of many years of research and investigation. So research has always been associated with innovation and answers to new questions, the result of which is human progress and excellence.

What is scientific research?

A: Statement and definition of the problem and review of sources

B: Hypothesizing:

H_0 Assumption zero assumption: based on the principle of innocence,

Hypothesis H_1 or the opposite hypothesis: it means that it affects, the guilt of the subject and...

C: Conducting an experiment: collecting evidence and information for and against the hypothesis

D: Analysis of the results (data):

Various simple and advanced methods can be used as needed to analyze, analyze and interpret the results.

For this, it is better to consult and exchange opinions with people familiar with the science of statistics.

E: Interpreting the results and generalizing them to society:

The last step in scientific research is to interpret the results and express them in the form of a relationship or scientific law. The generalization of the results to a more general state or states (society) is part of the goals of scientific research. This issue is one of the basic concepts in inferential statistics. In other words, the generalization of the results to society is called statistical inference or statistical induction.

Statistics as a scientific research tool provides the possibility of conducting, facilitating, and accelerating scientific research. This tool has developed so much that it has become a science and is taught as a University course. Therefore, statistics prevent the researcher from deviating from the main research path. The contents of this book, which was written based on the new chapters of the Ministry of Science, Research and Technology for students of agricultural fields, can increase the technical knowledge of students and farmers to achieve the above goals, as well as the prerequisites for various courses in agriculture, such as plant genetics, basics to plant species, the basics of plant biotechnology, experimental designs in agricultural sciences 1, experimental designs in agricultural sciences 2 and advanced statistical methods for undergraduate and graduate students in various fields of agriculture (especially agriculture) and also as a useful collection for researchers and agricultural researchers be used.

It is obvious that reflecting the opinions of experts, researchers and students will be grateful and will improve this work in future editions. We are extremely grateful to all those who helped us in the various stages of writing and publishing this work.

Written by:

Dr. Mohsen MIRZAPOUR

Dr. Saeid HEYDARZADEH

Dr. Harun GĪTARĪ

CHAPTER 1

**DEFINITIONS AND SOME TERMS OF
STATISTICAL DISTRIBUTION**

Assist. Prof. Dr. Mohsen MIRZAPOUR¹

¹ Siirt University, Faculty of Agriculture, Department of Agricultural Biotechnology, Siirt, Türkiye
ORCID ID: 0000-0002-2898-6903, e-mail: m.mirzapour@siirt.edu.tr

INTRODUCTION

Various definitions have been used for statistics. In general, statistics is defined as the science and art of collecting, classifying, and adjusting observations and analyzing and interpreting them, and interpreting the results of census and experiments. In other words, statistics refers to a set of techniques and methods that are used in collecting, classifying, purifying, analyzing, and interpreting statistical information.

Statistical methods can be divided into descriptive statistics and inferential statistics. Descriptive statistics is that part of statistics that deals with collecting, summarizing, and explaining the characteristics of a series of observations, without making any conclusions beyond that information. The purpose of descriptive statistics is to organize, summarize data (based on mean, median, median, standard deviation, and quartile deviation, etc.) and describe sample sizes. Setting up information clearly and understandable way is the task of descriptive statistics. The most important use of descriptive statistics is to summarize data. Descriptive statistics are methods that provide data processing.

While in this part of inferential statistics, the information obtained from descriptive statistics is analyzed and based on the analysis, conclusions and inferences are made and this conclusion is generalized to the whole society or similar cases. It should be known that the generalization of information from the sample to the population is the task of inferential statistics. This action is done using the process of deductive reasoning and is based on the theory of possibilities. In

other words, its goals are to find out the characteristics of the population by measuring the characteristics of the sample, and in this direction, the effect of chance errors is measured in order to reduce them.

In inferential statistics, how close the predicted value is to the actual value depends on two factors:

1- Sampling methods 2- Statistical methods used.

Inferential statistics are used to predict or estimate population parameters through sample sizes.

In contrast to parametric statistics, the basic assumption of which is the normality of the statistical population. Parametric statistics are used for distance and relative scales. In non-parametric statistics, which is mostly used in behavioral and human sciences, the distribution of the statistical population is free, and therefore it is called free distribution. It includes statistical methods whose acceptance is not based on precise and firm assumptions. It should be known that in parametric statistics, 1- the distribution of community variables should be normal, 2- a random sample should be selected, and 3- there are quantitative and continuous variables. If the distribution is not normal due to the lack of sample number and the sampling is biased (not random) and the variable is rank or nominal, non-parametric statistics are used.

A major factor in the difference between parametric and nonparametric statistical methods is their computational methods.

The advantage of using non-parametric methods is that:

1- Its formulas are simple.

2- Using them is simpler and less time-consuming compared to parametric methods. The most important weakness of non-parametric

methods is related to their power in rejecting the null hypothesis (when the hypothesis is false).

1.1. Measurement error

The difference between the actual value and the measured value is called measurement error. The error is usually denoted by E , and its calculation is "approximate value less actual value = E ". Note that the error value (E) can be positive or negative and must be less than the unit of measurement.

1.2. Population and statistical sample

Statistical population: a set of people or objects or elements that have at least one attribute in common about which we want to study a subject or subjects. The definition of the statistical population is important because the final results of the research can only be generalized to the statistical population. The number of elements in a community is called community size (n).

Limited statistical population: It is a population whose number of elements can be counted. Such as the University of Tehran student society.

Unlimited statistical population: a population whose number of elements is uncountable. Such as the society of stars in the sky.

Community size: The number of members of a community is called the size of that community.

As mentioned, in inferential statistics, the characteristics of the society are estimated from the characteristics of the sample. Therefore,

in statistics, it is valuable and valid to have the following characteristics:

1- Not being biased 2- Estimation stability 3- Estimation efficiency 4- Sufficient estimation (getting the best result).

Unbiased: An unbiased estimate in which the calculated statistical index does not have a regular tendency to be larger or smaller than the parameter. The sample mean is usually an unbiased estimate because if we select many samples from the population, the mean of their means is equal to the population mean, the opposite of the variance is a biased estimate.

Skewness in a distribution indicates its degree of skewness.

Consistency: The estimation method has consistency when the value of the estimated statistical index becomes closer to the population as the sample increases.

Increasing the sample size leads us to the stability index.

Efficiency: An efficient estimate is an estimate that gives the best result without losing information.

Sufficiency: the estimation is sufficient when the sampling variance of that method is smaller than another method. The greater the estimated efficiency, the greater its adequacy.

1.3. Adjective

An attribute is a quantity or quality that belongs to the elements of the statistical community.

Fixed trait: It is the trait that is common to all elements of the statistical population. For example, in the community of Urmia University students, being a student can be considered a fixed attribute.

Variable attribute: The characteristics that the researcher observes and measures are called variables. It is a category that may have a different status for different people. In other words, it is an attribute that can change from one person to another among the elements of the statistical community. In general, any quantity that changes from one value to another is a variable. For example, in the community of Urmia University students, height and age can be considered variable attributes.

Random variables: The subject or subjects under study are called random variables. Like the height of people or the amount of income of people.

We have two types of random variables: quantitative variables and qualitative variables.

Quantitative Variable Adjective: Whenever the size of an attribute can be measured with common tools and expressed with a single number, it is called a quantitative variable adjective. Such as age, height, and weight. Quantitative variables in society are either continuous or discontinuous. A quantitative variable attribute is called continuous when its numerical size is obtained through counting, that is, its changes can be expressed with a set of natural numbers. A quantitative variable attribute is called continuous when it is obtained through measurement, that is, its changes can be expressed with a set of real numbers such as weight, distance, length, etc.

We have two types of quantitative variables:

a) Continuous: It is a variable that can choose any point or value between two units or two points. For example, weight is a continuous variable that can be between zero and infinity. Continuous variables are inherently quantitative and have different degrees of measurement.

b) Discrete: A quantitative variable is discrete if it is not continuous. In other words, it is a variable that cannot choose any value between two units or two points. For example, gender, number of accidents, etc.

Qualitative variable attribute: When an attribute cannot be measured with common tools and expressed as a single number, it is called a qualitative variable attribute. Such as skill, beauty, and talent. Qualitative variables in society are either ordinal or nominal. Qualitative variables in which there is some kind of natural order are called ordinal qualitative variables. Such as stages of life or stages of education. A qualitative variable that is not ordered is called a nominal variable such as eye color and blood type.

Independent variable: It is a variable that the researcher manipulates to measure changes in the dependent variable through it.

Dependent variable: It is a variable whose value or value (its presence or absence) depends on the independent variable.

Nominal: Qualitative variables that have no order. Like blood group and...

Ordinal: Variables that have a natural order. Like the seasons, stages of life and...

1.4. Scales

It should be known that a basic principle in classification is that the variables themselves can be measured. Measurement is to determine the numerical value of an attribute according to a predetermined rule, measurement can be done at different levels. The meaning of scale is actually the level of measurement. There are four important types of measurement scales for variables, which are:

Class scales: this type of measurement is the act of classification. When we divide a set of objects into several classes, we can represent each class with the numbers 1, 2, 3, etc. These numbers have no concept other than calling the class and from abstract properties, for this reason, other conventional signs such as B, A and C can be used instead of numbers. In this scale, firstly, the classes are an obstacle to a gathering, i.e. one person cannot be placed in more than one class, secondly, the scale must be complete, that is, all people of the same type should be placed in each class. In this scale, only mathematical symbols = and \geq are used. In this scale, the frequency of classes can be analyzed with statistical methods.

There are two types of scales, nominal scale, and sorted class scale.

In the nominal scale, numbers are used only for naming and have no real quantitative value. This scale is used for qualitative variables. When we assign numbers to qualitative variables, nominal data is obtained. Nominal data are numerical in name only, as they have none of the properties of numbers that we deal with in ordinary arithmetic. The nominal scale has specific and distinct classes that have a

qualitative aspect and the only existing relationship is their difference from each other. Eye color is an example of a variable in the nominal scale.

The other type of class scale is related to each other in a quantitative sense and is also expressed in ordered classes. For example, in a questionnaire where each question has very high, high, medium, low, and very low categories.

Ordinal scale: This scale has the main characteristic of the nominal scale (different numbers mean different things) plus the characteristic of "greater than" or "smaller than". In this scale, there are no fixed classes to assign observations to. Observations are compared with each other and ranked in a certain order, for example, from largest to smallest. From this point of view, ordinal scale data is similar to ordered class scale data. In general, the numbers in the ordinal scale only show the hierarchy of objects or events in terms of a certain characteristic of a group and without saying anything about the distance between people, it specifies their relative position in terms of a characteristic.

Distance scale: In addition to classifying, naming, and sorting classes, the distance scale also provides the possibility of recognizing the distances between people, objects, or events. This scale, in addition to the two properties of the nominal scale and the order that was stated, the mathematical operations of addition and subtraction can also be performed, it also has the property that the distances between the numbers are equal. By "equal distances" it means that the distance

between things represented by 3 and 5 is equal to the distance between things represented by 7 and 9.

One of the limitations of the interval scale is not having an absolute zero. Therefore, it cannot be assumed that, for example, in an academic achievement test, the score of the third-ranked student is equal to twice the sixth-ranked one. And if a student scores zero on an intelligence test, it doesn't mean that he doesn't have intelligence.

Proportional scale: In addition to all the characteristics of nominal, ordinal, and interval scales, this scale also has the following characteristic: there is a true zero point for it, which indicates the complete absence of the measured quantity. On a proportional scale, zero means "nothing". The data obtained with the help of this scale include all measurements related to length, weight, area, pressure, time, sound intensity, speed, and the like.

The relative scale is the highest level of measurement that, due to having absolute zero, all mathematical operations including addition, subtraction, multiplication, division, etc. can be performed on this scale. The use of this scale has a more meaningful application in physical sciences.

Census: If we study all the members of the society, then we have a census. The problems of the census are:

- 1- Unavailability of all members of society
- 2- Taking time
- 3- Not being affordable
- 4- Destruction of society or part of it

1.5. Statistical sample

The sample is a small part of the statistical population, which is selected as a model of the target population in terms of the studied trait with scientific methods and should be a complete statement of the state of the society in terms of the studied trait. The number of elements in a sample is called sample size (n). In other words, we select a subset of the statistical population that represents all the main characteristics of the population. Sampling is done based on random principles and the purpose of studying the sample is to know more about the community. A good random sample has the following characteristics in selecting its members.

- 1- It is possible to choose any member of the society
- 2- The same chance of choice for all members of the society

The study of statistical indicators in a society is not possible due to a large number of people in the society and even the fact that some societies are hypothetical, so they are estimated based on the sample. Sampling should be random to get correct results. In addition, the more people in the sample, the more accurate the estimation of population characteristics will be. In statistics, if the sample size is more than 30, it is called a large sample and if it is less than 30, it is called a small sample.

Data or observation: These are the sizes of the variable attributes of the elements of society or a statistical sample obtained by using measurement, testing, observation, etc. In other words, it is called the raw results of an experiment, which statistically includes two types:

qualitative variables that have a discrete distribution and quantitative variables that have a continuous distribution.

Indicators related to society are called parameters, whose number is fixed.

The indicators related to the sample are called statistics whose value is variable.

Parameter: The results of community measurement are called parameters. In other words, information or measurements that describe the characteristics of society are called parameters.

The difference between statistic and parameter is that the statistic is related to the sample and the parameter is related to the society. The characteristic feature of a community is called a parameter. The statistic (characteristics of the sample) estimates the parameter (characteristics of the population).

1.6. Data collection methods

The results obtained from the measurement or examination of a sample are called data or observation.

Data collection methods:

- 1- Using pre-prepared data
- 2- Question
- 3- Observing and recording events
- 4- Performing the test

Addition rules in statistics

One of the signs that is widely used in statistics is the plus sign. This symbol is represented by Σ and is one of the Greek letters called Sigma. Sum limits are shown below and above the sigma mark.

Example 1-1: Consider the following numbers.

$$x_1 = 1 \quad x_2 = 3 \quad x_3 = 5 \quad x_4 = 7$$

$$\sum_{i=1}^n x_i = 1+3+5+7 = 16$$

Other algebraic operations are also used using the sigma sign.

$$-\sum_{i=1}^n C = nC$$

$$-\sum_{i=1}^n ax_i = a \sum x_i$$

$$-\sum_{i=1}^n (x_i + y_i) = \sum x_i + \sum y_i$$

Example 1-2: What is the result of the following expression? In the event that:

$$\sum_{i=1}^{10} x_i = 10$$

$$\sum_{i=1}^{10} (2x_i + 3) = 2 \sum_{i=1}^{10} x_i + \sum_{i=1}^{10} 3 = 2 \times 10 + 30 = 50$$

1.7. Classification and arrangement of observations

For statistical analysis of a data series, the best thing to do is to categorize it first. Sometimes the number of observations and their varieties are so great that placing an observation in a separate class leads to the production of tables with too many classes. In these cases, it is

better to use the classification in frequency distribution tables, and in this method, a series of observations is placed in a specific row. In other words, the organization of observations in statistics is called frequency distribution. Frequency distribution is a suitable tool for summarizing and specifying the main characteristics of raw research data.

Unclassified frequency distribution table: To set this table, we put the number of members of each state (absolute frequency) in one column of different variable states and in the other column. In other words, if we have a series of raw data, we do the following steps to draw the frequency table of the raw data.

1- In one column, we write all the numbers in order from the largest number (score) to the smallest number (score).

2- In the next column, we write the absolute frequency of numbers or f (the number of repetitions of each number).

3- In the next column, in front of each number, we put a slash (/) as many times as it is repeated. This is optional.

To set up the classified frequency distribution table, the following steps are performed.

1- Calculation of the range of changes

2- Calculate the number of floors

3- Calculate the length of the floors

4- Calculation of floor boundaries

5- Count the elements that are on each floor (abundance).

Range of changes: In a set of numbers, the overall range of changes is the difference between the smallest and largest observation.

Floor limits: The distance between the lowest (lower limit) and the highest (upper limit) observation in a floor is called the range of a floor. The range of classes can be equal or unequal, which is better.

The average grade of a class: It is used as a representative of that class and it is the average of the two upper and lower limits of the class, which are represented by X_C .

$$X_C = (\text{lower limit} + \text{upper limit})/2$$

To determine the number of classes and their limits, you can act as you like, but when the number of observations is large, it is better to follow the following method:

1- We determine the general scope of the changes.

$$R = \text{Max} - \text{Min}$$

2- We calculate the number of categories using the following relationship.

$$K = 1 + 3/3 \log N$$

K = number of categories

N = total number of observations

3- The distance between floors is obtained by dividing the following relationship.

$$C = \frac{R}{K}$$

If the number of categories is less than 10, the accuracy of information interpretation and review decreases, and if the number of categories is more than 20, the power of information review and interpretation is high, but the calculation becomes difficult.

If there is a decimal place, we round it. For example, we consider $3/7$ as 4.

The number of categories is never a decimal, and if it is a decimal, we should consider the higher number.

After determining the range of changes and the number of categories and the distance between classes, we draw a frequency table.

We write the classes in one column. In the next column, we write about the category. Each floor has two numbers. If we subtract half a unit from the number on the left side and add half a unit to the number on the right side, we get the size of the handle.

In the next column, we write the absolute frequency of numbers or f (the number of repetitions of each number).

The sum of the absolute frequency of data is equal to the total number of data.

To calculate the cumulative frequency (F_c), the absolute frequencies of each score are added from bottom to top.

To estimate the relative frequency of each grade, the f/N formula is used, that is, the simple frequency of each grade divided by the sum of grades.

The total number of data can be determined using the following formula:

Total number of data = (class of each absolute frequency) / (relative frequency)

Cumulative frequency percentage (accumulative) or percentage cumulative frequency, for this we add the relative frequencies from bottom to top. Or we divide the cumulative frequency of each class by the sum of numbers and multiply by 100.

The cumulative percentage frequency of the highest score must be equal to 1 or 100 percent.

We calculate the percentage of relative frequency (percentage relative frequency) and percentage of cumulative frequency (percentage cumulative frequency) and cumulative relative frequency (cumulative) as follows.

Relative frequency percentage = $(f \text{ (frequency)}) / (N \text{ (total numbers)}) \times 100$

Cumulative frequency percentage = $(F_c \text{ (cumulative frequency)}) / (N \text{ (total numbers)}) \times 100$

Cumulative frequency shows the position of single scores, that is, it shows how many scores are less than a certain score. Cumulative frequency percentage also shows the position of people in the class as a percentage. For example, if the relative cumulative frequency of a person in the class is 96, it means that 96% of the people in the class have scored less or equal to him.

(point) middle or median (X_c) of each floor through the smallest and largest number of each floor, we add and divide by 2.

To classify the raw data, the number of classes, the range of changes, and the distance between classes are required.

Example 1-3: The random variable investigated is the height of the children in the class, in the class, we have 160 to 180 cm tall. So the range of changes

$$R = \text{Max} - \text{Min} = 180 - 160 = 20 \text{ cm}$$

If we indicate the length of each group with c and the number of groups with k , it is obvious that, for example, if we divide $\{180, 160\}$ into 4 groups ($K = 4$), the length of each group will be ($C = 5$).

$$C = (R)/K = 20/4 = 5$$

Now 165 is the upper limit of the first category and the lower limit of the second category. That is, for example, the center of the third category becomes 172.5, which means:

$$(175 + 170)/2=172.5$$

The distance between the centers of two consecutive bunches is the same as the length of the bunch. Now, if you have the center of n and the center of m , the length of the bunch is obtained from the opposite relationship.

$$C = \frac{X_n - X_m}{n - m}$$

$$C = \frac{X_4 - X_1}{4 - 1} = \frac{177.5 - 162.5}{3} = \frac{15}{3} = 5$$

Batch width (h) is the difference between the upper limit and the lower limit of a batch. If the upper limit of one category and the lower limit of the next category is different. By halving this distance and adding it to the upper limit of each category and subtracting it from the lower limit of each category, the categories can be bordered, which is called the category limit or the actual category limits.

Example 1-4: We consider the following frequency table:

Category	4-7	8-11	12-15
Frequency	10	2	8

The upper limit of one category and the lower limit of the next category differ by 1 unit. Therefore, we add 0.5 units to the upper limit of each category and subtract 0.5 units from the lower limit of each category.

Category	3.7-5.5	5.5-7.11	5.5-11.15
Frequency	10	2	8

1.8. Abundance charts

It should be known that the information in the table cannot be understood quickly unless it is studied part by part. To interpret the data in a simple and objective way, they are displayed in the form of graphs or geometric shapes. Another method that is used in addition to frequency distribution tables to classify and adjust observations is the frequency chart method. The data display method is selected according to the type of data scale. If the data scale is interval and relative, quantitative charts (can be used for continuous and discrete statistical distribution) and if the data scale is nominal or ordinal, descriptive charts will be used for the geometrical representation of qualitative data. The most important quantitative diagrams are histogram diagrams, polygon diagrams, cumulative frequency diagrams, and exploratory data analysis (branch and leaf diagram and box diagram). The most important descriptive charts are column charts and pie charts as follows:

1.8.1. Bar chart

A bar chart is useful when the data collected is discrete, quantitative, or qualitative (such as interest rates, the number of men and women in a class, or passing and failing) and is measured using a nominal scale. When data are measured using a nominal scale. The best chart to display data is a bar chart. To draw this graph, we specify the variable under consideration on the x-axis and the absolute or relative frequency on the y-axis. This chart is used for discrete, qualitative, and quantitative variables.

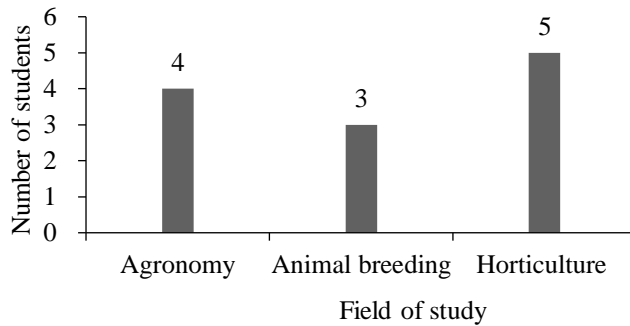
In a column chart, the columns of the chart are separate.

A pie chart can also be drawn for a bar or column chart.

A bar graph is used to graphically display quantitative and qualitative discrete variables.

Example 1-5: The bar graph of the academic fields of 12 students is as follows:

X_i = field of study	Agronomy	Animal breeding	Horticulture
f_i = number of students	4	3	5



1.8.2. Rectangular chart or histogram: A histogram chart is like a bar chart, with the difference that in a histogram chart, the columns are stuck together and it is used for continuous variables. In the histogram chart, connecting the columns has made this chart a suitable tool for displaying data resulting from the implementation of quantitative continuous variables (such as records of two speeds), that is, variables that are measured using an interval and relative scale.

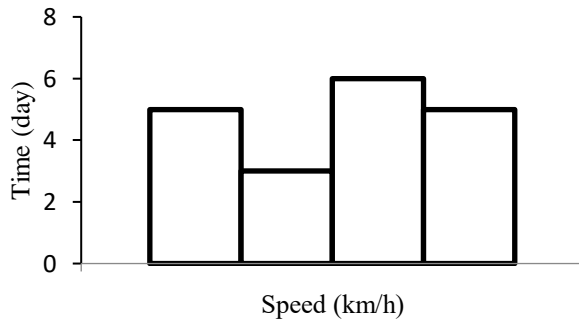
When data are measured using an interval or relative scale. The best chart to display data is a histogram chart.

To draw this graph, we specify the boundaries of the categories with equal lengths on the x-axis and the absolute or relative frequency of the category on the y-axis. In this diagram, the length of the groups is always equal and it is suitable for quantitative continuous variables.

A histogram chart is used to graphically display quantitative continuous variables.

Example 1-6: The following Table shows the wind speed for 19 consecutive days. Draw a histogram plot for the wind speed.

Speed (km/h)	1-3	3-5	5-7	7-9
Time (day)	5	3	6	5



If the area of a rectangle is given from the diagram, the relative frequency is obtained from the following relationship.

$$f_i = (\text{Rectangle of an area}) / (\text{Rectangle of the total area})$$

The concept of density is also used to draw a histogram.

The absolute density is obtained from the absolute abundance part per unit length of the bunch distance. That's mean:

$$di = \frac{F_i}{\Delta X_i}$$

The relative density is obtained from the outside part of the relative abundance per unit length of the bunch distance. that's mean:

$$di' = \frac{f_i}{\Delta X_i}$$

Therefore, in drawing a histogram, the variable attribute is on the x-axis and the absolute or relative density value is on the y-axis.

1.8.3. Multiple frequency chart: First, this chart is more suitable for quantitative continuous variables than rectangular.

A polygonal (linear) graph or frequency multiplier is in the form of broken lines and is used more than other graphs due to its ease of description and drawing.

The bottom surface of the polygon chart is equal to the bottom surface of the histogram chart.

The polygon diagram is mostly used to show the relationship between two continuous variables and relative frequencies.

The main reason for using a polygon chart is that it shows how well the records are distributed.

A polygon chart is used to compare multiple groups of data.

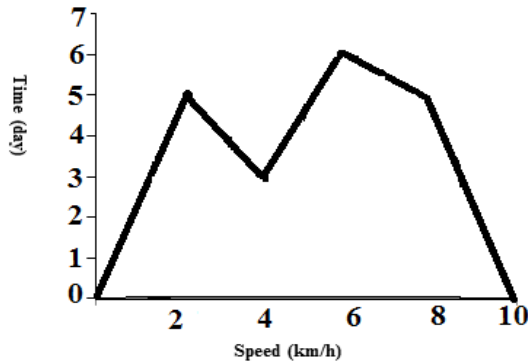
The shape of the polygon graph can be normal, positively skewed (the curve is tilted to the right), or negatively skewed (the curve is tilted to the left).

To draw this graph, we specify the center of categories on the x-axis and the absolute or relative frequency on the y-axis. So, if we connect these points (f_i, X) , the graph of multiple frequencies is obtained.

In the polynomial graph, we add two clusters with zero frequency to the beginning and end of the cluster so that the two ends of the graph are connected to the X-axis. This new diagram is called a full frequency diagram. By connecting this graph to the X-axis, the area under the

graph is multiplied by the total area of the rectangle in the rectangle graph.

Example 1-7: The graph of the frequency multiplier for the table of example 4 is drawn as follows.



According to the diagram above, each of the vertices of the frequency diagram is studied at the middle point of the upper side of a histogram rectangle related to the frequency Table.

1.8.4. Cumulative frequency chart: Cumulative frequency chart gives us information about the order of numbers in the table. In other words, this chart specifies the status and position of a grade compared to other grades or classes. For example, determining whether a grade is higher or lower than the percentage of grades.

The difference between this graph and the polygon graph is that the polygon graph is drawn based on simple frequency, but this graph is drawn based on density frequency.

In order to specify a certain percentage of different wheat cultivars that perform better than other wheat cultivars (determining a marker or position), a density frequency diagram is used.

Cumulative frequency is shown on the y-axis, but the upper limit of the categories is shown on the x-axis. We connect the obtained points to obtain a cumulative frequency diagram. This chart starts from the x-axis always ascending. If we have a relative cumulative frequency on the y-axis, the maximum height becomes one.

The cumulative frequency chart is less than always ascending and is used more in cases where the variables are continuous. Drawing the graph of cumulative frequency greater than is the same as the graph of cumulative frequency less than, with the difference that here on the y-axis we place values of cumulative frequency greater than. This graph is always downward.

To determine the status of a grade in relation to other grades, a density frequency chart is useful.

1.8.5. Pie chart: A pie chart is the simplest and at the same time the most suitable method for comparing and displaying data collected from discrete (such as gender) and continuous percentage variables (percentages of a whole). This diagram shows the frequency of variables and different ratios of a whole. To draw this graph, first, the frequencies are calculated as percentages or degrees, and then we draw a circle according to the frequency of each category. It is used to show the ratio of one series of observations among others.

If we want to find out the share of each class in terms of degrees from the circle. First, we draw a circle. The circumference of the circle

is 100%. To determine different levels of frequency, we divide one hundred percent of the circumference of the circle, i.e. 360 degrees, by the sum of numbers (N), multiply by the frequency of each data (f_i). Therefore, we have:

$$S_i = f_i \times 360 = F_i/N \times 360$$

F_i = relative frequency

N = number of observations

If we want to find out the share of each floor in terms of the percentage of the circle, we divide the frequency of each data by the total number of data and multiply by 100, so we have:

$$S_i = f_i \times 100 = F_i/N \times 100$$

If the percentage is given and we want to draw a circle, we convert the given percentages into hundredths or tenths and multiply by 360, as a result, the degree of the given percentage is obtained. For example, 40% is equal to an angle of 144 degrees ($360 \times 0.40 = 144$).

A pie chart is used to graphically display the percentage discrete variable.

Example 1-8: Consider the following example:

Variable	Diploma	Associate Degree	Bachelor's degree	Master's degree
Number of employees	20	25	50	5

$$S_1 = \frac{F_i}{N} \times 360^0 = \frac{20}{100} \times 360^0 = 72^0$$

$$S_2 = \frac{F_i}{N} \times 360^0 = \frac{25}{100} \times 360^0 = 90^0$$

$$S_3 = \frac{F_i}{N} \times 360^\circ = \frac{50}{100} \times 360^\circ = 180^\circ$$

$$S_4 = \frac{F_i}{N} \times 360^\circ = \frac{5}{100} \times 360^\circ = 18^\circ$$



■ Diploma ■ Associate Degree ■ Bachelors Degree ■ Masters Degree

Example 1-9: In an institution, 20 employees have a bachelor's degree, 50 have a diploma, and 20 have a middle school degree. How many degrees is the angle of the bachelor's sector to draw a diametrical diagram?

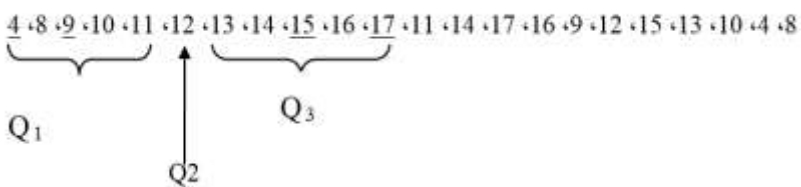
$$S_i = \frac{F_i}{N} \times 360^\circ = \frac{20}{90} \times 360^\circ = 80^\circ$$

1.8.6. Stem and leaf diagram: This diagram is used for quantitative variables where the difference between the smallest and largest data in terms of the number of digits is small. To prepare the stem and leaf diagram, we divide the statistical data figures into two parts. The stem contains one or more initial digits and the leaf contains the remaining digits. Unlike the texture chart, the original numbers are not lost in this chart. Repeated digits are written as many as there are. The advantage of this chart is that it includes all the data.

Example 1-10: If we have the numbers 21, 22, 23, 27, 29, 30, 35, 35, 38, and 39, then we will have

Stem	Leaf				
2	1	2	3	7	9
3	0	5	5	8	9

1.8.7. Box diagram: This diagram shows the quartiles and the minimum and maximum observations so that the box contains the difference between the first and third quartiles. In this chart, we first sort the data and call it Q_2 . The data is divided into two parts, and the median dimension is specified, and in the next step, we call the left part of the middle the first quartile (Q_1), and the right part of the middle the third quartile (Q_3). The line that divides the box into two parts is the midpoint of the observations. From each side of the box, a line continues as the minimum and maximum observations. A box plot shows the dispersion of data better than other plots.

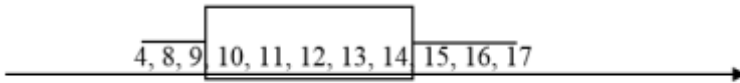


First from small to large:

Now:

Q_1, Q_2, Q_3

Then we close the box.



First, 9 and 15 are not inside the box. Second, it is a box plot that only shows the data inside the box and the 5 values (lowest, highest, 1st, 2nd, and 3rd quartile). If we want to draw an accurate graph, the units should be on the number axis with equal distances. Not like the diagram above, where the distance between 4 and 8 is not different from the distance between 8 and 9. Because the median is closer to the right side of the box, it means that the dispersion is more on the left side of the box. That is, the data on the left side of the box is the farthest. So this chart is best for comparing two sets of data.

1.8.8. Polygon diagram (abundance of multiple angles)

This chart is used when the data has an interval scale and continuous quantities. First, we draw the coordinate axes, and then we divide the horizontal axis (XX') according to the middle limit of the classes (X_i) and the numerical axis (yy') according to the highest absolute frequency (F_i) so that the length of the x and y axes is proportional to 3 Draw 2 or 4 by 3. If the frequency distribution table is related to two groups and therefore the classes have two absolute frequency columns (F_1 and F_2) and we want to plot the scores of the two groups on a coordinate system and then compare them, the best way is to convert the absolute frequencies of each group. It is in percentage or ratio and drawing them (to convert any absolute frequency into a percentage frequency, we use the relation $\%P = \frac{F_i}{N} \times 100$

A polygon chart is used to compare several groups of continuous data.

1.9. Shapes of a polygonal diagram

A frequency distribution diagram may have different shapes, it may be symmetrical or asymmetrical. A polygonal diagram is symmetrical when half of it is the same as the other half, otherwise, it is called asymmetrical. If in an asymmetric curve, the origin point (the point with the highest frequency) is not in the center of the curve and the curves are unequal, the curve is called skewed. In a curve with skewness, if the right sequence of the curve is longer than its left sequence, the curve is called right-skewed (positive skewness), in this case, more scores are at the end of the right sequence, and if the left sequence of the curve is longer than its right sequence, the curve will be skewed to the left (negative skewness), in which case the longer sequence of the curve ends towards very low scores.

1.10. Exponential in a polygon diagram

If the distribution has only one peak or point or maximum height, it is called monoexponential (Figure 1-1), if the distribution has two peaks or two points with maximum height, it is called biexponential, and in the same way, if the distribution curve has more than two peaks or two points with If it is the maximum height, it is called multidimensional.

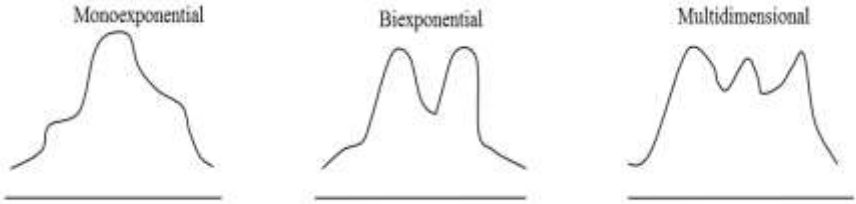


Figure 1-1: Exponential in a polygon diagram

1.11. Many frequencies

If we collect a series of values along with their number in a table, a frequency distribution table is obtained. In the frequency distribution table, the number of observations can be shown in different ways:

1.11.1. The absolute frequency of observations (F_i): It is the number of a specific observation among all observations or the number of times each data is repeated is called the absolute frequency of that data. The sum of the absolute frequency of all observations is equal to the total number of observations (statistical population).

In mathematical language:

$$\sum_{i=1}^n F_i = F_1 + F_2 + \dots + F_n = N$$

Example 1-11: Find the absolute frequency of the following data.

14, 17, 14, 14, 14, 14, 17, 20, 23, 20, 26, 20, 26, 26, 26

Variable	14	17	20	23	26
Absolute frequency	5	2	3	1	4

1.11.2. Relative frequency (f_i): If we divide the absolute frequency by the number of data, the relative frequency is obtained, which is always between zero and one. So that the sum of the relative

frequency of all the data is equal to one. This is best for comparing data in a table with different frequencies. If we multiply by 100, it shows the percentage of relative frequency. The sum of the relative abundance percentages is always one (i.e. 100%).

$$\frac{f_i}{N} \times 100$$

In mathematical language:

$$\sum_{i=1}^n f_i = f_1 + f_2 + \dots + f_n = 1$$

Example 1-12: In the previous example, find the relative frequency of the number 20.

$$\text{Percentage of relative frequency} = \frac{f_i}{N} \times 100$$

Percentage of the relative frequency of the number 20 = $\frac{3}{15} \times 100 = 20\%$

Example 1-13: Height measurements of 120 students are categorized in the table below. What is the frequency of the fourth category?

Category center	155	158	161	164	167	170
Percentage of relative abundance	10	15	18	x	20	12

If we add the relative frequency of all the categories together, their sum will be 100. So:

$$12 + 20 + x + 18 + 15 + 10 = 100$$

$$x = 25$$

$$\text{Percentage of relative abundance} = \frac{f_i}{N} \times 100$$

$$30 = \frac{f_i}{120} \times 100 \quad f_i = 25$$

Percentage of relative frequency (P_i): Whenever we multiply the relative frequency of each data (category) by 100. Its relative abundance percentage is obtained. That's mean:

$$P_i = f_i \times 100$$

The total percentage of relative frequencies is equal to 100. That's mean:

$$\sum_{i=1}^n P_i = P_1 + P_2 + \dots + P_n = 100$$

1.11.3. Cumulative frequency (F_{Ci}): The sum of the frequency of each category with the frequency of previous categories is called cumulative frequency. In fact, this abundance is very expected and includes all the abundances of the categories before it. So the cumulative frequency for the first category is the same as the absolute frequency of the first category because before the first category we have nothing to add to it, but the cumulative frequency of the last category is equal to the total data because they are added together.

Cumulative frequency should always be obtained by cumulative frequency of absolute frequency.

$$F_{Ci} = f_{Ci} - f_{Ci-1}$$

That is, for example, the absolute frequency of the fifth category becomes the cumulative frequency of the fifth category minus the cumulative frequency of the fourth category:

$$F_{C5} = f_{C5} - f_{C4}$$

The cumulative frequency of the first category is always equal to the absolute frequency of the first category: that is:

$$F_{C1} = F_1$$

The cumulative frequency of the last category is always equal to the size of the population. that's mean:

$$F_{C_N} = N$$

The relative cumulative frequency is obtained from the following relationship.

$$F_{C_i} = \frac{F_{C_i}}{N}$$

Cumulative frequencies are used to determine the number of observations that are at least or at most equal to a certain value.

Example 1-14: Find the cumulative frequency of the following table:

C - L	F_i	f_i	P_i	FC_i
0 -10	1	$\frac{1}{10} = 0.1$	$0.1 \times 100 = 10$	1
10 - 20	2	$\frac{2}{10} = 0.2$	$0.2 \times 100 = 20$	$1+2 = 3$
20 - 30	4	$\frac{4}{10} = 0.4$	$0.4 \times 100 = 40$	$1+2+4 = 7$
30 - 40	3	$\frac{3}{10} = 0.3$	$0.3 \times 100 = 30$	$1+2+4+3=10$
Total	10	1	100	

Example 1-15: In a frequency distribution table, 10 categories and $N=100$, the total relative frequency before the tenth category is equal to 0.96, find the absolute frequency of the last category.

$$\sum_{i=1}^n f_i = 1 \quad F_{10} = 0 - 1.96 = 0.04$$

$$f_{10} = \frac{F_{10}}{N} = \frac{F_{10}}{10} = \frac{4}{100} \quad F_{10} = 4$$

As we know, the cumulative frequency is equal to the sum of the absolute frequency of each category with the absolute frequency of its previous categories.

We have two types of density frequency which are:

1.11.3.1. Ascending density density

The ascending cumulative frequency of each observation is its absolute frequency + the absolute frequency of all categories (observations) before it, this type of frequency shows the number of observations that are at most equal to a certain value. Ascending density frequency shows how many percent of the data is less than and equal to a certain number or what percent of the data is at most equal to a certain number.

Example 1-16: What does the number 80 represent in the density frequency below?

About categories	16-19	13-16	10-13	7-10	4-7
Ascending density frequency percent		80	65		

The number 80 indicates that 80% of the data is smaller than 16.

1.11.3.2. Descending density frequency

The cumulative frequency of descent of each observation is its absolute frequency + the absolute frequency of all categories after it, this type of frequency shows the number of observations that are at least equal to a certain value.

Sometimes the relative abundance is expressed in the form of density, in which case it is called density relative abundance or relative density. The ascending density frequency of the last category is equal to the descending density frequency of the first category, which is equal

to the total number of observations. Also, the ascending density frequency of the first category is equal to its absolute frequency.

Example 1-17: What does the number 73 represent in the table below?

About categories	15-18	12-15	9-12	6-9	3-6
Descending density percentage		45	73		

The number 73 indicates that 73% of the data is greater than 9.

Exercises of chapter 1

1- In 56 statistical data, the largest and smallest are 86 and 65, respectively. These data are categorized into 7 categories. If the data in one category are considered the same, what is their common value in the fifth category?

2- 75 statistical data are categorized into 7 categories. The smallest data is 27 and the largest is 47.8. We know that 28% of the data is less than 36 and 40% of the data is less than 39. What is the absolute frequency of the middle class?

3- Cumulative frequency of 70 is equal to 40 percent, what does it mean?

4- In a frequency distribution table, the number of observations is 30, if the absolute frequency of the last class is 3, what is the cumulative frequency percentage of the penultimate?

5- In a collection, the minimum and maximum data are 321 and 520, respectively, and 10 floors (classes) have been selected. What is the distance between floors?

6- If $N=5$ and $X=\sum 30$, What is the result of the following expression?

$$=\sum_{i=1}^5 (4x_i + 6)$$

7- In case we have, the result of the following expression will be equal to:

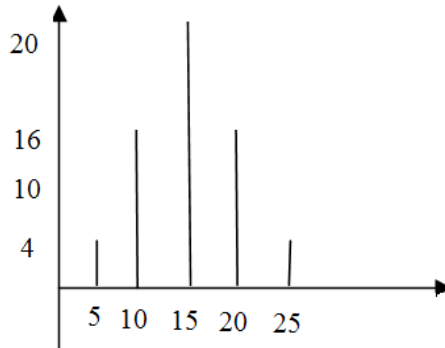
$$\begin{aligned} X_1 = 4 \quad X_2 = 5 \quad X_3 = 2 \quad X_4 = -5 \quad X_5 = -3 \\ = \sum_{i=3}^5 a^2(x_i - 4) \end{aligned}$$

8- If we have, the result of the following expression will be equal to:

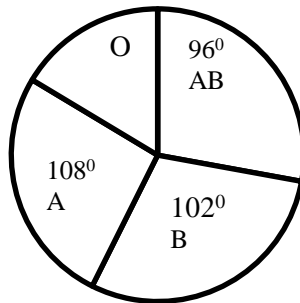
$$\sum_{i=1}^6 x_i = -4 \quad \sum_{i=1}^6 x_i^2 = 10$$

$$= \sum_{i=1}^6 x_i (x_i - 1)$$

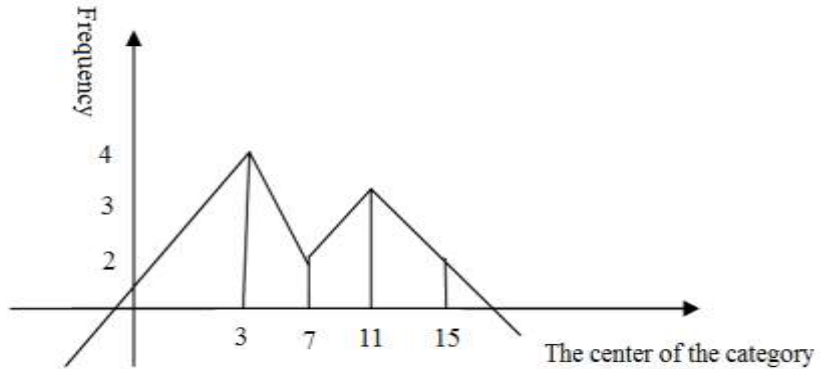
9- In the facing bar chart, what is the percentage of the relative frequency of the second category (center 10)?



10- The diagram of blood donation of people referring to a blood transfusion station is shown in the opposite figure. What percentage of these people are in blood type O?



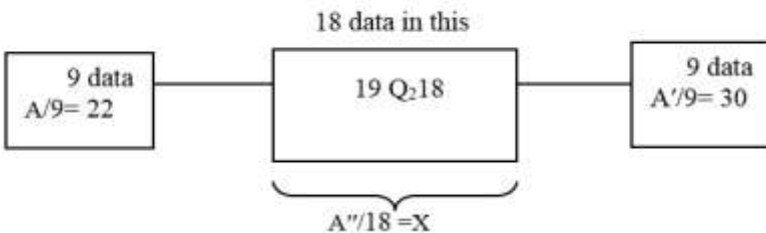
11- What is the area under the multi-faceted graph?



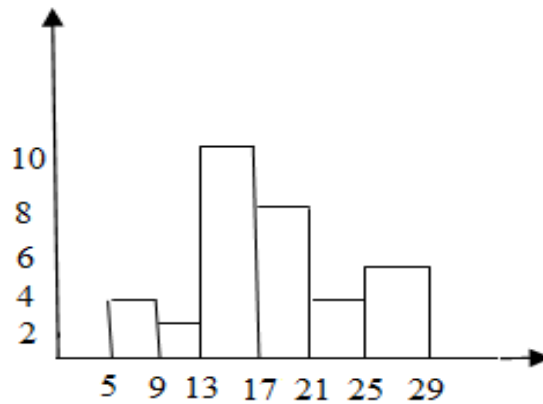
12- Statistical data are given with one decimal place with a stem and leaf diagram, what is their average?

Stem	Leaf							
8	0	0	1	2	2	5	6	7
9	1	1	2	3	3	4	5	5
10	1	1	2	2				

13- In the box diagram of 36 statistical data, the average data on both sides of the box are 22 and 30, respectively. If the mean of all the data is 27.5, then what is the mean of the data inside the box?



14- According to the diagram opposite, the relative frequency of the category with category 15 is equal to:



The 15-category symbol is placed on a category whose center is 15?

CHAPTER 2

CENTRAL TENDENCY AND DISPERSION

INDICATORS

Dr. Saeid HEYDARZADEH¹

¹ - Former Ph.D. Student of Urmia University, Faculty of Agriculture, Department of Plant Production and Genetics, Urmia, Iran
ORCID ID: 0000-0001-6051-7587, e-mail: s.heydarzadeh@urmia.ac.ir

INTRODUCTION

In many researches, it is necessary to provide a quantity or an index as a representative or general representative of the data and observations, and use it in reviewing the research results. Statistical indices are divided into two categories: center tendency (average value) and dispersion index. The most important indicators of tendency to the center are: median, view, parks (trenches) and various averages.

2.1. Mode

In a string of numbers, it is the observation that has the highest frequency. Mode or face is not a valid statistical indicator, because it is calculated only according to the frequency of observations and the quantity of data is not involved in its value. In a series of statistical data, the mode is the number that is repeated the most. The simplest indicator is central tendency. A statistical population may have more than one mode or no mode at all.

Among the mean, median, and mean, the mean is the fastest estimate, which does not have much value, but only expresses the number or numbers with the most frequency.

In data with a nominal scale, the mode or mode is the only indicator of central tendency that is used.

Mode is an unstable index, because its value changes by adding or subtracting a score. We should know that when the mean and median are determined, the face can be calculated using the following formula.

$$\text{Mode} = 3\text{median} - 2\text{Mean}$$

Calculation of mode was not classified in the frequency table: fashion is in the category that has the highest frequency.

It should be known that the data may have one mode (one mode), have two mode (two mode) or be duplicated to the same extent, that is, have no mode, which is called multi-view data in some sources.

The prominent feature of mode is that a series of data may not have mode or have more than one mode. This is despite the fact that there is always a mean and a median in a data series (Figure 2-1).

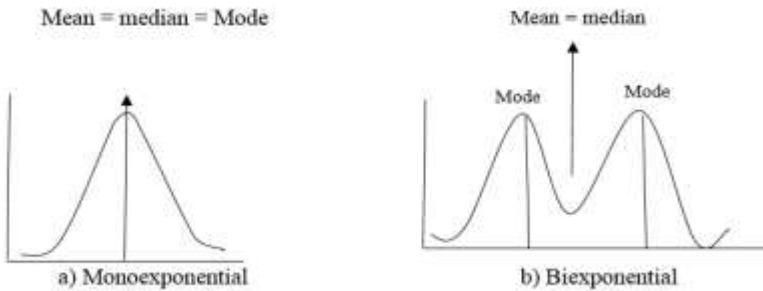


Figure 2-1: Mean, median and mean in two symmetrical distributions

Example 2-1: In the string of numbers below, the number 2 is mode.

1,1,2,2,2,2,2,3,3,4,4,5,5

Calculating the mode in the classified frequency table: There are two methods to calculate the face in the classified numbers, both of which end in the same answer, one is to select the class that has the highest frequency and use the middle point of that class as the mode.

In the second method, to calculate the expression in a frequency distribution table with category limits, the following method should be used:

1- If the range of the category is small, the average score of the class that has the highest frequency is considered as the mode.

2- In the case that the category is larger, the following formula should be used to calculate the mode.

$$\text{Mod} = L + \left(\frac{d_1}{d_1 + d_2} \right) j$$

The components of this formula are:

L: the lower limit of the category (the category is the category that has the highest absolute frequency).

d_1 : The difference in the frequency of the category compared to our previous category

d_2 : The difference in the frequency of the next category from our next category

j: limits of categories

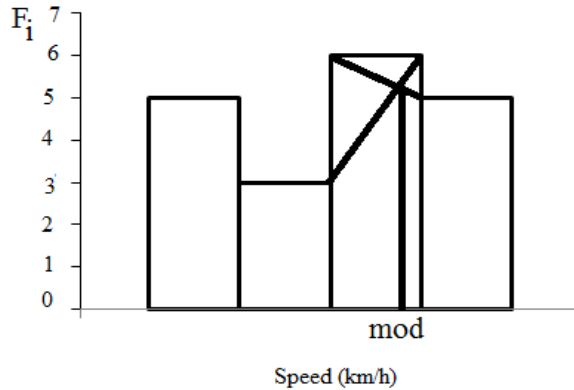
If there are several classes in a table that have the highest frequency, the mode must be calculated for each of them, that is, the table is multi- mode.

Example 2-2: Calculate the mode in the following example?

Category	f_i
2 - 4	2
6 - 8	3
10 -12	9
14 -16	6

$$\text{Mode} = 10 + \left(\frac{6}{6+3}\right) 2 \approx 11$$

To determine the mode in the histogram chart, we draw diagonal lines in the floor that has the highest frequency (height), like the figure below, the intersection of these two diagonal lines shows the mode.



2.2. Median

In the ordered series of observations, the median is the observation that divides this series into two equal parts in terms of

number, so that half of the numbers are less and the other half are more than it. If the number of observations is odd, the median is the median data, and when the number of observations is even, the median is obtained from the average of the two middle data.

Median in unsorted numbers: To find the median in raw numbers, we must first sort the data, that is, write them from smallest to largest. Then we find a number that has exactly half of the marks above it and half of the marks below it. In this case, two points should be noted:

A) If the data are odd, the middle number is the median.

b) If the data are even, the sum of the two middle numbers divided by 2 is the median (the average of the two middle numbers is the median).

It should be known that when the number that divides the distribution into two parts is repeated, then there are two cases.

1- If the number of numbers on both sides of the repeated number (middle number) is equal. In this case, the median is calculated in the same old way, that is, if the number of numbers is odd, the middle number is considered the median, and if the number of numbers is even, the sum of the two middle numbers divided by 2 is the median.

2- If the number of digits on both sides of the repeated number (middle number) is not equal. In this case, the method of real limits of repeated number or decomposition of repeated number is used. In this method, we do the following steps to find the middle:

a) We determine the true limits of the repeated number.

b) The unit of integers is 1. Therefore, we divide the number one by the number of repetitions of the repeated number. We add everything

obtained to the lower limit of the repeated number and continue until we reach the upper limit of the last repeated number. In this case, a series of new numbers is obtained, whose middle number is no longer repeated. Therefore, in these new numbers, the number in the middle or the average of the two middle numbers is considered the middle number.

The formula for calculating the median in repeated numbers

If the median of the score distribution is placed between repeated numbers, then the median is calculated according to the following formula.

$$\text{median} = \frac{N+1}{2}$$

The formula for calculating the median for a special case can be considered as follows.

$$M_n = \text{Repeat down the bank} + \frac{\frac{N}{2} \text{ The number of repetitions below}}{\text{The number of repetitions}}$$

Example 2-3: Determine the median in the following data?

10,8,8,7,7,2,2,3,1,8,7,4,5

First, we sort the data from small to large, then we specify the median:

10,8,8,8,7,7,7,5,4,3,2,2,1

Due to the fact that the data are odd, we call the number 7 as the middle data as the median.

Example 2-4: It is desirable to calculate the median of the numbers "5, 5, 5, 4, 3"

$$\text{median} = \frac{N+1}{2} = \frac{6+1}{2} = 3.5$$

Therefore, the median is placed in the 3.5th order, that is, the middle of the first and second 5, that is, between the repetition numbers.

Median in classified numbers: It should be known that if the data are classified (in the frequency distribution table), we do the following steps to obtain the median:

1- We divide the total number of numbers by 2 ($\frac{N}{2}$).

2- We write the density column. In the cumulative frequency column, we select the class that has a cumulative frequency equal to $\frac{N}{2}$ or the closest number greater than $\frac{N}{2}$. This is the median class.

3- We note the lower limit of the desired floor and obtain the median using the following formula:

$$\text{Me: } L + \left(\frac{\frac{N}{2} - F_c}{f_i} \right) j$$

The components of the formula are:

L: the lower limit of the median category (the median category is the first category whose density frequency is greater than half of the data).

F_c: density frequency of our group before the median group

f_i: absolute frequency of the median category

j: limits of categories or classes

In a table, the frequency of one or more classes may be equal to zero. If the midpoint is in the range of such a class, the average score is the median class.

Example 2-5: Find the median below the frequency distribution table?

Category	f_i	F_c
2-4	2	2
4-6	4	6
6-8	11	17
8-10	9	26
	$N=26$	$\frac{N}{2}=13$

$$\text{Me: } L + \left(\frac{\frac{N}{2} - F_c}{f_i} \right) j$$

$$\text{Me} = 6 + \left(\frac{13 - 6}{11} \right) 2 \approx 7$$

2.2.1. Medium features

- 1- There is only one median in each statistical population.
- 2- Unlike the average, the median is not affected by the large and very small numbers of the set of sizes.
- 3- When all variable values are not known, the median can still be calculated.
- 4- If the sizes of people located at the beginning or end of the distribution are significantly different from other sizes, it is better to use

the median as the average parameter, because the average will be far from the real average value.

5- Geometrically, the median length (meaning the first specific) is a line perpendicular to the x-axis that divides the histogram into two equal parts in terms of surface.

6- The most important feature of the median is that the sum of the absolute magnitudes of the differences of different values of the variable is the smallest of the median. that's mean:

$$\sum f_i |X_i - me| = \min \quad \text{or} \quad \sum F_i |X_i - me| = \min \quad \text{or} \quad \sum |X_i - me| = \min$$

So, the sum of the absolute value of the differences of the variables from the mean is smaller than the sum of the absolute value of the differences of the variables from any other number (such as a).

$$\sum |X_i - a| > \sum |X_i - me|$$

7- It is used with rank variables.

Due to not defining the median algebraically, it cannot be used in other calculations. as

For example, the median of different groups cannot be used to calculate the group that is obtained from the combination of these groups.

For example, suppose the average age of female and male students of a college is 22 and 25 years, respectively. There is no way

It is not possible to calculate the average age of all students of this faculty using the above averages to gain.

The average age of all students should be the age distribution of this group.

2.3. Quantiles or lots

Quantiles are values of observations that divide the range of changes into required quantile intervals, so that frequencies in each of these intervals constitute a certain percentage of the total frequency. In other words, an ordered string of numbers can be divided into $n+1$ equal parts with n indices. For example, we have divided a sorted string into 2 equal parts by means of an index (middle). Each of the n indices by which the string of numbers is divided is a quantile or a quotient, and the distance between two consecutive quantiles is a quantile range. Now, if we divide a string of numbers into 4 equal parts by 3 indices and each of them is 1 quarter, and if we divide it into 10 equal parts by 9 indices, each of them will have 1 decimal, and if by 99 indices Let's divide 100 parts and call each of them a percentage.

Obviously, if the observations are not classified, it is possible to determine the desired quantile after sorting them.

To calculate each of the types of quantiles in an abundance distribution table with category limits, the following should be done:

1- First, we calculate P_n .

The values of P_n for the following different quantiles are:

$$\text{first quarter} = \frac{1}{4} N \qquad \text{third quarter} = \frac{3}{4} N$$

$$\text{fourth decile} = \frac{4}{10} N \qquad \text{20th percentile} = \frac{20}{100} N$$

2- We determine the category in which the desired quantile is placed (this category is the first category whose density frequency is higher than P_n).

3- According to the following formula, we calculate the amount of the desired quantile, which we have in this formula:

$$\text{The amount of quantile} = L + \left(\frac{P_n - F_c}{f_i} \right) j$$

L: The lower limit of the category in which the desired quantile is located.

F_c: Density frequency of our category before the desired quantile category

f_i: absolute frequency of the desired quantile category

j: limits of categories

Example 2-6: Find the second quartile in the frequency table below?

Category	f _i	F _c
2-4	2	2
4-6	4	6
6-8	11	17
8-10	9	26
	N=26	$\frac{N}{2}=13$

$$P_n = \frac{2}{4} N = \frac{2}{4} \times 26 = 13$$

$$Q_2 = L + \left(\frac{P_n - F_c}{f_i} \right) j$$

$$Q_2 = 6 + \left(\frac{13 - 6}{11} \right) 2 \approx 7$$

Important notes and reminders about central indicators:

1- If there is a frequency table and it is with the center of the category, that is nothing, but if it is with the limits of the category, we use the center of the same category to calculate the average, etc.

2- If all the data are equal, all central indices will be equal.

3- If the data form a numerical progression, the mean will be equal to the median and the data will be the middle.

$$\bar{X} = \bar{X}$$

2.4. Mean

It should be known that the mean (\bar{X}) is the most useful measure that is used to describe the tendency to the center or the mean score distribution of a group of people or objects or events. To get the mean, we add all the scores together and divide by their number. The mean is the most important indicator of central tendency, which is the most widely used. There are different types of means, each of which has its own application. But in general, their most important types can be stated as follows:

2.4.1. Arithmetic mean

The arithmetic mean, which we will call the mean from now on, is the sum of the data divided by their number. If we denote the mean with the symbol μ (Greek letters mu) and the data with X and the total number of data or observations with \tilde{N} , the mean is calculated according to the following formula.

$$\bar{X} = \frac{\sum x_i}{N} \quad \text{or} \quad \mu = \frac{\sum x_i}{N}$$

If the distance between the numbers is the same, then the mean is equal to the sum of the largest and smallest number divided by 2.

Note that it is always expressed as the mean of a single experimental data series or measurement scale. When the data are set in the unclassified frequency table (that is, with $j \neq 1$), the following formula is used to calculate the mean.

$$\mu = \frac{\sum f_i X_c}{\sum f_i}$$

When the data are arranged in the classified frequency table, it should be known that there are three methods of calculating the mean in the classified numbers larger than 30 people.

The first method of calculating the mean in classified numbers is that I multiply each score by their frequency and then add them up and divide by the total numbers. This method is when there is only one number in each floor.

$$\mu = \frac{\sum f_i X_i}{N} \quad \text{or} \quad \mu = \frac{\sum f_i X_i}{N}$$

The second method is to calculate the mean in the classified numbers through the calculation of the middle scores of each class. In this way, we multiply the middle point of each class by the frequency of that class, then we divide the sum of the middle points by the total number of grades. This method is when there are two numbers in each floor.

To calculate the middle point, we add the smallest and largest number of each floor and divide by 2.

$$\text{mean} = \frac{(\sum f_{xc})}{(N)}$$

If the data in the table is not in order, the first thing to do is to sort it from bottom to top.

The third method of calculating the mean in classified numbers is done through the short method of calculating the hypothetical average. In this method, we perform the following steps.

A class is selected from among the classes in the table and the zero class is given in a column called d , and the upper classes are given $+1$ and $+2 \dots$ respectively, and the lower classes are given -1 and $-2 \dots$ respectively.

Usually, a class is chosen which is in the middle of the classes or a class is chosen which has the greatest frequency.

We multiply each class by the frequency of the same class and write it in the fd column.

We calculate the fd column and put it in the following formula:

$$\text{mean} = Xc + \frac{\sum fd}{N} \times i$$

Xc is the data or midpoint of the stratum in which the mean is assumed to lie.

Example 2-7: A person has bought a product at the price of 16, 18, 21 and 25 Rials per kilo during four consecutive years. What will be the average price of this product assuming that the purchase amount is 10 kg per year?

Note that this average does not depend on the purchase amount, so in this case, the arithmetic mean is used to calculate the average amount. so:

$$\mu = \frac{1}{4} (25+21+18+16) = 20$$

Mean features

1- There is only one mean in each statistical population.

2- The mean measurement unit is the same as the variable measurement unit.

3- The mean is the only parameter that if it is placed in place of all the data, their sum will not change. that's mean:

$$\sum X_i = n \cdot \bar{X}$$

4- The algebraic sum of the difference of the data from their mean is zero. that's mean:

$$\sum f_i (X_i - \bar{X}) = 0 \quad \text{or} \quad \sum F_i (X_i - \bar{X}) = 0 \quad \text{or} \quad \sum (X_i - \bar{X}) = 0$$

5- The sum of the squares of the deviation of the data from the mean is too small to use any other arbitrary number such as a instead of the mean. that's mean:

$$\sum f_i (X_i - \bar{X})^2 = \text{Min} \quad \text{or} \quad \sum F_i (X_i - \bar{X})^2 = \text{Min} \quad \text{or} \quad \sum (X_i - \bar{X})^2 = \text{Min}$$

That is, the sum of the squares of the deviation of the data from the mean is too small to use any other arbitrary number such as a instead of the mean. that's mean:

$$\sum (X_i - \bar{X})^2 < \sum (X_i - a)^2$$

6- If \bar{X}_1 is the mean of n_1 data, \bar{X}_2 is the mean of n_2 data, ..., \bar{X}_k is the mean of n_k data, the mean of all these numbers is equal to:

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_k \bar{X}_k}{n_1 + n_2 + \dots + n_k}$$

7- When x and y are two sets of numbers and the set z is obtained from the sum of the numbers of the sets x and y, the mean of the set z is equal to the average sum of the two sets x and y. that's mean:

$$z_i = x_i + y_i \qquad \bar{z} = \bar{x} + \bar{y}$$

8- The mean of a fixed number is a fixed number itself. that's mean:

$$a = \bar{a}$$

9- If we add a constant number like a to the variables, the amount of a will be added to the average. that's mean:

$$\bar{x} + \bar{a} = \bar{x} + a$$

10- If we subtract a fixed number like a from the variables, it will also be subtracted from the mean by the amount of a. that's mean:

$$\bar{x} - \bar{a} = \bar{x} - a$$

11- If we multiply the variables by a constant number like a, the average will be multiplied by the constant number a. that's mean:

$$\bar{ax} = a\bar{x}$$

12- If we divide the variables by a fixed number like a. The average will be divided by the constant number a. that's mean:

$$\left(\frac{\bar{x}}{a}\right) = \frac{\bar{x}}{a}$$

$$\bar{ax} \pm \bar{b} = a\bar{x} \pm b$$

It should be noted that the average depends on the numerical value of each observation. But this advantage is considered one of the weaknesses of the average in some situations. Consider a class in which one or two people have a very high score, and these one or two numbers affect the numerical value of the average.

Average is best used when the measurement scale is interval or relative.

The sum of the squared deviations of the scores from the mean is always smaller than or equal to the sum of the squared deviations of the scores from any other number.

The sum of the squared difference of each score from the mean is less than the sum of the squared difference of each score compared to the other point.

The algebraic sum of deviations from the mean is always equal to zero.

The sample mean is closer to the population mean than the sample mean is closer to the population mean.

2.4.2. Geometric mean

This index is rarely used in educational and psychological statistics and is used more in economic statistics, which is one of its applications in psychophysics. When we want to calculate the average of a series of observations in the form of percentage, ratio, index, growth rate, etc., it is better to use the geometric mean. The geometric mean of observation is obtained as follows. Whenever x_1, x_2, \dots, x_n are a set of n numbers, the geometric mean of these n numbers is equal to the n th root of the product of those numbers. that's mean:

$$M_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \dots}$$

The following relationships can be used to calculate the geometric mean of a series of observations in a frequency distribution table.

$$\text{Log } m_g = \frac{\sum f_i \log x_i}{\sum f_i}$$

$$\text{Log } m_g = \sum f_i \sqrt{f_1(x_1) \cdot f_2(x_2) \dots f_n(n)}$$

In a frequency distribution table with category limits, we will have:

$$= \sum f_i \sqrt{x_1^{f_1} \cdot x_2^{f_2} \dots x_n^{f_n}}$$

Example 2-8: During three consecutive semesters, a student increases his activity by 2, 3 and 4 times compared to his other classmates. What is the average increase in activity of this student compared to other students?

Whenever the data is expressed in terms of ratio or percentage, the geometric mean is used, therefore:

$$\text{Log } m_g = \frac{1}{N} (\log x_1 + \log x_2 + \dots + \log x_n)$$

$$\text{Log } m_g = \frac{1}{3} (\log 2 + \log 3 + \log 4)$$

$$\text{Log } m_g = \frac{1}{3} (1.38) = 0.46$$

$$m_g = \text{anti log } (0.46) = 2.88$$

If, instead of multiples, we have real values of the variable in different years, we obtain the average annual multiples from the following formula:

$$M_g = {}^{n-1}\sqrt{\frac{\text{Last year data}}{\text{First year data}}} \quad n = \text{number of years}$$

If we want to get the average annual growth percentage, we use the following formula.

$$(M_g - 1) \times 100$$

Whenever the data is weighted in this type of average, we use the weighted geometric mean.

$$Mg = \sqrt[n]{x_1^{w_1} \times x_2^{w_2} \times \dots \times x_n^{w_n}}$$

2.4.3. Harmonic average

When $X_1, X_2 \dots X_n$ are all non-zero and equal, or in other words, if we invert the values of X and calculate their average and then invert this average, the harmonic average or harmonic average is obtained. This average is used in cases where the intensity and value of the data are different. Also, if the data measurement scale is mixed (such as kilometers per hour, meters per second, trials per grain, people per hour, etc.), we use the harmonic mean to find their mean.

$$M_h = \frac{1}{\frac{\sum \frac{1}{x_i}}{N}} = \frac{N}{\sum 1/X_i}$$

Therefore, the main use of the mean coefficient is when we want to determine how much time is required to do a certain unit of work on mean. In general, for a group of positive numbers, there is the following relationship between the three mean.

$$m_h \leq m_g \leq \bar{x}$$

Only when all variable values are the same, we will have:

$$m_h = m_g = \bar{x}$$

If the data is weighted in this type of average, we use the weighted harmonic average:

$$\bar{x}_H = \frac{w_1 + w_2 + \dots + w_n}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \dots + \frac{w_n}{x_n}}$$

In a frequency distribution table including the category limits, the harmonic mean can be calculated as follows.

$$M_h = \frac{\sum f_i}{\sum \left(\frac{f_i}{x_i}\right)}$$

Example 2-9: It takes 200 days for an oil tanker to reach its destination with cargo and 150 days to return to the place of loading after unloading. How many days is the average round trip time of this tanker?

Note that if the intensity and value of the data are different, we use harmonic or harmonic average. so:

$$M_h = \frac{N}{\sum 1/x_i}$$

$$= \frac{2}{\frac{1}{150} + \frac{1}{200}} = \frac{2}{\frac{4+3}{600}} = \frac{1200}{7} = 43.171$$

Also, the general formula, which is known as the compound profit formula, is as follows:

$$A = P(1+r)^n$$

P = initial value

R = percentage increase per unit time

N = number of time units

A = final value

Example 2-10: The number of a type of bacteria in a laboratory culture has increased from one thousand to four thousand in three days. Calculate the average increase in the number of bacteria.

Since the increase in the number of bacteria from one thousand to four thousand in three days is equal to 300%, it may be mistaken that the average increase is therefore equal to 100%. Of course, if this answer is correct, the number of bacteria should reach 2,000 after one

day, 4,000 after 2 days, and 8,000 after three days, which is not the case in practice. Assume that the average increase per day is equal to r . So:

$$\text{number of bacteria after one day} = 1000 + 1000r = 1000(r + 1)$$

$$\begin{aligned} \text{number of bacteria after two days} &= 1000(r + 1)r + 1000(r + 1) \\ &= 1000(r + 1)^2 \end{aligned}$$

$$\begin{aligned} \text{number of bacteria after three days} &= 1000(r + 1)r^2 + 1000(r + 1)^2 \\ &= 1000(r + 1)^3 \end{aligned}$$

Therefore:

$$1000(1 + r)^3 = 4000 \quad \therefore (1 + r)^3 = 4$$

$$1 + r = \sqrt[3]{4} \quad \therefore r = 1.1-587 \quad \therefore r = 0.587 = 58.7 \%$$

2.4.4. Equalized mean

If we want to calculate the average of a series of observations, each of them has its own value, we can use the weighted average. As you can see, this method of calculating the average is similar to calculating the arithmetic mean of numbers in a frequency distribution table.

$$M_w = \frac{\sum x_i w_i}{\sum w_i}$$

If any of the 4 operations of multiplication, division, addition and subtraction are performed on the numbers obtained from the observations, these changes will appear in the resulting average.

Example 2-11: If the average language score of three classes whose number of students are 40, 35 and 25 respectively is 15, 14 and

16, the average of the total language score of these three classes is equal to:

$$\mu = \frac{(15 \times 40) + (14 \times 35) + (16 \times 25)}{15 + 14 + 16} = \frac{600 + 490 + 400}{45} = \frac{1490}{45} = 14.9$$

2.4.5. Moving mean

In some cases, data are collected in time series (weekly, monthly or yearly). For example, we can mention the temperature or the amount of rainfall and even the amount of plant growth in certain periods of time (usually at equal intervals). From a mathematical point of view, a time series is defined with different values of variable x (such as X₁, X₂, X₃ and... X_n) at times t₁, t₂, t₃ ...t_n. Therefore, x is a function of t, which is shown as x=f(t). Usually, in time series, an index is used to express the existing mobility, an example of which is the moving average.

According to the definition of moving averages with rank or order n for X₁, X₂, X₃ and... X_n is equal to:

$$, \frac{x_2 + x_3 + \dots + x_{n+1}}{n}, \frac{x_3 + x_4 + \dots + x_{n+2}}{n} \dots \frac{x_1 + x_2 + \dots + x_n}{n}$$

Example 2-12: The table below shows the amount of rainfall on the first day of September in Moghan city.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Rainfall amount	10	8	20	7	5	6	4	9	3

The 3-year moving mean are as follows.

$$, \frac{8+20+7}{3}, \frac{20+7+5}{3}, \frac{7+5+6}{3}, \frac{5+6+4}{3}, \frac{6+4+9}{3}, \frac{4+9+3}{3}, \frac{10+8+20}{3}$$

So that the results will be equal to 3.5, 3.6, 0.5, 0.6, 7.10, 7.11 and 7.12.

2.4.6. Compound average: There are two modes:

1- If the number of groups is equal, we divide the sum of averages by the number of groups (not the number of people):

$$(\bar{X}_t = \frac{\sum X}{N})$$

2- If the number of groups is unequal, we multiply the average of each group by the number of the same group and divide by the number of groups:

$$(\bar{X}_t = \frac{\sum X.n}{N})$$

2.5. Comparison of mean, median and mode

The basic difference between average, median and mode is that the average is based on the scale of the data and the frequency and quantity of the data are considered in its calculation. But the mean and mode are indicators or data among the observations that depend on the order and frequency of the data. Therefore, the average is not obtained from the combination of data and any increase or decrease in them will change its value. On the contrary, if increases or decreases do not change the order of the data, there will be no change in the median value and mode.

