University of Tamanghasset, Amine Elokkal El Hadj Moussa Eghamouk, Algeria

Faculty of Sciences and Technology

LECTURE NOTES

Introduction to Descriptive statistics and probability

Editing :
Dr. Thiziri SIFAOUI

Examiners
Dr. Arsalane Chouaib Guidoum
Dr. Djamel Boudaa

Academic year 2023/2024

Table of Contents

Introduction

Welcome to the world of statistics, probability, and combinatorics! In this course, you'll learn how data can give us useful information, how to measure uncertainty, and how to count and organize different possibilities. This course is designed for first-year students in mathematics and computer science. Statistics helps us collect, understand, and use data to make good decisions, even when we don't know everything. Combinatorics shows us how to count and arrange different outcomes, which is important for solving many problems. Probability helps us figure out how likely things are to happen in different situations.

We'll start by learning how to collect and present data, understand basic counting methods, and explore the basics of probability. You'll gain the tools you need to work with data and uncertainty effectively. Our expedition will traverse a spectrum of topics, including:

1. Vocabulary of Statistics and Basic Concepts
2. Methods of Data Collection and Presentation.
3. Measures of Central Tendency.
4. Measures of Dispersion.
5. Combinatorial Analysis.
6. Probability Space.

Upon conclusion of this course, you will not only possess a firm grasp of statistical principles but also wield the proficiency to apply them adeptly to real-world predicaments. Prepare to embark on an enthralling odyssey into the domain of statistics and probability, where data unveils its mysteries, and uncertainty is harnessed through the formidable prowess of mathematics and computation.

Chapter 1

Vocabulary of Statistics and Basic Concepts

1.1 Introduction

Statistics can be defined as the mathematical science involving the collection, analysis, interpretation, and presentation of data. It encompasses methods for gathering data, organizing it into meaningful patterns, summarizing it, and drawing conclusions or making predictions based on the data.

Statistics classifications include:

1. **Descriptive Statistics:** Descriptive statistics involves methods for summarizing and describing the main features of a dataset. It includes measures such as mean, median, mode, range, variance, and standard deviation. Descriptive statistics provide insights into the characteristics of the data without making any inferences beyond the dataset itself.

2. **Inferential Statistics:** Inferential statistics involves making inferences or predictions about a population based on a sample of data drawn from that population. It includes techniques such as hypothesis testing, confidence intervals, regression analysis, and analysis of variance (ANOVA). Inferential statistics allows researchers to generalize findings from a sample to a larger population.

3. **Parametric Statistics:** Parametric statistics assumes that the data follows a specific probability distribution, usually the normal distribution. Parametric tests require certain assumptions to be met, such as the data being normally distributed and having homogeneity of variance. Examples of parametric tests include t-tests, ANOVA, and linear regression.

4. **Non-parametric Statistics:** Non-parametric statistics do not rely on specific assumptions about the underlying probability distribution of the

data. These methods are often used when data do not meet the assumptions of parametric tests, such as when the data is skewed or contains outliers. Non-parametric tests include the Wilcoxon signed-rank test, Mann-Whitney U test, and Kruskal-Wallis test.

5. **Quantitative Statistics:** Quantitative statistics deals with numerical data that can be measured and expressed using numbers. It includes techniques for analyzing variables such as height, weight, temperature, and income.

6. **Qualitative Statistics:** Qualitative statistics deals with non-numerical data that can be categorized or described based on attributes or characteristics. It includes techniques for analyzing categorical variables such as gender, ethnicity, occupation, and political affiliation.

These classifications provide a framework for understanding the various approaches and techniques used in statistical analysis. In this context, our focus is on Descriptive Statistics, which encompass Quantitative and Qualitative Statistics.

1.2 Stages in Statistical Investigation

1. **Formulating the Problem:**
   - Clearly define the research question or problem to be investigated.
   - Specify the objectives and scope of the investigation.
2. **Collecting Data:**
   - Choose appropriate data collection methods (surveys, experiments, etc.).
   - Gather relevant data sources and ensure data quality.
3. **Exploratory Data Analysis (EDA):**
   - Explore the data using summary statistics, visualization, and graphical analysis.
   - Identify patterns, outliers, and initial insights.
4. **Formulating Hypotheses:**
   - Based on EDA, formulate specific hypotheses or research questions.
   - Clearly state null and alternative hypotheses.
5. **Choosing the Statistical Methods:**
   - Select appropriate statistical techniques for analysis (parametric tests, regression, etc.).

- Consider assumptions and conditions required for chosen methods.

6. **Data Analysis:**
   - Apply chosen statistical methods to analyze the data.
   - Calculate test statistics, perform hypothesis tests, and estimate parameters.
7. **Interpreting Results:**
   - Interpret the statistical findings in the context of the research question.
   - Assess the significance of results and implications for hypotheses.

8. **Drawing Conclusions:**
   - Draw conclusions based on the interpretation of results.
   - Discuss the practical implications and limitations of the findings.
9. **Communicating Findings:**
   - Prepare reports, presentations, or academic papers to communicate the findings.
   - Tailor the communication to the audience and convey the key insights effectively.

