MRCET CAMPUS

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS INSTITUTION – UGC, GOVT. OF INDIA)

Department of CSE

(Emerging Technologies)

(DATA SCIENCE)

B.TECH(R-20 Regulation)

(III YEAR – I SEM)

(2022-23)

STATISTICAL FOUNDATIONS IN DATA SCIENCE

LECTURE NOTES

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous Institution – UGC, Govt. of India)

Recognized under 2(f) and 12(B) of UGC ACT 1956

(Affiliated to JNTUH, Hyderabad, Approved by AICTE-Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)

Maisammaguda, Dhulapally (Post Via. Hakimpet), Secunderabad–500100, Telangana State, India

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

Vision

- "To be at the forefront of Emerging Technologies and to evolve as a Centre of Excellence in Research, Learning and Consultancy to foster the students into globally competent professionals useful to the Society."

Mission

The department of CSE (Emerging Technologies) is committed to:

- To offer highest Professional and Academic Standards in terms of Personal growth and satisfaction.
- Make the society as the hub of emerging technologies and thereby capture opportunities in new age technologies.
- To create a benchmark in the areas of Research, Education and Public Outreach.
- To provide students a platform where independent learning and scientific study are encouraged with emphasis on latest engineering techniques.

QUALITY POLICY

- To pursue continual improvement of teaching learning process of Undergraduate and Post Graduate programs in Engineering & Management vigorously.
- To provide state of art infrastructure and expertise to impart the quality education and research environment to students for a complete learning experiences.
- Developing students with a disciplined and integrated personality.
- To offer quality relevant and cost effective programmes to produce engineers as per requirements of the industry need.

For more information: www.mrcet.ac.in

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

Vision

- "To be at the forefront of Emerging Technologies and to evolve as a Centre of Excellence in Research, Learning and Consultancy to foster the students into globally competent professionals useful to the Society."

Mission

The department of CSE (Emerging Technologies) is committed to:

- To offer highest Professional and Academic Standards in terms of Personal growth and satisfaction.
- Make the society as the hub of emerging technologies and thereby capture opportunities in new age technologies.
- To create a benchmark in the areas of Research, Education and Public Outreach.
- To provide students a platform where independent learning and scientific study are encouraged with emphasis on latest engineering techniques.

QUALITY POLICY

- To pursue continual improvement of teaching learning process of Undergraduate and Post Graduate programs in Engineering & Management vigorously.
- To provide state of art infrastructure and expertise to impart the quality education and research environment to students for a complete learning experiences.
- Developing students with a disciplined and integrated personality.
- To offer quality relevant and cost effective programmes to produce engineers as per requirements of the industry need.

For more information: www.mrcet.ac.in

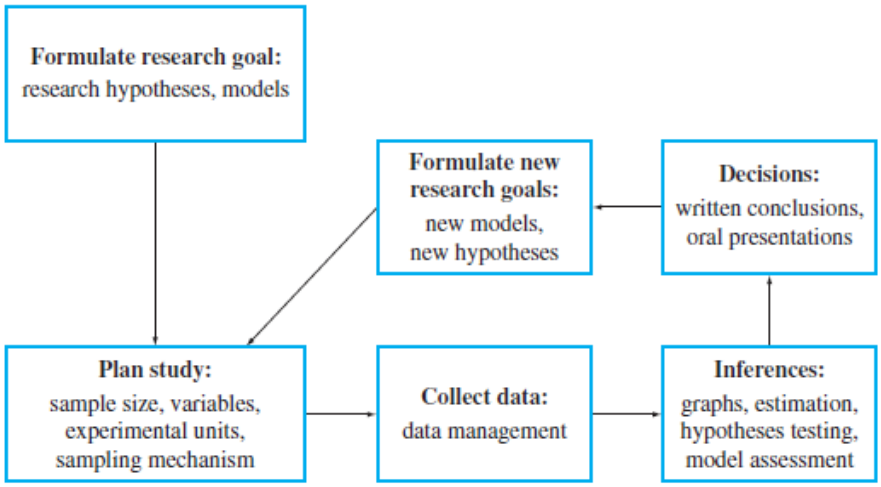
STATISTICAL FOUNDATION OF DATA SCIENCE

UNIT-I

Statistics is the scientific discipline that provides methods to help us make sense of data. Some people are suspicious of conclusions based on statistical analyses. Extreme skeptics, usually speaking out of ignorance, characterize the discipline as a subcategory of lying—something used for deception rather than for positive ends. However, we believe that statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us. Statistical methods are used in business, medicine, agriculture, social sciences, natural sciences, and applied sciences, such as engineering. The widespread use of statistical analyses in diverse fields has led to increased recognition that statistical literacy—a familiarity with the goals and methods of statistics—should be a basic component of a well-rounded educational program. The field of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

Statistics is the science of designing studies or experiments, collecting data and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both limited and variable. That is, statistics is the science of *Learning from Data.* Almost everyone—including corporate presidents, marketing representatives, social scientists, engineers, medical researchers, and consumers—deals with data. These data could be in the form of quarterly sales figures, percent increase in juvenile crime, contamination levels in water samples, and survival rates for patients undergoing medical therapy, census figures, or information that helps determine which brand of car to purchase. In this text, we approach the study of statistics by considering the four-step process in Learning from Data: (1) defining the problem, (2) collecting the data, (3) summarizing the data, and (4) analyzing data, interpreting the analyses, and communicating results. Through the use of these four steps in Learning from Data, our study of statistics closely parallels the Scientific Method, which is a set of principles and procedures used by successful scientists in their pursuit of knowledge. The method involves the formulation of research goals, the design of observational studies and/or experiments, the collection of data, the modeling/ analyzing of the data in the context of research goals, and the testing of hypotheses. The conclusions of these steps are often the formulation of new research goals for another study.

FIGURE 1.1
Scientific Method Schematic



Three Reasons to Study Statistics

Because statistical methods are used to organize, summarize, and draw conclusions from data, a familiarity with statistical techniques and statistical literacy is vital in today's society. Everyone needs to have a basic understanding of statistics, and many college majors require at least one course in statistics. There are three important reasons why statistical literacy is important: (1) to be informed, (2) to understand issues and be able to make sound decisions based on data, and (3) to be able to evaluate decisions that affect your life.

Statistics and the Data Analysis Process

Data and conclusions based on data appear regularly in a variety of settings: newspapers, television and radio advertisements, magazines, and professional publications. In business, industry, and government, informed decisions are often data driven. Statistical methods, used appropriately, allow us to draw reliable conclusions based on data. Once data have been collected or once an appropriate data source has been identified, the next step in the data analysis process usually involves organizing and summarizing the information. Tables, graphs, and numerical summaries allow increased understanding and provide an effective way to present data. Methods for organizing and summarizing data make up the branch of statistics called **descriptive statistics**. After the data have been summarized, we often wish to draw conclusions or make decisions based on the data. This usually involves generalizing from a small group of individuals or objects that we have studied to a much larger group.

For example, the admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2006 term failed to enroll at the university. The population of interest to the director consists of all accepted applicants who did not enroll in the fall 2006 term. Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 selected students. These 300 students constitute a sample.

The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

The Data Analysis Process

Statistics involves the collection and analysis of data. Both tasks are critical. Raw data without analysis are of little value, and even a sophisticated analysis cannot extract meaningful information from data that were not collected in a sensible way.

■ Planning and Conducting a Statistical Study Scientific studies are undertaken to answer questions about our world. Is a new flu vaccine effective in preventing illness?

Is the use of bicycle helmets on the rise? Are injuries that result from bicycle accidents less severe for riders who wear helmets than for those who do not? How many credit cards do college students have? Do engineering students pay more for textbooks than do psychology students? Data collection and analysis allow researchers to answer such questions.

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to informed conclusions based on the resulting data. The process can be organized into the following six steps:

1. Understanding the nature of the problem. Effective data analysis requires an understanding of the research problem. We must know the goal of the research and what questions we hope to answer. It is important to have a clear direction before gathering data to lessen the chance of being unable to answer the questions of interest using the data collected.

2. Deciding what to measure and how to measure it. The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious (e.g., in a study of the relationship between the weight of a Division I football player and position played, you would need to collect data on player weight and position), but in other cases the choice of information is not as straightforward (e.g., in a study of the relationship between preferred learning style and intelligence, how would you define learning style and measure it and what measure of intelligence would you use?). It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.

3. Data collection. The data collection step is crucial. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. Even if a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood and judged to be acceptable. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the subsequent conclusions that can be drawn depend on how the data are collected.

4. Data summarization and preliminary analysis. After the data are collected, the next step usually involves a preliminary analysis that includes summarizing the data graphically and

numerically. This initial analysis provides insight into important characteristics of the data and can provide guidance in selecting appropriate methods for further analysis.

5. Formal data analysis. The data analysis step requires the researcher to select and apply the appropriate inferential statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.

6. Interpretation of results. Several questions should be addressed in this final step—for example, what conclusions can be drawn from the analysis? How do the results of the analysis inform us about the stated research problem or question? And how can our results guide future research? The interpretation step often leads to the formulation of new research questions, which, in turn, leads back to the first step. In this way, good data analysis is often an iterative process.

Observational Studies

A study may be either observational or experimental. In an **observational study,** the researcher records information concerning the subjects under study without any interference with the process that is generating the information. The researcher is a passive observer of the transpiring events. Observational studies may be dichotomized into either a **comparative study** or **descriptive study.** In a comparative study, two or more methods of achieving a result are compared for effectiveness. For example, three types of healthcare delivery methods are compared based on cost effectiveness. Alternatively, several groups are compared based on some common attribute. For example, the starting income of engineers is contrasted from a sample of new graduates from private and public universities. In a descriptive study, the major purpose is to characterize a population or process based on certain attributes in that population or process—for example, studying the health status of children under the age of 5 years old in families without health insurance or assessing the number of overcharges by companies hired under federal military contracts. In an observational study, the factors (treatments) of interest are not manipulated while making measurements or observations. The researcher in an environmental impact study is attempting to establish the current state of a natural setting from which subsequent changes may be compared.

Observational studies are of three basic types:

- A **sample survey** is a study that provides information about a population at a particular point in time (current information).
- A **prospective study** is a study that observes a population in the present using a sample survey and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes.
- A **retrospective study** is a study that observes a population in the present using a sample survey and also collects information about the subjects in the sample regarding the occurrence of specific outcomes that have already taken place.

Experimental Studies

In an **experimental study** the researcher actively manipulates certain variables associated with the study, called the **explanatory variables,** and then records their effects on the **response variables** associated with the experimental subjects. A severe limitation of observational studies is that the recorded values of the response variables may be affected by variables other than the explanatory variables. These variables are not under the control of the researcher. They are called **confounding variables.** The effects of the confounding variables and the explanatory variables on the response variable cannot be separated due to the lack of control the researcher has over the physical setting in which the observations are made. In an experimental study, the researcher attempts to maintain control overall variables that may have an effect on the response variables.
An experimental study may be conducted in many different ways. In some studies, the researcher is interested in collecting information from an undisturbed natural process or setting In experimental studies, the researcher controls the crucial factors by one of two methods.
Method 1: The subjects in the experiment are randomly assigned to the treatments. For example, ten rats are randomly assigned to each of the four dose levels of an experimental drug under investigation.
Method 2: Subjects are randomly selected from different populations of interest. For example, 50 male and 50 female dogs are randomly selected from animal shelters in large and small cities and tested for the presence of heart worms.

In experimental studies, it is crucial that the scientist follows a systematic plan established prior to running the experiment. The plan includes how all randomization is conducted, either the assignment of experimental units to treatments or the selection of units from the treatment populations. There may be extraneous factors present that may affect the experimental units. These factors may be present as subtle differences in the experimental units or slight differences in the surrounding environment during the conducting of the experiment. The randomization process ensures that, on the average, any large differences observed in the responses of the experimental units in different treatment groups can be attributed to the differences in the groups and not to factors that were not controlled during the experiment. The plan should also include many other aspects on how to conduct the experiment. A list of some of the items that should be included in such a plan is listed here:
1. The research objectives of the experiment
2. The selection of the factors that will be varied (the treatments)
3. The identification of extraneous factors that may be present in the experimental units or in the environment of the experimental setting (the blocking factors)
4. The characteristics to be measured on the experimental units (response variable)
5. The method of randomization, either randomly selecting from treatment populations or the random assignment of experimental units to treatments
6. The procedures to be used in recording the responses from the experimental units

7. The selection of the number of experimental units assigned to each treatment may require designating the level of significance and power of tests or the precision and reliability of confidence intervals
8. A complete listing of available resources and materials

Types of Data and Some Simple Graphical Displays

A data set consisting of observations on a single attribute is a **univariate data set**. A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses. A univariate data set is **numerical** (or **quantitative**) if each observation is a number.

A numerical variable results in **discrete** data if the possible values of the variable correspond to isolated points on the number line. A numerical variable results in **continuous** data if the set of possible values forms an entire interval on the number line.

Frequency Distributions and Bar Charts for Categorical Data

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. When the data set is categorical, a common way to present the data is in the form of a table, called a *frequency distribution*.
A **frequency distribution for categorical data** is a table that displays the possible categories along with the associated frequencies and/or relative frequencies. The **frequency** for a particular category is the number of times the category appears in the data set. The **relative frequency** for a particular category is the fraction or proportion of the observations resulting in the category.
If the table includes relative frequencies, it is sometimes referred to as a **relative frequency distribution**.

Example 1.1 Motorcycle Helmets—Can You See Those Ears?

In 2003, the U.S. Department of Transportation established standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report "Motorcycle Helmet Use in 2005—Overall Results" (National Highway Traffic Safety Administration, August 2005) summarized data collected in June of 2005 by observing 1700 motorcyclists nationwide at selected roadway locations. Each time a motorcyclist passed by, the observer noted whether the rider was wearing no helmet, a noncompliant helmet, or a compliant helmet. Using the coding

N = no helmet
NH= noncompliant helmet
CH= compliant helmet

a few of the observations were

CH N CH NH N CH CH CH N N

There were also 1690 additional observations, which we didn't reproduce here! In total, there were 731 riders who wore no helmet, 153 who wore a noncompliant helmet, and 816 who wore a compliant helmet.
The corresponding frequency distribution is given in Table 1.1.

Table 1.1 Frequency Distribution for Helmet Use

Helmet Use Category	Frequency	Relative Frequency
No helmet	731	.430 ← *731/1700*
Noncompliant helmet	153	.090 ← *153/1700*
Compliant helmet	816	.480
	1700 ← *Total number of observations*	1.00 ← *Should total 1, but in some cases may be slightly off due to rounding*

From the frequency distribution, we can see that a large number of riders (43%) were not wearing a helmet, but most of those who wore a helmet were wearing one that met the Department of Transportation safety standard.

Bar Charts

A **bar chart** is a graph of the frequency distribution of categorical data. Each category in the frequency distribution is represented by a bar or rectangle, and the picture is constructed in such a way that the *area* of each bar is proportional to the corresponding frequency or relative frequency.

Bar Charts

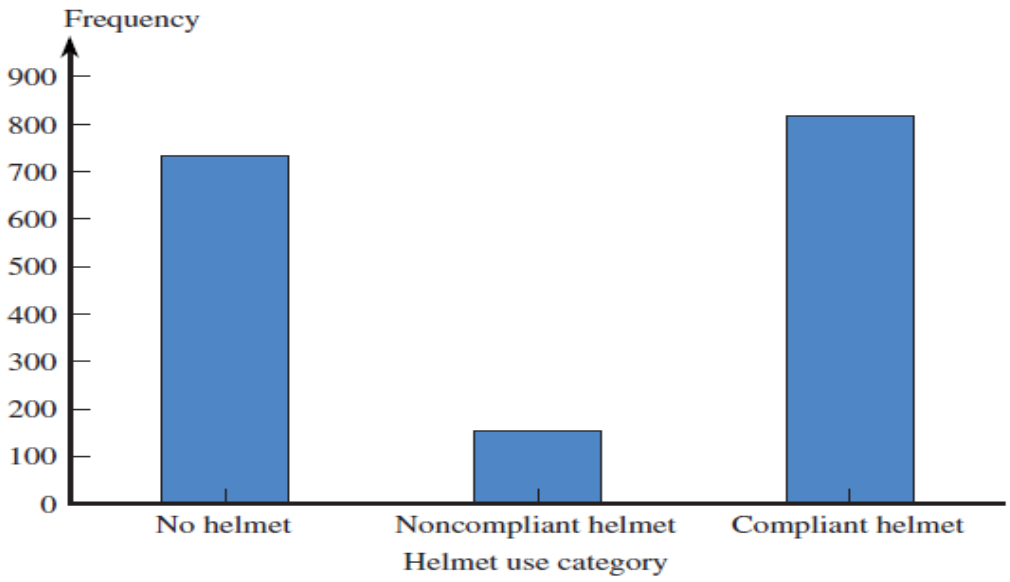
When to Use Categorical data.

How to Construct

1. Draw a horizontal line, and write the category names or labels below the line at regularly spaced intervals.
2. Draw a vertical line, and label the scale using either frequency or relative frequency.
3. Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to frequency and relative frequency.

What to Look For

- Frequently and infrequently occurring categories.



Bar Chart of Example 1.1

Dotplots for Numerical Data .

A dotplot is a simple way to display numerical data when the data set is reasonably small. Each observation is represented by a dot above the location corresponding to its value on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence and these dots are stacked vertically.

Dotplots

When to Use Small numerical data sets.

How to Construct

1. Draw a horizontal line and mark it with an appropriate measurement scale.
2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

What to Look For Dotplots convey information about:

- A representative or typical value in the data set.
- The extent to which the data values spread out.
- The nature of the distribution of values along the number line.
- The presence of unusual values in the data set.

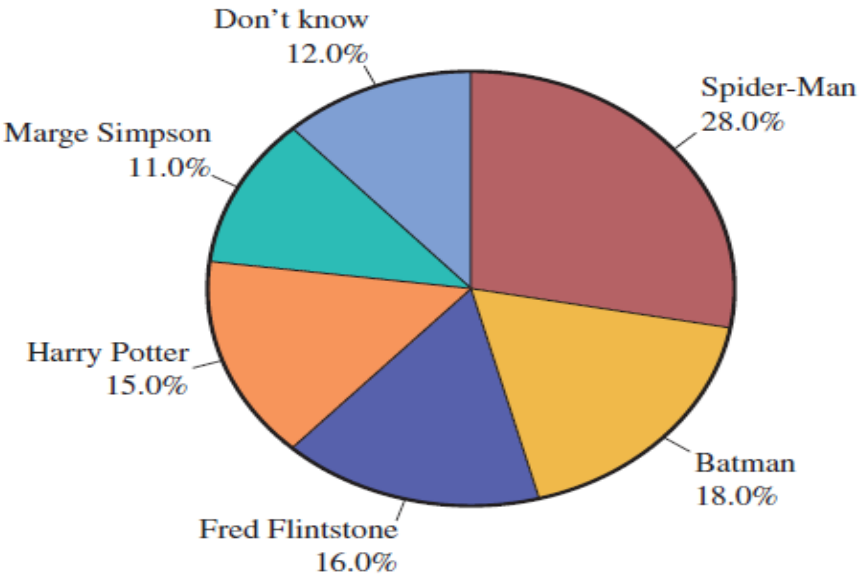
Pie Charts

A categorical data set can also be summarized using a pie chart. In a pie chart, a circle is used to represent the whole data set, with "slices" of the pie representing the possible categories. The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency. Pie charts are most effective for summarizing data sets when there are not too many different categories.

Example 1.2

Life Insurance for Cartoon Characters??

The article "Fred Flintstone, Check Your Policy" (*The Washington Post,* October 2, 2005) summarized the results of a survey of 1014 adults conducted by the Life and Health Insurance Foundation for Education. Each person surveyed was asked to select which of five fictional characters, Spider-Man, Batman, Fred Flintstone, Harry Potter, and Marge Simpson, he or she thought had the greatest need for life insurance.



