Bio/Statistics
SPHM 5011

Lecture Notes

Awol S.
Department of Public Health
St. Paul's Hospital Millennium Medical College
Addis Ababa, Ethiopia

Contents

Chapter 1

Introduction

1.1 Basic Terms

Before getting involved in the subject matter in detail, let us define some of the terms used extensively in the field of statistics.

1.1.1 Datum vs Data

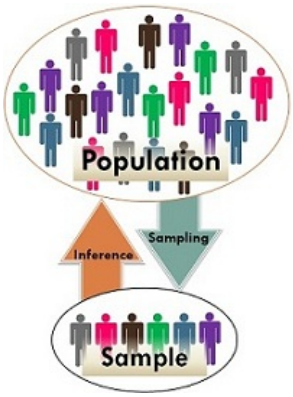
Datum is an observed value representing *one or more characteristics of an object*. It is also known as an *observation* or an *item* or a *case* or a *unit*. For example, if the height of an individual is 1.72m, then 1.72m is an observation. Similarly, if the height and age of a person are given as 1.65m and 27yrs, respectively, then (1.65m, 27yrs) is also a single observation.

Data is a collection of observed values (observations or cases) of *some* objects. For example, given the heights of two individuals as 1.72m and 1.69m. These values can be considered as data consisting of two observations. In addition, the height and age of two persons given as (1.65m, 27yrs) and (1.79m, 35yrs) are data.

1.1.2 Population vs Sample

A statistical *population* consists of *all* objects under study. Each object, upon which an observation is recorded, is called a *unit of analysis* or *study unit*. The total number of objects in a population is called *population size* (mostly denoted by N).

A *sample* is the *subset* of a population. The number of objects in a sample is also called *sample size* (mostly denoted by n).



Example 1.1. A study of 300 married women in Addis Ababa city administration showed that 99% of them have comprehensive knowledge of HIV/AIDS. Here, the population consists of "All married women in Addis Ababa city administration" and the sample consists of "The 300 married women selected from Addis Ababa city administration". The unit of analysis is a married woman.

Example 1.2. A study of 250 patients admitted to St. Paul's Hospital during the past year revealed that, on the average, the patients lived 150 kms away from the hospital. "All patients admitted to St. Paul's Hospital during the past year" is the population and "The 250 patients selected from those who admitted to St. Paul's Hospital during the past year" is the sample. Now the unit of analysis is a patient.

Example 1.3. Of 58 students who joined a nursing school in a certain time, only 5 failed to graduate. The population is "All the 58 students who joined the school" and all are taken into consideration. Hence, no sample is taken. The unit of analysis is " a student".

Example 1.4. Of all printed circuit boards manufactured this month, 5% are defective. The population consists of "All circuit boards manufactured this month" and the unit of analysis is "a circuit board".

1.2 Definitions of Statistics

Statistics can be defined in two senses: *plural* sense (as statistical data) and *singular* sense (as statistical methods).

1.2.1 Plural Sense: as Statistical Data

In the plural sense definition, statistics are collection of *facts* and *figures*. This meaning of the word is widely used when a reference is made to facts and figures on a certain characteristic. For example: malaria statistics, sales statistics, employment statistics, e.t.c. 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, the data must possess certain characteristics.

The data should be in an aggregate form of facts. Single or isolated fact cannot be called statistics as this cannot be compared or related to other figures within the same framework. That is, a single fact, even though numerically expressed, cannot be called statistics. Accordingly, if a new employee says that "I earn Birr 300000 per year", it would not be considered

as statistics. On the other hand, if we say that the average salary of a new employee is Birr 300000 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 new employees.

In addition, the data must be placed in relation to each other. The phrase, *placed in relation to each other* suggests that the facts should be comparable. The comparison of facts and figures is conducted regarding the same characteristics over a period of time from a single source or it may be from various sources at any one given time. For example, prices of different items in a store as such would not be considered statistics. However, prices of one product in different stores constitute statistical data since these prices are comparable. Also, the changes in the price of a product in one store over a period of time would also be considered as statistical data since these changes provide for comparison over a period of time. The general rule is, the comparisons must relate to the same phenomenon so that *likes are compared with likes* and *oranges are not compared with apples.*

Hence, the term statistics refers to collection of facts. However, statistics involves much more than numerical facts.

1.2.2 Singular Sense: as Statistical Method

In the singular sense definition, statistics refers to a discipline concerned with the extraction of *relevant information* from data with the aim to provide solutions to a problem; make more informed and better decision-making; and design new products and processes. In this meaning, statistics is concerned with the development and application of methods and techniques for collecting, organizing, presenting, analyzing data and interpreting the results of the analysis.

Accordingly, a statistical investigation involves five stages: data collection, data organization, data presentation, data analysis and interpretation of results.

1. **Collection of data**: Data collection is the first stage in any statistical investigation. It involves the process of obtaining (gathering) a set of related measurements or counts to meet predetermined objectives. The data might be obtained from either *primary* or *secondary* sources, see Section 1.7.2 for details.

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 stage of a statistical investigation is *describing the properties of the data in a summary form* called *data organization.* Since there may be omissions, inconsistencies, ambiguities, irrelevant answers and recording errors, the data should be corrected first. Hence, the first step in the organization of data is *editing.* Then, once the data is edited, the second step is *classification*, that is, the collected data should be arranged according to some common characteristics. The last step of the organization of data is presenting the classified data in *tabular form*, i.e., using rows and columns.

3. **Presentation of data**: The collected data is usually presented using charts (diagrams) and graphs. The purpose of data presentation is to have an overview of what the data actually looks like. Charts and graphs provide visually an intuitive understanding of

data. They have greater attraction and memorizing effect than figures, and also facilitate comparison.

4. **Analysis of data**: The analysis of data is the extraction of a *few summarized and comprehensive numerical values* with the goal of discovering useful information. This is the most important part of a statistical investigation. The analysis may require simple to sophisticated statistical techniques.

5. **Interpretation of results**: This is the last stage of a statistical investigation. Once the data is analyzed, the main job is attaching physical meaning or interpretation to those numerical results obtained from the analysis. The interpretation must be true in its meaning and sense. No pre-conceived ideas should be thrusted on the numerical results obtained out of the analysis. Also, no attempts should be made to draw more conclusions than the results are actually liable to.

1.3 Classification of Statistics

Based on the scope of the decision, statistics can be classified into two; *descriptive* and *inferential* statistics.

1.3.1 Descriptive Statistics

Descriptive statistics is concerned with *organizing and summarizing* the most important features of the data *without going beyond the data itself.* That is, descriptive statistics describes only the data that we have, without attempting to conclude anything that goes beyond the data. It includes *methods of data organization* like classification, tabulation and frequency distributions; *methods of data presentation* like charts (diagrams) and graphs; and certain indicators of data like *measures of central tendency* and *measures of variation.*

1.3.2 Inferential Statistics

Inferential statistics is concerned with drawing statistically *valid conclusions* about the characteristics of the population based on the results obtained from the sample. In this form of statistical analysis, *descriptive statistics is linked with probability theory.* Performing *hypothesis testing*, determining *relationship (association) between characteristics* and making *predictions* (*forecasting*) are also inferential statistics.

Example 1.5. Suppose a researcher is interested to know the average mark of a certain class in "Statistics" course. From a class of size 150, s/he took a random sample of 9 students and gave them an exam out of 100. Then, s/he got the average score 76. The statement "The average score of the 9 selected students is 76" is descriptive where as the statement "The average score of the class is 76" is inferential.

Example 1.6. We want to compare the average mark of boys and girls. Suppose we took a random sample of 7 boys of the total 85 boys and 6 girls of the total 60 girls, and gave them all the same exam. The average score of the 7 boys became 87 and that of the 6 girls became 92. Now the statements "The average score of the 7 boys is lower than that of the 6 girls" and "The 6 selected girls did better than the 7 selected boys" are both descriptive. But, saying "Girls did better in the exam than boys" is an inferential statistical statement.

Exercise 1.1. Classify each of the following statements as *descriptive* and *inferential* statistics.

1. The average age of the students in this class is 21 years.
2. There is a strong association between smoking and lung cancer.
3. The price of wheat will be increased by 5% in the coming year.
4. Teaching statistics by computer method is more effective than teaching by lecture method.
5. Of the students enrolled in St.PHMMC this year, 74% are female and 26% are male.
6. The chance of winning the Ethiopian National Lottery in any day is 1 out of 167000.

1.4 Applications, Uses and Limitations of Statistics

1.4.1 Applications of Statistics

There is almost no walk of life that has not been affected by statistics - ranging from a simple household to big business and the government. Statistics plays a very important role in a wide range of fields; natural, social and physical sciences. There is hardly any scientific research going on without the use of statistics in one form or another. Following are just a few examples illustrating the use of statistical inference in different situations.

The effectiveness of a new drug is determined by statistical experimentation and evaluation in medical and pharmaceutical research. Utilizing inferential statistics, researchers can design experiments with small randomly selected sample of patients and the results of tests of the drug may be used about the entire population of patients who may use the drug if it is introduced.

In botany, statistics is used in evaluating the effects of temperature and other climatic conditions and types of soil on the health of plants.

In agriculture, experiments about crop yields, types of fertilizers and types of soils under different types of environments are commonly designed and analyzed through statistical methods and concepts.

In economics, statistics is used for modeling functional relationships between different characteristics.

In marketing research, statistical tools are indispensable in studying consumer behavior, effects of various promotional strategies and so on. For instance, market researchers are often interested in the relationship between advertising and sales. A data set of randomly chosen sales and advertising figures for a given firm may be of some interest in itself, but the information in it is much more useful if it leads to implications about the underlying process - the relationship between the firm's level of advertising and the resulting level of sales. An understanding of the true relationship between advertising and sales - the relationship in the

population of advertising and sales possibilities for the firm - would allow us to predict sales for any level of advertising and thus to set advertising at a level that maximizes profits.

In education, statistical methods are used to study the effects of certain training.

In a public health study, epidemiologists want to know whether smoking is linked to a particular demographic class individuals.

In quality control, statistical methods help to check whether a product satisfies a given standard. A quality control engineer at a plant making disk drives for computers needs to make sure that no more than 3% of the drives produced are defective. The engineer may routinely collect random samples of drives and check their quality. Based on the random samples, the engineer may then draw a conclusion about the proportion of defective items in the entire population of drives.

In physical sciences, the science of meteorology uses statistics for analyzing the data gathered by satellites and for forecasting weather conditions.

Statistics helps to enhance the power of decision making in the face of uncertainty. For instance, it is immensely useful for politicians to determine their chances of winning and study the attitudes of their people on their policies.

Statistics is also used in other areas like insurance companies, banks, public utility companies and so on. A bank may be interested in assessing the popularity of a particular model of automatic teller machines. The machines may be tried on a randomly chosen group of bank customers. The conclusions of the study could then be generalized by statistical inference to the entire population of the bank's customers.

1.4.2 Uses of Statistics

- **Reduction and summarization of data**: It is generally not possible to draw any conclusions from the raw data that is voluminous and in haphazard manner. Statistics condenses and summarizes a large mass of data, and presents facts into a few presentable and understandable numerical figures.
- **Facilitating comparison of data**: Arrangement of data with respect to different characteristics facilitates comparison. Statistical devises such as averages, percentages, ratios, e.t.c. are used for this purpose.
- **Test of the validity of important conjectures**: For instance, hypothesis like whether a new medicine is effective in curing a disease can be tested using statistical tools.
- **Determining the relationship (association) between characteristics**: Statistical techniques assist in determining the degree of relationship and establishing cause-and-effect relationship between two or more characteristics.
- **Prediction and forecasting future behaviour**: Statistical methods are highly useful tools in analyzing past data and forecasting future trends,e.g., forecasting the number

of new HIV/AIDS cases in the coming year or determining the probability of occurrence of an outbreak in a certain area in the coming five years.

1.4.3 Limitations of Statistics

- Statistics does *not deal with a single observation*, rather, as discussed earlier, it only deals with aggregate of facts. For example, the mark 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 results are *true on average*; i.e. for the majority of cases. In other words, statistics is not an exact science and hence, statistical conclusions are not universally true. That is, statistical laws are *not* universally true unlike the laws of mathematics, chemistry or physics.
- Statistical methods are *liable to be misused or misinterpreted*. Statistical interpretation requires a high degree of skill and understanding of the subject. Often, misuse of statistics happens due to lack of knowledge.

1.5 Biostatistics

Biostatistics is the application of statistical methods to public health and biomedical sciences. As technology progresses, public health and medicine are becoming increasingly quantitative rather than descriptive information. Therefore, much of current medical research are becoming reliant on statistical methodologies.

Biostatistics covers applications and contributions not only from health, medicines and, nutrition but also from fields such as genetics, biology, epidemiology, and many others.

1.6 Variables

A variable is a characteristic or an attribute that can assume *different* values. For example: height, family size, gender, marital status, $\cdots$.

1.6.1 Types of Variables: Quantitative vs Qualitative

Based on the values that variables assume, variables can be classified into two as *quantitative* and *qualitative* (*categorical*).

- **Quantitative variables**: Quantitative variables are those variables which assume numeric values. These variables are numeric in nature. Height and family size are examples of quantitative variables.

  Quantitative variables are again further classified into two; *discrete* and *continuous* variables.

  - **Discrete variables**: Discrete variables are those variables that assume a countable number of distinct and recognizable whole number values. Family size, number of children in a family, number of cars at a traffic light, mother's history of number of births (parity) and pregnancies (gravidity), $\cdots$ are some examples of discrete variables. Such variables can assume a *finite* number of possible values or a *countably infinite* number of values. The values of these variables are obtained by counting $(0, 1, 2, \cdots)$.
  - **Continuous variables**: Continuous variables can take any value including decimals. One is not restricted, in principle, to particular values such as the integers of the discrete scale. The restricting factor is the degree of accuracy of the measuring instrument. These variables theoretically assume an *infinite* number of possible values. Their values are obtained by measuring. Examples of continuous variables are height, weight, time, temperature, $\cdots$

- **Qualitative variables**: Qualitative variables are, on the other hand, those variables that assume non-numeric values called *categories* or *levels* or *groups*. For example, gender is a qualitative variable with two categories (levels): male and female. Marital status is also qualitative with, say, four categories: single, married, divorced, other. Numerical codes might be used for facilitating data collection, entry and analysis. But, the varaible is qualitative even though the values appear as numeric.

  Based on the number of values that qualitative variables assume, they can be classified as *binary* (*dichotomous*) and *multinomial* (*polytomous*).

  - **Binary variables**: Binary variables often consist of 'either-or' type responses. That is, these variables have only two categories (levels). For example, gender (female, male), patient status (cured, not cured), pregnancy status (pregnant, not pregnant), exam result of a student (pass, fail), smoking status (smoker, non-smoker) are binary variables.
  - **Multinomial variables**: Multinomial variables are those qualitative variables with *three or more* categories. For example, blood type (A, B, AB, O), marital status (single, married, divorced, other), religion (orthodox, muslim, protestant, $\cdots$), color (blue, red, green, black, $\cdots$) are multinomial variables.

Example 1.7. Classify each of the following variable as qualitative or quantitative and if it is quantitative classify as discrete or continuous.

1. Color of automobiles in a dealer's show room
2. Age of patients seen in a dental clinic
3. Number of seats in a movie theater
4. Blood pressure of a patient
5. The distance between a hospital to a house
6. Classification of patients based on nursing care needed (complete, partial, safers)
7. Temperature in a class room

8. Number of tomatoes on each plant on a field
9. Weight of newly born babies in a hospital during a year
10. Number of heart attacks
11. Temperature (very cold, cold, hot, very hot)
12. Heart rate
13. Cholesterol level
14. A woman may never have conceived, conceived but spontaneously aborted, or given birth to a live infant.
15. Diastolic blood pressure (hypertension > 90 mmHg, normotension ≤ 90 mmHg)

1.6.2 Measurement Scales: Nominal, Ordinal, Interval and Ratio

Before explaining what measurement scales are, first, let us consider the following two cases:

Case 1:

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

Who plays better? What is the average t-shirt number?

Case 2:

- Mr A scored 5 in Stat quiz.
- Mr B scored 6 in Stat quiz.

Who did better? What is the average score?

Based on the number on the t-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 not possible to find the average t-shirt numbers because the numbers on the t-shirts are simply codes but it is possible to obtain the average test score.

Therefore, a scale of measurement shows the *amount of information* contained in the value of a variable, and what *mathematical operations* and *statistical analysis* are permissible to be done on the values of the variable. There are four levels of measurement. These levels, from the weakest to the strongest, in order are: *nominal*, *ordinal*, *interval* and *ratio*.

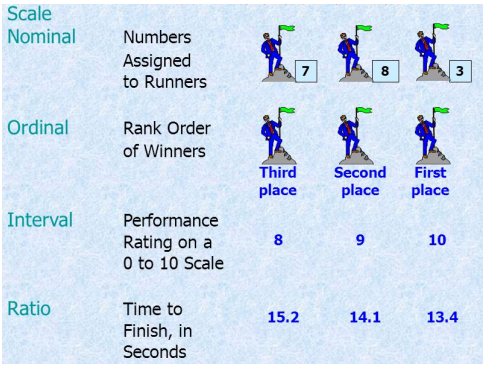
1. **Nominal variables**: Nominal variables are qualitative variables which show classification of individuals into *mutually exclusive* (*non-overlapping*) and *exhaustive* categories without any associated ranking. For example; gender, religion, ethnicity, eye color (black, brown, etc), $\cdots$ are nominal variables. Numbers may be assigned to the categories of these variables for coding purposes. But, it is not possible to compare individuals based on the numbers assigned to the categories. The only mathematical operation permissible to be done on the values of these variables is counting. Mobile number and patient's card number are also examples of nominal variables.

2. **Ordinal variables**: Ordinal variables are also qualitative variables whose values can be ordered and ranked. However, the ranks only indicate as to which category *greater* or *better* but there is *no precise difference* between the categories of the variable. Example: grade scores (A, B, C, D, F), academic qualifications (B.Sc., M.Sc., Ph.D.), strength (very weak, week, strong, very strong), health status (very sick, sick, cured), strength of opinions in likert scales (strongly agree, agree, neutral, disagree, strongly disagree), stages of breast cancer (I, II, III, IV).

3. **Interval variables**: Interval variables are quantitative variables and identify not only as to which category is *greater* or *better* but also by *how much*. The distances represented by the differences between consecutive values are equal; that is, interval variables have equal intervals. An interval scale is the stronger form of measurement but, there is *no true zero*. That is, the zero point is a matter of convention or convenience and not a natural or fixed zero point. Zero indicates *low* than *empty*. That is, zero does not show absence of a phenomenon. For example, a temperature of 0 °C does not mean there is no temperature but, rather, it is too cold. Similarly, if a student scores 0 in a certain course, it does not mean the student has no knowledge in the course at all.

4. **Ratio variables**: Ratio variables are quantitative variables in which the ratio of two values are meaningful. These scales are the highest form of measurements. Unlike interval variables, zero for ratio variables indicates that absence of the characteristic being studied. Hence, there is a *true (absolute) zero*, and the ratio of two values is meaningful. All mathematical operations are allowed to be operated on the values of these variables. Examples: height, weight, heart rate, $\cdots$.

Summary: All qualitative variables are either nominal or ordinal scales. And all quantitative variables are either interval or ratio scales. Most statistical analyses do not distinguish interval and ratio scale variables. As a result, in most practical aspects, they are grouped under metric (scale) variables.



Exercise 1.2. Identify the scale of measurement of the following variables.

1. The response after treatment of a patient (improved, the same or worse).
2. Women's choice of contraceptive methods.
3. Body Mass Index of HIV/AIDS patients.
4. Patient satisfaction by medical service in a scale of 0 to 4.

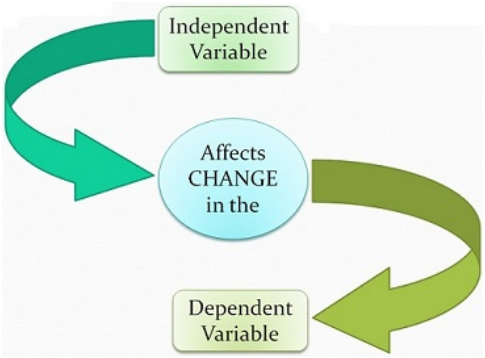
5. Length of stay in a hospital for medical treatment.
6. Type of unit stayed most when admitted in a hospital (coronary care, intensive care, maternity care, medical unit, pediatric unit, surgical unit).
7. Comfortability of a hospital's location (very comfortable, comfortable, somewhat comfortable, not at all comfortable).
8. Condition of a patient when admitted in a hospital (critical, serious, moderate, minor).
9. Skill of a doctor (excellent, very good, good, fair, poor).

If a characteristic has *only one value* in a particular study, it is *not a variable*; it is a constant. Thus, marital status is *not a variable if all participants are married.* Similarly, gender is *not a variable if all participants in a study are female.*

1.6.3 Role of Variables: Dependent vs Independent

Based on the role of variables in a statistical analysis, variables can be classified as *dependent* and *independent* variables.

- A *dependent* variable is a variable, that is, of primary interest to be determined as an outcome. For example, the outcome of a certain treatment or the educational achievement level can be considered dependent variables. The terms *outcome*, *response* and *dependent* are used interchangeably.
- An *independent* variable is a variable to be used to determine the value of the dependent variable. It is also called a *factor*, an *exposure*, a *predictor* or a *covariate.*



There are two types of independent variables: *attribute* (*measured*) and *active* (*manipulated*) variables.

  - An *attribute* independent variable is a variable whose values are *preexisting characteristics* of objects under study. The values of such a variable cannot be systematically changed or manipulated. For example, education, sex, socio-economic status, $\cdots$.
  - An *active* independent variable can be experimentally manipulated. Such an independent variable is a necessary (but not sufficient) condition to make *cause-and-effect* conclusions. For example, a researcher might investigate a new kind of

therapy compared to the traditional treatment (the treatment group each person is assigned to). A second example could be a design to evaluate the effect of different fertilizers on crop yields. A third example might be to study the effect of a new teaching method, such as cooperative learning, on student performance. Studies with active independent variables are experimental studies.

Even though a statistical analysis does not differentiate whether an independent variable is an attribute or active, there is a crucial difference in interpretation. For scientific researches in applied disciplines, the need to demonstrate that a given intervention or treatment causes change in behaviour or performance is extremely important. Only the approaches that have an active independent variable can allow one to infer that the change (difference) in the independent variable caused the change (difference) in the dependent variable. In contrast, a significant difference between or among persons with different values of an attribute independent variable should not lead one to conclude that the attribute independent variable caused the dependent variable to change.

Based on the type and role of variables, the common statistical methods are listed in the following table.

Dependent Variable	Independent Variable	Method
Continuous	Binary	t test
Continuous	Multinomial	ANOVA
Continuous	Continuous	Correlation
Continuous	Quantitative/Categorical/Both	Linear Regression
Categorical	Categorical	χ^2 test
Binary	Quantitative/Categorical/Both	Binary Logistic Regression
Multinomial	Quantitative/Categorical/Both	Multinomial Logistic Regression
Ordinal	Quantitative/Categorical/Both	Ordinal Logistic Regression
Discrete	Quantitative/Categorical/Both	Poisson Regression
Time-to-event	Quantitative/Categorical/Both	Survival Models

Note: For correlation and χ^2 test, there is no need to differentiate variables as dependent and independent.

Exercise 1.3. For each of the following objectives of a study, identify the dependent and independent variables, and the most appropriate model.

1. The effect of gender on blood pressure.
2. The effect of age on hypertension.
3. The effect of gender on hypertension.
4. Comparing the number of sexual partners between urban and rural resident women.
5. Investigating the effectiveness of a new therapy compared to the traditional treatment.
6. Evaluating the effect of different fertilizers on crop yields.
7. Comparing the blood pressure of male and female individuals.

8. Studying the effect of a new teaching method, such as cooperative learning, on student performance.
9. The effect of age (classified as $18-24$, $25-34$ and ≥ 35 years) on hypertension.

1.7 Types and Sources of Data

Data is presented in a rectangular form. All the values of a particular variable is organized to a column. Observation, i.e., measurements collected from a subject forms a row in a dataset.

1.7.1 Types of Data

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

1. **Cross-sectional data**: Cross-sectional data consists of a set of observations taken from different subjects at a single point in time.

Student	Gender	Age (in years)	GPA
1	Male	23	3.84
2	Female	27	4.00
3	Male	26	2.18
4	Female	25	3.25
5	Female	21	3.57
6	Male	24	3.01
7	Female	25	2.75

2. **Time-series data**: is a set of observations collected from a single subject at different times usually at equal intervals, such as daily, weekly, monthly, quarterly, annually, etc.

Year	2007	2008	2009	2010	2011
Number of RTA	5214	4845	8174	1052	9784

3. **Longitudinal data**: It is usually called as *cross-sectional time-series* data as it involves a collection of observations from different subjects at multiple instances.

StudentID	Female	Semester	GPA
251	0	1	3.51
251	0	2	3.25
251	0	3	3.63
251	0	4	3.70
251	0	5	3.65
251	0	6	3.20
257	1	1	3.67
257	1	2	3.90
257	1	3	3.78
257	1	4	3.50
257	1	5	3.82
257	1	6	3.90

Case-control (retrospective) and cohort (prospective) studies use longitudinal data.

1.7.2 Sources of Data

Depending upon the sources, a statistical data can be classified into two categories: *primary* and *secondary*.

1. **Primary data**: Primary data is the data collected, under the direct supervision and instruction of the researcher, either through direct personal observation or by enquiring individuals. The personal observation ranges from simple visual observations to those requiring special skills like clinical and microbiological examinations using radiographic, biochemical, X-ray machines, microscope, $\cdots$. And for obtaining data by enquiring individuals, a specially designed form called *questionnaire*[1] will be prepared prior to the actual data collection.

2. **Secondary data**: When an investigator uses data which has already been collected by others, the data is called *secondary* data. Such data could be obtained from different organizations like health institutions (health centers, hospitals), research organizations, Central Statistical Agency, Ministry of Health and other ministerial offices. This data is primary for the agency that collected it and becomes secondary for someone else who uses this data for his/her own purposes.

Before using secondary data for analysis, the data should be checked for whether it is *suitable*, *adequate* and *reliable* for the purpose of investigation. Whether the data are suitable can be judged in the light of the nature and scope of investigation. Whether the data are adequate can be judged in the light of the time and geographical area covered. Whether the data are reliable is checking for its accuracy, for example, whether the sample is a proper representative of the population.

1.8 Methods of Data Collection

The first and foremost task in statistical investigation is data collection. In fact, prior to the actual data collection, there are four important points to be considered. These are the *purpose* of data collection (why we need to collect data), the *type of variables* to be considered (what are the variables of interest), the *source of data* (from where we can get the data) and the *methods of data collection* (how we can collect the data).

Once these questions are answered, it becomes necessary to collect the data needed. There are two broad ways of collecting data. The first is *experimentation*, i.e., an actual experiment is conducted and then observations (measurements and counts) are recorded. Such experimental studies are common in natural sciences; agriculture, biology, medical science, industry, $\cdots$.

The other way of collecting data is by enquiring certain individuals, directly or indirectly. Such a technique is known as *survey* which is commonly used in social sciences, i.e., problems related to sociology, psychology and various economic studies.

For a survey, there are two common methods of data collection: face-to-face interview and self-administered questionnaire.

[1] see Section 1.8

1. **Face-to-face interview**: A trained interviewer (commonly called *enumerator*) asks a series of questions to the interviewee (commonly called *respondent*) and records responses on a specially designed form called *questionnaire*. In this approach, the quality of the data is affected both by the design of the questionnaire and the quality of the interviewer. This method of data collection has the advantage of obtaining correct information, since the enumerator can make some clarifications to the questions which are not clear for the respondent, and avoids incomplete responses.

   On the other hand, face-to-face interview is costly since it requires training of interviewers and other costs. In addition, the respondent may not tell us the real information for sensitive questions, since there is face-to-face interaction. For example, if the respondent's salary is small, then s/he may get ashamed of it and might not tell the correct one.

2. **Self-administered questionnaire**: The researcher distributes a copy of a single questionnaire to the respondents; the respondents complete the questionnaire and give back to the researcher.

   As compared to face-to-face interview, costs are low. Also, the responses are free from biases of the interviewer. On the other hand, the respondent might give inappropriate answers to questions, since there is no one with them, they may understand the question wrongly and respond it incorrectly. This method is not applicable for uneducated respondents. There is also a high degree of non- or partial-response, and low return rates.

Designing Questionnaire

In most survey methods of data collection, it is necessary to prepare a document, called *questionnaire*, which contains a set of questions to be answered by the respondents and used to record the responses. In addition to the set of related questions, a questionnaire should have also a *cover letter*.

The cover letter should introduce the person conducting the survey (researcher), state the *objective(s)* of the survey, put a *promise of the anonymity* (*confidentiality*), estimated *time* to complete the questionnaire, a *consent* information whether the respondent is willing to participate, and also include general *instructions* how to fill the questionnaire.

In designing a questionnaire, first, the contents (topics) needed to meet the objectives of the study should be described. Next, the set of questions under each content should be formulated. Lastly, the layout of the questionnaire should be formatted appropriately.

In preparing questions for a questionnaire, the following points should be kept in mind:

- The number of questions should be as few as possible. Once the objectives of the survey are clearly defined only questions pertinent to the objectives should be included. That is, *all questions in a questionnaire must have a relevance to the objectives* of the survey. If a lengthy questionnaire is unavoidable, it should preferably be divided into two or more parts depending on the contents.

- Questions should be *simple, short and easy to understand* and each questionnaire should convey one and only one idea. Technical terms should be avoided.
- Questions should be capable of *objective answers* and the answers *should not* require any calculation.
- *Sensitive questions* (questions of personal and financial nature) should be avoided if possible. Otherwise, such questions should be asked indirectly, by constructing a set of ranges, say, age in years ($\leq$17, 18-24, 25-34, 35-64, $\geq$65), salary in Birr ($\leq$1000, 1001-2000, 2001-3000, 3001-4000, $\geq$4001). For non-numeric sensitive questions like diseases with stigma should be posed as late as possible in the questionnaire.
- *Leading questions*, question that prompts the answer wanted, should be completely avoided. The fashion in which a question is asked may result in a response that may be biased in the direction in which the question is slanted. If we ask a person like "You do not smoke cigarette?" the person will automatically say 'Yes I do not smoke'.
- Questions should be *logically arranged* under the appropriate sequence of topics so that a natural and spontaneous reply is introduced. Topics should not be mixed up and questions should not skip back and forth. For example, it is undesirable to ask a person how many children s/he has before asking whether s/he is married or not. In general, questions related to background (identification and description of respondents like age, religion, education, marital status, occupation, $\cdots$) variables should be come first, followed by major information questions. If opinions are needed, such questions should usually be placed at the end of the list.

Chapter 2

Methods of Data Presentation

2.1 Tabulation of Data

A table is a systematic arrangement of data in rows and columns, which simplifies complex data and facilitates comparison. Tabulation should not be confused with classification, as the two differ in many ways. Classification is the separation of items according to similar characteristics and grouping them into various groups. The classification may be geographically (according to location differences): places, areas or regions; chronologically (according to the time period): weekly, monthly, quarterly, annually, $\cdots$; according to qualitative characteristics such as religion, sex, marital status, $\cdots$; or quantitative characteristics such as weight, height, income, $\cdots$. 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 table, in general, should have the following parts.

1. **Table Number**: Every table should be identified by a number. It facilitates easy referencing and identification. The number should be written in the center at the top of the table. Whenever referring to the table in the text, only the number of the table can be given.

2. **Title**: There should be a title at the top, which describes the content of the table. The title should be clear, concise and brief. The title should be also self-explanatory, meaning it should answer the questions: What kind of data? When the data is collected and from where it is collected?

3. **Caption**: Caption is a brief and self-explanatory heading for each column given the unit of measurement in parentheses.

4. **Stub**: Stub is a brief and self-explanatory heading for each row given the unit of measurement in parentheses.

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 in different cells. This arrangement of data remains according to the description of captions and stubs. 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.

6. **Footnote**: If any value 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 value should be explained at the beneath of the table with the same mark as a 'footnote'. This is a qualifying statement which is to be written below the table explaining certain points related to the data which have not been covered in title, caption, and stubs.

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 obtained at the end of the table. Reference of the source must be complete so that if the potential reader wants to consult the original source they may do so.

2.2 Frequency Distribution

The most convenient way of organizing data is constructing a frequency distribution. Frequency distribution is the organization of raw data in table form, using *classes* and *frequencies*. The term *class* stands for a description of a group of similar objects in a dataset and *frequency* is the number of times a variable value (category) is repeated.

There are three types of frequency distributions; *categorical*, *discrete* (*ungrouped*) and *grouped* (*continuous*) frequency distributions.

2.2.1 Categorical Frequency Distribution

Categorical frequency distribution is used when the variable 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.

Example 2.1. The blood type of a sample of 25 students is: A B B AB O AB O O B B B A B B AB O A O AB A O O O AB O. Construct categorical frequency distribution.

Solution: The variable of interest is blood type. This variable has four categories; A, B, AB and O. As a result, the frequency distribution will have four classes.

Class (Blood type)	Frequency (Number of students)
A	4
B	7
AB	5
O	9
Total	$n = 25$

This frequency distribution shows the actual number of observation (frequency) of each class and it is called *absolute* frequency distribution. However, an absolute frequency (count) is rarely useful.

Consider a study of a sample of 150 patients (99 female and 51 male) in which 49 female and 25 male are depressed. Then, it seems that more female patients are depressed than male patients. However, it turns out that there are simply more females in the study sample than males. It is only when this count is expressed as a *proportion* that it becomes useful.

Looking at the proportion of female patients who are depressed compared to the proportion of male patients who are depressed, it is found the proportions are almost equal ($\frac{49}{99} = 0.495$ and $\frac{25}{51} = 0.490$) and so females are not more likely to be depressed than males. Hence, in constructing a frequency distribution, it is essential to express the counts (the number of observations) in each class as *proportions* of the total sample size.

Note: *Proportion* is a special case of a *ratio*.

- A *ratio* is simply one number divided by another. Given the weight of a person (in kg) and the height (in metres), then the ratio of weight to height^2 is the *Body Mass Index* (BMI).
- *Proportion* is a ratio of counts where the numerator is a subset of the denominator. Of 40 patients, if 10 are cured, then the proportion is $\frac{10}{40} = 0.25$ (1:3) which indicates 25% of the patients are cured. {A *population proportion* is denoted by π while a *sample proportion* is denoted by p.} A proportion is known as a *risk* if the numerator counts bad events. Hence, if 300 students started nursing school and finally 15 failed to graduate, the risk of failing is $\frac{15}{300} = 0.05 = 5\%$.
- When a time period is involved in the denominator, then a proportion is known as a *rate*. If 500 persons die in one month, out of a population of 50000, the death *rate* is $\frac{500}{50000} = 0.01$ deaths per person per month. {A *population rate* is denoted by λ while a *sample rate* is denoted by r.}

Relative Frequency Distribution

A *relative* frequency distribution displays the proportion of observations in each group. The relative frequency is $\frac{f_i}{n}$ where f_i is frequency of the i^{th} class and n is the total number of observations. It can be converted into a percentage frequency distribution by multiplying the relative frequency by 100.

Considering example 2.1, the proportion and percentages frequencies are presented in the following frequency distribution.

Blood type	Number of students	Proportion	Percentage
A	4	0.16	16
B	7	0.28	28
AB	5	0.20	20
O	9	0.36	36
Total	25	1.00	100

The table shows, of the total number of 25 students, 4 (16%) of the students have blood type A, 7 (28%) of the students have blood type B, 5 (20%) of the students have blood type AB and the remaining 9 (36%) of students have blood type O.
Notes:

- The relative (percentage) frequencies are particularly helpful when comparing two or more distributions in which the number of observations are quite different. The percentage distributions make such a comparison more meaningful, since the total number in the sample or population under consideration becomes irrelevant.

- Consideration of the relative or percentage frequency is preparatory to the study of probability theory in Chapter 5. Indeed, if the students in the above table were selected randomly from, say, a certain school, the probability of selecting a student whose blood type is A would be 0.16 (16%), the probability of selecting a student whose blood type is B would be 0.28 (28%), the probability of selecting a student whose blood type is AB would be 0.20 (20%), and the probability of selecting a student whose blood type is O would be 0.36 (36%).

Example 2.2. Patients entering ICU in a given hospital may be classified as medical, surgical, cardiac and other ICT types. Thus, the distribution of 25 patients in the different ICUs can be presented as follows.

ICU Type	Frequency	Relative Frequency
Medical	12	0.48
Surgical	6	0.24
Cardiac	5	0.20
Other	2	0.08
Total	25	1.00

Exercise 2.1. Construct an absolute and a percentage frequency distributions for the following letter grades of 32 students in statistics course: A B C C C C B B A D A C C A B F C C A B A B C C C A D A C C A B.

2.2.2 Discrete Frequency Distribution

A frequency distribution is also used to summarize and present quantitative data. If the variable of interest is *discrete* that can take only a *few unique values*, then the frequency distribution can be constructed in the same way as that of for a qualitative variable. Such a frequency table is called *discrete* (*ungrouped*) frequency distribution. In a discrete frequency distribution, each numerical value of the quantitative variable represents a single class and the counts of each value represents the frequency of that class.

Example 2.3. The number of children for a sample of 21 families is given as: 2 3 5 4 3 3 2 3 1 0 4 3 2 2 1 1 1 4 2 2 2. Construct discrete frequency distribution.

Solution: The variable of interest here is number of children with 6 unique values; 0, 1, 2, 3, 4 and 5.

Class (No. of children)	Frequency (No. of families)	Percentage (%)
0	1	4.76
1	4	19.04
2	7	33.33
3	5	23.81
4	3	14.29
5	1	4.76
Total	21	100

Only 1 (4.76%) family has no child, 4 (19.04%) of the families have only 1 child, 7 (33.33%) of the families have 2 children, 5 (23.81%) of the families have 3 children, 3 (14.29%) of the families have 4 children and the remaining 5 (23.81%) of the families have 5 children.

Exercise 2.2. Suppose that a sample of 16 children from a primary school are taken and get the following data about the number of their decayed teeth as 3, 4, 2, 4, 0, 1, 3, 0, 2, 3, 2, 3, 3, 2, 4, 1. Construct a frequency distribution.

Cross-tabulation

In addition to tabulating each variable separately (one-way table), there may be an interest to cross-classify individuals based on two variables. Each of the two variables can be either qualitative or discrete assuming a few unique values. Such a classification using two variables is called a *two-way (contingency) table*. The primary value of a cross-tabulation lies in the insight it offers about the association between the variables.

Recall the example in which 49 female and 25 male patients are depressed of 150 (99 female and 51 male) patients. Here, there are two variables of interest; gender and depression. Each variable has two categories. Therefore, the data can be summarized in a 2×2 contingency table as follows.

Gender	Depression		Total
	Depressed	Not Depressed	
Male	25	26	51
Female	49	50	99
Total	74	76	150

This table shows, of the total 150 patients, 25 patients are male and depressed, 49 patients are female and depressed, 26 patients are male and not depressed, and 50 patients are female and not depressed.

The frequency distributions constructed from the margins of a cross-tabulation provide information about each of the variables individually, but these do not shed any light on the relationship between the variables. Converting the frequencies in a cross-tabulation into row percentages or column percentages can provide more insight into the relationship between the two variables.

Exercise 2.3. The age in years and hypertension test results of a sample of 9 women in the reproductive age, 15-49, is given in the following table:

Age	$15-24$	$25-34$	$35-49$	$35-49$	$25-34$	$15-24$	$35-49$	$35-49$	$25-34$
Hyp	+ve	-ve	+ve	+ve	+ve	-ve	+ve	-ve	+ve

Summarize the data using a contingency table.

2.2.3 Grouped Frequency Distribution

Like a discrete frequency distribution, *grouped* (*continuous*) frequency distribution is used for quantitative variable data. But, the frequency distribution of the individual values is not interesting for a *continuous variable* or *discrete variable with a lot of unique values.*

For example, consider the age distribution of 25 individuals. If a discrete frequency distribution is constructed, the table will look like:

Age	Frequency
15	1
19	2
29	1
31	1
34	1
39	1
52	2
53	2
45	2
65	1
71	1
74	1
75	2
76	2
77	1
78	1
79	1
84	1
85	1
Total	25

which has too many rows and is too complex to understand. Instead, it is better to group the different values of age into non-overlapping groups.

Hence, when the quantitative variable of interest has *a lot of different values*, a grouped frequency distribution is constructed. Unlike a discrete distribution, several values of a variable are grouped into one class in a grouped frequency distribution and the number of observations belonging to each class is the frequency of that class.

Example 2.4. Consider the following age distribution of a sample of 70 persons infected with malaria:

Class Limits (Age in years)	Class Boundaries (Age in years)	Frequency (No. of persons)	Percentage (%)
1 - 25	0.5 - 25.5	20	28.57
26 - 50	25.5 - 50.5	15	21.43
51 - 75	50.5 - 75.5	25	35.71
76 - 100	75.5 - 100.5	10	14.29
Total		70	100.00

From this frequency distribution, of the total 70 persons who have malaria, 20 (28.57%) of them are between 1-25 years, 15 (21.43%) of them are in between 26-50 years, 25 (35.71%) of them are in between 51-75 years and 10 (14.29%) of the persons are in between 76-100 years. It is also easy to observe that most (35.71%) of the persons are between 51-75 years of age.

Basic Terms

1. **Class Limits**: The smallest and highest values that can be included in a particular class are called *class limits*. The smallest values are called *lower* class limits (LCL) and

the highest values are called *upper* class limits (UCL). For example, class limit of the first class is 1 - 25, where 1 is the *lower* class limit and 25 is the *upper* class limit of the first class.

2. **Class Boundaries**: Class boundaries are also similar to class limits. The smallest values are called *lower* class boundaries (LCB) and the highest values are called *upper* class boundaries (UCB). For example, class boundary of the first class is 0.5 - 25.5 where the *lower* class boundary is 0.5 and the *upper* class boundary is 25.5. Note that the UCB of one class is also the LCB of the next class.

3. **Class Width** (w): It is the difference between UCB and LCB of a certain class. That is, $w = \text{UCB} - \text{LCB}$. Class width is also the difference between the lower limits (boundaries) of two consecutive classes ($w = \text{LCL}_i - \text{LCL}_{i-1}$) or it is the difference between upper limits (boundaries) of two consecutive classes ($w = \text{UCL}_i - \text{UCL}_{i-1}$). The class width of the above frequency distribution is $w = \text{UCB}_1 - \text{LCB}_1 = 25.5 - 0.5 = 25$ or $w = \text{LCL}_2 - \text{LCL}_1 = 26 - 1 = 25$ or $w = \text{UCL}_2 - \text{UCL}_1 = 50 - 25 = 25$.

4. **Class Mark**: Class mark is the mid-point of each class which can approximate all the values in that class. Or it is the half-way between the class limits (boundaries) of a certain class. That is, $CM_i = \frac{\text{LCL}_i + \text{UCL}_i}{2} = \frac{\text{LCB}_i + \text{UCB}_i}{2}$.

   For example, class marks of the above distribution are $\text{CM}_1 = 13$, $\text{CM}_2 = 38$, $\text{CM}_3 = 63$ and $\text{CM}_4 = 88$. Note also that $w = \text{CM}_i - \text{CM}_{i-1}$.

   Class mark is important, because all the observations in a particular class are approximated by its class mark in most statistical analyses.

Cumulative Frequency Distribution

A cumulative frequency distribution displays the total number of observations above (below) a certain value. When the interest focuses on the number of observations below a certain value, then the reference value should be an upper class boundary. It is known as *less than* cumulative frequency distribution. Similarly, when the interest lies in finding the number of observations above a certain value, then the reference values should be lower class boundaries and the corresponding frequency distribution is known as *more than* cumulative frequency distribution.

Less than cumulative frequency	
Age in years	F
< 25.5	20
< 50.5	20+15=35
< 75.5	20+15+25=60
<100.5	20+15+25+10=70

More than cumulative frequency	
Age in years	F
> 0.5	10+25+15+20=70
>25.5	10+25+15=50
>50.5	10+25=35
>75.5	10

Steps in Constructing Grouped FD

1. The first step in constructing a frequency distribution is to arrange the data in an array form (preferably increasing order).

2. The second step is to find the unit of measurement (u). The unit of measurement is the smallest numerical difference between any *two distinct values* of a dataset.

3. The third step is to determine the range (R). Range is the maximum numerical difference in the dataset, i.e. the difference between the largest and the smallest values.

4. The fourth step in constructing a frequency distribution is to determine how many classes it will contain. The number of classes (k) can be determined using Sturge's Rule, $k = 1 + 3.322 \times \log n$ where n is the total number of observations in the sample or population. Normally, the number is rounded up to the next whole number without considering the mathematical rounding rule.

5. After determining the number of classes, the researcher should determine the class width (w) of the frequency distribution. An approximation of the class width can be calculated by dividing the range by the number of classes, that is, $w = \frac{R}{k}$. This will be the same for all the classes.

6. The sixth step is to obtain the class limits. To do so, the frequency distribution should start at a value equal to the lowest number of the raw data.

   (a) For obtaining the class limits, the smallest value in the dataset can be taken as the LCL of the first class. If S is the smallest observation, then $\text{LCL}_1 = S$.

   (b) Then, the LCL of the second class is obtained by adding the class width w to the LCL of the first class. That is, $\text{LCL}_2 = S + w$. Similarly, $\text{LCL}_3 = \text{LCL}_2 + w$. Continue adding w until k classes are obtained: $\text{LCL}_i = \text{LCL}_{i-1} + w$ for $i = 2, 3, \cdots, k$.

   (c) Next, the UCLs of the frequency distribution are obtained by adding $w - u$ to the corresponding LCLs: $\text{UCL}_i = \text{LCL}_i + (w - u)$ for $i = 1, 2, \cdots, k$.

7. The last step is to generate the class boundaries. The LCBs are obtained by subtracting $\frac{1}{2}u$ from the corresponding LCLs and the UCB are obtained by adding $\frac{1}{2}u$ to the corresponding UCLs of the classes. That is, $\text{LCB}_i = \text{LCL}_i - \frac{1}{2}u$ and $\text{UCB}_i = \text{UCL}_i + \frac{1}{2}u$ for $i = 1, 2, \cdots, k$.

Example 2.5. The number of hours of 56 emergency patients spent in emergency room at a given hospital is given as: 31 33 33 34 34 35 35 17 31 36 17 18 19 25 26 27 27 19 20 22 31 36 38 13 22 22 35 36 28 28 29 30 30 36 11 13 16 17 17 22 22 23 23 23 23 24 24 24 25 27 27 28 28 30 13 16. Construct a grouped frequency distribution for this data.

Solution:

1. Increasing order: 11 13 13 13 16 16 17 17 17 17 18 19 19 20 22 . . . 36 38.

2. The unit of measurement is $u = 17 - 16 = 1$.

3. The range of the data is $R = L - S = 38 - 11 = 27$.

4. The number of classes is $k = 1 + 3.322 \log N = 1 + 3.322 \log 56 = 6.81 \approx 7$.

5. The class width is $w = \frac{R}{k} = \frac{27}{6.81} = 3.96 \approx 4$.

6. Class limits:
   (a) The smallest value is 11. Thus, the $\text{LCL}_1 = 11$.
   (b) Then, the LCLs of 2^{nd}, 3^{rd}, 4^{th}, $\cdots$, 7^{th} classes are 15, 19, 23, $\cdots$, 35, respectively.
   (c) The difference $w - u = 4 - 1 = 3$. Hence, the UCLs of the classes are 14, 18, 22, $\cdots$, 38 which are obtained by adding 3 to the corresponding LCLs.
7. $\frac{1}{2}u = 0.5$. Class boundaries: $\text{LCB}_i = \text{LCL}_i - 0.5$ and $\text{UCB}_i = \text{UCL}_i + 0.5$ for $i = 1, 2, \cdots, 7$.

Class	Class Limit	Class Boundary	Class Mark	Frequency (f)	Percentage (%)	LCF	MCF
1	11 - 14	10.5 - 14.5	12.5	4	7.14	4	56
2	15 - 18	14.5 - 18.5	16.5	7	12.50	11	52
3	19 - 22	18.5 - 22.5	20.5	8	14.29	19	45
4	23 - 26	22.5 - 26.5	24.5	10	17.86	29	37
5	27 - 30	26.5 - 30.5	28.5	12	21.43	41	27
6	31 - 34	30.5 - 34.5	32.5	7	12.50	48	15
7	35 - 38	34.5 - 38.5	36.5	8	14.28	56	8
Total				56	100		

Example 2.6. The following data is the body mass index of a sample of 70 adults in a certain place. Construct the grouped frequency distribution for the data.
25.4 26.6 27.5 18.1 21.9 23.0 24.3 28.8 30.9 34.8 19.2 21.9 23.1 24.3 25.6 26.9 27.5 28.8 30.9 34.9 19.8 21.9 23.1 24.5 25.7 27.1 27.6 28.9 31.0 35.0 20.2 22.3 23.3 24.6 25.7 27.3 28.2 29.3 31.1 35.5 20.7 22.3 23.4 24.6 25.8 27.3 28.3 29.5 31.3 35.8 20.8 22.3 23.5 24.7 25.8 27.3 28.3 29.8 31.6 35.9 21.1 22.4 24.0 24.7 25.9 27.3 28.3 30.0 31.6 36.6

Solution:

1. The data in an increasing order:
   18.1 19.2 19.8 20.2 20.7 20.8 21.1 21.9 21.9 21.9 22.3 22.3 22.3 22.4 23.0 23.1 23.1 23.3 23.4 23.5 24.0 24.3 24.3 24.5 24.6 24.6 24.7 24.7 25.4 25.6 25.7 25.7 25.8 25.8 25.9 26.6 26.9 27.1 27.3 27.3 27.3 27.3 27.5 27.5 27.6 28.2 28.3 28.3 28.3 28.8 28.8 28.9 29.3 29.5 29.8 30.0 30.9 30.9 31.0 31.1 31.3 31.6 31.6 34.8 34.9 35.0 35.5 35.8 35.9 36.6
2. The unit of measurement is $u = 23.5 - 23.4 = 0.1$.
3. The range of the data is $R = L - S = 36.6 - 18.1 = 18.5$.
4. The number of classes is $k = 1 + 3.322 \times \log n = 1 + 3.322 \log 70 = 7.13 \approx 8$.
5. The class width is $w = \frac{R}{k} = \frac{18.5}{7.13} = 2.594 \approx 2.6$.
6. Class limits:
   (a) The smallest value is 18.1. Thus, the $\text{LCL}_1 = 18.1$.
   (b) Then, the LCLs of 2^{nd}, $\cdots$, 8^{th} classes are 20.7, $\cdots$, 36.3, respectively.
   (c) As $w - u = 2.6 - 0.1 = 2.5$, the UCLs are 20.6, 23.2, 25.8, $\cdots$, 38.8.

7. $\frac{1}{2}u = 0.05$. Class boundaries: $\text{LCB}_i = \text{LCL}_i - 0.05$ and $\text{UCB}_i = \text{UCL}_i + 0.05$.

Class	Class Limit	Class Boundary	Class Mark	Frequency (f)	Percentage (%)	LCF	MCF
1	18.1 - 20.6	18.05 - 20.65	19.35	4	5.71	4	70
2	20.7 - 23.2	20.65 - 23.25	21.95	13	18.57	17	66
3	23.3 - 25.8	23.25 - 25.85	24.55	17	24.29	34	53
4	25.9 - 28.4	25.85 - 28.45	27.15	15	21.43	49	36
5	28.5 - 31.0	28.45 - 31.05	29.75	10	14.29	59	21
6	31.1 - 33.6	31.05 - 33.65	32.35	4	5.71	63	11
7	33.7 - 36.2	33.65 - 36.25	34.95	6	8.57	69	7
8	36.3 - 38.8	36.25 - 38.85	37.55	1	1.43	70	1
Total				70	100.00		

Exercise 2.4. The birth weights (in kilograms) of 30 infants were recorded as follows. Construct grouped frequency distribution for this data.

2.0 2.1 2.3 3.0 3.1 2.7 2.8 3.5 3.1 3.7 4.0 2.3 3.5 4.2 3.7
3.2 2.7 2.5 2.7 3.8 3.1 3.0 2.6 2.8 2.9 3.5 4.1 3.9 2.8 2.2

Summary

In constructing a grouped frequency distribution, the following general points should be kept in mind.

- The number of classes should be neither too large nor too small. Fewer classes would mean greater class width with consequent loss of accuracy. Too many classes result in greater complexity because it does not aggregate the data enough to be helpful.
- Classes should be complete (it should include all the values) and non-overlapping (no value should belong to two classes).
- Classes should be standardized (arranged in a logical and chronological (increasing) order) and continuous (a class must be included in the frequency distribution even if there are no values in that class).
- Open ended classes, where there is no lower limit of the first class or no upper limit of the last class, should be avoided since this creates difficulty in analysis and interpretation.

Grouped frequency distribution has the advantage of reducing a large mass of data into a comparatively small table, that is, it makes summarization easy. It also helps for further statistical analysis like central tendency, variation, skewness, kurtosis, $\cdots$. On the other hand, the identity of the observations is lost. That is, only the number of observations in a class is known but it is unknown what the values are in a class. Hence, the original data cannot be reconstructed from a grouped frequency distribution.

Exercise 2.5. Given the following frequency distribution of pathologic tumor size (in cm) for a sample of 110 cancer patients:

Tumor size in cm	Frequency
0.25 - 0.75	13
0.75 - 1.25	36
1.25 - 1.75	17
1.75 - 2.25	18
2.25 - 2.75	15
2.75 - 3.25	11
Total	110

1. What is the variable of interest?
2. What percent of patients is with an approximate level of pathologic tumor size of 2cm?
3. What number of cancer patients is with lowest pathologic tumor size?
4. What is the approximate size of pathologic tumor with highest percentage of patients?
5. What is the percentage of cancer patients with pathologic tumor size less than 0.25cm?
6. How many cancer patients have pathologic tumor size above 3.25cm?
7. How many patients have tumor size more than 1.25cm?
8. What percentage of the measurements are between 1.25cm and 2.25cm?
9. What proportion of the measurement is less than or equal 2.75cm?

2.3 Charts and Graphs

The second way and the most effective mechanisms for presenting data in a form meaningful to decision makers is through the use of charts and graphs. Charts and graphs have greater attraction and memorizing effect than mere figures. They also facilitate comparison, and are used to understand patterns and trends.

Charts and graphs give quick impression of the data. Hence, through charts and graphs, the decision maker can often get an overall picture of the data and reach some useful conclusions.

2.3.1 Pie Chart

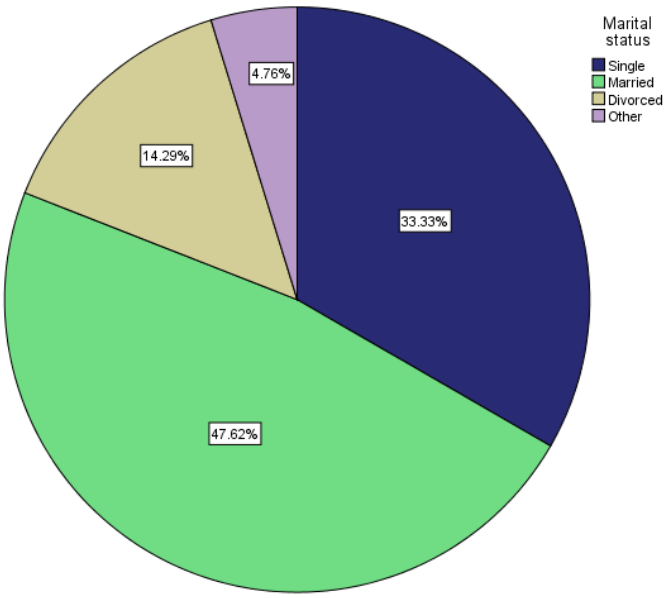
Pie chart is a circle used to display one-way frequency distributions (categorical or discrete) and also quantitative data given in different categories (values) of a qualitative or discrete variable. The total area of the circle represents 100% of the sum of the frequencies (magnitudes) of all the categories (values) of the variable. The circle is divided into a number of slices where the size of each slice corresponds to the percentage frequency (magnitude) corresponding to each category (value) of the variable. The slices are marked differently (different lines or dots) or by different colors to identify each category of the variable of interest.

Example 2.7. Construct pie chart for the following frequency distribution.

Marital status	Number of individuals
Single	70
Married	100
Divorced	30
Other	10
Total	210

Solution: Construction of the pie chart begins by determining the proportion of each category to the whole, and then the degree coverage of each category frequency should be calculated. Hence, to obtain the angle for any category, the relative frequency is multiplied by 360 degrees ($\frac{f_i}{n} \times 360$).

Marital status	Degree
Single	$(\ 70/210) \times 360 = 119.99$
Married	$(100/210) \times 360 = 171.43$
Divorced	$(\ 30/210) \times 360 = \ 51.44$
Other	$(\ 10/210) \times 360 = \ 17.14$
Total	360



Exercise 2.6. Construct a pie chart using the data from example 2.1 and example 2.3.

2.3.2 Bar Charts

A bar chart is another widely used chart for presenting categorical or discrete frequency distributions (both one-way and two-way frequency tables) and also quantitative data given in different categories (values) of a qualitative or discrete variable.

The categories or values of the variable are marked along the $x-$axis and the frequencies (proportions, percentages) corresponding to the categories of that variable are marked along the $y-$axis. As a result, the heights of the bars represent the frequencies corresponding to

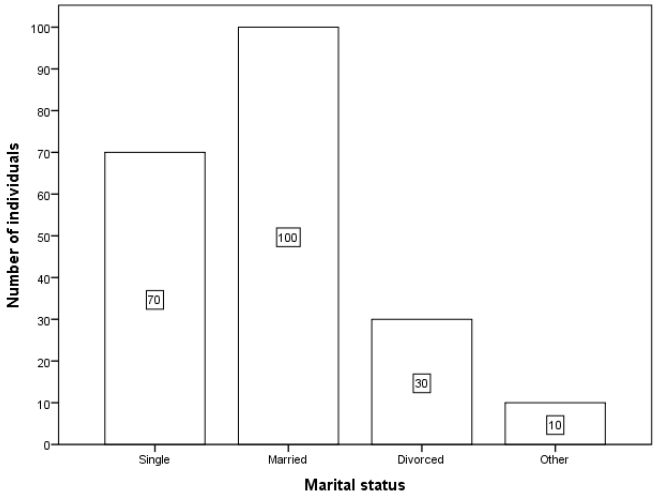
the classes. But, the width of the bars has no meaning; however, all the bars should have the same width to avoid confusion. In addition, the bars are separated by a constant distance so as not to imply continuity.

There are three types of bar charts: simple, component (stacked) and multiple (clustered) bar charts.

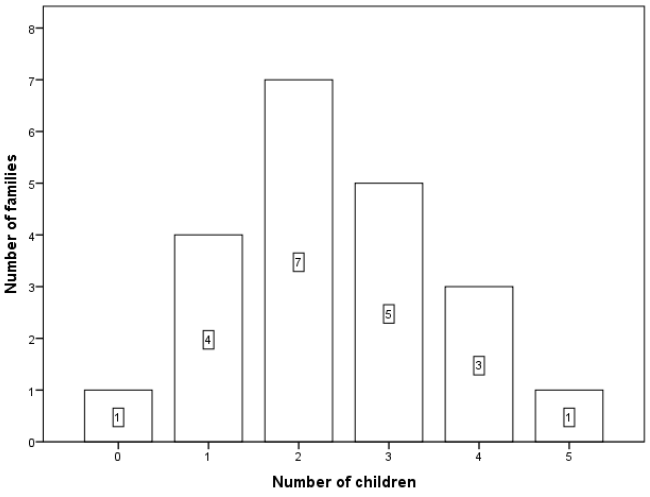
Simple Bar Chart

The first use of a simple bar-chart is for presenting one-way (categorical or discrete) frequency distribution. It is constructed from the same type of frequency distribution (categorical or discrete) and data that is used to produce a pie chart.

Example 2.8. Construct a simple bar chart for the data given in example 2.7.



Example 2.9. Construct a simple bar chart for the number of children given in example 2.3.

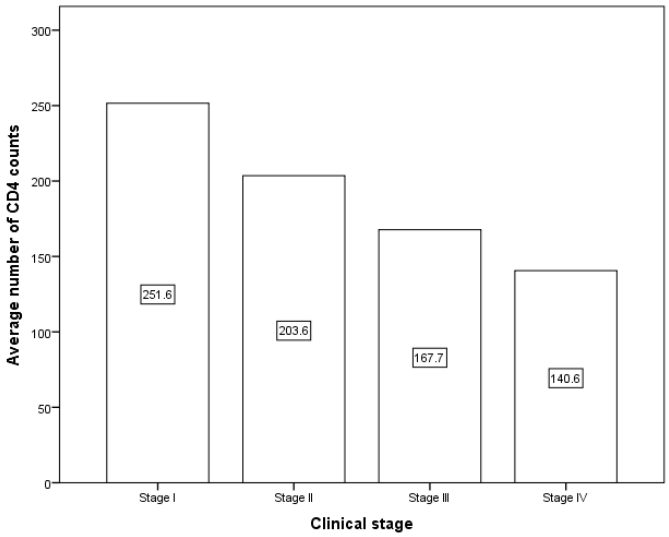


Exercise 2.7. Construct a simple bar chart using the data from example 2.1 and example 2.3.

In addition, like a pie chart, simple bar chart is also used to present quantitative data, that is, given along the different categories (values) of qualitative or discrete variable. The interest in constructing a simple bar chart is to explore a quantitative variable (usually, a dependent variable) given along the different categories (values) of a qualitative or discrete variable (usually, an independent variable). The categories (values) of the variable are marked along the $x-$axis and the magnitudes corresponding to the categories of that variable are marked along the $y-$axis.

Example 2.10. The following data shows 1464 HIV/AIDS patients classified by clinical stage when they started HAART treatment and also the average number of CD4 counts in each clinical stage.

Stage	Number of patients	Average CD4 counts
Stage I	347	251.6
Stage II	514	203.6
Stage III	496	167.7
Stage IV	107	140.6



This bar chart clearly shows that as the clinical stage (severity) of the HIV/AIDS patients increases, the average CD4 count decreases.

An advantage of using a bar chart over a pie chart for a given dataset is that for categories that are close in frequency or magnitude, it is considered easier to see the difference in the bars of a bar chart than discriminating between the slices of a pie chart.