1.3 Statistical Vocabulary

1. **Population:** The complete set of elements or individuals on which a statistical study is conducted.e.g. All registered voters in a country.
2. **Sample:** A subset of the population that is studied to draw conclusions about the larger population. e.g. A group of 1000 registered voters randomly selected from the population.
3. **Individual:** A unique element in a population, often used interchangeably with the term "unit" or "observation".e.g. Farid is an individual voter in the population.
4. **Characteristic:** A specific feature studied within a population or sample, synonymous with "variable".e.g. Income level is a characteristic studied in a population survey.
5. **Frequency:** The number of occurrences of a particular value in a data set.e.g. There are 150 voters in the sample aged between 30 and 40 years.
6. **Statistical series:** An organized data set listing the values taken by a variable, often accompanied by their frequencies.e.g. A statistical series lists the ages of 1000 voters in the sample along with the number of voters in each age group.
7. **Independent variable:** A variable that is manipulated or controlled in a study to see how it affects another variable.e.g. The amount of advertising expenditure in a marketing study.
8. **Dependent variable:** A variable whose values are influenced by other variables, often the independent variable in a study.e.g. Sales revenue in a marketing study.
9. **Mean:** The measure of central tendency that represents the average value of a set of data.e.g. The average age of the voters in the sample studied.

10. **Median:** The middle value in a set of data when arranged in ascending or descending order.e.g. The middle age of the voters in the sample is 42 years.

11. **Mode:** The value that appears most frequently in a set of data. e.g. The most common age among the voters in the sample is 38 years.

12. **Standard deviation:** A measure of the dispersion or variability of values in a set of data. e.g. The age distribution of the voters in the sample has a standard deviation of 10 years.

13. **Distribution:** The arrangement of values of a variable in a data set. e.g. The age distribution of the voters in the sample is normal.

14. **Correlation:** The measure of the relationship between two variables. e.g. There is a positive correlation between advertising expenditure and sales revenue.

15. **Regression:** A statistical technique used to study the relationship between a dependent variable and one or more independent variables. e.g. A regression analysis shows that advertising expenditure predicts 80% of the variance in sales revenue.

16. **Probability:** The measure of the likelihood that an event will occur, often expressed as a number between 0 and 1. e.g. The probability of rolling a six on a fair six-sided die is 1/6.

17. **Random sampling:** The process of selecting a sample in such a way that each individual or unit in the population has an equal chance of being included. e.g. Each registered voter in the country has an equal chance of being selected for the sample.

18. **Hypothesis testing:** A statistical method used to assess the validity of a statement about a population based on sample data. e.g. A study tests the hypothesis that there is no difference in voting preferences between males and females based on survey data.

19. **Types of Variables or Data:**

    In statistics, variables or data can broadly be categorized into two main types: qualitative (or categorical) (nominal or ordinal) and quantitative (or numerical) (discrete or continuous).

    - **Qualitative (or Categorical) Variables:**
      - **Nominal Variables:** These variables represent categories or groups with no inherent order or ranking. Examples include:
        * Gender (male, female, other)
        * Eye color (blue, brown, green)
        * Types of fruits (apple, banana, orange)

- **Ordinal Variables:** Unlike nominal variables, ordinal variables have a natural order or ranking. Examples include:
  * Education level (high school, college, graduate)
  * Economic status (low income, middle income, high income)
  * Satisfaction level (low, medium, high)

- **Quantitative (or Numerical) Variables:**
  - **Discrete Variables:** Discrete variables can only take on specific values, often whole numbers, and cannot be subdivided into smaller units. Examples include:
    * Number of children in a family (1, 2, 3, ...)
    * Number of cars in a parking lot
  - **Continuous Variables:** Continuous variables can take on any value within a given range and can be subdivided into smaller units. These variables are often measured rather than counted. Examples include:
    * Height (e.g., 170 cm, 170.5 cm, 171 cm, ...)
    * Weight
    * Temperature
    * Time

Understanding the type of variable is crucial for choosing appropriate statistical analyses and interpreting the results accurately.

1.4 Applications, Uses and Limitations of statistics

Statistics is a powerful tool with a wide range of applications across various fields. Here are some common applications, uses, and limitations of statistics:

1. **Applications and Uses of Statistics:**

   - Research and Data Analysis: Statistics is extensively used in scientific research to analyze data, test hypotheses, and draw conclusions. It helps researchers make sense of complex data sets and identify patterns or relationships.
   - Business and Economics: Statistics plays a crucial role in business and economics for market research, forecasting demand, analyzing financial data, and making informed business decisions. It helps businesses optimize processes, manage risks, and identify trends.
   - Quality Control and Process Improvement: In manufacturing and industry, statistics are used for quality control to monitor production processes, detect defects, and ensure product consistency. Statistical methods like Six Sigma are employed to improve process efficiency and reduce variability.