The survey results were quite different from an insurance expert's assessment. His opinion was that Fred Flintstone, a married father with a young child, was by far the one with the greatest need for life insurance. Spider-Man, unmarried with an elderly aunt, would need life insurance only if his aunt relied on him to supplement her income. Batman, a wealthy bachelor with no dependents, doesn't need life insurance in spite of his dangerous job!

Pie Chart for Categorical Data

When to Use Categorical data with a relatively small number of possible categories. Pie charts are most useful for illustrating proportions of the whole data set for various categories.

How to Construct

1. Draw a circle to represent the entire data set.
2. For each category, calculate the "slice" size. Because there are 360 degrees in a circle

   slice size = 360 (category relative frequency)

3. Draw a slice of appropriate size for each category. This can be tricky, so most pie charts are generated using a graphing calculator or a statistical software package.

What to Look For

- Categories that form large and small proportions of the data set.

Displaying Numerical Data: Stem-and-Leaf Displays

A stem-and-leaf display is an effective and compact way to summarize univariate numerical data. Each number in the data set is broken into two pieces, a stem and a leaf. The **stem** is the first part of the number and consists of the beginning digit(s). The **leaf** is the last part of the number and consists of the final digit(s). For example, the number 213 might be split into a stem of 2 and a leaf of 13 or a stem of 21 and a leaf of 3. The resulting stems and leaves are then used to construct the display.

Example 1.3

Should Doctors Get Auto Insurance Discounts?

● Many auto insurance companies give job-related discounts of between 5 and 15%. The article "Auto-Rate Discounts Seem to Defy Data" (*San Luis Obispo Tribune,* June 19, 2004) included the accompanying data on the number of automobile accidents per year for every 1000 people in 40 occupations.

Occupation	Accidents per 1000	Occupation	Accidents per 1000
Student	152	Banking-finance	89
Physician	109	Customer service	88
Lawyer	106	Manager	88
Architect	105	Medical support	87
Real estate broker	102	Computer-related	87
Enlisted military	99	Dentist	86
Social worker	98	Pharmacist	85
Manual laborer	96	Proprietor	84
Analyst	95	Teacher, professor	84
Engineer	94	Accountant	84
Consultant	94	Law enforcement	79
Sales	93	Physical therapist	78
Military officer	91	Veterinarian	78
Nurse	90	Clerical, secretary	77
School administrator	90	Clergy	76
Skilled labor	90	Homemaker	76
Librarian	90	Politician	76
Creative arts	90	Pilot	75
Executive	89	Firefighter	67
Insurance agent	89	Farmer	43

4	3
5	
6	7
7	56667889
8	44567788999
9	000013445689
10	2569
11	
12	
13	
14	
15	2

Stem: Tens
Leaf: Ones

The numbers in the vertical column on the left of the display are the **stems**. Each number to the right of the vertical line is a **leaf** corresponding to one of the observations in the data set. The legend

Stem: Tens
Leaf: Ones

tells us that the observation that had a stem of 4 and a leaf of 3 corresponds to an occupation with an accident rate of 43 per 1000 (as opposed to 4.3 or 0.43). Similarly, the observation with the stem of 10 and leaf of 2 corresponds to 102 accidents per 1000 (the leaf of 2 is the ones digit) and the observation with the stem of 15 and leaf of 2 corresponds to 152 accidents per 1000.

The above stem-leaf chart suggests that a typical or representative value is in the stem 8 or 9 row, perhaps around 90. The observations are mostly concentrated in the 75 to 109 range, but there are a couple of values that stand out on the low end (43 and 67) and one observation (152) that is far removed from the rest of the data on the high end.

From the point of view of an auto insurance company it might make sense to offer discounts to occupations with low accident rates—maybe farmers (43 auto accidents per 1000 farmers) or firefighters (67 accidents per 1000 firefighters) or even some of the occupations with accident rates in the 70s. The "discounts seem to defy data" in the title of the article refers to the fact that some insurers provide discounts to doctors and engineers, but not to homemakers, politicians, and other occupations with lower accident rates. Two possible explanations were offered for this apparent discrepancy. One is that it is possible that while some occupations have higher accident rates, they also have lower average cost per claim. Accident rates alone may not reflect the actual cost to the insurance company. Another possible explanation is that the insurance companies may offer the discounted auto insurance in order to attract people who would then also purchase other types of insurance such as malpractice or liability insurance.

Stem-and-Leaf Displays

When to Use

Numerical data sets with a small to moderate number of observations (does not work well for very large data sets)

How to Construct

1. Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

What to Look For The display conveys information about

- a representative or typical value in the data set
- the extent of spread about a typical value
- the presence of any gaps in the data
- the extent of symmetry in the distribution of values
- the number and location of peaks

Example 1.4

Median Ages in 2030

● The accompanying data on the Census Bureau's projected median age in 2030 for the 50 U.S. states and Washington D.C. appeared in the article "2030 Forecast: Mostly Gray" (*USA Today,* April 21, 2005). The median age for a state is the age that divides the state's residents so that half are younger than the median age and half are older than the median age.

Projected Median Age

41.0	32.9	39.3	29.3	37.4	35.6	41.1	43.6	33.7	45.4	35.6	38.7
39.2	37.8	37.7	42.0	39.1	40.0	38.8	46.9	37.5	40.2	40.2	39.0
41.1	39.6	46.0	38.4	39.4	42.1	40.8	44.8	39.9	36.8	43.2	40.2
37.9	39.1	42.1	40.7	41.3	41.5	38.3	34.6	30.4	43.9	37.8	38.5
46.7	41.6	46.4									

2H	9
3L	0234
3H	5567777778888899999999
4L	00000011111112223334
4H	56666

Stem: Tens
Leaf: Ones

Comparative Stem-and-Leaf Displays

Example 1.5

Progress for Children

● The UNICEF report "Progress for Children" (April, 2005) included the accompanying data on the percentage of primary-school-age children who were enrolled in school for 19 countries in Northern Africa and for 23 countries in Central Africa.

Northern Africa

54.6 34.3 48.9 77.8 59.6 88.5 97.4 92.5 83.9 96.9 88.9
98.8 91.6 97.8 96.1 92.2 94.9 98.6 86.6

Central Africa

58.3 34.6 35.5 45.4 38.6 63.8 53.9 61.9 69.9 43.0 85.0
63.4 58.4 61.9 40.9 73.9 34.8 74.4 97.4 61.0 66.7 79.6
98.9

Central Africa		Northern Africa
4854	3	4
035	4	8
838	5	49
6113913	6	
943	7	76
5	8	8386
87	9	7268176248

Stem: Tens
Leaf: Ones

From the comparative stem-and-leaf display you can see that there is quite a bit of variability in the percentage enrolled in school for both Northern and Central African countries and that the shapes of the two data distributions are quite different. The percentage enrolled in school tends to be higher in Northern African countries than in Central African countries, although the smallest value in each of the two data sets is about the same. For Northern African countries the distribution of values has a single peak in the 90s with the number of observations declining as we move toward the stems corresponding to lower percentages enrolled in school. For Central African countries the distribution is more symmetric, with a typical value in the mid 60s.

Displaying Numerical Data: Frequency Distributions and Histograms

A stem-and-leaf display is not always an effective summary technique; it is unwieldy when the data set contains a large number of observations. Frequency distributions and histograms are displays that are useful for summarizing even a large data set in a compact fashion.

Histogram for Discrete Numerical Data

When to Use

Discrete numerical data. Works well even for large data sets.

How to Construct

1. Draw a horizontal scale, and mark the possible values of the variable.
2. Draw a vertical scale, and mark it with either frequency or relative frequency.
3. Above each possible value, draw a rectangle centered at that value (so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on). The height of each rectangle is determined by the corresponding frequency or relative frequency. Often possible values are consecutive whole numbers, in which case the base width for each rectangle is 1.

What to Look For

- Central or typical value
- Extent of spread or variation
- General shape
- Location and number of peaks
- Presence of gaps and outliers

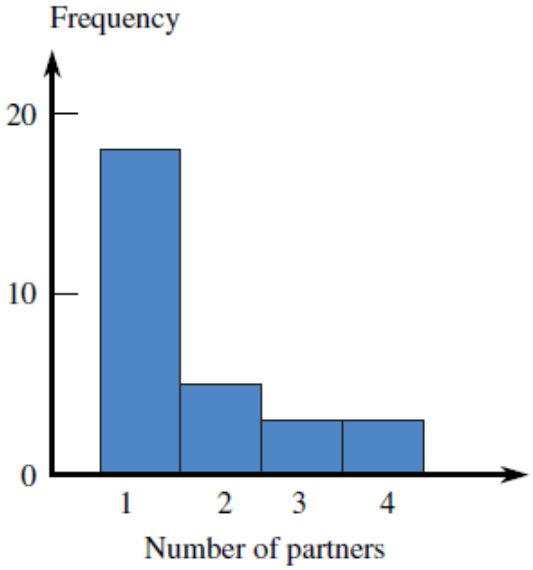
Example 1.6

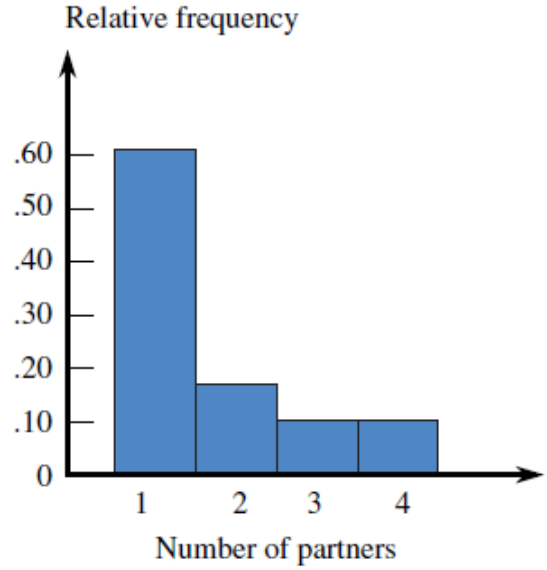
Promiscuous Raccoons!

● The authors of the article "Behavioral Aspects of the Raccoon Mating System: Determinants of Consortship Success" (*Animal Behaviour* [1999]: 593–601) monitored raccoons in southern Texas during three mating seasons in an effort to describe mating behavior. Twenty-nine female raccoons were observed, and the number of male partners during the time the female was accepting partners (generally 1 to 4 days each year) was recorded for each female. The resulting data were as follows:

1	3	2	1	1	4	2	4	1	1	1
3	1	1	1	1	2	2	1	1	4	1
1	2	1	1	1	1	3				

Number of Partners	Frequency	Relative Frequency
1	18	.621
2	5	.172
3	3	.103
4	3	.103
	29	.999 ← Differs from 1 due to rounding





Example 1.7

Enrollments at Public Universities

● States differ widely in the percentage of college students who are enrolled in public institutions. The National Center for Education Statistics provided the accompanying data on this percentage for the 50 U.S. states for fall 2002.

Percentage of College Students Enrolled in Public Institutions

86	96	66	86	80	78	62	81	77	81	77	76
73	69	76	90	78	82	70	83	46	80	78	93
66	91	76	86	58	81	91	57	81	88	71	86
84	57	44	82	79	67	86	75	55	75	80	80
85	69										

Table 1.2 Frequency Distribution for Percentage of College Students Enrolled in Public Institutions

Class Interval	Frequency	Relative Frequency
40 to <50	2	.04
50 to <60	4	.08
60 to <70	6	.12
70 to <80	14	.28
80 to <90	19	.38
90 to <100	5	.10
	50	1.00

Histogram for Continuous Numerical Data When the Class Interval Widths Are Equal

When to Use

Continuous numerical data. Works well, even for large data sets.

How to Construct

1. Mark the boundaries of the class intervals on a horizontal axis.
2. Use either frequency or relative frequency on the vertical axis.
3. Draw a rectangle for each class directly above the corresponding interval (so that the edges are at the class boundaries). The height of each rectangle is the frequency or relative frequency of the corresponding class interval.

What to Look For

- Central or typical value
- Extent of spread or variation
- General shape
- Location and number of peaks
- Presence of gaps and outliers

Constructing a Histogram for Continuous Data When Class Interval Widths Are Unequal

In this case, frequencies or relative frequencies should not be used on the vertical axis. Instead, the height of each rectangle, called the **density** for the class, is given by

$$\text{density} = \text{rectangle height} = \frac{\text{relative frequency of class interval}}{\text{class interval width}}$$

The vertical axis is called the **density scale**.

Time Series Plots

Data sets often consist of measurements collected over time at regular intervals so that we can learn about change over time. For example, stock prices, sales figures, and other socio-economic indicators might be recorded on a weekly or monthly basis.

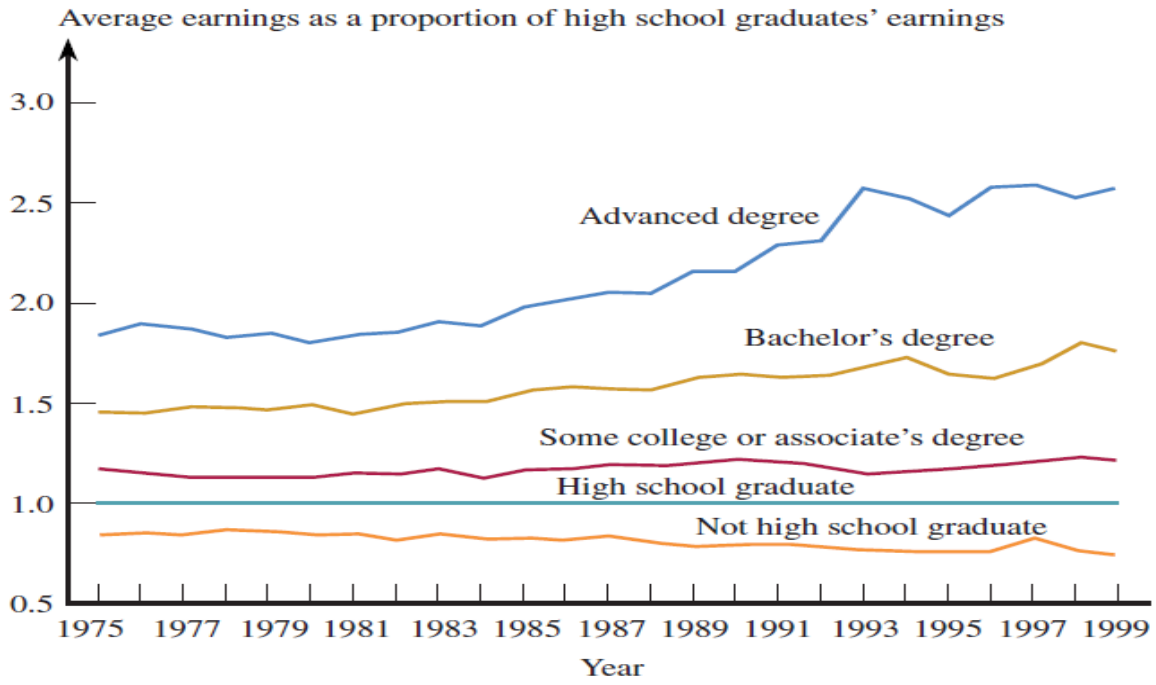
A **time-series plot** (sometimes also called a time plot) is a simple graph of data collected over time that can be invaluable in identifying trends or patterns that might be of interest.

A time-series plot can be constructed by thinking of the data set as a bivariate data-set, where y is the variable observed and x is the time at which the observation was made. These (x, y) pairs are plotted as in a scatterplot. Consecutive observations are then connected by a line segment; this aids in spotting trends over time.

Example 1.8

Education Level and Income—Stay in School!

The time-series plot shown in Figure below appears on the U.S. Census Bureau website. It shows the average earnings of workers by educational level as a proportion of the average earnings of a high school graduate over time. For example, we can see from this plot that in 1993 the average earnings for people with bachelor's degrees was about 1.5 times the average for high school graduates. In that same year, the average earnings for those who were not high school graduates was only about 75% (a proportion of .75) of the average for high school graduates. The time-series plot also shows that the gap between the average earnings for high school graduates and those with a bachelor's degree or an advanced degree widened during the 1990s.



Summarizing a Data Set: Boxplots

It would be nice to have a method of summarizing data that gives more detail than just a measure of center and spread and yet less detail than a stem-and-leaf display or histogram. A *boxplot* is one such technique. It is compact, yet it provides information about the center, spread, and symmetry or skewness of the data. We consider two types of boxplots: the skeletal boxplot and the modified boxplot.

Construction of a Skeletal Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and a right (or upper) edge at the upper quartile. The box width is then equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Extend horizontal (or vertical) line segments, called whiskers, from each end of the box to the smallest and largest observations in the data set.

Example 1.9

Revisiting Hospital Cost-to-Charge Ratios

Let's reconsider the cost-to-charge data for hospitals in Oregon. The ordered observations are:

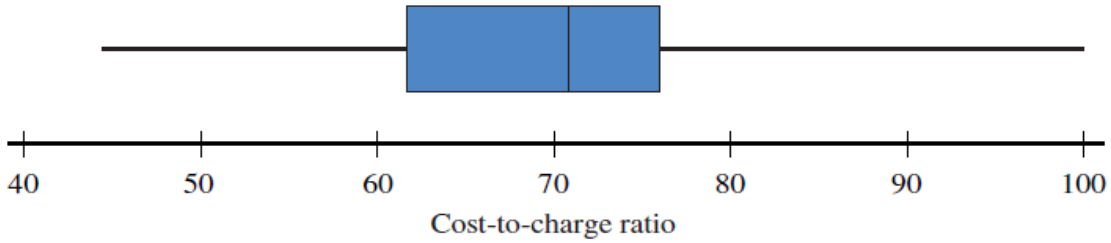
Ordered Data

Lower half	45	48	50	54	57	60	60	**62**	63	63	64	65	67	68	69
Median								**71**							
Upper half	71	72	72	74	74	75	75	**76**	80	83	84	88	100	100	100

To construct a boxplot of these data, we need the following information: the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation. This collection of summary measures is often referred to as the **five-number summary**. For this data set, we have-

smallest observation = 45
lower quartile = median of the lower half = 62
median = 16th observation in the ordered list = 71
upper quartile = median of the upper half = 76
largest observation = 100

The figure shows the corresponding boxplot. The median line is somewhat closer to the upper edge of the box than to the lower edge, suggesting a concentration of values in the upper part of the middle half. The upper whisker is a bit longer than the lower whisker.



DEFINITION

An observation is an **outlier** if it is more than 1.5(iqr) away from the nearest quartile (the nearest end of the box).

An outlier is **extreme** if it is more than 3(iqr) from the nearest end of the box, and it is **mild** otherwise.

A **modified boxplot** represents mild outliers by shaded circles and extreme outliers by open circles, and the whiskers extend on each end to the most extreme observations that are *not* outliers.

Construction of a Modified Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge is at the lower quartile and whose right (or upper) edge is at the upper quartile. The box width is then equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Determine if there are any mild or extreme outliers in the data set.
5. Draw whiskers that extend from each end of the box to the most extreme observation that is *not* an outlier.
6. Draw a solid circle to mark the location of any mild outliers in the data set.
7. Draw an open circle to mark the location of any extreme outliers in the data set.

