Introduction to Statistics

Contents

1

Introduction

Studying statistics is great!!

1.1. History and Definition of Statistics

All of us are familiar with statistics in everyday life. As a discipline of study and research it has a short history, but as a numerical information it has a long antiquity. There are various documents of ancient times containing numerical information about countries (states), their resources and composition of the people. This explains the origin of the word statistics as a factual description of a state. The term 'statistics' is derived from the Latin word *status*, meaning *state*, and historically statistics referred to the display of facts and figures relating to the demography of states or countries. Generally, it can be defined in two senses: plural (as *statistical data*) and singular (as *statistical methods*).

Plural sense: Statistics are collection of facts (figures). This meaning of the word is widely used when reference is made to facts and figures on sales, employment or unemployment, accident, weather, death, education, etc. In this sense the word Statistics serves simply as data. But not all numerical data are statistics. In order for the numerical data to be identified as statistics, it must possess certain identifiable characteristics. Some of these characteristics are described as follows:

1. **Statistics are aggregate of facts**. Single or isolated facts or figures cannot be called statistics as these cannot be compared or related to other figures within the same framework. Accordingly, there must be an aggregate of these figures. For example, if a person says that "I earn Birr 30,000 per year", it would not be considered as statistics. On the other hand if we say that the average salary of a professor at our university is Birr 30,000 per year, then this would be considered

as statistics since the average has been computed from many related figures such as yearly salaries of many professors.

2. **Statistics, generally, are not the outcome of a single cause but affected by multiple causes**. There are a number of forces working together that affect the facts and figures. For example, when we say the crime rate in a certain city has increased by 15% over the last year, a number of factors might affect these change. These factors may be general level of economy such as economic recession, unemployment rate, extent of use of drugs, extent of legal effectiveness and so on. While these factors can be isolated by themselves, the effect of these factors cannot be isolated and measured individually. Similarly, a marked increase in food grain production in a certain country may have been due to combined effect of many factors such as better seeds, more extensive use of fertilizers, governmental and banking support, adequate rainfall and so on. It is generally not possible to segregate and study the effect of each of these forces individually.

3. **Statistics are numerically expressed**. All statistics are stated in numerical figures only. Qualitative statements cannot be called statistics. For example, such qualitative statements as 'Ethiopia is a developing country' or 'Jack is very tall' would not be considered as statistical statements. On the other hand, comparing per capita income of Ethiopia with that of Kenya would be considered statistical in nature. Similarly, Jack's height in numbers compared to average height in Ethiopia would also be considered as statistics.

4. **Statistical data are collected in a systematic manner for predetermined purpose**. The purpose and objective of collecting pertinent data must be clearly defined, decided upon and determined prior to data collection. Also the procedures for collecting data should be predetermined and well planned. These would facilitate the collection of proper and relevant data.

5. **Statistics are enumerated or estimated according to reasonable standard of accuracy**. There are basically two ways of collecting data. One is the actual counting or measuring, which is the most accurate way. The second way of collecting data is by estimation and is used in situations where actual counting or measuring is not feasible or where it involves prohibitive costs. Estimates, based on samples cannot be as precise and accurate as actual counts or measurements,

but these should be consistent with the degree of accuracy desired.

Singular sense: Statistics is the science that deals with the methods of data collection, organization, presentation, analysis and interpretation of data. It refers the subject area that is concerned with extracting relevant information from available data with the aim to make sound decisions. According to this meaning, statistics is concerned with the development and application of methods and techniques for collecting, organizing, presenting, analyzing and interpreting statistical data.

According to the singular sense definition of statistics, a statistical study (statistical investigation) involves five stages: collection of data, organization of data, presentation of data, analysis of data and interpretation of data.

1. **Collection of Data:** This is the first stage in any statistical investigation and involves the process of obtaining (gathering) a set of related measurements or counts to meet predetermined objectives. The data collected may be primary data (data collected directly by the investigator) or it may be secondary data (data obtained from intermediate sources such as newspapers, journals, official records, etc).

2. **Organization of Data:** It is usually not possible to derive any conclusion about the main features of the data from direct inspection of the observations. The second purpose of statistics is describing the properties of the data in a summary form. This stage of statistical investigation helps to have a clear understanding of the information gathered and includes *editing (correcting)*, *classifying* and *tabulating* the collected data in a systematic manner. Thus, the first step in the organization of data is *editing.* It means correcting (adjusting) omissions, inconsistencies, irrelevant answers and wrong computations in the collected data. The second step of the organization of data is *classification* that is arranging the collected data according to some common characteristics. The last step of the organization of data is presenting the classified data in tabular form, using rows and columns *(tabulation).*

3. **Presenting of Data:** The purpose of data presentation is to have an overview of what the data actually looks like, and to facilitate statistical analysis. Data presentation can be done using Graphs and Diagrams which have great memorizing effect and facilitates comparison.

4. **Analysis of Data:** The analysis of data is the extraction of summarized and comprehensive numerical description in order to reach conclusions or provide answers to a problem. The problem may require simple or sophisticated mathematical expressions.

5. **Interpretation of Data:** This is the last stage of statistical investigation. Interpretation involves drawing valid conclusions from the data collected and analyzed in order to make rational decision.

1.2. Classification of Statistics

Based on the *scope of the decision making*, statistics can be classified into two: Descriptive and Inferential Statistics.

Descriptive Statistics: refers to the procedures used to organize and summarize masses of data. It is concerned with describing or summarizing the most important features of the data. It deals only the characteristics of the collected data without going beyond it. That is, this part deals with only describing the data collected without going any further: that is without attempting to infer(conclude) anything that goes beyond the data themselves.

The methodology of descriptive statistics includes the methods of organizing (classification, tabulation, frequency distributions) and presenting (graphical and diagrammatic presentation) data and calculations of certain indicators of data like measures of central tendency and measures of variation which summarize some important features of the data.

Inferential Statistics: includes the methods used to find out something about a population, based on the sample. It is concerned with drawing statistically valid conclusions about the characteristics of the population based on information obtained from sample. In this form of statistical analysis, descriptive statistics is linked with probability theory in order to generalize the results of the sample to the population. Performing hypothesis testing, determining relationships between variables and making predictions are also inferential statistics.

Examples: Classify the following statements as *descriptive* and *inferential* statistics.

(a) The average age of the students in this class is 21 years.

(b) At least 5% of the killings reported last year in city X were due to tourists.

(c) Of the students enrolled in Haramaya University in this year 74% are male and 26% are female.

(d) The chance of winning the Ethiopian National Lottery in any day is 1 out of 167000.

(e) The demand for automobiles may decline next year in Europe.

(f) It has been continuously raining in Harar from Monday to Friday. It will continue to rain in the weekend.

1.3. Application of Statistics

In this modern time, statistical information plays a very important role in a wide range of fields. Today, statistics is applied in almost all fields of human endeavor.

In Scientific Research: Statistics plays an important role in the collection of data through efficiently designed experiments, in testing hypotheses and estimation of unknown parameters, and in interpretation of results.

In Industry: Statistical techniques are used to improve and maintain the quality of manufactured goods at a desired level. Statistical methods help to check whether a product satisfies a given standard.

In Business: Statistical methods are employed to forecast future demand for goods, to plan for production, and to evolve efficient management techniques to maximize profit.

In Medicine: Principles of design of experiments are used in screening of drugs and in clinical trials. The information supplied by a large number of biochemical and other tests is statistically assessed for diagnosis and prognosis of disease. The application of statistical techniques has made medical diagnosis more objective by combining the collective wisdom of the best possible experts with the knowledge on distinctions between diseases indicated by tests. Beside, statistical methods are used for computation and interpretation of birth and death rates.

In Literature: Statistical methods are used in quantifying an author's style, which is useful in settling cases of disputed authorship.

In Archeology: Quantitative assessment of similarity between objects has provided a method of placing ancient artifacts in a chronological order.

In Courts of Law: Statistical evidence in the form of probability of occurrence of certain events is used to supplement the traditional oral and circumstantial evidence in judging cases.

In Detective Work: Statistics helps in analyzing bits and pieces of information, which individually may appear to be unrelated or even inconsistent, to see an underlying pattern.

There seems to be no human activity whose value cannot be enhanced by injecting statistical ideas in planning and by using statistical methods for efficient analysis of data assessment of results for feedback and control.

1.4. Uses of Statistics

- **To reduce and summarize masses of data and to present facts in numerical and definite form**. Statistics condenses and summarizes a large mass of data and presents facts into a few presentable, understandable and precise numerical figures. The raw data, as is usually available, is voluminous and haphazard. It is generally not possible to draw any conclusions from the raw data as collected. Hence it is necessary and desirable to express these data in a few numerical values.

- **To facilitate comparison**: statistical devises such as averages, percentages, ratios, etc are used for this purpose.

- **For formulating and testing hypotheses**: For instance, hypothesis like whether a new medicine is effective in curing a disease, whether there is an association between variables can be tested using statistical tools.

- **For forecasting**: Statistical methods help in studying past data and predicting future trends.

1.5. Limitation of Statistics

- It does not deal with a single observation, rather, as discussed earlier, it only deals with aggregate of facts. For example, the marks obtained by one student in a class does not carry any meaning in itself, unless it is compared with a set standard or with other students in the same class or with his own marks obtained earlier.

- Statistical methods are not applicable to qualitative characters and cannot be coded in numerical values.

- Statistical results are true on average; i.e. for the majority of cases. Since statistics is not exact science, statistical conclusions are not universally true. That is, statistical laws are not universally true like the laws of physics, chemistry and mathematics.

- Statistics are liable to be misused or misinterpreted. This may be due to incomplete information, inadequate and faulty procedures during data collection and sample selection and mainly due to ignorance (lack of knowledge).

1.6. Variable

Variable is any phenomena or an attribute that can assume different values. The most important single distinguishing feature of a variable is that it varies; that is, it can take on different values. Based on the values that variables assume, variables can be classified as

1. **Qualitative variables:** A qualitative variable has values that are intrinsically nonnumerical (categorical).

   Example: Gender, Religion, Color of automobile, etc.

2. **Quantitative variables:** A quantitative variable has values that are intrinsically numerical.

   Example: Height, Family size, Weight, etc.

   - **Discrete variable:** takes whole number values and consists of distinct recognizable individual elements that can be counted. It is a variable that assumes a finite or countable number of possible values. These values are obtained by counting $(0, 1, 2, ...)$.

Example: Family size, Number of children in a family, number of cars at the traffic light.

- ▷ **Continuous variable:** takes any value including decimals. Such a variable can theoretically assume an infinite number of possible values. These values are obtained by measuring.
  Example: Height, Weight, Time, Temperature, etc.

  Generally the values of a variable can be obtained either by counting for discrete variables, by measuring for continuous variables or by making categories for qualitative variables.

Example: Classify each of the following as qualitative and quantitative and if it is quantitative classify as *discrete and continuous.*

1. Color of automobiles in a dealer's show room.
2. Number of seats in a movie theater.
3. Classification of patients based on nursing care needed (complete, partial or safer).
4. Number of tomatoes on each plant on a field.
5. Weight of newly born babies.

1.7. Measurement Scales

The level of measurement is one way in which variables can be classified. Broadly, this relates to the level of information content implicit in the set of values and how each value may be interpreted (mathematically) relative to other values on the variable - an issue which dictates how the variable can be used and interpreted in statistical analysis. Consider the following illustrations.

- ▷ Mr A wears 5 when he plays foot ball and Mr B wears 6 when he plays foot ball.

  Who plays better?
  What is the average shirt number?

- ▷ Mr A scored 5 in Statistics quiz and Mr B scored 6 in Statistics quiz.

  Who did better?

What is the average score?

Based on the number on the shirts it is not possible to judge, whether Mr B plays better. But by using the test score, it is possible to judge that Mr B did better in the exam. Also it is not possible to find the average shirt numbers (or the average shirt number is nothing) because the numbers on the shirts are simply codes but it is possible to obtain the average test score. Therefore, scales of measurement

- shows the information contained in the value of a variable.
- shows also that what mathematical operations and what statistical analysis are permissible to be done on the values of the variable.

Different measurement scales allow for different levels of exactness, depending upon the characteristics of the variables being measured. The four types of scales available in statistical analysis are

1. ***Nominal Scales*** of variables are those qualitative variables which show category of individuals. They reflect classification in to categories (name of groups) where there is no particular order or qualitative difference to the labels. Numbers may be assigned to the variables simply for coding purposes. It is not possible to compare individual basing on the numbers assigned to them. The only mathematical operation permissible on these variables is counting. These variables

    - have mutually exclusive (non-overlapping) and exhaustive categories.
    - no ranking or order between (among) the values of the variable.

    Example: Gender (Male, Female), Political Affiliation (Labour, Conservative,Liberal), Ethnicity (White, Black, Asian, Other), etc.

2. ***Ordinal Scales*** of variables are also those qualitative variables whose values can be ordered and ranked. Ranking and counting are the only mathematical operations to be done on the values of the variables. But there is no precise difference between the values (categories) of the variable.

    Example: Academic Rank (BSc, MSc, PhD), Grade Scores (A, B, C, D, F), Strength (Very Weak, Week, Strong, Very Strong), Health Status (Very Sick, Sick, Cured), Economic Status (Lower Class, Middle Class, Higher Class), etc.

3. ***Interval Scales*** of variables are those quantitative variables when the value of the variables is zero it does not show absence of the characteristics i.e. there is no true zero. Zero indicates lower than empty. For example, for temperature measured in degrees Celsius, the difference between 5℃ and 10℃ is treated the same as the difference between 10℃ and 15℃. However, we cannot say that 20℃ is twice as hot as 10℃ , i.e. the ratio between two different values has no quantitative meaning. This is because there is no absolute zero on the Celsius scale; 0℃ not imply 'no heat'.

4. ***Ratio Scales*** of variables are those quantitative variables when the values of the variables are zero, it shows absence of the characteristics. Zero indicates absence of the characteristics. All mathematical operations are allowed to be operated on the values of the variables.

   For instance, a zero unemployment rate implies zero unemployment. Thus, we can also legitimately say an unemployment rate of 20 percent is twice a rate of 10 percent or one person is twice as old as another. In the case of temperature, we can use the Kelvin scale instead of the Celsius scale: the Kelvin scale is a ratio scale because 0 Kelvin is 'absolute zero' (-273℃) and this does imply no heat.

2

Methods of Data Collection and Presentation

2.1. Types of Data

Research results or findings reveal information's that are obviously an output of properly and carefully collected relevant data, after they are being analyzed through legitimate data analysis instruments. So, data are always a base (or an input) for research. This implies that the quality of our study is heavily dependent on the quality of our data. Data can be collected from different sources which are generally grouped under two major categories, namely, primary and secondary sources of data. Thus, despite their nature (i.e., qualitative or quantitative, discrete or continuous, etc), data are necessarily from:

1. **Primary Data**: Primary data is the one which is collected by the investigator himself for the purpose of a specific inquiry or study. These data are those data collected for the first time either through direct observation or by enquiring individuals under the direct supervision and instruction of the researcher. Such data is original in character and is generated in surveys conducted by individuals or research institutions.

2. **Secondary Data**: When an investigator uses the data which has already been collected by others, such data is called *secondary data*. This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes. The secondary data can be obtained from journals, official reports, government publications, publications of professional and research organizations and so on.

Based on the role of time, data can be classified as *cross-sectional* and *time series*.

1. **Cross-sectional data**: is a set of observations taken at a point of time.

2. **Time series data**: is a set of observations collected for a sequence of time usually at equal intervals.

2.2. Methods of Data Collection

The first and foremost task in statistical investigation is data collection. Before the actual data collection, four important points should be considered. These are the purpose of data collection (why we need to collect data?), the data to be collected (what kind of data to be collected?), the source of data (where we can get the data?) and the methods of data collection (how can we collect this data?).

Once it is decided what type of study is to be made, it becomes necessary to collect information about the concerned body. This information has to be collected from certain individuals directly or indirectly. Such a technique is known as *survey method.* The survey methods are commonly used in social sciences, i.e., problems related to sociology, political science, psychology and various economic studies.

Another way of collecting data is *experimentation*, i.e., an actual experiment is conducted and then observations (measurements and counts) will be recorded. Such experimental studies are common in natural sciences; agriculture, biology, medical science, industry,...etc.

2.2.1. Questionnaire

The most common methods of data collection for survey are personal interview and self-administered questionnaire. In these and other methods of data collection, it is necessary to prepare a document, called questionnaire, which contains a number of questions to be answered and is used to record the responses.

Questionnaire is a form containing a cover letter that explains about the person conducting the survey and the objectives of the survey, and a set of related questions which will be answered by the respondents. One of the most important points in preparing it is that *all questions in it must have relevance to the objectives of the survey.* In short, the following points should be kept in mind while designing a questionnaire:

- ▷ Questions should be simple, short and easy to understand and they should convey one and only one idea. Technical terms should be avoided.

- ▷ Sensitive questions (questions of personal and financial nature) should be avoided. Such questions should be obtained indirectly, by constructing a set of ranges and must put at the last part.

  Examples: age ($0-25, 26-50, 51-75, > 75$), salary (below $200, 200-500, 500-1000, > 1000$).

- ▷ Leading questions should be completely avoided. If you ask person like "Do not you smoke?" the person will automatically say 'Yes I do not'.

- ▷ Answers to the questions should not require any calculation.

- ▷ Questions should be capable of objective answers.

2.2.2. Secondary Data

Secondary data should be used with utmost care. The investigator, before using these data, must observe that they possess the following characteristics.

1. **Reliability of Data**: The data collected from other source should be reliable enough to be used by the investigator. Determining and testing the reliability of secondary data is the most important as well as difficult task. Reliability can be tested by answering questions like:

   - ⇝ Who collected them?

   - ⇝ What were the sources of data?

   - ⇝ What methods were used to collect them?

   - ⇝ At what time were they collected?

2. **Suitability of Data**: Before using secondary data, they must be evaluated whether they could serve for another purpose other than the one for which they were collected. The suitability of data can be evaluated from the point of the nature and scope of investigation view.

3. **Adequacy of Data**: Adequacy can be tested by evaluating the data in terms of area coverage, level of accuracy, number of respondents participated and so on.

Once the above points are observed in the secondary data, it is ready to be used for further analysis.

2.3. Data Organization

It is almost impossible for management to deal with all the collected data in the raw form as it is in a haphazard and unsystematic form. In order to describe situations and make inferences about the population even to describe the sample, the data must be organized into some meaningful way.

2.3.1. Editing Data

Before further analysis, the collected data should be edited for completeness, consistency, accuracy and homogeneity.

Completeness: If the answer to some questions is missing, it becomes necessary to contact the person again and complete the missing information.

Consistency: Some information given by the respondent may not be compatible in the sense that an information furnished by the individual either does not justify some other information or is contradictory to earlier one.

Accuracy: It is of vital importance. If the data are inaccurate, the conclusions drawn from it have no relevance. If the investigator has either made a false report or the respondent has deliberately supplied the wrong information, editing will be of no use. In recent times, checks have been evolved to attain accuracy example by sending supervisors to check the work of investigators or reinvestigating a few respondents after a certain gap of time.

Homogeneity: To maintain homogeneity, the information sheets are checked to see whether the unit of information or measurement is the same in all the questionnaires. If differences are there, it has to be converted to the same unit during editing.

2.3.2. Classification of Data

The next important step towards organizing data is classification. Classification is the separation of items according to similar characteristics and grouping them into various groups.

Data may be classified into four broad classes:

1. **Geographical classification**. This classification groups the data according to location differences; places, areas or regions among the items. The geographical areas are usually listed in alphabetical order for easy reference.

2. **Chronological classification**. Chronological classification includes data according to the time period; i.e., weekly, monthly, quarterly, annually, ... in which the items under consideration occurred.

3. **Qualitative classification**. In this type of classification, the data is grouped together according to some distinguished characteristic or attribute such as religion, sex, nation and so on. This classification simply identifies whether a given attribute is present or absent in a given population.

4. **Quantitative classification**. It refers to the classification of data according to some characteristics that has been measured such as classification according to weight, height, income and so on.

2.3.3. Tabulation of Data

A table is a systematic arrangement of data in rows and columns, which is easy to understand and makes data fit for further analysis and drawing conclusions. Tabulation should not be confused with classification, as the two differ in many ways. Mainly the purpose of classification is to divide the data into homogenous groups whereas the data are presented into rows and columns in tabulation. Hence, classification is a preliminary step prior to tabulation.
A statistical table, in general, should have the following parts.