- Healthcare and Medicine: Statistics is vital in healthcare for clinical trials, epidemiological studies, patient outcome's analysis, and public health research. It helps healthcare professionals make evidence-based decisions, assess treatment effectiveness, and identify health trends.
- Social Sciences: Statistics is applied in social sciences such as sociology, psychology, and political science to analyze survey data, study human behavior, and examine social phenomena. It helps researchers understand population trends, attitudes, and demographics.
- Environmental Studies: Statistics is used to analyze environmental data, assess environmental risks, and monitor ecological changes. It helps scientists study climate patterns, biodiversity, pollution levels, and natural resource management.
- Education: In education, statistics are used for educational research, assessment, and evaluation. It helps educators measure student performance, evaluate teaching methods, and identify areas for improvement in educational systems.

2. **Limitations of Statistics:**

- Data Quality Issues: Statistics relies on the quality and accuracy of data. Poorly collected or biased data can lead to inaccurate results and flawed conclusions.
- Sampling Errors: In sampling, there is a risk of sampling errors where the characteristics of the sample may not accurately represent the entire population, leading to incorrect inferences.
- Assumption Violations: Many statistical methods are based on certain assumptions about the data. If these assumptions are violated, the results may be unreliable.
- Causation vs. Correlation: Statistical analysis can establish correlations between variables, but it cannot always determine causation. Correlation does not imply causation, and other factors may be influencing the observed relationship.
- Interpretation Challenges: Statistical results can be misinterpreted or misunderstood, especially by those without a strong statistical background. Misinterpretation can lead to erroneous conclusions and decisions.
- Ethical Considerations: The use of statistics raises ethical concerns related to privacy, confidentiality, and data manipulation. Ethical guidelines must be followed to ensure the responsible and ethical use of statistics.
- Complexity and Misleading Statistics: Statistics can be complex, and misleading statistics or misuse of statistical methods can lead to misconceptions or misrepresentation of data, impacting decision-making and public perception.

Remark 1. *While statistics has numerous applications and can provide valuable insights, it is essential to recognize its limitations and exercise caution in its interpretation and use. Proper understanding, careful analysis, and critical thinking are crucial for effectively utilizing statistics in various fields.*

1.5 Exercises

1. In a study conducted in Tamanrasset to assess the level of satisfaction with local healthcare services, researchers randomly select households to survey. They visit 500 households scattered across the city. In this study, what is the statistics term for THE PEOPLE OF TAMANRASSET, and what is the statistics term for THE HOUSEHOLDS VISITED?

2. When an investigator uses data, which have already been collected by others, such data is called?

3. What technique is used to ensure that a sample is representative of a population?

4. Statistical methods may be categorized into

5. A researcher is gathering data from four Senatorial zones, designated Tamanrasset =1; Algiers =2; Oran = 3; Constantine = 4. The designated Senatorial zones represent what type of data?

A small-scale study in which all the operations intended to be used in the main study are used is called ?

6. The characteristic feature possessed by the units of the population that change during the period under consideration is called?

7. Memory lapse is a source of error in data collection. True/False?

8. Data on gender is categorized as

9. Indicate whether the following variables are qualitative or quantitative:

(i) Favorite food.
(ii) Favorite profession.
(iii) Number of goals scored by FC Barcelona last season.
(iv) Number of students in the University of Tamanghasset.
(v) The hair color of your course mates.
(vi) Level of comprehension in the class.
(vii) Mode of transportation used by students to commute to campus.
(viii) Favorite movie genre.

10. Indicate whether the following variables are discrete or continuous:

(i) Number of olives harvested daily in the Tizi Ouzou region of Algeria.
(ii) Hourly temperatures recorded at Algiers Observatory.
(iii) Lifetime of a traditional Algerian carpet.
(iv) The diameter of the wheels of Algerian-made cars.

(v) Number of children from 50 Algerian families.

(vi) Annual Census of Algerian citizens.

11. Classify the following variables as qualitative, quantitative discrete or continuous.

(i) The nationality of a person.

(ii) Number of liters of water contained in a tank.

(iii) Number of books on a library shelf.

(iv) Sum of points tallied from a set of dice.

(v) The profession of a person.

(vi) The area of the different tiles on a building.

12. Analysis of labor turnover rates, performance appraisal, training programs and planning of incentives are examples of role of

A. statistics in personnel management

B. statistics in finance

C. statistics in marketing

D. statistics in production

13. Focus groups, individual respondents and panels of respondents are classified as

A. pointed data sources

B. itemized data sources

C. secondary data sources

D. primary data sources

14. Variables whose measurement is done in terms such as weight, height and length are classified as

A. continuous variables

B. measuring variables

C. flowchart variables

D. discrete variables

15. One of the following is not an example of Ordinal data

A. rank

B. volume

C. statistics grade

D. satisfaction level

16. Numerical methods and graphical methods are specialized procedures used in

A. inferential statistics
B. military statistics
C. descriptive statistics
D. education statistics

17. Scale used in statistics which provides difference of proportions as well as magnitude of differences is considered as

A. satisfactory scale
B. ratio scale
C. goodness scale
D. exponential scale

18. Sample statistics are denoted by the

A. upper case Greek letter
B. associated roman alphabets
C. roman letters
D. lower case Greek letter

19. Which of the following points do not reflect statistics?

A. It is a while subject of study
B. They can be inferential
C. It describes methods of collecting, quantitative data
D. It describes ways of analyzing qualitative data

