Wollo University

COLLEGE OF BUSINESS AND ECONOMICS

DEPARTMENT OF ACCOUNTING

DISTANCE MODULE OF STATISTICS FOR FINANCE (AcFn1043)

Prepared By:

1. **GETAHUN WORKU (MSc.), DEPARTMENT OF STATISTICS, WOLLO UNIVERSITY**
2. **TILAYE MATEBE (MSc.), DEPARTMENT OF STATISTICS, WOLLO UNIVERSITY**

Editor:

1. **AMSALU AREGA (MSc.), DEPARTMENT OF STATISTICS, WOLLO UNIVERSITY**

APRIL, 2018

DESSIE, ETHIOPIA

Module Introduction

Hello dear student! Welcome to introduction to statistics for finance (AcFn1043) course module. This module is a four credit hour (6 ECTS) course deliver to accounting and finance students.

Dear student, this module takes you from the basic concepts of statistics to the application of their tools in financial data analysis to make evidence based decision in the area finance and business world. It presents the subject matter in a precise and simple manner, fully demonstrated examples, charts and diagrams. There are a number of solved problems and self test exercises.

We have divided this module into eight units. Further, the units subdivided in to sections. We want to advise you to deal on each activities and exercises available in each of the units and then you check answers for exercises at the end of the module.

Dear learner, this module will help you to simulate and provide knowledge of statistics in financial world. In addition to this, the module is prepared with the assumption that you will be able to study by yourself.

Course Objectives & Competences to be Acquired

The objective of this course is thus to discuss the theoretical aspects of statistics and then focus on its practical applications in business decision making, which modern managers and decision makers are expected to be armed with on the face of considerable uncertainty. Besides, it is also to create know-how to students on various application areas and benefit of statistical in business.

This course aims:

- ☞ To familiarize students about the use & application of various statistical tools in the field of financial decision making
- ☞ To enable students make valid inference from data
- ☞ To enable students to construct and test different types of hypothesis
- ☞ To enable students to find correlation between variables
- ☞ How to apply the statistical tests in the preparation of Research report.
- ☞ To enable students appreciate the application of statistics in every areas of activities in business and industry such as production, financial analysis, distribution, market research, manpower planning.
- ☞ Demonstrate the importance and usefulness of statistics in real life and on real data;
- ☞ Understand the methods of data collection, organization, presentation, analysis and interpretation;
- ☞ Apply statistical methods in Business researches, decision making and future career to solve standard problems from financial data;
- ☞ Apply statistics in every areas of activities in business and industry such as production, financial analysis, distribution, market research, manpower planning.

Course description

This course is designed to ensure that you are good with numbers; that you can interpret and analyze financial and economic data and develop a critical awareness of some of the pitfalls in collecting, presenting and using data. You must know the statistical methods, which rely on probability theory, to summarize the data, e.g. in estimates. You must be able to interpret the statistics or estimates in terms of your original purpose and the theory. Thus during this course we will be moving backwards and forwards between these elements: purpose, data, theory and statistical methods.

The use of statistical knowledge in the field of business aid dated many years back. In recent years, an understanding of statistical methods, techniques, and the skills to make use of them had widely been recognized more than before. It is essential for anyone making business decisions on the basis of data to possess a clear understanding of statistics.

Among other, the vast and fast changing technological, financial and economic setting has necessitated an organized use and extensive application of statistical tools to business decision making. Statistics has proved useful in many ways. Such as in establishing relationship, making predications, and providing solution to the many problems of business operations and managerial decision Statistics is widely applied in production and quality control, marketing research, manpower planning, finance, etc.

Roles of the students

The success of this course depends on the students' individual and collective contribution to the class discussions. Students are expected to participate voluntarily, or will be called upon, to contribute to set exercises and problems. Students are also expected to read the assigned readings and prepare the cases before each class so that they could contribute effectively to class discussions. Students must attempt assignments by their own. Proficiency in this course comes from individual knowledge and understanding. Copying the works of others is considered as serious offence and leads to disciplinary actions.

Table of Contents

Module Introduction ii
Course Objectives & Competences to be Acquired iii
Course description iii
Roles of the students iv
CHAPTER ONE 1
1. INTRODUCTION 1
1.1. Definition and classification of statistics 1
1.1. Classification of statistics 3
1.2. Basic statistical terms 4
1.3. Stages in statistical investigation 5
1.4. Application, Uses and limitation of statistics 6
1.5. Statistical data (meaning, types, sources and methods of obtaining data) 8
1.5.1. Qualitative and Quantitative Data 9
1.5.2. Cross-Sectional and Time Series Data 10
1.5.3. Organizing and summarizing data 12
1.6. Measures of Central Tendency 28
1.6.1. The Summation Notation 28
1.5.2 Properties of Measures of Central Tendency 30
1.7. Types of Measures of Dispersion 39
1.7.1. The Range (R) and Relative Range (RR) 39
1.7.2. The Variance, the Standard deviation and the coefficient of Variation 40
1.8. Measures of shape 44
1.8.1. Skewness 45
1.8.2. Kurtosis 46
Chapter Two 49
2. Probability and Probability Distribution 49
2.1. Basic definitions of probability 49
2.2. Fundamental concepts of probability 49
2.3. Common Discrete Probability Distributions 61
2.4. Common Continuous Probability Distributions 64
Chapter Three 69
Sampling & Sampling Distributions 69
3.1. Definition and Some Basic Terms 69

3.1.1. Types of Errors 69
3.1.2. Non sampling errors 70
3.2. Random (probability) sampling methods 70
3.3. Sampling Distribution of the Mean and Proportion 73
3.3.1 Sampling Distribution of the Mean 73
3.3.2. Sampling Distribution of The Proportion 74
CHAPTER FOUR 76
4. ESTIMATION 76
4.1. Point and Interval Estimators 76
4.1.1. Methods of Estimation 77
4.1.2. Statistics as Estimators for Parameters 78
4.1.3. Point estimator of the proportion 78
4.1.4. Point estimator of the mean 79
4.2. Interval estimators of the mean and proportion 81
4.2.1. Confidence Interval about One Proportion 81
4.2.2. Confidence Interval about One Mean 82
4.2.3. Confidence Interval about One Variance 85
4.3. Confidence Interval about two Parameters 87
4.3.1. Confidence Interval about the difference between two proportions 87
4.3.2. Confidence Interval about the difference between Two Means 88
4.4. The t distribution 91
4.5. Samples-Size Determination 93
Summary 95
CHAPTER FIVE 97
5. HYPOTHESIS TESTING 97
5.1. Concepts of Hypothesis Testing 97
5.1.1. THE NULL AND ALTERNATIVE HYPOTHESES 97
5.1.2. Possible Decisions 99
5.1.3. STEPS IN HYPOTHESIS TESTING: 100
5.1.4. TYPES OF ERRORS 101
5.1.5. The Test Statistic 103
5.1.6. The Critical Region or The Rejection Region (CR or RR) 103
5.1.7. Conclusion and interpretation 103
5.2. Hypothesis Testing about One Parameter 103
5.2.1. Hypothesis Testing about One Proportion, P 104

5.2.2. Hypothesis Testing About The Population Mean, $\boldsymbol{\mu}$: 106
5.2.3. Hypothesis Testing about One variance..... 111
5.3. Hypothesis Testing Concerning Two Parameters 114
5.3.1. HYPOTHESIS TEST FOR THE DIFFERENCE BETWEEN TWO MEANS..... 116
5.3.2. HYPOTHESIS TESTING FOR TWO POLULATION PROPORTION..... 119
5.3.3. TESTS ABOUT TWO VARIANCES 122
Summary 124
CHAPTER SIX 126
CHI-SQUARED DISTRIBUTIONS 126
Introduction 126
CHI-SQUARED DISTRIBUTIONS 126
Chi-Square as a Test of Goodness of Fit 128
Test of Association..... 130
UNIT SEVEN..... 135
ANALYSIS OF VARIANCE..... 135
Introduction 135
7.1. The F Distribution..... 135
7.2. Comparison of several means 136
7.3. Comparing Two Population Variances 139
CHAPTER EIGHT 141
8. REGRESSION and CORRELATION ANALYSIS..... 141
Introduction 141
8.1. Simple linear Regression..... 142
8.2. Correlation Analysis 144
8.2.1. Karl Pearson’s coefficient of (simple) correlation 144
8.2.2. Coefficient of determination..... 147
8.2.3. Rank Correlation..... 147
Exercise 150
References 151
Appendices 152

CHAPTER ONE

1. INTRODUCTION

This unit of the module introduces you the concept of statistics and methods of data collection, presentation and analysis. At the end of the this chapter, the learner be able to

- ✓ Define statistics and basic terminologies
- ✓ Identify the different types of scale of measurement
- ✓ Be familiar with data collection methods and presentation mechanisms
- ✓ Summarize data by using measures of central tendency and dispersion of the dataset
- ✓ Measure the extent to which the distribution of values in a data set deviate from symmetry

Peoples are familiar with statistics in their different day to day communication and activities. Every day, we have confronted with some form of statistical information through news papers, magazines, and other forms of communication. Besides, statistical tables, survey results, and the language of probability are used with increasing frequency by the media. Such statistical information has become highly influential in our lives.

Statistics can be considered as numerical statements of facts which are highly convenient and meaningful forms of communication. The subjects of statistics, as it seems, is not a new discipline but it is as old as the human society itself. The sphere of its utility, however, was very much restricted. The word statistics is derived from the Latin word “statis” which means a “political state” or government. It was originally applied in connection with kings and monarchs collecting data on their citizenry which pertained to state wealth, taxes collected population and so on. Thus, the scope of statistics in the ancient times was primarily limited to the collection of demographic, property and wealth data of a country by governments for framing military and fiscal policies.

Its usefulness has now spread to diverse fields such as agriculture, accounting, marketing, economics, management, medicine, political science, psychology, sociology, engineering, journal, metrology, tourism, etc. And hence, statistics is included in the curriculum of many professional and academic study programs.

1.1. Definition and classification of statistics

Dear learner, what is statistics to you? How can you define it? Why you study statistics?

Most people become familiar with statistics in reading newspapers, books, and heard on radio speeches and watch on TV presentations. For instance one may obtain the following results from reports, TV shows, newspapers, etc. “The average salary of teachers in country X is 630 Birr per month”; “the cost of living rose by 10% in the last three years”; “the car accident rate has gone up 15% since 5 years”; “Among older men, the mortality rate for smokers is twice the rate of those who never smoked”; “The agricultural production increased by 5 percent this year”. All the above statements are statistical conclusions in some form.

However, Statistics is not limit to some area of application rather applied in all disciplines to make a scientific decision based on data. In public health, an administrator would be concerned with the number of residents who contract a new strain of flu virus during a certain year. In pharmacy, it is used to study the efficacy and potency of drugs. To study plant life, a botanist has to relay on statistics to know the effect of temperature, rainfall and so on. In general, statistics can be applied in business, social sconces, natural sciences and engineering.

The word “**statistics**” has different meanings. Statistics could be singular or plural sense. Statistics is a **plural noun** which describes a collection of numerical data such as employment statistics, accident statistics, population statistics, economic statistics, agricultural statistics, statistics of business firms, of imports and exports, index numbers that help us understand a variety of business and economic situations etc. It is in this sense that the word 'statistics' is usually understood by a layman. However, as you will see, the field, or subject, of statistics involves much more than numerical facts. In other way, the word statistics defied as a **singular noun.** In a broader sense, **statistics** is defined as the art and science of collecting, analyzing, presenting, and interpreting data. Particularly in business and economics, the information provided by collecting, analyzing, presenting, and interpreting data gives managers and decision makers a better understanding of the business and economic environment and thus enables them to make more informed and better decisions.

In the **plural** sense: Statistics can be defined as data i.e. the raw data themselves such as statistics of income tax and economic statistics, statistics of Business Firms, employment statistics, accident statistics, statistics of imports and exports, and so on. Solving statistical problems begins with a problem and data. Averages, medians, percents, and index numbers that help us to understand a variety of business and economic situations are example of statistical results or figures. **Statistics** is a science that helps us to make better decisions in any fields.

In general Statistics teaches us how to summarize, analyze, and draw meaningful inferences from data that lead to improve decisions making. These decisions that we made help us to improve the running, for example, a department, a company, the entire economy, etc.

- Why you learn statistics? Because

1. Data are everywhere
2. Statistical techniques are used to make many decisions that affect our lives
3. No matter what your career, you will make professional decisions that involve data. Understanding of statistical methods will help you make these decisions effectively.

1.1. Classification of statistics

Statistics is broadly divided into two categories based on how the collected data are used.

A. Descriptive statistics: - It is concerned with summary calculations, graphs, charts and tables to describe the data that you have in hand. It deals with describing data without attempting to infer anything that goes beyond the given set of data. Descriptive statistics consists of collection, organization, summarization and presentation of data in the convenient and informative way.

Example:

- Presenting the starting salaries of 100 graduates of accounting and economics in different organizations.
- From sample we have 40% employee suggest positive attitude toward the management of the organization.

Inferential statistics: - Descriptive Statistics describe the data set that’s being analyzed, but doesn’t allow us to draw any conclusions or make any inferences about the data, other than visual “It looks like”. But inferential statistics includes the methods used to find out something about a population based on a sample. Inferential statistics utilizes sample data to make decision for entire data set.

In this form of statistical analysis, descriptive statistics is linked with probability theory so that an investigator can generalize the results of a study.

Inferential statistics is important because statistical data usually arises from sample. In inference, statistical techniques based on probability theory are required. For example, the average income from all tax payers of Ethiopia can be estimated from figures obtained from a few hundred (the sample) tax payer in probabilistic way.

Example:

- From past figures, it has been predicted that 90 of registered voters will vote in the November election.
- Assessing the effectiveness of a accountancy (on the basis of data obtained from few selected accountants)

Self test exercises:

a. Explain the difference between descriptive statistics and inferential statistics and give an example for each.

b. Suppose you have a data and your goal is summarizing and explaining a specific set of data, then what type of statistics you use? Explain your answer briefly in your own words

1.2. Basic statistical terms

Sampling: The process of selecting a sample from the population is called sampling.

Population*:* A population is a totality of things, objects, peoples, etc about which information is being collected. It is the totality of observations with which the researcher is concerned.

Sample: A sample is a subset or part of a population selected to draw conclusions about the population.

Census survey: It is the process of examining the entire population. It is the total count of the population.

Parameter: It is a descriptive measure (value) computed from the population. It is the population measurement used to describe the population.

Example: population mean and population standard deviation

Statistic: It is a measure used to describe the sample. It is a value computed from the sample like sample mean, proportion etc.

Sampling frame: A list of people, items or units from which the sample is taken.

Data: Data as a collection of related facts and figures from which conclusions may be drawn. It is the set of different values of a variable. Data can be classified in different ways depending on time, nature of variable and its source. It may be classified as primary or secondary data based on source of data.

Variable: A certain characteristic which changes from object to object and time to time. Variable may be classified as quantitative or qualitative variables.

Quantitative variable: A variable that can be measured numerically. The data collected on quantitative variable are called quantitative data. Examples include Salary, number of customer, height, number of students in a class, number of car accidents, e t c.

Qualitative variable: A variable that cannot assume a numerical value but can be classified into two or more non numerical categories. The data collected on such a variable are called qualitative or categorical data. Examples include sex, tax payer category (A, B, C), marital status, e t c.

Discrete variable: a variable whose values are countable. Examples include number clients in the town, number of students in the class, number of injuries insured in the insurance company, e t c.

Continuous variable: a variable that can assume any numerical value over a certain interval or intervals. Examples include **weight of new born babies, height of seedlings, temperature measurements e t c.**

Dear learner discuss, further on types variable with examples.

1.3. Stages in statistical investigation

Statistical investigation can be done by following the following five stages.

1. **Proper Collection of Data:** This is a stage where we gather information for our purpose.

If data are needed and if not readily available, then they have to be collected. Data may be collected by the investigator directly using methods like interview, questionnaire, experimentation and observation or may be available from published or unpublished sources.

Data gathering is the basis (foundation) of any statistical work and valid conclusions can only result from properly collected data.

2. **Data Organization**: After collecting the data using data collection methods seen in stag one, then just organize the collected data. It is a stage where we edit our data. A large mass of figures that are collected from surveys frequently need organization and edition of ambiguity, and inconsistency. After editing, we may classify (arrange) according to their common characteristics. Classification or arrangement of data in some suitable order makes the information easer for presentation.

3. **Data Presentation:** The organized data should be presented in clear and convenient form using tables, graphs and diagram. At this stage, large data will be presented in tables in a very summarized and condensed manner. The main purpose of data presentation is clarity and simplicity of the data user and to facilitate statistical analysis.
4. **Data Analysis:** This is the stage where we critically study the data to draw conclusions about the population parameter. The purpose of data analysis is to dig out information useful for decision making. Analysis usually involves highly complex and sophisticated mathematical techniques. However, in this material only the most commonly used methods of statistical analysis are included.

Data Interpretation: Interpretation means drawing conclusions from the data which form the basis for decision making. Correct interpretation requires a high degree of skill and experience and it is necessary in order to draw valid conclusions. Therefore, care should be taken in data analysis to make a valid interpretation and conclusion.

1.4. Application, Uses and limitation of statistics

Today the field of statistics is recognized as a highly applied and useful discipline in decision making process by managers of modern business, industry, frequently changing technology etc.

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client’s balance sheet fairly represents the actual amount of accounts receivable. Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample. After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client’s balance sheet is acceptable.

Financial analysts use a variety of statistical information to guide their investment and to give some recommendations investors, government officials and policy makers. In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is over- or underpriced.

Questions which need the application of statistics often rose in our day to day financial and business activities.

- Which of several brokerages has a reliable record of higher-than-average return on investment?
- Is the share price for this firm rising predictably enough for a day-trader to invest in it?
- What has our return on investment been for these two brands of computer equipment over the last four years?
- Should we pay the new premium being proposed by our insurance company for liability insurance?
- What is the premium that our actuaries are calculating for a fire-insurance policy on this type of building?
- Should we invest in bonds or in stock this year? What are the average rates of return? How predictable are these numbers?
- Which country has the greatest chance of maximizing our profit on investment over the next decade? Which industry? Which company?

In general**,** the following are some uses of statistics:

- It presents facts in a definite and precise form.
- Data reduction. Reduces mass of data. The original set of data (raw data) is normally voluminous and disorganized unless it is summarized and expressed in few numerical values.
- Measuring the magnitude of variations in data.
- Furnishes a technique of comparison of different sets of data. Statistical values such as averages, percentages, ratios, etc, are the tools that can be used for comparing sets of data.
- Estimating unknown population characteristics.
- Testing and formulating of hypothesis and new theories.
- Studying the relationship between two or more variables.
- Forecasting future events. Statistics is extremely useful for analyzing the past and present data and for prediction of future trends and events.
- Used for the government to formulate new polices and strategies. Statistical study results in the areas of taxation, on unemployment rate, on the performance of every sort of military

equipment, health and education sectors, etc, may convince a government to review its policies and plans with the view to meet national needs and aspirations.

Even though statistics is widely used in various fields of natural and social sciences, which closely related with human inhabitant, it has its own limitations as far as its application is concerned.

Some of the limitations are as follows:

- Statistics doesn't deal with single (individual) values. Statistics deals only with aggregate values.
- Statistical results are only true on average and in general but not individually true. Statistical conclusions are true only under certain condition or true only on average
- Statistics can't deal with qualitative characteristics
- Statistical interpretations require a high degree of skill and understanding of the subject.
- Statistics can be misused. It may be misused by ignorant persons or experts.

1.5. Statistical data (meaning, types, sources and methods of obtaining data)

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation and thus useful for further investigation. All the data collected in a particular study are referred to as the data set for the study. Data is called statistical data if is in mass or in aggregate and collected for predetermined purpose. Depending on the sources there are two types of data; Primary and secondary data.

Primary Data: Primary data are measurements observed and recorded as part of an original study. When the data required for a particular study can be found neither in the internal records of the enterprise, nor in published sources, it may become necessary to collect original data, i.e., to conduct first hand investigation. It is collected by the investigator for the first time. The work of collecting original data is usually limited by time, money and manpower available for the study. When the data to be collected are very large in volume, it is possible to draw reasonably accurate conclusions from the study of a small portion of the group called a sample. There are different methods of primary data collection.

Methods of collecting primary data

Primary data are collected for the first time through census or sample survey. Following are the methods for collecting primary data.

- Direct personal interview or observation. The investigator personally meets respondent and asks questions to gather the necessary information.
- Mailed Questionnaires. Under this method a list of questions called questionnaire is prepared and is sent to all the respondents to reply on.
- Any other methods like field observation, experimentation, focus group discussion etc based on the nature of our objectives and data to be collected.

Activity: Dear learners differentiate the applicability of each method of primary data collection.

Secondary Data: In statistics the investigator need not begin from the very beginning, he/she may use and must take into account what has already been discovered by others. When an investigator uses the data which has already been collected by others, such data are called secondary data. Secondary data can be obtained from journals, reports, government publications, publications of research organizations, etc.

However, secondary data must be used with utmost care. The reason is that such data may be full of errors because of bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors, etc. Even if there is no error, secondary data may not be suitable and adequate for the purpose of inquiry.

Before using secondary data the investigator should examine the following aspects:

- ✓ The accuracy or originality of the data
- ✓ Whether the data are suitable for the purpose of current investigation
- ✓ How the data has been collected and processed
- ✓ How far the data has been summarized
- ✓ How to interpret the data, especially when figures collected for one purpose is used for another

In general, secondary data are those which are collected by some other agency and are used for further investigation for the second time. It can be obtained from published and unpublished sources just like journals, magazines, newspapers, websites, official reports, etc.

1.5.1. Qualitative and Quantitative Data

According to **the nature of variables,** data can be classified as either qualitative or quantitative data. **Qualitative data** (also called **categorical data**) are data that include labels or names used to identify an attribute of each element. They might be nonnumeric or numeric. When the qualitative data use a numeric code, arithmetic operations such as addition, subtraction, multiplication, and division do not provide meaningful results. A qualitative variable is a variable with qualitative data. It might be nominal or ordinal data.

Examples of qualitative data: Gender, product type, blocks number, occupation type, category of tax payer, economic level (low, medium and high), academic ranks and geographic locations.

Quantitative data require numeric values that indicate how much or how many. They are numeric. Arithmetic operations often provide meaningful results for a quantitative variable. A quantitative variable is a variable with quantitative data. Examples: Number of Students, Salary, income tax, etc

Quantitative variables can be further classified into two groups: discrete and continuous. **Discrete variables** can assume only certain values, and there are usually "gaps" between the values. Example: number of children in a family. **Continuous Variables** can assume an infinite number of values in an interval between any two specific values. Example: price of gasoline per gallon, tax paid by one investor.

1.5.2. Cross-Sectional and Time Series Data

Data can be classified based on time reference as Cross-sectional and Time series data. **Cross-sectional data** are data collected at the same or approximately the same point in time. Example: number of unemployed person in African countries this year.

Time series data are data collected over several time periods. Example: electric consumption of a household from November 2009 through October 20116 E.C.

Scales of Measurement: Data may also classify based on their measurement scales. The scale of measurement determines the amount of information contained in the data and indicates the most appropriate data summarization and statistical analyses. It indicates how variables are categorized, counted, or measured. According to the scale of measurement, data can be classified as nominal, ordinal, interval and ratio.

Nominal Scale of measurement: When the data for a variable consist of labels or names used to identify values of variable and there is no ranking or order can be placed on the data is considered a nominal scale of measurement.

Examples: Sex (Male and Female) religion (Muslim, Protestant, Orthodox, Catholic, other), blood type (A, B, AB, O), marital status (unmarried, Married, Divorced, Widowed) etc are examples of nominal scale.

Ordinal Scale of measurement: Data measured at this level can be placed into categories, and these categories can be ordered, or ranked; however, precise differences between the ranks do not exist. In ordinal scale of measurement one is higher or less than the other.

Examples: service quality (Excellent, good, poor). Here hence, excellent indicate the best service, followed by good and then poor, letter grades (A, B, C, D, F,), living standard of a family (poor, medium, higher), tax payers category (A, B C) etc.

Interval Scale of measurement: data show the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure.

Examples: Intellectual question (IQ), Temperature in ^oC. One property is lacking in the interval scale: There is no true zero. For example, IQ tests do not measure people who have no intelligence. For temperature of zero degrees does not indicate lack of heat. The two common temperature scales are Celsius (C) and Fahrenheit (F). We can see that the same difference exists between 10^oC (50^oF) and 20^oC (68^OF) as between 25^oc (77^oF) and 35^oc (95^oF) i.e. the measurement scale is composed of equal-sized interval. But we cannot say that a temperature of 20^oc is twice as hot as a temperature of 10^oc. because the zero point is arbitrary.

Ratio scale of measurement: possesses all properties of interval data and the ratio of two values is meaningful.

Examples: distance, height, weight, and time. This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point. For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of 300,000 (in Birr) for one automobile to the cost of 150,000 (in Birr) for a second automobile, the ratio property shows that the first automobile is $300{,}000 \text{ Birr}/150{,}000 \text{ Birr} = 2 \text{ times}$, or twice the cost of the second automobile.

Self test activity:

1. Explain the difference between the following with example.

a. Qualitative and quantitative variables
b. Nominal and ordinal scale.
c. Interval and ration scale.

2. State the level of measurement for each of the following:

a. Individuals may be classified according to socio-economic as low, medium & high.
b. Dates of the week Monday, Tuesday, Friday
c. Patients may be characterized as unimproved, improved & much improved.
d. The height of students in Wollo University.
e. Your score on an individual intelligence test as a measure of your intelligence.

1.5.3. Organizing and summarizing data

Methods commonly used to summarize categorical and quantitative data are tabular diagrammatic and graphical methods. Tabular and graphical summaries of data can be found in annual reports, newspaper articles, and research studies. Everyone is exposed to these types of presentations. Hence, it is important to understand how they are prepared and how they should be interpreted.

Summarizing of categorical data

In this section, we discussion of how tabular and graphical methods can be used to summarize categorical data.

Frequency distribution

To make it easier for understanding is to put the data into a frequency distribution. Frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes or categories. Frequency is the number of times a value is repeated for the variable in the corresponding data operations. For example look the following table which shows working status of persons.

Work status	Code	Frequency	Relative frequency (%)
Working full time	1	912	46.2
Working part-time	2	226	11.5
Temporarily not working	3	40	2.0
Unemployed ,laid off	4	104	5.3
Retired	5	357	18.1

School	6	70	3.5
Keeping house	7	210	10.6
Others	8	54	2.7
Total		1973	100

Steps to construct categorical frequency distribution

Step 1: Prepare a table as shown below.

Category	Tally (2)	Frequency (3)	Percent (4)

Step 2: Tally the data and place the result in column (2).

Step 3: Count the tally and place the result in column (3).

Step 4: Find the percentages of values in each class by using: $\% = \frac{f}{n} \times 100$, where f= frequency of the class, n=total frequency of values. Percentages are not necessarily part of frequency distribution but they can be added since they are used in certain types of diagrammatic representations such as pie charts.

Step 5: Find the total for column (3) and (4).

Example: 30 tax-payers in a town are listed below according to their level of taxation (A, B, C).

A	C	A	C	B	B
C	A	C	B	A	A
C	C	C	C	C	B
A	B	A	B	B	C
B	C	A	B	A	C

Construct a frequency distribution for the above data set.

Solution:

1. Identify the categories and then prepare the table. You have three categories; A, B and C. By following the above procedures, you will get the following result.

Category	Tally	Frequency (Number of tax-payers)	Relative frequency	% relative frequency

A	~~////////~~	9	0.3	30%
B	~~///////~~/	9	0.3	30%
C	~~//////////~~/	12	0.4	40%
Total		30	1	100%

Pie charts and Bar Charts

A. Pie Chart

Pie chart is a circular method of data presentation in which the circle is divided into sections or wedges according to the percentage of frequencies in each category of the distribution. Since there are 360^0 in a circle, the frequency for each class must be converted into a proportional part of the circle. This conversion is done by using the formula;

$Degrees\ of\ a\ part = \frac{f}{n} * 360^o$, where f =frequency of each category and $n =$ sum of the frequencies.

Each frequency must be converted to percentage by using the formula, $\% = \frac{f}{n} * 100$

Example 1: A small business consultant is investigating the performance of several companies. The sales in 2015 (in thousands of dollars) for the selected companies were:

Corporation	A	B	C	D	E
Sales (in thousands $)	255	520	750	420	330

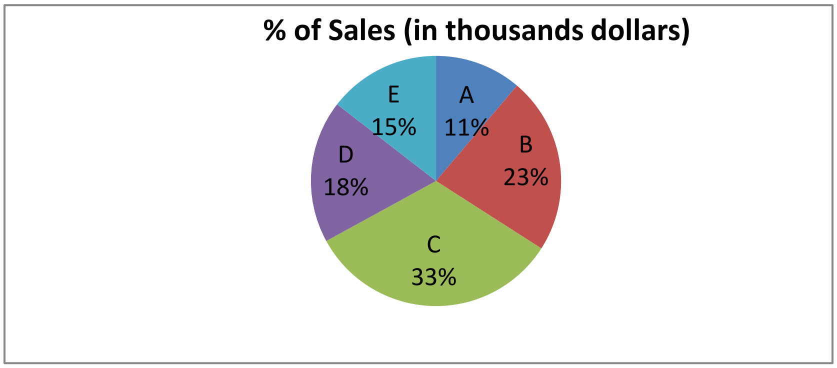
Solution:

Step 1: Find the number of degrees for each class using the formula.

Step 2: Find the percentage. The corresponding degree and percentage were computed as seen in the table.

Corporation	A	B	C	D	E	Total
Sales (in thousands $)	255	520	750	420	330	2275
Degree	40.4	82.3	118.7	66.5	52.2	360
Percentage (%)	11.2	22.9	33.0	18.5	14.5	100

Step 3: Draw the pie chart. First subdivide the circle using protractor. Then each part of the circle equals to the equivalent percentage.



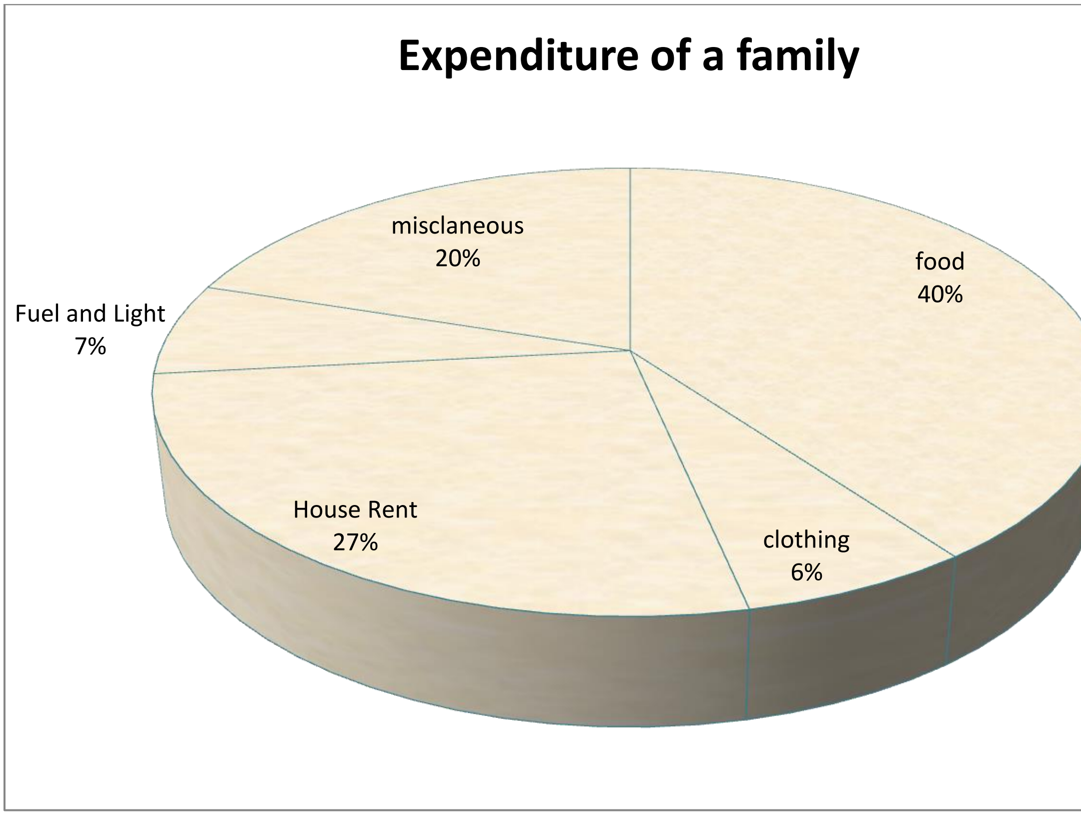
Example 2: The following table gives the details of monthly budget of a family. Represent these figures by using simple bar chart.

Item of expenditure	Family Budget($)
Food	600
Clothing	100
House rent	400
Fuel and lighting	100
Miscellaneous	300
Total	1500

***Solution**: The necessary computations are given below:*

Items	**Family budget**		
	Expenditure $	Angel of sector in ^{o}c	Percent
Food	600	144	40
Clothing	100	24	6.67
House rent	400	96	26.67
Fuel and lighting	100	24	6.67
Miscellaneous	300	72	20
Total	1500	360	100

➢ Thus the pie chart becomes as follows.