1. **Table Number**: Every table should be identified by a number. It facilitates easy reference. Whenever you refer to the table in the text, you can give the number of the table only.

2. **Title**: There should be a title at the top of every statistical table. The title should be clear, concise and adequate. The title should answer the questions : What is the data? where is the data? how is the data classified? and, what is the time period of data?

3. **Stub**: It is a title given to each row.

4. **Caption**: The caption labels the data presented in a column of the table. There may be sub-captions in each caption.

5. **Body**: The body of the table is the most important part. The information given in the rows and columns forms the body of the table. It contains the quantitative information to be presented.

6. **Footnote**: Any explanatory notes concerning the table itself, placed directly beneath the table, is called 'footnote'. The main purpose of footnote is to clarify some of the specific items given in the table or to explain the ambiguities, omissions, if any, about the data shown in the table.

7. **Source Note**: If the data is collected from secondary sources, a source note is given to disclose the sources from which the data is collected.

Example: Consider the following format.

Table 2.1: Title of your table.

	Caption 1	Caption 2	Caption 3
Stub 1	15	65	3.5
Stub 1	22	88	6.3
Stub 3	a*	78	5.3

* No caption is found for stub 3.

Though the format of a table has already been discussed, some guidelines for preparing a table are as follows:

- The table should contain the required number of rows and columns with stubs and captions and the whole data should be accommodated within the cells formed corresponding to these rows and columns.

- If the quantity is zero, it should be entered as zero. Leaving blank space or putting dash in place of zero is confusing and undesirable.

- The unit of measurement should either be given in parentheses just below the column's caption or in parentheses along with the stub in the row.

- If any figure in the table has to be specified for a particular purpose, it should be marked with an asterisk or another symbol. The specification of the marked figure should be explained at the beneath of the table with the same mark.

2.3.4. Frequency Distributions

The most convenient way of organizing numerical data is to construct a frequency distribution. Frequency distribution is the organization of raw data in table form, using classes and frequencies. Here the term ‘class’ is a description of a group of similar numbers in a data set while ‘frequency’ is the number of times a variable value is repeated. Hence, ‘class frequency’ is the number of observations belonging to a certain class.

There are three types of frequency distributions: categorical, ungrouped and grouped frequency distributions.

1. **Categorical Frequency Distribution**: the data is qualitative i.e. either nominal or ordinal. Each category of the variable represents a single class and the number of times each category repeats represents the frequency of that class (category).

   Example: The blood type of 22 students is given below. Construct categorical frequency distribution.

   $$\begin{array}{ccccccccccc} A & B & B & AB & O & A & O & O & B & AB & B \\ A & B & B & O & A & O & AB & A & O & O & AB \end{array}$$

Class (Blood type)	Frequency (no of students)
A	5
B	6
AB	4
O	7
Total	22

2. **Ungrouped Frequency Distribution:** A frequency distribution of numerical data (quantitative) in which each value of a variable represents a single class. The values of the variable are not grouped) and the number of times each value repeats represents the frequency of that class.

   Example: Number of children for 21 families is:

   2 3 5 4 3 3 2 3 1 0 4 3 2 2 1 1 1 4 2 2 2

Construct ungrouped frequency distribution.

Class (no of children)	Frequency (No of families)
0	1
1	4
2	7
3	5
4	3
5	1
Total	21

3. **Grouped (Continuous) Frequency Distribution:** A frequency of numerical data in which several values of a variable are grouped into one class. The number of observations belonging to the class is the frequency of the class.

Example: Consider age group and number of persons:

Class Limits	Class Boundaries	Frequency
1-25	0.5-25.5	20
26-50	25.5-50.5	15
51-75	50.5-75.5	25
76-100	75.5-100.5	10
Total		70

Basic Terms

Class Limits: the lowest and highest values that can be included in a class are called class limits. The lowest values are called lower class limits and the highest values are called upper class limits. For example: Class limit for the first class is 1-25, where 1 is the lower class limit and 25 is the upper class limit of the first class.

Class Boundaries: are class limits when there is no gap between the UCL of the first class and the LCL of the second class. The lowest values are called lower class boundaries and the highest values are called upper class boundaries. The class boundary for the first class 0.5-25.5 where the Lower class boundary is 0.5 and the Upper class boundary is 25.5. Note that the UCL of one class is the LCL of the next class.

Class Width: the difference between UCB and LCB of a class. It is also the difference between the lower limits of two consecutive classes or it is the difference between upper limits of two consecutive classes.

$$\begin{aligned} w &= UCB_i - LCB_i \\ &= LCL_i - LCL_{i-1} \\ &= UCL_i - UCL_{i-1} \\ &= CM_i - CM_{i-1} \end{aligned}$$

For the above example, $w = 25.5 - 0.5 = 26 - 1 = 50 - 25 = 25$.

Class Mark: is the half way between the class limits or the class boundaries.

$$cm_i = \frac{LCL_i + UCL_i}{2} = \frac{LCB_i + UCB_i}{2}$$

Relative Frequency

The absolute frequency distribution is a summary table in which the original data is condensed into groups and their frequencies, which is called absolute frequency distribution. But if a researcher would like to know the proportion or percentage of cases in each group, instead of simply, the number of cases, s/he can do so by constructing a relative frequency distribution table. The relative frequency distribution can be formed by dividing the frequency in each class of the frequency distribution by the total number of observations. It can be converted in to a percentage frequency distribution by simply multiplying each relative frequency by 100.

The relative frequencies are particularly helpful when comparing two or more frequency distributions in which the number of cases under investigation are not equal. The percentage distributions make such a comparison more meaningful, since percentages are relative frequencies and hence the total number in the sample or population under consideration becomes irrelevant.

Class Limits	Class Boundaries	Relative Frequency	Percentage Frequency
1-25	0.5-25.5	$20/70 = 0.2857$	28.57
26-50	25.5-50.5	$15/70 = 0.2143$	21.43
51-75	50.5-75.5	$25/70 = 0.3571$	35.71
76-100	75.5-100.5	$10/70 = 0.1429$	14.29
Total		1	100

Cumulative Frequency

The above frequency distributions tell us the actual number (percentage) of units in each class, it does not tell us directly the total number (percentage) of units that lie below or above the specified values of the classes. This can be determined from a cumulative frequency distribution. A cumulative frequency distribution displays the total number of observations above (below) a certain value. When the interest of the investigator focuses on the number of items below a specified value, then this specified value is the upper boundary of the class. It is known as *less than cumulative frequency distribution*. Similarly, when the interest lies in finding the number of cases above a specified value, then this value is taken as the lower boundary of the specified class and is known as *more than cumulative frequency distribution*.

Class Limits	Class Boundaries	Frequency	LCF	MCF
1-25	0.5-25.5	20	20	20+15+25+10=70
26-50	25.5-50.5	15	20+15=35	15+25+10=50
51-75	50.5-75.5	25	20+15+25=60	25+10=35
76-100	75.5-100.5	10	20+15+25+10=70	10
Total		70		

Steps for the Construction of Grouped Frequency Distribution

(a) Arrange the data in an array form (increasing or decreasing order).

(b) Find the unit of measurement (u). u is the smallest difference between any two distinct values of the data.

(c) Find the Range(R). R is the difference between the largest and the smallest values of the variable.

$$R = max - min$$

(d) Determine the number of classes (k) using Sturge's rule.

$$k = 1 + 3.322 \log N$$

where N is the total number of observations.

(e) Specify the class width (w).

$$w = \frac{R}{k} = \frac{R}{1 + 3.322 \log N}$$

(f) Put the smallest value of the data set as the LCL of the first class. To obtain the LCL of the second class add the class width w to the LCL of the first class. Continue adding until you get k classes.

Let x be the smallest observation.
$LCL_1 = x$
$LCL_i = LCL_{i-1} + w$ for $i = 2, 3, ..., k$.

Obtain the UCLs of the frequency distribution by adding $w-u$ to the corresponding LCLs.

$UCLi = LCL_i + (w - u)$ for $i = 2, 3, ..., k$.

(g) Generate the class boundaries.

$LCB_i = LCL_i - \frac{u}{2}$ and $UCB_i = UCL_i + \frac{u}{2}$ for $i = 2, 3, ..., k$.

Example: Mark of 50 students out of 40.

16 21 26 24 11 17 25 26 13 27 24 26 3 27 23 24 15 22 22 12 22 29 18 22 28
25 7 17 22 28 19 23 23 22 3 19 13 31 23 28 24 9 20 33 30 23 20 8 21 24

Construct grouped frequency distribution for the given data set.

Solution:

The array form of the data (increasing order).

3 3 7 8 9 11 12 13 13 15 16 17 17 18 19 19 20 20 21 21 22 22 22 22 22 22
23 23 23 23 23 24 24 24 24 24 25 25 26 26 26 27 27 28 28 28 29 30 31 33

$u = 9 - 8 = 1, R = max - min = 33 - 3 = 30$
$k = 1 + 3.322 \log N = 1 + 3.322 \log 50 = 6.64 \approx 7$
$w = R/k = 30/6.64 = 4.5 \approx 5$
$w - u = 5 - 1 = 4$

Hence, the grouped frequency distribution for score of 50 student is:

Class Limits	Class Boundaries	Class Mark	Frequency
3-7	2.5-7.5	5	3
8-12	7.5-12.5	10	4
13-17	12.5-17.5	15	6
18-22	17.5-22.5	20	13
23-27	22.5-27.5	25	17
28-32	27.5-32.5	30	6
33-37	32.5-37.5	35	1
Total			50

Advantages and disadvantages of grouped frequency distributions:

- **Advantages**:
  - It condenses a large mass of data into a comparatively small table.
  - It attracts the attention of even a layman and gives him an insight into the nature of the distribution.
  - It helps for further statistical analysis, like central tendency, scatter, symmetry, of the data.
- **Disadvantages**:
  - In the grouped frequency distributions, the identity of the observations is lost. We know only the number of observations in a class and do not know what the values are.
  - Because the selection of the class width and the lower class limit of the first class are to a certain extent arbitrary, different frequency distributions may be constructed for the same data and hence may give contradictory impressions.

2.4. Methods of Data Presentation

This section covers methods for organizing and displaying data. Such methods provide summary information about a data set and may be used to conduct exploratory data analyses. The methods for providing summary information are essential to the development of hypotheses and to establishing the groundwork for more complex statistical analyses.

Though the data presented in the form of table yields a good information, they are not always good for all. Showing data in the form of a graph can make complex and confusing information appear more simple and straightforward.

Graphic Display of Data

Bar Chart

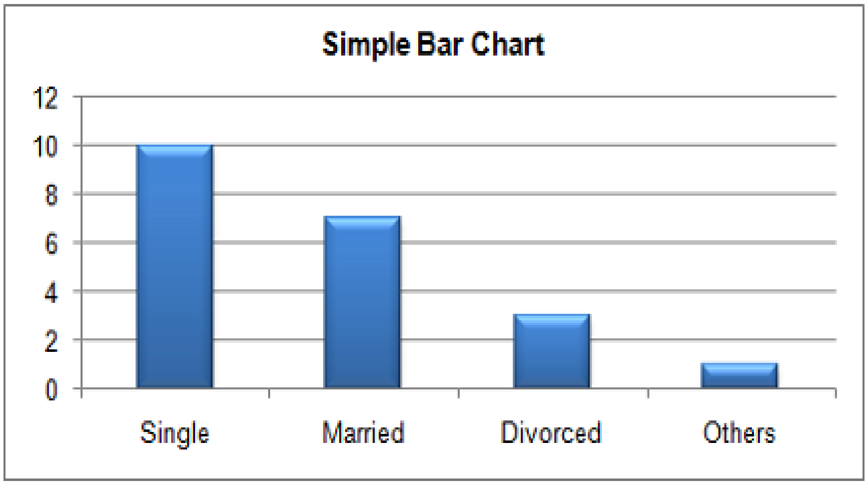
It is the simplest and most commonly used diagrammatic representation of a frequency distribution. It is the most common presentation for nominal, categorical or discrete data. It uses a serious of separated and equally spaced bars. The heights of the bars represent the frequency or relative frequency of the classes. But the width of the bars has no meaning; however, all the bars should be the same width to avoid distortion. And also the bars are separated by constant distance.

- ▷ **Simple Bar Chart**: is a diagram in which categories of a variable are marked on the X axis and the frequencies of the categories are marked on the Y axis. It is applicable for discrete variables, that is, for data given according to some period, places and timings. These periods and timings are represented on the base line (X axis) at regular interval and the corresponding frequencies are represented on the Y-axis.
  - The width of the bar represents nothing (it is meaningless), but it should be equal for all bars.
  - Each bar is separated by an equal space.
  - It can also represent some magnitude (on the Y axis) over time, space, groups, etc (on the X axis).

  Example:

  Construct simple bar chart for the following data.

Marital Status	Number of Individuals
Single	10
Married	7
Divorced	3
Others	1
Total	21



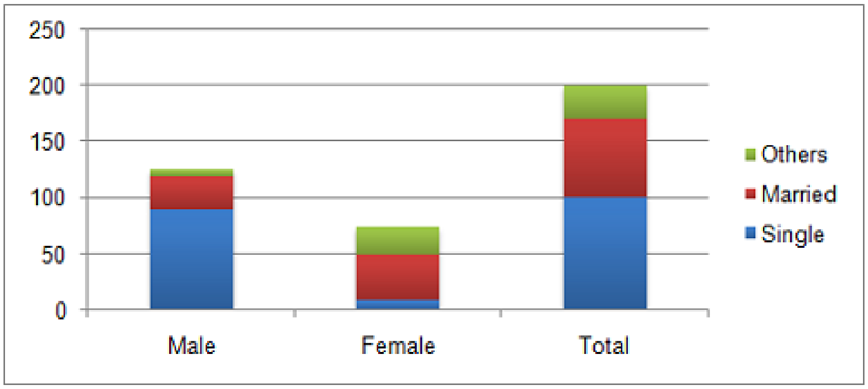
▷ **Component Bar Chart**: is used when there is a desire to show a total or aggregate is divided into its component parts. The bars represent total value of a variable with each total broken into its component parts and different colors are used for identification.

In such type of diagrams, a bar is subdivided into parts in proportion to the size of the subdivision. These subdivided rectangles are shaded differently by lines, dots and colors so that they will be very easy to compare the components. Sometimes the volumes of different attributes may be greatly different.

For making meaningful comparisons, the components of the attributes are reduced to percentages. In that case each attribute will have 100 as its maximum volume. This sort of component bar chart is known as percentage bar chart.

Example:
Consider the following table and the corresponding chart.

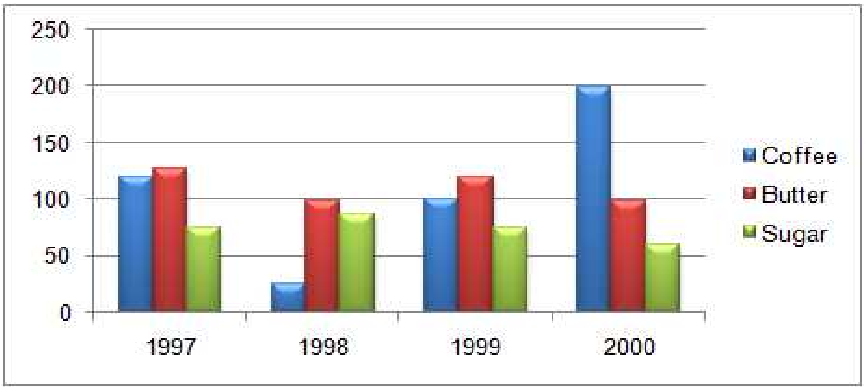
Marital	Male	Female	Total
Single	90	10	100
Married	30	40	70
Others	5	25	30



- ▷ **Multiple Bar Chart**: used to display data on more than one variable. In the multiple bars diagram two or more sets of inter-related data are interpreted.

  Example: Consider the following table which show the export of some item for a given country and the corresponding chart.

Year	Coffee	Butter	Sugar
1997	120	127	75
1998	25	98	87
1999	100	120	75
2000	198	98	60

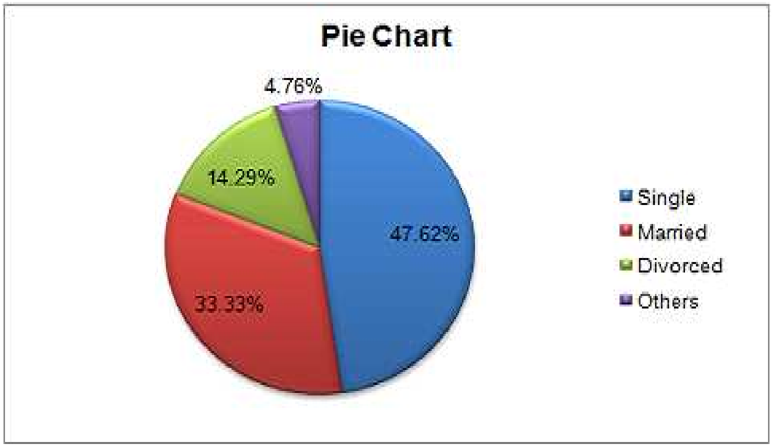


Pie Chart

Pie chart is popularly used in practice to show percentage break down of data. It is a circle representing a set of data by dividing the circle into sectors proportional to the number of items in the categories or it is a circle representing the total, cut into slices in proportional to the size of the parts that make up the total. It gives the proportional sizes of different data groups as slice of a pie or a circle.

Example: Construct pie chart for the following data.

Marital Status	Number of individuals	Percentage	Degree
Single	10	$\frac{10\times100}{21}=47.62$	$\frac{47.62\times360}{100}=171.43$
Married	7	$\frac{7\times100}{21}=33.33$	$\frac{33.33\times360}{100}=119.99$
Divorced	3	$\frac{3\times100}{21}=14.29$	$\frac{14.29\times360}{100}=51.44$
Others	1	$\frac{1\times100}{21}=4.76$	$\frac{4.76\times360}{100}=17.14$
Total	21	100	360



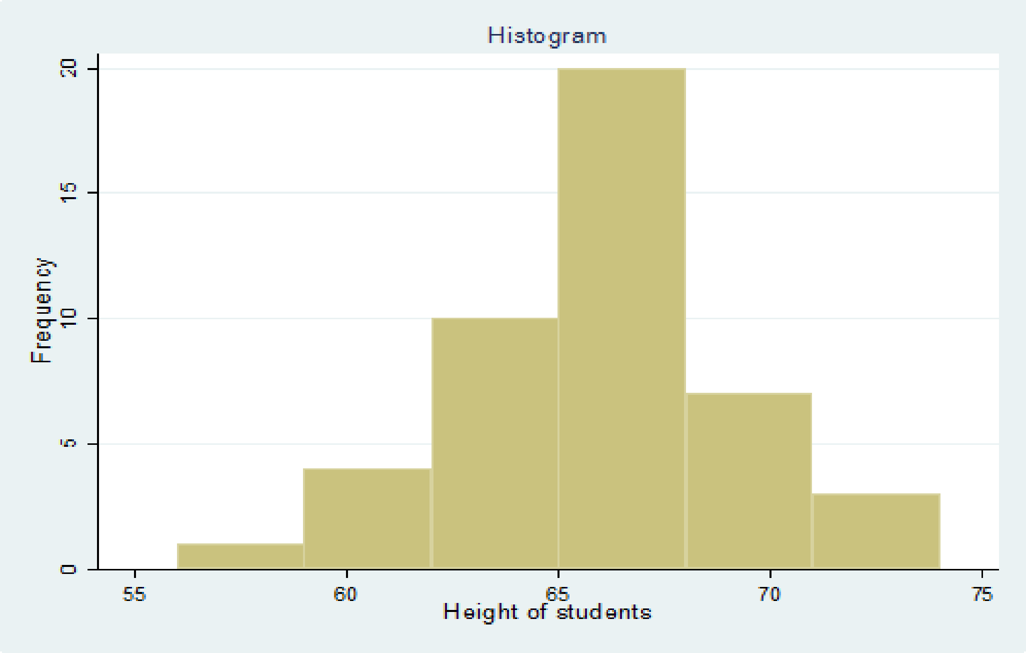
Histogram

Histogram is the most common graphical presentation of a frequency distribution for *numerical data*. It uses a series of adjacent bars in which the width of each bar represents the class width and the heights represent the frequency or relative frequency of the class. It is used for grouped data in which the class boundaries are marked on the X axis and the frequencies are marked along the Y axis.