20. Which of the following is a statistic?

A. Sample mean
B. Population variance
C. None of these
D. Population mean

21. What name is given to data which can be ranked?

A. Categorical data
B. Ordinal data
C. Interval data
D. Ratio data

22. What name is given to data which is on a continuous scale with a neutral zero?

A. Interval data
B. Ranked data
C. Ratio data
D. Ordinal data

23. What is the first stage in statistics?

A. Analyze data
B. Collect data
C. Organize data
D. Identify the group of people to be studied

24. Which of these is an example of a categorical variable?

A. flavor of soft drink ordered by each customer at a fast food restaurant
B. height, measured in inches, for each student in a class
C. points scored by each player on a team
D. color of a car

25. Numerical and pictorial information about variables are called

A. analytical statistics
B. descriptive statistics
C. inferential statistics
D. parametric statistics

26. The entire group of interest for a statistical conclusion is called the

A. data
B. population
C. sample
D. statistic

27. A subgroup that is representative of a population is called

A. a category
B. data
C. a sample
D. census

28. Statistical inference is

A. the process of estimates and conclusions carefully based on data from a sample

B. the process of estimates and conclusions carefully based on data from an entire population

C. pictorial displays that summarize data

D. tabulation of data.

29. Two types of statistical variables are

A. categorical and descriptive

B. categorical and numerical

C. descriptive and numerical

D. constant and numerical

Chapter 2

Methods of Data Collection and Presntation

2.1 Introduction to methods of data collection

There are two sources of data:

2.1.1 Primary Data:

Primary data involves the direct measurement or collection of data from its original source. This process typically encompasses two key activities: planning and measuring.

2.1.1.1 Planning:

- **Identify Source and Elements:** Begin by identifying the source of the data and the specific elements to be collected. This helps in ensuring the relevance and accuracy of the collected information.

- **Sampling Decision:** Determine whether to conduct a sample or census. Sampling involves selecting a subset of the population for study, while a census involves gathering data from the entire population. Factors such as resource constraints and population characteristics influence this decision.

- **Sample Size and Selection Method:** If sampling is chosen, determine the appropriate sample size and selection method. Various techniques, such as random sampling, stratified sampling, or cluster sampling, may be employed based on the research objectives and population characteristics.

- **Measurement Procedure:** Define the measurement procedures to be used for data collection. This includes specifying the instruments, scales,

and techniques for gathering the required information.

- **Organizational Structure:** Establish the necessary organizational structure for data collection, including assigning roles and responsibilities to team members, coordinating logistics, and ensuring compliance with ethical guidelines.

2.1.1.2 Measuring:

There are several options available for measuring primary data:

- **Focus Groups:** Small group discussions facilitated by a moderator to gather insights on specific topics or issues.
- **Telephone Interviews:** Conducting interviews over the phone, allowing for efficient data collection from geographically dispersed respondents.
- **Mail Questionnaires:** Sending questionnaires via mail to respondents, providing them with time to consider their responses.
- **Door-to-Door Surveys:** Visiting respondents' homes or locations to administer surveys in person.
- **Mall Intercept:** Conducting surveys or interviews with individuals in a mall or public place.
- **New Product Registration:** Gathering data through registration forms when individuals purchase or use a new product or service.
- **Personal Interviews:** Face-to-face interviews conducted by trained interviewers, allowing for in-depth exploration of responses.
- **Experiments:** Controlled studies designed to test hypotheses or evaluate the impact of interventions under controlled conditions.

2.1.2 Secondary Data:

The data utilized is sourced from published and unpublished sources or files. When using secondary data, it's crucial to consider the following:

- Assess the relevance of the data's type and objective to the current situation.
- Verify alignment between the original data collection purpose and the present problem.

- Ensure appropriateness of data character and classification for addressing the problem.

- Scrutinize for biases and misreporting in published sources to maintain integrity.

Remark 2. *Data considered primary by one party may be regarded as secondary by another, depending on their respective perspectives or needs.*

2.2 Methods of data presentation

After collecting and editing the data, the crucial next step involves organizing it effectively. This entails presenting the information in a condensed format that is easily understandable, facilitating the extraction of meaningful insights. Additionally, it is imperative to separate similar data points from dissimilar ones to enhance clarity and coherence in the analysis process. The presentation of data is broadly classified into the following two categories:

- **Visual Presentation:** This category involves representing data using graphical elements such as charts, graphs, maps, and diagrams. Visual presentations are effective for conveying trends, patterns, and comparisons in a clear and intuitive manner.

- **Numerical Presentation:** This category involves presenting data in numerical or tabular form. Numerical presentations include tables, spreadsheets, and lists, providing detailed information and allowing for precise analysis and calculations.

2.2.1 Classification

Classification is the process of arranging data into classes or categories based on similarities. It's a preliminary step that sets the stage for the proper presentation of data.

Definition

- **Raw data:** Original recorded information, whether counts or measurements.
- **Frequency:** The number of values in a specific class of the distribution.
- **Frequency distribution:** The organization of raw data in table form using classes and frequencies.

There are three basic types of frequency distributions:

- **Categorical frequency distribution:** Used for data that can be placed into specific categories, such as nominal or ordinal data.