Average is the most suitable central index to describe interval and ratio variables. But it does not work for rank and nominal variables. To describe ordinal variables, median and mode can be used (median is preferred) and only mode can be used to describe nominal variables.

2.6. Empirical relationship between mean, median and mode

In symmetrical distributions, all three central indices (mean, median, and mean) have the same value and are equal to each other, that is:

$$\bar{x} = me = mo$$

The normal distribution (normal, symmetric) has a symmetrical shape and has only one point, which is located in the middle of the distribution.

In normal distribution, any of the mean, median and mode indicators can be used to check the depth of scores or the overall estimation of score distribution. Because they have a value.

In data that has a normal distribution, the following relationships are true:

$$mo = 3md - 2\bar{x} \quad me = \frac{mo - 2\bar{x}}{3} \quad \bar{x} = \frac{3me - mo}{2}$$

If the mean, median, and mode are equal in a class, the population is normal.

In distributions with positive skewness: When the graph has a long sequence on the right side, it is said to be positively skewed.

Such a distribution occurs when the test is difficult and difficult so that only a few subjects answer the questions correctly and most of them do not do well.

In this curve, the mean is greater than the median and the median is greater than the mean (the mean is the largest and the mean is the smallest) and the following relationship is true.

$$\bar{x} > me > mo$$

In a curve with a positive skew, the density of scores is on the side of smaller scores (most of the data have low values and are below the mean) and the length of the curve is on the side of larger scores. In other words, in positive skewness, the frequency of smaller numbers is higher than larger numbers.

In a curve with a positive skewness, the mean tends to the long tail and to the right, and the exponent is on the density side and on the negative side of the coordinate axis, and the median is located between the mean and the mode.

In distributions with negative skewness: When the graph has a long sequence on the left side, it is called negative skewness. In this curve, the peak is greater than the median and the median is greater than the mean, that is, the mean is the smallest and the following relationship is true.

$$\bar{x} < me < mo$$

In a curve with negative skewness, the density of the curve is on the side of larger scores (most of the data have high values and are larger than the mean) and the long tail of the curve is on the side of smaller

scores. In other words, in negative skewness, the frequency of larger numbers is higher than smaller numbers.

In a curve with a negative skew or skewed to the left, the mean is inclined towards the long tail of the curve and to the left, and the plot is on the side of density and on the positive side of the coordinate axis, and the median is located between the mean and mode.

In negative skewness, the highest value belongs to the face and the lowest value belongs to the average.

In negative skewness, the mean is smaller than the mean and median and tends towards smaller scores.

In negative skewness, most learners do well.

In positive skewness, the mean tends to move away from the face towards extreme values.

The position of the average, median, and mean indices relative to each other determines the distribution of the graph.

In the distributions that have skewness, the mean is placed on the long tail side of the distribution.

An empirical relationship between the three parameters mean, median and mode is established as follows.

$$mo - \bar{x} = 3(me - \bar{x})$$

2.7. Comparison of central tendency indicators:

In a symmetric distribution, the measures of central tendency are equal. If the distribution is skewed, then the scores will be more concentrated on one side of the distribution and the three central indicators will not be equal and therefore will be placed in three

completely different points in the distribution. In general, it can be said that in a curve with a positive skew, the mean has the highest value, the median has the lowest value, and the mode is located between these two (at a distance of $1/3$ from the mean and $2/3$ of the mean), and in distributions that have a negative skew, the mode has the highest and the average has the lowest value, and the median is between these two (at a distance of $1/3$ from the average and $2/3$ of the view). According to the position of these three indicators, the following relationships can be extracted about them.

$$mo = 3 me - 2\bar{X}$$

$$me = \frac{mo + 2\bar{X}}{3}$$

$$\bar{X} = \frac{3me - mo}{2}$$

However, the mean is the most reliable and stable central index, if a population of many samples is selected, the difference between the mean of different samples is less than the difference between the medians of the same samples, and in the same way, the difference between the medians of different samples is less than the difference between the mean of the same samples. Therefore, to estimate the population from the sample, the mean is the most reliable estimate of the population mean. In a frequency distribution, if we sum the scores with a fixed number, the same fixed number is added to the mean, median, and mode of the scores. This rule of combination with constant value C is also true for subtraction, multiplication and division of numbers. Table (2-1) shows the characteristics and uses of each of the central tendency indicators.

Table 2-1: Features and uses of central tendency indicators

Indicator	Use cases	Features
mean	1-When the research question is quantitative and the researcher is interested in interpreting the data at a distance. 2- When other statistical calculations such as dispersion measures or correlation coefficient are required. 3- When the data distribution is normal or close to normal. 4- When we want to know the center of gravity of an example. 5- When the most reliable representative of the measures of tendency to the center is required.	1- It is a distance index that is used with relative and distance scales. 2- The total deviation from the mean is always zero. 3- In its calculation, all numbers or scores are used, and very small or large scores have an effect on it. 4- It is the most stable central index. 5-More mathematical operations are used in its calculation.
median	1- When the distribution is on a relative or interval scale and has a significant skew, it means that there are very large or very small numbers in the distribution of scores. 2- When the distribution of grades is incomplete. 3- When we want to know, the measurement result is placed in the upper half or the lower half. 4- When the measurement scale is ordinal.	1- It is an ordinal index that can be calculated with ordinal, interval and relative scales. 2- It is not affected by the numerical value of all grades. 3- Fewer mathematical operations are used in its calculation. 4- Its stability is less than the average but more than the view. 5- Its value is determined by the number in the middle of the distribution.
mode	1- When the measurement scale is nominal. 2- When an immediate estimate of the dimensions of the tendency to the center is necessary. 3- When the researcher wants to find the most repeated number	1- It is a nominal index that can be calculated in nominal, distance and relative scales. 2- It is a number that has the highest frequency and is often in the middle of the distribution of scores. 3- In some distributions there is more than one. 4- Mathematical operations cannot be done with it. 5- It has very little stability.

If the data are on an interval or relative scale, the best indicator is the central tendency of the mean. But if there is a score on the border (very large or very small score) in the distribution (the distribution has a curve), then the median index is more suitable.

2.8. Abundance curves

Abundance curves have certain shapes, which are explained in brief.

2.8.1. Symmetrical or bell-shaped curves

They represent the fact that groups equidistant from maxima have equal frequencies. The most important example of these curves is the normal curve. This curve has a maximum and the abundance on both sides of this maximum tends to zero uniformly. This curve is one of the most important curves in statistics and other curves are compared with this curve.

2.8.2. Chuleh curves

Skewness is a term used for asymmetric distributions, the curve of these distributions has a maximum point and the frequency tends to zero on both sides of it, but the speed of the tendency to zero on both sides of the maximum curve is not uniform, so the curve is not symmetrical. The skew direction is narrower towards the trail. Skewness to the right is called positive skewness and skewness to the left is called negative skewness.

2.8.3. J-shaped and inverted J curves

In these curves, the maximum point is at one end of the curve and tends to zero from there on.

2.8.4. U-shaped curves

These curves have two maximum points at both ends and the frequency becomes minimum for the value of the variable that is in the middle of the curve.

2.8.5. Biexponential curves

These curves have two maximum points.

2.8.6. Numerical indicators

Numerical indicators are parameters that are used for comparison between several societies and are divided into three parts:

- 1- Central indicators (central parameters)
- 2- Scattering indices (scattering parameters)
- 3- Relative dispersion indices (relative dispersion parameters)

2.8.7. Central parameters

Perhaps the most important issue in the study of any statistical population is determining the central value. It means determining the number of representatives around which the observations are distributed. Any numerical measure that represents the center of the data set is called central parameter. There are three central parameters namely mean, median and mode.

2.8.8. Characteristics of average value indicators

1- Among the indicators of central tendency, the importance of the average and the arithmetic average is more important in statistical studies. The basic difference between average, median and mode is that the average is based on the scale of the data and its calculation refers to the frequency and quantity of the data. But the median and the mean are

indicators or data among the observations that depend on the order and frequency of the data. Therefore, the average is obtained from the combination of data and any increase or decrease in them will change its value. On the contrary, if increases or decreases do not change the order of the data, there will be no change in the median value and the mode.

2- In a frequency distribution, the mean is the center of gravity of the observations, and the algebraic sum of the data differences is equal to zero. The distance of any statistical data from the average of those data is called the deviation of that data from the average, and in simpler terms in mathematics and statistics, it is called deviation. In a frequency distribution, the mean is the only index for which the sum of deviations is equal to zero, and no other index (for example, median, mode, geometric mean, etc.) has this characteristic (except in cases where the median or mode or both averages are the same). This can be proved as follows. If we denote the deviation of the data from the mean, i.e. $(X - \mu)$ with X , the sum of the deviations is as follows:

$$\begin{aligned}\sum X &= \sum (X - \mu) \\ &= \sum X - \sum \mu\end{aligned}$$

Because μ is a constant value, the sum of \tilde{N} times μ is equal to the product of μ times \tilde{N} .

$$\begin{aligned}&= \sum X - N\mu && \sum X \\ &= \sum X - N(\sum X/N) \\ &= \sum X - \sum X = 0\end{aligned}$$

3- In a frequency distribution, if the average and observations or other indicators are considered among the experimental data and the

deviation of the data is calculated with respect to each of them, the sum of the squared powers of the deviations from the average will have the lowest value. This is also one of the characteristics of the average that is used in statistics and many statistics laws are based on it. This principle is known as the law of least squares. In statistics, the sum of the squared deviations of the data from the mean is called the sum of squares, and it is denoted by SS.

$$\sum(X-\mu)^2 = \sum X^2 - N\left(\frac{\sum X}{N}\right)^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

4- There are simple relationships between mean, median, and mode, which are used to compare frequency distribution and convert one index to another.

In general, in the symmetrical graphs and tables, all three indicators are equal to the average value. In curves with left skewness, the relationship (mode \geq median \geq mean) and in curves with right skewness, the relationship (mean \geq median \geq mode) is established (Figure 2-2). For frequency distributions of a profile that are symmetrical, the relation (mean-median) 3 = (mode-median) is valid.

5- If $D_i = X_i - C$, so that C is a constant number, the average D, denoted by $D\mu$, is equal to:

$$\mu_D = \frac{\sum D_i}{N}$$

If we put its value instead of D:

$$= \frac{\sum(X_i - c)}{N} = \frac{\sum x_i}{N} - \frac{\sum c}{N}$$

$$= \mu_x - C \quad , \quad \mu_x = \mu_D + C$$

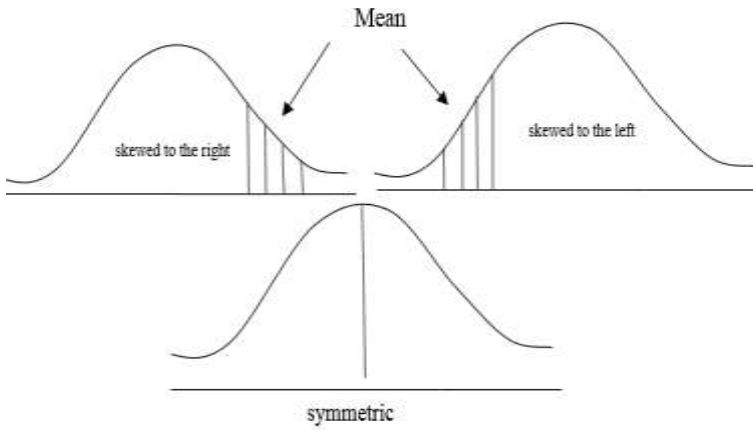


Figure 2-2: left, right and symmetric skewness index curves

This method is used to simplify the data (for example, converting multi-digit numbers to one or two digits), in order to facilitate the calculation of the average. So in a frequency distribution if we sum all the data with a fixed number (this number can be negative), the same fixed value is added to the mean of the data. Also, if the data of an experiment is multiplied or divided by a fixed number, the average of the data will increase and decrease in the same proportion.

In some cases, a new variable is obtained by combining the data of two variables such as x and y . If we denote the averages of two variables x and y by $X\mu$ and $Y\mu$ and $Z_i = X_i \pm Y_i$, the average of the new variable is equal to:

$$\mu_Z = \mu_X \pm \mu_Y$$

2.9. Scattering indices

Frequency distribution tables and average value indices express very important information about the research data, but they do not provide a complete picture of the nature of the data. Dispersion indices are criteria for determining the extent of data dispersion from each other or their dispersion relative to the average. In general, if the values of an attribute in the community differ more than the central index, it is a sign that the dispersion among the members of the community is greater. That is, there is no harmony and uniformity among the people of the society in terms of the studied trait, and if this difference between people is small, it is a sign that the dispersion among the people of the society is less in terms of the studied trait, and as a result, there is more harmony and uniformity among the people of the society. The most important dispersion indicators are: the overall range of changes, average quartile, average deviation and standard deviation, the latter index being the most important in terms of the principles and basics of statistics and application.

2.9.1. The total scope of data changes

This index is obtained from the difference of the largest and smallest observation and gives us an overview of the dispersion of observations. If the real concept of dispersion depends on the distance of the data from the indicators of the average value and their frequency.

$$R = X_{\text{Max}} - X_{\text{Min}}$$

The range of changes is the least important dispersion parameter, because it is only a function of the changes of two sizes and does not determine the status of the numbers that are in the middle, and the measure is unstable, because if its value changes dramatically with the change of the smallest number or the largest number. Using the range of changes requires having an interval scale, because the difference between the scores has a logical interpretation.

We use the range of variation when the scores have too much dispersion or we need general information about the dispersion. Using the range of changes requires having an interval and relative measurement scale. If the measurement scale is nominal or rank, the range of variation is not appropriate. Because the difference between the numbers in these scales has no logical meaning. The range of changes does not become negative and is always positive. If we want to have a quick estimate of the dispersion of the data, we use the range of changes.

2.9.2. The average quarter

It should be known that quartile in statistics is a concept that divides a distribution into four parts. For example, if we divide an interval of a series into 4 parts, each part is considered a quarter, that is, each quarter is equivalent to a quarter of a specific range. In this case, the first 25 cm point is called the first quarter (Q_1). The 50 cm point is called the median or the second quartile (Q_2). The 75 cm point is called the third quarter (Q_3). Based on percentage points, the first quartile is equivalent to the 25% point (25P), the second or middle quartile is

equivalent to the 50% point (50P), and the third quartile is equivalent to the 75% point (75P).

So that the size of the dispersion does not come from only two extreme values, it is possible to calculate the range of change between certain deciles or deciles.

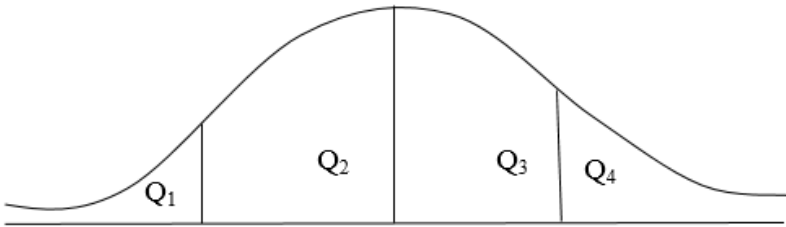
For example, we obtain the range of change between the 75% (third quartile = Q_3) and 25% (first quartile = Q_1) points and call it the range of quartiles. In general, the closer the observations are to the indicators of central tendency, the less dispersion they have. This index expresses the dispersion around the mean and its value is equal to half of the interquartile range. Each quartile contains a quarter or 25% of the abundance distribution data. The difference between the third quartile and the first quartile is called the range of the quartiles, which includes 50% of the middle sizes. that's mean:

$$Q_1 - Q_3 = \text{quartile interval}$$

$$50 \% \text{ point } Q_2 = \frac{Q_3 + Q_1}{2}$$

Finally, the quartile deviation in a distribution is defined as half the distance between 75% (third quartile) and 25% (first quartile). If we halve the range of the quartiles, the quartile deviation is obtained. The following formula is used to calculate it:

$$\text{Average quartile} = \frac{Q_3 - Q_1}{2}$$



Each quartile contains a quarter or 25% of the abundance distribution data. In a frequency distribution, the endpoints are in the first and fourth quartiles. Therefore, the interquartile range of 50% of normal data represents a frequency distribution. The median quartile of the specific dispersion is the median and is used and interpreted with it.

Example 2-13: Calculate the quartile deviation in the data 40, 38, 34, 28, 25, 21, 20, 15, 12?

Because the number of data is odd, the median (second quartile) is equal to the middle data, that is, $Q_2 = 25$, the median of the first half of the data is equal to the first quartile:

$$Q_1 = \frac{15+20}{2} = 17.5$$

The median of the second half of the data is equal to the third quartile:

$$Q_3 = \frac{34+38}{2} = 36$$

as a result:

$$\text{Average quartile} = \frac{Q_3 - Q_1}{2} = \frac{36 - 17.5}{2} = 9.25$$

Quartile median is used when the measurement scale is a minimum interval. If the mean of the quartiles is zero, then 50% of the measures in the middle are equal. As a result, the first, second and third quadrants are equal and vice versa.

To calculate the first and third quartiles, we use the median formula, only instead of $n/2$ we use $n/4$ and $3n/4$. that's mean:

$$Q_1 = L + \left(\frac{\frac{N}{4} - F_c}{f_i} \right) j$$

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F_c}{f_i} \right) j$$

When the graph of scores is skewed (positive or negative), the interquartile range or interquartile median is a good indicator of the dispersion of the scores.

The first quartile is called Q_1 , which is the 25% point, that is, Q_1 is the point where 25% of people are below this point and 75% of people are above it.

To calculate the quartile deviation in raw numbers, the first thing to do is to sort the numbers and get their median. In the next step, we get the middle point and the first number (the number of numbers is divided by 4, i.e. $\frac{N}{4}$) and then we calculate the middle point and the last number (the number of numbers is multiplied by 3 and divided by 4, that is, $\frac{3n}{4}$). We call this point Q_1 and Q_3 and then we put it in the quadratic deviation formula.

It should be known that if the quartile deviation and the median of a distribution are known, then the value of Q_1 and Q_3 as well as the middle 50% of the records can be calculated using the following formulas:

$$Q_3 = md + Q$$

$$Q_1 = md - Q$$

In some questions, if we are asked to calculate the mean deviation through the quartile deviation. In this case, we first get the quartile deviation and then multiply it by 1.5 to get the average deviation.

$$S = 1.5 \times Q$$

It should be known that there is no skew in the normal or natural (symmetric) curve and the skew is equal to zero. In fact, skew means the deviation of a curve from the point of symmetry. In normal (symmetric) distributions, the distance between the first quartile and the median is equal to the distance between the third quartile and the median. Therefore, if we consider Q_1 , Q_2 and Q_3 as the first quartile (25P), the second or middle quartile (50P) and the third quartile (75P), respectively, the following relationships are true about them. that's mean:

- 1- If $(Q_1 - Q_2) = (Q_2 - Q_3)$, then coefficient is zero.
- 2- If $(Q_1 - Q_2) < (Q_2 - Q_3)$, then the coefficient is positive.
- 3- If $(Q_1 - Q_2) > (Q_2 - Q_3)$, then the coefficient is negative.

In a positively skewed curve, the difference between the 75th percentile and the median is greater (larger) than the difference between the 25th percentile and the median.

In a curve with positive skewness, the density of scores from the first quartile to the median is higher than the density of scores from the third quartile to the median. Because most of the small scores are between the first quartile and the median.

In a curve with negative skewness, the difference between the 70th percentile and the median is (smaller) than the difference between the 25th percentile and the median.

In a curve with negative skewness, the concentration of scores from the first quartile to the median is less than the concentration of scores from the third quartile to the median. Because most of the big scores are between the third quartile and the median.

We also obtain the range of change between the 90% (ninth decile = D_9) and 10% (first decile = D_1) points and call it the range between the deciles.

$$D = D_9 - D_1$$

Therefore, the range of deciles is the range of changes in the middle 80% of the distribution.

Therefore, in a symmetrical distribution, $m_3 = 0$, and as a result, $SK = 0$, but in a distribution with a positive skewness, the sum of the cubes

The deviation of scores from the numerical mean is positive, and in the distribution with a negative skewness of the cube, the sum of the deviations of the scores from the mean will be a negative number. The skewness coefficient can be calculated using the following formulas known as Pearson's skewness formula.

$$G_1 = SK = \frac{\bar{X} - M_o}{s}$$

2.9.3. Average deviation: in a distribution, the frequency of each observation is either equal to the average, or higher or lower than it. As you know, the algebraic sum of these deviations is equal to zero. The more scattered the data is than the mean, the greater the deviations. This index shows the dispersion of observations around the mean. The average deviation is the average absolute value of the variables'

deviations from the average. We put the deviation of scores from the average in absolute value so that if they have a negative sign, their negative sign is neutralized. Therefore, the average deviation is never negative. In the raw data, we calculate the average deviation with the following formula:

$$AD = \frac{\sum |x_i - M|}{N}$$

The sum of the absolute value of the distance of the scores from the average of a statistical index is used to express the quality of the extent (dispersion) of the curve (diagram) of the scores.

It should be known that the sum of the deviations from the average of the total scores is equal to zero, that is, the sum of the deviations from the average of scores above the average and scores below the average, which are equal to each other, is equal to zero.

In the frequency distribution table, to calculate the average deviation, instead of calculating the difference of each score from the average, we calculate the difference of the midpoint of each point from the average and multiply it by the frequency of each class. Therefore, we will have:

$$AD = \sum f_i |X_i - M| \quad \text{or} \quad AD = \frac{\sum f_i |x_i - M|}{\sum f_i}$$

In a symmetrical distribution, about 80% of the observations fall within one standard deviation above and below the mean.

In calculating the average deviation, it should be noted that the absolute value of positive deviations from the mean is equal to the absolute value of negative deviations.

Example 2-14: According to the frequency distribution table below, what is the mean deviation or mean deviations of equal observations?

f_i	1	3	4	3	1
X_i	2	3	4	5	6

$$AD = \frac{\sum f_i |X_i - M|}{\sum f_i}$$

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{2+9+16+15+6}{12} = \frac{48}{12} = 4$$

$$AD = \frac{1(2-4)+3(3-4)+4(4-4)+3(5-4)+1(6-4)}{2} = 0.83$$

The reason for using AD as a measure of dispersion is that this parameter relies on all the data and observations, while the range of changes and the quartile mean rely on two or more specific measures. The major drawback of AD is to ignore algebraic signs.

When the statistical data form an arithmetic (numerical) expansion,

1- If the number of data is odd:

$$AD = \frac{(N^2-1)d}{4N}$$

2- If the number of data is even:

$$AD = \frac{Nd}{4}$$

In the above relations, d is the value of the arithmetic exponential ratio.

To use the mean deviation, the measurement scale must be at least interval. Because to interpret the results, it is necessary to calculate the distance of the scores from the average, and the condition that the average should also be calculated, it is necessary to have an interval scale.

2.9.3.1. Mean deviation characteristics

1- If all variables are equal, the average deviation is zero and vice versa. that's mean:

$$x_1 = x_2 = \dots = x_n \quad AD = 0$$

2- If we add a fixed number like a to the variables, their average deviation will not change. that's mean:

$$MD_{(x+a)} = MD_{(x)}$$

3- If we subtract a fixed number like a from the variables, their average deviation does not change. that's mean:

$$MD_{(x-a)} = MD_{(x)}$$

4- If we multiply the variables by a constant number such as a , the average deviation is multiplied by the absolute value of a . that's mean:

$$MD_{(ax)} = |a| MD_{(x)}$$

5- If we divide the variables by a fixed number like a , the average deviation is divided by the absolute value of a . that's mean:

$$MD\left(\frac{x}{a}\right) = \frac{MD(x)}{|a|}$$

6- The average deviation of constant number is zero.

7- It is not possible to perform algebraic operations with this index.

8- In average deviation, signs of numbers and in quartile deviation, all numbers are not studied.

9- It does not show the effect of large deviations in situations where a large number of small deviations are against a small number of large deviations (the most important shortcoming).

10- To use the average deviation, the measurement scale must be of the distance or relative type.

11- The advantage of the average deviation over the range of changes is that all scores (numbers) are involved in the calculation of the average deviation.

2.9.4. Standard deviation: Standard deviation is the most important and reliable statistical index to show the dispersion of observations compared to the average. In fact, we not only check the dispersion of a group of raw data and statistics by calculating the standard deviation, but also because most of the statistical indicators such as percentage, average, correlation coefficient, etc., are random variables, they also have a standard deviation, which can be used to interpreted these indicators and measured their statistical validity. Standard deviation is widely used in descriptive statistics and inferential statistics. In fact, the transition from description to statistical inference is done by generalizing and interpreting the concept of

standard deviation. On the other hand, compared to the rest of the dispersion indices, the standard deviation difference of a randomly selected sample from the population is less than the standard deviation of the population.

The biggest advantage and benefit of the standard deviation is the relationship between the standard deviation unit and the way the scores are placed in the natural curve, and it is because of this relationship that they use the standard deviation as a criterion for comparing different groups. The difference between standard deviation and average deviation is in how they are calculated.

If the standard deviation of a group of data or scores is zero, then all the data are equal.

The standard deviation, like the average deviation, is obtained based on the deviation of the observations from the average, and the average shows the dispersion of the observations around the average. In other words, the standard deviation is the positive root of the variance. that's mean:

$$\delta = \sqrt{\frac{\sum fi(x_i - \mu)^2}{N}} = \sqrt{\frac{\sum fi x_i^2}{N}}$$

The formula for calculating the standard deviation is as follows:

$$\delta = \sqrt{\frac{\sum x_i^2 - \frac{(\sum X)^2}{N}}{N}}$$

When the statistical measures form an arithmetic progression with the ratio d. The standard deviation is obtained from the following formula.

$$\delta_x = d \sqrt{\frac{N^2-1}{12}}$$

Example 2-15: The weight of 4 varieties of sorghum in grams is 50, 46, 42, 38. What is the standard deviation of their weights?

The obtained numbers form an arithmetic expansion with a ratio of 4, so we have:

$$\delta_x = d \sqrt{\frac{N^2-1}{12}} = 4 \sqrt{\frac{16-1}{12}} = 2\sqrt{5}$$

Standard deviation is the most stable index of dispersion. The greater the standard deviation, the greater the dispersion. Perhaps the most basic benefit of standard deviation is that it can be used to determine what proportion of grades are in different distances from the mean. When the distribution is highly skewed, the standard deviation should be used with caution. S is used when the mean is used as a central index. All dispersion indices (range of changes, average deviation, standard deviation) are used with minimum distance scale.

The stability of dispersion is in the order of standard deviation > variance > average deviation > average quartile > range of changes.

It should be known that the most important characteristic of dispersion (variability) is the standard deviation, that is, the standard deviation of the sign and represents the dispersion of the data.

The higher the standard deviation, the greater the variability of the numbers, and the lower the standard deviation, the lower the dispersion of the numbers.

The amount of standard deviation depends on the unit of measurement.

2.10. Characteristics of standard deviation

1- If all the variables are equal, the standard deviation is zero and vice versa. that's mean:

$$x_1 = x_2 = \dots = x_n \quad \delta_x = 0$$

2- If we add a fixed number like a to the variables, their standard deviation will not change. that's mean:

$$\delta_{(x+a)} = \delta_{(x)}$$

3- If we subtract a fixed number like a from the variables, their standard deviation will not change. that's mean:

$$\delta_{(x-a)} = \delta_{(x)}$$

4- If we multiply the variables by a constant number such as a, their standard deviation is multiplied by the absolute value of a. that's mean:

$$\delta_{(ax)} = |a| \delta_{(x)}$$

5- If we divide the variables by a fixed number such as a, their standard deviation is divided by the absolute value of a. that's mean:

$$\delta_{\left(\frac{x}{a}\right)} = \frac{\delta_{(x)}}{|a|}$$

6- The variance of a fixed number is zero.

7- The standard deviation, like the variance, is never negative, and this criterion is used when the measurement scale is at least a distance.

8- The standard deviation is not interpreted in terms of large or small, but the researcher interprets the collected scores with the help of this index. The advantage of the standard deviation over the variance is that the unit of measurement of the standard deviation is the same as the

unit of the data being measured, but the unit of measurement is raised to the power of 2 when calculating the variance.

Example 2-16: Calculate the standard deviation of the data in the following table?

x_i	0	1	2	3	4
f_i	3	2	12	6	1

$$\bar{x} = \frac{3(0)+(2)(1)+12(2)+6(3)+1(4)}{3+2+12+6+1} = \frac{48}{24} = 2$$

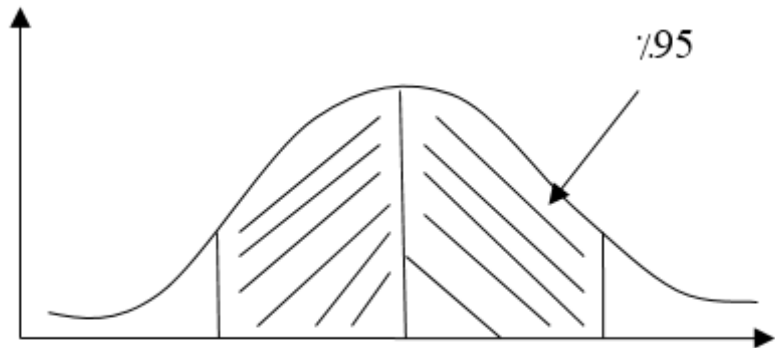
$$\delta = \sqrt{\frac{\sum f_i(x_i - \mu)^2}{f_i}} = \sqrt{\frac{3(0-2)^2 + 2(1-2)^2 + 12(2-2)^2 + 6(3-2)^2 + 1(4-2)^2}{3+2+12+6+1}}$$

$$= \sqrt{\frac{12+2+6+4}{24}} = \sqrt{\frac{24}{24}} = 1$$

Characteristics of standard deviation in a normal distribution:

$\frac{2}{3}$ of the data are within 1 standard deviation, more and less than the mean.

$\mu \pm \delta$	$\frac{2}{3} n$
$\mu \pm 1.64 \delta$	90% n
$\mu \pm 1.96 \delta$	95% n
$\mu \pm 2.58 \delta$	99% n
$\mu \pm 3 \delta$	99.7% n



In a symmetric distribution, the total range of variation is 6 times the standard deviation.

$$R = 6 \delta$$

The relationship between mean and standard deviation

The following points will be true in the normal curve and, in general, in the distributions of Chula between the mean (μ) and the standard deviation (δ).

1- About 68% of the examined cases are within the distance of $\mu \pm \delta$ (that is, one standard deviation away from the mean).

2- About 95% of the examined cases are within the distance of $\mu \pm 2\delta$ (that is, two standard deviations away from the mean).

2- About 99% of the examined cases fall within $\mu \pm 3\delta$ (that is, three standard deviations away from the mean).

In a normal curve, a deviation of more than 2δ is abnormal and a deviation of more than 3δ is very abnormal.

2.11. Variance

The standard deviation is the square root of the variance. In other words, the average is the squared deviation of the variables from the average. The variance does not have a unit and its absolute quantity is the criterion of action, because with the act of powering the measurement scale, it is squared, which is meaningless. If we calculate the difference of each number from the mean and raise it to the power of 2, then the sum of these scores divided by the total number of numbers is the variance. Variance depends on the value and difference between each number and the mean of the distribution of scores. If the variance of a group of data or scores is zero, then all the data are equal.

To calculate the variance in unclassified numbers, the first task is to calculate the mean, then we calculate the deviation (difference, difference) of each score from the mean and raise it to the power of 2 and divide the sum by the number of data. In calculating the variance, depending on whether we are dealing with a large population or a small sample, we use almost different relationships.

2.12. Variance of small communities

In cases where the number of observations is limited, the variance is equal to the sum of squares divided by N-1. It is called N-1degrees of freedom. On the other hand, according to the definition of the average, the standard deviation shows the dispersion around the average.

$$\text{Variance} = \frac{\text{sum of squares}}{\text{Degrees of freedom}}$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \text{ Theory}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{n-1} \text{ Functional}$$

Variance of large societies:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n} \text{ Theory}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{n} \text{ Functional}$$

If the number of observations is more than 30, the population and if it is less than 30, we take a sample.

Whenever N is large, the difference resulting from the application of N and N-1 in the variance formulas will be minor and the result of the two formulas (sample and population) will be almost the same. In many situations, N is not large and the difference in calculations from two formulas will not be insignificant. In this case, the value obtained by dividing the sum of squares by N is a biased estimate of σ^2 (population variance). This value describes $\hat{\sigma}^2$, which has a standard error, so the ratio $\frac{\sum X^2}{n-1}$ will give an unbiased estimate of σ^2 .

Example 2-17: Calculate the mean and variance of the numbers 18, 15, 12 and 21?

$$\mu_x = \frac{\sum x_i}{N} = \frac{66}{4} = 16.5$$

$$S^2 = \frac{\sum X_i^2 - \frac{(\sum X_i)^2}{N}}{N} = \frac{1134 - \frac{4356}{4}}{4} = \frac{1134 - 1089}{4} = 11.25$$

If the data is changed through subtraction or addition, their average will change accordingly and the deviations will remain

constant. which can be the basis for calculating the variance of data deviation from the average.

$$Y_i = X_i \pm C \quad \delta^2_Y = \delta^2_X$$

If we multiply or divide all the observations by a constant number, the variance of the new observations changes in proportion to the second power of that constant number.

$$Y_i = X_i \times C \quad \delta^2_Y = \delta^2_X \times C^2$$

To calculate the variance in classified numbers with a class distance greater than one, after calculating the average, we calculate the midpoint of the classes and subtract the average from it. We multiply the obtained deviation (difference, difference) to the power of 2 and then multiply by the frequency of each class and finally divide their sum by the number of data, that is, we use the following formula:

$$S^2 = \frac{\sum f_i(x_c - \bar{x})^2}{N}$$

Variance and Shepard's correction, the standard deviation obtained from the classified data of less than 12 classes are somewhat different from what is calculated on the original data, the reason is due to the use of the middle number of classes instead of the original data. In this direction, a correction called "Shepard's correction" is made on the calculated standard deviation. This variance is used in cases where, firstly, it is a continuous variable, secondly, the number of N is at least one thousand, and thirdly, it is a symmetric or slightly symmetric Freon distribution function. It is calculated from the following formula.

$$S^2 = S^2_x - \frac{c^2}{12}$$

2.13. Variance characteristics

1- If all the variables are equal, the variance is zero and vice versa. that's mean:

$$x_1 = x_2 = \dots = x_n \quad S^2_x = 0$$

2- If we add a constant number like a to the variables, their variance will not change. that's mean:

$$S^2_{(x+a)} = S^2_{(x)}$$

3- If we subtract a fixed number like a from the variables, their variance will not change. that's mean:

$$S^2_{(x-a)} = S^2_{(x)}$$

4- If we multiply the variables by a constant number like a, their variance is multiplied by the square of a. that's mean:

$$S^2_{(ax)} = a^2 S^2_{(x)}$$

5- If we divide the variables by a fixed number like a, their variance is divided by the square of a. that's mean:

$$S^2_{\left(\frac{x}{a}\right)} = \frac{S^2_{(x)}}{a^2}$$

$$S^2_{(ax \pm b)} = a^2 S^2_{(x)}$$

6- The variance of a fixed number is zero.

7- The greater the accumulation and concentration of scores around the average, the smaller the variance and vice versa.

8- Big or small variance is relative and variance is used when the measurement scale is at least a distance.

9- Variance is never negative.

2.14. Algebraic operation of mean and variance

If K population with the number of observations $N_1, N_2 \dots N_k$ with means $N_1, N_2 \dots N_\mu$ with variances $S_1^2, S_2^2 \dots S_k^2$ are combined as a single population and create a total population, the mean and variance of that total population with It is obtained using the following relations.

$$\mu = \frac{\sum N_i \mu_i}{N}$$

$$S^2 = \frac{\sum N_i S_i^2}{N} + \frac{\sum N_i (\mu_i - \mu)^2}{N}$$

2.15. Variance of differences and sums

Frequency distribution of differences is one of the common distributions in statistics. This distribution is used to compare differences between societies, which is a common topic in research.

for example:

$$d_i = x_i \pm y_i$$

Sum of two variables x_i and y_i $S_i = \delta^2_{x_i} + \delta^2_{y_i} + 2Cov_{x,y} = S_i \delta^2$

Difference of two variables x_i and y_i $D_i = \delta^2_{x_i} + \delta^2_{y_i} - 2Cov_{x,y} = D_i \delta^2$

2.16. Covariance (Cov)

Covariance is a statistical index that shows the intensity and direction of this common influence if two variables are affected by common factors. If two variables are independent, their covariance will be zero. It is obvious that if the covariance of x and y is not equal to zero, part of the variation observed in x and y comes from the influence of a series of common factors. If the value of Cov is positive, it indicates

the existence of a direct relationship, and if its value is negative, it indicates the existence of an inverse relationship between x and y . The covariance of two variables is an indicator that is used to measure the degree of relationship between them. The dependence between two variables can be one of these three modes:

1- When one variable increases, the other increases and when it decreases, the other decreases (direct dependence).

2- When one variable increases, the other decreases and when it decreases, the other increases (inverse relationship).

3- The increase or decrease of one variable has no effect on the other (two variables are independent).

We usually use the following formula to calculate covariance.

$$\text{Cov}_{x,y} = \frac{\sum(X-X)(Y-Y)}{n-1}$$

$$\text{Cov}_{x,y} = \frac{\sum XY - \sum X \sum Y / N}{n-1}$$

If X and Y are two discrete random variables, then the covariance of X and Y , expressed as $\text{Cov}(X, Y)$ or δ_{xy} is equal to:

$$\text{Cov}(X, Y) = \sum \sum (X - \mu_x)(Y - \mu_y) f(x, y)$$

Using the mathematical expectation, we can write:

$$\text{Cov}(X, Y) = E(X - \mu_x)(Y - \mu_y)$$

The above relationship can be expressed as follows.

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

In this regard, if X and Y are discrete.

$$E(XY) = \sum_i \sum_j x_i y_j \cdot f(x_i, y_j)$$

Covariance represents the simultaneous changes of X and Y . So:

If $0 > \text{Cov} (Y, X)$, then the changes of X and Y are in the same direction (direct dependence).

If $0 < \text{Cov} (Y, X)$, then the changes of X and Y are opposite (inverse relationship).

If $0 = \text{Cov} (Y, X)$, then X and Y have no linear relationship with each other (independent).

2.17. Properties of covariance

If a, b, c and d are fixed numbers, we have:

$$\text{Cov} (aX \pm b, Cy \pm d) = ac \text{Cov} (X, Y)$$

$$\text{Cov} (X, X) = \text{var} (X)$$

$$\text{Cov} (X, Y) = \text{Cov} (Y, X)$$

$$\text{Cov} (X, c) = 0$$

If X and Y are two random variables with joint distribution $f(y, x)$, then if X and Y are independent of each other, the covariance of X and Y will be zero.

$$\text{Cov} (X, Y) = 0$$

That is, two random variables X and Y do not have any linear relationship. But it cannot be said that if $\text{Cov} (X, Y) = 0$, those two variables are independent of each other. It is possible that two random variables X and Y are non-linearly related to each other and the covariance between them is zero.

Example 2-18: If the variance of X is equal to 4 and the variance of Y is equal to 3 and their covariance is 2, then the variance of $Z = 2X - 3Y$ is equal to:

$$Z = 2X - 3Y$$

$$Z^2 = (2X - 3Y)^2 = 4X^2 + 9Y^2 - 12XY$$

$$\delta^2 = 4\delta^2 + 9\delta^2 - 12 \text{Cov}_{xy}$$

$$\delta^2 = 4 \times 4 + 9 \times 3 - 12 \times 2 = 19$$

Analysis of covariance allows the researcher to equate the pre-test conditions of the groups in terms of the studied variables.

The difference between the initial positions of the groups can be statistically removed in such a way that the groups are comparable, as if they were equivalent and similar to each other at the beginning.

For example, if a researcher, after conducting an experiment, realizes that he did not control the disturbing variable that the related information was also available, he can remove the effect of this variable by using the covariance analysis method.

Example 2-19: What is the covariance value of the following joint probability function?

	y	
x	0.2	0.4
	0.1	0.3

$$\text{Cov}_{x,y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{n-1}$$

$$= \frac{\{(0.2 \times 0)(0.2 \times 0)\} + \{(0.2 \times 0)(0.1 \times 0)\} - \frac{0.4 \times 0.7}{2}}{2-1} = 0.44$$

2.18. Skewness coefficient

To compare two societies, in addition to the central and dispersion parameters, there is another criterion called skewness. The skewness of the distributions is determined in comparison with the symmetrical distribution. In other words, the skewness is a distribution deviation from the symmetry state and it is divided into two curves: skewed to the right and skewed to the left, which is represented by SK. In the following formula, asymmetry is known as Pearson's asymmetry coefficients.

$$SK_1 = \frac{Mean - Mode}{S}$$

$$SK_2 = \frac{3(Mean - median)}{S}$$

To calculate the skewness coefficient more accurately, the ratio of the central third order moment to the cube of the standard deviation is used. that's mean:

$$SK = \frac{\mu_3}{S^2}$$

That:

$$\mu_3 = \frac{\sum(X_i - \mu)^3}{N} \quad \text{or} \quad \mu_3 = \frac{\sum F_i(X_i - \mu)^3}{N} \quad \text{or} \quad \mu_3 = \sum f_i (x_i - \mu)^3$$

Skewness coefficient in terms of quartiles:

$$SK_Q = \frac{Q_3 - Q_2 + Q_1}{Q_3 - Q_1}$$

Skewness coefficient percentage:

$$SK_P = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Interpretation of the coefficient of skewness

1- If $SK = 0$, the distribution has no skewness (probability distribution).

2- If $0 > SK$, the distribution has negative skewness.

3- If $SK < 0$, the distribution has positive skewness.

4- If $0.1 \leq |SK|$, the distribution has a slight skewness (almost normal).

5- If $|SK| < 0.5$, the distribution is highly skewed (significantly different from the normal distribution).

6- If $0.5 < |SK| < 1$, the skewness of the distribution is low (it differs slightly from the normal distribution).

Example 2-20: If the mean, median, and variance values for a set of data are 6, 5, and 4, respectively. The degree of asymmetry or skewness of that data set will be equal to:

$$SK = \frac{3(\text{Mean} - \text{median})}{s}$$
$$SK = \frac{3(6-5)}{2} = \frac{3}{2} = 1.5$$

Example 2-21: Calculate and interpret the skewness coefficient for the assumed data?

$$Q_1 = 36 \quad , Q_2 = 43 \quad , Q_3 = 64$$

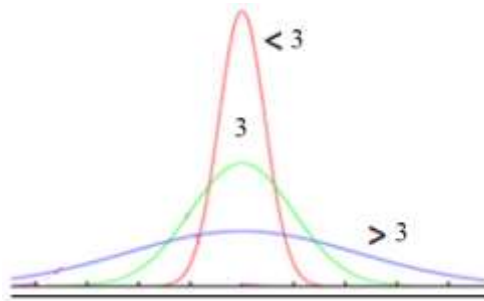
$$SK_Q = \frac{Q_3 - Q_2 + Q_1}{Q_3 - Q_1} = \frac{64 - (2 \times 43) + 36}{64 - 36} = 0.5$$

The skewness of the distribution is relatively low.

2.19. Slenderness ratio

This index, like the variance and standard deviation, is about the roundness of the shape or the width of the shape. In other words,

whenever the elongation of distributions is discussed, it actually refers to the length and the shortness of the abundance distribution. Sometimes, the mean, median, and mean of the studied group may be equal and the shape of the curves is symmetrical, but the dispersion around the mean, i.e., the standard deviation, is different in different groups, in this case, the curves will be different in length as shown in the figure below. The index to measure the dispersion of the society compared to the normal distribution is called "stretch coefficient" which is denoted by E and it is the ratio of the central fourth order moment to δ^4 and because the stretch of the distributions is compared with the normal distribution and the stretch of the normal distribution is equal to 3. A value less than 3 is high and a value greater than 3 is low.



Therefore:

$$E = \frac{\mu_4}{\delta^4} - 3$$

That:

$$\mu_4 = \frac{\sum(X_i - \mu)^4}{N} \quad \text{or} \quad \mu_4 = \frac{\sum F_i(X_i - \mu)^4}{N} \quad \text{or} \quad \mu_4 = \sum f_i(x_i - \mu)^4$$

The quantum elongation value can also be obtained from the following equation.

$$E_P = \frac{1/2(Q_3 - Q_1)}{P_{90} - P_{10}} - 0.263$$

In the normal distribution, which is a symmetrical distribution and the mean, median, and mode coincide, the coefficient of elongation is equal to 0.263. The constant number of 0.263 is the elongation of the normal distribution.

Interpretation of elongation factor

- 1- If $E=0$, the distribution is normal.
- 2- If $0 > E$, the distribution is shorter than the normal distribution.
- 3- If $E < 0$, the distribution is more elongated than the normal distribution.
- 4- If $|E| \leq 0.1$, the distribution is almost normal.
- 5- If $0.5 < |E|$, the difference between distribution and normal distribution is extreme.
- 6- If $0.5 < |E| < 0.1$, the difference between distribution and normal distribution is not extreme.

Elongation is a standard without a unit that represents height and has an inverse relationship with dispersion.

If $E = 0$, the extension is both the size and the height of the normal distribution, if $E < 0$, it is larger than normal and less spread, and if $E > 0$, it is shorter than normal and more spread.

Example 2-22: Calculate and interpret the elongation coefficient for a population with

$$N = 20 \text{ and } \sum(X_i - \mu)^4 = 400, \delta = 2?$$

$$\mu_4 = \frac{\sum(X_i - \mu)^4}{N} = \frac{400}{20} = 20$$

$$E = \frac{\mu_4}{\delta^4} - 3 = \frac{20}{16} - 3 = -1.75$$

$$E = |-1.75| > 0.5$$

The distribution is very different from the normal distribution.

Pearson Skewness: It should be known that a scientist named Pearson has provided a formula to detect the type of skewness, in which if the resulting number is positive, the curve has a positive skewness, and if the resulting number is negative, the curve has a negative skewness.

$$g_1 = \frac{\bar{x} - mo}{sd}$$

1- If $0.1 \leq |sk|$ Yes, there is almost no crookedness and the society is normal.

2- If $0.5 \leq |sk| \leq 0.1$, the existing skewness is small but not negligible. In fact, the society is slightly different from the normal distribution in terms of symmetry.

3- If $|sk| \geq 0.5$, the skewness is high and cannot be neglected. In other words, the community has a significant difference from the normal distribution in terms of color.

2.20. Relative scattering parameters

In this section, we will introduce parameters that can better understand the distribution of observations by presenting them along with the average and standard deviation. Relative dispersion indices are obtained by dividing a dispersion parameter by a central parameter of the same dimension or by dividing two dispersion indices of the same dimension, which are called "coefficients".

2.21. Coefficient of variation or dispersion

In some cases, comparing the dispersion of two or more frequency distributions is desired. So that the standard deviation is one of the most common dispersion indices that can be used to compare the dispersion of two distributions. But in cases where these two distributions have different measurement units, it is better to use the coefficient of variation (C.V) to calculate their dispersion. If the standard deviation of each distribution is divided by its mean, the relative standard deviation or relative dispersion is obtained. This index, which has no unit of measurement, is known as the coefficient of variation or coefficient of dispersion (C.V). The main disadvantage of the coefficient of variation is that it cannot be used when \bar{x} is close to zero. The lower the C.V, the more different the scores.

$$CV = \frac{s}{\mu} \times 100$$

2.22. Characteristics of the coefficient of variation

1- If all the data are equal, then $C.V = 0$.

2- If we sum all the data with a fixed number (opposite to zero), then the $C.V$ will decrease.

3- If we subtract all the data by a fixed number (opposite to zero), then the $C.V$. will increase.

2- If we multiply or divide all the data by a fixed number (opposite to zero), then the $C.V$. will not change.

2.23. Cases of use

It is used when two or more communities have heterogeneous observations in terms of measurement units, such as one community in meters and one community in inches, or several communities have different averages.

Sometimes the scale of the attribute to be measured is the same in two societies, but the size of their observations is significantly different. Like comparing the spread of profit and loss in handicraft industries with heavy industries.

The coefficient of variation of fixed numbers is zero.

The higher the coefficient of variation, the greater the differences and the lower the coefficient of variation, the smaller the differences.

Comparing two tests, the lower the coefficient of variation, the more difficult the test.

The coefficient of variation or dispersion is used to determine and compare the dispersion of two or more variables in one group and also

to compare the dispersion of one variable from two different communities.

2.24. Dispersion indices for nominal and rank scale data

In the case of nominal and rank scale data, it is possible to calculate the homogeneity or heterogeneity ratio of the classes, which is called dispersion ratio (VR).

$$VR = 1 - \frac{Fm}{N}$$

In this formula, Fm means the maximum frequency in the classes of the table.

As the value of VR tends to zero, it becomes scattered, in a special case where there is only one floor in the distribution, the value of VR will be equal to zero.

Example 2-23: Two traits of cluster length and seed weight were studied in a statistical survey; So that the standard deviation and mean for the trait of cluster length were 0.14 and 10 cm, respectively, while for the weight of 1000 seeds, the standard deviation was 0.56 and the mean was 40 grams. Which shows more dispersion?

To compare the dispersion of two traits, the coefficient of variation or cv is used.

$$CV_1 = \frac{0.14}{10} \times 100 = 1.4$$

$$CV_2 = \frac{0.56}{40} \times 100 = 1.4$$

Therefore, the dispersion of both methods is equal.

2.25. Standard score or sigma score

If the scale of the data and the range of changes in two or more frequency distributions are not the same, the data should be converted to relative numbers for comparison. To transform the data, one must calculate the deviation of each from the mean and then divide it by the unitless standard deviation. In other words, the deviations are expressed relative to the average dispersion. In other words, the standard score determines that a score is a few standard deviations higher or lower than the average. This relative number is called standard score or sigma data or z score. Standard scores determine the individual's position in the group. They are used with minimum distance scale data. We use the opposite formula to convert raw scores to standard scores. So:

$$Z = \frac{X - \mu}{S}$$

Z_i is actually the standard score of X_i , which determines its relative position compared to other x 's. Therefore, standard scores will be a new variable with zero mean and one variance. Therefore, to compare different frequency distributions, their observations can be converted to Z or standard variable.

This variable is a discrete quantity, that is, it is independent of the measurement units of the mean and standard deviation, and thus it is very suitable for comparing different distributions. If the raw score, X , is above the mean, the Z score will be positive. If the raw score, X , is below the mean, the Z score becomes negative.

Example 2-24: A student scored 84 in the statistics exam, the average score was 76 and the standard deviation of the scores was 10. He scored 90 in the course of machines, which has an average of 82 and

a standard deviation of 16. In which subject was his grade relatively better?

We convert raw scores to standard scores:

$$\text{Statistics} = \frac{x_1 - \mu_1}{s_1} = \frac{84 - 76}{10} = 0.8Z_1$$

$$\text{Machines} = \frac{x_2 - \mu_2}{s_2} = \frac{90 - 82}{16} = 0.5Z_2$$

$Z_1 > Z_2$ and the student's score in the statistics course has been better.

Mean and exponent in addition, subtraction, multiplication and division of a fixed number are the function of the mean and they change in the same way.

In addition, and subtraction, variance, standard deviation, average deviation, range of changes and quartile deviation are the same and do not change.

In multiplication and division, standard deviation, average deviation, quartile deviation and range of changes are multiplied or divided by the same number.

Variance is multiplied or divided by the square of that number.

The variance of fixed numbers is zero.

If several communities are combined, the mean and variance of the total combined community will be greater than the mean and variance of the constituent communities. Unless the mean of the communities is equal, in which case their variance is also equal:

$$S^2_{y_1} + S^2_{y_2} = S^2_{y_1} + S^2_{y_2} - 2\rho S_{y_1} S_{y_2}$$

The variance of differences in independent samples is equal to:

$$S^2_{\bar{X}_1 - \bar{X}_2} = S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2 = \frac{S^2_1}{N_1} + \frac{S^2_2}{N_2}$$

The total variance is equal to:

$$\text{Cov}(X, Y) = (S^2_X + S^2_Y) r S_X S_Y \pm S^2_X + S^2_Y = S^2_{(X+Y)}$$

Table 2-2: Comparison of characteristics of dispersion indices

Indicator	Weak points	Strengths points
Standard deviation	1- It is affected by too big or too small scores. 2- It cannot be calculated in the case of ordinal and nominal scale data.	1- It is the most stable dispersion. 2- All scores are used in its calculation. 3- The radius of rotation of the torque distribution device is around the center axis.
Range of quadrants	1- Compared to the standard deviation, it is less stable. 2- In the case of nominal scale data, it cannot be calculated.	1- It is not affected by big or small scores. 2- About 50 percent of the middle scores are relevant. 3- It is used for ordinal scale data.
Variation range	1- It is not affected by big or small scores. 2- About 50 percent of the middle scores are relevant. 3- It is used for ordinal scale data.	1- It is easy to calculate. 2- It gives us information about high and low scores. 3- In the case of distance and relative scale data, it can be calculated.

2.26. Torques

Generally, in any society, the size of different values is scattered around a point, which can be compared to the center of gravity of the society. Therefore, this center of gravity can be an arbitrary number or the origin of zero or the arithmetic mean.

General torques (relative to an arbitrary number a)

The torque of the nth order relative to the arbitrary number a is written in the following ways.

$$M_n(a) = \frac{\sum(X_i - a)^n}{N}, \quad M_n(a) = \frac{\sum F_i(X_i - a)^n}{N}, \quad M_n(a) = \sum f_i(x_i - a)^n$$

If we place $n = 1, 2, 3, 4$ in the general torques, the torques of the first, second, third and fourth orders will be obtained.

$$\begin{aligned} M_{1(a)} &= \frac{\sum(X_i - a)}{N} & , & & M_{2(a)} &= \frac{\sum(X_i - a)^2}{N} \\ M_{3(a)} &= \frac{\sum(X_i - a)^3}{N} & , & & M_{4(a)} &= \frac{\sum(X_i - a)^4}{N} \end{aligned}$$

2.27. Initial torques (relative to zero origin)

By placing $a = 0$ in the relation of general moments, initial moments are obtained. Therefore, the n th order torque relative to zero origin is defined as follows.

$$M_n = \frac{\sum X_i^n}{N} \quad , \quad M_n = \frac{\sum F_i X_i^n}{N} \quad , \quad M_n = \sum f_i x_i^n$$

If we place $n = 1, 2, 3, 4$ in the primary torques, the torques of the first, second, third and fourth order relative to the origin will be obtained.

$$\begin{aligned} M_1 &= \frac{\sum X_i}{N} = \mu & , & & M_2 &= \frac{\sum X_i^2}{N} \\ M_3 &= \frac{\sum X_i^3}{N} & , & & M_4 &= \frac{\sum X_i^4}{N} \end{aligned}$$

The first order moment relative to the zero origin is equal to the arithmetic mean.

2. 27. Central torques (relative to the average)

By placing $a = \mu$ in the relation of general torques, the central torques are obtained. Therefore, the torque of the n th order relative to the mean is defined as follows.

$$\mu_n = \frac{\sum(X_i - \mu)^n}{N} \quad , \quad \mu_n = \frac{\sum F_i (X_i - \mu)^n}{N} \quad , \quad \mu_n = \sum f_i (x_i - \mu)^n$$

If we place $n = 1, 2, 3, 4$ in the central torques, the first, second, third and fourth order torques will be obtained relative to the average.

$$\begin{aligned} \mu_1 &= \frac{\sum(X_i - \mu)}{N} = 0 & , & & \mu_2 &= \frac{\sum(X_i - \mu)^2}{N} = S^2 \\ \mu_3 &= \frac{\sum(X_i - \mu)^3}{N} & , & & \mu_4 &= \frac{\sum(X_i - \mu)^4}{N} \end{aligned}$$

The first-order torques relative to the mean is equal to zero, and the second-order torques relative to the mean is equal to the variance. The third torques is used to calculate skewness, that is:

$$SK = \frac{m_3}{m_3 \sqrt{m_3}}$$

The fourth torques is used to calculate elongation, that is:

$$K_g = \frac{m_3}{(m_3)^2} - 3$$

Example 2-25: The mean and variance of the frequency distribution of the variable attribute x is obtained as follows: $\delta^2_x = 400$ and $\mu_x = 5$, what is the second-order torques relative to zero?

$$\mu_2 = m_2 - m_1^2 \quad 400 = m_2 - (5)^2 \quad m_2 = 425$$

Exercises of chapter 2

1- 100 is subtracted from each of the primary data and the result is divided by 100. The mean and standard deviation of the changed data are equal to 10 and 5. What is the mean and variance of the original data?

2- 5 agricultural workers weed 100 square meters of land in 4, 5, 5, 6 and 7 hours, respectively. If these 5 workers want to weed 500 square meters of land, how many hours will it take?

3- If the mean of population A is equal to 4 and population B is equal to 6, what is the mean of the following linear function?

4- If the height of two wheat plants in a field is equal to 70 and 80 cm, respectively, and the corresponding standard grades of these two people are +1 and +2, respectively. What is the mean and standard deviation of the height of the plant?

5- A company manufacturing TV bulbs produces two types of bulbs A and B. Type A has a mean life of 1495 hours and a standard deviation of 280 hours. If type B has an average life of 1875 and a standard deviation of 310 hours. Which bulb shows more dispersion?

6- In a distribution

$$\sum \log \bar{X}_i F_i = 132.76$$

If the geometric mean of this distribution is 23.37, what is the number of data in this distribution?

7- According to the following data, answer the following questions:

X_i	1	5	9	13	17
F_c	5	11	19	25	30
f_i	5	6	8	6	5

- A): Find the arithmetic mean?
 - B): Which is the median?
 - C): Calculate the variance?
 - D): What is the average absolute value of the deviations from the mean?
 - E): Which is the fashion?
 - f): What is the approximate distribution of the said community?
- 8- If the variance of a sample $n = 10$ with 3 and $\sum x_i^2 = 49.5$, what is the mean of this sample?
- 9- Calculate the third quarter in the table below?

Categories	4-12	12-20	20-28	28-36	36-44	44-52	52-60	60-68	68-76
f_i	1	9	15	20.5	22	13	9	6	4.5

10- The following abundance table is available:

Category center	27	25	23	21	19	17	15
F_c	8	12	52	x	111	135	150

If the percentage relative frequency of the cluster whose center is 21 is equal to 24, what is the absolute frequency of the cluster whose center is 19?

11- If $\sum_{i=1}^4 X_i Y_i = 5$, $\sum_{i=1}^4 X_i = 7$, the numerical value is

$$\sum_{i=1}^4 (X_i - 3)(2Y_i + 1)$$

How much is it?

12- If X_i is a random variable with a mean of 15 and a standard deviation of 3. $d_i = \frac{X_i}{2}$ What is the coefficient of variation (CV) of variable d ?

13- Given that X_i is a continuous random variable and according to the opposite table, what is the variance of X_i ?

X_i	2	3	11
$P(i)$	1.3	1.2	1.6

CHAPTER 3

POSSIBILITIES

Dr. Harun GITARI¹

¹ - Kenyatta University, School of Agriculture and Enterprise Development, Department of Agricultural Science and Technology, Nairobi, Kenya
ORCID ID: 0000-0002-1996-119X, e-mail: harun.gitari@ku.ac.ke

INTRODUCTION

Phenomena whose outcome can be determined with certainty before they occur are called deterministic phenomena and phenomena whose outcome cannot be determined with certainty before they occur are called random phenomena.

In the science of probability, the act of collecting data is called an experiment, and if the result of this experiment cannot be determined with certainty in advance, it is called a random experiment. The set of all possible outcomes of a random experiment is called its sample space, which is usually represented by the letter S , and the number of elements of the sample space is represented by $n(s)$. Sample space is of two types:

1- If the sample space contains a limited number of elements or the number of elements is unlimited in terms of counting, it is called a discrete sample space.

2- If the sample space includes the set of all numbers between two specified limits, it is called a continuous sample space.

If a random experiment consists of m members and if we repeat this experiment n times, we have:

$$n(s) = m^n$$

Whenever r objects are selected from among n objects without placement and without order. we have:

$$n(S) = C(r, n) = \binom{n}{r}$$

Whenever r objects are selected from among n objects by placement and order. we have:

$$n(S) = n^r$$

When choosing r objects from among n objects without placement and in order. we have:

$$n(S) = P(r \cdot n)$$

3.1. Incident or event

Each subset of a sample space is called an event. Events are usually displayed with letters A, B, C, etc. Types of incident:

1- Simple event: It is an event that has only one member. Like coming up with the number 3, in throwing a dice {3}

2- Compound event: It is an incident that has more than one member. Like the occurrence of a lion, in the toss of two coins (heads or tails).

3- Impossible event: It is an event that does not have members. Like coming up with the number 7, in throwing a dice: { }

4- Certain event: It is an incident that has all the members of the sample space. Such as: the occurrence of a lion or a line in the toss of a coin: {face, back}

5- Random event: It is an event that can happen and also cannot happen. Like coming up with the number 5 when throwing a dice

3.2. Definition of probability

Probability is a topic of mathematics on which the principles of statistical theory are based. In addition, the practical aspect of probabilities is used in predicting future events based on available information and evidence. Definition of probability: The probability of a phenomenon is the number of possible states to the total number of

states. If S is a finite sample space, and the probability of occurrence of all members of S is equal, the probability of occurrence of any event like A is obtained from the following equation. In mathematical language

$$P(A) = \frac{n(A)}{n(S)}$$

In solving probability problems, first we find the set of all possible outcomes of a random experiment, then we determine the desired event as a subset of the sample space, and finally we obtain the probability of that event with the help of probability laws.

The probability function P defined by the above rule has the following properties.

1- $0 \leq P(A) \leq 1, S \subset A$

2- $P(S) = 1$

3- $P(\emptyset) = 0$

If event A causes event B to occur, then:

$$A \subset B \quad P(A) \leq P(B)$$

If event A causes event B to occur and at the same time event B also causes event A to occur, then:

$$A = B \quad P(A) = P(B)$$

If the results of a random experiment are not random, the property of the relative frequency of the event is used. The relative frequency of the event tends to a number in many repetitions of the experiments, which expresses the probability of the event occurring.

$$P(A) = \lim_{n \rightarrow \infty} f_n(A) = \frac{m}{n}$$

So if n is large, the ratio $\frac{m}{n}$ tends to a limit that we call probability.

3.3. Principles of probability

Consider a random experiment consisting of n random outcomes.

$$S = \{e_1, e_2, \dots, e_n\}$$

The probability of occurrence of each outlier in the sample space is a non-negative number. that's mean:

$$P(e_i) \geq 0$$

The sum of the probabilities of occurrence of all outcomes in the sample space is equal to one. that's mean:

$$P(S) = P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

The probability of occurrence of any given event such as A that has m members

$$A = \{e_1, e_2, \dots, e_m\}$$

Let $(B \subset A)$ be equal to the sum of the probability of occurrence of the outcomes that make up the said event. that's mean:

$$P(A) = P(e_1) + P(e_2) + \dots + P(e_m)$$

$$= \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = \frac{m}{n}$$

m times

If A_3, A_2, A_1 and ... are two-by-two incompatible events of S^1 , we have:

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

¹ - Both events that cannot occur at the same time are incompatible.

Random experiment (random experiment): It is an experiment whose result we cannot predict, but the set of possible results of that experiment can be predicted in advance.

For example, tossing a coin or a dice are considered random experiments.

Sample space (S): The collection of all possible outcomes (outcomes) of a random experiment is called the sample space of said experiment and is usually denoted by S.

For example, in tossing a coin:

$$S = (\text{front, back}), N(S) = 2$$

Whenever we toss a coin n times or n coins together, $n(s) = 2^n$. The gender sample space of n children also has 2^n members.

Whenever we throw a dice n times or n dice together, $n(s) = 6^n$.

If a bag containing n beads is distinct and we want to randomly remove k beads from this bag, the number of members of the sample space is equal to $n(s) = \binom{n}{k}$.

Example 3-1: There are 3 red balls and 5 distinct blue balls in a bag. We randomly choose a ball from this bag and after observing it, we put it aside and choose another ball again. In choosing these two balls, how many members does the sample space have?

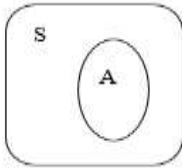
The random analysis we do is to first randomly select a ball from the bag, which can be done in $\binom{8}{1}=8$ ways. We leave the selected ball and choose another ball again. Because one of the number of balls is reduced, as a result, from the remaining 7 balls, we choose a ball at random, which can be done in $\binom{7}{1}=7$ ways. Let's be careful that

random experience includes these two actions that are performed one after the other (consecutively). As a result, according to the principle of multiplication, we write:

$$n(S) = \binom{8}{1} \times \binom{7}{1} = 8 \times 7 = 56$$

random event

Each subset of the sample space is a random event.



$$A \subset S \qquad 0 \leq n(A) \leq n(S)$$

For example, in throwing a dice, the first number appears as $A = (5,3,2)$.

Deterministic event: If S is the sample space of an experiment, then S is called a deterministic event.

Impossible event: An event that cannot happen is called an impossible event, in other words, an event that has no members is an impossible event. Any random experiment, φ , is called an impossible event.

3.4. Accidental occurrence

We say that event A has occurred when one of its members has occurred as a result of the experiment. As an example of a dice, the occurrence of an even number is $A = (6,4,2)$. Now, when we say that A has occurred, the number 2 or the number 4 or the number 6 appears

after throwing the dice. Therefore, as soon as we see any of the even numbers, we claim that event A has occurred.

The total number of random events in a random experiment whose sample space has $n(S)$ members is equal to $2^{n(S)}$.

$$\text{total number of random events} = 2^{n(s)}$$

Example 3-2: We throw the dice. We know that the number is greater than 3. How many random events are there in this random experiment?

$$S = (1,2,3,4,5,6) = (4,5,6) \quad n(s) = 3$$

$$\text{the total number of random events} = 2^{ns} = 2^3 = 8$$

The number of K-membered random events is equal to the number of K-membered subsets of the sample space $\binom{n(s)}{k}$.

$$\text{number of random occurrences of k members} = \binom{n(s)}{k}$$

Example 3-3: We throw the dice. In this random experiment, how many events of 3 members are there?

$$S = (1,2,3,4,5,6) \quad n(s) = 6$$

$$\text{the number of k member random events} = \frac{6 \times 5 \times 4 \times 3!}{6 \times 3!} = 20 \binom{6}{1} =$$

$$\frac{6!}{3! \times 3!}$$

3.5. The probability of a random event A

If S is the sample space of a random experiment and A is an event from this sample space (that is, A is a subset of S), the probability of occurrence of event A is represented by $P(A)$ and we define it as follows.

probability of occurrence of event A = $P(A) = \frac{n(A)}{n(S)}$

The number of event members A

The number of members of the sample space S

The probability number is always a number between zero and one, i.e. $0 \leq P(A) \leq 1$

If the probability number is equal to zero, it is impossible (impossible) and if it is equal to 1, the event is certain.

did not happen $P(\phi) = \frac{n(\phi)}{n(s)} = 0$

Certain event $P(\phi) = \frac{n(s)}{n(s)} = 1$

Example 3-4: 3 white mice and 5 black mice are kept in the laboratory. If 4 mice are randomly selected for testing, what is the probability that only one of the tested mice is white?

3 white mice and 5 black mice are kept in the laboratory. If we randomly take 4 mice from among them for testing, the number of members of the sample space is equal to $n(s) = \binom{8}{4}$. Now, to determine the probability that only one of the mice is white (and the other 3 mice are black), we need to determine $n(A)$. we have:

S selection of 4 mice from the total of 8 mice

$$N(S) = \binom{8}{4} = 70$$

A = choosing 1 white mouse and 3 black mice

$$N(A) = \binom{3}{1} \times \binom{5}{3} = 3 \times 10 = 30$$

$$P(A) = \frac{n(A)}{n(s)} = \frac{30}{70} = \frac{3}{7}$$

Example 3-5: In throwing two dice, what is the probability that the sum of the two dice is 6?

The number of members of the sample space is equal to $n(s) = 36$. To find the number of members of the random event "the sum of two dice equals 6", it is enough to write ordered pairs whose sum of the first and second components is equal to 6. To do this, we must use a special order. we have:

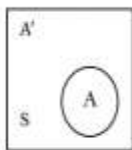
Sum of two dice equals 6 $A = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$

$$N(A) = 5 \quad P(A) = \frac{n(A)}{n(s)} = \frac{5}{36}$$

3.6. Complementary probability of event A

If we assume that A is an event, we denote its complement by A'. A' is an event that, if it occurs, A will never occur, and vice versa, that is, they are incompatible, and secondly, the sum of their probabilities is equal to 1. For example, if an even number comes up and an odd number comes up in throwing a dice, the probability of the occurrence of the two complementary events is zero.

Two complementary events are necessarily incompatible, but two incompatible events are not necessarily complementary to each other.



$$P(A') = 1 - P(A)$$

$$P(A) = 1 - P(A')$$

or

$$P(\text{favorable}) = 1 - P(\text{unfavorable})$$

In cases where the calculation of the unfavorable probability of the problem is easier than the calculation of the favorable probability.

In this situation, it is enough to calculate the unfavorable probability and subtract it from one.

Two complementary events are also incompatible because:

$$A \cap A' = \phi \qquad P(A' \cap A) = 0$$

The great advantage of the complementation rule is that it can sometimes significantly reduce the calculations required to solve the problem.

Example 3-6: We throw a pair of dice. What is the probability that the sum of the numbers turned up is less than 10?

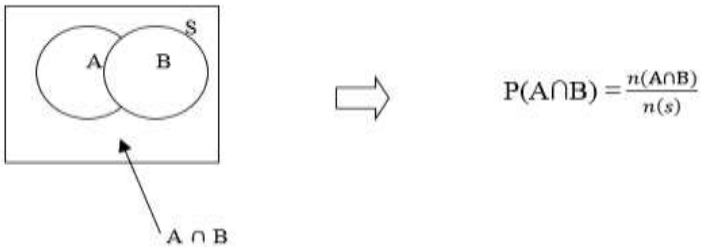
To calculate the probability of the sum of numbers smaller than 10, we must calculate the probability of the sum of 9 to the sum of 2 and add them together. This is very time consuming and tiring. Then we use the complementary probability of the event. we have:

$$A = \{ \text{The sum of the two numbers on the front is less than 10} \}$$

$$P(\text{Sum smaller than 10}) = 1 - P(\text{Sum of 12 or sum of 11 or sum of 10}) = 1 - \frac{3+2+1}{36} = 1 - \frac{6}{36} = \frac{5}{6}$$

3.7. Sharing two events

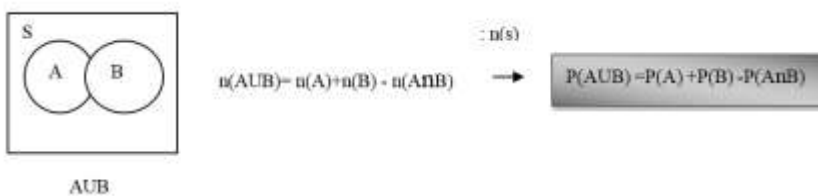
If A and B are two events, the event $A \cap B$ occurs when both events A and B occur. In other words, whenever we are asked to obtain the probability that A and B occur, we must calculate $P(A \cap B)$. we have:



3.8. The community of two events

If A and B are two events, then $A \cup B$ occurs when one or both events A and B occur. In other words, whenever we are asked to obtain the probability that A or B will occur, we must calculate $P(A \cup B)$. Note that the probability of event A or B is not necessarily equal to the sum of the probability of event A and the probability of event B (in other words, it does not mean that we necessarily add the probability of both events together), but we must:

In the case that two phenomena A and B are compatible. The following relationship is called the law of total probability.