Example:

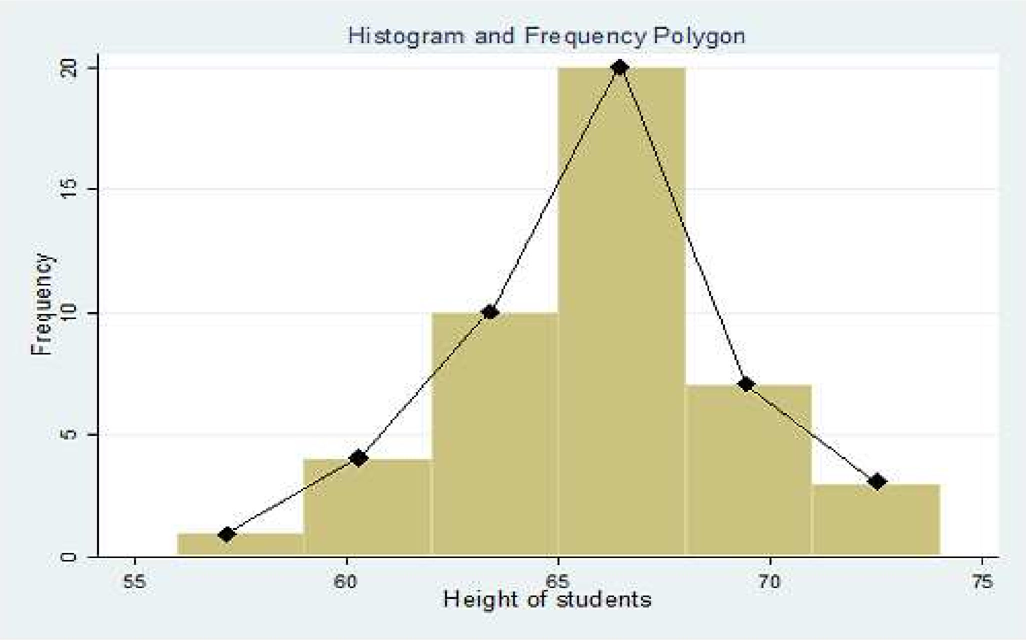
In the following, the heights of 45 female students at Haramaya University are recorded to the nearest inch. Construct a histogram by hand first. Check your result by using any statistical package.

67	67	64	64	74	61	68	71	69	61	65	64	62	63	59
70	66	66	63	59	64	67	70	65	66	66	56	65	67	69
64	67	68	67	67	65	74	64	62	68	65	65	65	66	67



Frequency Polygon

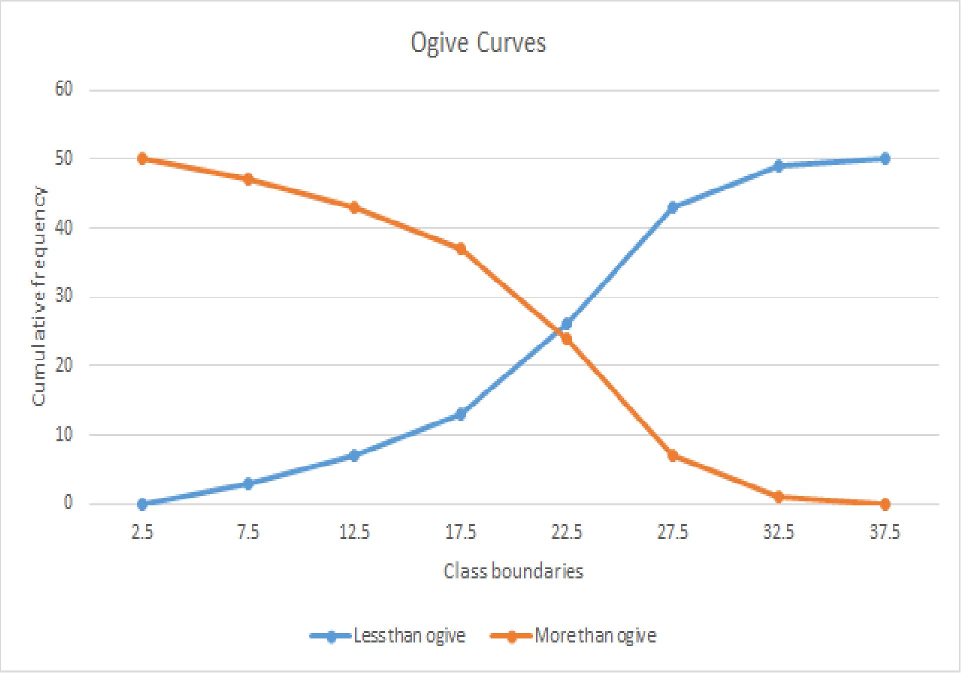
It is a graph that consists of line segments connecting the intersection of the class marks and the frequencies of a continuous frequency distribution. It can also be constructed from histogram by joining the mid-points of each bar.



Cumulative Frequency Curves (Ogive)

As there are two cumulative frequency distributions, there are two ogive (pronounced as "oh-jive") curves. These are the less than cumulative frequency which is a line graph joining the intersection points of the upper class boundaries and their corresponding less than cumulative frequencies and the more than cumulative frequency which is a line graph joining the intersection points of the lower class boundaries and their corresponding more than cumulative frequencies.

Example: Consider the following ogive curves for the marks of 50 students.



2.5. Exercises

1. A car salesman takes inventory and finds that he has a total of 125 cars to sell. Of these, 97 are the 2001 model, 11 are the 2000 model, 12 are the 1999 model, and 5 are the 1998 model. Which two types of charts are most appropriate to display the data? Construct one of the plots.

2. Define the following graphical methods and describe how they are used.

   a) Bar chart

   b) Histogram

   c) Relative frequency histogram

   d) Frequency polygon

   e) Ogive

3. The following are the ages of 30 patients in the emergency room of a hospital on a Friday night. Construct a histogram display from these data.

35	32	21	43	39	60
36	12	54	45	37	53
45	23	64	10	34	22
36	45	55	44	55	46
22	38	35	56	45	57

4. The final grades in Basic Statistics of 80 students at Haramaya University are recorded in the accompanying table.

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

Use these data to prepare:

(a) a frequency distribution.

(b) a relative frequency distribution.

(c) a cumulative frequency distribution.

(d) a histogram.

(e) a frequency polygon.

3

Measures of Central Tendency

3.1. Introduction

Usually the collected data is not suitable to draw conclusions about the mass from which it has been taken. Even though the data will be some what summarized after it is depicted using frequency distributions and presented by using graphs and diagrams, still we cannot make any inferences about the data since we have many groups. Hence, organizing a data into a frequency is not sufficient, there is a need for further condensation, particularly when we want to compare two or more distributions we may reduce the entire distribution into one number that represents the distribution we need. A single value which can be considered as a typical or representative of a set of observations and around which the observations can be considered as centered is called an average (or average value or center of location). Since such typical values tend to lie centrally within a set of observations when arranged according to magnitudes; averages are called *measures of central tendency (MCT).*

3.2. Objectives of MCT

- *To condense a mass of data in to one single value.* That is to get a single value which is best representative of the data (that describes the characteristics of the entire data). Measures of central tendency, by condensing masses of in to one single value enable us to get an idea of the entire data. Thus one value can represent thousands of data even more.

- *To facilitate comparison.* Statistical devices like averages, percentages and ratios used for this purpose. Measures of central tendency, by condensing masses of in to one single value, facilitates comparison. For instance, to compare two classes A and B, instead of comparing each student result, which is practically infeasible, we can compare the average mark of the two classes.

3.3. Desirable Properties of Good MCT

A measure of central tendency is good or satisfactory if it possesses the following characteristics.

1. It should be calculated based on all observations.
2. It should not be affected by extreme values.
3. It should be defined rigidly which means it should have a definite value.
4. It should always exist.
5. It should be easy to understand and calculate. It should not be subject to complicated and tedious calculations, though the advent of electronic calculators and computers has made it possible.
6. It should be capable of further algebraic treatment. By algebraic treatment, we mean that the measures should be used further in the formulation of other formulae or it should be used for further statistical analysis.

3.4. Summation Notation

Suppose we have variable x having successive values $x_1, x_2, ..., x_n$. The sum of these values can be written as $x_1 + x_2 + ... + x_n$. This can be written as using Greek letter $\sum$ as

$$x_1 + x_2 + ... + x_n = \sum_{i=1}^{n} x_i$$

By $\sum$ notation we can write

- ▷ $x_1^2 + x_2^2 + ... + x_n^2 = \sum_{i=1}^{n} x_i^2$
- ▷ $x_1y_1 + x_2y_2 + ... + x_ny_n = \sum_{i=1}^{n} x_iy_i$
- ▷ $(x_1 + x_2 + ... + x_n)^2 = (\sum_{i=1}^{n} x_i)^2$
- ▷ $\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{1}{x_4} = \sum_{i=1}^{4} \frac{1}{x_i}$

Rules of Summation

1. For two variables x and y we have
$$\sum_{i=1}^{n}(x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$$

2. If k is constant number, we have
$$\sum_{i=1}^{n} kx_i = k\sum_{i=1} x_i$$

3. For constant number k, we have
$$\sum_{i=1}^{n} k = nk$$

4. $\sum_{i=1}^{n}(x_i - k)^2 = \sum_{i=1}^{n} x_i^2 - 2k\sum_{i=1}^{n} x_i + nk^2$

From now onwards we will use $\sum x_i$ in place of $\sum_{i=1}^{n} x_i$ just for simplicity.

3.5. Types of Measures of Central Tendency

There are many types of measures of central tendency, each possessing particular properties and each being typical in some unique way. The most frequently encountered ones are

- ▷ Mean (computed average)
  - Arithmetic mean (simple arithmetic mean, weighted arithmetic mean and combined mean)
  - Geometric mean
  - Harmonic mean
- ▷ Mode (the most frequented value)
- ▷ Positional averages
  - Median
  - Quantiles (quartiles, deciles and percentiles)

3.6. Mean

3.6.1. Arithmetic Mean (AM)

Simple Arithmetic Mean

1. Suppose a variable x has observed values $x_1, x_2, ..., x_n$. The simple arithmetic mean denoted by $\bar{x}$ (for sample) and μ (for population) is the sum of these observations divided by the total number of observations. Symbolically,

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Simple AM is the most commonly used average.

2. Suppose the values $x_1, x_2, ..., x_n$ are accompanied by frequencies $f_1, f_2, ..., f_n$ respectively, then the simple AM is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_ix_i}{\sum f_i}$$

3. For data in grouped frequency distribution we use the class mark instead of each observed value and simple AM is given by

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + ... + f_nm_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_im_i}{\sum f_i}$$

where m_i is the class mark of the i^{th} class.

Example 1: The heights of 7 students selected from a class are given below in centimeter. $165, 160, 172, 168, 159, 170, 173$. Calculate the simple AM of heights.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_7}{7} = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{1167}{7} = 166.5\,cm$$

Example 2: The following is the frequency distribution of marks in Stat 1011 of 46 students (out of 20). Find the mean mark of this class.

Mark (x_i)	9	10	11	12	13	14	15	16	17	18	Total
No of students (f_i)	1	2	3	6	10	11	7	3	2	1	46
f_ix_i	9	20	33	72	130	154	105	48	34	18	623

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_ix_i}{\sum f_i} = \frac{623}{46} = 13.54$$

Example 3: Calculate the mean amount of yield of maize from the grouped frequency distribution given below.

Yield (in kg)	No of plots (f_i)	Class mark (m_i)	f_im_i
171-179	3	175	525
180-188	7	184	1288
189-197	12	193	2316
198-206	9	202	1818
207-215	4	211	844
216-224	4	220	880
225-233	1	229	229
Total	40		7900

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + ... + f_nm_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_im_i}{\sum f_i} = \frac{7900}{40} = 197.5\, kg\, per\, plot$$

Weighted Arithmetic Mean

It is an arithmetic mean used when all observations in data have unequal relative importance (technically termed as weight). Suppose $x_1, x_2, ..., x_n$ have weights $w_1, w_2, ..., w_n$ respectively, then weighted arithmetic mean ($\bar{x}_w$) is given by

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + ... + w_nx_n}{w_1 + w_2 + ... + w_n} = \frac{\sum w_ix_i}{\sum w_i}$$

Example: Semester grade point average (GPA) of a student is a good example of weighted arithmetic mean.

Course	Weights (Credit hours)	Grade (x)
Stat 281	4	B = 3
Math 261	4	B = 3
Math 224	3	C = 2
Phil 201	3	B = 3
Comp 201	3	C = 2

Calculate the GPA of this student?

$$GPA = \bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5}{w_1 + w_2 + w_3 + w_4 + w_5} = \frac{\sum w_ix_i}{\sum w_i} = \frac{45}{17} = 2.64$$

Combined Mean

If there are k different groups (having the same unit of measurement) with mean $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ and number of observations $n_1, n_2, ..., n_k$ respectively, then the mean of all the groups i.e. the combined mean is given by

$$\bar{\bar{x}} = \bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k} = \frac{\sum n_i\bar{x}_i}{\sum n_i}$$

Example: There are 49 students in a certain department. Among these 7 are seniors with average weight of 165 lbs, 9 are juniors with average weight of 160 lbs, 13 are sophomores with average weight of 152 lbs and 20 freshman with average weight of 150 lbs. Find the average weight of students in the department.

$$\begin{aligned}\bar{\bar{x}} &= \frac{n_s\bar{x}_s + n_j\bar{x}_j + n_{so}\bar{x}_{so} + n_f\bar{x}_f}{n_s + n_j + n_{so} + n_f}\\ &= \frac{7 \times 165 + 9 \times 130 + 3 \times 152 + 20 \times 150}{7 + 9 + 13 + 20}\\ &= 93.28\,lbs\end{aligned}$$

Properties of Arithmetic Mean

- If a constant k is added or subtracted from each value in a distribution, then the new mean will be

$$\bar{x}_{new} = \bar{x}_{old} \pm k$$

- If each value of a distribution is multiplied by a constant k, the new mean will be the original mean multiplied by k. That is,

$$\bar{x}_{new} = k\bar{x}_{old}$$

- Arithmetic mean can be calculated for any set of data (quantitative data), and it will be unique.We cannot calculate AM for open-ended grouped frequency distribution.
- It is highly affected by extreme values.
- It lends itself for further statistical analysis. For example, as combined mean.
- The algebraic sum of the deviations of each value from the arithmetic mean is zero. That is

$$\sum(x_i - \bar{x}) = 0$$

Example 1: The mean age of a group of 100 students was found to be 32.02 years. Later it was discovered that age of 57 was misread as 27. Find the correct mean.

Solution:

Let $\bar{x}_{cor}$ and $\bar{x}_{wr}$ are the correct and wrong means respectively. Thus, from the given problem $\bar{x}_{wr} = 32.02, n = 100, x_{wr} = 27$ and $x_{cor} = 57$.

$$\bar{x}_{wr} = \frac{(\sum x_i)_{wr}}{n}$$

$$(\sum x_i)_{wr} = \bar{x}_{wr} \times n$$

$$(\sum x_i)_{wr} = 32.02 \times 100 = 3202$$

$$(\sum x_i)_{cor} = (\sum x_i)_{wr} + x_{cor} - x_{wr}$$

$$(\sum x_i)_{cor} = 3202 + 57 - 27 = 3232$$

$$\bar{x}_{cor} = \frac{(\sum x_i)_{cor}}{n}$$

$$\bar{x}_{cor} = \frac{3232}{100} = 32.32 year$$

Example 2: The mean weight of 150 students in certain class is 60 kg. The mean weight of boys in the class is 70 kg and that of the girls is 55 kg. Find the number of boys and girls in the class.

Solution:

Let n_b and n_g are number of boys and girls in the class respectively. Further, suppose $\bar{\bar{x}} = 60kg$, $\bar{x}_b = 70kg$ and $\bar{x}_g = 55kg$ are the mean weight of both, boys and girls respectively.

$$n_b + n_g = 150 \tag{3.1}$$

Using combined mean formula

$$\bar{\bar{x}} = \frac{n_b\bar{x}_b + n_g\bar{x}_g}{n_b + n_g} = 60 = \frac{70n_b + 55n_g}{n_b + n_g}$$

$$n_g = 2n_b \tag{3.2}$$

Inserting equation (3.2) in equation (3.1) we obtain $n_b = 50$ and $n_g = 100$.

3.6.2. Geometric Mean (GM)

The geometric mean of n-positive numbers is the n^{th} root of their product. The geometric mean of $x_1, x_2, ..., x_n$ is given by the following for raw data, ungrouped and grouped frequency respectively.

$$GM = \sqrt[n]{x_1 \times x_2 \times ... \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$$GM = \sqrt[n]{x_1^{f_1} \times x_2^{f_2} \times ... \times x_n^{f_n}} = \sqrt[n]{\prod_{i=1}^{n} x_i^{f_i}}$$

$$GM = \sqrt[n]{m_1^{f_1} \times m_2^{f_2} \times ... \times m_n^{f_n}} = \sqrt[n]{\prod_{i=1}^{n} m_i^{f_i}}$$

We can also use logarithms to calculate GM

$$GM = \sqrt[n]{x_1 \times x_2 \times ... \times x_n} = (x_1 \times x_2 \times ... \times x_n)^{1/n}$$

$$\log GM = \frac{1}{n} \log(x_1 \times x_2 \times ... \times x_n)$$

$$\log GM = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$$

Taking antilog of both sides we get that

$$GM = anti\log\{\frac{1}{n}\log(x_1 + \log x_2 + ... + \log x_n)\} = anti\log(\frac{1}{n}\sum \log x_i)$$

If the variable values are measured as ratios, proportions or percentage and some values are larger in magnitude and others are small, then the geometric mean is a better representative of the data than the simple average. In a “geometric series”, the most meaning full average is the geometric mean. The arithmetic mean is very biased toward the large numbers in the series. The main disadvantage of geometric mean is that it cannot be calculated if one or more observations are zero or negative. It is also affected by extreme values but not to the extent of AM.

Examples

1. A given epidemic was spreading at the rate of 1.5 and 2.67 in two successive days. What is its average spread rate?

   Solution:

$$GM = \sqrt{x_1 \times x_2} = \sqrt{1.5 \times 2.67} = \sqrt{4.005} = 2.001$$

2. The price of a commodity increased by 5% from 1989 to 1990, 8% from 1990 to 1991 and by 77% from 1991 to 1992. Find the average price increase.

   Solution:

   For increment, take the base line value as 100% and then add the % increase so as to get the values in successive years.

Year	% increase	Value (x_i)	$\log x_i$
1989-1990	5	105	2.02
1990-1991	8	108	2.03
1991-1992	77	177	2.25
Total			$\sum \log x_i = 6.30$

   Then,

   $$GM = anti\log(\frac{1}{n}\sum \log x_i) = anti\log(\frac{1}{3} \times 6.30) = anti\log(2.1) = 125.89$$

   Therefore, the price increment is 25.89%.

3. A machine depreciated by 10% each in the first two years and by 40% in the third year. Find out the average rate of depreciation.

   Solution:

   Like the previous one, take the base line value of the machine as 100% and then deduct the % of depreciation so as to get the depreciated values in successive years.

Year	% depreciation	Value (x_i)	$\log x_i$
1	10	90	1.95
2	10	90	1.95
3	40	60	1.79
Total			$\sum \log x_i = 5.69$

   Then,

   $$GM = anti\log(\frac{1}{n}\sum \log x_i) = anti\log(\frac{1}{3} \times 5.69) = anti\log(1.70) = 50.12$$

   Therefore, the machine depreciated by is 49.88%.

3.6.3. Harmonic Mean (HM)

Harmonic mean is another specialized average which is useful in averaging variables expressed as rate per unit of time such as speed, number of units produced per day. Simple harmonic mean is the reciprocal of the arithmetic mean of the numbers.

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

- The simple HM is preferably used to calculate average speed for fixed distance, average price for fixed total cost, average time for fixed total distance.

For ungrouped frequency distribution,

$$HM = \frac{f_1 + f_2 + ... + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + ... + \frac{f_n}{x_n}} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$$

For grouped frequency distribution,

$$HM = \frac{f_1 + f_2 + ... + f_n}{\frac{f_1}{m_1} + \frac{f_2}{m_2} + ... + \frac{f_n}{m_n}} = \frac{\sum f_i}{\sum \frac{f_i}{m_i}}$$

The weighted HM of n non-zero observations $x_1, x_2, ..., x_n$ having weights $w_1, w_2, ..., w_n$ respectively is given by

$$HM_w = \frac{w_1 + w_2 + ... + w_n}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + ... + \frac{w_n}{x_n}} = \frac{\sum w_i}{\sum \frac{w_i}{x_i}}$$

- The weighted HM is used to compute mean speed to cover differing distances, mean prices when the total cost is not fixed, etc.