- **Ungrouped frequency distribution:** A way of organizing raw data into categories or classes along with their corresponding frequencies. Unlike a grouped frequency distribution where data is organized into intervals, in an ungrouped frequency distribution, each individual data point is listed separately along with its frequency.

- **Grouped frequency distribution:** A way to organize and display data by grouping individual data points into intervals or classes, and then counting how many data points fall into each interval. This method is useful when dealing with a wide range of values or continuous data.

2.2.2 Key Terms in Frequency Distributions:

- **Class Limits:** The boundaries that separate one class in a grouped frequency distribution from another. These limits could actually appear in the data and may have gaps between the upper limits of one class and the lower limit of the next.

- **Units of Measurement (U):** The distance between two possible consecutive measures. It is typically represented as 1, 0.1, 0.01, 0.001, and so on.

- **Class Boundaries:** The boundaries that separate one class in a grouped frequency distribution from another. These boundaries have one more decimal place than the raw data and therefore do not appear in the data. There is no gap between the upper boundary of one class and the lower boundary of the next class. The lower class boundary is found by subtracting $U/2$ from the corresponding lower class limit, and the upper class boundary is found by adding $U/2$ to the corresponding upper class limit.

- **Class Width:** The difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes, or the difference between any two consecutive class marks.

- **Class Mark (Midpoints):** The average of the lower and upper class limits, or the average of the upper and lower class boundaries.

- **Cumulative Frequency:** The number of observations less than/more than or equal to a specific value.

- **Cumulative Frequency Above:** The total frequency of all values greater than or equal to the lower class boundary of a given class.

- **Cumulative Frequency Below:** The total frequency of all values less than or equal to the upper class boundary of a given class.

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

- **Relative Frequency (rf):** The frequency divided by the total frequency.

- **Relative Cumulative Frequency (rcf):** The cumulative frequency divided by the total frequency.

2.2.3 Steps for Constructing Grouped Frequency Distribution:

1. Find the largest and smallest values.

2. Determine the range of the data: Range = Maximum value - Minimum value.

3. Choose the number of desired classes, typically ranging from 5 to 20. Alternatively, apply Sturges' rule: $k = 1 + 3.32 \log_{10}(n)$, where k represents the desired number of classes and n denotes the total number of observations.

4. Calculate the class width: Class width = Range / Number of classes.

5. Designate the starting point as the minimum value or a value less than it. This initial value becomes the lower limit of the first class. Then, incrementally add the class width to this starting point to determine the lower limits of subsequent classes.

6. Subtract the class width from the lower limit of the second class. This calculation yields the upper limit of the first class. Subsequently, continue adding the class width to this upper limit to ascertain the upper limits of the subsequent classes in the distribution.

7. Calculate boundaries for class intervals by subtracting $U/2$ from lower limits and adding $U/2$ to upper limits.

8. Tally the data: Count how many data points fall into each interval to determine the frequency for each class.

9. Tabulate the data: Create a table with columns for the intervals and their corresponding frequencies.

2.2.4 Example of Grouped Frequency Distribution:

The dataset represents the ages of 20 people who work in an enterprise

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

Let's construct a grouped frequency distribution for this dataset:

1. Find the highest and the lowest value ($H = 39$ and $L = 6$)

2. Determine the range of the data:

$$\text{Range} = \text{Maximum value} - \text{Minimum value} = 39 - 6 = 33$$

3. Choose the number of classes:

$$\begin{aligned} k &= 1 + 3.32 \log_{10}(n) \\ &= 1 + 3.32 \log_{10}(20) \\ &\approx 5.32 \end{aligned}$$

   Rounding up to the nearest whole number:

$$k \approx 6$$

   So, based on Sturges' rule, the recommended number of classes for this dataset is 6.

4. Calculate the class width:

$$\text{Class width} = \frac{\text{Range}}{\text{Number of classes}} = \frac{33}{6} \approx 5.5$$

   Since the class width should be a convenient number, let's round it up to 6.

5. Select the starting point, ensuring that it is the minimum observation.

$$6, 12, 18, 24, 30, 36$$

   .

6. Find the upper limit of the class; for example the first upper class = 12 - U = 12 - 1 = 11

$$11, 17, 23, 29, 35, 41$$

.

7. So combining 5 and 6, one can construct the following classes:

   - $6 - 11$
   - $12 - 17$
   - $18 - 23$
   - $24 - 29$
   - $30 - 35$
   - $36 - 41$

8. Find the class boundaries:

   For example, for class 1, Lower class limit $= 6 - U/2 = 5.5$, Upper class limit $= 11 + U/2 = 11.5$

   - $5.5 - 11.5$
   - $11.5 - 17.5$
   - $17.5 - 23.5$
   - $23.5 - 29.5$
   - $29.5 - 35.5$
   - $35.5 - 41.5$

9. Create a table with columns for the intervals and their corresponding frequencies:

2.3 Diagrammatic and Graphic Presentation of Data

Visual representation techniques, such as diagrams and graphics, offer effective means to present data in a compelling and easily comprehensible manner. Their

Class limit	Class boundary	Class Mark	Frequency	Cumulative Frequency
6 – 11	5.5 – 11.5	8.5	2	2
12 – 17	11.5 – 17.5	14.5	2	4
18 – 23	17.5 – 23.5	20.5	7	11
24 – 29	23.5 – 29.5	26.5	4	15
30 – 35	29.5 – 35.5	32.5	3	18
36 – 41	35.5 – 41.5	38.5	2	20

significance lies in their ability to attract attention, facilitate comparison, and enhance understanding. Diagrams are particularly suited for depicting discrete data. Among the commonly utilized methods for representing both discrete and qualitative data are pie charts and bar charts.