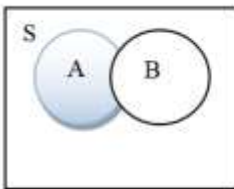
If two events A and B are inconsistent, then at least one of the two events occurs, i.e. A or B (i.e. $A \cup B$) has occurred. The probability of occurrence of one of them, A or B, is equal to the sum of their simple probabilities, i.e.:

$$P(A \cup B) = P(A) + P(B)$$

Whenever we say the occurrence of event A or B, we mean $B \cup A$. And when we say the occurrence of events A and B, we mean $A \cap B$.

3.9. Difference of two events

If A and B are two events, A-B occurs when event A occurs but event B does not occur. we have:



$$P(A-B) = P(A \cap B') = P(A) - P(A \cap B)$$

Event A-B occurs when only event A occurs and B-A occurs when only event B occurs.

Example 3-7: In rolling a pair of dice, what is the probability that two dice appear the same or their sum is greater than 9?

$$n(s) = 6^2$$

two dice appear $A = \{(6, 6), (5, 5), (4, 4), (3, 3), (2, 2), (1, 1)\}$

$$n(A) = 6$$

The sum of two dice must be greater than 9. $B = \{(6, 6), (5, 6), (6, 5), (4, 6), (5, 5), (6, 4)\}$

$$n(B) = 6$$

$$A \cap B = \{(6, 6), (5, 5)\}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{6}{36} + \frac{6}{36} - \frac{2}{36} = \frac{10}{36} = \frac{5}{18}$$

Example 3-8: The probability that a student will be accepted in the physics course is 0.55 and the probability of being accepted in the chemistry course is 0.6. If the probability of passing in at least one of the two courses is 0.75, with what probability will it be passed in both courses?

We consider the probability of being accepted in the physics course as A and the probability of being accepted in the chemistry course as B. we have:

$$P(A) = 0.55$$

$$P(B) = 0.6$$

The probability of passing in at least one of the two courses is equal to the probability of AUB, so $P(A \cup B) = 0.75$. Now, to determine the probability of passing in both subjects (passing physics and passing chemistry or $A \cap B$), we have:

$$P(\text{Passed in both courses}) = P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.55 + 0.6 - 0.75 = 0.4$$

Example 3-9: According to the previous problem, the probability of passing only in physics course is equal to:

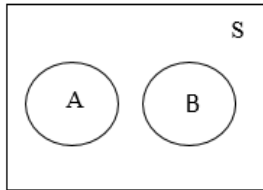
A pass in physics course means A-B, so we have:

$$P(A - B) = P(A) - P(A \cap B) = 0.55 - 0.4 = 0.15$$

3.10. Compatible and incompatible events

Two events A and B are said to be incompatible if they cannot occur together. In other words, the occurrence of one means the non-occurrence of the other. $A \cap B = \phi$ in two incompatible events.

For example, in throwing a dice, the events "even" and "odd" are inconsistent.



If two events are not incompatible, they are called compatible. Like the occurrence of "coming even" and "coming first number" which have only one number in common, that is the number "2". In two incompatible events A and B, the probability that A or event B will occur is equal to:

$$A \cap B = \emptyset \rightarrow n(A \cap B) = 0 \longrightarrow P(A \cap B) = \quad P(A) + P(B) = P(A \cup B)$$

This theorem can be generalized for more than two events.

If A and B are two incompatible events, then A and B' and B and A' are definitely compatible. If A and B are not complementary, A' and B' will also be compatible.

Example 3-10: If $P(A) = \frac{2}{3}$, $P(B) = \frac{1}{6}$ and $A \cap B = \emptyset$, then what is the probability of event A or B?

Because $A \cap B = \emptyset$, we conclude that two events A and B are incompatible with each other. In two incompatible events, the probability of occurrence of A or B is equal to:

$$\text{A and B are incompatible } P(A \cup B) = P(A) + P(B) = \frac{2}{3} + \frac{1}{6} = \frac{5}{6}$$

3.11. Two independent events

Two events A and B are said to be independent if the occurrence of one does not affect the probability of the other occurrence. In two independent events, we calculate the probability of event A and event B from the following equation. we have:

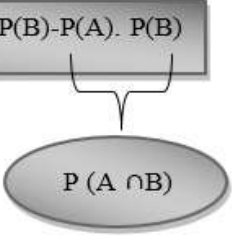
$$P(A \cap B) = P(A) \cdot P(B)$$

If the above equality is not established, A and B are called two dependent events.

The above theorem can be generalized for independent events, A_1, A_2, \dots, A_n

$$(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \dots P(A_n)$$

As a result, the probability that event A or event B will occur is equal to:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$


The diagram illustrates the formula for the probability of the union of two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. A horizontal rectangular box contains the equation. A bracket underneath the terms $P(A)$ and $P(B)$ in the equation points down to an oval containing the expression $P(A \cap B)$, indicating that this term is subtracted from the sum of $P(A)$ and $P(B)$.

If A and B are two events, the probability of occurrence of only event A or only event B is equal to:

$$P(A \cup B) = P(A) + P(B) - 2P(A \cap B)$$

When two events are independent, knowing the outcome of one of them cannot predict the outcome of the other. Therefore, the probability of each event is self-evident and has nothing to do with

others. For example, when we throw two dice, if it is announced that the first dice is a 5, how likely is it that the second dice is a 3? I say: $\frac{1}{6}$

Because the second dice has nothing to do with the first dice and they are independent of each other. Now, if they say how likely it is that the first dice will come up as 5 and the second dice as 3 (that is, they want to share them), we say:

$$P(A \cap B) \longrightarrow P(A) \cdot P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

If A and B are two independent events, then the events A and B', B and A', and A' and B' will also be two independent events.

If A and B are two independent events, we obtain the probability that only event A does not occur or the probability that only event B does not occur from the following formulas. we have:

B' and A are independent

$$P(A - B) = P(A \cap B') \longrightarrow P(A) \cdot P(B')$$

A' and B are independent

$$P(B - A) = P(B \cap A') \longrightarrow P(B) \cdot P(A')$$

Example 3-11: If A and B are two independent events and

$P(A \cup B) = 0.8$ and $P(A) = 0.5$, then $P(B')$ is equal to:

$$P(A \cup B) = 0.8 \quad P(A) + P(B) - P(A) \cdot P(B) = 0.8$$

$$0.5 + P(B) - 0.5P(B) = 0.8$$

$$0.5 P(B) = 0.3 \quad P(B) = \frac{0.3}{0.5} = 0.6 \quad P(B') = 1 - P(B) = 0.4$$

$$= 0.4$$

Example 3-12: When $P(A) = 0.2$, $P(B) = 0.3$ and A and B are independent of each other, $P(B-A)$ is equal to:

$$P(B-A) = P(B \cap A') \longrightarrow P(B) \cdot P(A') = 0.3 \times 0.8 = 0.24$$

$\underbrace{\hspace{10em}}_{1-P(A)}$

A' and B are independent

3.12. Conditional probability

When two events are dependent on each other, the occurrence or non-occurrence of one affects the occurrence or non-occurrence of the other. In this case, the occurrence of one is calculated after the other has occurred. Such probability is called conditional probability.

The probability of A occurring after B has occurred is called the conditional probability of A conditional on B and is represented by the symbol $P(A/B)$. In other words, if A and B are two events from the sample space S, when $P(B) \neq 0$, the probability of occurrence of event A provided that event B has occurred is equal to: $(P(B) \neq 0)$.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \text{Conditional probability of B provided that A occurs}$$

Obviously, when $P(B) = 0$, the probability of the condition cannot be defined.

In conditional probability questions, we consider the event that has occurred, i.e. event B, as a new sample space and select the members of the desired event, i.e. A, from within it. Then we also get

the conditional probability of A under the condition of occurrence of

$$B \text{ from the relation } P(A/B) = \frac{n(A)}{n(B)}.$$

The knowledge of the occurrence of a specific event acts as a condition and causes it to usually make the sample space smaller for finding outcomes belonging to another event. Therefore, the calculation of conditional probabilities becomes the calculation of a simple unconditional probability.

Whenever A and B are two arbitrary occurrences of S, we always have:

$$\begin{array}{ll} P(A/A') = 0 & P(A/B) \leq P(A) \\ P(A/A) = 1 & P(A'/A) = 0 \end{array}$$

Whenever A and B are two incompatible events, we have:

$$A \cap B = \phi \quad P(A \cap B) = 0 \quad P(A/B) = 0 \quad P(B/A) = 0$$

If A and B are two independent events, we have:

$$P(A \cap B) = P(A) P(B) \quad P(A/B) = P(A) \quad P(B/A) = P(B)$$

Example 3-13: Office employees are distributed according to the table below. What is the probability that a male employee has university education?

		Gender	
		Female	Man
Education	Academic	10	15
	Less than academic	80	90

In the text of the question, it is mentioned that "...a male employee..." after hearing this sentence, we come to the conclusion that

we know that the selected employee is definitely a man (it happened). So, we consider the event of being "male" as a new sample space and determine the members of the desired random event from within this sample space. we have:

$$\begin{aligned}
 & \text{Event B has occurred} \\
 B = (\text{being a man}) & \longrightarrow n(B) = n(S_{\text{new}}) = 15 + 90 = 105 \\
 A = (\text{Having an academic education}) & \longrightarrow n(A) = 15 \\
 P(A/B) = \frac{n(A)}{n(B)} & = \frac{15}{105} = \frac{1}{7}
 \end{aligned}$$

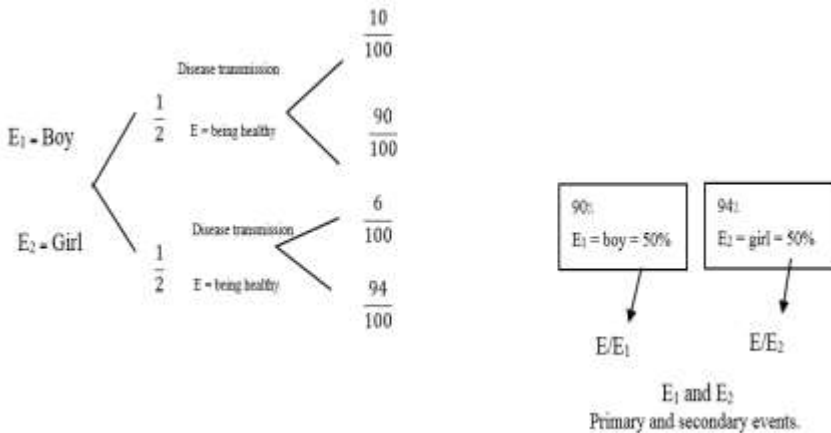
Sometimes the sample space turns into sets E_1, E_2, \dots, E_n , none of which have anything in common and one of them must occur (in other words, the sample space is divided into two incompatible events). Now, to obtain the probability of occurrence of event E, which is located within the above events, we use the following relationship:

$$P(E) = P(E_1).P(E/E_1) + P(E_2). P(E/E_2) + \dots + P(E_n). P(E/E_n)$$

In such cases, two or more nested possibilities are usually involved. Although the above formula exists to solve them, it is difficult to use the above formula. To solve this type of problem, the best method is to first specify the primary probability and the secondary probability and consider them in the form of a tree. Then we specify the branches that end in the desired state and multiply the probabilities on those branches together and add the results from the other branches.

Example 3-14: The probability of transmitting a hereditary disease from parents to a male child is 10% and to a female child is 6%. What is the probability that a child who is born will not have this type of disease?

If we represent the event of being a boy with E_1 , the event of being a girl with E_2 and the event of being healthy with E , we have: 90



$$P(E) = P(E_1).P(E/E_1)+P(E_2).P(E/E_2) = \frac{1}{2} \times \frac{90}{100} + \frac{1}{2} \times \frac{94}{100} = \frac{92}{100} = 0.92$$

Important note: If the probability of E_1 and E_2 events (initial events) is not mentioned in the problem, we assume that their probabilities are equal, that is:

$$P(E_1) = P(E_2) = \frac{1}{2}$$

If we assume the number of primary events to be n . In this case, the probability of each of the initial events is equal to $\frac{1}{n}$, that is:

$$P(E_1) = P(E_2) = \dots = P(E_n) = \frac{1}{n}$$

3.13. Random variable

If we assign a unique number to each test result in the experiment, we call this number a random variable and we usually show it with capital letters x , y , etc. In other words, a random variable is a function of a sample space of real numbers or a function whose domain is defined by members of the sample space and its range is a subset of real numbers.

$$X: S \rightarrow \mathbb{R}$$

For example, when we throw a coin, the sample space is $S = \{B, F\}$. If the random variable X is the number of occurrences, we have:

$$X: \text{Number of coming front} \rightarrow X = 0, 1$$

The front of the coin shows zero, which means that the back of the coin has appeared.

Example 3-15: We throw a coin 5 times. If we define the random variable x equal to the number of "fronts", the result $P(X=2)$ is equal to:

Considering that X is the number of fronts in 5 coin tosses, it can take the values $x = 5, 4, 3, 2, 1, 0$. Now, to calculate $P(X=2)$, we need to obtain the probability of 2 times "front" (and 3 with "back"). we have:

$$X = 2 \quad P(x = 2) = \frac{\binom{5}{2}}{2^5} = \frac{10}{32} = \frac{5}{16}$$

3.14. Random variable probability distribution table

The probability distribution table of a random variable is a table in which all the values of the random variable and the probability of each occurrence are given. In other words, calculate the probability of

different states of a random variable, as if we have obtained the probability distribution of the random variable. The purpose of probability distribution is to determine how the probability is distributed over the values of the random variable. For example, we throw a coin 3 times. The probability distribution table of the random variable "number of occurrences" is as follows: $x = 0, 1, 2, 3$

x	0	1	2	3
$P(x = x_i)$	$\frac{\binom{3}{0}}{2^3}$	$\frac{\binom{3}{1}}{2^3}$	$\frac{\binom{3}{2}}{2^3}$	$\frac{\binom{3}{3}}{2^3}$

 \Rightarrow

x	0	1	2	3
$P(x = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

In a probability distribution table, the set of probability values is always equal to one.

Example 3-16: If the probability distribution table of the random variable x is front-to-front, what is a ?

x	1	3	5
$P(x = x_i)$	a	$5a^2$	$3a$

To determine a , it is enough to set the sum of the values of the second line of the probability distribution table equal to one. we have:

$$a + 5a^2 + 3a = 1 \quad 5a^2 + 4a - 1 = 0 \quad \left\{ \begin{array}{l} a = -1 \text{ Unacceptable, the number of} \\ \text{possibilities is not between 0 and 1.} \\ \\ a = \frac{1}{5} \end{array} \right.$$

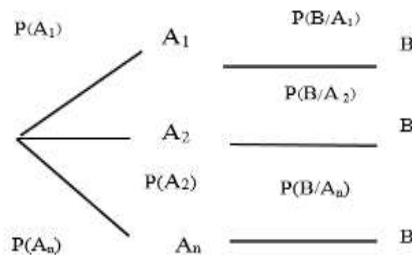
3.15. Bais theorem or law

Sometimes the possibility of a phenomenon can happen in different ways. Suppose that B_1 to B_n are n number of incompatible

events, among which only one state occurs, and also suppose that the event E is a phenomenon that can occur if any of the Bi occurs, so that P(E/Bi) can be obtained Bi/E by using Biss law as follows.

$$P(B_i/E) = \frac{P(B_i) \times P(E/B_i)}{\sum_{i=1}^n P(B_i) \times P(E/B_i)}$$

The denominator of the fraction is the same as the law of total probability. This formula can be shown as follows.



Example 3-17: Three machines A, B, and C produce 50%, 30%, and 20% of the total factory output, respectively. The percentage of defective products of these machines is 3%, 5% and 6% respectively. Among the products of this factory, a product is randomly selected. What is the probability that it was produced by machine C, if we know that the product is defective?

$$P(C \text{ car} / \text{defective goods}) = \frac{0.20 \times 0.06}{(0.5 \times 0.03) \times (0.03 \times 0.05) \times (0.20 \times 0.06)} = \frac{0.012}{0.042} = 0.28$$

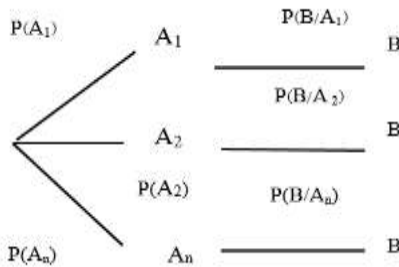
3.16. Law of Total Probability

Suppose $A_1, A_2, + \dots An$ are two-by-two incompatible events whose community is S, that is, $U Ai$. If the event B can occur with any

of the events A_i , in such a case, to calculate the probability of occurrence of B , we can use the relation used below:

$$P(B) = \sum_{i=1}^n P(A_i)P(B/A_i)$$

It displayed the above formula in the following tree diagram:



Example 3-18: In a factory, three machines C_1, C_2, C_3 produce a product. C_3 machine produces 45% of C_2 machine, 30% of C_1 machine, 25% of the total production of this product. 6%, 2% and 4% of C_3, C_2 and C_1 products have defects, respectively. We take one of this product out of the factory warehouse, what is the probability that this product has a defect?

$$P(B) = (0.45 \times 0.06) + (0.3 \times 0.02) + (0.25 \times 0.04) = 0.0405$$

3.17. The most tolerant event

An event that is more likely to occur than other events is called the most tolerant event. Therefore, in the expansion of $(p + q)^n$, there is a sentence whose probability is greater than other sentences of the expansion or at most equal to one of them that is located on both sides of it.

In n times of testing with constant probability p , we want to know which of the x values will have the greatest probability, for this we use the following formula:

$$np - q \leq x \leq np + p$$

1- If $p + np$ is an integer, then $p - np =$ an integer, in which case there will be two numbers for x values that will have the greatest probability.

2- If $p + np$ is a decimal number, then $p - np =$ decimal number, in which case the only integer between the above two decimal numbers will have the greatest probability.

3.18. Combined analysis

The purpose of composite analysis is to categorize objects or numbers or letters or people into several groups that differ according to specific conventions in terms of gender and type or placement of objects next to each other. Combinatorial analysis consists of 3 parts: permutation, order and combination

3.19. Factorial

Definition of $n!$ (n factorial): The product of natural numbers from n to 1 is called n factorial and we display it with the symbol $n!$.

$$n! = n(n-1)(n-2) \times \dots \times 3 \times 2 \times 1$$

Example 3-19: Find the result of the following expression?

$$3! = 3 \times 2 \times 1 = 6$$

According to the factorial convention, the numbers one and zero are equal to 1. that's mean:

$$0! = 1 \quad \text{and} \quad 1! = 1$$

The factorial of any number can be written as the product of the same number in the factorial of the consecutive number less than that number. For example:

$$n! = n(n-1)(n-2)!$$

Factorials cannot be added, subtracted, multiplied, or divided.

3.20. The principle of multiplication and addition

The principle of counting: suppose we want to perform two actions of choosing A and B one after the other. If the act of choosing A is performed in M ways and the act of choosing B is performed in n ways, then we can perform these two acts of selection together in $n \times m$ ways. In simple words, the number of actions of A and B is equal to $M \times N$ ways.

Generalization of the principle of counting: if an action is performed in N_1 different ways, and after it is performed, the second action is performed in N_2 different ways, and... and the next K action is performed in N_K different ways, these K actions together add up to $N_1 \times N_2 \times \dots \times N_k$ is done in different ways.

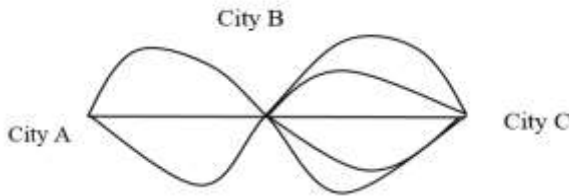
In principle, the multiplication of all A_i 's must be done, the conjunction "and" plays a decisive role in expressing the content of the principle of multiplication.

Addition principle: If the act of choosing A can be done in M ways and the act of choosing B can be done in N ways, then the number of states in which the act A or B is performed is equal to $M+N$.

If "and" is said between two or more selection actions, we multiply the number of their execution modes and if "or" is said, we add the number of their execution modes together.

In many problems, multiplication principle and addition principle are used together.

Example 3-20: There are three roads from city A to city B and four roads from city B to city C. In how many ways can you travel from city A to city C and back, provided that no route is taken twice?



With a little precision in the form of the problem, we see that going from city A to city C and then back to city A, includes 4 selection actions. These 4 actions consist of going from city B and going from city B to city C and returning from city C to city B and returning from city B to city A. Because there is a choice between these four actions and these actions It has been done consecutively, so according to the principle of multiplication, we count the different states of doing each of these actions and then multiply them together.

On the way from city A to city C, there is no problem to repeat the routes, so we can travel from city A to city C in $3 \times 4 = 12$ ways. But on the way back, one of the ways to return from city C to city B

and one of the ways to return from city B to city A will be removed (because we must have traveled one of the ways on the way) so to $2 \times 3 = 6$ ways from city C to city A. Therefore, in total, the number of round trip states, provided that no route is taken twice, is equal to:

$$3 \times 4 \times 3 \times 2 = 6 \times 12 = 72$$

Example 3-21: How many three-digit numbers greater than 300 can be written with the digits 0, 1, 2, 3, 4, 5?

Whenever there is a zero among the given numbers and the number of even numbers or multiples of 5 (divisible by 5) is asked, and the condition that repetition of numbers is not allowed is emphasized, we solve the issue in the following two cases:

The first case: zero digit in one. Or the second case: the number zero is not in one.

Because there are zeros among the given numbers and the condition that repetition of numbers is not allowed is emphasized, to determine the number of even three-digit numbers, we consider the following two situations. we have:

First state: zero in one:

$$5 \times 4 \times 1 = 20$$

Definitely non-zero Zero is located

The second case: zero is not in one:

$$4 \times 4 \times 2 = 20$$

Now, since we add "or" between these two situations, according to the addition principle, the answer is equal to:

$$\text{first mode} + \text{second mode} = 20 + 32 = 52$$

Definitely non-zero 2 or 4 is located
and the digit is one

3.21. Permutation

If we have n distinct objects, we call each state of placing these n objects next to each other, so that the order in which the objects are placed is important, a permutation of that object.

The number of permutations of n distinct objects is represented by the symbol P_n and is equal to $n!$. we have:

$$P_n = n!$$

For a better understanding, for example, suppose we want to put three people in a line together. The first person can be any of these three people. So the first person is selected in 3 ways. When one person is selected, the second person is selected in two ways and the third person is selected in one way. So, the total number of states in which these three people stand together in a row, according to the principle of multiplication, is:

3	2	1
---	---	---

$$P_3 = 3! = 6$$

Note: n people to $(n-1)!$ You can sit around the round table.

N gems (or n keys) can be made into $\frac{(n-1)!}{2}$ necklaces (or key chains).

Example 3-22: How many ways can 5 people be seated around the table?

The number of seating positions for 5 people = $(5-1)! = 4! = 24$
around the round table

If we want to replace n objects next to each other so that several specific objects are next to each other, we consider those several objects connected together and one. We use the following formula:

$$r! (n-r+1)!$$

Pay attention: if no order is mentioned in placing these specific objects, we multiply the permutation of these objects in the obtained answer.

Example 3-23: In how many possible ways can three distinct math books and four distinct literature books be placed together on a shelf, so that the math books are always next to each other?

We want to put three different math books and four different literature books together on a shelf so that the math books are next to each other. For this purpose, we tie math books together and consider them as one book. Now we have a math book and four distinct literary books that are permuted (placed together) in 5 ways. Since there is no mention of the order of the math books, we can change the order of the math books. In other words, replace three distinct math books together, which can be done in 3 ways. So finally, by multiplying these two states, the answer is equal to:

$$n = 7 \quad , \quad r = 3 \quad \quad 5! \times 3! = 120 \times 6 = 720$$

If we want to replace the members of two groups side by side, the number of members of the two groups should be equal or one of the groups should have one member more than the other. In the exchange, someone in the middle paid attention to the following points:

If the number of members of two groups is the same, we must assume that one of the members of the first group is at the head of the queue and once a member of the second group.

$$2 \times m! \times n!$$

2- If the number of members of one of the groups is one member more than the other, we start the queue with the group that has more members.

$$m! \times n!$$

Example 3-24: In how many ways can three different types of scientific books and four different types of literary books be placed together in a row so that the scientific books are placed one in the middle?

In order for scientific books to be placed together, only one literary book should be placed between them. So we have:

$$\begin{array}{cccccccc}
 \text{Scientific} & \text{Literary} & \text{Scientific} & \text{Literary} & \text{Scientific} & \text{Literary} & \text{Literary} & 3 \times 4 = 44 \\
 & & & & \text{or} & & & \\
 \text{Literary} & \text{Scientific} & \text{Literary} & \text{Scientific} & \text{Literary} & \text{Scientific} & \text{Literary} & 3 \times 4 = 44 \\
 & & & & \text{or} & & & \\
 \text{Literary} & \text{Literary} & \text{Scientific} & \text{Literary} & \text{Scientific} & \text{Literary} & \text{Scientific} & 3 \times 4 = 44
 \end{array}$$

Plural principle $3 \times 144 = 432$

If we want to put n distinct element categories that include $1n_1$ different elements from the first category and n_2 different elements from the second category and ... n_k different elements from the k th category,

there are many different ways that this is possible, from the following formula to It comes.

$$n_1 \times n_2 \times \dots \times n_k! \times k!$$

3.22. Permutation with repetition

Whenever we want to put n objects together so that n_1 objects are of the first type, n_2 objects are of the second type ... and n_k objects are of the k th type, the number of distinct permutations of these n objects is equal to:

$$\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$$

In other words, we first assume that n objects are distinct and determine the answer. Then we divide the obtained answer into permutations of repeated members.

$n-1$ permutations of n objects with repetition of objects or without repetition of objects are equal to its n permutations. In mathematical language, we have:

$$P(n, n-1) = P_n$$

If we want to replace m among n repeated objects, the number of cases of doing this can be calculated when first, with great patience and accuracy, we write all the m categories of those objects, and then in each of the categories, using We calculate the permutation relation of repetition of the number of states. In the end, because we have used "or" among the different states of the categories, we add the states together to determine the final answer.

Example 3-25: How many five-digit phone numbers can be made with the digits of the phone number "225755"?

Instead of counting the number of five-digit phone numbers, we calculate the same number of six-digit phone numbers. The phone number "225755" has six digits, three digits "5" and two digits "2" are repeated. So the number of distinct permutations of these six digits is equal to:

$$\frac{6!}{3! \times 2!} = \frac{6 \times 5 \times 4 \times 3!}{3! \times 2!} = \frac{120}{2} = 60$$

Example 3-26: How many four-digit numbers can be written with the digits 1,2,2,3,3,3?

Because we want the permutation of 4 of the 6 given repeated digits, we must first write down all the categories of 4 of the given digits and then calculate the number of states in each of the categories from the relation of permutation with repetition. we have:

Category 1:	{1, 3, 3, 3}	Number of Permutations	of	$\frac{4!}{3!} = 4$
Category 2:	{2, 3, 3, 3}			$\frac{4!}{3!} = 4$
			or	
Category 3:	{2, 2, 3, 3}			$\frac{4!}{2!2!} = \frac{24}{4} = 6$
			or	
Category 4:	{1, 2, 3, 3}			$\frac{4!}{2!} = \frac{24}{2} = 12$
			or	
Category 5:	{1, 2, 2, 3}			$\frac{4!}{2!} = \frac{24}{2} = 12$

We have the number of four-digit numbers according to the principle of addition

$$= 4 + 4 + 12 + 6 + 12 = 38$$

3.23. Circular permutation

In contrast to linear permutations, there is another concept called circular permutation. Circular permutation means an arrangement of placing objects around a circle. The difference between circular permutations and linear permutations is that the positions do not differ from each other and only the position of the objects relative to each other is important. Therefore, if we want to place the first object in the same position, since the positions do not have any difference, we do not have a particular choice.

The number of permutations of n distinct objects a_1, a_2, \dots , arranged on the circumference of a circle or any other closed surface is obtained from the following relation.

$$(n-1)!$$

3.24. Permutation (conversion or combination) of r objects from n objects

If we have n distinct objects, each of the states of placing r objects among n objects together, so that the order of placing objects in it is important, is called a permutation of r objects among n objects.

We represent the number of permutations of r objects among n objects (assuming $n \geq r$) with the symbol $P(n,r)$, which is equal to:

$$P(n,r) = P_n^r = (n)_r = \frac{n!}{(n-r)!}$$

The meaning of the arrangement of r of n objects is to form a group, each of which contains r distinct objects from these n objects, and the difference between each group and the other group is both in terms of the placement of objects and their type, each of these groups is a We call an order of r out of n objects.

The difference between order and permutation is that not all objects participate in each case.

The number of arrangements of r of n objects that do not have m special objects is obtained from the following formula.

$$P_{n-m}^r$$

The number of arrangements of r of n objects in all of which private objects are used is obtained from the following formula.

$$r \times P_{n-1}^{r-1}$$

Example 3-27: In how many ways can 4 people be placed in a row out of 6 people?

Because we want to place 4 out of 6 people in a row, so the order of placing people is important and actually the number of permutations of 4 out of 6 people is desired. Therefore, we have the formula for permuting r objects from n objects:

$$P(6, 4) = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{6 \times 5 \times 4 \times 3 \times 2!}{2!} = 360$$

3.25. Composition

If we have n distinct objects, each of the states of choosing r objects from n objects, so that the order of placing the selected objects does not matter, is called a combination of r objects among n objects.

Because with each selection of r objects from n objects, we can make $r!$ permutations of these n objects, as a result, the number of combinations of r objects among n objects (assuming $r \leq n$) which is denoted by the symbol $C(n, r)$ or $\binom{n}{r}$. We show that it is equal to:

permutation r object selects r object, permutation r object from n objects:

$$P(n,r) = C(n,r) \times r!$$

$$C(n,r) = C_n^r = \binom{n}{r} = \frac{n!}{(n-r)! r!}$$

If in solving the problem of the number of states, only choosing is relevant, then we use the topic of combination, but if in solving the problem of the number of states, choosing and placing (arranging) is relevant, we use the topic of permutation.

Example 3-28: In how many ways can 3 people be selected from among 8 employees of an office to be sent on a mission?

Since we choose only 3 out of 8 employees (and their order does not matter), we go to the formula of combining 6 objects out of 8 objects. we have:

$$C(8, 3) = \binom{8}{3} = \frac{8!}{(8-3)! 3!} = \frac{8 \times 7 \times 6 \times 5!}{6 \times 5!} = 56$$

The meaning of combining r of n objects is to form groups, each of which contains r objects, and the difference between the groups is only in the objects, not the way they are placed next to each other.

The difference between multiple combinations and multiple permutations (arrangements or multiple transformations) is that in multiple combinations the order of placing objects together is not

important, so in issues such as selecting multiple objects from different objects or mixing different colors to create new colors and In other words, the main problem in the problems related to permutation and combination is to distinguish whether the problem asks for the number of combinations or the number of permutations. It is better to first determine We check whether the change in the order of objects causes a change in the result or not, if the action causes a change, in this case, the number of permutations is asked in the problem, and if it does not affect, the number of combinations is of interest.

In many problems, combination and principle of multiplication and principle of addition are used together.

The following relationship exists between composition and order.

$$P(n, r) = r! C(n, r)$$

The following ties always hold.

$$C(0, n) + C(1, n) + C(2, n) + \dots + C(n, n) = 2^n$$

$$C(0, n) - C(1, n) + C(2, n) + \dots + (-1)^n C(n, n) = 0$$

The number of r combinations of n distinct objects that do not have m special objects is obtained from the following formula.

$$C(r, n-m)$$

The number of r combinations of n distinct objects that include m specific objects is obtained from the following formula.

$$C(r-m, n-m)$$

The number of r combinations of n distinct objects with the condition that each object can be repeated twice or three times or ... or r times, is obtained from the following formula.

$$C(r, n-1+r)$$

The number of ways that a set of n objects can be divided into k subsets with n_1 objects in the first set, n_2 objects in the second set, ... and n_k objects in the k_{th} set is equal to:

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! + n_2! + \dots + n_k!} \quad , n = n_1 + n_2 \dots + n_k$$

Consider the following 3 points to be able to separate such issues.

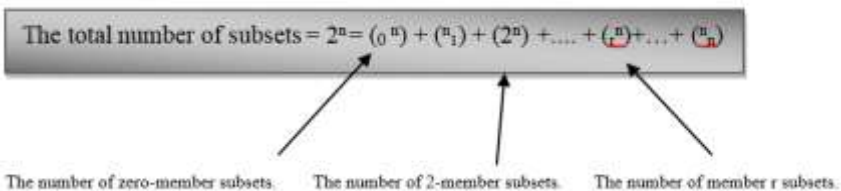
- Objects are distinct.
- Groups are different.
- It is known exactly how many members each group accepts.

The number of k -membered subsets of an n -membered set

As we know, moving members in sets does not change them. Now suppose a set contains n members. To form a subset of k members of this set, we must choose k members from among n members of this set and put them in a new set. So:

The number of sub k members of an n member set is equal to $\binom{n}{k}$

Now, since the total number of subsets of a set of n members is 2^n , we conclude that:



Example 3-29: The number of subsets of 4 members of the set {6,5,4,3,2,1} is equal to: $\binom{6}{4} = \frac{6 \times 5}{2} = 15$ subsets of 4 members

Example 3-30: If the number of 3-member subsets of a set is equal to 45, then the number of 2-member subsets is equal to:

The number of subsets of 2 members of a set (we assume this set has n members) is equal to 45. So:

$$\binom{n}{2} = 45 \rightarrow \frac{n(n-1)}{2} = 45 \rightarrow n(n-1) = 90 \rightarrow n = 10$$

Now, having the number of members of the set, the number of subsets of its 6 members is equal to:

The number of subsets of 3 members = $\frac{10!}{7!3!} = \frac{10 \times 9 \times 8 \times 7!}{7! \times 3 \times 2 \times 1} = 120 \binom{10}{3}$ members

3.26. Binomial probability distribution

Suppose we are faced with a random experiment that has two states of victory and failure. If we repeat this experiment n times and the results of each of these repetitions are independent of each other, the probability of winning (success) k times in these n repetitions is obtained from the following relationship (assume the probability of winning in one test is equal to p).

$$p^k (1-p)^{n-k}$$

If the probability of failure or victory in the experiment is equal to 0.5 (for example, in the coin toss experiment), the binomial

probability distribution formula for k times of success is calculated from the following equation.

$$P(x = k) = \frac{\binom{n}{k}}{2^n}$$

Example 3-31: In a factory, 60% of the workers are natives. If 4 of them are randomly selected, what is the true probability that 3 of them are natives?

If we consider being native as a success, the probability that among 4 people ($n = 4$), exactly 3 people ($k = 3$) are native, according to the binomial probability distribution formula is equal to:

$$P = \frac{60}{100} = \frac{6}{10} \quad 1 - P = \frac{4}{10}$$

$$N = 4, k = 3 \quad P(x = 3) = \binom{4}{3} \left(\frac{6}{10}\right)^3 \left(\frac{4}{10}\right)^1 = 4 \times \frac{216}{1000} \times \frac{4}{10} =$$

0.3456

Exercises of chapter 3

- 1- How many ways can 6 people sit around a round table?
- 2- There are 3 white and 2 black rabbits in cage A, 2 white and 4 black rabbits in cage B, and 2 white and 2 black rabbits in cage C. One rabbit is randomly selected from each cage. What is the probability of selecting at least two white rabbits?
- 3- In a region, the probability of rain on the first of December is 50%. The probability that it will rain on the second day is 40%. If it rains on the first day of December, what is the probability that it will rain on the second day?
- 4- 30% of the used tractors are made by factory A and 70% are made by factory B. The probability of technical defects of these two cars is 3 and 4% respectively. If a tractor is selected at random, what is the probability that it is broken?
- 5- From a group of 10 people, in how many ways can groups of 3 people be selected so that each group differs from the other groups in at least one person?
- 6- How many groups of 5 people including 3 plant breeders and 2 herbalists can be selected from 6 plant breeders and 5 herbalists?
- 7- From a family with three children, the eldest of whom is a boy. What is the probability of having at least one girl in this family?
- 8- Two red and white dice are thrown together; What is the probability that the white dice is an odd number and the red dice is a multiple of 3?
- 9- In how many ways can the flags of 4 European countries, 3 Asian countries and one African country be placed in a row, so that the flags of the Asian countries are next to each other?

The flags of three Asian countries count as one flag.

10- There are 5 red, 4 white and 6 black balls in a box. If we randomly draw 3 balls without replacement, what is the probability that one ball of each color is drawn?

There are six different combinations to remove the balls, and the probability of the six combinations is equal to:

11- If the probability of cloudiness is 0.3 and the probability of cloudiness and rain is 0.2, what is the probability of rain?

12- In the test of two balanced hexagons (dice), what is the probability that the sum of spots on two dice is at least 9?

13- 60% of the trees in a garden are modified and the rest are indigenous. 10% of native trees and 5% of modified trees are more than three meters high. If a tree is selected at random and it is more than three meters high, what is the probability that it is a modified variety?

14- Warehouse has 20 types of pens, 12 of which are healthy and the rest are unhealthy. A customer has selected 3 items. What is the probability that all 3 pens are of healthy?

15- There are 7 carnation branches in a vase, one of them is white and the others are red; There are also 5 evening primrose branches in this vase, 2 of them are white and the rest are red. A child is asked to choose a flower. What is the probability that it is a carnation or a red one?

16- Answer the following questions.

There are 4 blue pencils and 5 red pencils in a box. From this box, we choose 2 pencils consecutively without replacement.

a): What is the probability that both pencils are the same color?

b): What is the probability that one of the pencils is blue and the other is red?

c): What is the probability that only one of the pencils is blue?

d): What is the probability that the second pencil is only blue?

17- In a diverse plant population, 25% of the plants are short, 15% are late, and 10% are both short and late. A plant is chosen at random, what is the probability that it is late or short-legged?

18- We have five different math books, four different chemistry books and two different physics books. In how many ways can we arrange these books together as long as the books in each group stay together?

19- We have five different physics books, three different chemistry books and four different math books. How many ways can we put them together?

20- In how many ways can 10 people be divided into three groups of 3, 4 and 2 people?

21- Consider a balanced coin and an unbalanced coin (with probability equal to 0.6). If a coin is randomly tossed, what is the probability of seeing heads?

22- Consider four students, two of whom are brothers. How can these four students sit on the same bench if two brothers are supposed to be together?

23- Four men and three women enter the shop one by one, what is the probability that one of the people entered is a man and a woman?

24- Out of 7 math students and 5 experimental students, 4 students are randomly selected for a competition, what is the probability that at least 3 of them are math students?

- 25- In a sample space $P(A) = 0.4$ and $P(A \cup B) = 0.6$, for which value of $P(B)$ will the events of A and B be independent?
- 26- We throw a coin until it comes up heads for the third time. What is the number of states that can be reached in 10 times of throwing a coin?
- 27- Out of 742 students of a university, 241 failed in general mathematics, 271 failed in biology, and 111 failed in both courses. How many failed in none of the two mentioned courses?
- 28- In how many ways can 3 out of 5 math books and 4 out of 6 chemistry books be arranged in a shelf?
- 29- 40 students are enrolled in the statistics class, 30 of them are first year students and 10 are second year students, and half of the class students are girls. If a student is selected at random, what is the probability that this person is a freshman or a girl?
- 30- How many 4-letter words can be made with Motor letters?

CHAPTER 4

RANDOM VARIABLE, MATHEMATICAL EXPECTATION AND BINOMIAL DISTRIBUTIONS

Assist. Prof. Dr. Mohsen MIRZAPOUR¹

Dr. Saeid HEYDARZADEH²

¹ - Siirt University, Faculty of Agriculture, Department of Agricultural Biotechnology, Siirt, Türkiye

ORCID ID: 0000-0002-2898-6903, e-mail: m.mirzapour@siirt.edu.tr

² - Former Ph.D. Student of Urmia University, Faculty of Agriculture, Department of Plant Production and Genetics, Urmia, Iran

ORCID ID: 0000-0001-6051-7587, e-mail: s.heydarzadeh@urmia.ac.ir

4.1. Probability distribution

So that the probability of a variable is the relative frequency of that variable in the population. Therefore, the probability distribution of a random variable is obtained from the relative frequency distribution of the values of that variable in the population. If we have different values of a random variable like x_i , along with the corresponding probability of each value, this set will form a probability distribution of variable x_i . For example, in throwing 2 coins, 3 states (2 lions, 1 lion, 1 line, 2 lines) may occur, so the possible distribution of the number of heads in throwing 2 coins can be written as below.

X_i	0	1	2	
P_i	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\sum P_i = 1$

Distributions related to discrete random variables are called discrete probability distributions. Important discrete probability distributions are: uniform distribution, Bernoulli distribution, binomial distribution, polynomial distribution, geometric distribution, negative binomial distribution, Poisson distribution and hypergeometric distribution, among the probability distributions, there are three types of probability distribution including binomial, Poisson and normal distribution. which are widely used in data analysis.

4.2. Discrete uniform distribution

If a random experiment selects n possible outcomes x_1, x_2, \dots, x_n with equal probability $\frac{1}{n}$, the experiment is uniform.

4.3. Random variable and probability distribution function

Discrete uniform random variable assigns a distinct number to each of the discrete uniform test results. The probability function of a uniform distribution is defined as follows:

$$f_x(x) = \frac{1}{n} \quad x = x_1, x_2, \dots, x_n$$

The uniform distribution has a parameter n .

A uniform rectangular diagram is always a set of rectangles with equal height.

In general, n in the uniform distribution formula is calculated as $\binom{N}{K}$ when the total number of sample points is N and the number of points in the selected subset is k . that's mean:

$$n = \binom{N}{K}$$

Mathematical expectation, variance and moment generating function

$$E(x) = \frac{n+1}{2}$$

$$\text{Var}(x) = \frac{n^2-1}{12}$$

$$M_x(t) = \sum e^{tx} \cdot \frac{1}{n}$$

4.4. Bernoulli distribution

The random variable X_i that can take only 2 values 0 and 1 is called Bernoulli random variable. These results are called success and

failure. The probability of a lion is represented by p and the probability of a tail is represented by $q = p - 1$.

4.5. Random variable and probability function

In a Bernoulli experiment, the number of successes is called a Bernoulli random variable. The probability function of the Bernoulli distribution is defined as follows:

$$f_x(x) = p^x q^{1-x} \quad x = 0, 1$$

Bernoulli distribution has a parameter p .

The Bernoulli distribution probability function can also be shown as below.

$$f_x(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$

Mathematical expectation, variance and moment generating function

$$E(x) = p$$

$$\text{Var}(x) = pq$$

$$M_x(t) = (q + pe^t)^n$$

Example 4-1: The random quantity X is distributed according to Bernoulli's law with parameter $P = 0.75$, what is the generating function of the moments for the random variable X ?

$$P = 0.75 = \frac{3}{4} \quad q = 1 - \frac{3}{4} = \frac{1}{4} \quad M_x(t) = \frac{1}{4} + \frac{3}{4}e^t$$

4.6. Binomial distribution

If the probability of each event is known in advance and the events are incompatible and occur independently of each other, the probability of these variables can be calculated. The distribution of such probabilities is known as the binomial distribution. If we consider n Bernoulli experiments with probability of success P and probability of failure q in each experiment ($p+q=1$), then the number of success x is a random variable called binomial random variable. and has the following conditions:

- 1- Experiments should be repeated a certain number of times.
- 2- Experiments should be repeated independently of each other.
- 3- Each experiment leads to two possible outcomes of success and failure.
- 4- The success must be constant in all experiments.

Binomial random variable is the number of successes in a binomial experiment, in other words, it is a specific binomial distribution of events that have two states of occurrence (success) and non-occurrence (failure).

$$F_{(r)} = \frac{n!}{r!(n-r)!} P^r q^{n-r}$$

where r is the number of successful events, n is the number of people in the sample, p is the probability of the desired event, and $q=1-p$ is the probability of another event.

They also calculate the probability of the classes in the binomial distribution from the binomial expansion of $(p+q)^n$. n is the number of people in the sample. The binomial expansion with 3 examples will be as follows:

$$(p+q)^3 = p^3 + 3p^2q + q^3$$

The binomial $(p+q)^n$ contains $n + 1$ terms pq .

The power of p in the first sentence is equal to n and in the following sentences it is reduced by one unit each time until it reaches zero in the last sentence.

The sum of powers of p and q in each term must be equal to n .

The coefficient of each term is obtained using the combination of C_n^x .

$$\text{Sentence coefficient} = \frac{n!}{(p^{\text{power}})_!(q^{\text{power}})!}$$

In the binomial expansion $(p+q)^n$, when n is even (the number of terms is odd), the middle term has the highest coefficient, and the terms on its sides have equal coefficients and are symmetrical. In the case that n is odd (the number of sentences is even), the middle 2 sentences have the highest coefficients, and the sentences on their sides will have equal coefficients symmetrically.

Example 4-2: The expansion of the binomial distribution of this expression $(2x + y^2)^5$ is equal to:

$$(2x + y^2)^5 = (2x)^5 + C_5^1(2x)^4(y^2) + C_5^2(2x)^3(y^2)^2 + C_5^3(2x)^2(y^2)^3 + C_5^4(2x)(y^2)^4 + (y^2)^5$$

$$C_5^2(2x)^3(y^2)^2 = 10(8x^3)(y^4) = 80x^3y^4$$

Example 4-3: In a family of 6 children, the probability of this situation that the male children are less than the female children is equal to:

$P = P(4 \text{ girls and } 2 \text{ boys}) + P(5 \text{ girls and one boy}) + P(6 \text{ children are girls})$

$$P = \left(\frac{1}{2}\right)^6 + C_6^1 \left(\frac{1}{2}\right)^5 \frac{1}{2} + C_6^2 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2$$

$$P = \frac{1}{64} + \frac{6}{64} + \frac{15}{64} = \frac{22}{64}$$

This type of probability distribution is different from empirical distributions. The probability distribution is obtained through the known characteristics of the population. Fixed numbers such as p that characterize the theoretical (theoretical) distribution are called parameters. The value of p is obtained from the previously known information related to the mechanisms that control the events. Mendel's laws can be mentioned as an example in this context, by which the probability of an event can be determined in advance. If p is unknown, samples are taken from the population to estimate p . Usually, the estimation of p is indicated by p' . It should be noted that the resulting distribution in the latter case is an empirical distribution, not a theoretical one.

In the binomial distribution, the value of p and q can be equal $p = q = \frac{1}{2}$ or different from each other.

The equality of p and q is the condition of symmetry of binomial distribution, in other words, this distribution will be symmetric only if $p = q = \frac{1}{2}$.

In the binomial distribution, the mean is equal to np and the variance is equal to npq , so the variance and the mean are related to each other and are not independent of each other.

Binomial distribution is used in cases where p and q are almost close to each other (or equal) and n is not too large.

Example 4-4: In a binomial distribution, $\delta = 6$, $\mu = 144$. Is the value of n and p desirable?

$$\mu = np = 144 \qquad \delta = \sqrt{npq} = 6$$

$$\delta^2 = npq \qquad 36 = 144 \times q \qquad q = \frac{36}{144} = \frac{1}{4}$$

$$p = 1 - q = 1 - \frac{1}{4} = \frac{3}{4}$$

$$\mu = np \qquad 144 = n \frac{3}{4} \qquad n = 192$$

Mathematical expectation, variance and moment generating function

$$E(x) = np$$

$$\text{Var}(x) = npq$$

$$M_x(t) = (q + pe^t)^n$$

4.7. Negative binomial distribution (Pascal distribution)

The negative binomial test is a generalization of the Bernoulli test and has the following conditions:

- 1- Experiments should be repeated independently of each other.
- 2- Each experiment leads to two possible outcomes of success and failure.
- 3- The probability of success in all experiments is constant.
- 4- We repeat the experiments until we reach the k th success.

4.8. Random variable and probability function

In a negative binomial experiment, the random variable is the number of experiments until the k -th success, the probability that the k -th success occurs in the x -th experiment is obtained from the following

formula, which is called the probability function of the negative binomial distribution.

$$f_x(x) = \binom{x-1}{k-1} p^k q^{x-k} \quad x = k, k+1, \dots$$

Negative binomial probability distribution has two parameters p and k .

In the negative binomial distribution, if $k = 1$, the geometric distribution is obtained.

Mathematical expectation, variance and moment generating function

$$E(x) = \frac{k}{p}$$

$$\text{Var}(x) = \frac{kp}{p^2}$$

$$M_x(t) = \left(\frac{pe^t}{1-qe^t}\right)^k$$

Example 4-5: We toss a fair coin, if the random variable x represents the third tail in the seventh toss of the coin, calculate the mathematical expectation and variance of x .

$$E(x) = \frac{k}{p} = \frac{3}{\frac{1}{2}} = 6, \quad \text{Var}(x) = \frac{kp}{p^2} = \frac{3 \times \frac{1}{2}}{\frac{1}{4}} = 6$$

4.9. Polynomial distribution

If the experiment has more than two possible outcomes and the probability of each outcome is constant in different experiments and the experiments are independent of each other, the corresponding distribution is a polynomial distribution. The only difference between this distribution and the binomial distribution is that in the binomial

distribution there are two possible outcomes, but in this distribution there are more than two possible outcomes, the amount of factory production may be classified as excellent, good, average and poor.

4.10. Random variable and probability function

If the experiment is conducted n times independently and each experiment contains k distinct outcomes with constant probabilities p_1, p_2, \dots, p_k , then the probability of occurrence of outcome one is x_1 times and outcome two is x_2 times and x_k of consequence k has a polynomial distribution with the following probability:

$$f(x_1 = x_1 \cdot x_2 = x_2 \cdot x_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} \cdot p_1^{x_1} \cdot p_2^{x_2} \dots p_k^{x_k}$$

4.11. Poisson distribution

Poisson distribution is derived from binomial distribution. In other words, the random variable x that represents Poisson random events in a certain time or place is called Poisson random variable. In cases where n is large and p is very small in the binomial distribution so that the product of $n \times p$ is smaller than 5, then a distribution called Poisson distribution is used as an approximation of the binomial distribution. In other words, the Poisson distribution is equivalent to the binomial distribution.

$$F(r) = \frac{m^r \times e^{-m}}{r!}$$

where r, n and p are the same values in the binomial distribution. and e is equal to 2.718. Poisson distribution can be used especially in situations where n and p are unknown, but their product, i.e. m , is

available (in fact, m shows the mean and variance of the distribution), a distribution that, unlike the normal distribution, does not have a symmetrical state.

e is the NEP number and m is the mean of the Poisson distribution, which is equal to the product of np . So that this distribution is used for rare events and a rare event is an event for which the relation $np < 5$ is true. In Poisson distribution, mean and variance are equal and equal to m .

$$\mu = \sigma^2 = m = np$$

4.12. Poisson's test has the following conditions:

1- The probability of more than one incident occurring in a very small space or time interval is almost zero.

2- The probability of an incident occurring in any time interval or place is proportional to the length of that interval.

3- The probability of events in time intervals or places are independent of each other.

Mathematical expectation, variance and moment generating function

$$E(x) = np$$

$$\text{Var}(x) = np$$

$$M_x(t) = e^{np(e^t - 1)}$$

Example 4-6: The probability of a mutant genotype in a corn field is 5 in a thousand. If a sample consisting of 200 plants is selected, what is the probability of having 5 mutant genotypes in the selected sample?

$$P = 0.005, \quad n = 200$$

$$M = np = 200 \times 0.005 = 1$$

$$F = \frac{m^r \times e^{-m}}{r!} = \frac{(1)^5 e^{-1}}{5!} = \frac{0.368}{120} = 0.003$$

Example 4-7: The death rate of a disease is reported to be 4 per thousand. In a group of 350 people, the variance is equal to:

Because the current event is a rare event, it has a Poisson distribution and its variance is as follows:

$$\delta^2 = np = 350 \times 0.004 = 1.4$$

4.13. Normal distribution

Normal distribution is the most important type of continuous distribution in data analysis. This distribution was first studied by Gauss and it is also called Gauss distribution. The distribution of any variable that is caused by the accumulation of small incremental effects such as animal weight, grain yield and plant height can be justified by normal distribution. The normal probability distribution curve is expressed by the following formula:

$$F(x) = \frac{1}{\sqrt{2n\delta^2}} e^{-(x-\mu)^2/2\delta^2}$$

In cases where the number of "n" tests in a binomial distribution is very large, then the discontinuous binomial distribution becomes a continuous distribution, which is called a normal distribution

The equation of normal distribution is shown below.

$$x \sim N, \text{IND}, (\mu, \delta^2)$$

The normal distribution curve is a symmetric and bell-shaped curve and is proportional to the average of the population, where the mean, the mean and the median are equal, the total area under the curve is equal to one. The parameters μ and δ are the mean and standard deviation of this distribution, respectively, and the square of δ is called δ^2 , which means the variance of the normal population. Therefore, the location and shape of a normal curve are determined by two quantities μ and δ .

The normal distribution is a continuous distribution, in which the probability of an event is calculated in a continuous document, for example, the probability $P(X=7)$ is written as follows:

$$P(6.9 \leq X \leq 7.1)$$

In the normal distribution, the values of X vary between $+\infty$ and $-\infty$, while in the binomial distribution, the range of X changes is between 0 and n .

Calculation of probabilities related to a normal distribution is done by accessing standard normal tables. These tables are based on the standard normal distribution. The standard normal variable, which is also known as "standard normal deviation", has a normal distribution with a mean of zero and a variance of one. This variable is represented by the symbol z and is equal to:

$$Z = \frac{X - \mu}{\delta}$$

In the standard normal distribution, it is always necessary to remember the following divisions:

68% of Z_i is between -1 and +1.

90% of Z_i is between -1.64 and +1.64.

95% of Z_i is between -1.96 and +1.96.

99% of Z_i is between -2.58 and +2.58.

99.7% of Z_i is between the two numbers -3 and +3.

Example 4-8: Out of 100 students, how many male students at least and at most should be observed in order to be able to infer with a confidence factor of 95% that the probability of entering the university is the same for boys and girls?

$$P = \frac{1}{2} \quad q = 1-p = \frac{1}{2} \quad n = 100$$

$$\mu = np = 100 \times \frac{1}{2} = 50$$

$$\sigma = \sqrt{npq} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{25} = 5$$

$$Z = \frac{X-\mu}{\sigma}$$

$$\pm \frac{1}{96} = \frac{X-50}{5} \quad X = 40.60$$

4.14. Hypergeometric distribution

Suppose a community of N members consists of two parts: the first part contains K members and the second part contains $N-K$ members, we want to select a random sample consisting of n members from this community and see how many members of the sample have a certain characteristic. Such a test is called a hypergeometric test.

In the hypergeometric distribution, to calculate the probability of occurrence of x times of a specific event in n times of testing when the total number of states of that specific event is equal to k , we will have:

$$\text{for two cases: } \frac{C_k^x \times C_{n-k}^0}{C_n^x}$$

The mean and variance of the hypergeometric distribution are obtained as follows:

$$\mu = np \quad \sigma^2 = npq \left(\frac{N-n}{N-1} \right)$$

In this relationship, $P = \frac{K}{N}$, which means we determine the probability value before any choice is made, naturally, q will also be equal to $\frac{K}{N} - 1$.

N is the number of people in the community and n is the number of sample people.

In the case that N is large, the mean and variance of the hypergeometric distribution will be similar to the binomial distribution, that's why when a small sample is selected from a very large population and $\frac{n}{N}$ will be a very small number, so there is practically no difference between the sample. There will be no substitution and no substitution. Hypergeometric distribution has three parameters N , K and n .

The difference between binomial probability distribution and hypergeometric distribution is as follows:

1- In the binomial probability distribution, the population size is not given, i.e. N , but in the hypergeometric distribution, the population size is given, i.e. N , and it is usually not a large number.

2- In the binomial probability distribution, the probability of success is constant and equal to P every time the experiment is repeated, but the probability of success in the hypergeometric distribution changes every time.

Mathematical expectation and variance

$$E(x) = np = n \cdot \frac{k}{N}$$

$$\text{Var}(x) = npq \left(\frac{N-n}{N-1} \right)$$

Example 4-9: From a population that has 8 members and its variance is 42, all possible samples of 3 members have been selected and the average of the samples is available. The variance of the means is equal to:

$$\bar{\sigma}_x^2 = \frac{\sigma_x^2}{n} \left(\frac{N-n}{N-1} \right) \delta^2$$

$$\bar{\sigma}_x^2 = \frac{42}{3} \left(\frac{8-3}{8-1} \right) = 10 \delta^2$$

In the hypergeometric probability distribution, the community was divided into two parts. If the community is divided into several parts in the hypergeometric distribution, a more general form of the hypergeometric probability distribution is obtained. that's mean:

$$\begin{array}{ccccccc} N & k_1 & k_2 & \dots & K_t & & \\ n & X_1 & X_2 & \dots & X_t & & \end{array}$$

$$f_x(x_1, x_2, \dots, x_t) = \frac{\binom{K_1}{x_1} \cdot \binom{K_2}{x_2} \dots \binom{K_t}{x_t}}{\binom{N}{n}}$$

In which:

$$\sum_{i=1}^t x_i = n, \quad \sum_{i=1}^t k_i = N$$

4.15. Geometric distribution

As observed in binomial and hypergeometric distributions n is already known and their difference is that in binomial distribution p is

constant in all experiments but in hypergeometric distribution p varies from experiment to experiment. But in the situation where n is not known in advance, but p is constant in all experiments, the experiment is repeated until the first success is achieved.

$X=k$ is obtained from the following equation:

$$P(x: k) = pq^{k-1}$$

The geometric probability distribution has a parameter p .

In this relationship, p is the probability of success and q is the probability of failure. For example, we will have:

$$p(x=1) = p \qquad p(x=2) = pq \qquad p(x=3) = pq^2 \dots$$

The geometric test is a generalization of the Bernoulli test and has the following conditions:

- 1- Experiments should be repeated independently of each other.
- 2- Each experiment leads to two possible outcomes of success and failure.
- 3- The probability of success in all experiments is constant.
- 4- We repeat the experiments until we reach the first success.

Mathematical expectation, variance and moment generating function

$$E(x) = \frac{1}{p}$$

$$Var(x) = \frac{q}{p^2}$$

$$M_x(t) = \frac{pe^t}{1-qe^t}$$

Example 4-10: A person applies for a driver's license and passes the test with a probability of 0.8. If the random variable x indicates passing the third test. Calculate the arithmetic mean and variance of x .

$$E(x) = \frac{1}{p} = \frac{1}{0.8} = \frac{1}{25} \quad \text{Var}(x) = \frac{q}{p^2} = \frac{0.2}{(0.8)^2} = 0.3125$$

4.16. Mathematical expectation

The expected value of a random variable in the population is called the mathematical expectation of that variable. The mathematical expectation of the normal variable, x , is shown as:

$$E(X) = \mu$$

Example 4-11: There is a box containing 8 items, 2 of which are defective, we choose 3 items from the box. What is the mathematical probability of selecting defective items?

$$\text{Probability that none of them are defective} = \frac{6}{8} \times \frac{5}{7} \times \frac{4}{6} = \frac{5}{14}$$

$$\text{The probability that one is defective} = \frac{2}{8} \times \frac{6}{7} \times \frac{5}{6} = \frac{5}{28}$$

$$\text{The possibility that the first or second or third is defective} = 3 \times \frac{5}{28} = \frac{15}{28}$$

$$\text{The probability that two are defective} = \frac{2}{8} \times \frac{1}{7} \times \frac{6}{6} = \frac{1}{28}$$

$$\text{The possibility that the first two or the last two or the first and the last are defective} = 3 \times \frac{1}{28} = \frac{3}{28}$$

x	0	1	2
F(x)	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$

$$E(X) = \mu = \sum F_i x_i = 0 \times \left(\frac{5}{44}\right) + 1 \left(\frac{15}{28}\right) + 2 \left(\frac{3}{28}\right) = \frac{15}{28} + \frac{6}{28} = \frac{21}{28} = \frac{3}{4}$$

The mathematical expectation of X is the population mean, i.e. μ . To estimate μ , the average of n samples from a normal distribution can be calculated as follows:

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

The mathematical expectation of a random variable is equal to its mean, that is, it can be written:

$$E(X) = np$$

The variance of a random variable is defined as the mathematical expectation of the square of the deviation of that variable from its mathematical expectation. In relation to the normal variable, x, the variance formula by definition is equal to:

$$E\{X-E(X)\}^2 = \delta^2$$

$$E(X-\mu)^2 = \delta^2$$

δ^2 can be estimated through a random sample from the following formula:

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

The variance of the r variable of binomial and Poisson distribution is as follows:

$$\text{Binomial : } E\{r-E(r)\}^2 = npq$$

$$\text{Poisson : } E\{r-E(r)\}^2 = np$$

Mathematical expectation of linear combination of multiple random variables

Usually, the calculation of the mathematical expectation of a linear combination is expressed in terms of the characteristics of the primary random variables or the constituent components of that function:

The mathematical expectation of the difference and the sum of two or more random variables is equal to the difference and their sum.

$$E(X \pm Y) = E(X) \pm E(Y)$$

The mathematical expectation of a fixed number is equal to that fixed number itself.

$$E(C) = C$$

The mathematical expectation of the product of a fixed number and a random variable is as follows.

$$E(CX) = CE(X)$$

The mathematical expectation of a set of n while all x 's are taken from the same population is equal to:

$$E(\sum X) = \mu + \mu + \dots + \mu = n\mu$$

The average mathematical expectation of an n sample is equal to:

$$E(\bar{X}) = \mu$$

The mathematical expectation of the deviations of a random variable from its own mathematical expectation of the mean is equal to zero.

$$E(X_i - \mu) = 0 \quad E(X_i - E(x_i)) = 0$$

4.17. Variance of linear combination of several random variables

This variance can also be expressed in terms of the variance of the random variables that make up the function. Before dealing with the variance of different functions, the mathematical expectation of the product of the deviation of two random variables from the respective mean should be defined. If two variables are not independent, we will have:

$$E(X-\mu_X)(Y-\mu_Y) = \delta_{XY} = \text{Cov}(xy)$$

The quantity on the right is known as the covariance of x and y, and it shows the joint changes of the two variables. This covariance in a normal population is estimated through sampling and based on the following formula:

$$\delta_{xy} = S = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{n-1}$$

If two variables are independent, the covariance value will be zero.

$$E(X-\mu_X)(Y-\mu_Y) = 0$$

The total variance of several random variables with mean μ and variance δ^2 :

a) If the variables are independent

If the equation $D = X_1 + X_2$ is available, δ^2 can be calculated as follows:

$$\begin{aligned} \delta^2 &= E\{(X_1+X_2)-E(X_1+X_2)\}^2 \\ &= E(X_1+X_2-\mu-\mu)^2 \end{aligned}$$

$$= E\{(X_1 - \mu) + (X_2 - \mu)\}^2$$

$$= E(X_1 - \mu)^2 + E(X_2 - \mu)^2 + 2E(X_1 - \mu)(X_2 - \mu)$$

Because they are independent of each other, then the variance will be zero. As a result: the variance of several random variables is equal to the set of variance of the variables.

$$\delta^2 = \delta^2 + \delta^2 + 0 = 2\delta^2$$

b) In case of non-independence of variables:

We will have: equal to the sum of the variance of those two variables plus twice their covariance.

$$\delta^2 = \delta^2_X + \delta^2_Y + 2\delta_{XY}$$

The variance of the product of a random variable with a fixed number:

equal to the power of two fixed numbers multiplied by variable variance

$$d = CX_i$$

$$\delta^2 = E\{CX - E(CX)\}^2$$

$$= E\{CX - Ce(x)\}^2$$

$$= C^2 E\{X - E(X)\}^2$$

$$= C^2 E(X - \mu)^2$$

$$C^2 \delta^2 = \delta^2$$

4.18. How to calculate the variance of a linear combination?

To calculate a linear combination that is the sum or subtraction of several variables or several numbers, we proceed in the following order:

- 1- We remove the fixed values.
- 2- We raise the remainder of the functions to the power of two.
- 3- instead of the power of two variables δ and instead of their product we put Cov.
- 4- If x and y are independent variables, their Cov will be zero.
- 5- If x and y are from one community or from two communities with equal variance, they can be added together.

A) Assume that x 's are independent and random variables from a normal population with variance δ^2 . In this case, the variance of a function like d will be as follows:

$$\begin{aligned}
 D &= 2x_i + 3y_i + 10 \\
 &2x_i + 3y_i \\
 (2x_i + 3y_i)^2 &= 4x_i^2 + 9y_i^2 + 12 x_i y_i \\
 &4\delta^2 + 9\delta^2 + 0 \\
 \delta^2_d &= 13 \delta^2
 \end{aligned}$$

b) Average variance of a random sample:

$$\begin{aligned}
 \bar{x} &= \frac{\sum x_i}{n} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) \\
 1 &= \frac{1}{n} (x_1 + x_2 + \dots + x_n) \\
 2 &= \frac{1}{n^2} (x_1 + x_2 + \dots + x_n)^2 \\
 3 &= \frac{1}{n^2} (x_1^2 + x_2^2 + \dots + x_n^2 + 2x_1x_2 + 2x_2x_3 + \dots + 2x_n - \\
 &1x_n) \\
 4 &= \frac{1}{n^2} (\delta^2 + \delta^2 + \dots + \delta^2 + 0 + 0 + \dots + 0) = \frac{n\delta^2}{n^2} = \frac{\delta^2}{n}
 \end{aligned}$$

As a result, the average variance of a random sample from a normal population will be equal to the variance of the population divided by the number of people in the sample.

$$\delta^2_{\bar{x}} = \delta^2_{\bar{x}} = \frac{\delta^2}{n}$$

The variance of the difference between two independent averages, each with n observations:

So that the variance of the difference of two independent samples is the same as the sum of the variances of the two averages.

$$\bar{d} = \bar{x}_1 - \bar{x}_2$$

$$1 = \bar{x}_1 - \bar{x}_2$$

$$2 = \bar{x}_1^2 - \bar{x}_2^2 - 2 \bar{x}_1 \bar{x}_2$$

$$3 = \frac{\delta^2}{n} + \frac{\delta^2}{n} - 0$$

$$S^2_{\bar{d}} = \delta^2_{\bar{d}} = \frac{\delta^2}{n} + \frac{\delta^2}{n} = \frac{2\delta^2}{n}$$

while S^2 is the common variance of two samples and can be estimated from the following formula:

$$S^2 = \frac{S_1 + S_2}{2}$$

Therefore, the standard deviation of the difference between the two averages is as follows.

$$S_{\bar{d}} = \sqrt{\frac{2\delta^2}{n}} = \sqrt{S^2_{\bar{d}}}$$

If the number of people in two samples is different, $\delta^2_{\bar{d}}$ will be:

$$\delta^2_{\bar{d}} = \frac{\delta^2}{n_1} + \frac{\delta^2}{n_2} = \delta^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

The common variance is calculated through the following formula:

$$\delta^2 = S^2 = \frac{(n_1-1)\delta^2 + (n_2-1)\delta^2}{n_1+n_2-2}$$

Therefore, we can conclude that the unbiased estimate of δ^2 is obtained through a random sample from the formula $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$. That is, the mathematical expectation of the variance of the S^2 sample is representative of the population's δ^2 :

$$\begin{aligned} &= \frac{1}{n-1} E\{\sum_{i=1}^n (x_i - \bar{x})^2\} \\ &= \frac{1}{n-1} E\{\sum_{i=1}^n (x_i - \mu + \mu - \bar{x})^2\} \\ &= \frac{1}{n-1} E\{\sum_{i=1}^n (x_i - \mu) - (\bar{x} - \mu)^2\} \\ &= \frac{1}{n-1} E\{\sum_{i=1}^n (x_i - \mu)^2 + \sum (\bar{X} - \mu)^2 - 2\sum (\bar{x} - \mu)(\bar{X} - \mu)\} \\ &= \frac{1}{n-1} (n\delta^2 + \frac{n\delta^2}{n} - \frac{2n\delta^2}{n}) \\ &= \frac{1}{n-1} (n\delta^2 - \delta^2) \\ &= \delta^2 \left(\frac{1}{n-1}\right)(n-1) \\ &= \delta^2 \end{aligned}$$

Example 4-12: The variance of the combination $y = 2\bar{x} - \frac{3}{4}(\sum_{i=1}^n x_i)$ is equal to:

$$\begin{aligned} Y^2 &= 4\bar{x}^2 + \frac{9}{16} (\sum_{i=1}^n x_i)^2 \\ \delta^2_y &= 4\delta^2_{\bar{x}} + \frac{9}{16} n\delta_x^2 \\ \delta^2_y &= \frac{4\delta^2_{\bar{x}}}{n} + \frac{9n\delta^2_{\bar{x}}}{16} \end{aligned}$$

Example 4-13: The variance of the linear function $y = x_1 + 3\bar{x}_2 + 7$ is equal to:

$$Y = x_1 + 3\bar{x}_2 + 7$$

$$y^2 = x_1^2 + 9\bar{x}_2^2$$

$$\delta_y^2 = \delta_1^2 + \frac{9\delta_2^2}{n_2}$$

4.19. Sampling and estimation of parameters

In many applied fields, the goal of researchers is to determine the parameters of the population, but it is not possible to access them through statistical population counting. In such cases, a statistical sample is used to infer community parameters.

4.20. Random example

It is a random sample whose constituents are selected by chance and lottery from among the society. This means that the chance of all people to be a member of the sample should be equal and the people who make up it should have the characteristics of the people of the main society and it should remain constant in all stages of sampling. Any part of the society that is famous is called an example. The results obtained from such samples can be generalized to society. The more the number of people in the sample, the more similar it will be to the population, and the more reliable the generalization of results and statistical inference will be. Sampling is based on chance and lottery among people in the community.

In other words, the probability of choosing each person is an independent event compared to the probability of choosing or not choosing other people. Therefore, we can say: a random sample is a set of n independent quantities, x_1, x_2, \dots, x_n that has the same distribution as the original population.

An empirical law is true when it considers all possible aspects and cases. So that the complete components of these studies, i.e. the investigation of all aspects of the society, are not included in all cases due to time constraints and necessary costs. Therefore, according to statistical rules and principles, a part of the society was selected as a statistical sample and the survey was conducted in a limited way.

A community is a group of people who have at least one common trait. We have two types of society, limited and unlimited society. The society of the people of a city is limited. Theoretical and theoretical communities are unlimited in number.

4.21. Sampling methods

Simple random sampling

In this method, each element of the desired society has an equal chance to be selected. On the other hand, the selected people must have the same characteristics as the society from which they were selected. Two types of simple random sampling methods are as follows.

From a statistical population with N members, we want to select a simple random sample of size n :

1- If we return the first sample to the community after selecting and noting its size, and this procedure continues until the selection of

the n th member of the sample, it is known as simple random sampling with placement.

In this method, the probability of selecting each element of the sample from the statistical population is equal to $\frac{1}{N}$.

2- If the first sample is not returned to the community after selection and this procedure continues until the selection of the n th member of the sample, it is known as simple random sampling without replacement.

In this method, the probability of selecting the first member of the sample is $\frac{1}{N}$, the second member of the sample is $\frac{1}{N-1}$ and finally the n th member of the sample is $\frac{1}{N-n+1}$.

4.22. Regular sampling

In this method, sample elements are selected from the list of people or members of the statistical community prepared for this purpose. Therefore, with the determination of the first member of the sample, the other members of the sample will be determined. In this method, each element of the community has a known and non-zero chance of being selected in the sample, because we have chosen the starting point randomly.