Examples

1. A driver travels for 3 days at speed of 48 km/hr for about 10 hrs, 40 km/hr for 12 hrs, 32 km/hr for 15 hrs respectively. What is the average speed of the driver in 3 days?

   Solution:

   Using $d_i = s_i \times t_i$; $i = 1, 2, 3$ the distance covered in three days is fixed, which is $480km$. So simple HM is appropriate to compute the average speed.

$$\begin{aligned} HM &= \frac{3}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}} = \frac{3}{\sum \frac{1}{x_i}} \\ &= \frac{3}{\frac{1}{48} + \frac{1}{40} + \frac{1}{32}} = \frac{3}{0.0771} \\ &= 38.91 km/hr \end{aligned}$$

2. A driver travelled for 3 days on first days he derived for 10 hrs at speed of 48 km/hr, on the second day for 12 hrs at 45 km/hr, on third day for 15 hrs at 40 km/hr. What is the average speed?

Solution:

Using $d_i = s_i \times t_i;\ i = 1, 2, 3$ the distance covered in each day is not fixed, which is $480km, 540km$ and $600km$ respectively. So weighted HM is appropriate to compute the average speed.

$$\begin{aligned} HM_w &= \frac{w_1 + w_2 + w_3}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \frac{w_3}{x_3}} = \frac{\sum w_i}{\sum \frac{w_i}{x_i}} \\ &= \frac{10 + 12 + 15}{\frac{10}{48} + \frac{12}{45} + \frac{15}{40}} = \frac{37}{0.892} \\ &= 41.48km/hr \end{aligned}$$

Some Empirical Relationship among AM, GM and HM

▷ The GM of two numbers x_1 and x_2 is equal to the GM of their AM and HM. That is,

$$GM = \sqrt{x_1 \times x_2} = \sqrt{AM \times HM}$$

▷ For n positive numbers $HM \leq GM \leq AM$.

3.7. Mode

The mode (modal value) of data set is the value that occurs *most frequently.* When two values occur with the same greatest frequency, each one is a mode and the data set is *bimodal.* When more than two values occur with the greatest frequency, each is a mode and the data set is said to be *multimodal.* When no value is repeated or values are equally repeated, we say that there is *no mode.*

Example 1: Find the modes of the following data sets.

► 5 5 5 3 1 5 1 4 3 5

► 1 2 2 2 3 4 5 6 6 6 7 9

► 1 2 3 6 7 8 9 10

In a frequency distribution, the mode is located in the class with highest frequency and that class is the modal class. Then the formula for mode is

$$\hat{x} = L_{\hat{x}} + \left[\frac{f_{\hat{x}} - f_{\hat{x}-1}}{(f_{\hat{x}} - f_{\hat{x}-1}) + (f_{\hat{x}} - f_{\hat{x}+1})}\right] w$$

where

$L_{\hat{x}}$ is the lower class boundary of the modal class,

$f_{\hat{x}}$ is the frequency of modal class,

$f_{\hat{x}-1}$ is the frequency of the class which precedes the modal class,

$f_{\hat{x}+1}$ is the frequency of the class which is successor of the modal class and

w is the class width of the modal class.

Example: Use the frequency distribution of heights in the following table to find the mode of height of the 100 male students at XYZ university and interpret the result.

Height (in)	Frequency (f_i)
59.5-62.5	5
62.5-65.5	18
65.5-68.5	42
68.5-71.5	27
71.5-74.5	8

Solution:

A class having the highest frequency is considered as a modal class. Thus the 3^{rd} class (65.5-68.5) is the modal class.

$$\begin{aligned}\hat{x} &= L_{\hat{x}} + \left[\frac{f_{\hat{x}} - f_{\hat{x}-1}}{(f_{\hat{x}} - f_{\hat{x}-1}) + (f_{\hat{x}} - f_{\hat{x}+1})}\right] w \\ &= 65.5 + \left[\frac{42-18}{(42-18)+(42-27)}\right] \times 3 \\ &= 65.5 + \left[\frac{24}{39}\right] \times 3 \\ &= 65.5 + 1.846 \\ &= 67.346\end{aligned}$$

Mode is not affected by extreme values and can be calculated for open-ended classes. But it often does not exist and is value may not be unique. In such case mode is ill-defined.

Properties of Mode

1. It is simple to calculate and easy to determine.
2. It is not based on all observations.

3. The mode can be used for both qualitative (such as religious preference, gender, political affiliation, etc) and quantitative data types.

3.8. Median

A median is a value which divides set of data in to two equal parts such that the number of observations below it is the same as the number of observations above it. It is the middle value when the values are arranged in order of increasing (or decreasing) magnitude. To find the median, first sort the values (arrange them in order), then use one of the following procedures.

1. If the number of values is *odd*, the median is the number that is located in the exact middle of the list.

$$\tilde{x} = \left(\frac{n+1}{2}\right)^{th} value$$

Example: What is the median of 180, 201, 220, 191, 219, 209 and 220.

Solution:

First we should have to sort the data: 180, 191, 201, 209, 219, 220, 220. Since $n = 7$ is odd

$$\tilde{x} = \left(\frac{4+1}{2}\right)^{th} value = 4^{th} value = 209$$

2. If the number of values is *even*, the median is found by computing the mean of the two middle numbers.

$$\tilde{x} = \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n}{2}+1\right)^{th} value}{2}$$

Example: What is the median of 62, 63, 64, 65, 66, 66, 68 and 78.

Solution:

First we should have to sort the data: $62, 63, 64, \underbrace{65, 66}, 66, 68, 78$. Since $n = 8$ is even

$$\begin{aligned}\tilde{x} &= \frac{\left(\frac{n}{2}\right)^{th} value + \left(\frac{n}{2}+1\right)^{th} value}{2} \\ &= \frac{4^{th} value + 5^{th} value}{2} \\ &= \frac{65+66}{2} = 65.5\end{aligned}$$

3. For grouped frequency distributions median is given by the formula

$$\tilde{x} = L_{\tilde{x}} + \left(\frac{\frac{n}{2} - F_{\tilde{x}-1}}{f_{\tilde{x}}} \right) w$$

where

$L_{\tilde{x}}$ is the lower class boundary of the median class,
$F_{\tilde{x}-1}$ is the less than cumulative frequency just before the median class,
w is the class width of the median class,
$f_{\tilde{x}}$ is the frequency of the median class and $n = \sum f_i$.

■ The median class is the class which include $\left(\frac{n}{2}\right)^{th}$ *value*.

Example: The following table shows a frequency distribution of grades on a final examination in college algebra for 120 students. Then, obtain median and interpret the results.

Grade	No of students
30-39	1
40-49	3
50-59	11
60-69	21
70-79	43
80-89	32
90-99	9

Solution:

First we should do the following.

Class limits	Class boundaries	Frequency	LCF
30-39	29.5-39.5	1	1
40-49	39.5-49.5	3	4
50-59	49.5-59.5	11	15
60-69	59.5-69.5	21	37
70-79	69.5-79.5	43	80
80-89	79.5-89.5	32	112
90-99	89.5-99.5	9	120

The class which includes $\left(\frac{n}{2}\right)^{th}$ $value = 60^{th} value$ is considered as the median class. Hence, the 5^{th} class is the median class.

$$\begin{aligned}\tilde{x} &= L_{\tilde{x}} + \left(\frac{\frac{n}{2} - F_{\tilde{x}-1}}{f_{\tilde{x}}}\right) w \\ &= 69.5 + \left(\frac{\frac{120}{2} - 37}{43}\right) \times 10 \\ &= 74.849\end{aligned}$$

Therefore, out of 120 students 60 of them scored less than 74.849 and 60 of them scored greater than 74.849 on college algebra examination.

Properties of the Median

1. It is an average of location, not the average of the values in the data set.
2. It is more affected by the number of observations than the extreme values.
3. Median can be calculated even in the case open-ended interval.

3.9. Quantiles

The median gives us a value which divides the data set in to two equal parts. There are also *other positional measures* that divide a given data set into more than two equal parts. These measures are collectively known as *quantiles.* Quantiles include quartiles, deciles and percentiles.

Quartiles are some three points that divide the array in to four parts in away each portion contains equal number of observations. The first, second and third points are called the first (Q_1), second (Q_2) and third (Q_3) quartiles respectively. 25% of the data fall below Q_1, 50% below Q_2 and 75% below Q_3 and

$$Q_1 \leq Q_2 \leq Q_3$$

Deciles are nine points that divide the array in to ten equal parts.The first, second, . . . , ninth deciles are denoted by $D_1, D_2, ..., D_9$ respectively. 10% of the data fall below D_1, 20% below D_2, . . . , 90% below D_9 and

$$D_1 \leq D_2 \leq \ldots \leq D_9$$

Percentiles are ninety nine points that divide the array in to 100 equal parts. They are denoted by $P_1, P_2, ..., P_{99}$. Always

$$P_1 \leq P_2 \leq \ldots \leq P_{99}$$

Methods of Finding Quantiles

1. For raw data and data in ungrouped frequency distribution. After arranging data in ascending order, we apply the following formula.

$$Q_i = \left(\frac{i(n+1)}{4}\right)^{th} value, i = 1, 2, 3$$

$$D_i = \left(\frac{i(n+1)}{10}\right)^{th} value, i = 1, 2, ..., 9$$

$$P_i = \left(\frac{i(n+1)}{100}\right)^{th} value, i = 1, ..., 99$$

Example: Given the data 420, 430, 435, 438, 441, 449, 490, 500, 510 and 515. Find

(a) all quartiles.

$$\begin{aligned} Q_1 &= \left(\frac{1 \times (10+1)}{4}\right)^{th} value = 2.75^{th} value \\ &= 2^{nd} value + 0.75(3^{rd} value - 2^{nd} value) \\ &= 430 + 0.75(435 - 430) \\ &= 433.75 \end{aligned}$$

$$\begin{aligned} Q_2 &= \left(\frac{2 \times (10+1)}{4}\right)^{th} value = 5.5^{th} value \\ &= 5^{th} value + 0.5(6^{th} value - 5^{th} value) \\ &= 441 + 0.5(449 - 441) \\ &= 445 \end{aligned}$$

$$\begin{aligned} Q_3 &= \left(\frac{3 \times (10+1)}{4}\right)^{th} value = 8.25^{th} value \\ &= 8^{th} value + 0.25(9^{th} value - 8^{th} value) \\ &= 500 + 0.25(510 - 500) \\ &= 502.5 \end{aligned}$$

(b) the 1^{st} and 7^{th} deciles.

$$\begin{aligned}
D_1 &= \left(\frac{1\times(10+1)}{10}\right)^{th} value = 1.1^{th} value \\
&= 1^{st}\, value + 0.1(2^{nd} value - 1^{st} value) \\
&= 420 + 0.1(430 - 420) \\
&= 421
\end{aligned}$$

$$\begin{aligned}
D_7 &= \left(\frac{7\times(10+1)}{10}\right)^{th} value = 7.7^{th} value \\
&= 7^{th} value + 0.7(8^{th} value - 7^{th} value) \\
&= 490 + 0.7(500 - 490) \\
&= 497
\end{aligned}$$

(c) the 40^{th} and 75^{th} percentiles.

$$\begin{aligned}
P_{40} &= \left(\frac{40\times(10+1)}{100}\right)^{th} value = 4.4^{th} value \\
&= 4^{th} value + 0.4(5^{th} value - 4^{th} value) \\
&= 438 + 0.4(441 - 438) \\
&= 439.2
\end{aligned}$$

$$\begin{aligned}
P_{75} &= \left(\frac{75\times(10+1)}{100}\right)^{th} value = 8.25^{th} value \\
&= 8^{th} value + 0.25(9^{th} value - 8^{th} value) \\
&= 500 + 0.25(510 - 500) \\
&= 502.5
\end{aligned}$$

2. For data in grouped frequency distribution.

$$Q_i = L_{q_i} + \frac{(\frac{in}{4} - F_{q_{i-1}})}{f_{q_i}} w$$

$$D_i = L_{d_i} + \frac{(\frac{in}{10} - F_{d_{i-1}})}{f_{d_i}} w$$

$$P_i = L_{p_i} + \frac{(\frac{in}{100} - F_{p_{i-1}})}{f_{p_i}} w$$

where

$L_{q_i}, L_{d_i}, L_{p_i}$ are the lower class boundaries of the classes containing the concerned quantile points,

$F_{q_{i-1}}, F_{d_{i-1}}, F_{p_{i-1}}$ are the LCF of the class which precedes the class containing the concerned quantile points,

$f_{q_i}, f_{d_i}, f_{p_i}$ are frequencies of classes containing the concerned quantile points and

w is the class width of a class containing the concerned quantile point.

Note

- Q_i is found in the class containing the $\left(\frac{in}{4}\right)^{th}$ observation.
- D_i is found in the class containing the $\left(\frac{in}{10}\right)^{th}$ observation.
- P_i is found in the class containing the $\left(\frac{in}{100}\right)^{th}$ observation.

Example: Calculate all quartiles, the 5^{th} and 8^{th} deciles, and the 30^{th} and 80^{th} percentiles for the students score data and interpret the results.

Class boundaries	Frequency (f_i)	LCF
10.5-14.5	4	4
14.5-18.5	7	11
18.5-22.5	8	19
22.5-26.5	10	29
26.5-30.5	12	41
30.5-34.5	7	48
34.5-38.5	8	56

Solution:

Q_1 is found in the 3^{rd} class (18.5-22.5) because this class include the $\left(\frac{1\times56}{4}\right)^{th} = 14^{th} value$

$$\begin{aligned} Q_1 &= L_{q_1} + \frac{(\frac{1\times56}{4} - F_{q_0})}{f_{q_1}} \times 4 \\ &= 18.5 + \frac{(\frac{1\times56}{4} - 11)}{8} \times 4 \\ &= 18.5 + 1.5 = 20 \end{aligned}$$

Q_2 is found in the 4^{th} class (22.5-26.5) because this class include the $\left(\frac{2\times 56}{4}\right)^{th} = 28^{th} value$

$$\begin{aligned} Q_2 &= L_{q_2} + \frac{(\frac{2\times 56}{4} - F_{q_1})}{f_{q_2}} \times 4 \\ &= 22.5 + \frac{(\frac{2\times 56}{4} - 19)}{10} \times 4 \\ &= 22.5 + 3.6 = 26.1 \end{aligned}$$

Q_3 is found in the 6^{th} class (30.5-34.5) because this class include the $\left(\frac{3\times 56}{4}\right)^{th} = 42^{th} value$

$$\begin{aligned} Q_3 &= L_{q_3} + \frac{(\frac{3\times 56}{4} - F_{q_2})}{f_{q_3}} \times 4 \\ &= 30.5 + \frac{(\frac{3\times 56}{4} - 41)}{7} \times 4 \\ &= 30.5 + 0.57 = 31.07 \end{aligned}$$

D_5 is found in the 4^{th} class (22.5-26.5) because this class include the $\left(\frac{5\times 56}{10}\right)^{th} = 28^{th} value$

$$\begin{aligned} D_5 &= L_{d_5} + \frac{(\frac{5\times 56}{10} - F_{d_4})}{f_{d_5}} \times 4 \\ &= 22.5 + \frac{(\frac{2\times 56}{4} - 19)}{10} \times 4 \\ &= 22.5 + 3.6 = 26.1 \end{aligned}$$

D_8 is found in the 6^{th} class (30.5-34.5) because this class include the $\left(\frac{8\times 56}{10}\right)^{th} = 44.8^{th} value$

$$\begin{aligned} D_8 &= L_{d_8} + \frac{(\frac{8\times 56}{10} - F_{d_7})}{f_{d_8}} \times 4 \\ &= 30.5 + \frac{(\frac{8\times 56}{10} - 41)}{7} \times 4 \\ &= 30.5 + 2.17 = 32.67 \end{aligned}$$

P_{30} is found in the 3^{rd} class (18.5-22.5) because this class include the $\left(\frac{30\times 56}{100}\right)^{th} = 16.8^{th} value$

$$\begin{aligned} P_{30} &= L_{p_{30}} + \frac{(\frac{30\times 56}{100} - F_{p_{29}})}{f_{p_{30}}} \times 4 \\ &= 18.5 + \frac{(\frac{30\times 56}{100} - 11)}{19} \times 4 \\ &= 18.5 + 1.22 = 19.72 \end{aligned}$$

P_{90} is found in the 7^{th} class (34.5-38.5) because this class include the $\left(\frac{90\times56}{100}\right)^{th} = 50.4^{th} value$

$$\begin{aligned} P_{90} &= L_{p_{90}} + \frac{(\frac{90\times56}{100} - F_{p_{89}})}{f_{p_{90}}} \times 4 \\ &= 34.5 + \frac{(\frac{90\times56}{100} - 48)}{8} \times 4 \\ &= 34.5 + 1.2 = 35.7 \end{aligned}$$

3.10. Exercises

1. Define and compare the characteristics of the mean, the median and the mode.

2. Your statistics instructor tells you on the first day of class that there will be five tests during the term. From the scores on theses tests for each student he will compute a measures of central tendency that will serve as the student's final course grade. Before taking the first test you must choose whether you want your final grade to be the mean or the median of the five test scores. Which would you choose? Why? Justify your answer.

3. A student's final grades in mathematics, physics, chemistry and sport are, respectively, 82, 86, 90, and 70. If the respective credits received for these courses are 3, 5, 3, and 2, determine an appropriate average grade.

4. A large department store collects data on sales made by each of its sales people. The number of sales made on a given day by each of 20 sales people is shown below.

   9 6 12 10 13 15 16 14 14 16 17 16 24 21 22 18 19 18 20 17

   Then, find Q_3, D_8, P_{80} and P_{90} and interpret all results.

5. In a certain investigation, 460 persons were involved in the study, and based on an enquiry on their age, it was known that 75% of them were 22 or more. The following frequency distribution shows the age composition of the persons under study.

Mid age in years	13	18	23	28	33	38	43	48
Number of persons	24	f_1	90	122	f_2	56	20	33

   (a) Find the median and modal life of condensers and interpret them.

   (b) Find the values of all quartiles.

   (c) Compute the 5^{th} decile, 25^{th} percentile, 50^{th} percentile and the 75^{th} percentile and interpret the results.

6. Given the following frequency distribution,

Mid price of a commodity	15	25	35	45	55
Number of items sold	27	A	28	B	19

(a) If 75% of the items were sold in birr 45 or less and most items were sold in birr 34, find the missing frequencies.

(b) If 25% of the items were sold in greater than or equal to birr 45 and most items were sold in birr 34, find the missing frequencies.

4

Measures of Variation

4.1. Introduction

In the third chapter, we concentrated on a central value (measures of central tendency), which gives an idea of the whole mass that is a complete set of values. However the information so obtained is neither exhaustive nor comprehensive, as the mean does not lead us to know whether the observations are close to each other or far apart. Median is a positional average and has nothing to do with the variability of the observations in a data set. Mode is the largest occurring value independent of the other values in the set. This leads us to conclude that a measure of central tendency is not enough to have a clear idea about the data unless all observations are the same. Moreover two or more data sets may have the same mean and/or median but they may be quite different. So MCT alone do not provide enough information about the nature of the data. The table below displays the price of a certain commodity in four cities. Find the mean and median prices of the four cities and interpret it.

City A	30	30	30
City B	29	30	31
City C	15	30	45
City D	5	30	55

All the four data sets have mean 30 and median is also 30. But by inspection it is apparent that the four data sets differ remarkably from one another. So measures of central tendency alone do not provide enough information about the nature of the data. Thus, to have a clear picture of the data, one needs to have a measure of dispersion or variability among observations in the data set.

Variation or dispersion may be defined as the extent of scatteredness of value around the measures of central tendency. Thus, a measure of dispersion tells us the extent to which the values of a variable vary about the measure of central tendency.

4.2. Objectives of Measures of Variation

1. **To have an idea about the reliability of the measures of central tendency.** If the degree of scatterdness is large, an average is less reliable. If the value of the variation is small, it indicates that a central value is a good representative of all the values in the data set.
2. **To compare two or more sets of data with regard to their variability.** Two or more data sets can be compared by calculating the same measure of variation having the same units of measurement. A set with smaller value posses less variability or is more uniform (or more consistent).
3. **To provide information about the structure of the data.** A value of a measure of variation gives an idea about the spread of the observation. Further, one can summarise about the limits of the expansion of the values in the data set.
4. **To pave way to the use of other statistical measures.** Measures of variation especially variance and standard deviation lead to many statistical techniques like correlation, regression, analysis of variance,…etc.