2.3.1 Pie charts

A pie chart is a circular graphical representation divided into sections or wedges, each corresponding to the percentage of frequencies in a specific category of the distribution. The angle of each sector is determined using the formula:

$$\text{Angle of sector} = \frac{\text{Value of the part}}{\text{The whole quantity}} \times 360^\circ$$

Let's consider representing the population distribution in a town, categorized by gender (Men, Women, Girls, Boys):

Class	Frequency
Men	2500
Women	2000
Girls	4000
Boys	1500

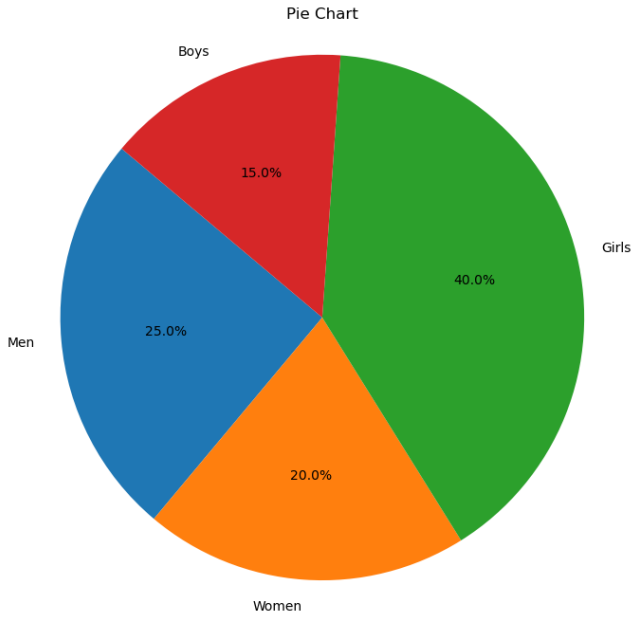
- **Step 1:** Calculate the percentage for each category.

Class	Frequency	Percent
Men	2500	25%
Women	2000	20%
Girls	4000	40%
Boys	1500	15%

- **Step 2:** Find the number of degrees for each category.

- **Step 3:** Using a protractor and compass, draw each section of the pie chart according to its corresponding percentage.

Class	Frequency	Degree
Men	2500	90°
Women	2000	72°
Girls	4000	144°
Boys	1500	54°



2.3.1.1 Bar Charts

A set of bars (thick lines or narrow rectangles) representing some magnitude over time space.

- They are useful for comparing aggregates over time space.
- Bars can be drawn either vertically or horizontally.
- There are different types of bar charts. The most common being:
  - Simple bar chart.
  - Deviation or two-way bar chart.
  - Broken bar chart.
  - Component or subdivided bar chart.

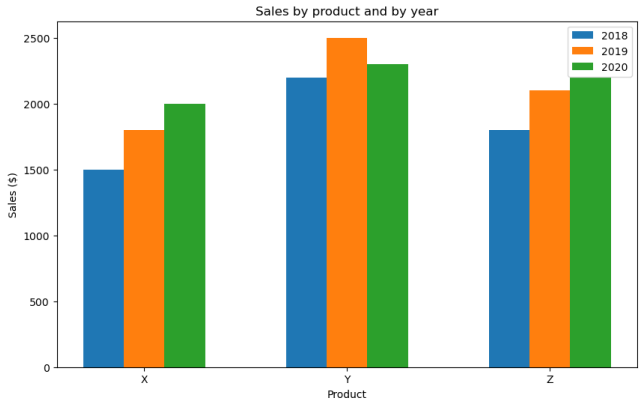
– Multiple bar charts.

2.3.1.2 Simple Bar Chart:

- Are used to display data on one variable.
- They are thick lines (narrow rectangles) having the same breadth. The magnitude of a quantity is represented by the height/length of the bar.

Consider the monthly sales figures (in dollars) of three different products, X, Y, and Z, for the years 2018-2020.

Product	Sales ($) in 2018	Sales ($) in 2019	Sales ($) in 2020
X	1500	1800	2000
Y	2200	2500	2300
Z	1800	2100	2200

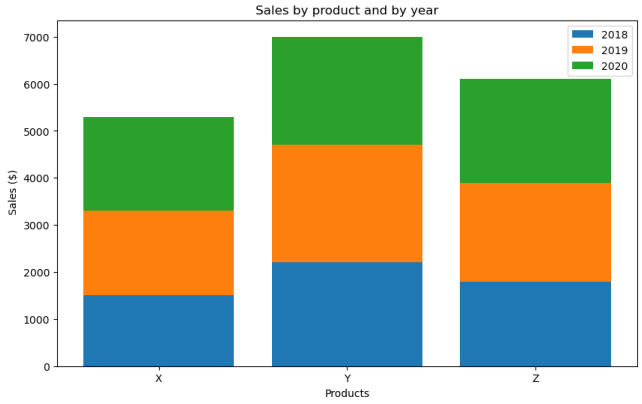


2.3.1.3 Component Bar chart

When there is a desire to show how a total (or aggregate) is divided into its component parts, we use a component bar chart.