B. Bar chart

A bar chart represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data. There are different types of bar charts.

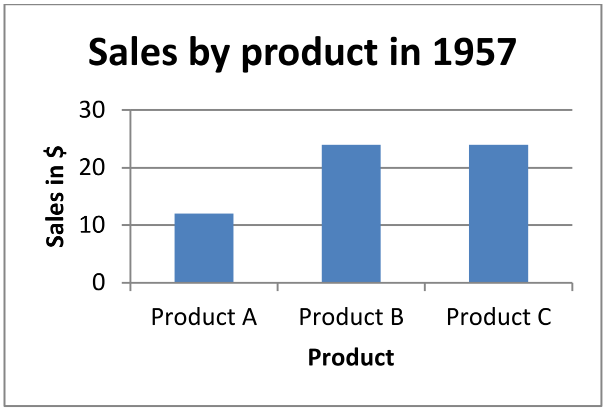
The most common are Simple bar chart, Component or sub divided bar chart and multiple bar charts.

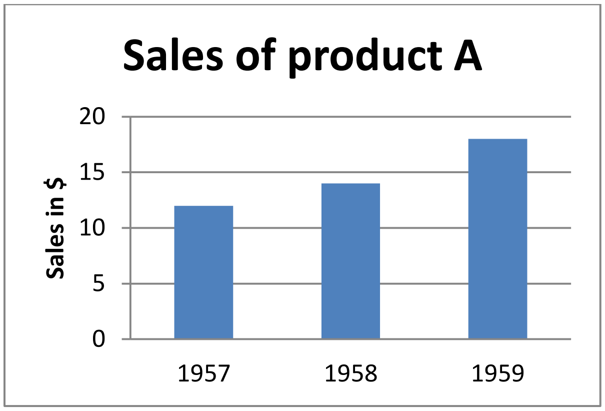
A. Simple bar chart

Simple bar chart can be drawn either on horizontal or vertical base, but bars on horizontal base more common. Bars must be uniform width and space between bars must be equal. It is used to display data on one variable.

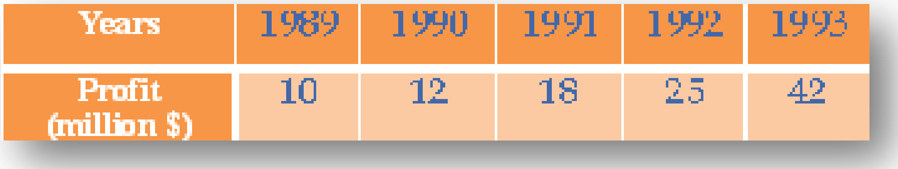
Example 1: The following data represent sale by product, 1957- 1959 of a given company for three products A, B, C. Present sale in 1957 and product A by year using simple bar chart.

Product	Sales ($) in 1957	Sales ($) in 1958	Sales ($) in 1959
A	12	14	18
B	24	21	18
C	24	35	54



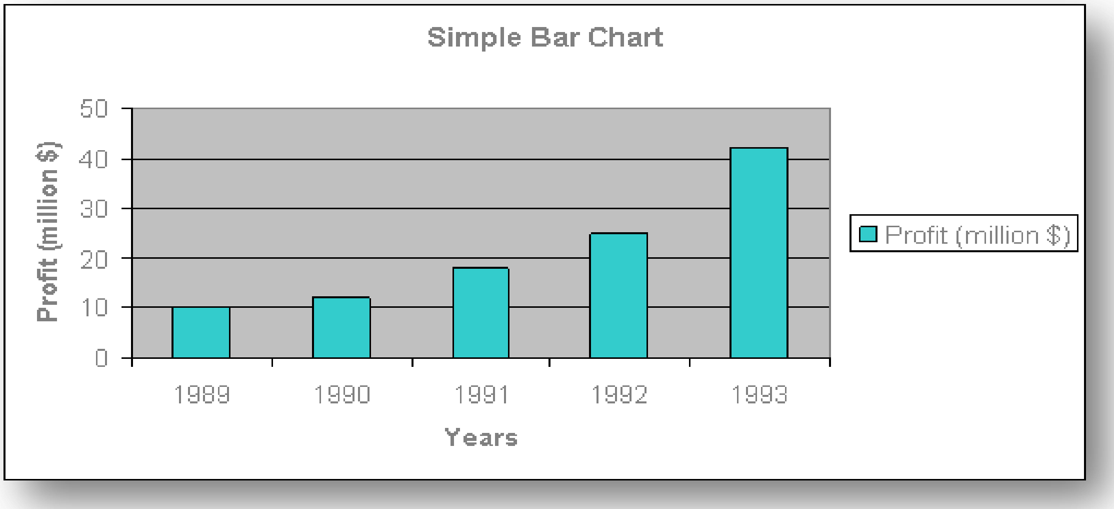


Example 2: Draw simple bar diagram to represent the profits of a bank for 5 years.



Years	1989	1990	1991	1992	1993
Profit (million $)	10	12	18	25	42

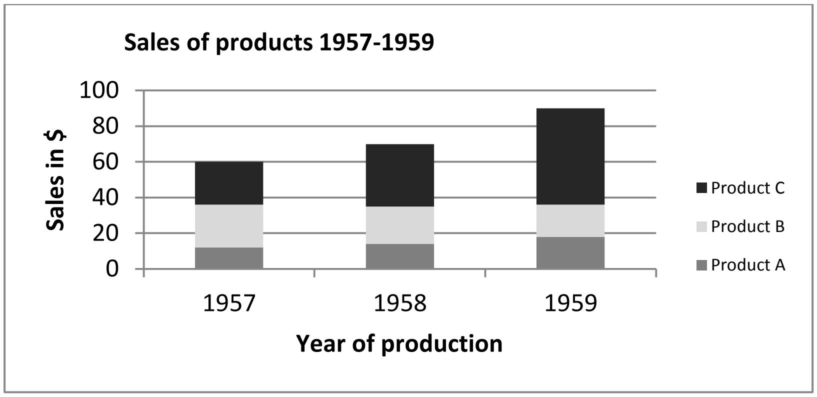
Solution: We can present as following



B. <u>Component or sub divided Bar chart</u>

In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total. The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

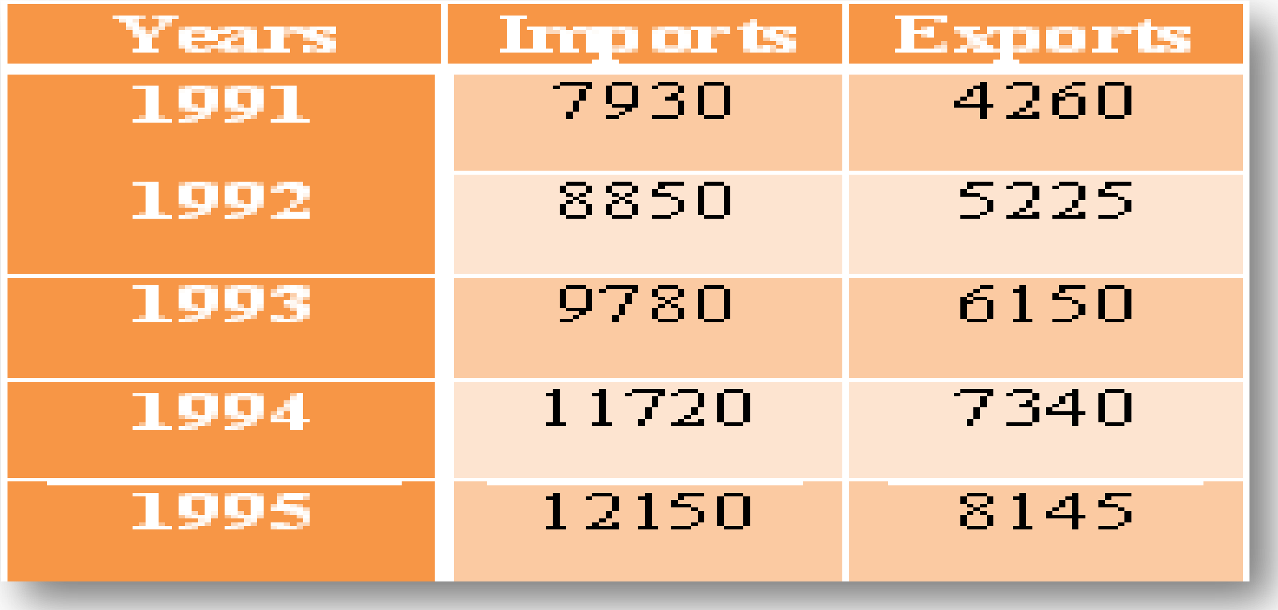
Example: For the above the draw component bars chart to represent the sales by product from 1957 to 1959?



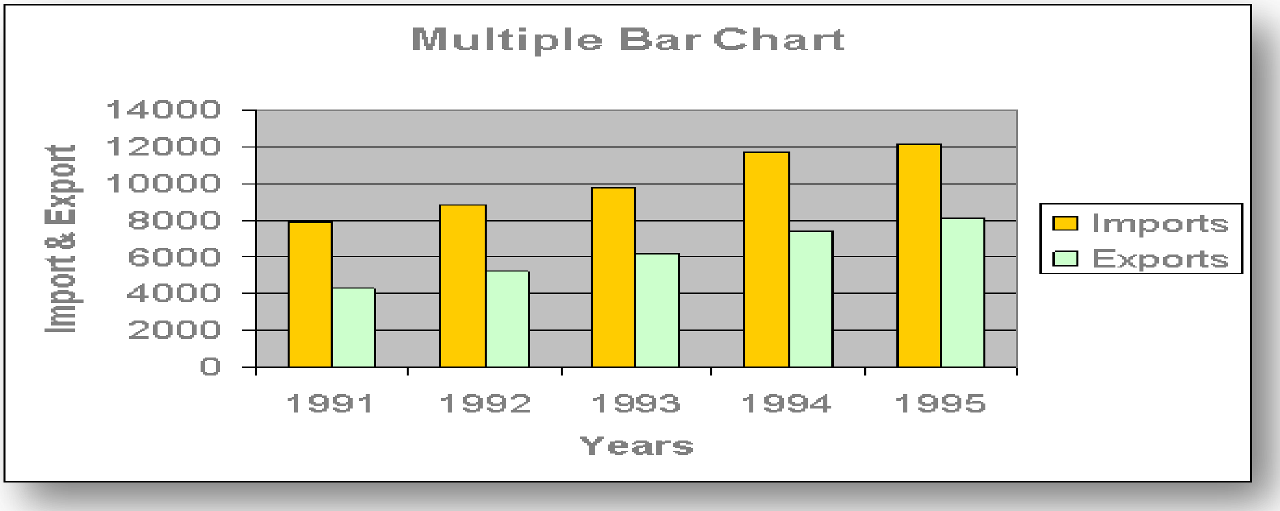
C. Multiple bar charts

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dotting.

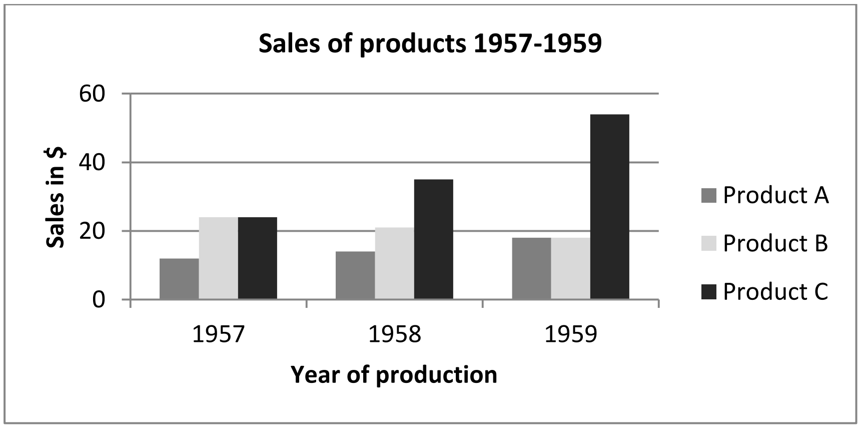
Example: Draw a multiple bar chart to represent the import and export of Canada (values in $) for the years 1991 to 1995.



Years	Imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145



Example: Draw a multiple bar chart to represent the sales by product from 1957 to 1959.



2. *Presentation of quantitative data*

As defined previously, a frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes. This definition holds for quantitative as well as qualitative data. However, with quantitative data there must be more careful in defining the non-overlapping classes to be used in the frequency distribution.

Before the procedure of constructing frequency distribution for quantitative data (Grouped frequency distribution) several things should be noted.

Class limits: the values of a variable which typically serve to identify the classes of a frequency distribution. The smaller and the larger values are known as the lower and the upper class limits, respectively. Class limits must be chosen so that each data item belongs to one and only one class.

Class boundaries: are used to separate the classes so that there are no gaps in the frequency distribution. The class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.

Class mark/Mid-point: the point which divides the class into two equal parts. This can be determined by dividing the sum of the two limits or the sum of the two boundaries by 2.

Class width: the difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes or the difference between any two consecutive class marks.

Cumulative frequency: is the number of observations less than/more than or equal to a specific value.

Cumulative frequency greater than type: it is the total frequency of all values greater than or equal to the lower class limit/ boundary of a given class.

Cumulative frequency less than type: it is the total frequency of all values less than or equal to the upper class boundary of a given class.

Cumulative Frequency Distribution (CFD): it is the tabular arrangement of class interval together with their corresponding cumulative frequencies. It can be more than or less than type, depending on the type of cumulative frequency used.

Relative frequency (rf): it is the frequency divided by the total frequency.

Relative cumulative frequency (rcf): it is the cumulative frequency divided by the total frequency.

Guidelines for Classes

1. There should be between 5 and 20 classes.
2. The classes must be mutually exclusive. This means that no data value can fall into two different classes
3. The classes must be all inclusive or exhaustive. This means that all data values must be included.
4. The classes must be continuous. There are no gaps in a frequency distribution.
5. The classes must be equal in width. The exception here is the first or last class. It is possible to have “below ..." or "... and above" class. This is often used with ages.

Steps for constructing frequency Distribution for quantitative data

1. Find the largest and the smallest value
2. Compute the Range (R) = Largest –Smallest

3. Decided the number of classes (K), usually between 5 and 20 or with the help of Sturges' Rule. According to him, the number of classes can be determined by the formula $K = 1 + 3.32\log_{10} n,$ Where **n** is number of observations

4. Find the class width (W) by dividing the range by the number of classes and rounding up to get an integer value. $W = {}^{R}/_{K}$

5. Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. Add the width to the first lower class limit to get the lower limits

6. Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

7. Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit.

8. Tally the data

9. Find the numerical frequencies from the tallies.

10. Find the cumulative frequencies (greater than and less than types). If necessary find relative frequencies and cumulative relative frequency (greater than and less than types).

Example: Construct a frequency distribution for the following data.

11	29	6	33	14	31	22	27	19	20
18	17	22	38	23	21	26	34	39	27

Solution:

Step 1: Find the highest and the lowest value H=39, L=6.

Step 2: Find the range; R=H-L=39-6=33.

Step 3: Select the number of classes desired using Sturges' formula:

k=1+3.32log (20) =5.32=6(rounding up).

Step 4: Find the class width; w=R/k=33/6=5.5=6 (rounding up)

Step 5: Select the starting point, let it be the minimum observation. Then,

6, 12, 18, 24, 30, 36 are the lower class limits.

Step 6: Find the upper class limit.

E.g. the first upper class=12-U=12-1=11. Then,

11, 17, 23, 29, 35, 41 are the upper class limits.

Step 7: Find the class boundaries.

E.g. for the first class, lower class boundary=6-U/2=5.5,

Upper class boundary =11+U/2=11.5.

Then, continue adding W on both boundaries to obtain the rest boundaries. By doing, so one can obtain the following class boundaries:

Step 8: Tally the data.

Step 9: Write the numeric values for the tallies in the frequency column.

Step 10: Find cumulative frequency.

Step 11: Find relative frequency or/and relative cumulative frequency.

The complete frequency distribution follows:

Class limit	Class boundary	Class Mark	Tally	Freq.	Cf (less than type)	Cf (more than type)	rf.	rcf (less than type
6 – 11	5.5 – 11.5	8.5	//	2	2	20	0.10	0.10
12 – 17	11.5 – 17.5	14.5	//	2	4	18	0.10	0.20
18 – 23	17.5 – 23.5	20.5	//////	7	11	16	0.35	0.55
24 – 29	23.5 – 29.5	26.5	////	4	15	9	0.20	0.75
30 – 35	29.5 – 35.5	32.5	///	3	18	5	0.15	0.90
36 – 41	35.5 – 41.5	38.5	//	2	20	2	0.10	1.00

Activity: The following data shows number of hours 40 employees spends on their job for the last 7 working days:

62 62 50 35 36 31 43 43 43 41 31 65 30 41 58 49 41 58 61 38 37

27 47 65 50 45 48 27 53 40 29 63 34 44 32 41 26 50 47 37. Construct grouped frequency distribution.

Graphical presentation of quantitative data

1. Histogram

It consists of a set of adjacent rectangles whose bases are marked off by class boundaries (not class limits) along the horizontal axis and whose heights are proportional to the frequencies associated with the respective classes.

To construct a histogram from a data set:

1. Arrange the data in increasing order.
2. Choose class intervals so that all data points are covered.

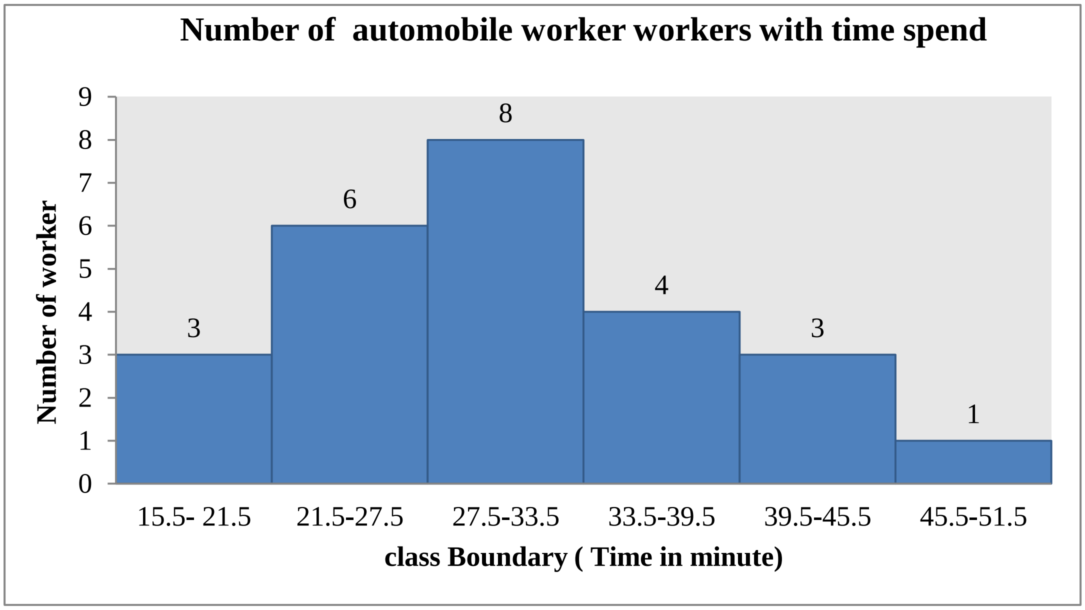
3. Construct a frequency table.
4. Draw adjacent bars having heights determined by the frequencies in step3.

The importance of a histogram is that it enables us to organize and present data graphically so as to draw attention to certain important features of the data. For instance, a histogram can often indicate how symmetric the data are; how spread out the data are; whether there are intervals having high levels of data concentration; whether there are gaps in the data; and whether some data values are far apart from others.

Example: Construct a histogram for the frequency distribution of the time spent by the automobile workers.

Time (in minute)	Class mark	Number of workers
15.5- 21.5	18.5	3
21.5-27.5	24.5	6
27.5-33.5	30.5	8
33.5-39.5	36.5	4
39.5-45.5	42.5	3
45.5-51.5	48.5	1

The histogram of the data can be constructed by using the frequency in the vertical axis and class boundaries in the horizontal axis.

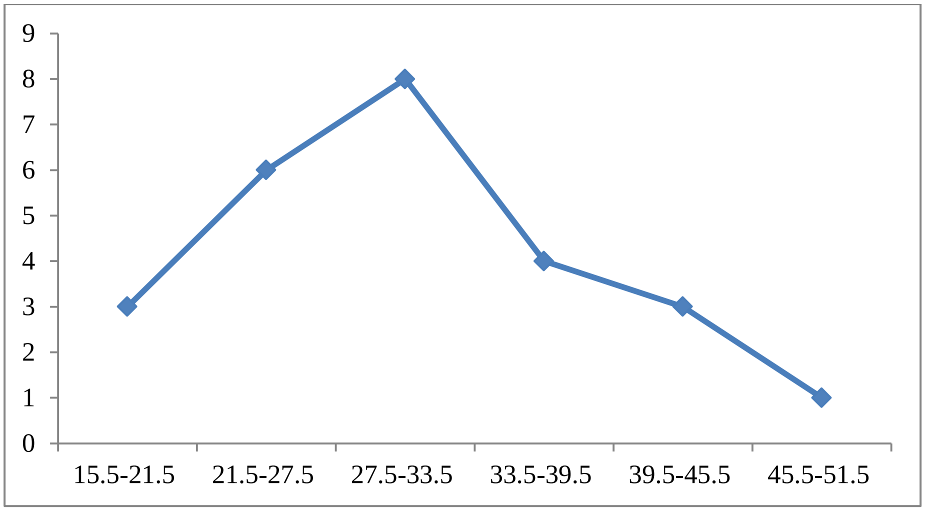


2. **Frequency Polygon:** A frequency polygon is a line graph drawn by taking the frequencies of the classes along the vertical axis and their respective class marks along the horizontal axis. Then join the cross points by a free hand curve.

Example: Draw a frequency polygon presenting the following data.

Time (in minute)	Class mark	Number of workers
15.5-21.5	18.5	3
21.5-27.5	24.5	6
27.5-33.5	30.5	8
33.5-39.5	36.5	4
39.5-45.5	42.5	3
45.5-51.5	48.5	1

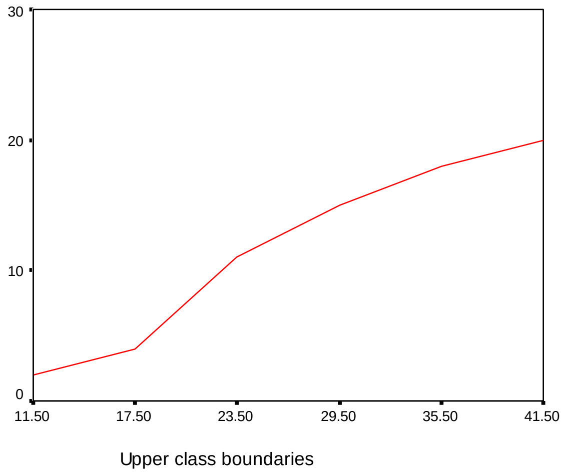
The frequency polygon drown as follows using class boundaries and frequency



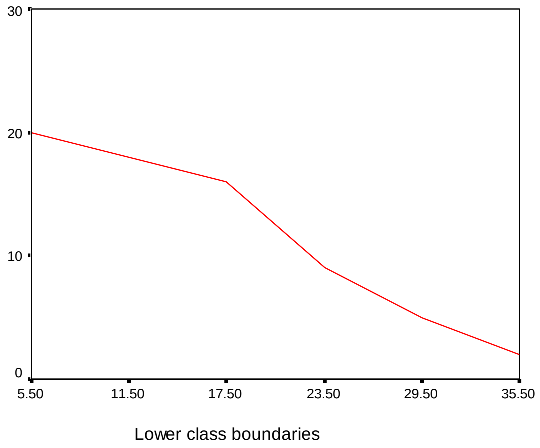
3. **Cumulative Frequency Polygon (Ogive)**

Cumulative frequency polygon can be traced on less than or more than cumulative frequency basis. Place the class boundaries along the horizontal axis and the corresponding cumulative frequencies (either less than or more than cumulative frequencies) along the vertical axis. Then join the cross points by a free hand curve.

Example: the data in the above example can be presented using either a less than or a more than cumulative frequency polygon as given below (i) and (ii) respectively.



(*i*) Less than type cumulative frequency curve



(ii) More than type cumulative frequency curve

1.6. Measures of Central Tendency

So far, we discussed how raw data can be organized in terms of tables, charts and frequency distributions in order to be easily understood and analyzed. Frequency distributions and their corresponding graphical displays roughly tell us some of the features of a data set. However, they don't condense the mass of data in a way that we can easily understand and interpret. Here, we will see how to summarize data using a descriptive measure called average. This will help us in condensing a mass of data into a single value which is in some sense representative of the whole data set. Measures of Central Tendency give us information about the location of the center of the distribution of data values. A single value that describes the characteristics of the entire mass of data is called measures of central tendency.

The main objectives of measuring central tendency are:

- To get a single value that represent(describe) characteristics of the entire data
- To summarizing/reducing the volume of the data
- To facilitating comparison within one group or between groups of data
- To enable further statistical analysis

1.6.1. The Summation Notation

Let X_1, X_2, $X_{3,\ldots,}$ X_N be a number of measurements where N is the total number of observation and X_i is i^{th} observation, then it is very often in statistics an algebraic expression of the form $X_1+X_2+X_3+...+X_N$ is used in a formula to compute a statistic. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the summation notation.

The symbol $\sum_{i=1}^{N} X_i$ is mathematical shorthand for $X_1+X_2+X_3+...+X_N$

- ➢ $\sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N$. The expression is read as "the sum of X_i where i run from 1 to N." It means "add up all the numbers."

Example: Suppose the following were prices of five different models of printers in a computer store (in $1,000): 5, 7, 7, 6, and 8. In this example to five numbers, where N=5, the summation could be written:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 5+7+7+6+8 = 33$$

The "N" in the upper part of the summation notation tells where to end the sequence of summation. If there were only three scores then the summation and example would be:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 5 + 7 + 7 = 19$$

Sometimes if the summation notation is used in an expression and the expression must be written a number of times, as in a proof, then a shorthand notation for the shorthand notation is employed. When the summation sign " $\sum$ " is used without additional notation, then "i=1" and "N" are assumed.

1.6.1.1. Properties of Summation

1. $\sum_{i=1}^{n} K = nK$, Where k is any constant

2. $\sum_{i=1}^{n} KX_i = K\sum_{i=1}^{n} X_i$, Where k is any constant

3. $\sum_{i}^{n} (a + bX_i) = na + b\sum_{i=1}^{n} X_i$, where a and b are any constant

4. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

5. $\sum_{i=1}^{N} (X_i \times Y_i) = X_1 \times Y_1 + X_2 \times Y_2 + - - - + X_N \times Y_N$

Example: Let X_1=1, X_2=2, X_3=3, X_4= 4 and X_5=5, then find

$$i)\sum_{i=1}^{3}\left(1 - 3x_i\right) = \sum_{i=1}^{3} 1 + 3\sum_{i=1}^{3} x_i = 3(1) + 3(1 + 2 + 3) = 21$$

$$ii)\sum_{i=3}^{5}\left(x_i - 2x_i^2\right) = \sum_{i=3}^{5} x_i + 2\sum_{i=3}^{5} x^2{}_i = 12 + 2(9 + 16 + 25) = 92$$

Activity: Considering the following data

X	5	7	7	6	8
Y	6	7	8	7	8

Determine a) $\sum_{i=1}^{5} X_i$ b) $\sum_{i=1}^{5} Y_i$ c) $\sum_{i=1}^{5} 11$ d) $\sum_{i=1}^{5} (X_i + Y_i)$ e) $\sum_{i=1}^{5} (X_i - Y_i)$

f) $\sum_{i=1}^{5} X^2{}_i$ g) $\sum_{i=1}^{5} X_i Y_i$ h) $\sum_{i=1}^{5} X_i + \sum_{i=1}^{5} Y_i$ i) $\sum_{i=1}^{5} X_i \sum_{i=1}^{5} Y_i$

1.5.2 Properties of Measures of Central Tendency

The Characteristics of a good measure of central tendency or a typical average should the following properties:

- It should be defined rigidly which means that it should have a definite value.
- It should be based on all observation under investigation.
- It should be not be affected by extreme observations.
- It should be capable of further algebraic treatment.
  - It should be as little as affected by fluctuations of sampling or should be stable with sampling.
- It should be ease to calculate and simple to understand.
- It should be unique and always exist.

Note: There is no measure satisfied all the above condition, we choose the one that satisfies most of the properties!

1.6.1.2. Types of Measures of Central Tendency

There are several different measures of central tendency; each having its own advantage and disadvantage, including the mean, median and mode. The choice of these averages depends up on which one best fits the property under discussion.

a. **Mean**: The mean is usually best measure of central tendency when there are no outliers/extreme values. It can calculate either from the population or sample values. However; most of the time, the population values are unknown and hence we may use sample values to estimate population mean. There are three types of means which are suitable for a particular type of data. These are arithmetic Mean, geometric Mean and harmonic Mean. In this module, the arithmetic mean is covered.

The Arithmetic Mean

Arithmetic means are two types; simple arithmetic mean and the weighted arithmetic mean.

Simple Arithmetic Mean:

The mean is defined as the sum of the magnitude of the items divided by the number of items

- The mean of X_1, X_2 ,X_3 …X_n is denoted by A.M, and is given by:

$$\overline{X} = \frac{X_1 + X_2 + K + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- When the data are arranged or given in the form of frequency distribution i.e. there are k variate values such that a value X_i has a frequency f_i (i=1,2,---,k) ,then the Arithmetic mean will be

$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$ Where k is the number of classes and $\sum_{i=1}^{k} f_i = n$.

Arithmetic Mean for Grouped Data

If data are given in the shape of a continuous frequency distribution, then the arithmetic mean is obtained as follows:

$\overline{X} = \frac{\sum_{i=1}^{n} f_i m_i}{\sum_{i=1}^{n} f_i}$, m_i =the class mark of the i^{th} class and f_i= the frequency of the i^{th} class

Example: Find the arithmetic mean for the following frequency distribution

Class interval	Frequency	Class mark
3-5	3	4
6-8	2	7
9-11	5	10
12-14	4	13

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i m_i}{\sum_{i=1}^{n} f_i} = \frac{3*4+2*7+5*10+4*13}{3+2+5+4} = \frac{128}{14} = 9.14$$

NB: If we take an entire population mean is denoted by Greek letter μ and given as

$$\mu = \begin{cases} \frac{\sum_{i=1}^{N} x_i}{N} \\ \frac{\sum_{i=1}^{k} x_{if_i}}{\sum_{i=1}^{k} f_i} \\ \frac{\sum_{i=1}^{k} M_{if_i}}{\sum_{i=1}^{k} f_i} \end{cases} \quad \text{Based on the nature of the data.}$$

Activity

1) Daily cash earnings of 15 workers working in different industries are as follows: 11.63,8.22,12.56,12.14,29.23,18.23,11.49,11.30,17.00,9.16,8.64,27.56,8.23,19.77,12.81. Find the average daily earning of a worker?

2) The cost (in $1,000) at first production of 130 cars were as given below

Cost(X): 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29.

No. of cars (f): 2, 1, 4, 8, 10, 12, 17, 19, 18, 14, 13, 12. Compute the average cost of cars at first production?

3) Calculate the mean for the following Frequency distribution.

Class	**6-10**	**11-15**	**16-20**	**21-25**	**26-30**	**31-35**
Frequency	35	23	15	12	9	6

Special properties of Arithmetic mean

1. The sum of the deviations of a set of items from their mean is always zero. i.e.

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

2. The sum of the squared deviations of a set of items from their mean is the minimum. i.e. $\sum_{i=1}^{n}(X_i - \overline{X})^2 < \sum_{i=1}^{n}(X_i - A)^2$, for any constant A.

3. If $\bar{X}_1$ is the mean of observations n_2, and $\bar{X}_2$ is the mean of observations n_2, etc, and $\bar{X}_k$ is the mean of observations n_k , then the mean of all the observation in all groups, often called the **combined mean**, is given by: $\bar{X}_c = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i}$.

4. If a wrong figure has been used when calculating the mean the correct mean can be obtained without repeating the whole process using:

$Corrected\ mean = Wrong\ \ mean + \frac{(correct\ \ value - Wrong\ value)}{n}$. Where n is total number of observations.

5. The effect of transforming original series on the mean.

a) If a constant *k* is added/ subtracted to/from every observation then the new mean will be *the old mean*± *k* respectively i.e. $\overline{X}_{new} = \overline{X}_{old} \pm K$.

b) If every observations are multiplied by a constant *k* then the new mean will be *k*old mean i.e.* $\overline{X}_{new} = \overline{X}_{old} \times K$.

Merits and Demerits of Arithmetic Mean

Merits:

- It is rigidly defined.
- It is based on all observations.
- It is suitable for further mathematical treatment.
- It is a stable average, i.e. it is not affected by fluctuations of sampling to some extent.
- It is easy to calculate and simple to understand.