4.3. Types of Measures of Variation

- **Absolute Measures of Variation**: A measure of variation is said to be an absolute form when it shows the actual amount of variation of an item from a measure of central tendency and are expressed in concrete units in which the data have been expressed.
- **Relative Measures of Variation**: A relative measure of variation is the quotient obtained by dividing the absolute measure by a quantity in respect to which absolute deviation has been computed. It is a pure number and used for making comparisons between different distributions.

Absolute Measures	**Relative Measures**
Range	Coefficient of Range
Quartile Deviation	Coefficient of Quartile Deviation
Mean Deviation	Coefficient of Mean Deviation
Variance	Coefficient of Variation
Standard Deviation	Standard Scores

Before giving the details of these measures of dispersion, it is worthwhile to point out that a measure of dispersion (variation) is to be judged on the basis of all those properties of good measures of central tendency. Hence, their repetition is superfluous.

4.3.1. Range and Relative Range

Range is the simplest and crudest/rough measure of dispersion. It is defined as the difference between the largest and the smallest values in the data.

- For raw data: $R = L - S$
- For grouped data: $R = UCL_{last} - LCL_{first}$

Coefficient of Range:

- For raw data: $CR = \frac{L-S}{L+S}$
- For grouped data: $CR = \frac{UCL_{last}-LCL_{first}}{UCL_{last}+LCL_{first}}$

Range hardly satisfies any property of good measure of dispersion as it is based on two extreme values only ignoring the others. It is not also liable to further algebraic treatment. The main advantage in using range is the simplicity of its computation.

4.3.2. Quartile Deviation and Coefficient of Quartile Deviation

Quartile deviation is sometimes known as Semi-Interquartile Range (SIR). The interquartile Range is $Q_3 - Q_1$. Thus,

$$QD = \frac{Q_3 - Q_1}{2}$$

The corresponding relative measure of variation, coefficient of quartile deviation is:

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

QD involves only the middle 50% of the observations by excluding the observations below the lower quartile and the observations above the upper quartile. Note also that it does not take into account all the individual values occurring between Q_1 and Q_2. It means that, no idea about the variation of even the 50% mid values is available from this measure. Anyhow it provides some idea if the values are uniformly distributed between Q_1 and Q_2.

4.3.3. Mean Deviation and Coefficient of Mean Deviation

The measures of variation discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation is a better measure than range and quartile deviation. Mean deviation is the arithmetic mean of the absolute values of the deviation from some measures of central tendency usually the mean and the median of a distribution. Hence we have mean deviation about the mean $MD(\bar{x})$ and mean deviation about the median $MD(\tilde{x})$.

Methods Obtaining Mean Deviation

- For raw data: $MD(\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n}$ and $MD(\tilde{x}) = \frac{\sum |x_i - \tilde{x}|}{n}$
- For grouped data: $MD(\bar{x}) = \frac{\sum f_i |m_i - \bar{x}|}{\sum f_i}$ and $MD(\tilde{x}) = \frac{\sum f_i |m_i - \tilde{x}|}{\sum f_i}$

MD is not much affected by extreme values. Its main drawback is that the algebraic negative signs of the deviations are ignored. MD is minimum when the deviation is taken from median. The coefficient of mean deviations are:

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}}$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}}$$

Examples

1. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the range, coefficient of range, quartile deviation, coefficient of quartile deviation, mean deviation about mean, mean deviation about median, coefficient of mean deviation about mean and coefficient of mean deviation about median.

Solution:

Data: 15, 20, 25, 25, 27, 28, 30, 34

$$R = max - min = 34 - 15 = 19, CR = \frac{max - min}{max + min} = \frac{34 - 15}{34 + 15} = 0.388$$

To find QD and CQD, we have to calculate Q_1 and Q_3 first.

$$\begin{aligned} Q_1 &= \left(\frac{1 \times (8+1)}{4}\right)^{th} value = 2.25^{th} value \\ &= 2^{nd} value + 0.25(3^{rd} value - 2^{nd} value) \\ &= 20 + 0.25 \times (25 - 20) \\ &= 21.25 \end{aligned}$$

$$\begin{aligned} Q_3 &= \left(\frac{3 \times (8+1)}{4}\right)^{th} value = 6.75^{th} value \\ &= 6^{th} value + 0.75(7^{th} value - 6^{th} value) \\ &= 28 + 0.75 \times (30 - 28) \\ &= 29.5 \end{aligned}$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{29.5 - 21.25}{2} = 4.125$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{29.5 - 21.25}{29.5 + 21.25} = \frac{8.25}{50.75} = 0.163$$

Beside to this to compute $MD(\bar{x}), MD(\tilde{x}), CMD(\bar{x})$ and $CMD(\tilde{x})$ we should obtain $\bar{x}$ and $\tilde{x}$.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{204}{8} = 25.5;\ \tilde{x} = 26$$

$$\begin{aligned} MD(\bar{x}) &= \frac{\sum |x_i - \bar{x}|}{n} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_8 - \bar{x}|}{8} \\ &= \frac{|15 - 25.5| + |20 - 25.5| + ... + |34 - 25.5|}{8} \\ &= \frac{34}{8} = 4.25 \end{aligned}$$

$$\begin{aligned} MD(\tilde{x}) &= \frac{\sum |x_i - \tilde{x}|}{n} = \frac{|x_1 - \tilde{x}| + |x_2 - \tilde{x}| + ... + |x_8 - \tilde{x}|}{8} \\ &= \frac{|15 - 26| + |20 - 26| + ... + |34 - 26|}{8} \\ &= \frac{32}{8} = 4 \end{aligned}$$

Thus,

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} = \frac{4.25}{25.5} = 0.1667$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}} = \frac{4}{26} = 0.154$$

2. Calculate the R, QD and CQD for the following frequency distribution.

Class limits	10-14	15-19	20-24	25-29	30-34	35-38
Frequency	8	10	22	35	15	10

Solution:

Previously, we have obtained the following quantities for the students score data:

$$\bar{x} = 25.64, \tilde{x} = 26.1, Q_1 = 20, Q_3 = 31.07$$

Class	m_i	f_i	$\|m_i - \bar{x}\|$	$f_i\|m_i - \bar{x}\|$	$\|m_i - \tilde{x}\|$	$f_i\|m_i - \tilde{x}\|$
10.5-14.5	12.5	4	13.14	52.56	13.6	54.4
14.5-18.5	16.5	7	9.14	63.98	9.6	67.2
18.5-22.5	20.5	8	5.14	41.12	5.6	44.8
22.5-26.5	24.5	10	1.14	11.40	1.6	16.0
26.5-30.5	28.5	12	2.86	34.32	2.4	28.8
30.5-34.5	32.5	7	6.86	48.02	6.4	44.8
34.5-38.5	36.5	8	10.86	86.88	10.4	83.2
Total		56		338.28		339.2

$$R = UCL_{last} - LCL_{first} = 38 - 11 = 27$$

$$CR = \frac{UCL_{last} - LCL_{first}}{UCL_{last} + LCL_{first}} = \frac{38 - 11}{38 + 11} = \frac{27}{49} = 0.551$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{31.07 - 20}{2} = \frac{11.07}{2} = 5.54$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{31.07 - 20}{31.07 + 20} = \frac{11.07}{51.07} = 0.22$$

$$MD(\bar{x}) = \frac{\sum f_i|m_i - \bar{x}|}{\sum f_i} = \frac{338.28}{56} = 6.04$$

$$MD(\tilde{x}) = \frac{\sum f_i|m_i - \tilde{x}|}{\sum f_i} = \frac{339.2}{56} = 6.06$$

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} = \frac{6.04}{25.64} = 0.24$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}} = \frac{6.06}{26.1} = 0.23$$

4.3.4. Variance and Standard Deviation

Variance and standard deviation are the most superior and widely used measures of dispersions and both measure the average dispersion of the observations around the mean. The variance of a data set is the sum of the squares of the deviation of each observation taken from the mean divided by total number of observations in the data set. The positive square root of variance is called standard deviation.

For a *population* containing N elements, the population standard deviation is denoted by the Greek letter σ (sigma) and hence the population variance is denoted by σ^2.

- For raw data: $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ and $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$
- For grouped data: $\sigma^2 = \frac{\sum f_i(m_i-\mu)^2}{N}$ and $\sigma = \sqrt{\frac{\sum f_i(m_i-\mu)^2}{N}}$

For a *sample* of n elements, the sample variance and standard deviation denoted by s^2 and s, respectively, are calculated as using the formulae:

- For raw data: $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ and $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
- For grouped data: $s^2 = \frac{\sum f_i(m_i-\bar{x})^2}{\sum f_i-1}$ and $s = \sqrt{\frac{\sum f_i(m_i-\bar{x})^2}{\sum f_i-1}}$

Examples

1. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.

   Solution:

   We are expected to compute the sample mean $\bar{x}$ first since the sample variance is a function the sample mean.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{10+20+12+17+16}{5} = \frac{75}{5} = 15$$

$$\begin{aligned} S^2 &= \frac{\sum(x_i-\bar{x})^2}{n-1} \\ &= \frac{(10-15)^2+(20-15)^2+(12-15)^2+(17-15)^2+(16-15)^2}{5-1} \\ &= \frac{64}{4} = 16 \end{aligned}$$

   Hence, $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{16} = 4.$

2. Calculate the variance and standard deviation for the following frequency distribution.

Class limits	10-14	15-19	20-24	25-29	30-34	35-38
Frequency	8	10	22	35	15	10

Solution:

The necessary calculation for calculating variance are as follows.

Class	m_i	f_i	$(m_i - \bar{x})$	$(m_i - \bar{x})^2$	$f_i(m_i - \tilde{x})^2$
10.5-14.5	12.5	4	-13.14	172.6596	690.6384
14.5-18.5	16.5	7	-9.14	83.5396	584.7772
18.5-22.5	20.5	8	-5.14	26.4196	211.3568
22.5-26.5	24.5	10	-1.14	1.2996	12.9960
26.5-30.5	28.5	12	2.86	8.1796	98.1552
30.5-34.5	32.5	7	6.86	47.0596	329.4172
34.5-38.5	36.5	8	10.86	117.9396	943.5168
Total		56		338.28	2870.8576

$$s^2 = \frac{\sum f_i(m_i - \bar{x})^2}{\sum f_i - 1} = \frac{2870.8576}{55} = 52.19$$

Therefore $s = \sqrt{52.19} = 7.22$.

- The main objection of mean deviation, removal of the negative signs, is removed by taking the square of the deviations from the mean. The first main demerit of variance is that its unit is the square of the unit of measurement of the variable values. For example, the sample variance of $2m$, $6m$ and $4m$ is $4m^2$. The interpretation is, on average each value differs from the mean by $4m^2$, which is completely wrong because one thing the unit of measurement of variance is not the same as that of the data set. The other disadvantage of variance is, the variation of the data is exaggerated because the deviation of the each value from the mean is squared. For the given example, the variation of the data is exaggerated from two to four since it is taking the square of the deviations. Variance also gives more weight the extreme values as compared to those which are near to the mean value.

- Standard deviation is considered to be the best measure of dispersion because the unit of measurement is the same as the data set and the exaggeration made by variance will be eliminated by taking the square root of it. In simple words, it explains the average amount of variation on either sides of the mean. If the standard deviation of the data is small the values are concentrated near the mean and if it large the values are scattered away from the mean.

Properties of Variance and Standard Deviation

1. If a constant is added (subtracted) to (from) each and every observation, the standard deviation as well as the variance remains the same.

2. If each and every value is multiplied by a nonzero constant k, the standard deviation is multiplied by k and the variance is multiplied by k^2.

3. If there are k different groups having the same units of measurement with sample means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$, number of sample observations $n_1, n_2, ..., n_k$ and sample variances $s_1^2, s_2^2, ..., s_k^2$ respectively, then the variance of all the groups called the *pooled variance* denoted by s_p^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)[s_1^2 + (\bar{x}_1 - \bar{x}_c)^2] + ... + (n_k - 1)[s_k^2 + (\bar{x}_k - \bar{x}_c)^2]}{n_1 + n_2 + ... + n_k - k}$$

$$s_p^2 = \frac{\sum (n_i - 1)[s_i^2 + (\bar{x}_i - \bar{x}_c)^2]}{\sum n_i - k}$$

If $\bar{x}_1 = \bar{x}_2 = ... = \bar{x}_k$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2}{n_1 + n_2 + ... + n_k - k} = \frac{\sum (n_i - 1)s_i^2}{\sum n_i - k}$$

Examples

1. The mean weight of 150 students is 60 kilograms. The mean weight of boys is 70 kg with a standard deviation of 10 kg. For the girls, the mean weight is 55 kg and the standard deviation 15 kg. Then,

   (a) Find the number of boys and girls.

   (b) Find the combined standard deviation.

2. A distribution consists of four parts characterized as follows. Find the mean and standard deviation of the distribution.

Part	No of items	Mean	S.D.
1	50	61	8
2	100	70	9
3	120	50	10
4	30	83	11

3. The arithmetic mean and standard deviation of a series of 20 items were computed as 20 and 5 respectively. While calculating these, an item 13 was misread as 30. Find the correct mean and standard deviation.

4. The following data are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	Girls
Number	100	50
Mean	60	45
Variance	9	4

   a) Find the mean and variance of the combined series.

   b) If one of the values is misread as 60 instead of 40 what is the correct standard deviation.

4.3.5. Coefficient of Variation

All absolute measures of dispersion have units. If two or more distributions differ in their units of measurement, their variability cannot be compared by any of the absolute measure of variation. Also, the size of the absolute measures of dispersion depends upon the size of the values. That is if the size of the values is larger, the value of the absolute measures will also be larger. Generally absolute measures of variation fail to be appropriate for comparing two or more groups if:

- The groups have different units of measurement.
- The size of the data between the groups is not the same.

Coefficient of variation is a relative measure of standard deviation. It is the ratio of the standard deviation to the mean and expressed as percent. Hence, it is a unitless measure of variation and also takes into account the size of the means of the distributions.

- ▶ For population: $cv = \frac{\sigma}{\mu} \times 100\%$
- ▶ For sample: $cv = \frac{s}{\bar{x}} \times 100\%$
- ■ The distribution having less cv is said to be less variable or more consistent or more uniform. For field experiments, cv , is generally reported. If it is small, it indicates more reliability of experimental findings.

Examples

1. Compare the variability of the following two sample data sets using standard deviation and coefficient of variation.

   A : 2 Meters, 4 Meters, 6 Meters

   B : 1000 Liters, 800 Liters, 900 Liters

2. The average IQ of statistics students is 110 with standard deviation 5 and the average IQ of mathematics students is 106 with standard deviation 4. Which class is less variable in terms of IQ?

4.4. Exercises

1. Find the range, quartile deviation, mean deviation about the mean, mean deviation about the median, mean deviation about the mode, variance, standard deviation and coefficient of variation for the following distribution.

Class	2-4	4-6	6-8	8-10
Frequency	2	5	4	7

2. Explain the rationale for using $n-1$ to compute the sample variance.

3. What is the purpose of coefficient variation?

4. Two persons participated in five shooting competition and were able to hit the target correctly out of fifteen shots as given below.

Competitor A	6	12	12	10	7
Competitor B	12	15	7	7	4

Which competitor is more uniform in shooting performance?

5

Elementary Probability

5.1. What is Probability?

Probability is a numerical description of chance of occurrence of a given phenomena under certain condition.

Probability theory plays a central role in statistics. After all, statistical analysis is applied to a collection of data in order to discover something about the underlying events. These events may be connected to one another. However, the individual choices involved are assumed to be *random*. Alternatively, we may sample a population at random and make inferences about the population as a whole from the sample by using statistical analysis. Therefore, a solid understanding of probability theory - *the study of random events* - is necessary to understand how the statistical analysis works and also to correctly interpret the results.

5.2. Concept of Set

In order to discuss the theory of probability, it is essential to be familiar with some ideas and concepts of mathematical theory of set. A *set* is a collection of well-defined objects which is denoted by capital letters like A, B, C, etc.

In describing which objects are contained in set A, two common methods are available. These methods are:

1. Listing all objects of A. For example, $A = \{1, 2, 3, 4\}$ describes the set consisting of the positive integers 1, 2, 3 and 4.

2. Describing a set in words, for example, set A consists of all real numbers between 0 and 1, inclusive. It can be written as $A = \{x : 0 \leq x \leq 1\}$, that is, A is the set of all $x's$

where x is a real number between 0 and 1, inclusive.

If $A = \{a_1, a_2, ..., a_n\}$, then each object $a_i; i = 1, 2, ..., n$ belonging to set A is called a member or an element of set A, i.e., $a_i \in A$. A set consisting all possible elements under consideration is called a *universal* set (denoted by $\cup$). On the other hand, a set containing no element is called an empty set (denoted by $\emptyset$ or $\{\}$).

If every element of set A is also an element of set B, A is said to be a subset of B and write as $A \subset B$. Every set is a subset of itself, i.e., $A \subset A$. Empty set is a subset of every set. If $A \subset B$ and $B \subset C$, then $A \subset C$. If $A \subset B$ and $B \subset A$, then A and B are said to be equal.

5.2.1. Set Operation

1. **Union (Or)**: A set consisting all elements in A or B or both is called the union set of$A$ and B, and write as $A \cup B$. That is, $A \cup B = \{x : x \in A, x \in B \text{ or } x \in \mathit{both}\}$. The set$A \cup B$ is also called the sum of A and$B$.

2. **Intersection (And)**: A set consisting all elements in both A and B is called an intersection set of A and B, and write as $A \cap B$. This is, $A \cap B = \{x : x \in A$ and $x \in B\}$. The intersection set of A and B is also called the the product of A and B.

3. **Complement (Not)**: The complement of a set A, denoted by A^c, is a set consisting all elements of $\cup$ that are not in A; i.e., $A^c = \{x : x \notin A\}$.

4. **Disjoint Set**: Sets A and B are disjoint set if $A \cap B = \emptyset$.

5. **Relative Complement**: The relative complement of B in A, denoted by $A \backslash B$ is a set of all elements of A which are not in B. It is written as $A \backslash B = \{x : x \in A$ and $x \notin B\} = A \cap B^c$.

Important Laws

- Commutative laws:
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
- Associative laws:
  - $A \cup (B \cup C) = (A \cup B) \cup C$

  - $A \cap (B \cap C) = (A \cap B) \cap C$

- Distributive laws:
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

- Identity laws:
  - $A \cup A = A, A \cap A = A$
  - $A \cup U = U, A \cap U = A$
  - $A \cup \emptyset = A, A \cap \emptyset = \emptyset$

5.3. Definition and Some Basic Concepts

1. ***Experiment*** (ξ): is any statistical process that can be repeated several times and in any trial of which the outcome is unpredictable.

   ▷ Tossing a coin only once, $S = \{$Head (H), Tail (T)$\}$

   ▷ Tossing a coin two times, $S = \{HH, HT, TH, TT\}$

   ▷ Rolling a die, $S = \{1, 2, 3, 4, 5, 6\}$

   ▷ Selecting an item from a production lot, $S = \{$Defective, Non-defective$\}$

   ▷ Introducing a new product, $S = \{$Success, Failure$\}$

2. ***Sample Space (S)***: is a set consisting all possible outcomes of a given experiment, ξ.

3. ***Event***: is an outcome or a set of outcomes (having some common characteristics) of an experiment.

   Simple Event (Elementary Event): is an event consisting a single outcome. The elementary events are the building blocks (or atoms) of a probability model. They are the events that cannot be decomposed further into smaller sets of events.

   Compound Event: is an event consisting two or more outcomes.

4. ***Independent Event***: two or more events are independent if the occurrence of one event has no effect on the probability of occurrence of the other.

5. ***Mutually Exclusive Events***: two or more events are mutually exclusive, if they have no outcome in common. They cannot occur together simultaneously.
6. ***Complementary Event***: Two mutually exclusive events are complementary if there are no common elements between themselves and both of them contain all possible outcomes. To be complementary, first they should be mutually exclusive events.

5.4. Counting Rules

Counting techniques are mathematical models which are used to determine the number of possible ways of arranging or ordering objects. They are used to find a solution to fix the size of the sample space that is extremely large. To count possible outcomes of a sample space or/and an event we use the following counting techniques.