The advantage of regular sampling method is that firstly: it is easy to do and secondly: it is not expensive. Regular sampling can be used for populations that have some order (such as employee numbers and student numbers).

Regular sampling method is a modified simple random sampling method.

4.23. Stratified sampling

In this method, they first classify the society and then select a random sample from each class. The reasons for using stratified sampling are that: First, we consider a suitable sample size for each stratum. Second: it provides a more accurate estimate of the community parameters than other sampling methods.

4.24. Cluster sampling

In this method, first clusters (groups) of community elements are selected and then the elements of each cluster form a part of the samples. The reason for using this method is mostly due to its low cost. In this method, the condition of community homogeneity plays an important role.

4.25. Parameter and statistics

1- The numerical characteristics of society are called parameters. Parameters are fixed and unique values, but they are usually unknown. For example, the population mean is a population parameter.

2- The numerical characteristics of the sample are called statistics. Statistics are random variables and its value depends on the selected sample. For example, if we take a sample from the community and obtain the average of the sample, this average is a statistic. The statistic is usually shown with one of the Latin letters and we put a sign on it

like \bar{X} , \tilde{X} , \hat{p} ... And we observed the values of x_1, x_2, \dots, x_n then we show the similar value of \bar{X} to \bar{x} .

4.26. 1. Some useful statistics

If x_1, x_2, \dots, x_n which are not necessarily distinct, represent a random sample of n , then the mean of the sample is equal to:

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

2- If x_1, x_2, \dots, x_n are not necessarily distinct, the display is a sample of n , then the mode or face is the value of the sample that happened the most. Fashion may not exist, and when it does, it is not necessarily unique.

3- If x_1, x_2, \dots, x_n represents a random sample of n , arranged in ascending order, then the mean of the sample is equal to the middle number if n is odd, or equal to the arithmetic mean of the two middle numbers if n is even.

4- If x_1, x_2, \dots, x_n represents a random sample of n , then the variance (difference) of the sample will be calculated from the following formula.

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

The sample standard deviation, represented by the S statistic, is the positive square root of the sample variance.

In order to estimate the community parameters μ and δ^2_x , first a random sample of the community is selected and then the community parameters are estimated based on the quantities obtained from the

"statistical" sample. And in this way, the average of the society is estimated. The accuracy of this statistical estimate depends on the following factors.

1- The sample should be representative of the mentioned society and at the same time it should be large enough.

2- The estimate must be unbiased.

In terms of mathematical statistics, an estimate is called unbiased if its mathematical expectation is equal to the population parameter. In mathematical language:

$$E(t) = \Theta$$

$$\text{Community parameter} = (\text{theta}) \Theta$$

$$\text{Estimated } \Theta = t$$

$E(t)$ = the mathematical expectation of t , and it means that if we perform an infinite number of random sampling operations (with replacement) and each time an estimate of Θ is obtained, (t_1, t_2, \dots, t_n), the average of these estimates, or in other words, the Mathematical expectation of this estimate is equal to the main number of the community, i.e. Θ . If $E(t) \neq \Theta$, t is called a biased estimate of Θ .

In order for the average of a randomly selected attribute to be generalized to the society as a good estimate of the desired attribute, the following relationship must be established.

$$E(\bar{X}) = \mu$$

3- The variance of the estimates should be small.

If the estimate obtained from sample i is denoted by t_i , the difference between the estimate and the population parameter ($t_i - \Theta$) is called error. Mathematical expectation $(t_i - \Theta)^2$, which we denote by E

$(t_i - \Theta)^2$, means the average of the squared powers of the deviations of $t_i - \Theta$, which by definition is equal to the variance of t .

$$\delta_t^2 = E(t_i - \Theta)^2$$

Therefore, the smaller this variation is, the closer it is to the actual value of the garment and the fewer errors there are.

The variance of limited communities or very large samples can be estimated by the following formula.

$$\delta_x^2 = \frac{\sum(x_i - \mu)^2}{N}$$

To calculate the variance of the unlimited population, we select a random sample from that population. We estimate the desired variance with δ_x^2 . Variance estimation should be accurate. The variance of a small sample is less than the variance of a large community because as we know, the limits of community changes are greater than the limits of sample changes. Therefore, it is expected that δ_x^2 is also greater than S_x^2 .

If all n samples of possible members are taken from the unlimited study population and the variance of each is calculated and we denote them by S_i^2 , because the different values of S_i^2 are close to each other, so the average variances or Their Mathematical expectation is equal to:

$$E(S_i^2) = \frac{n-1}{n} S_i^2$$

Example 4-14: If a random sample with 5 members is selected from a population of 1000 individuals and the mean and variance of this sample are estimated to be 20 and 25 respectively, the true variance of the population is equal to:

$$E(S_i^2) = \frac{n-1}{n} s_i^2$$

$$5 = \frac{4}{5} s_i^2 s_i^2 = \frac{5 \times 25}{4} = 31.25$$

4.27. Frequency distribution of means

If from a statistical population consisting of N members, with mean μ and variance δ^2 , we randomly select all n possible samples and measure attribute x in them, the sum of the averages of these samples ($\bar{x}_1, \bar{x}_2, \dots \dots \bar{x}_n$) forms a random variable that will have a frequency distribution with the following characteristics.

1- If the frequency distribution of the population is normal, it is proved that the frequency distribution of the averages will also be normal. This is also true for sampling from populations with non-normal distribution. In general, for large samples ($n \geq 30$), the frequency distribution of the averages is almost normal.

If we set the population with unknown frequency distribution and mean μ and variance δ^2 as the initial population, and randomly take n samples from it, if the number of samples is pushed to infinity, the population that is from the average These samples are formed, the distribution will be normal, its mean will be μ and its variance will be $\frac{\delta^2}{n}$.

$$\delta^2_{\bar{x}} = \frac{\delta^2_x}{n}$$

Example 4-15: From the population consisting of 1000 members, all possible samples of 7 members have been selected. If the variance of the sample means is equal to 4, what is the standard deviation of the population?

$$\delta^2 = \frac{\delta_{\bar{x}}^2}{n} = \frac{2}{2} = 1\delta$$

$$\delta^2 = \frac{\delta_{\bar{x}}^2}{n} \quad \delta_{\bar{x}}^2 = 4 \times 7 = 28\delta$$

$$\delta_x = \sqrt{\delta_{\bar{x}}^2} = \sqrt{28} = 5.29\delta$$

In general, if all n possible samples are extracted from a finite population of n without replacement, it is proved that:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$\delta_{\bar{X}} = \frac{\delta_{\bar{x}}^2}{n} \left(\frac{N-n}{N-1} \right)^2$$

If the above population is unlimited and sampling is done with replacement, the value of $\frac{N-n}{N-1}$ will be close to one.

Example 4-16: From a community consisting of 10 members, all 7 samples have been selected. If the variance of the averages is 4, how much is the variance of the population:

$$\delta_{\bar{x}}^2 = 4, n = 7, N = 10$$

$$\bar{x}^2 = \frac{\delta_{\bar{x}}^2}{n} \left(\frac{N-n}{N-1} \right) \delta$$

$$4 = \frac{\delta_{\bar{x}}^2}{7} \left(\frac{10-7}{10-1} \right)$$

$$\delta_{\bar{x}}^2 = 84$$

If the variance of the initial population is not known, we calculate its estimate by choosing a random sample.

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

Because it is not possible to select all samples and calculate their average variance, we use the following formula.

$$S_{\bar{X}}^2 = \frac{S_x^2}{n}$$

In the community distribution of averages, the criterion score is calculated using the following formula.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\delta_{\bar{X}}} = \frac{\bar{X} - \mu}{\delta_{\bar{X}}}$$

Also, the variable Z will have a mean of zero and a standard deviation of one.

The standard deviation of the distribution of means is called standard error or standard error of the mean. The standard error is used in statistical inferences, and the standard error does not require repeated sampling, calculating the average of the samples and finally calculating their variance. The standard error has an inverse relationship with the number of sample members. Therefore, the larger the sample, the smaller the standard error, or in other words, the higher the accuracy of statistical judgments.

Example 4-17: Suppose a sample of 16 members is randomly extracted from a population with a variance of 64 and its mean is estimated to be 15. It shows the Z value (standard number) calculated for testing the null hypothesis $H_0: \mu \leq 10$ against the one hypothesis

$$H_1: \mu > 10?$$

$$\delta_{\bar{X}} = \sqrt{\frac{\delta_x^2}{n}} = \sqrt{\frac{64}{16}} = 2$$

$$p(z) = \frac{\bar{X} - \mu_{\bar{X}}}{\delta_{\bar{X}}} = \frac{10 - 15}{2} = -2.5$$

$$|z| = 2.5$$

4.28. Sampling distribution

The probability function of a statistic is a function that is obtained based on n random samples that are repeatedly selected from the statistical population. This function is called the sampling distribution of the statistic. For example, we call the distribution of the \bar{x} statistic the sample distribution of the mean, and the distribution of the S^2 statistic the sample distribution of the variance.

4.29. Sampling distribution of the mean (distribution of \bar{x})

Sampling distribution is the average of a distribution obtained by using the averages calculated from samples of a certain size obtained from a population.

If all possible random samples of size n are selected from a population with mean μ_x and standard deviation δ_x , then the sampling distribution of the mean has the following properties:

The average of the samples is equal to the average of the population, that is:

$$\text{For infinite communities } \mu_x = \mu_{\bar{x}}$$

$$\text{For finite communities } \mu_x = \mu_{\bar{x}}$$

2- The standard deviation of the average samples or the standard error is equal to:

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} \quad \text{For infinite communities}$$

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} \times \sqrt{\frac{N-n}{n-1}} \quad \text{For finite communities}$$

The expression $\sqrt{\frac{N-n}{n-1}}$ is called the finite Friday correction factor. When $n \leq 0.05N$, we assume that the population is finite and this coefficient is equal to 1 in the calculation of the standard error.

4.30. Cases of normality of \bar{x} distribution

If the community distribution is normal, then for samples with any given volume, the distribution of \bar{x} is also normal. If the community distribution is not normal (indeterminate), if the sample size is greater than 30 ($n > 30$), the distribution of \bar{x} is almost normal.

4.31. Sampling error

It should be known that if we choose different samples at random, we will realize that not all of them have the same characteristics. The reason for this is two things.

- 1- Sampling error
- 2- Huge difference between subjects

If we select different and relatively large samples from a community and calculate their average together, the calculated averages will be different. Some of them may be much higher and some may be lower and some of them may be in the middle. This phenomenon is justified by the central limit theorem.

4.32. Central limit theorem

The central limit theorem refers to the point that if we select several samples with equal size from a population, the distribution of

the averages of these samples in a specific variable such as height will be normal and the average of the selected averages is almost equal to the average of the population that the sample is from. It is selected.

If there is a difference between the averages of the two experimental groups, this difference is attributed to the sampling error before the test. The most important reason for sampling error is the non-identical and different characteristics of the samples.

Sampling error is not necessarily the result of sampling error, but may be due to differences between subjects.

If the number tends to infinity in the central limit theorem, the standard error becomes zero, the distribution of the sample mean becomes normal, and t becomes z .

If simple random samples are selected from a population with unknown frequency distribution and mean μ and variance δ^2 , as the sample size increases ($n \rightarrow \infty$), the distribution of \bar{x} tends towards the normal distribution, with mean μ and variance $\frac{\delta^2}{n}$ is as a result:

$$\begin{aligned}x_1, x_2, \dots, x_n & \quad \tilde{N}(\mu, \delta^2) \\ \bar{x}_1, \bar{x}_2, \dots, \bar{x}_n & \quad \tilde{N}\left(\mu, \frac{\delta^2}{n}\right)\end{aligned}$$

The sampling error of the mean follows the following principles:

The sampling error of the expected mean is equal to zero.

Sampling error is inversely related to sample size.

The sampling error is directly related to the standard deviation.

The distribution of sampling errors is normal.

It should be known that the standard deviation (standard) of the theoretical distribution of averages is an index that is measured by the sampling error, and it is called the standard error of the average or the standard error.

By using the standard or standard error of the mean and normal distribution, it is possible to determine the area of rejection or confirmation of the null hypothesis.

The standard error of the mean is calculated using the following formula.

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

$S_{\bar{x}}$ = standard error of the mean

S = standard deviation

n = group or sample size

Using the \bar{x} distribution

If a random sample of n is selected from a large or infinite population with mean μ_x and standard deviation δ_x , then the sample distribution of the mean will have an approximate normal distribution with mean $\mu_x = \mu_{\bar{x}}$ and standard deviation $\frac{\delta_x}{\sqrt{n}} = \delta_{\bar{x}}$ Was. as a result:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\delta_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\delta_x / \sqrt{n}}$$

4.33. Sampling distribution of mean difference (distribution $\bar{x}_1 - \bar{x}_2$)

The sampling distribution of the difference between the averages of two samples is a distribution obtained by using the difference of

averages calculated from random samples obtained from two populations. If n_1 and n_2 independent samples are taken from two large or infinite populations with means μ_1 and μ_2 and variances δ_1^2 and δ_2^2 , then the sampling distribution of the difference of the means ($\bar{x}_1 - \bar{x}_2$) will have approximately a normal distribution with the following parameters .

$$\mu(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

$$\delta(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}$$

And:

$$Z = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}}$$

Example 4-18: Television picture lamps produced by manufacturer A have a mean life span of 6.5 years and a standard deviation of 0.9 years, while producer B has a mean life span of 6 years and a standard deviation of 0.8 years. What is the probability that a random sample of 36 bulbs from factory A will have an average lifetime that is at least one year longer than the average lifetime of a sample of 49 bulbs from factory B?

$$\mu(\bar{x}_1 - \bar{x}_2) = 6.6 - 5 = 0.5$$

$$\delta(\bar{x}_1 - \bar{x}_2) = \frac{0.81}{36} + \frac{0.64}{49} = 0.035 \quad \delta(\bar{x}_1 - \bar{x}_2) = 0.189$$

$$Z = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_1 - \mu_2)}{\delta(\bar{x}_2 - \bar{x}_1)} = \frac{1 - 0.5}{0.189} = 2.646$$

$$P(\bar{x}_2 \geq 1) = P(Z \geq 2.4646) = 1 - P(Z < 2.4646) = 1 - 0.9959 = 0.0041$$

$$P(\bar{x}_1 -$$

4.34. Sample distribution of ratios (\bar{P} distribution)

Sampling distribution of a sample proportion is a distribution obtained by using proportions calculated from random samples of a certain size obtained from a population.

Suppose we represent the people of a society who have a certain trait with P and the value of P is unknown. In this case, to draw a conclusion about the real value of P , we first obtain a sample of size n from the selection population and the proportion of people who have the desired characteristic in this sample. The obtained ratio is called a sample ratio and is denoted by \bar{P} .

$$\bar{P} = \frac{X}{n}$$

where X is the number of individuals with the desired characteristic, or the number of specified successes in the sample, and n is the sample size. In this case, the distribution of \bar{P} , It has the following features:

1- The average ratio of the samples is equal to the real ratio of the population. that's mean:

$$\mu_{\bar{p}} = p$$

2- The standard deviation of the proportions of the samples is equal to:

$$\delta_{\bar{p}} = \sqrt{\frac{pq}{n}}$$

Application of \bar{P} distribution

Suppose n samples are selected from a population where the true proportion of the desired characteristic is equal to p . In this case, with

the condition $5 > np$ and $nq > 5$, the distribution of \bar{P} will be almost normal distribution with $p = \mu_{\bar{p}}$ and $\delta_{\bar{p}} = \sqrt{\frac{pq}{n}}$. as a result:

$$Z = \frac{\bar{P} - \mu_{\bar{p}}}{\delta_{\bar{p}}} = \frac{\bar{P} - p}{\sqrt{\frac{pq}{n}}}$$

Example 4-19: 80% of the employees of the Agricultural Jihad Organization of West Azerbaijan province are men. We select a random sample of size $n = 100$. Calculate the probability that at least 70% of the people in the sample are male?

$$n=100 \quad p=0.8 \quad q=0.2 \quad P(\bar{P} \geq 0.7) = ?$$

$$\mu_{\bar{p}} = 0.8, \delta_{\bar{p}} = \sqrt{\frac{0.8 \times 0.2}{100}} = \sqrt{0.0016}$$

$$Z = \frac{\bar{P} - \mu_{\bar{p}}}{\delta_{\bar{p}}} = \frac{0.7 - 0.8}{\sqrt{0.0016}} = -2.5$$

$$P(\bar{P} \geq 0.7) = P(Z \geq -2.5) = 1 - 0.0049 = 0.9951$$

4.35. Sample distribution of the difference of two ratios ($\bar{P}_1 - \bar{P}_2$ distribution)

The sampling distribution of the difference between the proportions of two samples is a distribution that is obtained by using the difference of proportions calculated from random samples from two populations.

If all possible random samples of sizes n_1 and n_2 are selected from two populations with assumed ratios of P_1 and P_2 , then the

sampling distribution of the difference in ratios ($\bar{P}_1 - \bar{P}_2$) will have a normal distribution with the following parameters.

$$\mu(\bar{p}_1 - \bar{p}_2) = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Sampling distribution of variance (S^2 distribution)

The sampling distribution of the variance is a distribution obtained by using the variances calculated from samples of a certain size obtained from a population.

If from a normal population with mean μ and variance δ^2 , independent samples of size n are selected and sample variance:

$$S^2 = \frac{\sum(\bar{x}_1 - \bar{x}_2)}{n-1}$$

calculate for each sample, the values of the S^2 statistic, the sample variance, are obtained. The distribution of the S^2 statistic is not known to us, but the following theorem defines the distribution of the S^2 statistic.

If S^2 is the variance of a random sample of size n from a normal population with variance S^2 , then the random variable:

$$\chi^2 = \frac{(n-1)S^2}{\delta^2}$$

The variance of the chi-square distribution is $(n-1)$ degrees of freedom. This statistic varies from sample to sample as S^2 varies.

Example 4-20: A company that manufactures campaign batteries guarantees that its batteries will last an average of 3 years with a standard deviation of 1 year. If 5 of these batteries have a lifetime of

1.2, 2.8, 1.3, 3.4 and 4.5. Can the manufacturer still be sure that the standard deviation of his batteries is one year?

$$S^2 = \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)} = \frac{(5 \times 53.67) - (15.9)^2}{5 \times 4} = 0.777$$

$$\chi^2 = \frac{(n-1)S^2}{\delta^2} = \frac{4 \times 0.777}{1} = 3.108$$

Considering that 95% of the χ^2 values with 4 degrees of freedom are between 0.484 and 11.43, so the value calculated with δ^2 is acceptable and correct.

4.36. Distribution of ratio of sample variances

In some problems, we need to estimate the ratio $\frac{\delta_1^2}{\delta_2^2}$, in such problems, we should compare the variances of two societies with each other. $\frac{S_1^2}{S_2^2}$ statistic is used for comparison. But the distribution of this statistic is not known to us. To check the distribution of the statistic $\frac{S_1^2}{S_2^2}$ from the statistic $(\frac{S_1^2/\delta_1^2}{S_2^2/\delta_2^2})$ which has an F distribution with The degrees of freedom are $r_1 = n_1 - 1$ and $r_2 = n_2 - 1$, we use

4.37. Determination of community parameters

The purpose of statistical studies is to generalize sample results to society and infer its parameters. Determining the parameters of society, which is the ultimate goal of the research, is possible in two ways:

1- Census of all elements of society and parameter calculation, in which case descriptive statistical techniques will be used.

2- Estimating or estimating the parameters according to the sample index values and with the methods that are discussed in inferential statistics.

Statistical inference theory, which is newly known as decision theory, includes "estimation" and "hypothesis testing". If the research is of a question type and only contains questions about the parameter, "estimation" is used to answer the questions, and if it contains hypotheses and has passed the question stage, "hypothesis testing" and statistical techniques are used.

The size of the parameters obtained by using the measurements obtained from sampling is called statistical estimation. Any kind of estimation starts from the selected statistic and its sampling distribution. In general, there are two types of estimation: point estimation and interval estimation.

4.38. Point estimate

In point estimation, the value of the desired index in the population is calculated from the sample, and in fact, in this way, an estimate of the index related to the statistical population is obtained.

For each parameter, there are different estimators, the best of which should be selected to infer the parameter. If Θ is a true and unknown parameter of the community, its estimator, represented by $\hat{\Theta}$, must have the following characteristics.

- To be an unbiased (unbiased) estimator

If the average of the delta feather exactly matches its true value, delta, the delta feather is called an unbiased (unbiased) estimator. that's mean:

$$E(\hat{\Theta}) = \Theta$$

If $E(\hat{\Theta})$ is not equal to Θ , the estimator of $\hat{\Theta}$ is called biased. In fact, biased is defined as this difference. that's mean:

$$\text{biased } E(\hat{\Theta}) \neq \Theta$$

$\hat{\Theta}$ may be larger or smaller than the actual value of Θ , in which case the skew value can be positive or negative.

In a normal population, the sample mean and sample median are both unbiased estimates of the population mean.

$S^2 = \frac{\sum(x_i - \bar{x})^2}{n}$ will estimate the variance of the community less than it actually is (biased estimator), but if $(n-1)$ is used instead of n in the denominator, the relationship $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ is obtained, which is an unbiased estimator for the population variance.

4.39. Efficiency (minimum variance)

The estimation efficiency of women is measured by their variance. That is, an estimate that has less variance (dispersion) is more efficient or more efficient. The reason we prefer the efficient estimator is that it is more likely to obtain an estimate within a certain distance of the population parameter. Define the relative efficiency of two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ as follows.

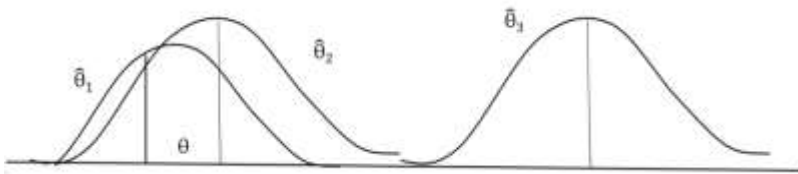
$$\text{relative efficiency of } \hat{\Theta}_1 \text{ compared to } \hat{\Theta}_2 = \frac{\hat{\Theta}_2^2}{\hat{\Theta}_1^2}$$

An estimator that is more efficient than any other estimator is called efficient.

In a normal population, the sample mean and sample median are both estimators unbiased from the population mean. But because the variance of the sample mean is less than the variance of the sample median, so the sample mean is more efficient than the sample median.

4.40. Minimum Mean Square Error (MSE)

Suppose we want to compare a biased estimator with an unbiased estimator.



The estimator $\hat{\Theta}_3$ has a low variance, but it is not satisfactory because it is highly biased. On the other hand, the estimator $\hat{\Theta}_1$ is unbiased, but it does not seem satisfactory due to its high variance. But the estimator $\hat{\Theta}_2$ seems to be better overall than $\hat{\Theta}_1$ and $\hat{\Theta}_3$ because it has the best combination in terms of small biased.

A measure that can take into account both biased and variance at the same time is the mean squared error (MSE), which is:

$$MSE = E(\hat{\Theta} - \Theta)^2 = E(\hat{\Theta})^2 - \Theta^2$$

In other words, it proves that:

$$MSE = \sigma_{\theta}^2 + (\text{biased})^2$$

After the two estimators, the one with lower MSE is better

The smaller the skewness value, the MSE will be closer to $\delta_{\hat{\theta}}^2$. So that if the value of skewness is equal to zero, then MSE is the same as $\delta_{\hat{\theta}}^2$. If two estimators have equal variance, the one with less skewness is better.

By using MSE, the efficiency of two estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ "either biased or unbiased" can be obtained as follows.

$$\text{Relative efficiency of } \hat{\Theta}_1 \text{ compared to } \hat{\Theta}_2 = \frac{MSE(\hat{\Theta}_2)}{MSE(\hat{\Theta}_1)}$$

Because the above relationship has less bias and less variance, it is the most important criterion for judging estimators.

Example 4-21: If $\hat{\Theta}_1 = \frac{X_1+X_2}{4}$, $\hat{\Theta}_2 = \frac{4X_1+X_2}{3}$ are two estimators, calculate the efficiency of $\hat{\Theta}_1$ relative to $\hat{\Theta}_2$.

$$\delta_{\hat{\Theta}_1}^2 = \text{Var}(\hat{\Theta}_1) = \text{Var}\left(\frac{X_1+X_2}{4}\right) = \frac{\text{Var}(X_1)+\text{Var}(X_2)}{4} = \frac{\delta^2+\delta^2}{4} = \frac{1}{4}\delta^2$$

$$\begin{aligned} \delta_{\hat{\Theta}_2}^2 &= \text{Var}(\hat{\Theta}_2) = \text{Var}\left(\frac{4X_1+X_2}{3}\right) = \\ &= \frac{16\text{Var}(X_1)+\text{Var}(X_2)}{9} = \frac{16\delta^2+\delta^2}{9} = \frac{17}{9}\delta^2 \end{aligned}$$

$$= \frac{68}{9} = 7.55 \quad \frac{\hat{\Theta}_2^2}{\hat{\Theta}_1^2} = \frac{\frac{17}{9}\delta^2}{\frac{1}{4}\delta^2}$$

Relative efficiency of $\hat{\Theta}_1$ compared to $\hat{\Theta}_2$

4.41. Compatibility (sustainability)

It is a consistent estimator that the variance of the sampling distribution tends to zero when the sample size tends to infinity ($n \rightarrow \infty$). Therefore, the estimator $\hat{\Theta}$ is consistent if its variance tends to zero when $n \rightarrow \infty$, that is:

$$\lim_{n \rightarrow \infty} (\hat{\Theta}) \text{Var} = 0$$

If $\lim_{n \rightarrow \infty} (\hat{\Theta}_2) \text{Var} < \lim_{n \rightarrow \infty} (\hat{\Theta}_1) \text{Var}$, then $\hat{\Theta}_1$ is asymptotically more efficient than $\hat{\Theta}_2$.

4.42. Distance estimation

The magnitude of the point estimation error is that we do not know the error of the estimator, because its value will change with the change of the sample, so they do not have high reliability, as a result, interval estimates are used. Interval estimation is done by the confidence interval method and consists of determining two values and specifying the probability that the unknown parameter of the society is located in that interval. The special advantage of the distance estimate over the point estimate is that the distance estimate is reliable, while the point estimate cannot be trusted. In general, when a community parameter is estimated by sampling, the estimated value is not completely equal to the true value of the parameter, and there is uncertainty in the generalization of the results. Therefore, we should consider the two numbers L and L and judge that the true value of the parameter is between those two numbers. Obviously, this judgment

may be wrong. The values of L and L are the "confidence limits" and the distance between them is the "confidence interval". The large number that makes the upper limit of the interval is called the "upper limit of confidence L " and the small number that makes the lower limit of the interval is called the "lower limit of confidence L ". If the percentage of possible error in judgment is equal to a ($a < 10$), the percentage of accuracy of judgment (confidence coefficient) will be equal to $(1 - a)$, in which case the distance L^- and L is called the confidence interval $(1 - a)$ It is called percentage. In general, if we calculate the confidence interval $(1 - a)$ percent for the parameter Θ , it means that if we calculate a large number of L and L with the same method, the distance between these two limits will be $(a-1)$ percent of the times. It includes the real Θ , and on the contrary, it does not include the real value of Θ in a percentage of times.

At a constant confidence level (confidence coefficient), a good confidence interval is an interval that is "correct" with a smaller length. The smaller the length of the estimate containing the parameter, the higher the "accuracy". What increases accuracy and precision in a confidence interval is a larger sample.

4.43. Interval estimation for the population mean

Distance estimation of μ_x parameter is the estimation of the distance in which the unknown average of the population is expected to be located. The distance estimate μ_x is defined as follows:

$$e \pm \bar{x}$$

e is a constant value and is called the precision (marginal error) of the estimate. The interval estimation of μ_x or the value of e is influenced by the confidence level and the distribution of \bar{x} . The distribution of \bar{x} is determined by these conditions.

1- The type of distribution of the statistical population (normal or non-normal)

2- The quality of society's standard deviation (known or unknown)

3- Sample size (small or large)

Combinations of the above conditions will bring different relations for the distance estimation \bar{x} of the phenomenon, which will be explained in order.

Normal statistical population distribution and known ∂_x

If the population is normally distributed and ∂_x is known, the normal distribution is used to set the confidence interval. In this case, the sample size is of any size, the \bar{X} distribution is a normal distribution and will be standardized as follows:

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\partial_{\bar{x}}}$$

We had that $\mu_{\bar{x}} = \mu_x$ and $\partial_{\bar{x}} = \frac{\partial_x}{\sqrt{n}}$, so we will have:

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\partial_x}{\sqrt{n}}}$$

On the other hand, according to the value of \bar{x} , the sign of Z may be positive or negative.

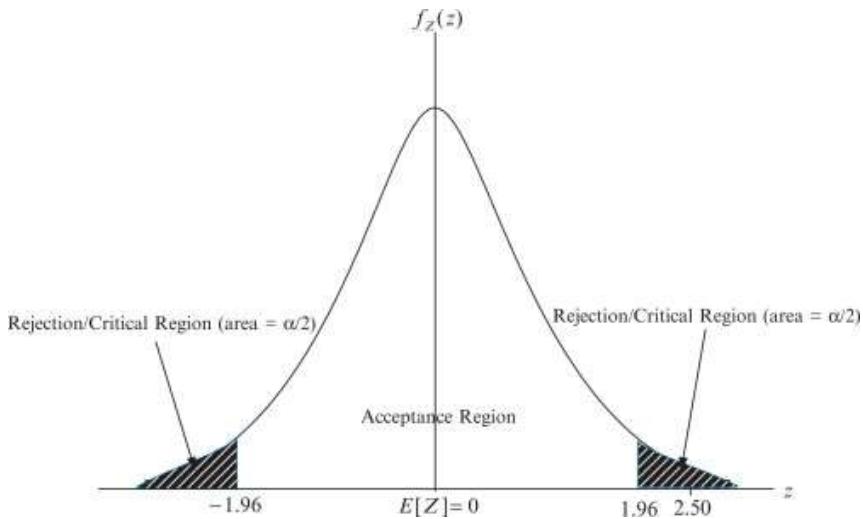
If we choose a random sample of size n from a normal population with a known standard deviation σ_x , the confidence interval $(1 - \alpha)$ percentage for μ_x is as follows:

$$\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} < \mu_x < \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$$

The above formula is shown below.

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$$

where \bar{x} is the mean value of the random sample and $Z_{\frac{\alpha}{2}}$ is the value of the standard normal variable, which is the area under the curve on the right side of $\frac{\alpha}{2}$.



If the sample size (n) and confidence factor $(1 - \alpha)$ remain constant, the interval length is always constant and equal to $2Z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$.

To calculate the confidence interval $(1 - \alpha)$ percent for μ_x , we assumed that σ_x is known. Therefore, in general, if we come across cases where σ_x

is unknown, then instead of σ_x , we use the standard deviation of the sample S_x , provided that $n > 30$. It means that we use the following relationship to estimate the distance μ_x .

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}}$$

Usually, we use confidence coefficients of 90%, 95% and 99%. Therefore, you can remember the following table.

$1 - \alpha$	α	$Z_{\frac{\alpha}{2}}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

95% is the usual choice. Because it seems to provide a good balance between precision (as reflected in the width of the confidence interval) and reliability (as expressed by the degree of confidence).

Example 4-22: A company manufacturing light bulbs for agricultural machinery produces light bulbs whose life spans are normally distributed with a standard deviation of 50 hours. If a random sample of 25 has a mean life of 500 hours, find a 96% confidence interval for the population mean of all light bulbs produced by this company.

$$\partial_x = 50 \quad , \quad n = 25 \quad , \quad \bar{x} = 500$$

$$1 - a = 0.96 \quad a = 0.04 \quad , \quad Z_{\frac{a}{2}} = Z_{0.02} = 2.05$$

$$\bar{x} \pm Z_{\frac{a}{2}} \frac{\partial_x}{\sqrt{n}} = 500 \pm \left(2.05 \times \frac{50}{\sqrt{25}} \right) = 500 \pm 20.5$$

$$= (479.5, 520.5) = (480, 520)$$

4.44. Normal statistical population distribution and unknown ∂_x

If the population is normally distributed and ∂_x is unknown, the t distribution is used to set the confidence interval, which is a similar distribution and has less accuracy (more dispersion) than the normal distribution. So, in this case, \bar{x} is a symmetric distribution that has more dispersion due to the unknown ∂_x . In distance estimation μ_x , when the standard deviation of the population is unknown, $S_{\bar{x}}$ will be replaced by $\partial_{\bar{x}}$. Therefore, the variable:

$$t = \frac{\bar{x} - \mu_x}{S_{\bar{x}}}$$

It is no longer normal but has a distribution called t-Student.

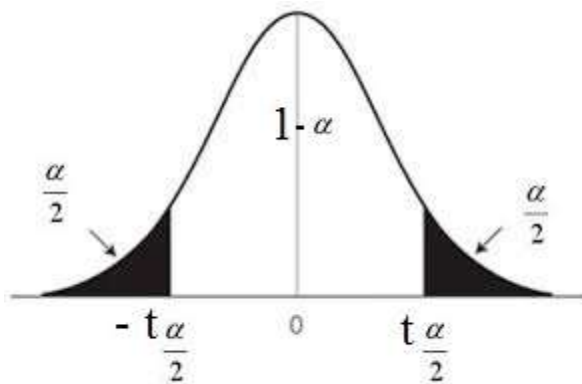
If we choose a random sample of size n from the normal population with unknown standard deviation, the confidence interval (1-a) percent for μ_x is as follows:

$$\bar{x} - t_{\frac{a}{2}, n-1} \frac{S_x}{\sqrt{n}} < \mu_x < \bar{x} + t_{\frac{a}{2}, n-1} \frac{S_x}{\sqrt{n}}$$

The above formula is shown below.

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

where \bar{x} and S_x are respectively the mean and standard deviation of a sample $n < 30$ of from a normal population and $t_{\frac{\alpha}{2}}$ is the value of the t distribution with $(n-1)$ degrees of freedom with an area equal to $\frac{\alpha}{2}$ has left.



Example 4-23: A random sample of 20 from a normal distribution has mean $\bar{x} = 40$ and standard deviation $S = 4$. You obtained a 95% confidence interval for μ .

$$1 - \alpha = 0.95 \quad \alpha = 0.05 \quad \therefore t_{\frac{\alpha}{2}, n-1} = t_{0.025, 19} = 2.093$$

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S_x}{\sqrt{n}} = 40 \pm \left(2.093 \times \frac{4}{\sqrt{20}} \right) = 40 \pm 1.77 = (38.23, 41.77)$$

4.45. Abnormal statistical population distribution

If the sample size is small ($n < 30$) and the population is not normally distributed, Chi Besheff theorem is used to adjust the

confidence interval. According to this theorem, the probability of the sample being within k standard deviations, $\partial_{\bar{x}}$, is equal to:

$$P(|\bar{x} - \mu_x| \leq k \cdot \partial_{\bar{x}}) \geq 1 - \frac{1}{k^2}$$

To make a confidence interval, first, $1 - \frac{1}{k^2}$ is set as the desired degree of confidence and K is obtained.

$$1 - \frac{1}{k^2} = 1 - a \qquad K = \frac{1}{\sqrt{a}}$$

Then the following relationships are used in the confidence interval based on whether ∂_x is known or unknown.

$$\bar{x} \pm k \frac{\partial_x}{\sqrt{n}}$$

and or:

$$\bar{x} \pm k \frac{S_x}{\sqrt{n}}$$

Example 4-24: If we select a sample of size $n = 20$ from a non-normal population. $\bar{x} = 50$ and $S = 5$ are obtained. It is desirable to obtain a 90% confidence interval for the population mean.

$$1 - a = 0.9 \qquad a = 0.1 \qquad K = \sqrt{\frac{1}{a}} = \sqrt{\frac{1}{0.1}} = 3.16$$

$$\bar{x} \pm k \frac{S_x}{\sqrt{n}} = 50 \pm \left(3.16 \times \frac{5}{\sqrt{20}} \right) = 50 \pm 3.53 = (46.48, 53.53)$$

4.46. Determining the sample size to estimate the population mean

The confidence interval for μ_x was:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\partial_x}{\sqrt{n}}$$

Certainly, the smaller the term that comes after \pm , the higher the accuracy of the confidence interval (the confidence interval is smaller). We call this sentence the limit error and we display it with the letter e . So we have:

$$e = Z_{\frac{\alpha}{2}} \frac{\partial_x}{\sqrt{n}}$$

If we raise the above sides to the power of two, we have:

$$e^2 = Z_{\frac{\alpha}{2}}^2 \frac{\partial_x^2}{n}$$

If we calculate n from this relationship, we will have:

$$e^2 = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \partial_x}{e} \right)^2$$

It can be seen that the larger ∂_x is, the larger sample size is needed. In the above relation, we do not have the value of ∂_x , so ∂_x must be estimated. The most common methods of estimating ∂_x are:

Based on past studies and experiences, find out the value of ∂_x .

Let's choose a preliminary sample, use it to calculate the standard deviation of the population and use it as an approximation for ∂_x .

If there is evidence that the sampled population is approximately normally distributed, we can use the fact that the width of the range is

approximately equal to 4 or 6 standard deviations and $\partial_x = \frac{R}{4}$ or $\partial_x = \frac{R}{6}$ will be R is the range of variation.

Example 4-25: To estimate the mean of a population, what is the sample size so that the 99% margin of error is $\frac{1}{6}$ of the standard deviation of the population?

$$e = \frac{1}{6} \partial_x \quad , 1-a = 0.99 \quad a = 0.01 \quad Z_{\frac{a}{2}} = 2.575$$

$$n = \left(\frac{Z_{\frac{a}{2}} \partial_x}{e} \right)^2 = \left(\frac{2.575 \times \partial_x}{\frac{1}{6} \partial_x} \right)^2 = (2.575 \times 6)^2 = 238.70$$

4.47. Interval estimation of the difference between the average of two communities

Usually, to compare two statistical populations, we compare the average of the two populations with each other, or in other words, we pay attention to the difference between the two averages. The average difference of two communities is $\mu_1 - \mu_2$. But instead of the average of the statistical communities, the sample average of n of them is available, so the comparison of two statistical communities should be done using the average of the samples.

If μ_1 and ∂_1^2 are the mean and variance of the first population, respectively, and μ_2 and ∂_2^2 are the mean and variance of the second population, respectively, then n_1 and n_2 are two random samples from the first and second population, respectively. It is assumed that these samples are independent from each other, each of these samples have

statistics \bar{x}_1 and s_1^2 for the first population and \bar{x}_2 and s_2^2 for the second population.

In order to make a confidence interval for the difference in the mean of two populations, it is necessary that the probability distributions of the random variable \bar{x}_1 and \bar{x}_2 are independent of each other, then:

$$\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$

that's mean:

$$\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$$

And:

$$\partial^2_{(\bar{x}_1 - \bar{x}_2)} = \partial^2_{\bar{x}_1} - \partial^2_{\bar{x}_2}$$

that's mean:

$$\partial^2_{(\bar{x}_1 - \bar{x}_2)} = \frac{\partial_1^2}{n_1} + \frac{\partial_2^2}{n_2}$$

Therefore:

$$\partial_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\partial_1^2}{n_1} + \frac{\partial_2^2}{n_2}}$$

The distance estimate $\mu_1 - \mu_2$ is defined as follows.

$$(\bar{x}_1 - \bar{x}_2) \pm e$$

The value of e regardless of the confidence level depends on these conditions:

The type of statistical distribution of the two sampled populations (normal or abnormal)

The quality of the standard deviation of the two sampled communities (known or unknown, equal or unequal)

Degree of freedom value, $(n_1 + n_2 - 2)$, (large or small)

Combinations of the above conditions will create different relationships for distance estimation $\mu_1 - \mu_2$, which will be explained respectively.

The distribution of two normal statistical populations and σ_2^2 and σ_1^2 are known

If \bar{x}_1 and \bar{x}_2 are the averages of two independent random samples of size n_1 and n_2 from normal communities, which have mean μ_1 and μ_2 and variance σ_1^2 and σ_2^2 , then $\bar{x}_1 - \bar{x}_2$ is a random variable with normal distribution, so the standard variable Z can be defined as follows:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The percentage confidence interval $(1 - \alpha)$ for $\mu_1 - \mu_2$ is as follows.

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where \bar{x}_1 and \bar{x}_2 are the averages of random samples n_1 and n_2 of communities with variance σ_1^2 and σ_2^2 and Z is the value of the standard variable whose level $\frac{\alpha}{2}$ is on the right side.

To interpret $\mu_1 - \mu_2$, we use the estimate length sign:

If both amplitudes are positive, μ_1 is greater than μ_2 at the desired confidence level.

If both amplitudes are negative, μ_1 is smaller than μ_2 at the desired confidence level.

Otherwise, there is no significant difference between μ_1 and μ_2 .

Example 4-26: A random sample of $n_1 = 20$ from a normal distribution with a standard deviation of $\sigma_1 = 4$ has a mean of $\bar{x}_1 = 50$. Another random sample of $n_2 = 25$ from another normal population with a standard deviation of $\sigma_2 = 5$ has a mean of $\bar{x}_2 = 40$. It is desirable to determine the 96% confidence interval for the population mean difference.

$$1 - \alpha = 0.96 \quad \alpha = 0.04 \quad \left\{ Z_{\frac{\alpha}{2}} = Z_{0.02} = 2.05 \right.$$

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (40 - 50) \pm 2.05 = \sqrt{\frac{16}{20} + \frac{25}{25}} = 10 \pm 2.75 = (7.25, 12.75)$$

4.48. The distribution of two statistical populations is normal and $\sigma_1^2 = \sigma_2^2$ but unknown

In cases where $\sigma_1^2 = \sigma_2^2$ is unknown and n_1 and n_2 are also small, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is a normal variable We will have the governor as follows:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

that ∂^2 should be estimated by integrating the sample variances, S_1^2 and S_2^2 , which is called integrated S^2 or S_p^2 and obtained from the following relationship comes.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

With this standard variable description, it will change like this:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has t with $(n_1 + n_2 - 2)$ degrees of freedom.

The percentage confidence interval $(1 - a)$ for $\mu_1 - \mu_2$ is as follows.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{a}{2}} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where \bar{x}_1 and \bar{x}_2 are the means of selected independent samples from normal distributions, which are n_1 and n_2 samples, respectively, and S_p is the selected standard deviation, and $t_{\frac{a}{2}}$ is the value of the t distribution, the number of degrees of freedom of which is $(n_1 + n_2 - 2)$ and leaves an area equal to $\frac{a}{2}$ on the right side of the curve.

Example 4-27: Two random samples $n_1 = 9$ and $n_2 = 16$ from two independent normal populations are given with $\bar{x}_1 = 25$ and $\bar{x}_2 = 20$ and $S_1 = 4$ and $S_2 = 3$. Obtain a 95% confidence interval for $\mu_1 - \mu_2$ assuming $\partial_1 = \partial_2$.

$$1 - a = 0.95 \quad a = 0.05 \quad t_{n_1 + n_2 - 2} = t_{0.025, 23} = 2.069$$

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(9-1)(4)^2 + (16-1)(3)^2}{9+16-2} = 11.43$$

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \cdot S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= (25 - 20) \pm 2.069 \times 3.38 \sqrt{\frac{1}{9} + \frac{1}{16}} \\ &= 5 \pm 2.91 = (2.09, 7.91) \end{aligned}$$

4.49. The distribution of two statistical populations is normal and $\sigma_1^2 \neq \sigma_2^2$ but unknown

In cases where σ_1^2 and σ_2^2 are unknown and n_1 and n_2 are also small. If $\sigma_1^2 \neq \sigma_2^2$ then $S_{(\bar{x}_1 - \bar{x}_2)}$ is defined as:

$$S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

The distribution $(\bar{x}_1 - \bar{x}_2)$ has a t distribution, but in terms of definition, we denote it by t' , which consists of:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The percentage confidence interval $(1 - \alpha)$ for $\mu_1 - \mu_2$ is as follows:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where \bar{x}_1 and \bar{x}_2 mean S_1^2 and S_2^2 are small sample variances of size n_1 and n_2 which are selected from normal communities and $t_{\frac{\alpha}{2}}$ is a value of t distribution with degrees of freedom under:

$$r = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

which leaves a surface of size $\frac{a}{2}$ on its right side.

Example 4-28: The following information is available on the monthly production of two paste factories.

$$n_1 = 16, S_1 = 4, \bar{x}_1 = 20 \quad : \text{Factory A}$$

$$n_2 = 9, S_2 = 2, \bar{x}_2 = 15 \quad : \text{Factory B}$$

Assuming that our observations are from normal populations with different variances, calculate a 95% confidence interval for the difference in the true means of paste production in these two factories.

$$r = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left(\frac{16}{16} + \frac{4}{9}\right)^2}{\frac{\left(\frac{16}{16}\right)^2}{16-1} + \frac{\left(\frac{4}{9}\right)^2}{9-1}} = 22.60$$

$$1 - a = 0.95 \quad a = 0.05 \quad t_{n_1+n_2-2} = t_{0.025, 23} = 2.069$$

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{a}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} &= (20 - 15) \pm 2.069 \times \sqrt{\frac{16}{16} + \frac{4}{9}} \\ &= 5 \pm 2.26 = (2.73, 7.26) \end{aligned}$$

4.50. Estimate the confidence interval for the population proportion

The ratio of successes in society is:

$$P = \frac{x}{N}, \quad q = 1 - p$$

The sampling distribution \bar{P} in large samples has a normal approximation and its standard variable is as follows:

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where $P = \mu_{\bar{p}}$ and $S_{\bar{P}} = \sqrt{\frac{pq}{n}}$

Therefore, the percentage confidence interval $(1 - a)$ for parameter P is as follows:

$$\bar{P} \pm Z_{\frac{a}{2}} \sqrt{\frac{pq}{n}}$$

where \bar{P} is the success ratio in n random sample and $\bar{q} = 1 - \bar{P}$ and $Z_{\frac{a}{2}}$ is a value of the standard normal distribution.

4.51. Determining the sample size to estimate the population proportion

The confidence interval for p was:

$$\bar{P} \pm Z_{\frac{a}{2}} \sqrt{\frac{pq}{n}}$$

Certainly, the smaller the sentence after \pm is, the higher the confidence accuracy (the confidence interval is smaller). We call this statement the limit error and we display it with the letter e, so we have:

$$e = Z_{\frac{a}{2}} \sqrt{\frac{pq}{n}}$$

If we raise the sides of the above relationship to the power of two, we have:

$$e^2 = Z_{\frac{\alpha}{2}}^2 \cdot \frac{\bar{p}\bar{q}}{n}$$

If we calculate n from this relationship, we will have:

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \cdot \bar{p}\bar{q}}{e^2}$$

Therefore, it is necessary to have the value of \bar{p} and \bar{q} . But it is not possible to know these values before sampling. As a result, we need to make a guess for p in the original population.

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \cdot pq}{e^2}$$

Since when $q = p = \frac{1}{2}$, the fraction will have its maximum value, in cases where a better estimate for p cannot be obtained, it can be taken equal to 50% and calculate n, because If $p = \frac{1}{2}$, then in the above formula, n will find its maximum value.

Example 4-29: A study to introduce a new variety of wheat is under research. Research shows that the ratio is not greater than 0.8. Also, the estimation accuracy is 0.06, calculate the sample size at the 1% error level.

$$1 - \alpha = 0.99 \quad \alpha = 0.01 \quad Z_{\frac{\alpha}{2}} = 2.575$$

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \cdot pq}{e^2} = \frac{(2.575)^2(0.8)(0.2)}{0.06^2} = 297$$

4.52. Interval estimation of the difference in success rate in two societies

If the data collected from two statistical populations are of a qualitative type, the comparison of the ratio of the two statistical populations is used. If we consider P_1 as the success ratio in the first society and P_2 as the success ratio in the second society, $\bar{p}_1 = \frac{x_1}{n_1}$ and $\bar{p}_2 = \frac{x_2}{n_2}$ are its statistics, respectively. For sufficiently large n_1 and n_2 and independent random samples, it can be accepted that the statistic $\bar{p}_1 - \bar{p}_2$ will be an unbiased estimator of $p_1 - p_2$ with the lowest variance. we know.

$$\mu_{(\bar{p}_1 - \bar{p}_2)} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = p_1 - p_2$$

And:

$$\sigma^2_{(\bar{p}_1 - \bar{p}_2)} = \sigma_{\bar{p}_1}^2 + \sigma_{\bar{p}_2}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

Therefore, the statistic $\bar{p}_1 - \bar{p}_2$ has a normal distribution, which is standardized as follows:

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{(\bar{p}_1 - \bar{p}_2)}}$$

Therefore, the percentage confidence interval $(1 - \alpha)$ for the difference of parameters $p_1 - p_2$ is approximately as follows:

$$(\bar{p}_1 - \bar{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1 \bar{q}_1}{n_1} + \frac{\bar{p}_2 \bar{q}_2}{n_2}}$$

where p_1 and p_2 are the success rates in random samples of n_1 and n_2 and $\bar{q}_1 = 1 - \bar{P}_1$ and $\bar{q}_2 = 1 - \bar{P}_2$ and $Z_{\frac{\alpha}{2}}$ is a value of the standard normal distribution which is a level for a It leaves $\frac{\alpha}{2}$ on its right side.

4.53. Interval estimation for community variance

The point estimate of community variance, i.e. ∂^2_x , is dependent on the sample variance s^2_x . Usually confidence limits for ∂^2_x based on sampling distribution:

$$\chi^2 = \frac{(n-1)S_X^2}{\partial^2_x}$$

If a sample of n is selected from the population with a normal distribution, the said value is called Chi-2 distribution with $(n-1)$ degrees of freedom.

Based on this, the confidence interval $(1 - \alpha)$ is a percentage for the variance ∂^2_x of a normal population as follows:

$$\frac{(n-1)S_X^2}{\chi^2_{\frac{\alpha}{2}}} < \partial^2_x < \frac{(n-1)S_X^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

where S_X^2 is the variance of n random sample and $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$ are values of chi-square distribution with $(n-1)$ degree is the freedom that leaves $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ levels on its right, respectively.

4.54. Interval estimation of the variance ratio of two statistical populations

The point estimate of the variance ratio of two populations, i.e. $\frac{\partial_1^2}{\partial_2^2}$, is given by the ratio of sample variances i.e. $\frac{s_1^2}{s_2^2}$. Hence, $\frac{s_1^2}{s_2^2}$ is the estimator of $\frac{\partial_1^2}{\partial_2^2}$. If ∂_1^2 and ∂_2^2 are normal population variances, we can construct an interval estimate for $\frac{\partial_1^2}{\partial_2^2}$ by the following statistic.

$$F = \frac{\frac{S_1^2}{\partial_1^2}}{\frac{S_2^2}{\partial_2^2}} = \frac{S_1^2 \partial_2^2}{S_2^2 \partial_1^2}$$

The random variable F has an F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. Therefore, the percentage confidence interval $(1 - a)$ for the ratio $\frac{\partial_1^2}{\partial_2^2}$ is as follows:

$$\frac{S_1^2}{S_2^2} \cdot f_{1-\frac{a}{2}}(r_1, r_2) < \frac{\partial_1^2}{\partial_2^2} < \frac{S_1^2}{S_2^2} \cdot f_{\frac{a}{2}}(r_1, r_2)$$

where s_1^2 and s_2^2 are respectively the variance of n_1 and n_2 independent samples selected from normal populations. $f_{\frac{a}{2}}(r_1, r_2)$ is a value of the F distribution with $r_1 = n_1 - 1$ and $r_2 = n_2 - 1$ degrees of freedom, on the right side of which there is a surface of the size of $\frac{a}{2}$.

Example 4-30: We have selected two independent random samples from two paste factories. The results are calculated as follows. Calculate a 90% confidence interval for $\frac{\partial_1^2}{\partial_2^2}$.

$$S_1 = 2 \quad n_1 = 12$$

$$S_2 = 4 \quad n_2 = 12$$

$$r_2 = 12 - 1 = 11 \text{ and } r_1 = 12 - 1 = 11. \quad f_{0.05,11,11} = 2.82$$

$$1 - a = 0.9 \Rightarrow a = 0.1 \text{ and}$$

$$\frac{S_1^2}{S_2^2} \cdot f_{1-\frac{a}{2}}(r_1, r_2) < \frac{\partial_1^2}{\partial_2^2} < \frac{S_1^2}{S_2^2} \cdot f_{\frac{a}{2}}(r_1, r_2)$$

$$\frac{2^2}{4^2} \cdot \frac{1}{2.82} < \frac{\partial_1^2}{\partial_2^2} < \frac{2^2}{4^2} \cdot 2.82 \quad 0.109 < \frac{\partial_1^2}{\partial_2^2} < 0.705$$

Exercises of chapter 4

1- If there is a normal distribution in a class and the average of the class is equal to 12 and its standard deviation is equal to 4, in this class, 2.5% of the people at the top of the class have at least what grade?

2- In a frequency distribution of scores of 200 candidates, the mean and standard deviation were 40 and 15, respectively. Later, we realize that we mistakenly counted the score 43 as 53. What will be the correct mean and standard deviation?

3- If the grades in a class have a normal distribution with a mean of 12 and a variance of 16, what is the maximum value of the bottom 50% of the class?

4- What is the mathematical expectation of the following equation?

$$Y_i = \sum X_i + 5 \quad S_x^2 = 10 \quad \mu_x = 6 \quad n = 8$$

5- Scores of students in a math test have a distribution of $N(60, 25)$. The professor gives an A grade to the students whose score is more than 60. How many students get an A in a class of 200 students?

6- The average and standard deviation of the height of 1000 wheat plants are 110 and 25 cm, respectively. The standard score of the average height of 16 random plants is -2. What is the average height of these 16 plants?

7- The weight of tea North packages has a normal distribution with a standard deviation of 8 grams. If 5% of these packages weigh more than 350 grams, what is the average of these packages?

$$P(x \geq 350) = 5\% \quad Z = 1.64$$

8- A dice is thrown 20 times. What is the variance of the number of times a number greater than 4 appears?

9- In a binomial distribution, $\delta=6$, $\mu=144$. What is the desired value of n and p ?

10- If in a frequency distribution the mean is 5.5 and the standard deviation is 1, how many percent of the data are between 5.5 and 7.5 (if $P_Z \geq 2 = 0.0228$)?

11- The average and standard deviation of the height of 2000 wheat plants are 115 and 26 cm respectively, the standard score of the average height of 20 random plants was -2. What is the average height of these 20 plants?

12- If the average of the first population is equal to 3 and the average of the second population is equal to 8, the mathematical expectation of the opposite function is equal to:

$$Y = 2\bar{x}_1 + 3\bar{x}_2 - 3$$

13- The Mathematical expectation of the following table is equal to:

X_i	10	20	30
P_i	1.5	2.5	3.5

14- The diameter of sunflower varieties has a normal distribution with a standard deviation of 2 cm. If 5% of the floors have a diameter greater than 23.28 cm, the average distribution of the diameter of the floors is equal to?

15- The average score of a test was 11.5 and its standard deviation was 5. It is desirable to calculate the lowest score belonging to the 20% of the highest scores of this test.

16- If the average height of a variety of wheat is 30 cm with a standard deviation of 6 cm, how many percent of these plants in a field are expected to be between 24 and 36 cm tall?

$$P(Z \geq 1) = 0.16$$

$$P(Z \geq 1.64) = 0.05$$

17- The germination power of a particular plant is 60%. If 8 seeds are planted from this plant, what is the probability that 6 or more seeds will germinate?

18- The average weight of tea bags has a normal distribution with a standard deviation of 10. If we want to determine the population mean with 90% confidence that the maximum error in the estimate is equal to 2, what should be the sample size given the following information?

19- If we have the following expression. Determine the positive and negative Z_1 and Z_2 ?

$$P(Z \geq Z_1) = 0.6 \quad P(Z \leq Z_2) = 0.8$$

20- It has been claimed that a new variety of sunflower has 40% oil with a standard deviation of 4%. In this test to determine the validity of this claim, the average oil from 16 test plots was calculated as 35%. Is the claim true? Why?

21- The lifespan of the computers produced by the computer factory has a normal distribution with an average of 10 years and a standard deviation of 3 years. If the factory wants to guarantee its computers in such a way that it covers 0.05 of the computers sold, for how many years should it guarantee them?

$$(Z_{0.05} = 1.64)$$

22- The average yield of fodder corn is equal to 18 with a standard deviation of 1 ton per hectare. A new hybrid was purchased and its average yield was estimated to be 17 tons per hectare by evaluating 100 test plots. Is it possible to show the inferiority of the new hybrid at the probability level of 0.01?

23- In a family with 6 children, what is the probability that the number of boys is less than the number of girls?

24- Number of table t with 5 degrees of freedom per $\alpha = 5\%$ From number of table t with 15 degrees of freedom per $\alpha = 5\%$ and from table Z with $\alpha = 5\%$ Is.

25- In a study for a certain attribute, 30 people were randomly used. If 95% of people for this attribute are between 14 and 26, what is the variance of the studied people? (Assume Z 0.05 equal to 2).

26- In a 10-question test with two true and false options, what is the probability that the candidate can guess and answer 6 questions correctly?

27- If for one degree of freedom, $P(X \geq 3.841) = 0.05$, then what is the value of Z_1 in the expression $P(-Z_1 \geq Z \geq Z_1)$?

28- If $P(|Z| \geq 1.96) = 0.05$, then $P(|Z| \leq 1.96)$ is equal to:

29- The oil percentage of a rapeseed variety is 40% with a variance of 9. In an experiment with 9 plots, the oil content of this cultivar was 36%. If $P(Z \geq 2.33) = 0.01$, can we conclude with 99% confidence that the percentage of oil in this figure is lower than the claimed amount?

30- The average grade of statistics course in a class of 40 students is 16. The standard error of a sample of 10 people is equal to 0.4. What is the rate of change in the statistics of that class?

31- A sample of one hundred has been randomly extracted from a large society. If the sample mean is equal to 50 and its standard deviation is equal to 3.3, what is the mean and variance of the population?

CHAPTER 5

STATISTICAL ASSUMPTION TEST

Assist. Prof. Dr. Mohsen MIRZAPOUR¹

Dr. Harun GİTARİ²

¹ - Siirt University, Faculty of Agriculture, Department of Agricultural Biotechnology, Siirt, Türkiye

ORCID ID: 0000-0002-2898-6903, e-mail: m.mirzapour@siirt.edu.tr

² - Kenyatta University, School of Agriculture and Enterprise Development, Department of Agricultural Science and Technology, Nairobi, Kenya
ORCID ID: 0000-0002-1996-119X, e-mail: harun.gitari@ku.ac.ke

INTRODUCTION

Research usually starts with questions and hypotheses. Any statement or claim about the statistical community is called a statistical assumption. Statistical hypothesis testing techniques are used for the validity of a statistical assumption. In fact, the statistical hypothesis test is a test to determine whether the statistical hypothesis is strongly confirmed or not according to the information obtained from the sample data. In other words, any statement about the population distribution or the population parameter is called a statistical hypothesis and it may be true or false. The truth or falsity of an assumption must be checked based on the information obtained from sampling from the community, this practice is called hypothesis testing.

Because the claim may be true or false, two complementary assumptions arise: one for the claim to be true and the other for the claim to be false. Therefore, the beginning of a hypothesis test always includes two statistical hypotheses that are opposed to each other.

5.1. Statistical judgments

Any statistical estimate such as \bar{X} has error bounds that must be controlled and calculated. In general, when a community parameter is estimated by sampling, the estimated value is almost equal to the true value of the parameter, and in the generalization of the results, there is a state of uncertainty, or in other words, there is a margin of error that must be determined and evaluated.

As explained, in order to calculate the average of the community μ , we use its estimate, \bar{X} , but it is obvious that \bar{X} is not necessarily the

same as μ , and the real average of the community may be slightly larger or smaller than \bar{X} . For this reason, we usually consider the two limits L and U and say that the average of the society is located between the two limits L and U. On the other hand, for our judgment, we also consider a degree of probability because there is a possibility that the real average of the community is outside the distance between L and U. So, here we consider an amount of numbers and a distance, which we denote by e. The degree of probability that one's judgment is correct is the degree of probability and the degree of probability that it is wrong is called the level of probability and the values of L and U are called confidence limits of the average of the community.

L and U are called the upper limit and the lower limit, respectively, which are obtained as follows:

$$U = \bar{X} + e \qquad L = \bar{X} - e$$

e: is called the maximum acceptable error, whose value is equal to:

$$e = Z\delta_{\bar{x}}$$

so that:

$$U = \bar{X} + Z\delta_{\bar{x}}$$

$$L = \bar{X} - Z\delta_{\bar{x}}$$

$\delta_{\bar{x}}$ is also the standard deviation of the mean distribution, which is equal to:

N is the number of sample members

$$\delta_{\bar{x}} = \frac{\delta x}{\sqrt{n}}$$

According to the relations defined for e and $\delta\bar{x}$. The formula for calculating the sample size can be obtained as follows:

$$e = Z \delta\bar{x} \qquad e = \frac{Z \delta\bar{x}}{\sqrt{n}} \qquad e^2 = \frac{Z^2 \delta\bar{x}^2}{n} \qquad n = \frac{Z^2 \delta\bar{x}^2}{e^2}$$

As it was said in the topic of research method, after determining the context and topic of the research, the expected results should be predicted. Such predictions are called research hypotheses. In other words, the hypothesis is the statement of the researcher's guess about the result of the research. The information and evidence obtained through experiments and research and finally expressed as statistical indicators are measured against these assumptions.

Every scientific research begins with the observation of a problem. The next stage of the research is that the researcher has assumptions to conduct the research.

For example, if the average weight of tea packages is less than 100 grams, you should complain to the manufacturer, it can be assumed that the average weight of tea packages is 100 grams and there is no need to complain. The reason for this is that because in inductive arguments, including statistical problems, the results are generalized according to observations and considering probabilities, we are never trying to prove something, but we are in the position of rejecting or not rejecting assumptions. In the terminology, we call such assumptions that we are trying to reject or acquit as zero or null and baseless assumptions, and we denote them with H_0 . Therefore, in the case of the discussed example, the null hypothesis is:

$$H_0: \mu = 100$$

So that in front of the null hypothesis, there is another hypothesis called the opposite hypothesis $1H$, which is accepted if the null hypothesis $0H$ is rejected, or it is not rejected statistically. Therefore, in the previous example, if the weight of the tea bags is less than 100, the assumption that it is equal to 100 and not complaining is rejected. Therefore, hypothesis H_1 is written as follows.

$$H_1: \mu < 100$$

In general, zero and one assumptions are logically opposed to each other, but in terms of probabilities, they are complementary. So that with the decrease of the first type error, the second type error increases, and with the decrease of the second type error, the first type error increases. So that the only way to simultaneously reduce both type 1 and type 2 errors is to increase the number of sample members by "increasing the sample size". Basically, hypothesis one is the hypothesis that the researcher wants to test. While the null hypothesis that we seek to reject or accept by conducting the experiment.

The next stage of the research is to collect information through sampling and averaging for or against the research assumptions.

Therefore, depending on the type of research, personal taste, each researcher can set a limit for decision-making, and this action does not have a scientific or standard aspect. Decision-making in a scientific research is not outside of one of the four states mentioned in the case of tea bags.

$1-H_0$ is correct and reject it based on the test result or if we accept H_1 as wrong, the reason for this can be wrong sampling or measurement error. In other words, the hypothesis H_0 is correct, but the researcher

decides to reject H_0 based on the results of the statistical sample. This type of wrong decision is called **type 1 error**. Type I error is also called significance level or detection level. The probability corresponding to the first type error is denoted by and defined as follows.

$$\alpha = P(\text{error of the first type}) = P(\text{reject } H_0 / H_0 \text{ is true})$$

2- H_0 is correct and based on the test result, we accept it or if H_1 is false, we reject it. This is where the right decision is made.

3- H_0 is wrong and we reject it based on the test result or if we accept H_1 as correct. This is where the right decision is made.

4- H_0 is wrong and based on the test result, we accept it or if H_1 is correct, we reject it. In other words, the hypothesis H_0 is wrong, but the researcher decides not to reject H_0 based on the results of the statistical sample. Here, a wrong decision has occurred, as well as a **type II error**. The reason for this is that sampling has not been done. The probability corresponding to the second type error is denoted by β and defined as follows.

$$\beta = P(\text{Type II error}) = P(\text{accept } H_0 / H_1 \text{ is true})$$

Here, α and β are used to indicate both the type of errors and the probability of committing those errors.

Here, α and β are used to indicate both the type of errors and the probability of committing those errors.

Decision	The real state of society	
	H ₀ correct	H ₀ Incorrect
Decision to accept H ₀	1 - α	β
Decision to reject H ₀	α	β-1

According to the above content, the method of standardizing the decision-making rules becomes clear. In a scientific research, it is possible to determine the maximum probability of committing the first type of error and use it as a decision criterion. This probability is known as the "significance level of judgment". It is clear that a good test method is one where the values of β and α are small.

- There are two types of errors in each test, first type error and second type error

- α and β are dependent on each other and there is an inverse relationship between them. Therefore, if the selected sample size is fixed, the reduction of each type of error will increase the other and vice versa.

$$\lim_{n \rightarrow \infty} \alpha = 1 \qquad \lim_{n \rightarrow \infty} \beta = 1$$

What is certain is that the sum of α and β is not necessarily one.

- By increasing the value of the sample size (n), both errors can be reduced at the same time.

$$\lim_{n \rightarrow \infty} \alpha = 0 \qquad \lim_{n \rightarrow \infty} \beta = 0$$

- If the assumption H_0 is false, then the value of β will reach its maximum if the true value of the parameter is chosen as the opposite assumption (H_1).

Reducing the first type of error, or in other words, reducing the detection level of the test, α , causes the H_0 hypothesis to be rejected with less chance if it is correct, or the probability of error in rejecting the H_0 hypothesis is reduced. On the other hand, reducing the probability of the second type of error, β , causes the H_0 assumption to be accepted with less chance if it is an error, or the probability of error in accepting the H_0 assumption is reduced.

5.2. Statistical hypothesis testing

A hypothesis is a researcher's guess about the relationship or difference between variables. Testing a statistical hypothesis consists of using a set of specific rules to determine whether to accept the null hypothesis or reject it in favor of the opposite hypothesis. In other words, hypothesis testing refers to activities that confirm or reject the existence of a relationship between variables with a certain degree of certainty. The main purpose of the hypothesis test is to estimate the parameter of the society from which the sample was extracted. In the hypothesis, it may be to find the difference between the desired variables, not the correlation between them. The proper method for hypothesis testing has different steps. Before mentioning the desired steps, we will explain the concepts and terms used in the assumption test.

5.3. Test statistics

For each parameter, a test statistic can be calculated based on sample observations. According to the value obtained for this statistic in the observed random sample, we will decide to reject or accept the H_0 hypothesis.

The test statistic either has a well-known distribution (such as Z, t, F, etc.) or its distribution can be obtained. Table 5-1 shows the appropriate test statistic for testing each parameter.

Table 5-1: test statistics of each parameter

Parameter	Hypothesis statistical symbol	Statistics	Sampling distribution	Test statistics
Average community	μ_0	\bar{X}	Normal student or t	Z or t
Comparison of the average of two societies	$\mu_1-\mu_2$	$\bar{X}_1- \bar{X}_2$	Normal student or t	Z or t
Community success ratio	P_0	\bar{P}	Normal	Z
Comparison of the success ratio of society	P_1-P_2	$\bar{P}_1- \bar{P}_2$	Normal	Z
Community variance	δ^2	S^2	Chi-squared test	χ^2
Comparison of the variance of two communities	$\frac{\delta_1^2}{\delta_2^2}$	$\frac{S_1^2}{S_2^2}$	Fisher	F

5.4. Rejection zone structure and critical values

The rejection region of a test is determined based on the structure of the opposite hypothesis, H_1 . The distribution of the test statistic is divided into two areas according to Figure 5-1, one is the H_0 rejection area and the other is the H_0 acceptance (non-rejection) area. We call the region of rejecting the hypothesis H_0 as the critical region and the

border between the rejection and acceptance regions as critical values. In fact, critical values separate the area of rejection from acceptance.

If the statistic is in the acceptance area, we assume the null hypothesis to be true, and otherwise, we reject it in favor of the opposite hypothesis. Usually, the region rejecting the null hypothesis is called the critical region.



Figure 5-1: Structure of rejection region and critical values

5.5. Test power

The power of the test is equal to the probability of rejecting the hypothesis H_0 when this hypothesis is really false and the probability of accepting the hypothesis H_1 when this hypothesis is really true. If β is the probability of type II error in a test, then the power of the test is defined as $(1-\beta)$.

$$\text{Test power} = 1 - \beta = 1 - P(\text{accept } H_0 / H_0 \text{ is not correct})$$

The test can be considered more powerful if it has a lower type II error β according to a certain value of α . The power function of a statistical hypothesis test of H_0 against the opposite hypothesis of H_1 is defined as follows.

$$\pi(H_0/H_1) = P(H_0 \text{ is rejected} / H_1 \text{ is true})$$

It should be known that any factor that reduces the second type error (β) will increase the power of the test. Therefore, the following increase the statistical power of the test:

Increasing the influence of the independent variable

Maximizing the variance of the dependent variable (or maximizing the variance of the error variables).

Increasing the size (volume) of the sample

The statistical power of a test can be increased by increasing the sample size, increasing the influence of the independent variable and maximizing the dispersion of the dependent variable.

Increasing the sample size reduces the values of α and β (probability of committing type 2 error) at the same time.

• If the statistical hypotheses are as follows:

$$H_0 : \mu = \alpha_0$$

$$H_1 : \mu \neq \alpha_0$$

Then the power function of the test will be equal to:

$$\pi = \beta - 1 (H_0/H_1) = F(\frac{\delta + S_{\alpha}}{2}) + F(\frac{\delta - S_{\alpha}}{2})$$

where is $\delta = \frac{a - a_0}{\frac{\delta}{\sqrt{n}}}$

If the statistical hypotheses are as follows:

$$H_0 : \mu = a_0$$

$$H_0 : \mu = a_0$$

$$H_1 : \mu > a_0$$

$$H_1 : \mu < a_0$$

Then the power function of the test will be equal to:

$$\pi = \beta - 1 (H_0/H_1) = F(Z_{\alpha} + \delta)$$

Example 5-1: In a hypothesis test $H_0 : \mu = 32.5$ and $H_1 : \mu \neq 32.5$ and H_0 is rejected, if based on a statistical sample size $n = 50$ and standard deviation $S = 5.5$, we have $\bar{X} < 34.8$ or $-34.8 > \bar{X}$, the probability of the first type error find out

$$\alpha = P(\bar{X} > 34.8 \quad \text{or} \quad \bar{X} < -34.8) = P(\bar{X} > 34.8) + P(\bar{X} < -34.8)$$

$$\alpha = P\left(\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} > \frac{34.8-32.5}{\frac{5/5}{\sqrt{50}}}\right) + P\left(\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < -\frac{34.8-32.5}{\frac{5/5}{\sqrt{50}}}\right)$$

$$\begin{aligned} &= P(Z > 2.96) + P(Z < -2.96) = 1 - P(Z < 2.96) + 1 - P(Z < 2.96) \\ &= 2 - 2 P(Z < 2.96) = 2 - 2(0.9985) = 0.003 \end{aligned}$$

The first type error is more important than the second type error, for this reason, α is called the significance level of the test.

There is an inverse relationship between α and β , but it is not necessary that their sum becomes one.

When the sample size increases, the type I and type II errors decrease and the power of the test increases.

Why is the first type error more than the second type error?