Demerits:

- It is affected by extreme observations.
- It cannot be used in the case of open end classes.
- It cannot be determined by the method of inspection.
- It cannot be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.
- It can be a number which does not exist in a series of data.
- Sometimes it leads to wrong conclusion if the details of the data from which it is obtained are not available.
- It gives high weight to high extreme values and less weight to low extreme values.

The Weighted arithmetic mean:

In some cases the data in the sample or population should not be weighted equally, and each value weighted according to its importance. There is a measure of average for such problems known as weighted Arithmetic mean. Weighted arithmetic mean is used to calculate the average when the relative importance of the observations differs. This relative importance is technically known as weight. Weight could be a frequency or numerical coefficient associated with observations.

Definition: If X_1, X_2,…, X_n have weights W_1, W_2, …, Wn, respectively. Then the weighted arithmetic mean is denoted by $\bar{x}_w$ is defined as:

$$\bar{x}_w = \frac{X_1W_1+X_2W_2+\cdots+X_nW_n}{W_1+W_2+\cdots+W_n} = \frac{\sum_{i=1}^{n} w_iX_i}{\sum_{i=1}^{n} w_i} \ .$$

Example 1: The GPA or CGPA of students is a good example of a weighted arithmetic mean. Suppose that Solomon obtained the following grade in the first semester of freshman program at Wollo University in 2007.

Course	Credit hour(W_i)	Grade
Math101	4	A=4
Bio101	3	C=2
Chem101	3	B=3
Phys101	4	B=3
Flen101	3	C=2

Find the GPA of Solomon.

Solution: $\bar{x}_w = \frac{X_1W_1+X_2W_2+\cdots+X_nW_n}{W_1+W_2+\cdots+W_n} = \frac{\sum_{i=1}^{n} w_iX_i}{\sum_{i=1}^{n} w_i} = \frac{4x4+3x2+3x3+4x3+3x2}{4+3+3+4+3} = \frac{49}{17} = 2.88$

Example 2: In a vacancy for a position of animal breeder in an organization, then criteria of selection were work experience, entrance exam, and interview results. The relative importance of these criteria was regarded to be different. The weights of these criteria and scores obtained by 3 candidates (out of 100 in each criterion) are given in the following table. In addition, the selection of a candidate is based on the average result on these criteria.

Criterion	Weight	Candidates		
		Dechasa	Gutema	Engdawork
Work experience	4	70	89	85
Entrance exam	3	78	83	89
Interview result	2	90	92	90

Who is the appropriate candidate for this position based on the criteria?

Solution: We use the weighted mean since the relative importances of the criterion are different:

Criterion	Weight (Wi)	Candidates					
		Dechasa		Gutema		Engdawork	
		Xi	Xi Wi	Xi	XiWi	Xi	XiWi
Work experience	4	70	280	89	356	85	340
Entrance exam	3	78	234	83	249	89	267
Interview	2	90	180	92	184	90	180
Total	9	238	694	264	789	264	787

The weighted mean and the simple arithmetic mean for the applicants are as follows:

Applicants	Weighted mean	Simple arithmetic mean
Dechasa	694/9 =77.1	238/3 = 79.33
Gutema	789/9 = 87.67	264/3 = 88
Engdawork	787/9 = 87.44	264/3 = 88

If we use the simple arithmetic mean of the scores, both Gutema and Engdawork have got equal chances to be recruited. However, the relative importance of the criteria is different. So we have to use the weighted mean for discriminating among the candidates. The weighted mean of the scores obtained by Gutema is larger than the others. So Gutema should be recruited for the job.

Activity: A student obtained the following percentage in an examination: English 60, management 75, Mathematics 63, accounting 59, and economics 55.Find the students weighted arithmetic mean if weights 1, 2, 1, 3, 3 respectively are allotted to the subjects.

The Mode

The mode is a value which occurs most frequently in a set of values, and which occurs more than once. The mode may not exist and even if it does exist, it may not be unique. In case of discrete distribution, the value having the maximum frequency is the modal value. If in a set of observed values, all values occur once or equal number of times, then, there is no mode.

Example

a) Find the mode of 5, 3, 5, 8, 9

Solution: Mode =5

b) Find the mode of 8, 9, 9, 7, 8, 2, and 5. It is a bimodal Data: 8 and 9

c) Find the mode of 4, 12, 3, 6, and 7. No mode for this data.

The mode of a set of numbers X_1, X_2, …, X_n is usually denoted by $\hat{X}$.

Mode for Grouped data

If data are given in the shape of continuous frequency distribution, the mode is defined as:

$$\hat{X} = L_{\text{mod}} + (\frac{\Delta_1}{\Delta_1 + \Delta_2})W$$

Where: $\hat{X}$ = the mode of the distribution

L_{mod}= the lower class boundary of the modal class

$$\Delta_1 = f_{mo} - f_1$$
$$\Delta_2 = f_{mo} - f_2$$

f_{mod}= frequency of the modal class

f_1= frequency of the class preceding the modal class

f_2= frequency of the class succeeding the modal class

W=the size of the modal class

Note: The modal class is a class with the highest frequency

Example: Find the mode for the frequency distribution given by below.

Class interval	3-6	6-9	9-12	12-15

Frequency	4	8	10	3

$$\operatorname{mod} e = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)cw$$
$$= 9 + \left(\frac{2}{2+7}\right)3 = 9 + \left(\frac{2}{9}\right)3$$
$$= \frac{29}{3}$$

Activity: The following is the distribution of the monthly bills for electricity of certain households selected at random from a district. Calculate the mode of the distribution.

Monthly payment	**6-15**	**16-25**	**26-35**	**36-45**	**46-55**	**56-65**	**66-75**
No. of households	8	12	17	29	31	5	3

Merits and Demerits of Mode

Merits:

- It is not affected by extreme observations.
- Easy to calculate and simple to understand.
- It can be calculated for distribution with open end class.
- Can be used for qualitative data as well.

Demerits:

- It is not rigidly defined.
- It is not based on all observations
- It is not suitable for further mathematical treatment.
- It is not stable average, i.e. it is affected by fluctuations of sampling to some extent.
- Often its value is not unique.

The Median

In a distribution, median is the value of the variable which divides the data in to two equal halves. In an ordered series of data, the median is an observation lying exactly in the middle of the series. It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

Let X_1, X_2, …, X_n be the observations, then the numbers arranged in ascending order will be $X_{[1]}$, $X_{[2]}$, …$X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value.

Here, we find that $X_{[1]} < X_{[2]} < \ldots < X_{[n]}$

Median is denoted by $\tilde{X}$.

Median for ungrouped data

$$Median = \begin{cases} X_{\frac{n+1}{2}} & ,if\ n\, is\ odd \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}), & ,if\ n\ is\ even \end{cases}$$

Example 1.19: Find the median of the following data.

a) 3,8,4,7,7,5,6,8,7,4,6,8,9,7,6

Arrange the given data in either increasing or decreasing order:

3,4,4,5,6,6,7,7,7,7,8,8,8,9

Median = 7

b) 3,4,4,5,6,6,6,7,7,7,7,8,8,8

Median= $\frac{6+7}{2} = 6.5$

Activity

a) Actual waiting time for the first job on the selected sample of nine people having different field of specializations was given below. Waiting time (in month):11.6, 11.3, 10.7, 18.0, 3.3, 9.2, 8.3, 3.8, 6.8. Calculate the median of the waiting time?

b) The export of agricultural products in million dollars from a country during eight quarters in 1974 and 1975 was, 29.7, 16.6, 2.3, 14.1, 36.6, 18.7, 3.5, 21.3. Find the median of the given set of values?

Median for grouped data: If data are given in the shape of continuous frequency distribution, the median is defined as:

$$\tilde{X} = L_{med} + \frac{W}{f_{med}}(\frac{n}{2} - f_c)$$

Where: L_{med} =lower class boundary of the median class.

$f_{med} = The\ \ frequency\, of\ \ the\ \ median\ \ class$

$f_c = The\ comulative\ frequency\left(less\ than\ type\right) preceding\ the\ \ median\, class.$

W=the size of the median class and n=total number of observation.

Note: The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to n/2.

Example: Find the median wage of the following distribution

Wages(in Rs)	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No.of workers	3	5	20	10	5

Solution:

Wages(in Rs)	No.of workers	Cf
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-3000	5	43

Here N/2 =43/2=21.5. So, cf> 21.5 is 28 and the corresponding class is 4,000-5,000, so the median class is 4,000-5,000, and

$$median = 4000 + \frac{1000}{20}(21.5 - 8) = 4,675$$

$$so\, the\, wage\, is\, 4,675$$

Activity: Find the median of the following distribution.

Class	40-44	45-49	50-54	55-59	60-64	65-69	70-74
Frequency	7	10	22	15	12	6	3

Merits and Demerits of Median

Merits:

- Median is a positional average and hence not influenced by extreme observations.
- Can be calculated in the case of open end intervals.
- Median can be located even if the data are incomplete.

Demerits:

- It is not a good representative of data if the number of items is small.

- It is not amenable to further algebraic treatment.
- It is susceptible to sampling fluctuations.

Empirical relationship between $\overline{X}, \hat{X}, and\ \tilde{X}$

$\overline{X} = \hat{X} = \tilde{X}$, for symmetrical distribution and $\overline{X} - \hat{X} = 3(\overline{X} - \tilde{X})$, for unimodal skewed or asymmetrical frequency distribution

1.7. Measures of Dispersion (Variation)

Measure of central tendency alone does not adequately describe a set of observation unless all observations are the same. So we need some additional information like

1) The extent to which the items in a particular distribution are scatters around the central tendency i.e. measure of dispersion.
2) The direction of scatteredness whether more items are attached towards higher or lower values i.e. measure of skewness.

A measure of scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data. Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.

Objectives of measuring Variation:

- ➢ To judge the reliability of measures of central tendency
- ➢ To control variability itself.
- ➢ To compare two or more groups of numbers in terms of their variability.
- ➢ To make further statistical analysis.

1.7. Types of Measures of Dispersion

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

1) Range and relative range
2) Variance, Standard deviation and coefficient of variation.

1.7.1. The Range (R) and Relative Range (RR)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

istribution 1	2	5	6	6	7	8	0	2	2	3	3	5
istribution 2	2	2	3	3	3	4	4	4	4	4	5	5

For this reason, among others, the range is not the most important measure of variability.

For ungrouped data: R= $X_{\max} - X_{\min}$ where $X_{\max}$ is the maximum observation and $X_{\min}$ is the minimum observation. For grouped data: $R = UCL_{last} - LCL_{first}$ where UCL_{last} is the last upper class limit and LCL_{first} is the first lower class limit.

Relative Range (RR): It is also sometimes called coefficient of range and given by:

For ungrouped data: $RR = \dfrac{X_{\max} - X_{\min}}{X_{\max} + X_{\min}}$

For grouped data: $RR = \dfrac{UCL_{last} - LCL_{first}}{UCL_{last} + LCL_{first}}$

Activity

1) Find the R and RR and then identify which data is more dispersed?

a) For the month income of 10 workers X_i: 347, 420, 500,600,696,710, 835, 850, and 900.

b) For the following age distribution.

Class	**6-10**	**11-15**	**16-20**	**21-25**	**26-30**	**31-35**
Frequency	35	23	15	12	9	6

2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of:

a) Smallest observation

b) Largest observation

1.7.2. The Variance, the Standard deviation and the coefficient of Variation

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

For ungrouped data: Population variance= $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ $\quad Where \ \ \mu \ is \ the \ \ population mean$

For grouped data: Population variance = $\sigma^2 = \frac{\sum_{i=1}^{N} f_i (X_i - \mu)^2}{N}$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

For ungrouped data:

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

For grouped data:

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

We usually use the following computational formula.

$$S^2 = \frac{\sum_{i=}^{n} x^2{}_i - n\bar{x}^2}{n-1}, \ for \ ungrouped data$$

$$S^2 = \frac{\sum_{i=}^{n} f_i x^2{}_i - n\bar{x}^2}{n-1}, \ for \ \ frequency \ \ distribution \cdot$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X - \mu)^2}{N}}$ with class frequency

$\sigma = \sqrt{\frac{\sum_{i=1}^{N} f_i (X-\mu)^2}{N}}$ for grouped frequency distribution.

Population standard deviation= $\sigma = \sqrt{\sigma^2}$

Sample standard deviation = $S = \sqrt{S^2}$

Remark: If the standard deviation of a set of data is small the values are more concentrated around the mean and if the standard deviation is large, the value is more scattered widely around the mean.

Properties of standard deviation

1) It is considered to be the best measure of dispersion and is used widely.

2) There is however one difficulty with it. If the unit of measure of variables of two series is not same, and then the variability cannot be compared by comparing the values of standard deviation**.**

3) If the standard deviation of $x_1, x_2, ..., x_n$ observation is S, then the standard deviation of

a) $x_1 + k, x_2 + k, ..., x_n + k$ is also S. Where k is a constant number.

b) $x_1 - k, x_2 - k, ..., x_n - k$ is also S.

c) $x_1 \times k, x_2 \times k, ..., x_n \times k$ is $|k|S$

Example : Find the variance and standard deviation of the following sample data 5, 17, 12, 10.

Solution: $\boldsymbol{\bar{X}} = \frac{\sum_{i=1}^{4} X_i}{4} = 11$

X_i	5	10	12	17	Total
$(X_i - \bar{X})^2$	36	1	1	36	74

$$S2 = \frac{\sum_{i=1}^{4}(x_i - \bar{X})^2}{4-1} = 74/3 = 24.67 \;=>\; S = \sqrt{S2} = \sqrt{24.67} \;=\; 4.97$$

Activity

i) The sample data is given in the form of frequency distribution as follows, then find the variance and standard deviation.

Class	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65-69**	**70-74**
Frequency	7	10	22	15	12	6	3

ii) The mean and the standard deviation of a set of numbers are respectively 500 and 10.

a) If 10 is added to each of the numbers in the set, then what will be the variance and standard deviation of the new set?

b) If each of the numbers in the set are multiplied by -5, then what will be the variance and standard deviation of the new set?

The Coefficient of Variation (C.V) :- Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{\sigma}{\mu} \times 100 \quad for \ \ population$$

$$= \frac{S}{\overline{X}} \times 100 \quad for \ \ the \ \ sample$$

Remark: Smaller the value of C.V, more consistent is the data and vice versa.

Example: Consider the distribution of per dim (per employee) of two companies. For the first company, the mean and S.d are 60 and 10 respectively. For the second company the mean and S.d are 50 and 9, respectively

$$Cv_1 = \frac{10}{60} x\, 100 = 16.7\%$$

$$Cv_2 = 9/50\, x\, 100 = 18.0\%$$

This shows that the variability in first company is less as compared to that in the second company

Activity: Two distributions A& B have mean 80 inch and 20 kg and s. deviation is 10 inch and 1.5 kg respectively. Which distribution has greater variability?

The Standard Scores (Z-scores): If X is a measurement from a distribution with mean $\overline{X}$ and standard deviation S, then its value in standard units is

$$Z = \frac{X_i - \mu}{\sigma} \quad for \ \ population$$

$$= \frac{X_i - \overline{X}}{S} \quad for \ \ the \ \ sample$$

- Z gives the deviations from the mean in units of standard deviation and it tell us how many S.D a given value lie above or below the mean.
- It also helps in hypothesis testing. It is used to compare two observations coming from different groups.

Example: Two groups of children were trained to perform a certain task for a month and then tested to find out which group is faster to learn the task. The average time taken to perform the task was 10.4 minutes with s.d of 1.2 min &11.9 min with as.d. of 1.3 min for the 2nd group .Child A form group 1 took 9.2 min. while child B from group 2 took 9.3 min, who was faster in performing the task relative to the other

Group I	Group II
Mean = 10.4	Mean = 11.9
S. d = 1.2	S. d = 1.3
$X_A = 9.2$	$X_B = 9.3$

$$Z_A = \frac{X_A - \overline{x_1}}{s_1} = \frac{9.2 - 10.4}{1.2} \qquad Z_B = \frac{X_B - \overline{x}_B}{s_2} = \frac{9.3 - 11.9}{1.3} = -2$$

These values indicate that the time taken, by child A is one Standard deviation below the average time taken by the group. The time taken by child B is two St.dev below the mean time taken by his/her group, child B is therefore, faster in performing the task relative to the other.

1.8. Measures of shape

Moments: If X is a variable that assume the values $X_1, X_2, \ldots, X_n$,then

1. The r^th^ moment is defined as:

$$\bar{X}^r = \frac{X^r{}_1 + X^r{}_2 + \ldots + X^r{}_n}{n} = \frac{\sum_{i=1}^{n} X^r{}_i}{n}$$

- For the case of frequency distribution, this expressed as $\bar{X}^r = \frac{\sum_{i=1}^{k} f_i X^r{}_i}{n}$

If r =1, it is the simple arithmetic mean, this is called the first moment.

2. The r^th^ moment about the mean (the r^th^ central moment)

Denoted by M_r and defined as:

$$M_r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^r}{n} = \frac{n-1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^r}{n-1}$$

For the case of frequency distribution, this expressed as: $M_r = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^r}{n}$

If r= 2, then it is the population variance, this is called second central moment. If we assume $n - 1 \approx n, it\ is\ also\ the\ sample\ varaince$.

3. The r^th^ moment about any number A is defined as: Denoted by M'_r and

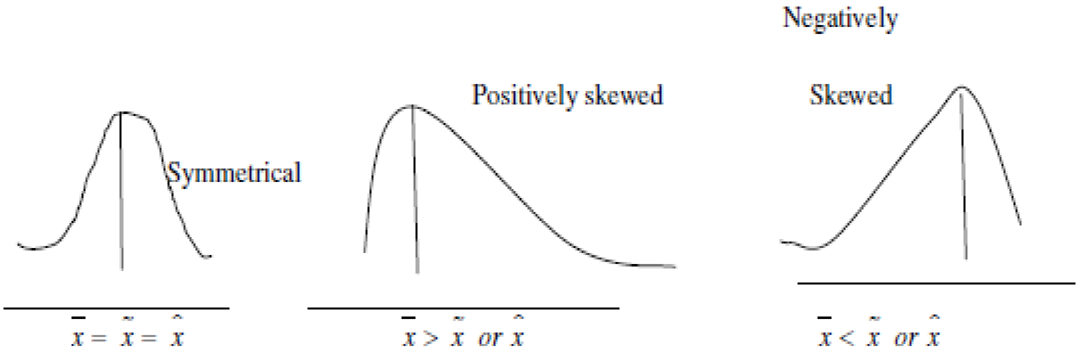
$$M'_r = \frac{\sum_{i=1}^{n}(X_i - A)^r}{n}$$

1.8.1. Skewness

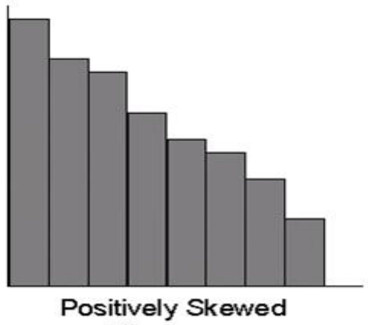
Skewness is the degree of asymmetry or departure from symmetry of a distribution. A skewed frequency distribution is one that is not symmetrical. it is concerned with the shape of the curve not size. If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.

- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

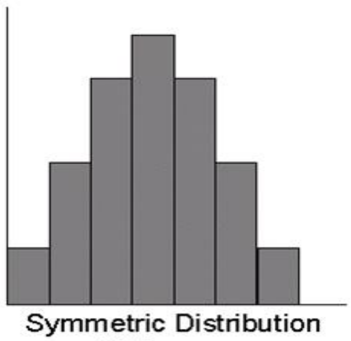
Mean- Mode = 3*(Mean-Median)


Negatively
Skewed
Positively skewed
Symmetrical
$\bar{x} = \tilde{x} = \hat{x}$
$\bar{x} > \tilde{x}$ or $\hat{x}$
$\bar{x} < \tilde{x}$ or $\hat{x}$

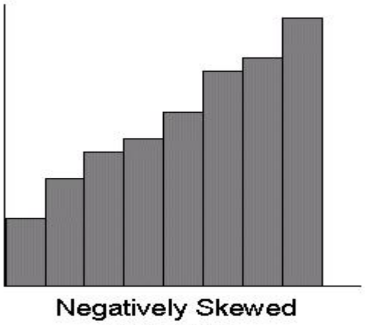
Using histogram, we can express as follows.



Positively Skewed Histogram



Symmetric Distribution Histogram



Negatively Skewed Histogram

Measures of Skewness

There are various measures of skewness.

1. The Pearsonian coefficient of skewness: The skewness of the distribution can be measured by karl Pearson's Coefficient of skewness.

$$P_{csk} = \frac{\text{Mean} - \text{Mode}}{\text{standard devaition}} = \frac{\overline{\text{X}} - \widehat{\text{X}}}{\text{S}}.$$

It is usually between -3 and 3.

$$If\ P_{csk} = \begin{cases} P_{csk} > 0, then\ the\ distribution\ is\ positively\ skewed\ (mean\ is\ greater\ than\ median). \\ \quad P_{csk} = 0, then\ the\ distribution\ is\ symetric\ (mean\ equals\ to\ median) \\ \quad P_{csk} < 0\ , the\ distribution\ is\ negatively\ skewed (mean\ is\ less\ than\ median) \end{cases}$$

2. The Bowley’s coefficient of skewness (coefficient of skewness based on quartiles)

It is usually between -1 and 1.

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3. The moment coefficient of skewness.

$\alpha_3 = \frac{M_3}{M_2{}^{\frac{3}{2}}} = \frac{M_3}{(\sigma^2)^{\frac{3}{2}}} = \frac{M_3}{\sigma^3}$, where σ is the population standard deviation.

The shape of the curve is determined by the value of α_3

$$If\ \alpha_3 = \begin{cases} \alpha_3 > 0, then\ the\ distribution\ is\ positively\ skewed. \\ \alpha_3 = 0, then\ the\ distribution\ is\ symetric \\ \alpha_3 < 0\ , the\ distribution\ is\ negatively\ skewed \end{cases}$$

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.

Activity:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?
2. In a frequency distribution, the coefficient of skewness based on the quartiles is given to be 0.5. If the sum of the upper and lower quartile is 28 and the median is 11, find the values of the upper and lower quartiles.

1.8.2. Kurtosis

Kurtosis is the degree of peakdness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called leptokurtic. If a curve representing a distribution is flat topped, it is called platykurtic. The normal distribution which is not very high peaked or flat topped is called mesokurtic.

Measures of kurtosis

1. The moment coefficient of kurtosis:
   - Denoted by α_4 and given by $\alpha_4 = \frac{M_4}{{M_2}^2} = \frac{M_4}{\sigma^4}$

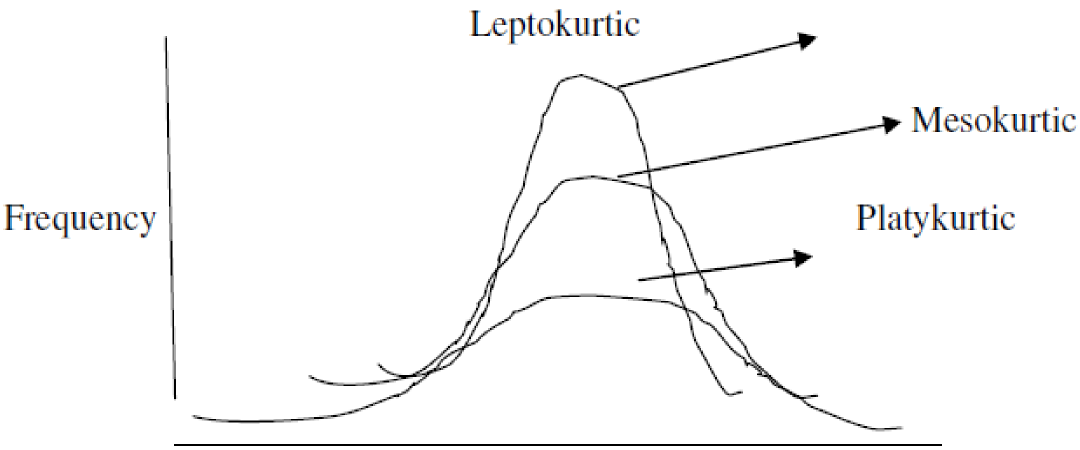
   Where : M_4 is the fourth momnet about the mean

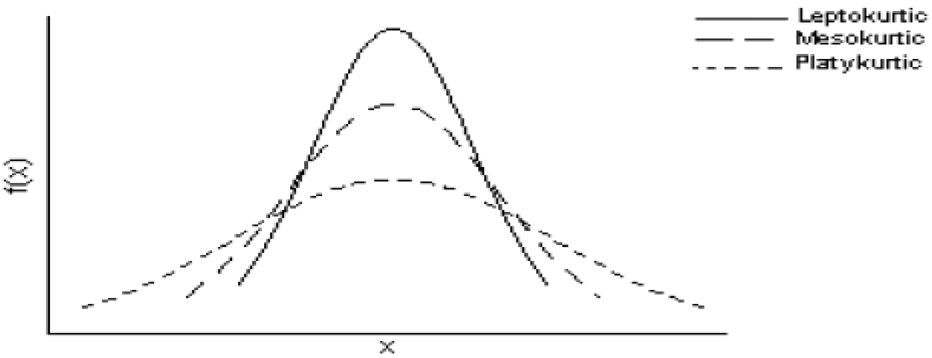
   M_2 is second moment about the mean.

   σ is the population standard deviation

The peakdness of distribution depends on the value of α_4.

$$If \ \ \alpha_4 = \begin{cases} > 3, \text{ then the curve is leptokurtic.} \\ = 3, \text{then the curve is mesokurtic.} \\ < 3, then\ the\ curve\ is\ platykurtic. \end{cases}$$





2. **Percentile coefficient of kurtosis is given by:**

$$P_{ck} = \frac{QD}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}.$$

- It has been shown that $P_{ck} = 0.263$ for a normal distribution.

If $P_{ck} = \begin{cases} < 0.263, & the\ distribution\ is\ leptokurtic\ distribution. \\ > 0.263, & the\ distribution\ is\ paltykurtic\ distribtuion. \end{cases}$

Activity:

1. If the first four moments of a distribution are :

M_1=0, M_2 = 16, M_3 = -60, and M_4= 162

a. Compute a measure of skewness

b. Compute a measure of kurtosis and give your interpretation.

2. The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4th moment about the mean is 243. Compute the Pearsonian coefficient of skewness and identify the type of skewness. Assume (n-1 = n).

Chapter Two

2. Probability and Probability Distribution

2.1. Basic definitions of probability

Knowledge of the properties of theoretical probability distributions is an important part of the decision making process in the various areas of the applied and basic sciences.

In managerial decision making procedure, probability has significant role. Probability and statistical information are used for operational and marketing decisions.

For instance, a time series showing monthly sales is used to track the company's growth and to set future target sales levels. Statistics such as the mean customer order size and the mean number of days a customer takes to make payments help identify the firm's best customers as well as provide benchmarks for handling accounts receivable issues. In addition, data on monthly inventory levels are used in the analysis of operating profits and trends in product sales. Probability analysis helps to determine reasonable and profitable prices for its products.

Managers and business accountant often base their decisions on an analysis of uncertainties such as:-

- What are the chances that sales will decrease if we increase prices?
- What is the likelihood a new assembly method will increase productivity?
- How likely is it that the project will be finished on time?
- What is the chance that a new investment will be profitable?
- What are the chances that women participation will increase if they are educated?
- What is the chance that crime will decrease if we increase society mobilization?

In business organization such questions often raised and hence careful statistical and managerial decisions are needed using probability theory.

2.2.Fundamental concepts of probability

To determine probability, you have to know the definition and concepts of the following terms.

- Experiment
- Sample space and Event
- Outcome and Event
- Equally likely and ME events
- Complement of an event
- Independent and dependent event

Experiment: Any process of observation or measurement which generates well defined outcome. It may be probabilistic/random or non probabilistic experiment.

Probability Experiment: It is an experiment that can be repeated any number of times under similar conditions and it is possible to enumerate the total number of outcomes without predicting an individual outcome. It is also called random experiment. On any single repetition of an experiment, one and only one of the possible experimental outcomes will occur. Several examples of experiments and their associated outcomes follow.

Experiment	**Experimental Outcomes**
Toss a coin	Head, tail
Select a part for inspection	Defective, non defective
Conduct a sales call	Purchase, no purchase
Roll a die	1, 2, 3, 4, 5, 6
Play a football game	Win, lose, tie

Sample Space: Set of all possible outcomes of a probability experiment. It is the set of all experimental outcomes. Sample space can be Countable (finite or infinite) or Uncountable. An experimental outcome is also called a **sample point** to identify it as an element of the sample space.

Why is knowledge of probability necessary for the study of statistical inference? In order to be able to say something about a population on the basis of some sample evidence we must first examine how the sample data are collected. In many cases, the sample is a random one, i.e. the observations making up the sample are chosen at random from the population. If a second sample were selected it would almost certainly be different from the first. Each member of the population has a particular probability of being in the sample (in simple random sampling the probability is the same for all members of the population). To understand sampling procedures, and the implications for statistical inference, we must therefore first examine the theory of probability.

As an illustration of this, suppose we wish to know if a coin is fair, i.e. equally likely to fall heads or tails. The coin is tossed 2 times and 2 heads are recorded. This constitutes a random sample of tosses of the coin. What can we infer about the coin? If it is fair, the probability of getting 2 heads is 1 in 4.

Event: It is a subset of sample space. It is a statement about one or more outcomes of a random experiment. They are denoted by capital letters.

- **Outcome**: The result of a single trial of a random experiment. An experimental outcome is also called a sample point to identify it as an element of the sample space.

Example: suppose you roll a die and let **A** be the event of odd numbers, **B** be the event of even numbers, and **C** be the event of number 8.

Solution: since the die has six sides you have, $A = \{1,3,5\}$, $B = \{2,4,6\}$ and $C = \{\}$.

Remark: If S (sample space) has **n** members then there are exactly $\mathbf{2^n}$ subsets or events.

- **Equally Likely Events**: Events which have the same chance of occurring.
- **Mutually Exclusive Events**: Two events which cannot happen at the same time/have no the same chance of occurrence.
- **Complement of an Event:** The complement of an event A means non-occurrence of A and is denoted by $A'\ or A^c\ or\ \bar{A}$ and contains those points of the sample space which don't belong to A.
- **Elementary Event**: an event having only a single element or sample point.
- **Compound events:** Most practical problems require the calculation of the probability of a set of outcomes rather than just a single one, or the probability of a series of outcomes in separate trials.
- **Intersection:** The intersection of two events say A and B denoted by $A \cap B$ and read as the intersection of A and B or "A and B". The intersection of A and B are all elements which are common both in A and B.
- **Union:** The Union of A and B denoted by $A \cup B$ or "A or B" and read as A union B which is the set of elements either in A or B or in both.
- **Independent Events**: Two events are independent if the occurrence of one does not affect the probability of the other occurring.
- **Dependent Events**: Two events are dependent if the first event affects the outcome or occurrence of the second event in a way the probability is changed.

Example: What is the sample space for the following experiment?

- Roll a die.
- Toss a coin two times.
- A light bulb is manufactured. It is tested for its life length by time.

Probability is a numerical measure of the likelihood that an event will occur. Thus, probabilities can be used as measures of the degree of uncertainty associated with the four events previously listed. If probabilities are available, we can determine the likelihood of each event occurring. Probability values are always assigned on a scale from 0 to 1. A probability near zero indicates an event is unlikely to occur; a probability near 1 indicates an event is almost certain to occur. Other probabilities between 0 and 1 represent degrees of likelihood that an event will occur. For example, if we consider the event "the price of good increase next month," we understand that when the exchange rate report indicates "a near-zero probability of increase in price," it means almost no chance of price change. However, if a 0.90 probability of change in price is reported, we know that the price is likely to increase. A 0.50 probability indicates that the price is just as likely to increase as not. You can consider another example. Consider an event that probability that the rain will rain tomorrow which is uncertain. If the probability is near to zero, it shows that the rain has little/ no chance of raining; while when the probability near to one, it implies higher chance of raining tomorrow.

Approaches of measuring probability

➢ There are four different conceptual approaches to the study of probability theory and computation.

- The classical approach.
- The relative frequency (frequentist) approach.
- The axiomatic approach and the subjective approach.

1. **The classical approach**

This approach is used when

- All outcomes are equally likely and mutually exclusive
- Total number of outcome is finite, say N.

Definition: If a random experiment with N equally likely outcomes is conducted and out of these N_A outcomes are favorable to the event A, then the probability that event A occur denoted P (A) *is* defined as:

$$P(A)=\frac{N(A)}{N(S)}$$ [illegible]

Example:-Consider the following sample space S, or as it is called the universal set, where $S = \{1,2,3,4,5,6,7,8,9,10,11,12\}$. let $A = \{1,2,3,4,5,6\}, B = \{3,4,5,6,7,8,9\}, E = \{2,4,6,8,10,12\}, and\ F = \{3,6,9,12\}, and\ G = \{5,7,11\}$. Furthermore, assume that the elements in S are all equally likely to occur. Find

i. $AUB, and\ P(A \cup B),$
ii. $A \cap B, and\ P(A \cap B),$
iii. $E \cap F, and\ P(E \cap F).$

Solution:-we can see the solution of each as;

i. $AUB = \{1,2,3,4,5,6,7,8,9\}. Hence\ P(AUB) = \frac{9}{12} = 0.75$

ii. $A \cap B = \{3,4,5,6\}, and\ P(A \cap B) = \frac{4}{12} = \frac{1}{3}$.

iii. $E \cap F = \{6,12\}, with\ P(E \cap F) = \frac{2}{12} = \frac{1}{6}$.