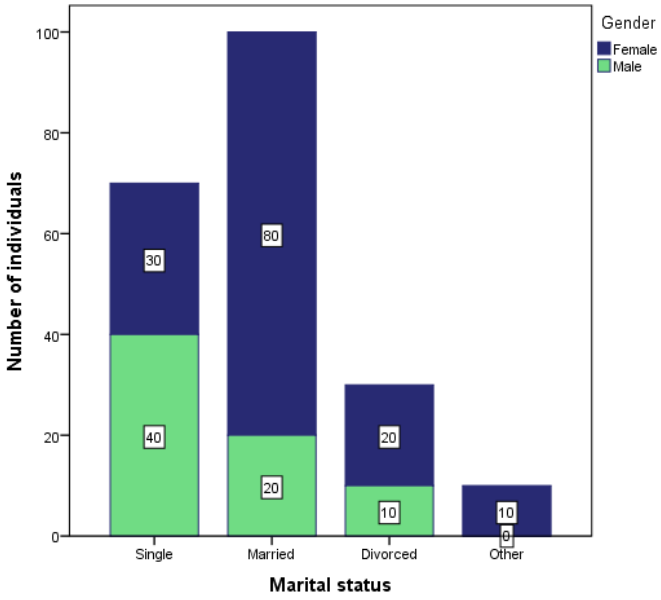
Component Bar Chart

Component bar chart is used to display two-way frequency distributions and it is also used to present the magnitude of a quantitative variable (usually, a dependent variable) given along a combinations of the different categories (values) of two qualitative or discrete variables (usually, independent variables). Here, there is a desire to show the frequency or magnitude of each category of the first variable divided into its component parts based on the categories

of the second variable. In such type of charts, each bar is subdivided into parts in proportion to the frequency or magnitude of each category of the second variable. The subdivided bars are shaded by different colors, lines or dots for identifying the components (categories of the second variable).

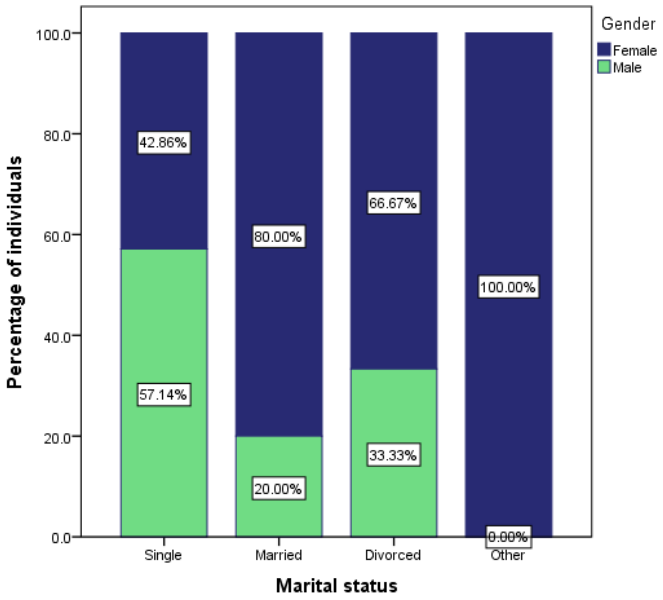
Example 2.11. Construct component bar chart for the following data.

Marital status	Male	Female	Total
Single	40	30	70
Married	20	80	100
Divorced	10	20	30
Other	0	10	10
Total	70	140	210



Sometimes, the total number of observations or magnitudes corresponding to the different categories of a variable may be greatly different. For making meaningful comparisons, the frequencies or magnitudes of the categories are converted to percentages. In that case, each category will have 100 as its maximum frequency. This sort of bar chart is known as *percentage component bar chart*.

Example 2.12. Construct percentage component bar chart for the above data, example 2.11.

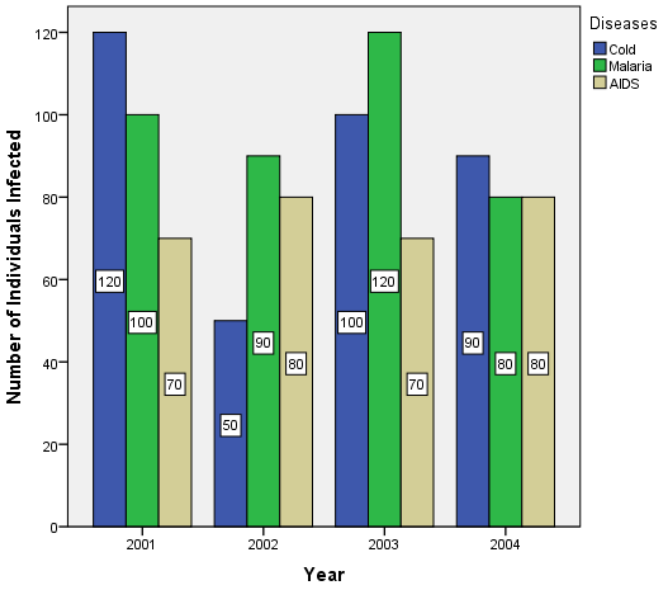


Multiple Bar Chart

Multiple bar chart is used to display the frequencies (magnitudes) of two or more variables data given at different places, periods or timings. Different bars are used represent the different variables at the same place, period or timing. The bars at different places, periods or timings are separated by a constant (an equal) space. The different bars representing the different variables are shaded by different colors, lines or dots for identifying the variables.

Example 2.13. The number of new cases of cold, malaria and HIV/AIDS from 2001 to 2004 in a certain place is shown below. Construct multiple bar chart for this data.

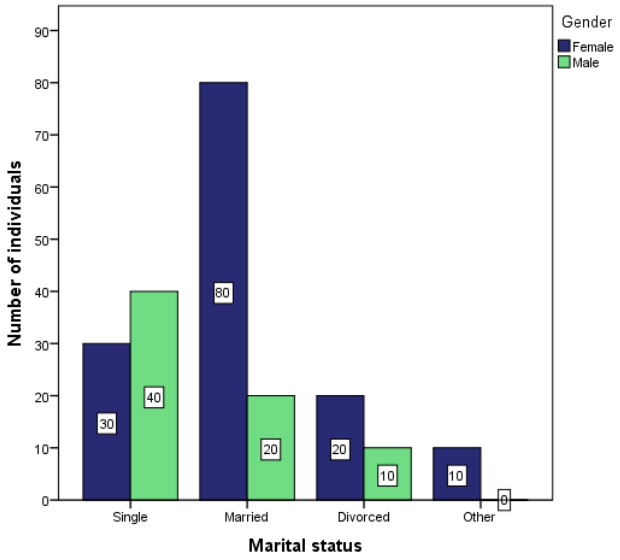
Year	Cold	Malaria	AIDS
2001	120	100	70
2002	50	90	80
2003	100	120	70
2004	90	80	80



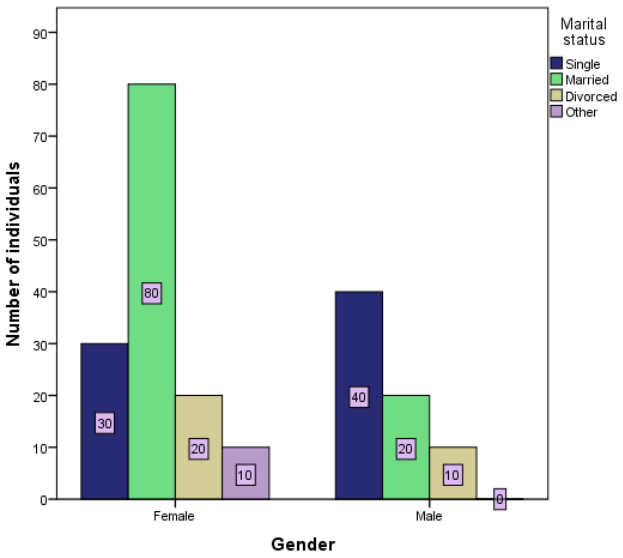
Multiple bar chart can also be used for displaying two-way frequency tables.

Example 2.14. Present the data given in example 2.11 using multiple bar chart.

Solution: Two possible multiple bar charts can be constructed. One is by marking the marital status categories along the x-axis and using the categories of gender as clusters.



The other is by marking the gender categories along the x-axis and using the categories of marital status as clusters.

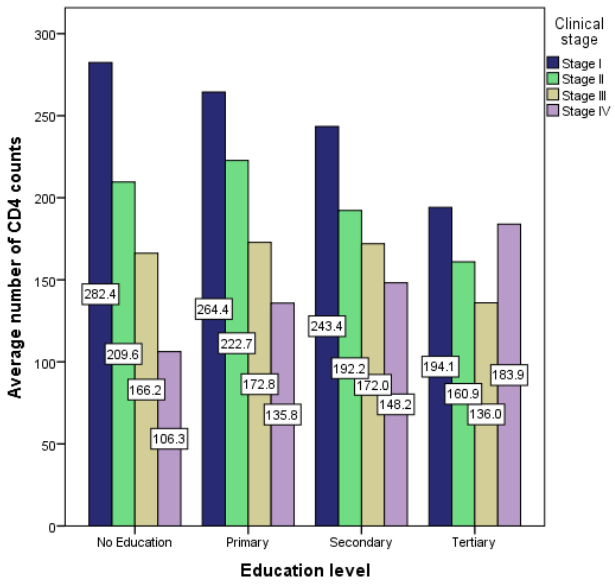


In addition, a multiple bar chart is used to present the magnitude of a quantitative variable (usually, a dependent variable) along the categories (values) of two qualitative or discrete variables (usually, independent variables).

Example 2.15. The following table presents the average CD4 counts of 1464 HIV/AIDS patients classified by their educational level and clinical stage (values in parenthesis are frequencies). Construct multiple bar chart for this data.

Education Level	Clinical Stage			
	Stage I	Stage II	Stage III	Stage IV
No Education	282.4 (65)	209.6 (95)	166.2 (110)	106.3 (27)
Primary	264.4 (119)	222.7 (194)	172.8 (168)	135.8 (34)
Secondary	243.4 (116)	192.2 (169)	172.0 (173)	148.2 (34)
Tertiary	194.1 (45)	160.9 (55)	136.0 (45)	183.9 (10)

Solution: By marking the education level categories in the x-axis and using the categories of clinical stage as clusters, the multiple bar chart showing the average number of CD4 counts is shown below.



2.3.3 Line Graph

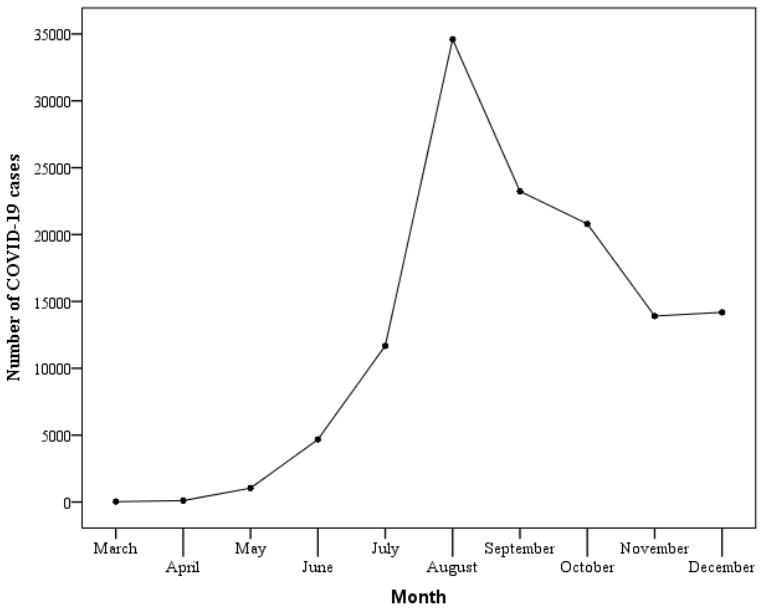
A line graph is used for displaying the magnitude of a quantitative variable measured over time (time-series data). The time variable (in weeks, months or years) is marked along the horizontal $x-$axis, and the values of the variable being studied is marked on the vertical $y-$axis. Then the intersection points of the value of the variable and the corresponding time points are joined by a line.
Such a graph is useful for assessing and monitoring the trend of a particular situation, like epidemics, overtime. It is also used to examine the relationship between the variable of interest and time.

Example 2.16. The first case of COVID-19 case was found in Ethiopia on March 13, 2020 and the total number of cases in the year was 124264. The distribution of the number of cases from March to December 2020 is presented in the following table.

Month	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Cases	26	105	1041	4674	11684	34601	23237	20801	13905	14190	124264

Construct a line graph for this data.

Solution: The months are marked in the $x-$ axis and the number of cases are marked along in the $y-$axis. Then, the line graphs looks as follows.



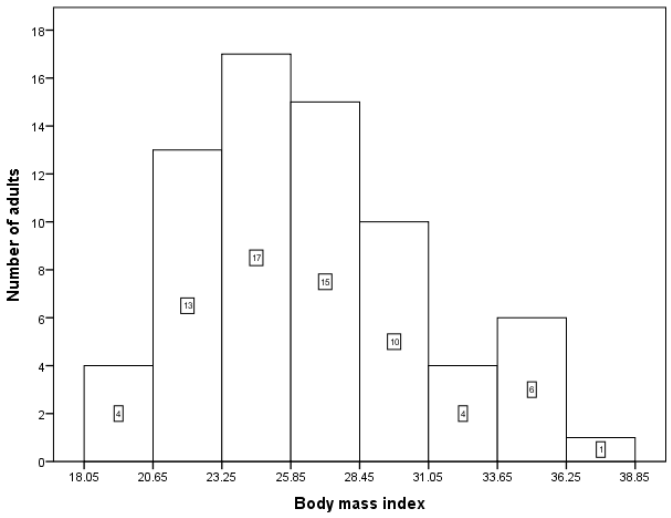
2.3.4 Histogram

Histogram is the most common and widely used graphical presentation used for quantitative data (grouped frequency distribution). It is constructed by marking the values (class boundaries) of the variable of interest on the $x-$axis, and the frequencies (preferably relative or percentage frequencies) along the $y-$axis.

Unlike a bar chart, a histogram uses a series of adjacent or contiguous bars. The base of each bar is determined by the class boundaries on the horizontal axis and the height of each bar is the frequency, relative frequency and percentage frequency of the corresponding class. As a result, the width of each bar represents the class width of the frequency distribution and the height of each bar represents the frequency, relative, or percentage frequency of each class.

Example 2.17. Construct a histogram for the BMI frequency distribution of adults, example 2.6.

Solution: The histogram of the BMI frequency distribution is constructed by marking the class boundaries in the x-axis and the corresponding frequencies along the y-axis.

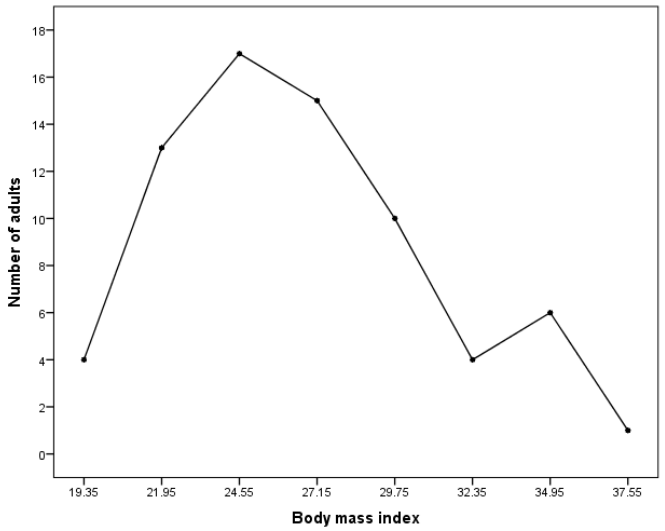


2.3.5 Frequency Curve

A frequency curve, like the histogram, is a graphical display of a continuous frequency distribution. It is a smooth line graph connecting the intersection points of class marks and frequencies, relative or percentage frequencies of the classes. Construction of a frequency curve begins by labeling the class marks along the horizontal axis and the frequencies along the vertical axis. The rightmost and leftmost points are zero, that is, the curve starts at zero and ends at zero.

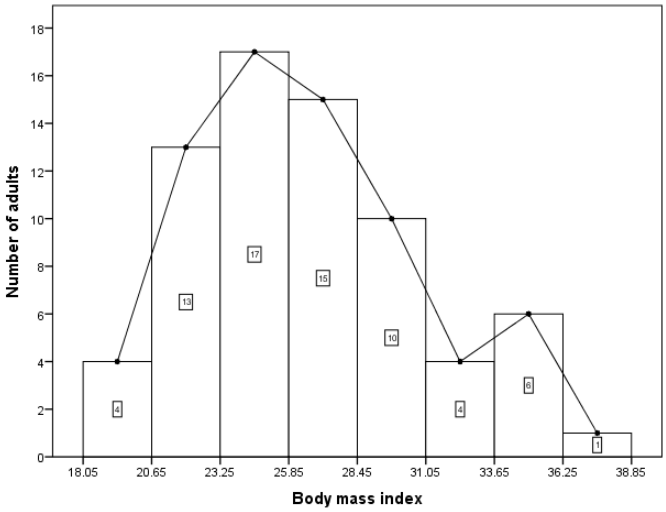
Example 2.18. Construct frequency polygon for the BMI frequency distribution, example 2.6.

Solution: The frequency curve is constructed by marking the class marks in the x-axis and the corresponding frequencies along the y-axis.



Clearly, a frequency curve is similar to a histogram except that there are no bars, only a point in the midpoint of each class at a height proportional to the frequency, relative or percentage

frequency of the class. It can also be constructed by joining the mid-points of the bars of a histogram.



Notes

Effective charts and graphs are simple and clean: thus, it is important that the chart or graph be self-explanatory (i.e., have a figure number for ease of referencing, a descriptive title, properly labeled axes, and an indication of the units of measurement).

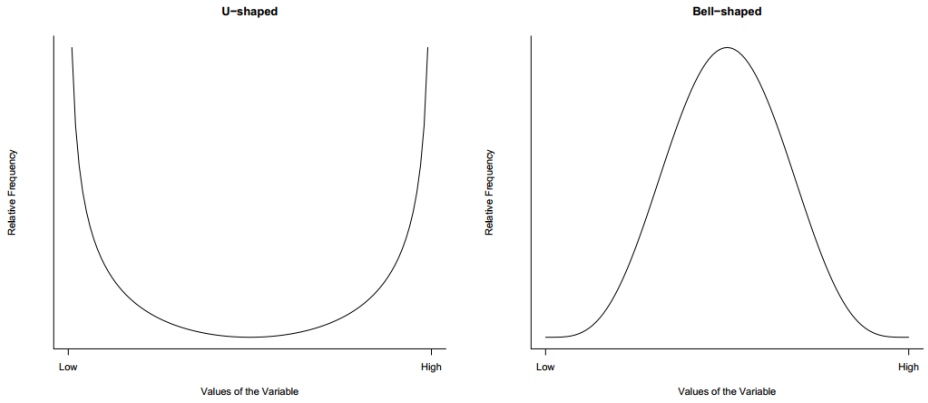
A chart or graph is indeed worth a thousand words and is a powerful means of communicating a great deal of information. But, often some individuals present the data on a stretched or compressed scales of the axes of a chart or graph with the aim of showing whatever they want to show. Such a display affects the user's impression of what the graph represents and is misleading. This is one important argument against a merely descriptive approach to data analysis and an argument for statistical inference. Hence, it is important as a user to clearly understand the scales used for the axes of the chart or graph.

2.4 Shapes of Distributions

One of the most important uses of a histograms and/or frequency curve is to examine how the frequencies are distributed over the values of a quantitative variable and hence, get some idea of the shape or form of a distribution. Hence, examination of a histogram or a frequency curve reveals which values of the variable are highly frequent or which values are less frequent. Had the frequencies of the classes of a frequency distribution been all equal, the frequency curve would be a straight line.

2.4.1 Symmetric Distributions

A frequency curve is said to be *symmetric* when it looks the same to the left and right of the central point. There can be a U-shaped, a Bell-shaped or other possible symmetric curves.



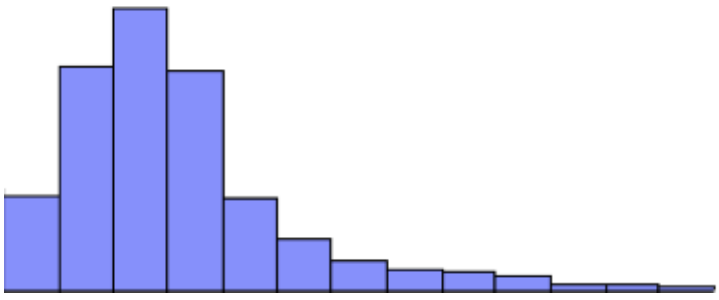
In symmetric curves, the frequency distribution spread around a central value in a similar pattern, that is, the lengths of both tails looks the same. In such symmetric distributions, the number of values below and above the central point are equal.

The distributions of data in applications are never perfectly symmetric, but may be roughly symmetric.

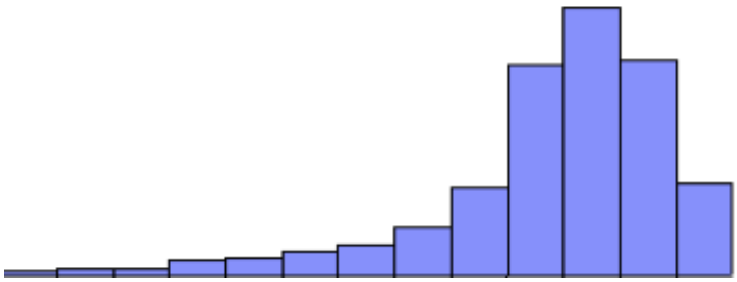
2.4.2 Skewed Distributions

Skewness is the *lack of symmetry* of a distribution. If the frequency curve is *symmetrical*, then it has no skewness.

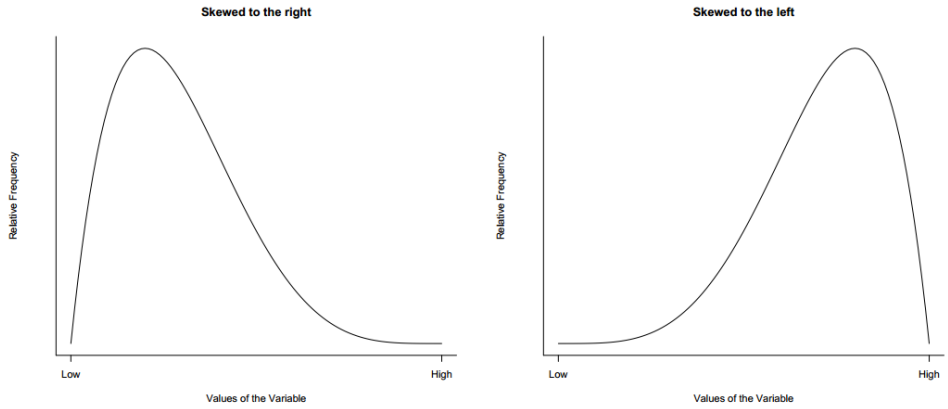
In a *skewed* (*asymmetrical*) distribution, the frequency of large and small values are different. For example, if a few values are extremely large, the right tail of the frequency curve is more elongated. In such case, the distribution is said to be *positively skewed* (*skewed to the right*). The histogram of a positively skewed distribution looks:



On the other hand, if there are a few extremely small values, the left tail is more elongated, and the distribution is said to be *negatively skewed* (*skewed to the left*). The histogram of a negatively skewed distribution looks:



The frequency curves of a positively and negatively skewed distributions look:



Chapter 3

Measures of Central Tendency

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, somewhat summarized after it has been organized into a frequency distribution and presented using charts and graphs, they do not allow us to make concise statements that characterize the distribution of values as a whole. Therefore, there is a need for further condensation, particularly to compare two or more distributions. That is, the entire distribution should be reduced into a single value which can be considered as typical or representative of the set of values. Such a typical value that tends to lie centrally within a set of values of a particular quantitative variable in a dataset is called a *measure of central tendency*.

3.1 Objectives of MCT

1. **To condense a mass of data into one single value**: A measure of central tendency, by condensing masses of data into one single value, enables us to get an idea of the characteristics of the entire dataset. Thus, one value can represent thousands of observations of a variable, even more.

2. **To facilitate comparison**: Statistical devises like averages, percentages and ratios are used for this purpose. For example, to compare the performances of two classes, A and B, instead of comparing each student result, which is infeasible, we can compare the average marks of the two classes.

3.2 Characteristics of Good MCT

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:

- Computed averages: Mean (Arithmetic Mean, Geometric Mean and Harmonic Mean)
- Positional averages: Median and Quantiles (Quartiles, Deciles and Percentiles)
- Mode

However, a measure of central tendency is good or satisfactory if it possesses the following characteristics. Of course, there is no measure which satisfy all these properties:

1. It should be calculated based on all the values of the variable.
2. It should not be affected by extreme values. It should be as close to the maximum number of observed values as possible.
3. It should be defined rigidly which means it should have a definite (unique) value.
4. It should always exist.
5. It should be stable with regard to sampling. This means that if a number of samples of the same size are drawn from a population, the measure of central tendency with a minimum variation should be preferred.

3.3 Summation Notation

The sum of n values, $x_1 + x_2 + \cdots + x_n$, is denoted by the Greek letter Σ (Sigma) as $\sum_{i=1}^{n} x_i$ and it is called the *summation* notation. Let us see some of its properties:

- $\sum_{i=1}^{n} c = nc$ where c is a constant.
- $\sum_{i=1}^{n} (x_i \pm c) = \sum_{i=1}^{n} x_i \pm nc$
- $\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$
- $\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$
- $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$
- $\sum_{i=1}^{n} (x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$
- $\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$
- $\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = (x_1 + x_2 + \cdots + x_n)(y_1 + y_2 + \cdots + y_n)$
- $\sum_{i=1}^{n} x_i y_i \neq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$
- $\sum_{i=1}^{n} (x_i \pm y_i)^2 = \sum_{i=1}^{n} x_i^2 \pm 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$

3.4 Mean

3.4.1 Simple Arithmetic Mean

The *arithmetic mean* is the simplest but most useful measure of central tendency. It is the 'average' which we compute in our high school arithmetic. It is defined as the sum of all the values of a variable divided by the number of values.

Raw Data

- For a population of N values, $x_1, x_2, \cdots, x_N$ of a particular variable; the population mean (denoted by the Greek letter mu, μ) is given by:

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

- For a sample of n raw values, $x_1, x_2, \cdots, x_n$ of a varaible; the sample mean (denoted by $\bar{x}$ and read as x bar) is:

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

Example 3.1. Find the mean of a sample of 3 values given as 2, 4 and 8.

Solution: There are 3 values, adding all the values and dividing by 3 gives the mean.

$$\bar{x} = \frac{\sum_{i=1}^{3} x_i}{3} = \frac{2+4+8}{3} = \frac{14}{3} = 4.67$$

Example 3.2. The heart rates of 10 patients is 60, 70, 64, 55, 70, 80, 70, 74, 51, 80. Calculate the mean heart rate.

Solution: Adding all the 10 values and then dividing by 10 gives the mean heart rate.

$$\mu = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{60+70+64+55+70+80+70+74+51+80}{10} = \frac{674}{10} = 67.4$$

The average heart rate of the patients is about 67.

Note: The calculation of a sample mean uses the same algorithm as for a population mean and will produce the same answer if computed on the same data. However, it is inappropriate to compute a sample mean for a population or a population mean for a sample. Because both population and sample are important in statistics, a separate symbol is necessary for the population mean and for the sample mean.

Grouped Data

For discrete or grouped frequency distributions with k classes, the sample mean is determined as:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k}$$

where x_i is the i^{th} class value for a discrete frequency distribution or the i^{th} class mark for a grouped frequency distribution and f_i is the corresponding frequency where $n = \sum_{i=1}^{k} f_i$.

Example 3.3. Find the mean number of children for the discrete frequency distribution constructed in example 2.3.

Solution: To find the mean number of children of the discrete frequency distribution, the necessary calculations are as follows:

Class	Number of children (x_i)	Number of families (f_i)	$f_i x_i$
1	0	1	0
2	1	4	4
3	2	7	14
4	3	5	15
5	4	3	12
6	5	1	5
Total		$\sum f_i = 21$	$\sum f_i x_i = 50$

Hence, $\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \frac{50}{21} = 2.38$. On average, each family has about 2.38 children.

Example 3.4. Find the mean BMI for the frequency distribution constructed in example 2.6.

Solution: To find the mean BMI of the frequency distribution, the necessary calculations are as follows:

Class	Class Boundary	Class Mark (x_i)	Frequency (f_i)	$f_i x_i$
1	18.05 - 20.65	19.35	4	77.40
2	20.65 - 23.25	21.95	13	285.35
3	23.25 - 25.85	24.55	17	417.35
4	25.85 - 28.45	27.15	15	407.25
5	28.45 - 31.05	29.75	10	297.50
6	31.05 - 33.65	32.35	4	129.40
7	33.65 - 36.25	34.95	6	209.70
8	36.25 - 38.85	37.55	1	37.55
Total			$\sum f_i = 70$	$\sum f_i x_i = 1861.50$

Thus, $\bar{x} = \frac{\sum_{i=1}^{8} f_i x_i}{\sum_{i=1}^{8} f_i} = \frac{1861.50}{70} = 26.59$. The mean BMI of the 70 adults is 26.59 kg/m^2.

Properties of Arithmetic Mean

- The algebraic sum of the deviations of each value from the arithmetic mean is always zero. That is, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
- The sum of the squares of the deviations from the mean is less than the sum of the squares of the deviations from other value. That is, $\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - a)^2, a \neq \bar{x}$.
- If a constant c is added to (subtracted from) each value in a distribution, then the new mean will be increased by c, that is, $\bar{x}_{new} = \bar{x}_{old} \pm c$.
- If each value of a distribution is multiplied by a nonzero constant c, the new mean will be the original mean multiplied by c, that is, $\bar{x}_{new} = c \times \bar{x}_{old}$.

Exercise 3.1. The mean of 100 values was found to be 40. It was latter discovered that a value was misread as 83 instead of 53. Find out the correct mean.

Exercise 3.2. The mean of 200 items was found to be 50. Later on it was discovered that two items were wrongly read as 92 and 8 instead of the correct values 192 and 88 respectively. Find the correct mean.

3.4.2 Weighted Arithmetic Mean

While calculating the simple arithmetic mean, equal importance is given to all values. But, there are cases where the relative importance (weight) is not the same for all values. When this is the case, it is necessary to assign the observations different weights and then calculate the mean called *weighted arithmetic mean.*

Let $x_1, x_2, \cdots, x_n$ be the values and $w_1, w_2, \cdots, w_n$ be the corresponding weights. Then, the weighted arithmetic mean is denoted by $\bar{x}_w$ and is given by:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}$$

Example 3.5. A teacher attaches 2 to quiz, 3 to midterm and 5 for final exam. If a student gets 90, 50 and 60 for quiz, midterm and final exam, respectively, what is the average academic performance of the student?

Solution: The variable of interest is score with three values $x_i = 90, 50, 60$ and their corresponding weights are $w_i = 2, 3, 5$.

$$\bar{x}_w = \frac{\sum_{i=1}^{3} w_i x_i}{\sum_{i=1}^{3} w_i} = \frac{2(90) + 3(50) + 5(60)}{2 + 3 + 5} = \frac{630}{10} = 63$$

The average academic performance of the student is 63.

3.4.3 Combined Mean

If there are g different groups having the same units of measurement with mean $\bar{x}_1$, $\bar{x}_2$, $\cdots$, $\bar{x}_g$; and number of sample observations n_1, n_2, $\cdots$, n_g; respectively, then the mean of all the groups called *combined mean* (denoted by $\bar{x}_c$) is given by:

$$\bar{x}_c = \frac{\sum_{i=1}^{g} n_i \bar{x}_i}{\sum_{i=1}^{g} n_i} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_g\bar{x}_g}{n_1 + n_2 + \cdots + n_g}.$$

Example 3.6. The mean weight of 50 women working in a hospital is 48 kilograms. The mean weight of 75 men working in the same hospital is 58 kilograms. Find the mean weight of all workers in the hospital.

Solution: Given $n_w = 50$, $\bar{x}_w = 48$, $n_m = 75$, $\bar{x}_m = 58$. Required $\bar{x}_c =$?

$$\bar{x}_c = \frac{n_w\bar{x}_w + n_m\bar{x}_m}{n_w + n_m} = \frac{50(48) + 75(58)}{50 + 75} = \frac{6750}{125} = 54.$$

The mean weight of all the 125 persons working in the hospital is 54 kg.

Example 3.7. The mean mark in statistics of 50 students in a class was 72 and that of the 35 boys was 75. Find the mean mark of the girls in the class.

Solution: Given $n = 50$, $\bar{x}_c = 72$, $n_b = 35$, $\bar{x}_b = 75$, $\Rightarrow n_g = n - n_b = 50 - 35 = 15$. Required $\bar{x}_g =$?.

$$\bar{x}_c = \frac{n_b\bar{x}_b + n_g\bar{x}_g}{n} \Rightarrow \bar{x}_g = \frac{n\bar{x}_c - n_b\bar{x}_b}{n_g} = \frac{50(72) - 35(75)}{15} = 65.$$

The mean mark of the 15 girls in the class is 65.

Note: The arithmetic mean fulfils all characteristic of good measures of central tendency with the exception that it is highly affected by extreme values (a few very large or very small values). For example, the mean for the values 115, 110, 119, 117, 121 and 126 is $\bar{x} = 118$. Similarly, the mean for the values 85, 65, 87, 73 and 280 is again 118. But, this mean value is not a good measure of central tendency for the second dataset because out of the five values, most (four) of the values are 87 or less. Hence, this mean is not representative of the second dataset as a whole.

In addition, a mean cannot be calculated for a frequency distribution with open-ended classes (a frequency distribution with no lower class boundary of the first class or with no upper class boundary of the last class or with both).

Mean is not appropriate for qualitative (either nominal or ordinal) variables' data. Such data is summarized by obtaining the frequency and percentage of each category of the variable.

3.5 Median

It has been pointed out that mean is affected by extreme values to a great extent. Hence, some better measure is preferable and median is one of them. *Median* (denoted by $\tilde{x}$) ia a

single point that divides a set of ordered values of a particular variable into *two equal parts* such that half (50%) of the values of the variable are less than it and the remaining half (50%) of the values are greater than it.

Raw Data or Discrete FD

The median for a set of n values is the middle value if n is odd or the arithmetic mean of the middle two values if n is even. That is, if n is odd, $\tilde{x} = \left(\frac{n+1}{2}\right)^{th}$ value; if n is even, $\tilde{x} = \frac{\left(\frac{n}{2}\right)^{th} \text{ value} + \left(\frac{n}{2}+1\right)^{th} \text{ value}}{2}$. Note here, before using the these formula, the values should be arranged in an ascending order of their magnitude.

Example 3.8. Find the median of 180, 201, 220, 191, 219, 209 and 220.

Solution: There are $n = 7$ (n is odd) values. Hence, the median is the middle value after arranging in an increasing order. That is, $\tilde{x} = 4^{th}$ value=209. This means, about 50% of the values are below 209 or about 50% of the values are above 209.

Example 3.9. Consider 62, 63, 64, 65, 66, 66, 68 and 78; and calculate the median.

Solution: There are $n = 8$ (n is even) values. Thus, the median is the average of the middle two values (4^{th} and 5^{th} values). Thus, $\tilde{x} = \frac{4^{th} \text{ value} + 5^{th} \text{ value}}{2} = \frac{65+66}{2} = 65.5$. About 50% of the values are below 65.5 or about 50% of the values are above 65.5.

Exercise 3.3. Find the median number of children using the discrete frequency distribution constructed under example 2.3 and interpret it.

Grouped FD

To find median for a grouped frequency distribution, the median class should be identified first. The median class is the class corresponding to the minimum *less than* cumulative frequency that contains the value $\frac{n}{2}$ where n is the total number of observations.

Then, the median value is given by the formula:

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

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 (it is the sum of all the frequencies up to but not including the median class), $f_{\tilde{x}}$ is frequency of the median class and w is the class width of the median class.

Example 3.10. Find the median of the BMI frequency distribution constructed in example 2.6 and interpret it.

Solution: First, calculate less than cumulative frequencies and identify the median class.

Class	Class Boundary	f_i	LCF (F_i)
1	18.05 - 20.65	4	4
2	20.65 - 23.25	13	17
3	23.25 - 25.85	17	34
4	25.85 - 28.45	15	49
5	28.45 - 31.05	10	59
6	31.05 - 33.65	4	63
7	33.65 - 36.25	6	69
8	36.25 - 38.85	1	70
Total		70	

The median class is the class having the less than cumulative frequency containing the value $\frac{n}{2} = \frac{70}{2} = 35$. This implies, the 4^{th} class is the median class: 25.85 - 28.45.

$$\tilde{x} = 25.85 + \left(\frac{35 - 34}{15}\right) \times 2.6 = 25.85 + 0.17 = 26.02.$$

Therefore, 50% of the adults have BMI below 26.02 kg/m^2 or 50% of the adults have BMI above 26.02 kg/m^2.

Exercise 3.4. Find the median of the following data.

Class	13.5-22.5	22.5-31.5	31.5-40.5	40.5-49.5	49.5-58.5
Frequency	3	9	12	20	3

Note: Median is not sensitive to extreme values, that is, it is *robust*. It can also be calculated for frequency distributions with open-ended classes.

3.6 Other Measures of Location: Quantiles

As discussed before, median divides a set of values, arranged in order, into two equal parts. There are also other positional measures that divide a set of values arranged in order into more than two equal parts. These measures are collectively known as *quantiles*, which include *quartiles*, *deciles* and *percentiles*.

3.6.1 Quartiles

Quartiles are points that divide a set of values, arranged in order, of a variable into *four* equal parts. Thus, there are 3 points denoted by Q_1 (called *lower quartile*), Q_2 (called *middle quartile*) and Q_3 (called *upper quartile*). This means, 25% of the values of the variable are below Q_1, 50% of the values are below Q_2 and 75% of the values are below Q_3.

Raw Data or Discrete FD

Let Q_i be the i^{th} quartile ($i = 1, 2, 3$), then $Q_i = \left[\frac{i(n+1)}{4}\right]^{th}$ value, $i = 1, 2, 3$.

Example 3.11. Consider the observations 62, 63, 64, 65, 66, 66, 68; and determine Q_1, Q_2 and Q_3.

Solution: The $n = 7$ observations are already arranged in an increasing order. Then,
$Q_1 = \left[\frac{(7+1)}{4}\right]^{th}$ value $= 2^{nd}$ value $= 63$
$Q_2 = \left[\frac{2(7+1)}{4}\right]^{th}$ value $= 4^{th}$ value $= 65$
$Q_3 = \left[\frac{3(7+1)}{4}\right]^{th}$ value $= 6^{th}$ value $= 66$

Note: If the i^{th} quartile position index is not a whole number or if it has decimals, linear interpolation is used. For instance, if $Q_i = (I.D)^{th}$ value, it can be obtained as $Q_i = I^{th}$ value $+ 0.D\{(I+1)^{th}$ value $- I^{th}$ value$\}$).

Example 3.12. Given 420, 430, 435, 438, 441, 449, 490, 500, 510 and 515. Find all quartiles.

Solution: As usual, the observations should be arranged in an increasing order.
$Q_1 = \left[\frac{(10+1)}{4}\right]^{th}$ value $= 2.75^{th}$ value $= 2^{nd} + 0.75$ (3^{rd} - 2^{nd}) = 430+0.75(435-430) = 433.75
$Q_2 = \left[\frac{2(10+1)}{4}\right]^{th}$ value $= 5.5^{th}$ value $= 5^{th} + 0.5$ (6^{th} - 5^{th}) = 441+0.5(449-441) = 445
$Q_3 = \left[\frac{3(10+1)}{4}\right]^{th}$ value $= 8.25^{th}$ value $= 8^{th} + 0.25$ (9^{th} - 8^{th}) = 500+0.25(510-500)= 502.5

Grouped FD

Like the median, to find the i^{th} quartile for a grouped frequency distribution, its class should be identified first. The i^{th} quartile class is the class corresponding to the minimum *less than* cumulative frequency that contains the value $\frac{in}{4}$. Then, the Q_i value is given by:

$$Q_i = L_{Q_i} + \left(\frac{\frac{in}{4} - F_{Q_i-1}}{f_{Q_i}}\right) \times w,\ i = 1, 2, 3$$

where L_{Q_i} is the lower class boundary of the i^{th} quartile class, F_{Q_i-1} is the less than cumulative frequency just before the i^{th} quartile class, f_{Q_i} is frequency of the i^{th} quartile class and w is the class width of the i^{th} quartile class.

Example 3.13. Calculate all quartiles for the BMI frequency distribution constructed in example 2.6 and interpret the results.

Solution: Calculate the less than cumulative frequencies first. These are already obtained in the solution section of example 3.10.

Q_1 class: $\frac{n}{4} = \frac{70}{4} = 17.5$. The Q_1 class is the 3^{rd} class: $23.25 - 25.85$.

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}}\right) \times w = 23.25 + \left(\frac{17.5 - 17}{17}\right) \times 2.6 = 23.25 + 0.08 = 23.33.$$

$\rightarrow$ 25% of the adults have BMI below 23.33 kg/m^2 or 75% of the adults have BMI above 23.33 kg/m^2.

Q_2 class: $\frac{2n}{4} = \frac{2(70)}{4} = 35$. The Q_2 class is the 4^{th} class: $25.85 - 28.45$.

$$Q_2 = L_{Q_2} + \left(\frac{\frac{2n}{4} - F_{Q_2-1}}{f_{Q_2}}\right) \times w = 25.85 + \left(\frac{35 - 34}{15}\right) \times 2.6 = 25.85 + 0.17 = 26.02.$$

$\rightarrow$ 50% of the adults have BMI below 26.02 kg/m^2 or 75% of the adults have BMI above 26.02 kg/m^2.

Q_3 class: $\frac{3n}{4} = \frac{3(70)}{4} = 52.5$. The Q_3 class is the 5^{th} class: $28.45 - 31.05$.

$$Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F_{Q_3-1}}{f_{Q_3}}\right) \times w = 28.45 + \left(\frac{52.5 - 49}{10}\right) \times 2.6 = 28.45 + 0.61 = 29.06.$$

$\rightarrow$ 75% of the adults have BMI below 29.06 kg/m^2 or 25% of the adults have BMI above 29.06 kg/m^2.

3.6.2 Deciles

Deciles are points that divide a set of values, arranged in order, of a variable into *ten* equal parts. Here, there are 9 points denoted by D_1, D_2, $\cdots$, D_9. This means, 10% of the values of the variable are below D_1, 20% of the values are below D_2, $\cdots$, 90% of the values are below D_9.

Raw Data or Discrete FD

Let D_i be the i^{th} decile ($i = 1, 2, \cdots, 9$), then $D_i = \left[\frac{i(n+1)}{10}\right]^{th}$ value, $i = 1, 2, \cdots, 9$. Like the quartiles, if the i^{th} decile position index is not a whole number, linear interpolation is used.

Example 3.14. Given the data: 420, 430, 435, 438, 441, 449, 490, 500, 510 and 515. Find the 1^{st} and 7^{th} deciles.

Solution: As before, the observations of the variable should be arranged in an increasing order.
$D_1 = \left[\frac{(10+1)}{10}\right]^{th}$ value = 1.1^{st} value = 1^{st} + 0.1 (2^{nd} - 1^{st}) = 420+0.1(430-420) = 421
$D_7 = \left[\frac{7(10+1)}{10}\right]^{th}$ value = 7.7^{th} value = 7^{th} + 0.7 (8^{th} - 7^{th}) = 490+0.7(500-490)= 497

Grouped FD

For a frequency distribution, the i^{th} decile class is the class corresponding to the minimum *less than* cumulative frequency that contains the value $\frac{in}{10}$. Thus,

$$D_i = L_{D_i} + \left(\frac{\frac{in}{4} - F_{D_i-1}}{f_{D_i}}\right) \times w, \ i = 1, 2, \cdots, 9$$

where L_{D_i} is the lower class boundary of the i^{th} decile class, F_{D_i-1} is the less than cumulative frequency just before the i^{th} decile class, f_{D_i} is frequency of the i^{th} decile class and w is the class width of the i^{th} decile class.

Example 3.15. Calculate the 5^{th} and 8^{th} deciles for the BMI frequency distribution constructed in example 2.6 and interpret the results.

Solution: The less than cumulative frequencies are already presented in the solution section of example 3.10.
D_5 class: $\frac{5n}{10} = \frac{5(70)}{10} = 35$. The D_5 class is the 4^{th} class: $25.85 - 28.45$.

$$D_5 = L_{D_5} + \left(\frac{\frac{5n}{10} - F_{D_5-1}}{f_{D_5}}\right) \times w = 25.85 + \left(\frac{35-34}{15}\right) \times 2.6 = 25.85 + 0.17 = 26.02.$$

$\rightarrow$ 50% of the adults have BMI below 26.02 kg/m^2 or 50% of the adults have BMI above 26.02 kg/m^2.

D_8 class: $\frac{8n}{10} = \frac{8(70)}{10} = 56$. The D_8 class is the 5^{th} class: $28.45 - 31.05$.

$$D_8 = L_{D_8} + \left(\frac{\frac{8n}{10} - F_{D_8-1}}{f_{D_8}}\right) \times w = 28.45 + \left(\frac{56-49}{10}\right) \times 2.6 = 28.45 + 0.7 = 29.15.$$

$\rightarrow$ 80% of the adults have BMI below 29.15 kg/m^2 or 20% of the adults have BMI above 29.15 kg/m^2.

3.6.3 Percentiles

Percentiles are points that divide a set of values, arranged in order, of a varaible into 100 equal parts. The 99 points are denoted by P_1, P_2, $\cdots$, P_{99}.

Raw Data or Discrete FD

Let P_i be the i^{th} percentile ($i = 1, 2, \cdots, 99$), then $P_i = \left[\frac{i(n+1)}{100}\right]^{th}$ value, $i = 1, 2, \cdots, 99$.

Example 3.16. Given 420, 430, 435, 438, 441, 449, 490, 500, 510 and 515. Find the 40^{th} and 75^{th} percentiles.

Solution: The observations should be arranged in an increasing order.
$P_{40} = \left[\frac{40(10+1)}{100}\right]^{th}$ value = 4.4^{th} value = 4^{th} + 0.4 (5^{th} - 4^{th}) = 438+0.4(441-438) = 439.2
$P_{75} = \left[\frac{75(10+1)}{100}\right]^{th}$ value = 8.25^{th} value = 8^{th} + 0.25 (9^{th} - 8^{th}) = 500+0.25(510-500) = 502.5

Grouped FD

For a frequency distribution, the i^{th} percentile class is the class corresponding to the minimum *less than* cumulative frequency that contains the value $\frac{in}{100}$. Then, the i^{th} percentile value is given by:

$$P_i = L_{P_i} + \left(\frac{\frac{in}{4} - F_{P_i-1}}{f_{P_i}}\right) \times w, \; i = 1, 2, \cdots, 99$$

where L_{P_i} is the lower class boundary of the i^{th} percentile class, F_{P_i-1} is the less than cumulative frequency just before the i^{th} percentile class, f_{P_i} is frequency of the i^{th} percentile class and w is the class width of the i^{th} percentile class.

Example 3.17. Calculate the 30^{th} and 80^{th} percentiles for the BMI frequency distribution constructed in example 2.6 and interpret the results.

Solution: Recall the cumulative frequencies presented in the solution section of example 3.10. P_{30} class: $\frac{30n}{100} = \frac{30(70)}{100} = 21$. The P_{30} class is the 3^{rd} class: $23.25 - 25.85$.

$$P_{30} = L_{P_{30}} + \left(\frac{\frac{30n}{100} - F_{P_{30}-1}}{f_{P_{30}}} \right) \times w = 23.25 + \left(\frac{21 - 17}{17} \right) \times 4 = 23.25 + 0.61 = 23.86.$$

$\rightarrow$ 30% of the adults have BMI below 23.86 kg/m^2 or 70% of the adults have BMI above 23.86 kg/m^2.
P_{90} class: $\frac{90n}{100} = \frac{90(70)}{100} = 63$. The P_{90} class is the 6^{th} class: $31.05 - 33.65$.

$$P_{90} = L_{P_{90}} + \left(\frac{\frac{90n}{100} - F_{P_{90}-1}}{f_{P_{90}}} \right) \times w = 31.05 + \left(\frac{63 - 59}{4} \right) \times 4 = 31.05 + 2.6 = 33.65.$$

$\rightarrow$ 90% of the adults have BMI below 33.65 kg/m^2 or 10% of the adults have BMI above 33.65 kg/m^2.

Relationship between Median, Quartiles, Deciles and Percentiles

$\tilde{x} = Q_2 = D_5 = P_{50}$, $Q_i = P_{i\times 25},\ i = 1, 2, 3$, $D_i = P_{i\times 10},\ i = 1, 2, \cdots, 9$

3.7 Mode

Mode (denoted by $\hat{x}$) is another measure of central tendency. The *mode* of a set of values is the value(s) that occurs *most frequently*. For instance, if a shoe size of 41 has the maximum demand by males, size number 41 is the modal shoe size. A dataset may have one mode (uni-modal) or two modes (bi-modal), more than two modes (multi-modal) or no mode at all (i.e. when all observations are equally frequent).

Raw Data or Discrete FD

In individual series cases or discrete frequency distributions, the mode can be found by inspection.

Example 3.18. Find the mode of the following datasets.

a. 110, 113, 116, 116, 118, 118, 118, 121 and 123.

b. 2, 3, 5, 7 and 8.

c. 15, 18, 18, 18, 20, 22, 24, 24, 24, 26 and 26.

d. 5, 6, 6, 7, 9, 9, 10, 12 and 12.

e. 1, 1, 0, 1, 0, 0, 0, 2, 4 and 3.

Solution: The modal value of each dataset is just the value with the highest frequency.

a. Since 118 occurs more than other values, the mode is 118.

b. Each value occurs once (equally frequent), the data has no mode.

c. 18 and 24 occur three times, hence the modal values are 18 and 24 (bi-modal).

d. Tri-modal (multi-modal): 6, 9 and 12.

e. The modal value here is 0 as it occurs more number of times than other values.

Exercise 3.5. Find the modal values of the categorical frequency distribution given in example 2.1 and discrete frequency distribution given in 2.3, and interpret them.

Note: If a dataset is not exactly bi-modal (multi-modal) but contains two (more than two) values that are more dominant than others, some researchers take the liberty of referring to the dataset as bi-modal (multi-modal) even without an exact tie for the mode.

Grouped FD

In a grouped frequency distribution, the modal value is located in the class with *highest frequency* and that class is the modal class. Its value is given by:

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

where $L_{\hat{x}}$ is the lower class boundary of the modal class, $f_{\hat{x}}$ is frequency of the modal class, $f_{\hat{x}-1}$ is the frequency just before the modal class, $f_{\hat{x}+1}$ is the frequency just after the modal class and w is the class width of the modal class.

Example 3.19. Find the modal score of the BMI frequency distribution constructed in example 2.6.

Solution: The class having highest frequency is the 3^{rd} class $(23.25 - 25.85)$, hence it is the modal class.

$$\hat{x} = 23.25 + \left[\frac{17 - 13}{(17 - 13) + (17 - 15)}\right] \times 2.6 = 23.25 + 1.73 = 24.98$$

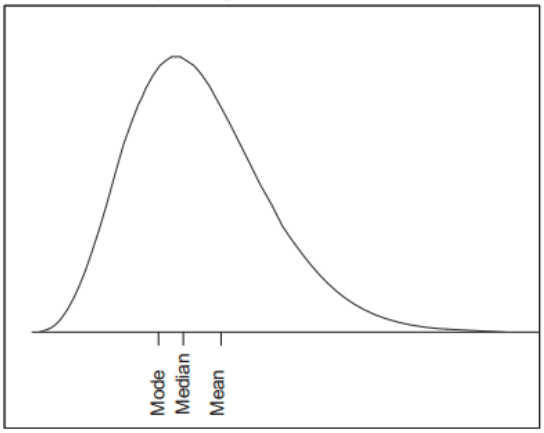
Note: Mode is applicable for both quantitative and qualitative variables. It is not affected by extreme values. But, it often does not exist and its value may not be unique.

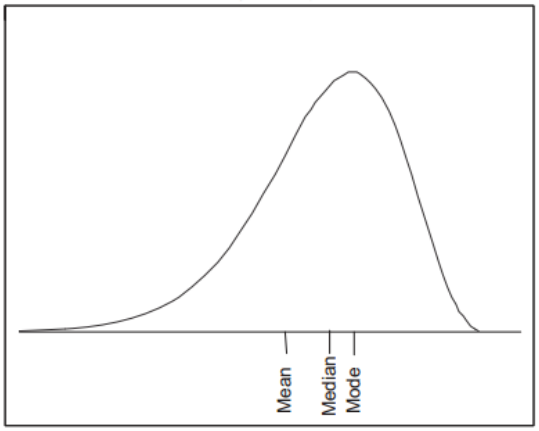
3.8 Relationship Between Mean, Median and Mode

In a symmetric bell-shaped (uni-modal) distribution; the mean, median and modal values are approximately equal. Hence, the number of observations below and above the mean are equal. In addition, the corresponding pairs of quartiles, deciles and percentiles are at equidistance from the median. For example, first quartile and third quartile have the same distance from the median.

In case of skewed distributions, median is a better measure of central tendency than the mean. This is related to the fact that the mean can be highly influenced by an *extreme value* (*outlier*).

For example, in a *positively skewed* distribution, a few observations are extremely large, the mean of the distribution becomes greater than the median and mode (mean > median > mode). Therefore, the number of observations below the mean is greater than the number of observations above the mean.





In a *negatively skewed* distribution, there are a few extremely small observations, then the mean will be the smallest of the other two averages (mean < median < mode). In this case, the number of observations below the mean is less than the number of observations above the mean.

Note: In a moderately skewed distribution, Mean - Mode = 3 (Mean - Median).

Exercise 3.6. In a certain moderately skewed distribution, the mean is 74 and the mode is 60. What is the skewness type? Compute the median?

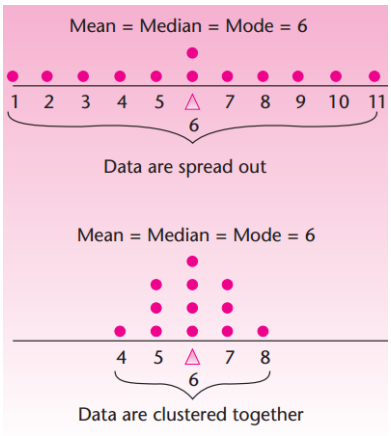
Chapter 4

Measures of Variation

Let us consider the following three datasets, each with a sample of 12 observations. All the three datasets have the same mean 6 and the same median 6. Are all the datasets the same?

A	1	2	3	4	5	6	6	7	8	9	10	11
B	4	5	5	5	6	6	6	6	7	7	7	8
C	6	6	6	6	6	6	6	6	6	6	6	6

But, by inspection, it is apparent that the observations in the three datasets differ remarkably from one another.



Therefore, two or more datasets may have the same measures of central tendency but they may be quite different. Thus, a measure of central tendency alone does not provide sufficient information about the nature of the dataset. Thus, to have a clear picture of the characteristics of the entire distribution, one needs to have a measure of variability among the values of the variable, whether the values are clustered to each other or scattered away.

Variation or *dispersion* may be defined as the extent of scatteredness of the values of a variable around a measure of central tendency. When the variation of a dataset is smaller, the values of the variable are concentrated near a particular measure of central tendency, that is, the values are more homogeneous. A dataset with a larger variation exhibits a greater amount of heterogeneity among the values of the variable, that is, the values are scattered away from a particular measure of central tendency. If all the values are the same, there is

no variability and vice versa.

Thus, a *measure of variation* tells us the extent to which the values of a variable in a dataset vary about a particular measure of central tendency. For example, the values in dataset A above are more variable than the values in dataset B and C, the values in dataset B are also more variable than the values in dataset C in which all the values are the same (no variation).

Note: Variability is encountered in all our everyday lives, and statistical thinking can give us a useful way to incorporate this variability into our decision-making processes. By variability, it means that successive observations of a system or phenomenon do not produce exactly the same result or the values of the variable of interest are not the same for different objects.

- For example, consider the gasoline mileage performance of a car. Do we always get exactly the same mileage performance on every tank of fuel? Of course not-in fact, sometimes the mileage performance varies considerably. This observed variability in gasoline mileage depends on many factors, such as the type of road mostly used, the changes in condition of the vehicle over time (which could include factors such as tire inflation, engine compression, or valve wear), the brand and/or octane number of the gasoline used, or possibly even the weather conditions that have been recently experienced. These factors represent potential sources of variability in the system.
- Similarly, consider the academic performance of students in statistics course. Do all the students in a class score the same result? Of course not. The academic performance varies from student to student. Like the previous one, this observed variability in the students' performance depends on many factors like gender, number of hours studying per week, $\cdots$. These factors represent potential sources of variability among the students.

Therefore, statistics gives us a framework for describing such variability and for identifying about which potential sources of variability are the most important or which have the greatest impact on the variable of interest.

4.1 Objectives of Measures of Variation

1. **To have an idea about the reliability of a measure of central tendency**. If the degree of scatterdness is large, a measure of central tendency is less reliable. If the variation is smaller, it indicates that the measure of central tendency is a good representative of all the values in a dataset.
2. **To compare two or more datasets with regard to their variability**. A dataset with smaller variation posses more uniformity (consistency) among the values of a variable or it is less variable.
3. **To provide information about the structure of the observations in a dataset**. A measure of variation gives an idea about the limits of the expansion of the values of a dataset.

4.2 Types of Measures of Variation

There are two types of measures of variation: *absolute* and *relative*. A measure of variation is said to be in an *absolute* form when it shows the actual amount of variation of the values of a variable of a dataset in concrete units in which the data has been already expressed. In other words, all absolute measures of variation have units. As a result, if two or more distributions differ in their units of measurement, their variability cannot be compared by using an absolute measure of variation. Also, the magnitude of the absolute measures of variation depends on the magnitude of the observed values in a dataset. That is, if the magnitude of the values is large in a dataset, the value of the absolute measures will also be large. Therefore, an absolute measure of variation fails to be appropriate for comparing two or more datasets if the magnitudes of the values among the datasets are too different.

A *relative* measure of variation shows the actual amount of variation of the observations in a dataset in a unit-less pure number form. It also takes into account the differences in the sizes of observed values between two or more datasets. Hence, it can be used for making comparisons between different distributions.

Absolute Measures of Variation	Relative Measures of Variation
Range	Coefficient of Range
Inter-Quartile Range	Coefficient of Inter-Quartile Range
Variance and Standard Deviation	Coefficient of Variation
	Standard Scores

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

4.2.1 Range

Range is the simplest and crudest measure of variation. It is defined as the difference between the largest and the smallest values in a dataset. That is, $R = L - S$. Its corresponding relative measure is called *coefficient of range* which is defined as $CR = \frac{L-S}{L+S}$.

Example 4.1. The hemoglobin level (in g/dl) of a sample of 13 apparently healthy men aged 20-24 years is given as: 17.5, 15.7, 15.8, 16.2, 15.5, 15.3, 17.4, 13.5, 18.8, 17.5, 15.8, 16.2, 14.3. Find the range and coefficient of range of this dataset.

Solution: The ordered values are: 13.5, 14.3, 15.3, 15.5, 15.7, 15.8, 15.8, 16.2, 16.2, 17.4, 17.5, 17.5, 18.8. This implies $R = 18.8 - 13.5 = 5.3$ and $CR = \frac{18.8-13.5}{18.8+13.5} = 0.16$.

Note: Range hardly satisfies any property of a good measure of dispersion as it is based on the two extreme values only, ignoring the others.

4.2.2 Inter-Quartile Range

The inter-quartile range is the difference between the first (lower) and third (upper) quartiles, that is, $IQR = Q_3 - Q_1$. Consequently, the relative measure of the inter-quartile range is called *coefficient of inter-quartile range* and defined as $CIQR = \frac{Q_3-Q_1}{Q_3+Q_1}$.

Example 4.2. Find the inter-quartile range the hemoglobin data given in example 4.1.

Solution: First, let us find the lower and upper quartiles of the dataset.

- $Q_1 = \left(\frac{13+1}{4}\right)^{th}$ value $= 3.5^{th}$ value $= 15.3 + 0.5(15.5 - 15.3) = 15.4$.
- $Q_3 = \left[\frac{3(13+1)}{4}\right]^{th}$ value $= 10.5^{th}$ value $= 17.4 + 0.5(17.5 - 17.4) = 17.45$.

Thus, the inter-quartile range of the dataset is $IQR = 17.45 - 15.4 = 2.05$ and the coefficient of the inter-quartile range is $CIQR = \frac{17.45-15.4}{17.45+15.4} = 0.06$.

Note: Inter-quartile range involves only the middle 50% of the values by excluding the values below the lower quartile and the values above the upper quartile. Note also that, it does not take into account all the values between Q_1 and Q_3. It means that, no idea about the variation of even the 50% mid values is available from this measure. Anyhow, it is a preferred measure of variation when the median is used as a measure of center (i.e., in case of skewed distribution) and is used to identify an outlier (extreme values) in a dataset.

- An observed value is said to be an *outlier* if it is less than $Q_1 - 1.5(Q_3 - Q_1)$ or greater than $Q_3 + 1.5(Q_3 - Q_1)$.
- An observed value is said to be an *extreme value* if it is less than $Q_1 - 3(Q_3 - Q_1)$ or greater than $Q_3 + 3(Q_3 - Q_1)$.

Example 4.3. The heart rates (beats per minute) for ten asthmatic patients in a state of respiratory arrest are 167, 150, 125, 120, 150, 151, 40, 136, 120, 150. Are there any outliers or extreme values in this dataset?

Solution: The ordered heart rates are 40, 120, 120, 125, 136, 150, 150, 150, 151, 167.

- $Q_1 = \left(\frac{10+1}{4}\right)^{th}$ value $= 2.75^{th}$ value $= 120 + 0.75(120 - 120) = 120.00$.
- $Q_3 = \left[\frac{3(10+1)}{4}\right]^{th}$ value $= 8.25^{th}$ value $= 150 + 0.25(151 - 150) = 150.25$.

For identifying outliers: $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)] = (74.625, 195.625)$ and for identifying extreme values: $[Q_1 - 3(Q_3 - Q_1), Q_3 + 3(Q_3 - Q_1)] = (29.25, 241.00)$. Therefore, the observation 40 is an outlier in the above dataset but there is no extreme value.

The mean heart rate of the ten asthmatic patients is $\bar{x} = \frac{1}{10}(40+120+120+125+\cdots+167) = 130.9$ beats per minute. In this data set, the heart rate of one patient is considerably lower (it is an outlier) than those of the other patients. What would happen if this observation were removed? In this case, $\bar{x} = \frac{1}{9}(120 + 120 + 125 + \cdots + 167) = 141.00$ beats per minute. The mean has increased by approximately 10 beats per minute; this change demonstrates how much influence a single unusual (outlier) observation can have on the mean.

4.2.3 Variance

Variance is the most superior and widely used measure of variation. It is the sum of the squares of the deviation of each value taken from the mean divided by the total number of values in a dataset. Thus, it measures the average dispersion of the values of a variable around

the mean. If the variance of a dataset is smaller, the values are concentrated near the mean and if it larger, the values are scattered away from the mean.

For a population containing N values, the population variance is denoted by the square of the Greek letter sigma, σ^2. For a raw dataset, the population variance formula is:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}\right]$$

where μ is the population mean of the dataset.

For a sample of n values, the sample variance is denoted by s^2 and calculated using the formulae:

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

where $\bar{x}$ is the sample mean of the dataset.

Example 4.4. Find the variance of 20, 28, 40, 12, 30, 15 and 50.