Suppose a mass production in warranty or production management stated that the rate (percentage) of defective goods in the production line is 5% ($H_0 = 0.05$). from 5% to 1%. The first type of error is more important because it has a high cost, but in the second type of error, we lose the opportunity. If the error of the first type is reduced, the power of the test will increase $1 - \downarrow \beta = \uparrow \gamma$.

5.6. Significant level

The power of each test depends on the probability of rejecting the null hypothesis, which is directly related to the degree of doubt in the correctness of the judgment I make. The significance levels that are used in the hypothesis tests are 5% and 1%. For example, when we accept an attribute with a probability of 95% confidence, we have only 5% doubt in accepting it. It means that the probability of making a mistake about this measured indicator was 5%. In other words, the null hypothesis is correct and accurate and we have to accept it, while there is a 5% chance of wrong judgment (due to the inability to estimate correctly). In other words, it should be known that the degree of confidence changes is called confidence levels. Confidence levels indicate how far the sample mean is from the population mean. The further the sample mean is from the population mean, the less we can trust it. For example, if the sample mean is outside the 0.01 interval, it means that there is a 1% chance that the sample mean is representative of the population mean, in other words, there is a 99% chance that the sample mean is not truly representative of the population mean.

Therefore, the maximum probability that is considered for the first type of error risk is called the significance level of the test. which we show with the symbol α . In order not to have an impact on statistical judgment, it is determined before the start of the experiment, like research hypothesis tests.

The calculated mean is a member of the distribution which is determined according to the central limit theorem. As mentioned earlier, it is called standard deviation, standard error or standard error.

This standard deviation arises from the differences between the means, which are classified as measurement and sampling errors. But the occurrence of less than that cannot be attributed to luck and coincidence, but its cause can be found in other cases.

$$\mu = \bar{x} \pm Z\check{\sigma}_{\bar{x}}$$
$$25 \pm 100$$

5.7. Significant difference

The difference between two numbers is said to be significant if its magnitude is statistically compatible with the laws of probability and we cannot consider this difference as a result of chance or mistake. When the null hypothesis is rejected at a significant level of 5%, the results are observed and therefore the test is called significant, and the statistical index used (in this example Z) is marked with an asterisk. If the null hypothesis is rejected at the level of 1%, the results are observed and therefore we call the test highly significant and indicate the statistical index with two stars. If the null hypothesis is not rejected in these two statistical levels, the test results are not significant in the conditions of the test and the statistical index is indicated by (ns).

When the results of a statistical test are not significant, the following two interpretations are acceptable.

1- The experimenter has not been able to achieve real results under the conditions of the experiment.

2- The obtained results are not correct. Of course, it is clear that there is a difference between the calculated average and the real average of the society, but the amount is not so much that it is statistically significant.

5.8. One-tailed and two-tailed tests

Each function of the parameters describing the statistical sample is called a statistic. The region rejecting the null hypothesis is called the critical region in a hypothesis test. In general, in the case that only the value obtained is less than a certain value is important for us, in this case, the critical area (rejecting the null hypothesis) is located on the left side of the curve and the test is one-tailed. For example, the weight of tea packages, which we test for the lower weight of the packages. The assumptions to be tested are:

$$H_0 = \mu = 100$$

$$H_1 = \mu < 100$$

The probability of committing the first type of error is equal to:

$$P(\bar{x} \leq \mu_0) = ?$$

Also, when the magnitude of the value obtained from a certain value is important to us (if the value obtained from a certain value is greater or less than the null hypothesis is rejected), the test is still one-tailed, in this case the region A critical point is located on the right or left side of the curve.

Example 5-2: In the research hypothesis, it is stated as follows: the average duration of use of a healthy light bulb is less than 200.

The above hypothesis is transformed into a statistical hypothesis that shows the contradiction of the claim and shows the research hypothesis. So:

$$\begin{cases} H_0: \mu \leq 200 \\ H_1: \mu < 200 \end{cases}$$

It is clear that the size of α will be defined on the left side of the sampling distribution. This type of test is called a "left-tailed test", Figure 5-2.

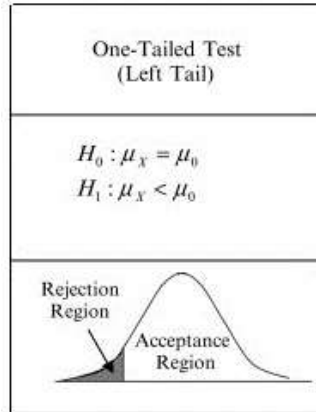


Figure 5-2: Left one-tailed test

Let's say in the above example: "The average working time of a healthy light bulb is more than 200 hours". The above hypothesis can be converted into a statistical hypothesis as follows:

$$\begin{cases} H_0: \mu \leq 200 \\ H_1: \mu > 200 \end{cases}$$

This type of test is called a "right-tailed test", Figure 5-3.

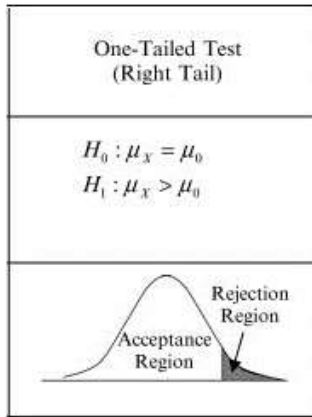


Figure 5-3: Right-tailed test

If it is not important for us that the obtained value is smaller or larger than certain values, in other words, if the obtained value is not equal to a certain value, the null hypothesis is rejected, the test is two-tailed and the critical region is in two. The end of the curve is located (Figure 5-4). Example: The thickness of the car cylinder head gasket that is larger or smaller than the desired value causes a problem, which raises both sides of the issue (i.e. smaller or larger than the actual value) and both cases are involved in committing the first type of error. Therefore, the hypotheses to be tested are:

$$H_0: \mu = 200$$

$$H_1: \mu < 200, \mu > 200, \mu \neq 200$$

In two-tailed tests, the significance level is halved ($\frac{\alpha}{2}$). If the significant level is equal to 5%, half of it is assigned to the probability of occurrence of each of the two cases (smaller and larger thickness of washers). In other words, the calculated probability of committing the first type of error is compared with 0.025 and in general with $\frac{\alpha}{2}$. The two-tailed test of the presence or absence of a difference between two

numbers is tested regardless of the direction of the difference (positive or negative).

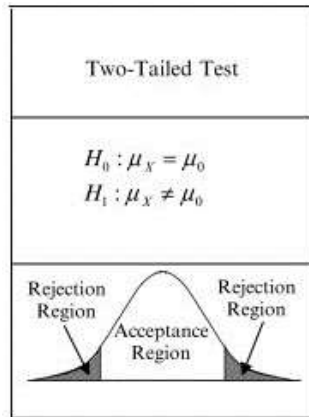


Figure 5-4: Two-tailed test

If the test statistic falls in the critical region, we reject the null hypothesis H_0 , because this indicates a significant difference between the null hypothesis H_0 and the sample data. Therefore, if the average of the sample is in the "critical area" in the above figures, the null hypothesis H_0 is rejected.

By examining H_1 , we must understand whether it is a right-tailed, left-tailed, or two-tailed test. The table 5-2 below shows how the inequality signs in H_1 points towards the critical region.

Table 5-2: the inequality signs in H_1 points towards the critical region

Test type	The sign used in H_1
Right domain	>
Left domain	<
Two domains	\neq

One-tailed tests are more powerful than two-tailed tests because one-tailed tests place the entire rejection region on one side of the distribution.

In general, with the increase of the probability level (α), the first type error of the test increases, but comparing a two-tailed test and a one-tailed test at a fixed probability level, the probability of the first type error in the one-tailed test will be higher.

In the two-tailed test, α is divided into two parts. For example, if $\alpha = 0.05$, the curve will be 0.025 on the right side (that is, 1.96 standard deviation units above the mean) and 0.025 on the left side (that is, 1.96 standard deviation units below the mean). contract.

The difference between one-tailed and two-tailed test is the level of alpha (α).

The probability of rejecting the null hypothesis is lower in a two-tailed test with a significance level of 0.08.

The power of statistical tests ($1-\beta$) depends on the probability of rejecting the H_0 hypothesis. The probability of committing type I and type II errors is indicated by α and β . So that for a given example, if α decreases, β increases. That is, for a given sample, if the probability of making the first type of error increases, the probability of making the second type of error decreases, and vice versa. Therefore, if the number of sample members increases, we can reduce the probability of committing two types of mistakes at the same time. In mathematical language, it can be written that:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad , \quad \frac{\partial^2 \bar{X}}{\partial n} = \frac{\sigma^2}{n}$$

Steps to perform the statistical hypothesis test

- 1- Formation of hypothesis zero and hypothesis one
- 2- Choosing the significance level of the test or the same as α
- 3- Determining the critical area
- 4- Determining the test statistic and calculating it with statistical sample values
- 5- Conclusion: If the test statistic is in the critical area, the null hypothesis is rejected, and otherwise, the null hypothesis is not rejected.

5.9. Hypothesis testing for the mean of a population

The test of the equality of the mean with a fixed number when the standard deviation is known

The statistical hypotheses for the population mean μ are considered as follows.

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

If the sample is selected from a normal population with a known standard deviation, the distribution \bar{X} is normal regardless of the sample size, and as a result, the test statistic will be Z. When the sample size is large ($30 < n$), the normality of the original population is not necessary. If the standard deviation of the population δ is not known, its estimator, the standard deviation (S), is used.

$$Z = \frac{\bar{X} - \mu}{\frac{\delta}{\sqrt{n}}}$$

In the above relation, μ is the average of the test item and the denominator of the variable Z is $\delta_{\bar{x}}$.

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, the value of μ at the detection level α cannot be accepted as the average of the society. If the numerical value of the test statistic is not in the critical region, we will have no reason to reject the H_0 hypothesis. That is, in the sample, there is no reason to reject the number (μ_0) as the community average, and it can be accepted.

Example 5-3: It has been claimed that the average electricity consumption of an area of Urmia in the month of September was at least 650 kilowatt hours. For this purpose, a random sample of 200 households was selected from the region, whose average and standard deviation of electricity consumption is 626 and 128 kilowatt hours. Consider the level of error of 1% and check the accuracy of the claim.

$$H_0 : \mu \geq 650 \qquad = 0.01$$

$$H_1 : \mu < 650 \qquad \quad \alpha$$

Since $n < 30$, the distribution has a normal approximation, so:

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{626 - 650}{\frac{128}{\sqrt{200}}} = - 2.65$$

$$-Z_\alpha = -Z_{0.01} = - 2.325$$

Because the numerical value of the test - 2.65 is smaller than the numerical value of the table - 2.325, therefore the hypothesis H_0 is rejected and the hypothesis H_1 is accepted.

5.10. The test of the equality of the mean with a fixed number when the standard deviation is unknown.

1- The statistical hypotheses for the community average (μ) are considered as follows.

$$H_0 : \mu = \mu_0$$

$$H_0 : \mu = \mu_0$$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu \neq \mu_0$$

2- If the sample is selected from a normal population with an unknown standard deviation, if the sample size is small ($30 > n$), the \bar{X} distribution has t-Stevens and the test statistic is:

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. If the numerical value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected.

Example 5-4: A random sample of 20 from a normal watermelon farm has an average weight (kg) of $\bar{X} = 23.9$ and a standard deviation of $S = 5$. Will this result suggest at the detection level of 0.05 that the average weight of the watermelon farm is more than 20 kg?

$$H_0 : \mu \leq 20 \quad = 0.05$$

$$H_1 : \mu > 20 \quad \alpha$$

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{23.9 - 20}{\frac{5}{\sqrt{20}}} = 3.48$$

$$t_{\alpha, n-1} = t_{0.05, 19} = 1.833$$

Because the numerical value of the test 3.48 is greater than the numerical value of the table 1.833, therefore, the hypothesis H_0 is rejected and the hypothesis H_1 is accepted.

5.11. Hypothesis test for the average of two statistical populations

The test of the sameness of means in two societies when the standard deviation is known

1- The statistical hypotheses for the averages of two communities (μ_1, μ_2) are considered as follows.

$$\begin{array}{lll}
 H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 \\
 H_1 : \mu_1 > \mu_2 & H_1 : \mu_1 < \mu_2 & H_1 : \mu_1 \neq \mu_2
 \end{array}$$

When samples are selected from two normal populations with a known standard deviation, the distribution $(\bar{X}_1 + \bar{X}_2)$ will be normal and the test statistic will be Z. When the sample size is large ($n_1, n_2 > 30$), the normality of the original population is not necessary. If the standard deviation of two populations δ_1 and δ_2 is not known, its estimator (S_1, S_2) is used.

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}}$$

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, the difference between the averages is significant and the average of two communities cannot be considered the same at the detection level α .

Example 5-5: In a sugar beet factory, there are two sugar production machines, and the packing samples were selected from each machine. The results of both friendships are obtained as follows. we have:

Device A	Device B
$n_1 = 25$	$n_2 = 50$
$\bar{X}_1 = 50.56 \text{ Kg}$	$\bar{X}_2 = 50.01 \text{ Kg}$
$\delta_1 = 0.4 \text{ Kg}$	$\delta_2 = 0.2 \text{ Kg}$

Test the claim that the average amount of sugar filled by machine A is equal to the average for machine B. Consider a significant level of 0.05.

$$H_0 : \mu_1 = \mu_2 \quad \alpha = 0.05$$

$$H_1 : \mu_1 \neq \mu_2$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}} = \frac{(50.56 - 50.01)}{\sqrt{\frac{0.16}{25} + \frac{0.2}{50}}} = 5.39$$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96 \quad \text{and} \quad -Z_{\frac{\alpha}{2}} = -Z_{0.025} = -1.96$$

Because the numerical value of the test 5.39 is greater than the numerical value of the table 1.96, therefore, the hypothesis H_0 is rejected and the hypothesis H_1 is accepted. As a result, it seems that device A fills packets significantly more than device B, and the average communities of the two devices are not equal.

5.12. The test of the sameness of means in two societies when standard deviations are unknown

1- The statistical hypotheses for the averages of two communities (μ_1, μ_2) are considered as follows.

$$\begin{array}{lll}
 H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 \\
 H_1 : \mu_1 > \mu_2 & H_1 : \mu_1 < \mu_2 & H_1 : \mu_1 \neq \mu_2
 \end{array}$$

2- When samples are selected from two normal populations with a known standard deviation, the sampling distribution of the statistic will depend on (n_1+n_2-2) . If the degrees of freedom are smaller than 30, the sampling distribution is the Student's t-statistic. In this case, the standard error will be influenced by the assumption of equality and inequality of variance of the two populations. So that it is $\delta_1^2 = \delta_2^2$, the test statistic is:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}, \quad t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In this case, the t distribution will have n_1+n_2-2 degrees of freedom. If $\delta_1^2 = \delta_2^2$ it is true, the test statistic is:

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

In this case, the t' distribution will have degrees of freedom based on the following relationship.

$$r' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. If the numerical value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected.

Example 5-6: Assuming that the variance of two populations is the same, two random samples are selected from two independent normal populations and the data is displayed as follows. At the detection level of 0.05, calculate the hypothesis $\mu_1 = \mu_2$ against the hypothesis $\mu_1 \neq \mu_2$.

Society B	Society A
$n_1 = 9$	$n_2 = 16$
$\bar{X}_1 = 30$	$\bar{X}_2 = 20$
$\check{d}_1 = 6.6$	$\check{d}_2 = 5.4$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(9-1)(6.6)^2 + (16-1)(5.4)^2}{9+16-2} = 34.16 \quad S_p = 5.84$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{30-20}{5.84 \sqrt{\frac{1}{9} + \frac{1}{16}}} = 4.11$$

$$t_{\frac{\alpha}{2}, n_1 + n_2 - 2} = t_{0.025, 23} = 2.069$$

Because the numerical value of the test 4.11 is greater than the numerical value of the table 2.069, therefore, the hypothesis H_0 is rejected and the hypothesis H_1 is accepted.

5.13. Hypothesis test of paired samples (comparison of pairs)

In the mean test of two populations, it is assumed that the samples of each group are independent from each other. Hypothesis testing based on non-independent (dependent) samples is known as paired samples test. In such cases, we act as follows.

When we are working with two dependent samples, extracting and recording $\bar{X}_1, S_1, n_1, \bar{X}_2, S_2, n_2$, is a futile task, because the relationship between the matched values is lost. Instead, we calculate the differences (d) between pairs of data according to the table 5-3 below.

Table 5-3: the differences between pairs of data

Pair	Behavior (x_i) 1	Behavior (y_i)2	Difference $d_i = x_i - y_i$
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_2 - y_2$
.	.	.	.
.	.	.	.
.	.	.	.
n	x_n	y_n	$d_n = x_n - y_n$

The distribution of d_i will be normal, we calculate the values of \bar{d} and S_d for the sample. \bar{d} is the mean of d values, S_d is the standard deviation of d values and n is the number of paired data.

$$\bar{d} = \frac{\sum d_i}{n} \qquad S_d^2 = \frac{\sum (d_i - \bar{d})^2}{n - 1}$$

1- Statistical hypotheses for paired data are considered as follows.

$H_0 : \mu_d = 0$	$H_0 : \mu_d = 0$	$H_0 : \mu_d = 0$
$H_1 : \mu_d > 0$	$H_1 : \mu_d < 0$	$H_1 : \mu_d \neq 0$

2- In random sampling from two dependent normal populations in which μ_d is the mean of paired data, the following test statistic has a Student's t-distribution with n-1 degrees of freedom.

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted, that is, at the detection level α , the difference between the values of x and y before and after the period is significant, and otherwise, the values of x and y before and after It has not changed since the period.

Example 5-7: The following table 5-4 shows the dry weight of weeds in the safflower field in the treatment of control and without control of weeds with a cover plant. Is it possible to claim that the planting of cover crops reduces the weed population at the error level of 1%?

Table 5-4: Dry weight of weed (gram)

Weed	with cover crop (WC)	No cover crop (NC)	d = NC - WC
<i>Convolvulus arvensis</i>	10	20	10
<i>Amaranthus retroflexus</i>	5	12	7
<i>Sorghum halepense</i>	8	16	8
<i>Chenopodium album</i>	5	10	5
<i>Carthamus tinctorius</i>	2	4	2
<i>Raphanus</i>	2	5	3

$$H_0 : \mu_d \leq 0$$

$$\alpha = 0.01$$

$$H_1 : \mu_d > 0$$

$$\bar{d} = \frac{\sum d_i}{n} = \frac{35}{7} = 5$$

$$S_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n_1 - 1}} = \sqrt{\frac{(10-5)^2 + (7-5)^2 + \dots + (3-5)^2}{7-1}} = 2.92$$

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{5}{\frac{2.92}{\sqrt{7}}} = 4.53 \quad \text{‘ } t_{0.01, 6} = 3.43$$

Because the numerical value of the test 4.53 is greater than the numerical value of the table 3.43, therefore, the hypothesis H_0 is rejected and the hypothesis H_1 is accepted. That is, the use of a cover plant will control weeds in the safflower field.

5.14. Hypothesis test of success ratio in a society

1- The statistical hypotheses for the success ratio in a society (p) are considered as follows.

$H_0 : p \leq p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
$H_1 : p > p_0$	$H_1 : p < p_0$	$H_1 : p \neq p_0$

2- The sampling distribution \bar{p} is normal and the test statistic is z.

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

In the above relationship, p_0 is the test ratio.

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, the number p_0 cannot be accepted as a population ratio at the detection level α . If the numerical value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected. That is, in the sample, there is no reason to reject the number H_0 as a proportion of society.

Example 5-8: It has been claimed that the sanitary condition of poultry farms is inappropriate. For this purpose, a random sample of 200 poultry farms has been selected, 30 of them have complained about the health situation. Calculate at 2% error level, can we say that the above claim is true?

$$n = 200, x = 30$$

$$\bar{p} = \frac{x}{n} = \frac{30}{200} = 0.15$$

$$H_0 : p \leq 0$$

$$H_1 : p > 0$$

$$\alpha = 0.02$$

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.15 - 0}{\sqrt{\frac{0(1-0)}{200}}} = \infty$$

$$Z_\alpha = Z_{0.02} = 2.055$$

Therefore, the numerical value of the test statistic is placed in the critical area, so the hypothesis H_0 is rejected and the hypothesis H_1 is accepted. It means that the sanitary condition of poultry farms is suitable.

5.15. Hypothesis test comparing the ratio in two statistical populations

1- The statistical hypotheses for success ratios in two societies (p_1, p_2) are considered as follows.

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2$$

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 < p_2$$

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

2- The distribution (p_1 , p_2) is normal and the test statistic will be Z.

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (P_1 - P_2)}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}$$

Sometimes the denominator of the test statistic is defined based on the common \bar{p} of two samples. The common \bar{p} is:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Now the previous relationship can be expressed as follows:

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, at the detection level α , the hypothesis of the same proportion of two societies can be accepted. If the numerical value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected. That is, the difference between the two ratios is significant and the ratios of the two societies cannot be considered the same.

Example 5-9: A random sample of 80 male students and 80 female students of the Faculty of Agriculture was selected and they were asked about having a personal car. If 25 of the boys and 20 of the girls say that they own a private car, can we conclude that male students have more private cars than female students? Use a detection level of 0.01.

$$\bar{P}_1 = \frac{x_1}{n_1} = 0.31 \quad n_1 = 80, \quad x_1 = 25$$

$$\bar{p}_2 = \frac{x_2}{n_2} = 0.25 \quad n_2 = 80, \quad x_2 = 20$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{25 + 20}{80 + 80} = 0.281 \quad \bar{q} = 1 - \bar{p} = 0.719$$

$$H_0 : p_1 = p_2 \quad \alpha = 0.01$$

$$H_1 : p_1 > p_2$$

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.31 - 0.25}{\sqrt{(0.281 \times 0.719)\left(\frac{1}{80} + \frac{1}{80}\right)}} = 0.845$$

$$Z_\alpha = Z_{0.01} = 2.33$$

Because the numerical value of the test 0.845 is smaller than the numerical value of table 2.33, therefore, there is no reason to reject the hypothesis H_0 and male students have more private cars than female students.

5.16. Statistical hypothesis testing for variance in a population

1- The statistical hypotheses for the variance in the community (δ^2) are considered as follows.

$$H_0 : \delta^2 = \delta_0^2 \quad H_0 : \delta^2 = \delta_0^2 \quad H_0 : \delta^2 = \delta_0^2$$

$$H_1 : \delta^2 > \delta_0^2 \quad H_1 : \delta^2 < \delta_0^2 \quad H_1 : \delta^2 \neq \delta_0^2$$

We know that the statistic δ^2 is the variance of the sample S^2 . The S^2 sampling distribution, if it is selected from a normal population, is a χ^2 distribution with $n-1$ degrees of freedom, whose test statistic is defined as follows.

$$\chi^2 = \frac{(n-1)S^2}{\delta_0^2}$$

In the above relation, δ_0^2 is the variance of the test.

3- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, at the detection level α , the number δ_0^2 cannot be accepted as the community variance. If the numerical value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected, that is, there is no reason to reject the number δ_0^2 as the population variance in the sample.

Example 5-10: A random sample of 10 fennel essential oils has an average limonene of 11.77% with a standard deviation of 2.2%. Calculate the hypothesis $\delta^2 = 2.1$ against the opposite hypothesis $\delta^2 \neq 2.1$ at the detection level of 0.05.

$$\begin{aligned} H_0 : \delta^2 &= 2.1 \\ H_1 : \delta^2 &\neq 2.1 \end{aligned} \quad \alpha = 0.05$$

$$\chi^2 = \frac{(n-1)S^2}{\delta_0^2} = \frac{(10-1)(2.2)^2}{2.1} = 20.74, \quad \chi^2_{\frac{\alpha}{2}, n-1} = \chi^2_{0.025, 9} = 19.02$$

Because the numerical value of the test 20.74 is greater than the numerical value of the table 19.02, therefore the hypothesis H_0 is rejected.

5.17. Statistical hypothesis test to compare the variance of two populations

1- The statistical hypotheses for the variances of two communities (δ^2_1, δ^2_2) are considered as follows.

$$\begin{array}{lll} H_0 : \delta^2_1 = \delta^2_2 & H_0 : \delta^2_1 = \delta^2_2 & H_0 : \delta^2_1 = \delta^2_2 \\ H_1 : \delta^2_1 > \delta^2_2 & H_1 : \delta^2_1 < \delta^2_2 & H_1 : \delta^2_1 \neq \delta^2_2 \end{array}$$

2- If the two communities being compared have a normal or approximate normal distribution, the test statistic for the ratio of the two variances of the community ($\frac{\delta^2_1}{\delta^2_2}$) has an F distribution.

$$F = \frac{S_1^2 \cdot \delta_2^2}{S_2^2 \cdot \delta_1^2}$$

And if the assumption of equality of two variances is accepted, the above relationship can be written as follows:

$$F = \frac{S_1^2}{S_2^2}$$

If two populations really have equal variances, then $F = \frac{S_1^2}{S_2^2}$ tends to 1 because S_1^2 and S_2^2 tend to values close to but if two communities basically have different variances, S_1^2 and S_2^2 will have different values.

3-The critical value is determined based on α and the degree of freedom ($r_1 = n_1 - 1$) and the degree of freedom of the denominator ($r_2 = n_2 - 1$) from Table F.

The critical value of $F_{1-\alpha}$ can be obtained from the table based on the following relationship.

$$F_{1-\alpha, r_1, r_2} = \frac{1}{F_{\alpha, r_1, r_2}}$$

4- If the numerical value of the test statistic is in the critical area, the H_0 hypothesis is rejected and the H_1 hypothesis is accepted. That is, at the detection level α , the hypothesis of the same variance of two communities can be accepted. If the value of the test statistic is not in the critical area, the hypothesis H_0 is not rejected. That is, the difference between the two variances is significant and the variances of the two societies cannot be considered the same.

Example 5-11: It shows the duration of storage (in months) of apples stored in a cold storage, which are stored in two different ways. Assume that each data set is drawn from a normally distributed population. At the detection level of 0.05, test the claim that the variance of the duration of apple storage with method A is equal to the variance of method B.

Storage with method A	Storage with method B
$n_1 = 10$	$n_2 = 12$
$\bar{X}_1 = 6$	$\bar{X}_2 = 6$
$S_1^2 = 4.06$	$S_2^2 = 0.4$

$$\begin{aligned}
 H_0: \delta_1^2 &= \delta_2^2 \\
 H_1: \delta_1^2 &\neq \delta_2^2
 \end{aligned}
 \quad \alpha = 0.05$$

$$F = \frac{S_1^2}{S_2^2} = \frac{4.06}{0.4} = 10.15$$

$$F_{\frac{\alpha}{2}, n_1-1, n_2-1} = F_{0.025, 9, 11} = 3.579$$

Because the numerical value of the test is 10.15 greater than the numerical value of the table 3.579, therefore, the hypothesis H_0 is rejected and the hypothesis H_1 is accepted. That is, there is enough evidence to reject the claim that the variances are equal.

Exercises of chapter 5

1- If you accept a hypothesis that is really wrong due to the inaccuracy of the test, which mistake did you make?

2- It is a society consisting of numbers 3, 6, 5, 7 and 4. If all possible binary samples are extracted from this population with replacement, what is the variance of the frequency distribution of the means?

3- A society has a normal distribution with a mean of 10 and a variance of 25. A sample of 64 individuals from this community was randomly selected and \bar{x} was calculated. What is the distribution of \bar{x} ?

4- From the statistical population consisting of 20 observations, all 5 samples have been selected with replacement. The mean and variance of these samples are 8 and 4.2, respectively. What is the mean and variance of the initial population from right to left?

5- A factory has claimed that its production car consumes 5.5 liters of gasoline per 100 km. After measuring the gasoline consumption of 35 cars, the sales representative of this factory estimated their average consumption per 100 kilometers to be 5.65 liters with a standard deviation of 35%. Do these results contradict the claim of the mentioned factory?

$$P(Z \geq 1.64) = 0.05$$

$$P(Z \geq 2.23) = 0.01$$

6- If all 9 possible samples are taken from a society with the distribution (3.27) $X \approx N$, what is the standard error of the averages?

7- All 25 samples are selected from the normal population with an average of 10 and a standard deviation of 2. What is the range that includes 90% of the sample means?

8- From a population consisting of 10 sample members, 4 samples were extracted and its mean and variance were calculated as 18 and 30, respectively. What is the standard error of the sample mean?

9- The average weight of oranges has a normal distribution with a mean of 127 and a standard deviation of 8. What is the probability that the mean weight of a sample of 64 is greater than 126?

CHAPTER 6

**NORMAL DISTRIBUTIONS, T, CHI-SQUARE, F
AND ANALYSIS OF VARIANCE**

Dr. Saeid HEYDARZADEH¹

Dr. Harun GĪTARĪ²

¹ - Former Ph.D. Student of Urmia University, Faculty of Agriculture, Department of Plant Production and Genetics, Urmia, Iran

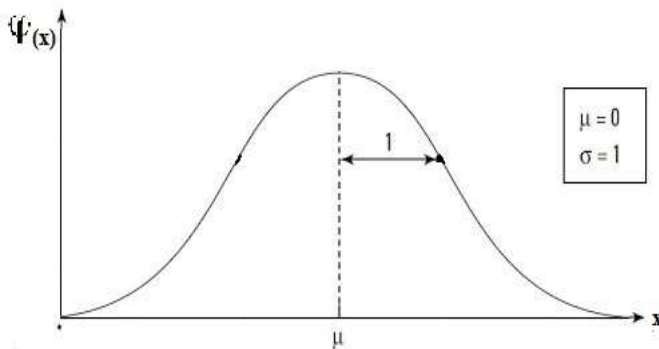
ORCID ID: 0000-0001-6051-7587, e-mail: s.heydarzadeh@urmia.ac.ir

² - Kenyatta University, School of Agriculture and Enterprise Development, Department of Agricultural Science and Technology, Nairobi, Kenya
ORCID ID: 0000-0002-1996-119X, e-mail: harun.gitari@ku.ac.ke

6.1. Normal distribution

The most important continuous probability distribution throughout the science of statistics is the normal distribution. Its graph is called a normal curve and it is bell-shaped, and measures such as strength and endurance and other physical capabilities are almost consistent with this natural pattern. In other words, if we choose several samples with equal volume from a community. The distribution of these samples will be normal, which is represented by a bell-shaped curve called the normal curve.

The meaning of normality or averageness of a phenomenon is its adaptation to the characteristics of the normal curve. Also, the normal distribution is based on an infinite number of observed cases.



A random variable x that has a bell-shaped distribution is called a normal random variable. If variable x is a normal random variable with mean μ and variance δ^2 , its density function will be as follows.

$$-\infty < x < \infty \rightarrow +\infty$$

$$\varphi_x(x) = \frac{1}{\delta\sqrt{\pi 2}} e^{-\frac{(x-\mu)^2}{2\delta^2}}$$

where $\pi = 3.14159000$ and $e = 2.71828000$. When the values of μ and δ are specified, the normal curve is precisely specified. Usually, when the random variable x has a normal distribution with mean μ and variance δ^2 , its density function will be as follows.

$$\bar{X} \sim N(\mu, \delta^2)$$

The normal probability distribution has two parameters and μ as δ^2 the mean increases, the curve shifts to the right, and as the variance increases, the curve becomes shorter.

Points to know about the surface under the curve:

The surface under the curve consists of 6 parts. In this way, they put zero in the place of the average and divide both sides of the curve into three standard deviations (6 parts) (Figure 1-6).

It is about 34.13% from zero to +1 region.

From zero to + 2 about 47.72 (47.72 = 13.59 + 34.13).

From zero to +3, it is about 49.78 (49.78 = 2.14 + 13.59 + 34.13).

The number of regions zero to 1 is about 34.13 percent, zero to 2 is about 47.72 percent, and zero to 3 is about 49.87 percent.

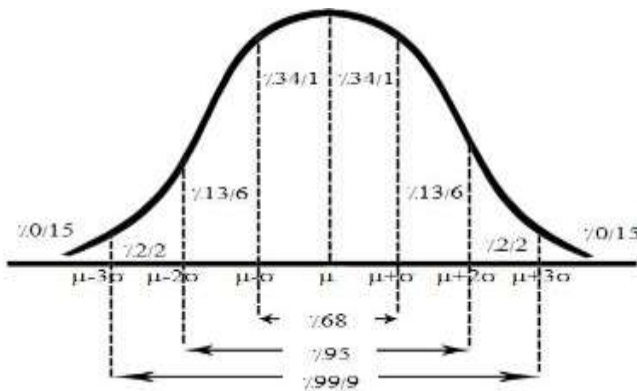


Figure 6 -1: Surfaces under the curve

It should be noted that the surface of curved areas may be rounded. For example, the level of -2 to -3 or +2 to +3 is about 2%, and the region of +1 to +3 or -1 to -3 is about 16%, and the entire area under the curve from -3 to +3 is about 100%.

In the range of one standard deviation unit higher ($Z = +1$) and lower than the average ($Z = -1$) about 68% of the data are under the curve.

In the range of two standard deviation units higher ($Z = +2$) and lower than the average ($Z = -2$), about 95% of the data are under the curve.

In the range of three standard deviation units higher ($Z = 3$) and lower than the average ($Z = -3$) in the curve, about 99% of the data are located.

Mathematical expectation, variance and moment generating function

$$E(X) = \mu$$

$$Var(x) = \sigma^2$$

$$M_x(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

6.2. Properties of normal distribution

The normal distribution is symmetrical about the line $x = \mu$. $(x + \mu) = \varphi(-x + \mu)$ φ

In normal distribution, the central parameters, mean, median and mode are equal to each other (they are on top of each other). So:

$$\mu = me = mo$$

The maximum value of the function $x = \mu$ is determined, so at $x = \mu$, first: $\varphi(x) = 0$, secondly: $\varphi(x) < 0$.

In the middle of the curve, there is the highest frequency and it gradually decreases on the sides, but never reaches zero.

The normal distribution has a maximum point for $x = \mu$, the maximum value of which is equal to $\frac{1}{\delta\sqrt{\pi 2}}$.

The normal distribution has two turning points for $\delta \pm x = \mu$, which has a width equal to $\frac{1}{\delta\sqrt{\pi 2}} e^{-\frac{1}{2}}$.

On both sides of the mean, the curve approaches its asymptote, that is, the x-axis (line $y = 0$). that's mean:

$$\lim_{n \rightarrow +\infty} \varphi(x) = 0 \qquad \lim_{n \rightarrow -\infty} \varphi(x) = 0$$

The area under the normal curve and the x-axis is equal to 1.

The area under the normal curve is divided by the line $x = \mu$ into two equal parts, each equal to $\frac{1}{2}$. That is, always more than 50% of the sizes are greater than the average and 50% of the sizes are less than the average.

The probability of an interval equal to "1 standard deviations on each side of the mean" equals 0.68, "2 standard deviations on each side of the mean" equals 0.95, and "3 standard deviations on each side of the mean" equals 0.99, that is:

$$p(\mu - \delta \leq x \leq \mu + \delta) = 0.68$$

$$p(\mu - 2\delta \leq x \leq \mu + 2\delta) = 0.95$$

$$p(\mu - 3\delta \leq x \leq \mu + 3\delta) = 0.99$$

These concepts are shown in the Figure 6-1:

6.3. Standard normal distribution

It should be known that the standard grades are the grades that we can use to calculate a grade for each person and compare the records. These scores determine the individual's position relative to the average and determine whether a score is several standard deviations above or below the average.

It should be known that the distribution of standardized numbers is normal if their number is very high and their mode, median, and average match each other, that is, their distribution is normal.

Standard scores are used with an interval scale.

The purpose of all standard scores is to determine the individual's status and success in the group.

The most accurate and valid educational grades are standardized grades, because they are calculated and determined in relation to the grades of the rest of the group.

The important difference between the normal curve and the standard normal curve is in the average and standard deviation indicators.

The real value of each grade is determined in relation to the statistical indicators of average and standard deviation.

The sum of the absolute value of the distance of the scores from the average is a statistical index to express the quality of the extent of the curve (diagram) of the scores.

The best way to get the standard score is to display the records with their mean and standard deviation, such scores are called the standard Z score. In other words, the criterion score or Z score tells us

where a certain score is in relation to the mean of the corresponding distribution. Therefore, if a random variable like x has mean μ and variance δ^2 , then if we subtract the mean of this random variable from it and divide it by its standard deviation, we have:

$$z = \frac{x - \mu}{\delta}$$

The resulting random variable, which is usually denoted by the letter z , is called "standardized variable" and this act of standardization is called. The mean of the standardized variable is zero and its variance is one, and they are shown as $\tilde{X} X(0, 1)$. It is proved that if the distribution of x is normal, the distribution of z will also be normal, and therefore its density function is as follows:

$$\phi_z(z) = \frac{1}{\delta\sqrt{\pi^2}} e^{-\frac{z^2}{2}} \quad -\infty < Z < +\infty$$

Such a distribution is called standard normal distribution.

In Z scores, because all the records are minus their average, so the average of these scores is zero.

In the standard Z score, the mean will be zero and the standard deviation will be 1, and the distance between the original numbers will not change.

If a student's score in a test is equal to the average, his Z score is zero.

In the normal curve, the peak of the curve, which has the highest frequency, is at the point $Z = 0$, that is, the peak of the curve is at the mean location.

If the distribution of scores is normal, the maximum frequency of scores corresponds to the point $Z = 0$.

The curve of two or more classes can be compared by averages and dispersion of scores.

It should be known that one of the disadvantages of Z score is that sometimes this score is negative and decimal, and this issue makes it difficult to understand this score.

Mathematical expectation, variance and moment generating function

$$E(Z) = 0$$

$$Var(z) = 1$$

$$M_z(t) = e^{\frac{t^2}{2}}$$

Example 6-1: If the protein percentage of two wheat genotypes from the normal population is 12 and 14, and the protein percentage of these two genotypes is 0 and 2 according to the standard z variable, find the mean and standard deviation.

$$z = \frac{x - \mu}{\delta} \quad \left\{ \begin{array}{l} 0 = \frac{12 - \mu}{\delta} \quad \mu = 12 \\ 2 = \frac{14 - \mu}{\delta} \quad 2\delta = 14 - \mu \quad 2\delta = 2 \quad \delta = 1 \end{array} \right.$$

6.4. The area under the normal curve

To calculate the probability that the random variable x takes a quantity between x_1 and x_2 , as mentioned in the discussion of continuous distributions, it is:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\delta\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}} dx$$

Calculating the above integral is a difficult and time-consuming task. To solve this problem, it is possible to express the measurements

in terms of standard units or standardized scores (z) by changing the origin and scale.

When x is between the values x_1 and x_2 , the random variable z between the corresponding values

$$z_1 = \frac{x_1 - \mu}{\delta} \text{ and } z_2 = \frac{x_2 - \mu}{\delta} \text{ will be placed. The area under the } x$$

curve between x_1 and x_2 according to the figure above will be equal to the area under the z curve between z_1 and z_2 . Hence we have:

$$P(x_1 < x < x_2) = P(z_1 < z < z_2)$$

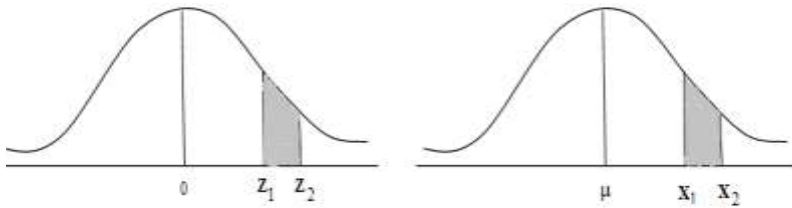
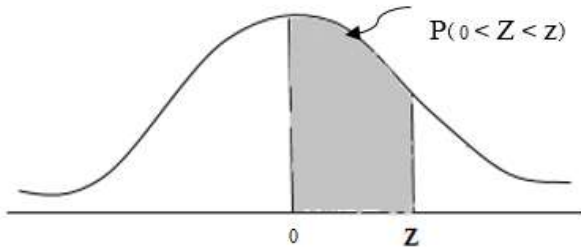


Table of the area under the distribution curve of the standard normal distribution

The numbers in the text of the table represent the level under the standard normal curve from zero to z (Appendix at the end of the book).



$$P(z > 3.59) = p(z < -3.59) \approx 0$$

$$P(z < 3.59) = P(z > -3.59) \approx 1$$

$$P(z \geq a) = 1 - p(z \leq a) \approx 0$$

$$P(a \leq x \leq b) = P(z \leq b) - p(z \leq a)$$

$$P(z \geq a) = p(z \leq -a)$$

In calculating all 3 possibilities in the normal distribution, it is useful to use the following method.

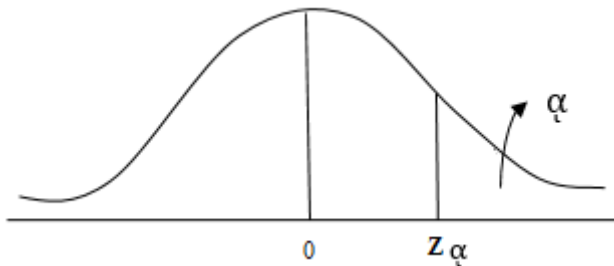
Write the possible equivalent statement.

Draw the normal curve.

Shade the desired surface.

Find the shaded area using the standard normal distribution table.

The percentage points of the normal distribution, if α is a number between zero and one ($0 \leq \alpha \leq 1$), then according to the definition Z_α is the point from which the surface of the curve is equal to α .



Example 6-2: If $P(0 \leq Z \leq 1) = 0.3413$, $P(0 \leq Z \leq 0.5) = 0.1915$, Find $P(-1 \leq Z \leq 0.5)$.

$$\begin{aligned} P(-1 \leq Z \leq 0.5) &= P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 0.5) \\ &= P(0 \leq Z \leq 1) + P(0 \leq Z \leq 0.5) \\ &= 0.3413 + 0.1915 = 0.5238 \end{aligned}$$

6.5. Approximation of binomial distribution by normal distribution

If X is a random variable of binomial distribution with mean $\mu = np$ and variance $\delta^2 = npq$, then the limiting form of the distribution is:

$$z = \frac{x - np}{\sqrt{npq}}$$

When $n \rightarrow \infty$, the standard normal distribution will be $NZ(0, 1)$.

If in the binomial distribution n is large and p is not close to zero, so that $5 > np$ and $5 < nq$, the normal distribution can be used approximately. Because the binomial distribution is discrete and the normal distribution is continuous, if necessary, you can use the continuity correction, that is, depending on the case, use $x \pm 0.5$ instead of x .

Example 6-3: Assume that the proportion of engines that have a defect in the assembly flow is 0.1. There are 200 engines in a particular shipment, what is the probability that at least 30 of these 200 engines are defective?

$$P = 0.1 \quad q = 0.9 \quad np = 200 \times 0.1 = 20 \quad nq = 200 \times 0.9 = 180$$

For continuity $30 - 0.5 = 29.5$

$$z = \frac{x - np}{\sqrt{npq}} = \frac{29.5 - 200 \times 0.1}{\sqrt{200 \times 0.1 \times 0.9}} = 2.24$$

$$= P(Z > 2.24) = 0.5 - P(0 < Z < 2.24) - 0.5 - 0.4875 = 0.0125$$

$$P(X \geq 30)$$

6.6. Approximation of Poisson distribution by normal distribution

When the mean of the Poisson distribution (λ) becomes relatively large, the normal approximation can be used for the Poisson distribution. The more λ increases, the closer the distribution is to the normal distribution. In general, if $10 < \lambda$, the normal approximation is a good approximation for Poisson. In this case, the mean and standard deviation will be equal to $\lambda = \mu$ and $\delta = \sqrt{\lambda}$.

$$P(X = x) = \frac{X - \lambda}{\sqrt{\lambda}}$$

Due to the discreteness of the Poisson distribution and the continuousness of the normal distribution, the continuity correction should be used as the binomial approximation by normal.

Example 6-4: If the number of customers of a bank is a Poisson random variable with a rate of 36 people per hour, what is the probability that at least 20 people visit this bank in a certain hour?

$$P(X \geq 20) = P(X \geq 19.5)$$

$$z = \frac{x - \mu}{\delta} = \frac{19.5 - 36}{6} = -2.75$$

$$\begin{aligned} P(Z \geq -2.75) &= P(Z \leq 2.75) = 0.5 + P(0 \leq Z \leq 2.75) \\ &= 0.5 + 0.4970 = 0.9970 \end{aligned}$$

6.7. Student's t-distribution

The standard normal variable "z" was previously defined as the ratio of the deviation of a random variable from its mathematical

expectation to the standard deviation of that variable. For example, for the random variable x , we can write:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

The variable Z has a normal distribution with a mean of zero and a variance of one.

Student's t -tests are used to compare the average of groups and are of three types: t -test between the average of a sample and the average of a population, t -test for independent groups and t -test for dependent groups.

t -test between the mean of a sample and the mean of a population: The following formula is used to check the difference between a sample and a larger population.

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}} \quad \text{or} \quad t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

\bar{X} = mean of t studied variable in the sample

μ = average of the desired community

$S_{\bar{X}}$ or $\frac{S}{\sqrt{n}}$ = standard error of the mean estimate

Independent t -test: If we want to compare two independent groups, we use independent t -test. For example, if we want to compare the experimental group with the control group in a research, we use independent t -test.

The assumptions of the independent t test are:

- Groups should be independent and randomly selected.
- The distribution of grades in society should be normal.
- The variances should be homogeneous (equal).

• The following formula is used to calculate the independent t value:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If the number of two groups is equal, the above formula can be used, but for convenience, the general form of the formula becomes as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

The following formula is a form of the t-test formula that can be easily calculated with a calculator.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\left[\sum X_1^2 - \frac{(\sum X_1)^2}{n_1} + \sum X_2^2 - \frac{(\sum X_2)^2}{n_2} \right] \times \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}{n_1 + n_2 - 2}}}$$

The degree of freedom of the independent t test is calculated based on the following formula:

$$df = n_1 + n_2 - 2$$

Independent t-test is used to check the difference between the average of two independent samples.

Dependent t-test: If we want to compare the mean with a predetermined number or if we measure the data on two separate occasions, we use the one-sample t-test (dependent t). The t ratio in the test comparing the averages of two samples is calculated using the following formula:

$$t = \frac{\bar{D}}{S_{\bar{D}}} = \frac{\bar{D}}{\frac{S}{\sqrt{n}}} \quad \bar{D} = \frac{\sum D}{n} \text{ Average difference between scores}$$

$S_{\bar{D}}$ = standard error of the difference between the scores of each subject

To calculate the dependent t value if the average difference of numbers and the standard error of their difference are equal, the following formula is used:

$$t = \frac{|\sum D|}{\sqrt{\frac{n \times \sum D^2 - (\sum D)^2}{n-1}}}$$

D = is the difference between pre-test and post-test score for each person in the group.

N = the number of pairs of individuals in the group.

The assumptions of the dependent t test are:

- Collected data belong to two different research conditions.
- The measurement scale should be of interval type.

The degree of freedom of the dependent t test is calculated based on the following formula:

$$df = n - 1$$

In the t test, the null hypothesis is confirmed when the calculated t is smaller than the table t, and the null hypothesis is rejected when the calculated t is greater than the table t.

t distribution is used to solve the problem of decimal or negative Z distribution.

The t distribution is always positive and the range of its changes is between 20 and 80, that is, it divides the 6 parts under the curve (+3 to -3) into 20 to 80 (Figure 6-2).

Sometimes the range of changes of T score is considered between zero and 100.

The t score is called the aligned score. The most common standard distribution is the Z distribution, but the t distribution is more useful than the Z distribution.

There is no difference between the standard t and Z distributions in terms of determining the rank or position of people in the group.

The shape of the distribution of Z numbers is the same as the shape of the distribution of the original raw numbers.

The standard Z and t distribution is used to compare people, that is, the information obtained from the Z and t distribution is essential.

In the standard t distribution, the mean is 50 and the Standard deviation (criterion) is 10, and the range of these scores is 20-80.

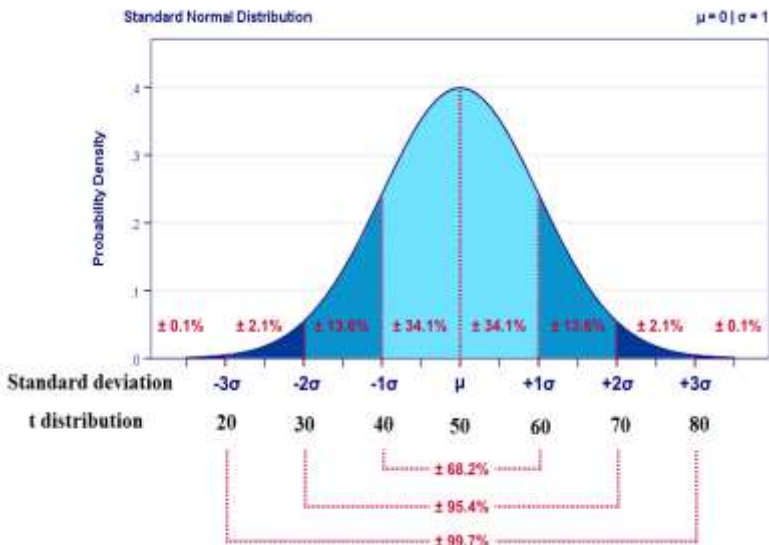


Figure 2-6: t- distribution and standard deviations

One standard deviation unit Z is equal to one tenth of standard unit t .

The standard t distribution has a constant mean and standard deviation, and we need other standard distributions, such as the Z distribution, to calculate this score.

The standard deviation of the standard t distribution is always equal to the Z factor. Because $Z = Z \times 10 + 50$

T scores are multiplied by 10 to remove the Z score decimal. Also, to remove negative Z scores, t scores are added to 50.

In a normal distribution, the mean is equal to $t = 50$ and $Z = 0$.

In some cases, the variance of the initial population is not known and we are dealing with small samples ($n < 30$). The frequency distribution used in small samples is called t -distribution or Student's t -distribution. If the true value of σ^2 is unknown and S^2 , i.e. the estimate of σ^2 , is placed in the formula instead, the resulting distribution will follow another distribution called t :

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

Because the estimate of the community variance S^2 is based on the degree of freedom ($df = n-1$) and depending on the sample size, we get different estimates for the community variance, the variable t for each value of the degree of freedom has a different frequency distribution and is proportional to the variance estimate. It is society. As the degree of freedom increases, the number t becomes closer to z . So that in the lower degree of freedom, the symmetrical t distribution is wider and the surface under the curve is more on the sides than the

normal distribution. While there is a slight difference between them as the degree of freedom increases.

A particular mean is only converted to a Z, but depending on which degree of freedom the population variance is calculated, a particular mean is converted to different values of t.

The terms of the t distribution are:

The random sample should be small ($30 \geq n$).

The population variance is unknown.

The distribution of the main population is normal.

6.8. Characteristics of t distribution

The t distribution is similar to the standard normal distribution, symmetric with zero mean. That is, the vertical axis is its axis of symmetry.

The t curve is shorter than the z curve in the central parts and wider than it in the tails (surroundings).

The t and z distributions are bell-shaped, but the t distribution is more variable, because its distribution depends on the two parameters \bar{x} and S^2 , while the values of Z only depend on the changes of \bar{x} from one sample to another.

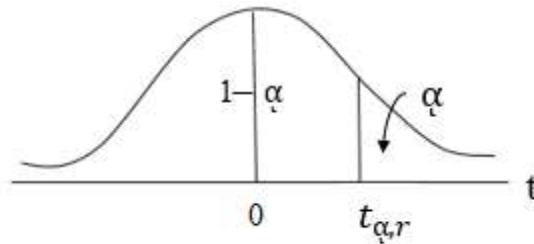
The standard deviation of the t distribution is greater than one, as a result, the dispersion of the t distribution is greater than that of the z distribution, and therefore the t distribution is different from the z distribution.

The t distribution is an approximation of the standard normal distribution, and the higher the degree of freedom of the t distribution,

the closer this approximation becomes. so that in the limit we can say that when $r \rightarrow \infty$ the distribution of t tends to the distribution of z .

In the sample size of about 30, the t -distribution almost coincides with the normal distribution.

t distribution table: for a certain degree of freedom, the number in the text of the table represents the value of t corresponding to the level under the curve α in the right tail of the distribution. (Appendix at the end of the book).



Unlike the z distribution, we can have a certain t that has different probabilities of occurrence depending on the degree of freedom. Also, for a certain probability, unlike the distribution of z , different values of t are observed.

In the t table in the last row ($r = \infty$), the values of the t table match the values of the standard normal table.

The table determines how many times the observed difference must be the error in order to reject H_0 and be confident that the maximum probability of committing a type I error is α percent.

The confidence limits of the community average:

$$P\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right) = (1 - \alpha)\%$$

Example 6-5: The amount of tractor fuel per hundred kilometers has a normal distribution. A sample of 9 tractors was randomly selected

and the average and variance of fuel consumption of samples 14 and 4 were obtained. The necessary statistic for test $H_0: \mu = 12$ is:

$$\bar{X}=14$$

$$S_{\bar{X}} = \sqrt{\frac{S_x}{n}} = \sqrt{\frac{4}{9}} = \frac{2}{3} = 0.667$$

$$t = \frac{\bar{x}-\mu}{S_{\bar{x}}} = \frac{14-12}{0.667} = 3$$

6.9. Comparing the averages of two samples

One of the important cases of statistical judgment is comparing the averages of several societies. In this section, we examine the method of comparing two averages. There are two modes in this section:

1- Paired measurements are:

The act of pairing observations is used in cases where the variation between two series of statistical data does not have a particular trend or when each individual or member of the experiment is subjected to two types of treatment.

In the case that observations are paired, the number of n_1 must be equal to n_2 ($n_1 = n_2$). To compare the average of two samples, the t-test is used as follows:

$$t = \frac{\bar{d}-\mu_0}{S_{\bar{d}}}$$

And if the null hypothesis is $H_0: \mu_1 - \mu_2 = 0$

$$t = \frac{\bar{d}}{S_{\bar{d}}} \qquad \bar{d} = \frac{\sum d_i}{n}$$

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

$$S_d = \sqrt{\frac{S^2_d}{n}} = \frac{S_d}{\sqrt{n}}$$

2- The measurement was not paired.

a): The variances should be equal.

If there are two samples with the averages of \bar{X}_1 and \bar{X}_2 , it is possible to check whether the two samples are obtained from the same population with the average μ or whether each of them belongs to two populations with the average μ_1 and μ_2 . If the mathematical expectation \bar{X}_1 and \bar{X}_2 are μ_1 and μ_2 respectively, the null hypothesis and the opposite hypothesis will be:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad \mu_1 = \mu_2 = \mu$$

As you know, the variance of two independent variables is equal to the sum of their variances. The average difference of two samples will have variance $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. n_1 and n_2 here show the sample size, i.e. the number of observations in the sample.

$$S^2_{\bar{X}_1 - \bar{X}_2} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

The ratio of the difference of a quantity from its mathematical expectation to the square root of the variance of that quantity has a t distribution. So:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Assuming zero, that is $\mu_1 = \mu_2$, the right side of the fraction will be equal to zero, therefore:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|\bar{d}|}{s_{\bar{d}}}$$

After calculating t , it is compared with t in the table at the desired probability level and the degree of freedom related to s^2 . If the calculated t is smaller than the table t , it is said that the null hypothesis is true. It means that the probability of occurrence of t calculated in population t is high and it can be considered as part of this population. The opposite assumption is $\mu_1 \neq \mu_2$. Larger differences may result in the calculated t being greater than the table t , and as a result, the null hypothesis will be rejected in favor of the opposite hypothesis.

When it is observed in the assumption that the variances are equal or it is determined in the test that there is no statistically significant difference between the variances of two samples, the average of the two variances should be taken, in this case:

$$s_p^2 = \frac{s_1^2 + s_2^2}{2}$$

And since each of these variances (s_1^2 , s_2^2) can be calculated according to a different number of observations, then:

$$s_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$s_1^2 = \frac{\sum(X_{1j} - \bar{X}_1)^2}{n_1 - 1} = \frac{\sum X_{1j}^2 - \frac{(\sum X_{1j})^2}{n_1}}{n_1 - 1}$$

as:

$$s_P^2 = \frac{\sum(X_{1j} - \bar{X}_1)^2 + \sum(X_{2j} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\left\{ \sum X_{1j}^2 - \frac{(\sum X_{1j})^2}{n_1} \right\} + \left\{ \sum X_{2j}^2 - \frac{(\sum X_{2j})^2}{n_2} \right\}}{n_1 + n_2 - 2}$$

s_P^2 It is an estimate of the variance of both societies, and their equality is a necessary condition for conducting the study, therefore:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{s_P^2}{n_1} + \frac{s_P^2}{n_2}} = \sqrt{s_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

If the size of the two samples is equal, i.e. $n_1 = n_2 = n$, the calculation of $s_{\bar{X}_1 - \bar{X}_2}$ can be easily calculated from the average variance of the two samples:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{2s_P^2}{n}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

When the variances of two samples are not the same, the variance of each sample is put in place of the common s^2 of the t formula. In this case, the t distribution becomes the t' distribution.

$$t' = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Example 6-6: A genotype of wheat has been cultivated in 9 plots of land and it is claimed that the average protein percentage of this genotype is 13.5%. If the calculated protein percentage values for 9 pieces are as follows, test the accuracy of this claim at the 5% probability level.

$$12.9, 13.4, 12.4, 12.8, 13, 12.7, 12.4, 13.5, 13.9 \quad \bar{X} = 13$$

$H_0 : \mu = 13.5$ $H_1 : \mu \neq 13.5$ (two-domain test)

$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}} = \sqrt{\frac{1523.08 - \frac{(117)^2}{9}}{8}} = 0.5099$$

$$t = \frac{\bar{x} - \mu}{\frac{S_x}{\sqrt{n}}} = \frac{13 - 13.5}{\frac{0.5099}{\sqrt{9}}} = -2.94$$

Because the test is of two domains and the calculated t is not between + t and - t of the table, then the null hypothesis ($\mu = 13.5$) is rejected. That is, the average percentage of protein was not equal to the claimed value of 13.5% (t table with degree of freedom $n - 1 = 8$ and $\alpha = 5\%$ is equal to 2.31).

$$\frac{\alpha}{2} 2.5\% = t_{0.975}$$

Example 6-7: Suppose we want to replace the existing variety of a plant with a new variety and the following information is available:

New variety	$\bar{X}_1 = 4000$	kg	$s_1^2 = 36000$	$n_1 = 11$
Variety available in the area	$\bar{X}_2 = 3700$	kg	$s_2^2 = 4000$	$n_2 = 11$

Do the results justify this substitution?

At the beginning, we test the uniformity of variances of two samples:

$$F = \frac{36000}{4000} = 9$$

Considering that the F obtained is greater than the F in the table (4.85) at the 1% probability level, then the variances of the two samples are significantly different from each other and the t' test should be used:

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4000 - 3700}{\sqrt{\frac{36000}{11} + \frac{4000}{11}}} = \frac{300}{60.3} = +4.97$$

Because the calculated t is greater than the one-sided t table number (1.81) with a degree of freedom of 10, therefore, the H_0 hypothesis is rejected in favor of H_1 .

Due to the fact that the table related to the t' distribution is rarely included in the classic statistics books, usually the numbers of the t table are used approximately to determine the significance of the t' statistic. If $n_1 = n_2$, the degree of freedom to determine the table number t will be $n-1$. If $n_1 \neq n_2$, two methods can be used:

The first method: using the smallest degree of freedom (at least n_1-1 and n_2-1).

The second method: calculating the average degree of freedom through Sattervit's formula.

$$df' = \frac{\left\{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right\}^2}{\frac{\left\{ \frac{s_1^2}{n_1} \right\}^2}{n_1-1} + \frac{\left\{ \frac{s_2^2}{n_2} \right\}^2}{n_2-1}}$$

Example 6-8: Two data samples of the diameter of mycelium colonies were prepared. Sample A included 11 observations and its mean and variance were estimated as 6.65 and 5.2824 respectively. Sample B includes 16 observations and its mean and variance were calculated as 4.28 and 8.0275, respectively. Assuming normal distribution and equal variance of these two populations, the Student's t -statistic to compare the average of these two samples is equal to:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1) + (n_2-1)} = \frac{(11-1)5.2824 + (16-1)8.0275}{(11-1) + (16-1)} = 10.929$$

$$s^2_{\bar{X}_1 - \bar{X}_2} = s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 10.929 \left(\frac{1}{11} + \frac{1}{16} \right) = 1.677$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{1.677} = 1.295$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{6.65 - 4.28}{1.295} = \frac{2.37}{1.295} = 1.84$$

Example 6-9: In the previous question, the degree of freedom for referring to the t-Student table is equal to:

$$df = (n_1 - 1) + (n_2 - 1) = (11 - 1) + (16 - 1) = 25$$

Note: In order to determine the uniformity of the variances of two samples, the larger variance is divided into the smaller variance. This ratio with F distribution obtained is compared with F in the table. If F is not significant, the t test is used, and if it is significant, the t' test is used.

6.10. Estimate confidence limits of means

As you know, the sample mean, \bar{X} , is an unbiased estimate of μ . Therefore, for the usefulness of the estimation of μ through \bar{X} , we use the confidence limits of the average. The t formula can be used to estimate confidence limits. If n samples are taken from a normal population with mean μ and variance δ^2 , t will be equal to:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}}$$

The value of t_p with n - 1 degrees of freedom can be obtained from the t table. 0.05 t is considered here. So:

$$P\{|\bar{X} - \mu| \leq t_{5\%} \sqrt{\frac{S^2}{n}}\} = 0.95$$

That is, the probability that $|\bar{X} - \mu|$ on the right side, the inequality is equal to 5%. In other words, the absolute value of $\bar{X} - \mu$ in 95% of the samples from this population is equal to or less than $t_{5\%} \sqrt{\frac{S^2}{n}}$. This inequality is shown as follows:

$$P\{-t_{5\%} \sqrt{\frac{S^2}{n}} \leq \bar{X} - \mu \leq +t_{5\%} \sqrt{\frac{S^2}{n}}\} = 0.95$$

By adding \bar{X} to the sides of the inequality, we will have:

$$P\{\bar{X} - t_{5\%} \sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + t_{5\%} \sqrt{\frac{S^2}{n}}\} = 0.95$$

That is, in 95% of the samples from a population, the mean will be in the range of $\bar{X} \pm t_{5\%} \sqrt{\frac{S^2}{n}}$. This interval is called 95% confidence interval for μ . Similarly, 99% confidence limits are obtained using $t_{\%1}$. In general, the average confidence limits at the probability level p are equal to $\bar{X} \pm t_p S_{\bar{x}}$. The confidence limits of the average will be different from one sample to another because \bar{X} and S^2 are not fixed values and vary from one sample to another. For each sample, the μ value is either within this interval or outside it. It can only be said that by repeating the sampling, μ is expected to be in this interval in 95% of cases. It can be seen that the confidence limits of the mean provide more information than \bar{X} in the estimation of the μ parameter.

The confidence limits of the difference between two means can be shown in the form of the following general relationship:

$$P\{(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} (S_{\bar{X}_1 - \bar{X}_2}) \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} (S_{\bar{X}_1 - \bar{X}_2})\} = (1-\alpha)\%$$

Example 6-10: Examining 11 colonies of snails living in a special environment in terms of the lack of stripe trait gave the following results:

$$\bar{x} = 0.34 \quad , \quad S^2 = 0.024$$

The question is, what is the distance between the population mean at the 5% and 1% probability levels? First, the table t for 11- 1= 10 degrees of freedom is obtained at the probability level of 5% and 1%, which is equal to 2.23 and 3.17, respectively, and $S_{\bar{x}}$ is also calculated from the following formula:

$$S_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.024}{11}} = 0.047$$

Now, the confidence limits of the mean can be calculated:

$$P[0.34-(2.23)(0.047) \leq \mu \leq 0.34 + (2.23)(0.047)] = 0.95$$

$$P(0.24 \leq \mu \leq 0.44) = 0.95$$

That is, with 95% confidence, the average population is between the numbers 0.24 and 0.44.

Similarly, the 99% confidence limits will be:

$$P[0.34-(3.17)(0.047) \leq \mu \leq 0.34 + (3.17)(0.047)] = 0.99$$

$$P(0.19 \leq \mu \leq 0.49) = 0.99$$

6.11. Special modes of comparing two averages:

1- The samples should be large and the population variance (δ^2) should be known.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\delta_{\bar{X}_1 - \bar{X}_2}} \delta^2_{\bar{X}_1 - \bar{X}_2} = \delta^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

2- The population variance is unknown and $n_1 \neq n_2$.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

$$S^2_{\bar{X}_1 - \bar{X}_2} = Sp^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$Sp^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

3- The population variance is unknown and $n_1 = n_2 = n$.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

$$S^2_{\bar{X}_1 - \bar{X}_2} = Sp^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = \frac{2Sp^2}{n}$$

4- In cases where the observations are paired.

$$t = \frac{\bar{d}}{S_d}$$

$$S^2_d = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1} \quad S^2_d = \frac{Sd^2}{n}$$

5- In cases where two samples are selected from communities with different variances ($\delta_2^2 \neq \delta_1^2$).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

$$S_{\bar{X}_1 - \bar{X}_2}^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

In this case, the obtained t is compared with t' .

6.12. Chi-score or chi-square distribution

The chi square test (χ^2) or chi 2 is used to evaluate the relationship between two nominal variables, regardless of the sign. For example, the relationship between gender and smoking addiction is measured using this statistical method.

In some cases, chi-square test is used to compare the frequency of an experimental event against a theoretical event. For example, flipping a coin for 100 may be 70 times milk and 30 times gold. In this situation, the frequency of this experimental event can be tested against the theoretical frequency, which says that under normal conditions, 50 times a tap and 50 times a line should be observed.

In Chi-square test, the collected information may belong to one variable (one variable is divided into different classes), or the collected information may belong to two variables.

Chi-square test is used to calculate the difference between two or more variables with a nominal scale.

The presuppositions of the chi-square test are:

- Data should be of abundant type.

- The sample size should be sufficient.
- Measurements should be independent of each other.
- There should be a theoretical basis for the classification of variables.

Our goal is to test the significance of the difference between the observed frequencies and the frequencies we theoretically expect. Because we test how well the observed frequency distribution (experimental frequency) matches (or fits) the expected frequency distribution (theoretical frequency³), this method is often known as a goodness-of-fit test.

χ^2 distribution is a continuous distribution and variables that have such a distribution are called continuous variables. But the number of occurrences in the χ^2 test is not continuous and is limited to real numbers. Therefore, it is recommended that the expected number remain in decimal form and not be converted into integers. In this way, the χ^2 statistic has been adjusted for the actual number of events. The use of relative frequencies and percentages in the χ^2 formula should be avoided. A common mistake is to use percentages instead of actual numbers in calculations.

The statistical hypotheses are as follows:

Experimental and theoretical frequencies are

H_0 :the same

Experimental and theoretical abundances are

H_1 :not the same

³ - The expected frequency of an event is the product of the probability of that event multiplied by the total number of trials

Suppose N statistical data are classified in C groups and the observed frequency of each class is o_1, o_2, \dots, o_c respectively. Also, suppose that based on theoretical principles or some assumptions, including the null hypothesis, the expected frequency in each of the classes is e_1, e_2, \dots, e_c respectively. In that case, it is proven that if the deviation of each observed frequency from the expected frequency reaches the power of two and is divided by the expected frequency and the sum of this expression is obtained for all frequencies, the resulting sum has a chi-square distribution.

$$\chi^2 = \sum_{i=1}^C \frac{(o_i - e_i)^2}{e_i}$$

According to the above formula, if the probability of obtaining a χ^2 value is small, the null hypothesis (the presence of agreement between the observed and expected frequencies) can be rejected, and it is certain that the probability of committing the first type of error (rejecting the null hypothesis is correct) it's small. The correct use of this distribution requires that the frequency of each group is at least equal to 5. In cases where some frequencies are less than 5, two classes or groups can be merged. If the sum of observed and expected frequencies are equal, the degree of freedom χ^2 of the table is equal to the number of classes minus one (k-1). If we use community parameters to calculate expected frequencies, the degree of freedom χ^2 is equal to (k-1) minus the number of parameters used to calculate those frequencies. In the Chi -square test, if the collected information belongs to two variables, we get the degrees of freedom from the formula $df = (k-1)(r-1)$.

A one-domain chi-square test is synonymous with a two-domain Z-test.

If the observed frequencies are close to the corresponding expected values, the χ^2 value is small and that is a good indication of the adaptation (or fit). If the observed values are significantly different from the expected values, the χ^2 value will be large and the fit (or fit) will be poor. Goodness of fit (large χ^2 value) will lead to rejection of H_0 . As a result, the critical region will be located on the right side of the chi-square distribution (the right is the range).

If $(1 - \text{number of groups}), a \chi^2 < \chi^2$, we reject the hypothesis H_0 at the significance level of α .

If $(1 - \text{number of groups}), a \chi^2 < \chi^2$, the hypothesis H_0 is not rejected at the significance level of α and can be accepted.

Therefore, the chi2 distribution for the degrees of freedom is obtained and the desired confidence level (usually 0.05 or 0.01) is compared with the chi2 of the special tables attached at the end of the book. If the calculated χ^2 was smaller than the χ^2 of the table, the null hypothesis is confirmed, and if it was greater than the χ^2 of the table, the null hypothesis is rejected.

6.13. Corrected Chi-square

It should be known that in chi -square analysis, if the number of classes or levels is only 2, its degrees of freedom will be equal to 1. In such a situation, the test result is subject to error. In order to avoid this error, we make a change in the formula of χ^2 . This change consists of subtracting 0.5 from the absolute value of the formula, this action is

called Yates correction. In other words, the chi-square distribution is a continuous distribution. If the frequency of classes in the agreement table is less than 5, the continuity of the χ^2 distribution curve is not established, and Yates correction is used to remove this violation. When the degree of freedom is $r = 1$, in this case Yates correction is used and the χ^2 formula is as follows.

$$\chi^2 = \sum_{i=1}^C \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

Chi-square test is a non-parametric test. In this test, the measurement scale is nominal and the two groups are independent.

The use of Chi-square test requires random selection of a sample of appropriate size.

In the Chi-square test, if the expected frequency of each class is less than 5, the reliability of the test decreases.

The power of Chi-square test is not reported. Because this test is used if there is no suitable parametric test to analyze the same data.

When the data are abundant and discrete and stratified. Chi-square test is used.

Chi-square test is a common method to calculate the difference between two variables with a nominal scale.

In a 2×2 consensus table, the total number (N) must be at least 20 in order to be able to accurately calculate the Chi-square.

It should be known that when using the Chi-square test, a large sample size should be chosen as much as possible, so that the expected

frequencies in tables with more than two rows and columns are less than 5 and in tables with two rows and two columns. are not less than 10.

Example 6-11: 120 people have graduated in a university over a period of several years, 50 of them are non-native and the rest are native. If the test is performed, its value and degree of freedom will be equal to:

Since the degree of freedom is equal to one, Yates correction is used.

$$\chi^2 = \sum_{i=1}^C \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

$$\chi^2 = \sum_{i=1}^C \frac{(|o_i - e_i| - 0.5)^2}{e_i} = \frac{(|50 - 60| - 0.5)^2}{60} + \frac{(|70 - 60| - 0.5)^2}{60}$$

$$\chi^2 = 1.5 + 1.5 = 3$$

$$df = 2 - 1 = 1$$

We use this test only when none of the theoretical frequencies is less than 5, when one of the theoretical frequencies is not less than 5, in this case we merge that category into another category, and as a result, the degree of freedom will be one unit less. became.

6.14. Independence test - double χ^2 test

In some cases, it is necessary to examine classification with two variables, which is called double χ^2 . In the double χ^2 test, we are in contact with consensus tables. Consensus table is used as a useful relationship to analyze the dependence of one variable with another

variable, it is only a statistical dependence and cannot be used as an inherent cause and effect relationship.