- The bars represent the total value of a variable, with each total broken down into its component parts.
- Different colors or designs are used for identifications.

Consider the monthly sales figures (in dollars) of three different products, X, Y, and Z, for the years 2018-2020.



2.3.2 Histogram:

- Draw and label the X and Y axes.
- Choose a suitable scale for the frequencies and label it on the Y-axis.
- Divide the range of the data into intervals called bins.
- Count the number of data points that fall into each bin.
- Draw rectangles (bars) on the X-axis corresponding to each bin, with heights proportional to the frequency of data points in that bin.

2.3.3 Frequency Polygon

- Draw and label the X and Y axes.
- Choose a suitable scale for the frequencies and label it on the Y-axis.
- Plot the midpoints of each interval (bin) on the X-axis.
- Plot the frequency of data points in each bin as points on the graph.
- Connect the points with line segments to form the frequency polygon.

2.3.4 Cumulative Frequency Graph (Ogive)

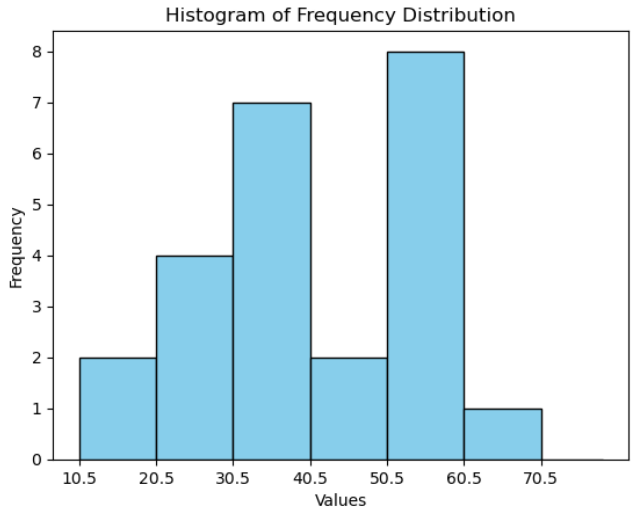
- Draw and label the X and Y axes.

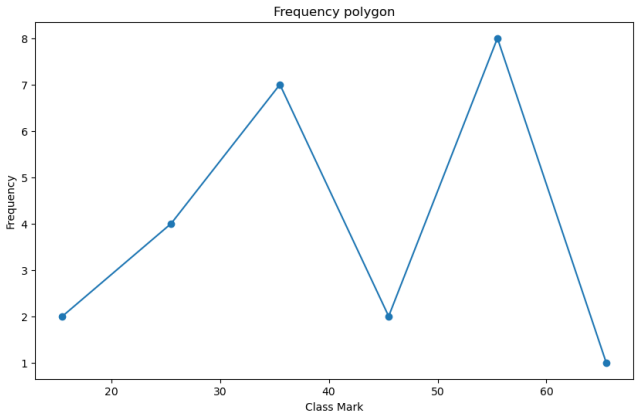
- Choose a suitable scale for the cumulative frequencies and label it on the Y-axis.
- Plot the upper class boundaries of each bin on the X-axis.
- Calculate the cumulative frequency for each bin, which is the sum of frequencies up to that bin.
- Plot the cumulative frequencies as points on the graph.
- Connect the points with line segments to form the Ogive.

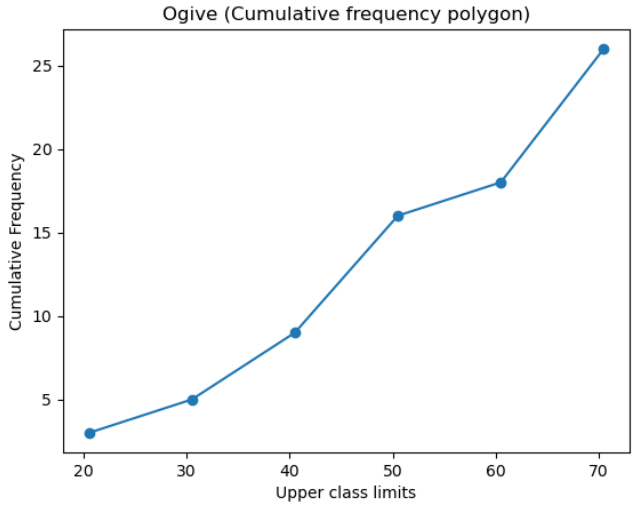
Example.