Solution:

- Considering as a population: $N = 7$, $\mu = 27.86$;

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{7}[(20 - 27.86)^2 + \cdots + (50 - 27.86)^2] = \frac{1}{7}(1120.86) = 160.12$$

- Considering as a sample: $n = 7$, $\bar{x} = 27.86$;

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{6}[(20 - 27.86)^2 + \cdots + (50 - 27.86)^2] = \frac{1}{6}(1120.86) = 186.81$$

Example 4.5. Re-calculate the population and sample variances of the dataset given in example 4.4 using the expanded formula.

Solution: The summary statistics required to use the expanded formula of the variance are $\sum_{i=1}^{7} x_i = 20+28+40+12+30+15+50 = 195$ and $\sum_{i=1}^{7} x_i^2 = 20^2+28^2+40^2+12^2+30^2+15^2+50^2 = 6553$. Thus, the population and sample variances, respectively, are:

$$\sigma^2 = \frac{1}{7}\left[6553 - \frac{(195)^2}{7}\right] = 160.12 \text{ and } s^2 = \frac{1}{6}\left[6553 - \frac{(195)^2}{7}\right] = 186.81.$$

Note: The first main demerit of variance is that its unit is the square of the unit of measurement of the observations in the dataset. For example, the sample variance for the observations $2m$, $6m$ and $4m$ is $4m^2$. The interpretation seems, on average each value differs from the

mean by $4m^2$. This is completely non-sense because the unit of measurement of the variance is not the same as that of the dataset. The other disadvantage of variance is, the variation of the data is exaggerated because the deviation of 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.

4.2.4 Standard Deviation

Standard deviation is the positive square root of variance. For a population containing N elements, the population standard deviation is denoted by the Greek letter sigma, σ and its formula is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

For a sample of n elements, the sample standard deviation is denoted by s and calculated using the formulae:

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

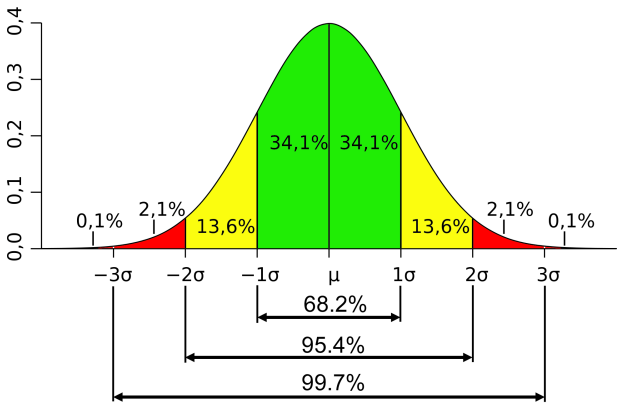
Example 4.6. Find the standard deviation of the dataset given in example 4.4.

Solution: The population standard deviation is $\sigma = \sqrt{160.12} = 12.65$ and the sample standard deviation $s = \sqrt{186.81} = 13.67$.

Note: Standard deviation is considered to be the best measure of variation because the unit of measurement is the same as with the dataset and the exaggeration made by variance is eliminated by taking the square root of it. Thus, in simple words, standard deviation explains the average amount of variation on either sides of the mean. If the standard deviation is smaller, the values are concentrated near the mean and if it large, the values are scattered away from the mean.

For a bell-shaped symmetric (called *normal*) distribution, empirical rule relates the standard deviation (σ) to the proportion of the observed values in a dataset that lie in an interval around the mean (μ):

- 68.2% of the values are within one standard deviation of the mean ($\mu \pm \sigma$),
- 95.4% of the values are within two standard deviations of the mean ($\mu \pm 2\sigma$), and
- 99.7% of the values are within three standard deviations of the mean ($\mu \pm 3\sigma$).



Properties of Variance and Standard Deviation

1. If a constant is added to (subtracted from) each value of a variable in a dataset, the standard deviation as well as the variance remains the same.

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

Pooled Variance and Standard Deviation

If there are g groups having the same units of measurement with sample means $\bar{x}_1$, $\bar{x}_2$, $\cdots$, $\bar{x}_g$; number of sample observations n_1, n_2, $\cdots$, n_g; and sample variances s_1^2, s_2^2, $\cdots$, s_g^2, respectively, then the sample variance of all the g groups called *pooled variance* (denoted by s_p^2) is given by:

$$s_p^2 = \frac{\sum_{i=1}^{g}(n_i - 1)[s_i^2 + (\bar{x}_i - \bar{x}_c)^2]}{\sum_{i=1}^{g} n_i - g} = \frac{(n_1 - 1)[s_1^2 + (\bar{x}_1 - \bar{x}_c)^2] + \cdots + (n_g - 1)[s_g^2 + (\bar{x}_g - \bar{x}_c)^2]}{n_1 + n_2 + \cdots + n_g - g}$$

where $\bar{x}_c$ is the combined (sample) mean of all the g groups. If all the sample means of the groups are equal ($\bar{x}_1 = \bar{x}_2 = \cdots = \bar{x}_g$), then the pooled sample variance formula becomes:

$$s_p^2 = \frac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g} n_i - g} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_g - 1)s_g^2}{n_1 + n_2 + \cdots + n_g - g}$$

Similarly, the pooled population variance can be calculated using the formula:

$$\sigma_p^2 = \frac{\sum_{i=1}^{g} N_i[\sigma_i^2 + (\mu_i - \mu_c)^2]}{\sum_{i=1}^{g} N_i} = \frac{N_1[\sigma_1^2 + (\mu_1 - \mu_c)^2] + \cdots + N_g[\sigma_g^2 + (\mu_g - \mu_c)^2]}{N_1 + N_2 + \cdots + N_g}$$

where μ_c is the combined (population) mean of all the g groups. If $\mu_1 = \mu_2 = \cdots = \mu_g$, then the pooled population variance formula becomes:

$$\sigma_p^2 = \frac{\sum_{i=1}^{g} N_i \sigma_i^2}{\sum_{i=1}^{g} N_i} = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \cdots + N_g\sigma_g^2}{N_1 + N_2 + \cdots + N_g}$$

Exercise 4.1. The mean weight of 150 students is 60 kilograms. The mean weight of boys is 70 kilograms with a standard deviation of 10 kilograms. For the girls, the mean weight weight is 55 kilograms and the standard deviation 15 kilograms. Determine the number of boys and girls. Also, find the combined standard deviation.

Exercise 4.2. A distribution consists of four groups characterized as follows. Find the mean and standard deviation of the distribution. Ans: $\mu_c = 73.8$ and $\sigma = 11.93$

Group	Number of items	Mean	Standard deviation
1	50	61	8
2	100	70	9
3	120	50	10
4	30	83	11

Exercise 4.3. 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.

Example 4.7. The following data are some of the particulars of the distribution of weights of boys and girls in a class. Find the mean and variance of the combined series.

	Boys	Girls
Number	100	50
Mean	60	45
Variance	9	4

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

4.2.5 Coefficient of Variation

Coefficient of variation is a relative measure of standard deviation. It measures how large the standard deviation with respect to the mean. It is defined as the ratio of the standard deviation to the mean and expressed as percent. That is, CV $= \frac{\sigma}{\mu} \times 100\%$ for a population and CV $= \frac{s}{\bar{x}} \times 100\%$ for a sample.

For example, if the sample mean of a dataset is 44 and the standard deviation is 8, then the coefficient of variation is CV $= (\frac{8}{44})100\% = 18.2\%$. It indicates that the sample standard deviation is 18.2% of the value of the sample mean.

The coefficient of variation is a unit-less measure of variation and also takes into account the size of the mean of a distribution. Hence, it is a useful measure for comparing the variability

of variables that have different standard deviation and different means.

If a distribution has a smaller CV, then it has more consistent or more uniform or less variable observations. For field experiments, a smaller CV also indicates more reliability of experimental findings.

Example 4.8.

Given the following summaries on systolic blood pressure (SBP) and cholesterol level.

Variable	Mean	SD
SBP	130mmHg	25mmHg
Cholesterol	200mg/dl	30mg/dl

Which variable's values is more variable/less consitent/less reliable?

Solution: As the standard deviation of systolic blood pressure (SBP) is 25 and cholesterol is 30, it is not correct to say the values of SBP are less variable (more uniform) that of the values of cholesterol. Hence, the appropriate measure of variability for these two different types of variables is to obtain the coefficient of variation for each of the two variables and compare both.

Variable	Mean	SD	CV
SBP	130mmHg	25mmHg	19.23%
Cholesterol	200mg/dl	30mg/dl	15.00%

The values of cholesterol is more uniform/more consistent/more reliable than the values of systolic blood pressure (SBP).

Exercise 4.4. Compare the variability of the following two sample datasets using standard deviation and coefficient of variation:

A. 2 meters, 4 meters, 6 meters

B. 600 liters, 400 liters, 500 liters

Exercise 4.5. 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?

4.2.6 Standard (z) Score

The standard score (called $z-$score) is a measure relative standing. It measures how many standard deviations a given value x_i is above or below the mean depending on whether the $z-$score is positive or negative. The formulas are $z = \frac{x-\mu}{\sigma}$ for a population and $z = \frac{x-\bar{x}}{s}$ for a sample.

Example 4.9. Two public health experts from two different areas were assessed for the time they have taken to accomplish a given task as presented in the table below.

Expert	Actual time	Mean	SD
A	16hr	13hr	2hr
B	28min	20min	4min

Who is better relative to his area?

Solution: The standard score of the two public health experts are calculated first.

Expert	Actual time	Mean	SD	$z-$score
A	16hr	13hr	2hr	1.5
B	28min	20min	4min	2.0

Since the $z-$score for expert A is less than the $z-$score for expert B, expert A performs better (in a shorter time) relative to his area.

Exercise 4.6. Suppose Yoseph scored 90 on a test in which the mean and standard deviation of the class were 70 and 10, respectively. In another test, Helen scored 600 on which the mean and standard deviation of the class were 560 and 40, respectively. Who is better of relative to his/her class?

Note: The standard score ($z-$score) is useful to transform a given data sets in to a new distribution whose mean is 0 and variance is 1.

4.3 Skewness and Kurtosis

4.3.1 Measure of Skewness

Recall that skewness is the lack of symmetry. That is, if the tails of a frequency curve are not equally distributed, the curve is *asymmetric* (*skewed*), see Section 2.4.2. The measure of such degree of asymmetry is called a *measure of skewness*. It is denoted by α_3 and given by the formula:

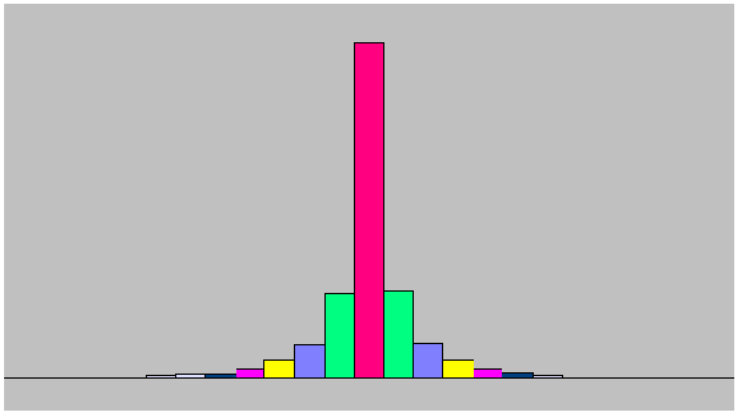
$$\alpha_3 = \frac{\mu_3}{\sqrt{(\mu_2)^3}} \text{ where } \mu_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^r;\ r = 2, 3.$$

If the distribution is *symmetric*, then $\alpha_3 = 0$. On the contrary, if $\alpha_3 > 0$, the distribution is *positively skewed* and if $\alpha_3 < 0$, the distribution is *negatively skewed*.
Two distributions that have the same mean, variance, and skewness could still be significantly different in their shape. We may then look at their kurtosis.

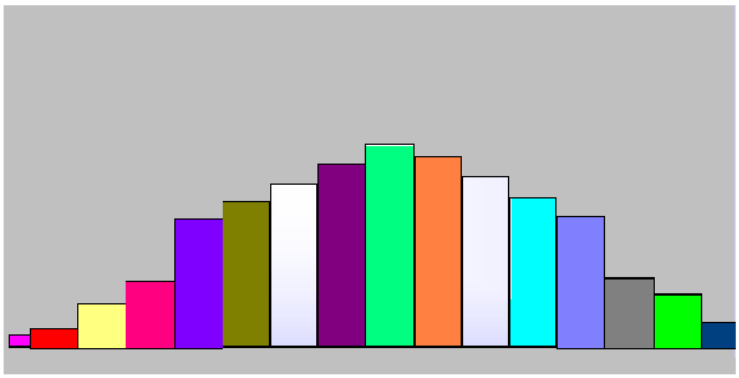
4.3.2 Measure of Kurtosis

Kurtosis refers to the *peakedness or flatness* of a certain distribution. It describes the degree of concentration of the values of a variable around the mode of a distribution, whether the values are concentrated more around the mode (a peaked curve) or scattered away from the mode toward the tails of a frequency curve (a flatter curve).

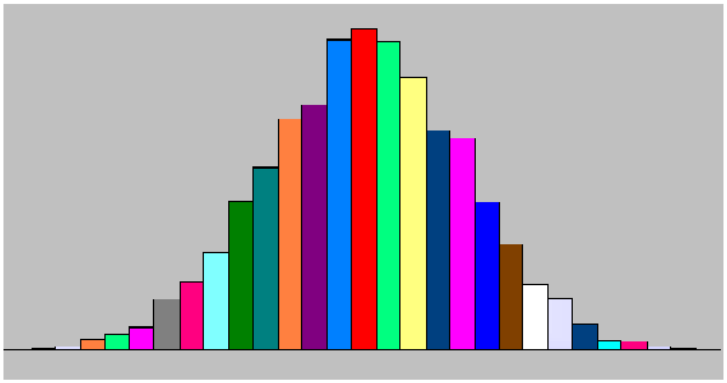
Two or more distributions may have identical measures of central tendency and skewness, but, they may show different degrees of peakedness. A distribution that is more picked is called a *leptokurtic* distribution. In such distribution, most of the values are concentrated to the mode and hence, the variance is the smaller.



If a distribution is more flat toped, it is called *platykurtic* in which the observations are scattered away from the modal value (the variance is now larger).



A distribution that is neither more peaked nor flat topped is called *mesokurtic*.



The moment measure of kurtosis is denoted by β and given by:

$$\beta = \frac{\mu_4}{(\mu_2)^2} \text{ where } \mu_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^r;\ r = 2, 4.$$

If $\beta = 3$, the distribution is *mesokurtic*. If $\beta > 3$, the distribution is *leptokurtic* and if $\beta < 3$, the distribution is *platykurtic*.

Chapter 5

Elementary Probability

5.1 Deterministic and Nondeterministic Models

Upon attempting to tackle real life problems, mathematical models can be viewed as *deterministic* and *nondeterministic* models.

- **Deterministic Model**: A deterministic model is a model which determines the *exact outcome* of the activity, that is, it explains a condition with no variability. For example, to know the BMI of an individual from his/her weight (in kg) W and height (in meter) H, one can determine the BMI by BMI $= \frac{W}{H^2}$.
- **Nondeterministic Model**: A nondeterministic model is a model in which the conditions of experimentation determine only the *random behavior* of the outcome. For example, the number of defective light bulbs in a factory cannot be determined with certainty. This model is also called *probabilistic* or *stochastic* model.

5.2 The Concept of Set Theory

As a general concept, *probability* is a quantitative measure of uncertainty on a scale of 0 (0%) to 1 (100%). It measures of the chance that something will occur.

In order to discuss the detail theory of probability, it is essential to be familiar with some ideas and concepts of set theory.

5.2.1 Definition of Set

A *set* is a collection of well-defined objects and denoted by capital letters like A, B, C, etc. If set A consists of n objects a_1, a_2, $\cdots$, a_n; it is written as $A = \{a_1, a_2, \cdots, a_n\}$. Each object '$a_i$; $i = 1, 2, \cdots, n$' is called an *element* of set A and written as $a_i \in A$. Hence, set A has n elements, $n(A) = n$.

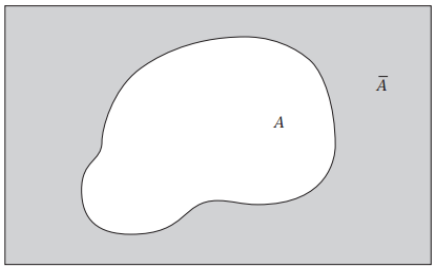
The number of elements in a set may be finite (the set is called a finite set) or infinite (the set is called an infinite set). For example, a set consisting the numbers, say, 1, 2, 3 and 4; is written as $B = \{1, 2, 3, 4\}$ and it is finite. Similarly, a set consisting of the numbers between

0 and 1 is written as $C = \{x : 0 \leq x \leq 1\}$ and it is infinite.

A set consisting all possible elements under consideration is called a *universal* set (denoted by U) and let $n(U) = N$ for a finite number of elements. On the other hand, a set containing no element is called an *empty* set (denoted by $\emptyset$ or $\{\}$). Here, $n(\emptyset)$=0.

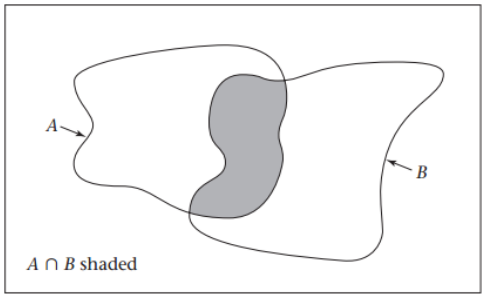
Now let us see the three basic and common set operations.

1. **Complement Set** (A'): The *complement* of a set A, denoted by A', is a set consisting all elements of U that are not in A, i.e., $A' = \{x : x \notin A\}$.

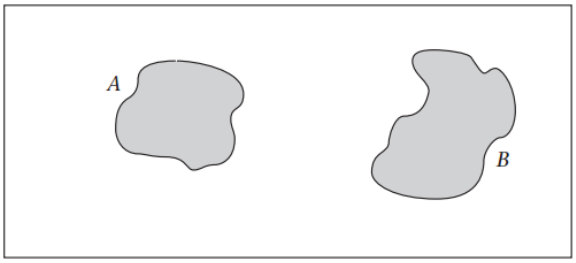


Note: $\emptyset' = U$ and $U' = \emptyset$.

2. **Intersection Set** ($A \cap B$): A set consisting all common elements from two sets A and B is called the *intersection* set or *product* of A and B, and written as $A \cap B$. That is, $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

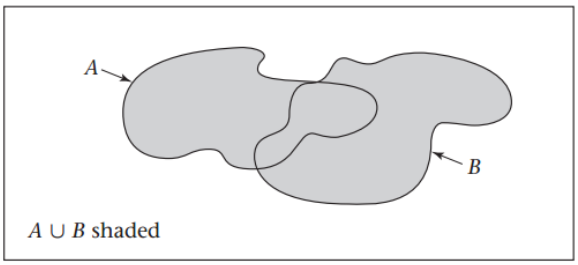


Sets having no element in common are called *mutually exclusive* or *disjoint* sets. If A and B are mutually exclusive, $A \cap B = \emptyset$ or $n(A \cap B) = 0$.



Note: A and A' are mutually exclusive sets ($A \cap A' = \emptyset$), and also any set A and $\emptyset$ ($A \cap \emptyset = \emptyset$).

3. **Union Set** $(A \cup B)$: A set consisting all elements from A or from B or from both is called the *union* set or *sum* of A and B, and written as $A \cup B$. That is, $A \cup B = \{x : x \in A,\ x \in B \text{ or } x \in A \cap B\}$.



Note: $A \cup A' = U$.

Properties of sets

- 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: $A \cup A = A$, $A \cup U = U$, $A \cup \emptyset = A$, $A \cap A = A$, $A \cap U = A$, $A \cap \emptyset = \emptyset$.
- De-Morgan's laws: $(A \cup B)' = A' \cap B'$, $(A \cap B)' = A' \cup B'$.

Example 5.1. Let $U = \{a, b, c, d, e, 1, 2, 3\}$. Let $A = \{a, d, e\}$, $B = \{d, e, 2, 3\}$ and $C = \{a, d, c, e, 3\}$. Find $A \cup B$, $A \cap B$, $A \cap B'$, $A' \cap B$, $(A \cup B)'$, $(A \cap B)'$, $A \cap (B \cup C)$, $A \cup (B \cap C)$.

Example 5.2. Consider the universal set $U = \{x : x \geq 0\}$ and three sets $A = \{x : x \leq 100\}$, $B = \{x : 50 \leq x \leq 200\}$ and $C = \{x : x \geq 150\}$. Find i) $A \cup B$ ii) $A \cap B$ iii) $B \cap C$ iv) $(A \cap B)'$

Let A and B be finite sets. Then, $n(A \cup B) = n(A) + n(B) - n(A \cap B)$. Also, $n(A \cup B) = n(A) + n(B)$ if A and B are mutually exclusive.

Example 5.3. In a survey conducted among 200 statistics major students, the number of students who visited historical, religious and both sites are found to be 150, 130 and 80 respectively. Find the number of students who visited none of the sites.

5.2.2 The Subset of a Set

If every element of set A is also an element of set B, A is said to be the *subset* of B and is written as $A \subset B$. If $A \subset B$, then $n(A) \leq n(B)$.

- Every set is a subset of itself, i.e., $A \subset A$.
- Every set is a subset of a universal set, i.e., $A \subset U$.
- Empty set is a subset of every set, i.e., $\emptyset \subset A$.
- Empty set is a subset of a universal set, i.e., $\emptyset \subset U$.

Note $A \subset B$ if and only if $B' \subset A'$. Then, $A \cup B = B$ and $A \cap B = A$. If $A \subset B$ and $B \subset A$, then A and B are said to be *equal*. If $A \subset B$ and $B \subset C$, then $A \subset C$.

5.3 Basic Probability Terms

1. **Experiment**: A probability *experiment* is an activity or a trial that leads to well-defined results. But, it is uncertain to which result will occur. Hence, a probability experiment is identified by two properties. First, each experiment has several (at least two) possible results and all these results are known in advance and second, none of these results can be predicted with certainty. For example, for the experiment of tossing a coin, there are two possible results: head and tail. But, we cannot be certain whether the result will be a head.

2. **Outcome**: An *outcome* is a result of a single trial (experiment). For the experiment of tossing a coin, 'head' is one possible outcome and 'tail' is also another possible outcome.

3. **Sample Space** (S): *Sample space* is a collection of all possible outcomes of an experiment. (In this context, S represents the universal set U described previously.)

   Example 5.4. Describe all the outcomes for the following probability experiments and also determine the total number of outcomes in the sample space.

   (a) Tossing a coin: S={Head (H), Tail (T)}, $n(S) = 2$.
   (b) Testing a patient for HIV: S={Positive, Negative}, $n(S) = 2$.
   (c) Playing a football game: S={Win, Lose, Tie}, $n(S) = 3$.
   (d) Tossing two coins: $S = \{HH, HT, TH, TT\}$, $n(S) = 4$.
   (e) Rolling a die: S={1, 2, 3, 4, 5, 6}, $n(S) = 6$.
   (f) Selecting an item from a production lot: S={Defective, Non-defective}, $n(S) = 2$.
   (g) Introducing a new product into a market: S={Success, Failure}, $n(S) = 2$.

4. **Event** (E): An *event* is an outcome or a set of outcomes (of interest) of an experiment. For example in the experiment of tossing two coins simultaneously if the event E is defined as getting one head, then $E = \{HT, TH\}$ and $n(E) = 2$. Similarly, $E = \{2, 4, 6\}$ is an event with $n(E) = 3$ defined for getting an even number in the experiment of rolling a die.

Note:

- Since $E \subset S$ $\{n(E) \leq n(S)\}$, it follows that S and $\emptyset$ are also events.
- S is called *certain* (*sure*) event because every outcome is an element of S.
- The event $\emptyset$ is an *impossible* event because no outcome is an element of $\emptyset$.
- If $n(S) = N$ and $n(E) = n$, then $0 \leq n \leq N$.

5.4 Counting Techniques

Counting techniques are used to fix the size of a sample space that is extremely large. In addition, they are used to determine the number of possible ways of arranging or selecting different objects.

1. **Addition Rule**: Suppose there are k procedures $(p_1, p_2, \cdots, p_k)$, in which the i^{th} procedure can be done in n_i; $i = 1, 2, \cdots, k$ ways. Hence, the total number of ways of performing p_1 or p_2 or $\cdots$ or p_k is $n_1 + n_2 + \cdots + n_k$, provided that no two procedures can be performed at the same time or one after the other.

   Example 5.5. There are 2 bus and 3 train routes from city X to city Z. In how many ways can a person go from city X to city Z? Ans: $2 + 3 = 5$ ways

2. **Multiplication Rule**: Suppose there are a sequence of k procedures, in which the i^{th} procedure has n_i; $i = 1, 2, \cdots, k$ possibilities. Then, the total number of possibilities of the whole sequence is $n_1 \times n_2 \times \cdots \times n_k$.

   Example 5.6. 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? Ans: 2×3=6 ways

   Example 5.7. Consider the following examples.

   (a) There are 3 questions. Each question has 2 choices. How many answer keys must be made? Ans: $2 \times 2 \times 2 = 2^3 = 8$

   (b) There are 5 patients in a clinic. If 4 doctors examine a different patient, in how many ways can this be done? Ans: $5 \times 4 \times 3 \times 2 = 120$

   (c) In how many ways can 6 persons be sit in a row? Ans: $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

   (d) Seven dice are rolled. How many different outcomes are there? Ans: $6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 = 6^7 = 279986$

3. **Permutation**: *Permutation* is the arrangement or selection of objects in a *specific order*.

   (a) **Permutation Rule 1**: The number of permutations of n *distinct* objects taking all together is $n! = n \times (n-1) \times (n-2) \times \cdots \times (1)$. By definition, $1! = 0! = 1$.

   Example 5.8. In how many ways can 6 persons be sit in a row? Ans: $6! = 720$

   Example 5.9. Suppose a photographer must arrange 4 persons in a row for a photograph. In how many different ways can the arrangement be done? Ans: $4! = 24$

   (b) **Permutation Rule 2**: The arrangement of n *distinct* objects in a specific order using r objects at a time is called a *permutation of n objects taking r objects at a time* and written as nP_r where

$$nP_r = \frac{n!}{(n-r)!},\ 0 < r \leq n.$$

   Example 5.10. In how many ways can 9 books be arranged on a shelf having 4 places? Ans: $9P_4 = \frac{9!}{(9-4)!} = 9 \times 8 \times 7 \times 6 = 3024$

   Example 5.11. How many 5 letter permutations can be formed from the letters in the word 'DISCOVER'? Ans: $8P_5 = \frac{8!}{(8-5)!} = 8 \times 7 \times 6 \times 5 \times 4 = 6720$

(c) **Permutation Rule 3**: The number of permutations of n objects in which n_1 are alike, n_2 are alike, $\cdots$, n_r are alike is given by

$$\frac{n!}{n_1! \times n_2! \times \cdots \times n_r!} \text{ where } n_1 + n_2 + \cdots + n_r = n.$$

Example 5.12. How many different permutations can be made from the letters in the word

a. STATISTICS. Ans: $\frac{10!}{3!\times 3!\times 1!\times 2!\times 1!}$
b. MISSISSIPPI. Ans: $\frac{11!}{1!\times 4!\times 4!\times 2!}$
c. EXERCISES. Ans: $\frac{9!}{3!\times 1!\times 1!\times 1!\times 1!\times 2}$

4. **Combination**: *Combination* is the arrangement or selection of objects *without regard to order*. Here, order does not matter.

- The number of *combinations of n objects taking r objects at a time* is denoted by $nC_r = \binom{n}{r}$ where

$$nC_r = \binom{n}{r} = \frac{n!}{(n-r)! \times r!};\ 0 < r \leq n.$$

Exercise 5.1. In how many different ways can a secretary, a president and a manager be selected from 5 persons?

Exercise 5.2. A committee of 3 persons is to be selected from 5 persons. In how many different ways can this be done?

Exercise 5.3. There are 12 new GP applicants for a certain vacant position in a hospital. How many different ways can the selection be done if there

1. is only one vacant position.
2. are two vacant positions.
3. are three vacant positions.
4. are ten vacant positions.

Exercise 5.4. A committee of 5 persons must be selected from 5 men and 8 women. How many ways can the selection be done if there are

1. exactly 4 men in the committee?
2. at least 3 women in the committee?

5.5 Definitions of Probability

There are three common approaches for the definitions of probability: classical/mathematical, empirical/frequentist and subjective/axiomatic approaches.

5.5.1 Classical Probability

The classical definition uses sample space to determine the probability of an event. Suppose there are N possible outcomes in the sample space S of an experiment. Of these, suppose, only n are favorable to the event E, then the probability that event E will occur is:

$$P(E) = \frac{n(E)}{n(S)} = \frac{n}{N}.$$

Example 5.13. What is the probability of getting number 6 in rolling a die?

Solution: The sample space for the experiment of rolling a die is $S = \{1, 2, 3, 4, 5, 6\}$. Hence, $n(S) = 6$. Let $E =$ getting number 6 in rolling a die. Thus, $E = \{6\}$ and $n(E) = 1$. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{1}{6}$.

Example 5.14. What is the probability of getting one head in tossing two coins?

Solution: The sample space for the experiment of two coins is $S = \{HH, HT, TH, TT\}$. Hence, $n(S) = 4$. Let $E =$ getting one head in tossing two coins. Thus, $E = \{HT, TH\}$ and $n(E) = 2$. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{2}{4} = 0.5$.

Example 5.15. A die is rolled. What is the probability of getting

1. an odd number?
2. a number greater than 4?

Solution: The sample space for the experiment of rolling a die is $S = \{1, 2, 3, 4, 5, 6\}$. Hence, $n(S) = 6$.

1. Let $E =$ getting an odd number in rolling a die. Thus, $E = \{1, 3, 5\}$ and $n(E) = 3$. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{3}{6} = 0.5$.
2. Let $E =$ getting a number > 4 in rolling a die. Thus, $E = \{5, 6\}$ and $n(E) = 2$. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{2}{6}$.

Example 5.16. An urn contains 7 white and 3 black balls.

1. If one ball is selected, what is the probability that the selected ball is black?
2. If two balls are selected, what is the probability that both balls are black?

Solution: There are 7W+3B=10 balls.

1. Let $E =$ selecting a black ball from 10 balls. Thus, $n(E) = 3$. For selecting one ball, there are a total of $n(S) = 10$ possibilities. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{3}{10}$.
2. Let $E =$ selecting two black balls from 10 balls. Thus, $n(E) = 3C_2$. For selecting two balls, there are a total of $n(S) = 10C_2$ possibilities. Therefore, $P(E) = \frac{n(E)}{n(S)} = \frac{3C_2}{10C_2}$.

Example 5.17. An urn contains 6 white, 4 red and 9 black balls. If three balls are drawn at random, what is the probability that:

1. 1 is of each color.

2. 2 of the balls drawn are white.
3. none is red.
4. at least one is white.

Solution: The number of ways of selection of 3 balls of the total 19 is $n(S) = 19C_3$.

1. Let E = selecting 1 ball of each color. Thus, $n(E) = 6C_1 4C_1 9C_1$. Hence,
$$P(E) = \frac{6C_1 4C_1 9C_1}{19C_3}.$$
2. Let E = selecting 2 white balls. Thus, $n(E) = 6C_2 13C_1$. Hence,
$$P(E) = \frac{6C_2 13C_1}{19C_3}.$$
3. Let E = selecting no red ball. Thus, $n(E) = 4C_0 15C_3$. Hence,
$$P(E) = \frac{4C_0 15C_3}{19C_3}.$$
4. Let E = selecting at least one white ball. Thus, $n(E) = 6C_1 13C_2 + 6C_2 13C_1 + 6C_3 13C_0$. Hence,
$$P(E) = \frac{6C_1 13C_2 + 6C_2 13C_1 + 6C_3 13C_0}{19C_3}.$$

Example 5.18. A committee of 5 persons must be selected from 5 men and 8 women. What is the probability that the committee consists of at least 3 women?

Solution: The number of ways of selecting 5 persons for the committee of the total 13 persons is $n(S) = 13C_5$. Let E be the event that the committee consists of at least three women. Thus, $n(E) = 8C_3 5C_2 + 8C_4 5C_1 + 8C_5 5C_0$. Therefore,
$$P(E) = \frac{8C_3 5C_2 + 8C_4 5C_1 + 8C_5 5C_0}{13C_5}.$$

Exercise 5.5. A family plans to have three children. Describe the sample space for all possible gender combinations. What is the probability that the family will have two boys?

Exercise 5.6. Two dice are rolled. Describe the sample space. What is the probability of getting i) a sum of 10 or more, ii) a pair which at least one number is 3, iii) a sum of 8, 9 or 10, iv) one number less than 4.

Note: The classical definition of probability is appropriate when all outcomes of an experiment are equally likely. If there are N outcomes in the sample space S, then the probability of each outcome is $\frac{1}{N}$.

5.5.2 Empirical Probability

The empirical probability is calculated based on a relative frequency. Given a frequency distribution, the probability of an event being in a given class is:

$$P(E) = \frac{f}{n}$$

where f is the class frequency and $n = \sum_{i=1}^{k} f_i$ is the total number of observations.

Example 5.19. Consider a study of waiting times in the X-ray department for a certain hospital. A clerk recoded the number of patients waiting for service at 9:00 A.M. on 20 successive days and obtained the following discrete frequency distribution.

Number of patients waiting	0	1	2	3	4
Number of days	2	5	6	4	3

The frequency distribution shows that on 2 of the 20 days, 0 patients were waiting for service; on 5 of the days, one patient was waiting for service; so on. Using a relative frequency, the probability of no patient waiting for service is $\frac{2}{20} = 0.10$, the probability of one patient waiting for service is $\frac{5}{20} = 0.25$, the probability of two patients waiting for service is $\frac{6}{20} = 0.30$, the probability of three patients waiting for service is $\frac{4}{20} = 0.20$ and the probability of four patients waiting for service is $\frac{3}{20} = 0.15$.

5.5.3 Subjective Probability

A subjective approach calculates probability based on an educated guess, experience or evaluation of a problem. It expresses a person's *degree of belief* for the occurrence of an event. For example, a physician might say that on the basis of his/her diagnosis, there is a 30% chance a patient will need an operation. When a patient presents with chest pains, a clinician may say that the probability that the patient has heart disease is about 20%. Also, an epidemiologist might say there is an 80% probability that an outbreak will occur in certain area.

A subjective probability is personal. Different people can be expected to assign different probabilities for the same event of interest.

5.6 Probabilistic Rules and Notations

The probability of an event always lies in between 0 and 1, that is, $0 \leq P(E) \leq 1$. If $P(E) = 0$, then it is sure that E can never happen. On the other hand, if $P(E) = 1$, the event E is certain to occur.

Example 5.20. In the experiment of rolling a die, the probability of getting number 9 is 0, and the probability of getting a number less than 7 is 1. How?

The sum of the probabilities of each outcome, $s_i; i = 1, 2, \cdots, N$, in a sample space S is 1, that is, $\sum_{i=1}^{N} p(s_i) = 1$. For example, there are six outcomes in the experiment of rolling a die, each with probability $\frac{1}{6}$.

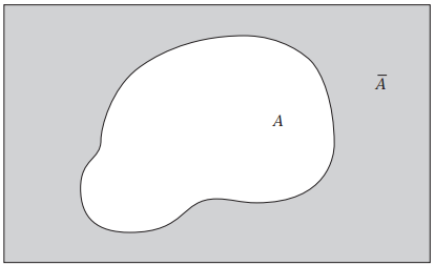
Outcome (s_i)	1	2	3	4	5	6
Probability $\{P(s_i)\}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Therefore, $\sum_{i=1}^{6} p(s_i) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{6}{6} = 1$.

Example 5.21. Suppose that only three outcomes are possible in an experiment: a_1, a_2 and a_3. Suppose further more that a_1 is twice as likely to occur as a_2 which is four times as likely to occur as a_3. Find p_1, p_2 and p_3.

Complement of an Event

The *complement* event of A consists of all outcomes that are not in A and it is denoted by A'. It is the event that A does not occur.



If there are n outcomes in favor of an event A of the total N outcomes in S, then there will be $N - n$ outcomes against the event A (in favor of the complement event A'). Thus, the probability of the complement event A' is:

$$P(A') = \frac{n(A')}{n(S)} = \frac{N-n}{N} = 1 - P(A).$$

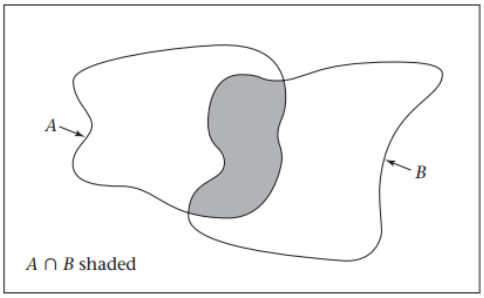
Event, A', occurs only when A does not occur. Therefore, $P(A) + P(A') = 1$.

Example 5.22. The probability of a patient being HIV positive is 0.02. What is the probability of being negative?

Solution: $P(\text{HIV}^+) + P(\text{HIV}^-) = 1$. Thus, $P(\text{HIV}^-) = 1 - P(\text{HIV}^+) = 1 - 0.02 = 0.98$.

Intersection of Two Events

The *intersection* of events A and B ($A \cap B$) is an event containing outcomes belonging to both A and B.



Therefore, $P(A \cap B)$ represents the probability that both events will occur at the same time, and is calculated as:

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)}.$$

Example 5.23. In an undergraduate class consisting of 30 girls and 20 boys, it is observed that 4 girls and 6 boys wear eyeglasses because of vision problems. If a student is selected at random, what is the probability that the student is

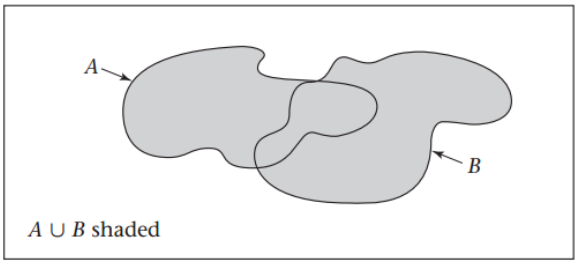
1. a girl and has a vision problem?
2. a boy and has a vision problem?

Solution:

1. Let G be a girl, V be a student with a vision problem. Hence, $n(G \cap V) = 4$ (4 students are girls and have vision problem). Total number of students in the class is $n(S) = 50$. Thus, $P(G \cap V) = \frac{n(G \cap V)}{n(S)} = \frac{4}{50} = 0.08$ (8% of the students are girls and have vision problem).
2. Let B be a boy, V be a student with a vision problem. Hence, $n(B \cap V) = 6$ (6 students are boys and have vision problem). Total number of students in the class is $n(S) = 50$. Thus, $P(E) = \frac{n(B \cap V)}{n(S)} = \frac{6}{50} = 0.12$ (12% of the students are boys and have vision problem).

Union of Two Events

If there are two events A and B, the *union* set of A and B $(A \cup B)$ is an event containing all outcomes from A or from B or from both.



Therefore, $P(A \cup B)$ is the probability that at least one of the two events (either A or B or both) will occur and is the sum of the probability that each event will occur minus the probability that both events will occur at the same time. That is, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Recall that $n(A \cup B) = n(A) + n(B) - n(A \cap B)$. Thus, dividing both sides of this equality by $n(S)$ provides the probability.

Example 5.24. A patient is suspected to have two diseases: TB and HIV. The probability that the patient will have TB is 0.60 and the probability of HIV is 0.70. The probability that the patient will have both diseases is 0.50. Find the probability that the patient will have at least one of the two diseases.

Solution: $P(\text{TB} \cup \text{HIV}) = P(\text{TB}) + P(\text{HIV}) - P(\text{TB} \cap \text{HIV}) = 0.60 + 0.70 - 0.50 = 0.80$. Thus, 80% of the patients have at least one of the diseases (TB or HIV).

Example 5.25. Recall example 5.23. If one student is selected randomly, what is the probability that the student is a girl or has a vision problem.

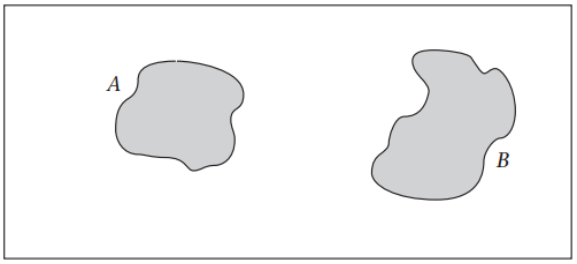
Solution: Let G be a girl student, V be a student with vision problem. Then, the required is $P(G \cup V) =?$. $P(G \cup V) = P(G) + P(V) - P(G \cap V) = \frac{30}{50} + \frac{10}{50} - \frac{4}{50} = \frac{36}{50} = 0.72$. The probability that a student is a girl or has a vision problem is 72%.

Note: Consider events A and B. Let s be an outcome.

- If $s \in A$, then event A occurs. But, if $s \notin A$, then event A does not occur.
- If $s \in A \cap B$, then $A \cap B$ represents the event that both A and B occur.
- If $s \in A \cup B$, then $A \cup B$ represents the event that A occurs, or B occurs or both occur.
- If $s \in A \cap B'$, then $A \cap B'$ represents the event that A occurs but B does not.
- If $s \in A' \cap B$, then $A' \cap B$ represents the event that B occurs but A does not.
- If $s \in A' \cap B' = (A \cup B)'$, then $A' \cap B' = (A \cup B)'$ represents the event neither A nor B occurs.
- If $s \in (A \cap B') \cup (A' \cap B)$, then $(A \cap B') \cup (A' \cap B)$ represents exactly only one of the two events occurs.
- If $s \in (A \cap B)' = A' \cup B'$, then $(A \cap B)' = A' \cup B'$ represents the event that both events do not occur.

Mutually Exclusive Events

Two events are said to be *mutually exclusive* if the events have no outcome in common. That is, if A and B are mutually exclusive, then both events cannot occur simultaneously which means the occurrence of one stops the occurrence of the other.



For example, in the experiment of rolling a die, odd numbers and even numbers cannot occur at the same time. Hence, odd and even numbers are mutually exclusive events.

Tossing a coin also cannot result both a head and a tail simultaneously. Thus, head and tail are also mutually exclusive outcomes.

In addition, weight of an individual cannot be classified simultaneously as "underweight", "normal" and "overweight".
If event A and B are *mutually exclusive*, then $A \cap B = \emptyset$. This implies $P(A \cap B) = 0$ and $P(A \cup B) = P(A) + P(B)$.

5.7 Marginal and Joint Probabilities

Consider again example 5.23. Let G = event a student is a girl, B = event a student is a boy, V = event a student has a vision problem and V' = event a student has not a vision problem. Then, the data can be presented in a 2×2 table as:

Gender	Vision		Total
	V	V'	
Girls (G)	$n(G \cap V) = 4$	$n(G \cap V') = 26$	$n(G) = 30$
Boys (B)	$n(B \cap V) = 6$	$n(B \cap V') = 14$	$n(B) = 20$
Total	$n(V) = 10$	$n(V') = 40$	$n(S) = 50$

Dividing the frequencies by the total number of students, $n(S) = 50$, enables us to determine the following probabilities:

Gender	Vision		Total
	V	V'	
Girls (G)	$P(G \cap V)$=0.08	$P(G \cap V')$=0.52	$P(G)$=0.60
Boys (B)	$P(B \cap V)$=0.12	$P(B \cap V')$=0.28	$P(B)$=0.40
Total	$P(V)$=0.20	$P(V')$=0.80	1.00

For gender, girls G and boys B:

- $P(G) = \frac{30}{50} = 0.60$ is the probability that a randomly selected student is a girl (60% of the students are girls).
- $P(B) = \frac{20}{50} = 0.40$ is the probability that a randomly selected student is a boy (40% of the students are boys).

For vision problem, vision problem V and no vision problem V':

- $P(V) = \frac{10}{50} = 0.20$ is the probability that a randomly selected student has a vision problem (20% of the students have vision problem).
- $P(V') = \frac{40}{50} = 0.80$ is the probability that a randomly selected student has no vision problem (80% of the students have no vision problem).

These values provide probabilities for gender and vision problem separately, and are called *marginal* probabilities.

Also,

- $P(G \cap V) = \frac{4}{50} = 0.08$ is the probability that a randomly selected student is a girl and has a vision problem.
- $P(G \cap V') = \frac{26}{50} = 0.52$ is the probability that a randomly selected student is a girl and has no vision problem.
- $P(B \cap V) = \frac{6}{50} = 0.12$ is the probability that a randomly selected student is a boy and has a vision problem.
- $P(B \cap V') = \frac{14}{50} = 0.28$ is the probability that a randomly selected student is a boy and has no vision problem.

Because each of these values gives the probability of the intersection of two events, these probabilities are called *joint* probabilities.

5.8 Conditional Probability

5.8.1 Conditional Events

When the occurrence of an event affects the probability of occurrence of another event, the two events are said to be *conditional* (*dependent*) events. If the events A and B are conditional to each other, then the probability of event A occurring after event B has occurred is said to be the *conditional probability of A given B*, and is written as

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B) > 0.$$

This implies, the probability that both events will occur is $P(A \cap B) = P(B)P(A|B)$.

Similarly, the probability of event B occurring knowing that event A has already occurred is said to be the *conditional probability of B given A*, and is written as

$$P(B|A) = \frac{P(A \cap B)}{P(A)},\ P(A) > 0.$$

This implies, the probability that both events will occur is $P(A \cap B) = P(A)P(B|A)$.

Therefore, $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$. If A and B are mutually exclusive, then $P(A|B) = P(B|A) = 0$.

Example 5.26. Recall example 5.24. Find $P(\text{TB}|\text{HIV})$ and $P(\text{HIV}|\text{TB})$.

Solution: Here $P(\text{TB}|\text{HIV}) = \frac{0.50}{0.70} = 0.714$. This means, of those persons who have HIV, about 71% will have TB. And, $P(\text{HIV}|\text{TB}) = \frac{0.50}{0.60} = 0.833$. Thus, about 83% of those persons who have TB will have also HIV.

Example 5.27. If the probability of a patient with chest pains having heart disease is 20%, then what is the probability of a patient with chest pains not having heart disease?

Solution: Let $P(\text{HD}|\text{CP}) = 20$. Then, $P(\text{HD}'|\text{CP}) = 1 - P(\text{HD}|\text{CP}) = 1 - 0.20 = 0.80$

Example 5.28. Recall the 2×2 table constructed from data given in example 5.23.

Gender	Vision		Total
	V	V'	
Girls (G)	$n(G \cap V) = 4$	$n(G \cap V') = 26$	$n(G) = 30$
Boys (B)	$n(B \cap V) = 6$	$n(B \cap V') = 14$	$n(B) = 20$
Total	$n(V) = 10$	$n(V') = 40$	$n(S) = 50$

Dividing the frequencies by the marginal total of girls and boys enables us to determine two probabilities (one for girls and the other for boys). Because each of these values gives the probability of vision problem given the gender of a student, the probabilities are called *conditional* probabilities.

Gender	Vision		Total
	V	V'	
Girls (G)	$P(V\|G)$=0.13	$P(V'\|G)$=0.87	1.00
Boys (B)	$P(V\|B)$=0.30	$P(V'\|B)$=0.70	1.00

For girls:

- $P(V|G) = \frac{4}{30} = 0.13$ is the probability that a randomly selected girl has a vision problem. Or it is the (conditional) probability of having a vision problem if the student is a girl (13% of the girls have vision problem).
- $P(V'|G) = \frac{26}{30} = 0.87$ is the probability that a randomly selected girl has no vision problem. Or it is the (conditional) probability of not having a vision problem if it is known that the student is a girl (87% of the girls do not have vision problem).

For boys:

- $P(V|B) = \frac{6}{20} = 0.30$ is the probability that a randomly selected boy has a vision problem. Or it is the (conditional) probability of having a vision problem if the student is a boy (30% of the boys have vision problem).
- $P(V'|B) = \frac{14}{20} = 0.70$ is the probability that a randomly selected boy has no vision problem. Or it is the (conditional) probability of not having a vision problem if the student is a boy (70% of the boys do not have a vision problem).

5.8.2 Total Probability Theorem

Suppose, a sample space is partitioned into k parts. The events B_1, B_2, $\cdots$, B_k represent partitions of a sample space S if

- $B_1 \cup B_2 \cup \cdots \cup B_k = S$,
- $B_i \cap B_j = \emptyset, \forall i \neq j = 1, 2, \cdots, k$ and
- $P(B_i) > 0, \forall i$.

Let E be some event with respect to S and let B_1, B_2, $\cdots$, B_k be partitions of S. Hence, E may be written as $E = (E \cap B_1) \cup (E \cap B_2) \cup \cdots \cup (E \cap B_k)$. This implies, the unconditional probability of E as $P(E) = P(E \cap B_1) + P(E \cap B_2) + \cdots + P(E \cap B_k)$. Therefore,

$$\begin{aligned} P(E) &= P(B_1)P(E|B_1) + P(B_2)P(E|B_2) + \cdots + P(B_k)P(E|B_k) \\ &= \sum_{i=1}^{k} P(B_i)P(E|B_i). \end{aligned}$$

This is called *total probability* theorem.

Example 5.29. In a study of smoking and lung cancer, three groups of individuals (non-smokers, former smokers and current smokers) are involved. The percentage of non-smokers, former smokers and current smokers are 50%, 30% and 20%, respectively. Also, the percentage of lung cancer in each group is 3%, 4% and 5%, respectively. If one of the study participants becomes sick, what is the probability that s/he has the disease?

Solution: Let E be a person with lung cancer (an event consisting of lung cancer patients). Let B_1 be an event consisting of non-smokers, B_2 be an event consisting of former smokers and B_3 be a set consisting of current smokers.

The required probability is $P(E)$:

$$\begin{aligned} P(E) &= P(E \cap B_1 \cup E \cap B_2 \cap E \cap B_3) \\ &= P(E \cap B_1) + P(E \cap B_2) + P(E \cap B_3) \\ &= P(B_1)P(E|B_1) + P(B_2)P(E|B_2) + P(B_3)P(E|B_3) \\ &= 0.50(0.03) + 0.30(0.04) + 0.20(0.05) \\ &= 0.037 \end{aligned}$$

Therefore, the (unconditional) probability that the person, who was sick, has lung cancer is 3.7%.

Example 5.30. Among 100000 women with negative mammograms 20 will be diagnosed with breast cancer within 2 years, whereas 1 woman in 10 with positive mammograms will be diagnosed with breast cancer within 2 years. Suppose that 7% of the general population of women will have a positive mammogram. What is the probability of developing breast cancer over the next 2 years among women in the general population?

Solution: Let E be an event of breast cancer, $P(E) =?$. Let M be a positive mammogram and hence, M' be a negative mammogram. Thus, $P(M) = 0.07$ and $P(M') = 1 - P(M) = 0.93$. Also, $P(E|M) = 0.10$ and $P(E|M') = 0.0002$.

Now the required probability is $P(E)$:

$$\begin{aligned} P(E) &= P(E \cap M) + P(E \cap M') \\ &= P(M)P(E|M) + P(M')P(E|M') \\ &= 0.07(0.10) + 0.93(0.0002) \\ &= 0.0072 \end{aligned}$$

Therefore, the (unconditional) probability of developing breast cancer over the next 2 years among women in the general population is 0.0072.

Exercise 5.7. A 5-year study of cataract in a population of 5000 people 60 years of age and older is planned. It is known from census data that 45% of this population is 60-64 years of age, 28% are 65-69 years of age, 20% are 70-74 years of age, and 7% are 75 or older. It is also known that 2.4%, 4.6%, 8.8%, and 15.3% of the people in these respective age groups will develop cataract over the next 5 years. What percentage of the population in the study will develop cataract over the next 5 years, and how many people with cataract does this percentage represent?

Solution: 5.2% of the population will develop cataract over the next 5 years, which represents a total of $5000 \times 0.052 = 260$ people with cataract.

5.8.3 Bayes' Theorem

Let B_1, B_2, $\cdots$, B_k be partitions of a sample space S and let E be an event associated with S. Then,

$$\begin{aligned} P(B_i|E) &= \frac{P(B_i \cap E)}{P(E)} \\ &= \frac{P(B_i)P(E|B_i)}{\sum_{i=1}^{k} P(B_i)P(E|B_i)};\ i = 1, 2, \cdots, k. \end{aligned}$$

This is called *Bayes'* theorem.

Example 5.31. Recall example 5.29 about the distribution of lung cancer patients among non-smokers, former smokers and current smokers. If the sick person is found to have the disease, what is the probability that s/he is a current smoker?

Solution: Here, the required probability is $P(B_3|E)$:

$$P(B_3|E) = \frac{P(B_3)P(E|B_3)}{P(E)} = \frac{0.20(0.05)}{0.037} = 0.27$$

Therefore, the probability that the sick person is a current smoker if s/he has lung cancer is 27%.

Example 5.32. Consider example 5.30. If a woman is diagnosed to have breast cancer, find the probability that she has a positive mammogram?

Solution: Now the required probability is $P(M|E)$:

$$P(M|E) = \frac{P(M)P(E|M)}{P(E)} = \frac{0.07(0.10)}{0.0072} = 0.9722$$

Thus, 97.22% of women have positive mammograms if they are diagnosed to have breast cancer.

Screening Tests

There are two possible types of errors in a screening test: False Positive and False Negative.

Test	Disease		Total
	D^+	D^-	
T^+	TP	FP	TP+FP
T^-	FN	TN	FN+TN
Total	TP+FN	FP+TN	

- The *sensitivity* of a screening test is the probability that the test is positive given that the person has a disease, $P(T^+|D^+) = \frac{TP}{TP+FN}$. A high sensitivity indicates there are a few false negative results.

- The *specificity* of a screening test is the probability that the test is negative given that the person does not have a disease, $P(T^-|D^-) = \frac{TN}{FP+TN}$. A high specificity indicates there are a few false positive results.

Example 5.33. Find the sensitivity and specificity of the mammography considered in example 5.30.

Solution: The sensitivity of mammography is $P(M|E) = \frac{P(M)P(E|M)}{P(E)} = \frac{0.07(0.10)}{0.0072} = 0.9722$. This means, 97.22% of women who have breast cancer will have positive mammograms. Also, the specificity of the mammography is $P(M'|E') = \frac{P(M'\cap E')}{P(E')} = \frac{P(M')P(E'|M')}{P(E')} = \frac{0.93(1-0.0002)}{1-0.0072} = 0.9366$. Of those women who do not have breast cancer, 93.66% will have negative mammograms.

- The *positive predictive value* of a screening test is the probability that a person has a disease given that the test is positive, $PV^+ = P(D^+|T^+) = \frac{TP}{TP+FP}$.
- The *negative predictive value* of a screening test is the probability that a person does not have a disease given that the test is negative, $PV^- = P(D^-|T^-) = \frac{TN}{FN+TN}$.

Example 5.34. Find the PV^+ and PV^- of mammography considered in example 5.30.

Solution: The predictive value of a positive mammogram is $PV^+ = P(E|M) = \frac{P(E\cap M)}{P(M)} = \frac{P(E)P(M|E)}{P(M)} = \frac{0.0072(0.9722)}{0.07} = 0.10$. Thus, if the mammogram is positive, a woman has a 10% chance of developing breast cancer over the next 2 years. The predictive value of a negative mammogram is $PV^- = P(E'|M') = 1 - P(E|M') = 1 - 0.0002 = 0.9998$. Now, if the mammogram is negative, the woman is virtually certain not to develop breast cancer over the next 2 years.

5.9 Independence

Two events are said to be *independent* if the occurrence of one does not affect the probability of the occurrence of the other. If event A and B are independent, the probability of A occurring is in no way affected by the occurrence of event B or vice versa, hence, $P(A\cap B) = P(A)P(B)$.

If the events A and B are independent, then $P(A|B) = P(A)$, $P(B) > 0$ and $P(B|A) = P(B)$, $P(A) > 0$.

Example 5.35. A coin is tossed and a die is rolled. What is the probability of getting a head on the coin or number 4 on the die?

Solution: Let A be a head on the coin, B be number 4 on the die. Thus, $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{6}$. Hence, $P(A\cup B) = P(A) + P(B) - P(A\cap B)$. But, $P(A\cap B) = P(A)P(B) = \frac{1}{12}$. Therefore, $P(A\cup B) = \frac{7}{12}$.

Example 5.36. Suppose two doctors, A and B, test all patients coming into a clinic for syphilis. Suppose doctor A diagnoses 15% of all patients as positive, doctor B diagnoses 20% of all patients as positive, and both doctors diagnose only 3% of all patients as positive. Are the two diagnosis results independent?

Solution: $P(A) = 0.15$, $P(B) = 0.20$ and $P(A \cap B) = 0.03$. Since $P(A) \cdot P(B) = 0.15(0.20) = 0.03$ equals to $P(A \cap B) = 0.03$, the two diagnosis results are independent.

Example 5.37. Suppose a patient is referred for further lab tests if either doctor A or B makes a positive diagnosis as given in example 5.36 above. What is the probability that a patient will be referred for further lab tests?

Solution: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.15 + 0.20 - 0.03 = 0.32$. Thus, 32% of all patients will be referred for further lab tests.

Exercise 5.8. Let A and B be two events associated with an experiment. Suppose that $P(A) = 0.4$, $P(A \cup B) = 0.7$ and $P(B) = p$.

1. For what choice of p, are A and B independent.
2. For what choice of p, are A and B mutually exclusive.

Chapter 6

Probability Distributions

6.1 Random Variable

Random variable is a variable whose values are determined by chance (with some probability). For example, number of females in a class of 30 students, number of patients coming to a certain hospital in a day, weight (in kg) of newly born babies can be considered as random variables.

Mostly, a random variable is denoted by capital letters, for example, Y and its value is denoted by the corresponding small letter; y_i. The set consisting of all possible values of a random variable is called *range space* (R_Y). For instance, the range space for the random variable Y=number of females in a class of 30 students is $R_Y = \{0, 1, 2, \cdots, 30\}$. Similarly, the range space for the random variable Y=number of patients coming to a certain hospital in a day is $R_Y = \{0, 1, 2, \cdots\}$. Also, the range space for Y=weight (in kg) of newly born babies can be written as $R_Y = \{y : 2.5 \leq y \leq 4.0\}$.

If the number of possible values of a random variable Y (that is, R_Y) is finite or countable infinite, the random variable is called *discrete* random variable. Hence, the possible values of a discrete random variable Y may be listed as $y_1, y_2, \cdots, y_n, \cdots$. In the finite case, the list terminates and in the countably infinite case the list continues indefinitely. As a result, the random variables 'number of females in a class of 30 students' and 'number of patients coming to a certain hospital in a day' are examples of discrete random variables with finite and countably infinite possible values, respectively.

On the other hand, if a random variable assumes an uncountable infinite number of possible values (that is, R_Y is infinite), the random variable is called *continuous* random variable. The random variable 'weight (in kg) of newly born babies' could take an uncountable infinite number of values between, say, 2.5kg and 4.0kg. Therefore, this random variable is an example of continuous random variable.

6.2 Probability Distribution

A *probability distribution* describes how probabilities are distributed over the values of a random variable. It consists of the values a random variable and their corresponding prob-

abilities. Based on the type of a random variable, a probability distribution can be *discrete* and *continuous*.

6.2.1 Discrete Probability Distribution

With each possible value y_i of a discrete random variable, a number $p(y_i) = P(Y = y_i)$, called *probability of* y_i, is associated. The numbers $P(Y = y_i) = p(y_i), i = 1, 2, \cdots$ must satisfy the following conditions:

i. $0 \leq P(Y = y_i) \leq 1;\ i = 1, 2, \cdots$

ii. $\sum_{i=1}^{\infty} P(Y = y_i) = 1$

The function p is called *probability mass function* (pmf) of a random variable Y.

y_i	y_1	y_2	$\cdots$	y_n	$\cdots$
$P(Y = y_i)$	$p(y_1)$	$p(y_2)$	$\cdots$	$p(y_n)$	$\cdots$

These collection of pairs $[y_i, p(y_i)], i = 1, 2, \cdots$ is also called *discrete* probability distribution of Y.

Example 6.1. Construct a probability distribution for the number of girls to be born if a family plans to have three children. Plot the probability distribution using bar chart.

Solution: First list all the possible values that Y can assume. Then, calculate the probability of each possible distinct value of Y and present it in the form of a discrete frequency distribution.

The sample space for the possible gender combinations of three children to be born is $S = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$. Let Y be the number of girls among three children. As a result, $R_Y = \{0, 1, 2, 3\}$.

y_i	0	1	2	3	Total
$p(y_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Example 6.2. A random variable Y assumes six values with probabilities shown in the following probability distribution:

y_i	-2	-1	0	1	2	3	Total
$p(y_i)$	0.05	0.35	0.15	0.20	0.15	0.10	1.00

Find $P(-1 \leq Y \leq 2)$, $P(-1 < Y < 2)$, $P(Y > 1)$ and $P(Y \leq 1)$.

Solution:

- $P(-1 \leq Y \leq 2) = P(Y = -1) + P(Y = 0) + P(Y = 1) + P(Y = 2) = 0.85$
- $P(-1 < Y < 2) = P(0 \leq Y \leq 1) = P(Y = 0) + P(Y = 1) = 0.35$
- $P(Y > 1) = P(Y \geq 2) = P(Y = 2) + P(Y = 3) = 0.25$

- $P(Y \leq 1) = P(Y = -2) + P(Y = -1) + P(Y = 0) + P(Y = 1) = 0.75$

Example 6.3. The instructor of a large class gives 15% each of A, 25% each of B, 35% each of C, 20% each of D and 5% each of F. If a student is chosen at random from this class, the student's grade is a random variable Y with 5 possible values: R_Y={A, B, C, D, F}.

1. Construct the probability distribution of Y?
2. Draw a bar chart for the probability distribution of Y.
3. What is the probability that the student got B or better?

6.2.2 Continuous Probability Distribution

As continuous random variables differ from discrete random variables, consequently continuous probability distributions differ from discrete ones. Since a continuous random variable assumes any value in an interval $[a, b]$, the graph of its corresponding continuous distribution (equivalent of a bar chart for a discrete distribution) is usually a smooth frequency curve described by a mathematical function, $f(y)$, where $a \leq y \leq b$. The function f is called *probability density function* (*pdf*) if it satisfies the following two conditions:

i. $f(y) \geq 0$ for $a \leq y \leq b$

ii. $\int_a^b f(y)dy = 1.$

Remarks:

- The value of $f(y)$ is not a probability at all; hence $f(y)$ can take any nonnegative value including values greater than 1.
- The total area under the *pdf* curve over the entire range of possible values of a continuous random variable is 1.

Example 6.4. Show that $f(y) = \begin{cases} 1, & 0 \leq y \leq 1; \\ 0, & \text{otherwise} \end{cases}$ is a *pdf*.

Example 6.5. Show that $f(y) = \begin{cases} 2y, & 0 \leq y \leq 1; \\ 0, & \text{otherwise} \end{cases}$ is a *pdf*.