In general, a consensus table $r \times c$ has r rows that represent different states x_1, x_2, \dots, x_r of a variable attribute and has c columns that represent different states y_1, y_2, \dots, y_c of an attribute. is another variable.

The statistical hypotheses are as follows:

H_0 : Two variables in question are independent

H_1 : The two variables in question are not independent

The statistics of the independence test are as follows:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(|o_{ij} - e_{ij}|)^2}{e_{ij}}$$

where o_{ij} is the frequency of the observed house (experimental frequency) and e_{ij} is the expected frequency (theoretical frequency) for the house located at the intersection of the i -th row and the j -th column. For each house in the table, e_{ij} can be determined as follows.

$$e_{ij} = \frac{(\text{Frequency of lines } i \text{ th})(\text{Sum of column frequencies } j \text{ th})}{(\text{total sum})}$$

In a consensus table with r rows and c columns, the degree of freedom is as follows.

$$df = (r-1)(c-1)$$

Independence tests with agreement tables only involve the critical regions of the right domain, because small values of χ^2 confirm the claim of independence of two variables. In other words, if the observed and expected frequencies are close, χ^2 is small. Large values of χ^2 lie

on the right side of the chi-square distribution and reflect a significant difference between the observed and expected frequencies.

We reject the carpet H_0 if the numerical value of the test statistic is greater than $\chi^2_{\alpha, (r-1)(c-1)}$

If the analysis of double χ^2 test shows that there is a relationship between two variables, the strength of this relationship can be determined with the help of the following agreement coefficient.

$$c = \sqrt{\frac{\chi^2}{\chi^2+n}}$$

where c is a value calculated according to the formula and n is the total number of observations.

Calculation of correlation coefficient through chi 2

Two types of correlation coefficient can be calculated directly through chi2:

- 1- Phi coefficient (ϕ)
- 2- Consensus Coefficient C (C)

It should be known that if the χ^2 is statistically significant, the researcher can calculate the correlation coefficient to determine the intensity of the relationship, but if the χ^2 is not significant, the difference between the frequencies may be due to chance, and in this situation, it is not logical to convert the χ^2 into a correlation coefficient.

1- Phi coefficient: Chi2 calculated from a 2×2 consensus table can be converted to Phi coefficient. The phi coefficient is the square root of the χ^2 divided by the total number of subjects. To calculate phi, you need to calculate Chi 2 without Yates correction. The following formula is used for this conversion.

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Chi-square determines the significance of the correlation coefficient and the phi coefficient determines the intensity of this correlation.

2- Consensus factor C: If the number of houses in the consensus table is more than 2×2 . Phi coefficient cannot be used to calculate the correlation coefficient. In such a situation, the coefficient of resistance C should be used. C is an index that calculates the degree of correlation between two nominal variables arranged in the form of a consensus table. This coefficient is calculated using the following formula.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The value of C depends on the number of houses in the table. Therefore, it is possible to compare different values of C if their tables have the same rows and columns.

6.15. Table of agreement between frequencies

Agreement table is a table in which a series of statistical data are classified based on two characteristics. If the first characteristic or attribute has C different levels and the second characteristic or attribute has r different levels, the said table is called $r \times c$ consensus table. The purpose of testing the assumption about the existence or non-existence of the relationship between two classification criteria (attributes and characteristics) is. The hypothesis that one factor is independent of

another or not can be tested with the chi-square test. Here, the expected frequency is calculated according to the probability of occurrence of the combination in the independent event. The degree of freedom χ^2 in the agreement table is equal to $(r-1)(c-1)$ and in cases where the population parameter is used to calculate the expected frequencies, it is equal to $(r-1)(c-1)-m$ that m is the number of population parameters to calculate expected frequencies.

Example 6-12: In the agreement table 3×4 degrees of freedom is equal to:

$$df = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$$

If $Z_1, Z_2 \dots Z_n$ are standard normal independent random variables, then the distribution of the random variable $Z_1^2 + Z_2^2 \dots Z_n^2$ in a chi-square distribution with n degree They say freedom and show it with the symbol χ_n^2 . We say that the random variable x has a chi-square distribution with r degrees of freedom. If its density function is as follows:

$$f_x(x) = \frac{1}{2^{\frac{r}{2}} \sqrt{\Gamma(\frac{r}{2})}} x^{\frac{r-2}{2}} \cdot e^{-\frac{x}{2}} \quad (x > 0)$$

It is noted that the density function χ^2 follows only one parameter r , i.e. the degree of freedom. That is, if the degree of freedom changes, the shape of the curve also changes.

Mathematical expectation, variance and moment generating function

$$E(X) = r$$

$$\text{Var}(X) = 2r$$

$$M_x(t) = (1 - 2t)^{\frac{r}{2}}$$

If S^2 is the variance of an n random sample selected from a normal population with variance δ^2 , then:

$$\chi^2 = \frac{(n-1)S^2}{\delta^2}$$

It is a value of random variable χ^2 with chi-square distribution with $r = n-1$ degrees of freedom. The general form of the chi-square distribution is $\frac{rS^2}{\delta^2}$, where r represents the degree of freedom.

Considering that the mathematical expectation of chi distribution is twice the degree of freedom, then:

$$E\left(\frac{(n-1)S^2}{\delta^2}\right) = r \quad \left(\frac{(n-1)E(S^2)}{\delta^2}\right) = n-1 \quad E(S^2) = \delta^2$$

Given that the variance of the chi distribution is twice its degree of freedom, then:

$$\text{Var}\left(\frac{(n-1)S^2}{\delta^2}\right) = 2r \quad \frac{(n-1)}{\delta^4} \cdot \text{Var}(S^2) = 2(n-1) \quad \text{Var}(S^2) = \frac{2\delta^4}{n-1}$$

6.16. Properties of Chi-square

The values of χ^2 cannot be negative because χ^2 is the sum of a number of squares and therefore its range of variation is from zero to infinity.

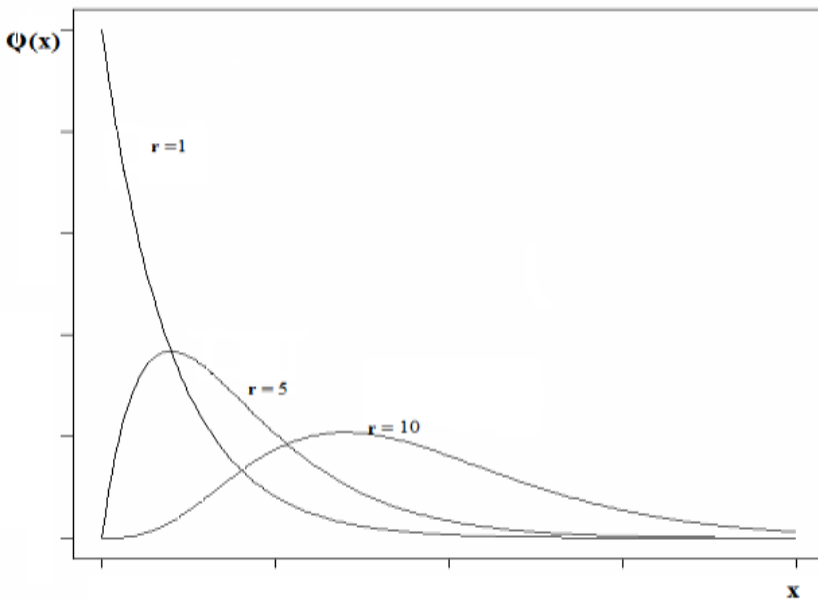
The χ^2 distribution is an exponential.

The number of degrees of freedom of the χ^2 distribution is equal to the number of observations.

The shape of the χ^2 distribution is determined by its degree of freedom, r .

For a low degree of freedom, the χ^2 distribution has a positive skewness, and as r increases, the skewness of the distribution decreases and goes towards the normal distribution. When $10 < r$, the distribution is almost approximate.

When $r=1$, the curve becomes the right half of the normal curve.



When $30 < r$, the $2\sqrt{\chi_r^2}$ distribution tends towards the normal distribution with mathematical expectation $\sqrt{2r - 1}$ and variance 1, that is $N(\sqrt{2r - 1}, 1)$ and standard It has a random variable as:

$$Z = \sqrt{2\chi_r^2} - \sqrt{2r - 1}$$

which has a normal distribution of $N(0, 1)$.

If $\chi_{r_1}^2$, $\chi_{r_2}^2$, ... , $\chi_{r_k}^2$ and ... and χ^2 are independent random variables, each according to the χ^2 distribution with degrees The freedom of r_1 , r_2 , ... , r_k are distributed, then their sum.

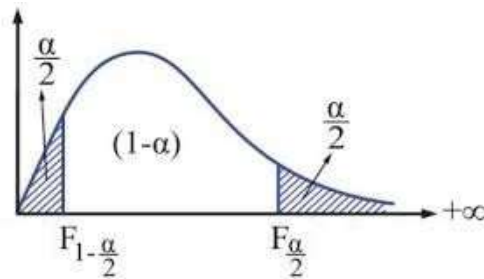
$$\chi_{r_1}^2 + \chi_{r_2}^2 + \dots + \chi_{r_k}^2 = \sum_{i=1}^k \chi_{r_i}^2$$

It is also distributed according to the χ^2 distribution with the degree of freedom r, which is equal to:

$$r = r_1 + r_2 + \dots + r_k$$

$\chi_{\frac{\alpha}{2}}$ is such that the area under the curve on the right is equal to $\frac{\alpha}{2}$,

and for $\chi_{1-\frac{\alpha}{2}}$ the area under the curve on the left is equal to $\frac{\alpha}{2}$.



Example 6-13: A normal population has a variance of 6. If a random sample of 25 is selected from this population, what is the probability that the sample variance is between 3.45 and 10.75?

$$P(3.45 < S^2 < 10.75) = ?$$

$$P(3.45 < S^2 < 10.75) =$$

$$P(3.45 \frac{24}{6} < \frac{(n-1)S^2}{\delta^2} < 10.75 \frac{24}{6}) =$$

$$P(13.8 < \chi^2 < 43) = P(\chi^2 < 43) - P(\chi^2 < 13.8) = 0.99 - 0.05 = 0.94$$

Example 6-14: A normal population has a variance of δ^2 and a mean of μ . We select a sample of 11. Consider the random variable $Y = \frac{10S^2}{\delta^2}$.

a) Calculate the probability $P(3.94 < Y < 18.31)$.

b) Using the above probability, determine the relationship between S^2 and δ^2 .

$$P(3.94 < Y < 18.3) = P(Y < 18.3) - P(Y < 3.94)$$

$$P(Y < 18.3) - P(Y < 3.94) = 0.95 - 0.05 = 0.9$$

$$P(3.94 < Y = \frac{10S^2}{\delta^2} < 18.3) = P\left(\frac{3.94 \delta^2}{10} < S^2 < \frac{18.3 \delta^2}{10}\right)$$

$$P(0.394 \delta^2 < S^2 < 1.83 \delta^2) = 0.9$$

6.17. Comparison of two variances and F distribution

If two independent samples of sizes n_1 and n_2 are equally extracted from a normal population or from two normal populations with equal variance:

$$S_1^2 = \frac{\sum(x_{1j} - \bar{x}_1)^2}{n-1}, \sum(x_{1j} - \bar{x}_1)^2 = (n_1 - 1)S_1^2$$

$$S_2^2 = \frac{\sum(x_{2j} - \bar{x}_2)^2}{n-1}, \sum(x_{2j} - \bar{x}_2)^2 = (n_2 - 1)S_2^2$$

Remembering that the deviation of the observations from the mean has a mean of zero, if the sum of their squares is divided by the variance of the population, their variance will also be equal to one, and therefore they will become the chi-square variable.

$$X_1^2 = \frac{(n_1-1)S_1^2}{\delta^2}, \quad X_2^2 = \frac{(n_2-1)S_2^2}{\delta^2}$$

Therefore:

$$\frac{X_1^2}{X_2^2} = \frac{\frac{(n_1-1)S_1^2}{\delta^2}}{\frac{(n_2-1)S_2^2}{\delta^2}} = \frac{(n_1-1)S_1^2}{(n_2-1)S_2^2} = \frac{\sum(x_{1j}-\bar{x}_1)^2}{\sum(x_{2j}-\bar{x}_2)^2}$$

And if $n_1 = n_2 = n$:

$$\frac{X_1^2}{X_2^2} = \frac{S_1^2}{S_2^2}$$

If multiple samples are extracted from a community and the values $\frac{X_1^2}{X_2^2}$ are calculated, the above ratios have F frequency distribution.

The frequency distribution of F and its curve shape is similar to the X^2 distribution and its formula is the ratio of the variances of two independent samples or:

$$F = \frac{X_1^2}{X_2^2} = \frac{S_1^2}{S_2^2}$$

The t test is used to compare two means and the F test is used to compare two variances.

Therefore, F distribution follows two parameters X_1 and X_2 .

If the random variable $\bar{X}_1 \chi^2_{r_1}$ and the random variable $\bar{X}_2 \chi^2_{r_2}$ as well as the aforementioned random variables are independent of each other, then:

$$F = \frac{\frac{\chi^2_{r_1}}{r_1}}{\frac{\chi^2_{r_2}}{r_2}} = \frac{\frac{X_1}{r_1}}{\frac{X_2}{r_2}} \sim f_{r_1, r_2}$$

it will be Therefore, theoretically, F distribution can be defined as the result of dividing two independent chi-square distributions, each of which is divided by its degrees of freedom. It follows from this theorem that if the random variables are $F \sim f_{r_1, r_2}$, then:

$$E(F) = \frac{r_2}{r_2 - 2}, \quad Var(F) = \frac{2r_2^2(r_1 + r_2 - 2)}{r_1(r_2 - 4)(r_2 - 2)^2}$$

The F distribution is used to compare variances in two or more populations.

If T has a t distribution with degree of freedom r:

$$T^2 \sim t_r \quad T^2 \sim F_{1, r}$$

Consider two independent normal populations with variances δ_1^2 and δ_2^2 , if we choose a sample size of n_1 from the first population and n_2 from the second population and the sample variances are S_1^2 and S_2^2 , then:

$$F = \frac{\frac{S_1^2}{\delta_1^2}}{\frac{S_2^2}{\delta_2^2}} = \frac{S_1^2 \delta_2^2}{S_2^2 \delta_1^2}$$

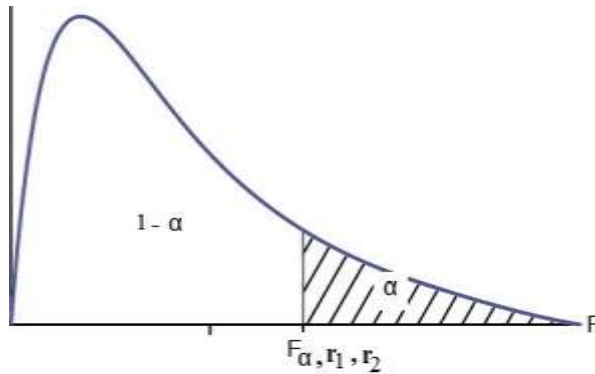
It will have F distribution with degrees of freedom $r_1 = n_1 - 1$ and $r_2 = n_2 - 1$.

6.18. Characteristics of F distribution

F distribution also has the characteristics of χ^2 distribution, that is, firstly, it is non-negative, secondly, it has an exponential and thirdly, it has a positive skewness.

The distribution curve F not only depends on two parameters r_1 and r_2 , but also depends on the order of their declaration.

F distribution table, for a specific combination of numerator and denominator degrees of freedom, the number in the table represents the value of F corresponding to the area under curve a in the right tail of the distribution. (Appendix at the end of the book).



Regarding the values related to the F distribution table, the following relationship always holds.

$$f_{1-a.r_1.r_2} = \frac{1}{f_{a.r_1.r_2}}$$

The values of the quantiles of F distribution are obtained from the above equation.

Conversion relation between F distribution with χ^2 and t distributions

If in the F distribution, the degree of freedom of the denominator, i.e. r_2 , tends towards ∞ , then:

$$f_{1-a.r_1.\infty} = \frac{\chi^2_{1-a.r_1}}{r_1}$$

If in the F distribution, the degree of freedom of the face, i.e. r_1 tends to ∞ , then:

$$f_{1-a.r_2.\infty} = \frac{r_2}{\chi^2 a. r_2}$$

If in the F distribution, the degree of freedom of the face, i.e. r_1 is equal to 1, then:

$$f_{1-a.1. r_2} = t_{1-\frac{a}{2}, r_2}$$

If in F distribution, the degree of freedom of the face, i.e. r_1 , tends towards 1 and the degree of freedom of the denominator, i.e. r_2 , tends towards ∞ , then:

$$f_{1-a.1. \infty} = Z_{1-\frac{a}{2}}$$

Example 6-15: The values of sum of squares for two wheat varieties evaluated in 21 completely uniform experimental plots are estimated as 20 and 50. If F in the table is equal to 2, can the assumption of uniformity of variance of these two varieties be rejected?

Because the calculated F is greater than the table F. So the hypothesis of uniformity of variance of these two varieties is rejected.

$$F = \frac{S_1^2}{S_{22}^2} = \frac{S^2}{S^2} = \frac{50}{20} = 2.5 > 2 \text{ Table}$$

6.19. Variance hypothesis test

Just as the population mean estimate (\bar{X}) was tested in relation to assumptions about the population mean, the population variance estimate is also tested as follows. In this regard, the chi-square distribution is used and the size and smallness of the sum of squares is used as a test.

$$H_0 : \delta^2 = \delta_0^2$$

$$H_1 : \delta^2 \neq \delta_0^2$$

$$\chi^2_1 = \frac{(n-1)S^2}{\delta^2} = \frac{\sum(x-\bar{x})^2}{\delta^2}$$

If the calculated χ^2 is greater than the Chi-square value of the table with the degree of freedom (n-1) and the confidence level of α percent, the null hypothesis is rejected.

Example: suppose that a researcher uses a breed of mice whose weight has a standard deviation of 26 grams for his laboratory experiments, in order to improve the condition of his experiments, this researcher decides to use another breed that has greater weight changes. Before this selection, and for more certainty, he conducts a test to find out whether the new breed has statistically more variation than the old breed. 20 mice of a new breed were randomly selected and their weight was measured. If $\sum(x - \bar{x})^2 = 23000$ and the standard deviation is equal to 34.79 grams, can this researcher use the new breed with 95% confidence that his desire will be met?

$$H_0: \delta^2 = (26)^2$$

$$H_1: \delta^2 \geq (26)^2$$

$$\chi^2 = \frac{23000}{(26)^2} = 34.02$$

Here too, the probability of committing the first type of error, i.e. $P(\chi^2 \geq 02.34)$ should be determined and compared with 0.05 or 0.01,

but by calculating the chi score (χ^2_c) and comparing it It is enough with the chi-square of the table ($\chi^2_{\alpha, df}$).

$$\chi^2_{0.05,19} = 30.14 < \chi^2_c = 34.02$$

Therefore, since the calculated chi-square is greater than χ^2 in the table, the null hypothesis is rejected and it is concluded that the new breed has a standard deviation greater than 26 with a probability of 95% and it can be replaced by the previous breed. The assumptions of the test can also be determined in terms of standard deviation, and in case of subtracting χ^2 , put the root of the sum of squares.

6.20. Confidence limits of community variance:

If a random sample is selected from a population and its variance is estimated, the confidence limits of this estimate will be as follows:

$$P(X^2 \leq X^2 \leq X^2) = (1-\alpha) \%$$

$$P(X^2 \leq \frac{(n-1)S^2}{\delta^2} \leq X^2) = (1-\alpha) \%$$

$$P(X^2 \geq \frac{(n-1)S^2}{\delta^2}) = \frac{\alpha}{2} \%, \quad P(X^2 \leq \frac{(n-1)S^2}{\delta^2}) = (1-\frac{\alpha}{2}) \%$$

$$P(\frac{X^2}{(n-1)S^2} \leq \frac{1}{\delta^2} \leq \frac{X^2}{(n-1)S^2}) = (1-\alpha) \%$$

$$P(\frac{(n-1)S^2}{X^2} \leq \delta^2 \leq \frac{(n-1)S^2}{X^2}) = (1-\alpha) \%$$

$$P(\frac{\sum(X-\bar{X})^2}{X^2 \text{ df}(\frac{\alpha}{2})} \leq \delta^2 \leq \frac{\sum(X-\bar{X})^2}{X^2 \text{ df}(1-\frac{\alpha}{2})}) = (1-\alpha) \%$$

Example 6-16: The 95% confidence limit of the amount of protein of a variety of wheat whose 9 random samples have a mean of 13 and a standard deviation of 0.51 is equal to:

$$Z_{0.025} = 1.96, t_{8, \alpha = 0.05} = 2.306$$

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} = \frac{0.51}{\sqrt{9}} = 0.17$$

$$\bar{x} \pm t S_{\bar{x}} \quad 13 \pm (2.036 \times 0.17) \quad 12.61 \leq \mu \leq 13.39$$

6.21. Proportions and frequencies test

Due to the fact that chi-square distribution was discussed for quantitative (continuous) variables, but this distribution is mostly used for testing qualitative (discrete) variables. In some cases, it is desirable to compare the frequency of the observed ratios with theoretical ratios and frequencies, in terms of having a significant difference. In situations where we are faced with two frequencies or two ratios, the test of the significance of the difference between the ratios is performed using the z distribution, and the chi-square test is used to compare more than two frequencies or ratios. If in a limited society with N members, the probability of occurrence of event A is equal to P and the probability of its non-occurrence is equal to $q = 1 - P$, the distribution of the sentence is binomial, and its mean and standard deviation are equal to:

$$\mu = np \quad , \quad \delta = \sqrt{np(1 - p)}$$

If all n samples are extracted and the estimate of p is calculated in each sample, these values form the frequency distribution of the ratios, whose mean and standard deviation are equal to:

$$\mu_p = p \quad , \quad \delta_p = \sqrt{\frac{p(1-p)}{n} \left(\frac{N-n}{N-1} \right)}$$

And in order to test the comparison of two ratios, the observed ratio or frequency is standardized according to the following formula, and the probability of occurrence of its different values is determined according to the Z table.

$$Z = \frac{n-np}{\delta} \qquad Z = \frac{p-p_0}{\delta p}$$

Example 6-17: In a public opinion survey, 80 people out of 400 people gave a positive answer to a specific issue, the 95% confidence limit for estimating the proportion of people who give a positive answer is equal to:

$$P = \frac{80}{400} = 0.2 \qquad q = 1 - p = 1 - 0.2 = 0.8 \qquad n = 400$$

$$\delta_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.2 \times 0.8}{400}} = \sqrt{0.0004} = 0.02$$

$$Z = \frac{p-p_0}{\delta_p} \qquad \pm 1.96 = \frac{p-0.2}{0.02}$$

$$P - 0.2 = \pm 0.0392 \qquad 0.1608 \leq P \leq 0.2392$$

6.22. Distribution of ratio differences

If samples consisting of n_1 and n_2 members are randomly extracted from two populations in which the ratio of a certain characteristic is p_1 and p_2 respectively, and the desired ratio is estimated for each sample and the difference of the ratios is calculated, it is proven that the average The difference of ratios is equal to the difference of two ratios at the community level ($p_1 - p_2$) and the variance of the difference of ratios is equal to:

$$\delta^2_{P_1 - P_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

q_1 and q_2 are the probability (ratio) of non-occurrence of that characteristic in two societies. If the ratios are not known at the community level, they are estimated by sampling each community once.

It is also possible to calculate the ratio estimate in two populations first, like paired observations:

$$\begin{aligned}\hat{p} &= \frac{p_1 + p_2}{2} = \frac{n_1 + n_2}{N} \\ \hat{q} &= 1 - \hat{p}\end{aligned}$$

And then the variance of the ratio difference distribution is equal to:

$$\delta^2_{P_1 - P_2} = \hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

In this case, the observed difference can be standardized and the probability of its occurrence can be calculated.

$$Z = \frac{(P_1 - P_2) - (p_1 - p_2)}{\delta_{P_1 - P_2}}$$

$(p_1 - p_2)$ is the difference assumed in the null hypothesis.

Example 6-18: Two groups of 100 patients have been subjected to a medical test. People of group A were injected with recovery serum, but people of group B were not injected with serum. These two groups were under equal conditions in terms of other aspects. From groups A and B, 75 and 65 people have recovered, respectively. Has this serum been effective in the recovery above? Write down the hypotheses to be tested and test them at the significance levels of ten, five and one

percent. Discuss the results and state which of the above significance levels is better.

P_1 = A proportion of society that injection serum

P_2 = A proportion of society that does not injection serum

$$H_0: P_1 = P_2$$

$$H_1: P_1 \geq P_2$$

$$\hat{p} = \frac{75+65}{200} = 0.7. \quad \hat{q} = 1-0.7 = 0.3$$

$$\delta_{P_1-P_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.7)(0.3)\left(\frac{1}{100} + \frac{1}{100}\right)} = 0.0648$$

$$Z_C = \frac{P_1 - P_2}{\delta_{P_1, P_2}} = \frac{0.75 - 0.65}{0.0648} = 1.54$$

$$a) Z_c = 1.54 < Z_{0.01} = 2.326$$

$$b) Z_c = 1.54 < Z_{0.05} = 1.648$$

$$c) Z_c = 1.54 < Z_{0.1} = 1.282$$

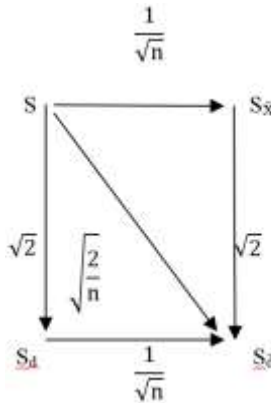
According to the results, the null hypothesis is rejected at 10 percent confidence level and accepted at 5 and 1 percent confidence levels. In other words, there is a 90% probability that this serum is effective in the recovery of people, but there are not enough reasons to adopt such a result with 95 or 99% certainty.

The mean and standard deviation of the ratio distribution is equal to $\frac{1}{n}$ of the mean and standard deviation of the binomial distribution, because each ratio is $\frac{1}{n}$ of frequency.

The confidence limits of the ratio difference are calculated as follows:

$$P\{(p_1 - p_2) - Z \delta q_1 - q_2 \leq p_1 - p_2 \leq (p_1 - p_2) + Z \delta q_1 - q_2\} = (1 - \alpha)\%$$

The relationship between the formulas of standard deviation and variance between means:



Here, a brief explanation about variance analysis is given. For further study, refer to specialized books on experimental designs.

6.23. Analysis of variance

ANOVA is a statistical method used to analyze the differences between the means of several groups. In the analysis of variance (ANOVA), the changes between the averages of the groups (between-group changes) are compared with the changes in the scores within the

groups (intra-group changes). Here, a ratio called F is calculated, if it is significant, it indicates that there is a difference between the compared means, and the post hoc test is used to find the location of the differences. If we want to compare the average of several groups, we use the analysis of variance test. For example, suppose we want to compare the sameness of physical fitness in 3 age groups. In this case, if it is shown that no group is different from another group, then our work is finished and the groups are similar. But if it is shown that one group is different from another group, it should be checked which group it is different from. For this purpose, using post hoc tests (such as Tukey's test), pairwise comparisons are made, that is, groups are compared two by two. The number obtained from paired comparisons should be compared with a standard number called 0.05. If the obtained number is greater than 0.05, it is not significant and if it is less than 0.05, the difference is significant.

Variance analysis is based on the analysis of known and unknown factors that explain the dispersion of scores.

It should be known that the total dispersion or variance of a set of scores is divided into regular variance and error variance.

Regular variance: The regular variance is that part of the variance or dispersion that is caused by an independent variable or a factor that moves the scores in a certain direction.

Error variance: The error variance is that part of the total variance that is caused by random changes in measurement.

Variance analysis deals with the contribution of regular variance and error variance.

It should be known that the difference between the averages is determined by calculating the F ratio. The main purpose of the F ratio is to compare the dispersion between means and the dispersion between the scores of individuals within groups.

To use variance analysis, it is necessary to observe the following assumptions:

The measurement scale should be at least at the interval level.

The sample is randomly selected.

Samples should be selected from a population that has a normal distribution.

Community variances should be equal.

In situations where the assumptions of variance analysis (for example, the condition of normal distribution in numbers with an interval scale or the condition of homogeneity of variances) are not met for any reason, it is better to use non-parametric tests equivalent to this test.

In variance analysis, if the calculated F value is lower than the F value of the table, the null hypothesis is confirmed, and if the calculated F value is greater than the F value of the table, the null hypothesis is rejected.

In the analysis of variance, the F ratio tells us that there is a significant difference between at least two means, but it does not specify which groups these means belong to. If F is significant, the researcher uses post hoc tests (for example, Tukey's test) to know which group's average difference is significant. With Tukey's test, it is possible to compare only one pair of averages, and this estimate is more

appropriate when the sample size is equal. For this reason, in situations where the researcher intends to compare averages with unequal volumes or a combination of them, the Scheffe method is the most appropriate test. The disadvantage of this method is that it is cautious or conservative.

Scheffe test due to:

Ease of calculation, its use in groups with unequal size and insensitivity to the assumption of normality of the distribution of the studied trait, have a lot of validity and use.

In the analysis of variance, the post hoc test is used if the F ratio is significant in at least two groups.

In the analysis of variance, when the null hypothesis (which negated the differences) is rejected and there is a difference between the means, post hoc tests are used to investigate the location of the differences.

A posteriori test of Scheffe and Tukey differs in the way of determining the critical number. In Scheffe method, the critical value is calculated directly, and in Tukey's method, it is extracted from the table.

Follow-up tests include Tukey's, Duncan's, Scheffe, Bonferroni's and....

The most conservative and famous post hoc test is Scheffe's test, which detects less significant differences. That is, it rejects the null hypothesis less. Duncan's method is the most liberal method that detects more significant differences. Scheffe's method has a fixed critical value for the post hoc comparison of all means when the F ratio obtained from the analysis of variance is significant. This method controls type 1 error

(increased alpha) for each number of comparisons. The following formula is used to calculate the critical difference (necessary difference value) for significance.

$$C.V = sch = \sqrt{(k - 1) \times F_{\alpha}}$$

sch = critical Scheffe.

K = number of means to be compared.

F_{α} = critical F ratio based on selected alpha.

The F ratio is used to estimate the population variance, derived from the between-groups data (MS_b), or the estimate of the population variance, derived from the within-groups data (MS_w), and is defined as:

$$F = \frac{MS_b}{MS_w}$$

MS_b is affected by differences between groups.

MS_w indicates the error caused by the dispersion within the groups.

MS_b (between-group mean of squares) and MS_w (within-group mean of squares) of the total sum of squares are calculated using the following relations.

$$MS_b = \frac{SS_b}{k-1}$$

$$MS_w = \frac{SS_w}{N-K}$$

$$SS_t = SS_b + SS_w$$

As the difference between the groups increases, the F ratio increases.

In analysis of variance, the degrees of freedom between groups (face) are calculated from the formula $k-1$, where k is equal to the number of groups. Also, degrees of freedom within groups (denominator) are calculated from the $N-K$ formula, where N is the total number of scores and K is the number of groups.

6.24. Degrees of freedom

Degrees of freedom means the number of subjects, minus the number of parameters that we intend to estimate, for example, a sample with size n that we intend to estimate the average of the population. The degrees of freedom are $n-1$. Because the community average is considered as a parameter.

6.25. One-factor analysis of variance

It is a statistical method that is used to investigate the differences between the averages of two or more different groups. For example, to examine the difference between the flexibility of four independent groups, one-way analysis of variance (one-factor) is used.

6.26. Multivariate analysis of variance

It is a method in which the effects of two or more factors on the dependent variable are simultaneously determined. For example, factorial variance analysis is used to investigate the effects of different exercise programs (action A) and gender (factor B) on strength (dependent variable).

6.27. Analysis of variance with repeated measures

It is a method in which a dependent variable (one subject) is exposed to more than one independent variable. For example, variance or repeated measurement is used to investigate the effect of exercise type on motor performance. In this example, factor A (intergroup) is the type of training (blocked, chain, random) and factor B (within group) is the individual's performance over a period of time in the pre-test, acquisition test, and memory test. Factor A is a between-group factor because each group consists of a different number of subjects, while factor B is a within-group (repeated measures) factor. Because all subjects are tested three times.

Analysis of variance test with repeated measurement is used to determine the effect of the independent variable both in independent groups and in dependent groups.

The advantage of repeated measurement is that a group of subjects are repeatedly tested and in this way they control their own work.

One of the interpretations of the null hypothesis of $H_0: \mu_1 - \mu_2$ is that this test, in addition to comparing two averages, also shows the equality or inequality of variances. In other words, can two samples be samples from one society or two societies with equal variances? The above test is performed using the t distribution and the following formula:

$$t = \frac{(\bar{x}_1 - \mu) - (\bar{x}_2 - \mu)}{S_{\bar{x}_1 - \bar{x}_2}}$$

And if the above formula reaches the power of 2, the result is the ratio of two variances, which according to the definition of the statistical criterion is F.

$$t^2 = \frac{\{(\bar{x}_1 - \mu) - (\bar{x}_1 - \mu)\}^2}{S_{\bar{x}_1 - \bar{x}_2}^2} = \frac{\sum(\bar{x}_1 - \mu)^2}{S_{\bar{x}_1 - \bar{x}_2}^2}$$

Therefore, t^2 is equal to F with degrees of freedom of one (for the numerator of the fraction) and $n-1$ (for the denominator of the fraction). $t^2_{\alpha,df} = F_{\alpha,(1,df)}$ is the sum of the sum of the squares of the variances in the numerator and the denominator of the fraction F (the sum of squares between groups (SSB) and the sum of squares within groups (SSW) is equal to the sum of the squares of the deviations of each of X_{ij} values is from the total mean and it is called total sum of squares (SST).

The mean square between the groups is called the treatment variance and the mean square within the groups is called the error variance.

Comparison of variances and general principles of variance analysis

In many cases, the comparison of the averages of more than two societies has been considered, and the calculations are as follows.

1- Mean square within groups (MSW)

$$S^2_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \dots + (n_t-1)S_t^2}{(n_1-1) + (n_2-1) + \dots + (n_t-1)}$$

In general, the degree of freedom (the denominator of the above fraction) is equal to $(n_1 + n_2 + \dots + n_t - t)$ and if $n_1 = n_2 = \dots = n_t$ then it is equal to $n_t - t$ or $t(n-1)$ is.

$$S^2_p = \frac{SSW}{t(n-1)} = MSW$$

2- The variance between the average samples ($\delta^2_{\bar{X}}$)

$$\delta^2_{\bar{X}} = \frac{\sum^t (\bar{x}_i - \bar{x}_{..})^2}{t-1}$$

And since $\delta^2_{\bar{X}} = \frac{\delta^2}{n}$, $n \delta^2_{\bar{X}}$, the other estimate will be δ^2 .

$$nS^2_{\bar{X}} = \frac{n \sum^t (\bar{x}_i - \bar{x}_{..})^2}{t-1} = \frac{SSB}{t-1} = MSB$$

And the total sum of squares (SST) is equal to:

$$SST = SSB + SSW$$

The variance analysis table is as follows.

S.V	df	Sum of squares	Mean square	F
between groups (treatment)	t-1	$n \sum (\bar{x}_i - \bar{x}_{..})^2$	MSB	$\frac{MSB}{MSW}$
Inside groups (error)	t(n - 1)	$\sum \sum (\bar{x}_{ij} - \bar{x}_i)^2$	MSW	
Total	tn - 1	$\sum \sum (\bar{x}_{ij} - \bar{x}_{..})^2$		

Example 6-19: Assume that two wheat varieties A and B are randomly planted in five plots and the following data are obtained after measuring one trait from each plot. Are the two varieties different in terms of this attribute or not?

Variety	Observations for each variety					$\sum X_j$	\bar{X}_j
A	19	14	15	17	20	85	17
B	23	19	19	21	18	100	20
						$\sum X_{ij}=185$	$\bar{X}_{..}=18.5$

This investigation can be done through the F test. The null hypothesis and the opposite hypothesis are:

$$H_0: \mu_A = \mu_B$$

$$H_1: \mu_A \neq \mu_B$$

First, the sum of squares of different sources is calculated:

$$SS_G = \sum \sum (\bar{x}_{ij} - \bar{x}_{..})^2$$

$$= (19-18.5)^2 + (14 -18.5)^2 + \dots + (21 - 18.5)^2 + (18-18.5)^2 =$$

64.5

$$SS_T = n \sum (\bar{x}_i - \bar{x}_{..})^2$$

$$= 5 \{ (17-18.5)^2 + (20- 18.5)^2 \} = 22.5$$

$$SS_W = \sum \sum (\bar{x}_{ij} - \bar{x}_i)^2$$

$$= (19 - 17)^2 + (14-17)^2 + \dots + (20 - 17)^2 + (23 -20)^2 + (19 - 20)^2 + \dots + (18 - 20)^2 = 42$$

Considering that the total sum of squares is composed of the sum of within-group and between-group squares, the simpler way to calculate the intra-group sum of squares is the total sum of squares minus the sum of between-group squares, that is, here:

$$SS_W = SS_G - SS_T = 64.5 - 22.5 = 42$$

Now form the variance analysis table:

S.V	df	SS	MS	F
between groups (treatment)	t -1= 2-1 = 1	22.50	22.50	4.29
Inside groups (error)	t(r -1)= 2(5-1) = 8	42	5.25	
Total	tr -1= 10 -1 = 9	64.50		

The numerator variance means the variance between the varieties has 1 degree of freedom and the denominator variance means the

variance within the varieties has 8 degrees of freedom. The F table number at the 5% probability level for 1 and 8 degrees of freedom is 5.32. Because the calculated F, which means 4.29, is smaller than the F in the table, which is 5.32, then the null hypothesis is true and there is no significant difference between the varieties in terms of the measured trait. So two varieties belong to the same statistical population.

6.28. Statistical models

A statistical model is a linear function that shows the relationship between the components of statistical data by mentioning their assumptions.

Statistical model in one-way classification:

$$X_{ij} = \mu + \alpha_i + e_{ij}$$

Statistical model in two-way classification:

$$X_{ijk} = \mu + \alpha_i + p_j + e_{ijk}$$

6.29. Means comparison method

If the F treatment or groups become significant, it means that there is a significant difference between the two treatments. One of the methods of comparing means is the least significant difference test or LSD.

$$LSD = t_{\alpha/2, df} S_{\bar{d}}$$

$$S_{\bar{d}} = \sqrt{\frac{2MSE}{n}}$$

In the following cases, the LSD test is used.

1- In cases where a witness or control is used in the experiment and the purpose is to compare the rest of the samples or groups with this witness.

2- In cases where the experimenter has considered some special comparisons that are important to him before conducting the experiment, and after conducting the experiment, only those comparisons are made.

6.30. Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a non-parametric test that enables the researcher to compare and test the degree of agreement between the distribution of observed numbers with a theoretical distribution.

The assumptions of this test are:

- 1- The basic dimensions or columns used must be continuous.
- 2- The scale should be nominal.

If the significance test of the correlation coefficient is performed using the t distribution. In this case, the t ratio in this test is calculated using the following formula.

$$t = r \sqrt{\frac{n - 2}{r - r_{xy}^2}}$$

Exercises of chapter 6

- 1- If we have a frequency distribution table with k categories or classes, what will be the degree of freedom for the normality test?
- 2- What is the degree of freedom in a $C \times R$ consensus table?
- 3- In a sample of 50 individuals, if the mean of the sample is equal to 12 and its variance is 1000, the probability of this sample occurring in the t distribution, assuming $\mu = 0$, is the probability level closest to what percentage?
- 4- What is common in the t test?
- 5- From a sample of 5, the 95% confidence level of the average weight of apples in a shipment is 128.9 to 171.1 grams. How many grams was the average weight of apples in this sample?

$$P(128.9 \leq \mu \leq 171.1) = 95 \%$$

- 6- Suppose the value of t with 8 and 9 degrees of freedom is equal to 3.355 and 3.250, respectively. If the values of $\bar{x} = 1$ and $S = 3$ are obtained in a sample of 9, what is the distance estimate from the average of this population?
- 7- 40% of all university students have televisions. If a random sample of 100 students is selected, what is the probability that the proportion of students with televisions is between 32% and 47%?

$$P(Z \geq 1.43) = 0.0764$$

$$P(Z \geq 1.96) = 0.025$$

$$P(Z \geq 1.63) = 0.0516$$

$$P(Z \geq 1.64) = 0.05$$

- 8- From a sample of 9 oranges, the average weight of oranges in a shipment is 125 to 170 grams with a confidence of 95%. What is the mean and standard deviation of 9 samples of oranges from right to left?

0.05	0.05
$t = 2.306$	$t = 1.86$
Table	
Two-sided	One sided

9- A researcher believes that the standard deviation of the concentrator in his laboratory is equal to 2. In measuring a solution with a certain concentration, the values of 1.4, 2.5, 2.10 have been recorded. Can the researcher's claim be rejected at the 5% probability level?

10- If it is assumed that the height of men has a normal distribution with a standard deviation of 2.5 inches, how many sizes should the sample be selected so that with a 99% probability the absolute value of the sample mean does not differ from the true community mean by more than 0.5?

11- A group of 24 men and 16 women were asked to express their opinion about transgenic products. The results of the investigation were as follows:

	Non-transgenic	Transgenic
Men	16	8
Women	14	2

12- Two normal communities with equal variance $\delta^2 = 4$ have averages of 66 and 65.5. What is the probability that the mean of a sample of 50 from the first population is greater than the mean of a sample of 50 from the second population?

$$Pz < - 1.25 = 0.1056$$

$$Pz > 1.25 = 0.1093$$

$$Pz > 0.16 = 0.4364$$

13- A psychologist has obtained a standard deviation of 0.05 by studying students. This psychologist wants to determine the average IQ of the society by selecting a sample of people in such a way that the maximum error of estimation does not exceed 0.01 with 99% confidence, the required number of samples is equal to:

$$P(Z \geq 2.58) = 0.005$$

$$P(Z \geq 1.96) = 0.025$$

14- Out of 64 *Drosophila* flies, by observing the minimum and maximum number of male flies, it is possible to infer with 95% confidence that the probability of being male and female is equal? ($Z = 2$)

15- What is the degree of freedom for testing the normality of 100 data using a frequency distribution table with 8 classes (category or class)?

16- To compare the average of 6 groups each with 3 repetitions, how much is the degree of freedom between groups and within groups, respectively?

17- A gardener has sent 1000 boxes of pears to a compote factory. If each box contains 100 pears and 10% of them are damaged, how many boxes are expected to contain less than 83 healthy pears?

$$P(Z \geq 2.23) = 0.0099$$

$$P(Z \geq 1.05) = 0.146$$

18- To compare the effect of a type of hormone in increasing the flowering of tomatoes, 10 pots were randomly selected and plants

planted with the hormone were sprayed and the other half was considered as a control. What is the degree of freedom to refer to table t?

19- In an experiment, the number of claws was studied in two indigenous groups of wheat. For this purpose, a random sample including 5 plants in each group was selected and their average was estimated as $\bar{x} = 12$, $\bar{y} = 6$. If the variance of the number of claws per plant in both wheat masses is the same and equal to 2.5, what is the numerical value of the t statistic for the test of the null hypothesis $H_0: \mu_1 = \mu_2$ against the one hypothesis $H_1: \mu_1 \geq \mu_2$?

CHAPTER 7

THE RELATIONSHIP BETWEEN VARIABLES

Assist. Prof. Dr. Mohsen MIRZAPOUR¹

Dr. Saeid HEYDARZADEH²

Dr. Harun GĪTARĪ³

¹ - Siirt University, Faculty of Agriculture, Department of Agricultural Biotechnology, Siirt, Türkiye

ORCID ID: 0000-0002-2898-6903, e-mail: m.mirzapour@siirt.edu.tr

² - Former Ph.D. Student of Urmia University, Faculty of Agriculture, Department of Plant Production and Genetics, Urmia, Iran

ORCID ID: 0000-0001-6051-7587, e-mail: s.heydarzadeh@urmia.ac.ir

³ - Kenyatta University, School of Agriculture and Enterprise Development, Department of Agricultural Science and Technology, Nairobi, Kenya
ORCID ID: 0000-0002-1996-119X, e-mail: harun.gitari@ku.ac.ke

INTRODUCTION

Many daily occurrences and natural phenomena are subject to some factors and their quantity. One of the main goals in scientific research is to discover relationships between phenomena, so calculating and analyzing the relationship between variables is of particular importance. The meaning of random variable in agricultural experiments is that their values are influenced by uncontrolled and unknown factors and sampling and measurement errors.

In general, regression and correlation methods are used to determine the presence or absence of a relationship between variables. The correlation index shows the relationship between two variables that are both affected by common factors. Therefore, the correlation index is used to measure and determine the degree of mutual relationship between the changes of two random variables. The regression index is used to determine the cause and effect relationship between the changes of a random variable and a fixed variable. In other words, in regression, the amount of one variable (function) is determined by each unit of change in another variable (constant).

In this chapter, the bivariate community is examined. But in some cases, it is desired to determine the relationship between a variable and a set of variables, and in these cases, regression and multiple correlation are used, and in this case, the indicators that show the relationship between the variables, in the case where it is calculated that other variables have an effect on the amount Do not have a relationship.

7.1. Scatter plot

If we consider each ordered pair (x_i, y_i) as the coordinates of a point and plot them on the coordinate plane, the scatter plot will be obtained. A scatter diagram can provide us with three types of information.

Whether or not there is a pattern that indicates some kind of relationship between the observations

If there is some kind of relationship, is it linear or non-linear?

If the relationship is linear, what is the type of relationship?

7.2. Linear regression

Although correlation analysis is closely related to regression, conceptually, these two analyzes have significant differences with each other. In correlation analysis, the primary goal is to measure the degree of linear relationship between two variables. Instead, we are trying to estimate or predict the average value of a variable based on the fixed values of other variables. Basically, regression analysis consists of finding the relationship between two variables, one of which is a function of the other. A variable that is a function or dependent is indicated by y and is also placed on the y axis in the coordinate system. The second variable that determines the first variable is called the independent variable, it is indicated by x and it is placed in the coordinate system on the x axis. In this way, with regression analysis, we try to find the most suitable relationship between these two variables. Obviously, the relationship makes sense where there is a logical connection between y and x , or in other words, the changes in y

depend on the changes or values of x . Instead of relationship, other words such as equation, model or function are also used, because the regression analysis for the relationship between y and x finally leads to an equation or model that shows the dependence of y on x in a quantitative way. Now the question is, which variable is y and which variable is x in the experiments, y consists of variables such as grain yield, the amount of weight gains of livestock or poultry during the breeding period, the amount of milking of livestock, the amount of dry matter production by plants, the amount of tree yield, etc. These are the ones that are actually interesting to the experimenter and their changes or magnitudes are important. But, x is a variable that affects y and causes changes or up and down of y . The levels of the factor investigated in the experiment or the treatments such as fertilizer can be considered as x . In fact, in the application of regression analysis as a postvariance method, the treatments are always x . In order to check the equation or model of relationship between y and x , it is briefly explained in steps.

The first step is to plot the values of y and x against each other in a figure. The resulting figure is called a data distribution diagram. Figure 7-1A is an example of a scatterplot of y versus x data. In the first step, by seeing or examining the distribution diagram, one must guess which equation or model best explains the dependence of y on x . The appropriate equation or model is the one that fits well between the points in the scatter plot of the data. At this stage, the experimenter may come up with different equations or modes that he thinks are appropriate. Unfortunately, regression analysis does not tell us which

equation to choose for the data, so prior knowledge of different equations or models, or the help of a familiar person, must be available. Suppose at this stage, seeing the scatter diagram in Figure 7-1a, we think that a straight line model or a 2nd degree model is suitable.

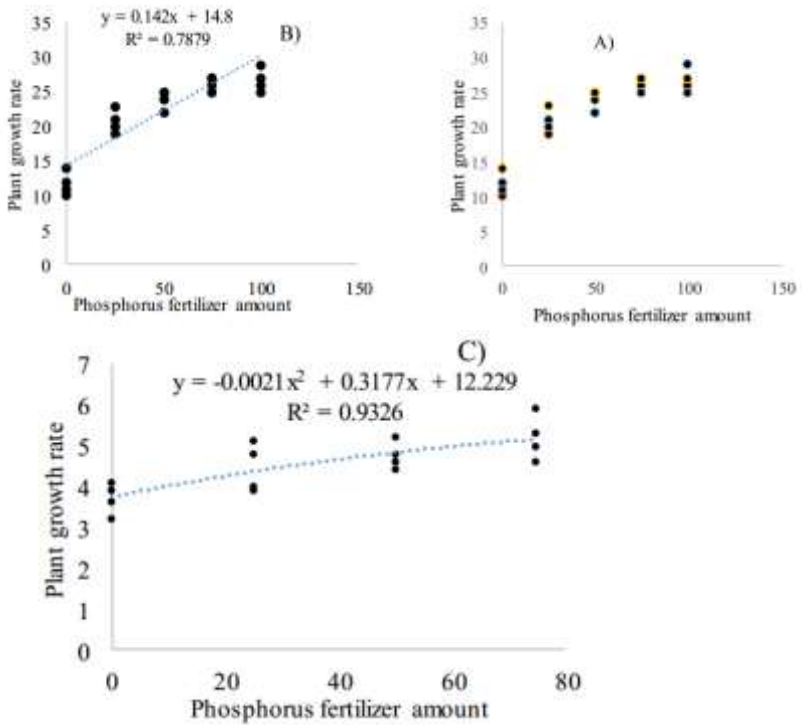


Figure 7-1: a) data distribution diagram, b) fitting a straight line model and c) fitting a 2nd degree model

The second step is to fit the model to the data. In this step, the number and digit are found for the coefficients or parameters of the model. As we know, the straight line model is:

$$\hat{y} = b_0 + b_1 X$$

And the 2nd grade model consists of:

$$\hat{y} = b_0 + b_1 X + b_2 x^2$$

In the above equations or models, b_0 and b_1 in the straight line model and b_0 , b_1 and b_2 in the quadratic model are the coefficients of the model. If we want to pass a straight line through the data in Figure 1-7A, infinitely many straight lines can be crossed or fitted. This method yields the best possible straight line and estimates values for its coefficients. If we pass a quadratic model from the data in Figure 7-1 A, the quadratic line can be crossed infinitely, and here again the method of least squares can pass the best quadratic line and determine its coefficient. In the data of Figure 7-1, for the coefficients b_0 and b_1 in the straight line model, the values are estimated as 14.8 and 0.142, and for the coefficients b_0 , b_1 and b_2 in the quadratic model, the values are 12.23, 0.348 and - 0.0021 have been estimated.

The third step is to check whether it fits. It is not enough to simply find values for the coefficients of a model or equation. It should be checked whether the fitted model justifies the changes of y versus x well enough or not? For this purpose, in the regression analysis, the changes of y around the mean of y (i.e. \bar{y}) are calculated and divided into two parts: one is the changes due to the relationship of y with x (set of regression squares) and the other is the changes due to the influence of other factors (set of residual squares or error). Then, the mean square of the regression and the residual are calculated and the F test is performed. A good model should lead to a significant F, but it is not enough that the F is significant. At this stage, statistics such as R^2 and C.V are also calculated, and the residuals chart is examined to determine whether the fitted model is appropriate and acceptable. By

comparing Figure 1-7b and Figure 1-7c, we can see that the quadratic model is more suitable.

As you can see, for the right line model, the points at the beginning of the graph (related to low values of X) are all below the line, in the middle of the graph, the points are located above the line, and again at the end of the graph, they are below the line corresponding to the model.

The fourth step is to use the resulting model, its coefficients and the wrong values of coefficients and \hat{y} for further analysis. The regression analysis model can be used for different purposes. These uses or applications are: 1) description, 2) prediction, 3) parameter estimation, and 4) control.

There are different ways to describe the relationship or dependence between two variables (such as x and y). Regression analysis summarizes this relationship by presenting an equation or model. Using regression analysis as a postvariance method is actually an example of this application and its expansion.

In forecasting, the equation obtained from regression analysis is used to estimate the value of y for a certain value of x. For example, if y is the yield of a plant and x is the amount of nitrogen fertilizer, the resulting equation can be used to predict the amount of yield when, for example, the amount of fertilizer is 50 kg.

Parameter estimation with the help of regression equation is in the condition that the coefficients of the equation are special characteristics of the studied system. For example, if x are different amounts of water and y is the yield of the plant, then the coefficient b_1 in the equation of

the straight line describing y against x is the efficiency of water consumption, which can be estimated with the help of the regression model.

In control, the regression equation is used for inverse prediction. For example, suppose in a study the relationship between the degree of fabric strength (y) and the percentage of cotton in it (x) is obtained by regression analysis. Now, if we use this equation to see how much cotton percentage should be in order to obtain a fabric with a certain resistance, we have actually used the regression model to control (fabric strength).

7.3. Application of regression to find optimal treatment

In regression analysis to find optimal treatments, x is always equal to the levels of the tested factor, which is small, and y can be any of the variables measured in the experiment, such as the amount of product, the amount of weight gain, and so on. However, how the regression equation or model is analyzed and interpreted to obtain a favorable result regarding the treatments depends on the model fitted to the data.

7.4. Considerations in using regression to find the optimal treatment

In this section, there are some points that will be useful to pay attention to in applying the regression method as a postvariance analysis:

If the experimental treatments are small and regression analysis is to be used, it is necessary to have sufficient number of treatments or levels of the tested factor. At least 2, 3 and 4 points are needed to fit 1st, 2nd and 3rd degree models. If the number of treatment levels is, for example, 3 or 4, it will not be possible to use many regression models. It is also important how to choose levels. Levels should be chosen to cover the full range of the value of the factor under consideration. For example, when fertilizer treatments include only small amounts, it is not possible to find the optimal treatment.

In the situation where the factor under investigation is quantitative, it can be reduced from the number of test repetitions and added to the number of levels of the factor. It is even possible to use one repetition for each treatment. Of course, in this situation, let's analyze the changes of the desired variable (y) against the desired factor (x) with the regression model.

In applying regression analysis as a postvariance analysis, there are often multiple replications for each treatment. Now, an important question is whether the regression equation should be fitted to the total points including repetitions or whether the average of each treatment should be calculated and the regression equation should be fitted to the average of the treatments. The main point is that in both cases the regression equation is obtained with the same coefficient and in both cases the conclusion regarding the desired treatment will be the same. But, in the situation where repetitions are also used, the regression coefficients of the standard error are smaller, the R^2 value is lower and the C.V value is higher. In the conditions of using repetitions instead of

averages, it will be possible to use the lack of fit test, which is not possible if averages are used.

Sometimes, when choosing the right regression model, there is no need to compare different models, but the right model can be chosen according to the theory of the tested problem. The theory of the problem is related to the subject of the experiment and is not related to statistics. For example, if the treatments consist of different amounts of the same input (such as fertilizer), the asymptotic exponential regression model or the monomolecular model can be used.

Sometimes there are two factors in the experiment, and the purpose of the study is to investigate the different levels of these two factors on the desired variable. If both of these factors are small, then we are dealing with regression relationships in which the relationship of y with more than one x is examined. For example, suppose in an experiment, the effect of negative amounts of nitrogen fertilizer as well as phosphorus fertilizer on wheat yield has been investigated. In this case, wheat yield y and nitrogen fertilizer are x_1 and phosphorus fertilizer is x_2 , and the resulting regression equation will not actually describe a line, but it will describe a plane in the coordinate system with three axes. Regression decomposition in this situation to find the optimal values of x_1 and x_2 is called reaction surface decomposition.

7.5. Basics of regression with simple linear model

In this part, the principles of regression are explained with a simple example. The regression model that will be fitted to the data is a simple linear model (straight line), which is the simplest regression

model. Of course, the discussed principles can be generalized for other regressions in the same way.

In an experiment, the effect of 4 amounts of nitrogen fertilizer on wheat grain yield has been investigated. The data are given in Table 7-1 and the results of analysis of variance are included in Table 7-2. The adequacy and correctness of variance analysis was also checked and it did not indicate any significant problem. The results of the analysis of variance indicate a significant difference between the levels of the desired factor, and because these levels are small, we intend to use regression analysis to further investigate the treatments and find the appropriate fertilizer treatment.

Table 7-1: test data on the effect of nitrogen fertilizer on wheat yield

Repetition	Fertilizer treatment			
	0	25	50	75
1	3.6	4.8	4.4	5.3
2	4.1	5.1	5.2	5.9
3	3.2	4.0	4.6	4.6
4	3.9	3.9	4.8	5.0

Table 7-2: results of analysis of variance of the example of nitrogen fertilizer on wheat yield

Sources of variation	SS	df	MS	F
Treatment	4.77	3	1.59	6.91**
Experimental error	2.76	12	0.23	
Total	7.53	15		

$$R^2 = 0.63$$

$$CV = 10.6$$

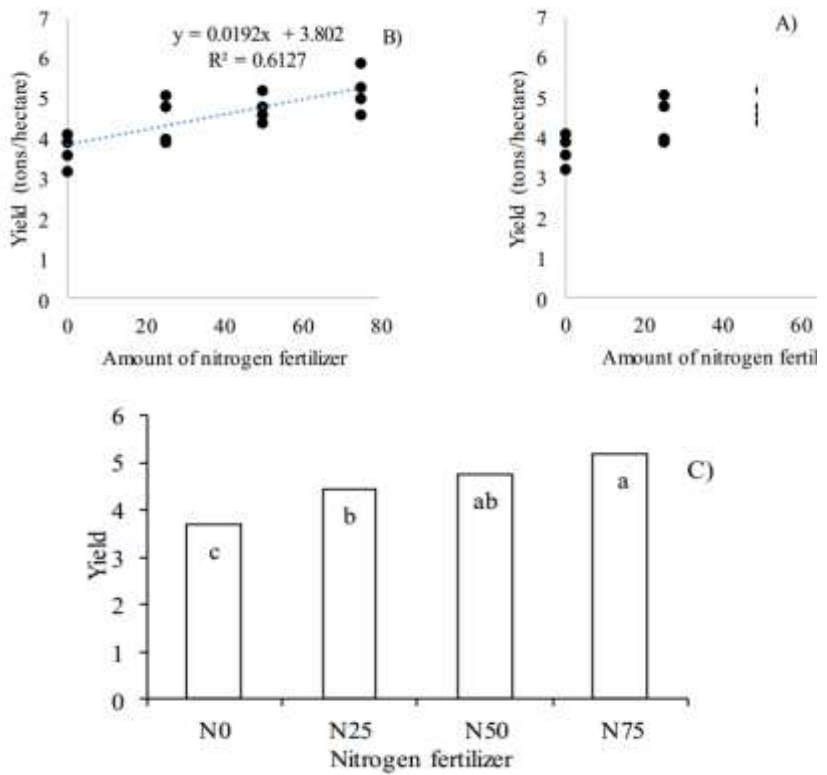


Figure 7-2: a) Data distribution diagram b) Fitting the 1st degree model to changes in wheat yield under the application of different levels of nitrogen fertilizer c) Comparing the average nitrogen fertilizer treatments

7.6. The steps of regression and relevant conclusions are as follows:

1- Examining the data distribution chart: Figure 7-2 shows the test data distribution chart. It seems that the yield increased linearly with the increase in the amount of fertilizer. Therefore, we decide to fit a straight line model to the data.

2- Fitting the model with the method of the least square powers: a large number of straight lines can be passed between the points of Figure 7-2. The method of the least square powers of the straight line passes between the points so that the sum of the square deviations of the points from the line (S) is minimal. Figure 7-3 shows a sample of data and a straight line passing through the points. Each point of the regression line has a deviation or a distance, and in this figure, these deviations are shown as vertical lines that connect the points to the regression line.

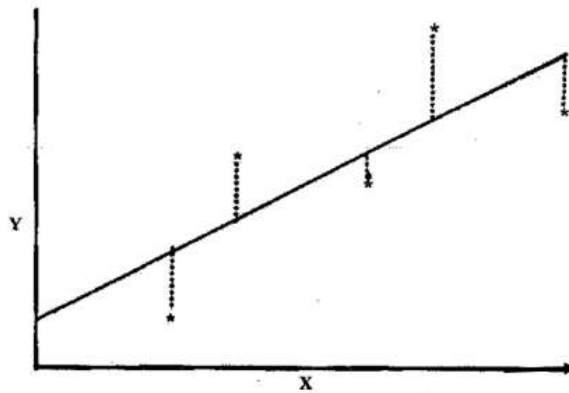


Figure 7-3: The method of the least squares sum of square deviations from the line

If we denote each point by y and its predicted value for the same value by \hat{y} , for each point, the value of deviation is equal to:

$$e = y - \hat{y}$$

Since we are dealing with simple linear regression and

$$\hat{y} = b_0 + b_1 X$$

So we can write:

$$e = y - (b_0 + b_1 X)$$

And the value of S is equal to:

$$S = \sum e^2 = \sum (y - b_0 - b_1 X)^2$$

In the method of the lowest second power (which is actually the lowest S), the values of b_0 and b_1 are obtained by deriving the component of S with respect to b_0 and b_1 and setting them equal to zero. The result of this action will be two equations known as normal equations (n is the number of points):

$$nb_0 + b_1 \sum x = \sum y$$

$$b_0 \sum x + b_1 \sum x^2 = \sum xy$$

And by solving the normal equations, b_0 and b_1 will be obtained:

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

In the above equations, S_{xy} is the sum of the product of deviations of x and y, and S_{xx} is the sum of the squares of the deviations of X from its mean. where S_{xx} is the sum of squares calculated for x. It can be shown that S_{xy} and S_{xx} can also be calculated from the following equations, which are easier to work with a calculator:

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 / n$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - [(\sum x)(\sum y)] / n$$

For the example data of nitrogen fertilizer in wheat, the value of b_0 and b_1 can be found with the following calculations:

$$\begin{aligned}
 n &= 16 & \sum y &= 72.4 & S_{xx} &= 12500 \\
 \sum x &= 600 & \bar{y} &= 4.525 & S_{xy} &= 240 \\
 \bar{x} &= 37.5 & \sum xy &= 2955 & b_1 &= 0.0192 \\
 \sum x^2 &= 35000 & b_0 &= 3.805 & &
 \end{aligned}$$

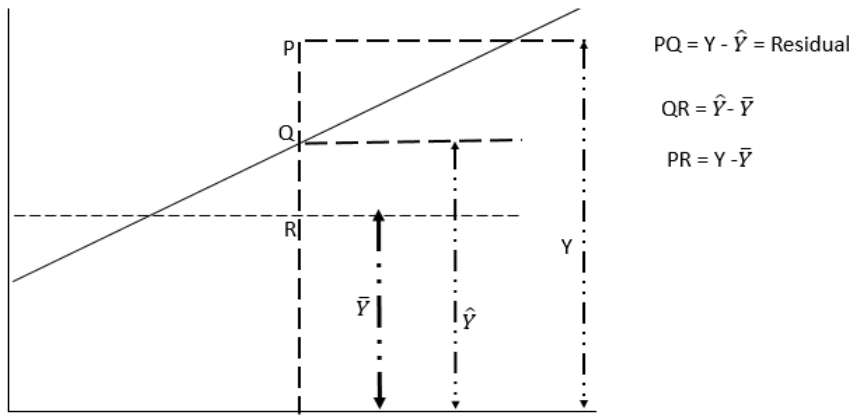


Figure 7-4: Relationship between observed, predicted, residual and average value y .

In this way, the values of b_0 and b_1 are calculated and the resulting straight line model will be:

$$\hat{y} = 3.805 + 0.0192x$$

In this model, b_0 is the amount of wheat crop when no nitrogen fertilizer is used ($x = 0$). The b_0 unit is the same as the y unit. b_1 is the slope of the line and is equal to the tangent of the angle of the regression line with the X axis. b_1 shows how much the yield increases for each kilogram of nitrogen fertilizer increase. In this example, for each

kilogram of nitrogen fertilizer (unit x), the yield has increased by 0.0188 tons per hectare (unit y), or in other words, 18 kilograms per hectare. Figure 7-2 b shows the data along with the fitted line. For this reason, it is not possible to determine the optimal amount of nitrogen from the equation. To determine the optimal amount of fertilizer, another test should be performed with higher amounts of nitrogen fertilizer. Basically, in any situation where the 1st degree model fits the data better, it means that the desired value of x (treatment) cannot be determined.

Figure 7-3 c shows the histogram of experimental treatments, where the mean of each treatment is given. On each column of the graph, one letter or (for the 50 kg treatment) two letters are written for each treatment, which are the same letters used to group the treatments using the LSD method. Based on this figure and if the method of multiple comparison of averages was used, the experimenter comes to the conclusion that the treatments of 75 and 50 kg of nitrogen fertilizer are desirable because they have produced the highest yield and there is no statistically significant difference from each other. On the other hand, because the 50 kg fertilizer treatment has achieved the same performance with a lower amount of fertilizer, it is better. By comparing Figures 6-7 b and c, it is easy to understand that the method of multiple comparison of the average has led to the wrong answer, because in the range of 0 to 75 kg of fertilizer, the yield has increased linearly, and it is possible that the optimal amount of fertilizer that gives the maximum yield gives more than 75 kg of fertilizer. Maybe even with the use of higher amounts of fertilizer, the yield will not increase, and for

example, the same 75 kg of fertilizer is the optimal treatment, but this is not known and its definitive determination requires another experiment.

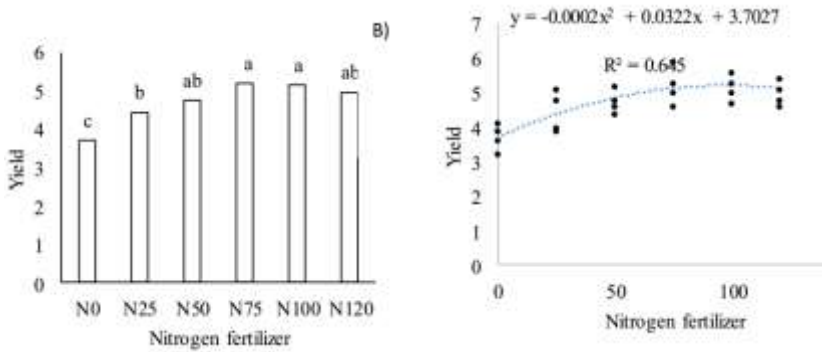


Figure 7-5: a) Fitting the 2nd order function to the yield changes in the nitrogen fertilizer treatments in the repetition of this experiment with higher fertilizer levels, b) Comparison of the average nitrogen fertilizer treatments.

For an example of the application of the 2nd degree model, pay attention to Figure 7-5 A. This figure shows the average yield of wheat plant for different treatments of nitrogen fertilizer (6 amounts, 0, 25, 50, 75, 100 and 120 kg of nitrogen per hectare), which was done in the form of a completely randomized design. You can assume that the previous experimenter repeated the same experiment with 2 more nitrogen fertilizer treatments. As can be seen, a 2nd order model fits the data well:

$$\hat{y} = 3.7 + 0.032x - 0.000176X^2 \quad (R^2 = 0.98, CV = 1.90 \%)$$

Based on the 2nd degree function, the maximum performance is located where the derivative of this model is equal to zero. We can use the same rule and find the optimal amount of fertilizer to maximize yield:

$$\hat{y} = b_0 + b_1 + b_2x^2$$

$$\dot{y} = b_1 + 2b_2x$$

$$b_1 + 2b_2x = 0$$

$$x_0 = -b_1/2b_2$$

In the above relations, \dot{y} is the derivative of y and X_0 is the value of X in which the derivative of the 2nd degree equation becomes zero. Now, by placing the values of the coefficients in the relation X_0 , we can calculate its numerical value:

$$x_0 = -0.032 \div (2 \times -0.000176) = 90.48$$

Therefore, we can conclude that to achieve the maximum yield of wheat in the conditions of this experiment, 90.5 kg of nitrogen fertilizer per hectare should be used. Figure 7-5 b shows the performance histogram for the treatments of this experiment and their grouping based on the LSD test. Based on the method of multiple comparison of averages, the optimal amount of fertilizer is 50 kg of fertilizer per hectare, which is the lowest fertilizer that has produced the highest yield, and the performance of this treatment with treatments of 75, 100 and 125 is not statistically significant. It is clear that this is a wrong conclusion from a wrong method.

Basically, in situations where the amount of fertilizer or water or other inputs constitute experimental treatments, it is incorrect to compare the average and find the amount of entities that have achieved

the highest performance. In this situation, the function or model of the performance reaction to the input should be obtained in order to determine exactly at what value of the input (X_0) the maximum performance is obtained. Of course, this amount of input that produces the maximum yield, such as the amount of 90.5 kg of nitrogen per hectare, is not recommended to farmers, but considering this amount and the cost of each fertilizer unit and increasing the yield of consecutive fertilizer units, the appropriate amount of fertilizer should be found. and is recommended. The appropriate amount of fertilizer will be determined according to the law of diminishing returns, and it is the amount of fertilizer that increases the yield of the last unit in terms of the price equal to the cost of the last unit of fertilizer. Sometimes, when the input used is scarce, such as pits or water, but the land is not too restrictive for producing products, other rules than the law of diminishing returns are used to find the appropriate amount of input. In this situation, the use of even lower amounts of inputs leads to higher production not per unit area but in the field. This is the basis of the low irrigation method.

3- Checking the significance of the regression and improving the fit: it is not enough to find the values of b_0 and b_1 and it is necessary to check whether the regression is significant or not and how well it fits the data. To check the significance of the regression, it is necessary to see how much the fitted regression equation justifies the changes of y values around the mean of y (i.e. \bar{y}). At this stage, variance analysis is actually used. The total changes in y are divided into two parts: 1-

changes related to the relationship between y and x (regression) and 2-residual changes or changes caused by other factors (error).

Figure 7-4 shows the deviation of a point from \bar{y} and its reasons. The desired point has a deviation from the average y ($y - \bar{y}$). Part of this deviation is due to the relationship between y and x: the desired point has a value of x greater than the average of x (i.e. \bar{x}) and due to the relationship of y and x, the value of this point or its numerical value is greater than the average of y (\bar{y}) has been This deviation \hat{y} from the corresponding mean (\bar{y}) can be expressed as $(\hat{y} - \bar{y})$. Another meaning of this article is that because this point has x greater than \bar{x} and due to the relationship between x and y, it is expected to be somewhere higher than the average y (\bar{y}) on the regression line (the line connecting y and x). But the desired point is not on the regression line, but rather higher. This point deviation from the regression line (ie $y - \hat{y}$) can be caused by other factors or random factors (error). The above content can be summarized as follows:

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

By squaring the sentences above and adding them from 1 to n, we will have:

$$\sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 = \sum (y - \bar{y})^2$$

$$SSG = SSR + SSE$$

in other words:

Sum of squares of deviations from regression (error) + sum of squares of deviations due to relationship with x (regression) = sum of squares of deviations of y around the mean

In the above equations, SSG is the total change in y or the sum of total squares, SSR is the total change due to the relationship of y with x, the sum of the regression squares, and SSE is the total remaining change due to random or unaccounted factors.

It can be shown that:

$$SSG = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = b_1 S_{xy} = b_1^2 S_{xx}$$

$$SEE = SSG - SSR$$

$$df_G = n - 1$$

$$df_R = p - 1$$

$$df_E = n - p$$

$$MSR = SSR / df_R$$

$$MSE = SSE / df_E$$

where n is the total number of points and p is the number of parameters in the regression model. The straight line regression model has two parameters (b_0 and b_1), so $p = 2$.