Example 1.10

Golden Rectangles

● The accompanying data came from an anthropological study of rectangular shapes (*Lowie's Selected Papers in Anthropology,* Cora Dubios, ed. [Berkeley, CA: University of California Press, 1960]: 137–142). Observations were made on the variable x = width/length for a sample of $n = 20$ beaded rectangles used in Shoshoni Indian leather handicrafts:

.553	.570	.576	.601	.606	.606	.609	.611	.615	.628
.654	.662	.668	.670	.672	.690	.693	.749	.844	.933

The quantities needed for constructing the modified boxplot follow:

median = .641
lower quartile = .606
upper quartile = .681

iqr = .681 − .606 = .075
1.5(iqr) = .1125
3(iqr) = .225

Thus,

$$\text{(upper quartile)} + 1.5(\text{iqr}) = .681 + .1125 = .7935$$
$$\text{(lower quartile)} - 1.5(\text{iqr}) = .606 - .1125 = .4935$$

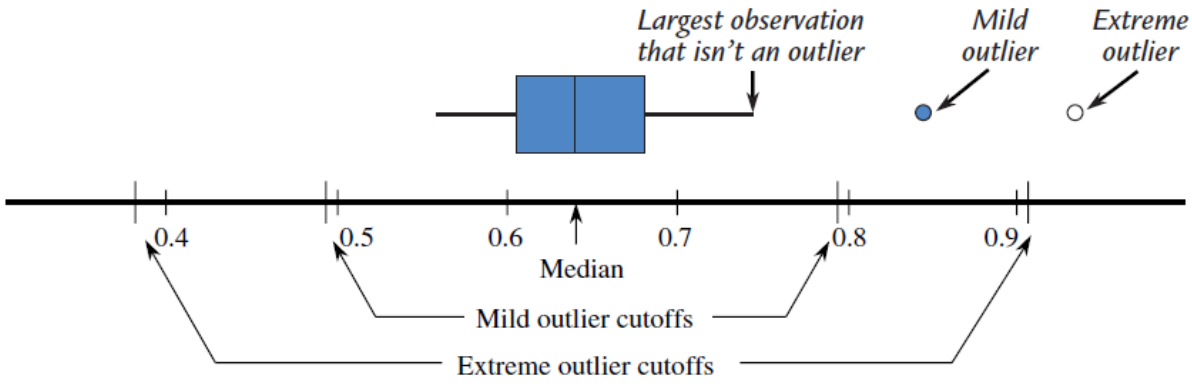
So 0.844 and 0.933 are both outliers on the upper end (because they are larger than 0.7935), and there are no outliers on the lower end (because no observations are smaller than 0.4935). Because

$$\text{(upper quartile)} + 3(\text{iqr}) = 0.681 + 0.225 = 0.906$$

0.933 is an extreme outlier and 0.844 is only a mild outlier. The upper whisker extends to the largest observation that is not an outlier, 0.749, and the lower whisker extends to 0.553. The boxplot is presented in Figure

The median line is not at the center of the box, so there is a slight asymmetry in the middle half of the data.

However, the most striking feature is the presence of the two outliers. These two x values considerably exceed the "golden ratio" of 0.618, used since antiquity as an aesthetic standard for rectangles.



UNIT-II

2.1 PROBABILITY: Probability is the science of uncertainty. It provides precise mathematical rules for understanding and analyzing our own ignorance. It does not tell us tomorrow's weather or next week's stock prices; rather, it gives us a framework for working with our limited knowledge and for making sensible decisions based on what we do and do not know.

A formal definition of probability begins with a *sample space*, often written S. This sample space is any set that lists all possible *outcomes* (or, *responses*) of some unknown experiment or situation. For example, perhaps

$$S = \{\text{rain, snow, clear}\}$$

when predicting tomorrow's weather. Or perhaps S is the set of all positive real numbers, when predicting next week's stock price. The point is, S can be any set at all, even an infinite set. We usually write s for an element of S, so that $s \in S$. Note that S describes only those things that we are interested in; if we are studying weather, then rain and snow are in S, but tomorrow's stock prices are not.

Classical Approach to Probability for Equally Likely Outcomes

When the outcomes in the sample space of a chance experiment are equally likely, the **probability of an event E**, denoted by $P(E)$, is the ratio of the number of outcomes favorable to E to the total number of outcomes in the sample space:

$$P(E) = \frac{\text{number of outcomes favorable to } E}{\text{number of outcomes in the sample space}}$$

According to this definition, the calculation of a probability consists of counting the number of outcomes that make up an event, counting the number of outcomes in the sample space, and then dividing.

Properties of Probability:

1. $P(A)$ is always a nonnegative real number, between 0 and 1 inclusive.
2. $P(\emptyset) = 0$, i.e., if A is the empty set $\emptyset$, then $P(A) = 0$.
3. $P(S) = 1$, i.e., if A is the entire sample space S, then $P(A) = 1$.
4. P is *(countably) additive*, meaning that if $A_1, A_2, \ldots$ is a finite or countable sequence of disjoint events, then

$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.

Theorem 2.1 (*Law of total probability, unconditioned version*) Let $A_1, A_2, \ldots$ be events that form a partition of the sample space S. Let B be any event. Then

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots.$$

Theorem 2.2 Let A and B be two events with $A \supseteq B$. Then

$$P(A) = P(B) + P(A \cap B^C).$$

Corollary 2.2.1 (*Monotonicity*) Let A and B be two events, with $A \supseteq B$. Then

$$P(A) \geq P(B).$$

Corollary 2.2.2 Let A and B be two events, with $A \supseteq B$. Then

$$P(A \cap B^C) = P(A) - P(B) .$$

Theorem 2.3. (*Principle of inclusion–exclusion, two-event version*) Let A and B be two events. Then

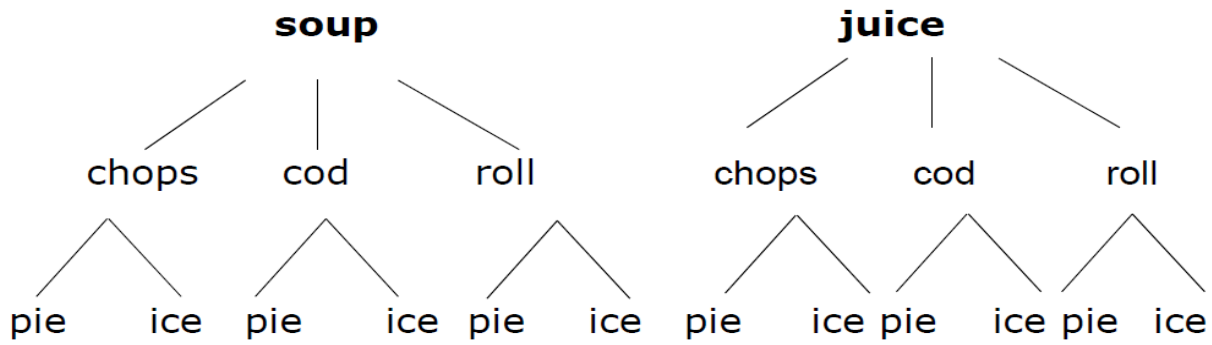
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

2.2 COMBINATORIAL PRINCIPLES

The science of counting is called *combinatorics*, and some aspects of it are very sophisticated. Combinatorics concerns itself with finite collections of discrete objects. With the growth of digital devices, especially digital computers, discrete mathematics has become more and more important. Counting problems arise when the combinatorial problem is to count the number of different arrangements of collections of objects of a particular kind. Such counting problems arise frequently when we want to calculate probabilities and so they are of wider application than might appear at first sight. Some counting problems are very easy, others are extremely difficult.

PROBLEM I: A café has the following menu:
Tomato soup, Fruit juice --- Lamb chops, Baked cod, Nut roll --- Apple pie, Strawberry ice. How many different three course meals could you order?

Solution:



We would obtain 2x3x2=12 as the total of possible meals.

The problems we study are:

✦ Counting assignments. The paradigm problem is how many ways can we paint a row of n houses, each in any of k colors.

✦ Counting permutations. The paradigm problem here is to determine the number of different orderings for n distinct items.

✦ Counting ordered selections that is, the number of ways to pick k things out of n and arrange the k things in order. The paradigm problem is counting the number of ways different horses can win, place, and show in a horse race.

✦ Counting the combinations of m things out of n, that is, the selection of m from n distinct objects, without regard to the order of the selected objects. The paradigm problem is counting the number of possible poker hands.

✦Counting permutations with some identical items. The paradigm problem is counting the number of anagrams of a word that may have some letters appearing more than once.

✦ Counting the number of ways objects, some of which may be identical, can be distributed among bins. The paradigm problem is counting the number of ways of distributing fruits to children.

2.3 CONDITIONAL PROBABILITY

Sometimes the knowledge that one event has occurred changes the likelihood that another event will occur. One example involves a population in which 0.1% of all individuals have a certain disease. The presence of the disease cannot be discerned from outward appearances, but there is a diagnostic test available. Unfortunately, the test is not infallible: 80% of those with positive test results actually have the disease; the other 20% who show positive test results are false-positives. To cast this in probability terms, consider the chance experiment in which an individual is randomly selected from the population. Define the following events:

E = event that the individual has the disease
F = event that the individual's diagnostic test is positive

Let $P(E|F)$ denote the probability of the event E *given that* the event F is known to have occurred. A new symbol has been used to indicate that a probability calculation has been made conditional on the occurrence of another event. The standard symbol for this is a vertical line, and it is read "given." We would say, "the probability that an individual has the disease *given* that the diagnostic test is positive," and represent this symbolically as $P(\text{has disease}|\text{positive test})$ or $P(E|F)$. This probability is called a **conditional probability**.

The information provided then implies that

$$P(E) = .001$$
$$P(E|F) = .8$$

That is, before we have diagnostic test information, the occurrence of E is unlikely, whereas once it is known that the test result is positive, the likelihood of the disease increases dramatically. (If this were not so, the diagnostic test would not be very useful!)

DEFINITION

Let E and F *be two events with* $P(F) > 0$. The **conditional probability of the event E given that the event F has occurred**, denoted by $P(E|F)$, is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

PROBLEM II

Consider the population of all families with two children. Representing the gender of each child using G for girl and B for boy results in four possibilities: BB, BG, GB, GG. The gender information is sequential, with the first letter indicating the gender of the older sibling. Thus, a family having a girl first and then a boy is denoted GB. If we assume that a child is equally likely to be male or female, each of the four possibilities in the sample space for the chance experiment that selects at random from families with two children is equally likely. Consider the following two questions:

1. What is the probability of obtaining a family with two girls, given that the family has at least one girl?
2. What is the probability of obtaining a family with two girls, given that the older sibling is a girl?

To many people, these questions *appear* to be identical. However, by computing the appropriate probabilities, we can see that they are indeed different.

For Question 1 we obtain

$$\begin{aligned} &P(\text{family with two girls} \,|\, \text{family has at least one girl}) \\ &\quad = \frac{P(\text{family with two girls and family with at least one girl})}{P(\text{family with at least one girl})} \\ &\quad = \frac{P(GG)}{P(GG \text{ or } BG \text{ or } GB)} \\ &\quad = \frac{1/4}{3/4} = \frac{.25}{.75} = .3333 \end{aligned}$$

For question 2 we calculate

$$
\begin{aligned}
&P(\text{family with two girls} \mid \text{family with older sibling a girl}) \\
&\quad = \frac{P(\text{family with two girls } \textit{and} \text{ family with older sibling a girl})}{P(\text{family with older sibling a girl})} \\
&\quad = \frac{P(GG)}{P(GB \textit{ or } GG)} \\
&\quad = \frac{1/4}{1/2} = \frac{.25}{.50} = .50
\end{aligned}
$$

2.4 INDEPENDENCE OF EVENTS

When two events E and F are such that $P(E|F) = P(E)$, the likelihood that event E has occurred is the same after we learn that F has occurred as it was before we had information about F's occurrence. We then say that E and F are independent of one another.

DEFINITION

Two events E and F are said to be **independent** if

$$P(E|F) = P(E)$$

If E and F are not independent, they are said to be **dependent** events.

If $P(E|F) = P(E)$, it is also true that $P(F|E) = P(F)$, and vice versa.

Multiplication Rule for Two Independent Events

The events E and F are independent if and only if

$$P(E \cap F) = P(E)P(F)$$

2.5 BAYE'S THEOREM:

Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different color balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

Let E_1, E_2…, E_n be a set of events associated with a sample space S, where all the events E_1, E_2…, E_n have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem-

$$P(E_i/A) = \frac{P(E_i)P(A/E_i)}{\sum_{k=1}^{n} P(E_k)P(A/E_k)}$$

PROBLEM III:

A bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from Bag I.

Solution:

Let E_1 be the event of choosing bag I, E_2 the event of choosing bag II, and A be the event of drawing a black ball.

Then,

$$P(E_1) = P(E_2) = \frac{1}{2}$$

Also, $P(A|E_1)$ = P(drawing a black ball from Bag I) = 6/10 = 3/5

$P(A|E_2)$ = P(drawing a black ball from Bag II) = 3/7

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E_1/A) = \frac{P(E_1)P(A/E_1)}{P(E_1)P(A/E_1)+P(E_2)P(A/E_2)}$$

$$= \frac{\frac{1}{2}\times\frac{3}{5}}{\left(\frac{1}{2}\times\frac{3}{5}\right)+\left(\frac{1}{2}\times\frac{3}{7}\right)}$$

$$= \frac{7}{12}$$

2.5.1 BAYES THEOREM APPLICATIONS

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. Bayesian inference has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc. For example, we can use Bayes' theorem to define the accuracy of medical test results by considering how likely any given person is to have a disease and the test's overall accuracy. Bayes' theorem relies on consolidating prior probability distributions to generate posterior probabilities. In Bayesian statistical inference, prior probability is the probability of an event before new data is collected.

2.6 RANDOM VARIABLES

DEFINITION

A numerical variable whose value depends on the outcome of a chance experiment is called a **random variable**. A random variable associates a numerical value with each outcome of a chance experiment.

A random variable is **discrete** if its set of possible values is a collection of isolated points on the number line. The variable is **continuous** if its set of possible values includes an entire interval on the number line.

2.6.1 PROBABILITY DISTRIBUTIONS FOR DISCRETE RANDOM VARIABLES

DEFINITION

The **probability distribution of a discrete random variable *X*** gives the probability associated with each possible *x* value. Each probability is the limiting relative frequency of occurrence of the corresponding *x* value when the chance experiment is repeatedly performed.

Common ways to display a probability distribution for a discrete random variable are a table, a probability histogram, or a formula.

Properties of Discrete Probability Distributions

1. For every possible x value, $0 \le p(x) \le 1$.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

2.6.2 PROBABILITY DISTRIBUTIONS FOR CONTINUOUS RANDOM VARIABLES

DEFINITION

A **probability distribution for a continuous random variable *X*** is specified by a mathematical function denoted by $f(x)$ and called the **density function**. The graph of a density function is a smooth curve (the **density curve**). The following requirements must be met:

1. $f(x) \ge 0$ (so that the curve cannot dip below the horizontal axis).
2. The total area under the density curve is equal to 1. The probability that X falls in any particular interval is the area under the density curve and above the interval.

2.7 JOINT PROBABILITY

Joint Probability Function:

1-Discrete Random Variables: If (X,Y) is a 2-dimensional discrete random variable such that

$$P(X = x_i, Y = y_j) = p_{ij}$$

Then p_{ij} is called the joint probability mass function provided the following conditions satisfied

1. $p_{ij} \ge 0$
2. $\sum_j \sum_i p_{ij} = 1$

2-Continuous Random Variable: If X and Y are continuous random variables with $f(x,y)$ is called the joint probability density function provided the following conditions are satisfied

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

Marginal Distribution Function

1-Discrete Random Variable: The collection of pairs $\{x_i, p_{i*}\}$ where i takes the value 1, 2, 3, … etc. is called the Marginal distribution of X. Similarly the collection of pairs $\{y_j, p_{*j}\}$ where j takes the value 1, 2, 3, … etc. is called the Marginal distribution of Y.

2-Continuous Random Variable: If the joint distribution of the random variables X, Y is F[x, y], then

the marginal distribution of X is $F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(x,y)dydx$

the marginal distribution of Y is $F_Y(y) = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f(x,y)dxdy$

Conditional Probability Density Function

Discrete Random Variable

Conditional Probability of X given $Y = y_j$ is given by

$$P\left(X = x_i \middle/ Y = y_j\right) = \frac{p(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{*j}}$$

Conditional Probability of Y given $X = x_i$ is given by

$$P\left(Y = y_j \middle/ X = x_i\right) = \frac{p(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i*}}$$

Continuous Random Variable

The conditional probability function of X given Y

$$f\left(\frac{x}{y}\right) = \frac{f(x,y)}{f(y)}$$

The conditional probability function of Y given X

$$f\left(\frac{y}{x}\right) = \frac{f(x,y)}{f(x)}$$

2.8 EXPECTATION:

Let *X* be a discrete random variable, taking on distinct values $x_1, x_2, \ldots x_n$ with $p_i = P(X = x_i)$. Then the *expected value* of *X* is given by-

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

Let *X* be an absolutely continuous random variable, with density function f_X. Then the *expected value* of *X* is given by-

$$E(X) = \int_{-\infty}^{\infty} x f_X \, dx$$

2.9 VARIANCE:

The *variance* of a random variable *X* is the quantity-

$$\sigma_x^2 = \mathrm{Var}(X) = E\left((X - \mu_X)^2\right),$$

where $\mu_X = E(X)$ is the mean of X.

The *standard deviation* of a random variable X is the quantity

$$\sigma_X = \mathrm{Sd}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{E((X - \mu_X)^2)}.$$

The *covariance* of two random variables X and Y is given by

$$\mathrm{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right),$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

Let X and Y be two random variables. Then

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

If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

(a) For any random variables X and Y,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).$$

(b) More generally, for any random variables $X_1, \ldots, X_n$,

$$\text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).$$

(a) If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
(b) If $X_1, \ldots, X_n$ are independent, then $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$.

Conditional Expectation: Let f(x,y) be the joint probability density function of X,Y and u(Y) is a function of y. Then the conditional expectation of u(Y), given X = x is

$$E[u(y)/x] = \int_{-\infty}^{\infty} u(y) f(y/x)\,dy$$

In particular,

Conditional Mean,

$$E[Y/x] = \int_{-\infty}^{\infty} y f(y/x)\,dy$$

Conditional Variance,

$$E\left[[Y - E(Y/x)]^2/x\right] = E[Y^2/x] - E[Y/x]^2$$

PROBLEM IV: Find the marginal distributions given that the joint distribution of X and Y is,

$$f(x,y)=\frac{x+y}{21}$$

where X = 1, 2, 3; $Y = 1, 2$. Also determine the conditional mean and variance of X, given Y=1.

Solution:

X \ Y	1	2	
1	2/21	3/21	5/21
2	3/21	4/21	7/21
3	4/21	5/21	9/21
	9/21	12/21	1

- Marginal Distribution of X

$$P(X=x_i)=p_{i*}=\sum_j p_{ij}$$

X	1	2	3
p_{i*}	5/21	7/21	9/21

- Marginal Distribution of Y

$$P(Y=y_j)=p_{*j}=\sum_i p_{ij}$$

Y	1	2
p_{*j}	9/21	12/21

X	$f\left(\frac{y}{x}\right)=\frac{f(x,y)}{f(x)}$
1	$\frac{2/21}{5/21}=\frac{2}{5}$
2	$\frac{3/21}{5/21}=\frac{3}{5}$

- Conditional mean

$$E[Y/x]=\sum_{y=1}^{2} yf(y/x)=1\times f(1/1)+2\times f(2/1)=7/5$$

$$E[Y^2/x]=\sum_{y=1}^{2} y^2 f(y/x)=1^2\times f(1/1)+2^2\times f(2/1)=13/5$$

- Conditional Variance

$$E[Y^2/x]-E[y/x]^2=\frac{13}{5}-\frac{49}{25}=\frac{16}{25}$$

PROBLEM V: Joint probability mass function of X,Y is given by

$P(x,y)=k(2x+3y)$; $x=0,1,2$; $y=1,2,3$.

Find all the marginal and conditional probability and also find probability distribution of X+Y.