Note:

- The *pdf* $f(y)$ of a continuous random variable does not give the probability $P(Y = y)$. This is because Y can take an infinite number of values and, therefore, it is not possible to assign a probability for each value y. Therefore, the probability corresponding to a single point is assumed to be zero, that is, $P(Y = y) = 0$ for all y. Consequently, $P(c \leq Y \leq d)$, $P(c < Y \leq d)$, $P(c \leq Y < d)$ and $P(c < Y < d)$ are all equivalent, which is certainly not true for discrete distributions.
- The area under the curve between any two points c and d equals to the probability that the random variable Y assumes values between c and d. That is, $P(c \leq Y \leq d) = \int_c^d f(y)dy$, if $a \leq c \leq d \leq b$.

Example 6.6. Find $P(0 \leq Y \leq 0.25)$ using the *pdf* given in example 6.5.

Solution: $P(0 \leq Y \leq 0.25) = \int_0^{0.25} 2ydy = y^2|_0^{0.25} = (0.25)^2 - 0^2 = 0.0625$

6.3 Expectations

6.3.1 Mean and Variance of a Random Variable

A probability distribution of a random variable has parameters describing its central tendency and variability.

The mean, μ, of a random variable Y is known as the *expected value* of Y, denoted by $E(Y)$. It is defined as:

$$E(Y) = \mu = \begin{cases} \sum_{i=1}^{\infty} y_i p(y_i) & \text{if } Y \text{ is a discrete random variable;} \\ \int_{-\infty}^{\infty} y f(y) dy & \text{if } Y \text{ is a continuous random variable.} \end{cases}$$

The variance, σ^2, of the random variable Y is the expected value of the square of the deviation of Y from its mean.

$$V(Y) = \sigma^2 = E(Y-\mu)^2 = \begin{cases} \sum_{i=1}^{\infty} (y_i - \mu)^2 p(y_i) & \text{if } Y \text{ is a discrete random variable;} \\ \int_{-\infty}^{\infty} (y-\mu)^2 f(y) dy & \text{if } Y \text{ is a continuous random variable.} \end{cases}$$

Example 6.7. Find the expected number of girls to be observed for a family planning to have three children. Also, calculate the variance and standard deviation.

Solution: Recall the probability distribution for number of girls is:

y_i	0	1	2	3	Total
$p(y_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

$\mu = E(Y) = \sum_{i=1}^{4} y_i p(y_i) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5.$

$\sigma^2 = \sum_{i=1}^{4} (y_i - \mu)^2 p(y_i) = (0-1.5)^2 \times \frac{1}{8} + (1-1.5)^2 \times \frac{3}{8} \times + (2-1.5)^2 \times \frac{3}{8} + (3-1.5)^2 \times \frac{1}{8} = 0.75.$

$\Rightarrow \sigma = \sqrt{0.75} = 0.86.$

6.3.2 Properties of Expectations

- If Y is a random variable and c is a non-zero constant, then
  - $E(c) = c$ and $V(c) = 0$
  - $E(Y \pm c) = \mu \pm c$ and $V(Y \pm c) = \sigma^2$
  - $E(cY) = c\mu$ and $V(cY) = c^2\sigma^2$
- If X and Y are random variables, then
  - $E(X \pm Y) = E(X) \pm E(Y)$
  - $E(X \pm Y)^2 = E(X^2) \pm 2E(XY) + E(Y^2)$

Note: $V(Y) = \sigma^2 = E(Y-\mu)^2 = E(Y^2) - [E(Y)]^2$.

Example 6.8. Find the mean and variance of the *pdf* $f(y) = \begin{cases} 2y, & 0 \leq y \leq 1; \\ 0, & \text{otherwise.} \end{cases}$

Solution:

- The mean is $E(Y) = \int_0^1 yf(y)dy = \int_0^1 y(2y)dy = \int_0^1 2y^2dy = \frac{2}{3}y^3|_0^1 = \frac{2}{3}$.
- Similarly, $E(Y^2) = \int_0^1 y^2f(y)dy = \int_0^1 y^2(2y)dy = \int_0^1 2y^3dy = \frac{1}{2}y^4|_0^1 = \frac{1}{2}$.
- Thus, the variance is $\sigma^2 = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$.

6.4 Common Discrete Distributions

6.4.1 The Binomial Distribution

A binomial distribution is one of most frequently used discrete distribution that is very useful in many practical situations involving only two types of outcomes; dead or alive, sick or well, male or female. For example, if 72% of new born infants survive up to 70 years, then the random variable Y represents the survival status of the infant at age of 70 years where $Y = 1$ with $P(Y = 1) = 0.72$ if the infant survives and $Y = 0$ with $P(Y = 0) = 0.28$ if infant does not survive. Thus, the probability distribution of Y is:

y	$P(Y = y)$
1	0.72
0	0.28
Total	1.00

This distribution is associated with an experiment called *Bernoulli* trial which exhibits the following properties.

1. Each trial has only *two* mutually exclusive outcomes or outcomes that can be reduced to two. One of the outcomes is labeled as *success* and the other is labeled as *failure*.
2. The outcome of each trial is *independent*. That is, the outcome of one trial does not affect the outcome of another.
3. The trials are *identical*. That is, the probability of success, denoted by π, remains the same from trial to trial. Consequently, the probability of failure, $1 - \pi$, does not change from trial to trial.
4. The experiment is performed for fixed number of times, say n.

When the Bernoulli experiment is repeated for n independent and identical times, the experiment is called *Binomial* experiment. In a Binomial experiment, the interest is in the *number of successes to be occurred in n Bernoulli trials*. Let Y be the number of successes to be occurred in n trials. Hence, $R_Y = \{0, 1, 2, \cdots, n\}$. Because the range space is finite, Y is a discrete random variable. The probability distribution associated with this random variable is called *Binomial* probability distribution which is characterized by two parameters:

the number of trials n (sample size) and probability of success π. Then, it is written as $Y \sim \text{Bin}(n, \pi)$. As a result, the probability of obtaining y successes in n trials is given by:

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \cdots, n$$

where n is number of trials, y is number of successes to be occurred in n trials, π is the probability of success and $1 - \pi$ is the probability of failure.

Notes:

1. The Binomial distribution is a legitimate discrete probability distribution:
   (a) $0 \leq P(Y = y) \leq 1$; $y = 0, 1, 2, \cdots, n$.
   (b) $\sum_{y=0}^{n} P(Y = y) = 1$.
2. The expected number of successes in n trials of a binomial random variable is $\mu = n\pi$ and its variance is $\sigma^2 = n\pi(1-\pi)$. Note that the mean is greater than the variance.

Example 6.9. An investigator reveals that the probability that infants develop chronic bronchitis in the first year of life is 0.06 in which both parents have chronic bronchitis. A random sample of 20 infants are selected from the same study area.

1. What is the probability there will be exactly 2 bronchitic infants?
2. What is the probability of getting no bronchitic infant?
3. What is the probability of getting at most 2 bronchitic infants?
4. What is the probability of getting at least 3 bronchitic infants?
5. What is the mean, variance and standard deviation of the number of bronchitic infants?

Solution: Let Y be the number of bronchitic infants among $n = 20$ infants. $R_Y = \{0, 1, 2, \cdots, 20\}$. Let π be the probability of an infant developing bronchitis in the first year of life. Hence, $Y \sim \text{Bin}(n = 20, \pi = 0.06)$.

Thus, the probability of getting y bronchitic infants among 20 infants is given by:

$$P(Y = y) = \binom{20}{y}(0.06)^y(0.94)^{20-y};\ y = 0, 1, 2, \cdots, 20$$

1. The probability there will be exactly 2 bronchitic infants is $P(Y = 2)$:

$$P(Y = 2) = \binom{20}{2}(0.06)^2(0.94)^{20-2} = \binom{20}{2}(0.06)^2(0.94)^{18} = 0.2246$$

2. The probability of getting no bronchitic infant is $P(Y = 0)$:

$$P(Y = 0) = \binom{20}{0}(0.06)^0(0.94)^{20-0} = \binom{20}{0}(0.06)^0(0.94)^{20} = 0.2901$$

3. The probability of getting at most 2 bronchitic infants is $P(Y \leq 2)$:

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= \binom{20}{0}(0.06)^0(0.94)^{20} + \binom{20}{1}(0.06)^1(0.96)^{19} + \binom{20}{2}(0.06)^2(0.94)^{18} \\ &= 0.2901 + 0.3703 + 0.2246 \\ &= 0.8850 \end{aligned}$$

4. The probability of getting at least 3 bronchitic infants is $P(Y \geq 3)$:

$$\begin{aligned} P(Y \geq 3) &= P(Y = 3) + P(Y = 4) + P(Y = 5) + \cdots + P(Y = 20) \\ &= 1 - P(Y < 3) \\ &= 1 - P(Y \leq 2) \\ &= 1 - \{P(Y = 0) + P(Y = 1) + P(Y = 2)\} \\ &= 1 - (0.2901 + 0.3703 + 0.2246) \\ &= 0.1150 \end{aligned}$$

5. The expected number of bronchitic infants among the 20 infants is $\mu = 20(0.06) = 1.2$. Also, the variance of the number of bronchitic infants is $\sigma^2 = 20(0.06)(0.94) = 1.128$ and the standard deviation is $\sigma = 1.062$.

Example 6.10. In observing patients administered a new drug product in a properly conducted clinical trial, the number of persons experiencing a particular side effect is 1 in 1000. What is the probability that 4 of a random sample of 1000 patients experience a particular side effect? What is the expected number of patients experiencing the side effect of the drug?

Solution: Let Y be the number of patients experiencing a particular side effect of the drug among $n = 1000$ patients. $R_Y = \{0, 1, 2, \cdots, 1000\}$. Hence, $Y \sim \text{Bin}(n = 1000, \pi = 0.001)$. Thus, the probability that 4 of 1000 patients experiencing a particular side effect is:

$$P(Y = 4) = \binom{1000}{4}(0.001)^4(1 - 0.001)^{1000-4} = \binom{1000}{4}(0.001)^4(1 - 0.001)^{996} = 0.0153.$$

The expected number of patients experiencing the side effect of the drug is $\mu = n\pi = 1000(0.001) = 1$.

Exercise 6.1. What is the probability of 2 lymphocytes out of 10 white blood cells if the probability that any one cell is a lymphocyte is 0.2? Also, find the expected number of lymphocytes.

6.4.2 The Poisson Distribution

Poisson distribution is another theoretical discrete probability distribution, which is useful for modeling the number of successes to be occurred independently and randomly in a certain time, space, $\cdots$. It differs from binomial distribution in the sense that it is not possible to count the number of failures even though the number of successes is known. For example, in the case of 'number of deaths from a particular disease per day', only the number of patients

who died in a given day is known but it is not possible to count the number of patients who did not die in that day.

Accordingly, it is not possible to determine the number of trials (total number of outcomes - successes and failures) and hence binomial distribution cannot be applied as a decision making tool. In such situation, the poisson distribution should be used given the events occur randomly and independently at a constant average rate of successes.

Other examples include number of patients coming to hospital for emergency treatment, number of telephone calls going to a switch board system, number of cars in a certain parking lot, number of customers coming to a bank for service and so on.

Like a Binomial distribution, the interest in Poisson distribution is the *number of successes to be occurred in a specified unit of time or space.* Let Y be the number of successes in a specific unit of time or space. Hence, $R_Y = \{0, 1, 2, \cdots\}$. Then, Y follows a poisson distribution with a single parameter λ, average rate of successes (number of successes in a specified unit of time or space), and it is written as $Y \sim \text{Poisson}(\lambda)$. Hence, the probability of getting y successes in the same unit of time or space is:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \cdots$$

where λ is the average number of successes per unit of time or space. Here $e = 2.71828$.

Notes:

1. The poisson distribution is a legitimate discrete probability distribution:
   (a) $0 \leq P(Y = y) \leq 1$; $y = 0, 1, 2, \cdots$.
   (b) $\sum_{y=0}^{\infty} P(Y = y) = 1$.
2. The expected number of successes in a specified unit of time or space is $\mu = \lambda$ and its variance is $\sigma^2 = \lambda$. Here, the mean and variance are equal.
3. The poisson random variable has no theoretical maximum value, but the probabilities tail off towards zero very quickly.

Example 6.11. On average, there are 4.6 deaths from typhoid fever over a 1-year period. Find the mean and variance of the number of deaths from typhoid fever. In addition, what is the probability of

1. 2 deaths from typhoid fever over a 1-year period?
2. at least 1 death from typhoid fever over a 1-year period?

Solution: Let Y be the number of deaths from typhoid fever over a 1-year period. $R_Y = \{0, 1, 2, \cdots\}$. Let λ be the average number of deaths from typhoid fever over a 1-year period. Hence, $Y \sim \text{Poisson}(\lambda = 4.6)$.

Both the mean and variance of the number of deaths from typhoid fever in a 1-year period is 4.6. The probability of y deaths from typhoid fever over a 1-year period is given by:

$$P(Y = y) = \frac{e^{-4.6}\ 4.6^y}{y!}, \quad y = 0, 1, 2, \cdots .$$

1. The probability of 2 deaths from typhoid fever over a 1-year period:

$$P(Y = 2) = \frac{e^{-4.6}\ 4.6^2}{2!} = 0.1063$$

2. The probability of at least 1 death from typhoid fever over a 1-year period:

$$\begin{aligned} P(Y \geq 1) &= P(Y = 1) + P(Y = 2) + \cdots \\ &= 1 - P(Y < 1) \\ &= 1 - P(Y = 0) \\ &= 1 - 0.0101 \\ &= 0.9899 \end{aligned}$$

Exercise 6.2. Suppose a hospital Accident and Emergency department has an average of 10 new emergency cases per hour. Calculate the probability of observing exactly 10 new emergency cases in any given hour.

Poisson approximation to Binomial Distribution

Another important use of the poisson distribution is its approximation to the binomial distribution. Consider the binomial distribution for large number of trials (sample size) n and small probability of success π. Recall the mean of the distribution is $n\pi$ and the variance is $n\pi(1-\pi)$. If the probability of success π is small, then the probability of failure $1-\pi$ is approximately 1. Thus, in this particular case, the mean and variance are almost equal (that is, $n\pi(1-\pi) \approx n\pi$) like the poisson distribution. Therefore, the binomial distribution with a large number of trials n and small probability of success π can be approximated by a poisson distribution with parameter $\lambda = n\pi$.

Example 6.12. Consider again example 6.10. Using a poisson distribution approximation, find the probability that 4 of a random sample of 1000 patients experience a particular side effect of the drug?

Solution: Let Y be the number of patients experiencing a particular side effect of the drug among $n = 1000$ patients. The expected number of patients experiencing the side effect of the drug is $\lambda = n\pi = 1000(0.001) = 1$. Hence, now $Y \sim \text{Poisson}(\lambda = 1)$. Thus, the probability that 4 of 1000 patients experiencing a particular side effect is:

$$P(Y = 4) = \frac{e^{-1}\ 1^4}{4!} = 0.0153.$$

6.5 Common Continuous Distributions

6.5.1 The Normal Distribution

The most often used continuous probability distribution is the *normal* (also called *gaussian*) distribution. This distribution plays a very important and pivotal role in the area of statistical inference (estimation and hypothesis testing), which is the major topic of the remainder of this lecture note.

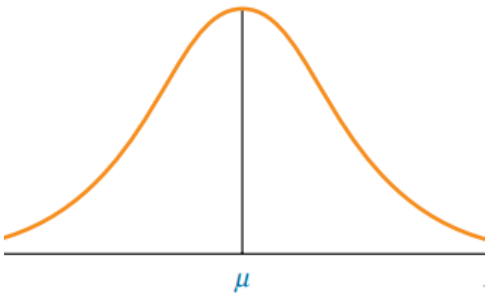
The Normal Distribution Curve

The *pdf* of a continuous random variable Y with a normal distribution is given by:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty$$

where μ is the mean and σ^2 is the variance of the random variable. Hence, normal distribution is determined by two parameters: mean μ and variance σ^2, and it is written as $Y \sim \mathcal{N}(\mu, \sigma^2)$. Note the constants $\pi = 3.14159$ and $e = 2.71828$.

The curve of a normal distribution is symmetric and bell-shaped as shown in the following figure.



Notes:

1. The normal distribution is a legitimate continuous probability distribution:
   (a) $f(y) \geq 0$ for $-\infty < y < \infty$.
   (b) $\int_{-\infty}^{\infty} f(y)dy = 1$
2. $E(Y) = \mu$ and $V(Y) = \sigma^2$.

Properties of Normal Distribution

Some of the interesting features of a normal distribution are:

1. The random variable Y can take on *any value*: negative, zero, or positive. That is $R_Y = \{y : -\infty < y < \infty\}$.

2. The curve is *symmetric at the mean*, with the shape of the curve to the left of the mean a mirror image of the shape of the curve to the right of the mean. In other words, that the number of observations below the mean is the same as the number of observations above the mean. This means the mean and median are equal.

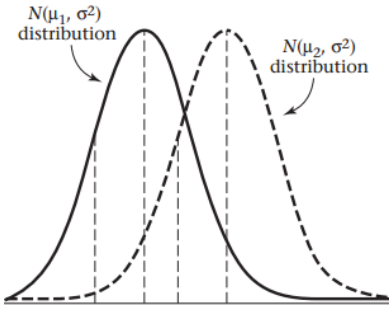
3. The height of the curve is *maximum at the mean*. Thus, the mean and mode coincide. This means the normal distribution has the same value for the mean, median and mode.

4. The curve declines and extends indefinitely in both directions from the mean to infinity. But, theoretically, it never touches the horizontal axis.

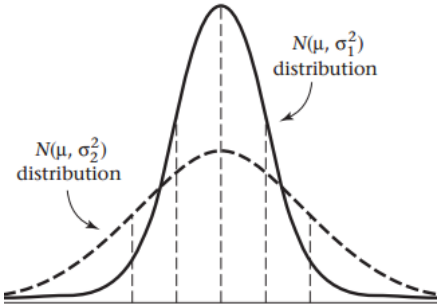
5. The corresponding deciles, quartiles and percentiles are at equidistant from the mean.

Mean and Variance of a Normal Distribution

The *pdf* of a normal distribution describes a family of curves which may differ only with regard to μ and σ^2, but have the same characteristics. The two parameters, μ and σ^2, determine the location and peakedness of a normal distribution. The mean μ determines whether the curve is located to the right or left side. Large values of the mean μ indicates the curve is located to the right side and small values of the mean μ indicates the curve is located to the left. For illustration, two normal distributions with the same variance but with different means are presented below.

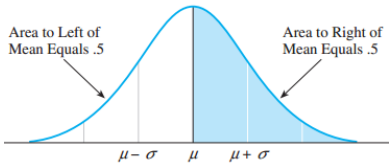


The variance (standard deviation) of a normal distribution also determines the flatness (wideness) or peakedness (narrowness) of the curve. The variance σ^2 (standard deviation σ) of a normal distribution also determines the flatness or wideness of the curve. Larger value of the variance σ^2 result in a wider or flatter curve showing more variability. On the other hand, smaller value of the variance σ^2 result in a more picked curve showing more uniformity. Two normal distributions with the same mean but with different variances are shown below.



Finding probabilities in the Normal Distribution

The total area (representing maximum value for probability) under the *pdf* of a normal curve over the entire range from $-\infty$ to ∞ is 1. Also, since the curve is symmetric at the mean, the area to the right and left of the mean is equal. Specifically, the area (probability) to the left of the mean is 0.5 and the area (probability) to the right of the mean is 0.5. That is, $P(Y > \mu) = P(Y < \mu) = 0.5$.



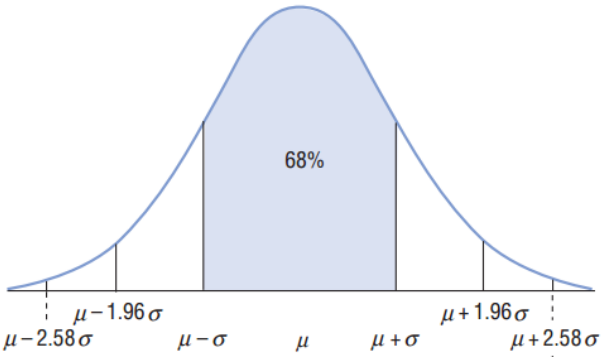
$$P(Y < \mu) = \int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = 0.5 \text{ and } P(Y > \mu) = \int_{\mu}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = 0.5$$

In general, the area (probability) of Y between two values y_1 and y_2 is defined as:

$$P(y_1 < Y < y_2) = \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy.$$

Empirical rule also relates the standard deviation (σ) of a normal distribution to the proportion of observations around the mean (μ):

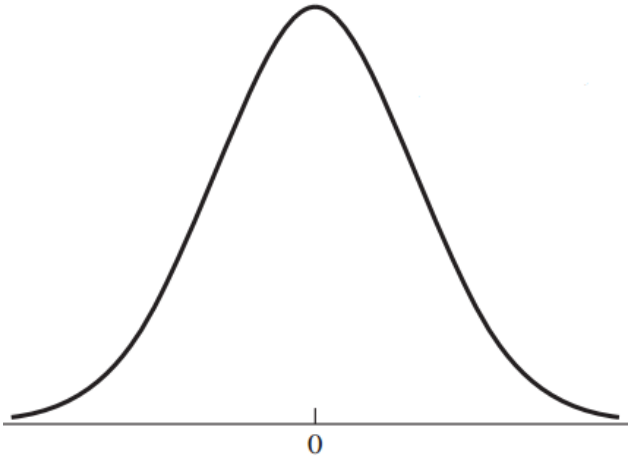
- About 68.2% of the observations are within 1 standard deviation of the mean ($\mu \pm \sigma$), that is, $P(\mu - \sigma < Y < \mu + \sigma) = 0.682$.
- About 95.4% of the observations are within 1.96 (≈ 2) standard deviations of the mean ($\mu \pm 2\sigma$), that is, $P(\mu - 1.96\sigma < Y < \mu + 1.96\sigma) = 0.954$.
- About 99.7% of the observations are within 2.58 (≈ 3) standard deviations of the mean ($\mu \pm 3\sigma$), that is, $P(\mu - 2.58\sigma < Y < \mu + 2.58\sigma) = 0.997$.



Standard Normal Distribution

The *pdf* of a normal distribution to be integrated for evaluating probabilities is different for different values of μ and σ^2. Fortunately, a normal distribution can easily be *standardized*, which allows to integrate a single function for any normal distribution.

Suppose Y has a normal distribution with mean μ and variance σ^2, i.e, $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then, the Z score defined as $Z = \frac{Y-\mu}{\sigma}$ will have also a normal distribution with mean 0 and variance 1, that is, $Z \sim \mathcal{N}(0, 1)$.



Such a random variable that has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called *standard* normal distribution. Hence, the *pdf* of the standard normal variate Z is given by:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty.$$

Notes:

1. The standard normal distribution is a legitimate continuous probability distribution:
   (a) $f(z) \geq 0$ for $-\infty < z < \infty$.
   (b) $\int_{-\infty}^{\infty} f(z)dz = 1$
2. $E(Z) = 0$ and $V(Z) = 1$.

Finding probabilities, given the $z-$scores, in the Standard Normal Distribution

The total area (representing maximum value for probability) under the *pdf* of the standard normal curve over the entire range from $-\infty$ to ∞ is also 1. The area to the right and left of the central value ($\mu = 0$) is 0.5 (as it is symmetric about 0):

$$P(Z < 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0.5 \text{ and } P(Z > 0) = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0.5$$

Similarly for any standard normal variate Z, the area (probability) between two values z_1 and z_2 is defined as:

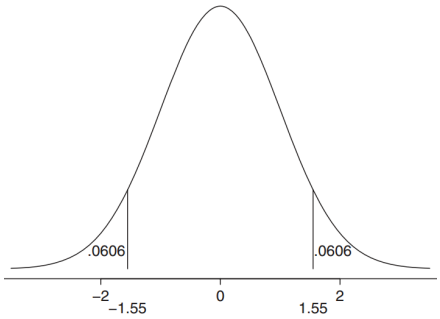
$$P(z_1 < Z < z_2) = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

However, manual integration and evaluation of the *pdf* of the (standard) normal distribution is quite complicated. Rather, computer software is used for evaluating the corresponding

probabilities, and a standard normal table is already tabulated.

The standard normal table to be used here provides probability values above a certain $z-$score value, say z. That is, the provides the integrations of the standard normal distribution from a positive number z to ∞, $P(Z > z) = \int_z^{\infty} f(z)dz$.

Because the standard normal curve is symmetric about zero, the probability that Z greater than z is the same as the probability that Z less than $-z$. That is, $P(Z > z) = P(Z < -z)$. For example, $P(Z > 1.55 = 0.0606) = P(Z < -1.55)$ as shown in the figure below.



In applied work, there are three types of probabilities that need to be determined:

1. $P(Z > z$, the probability that a standard normal random variable is greater than z,
2. $P(Z < z)$, the probability that a standard normal random variable is less than z, and
3. $P(z_1 < Z < z_2)$, the probability that a standard normal random variable is between the values z_1 and z_2.

The first of these is determined from table to be provided. Because the area under the curve is one, the second is given by $P(Z < z) = 1 - P(Z > z)$. The third is given by $P(z_1 < Z < z_2) = P(Z > z_1) - P(Z > z_2)$.

Example 6.13. Find the area, of the standard normal distribution,

1. to the right of 1.96; $P(Z > 1.96)$.
2. between 0 and 1.96; $P(0 < Z < 1.96)$.
3. to the right of -2; $P(Z > -2)$.
4. to the left of -0.5; $P(Z < -0.5)$.
5. between -1 and 1.5; $P(-1 < Z < 1.5)$.

Solution:

1. $P(Z > 1.96) = 0.0250 = P(Z < -1.96)$
2. $P(0 < Z < 1.96) = P(Z > 0) - P(Z > 1.96) = 0.5 - 0.0250 = 0.4750$
3. $P(Z > -2) = 1 - P(Z < -2) = 1 - P(Z > 2) = 1 - 0.0228 = 0.9772$
4. $P(Z < -0.5) = P(Z > 0.5) = 0.3085$
5. $P(-1 < Z < 1.5) = 1 - P(Z < -1) - P(Z > 1.5) = 1 - 0.1587 - 0.0668 = 0.7745$

Converting any Normal Distribution to Standard Normal Distribution

The general principle for converting any probability expression concerning normal random variable, Y, of the form $P(y_1 < Y < y_2)$ to an equivalent probability expression of the standard normal random variable, Z, of the form $P(z_1 < Z < z_2)$ is subtracting the population mean μ from each boundary point of Y and dividing by the standard deviation σ. That is,

$$\begin{aligned} P(y_1 < Y < y_2) &= \left(\frac{y_1 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{y_2 - \mu}{\sigma} \right) \\ &= P(z_1 < Z < z_2) \end{aligned}$$

Then, the standard normal table is then used to evaluate this latter probability.

Example 6.14. The IQ score of students is normally distributed with a mean of 120 and variance 400. What is the probability that a student will have an IQ

1. above 140?
2. below 150?
3. between 100 and 130?
4. between 140 and 150?

Solution: Let Y be IQ score. Thus, $Y \sim \mathcal{N}(120, 400)$.

$$P(y_1 < Y < y_2) = P\left(\frac{y_1 - 120}{20} < Z < \frac{y_2 - 120}{20} \right)$$

1. $P(Y > 140) = P\left(Z > \frac{140-120}{20}\right) = P(Z > 1) = 0.1587$
2. $P(Y < 150) = P\left(Z < \frac{150-120}{20}\right) = P(Z < 1.5) = 1 - P(Z > 1.5) = 1 - 0.0668 = 0.9332$
3. $P(100 < Y < 130) = P\left(\frac{100-120}{20} < Z < \frac{130-120}{20}\right) = P(-1 < Z < 0.5) = 1 - P(Z > 1) - P(Z > 0.5) = 1 - 0.1587 - 0.3085 = 0.5328$
4. $P(140 < Y < 150) = P\left(\frac{140-120}{20} < Z < \frac{150-120}{20}\right) = P(1 < Z < 1.5) = P(Z > 1) - P(Z > 1.5) = 0.1587 - 0.0668 = 0.0919$

Finding $z-$scores, given the probabilities, in the Standard Normal Distribution

If the concern is to find the $z-$scores for given probability values, the z_α notation is adopted. According to this notation, z_α is a value such that $P(Z > z_\alpha) = \alpha$. For instance, $P(Z > 1.55 = 0.0606)$ and hence $z_{0.0606} = 1.55$. Because of the symmetry of the (standard) normal distribution, $P(Z > z_\alpha) = P(Z < -z_\alpha) = \alpha$ and $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Example 6.15. Find the $z-$score values of the following z_α notations: $z_{0.05}$, $z_{0.01}$, $z_{0.10}$, $z_{0.005}$, $z_{0.025}$.

Solution: The value of z_α is associated with $P(Z > z_\alpha) = \alpha$.

1. $z_{0.05} \Rightarrow P(Z > z_{0.05}) = 0.05 \Rightarrow z_{0.05} \approx 1.65$.

2. $z_{0.01} \Rightarrow P(Z > z_{0.01}) = 0.01 \Rightarrow z_{0.01} \approx 2.33$

3. $z_{0.10} \Rightarrow P(Z > z_{0.10}) = 0.10 \Rightarrow z_{0.10} \approx 1.28$

Example 6.16. Let Y be the variable representing the distribution of scores in biostatistics course. It can be assumed that these scores are normally distributed with $\mu = 75$ and $\sigma = 10$. If the instructor wants no more than 10% of the class to get an A, what should be the cutoff grade? That is, what is the value of y such that $P(Y > y) = 0.10$?

Solution: $P(Y > y) = P(Z > z_{0.10}) = P\left(Z > \frac{y-\mu}{\sigma}\right) = P\left(Z > \frac{y-75}{10}\right) = 0.10$

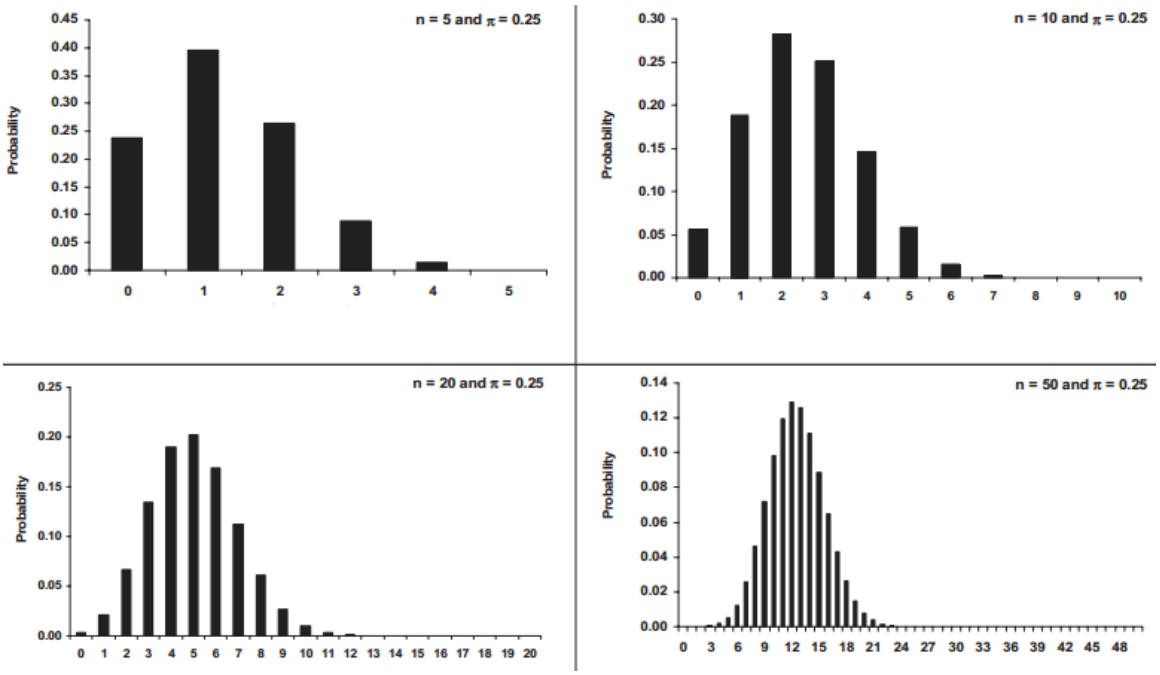
$$\Rightarrow z_{0.10} = \frac{y - 75}{10} = 1.28 \Rightarrow y = 87.8.$$

Therefore, the instructor should assign an A grade to those students with scores 87.8 or higher.

Normal approximation to Binomial Distribution

The binomial distribution is always *symmetric* when the probability of success is 50%, that is, $\pi = 0.50$. For fixed number of trials (sample size) n, the distribution becomes skewed as the probability of success π moves toward 0 ($\pi < 0.50$) or 1 ($\pi > 0.50$). Specifically, the distribution is *positively skewed* when the probability of success is less than 50% ($\pi < 0.5$) and it is *negatively skewed* when the probability of success is greater than 50% ($\pi > 0.5$).

For fixed probability of success π, it becomes *symmetric* as the sample size n increases.

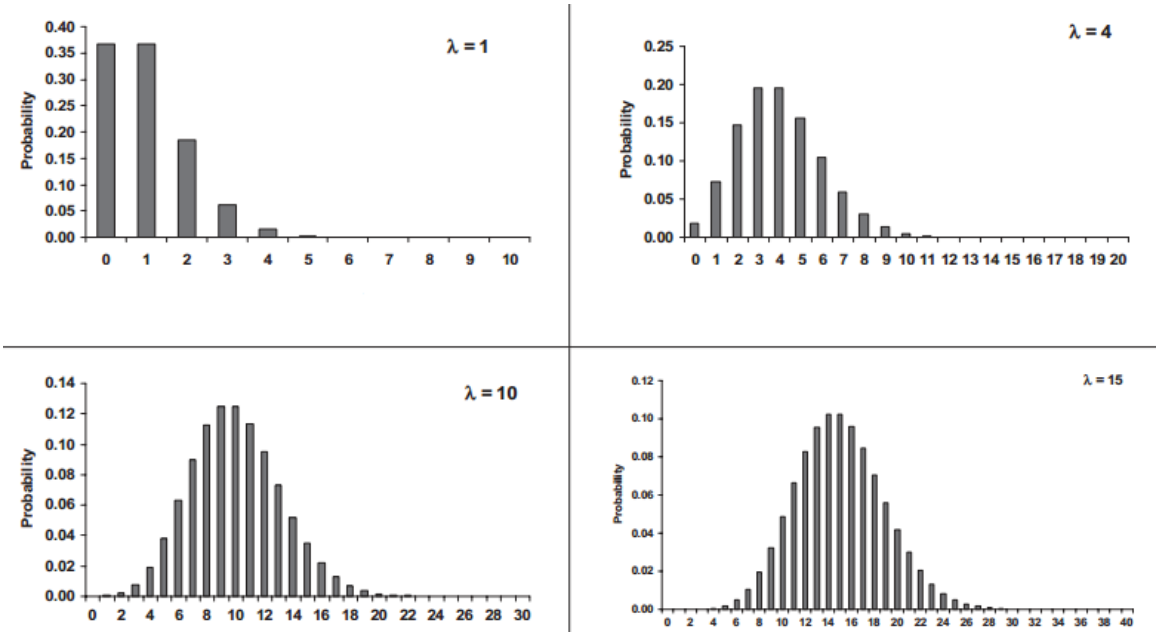


When the sample size n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma^2 = n\pi(1 - \pi)$. For a better normal approximation, a guideline is that the expected number of both outcomes, expected number of successes $n\pi$ and expected number of failures $n(1 - \pi)$, should both be at least 5. For a probability of success $\pi = 0.50$, the sample size

required is only $n \geq 10$. For probability of success $\pi = 0.10$ (or $\pi = 0.90$), the sample size required is at least 50 ($n \geq 50$). When probability of success π far from 0.5 ($\pi \neq 0.50$), larger samples are needed to attain normality.

Normal approximation to Poisson Distribution

In a poisson distribution, for small values of the average number of successes in a given time or space λ, the distribution is *positively skewed*. But, for large values of the average rate of successes, it becomes *symmetrical*.



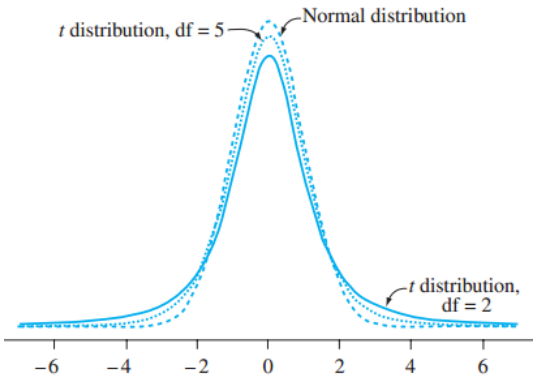
Therefore, when the average rate of success is large ($\lambda \to \infty$), the poisson distribution can be approximated by a normal distribution with mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$.

6.5.2 Other Continuous Distributions

The Student's t Distribution

The t distribution is quite similar to the standard normal in that it is *symmetric* about 0 and bell shaped. However, the curve of a t distribution is "flatter" than the normal.

The t distribution has only one parameter called *degrees of freedom*(df). The degrees of freedom (df) is the number of independent observations that are free to vary. Then the distribution is written as $t(df)$.



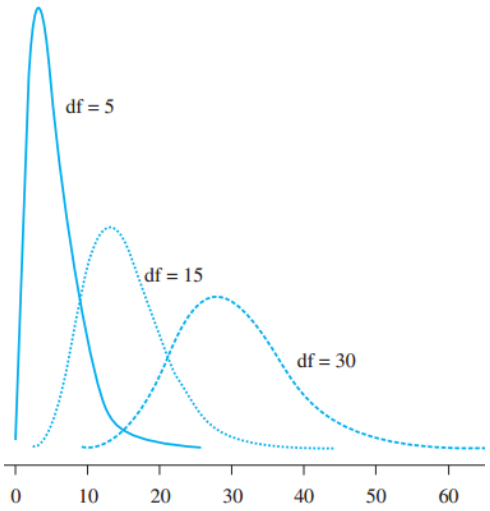
The t values for given probability values is denoted by $t_\alpha(df)$. The notation $t_\alpha(df)$ is the value of the t variable such that the area (probability) to the right is α.

Exercise 6.3. Find the following t values from the t distribution table: $t_{0.05}(5)$, $t_{0.01}(20)$, $t_{0.10}(10)$.

Note: When the degrees of freedom is large, the t distribution is identical to the standard normal distribution. That is, $t(\infty) \approx Z$ where $Z \sim \mathcal{N}(0, 1)$.

The Pearson χ^2 Distribution

Chi-square (χ^2) distribution is a non-negative *positively skewed* distribution. Like the t distribution, it has only one parameter, *degrees of freedom* (df), and it is usually denoted as $\chi^2(df)$.



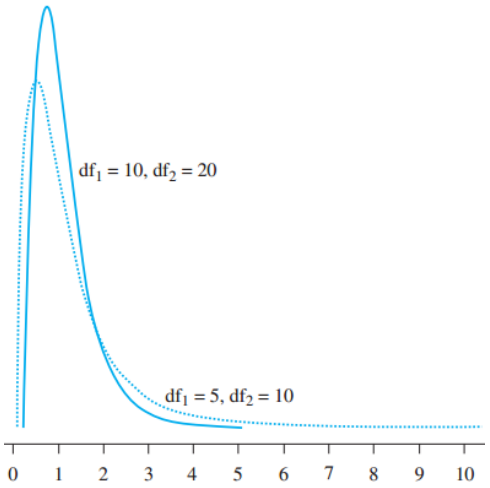
The chi-square values for given probability values is denoted by $\chi^2_\alpha(df)$. The notation $\chi^2_\alpha(df)$ is the value of the chi-square variable such that the area (probability) to the right is α.

Exercise 6.4. Find the following chi-square values from the chi-square probability table: $\chi^2_{0.05}(5)$, $\chi^2_{0.01}(20)$, $\chi^2_{0.10}(10)$.

Note: The square of a standard normal variable is a χ^2 distribution with 1 degrees of freedom. That is, $Z^2 = \chi^2(1)$ where $Z \sim \mathcal{N}(0, 1)$.

The F Distribution

F distribution another non-negative and *right skewed* continuous distribution like the chi-square distribution. It is the ratio of two chi-square distribution and hence, is identified by a set of two degrees of freedom, the first called "numerator degrees of freedom" v_1 and the second called "denominator degrees of freedom" v_2. It is written as $F(v_1, v_2)$.



The F values for given probability values is denoted by $F_\alpha(v_1, v_2)$.

Exercise 6.5. Find the following F values from the F probability table: $F_{0.05}(5, 15)$, $F_{0.01}(20, 18)$, $F_{0.10}(9, 9)$.

Notes:

- If the "numerator degrees of freedom" is 1, the distribution is reduced to the square of t distribution with v_2 degrees of freedom. That is, $F(1, v_2) = t^2(v_2)$.
- If the "denominator degrees of freedom" is large, the distribution is reduced to χ^2 distribution with v_1 degrees of freedom. That is, $F(v_1, \infty) = \chi^2(v_1)$.

Chapter 7

Sampling and Sampling Distributions

7.1 Census Vs Sample Survey

There are two broadly classified statistical investigations: *census survey* and *sample survey*. In the census method, a 100% inspection of each unit of the population, known as *sampling unit*, is made. The latter method is a study in which some elements which are assumed representatives of the population are investigated. It is a statistical process in which we select and examine a sample instead of considering the whole population.

In practice, it may not be possible to collect information on all units of the population. One reason is lack of resources in terms of money, personnel and equipment. Another reason is that sample survey enables us to obtain results on time. Hence, for getting quick results sampling is preferred. Moreover, complete investigation may be destructive in nature. And samples reduce the damages caused by some tests in quality control. For example, in cooking food mothers check whether the food has enough amount of salt, spices, butter and so on, by taking a small amount and testing it. What would happen if the test is all what is in the dish?

7.2 Sampling Techniques

The process of selecting a sample from the population is known as *sampling* and the method of selecting a sample is known as *sampling technique*. In the selection of a sample, always the effort is to make the sample representative of the population. There several sampling methods which can be broadly classified into two categories; *probability* and *non-probability* sampling methods.

In probability sampling, each unit in the population has an *equal chance* of being included in the sample. In the non-probability sampling, the units are drawn using ceratin amount of judgement.

7.2.1 Probability Sampling Techniques

1. **Simple Random Sampling**: In simple random sampling, each and every member of the population has an *equal* and *independent* chance of being selected in the sample. The items that get selected will be purely a matter of chance. Before applying this method, a complete list of all members, called *sampling frame*, should be prepared so that each member can be identified by a distinct number. There are two methods that can be used in order to ensure the randomness of the selection. These are:

   (a) **Lottery Method**: This method is useful in comparatively small size (mostly $N \leq 100$) of population. All members in the population are numbered uniquely on separate pieces of paper of identical size and shape. These slips of paper are then identically folded and mixed up in a container. The probability of the first item being selected out of the total number of N slips of paper is $\frac{1}{N}$, for the second particular piece, this probability is $\frac{1}{N-1}$, since $N-1$ slips of papers left in the container after the first slip has been drawn. Similarly, the probability of the third slip being picked up is $\frac{1}{N-2}$ and so on. The items from the container are selected successively until the desired sample size reached. This would constitute a random sample called *simple random sample*.

   (b) **Random Number Table Method**: A random number table is giving numbers in a random order which are generated using computer. In the lottery method, the selection may subject to human bias as people may identify the slips (chits) in many ways. The inconvenience of preparing slips of paper, shuffling them and choosing the items one by one may be avoided by the use of random number table.

   Suppose N is a k digit number. Choose k digit numbers from the random number table and read out the numbers continuously, vertically or horizontally. If the number is greater than N but less than the biggest multiple of N which has k figures, divide that number by N and take the remainder r and include the r^{th} unit in the sample. Discard random numbers which are greater than the biggest multiple of N with k figures. For example, if $N = 43$ take 2 digit random numbers. If the number is, say, 23 include the unit with number 23 in the sample. If the second number is 68, since it is less than 86, the biggest 2 digit multiple of 43, divide 68 by 43 and take the remainder, 25 and include the unit with number 25 in the sample. If the number obtained is greater than 86, discard the number and go to the next in the table. This process continues until n sampling units are selected.

2. **Systematic Random Sampling**: A systematic sample is formed by selecting the first unit at random, and the remaining units in the sample are automatically selected in some predetermined pattern. The process requires that the members of the population be presented in some kind of order; alphabetically or numerically or in any other order, and every k^{th} unit ($k = \frac{N}{n}$ is called *sampling interval*) is included in the sample after the first item has been selected randomly. This may be considered representative as the sample is evenly distributed over the whole population. There are two methods of systematic sample selection. These are:

   (a) **Linear Systematic Sampling**: Suppose N is a multiple of n, that is, $N = nk$. The procedure is to select a random number, say, j such that $1 \leq j \leq k$ and then

select the j^{th} and every subsequent $j+k$, $j+2k$, $\cdots$, $[(n-1)k]^{th}$ positional units. This sampling plan is known as *linear systematic sampling.*

(b) **Circular Systematic Sampling**: In linear systematic sampling, the situation that N is a multiple of n does not always hold, in such case a sample of $n-1$ units, instead of n, will be obtained. As a result, *circular systematic sampling* is applied when $N \neq nk$. Hence, take $\frac{N}{n}$ as k by rounding to the nearest integer. Select a random number from 1 to N, let the number be m. Now select every $(m+jk)^{th}$ unit when $m+jk < N$ and select every $(m+jk-N)^{th}$ unit when $m+jk > N$ putting $j = 1, 2, \cdots$ till n units are selected. By this method always a sample size of n will be obtained.

3. **Stratified Random Sampling**: When the population is *heterogenous* with respect to the characteristic in which one is interested, the population should be divided in *homogeneous groups*, called *strata* (e.g., gender, region, $\cdots$). This ensures maximum uniformity (homogeneity) within each stratum and largest degree of variability (heterogeneity) among strata. From each stratum, a separate sample is selected using *simple random sampling.* This sampling method is known as *stratified sampling.* The total sample size might be allocated to each stratum equally (*equal allocation*) or proportionally (*proportional allocation*).

4. **Cluster Sampling**: In this sampling technique, the population is divided into subpopulations known as *clusters.* But, unlike a stratum, the units within a cluster are relatively heterogenous. The number of clusters to be selected depends on how representative each cluster is of the entire population. If all clusters are similar in this regard, then sampling a small number of clusters will provide good estimates of the population parameters. Then, from each cluster, a random sample of the desired size will be selected.

5. **Multistage Sampling**: This method of sampling is useful when the population is very large and widely spread. In a multistage sampling technique, the population is divided into a number of successive stages. The sample size at each stage is determined by the relative population size at each stage.

7.2.2 Non-probability Sampling Techniques

Nonprobability sampling gives rise to those methods where the units are selected deliberately. No probability is attached or can be computed for an item being selected.

1. **Quota Sampling**: In case of stratified sampling if the cost of selecting sampling units from each stratum is very high, then the investigator is assigned a quota (fixed number of subjects) in each stratum. Then the actual selection of persons is left at the discretion of the investigator.

2. **Judgment Sampling**: In this method, sampling units are selected on the judgement of the person doing the study. The underlying assumption is that the unit selected truly represent the entire population. For example to find out the potential of drip irrigation technology, a researcher may go the teachers of an Agricultural University.

3. **Convenience Sampling**: Here, an investigator selects the sample at his own convenience. This method is based on the assumption that the population is homogeneous and the individuals selected and interviewed similar information with regard to the characteristic under study. For example, persons selected from gas stations or petrol pumps to collect information about the quality of gas or petrol, service or correctness of the measurement, e.t.c are supposed to represent the population of gasoline buyers.

4. **Snowball Sampling**: Snowball sampling technique involves the practice of identifying set of respondents who can, in turn, help the investigator to identify some other person who will be included in the study. After interviewing this person, s/he will contact the other person and interview him/her. In this way, a chain process continuous till the required number of persons are interviewed. This type of sampling is most suitable for rare subjects, for example, a study involving commercial sex workers.

7.3 Errors In Surveys

1. **Sampling Errors**: Sampling errors are the errors which are introduced due to errors in the selection of a sample or the discrepancies between population parameters and estimates which are derived from random sample. That is, the absolute value of the difference between a point estimate and the corresponding population parameter is called *sampling error*. This error is due to sampling fluctuations which are the outcome of the random sampling process. These errors can be controlled by proper choice of sampling methods and increasing sample size.

2. **Nonsampling Errors**: It is experienced that studies based on complete enumeration do not yield similar results in repeated enumerations. Such a discrepancy occurs due to many errors which are termed as nonsampling errors. Some of the sources of such errors are observation error or response error, measurement errors or errors in editing and tabulation of data. These errors can be minimized through superior management of survey, employing benefiting personnel and by using modern computational aids.

7.4 Concepts of Statistical Inference

The primary objective of a statistical analysis is drawing statistically valid conclusions about the characteristics of the population based on the results obtained from sample. There are two important terms that are key to statistical inference. These are population quantities (*parameters*) and their sample counterparts (*statistics*. A *parameter* is a fixed (but usually unknown) summary measure of the characteristic of a population. For example; population mean (μ), population proportion (π), population rate λ, population variance (σ^2), population standard deviation (σ) are examples of parameters.

On the other hand, a *statistic* is a known summary measure of the characteristic of a sample. For example; sample mean ($\bar{y}$), sample proportion (p), sample rate r, sample variance (s^2), sample standard deviation (s) are examples of statistics. These measures, unlike parameters, are random because their values vary from sample to sample.

Generally, statistical inference can be defined as the process of making conclusions for population parameters using sample statistics. It generally takes two forms, namely, *estimation* of a parameter and *testing of a hypothesis*. *Estimation* is concerned with determining the values of specific population parameters using sample data; *hypothesis testing* is concerned with testing whether a particular value of a population parameter is plausible or not.

7.4.1 Estimation of Parameters

For the purpose of general discussion, let θ be a population parameter and $\hat{\theta}$ be the corresponding statistic. The statistic $\hat{\theta}$ intended for estimating the parameter θ is called an *estimator* of θ. A specific numerical value of an estimator calculated from the sample is called an *estimate*. The process of obtaining an estimate of the unknown value of a parameter by a statistic is called *estimation*. There are two types of estimations. One is *point* estimation and the other is *interval* estimation.

Point Estimation

Point estimation is the process of obtaining a *single* sample value that is used to estimate the desired population parameter. The estimator is known as *point estimator*. For example:

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a point estimator of μ.
- $P = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a point estimator of π.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is a point estimator of σ^2.

Properties of Point Estimators

The best estimator should be highly reliable and have desirable properties like unbiasedness, consistency, efficiency and sufficiency. These criteria are described as follows:

1. **Unbiasedness**: An estimator is a random variable since it is always a function of the sample values. A sample statistic is considered to be an *unbiased* estimator if its expected value equals the population parameter which is being estimated. This means if $E(\hat{\theta}) = \theta$, then $\hat{\theta}$ is an unbiased estimator of θ.

   For example, $\bar{Y}$ is an unbiased estimator of μ, P is an unbiased estimator of π and S^2 is an unbiased estimator of σ^2 (but S is a biased estimator of σ).

   Also, $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is a biased estimator of σ^2 since $E(S_n^2) \neq \sigma^2$.

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

   For example, $\bar{Y}$ is a consistent estimator of μ, P is also a consistent estimator of π.

3. **Efficiency**: An estimator is considered to be *efficient* if its value remains stable from sample to sample. The best estimator would be the one which would have the *least variance* from sample to sample. From the three point estimators of central tendency, namely, the mean, median and mode, the mean is considered the least variant and hence is a better estimator for the population mean.

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 sample mean uses all the sample values in its computation while median and mode do not. Hence, mean is the better estimator in this sense.

Examples of point estimates:

- The proportion of smokers among patients having a respiratory problem is 0.72. The sample proportion,, $p = 0.72$ is a point estimate of the population proportion, π, smokers among patients having a respiratory problem.

- The mean SBP of patients under antihypertensive treatment in St. Paul's Hospital is 120mmHg. The sample mean SBP of patients, $\bar{x} = 120$mmHg, is a point estimate of the population mean μ.

- The current prevalence of HIV in Addis Ababa is 25%. Here, the sample prevalence $p = 0.25$ estimates the population prevalence of HIV in Addis Ababa, π.

Interval Estimation

Point estimator has some drawbacks. First, a point estimator *may not exactly locate* a population parameter, that is, the value of a point estimator *is not likely to be equal* to the value of a parameter, resulting in some margin of uncertainty. If the sample value is different from the population value, the point estimator *does not indicate the extent of the possible error.* Second, a point estimate *does not specify as to how confident we can be that the estimate is close to the parameter* it is estimating. That is, we *cannot attach any degree of confidence* to such an estimate as to what extent it is closer to the value of a parameter. Because of these limitations of point estimation, interval estimation is considered desirable. *Interval estimation* involves the determination of an *interval* (*a range of plausible values*) on both sides of a point estimate within which a population parameter is assumed to lie with a *specified degree of confidence.* Therefore, an interval estimate of a parameter is of the form: (lower bound, upper bound).

Examples of interval estimates:

- The proportion of smokers among patients having a respiratory problem is in between 0.67 and 0.77, $0.67 < \pi < 0.77$.

- The mean SBP of patients under antihypertensive treatment in St. Paul's Hospital is between 110mmHg and 130mmHg, $110\text{mmHg} < \mu < 130\text{mmHg}$.

- The current prevalence of HIV in Addis Ababa is between 21% and 29%, $0.21 < \pi < 0.29$.

7.5 Sampling Distributions

As said before, sample statistics are known values but vary from sample to sample taken from the same population. This variability of sample statistics is always present and must be accounted for in any inferential procedure. In fact, a sample statistic is a random variable and, like any other random variable, it has a probability distribution. The probability distribution of a statistic is called *sampling distribution*. Therefore, the *sampling variation* of a statistics is accounted for by identifying its probability distributions.

The expected value (mean) of the sampling distribution of a statistic is the population parameter corresponding to that statistic and the measure of variability of sample statistics is the variance. The positive square root of the variance of a statistic is called *standard error (SE)* which is the standard deviation of the distribution of the sample mean.

7.5.1 Sampling Distribution of the Sample Mean

The sampling distribution of $\bar{Y}$ is the probability distribution of all possible values of the sample mean $\bar{y}$. The idea is that if a number of repeated samples of fixed size n is drawn from population having a mean μ and variance σ^2, each sample mean $\bar{y}$ will have a different value. Thus, $\bar{Y}$ itself is a random variable and hence it has a probability distribution.

For example, suppose there are $N = 5$ patients in a population labeled alphabetically A, B, C, D, and E. If a random sample of $n = 2$ patients is decided to be selected, then there are $5C_2 = 10$ possible samples of size $n = 2$:

Number	1	2	3	4	5	6	7	8	9	10
Sample	(A,B)	(A,C)	(A,D)	(A,E)	(B,C)	(B,D)	(B,E)	(C,D)	(C,E)	(D,E)

Note here that each possible sample has an equal probability of $\frac{1}{10}$. A sample of size n selected from a population containing N sampling units ($n < N$) is said to be a *random sample* if every different sample of size n from the population has an equal probability of being selected.

Let the observed values representing a certain characteristic of the five patients be {2, 3, 4, 5, 6}. Then, the population mean is $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{5}\sum_{i=1}^{5} y_i = 4$ and the population standard deviation is $\sigma = 1.58$.

Now let us find the sample means for all the ten possible samples:

Sample	(A,B)	(A,C)	(A,D)	(A,E)	(B,C)	(B,D)	(B,E)	(C,D)	(C,E)	(D,E)
$\bar{y}$	2.5	3	3.5	4	3.5	4	4.5	4.5	5	5.5

Thus, the sampling distribution of the sample mean is:

$\bar{y}$	2.5	3	3.5	4	4.5	5	5.5
$p(\bar{y})$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Clearly, the sample mean takes different values, the smallest value is 2.5 and the largest is 5.5, with different probabilities.

Expected Value of the Sample Mean $\bar{Y}$

Let us consider the above example again. The *expected value* of the sample mean can be calculated from the sampling distribution as: $E(\bar{Y}) = \mu_{\bar{Y}} = \sum_{i=1}^{10} \bar{y}p(\bar{y}) = 2.5\times\frac{1}{10}+3\times\frac{1}{10}+3.5\times\frac{2}{10}+4\times\frac{2}{10}+4.5\times\frac{2}{10}+5\times\frac{1}{10}+5.5\times\frac{1}{10} = 4$. This value is exactly equal to the population mean $\mu = 4$.

In general, assume a population with mean μ and variance σ^2. If a simple random sample of size n (Y_1, Y_2, $\cdots$, Y_n) is selected from the population, the sample mean is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then, $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$. The mean of the distribution of the sample mean is equal to the population mean μ. That is, $E(\bar{Y}) = \mu$. Therefore, $\bar{Y}$ is an *unbiased* estimator of μ.

Standard Error of the Sample Mean $\bar{Y}$

Standard error of the mean is the measure of variability of sample means, that is, the standard deviation of the distribution of the sample mean. In general, a *standard error* is the standard deviation of the sampling distribution of an estimator.

Let us consider the above example once more. The variance of the sampling distribution of the sample mean is:
$E(\bar{Y} - \mu_{\bar{Y}})^2 = \sigma^2_{\bar{Y}} = \sum_{i=1}^{10}(\bar{y} - \mu_{\bar{Y}})^2 p(\bar{y}) = (2.5-4)^2 \times \frac{1}{10} + \cdots + (5.5-4)^2 \times \frac{1}{10} = 0.555.$
This implies, the standard deviation of the sample mean is $\sigma_{\bar{Y}} = 0.745$. This value is the ratio of the population standard deviation to the square root of sample size,???.

Assuming a population with mean μ and variance σ^2, the variance of the sample mean $\bar{Y}$ is:

$$V(\bar{Y}) = \sigma^2_{\bar{Y}} = V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}\sigma^2.$$

Thus, the standard error of the sample mean $\bar{Y}$ is the positive square root of its variance and denoted as $\text{SE}(\bar{Y}) = \sigma_{\bar{Y}} = \sigma/\sqrt{n}$. Therefore, the standard error of the distribution of the sample mean is equal to the standard deviation of the population divided by the square root of the sample size.

In reality the population standard deviation σ is unknown, hence the sample standard deviation s is used in place of σ. This provides the estimated standard error of the mean, that is, $\widehat{\text{SE}}(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s/\sqrt{n}$.

If the sample size increases, the standard error of the sample mean $\bar{Y}$ decreases. Thus, $\bar{Y}$ is a *consistent* estimator of the population mean μ.

7.5.2 Sampling Distribution of the Sample Proportion

Consider the number of successes to be occurred in n binomial experiments with probability of success π. If a random sample of size n (Y_1, Y_2, $\cdots$, Y_n) is selected from the population,

then the mean and variance of the number of successes Y_i are π and $\pi(1-\pi)$, respectively. The sample proportion of successes P is also the ratio of the sum of the observed number of successes Y_i to the total number of outcomes n, that is, $P = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Expected Value of the Sample Proportion P

The expected value (mean) of the sampling distribution of the sample proportion P is:

$$E(P) = \mu_P = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\pi = \pi$$

This shows that the mean of all possible the sample proportion P values is the same as the population proportion π. That is, $E(P) = \pi$. Thus, the sample proportion P is an *unbiased* estimator of the population proportion π.

Standard Error of the Sample Proportion P

The variance of the sampling distribution of the sample proportion P is:

$$V(P) = \sigma_P^2 = V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\pi(1-\pi) = \pi(1-\pi)/n.$$

Thus, the standard error of P is $\text{SE}(P) = \sigma_P = \sqrt{\pi(1-\pi)/n}$. This provides the estimated standard error of the proportion as $\widehat{\text{SE}}(P) = \hat{\sigma}_P = \sqrt{p(1-p)/n}$.

Since the standard error of P decreases to zero when the sample size increases, the sample proportion P is a *consistent* estimator of the population proportion π.

7.5.3 Central Limit Theorem

If random samples of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{Y}$ will have approximately a normal distribution with mean μ and variance σ^2/n for large sample size n. This is the most celebrated theorem called *central limit theorem*. When the population distribution is normal, the sampling distribution of $\bar{Y}$ is exactly normal for any sample size n.

Remarks:

- Suppose the sample mean $\bar{Y}$ has a normal distribution with mean μ and variance σ^2/n $\{\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)\}$.
  - If the population variance σ^2 (standard deviation σ) is known, then the $Z-$score of the sample mean will have a standard normal distribution. That is,

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- If the population standard deviation σ is unknown, it can be approximated by the sample standard deviation s. Then, as long as the sample size n is large, the $z-$score, based on the estimated standard error, of the sample mean will also have a standard normal distribution. That is:

$$Z = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Example 7.1. A person visits his/her doctor with concerns about his/her blood pressure. If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure and medication may be prescribed. A patient's blood pressure readings often have a considerable variation during a given day. Suppose a patient's systolic blood pressure readings during a given day have a normal distribution with a mean $\mu = 160$ mmHg and a standard deviation $\sigma = 20$ mmHg.

1. What is the probability that a single blood pressure measurement will fail to detect that the patient has high blood pressure?
2. If five blood pressure measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 150 and hence fail to indicate that the patient has high blood pressure?
3. How many measurements would be required in a given day so that there is at most 1% probability of failing to detect that the patient has high blood pressure?

Solutions: Let Y be the blood pressure measurement of the patient. Thus, $Y \sim \mathcal{N}(160, 400)$.

1. $P(Y < 150) = P(Z < \frac{y-\mu}{\sigma}) = P(Z < \frac{150-160}{40}) = P(Z < -0.5) = 0.3085$. Thus, there is over about a 31% chance of failing to detect that the patient has high blood pressure if only a single measurement is taken.
2. Since $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$, $\bar{Y} \sim \mathcal{N}(160, 80)$. $P(\bar{Y} < 150) = P(Z < \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}) = P(Z < \frac{150-160}{8.944}) = P(Z < -1.12) = 0.1314$. Therefore, by using the average of five measurements, the chance of failing to detect the patient has high blood pressure has been reduced from over 31% to about 13%.
3. Given $P(\bar{Y} < 150) = 0.01$. Thus $P(Z < \frac{150-160}{40/\sqrt{n}}) = 0.01$ This implies $\frac{150-160}{40/\sqrt{n}} = -2.33$. Solving for n, yields $n = 21.64$. It would require at least 22 measurements in order to achieve the goal of at most a 1% chance of failing to detect high blood pressure.

- If the number of successes Y has a binomial distribution with number of trials n and probability of success π $\{Y \sim \text{Bin}(n\pi, n\pi(1-\pi))\}$, then the sample proportion of successes P will also have a normal distribution with mean π and variance $\pi(1-\pi)/n$, $\{P \sim \mathcal{N}(\pi, \pi(1-\pi)/n)\}$.