2. The Frequencies or a posteriori Approach

This is based on the relative frequencies of outcomes belonging to an event.

Definition: The probability of an event A is the proportion of outcomes favorable to A in the long run when the experiment is repeated under same condition.

$$P(E) = \lim_{n \to \infty} \frac{N}{N}.$$

Example: If the records of Ethiopian Air Lines show that 468 of 600 of its flights from B/Dar to Addis arrived on time, what is the probability that any one of similar flights will arrive on time?

- **Solution**: If E =The event that the plane will arrive on time, then: $P(E) = \frac{m}{n} = \frac{468}{600} = 0.78.$

Exercise: If records show that 60 out of 100,000 bulbs produced are defective. What is the probability of a newly produced bulb to be defective?

3. Axiomatic Approach:

Let E be a random experiment and S be a sample space associated with E. With each event A a real number called the probability of A satisfies the following properties called axioms of probability or postulates of probability.

1. $0 \le P(A) \le 1$
2. $P(s) = 1$
3. $P(\overline{A}) = 1 - P(A)$, where $\overline{A}$ is the complement of event A.
4. If A and B are mutually exclusive events, the probability that one or the other occur equals the sum of the two probabilities. i. e. $P\ (AUB) = P\ (A) + P\ (B)$

Note: For any two events A and B, $P(AUB) = P(A) + P(B) - P(A \cap B)$; this is the general addition rule.

Example: Using Statistics follow-up survey, additional questions were asked of the 300 households that actually purchased big-screen televisions. Table below indicates the consumers' responses to whether the television purchased was a plasma screen and whether they also purchased a DVR (digital video recorder) in the past 12 months.

Purchased Plasma Screen	Purchased DVR		
	Yes	No	Total

Plasmas screen	38	42	80
Not Plasma Screen	70	150	220
Total	108	192	300

A. Find the probability that if a household that purchased a big-screen television is randomly selected, the television purchased is a plasma screen. **Answer**: $\mathrm{P(plasma\ screen)} = \frac{80}{300} = 0.267$.

B. Find the probability that a randomly selected household that purchased a big-screen television also purchased a plasma screen television and a DVR. Answer: $\mathrm{P(plasma\ screen\ and\ DVR)} = 38/300 = 0.127$.

C. Find the probability that among households that purchased a big-screen television, they purchased a plasma-screen television or a DVR.

D. What the probability that they purchased plasma screen neither plasma screen nor DVR?

E. What the probability they purchased DVR?

Activity: A sample of 500 respondents was selected in a large metropolitan area to study consumer behavior. Among the questions asked was Do you enjoy shopping for clothing? Of 240 males, 136 answered yes. Of 260 females, 224 answered yes. Construct a contingency table to evaluate the probabilities. What is the probability that a respondent chosen at random

A. Enjoys shopping for clothing?

B. Is a female or enjoys shopping for clothing?

C. Is a male or a female?

4. Subjective Approach

It is always based on some prior body of knowledge. Hence subjective measures of uncertainty are always conditional on this prior knowledge. The subjective approach accepts unreservedly that different people (even experts) may have vastly different beliefs about the uncertainty of the same event.

Conditional Probability and Independence

Conditional Events: If the occurrence of one event has an effect on the next occurrence of the other event then the two events are conditional or dependant events.

Conditional probability of an event

The conditional probability of an event A given that B has already occurred, denoted by P(A/B).Since A is known to have occurred, it becomes the new sample pace replacing the original sample space.

From this we are led to the definition

$P(A/B) = \frac{p(A \cap B}{P(B)}$, P(B)$\neq 0$ or P(A∩ B)=P(A/B).P(B)

Activity: For a student enrolling at freshman in a certain university, the probability is 0.25 that he/she will get scholarship and 0.75 that he/she will graduate. If the probability is 0.2 that he/she will get scholarship and will also graduate. What is the probability that a student who get a scholarship will graduate?

Probability of Independent Events

The probability of B occurring is not affected by the occurrence or nonoccurrence of A, then we say that A and B are independent events i.e. P (B/A) =P(B). This is equivalent to $P(A\cap B)=P(A)P(B)$

Example 2.10: Given that P (A) = 0.4, P (B) = 0.2, P (A ∩ B) = 0.08,

P(C) = 0.5, P (D) = 0.3, P(C ∩ D) = 0.10.

a) Are A and B independent? b) Are C and D independent?

Solution: a) P (A) P (B) = (0.4) (0.2) = 0.08 = P (A ∩ B).Hence, A and B are independent.

b) P(C) P (D) = (0.5) (0.3) = 0.15 ≠ P(C ∩ D) = 0.10. Hence, C and D are dependent.

Activity: 1. a machine is drawn at random from a box containing 6 copying machines and 4 printing machine. Find the probability that they are drawn in the order copying and printing machine if the first machine is

a) Replaced b) not replaced

2. If the probability that a research project will be well planned is 0.60 and the probability that it will be well planned and well executed is 0.54, what is the probability that it will be well executed given that it is well planned?

2.2. Definition of random Variable and Probability Distribution

Definition and types of Random Variables

Random variable: - is numerical valued function defined on the sample space. It assigns a real number for each element of the sample space. Generally random variables are denoted by capital letters and the values of the random variables are denoted by small letters.

There two types of random variables: **Discrete** and **Continuous** random variable.

Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted.

Examples:

- number of customers in restaurant per day
- Toss a coin n time and count the number of heads.

- Number of children in a family.
- Number of car accidents per week.
- Number of defective items in a given company.
- Number of consumers for sales.

Continuous random variable: are variables that can assume all values between any two give values or intervals.

Examples

- Time between customer arrivals in minutes
- Percentage of project complete after
- Height of students at certain college.
- Life time of light bulbs.
- Length of time required to complete a given training.

Probability distribution: - consists of a value a random variable can assume and the corresponding probabilities of the values or it is a function that assigns probability for each element of random variable. Probability distribution can be discrete or continuous.

Discrete probability distribution: - is a formula, a table, a graph or other devices used to specify all possible values of the discrete random variable (R.V) X along with their respective probabilities.

Properties of discrete probability distribution

1) $\sum_{i=1}^{n} P(X = x_i) = 1$

2) $P(X = x_i) \geq 0 \; or \; 0 \leq P(X = x_i) \leq 1$

3) If X is discrete random variable, then

$$P(a < X < b) = \sum_{X=a+1}^{b-1} P(x) \neq P(a \leq X < b) = \sum_{X=a}^{b-1} P(x) \neq P(a < X \leq b) = \sum_{X=a+1}^{b} P(x) \neq P(a \leq X \leq b) = \sum_{X=a}^{b} P(x)$$

Activity: The daily exchange rate of one dollar in euros during the first three months of 2007 can be inferred to have the following distribution.

X	0.73	0.74	0.75	0.76	0.77	0.78
P(x)	0.05	0.10	0.25	0.40	0.15	0.05

a. Show that P(x) is a probability distribution.

b. What is the probability that the exchange rate on a given day during this period will be at least 0.75?

c. What is the probability that the exchange rate on a given day during this period will be less than 0.77?

d. If daily exchange rates are independent of one another, what is the probability that for two days in a row the exchange rate will be above 0.75?

2.3. Continuous probability distribution

Definition: let X be a continuous random variable with a non negative function f(x) called probability density function (Pdf). Then f(x) satisfies the following legitimates:

i. $f(x) \geq 0, for\ all\ x\ i.e. f(x)\ should\ be\ non-negative\ function.$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1$. I.e. total area under the normal cure equals to one.

Properties of continuous probability distribution

The total area under the curve is one i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$

$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx =$ The area under the curve between the point a and b.

$P(X = a) = 0$

Note: $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$

Example 2.15: Suppose that the r-v X is continuous with the following pdf:

$$f(x) = \begin{cases} 2x, o < x < 1, \\ 0, otherwise \end{cases}$$

a) Check that $f(x)$ satisfies the two conditions of being a p.d.f.;

b) Evaluate P(X<0.5) and $P(0.8 < X < 1.2)$.

Solution: a) Obviously, for 0 < X< 1, f(x) >0, and

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} f(x)dx = \int_{0}^{1} 2xdx = x^2\Big|_{0}^{1} = 1.$$

Hence, $f(x)$ is the pdf of some random variable X.

Note: $\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} f(x)dx,$ since f(x) is zero in the other two intervals: $(-\infty, 0] \cup [1, \infty)$.

b) $P(X < 0.5) = \int_{0}^{0.5} f(x)dx = \int_{0}^{0.5} 2xdx = x^2\Big|_{0}^{0.5} = 0.25.$

Properties of continuous probability distribution

- The total area under the curve is one i.e. $\int_{-\infty}^{\infty} f(x) = 1$

- $P(a \leq X \leq b) =$ the area under the curve between the point a and b.
- $P(X) \geq 0$
- $P(X = a) = 0$

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

Introduction to expectation

1. Let a discrete random variable X assume the values X1, X2, ….,Xn with the probabilities P(X_1), P(X_2), ….,P(Xn) respectively. Then the expected value of X, denoted as E(X) is defined as: E(X) =X_1.P (X_1) +X_2.P(X_2) +…. +X_n.P (X_n)

➢ $E(x) = \sum_{i=1}^{n} X_i.P(X_i).$

2. Let X be a continuous random variable assuming the values in the interval (a, b) such that $\int_a^b f(x)d(x) = 1, then\ \ E(X) = \int_a^b X.f(x)d(x)$

Mean and Variance of a random variable

Let X is a random variable.

1. The expected value of X is its mean

Mean of X=E(X)

2. The variance of X is given by:

Variance of X=Var(x) = $E(X - E(X))^2 = E(X^2) - (E(X))^2$

Where $E(X^2) = \sum_{i=1}^{n} X_i^2.P(X_i) \quad If\ \ X\ \ is\ \ discrete$

$= \int_x X^2 f(x)d(x) \quad if\ \ X\ \ is\ \ continuous$

Rules of Expectation

1) Let X be a R.V and k be a real number, then

a) E (kX) =kE(X) $\rightarrow \operatorname{var}(kX) = k^2.\operatorname{var}(X)$

b) E(X±k) =E(X) ± k $\rightarrow \operatorname{var}(X \pm k) = \operatorname{var}(X)$

2) Let X and Y be R.V on the sample space, then

a) $E(X \pm Y) = E(X) \pm E(Y)$

b) $\operatorname{var}(X \pm Y) = \operatorname{var}(X) + \operatorname{var}(Y) \pm 2.\operatorname{cov}(X,Y)$

Where Cov(X, Y) =the covariance between X and Y=E (XY) - E(X).E(Y)

3) Let X and Y be independent R.V, then

a) E (XY) =E(X).E(Y)

b) $\operatorname{var}(X \pm Y) = \operatorname{var}(X) + \operatorname{var}(Y)$

c) Cov(X, Y) =0

Example: Using historical records, the personnel manager of a plant has determined the probability distribution of X, the number of employees absent per day. It is

X	0	1	2	3	4	5	6	7
P(x)	0.005	0.025	0.31	0.34	0.22	0.08	0.019	0.001

Find the following:

A. $P(2 \leq X \leq 5)$

B. $P(X > 5)$

C. $P(x < 4)$

D. The expected number of employee absent per day.

E. Variance.

Solution: - Since the variable X is discrete,

a. $P(2 \leq X \leq 5) = \sum_{I=2}^{5} P(X = x_i) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 0.31 + 0.34 + 0.22 + 0.08 = 0.95$. This is the probability that the number of employee absent per day will be between 2 and 5 employee including end points.

b. $P(X > 5) = P(x = 6) + P(X = 7) = 0.019 + 0.001 = 0.02.$

D. $E(X) = \sum_{I=0}^{7} X_i P(X = x_i) = 0x0.005 + 1x0.025 + 2x0.31 + \cdots + 7x0.001 = 3.066 \approx 3$

Activity:

1. The manager of a bookstore recorded the number of customers who arrive at a checkout counter every 5 minutes from which the following distribution was calculated. Calculate the mean and standard deviation of the random variable.

X	0	1	2	3	4
P(x)	0.1	0.2	0.25	0.25	0.2

2. Suppose the two largest cable providers in Ethiopia are Euro cable, with 21.5 million subscribers, and Elsewedy, with 11.0 million subscribers. Suppose that the management of Elsewedy subjectively assesses a probability distribution for the number of new subscribers in the next year of Ethiopia as follows.

X	**P(x)**

100,000	**0.1**
200,000	**0.2**
300,000	**0.25**
400,000	**0.3**
500,000	**0.1**
600,000	**0.05**

a. Is this probability distribution valid? Explain.

b. What is the probability Time Warner will obtain more than 400,000 new subscribers?

c. What is the probability Time Warner will obtain fewer than 200,000 new subscribers?

3. To examine the effectiveness of its four annual advertising promotions, a mail-order company has sent a questionnaire to each of its customers, asking how many of the previous year's promotions prompted orders that would not otherwise have been made. The table lists the probabilities that were derived from the questionnaire, where *X* is the random variable representing the number of promotions that prompted orders. If we assume that overall customer behavior next year will be the same as last year, what is the expected number of promotions that each customer will take advantage of next year by ordering goods that otherwise would not be purchased?

x	**0**	**1**	**2**	**3**	**4**
P(x)	**0.1**	**0.25**	**0.40**	**0.20**	**0.05**

4. Let X be a continuous R.V with distribution

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \le x \le 2 \\ 0, & otherwise \end{cases}$$

Then find a) P (1<x<1.5; b) E(x); c) Var(x); and d) $E(3x^2 - 2x)$.

5. Delta Airlines quotes a flight time of 2 hours, 5 minutes for its flights from Cincinnati to Tampa. Suppose we believe that actual flight times are uniformly distributed between 2 hours and 2 hours, 20 minutes.

A. What is the probability that the flight will be no more than 5 minutes late?

B. What is the probability that the flight will be more than 10 minutes late?

C. What is the expected flight time?

2.3.Common Discrete Probability Distributions

In this section, we shall study two common discrete probability distributions, namely, the Binomial and Poisson distributions.

Binomial Distribution: A binomial experiment is a probability experiment that satisfies the following four requirements called assumptions of a binomial distribution.

1. The experiment consists of n identical trials.
2. Each trial has only one of the two possible mutually exclusive outcomes, success or a failure.
3. The probability of each outcome does not change from trial to trial, and
4. The trials are independent, thus we must sample with replacement.

Examples of binomial experiments

- Tossing a coin 20 times to see how many tails occur.
- Asking 200 people if they watch BBC news.
- Registering a newly produced product as defective or non-defective.
- Asking 100 people if they favor the ruling party.
- Rolling a die to see if a 5 appears.

Definition: The outcomes of the binomial experiment and the corresponding probabilities of these outcomes are called Binomial Distribution.

Let p=probability of success q= 1-p=probability of failure on any given trials

Then the probability getting x success in n trials becomes

$$P(X = x) = \begin{cases} \binom{n}{x} . p^x q^{n-x} & x = 0,1,3,....n \\ 0 \quad otherwise \end{cases}$$

And this sometimes written as

$X \sim Bin(n, p)$

When using the binomial formula to solve problems, we have to identify three things:

- The number of trials (n)
- The probability of a success on any one trial (P) and
- The number of successes desired (X).

Remark: If X is a binomial random variable with parameters n and p then

$E(X) = np$ and $var(X) = npq$

Example: If the probability is 0.20 that people traveling on Ethiopian airline flight will a business man, find the probability that 3 of 10 people on such flight will be a Business man?

Solution: Let X be the number of vegetarians. Given n = 10, p = 0.20, k = 3; then,

$$P(X = 3) = \binom{10}{3}(0.2)^3(0.8)^7 = 0.201.$$

Activity:

2. Suppose that an examination consists of six true and false questions, and assume that a student has no knowledge of the subject matter. The probability that the student will guess the correct answer to the first question is 30%. Likewise, the probability of guessing each of the remaining questions correctly is also 30%.

a) What is the probability of getting more than three correct answers?

b) What is the probability of getting at least two correct answers?

c) What is the probability of getting at most three correct answers?

d) What is the probability of getting less than five correct answers?

3. According to the last census, 45% of working women held full-time jobs in 2010. If a random sample of 50 working women is drawn, what is the probability that 2 or more hold full-time jobs?

Poisson distribution

Another useful discrete probability distribution is the **Poisson distribution**, named after its French creator. Like the binomial random variable, the **Poisson random variable** is the number of occurrences of events, which we'll continue to call *successes*. The difference between the two random variables is that a binomial random variable is the number of successes in a set number of trials, whereas a Poisson random variable is the number of successes in an interval of time or specific region of space. It is used as a distribution of rare events, such as:

- Natural disasters like earth quake.
- Accidents. The number of accidents in 1 day on a particular stretch of highway.
- Arrivals. The number of costumers arriving at a service station in 1 hour.
- Number of misprints pages.
- Number of claims per day in an insurance company

A random variable X is said to have a Poisson distribution if its probability distribution is given by:

$$P(X = x) = \begin{cases} \dfrac{\lambda^x . e^{-\lambda}}{x!} & x = 0,1,2..... \\ 0 & otherwise \end{cases}$$

Where: λ is the average number occurrence of an event in the unit length of interval or distance and x is the number of occurrence in a Poisson process.

The Poisson distribution depends only on the average number of occurrences per unit time of space.

The process that gives rise to such events is called Poisson process.

Note: - If X is a Poisson random variable with parameters λ then $E(x) = \lambda$ and $var(x) = \lambda$.

Example:

1**.** Suppose that customers enter a waiting line at random at a rate of 4 per minute. Assuming that the number entering the line during a given time interval has a Poisson distribution, find the probability that:

a) One customer enters during a given one-minute interval of time;

b) At least one customer enters during a given half-minute time interval.

Solution:

a) Given $\lambda = 4$ per min,

$$P(x=1) = \frac{4^1 e^{-4}}{1!} = 4e^{-4} = 0.0733.$$

b) Per half-minute, the expected number of customers is 2, which is a new parameter.

$P(X \geq 1) = 1 - P(X = 0)$, but $P(X = 0) = e^{-2} = 0.1353.$

$\therefore P(X \geq 1) = 1 - 0.1353 = 0.8647.$

1. If there are 500 customers per eight-hour day in a checkout line, what is the probability that there will be exactly 3 in line during any five-minute period?

Solution: The expected value during any one five minute period would be 500 / 96 = 5.2083333. The 96 is because there are 96 five-minute periods in eight hours. So, you expect about 5.2 customers in 5 minutes and want to know the probability of getting exactly 3.

$$P(x=3, \lambda = \frac{500}{96}) = \frac{\left(e^{-\frac{500}{96}}\right)\left(\frac{500}{96}\right)^3}{3!} = 0.1288$$

Activity

1. The number of bank robberies that occur in a large city is Poisson distributed with a mean of 2 per day. Find the probabilities of the following events.

A. Three or more bank robberies in a day

B. Between 10 and 15 (inclusive) robberies during a 5-day period

2. If 2 accidents can be expected an intersection on any given day, what is the probability that there will be 3 accidents on any given day?

3. A sale firm receives, on the average, 3 calls per hour on its toll-free number. For any given hour, find the probability that it will receive the following.

   A. At most 3 calls

   B. At least 3 calls

   C. Five or more calls

4. A ban manager wants to provide prompt service for customers at the bank drive up window. The bank currently can serve up to 10 customers per 15- minute's period without significant delay. The average number arrival rate is 7 customers per 15 minutes period. Assuming that X has a Poisson distribution, find the probability that 10 customers will arrive in a particular 15 minutes period?

5. If approximately 2% of the employers of the company are left-handed, find the probability that in a company of 200 employers, there are exactly 5 people who are left-handed?

6. The number of users of an automatic banking machine is Poisson distributed. The mean number of users per 5-minute interval is 1.5. Find the probability of the following events.

   A. No users in the next 5 minutes

   B. Five or fewer users in the next 15 minutes

   C. Three or more users in the next 10 minutes

2.4.Common Continuous Probability Distributions

Normal Distribution: A random variable X is said to have a normal distribution if its probability density function is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad where\ -\infty<x<\infty,\ -\infty<\mu<\infty,\quad \sigma>0$$

$\mu=E(x)$ *and* $\sigma^2=\text{var}\,iance(x)$ -are parameters of the normal distribution.

Properties of Normal Distribution:

1. It is bell shaped and is symmetrical about its mean

2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.

3. It is completely described by two parameters: mean and standard deviation.

4. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the mean is 0.5 $\Rightarrow \int_{-\infty}^{\infty} f(x)d(x) = 1$

5. It is uni-modal, i.e., values mound up only in the center of the curve.

6. Median=Mean=mode =μ and located at the center of the distribution.

7. The probability that a random variable will have a value between any two points is equal to the area under the curve between those points.

- To calculate the probability that a normal random variable falls into any interval, we must compute the area in the interval under the curve. Unfortunately, the function is not as simple as the uniform precluding the use of simple mathematics or even integral calculus. Instead we will resort to using a probability tables through standardization i.e. using standard normal distribution table.

Note: The following distribution known as the standard normal distribution was derived by using the transformation;

$$Z = \frac{X-\mu}{\sigma} \;\Rightarrow\; f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}.$$

That is, $if \;\; X \sim N(\mu, \sigma^2) \; then \;\; Z \sim (0,1)$.

Properties of the Standard Normal Distribution:

Same as a normal distribution, but

- Mean is zero
- Variance and Standard Deviation is one

- Areas under the standard normal distribution curve have been tabulated in various ways. The most common ones are the areas between Z=0 and a positive value of Z.

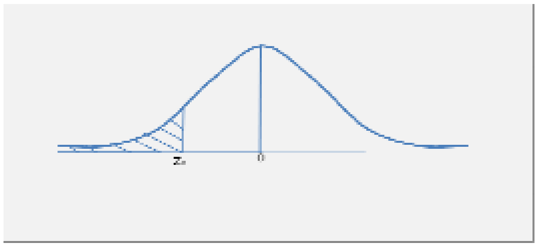
- Given a normal distributed random variable X with Mean μ and standard deviation σ

$$\boldsymbol{P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)}$$

$$\boldsymbol{\Rightarrow P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)}$$

i. Let Z_0 be negative number, then $P(Z < z_0) = P(Z < 0) - P(z_0 < Z < 0)$

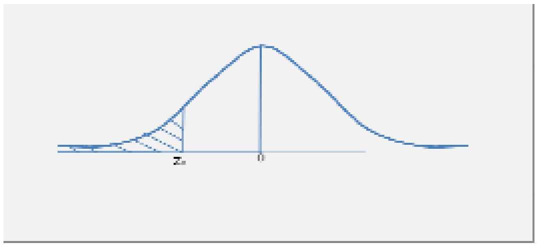
$= 0.5 - P(z_0 < Z < 0)$.



ii. If Z0 is positive real number, then $P(Z > z_0) = 0.5 - P(0 < Z < z_0)$.

iii. Let Z_0 be negative number, then $P(Z < z_0) = P(Z < 0) - P(z_0 < Z < 0)$

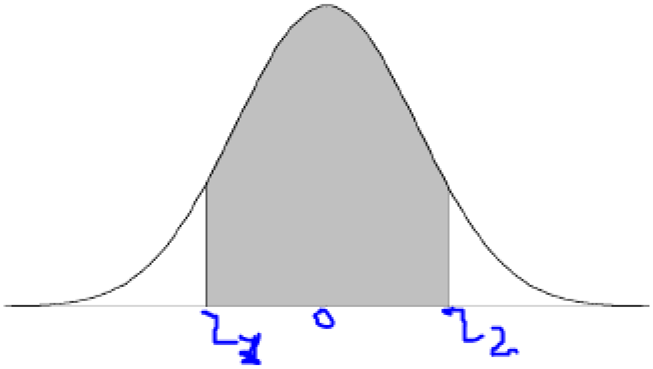
$= 0.5 - P(z_0 < Z < 0)$.



iv. If Z0 is positive real number, then $P(Z > z_0) = 0.5 - P(0 < Z < z_0)$.

v. Let Z1 be a negative number and Z2 be positive real number, then

$$P(z_1 < Z < z_2) = P(z_1 < Z < 0) + P(0 < Z < z_2)$$



Example: Calculate probability using Z-table.

1. Find the probabilities that a r-v having the standard normal distribution will take on a value

a) Less than 1.72

b) less than -0.88

c) Between 1.30 & 1.75

d) Between -0.25 & 0.45.

Solution: Making use of the Z table, we find that

a) P (Z<1.72)=P(Z<0)+P(0<Z<1.72)=0.5+0.4573=0.9573.

b) P (Z < -0.88) = P(Z > 0.88) =0.5 - P(0 < Z < 0.88) =0.5- 0.3106 = 0.1894.

c) P (1.30 < Z <1.75)= P(0 < Z < 1.75) – P(0 < Z < 1.30) = 0.4599 – 0.4032)=0.0567.

d) P (-0.25 < Z < 0.45)= P(-0.25 < Z < 0) + P(0 < Z < 0.45) = 0.0987 + 0.1736=0.2723.

2. Consider an investment whose return is normally distributed with a mean of 10% and a standard deviation of 5%.

a. Determine the probability of losing money.

b. Find the probability of losing money when the standard deviation is equal to 10%.

Solution:

a. The investment loses money when the return is negative. Thus, we wish to determine

$P(X < 0)$

The first step is to standardize both X and 0 in the probability statement: $P(x < 0) = P\left(\frac{X-\mu}{\sigma} < \frac{0-\mu}{\sigma}\right) = P\left(Z < \frac{0-10}{5}\right) = P(Z < -2) = 0.0228.$

Therefore, the probability of losing money is .0228.

b. $P(x < 0) = P\left(\frac{X-\mu}{\sigma} < \frac{0-\mu}{\sigma}\right) = P\left(Z < \frac{0-10}{10}\right) = P(Z < -1) = 0.1587.$

- As you can see, increasing the standard deviation increases the probability of losing money. Note that increasing the standard deviation will also increase the probability that the return will exceed some relatively large amount. However, because investors tend to be risk averse, we emphasize the increased probability of negative returns when discussing the effect of increasing the standard deviation.

Activity: Find the following:

1. a) P (-0.45 < Z < -0.25) b) P (Z>1.75).

2. Let Z be the standard normal random variable. Calculate the following probabilities using the standard normal distribution table:

b)	P (0<Z<1.2)	f)	P (Z≤0)
c)	P (Z<-1.43)	g)	P (Z>1.52)
d)	P (0<Z<1.43)	h)	P (-1.2<Z<0)
e)	P (-1.43<Z<1.2)	i)	P (Z<-1.52)

3. Find the following values of z* of a standard normal random variable based on the given probability values:

a) P (Z > z*) =0.1446 b) P (Z>z*) = 0.8554

4. The weekly incomes of a large group of middle managers are normally distributed with a mean of 1000 birr and standard deviation of 100 birr, and then find the probability that the income of managers will be

A. Less than 1100 birr? B. Between 900 birr and 1250 birr?

5. Under the system of floating exchange rates, the rate of foreign money to the U.S. dollar is affected by many random factors, and this leads to the assumption of a normal distribution of small daily fluctuations. The rate of U.S. dollar per euro is believed in April 2016 to have a mean of 1.36 and a standard deviation of 0.03.

a) Find the probability that tomorrow's rate will be above 1.42.

b) Find the probability that tomorrow's rate will be below 1.35.

6. The long-distance calls made by the employees of a company are normally distributed with a mean of 6.3 minutes and a standard deviation of 2.2 minutes. Find the probability that a call

a) Lasts between 5 and 10 minutes. b) Lasts more than 7 minutes.

c) Lasts less than 4 minutes.

7. Because of the relatively high interest rates, most consumers attempt to pay off their credit card bills promptly. However, this is not always possible. An analysis of the amount of interest paid monthly by a bank's Visa cardholders reveals that the amount is normally distributed with a mean of $27 and a standard deviation of $7.

a) What proportion of the bank's Visa cardholder pay more than $30 in interest?

b) What proportion of the bank's Visa cardholder pay more than $40 in interest?

c) What proportion of the bank's Visa cardholder pay less than $15 in interest?

d) What interest payment is exceeded by only 20% of the bank's Visa cardholders?

8. Assume that the time taken to produce a computer in a given company is normally distributed with average time $\mu = 200$ minutes and $\sigma = 20$ minutes. Determine the probability that the time taken to produce a computer is:

a) More than 230 minutes

b) Less than 230 minutes

Chapter Three

Sampling & Sampling Distributions

3.1. Definition and Some Basic Terms

The probability studied in chapter two is important to discuss inferential statistics. We are going to analyze and interpret data to draw conclusions not about the data but about the source of the data (population consisting of all elements being studied). We collect a sample of data from the population and use it to make inferences about the population. Very often we will be interested in estimating a population parameter. In order to estimate this we need to define our terms carefully:

Population: the entire group of individuals or objects of interest under investigation or study.

Unit: An element of the population. This will be a person or object on which observations can be made or from which information can be obtained.

Sampling Frame: The list of all the units in the population.

Target population: the population about which one wishes to make an inference.

Sample size: - the number of individuals in the sample.

Sampling: - It is the process of selecting a sample from the population.

Major reasons why sampling is necessary

1) the destructive nature of certain tests/studies
2) physical impossibility of checking all items in the population/ infinite population
3) cost of studying all items in the population is often prohibitive/ sampling reduce cost/
4) The adequacy of sample result. Sampling has greater adequacy and accuracy.
5) In terms of time/ sampling has greater speed/

3.1.1. Types of Errors

An estimate based on a sample will not be exact; there will be an error involved. In general, errors which occur during estimation based on a sample can be categorized into two:

- Sampling errors
- Non sampling errors

Sampling errors: It is the discrepancy between population parameter and the sample statistic. The error which arise due to only a sample being used (technique and sample size) to estimate population parameter. Even if we have a representative sample will also introduce errors if the sample size is small. On the other hand, our estimates of parameters will often be inaccurate if our sample is not representative of the population. Because of this we need to know how to choose a sample. Sampling error is the difference b/n an estimate and the true value of the parameter being evaluated.

3.1.2. Non sampling errors

Suppose we have a representative sample and have chosen a sample large enough to ensure our parameter estimates are accurate to a good degree of precision, errors may still arrive such as measurement errors, recording errors, non-response errors, respondent bias, interviewer error, errors in processing the data, and reporting error. Measurement errors and recording errors occur if there is an error in measuring the item being studied or in recording its result. Interviewer errors can occur in surveys when an interviewer introduces bias into an interview or when a questionnaire is badly designed. Another common form of error is the non-response error. Non responses can be due to refusals.

Sampling Methods: - Sampling techniques can be grouped into two categories:

- **Random (probability)** sampling methods, and
- **Non-random (non-probability)** sampling methods.

3.2. Random (probability) sampling methods

Random sampling: sampling method in which the items are included in the sample in a random basis.

Simple random sample: a sampling technique in which member of the population is equally likely to be included in the sample. It might be done in different ways.

Lottery method – the units to be included in the sample are chosen by a lottery. Assign numbers to each element in the population. Write each number in a split of paper, toss then draw one number at a time. This method can only be used if the population is not very large otherwise it is cumbersome.

Table of random number: used to select representative sample from a large size population. To select the sample use random digit techniques. We proceed with the following steps.

Step 1: each element numbered for example for a population of size 500 we assign 001 to 500.

Step 2: select a random starting point

Step 3: we need only respective number of digits. Proceed in this fashion until the required number of sample selected.

Suppose that a sample of 6 study centers is to be selected at random from a serially numbered population of 60 study centers. The following table is portion of a random numbers table used to select a sample.

Row>	1	2	3	4	5	–––	n

Column∨							
1	2315	7548	5901	8372	5993	– – –	6744
2	0554	5550	4310	5374	3508	– – –	1343
3	1487	1603	5032	4043	6223	– – –	0834
4	3897	6749	5094	0517	5853	– – –	1695
5	9731	2617	1899	7553	0870	– – –	0510
6	1174	2693	8144	3393	0862	– – –	6850
7	4336	1288	5911	0164	5623	– – –	4036
8	9380	6204	7833	2680	4491	– – –	2571
9	4954	0131	8108	4298	4187	– – –	9527
10	3676	8726	3337	9482	1569	– – –	3880
11	– – –	– – –	– – –	– – –	– – –	– – –	– – –
12	– – –	– – –	– – –	– – –	– – –	– – –	– – –
13	– – –	– – –	– – –	– – –	– – –	– – –	– – –
N	3914	5218	3587	4855	4888	– – –	8042

If you start in the first row and first column, centers numbered 23, 05, 14,..., will be selected. However, centers numbered above the population size (60) will not be included in the sample. In addition, if any number is repeated in the table, it may be substituted by the next number from the same column. Besides, you can start at any point in the table. If you chose column 4 and row 1, the number to start with is 83. In this way you can select first 6 numbers from this column starting with 83.