Addition Rule: states that if a task can be done (accomplished) by any of the k procedures, where i^{th} procedures has n_i alternatives, then the total number of ways of doing the task is

$$n_1 + n_2 + ... + n_k = \sum_{i=1}^{k} n_i$$

Example: Suppose a lady wants to make journey from Harar to Dire Dawa. If she can use either plane, bus, cycle, horse, and there are 3 flights, 4 buses, 2 cycles and 3 horses available. In how many different ways can she make her journey?

Solution:

From the given problem $n_f = 3, n_b = 4, n_c = 2$ and $n_h = 3$. So she has

$$n_f + n_b + n_c + n_h = 3 + 4 + 2 + 3 = 12$$

different ways to make her trip from Harar to Dire Dawa.

Multiplication Rule: states that if a choice consists k steps where the first step can be done in n_1 ways, for each of which second can be done in n_2 ways, ..., for each of those k^{th} steps can be done in n_k ways. Then, the total number of distinct ways to accomplish the task/choice is equal to

$$n_1 \times n_2 \times ... \times n_k = \prod_{i=1}^{k} n_i$$

Example 1: Suppose a cafeteria provides 5 kinds of cake which it serves with tea, coffee, milk and coca cola. Then, in how many different ways can you order your breakfast of cake with a drink?

Solution:

The work has two steps. First, we order a type of cake $n_1 = 5$ and then we order kind of drink through $n_2 = 4$. Thus,one can have

$$n_1 \times n_2 = 5 \times 4 = 20$$

different ways to order his/her breakfast.

Example 2: There are 2 bus routes from city X to city Y and 3 train routes from city Y to city Z. In how many ways can a person go from city X to city Z?

Solution:

$$n_1 \times n_2 = 2 \times 3 = 6$$

So the person can go from city X to city Z in 6 ways.

Permutation: is arrangement of objects with attention to order of appearance.

Rule 1: The number of permutations of n distinct objects taking all together is

$$n! = n \times (n-1) \times (n-2) \times ... \times (1)$$

By definition $1! = 0! = 1$.

Example 1: In how many different ways can 3 persons sleep in a bed?

Solution:

$$n! = 3! = 3 \times 2 \times 1 = 6\, ways.$$

Example 2: Suppose a photographer must arrange 4 persons in a row for a photograph. In how many different ways can the arrangement be done?

Solution:

$$n! = 4! = 4 \times 3 \times 2 \times 1 = 24\, ways.$$

Rule 2: Given n distinct objects, the number of permutations of r objects taken from n objects is denoted by nPr and given by

$$nPr = \frac{n!}{(n-r)!}; \quad r \leq n$$

Example 1: In how many ways can 10 people be seated on a bench if only 4 seats are available?

Solution:

$$nPr = 10P4 = \frac{10!}{(10-4)!} = \frac{10 \times 9 \times 8 \times 7 \times 6!}{6!} = 5040\, ways.$$

Example 2: How many 5 letter permutations can be formed from the letters in the word DISCOVER?

Solution:

$$nPr = 8P5 = \frac{8!}{(8-5)!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3!}{3!} = 6270$$

Rule 3: Given n objects in which n_1 are alike, n_2 are alike, ..., n_r are alike is given by

$$\frac{n!}{n_1! \times n_2! \times ... \times n_r!}$$

Example: How many different permutations can be made from the letters in the word:

- STATISTICS

  Solution:

  $n_1 = n(s) = 3, n_2 = n(t) = 3, n_3 = n(a) = 1, n_4 = n(i) = 2$ and $n_5 = n(c) = 1$. Thus,

  $$\frac{n!}{n_1! \times n_2! \times n_3! \times n_4! \times n_5!} = \frac{10!}{3! \times 3! \times 1! \times 2! \times 1!} = 50400$$

- MISSISSIPPI

  Solution:

  $n_1 = n(m) = 1, n_2 = n(i) = 4, n_3 = n(s) = 4$ and $n_4 = n(p) = 2$. Thus,

  $$\frac{n!}{n_1! \times n_2! \times n_3! \times n_4! \times n_5!} = \frac{11!}{1! \times 4! \times 4! \times 2!} = 34650$$

Combination: A set of n distinct objects considered without regard to the orders of appearance is called combination. For example, *abc, bac, acb, cab, cba* are six different permutations but they are the same combination.

Rule 1: The number of ways of selecting r objects from n distinct objects is called combination of r objects from n objects denoted by nCr or $\binom{n}{r}$ and given by

$$nCr = \binom{n}{r} = \frac{n!}{(n-r)! \times r!}; \quad r \leq n$$

Example: In how many ways can student choose 3 books from a list of 12 different books?

Solution:

$$\begin{aligned}\binom{n}{r} = \binom{12}{3} &= \frac{n!}{(n-r)! \times r!} \\ &= \frac{12!}{(12-3)! \times 3!} \\ &= \frac{12!}{9! \times 3!} = \frac{12 \times 11 \times 10 \times 9!}{9! \times 3!} \\ &= 220\end{aligned}$$

Rule 2: If the selection has k steps, by selecting r_1 of n_1 objects, r_2 of n_2, ..., r_k of n_k objects, then the total number of ways of doing this selection is equal to

$$\binom{n_1}{r_1} \times \binom{n_2}{r_2} \times ... \times \binom{n_k}{r_k}$$

Example: Out of 5 male workers and 7 female workers of some factory a committee consisting 2 male and 3 female workers to be formed. In how many ways can this done if

(a) all workers are eligible.

$$\binom{n_1}{r_1} \times \binom{n_2}{r_2} = \binom{5}{2} \times \binom{7}{3} = 10 \times 35 = 350$$

(b) one particular female must be a member.

$$\binom{n_1}{r_1} \times \binom{n_2}{r_2} = \binom{5}{2} \times \binom{6}{2} = 10 \times 15 = 150$$

(c) two particular male workers cannot be members for some reason.

$$\binom{n_1}{r_1} \times \binom{n_2}{r_2} = \binom{3}{2} \times \binom{7}{3} = 3 \times 35 = 105$$

■ The difference between permutation and combination is that in combination the order of objects being selected (arranged) is not important, but order matters in permutation.

5.5. Approaches in Probability Definition

1. **The Classical Approach** (also called **Mathematical Approach**): Suppose there are N possible outcomes in the sample space S of an experiment. Out of these N outcomes,

only n are favorable to the event E, then the probability that the event E will occur is:

$$P(E) = \frac{No\,of\,favourable\,outcomes\,to\,E}{total\,no\,of\,outcomes} = \frac{n(E)}{n(S)} = \frac{n}{N}$$

Example 1: Consider an experiment of tossing a die. Then, what is the probability that

(a) odd numbers occur.

Solution:

The sample space of the given experiment is $S = \{1, 2, 3, 4, 5, 6\}$. Further let A be an event of getting odd numbers in rolling a die only once.

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = 0.5$$

b) number 4 occurs.

Solution:

Let B be an event of getting number 4 in rolling a die only once.

$$P(B) = \frac{n(B)}{n(S)} = \frac{1}{6} = 0.167$$

(c) number 8 occurs.

Solution:

Let C be an event of getting number 8 in rolling a die only once.

$$P(C) = \frac{n(C)}{n(S)} = \frac{0}{6} = 0$$

(d) numbers between 1 and 6 inclusive occur.

Solution:

Let D be an event of getting numbers between 1 and 6 inclusive occur.

$$P(D) = \frac{n(D)}{n(S)} = \frac{6}{6} = 1$$

- ▷ Events with zero probability of occurrence are known as null or impossible events.
- ▷ Events with probability equal to unity are known as sure events.

Example 2: What is the probability of getting one head in tossing two coins?

Solution:

$S = \{HH, HT, TH, TT\}$ and suppose E be the event getting one head in an experiment of tossing two coins.

$$P(E) = \frac{n(E)}{n(S)} = \frac{2}{4} = 0.5$$

2. **The Empirical Approach** (also called **Frequentist Approach**): It is based on a relative frequency. Given a frequency distribution, the probability of an event being in a given class is

$$P(E) = \frac{f_i}{\sum f_i}$$

The difference between classical and empirical probability is that the former uses sample space to determine the numerical probability while the latter is based on frequency distribution.

3. **Subjective Approach**: calculates probability based on an educated guess or experience or evaluation of a problem. For example a physician might say that on the basis of his/her diagnosis, there is a 30% chance the patient will need an operation.

5.6. Some Probability Rules/Axioms

Let S be a sample space associated with a random experiment. Then with any event E, in this sample space, we associate a real number called probability of E satisfying the following properties (axioms).

- $0 \leq P(E) \leq 1$
- $P(S) = 1$
- If A and B are mutually exclusive events, then

$$P(A\, or\, B) = P(A \cup B) = P(A) + P(B)$$

- If $A_1, A_2, ..., A_n$ are pairwise mutually exclusive events, then

$$P\left(\bigcup_{i=n}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

- $P(A \cup A^c) = P(A) + P(A^c)$

- ► $P(\phi) = 0$

Using the above axioms, it can be shown that for any two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example 1: A box of 20 candles consists of 5 defective and 15 non-defective candles. If 4 of these candles are selected at random, what is the probability that

(a) all will be defective.

Solution:

Let A be an event of all candles are defective.

$$P(A) = \frac{n(A)}{n(S)} = \frac{\binom{5}{4} \times \binom{15}{0}}{\binom{20}{4}} = 0.001032$$

(b) 3 will be non-defective.

Let B be an event of 3 candles are non-defective.

$$P(B) = \frac{n(B)}{n(S)} = \frac{\binom{5}{1} \times \binom{15}{3}}{\binom{20}{4}} = 0.4696$$

(c) all will be non-defective.

Let C be an event of all candles are non-defective.

$$P(C) = \frac{n(C)}{n(S)} = \frac{\binom{5}{0} \times \binom{15}{4}}{\binom{20}{4}} = 0.2817$$

Example 2: An urn contains 6 white, 4 red and 9 black balls. If 3 balls are drawn at random, find the probability that

(a) two of the balls drawn are whites.

Let E_1 be an event two of the balls drawn are whites.

$$P(E_1) = \frac{n(E_1)}{n(S)} = \frac{\binom{6}{2} \times \binom{13}{1}}{\binom{19}{3}} = 0.2012$$

(b) one is from each colour.

Let E_2 be an event of one from each colour.

$$P(E_2) = \frac{n(E_2)}{n(S)} = \frac{\binom{6}{1} \times \binom{4}{1} \times \binom{9}{1}}{\binom{19}{3}} = 0.2229$$

(c) none is red.

Let E_3 be an event of none is red.

$$P(E_3) = \frac{n(E_3)}{n(S)} = \frac{\binom{15}{3} \times \binom{4}{0}}{\binom{19}{3}} = 0.4695$$

(d) at least one is white.

Let E_4 be an event of at least one is white.

$$P(E_4) = \frac{n(E_4)}{n(S)} = \frac{\binom{6}{1} \times \binom{13}{2}}{\binom{19}{3}} + \frac{\binom{6}{2} \times \binom{13}{1}}{\binom{19}{3}} + \frac{\binom{6}{3} \times \binom{13}{0}}{\binom{19}{3}} = 0.7048$$

5.7. Conditional Probability

When the outcome or occurrence of an event affects the outcome or occurrence of another event, the two events are said to be dependent (conditional). If two events, A and B, are dependent to each other, the probability of event A occurring knowing that event B has already occurred is said to be the conditional probability of A given that event B has already occurred,

$$P(A/B) = \frac{P(A \cap B)}{P(B)};\ P(B) \neq 0$$

The probability of event B occurring knowing that event A has already occurred is said to be the conditional probability of B given that event A has already occurred,

$$P(B/A) = \frac{P(A \cap B)}{P(A)};\ P(A) \neq 0$$

Remarks

(i) $0 \leq P(A/B) \leq 1$

(ii) $P(S/B) = 1$

(iii) For mutually exclusive events A_1 and A_2,

$$P(A_1 \cup A_2/B) = P(A_1/B) + P(A_2/B)$$

(iv) For pairwise mutually exclusive events $A_1, A_2, ..., A_n$

$$P\left(\bigcup_{i=n}^{n} A_i/B\right) = \sum_{i=1}^{n} P(A_i/B)$$

Example: If the probability that a research project will be well planned is 0.6, and the probability that it will be well planned and well executed is 0.54. Then, what is the probability that it will be

(a) well executed given that it is well planned.

Solution:

Let D and E be an events of the research project is well planned and well executed respectively. Then $P(D) = 0.6$ and $P(D \cap E) = 0.54$.

$$P(E/D) = \frac{P(D \cap E)}{P(D)} = \frac{0.54}{0.6} = 0.9$$

(b) will not be well executed given that it is well planned.

Solution:

$$P(E^c/D) = \frac{P(D \cap E^c)}{P(D)} = \frac{P(D) - P(D \cap E)}{P(D)} = 1 - \frac{P(D \cap E)}{P(D)}$$
$$P(E^c/D) = \frac{P(D \cap E^c)}{P(D)} = 1 - P(E/D) = 1 - 0.9 = 0.1$$

5.8. Independence

Recall mutually exclusive events A and B, $A \cap B = \phi$, which implies that $P(A \cap B) = 0$.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = 0$$

If B occurs A will never occur at the same time. That means, they are dependent. Again recall that if $A \subset B$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)}{P(A)} \leq 1$$

Definition: Two events, A and B are said to be statistically independent if

$$P(A \cap B) = P(A) \times P(B)$$

Example: Consider an experiment of tossing two dice. Then, let

A - the first die show an even number.

B - the second die show an odd number.

C - both dice show even number.

Thus check whether A and B, A and C, B and C are independent events.

Solution:

Use the following sample space, S.

→	**1**	**2**	**3**	**4**	**5**	**6**
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

$$P(A) = \frac{n(A)}{S} = \frac{18}{36}, \; P(A \cap B) = \frac{n(A \cap B)}{S} = \frac{9}{36}$$

$$P(B) = \frac{n(B)}{S} = \frac{18}{36}, \; P(A \cap C) = \frac{n(A \cap C)}{S} = \frac{9}{36}$$

$$P(C) = \frac{n(C)}{S} = \frac{9}{36}, \; P(B \cap C) = \frac{n(B \cap C)}{S} = \frac{0}{36}$$

$$P(A \cap B) = P(A) \times P(B)$$

$$\frac{9}{36} = \frac{18}{36} \times \frac{18}{36}$$

$$P(A \cap C) \neq P(A) \times P(C)$$

$$\frac{9}{36} \neq \frac{18}{36} \times \frac{9}{36}$$

$$P(B \cap C) \neq P(B) \times P(C)$$

$$\frac{0}{36} \neq \frac{18}{36} \times \frac{9}{36}$$

Therefore, based on the above results A and B are statistically independent events. However, events A and C and B and C are not statistically independent.

If A and B are independent, then the following holds true.

(i) $P(A/B) = P(A)$

(ii) $P(B/A) = P(B)$

(iii) A^c and B^c are independent.

(iv) A^c and B, B^c and A are independent.

5.9. Exercises

1. Provide definitions for each of these terms:

   (a) Random experiment

   (b) Events

   (c) Mutually exclusive events

   (d) Equally likely events

   (e) Independent events

2. A package contains 12 resistors, 3 of which are defective. If 3 are selected, find the probability of getting

   (a) no defective resistor.

   (b) one defective resistor.

   (c) all defective resistors.

3. Let A and B be two events associated with an experiment and suppose that $P(A) = 0.4$ while $P(A \cup B) = 0.7$. Let $P(B) = p$. For what choice of p

   (a) A and B are mutually exclusive?

   (b) A and B are independent?

4. The personnel department of a company has records which show the following analysis of its 200 accountants.

Age	Bachelor's	Master's
Under 30	90	10
30 to 40	20	30
Over 40	40	10

   If one accountant is selected at random from the company, find the probability that

   (a) he has only a bachelor's degree.

   (b) he has mater's degree given that he is over 40.

   (c) he is under 30 given that he has a bachelor's degree.

6

Probability Distributions

Consider the following illustrations.

Example: Consider the experiment, ξ, of tossing a coin twice.

$$S = \{HH, HT, TH, TT\}$$

Let X be number of heads. Thus, another sample space with respect to X (also called the range space of X) is

$$R_x = \{0, 1, 2\}$$

Definition: A function X which assigns a real numbers to all possible values of a sample space is called a random variable. A random variable is a variable that has a single numerical value (determined by chance) for each outcome of a procedure.

6.1. Type of Random Variables

A random variable can be classified as being either *discrete* or *continuous* depending on the numerical values it assumes.

A ***discrete random variable*** has either a finite number of values or a countable number of values; that is, they result from counting process. The possible value of X may be $x_1, x_2, ..., x_n$. For any discrete random variable X the following will be true.

(i) $0 \leq p(x_i) \leq 1$

(ii) $\sum_{i=1}^{n} p(x_i) = 1$ for finite and $\sum_{i=1}^{\infty} p(x_i) = 1$ for countably infinite.

$p(x_i)$ is called probability function or point probability function or mass function. The collection of pairs $(x_i, p(x_i))$ is called probability distribution. It gives the probability for each value or range of values of the random variable.

Example 1: Construct a probability distribution for getting heads in an experiment of tossing a coin two times.

Solution:

$S = \{HH, HT, TH, TT\}$. Let X be a random variable of getting head in tossing a coin two times. Then $R_x = \{0, 1, 2\}$.

$$P(X = 0) = P(TT) = \frac{1}{4},\ P(X = 1) = P(HT, TH) = \frac{2}{4},\ P(X = 2) = P(HH) = \frac{1}{4}$$

Hence the probability distribution for X is given by:

X	0	1	2
$P(X = x_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

Example 2: The probability distribution of a discrete random variable Y is given by

$$P(Y = y) = cy^2,\ y = 0, 1, 2, 3, 4$$

Then find the value of c.

Solution:

First we should have to compute the point probabilities find the value of c.

$$P(Y = 0) = 0,\ P(Y = 1) = c,\ P(Y = 2) = 4c,\ P(Y = 3) = 9c,\ P(Y = 4) = 16c$$

$$\sum_{i=0}^{4} P(Y = y_i) = 1$$

$$0 + c + 4c + 9c + 16c = 1$$

$$c = \frac{1}{30}$$

A ***continuous random variable*** has infinitely many values, and those values can be associated with measurements on a continuous scale in such a way that there are no gaps or interruptions. That means, if it assumes all possible values in the interval (a, b), where $a, b \in \Re$ and there exist a function called probability density function (pdf) satisfying the following conditions.

(i) $f(x) \geq 0,\ \forall x$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

For any two real numbers a and b such that $-\infty < a < b < \infty$ then

$$P(a < X < b) = \int_a^b f(x)dx$$

If X is a continuous random variable, then:

- $P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0$
- $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x)dx$

Example: Let X be a continuous random variable and its pdf is given by:

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

(a) Verify whether *f(x)* is a pdf or not.

- $f(x) = 2x \geq 0 \, \forall x$
- $\int_0^1 f(x)dx = \int_0^1 2x = 1$

(b) Find $P(0.5 < X < 0.75)$.

$$P(0.5 < X < 0.75) = \int_{0.5}^{0.75} 2xdx = 2\int_{0.5}^{0.75} xdx = 2\left[\frac{x^2}{2}\right]_{0.5}^{0.75} = 0.315$$

6.2. Introduction to Expectation of Random Variables

6.2.1. Expectation of Random Variables and Its Properties

Definition: If X is discrete random variable with possible values of $x_1, x_2, ..., x_n$ having the probabilities of $p(x_1), p(x_2), ..., p(x_n)$, then the mean value of X denoted by $E(X)$ or μ is defined as:

$$E(X) = \mu = \sum_{i=1}^{\infty} x_i p(x_i)$$

if the series converges.

Definition: If X is continuous random variable with pdf of *f(x)*, its mean is given by

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

Properties of Expectation

Assume that the expected value of a random variable exists.

1. For any constant *"a"* we have

$$E(aX) = aE(X) = a\mu$$

2. If $X = a$, then $E(X) = a$.

6.2.2. Variance of Random Variables and Its Properties

Definition: Let X be a random variable. Then variance of X denoted by $var(X)$ or σ_x^2 is defined as

$$var(X) = \sigma_x^2 = E[X - E(X)]^2 = E[X - \mu]^2 = E(X^2) - \mu^2$$

Thus, the standard deviation of X is given by $\sigma_x = \sqrt{\sigma_x^2}$.