X \ Y	1	2	3	
0	3 k	6 k	9 k	18k
1	5 k	8 k	11 k	24k
2	7 k	10 k	13 k	30k
	15k	24k	33k	72k

Solution: Since $\sum_j \sum_i p_{ij}=1$, k = 1/72

- Marginal Distribution of X $P(X = x_i) = p_{i*} = \sum_j p_{ij}$

X	0	1	2
p_{i*}	18/72	24/72	30/72

- Marginal Distribution of Y $P(Y = y_j) = p_{*j} = \sum_i p_{ij}$

Y	1	2	3
p_{*j}	15/72	24/72	33/72

Conditional probability distribution of X given Y = y_j

$$P\left(X = x_i / Y = y_j\right) = \frac{p(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p*j}$$

X	$P(X \le x_i/Y = 1)$	$P(X \le x_i/Y = 2)$	$P(X \le x_i/Y = 3)$
0	$\frac{3/72}{15/72} = \frac{3}{15}$	$\frac{6/72}{24/72} = \frac{6}{24}$	$\frac{9/72}{33/72} = \frac{9}{33}$
1	$\frac{5/72}{15/72} = \frac{5}{15}$	$\frac{8/72}{24/72} = \frac{8}{24}$	$\frac{11/72}{33/72} = \frac{11}{33}$
2	$\frac{7/72}{15/72} = \frac{7}{15}$	$\frac{10/72}{24/72} = \frac{10}{24}$	$\frac{13/72}{33/72} = \frac{13}{33}$

Conditional probability distribution of Y given X = x_i

$$P\left(Y = y_j \middle/ X = x_i\right) = \frac{p(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i*}}$$

Y	$P(Y \le x_i/X = 0)$	$P(Y \le x_i/X = 1)$	$P(Y \le x_i/X = 2)$
1	$\frac{3/72}{18/72} = \frac{3}{18}$	$\frac{5/72}{24/72} = \frac{5}{24}$	$\frac{7/72}{30/72} = \frac{7}{30}$
2	$\frac{6/72}{18/72} = \frac{6}{18}$	$\frac{8/72}{24/72} = \frac{8}{24}$	$\frac{10/72}{30/72} = \frac{10}{30}$
3	$\frac{9/72}{18/72} = \frac{9}{18}$	$\frac{11/72}{24/72} = \frac{11}{24}$	$\frac{13/72}{30/72} = \frac{13}{30}$

- Probability distribution of X + Y

X + Y	Probabilities
1	3/72
2	11/72
3	24/72
4	21/72
5	13/72

PROBLEM VI:

Given joint probability density function

$$f(x,y)=\begin{cases} e^{-y}, & x>0, y>x \\ 0, & otherwise \end{cases}$$

1. Find $P\left(X>1/Y<5\right)$
2. Find the marginal density function of X and Y.

Solution: 1.

$$P\left(X>1/Y<5\right)=\frac{P(X>1,Y<5)}{P(Y<5)}$$

$$P(X>1,Y<5)=\int_1^5\int_1^y e^{-y}\,dx\,dy$$

$$=\int_1^5 e^{-y}\left[x\right]_1^y dy$$

$$=\int_1^5 [1-y]e^{-y}\,dy$$

$$=-5e^{-5}+e^{-1}$$

0<1<x<y<5

$$P(Y<5)=\int_0^5\int_0^y e^{-y}\,dx\,dy$$

$$=\int_0^5\left[e^{-y}x\right]_0^y dy$$

$$=\int_0^5 e^{-y}y\,dy$$

$$=-6e^{-5}+1$$

0<x<y<5

$$\therefore\ P\left(X>1/Y<5\right)=\frac{P(X>1,Y<5)}{P(Y<5)}=\frac{e^4-5}{e^5-6}$$

2. Marginal density function of X

$$f_X(x) = \int_x^\infty e^{-y} dy = e^{-x}$$

Marginal density function of Y

$$f_Y(y) = \int_0^y e^{-y} dx = ye^{-y}$$

PROBLEM VII: Let X be the random variable that denotes the life in hours of a certain electronic device. The probability density function is-

$$f(x) = \begin{cases} \frac{20{,}000}{x^3}, & x > 100 \\ 0, & eleswhere \end{cases}$$

Find the expected life of this type of device.

Solution: We have,

$$\mu = E[X] = \int_{100}^{x} x \frac{20{,}000}{x^3} dx = \int_{100}^{x} \frac{20000}{x^2} dx = 200$$

Therefore, we can expect this type of device to last, on average 200 hours.

PROBLEM VIII:

Suppose that the number of cars X that pass through a carwash between 4.00p.m and 9.00p.m.on any day has the following distribution:

x	4	5	6	7	8	9
P(X = x)	1/12	1/12	1/4	1/4	1/6	1/6

Let g(X)=2X–1 represent the amount of money in rupees, paid to the attendant by the manager. Find the attendant's expected earnings for this particular time-period.

Solution: The attendant can expect to receive-

$$E[g(X)] = E[2X - 1]$$

$$= \sum_{x=4}^{9} (2x-1)\,p(x)$$
$$= 7\left(\frac{1}{12}\right)+9\left(\frac{1}{12}\right)+11\left(\frac{1}{4}\right)+13\left(\frac{1}{4}\right)+15\left(\frac{1}{6}\right)+17\left(\frac{1}{6}\right)$$
$$= 12.67$$

PROBLEM IX:

The weekly demand for a drinking-water product, in thousands of liters, from a local chain of efficiency stores is a continuous random variable *X* having the probability density-

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

Find the mean and variance of *X*.

Solution: Calculating *E(X) and E(X²), we have*

$$\mu = E[X] = 2\int_1^2 x(x-1)\,dx = \frac{5}{3}$$

$$E[X^2] = 2\int_1^2 x^2(x-1)\,dx = \frac{17}{6}$$

$$\therefore Var[X] = \frac{17}{6} - \left[\frac{5}{3}\right]^2 = \frac{1}{18}$$

2.10 MOMENTS:

Statistical Moments plays a crucial role while we specify our probability distribution to work with since, with the help of moments, we can describe the properties of statistical distribution. Therefore, they are helpful to describe the distribution.

In Statistical Estimation and Testing of Hypothesis, which all are based on the numerical values arrived for each distribution, we required the statistical moments.

Moment word is very popular in mechanical sciences. In science moment is a measure of energy which generates the frequency. In Statistics, moments are the arithmetic means of first, second, third and so on, i.e. r^{th} power of the deviation taken from either mean or an arbitrary point of a distribution. In other words, moments are statistical measures that give certain characteristics of the distribution. In statistics, some moments are very important. Generally, in any frequency distribution, four moments are obtained which are known as first, second, third and fourth moments. These four moments describe the information about mean, variance, skewness and kurtosis of a frequency distribution. Calculation of moments gives some features of a distribution which are of statistical importance. Moments can be classified in raw and central moment. Raw moments are measured about any arbitrary point A (say). If A is taken to be zero then raw moments

are called moments about origin. When A is taken to be Arithmetic mean we get central moments. The first raw moment about origin is mean whereas the first central moment is zero. The second raw and central moments are mean square deviation and variance, respectively. The third and fourth moments are useful in measuring skewness and kurtosis.
Three types of moments are:
1. Moments about arbitrary point,
2. Moments about mean, and
3. Moments about origin

2.10.1 Moments about Arbitrary Point

When actual mean is in fraction, moments are first calculated about an arbitrary point and then converted to moments about the actual mean. When deviations are taken from arbitrary point, the formulas are:

For Ungrouped Data

If $x_1, x_2, \ldots x_n$ are the n observations of a variable X, then their moments about an arbitrary point A are-

Zero order moment A $$\mu_0' = \frac{\sum_{i=1}^{n}(x_i - A)^0}{n} = 1$$

First order moment $$\mu_1' = \frac{\sum_{i=1}^{n}(x_i - A)^1}{n}$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{n}(x_i - A)^2}{n}$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{n}(x_i - A)^3}{n}$$

Fourth order moment $$\mu_4' = \frac{\sum_{i=1}^{n}(x_i - A)^4}{n}$$

In general, the r^{th} order moment about arbitrary point A is given by-

$$\mu_r' = \frac{\sum_{i=1}^{n}(x_i - A)^r}{n}; \; for \; r = 1, 2 \ldots$$

For Grouped Data

If $x_1, x_2, \ldots x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, \ldots f_k$, then moments about an arbitrary point A are-

Zero order moment A $$\mu_0' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^0}{N} = 1, N = \sum_{i=1}^{k} f_i$$

First order moment $$\mu_1' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^1}{N}$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^2}{N}$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^3}{N}$$

Fourth order moment $$\mu_4' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^4}{N}$$

In general, the r^{th} order moment about arbitrary point A is given by-

$$\mu_r' = \frac{\sum_{i=1}^{k} f_i (x_i - A)^r}{N} \quad r = 1, 2 \ldots$$

2.10.2 Moments about Origin

In case, when we take an arbitrary point A = 0 then, we get the moments about origin.

For Ungrouped Data

First order moment $$\mu_1' = \frac{\sum_{i=1}^{n} (x_i - 0)^1}{n} = \bar{x}$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{n} (x_i - 0)^2}{n}$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{n} (x_i - 0)^3}{n}$$

Fourth order moment $$\mu_4' = \frac{\sum_{i=1}^{n}(x_i - 0)^4}{n}$$

In general, the r^{th} order moment about origin is given by-

$$\mu_r' = \frac{\sum_{i=1}^{n}(x_i - 0)^r}{n} \quad r = 1, 2...$$

For Grouped Data

r^{th} order moment $$\mu_r' = \frac{\sum_{i=1}^{k} f_i(x_i - 0)^r}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^r$$

First order moment $$\mu_1' = \frac{\sum_{i=1}^{k} f_i(x_i - 0)^1}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{k} f_i(x_i - 0)^2}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{k} f_i(x_i - 0)^3}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^3$$

Fourth order moment $$\mu_4' = \frac{\sum_{i=1}^{k} f_i(x_i - 0)^4}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^4$$

2.10.3 Moments about Mean

When we take the deviation from the actual mean and calculate the moments, these are known as moments about mean or central moments. The formulae are:

For Ungrouped Data

Zero order moment $$\mu_0 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^0}{n} = 1$$

First order moment $$\mu_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^1}{n} = 0$$

Thus, first order moment about mean is zero, because the algebraic sum of the deviation from the mean is zero $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Second order moment $$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \sigma^2$$

Therefore, second order moment about mean is variance.

Third order moment $$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}$$

For Grouped Data

In case of frequency distribution, the r^{th} order moment about mean is given by:

$$\mu_r = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^r}{N}, \text{ for } r = 0,1,2,...$$

Zero order moment $$\mu_0 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^0}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i = 1$$

First order moment $$\mu_1 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^1}{N} = 0$$

Because $$\sum_{i=1}^{n} f_i(x_i - \bar{x}) = 0$$

Second order moment $$\mu_2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{N} = \sigma^2$$

Third order moment $$\mu_3 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^3}{N}$$

Fourth order moment $$\mu_4 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^4}{N}$$

2.11 MOMENT GENERATING FUNCTIONS:

*The **moment generating function (mgf)** of a random variable X*

More explicitly, the mgf of X can be written as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \quad \text{if } X \text{ is continuous,}$$

$$M_X(t) = \sum_{x \in \mathcal{X}} e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

Theorem 2.11.1:

If X has mgf $M_X(t)$, then

$$\mathrm{E}(X^n) = M_X^{(n)}(0),$$

where

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_0.$$

That is, the n-th moment is equal to the n-th derivative of the mgf evaluated at $t = 0$.

Theorem 2.11.2

Suppose that the random variable Y has the mgf $m_Y(t)$.

Then mgf of the random variable $W = aY + b$, where a and b are constants, is given by

$$m_W(t) = e^{tb} m_Y(at).$$

Theorem 2.11.3

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with mgfs $m_{Y_1}(t), m_{Y_2}(t), \ldots, m_{Y_n}(t)$. Then the mgf of their sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is given by

$$m_Y(t) = m_{Y_1}(t) \times m_{Y_2}(t) \times \cdots \times m_{Y_n}(t).$$

Example 1: For the following distribution calculate first four moments about mean and also find β_1, β_2, γ_1 and γ_2:

Marks	5	10	15	20	25	30	35
Frequency	4	10	20	36	16	12	2

Solution: First we construct the following frequency distribution for calculation of moments:

Marks (x)	**f**	$\mathbf{d = \frac{(x-20)}{5}}$	**fd**	$\mathbf{fd^2}$	$\mathbf{fd^3}$	$\mathbf{fd^4}$
5	4	−3	−12	36	−108	324
10	10	−2	−20	40	−80	160
15	20	−1	−20	20	−20	20
20	36	0	0	0	0	0
25	16	1	16	16	16	16
30	12	2	24	48	96	192
35	2	3	6	18	54	162
			$\sum fd = -6$	$\sum fd^2 = 178$	$\sum fd^3 = -42$	$\sum fd^4 = 874$

$$\mu_1' = \frac{\sum fd}{N} \times h = \frac{-6}{100} \times 5 = -0.3$$

$$\mu_2' = \frac{\sum fd^2}{N} \times h^2 = \frac{178}{100} \times 25 = 44.5$$

$$\mu_3' = \frac{\sum fd^3}{N} \times h^3 = \frac{-42}{100} \times 125 = -52.5$$

$$\mu_4' = \frac{\sum fd^4}{N} \times h^4 = \frac{874}{100} \times 625 = 5462.5$$

Moments about mean

$$\mu_2 = \mu_2' - \mu_1'^2 = 44.5 - 0.09 = 44.41 = \sigma^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + \mu_1'^3$$

$$= -52.5 - 3 \times 44.5 \times - 0.3 + 2(-0.3)^3$$

$$= -52.5 + 40.05 - 0.054 = -12.504$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 5462.5 - 4\,(-0.3 \times -52.5) + 6\,(44.5)\,(-0.3)^2 - 3\,(-0.3)^4$$

$$= 5462.5 - 63 + 24.03 - 0.0243$$

$$= 5423.5057$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\left(-12.504\right)^2}{\left(44.41\right)^3} = 0.001785$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-12.504}{\left(6.6641\right)^3} = -0.0422$$

$$\beta_2 = \frac{\mu_4}{\sigma_2^2} = \frac{5423.5057}{\left(44.41\right)^2} = 02.7499$$

$$\gamma_2 = \beta_2 - 3 = 2.7499 - 3 = -0.2501$$

2.12 LAW OF LARGE NUMBERS

(*Weak law of large numbers*) Let $X_1, X_2, \ldots$ be a sequence of independent random variables, each having the same mean μ and each having variance less than or equal to $\upsilon < \infty$. Then for all $\epsilon > 0$, $\lim_{n\to\infty} P(|M_n - \mu| \geq \epsilon) = 0$. That is, the averages converge in probability to the common mean μ or $M_n \xrightarrow{P} \mu$.

Consider flipping a sequence of identical fair coins. Let M_n be the fraction of the first n coins that are heads. Then $M_n = (X_1 + \cdots + X_n)/n$, where $X_i = 1$ if the ith coin is heads, otherwise $X_i = 0$. Hence, by the weak law of large numbers, we have

$$\begin{aligned}\lim_{n\to\infty} P(M_n < 0.49) &= \lim_{n\to\infty} P(M_n - 0.5 < -0.01) \\ &\leq \lim_{n\to\infty} P(M_n - 0.5 < -0.01 \text{ or } M_n - 0.5 > 0.01\ ) \\ &= \lim_{n\to\infty} P(|M_n - 0.5| > 0.01) = 0\end{aligned}$$

and, similarly, $\lim_{n\to\infty} P(M_n > 0.51) = 0$. This illustrates that for large n, it is very likely that M_n is very close to 0.5. ∎

(*Strong law of large numbers*) Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having finite mean μ. Then

$$P\left(\lim_{n\to\infty} M_n = \mu\right) = 1.$$

That is, the averages converge with probability 1 to the common mean μ or $M_n \xrightarrow{a.s.} \mu$.

NOTE:

1) i.i.d = Identically and independently distributed
2) Converges in Probability

Let $X_1, X_2, \ldots$ be an infinite sequence of random variables, and let Y be another random variable. Then the sequence $\{X_n\}$ *converges in probability* to Y, if for all $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - Y| \geq \epsilon) = 0$, and we write $X_n \xrightarrow{P} Y$.

3) Converges with Probability 1

Let $X_1, X_2, \ldots$ be an infinite sequence of random variables. We shall say that the sequence $\{X_i\}$ *converges with probability 1* (or *converges almost surely (a.s.)*) to a random variable Y, if $P(\lim_{n\to\infty} X_n = Y) = 1$ and we write $X_n \xrightarrow{a.s.} Y$.

2.13 MONTE-CARLO APPROXIMATIONS

The laws of large numbers say that if $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables with mean μ, and

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

then for large n we will have $M_n \approx \mu$.

Suppose now that μ is *unknown*. Then, it is possible to change perspective and use M_n (for large *n*) as an *estimator* or approximation of μ. Any time we approximate or estimate a quantity, we must also say something about how much error is in the estimate. Of course, we cannot say what this error is exactly, as that would require knowing the exact value of μ. However, we showed how the central limit theorem leads to a very natural approach to assessing this error, using three times the standard error of the estimate i.e. for large *n*,

$$\left(M_n - 3\sigma_n/\sqrt{n},\, M_n + 3\sigma_n/\sqrt{n}\right) = \left(M_n - 3S/\sqrt{n},\, M_n + 3S/\sqrt{n}\right)$$

contains the true value of μ with virtual certainty (also with probability about 0.9974). Therefore, the half-length is a measure of the accuracy of the estimate M_n — notice that this can be computed from the values $X_1, \ldots, X_n$. The quantity $S/\sqrt{n}$ is referred to as the *standard error* of the estimate M_n.

Note: (*The central limit theorem*) Let $X_1, X_2, \ldots$ be i.i.d. with finite mean μ and finite variance σ^2. Let $Z \sim N(0, 1)$. Then as $n \to \infty$, the sequence $\{Z_n\}$ converges in distribution to Z, i.e., $Z_n \xrightarrow{D} Z$.

Example 1:

Consider flipping a sequence of identical coins, each of which has probability θ of coming up heads, but where θ is unknown. Let M_n again be the fraction of the first n coins that are heads. Then we know that for large n, it is very likely that M_n is very close to θ. Hence, we can use M_n to *estimate* θ.

Example 2:

Suppose we believe a certain medicine lowers blood pressure, but we do not know by how much. We would like to know the mean amount μ, by which this medicine lowers blood pressure.

Suppose we observe n patients (chosen at random so they are i.i.d.), where patient i has blood pressure B_i before taking the medicine and blood pressure A_i afterwards. Let $X_i = B_i - A_i$. Then

$$M_n = \frac{1}{n}\sum_{i=1}^{n}(B_i - A_i)$$

is the average amount of blood pressure decrease. (Note that $B_i - A_i$ may be negative for some patients, and it is important to also include those negative terms in the sum.)

Then for large n, the value of M_n is a good *estimate* of $E(X_i) = \mu$.

What Are Monte Carlo Methods?

Monte Carlo methods, or MC for short, are a class of techniques for randomly sampling a probability distribution.

There are three main reasons to use Monte Carlo methods to randomly sample a probability distribution; they are:

- **Estimate density**, gather samples to approximate the distribution of a target function.
- **Approximate a quantity**, such as the mean or variance of a distribution.
- **Optimize a function**, locate a sample that maximizes or minimizes the target function.

Monte Carlo methods are named for the casino in Monaco and were first developed to solve problems in particle physics at around the time of the development of the first computers and the Manhattan project for developing the first atomic bomb.

"This is called a Monte Carlo approximation, named after a city in Europe known for its plush gambling casinos. Monte Carlo techniques were first developed in the area of statistical physics – in particular, during development of the atomic bomb – but are now widely used in statistics and machine learning as well"- quoted from the book Machine Learning: A Probabilistic Perspective, 2012.