~~**83**~~	~~**75**~~
53	**33**
40	**01**
05	**26**

The sample, then, is as follows: Hence, the study centers numbered 53, 40,05,33,01 and 26 will be in the sample.

Stratified random sampling: is often used when the population is split into subgroups or "strata". The different subgroups are believed to be very different from each other, but it is thought that the individuals who make up each subgroup are similar. The number of units to be chosen from each sub-group is fixed in advance and the units are chosen by simple random sampling within the sub group.

Example: An investigator is interested in securing a particular response that would be representative of under graduate college student. He might stratify the population into four groups: freshman, sophomore, junior and senior and then select from each group.

Cluster sampling: in some case the identification and location of an ultimate unit for sampling may require considerable time and cost in such cases cluster sampling is used. In cluster sampling the population is subdivided into groups or clusters and a probability of these clusters is then drawn and studied. Clusters may be Region, Zones, Weredas, Kebeles etc. This method of sampling has less cost, faster and more convenient but it may not be very efficient and representative due to the usual tendency of the units in different cluster be similar.

Example: If we want to study the auditing habit of families in Ethiopia which is divided in to Regions and Zones. We shall first draw a random sample from the Zones to be studied and then from these selected Zones or clusters, we draw random sample of households for the purpose of investigation.

Systematic sampling: the items or individuals of the population are arranged in some way alphabetically, in file drawer by data received or some other method. A random starting point is selected and then every Kth member of the population is selected for the sample. For example if we want select n items from the population of size N using systematic sampling, we divide N by n $(N/n = K)$ and choose one b/n 1 and K then we take every Kth member. So the samples will be $i, i+K, i+ 2K, i+ 3K$, etc. where 0< i < K. For instance, to study the average monthly expenditure of households in a city, you may randomly select every fourth households from the household listings.

Example: Suppose we want to choose a sample of about 20 students out of a class of 100 students. First we put the class in order (may be alphabetical order, or by ID number) and give each a number between 1 and 100. Next we divide 100 by 20 and we get 100/20 = 5. We now choose a number at random between 1 and 5. The student corresponding to that number is the first student in the sample, and we then take every 5th student. So if, for example, we choose the number 2 the sample will consist of the 2nd, 7th, 12th, 17th, ..., 92nd and 97th students on the list.

Non probability sampling: In non-probability sampling, the sample is not based on chance. It is rather determined by personal judgment of the researcher. This method is cost effective; however, we cannot make objective statistical inferences. Depending on the technique used, non-probability samples are classified into quota, judgment or purposive and convenience samples.

Judgment sampling: the subjective judgment of the researcher is the basis for selecting items to be included in a sample. Judgment sampling often used to pre-test the questionnaire.

Quota sampling: In this sampling technique major population characteristics play an important role in selection of the sample. It has some aspects in common with stratified sampling, but has no randomization. Here the population may be divided in two groups like stratified sampling to give quota and select from each group.

Convenient sampling: this technique of selecting sample which is simply convenient to the researcher in terms of time, money and administration.

3.3. Sampling Distribution of the Mean and Proportion

3.3.1 Sampling Distribution of the Mean

Suppose we have a simple random sample of size *n*, picked up from a population of size *N*. We take measurements on each sample member in the characteristic of our interest and denote the observation as x_1, x_2, …,x_n, respectively. The sample mean for this sample is defined as:

$$\bar{X} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

If we pick up another sample of size *n* from the same population, we might end up with a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance. In other words, ***the sample mean is a random variable.***

The possible values of this random variable depends on the possible values of the elements in the random sample from which sample mean is to be computed. The random sample, in turn, depends on the distribution of the population from which it is drawn. As a random variable, $\bar{X}$ has a **probability distribution**. This probability distribution is the sampling distribution *of* $\bar{X}$.

The sampling distribution of $\bar{X}$ is the probability distribution of all possible values the random variable $\bar{X}$ may take when a sample of size n is taken from a specified population.

There are commonly three properties of interest of a given sampling distribution.

- It's Mean,
- Its Variance,
- Its Functional form.

When sampling without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on.

In other words, the *n* random variables representing the *n* sample members do not remain independent, the expression for the variance of $\bar{X}$ changes. The results in this case will be:

$\mu_{\bar{X}} = E(\bar{X}) = \mu$ and $\sigma^2{}_{\bar{X}} = var(\bar{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$. By comparing these expression with the ones derived above, we find that the variance of $\bar{X}$ is the same but further multiplied by a factor $\frac{N-n}{N-1}$. This factor is therefore known as the finite population multiplier or the correction factor.

Remark:

1. In general if sampling is with replacement, then $Var(\bar{X}) = \sigma^2{}_{\bar{X}} = \frac{\sigma^2}{n}$
2. If sampling is without replacement, then $\sigma^2{}_{\bar{X}} = var(\bar{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$.
3. In any case the sample mean is unbiased estimator of the population mean. i.e. $\mu_{\bar{X}} = \mu \gg E(\bar{X}) = \mu$.

➢ When sampling is from a normally distributed population, the distribution of $\bar{X}$ will also be normal. i.e. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Central limit theorem: given the population of any functional form with mean $\boldsymbol{\mu}$ and finite variance σ^2, the sampling distribution of $\bar{X}$ computed from the sample size of n from the population will be approximately normally distributed with mean $\boldsymbol{\mu}$ and variance $\frac{\boldsymbol{\sigma^2}}{\mathbf{n}}$, when the sample size will be approximately large.

3.3.2. Sampling Distribution of The Proportion

Let us assume we have a binomial population, with a proportion p of the population possesses a particular attribute that is of interest to us. This also implies that a proportion $q = 1 - p$ of the population does not possess the attribute of interest. If we pick up a sample of size $\boldsymbol{n}$ with replacement and found **x** successes in the sample, the sample proportion of success $\widehat{\boldsymbol{P}}$ is given by $\hat{P} = \frac{x}{n}$. **X** is a binomial random variable, the possible value of this random variable depends on the composition of the random sample from which $\hat{P}$ is computed. The probability of **x** successes in the sample of size **n** is given by a binomial probability distribution, with $P(x) = nCx\ P^x q^{n-x}$. Since $\hat{P} = \frac{x}{n}$ and **n** is fixed (determined before the sampling) the distribution of the number of successes (*x*) leads to the distribution of $\hat{P}$.

The sampling distribution of $\hat{P}$ is the probability distribution of all possible values the random variable $\hat{P}$ may take when a sample of size **n** is taken from a specified population.

The expected value and the variance of *x i.e.* number of successes in a sample of size *n* is known to be: *E(x) = n p, Var (x) = n p q.* Finally we have mean and variance of the sampling distribution of $\hat{P}$, $\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$ and variance of $\hat{P}$ is $\sigma^2{}_{\hat{P}} = var(\hat{P}) = var\left(\frac{x}{n}\right) = \frac{pq}{n}$.

When sampling is without replacement, we can use the finite population correction factor, so sampling distribution of $\hat{P}$ has $\mu_{\hat{P}} = p$ and $\sigma^2{}_{\hat{P}} = \frac{pq}{n}.\left(\frac{N-n}{N-1}\right)$,

- When sampling is done from a population with proportion **p**, the sampling distribution of the sample proportion $\hat{P}$ approaches to a normal distribution with proportion p and variance $\frac{pq}{n}$ as the sample size n increases.

CHAPTER FOUR

4. ESTIMATION

This chapter assumes information about the population, such as the mean, the standard deviation, or the shape of the population. In most business situations, such information is not available. In fact, the purpose of sampling may be to estimate some of these values. For example, you select a sample from a population and use the mean of the sample to estimate the mean of the population. This chapter considers several important aspects of sampling. We begin by studying point estimates. A point estimate is a particular value used to estimate population value. For example, suppose we select a sample of 50 junior executives and ask each the number of hours they worked last week. Compute the mean of this sample of 50 and use the value of the sample mean as a point estimate of the unknown population mean. However, a point estimate is a single value. A more informative approach is to present a range of values in which we expect the population parameter to occur. Such a range of values is called a confidence interval.

Frequently in business we need to determine the size of a sample. How many voters should a polling organization contact to forecast the election outcome? How many products do we need to examine to ensure our quality level? This chapter also develops a strategy for determining the appropriate size of the sample.

The objective of estimation is to determine the approximate value of a population parameter on the basis of a sample statistic. For example, the sample mean is employed to estimate the population mean. We refer to the sample mean as the *estimator* of the population mean. Once the sample mean has been computed, its value is called the estimate.

4.1. Point and Interval Estimators

☞ **Point Estimator**

A point estimate is a single statistic used to estimate a population parameter. Suppose Best Buy, Inc. wants to estimate the mean age of buyers of high-definition televisions. They select a random sample of 50 recent purchasers, determine the age of each purchaser, and compute the mean age of the buyers in the sample. The mean of this sample is a point estimate of the mean of the population. Generally, *Point estimate* is the statistic, computed from sample information, which is used to estimate the population parameter.

There are three drawbacks to using point estimators.

- It is virtually certain that the estimate will be wrong. (The probability that a continuous random variable will equal a specific value is 0; that is, the probability that x will exactly equal μ is 0.)
- Often need to know how close the estimator is to the parameter.
- In drawing inferences about a population, it is intuitively reasonable to expect that a large sample will produce more accurate results because it contains more information than a smaller sample does. But point estimators don't have the capacity to reflect the effects of larger sample sizes. As a consequence, we use the second method of estimating a population parameter, the *interval estimator*.

4.1.1. Methods of Estimation

Let us outline the procedures by which we can find the point estimators of a parameter. The procedures to be used here are: (1) the method of moments, and (2) the maximum likelihood method.

A) The Method of Moments

Let $x_1\ x_2, \dots, x_n$ be a random sample of a random variable X. The average value of the kth power of $x_1\ x_2, \dots, x_n$.

$$M_k = \sum_{i}^{n} {x_i}^k \text{ z}$$

is called the kth sample moment, for $k = 1, 2, 3, \dots.\ while\ E\ [X^k]$ is called kth population moment, for $k = 1, 2, 3, \dots.$ Thus the method of moments estimators of the parameters is given by setting the sample moments equal to population moments and solving the resulting equations simultaneously, for the parameters of the population.

Maximum likelihood estimators

The essential feature of the principle of maximum likelihood estimation, as it applies to the problem of estimation, is that it requires the investigator to choose as an estimate of the parameter that value of the parameter for which there is the prior probability of obtaining the sample point actually observed, is as large as possible. This probability will in general depend on the parameter, which is then given that value for which this probability is as large as possible.

Suppose that the population random variable X has a probability function which depends on some parameter $\theta \ = Pr\ [X = x]\ =\ f\ (x;\ \theta)$. We suppose that the form of the function f is known, but not the value of 0. The joint probability function of the sample random variables, evaluated at the sample point $(x_1\ x_2, \dots, x_n)$, is

$$L\ (\theta)\ =\ f\ (x_1\ x_2, \dots, x_n;\ \theta)\ =\ \prod_{i}^{n} f\ (xi, \theta).$$

This function is also known as the likelihood function of the sample. We are considering it as a function of 0 when the sample values $x_1\ x_2, \dots, x_n$ are fixed.

The principle of maximum likelihood requires us to choose as an estimate of the unknown parameter that value of 0 for which the likelihood function assumes its maximum value.

4.1.2. Statistics as Estimators for Parameters

It is clearly visible that we use statistics to estimate parameters due to the lack of time, energy, resources, and infinite populations. Statistics, from the sample, can be listed as: proportions, Arithmetic averages, ranges, quartiles, deciles, percentiles, variances, and standard deviations. It will become clear enough what each one means and what it will stand for.

4.1.3. Point estimator of the proportion

Selecting an individual from this population is an experiment that has ONLY two outcomes: either that individual has the characteristic we are interested in or does not. This kind of an experiment is called a Bernoulli experiment that has two outcomes, a success or a failure. Let a random sample of size ***n*** be taken from this population. The sample proportion, denoted by $\hat{p}$ (read “p-hat”), given by

$$\hat{p} = \frac{X}{n}$$

is a point estimator of a population proportion p whereas a specific value $\hat{p} = \frac{X}{n}$ is a point estimate of the population proportion.

Here X is the number of individuals in the sample with the specified characteristic. The sample proportion $\hat{p}$ is a statistic that estimates the population proportion, p, which is a parameter.

Example 4.1

A random sample of 50 households was selected for a telephone survey. The key question asked was, do you or any member of your household own a cellular telephone with a built-in camera? Of the 50 respondents, 15 said “yes” and 35 said “no”. Find the point estimate of the true proportion, p, of households with cellular telephones with built-in cameras.

Solution:

The first group answering “yes” and the second group answering “No”, such that the proporition of saying yes is

$$\hat{p} = \frac{X}{n} = \frac{15}{50} = 0.3$$

4.1.4. Point estimator of the mean

Let there be a population with unknown mean $“\mu”, (lower\ case\ Greek\ mu)$ and unknown standard deviation "σ" (lower case Greek sigma). To estimate μ and σ, we draw a simple random sample of size n: $x_1\ x_2, \dots, x_n$. Then compute the **sample mean**

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is a ***point estimator*** of the population mean μ whereas the specific value of this estimator $\bar{x}$ is called a ***point estimate*** of μ.

and compute the sample variance

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

Then, **μ** is estimated by $\bar{x}$ and σ is estimated by s, called the sample standard deviation.

Example 4.2

A branch manager of a bank located in a commercial district of a city has developed an improved process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 15 customers during this hour is recorded over a period of one week are given below.

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20, 4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79

Find the estimate of the average waiting time, in minutes of customers in this branch.

Solution:

$$\bar{x} = average\ waiting\ time = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{4.21 + 5.55 + \cdots + 3.79}{15} = \frac{64.3}{15} = 4.29\ minutes$$

Interval Estimator

An **interval estimator** draws inferences about a population by estimating the value of an unknown parameter using an interval. **An Interval Estimation** is a range of values, calculated based on the information in the sample that the parameter in a population will be within that range with some degree of confidence.

The purpose of an interval estimate is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter. The general form of an interval estimate of a population mean is

$$\bar{x} \pm Margion\ of\ error$$

Similarly, the general form of an interval estimate of a population proportion is

$$\hat{p} \pm Margion\ of\ error$$

The sampling distributions of and play key roles in computing these interval estimates.

For example:

Television network executives want to know the proportion of television viewers who are tuned in to their networks; an economist wants to know the mean income of university graduates; in each of these cases, to accomplish the objective exactly, the statistics practitioner would have to examine each member of the population and then calculate the parameter of interest. For instance, network executives would have to ask each person in the country what he or she is watching to determine the proportion of people who are watching their shows. Because there are millions of television viewers, the task is both impractical and prohibitively expensive. An alternative would be to take a random sample from this population, calculate the sample proportion, and use that as an estimator of the population proportion. The use of the sample proportion to estimate the population proportion seems logical. The selection of the sample statistic to be used as an estimator, however, depends on the characteristics of that statistic.

Most desirable qualities for our purposes

1. ***Unbiased estimator****: An* ***unbiased estimator of*** a population parameter is an estimator whose expected value is equal to that parameter.

☞ The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ, because $E(\bar{X}) = \mu$.

☞ The sample proportion is an unbiased estimator of the population proportion because $E(\hat{P}) = p$.

☞ The difference between two sample means is an unbiased estimator of the difference between two population means because $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_1$.

2. **Consistency *estimator***: An unbiased estimator is said to be **consistent** if the difference between the estimator and the parameter grows smaller as the sample size grows larger. Thus, $\bar{X}$ is a consistent estimator of μ because the variance of $\bar{X}$ is $\frac{\sigma^2}{n}$. This implies that as n grows larger, the variance of $\bar{X}$ grows smaller. As a consequence, an increasing proportion of sample means falls close to μ.

3. **Relative Efficiency**: If there are two more unbiased estimators of a parameter, the one whose variance is smaller is said to have relative efficiency.

Example 4.2

A branch manager of a bank located in a commercial district of a city has developed an improved process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 15 customers during this hour is recorded over a period of one week are given below.

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20, 4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79

Find the estimate of the average waiting time, in minutes of customers in this branch.

Solution:

$$\bar{x} = average\ waiting\ time = \frac{\sum_{i=1}^{15} x_i}{15} = \frac{4.21 + 5.55 + \cdots + 3.79}{15} = \frac{64.3}{15} = 4.29\ minutes$$

4.2. Interval estimators of the mean and proportion

4.2.1. Confidence Interval about One Proportion

It is of interest to estimate the proportion of employees, who favor a certain type of work, or the proportion of defective items in a certain lot, or the proportion of rats having a certain kind of symptoms. Let P be the true proportion of elements that have attribute A (a certain characteristic of interest) in a population. We draw a simple random sample of size n from this population and let X equal number of elements in the sample that have attribute A. Thus the point estimate of the true proportion p in the population is given by p = X/n, where X has a binomial distribution with parameters n and p. Recall that $E(X) = np$, and $Var\ (X) = n.p.(1-p)$. Hence we find that $E\ (\hat{\mathrm{p}}) = p$, and $Var\ (\hat{\mathrm{p}}) = p.(1-p)/n.$ By the

Central Limit Theorem, the random variable given $Z = \frac{\hat{\mathrm{p}}-\mathrm{p}}{\sqrt{p.(1-p)/n}}$ has a standard normal distribution

as n increases. Thus

$$P\ (-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

where Z is the value of the standard normal variable comprising a probability of alpha on its right. This, with some algebra manipulations we reach the $100(1 - a)$ % C.I. (Confidence Interval) on p to be given by

$$\hat{P} - Z_{\alpha/2}\hat{p}(1 - \hat{p})/n < p < \hat{P} + Z_{\alpha/2}\hat{p}(1 - \hat{p})/n\ .$$

Based on the above confidence interval, when the two limits are given, we can find the sample proportion by the following equation:

$$The\ lower\ limit + the\ upper\ limit = 2(sample\ proportion).$$

EXAMPLE 4.3

In a simple random sample of 500 employees, 160 preferred to take training classes in the morning rather than in the afternoon. Construct a 95% C.I. on the true proportion of employees who favor morning training classes.

Solution:

From the information on hand we have; $x = 160, n = 500$, the confidence level is 0.95, and so $\hat{\mathrm{p}} = 0.32$, $alpha = 0.05, and\ Z_{0.025} = 1.96$. By substitution in the above interval we find that $0.28 < p < 0.36$.

This means that the true proportion, of the employees who favor the morning training classes, is between 28% and 36%.

4.2.2. Confidence Interval about One Mean

Let there be a population with mean μ and variance σ^2. We wish to construct a confidence interval about, p with $100(1 - a)$ % confidence level, where $0 < a < 1$. There are three cases to be considered here. For the following cases, we will have a simple random sample of size n from the original population. Let that sample be $X_i, i = 1, 2, \ldots, n$.

There are different cases to be considered to construct confidence intervals.

☞ **Case 1:** If sample size is large or if the population is normal with known variance

Recall the Central Limit Theorem, which applies to the sampling distribution of the mean of a sample. Consider samples of size n drawn from a population, whose mean is μ and standard deviation is σ with replacement and order important. The population can have any frequency distribution. The sampling distribution of $\overline{X}$ will have a mean $\mu_{\bar{x}} = \mu$ and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and approaches a normal distribution as n gets large. This allows us to use the normal distribution curve for computing confidence intervals.

$$\Rightarrow Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \quad has\, a\, normal\; distribution\, with\, mean = 0\; and\, \text{var}\, iance = 1$$
$$\Rightarrow \mu = \overline{X} \pm Z\,\sigma/\sqrt{n}$$
$$= \overline{X} \pm \varepsilon, \quad where\, \varepsilon\; is\, a\; measure\, of\; error.$$
$$\Rightarrow \varepsilon = Z\,\sigma/\sqrt{n}$$

☞ For the interval estimator to be good the error should be small. How it be small?

- ✓ By making n large
- ✓ Small variability
- ✓ Taking Z small

1. To obtain the value of Z, we have to attach this to a theory of chance. That is, there is an area of size $1-\alpha$ such

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$$
$$Where\;\alpha = is\, the\; probability\; that\, the\;\; parameter\, lies\, outside\, the\, \text{int}\, erval$$
$$Z_{\alpha/2} = s\tan ds\;\; for\, the\, s\tan dard\, normal\, \text{var}\, iable\, to\, the\, right\, of\;\; which$$
$$\alpha/2\; probability\, lies, i.e\, P(Z > Z_{\alpha/2}) = \alpha/2$$
$$\Rightarrow P(-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1-\alpha$$
$$\Rightarrow P(\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

$\Rightarrow (\bar{X} - Z_{\alpha/2}\,\sigma/\sqrt{n},\ \ \bar{X} + Z_{\alpha/2}\,\sigma/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ interval\ \ for\,\mu$ But usually σ^2 is not known, in that case we estimate by its point estimator S^2.

$\Rightarrow (\bar{X} - Z_{\alpha/2}\,S/\sqrt{n},\ \ \bar{X} + Z_{\alpha/2}\,S/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ interval\ \ for\,\mu$ Here are the z values corresponding to the most commonly used confidence levels.

$100(1-\alpha)$ %	α	$\alpha/2$	$Z_{\alpha/2}$
90	0.10	0.05	**1.645**
95	0.05	0.025	**1.96**
99	**0.01**	**0.005**	**2.58**

Case 2:

If sample size is small and the population variance, σ^2 is not known.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad has\ t\ distribution\ with\ n-1\ degrees\ of\ freedom$$

$\Rightarrow (\bar{X} - t_{\alpha/2}\,S/\sqrt{n},\ \ \bar{X} + t_{\alpha/2}\,S/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ interval\ \ for\,\mu$ The unit of measurement of the confidence interval is the standard error. This is just the standard deviation of the sampling distribution of the statistic.

Examples:

1. From a normal sample of size 25 a mean of 32 was found .Given that the population standard deviation is 4.2. Find

    a) A 95% confidence interval for the population mean.

    b) A 99% confidence interval for the population mean.

Solution:

$$a)\overline{X} = 32, \quad \sigma = 4.2, \quad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05, \alpha/2 = 0.025$$

$$\Rightarrow Z_{\alpha/2} = 1.96 \quad from\ \ table.$$

$$\Rightarrow The\ \ required\,\mathrm{int}\,erval\ will\,be\,\overline{X} \pm Z_{\alpha/2}\ \sigma/\sqrt{n}$$

$$= 32 \pm 1.96 * 4.2/\sqrt{25}$$

$$= 32 \pm 1.65$$

$$= \underline{\underline{(30.35,\ 33.65)}}$$

$$b)\overline{X} = 32, \quad \sigma = 4.2, \quad 1-\alpha = 0.99 \Rightarrow \alpha = 0.01, \ \alpha/2 = 0.005$$

$$\Rightarrow Z_{\alpha/2} = 2.58 \quad from\ \ table.$$

$$\Rightarrow The\ \ required\,\mathrm{int}\,erval\ will\,be\,\overline{X} \pm Z_{\alpha/2}\ \sigma/\sqrt{n}$$

$$= 32 \pm 2.58 * 4.2/\sqrt{25}$$

$$= 32 \pm 2.17$$

$$= \underline{\underline{(29.83,\ 34.17)}}$$

2. An insurance company is testing a new service system which is supposed to reduce wasting time. From the six people who are used as subjects, it is found that the average waiting time is 2.28 minutes, with a standard deviation of .95 minutes. What is the 95% confidence interval for the mean change in minutes?

Solution:

That is, we can be 95% confident that the mean decrease in blood pressure is between 1.28 and 3.28 points.

$$\overline{X} = 2.28, \quad S = 0.95, \quad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05, \ \alpha/2 = 0.025$$

$$\Rightarrow t_{\alpha/2} = 2.571\ with\ df = 5 \quad fromtable.$$

$$\Rightarrow The\ \ required\,\mathrm{int}\,erval\ will\,be\,\overline{X} \pm t_{\alpha/2}\ S/\sqrt{n}$$

$$= 2.28 \pm 2.571 * 0.95/\sqrt{6}$$

$$= 2.28 \pm 1.008$$

$$= \underline{\underline{(1.28,\ 3.28)}}$$

4.2.3. Confidence Interval about One Variance

In studying the precision of measuring instruments, and in studying variability in populations, we face the problem of estimating the population variance, or its standard deviation from a random sample. In this section we will investigate how to construct a confidence interval on either the population variance or the standard deviation. For to have a good confidence and excellent estimation for a $100(1 - \alpha)$ % C.I. on the population variance we need to assume that that population has a normal distribution with some variance σ^2.

From a random sample of size n; $X_1, X_2, \dots, X_n$, taken from that normal population, we can see that

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

is the sample variance. Thus the random variable given by

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

will have a chi-square (kigh-square) distribution with n-1 degrees of freedom. Let $X^2_{\alpha/2}$, and $X^2_{1-\alpha/2}$be those values on the Chi-square axis such that the area to the right of that value is $\alpha/2$ and $1 - \alpha/2$ respectively. Hence

$$P\left(X^2_{1-\frac{\alpha}{2}} < X^2 < X^2_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Where

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

As it was done before, with little algebra, we have the $100(1 - \alpha)$ % C.I. on σ^2 given by

$$\frac{(n-1)\, s^2}{X^2_{\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)\, s^2}{X^2_{1-\frac{\alpha}{2}}}$$

In addition to the C.I. on σ^2, we can get the 100(1- α) % C.I. on σ, and thus it is given by just taking the square root of every term in the above interval.

$$\left(\frac{(n-1)\, s^2}{X^2_{\frac{\alpha}{2}}}\right)^{\frac{1}{2}} \le \sigma \le \left(\frac{(n-1)\, s^2}{X^2_{1-\frac{\alpha}{2}}}\right)^{\frac{1}{2}}$$

EXAMPLE

Human beings vary in the time it takes them to respond to driving hazards. In one experiment in which 100 healthy adults between age 21 and 30 years were subjected to a certain driving hazard, and the sample variance of the observed times it took them to respond was 0.0196 second squared. Assuming that the times to respond are normally distributed, estimate the variability in the time response of the given age group using a 95% C.I.

Solution

The confidence level is 0.95, so that $\alpha/2 = 0.025$. Reading the X^2-Table with $100 - 1 = 99$ degrees of freedom we find that $X^2_{0.025} = 128.45$, $X^2_{0.975} = 128.45$. Substituting in the C.I. for σ^2, we obtain the following interval

$$0.0151 < \sigma^2 < 0.0265.$$

Moreover the 95% C.I. on σ is given by

$$0.123 < \sigma < 0.163.$$

4.3. Confidence Interval about two Parameters

The confidence intervals to be calculated on two parameters will involve: a) two means, b) two proportions, c) two variances, and two standard deviations. For the means and proportions the confidence intervals will be established on the difference between the two parameters, while in the case of two variances or two standard deviations, the confidence interval will be on the ratio between the two parameters.

4.3.1. Confidence Interval about the difference between two proportions

Comparisons of proportions, in different groups, are a common practice. A whole-seller compares the proportions of defective items found in two separate sources of supply from which he buys these items. Consider two independent samples of sizes n_1 and n_2 that are drawn from two binomial populations with parameters (i.e. probabilities of successes) p_1 and p_2. A $100(1 - \alpha)$ % confidence interval will derived on the difference between p_1 and p_2 using the central limit theorem and the normal approximation to the binomial distribution.

Let x_1 and x_2 be the number of successes obtained in sample 1 and sample 2 respectively. We then have $\hat{P}_1 = x_1/n_1$ and $\hat{P}_2 = x_2/n_2$ is approximately normal with mean as the point estimates of p_1 and p_2 respectively. Moreover we have

$$E(\hat{P}_1) = p_1, \mathrm{E}(\hat{P}_2) = p_2 \text{ with } \mathrm{Var}(\hat{P}_1) = \frac{p_1(1-p_1)}{n_1}, \mathrm{Var}(\hat{P}_2) = \frac{p_2(1-p_2)}{n_2}$$

Because of the independence of the two samples, we can write

$$E(\hat{P}_1 - \hat{P}_2) = P_1 - P_2, \text{ with } Var(\hat{P}_1 - \hat{P}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Now by, the Central Limit Theorem, the random variable Z given by

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a standard normal distribution with mean 0 and variance 1, i.e. $Z \cong$ N (0, 1). Hence

$P\ (-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha,$

where Z is as given above. With little algebra manipulations we reach at the 100(1- α) % confidence interval on the difference between the two proportions, $(P_1 - P_2)$ as given by the two limits

Lower Limit: $\left(\hat{P}_1 - \hat{P}_2\right) - Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Upper Limit: $\left(\hat{P}_1 - \hat{P}_2\right) + Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

EXAMPLE

A certain change in a manufacturing procedure for component parts is being considered. Samples were taken using both the old and the new procedures in order to determine the difference induced by the new procedure. Suppose that 75 out 1500 items, from the existing procedure, were found defective, and 80 out 2000 items, from the new procedure were found to be defective. Find a 90% C.I. on the true difference in the fraction of defectives between the two procedures.

Solution:

Let p_1 and p_2 be the true proportions of defectives in the existing and new procedures, respectively. Thus we have x_1 = 75 and x_2 = 80, with n_1 = 1500 and n_2 = 2000. With 1 – α = 0.90, we see that $Z_{\ 0.05}$ = 1.645, using the Standard Normal Table, and with the above limits on the difference between the fraction of defectives between the two procedures we found that the 90% C.I. is given as

$$-0.0017 < P_1 - P_2 < 0.0217$$

Notice as the interval contains zero, there is no reason to believe that the new procedure reduces the proportion of defectives.

4.3.2. Confidence Interval about the difference between Two Means

It is quite often the following question is raised: Which average of those two means which are under investigation is better, or higher or smaller, or worse? In comparative experiments the investigator wishes to estimate the difference between two processes based on the difference between their means.

Assume that there are two populations with their means and variances given μ_i and s_i^2 for i =1, 2 respectively. These populations could be normally distributed or not, as the discussion will reveal the cases below. We will select two simple random samples of sizes n_i, i = 1, 2, and denote them by X_j and Y_j, j = 1,2, …. n_i. Based on the data, from the samples, we can compute the mean and the variance for each sample, which are given by $\bar{X}$, S_1^2 and $\bar{Y}$ and S_2^2.

Case I: The two population variances, s_i^2 for $i = 1, 2$, are known.

We know from earlier discussion that $\bar{X}$ and $\bar{Y}$ are normally distributed each has a normal distribution with mean and variance given by μ_1 , S_1^2/n_1 and μ_2 , S_2^2/n_2 respectively, and thus the random variables

$$Z_1 = \frac{\bar{X} - \mu_1}{\sqrt{S_1^2/n_1}}, and\ Z_2 = \frac{\bar{X} - \mu_2}{\sqrt{S_2^2/n_2}}$$

each will have a standard normal distribution with mean 0 and variance 1. Because of the independence of the two samples we have $\bar{X} - \bar{Y}$ as the point estimate $\mu_1 - \mu_2$ for which it has a normal distribution with mean equal to $\mu_1 - \mu_2$ and variance given by $\frac{S_1^2}{n_1} + S_2^2/n_2$, and thus the random variable

$$Z_{cal} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Has a standard normal distribution with mean 0 and variance 1. Hence we have

$$P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Now as it was the case when we derived the C.I. on the difference between two proportions, and with little algebra manipulation we reach at the $100(1 - \alpha)$ % C.I. on the difference $\mu_1 - \mu_2$ to be given by the following two limits

Lower Limit: $(\bar{x}_1 - \bar{x}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Upper limit: $(\bar{x}_1 - \bar{x}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

In other words, the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

It is to be noted here that the above Confidence interval applies for any values for the sample sizes.

Case II: The two population variances s_i^2 for $i = 1, 2$, are unknown, Large Sample Sizes.

In this case the question that will be asked is: What are the sample sizes? For sample sizes of greater than 30 each, by using the Central Limit Theorem, and replacing the population variances by their estimates, from the sample we see that the random variable given by

$$Z_{cal} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has a standard normal distribution. Hence the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ will be given by

$$(\bar{x}_1 - \bar{x}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Case III: The two populations variances, s_i^2 for $i = 1, 2$, are unknown, Small sample sizes.

We will assume in this case that the populations, from which the two samples are randomly drawn, are normally distributed with means $\mu_1$and μ_2. Again there is a question to be asked about the two population variances: Are they equal or unequal? When the two population variances are equal, we can pool the samples' variances to estimate the common value. This value is termed **the pooled Variance**, and it is given by

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

In this case we have a new random variable given by

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

That has a student t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus the 100(1 – α) % C.I. on the difference $\mu_1 - \mu_2$ will be given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\, s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\, s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

On the other hand, when the two population variances are not equal, we still have a student t-distribution for the random variable T given by

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In this case the degrees of freedom, v will be calculated from the following formula, and it is rounded down to the nearest whole number,

$$n = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

In this case the 100(1 – α) % C.I., on the difference $\mu_1 - \mu_2$, will be given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\mathrm{n}_1} + \frac{s_2^2}{\mathrm{n}_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\mathrm{n}_1} + \frac{s_2^2}{\mathrm{n}_2}}$$

Consider this example but with small samples, and sample standard deviations as given below:

Sample	Mean	Size	Sample Standard
I	85	12	5
II	81	10	4

Calculate the 95% C. I on the difference between the two means, by

a) Pooling for the common variance of the two populations,

b) Not pooling

Solution:

a) In the pooled case we have this interval

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\, s_{pooled}\sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\, s_{pooled}\sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}}$$

By using the data we get the 95% C.I. on $\mu_1 - \mu_2$ to be given by (-0.088, 8.088),

With df = 20, the pooled standard deviation = 4.577.

b) In the non –pooled we have to use this interval

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\mathrm{n}_1} + \frac{s_2^2}{\mathrm{n}_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\mathrm{n}_1} + \frac{s_2^2}{\mathrm{n}_2}}$$

The degrees of freedom are given by $n = \frac{\left(\frac{s_1^2}{\mathrm{n}_1} + \frac{s_2^2}{\mathrm{n}_2}\right)^2}{\frac{\left(\frac{s_1^2}{\mathrm{n}_1}\right)^2}{\mathrm{n}_1 - 1} + \frac{\left(\frac{s_2^2}{\mathrm{n}_2}\right)^2}{\mathrm{n}_2 - 1}}$ Based on the interval cited and the data provided, we have these limits for the 95% C.I.