  - If the population proportion π is known, then the $z-$score of the sample proportion will have a standard normal distribution. That is,

$$Z = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}} \sim \mathcal{N}(0, 1).$$

– If the population proportion π is unknown, it can be approximated by the sample proportion p. Then, as long as the sample size n is large, the $z-$score, based on the estimated standard error, of the sample proportion will also have a standard normal distribution. That is:

$$Z = \frac{P - \pi}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0, 1).$$

The t Distribution

The t distribution describes the distribution of a normally distributed random variable for small sample size. Then, if σ is unknown and the sample size is small, then the standardized sample mean will have a t distribution with $n - 1$ degrees of freedom:

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t(n-1).$$

Similarly for the sample proportion, if π is unknown and the sample size is small , then the sample proportion will have a t distribution with $n - 1$ degrees of freedom:

$$T = \frac{P - \pi}{\sqrt{p(1-p)/n}} \sim t(n-1).$$

The χ^2 Distribution

The χ^2 describes the distribution of the sample variance.

The F Distribution

F distribution describes the distribution of the ratio of two variances.

7.5.4 Hypothesis Testing

A statistical hypothesis is an assumption (a conjecture) about a population parameter. Examples:

- Is the proportion of smokers among patients having a respiratory problem 0.72? Is $\pi = 0.72$?
- Is the mean SBP of patients under antihypertensive treatment in St. Paul's Hospital 120mmHg? Is $\mu = 120$mmHg?
- Is the current prevalence of HIV in Addis Ababa 25%? Is $\pi = 0.25$?

Such an assumption usually results from speculation concerning observed behavior, natural phenomena, or established theory. Hence, hypothesis testing is a statistical procedure that leads to take a decision about a statistical hypothesis for being supported or not by the sample data.

Statistical Hypothesis

A statistical hypothesis testing starts by making a set of two mutually-exclusive and exhaustive hypotheses about the parameter(s) in question. The first hypothesis is called a *null* hypothesis (denoted by H_0) which states there is no difference between a parameter and a hypothesized value. This hypothesis always states there is no difference, no effect, no impact, or no association. In other words, the statement under H_0 is similar to the notion of innocent until proven guilty. For any parameter θ and an assumed value θ_0, the null hypothesis is written as $H_0 : \theta = \theta_0$.

The second hypothesis, is called an *alternative* hypothesis (denoted by H_1), contradicts the null hypothesis and states there is a difference, an effect, an impact or an association. That is, it states there is a difference between a parameter and a hypothesized value. An alternative hypothesis has either of the following three different forms:

- Two-sided alternative; $H_1 : \ \theta \neq \theta_0$.
- One-sided (right tailed) alternative; $H_1 : \ \theta > \theta_0$.
- One-sided (left tailed) alternative; $H_1 : \ \theta < \theta_0$.

Errors in Hypothesis Testing

There are two types of errors in hypothesis testing.

- **Type I Error**: Type I error is an error occurred if one rejects the null hypothesis which is actually true. The probability of making Type I error is called *significance level* (denoted by α). Consequently, the probability of not rejecting a true null hypothesis is $1 - \alpha$ and called *confidence level*.
- **Type II Error**: Type II error is an error occurred if one failed to reject the null hypothesis which is actually false. The probability of making Type II error is denoted by β. The probability of correctly rejecting the null hypothesis which is actually false, called *power* of a test, is, therefore, $1 - \beta$. In other words, the power of a test is the probability of detecting a significant difference, if any, between θ and θ_0.

	Decision about H_0	
Null Hypothesis (H_0)	Do not reject H_0	Reject H_0
True	Correct decision	Type I error
False	Type II error	Correct decision

In statistical hypothesis testing, the maximum acceptable probability of rejecting a true null hypothesis, the significance level (α), is specified first. The common choices are $\alpha = 10\%$ (means 90% confidence level), $\alpha = 5\%$ (means 95% confidence level) and $\alpha = 1\%$ (means 99% confidence level).

Steps in Hypothesis Testing

A statistical hypothesis test can be formally summarized as a five-step process:

- **Step 1**: State both the null and alternative hypotheses, that is, H_0 and H_1.
- **Step 2**: Specify the maximum acceptable level of significance (α). Then, obtain the critical (tabulated) value (T_{tab}) which is used to define the *rejection* (critical) region of the null hypothesis (H_0). For a one sided test, the critical (tabulated) value is $T_{tab} = T_{\alpha}$. And for a two sided test, the critical (tabulated) value is $T_{tab} = T_{\alpha/2}$.
- **Step 3**: Define the appropriate test statistic and then find its calculated value (T_{cal}).
- **Step 4**: Decision about H_0. There are two possible methods for deciding whether to reject H_0 or not.
  - **Critical value method**: If the calculated value of the test statistic falls in the critical (rejection) region (that is, $|T_{cal}| \geq T_{tab}$), the null hypothesis can be rejected.
  - **The $p-$value method**: A p-value is the probability of obtaining values of a test statistic as extreme as that observed if the null hypothesis is true. For a one-sided test, $p-$value $= P(T \geq |T_{cal}|)$ and for a two-sided test, $p-$value $= 2 \times P(T \geq |T_{cal}|)$. If the $p -$ value $\leq \alpha$, then H_0 can be rejected.
- **Step 5**: Conclusion.

Notes: In a statistical test, the null hypothesis H_0 is rejected when there is sufficient evidence against it and the test is said to be *significant*. But if H_0 is not rejected, the test is said to be *insignificant* (*not significant*).

If a statistical test is found significant, we need to answer two or more questions.

1. What is the direction of the effect? Difference inferential questions (t test or analysis of variance) compare two or more groups so it is necessary to state which group performed better. For associational inferential questions (eg, correlation, regression), the sign is very important, so we must indicate whether the relationship is positive or negative.
2. What is the size of the effect? We should include the effect size, confidence intervals or both in the description of our results.
3. With large samples, it is possible to find statistical significance even when the difference is very small (i.e., has a small effect size). A significant result with a small effect size means that it is sure that there is at least a little difference, but it may not be of any practical importance. Hence, a statistical significant test does not indicate practical (clinical) importance (significance). Therefore, the researcher or consumer of the research should make a judgment whether the result has practical (clinical) importance

(significance). To do so, they need to take into account the effect size, the cost of implementing the change, and the probability and severity of any side effect or unintended consequence.

Chapter 8

Inference for Continuous Responses

8.1 Inference about a Single Population Mean

Suppose a sample is selected from a single group (population) and there is one quantitative variable which is assumed to be normally distributed.

That is, given a random sample $y_1, y_2, \cdots, y_n$ of size n from a normal population with mean μ and variance σ^2; $Y \sim \mathcal{N}(\mu, \sigma^2)$.

The point estimator of the population mean μ is the sample mean $\bar{y}$ ($\bar{y}$ estimates μ). The mean of the sampling distribution of the sample mean $\bar{y}$ is $E(\bar{y}) = \mu$. Also, the variance of the sample mean is $\widehat{V}(\bar{y}) = \sigma^2_{\bar{y}} = \sigma^2/n$. Thus, the sampling distribution of the sample mean is identical as normal with mean μ_1 and variance σ^2/n; $\bar{y} \sim \mathcal{N}(\mu, \sigma^2/n)$.

Therefore, the standard error of the sample mean $\bar{y}$ is $\text{SE}(\bar{y}) = \sigma_{\bar{y}} = \sigma/\sqrt{n}$. Consequently, the estimated standard error of the mean is $\widehat{\text{SE}}(\bar{y}) = \hat{\sigma}_{\bar{y}} = s/\sqrt{n}$.

8.1.1 Testing for a Population Mean μ

The interest here is whether the population average μ of the variable of interest Y takes a particular value, say μ_0.

Step 1: State both the null and alternative hypotheses. There three options are:

Option 1: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ - two sided test

Option 2: $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$ - one sided (left tailed) test

Option 3: $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ - one sided (right tailed) test

Step 2: Specify the level of significance α and obtain the corresponding critical (tabulated) value. The critical (tabulated) value for a two sided test is $T_{tab} = T_{\alpha/2}$ whereas the critical (tabulated) value for a one sided test is $T_{tab} = T_{\alpha}$.

Step 3: Use the appropriate test statistic and obtain its calculated value T_{cal}. Here, there are three possible cases for selecting the appropriate test statistic:

Case 1: When σ is known. If σ is known, the appropriate test statistic is the z test statistic. That is,

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Case 2: When σ is not known but n is large ($n \geq 30$). If σ is not known but n is large, again the appropriate test statistic is the z test statistic. That is defined as:

$$Z = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Case 3: When σ is not known and n is small ($n < 30$). If σ is not known and n is small, the appropriate test statistic is t statistic with $n - 1$ degrees of freedom. That is,

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t(n - 1).$$

Step 4: Decision: If $|T_{cal}| \geq T_{tab}$ or $p-$value $< \alpha$, H_0 can be rejected.

Step 5: Conclusion.

Example 8.1. The life expectancy of HIV/AIDS patients is expected to be 50 years. A survey was conducted on eleven patients of a certain hospital and the data obtained as: 54.2, 58.2, 56.6, 50.4, 44.2, 61.9, 57.5, 49.7, 55.4, 53.4, 57.0. Does the data confirm the expected view?

Solution: Sample size $n = 11$, sample mean $\bar{y} = 54.41$, sample variance $s^2 = 23.607$, sample standard deviation $s = 4.859$.

- Let μ be the mean life expectancy of HIV/AIDS patients.

Step 1: Hypothesis:

$H_0 : \mu = 50$. The mean life expectancy of HIV/AIDS patients is not significantly different from 50 years.

$H_1 : \mu \neq 50$. The mean life expectancy of HIV/AIDS patients is significantly different from 50 years.

Step 2: Let us assume $\alpha = 0.05$. Since the sample size $n < 30$, the critical (tabulated) value is determined using the t distribution. Thus, $t_{tab} = t_{\alpha/2}(n - 1) = t_{0.025}(10) = 2.228$.

Step 3: The calculated test statistic is $t_{cal} = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{54.41-50}{4.859/\sqrt{11}} = 3.01$.

Step 4: Decision: Since $|t_{cal}| = 3.01 > t_{tab} = t_{0.025}(10) = 2.228$, H_0 should be rejected. Or using the p-value method, p-value$=2 \times P[t(10) > 3.01] = 2 \times 0.0066 = 0.0132$ which is less than $\alpha = 0.05$, H_0 can be rejected.

Step 5: Conclusion: Therefore, the conclusion is there is sufficient evidence to reject the null hypothesis and hence "the mean life expectancy of HIV/AIDS patients is significantly different from 50 years at 5% significance level". In particular, since the difference is positive, 'the mean life expectancy of HIV/AIDS patients is significantly larger than 50 years at $\alpha = 5\%$'.

Example 8.2. The thermostat in a classroom is set at 72°F, but we think the thermostat is not working well. On seven randomly selected days, we measure the temperature at our seat. The measurements (in degree Fahrenheit) are 71, 73, 69, 68, 69, 70, and 71. Test whether the mean temperature at our seat is different from 72°F at 99% confidence level.

Solution: Sample size $n = 7$, sample mean $\bar{y} = 70.14$, sample variance $s^2 = 2.81$, sample standard deviation $s = 1.68$.

- Let μ be the mean temperature (in degree Fahrenheit) in the classroom.

Step 1: Hypothesis:

$H_0 : \mu = 72°F$. The mean temperature of the classroom is not significantly different from $72°F$.

$H_1 : \mu \neq 72°F$. The mean temperature of the classroom is significantly different from $72°F$.

Step 2: It is given $\alpha = 0.01$. Since the sample size is $n < 30$, the critical value is determined using the t distribution. Thus, $t_{tab} = t_{\alpha/2}(n-1) = t_{0.005}(6) = 3.707$.

Step 3: The calculated test statistic is $t_{cal} = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{70.14-72}{1.68/\sqrt{7}} = -2.929$.

Step 4: Decision: Since $|t_{cal}| = 2.929 < t_{tab} = t_{0.005}(6) = 3.707$, H_0 cannot be rejected. Or p-value$=2 \times P[t(6) > 2.929] = 2 \times 0.0132 = 0.0264$ which is greater than $\alpha = 0.01$, H_0 cannot be rejected.

Step 5: Conclusion: Therefore, there is not sufficient evidence to say "the mean temperature of the class room is significantly different from 72°F at 1% significance level". In other words, "the mean temperature of the class room is not significantly different from 72°F at 1% significance level".

Example 8.3. A researcher is interested to examine the cholesterol levels in women of aged 18-30 but s/he did not have any prior information about the distribution of cholesterol levels in such women. It was known that the distribution of cholesterol levels in women aged 31-50 is known to be approximately normal with a mean of 190 mg/dL. Thus, the researcher wanted to evaluate whether the mean cholesterol level differed from the mean cholesterol level of middle-aged women. A random sample of 100 females aged 18-30 was selected and the selected females were administered blood tests that yielded cholesterol levels having a mean of 178.2 mg/dL and a standard deviation of 45.3 mg/dL. Is there significant evidence in the data to demonstrate that the mean cholesterol level of females of aged 18-30 differs from 190 mg/dL?

Solution: Let μ be the mean cholesterol level of females of aged 18-30.

Step 1: Hypothesis:

$H_0 : \mu = 190$. The mean cholesterol level of females of aged 18-30 is not significantly different from 190 mg/dL.

$H_1 : \mu \neq 190$. The mean cholesterol level of females of aged 18-30 is significantly different from the mean cholesterol level of females of aged 31-50.

Step 2: Assume $\alpha = 0.05$ level of significance. Since the sample size is large ($n = 100 > 30$), the critical (tabulated) value is determined using the z distribution, $z_{tab} = z_{\alpha/2} = z_{0.025} = 1.96$.

Step 3: The calculated test statistic is $z_{cal} = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{178.2-190}{45.3/\sqrt{100}} = -2.60$.

Step 4: Decision: Since $|z_{cal}| = 2.60 > z_{tab} = z_{0.025} = 1.96$, H_0 can be rejected. Or using the $p-$value method, $p-$value$= 2 \times P(Z > 2.60) = 2 \times 0.0047 = 0.0094$. This $p-$value is less than $\alpha = 0.05$ indicating sufficient evidence against H_0.

Step 5: Conclusion: Therefore, the mean cholesterol level of females of aged 18-30 is significantly different from 190 mg/dL at 5% significance level. In particular, since the difference is negative, the mean cholesterol level of females of aged 18-30 is significantly less than 190 mg/dL at $\alpha = 5\%$.

Example 8.4. As the public concern about bacterial infections increases, a soap manufacturer quickly promoted a new product to meet the demand for an antibacterial soap. This new product has a substantially higher price than the 'ordinary soaps'on the market. A consumer testing agency notes that ordinary soap also kills bacteria and questions whether the new antibacterial soap is a substantial improvement over the ordinary soap. From previous studies using many different brands of ordinary soaps, the mean bacteria count is 33 for ordinary soap products. The consumer group runs a test on the antibacterial soap using 35 petri dishes and it yielded a mean bacterial count of 31.2 with a standard deviation of 8.4. Do the data provide sufficient evidence that the antibacterial soap is more effective than the ordinary soap in reducing bacteria counts?

Solution: Let μ be the mean bacterial count for the antibacterial soap.

Step 1: Hypothesis:

$H_0 : \mu = 33$. The mean bacterial count for the antibacterial soap is not significantly different from 33.

$H_1 : \mu < 33$. The mean bacterial count for the antibacterial soap is not significantly different from 33 (the antibacterial soap is more effective than the ordinary soap).

Step 2: Assume $\alpha = 0.05$ level of significance. Since the sample size is relatively large ($n = 35 > 30$), the critical (tabulated) value is determined using the z distribution, $z_{tab} = z_{\alpha} = z_{0.05} = 1.645$.

Step 3: The calculated value of the test statistic is $z_{cal} = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{31.2-33}{8.4/\sqrt{35}} = -1.27$.

Step 4: Decision: Since $|z_{cal}| = 1.27 < z_{tab} = z_{0.05} = 1.645$, H_0 cannot rejected. Or using the $p-$value method, $p-$value $= P(Z > 1.27) = 0.1020$. Clearly, this $p-$value is greater than $\alpha = 0.05$ indicating no sufficient evidence against H_0.

Step 5: Conclusion: Therefore, the conclusion is "the mean bacterial count for the antibacterial soap is not significantly less than 33 at 5% significance level". In other words, "there is not sufficient evidence that the antibacterial soap is more effective than the ordinary soap at 5% significance level".

Example 8.5. A research done by a graduating student reports that the average score of SPHMMC students in biostatistics course is greater than 80. To test this claim, a random sample of 10 students was taken and their scores in the course are recorded as: 78, 88, 93, 68, 98, 73, 88, 83, 93, 98. At 0.05 level of significance, test the validity of this claim.

Solution: Sample size $n = 10$, sample mean $\bar{y}86$, sample variance $s^2 = 106.67$, sample standard deviation $s = 10.33$.

- Let μ be the mean score of SPHMMC students in biostatistics course.

Step 1: Hypothesis:

$H_0 : \mu = 80$. The mean score of SPHMMC students in biostatistics course is not significantly different from 80.

$H_1 : \mu > 80$. The mean score of SPHMMC students in biostatistics course is significantly greater than 80.

Step 2: Given $\alpha = 0.05$. Since the sample size $n < 30$, the critical (tabulated) value is determined using the t distribution. Thus, $t_{tab} = t_\alpha(n-1) = t_{0.05}(9) = 1.833$.

Step 3: The calculated test statistic is $t_{cal} = \frac{\bar{y}-\mu}{s/\sqrt{n}} = \frac{86-80}{10.33/\sqrt{10}} = 1.837$.

Step 4: Decision: Since $|t_{cal}| = 1.837 > t_{tab} = t_{0.05}(9) = 1.833$, H_0 can be rejected. Or using the p-value method, p-value$=P[t(9) > 1.837] = 0.0497$ which is less than $\alpha = 0.05$, H_0 can be rejected.

Step 5: Conclusion: Therefore, it can be concluded that the average score of SPHMMC students in biostatistics course is significantly greater than 80 at 5% level of significance.

8.1.2 Interval Estimation for a Population Mean μ

A statistical test merely indicates whether a particular value for a parameter is plausible or not. The construction of a confidence interval determines the range of plausible values for which H_0 is "not rejected".

The $(1-\alpha)100\%$ confidence interval for a population mean μ is constructed by solving the equation

$$P\left(\left|\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} < \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = (1-\alpha)100\%$$

for μ. The confidence interval for μ is given by $P\left(\bar{y} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right) = (1-\alpha)100\%$.

Case 1: When σ is known. The $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Case 2: When σ is not known but n is large ($n \geq 30$). The $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\left(\bar{y} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{y} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Case 3: When σ is not known and n is small ($n < 30$). The $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\left[\bar{y} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{y} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right]$$

Note: The half-width (plus-or-minus term) of the confidence interval for a population mean μ is $d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and called *margin of error*. And, the entire width of the confidence interval $w = 2d$ is also called *tolerable error*.

Example 8.6. Consider example 8.1. Find the point estimate for the mean life expectancy of HIV/AIDS patients and also construct the 95% confidence interval. What is the value of the margin of error and tolerable error?

Solution: From the example point estimate of the population mean life expectancy of HIV/AIDS patients is $\bar{y} = 54.41$. The sample standard deviation of the life expectancy of HIV/AIDS patients is $s = 4.859$.

Here, the sample size is small, the critical value corresponding to the 95% confidence level is $t_{tab} = t_{\alpha/2}(n-1) = t_{0.025}(10) = 2.228$. Therefore, the 95% confidence interval for the population mean life expectancy of HIV/AIDS patients is:

$$\begin{aligned}\left[\bar{y} \pm t_{0.025}(10)\frac{s}{\sqrt{n}}\right] &= \left[54.41 - 2.228\left(\frac{4.859}{\sqrt{11}}\right),\ 54.41 + 2.228\left(\frac{4.859}{\sqrt{11}}\right)\right] \\ &= (54.41 - 3.2641,\ 54.41 + 3.2641) \\ &= (51.15, 57.67)\end{aligned}$$

Hence, the margin of error is $d = 2.228\left(\frac{4.859}{\sqrt{11}}\right) = 3.2641$ and the tolerable error is $w = 2d = 6.5282$.

Example 8.7. Recall example 8.3 and construct the 95% confidence interval for the mean cholesterol level in women of aged 18-30. Also, determine the value of the margin of error and tolerable error.

Solution: The point estimate of the population mean cholesterol level in women of aged 18-30 is $\bar{y} = 178.2$ mg/dL. The sample standard deviation of the cholesterol levels in women of aged 18-30 is $s = 45.3$ mg/dL.

Since the sample size is large, the critical value corresponding to the 95% confidence level is $z_{\alpha/2} = z_{0.025} = 1.96$. Therefore, the 95% confidence interval for the mean cholesterol level in women of aged 18-30 is:

$$\left(\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\right) = \left[178.2 - 1.96\left(\frac{45.3}{\sqrt{100}}\right),\ 178.2 + 1.96\left(\frac{45.3}{\sqrt{100}}\right)\right] = (169.321,\ 187.079).$$

Hence, the margin of error is $d = 1.96\left(\frac{45.3}{\sqrt{100}}\right) = 8.879$ and the tolerable error is $w = 2d = 17.758$.

Note: If the $(1-\alpha)100\%$ confidence interval for a population mean of a two-sided test includes the hypothesized value μ_0, the null hypothesis should not be rejected at α level of significance. If, however, the confidence interval does not include μ_0, then the null hypothesis should be rejected at that level.

Example 8.8. Recall example 8.5. Construct the 95% confidence interval for the population mean score of SPHMMC students in biostatistics course.

Solution: It is given $\bar{y} = 86$ and $s = 10.33$. Since the sample size is small, the critical value corresponding to the 95% confidence level is $t_{\alpha/2}(10-1) = t_{0.025}(9) = 2.262$. Consequently, the 95% confidence interval is for the population mean score of SPHMMC students in biostatistics course is:

$$\left[\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right] = \left[86 - 2.262\left(\frac{10.33}{\sqrt{10}}\right),\ 86 + 2.262\left(\frac{10.33}{\sqrt{10}}\right)\right] = (78.611,\ 93.389).$$

8.2 Comparing Two Population Means: Paired Samples

The inferences in the previous section have concerned about a parameter from a single population. There may be situations that need comparison of parameters from two or more groups (populations). In studies involving the comparison of two groups, there are two ways of taking the samples: paired and independent.

In paired comparison, the presence and absence of a single treatment or two treatments are compared. Here a pair of *same* (e.g., persons) or *identical* (e.g., plots) experimental units are selected, and one type of treatment is applied on one member of each pair and another type of treatment is applied on the second member of each pair. Then the response of interest is recorded on each pair. The paired responses are then analysed by computing their differences.

A common application occurs when the response is measured on two different occasions (appropriate for pre-post treatment responses). The aim of pairing sample is to make the comparison more accurate by having experimental units in each pair as likely as possible except the treatment difference.

Suppose there are two paired normally distributed random variables y_1 and y_2 with mean μ_1 and μ_2, and variance σ_1^2 and σ_2^2, respectively. Thus, the difference of the two variables, $d_i = y_{1i} - y_{2i}; i = 1, 2, \cdots, n$, is treated as if it were a single sample.

If the observed differences are $d_i = y_{1i} - y_{2i}$; $i = 1, 2, \cdots, n$, then the population mean of the differences μ_d is estimated by the sample mean of the differences $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$. Similarly, the population variance of the differences σ_d^2 is estimated by sample variance of the differences $s_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$.

The point estimator of the population mean of the differences μ_d is the sample mean of the differences $\bar{d}$ ($\bar{d}$ estimates μ_d). The mean of the sampling distribution of the difference $\bar{d}$ is $E(\bar{d}) = \mu_d$. Also, the variance of the differences is $\sigma^2_{\bar{d}} = \sigma^2_d/n$. Thus, the sampling distribution of the difference $d = y_1 - y_2$ is normal with mean μ_d and variance σ^2_d/n, that is, $\bar{d} \sim \mathcal{N}(\mu_d, \sigma^2_d/n)$.

The standard error of the sample mean of the differences $\bar{d}$ is $\sigma_{\bar{d}} = \sigma_d/\sqrt{n}$. Consequently, the estimated standard error of the differences is $\hat{\sigma}_{\bar{d}} = s_d/\sqrt{n}$.

8.2.1 Testing for the Population Mean of the Differences μ_d

The interest in paired samples is whether the average of the differences μ_d of the variables of interest Y_1 and Y_2 takes a particular value, say μ_{d0}. The difference is typically assumed to be zero unless explicitly specified.

The steps to be followed is similar to the one used for the one sample case:

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \mu_d = 0$ vs $H_1 : \mu_d \neq 0$ - two sided test

Option 2: $H_0 : \mu_d = 0$ vs $H_1 : \mu_d < 0$ - one sided (left tailed) test

Option 3: $H_0 : \mu_d = 0$ vs $H_1 : \mu_d > 0$ - one sided (right tailed) test

Step 2: Specify the level of significance α and obtain the critical (tabulated) value. The critical (tabulated) value for a two sided test is $T_{tab} = T_{\alpha/2}$ whereas the critical value for a one sided test is $T_{tab} = T_\alpha$.

Step 3: Use the appropriate test statistic and obtain its calculated value T_{cal}. As usual, there are three possible cases for selecting the appropriate test statistic.

Case 1: When σ_d is known. If σ_d is known, the appropriate test statistic is the z test statistic. That is,

$$Z = \frac{\bar{d} - \mu_d}{\sigma_d/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Case 2: When σ_d is not known but n is large ($n \geq 30$). If σ_d is not known but n is large, again the appropriate test statistic is the z test statistic. That is defined as:

$$Z = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Case 3: When σ_d is not known and n is small ($n < 30$). If σ_d is not known and n is small, the appropriate test statistic is t statistic with $n - 1$ degrees of freedom. That is,

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \sim t(n - 1).$$

Step 4: Decision: If $|T_{cal}| \geq T_{tab}$ or $p-$value $< \alpha$, H_0 can be rejected.

Step 5: Conclusion.

Example 8.9. A medical researcher wishes to determine if a pill has an effect on reducing the blood pressure of individuals. The study involves recording the initial blood pressure of 15 women. After they took the pill for six months, their blood pressure are again recorded. The data is:

Women	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before (y_{1i})	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
After (y_{2i})	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74

Do the data substantiate the claim that the pill reduced blood pressure?

Solution: Let μ_d be the population mean of the differences in the blood pressure of women. The observed differences of the before - after blood pressures are:

Women	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$d_i = y_{1i} - y_{2i}$	2	8	10	6	18	10	4	26	18	-8	0	32	0	-4	10

The sample mean of the differences is $\bar{d} = \frac{1}{15}(2 + 8 + \cdots + 10) = \frac{1}{15}(132) = 8.8$. The sample variance of the differences is $s_d^2 = \frac{1}{15-1}\{(2-8.8)^2+(8-8.8)^2+\cdots+(10-8.8)^2\} = \frac{1}{14}(1686.4) = 120.457$. This implies the sample standard deviation of the differences is $s_d = 10.98$.

Step 1: Hypothesis:

$H_0 : \mu_d = 0$. The mean of the differences in the blood pressure of women is not significantly different from 0.

$H_1 : \mu_d > 0$. The mean of the differences in the blood pressure of women is significantly larger than 0 (the pill has a significant decreasing effect in the blood pressure of women).

Step 2: Assume $\alpha = 0.05$. Since the sample size $n < 30$, the critical value is determined using the t distribution. Thus, $t_{tab} = t_\alpha(n-1) = t_{0.05}(14) = 1.761$.

Step 3: The calculated test statistic is $t_{cal} = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} = \frac{8.8-0}{10.98/\sqrt{15}} = 3.10$.

Step 4: Decision: Since $|t_{cal}| = 3.10 > t_{tab} = t_{0.05}(14) = 1.761$, H_0 can be rejected. Or $p-$value $= P[t(14) > 3.10] = 0.0039$ less than $\alpha = 0.05$.

Step 5: Conclusion: Therefore, there is sufficient evidence to reject the null hypothesis and conclude that "the mean of the differences in the blood pressure of women is significantly larger than 0 at 5% level of significance". That is, the pill has a significant decreasing effect in the blood pressure of women at $\alpha = 5\%$.

8.2.2 Interval Estimation of the Population Mean of the Differences μ_d

The $(1-\alpha)100\%$ confidence interval for the population mean of the differences μ_d is constructed by solving the equation

$$P\left(\left|\frac{\bar{d}-\mu_d}{\sigma_d/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} < \frac{\bar{d}-\mu_d}{\sigma_d/\sqrt{n}} < z_{\alpha/2}\right) = (1-\alpha)100\%$$

for μ_d. The confidence interval for μ_d is given by $P\left[\bar{d} - z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right) < \mu_d < \bar{d} + z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)\right] = (1-\alpha)100\%$.

Case 1: When σ_d is known. The $(1-\alpha)100\%$ confidence interval for μ_d is given by:

$$\left[\bar{d} - z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right),\ \bar{d} + z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)\right]$$

Case 2: When σ_d is not known but n is large ($n \geq 30$). The $(1-\alpha)100\%$ confidence interval for μ_d is given by:

$$\left[\bar{d} - z_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right),\ \bar{d} + z_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)\right]$$

Case 3: When σ_d is not known and n is small ($n < 30$). The $(1-\alpha)100\%$ confidence interval for μ_d is given by:

$$\left[\bar{d} - t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}},\ \bar{d} + t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}\right]$$

Example 8.10. Construct the 95% confidence interval for the mean difference of blood pressure, given in example 8.9.

Solution: The 95% confidence interval is:

$$\left(\bar{d} \pm t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}\right) = \left[8.8 - 2.145\left(\frac{10.98}{\sqrt{15}}\right),\ 8.8 + 2.145\left(\frac{10.98}{\sqrt{15}}\right)\right] = (2.72,\ 14.88).$$

Note: If the $(1-\alpha)100\%$ confidence interval for the population mean difference μ_d includes 0, the null hypothesis of the two-sided test cannot be rejected at α level of significance. If, however, the confidence interval does not include 0, then the null hypothesis of no difference can be rejected at that level.

8.3 Comparing Two Population Means: Independent Samples

Suppose a random sample of $y_{11}, y_{12}, \cdots, y_{1n_1}$ is drawn from a normal population with mean μ_1 and variance σ_1^2, and another random sample $y_{21}, y_{22}, \cdots, y_{2n_2}$ is drawn from a normal population with mean μ_2 and variance σ_2^2. Thus, $\bar{y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1)$ and $\bar{y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2)$, where $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the samples drawn from the first and second populations, respectively.

The interest in such independent samples is to test whether the difference of the population means is zero, that is, $\mu_1 - \mu_2 = 0$.

Thus, the difference of the sample means $\bar{y}_1 - \bar{y}_2$ is an estimator for the difference of the population means $\mu_1 - \mu_2$ ($\bar{y}_1 - \bar{y}_2$ estimates $\mu_1 - \mu_2$). The mean of the sampling distribution of the difference of the sample means $\bar{y}_1 - \bar{y}_2$ is $E(\bar{y}_1 - \bar{y}_2) = \mu_1 - \mu_2$. Also, the variance of the difference of the sample means is $\widehat{V}(\bar{y}_1 - \bar{y}_2) = \sigma^2_{\bar{y}_1 - \bar{y}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Thus, the sampling distribution of the sample means difference $\bar{y}_1 - \bar{y}_2$ is identical as being normal with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, that is, $\bar{y}_1 - \bar{y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Therefore, the standard error of the sample means difference $\bar{y}_1 - \bar{y}_2$ is $\text{SE}(\bar{y}_1 - \bar{y}_2) = \sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The estimated standard error of the sample means difference $\bar{y}_1 - \bar{y}_2$ is $\widehat{\text{SE}}(\bar{y}_1 - \bar{y}_2) = \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Note that $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2$ is the sample variance of the first group and $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$ is the sample variance of the second group; and the sample mean of the first group is $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}$ and the sample mean of the second group is $\bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i}$.

If the two groups have equal population variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, that is, $\bar{y}_1 \sim \mathcal{N}(\mu_1, \sigma^2/n_1)$ and $\bar{y}_2 \sim \mathcal{N}(\mu_2, \sigma^2/n_2)$, then $\bar{y}_1 - \bar{y}_2 \sim \mathcal{N}\left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$.

Now the standard error of the sample means difference $\bar{y}_1 - \bar{y}_2$ is $\text{SE}(\bar{y}_1 - \bar{y}_2) = \sigma_{\bar{y}_1 - \bar{y}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. The estimated standard error of the sample means difference $\bar{y}_1 - \bar{y}_2$ is $\widehat{\text{SE}}(\bar{y}_1 - \bar{y}_2) = \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where s_{pooled} is the pooled standard deviation of both groups. That is, $s_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is an estimate of the common variance σ^2.

8.3.1 Testing for the Difference of Two Population Means

Case 1: When σ_1^2 and σ_2^2 are equal

Step 1: State both the null and alternative hypotheses. There three possible options are:

Options 1: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ - two sided test
Options 2: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$ - one sided (left tailed) test
Options 3: $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ - one sided (right tailed) test

Step 2: Specify the level of significance α and obtain the critical value.

- If both sample sizes (n_1 and n_2) are large, the critical value (tabulated) for a two sided test is $z_{tab} = z_{\alpha/2}$ whereas the critical value for a one sided test is $z_{tab} = z_\alpha$.
- If the sample sizes (n_1 and n_2) are small, the critical (tabulated) value for a two sided test is $t_{tab} = t_{\alpha/2}(n_1 + n_2 - 2)$ whereas the critical value for a one sided test is $t_{tab} = t_\alpha(n_1 + n_2 - 2)$.

Step 3: Use the appropriate test statistic and obtain the calculated value.

- If the sample sizes are large, the z test statistic is used. That is,

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0, 1)$$

- If the sample sizes are small, the t test statistic is used. That is.

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Step 4: Decision: H_0 can be rejected if $|T_{cal}| > T_{tab}$ or $p-\text{value} < \alpha$.

Step 5: Conclusion.

Example 8.11. Company officials were concerned about the length of time a particular drug product retained its potency. A random sample of 8 bottles of the product was drawn from the production line and measured for potency. A second sample of 10 bottles was obtained and stored in a regulated environment for a period of one year. The potency readings obtained from each sample are given below.

Sample 1 (Fresh)	10.2	10.5	10.3	10.8	9.8	10.6	10.7	10.2		
Sample 2 (Stored)	9.8	9.6	10.1	10.2	10.1	9.7	9.5	9.6	9.8	9.9

Test the null hypothesis that the drug product retains its potency.

Solution: Let μ_1 be the mean potency of the (fresh) drug product taken from the production line and μ_2 be the mean potency of the (stored) drug product that was retained for a year. The summary statistics of the data are:

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \frac{1}{8}(83.1) = 10.388 \text{ and } \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \frac{1}{10}(98.3) = 9.830$$

$$s_1^2 = \frac{1}{n_1}\left[\sum_{i=1}^{n_1} y_{1i}^2 - \frac{1}{n_1}(\sum_{i=1}^{n_1} y_{1i})^2\right] = \frac{1}{7}[863.95 - \frac{1}{8}(83.1)^2] = 0.107$$

$$s_2^2 = \frac{1}{n_2}\left[\sum_{i=1}^{n_2} y_{2i}^2 - \frac{1}{n_2}(\sum_{i=1}^{n_2} y_{2i})^2\right] = \frac{1}{9}[966.81 - \frac{1}{10}(98.3)^2] = 0.058$$

$$s_{pooled}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(8-1)(0.107)+(10-1)(0.058)}{8+10-2} = 0.079$$

$$s_{pooled} = 0.281.$$

Step 1: Hypothesis:

$H_0 : \mu_1 = \mu_2$. The mean potency of the drug product taken from the production line and the mean potency of the drug product stored for a period of one year are not significantly different.

$H_1 : \mu_1 \neq \mu_2$. The mean potency of the drug product taken from the production line and the mean potency of the drug product stored for a period of one year are significantly different.

Step 2: Assume $\alpha = 0.05$. Since the sample sizes are small, the critical value is $t_{tab} = t_{\alpha/2}(n_1+n_2-2) = t_{0.025}(16) = 2.12$.

Step 3: The calculated test statistic is $t = \frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(10.388-9.830)-0}{0.281\sqrt{\frac{1}{8}+\frac{1}{10}}} = 4.186$.

Step 4: Decision: Since $|t_{cal}| = 4.186 > t_{tab} = 2.12$, H_0 can be rejected. Or $p-$value $= 2 \times P[t(16) > 4.186] = 2 \times 0.0003 = 0.0006 < \alpha = 0.05$.

Step 5: Therefore, there is a significant difference in the mean potency of the drug product from the production line and the drug that was retained for one year at $\alpha = 5\%$. In particular, the mean potency of the drug product that was retained for one year is significantly lower than the one taken from the production line.

Case 2: When σ_1^2 and σ_2^2 are not equal

The above test statistic is only used when the two distributions have the same variance. When the variances are different ($\sigma_1^2 \neq \sigma_2^2$), they are estimated separately. That is, the sample variance of the first group s_1^2 estimates σ_1^2 and the sample variance of the second group s_2^2 estimates σ_2^2. Therefore, the z test statistic for large sample sizes is:

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

For small sample sizes, the t test statistic is given as:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(v)$$

where v is the degrees of freedom defined as $v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$.

Example 8.12. A quick but impressive method of estimating the concentration of a chemical in a rat has been developed. The sample from this method has 8 observations and the sample from the standard method has 4 observations. Assuming different population variances, test whether the quick method gives under-estimate result. The data in the two samples are:

Standard Method	25	24	25	26				
Quick Method	23	18	22	28	17	25	19	16

Solution: Let μ_1 be the population mean concentration of the chemical from the standard method and μ_2 be the population mean concentration of the chemical from the quick method.

Step 1: Hypothesis:

$H_0 : \mu_1 = \mu_2$. The means of the concentrations of the chemical in the standard and quick methods are not significantly different.

$H_1 : \mu_1 > \mu_2$. The mean concentrations of the chemical in the standard method is significantly larger than that of the quick method.

Step 2: Assume $\alpha = 0.05$. Since the samples are small, t distribution will be used. The degrees of freedom for unequal variances assumption is $v = \frac{(0.67/4 + 17.71/8)^2}{(0.67/4)^2/(4-1) + (17.71/8)^2/(8-1)} \approx 8$. Thus, $t_{tab} = t_{0.05}(8) = 1.86$.

Step 3: The calculated test statistic is $t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(25-21)-0}{\sqrt{\frac{0.67}{4} + \frac{17.71}{8}}} = 2.60$.

Step 4: Decision: Since $|t_{cal}| = 2.60 > t_{tab} = 1.86$, H_0 can be rejected. Or $p-$value $= P[t(8) > 2.60] = 0.0158 < \alpha = 0.05$.

Step 5: Conclusion: Therefore, the quick method gives an under-estimate result at $\alpha = 5\%$.

Example 8.13. For a random sample of 120 adult female born in country A, the mean height was 62.7 inches with standard deviation 2.50 inches. For another random sample of 150 adult female born in country B the mean height was 61.8 inches with standard deviation 2.62 inches. Would you reject the null hypothesis that there is no difference in height between adult female born in the two countries at 1% level of significance.

Solution: Let μ_1 be the mean height of adult female born in country A and μ_2 be the mean height of adult female born in country B.

Step 1: Hypothesis:

$H_0 : \mu_1 = \mu_2$. There is no significant difference in the mean heights of adult females born in the two countries.

$H_1 : \mu_1 \neq \mu_2$. There is a significant difference in the mean heights of adult females born in the two countries.

Step 2: Given $\alpha = 0.01$. Since, the sample sizes are large, $z_{tab} = z_{0.01/2} = z_{0.005} = 2.58$.

Step 3: The calculated test statistic is $z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(62.7 - 61.8) - 0}{\sqrt{\frac{(2.50)^2}{120} + \frac{(2.62)^2}{150}}} = 2.88$.

Step 4: Decision: Since $|z_{cal}| = 2.88 > z_{tab} = 2.58$, H_0 can be rejected. Or $p-$value $= 2 \times P(Z > 2.88) = 2 \times 0.0020 = 0.0040 < \alpha = 0.01$.

Step 5: Conclusion: Therefore, there is a difference in the population mean heights of adult females in the two countries at $\alpha = 1\%$. Particularly, the mean height of adult females born in country A is higher than the mean height of adult females born in country B at $\alpha = 1\%$.

8.3.2 Interval Estimation for the Difference of Population Means $\mu_1 - \mu_2$

The $(1 - \alpha)100\%$ confidence interval for the difference of the population means under the common variance assumption is:

$$\left[(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\ s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Similarly, the $(1-\alpha)100\%$ confidence interval for the difference of the population means when the population variances are different is:

$$\left[(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right].$$

For small sample sizes, $z_{\alpha/2}$ is replaced by $t_{\alpha/2}(n_1 + n_2 - 2)$ and $t_{\alpha/2}(v)$ in the above two equations respectively.

Example 8.14. Construct the 95% confidence interval for the difference of the population means of the potency of the two types of drugs (fresh and stored) considered in example 8.11.

Solution: The 95% confidence interval for the difference of the population means, $\mu_1 - \mu_2$ is:

$$\left[(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)\ s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = \left[(10.388 - 9.830) \pm 2.12(0.281)\sqrt{\frac{1}{8} + \frac{1}{10}}\right]$$
$$= (0.275, 0.841).$$

Note: If the $(1 - \alpha)100\%$ confidence interval for the difference of the population means includes 0, the null hypothesis of no difference of the two means of a two-sided test cannot be rejected at α level of significance. If, however, the confidence interval does not include 0, then the two means are significantly different at that level.

8.4 Comparison of More Than Two Population Means

The t and Z tests have been used for testing the hypothesis of a single population mean equal to a specified value or equality of two populations means when the sample size is small and large respectively. However in testing the equality of more than two population means, the technique of *analysis of variance* (ANOVA) will be used.

8.4.1 Analysis of Variance (ANOVA)

Suppose random samples representing a certain continuous (quantitative) variable are selected from g different groups (populations). Assume the samples are independently and randomly drawn from a reasonably normal distribution and have approximately equal variances.

Group 1	$y_{11}, y_{21}, \cdots, y_{n_1 1} \sim \mathcal{N}(\mu_1, \sigma^2)$
Group 2	$y_{12}, y_{22}, \cdots, y_{n_2 2} \sim \mathcal{N}(\mu_2, \sigma^2)$
$\vdots$	$\vdots$
Group g	$y_{1g}, y_{2g}, \cdots, y_{n_g g} \sim \mathcal{N}(\mu_g, \sigma^2)$

The interest is whether all the g group means are equal. For this purpose, ANOVA is used. The principle in ANOVA is decomposing the total variability of the dataset into two components representing different source of variation; one is the variability *within* groups and the other is the variability *between* groups. The *within* groups variation is the variation of the observation in each group resulting from uncontrolled biological variation, technical and measurement errors; and the *between* variation is the variation among the group means.

The null hypothesis to be tested is that the all the g group means are equal and the alternative hypothesis is at least one of the group mean is significantly different from the other. That is,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$$
$$H_1 : \text{not } H_0$$

To construct the test statistics, the total variation of the dataset (called *total sum squares* - TSS) is decomposed into the between group variation (called *between sum squares* - BSS)

and the within group variation (called *within sum squares* - WSS).

$$\begin{aligned}
y_{ij} - \bar{y} &= y_{ij} - \bar{y} + \bar{y}_j - \bar{y}_j \\
(y_{ij} - \bar{y}) &= (y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y}) \\
\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 &= \sum_{j=1}^{n_j}(\bar{y}_j - \bar{y})^2 + \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 \\
\underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2}_{\text{Total Sum Squares}} &= \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2}_{\text{Between Sum Squares}} + \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}_{\text{Within Sum Squares}}
\end{aligned}$$

The TSS has $n-1$ degrees of freedom, the BSS has $g-1$ degrees of freedom and the WSS has $n-g$ degrees of freedom.

$$\underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2}_{df=n-1} = \underbrace{\sum_{j=1}^{g} n_j(\bar{y}_j - \bar{y})^2}_{df=g-1} + \underbrace{\sum_{j=1}^{g}(n_j - 1)s_j^2}_{df=n-g}$$

The ratio of BSS and its degrees of freedom is called *between mean squares* (BMS); and the ratio of WSS and its degrees of freedom is called *within mean squares* (WMS). Therefore, the test statistic is called an F statistic which is based on a variance ratio test (the ratio of BMS to WMS).

$$F = \frac{\text{BSS}/(g-1)}{\text{WSS}/(n-g)} = \frac{\text{BMS}}{\text{WMS}} \sim F(g-1, n-g)$$

Consequently, the critical value is determined using F distribution as $F_{tab} = F_\alpha(g-1, n-g)$. Therefore, if $F_{cal} > F_\alpha(g-1, n-g)$ or p-value$= P[F_(g-1, n-g) > F_{cal}] < \alpha$, the null hypothesis of no difference in all the g population means can be rejected. Hence, at least one group mean will be different from the others.

The ANOVA table is presented as follows.

Source of Variation	Sum Squares	*df*	MS	F
Between	$\text{BSS} = \sum_{j=1}^{g} n_j(\bar{y}_j - \bar{y})^2$	$g-1$	$\text{BMS} = \frac{\text{BSS}}{g-1}$	$F = \frac{\text{BMS}}{\text{WMS}}$
Within	$\text{WSS} = \sum_{j=1}^{g}(n_j - 1)s_j^2$	$n-g$	$\text{WMS} = \frac{\text{WSS}}{n-g}$	
Total	$\text{TSS} = \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$	$n-1$		

Note: ANOVA quite robust to (relatively unperturbed by) the violation of normality especially if the samples are equal size. But if the samples have different variances, the appropriate non-parametric alternative for one-way ANOVA which is called *Kruskal-Wallis* test should be used.

Example 8.15. Suppose a university wishes to compare the effectiveness of four teaching methods (Slide, Self Study, Lecture, Discussion) for a particular course. Twenty four students

are randomly assigned to the teaching methods, with 5, 6, 6 and 7 respectively. At the end of teaching the students with their assigned method, a test (out of 20%) was given and the performance of the students were recorded as follows:

Slide	Self Study	Lecture	Discussion
9	10	12	9
12	6	14	8
14	6	11	11
11	9	13	7
13	10	11	8
	5	16	6
			7

Test the hypothesis that there is no difference among the four teaching methods and also construct the ANOVA table.

Solution: The hypothesis to be tested is:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. All the four teaching methods are equally effective.

$H_1 :$ **not** H_0. At least one teaching method significantly differs from the others.

The summary statistics needed are obtained as:

	Slide (y_{i1})	Self Study (y_{i2})	Lecture (y_{i3})	Discussion (y_{i4})
	9	10	12	9
	12	6	14	8
	14	6	11	11
	11	9	13	7
	13	10	11	8
		5	16	6
				7
Sample size	$n_1 = 5$	$n_2 = 6$	$n_3 = 6$	$n_4 = 7$
Mean	$\bar{y}_1 = 11.800$	$\bar{y}_2 = 7.667$	$\bar{y}_3 = 12.833$	$\bar{y}_4 = 8.000$
Variance	$s_1^2 = 3.700$	$s_2^2 = 5.067$	$s_3^2 = 3.767$	$s_4^2 = 2.667$

Also, the total sample size is $n = \sum_{j=1}^{4} n_j = 24$ and the grand mean is $\bar{y} = \frac{1}{n}\sum_{j=1}^{4}\sum_{i=1}^{n_j} y_{ij} = \frac{1}{24}(231) = 9.625$.

Therefore, BSS $= \sum_{j=1}^{g} n_j(\bar{y}_j - \bar{y})^2 = 126.89$ with $df = 3$ and WSS $= \sum_{j=1}^{g}(n_j - 1)s_j^2 = 74.97$ with $df = 20$. Therefore, the calculated value of the F test statistic is

$$F = \frac{\text{BMS}}{\text{WMS}} = \frac{126.89/3}{74.97/20} = 11.28.$$

The ANOVA table is also constructed as follows.

S.V.	SS	df	MS	F
Between	BSS = 126.89	$4-1=3$	BMS $=\frac{126.89}{3}=42.2967$	$F=\frac{42.2967}{3.7485}=11.28$
Within	WSS = 74.97	$24-4=20$	WMS $=\frac{74.97}{20}=3.7485$	
Total	TSS = 201.85	$24-1=23$		

The critical value is $F_{cal}=F_{0.05}(3,20)=3.0984$. Since $F_{cal}=11.28>F_{0.05}(3,20)=2.38$ or $p-$value$=P[F(3,20)>11.28]=0.0001$ is less than 5%, H_0 can be rejected. This means at least one of the four teaching methods is significantly different at 5% significance level. But which method is different from the others?

8.4.2 Mean Separation - Multiple Comparison

In ANOVA, once the null hypothesis is rejected, then there is a need to identify which pair of group means are significant and which are not. There are several multiple comparison methods used for this purpose. Of them, the Fisher's Least Significant Difference (LSD) mean separation method will be considered in this course.

The LSD method compares two means at a time. For example, for comparing the j^{th} and k^{th} group means, the LSD statistic is:

$$\text{LSD}_{jk}=t_{\alpha/2}(n-g)\sqrt{\text{WMS}\left(\frac{1}{n_j}+\frac{1}{n_k}\right)}$$

where $j\neq k$. Then, if $\bar{y}_j-\bar{y}_k>\text{LSD}_{jk}$, there is a significant difference between μ_j and μ_k at α significance level. Otherwise, no significant difference is observed at that significance level.

Example 8.16. Recall example 8.15 and identify the significant pair of teaching method using LSD.

Solution: First, it is better to sort the groups (teaching methods) based on the sample means. Hence, the ranks from first to the fourth are lecture: $\bar{y}_3=12.833$, slide: $\bar{y}_1=11.800$, discussion: $\bar{y}_4=8.000$, and self study: $\bar{y}_2=7.667$, respectively. Then, the significance can be checked by taking each pair of successive groups. As a result, there will be three pairs of comparison: (Lecture vs Slide), (Slide vs Discussion) and (Discussion vs Self study). The critical value is $t_{0.025}(20)=2.086$.

Methods	$\bar{y}_j-\bar{y}_k$	LSD_{jk}	Significance
Lecture vs Slide	12.833-11.800=1.033	$2.086\sqrt{3.7485(\frac{1}{6}+\frac{1}{5})}=2.446$	No
Slide vs Discussion	11.800- 8.000=3.800	$2.086\sqrt{3.7485(\frac{1}{5}+\frac{1}{7})}=2.365$	Yes
Discussion vs Self study	8.000- 7.667=0.333	$2.086\sqrt{3.7485(\frac{1}{7}+\frac{1}{6})}=2.247$	No

Therefore, lecture and slide are better teaching methods than the other two.

Chapter 9

Inference for Categorical Responses

9.1 Inference about a Population Proportion

Recall a binary variable is a variable having only two categories, for example: patient outcome (cured or dead), development of cancer (yes or no). One of the categories is labeled as success and the other as failure. Mostly, the success outcome is coded by 1 and the failure is coded by 0.

The probability of a success is denoted by π and the probability of a failure is denoted by $1-\pi$. Then the probability distribution for the number of successes y in n independent and identical trials, is:

$$P(Y=y)=\binom{n}{y}\pi^y(1-\pi)^{n-y};\ y=0,1,2,\cdots,n.$$

Recall the mean and variance of the number of successes y are $n\pi$ and $n\pi(1-\pi)$, respectively. If both the expected number of outcomes are at least 5, then a normal distribution with mean $n\pi$ and variance $n\pi(1-\pi)$ can be used as an approximation for the binomial. If $Y\sim \text{Bin}(n,\pi)$, then $Y\sim\mathcal{N}(n\pi, n\pi(1-\pi))$. The approximation becomes more precise for large n.

In a random sample of n from a population, if there are y successes, then the sample proportion of successes is $p=\frac{y}{n}$ (alternatively, it can be denoted by $\hat{\pi}$). The point estimator of the binomial parameter π is the sample proportion of successes p (p estimates π). The mean of the sampling distribution of a sample proportion p is $E(p)=\mu_p=\pi$. Also, the variance of the sample proportion of successes is $V(p)=\sigma_p^2=\pi(1-\pi)/n$. Hence, for large sample size, the sampling distribution of a sample proportion is normal with mean π and variance $\pi(1-\pi)/n$. That is, $p\sim\mathcal{N}[\pi,\pi(1-\pi)/n]$.

Therefore, the standard error of the sample proportion p is $\text{SE}(p)=\sigma_p=\sqrt{\pi(1-\pi)/n}$. Consequently, the estimated standard error of the sample proportion p is $\widehat{\text{SE}}(p)=\hat{\sigma}_p=\sqrt{p(1-p)/n}$.

9.1.1 Testing for a Population Proportion π

The interest here is whether the population proportion of success π takes a particular value, say π_0.

The Wald Test

The Wald test uses the sample proportion p for estimating the standard error of the sample proportion p. That is, the estimated standard error is $\widehat{\text{SE}}(p) = \hat{\sigma}_p = \sqrt{p(1-p)/n}$.

Step 1: State both the null and alternative hypotheses. There three options are:

Option 1: $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$
Option 2: $H_0 : \pi = \pi_0$ vs $H_1 : \pi < \pi_0$
Option 3: $H_0 : \pi = \pi_0$ vs $H_1 : \pi > \pi_0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value is $z_{\alpha/2}$ for a two sided test and z_α for a one sided test.

Step 3: The Wald test statistic defined as:

$$Z = \frac{p - \pi}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: H_0 can be rejected if $|z_{cal}| > z_{crt}$ or p-value$< \alpha$.

Step 5: Conclusion.

Example 9.1. Of 1464 HIV/AIDS patients under HAART treatment in Jimma University Specialized Hospital from 2007-2011, 331 defaulted. Did the proportion of defaulter patients different from one fourth?

Solution: Let π denote the proportion of defaulter patients. The sample proportion of defaulters is $p = \frac{331}{1464} = 0.226$. For a sample of size $n = 1464$, the estimated standard error of p is $\widehat{\text{SE}}(P) = \sqrt{0.226(1 - 0.226)/1464} = 0.011$.

Step 1: Hypothesis:

$H_0 : \pi = 0.25$ The proportion of defaulter patients is not significantly different from 25%.

$H_1 : \pi \neq 0.25$ The proportion of defaulter patients is significantly different from 25%.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$

Step 3: The calculated value of the Wald test statistic is:

$$z = \frac{p - \pi}{\sqrt{p(1-p)/n}} = \frac{0.226 - 0.25}{\sqrt{0.226(1 - 0.226)/1464}} = -2.18$$

Step 4: Decision: Since $|z| = 2.18 > 1.96$, H_0 can be rejected. Or it is easy to find the two-sided p-value which is the probability that the absolute value of a standard normal variate exceeds 2.18, that is, $p - \text{value} = 2P(Z > 2.18) = 2(0.0146) = 0.0292$.

Step 5: Conclusion: Since, the one-sided p-value is 0.0146, there is a strong evidence that, $\pi < 0.25$, that is, the proportion of defaulter patients is fewer than a quarter at 5% level of significance.

The Score Test

The *Score* test is an alternative possible test which uses a known standard error. This known standard error is obtained by substituting the assumed value under the null hypothesis π_0. That is, $\hat{\sigma}_P = \sqrt{\pi_0(1-\pi_0)/n}$. Hence, the Score test statistic for a binomial proportion is:

$$Z = \frac{P - \pi}{\sqrt{\pi_0(1-\pi_0)/n}} \sim \mathcal{N}(0,1).$$

Example 9.2. Recall example 9.1. Test the hypothesis using the Score test.

Solution: Let π denote the proportion of defaulter patients. The sample proportion of defaulters is $p = \frac{331}{1464} = 0.226$. For Score test, the known standard error of P is $\widehat{\text{SE}}(P) = \sqrt{0.25(1-0.25)/1464} = 0.0113$.

Step 1: Hypothesis:

$H_0 : \pi = 0.25$ The proportion of defaulter patients is not significantly different from 25%.

$H_1 : \pi \neq 0.25$ The proportion of defaulter patients is significantly different from 25%.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$

Step 3: The calculated value of the Score test statistic is:

$$z = \frac{p - \pi}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.226 - 0.25}{\sqrt{0.25(1-0.25)/1464}} = -2.12$$

Step 4: Decision: Since $|z| = 2.12 > 1.96$, H_0 should be rejected. Also, the two-sided p-value is $2P(Z > 2.12) = 2(0.0170) = 0.034$ which leads to the rejection of H_0.

Step 5: Conclusion: There is a strong evidence that, $\pi < 0.25$, that is, the proportion of defaulter patients is fewer than a quarter at 5% level of significance..

9.1.2 Interval Estimation for a Population Proportion π

Wald CI: The $(1-\alpha)100\%$ (Wald) confidence interval for the population proportion π is given by:

$$\left[p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right].$$

This is a large sample confidence interval for the population proportion π which uses the sample proportion p as the mid-point of the interval.

Example 9.3. Recall example 9.1. Construct the 95% CI for the population proportion of HIV/AIDS patients who were defaulted.

Solution: For $n = 1464$ observations, $p = 0.226$. And $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval is $[0.226 \pm 1.96\frac{0.226(1-0.226)}{1464}] = (0.204, 0.248)$. Therefore, the proportion of HIV/AIDS patients who were defaulted is between 0.204 and 0.248 at 0.05 level of significance.

Note: The Wald confidence interval for π is based on a normal approximation to the binomial distribution. The rule is that both $n\pi$ and $n\pi(1-\pi)$ should be at least 5. Unless π is close to 0.50, it does not work well if n is not very large. That is, it works poorly to use the sample proportion as the mid-point of the confidence interval when π is near 0 or 1.

Score CI: The Score confidence interval uses a duality with significance tests. It is constructed by inverting results of a significance test using the null standard error. This confidence interval consists of all values π_0's for the null hypothesis parameter that are 'not rejected' at a given significance level.

For a binomial proportion, given n and p with a critical value $\pm z_{\alpha/2}$, the π_0 solutions for the equation

$$\frac{|p-\pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}} = \pm z_{\alpha/2}$$

are the end points of the Score confidence interval for π. Squaring both sides gives an equation which is quadratic in π_0. This method does not require estimation of π in the standard error, since the standard error in the test statistic uses the null value π_0.

Example 9.4. A clinical trial is conducted to evaluate a new treatment. This experiment has nine successes in the first 10 trials. Construct the 95% Score and Wald CIs.

Solution: The sample proportion of successes $p = 0.90$ based on $n = 10$ trials. The solutions for $n(p-\pi_0)^2 = \pi_0(1-\pi_0)z^2_{\alpha/2}$ are 0.596 and 0.982. Thus, the 95% Score CI is (0.596, 0.982).

By contrast, using the estimated standard error gives confidence interval (0.714, 1.086) in which the upper limit is greater than 1. That is why, it is said Wald CI works poorly when the parameter may fall near the boundary values of 0 or 1.

Example 9.5. Of $n = 16$ students, $y = 0$ answered "yes" for the question "Did you ever smoke cigarette?". Construct the 95% Wald and Score confidence intervals for the population proportion of smoker students.

Solution: Let π be the population proportion of smoker students. Since $y = 0$, $p = \frac{0}{16} = 0$. The 95% Wald CI is given by $(p \pm z_{\alpha/2}\sqrt{p(1-p)/n}) = (0 \pm 1.96\sqrt{0(1-0)/16}) = (0,\ 0)$. As said before when the number of successes is near 0 or near n, Wald methods do not provide sensible results.

The 95% Score confidence interval is obtained by solving $|0-\pi_0| = \pm 1.96\sqrt{\pi_0(1-\pi_0)/16}$ for π_0. By contrast this provides the interval (0, 0.316) which is sensible than the Wald interval (0, 0).

9.2 Comparing Two Population Proportions

For comparisons of two population proportions, independent random samples are assumed to be drawn from two binomial populations with parameters π_1 and π_2. If y_1 is the number of successes to be observed for a random sample of size n_1 from population (group) 1 and y_2 is the number of successes to be observed for a random sample of size n_2 from population

(group) 2, then the point estimators of π_1 and π_2 are the sample proportions $p_1 = \frac{y_1}{n_1}$ and $p_2 = \frac{y_2}{n_2}$, respectively.

The interest is whether the two population proportions are equal $\pi_1 = \pi_2$, that is, whether the difference between the two population proportions (absolute risk) is zero $\pi_1 - \pi_2 = 0$. The point estimator of the difference of the population proportions $\pi_1 - \pi_2$ is $p_1 - p_2$. The mean of the sampling distribution of the difference of the sample proportions $p_1 - p_2$ is $E(p_1 - p_2) = \mu_{p_1-p_2} = \pi_1 - \pi_2$. The variance of the sampling distribution of the difference of the population proportions $p_1 - p_2$ is also given as $V(p_1 - p_2) = \sigma^2_{p_1-p_2} = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$. Thus, $p_1 - p_2 \sim \mathcal{N}\left[\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]$.

The standard error is $\text{SE}(p_1 - p_2) = \sigma_{p_1-p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$. The estimated standard error is $\widehat{\text{SE}}(p_1 - p_2) = \hat{\sigma}_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

9.2.1 Testing for Difference of Two Population Proportions

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 \neq 0$

Option 2: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 < 0$

Option 3: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 > 0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value for a two sided test is $z_{\alpha/2}$ whereas the critical value for a one sided test is z_α.

Step 3: Use the z test statistic and obtain its calculated value:

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_2)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{tab}$ $(p - \text{value} < \alpha)$, H_0 can be rejected.

Step 5: Conclusion.

Example 9.6. A study looked at the effects of OC use on heart disease in women 40-44 years of age. The researchers found that among 50 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3 year period, whereas among 100 non-OC users, 7 developed an MI over a 3-year period. Assess the statistical significance of the results.

Solution: Let π_1 be the proportion of MI among OC users and π_2 be the proportion of MI among non-OC users. The sample proportion of MI among OC users is $p_1 = \frac{13}{50} = 0.26$ and the sample proportion of MI among non-OC users is $p_2 = \frac{7}{100} = 0.07$.

Step 1: Hypothesis:

$H_0 : \pi_1 - \pi_2 = 0$. The proportions of MI among OC users and non-OC users are not significantly different. That is, OC has not a significant effect.

$H_1 : \pi_1 - \pi_2 \neq 0$. The proportions of MI among OC users and non-OC users are significantly different. That is, OC has a significant effect.

Step 2: Assuming $\alpha = 0.05$, $z_{0.025} = 1.96$.

Step 3: The calculated value of the z test statistic is:

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_2)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(0.26 - 0.07) - 0}{\sqrt{\frac{0.26(1-0.26)}{50} + \frac{0.07(1-0.07)}{100}}} = \frac{0.19}{0.067} = 2.836$$

Step 4: Decision: Since $z_{cal} = 2.836 > z_{0.025} = 1.96$, H_0 can be rejected. Or $p-$value $= 2 \times P[Z > 2.836] = 2 \times 0.0023 = 0.0046 < \alpha = 0.05$.

Step 5: Conclusion. The proportions of MI among OC users and non-OC users are significantly different at 5% level of significance. That is, OC use has a significant positive effect to develop MI at 5% level of significance.

9.2.2 Interval Estimation for $\pi_1 - \pi_2$

The $(1 - \alpha)100\%$ confidence interval for the difference of the two population proportions $\pi_1 - \pi_2$ are given by:

$$\left\{ (p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right\}.$$

Example 9.7. Consider again example 9.6 and construct the 95% confidence interval for the difference in the proportions of MI between OC and non-OC users.