By dividing the regression variance by the error variance, the F value for the regression can be obtained:

$$F = MSR/MSE$$

By comparing the F of the regression with the F of the table at level α , the degree of freedom of the numerator equal to df_R and the degree of freedom of the denominator df_E , it is possible to find out

whether the regression is significant or not. If regression F is smaller than table F, it means that the relationship between y and X is not significant in the form of a straight line, or in other words, $b_1 = 0$. However, if the regression F is equal to or greater than the table F, it means that there is a straight line relationship between y and x.

$$SSG = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 335.14 - (72.4^2 \div 16) = 7.530$$

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = b_1 S_{xy} = 0.0192 \times 240 = 4.608$$

$$SEE = SSG - SSR = 7.530 - 4.608 = 2.922$$

$$df_G = n - 1 = 16 - 1 = 15$$

$$df_R = p - 1 = 2 - 1 = 1$$

$$df_E = n - p = 16 - 2 = 14$$

$$MSR = SSR / df_R = 4.608 \div 1 = 4.608$$

$$MSE = SSE / df_E = 2.922 \div 14 = 0.20871$$

$$F = MSR / MSE = 4.608 \div 0.20871 = 22.08$$

$$F_{0.01, 1, 14} = 4.600$$

$$F_{0.05, 1, 14} = 8.862$$

As it can be seen that the regression F is greater than the table's F at the 1% level, we can put two stars on the regression F, which means "significance at the 1% probability level". The results of variance analysis of regression analysis are included in Table 7-3.

To determine goodness of fit, there are various statistics, the simplest of which are coefficient of explanation (R^2) and coefficient of variation (C.V):

$$R^2 = 1 - \frac{SSE}{SSG} = \frac{SSR}{SSG} = \frac{4.608}{7.530} = 0.61$$

$$CV = \sqrt{MSE} / \bar{y} \times 100 = \sqrt{0.20871} \div 4.525 \times 100 = 10.10 \%$$

Table 7-3: results of analysis of variance related to the relationship between wheat yield and the amount of fertilizer based on the straight line model

Sources of variation	df	SS	MS	F
Regression (R)	1	4.608	4.608	22.08**
Error (E)	14	2.922	0.20871	
Total (G)	15	7.530		

The coefficient of explanation says that 61% of the total variation of y's around their mean was due to different amounts of fertilizer and the linear relationship between yield and amount of fertilizer. In other words, 61% of the total variation of y's around their mean is explained by the amount of fertilizer. The remaining 39% are caused by random factors and unaccounted factors. The coefficient of variation shows that the standard deviation is slightly more than 10 percent of the average, which is not a high figure, although it is not considered a desirable figure either.

4- Controlling the adequacy and correctness of regression analysis: As with variance analysis, there are conditions in regression analysis that if not met, the resulting regression will be flawed or wrong.

In short, the most important assumptions or regression conditions are that the residuals (e) have a normal distribution and are also random and independent. Another assumption or another condition is that there are no outliers in the data, and another important assumption is that the straight line model fits the data well or gives a good description of the relationship between y and x .

In short, in the normal probability diagram, the points should be located around the right line of this diagram, and if the deviations from this line are large, it indicates the existence of a problem, which is the non-normality of the residuals. Small and medium deviations are not important in this chart. In the diagram of the residuals, the existence of a pattern can indicate that the assumption of randomness and independence of the residuals is flawed. Also, if there are outliers in the data or the model does not fit the data, it will be shown in the residual plot.

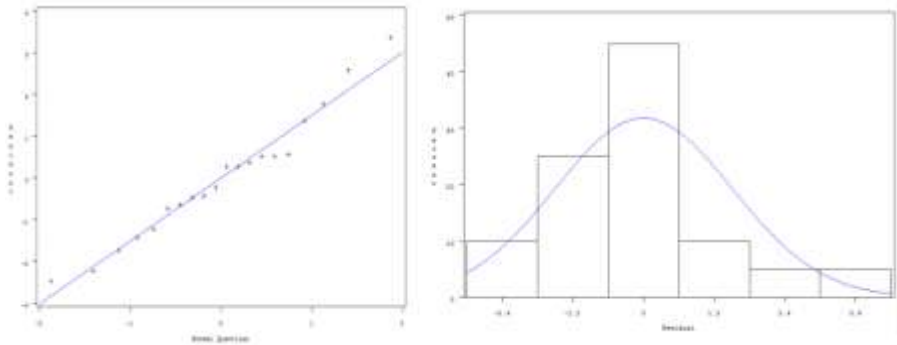


Figure 7-6: Normal histogram, normal probability versus predicted value

Figure 7-6 shows the normal histogram and normal probability against \hat{y} for example nitrogen fertilizer and wheat. These graphs do not indicate that the residuals are not normal or that they are not random. It also seems that there are no outliers between the data and a straight line model fits the data because the residuals do not have a pattern.

5- Assumption testing and confidence limits in regression: In regression analysis, it is possible to test different assumptions for the coefficients of the model as well as for \hat{y} and, if necessary, calculate confidence limits for them. In postvariance analysis with the help of regression analysis, these items are usually not needed, but they will be briefly explained.

To test the assumption about any parameter (such as b_0 , b_1 or \hat{y}), you can use the t-test as follows:

H_0 : Hypothetical value = parameter estimate

H_1 : Hypothetical value \neq parameter estimate

$$t = \frac{(\text{Parameter estimation}) - (\text{Default parameter value})}{\text{Standard error of statistic}}$$

The value of t calculated by the above formula is compared with the t of the specification table $t_{\frac{\alpha}{2}, n-p}$ and if the calculated t is equal to or greater than the t of the table, the null hypothesis is rejected. To use the above equation, the deviation of the regression coefficients and \hat{y} are needed, which can be calculated from the following relations:

$$\text{SEM}(b_1) = \sqrt{MSE/S_{xx}}$$

$$\text{SEM}(b_0) = \text{SEM}(\hat{y}) = \sqrt{MSE \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$$

It is worth mentioning that b_0 and \hat{y} are of the same genus and b_0 in simple linear regression is the same as \hat{y} when $x = 0$.

To calculate the confidence limits of b_0 , b_1 and \hat{y} , the following general relationship can be used:

Confidence limits = statistic $\pm (t_{\frac{\alpha}{2}, n-p})$ (standard error of statistic)

In the example of nitrogen fertilizer and wheat, the equality of b_1 with zero can be tested:

$$H_0 : b_1 = 0$$

$$H_1 : b_1 \neq 0$$

$$\text{SEM}(b_1) = \sqrt{0.20871/125000} = 0.00409$$

$$t = \frac{0.0192 - 0}{0.00409} = 4.70$$

$$t_{0.05/2, 14} = 2.145$$

$$t_{0.01/2, 14} = 2.977$$

Therefore, the slope of the line is not equal to zero and the null hypothesis is rejected. 95% confidence limits can also be calculated for b_1 :

$$\text{Confidence limits} = \pm 0.0192(2.145) (0.00409)$$

$$0.02797 \text{ upper limit}$$

$$0.01043 \text{ lower limit}$$

Conclusion: The regression analysis in this example shows that the relationship between wheat yield and the amount of fertilizer used is linear in the range of values from 0 to 75 kg per hectare, so that in the condition of not using fertilizer, the average yield is 3.805 tons per hectare and after It increases the yield by 0.0192 tons per hectare (equivalent to 19 kg per hectare) per kilogram of fertilizer used. This increase is statistically significant. Therefore, there is no optimal amount of nitrogen fertilizer in this range. It is necessary to repeat the experiment with higher values of nitrogen fertilizer in order to determine the optimal amount of fertilizer.

In the case of data from different experiments, the observations Y at given values of X never form a straight line. For each specific value of X, there is not a single value of Y, but rather a distribution of Ys, of which the observed value of Y is an example. Averages of Ys with different values of X may form a straight line. After estimating a and b, the equation of the hypothetical line can be written as follows:

$$Y_i = \alpha + \beta X_i$$

Having the above equation in hand, it is possible to predict the value of Y_i for each value of X. Considering that the estimates a and b are obtained in connection with the range of X changes under investigation, the prediction of Y_i should not be extended to X values higher and lower than the range of X changes available in the experiment. Because there may not be a linear relationship between Y and X in these values, and wrong results may be obtained.

The equation of this line shows the relationship between the X and Y values for the mentioned points and their position on the

coordinate system. As we know, in statistical problems, variable Y is a random variable, and on the other hand, we are always faced with more points. It is rare to get a straight line from drawing lines. As a result, the only way to show the relationship between X and Y values is to draw and obtain the equation of the best line that can exist.

Coefficients a and b are called regression coefficients. In order to determine the regression equation, a and b should be determined so that the sum of the squares of the deviations of the observations from the estimated regression line is the minimum value.

$$D = \sum(Y_i - \bar{Y}_i)^2 = \text{Min}$$

\bar{Y} is the mean of Y for each value of X, therefore:

$$D = \sum(Y_i - a - bX_i)^2 = \text{Min}$$

Since there are two unknowns a and b, the derivatives of component D with respect to a and with respect to b are set equal to zero and finally the values of a and b are:

$$b = \frac{\sum Y_i X_i - \frac{\sum Y_i \sum X_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{\sum (Y - \bar{Y})(X - \bar{X})}{\sum (X - \bar{X})^2} = \frac{\sum XY}{\sum X^2}$$

In statistical terms, b is called regression coefficient. It represents the amount of change that occurs in the function variable (Y) as a result of a unit change in the constant variable (X). After calculating the equation of the regression line, it is possible to put the value of variable X in the said equation and calculate the real value of Y for each value of X. These values are called the estimate of Y (\bar{Y} Het) and they are the values of Y in the case that there is a completely linear relationship with the characteristics of the regression line equation between two variables. Another characteristic of this type of relationship is that it is

possible to predict and estimate the unknown Y values by knowing the X values. Other applications of regression can be described, predicted, estimated and controlled.

If the regression line fits the data well, then it is appropriate to use its equation for prediction. Provided that we do not leave the field of numbers. However, we should only use the regression line equation when r shows a significant linear correlation. When there is no significant linear correlation, we should not use the regression equation for prediction. In this case, the best estimate of the second variable is the mean of the random sample. In general, if r is close to +1 or -1, then the regression line is a good fit to the data, but if r is close to zero, the linear regression fit is poor.

Example 7-1: According to the following information, the regression estimation equation is equal to:

$$N = 10 \quad \sum xy = 15$$

$$\sum x^2 = 140 \quad \sum y^2 = 35$$

$$\sum x = 20 \quad \sum y = 10$$

$$b = \frac{\sum Y_i X_i - \frac{\sum Y_i \sum X_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{15 - \frac{(20 \cdot 10)}{10}}{35 - \frac{(10)^2}{10}} = \frac{-5}{25} = -0.2$$

$$\bar{x} = a + b\bar{y} \quad 2 = a + (-0.2 \times 1) \quad a = 2.2$$

$$x = 2.2 - 0.2 y$$

7.7. Regression line equation

It should be known that the value of the relationship between the predicted variable and the predicting variable is a function of the sign and intensity of the correlation coefficient. If the correlation is positive,

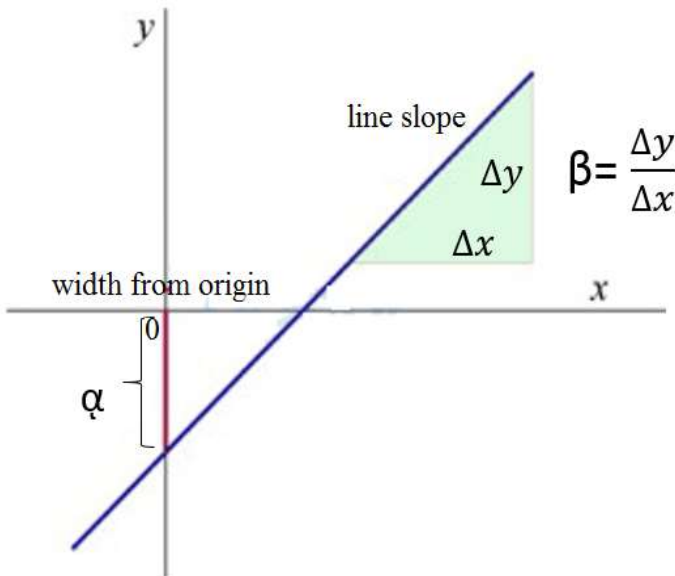
the direction of prediction y will be the same as the direction of standard score x , and if the correlation is negative, the direction of prediction y will be opposite to the direction of standard score x .

If we plot the predicted standard scores (y) in the coordinate system, they will be on a line, which is called the regression line, which is related to the predicted scores. In other words, regression line equation is used to predict one variable from another variable.

The mathematical form of a simple linear function is as follows:

$$Y_i = \alpha + \beta X_i$$

where Y is the dependent variable and X is the independent variable. α and β are constant values. α is called the width from the origin and represents the value of Y for $X = 0$. The coefficient β , which represents the slope of the line, shows the amount of changes in Y per unit change in X .



High and positive statistical correlation between x and y means that:

- 1- The regression line can be used.
- 2- The regression line gives a more accurate estimate of y for a given x.
- 3- Draw the regression line with positive coefficient b.

The method of calculating coefficients a and b:

$$b_{YX} = \frac{Cov_{XY}}{S^2_X} \qquad b_{yx} = r_{xy} \frac{S_Y}{S_X} \qquad b_{xy} = r_{xy} \frac{S_X}{S_Y}$$

Slope of the regression line Standard deviation $S_Y = y$
 = b_{yx} and b_{xy}

Correlation coefficient = r_{xy} Standard deviation $S_X = x$

Calculation of a and b from raw data:

$$b_{yx} = \frac{N\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sum X_i^2 - (\sum X_i)^2} \qquad a_{yx} = \frac{\sum Y_i - b_{yx} \sum X_i}{N}$$

The most basic method used in using the Pearson correlation coefficient for prediction is the method of predicting standard scores.

Correlation specifies how far the predicted scores are from the mean.

In Pearson's correlation coefficient, the following formula is used to obtain the number of pairwise comparisons.

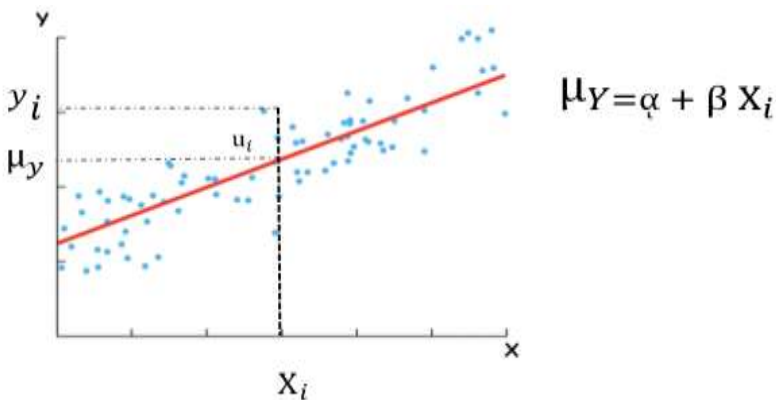
$$C = \frac{n(n - 1)}{2}$$

7.8. Community Regression Function (PRF)

The regression line of the main population is shown below.

$$\mu_Y = \alpha + \beta X_i$$

where α and β are unknown but fixed parameters that are regression coefficients, α and β are also called width from origin and slope (angle coefficient) respectively. We define the expression Y/X_i as a random variable Y for the given value of X , and μ_{Y/X_i} is the expected value (average value or community average) of the assumed Y in which X takes the value of X_i . Because for every X_i there will be many Y_i as seen in the figure below.



The curve connecting the mean values (bold dots) is called the (population) regression curve of Y in terms of X . If the mean values lie on a straight line like the straight line shown in the figure above, the regression curve is linear.

For each X_i there are many different values of Y_i . Therefore, we can define the deviation of a particular Y_i around its mean value as follows.

$$u_i = Y_i - \mu_Y$$

$$Y_i = \mu_Y + u_i$$

According to the relation $\mu_Y = \alpha + \beta X_i$, we will have:

$$Y_i = \alpha + \beta X_i + u_i$$

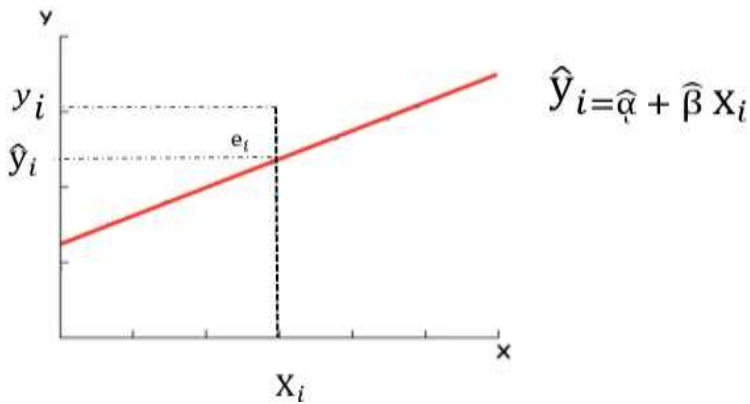
where u_i is the standard deviation of an unobservable random variable that takes positive and negative values. In other words, u_i is a representative or surrogate for all omitted or forgotten variables that affect Y_i but are not present in the model. u_i is a random variable with zero mean and variance σ_u^2 .

7.9. Sample Regression Function (SRF)

In most practical situations, all that is available is a sample of values of Y corresponding to some given value of X . Therefore, our task is to estimate the population regression function (PRF) based on sample data. The regression line based on a random sample of the original population is as follows:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

The constant value $\hat{\alpha}$ is an estimate of α , the slope $\hat{\beta}$ is an estimate of the slope β and \hat{Y}_i is also an estimate of μ_Y of the original population.



Now, just as we stated in the community regression function, the sample regression function (SRF) can also be expressed in its probabilistic random form as below.

$$e_i = Y_i - \hat{Y}_i$$

$$Y_i = \hat{Y}_i + e_i$$

According to the relation $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, we will have:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

where e_i also indicates the sample error component. Conceptually, this symbol is comparable to u_i and can be considered as an estimate of u_i .

By summing up the above information, we find that the first issue in regression analysis is PRF estimation:

$$Y_i = \alpha + \beta X_i + u_i$$

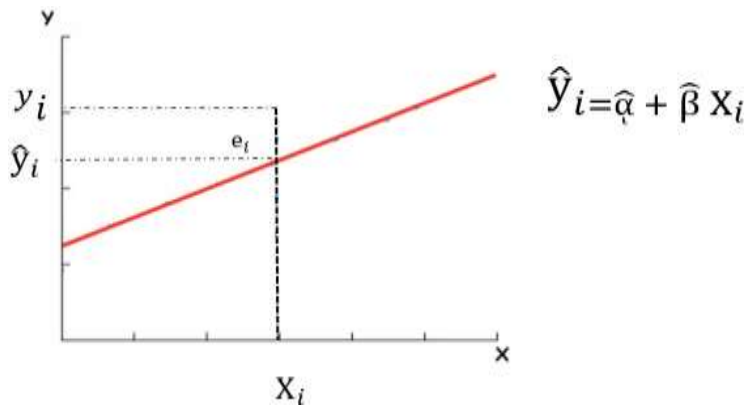
According to SRF:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + e_i$$

Is. In most cases, our analysis is based on only one sample of an assumed society. But due to sampling fluctuations, our estimate of PRF is at best only an approximation based on SRF.

7.10. Ordinary least squares method

The least squares method is used to obtain the parameters $\hat{\alpha}$ and $\hat{\beta}$ in such a way that the regression equation is the best indication of the relationship between the values of X_i and the dependent variable \hat{Y}_i .



Error (e_i) is the vertical distance between the actual observed value (Y_i) and the value obtained for it from the fitted line (\hat{Y}). that's mean:

$$e_i = Y_i - \hat{Y}_i$$

The best fitting line is the line in which the sum of squared errors ($\sum e^2_i$) which is represented by SSE, has the smallest value, because:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i. \quad \sum e^2_i = \sum (Y - \hat{Y}_i)^2$$

we will have:

$$SSE = \sum e^2_i = \sum (Y_i - \hat{\alpha} + \hat{\beta} X_i)^2$$

$$= \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

It is enough to derive SSE once in terms of $\hat{\alpha}$ and once in terms of $\hat{\beta}$ and set them equal to zero.

$$\frac{\partial(SSE)}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} + \hat{\beta} X_i) = 0$$

$$\frac{\partial(SSE)}{\partial \hat{\beta}} = -2 \sum (Y_i - \hat{\alpha} + \hat{\beta} X_i) X_i = 0$$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta} \sum X_i \quad \text{Equation (1)}$$

$$\sum X_i Y_i = \hat{\alpha} \sum X_i + \hat{\beta} (\sum X_i^2) \quad \text{Equation (2)}$$

By dividing equation 1 on n (the number of observations on X and Y), we have:

$$\frac{\sum Y_i}{n} = \hat{\alpha} + \hat{\beta} \frac{\sum X_i}{n} \quad \bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}$$

In this case, the value of $\hat{\alpha}$ is equal to:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

We insert the value of $\hat{\alpha}$ in equation (2), the value of $\hat{\beta}$ is obtained as follows:

$$\hat{\beta} = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \hat{\alpha} + \hat{\beta} \frac{\sum X_i}{n}$$

where \bar{X} and \bar{Y} are arithmetic averages of X and Y values, respectively. In statistical terms, $\hat{\beta}$ is called regression coefficient.

Note that the estimation of two parameters α and β from the sample, i.e. $\hat{\alpha}$ and $\hat{\beta}$, means that two degrees of freedom are lost, so if the sample size is n, the number of degrees of freedom will be equal to $r = n - 1$.

Other formulas for calculating $\hat{\beta}$ are as follows:

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}_i \bar{Y}_i}{\sum X_i^2 - n\bar{X}^2} \qquad \hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y}) - n\bar{X}_i \bar{Y}_i}{\sum (X_i - \bar{X})^2}$$

$$= \frac{SPD_{X,Y}}{SSD_X}$$

We had $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ and also $\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}$, if we subtract the above two relationships,

$$\hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X}) \qquad \text{or} \qquad \hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X})$$

The regression line equation for when X is the independent variable is:

$$\frac{y - \bar{y}}{S_t} = r \frac{x - \bar{x}}{S_x}$$

7.11. Sources of variation and statistical mode

After conducting the experiment and calculating the values of a and b, since these two indexes are random variables that each have variance, it is necessary to make statistical inferences about the parameters of the society (α , β), along with the statement of the probability of correct judgment. Fixed variable X means that its values are known and fixed in advance. The Y function variable means that its values are random and there is a possibility of making a mistake in their measurement. Here too, one should pay attention to the possibility of committing the first type of mistake. The first type of error in regression problems is that there is no relationship between two variables and it is judged based on the calculated b that there is a relationship (correct H_0 rejection). Therefore, the values of Y for each X have a normal frequency distribution with mean $\mu_{Y,X}$ and variance $S_{Y,X}^2$.

The statistical model of data in regression problems is investigated under two assumptions.

1- If there is no relationship between two variables. The statistical model will be as follows:

$$Y_i = \mu_{Y.X} + e_i$$

Y_i = the value of the function variable for level i of the fixed variable

$\mu_{Y.X}$ = the average of the hypothetical Y for each of the values of X , which is the same as \bar{Y} .

e_i = deviation of Y from the mean ($-\mu_{Y.X}$), which is an independent variable with a normal frequency distribution and zero mean and a certain variance, which is represented by $\delta_{X.Y}^2$.

2- If there is a relationship between two variables, the statistical model of the data will be as follows:

$$Y_i = \alpha + \beta X_i + e_i$$

α and β are community parameters that are estimated by sampling from the community with a and b , respectively.

In the case of a random sample, the deviation of each random value Y_i from the sample mean \bar{Y} consists of the following two components.

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

a): Deviation of Y from the regression line, $(Y_i - \hat{Y}_i)$

b) Deviation of points on the regression line from the mean, $(\hat{Y}_i - \bar{Y})$

If the above relation is written for each Y and then it is raised to the power of two and then added for all values, the following sum of squares will be obtained:

$$SST = \sum(Y_i - \bar{Y})^2 \quad \text{Total sum of squares}$$

$$SSD = \sum(Y_i - \hat{Y}_i)^2 \quad \text{Sum of squared deviation from regression}$$

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2 = b^2 \sum(X - \bar{X})^2 \quad \text{Regression sum of squares}$$

The regression variance analysis table is as follows:

Sources of variation	df	Sum of squares		Mean square	F
		Application formula	Theoretical formula		
Regression	1	$b^2 \sum(X - \bar{X})^2$	$\sum(\hat{Y}_i - \bar{Y})$	NSR	$\frac{MSR}{MSD}$
Deviation from regression	n-2	$\sum(Y_i - \bar{Y})^2 - b^2 \sum(X - \bar{X})^2$	$\sum(Y_i - \hat{Y}_i)$	$MSD = S_{XY}^2$	
Total	n-1	$\sum Y^2 - \frac{(Y)^2}{n}$	$\sum(Y_i - \bar{Y})^2$		

In statistical terms, b is called the regression coefficient and it represents the amount of change that occurs in the function variable (Y) as a result of a unit change in the constant variable (X). In the variance analysis table, two estimates of the two-variable community variance can be seen, (1) regression variance indicates the distribution of points in the range of \bar{Y} to the regression line. The larger the variance, the closer the concentration of X and Y points to the regression line. (2) The variance of the deviation from the regression, which is the variation that is not justified by the regression line, and shows the part of the dispersion located outside the range of \bar{Y} to the point Y. The magnitude of this variance is a justification that there is no close relationship between the two variables.

Example 7-1: check the significance of the regression coefficient according to the following data:

X	1	2	4	6	7
Y	3	5	8	9	10

$$SSR = b^2 \sum (x - \bar{x})^2 = 2 (1.115)(26) = 32.324$$

$$SSG = \sum (Y_i - \bar{Y})^2 = (7 - 3)^2 + (7 - 5)^2 + \dots + (7 - 10)^2 = 34$$

$$SSD = \sum (Y_i - \bar{Y})^2 - \sum (x - \bar{x})^2 = 34 - 32.324 = 1.676$$

S.V	df	SS	MS	F
Caused by regression	1	32.324	32.324	57.86 ^{**}
Deviation from regression	3	1.676	0.5587	
Total	4	34		

The calculated F is higher than the F in the table at the probability level of 1 percent, i.e. 34.12. Therefore, the hypothesis of $b = 0$ is rejected and it can be said that the regression coefficient is significantly different from zero.

7.12. Assumptions of linear regression analysis

Linear regression analysis is based on assumptions that if these assumptions are not true, the results will be biased. The most important assumptions in linear regression analysis are as follows:

A) The linearity of the regression relationship

One of the easiest ways to determine the linearity of the regression relationship is to draw the graph of e in terms of X_i or \bar{Y}_i . If the distribution of points in the coordinate axis is rectangular and does

not have a particular trend, then the relationship can be assumed to be linear. If the regression relationship is non-linear, non-linear relationships such as polynomial regression should be used to fit the data.

b) Uniformity of variances

The uniformity of the variance of Y values within X levels is one of the basic assumptions in regression analysis. To check the uniformity of the variances, the deviations from the regression line can be drawn in terms of \bar{Y}_i or X_i . If the distribution of points is rectangular and has no particular trend, it can be said that the assumption of uniformity of variance is true. The existence of the trend, especially the trumpet mode, in the graph indicates the non-uniformity of the variances. In case of non-uniformity of the variances, it is better to transform the data by one of the methods of logarithmic transformation, square root, etc. If the data transformation does not work, another solution is to use the weighted least squares method.

$$W_i = \frac{1}{S_i^2}$$

$$b' = \frac{\sum_{i=1}^n W_i (X_i - \bar{X}') (Y_i - \bar{Y}')}{\sum_{i=1}^n W_i (X_i - \bar{X}')^2}$$

$$a' = \bar{Y}' - b' \bar{X}'$$

$$\bar{Y}' = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i}$$

$$\bar{X}' = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n W_i}$$

However, the condition for using this method is to measure more than one Y for each X level. In the weighted least squares method, the

variance image within each X is used as a weight in the calculation of the linear regression coefficient. Therefore, any level of X that has a smaller variance will have more weight in explaining the regression coefficient. As you know, statistics that have a smaller variance are more reliable. The formulas related to the weighted least squares method are listed below. The prime sign indicates the weighted statistics.

$$SSReg' = b'^2 \sum_{i=1}^n W_i (X_i - \bar{X}')^2$$

$$SSe' = SSY' - SSReg'$$

$$SSY' = \sum_{i=1}^n W_i (Y_i - \bar{Y}')^2$$

$$SSX' = \sum_{i=1}^n W_i (X_i - \bar{X}')^2$$

$$t_b = \frac{b'}{\sqrt{\frac{MSe'}{SSX'}}$$

$$t_a = \frac{a'}{\sqrt{MSe' \left(\frac{1}{n} + \frac{X'^2}{SSX'} \right)}}$$

In these formulas:

S_i^2 = The variances y within each X level

W_i = The weight of each X level

c) The normality of the distribution of deviations from the regression line

The normality of the residuals can be checked by tests such as skewness and kurtosis as well as Kemolmogrov-Smirnov. If the distribution is not normal, they transform the data. Considering that

non-uniformity of variances and non-linearity of the relationship may falsely cause non-normality of the distribution of deviations, it is better to examine the assumptions of uniformity of variances and linearity of the regression relationship.

d) independence of deviations from the regression line

In regression analysis, it is assumed that the e_i 's are independent from each other. From the graph related to the distribution of deviations from the regression line, it can be seen that the e_i 's are independent. If the deviations are independent, no particular trend will be observed and the distribution of points will be rectangular. If the deviations from the regression line are not independent, this phenomenon is called serial correlation or autocorrelation. Serial correlation can be positive or negative. In the positive case, an increase in one e leads to a decrease in the adjacent e .

Another method to check autocorrelation is Durbin-Watson test. Durbin-Watson statistic (D) is calculated from the following formula:

$$D = \frac{\sum_{i=1}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

D = 2 Non-correlation of the series

D = 4 Negative series correlation

D = 0 Positive serial correlation

Conventionally, if D is in the range between 1.5 and 2.5, it is said that there is no serial correlation.

If the e 's are not independent, the weighted least squares method can be used under certain conditions. If the X 's are not the same distance, another solution is to calculate the linear regression coefficient through the pair of data $X_i - X_{i-1}$ and $Y_i - Y_{i-1}$ instead of X_i and Y_i .

In some cases, analysis of time series is used to eliminate the effect of "autocorrelation".

7.13. Comparison of two regression coefficients

The t-test can be used to compare two regression coefficients as follows:

$$t = \frac{b_1 - b_2}{\sqrt{\frac{MSE_1}{SSX_1} + \frac{MSE_2}{SSX_2}}}$$

The joint MSE is obtained from the weighted average of the error variance of two experiments:

$$MSE = \frac{(MSE_1)(n_1 - 2) + (MSE_2)(n_2 - 2)}{(n_1 - 2) + (n_2 - 2)}$$

So, the calculated t is compared with the table t for the degree of freedom $n_1 + n_2 - 4$, and if it becomes significant, it can be stated that there is a significant difference between the two regression slopes.

7.14. A simple method in statistical regression calculations

If the values of X and Y or one of them decrease, the regression coefficient and variance analysis table calculations will not change, but the value of a will change. One of the simplest methods is to express two variables in terms of deviation from the mean. Also, if multiplication and division are used to simplify the data, the regression

coefficient will not change, but the value of a will be smaller and larger according to the simplification coefficient.

Another method for simplifying the data is common and it is used in cases where the values of X change to an equal distance from each other. In this case, the results of the variance analysis table are not different from the case where the original Xs are used, but the values of a and b change. In this case, the simplified values are represented by xi and the sum of xi is equal to zero.

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{\sum x_i Y_i}{\sum x_i^2} = \frac{Q}{\sum x_i^2}$$

$$SSR = \frac{Q^2}{\sum x_i^2}$$

In this method, it is usually not necessary to obtain the regression line in the main scale of the desired data, and only the presence or absence of a relationship between two variables is investigated through the analysis of variance table.

Example 7-2: According to the table below, if 20 grams of nitrogen fertilizer is used in the pot, what will be the weight of the plant?

Nitrogen fertilizer	7	9	11	13	15
Plant weight	20	25	40	40	50

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{2075 - \frac{9625}{5}}{645 - \frac{(55)^2}{5}} = 3.75$$

$$a = \bar{y} - b\bar{x} = 35 - (3.75 \times 11) = -6.25$$

$$y = a + bx \quad x = 20 \quad y = -6.25 + (3.75 \times 20) = 68.75$$

Regression line from the coordinate origin

In this case, $a = 0$ and the equation of the regression line and its slope are as follows:

$$Y = b x$$

$$b = \frac{\sum X_i Y_i}{\sum X_i^2}$$

Example 7-3: In the following example, if the regression equation passes through the origin of coordinates, it is desirable to obtain the equation of the regression line?

x	13.6	13.9	21.1	25.6	26.4	39.8	40.1	43.9	51.9	53.2	65.2	66.4	67.7
y	52	48	72	89	80	130	139	173	208	225	259	199	255

In cases where the regression line passes through the coordinate origin, the value of a is equal to zero and its equation is $\hat{y} = bx$. The value of b and the equation of the regression line are as follows:

$$b = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{(13.6 \times 52) + (13.9 \times 48) + \dots + (67.7 \times 255)}{(13.6)^2 + (13.9)^2 + \dots + (67.7)^2} =$$

$$b = \frac{95762.9}{26066.33} = 3.67$$

$$\hat{y} = 3.67 x$$

7.15. lack of fit

In examining the relationship between two variables, all X values or some of them can be repeated equal or unequal times. In such cases, in addition to the variance of deviation from regression, the error variance was also calculated, which is known as pure error.

The difference between the sum of the squares of this error and the sum of the squares of the deviation from the regression is known as the sum of the squares of the lack of fit (LOF), which shows the deviation from the true regression.

The obvious use of this method is in variance analysis problems to compare the averages of several groups. If F is related to a significant lack of fit, it means that there is another relationship between the two variables other than linear, and therefore the hypothesis $H_0: \beta = 0$ cannot be tested. In other words, the significance of the lack of fit may mean that non-linear expressions of X (such as X^2) should be added to the model or some non-linear models such as exponential function may be more suitable.

7.16. Correlation

Correlation is the study of the relationship between two variables with several variables, that is, we are looking for whether the increase or decrease of one variable has an effect on the other variable. Descriptive techniques in such studies (in which only the relationship between variables is examined, without any of them being manipulated or controlled) is the degree of agreement (such as Phi coefficient or other similar coefficients) for cases where two variables are measured.

be a floor Spearman's rank coefficient for when the two measured variables are expressed in an ordinal scale, and finally Pearson's correlation coefficient for when the two variables in question have an interval scale. Each of these shows the strength of the relationship between two variables.

If the goal is to study one variable such as X on another variable such as y , it is a single variable correlation study. So-called single-variable studies are called artificial situations. If the goal is to investigate several variables such as x_1, x_2, \dots, x_n on another variable, i.e. y , then it is an investigation of multivariable correlation. If the purpose of examining several variables x_1, x_2, \dots, x_n on several y_1, \dots, y_n , then it is focal correlation. In these studies, the independent variable is called predictor variable and the dependent variable is called predictor variable.

For a clearer, more objective analysis and to make a decision about whether there is a statistically significant relationship between two variables, we use the correlation coefficient. In other words, correlation, covariance between two variables shows the direction and value of the relationship between the two variables. But in practice, covariance is not used to determine correlation, because the value of covariance changes with respect to the measurement unit. Therefore, to compare the relationships between variables, it is better to use a quantity that shows the relationship between two attributes independent of the effect of the measurement unit. For this purpose, first two variables are standardized so that each has a mean of zero and a variance of one.

7.17. The correlation coefficient

The ratio of covariance of two variables to the product of standard deviations is called correlation coefficient according to the definition and it is denoted by r . This coefficient varies from -1 to +1 and the closer the correlation coefficient is to +1. The correlation increases and the closer this value is to -1, the correlation decreases. If the correlation coefficient of two attributes is +1, the changes of the two attributes are completely dependent and aligned. If the correlation coefficient is -1, their changes are completely interdependent but non-aligned. Values close to zero indicate the absence of correlation between two variables. In fact, the correlation coefficient expresses the intensity of the relationship between two traits in a standardized way without units.

$$r = \frac{\bar{d}_{x,y}}{S_x S_y} = \frac{\sum(x-\bar{x})(Y-\bar{Y})}{\sqrt{(x-\bar{x})^2(Y-\bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2 Y^2}}$$

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum X)^2}{n}\} \{\sum Y^2 - \frac{(\sum Y)^2}{n}\}}}$$

Usually, to determine the significance of the obtained correlation coefficient, it is tested statistically. If the calculated r is greater than the r of the table with $n - 2$ degrees of freedom, it is said that the null hypothesis $r = 0$ is rejected and the correlation coefficient is significantly greater than zero, which means that there is a significant relationship between the two variables. A correlation coefficient smaller than r in the table, even if its value is greater than zero, is considered non-significant and lacks statistical significance.

To calculate the correlation, you must first draw a scatter diagram. If the points are close to the hypothesized line, the correlation is high.

The four main operations ($\div + - \times$) of a score in all the scores of a distribution have no effect on the correlation coefficient. For example, if the correlation between (y, x, height and weight) is 0.5, the correlation between $2x$ and y will be 0.5.

The direction of correlation is determined by the sign of the correlation coefficient (negative or positive) and the intensity of the correlation is determined by the absolute value of the correlation coefficient, and it does not indicate the cause and effect relationship between the variables in any way.

The intensity (strength) of correlation is independent of the sign of the correlation coefficient, so that the highest intensity of correlation between two variables is when the calculated correlation coefficient (r) is $+1$ or -1 .

The purpose of correlation studies is to investigate the type and degree of relationship between the studied variables.

In correlation studies, the relationship between two traits is investigated. The correlation value shows the relationship and direction between two attributes.

If a constant numerical value is added or subtracted to all numbers or scores, or multiplied and divided, the correlation coefficient between the two variables does not change. Because when we change all the numbers equally, there is no change in their general relationship.

It should be known that the correlation coefficient is used with an interval and relative scale. The correlation coefficient can be used as a ratio (decimal), but its interpretation should not be in terms of

percentage. For example, it cannot be said that the correlation coefficient of 90% is exactly twice 45%.

Lack of communication	Correlation coefficient = 0
Low communication	Correlation coefficient = ± 0.01 until ± 0.2
Good communication	Correlation coefficient = ± 0.21 until ± 0.5
Eaningful communication	Correlation coefficient = ± 0.51 until ± 0.7
High communication	Correlation coefficient = ± 0.71 until ± 0.9
Excellent communication	Correlation coefficient = ± 0.91 until ± 1

When estimating the correlation coefficient, we must consider the nature of the society in which the two variables were observed or measured. Because the interpretation of the correlation coefficient is different from one society to another. Because:

The basis of the relationship differs from one society to another.

The distribution of variables is different in different societies.

The correlation between two variables can be influenced by their correlation with a third variable.

It should be known that the correlation between two variables in a society that is heterogeneous based on the investigated variables is more than the correlation of the same variables in a homogeneous society.

If the correlation coefficient is 0.5, it means that the relationship is relatively (somewhat) good and the relationship between the two variables is direct.

A correlation coefficient higher than 0.7 indicates a high correlation between the variables, but the actual interpretation depends on the subject of the research.

It should be known that the correlation between two variables can be shown using a scatter diagram.

The stronger the correlation coefficient, the smaller the scatter of points on the fitted line. In perfect correlation, all points lie on the fitted line.

7.18. Types of correlation

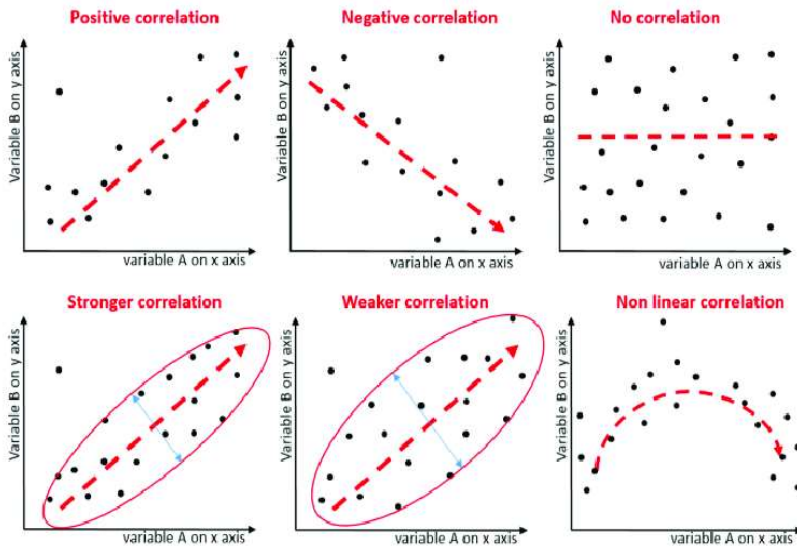
The nature of the relationship between two random variables can also be displayed by plotting the points related to the pair of data in a coordinate axis. The general distribution of points shows the intensity and direction of correlation.

If there is a positive correlation between two variables, the changes of the two variables are in the same direction, that is, X and Y both increase or both decrease. When there is perfect positive correlation ($r = 1$), accurate predictions can be made.

A strong positive linear correlation between X and Y is reflected by a number close to +1.

If there is a negative correlation between two variables, the changes of the two variables are in the opposite direction, that is, if X increases, Y decreases and vice versa. When there is perfect negative correlation ($r = -1$), accurate predictions can be made.

A strong negative linear correlation between X and Y is reflected by a number close to -1.



The two correlation coefficients of 0.64 and -0.64 indicate the same degrees of correlation and when we discuss the correlation coefficient, it cannot be said that positive is better than negative. In general, the larger the absolute value of the number, the stronger the correlation. The + or - sign only indicates the direction of correlation.

If $r = 0.3$ or $r = 0.9$, the correlation is three times higher than $r = 0.3$.

Lack of correlation ($r = 0$) between two variables means that the high and low values of the two variables X and Y are not related to each other in a predictable way.

If r is close to zero, we conclude that there is no significant linear correlation between X and Y, and this does not mean independence.

If r between two random variables is equal to zero, it means that these two variables do not have a linear relationship, but it is not unlikely that they are related to each other in a non-linear way.

The range of correlation is between +1 and -1, i.e., the correlation of -1 is inverse and complete, the correlation of +1 is direct and complete. The correlation value is zero to 1. Any data between these two limits is incomplete. It means that the correlation of 0.99 is incomplete. The closer the correlation is to the absolute value of 1, the more intense it is.

If the correlation coefficient is higher than 0.70, we say that the relationship is linear. If it is lower, the relationship is non-linear, and if the relationship between two variables is non-linear, use the correlation coefficient η .

$$\eta = 1 - \frac{SS_{reg}}{SS_{res}}$$

7.19. Correlation coefficient properties

always $+1 \leq r \leq -1$

The correlation coefficient is relative in nature, that is, the correlation coefficient between x and y (r_{XY}) is equivalent to the correlation between Y and X (r_{YX}).

The correlation coefficient is independent of the origin and scale of measurement. That is, if we define that $X_i^* = aX_i + c$ and $Y_i^* = bY_i + d$ where b, a, c, d > 0 are constant, then the correlation coefficient between Y^* and X^* is similar to the correlation between the main variable is X and Y.

If X and Y are statistically independent, the correlation coefficient between them is zero, but if $r = 0$, it does not mean that the two variables

are independent. In other words, zero correlation is not necessarily proof of independence.

The correlation coefficient is a measure of linear correlation or linear dependence and cannot be used to describe non-linear relationships.

Although r is a measure of linear correlation between two variables, it does not necessarily indicate any cause and effect relationship.

The methods of calculating the correlation coefficient are:

a) Pearson's correlation coefficient

If the measurement scale is distance or relative, then Pearson's moment correlation coefficient is used to calculate the correlation coefficient of two variables.

To use Pearson's correlation coefficient, the following assumptions must exist:

The relationship between two variables must be linear (a relationship whose scatter diagram is around a line).

The distributions are normal or have similar shapes (assuming two variables are skewed to the right or left).

The scatter diagram is the same.

There are two methods for calculating Pearson's moment correlation coefficient, which is the most famous type of correlation coefficient.

The first method: This method is used when the data is in the form of raw numbers and the desired measurement scale is relative or distance. In this method, the following formula is used.

$$\text{(Pearson correlation coefficient) } r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

In the second method, which is shorter and simpler, to calculate the Pearson correlation coefficient, we calculate the average of the data, and according to the deviation of each of the x and y scores from the average, we obtain the correlation coefficient using the following formula:

$$r_{xy} = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

7.20. Introduction to covariance

It should be known that covariance means the change of one variable with another variable and is closely related to correlation. In physics, Pearson's correlation coefficient is called dimensionless covariance.

The following formula is used to calculate the covariance of two variables.

$$\text{Cov} [x, y] = S_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\sum xy}{n - 1}$$

The amount of covariance is affected by the measurement unit, but the small or large covariance is not the reason for the low or high correlation. For this reason, covariance is not used to determine the presence or absence of correlation.

If covariance can be used to correlate one variable with another variable, if its value is divided by the standard deviation of two variables, that is, its formula should be converted to Pearson's correlation coefficient.

$$r_{XY} = \frac{\text{Covariance} \left(\frac{\sum xy}{n-1} \right)}{\text{Multiply the product of standard deviation of two variables } (S_x \cdot S_y)}$$

After determining the correlation coefficient, it is possible to determine the probability and percentage of prediction of one variable from another variable. For this purpose, the correlation coefficient is multiplied by two and multiplied by 100. that's mean:

$$\text{Prediction coefficient} = (r_{xy})^2 \times 100$$

Covariance is a concept closely related to correlation.

In the case of Pearson's correlation coefficient, the common variance (covariance) of two variables is always smaller than or equal to the product of the standard deviation of the two variables.

b) Spearman's rank correlation coefficient (r_s)

If the measurement comparison of the studied variables is of ordinal (rank) type (that is, numbers show the order of objects or people), or the condition of normal distribution in numbers with an interval scale for the Pearson correlation coefficient does not exist, in this case to calculate the relationship between Spearman rank-order correlation coefficient is used for two variables.

The following formula is used to calculate Spearman's non-parametric rank correlation coefficient.

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

d = difference between a pair of ranks

N = number of a batch of data

Using this formula requires ranking the numbers and calculating their difference (D). To rank each of x and y variables:

The first thing is to sort the records in the column from the largest to the smallest, and then in the next column, we assign the row number to each of the records from the top to the bottom (from the largest to the smallest).

In the next step, people who have the same record, to find a common rank between them, we first add their row numbers together and then divide by their number. The obtained number is the common rank for them, and other people who have the same records, their row numbers are considered their records.

Spearman's correlation coefficient is used for qualitative cases.

c) Kendall's rank correlation coefficient

Kendall's correlation coefficient is a method that can be used instead of Spearman's correlation coefficient to calculate correlation. This method is used when the studied variables are measured using a rank scale or higher.

d) serial correlation coefficient

If one of the variables is measured with an interval or relative scale and the other with a nominal scale, the serial correlation coefficient is used to calculate the correlation coefficient.

e) Correlation coefficient of nominal variables

When two variables are nominal, the agreement coefficient, Phi coefficient and tetrachoric coefficient are used to calculate the correlation coefficient.

Since both distributions are continuous variables, we use the contingency coefficient to measure the correlation between the variables.

The phi coefficient (ϕ) is used only when there is a database table and the variables are two real values (female-male, married-single).

When the variables are nominal (non-real or fictitious two-valued and only two types of classification are suitable), the tetrachoric coefficient is used.

When there is a continuous variable and a dichotomous variable, the bivariate correlation coefficient is used.

Average plays a key role in statistical calculations such as correlation coefficient, average deviation, standard deviation and variance.

7.21. Correlation coefficient significance test

We know that r is the linear correlation coefficient, which measures the intensity of correlation between X and Y in the sample, so r is a sample statistic. But p is a sample statistic. But p is the linear correlation coefficient, which measures the intensity of correlation between X and Y in society. So P is a community parameter. As a result, the correlation coefficient calculated from the sample (r) will be an estimate of the population correlation coefficient (p).

Sometimes it is possible that the two variables X and Y do not have any linear relationship and the correlation coefficient of these two variables in the society is equal to zero ($P = 0$), but the correlation

coefficient calculated in the sample shows a non-zero quality ($r \neq 0$) for to clarify the issue, we must do the assumption test.

$$H_0 : p = 0$$

$$H_1 : p \neq 0$$

In order to perform this test, it is necessary to add the assumption that X is a random variable with a normal distribution in addition to the assumption that the random variable Y is normal. The appropriate statistic for the test regarding the social correlation coefficient being zero ($p \neq 0$) is:

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

which has a t distribution with $n - 2$ degrees of freedom and the critical region is:

$$t \leq t_{\frac{\alpha}{2}, n-2} \quad \text{,} \quad t \geq t_{1-\frac{\alpha}{2}, n-2}$$

So, by calculating the numerical value of the test statistic, the ratio of correctness or incorrectness of the H_0 hypothesis is judged and decided.

If we reject the hypothesis $H_0: P = 0$ and r is positive, we conclude that there is a significant positive linear correlation between X and Y.

If we reject the hypothesis $H_0: P = 0$ and r is negative, we conclude that there is a significant negative linear correlation between X and Y.

If we do not reject the hypothesis $H_0: P = 0$, we conclude that there is no significant linear correlation between X and Y.

Example 7-4: According to the information, the correlation coefficient between two variables is equal to:

$$\begin{aligned} N &= 10 & \sum xy &= 15 \\ \sum x^2 &= 140 & \sum y^2 &= 35 \\ \sum x &= 20 & \sum y &= 10 \end{aligned}$$

$$r = \frac{\sum XY - \frac{(\sum X \sum Y)}{n}}{\sqrt{\left\{ \sum x^2 - \frac{(\sum x)^2}{n} \right\} \left\{ \sum y^2 - \frac{(\sum y)^2}{n} \right\}}} = \frac{15 - \frac{(20 \times 10)}{10}}{\sqrt{\left\{ 140 - \frac{(20)^2}{10} \right\} \left\{ 35 - \frac{(10)^2}{10} \right\}}}$$

$$r = \frac{-5}{\sqrt{100 \times 25}} = \frac{-5}{50} = -0.1$$

There is a negative correlation between the two variables.

7.22. The relationship between regression and correlation

The slope of the line between X and Y is as follows:

$$b_{Y/X} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

If the numerator and denominator are divided by n-1, the regression coefficient is the ratio of covariance to variance of X:

$$b_{Y/X} = \frac{\delta_{\bar{X}Y}}{\delta_{\bar{X}}^2}$$

If X is also assumed to be a random variable like Y, then the slope of the line can be calculated, only the interpretation of the results will be different. In this case, two types of regression are possible, regression of Y on X and regression of X on Y:

$$b_{Y/X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

$$b_{X/Y} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_Y^2}$$

The correlation coefficient between two variables Y and X is not quantitative but the geometric mean of these two regression coefficients. Therefore, the product of the above two regression coefficients is equal to the square of the correlation coefficient (r^2), which is called the detection, explanation or determination coefficient.

$$b_{Y/X} \cdot b_{X/Y} = \frac{\{(X-\bar{X})(Y-\bar{Y})\}^2}{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2} = r^2$$

$$\frac{SSR}{SST} = \frac{\sum(\bar{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{b^2 \sum(X - \bar{X})^2}{\sum(Y - \bar{Y})^2} = r^2$$

The detection coefficient represents the portion of the total variation (or total variance) that is explained by the linear relationship between X and Y. r^2 is a variable between zero and one and is expressed as a percentage. The relationships between the detection coefficient and the sum of squares of regression and deviation from regression are as follows:

$$\sum(Y - \bar{Y})^2 = SST$$

$$\sum(Y - \hat{Y})^2 = SSD = (1 - r^2)SST$$

$$\sum(\bar{Y} - \bar{Y})^2 = SSR = r^2SST$$

Example 7-5: In an experiment, the number of clusters and the weight of 200 seeds were measured in a variety of rice. According to the following observations obtained from a random sample of 5

individuals, what percentage of the changes in the weight of 200 seeds is justified by the number of clusters?

Number of clusters	2	4	2	1	3
Weight 200 grains	6	7	5	4	8

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{SS_x SS_y}} = \frac{78 - \frac{2 \times 30}{5}}{\sqrt{\left(34 - \frac{(12)^2}{5}\right) \left(190 - \frac{(30)^2}{5}\right)}} = 0.832$$

$$r^2 = (0.832)^2 = 0.69$$

7.23. Interpreting the correlation coefficient

From a descriptive point of view, the different values of correlation coefficients can be roughly and generally introduced as follows:

- 1- Very weak and insignificant correlation ($r < 0.2$)
- 2- Weak correlation, ($0.2 < r < 0.4$)
- 3- Medium correlation, ($0.4 < r < 0.6$)
- 4- Strong correlation, ($0.6 < 0.8 < r$)
- 5- Very strong correlation ($r < 0.8$)

In the interpretation of correlation coefficients, it is necessary to pay attention to the following points:

- 1- The correlation coefficient is not a function of simple linear changes. For example, it can be said that the coefficient of 0.8 is twice the coefficient of 0.4, but it can be said that the coefficient of 0.8 or 0.6 is stronger than the coefficient of 0.4.

2- The description of the correlation coefficient depends on the subject of the research and it should be interpreted according to the specific context and conditions of the research.

In the inferential interpretation of the correlation coefficient, the ultimate goal is to generalize the results obtained from the sample to the population. If 8 samples are extracted from a two-variable population and the value of r is calculated for each sample, because the values of r are between $+1$ and -1 , their distribution cannot be considered symmetric and testing is done using models such as t and z distribution. He made assumptions. On the other hand, only in the case that the null hypothesis is as follows, the distribution of r can be considered symmetrical and the frequency distribution of t can be used for this test.

There is no correlation between two variables. $H_0 : P = 0$

There is a correlation between two variables. $H_0 : P \neq 0$

$$t = \frac{r-0}{S_r} = \frac{r}{S_r}$$

In cases where the number of sample members is large and the calculated r value is not close to -1 and $+1$, the standard deviation of r can be calculated using this formula:

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

In cases where a hypothesis other than $P = 0$ is tested, because under the hypothesis p is considered equal to the mean of the distribution, the t distribution is not symmetrical and the t table cannot be used for statistical inference. In such cases, if r and p are converted to Z' using the following formula, they will become a new variable

whose distribution is almost normal, and the Z table can be used to perform the hypothesis test.

$$Z' = 0.5 \ln \frac{1+r}{1-r}$$

The Z' values calculated using the above formula have an almost normal distribution with a mean of $\ln 0.5 \ln \frac{1+r}{1-r}$ and a variance of $\frac{1}{n-3}$.

7.24. Coefficient of determination (detection)

It should be known that the correlation coefficient shows the size (value) of correlation between two variables and does not give us much information about the nature of this correlation. To determine the value (size) of sharing between the dispersion of two variables x and y , the coefficient of determination is used and calculated with the following formula.

$$V = (r_{xy})^2 \times 100$$

By calculating the coefficient of determination, it is possible to determine how many percent of the total variance of x is caused by the variance of y and vice versa.

If the correlation coefficient is expressed as a percentage, it is called the coefficient of determination.

The coefficient of determination is considered an important generalized (general) or developmental index between skills and is estimated as a percentage.

The detection or determination coefficient will never be negative. Because the square (power of 2) is the correlation coefficient.

The coefficient of determination is used when we need to determine the amount (severity) of change of one variable from another variable. It is also necessary to predict one variable from another variable.

If the correlation coefficient of two variables is zero, the coefficient of determination will be zero.

If the variances of two categories are homogeneous, the correlation coefficient will be higher.

When one correlation coefficient is twice as large as the other, the smaller coefficient actually represents twice the amount of dispersion that the smaller coefficient represents.

Using the following formula, you can calculate the standard error of prediction for the scores of a variable.

$$S_{yx} = S_y \sqrt{1 - r^2}$$

S_{yx} = error of prediction criterion

S_y = standard deviation of y scores

r^2 = squared correlation coefficient (determination coefficient)

The numerical value of r^2 is known as the coefficient of determination (sample) and it is used as the most important criterion for the goodness of fit of the regression line. In other words, r^2 gives the proportion or percentage of the total variation in the dependent variable Y that is explained by the independent variable X.

Note the two properties of r^2 :

r^2 is a non-negative quantity.

Its limit is $0 < r^2 \leq 1$.

r^2 equal to one indicates that the regression line was able to accurately relate changes in Y to independent changes in X (perfect fit), and r^2 equal to zero indicates that the regression line was never able to relate changes in Y to independent changes in X (Lack of relationship between dependent and independent variable) Other values are between these two limits.

r^2 can be obtained faster from the following formulas.

$$r^2 = \frac{(n \sum XY - \sum X \sum Y)^2}{(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)}$$

$$r^2 = \frac{\hat{\alpha} \sum Y + \hat{\beta} \sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

In regression problems, the correlation coefficient related to the sample, r , can be easily determined by taking the square root of the coefficient.

$$r = \pm\sqrt{r^2}$$

In simple linear regression, if $\hat{\beta}$ is positive, r is also positive, and if $\hat{\beta}$ is negative, r is also negative.

The coefficient of determination is useful for comparing correlation coefficients. When one compares an r value of 0.8 with another r value of 0.4, there is a tendency to think of 0.8 as twice 0.4. This comparison is not correct; the correlation coefficients should be compared according to the common change of two variables (= coefficient of determination).

$$r_1^2 = (0.8)^2 = 0.64$$

$$\frac{0.64}{0.16} = 4$$

$$r_1^2 = (0.4)^2 = 0.16$$

In the context of regression, r^2 is a more meaningful measure than r , because r^2 expresses the ratio of the changes of the dependent variable Y by the independent variable X , and therefore provides a broader action criterion in relation to the explanation of the changes of the dependent variable by the independent variables. gives, while r is such a feature.

Factors that affect the correlation coefficient:

1- The basis of the relationship differs from one society to another. For example, there is a high correlation between calendar age and physiological ability in humans aged 10-16. But there is no correlation between these two variables at the age of 20-26.

2- The distribution of variables is different in different societies. This means that the greater the homogeneity (the lower the variance), the lower the correlation. For example, in a study, if everyone is smart, the range is limited and the value of r is reduced.

3- Correlation between two variables is under the influence of their correlation with the third variable. For example, the correlation between physics and mathematics may be due to the correlation of these variables with intelligence.

4- The use of final groups increases the correlation coefficient. Final scores in small volumes have a great effect on correlation.

5- If n increases, the significance probability of r is higher.

6- When the variables are unrelated, r decreases.

7- Non-linear relationship can make Pearson's correlation coefficient close to zero.

8- The limitation in the range of changes reduces the correlation coefficient. For the existence of a limitation in the scope of change, the variances or standard deviations of the variables whose correlation coefficient is calculated should be investigated. The smallness of the variances can be a sign of limitation in the scope of change.

9- If the variance is small, the correlation coefficient should be interpreted with due caution.

The significance test for correlation in two independent variables is Friedel's t test, and if the groups are independent, Fisher's Z test, and if we compare the correlation coefficient in two dependent groups, it is Student's t test.

The following relationship exists between correlation coefficient (r) and regression coefficient ($\hat{\beta}$).

$$r = \hat{\beta} \frac{S_X}{S_Y} \quad \text{or} \quad \hat{\beta} = r \frac{S_Y}{S_X}$$

Therefore, two standard deviations S_X and S_Y are always positive. r and $\hat{\beta}$ are always similar in algebraic sign.

The regression coefficient ($\hat{\beta}$) can be calculated from the following equation.

$$\hat{\beta} = \frac{\text{Cov}(X, Y)}{\sigma_X^2}$$

If the linear regression equations of Y in terms of X and X in terms of Y are:

$$\hat{Y} = \hat{\alpha}_1 + \hat{\beta}_1 X \quad , \quad \hat{X} = \hat{\alpha}_2 + \hat{\beta}_2 Y$$

We will have:

$$r^2 = \hat{\beta}_1 \cdot \hat{\beta}_2$$

Example 7-6: To study the age and weight of people, four ages of 60, 62, 64 and 70 years were selected and the weight of a number of randomly selected people was determined at each age. According to the statistics in the table below, is there a linear relationship between age and weight?

Age	60	60	60	62	62	62	64	64	64	70	70	70
Weight (pounds)	110	135	120	120	140	130	135	150	145	170	185	160

Test the hypothesis $H_0: B = 0$ using table t and table F:

$$SST = 110^2 + \dots + 160^2 - \frac{1700^2}{12} = 246100 - 240833.3 = 5266.7$$

$$SSR = (5.029)^2(171.67) = 4341.7$$

$$SSD = 5266.7 - 4341.7 = 925$$

$$b = \frac{109380 - \left(\frac{766 \times 1700}{2}\right)}{49068 - \frac{(766)^2}{2}} = 5.029$$

S.V	df	Sum square	Mean square	F
Regression	1	4341.7	4341.7	46.9**
Deviation from regression	10	925	92.5	
Total	11	5266.7		

$$t = \frac{5.029}{\sqrt{\frac{92.5}{171.67}}} = 6.9^{**} > t_{0.01.10} = 3.169$$

Example 7-7: The following table shows the weight and diameter of 10 onions:

Diameter (mm)	51.0	66.2	69.2	69.5	56.9	67.1	58.1	53.9	63.0	60.0
Weight (gram)	63.4	115.3	146.6	132.6	80.7	125.6	80.0	78.7	112.8	96.2

Obtain the 95% confidence interval for α .

$$P(r \leq \alpha \leq r) = 95\%$$

$$Z' = 0.5 \ln \frac{1+0.97}{1-0.97} = 0.5 \ln(65.7) = 0.5(4.18) = 2.09$$

$$Z' = \sqrt{\frac{1}{7}} = 0.38 \sigma$$

$$P(2.09 - 1.96(0.38) \leq Z \leq 2.09 + 1.96(0.38)] = 0.95$$

$$P(1.3 \leq Z \leq 2.8) = 0.95$$

And using the table

$$P(0.86 \leq \alpha \leq 0.99) = 0.95$$

7.25. Correlation analysis of prediction type: regression and prediction

When there is a correlation between two variables, the value of one variable (Y) can be predicted or estimated from another variable (X) through regression, and the higher the correlation between the

variables, the more accurate the prediction. (We predict the grade point average of university courses based on the aptitude test score) This prediction is obtained from the regression line. This line is the best fit among the set of scatter plot points. Another way that does not depend on subjective judgment is the principle of least squares, which minimizes the square of deviations around the regression line. You can get the equation of this line and then draw its graph. The regression line is actually a moving average or least squares line. Regression is actually a line equation. $y = a + bx$

The relationship between the predicted variable (criterion) (Y) and the predictor variable (X) is a function of the sign and intensity of the correlation coefficient. If $r_{xy} = 1$, the regression is zero and if $r_{xy} = 0$, the correlation is complete.

7.26. Prediction of Y from X

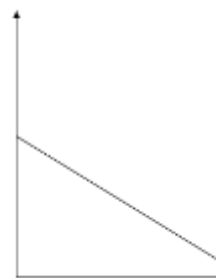
Intuitive method: which is possible by drawing a diagram. Its interpretation is up to the observer.



a)



b)



c)

Figure a) because it is a horizontal line, its slope height is zero. In this case, accurate prediction is possible.

This form has two important factors in prediction:

1- The mean of variable Y, that is, \bar{Y} and the slope of the hypothetical line in the scatter diagram. 2- This graph shows that if the slope of the line (angle coefficient) is zero, for each value of X, the score of Y is equal to the average score of Y.

Figure b) There is a positive and complete relationship between X and Y. If X increases, the value of Y increases by the same amount.

Figure c) There is a negative and complete relationship between X and Y. If X increases, the value of Y decreases by the same amount.

$$\text{Slope and angle coefficient} = \frac{\text{change in } Y}{\text{change in } X} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

In order to predict in general, we must have the following information:

- A) Average scores of Y variable
- b) The slope of the hypothetical line that describes the scatter plot.
- c) The relative position of the individual in variable X (assumed score). This value must be added to or subtracted from the average of Y grades.

$$Y = \bar{Y} + \frac{Y_2 - Y_1}{X_2 - X_1} (X - \bar{X})$$

7.27. Prediction using linear regression equation

$$Y = a + bx$$

a = width from origin (Y value when X is zero).

b = slope of the line (value of variable in Y per unit change in X)

$$b = \frac{Y_2 - Y_1}{X_2 - X_1}$$

This equation is used when all points of a common distribution lie on a straight line.

In the formula of the regression line, a is the point where the line intersects with the vertical axis, and b represents the value that the line climbs for each unit increase in x , and as a result, a is the width from the origin and b is the slope of the line. By knowing these fixed numbers, different x values can be substituted in the equation of the line. and obtained the y values.

In regression analysis, there is a one-way relationship between the variables. Therefore, the predictor variable is unmanipulated. When the regression analysis between a predicted variable (such as literacy) and a predictive variable or criteria (literacy conditions) is considered, it is called simple regression.

When regression analysis is performed between two or more predictor variables with a criterion variable, it is called multivariate regression.

7.28. Types of regression

7.28.1. Simple regression with one variable

In prediction with simple regression (one variable), the individual's uncertain score (Y') is estimated using his performance (X').

The following formula is used to calculate simple regression:

$$Y' = \left[r \frac{S_y}{S_x} \right] (X - \bar{X}) + \bar{Y}$$

Y' = predicted score for the individual

r = correlation between X and Y scores

S_y = standard deviation of Y scores

X = known score of the person in X index

\bar{X} = average scores of X

\bar{Y} = average scores of Y

In regression analysis, variable x is considered as predictor and variable y is considered predicted.

The regression defaults are:

The collected sample should be a true representative of the intended community.

The distribution of both variables should be normal.

There must be a correlation between two variables x and y, and for each quantity x, there must be quantities y with a distribution with relatively equal variance to other x values.

The relationship between variable x and y must be linear (that is, there is no power in the regression equation).

Univariate regression is the basis of regression analysis. In regression analysis, there is an x variable and a y variable.

7.28.2. Multivariate regression

Multiple correlation-regression, y score is predicted using several x scores.

Multiple: One variable is predicted from multiple variables. Academic progress is predicted based on self-confidence, anxiety and creativity.

Multiple variables: multiple independent variables predict multiple dependent variables.

When we calculate the correlation of several x variables with several y variables, it is called central or legal correlation.

The value of multiple correlation is usually higher than the correlation between each of the predictor and criterion variables.

7.29. Prediction of standard scores (Z)

The most preliminary method used in using Pearson's correlation coefficient for prediction is standard scores. This method uses Pearson's correlation coefficient to predict. So that the variable we intend to predict is represented by y and the variable through which the prediction is made is represented by x.

If the correlation between two variables x and y and the standard Z scores for variable x (predictor variable) is known, then the standard Z score for variable y can be calculated using the following formula.

$$Z_Y = (Z_X)(r_{XY})$$

Z_Y = predicted standardized Z score for Y variable

Z_X = standardized Z score for variable X (predictor)

r_{XY} = correlation coefficient between two variables

If the correlation is complete (1), then the prediction standard score is also complete, otherwise it is incomplete. For example, if the

standard Z score is 2 and the complete correlation is 1, then the standard prediction score is 2.

When the correlation between two variables is low, the standard score we predict will be close to the average.

When the correlation between two variables is low, the predicted scores are closer to the mean of the predicted score than the actual score, this phenomenon is called regression.

The degree of correlation between two variables determines the limits or amount of regression.

The phenomenon of regression was first used by Galten. According to Galten's studies, the children of tall parents are tall but not as tall as their parents, similarly the children of short parents are short but not as short as their parents.

The regression line is the line that minimizes the prediction errors (least squares line). It means that the sum of the squared distances of the Ys from the regression line is smaller than the distance of any other line to the Y axes. The regression line is also called the graceful line.

The difference between the actual score (Y) and the predicted score (y') is called the prediction error e .

$$e = y - y'$$

Regression happens when the correlation is not perfect. It means that the grades are selected from the upper and lower groups of the society (high and low range).

If the correlation is zero, the regression is complete and if it is one, the regression is zero.

There is an inverse correlation between the intensity of correlation and regression.

The deviation of the score from the regression line is smaller than the deviation of the score from any other line.

If the distribution of points around the regression line is oval, the correlation between two variables is moderate and its value is close to 0.5.

7.30. Standard error of estimation

Most of the time, there is a difference between the predicted scores and the observed scores. In this case, the difference is called the standard error of estimation.

The higher the correlation between the variables, the lower the prediction error. This error is calculated in two ways.

$$\text{Standard } S_{YX} = \sqrt{\frac{\sum e^2}{N}} \quad \text{or} \quad S_{YX} = \sqrt{\frac{(Y-\bar{Y})^2}{N-2}}$$

$$S_{YX} = S_y \text{ error of prediction } \sqrt{1 - r^2_{xy}}$$

r = Correlation coefficient e = Prediction error

$N-2$ is called the degree of freedom of the standard error of prediction.

The relationship between reliability and error

$\sqrt{1 - r_{xx}}$ $S_E = S$ as it can be seen, r_{xx} is the reliability coefficient.

If the correlation is perfect, the error is zero. It means that the correlation with the error is inversely related.

The standard deviation of the sample and the prediction error have a direct relationship.

There is an inverse relationship between measurement error and reliability coefficient.

Exercises of chapter 7

- 1- If the correlation coefficient of x and y is equal to 0.3 and $z = 3 - 2x$, what is the correlation coefficient of y and z?
- 2- If the regression coefficient is $b = 1.2$ and the sum of the squares of the variable x is equal to 10, what is the sum of the squares of the regression?
- 3- If the total sum of squares is 200 and 80% of the changes in y are explained by x, for $n = 10$, what is the variance of the regression deviation?
- 4- If the sum of squares of deviation from regression between two variables x and y is equal to 2.5 and the total sum of squares is equal to 10, what is the correlation coefficient between the two variables?
- 5- If $SP_{xy} = 15$, $SS_x = 3$, $\bar{X} = 1.5$, and $\bar{Y} = 4.5$, write the equation of the regression line of Y on X?
- 6- If $\sum_{i=1}^n X_i = 56$, $\sum_{i=1}^n X_i^2 = 524$, $\sum_{i=1}^n Y_i = 40$, $\sum_{i=1}^n Y_i^2 = 256$, $\sum_{i=1}^n X_i Y_i = 364$ and y is the independent variable, write the equation of the variable regression line?
- 7- If the sum of squares of regression is equal to 32 and the sum of squares of deviation from regression is equal to 6, what is the value of detection coefficient (r^2)?
- 8- According to the following information, the regression sum of squares is equal to:

$$N = 10 \quad \sum xy = 15$$

$$\sum x^2 = 140 \quad \sum y^2 = 35$$

$$\sum x = 20 \quad \sum y = 10$$

9- The sum of regression and total squares in a study is 100 and 160, respectively. What is the detection or explanation coefficient?

10- The relationship between students' grades and study rate is $\bar{y} = 6+5x$. If $\bar{y}=60$, what is \bar{x} ?

11- If the value of y is $\bar{x} = 5 +0.5y$ and the regression coefficient is $b_{xy} = 0.2$, what is the detection coefficient?

12- In a sample of 27 pairs, a correlation coefficient of 50% has been obtained. Can we consider this value of the correlation coefficient to be significant at the 5% level?

$$t_{26, 0.05} = 2.056 \qquad t_{26, 0.05} = 1.706$$

two-sided

one-sided

13- If the slope of the regression line is equal to $\sum_{i=1}^6 x = 6$, $\sum X^2 = 100$ and 0.5, what is the sum of the squares of the regression?