4.4. The t distribution

So far only large samples (defined as sample sizes in excess of 25) have been dealt with, which means that (by the Central Limit Theorem) the sampling distribution of $\bar{x}$ follows a Normal distribution, whatever the distribution of the parent population. Remember, from the two theorems, that:

☞ if the population follows a Normal distribution, $\bar{x}$ is also Normally distributed; and

☞ if the population is not Normally distributed, $\bar{x}$ is approximately Normally distributed in large samples ($n \geq 25$).

In both cases, confidence intervals can be constructed based on the fact that

$$\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

and so the standard Normal distribution is used to find the values which cut off the extreme 5% of the distribution ($z = \pm 1.96$). In practical examples, we had to replace o by its estimate, s. Thus the confidence interval was based on the fact that

$$\frac{\bar{x}-\mu}{\sqrt{s^2/n}} \sim N(0,1)$$

in large samples. For small sample sizes, the above equation is no longer true. Instead, the relevant distribution is the $\boldsymbol{t}$ distribution and we have

$$\frac{\bar{x}-\mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

The random variable defined in the above equation has a t distribution with $n - 1$ degrees of freedom. As the sample size increases, the t distribution approaches the standard Normal, so the latter can be used for large samples.

The t distribution was derived by *W.S. Gossett in 1908* while conducting tests on the average strength of Guinness beer (who says statistics has no impact on the real world?). He published his work under the pseudonym 'Student', since the company did not allow its employees to publish under their own names, so the distribution is sometimes also known as the Student distribution.

The t distribution is in many ways similar to the standard Normal, insofar as it is:

☞ unimodal;

☞ symmetric;

☞ centred on zero;

☞ bell-shaped;

☞ extends from minus infinity to plus infinity

☞

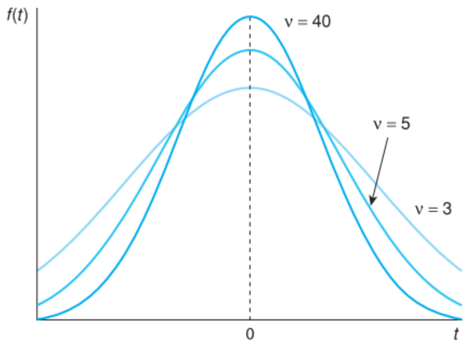


Figure: The t distribution drawn for different degrees of freedom 1

The differences are that it is more spread out (has a larger variance) than the standard Normal distribution, and has only one parameter rather than two: the **degrees of freedom**, denoted by the Greek letter ν (pronounced 'nu'). In problems involving the estimation of a sample mean the degrees of freedom are given by the sample size minus one, i.e. $\nu = n - 1$.

The *t* distribution is drawn in Figure 4.6 for various values of the parameter ν. Note that the fewer the degrees of freedom (smaller sample size) the more dispersed is the distribution. To summarize the argument so far, when

☞ the sample size is small, *and*

☞ the sample variance is used to estimate the population variance,

then the *t* distribution should be used for constructing confidence intervals, not the standard Normal. This results in a slightly wider interval than would be obtained using the standard Normal distribution, which reflects the slightly greater uncertainty involved when s^2 is used as an estimate of σ^2 if the sample size is small.

Apart from this, the methods are exactly as before and are illustrated by the examples below. We look first at estimating a single mean, then at estimating the difference of two means. The t distribution cannot be used for small sample proportions (explained below) so these cases are not considered.

4.5. Samples-Size Determination

Frequently, we wish to determine how large a sample should be in order to ensure that the error in estimating the population mean, or the population proportion, is less than a specified value of the error. As it was shown in the derivation of the confidence interval for μ and P, the margin of error was given by the following two formulas respectively

$$E = \begin{cases} z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}, & for\ continous\ response\ variable \\ z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, & for\ binomial\ response\ variable \end{cases}$$

when the confidence level is taken to be 100(1 – α) % in both cases.

It is quite clear when the sample is too small; the required precision is not achieved. On the other hand, when the sample size is large, then some resources have been wasted. In order to meet the criterion of a specified margin of error calculations can be made to approximate the sample size needed in both cases of the mean and the proportion. In case of finding the sample size, the rounding will be up to the nearest whole number in order to meet the error criterion.

☞ Minimum required sample size in estimating the population mean, μ is:

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

Where, E is the half-width of a (1-α) 100% confidence interval for μ called **margin of error.**

☞ Minimum required sample size in estimating the population proportion *P* is:

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \hat{p}(1-\hat{p})}{E^2}$$

Where, E is the half-width of a (1-α) 100% confidence interval for *P* called **margin of error.**

Example 4.3

A market research firm wants to conduct a survey to estimate the average amount spent on entertainment by each person visiting a popular resort. The people who plan the survey would like to be able to determine the average amount spent by all people visiting the resort to within $120, with 95% confidence. From past operation of the resort, an estimate of the population standard deviation is $400. What is the minimum required sample size?

Solution: from the information given above, we know

$\alpha = 0.05$ *then* $Z_{\alpha/2} = Z_{0.025} = 1.96, \sigma = 400,$ *and the margin error,* $d = 120$ *then insert the value in the equation below,*

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = (1.96)^2 * \frac{(400)^2}{120^2}$$

EXAMPLE

An economist wants to know if the proportion of the population who commutes to work via carpooling is on the increase due to gas prices. What sample size should be obtained if the economist wants to estimate with 2 percentage points of the true proportion with 90% confidence if

a) The economist uses the 2006 estimate of 10.7% obtained from the American Community Survey?

b) The economist does not use any prior estimates?

Solution:

In both cases we have $E = 0.02, and\ z_{\alpha/2} = 1.645$. Hence for

a) We use $n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 \hat{p}(1-\hat{p})}{E^2}$, and find that $n = 647.$

b) We use $n = 0.25\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$, and find that $n = 1692.$

We can see the effect of not having a prior estimate of p. In this case, the required sample size more than doubled of the case when we used a prior estimate.

Summary

☞ Estimation is the process of using sample information to make good estimates of the value of population parameters, for example using the sample mean to estimate the mean of a population.

☞ There are several criteria for finding a good estimate. Two important ones are the (lack of) bias and precision of the estimator. Sometimes there is a tradeoff between these two criteria – one estimator might have a smaller bias but be less precise than another.

☞ An estimator is unbiased if it gives a correct estimate of the true value on average. Its expected value is equal to the true value.

☞ The precision of an estimator can be measured by its sampling variance (e.g. s^2/n for the mean of a sample).

☞ Estimates can be in the form of a single value (point estimate) or a range of values (confidence interval estimate). A confidence interval estimate gives some idea of how reliable the estimate is likely to be.

☞ For unbiased estimators, the value of the sample statistic (e.g. X) is used as the point estimate.

☞ In large samples the 95% confidence interval is given by the point estimate plus or minus 1.96 standard errors (e.g. $\bar{x} \pm 1.96$ for the mean).

For small samples the *t* distribution should be used instead of the Normal (i.e. replace 1.96 by the critical value of the *t* distribution) to construct confidence intervals of the mean.

CHAPTER FIVE

5. HYPOTHESIS TESTING

Outline

- Introduction
- Fundamental Concepts
- Methods and Steps in Testing a Statistical Hypothesis
- Hypothesis Testing about One Parameter
  - Hypothesis Testing about One Proportion
  - Hypothesis Testing about One Mean
  - Hypothesis Testing about One Variance
- Hypothesis Testing about Two Parameters
  - Tests About Two Proportions
  - Tests About Two Means
  - Tests About Two Variances

5.1.Concepts of Hypothesis Testing

Testing a statistical hypothesis is the second main and major part of inferential statistics. A statistical hypothesis is an assumption or a statement, about one or two parameters and involving one or more than one population. A statistical hypothesis may or may not be true. We need to decide, based on the data in a sample, or samples, whether the stated hypothesis is true or not. If we knew all the members of the population, then it is possible to say with certainty whether or not the hypothesis is true. However, in most cases, it is impossible, and impractical to examine the entire population. Due to scarcity of resources, lack of time, and tedious calculations based on a population, we can only examine a sample that hopefully represents that population very well. So the truth or falsity of a statistical hypothesis is never known with certainty.

Testing a statistical hypothesis is a technique, or a procedure, by which we can gather some evidence, using the data of the sample, to support, or reject, the hypothesis we have in mind. This is also one way of making inference about population parameter, where the investigator has *prior notion* about the value of the parameter.

5.1.1. THE NULL AND ALTERNATIVE HYPOTHESES

The first step, in testing a statistical hypothesis, is to set up a null hypothesis and an alternative hypothesis. When we conjecture a statement, about one parameter of a population, or two parameters of two populations, we usually keep in mind an alternative conjecture to the first one. Only one of the conjectures can be true. So, in essence we are weighing the truth of one conjecture against the truth of the other. This idea is the first basic principle in testing a statistical hypothesis. For Example, a person is accused of a crime; he/she faces a trial. The prosecution presents its case, and a jury must make a decision on the basis of the evidence presented. In fact, the jury conducts a test of hypothesis.

Typically, the *question of interest will be represented by the alternative hypothesis*, as illustrated in the following examples, note how consistently what is interesting to the analyst is the alternative hypothesis in the following examples of some questions we might encounter and the corresponding statistical hypotheses that might be framed:

1. An accountant doing an audit is becoming suspicious of the figures shown in the books of a *big company called Unron*; she/he extracts the data from several hundred transactions and wants to know if the frequencies of the ten digits $(0, 1, \dots 9)$ in the last portions of the entries are equal (radical deviation from equality would suggest that the numbers were fraudulently invented, since people aren’t very good at making up numbers that fit the uniform probability distribution).

H_0: the frequencies *are* all 0.1;

H_1: the frequencies *are not* all 0.1.

2. Suppose a stock broker has become interested in the performance of the shares for *DASHEN BANK*; he wants to know if the data for the last three years support the view that the growth rate is at least 6% per year. If it is, he will recommend to a client interested in long-term investments that the investment fits the client’s profile.

H0: the regression coefficient is *less than* 6% per year.

H1: the regression coefficient of stock price versus year is *6% per year or more*;

3. A marketing firm has three different variations on an advertising campaign. They are trying them out in six different regions of the target market by counting the number of answers to questions using a Likert scale (1: strongly disagree, 2: disagree… 5: strongly agree).

i. Are there differences among the ad versions in the way people respond in the six regions overall?

H0: there are *no* main (overall) effects of ad versions on responses;

H1: there *are* main effects of ad versions on responses.

Are there differences among the regions of the market in the way people respond?

H0: there are *no* main effects of region on responses;

H1: there *are* main effects of region on responses.

ii. Are there any regional variations in the way people respond to the different campaigns

H0: there are *no interactions* between the ad variations and the regions of the market in the way people respond;

H1: there *are* interactions between the ad variations and the regions of the market.

4. An investor is looking at two different manufacturers of plant as potential investments. One of the steps in due diligence is to examine the reliability of quality control of the two factories' production lines by comparing the variances of the products.

H0: there is no difference in the variances of the two production lines;

H1: there is a difference in the variances of the two production lines.

In general, the null hypothesis, H0, is the one that posits no effect, no difference, no relationship, or a default state. H1, the alternative hypothesis, is the one that posits an effect, a difference, a relationship, or deviation from a ho-hum uninteresting state. There is nothing absolute or inevitable about the framing of H_0 and H_1: the decisions depend on the interests of the investigator.

5.1.2. Possible Decisions

The test procedure will lead to either one of the following decisions:

1. Reject the Null Hypothesis, H_o, i.e., conclude that H_o is a false statement and this will lead to take, or accept, that the alternative hypothesis H_1 as a true statement.

2. Do not reject the Null hypothesis, H_o. This means that there is no evidence from the sample, to disprove the null hypothesis. The non-rejection of H_o should not imply that it is true. This is because the objective of testing a statistical hypothesis is to disprove, or reject, the null hypothesis with a high certainty, rather than to prove it. Thus if the statistical test rejects H_o then we are highly certain that it is false. However, if the test does not reject H_o, then we interpret the non-rejection as that the sample does not give enough evidence to disprove the null hypothesis. In other words, the rejection of the null hypothesis is the decisive conclusion that we can depend on.

Based on the decision, whether to "Reject H0" or "Do Not Reject H0", we should be careful in stating the null and alternative hypotheses. This is due to the fact that originally we have two statements to be examined against each other and we may call either one of them the null hypothesis. But since we are only highly confident about the conclusion of rejecting the null hypothesis, we take H0 as the statement that the sample will reject. On the other hand, the alternative hypothesis will be that statement which we hope that the data will support. In the prosecution case example above,

H_0: The defendant is innocent.

H_1: The defendant is guilty.

5.1.3. STEPS IN HYPOTHESIS TESTING:

There are two methods to test a statistical hypothesis, namely The *Classical* or *Traditional* method, and the *P-value* method. Both of these methods will be introduced and used in this text. Based on the notation and definitions that were set above, we will list the steps in the Classical method, in general, first and then the steps for the *p-value* method next. More detailed steps will be outlined later based on the parameter, or parameters, involved, or stated in the hypotheses.

5.1.3.1. Classical Method Steps

1. Determine, and clearly, state the two hypotheses: H0 and H1. Equality to the assigned parameter should be included under the Null Hypothesis.
2. Decide on the significance level α. Find the critical value or values, and locate the rejection region or regions (all based on the parameter and distribution under consideration).
3. Choose the appropriate Test statistic for the hypotheses based on the parameter on hand.
4. Using the information provided by the data in the sample, and the computed statistics, calculate the test statistic that was chosen in Step 3.
5. Make your statistical decision, whether to reject, or not to reject, H0 based on the comparison between the computed value of the test statistic and the critical value(s) found in Step 2, and as outlined earlier.
6. Give the conclusion, or the answer, in a statement that anyone can understand without any mathematical jargons or statistical ambiguity.

5.1.3.2. P-value Method Steps

1. Determine and clearly state the two hypotheses: H0 and H1. Equality to the assigned parameter should be included under the Null Hypothesis.

2. Decide on the significance level α.

3. Choose the appropriate Test statistic for the hypotheses based on the parameter on hand.

4. Using the information provided by the data in the sample, and the computed statistics, calculate the test statistic that was set up in Step 3.

5. Make your statistical decision, whether to reject, or not to reject, H0 based on the comparison between the theoretical significance level α, (that was set up above) and the calculated p-value. (This p-value is the practical, or attained, significance level, based on the type of the test and the distribution of the parameter involved). A p-value less than α will lead to the rejection of H0, otherwise do not reject H0.

6. Give the conclusion, or the answer to the question, in a statement that anyone can understand without any mathematical jargons or statistical ambiguity.

The above steps will be applied to test on one parameter or two parameters whether the test was two tailed test or one tailed, left or right test. In the next section we will introduce the test of a statistical hypothesis on one parameter. The one parameter case will involve; one proportion, one mean and one standard deviation, or one variance.

5.1.4. TYPES OF ERRORS

The procedure, in testing a statistical hypothesis, either rejects the null hypothesis or not. Of course the truth is never known, i.e. we do not know whether H_o is true or not. The "true state of nature" may then be that H_o is true or H_o is false. We make the decision of rejecting the null hypothesis or not rejecting it, without knowing the true state of nature. In making a decision about testing a statistical hypothesis, two types of errors may be committed:

☞ **Type I Error:** A Type I error has been committed if the test rejects the null hypothesis, H_o when in fact it is true. The probability of making such an error will be denoted by α, (The Greek letter Alpha). For sure, it is clear that $0 \leq \alpha \leq 1$.

☞ **Type II Error:** A Type II error has been committed if the test does not reject $H_o$when $H_o$is false. The probability of making such an error will be denoted by β, (the Greek letter Beta) with $0 \leq \beta \leq 1$. What is more important is that we do not like to make such errors with high probabilities.

In either one of the other two cases, there is no error committed, as shown by Table I

Table 1 Decision making with incomplete information

		Nature State of $\mathrm{H_o}$	
Based on data		Ho true	Ho false
	Reject H_o	Type I error (α)	Correct (1 – β)
	Do not reject H_o	Correct (1 – α)	Type II error (β)

Each type of error has a certain probability of being committed. These probabilities are given specific names, and values, due to their importance and the severity of the decision.

Level of Significance: The probability of committing a type I error is denoted by (Alpha) α. It is called the theoretical level of significance for the test. The most common used values for α are: $0.01, 0.05$, or 0.10. Other values for α are at the discretion of the researcher. More expressions for α are:

$$\alpha \; = \; P \; (committing \; a \; type \; I \; error),$$

$$= \; P \; (rejecting \; H_o \; when \; H_o is \; true),$$

$$= \; P \; (rejecting \; H_o \; when \; H_1 \; is \; false).$$

The probability of committing a type II error is denoted by (beta) β. Other labels for β are:

$$\beta = P \text{ (committing a type II error)},$$

$$= P \text{ (not rejecting } H_o \text{when } H_o \text{is false)},$$

$$= P \text{ (not rejecting } H_o \text{ when } H_1 \text{ is true)}.$$

Power of the Test: The value of 1– β, which stands for $P\;(rejecting\; H_o \; when\; H_o\; is\; false)$ is the power of the test. The probabilities of type I and Type II errors tell us how good the test is. Clearly we do not like to make a type I error with high probability, as well as we like to make a correct decision with a very high power. The smaller these probabilities (of type I error and Type II error) are the better is the test. Ideally, we like to use a test procedure for which both type I and type II errors have small probabilities. However, it turns out that the type I error and the type II error are related in such a way that a decrease in the probability of one of them generally results in an increase in the probability of the other. It is not possible to control both probabilities based on a fixed sample size. Traditionally, or by convention, statisticians have adopted to fix the level of significance of the test in advance, and to search for a test procedure that will minimize the probability of making a type II error, and consequently maximize the power for the test. Practically, we can make those probabilities, namely α and β, smaller by taking a larger sample if possible.

5.1.5. The Test Statistic

The test statistic is a quantity that depends on the information, or statistics, that the sample will provide. It is a function of the sample statistics, and the value(s) of the parameter(s) under the null hypothesis. Thus a statistic is a random variable until we get some values from the sample. The numerical value of the test statistic (large or small) leads us to decide whether or not to reject the Null Hypothesis when it is compared to the critical value(s) of the test.

5.1.6. The Critical Region or The Rejection Region (CR or RR)

The critical region is an interval, or a union of intervals, which is determined by using special and certain distributions with the appropriate Table values. It depends on the distribution of the test statistic when H0 is true, on the form of the alternative Hypothesis, and on the level of significance that was set for the test.

5.1.7. Conclusion and interpretation

The final conclusion, of the test procedure, is based on whether or not the computed value of the test statistic falls inside the critical region, or not, as follows:

i. Reject H0 if the computed value of the test statistic falls in the critical region.

ii. Do not reject H0 if the computed value of the test statistic does not fall inside the critical region.

In either of the two cases detailed above, an interpretation and a practical statement are due in order to answer the question that was raised before the test procedure started.

5.2.Hypothesis Testing about One Parameter

In this section we will discuss, and display, the procedures for testing a statistical hypothesis about one population parameter. The one population parameter, which is of interest, will include: one proportion, one mean, and one variance or a standard deviation. In general, let that parameter be θ, and its assumed value be θ_0. Following the steps that were set earlier, we will give the steps in more details for the case when one proportion is under investigation.

It is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistic based on the null hypothesis being true.

Thus the steps will go as follows.

5.2.1. Hypothesis Testing about One Proportion, P

Recall that the best point estimate of p, the proportion of the population with a certain characteristic, is given by

$$\hat{p} = \frac{x}{n},$$

where x is the number of individuals in the sample with the specified characteristic of interest and n is the sample size. Recall that the sampling distribution of $\hat{p}$ is approximately normal, with mean, $E(\hat{p}) = p$ and standard deviation

$$sd(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

In addition to the above two criteria, the following requirements should be satisfied:

1. The sample is a simple random sample
2. $np(1 – p) \geq 10.$
3. The sample values are independent of each other.

Classical Method Steps

For a specified value of the proportion P_o we have (We are using the z-test on one proportion):

1. State the null and alternative hypotheses:
   a) There are three ways to set up the null and alternative Hypotheses. Equal hypothesis versus not equal hypothesis: $H_o: P = P_0$ versus $H_1: P \neq P_0$, two-tailed test.
   b) At least versus less than: $H_o: P \geq P_0$ versus $H_1: P < P_0$, left-tailed test.
   c) At most versus greater than: $H0: P \leq P_0$versus H_1: P > P_0, right-tailed test.
2. Let α (the most used values for the level of significance are: $0.01, or$ $0.05, or$ 0.10) be the significance level. Based on the three cases in step 1, we have the following three cases that will go along with that:
   a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|z| > Z_{\alpha/2}$ to include both areas in this case each will be $\alpha/2$.

b) For the left-tailed test, the critical value is $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$, the area to the left of $-Z_{\alpha}$is α.

c) For the-right tailed test, again, there is one critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$, the area to the right t of $Z_{\alpha}\ is\ \alpha$.

3. For the test statistic we have $z = \frac{\hat{p}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{x-nP_0}{\sqrt{nP_0(1-P_0)}}$, where n is the sample size and $\hat{p} = \frac{x}{n}$.

4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed, or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

6. The interpretation and conclusion are due to answer the question that was raised

EXAMPLE

The government of a wealthy country intends to institute a program to discourage investment in foreign countries by its citizens. It is known that in the past 35% of the country's adult citizens held investment in foreign countries. The government wishes to determine if the current percentage of adult citizens, who own foreign investment is greater than this long term figure of 35%. A random sample of 800 adults is selected, and it is found that 320 of these citizens hold foreign assets. Is this percentage greater than 35%? Use a 10% significance level for testing this claim.

Solution:

Using the setup above for the classical method on testing a statistical hypothesis on one proportion we proceed as follows:

1. $H_o: P \leq 0.35\ Versus\ H_1: P > .35$. This is a right–tailed test.

2. The level of significance is given to be $10\%, or\ 0.10$. Thus $\alpha = 0.10$. Since the test is onetailed, on the right side, we have one critical value given by: $C.V. = Z_{\alpha} = Z_{0.10} = 1.282$ and the rejection region is given by: $Z > 1.282$.

3. The test statistic is, $z = \frac{\hat{p}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{x-nP_0}{\sqrt{nP_0(1-P_0)}}$, where $n = 800, x = 320, and\ P_0 = 0.35$.

4. From the above values, and form for the test statistic, we find $Z_{cal} = 2.965\ where\ \hat{p} = 0.40$.

5. Since $Z_{cal} = 2.965 > C.V. = Z_{0.10} = 1.282, we\ reject\ H_o.$

6. It is concluded that the percentage of adult citizens, who own investment in a foreign country, is greater than 35%.

EXAMPLE 2

Apply the P-value method to check on the test in EXAMPLE above

Solution:

Using the setup above for the p-value method on testing a statistical hypothesis on one proportion we proceed as follows:

1. 1. $H_o: P \leq 0.35\ Versus\ H_1: P > .35$. This is a right–tailed test.
2. The level of significance is given to be 10%, or 0.10.
3. The test statistic is $z = \frac{\hat{p}-P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{x-nP_0}{\sqrt{nP_0(1-P_0)}}$, where $n = 800, x = 320, and\ P_0 = 0.35$.
4. From the above values, and form for the test statistic, we find $Z_{cal} = 2.965\ where\ \hat{p} = 0.40$.

5. Since $Z_{cal} = 2.965$. Since we have a right-tailed test, then the p-value for the test will be calculated by finding $P\ (Z > 2.965)$. This is done by applying step No. 5. Part c) in the p-value method steps. Using the table for standard normal distribution we find ourselves trapped in rounding to two decimal places.

First, let us take $Z_{cal} = 2.97$. Based on that we see then $P\ (Z > 2.97) = 1 – P(Z < 2.97)$, and from the Standard Normal Table, we have $P\ (Z > 2.97) = 1 – 0.9985 = 0.0015 < 0.10$, hence the Null hypothesis is rejected.

Second, let us take $Z_{cal} = 2.96$. Based on that we see then $P\ (Z > 2.96) = 1 – P\ (Z < 2.96)$, and from Standard Normal Table, we have $P\ (Z > 2.96) = 1 – 0.9985 = 0.0015 < 0.10$, hence The Null hypothesis is rejected. In this case it did not make a difference whether you rounded up or down the calculated value for the test statistic. Using a graphing calculator, and testing the same hypothesis, we find that the p-value, to 4 decimal places is, again, 0.0015. Thus the null hypothesis is rejected.

6. It is concluded that the percentage of adult citizens, who own investment in a foreign country, is greater than 35%.

5.2.2. Hypothesis Testing About The Population Mean, μ:

Dear learner, this section will display the procedure, by using the two methods outlined above for Testing a statistical hypothesis about one population mean. There are three cases to be considered in this section. Again, as it was stated above for one proportion, it is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistics based on the null hypothesis being true.

Case I: Testing on One Mean, μ when sampling is from a normal distribution with Population variance, σ^2 known

Classical Method Steps: For a specified value of the population mean $\mu0$ we have (We are using the z-test):

1. State the null and alternative hypotheses.

There are three ways to set up the null and alternative Hypotheses:

a) Equal hypothesis versus not equal hypothesis: $H_0: \mu = \mu_o$versus $H_1: \mu \neq \mu_o$, two-tailed test

b) At least versus less than: $H_0: \mu \geq \mu_o$versus $H_1: \mu < \mu_o$ left-tailed test

c) At most versus greater than: $H_0: \mu \leq \mu_o$versus $H_1: \mu > \mu_o$, right-tailed test

2. Let α be the significance level, and based on the three cases in step1, we have the following three cases that will go along for finding the critical values and rejection regions:

   a) For the two-tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$, with the area on each is $\alpha/2$.

   b) For the left tailed test, there is the critical value of $-Z_\alpha$, and the rejection region is given by $Z < -Z_\alpha$, with the area to the left of $-Z\alpha$ is equal to α.

   c) For the right tailed test, again, there is on critical value given by Z_α, and the rejection region is $Z > Z_\alpha$, with the area to the right of Z_α, is equal to α.

3. The test statistic is given by $z = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$ = , where n is the sample size, x is the mean, and Z has the standard normal distribution, $N(0, 1)$.
4. The above test statistic is computed based on the information provided to us by the sample data.
5. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.
6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.3

To test $H_o: \mu \geq 50$ versus $H_1: \mu < 50$, a random sample of n = 24 is obtained from a population that is known to be normally distributed with σ = 12, and we got a sample mean of 47.1. Will the null hypothesis be rejected?

Solution:

Applying the classical Method steps for test on one mean we have

1. State the null and alternative hypotheses.

$H_o: \mu \geq 50\ versus\ H_1: \mu < 50$, left-tailed test

2. Let α = 0.05 be the significance level.

   a) For the left tailed test, there is the critical value of $-Z_{0.05} = -1.645$, and the rejection region is given by $Z < -1.645$, the green area. The green area now is equal α.

3. The test statistic is given by $z = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$

4. The above test statistic is computed as $z = \frac{47.1-50}{12\ /\ 24}$, based on the information provided to us

by the sample data. Thus $Z = -1.1839$.

5. Since the calculated value of the test statistic, namely -1.1839, does not fall in the rejection region, then $H_o$is not rejected.

6. The conclusion is that μ is not less than 50. Therefore $\mu \geq 50$.

EXAMPLE

Using the information in Example above, test the above hypothesis there by the $P - value$ method**.**

Solution:

As in the steps we do not need to find a critical value for this method. The significance level was given to be 0.05. Following the steps as if it were the classical method, we find that the test statistic has the value of -1.1839. Let us find the p-value for this left-tailed test. The $p.value = P\ (Z < -1.18)$ for using the Standard Normal Table, we get $P - value = 0.1190$. It is greater than Alpha. We do not reject the null hypothesis. Hence we conclude that μ is not less than 50, therefore $\mu \geq 50$.

It is to be recalled that the two methods used above lead to the same conclusion. In case there is a contradiction between them, i.e. if you reject the Null hypothesis by using one of them while you did not reject the Null Hypothesis by using the other method. It is for sure you have made a mistake in one of them.

Case II: Testing on One Mean, μ When Population Variance, σ^2 Unknown

Since the population variance, σ^2 is not known, it is traditionally reasonable to ask about the sample size. This is based on the earlier presentations done in Chapter 3, when we compare the standard normal distribution with the Student's t-distribution. We found out there that when n, the sample size is large, usually n ≥ 30, is suitable to use the standard normal distribution for the test statistic involving one mean. Based on this discussion we have two cases to consider.

A) large Sample Size

In this part, since the sample size is large, n ≥ 30, we will use the same procedure for testing a statistical hypothesis about one population mean with one change. That change will take place in calculating the test statistic Z, when the standard deviation of the population is replaced by that of the sample. Based on that, the test statistic will be given by

$$z = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

The test that will be used is the Z-Test. The steps, in the classical method and the P-value method, are the same as above for this case of testing about one mean when the population variance is unknown.

B) Small Sample Size

In this part, since the sample size is small, $n < 30$, we will use the same procedure for testing a statistical hypothesis about one population mean with one change. That change will take place in replacing Z as the test statistic with T, where T will have a student t-distribution with degrees of freedom ν = n-1. Based on that, the test statistic will be given by

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

The t-distribution is another continuous distribution that is widely used in statistics. Because of that let us describe that distribution before getting to use it here.

In case I, above, we discussed testing a statistical hypothesis when the population variance was known. We now have another case on hand when the population variance, or the standard deviation, is unknown, and we have a small sample. The Z-Test discussed in Case I and Case II A) does not apply any more. We have to appeal for another distribution. This distribution is the t-distribution. In this case we do not replace σ by s anymore, and say that

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

is normally distributed, with mean 0 and variance 1. Instead 0 and this random variable follows **Student's t-distribution** with $n - 1$ degrees of freedom. So, let us have the properties of the t-distribution as listed below.

1. The t-distribution is controlled by its degrees of freedom. It is different for different degrees of freedom.
2. The mean of the distribution is 0, and it is symmetric about its mean.
3. As it was the case with the standard normal distribution, the total area under the curve is 1.
4. The horizontal axis acts like a horizontal asymptote, i.e., as t increases (or decreases) without any bound; the graph approaches the horizontal axis but never intersects it.
5. Compared with the standard normal distribution, and if drawn on the same scale, we find that the peak for standard normal distribution is higher than that of the t-distribution. This makes the tails for the t-distribution thicker than those for the standard distribution.
6. The variance for the t-distribution is greater than 1.
7. As the number of degrees of freedom increases (i.e. as the sample size n increases) the t-distribution gets closer to Z, the standard normal distribution. In this case the two curves for the two distributions will look almost alike. That is, and based on the law of large numbers, the estimator S, of σ, gets closer and closer

Classical Method Steps: For a specified value of the population mean μ0 we have (We are using the T-test on one mean):

1. State the null and alternative hypotheses:

There are three ways to set up the null and the alternative Hypotheses.

a) Equal hypothesis versus not equal hypothesis: $H_0 : \mu = \mu_0 \; versus \; H_1 : \mu \neq \mu_0$, , two-tailed test

b) At least versus less than: $H_0 : \mu \geq \mu_0 \; versus \; H_1 : \mu < \mu_0$, left-tailed test

c) At most versus greater than: $H_0 : \mu \leq \mu_0 \; versus \; H_1 : \mu > \mu_0$, right-tailed test

It is to be noted here that the two hypotheses are mutually exclusive sets on the real line for the values of the parameter, with the equal sign always set to go with the null hypothesis. This is so chosen in order to compute the value of the test statistics based on the null hypothesis being true.

1. Let α be the significance level for the test. Based on the three cases in step 1, we have the following three cases that will go along, for finding the critical value(s)and the rejection region(s)

a) For the two tailed test there are two critical values: $-t_{\alpha/2}$ & $t_{\alpha/2}$, and the critical region is given by $|T| > t_{\alpha}$.

b) For the left tailed test, there is the critical value of $-\ t_{\alpha}$, and the rejection region is given by $T\ <\ -t_{\alpha}$,

c) For the right tailed test, again, there is on critical value given by t_{α}, and the rejection region is $T\ >\ \ t_{\alpha}$

2. For the test statistic we have $0 = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$, where n is the sample size, s is the standard deviation of the sample, and T has the student t-distribution with n-1 degrees of freedom.