Solution: The 95% confidence interval for the difference in the proportions of MI between OC and non-OC users $\pi_1 - \pi_2$ is:

$$\left\{ (0.26 - 0.07) \pm 1.96 \sqrt{\frac{0.26(1-0.26)}{50} + \frac{0.07(1-0.07)}{100}} \right\} = (0.059,\ 0.321).$$

Since the confidence interval is greater than 0, OC use has a significant positive effect to develop MI at 5% level of significance.

9.3 Contingency Table Method

Let X and Y denote two categorical variables with I and J categories (levels), respectively. Then, classifications of subjects on both variables have IJ possible combinations and the contingency table is called a *two-way* table or an $I \times J$ (read as I-by-J) table.

Suppose N subjects are classified on both X and Y as shown in Table 9.1. Then N_{ij} represents the number of subjects belonging to the i^{th} category of X and j^{th} category of Y.

Table 9.1: Layout of an $I \times J$ Contingency Table

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1J}	N_{1+}
2	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2J}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{iJ}	N_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	N_{I1}	N_{I2}	$\cdots$	N_{Ij}	$\cdots$	N_{IJ}	N_{I+}
Total	N_{+1}	N_{+2}	$\cdots$	N_{+j}	$\cdots$	N_{+J}	N

Here, N_{i+} and N_{+j} are the marginal totals representing the number of subjects belonging to the i^{th} category of X and the j^{th} category of Y, respectively. Note that $N_{i+} = \sum_{j=1}^{J} N_{ij}$ and $N_{+j} = \sum_{i=1}^{I} N_{ij}$. Also, the population size $N = \sum_{i=1}^{I} N_{i+} = \sum_{j=1}^{J} N_{+j} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$.

9.3.1 Probability Structures for Contingency Tables

The joint probability distribution of the responses (X, Y) of a subject chosen randomly from some population can be determined from the contingency table. This joint distribution determines the relationship between the two categorical variables. Also, from this distribution, the marginal and conditional distributions can be determined.

Joint and Marginal Probabilities

The (true) probability of a subject being in the i^{th} category of X and j^{th} category of Y is

$$P(X = i, Y = j) = \pi_{ij} = \frac{N_{ij}}{N}.$$

The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y shown in Table 9.2. The marginal distribution of each variable is the sum of the joint probabilities over all the categories of the other variable. That is,

$$P(X = i) = \pi_{i+} = \sum_{j=1}^{J} \pi_{ij} = \frac{N_{i+}}{N} \text{ and } P(Y = j) = \pi_{+j} = \sum_{i=1}^{I} \pi_{ij} = \frac{N_{+j}}{N}.$$

Table 9.2: Joint and Marginal Distributions X and Y

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	1

Thus, $\{\pi_{i+}\}$ is the marginal distribution of X and $\{\pi_{+j}\}$ is the marginal distribution of Y. The marginal distributions provide single-variable information. Note also that $\sum_{i=1}^{I} \pi_{i+} = \sum_{j=1}^{J} \pi_{+j} = \sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1$.

Conditional Probabilities

The joint distribution of X and Y is more useful if both variables are responses. But if one of the variable is explanatory (fixed), the notion of the joint distribution is no longer useful.

If X is fixed, for each category of X, Y has a probability distribution. Hence, it is important to study how the distribution of Y changes as the category of X changes.

Given that a subject is belong to the i^{th} category of X, then

$$P(Y = j|X = i) = \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

denotes the conditional probability of that subject belonging to the j^{th} category of Y. In other words, $\pi_{j|i}$ is the conditional probability of a subject being in the j^{th} category of Y if it is in the i^{th} category of X. Thus, $\{\pi_{j|i};\ j = 1, 2, \cdots, J\}$ is the conditional distribution of Y at the i^{th} category of X. Note also that $\sum_{j=1}^{J} \pi_{j|i} = 1$.

Table 9.3: Conditional Distributions of Y Given X

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	$\pi_{1\|1}$	$\pi_{2\|1}$	$\cdots$	$\pi_{j\|1}$	$\cdots$	$\pi_{J\|1}$	1
2	$\pi_{1\|2}$	$\pi_{2\|2}$	$\cdots$	$\pi_{j\|2}$	$\cdots$	$\pi_{J\|2}$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\pi_{1\|i}$	$\pi_{2\|i}$	$\cdots$	$\pi_{j\|i}$	$\cdots$	$\pi_{J\|i}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	$\pi_{1\|I}$	$\pi_{2\|I}$	$\cdots$	$\pi_{j\|I}$	$\cdots$	$\pi_{J\|I}$	1

The probabilities $\{\pi_{1|i}, \pi_{2|i}, \cdots, \pi_{j|i}, \cdots, \pi_{J|i}\}$ form the conditional distribution of Y at the i^{th} category of X. A principal aim in many studies is to compare the conditional distribution of Y at various level of X.

Example 9.8. In the HAART Data used by **?**, there were 1464 HIV/AIDS patients. Of these 22.6% were defaulters. 63.5% of these patients were females including 189 defaulters.

1. Construct the contingency table.
2. Find the joint and marginal distributions.
3. If a patient is selected at random, what is the probability that the patient is
   (a) a female and defaulter?
   (b) a male?
   (c) defaulter if the patient is female?

Solution:

1. The contingency table is

	Defaulter		
Gender	Yes (1)	No (2)	Total
Female (1)	$N_{11} = 189$	$N_{12} = 741$	$N_{1+} = 930$
Male (2)	$N_{21} = 142$	$N_{22} = 392$	$N_{2+} = 534$
Total	$N_{+1} = 331$	$N_{+2} = 1133$	$N = 1464$

2. The joint and marginal distributions are

	Defaulter		
Gender	Yes (1)	No (2)	Total
Female (1)	$\pi_{11} = 0.129$	$\pi_{12} = 0.506$	$\pi_{1+} = 0.635$
Male (2)	$\pi_{21} = 0.097$	$\pi_{22} = 0.268$	$\pi_{2+} = 0.365$
Total	$\pi_{+1} = 0.226$	$\pi_{+2} = 0.774$	1.000

3. If a patient is selected at random,

(a) $P(\text{Gender} = 1, \text{Defaulter} = 1) = \frac{N_{11}}{N} = \frac{189}{1464} = 0.1291.$

(b) $P(\text{Gender} = 2) = \frac{N_{2+}}{N} = \frac{534}{1464} = 0.3648.$

(c) $P(\text{Defaulter} = 1|\text{Gender} = 1) = \frac{N_{11}}{N_{1+}} = \frac{189}{930} = 0.2032.$

9.3.2 Statistical Independence

Statistical independence is a condition of no relationship between two variables in a population. In probability terms, two categorical variables are defined to be independent if all joint probabilities are the product of their marginal probabilities. That is, if X and Y are independent then $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j.

Also, when X and Y are independent, each conditional distribution of Y is identical to the marginal distribution of Y. That is, $\pi_{j|i} = \pi_{+j}$ for all i. Thus, two categorical variables are independent when $\pi_{j|1} = \pi_{j|2} = \cdots = \pi_{j|I}$ for $j = 1, 2, \cdots, J$; that is, the probability of any category of Y is the same in each category of X which is often referred as *homogeneity* of conditional distributions. This is a more better definition of independence than $\pi_{ij} = \pi_{i+}\pi_{+j}$ when one of the variables is explanatory.

Example 9.9. Recall example 9.9. Are the sex of the patient and defaulting statistically independent? The answer is No. Why?

9.4 Chi-squared Tests of Independence

For a multinomial sampling with probabilities π_{ij} in an $I \times J$ contingency table, the null hypothesis of statistical independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. For independent multinomial samples, independence corresponds to homogeneity of each outcome probability among the categories of the fixed variable. The marginal probabilities then determine the joint probabilities.

Under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, the expected values of cell counts are $\{\mu_{ij} = n\pi_{i+}\pi_{+j}\}$. That is, μ_{ij} is the expected number of subjects in the i^{th} category of X and j^{th} category of Y. Since $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, their maximum likelihood estimates, respectively, are $\left\{p_{i+} = \frac{n_{i+}}{n}\right\}$ and $\left\{p_{+j} = \frac{n_{+j}}{n}\right\}$. which are the sample marginal proportions. Hence, the estimated expected frequencies are $\left\{\hat{\mu}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}\right\}$.

Table 9.4: Observed and Expected Frequencies in an $I \times J$ Table

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	n_{11} $(\hat{\mu}_{11})$	n_{12} $(\hat{\mu}_{12})$	$\cdots$	n_{1j} $(\hat{\mu}_{1j})$	$\cdots$	n_{1J} $(\hat{\mu}_{1J})$	n_{1+}
2	n_{21} $(\hat{\mu}_{21})$	n_{22} $(\hat{\mu}_{22})$	$\cdots$	n_{2j} $(\hat{\mu}_{2j})$	$\cdots$	n_{2J} $(\hat{\mu}_{2J})$	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1} $(\hat{\mu}_{i1})$	n_{i2} $(\hat{\mu}_{i2})$	$\cdots$	n_{ij} $(\hat{\mu}_{ij})$	$\cdots$	n_{iJ} $(\hat{\mu}_{iJ})$	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1} $(\hat{\mu}_{I1})$	n_{I2} $(\hat{\mu}_{I2})$	$\cdots$	n_{Ij} $(\hat{\mu}_{Ij})$	$\cdots$	n_{IJ} $(\hat{\mu}_{IJ})$	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

9.4.1 The Chi-square Test Statistic

The Pearson chi-squared statistic for testing independence of two categorical variables is defined as:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2[(I-1)(J-1)].$$

Step 1: Hypothesis:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j$. The two variables have no significant association.

$H_1 :$ **not** H_0. The variables are significantly associated.

Step 2: Obtain the critical value $\chi^2_\alpha[(I-1)(J-1)]$.

Step 3: The calculated value of the X^2 test statistic is:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Step 4: Decision: If $X^2_{cal} > \chi^2_\alpha[(I-1)(J-1)]$, the null hypothesis H_0 of no statistical association can be rejected. Or if $p - \text{value} = P(\chi^2[(I-1)(J-1)] > X^2_{cal})$ is smaller than α, H_0 can be rejected.

Step 5: Conclusion.

9.4.2 The Likelihood-Ratio Test Statistic

The likelihood-ratio test statistic is an alternative test for independence that uses likelihood values. A likelihood-ratio statistic is defined as $G^2 = -2\log(\ell_0/\ell_1)$ where ℓ_0 is the maximized value of the likelihood function under H_0 and ℓ_1 is the maximized value of the likelihood function in general. Therefore, the likelihood-ratio test statistic for independence can be easily derived as

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) \sim \chi^2[(I-1)(J-1)].$$

When H_0 holds, the Pearson X^2 and likelihood-ratio G^2 statistics both have asymptotic chi-squared distributions with $[(I-1)(J-1)]$ degrees of freedom. For a better approximation, the general rule is that the smallest expected frequency should be at least 5. In general, if more than 20% of the expected frequencies are less than 5, the approximation worsens (that is, the test is not valid).

Example 9.10. The table below shows the distribution of HIV/AIDS patients by the survival outcome (active, dead, transferred to other hospital and lost-to-follow) and gender.

	Survival Outcome				
Gender	Active	Dead	Transferred	Lost-to-follow	Total
Female	741	25	63	101	930
Male	392	20	52	70	534
Total	1133	45	115	171	1464

Test whether or not the survival outcome depends on gender using both the Pearson chi-square and likelihood-ratio tests.

Solution: First let us find the expected cell counts, $\hat{\mu}_{ij} = \dfrac{n_{i+}n_{+j}}{n}$.

	Survival Outcome				
Gender	Active	Dead	Transferred	Lost-to-follow	Total
Female	741 (719.7)	25 (28.6)	63 (73.1)	101 (108.6)	930
Male	392 (413.3)	20 (16.4)	52 (41.9)	70 (62.4)	534
Total	1133	45	115	171	1464

Step 1: Hypothesis:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j$. Survival outcome and gender have no significant association.

$H_1 :$ **not** H_0. Survival outcome depends on gender.

Step 2: The critical value $\chi^2_\alpha[(2-1)(4-1)] = \chi^2_{0.05}(3) = 7.8147$.

Step 3: The calculated value of the X^2 and G^2 test statistics, respectively, are:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \frac{(741-719.7)^2}{719.7} + \frac{(25-28.6)^2}{28.6} + \cdots + \frac{(70-62.4)^2}{62.4}$$

$$= 8.2172$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) = 2\left[741\log\left(\frac{741}{719.7}\right) + 25\log\left(\frac{25}{28.6}\right) + \cdots + 70\log\left(\frac{70}{62.4}\right)\right]$$

$$= 8.0720$$

Step 4: Decision: Since both statistics have larger values than $\chi^2_{0.05}(3) = 7.8147$, the null hypothesis H_0 can be rejected. Also, $p-$value $= P(\chi^2(3) > 8.0720) = 0.0445$ suggests rejection of no association between the two variables.

Step 5: Conclusion: The survival outcome of patients depends on gender at 5% level of significance.

9.5 Measuring Strength of Association

There are many situations where both the independent and dependent variables have two levels. Let X (explanatory) and Y (response) be binary variables. The data can be displayed in a 2×2 contingency table in which the rows are the levels of X and the columns are the levels of Y. Let us use the generic terms success and failure for the outcome categories of Y.

	Y		
X	Success (1)	Failure (2)	Total
1	N_{11}	N_{12}	N_{1+}
2	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N

For each category $i;\ i = 1, 2$ of X, $P(Y = j|X = i) = \pi_{j|i};\ j = 1, 2$. Then, the conditional probability structure is as follows.

	Y		
X	Success (1)	Failure (2)	Total
1	$\pi_{1\|1}$	$\pi_{2\|1}$	1
2	$\pi_{1\|2}$	$\pi_{2\|2}$	1

Here, $\pi_{1|1}$ and $\pi_{1|2}$ are the proportions of successes in category 1 and 2 of X, respectively. From now onwards, let us use π_1 and π_2 are the proportions of successes in category 1 and 2 of X, respectively.

	Y		
X	Success (1)	Failure (2)	Total
1	π_1	π_1'	1
2	π_2	π_2'	1

In chi-square test, the question of interest is whether there is a statistical association between the explanatory (X) and the response (Y) variables. The hypothesis to be tested is

$$H_0 : \pi_1 = \pi_2 \text{ (There is no association between } X \text{ and } Y)$$
$$H_1 : \pi_1 \neq \pi_2 \text{ (There is an association between } X \text{ and } Y)$$

A significant chi-squared test merely tells the existence of the association between the variables. If an association exists, the next task is identifying the category of X which has a larger (smaller) proportion of successes. This can be done by calculating the *difference of proportions*, a *relative risk* and an *odds ratio*.

9.5.1 Difference of Proportions (Absolute Risk)

The difference of proportions (absolute risk) is a simple procedure which compares the probability of success between two groups. It is calculated as $\pi_1 - \pi_2$. It is interesting that the

difference in proportions ranges between -1 and +1. If $\pi_1 - \pi_2 \approx 0$, the proportion of successes in both categories of X are almost the same (0 is a baseline for comparison). That is, if $\pi_1 - \pi_2 \approx 0$, categories of X have identical conditional distributions. On the contrary, if $\pi_1 - \pi_2 \approx \pm 1$, the association between X and Y is strong (indicates a high level of association).

Let p_1 and p_2 be the sample proportion of successes in category 1 and 2 of X, respectively. The difference of the sample proportion of successes $p_1 - p_2$ estimates the difference of the population proportion of successes $\pi_1 - \pi_2$. (Details are already discussed in Section 9.5.1).

Example 9.11. An educational researcher designs a study to compare the effectiveness of teaching English to non-English speaking people by a computer software program and by the traditional classroom system. The researcher randomly assigns 35 students from a class of 100 to instruction using the computer. The remaining 65 students are instructed using the traditional method. At the end of a 6-month instructional period, all 100 students are given an examination with the results reported in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	45	20	65
Computer	32	3	35
Total	77	23	100

Find the difference of the pass proportions and interpret. Also test the significance using the 95% confidence interval.

Solution: The conditional probabilities for each instruction method are shown in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	$p_1 = 0.692$	$p_1' = 0.308$	1
Computer	$p_2 = 0.914$	$p_2' = 0.086$	1

The difference in the sample pass proportions is $p_1 - p_2 = 0.692 - 0.914 = -0.222$. Since the difference is less than 0, computer instruction *seems* to be a better way to improve the academic performance of students in English course. The probability of passing in the traditional instruction method *decreases by* 0.222 as compared to passing in the computer instruction method. Or, the probability of passing in the computer instruction method *increases by* 0.222 as compared to passing in the traditional instruction method.

The 95% confidence interval for the difference in the pass proportions between the traditional and computer instruction methods $\pi_1 - \pi_2$ is

$$\left[(0.692 - 0.914) \pm 1.96\sqrt{\frac{0.692(1-0.692)}{65} + \frac{0.914(1-0.914)}{35}}\right]$$

$$= (-0.222 \pm \sqrt{0.0033 + 0.022}) = (-0.222 \pm 0.0742) = (-0.2962, -0.1478).$$

Thus, since the confidence interval is less than 0, the difference of the pass proportions in the two instruction methods is significantly different (particularly computer instruction is better than traditional instruction). Specifically, the probability of passing in the traditional instruction method *decreases by* between 0.1478 and 0.2962 at 5% significance level as compared to passing in the computer instruction method.

9.5.2 Relative Risk

Relative risk is the ratio of the probability of successes in two groups. That is,

$$r = \frac{\pi_1}{\pi_2} = \frac{N_{11}N_{12}}{N_{1+}N_{2+}}.$$

The value of a relative risk is non-negative, that is, $r \geq 0$. If $r \approx 1$, the proportion of successes in the two categories of X are approximately the same. This corresponds to independence or it is baseline for comparison. On the other hand, values of the relative risk r farther from 1 in a given direction represent stronger association. A relative risk of 4 is farther from independence than a relative risk of 2, and a relative risk of 0.25 is farther from independence than a relative risk of 0.50. Two values for relative risk (for example, 4 and 0.25) represent the same strength of association, but in opposite directions, when one value is the inverse of the other.

The sample relative risk $\hat{r} = \frac{p_1}{p_2}$ estimates the population relative risk r.

Example 9.12. Find the relative risk for the data given on example 9.11 and interpret it.

Solution: The conditional probabilities for each instruction method are shown in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	$p_1 = 0.692$	$p_1' = 0.308$	1
Computer	$p_2 = 0.914$	$p_2' = 0.086$	1

The estimate of the relative risk is $\hat{r} = \frac{p_1}{p_2} = \frac{0.692}{0.914} = 0.757$. It can be interpreted as follows:

- The proportion of passing in the traditional instruction method is 0.757 *times* the proportion of passing in the computer instruction method.
- The traditional instruction method *reduces* the probability of passing by $(1 - \hat{r})100\% = (1 - 0.757)100\% = 24.3\%$ relative to computer instruction method.
- Or, by inverting, the probability of passing in the computer instruction method is 1.321 *times* the probability of passing in the traditional instruction method.
- This means, computer instruction method (relative to traditional instruction method) *increases* the probability of passing the exam by $(\hat{r}-1)100\% = (1.321-1)100\% = 32.1\%$.

Note: Relative risk is a widely reported measure of association between exposure status and disease state for prospective studies (cohort and randomized clinical trials). In such case, the levels of the explanatory variable are being exposed (E) and being unexposed (E'), and the levels of the response variable are having a disease (D) and not-having a disease (D').

	Disease		
Exposure	Present (D)	Absent (D')	Total
Exposed (E)	n_{11}	n_{12}	n_{1+}
Unexposed (E')	n_{21}	n_{22}	n_{2+}
Total	?	?	n

For this particular case, relative risk is a ratio of the probability of having a disease among those exposed to the probability of having the disease among those unexposed:

$$r = \frac{P(D|E)}{P(D|E')}.$$

- A relative risk of 1.0 implies that the risk of a disease is the same in both exposed and unexposed groups (no association between the exposure and the disease).
- A relative risk greater than 1.0 implies the exposed group have a higher probability of having a disease than the unexposed group (the exposure is a *risk* factor).
- A relative risk less than 1.0 implies that the exposed group has a lower chance of having disease than unexposed group (it is expected in drug efficacy studies, the exposure is a *protective* factor).

Testing for a Relative Risk

To infer about a relative risk r, the sampling distribution of the sample relative risk $\hat{r}$ should be determined. The values of the relative risk are highly skewed to the right. As a result, by taking the logarithm of $\hat{r}$, it turns out that $\log(\hat{r})$ is approximately normally distributed for large values of n. If the probability of successes are approximately equal in the two groups, then $r = 1$ or $\log(r) = 0$ indicating no statistical association between the two variables.

The standard error of $\log(\hat{r})$ is determined to be:

$$\text{SE}[\log(\hat{r})] = \sqrt{\frac{1}{N_{11}} - \frac{1}{N_{1+}} + \frac{1}{N_{21}} - \frac{1}{N_{2+}}}$$

which can be estimated by:

$$\widehat{\text{SE}}[\log(\hat{r})] = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}.$$

Step 1: Hypothesis:

$H_0 : \log(r) = 0$ The two variables have no significant association.

$H_1 : \log(r) \neq 0$ The two variables are significantly associated.

Step 2: Obtain the critical value $z_{\alpha/2}$.

Step 3: Under $H_0 : \log(r) = 0$, for large values of n the test statistic is defined as:

$$Z = \frac{\log(\hat{r}) - \log(r)}{\widehat{\text{SE}}[\log(\hat{r})]} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{\alpha/2}$, H_0 should be rejected.

Step 5: Conclusion.

Example 9.13. Test the significance of the relative risk for the data given on example 9.11.

Solution: The estimate of the relative risk is $\hat{r} = \frac{p_1}{p_2} = \frac{0.692}{0.914} = 0.757$ which implies $\log(\hat{r}) = \log(0.757) = -0.2784$ and the estimated standard error of $\log(\hat{r})$ is $\widehat{\text{SE}}[\log(\hat{r})] = 0.0975$.

Step 1: Hypothesis:

$H_0 : r = 1 \Rightarrow \log(r) = 0$. Instruction method and exam result have no significant association.

$H_1 : r \neq 1 \Rightarrow \log(r) \neq 0$. Instruction method and exam result have a significant association.

Step 2: Using $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The calculated value of the z test statistic is:

$$z = \frac{\log(0.757) - 0}{\sqrt{\frac{1}{45} - \frac{1}{65} + \frac{1}{32} - \frac{1}{35}}} = -2.86.$$

Step 4: Decision: Since $|z_{cal}| = 2.86 > z_{0.025} = 1.96$, H_0 should be rejected.

Step 5: Conclusion: Therefore, the relative risk is significantly different from 1. Instruction method has a significant effect on examination result at 5% significance level. Specifically, the computer instruction method has a positive effect in passing the examination.

Confidence Interval for a Relative Risk

The $(1 - \alpha)100\%$ confidence interval for the log of a relative risk $\log(r)$ is given by

$$\{\log(\hat{r}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{r})]\}.$$

Taking the exponentials of the end points this confidence interval provides the confidence interval for a relative risk r, that is,

$$\exp\{\log(\hat{r}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{r})]\}.$$

Example 9.14. An efficacy study was conducted for the drug pamidronate in patients with Paget's disease of bone. In this randomized clinical trial, patients were assigned at random to receive either pamidronate (E) or placebo (E'). One end point was the occurrence of any skeletal events after 9 cycles of treatment D and non-occurrence D'. The results are given in the following table.

	Skeletal Event		
Exposure	Yes (D)	No (D')	Total
Pamidronate (E)	47	149	196
Placebo (E')	74	107	181
Total	121	256	377

Compute a 95% confidence interval for the relative risk of suffering skeletal events (in a time period of this length) for patients on pamidronate relative to patients not on the drug.

Solution: Let $\pi_1 = P(D|E)$ and $\pi_2 = P(D|E')$. Thus, the estimated probability of patients suffering skeletal events among those receiving the drug, and among those receiving the placebo are $p_1 = \frac{47}{196} = 0.240$ and $p_2 = \frac{74}{181} = 0.409$, respectively.

Then, the estimated relative risk r is $\hat{r} = \frac{0.240}{0.409} = 0.587$ and its log value is $\log(\hat{r}) = -0.533$. The estimated standard error of log of the estimated relative risk $\log(\hat{r})$ is $\widehat{\text{SE}}[\log(\hat{r})] = \sqrt{\frac{1}{47} - \frac{1}{196} + \frac{1}{74} - \frac{1}{181}} = 0.155$.

The 95% confidence interval for the log of the relative risk $\log(r)$ is $-0.533 \pm 1.96(0.155) = (-0.837,\ -0.229)$. Therefore, the 95% confidence interval for the relative risk r is

$$\{\exp(-0.837),\ \exp(-0.229)\} = (0.433,\ 0.795).$$

Thus, the relative risk of suffering a skeletal event (in this time period) for patients on pamidronate (relative to patients not on pamidronate) is between 0.433 and 0.795 at 5% significance level. Since this entire interval is below 1, it can be concluded that pamidronate is effective in reducing the risk of skeletal events. Furthermore, pamidronate reduces the risk of skeletal events by $(1 - \hat{r})100\% = (1 - 0.587)100\% = 41.3\%$.

9.5.3 Odds Ratio

Before defining an odds ratio, let us define what an odds is? An odds (Ω) is the ratio of the probability of success to the probability of failure in a particular group.

$$\Omega = \frac{p(\text{success})}{p(\text{failure})} = \frac{\pi}{1-\pi} = \frac{\text{number of successes}}{\text{number of failures}}$$

Like a relative risk, an odds is a nonnegative number ($0 \leq \Omega < \infty$). If $\Omega = 1$, a successes is as likely as a failure. If $\Omega < 1$, a success is less likely and if $\Omega > 1$, a success is more likely to occur than a failure. Inversely,

$$\pi = \frac{\Omega}{1+\Omega}.$$

Odds ratio is the ratio of two odds. For a 2×2 table, for each group i of X, the odds of successes (instead of failures) is

$$\Omega_i = \frac{\pi_i}{1-\pi_i} = \frac{\pi_i}{\pi_i'};\ i = 1, 2.$$

Thus, the odds ratio is

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1\pi_2'}{\pi_2\pi_1'} = \frac{N_{11}N_{22}}{N_{12}N_{21}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Like a relative risk and an odds, an odds ratio is also non negative, that is, $\theta \geq 0$. An odds ratio of 1 implies independence of X and Y which is a baseline for comparison. If it larger than 1 ($\Omega_1 > \Omega_2$), a success is more likely to occur in category 1 of X than in category 2. If the odds ratio is near zero ($\Omega_1 < \Omega_2$), then a success is less likely to occur in category 1 than category 2.

Similar to a relative risk, values of an odds ratio θ farther from 1 in a given direction represent stronger association, that is, an odds ratio of 6 is farther from independence than an odds ratio of 2, and an odds ratio of 0.20 is farther from independence than an odds ratio of 0.60. Also, two values for odds ratio, when one value is the inverse of the other (for example, 5 and 0.20) represent the same strength of association, but in opposite directions.

The sample odds ratio $\hat{\theta}$ is used to estimate the population odds ratio θ which is given by

$$\hat{\theta} = \frac{\widehat{\Omega}_1}{\widehat{\Omega}_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Example 9.15. Again recall example 9.11. Find the odds ratio and interpret.

Solution: The estimated probability of passing in the traditional instruction method is $p_1 = 0.692$. Then, the estimated odds of passing in this group is $\widehat{\Omega}_1 = \frac{0.692}{1-0.692} = 2.247$ which means the probability of passing in the traditional instruction group is 2.247 times the probability of failing in that group.

Similarly, the estimated probability of passing in the computer instruction group is $p_2 = 0.914$. Hence, the estimated odds of passing in this group is $\widehat{\Omega}_2 = \frac{0.914}{1-0.914} = 10.628$ which means the probability of passing in the computer instruction group is 10.627 times the probability of failing.

Therefore, the odds ratio of passing the exam (instead of failing) is the ratio of the odds of passing in the traditional instruction method to the odds of passing in the computer instruction group, that is, $\hat{\theta} = \frac{\widehat{\Omega}_1}{\widehat{\Omega}_2} = \frac{2.247}{10.628} = 0.211$. This value can be interpreted in different ways as follows.

- The odds of passing (instead of failing) the exam in the traditional instruction method is 0.211 *times* the odds of passing in the computer instruction method.
- The odds of passing (instead of failing) in the traditional instruction group *decreases by a factor of* 0.211 relative to the odds of passing in the computer instruction group.
- That is, the odds of passing (instead of failing) in the traditional instruction group is $(1-\hat{\theta})100\% = (1-0.211)100\% = 78.9\%$ *lower than* the odds of passing in the computer instruction group.
- Those in the traditional instruction method group are 0.211 times *less likely* to pass the exam (instead of failing) than those in the computer instruction group.
- Or inversely, the odds of passing (instead of failing) the exam in the computer instruction group is 4.739 *times* the odds of passing in the traditional instruction group.

- The odds of passing (instead of failing) the exam in the computer instruction group *increases by a factor of* 4.739 as compared to those in the traditional instruction group.
- This means, the odds of passing (instead of failing) the exam in the computer instruction method is $(\hat{\theta} - 1)100\% = (4.739 - 1)100\% = 373.9\%$ *higher than* the odds of passing in the traditional instruction group.
- Those in the computer instruction group are 4.739 times *more likely* to pass the exam (instead of failing) than those in the traditional instruction method group.

Example 9.16. Given the following contingency table for the variable "death penalty for crime".

	Race		
Penalty	Blacks	Nonblacks	Total
Death Sentence	28	22	50
Life Imprisonment	45	52	97
Total	73	74	147

Find the odds of receiving a death sentence and interpret. Also, calculate the odds ratio for receiving a death penalty and interpret.

Solution: The estimated probability of receiving a death sentence is $\frac{50}{147} = 0.34$ (34%). Then, the estimated odds of receiving a death sentence (instead of a life imprisonment sentence) is $\frac{50}{97} = 0.516$ (51.6%). Receiving a death sentence is *half as likely as* life imprisonment or receiving a life imprisonment sentence is *twice as likely as* receiving a death penalty.

The odds ratio for receiving a death penalty (instead of life imprisonment) is the ratio of the odds if black to the odds if nonblack. It is estimated as 1.47 which means blacks are 1.47 *times more likely* to receive a death sentence (instead of life imprisonment) than nonblacks. This means, the risk (odds) of death sentence (instead of life imprisonment) for blacks *increases by a factor of* 1.47 as compared to nonblacks. Or the risk (odds) of death sentence for blacks are 47% *higher than* the risk (odds) of a death sentence for nonblacks.

Note: For retrospective (case-control) studies, subjects are identified as cases (D) or controls (D'), and it is observed whether the subjects had been exposed to the risk factor (E) or not (E'). Since the samplings are not from the populations of exposed and unexposed, and observing whether or not disease occurs (as in prospective studies), $P(D|E)$ or $P(D|E')$, cannot be estimated.

	Disease		
Exposure	Present (D)	Absent (D')	Total
Exposed (E)	n_{11}	n_{12}	?
Unexposed (E')	n_{21}	n_{22}	?
Total	n_{+1}	n_{+2}	n

- If the odds ratio is 1.0, the odds (and thus probability) of disease is the same for both groups (no association between an exposure and a disease).

- If the odds ratio is greater than 1.0, the odds (and thus probability) of disease is higher among exposed than unexposed (the exposure is a *risk* factor).
- If the odds ratio is less than 1.0, the odds (and thus probability) of disease is lower among exposed than unexposed (the exposure is a *protective* factor).

Testing for an Odds Ratio

To infer about an odds ratio θ, the sampling distribution of $\log(\hat{\theta})$ is used due to the similar reasons used for a relative risk. If the odds of successes are equal in the two groups, then $\theta = 1$ or $\log(\theta) = 0$ indicating independence (no statistical association).

The standard error of the log of an odds ratio $\log(\hat{\theta})$ can be determined using statistical theory as:

$$\text{SE}[\log(\hat{\theta})] = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}$$

which can be estimated by:

$$\widehat{\text{SE}}[\log(\hat{\theta})] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Step 1: Hypothesis:

$H_0 : \mathbf{OR} = 1 \Rightarrow \log(\mathbf{OR}) = 0$. The two variables have no significant association.

$H_1 : \mathbf{OR} \neq 1 \Rightarrow \log(\mathbf{OR}) \neq 0$. The two variables are significantly associated.

Step 2: Obtain the critical value $z_{\alpha/2}$.

Step 3: Under $H_0 : \log(\theta) = 0$, for large values of n the test statistic is defined as:

$$Z = \frac{\log(\hat{\theta}) - \log(\theta)}{\widehat{\text{SE}}[\log(\hat{\theta})]} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{\alpha/2}$, H_0 can be rejected.

Step 5: Conclusion.

Example 9.17. Test the significance of the odds ratio for the data given at example 9.16.

Solution: It is easily to calculate that $\hat{\theta} = 1.47$ and $\log(\hat{\theta}) = 0.385$. Also, the standard error of $\log(\hat{\theta})$ is $\widehat{\text{SE}}[\log(\hat{\theta})] = 0.349$.

Step 1: Hypothesis:

$H_0 : \theta = 1 \Rightarrow \log(\theta) = 0$. Death penalty and race have no significant association.

$H_1 : \theta \neq 1 \Rightarrow \log(\theta) \neq 0$. Death penalty and race have a significant association.

Step 2: Using $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The calculated value of the Z test statistic is:

$$z = \frac{\log(\hat{\theta}) - 0}{\sqrt{\frac{1}{28} + \frac{1}{22} + \frac{1}{45} + \frac{1}{52}}} = 1.103.$$

Step 4: Decision: Since $|z_{cal}| = 1.103 < z_{0.025} = 1.96$, H_0 cannot be rejected.

Step 5: Conclusion: Therefore, there is not much evidence of association between penalty for crime and race at 5% significance level.

Confidence Interval for an Odd Ratio

The $(1-\alpha)100\%$ confidence interval for an odds ratio θ is given by

$$\exp\{\log(\hat{\theta}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{\theta})]\}.$$

Example 9.18. An epidemiological case-control study was reported, with cases being 537 people diagnosed with lung cancer (D) and controls being made up of 500 people with no lung cancer (D'). One risk factor measured was whether or not the subject had smoked a cigarette (a smoker - E, a non-smoker - E'). The following table gives the numbers of subjects falling in each possible combination.

	Lung Cancer		
Exposure	Yes (D)	No (D')	Total
Smoker (E)	339	149	488
Nonsmoker (E')	198	351	549
Total	537	500	1037

Compute a 95% confidence interval for the population odds ratio, and determine whether or not cigarette smoking is associated with higher (or possibly lower) odds (and probability) of developing lung cancer.

Solution: The estimated odds ratio for developing cancer cancer in smokers and non-smokers is $\hat{\theta} = \frac{339(351)}{149(198)} = 4.03$. This implies $\log(\hat{\theta}) = 1.394$ and its estimated standard error is $\widehat{\text{SE}}\{\log(\hat{\theta})\} = \sqrt{\frac{1}{339} + \frac{1}{149} + \frac{1}{198} + \frac{1}{351}} = 0.133$. Therefore, the 95% confidence interval for the odds ratio θ is

$$\{\exp[1.394 - 1.96(0.133)],\ \exp[1.394 + 1.96(0.133)]\} = (3.110,\ 5.231).$$

That is, the risk of developing lung cancer is between 3.11 and 5.231 times higher among smokers than non-smokers at $\alpha = 0.05$.

Odds Ratios in an $I \times J$ Table

For a 2×2 table, a single number such as an odds ratio can summarize the association. For an $I \times J$ table, it is rarely possible to summarize association by a single number without some loss of information. However, a set of $(I-1)(J-1)$ local odds ratios can describe certain features of the association (the rest odds ratios can be determined from these odds ratios).

Consider category i and $i+1$ of X, and category j and $j+1$ of Y in an $I \times J$ contingency table. Then, the odds ratio:

$$\theta_{ij} = \frac{N_{ij}N_{i+1,j+1}}{N_{i,j+1}N_{i+1,j}} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}};\ i = 1, 2, \cdots, I-1,\ j = 1, 2, \cdots, J-1$$

compares the probability of category j (instead of $j+1$) of Y in category i of X as compared to category $i+1$ of X.

As usual, the estimated odds ratio for comparing category j (instead of $j+1$) of Y between category i and $i+1$ of X is:

$$\hat{\theta}_{ij} = \frac{n_{ij}n_{i+1,j+1}}{n_{i,j+1}n_{i+1,j}} = \frac{p_{ij}p_{i+1,j+1}}{p_{i,j+1}p_{i+1,j}};\ i = 1, 2, \cdots, I-1,\ j = 1, 2, \cdots, J-1.$$

Independence is equivalent to all odds ratios equal to 1 (that is, non-significance of all odds ratios).

Example 9.19. Suppose 980 individuals are classified according to their favorite soft drink preference (Fanta, Coca and Sprite) and gender as shown below.

	Soft Drink			
Gender	Fanta	Coca	Sprite	Total
Females	279	225	73	577
Males	165	191	47	403
Total	444	416	120	980

By looking at the frequencies in the table, guess which gender (male or female) seems more likely to prefer coca? Why? Find all (local) odds ratios and test their significance.

Solution: The association between gender and soft drink preference can be checked using the chi-square or likelihood-ratio tests.

Step 1: Hypothesis:

H_0 : There is no significant association between soft drink preference and gender.

H_1 : Soft drink preference significantly depends on gender.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The z test statistic is used for testing each odds ratio:

	Fanta versus Coca	Fanta versus Sprite	Coca versus Sprite
Odds Ratio ($\hat{\theta}_{ij}$)	$\frac{279(191)}{225(165)} = 1.435$	$\frac{279(47)}{73(165)} = 1.089$	$\frac{225(47)}{73(191)} = 0.758$
Log Odds Ratio $\{\log(\hat{\theta}_{ij})\}$	$\log(1.435) = 0.361$	$\log(1.089) = 0.085$	$\log(0.758) = -0.120$
$\widehat{\text{SE}}[\log(\hat{\theta}_{ij})]$	0.139	0.211	0.211
Test Statistic (z)	2.597	0.402	-0.569
Decision	Reject H_0	Do not reject H_0	Do not reject H_0

Step 4: Decision: Since one of the three odds ratios is significant at 5% significance level, the null hypothesis of no significant association is rejected.

Step 5: Conclusion: Therefore, there is a significant difference in the preference of Fanta (instead of Coca) by females as compared to males at 5% level of significance. Hence, from this analysis, it can be concluded that:

- Females are 1.435 times *more likely* to prefer Fanta (instead of Coca) than that of males.
- The odds of preferring Fanta (instead of Coca) by females is 43.5% *higher than* that of males.
- Males are 0.697 times *less likely* to prefer Fanta (instead of Coca) than females.
- The odds of preferring Fanta (instead of Coca) by males is 30.3% *lower than* that of females.

9.6 Exact Inference for Small Samples

The inferential methods of the previous sections are all large sample methods. The Pearson chi-square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When there are very small expected frequencies, the possible values of the chi-square statistic are quite discrete. For example, for a 2×2 table with only 4 observations in each row and column, the only possible values of chi-square are 8, 2, and 0. It should be clear that a continuous chi-square distribution is not a good match for a discrete distribution having only 3 values. In such cases, when n is small, alternative methods use exact distributions rather than large sample approximations.

In this section, small sample test of independence for 2×2 tables, which is called Fisher's exact inference is discussed. As described in Section **??**, in poisson sampling - the sample size is not fixed unlike multinomial sampling, and in independent multinomial (binomial) sampling only one set of the marginal totals are fixed. In addition, in a 2×2 table, if both sets of the marginal total are fixed, it yields a hypergeometric distribution, that is,

$$P(Y_{11} = n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}}.$$

Given the marginal totals, n_{11} determines the other three cell counts. The exact p-value is determined using the hypergeometric distribution. The procedure to calculate the p-value for testing $H_0 : \theta = 1$ is as follows. Of the four marginal totals, select the smallest one and create ordered pair of integers with that sum. Next complete the 2×2 table for each of the ordered pair. Then, the two-sided p-value is given by $P(Y_{11} \leq n_{11})$ where n_{11} is the observed frequency in cell $(1, 1)$. For a one sided test, the p-value is found by comparing the observed frequency n_{11} to its expected value $\hat{\mu}_{11}$. If $n_{11} > \hat{\mu}_{11}$, then the onesided (right-sided alternative: $H_1 : \theta > 1$) p-value is $P(Y_{11} \geq \hat{\mu}_{11})$ and if $n_{11} < \hat{\mu}_{11}$, then the onesided (left-sided alternative: $H_1 : \theta < 1$) p-value is $P(Y_{11} \leq n_{11})$.

Example 9.20. Suppose A and B are two small colleges, the results of the beginning Statistics course at each of the two colleges are given below.

	Statistics		
Colleges	Pass	Fail	Total
A	8	14	22
B	1	3	4
Total	9	17	26

Do the data provide sufficient evidence to indicate that the proportion of passing Statistics differs for the two colleges?

Solution: The hypothesis to be tested is, $H_0 : \pi_{1|A} = \pi_{1|B}$, the proportion of passing Statistics do not differ significantly for the two colleges. Since the sample sizes are small, Fisher's exact test will be used. Since $n_{2+} = 4$ is the smallest marginal total, the following ordered pairs for (n_{21}, n_{22}) can be determined: (0, 4), (1, 3), (2, 2), (3, 1) and (4,0). For each pair, the 2×2 table is completed and the corresponding probability is computed using

$$P(Y_{11} = n_{11}) = \frac{n_{1+}!\ n_{2+}!\ n_{+1}!\ n_{+2}!}{n!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!}.$$

For (n_{21}, n_{22})=(0, 4):

$$\begin{array}{|c|c|}\hline 9 & 13 \\ \hline 0 & 4 \\ \hline \end{array} \Rightarrow P(Y_{11} = 9) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 9!\ 13!\ 0!\ 4!} = 0.159197$$

For (n_{21}, n_{22})=(1, 3):

$$\begin{array}{|c|c|}\hline 8 & 14 \\ \hline 1 & 3 \\ \hline \end{array} \Rightarrow P(Y_{11} = 8) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 8!\ 14!\ 1!\ 3!} = 0.409365$$

For (n_{21}, n_{22})=(2, 2):

$$\begin{array}{|c|c|}\hline 7 & 15 \\ \hline 2 & 2 \\ \hline \end{array} \Rightarrow P(Y_{11} = 7) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 7!\ 15!\ 2!\ 2!} = 0.327492$$

For (n_{21}, n_{22})=(3, 1):

$$\begin{array}{|c|c|}\hline 6 & 16 \\ \hline 3 & 1 \\ \hline \end{array} \Rightarrow P(Y_{11} = 6) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 6!\ 16!\ 3!\ 1!} = 0.095518$$

For (n_{21}, n_{22})=(4, 0):

$$\begin{array}{|c|c|}\hline 5 & 17 \\ \hline 4 & 0 \\ \hline \end{array} \Rightarrow P(Y_{11} = 5) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 5!\ 17!\ 4!\ 0!} = 0.008428$$

Since the observed frequency $n_{11} = 8$, the two sided p-value is $P(Y_{11} \leq 8) = P(Y_{11} = 5) + P(Y_{11} = 6) + P(Y_{11} = 7) + P(Y_{11} = 8) = 1$. Hence, there is not enough evidence to conclude that the proportion of passing Statistics differs for the two colleges.

Since the observed frequency $n_{11} = 8 > \hat{\mu}_{11} = 7.6$, the alternative hypothesis is $(H_1 : \pi_{1|A} > \pi_{1|B})$. Then the onesided p-value is $P(Y_{11} \geq 7.6) = P(Y_{11} = 8) + P(Y_{11} = 9) = 0.159197 + 0.409365 = 0.568562$. Again, there is not enough evidence to indicate that the probability of passing Statistics is higher at college A than at college B.

9.7 Measures of Linear Association for Ordinal Variables

In situations where both the explanatory and response variables are ordinal, the X^2 and G^2 tests ignore the fact that the levels of the variables have distinct orderings. When both variables are ordinal, there will be an interest to examine whether individuals with high levels of an explanatory variable tend to have high (low) levels of the corresponding response variable. For instance, suppose that the explanatory variable is dose, with increasing (possibly numeric) levels of amount of drug given to a patient, and the response variable is categorical measuring the degree of improvement. Then, it is essential to determine if as dose increases, the degree of improvement increases.

Many measures have been developed for this type of ordinal variables classification. Most analytical techniques are based on concordant and discordant pairs. A *concordant* pair involves a pair where a subject is higher on both variables than other subject. A *discordant* pair is a pair where a subject is higher on one variable, but lower on the other variable, than other subject. If a pair is said to be *tied* if a subject is in the same category of a variable.

More concordant pairs than discordant pairs indicates a *positive association* between the two variables whereas more discordant pairs than concordant pairs indicates *negative association* between the variables.

Consider the following table

	Income Level		
Education Level	Low	High	Total
High School	N_{11}	N_{12}	
College	N_{21}	N_{22}	
Total			

Looking at the above table, it is easy to observe that income category is ordered by low and high. Similarly education category is ordered, with education ending at high school being the low category and education ending at college being the high category. All N_{11} observations represent individuals in low income and low education category and all N_{22} observations represent individuals in high income and high education category. Thus, there are $C = N_{11}N_{22}$ concordant pairs. On the other hand, all N_{12} observations are higher on the income variable and lower on the education variable, while all N_{21} observations are lower on the income variable and higher on the education variable. Thus, there are $D = N_{12}N_{21}$ discordant pairs.

9.7.1 The Gamma Measure

The strength of the association can be measured by calculating the difference in the proportions of concordant and discordant pairs. This is called the *gamma* (γ) measure which is defined as

$$\gamma = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}.$$

Since γ represents the difference in proportions, its value is between -1 and 1. A positive value of gamma indicates a positive association while a negative value of gamma indicates a

negative association. A value close to zero indicates weak association.

Let us consider again the above 2×2 table. Let $n_{11} = 25$, $n_{12} = 12$, $n_{21} = 11$ and $n_{22} = 14$. The number of concordant pais is $\widehat{C} = n_{11}n_{22} = 25(14) = 350$; the number of discordant pairs is $\widehat{D} = n_{12}n_{21} = 12(11) = 132$. Therefore, $\widehat{\gamma} = 0.45$ which indicates that the association between education level and income is medium-positive.

For an $I \times J$ table, the number of concordant pairs is $C = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}(\sum_{h=i+1}^{I}\sum_{k=j+1}^{J} N_{hk})$ and the number of discordant pairs is $D = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}(\sum_{h=i+1}^{I}\sum_{k=1}^{j-1} N_{hk})$.

Example 9.21. Find the gamma measure of association for the following cross-classification of HIV/AIDS patients by Clinical Stage and Functional Status.

	Functional Status			
Clinical Stage	Bedridden	Ambulatory	Working	Total
Stage I	0	23	324	347
Stage II	11	96	407	514
Stage III	28	233	235	496
Stage IV	18	52	37	107
Total	57	404	1003	1464

Solution: The total number of concordant pairs is

$$\begin{aligned}\widehat{C} =& 0(96 + 407 + 233 + 235 + 52 + 37) + 23(407 + 235 + 37) \\ &+ 11(233 + 235 + 52 + 37) + 96(235 + 37) + 28(52 + 37) + 233(37) \\ =& 58969\end{aligned}$$

The total number of discordant pairs is

$$\begin{aligned}\widehat{D} =& 23(11 + 28 + 18) + 324(11 + 96 + 28 + 233 + 18 + 52) + 96(28 + 18) \\ &+ 407(28 + 233 + 18 + 52) + 233(18) + 235(18 + 52) \\ =& 303000\end{aligned}$$

In this example, $\widehat{C} < \widehat{D}$, suggesting a tendency for low clinical stage to occur with high functional status of patients and higher clinical stages with lower functional status.

$$\widehat{\gamma} = \frac{\widehat{C} - \widehat{D}}{\widehat{C} + \widehat{D}} = \frac{58969 - 303000}{58969 + 303000} = -0.674$$

Of the untied pairs, the proportion of concordant pairs is 0.674 lower than the proportion of discordant pairs. This indicates that there is a medium negative linear association between clinical stage and functional status of HIV/AIDS patients. That is, as the clinical stage (severity) of the patient increases, the functional status of the patient decreases and vice versa.

9.7.2 The Kendall's tau-b

Kendall's tau-b, denoted τ_b, is a more sensitive measure of association between two ordinal variables. The formula for calculating Kendall's tau-b τ_b is:

$$\tau_b = \frac{C - D}{0.5\sqrt{\left(N^2 - \sum_{i=1}^{I} N_{i+}^2\right)\left(N^2 - \sum_{j=1}^{J} N_{+j}^2\right)}}.$$

The estimated value of Kendall's tau-b $\hat{\tau}_b$ is also obtained by substituting the sample frequencies in place of the population frequencies as:

$$\hat{\tau}_b = \frac{\widehat{C} - \widehat{D}}{0.5\sqrt{\left(n^2 - \sum_{i=1}^{I} n_{i+}^2\right)\left(n^2 - \sum_{j=1}^{J} n_{+j}^2\right)}}.$$

This measure has the advantage of adjusting for ties. The result of adjusting for ties is that the value of τ_b is always a little closer to 0 than the corresponding value of gamma.

Example 9.22. Find the Kendall's tau-b τ_b for the data given in example 9.21.

Solution:

$$\begin{aligned}
\hat{\tau}_b &= \frac{58969 - 303000}{0.5\sqrt{[1464^2 - (57^2 + 404^2 + 1003^2)][1464^2 - (347^2 + 514^2 + 496^2 + 107^2)]}} \\
&= \frac{-244031}{0.5\sqrt{(2143296 - 1172474)(2143296 - 642070)}} \\
&= -0.404
\end{aligned}$$

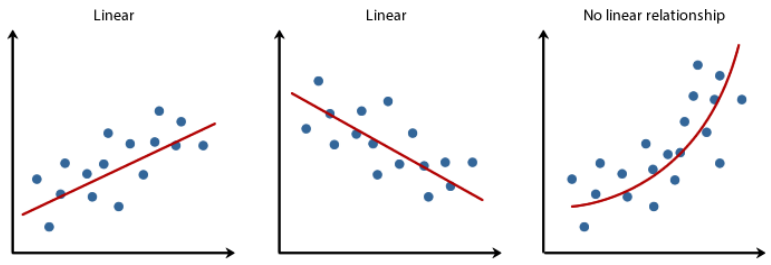
Chapter 10

Correlation and Regression

10.1 Measures of Correlation

Many times in research, it is important to explore the relationship between two quantitative variables. *Correlation* is a statistical tool desired towards measuring the degree of linear relationship (association) between two quantitative variables. If the value of one variable changes when the value of another variable changes, then the variables are said to be *correlated*.

Before looking at a more detailed statistical approach, let us present a graphical mechanism for examining the relationship between two quantitative variables. The simplest way is to plot the pair of values (x_i, y_i), $i = 1, 2, \cdots, N$ on the xy plane, known as *scatter plot* (or *scatter diagram*). If the relationship between the two variables can be described by a straight line, then the relationship is called *linear* otherwise it is known as *non-linear*.



Such a plot gives some idea about the presence and absence of correlation, and the nature (direct or inverse) of correlation. But, it will not indicate about the strength or degree of relationship between the variables.

Example 10.1. Based of the Ethiopian DHS 2016, the total fertility rate is highest in Somali regional state (7.2 children per woman) and lowest in Addis Ababa city adminstration (1.8 children per woman). Following table presents the distribution of modern contraception usage among married women aged 15-49 and the total fertility rate for the 3 years preceding the survey, Ethiopia DHS 2016, in each of the 9 administrative regions and 2 city administrations of the country.

Region	% Use of Contraception	TFR
Tigray	35.2	4.7
Affar	11.6	5.5
Amhara	46.9	3.7
Oromia	28.1	5.4
Somali	1.4	7.2
Benishangul-Gumuz	28.4	4.4
SNNPR	39.6	4.4
Gambela	34.9	3.5
Harari	29.3	4.1
Addis Ababa	50.1	1.8
Dire Dawa	29.1	3.1

Obtain the scatter plot of TFR versus the percentage use of modern contraception, and try to identify their relationship.

10.1.1 Covariance

Covariance is a measure of the joint variation between two quantitative variables. That is, it measures the way in which the values of the two variables vary together.

Recall the population variance of a certain variable x is defined as $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sigma_{xx}$ which is estimated by the sample variance given by $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_{xx}$.

Similarly the population covariance between two variables x and y is defined as

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\left(\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right).$$

Consequently, the sample covariance is given by:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right).$$

The value of a covariance can be any real number (negative, zero or positive). If the covariance is zero, there is *no linear* relationship between the two variables. If the covariance is positive, there is a *direct linear* relationship (an increase in the value of one variable leads to an increase in the value of the other variable). For example, the relationship between birth weight of infants and gestational age is expected to be positive.

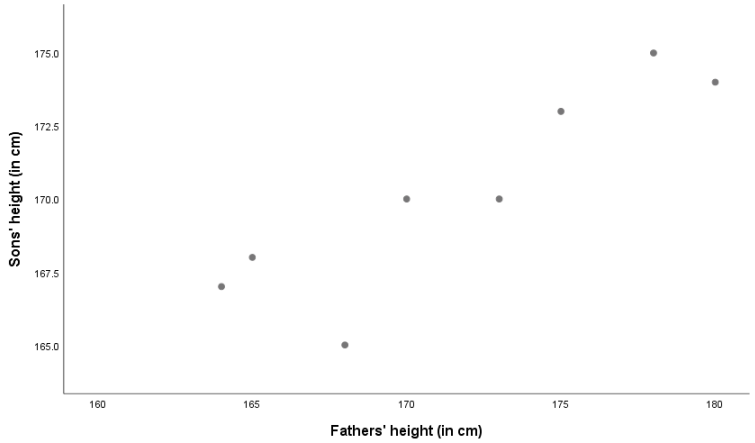
If the covariance is negative, there is an *inverse linear* relationship between the variables (an increase in the value of one variable implies a decrease in the value of the other variable). For example, the relationship between heart rate and age of individuals is expected to be negative.

Example 10.2. A researcher wants to find out if there is a relationship between the height of sons and the height of their fathers. In other words, do taller fathers have taller sons? The researcher took a random sample of 8 fathers and their 8 sons. Their height in centimeters is given below.

Father Height (x)	168	173	165	170	175	178	180	164
Son Height (y)	165	170	168	170	173	175	174	167

Draw the scatter plot and what can you say from the plot about the the relationship between the two variables. Also, find the covariance and interpret the result.

Solution: The scatter plot of the height of sons and the height of their fathers is:



The graph clearly shows there is a positive linear relationship between the height of sons and the height of their fathers.

To find the covariance between the two variables, the necessary calculations are presented in the following table:

No.	x	y	xy
1	168	165	27720
2	173	170	29410
3	165	168	27720
4	170	170	28900
5	175	173	30275
6	178	175	31150
7	180	174	31320
8	164	167	27388
Total	$\sum_{i=1}^{8} x_i = 1373$	$\sum_{i=1}^{8} y_i = 1362$	$\sum_{i=1}^{8} x_i y_i = 233883$

Therefore, the estimated (sample) covariance is:

$$s_{xy} = \frac{1}{8-1}\left(\sum_{i=1}^{8} x_i y_i - \frac{1}{8}\sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i\right) = \frac{1}{7}\left[233883 - \frac{1}{8}1373(1362)\right] = 18.54.$$

Since the estimated covariance is about 19 which is greater than 0, there is a positive linear relationship between the height of sons and their fathers for the sample data. That is, it seems taller fathers have taller sons.

Note: The value of a covariance depends on the size of the observed values. For example, had the observed values of the height of sons and fathers been measured in meters, the sample covariance would be 0.001854. Therefore, the strength of the linear relationship between two variables could not be determined from the covariance value.

10.1.2 Correlation Coefficient

The *coefficient of correlation*, which was developed by Karl Pearson, is a measure of the *degree* or *strength* of the linear association between two variables. It is defined as a ratio of the *covariance* between the two variables and the *product of the standard deviations* of each variable. The population correlation coefficient is denoted by the Greek letter ρ, rho:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Depending on the sign of a covariance, a correlation coefficient can be positive or negative. But, the value lies between the limits -1 and +1; that is, $-1 \leq \rho \leq 1$. The sign indicates the direction of the relationship and the absolute value indicates the strength of the relationship.

- If $\rho = 0$, there is no linear relationship between the two quantitative variables.
- If $\rho > 0$, there is a positive (direct) linear relationship between the variables (as the value of one increases, the value of the other variable increases). For example, the relationship between birth weight of infants and gestational age is expected to be positive.
- If $\rho < 0$, there is a negative (inverse) linear relationship between the variables (as the value of one increases, the value of the other variable decreases). For example, the relationship between heart rate and age of individuals is expected to be negative.
- If $\rho \approx \pm 1$, there is a strong positive ($\rho \approx 1$) or negative ($\rho \approx -1$) linear relationship between the variables.

The sample correlation coefficient is denoted by r:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This can also be written as:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Example 10.3. Recall example 10.2 and find the correlation coefficient and interpret.

Solution: The necessary calculations to calculate the correlation coefficient are presented as follows:

No.	x	y	x^2	y^2	xy
1	168	165	28224	27225	27720
2	173	170	29929	28900	29410
3	165	168	27225	28224	27720
4	170	170	28900	28900	28900
5	175	173	30625	29929	30275
6	178	175	31684	30625	31150
7	180	174	32400	30276	31320
8	164	167	26896	27889	27388
Total	$\sum_{i=1}^{8} x_i = 1373$	$\sum_{i=1}^{8} y_i = 1362$	$\sum_{i=1}^{8} x_i^2 = 235883$	$\sum_{i=1}^{8} y_i^2 = 231968$	$\sum_{i=1}^{8} x_i y_i = 233883$

$$
\begin{aligned}
r &= \frac{8\sum_{i=1}^{8} x_i y_i - \sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i}{\sqrt{8\sum_{i=1}^{8} x_i^2 - \left(\sum_{i=1}^{8} x_i\right)^2}\sqrt{8\sum_{i=1}^{8} y_i^2 - \left(\sum_{i=1}^{8} y_i\right)^2}} \\
&= \frac{8(233883) - 1373(1362)}{\sqrt{8(235883) - (1373)^2}\sqrt{8(231968) - (1362)^2}} \\
&= 0.892
\end{aligned}
$$

Since the sample correlation coefficient is positive and large (near to 1), there is a strong positive linear relationship between the height of sons and their fathers. In other words, taller fathers have taller sons for the sample data.

Notes:

- Although, the sign of the correlation and covariance are the same, the correlation is ordinarily easier to interpret as:
  - its *magnitude is bounded*, that is, $-1 \leq r \leq 1$.
  - it is *unitless* (its value is independent of the measurement units of the variables).
  - it *takes the variability into account.*
- But, it has also some disadvantages:
  - Correlation does not measure *non-linear* relationships. Two variables might have a perfect non-linear relationship, but, the correlation coefficient would still be zero. For example, even if $y = x^2; -4 < x < 4$ is an exact quadratic relationship, yet r is nearly zero. (Why?)
  - Zero correlation does not necessarily indicate independence of two variables. For example, the correlation between y and x would be zero if $y = x^2; -4 < x < 4$, but, it does not mean y and x have no relationship at all. If x and y are statistically

independent, the correlation coefficient between them will be zero; but the converse is not always true. In other words, *zero correlation does not necessarily imply independence.*

- A strong correlation does not necessarily imply *cause and effect* relationship.
- It *cannot be extrapolated beyond the observed range of values* of the variables.
- It is also *highly affected by extreme values (outliers)* and thus can sometimes be misleading.

Inference for Correlation Coefficient, *rho*

Having plotted the data, and established that it is plausible the two variables are associated linearly, it is necessary to decide whether the observed correlation could have arisen by chance or really significant.

Like other sample statistics, the sample correlation coefficient r is a random variable with mean $E(r) = \rho$. The standard error is $\text{SE}(r) = \sigma_r = \sqrt{\frac{1-\rho^2}{n-2}}$ which is estimated by $\widehat{\text{SE}}(r) = \hat{\sigma}_r = \sqrt{\frac{1-r^2}{n-2}}$.

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

Option 2: $H_0 : \rho = 0$ vs $H_1 : \rho < 0$

Option 3: $H_0 : \rho = 0$ vs $H_1 : \rho > 0$

Step 2: Specify the level of significance α and obtain the critical value based on the t distribution. The critical value for a two sided test is $t_{\alpha/2}(n-2)$ whereas the critical value for a one sided test is $t_{\alpha}(n-2)$.

Step 3: Use the t test statistic and obtain its calculated value:

$$t = \frac{r - \rho}{\sqrt{(1-r^2)/(n-2)}} \sim t(n-2).$$

Step 4: Decision: If the absolute value of the calculated value is greater than the critical value, H_0 should be rejected.

Step 5: Conclusion.

Example 10.4. Consider the data in example 10.2 and using the summary statistics given in example 10.3, test the significance of the correlation coefficient.

Solution: The estimated correlation coefficient is given as $r = 0.892$. The estimated standard error of the sample correlation coefficient is $\widehat{\text{SE}}(r) = \sqrt{\frac{1-0.892^2}{8-2}} = 0.1845$.

Step 1: Hypothesis:

$H_0 : \rho = 0$. There is no significant linear relationship between the heights of sons and their fathers.

$H_1 : \rho \neq 0$. There is a significant linear relationship between the heights of sons and their fathers.

Step 2: Assuming $\alpha = 0.05$, $t_{0.025}(6) = 2.447$.

Step 3: The calculated value of the t test statistic is:

$$t = \frac{0.892 - 0}{0.1845} = 4.835.$$

Step 4: Decision: As $t_{cal} = 4.835 > t_{0.025}(6) = 2.447$, the null hypothesis of no significant linear association should be rejected.

Step 5: Conclusion: There is a positive linear relationship between the height of sons and their fathers at 5% level of significance. In particular, it can be concluded that taller fathers have taller sons at 5% level of significance.

The $(1 - \alpha)100\%$ confidence interval for the population correlation coefficient ρ is given by:

$$\left[r \pm t_{\alpha/2}(n-2)\ \widehat{\text{SE}}(r)\right] = \left[r \pm t_{\alpha/2}(n-2)\sqrt{\frac{1-r^2}{n-2}}\right].$$

Example 10.5. Consider again the data in example 10.2 and construct the 95% confidence interval for the correlation coefficient.

Solution: The 95% confidence interval for the correlation coefficient for the correlation coefficient of the height of sons and their fathers is:

$$\left[0.892 \pm 2.447\sqrt{\frac{1-0.892^2}{8-2}}\right] = (0.4405,\ 1.3435) = (0.4405,\ 1).$$

Clearly, since the confidence interval is larger than 0, there is a positive linear relationship between the height of sons and the height of their fathers.

Exercise 10.1. Recall example 10.1. Determine the covariance and correlation coefficient, and interpret the results. Also, test the significance of the correlation coefficient.