Properties of Variance

1. If *"a"* is constant, then

   ▷ $var(X + a) = var(X)$

   ▷ $var(aX) = a^2 var(X)$

Example 1: A coin is tossed two times. Let X be the number of heads. Find the mean value and the standard deviation of X.

Solution:

We already constructed a probability distribution for number of heads in previous example.

X	0	1	2
$P(X = x_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

$$E(X) = \mu = \sum_{i=0}^{2} x_i p(x_i) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

To get the standard deviation of X, first we compute σ_x^2.

$$E(X^2) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^{\times} \frac{1}{4} = \frac{6}{4} = 1.5$$

$$\sigma_x^2 = E[X - E(X)]^2 = E[X - \mu]^2 = E(X^2) - \mu^2$$

$$\sigma_x^2 = E(X^2) - \mu^2 = 1.5 - 1^2 = 0.5$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{0.5} = 0.707$$

Example 2: Suppose that X is a continuous random variable with pdf of

$$f(x) = \begin{cases} 1 + x, & -1 \leq x \ < 0 \\ 1 - x, & 0 \leq x \leq 1 \end{cases}$$

then find the mean value and variance of X.

Solution:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-1}^{0} x(1 + x)dx + \int_{0}^{1} x(1 - x)dx$$

$$E(X) = \left[\frac{x^2}{2} + \frac{x^3}{3}\right]_{-1}^{0} + \left[\frac{x^2}{2} - \frac{x^3}{3}\right]_{0}^{1} = -\frac{1}{6} + \frac{1}{6} = 0$$

$$E(X^2) = \int_{-1}^{0} x^2(1 + x)dx + \int_{0}^{1} x^2(1 - x)dx$$

$$E(X^2) = \left[\frac{x^3}{3} + \frac{x^4}{4}\right]_{-1}^{0} + \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_{0}^{1} = \frac{1}{12} + \frac{1}{12} = 0.167$$

$$\sigma_x^2 = E[X - E(X)]^2 = E[X - \mu]^2 = E(X^2) - \mu^2$$

$$\sigma_x^2 = E(X^2) - \mu^2 = 0.167 - 0^2 = 0.167$$

6.3. Common Discrete Probability Distributions

6.3.1. Binomial Probability Distribution

The binomial probability distribution is a discrete probability distribution that provides many applications. It is associated with a multiple-step experiment that we call the binomial experiment. A binomial experiment exhibits the following four properties.

1. The procedure has a fixed number of trials.

2. The trials are independent. (The outcome of any individual trial does not affect the probabilities in the other trials.)

3. The outcome of each trial must be classifiable into one of two possible categories (*success* or *failure*).

4. The probability of a success, denoted by p, does not change from trial to trial.

If a procedure satisfies these four requirements, the distribution of the random variable (X) is called a binomial probability distribution (or binomial distribution). To calculate probabilities we use the following formula.

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \, for \, x = 0, 1, 2, ..., n$$

where

$x =$ the number of successes
$p =$ the probability of a success on one trial
$q =$ the probability of failure on one trial ($q = 1 - p$)
$n =$ the number of trials
$p(x) =$ the probability of x successes in n trials.

Expected value and variance of binomially distributed random variable [$X \sim Bin(n, p)$] can be obtained using the following.

$$E(X) = \mu = np$$

$$var(X) = \sigma^2 = np(1 - p) = npq$$

$$SD(X) = \sigma = \sqrt{np(1 - p)} = \sqrt{npq}$$

Example: A university found that 20% of its students withdraw without completing the introductory statistics course. Assume that 20 students registered for the course. Compute the probability

(a) exactly four will withdraw.

Let X be number of students who will withdraw without completing the introductory statistics course. From the given problem $p = 0.2 = 20\%, n = 20$ and $X \sim Bin(20, 0.2)$.

$$P(X = 4) = \binom{20}{4} 0.2^4 0.8^{16} = \frac{20!}{4!(20 - 4)!} 0.2^4 0.8^{16} = 0.2182$$

(b) at most two will withdraw.

$$\begin{aligned}P(X \leq 2) &= P(X=0)+P(X=1)+P(X=2) = \sum_{i=0}^{2} P(X=x_i)\\ &= \binom{20}{0}0.2^0 0.8^{20} + \binom{20}{1}0.2^1 0.8^{19} + \binom{20}{2}0.2^2 0.8^{18}\\ &= \frac{20!}{0!(20-0)!}0.2^0 0.8^{20} + \frac{20!}{1!(20-1)!}0.2^1 0.8^{19} + \frac{20!}{2!(20-2)!}0.2^2 0.8^{18}\\ &= 0.2061\end{aligned}$$

(c) more than three will withdraw.

$$\begin{aligned}P(X > 3) &= P(X=4)+P(X=5)+\ldots+P(X=20) = \sum_{i=3}^{20} P(X=x_i)\\ &= P(X \leq 3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)\\ &= 0.2061 + \frac{20!}{3!(20-3)!}0.2^3 0.8^{17}\\ &= 0.2054\end{aligned}$$

(d) the expected and standard deviation of withdrawals.

$$E(X) = np = 20 \times 0.2 = 4$$

$$var(X) = \sigma^2 = np(1-p) = npq = 20 \times 0.2 \times 0.8 = 3.2$$

$$SD(X) = \sqrt{np(1-p)} = \sqrt{npq} = \sqrt{20 \times 0.2 \times 0.8} = 1.788$$

6.3.2. Poisson Distribution

In this section we consider a discrete random variable that is often useful *in estimating the number of occurrences over a specified interval of time or space.* For example, the random variable of interest might be the number of arrivals at a car wash in one hour, the number of repairs needed in 10 miles of highway, or the number of leaks in 100 miles of pipeline. If the following two properties are satisfied, the number of occurrences is a random variable described by the Poisson probability distribution.

Properties of a Poisson Experiment

1. The probability of an occurrence is the same for any two intervals of equal length.

2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

The Poisson probability function is defined by the following equation.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where

$p(x) =$ the probability of x occurrences in an interval

$\lambda =$ expected value or mean number of occurrences in an interval.

For the Poisson probability distribution, X is a discrete random variable indicating the number of occurrences in the interval. Since there is no stated upper limit for the number of occurrences, the probability function $p(x)$ is applicable for values $x = 0, 1, 2, ...$ without limit. In practical applications, x will eventually become large enough so that $p(x)$ is approximately zero and the probability of any larger values of x becomes negligible.

A property of the Poisson distribution is that the mean and variance are equal. That is,

$$\boxed{E(X) = var(X) = \lambda}$$

Example: A student finds that the average number of amoeba in 10 ml of pond water is 4. Find the probability that in 10 ml of water from that pond there are

(a) exactly 5 amoeba.

Let Y be the number of amoeba found in 10 ml pond water. From the given question $\lambda = 4$ which implies that $Y \sim Poisson(\lambda)$.

$$P(X = 5) = \frac{e^{-4}4^5}{5!} = 0.156$$

(b) no amoeba.

$$P(X = 0) = \frac{e^{-4}4^0}{0!} = e^{-4} = 0.0183$$

(c) fewer than 3 amoeba.

$$\begin{aligned} P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) &= \sum_{i=0}^{2} P(X = x_i) \\ &= \frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{2!} \\ &= e^{-4} + 4e^{-4} + 8e^{-4} \\ &= 0.238 \end{aligned}$$

Exercise 1: If $X \sim Poisson(\lambda)$ with standard deviation of 2, then find $P(X = 3)$.

Exercise 2: If $X \sim Poisson(\lambda)$ and $P(X = 1) = P(X = 3)$, then find λ.

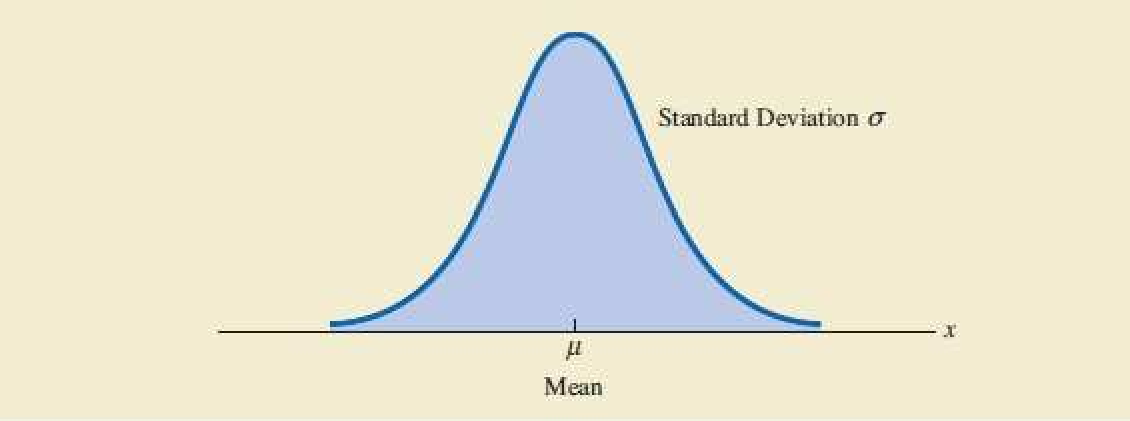
6.4. Common Continuous Distributions

6.4.1. Normal Distribution

The most important probability distribution for describing a continuous random variable is the normal probability distribution. The normal distribution has been used in a wide variety of practical applications in which the random variables are heights and weights of people, test scores, scientific measurements, amounts of rainfall, and other similar values. It is also widely used in statistical inference. In such applications, the normal distribution provides a description of the likely results obtained through sampling.

Normal Curve

The form or shape of the normal distribution is illustrated by the bell-shaped normal curve in the following figure. The probability density function (pdf) that defines the bell-shaped curve of the normal distribution follows.



If a random variable $X \sim N(\mu, \sigma^2)$ its probability density function (pdf) is given by:

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

where μ = mean, σ = standard deviation.

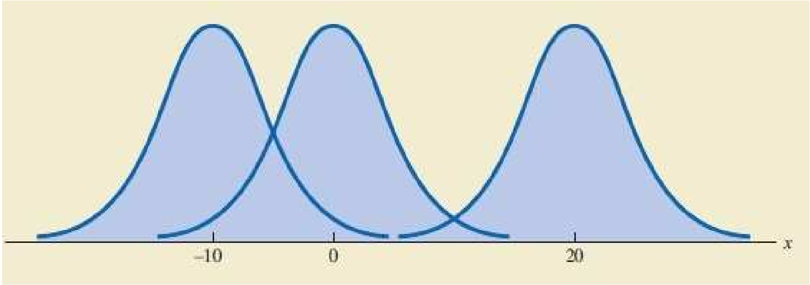
The normal curve has two parameters, μ and σ. They determine the location and shape of the normal distribution.

Properties of Normal Distribution

1. The entire family of normal distributions is differentiated by two parameters: the mean μ and the standard deviation σ.

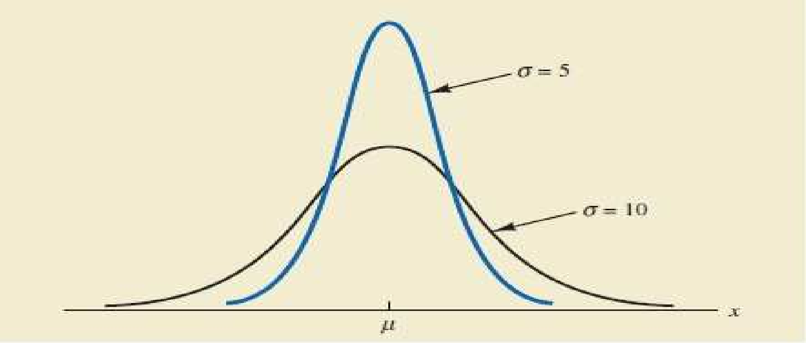
2. The highest point on the normal curve is at the mean, which is also the median and mode of the distribution.

3. The mean of the distribution can be any numerical value: negative, zero, or positive. Three normal distributions with the same standard deviation but three different means (-10, 0, and 20) are shown here.



4. The normal distribution is symmetric, with the shape of the normal curve to the left of the mean a mirror image of the shape of the normal curve to the right of the mean. The tails of the normal curve extend to infinity in both directions and theoretically never touch the horizontal axis. Because it is symmetric, the normal distribution is not skewed; its skewness measure is zero.

5. The standard deviation determines how flat and wide the normal curve is. Larger values of the standard deviation result in wider, flatter curves showing more variability in the data. Two normal distributions with the same mean but with different standard deviations are shown here.



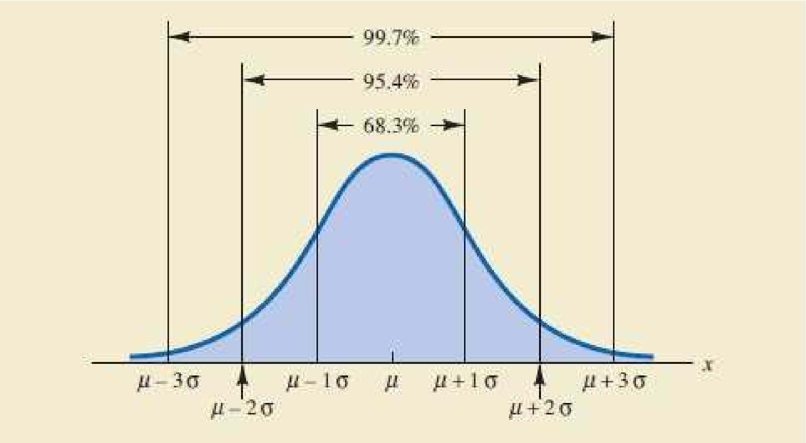
6. Probabilities for the normal random variable are given by areas under the normal curve. The total area under the curve for the normal distribution is 1. Because the distribution is symmetric, the area under the curve to the left of the mean is 0.50 and the area under the curve to the right of the mean is 0.50.

7. The percentage of values in some commonly used intervals are

   a) 68.3% of the values of a normal random variable are within plus or minus one standard deviation of its mean.

   b) 95.4% of the values of a normal random variable are within plus or minus two standard deviations of its mean.

   c) 99.7% of the values of a normal random variable are within plus or minus three standard deviations of its mean.



Standard Normal Probability Distribution

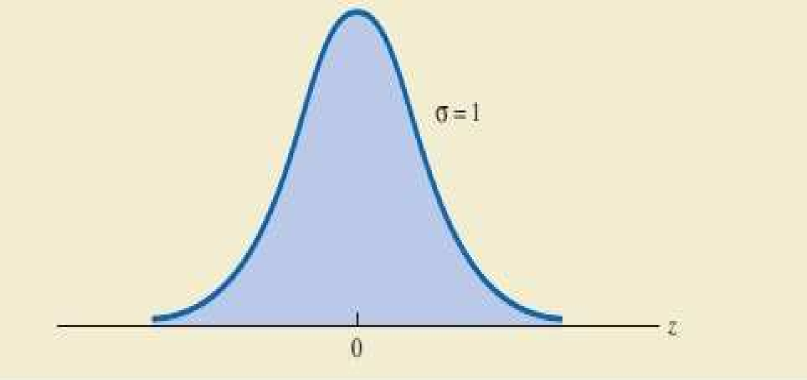
A random variable that has a normal distribution with a mean of zero and a standard deviation of one is said to have a standard normal probability distribution. The letter z is commonly used to designate this particular normal random variable, that is $z \sim N(0, 1)$. The reason for discussing the standard normal distribution so extensively is that probabilities for all normal distributions are computed by using the standard normal distribution. That is, when we have a normal distribution with any mean μ and any standard deviation σ, we answer probability questions about the distribution by first converting to the standard normal distribution. Then we can use the standard normal probability table and the appropriate z values to find the desired probabilities. Thus, we can convert using the following formula.

$$\boxed{z = \frac{x - \mu}{\sigma}}$$

Consequently, the standard normal density is given by:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp^{-z^2/2} \qquad -\infty < z < \infty$$

which is graphically shown below.



Example 1: Given that z is a standard normal random variable, compute the following probabilities.

a) $P(0 \leq z \leq 2.5) = 0.4938$

b) $P(z \geq 1) = P(z > 0) - P(0 < z < 1) = 0.5 - 0.3413 = 0.1587$

c) $P(z \leq 1) = P(z < 0) + P(0 < z < 1) = 0.5 + 0.3413 = 0.8413$

d) $P(1 \leq z \leq 1.5) = P(0 < z \leq 1.5) - P(0 < z \leq 1) = 0.4332 - 0.3413 = 0.0919$

e) $P(-1 < z < 1.5)$

$$\begin{aligned} P(-1 < z < 1.5) &= P(-1 < z < 0) + P(0 < z < 1.5) \\ &= P(0 < z < 1.5) = P(0 < z < 1) + P(0 < z < 1.5) \\ &= 0.3413 + 0.4332 = 0.7745 \\ &= 0.7745 \end{aligned}$$

Example 2: The college boards, which are administered each year to many thousands of high school students, are scored so as to yield a mean of 500 and a standard deviation of 100. These scores are close to being normally distributed. What percentage of the scores can be expected to satisfy each condition?

a) Greater than 600.

Let X be the score of students with mean $\mu = 500$, $\sigma = 100$ and $X \sim N(500, 100)$.

$$\begin{aligned}
P(X > 600) &= P\left[\frac{X-\mu}{\sigma} > \frac{600-\mu}{\sigma}\right] \\
&= P\left[z > \frac{600-500}{100}\right] \\
&= P\left[z > 1\right] \\
&= 0.1587
\end{aligned}$$

b) Less than 450.

$$\begin{aligned}
P(X < 450) &= P\left[\frac{X-\mu}{\sigma} < \frac{450-\mu}{\sigma}\right] \\
&= P\left[z < \frac{450-500}{100}\right] \\
&= P\left[z < -0.5\right] = P\left[z < 0\right] - P\left[-0.5 < z < 0\right] \\
&= P\left[z < 0\right] - P\left[0 < z < 0.5\right] \\
&= 0.5 - 0.1915 \\
&= 0.3085
\end{aligned}$$

c) Between 450 and 600.

$$\begin{aligned}
P(450 < X < 600) &= P\left[\frac{450-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{600-\mu}{\sigma}\right] \\
&= P\left[\frac{450-500}{100} < z < \frac{600-100}{100}\right] \\
&= P\left[-0.5 < z < 1\right] \\
&= P\left[-0.5 < z < 0\right] + P\left[0 < z < 1\right] \\
&= P\left[0 < z < 0.5\right] + P\left[0 < z < 1\right] \\
&= 0.1915 + 0.3413 \\
&= 0.5328
\end{aligned}$$

6.5. Exercises

1. State the conditions (assumptions) under which random variable can have a binomial distribution.

2. The probability that a freshman entering Haramaya University will survive first semester is 0.92. Assuming this pattern remain unchanged over the subsequent years, what is the probability that among 100 randomly selected freshmen in first semester,

   a) None will survive?

   b) Exactly 97 will survive?

   c) At least three will survive?

3. A normal distribution has mean $\mu = 62.4$. Find its standard deviation if 20% of the area under the curve lies to the right of 79.2.

4. Show that 65.24% of the observations in a normally distributed population lie between $\mu - 1.1\sigma$ and $\mu + 0.8\sigma$.

5. If a set of marks on a statistics examination are approximately normally distributed with a mean of 74 and a standard deviation of 7.9, then find

   (a) the lowest mark if the lowest 10% of the students are given F's.

   (b) the lowest mark to get grade A if the top 5% of the students are given A's.

   (c) the lowest mark to get grade B if the top 10% of the students are given A's and the next 25% are given B's.

7

One Sample Statistical Inference

One of the principal objectives of statistical analysis is to draw inference about the population on the basis of data collected by sampling from population. In other words, one is required to draw inference (or to generalize) about the population from the sample data. The inference to be drawn relates to some parameters of the population, such as the mean, standard deviation or some other feature like the proportion of an attribute occurring in the population. The two most important types of problems of inference in statistics are:

- Estimation of parameter or parameters and
- Testing of statistical hypothesis or hypotheses

In the absence of the complete data or information about the population, it would not be possible to determine the exact or true value of a parameter. It would be worthwhile to obtain from the sample data an *estimate* of the unknown true or exact value of the parameter or an interval of values in which the parameter lies, and also determine a procedure for determining the *accuracy* of the estimate. This type of inference is known as *estimation of parameters*. There are two types of estimation.

1. **Point estimation**: here the objective is to find *a single value* for the unknown parameter.
2. **Interval estimation**: here the objective is to find *an interval or range* of plausible values in which the unknown parameters lies.