The concept was invented by Stanislaw Ulam, a mathematician who devised these methods as part of his contribution to the Manhattan Project. He used the tools of random sampling and inferential statistics to model likelihoods of outcomes, originally applied to a card game (Monte Carlo Solitaire). Ulam later worked with collaborator John von Neumann, using newly developed computer technologies to run simulations to better understand the risks associated with the nuclear project. As you can imagine, modern computational technology allows us to model much more complex systems, with a larger number of random parameters, like so many of the scenarios that we encounter during our everyday lives.

UNIT-III

3.1 SAMPLING

Many studies are conducted in order to generalize from a sample to the corresponding population. As a result, it is important that the sample be representative of the population. To be reasonably sure of this, the researcher must carefully consider the way in which the sample is selected. It is sometimes tempting to take the easy way out and gather data in a haphazard way; but if a sample is chosen on the basis of convenience alone, it becomes impossible to interpret the resulting data with confidence.

For example, it might be easy to use the students in your statistics class as a sample of students at your university. However, not all majors include a statistics course in their curriculum, and most students take statistics in their sophomore or junior year. The difficulty is that it is not clear whether or how these factors (and others that we might not be aware of) affect inferences based on information from such a sample.

There are many reasons for selecting a sample rather than obtaining information from an entire population (a census). Sometimes the process of measuring the characteristics of interest is destructive, as with measuring the lifetime of flashlight batteries or the sugar content of oranges, and it would be foolish to study the entire population. But the most common reason for selecting a sample is limited resources; restrictions on available time or money usually prohibit observation of an entire population.

3.2 BIAS IN SAMPLING

Bias in sampling is the tendency for samples to differ from the corresponding population in some systematic way. Bias can result from the way in which the sample is selected or from the way in which information is obtained once the sample has been chosen. The most common types of bias encountered in sampling situations are selection bias, measurement or response bias, and nonresponse bias.

Selection bias (sometimes also called under coverage) is introduced when the way the sample is selected systematically excludes some part of the population of interest.

For example, a researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may exclude the homeless or those without telephones. If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample is virtually guaranteed to be unrepresentative of the population. If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid for the population of interest. Selection bias also occurs if only volunteers or self-selected individuals are used in a study, because self-selected individuals (e.g., those who choose to participate in a call-in telephone poll) may well differ from those who choose not to participate.

Measurement or response bias occurs when the method of observation tends to produce values that systematically differ from the true value in some way. This might happen if an improperly calibrated scale is used to weigh items or if questions on a survey are worded in a way that tends to influence the response. For example, a Gallup survey sponsored by the

American Paper Institute (Wall Street Journal, May 17, 1994) included the following question: "It is estimated that disposable diapers account for less than 2 percent of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard waste are estimated to account for about 21 percent of trash in landfills. Given this, in your opinion, would it be fair to tax or ban disposable diapers?" It is likely that the wording of this question prompted people to respond in a particular way.

Other things that might contribute to response bias are the appearance or behaviour of the person asking the question, the group or organization conducting the study, and the tendency for people not to be completely honest when asked about illegal behaviour or unpopular beliefs. Although the terms measurement bias and response bias are often used interchangeably, the term measurement bias is usually used to describe systematic deviation from the true value as a result of a faulty measurement instrument (as with the improperly calibrated scale).

Nonresponse bias occurs when responses are not obtained from all individuals selected for inclusion in the sample. As with selection bias, nonresponse bias can distort results if those who respond differ in important ways from those who do not respond. Although some level of nonresponse is unavoidable in most surveys, the biasing effect on the resulting sample is lowest when the response rate is high. To minimize nonresponse bias, it is critical that a serious effort be made to follow up with individuals who do not respond to an initial request for information.

The components that are necessary for a sample to be effective, the following definitions are required:

Target population: The complete collection of objects whose description is the major goal of the study. Designating the target population is a crucial but often difficult part of the first step in an observational or experimental study. For example, in a survey to decide if a new storm-water drainage tax should be implemented, should the target population be all persons over the age of 18 in the county, all registered voters, or all persons paying property taxes? The selection of the target population may have a profound effect on the results of the study.

Sample: A subset of the target population. Sampled population: The complete collection of objects that have the potential of being selected in the sample; the population from which the sample is actually selected. In many studies, the sampled population and the target population are very different. This may lead to very erroneous conclusions based on the information collected in the sample. For example, in a telephone survey of people who are on the property tax list (the target population), a subset of this population may not answer their telephone if the caller is unknown, as viewed through caller ID. Thus, the sampled population may be quite different from the target population with respect to some important characteristics such as income and opinion on certain issues.

Observation unit: The object upon which data are collected. In studies involving human populations, the observation unit is a specific individual in the sampled population. In ecological studies, the observation unit may be a sample of water from a stream or an individual plant on a plot of land.

Sampling unit: The object that is actually sampled. We may want to sample the person who pays the property tax but may only have a list of telephone numbers. Thus, the households in the sampled population serve as the sampled units, and the observation units are the individuals residing in the sampled household. In an entomology study, we may sample 1-acre plots of land and then count the number of insects on individual plants residing on the sampled plot. The sampled unit is the plot of land, the observation unit would be the individual plants.

Sampling frame: The list of sampling units. For a mailed survey, it may be a list of addresses of households in a city. For an ecological study, it may be a map of areas downstream from power plants.

3.3 SAMPLING TECHNIQUES/DESIGNS:

Simple Random Sampling: The most straightforward sampling method is called simple random sampling. A simple random sample is a sample chosen using a method that ensures that each different possible sample of the desired size has an equal chance of being the one chosen. For example, suppose that we want a simple random sample of 10 employees chosen from all those who work at a large design firm. For the sample to be a simple random sample, each of the many different subsets of 10 employees must be equally likely to be the one selected. A sample taken from only full-time employees would not be a simple random sample of all employees, because someone who works part-time has no chance of being selected. Although a simple random sample may, by chance, include only full-time employees, it must be selected in such a way that each possible sample, and therefore every employee, has the same chance of inclusion in the sample.
It is the selection process, not the final sample, which determines whether the sample is a simple random sample. The letter n is used to denote sample size; it is the number of individuals or objects in the sample.

DEFINITION

A **simple random sample of size *n*** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

Selecting a Simple Random Sample: A number of different methods can be used to select a simple random sample. One way is to put the name or number of each member of the population on different but identical slips of paper. The process of thoroughly mixing the slips and then selecting n slips one by one yields a random sample of size n.
This method is easy to understand, but it has obvious drawbacks. The mixing must be adequate, and producing the necessary slips of paper can be extremely tedious, even for populations of moderate size.
A commonly used method for selecting a random sample is to first create a list, called a sampling frame, of the objects or individuals in the population. Each item on the list can then

be identified by a number, and a table of random digits or a random number generator can be used to select the sample. A random number generator is a procedure that produces a sequence of numbers that satisfies properties associated with the notion of randomness. Most statistics software packages include a random number generator, as do many calculators.

When selecting a random sample, researchers can choose to do the sampling with or without replacement. Sampling with replacement means that after each successive item is selected for the sample, the item is "replaced" back into the population and may therefore be selected again at a later stage. In practice, sampling with replacement is rarely used. Instead, the more common method is to not allow the same item to be included in the sample more than once. After being included in the sample, an individual or object would not be considered for further selection. Sampling in this manner is called sampling without replacement.

Sampling without Replacement

Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n distinct individuals from the population.

Sampling with Replacement

After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Although these two forms of sampling are different, when the sample size n is small relative to the population size, as is often the case, there is little practical difference between them. In practice, the two can be viewed as equivalent if the sample size is at most 5% of the population size. (A less conservative approach that allows the sample size to be up to 10% of the population size is used in some texts.)

Stratified Random Sampling: When the entire population can be divided into a set of nonoverlapping subgroups, a method known as stratified sampling often proves easier to implement and more cost-effective than simple random sampling. In stratified random sampling, separate simple random samples are independently selected from each subgroup. For example, to estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular metropolitan area as being made up of four subpopulations: (1) surgeons, (2) internists and family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization. Rather than taking a random simple sample from the population of all doctors, the researcher could take four separate simple random samples—one from the group of surgeons, another from the internists and family practitioners, and so on. These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

When the population is divided in this way, the subgroups are called strata and each individual subgroup is called a stratum (the singular of strata). Stratified sampling entails selecting a separate simple random sample from each stratum. Stratified sampling can be used instead of simple random sampling if it is important to obtain information about characteristics of the individual strata as well as of the entire population, although a stratified sample is not required to do this—subgroup estimates can also be obtained by using an appropriate subset of data from a simple random sample.
The real advantage of stratified sampling is that it often allows us to make more accurate inferences about a population than does simple random sampling. In general, it is much easier to produce relatively accurate estimates of characteristics of a homogeneous group than of a heterogeneous group.

Cluster Sampling: Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves. Cluster sampling involves dividing the population of interest into nonoverlapping subgroups, called clusters. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample. For example, suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom. There are 24 senior homerooms, each with approximately 25 students. If school administrators wanted to select a sample of roughly 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample. In this way, an evaluation survey could be administered to all students in the selected homerooms at the same time—certainly easier logistically than randomly selecting 75 students and then administering the survey to the individual seniors. Because whole clusters are selected, the ideal situation occurs when each cluster mirrors the characteristics of the population. When this is the case, a small number of clusters results in a sample that is representative of the population. If it is not reasonable to think that the variability present in the population is reflected in each cluster, as is often the case when the cluster sizes are small, then it becomes important to ensure that a large number of clusters are included in the sample.
Be careful not to confuse clustering and stratification. Even though both of these sampling strategies involve dividing the population into subgroups, both the way in which the subgroups are sampled and the optimal strategy for creating the subgroups are different. In stratified sampling, we sample from every stratum, whereas in cluster sampling, we include only selected whole clusters in the sample. Because of this difference, to increase the chance of obtaining a sample that is representative of the population, we want to create homogeneous (similar) groups for strata and heterogeneous (reflecting the variability in the population) groups for clusters.

Systematic Sampling: Systematic sampling is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value k is specified (e.g., $k = 50$ or $k = 200$). Then one of the first k individuals is selected at random, after which every kth individual in the sequence is included in the sample. A sample selected in this way is called a 1 in k systematic sample.

For example, a sample of faculty members at a university might be selected from the faculty phone directory. One of the first k=20 faculty members listed could be selected at random, and then every 20th faculty member after that on the list would also be included in the sample. This would result in a 1 in 20 systematic sample. The value of k for a 1 in k systematic sample is generally chosen to achieve a desired sample size. For example, in the faculty directory scenario just described, if there were 900 faculty members at the university, the 1 in 20 systematic sample described would result in a sample size of 45. If a sample size of 100 was desired, a 1 in 9 systematic sample could be used.
As long as there are no repeating patterns in the population list, systematic sampling works reasonably well. However, if there are such patterns, systematic sampling can result in an unrepresentative sample.

Purposive Sampling: The term 'purposive sampling' has been used in several slightly different senses in connection with subjective method of sampling. In most general sense, it means selecting individuals according to some purposive principle. For example, an observer who wishes to take a sample of oranges from a lot runs his eyes over the whole lot and then chooses average oranges-average in size, shape, weight or whatever other quality he may consider important. It has been claimed that the purposive method is more likely to give a typical or representative sample. But it may be pointed out that the method in most cases may involve some bias of unknown magnitude. Moreover, the method cannot provide an estimate of the error involved. Also, the method, although it may tell more about the mean of the population, would probably give a wrong idea about the variability since the observer has deliberately chosen values near the mean.

ANALYSIS OF VARIANCE

The total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called the *analysis of variance* (ANOVA). With the help of the technique of analysis of variance, it will be possible for us to perform certain tests of hypotheses and to provide estimates for components of variation.

Linear Model:

Let $y_1, y_2, \dots, y_n$ be n observable quantities. In all cases, we shall assume the observed value to be composed of two parts:

$$y_i = \mu_i + e_i \qquad (3.1)$$

where, μ_i is the *true value* and e_i the *error*. The true value μ_i which is due to assignable causes and the portion that remains is error, which is due to various chance causes. The true value μ_i is again assumed to be a linear function of k unknown quantities $\beta_1, \beta_2, \dots, \beta_k$, called effects. Thus-

$$\mu_i = a_{i1}\beta_1 + a_{i2}\beta_2 + \cdots + a_{ik}\beta_k \qquad (3.2)$$

where a_{ij} are known, each being usually taken to be 0 or 1.
This set-up, which is fundamental to analysis of variance, is called the linear model.

One-Way ANOVA: Assumptions

For the results of a one-way ANOVA to be valid, the following assumptions should be met:

1. Normality – Each sample was drawn from a normally distributed population.

2. Equal Variances – The variances of the populations that the samples come from are equal.

3. Independence – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.

Note:

1. Errors e_i are always independent random variables.
2. Errors are also assumed to have zero expectations and to be homoscedastic.
3. A model in which all the effects β_j are unknown constants or parameters, is called a fixed-effects model.
4. A model in which all the effects β_j are random variables, is called a random-effects model.
5. A model in which at least one β_j is a random variable and at least one β_j is a constant is called mixed model.

One-Way ANOVA: The Process

A one-way ANOVA uses the following null and alternative hypotheses:

- **H_0 (null hypothesis):** $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (all the population means are equal)
- **H_1 (alternative hypothesis):** at least one population mean is different from the rest.

NOTATIONS:

Number of samples (or levels)	$= k$
Number of observations in ith sample	$= n_i, \quad i = 1, 2, \ldots, k$
Total number of observations	$= n = \sum_i n_i$
Observation j in ith sample	$= x_{ij}, \quad j = 1, 2, \ldots, n_i$
Sum of n_i observations in ith sample	$= T_i = \sum_j x_{ij}$
Sum of all n observations	$= T = \sum_i T_i = \sum_i \sum_j x_{ij}$

Computational Formulae:

Total sum of squares,	$SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{n}$
Between samples sum of squares,	$SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}$
Within samples sum of squares,	$SS_W = SS_T - SS_B$

Total mean square,	$MS_T = \frac{SS_T}{n-1}$
Between samples mean square,	$MS_B = \frac{SS_B}{k-1}$
Within samples mean square,	$MS_W = \frac{SS_W}{n-k}$

ANOVA ONE-WAY TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ratio
Between samples	SS_B	$k-1$	MS_B	$\frac{MS_B}{MS_W}$
Within samples	SS_W	$n-k$	MS_W	
Total	SS_T	$n-1$		

NOTE: FOLLOW CLASSNOTES FOR PROBLEM ON ANOVA

UNIT-IV

Statistical inference is the process by which we infer population properties from sample properties. Statistical procedures use sample data to estimate the characteristics of the whole population from which the sample was drawn. Scientists typically want to learn about a population. When studying a phenomenon, such as the effects of a new medication or public opinion, understanding the results at a population level is much more valuable than understanding only the comparatively few participants in a study.

Unfortunately, populations are usually too large to measure fully. Consequently, researchers must use a manageable subset of that population to learn about it. By using procedures that can make statistical inferences, you can estimate the properties and processes of a population. More specifically, sample statistics can estimate population parameters.

There are two types of statistical inference:

• Estimation

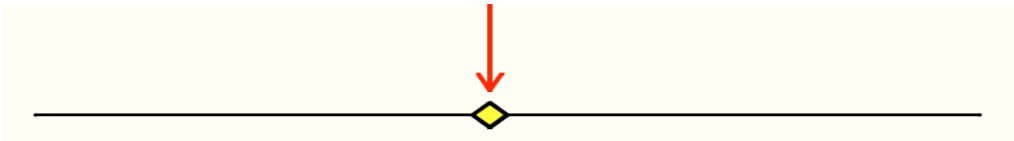
• Hypotheses Testing

ESTIMATION:

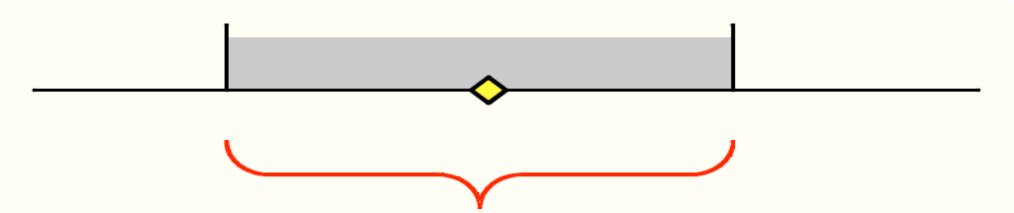
The objective of estimation is to approximate the value of a population parameter on the basis of a sample statistic. For example, the sample mean $\bar{X}$ is used to estimate the population mean μ.

There are two types of estimators:

1. Point Estimator



2. Interval Estimator



POINT ESTIMATION

Definition: A random sample of size 'n' from the distribution of X is a set of independent and identically distributed random variables $\{x_1, x_2, \ldots, x_n\}$ each of which has the same distribution as that of X. The probability of the sample is given by-

$f_0(x_1, x_2, \ldots, x_n) = f_0(x_1), f_0(x_2), \ldots f_0(x_n)$

Definition: A statistic T = T $(x_1, x_2, \ldots, x_n)$ is any function of the sample values, which does not depend on the unknown parameter θ. Evidently, T is a random variable which has its own probability distribution (called the 'Sampling distribution' of T)

For example,

$\bar{x} = \frac{1}{n}\sum_i^n x_i;\ s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2\ X_{(1)} = \min(x_1, x_2, \dots x_n), X_{(n)} = max(x_1, x_2, \dots x_n)$

are some statistics.

If we use the statistic T to estimate the unknown parameter θ, it is called the estimator (or point estimators) of θ and the value of T obtained from a given sample is its 'estimate'

Remark: Obviously, for T to be a good estimator of θ, the difference $[T - \theta]$ should be as small as possible. However, since T is itself a random variable all that we can hope for is that it is close to θ with high probability.

CRITERIA OF A GOOD ESTIMATOR:

Unbiasedness

An unbiased estimator of a population parameter is an estimator whose expected value is equal to that parameter. An estimator T of an unknown parameter θ is called unbiased if-

$E(T) = \theta$ for all $\theta \in \Omega$

Example: The sample mean $\bar{X}$ is an unbiased estimator for the population mean μ, since

$$E(\bar{X}) = \mu$$

It is important to realize that other estimators for the population mean exist: maximum value in a sample, minimum value in a sample, average of the maximum and the minimum values in a sample. Being unbiased is a minimal requirement for an estimator. For example, the maximum value in a sample is not unbiased, and hence should not be used as an estimator for μ.

Remarks: (i) An unbiased estimator may not exist.

(ii) Unbiased estimator may be assured.

(iii) Instead of the parameter θ we may be interested in estimating a function $g(\theta)$. $g(\theta)$ is said to be 'estimable' if there exists an estimator T Such that E(T)= $g(\theta)$, $\theta \in \Omega$.

Minimum Variance Unbiased (MVU) estimators: The class of unbiased estimators may, in general, be quite large and we would like to choose the best estimator from this class. Among two estimators of θ which are both unbiased, we would choose the one with smaller variance. The reason for doing this rests on the interpretation of variance as a measure of concentration about the mean. Thus, if T is unbiased for θ, then by Chebyshev's inequality-

$$P\{[T - \theta] \leq \varepsilon\} > 1 - \frac{Var(T)}{\varepsilon^2}$$

Therefore, the smaller $Var(T)$ is, the larger the lower bound of the probability of concentration of T about θ becomes. Consequently, within the restricted class of unbiased estimators we would choose the estimator with the smallest variance.

Definition: An estimator $T = T(X_1, X_2, \dots X_n)$is said to be a uniformly minimum variance unbiased (UMVU) estimator of θ (or an estimator for $g(\theta)$ if it is unbiased and has the smallest variance within the class of unbiased estimators of θ (or $g(\theta)$,) of all $\theta \in \Omega$. That is if T is any other unbiased estimator of θ, then-

$$Var(T) \leq Var(T')\ for\ \text{all}\ \theta \in \Omega$$

Consistency

An unbiased estimator is said to be consistent if the difference between the estimator and the target population parameter becomes smaller as we increase the sample size.