10.1.3 Spearman's Rank Correlation

It is not always possible to take measurements on units or objects. Many characters are expressed in comparative terms such as beauty, smartness, temperament, $\cdots$. In such cases the units are ranked pertaining to that particular character instead of taking measurements on them. Sometimes, the units are also ranked according to their quantitative measure. In these type of studies, two situations arise, (i) the same set of units is ranked according two characters, (ii) two judges give ranks to the same set of units independently pertaining to one character. In both these situations we get paired ranks for a set of units. For example, the students are ranked according to their marks in Mathematics and Statistics. Two judges rank the girls independently in a beauty competition. In all these situations, the usual Pearson's correlation coefficient cannot be obtained.

Suppose that a group of n individuals is given grades or ranks with respect to two characteristics separately. Let R_{x_i} and R_{y_i} be the ranks of the i^{th}, $i = 1, 2, \cdots, n$, individual on the two characteristics. Then, the Spearman's rank correlation coefficient is given by:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \text{ where } d_i = R_{x_i} - R_{y_i}.$$

Note that $-1 \leq r_s \leq 1$.

Example 10.6. The ranks of 10 students in two courses Biostatistics and Epidemiology are given below. Calculate the rank correlation and interpret.

Biostatistics	5	2	9	8	1	10	3	4	6	7
Epidemiology	6	4	7	9	3	8	2	1	10	5

Note: The above formula is used when all the ranks within each characteristics are unique. But, for repeated ranks, a correction factor is required. Specifically, tied values should be assigned the average of the ranks they would receive if the values were unequal. For example, if the forth and fifth-ranked values are equal, then each value should be assigned a rank of $\frac{4+5}{2} = 4.5$. Similarly, if the seventh, eighth and ninth-ranked values are tied, then each value should be assigned a rank of $\frac{7+8+9}{3} = 8$.

Example 10.7. Obtain the rank correlation for the following data.

x	85	74	85	50	65	78	74	60	74	90
y	78	91	78	58	60	72	80	55	68	70

Ans: $r_s = -0.545$

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

Option 2: $H_0 : \rho = 0$ vs $H_1 : \rho < 0$

Option 3: $H_0 : \rho = 0$ vs $H_1 : \rho > 0$

Step 2: Specify the level of significance α and obtain the critical value based on the t??? distribution. The critical value for a two sided test is $rs_{\alpha/2}(n-2)$ whereas the critical value for a one sided test is $rs_{\alpha}(n-2)$.

Step 3: Use the t test statistic and obtain its calculated value:

$$rs = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \sim t(n-2).$$

Step 4: Decision: If the absolute value of the calculated value is greater than the critical value, H_0 should be rejected.

Step 5: Conclusion.

10.2 Simple Linear Regression

Regression may be defined as the estimation of the unknown value of one variable from the known value(s) of one or more variables. The variable whose values are to be estimated is known as *dependent*[1] variable while the variables whose are used in determining the value of the dependent variable are called *independent*[2] variables.

If the relationship between the two variables can be described by a straight line then the regression is known as *linear regression* otherwise it is called *non-linear*.

A linear regression involving only two variables (one dependent and one independent) is called *simple linear regression* and a linear regression analysis that involves more than two variables (one dependent and two or more independents) is called *multiple linear regression*.

10.2.1 Representation of the Model

A *simple linear regression* model is a linear function of a single explanatory variable. The model which is used to estimate the expected value (mean) of a dependent variable Y for any given value of an independent variable X is called a *linear regression of* Y *on* X and can be written as:

$$y_i = \alpha + \beta x_i + \varepsilon_i;\ i = 1, 2, \cdots, N$$

where

y_i is the i^{th} actual value of the dependent variable,

x_i is the i^{th} actual value the independent variable,

α is the intercept (constant) of the model,

β is the slope (rate of change) of the model,

ε_i is the i^{th} value of the error term.

This model is called *population regression model*. The term $\alpha + \beta x_i$ is the fixed (deterministic) part of the model. But, the response variable Y and the error term ε_i are random variables.

The error term ε_i is generally assumed to be normally distributed with mean 0, and variance σ^2 (called *error variance*), that is, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Consequently, $E(Y_i|X_i) = \mu_{Y_i|X_i} = \alpha + \beta X_i$ and $V(Y_i|X_i) = \sigma_{Y_i|X_i} = V(\varepsilon_i) = \sigma^2$. Therefore, $Y_i|X_i \sim \mathcal{N}(\alpha + \beta X_i, \sigma^2)$.

In this setting, α, β and σ^2 are population parameters to be estimated based on sample data. They are interpreted as follows:

- α is the expected value (mean) of the dependent variable when the value of the independent variable is zero.
- β is the change (increment or decrement) in the expected value (mean) of the dependent variable when the value of the independent variable increases by 1 unit.

[1] A dependent variable is also called an outcome or a response variable

[2] An independent variable is also called a factor, an exposure variable, covariate or predictor

Note: The sign of β is the same as to that of the covariance and correlation coefficient. If β is positive, then there is a direct linear relationship between the two variables (that is, the expected value of the dependent variable increases as the value of the independent variable increases). If β is negative, there is an inverse linear relationship between the two variables (that is, the expected value of the dependent variable decreases as the value of the independent variable increases). But, if β is zero, it shows there is no linear relationship between the two variables (that is, the mean value of the dependent variable cannot be determined from the value of the independent variable).

10.2.2 Estimation of the Intercept α and Slope β

Assume a sample of n subjects is taken observing values y_i of the response variable and x_i of the explanatory variable. Then, the interest is to choose values $\hat{\alpha}$ and $\hat{\beta}$ that estimate the intercept and slope parameters α and β, respectively. The fitted (estimated) regression model is, therefore, written as:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i;\ i = 1, 2, \cdots, n$$

where

- $\hat{y}_i$ is the i^{th} fitted (estimated) value of the dependent variable.
- x_i is the i^{th} actual value of the independent variable.
- $\hat{\alpha}$ is the estimated intercept.
- $\hat{\beta}$ is the estimated slope.

The most common and widely used method of estimation is called *ordinary least squares* (OLS) that minimize the distances of the data points to the fitted line. Now, for each observed response y_i, with a corresponding independent variable x_i, the fitted (estimated) value is $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. The objective is to minimize the sum of the squared distances of each observed response to its fitted value, called *Sum Squares of Error* (SSE):

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2$$

The estimates of the parameters can be obtained as:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n}x_i$$

Example 10.8. The weight (in kilograms) of a random sample of 10 infants of age up to 6 weeks is recorded as follows.

Age (in weeks)	1	2	2	3	5	4	5	4	6	3
Weight (in kilograms)	4.1	3.5	3.6	4.0	5.0	5.2	4.5	4.6	4.8	3.8

1. Estimate the regression model of weight of infants on their age.
2. Interpret the estimated intercept and slope.
3. What would be the predicted weight of an infant if the age is 8 weeks?

Solutions: The necessary calculations are presented in the following table:

No	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	1	4.1	1	16.81	4.1
2	2	3.5	4	12.25	7.0
3	2	3.6	4	12.96	7.2
4	3	4.0	9	16.00	12.0
5	5	5.0	25	25.00	25.0
6	4	5.2	16	27.04	20.8
7	5	4.5	25	20.25	22.5
8	4	4.6	16	21.16	18.4
9	6	4.8	36	23.04	28.8
10	3	3.8	9	14.44	11.4
Total	$\sum_{i=1}^{10} x_i = 35$	$\sum_{i=1}^{10} y_i = 43.1$	$\sum_{i=1}^{10} x_i^2 = 145$	$\sum_{i=1}^{10} y_i^2 = 188.95$	$\sum_{i=1}^{10} x_iy_i = 157.2$

1. Thus, using the formula for estimating the slope and intercept parameters:

$$\hat{\beta} = \frac{10\sum_{i=1}^{10} x_iy_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10\sum_{i=1}^{10} x_i^2 - \left(\sum_{i=1}^{10} x_i\right)^2} = \frac{10(157.2) - 35(43.1)}{10(145) - (35)^2} = 0.282$$

and

$$\hat{\alpha} = \frac{1}{10}\sum_{i=1}^{10} y_i - \hat{\beta}\frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}(43.1) - (0.282)\frac{1}{10}(35) = 3.323.$$

The estimated regression model is: $\hat{y}_i = 3.323 + 0.282x_i;\ i = 1, 2, \cdots, 10.$

2. Based on the sample data, the mean weight of a newly born infant is about 3.323 kilograms. Also, as the age of an infant increases by 1 week, its mean weight will increase by 0.282 kilograms.
3. The predicted mean weight of an infant if the age is 8 weeks is $\hat{y}_i = 3.323 + 0.282(8) = 5.579$ kilograms.

10.2.3 Estimation of the Error Variance σ^2

The error variance σ^2 is estimated by the sample error variance s^2 is:

$$s^2 = \text{MSE} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\text{SSE} \quad \text{where} \quad \text{SSE} = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right).$$

This estimated variance s^2 can be thought of as the 'average' squared distance from each observed response to the fitted line. The word average is in quotes since the denominator is $n-2$ and not n. The smaller s^2, the closer the observed responses fall to the line and the better the predicted values will be.

Example 10.9. For the simple linear regression model in example 10.8, find the estimated error variance.

Solution: First, let us calculate the variances of each variable and the covariance between the two variables: $s_x^2 = \frac{1}{9}[145 - \frac{1}{10}(35^2)] = 2.5$, $s_y^2 = \frac{1}{9}[188.95 - \frac{1}{10}(43.1^2)] = 0.354$ and $s_{xy} = \frac{1}{9}[157.2 - \frac{1}{10}(35)(43.1)] = 0.706$. Thus, $SSE = (10-1)(0.354 - \frac{0.706^2}{2.5}) = 1.392$. Therefore, the estimated error variance of the estimate is $s^2 = \frac{SSE}{10-2} = \frac{1.392}{8} = 0.174$ and the standard error of the estimate is $s = \sqrt{0.174} = 0.417$.

10.2.4 Inferences for the Slope β

Recall that in the simple linear regression model, $E(Y_i|X_i) = \alpha + \beta x_i$. In this model, β represents the change in the mean of the response variable Y_i, as the independent variable x_i increases by 1 unit. Note that if $\beta = 0$, $E(Y_i|X_i) = \alpha$, which implies the mean of the response variable is the same at all values of x_i. This implies that knowledge of the level of the independent variable does not help predict the response variable.

Under the assumptions stated previously, namely that $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$, the estimator $\hat{\beta}$ has a sampling distribution that is normal with mean β (the true value of the parameter), and variance $V(\hat{\beta}) = \sigma_{\hat{\beta}}^2 = \sigma^2/[(n-1)s_x^2]$. That is, $\hat{\beta} \sim \mathcal{N}\{\beta, \sigma^2/[(n-1)s_x^2]\}$.

The standard error of $\hat{\beta}$ is, therefore, $\text{SE}(\hat{\beta}) = \sigma_{\hat{\beta}} = \sigma/\sqrt{(n-1)s_x^2}$ which is estimated by $\widehat{\text{SE}}(\hat{\beta}) = \hat{\sigma}_{\hat{\beta}} = s/\sqrt{(n-1)s_x^2}$. Therefore, here is the procedure if β is equal to some fixed value, say β_0. In virtually all real-life cases, $\beta_0 = 0$.

Testing for the Slope β

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

Option 2: $H_0 : \beta = 0$ vs $H_1 : \beta < 0$

Option 3: $H_0 : \beta = 0$ vs $H_1 : \beta > 0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value for a two sided test is $t_{\alpha/2}(n-2)$ whereas the critical value for a one sided test is $t_\alpha(n-2)$.

Step 3: Use the t test statistic and obtain its calculated value:

$$t = \frac{\hat{\beta} - \beta}{s/\sqrt{(n-1)s_x^2}} \sim t(n-2).$$

Step 4: Decision: If the absolute value of the calculated value is greater than the critical value, H_0 should be rejected.

Step 5: Conclusion.

Confidence Interval for the Slope β

The $(1-\alpha)100\%$ confidence interval for β is given by: $[\hat{\beta} \pm t_{\alpha/2}(n-2)\ \widehat{\text{SE}}(\hat{\beta})]$. That is,

$$\left[\hat{\beta} \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{(n-1)s_x^2}}\right].$$

Example 10.10. Test the significance of the slope parameter and construct the 95% confidence interval using the data in example 10.8.

Solution: The estimated standard error is obtained as $s = 0.417$ and variance of x is $s_x^2 = 2.5$.

Step 1: Hypothesis:

$H_0 : \beta = 0$. There is no significant linear relationship between weight of infants and their age.

$H_1 : \beta \neq 0$. There is a significant linear relationship between weight of infants and their age.

Step 2: Assuming $\alpha = 0.05$, $t_{0.025}(8) = 2.306$.

Step 3: The t test statistic is:

$$t = \frac{0.282 - 0}{0.417/\sqrt{(10-1)2.5}} = 3.208.$$

Step 4: Decision: Since, $t_{cal} = 3.208 > t_{0.025}(8) = 2.306$, H_0 should be rejected.

Step 5: Conclusion: There is a significant positive linear relationship between weight of infants and their age at 5% significance level.

Also, the 95% confidence interval for the effect (β) of age on the weight of infants is:

$$[0.282 \pm 2.306(0.0879)] = (0.282 \pm 0.19883) = (0.0793,\ 0.4847).$$

Therefore, as the age of the infant increases by 1 week, the mean weight of the infant increases by between 0.0793 and 0.4847 at 5% significance level.

10.2.5 The ANOVA approach to Regression

Consider the deviations of the individual responses y_i from their overall mean $\bar{y}$. The deviations could be broken into two parts; the deviation of the fitted value $\hat{y}_i$ from the overall mean $\bar{y}$, and the deviation of the observed value y_i from its fitted value $\hat{y} = \hat{\alpha} + \hat{\beta}x_i$. This is similar in nature to the way of partitioning the total variation in the one-way ANOVA case. It can be written as:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Squaring both sides and then summing over all the n observed and fitted values yields:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST: } df=n-1} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR: } df=1} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE: } df=n-2}.$$

These three pieces are called the *sum squares of total* (SST), the *sum squares of regression* (SSR), and the *sum squares of error* (SSE) respectively. The SST represents the total variation in the observed responses, SSR represents the amount of the total variation that is 'accounted for' (explained) by taking into account the explanatory variable x, and SSE represents the variation in the observed responses around the fitted regression model (unexplained variation).

Hence, SST=SSR+SSE. This decomposition can be used to test the hypothesis $H_0 : \beta = 0$ vs $H_1 : \beta = 0$ and is also useful in subsequent sections when there are more than one independent variable. The setup of the ANOVA table is as follows.

Source of variation	Sum squares (SS)	*df*	Mean squares (MS)	*F*
Regression	$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Note that SST=$(n-1)s_y^2$ and also SSE=$(n-2)s^2$. Then, SSR=SST-SSE. The testing procedure is similar to as usual.

Step 1: State both the null and alternative hypotheses:

$H_0 : \beta = 0.$
$H_1 : \beta \neq 0.$

Step 2: Specify the level of significance α and obtain the critical value $F_\alpha[1, n-2]$.

Step 3: Use the F test statistic and obtain its calculated value:

$$F = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2).$$

Step 4: Decision: If $F_{cal} > F_\alpha(1, n-2)$, the null hypothesis $H_0 : \beta = 0$ has to be rejected.

Step 5: Conclusion.

Note that there already have a procedure for testing the hypothesis $H_0 : \beta = 0$ vs $H_1 : \beta = 0$ (see section 10.2.4 on Inferences for the Slope β), but this is an important lead-in to multiple linear regression.

Example 10.11. Recall example 10.8 and test the significance of the slope parameter using the ANOVA approach.

Solution: Previously, it is calculated that $s_y^2 = 0.354$ and SSE=1.392. Thus, SST=9(0.354) = 3.186. Therefore, SSR=SST-SSE=3.186-1.392=1.794. Therefore, the ANOVA table is:

Source of variation	Sum squares (SS)	*df*	Mean squares (MS)	*F*
Regression	1.794	1	1.794	10.31
Error	1.392	8	0.174	
Total	3.186	9		

For testing the slope parameter, the test procedure is:

Step 1: Hypotheses:

$H_0 : \beta = 0.$

$H_1 : \beta \neq 0.$

Step 2: $\alpha = 5\%$ and $F_{0.05}(1, 8) = 5.318$.

Step 3: The calculated value of the F test statistic is:

$$F = \frac{1.794/1}{1.392/8} = 10.31.$$

Step 4: Decision: Since $F_{cal} = 10.31 > F_{0.05}(1, 8)$, the null hypothesis $H_0 : \beta = 0$ has to be rejected $\alpha = 5\%$.

Step 5: Conclusion. The age of infants significantly determines their weight.

10.2.6 Coefficient of Determination

Another measure of association in regression analysis is a goodness-of-fit of the model called *coefficient of determination*. The *coefficient of determination* represents the proportion of the total variation in the response variable that is 'accounted' for by fitting the regression on the independent variable(s). It is defined in a general formula as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Particularly, for a simple linear regression (two variables case), the coefficient of determination is just the square of the sample correlation coefficient, r^2. This measure is between 0 and 1 ($0 \leq r^2 \leq 1$) and it measures the proportion of the total variation in the dependent variable explained by the independent variable x. If the value of r^2 approaches to 1, it means the regression model is a good fit and if it approaches 0, the model is a bad fit to the data.

In example 10.3, the correlation coefficient between the height of sons and their fathers is 0.892. Then, the coefficient of determination is $r^2 = 0.796$. This means about 80% of the variation in the height of sons is explained by the height of their fathers.

Example 10.12. A study was reported in a medical journal suggesting that the peak heart rate of an individual can reach during intensive exercise decreases with age. A cardiologist wants to do his own study. The next 9 patients were given a stress test on the treadmill at 6 miles per hour and their ages and their heart rates were recorded as follows:

Age	30	30	40	20	20	45	30	45	50
Heart Rate	190	180	180	200	195	170	185	175	165

1. Identify the dependent and independent variables.
2. Estimate the regression model.
3. What is the peak heart rate of an 80 year old man who is given a similar stress test?
4. Calculate the coefficient of correlation and coefficient of determination, and interpret the results.

Solutions: The required summary statistics are shown in the following table.

No.	x	y	x^2	y^2	xy
1	30	190	900	36100	5700
2	30	180	900	32400	5400
3	40	180	1600	32400	7200
4	20	200	400	40000	4000
5	20	195	400	38025	3900
6	45	170	2025	28900	7650
7	30	185	900	34225	5550
8	45	175	2025	30625	7875
9	50	165	2500	27225	8250
Total	$\sum_{i=1}^{9} x_i = 310$	$\sum_{i=1}^{9} y_i = 1640$	$\sum_{i=1}^{9} x_i^2 = 11650$	$\sum_{i=1}^{9} y_i^2 = 299900$	$\sum_{i=1}^{9} x_i y_i = 55525$

1. The dependent variable is heart rate and the independent variable is age.
2. The estimated model is $\hat{y}_i = 216.37 - 0.99x_i$; $i = 1, 2, \cdots, 9$.
3. The estimated peak heart rate of an 80 year old man is $\hat{y} = 216.37 - 0.99 \times 80 = 137.17$.
4. The correlation coefficient and coefficient of determination are $r = -0.95$ and $r^2 = 0.90$, respectively.

10.3 Multiple Linear Regression

Outcome variables in medical research are usually affected by a multitude of factors. Fortunately, the simple linear regression is easily extended to multiple regression. A multiple linear regression is a regression of a continuous dependent variable on many, say k, independent

variables (quantitative, qualitative or a mixture of both). In that case, the corresponding model is written as:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i;\ i = 1, 2, \cdots, N$$

where $x_{ij}; j = 1, 2, \cdots, k$ is the i^{th} value of the j^{th} independent variable. The same assumptions are made as before in terms of ε, specifically, it is normally distributed with mean 0 and variance σ^2.

Just as before, α, β_1, β_2, $\cdots$, β_k, and σ^2 are unknown parameters that must be estimated from sample data. The parameter α is the usual intercept of the model representing the mean response when all the independent variables are zero. The parameter β_j is the (partial) slope corresponding to the j^{th} explanatory variable and it represents the change in the mean response when the j^{th} explanatory variable changes by 1 unit assuming all other explanatory variables are held constant.

The estimated model will be of the form:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik};\ i = 1, 2, \cdots, n.$$

Note: In a regression analysis, it is valid to include binary independent variables (that take only two values, say 0 and 1) directly into the model. The test of significance of the parameter corresponding to a binary independent variable is equivalent with a t test for testing the difference between two population means. One mean is the mean value of the dependent variable for those subjects with $x = 0$ and the other mean value of the dependent variable is for those with $x = 1$.

Example 10.13. Consider a survey of anaemia in women. They had a blood sample taken and their haemoglobin (Hb) level and packed cell volume (PCV) measured. They were also asked their age, and whether or not they had experienced the menopause. The results from a random sample of 20 women are given in the following table.

Number	Hb (g/dl)	PCV (%)	Age (years)	Menopause (1=Yes, 0=No)
1	11.1	35	20	0
2	10.7	45	22	0
3	12.4	47	25	0
4	14.0	50	28	0
5	13.1	31	28	0
6	10.5	30	31	0
7	9.6	25	32	0
8	12.5	33	35	0
9	13.5	35	38	0
10	13.9	40	40	1
11	15.1	45	45	0
12	13.9	47	49	1
13	16.2	49	54	1
14	16.3	42	55	1
15	16.8	40	57	1
16	17.1	50	60	1
17	16.6	46	62	1
18	16.9	55	63	1
19	15.7	42	65	1
20	16.5	46	67	1

The parameter estimates of the multiple linear regression of haemoglobin on PCV, Age and Menopause are provided in the following table.

Variable	Parameter Estimate	Standard Error
Constant	5.215	1.572
PCV	0.097	0.035
Age	0.111	0.030
Menopause	-0.024	0.954

Write out the estimated model and interpret the parameter estimates for the given sample.

Solution: The estimated model is $\widehat{\text{Hb}}_i = 5.215 + 0.097\ \text{PCV}_i + 0.111\ \text{Age}_i - 0.024\ \text{Menopause}_i;\ i = 1, 2, \cdots, 20$

- As the PCV increases by 1%, the mean haemoglobin level of anemic women increases by 0.097g/dl assuming the age and menopause status remain constant.
- Controlling for PCV and menopause, the mean haemoglobin level of anemic women increases by 0.111g/dl as their age increase by 1 year.
- The mean haemoglobin level of those anemic women who experienced menopause (who are in the post menopause) decreases by 0.024g/dl as compared to those who did not experience menopause (who are in the pre-menopause) provided that the PCV and age are held constant.

10.3.1 Testing the Joint Significance all Predictors

After fitting a multiple linear regression model, the next task is to test whether all the independent variables included in the model are jointly significant. That is, the hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (all the k variables are not jointly significant) vs $H_1 :$ not H_0 (at least one of the k variables is significant).

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST: } df=n-1} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^n}_{\text{SSR: } df=k} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE: } df=n-p}.$$

The test statistic is based on the ANOVA approach to regression:

Source of variation	Sum squares (SS)	*df*	Mean squares (MS)	F
Regression	$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$\text{MSR} = \frac{\text{SSR}}{k}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-(k+1)$	$\text{MSE} = \frac{\text{SSE}}{n-(k+1)}$	
Total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

The testing procedure is similar to as usual.

Step 1: State both the null and alternative hypotheses:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. All the k predictors are not jointly significant.

$H_1 :$ **not** H_0. At least one of the k predictors is significant.

Step 2: Specify the level of significance α and obtain the critical value $F_\alpha[k, n-(k+1)]$.

Step 3: Use the F test statistic and obtain its calculated value:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n-(k+1)]} = \frac{\text{MSR}}{\text{MSE}} \sim F[k, n-(k+1)].$$

Step 4: Decision: If $F_{cal} > F_\alpha[k, n-(k+1)]$, the null hypothesis of no significant contribution of the k explanatory variables has to be rejected.

Step 5: Conclusion.

Example 10.14. For the linear regression model estimated using the data in example 10.13, SSR=93.304 and SSE=16.308. Construct the ANOVA table and then test the joint significance of all the three independent variables.

Solution: The population model is of the form $\text{Hb}_i = \alpha + \beta_1\ \text{PCV}_i + \beta_2\ \text{Age}_i + \beta_3\ \text{Menopouse}_i + \varepsilon_i;\ i = 1, 2, \cdots, N$. Thus, the ANOVA table for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is:

Source of variation	Sum squares (SS)	*df*	Mean squares (MS)	F
Regression	93.304	3	31.101	30.515
Error	16.308	16	1.019	
Total	109.612	19		

The critical value is $F_{0.05}(3, 16) = 3.2389$ which is smaller than the calculated value $F = 30.515$. As a result, at least one of the three parameters is significantly different from zero at $\alpha = 0.05$. That means, at least one of the independent variables is significantly associated with the haemoglobin level of women at $\alpha = 0.05$.

10.3.2 Testing the Significance each Parameter

In a regression analysis, once the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (all the k variables are not significant) is rejected, the next task is to identify which parameter(s) {variable(s)} is (are) significant which is (are) not. For this purpose, an individual t test is performed for each $\beta_j; j = 1, 2, \cdots, k$.

Step 1: State both the null and alternative hypotheses. The three possible options are:

Option 1: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

Option 2: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j < 0$

Option 3: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j > 0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value for a two sided test is $t_{\alpha/2}[n - (k + 1)]$ whereas the critical value for a one sided test is $t_{\alpha}[n - (k + 1)]$.

Step 3: Use the t test statistic and obtain its calculated value:

$$t = \frac{\hat{\beta}_j - \beta_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim t[n - (k + 1)].$$

Step 4: Decision: If $|t_{cal}| > t_{\alpha}[n - (k + 1)]$, $H_0 : \beta_j = 0$ should be rejected.

Step 5: Conclusion.

The $(1 - \alpha)100\%$ confidence interval for $\beta_j; j = 1, 2, \cdots, k$ is given by: $\{\hat{\beta}_j \pm t_{\alpha/2}[n - (k + 1)]\ \widehat{\text{SE}}(\hat{\beta}_j)\}$.

Example 10.15. Test the significance of each parameter considered in example 10.13.

Solution: The critical value for testing the significance of each parameter is $t_{0.025}(16) = 2.120$.

Variable	Estimate	Standard Error	t Statistic	95% CI
Constant	5.215	1.572	3.318*	(1.8824, 8.5476)*
PCV	0.097	0.035	2.815*	(0.0228, 0.1712)*
Age	0.111	0.030	3.661*	(0.0474, 0.1746)*
Menopause	-0.024	0.954	-0.025	(-2.0465, 1.9985)

Therefore, of the three independent variables, menopause is not significantly associated with the haemoglobin level of women, in the presence of PCV and age in the model, at 5% significance level.

Remark: The mean and standard deviation of haemoglobin level for pre-menopausal women are 12.29 and 1.57 g/dl, whereas those for post-menopausal women are 16.36 and 0.63 respectively. Now the null hypothesis $H_0 : \mu_{pre} = \mu_{post}$ can be tested using a two independent samples t test (as described in Chapter ?). The value of the test statistic becomes $t = 7.3$ which has to be compared with the critical value $t_{0.025}(18) = 2.101$. Obviously, the test is significant at $\alpha = 0.05$ in contradiction with the regression analysis results. Why?

Women who have experienced the menopause clearly will be older than women who have not. Therefore, because of the confounding effect of age with menopause, menopause becomes insignificant in the presence of age. The next task is to remove menopause from the model and then refit the model with PCV and Age only. Note also that if the model is fitted excluding the age variable, menopause will be significant.

Such cross-sectional data are not good to examine whether there is a difference in the haemoglobin level of post-menopausal and pre-menopausal women. Best would be a longitudinal study which measured haemoglobin levels in women before and after their menopause (paired sample).

10.3.3 Coefficient of Multiple Determination

Recall the coefficient of determination is defined in a general formula as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{s^2}{s_y^2}.$$

This measure for a multiple linear regression is called *coefficient of multiple determination* and shows the proportion of the total variation in the responses explained by the set of independent variables. This R^2 does not take into account the loss of degrees of freedom from the inclusion of additional explanatory variables in the model.

The coefficient of multiple determination adjusted for the degrees of freedom is called *adjusted coefficient of multiple determination*(R^2_{adj}) and the formula is a little bit modified as:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right].$$

Example 10.16. In the analysis of the data in example 10.13, menopause was not found significant in the presence of the other two variables, PCV and age. Consequently, menopause is excluded from the model. That is, the final model is fitted consisting only PCV and age, and the parameter estimates are presented in the following table. The SSR of this final model is 93.304 and the SST is 109.612.

Variable	Parameter Estimate	Standard Error
Constant	5.239	1.206
PCV	0.097	0.033
Age	0.110	0.016

Write out the estimated model. Test the overall significance of the model and construct the ANOVA table. Then, test for the significance of each individual parameter. Find out the

coefficient of multiple determination and the adjusted one.

Solution: The coefficient of multiple determination of the model is $R^2 = \frac{93.304}{109.612} = 0.851$. The adjusted coefficient of multiple determination is $R^2_{\text{adj}} = 1 - (1 - 0.851)\left(\frac{19}{17}\right) = 0.8335$. Therefore, about 83.35% of the variation in the haemoglobin levels among anemic women is explained by both the PCV and age. For this particular example, there is no even a small change in the SSR (consequently R^2) by adding menopause in the regression model.

10.3.4 Including Multinomial Predictors

It would not be appropriate to include a categorical explanatory variable with more than two categories in a regression model as if it were quantitative or binary. This is because the codes used to represent the various categories are merely identifiers and have no numeric significance. In such case, a set of binary variables, called *design* (*dummy*, *indicator*) variables, should be created to represent such a polytomuous variable.

Suppose, for example, that one of the explanatory variable is marital status with three categories: "Single", "Married", "Separated". In this case, taking one of the categories as a reference (comparison group), two design variables (d_1 and d_2) are required to represent marital status in a regression model. For example, if the category "Single" is taken as a reference, the two design variables, d_1 and d_2 are set to 0; when the subject is "Married", d_1 is set to 1 while d_2 is still 0; when the marital status of the subject is "Separated", $d_1 = 0$ and $d_2 = 1$ are used. The following table shows this example of design variables for marital status:

	Design Variables	
Marital Status	Married (d_1)	Separated (d_2)
Single	0	0
Married	1	0
Separated	0	1

In general, if a polytomuous variable X has m categories, then $m - 1$ design variables are needed. The $m - 1$ design variables are denoted as d_u and the coefficients of those design variables are denoted as β_u, $u = 1, 2, \cdots, m - 1$. Thus, for a simple linear regression on a multinomial independent variable would be:

$$y_i = \alpha + \beta_1 d_{i1} + \beta_2 d_{i2} + \cdots + \beta_{m-1} d_{i,m-1}.$$

Example 10.17. A short survey was conducted on a random sample of 23 patients to know the percentage level of their satisfaction by the medical treatment they were given. The patients were asked about three additional variables: their age, gender and education level. The recorded data are presented as follows where y= patient satisfaction in percentage, x_1= age in years, x_2= gender (0=male, 1=female) and x_3= education level (1= uneducated, 2=primary, 3=secondary, 4=tertiary).

No	y_i	x_{i1}	x_{i2}	x_{i3}	Indicator variables for education level (x_{i3})		
					Uneducated (d_{i31})	Primary (d_{i32})	Secondary (d_{i33})
1	26.1	52	0	1	1	0	0
2	36.5	49	0	1	1	0	0
3	46.1	42	0	1	1	0	0
4	47.2	38	0	1	1	0	0
5	49.0	55	0	1	1	0	0
6	51.0	34	0	1	1	0	0
7	52.5	44	0	2	0	1	0
8	66.4	36	0	2	0	1	0
9	48.0	50	0	2	0	1	0
10	54.6	45	0	2	0	1	0
11	66.7	40	1	2	0	1	0
12	57.9	36	0	3	0	0	1
13	57.0	53	1	3	0	0	1
14	60.5	43	1	3	0	0	1
15	89.4	28	1	3	0	0	1
16	89.1	29	1	3	0	0	1
17	60.3	33	1	4	0	0	0
18	67.5	43	0	4	0	0	0
19	70.7	41	0	4	0	0	0
20	77.7	29	1	4	0	0	0
21	77.0	29	1	4	0	0	0
22	79.2	33	1	4	0	0	0
23	88.6	29	1	4	0	0	0

The parameter estimates and their estimated standard errors of the linear regression of patient satisfaction on age, gender and education level are presented in the following table.

Variable	Parameter Estimate	Standard Error
Constant	103.043	10.796
Age (x_1)	-0.956	0.258
Female (x_2)	5.246	4.909
Uneducated (d_{31})	-17.383	5.912
Primary (d_{32})	-5.354	5.572
Secondary (d_{33})	-0.331	4.919
Tertiary	Ref.	

In addition, the SST and SSE of the model are 6194.496 and 1126.660, respectively.

1. Write out the estimated linear regression.
2. Construct the ANOVA table and test the overall significance of the model.
3. Determine the coefficient of multiple determination and adjusted coefficient of multiple determination, and interpret.
4. Examine the significance of each individual independent variable.
5. Interpret the (partial) slope of each independent variable.

Solution: The population regression model is of the form $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{31} d_{i31} + \beta_{32} d_{i32} + \beta_{33} d_{i33} + \varepsilon_i;\ i = 1, 2, \cdots, N$ where y_i is the satisfaction level of the i^{th} patient.

1. The estimated model is $\hat{y}_i = 103.043 - 0.956 x_{i1} + 5.246 x_{i2} - 17.383 d_{i31} - 5.354 d_{i32} - 0.331 d_{i33};\ i = 1, 2, \cdots, 23$

2. For testing the overall significance of the model, the null hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_{31} = \beta_{32} = \beta_{33} = 0$. The ANOVA table is:

Source of variation	SS	df	MS	F
Regression	SSR=5067.836	5	MSR=1013.567	$F = 15.294$
Error	SSE=1126.660	17	MSE= 66.274	
Total	SST=6194.496	22		

   The critical value is $F_{0.05}(5, 17) = 2.8100$. Since $F_{cal} = 15.294 > F_{0.05}(5, 17)$, it indicates at least one of the three independent variables is significantly associated with patient satisfaction at 5% level of significance.

3. The coefficient of multiple determination is $R^2 = \frac{5067.836}{6194.496} = 0.8181$ and the adjusted coefficient of multiple determination is $R^2_{\text{adj}} = 1 - (1 - 0.8597)\left(\frac{22}{17}\right) = 0.7646$. About 76.46% of the variation in the patients satisfaction is explained by the three variables jointly (age, gender and education level)

4. For identifying the significant and insignificant parameters of the model, the t test statistics is used as usual. At $\alpha = 0.05$, the critical value for each individual parameter is $t_{0.025}(17) = 2.110$.

Variable	Estimate	Standard Error	t Statistic	95% CI
Constant	103.043	10.796	9.545*	(80.2634, 125.8226)*
Age (x_1)	-0.956	0.258	-3.707*	(-1.5004 -0.4116)*
Female (x_2)	5.246	4.909	1.069	(-5.1120, 15.6040)
Uneducated (d_{31})	-17.383	5.912	-2.940*	(-29.8573, -4.9087)*
Primary (d_{32})	-5.354	5.572	-0.961	(-17.1109, 6.4029)
Secondary (d_{33})	-0.331	4.919	-0.067	(-10.7101, 10.0481)
Tertiary	Ref.			

   Hence, age is significant but gender is not, at $\alpha = 0.05$. Also, since one of the three design (dummy) variables of education level is significant, education level is a significant factor of patient satisfaction at $\alpha = 0.05$.

5. Controlling for all other variables included in the model:

   - As the age of a patient increases by 1 year, his/her mean satisfaction level decreases by 0.956%.
   - For the given sample, the mean satisfaction level of female patients (relative to male patients) increases by 5.246%. But, this difference is not significant for the population in general at $\alpha = 5\%$.

- The interpretation of each design variable of education level is made relative to the tertiary education which is the reference category. The mean level of satisfaction between uneducated patients and tertiary educated patients is significantly different, but the other two design variables representing primary vs tertiary and secondary vs tertiary are not significant at $\alpha = 5\%$. Therefore, it can be concluded the mean level of satisfaction of uneducated patients decreases by 17.383% relative to educated (primary, secondary or tertiary) patients.

Chapter 11

Logistic Regression

11.1 Binary Logistic Regression

A binary logistic regression predicts the probability of success in a dichotomous dependent variable, for example, whether a person will develop a disease or whether a certain patient will survive a surgical procedure. There could be one or more independent variables which can be, as usual, either continuous, categorical or both.

11.1.1 The Logistic Function

Recall the logistic function is

$$f(z) = \frac{1}{1 + \exp(-z)};\ -\infty < z < \infty.$$

When $z = -\infty$, $f(-\infty) = 0$ and when $z = \infty$, $f(\infty) = 1$. Note also that $f(0) = \frac{1}{2}$.

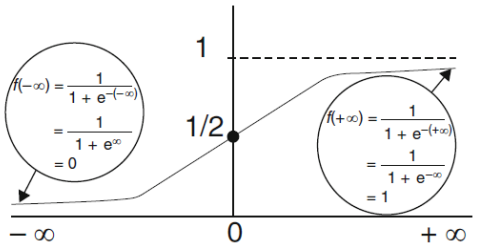


Figure 11.1: Plot of the Logistic Function

Thus, as the figure describes the range of $f(z)$ is between 0 and 1 (that is, $0 \leq f(z) \leq 1$) regardless of the value of z. Therefore, it is suitable for use as a probability model. Hence, to indicate that $f(z)$ is a probability value, the notation $\pi(z)$ can be used instead. That is,

$$\pi(z) = \frac{1}{1 + \exp(-z)};\ -\infty < z < \infty$$

where $\pi(z) = P(Y = 1|Z = z)$.

11.1.2 The Simple Logistic Regression

To begin with the simplest model, consider the case of a binary outcome and a single predictor variable x. Hence, in the logistic function, z is expressed as a function (mostly linear function) of the explanatory variable. That is, $z_i = g(x_i) = \alpha + \beta x_i$. As a result, the simple logistic probability model is:

$$\pi(x_i) = \frac{1}{1 + \exp[-(\alpha + \beta x_i)]}$$

where $\pi(x_i) = P(Y_i = 1|X_i = x_i) = 1 - P(Y_i = 0|X_i = x_i)$. It can also be written as

$$\pi(x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

As can be seen from this model, the relationship between the response variable (probability of success) and the explanatory variable is not linear. However, it can be linearized by using different transformations of the probability of success and the most common one is called a *logit* or *log-odds* transformation.

The Logit Transformation

In the previous chapter, odds is defined as the ratio of the probability of success to the probability of failure. Hence, the odds of successes at a particular value x_i of the explanatory variable is

$$\Omega(x_i) = \frac{\pi(x_i)}{1 - \pi(x_i)}.$$

Thus, the odds of successes for a simple logistic regression model is $\Omega(x_i) = \exp(\alpha + \beta x_i)$. If $\Omega(x_i) = 1$, then a success is as likely as a failure at the particular value x_i of the explanatory variable. If $\Omega(x_i) > 1$, then $\log \Omega(x_i) > 0$, a success is more likely to occur than a failure. On the other hand, if $\Omega(x_i) < 1$, then $\log \Omega(x_i) < 0$, a success is less likely than a failure.

The *logit* of the probability of success is given by the natural logarithm of the odds of successes. Therefore, the logit of the probability of success is a linear function of the explanatory variable. Thus, the simple logistic model is

$$\text{logit } \pi(x_i) = \log\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \alpha + \beta x_i$$

This is particulary called the *logit* model as it uses the *logit* transformation or the *log-odds scaling* (or *logit link function*) which is a reasonable choice for binary response models.

To clarify the relationship between probabilities, odds, and the natural log of the odds (logit), the following table includes probability values along with their corresponding odds as well as the natural log of the odds, log(odds). The table demonstrates that as the probability gets smaller and approaches 0, the odds also approach 0 while the log odds approach $-\infty$(negative infinity), and as the probability gets larger and approaches 1, the odds also get larger while the log odds approach $+\infty$(positive infinity). Therefore, while probabilities can theoretically vary from 0 to 1 with a midpoint of 0.5, the corresponding odds can theoretically vary from 0 to $+\infty$ with 1 corresponding to the probability midpoint, and the natural log of the odds can theoretically vary from $+\infty$ to $+\infty$ with 0 corresponding to the probability midpoint.

$\pi(x_i)$	$1-\pi(x_i)$	$\Omega(x_i)$	logit $\pi(x_i)$
0.001	0.999	0.001	-6.908
0.010	0.990	0.010	-4.605
0.100	0.900	0.111	-2.198
0.200	0.800	0.250	-1.386
0.300	0.700	0.429	-0.846
0.400	0.600	0.667	-0.405
0.500	0.500	1.000	0.000
0.600	0.400	1.500	0.405
0.700	0.300	2.333	0.847
0.800	0.200	4.000	1.386
0.900	0.100	9.000	2.197
0.990	0.010	99.000	4.595
0.999	0.001	999.000	6.907

Thus, the range of the log(odds) more closely resembles the standard normal distribution in that it is unbounded, has a midpoint of 0, and is symmetric around the midpoint.

There are also other models that are used in practice. The probit model or the complementary log-log model might be appropriate when the logit model does not fit the data well.

Interpretation of the Parameters

The parameters, α and β, are the intercept and slope of the logit model, respectively. Because the predicted value, probability, in logistic regression is different from the predicted value, mean, in linear regression, the interpretations of the intercept, α, and slope, β, are also somewhat different as these must be interpreted in the context of the predicted response.

The logit model is monotone depending on the *sign* of the parameter β. Its sign determines whether the probability of success is increasing or decreasing, as shown in figure 11.2, when the value of the explanatory variable increases. When the parameter β is zero, Y is independent of X. Then, $\pi(x_i) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ which is identical for all x_i, so the curve becomes a straight (horizontal) line.

The slope parameter of a logit model can be interpreted in terms of an odds ratio. From logit $\pi(x_i) = \alpha + \beta x_i$, an odds is an exponential function of x_i. This provides a basic interpretation for the *magnitude* of the slope parameter β. The odds at x_i is $\Omega(x_i) = \exp(\alpha + \beta x_i)$ and the odds at $x_i + 1$ is $\Omega(x_i + 1) = \exp[\alpha + \beta(x_i + 1)]$. Thus, the odds ratio is

$$\theta = \frac{\Omega(x_i + 1)}{\Omega(x_i)} = \exp(\beta).$$

This value is the multiplicative effect of the odds of successes due to a unit change in the explanatory variable. That is, for every one unit increase in x_i, the odds changes by a factor of $\exp(\beta)$. Similarly, for an m units increase in x_i, say $x_i + m$ versus x_i, the corresponding odds ratio becomes $\exp(m\beta)$.

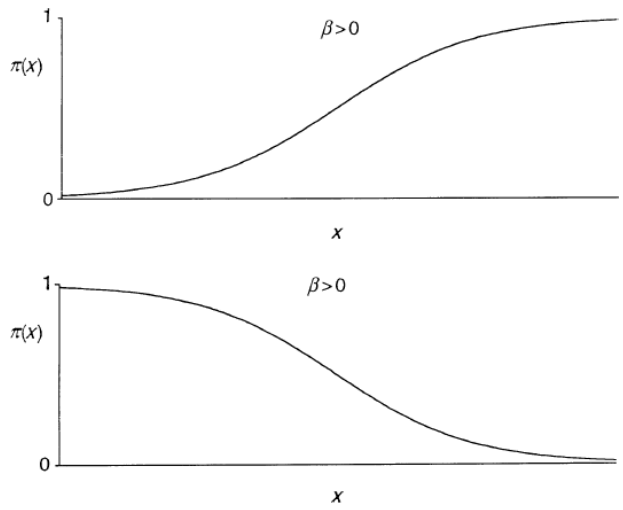


Figure 11.2: Plot of the Logistic Probability

Also, the parameter β determines the *slope* (*rate of change* or *marginal effect*) of the probability of success at a certain value of the explanatory variable. This *rate of change* (*marginal effect*) at a particular x_i value is described by drawing a straight line tangent to the curve at that point. That line will have a *slope* of $\pi(x_i)[1-\pi(x_i)]\beta$. This is the *rate of change* (*slope* or *marginal effect*) of $\pi(x_i)$ at a particular value of x_i. For example, the line tangent to the curve at x_i for which $\pi(x_i) = 0.5$ has a *slope* $(0.5)(1-0.5)\beta = 0.25\beta$. If $\pi(x_i)$ is 0.9 or 0.1, it has a *marginal effect* 0.09β. As the probability of success approaches either 0 or 1, the *rate of increment* (*decrement*) of the curve approaches to 0. The steepest *slope* of the curve is attained at x_i for which the probability of success is 50%. Thus, solving

$$\frac{1}{1+\exp[-(\alpha+\beta x_i)]} = 0.5$$

for x_i implies $x_i = -\frac{\alpha}{\beta}$. This x_i value is called *medial effective level* (EL_{50}). At this value, each outcome has a 50% chance of occurring.

The intercept α is, not usually of particular interest, used to obtain the odds (probability) at $x_i = 0$. Also, by centering the explanatory variable at 0 {that is, replacing x_i by $(x_i - \bar{x})$}, α becomes the logit at that mean, and thus $\pi(\bar{x}) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$.

The estimated logistic regression model is written as:

$$\text{logit } \hat{\pi}(x_i) = \log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = \hat{\alpha} + \hat{\beta}x_i.$$

Example 11.1. For studying the effect of age (continuous variable) on the occurrence of hypertension (coded as 1 for presence and 0 for absence), a sample of 13 individuals were examined. The ages (in years) of persons having hypertension are 45, 60, 60, 60, 55, 55, 20 and those who do not have hypertension are 20, 20, 18, 30, 55, 18. For these data, the following parameter estimates were obtained.

Variable	Parameter Estimate
Intercept	-3.4648
Age	0.0931

1. Write the model that allows the prediction of the probability of having hypertension at a given age.
2. What is the estimated probability of having hypertension at the minimum and maximum ages of this study.
3. What is the estimated probability of having hypertension at the age of 35. Also find the odds of having hypertension at this age.
4. Find the estimated probability of success at the sample mean and determine the incremental change (marginal effect) at that point.
5. Write out the estimated logit model.
6. Find the estimated odds ratio of having hypertension and interpret.
7. Determine the estimated median effective level (EL_{50}) and interpret.

Solution: Let $Y=$ hypertension and $X=$ age. Then $\hat{\pi}(x_i) = \widehat{P}(Y = 1|x_i)$ is the estimated probability of having hypertension, $Y = 1$, given the age x_i of an individual i.

1. The estimated probability of hypertension at a given age is given by:
$$\hat{\pi}(x_i) = \frac{\exp(-3.4648 + 0.0931x_i)}{1 + \exp(-3.4648 + 0.0931x_i)}.$$
2. The estimated probability of having hypertension at the age of 35 years is $\hat{\pi}(35) = 0.4486$ and its estimated odds is $\widehat{\Omega}(35) = 0.8136$.
3. The mean age of the sample is 39.69 years. The estimated probability of having hypertension at this mean age is $\hat{\pi}(39.69) = 0.5573$ and the rate of change (marginal effect) at this mean value is $\hat{\pi}(39.69)[1 - \hat{\pi}(39.69)]\hat{\beta} = 0.5573(1 - 0.5573)(0.0931) = 0.0230$. The probability of having hypertension at the age of 39.69 years increases by 2.30%.
4. The estimated logit model is written as
$$\log\left[\frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)}\right] = -3.4648 + 0.0931x_i.$$
5. The estimated odds ratio is $\exp(\hat{\beta}) = \exp(0.0931) = 1.0976$. Hence, the odds (risk) of having hypertension is 1.0976 times larger for every year older an individual is. In other words, as the age of an individual increases by one year, the odds (risk) of developing hypertension increases by a factor of 1.0976. Or the odds (risk) of having hypertension increases by $[\exp(0.0931) - 1] \times 100\% = 9.76\%$ every year.
6. The estimated median effective level, the estimated age in years at which an individual has a 50% chance of having hypertension, is $\widehat{\text{EL}}_{50} = -\hat{\alpha}/\hat{\beta} = -(-3.4648)/0.0931 = 37.2159$.

11.1.3 Logit Models with Categorical Predictors

Like ordinary regression, logistic regression extends to include qualitative explanatory variables, often called *factors*.

Binary Predictors

For simplicity, let us consider a binary predictor, X, representing an exposure which refers to a risk factor such as smoking (smoker, nonsmoker) or patient characteristics like sex (male, female), residence (urban, rural). The simple logit model is

$$\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \alpha + \beta x_i \text{ where } x_i = \begin{cases} 1, & \text{exposed group;} \\ 0, & \text{unexposed group.} \end{cases}$$

From this model, the odds in the exposed group is given by $\Omega(1) = \exp(\alpha + \beta)$ and the odds in the unexposed group is $\Omega(0) = \exp(\alpha)$. This implies, $\exp(\beta)$ as the odds ratio associated with an exposure (exposed $x_i = 1$ versus unexposed $x_i = 0$), which is equivalent to the odds ratio in a 2×2 table.

In other words, the estimates of the parameters of a logit model for a 2×2 table can be easily determined from the cell frequencies. Consider the 2×2 table below. Setting $x_i = 0$ for the

	Response		
Exposure	Success (1)	Failure (0)	Total
Exposed (1)	n_{11}	n_{10}	n_{1+}
Unexposed (0)	n_{01}	n_{00}	n_{0+}
Total	n_{+1}	n_{+0}	n

unexposed group and then solving for α gives the estimated intercept of the logit model in terms of the natural logarithm of the odds of successes in the unexposed group. That is,

$$\hat{\alpha} = \log\left[\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)}\right] = \log\left(\frac{n_{01}}{n_{00}}\right).$$

Similarly, the estimate of the slope of the logit model is derived as the natural logarithm of the odds ratio associated with an exposure by setting $x_i = 1$ for the exposed group,

$$\hat{\beta} = \log\left[\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}\right] - \hat{\alpha} = \log\left[\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}\right] - \left[\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)}\right] = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right).$$

As discussed before, the *marginal effect* of a continuous explanatory variable, which is very useful when interpreting a binary logit model, is the partial derivative of the probability of success with respect to that variable.

Similarly, the *discrete change* of a binary explanatory variable is the difference in estimated probabilities when the variable value is 1 and when it is 0. Note that *marginal effects* and *discrete changes* look similar but are not equal in conceptual and numerical senses.

Example 11.2. In a study of cigarette smoking and risk of lung cancer, a logistic regression analysis is used to determine how much greater the odds are finding cases of the diseases among subjects who have ever smoked than among those who have never smoked.

	Lung Cancer		
Smoking	Case (1)	Control (0)	Total
Yes (1)	77	123	200
No (0)	54	171	225
Total	131	294	425

Given the parameter estimates from a statistical software as follows:

Variable	Parameter Estimate
Intercept	-1.1527
Smoking	0.6843

Write out the estimated model and interpret the slope estimate. Also find the discrete change.

Solution: Let $Y=$ lung cancer where

$$y_i = \begin{cases} 1, & \text{if the subject develops lung cancer - Case;} \\ 0, & \text{otherwise (if the subject does not develop lung cancer) - Control.} \end{cases}$$

For the explanatory variable, let $X=$ smoking status where

$$x_i = \begin{cases} 1, & \text{if the subject had ever smoked - Smoker;} \\ 0, & \text{otherwise (if the subject had never smoked) - Nonsmoker.} \end{cases}$$

Thus, $\hat{\pi}(x_i)$ is the estimated probability of developing lung cancer, $Y = 1$, given the smoking status, $x_i = 1$ for smokers and $x_i = 0$ for non-smokers. The parameter estimates can also be obtained manually. The estimates are

$$\hat{\alpha} = \log\left(\frac{n_{01}}{n_{00}}\right) = \log\left(\frac{54}{171}\right) = -1.1527$$

and

$$\hat{\beta} = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right) = \log\left[\frac{77(171)}{123(54)}\right] = 0.6843.$$

Thus, the estimated model is

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.1527 + 0.6843x_i.$$

The estimated odds ratio is $\exp(0.6843) = 1.9824$. Thus, smokers are 1.9824 times (98.24%) more likely to develop lung cancer as compared to nonsmokers. Or the odds (risk) of developing lung cancer is 98.24% higher for smokers than for nonsmokers {the odds (risk) of developing lung cancer among smokers is 98.24% higher than that of among nonsmokers}.

The discrete change is $\hat{\pi}(1) - \hat{\pi}(0) = 0.3850 - 0.2400 = 0.1450$. The probability of developing lung cancer increases by 14.50% for smokers relative to nonsmokers.

Example 11.3. The following table presents the cross-classification of 1464 HIV/AIDS patients involved in **?** study by defaulting (Yes, No) and gender (Female, Male).

	Defaulter		
Gender	Yes (1)	No (0)	Total
Female (1)	189	741	930
Male (0)	142	392	534
Total	331	1133	1464

The parameter estimates are provided in the following table:

Variable	Parameter Estimate
Intercept	-1.0154
Smoking	-0.3508

Write out the estimated model and interpret the estimated slope.

Solution: Let $Y=$ defaulter where $y_i = 1$ if the patient was defaulted from the HAART treatment and $y_i = 0$ otherwise (if the patient was active on the treatment). Let $X=$ gender of the patient where $x_i = 1$ if the patient is female and $x_i = 0$ otherwise (if the patient is male).

Then $\hat{\pi}(x_i)$ is the estimated probability of the patient being defaulted from the HAART treatment. The estimated model is

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.0154 - 0.3508x_i.$$

The odds ratio is $\exp(-0.3508) = 0.7041$. This means that female patients are 0.7041 times (29.59%) less likely to default from HAART treatment as compared to male patients. Or, the risk of being defaulted is 29.59% lower for female patients than for male patients (the risk of being defaulted for male patients is 42.02% higher than the risk of being defaulted for female patients).

Polytomous Explanatory Variables

When there is a binary response variable and a polytomous explanatory variable, the data can be presented using a $2 \times m$ table. Taking one of the category of the explanatory variable as a reference, $m-1$ stratified 2×2 tables can be constructed. Then the parameter estimates corresponding to each design variable can be easily determined from each table. If category m is taken as a reference, then $\hat{\alpha} = \log\left(\frac{n_{m1}}{n_{m0}}\right)$ and $\hat{\beta}_u = \log\left(\frac{n_{u1}n_{m0}}{n_{u0}n_{m1}}\right)$; $u = 1, 2, \cdots, m-1$.

Example 11.4. Given the following cross-classified data on race and coronary heart disease for 100 subjects.

	Race				
CHD	White	Black	Hispanic	Other	Total
Present (1)	5	20	15	10	50
Absent (0)	20	10	10	10	50
Total	25	30	25	20	100

Software provides the following parameter estimates.

Variable	Parameter Estimate
Intercept	-1.386
Black (d_1)	2.079
Hispanic (d_2)	1.792
Other (d_3)	1.386

Specify the design variables for race using "white" as a reference group. Calculate the parameter estimates manually from the cell counts of the contingency table and compare them with the software estimates. Write out the estimated model and interpret.

Solution: Since the variable "Race" has four categories, three design variables are needed.

	Design Variables		
Race	Black (d_1)	Hispanic (d_2)	Other (d_3)
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1

Let $\hat{\pi}(x_i)$ be the estimated probability of developing coronary heart disease given the race of an individual. Thus,

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.386 + 2.079d_{i1} + 1.792d_{i2} + 1.386d_{i3}.$$

Blacks are about 8 $\{\exp(2.079) = 7.996\}$ times more likely to develop coronary heart disease as compared to whites. Similarly, the odds (risk) of coronary heart disease for hispanics is about 6 $\{\exp(1.792) = 6.001\}$ times that of whites. The odds (risk) of coronary heart disease for other (neither blacks nor hispanics) races is about 4 $\{\exp(1.386) = 3.999\}$ times that of whites.

11.1.4 Multiple Logistic Regression

Suppose there are k explanatory variables (categorical, continuous or both) to be considered simultaneously. Then, the multiple logit model is written as:

$$\text{logit } \pi(\boldsymbol{x}_i) = \log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

Similar to the simple logistic regression, $\exp(\beta_j)$ represents the (partial) odds ratio associated with an exposure if X_j is binary (exposed $x_{ij} = 1$ versus unexposed $x_{ij} = 0$); or it is the odds ratio due to a unit increase if X_j is continuous ($x_{ij} = x_{ij} + 1$ versus $x_{ij} = x_{ij}$).

If the j^{th} explanatory variable, X_j, has m_j levels, then the multiple logit model with k variables would be

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \sum_{u=1}^{m_j-1} \beta_{ju} d_{iju} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik}$$

where the d_{ju}'s are the $m_j - 1$ design variables and $\beta_{ju}, u = 1, 2, \cdots, m_j - 1$ are their corresponding parameters.

Note: Odd ratios obtained from a simple logistic regression (one independent variable) are called *crude odds ratios* (COR) and odd ratios obtained from a multiple logistic regression (two or more independent variables) are called *adjusted odds ratios* (AOR).

Example 11.5. To determine the effect of vision status (1=vision problem, 0=no vision problem) and driver education (1=took driver education, 0=did not take driver education) of a driver on car accident (did the subject had an accident in the past year?), the following parameter estimates are obtained from a sample of 210 individuals. Interpret the results.

Variable	Parameter Estimate
Intercept	0.1110
Vision	1.7139
Education	-1.5001

Solution: Let $Y=$ car accident ($y_i = 1$ if a subject had an accident in the past year and $y_i = 0$ if a subject had not an accident in the past year). Let $X_1=$ vision problem ($x_{i1} = 1$ if a subject had a vision problem and $x_{i1} = 0$ if a subject had not a vision problem). Let $X_2=$ driver education ($x_{i2} = 1$ if a subject took driver education, $x_{i2} = 0$ if a subject did not take driver education).

The estimated logit model is $\log\left[\frac{\hat{\pi}(\boldsymbol{x}_i)}{1-\hat{\pi}(\boldsymbol{x}_i)}\right] = 0.1110 + 1.7139x_{i1} - 1.5001x_{i2}$. The estimated odds ratio associated with vision problem is $\exp(1.7139) = 5.551$. The odds of having accident for a person with vision problem is 5.551 times that of a person with no vision problem assuming driver education the same. In other words, drivers who have vision problem are 5.551 times more likely to have an accident as compared to those with no vision problem.

Also, the estimated odds ratio associated with education problem is $\exp(-1.5001) = 0.223$. Drivers who took driving education are 0.223 times less likely to have an accident as compared to those who did not take driving education assuming the same vision status, that is, the risk of having an accident for those who took a driving education is 77.7% lower than those who did not take a driving education.

11.2 Inference for Logistic Regression

Recall the binary response probability given the values of the explanatory variables is

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\sum_{j=0}^{k} \beta_j x_{ij})}{1 + \exp(\sum_{j=0}^{k} \beta_j x_{ij})} \tag{11.1}$$

where $x_{i0} = 1$ for all $i = 1, 2, \cdots, n$. Equivalently using the logit transformation, it can be written as

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right] = \sum_{j=0}^{k} \beta_j x_{ij}. \tag{11.2}$$

11.2.1 Parameter Estimation

The goal of logistic regression model is to estimate the $k+1$ unknown parameters of the model. This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is largest.

Given a data set with n independent observations. Suppose these responses are grouped into m unique covariate patterns (called populations). Then each binary response $Y_i;\ i = 1, 2, \cdots, m$ has an independent Binomial distribution with parameter n_i and $\pi(\boldsymbol{x}_i)$, that is,

$$P(Y_i = y_i) = \binom{n_i}{y_i}\pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i};\ y_i = 0, 1, 2, \cdots, n_i$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$ for population i and $\sum_{i=1}^{m} n_i = n$. Then, the joint probability mass function of the vector of m Binomial random variables, $\boldsymbol{Y}^t = (Y_1, Y_2, \cdots, Y_m)$, is the product of the m Binomial distributions

$$P(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{m}\binom{n_i}{y_i}\pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i}. \tag{11.3}$$

The joint probability mass function in equation (11.3) expresses the values of $\boldsymbol{y}$ as a function of known, fixed values for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^t$. The likelihood function has the same form as the probability mass function, except that it expresses the values of $\boldsymbol{\beta}$ in terms of known, fixed values for $\boldsymbol{y}$. Thus,

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m}\binom{n_i}{y_i}\pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i} \tag{11.4}$$

Note that the combination term does not contain any of the $\pi(\boldsymbol{x}_i)$. As a result, it is essentially constant that can be ignored: maximizing the equation without the combination term will come to the same result as if it was included. Therefore, equation (11.4) can be written as:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m}\pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i} \tag{11.5}$$

and it can be re-arranged as:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m}\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right]^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i} \tag{11.6}$$

By substituting the odds of successes and probability of failure in equation (11.6), the likelihood function becomes

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m}\left[\exp\left(y_i\sum_{j=0}^{k}\beta_j x_{ij}\right)\right]\left[1+\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right]^{-n_i} \tag{11.7}$$

Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log-likelihood function and vice versa. Thus, taking the natural logarithm of equation (11.7) gives the log-likelihood function:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{m} \left\{ y_i \sum_{j=0}^{k} \beta_j x_{ij} - n_i \log \left[1 + \exp \left(\sum_{j=0}^{k} \beta_j x_{ij} \right) \right] \right\} \tag{11.8}$$

To find the critical points of the log-likelihood function, first, equation (11.8) should be partially differentiated with respect to each $\beta_j; j = 0, 1, \cdots, k$ which results in a system of $k + 1$ nonlinear equations with the $k + 1$ unknown parameters as shown in equation (11.9) below:

$$\frac{\partial L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{m} [y_i x_{ij} - n_i \pi(\boldsymbol{x}_i) x_{ij}] = \sum_{i=1}^{m} [y_i - n_i \pi(\boldsymbol{x}_i)] x_{ij}; \quad j = 0, 1, 2, \cdots, k. \tag{11.9}$$

The maximum likelihood estimates for $\boldsymbol{\beta}$ can be, then, found by setting each of the $k + 1$ equation equal to zero and solving for each β_j. Since the second partial derivatives of the log-likelihood function:

$$\frac{\partial^2 L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j \partial \beta_h} = -\sum_{i=1}^{m} n_i \pi(\boldsymbol{x}_i)[1 - \pi(\boldsymbol{x}_i)] x_{ij} x_{ih}; \quad j, h = 0, 1, 2, \cdots, k \tag{11.10}$$

is negative semidefinite, the log-likelihood is a concave function of the parameter $\boldsymbol{\beta}$. In addition, equation (11.10) represents the variance-covariance matrix of the parameter estimates which is a function of $var(Y_i) = n_i \pi(\boldsymbol{x}_i)[1 - \pi(\boldsymbol{x}_i)]$.

These equations do not have a closed form solution. Several optimization techniques are available for finding the maximizing estimates of the parameters. Of these, the Newton-Raphson method is the one which is commonly used.

11.2.2 Overall Significance of the Model

Once a logistic regression model is estimated, the next task is to answer the question "Does the entire set of explanatory variables contribute significantly to the prediction of the response?". In this case, two models are to be fitted; one with all explanatory variables (full model) and the other with no explanatory variable (null model).

Likelihood-Ratio/Deviance Test

If the model has k explanatory variables (either binary or continuous), the null hypothesis of no contribution of all the k explanatory variables is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Let ℓ_0 denote the maximized value of the likelihood function of the null model which has only one parameter, that is, the intercept. That is, $\ell_0 = \ell(\hat{\beta}_0)$. Also let ℓ_M denote the maximized value of the likelihood function of the model M with all explanatory variables (having $k + 1$ parameters). Here, $\ell_M = \ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)$.