3. The above test statistic is computed based on the information provided to us by the sample data.

4. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

5. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE 4.5

A colony of laboratory mice consisted of several hundred animals. Their average weight was believed to 30 gm. An experiment was conducted to check on this belief. A simple random sample of 25 animals was taken. The average weight for this sample turned up to be 33 grams with a sample standard deviation of 5 gm. What conclusion can be made using if the level of significance will be 5%?

Solution:

Using the classical method steps for testing on one mean, using the T-test, we have:

1. $H_0 : \mu =\ \mu_0\ versus\ H_1 : \mu\ \neq\ \mu_0$Two-tailed test.

2. It is assumed that the significance level is 0.05. Since we have a two-tailed test, we have the

following critical values and the corresponding Rejection regions.Critical values are $\pm\ 2.064$, and the rejection regions are given by $|T|\ >\ 2.064$.

3. The test statistic is $t = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$.

4. Using the information given on hand, by calculating the above expression for T we have $T\ =\ 3$.

5. Since the value of the test statistic falls in the rejection region, we reject H0.

6. Based on the data provided the average weight is greater than 30.

5.2.3. Hypothesis Testing about One variance

Testing a statistical hypothesis about a population variance, or standard deviation, is surely different from testing about one population mean or proportion. The difference lies in the distribution of the statistic involved in the test. The statistic we are talking about here is the sample variance, or standard deviation, and its distribution. The test on one population variance is carried out based on the interest to check on the variability in the population. As it was the case in finding the confidence interval for one variance, we appealed to the random variable for the statistic given by

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

This Random variable has a chi-squared distribution with $n-1$ degrees of freedom, and S2 is the sample variance, σ^2 is the population variance, and n is the sample size.

Activity:

Let $\alpha = 0.05, and\ n = 15$. Find the following values $X^2_{\alpha/2}, X^2_{1-\alpha/2}, X^2_{\alpha}, X^2_{1-\alpha}$.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from a normal population with mean μ and variance σ^2. The steps for the statistical test on σ^2, for a specified value of the population variance σ^2 are as follows:

Classical Method Steps:

1. State the null and alternative hypotheses:

   There are three ways to set up the null and the alternative Hypotheses.

   a) Equal hypothesis versus not equal hypothesis: $H_0: \sigma^2 = \sigma_0^2 versus\ H_0: \sigma^2 \neq \sigma_0^2$, two-tailed test
   b) At least versus less than: $H_0: \sigma^2 \geq \sigma_0^2 versus\ H_0: \sigma^2 < \sigma_0^2$, left-tailed test
   c) at most versus greater than: $H_0: \sigma^2 \leq \sigma_0^2 versus\ H_0: \sigma^2 > \sigma_0^2$, right-tailed test

2. Let α be the significance level. Based on the three cases in step 1, we have the following three cases that will go along to calculate the critical values and the rejection regions.

   a) For the two tailed test there are two critical values: $X^2_{1-\frac{\alpha}{2}}$ & $X^2_{\frac{\alpha}{2}}$, and the critical regions are given by $X^2 > X^2_{\frac{\alpha}{2}}, or\ X^2 < X^2_{1-\frac{\alpha}{2}}$

   b) For the left tailed test, there is the critical value of $X^2_{1-\alpha}$, and the rejection region is given by $X^2 < X^2_{1-\alpha}$

c) For the right tailed test, again, there is on critical value given by X_{α}^{2} and the rejection region is $X^2 > X_{\alpha}^{2}$.

3. The test statistic is given by $X^2 = \frac{(n-1)s^2}{\sigma^2}$, where n is the sample size, s^2 is the sample variance, and X^2 has the Chi-Square distribution, with $n-1$ degrees of freedom.

4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a twotailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE

Consider a random sample of size 15 is taken from a normal population that yielded $s^2 = 3$. Test whether $\sigma > 1$, by using

a) the classical method,

b) the p-value method, by taking the significance level of 0.05.

Solution:

a) **Using the classical method step**s for testing the hypothesis on one variance we have:

1. $H_0: \sigma^2 \leq 1 \; versus \; H_1: \sigma^{\,2} > 1$, right-tailed test.
2. $\alpha = 0.05, n = 15, \nu = 14$, we see that the critical value is $X_{\;0.05}^{\;2} = 23.685$, Thus the rejection region is given by$X^2 > 23.685$.
3. The test statistic is $X^2 = \frac{(n-1)s^2}{\sigma^2}$
4. The calculated value of the test statistic is 126.
5. Since the calculated value of the test statistic is greater than 23.685, $H_0$is rejected.
6. We conclude that $\sigma^2 > 1$ and thus $\sigma > 1$.

b) By using the P-value Method, we have

1. $H_0: \sigma^2 \leq 1 \; versus \; H_1: \sigma^{\,2} > 1$, right-tailed test.
2. $\alpha = 0.053$.
3. The test statistic is

4. The calculated value of the test statistic is 126.

5. We need to calculate the p-value. From the Chi-square table it is hard to find exactly how much the p-value is. Never the less we can put a range on the p value. How? Since the degrees of freedom = 14, and our test statistic value is 126, we go look along 14 for a number close to, or greater than, 126. Doing so we find the largest value in the table along 14 degrees of freedom is 31.319, which fall under 0.005, in the Table. Hence the P-value is < 0.005 < 0.05, and H 0 is rejected. On the other hand, using a graphing calculator, we find that the p-value will be given precisely as

6. We conclude that $\sigma^2 > 1$ and thus $\sigma > 1$.

5.3.Hypothesis Testing Concerning Two Parameters

In this section we will discuss, and display, the procedures for testing a statistical hypothesis about two populations' parameters. In general, let those parameters be denoted by θ_1, and θ_2. Thus the steps go as follows:

Example 3.1

1. The auditor of a department store is thinking about establishing a new billing system for the store's credit customers. After a thorough financial analysis, the auditor determines that the new system will be cost-effective only if the mean monthly account is more than $170. A random sample of 400 monthly accounts is drawn, for which the sample mean is $178. The auditor knows that the accounts are approximately normally distributed with a standard deviation of $65. Can the auditor conclude from this that the new system will be cost-effective?

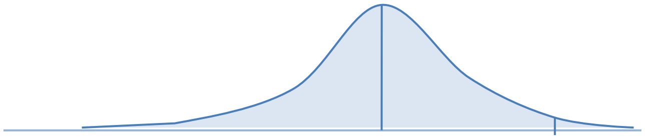
Solution

To conclude that the system will be cost-effective requires the manager to show that the mean account for all customers is greater than $170. Consequently, we set up the alternative hypothesis to express this circumstance:

H_1: $\mu > 170$ (Install new system)

If the mean is less than or equal to 170, then the system will not be cost-effective. The null hypothesis can be expressed as

H_0: $\mu \leq 170$ (Do not install new system)



$$\alpha = P(rejecting\ H0\ given\ that\ H0\ is\ true)$$

$$= P(\bar{x} > \bar{x}L\ \ given\ that\ H0\ is\ true)$$

$$= p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} > \frac{\overline{xL}-\mu}{\sigma/\sqrt{n}}\right) = p\left(Z > \frac{\bar{x}l-\mu}{\sigma/\sqrt{n}}\right) = \alpha$$

1. We know that $\sigma = 65, n = 400,\ \mu = 170$ and suppose that the manager chose α to be 5%, then the rejection region is $z_\alpha = z_{0.05} = 1.645$ from the table.

2. Z- Statistic is appropriate because population variance is known and identify the critical region.

$The\,critical\,region\,is\ Z_{cal} < -Z_{0.05} = -1.645$

$\Rightarrow (-1.645, \infty)\ \ is\,accep\tan ce\ region.$

3. Computations:

$$z_{cal} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{178-170}{65/\sqrt{400}} = 2.46$$

4. **Decision:**

Because 2.46 is greater than 1.645, reject the null hypothesis and conclude that there is enough evidence to infer that the mean monthly account is greater than \$170.

Example 5.2

The mean life time of a sample of 16 fluorescent light bulbs produced by a company is computed to be 1570 hours. The population standard deviation is 120 hours. Suppose the hypothesized value for the population mean is 1600 hours. Can we conclude that the life time of light bulbs is decreasing?

(Use $\alpha = 0.05$ and assume the normality of the population)

Solution:

Let $\mu = Population\,mean.\ ,\qquad \mu_0 = 1600$

Step 1: Identify the appropriate hypothesis

$H_0 : \mu = 1600 \qquad vs \qquad H_1 :\ \mu < 1600$

Step 2: select the level of significance, $\alpha = 0.05(given)$

Step 3: Z- Statistic is appropriate because population variance is known and identify the critical region.

$$\text{The critical region is } Z_{cal} < -Z_{0.05} = -1.645$$
$$\Rightarrow (-1.645, \infty) \text{ is acceptance region.}$$

Step 4: Computations:

$$Z_{cal} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1570 - 1600}{120/\sqrt{16}} = -1.0$$

Step 5: Decision

Accept H_0, since Z_{cal} is in the acceptance region. That is at 5% level of significance, we have no evidence to say that that the life time of light bulbs is decreasing, based on the given sample data

Example 5.3

A sample of 100 workers found the average overtime hours worked in the previous week was 7.8, with standard deviation 4.1 hours. Test the hypothesis that the average for all workers is 5 hours or less.

We can set out the five steps of the answer as follows:

1) $H_0 : \mu = 5$

$H_1 : \mu > 5$

2) Significance level, $a = 5\%$.

3) Critical value $z* = 1.64$.

4) Test statistic

$$Z_{cal} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{7.8 - 5}{\sqrt{\frac{4.1^2}{100}}} = 6.8$$

5) **Decision rule:**

6.8 > 1.64 so we reject H_0 in favor of H_1. Note that in this case we are dealing with the right-hand tail of the distribution (positive values of z and z*). Only high values of X reject H0.

5.3.1. HYPOTHESIS TEST FOR THE DIFFERENCE BETWEEN TWO MEANS

In order to test and estimate the difference between two population means, the statistician draws random samples from each of two populations.

The best estimator of the difference between two population means, $\mu_1 - \mu_2$, is the difference between two sample means, $\bar{x}_1 - \bar{x}_2$.

Hypothesis test:

1) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

2) $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 > 0$

3) H0 : $\mu_1 - \mu_2 = 0$ versus H1: $\mu_1 - \mu_2 < 0$

Classical Method Steps:

1. State the null and alternative hypotheses:

There are three ways to set up the null and alternative Hypotheses.

a) Equal hypothesis versus not equal hypothesis: $H_0 : \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$, two-tailed test.

b) At least versus less than: $H_0 : \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$, left-tailed test.

c) At most versus greater than: $H_0 : \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$, right-tailed test.

2. Let α be the significance level, and based on the three cases in step1, we have the following three cases that will go along for finding the critical values and rejection regions:

a) For the two-tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$, with the area on each is $\alpha/2$.

b) For the left tailed test, there is the critical value of $-Z_\alpha$, and the rejection region is given by $Z < -Z_\alpha$, with the area to the left of $-Z\alpha$ is equal to α.

c) For the right tailed test, again, there is on critical value given by Z_α, and the rejection region is $Z > Z_\alpha$, with the area to the right of Z_α, is equal to α.

3. The test statistic we have $Z_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

6. The interpretation and conclusion are due to answer the question that was raised

Case1: when $\bar{x}_1 - \bar{x}_2$ is normally distributed if the populations are normal or approximately normal, the sample sizes are large and /unknown.

The expected value of $\bar{x}_1 - \bar{x}_2$ is $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$

The variance of $\bar{x}_1 - \bar{x}_2$ is $V(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}$

Thus,

$$z_{cal} = \begin{cases} \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}, \text{for the two population variance are known} \\ \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}, \text{for the two population variance are unknown} \end{cases}$$

is a standard normal (or approximately normal) random variable.

Case 2: when the two population variance are unknown and small sample size

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

The test statistic for this case test statistic:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{pooled}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

Follows the t-distribution with the degree of freedom $n_1 + n_2 - 2$. Note that, the decision rule is similar to hypothesis testing about the population mean, μ.

Example 3.4:

In a market study for BGI, a local department store, you select a sample of 60 actual and potential clients to interview. Among the questions you wish to answer is whether the clients and non- clients differ in their incomes. The table below gives summary statistics. Noting the rather large difference in sample standard deviations, you decide that you must assume that the population standard deviations are unequal. Can you conclude that there is a difference in the mean incomes of clients and non- clients? Use α = 0.05.

	Clients	Non-clients
Mean income (in $ 1000s)	58.7	50.4
Standard deviation (in $ 1000s)	16.8	9.8
Number	27	33

1) $H_O: \mu_{clients} - \mu_{non-clients} = 0$ $\quad or \quad$ $H_O: \mu_{clients} = \mu_{non-clients}$

$H_a: \mu_{clients} - \mu_{non-clients} \neq 0$ $\quad$ $H_a: \mu_{clients} \neq \mu_{non-clients}$

2) Significance level, $a = 5\%$.

Critical value $df = n_1 + n_2 - 2 = 27 + 33 - 2 = 58$, then $t_{0.025,\ 58} = 2$.

3) Test statistic

$$t_{cal} = \frac{(58.7 - 50.4) - 0}{\frac{16.8^2(27-1) + 9.8^2(33-1)}{27+33-2}\sqrt{\frac{1}{27}+\frac{1}{33}}} = \frac{8.3}{179.51\sqrt{\frac{1}{27}+\frac{1}{33}}} = \frac{8.3}{46.58} = 0.178$$

$$df = n_1 + n_2 - 2 = 27 + 33 - 2$$

$$= 58$$

Step 5: Decision:

Accept H$_0$ or fail to reject H$_0$, since t_{cal} is in the acceptance region. That is at 5% level of significance, we have no evidence to say that there is a difference in the mean incomes of clients and non-clients based on the given sample data

5.3.2. HYPOTHESIS TESTING FOR TWO POLULATION PROPORTION

Suppose that two simple random samples were taken from two different populations with proportions of P_1 and P_2 for a certain property that we are interested in. The First sample is of size n_1 that produced x individuals having that interesting characteristic, while sample 2 is of size $n_2$that produced y individuals having the specified characteristic. Thus the sampling distribution of $\hat{P}_1 - \hat{P}_2$ where $\hat{P}_1 = x/n_1$ and $\hat{P}_2 = x/n_2$ is approximately normal with mean

To conduct the test, we assume each sample is large enough that the normal distribution will serve as a good approximation of the binomial distribution. The test statistic follows the standard normal distribution. We compute the value of z from the following formula:

$$Z_{CAL} = \frac{\left(\hat{P}_1 - \hat{P}_2\right) - (P_1 - P_2)}{\sqrt{\frac{P_C(1-P_C)}{n_1} + \frac{P_C(1-P_C)}{n_2}}}$$

This has an approximate standard normal distribution. With the respective sample proportions replacing the sample means and $p_c(1-p_c)$ replacing the two sample standard deviations. Where,

- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.
- $\hat{p}_1$ is the proportion in the first sample possessing the trait.
- $\hat{p}_2$ is the proportion in the second sample possessing the trait.
- $\hat{p}_c$ is the pooled proportion possessing the trait in the combined samples. It is called the pooled estimate of the population proportion and is computed from

The following formula.

$$P_C = \frac{x_1 + x_2}{n_1 + n_2}$$

Classical Method Steps: We are using the z-test on two proportions (2-prop Z-Test): 1. State the null and alternative hypotheses:

There are three ways to set up the null and the alternative Hypotheses.

a) Equal hypothesis versus not equal hypothesis: Ho: $P_1 = P_2$ versus : $P_1 \neq P_2$, two-tailed test.

b) At least versus less than: Ho: $P_1 \geq P_2$ versus : $P_1 < P_2$, left-tailed test.

c) At most versus greater than: Ho: $P_1 \leq P_2$ versus : $P_1 > P_2$, right-tailed test.

2. Let α be the significance level. Based on the three cases in step 1, we have the following three cases that will go along for finding the critical values and rejection regions.

a) For the two tailed test there are two critical values: $-Z_{\alpha/2}$ & $Z_{\alpha/2}$, and the critical region is given by $|Z| > Z_{\alpha/2}$

b) For the left tailed test, there is the critical value of $-Z_{\alpha}$, and the rejection region is given by $Z < -Z_{\alpha}$

c) For the right tailed test, again, there is on critical value given by Z_{α}, and the rejection region is $Z > Z_{\alpha}$

3. For the test statistic we have $Z_{CAL} = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{P_C(1-P_C)}{n_1} + \frac{P_C(1-P_C)}{n_2}}}$

4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

6. The interpretation and conclusion are due to answer the question that was raised.

Example 3.5:

A food manufacturer has two canning plants. The company's management wants to know whether the mean defect rate of a canned food from the new plant is different than that of the same canned food from the old plant. The canned food is packed in a carton that holds 24 cans. There are 500 cartons in each lot. The Table below gives the sample data obtained from each plant.

plant	Mean defect rate from each lot	Sample size

New	$\hat{p}_1 = 0.065$	$n_1 = 50$
Old	$\hat{p}_2 = 0.052$	$n_2 = 40$

1. The hypotheses:

   Ho: $P_1 - P_2 = 0$

   Ha: $P_1 - P_2 \neq 0$

2) Significance level, $a = 5\%$.

3) Critical value $z_{\alpha/2} = 1.96$.

4) Test statistic

$$P_C = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n1 + n2} = \frac{0.065 * 50 + 0.052 * 40}{50 + 40} = 0.059$$

Where, $\hat{P}_i = \frac{x_i}{n_i}, i = 1,2$

$$Z_{cal} = \frac{(0.065 - 0.052) - 0}{\sqrt{\frac{0.059(1-0.059)}{50} + \frac{0.059(1-0.059)}{40}}}$$

$$= \frac{0.013}{0.05} = 0.26$$

5) **Decision rule:**

Since Z$_{cal}$ is smaller than 1.96, so we cannot reject Ho. Based on the given data, the management confirms that the mean defect rate of the new plant is not statistically different from the mean defect rate of the old plant.

Exercise 1.

In recent years, the number of people who use the Internet to obtain political news has grown. Often the political Web sites ask Internet users to register their opinions by participating in online surveys. A Research Center conducted a survey of its own to learn about the participation of Republicans and Democrats in online surveys. The following sample data apply.

Political Party	Sample Size	Participate in Online Surveys
Republican	250	115
Democrat	350	98

a. Compute the point estimate of the proportion of Republicans who indicate they would participate in online surveys. Compute the point estimate for the Democrats.

b. What is the point estimate of the difference between the two population proportions?

c. At 95% confidence, what is the margin of error?

d. Representatives of the scientific polling industry claim that the profusion of online surveys can confuse people about actual public opinion. Do you agree with this statement? Use the 95% confidence interval estimate of the difference between the Republican and Democrat population proportions to help justify your answer.

Table 2: Four Commonly Used Confidence Levels and $\mathbf{Z_{a/2}}$

$\mathbf{1 - a}$	$\boldsymbol{a}$	$\boldsymbol{a/2}$	$\boldsymbol{Z_{a/2}}$
0.90	0.10	0.05	$Z_{\ 0.05} = 1.645$
0.95	0.05	0.025	$Z_{0.25} = 1.96$
0.98	0.02	0.01	$Z_{0.01} = 2.33$
0.99	0.01	0.005	$Z_{\ .005} = 2.575$

5.3.3. TESTS ABOUT TWO VARIANCES

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from a normal population with mean $\mu_1$and variance σ_1^2. Also let $Y_1, Y_2, \ldots, Y_m$ be another simple random sample from another normal population with mean $\mu_2$and variance σ_2^2. Here it is assumed that the populations' variances are unknown. In other words we let $X \sim N\ (\mu_1\ , \sigma_1^2\ )$, Y ~ N $(\mu_2, \sigma_2^2\ )$, respectively be the two independent variables with the assigned distribution that the samples were taken from. The steps for testing the statistical hypothesis on the ratio between the two variances need another random variable to be introduced.

Classical Method Steps: One more time the steps will go as follows:

1. State the null and alternative hypotheses:

There are three ways to set up the null and alternative Hypotheses.

a) Equal hypothesis versus not equal hypothesis: $H_0: \sigma_1^2 = \sigma_2^2 versus\ H_1: \sigma_1^2 \neq \sigma_2^2$ two-tailed test.

b) At least versus less than: $H_0: \sigma_1^2 \geq \sigma_2^2 versus\ H_1: \sigma_1^2 < \sigma_2^2$, left-tailed test.

c) at most versus greater than: $H_0: \sigma_1^2 \leq \sigma_2^2 versus\ H_1: \sigma_1^2 > \sigma_2^2$, right-tailed test

2. Let α be the significance level, and based on the three cases in step 1, we have the following three cases that will go along in order to find the critical values and the rejection region.

a) For the two tailed test there are two critical values:$F_{\ 1-\alpha/2}$& $F_{\alpha/2}$, and the critical regions are given by F

b) For the left tailed test, there is the critical value $F_{1-\alpha}$, and the rejection region is given by $F\ < F_{1-\alpha}$

c) For the right tailed test, again, there is on critical value given by F_{α}, and the rejection region is $F\ > F_{\alpha}$

3. The test statistic is $F = \frac{S_1^2}{S_2^2}$, where the larger of the two samples' variances is S_1^2. Then F has the F-distribution, with $n_1 - 1$ degrees of freedom for the numerator and $n_2 - 1$degrees of freedom for the denominator, with n_1 and n_2 are the samples' sizes.

4. The above test statistic is computed based on the information provided to us by the sample data.

5. The statistical decision will be made based on the case on hand whether we have a two tailed, a left-tailed or a right-tailed test, by comparing the computed value of the test statistic to the critical value based on the test being chosen.

6. The interpretation and conclusion are due to answer the question that was raised.

EXAMPLE

A company produces machined engine parts that are supposed to have a diameter variance no larger than 0.0002 (diameters are measured in inches). The company wishes to compare the variation in diameters produced by the company with the variation in diameters parts produced by another competitor. Our company had a sample of size n=10 that produced a sample variance S_1^2= 0.0003. In contrast, the sample variance of the diameter measurements for 20 of the competitor's parts was S_2^2 = 0.0001. Do the data provide sufficient information to indicate a smaller variation in diameters for the competitor? Test with α = 0.05, by using the classical method.

Solution:

1. $H_0: \sigma_1^2 \leq\ \sigma_2^2 versus\ H_1: \sigma_1^2 >\ \sigma_2^2$, right-tailed test

2. $\alpha\ =\ 0.05, and\ n_1\ = 10, n_1\ - 1 =\ 9, n_2\ = 20,\ n_2 - 1\ =\ 19$. Thus we have an F-distribution with 9 and 19 degrees of freedom for the numerator and denominator respectively, the CV is given by $F\ (9, 19;\ .05)\ =\ 2.42$. Moreover, the rejection region is given by $F\ >\ 2.42$.

3. The Test statistic is given by (since under $H_0: \sigma_1^2\ =\ \sigma_2^2$) $F = \frac{S_1^2}{S_2^2}$

4. The observed value of the test statistic, based on what we have on hand, is $F = 3$.

5. We see that $F > F\ (9, 19;\ .05)$, therefore at the $\alpha = 0.05$ level of significance, we reject the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ in favor of the alternative hypothesis, namely $H_1: \sigma_1^2 > \sigma_2^2$.

6. We conclude that the competitor produces parts with smaller variation in their diameters.

Summary

☞ Hypothesis testing is the set of procedures for deciding whether a hypothesis is true or false. When conducting the test we presume the hypothesis, termed the null hypothesis, is true until it is proved false on the basis of some sample evidence.

☞ If the null is proved false, it is rejected in favor of the alternative hypothesis. The procedure is conceptually similar to a court case, where the defendant is presumed innocent until the evidence proves otherwise.

☞ Not all decisions turn out to be correct and there are two types of error that can be made. A Type I error is to reject the null hypothesis when it is in fact true. A Type II error is not to reject the null when it is false.

☞ Choosing the appropriate decision rule (for rejecting the null hypothesis) is a question of trading off Type I and Type II errors. Because the alternative hypothesis is imprecisely specified, the probability of a Type II error usually cannot be specified.

☞ The rejection region for a test is therefore chosen to give a 5% probability of making a Type I error (sometimes a 1% probability is chosen). The critical value of the test statistic (sometimes referred to as the critical value of the test) is the value which separates the acceptance and rejection regions.

☞ The decision is based upon the value of a test statistic, which is calculated from the sample evidence and from information in the null hypothesis.

$$e.g\ Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

☞ The null hypothesis is rejected if the test statistic falls into the rejection region for the test (i.e. it exceeds the critical value).

☞ For a two-tail test there are two rejection regions, corresponding to very high and very low values of the test statistic.

☞ Instead of comparing the test statistic to the critical value, an equivalent procedure is to compare the Prob-value of the test statistic with the significance level. The null is rejected if the Prob-value is less than the significance level.

☞ The power of a test is the probability of a test correctly rejecting the null hypothesis. Some tests have low power (e.g. when the sample size is small) and therefore are not very useful.

CHAPTER SIX

CHI-SQUARED DISTRIBUTIONS

Introduction

Dear learner, χ^2 distribution allows us to establish confidence interval estimates for a variance, just as the Normal and t distributions were used in the case of a mean. Further, just as the Binomial distribution was used to examine situations where the result of an experiment could be either ‘success’ or ‘failure’, the χ^2 distribution allows us to analyses situations where there are more than two categories of outcome.

CHI-SQUARED DISTRIBUTIONS

The χ^2 distribution has a number of uses. In this chapter we make use of it in three ways:

- To calculate a confidence interval estimate of the population variance.
- To compare actual observations on a variable with the (theoretically) expected values.
- To test for association between two variables in a contingency table.

The use of the distribution is in many ways similar to the Normal and t distributions already encountered. Once again, it is actually a family of distributions depending upon one parameter, the degrees of freedom, similar to the t distribution. The number of degrees of freedom can have slightly different interpretations, depending upon the particular problem, but is often related to sample size in some way. Some typical χ^2 distributions are drawn in Figure 6.1 for different values of the parameter. Note the distribution has the following characteristics:

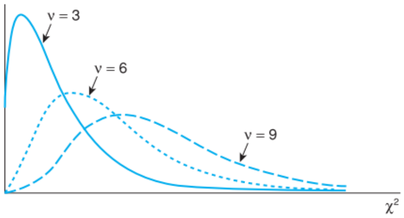


Figure 2: The □ distribution with different degrees of freedom

- it is always non-negative
- it is skewed to the right

- it becomes more symmetric as the number of degrees of freedom increases.

Using the χ^2distribution to construct confidence intervals is done in the usual way, by using the critical values of the distribution which cut off an area $\alpha/2$ in each tail of the distribution. For hypothesis tests, a rejection region is defined which cuts off an area α in either one or both tails of the distribution, whichever is appropriate.

Sample variance is an unbiased estimator of the population variance, $E(s^2) = \sigma^2$. To construct the confidence interval around this we need to know about the distribution of s^2. Unfortunately, this does not have a convenient probability distribution, so we transform it to

$$x^2 = \frac{(n-1)s^2}{\sigma^2}$$

which does have a x^2distribution, with $\nu = n - 1$ degrees of freedom.

To construct the 95% confidence interval around the point estimate we proceed in a similar fashion to the Normal or *t* distribution. First, we find the critical values of the x^2 distribution which cut off 2.5% in each tail. These are no longer symmetric around zero as was the case with the standard Normal and *t* distributions. Like the *t* distribution, the first column gives the degrees of freedom, so we require the row corresponding to ν = *n* − 1.

- ✓ For the *left-hand* critical value (cutting off 2.5% in the left-hand tail) we look at the column headed ‘0.975’, representing 97.5% in the right-hand tail.
- ✓ For the *right-hand* critical value we look up the column headed ‘0.025’ (2.5% in the right-hand tail).

Example:

Given a sample of size $n = 51$ yielding a sample variance $s^2 = 81$, we may calculate the 95% confidence interval for the population variance as follows. Since we are using the 95% confidence level the critical values cutting off the extreme 5% of the distribution are 32.36 and 71.42, from Table. We can therefore to find the interval

$$\frac{(n-1)\ s^2}{71.42} \le \sigma^2 \le \frac{(n-1)\ s^2}{32.36}$$

Substituting in the values gives

$$\frac{(51-1)81}{71.42} \leq \sigma^2 \leq \frac{(51-1)81}{32.36}$$

Yielding a confidence interval of [56.71, 125.15].

Note that if we wished to find a 95% confidence interval for the standard deviation we can simply take the square root of the result to obtain [7.53, 11.19].

The 99% CI for the variance can be obtained by altering the critical values. The values cutting off 0.5% in each tail of the distribution are (again from Table) 27.99 and 79.49. Using these critical values results in an interval [50.95, 144.69]. Note that, as expected, the 99% CI is wider than the 95% interval.

Chi-Square as a Test of Goodness of Fit

In this section, we will show how the chi-square statistic can be used to test the appropriateness of a distribution and its goodness of fit for a set of data. Goodness-of- fit tests are designed to study the frequency distribution to determine whether a set of data are generated from a certain probability distribution, such as the uniform, binomial, Poisson, or normal distribution.

Among the goodness-of-fit tests, the chi-square test is used to test the equality of more than two proportions if a probability distribution is assumed to be uniform.

If a manager is interested in knowing whether 4 different brands of painkillers are recommended equally often by doctors (or enjoy the same market shares), the manager can set up the following hypotheses:

H_0: Same market shares

H_1: Different market shares

To test this hypothesis, the manager can send out questionnaires to 1,000 doctors asking what painkiller they usually recommend to their patients. The responses can be tallied to obtain the observed sample frequency distribution. The tallied responses are called the observed frequency. If the null hypothesis of equal market shares is true, we would expect to see that roughly 250 doctors recommended each brand. This frequency distribution is called the expected frequency because it is anticipated when the null hypothesis is true. In applying the goodness-of-fit test, we compare the expected frequency with the observed frequency to determine whether the observed frequency conforms to the expected frequency—and hence supports the null hypothesis. If the null hypothesis is true, the frequencies for four brands of painkillers will be equal. Therefore, we can regard this example as a test of uniform distribution.

To take another example, many statistical inferences drawn in studying stock rates of return are based on the assumption that the rates of return of a stock follow a normal distribution. It should be interesting to test whether the rates of return are really generated from a normally distributed population. Here the null hypothesis is that the data are from a normally distributed population, and the alternative hypothesis is that the data are not from a normally distributed population. Again we perform the goodness-of-fit test by comparing the anticipated frequency distribution when the null hypothesis is true with the frequency distribution that is actually observed.

The chi-square statistic for determining whether the data follow a specific probability distribution is

$$X^2{}_{cal} = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

where

- $O_i \ = observed\ frequency$
- $e_i \ = expected\ frequency$
- $k \ = number\ of\ groups$
- $X^2_{k-1} = chi - square\ statistic\ with\ (k\text{–}1)\ degrees\ of\ freedom$

Example6.1:

A company manager wants to test his belief that four different categories of cars share the auto market equally. These four categories of cars are brand A, brand B, brand C, and imported cars. He sends out 2,000 questionnaires to car owners throughout the nation and receives the following responses:

Brand	*Number of owners (observed frequency)*
A	475
B	505
C	495

Imported	525
Total	**2,000**

Armed with these data, we can help him solve the problem. We first set up the hypotheses

H_0 : Same market shares (uniform distribution)

H_1 : Different market shares (nonuniform distribution)

When the null hypothesis is true, there should be 500 responses for each category of product. This implies that the expected frequency should be 500 for each category of product.

we divided the total sample into four groups. The frequencies of these four groups must add up to 2,000. This means that when any three groups' frequencies are known, the fourth group's frequency is also set. The number of degrees of freedom is therefore $(k-1)$, so here it is $4-1=3$. $X^2_{0.05,3}=7.81$

$$X^2{}_{cal}=\sum_{i=1}^{k}\frac{(O_i-e_i)^2}{e_i}$$

$$\sum_{i=1}^{4}\frac{(O_i-500)^2}{500}=\frac{1{,}300}{500}=2.6$$

We conclude that we do not have enough evidence to argue that the frequency distribution of different car brands is not uniformly distributed. In other words, the differences in market share among these four different brands of automobiles are not statistically significant.

Test of Association

Suppose we have a population consisting of observations having two attributes or qualitative characteristics say A and B, and if the attributes are independent then the probability of possessing both A and B is $P_A * P_B$

Where, P_A is the probability that a number has attribute A.

P_B is the probability that a number has attribute B.

☞ Suppose A has R mutually exclusive and exhaustive classes.

- B has C mutually exclusive and exhaustive classes

☞ The entire set of data can be represented using $R * C$ contingency table.

				B				
A	***B1***	***B2***	**.**	**.**	***Bj***	**.**	***Bc***	*Total*
A1	***O11***	***O12***			***O1j***		***O1c***	*R1*
A2	***O21***	***O22***			***O2j***		***O2c***	*R2*
.								
.								
Ai	***Oi1***	***Oi2***			***Oij***		***Oic***	*Ri*
.								
.								
Ar	***Or1***	***Or2***			***Orj***		***Orc***	
Total	*C1*	*C2*			*Cj*			*N*

The chi-square procedure test is used to test the hypothesis of independency of two attributes. For instance, we may be interested

☞ Whether the size of the family is independent of the level of education attained by the mothers.