14- The mean and variance of variable X are 60 and 36, respectively, and for variable Y are equal to 50 and 25, respectively. The covariance of two variables is equal to -15. What is the correlation coefficient of two variables X and Y?

15 - According to the following data, answer the following questions?

$$\sum XY = 103.2, \sum X^2 = 126, \sum X = 30, \sum Y^2 = 104, \sum Y = 20, n = 10$$

- a) Calculate the correlation coefficient?
- b) How much is SSreg?
- c) Get MSe?

16- The variance of two variables x and y obtained from 5 samples was equal to 3 and 12 respectively. If (sp) between two variables is -12.

What is the estimated correlation coefficient between two variables x and y ?

17- If for two traits dry matter amount (y) and day after planting (x), $sp_{xy} = 300$, $ss_x = 200$, $\bar{y} = 3$, find the equation of the regression line?

18- The table below shows the dry weight of the plant (y) in grams on different days (x).

X (Day)	1	2	3	4	5
Y (g)	7	8	14	15	16

What is the estimated weight of the plant after 6 days and the estimated number of days to reach the weight of 19.5 grams?

19- In a study, the regression coefficient was found to be 0.75. If the variance of the error is equal to 2 and the sum of squares of variable x is equal to 8. Is it possible to reject the assumption that there is no relationship between two variables? (table $t = 2$).

20- When the sum of squares of the dependent variable y is equal to 50 ($ss_y = 50$) and the independent variable x is equal to 10 ($ss_x = 10$), assuming that the slope of the regression line of y on x is equal to 2. What percentage of the variation of the dependent variable y is explained by the independent variable x ?

21- If $sp_{xy} = 42$, $ss_y = 64$, and $ss_x = 36$, What is the correlation coefficient between x and y and the regression coefficient of the x line with respect to y respectively?

CHAPTER 8

ANSWERS TO THE EXERCISES

Dr. Saeid HEYDARZADEH¹

¹ - Former Ph.D. Student of Urmia University, Faculty of Agriculture, Department of Plant Production and Genetics, Urmia, Iran
ORCID ID: 0000-0001-6051-7587, e-mail: s.heydarzadeh@urmia.ac.ir

ANSWERS TO THE EXERCISES**CHAPTER 1**

1-

$$R = X_{\max} - X_{\min} = 86 - 65 = 21$$

$$R = CK \quad 21 = 7C \quad C = 3$$

$$\text{The center of the fifth category} = \frac{77+80}{2} = 78.5$$

2-

$$R = X_{\max} - X_{\min} = 47.27 - 8 = 20.8$$

$$R = CK \quad 20.8 = 7C \quad C = 3$$

12% = %28 - %40 The percentage of middle data

$$\frac{f_i}{75} = \frac{12}{100} \quad f_i = 9$$

3- It means that 40% of the data is equal to 70 at most.

4- Our cumulative frequency before the last is equal to $30-3=27$ and the cumulative frequency is obtained from the following relationship:

$$\frac{27}{30} \times 100 = 90$$

5-

$$R = 321 - 520 = 199$$

$$J = \frac{R}{K} = \frac{199}{10} = 19.9$$

6-

$$\sum_{i=1}^5 (4x_i + 6) = 4\sum X_i + (5 \times 6) = 150$$

7-

$$= \sum_{i=3}^5 (a^2 x_i - 4a^2) = a^2 \sum_{i=3}^5 x_i - (3 \times 4a^2)$$

$$= a^2(3-5-2) - 12a^2 = -6a^2 - 12a^2 = -18a^2$$

8-

$$\sum_{i=1}^6 x_i(x_i - 1) =$$

$$= \sum_{i=1}^6 X_i^2 - \sum_{i=1}^6 X_i = 10 - (-4) = 14$$

9-

$$n = 4 + 16 + 20 + 16 + 4 = 60$$

The relative abundance percentage of the second category

$$= \frac{16}{60} \times 100 = 26.6$$

10-

$$\acute{o} = 360^0 - (108^0 + 102^0 + 96^0) = 54^0$$

$$\frac{54}{360} = \frac{X}{100} \quad X = 15$$

11-

$$4(4 + 2 + 3 + 2) = 44$$

12-

First, we write the data in order:

8.0	8.0	8.1	8.2	8.2	8.5	8.6	8.7
9.1	9.1	9.2	9.3	9.3	9.4	9.5	9.5
10.1	10.1	10.2	10.2				

Now we are averaging. In the first row we have 8 to 8 and 8 decimals, in the second row we have 8 to 9 and 8 decimals, in the third row we have 4 to 10 and 4 decimals:

$$\bar{X} = 9 + \{(8 \times 8) + (2.3) + (8 \times 9) (2.7) + (4 \times 10)(0.6) / 20\} = 9.07$$

13-

14-

To find the third category, add the upper and lower bounds of the category together and divide by 2:

$$\frac{17+13}{2} = 15$$

$$n = 4 + 2 + 10 + 8 + 4 + 6 = 34$$

$$\text{Relative frequency} = \frac{10}{34} = \frac{5}{17}$$

CHAPTER 2

1-

According to the problem, we have:

$$Y_i = \frac{(x_i - 100)}{100}$$

$$\mu_y = (\mu_x \times 100) + 100 = (10 \times 100) + 100 = 1100$$

$$\delta_y^2 = \delta_x^2 \times (100)^2 = 25 \times 10000 = 250000$$

2-

$$M_h = \frac{N}{\sum \frac{1}{X_i}} = \frac{5}{\frac{1}{4} + \frac{1}{5} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}} = 5.21$$

3-

$$Y = 2X_A + 3X_B - 3$$

$$\mu_y = 2\mu_A + 3\mu_B - 3$$

$$\mu_Y = 2 \times 4 + 3 \times 6 - 3 = 8 + 18 - 3 = 23$$

4-

$$Z = \frac{X - \mu}{S}$$

$$\begin{array}{l}
 1 = \frac{70 - \mu}{S} \quad 70 - \mu = S \\
 2 = \frac{80 - \mu}{S} \quad 80 - \mu = 2S
 \end{array}
 \quad
 \begin{array}{c}
 \boxed{(-1) X} \rightarrow \\
 \left\{ \begin{array}{l} -70 + \mu = -S \\ 80 - \mu = 2S \end{array} \right.
 \end{array}$$

$$\delta = 10$$

$$70 - \mu = S$$

$$70 - \mu = 10$$

$$\mu = 60$$

5-

$$CV_A = \frac{280}{1495} \times 100 = 18.7$$

$$CV_B = \frac{310}{1875} \times 100 = 16.53$$

Therefore, lamp A has more changes than lamp B.

6-

$$\text{Log } m_g = \frac{\sum \log \bar{X}_i F_i}{N}$$

$$N = \frac{\sum \log \bar{X}_i F_i}{\text{Log } m_g}$$

$$N = \frac{132.76}{\text{Log } 23.37} = 97$$

7-

a):

$$\mu = \frac{\sum f_i x_i}{\sum f_i} = \frac{270}{30} = 9$$

b): 9

c):

$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{\sum f_i} = \frac{5 \times (1-9)^2 + 6 \times (5-9)^2 + \dots + 5 \times (17-9)^2}{30} = \frac{832}{30} = 73.27$$

d):

$$M = \frac{\sum f_i |x_i - \text{median}|}{\sum f_i} = \frac{5 \times |1-9| + 6 \times |5-9| + \dots + 5 \times |17-9|}{30} = \frac{128}{30} = 4.27$$

e): 9

Because the distribution is normal, therefore the median, mode and mean will be equal.

f): Normal distribution

8-

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

$$3 = \frac{49.5 - \frac{(\sum X_i)^2}{10}}{9}$$

$$\frac{(\sum X_i)^2}{10} = 49.5 - 27 = 22.5$$

$$\sum X = \sqrt{225} = 15$$

$$\bar{X} = \frac{\sum X}{N} = \frac{15}{10} = 1.5$$

9-

$$Q_3 = L + \left(\frac{P_n - F_c}{f_i} \right) j$$

$$P_n = \frac{3}{4} N = \frac{3}{4} \times 100 = 75$$

$$Q_3 = 44 + \left(\frac{75 - 67.5}{13} \right) \times 8 = 43.9$$

10-

Relative frequency $f_i = 24$

$$f = \frac{f_i}{\sum f_i} \times 100$$

$$24 = \frac{x}{150} \times 100 \quad f_i = 36$$

11-

$$\sum_{i=1}^4 (X_i - 3)(2Y_i + 1)$$

$$\sum_{i=1}^4 (2xy + x_i - 6y - 3) = \sum_{i=1}^4 2xy + \sum_{i=1}^4 x_i + \sum_{i=1}^4 -6y + \sum_{i=1}^4 -3$$

$$2 \sum_{i=1}^4 xy + \sum_{i=1}^4 x_i - 6 \sum_{i=1}^4 y_i - 3 = 10 + 7 + 18 - 3 = 32$$

12-

When all data is multiplied or divided by a certain value, its C.V value remains constant.

Therefore: $20 \text{ CV} = \frac{3}{15} = 0.2 \quad 20\%$

13-

$$\bar{x} = \sum P_i X_i = \left(\frac{1}{3} \times 2\right) + \left(\frac{1}{2} \times 3\right) + \left(\frac{1}{6} \times 11\right) = 4$$

$$\delta^2_x = \sum P_i (X_i - \bar{X})^2 = \left\{\frac{1}{3} \times (2 - 4)^2\right\} + \left\{\frac{1}{2} \times (3 - 4)^2\right\} + \left\{\frac{1}{6} \times (11 - 4)^2\right\}$$

$$\delta^2_x = \frac{4}{3} + \frac{1}{2} + \frac{49}{6} = \frac{60}{6} = 10$$

CHAPTER 3

1-

$$(n-1)! = (6-1)! = 5! = 120$$

2-

The probability that the rabbit is white in cages A and B

$$= \frac{3}{5} \times \frac{2}{6} \times \frac{2}{4} = \frac{12}{120}$$

The probability that the rabbit is white in cages A and C

$$= \frac{3}{5} \times \frac{4}{6} \times \frac{2}{4} = \frac{24}{120}$$

The probability of the rabbit being white in cages B and C

$$= \frac{2}{5} \times \frac{2}{6} \times \frac{2}{4} = \frac{8}{120}$$

The probability that the rabbit is white in all three cages

$$= \frac{3}{5} \times \frac{2}{6} \times \frac{2}{4} = \frac{12}{120}$$

The probability of choosing at least two white rabbits

$$= \frac{12}{120} + \frac{24}{120} + \frac{8}{120} + \frac{12}{120} = \frac{56}{120} = 0.47$$

3-

$$P = (\text{Rainfall on the first day}) = 0.5$$

$$P = (\text{Rainfall on the second day} \cap \text{Rainfall on the first day}) = 0.4$$

$$P = (\text{Rainfall on the first day} / \text{rainfall on the second day}) = \frac{P(A \cap B)}{P(A)} = \frac{0.4}{0.5} = 0.8$$

4-

$$P(\text{defective} / A) + P(B) P(\text{defective} / B) = (0.3 \times 0.03) + (0.7 \times 0.04) = 0.03$$

$$P = (\text{defective})P(A)$$

5-

$$C^3_{10} = \frac{10!}{3!(10-3)!} = 120$$

6-

$$C^3_6 \times C^2_5 = \frac{6!}{3!(6-3)!} \times \frac{5!}{2!(5-2)!} = 20 \times 10 = 200$$

7 -

Because there is a boy in this family, so the probability of having one or two girls in this family should be calculated.

The possibility of having a daughter $C^1_2 \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \frac{1}{2}$

The possibility of having two daughters $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$

The probability of having at least one daughter $\frac{2}{4} + \frac{1}{4} = \frac{3}{4}$

8-

$$P(\text{White dice being odd}) = \frac{3}{6} = \frac{1}{2}$$

$$P(\text{The number of the red dice is a multiple of 3}) = \frac{2}{6} = \frac{1}{3}$$

$$P = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

9-

The flags of three Asian countries count as one flag.

$$3! \times (4+1+1)! = 3! \times 6!$$

10-

There are six different combinations to remove the balls, and the probability of the six combinations is equal to:

$$P(\text{Red-black-white}) = \frac{5}{15} \times \frac{6}{14} \times \frac{4}{13} = \frac{120}{2720}$$

$$P = 6 \times \left(\frac{120}{2720}\right) = 0.264$$

11-

The probability of cloudy weather is $P(A) = 0.3$

The probability of cloudiness and rain $P(A \cap B) = 0.2$

$$P(A \cap B) = P(A) - P(B)$$

$$P(B) = \frac{P(A \cap B)}{P(A)} = \frac{0.2}{0.3} = 0.67$$

12-

Number of possible states

$$n(A) = (6 \cdot 6) + (6 \cdot 5) + (6 \cdot 4) + (6 \cdot 3) + (5 \cdot 6) + (5 \cdot 5) + (5 \cdot 4) + (4 \cdot 6) + (4 \cdot 5) + (3 \cdot 6)$$

$$P = \frac{n(A)}{n(S)} = \frac{10}{36}$$

13-

Modified tree probability $P(B) = 0.6$

The probability of native trees is $P(N) = 0.4$

The probability of the height of modified trees exceeding 3 meters $P(H/B) = 0.05$

The probability of the height of native trees exceeding 3 meters $P(H/N) = 0.1$

$$P(B/H) = \frac{P(H/B) \cdot P(B)}{P(H/B) \cdot P(B) + P(H/N) \cdot P(N)} = \frac{(0.05)(0.6)}{(0.05)(0.6) + (0.1)(0.4)} = \frac{0.03}{0.07} = \frac{3}{7} = 0.429$$

14-

$$P(\text{All three healthy}) = \frac{12}{20} \times \frac{11}{19} \times \frac{10}{18} = \frac{11}{57}$$

15-

$$P(\text{cloves}) = \frac{7}{12}$$

$$P(\text{Red}) = \frac{9}{12}$$

$$P(\text{red} \cap \text{Clove}) = \frac{6}{12}$$

$$P(\text{Red} \cup \text{cloves}) = \frac{7}{12} + \frac{9}{12} - \frac{6}{12} = \frac{10}{12} = \frac{5}{6}$$

16-

a):

$$\text{Probability that both pencils are blue} = \frac{4}{9} \times \frac{3}{8} = \frac{12}{72}$$

$$\text{Probability that both pencils are red} = \frac{5}{9} \times \frac{4}{8} = \frac{20}{72}$$

The probability that both pencils are the same color

$$= \frac{12}{72} + \frac{20}{72} = \frac{32}{72} = \frac{4}{9}$$

b):

The probability that the first is blue and the second is red

$$= \frac{4}{9} \times \frac{5}{8} = \frac{20}{72}$$

The probability that the first is red and the second is blue

$$= \frac{5}{9} \times \frac{4}{8} = \frac{20}{72}$$

The probability that one of the pencils is blue and the other is red

$$= \frac{20}{72} \times \frac{20}{72} = \frac{40}{72} = \frac{5}{9}$$

c):

$$\text{Probability that both pencils are red} = \frac{5}{9} \times \frac{4}{8} = \frac{20}{72}$$

1- The probability that both pencils are red = the probability that only one of the pencils is blue

$$= 1 - \frac{20}{72} = \frac{52}{72} = \frac{13}{18}$$

d):

The probability that the first is red and the second $= \frac{5}{9} \times \frac{4}{8} = \frac{20}{72}$

is blue

17-

Short $P(A) = 0.25$

Late $P(B) = 0.15$

$P(A \cap B) = 0.1$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = 0.25 + 0.15 - 0.1 = 0.3$

18-

$$5! \times 4! \times 3! \times 2! = 34560$$

19-

$$(4 + 3 + 5)! = 12!$$

20-

$$P_n^{r_1, r_2, r_3} = \frac{n!}{r_1! r_2! r_3!}$$

$$P_{10}^{(3,4,2)} = \frac{10!}{3! 4! 2!} = 12600$$

21-

Probability of balanced coin and front observation

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The possibility of an unbalanced coin and front observation

$$= \frac{1}{2} \times \frac{6}{10} = \frac{3}{10}$$

The probability of front observation

$$= \frac{1}{4} + \frac{3}{10} = \frac{5+6}{20} = \frac{11}{20}$$

22-

$$3! \times 2! = 3 \times 2 \times 1 \times 2 \times 1 = 12$$

23-

$$\begin{array}{cccccc} \text{M} & \text{F} & \text{M} & \text{F} & \text{M} & \text{F} & \text{M} \\ \frac{4}{7} & \times \frac{3}{6} & \times \frac{3}{5} & \times \frac{2}{4} & \times \frac{2}{3} & \times \frac{1}{2} & \times 1 = \frac{144}{5070} = \frac{1}{35} \end{array}$$

24-

$$1 \text{ experimental and 3 mathematical} = \frac{C_7^3 C_5^1}{C_{12}^4} = \frac{35 \times 5}{495}$$

$$4 \text{ Mathematics} = \frac{C_7^4}{C_{12}^4} = \frac{35}{495}$$

$$\text{At least 3 mathematics} = \frac{175+35}{495} = \frac{14}{33}$$

25-

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$0.6 = 0.4 + P(B) - P(A \cap B)$$

$$0.6 = 0.4 + 0.3 - 0.12$$

$$0.6 = 0.6$$

26-

$$C^2_9 = \frac{9!}{2!7!} = 36$$

27-

$$\begin{aligned} n(A \cup B) &= n(A) + n(B) - n(A \cap B) \text{ Both rejected} \\ &= 241 + 271 - 111 = 401 \end{aligned}$$

$$\text{Rejected people} = 742 - 401 = 341$$

28-

$$C^3_5 C^4_6 = 3! 4!$$

29-

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{30}{40} + \frac{20}{40} - \frac{\frac{30}{2}}{40} = \frac{35}{40} = \frac{7}{8}$$

30-

$$4! = 24$$

CHAPTER 4

1-

$$Z = \frac{x-\mu}{\delta} \qquad 1.96 = \frac{x-12}{4} \qquad x = 19.84$$

2-

$$\bar{x} = \frac{\sum x}{n} \qquad 40 = \frac{\sum x}{200} \qquad \sum x = 800 \quad \text{Sum of primary scores}$$

$$800 - (53 - 43) = 7990 \quad \text{Total correct scores}$$

$$\bar{x} = \frac{7990}{200} = 39.95$$

3-

Considering that the scores have a normal distribution and the normal distribution is divided into two equal parts at the point $Z = 0$, then:

$$Z = \frac{x-\mu}{\delta} \qquad 0 = \frac{x-12}{4} \qquad x = 12$$

4-

$$Y_i = \sum X_i + 5 \qquad S_x^2 = 10 \qquad \mu_x = 6 \qquad n = 8$$

$$E(\sum X_i + 5) = E X_i + E(5)$$

$$E(\sum X_i) = EX_1 + EX_2 + \dots + EX_n = n\mu$$

$$E(\sum X_i + 5) = n\mu + 5 = (8 \times 6) + 5 = 53$$

5-

$$\mu = 60$$

$$\delta = \sqrt{\delta^2} = \sqrt{25} = 5$$

$$Z = \frac{x - \mu}{\delta} = \frac{60 - 60}{5} = 0$$

$$P(Z \geq 0) = 0.5$$

$$N = 0.5 \times 200 = 100$$

6-

$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{25}{\sqrt{16}} = \frac{25}{4} = 6.25$$

$$Z = \frac{\bar{x} - \mu}{\delta_{\bar{x}}} \quad -2 = \frac{\bar{x} - 110}{6.25} \quad \bar{x} - 110 = -12.5 \quad \bar{x} = 97.5$$

7-

$$P(x \geq 350) = 5\% \quad z = 1.64$$

$$Z = \frac{x - \mu}{\delta} \quad 1.64 = \frac{350 - \mu}{8} \quad \mu = 336.88$$

8-

$$P(\text{A number greater than 4}) = \frac{2}{6} = \frac{1}{3} \quad . \quad q = \frac{2}{3}$$

$$\delta^2 = npq = 20 \times \frac{1}{3} \times \frac{2}{3} = \frac{40}{9}$$

9-

$$\mu = np = 144 \qquad \delta = \sqrt{npq} = 6$$

$$2 = npq \quad 36 = 144 \times q \quad q = \frac{36}{144} = \frac{1}{4} \delta$$

$$P = 1 - q = 1 - \frac{1}{4} = \frac{3}{4}$$

$$= np \quad 144 = n \times \frac{3}{4} \quad n = 192 \mu$$

10-

$$P\left(\frac{7.5-5.5}{1} \leq z \leq \frac{5.5-5.5}{1}\right)$$

$$P(2 \leq z \leq 0)$$

$$P(z \leq 0) - p(z \geq 2) = 0.5 - 0.0228 = 0.4772$$

11-

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} = \frac{26}{\sqrt{20}} = 5.81$$

$$Z = \frac{\bar{x} - \mu}{\delta_{\bar{x}}} \quad -2 = \frac{\bar{x} - 115}{5.81} \quad \bar{x} - 115 = -11.62 \quad \bar{x} = 103.38$$

12-

$$Y = 2\bar{X}_1 + 3\bar{X}_2 - 3$$

$$E(Y) = 2E(\bar{X}_1) + 3E(\bar{X}_2) - 3$$

$$E(Y) = (2X \mu_1) + (3X \mu_2) - 3$$

$$E_y = (2 \times 3) + (3 \times 8) - 3$$

$$E_y = 27$$

13-

$$E(X) = \mu = \sum P_i X_i = \left(\frac{1}{5} \times 10\right) + \left(\frac{2}{5} \times 20\right) + \left(\frac{3}{5} \times 30\right) = 2 + 8 + 18 = 28$$

14-

$$P(Z \geq 1.64) = 0.05 \quad p(Z \geq 1.96) = 0.025$$

$$Z = \frac{x - \mu}{\delta} \quad 1.64 = \frac{23.28 - \mu}{2} \quad \mu = 20$$

15-

$$P(Z \geq 0.8) = 0.2$$

$$P(Z \geq 0.27) = 0.1$$

$$Z = \frac{x - \mu}{\delta} \quad 0.8 = \frac{x - 115}{5} \quad x = 15.5$$

16-

$$P(Z \geq 1) = 0.16$$

$$P(Z \geq 1.64) = 0.05$$

$$Z = \frac{x - \mu}{\delta}$$

$$Z = \frac{24 - 30}{6} = -1$$

68% $\mu \pm 1\delta$

$$Z = \frac{36-30}{6} = 1$$

17-

$$P = 0.6, \quad q = 1 - p = 0.4$$

$$P = C_8^6(0.6)^6(0.4)^2 + C_8^7(0.6)^7(0.4) + (0.6)^8$$

$$P = 0.21 + 0.09 + 0.027 = 0.315$$

18-

$$P(|z| \geq 1.64) = 0.1$$

$$P(|z| \geq 1.96) = 0.05$$

Wrong centers: $Z\check{\sigma}_{\bar{x}} = 2$

$$1.64 \times \check{\sigma}_{\bar{x}} = 2 \quad \check{\sigma}_{\bar{x}} = 0.82$$

$$\check{\sigma}_{\bar{x}} = \frac{\check{\sigma}_x}{\sqrt{n}} \quad 0.82 = \frac{10}{\sqrt{n}} \quad n = 67.24$$

19-

$$P(Z \geq Z_1) = 0.6 \quad P(Z \leq Z_2) = 0.8$$

Because the value of $P(Z \geq Z_1) = 0.6$ is greater than 0.5, then Z_1 is on the left side of $Z = 0$ and its value is negative, and because $P(Z \leq Z_2) = 0.8$ is greater than 0.5, so Z_2 is on the right side of $Z = 0$ and its value is positive.

20-

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} = \frac{4}{\sqrt{16}} = 1$$

$$Z = \frac{\bar{x} - \mu}{\delta_{\bar{x}}} = \frac{0.35 - 0.4}{1} = 0.05$$

The claim is rejected because the absolute value of the calculated z is greater than the table z .

21-

$$(Z_{0.05} = 1.64)$$

According to the form of the problem, computers should be guaranteed whose performance is lower than the average of the distribution, therefore:

$$Z = \frac{x - \mu}{\delta} \quad -1.64 = \frac{x - 10}{3} \quad x = 5$$

22-

$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{1}{100} = \frac{1}{10} = 0.1$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad z = \frac{17 - 18}{0.1} = -10$$

Yes, because $|z|$ Calculated is greater than any Z value in the table.

23-

$P = P(\text{six girls}) + P(\text{one boy and six girls}) + P(\text{two boys and one girl})$

$$P = \binom{1}{2}^6 + c_6^5 \binom{1}{2}^1 \binom{1}{2}^5 + c_6^2 \binom{1}{2}^2 \binom{1}{2}^4$$

$$P = 64 + \frac{6}{64} + \frac{15}{64} = \frac{22}{64} = \frac{11}{32}$$

24-

$$t_{df} = 5, 0.05 > t_{df} = 15, 0.05$$

$$Z_{df} = 5, 0.05 > Z_{df} = 5, 0.05$$

25-

$$n = 30$$

$$95\% \quad 2$$

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad 2 = \frac{26 - 14}{\frac{2}{X}} \quad X = 270$$

26-

$$P = \frac{1}{2} \quad q = \frac{1}{2} \quad x = 6 \quad n = 10$$

$$C_{10}^6 P^6 q^4 = \frac{10!}{6! 4!} \times \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4$$
$$= \frac{10 \times 9^3 \times 8 \times 7}{4 \times 3 \times 2} = 210 \times \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4 = 0.205$$

27-

$$Z_1 = 1.6$$

28-

29-

$$\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}} = \frac{3}{\sqrt{9}} = 1$$

$$Z = \frac{\bar{x} - \mu}{\delta_{\bar{x}}} = \frac{36 - 40}{1} = -4 \quad |\Sigma| > 2.33$$

30-

$$S_{\bar{x}} = 0.4, \quad n = 10, \quad \mu = 16$$

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} \quad 0.4 = \frac{S_x}{\sqrt{10}} \quad S_x = 1.26$$

$$C.V\% = \frac{S_x}{\mu} \times 100 = \frac{1.26}{16} \times 100 = 7.9$$

31-

$$\bar{X} = 50, S = 3.3, S^2 = 10.89, n = 100$$

$$\mu = \bar{x} = 50$$

$$\delta^2 = \frac{n}{n-1} S^2 = \left(\frac{100}{99}\right) (10.89) = 11$$

CHAPTER 5

1-

The first type of error occurs when we reject the correct hypothesis zero (H_0) and in other words accept the false hypothesis one (H_1).

2-

$$\mu = \frac{3+6+5+7+4}{5} = \frac{25}{5} = 5$$

$$\delta^2 = \frac{\sum(X_i - \mu)^2}{n} = \frac{(3-5)^2 + (6-5)^2 + (5-5)^2 + (7-5)^2 + (4-5)^2}{5} = \frac{10}{5} = 2$$

According to the central limit theorem, if all binary samples are extracted with replacement, the variance of the frequency distribution of the means is equal to:

$$\delta_{\bar{X}}^2 = \frac{\delta_x^2}{n} = \frac{2}{5} = 0.4$$

3-

$$\mu_{\bar{x}} = \mu = 10$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} = \frac{25}{64}$$

4-

$$\bar{x} = 8 \qquad SX^2 = 4.2$$

$$\mu = \bar{x} = 8$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \quad 4.2 = \frac{\sigma_x^2}{5} \quad \sigma_x^2 = 21$$

5-

$$P(Z \geq 1.64) = 0.05$$

$$P(Z \geq 2.23) = 0.01$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad Z = \frac{5.65 - 5.5}{0.35} = 0.43$$

No, because the probability of Z's greater than the calculated Z is less than 0.01.

6-

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}} = \sqrt{\frac{27}{9}} = \sqrt{3}$$

7-

$$P(|Z| \geq 1.64) = 0.1, \quad P(|Z| \geq 1.96) = 0.05$$

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} = \frac{2}{\sqrt{25}} = \frac{2}{5} = 0.4$$

$$\pm Z\delta_{\bar{x}} \quad 10 \pm (1.64 \times 0.4) \quad 10 \pm 0.66 \quad \left\{ \begin{array}{l} 10.66 \\ 9.34 \end{array} \right.$$

8-

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1$$

$$\bar{X} - ZS_{\bar{x}} \leq \mu \leq \bar{X} + ZS_{\bar{x}}$$

$$40 - (2X_1) \leq \mu \leq 40 + (2X_1)$$

$$38 \leq \mu \leq 42$$

9-

$$\delta_{\bar{x}} = \frac{\delta_x}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1$$

$$P(\bar{X} \geq 126) = P(Z \geq \frac{\bar{X} - \mu}{\delta_{\bar{x}}} = \frac{126 - 127}{1} = -1$$

$$P(Z \geq -1) = 0.5 + 0.3413 = 8413$$

CHAPTER 6

1-

Because two parameters (mean and standard deviation) are used to calculate the expected values, the degree of freedom is equal to:

$$df = k - 1 - 2 = k - 3$$

2-

$$df = (c - 1)(r - 1)$$

3-

$$S_{\bar{X}} = \sqrt{\frac{S_X^2}{n}} = \sqrt{\frac{1000}{50}} = \sqrt{20}$$

$$t = \frac{\bar{x} - \mu}{S_{\bar{X}}} = \frac{12 - 0}{\sqrt{20}} = 2.67 \quad \alpha \approx 0.01$$

4-

$$S_P^2 = \frac{S_1^2 + S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

The sum of intragroup ss divided by the sum of intragroup degrees of freedom.

5-

$$P(128.9 \leq \mu \leq 171.1) = 95 \%$$

$$\begin{cases} \bar{x} - tS_{\bar{X}} = 128.9 \\ \bar{x} + tS_{\bar{X}} = 171.1 \end{cases} \quad (\bar{X} - tS_{\bar{X}}) + (\bar{X} + tS_{\bar{X}}) = 128.9 + 171.1$$

$$2\bar{X} = 300 \quad \bar{X} = 150$$

6-

$$S_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{9}} = 1$$

$$\bar{x} \pm t_{(n-1)} S_{\bar{X}}$$

$$1 \pm (3.355 \times 1) \quad 4.355 \text{ and } -2.355$$

7-

$$\delta_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.4)(0.6)}{100}} = 0.049$$

$$P(0.32 \leq p \leq 0.47) = P\left(\frac{0.32-0.4}{0.049} \leq Z \leq \frac{0.47-0.4}{0.049}\right) = P(-1.63 \leq Z \leq 1.43)$$

$$= 1 - \{P(Z \geq 1.63) + P(Z \geq 1.43)\} = 1 - \{0.0516 + 0.0764\} = 0.872$$

8-

$$\bar{x} + t S_{\bar{X}} = 170$$

$$\bar{x} - t S_{\bar{X}} = 125$$

$$\bar{x} + t S_{\bar{X}} + (\bar{x} - t S_{\bar{X}}) = 170 + 125$$

$$2\bar{X} = 295 \quad \bar{X} = 147.5$$

$$\bar{x} + t S_{\bar{X}} = 170 \quad 147.5 + (2.306 \times S_{\bar{X}}) = 170 \quad S_{\bar{X}} = 9.76$$

$$S_{\bar{X}} = \frac{S_x}{\sqrt{n}} \quad 9.76 = \frac{S_x}{\sqrt{9}} \quad S_x = 27.29$$

9-

$$\chi^2_{0.975, 2} = 0.0506, \quad \chi^2_{0.025, 2} = 7.38$$

$$\mu = \frac{\sum x}{n} = \frac{4.1 + 5.2 + 10.2}{33} = 6.5$$

$$\chi^2 = \frac{\sum (X_i - \mu)^2}{\sigma_0^2} = \frac{(4.1 - 6.5)^2 + (5.2 - 6.5)^2 + (10.2 - 6.5)^2}{4} = \frac{21.14}{4} = 5.29$$

No, because the calculated K-score is between the two K-scores in the table.

10-

$$Z = \frac{\bar{X} - \mu}{\delta_{\bar{X}}} \quad 2.58 = \frac{0.5}{\delta_{\bar{X}}} \quad \delta_{\bar{X}} = 0.1938$$

$$\delta_{\bar{X}} = \frac{\delta_x}{\sqrt{n}} \quad 0.1938 = \frac{2.5}{\sqrt{n}} \quad n = 166$$

11-

Non-transgenic	Transgenic	
16	8	Men
14	2	Women

Since the degree of freedom is equal to one, Yates correction is used.

$$\chi^2 = \sum \frac{(|o_i - e_i| - 0.5)^2}{e_i} = \frac{(|8 - 5| - 0.5)^2}{5} + \frac{(|2 - 5| - 0.5)^2}{5} + \frac{(|16 - 15| - 0.5)^2}{15} + \frac{(|14 - 15| - 0.5)^2}{15} = \frac{38}{15} = 2.53$$

Because the calculated chi-square is smaller than the chi-square of the table, the results are not significant at the 5% level, and therefore there is no sufficient reason to reject the hypothesis H_0 , that is, the preference of a product depends on gender.

12-

$$Pz < -1.25 = 0.1056$$

$$Pz > 1.25 = 0.1093$$

$$Pz > 0.16 = 0.4364$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\begin{aligned} P(\bar{x}_1 - \bar{x}_2 > 0) &= P\left(z \geq \frac{0-0.5}{\sqrt{4\left(\frac{1}{50}+\frac{1}{50}\right)}}\right) = P(Z \geq -1.25) = 1 - P(Z \leq -1.25) \\ &= 1 - (0.1056) = 0.8944 \end{aligned}$$

13-

$$P(Z \geq 2.58) = 0.005$$

$$P(Z \geq 1.96) = 0.025$$

$$= Z \delta\bar{X} \quad 2.58 \times \delta\bar{X} = 0.01 \quad \delta\bar{X} = 0.0039 \quad \text{Maximum error}$$

$$\delta\bar{X} = \frac{\delta\bar{x}}{\sqrt{n}} \quad 0.0039 = \frac{0.05}{\sqrt{n}} \quad n = 167$$

14-

$$n = 64, \quad p = \frac{1}{2}, \quad q = \frac{1}{2}$$

$$\mu = np = 64 \times \frac{1}{2} = 32$$

$$= \sqrt{npq} = \sqrt{64 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{16} = 4\delta$$

$$Z = \frac{X \pm \mu}{\delta} \quad 2 = \frac{X \pm 32}{4} \quad X \pm 32 = 8 \quad X = 40 \quad X = 24$$

15-

For the normality test, X_2 is used, whose degree of freedom is equal to $df = 8 - 1 - 2 = 5$, where the number 8 is the number of classes and the number 2 indicates that in calculating the expected values of two parameters (mean and standard deviation) used.

16-

$$df_B = t - 1 = 6 - 1 = 5$$

$$df_w = t(r - 1) = 6(3 - 1) = 12$$

17-

$$P(Z \geq 2.23) = 0.0099$$

$$P(Z \geq 1.05) = 0.146$$

$$P = 0.9, q = 0.1, n = 100$$

$$\delta_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.9)(0.1)}{100}} = 0.03$$

$$P(Z \leq 0.83) = P\left(Z \leq \frac{0.83 - 0.9}{0.03}\right) = P(Z \leq -2.33) = P(Z \geq 2.33) = 0.0099$$

$$N = (1000)(0.0099) = 9.9 \approx 10$$

18-

The t-test is performed with paired measurements, the degree of freedom of which is $n-1$.

$$n - 1 = 10 - 1 = 9$$

19-

$$n_1 = n_2$$

$$S^2_{\bar{X}_1 - \bar{X}_2} = S^2_P \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 2.5 \left(\frac{1}{5} + \frac{1}{5} \right) = 1 \quad S_{\bar{X}_1 - \bar{X}_2} = \sqrt{1} = 1$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{12 - 6}{1} = 6$$

CHAPTER 7

1-

Adding and subtracting a constant value with a variable or multiplying and dividing a positive constant value in a variable has no effect on the correlation coefficient of that variable with another variable. But multiplying or dividing a negative constant by a variable changes the sign of the correlation coefficient.

2-

$$SSR = b^2 SSX$$

$$SSR = (1.2)^2 \times 10 = 14.4$$

3-

$$SSR = \frac{80}{100} \times 200 = 160$$

$$SSD = SST - SSR = 200 - 160 = 40$$

$$MSD = \frac{SSD}{n-2} = \frac{40}{10-2} = 5$$

4-

$$SSR = SST - SSD = 10 - 2.5 = 7.5$$

$$R^2 = \frac{SSR}{SST} = \frac{7.5}{10} = 0.75$$

$$r = \sqrt{r^2} = \sqrt{0.75} = 0.87$$

5-

$$b = \frac{SP_{xy}}{SS_x} = \frac{15}{3} = 5$$

$$a = \bar{y} - b \bar{x} = 4.5 - (5 \times 1.5) = -3$$

$$y = -3 + 5x$$

6-

$$b = \frac{SP_{xy}}{SS_y} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}} = \frac{364 - \frac{(256)(40)}{8}}{256 - \frac{(40)^2}{8}} = \frac{364 - 280}{56} = 1.5$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{56}{8} = 7$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{40}{8} = 5$$

$$\bar{x} = a + b\bar{y} \quad 7 = a + (1.5)(5) \quad a = -0.5$$

7-

$$SST = SSR + SSD = 32 + 6 = 38$$

$$R^2 = \frac{SSR}{SST} = \frac{32}{38} = 0.84$$

8-

$$\begin{aligned} N &= 10 & \sum xy &= 15 \\ \sum x^2 &= 140 & \sum y^2 &= 35 \\ \sum x &= 20 & \sum y &= 10 \end{aligned}$$

$$SSR = \frac{\left\{ \sum XY - \frac{(\sum X \sum Y)}{n} \right\}^2}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{\left\{ 15 - \frac{(20 \times 10)}{10} \right\}^2}{140 - \frac{(20)^2}{10}} = \frac{(-5)^2}{100} = \frac{25}{100} = 0.25$$

9-

$$r^2 = \frac{SSR}{SST} = \frac{100}{160} = 0.625 = 62.5\%$$

10-

$$a = 5, b = 6, \bar{y} = 60$$

$$\bar{y} = a + b\bar{x}$$

$$60 = 5 + 6\bar{x} \quad 6\bar{x} = 55 \quad \bar{x} = 9.17$$

11-

$$b_{xy} = 0.5, b_{yx} = 2$$

$$r^2 = b_{xy} \cdot b_{yx} = 0.5 \times 0.2 = 0.1$$

12-

$$r = 5$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-25}{27-2}} = 0.9797$$

$$t = \frac{r}{S_r} = \frac{5}{0.9797} = 5.10$$

13-

$$SS_X = \sum X^2 - \frac{(\sum X)^2}{8} = 100 - \frac{(6)^2}{6} = 94$$

$$SSR = b^2 SS_X = (0.5)^2 \times 94 = 23.5$$

14-

$$R = \frac{6_{xy}}{\sqrt{S_X^2 S_Y^2}} = \frac{-15}{\sqrt{36 \times 25}} = \frac{-15}{30} = -0.5$$

15-

$$\sum XY = 103.2, \quad \sum X^2 = 126, \quad \sum X = 30, \quad \sum Y^2 = 104, \quad \sum Y = 20, \quad n = 10$$

a)

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{SS_X SS_Y}} = \frac{103.2 - \frac{20 \times 30}{10}}{\sqrt{\left(126 - \frac{(30)^2}{10}\right) \left(104 - \frac{(20)^2}{10}\right)}}$$

$$r = \frac{103.2 - 60}{\sqrt{36 \times 64}} = \frac{43.2}{48} = 0.9$$

b)

$$SS_{\text{reg}} = \frac{\left[\sum xy - \frac{\sum x \sum y}{n} \right]^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$SS_{\text{reg}} = \frac{\left[103.2 - \frac{30 \times 20}{10} \right]^2}{126 - \frac{(30)^2}{10}}$$

$$SS_{\text{reg}} = \frac{(43.2)^2}{36} = 51.84$$

c)

$$R^2 = \frac{SSR}{SST} (0.9)^2 = \frac{51.84}{SST} \quad SST = 64$$

$$SST = SS_{\text{reg}} + SSe \quad SSe = 64 - 51.84 = 12.16$$

$$MSe = \frac{SSe}{n-2} = \frac{12.16}{10-2} = 1.52$$

16-

$$\sigma_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{sp}{5-1} = \frac{-12}{4} = -3$$

$$r = \frac{\sigma_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{-3}{\sqrt{3 \times 12}} = \frac{-3}{6} = -0.5$$

17-

Because the amount of dry matter (Y) on the first day after planting (X) is equal to zero, the value of a = 0. The value of b and the equation of the regression line is equal to:

$$b = \frac{SP_{xy}}{SS_x} = \frac{300}{200} = 1.5$$

$$\bar{y} = 1.5x$$

18-

$$\bar{y} = 12 \quad y = a + bx$$

$$\bar{x} = 3 \quad b = \frac{\text{cov } x}{s_x^2} \quad s_x^2 = \frac{6.25}{2.5} = 2.5$$

$$\frac{10 + 4 + 3 + 8}{4} = 6.25$$

$$12 = a + 2.5x \quad a = 4.5 \quad y = 4.5 + 2.5x$$

$$Y = 4.5 + 2.5(6) = 19.5$$

19-

$$t = \frac{b}{\frac{\text{MSD}}{\sqrt{\sum(X-\bar{X})^2}}} = \frac{0.75}{\sqrt{\frac{2}{8}}} = 1.5 < t \text{ Table}$$

No, because the calculated t is smaller than the table t.

20-

$$SS_y = 50 \quad SS_x = 10 \quad b_{yx} = 2$$

$$b_{yx} = r \frac{SS_y}{SS_x}$$

$$2 = r \frac{50}{10} \quad 2 = 5r \quad r = 0.4$$

21-

$$r = \frac{sp_{xy}}{\sqrt{SS_x SS_y}} = \frac{42}{\sqrt{36 \times 64}} = \frac{42}{6 \times 8} = 0.875$$

$$b_{xy} = r \frac{SS_x}{SS_y} = 0.875 \frac{6}{8} = 0.75$$

References

- Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.
- Agresti, A., & Finlay, B. (2009). *Statistical Methods for the Social Sciences*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall. (This book includes a good discussion of measures of association for two-way frequency tables.)
- Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(1): 262-270.
- Anderson, A.A. (2019). Assessing statistical results: Magnitude, precision, and model uncertainty. *The American Statistician*, 73(1): 118-121.
- Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *The Journal of Wildlife Management*, pp. 912-923.
- Archontoulis, S.V., & Miguez, F.E. (2015). Nonlinear regression models and applications in agricultural research. *Agronomy Journal*, 107:786–798.
- Archontoulis, S.V., Huber, I., Miguez, F.E., Thorburn, P.J., Rogovska, N., & Laird, D.A. (2016). A model for mechanistic and system assessments of biochar effects on soils and crops and trade-offs. *GCB Bioenergy*.

- Archontoulis, S.V., Yin, X., Vos, J., Danalatos, N.G., & Struik, P.C. (2012). Leaf photosynthesis and respiration of three bioenergy crops in relation to temperature and leaf nitrogen: How conservative are biochemical model parameters among crop species? *Journal of Experimental Botany*, 63:895–911.
- Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S.S., Sohl-Dickstein, J., & Ganguli, S. (2020). Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*.
- Basiri, A. (2017). *Statistical designs in agricultural sciences*. Shiraz University Press. (In Persian).
- Begley, C.G., & Ellis, L.M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391): 531-533.
- Belsley, D.A., Kuh, E., & Welsch, R.E. (2004). *Regression diagnostics. Identifying influential data and sources of collinearity*. Wiley, New York.
- Beres, B.L., Turkington, T.K., Kutcher, H.R., Irvine, B., Johnson, E.N., O'Donovan, J.T., Harker, K.N., Holzapfel, C.B., Mohr, R., Peng, G., & Spaner, D.M. (2016). Winter wheat cropping system response to seed treatments, seed size, and sowing density. *Agronomy Journal*, 108(3): 1101-1111.
- Blei, D.M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, 114(33): 8689-8692.
- Bolker, B.M. (2008). *Ecological models and data in R*. Princeton University. Press, Princeton, NJ.
- Bosker, T., Mudge, J.F., & Munkittrick, K.R. (2013). Statistical reporting deficiencies in environmental

toxicology. *Environmental toxicology and chemistry*, 32(8):1737-1739.

- Brien, C.J., Berger, B., Rabie, H., & Tester, M. (2013). Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems. *Plant Methods*, 9: 5-10.
- Brien, C.J., Harch, B.D., Correll, R.L., & Bailey, R.A. (2011). Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *Journal of Agricultural, Biological and Environmental Statistics*, 16: 422–450.
- Burgueño, J. (2018). Spatial analysis of field experiments. In: B. Glaz and K.M. Yeater, editors, *Applied statistics in agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Burgueño, J., Crossa, J., Rodríguez, F., & Yeater, K.M. (2018). Augmented design—experimental design with treatments replicated once. In: B. Glaz and K.M. Yeater, *Applied statistics in agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Casler, M.D. (2013). Finding hidden treasure: A 28-year case study for optimizing experimental designs. *Commun. Biometry Crop Science*, 8: 23–28.
- Casler, M.D. (2015). Fundamentals of experimental design: Guidelines for designing successful experiments. *Agronomy Journal*, 107: 692–705.

- Casler, M.D. (2018). Power and replication—designing powerful experiments. In: B. Glaz and K.M. Yeater, editors, *Applied statistics in agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Casler, M.D., Fales S.L., Undersander, D.J., & McElroy, A.R. (2001). Genetic progress from 40 Years of orchardgrass breeding in North America measured under management intensive rotational grazing. *Canadian Journal of Plant Science*, 81: 713–721.
- Casler, M.D., Fales, S.L., McElroy, A.R., Hall, M.H., Hoffman, L.D., & Leath, K.T. (2000). Genetic progress from 40 years of orchardgrass breeding in North America measured under hay management. *Crop Science*, 40: 1019–1025.
- Conaghan, P., Casler, M.D., O’Keily, P., & Dowley, L.J. (2008). Efficiency of indirect selection for dry matter yield based on fresh matter yield in perennial ryegrass sward plots. *Crop Science*, 48: 127–133.
- Cox, D.R. (2014). Statistics, an overview. *Wiley StatsRef: Statistics Reference Online*.
- Darlington, R.B., & Hayes, A.F. (2017). *Regression analysis and linear models*. New York, NY: Guilford.
- Devore, L. (2008). *Probability and Statistics for Engineering and the Sciences*, 7th ed. Belmont, CA: Brooks/Cole Cengage Learning. (The treatment of probability in this source is more comprehensive and at a somewhat higher mathematical level than ours is in this textbook.).

- Dey, A. (2010). *Incomplete block designs*. World Scientific Publishing Co Pte Ltd., Hackensack, NJ
- Dörre, A., & Emura, T. (2019). *Analysis of Doubly Truncated Data: An Introduction*. Springer Singapore.
- Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology*, 20(5): 1539-1544.
- Fitzmaurice, G.M., Lipsitz, S.R., & Parzen, M. (2007). Approximate median regression via the Box–Cox transformation. *American Statistical*, 61: 233–238.
- Foereid, B., Lehmann, J., & Major, J. (2011). Modeling black carbon degradation and movement in soil. *Plant and Soil*, 345: 223.
- Fontana, R., & Sampo, S. (2013). Minimum-size mixed-level orthogonal fractional factorial designs generation: A SAS-based algorithm. *Journal of Statistical Software*, 53(10): 1-18.
- Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*, 4th ed. New York: W. W. Norton, (An excellent, informal introduction to concepts, with some insightful cautionary examples concerning misuses of statistical methods.)
- Freeman, J., & Modarres, R. (2006). Inverse Box–Cox: The power-normal distribution. *Statistics & Probability Letters*, 76: 764–772.

- Freund, R.J., Wilson, W.J., & Mohr, D.L. (2010). *Statistical methods*, Third ed. Academic Press/ Elsevier, Amsterdam, the Netherlands.
- Galeano, P., & Peña, D. (2019). Data science, big data and statistics. *TEST*, 28(2): 289-329.
- Garland-Campbell, K. (2018). Errors in statistical decision making. In: B. Glaz and K.M. Yeater, editors, *Applied statistics in agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Gbur, E.E., Stroup, W.W., McCarter, K.S., Durham, S., Young, L.J., Christman, M., West, M., & Kramer, M. (2012). *Generalized linear mixed models* (No. analysisofgener, pp. 109-197). American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.
- Gezan, S.A., & Carvalho, M. (2018). Analysis of repeated measures for the biological and agricultural sciences. In: B. Glaz and K.M. Yeater, editors, *Applied statistics in the agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Gillis, J.D., & Price, G.W. (2011). Comparison of a novel model to three conventional models describing carbon mineralization from soil amended with organic residues. *Geoderma*, 160: 304–310.
- Glaz, B., & Yeater, K.M. (2020). *Applied statistics in agricultural, biological, and environmental sciences* (Vol. 172). John Wiley & Sons.
- Gupta, B.C., & Guttman, I. (2014). *Statistics and probability with applications for engineers and scientists*. John Wiley & Sons.

- Gupta, S.C., & Kapoor, V.K. (2020). *Fundamentals of mathematical statistics*. Sultan Chand & Sons.
- Heard, D.P. (2016). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Tailen Hsing. *Exposure-Response Modeling: Methods and Practical Implementation*, p.463.
- Heiberger, R.M., Heiberger, R.M., & Burt Holland, B.H. (2015). *Statistical Analysis and Data Display An Intermediate Course with Examples in R*. Springer.
- Hernández, R., & Kubota, C. (2016). Physiological responses of cucumber seedlings under different blue and red photon flux ratios using LEDs. *Environmental and Experimental Botany*, 121: 66-74.
- Heydarzadeh, S., Jalilian, J., Pirzad, Jamei, R., & Petrusa, E. (2021). Fodder value and physiological aspects of rainfed smooth vetch affected by biofertilizers and supplementary irrigation in an agri-silviculture system. *Agroforestry Systems*. 95(7): 1-13.
- Hicks, S.C., & Irizarry, R.A. (2018). A guide to teaching data science. *The American Statistician*, 72(4): 382-391.
- Hodges, C.B., Lindsey, H.M., Johnson, P., & Stone, B.M. (2020). Researcher degrees of freedom and a lack of transparency contribute to unreliable results of nonparametric statistical analyses across SPSS, SAS, Stata, and R.
- Hoffmann-Jørgensen, J. (2017). *Probability with a view toward statistics*. Routledge.

- Homer, M.S. (2018). An introduction to secondary data analysis with IBM SPSS statistics. *Educational Review*, 70(2): 251-252.
- Hsing, T., & Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators* (Vol. 997). John Wiley & Sons.
- Jansson, A. (2021). Statistics, Classification, and the Standardisation of Melancholia. In *From Melancholia to Depression* (pp. 123-171). Palgrave Macmillan, Cham.
- Jones, M., Woodward, R., & Stoller, J. (2015). Increasing precision in agronomic field trials using Latin square designs. *Agronomy Journal*, 107: 20–24.
- Kaufman, L., & Rousseeuw, P.J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.
- Kenward, M.G., & Roger, J.H. (2009). An improved approximation to the precision of fixed effects from Restricted Maximum Likelihood. *Computational Statistics & Data Analysis*, 53: 2583–2595.
- Kirk, R.E. (2013). Experimental design. *Handbook of Psychology, Second Edition*.
- Kleinman, K., & Horton, N.J. (2014). *SAS and R: Data management, statistical analysis, and graphics*. CRC Press.
- Koenker, R. (2013). Quantile regression. *encyclopedia of environmetrics*. John Wiley & Sons, New York.
- Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. 5th ed. McGraw-Hill, Irwin, NY

- Kvanli, A.H., Pavur, R.J., & Guynes, C.S. (1999). Introduction to business statistics: a computer integrated, data analysis approach. Dryden Press.
- Lawal, B. (2014). Applied statistical methods in agriculture, health and life sciences. Springer.
- Loehlin, J.C. (2004). Latent variable models: An introduction to factor, path, and structural analysis. 4th ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Lohr, S. (2010). Sampling Design and Analysis, 2nd edition. Belmont, CA: Duxbury Cengage Learning, 2010. (A nice discussion of sampling and sources of bias at an accessible level.)
- MacFarland, T.W. (2013). Introduction to data analysis and graphical presentation in biostatistics with R: statistics in the large. Springer Science & Business Media.
- MacInnes, J. (2016). An introduction to secondary data analysis with IBM SPSS statistics. Sage.
- Manjarin, R., Maj, M.A., La Frano, M.R., & Glanz, H. (2020). % polynova_2way: A SAS macro for implementation of mixed models for metabolomics data. Plos one, 15(12): 0244013.
- Matlock Cole, K., & Paek, I. (2017). PROC IRT: A SAS procedure for item response theory. Applied Psychological Measurement, 41(4): 311-320.
- McIntosh, M. (2018). Analysis of variance. In: B. Glaz and K.M. Yeater, Applied statistics in the agricultural, biological, and environmental sciences. ASA, CSSA, SSSA, Madison, WI.

- McShane, B.B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519): 885-895.
- McShane, B.B., Gal, D., Gelman, A., Robert, C., & Tackett, J.L. (2019). Abandon statistical significance. *The American Statistician*, 73(1): 235-245.
- Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Non-linear regression models and applications. In: B. Glaz and K.M. Yeater, editors, *Applied Statistics in Agricultural, Biological, and Environmental Sciences*. ASA, CSSA, SSSA, Madison, WI.
- Miguez, F., Archontoulis, S., & Dokoohaki, H. (2018). Nonlinear regression models and applications. In: B. Glaz and K.M. Yeater, *Applied statistics in the agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Milliken, G.A., & Johnson, D.E. (2009). *Analysis of messy data*. Vol. 1. *Designed experiments*. 2nd ed. CRC Press, Boca Raton, FL.
- Moehring, J., Williams, E.R., & Piepho, H.P. (2014). Efficiency of augmented p-rep designs in multienvironmental trials. *Theoretical and Applied Genetics*, 127: 1049–1060.
- Montgomery, D.C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.
- Moody, H.R. (2006). *Aging: Concepts and controversies*. Pine Forge Press.
- Moore, D., & William, N. (2009). *Statistics: Concepts and Controversies*, 7th ed. New York: W. H. Freeman, 2009. (Contains an excellent chapter on the advantages and pitfalls of

experimentation and another chapter in a similar vein on sample surveys and polls.).

- Morota, G., Ventura, R.V., Silva, F.F., Koyama, M., & Fernando, S.C. (2018). Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of Animal Science*, 25-36.
- Mudge, J.F. (2013). Explicit consideration of critical effect sizes and costs of errors can improve decision-making in plant science. *New Phytologist*, 199: 876–878.
- Mudge, J.F., Baker, L.F., Edge, C.B., & Houlahan, J.E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One*, 7: 32734.
- Mudge, J.F., Barrett, T.J., Munkittrick, K.R., & Houlahan, J.E. (2012). Negative consequences of using $\alpha = 0.05$ for environmental monitoring decisions: A case study from a decade of Canada's Environmental Effects Monitoring Program. *Environmental Science & Technology*, 46(17): 9249–9255.
- Nakagawa, S., & Cuthill, I.C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological reviews of the Cambridge Philosophical Society*, 82: 591–605.
- Naveau, P., Hannart, A., & Ribes, A. (2020). Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, 7: 89-110.
- Neter, J., William, W., & Michael, K. (2005). *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill, (The first half of this

book gives a comprehensive treatment of regression analysis without overindulging in mathematical development; a highly recommended reference.).

- Papaspiliopoulos, O. (2020). *High-Dimensional Probability: An Introduction with Applications in Data Science*.
- Parolini, G. (2015). The emergence of modern statistics in agricultural science: analysis of variance, experimental design and the reshaping of research at Rothamsted Experimental Station, 1919–1933. *Journal of the History of Biology*, 48(2): 301-335.
- Peck, R., Olsen, C., & Devore, J.L. (2015). *Introduction to statistics and data analysis*. Cengage Learning.
- Piepho, H.P., Möhring, J., & Williams, E.R. (2013). Why randomize agricultural experiments? *Journal of Agronomy and Crop Science*, 199: 374–383.
- Rafique, R. (2011). *Measurements and modeling of nitrous oxide emissions from Irish grasslands*. Ph.D. diss. National University of Ireland, Cork.
- Raudonius, S. (2017). *Application of statistics in plant and crop research: important issues*. *Zemdirbyste-Agriculture*, 104(4): 25-36.
- Richter, C., & Piepho, H.P. (2018). *Linear regression techniques*. In: B. Glaz and K.M. Yeater, *Applied statistics in the agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Richter, C., Kroschewski, B., Piepho, H. P., & Spilke, J. (2015). *Treatment comparisons in agricultural field trials accounting for*

- spatial correlation. *The Journal of Agricultural Science*, 153: 1187–1207.
- Ritz, C., & Streibig, J.C. (2008). *Nonlinear regression with R*. Springer, New York.
- Ritz, C., Pipper, C.B., & Streibig, J.C. (2013). Analysis of germination data from agricultural experiments. *European Journal of Agronomy*, 45:1–6.
- Sahu, P.K. (2016). *Applied statistics for agriculture, veterinary, fishery, dairy and allied fields* (pp. 133-194). India:: Springer.
- Sahu, P.K., & Das, A.K. (2017). *Agriculture and Applied Statistics-II*. Kalyani Publishers.
- Saraiva, T., & Slaton, A.E. (2018). Statistics as Service to Democracy: Experimental Design and the Dutiful American Scientist. In *Technology and Globalisation* (pp. 217-255). Palgrave Macmillan, Cham.
- Scheaffer, R.L., William, M., & Lyman, O. (2006). *Elementary Survey Sampling*, 6th ed. Belmont, CA: Duxbury Cengage Learning, (An accessible yet thorough treatment of the subject.).
- Schützenmeister, A., Jensen, U., & Piepho, H.P. (2012). Checking normality and homoscedasticity in the general linear model using diagnostic plots. *Communications in Statistics - Simulation and Computation*, 41: 141–154.
- Selya, A.S., Rose, J.S., Dierker, L.C., Hedeker, D., & Mermelstein, R.J. (2012). A practical guide to calculating Cohen's f^2 , a measure of local effect size, from Proc Mixed. *Frontiers in Psychology*, 3: 111- 123.

- Sharma, R., Singh, J., & Verma, N. (2020). Statistical optimization and comparative study of lipopeptides produced by *Bacillus amyloliquefaciens* SAS-1 and *Bacillus subtilis* BR-15. *Biocatalysis and Agricultural Biotechnology*, 25: 101575.
- Shrestha, J. (2019). P-Value: A true test of significance in agricultural research.
- Simmonds, J., Gómez, J.A., & Ledezma, A. (2017). November. knowledge inference from a small water quality dataset with multivariate statistics and data-mining. In international conference of ict for adapting agriculture to climate change (pp. 1-15). Springer, Cham.
- Smiley, R.W., Machado, S., Rhinhardt, K.E.L., Reardon, C.L., & Wuest, S.B. (2016). Rapid quantification of soilborne pathogen communities in wheat-based long-term field experiments. *Plant Disease*, 10 pp.
- Smith, A.B., Lim, P., & Cullis, B.R. (2006). The design and analysis of multi-phase plant breeding experiments. *The Journal of Agricultural Science*, 144: 393–409.
- Stokes, M., Chen, F., & Gunes, F. (2014). March. An introduction to Bayesian analysis with SAS/STAT® software. In Proceedings of the SAS Global Forum 2014 Conference, SAS Institute Inc, Cary, USA (available at <https://support.sas.com/resources/papers/proceedings14/SAS400-2014.pdf>).
- Stroup, W. (2018). Analysis of non-Gaussian data. In: B. Glaz and K.M. Yeater, *Applied statistics in the agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.

- Stroup, W.W. (2015). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, 107: 811–827.
- Stroup, W.W. (2018). Analysis of non-Gaussian data. In: B. Glaz and K.M. Yeater, editors, *Applied Statistics in Agricultural, Biological, and Environmental Sciences*. ASA, CSSA, SSSA, Madison, WI.
- Sullivan, G.M., & Feinn, R. (2012). Using effect size-Or why the P value is not enough. *Journal of Graduate Medical Education*, 4: 279–282.
- Sun, F., Roderick, M.L., & Farquhar, G.D. (2018). Rainfall statistics, stationarity, and climate change. *Proceedings of the National Academy of Sciences*, 115(10): 2305-2310.
- Taylor, J.M.G. (2006). Transformations-II. In: S. Kotz et al., editors, *Encyclopedia of statistical science*. Vol. 14. John Wiley and Sons, New York.
- Tempelman, R.J. (2008). Statistical analysis of efficient unbalanced factorial designs for two-color microarray experiments. *International Journal of Plant Genomics*, 584360.
- Urkude, P., & Lade, S. (2020). Automation of Reliability Warranty Report Using SAS Software for Data Analysis. In *Machine Learning and Information Processing* (pp. 413-423). Springer, Singapore.
- Utts, J. (2005). *Seeing Through Statistics*, 3rd ed. Belmont, CA: Duxbury Press, 2005. (A nice introduction to the fundamental ideas of statistical reasoning.).

- Van Es, H.M., Gomes, C.P., Sellman, M., & Van Es, C.L. (2007). Spatially-balanced complete block designs for field experiments. *Geoderma*, 140: 346–352.
- Vargas, M., Glaz, B., Crossa, J., & Morgounov, A. (2018). Analysis and interpretation of interactions of fixed and random effects. In: B. Glaz and K.M. Yeater, *Applied statistics in the agricultural, biological, and environmental sciences*. ASA, CSSA, SSSA, Madison, WI.
- Wallach, D. (2006). Evaluating crop models. In: D. Wallach, D. Makowski, and J.W. Jones, editors, *Working with dynamic crop models—Evaluations, analysis, parameterization, and applications*. Elsevier, New York. p. 11–53.
- Welham, S.J., Gezan, S.A., Clark, S.J., & Mead, A. (2014). Replication and power In: *Statistical methods in biology: Design and analysis of experiments and regression*. CRC Press, Boca Raton, FL. p. 241–256.
- West, B.T., Welch, K.B., & Galecki, A.T. (2014). *Linear mixed models: a practical guide using statistical software*. Crc Press.
- Williams, E.R., Piepho, H.P., & Whitaker, D. (2011). Augmented p-rep designs. *Biometrical Journal*, 53: 19–27.
- Wright, K. (2012). *AgriDat: Agricultural Datasets*. R package version 1.4. <https://cran.r-project.org/web/packages/agriDat/index.html> (accessed 30 June 2016).
- Yazdi Samadi, B. (2018). *Plans of agricultural experiments 1*. Faculty of Agriculture. University of Tehran. (In Persian).

Ye, Z., & Zhao, Z. (2010). A modified rectangular hyperbola to describe the light-response curve of photosynthesis of *Bidens pilosa* L. grown under low and high light conditions. *Frontiers of Agriculture in China*, 4: 50–55.

APPENDIX TABLES

Table 1- Ft values at 5 % probability level for use in analysis of variance and F tests

The degree of freedom of the denominator of the fraction F (experimental error)	Degree of freedom of fraction F							
	F							
	1	2	3	4	5	6	7	8
1	161	200	216	225	230	234	237	239
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	5.99	5.14	5.76	4.53	4.39	4.28	4.21	4.15
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94

Continued table 1

The degree of freedom of the denominator of the fraction F (experimental error)	Degree of freedom of fraction F							
	9	10	11	12	14	16	20	25
1	241	242	243	244	245	246	248	249
2	19.38	19.40	19.40	19.41	19.41	19.43	19.45	19.45
3	8.81	8.79	8.76	8.74	8.71	8.69	8.66	8.64
4	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77
5	4.77	4.74	4.70	4.68	4.64	4.60	4.56	4.54
6	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84
7	3.68	3.64	3.60	3.57	3.52	3.49	3.44	3.40
8	3.39	3.35	3.31	3.28	3.23	3.20	3.15	3.11
9	3.18	3.14	3.10	3.07	3.02	2.98	2.94	2.90
10	3.02	2.98	2.94	2.91	2.86	2.82	2.77	2.74
11	2.90	2.85	2.82	2.79	2.74	2.70	2.65	2.61
12	2.80	2.75	2.72	2.69	2.64	2.60	2.54	2.51
13	2.71	2.67	2.63	2.60	2.55	2.51	2.46	2.42
14	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35
15	2.59	2.54	2.51	2.48	2.43	2.39	2.33	2.29
16	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.23
17	2.49	2.45	2.41	2.38	2.33	2.29	2.23	2.19
18	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15
19	2.42	2.38	2.34	2.31	2.26	2.21	2.16	2.10
20	2.39	2.35	2.31	2.28	2.23	2.18	2.12	2.08
22	2.34	2.30	2.26	2.23	2.18	2.13	2.07	2.02
24	2.30	2.25	2.22	2.18	2.13	2.09	2.03	1.98
26	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.94
28	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.90
30	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.88
40	2.12	2.08	2.04	2.00	1.95	1.90	1.84	1.78
60	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.69
120	1.96	1.91	1.86	1.83	1.77	1.72	1.66	1.60
∞	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.51

Table 2- Ft values at 1% probability level for use in analysis of variance and F tests

The degree of freedom of the denominator of the fraction F (experimental error)	Degree of freedom of fraction F							
	1	2	3	4	5	6	7	8
1	4052	5000	5403	5625	5764	5859	5928	5982
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99
60	7.08	4.98	4.13	3.65	3.34	3.12	2.94	2.82
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51

Continued table 2

The degree of freedom of the denominator of the fraction F (experimental error)	Degree of freedom of fraction F							
	9	10	11	12	14	16	20	25
1	6022	6056	6082	6106	6142	6169	62.09	6240
2	99.39	99.40	99.41	99.42	99.43	99.44	99.45	99.46
3	27.35	27.23	27.13	27.05	26.92	26.83	26.69	26.58
4	14.66	14.55	14.45	14.37	14.24	14.15	14.02	13.90
5	10.16	10.05	9.96	9.89	9.77	9.68	9.55	9.45
6	7.98	7.87	7.79	7.72	7.60	7.52	7.40	7.30
7	6.72	6.62	6.54	6.46	6.35	6.27	6.16	6.05
8	5.91	5.81	5.74	5.67	5.56	5.48	5.36	5.26
9	5.35	5.26	5.18	5.11	5.00	4.92	4.81	4.71
10	4.94	4.85	4.78	4.71	4.60	4.52	4.41	4.31
11	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.01
12	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.77
13	4.19	4.10	4.02	3.96	3.85	3.78	3.66	3.58
14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.42
15	3.89	3.80	3.73	3.67	3.56	3.48	3.37	3.28
16	3.78	3.69	3.61	3.55	3.45	3.37	3.26	3.16
17	3.68	3.59	3.52	3.46	3.35	3.27	3.16	3.07
18	3.60	3.51	3.44	3.37	3.27	3.19	3.08	2.99
19	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.90
20	3.46	3.37	3.30	3.23	3.13	3.05	2.94	2.85
22	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.73
24	3.26	3.17	3.09	3.03	2.93	2.85	2.74	2.64
26	3.18	3.09	3.02	2.96	2.86	2.77	2.66	2.56
28	3.12	3.03	2.95	2.90	2.80	2.71	2.60	2.50
30	3.07	2.98	2.90	2.84	2.74	2.66	2.55	2.85
40	2.99	2.80	2.73	2.66	2.56	2.49	2.37	2.27
60	2.72	2.63	2.56	2.50	2.40	2.32	2.20	2.10
120	2.56	2.47	2.40	2.34	2.23	2.15	2.03	1.93
∞	2.41	2.32	2.24	2.18	2.07	1.99	1.88	1.77

Table 3: t values at different levels of probability for use in LSD test

degree of freedom of experimental error	Probability level			
	5%	1%	0.5%	0.1%
1	12.706	65.657	127.320	636.620
2	4.303	9.925	14.089	31.598
3	3.182	5.841	7.453	12.924
4	2.776	4.604	5.598	8.610
5	2.571	4.032	4.773	6.869
6	2.447	3.707	4.317	5.959
7	2.365	3.499	4.029	5.408
8	2.306	3.355	3.833	5.041
9	2.262	3.250	3.690	4.781
10	2.228	3.169	3.581	4.587
11	2.201	3.106	3.497	4.437
12	2.179	3.055	3.427	4.318
13	2.160	3.012	3.372	4.221
14	2.145	2.977	3.326	4.140
15	2.131	2.947	3.286	4.073
16	2.120	2.921	3.252	4.015
17	2.110	2.898	3.223	3.965
18	2.101	2.878	3.197	3.922
19	2.093	2.861	3.174	3.883
20	2.086	2.845	3.153	3.850
24	2.064	2.797	3.090	3.745
30	2.042	2.750	3.030	3.646
40	2.021	2.704	2.971	3.551
60	2.000	2.660	2.915	3.460
120	1.980	2.617	2.860	3.373
∞	1.960	2.576	2.807	3.291

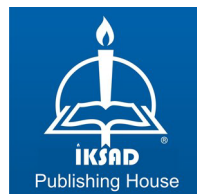
Table 4 chi score (χ^2)
Probability of χ^2 values greater than χ^2 in the table (probability level)

df	.995	0.990	.975	.950	.900	.750	.500	.25	.10	.05	.02	.01	.005
1	0.0 ³ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.397	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.1156	0.2106	0.352	0.584	1.21	2.371	4.11	6.25	7.81	9.35	11.34	12.8
4	0.207	0.2974	0.484	0.711	1.064	1.92	3.369	5.38	7.78	9.49	11.14	13.28	14.9
5	0.412	0.5541	0.831	1.15	1.61	2.67	4.353	6.63	9.24	11.08	12.83	15.09	16.7
6	0.676	0.8721	1.24	1.64	2.20	3.45	5.353	7.88	10.64	12.59	14.45	16.75	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.354	9.0	12.01	14.07	16.01	18.47	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.342	10.2	13.12	15.51	17.53	20.09	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.341	11.4	14.19	16.92	19.02	21.67	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.342	12.5	15.19	18.31	20.48	23.02	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.31	13.7	16.22	19.68	21.92	24.15	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.03	14.3	17.28	20.53	22.99	25.22	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.31	15.4	18.48	21.91	24.29	26.21	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.31	16.6	19.68	23.21	25.79	27.19	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.31	17.8	20.91	24.6	27.33	28.31	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.31	19.0	22.15	26.01	28.59	30.19	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.31	20.2	23.39	27.21	30.19	32.01	35.7
18	6.25	7.01	8.23	9.39	10.9	13.7	17.31	21.5	24.6	28.31	31.41	33.41	37.2
19	6.84	7.61	8.91	10.1	11.7	14.6	18.31	22.8	25.91	30.19	32.59	34.81	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.31	24.1	27.2	31.59	33.91	36.19	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.31	25.4	28.59	33.19	35.41	37.59	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.31	26.8	30.01	34.6	36.91	38.91	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.31	28.1	31.41	35.91	38.41	40.41	44.0
24	9.89	10.9	12.4	13.8	15.7	19.0	23.31	29.5	32.91	37.41	40.01	42.01	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.31	30.8	34.41	38.91	41.61	43.61	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.31	32.2	35.91	40.01	43.19	45.19	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.31	33.6	37.41	41.61	44.61	46.61	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.31	35.0	38.91	43.19	46.19	48.19	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.31	36.4	40.01	44.61	47.61	49.61	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.31	37.8	41.61	46.19	49.19	51.19	53.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.31	48.3	51.01	55.41	63.19	70.41	66.8
50	28.0	29.7	32.4	34.8	37.7	42.9	49.31	56.3	63.19	67.19	76.19	83.19	79.5
60	35.5	37.5	40.5	43.2	46.5	52.3	59.31	67.0	74.41	79.19	88.19	92.0	

Statistical Methods and Probabilities in Agricultural Science



Written by:
Dr. Mohsen MIRZAPOUR
Dr. Saeid HEYDARZADEH
Dr. Harun GİTARİ



ISBN: 978-625-8246-62-9