Then, the likelihood-ratio test statistic is $G^2 = -2\log(\ell_0/\ell_M) = -2(\log \ell_0 - \log \ell_M) \sim \chi^2(k)$. Deviance is -2 times the log-likelihood value of a model. Thus, $G^2 = D_0 - D_M \sim \chi^2(k)$.

Rejection of the null hypothesis, has an interpretation analogous to that in multiple linear regression using F test, indicates at least one of the k parameters is significantly different from zero.

Example 11.6. Suppose, a study was conducted with the objective of identifying the risk factors associated with HIV/AIDS HAART treatment defaulter patients. Of 1464 patients, 331 were defaulted and the remaining 1133 were actively following the treatment. Five variables which were considered as explanatory variables are age in years (Age), weight in kilograms (Weight), Gender (0=Female, 1=Male), Functional Status (0=Working, 1=Ambulatory, 2=Bedridden) and number of baseline CD4 counts (CD4). The parameter estimates and their corresponding standard errors are presented in the following table.

Variable	Parameter Estimate	Standard Error
Intercept	-0.3120	0.4299
Age	-0.0282	0.0080
Weight	-0.0051	0.0071
Gender	0.5372	0.1438
Ambulatory	0.4959	0.1448
Bedridden	1.2610	0.2882
Working	Ref.	
CD4	-0.0007	0.0004

The log-likelihood value of the null model is -782.5257 and the log-likelihood value of the full model is -753.2892. Test the significance of the entire five variables altogether.

Solution: The response variable takes the value $y_i = 1$ if the patient was defaulted and $y_i = 0$ otherwise (if the patient was on the treatment).

The design variables for Functional Status are:

	Design Variables	
Functional Status	Ambulatory (d_{41})	Bedridden (d_{42})
Working	0	0
Ambulatory	1	0
Bedridden	0	1

Now the model can be written as

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1\ \text{Age}_i + \beta_2\ \text{Weight}_i + \beta_3\ \text{Gender}_i + \beta_{41}\ \text{Ambulatory}_i + \beta_{42}\ \text{Bedridden}_i + \beta_5\ \text{CD4}_i$$

The null hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_{41} = \beta_{42} = \beta_5 = 0$. The test statistic value is $G^2 = -2(\log \ell_0 - \log \ell_M) = -2[-782.5257 - (-753.2892)] = 58.473$ which is greater than $\chi^2_{0.05}(6) = 12.592$. Therefore, H_0 should be rejected. At least one of the parameter is significantly different from zero.

11.2.3 Significance Test for Parameters

Once the null hypothesis of no contribution of all the explanatory variables to the model is rejected, there is a need to look at which of the variables are significant and which are not. The Wald test is used to identify the statistical significance of each coefficient (β_j) of the logit model. That is, it is used to test the null hypothesis $H_0 : \beta_j = 0$ which states that factor X_j does not have significant value added to the prediction of the response given that other factors are already included in the model. The test statistic for large sample size is, therefore,

$$Z_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim N(0,1).$$

Example 11.7. Recall example 11.6. Write out the estimated model and identify the significant explanatory variables using Wald test, and interpret the results.

Solution: We have that the estimated model is:

$$\begin{aligned}\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = &-0.3120 - 0.0282\ \text{Age}_i - 0.0051\ \text{Weight}_i + 0.5372\ \text{Gender}_i \\ &+ 0.4959\ \text{Ambulatory}_i + 1.2610\ \text{Bedridden}_i - 0.0007\ \text{CD4}_i\end{aligned}$$

The Wald test help us to identify those parameters which are responsible for rejection of the null hypothesis of all the parameters are zero. The value of the Wald test for each parameter which is obtained by dividing each parameter estimate by the corresponding standard error estimate is given in the following table.

Variable	Parameter Estimate	Standard Error	Wald Test
Intercept	-0.3120	0.4299	-0.7258
Age	-0.0282	0.0080	-3.5250*
Weight	-0.0051	0.0071	-0.7183
Gender	0.5372	0.1438	3.7357*
Ambulatory	0.4959	0.1448	3.4247*
Bedridden	1.2610	0.2882	4.3754*
Working	Ref.		
CD4	-0.0007	0.0004	-1.7500

As it can be seen from this table, age, gender and functional status (since both of the design variables are significant) are significant at 5% level of significance. When the age of the patient increases by one year, the odds of being defaulted decreases by a factor of $\exp(-0.0282) = 0.9723$ assuming all other variables are same. Also, males are $\exp(0.5372) = 1.7112$ times more likely to default than females, that is, the odds of being defaulted for males is 71.12% higher than that of females assuming the other variables constant. Again, assuming all other variables constant, ambulatory and bedridden patients are 1.6420 and 3.5290 times more likely to be defaulted than working patients, respectively.

Significance of a Polytomous Predictor

The Wald test considered above is used to identify the statistical significance of a binary or continuous explanatory variable. Whenever a multinomial explanatory variable is included

(excluded) in (from) the model, all of its design variables should be included (excluded); to do otherwise implies the variables are recorded. By just looking at the Wald statistics of the design variables, the contribution of the variable could not be determined. Hence, the Wald test can be not used to check the significance of such a variable, rather the likelihood-ratio test should be used.

If X_j has m categories, then the null hypothesis of no contribution of this multinomial variable is $H_0 : \beta_{j1} = \beta_{j2} = \cdots = \beta_{j,m-1} = 0$. The likelihood-ratio test statistic is $G^2 = -2(\log \ell_R - \log \ell_M) \sim \chi^2(m-1)$ where ℓ_R is the maximized likelihood value under H_0 (excluding the multinomial variable X_j) and ℓ_M is the maximized likelihood value of the full model.

Example 11.8. Again recall example 11.6. Test the significance of functional status.

Solution: Since functional status is a multinomial variable with $m = 3$ categories, wald test cannot be used for checking its significance. The null hypothesis is $H_0 : \beta_{41} = \beta_{42} = 0$. Here, β_{41} and β_{42} are the parameters associated with the two design variables of functional status; ambulatory and bedridden, respectively. Therefore, the model in example 11.6 is re-fitted without the two design variables of marital status. When fitted, the log-likelihood value becomes -765.7410.

The likelihood-ratio test statistic is $G^2 = -2(\log \ell_R - \log \ell_M) = -2[-765.7410 - (-753.2892)] = 24.9036$. Since this value is greater than $\chi^2_{0.05}(2) = 5.9915$, functional status has a significant contribution to the model.

11.2.4 Confidence Intervals

Confidence Intervals for Parameters

Confidence intervals are more informative than tests. A confidence interval for β_j results from inverting a test of $H_0 : \beta_j = \beta_{j0}$. The interval is the set of β_{j0}'s for which the z test statistic is not greater than $z_{\alpha/2}$. This means $|\hat{\beta}_j - \beta_{j0}| \leq z_{\alpha/2}|\widehat{\text{SE}}(\hat{\beta}_j)|$. This yields the confidence interval

$$\left[\hat{\beta}_j \pm z_{\alpha/2}\widehat{\text{SE}}(\hat{\beta}_j)\right]$$

for β_j; $j = 1, 2, \cdots, k$. As the point estimate of the odds ratio associated to X_j is $\exp(\hat{\beta}_j)$ and its confidence interval is

$$\left\{\exp\left[\hat{\beta}_j \pm z_{\alpha/2}\widehat{\text{SE}}(\hat{\beta}_j)\right]\right\}.$$

Example 11.9. Recall example 11.6 and construct the 95% confidence interval for each parameter and the corresponding odds ratio.

Solution: The critical value $z_{0.025} = 1.96$

Variable	$\hat{\beta}_j$	$\widehat{\text{SE}}(\hat{\beta}_j)$	95% CI for β_j	95% CI for $\text{OR}_j = \exp(\beta_j)$
Intercept	-0.3120	0.4299		
Age	-0.0282	0.0080	(-0.0439, -0.0125)*	(0.9570, 0.9876)*
Weight	-0.0051	0.0071	(-0.0190, 0.0088)	(0.9812, 1.0088)
Gender	0.5372	0.1438	(0.2554, 0.8190)*	(1.2910, 2.2682)*
Ambulatory	0.4959	0.1448	(0.2121, 0.7797)*	(1.2363, 2.1808)*
Bedridden	1.2610	0.2882	(0.6961, 1.8259)*	(2.0059, 6.2084)*
Working	Ref.			
CD4	-0.0007	0.0004	(-0.0015, 0.0001)	(0.9985, 1.0001)

Confidence Intervals for Predicted Probabilities

For summarizing the relationship, other characteristics may have greater importance such as $\pi(\boldsymbol{x}_i)$ at various $\boldsymbol{x}_i$ values. Consider the simple logistic model, logit $\hat{\pi}(x_i) = \hat{\alpha} + \hat{\beta}x_i$. For a fixed $x_i = x_0$, logit $\hat{\pi}(x_0) = \hat{\alpha} + \hat{\beta}x_0$ has a large standard error given by

$$\sqrt{\text{var}(\hat{\alpha}) + x_0^2 \text{ var}(\hat{\beta}) + 2x_0 \text{ cov}(\hat{\alpha}, \hat{\beta})}.$$

A $(1-\alpha)100\%$ confidence interval for logit $\pi(x_0)$ is

$$\left[(\hat{\alpha} + \hat{\beta}x_0) \pm z_{\alpha/2}\sqrt{\text{var}(\hat{\alpha} + \hat{\beta}x_0)}\right].$$

Substituting each end point into the inverse transformation

$$\pi(x_0) = \frac{\exp\{\text{logit}[\hat{\pi}(x_0)]\}}{1 + \exp\{\text{logit}[\hat{\pi}(x_0)]\}}$$

gives the corresponding interval for $\pi(x_0)$.

Example 11.10. Recall example 11.6, in which the estimated model is logit $\hat{\pi}(x_i) = -3.4648 + 0.0931x_i$. The variance-covariance matrix of the estimated parameters is:

$$\begin{pmatrix} 3.4037 & -0.0744 \\ & 0.0019 \end{pmatrix}$$

Find the 95% confidence interval for the odds ratio and for the probability of success at the age of 39.6923 years ($x_i = 39.6923$).

Solution: $\hat{\beta} = 0.0931$, $\widehat{\text{var}}(\hat{\alpha}) = 3.4037$, $\widehat{\text{var}}(\hat{\beta}) = 0.0019$ and $\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) = -0.0744$.

The 95% confidence interval for β is

$$\left[\hat{\beta} \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}\right] = \left(0.0931 \pm 1.96\sqrt{0.0019}\right) = (0.0077, 0.1785).$$

This implies, the confidence interval for the odds ratio is

$$[\exp(0.0077, 0.1785)] = [\exp(0.0077), \exp(0.1785)] = (1.0077, 1.1954).$$

Also, to construct the confidence interval for the proportion of having hypertension at the age of 39.6923 years, the estimated probability of having hypertension at the age of 39.6923 years is logit $\hat{\pi}(39.6923) = -3.4648 + 0.0931(39.6923) = 0.2306$ and its estimated variance is

$$\begin{aligned}\widehat{\text{var}}\{\text{logit } [\hat{\pi}(39.6923)]\} &= \widehat{\text{var}}(\hat{\alpha}) + 39.6923^2\,\widehat{\text{var}}(\hat{\beta}) + 2(39.6923)\,\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) \\ &= 3.4037 + 39.6923^2(0.0019) + 2(39.6923)(-0.0744) \\ &= 0.4909\end{aligned}$$

The 95% confidence interval for logit $\pi(39.6923)$ is $(0.2306 \pm 1.96\sqrt{0.4909}) = (-1.1427, 1.6039)$. Thus, the 95% confidence interval for the probability of hypertension at the age of 39.6923 years is

$$\left[\frac{\exp(-1.1427)}{1+\exp(-1.1427)}, \frac{\exp(1.6039)}{1+\exp(1.6039)}\right] = (0.2418, 0.8326).$$

This confidence interval is very wide which may be due to the small sample size, $n = 13$.

11.3 Model Checking

Once the variable selection process is addressed, then the selected model should be explored for assessing whether the assumptions of the probability model are satisfied. The diagnostic methods for logistic regression, like that of linear regression, mostly rely residuals which compare observed and predicted values. Goodness-of-fit statistics are often computed as an objective measures of the overall fit of a model. A model checked and if it is found lacking the fit, a new model is proposed - fitted and then checked. And this process is repeated until a satisfactory model is found.

Similar to grouping the observations by the unique covariate patterns for the purpose of estimating the parameters, again here for the purpose of checking the goodness-of-fit of a model, the n independent responses are grouped into m unique covariate patterns (populations) each with $n_i; i = 1, 2, \cdots, m$ observations where $\sum_{i=1}^{m} n_i = n$. Of the n_i observations in each covariate pattern, if n_{1i} successes are observed, then $n_{0i} = n_i - n_{1i}$ of them are failures. Thus, the raw residual is the difference between the observed number of successes n_{1i} and expected number of successes $\hat{\mu}(\boldsymbol{x}_i) = n_i\hat{\pi}(\boldsymbol{x}_i)$ for each value of the covariate $\boldsymbol{x}_i$.

11.3.1 The Pearson Chi-squared Goodness-of-fit Statistic

The Pearson residual is the standardized difference between the observed and expected number of successes. That is,

$$r_i = \frac{n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)}{\sqrt{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]}}; \quad i = 1, 2, \cdots, m.$$

Thus, the Pearson chi-squared statistic is the sum of the square of standardized residuals:

$$X^2 = \sum_{i=1}^{m} \frac{[n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)]^2}{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]} \sim \chi^2(m-k).$$

When this statistic is close to zero, it indicates a good model fit to the data. When it is large, it is an indication of lack of fit. Often the Pearson residuals r_i are used to determine exactly where the lack of fit occurs.

Example 11.11. Recall again example 11.1. Test the adequacy of the model using the Pearson chi-squared test.

Solution: The fitted probabilities are obtained from the fitted model. Note here the number of populations (aggregate values of the explanatory variable) is $m = 6$. Thus,

$$r_i = \frac{n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)}{\sqrt{n_i\hat{\pi}(\boldsymbol{x}_i)[1 - \hat{\pi}(\boldsymbol{x}_i)]}}; \quad i = 1, 2, \cdots, 6$$

Group (x_i)	Frequency (n_i)	Successes (n_{1i})	Probability [$\hat{\pi}(x_i)$]	r_i	r_i^2
18	2	0	0.1432	-0.5782	0.3343
20	3	1	0.1676	0.7685	0.5906
30	1	0	0.3381	-0.7147	0.5108
45	1	1	0.6736	0.6961	0.4846
55	3	2	0.8397	-0.8169	0.6673
60	3	3	0.8929	0.5999	0.3599
Total	13	7			2.9475

The Pearson chi-squared test statistic becomes $X^2 = \sum_{i=1}^{6} r_i^2 = 2.9475$ which is smaller than $\chi^2_{0.05}(6-2) = \chi^2_{0.05}(4) = 9.4877$, indicating that the model is a good fit to the data.

11.3.2 The Deviance Statistic

The deviance, like the Pearson chi-squared, is used to test the adequacy of the logistic model. As shown before, the maximum likelihood estimates of the parameters of the logistic regression are estimated iteratively by maximizing the Binomial likelihood function. Maximizing the likelihood function is equivalent to minimizing the deviance function. The choices for $\hat{\beta}_j;\ j = 0, 1, \cdots, k$ that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a 'likelihood sense'. The deviance is given by:

$$D = 2\sum_{i=1}^{m}\left\{n_{1i}\log\left[\frac{n_{1i}}{n_i\hat{\pi}(\boldsymbol{x}_i)}\right] + (n_i - n_{1i})\log\left[\frac{n_i - n_{1i}}{n_i[1 - \hat{\pi}(\boldsymbol{x}_i)]}\right]\right\} \sim \chi^2(m-k)$$

where the fitted probabilities $\hat{\pi}(\boldsymbol{x}_i)$ satisfy logit $\hat{\pi}(\boldsymbol{x}_i) = \sum_{j=0}^{k}\hat{\beta}_j x_{ij}$ and $x_{i0} = 1$. The deviance is small when the model fits the data, that is, when the observed and fitted proportions are close together. Large values of D (small p-values) indicate that the observed and fitted proportions are far apart, which suggests that the model is not good.

11.3.3 The Hosmer-Lemeshow Test Statistic

The Pearson chi-squared goodness-of-fit test cannot be readily applied if there are only one or a few observations for each possible value (combination of values) of the explanatory variable(s). Consequently, the Hosmer-Lemeshow statistic, the best goodness-of-fit test with continuous

explanatory variables, was developed to address this problem. The idea is to aggregate similar observations into (mostly 10 - decile) groups that have large enough samples so that a Pearson statistic is computed on the observed and predicted counts from the groups. That is,

$$HL = \sum_{i=1}^{m} \frac{[n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)]^2}{n_i\hat{\pi}(\boldsymbol{x}_i)[1 - \hat{\pi}(\boldsymbol{x}_i)]} \sim \chi^2(m-2).$$

11.4 Multinomial Logistic Regression

Multinomial logistic regression is used to predict a nominal dependent variable given one or more independent variables. It is an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

Let Y be a categorical response with J categories. Let $P(Y = j|\boldsymbol{x}_i) = \pi_j(\boldsymbol{x}_i)$ at a fixed setting $\boldsymbol{x}_i$ for explanatory variables with $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}_i) = 1$. Thus, Y has a multinomial distribution with probabilities $\{\pi_1(\boldsymbol{x}_i), \pi_2(\boldsymbol{x}_i), \cdots, \pi_J(\boldsymbol{x}_i)\}$.

Multinomial (also called *polytomous*) logit models simultaneously describe log odds for all $\binom{J}{2}$ pairs of categories. Of these, a certain choice of $J-1$ are enough to determine all, the rest are redundant. An odds for a multinomial response can be defined to be a comparison of *any pair* of response categories. For example, the odds of category 1 relative to category 3 is simply the ratio $\frac{\pi_1}{\pi_3}$.

Logit models for multinomial responses are developed by selecting one response category, often the first (last) category or the most common one, as a baseline (reference) and forming the odds of the remaining $J-1$ categories against this category. For example, the *multinomial logit* model (also called *baseline category logit* model) pairing each response category with the last category,

$$\log\left[\frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right] = \beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik};\ j = 1, 2, \cdots, J-1$$

simultaneously describes the effects of the explanatory variables on the $J-1$ logit models (if $J = 2$, it simplifies to binary logistic regression model). The intercepts and effects vary according to the response paired with the baseline. That is, each model has its own intercept and slope. Also note that for the reference category, $\beta_{J0} = \beta_{J1} = \beta_{J2} = \cdots = \beta_{Jk} = 0$.

The $J-1$ equations also determine parameters for logit models with other pairs of response categories, since

$$\log\left[\frac{\pi_1(\boldsymbol{x}_i)}{\pi_2(\boldsymbol{x}_i)}\right] = \log\left[\frac{\pi_1(\boldsymbol{x}_i)/\pi_J(\boldsymbol{x}_i)}{\pi_2(\boldsymbol{x}_i)/\pi_J(\boldsymbol{x}_i)}\right] = \log\left[\frac{\pi_1(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right] - \log\left[\frac{\pi_2(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right].$$

Example 11.12. Based on the survival outcome of HAART treatment, HIV/AIDS patients were classified into four categories (0= Active, 1= Dead, 2= Transferred to other hospital, 3= Lost-to-follow). To identify factors associated with these survival outcomes, a multinomial logit model was fitted. Three explanatory variables that were considered are Age, Gender

(0= Female, 1= Male) and Functional Status (0= Working, 1= Ambulatory, 2= Bedridden). The parameter estimates are presented as follows (values in brackets are standard errors).

				Functional Status	
logit	Intercept	Age	Gender	Ambulatory	Bedridden
$\log(\hat{\pi}_D/\hat{\pi}_A)$	-3.271 (0.624)	-0.020 (0.018)	0.564 (0.325)	0.940 (0.333)	2.280 (0.479)
$\log(\hat{\pi}_T/\hat{\pi}_A)$	-1.882 (0.413)	-0.030 (0.012)	0.635 (0.211)	0.833 (0.209)	1.584 (0.393)
$\log(\hat{\pi}_L/\hat{\pi}_A)$	-1.116 (0.343)	-0.031 (0.010)	0.455 (0.178)	0.292 (0.183)	0.828 (0.395)

Write the estimated multinomial logit models and interpret. Also, find the estimated logit model for the log odds of dead instead of transferred to other hospital.

Solution: Let Y = survival outcome, X_1 = age of the patient, X_2 = gender and X_3= functional status.

Each model is written as:

$$\log\left[\frac{\hat{\pi}_j(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_{i1} + \hat{\beta}_{j2}x_{i2} + \hat{\beta}_{j31}d_{i31} + \hat{\beta}_{j32}d_{i32};\ j = D, T, L.$$

For example, the estimated model for the log odds of being dead instead of active is

$$\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -3.271 - 0.020x_{i1} + 0.564x_{i2} + 0.940d_{i31} + 2.280d_{i32}.$$

An increase in the age of a patient by one year decreases the odds of being dead (instead of active) by 2% (a factor of $\exp(-0.020) = 0.98$). The odds that male patients being dead (instead of active) is $\exp(0.565) = 1.759$ times that of females, or the odds of being dead (instead of active) is 75.9% higher for males than for females. In other words, relative to female patients, male patients are 1.759 times (75.9%) more likely to be dead (instead of active). Also, ambulatory patients are $\exp(0.941) = 2.563$ times more likely to be dead (instead of active) as compared to working patients. Similarly, bedridden patients are $\exp(2.280) = 9.777$ times more likely to be dead (instead of active) relative to working patients. The functional status effects indicate that the odds of being dead (instead of active) are relatively higher for bedridden patients relative to ambulatory patients.

The estimated model for the log odds of being transferred instead of active is

$$\log\left[\frac{\hat{\pi}_T(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -1.882 - 0.030x_{i1} + 0.635x_{i2} + 0.833d_{i31} + 1.584d_{i32}.$$

An increase in the age of a patient by a year decreases the odds of being transferred to other hospital (instead of active) by 3% (a factor of $\exp(-0.030) = 0.970$). The odds that male patients being transferred to other hospital (instead of active) is $\exp(0.635) = 1.887$ times that of females, or the odds of being transferred to other hospital (instead of active) is 88.7% higher for males than for females. In other words, male patients are 1.887 times (88.7%) more likely to be transferred to other hospital (instead of active) as compared to female patients. Also, relative to working patients, ambulatory patients are $\exp(0.833) = 2.300$ times more likely to be transferred to other hospital (instead of active). Similarly, bedridden patients are

$\exp(1.584) = 4.874$ times more likely to be transferred to other hospital (instead of active) as compared to working patients.

Also, the estimated model for the log odds of being lost-to-follow instead of active is

$$\log\left[\frac{\hat{\pi}_L(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -1.116 - 0.031x_{i1} + 0.455x_{i2} + 0.292d_{i31} + 0.828d_{i32}.$$

The odds of being lost-to-follow (instead of active) decreases by 3.1% (a factor of $\exp(-0.031) = 0.969$) every year older an individual is. Male patients are $\exp(0.455) = 1.576$ times (57.6%) more likely to be lost-to-follow (instead of active) relative to female patients. As compared to working patients, ambulatory patients are $\exp(0.292) = 1.339$ times (33.9%) more likely to be lost-to-follow (instead of active). Similarly, bedridden patients are $\exp(0.828) = 2.289$ times more likely to be lost-to-follow (instead of active) as compared to working patients.

The estimated model for being dead instead of transferred to other hospital is

$$\begin{aligned}
\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_T(\boldsymbol{x}_i)}\right] &= \log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] - \log\left[\frac{\hat{\pi}_T(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] \\
&= -3.271 - 0.020x_{i1} + 0.564x_{i2} + 0.940d_{i31} + 2.280d_{i32} \\
&\quad - (-1.882 - 0.030x_{i1} + 0.635x_{i2} + 0.833d_{i31} + 1.584d_{i32}) \\
&= -1.389 + 0.010x_{i1} - 0.071x_{i2} + 0.107d_{i31} + 0.696d_{i32}.
\end{aligned}$$

Therefore, the estimated model for the log odds of dead instead of transferred to other hospital is

$$\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_T(\boldsymbol{x}_i)}\right] = -1.389 + 0.010x_{i1} - 0.071x_{i2} + 0.107d_{i31} + 0.696d_{i32}.$$

11.5 Ordinal Logistic Regression

Many categorical response variables have a natural ordering to their categories or called *levels*. For example, a response variable (like amount of agreement) may be measured using a Likert scale with categories 'strongly disagree', 'disagree', 'neutral', 'agree' or 'strongly agree'. Ordinal logistic regression is used to predict such an ordinal dependent variable given one or more independent variables.

Let Y is an ordinal response with J categories. Then there are $J-1$ ways to dichotomize these outcomes. These are $Y_i \leq 1$ ($Y_i = 1$) versus $Y_i > 1$, $Y_i \leq 2$ versus $Y_i > 2, \cdots, Y_i \leq J-1$ versus $Y_i > J-1$ ($Y_i = J$). With this categorization of Y_i, $P(Y_i \leq j)$ is the cumulative probability that Y_i falls at or below category j. That is, for outcome j, the cumulative probability is

$$P(Y_i \leq j|\boldsymbol{x}_i) = \pi_1(\boldsymbol{x}_i) + \pi_2(\boldsymbol{x}_i) + \cdots + \pi_j(\boldsymbol{x}_i);\ j = 1, 2, \cdots, J$$

where $P(Y_i \leq j|\boldsymbol{x}_i) = 1$. Each cumulative logit model uses all the J response levels. A model for logit $[P(Y \leq j|\boldsymbol{x}_i)]$ alone is the usual logit model for a binary response in which categories

from 1 to j form one outcome and categories from $j+1$ to J form the second. That is,

$$\begin{aligned}\text{logit } P(Y_i \leq j) &= \log\left[\frac{P(Y_i \leq j)}{1-P(Y_i \leq j)}\right] \\ &= \log\left[\frac{P(Y_i \leq j)}{P(Y_i > j)}\right] \\ &= \log\left[\frac{\pi_1(\boldsymbol{x}_i)+\pi_2(\boldsymbol{x}_i)+\cdots+\pi_j(\boldsymbol{x}_i)}{\pi_{j+1}(\boldsymbol{x}_i)+\pi_{j+2}(\boldsymbol{x}_i)+\cdots+\pi_J(\boldsymbol{x}_i)}\right];\ j=1,2,\cdots,J-1.\end{aligned}$$

Formally, a model that simultaneously uses all cumulative logits assuming linear relationship with the explanatory variables is

$$\text{logit } P(Y_i \leq j|\boldsymbol{x}_i) = \beta_{j0}+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik};\ j=1,2,\cdots,J-1.$$

Each cumulative logit has its own intercept which usually are not of interest except for computing response probabilities. Since logit $[P(Y_i \leq j|\boldsymbol{x}_i)]$ increases in j for a fixed $\boldsymbol{x}_i$ and the logit is an increasing function of this probability, each intercept increases in j.

But, the model assumes the *same slope* (its associated odds ratio called *cumulative odds ratio*) regardless of the category j. This is called *proportional odds* assumption which means the distance between each category is equivalent (proportional odds). That is, each model has the same effect associated with each explanatory variable (the effects of the explanatory variables are the same regardless of which cumulative probabilities are used).

The slope parameters can be interpreted in the same way as a binary logistic regression parameters - except in this case, there are three transitions estimated instead of one transition - as there would be with a dichotomous dependent variable. A positive parameter indicates an increased chance that a subject with a higher score on the independent variable will be observed in a higher category. A negative parameter indicates that the chances that a subject with a higher score on the independent variable will be observed in a lower category.

The intercepts can be used to calculate predicted probabilities for a person with a given set of characteristics of being in a particular category.

Example 11.13. To determine the effect of Age and Gender (0= Female, 1=Male) on the Clinical Stage of HIV/AIDS patients (1= Stage I, 2= Stage II, 3= Stage III and 4= Stage IV), the following parameter estimates of ordinal logistic regression are obtained. The loglikelihood values of the null and the full models are -1854.3173 and -1852.1351, respectively.

Variable	Parameter Estimate	Standard Error
Intercept 1	-0.9905	0.1884
Intercept 2	0.5383	0.1870
Intercept 3	2.7246	0.2066
Age	0.0034	0.0055
Gender	0.1789	0.1028

Obtain the cumulative logit model and interpret.

Solution: Let Y= Clinical Stage of patients (1= Stage I, 2= Stage II, 3= Stage III and 4= Stage IV), X_1= Age and X_2= Gender (0= Female, 1=Male).

Hence, the model has the form logit $\widehat{P}(Y_i \leq j|\boldsymbol{x}_i) = \hat{\beta}_{j0} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2};\ j = 1, 2, 3$. With $J = 4$ categories, the model has three cumulative logits. These are:

$$\begin{aligned}
\text{logit } \widehat{P}(Y_i \leq 1|\boldsymbol{x}_i) &= -0.9905 + 0.0034x_{i1} + 0.1789x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) &= \ \ 0.5383 + 0.0034x_{i1} + 0.1789x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 3|\boldsymbol{x}_i) &= \ \ 2.7246 + 0.0034x_{i1} + 0.1789x_{i2}.
\end{aligned}$$

The cumulative estimate $\hat{\beta}_1 = 0.0034$ suggests an increase in the age of the patient leads to be in higher clinical stages given the gender. Being in smaller ages reduces the likelihood of being in a higher clinical stage category. Also, the estimate $\hat{\beta}_2 = 0.1789$ males are more likely to be in higher clinical stages as compared to females given the age of the patient. That is, being male increases the likelihood of being in a higher clinical stage category.

Chapter 12

Count Regression Model

Count regressions such as poisson and negative-binomial models are used for modelling *count (discrete)* response variables: for example, the number of hospital admissions or the number of accidents over some period of time. The unit of analysis may be a person (e.g., number of infections per patient per year), an institution (e.g., number of admissions per hospital per month) or a place (e.g., number of car accidents per city per day). As a first pass, such a dependent variable could be analyzed as a continuous outcome. However, unlike a continuous variable, with counts there cannot be negative numbers. Also, the distribution of counts often tend to be skewed to the right and does not fit a normal distribution.

Count regression models are also used to model *incidence rate* or incidence of rare diseases. Incidence rate measures the rate at which a group of people develops a disease or condition. Often it is of interest to compare incidence rates. For example, is the incidence of diabetes higher in one city than another or is higher among men than women. As is true of counts, incidence rates cannot be negative. As a result, in situations such as these, analyzing the data with a technique such as linear regression is not appropriate.

12.1 The Exponential Function

Count regression models are modeled based on the exponential function. For any real number z, the exponential function is $f(z) = \exp(z)$. This function is nonnegative for all values of z. That is, if $z = -\infty$, then $f(-\infty) = 0$, if $z = 0$, then $f(0) = 1$ and if $z = \infty$, then $f(\infty) = \infty$.

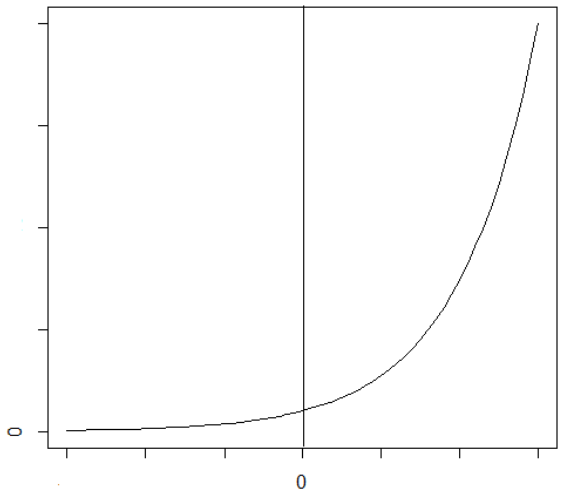


Figure 12.1: Plot of the Exponential Function

The figure also shows that the range of f is in between 0 and ∞ for every real number z. Therefore, $0 \leq f(z) < \infty$.

12.2 The Poisson Regression Model

To obtain the poisson regression model from the exponential function, z should be expressed as a function (mostly linear function) of the explanatory variable(s). That is, $z_i = g(x_i) = \alpha + \beta x_i$ for a single explanatory variable X. As a result, the simple poisson regression model can be written as $f(x_i) = \exp(\alpha + \beta x_i)$. Here, since $f(x_i)$ represents the mean response, let us use the notation $\mu(x_i)$. That is, $\mu(x_i) = \exp(\alpha + \beta x_i)$. This model can be linearized using the natural logarithm transformation as:

$$\log \mu(x_i) = \alpha + \beta x_i.$$

Here α and β are the intercept and slope parameters of the log-linear model. The slope parameter is commonly interpreted in terms of an incidence rate ratio (IRR). A one unit increase in x_i has a multiplicative impact of $\exp(\beta)$ on the mean response, that is, the mean of Y_i at $x_i + 1$ is the mean of Y_i at x_i multiplied by $\exp(\beta)$. If $\beta = 0$, then the multiplicative factor is 1, the mean of Y_i does not change as x_i changes. If $\beta > 0$, then $\exp(\beta) > 1$ and the mean of Y_i increases as x_i increases. If $\beta < 0$, the mean decreases as x_i increases.

Similarly, if there are k explanatory variables, the multiple poisson regression model is written as:

$$\log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=0}^{k} \beta_j x_{ij} \tag{12.1}$$

where $x_{i0} = 1$ for all $i = 1, 2, \cdots, n$. Here, $\mu(\boldsymbol{x}_i)$ is the conditional mean of Y_i given $\boldsymbol{x}_i$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$.

The sample poisson regression model is:

$$\log \hat{\mu}(\boldsymbol{x}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \sum_{j=0}^{k} \hat{\beta}_j x_{ij} \tag{12.2}$$

- $\hat{\mu}(\boldsymbol{x}_i)$ is the estimated mean response.
- $\hat{\beta}_0$ is the estimated intercept of the log-linear model.
- $\hat{\beta}_j; j = 1, 2, \cdots, k$ is the j^{th} estimated (partial) slope associated with the j^{th} independent variable.

Example 12.1. Suppose a study is conducted in identifying factors associated with CD4 counts of 1464 HIV/AIDS patients at the start of HAART treatment. Here the response variable is CD4 count of a patient and the explanatory variables were Age in years (Age), Gender (0=Female, 1=Male) and Functional Status (0=Working, 1=Ambulatory, 2=Bedridden). The parameter estimates and their corresponding standard errors of the poisson regression model are given in the following table.

Variable	Parameter Estimate	Standard Error
Intercept	5.4625	0.0079
Age	0.0060	0.0002
Gender	-0.1982	0.0041
Ambulatory	-0.3783	0.0046
Bedridden	-0.6296	0.0123

Obtain the estimated model and interpret the estimates.

Solution: Let Y= CD4 count, X_1= Age, X_2= Gender (0=Female, 1=Male) and X_3= Functional Status (0=Working, 1=Ambulatory, 2=Bedridden). The estimated model is:

$$\log \hat{\mu}(\boldsymbol{x}_i) = 5.4625 + 0.0060 x_{i1} - 0.1982 x_{i2} - 0.3783 d_{i31} - 0.6296 d_{i32}.$$

As the age of the patient increases by one year, the mean CD4 count increases by 0.60% [$\exp(0.0060) - 1 = 0.60\%$]. The mean CD4 count of male patients decreases by 17.98% [$1 - \exp(-0.1982) = 17.98\%$] than female patients. Similarly the mean CD4 counts of ambulatory and bedridden patients decreases by 31.50% and 46.72% than working patients, respectively.

12.2.1 Estimation

Inference on the model and its parameters follows exactly the same approach as used for logistic regression. Like other regression modeling, the goal of poisson regression is to estimate the $k + 1$ unknown parameters of the model. The method of maximum likelihood is used to estimate the parameters which follows closely the approach used for logistic regression.

Consider a random variable Y that can take on a set of count values. Given a dataset with a sample size of n where each observation is independent. Thus, $\boldsymbol{Y}$ can be considered as a vector of n poisson random variables. That is, each individual count response $Y_i;\ i = 1, 2, \cdots, n$ has an independent poisson distribution with parameter $\mu(\boldsymbol{x}_i)$, that is,

$$P(Y_i = y_i) = \frac{\mu(\boldsymbol{x}_i)^{y_i} \exp[-\mu(\boldsymbol{x}_i)]}{y_i!};\ y_i = 0, 1, 2, \cdots.$$

Then, the joint probability mass function of $\boldsymbol{Y}^t = (Y_1, Y_2, \cdots, Y_n)$ is the product of the n poisson distributions. Thus, the likelihood function is:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\mu(\boldsymbol{x}_i)^{y_i} \exp[-\mu(\boldsymbol{x}_i)]}{y_i!} \tag{12.3}$$

where $\mu(\boldsymbol{x}_i) = \exp(\sum_{j=0}^{k} \beta_j x_{ij})$. Also, the log-likelihood function becomes:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log\,[\mu(\boldsymbol{x}_i)] - \sum_{i=1}^{n} \mu(\boldsymbol{x}_i) - \sum_{i=1}^{n} \log\,(y_i!). \tag{12.4}$$

Then, partially differentiating the log-likelihood with respect to $\beta_j;\ j = 0, 1, 2, \cdots, k$ and setting it equal to zero results $k+1$ equations with $k+1$ unknown parameters. That is,

$$\frac{\partial L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} [y_i - \mu(\boldsymbol{x}_i)] x_{ij} = 0;\ j = 0, 1, 2, \cdots, k. \tag{12.5}$$

which is usually solved with some numerical method like the Newton-Raphson algorithm.

Also, the second partial derivative of the log-likelihood function yields the variance-covarince matrix of the estimated parameters:

$$\frac{\partial^2 L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j \beta_h} = -\sum_{i=1}^{n} \mu(\boldsymbol{x}_i) x_{ij} x_{ih}; \quad j = h = 0, 1, 2, \cdots, k. \tag{12.6}$$

12.2.2 Significance Tests

Let ℓ_M denote the maximized value of the likelihood function for the fitted model M with all the k explanatory variables. Let ℓ_0 denote the maximized value of the likelihood function for the fitted model with no explanatory variables (having only one parameter, that is, the intercept). The likelihood-ratio test statistic is $G^2 = -2(\log \ell_0 - \log \ell_M) = D_0 - D_M \sim \chi^2(k)$. Rejection of the null hypothesis implies at least one of the parameter is significantly different from zero. Then, Wald test can be used to look at the significance of each variable ($H_0 : \beta_j = 0$) using a Z statistic in which

$$Z_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \sim N(0, 1)$$

for large sample size.

Example 12.2. The log-likelihood value of the model given in example 12.1 is -85956.40 and the corresponding null model is -92061.31. Test the overall significance of the model and also identify the significant variables using wald test.

Solution: The model is of the form $\log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{31} d_{i31} + \beta_{32} d_{i32}$. For testing the significance of the model, the hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_{31} = \beta_{32} = 0$. Thus, the likelihood-ratio statistic is $G^2 = -2(\log \ell_0 - \log \ell_M) = -2[-92061.31 -$

$(-85956.40)] = 12209.82$ which is very larger than $\chi^2_{0.05}(4) = 1.145$. Therefore, at least one of the explanatory variable is significant.

To identify the significant explanatory variables one by one, the Wald statistics are calculated as shown in the following table.

Variable	t Statistic	95% CI for β	$\widehat{\text{IRR}}$	95% CI for IRR
Intercept	691.46*	(5.4470, 5.4780)*		
Age	30.00*	(0.0056, 0.0064)*	1.0060	(1.0056, 1.0064)*
Gender	-48.34*	(-0.2062, -0.1902)*	0.8202	(0.8137, 0.8268)*
Ambulatory	-82.33*	(-0.3877, -0.3697)*	0.6848	(0.6786, 0.6909)*
Bedridden	-30.76*	(-0.4024, -0.3542)*	0.6850	(0.6687, 0.7017)*

As can be seen, all the three explanatory variables are significantly associated with the CD4 counts of HIV/AIDS patients.

12.2.3 Model Diagnostics

Just as in any model fitting procedure, analysis of residuals is important in fitting poisson regression. Residuals can provide guidance concerning the overall adequacy of the model, assist in verifying assumptions, and can give an indication concerning the appropriateness of the selected link function.

The ordinary or raw residuals are just the differences between the observations and the fitted values, $e_i = y_i - \mu(\boldsymbol{x}_i)$, which have limited usefulness. The Pearson residuals are the standardized differences

$$r_i = \frac{y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}}.$$

These residuals fluctuate around zero, following approximately a normal distribution when $\mu(\boldsymbol{x}_i)$ is large. When the model holds, these residuals are less variable than standard normal, however, because the numerator must use the fitted value $\hat{\mu}(\boldsymbol{x}_i)$ rather than the true mean $\mu(\boldsymbol{x}_i)$. Since the sample data determine the fitted value, $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ tends to be smaller than $[y_i - \mu(\boldsymbol{x}_i)]$.

Since, the standardized residual takes $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ and divides it by its estimated standard error $\sqrt{\hat{\mu}(\boldsymbol{x}_i)}$, it does have an approximate standard normal distribution when $\mu(\boldsymbol{x}_i)$ is large. With standardized residuals, it is easier to tell when a deviation $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ is "large".

Components of the deviance are alternative measures of lack of fit. The deviance residuals are $d_i = \pm\sqrt{y_i \log\left[y_i/\hat{\mu}(\boldsymbol{x}_i)\right] - [y_i - \hat{\mu}(\boldsymbol{x}_i)]}; i = 1, 2, \cdots, n$ where the sign is the sign of the ordinary residual. The deviance residuals approach zero when the observed values of the response and the fitted values are closer to each other.

12.3 The Negative-Binomial Regression Model

For a poisson distribution, the variance and the mean are equal. Often count data vary more than the expected. The phenomenon of the data having greater variability than expected

is called *over-dispersion*. But, over-dispersion is not an issue in ordinary regression models assuming normally distributed response, because the normal distribution has a separate parameter to describe the variability.

In the presence of over-dispersion, a negative binomial model is should be applied. Like a poisson model, a negative binomial model expresses the log mean response in terms of the explanatory variables. But a negative binomial model has an additional parameter called a *dispersion parameter*. That is, because, the negative binomial distribution has mean $E(Y) = \mu$ and variance $\text{Var}(Y) = \mu + \psi\mu^2$ where $\psi > 0$. The index ψ is a dispersion parameter. As ψ approaches 0, $Var(Y)$ goes to μ and the negative binomial distribution converges to the poisson distribution. The farther ψ falls above 0, the greater the over-dispersion relative to poisson variability.

Example 12.3. Consider example 12.1. The parameter estimates and their corresponding standard errors of the negative binomial regression are given below.

Variable	Parameter Estimate	Standard Error
Intercept	5.4202	0.0867
Age	0.0067	0.0023
Gender	-0.1841	0.0443
Ambulatory	-0.3743	0.0460
Bedridden	-0.6332	0.1066
$\hat{\psi}$	0.6022 [CI: (0.5628,0.6443)]	0.0208

The log-likelihood value of this model is -9083.73 and that of the null model is -9135.30. Compare and contrast the parameter estimates with that of the poisson regression. In addition, compare both models by finding their corresponding AIC values.

Solution: As the dispersion parameter ψ is significantly larger than 0, it assures that the negative binomial regression model is more appropriate than the poisson regression model.

Chapter 13

Survival Analysis

Survival analysis deals with the analysis of the time from a certain origin to the occurrence of an event such as death, the appearance of a tumor, the occurrence of a certain disease, equipment breakdown, cessation of breast feeding, and so forth. The event of interest to be considered is not necessarily a bad event like death. It can also denote a good event such as time-to-cure from a certain disease after some treatment, cessation of smoking, and so forth.

The problem of analyzing time to event data arises in a number of applied fields, such as medicine/public health (e.g., time-to-relapse of a certain disease, time-to-death of HIV patients after HAART treatment, time-to-reoccurrence of a particular symptom, time-to-cure from a certain disease), agriculture (e.g., length of time required for a cow to conceive after calving, time until a farm experiences its first case of an exotic disease), sociology (called duration analysis) (e.g., time-to-find a job after graduation, time until re-arrest after release from prison), engineering (called reliability analysis) (e.g., time-to-the failure of a machine) and other.

Although the statistical tools for time-to-event data are applicable to all these disciplines, the focus is on applying the techniques to public health and medicine.

13.1 The Survival and Hazard Functions

Let T be a nonnegative random variable denoting the time until some specified event and let t be a specific point in time. Survival and hazard functions are the two functions of central interest in summarizing survival data. The *survival function* is the probability of an individual surviving to time t. That is, it reports the probability of surviving beyond time t: $S(t) = P(T > t)$. For example, if the event of interest is death and $t = 50$ years, then $S(t = 50) = \frac{\text{number of persons survived up to 50 years}}{\text{total number of persons in the study}}$ is probability of surviving beyond 50 years. Here, if $S(t = 50) = 0.74$, then the 50-years survival probability is 74% or 26% of the individuals will die below 50-years.

Said differently, the survival function is the probability that there is no event prior to t. The function is equal to 1 at $t = 0$ ($S(0) = 1$) and decreases toward 0 as t goes to ∞ ($S(\infty) = 0$). The survival function is a monotone, decreasing function of time.

The *hazard rate (or function)* is the instantaneous risk of experiencing the event of interest in $(t, t+\Delta t)$ given that the individual is still alive at t. It is the (limiting) probability that an event occurs in a given interval, conditional upon the subject having survived to the beginning of that interval, divided by the width of the interval:

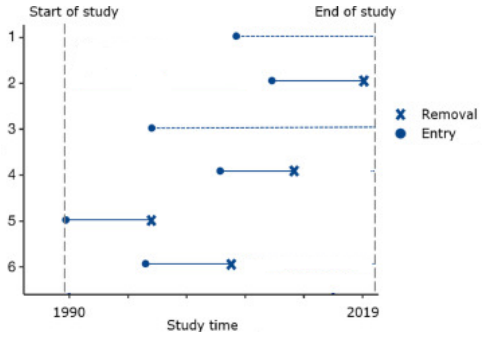
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t}.$$

For example, if the event of interest is death and $t = 50$ years, then $h(t = 50)$ is the hazard rate at 50 years. Here, if $h(t = 50) = 6.1$, then, at 50-years, individuals die at rate of 6.1 per year.
The hazard rate (or function) can vary from 0 (meaning no risk at all) to ∞ (meaning the certainty of event at that instant) and has unit $1/t$. Over time, the hazard rate can increase, decrease, remain constant, or even take on more serpentine shapes. The human mortality pattern related to aging generates a falling hazard for a while after birth, and then a long, flat plateau, and thereafter constantly rising and eventually reaching, one supposes, values near infinity at about 100 years. The risk of post-operative wound infection falls as time from surgery increases, so the hazard function decreases with time.

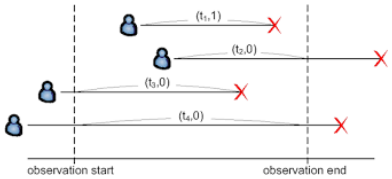
Calling the probability of a bad event of interest, like death, "hazard" might not be strange. But, it feels strange to think of the hazard of a positive outcome, like recurring from a disease. But technically, it is the same thing.

13.2 Survival Data Format

In survival analysis, it is realistic to record the start and end of study in calendar dates. Individuals enter in the study at different time points. Some individuals might not experience the event when the study ends. And other individuals might drop out or get lost in the middle of the study. For such individuals, only a lower bound of the true survival time is known (called right *censoring*).



Hence, the time-to-event variable actually records either the actual survival time for those subjects who experience the event of interest or the minimum survival time for subjects who do not experience the event of interest.



Consider the following survival data. The Entry Date is the date when each subject was at risk (the date when each subject was exposed) and the Exit Date is the date when each subject experienced the event of interest or censored. Hence, each record documents a span of time, from Entry to Exit Dates, for a subject. In the data below, the entry and exit times are in calendar dates, and the survival analysis time for each subject is, then, the difference in the number of days between each subject's Entry and Exit Dates.

ID	Entry Date	Exit Date	Time (t)	Event	Sex (1=Male, 2=Female)	Age
1	20 Jan 2000	21 Jan 2000	1	1	1	30
2	15 Dec 1999	20 Dec 1999	5	1	2	28
3	04 Jan 2000	13 Jan 2000	9	1	2	34
4	31 Jan 2000	19 Feb 2000	20	1	2	28
5	10 Feb 2000	04 Mar 2000	22	0	1	20

The value of the instantaneous variable (Event) is the value it had at the end of the span, that is, Exit Date. It is coded as 1 to indicate the occurrence of the event of interest and else coded as 0 (for censoring).

The values of the enduring variables (here Sex, Age) are the values they had on the Entry Date.

13.3 Non-Parametric Analysis

The analysis of survival data can take one of three forms - nonparametric, semiparametric, and parametric - all depending on the form of the survival function and about how the survival experience is affected by covariates. Nonparametric analysis follows the philosophy of making no assumption about the functional form of the survival function. The effects of covariates are not modeled, either - the comparison of the survival experience is done at a qualitative level across the values of the covariates.

The most basic of analyses would be on analysing the mean time-to-event or the median time-to-event. However, the typical preliminary data analysis tools do not translate well into the survival analysis paradigm. Estimates of the median survival time are similarly not possible to obtain using standard nonsurvival tools. The standard way of calculating the median is to order the observations and to report the middle one as the median. In the presence of censoring, that ordering is impossible to ascertain.

13.3.1 The Kaplan and Meier Estimator

The estimator of Kaplan and Meier (1958) is a nonparametric estimate of the survival function $S(t)$, which is the probability of survival past time t or, equivalently, the probability of experiencing the event after t. For a dataset with observed event times, t_1, t_2, $\cdots$, t_k, where k is the number of distinct event times observed in the data, the Kaplan-Meier estimate [also known as the *product limit* estimate of $S(t)$] at any time t is given by:

$$\widehat{S}(t) = \prod_{j|t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right)$$

where n_j is the number of individuals at risk at time t_j and d_j is the number of events at time t_j. The product is over all observed event times less than or equal to t.

Consider the following dataset of 6 subjects given in the usual format:

ID	t	Event
1	2	1
2	4	1
3	4	1
4	5	0
5	7	1
6	8	0

At $t = 2$, the earliest time in the data, all six subjects were at risk, but at that instant, only one event occurred (ID=1). At the next time, $t = 4$, five subjects were at risk, but at that instant, two events occurred. At $t = 5$, three subjects were left, and none occurred, but one subject was censored. This left us with two subjects at $t = 7$, of which one occurred. Finally, at $t = 8$, we had one subject left at risk, and this subject was censored at that time.

Now let us form a table that summarizes what happens at each time in the data (whether an event time or a censored time):

t	No. at Risk (n_j)	No. of Events (d_j)	No. of Censored ($n_j - d_j$)
2	6	1	0
4	5	2	0
5	3	0	1
7	2	1	0
8	1	0	1

Now we ask the following:

- What is the probability of survival beyond $t = 2$, the earliest time in the data? Because five of the six subjects survived beyond this point, the estimate is $\frac{5}{6} = 0.8333$.
- What is the probability of survival beyond $t = 4$ given survival right up to $t = 4$? Because we had five subjects at risk at $t = 4$, and two occurred, we estimate this probability to be $\frac{3}{5} = 0.60$.
- What is the probability of survival beyond $t = 5$ given survival right up to $t = 5$? Because three subjects were at risk, and none occurred, the probability estimate is $\frac{3}{3} = 1$.
- What is the probability of survival beyond $t = 7$ given survival right up to $t = 7$? Because two subjects were at risk, and one occurred, the probability is estimated to be $\frac{1}{2} = 0.50$.
- What is the probability of survival beyond $t = 8$ given survival right up to $t = 8$? Because we had only one subject at risk at $t = 4$, and it was censored, the probability is $\frac{1}{1} = 1$.

We can now augment a table with these component probabilities:

t	No. at Risk	No. of Events	No. of Censored	Probability
2	6	1	0	$\frac{5}{6}$
4	5	2	0	$\frac{3}{5}$
5	3	0	1	1
7	2	1	0	$\frac{1}{2}$
8	1	0	1	1

- The first probability value, $\frac{5}{6}$, is the probability of survival beyond $t = 2$.
- The second value, $\frac{3}{5}$, is the (conditional) probability of survival beyond $t = 4$ given survival up until $t = 4$, which in these data is the same as survival beyond $t = 4$ given survival beyond $t = 2$. Thus unconditionally, the probability of survival beyond $t = 4$ is $(\frac{5}{6})(\frac{3}{5}) = 1/2$.
- The third value, 1, is the conditional probability of survival beyond $t = 5$ given survival up until $t = 5$, which in these data is the same as survival beyond $t = 5$ given survival beyond $t = 4$. Unconditionally, the probability of survival beyond $t = 5$ is thus equal to $(\frac{1}{2})(1) = 1/2$.

Thus the Kaplan-Meier estimate is the running product of the probability values that we have previously calculated,

t	No. at Risk	No. of Events	No. of Censored	Probability	$\widehat{S}(t)$
2	6	1	0	$\frac{5}{6}$	$\frac{5}{6}$
4	5	2	0	$\frac{3}{5}$	$\frac{1}{2}$
5	3	0	1	1	$\frac{1}{2}$
7	2	1	0	$\frac{1}{2}$	$\frac{1}{4}$
8	1	0	1	1	$\frac{1}{4}$

Because the Kaplan-Meier estimate operates only on observed event times (and not at censoring times), the net effect is simply to ignore the cases where the probabilities are 1 in calculating the product; ignoring these changes nothing.

13.3.2 The Nelson-Aalen Estimator

The cumulative hazard function ($H(t)$) measures the total amount of risk that has been accumulated up to time t. The theoretical relationship between $H(t)$ and $S(t)$ is $H(t) = -\ln\{S(t)\}$ where for $S(t)$. There is, however, another nonparametric method for estimating $H(t)$ that has better small-sample properties. The estimator is from Nelson (1972) and Aalen (1978),

$$\widehat{H}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}$$

where n_j is the number at risk at time t_j, d_j is the number of events at time t_j, and the sum is over all distinct failure times less than or equal to t.

Thus, given some data

ID	t	Event
1	2	1
2	4	1
3	4	1
4	5	0
5	7	1
6	8	0

and we can construct the risk table

t	n_j	d_j	$n_j - d_j$
2	6	1	0
4	5	2	0
5	3	0	1
7	2	1	0
8	1	0	1

We calculate the number of events per subject at each observed time, $\frac{d_j}{n_j}$, and then sum these to obtain $\widehat{H}(t)$:

t	n_j	d_j	$n_j - d_j$	$\frac{d_j}{n_j}$	$\widehat{H}(t)$
2	6	1	0	0.1667	0.1667
4	5	2	0	0.4000	0.5667
5	3	0	1	0.0000	0.5667
7	2	1	0	0.5000	1.0667
8	1	0	1	0.0000	1.0667

$\widehat{H}(t)$ is the Nelson-Aalen estimator of the cumulative hazard.

13.4 Cox-PH Model

Cox Proportional Hazards (PH) model is a semi-parametric regression that models the natural logarithm of the *relative hazard* of the event of interest. In a general regression model, the hazard function $h(t)$ depends on time t and covariates x_l, x_2, $\cdots$, x_k. In a simpler model, called the Cox PH model do not depend on time, the hazard function has the following form:

$$h(t|\boldsymbol{x}_i) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k).$$

The nice thing about this model is that $h_0(t)$, the baseline hazard, is given no particular parameterizations and, in fact, can be left unestimated. The model makes no assumptions about the shape of the hazard over time - it could be constant, increasing, decreasing, or any other; what is assumed is that, whatever the general shape, it is the same for every subject. One subject's hazard is a multiplicative replica of another's; comparing subject i to subject j, the model states that

$$\frac{h(t|\boldsymbol{x}_i)}{h(t|\boldsymbol{x}_j)} = \exp\{\beta_1(x_{i1} - x_{j1}) + \beta_2(x_{i2} - x_{j2}) + \cdots + \beta_p(x_{ik} - x_{jk})\}$$

which is constant, assuming the covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ do not change over time.

The estimated model is:

$$\hat{h}(t|\boldsymbol{x}_i) = \hat{h}_0(t)\exp(\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k).$$

The model can also be written as:

$$\log \hat{h}(t|\boldsymbol{x}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

where $\hat{h}(t|\boldsymbol{x}_i)$ is the estimated hazard rate at time t, $\hat{\beta}_0$ is the estimated intercept of the log hazard model and $\hat{\beta}_j; j = 1, 2, \cdots, k$ is the j^{th} slope parameter estimate associated with the j^{th} covariate.

The Cox proportional hazards model yields the hazard ratio interpretation of the regression coefficients. The quantity $\widehat{\text{HR}} = \exp(\hat{\beta}_j)$ is the change {equivalently, $100[\exp(\hat{\beta}_j) - 1]\%$ is the percentage change} in the hazard function for each unit increase in the covariate, provided the other covariates stay fixed.

Example 13.1. A study of the factors associated with the survival of HIV/AIDS patients (time-to-death) under HAART treatment was conducted. The study involved 1464 patients and the effect of baseline characteristics like age (years), gender (1=male, 0=female), weight (kg), functional status (1=working, 2=ambulatory, 3=bedridden) and CD4 counts were examined. The parameter and hazard ratio estimated estimates together their 95% confidence interval are presented as follows:

Variable	$\hat{\beta}$	$\widehat{\text{SE}}(\hat{\beta})$	t-Statistic	95% CI for β	$\widehat{\text{HR}}$	95%CI for HR
Age	-0.0079	0.0188	-0.42	(-0.0447, 0.0289)	0.9921	(0.9563, 1.0293)
Male	0.6343	0.3356	1.89	(-0.0235, 1.2921)	1.8857	(0.9768, 3.6404)
Weight	-0.0397	0.0182	-2.18	(-0.0754,-0.0040)*	0.9611	(0.9274, 0.9960)*
Working	-1.2826	0.4821	-2.66	(-2.2275,-0.3377)*	0.2773	(0.1078, 0.7134)*
Ambulatory	-0.9813	0.4728	-2.08	(-1.9080,-0.0546)*	0.3748	(0.1484, 0.9469)*
Bedridden	Ref					
CD4	-0.0071	0.0019	-3.74	(-0.0108,-0.0034)*	0.9929	(0.9893, 0.9966)*

Identify the significant variables and interpret.

Solution: Of the five candidate predictors of the survival of HIV/AIDS patients, only weight, functional status and CD4 count are significant at 5% level of significance. Hence,

- A one kg increase in the weight of an HIV/AIDS patients reduces the hazard of death by 3.89% (a factor of 0.9611) (AHR=0.9611; 95%CI: 0.9274, 0.9960) assuming all the other variables constant.
- The hazard of death for working and ambulatory patients decreases by 72.27% (a factor of 0.2773) (AHR=0.2773; 95%CI: 0.1078, 0.7134) and 62.52% (a factor of 0.3748) (AHR=0.3748; 95%CI: 0.1484, 0.9469), respectively, relative to bedridden patients assuming all the other variables constant. In other words, the hazards of death for working and ambulatory patients are 0.2773 (AHR: 0.2773; 95%CI: 0.1078, 0.7134) and 0.3748 (AHR=0.3748; 95%: 0.1484, 0.9469) times lower than the hazard of bedridden patients assuming all the other variables constant.

- Holding the other variables constant, the hazard of death decreases as the CD4 count increases (AHR=0.9929; 95%CI: 0.9893, 0.9966).

Note: In the Cox modelling, there is an assumption of *proportional hazards*. The assumption is that the hazards for persons with different patterns of factors (covariates) are constant over time. For example, if the relative hazard of heart attack among diabetics is three times higher than among nondiabetics in the first year of the study, the relative hazard of heart attack must also be (about) three times higher among diabetics than nondiabetics in the second year of the study. Note that the hazard for a heart attack can be very different in the first year than in the second year (e.g., much higher in the first year than in the second year), but the difference between the hazards for diabetics and nondiabetics must be constant throughout the study period.

If this assumption of proportional hazards is not fulfilled, Cox-PH Model will not be appropriate and other type of models like Accelerated Failure Time Models are recommended.

13.5 Accelerated Failure Time (AFT) Model

An accelerated failure-time (AFT) model, also known as accelerated-time model follows log-time parametrization. It models the log mean-time as a function (mostly, linear) of the independent variables. That is,

$$\log(t_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The parameters are interpreted in terms of time ratios (TR), that is, $\text{TR}_j = \exp(\beta_j)$. Here, if TR = 1, then time passes at its "normal" rate. If TR < 1, then time passes more quickly for the subject (time is accelerated), so the event would be expected to occur sooner. And if TR > 1, then time passes more slowly for the subject (time is decelerated), so the event would be expected to occur later.

Example 13.2. A Weibull AFT model was fitted using the data from the example . Interpret the results.

Variable	$\hat{\beta}$	$\widehat{\text{SE}}(\hat{\beta})$	t-Statistic	95% CI for β	$\widehat{\text{TR}}$	95%CI for TR
Constant	1.1779	1.8392	0.64	(-2.4269,4.7827)		
Age	0.0138	0.0188	0.73	(-0.0230, 0.0506)	1.0139	(0.9773, 1.0519)
Male	-1.0060	0.3356	3.00	(-1.6638,-0.3482)*	0.3657	(0.1894, 0.7060)*
Weight	0.0628	0.0182	3.45	(0.0271, 0.0985)*	1.0648	(1.0275, 1.1035)*
Working	2.0018	0.4821	4.15	(1.0569, 2.9467)*	7.4024	(2.8774, 19.0430)*
Ambulatory	1.5492	0.4728	3.28	(0.6225, 2.4759)*	4.7077	(1.8636, 11.8924)*
Bedridden						
CD4	0.0112	0.0019	5.89	(0.0075, 0.0149)*	1.0113	(1.0075, 1.0150)*
ρ	0.6332	0.0852	7.43	(0.4662, 0.8002)*		

Solution: Of the five candidate predictors used in the Weibull AFT model, only age is not significant at 5% level of significance. Hence,

- Male patients are 0.3657 (ATR=0.3657; 95%CI: 0.1894, 0.7060) times faster to die than female patients. Or the mean time-to-death of male patients decreases by 63.43% (by a factor of 0.3657) (ATR=0.3657; 95%CI: 0.1894, 0.7060) relative to female patients assuming the other variables constant. The effect of being male is to accelerate time or the effect of being female is to slow down it, or; equivalently, the effect of being male is to hasten death, or that of being female is to delay it.

- A one kg increase in the weight of an HIV/AIDS patients increases their survival time by 6.48% (by a factor of 1.0648) (AHR=1.0648; 95%CI: 1.0275, 1.1035) assuming all the other variables constant.

- The mean survival time of working and ambulatory patients increases by a factor of 7.4024 (AHR=7.4024; 95%CI: 2.8774, 19.0430) and 4.7077 (AHR=4.7077; 95%CI: 1.8636, 11.8924), respectively, relative to bedridden patients assuming all the other variables constant.

- Holding the other variables constant, the mean time-to-death of HIV/AIDS patients increases by 1.13% (by a factor of 1.0113) (AHR=1.0113; 95%CI: 1.0075, 1.0150) as the CD4 count increases by one.

END OF THE COURSE!