☞ Whether there is association between father and son regarding boldness.

☞ Whether there is association between stability of marriage and period of acquaintance ship prior to marriage.

The χ^2 statistic is given by:

$$\chi^2{}_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] \sim \chi^2{}_{(r-1)(c-1)}$$

$$\begin{aligned} \mathit{Where}\quad O_{ij} &= \mathit{the\,number\,of\ units\,that\,belong\,to\,category\,i\ of\ A\,and\ j\,of\ B.} \\ e_{ij} &= \mathit{Expected\ frequency\,that\,belong\,to\,category\,i\ of\ A\,and\ j\,of\ B.} \end{aligned}$$

☞ The e_{ij} is given by :

$$e_{ij} = \frac{R_i * C_j}{n} \qquad \mathit{Where}\quad R_i = \mathit{the}\ i^{th}\,\mathit{row\,total.}$$

$$C_j = \mathit{the}\ j^{th}\ \mathit{column\ total.}$$

$$n = \mathit{total\ number\ of\ oservations}$$

Remark:

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} e_{ij}$$

The null and alternative hypothesis may be stated as:

The null hypothesis will specify that there is no relationship between the two variables. We state this in the following way:

***Ho**: The two variables are independent*

The alternative hypothesis specifies one variable affects the other, expressed as

***Ha**: The two variables are dependent*

Decision Rule:

Reject H_0 for independency at α level of significance if the calculated value of χ^2 exceeds the tabulated value with degree of freedom equal to $(r-1)(c-1)$.

$$\Longrightarrow \text{Reject } H_0 \; if \;\; \chi^2{}_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] > \chi^2{}_{(r-1)(c-1)} \; at \; \alpha$$

Examples:

4.2 The trustee of a company's pension plan has solicited the opinions of a sample of the 200 company's employees about a proposed revision of the plan. A breakdown of the responses is shown in the accompanying table. Is there enough evidence to infer that the responses differ between the three groups of employees?

Responses	***Workers***		
	Blue-collars	*White collars*	*Managers*
For	14	37	32
Against	31	59	27

Test the hypothesis that the three groups of workers are independent of the breakdown of responses. (Use 5% level of significance)

Solution:

H$_o$: There is no association between the breakdowns of responses the three groups of workers.

H$_a$: not H$_o$.

First calculate the row and column totals

$R_1 = 83, \ R_2 = 117, \ C_1 = 45, \ C_2 = 96, C_3 = 59$

Then calculate the expected frequencies (e$_{ij}$'s)

$$e_{ij} = \frac{R_i * C_j}{n}$$

$$\Rightarrow e_{11} = 18.675, e_{12} = 39.84, e_{13} = 24.485$$
$$e_{21} = 26.325, e_{22} = 56.16, e_{23} = 34.515$$

Obtain the calculated value of the chi-square.

$$\chi^2{}_{cal} = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right]$$
$$= \frac{(14-18.675)^2}{18.675} + \frac{(37-39.84)^2}{39.84} + ... + \frac{(27-34.515)^2}{34.515} = 6.3$$

Obtain the tabulated value of chi-square

$$\alpha = 0.05$$
$$Degrees of \ freedom = (r-1)(c-1) = 1*2 = 2$$
$$\chi^2_{0.05}(2) = 5.99 \ \ from\, table.$$

The decision is to reject H$_0$ since $\chi^2{}_{cal} > \chi^2_{0.05}(2)$

Conclusion: At 5% level of significance we have evidence to say there is association between breakdown of responses and the groups of workers, based on this sample data.

Exercise 6.3

The operations manager of a company that manufactures shirts wants to determine whether there are differences in the quality of workmanship among the three daily shifts. She randomly selects 600 recently made shirts and carefully inspects them. Each shirt is classified as either perfect or flawed, and the shift that produced it is also recorded. The accompanying table summarizes the number of shirts that fell into each cell. Do these data provide sufficient evidence to infer that there are differences in quality between the three shifts? $at\ \alpha = 0.05$

	Shift		
Shirt condition	**1**	**2**	**3**
Perfect	240	191	139
Flawed	10	9	11

UNIT SEVEN

ANALYSIS OF VARIANCE

Outlines

- ✓ Areas of application
- ✓ Comparison of the mean of more than two populations
- ✓ Variance test
- ✓ Areas of application

Introduction

We described the case where a large sample was selected from the population. We used the Z distribution (the standard normal distribution) to determine whether it was reasonable to conclude that the populations mean or proportion was equal to a specified value. We tested whether two population means are the same. We described methods for conducting tests of means where the populations were assumed normal but the samples were small (contained fewer than 30 observations). In that case the t distribution was used as the distribution of the test statistic. In this chapter we expand further our idea of hypothesis tests. We describe a test for variances and then a test that simultaneously compares several means to determine if they came from equal populations. The technique when F test is used to compare three or more populations is called the analysis of variance, and it is an extremely powerful and commonly used procedure. The analysis of variance technique determines whether differences exist between population means. The procedure works by analyzing the sample variance, hence the name.

7.1. The F Distribution

The probability distribution used in this chapter is the F distribution. It was named to honor Sir Ronald Fisher, one of the founders of modern-day statistics. This probability distribution is used as the distribution of the test statistic for several situations. It is used to test whether two samples are from populations having equal variances, and it is also applied when we want to compare several population means simultaneously. The simultaneous comparison of several population means is called analysis of variance (ANOVA).

Assumptions for the F Test for Comparing Three or More Means

1. The populations from which the samples were obtained must be normally or approximately normally distributed.
2. The samples must be independent of one another.
3. The variances of the populations must be equal.

The characteristics of the F distribution

1. The F distribution is continuous.
2. The F distribution cannot be negative.
3. The smallest value F can assume is 0.
4. It is positively skewed.
5. The long tail of the distribution is to the right-hand side. As the number of degrees of freedom increases in both the numerator and denominator the distribution approaches a normal distribution.
6. It is asymptotic.
7. As the values of X increase, the F curve approaches the X-axis but never touches it.

7.2.Comparison of several means

Another use of the F distribution is the analysis of variance (ANOVA) technique in which we compare three or more population means to determine whether they could be equal. To use ANOVA, we assume the following:

1. The populations follow the normal distribution.
2. The populations have equal standard deviations (σ).
3. The populations are independent.

When these conditions are met, F is used as the distribution of the test statistic. For a test of the difference among three or more means, the following hypotheses should be used:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ Vs.

H_1: At least one mean is different from the others

The degrees of freedom for this F test are, for nominator, $dfN = k - 1$, where k is the number of groups and for denominator, $dfD = n - k$ where n is the sum of the sample sizes of the groups $n = n_1 + n_2 + \cdots + n_k$. The sample sizes need not be equal.

The Steps for the ANOVA technique for comparing three or more means,

Step 1 State the hypotheses.

Step 2 Find the critical value. ${F_\alpha}^{(k-1,n-k)}$

Step 3 Compute the test value, using the procedure:

a. Find the mean and variance of each sample

b. Find the grand mean. The grand mean, denoted by $\bar{x}_{GM}$, is the mean of all values in the samples. $\bar{x}_{GM} = \frac{n_1\bar{x}_1+n_2\bar{x}_2+\cdots+n_k\bar{x}_k}{n_1+n_2+\cdots+n_k}$

c. Find the sum square between samples, denoted by SSB

$$SSB = \sum_{i=1}^{k} n_k(\bar{x}_i - \bar{x}_{GM})^2.$$

and the mean square between samples, denoted by ${s^2}_B$

$$MSB = {s^2}_B = \frac{\sum_{i=1}^{k} n_k(\bar{x}_i - \bar{x}_{GM})^2}{k-1} = \frac{SSB}{k-1}$$

d. Find the sum square within samples, denoted by SSW

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2.$$

and mean square within samples, denoted by ${s^2}_w$

$$MSW = {s^2}_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2}{n-k} = \frac{SSW}{n-k}$$

e. Find the F test value.

$$F = \frac{{s^2}_B}{{s^2}_w}$$

Step 4 Make the decision. Reject H_0 if $F > {F_\alpha}^{(k-1,n-k)}$

Step 5 Draw Conclusion.

The summarized table for analysis of variance is:

ANOVA summary table

Source	Sum of squares	df	Mean square	F
Between	SSB	$k-1$	$MSB = \frac{SSB}{k-1}$	$\frac{MSB}{MSW} = \frac{{s^2}_B}{{s^2}_w}$
Within (error)	SSW	$n-k$	$MSW = \frac{SSW}{n-k}$	
Total	SST	$n-1$		

$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{GM}\right)^2 = \text{SSB} + \text{SSW}.$

Example: A company wishes to purchase one of five different machines: A, B, C, D, or E. In an experiment designed to test whether there is a difference in the machines' performance, each of five experienced operators works on each of the machines for equal times. The following table shows the numbers of units produced per machine. Test the hypothesis that there is no difference between the machines at 5% significance level.

A	**68**	**72**	**77**	**42**	**53**
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Solution:

Step 1 $H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E\ Vs\ H_1$: At least one mean is different

Step 2 $F_{\alpha}{}^{(k-1,n-k)} = F_{0.05}{}^{(5-1,25-5)} = F_{0.05}{}^{(4,20)} = 2.87$

Step 3

a. $\bar{x}_A = 62.4,\ \bar{x}_B = 57.8,\ \bar{x}_C = 70.6,\ \bar{x}_D = 56,\ \bar{x}_E = 64$

b. $\bar{x}_{GM} = 62.16$

c. $SSB = \sum_{i=1}^{k} n_k(\bar{x}_i - \bar{x}_{GM})^2 = 658.16,\ MSB = s^2{}_B = \frac{SSB}{k-1} = \frac{658.16}{4} = 164.54$

d. $SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2 = 1883.20,\ MSW = s^2{}_w = \frac{SSW}{n-k} = \frac{1883.20}{20} =$ 94.16

e. $F = \frac{s^2{}_B}{s^2{}_w} = \frac{MSB}{MSW} = 1.75$

ANOVA summary table

Source	Sum of squares	df	Mean square	F
Between	658.16	4	164.54	1.75
Within (error)	1883.2	20	94.16	
Total	2541.36	24		

Step 4 Decision: since $F = 1.75 \ngtr 2.87$, Do Not Reject H_0

Step 5 Conclusion: at 5% level of significance we have enough evidence that there is no difference between the machines, based on the sample data.

7.3. Comparing Two Population Variances

The F distribution is used to test the hypothesis that the variance of one normal population equals the variance of another normal population. It also used to test assumptions for some statistical tests. Recall that, in the previous chapter when small samples were assumed, we used the t test to investigate whether the means of two independent populations differed. The F distribution provides a means for conducting a test regarding the variances of two normal populations.

Regardless of whether we want to determine if one population has more variation than another population or validate an assumption for a statistical test, we first state the null hypothesis. The null hypothesis could be that the variance of one normal population, ${\sigma_1}^2$, equals the variance of the other normal population, ${\sigma_2}^2$. The alternate hypothesis is that the variances differ. In this instance the null hypothesis and the alternate hypothesis are:

$$H_0: {\sigma_1}^2 = {\sigma_2}^2 \; vs \; H_1: {\sigma_1}^2 \neq {\sigma_2}^2$$

To conduct the test, we select a random sample of n_1 observations from one population, and a sample of n_2 observations from the second population. The test statistic is defined as:

$$F = \frac{{s_1}^2}{{s_2}^2} \sim {F_{\alpha/2}}^{(n_1-1,n_2-1)}$$

In order to reduce the size of the table of critical values, the larger sample variance is placed in the numerator. We made a decision that, reject H_0 if $F > {F_{\alpha/2}}^{(n_1-1,n_2-1)}$

Example: A soft-drink firm is evaluating an investment in a new type of canning machine. The company has already determined that it will be able to fill more cans per day for the same cost if the new machines are installed. However, it must determine the variability of fills using the new machines, and wants the variability from the new machines to be equal to or smaller than that currently obtained using the old machines. A study is designed in which random samples of 61 cans are selected from the output of both types of machines and the amount of fill (in ounces) is determined. The data are summarized in the following table

Summary Data for Canning Experiment			
Machine Type	Sample Size	Mean	Standard Deviation
Old	61	12.284	0.231
New	61	12.197	0.162

Do these data present sufficient evidence to indicate that the new type of machine has less variability of fills than the old machine?

Solution:

The hypothesis here is

$$H_0: {\sigma_{old}}^2 \leq {\sigma_{new}}^2 \; vs \; H_1: {\sigma_{old}}^2 > {\sigma_{new}}^2$$

The value of the test statistic is $F = \frac{{\sigma_{old}}^2}{{\sigma_{new}}^2} = \frac{(0.231)^2}{(0.162)^2} = 2.033$

Since the hypothesis is one-sided the critical value is ${F_\alpha}^{(n_{old}-1, n_{new}-1)} = {F_{0.05}}^{(61-1,61-1)} = {F_{0.05}}^{(60,60)} = 1.53$

Decision: since $F = 2.033 > 1.53$, Reject H_0

Conclusion: At 5% level of significance, the sample data support the statement that the new type of machine has less variability of fills than the old machine.

Example: A stockbroker at Critical Securities reported that the mean rate of return on a sample of 10 oil stocks was 12.6 percent with a standard deviation of 3.9 percent. The mean rate of return on a sample of 8 utility stocks was 10.9 percent with a standard deviation of 3.5 percent. At the .05 significance level, can we conclude that there is a difference in variation in the oil stocks utility stocks?

Solution:

The hypothesis here is

$$H_0: {\sigma_{os}}^2 = {\sigma_{us}}^2 \; vs \; H_1: {\sigma_{os}}^2 \neq {\sigma_{us}}^2$$

The value of the test statistic is $F = \frac{{\sigma_{os}}^2}{{\sigma_{us}}^2} = \frac{(3.9)^2}{(3.5)^2} = \frac{15.21}{12.25} = 1.24$

Since the hypothesis is one-sided the critical value is ${F_{\alpha/2}}^{(n_{os}-1, n_{us}-1)} = {F_{0.05/2}}^{(10-1,8-1)} = {F_{0.025}}^{(9,7)} = 4.82$

Decision: since $F = 1.24 \ngtr 4.82$, do not reject H_0

Conclusion: At 5% level of significance, we conclude that there is a **no** difference in variation in the oil stocks utility stocks, based on the sample data.

CHAPTER EIGHT

8. REGRESSION and CORRELATION ANALYSIS

General objectives

At the end of this Unit, you should be able to

- ✓ Understand what a simple linear regression and correlation
- ✓ Know Significance of regression and correlation analysis
- ✓ Use least square method to fit regression line of the dependent variable Y on a independent variable X
- ✓ Compute and interpret the simple, rank correlation coefficient between two variables and coefficient of determination.
- ✓ Understand the difference between simple and rank correlation coefficient

Introduction

Regression analysis, in the general sense, means the estimation of or prediction of the Unknown values of one variable from known values of the other variable .In Regression analysis there are two types of variables. The variable whose value is influenced or to be predicted is called dependent (regressed or explained) variable and the variable which influences the values or is used for prediction, is called independent variable (regressor or Predictor or explanatory). If the Regression curve is a straight line, we say that there is linear relationship between the variables under study, non-linear else where.

When only two variables are involved, the functional relationship is known as simple regression. If the relationship between the two variables is a straight line, it is known as simple linear regression; otherwise it is called as simple non-linear regression. When there are more than two variables and one of them is assumed dependent upon the other, the functional relationship between the variables is known as multiple regressions. More over, correlation analysis is concerned with mathematical measure of the extent or degree of relationship between two variables.

Example

Regression analysis is performed if one wants to know relationship between

1. Income –consumption

2. Sales of ice-cream –with temperature of the day
3. Industrial production and consumption of electricity
4. The yield of crops, amount of rainfall, type of fertilizer, humidity,..., etc.

8.1. Simple linear Regression

The simple linear regression of Y on X in the population is given by

$$Y=\alpha+\beta X+\varepsilon$$

Where, α = y intercept,

β = slope of the line or regression coefficient,

ε (Error term) = (y- $\hat{y}$).

Basic Assumptions of SLR

1. There is linear relation ship between dependent variable y and explanatory variable x
2. Expected value of error term is zero and its variance is constant (δ2) Hence error term is approximately normally distributed with mean zero and constant variance (δ^2).

The y-intercept α and the slope β are the population parameters. We generally obtain the estimates of $\alpha\ and\ \beta$ from the sample. The estimator of $\alpha\ and\ \beta$ are denoted by a and b, respectively. Thus the fitted regression line is

$$\hat{y} = a + \beta x$$

The values of a and b are obtained using the method of least squares. According to the principle of least squares, one should select a and b such that $\sum e^2$ will be as small as possible, that is, we minimize SSE= $\sum e^2 = S = \sum[y-(a+bx)]^2$

To minimize this function, first we take the partial derivatives of SSE with respect to a and b. Then the partial derivatives are equal to zero separately. These will result in the equations known as *normal equations.*

For the straight line, y= a+ bx the normal equations are

$$\sum y = na + bx$$

$$\sum xy = a\sum x + b\sum x^2$$

By solving these normal equations, we can get the values of a and b.

The best estimate of b is given by

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

And

The best estimate of α is given by

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

The regression line indicates the average value of the dependent variable Y associated with a particular value of the independent variable X. The slope b, hereafter referred to as regression coefficient which indicates the change in Y with a unit change in X.

Example

The following table gives the ages and blood pressure of 10 women

Age(x)	56	42	36	47	49	42	60	72	63	55
Blood pressure(y)	147	125	118	128	145	140	155	160	149	150

a) Determine the least square regression equation of blood pressure on age of women

b) Estimate the blood pressure of a women whose age is 45 years.

Solution

$\sum x =$ 522 $\sum y =$ 1417 $\sum xy =$75188 $\bar{x} =$ 52.2

$\sum x^2 =$ 28348 $\sum y^2 =$ 202493 $\bar{y} = 141.7$

The estimated value of a and b can be obtained as respectively

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 * 75188 - 522 * 1417}{10 * 28348 - (522)^2} = 1.11$$

The best estimate of α is given by

$$a = \bar{y} - b\bar{x} = 141.7 - (1.11)(52.5) = 83.76$$

a) The least square regression equation of blood pressure on age of women is given by

$\hat{y}$ =a + bx , where $\hat{y}$ is estimated blood pressure and x is age of woman

=83.76+1.11x

b) Estimated blood pressure of a woman whose age 45 years is given as follows.

When x= 45

$\hat{y}$ =a + bx

=83.76+1.11*45=133.71

Example2. From the following data obtain the regression equation of Y on X

Sales(X) :	91	97	108	121	67	124	51	73	111	57
Purchase(Y)):	75	75	69	97	70	91	39	61	80	47

Solution

n= 10 , $\sum x = 900$, $\sum y = 700$, $\sum xy = 66900$, $\sum x^2 = 87360$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 * 66900 - 900 * 700}{10 * 87360 - (900)^2} = 0.61$$

$$a = \bar{y} - b\bar{x} = \frac{1}{n}(\sum_{i=}^{n} y_i - b\sum_{i=1}^{n} x_i) = \frac{1}{10}(700 - 0.61 \times 900) = 15.1$$

$$\hat{y} = 15.1 + 0.61x$$

8.2. Correlation Analysis

Correlation analysis is concerned with measuring the strength (degree) of the relationship between two or more variables. It is used if we are interested in knowing the extent of interdependence between two or more variables.

8.2.1. Karl Pearson's coefficient of (simple) correlation

The Karl Pearson correlation coefficient denoted by $r(x, y)$ or r_{xy} or simply r, is defined as the ratio of the covariance between X and Y to the product of their standard deviations:

$$r = \frac{\text{cov}(x, y)}{\delta_x \delta_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The simplified formula used for computational purpose is

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

An increase in one variable may cause an increase in the other variable, or a decrease in one variable may cause decrease in the other variable. When the variables move in the same direction like this they are said to be positively correlated. The positive correlation may be termed as direct correlation. If a decrease in one variable causes an increase in the other variable or visa versa, the variables are said to be negatively correlated. The negative correlation may be termed as inverse correlation. In case the two variables are not at all related they are said to be independent or uncorrelated.

Example

i. Amount of rainfall and yield of crop(up to a point) has positive correlation

ii. Price and demand of a commodity has negative correlation.

Properties of simple correlation coefficient

- Coefficient of correlation lies between $-1 \leq r \leq 1$
- If $r = 0$ indicate that there is no linear relation ship between two variables.
- If $r = -1\,or\,+1$ indicate that there is perfect negative (inverse) 0r positive (direct) linear relationship between two variables respectively.
- A coefficient of correlation(r) that is closes to zero shows the relationship is quite weak, where as r is closest to $+1\,or\,-1$,shows that the relationship is strong.

Note that

❖ The strength of correlation does not depend on the positiveness and negativeness of r.

❖ The slope of simple linear regression (coefficient of regression) and correlation coefficient should be the same in sign.

The correlation between two variables is linear if a unit changes in one variable result in a constant change in the other variable. Correlation can be studied through plotting scattered diagrams

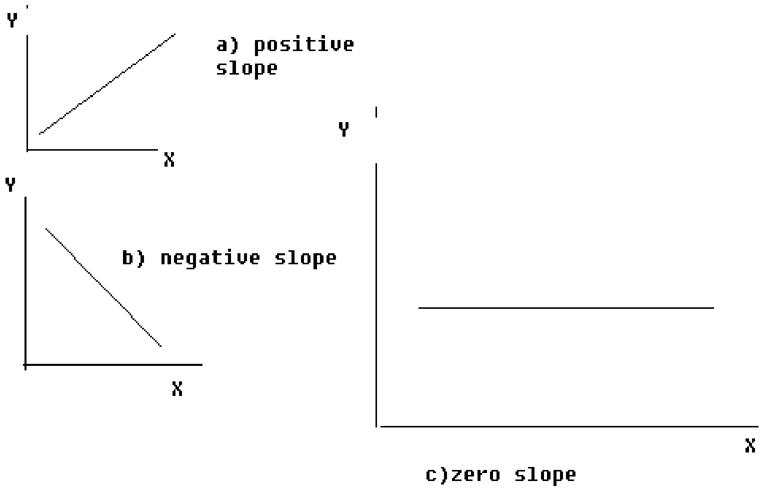


Figure: The slopes of liner regression lines.

Example1:

Calculate simple correlation coefficient (r) for the data on advertising and sales expenditure and interpret it.

Advertising(x):	39	65	62	90	82	75	25	98	36	78
Sales (y) :	47	53	58	86	62	68	60	91	51	84

Solution

$\sum x = 650$, $\sum y = 660$, $\sum xy = 45604$, $\sum x^2 = 47648$, $\sum y^2 = 45784$

$$r = \frac{10*45604 - 650*660}{\sqrt{(10*47648 - (650)^2)(10*45784 - (660)^2)}} = \frac{27040}{\sqrt{53980*22240}} = 0.78 \approx 0.8$$

There is strong positive (direct) linear relation ship between sales and advertisement since simple correlation coefficient approaches to 1

Example2.

Calculate and interpret simple correlation coefficient for data on blood pressure and age of 10 women

Age(x)	56	42	36	47	49	42	60	72	63	55

Blood pressure(y)	147	125	118	128	145	140	155	160	149	150

Solution

$\sum x = \ 522$ $\sum y = 1417$ $\sum xy = 75188$ $\bar{x} = 52.2$

$\sum x^2 = 28348$ $\sum y^2 = 202493$ $\bar{y} = 141.7$

$$r = \frac{10*75188 - 522*1417}{\sqrt{(10*28348-(522)^2)(10*202493-(1417)^2)}} = \frac{12206}{\sqrt{10996*17041}} = 0.89 \approx 0.9$$

- There is strong direct linear relationship between blood pressure and age of women Since correlation coefficient approaches to +1.

8.2.2. Coefficient of determination

It is defined as the proportion of the variation in the dependent variable Y that is explained, or accounted for, by the variation of the independent variable X. Its value is the square of the coefficient of correlation, thus we denote it by r^2 and it is usually expressed in the form of percentage.

Example1.compute and interpret coefficient of determination for above example on age and blood pressure.

Solution

Given that simple correlation coefficient between blood pressure and age is 0.89, hence coefficient of determination is square of the coefficient of correlation $(r^2)=(0.89)^2=79.21\%$ which implies that 79.21% variation in the blood pressure of women is accounted for, by the variation of the age of women.

8.2.3. Rank Correlation

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement, but can be arranged in serial order. This happens when we dealing with qualitative characteristics(attributes) such as beauty, efficient ,honest ,intelligence ,….etc., in such case one may rank the different items and apply the spearman method of rank difference for finding out the degree of relationship. The greatest use of this method(rank correlation) lies in the fact that one could use it to find correlation of qualitative variables, but since the method reduces the amount of labor of calculation ,it is sometimes used also where quantitative data is available. It is used when statistical series are ranked according to their magnitude and the exact size of individual item is not known. Spearman's correlation coefficient is denoted by r_s. If the ranks are given, denote the difference $R_{1i} - R_{2i}$ by di and obtain the total of d_i. Then the following formula is applied

$$r_s = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right]$$

If the actual data is given, rank it in ascending or descending order and follow the above procedures.

❖ Note that the values of rank correlation (r_s),.also lies between -1 and +1 inclusive.

Example1.

Ten competitors in a beauty contest are ranked by two judges in the following order. Compute and interpret opinion of two judges with regard to beauty out looking.

1st judge(x)	1	6	5	10	3	2	4	9	7	7
2nd judge(y)	3	5	8	4	7	10	2	1	6	9

Solution

d=(x-y)	-2	1	-3	6	-4	-8	2	8	1	-2
d^2	4	1	9	36	16	64	4	64	1	4

$$\sum d^2 = 203$$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*203}{10(10^2-1)} = -0.2303$$

➢ Hence the pair of judges has opposite (divergent) tastes for beauty since rank correlation coefficient is negative.

Example 2.

Calculate rank correlation coefficient between advertisement cost and sales from the following data and interpret it.

Advertisement(x): 39 65 62 90 82 75 25 98 36 78

Sales (y) :47 53 58 86 62 68 60 91 51 84

Solution

Rank of X:	**8**	**6**	**7**	**2**	**3**	**5**	**10**	**1**	**9**	**4**
Rank of Y:	10	8	7	2	5	4	6	1	9	3
$\boldsymbol{d = x - y}$**:**	-2	-2	0	0	-2	1	4	0	0	1
$\boldsymbol{d^2}$ **:**	4	4	0	0	4	1	16	0	0	1

$\sum d^2 = 30$

$$n = 10 \Rightarrow r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6\ x\ 30}{10(10^2 - 1)} = 1 - \frac{2}{11} = \frac{9}{11} = 0.82 \qquad \mathrm{r_s}$$

$= 0.82$

It implies that there is strong positive linear relation ship between advertisement cost and sales since correlation coefficient approaches to +1.

Exercise

1. Explain the meaning and significance of the concept of simple linear regression and correlation analysis.

2. How do you interpret a calculated value of Karl person's correlation coefficient? Discuss in particular the values of $r = 0, r = -1\ and\ +1$.

3. calculate and interpret the Karl Pearson's correlation coefficient for the ages of husband and wife for the data given below

Age of husband	23	27	28	29	30	31	33	35	36	39
Age of wife	18	22	23	24	25	26	28	29	30	32

4. Obtain the regression equation for costs related to age of cars for the following data on the ages of cars of a certain make and annual maintenance costs. Estimate the maintenance cost of cars when age of cars is 12 years.

Age of cars(in years)(x)	2	4	6	8
Maintenance cost(in 100 birr)(y)	10	20	25	30

5. Assuming that we conduct an experiment with eight fields planted with corn, amount of nitrogen fertilizer applied is given in kgs and corn yield per hectare, the resulting corn yield and amount of fertilizer applied shown in the table below .

Amount of Nitrogen fertilizer(kg)(x)	22	26	23	29	20	15	18	32
Corn yield/hectare(y)	120	130	160	180	120	110	118	190

a) Compute a linear regression equation of corn yield per hectare on amount of nitrogen fertilizer applied and also by using equation predict corn yield for a field treated with 34kgs of fertilizer.

Calculate and interpret simple correlation coefficient between amount of fertilizer applied and corn yield obtained, also obtain coefficient of determination

6 . The ranks of 15 students in two subjects A and B are given below. Calculate and interpret the rank correlation coefficient.

Rank in A (x)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Rank in B (y)	10	7	2	6	4	8	3	1	11	15	9	5	14	12	13

References

1] Anderson, D.R, Sweeney, D.j., and Williams, T.A.: *Statistics for Business and Economics*, 11th edition, 2011.

2] Lino Douglas A. and Robert D. mason. *Basic statistics for Business and Economics.*

3] Barrow, Michael (2009) *Statistics For Economics Accounting And Business Studies*, 5th Edition, Ft-Prentice Hall.

4] Earl Bowen: *Basic Statistics for Business and Economics*

5] Gupta C.B and Gupta, V.: *An Introduction to Statistics Methods*

6] Gupta C. P: *Statistics Methods*

7] H. Frank and S.C Althoen: *Statistics Concepts and Application*

8] Hanke and Reitsch: *Understanding Business Statistics*

9] J.G.V. an matre and G.H Gibreath: *Statistics for Management.*

10] Kinnfe Abraha: basic Statistics, *A textbook for Quantitative Method II*

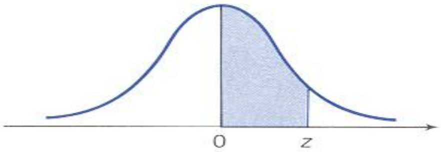
11] Levin, R.I., and Rubin, D.S.: *Statistics for Management*

12] Mason, R.D. and Lind, D.A.: *Statistics Techniques in Business and Economics*

13] Walpole, H.: *Introduction to Statistics*

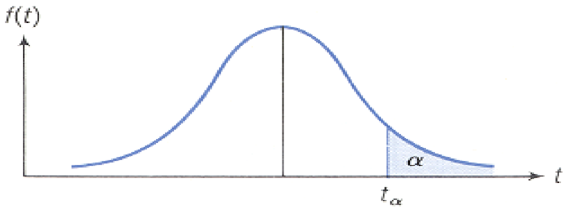
Appendices

Table 1: A standard normal distribution



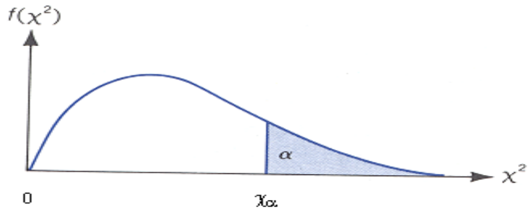
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table 2: Critical Values for Student's *t*



n	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.102	3.852	4.221
14	1.345	1.760	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.528	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
¥	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table 3:
Critical Values of chi-square statistic

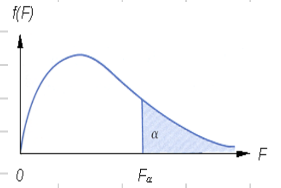


Degrees of freedom	$c^2_{.995}$	$c^2_{.990}$	$c^2_{.975}$	$c^2_{.950}$	$c^2_{.900}$	Degrees of freedom	$c^2_{.100}$	$c^2_{.050}$	$c^2_{.025}$	$c^2_{.010}$	$c^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	1	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	2	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	3	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	4	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	5	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	6	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	7	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	8	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	9	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	10	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	11	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	12	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	13	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	14	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	15	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	16	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	17	24.769	27.587	30.191	33.409	35.719
18	6.265	7.015	8.231	9.390	10.865	18	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	19	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	20	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	21	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	22	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	23	32.007	35.173	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	24	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	25	34.382	37.653	40.647	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	26	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	27	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	28	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	29	39.088	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	30	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	40	51.805	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	50	63.167	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	46.459	60	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	70	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.392	64.278	80	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	90	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.930	82.358	100	118.498	124.342	129.561	135.807	140.169

Table 4

Appendix C

A. Critical values for *F* statistic: $F_{.10}$



	ν_1	Numerator degrees of freedom																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Denominator degrees of freedom	1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
	2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
	4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
	6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
	7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
	8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
	9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
	11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
	13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
	14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
	16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
	17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
	18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
	19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
	21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
	22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
	23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
	24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
	25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
	26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
	27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
	28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
	29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
	40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
	120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
	∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

B. Critical values for *F* statistic: $F_{.05}$

	ν_1	Numerator degrees of freedom																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Denominator degrees of freedom	1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.21	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

C. Critical values for *F* statistic: $F_{.025}$

	ν_1	Numerator degrees of freedom																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Denominator degrees of freedom	1	647.80	799.50	864.20	899.60	921.80	937.10	948.20	956.70	963.30	968.60	976.70	984.90	993.10	997.20	1001.00	1006.00	1010.00	1014.00	1018.00
	2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	4.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
	26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
	27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
	28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
	29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
	∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

D. Critical values for *F* statistic: $F_{.01}$

	ν_1	Numerator degrees of freedom																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Denominator degrees of freedom	1	4,052.00	4,999.50	5,403.00	5,625.00	5,764.00	5,859.00	5,928.00	5,982.00	6,022.00	6056.00	6106.00	6157.00	6209.00	6235.00	6261.00	6287.00	6313.00	6339.00	6366.00
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	13	4.19	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00