7.1. Point Estimation

Suppose that we are concerned with the estimation of a parameter of a population from a given sample of the population. The procedure of point estimation consists of determining a single quantity from the sample values given such that the single number of fairly close to the unknown value of the parameter of the population. Suppose that the sample (of size n) drawn from the population is denoted by $x_1, x_2, ..., x_n$, and that the unknown parameter is denoted by θ. The point estimation of θ will be based on the sample observations $x_1, x_2, ..., x_n$. It will be a function of the sample observations $x_1, x_2, ..., x_n$, that is, a *statistic*. The statistic to be used for point estimation of θ is called *a point estimator* and is denoted by $\hat{\theta}$. When an actual set of sample values is given, we can compute a numerical value, which is called *a point estimate of* $\hat{\theta}$. The estimator $\hat{\theta}$ of the parameter θ is a function of the sample observations $x_1, x_2, ..., x_n$, and will assume different values corresponding to different sets of sample observations $x_1, x_2, ..., x_n$. For a given set of sample observations, we get point estimate of θ; this is one of the possible values of $\hat{\theta}$.

The point estimator of μ can assume an infinite number of values corresponding to the infinite set of (the numerical) sample values that $x_1, x_2, ..., x_n$ take. From one given set of sample values, that is, a *particular* set of numerical values one can compute one particular value of the estimator μ and this value is a point estimate of μ. Besides the mean $\bar{x} = \frac{x_1+x_2+...+x_n}{n} = \frac{\sum x_i}{n}$, there *may* be other types of estimator of μ, based on some other function of the same set of sample observations $x_1, x_2, ..., x_n$; in fact the sample median is also an estimator of μ. The question then arises: *which of the sample estimators is to be preferred and why.* This raises another question: *what should be the basis of selecting an estimator*, or in other words what should be the *criteria of a good estimator.* Without going into details, we would like to state below four desirable properties that an estimator should possess.

1. **Unbiasedness**: an estimator $\hat{\theta}$ of θ is said to be unbiased if $E(\hat{\theta}) = \theta$, i.e. its mean equal to the parameter value, otherwise it is said to be biased. The property of unbiasedness ensures that there will not be overestimation or underestimation.

2. **Minimum Variance (Efficiency)**: an estimator $\hat{\theta}$ is a function of the sample observations $x_1, x_2, ..., x_n$. An estimator with a smaller variance will have greater concentration near the parameter to be estimated. It will therefore be appropriate to select the estimator with the smallest variance. This would ensure greater accuracy.

3. **Consistency**: It refers to the effect of sample size on the accuracy of the estimator. A statistic is said to be consistent estimator of the population parameter if it approaches the parameter as the sample size increases, i.e. $\hat{\theta} \to \theta$ as $n \to N$.

4. **Sufficiency**: An estimator is said to be sufficient if it uses all the information about the

population parameter contained in the sample. For example, the statistic mean uses all the sample values in its computation while median and mode do not. Hence, the mean is a better estimator in this sense.

7.2. Interval Estimation

In previous section, we stated that a point estimator is a sample statistic used to estimate a population parameter. For instance, the sample mean $\bar{x}$ is a point estimator of the population mean μ and the sample proportion $\hat{p}$ is a point estimator of the population proportion p. Because a point estimator cannot be expected to provide the exact value of the population parameter, an interval estimate is often computed by adding and subtracting a value, called *the margin of error*, to the point estimate. The general form of an interval estimate is as follows:

$$\boxed{point\, estimate \pm margin\, of\, error}$$

Thus, a confidence interval (or interval estimate) is a range (or an interval) of values that is likely to contain the true value of the population parameter. A confidence interval associated with a degree of confidence, which is a measure of how certain we are that our interval contains the population parameter.

The degree of confidence is the probability $1-\alpha$ (often expressed as the equivalent percentage value) that the confidence interval contains the true value of population parameter. The degree of confidence is also called the level of confidence or the confidence coefficient.

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. Thus, the general form of an interval estimate of a population mean is

$$\bar{x} \pm margin\, of\, error$$

7.2.1. Interval Estimation for the Population Mean

1. Consider when σ is *known*

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $1-\alpha$ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution.

2. Consider when σ is *unknown*

   When developing an interval estimate of a population mean we usually do not have a good estimate of the population standard deviation either. In this case, we should have to estimate from sample values. When s is used to estimate σ, the margin of error and the interval estimate for the population mean are based on a probability distribution known as the t *distribution.* Thus, the interval estimate of μ becomes

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

   where s is the sample standard deviation, $1-\alpha$ is the confidence coefficient, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with $n-1$ degrees of freedom. The reason the number of degrees of freedom associated with the t value in above expression is $n-1$ concerns the use of s as an estimate of the population standard deviation σ. The expression for the sample standard deviation is

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

7.2.2. Interpretation of the Confidence Interval

1. If all possible samples of size n were drawn, then on an average $100(1-\alpha)\%$ of these samples would include the population mean within the interval around there sample means bounded by L and U.

2. If we took a random sample of size n from a given population, the probability is $1-\alpha$ that the population mean would lie between the interval L and U around the sample mean.

3. If a random sample of size n was taken from the population, we can be $100(1-\alpha)\%$ confident in our assertion that the population mean would lie around the sample mean in the interval bounded by the values L and U.

Example 1: Haramaya University wishes to estimate the average age of students who graduate with B.Sc. degree. A random sample of 625 graduating students showed that the average age was 24 with a standard deviation of 5 years. Construct the 95% confidence interval for the true average age of all such graduating students at the University and interpret it.

Solution:

Let μ is the true average age of all graduating students with B.Sc. degree from the University. From the sample data we have $n = 625$, $\bar{x} = 24$ and $s = 5$.

$$100 \times (1 - \alpha)\% = 95\%$$

$$\alpha = 0.05,\ \alpha/2 = 0.025,\ z_{\alpha/2} = z_{0.025} = 1.96$$

Thus $100 \times (1 - \alpha)\%$ confidence interval for μ is

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right] = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\left[24 - 1.96 \times \frac{5}{\sqrt{625}},\ 24 + 1.96 \times \frac{5}{\sqrt{625}}\right]$$

$$[24 - 0.392,\ 24 + 0.392] = [23.608,\ 24.392]$$

Interpretation:

On average the true average age of all graduating students with B.Sc. degree from the University is found between 23.608 and 24.392 at 95% confidence level.

Example 2: An airline wants to evaluate the depth perception of its pilots over the age of 50. A random sample of 14 airline pilots over the age of 50 are asked to judge the distance between two markers placed 20 feet apart at the opposite end of the laboratory. The sample data listed here are the pilots' error (recorded in feet) in judging the distance.

2.7	2.4	1.9	2.6	2.4	1.9	2.3
2.2	2.5	2.3	1.8	2.5	2.0	2.2

Use the sample data to place a 95% confidence interval on μ, the average error in depth perception for the company's pilots over the age of 50.

Solution:

Since the sample size small, it is appropriate to construct the confidence interval based on the t distribution. Let y be the average error in depth perception for the company's pilots over the age of 50. We can verify that $\bar{y} = 2.26$ and $s = 0.28$.

Referring to t table, the t-value corresponding to a $\alpha = 0.025$ and $n - 1 = 13$ is 2.160. Hence, the 95% confidence interval for μ is

$$\left[\bar{y} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{y} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right] = \bar{y} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

$$\left[2.26 - 2.16 \times \frac{0.28}{\sqrt{14}},\ 2.26 + 2.16 \times \frac{0.28}{\sqrt{14}}\right] = [2.10,\ 2.42]$$

Interpretation:

Therefore, we are 95% confident that the average error in the pilots' judgment of the distance is between 2.10 and 2.42 feet.

7.3. Hypothesis Testing

7.3.1. Introduction

A statistical hypothesis is a conjecture (an assumption) about a population parameter which may or may not be true. Hypothesis testing is a statistical procedure which leads to take a decision about such an assumption for the population parameter being correct or not, by using data obtained from the sample. In hypothesis testing, the researcher must define the population under study, state the particular hypothesis that will be checked, give the significance level, select sample from the population, perform calculations required for statistical test and reach conclusion. It is already expressed that a statistical hypothesis may or may not true. For each situation, there two types of statistical hypotheses.

1. **Null Hypothesis (H_0):** is a statistical hypothesis that states there is no difference between a parameter and a specific value or hypothesized value. The statement is not rejected unless there is convincing sample evidence that it is false. Often represents the status quo situation or an existing belief.

2. **Alternative Hypothesis** (H_a or H_1): is a statistical hypothesis that states there exists a difference between a parameter and a specific value or hypothesized value. It is the assertion of all situations not covered by the null hypothesis.

7.3.2. Errors in Hypothesis Testing

There are two types of error in hypothesis testing.

1. **Type-I Error:** is a mistake occurred if one rejects the null hypothesis which is actually true.

2. **Type-II Error:** is a mistake occurred if one failed to reject the null hypothesis which is actually false.

There are four possible outcomes of any hypothesis test, two of which are correct and two of which are incorrect. The incorrect ones are called type I and type II.

	State of Nature	
Decision	H_0 True	H_0 False
Do not reject H_0	Correct decision $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1-\beta)$

7.3.3. Hypothesis Testing About the Population Mean

1. State (formulate) the null and alternative hypotheses. The hypotheses may be either of the following.

   ▷ $H_0 : \mu = \mu_0 \, versus \, H_a : \mu \neq \mu_0$ - Two tailed test

   ▷ $H_0 : \mu = \mu_0 \, versus \, H_a : \mu > \mu_0$ - Right tailed test

   ▷ $H_0 : \mu = \mu_0 \, versus \, H_a : \mu < \mu_0$ - Left tailed test

2. Choose the level of significance α, the probability of making a Type I Error if H_0 is true.

3. Calculate *the appropriate test statistic.* The following is a general formula for a test statistic that will be applicable in many of hypothesis tests.

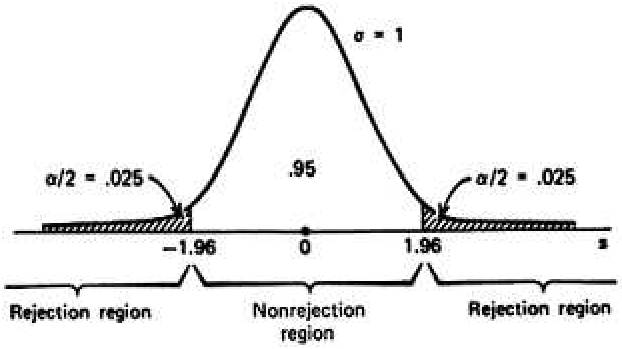
$$\boxed{test\,statistic = \frac{statistic - hypthesized\,parameter}{standard\,error\,of\,statistic}}$$

   ▷ Use t statistic if n is small, $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_\alpha(n-1)$.

   ▷ Use z statistic if n is large enough, $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$.

4. Obtain the tabulated (critical) value. For *two tailed test* the critical value is $z_{\alpha/2}(t_{\alpha/2})$, for *right tailed* $z_\alpha(t_\alpha)$ and for *left tailed* $-z_\alpha(-t_\alpha)$ respectively.

5. Define the critical (rejection) region. If the value of the test statistic falls in the critical region (rejection region), reject the null hypothesis; otherwise do not reject it.



6. State the conclusion.

Examples

1. A professor wants to know if the introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score at least 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor certain that the mean score for the class on the test would be at least 70 at 0.05?

   Solution:

   First, compute the sample mean and standard deviation.

$$\bar{x} = \frac{\sum x_i}{n} = 79.17, \quad s = \sqrt{\frac{\sum(x_i - \bar{x})}{n-1}} = 13.17$$

   ✓ Formulate the null and alternative hypotheses.

$$H_0 : \mu = 70$$

$$H_a : \mu > 70$$

   ✓ Specify the level of significance ($\alpha = 0.05$).

   ✓ Compute appropriate test statistics. Since the sample size is small t is appropriate in this case.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{79.17 - 70}{13.17/\sqrt{6}} = 1.71$$

   ✓ Obtain the tabulated value from the t-table which is $t_\alpha(n-1)$.

$$t_\alpha(n-1) = t_{0.05}(6-1) = t_{0.05}(5) = 2.015$$

- ✓ Here we define the rejection region. Reject H_0 if $t > 2.015$ otherwise do not reject H_0. Since the computed t value 1.71 is not greater than critical value 2.015, we fail to reject the null hypothesis.
- ✓ Interpretation:

  Hence, the professor is not certain on the math test of the class which states that it is at least 70 at 5% level of significance.

2. A merchant believes that the average age of customers who purchase a certain brand of wears is 13 years of age. A random sample of 35 customers had an average age of 15.6 years. At 1% should this conjecture be rejected. The standard deviation of the population is 1 year.

   Solution:
   Suppose x be the age of customers who purchase a certain brand of wear. Given $\mu_0 = 13$, $n = 35$, $\bar{x} = 13$ and $s = 1$.

   - ✓ Formulate the null and alternative hypotheses.

   $$H_0 : \mu = 13$$
   $$H_a : \mu \neq 13$$

   - ✓ Specify the level of significance ($\alpha = 0.01$). This test is a two tailed test, so you divide the alpha level by two, $\alpha/2 = 0.005$.
   - ✓ Compute appropriate test statistics. Since the sample size is large ($n > 30$) z is appropriate in this case.

   $$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15.6 - 13}{1/\sqrt{35}} = \frac{2.6}{0.169} = 15.38$$

   - ✓ Obtain the tabulated value from the standard normal table which is $z_{\alpha/2}$.

   $$z_{\alpha/2} = z_{0.005} = 2.575$$

   - ✓ Define the rejection region. Reject H_0 if $|z| > z_{\alpha/2}$ otherwise do not reject H_0. Since the calculated z value 15.38 is much greater than tabulated value 2.575, the null hypothesis is rejected.
   - ✓ Interpretation:

     Therefore, the merchants believe is not correct at 1% level of significance.

7.4. Exercises

1. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the sample mean and sample standard deviation of number of concurrent users at 100 randomly selected times is 37.7 and 9.2, respectively.

   (a) Construct a 90% confidence interval for the mean number of concurrent users.

   (b) Do these data provide significant evidence, at 1% significance level, that the mean number of concurrent users is greater than 35?

2. To assess the accuracy of a laboratory scale, a standard weight that is known to weigh 1 gram is repeatedly weighed 4 times. The resulting measurements (in grams) are: 0.95, 1.02, 1.01, 0.98. Assume that the weighings by the scale when the true weight is 1 gram are normally distributed with mean μ.

   (a) Use these data to compute a 95% confidence interval for μ.

   (b) Do these data give evidence at 5% significance level that the scale is not accurate?

3. A local juice manufacturer distributes juice in bottles labeled 32 ounces. A government agency thinks that the company is cheating its customers. The agency selects 35 of these bottles, measures their contents, and obtains a sample mean of 31.7 ounces with a standard deviation of 0.70 ounce. Use a 0.01 significance level to test the agency's claim that the company is cheating its customers.

8

Simple Correlation and Linear Regression Analysis

8.1. Correlation Analysis

Correlation is a mathematical tool desired towards measuring the degree of linear relationship (degree of association) between the variables. Correlation that involves only two variables is called simple correlation and which involves more than two variables is called multiple correlations.

Covariance is a measure of the joint variation in two variables, i.e. it measures the way in which the values of the two variables vary together.

1. If the covariance is zero, there is no linear relationship between the two variables.
2. If it is negative, there is an indirect linear relationship between them.
3. If the covariance is positive, there is a direct linear relationship between the variables.

Pearson's Coefficient of Correlation

Pearson's coefficient of correlation (r) is used to measure the strength of the linear relationship between two variables.

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

The value of r is always between -1 and +1 inclusive.

Interpretation of r

1. If the value of r is -1 or +1, there is perfect negative or perfect positive linear relationship between the variables.

2. If the value of r is approximately -1 or +1, there is a strong negative or strong positive linear relationship between the variables.

3. If r is -0.5 (or approximately -0.5) or +0.5 (or approximately +0.5), there is moderate negative or moderate positive linear relationship between the variables.

4. If r is 0, there is no linear relationship.

8.2. Simple Linear Regression

Regression is defined as the estimation or prediction of the unknown value of one variable from the known values of one or more variables. It is also functional relationship between two or more variables. The variable whose values are to be estimated or predicted is known as dependent or explained variable while the variable which are used in determining the value of the dependent variable are called independent or predictor variables. The regression study that involves only two variables is called simple regression and the regression analysis that studies more than two variables is called multiple regression.

Regression Equation: is a mathematical equation that defines the relationship between two variables. Regression of y on x is given by

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

where

y is the dependent variable,
x is the independent variable,
α is constant term (intercept),
β is slope (change in y for a unit change in x) and
ε is the error term.

To estimate the regression coefficients ($\hat{\alpha}$ and $\hat{\beta}$), the procedure is minimizing the sum of the squares of the errors.

Let the estimated model be

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Then, from sample data the values of $\hat{\alpha}$ and $\hat{\beta}$ can be obtained as follows:

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}; \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Interpretation of the slope

1. If $\hat{\beta}$ is positive, there is a direct relationship between the two variables.

2. If $\hat{\beta}$ is zero, there is no linear relationship between the two variables.

3. If $\hat{\beta}$ is negative, there is indirect linear relationship between the two variables.

8.3. Coefficient of Determination (R^2)

The coefficient of determination tells *how well the estimated model fits the data.* For simple linear regression (two variables case), it is defined as the square of the sample correlation coefficient, and denoted by r^2. Hence r^2 measures the proportion or percentage of the variation in the dependent variable explained by the independent variable. r^2 is a nonnegative quantity which lies in the limits 0 and 1. If it approaches to 1, it means a good fit and if it approaches 0, no relationship between the variables.

Example

A researcher wants to find out if there is a relationship between the heights of sons and the heights of their fathers. In other words, do taller fathers have taller sons? The researcher took a random sample of 6 fathers and their 6 sons. Their height in inches is given below in an ordered array.

Father (X)	63	65	66	67	67	68
Son (Y)	66	68	65	67	69	70

(a) Find the correlation coefficient and interpret.

$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}} \\ &= \frac{6 \times 26740 - 396 \times 405}{\sqrt{6 \times 26152 - (396)^2} \times \sqrt{6 \times 27355 - (405)^2}} \\ &= 0.597 \end{aligned}$$

(b) Estimate the regression model of height of sons on height of fathers and interpret the estimated parameters.

$$\begin{aligned}\hat{\beta} &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \\ &= \frac{6 \times 26740 - 396 \times 405}{6 \times 26152 - (396)^2} \\ &= 0.625 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 67.5 - 0.625 \times 66 \\ &= 26.25\end{aligned}$$

Hence, the estimated regression model is:

$$\hat{y} = 26.25 + 0.625x$$

For one inch increment in fathers height, the height of the son is increased by 0.625 inches.

(c) Compute coefficient of determination and interpret the result.

$$R^2 = (r)^2 = 0.597^2 = 0.357$$

Thus 35.7% of variation in the dependent variable (son height) is accounted for by the variation of the independent variable (father height).

8.4. Exercises

1. Given the following data:

AGE	SBP
15	116
20	120
25	130
30	132
40	150
50	148

a) Compute regression a line (systolic blood pressure (SBP) on AGE) and interpret the results.

b) Compute correlation coefficient between SBP and AGE and also interpret the result.

c) How much the variance of SBP can be explained by the fact that there is variability in AGE?

2. An experiment was conducted to study the effect on sleeping time of increasing the dosage of a certain barbiturate. Three readings were made at each of three dose levels:

Sleeping time (Hrs)	Dosage
4	3
6	3
5	3
9	10
8	10
7	10
13	15
11	15
9	15

a) Plot the scatter diagram.

b) Determine the regression line relating dosage (X) to sleeping time (Y).

c) What is the predicted sleeping time for a dose of 12?

THE END!!

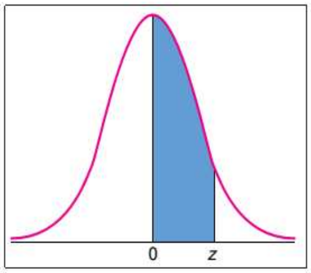


TABLE A Areas of a Standard Normal Distribution (Alternate Version of Appendix I Table 4)

The table entries represent the area under the standard normal curve from 0 to the specified value of *z*.

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
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999

For values of *z* greater than or equal to 3.70, use 0.4999 to approximate the shaded area under the standard normal curve.