A sequence of estimator $\{Tn\}$. n = 1,2, … of a parameter θ is said to be consistent if, as n→∞ $Tn \to p\ \theta$ for each fixed $\theta\ \epsilon\ \Omega$ i.e. , for any ϵ(> 0)

$Tn\ converges\ to\ \theta\ in\ probablity$

Or $P\{|Tn - \theta| > \epsilon\} \to 0$

Or $P\{|Tn - \theta| \leq \epsilon\} \to 1$

$as\ n \to \infty$

Remarks:

($\boldsymbol{i}$) For increase in sample size a consistent estimator will become more and more close to θ

(ii)Consistency is essentially a large sample property. We speak of the consistency of a sequence of estimators rather than that of one estimator.

(iii) If $\{Tn\}$ is a sequence of estimator which is consistent for θ and $\{Cn\}$, $\{gn\}$ are sequence of constants such that $Cn \to 0\ g \to 1$ as $n \to \infty$ then $\{Tn + Cn\ \}\ and\ \{gnTn\}$ are sequences of consistent estimators also.

(iv) We will show later that if $\{Tn\}$ is a sequence of estimators such that $E(Tn) \to \theta$ and $V(Tn)$ →0 and $n \to \infty$ then $\{Tn\}$ is consistent.

Efficiency:

If T_1 and T_2 are two unbiased estimators of a parameter θ, each having finite variance T_1 is said to be more efficient then T_1 if $V(T_1)$ >$V(T_1)$. The (relative) efficient of T_1 relative to T_1 is defined by-

$$Eff\left({}^{T_1}/_{T_2}\right) = \frac{V(T_2)}{V(T_1)}$$

It is used to judge the efficiency of an unbiased estimator by comparing its variance with the Cramer- Rao lower bound (C R B).

Definition: Assume that the regularity condition of CR inequality hold (we call it a regular situation) for family $\{f(x, \theta), \theta \in \Omega\}$. An unbiased estimator T* of θ is called most efficient if $V(T*)$ equals the CRB. In this situation, the 'efficiency' of any other unbiased estimator T of θ is defined by-

$$Eff(T) = \frac{V(T^*)}{V(T)}$$

Where T* is the most efficient estimator defined above

Remarks:

(i)The above definition not proper in−

(a) regular situation when there is no unbiased estimator whose variance equals the CRB but an UMVUE exists and maybe found by other methods.

(b) Non-regular situations when an UMVUE exists and may be found by other methods

(ii)The UMVUE is 'most efficient' estimator in the examples considered earlier all UMVUE, whose variances equalled CRB are most efficient,

Sufficiency Criterion:

A preliminary choice among statistics for estimating θ, before having for a UMVUE as BAN estimator, can be made on the basic of another enter on suggested by R.A fisher. This is called 'sufficiency' criterion.

Definition: Let $(x_1, x_2,\dots, x_n)$ be a random sample from the distribution of X having þ, d, f $f(x, \theta)$ $\theta \epsilon \Omega$. A statistic $T = T(x_1, x_2,\dots, x_n)$ is defined to be **sufficient statistic** if and only if the conditional distribution of $(x_1, x_2,\dots, x_n)$ given T=t does not depend on θ, for any value t.

Note: In such a case if we know the value of the sufficient statistic T, then the sample values are not needed to tell us anything more about θ].

Also, the conditional distribution of any other statistic T (which is not for Ω tray) given T is independent of θ.

A necessary and sufficient condition for T to be sufficient for θ is that the joint þ. d, f of $(x_1, x_2,\dots, x_n)$ should be of the form $f(x_1, x_2,\dots, x_n; \theta) = g(T, \theta)h(x_1, x_2,\dots, x_n)$

where the first term on r, h, s., depends on T and θ and the second them is independent of θ. *T*his is known as Nyman's Factorisation Theorem which provides a simple method of judging whether a statistic T is sufficient

Remark: Any one to one function of a sufficient statistic is also a sufficient statistic.

METHODS OF ESTIMATION:

For important methods of obtaining estimators are (I) methods of moments,(II) methods of maximum likelihood (III)method of minimum $\chi 2$ and (IV) method of least squares.

(I) Method of moments

Suppose the distribution of a random variable X has K parameters $(\theta_1, \theta_2, \dots \theta_k)$ which have to be estimated.

Let $\mu_r = E(X^r)$ denote the rth moment of about O. In general, μ_r' is a known function of $\theta_1, \theta_2, \dots \theta_k$ so that = μ_r $(\theta_1, \theta_2, \dots \theta_k)$. Let $(x_1, x_2,\dots, x_n)$ be a random sample from the distribution of X and let $m_r = \sum_i^n \frac{x_i^r}{n}$ be the rth sample moment from the equation-

$$m'_r = \mu'_r(\theta_1, \dots \theta_k), r = 1, \dots k$$

Whose solution is say $\widehat{\theta_1}$, $\widehat{\theta_2}\dots$ $\widehat{\theta_k}$, where $\widehat{\theta_\iota}$ is the estimate of θi (i = 1, . . k) Those are the method of moments estimators of the parameters.

Remark: (a) the method of moments estimators is not uniquely defined. We may equate the central moments instead of the raw moments and obtain solutions.

(b) These estimators are not, in general, consistent and efficient but will be so only if the parent distributions are of particular form.

(c) When population moments do not exist ($e. g.$ Cauchy population) this method of estimation is inapplicable.

(II) METHOD OF MAXIMUM LIKELIHOOD

Consider $f(x_1, x_2,\dots, x_n, \theta)$, the joint þ, d, f of sample $(x_1, x_2,\dots, x_n)$ of observations of a, r, s. X having the p.d.f $f(x, \theta)$whose parameters θ is to be estimated. When the values $(x_1, x_2,\dots, x_n)$ are given, $f(x_1, x_2,\dots, x_n, \theta)$ may be looked upon as a function of θ which is called the likelihood function of θ and is denoted by $L(\theta) = L(\theta, x_1, x_2,\dots, x_n)$ it gives the likelihood that the r, v. $(x_1, x_2,\dots, x_n)$ assumes the value $(x_1, x_2,\dots, x_n)$ when θ is the parameter.

We want to know from which distribution ($i.\ e.$ for what value of θ) is the likelihood largest for this set of observations. In other words, we want to find the value of θ, denoted by $\hat{\theta}$ which maximizes $L(x_1, x_2,\dots, x_n, \theta)$. The value θ maximizes the likelihood function is in general, a function of $x_1, x_2,\dots, x_n$ say $\hat{\theta} = \hat{\theta}(x_1, x_2,\dots, x_n)$ Such that

$$L(\hat{\theta}) = \max L(\theta, x_1, x_2,\dots, x_n)\ \theta \in \Omega$$

Then $\hat{\theta}$ is called the maximum likelihood estimator or MLE.

In many cases it would be more convenient to deal with log $L(\theta)$, rather than $L(\theta)$, since log $L(\theta)$ is maximized for some value of θ as $L(\theta)$. For obtaining $m.\ \ell.\ e$ we find the value of θ for which-

$$\frac{\partial}{\partial\theta}\log L(\theta) = 0 \quad \dots\dots\dots (1)$$

We must however, check that this provides the absolute maximum. It the derivate dose not exists at $\theta = \hat{\theta}$ or equation (1) is not solvable this method of solving (1) will fail.

Example: Let $(x_1, .., x_n)$ be a, r, s from the Poisson's distribution

$$f(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2, ...$$

Then $$L(\lambda, x_1, .., x_n) = \frac{e^{-nx}\lambda^{\sum_i^n x_i}}{\prod_i^n x_i!}$$

And $\log L(\lambda) = nx + (\sum_i^n x_i)\log \lambda - \sum_i^n log x_i!$

Or $$\frac{\partial}{\partial\theta}\log L(\lambda) = -n + \frac{\sum_i^n x_i}{\lambda}$$

Equating to zero we get $\lambda = \bar{x}$

$m.\ell.e$ of λ is $\hat{\lambda} = \bar{x}$

Optimum properties of MLE: (i) If $\hat{\theta}$ is $m.\ l.\ e$ of θ and $\psi(\theta)$is a simple valued function of θ with unique inverse, then $\psi(\hat{\theta})$ is the $m.\ l.\ e$ of $\psi(\theta)$.
(ii) If a sufficient statistic exists for θ $m.\ l.\ e$ $\hat{\theta}$ is a function of this sufficient statistics.
(iii) Suppose $f(x, \theta)$ statistics certain regularity conditions and $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \ldots x_n)$ is the $m.\ l.\ e$ of a random sample of size n from $f(x, \theta)$
Then- (a) $\{\hat{\theta}_n\}$ is consistent sequence of estimators of θ.

(b) $\hat{\theta}_n$ is asymptotically normally distributed with mean θ variance $\frac{1}{nE\left[\frac{\partial}{\partial\theta}log f(x,\theta)\right]^2}$

(c)The sequence of estimators $\hat{\theta}_n$ has the smallest asymptotic variance among all consistent, asymptotically normally distributed estimate of θ,i.e. $\hat{\theta}_n$ is BAN or CANE or most efficient.

(III) METHOD OF MINIMUM χ^2: Let X be $a.\ r.\ v$ with þ. $d.\ f$ $f(x, \theta)$ where parameter to be estimated $\theta = (\theta_1, \theta_2, \ldots \theta_r)$ Suppose $S_1, S_2, \ldots S_k$ are $\hbar$ mutually exclusive classes which from a partition of the range of X. Let the profanity at X falls in S_J be-

$$\text{þ}_j(\theta) = \int_{S_j} f(x,\theta)dx \,, j = 1,2, \ldots k$$

Where $\sum_{j=1}^{k} \text{þ}_j(\theta) = 1$

Suppose ,in practice ,corresponding to a random sample of n observations from the distribution of X we are given the frequencies $(N_1, \ldots\ldots, N_k)$ where N_j=observed number of sample observations falling in the class $S_J (j = 1,2 \ldots\ldots k)$ such that $\sum_i^k N_j = n$ then the expected number of observation in S_J is $n\text{þ}_j(\theta)$, Define

$$\chi 2 = \sum_{j=1}^{k} [n_j - n\text{þ}_j(\theta)]^2 / n\text{þ}_j(\theta)$$

Where n_j is the observed value of $N_j (j = 1,2 \ldots\ldots k)$ Evidently x^2 will be a function of $\theta (or\ \theta_i, .. \theta_r)$to obtain the estimator of θwe minimise $x^2 w.r,t\ \theta$. The minimise x^2 estimator of θ is that $\hat{\theta}$ which minimise above $\chi 2$.

The equation (s) for determining the estimator(s) by this me that are

$$\frac{\partial \chi 2}{\partial \theta} = \theta \ \ or \frac{\partial \chi 2}{\partial \theta} = 0 \ \ (i = 1, \ldots\ldots r)$$

Remarks:

(i)Often it is difficult to obtain $\widehat{\theta}$ which minimum $\chi 2$, hence $\chi 2$ is changed to modified

$$\chi 2 = \sum_{j=1}^{k} [n_j - n\text{þ}_j(\theta)]^2 \, n_j$$

(If $n_j = 0$, unity is used). The modified minimum $\chi 2$ estimator of θ is $\hat{\theta}$ which minimises the modified $\chi 2$

(ii) For large n, the minimum $\chi 2$ and likelihood equations are identical and, consequently, provide identical minimum $\chi 2$maximum likelihood estimators.

(iii) The minimum $\chi 2$ estimators are consistent asymptotically normal and efficient .

(IV) METHOD OF LEAST SQUARES Suppose Y is a random variable whose value depends on the value of a (non-random) variable X. For example, the weight of a baby (Y) depends on its age(X) , the temperature (Y) of a place at a given time depends on its altitude (X), or the salary (Y) of an individual at a given age depends on the number of years (X) of formal education which he has had the maintenance cost (Y) per year of an automobile depends on its age (X) etc.
We assume that the distribution of the $r.\ v\ Y$ is such that for a given x, E(Y/X) is a linear function of X while the variance and higher moments of Y are independent of X. It means that we assume the linear model-

$$E(\mathrm{Y}/x)=\alpha+\beta x$$

Where d and β and two parameters .We also write

$$\mathrm{Y}=\alpha+\beta x+\epsilon$$

Where ϵ is a, r, u such that E $(\epsilon)=\theta, V(\theta)=\sigma^2$

The problem is to estimate the parameters $\mathcal{d}$ and β on the basic of a random sample of n observations (Y, X_1), (Y, X_2), … … (Y, X_n)
The method of least squares estimations of α and β specifies that we should take as our estimates of $\mathcal{d}$ and β those values that minimise-

$$\sum_{i=1}^{n}[y_i - \alpha - x_i]^2$$

Where yi is the observed value of Y_i and X_i are the associated values of x. This we minimise the sum of squares of the residuals when applying the method of least squares.
The least squares estimators of α and β

Are $$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

And $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Remarks:
The least square estimator does not have any optimum properties ever asymptotically However in linear estimation this method provides good estimation in small simples. These estimators are minimum variance unbiased estimators among the class of linear function of Y′s.

TESTING OF HYPOTHESIS

(NEYMAN PEARSON THEORY)

Let x be a, r, u with þ.$d.f$ $f(x,\theta)$ where θ (may be a vector $(\theta_1, \ldots\ldots\theta_k)$ is an unknown parameter. A random sample of n observation denoted by $E = (x_1 \ldots\ldots x_k)$ which takes values in general, in the n-dimensional real space R_n the parameter space (all possible values of the parameter is denoted by Ω, say . For any subset AC R_n we can calculate.

$$p_o(E\ E\ A) = \int_A \left[\prod_{i=1}^{n} f(x,\theta)\right] dx, \ldots d_{xn}$$

Which will depend on θ.

Definition: A statistical hypothesis is a statement about the parameter θin the form $H: \theta \in \omega (< \Omega)$

For example consider

$$H: \theta = \theta_0$$

$$or\ H: \theta \geqslant \theta_0\ or\ H: \theta \neq \theta_0\ H: \theta_1 < \theta < \theta_2$$

Definition If a hypotheses specifics an exact value of the parameterθ, it is called a simple hypothesis

$e, g\ .\ H: \theta = \theta_0$ in this case ω in $H: \theta \in \omega$ is a set of a single point

If a hypothesis does not fully specify the value of θ(but gives a set of possible values only) it is called a composite hypothesis $e, g\ H: \theta \neq \theta_0$ or $H: \theta \geqslant \theta_0$ etc. In this case ω in $H: \theta \in \omega$ is set of more than one point.

Definition the hypothesis which is being actually tested is called the null hypothesis and other hypothesis which is stated as the alternative to the null hypothesis is called alternative hypothesis. For example, null hypothesis may be $H_o: \theta = \theta_0$ and the alternative may be $H_i: \theta \neq \theta_0$ or $H_i: \theta > \theta_0$ or $H_i: \theta \leq \theta_0$ etc.

Both null and alternative hypothesis may be simple or composite .For our study, we shall usually take null hypothesis to be simple .

Suppose we want to test a null hypothesis H_O against an alternative hypothesis H_1 on the basis of a random sample $E = (X_1, .. X_n)$ in the sense that we have to decide when to reject or accept H_O

Definition A Statistical test of a (null) hypothesis H_O against an alternative hypothesis H_1 is a rule or procedure for deciding when to reject or accept H_O on the basis of the sample $E = (X_I, .. X_n)$ .It specifies a position of the sample space R_n into two disjoint subsets W and $\overline{W} = R_n - W$ such that we reject H_O when $E \epsilon W$ and accept H_O when $E \epsilon \overline{W}$ [We note that the rejection of H_O amounts to acceptance of H_1 and vice-versa]

Definition The set W, corresponding to a test T, which is that we reject H_O when $E \epsilon W$ is called the critical region of the test while $\overline{W}$ is called its acceptance region. For different test the critical regions are different.

Two types of errors: In a testing problem we are liable to commit two types of error. Suppose H_O is true and get $E \epsilon \overline{W}$ so that we reject H_O this is called the Type I error which occasion when we reject the null hypothesis when it is actually true. On the other hand, suppose H_O is false and H_i is true and yet $x \epsilon \overline{w}$ so that we accept H_O this is called the types II errors which occurs when we accept the null hypothesis when it is actually false. We denote by α and β the probability of type I error and type II error, respectively, i, e

$$\alpha = P\{H_o / H_o \text{is true}\}$$

$$= P\{E \epsilon W / \theta \ \epsilon H_o\}$$

And $$\beta = P\{AccepH_o / H_o \text{ is false}$$

$$= P\{E\epsilon\bar{W} / \theta \epsilon H_o\}$$

Definition The probability of type I error for a test T, denoted by $\propto$ is called the "size" or level of significance of the test T.

Remark If H_o is simple (say H_o:θ=θ_o) is clearly defined ,when H_o is composite (say H_o: $\theta \epsilon W$)we take

$$\alpha = supP_T\{E \epsilon W /: \theta\} \theta\epsilon H_o$$

Definition For a test T having the co region w_2 the power function $P_T()$ is defind by

$$P_T(\theta) = P\{RejectH_o / \theta\}$$

$$= \mathrm{P}_\theta\{\mathrm{E} \epsilon W\}$$

As a function of θ

Evidently,

$$P_T(\theta) = \alpha \ for \ \theta \ \epsilon H_o$$

$$P_T(\theta) = 1 - \beta \ \theta \ \epsilon H_1$$

If we would find a test of the given hypothesis for which both $\propto$and $\beta$are minimum it would be the best. Unfortunately, it is not possible to minimise both error simultaneously for a fixed sample size test. Consists two tests T and T_2 defined as follows

T_I always rejects H_0 i,e its critical region $W_1 = R_n$, while T_2 always accepts H_o i,e its cur region $W_2 = \emptyset$ then for T_1, $\propto = 0$ and $\beta = 1$ this shows that if the probability of type I error becomes minimum than the probability of type II error becomes maximum and vice-versa what is done is to fix $\propto$, taking $\propto$ to be quite small (in practical $\propto = .05$ or $.01$)so that all test of size $\propto$ are only considered. Among all test of a given size$\propto$ comparison made on the basic of their power function. If T and T are two tests (for the same testing problem) of same size $\propto$, T is said to be better than T if its power is greater than the power of T for alteration hypothesis (equivalently the probability of type II error for T is less then the probability of type II for T,)

Example 18.3. *Let p be the probability that a coin will fall head in a single toss in order to test* $H_0 : p = \frac{1}{2}$ *against* $H_1 : p = \frac{3}{4}$. *The coin is tossed 5 times and* H_0 *is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.*

Solution. Here $H_0 : p = \frac{1}{2}$ and $H_1 : p = \frac{3}{4}$.

If the $r.v.$ X denotes the number of heads in n tosses of a coin then $X \sim B(n, p)$ so that

$$P(X = x) = {}^nC_x\, p^x (1-p)^{n-x} = {}^5C_x\, p^x (1-p)^{5-x} \quad \ldots (*)$$

The critical region is given by : $W = \{x : x \geq 4\} \Rightarrow \overline{W} = \{x : x \leq 3\}$

α = Probability of type I error $= P[X \geq 4 \mid H_0]$

$$= P\left(X = 4 \mid p = \frac{1}{2}\right) + P\left(X = 5 \mid p = \frac{1}{2}\right) = {}^5C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{5-4} + {}^5C_5 \left(\frac{1}{2}\right)^5 \text{ [From (*)]}$$

$$= 5\left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = 6\left(\frac{1}{2}\right)^5 = \frac{3}{16}$$

β = Probability of Type II error $= P(x \in \overline{W} \mid H_1) = 1 - P(x \in W \mid H_1)$

$$= 1 - \left[\left\{P\left(X = 4 \mid p = \frac{3}{4}\right) + P\left(X = 5 \mid p = \frac{3}{4}\right)\right] = 1 - \left\{{}^5C_4 \left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right) + {}^5C_5 \left(\frac{3}{4}\right)^5\right\}$$

$$= 1 - \left(\frac{3}{4}\right)^4 \left\{\frac{5}{4} + \frac{3}{4}\right\} = 1 - \frac{81}{128} = \frac{47}{128}$$

$\therefore$ Power of the test $= 1 - \beta = \frac{81}{128}$.

LIKELIHOOD RATIO TEST:

Neyman-Pearson Lemma based on the magnitude of the ratio of two probability density functions provides best test for testing simple hypothesis against simple alternative hypothesis. The best test in any given situation depends on the nature of the population distribution and the form of the alternative hypothesis being considered.

A general method of test construction called *Likelihood Ratio (L.R) Test* by Neyman and Pearson for testing a hypothesis, simple or composite, against a simple or composite alternative hypothesis. This test is related to maximum likelihood estimates.

Parameter Space. Let us consider a random variable X with p.d.f. $f(x, \theta)$. In most common applications, though not always, the functional form of the population distribution is assumed to be known except for the value of some unknown parameter(s) θ which may take any value on a set Θ. This is expressed by writing the p.d.f. in the form $f(x, \theta)$, $\theta \in \Theta$. The set Θ, which is the set of all possible values of θ is called the *parameter space*. Such a situation gives rise not to one probability distribution but a family of probability distributions which we write as $\{f(x, \theta) = \theta \in \Theta\}$. For example if $X \sim N(\mu, \sigma^2)$, then the parameter space is :

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma < \infty\}$$

In particular, for $\sigma^2 = 1$, the family of probability distributions is given by

$$\{N(\mu, 1); \mu \in \Theta\}, \text{ where } \Theta = \{\mu : -\infty < \mu < \infty\}$$

In the following discussion we shall consider a general family of distributions:

$$\{f(x : \theta_1, \theta_2, \ldots, \theta_k) : \theta_i \in \Theta, i = 1, 2, \ldots, k\}$$

The null hypothesis H_0 will state that the parameters belong to some subspace Θ_0 of the parameter space Θ.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size $n > 1$ from a population with p.d.f. $f(x, \theta_1, \theta_2, \ldots, \theta_k)$, where Θ, the parameter space is the totality of all points that $(\theta_1, \theta_2, \ldots, \theta_k)$ can assume. We want to test the null hypothesis :

$$H_0 : (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta_0$$

against all alternative hypotheses of the type :

$$H_1 : (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta - \Theta_0$$

The likelihood function of the sample observations is given by

Using (18·17), we can obtain the maximum likelihood estimates for the parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ as they are allowed to vary over the parameter space Θ and the subspace Θ_0. Substituting these estimates in (18·16), we obtain the maximum values of the likelihood function for variation of the parameters in Θ and Θ_0 respectively. Then the criterion for the likelihood ratio test is defined as the quotient of these two maxima and is given by

$$\lambda = \lambda(x_1, x_2, \ldots, x_n) = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\underset{\theta \in \Theta_0}{\text{Sup}}\, L(x, \theta)}{\underset{\theta \in \Theta}{\text{Sup}}\, L(x, \theta)}, \quad \ldots(18\cdot19)$$

where $L(\hat{\Theta}_0)$ and $L(\hat{\Theta})$ are the maxima of the likelihood function (18·16) with respect to the parameters in the regions Θ_0 and Θ respectively.

The quantity λ is a function of the sample observations only and does not involve parameters. Thus λ being a function of the random variables, is also a random variable. Obvious $\lambda > 0$. Further

$$\Theta_0 \subset \Theta \quad \Rightarrow \quad L(\Theta_0) \leq L(\Theta) \quad \Rightarrow \quad \lambda \leq 1 \quad \ldots(18\cdot19)$$

Hence, we get $0 \leq \lambda \leq 1$

The critical region for testing H_0 (against H_1) is an interval

$$0 < \lambda < \lambda_0 \quad \ldots(18\cdot20)$$

where λ_0 is some number (< 1) determined by the distribution of λ and the desired probability of type 1 error, *i.e.*, λ_0 is given by the equation :

$$P(\lambda < \lambda_0 \mid H_0) = \alpha \quad \ldots(18\cdot21)$$

For example, if $g(.)$ is the p.d.f. of λ then λ_0 is determined from the equation :

$$\int_0^{\lambda_0} g(\lambda \mid H_0)\, d\lambda = \alpha \quad \ldots(18\cdot21a)$$

A test that has critical region defined in (18·20) and (18·21) is a *likelihood ratio test* for testing H_0.

18·6·1. Properties of Likelihood Ratio Test. Likelihood ratio (*L.R.*) test principle is an intuitive one. If we are testing a simple hypothesis H_0 against a simple alternative hypothesis H_1 then the *LR* principle leads to the same test as given by the Neyman-Pearson lemma. This suggests that *LR* test has some desirable properties, specially large sample properties.

In *LR* test, the probability of type *I* error is controlled by suitably choosing the cut off point λ_0. *LR* test is generally *UMP* if an *UMP* test at all exists. We state below, the two asymptotic properties of *LR* tests.

1. Under certain conditions, $-2 \log_e \lambda$ has an asymptotic chi-square distribution.
2. Under certain assumptions, *LR* test is consistent.

Now we shall illustrate how the likelihood ratio criterion can be used to obtain various standard tests of significance in Statistics.

BAYESIAN VERSUS FREQUENTIST INFERENCE:

- There are 2 main approaches to statistical inference, frequentist and Bayesian, differing in their interpretation of uncertainty.
- The frequentist approach deals with long-run probabilities (ie, how probable is this data set given the null hypothesis), whereas the Bayesian approach deals with the probability of a hypothesis given a particular data set.
- Bayesian analysis incorporates prior information into the analysis, whereas a frequentist analysis is purely driven by the data.
- The Bayesian approach can calculate the probability that a particular hypothesis is true, whereas the frequentist approach calculates the probability of obtaining another data set at least as extreme as the one collected (giving the *P* value).
- Interpretation of results is more intuitive with a Bayesian approach compared with the frequentist approach, which can often be misinterpreted.

Just like Liverpool versus Manchester United, the Yankees versus the Red Sox, or Coke versus Pepsi, there are 2 main schools of statistics, and you may have heard the proponents of each noisily arguing their respective benefits. The first is the frequentist approach, which dominates the medical literature and consists of null hypothesis significance testing (think *P* values and confidence intervals). The other is the Bayesian approach, governed by Bayes' theorem. The fundamental difference between these 2 schools is their interpretation of uncertainty and probability: the frequentist approach assigns probabilities to data, not to hypotheses, whereas the Bayesian approach assigns probabilities to hypotheses. Furthermore, Bayesian models incorporate prior knowledge into the analysis, updating hypotheses probabilities as more data become available. The goal of this article is to educate readers about the differences between frequentist and Bayesian inference, discuss potential advantages or disadvantages of each approach, and use the case vignette to highlight how these 2 methods may be implemented in a real-world example in a radiation therapy clinic.

Introduction to Frequentist Statistics

Frequentist statistics is all about probability in the long run; the data set collected and analyzed is one of many hypothetical data sets addressing the same question, and uncertainty is due to sampling error alone. For example, the probability of getting heads when flipping a coin in the long run is 0.5; if we flip the coin many times, we would expect to see heads 50% of the time, whereas if we had flipped the coin only a few times we could reasonably expect to observe a different distribution (eg, all heads) just by chance.

Frequentist inference begins by assuming a null hypothesis to be true before data are collected (eg, that there is no effect of a particular treatment on survival). Investigators then collect data, analyze them, and ask, "How surprising is my result if there is actually no effect of the treatment on survival?" The data would be surprising if there was a low probability by chance alone of obtaining another data set at least as extreme (ie, far away from the null hypothesis and unlikely to occur by chance) than that collected (eg, showing a large difference in survival between patients having different treatments when our null hypothesis states there is no difference). If this is the case, then the collected data are considered unlikely under the null hypothesis and we can reject it, inferring that the null hypothesis does not adequately explain our data and that something else (eg, the new treatment) must account for our results.

This probability of obtaining another data set as extreme as the one collected is known as the *P* value. The *P* value is often criticized for being misunderstood and misused in the field of medicine. For example, in contrast to popular belief, the *P* value is not a measure of how correct a hypothesis is, nor is it a measure of the size or importance of an effect. In particular, large *P* values do not provide evidence of no effect. A *P* value is simply the probability of obtaining another data set at least as extreme as the one collected by chance alone.

- For example, suppose we run an analysis and get a highly significant *P* value of .001 for our treatment variable—how do we interpret this? Formally, there is a 0.1% chance of collecting data equal to or more extreme than this result if the null hypothesis were true—it would be surprising to collect these data if there is, in fact, no effect of treatment on survival. If we had run the analysis and got a *P* value of .46, then there is a 46% chance of collecting data equal to or more extreme than this if the null hypothesis is true—it would not be surprising to obtain these results if there is no effect of treatment on survival. This result is not evidence of no effect, however, because the inference began by assuming there would be no effect of treatment. The key here is that probabilistic statements (ie, *P* values) can only be made about the data, not about hypotheses or parameters (ie, the treatment effect).
- Reporting confidence intervals can improve the interpretation of results compared with a *P* value alone and can give information on the size and direction of an effect.
- A 95% confidence interval tells us that if we were to repeat the experiment over and over (remember, frequentist statistics are long run), 95% of the computed confidence intervals would contain the true mean.
- This is different than saying there is 95% chance the true mean lies within the interval, because frequentist statistics cannot assign probabilities to parameters—the true mean either lies within the interval or it does not.

- **Introduction to Bayesian Statistics**
- Bayesian statistics are named after the Reverend Thomas Bayes, whose theorem describes a method to update probabilities based on data and past knowledge. In contrast to the frequentist approach, parameters and hypotheses are seen as probability distributions and the data as fixed. This idea is perhaps more intuitive because generally the data we collect are the only data set we have, so it does not necessarily make sense to perform statistical analysis assuming it is one of many potential data sets. Probability

distributions summarize the current state of knowledge about a parameter or hypothesis and can be updated as more data becomes available using Bayes' theorem, presented in Equation 1.

$$p\left(\theta|Data\right) = \frac{p(Data|\theta)\cdot p(\theta)}{p(Data)}$$

The probability distribution that summarizes what is known about an outcome before a test or piece of information is obtained is known as the prior distribution, often just dubbed "the prior." Such an outcome may be the prevalence of disease or a specific diagnosis. The prior probability of the outcome θ (eg, of having the disease) is labeled $P(\theta)$ in Equation 1. The prior is one of the key differences between frequentist and Bayesian inference; frequentist analyses base their results only on the data they collect. The prior could be formulated by expert beliefs, historical data, or a combination of the two.

Consider now that one has a positive test for the disease. What is the probability of θ (having the disease), given these new data (ie, that the test was positive)? The notation for this scenario is $P(\theta|$ test +), where the "|" can be translated as "given" and we set the *Data* variable in Equation 1 to *test* +. In Bayesian terminology, this probability is known as the posterior distribution (or posttest distribution) and summarizes what is known about the outcome using both the prior information and the new data.

Among all potential patients, regardless of whether they have the disease, the probability of a positive test (ie, true positive plus false positive) is P(test+). Therefore, the relative probability of having a positive test if a patient is truly positive versus the probability of someone randomly selected from the population (who may or may not have the disease) having a positive test is the ratio $P(\text{test+}|\theta)$ / P(test+). The posterior probability of having the disease if you have a positive test is then the baseline prevalence of the disease in the population (the prior probability) multiplied by this factor.

We can further illustrate how Bayes' theorem and the use of prior information can help to answer important questions using the example of COVID-19 testing. For this example, let's assume that a test for COVID-19 infection is guaranteed (100% chance) to detect the COVID-19 virus in someone who has the infection (P[test+| virus] = 1.0) and has a 99.9% chance of correctly identifying that someone does not have the virus (or a 0.1% chance of a false positive: P[test- | no virus] = 0.999 and P[test + | no virus] = 0.001).

[9]

That sounds like a high probability, but what we are really interested in is if you have a positive test, how likely is it that you actually have the virus (P[virus | test+]). We can calculate this probability using Bayes' theorem. As discussed, this probability depends on some prior information—how likely you were to have the virus (ie, P[virus]) before taking the test.

$$p\left(virus|test+\right) = \frac{p(test+|virus)\cdot p(virus)}{p(test+)}$$

If we assume the prior probability is the prevalence of the virus in the population at the height of the pandemic, 2%, then P(virus) = 0.02. Out of 1 million people, 20,000 will have the virus and 980,000 will not. If we test them all, 20,000 will have a true positive test, 979,020 will

have a true negative test, and 980 will have a false positive test (980,000 × 0.001). Thus, there would be 20,980 positive tests in total (P[test+] = 20,980/1,000,000 = 0.02098), and the chance of having the virus given a positive test could be calculated using Equation 2: P(virus | test +) = (1.0 × 0.02)/0.02098 = 0.953, which can be approximated as 95%. In this setting, if the test is positive, one should believe the test.

However, what if the prevalence of COVID-19 was estimated to be much less, say, 0.2% (P[virus] = 0.002), such as during a period of a governmental stay-at-home order? In this scenario, when we test 1 million people, 2000 will have a true positive test, 997,002 will have a true negative test, and 998 will have a false positive test (998,000 × 0.001), giving 2998 positive test results, and P(test+) = 2988/1,000,000 = 0.002988. The probability of having the virus after a positive test now becomes P(virus | test+) = (1.0 × 0.002)/0.002988 = 0.669. In this scenario, one may truly question whether a positive test is diagnostic of infection, because the likelihood of a false positive is approximately 1 in 3.

Therefore, knowledge of the prior information (in this case, the prevalence of COVID-19 in a population) can alter the chance of having a virus when a test is positive from 95% to 67% when the sensitivity and specificity of the test remain unaltered. On the other hand, if we were testing hospitalized patients, for example, the prevalence would be expected to be much higher, and there would be a lower chance of obtaining a false positive result. This simple calculation highlights the importance of taking into account prior information, something a frequentist analysis does not do.

Sometimes formulating a prior is not that easy or clear. In such cases, the use of priors can be seen as a drawback to Bayesian inference, particularly when results depend on the chosen prior, and thus the analysis could be manipulated to get a positive result. In such cases, an "uninformative" prior that provides no additional information could be used, or multiple priors (eg, with either optimistic or skeptical assumptions) could be tested to determine the sensitivity of the results to particular priors. As with all analyses, it is vital that researchers are transparent in their methods and assumptions.

The posterior distribution captures our "updated" estimate of the probability of the outcome after incorporating new data, including our uncertainties, and can be analyzed to give various statistics, such as the mean and 95% credible interval. The 95% credible interval is different from a frequentist 95% confidence interval; it is the parameter range that has a 95% probability of including the true parameter value. The frequentist confidence interval is often misinterpreted in this way; however, one must remember that the 95% confidence interval assumes that the experiment is hypothetically repeated over and over and that 95% of the computed confidence intervals would contain the true mean. Perhaps more importantly, and one of the big advantages of Bayesian inference, is that the posterior distribution can also be used to directly calculate the probability of different hypotheses (eg, that one treatment is superior to another or that survival is improved by at least 3 months).

In this respect, Bayesian inference is more intuitive at its core and in closer alignment with our natural mode of probabilistic reasoning than frequentist inference. For example, we are more interested in the probability that 1 treatment is superior to another (Bayesian probability) than in the probability of obtaining certain data assuming the treatments are equal (frequentist null hypothesis). This advantage in interpretability remains even if our analysis uses an uninformative prior.

APPLICATION OF BAYESIAN INFERENCE IN DATA SCIENCE:

As a data science professional, you likely hear the word "**Bayesian**" a lot. Whether it's Bayesian Inference, Bayesian Statistics, or other Bayesian methods, knowing how to use and apply these methods is almost a necessity for any practicing professional. There are many different applications that one may use, and through a variety of industries. As such, here are a few examples of when you might use Bayesian methods.

Business and Commerce

Bayesian Inference is a popular data science technique for businesses, especially for pricing decisions. Businesses can determine prices for products based on field information like retail and wholesale prices, the size of the market, and market share.

Bayesian methods can also be used for new product development as a whole. Mainly, one would look at project risk by weighing uncertainties and determining if the project is worth it.

However, when it comes to Bayesian inference and business decisions, the most common application relates to product ranking. Wayfair developed its own Bayesian system to help customers have an ideal and customized shopping experience. Online shopping giants like Amazon use it to make ratings appear natural when searching for products, as opposed to displaying rankings in order the default option.

Marketing

Marketing decisions come in a number of ways. Some teams throw ideas at a dartboard and hope that one sticks. Others will use Bayesian methods to look at previous campaigns and marketing information to improve existing campaigns or to make new ones. You can also use Bayesian approaches in A/B testing to play around with email strategies, website designs, and so on. Market research also benefits from a Bayesian approach by examining probably insights to target the right audience.

Stock Markets

In a field all about predictions based on trends, Bayesian networks can be used to identify future trends in stocks based on previous trends. This has been a popular trend over the past year, especially as COVID was a new variable for investors to be concerned with.

Weather Prediction

While there's nothing wrong with traditional forecasting methods, Bayesian deep learning can be used to predict the weather too. The research is still in its early stages, but some researchers

are aiming to use recent computer vision imagery and recent weather patterns to predict what will happen next.

Disease Risk

Bayesian methods can be used in disease mapping to determine underlying disease risk in individuals and groups. With Bayesian hierarchical models, researchers can look at individual risk factors, subregions with risks and predispositions to disease, and other variables to determine if one may succumb to a particular disease.

Medical Diagnosis

Most doctors use Bayesian inference without realizing it. They get a sick patient, look at their history, their lifestyle, and other factors to determine what problem the patient may have. Bayesian analysis can even be used to fill in incomplete medical records based on the history and trends of the individual.

UNIT-V

STOCHASTIC PROCESS:

Definition: A **stochastic process** is a *family of random variables,* $\{X(t) : t \in T\}$, *where* t *usually denotes time. That is, at every time* t *in the set* T, *a random number* $X(t)$ *is observed.*

Definition: $\{X(t) : t \in T\}$ is a **discrete-time process** if *the set* T *is finite or countable.*

In practice, this generally means $T = \{0, 1, 2, 3, \ldots\}$

Thus a discrete-time process is $\{X(0), X(1), X(2), X(3), \ldots\}$: *a random number associated with every time 0, 1, 2, 3, . . .*

Definition: $\{X(t) : t \in T\}$ is a **continuous-time process** if T *is not finite or countable.*

Definition: The **state space**, S, is *the set of real values that* $X(t)$ *can take.*

Every $X(t)$ takes a value in $\mathbb{R}$, but S will often be a smaller set: $S \subseteq \mathbb{R}$. For example, if $X(t)$ is the outcome of a coin tossed at time t, then the state space is $S = \{0, 1\}$.

Definition: The state space S is **discrete** if it is *finite or countable.* Otherwise it is *continuous.*

The state space S is the set of **states** that the stochastic process can be in.

MARKOV CHAIN:

Suppose that a random process passes through a discrete sequence of **steps** or **trials** numbered $n = 0, 1, 2, \ldots$, where the outcome of the n-th trial is the random variable X_n; X_0 is the initial position of the process. This discrete random variable can take one of the values $i = 1, 2, \ldots m$. The actual outcomes are called the **states** of the system, and are denoted by E_i $(i = 1, 2, \ldots m)$ (states can be any consecutive sequence of integers, say starting with $i = 0$ or some other integer, but they can be renumbered to start with $i = 1$). In most but not quite all cases in this text, we shall investigate systems with a *finite* number, m, of states $E_1, E_2, \ldots E_m$, which are independent and exhaustive.

If the random variables $X_{n-1} = i$ and $X_n = j$, then the system has made a **transition** $E_i \to E_j$, that is, a transition from state E_i to state E_j at the n-th trial. Note that i can equal j, so that transitions within the same state may be possible. We need to assign probabilities to the transitions $E_i \to E_j$. This discrete process is known as a **chain**. Generally in chains the probability that $X_n = j$ will depend on the whole sequence of random variables starting with the initial value X_0. The **Markov chain** has the characteristic property that the probability that $X_n = j$ depends *only* on the immediate previous state of the system. Formally this means that we need no further information at each step other than, for each i and j,

$$\mathbf{P}\{X_n = j | X_{n-1} = i\},$$

which means the probability that $X_n = j$ given that $X_{n-1} = i$: this probability is *independent* of the values of $X_{n-2}, X_{n-3}, \ldots, X_1$. Put alternatively, conditional on the present state of the chain, its future and present are independent.

In some chains the probabilities $\mathbf{P}\{X_n = j | X_{n-1} = i\}$ are functions of n, the step or trial number. If this is not the case, so that the probabilities are the same at every step, then the chain is said to be **homogeneous**.

Transition probabilities

For a finite Markov chain with m states $E_1, E_2, \ldots, E_m$, introduce the notation

$$p_{ij} = \mathbf{P}\{X_n = j | X_{n-1} = i\},$$

where $i, j = 1, 2, \ldots, m$ to represent the probability of a transition from state E_i to E_j. The numbers p_{ij} are known as the **transition probabilities** of the chain, and must satisfy

$$p_{ij} \geq 0, \quad \sum_{j=1}^{m} p_{ij} = 1$$

for each $i = 1, 2, \ldots, m$. If $p_{ij} > 0$, then we say that state E_i can **communicate** with E_j: two-way communication is possible if additionally $p_{ji} > 0$

Obviously for each fixed i, the list $\{p_{ij}\}$ is a **probability distribution**, since at any step one of the outcomes $E_1, E_2, \ldots, E_m$ must occur: the states E_i, $(i = 1, 2, \ldots m)$.

Transition probabilities form an $m \times m$ array which can be assembled into a **transition matrix** T, where

$$T = [p_{ij}] = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}.$$

Note that each row of T is a probability distribution. Any square matrix for which $p_{ij} \geq 0$ and $\sum_{j=1}^{m} p_{ij} = 1$ is said to be **row-stochastic**.

TRANSITION PROBABILITIES

(i) **The absolute probability** $p_j^{(n)}$

One further probability which will be of interest is the probability of outcome E_j after n steps, given an **initial probability distribution** $\{p_i^{(0)}\}$. Here $p_i^{(0)}$ is the probability that initially the system occupies state E_i. Of course we must have $\sum_{i=1}^{m} p_i^{(0)} = 1$. Let $p_j^{(1)}$ be the probability E_j is occupied after one step. Then, by the law of total probability

$$p_j^{(1)} = \sum_{i=1}^{m} p_i^{(0)} p_{ij}.$$

We can express this more conveniently in vector form. Let $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ be the **probability (row) vectors** given by

$$\mathbf{p}^{(0)} = \begin{bmatrix} p_1^{(0)} & p_2^{(0)} & \ldots & p_m^{(0)} \end{bmatrix}$$

and

$$\mathbf{p}^{(1)} = \begin{bmatrix} p_1^{(1)} & p_2^{(1)} & \ldots & p_m^{(1)} \end{bmatrix}.$$

Here $\mathbf{p}^{(0)}$ is the initial distribution, and the components of $\mathbf{p}^{(1)}$ will be the probabil-

ities that each of the states $E_1, E_2, \ldots, E_m$ is reached after one step. Equation can be represented as a matrix product as follows:

$$\mathbf{p}^{(1)} = \left[p_j^{(1)}\right] = \left[\sum_{i=1}^{m} p_i^{(0)} p_{ij}\right] = \mathbf{p}^{(0)} T,$$

where T is the transition matrix given by
If $\mathbf{p}^{(2)}$ is the distribution after two

steps, then

$$\mathbf{p}^{(2)} = \mathbf{p}^{(1)} T = \mathbf{p}^{(0)} TT = \mathbf{p}^{(0)} T^2.$$

Hence after n steps by repeating the process

$$\mathbf{p}^{(n)} = \mathbf{p}^{(n-1)} T = \mathbf{p}^{(0)} T^n,$$

where

$$\mathbf{p}^{(n)} = \begin{bmatrix} p_1^{(n)} & p_2^{(n)} & \ldots & p_m^{(n)} \end{bmatrix}.$$

More generally,

$$\mathbf{p}^{(n+r)} = \mathbf{p}^{(r)} T^n.$$

the component $p_i^{(n)}$ is the **absolute** or **unconditional probability of outcome** E_j at the n-th step given the initial distribution $\mathbf{p}^{(0)}$, that is, $\mathbf{P}\{X_n = j\} = p_j^{(n)}$. Note that

$$\sum_{j=1}^{m} p_j^{(n)} = 1.$$

(ii) *The n-step transition probability $p_{ij}^{(n)}$*

We now define $p_{ij}^{(n)}$ as the probability that the chain is in state E_j after n steps *given that the chain started in* state E_i. The first step transition probabilities $p_{ij}^{(1)} = p_{ij}$ are simply the elements of the transition matrix T. We intend to find a formula for $p_{ij}^{(n)}$. Now, by definition,

$$p_{ij}^{(n)} = \mathbf{P}(X_n = j | X_0 = i),$$

and also

$$p_{ij}^{(n)} = \sum_{k=1}^{m} \mathbf{P}(X_n = j, X_{n-1} = k | X_0 = i)$$

for $n \geq 2$, since the chain must have passed through one of all the m possible states at step $n - 1$.

For any three events A, B, and C, we have available the identity

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|B \cap C)\mathbf{P}(B|C)$$

Interpreting A as $X_n = j$, B as $X_{n-1} = k$, and C as $X_0 = i$,

$$\begin{aligned} p_{ij}^{(n)} &= \mathbf{P}(A \cap B|C) = \mathbf{P}(X_n = j, X_{n-1} = k|X_0 = i) \\ &= \sum_{k=1}^{m} \mathbf{P}(X_n = j|X_{n-1} = k, X_0 = i)\mathbf{P}(X_{n-1} = k|X_0 = i) \\ &= \sum_{k=1}^{m} \mathbf{P}(X_n = j|X_{n-1} = k)\mathbf{P}(X_{n-1} = k|X_0 = i) \\ &= \sum_{k=1}^{m} p_{kj}^{(1)} p_{ik}^{(n-1)}, \end{aligned}$$

using the Markov property again. These are known as the **Chapman–Kolmogorov equations**

Putting n successively equal to $2, 3, \ldots$, we find that the matrices with

these elements are, using the product rule for matrices,

$$\left[p_{ij}^{(2)}\right] = \left[\sum_{k=1}^{m} p_{ik}^{(1)} p_{kj}^{(1)}\right] = T^2,$$

$$\left[p_{ij}^{(3)}\right] = \left[\sum_{k=1}^{m} p_{ik}^{(2)} p_{kj}^{(1)}\right] = T^2 T = T^3,$$

since $p_{ik}^{(2)}$ are the elements of T^2, and so on. Generalising this rule,

$$[p_{ij}^{(n)}] = T^n.$$

EXAMPLE 5.1:

In a certain region the weather patterns have the following sequence. A day is described as sunny (S) if the sun shines for more than 50% of daylight hours and cloudy (C) if the sun shines for less than 50% of daylight hours. Data indicate that if it is cloudy one day then it is equally likely to be cloudy or sunny on the next day; if it is sunny there is a probability $\frac{1}{3}$ that it is cloudy and $\frac{2}{3}$ that it is sunny the next day.
(i) Construct the transition matrix T for this process.
(ii) If it is cloudy today, what are the probabilities that it is (a) cloudy, (b) sunny, in three days' time?
(iii) Compute T^5 and T^{10}. How do you think that T^n behaves as $n \to \infty$? How does $\mathbf{p}^{(n)}$ behave as $n \to \infty$? Do you expect the limit to depend on $\mathbf{p}^{(0)}$?

(i) It is assumed that the process is Markov and homogeneous so that transition probabilities depend only on the state of the weather on the previous day. This is a two-state Markov chain with states

$$E_1 = (\text{weather cloudy}, C), \qquad E_2 = (\text{weather sunny}, S).$$

The transition probabilities can be represented by the table below which defines the transition matrix T:

	C	S
C	$\frac{1}{2}$	$\frac{1}{2}$
S	$\frac{1}{3}$	$\frac{2}{3}$

or $$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

The actual transition probabilities are

$$p_{11} = \tfrac{1}{2}, \quad p_{12} = \tfrac{1}{2}, \quad p_{21} = \tfrac{1}{3}, \quad p_{22} = \tfrac{2}{3}.$$

(ii) Measuring steps from today, we define

$$\mathbf{p}^{(0)} = \begin{bmatrix} p_1^{(0)} & p_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

which means that it is cloudy today. In three days' time,

$$\begin{aligned} \mathbf{p}^{(3)} = \mathbf{p}^{(0)} T^3 &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}^3 \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 29/72 & 43/72 \\ 43/108 & 65/108 \end{bmatrix} \\ &= \begin{bmatrix} 29/72 & 43/72 \end{bmatrix} = \begin{bmatrix} 0.403 & 0.600 \end{bmatrix}. \end{aligned}$$

Hence the probabilities of cloudy or sunny weather in three days' time are respectively:

(a) $p_1^{(3)} = 29/72$
(b) $p_2^{(3)} = 43/72$.
(iii) The computed values of T^5 and T^{10} are (to 6 decimal places):

$$T^5 = \begin{bmatrix} 0.400077 & 0.599923 \\ 0.399949 & 0.600051 \end{bmatrix}, \qquad T^{10} = \begin{bmatrix} 0.400000 & 0.600000 \\ 0.400000 & 0.600000 \end{bmatrix}.$$

Powers of matrices can be easily computed using software such as *R* or *Mathematica*. It appears that

$$T^n \to \begin{bmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix} = Q,$$

$$\mathbf{p}^{(n)} = \mathbf{p}^{(0)} T^n.$$

If $T^n \to Q$ as $n \to \infty$, then we might expect

$$\begin{aligned} \mathbf{p}^{(n)} \to \mathbf{p}^{(0)} Q &= \begin{bmatrix} p_1^{(0)} & p_2^{(0)} \end{bmatrix} \begin{bmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix} \\ &= \begin{bmatrix} (p_1^{(0)} + p_2^{(0)})0.4 & (p_1^{(0)} + p_2^{(0)})0.6 \end{bmatrix} \\ &= \begin{bmatrix} 0.4 & 0.6 \end{bmatrix} \end{aligned}$$

since $p_1^{(0)} + p_2^{(0)} = 1$. Note that $\lim_{n\to\infty} \mathbf{p}^n$ is independent of $\mathbf{p}^{(0)}$. The limit indicates that, in the long run, 40% of days are cloudy and 60% are sunny.

Classification of states

Let us return to the general m-state chain with states $E_1, E_2, \ldots, E_m$ and transition matrix

$$T = [p_{ij}], \qquad (1 \le i, j \le m).$$

For a homogeneous chain, recollect that p_{ij} is the probability that a transition occurs between E_i and E_j at any step or change of state in the chain. We intend to investigate and classify some of the more common types of states which can occur in Markov chains. This will be a brief treatment, using mainly examples of what is an extensive algebraic subject.

(a) Absorbing state

We have already met one type of state—namely the **absorbing state** Once entered there is no escape from an absorbing state. An absorbing state E_i is characterized by the probabilities

$$p_{ii} = 1, \qquad p_{ij} = 0, (j \neq i, \quad j = 1, 2, \ldots m),$$

in the i-th row of T.

(b) **Periodic state**

The probability of a return to E_i at step n is $p_{ii}^{(n)}$. Let t be an integer greater than 1. Suppose that

$$p_{ii}^{(n)} = 0 \text{ for } n \neq t, 2t, 3t, \ldots$$
$$p_{ii}^{(n)} \neq 0 \text{ for } n = t, 2t, 3t, \ldots.$$

In this case the state E_i is said to be **periodic** with period t. If, for a state, no such t exists with this property, then the state is described as **aperiodic**.

Let

$$d(i) = \gcd\{n | p_{ii}^{(n)} > 0\},$$

that is, the greatest common divisor (gcd) of the set of integers n for which $p_{ii}^{(n)} > 0$. Then the state E_i is said to be **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$. This definition, which includes the earlier case, covers the appearance of *delayed* periodicity in which a sequence of initial terms are zero

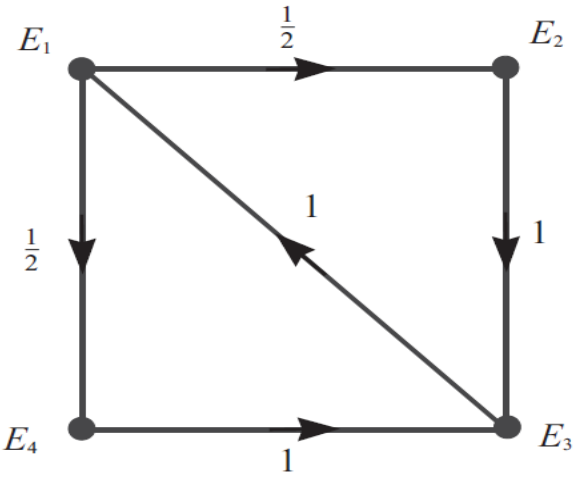
EXAMPLE 5.2:

A four-state Markov chain has the transition matrix

$$T = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Show that all states have period 3.

The transition diagram is-



The analysis of chains with periodic states can be complicated. However, one can check for a suspected periodicity as follows. By direct computation

$$S = T^3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

In this example,

$$S^2 = T^6 = SS = S,$$

so that

$$S^r = T^{3r} = S, \qquad (r = 1, 2, \ldots),$$

which always has nonzero elements on its diagonal. On the other hand,

$$S^{r+1} = S^r S = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad S^{r+2} = S^r S^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

and both these matrices have zero diagonal elements for $r = 1, 2, 3, \ldots$. Hence, for $i = 1, 2, 3, 4$,

$$p_{ii}^{(n)} = 0 \text{ for } n \neq 3, 6, 9, \ldots,$$
$$p_{ii}^{(n)} \neq 0 \text{ for } n = 3, 6, 9, \ldots,$$

which means that all states are period 3. In this example

$$d(i) = \gcd\{3, 6, 9, \ldots\} = 3, \text{ for } i = 1, 2, 3, 4.$$

(c) **Persistent state**

Let $f_j^{(n)}$ be the probability that the **first return** or **visit** to E_j occurs at the n-th step. This probability is not the same as $p_{jj}^{(n)}$, which is the probability that a return occurs at the n-th step, and includes possible returns at steps $1, 2, 3, \ldots, n-1$ also. It follows that

$$p_{jj}^{(1)} (= p_{jj}) = f_j^{(1)},$$
$$p_{jj}^{(2)} = f_j^{(2)} + f_j^{(1)} p_{jj}^{(1)},$$
$$p_{jj}^{(3)} = f_j^{(3)} + f_j^{(1)} p_{jj}^{(2)} + f_j^{(2)} p_{jj}^{(1)},$$

and, in general,

$$p_{jj}^{(n)} = f_j^{(n)} + \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)} \qquad (n \geq 2).$$

$$f_j^{(1)} = p_{jj},$$

$$f_j^{(n)} = p_{jj}^{(n)} - \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)} \qquad (n \geq 2).$$

The probability that a chain returns at some step to the state E_j is

$$f_j = \sum_{n=1}^{\infty} f_j^{(n)}.$$

If $f_j = 1$, then a return to E_j is certain, and E_j is called a **persistent state**.

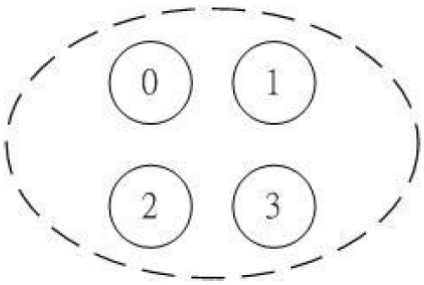
Ergodic :

- Recurrent states that are aperiodic are called ergodic states.
- A Markov chain is said to be ergodic if all its states are ergodic.

- An inventory example :
  - The process is irreducible and ergodic and therefore, has steady-state probability.

P =

state	0	1	2	3
0	0.080	0.184	0.368	0.368
1	0.632	0.368	0.000	0.000
2	0.264	0.368	0.368	0.000
3	0.080	0.184	0.368	0.368



- First Passage time :
  - The first passage time from state i to state j is the number of transitions made by the process in going from state i to state j for the first time.

- Recurrence time :
  - When j = i, the first passage time is just the number of transitions until the process returns to the initial state i and called the recurrence time for state i.

- Example :
  - $X_0 = 3, X_1 = 2, X_2 = 1, X_3 = 0, X_4 = 3, X_5 = 1$
  - The first passage time from state 3 to state 1 is 2 weeks.
  - The recurrence time for state 3 is 4 weeks.

- $f_{ij}^{(n)}$:
  - denotes the probability that the first passage time from state i to state j is n.

- Recursive relationship :

$$f_{ij}^{(n)} = \sum_{k \neq j} p_{ik} f_{kj}^{(n-1)}$$

$$f_{ij}^{(1)} = p_{ij}^{(1)} = p_{ij}$$

$$f_{ij}^{(2)} = \sum_{k \neq j} p_{ik} f_{kj}^{(1)}$$

- The inventory example :
  - $f_{30}^{(1)} = p_{30} = 0.080$
  - $f_{30}^{(2)} = p_{31} f_{10}^{(1)} + p_{32} f_{20}^{(1)} + p_{33} f_{30}^{(1)}$
    $= 0.184(0.632) + 0.368(0.264) + 0.368(0.080) = 0.243$
  - … …

- Sum :

$$\sum_{n=1}^{\infty} f_{ij}^{(n)} \leqq 1$$

- Expected first passage time :

- $\mu_{ij} = \begin{cases} \infty & \text{if } \sum_{n=1}^{\infty} f_{ij}^{(n)} < 1 \\ \sum_{n=1}^{\infty} n f_{ij}^{(n)} & \text{If } \sum_{n=1}^{\infty} f_{ij}^{(n)} = 1 \end{cases}$

- $$\mu_{ij} = 1 + \sum_{k \neq j} p_{ik} \mu_{ki}$$

- The inventory example :
  - $\mu_{30} = 1 + p_{31}\mu_{10} + p_{32}\mu_{20} + p_{33}\mu_{30}$
  - $\mu_{20} = 1 + p_{21}\mu_{10} + p_{22}\mu_{20} + p_{23}\mu_{30}$
  - $\mu_{10} = 1 + p_{11}\mu_{10} + p_{12}\mu_{20} + p_{13}\mu_{30}$

  - μ_{10} = 1.58 weeks, μ_{20} = 2.51 weeks, μ_{30} = 3.50 weeks

- Absorbing states :
  - A state k is called an absorbing state if $p_{kk} = 1$, so that once the chain visits k it remains there forever.

- An gambling example :
  - Suppose that two players (A and B), each having $2, agree to keep playing the game and betting $1 at a time until one player is broke. The probability of A winning a single bet is 1/3.

- The transition matrix form A's point of view

$$\mathbf{P} = \begin{array}{c|ccccc} \text{state} & 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2/3 & 0 & 1/3 & 0 & 0 \\ 2 & 0 & 2/3 & 0 & 1/3 & 0 \\ 3 & 0 & 0 & 2/3 & 0 & 1/3 \\ 4 & 0 & 0 & 0 & 0 & 1 \end{array}$$

- Probability of absorption :
  - If k is an absorbing state, and the process starts in state i, the probability of ever going to state k is called the probability of absorption into state k, given the system started in state i.

$$f_{ik} = \sum_{j=0}^{M} p_{ij} f_{jk} \qquad \text{for i = 0, 1, 2, ..., M}$$

subject to the conditions
$f_{kk} = 1$,
$f_{ik} = 0$, if state i is the recurrent and $i \neq k$.

- The gambling example :
  - $f_{20} = 4/5$, $f_{24} = 1/5$