A researcher wants to study the distribution of ages in a given population. They collect data on the ages of 24 individuals and wish to visualize this data in the form of a histogram, frequency polygon and cumulative frequency graph (Ogive). The collected ages are given in the following tables:

Class limit	Class boundary	Class Mark	Frequency	Cumulative Frequency
11 - 20	10.5 -20.5	15.5	2	2
21 - 30	20.5 -30.5	25.5	4	6
31 - 40	30.5 -40.5	35.5	7	13
41 - 50	40.5 -50.5	45.5	2	15
51 - 60	50.5 - 60.5	55.5	8	23
61 - 70	60.5 - 70.5	65.5	1	24







Remark 3. *The histogram, frequency polygon and cumulative frequency graph or ogive are the most commonly applied graphical representation for continuous data.*

2.4 Exercises

Exercise 1

The following sample data set lists the number of minutes 50 Internet subscribers spent on the internet during their most recent session. Construct a frequency distribution that has seven classes

$$50, 40, 41, 17, 11, 7, 22, 4, 28, 21, 19, 23, 37, 51, 54, 42, 86, 41, 78, 56,$$
$$72, 56, 17, 7, 69, 30, 80, 56, 29, 33, 46, 31, 39, 20, 18, 29, 34, 59, 73, 77,$$
$$36, 39, 30, 62, 54, 67, 39, 31, 53, 44$$

1. What is the population being studied?
2. What is the variable being studied and its nature?
3. Fill in the following table:

Classes	Classes boundary	Frequency	Midpoint	Relative frequency	Cumulative frequency

Exercise 2

The pupils in a class were asked how they got to school

Classes	Frequency
Walk	9
Bike	3
Car	6
Bus	12
Total	30

1. What is the population being studied?
2. What is the variable being studied and its nature?
   Illustrate the data using:
   - bar charts.
   - pie charts.

Exercise 3

The following dataset represents the scores of 25 students in an exam:

$$45, 50, 65, 77, 94, 83, 74, 65, 80, 58, 65,$$
$$90, 88, 68, 55, 79, 83, 72, 74, 76, 81, 53,$$
$$90, 87, 66$$

- What is the population being studied?
- What is the variable being studied and its nature?
- Construct a frequency table using six classes.
- Create a histogram representing the frequency distribution.
- Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
- Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Exercise 4

The following dataset represents the scores of 50 students in an exam:

$$45, 50, 65, 77, 94, 83, 74, 65, 80, 58, 65,$$
$$90, 88, 68, 55, 79, 83, 72, 74, 76, 81, 53,$$
$$90, 87, 66, 70, 78, 62, 91, 85, 76, 79, 92,$$
$$88, 78, 75, 71, 89, 84, 70, 86, 84, 89, 77,$$
$$93, 85, 78, 73, 80, 82, 75$$

- What is the population being studied?
- What is the variable being studied and its nature?
- Construct a frequency table using eight classes.
- Create a histogram representing the frequency distribution.
- Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
- Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Exercise 5

The following dataset represents the monthly sales (in thousands of dollars) for a company over two years:

$$120, 130, 140, 150, 160, 170, 180, 190, 200, 210,$$
$$220, 230, 240, 250, 260, 270, 280, 290, 300, 310,$$
$$320, 330, 340, 350, 360, 370, 380, 390, 400, 410,$$
$$420, 430, 440, 450, 460, 470, 480, 490, 500, 510,$$
$$520, 530, 540, 550, 560, 570, 580, 590, 600$$

- What is the population being studied?
- What is the variable being studied and its nature?
- Construct a frequency table using ten classes.
- Create a histogram representing the frequency distribution.
- Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
- Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Exercise 6

The following dataset represents customer satisfaction scores (on a scale from 1 to 10) for a product:

$$7, 8, 6, 9, 7, 5, 6, 8, 7, 8, 9, 6, 7, 8, 5, 6, 7, 8, 9,$$
$$7, 8, 6, 5, 6, 7, 8, 9, 6, 7, 8, 5, 6, 7, 8, 9, 7, 8, 6, 5,$$
$$6, 7, 8, 9, 7, 8, 6, 5, 6, 7, 8, 9$$

- What is the population being studied?
- What is the variable being studied and its nature?
- Construct a frequency table using six classes.
- Create a histogram representing the frequency distribution.
- Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.

- Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Chapter 3

Measures of Central Tendency

3.1 Introduction

Measures of central tendency serve as fundamental tools in statistical analysis, providing concise summaries of datasets by pinpointing a single value that represents the "center" or "typical" value of a distribution. Often referred to as averages, these measures offer valuable insights into the central tendencies of data, making them easier to understand and interpret.

Whether we're analyzing exam scores, income distributions, or household sizes, measures of central tendency help us grasp the essence of a dataset amidst its complexity. By distilling vast arrays of numbers into a single representative value, they simplify comparisons between different groups, aid in decision-making processes, and form the basis for further statistical analysis.

In this exploration of measures of central tendency, we'll delve into the various types of averages, their characteristics, and their applications in real-world scenarios. From the familiar arithmetic mean to the robust median and the mode representing the most frequent value, each measure offers unique insights into the nature of data, guiding us towards a deeper understanding of the phenomena under study.

3.2 Types of Measures of Central Tendency

Various measures of central tendency, including mean (Arithmetic, Geometric, and Harmonic), mode, median, and quartile, each offer unique strengths and weaknesses. The choice depends on data characteristics and analysis objectives. The mean is sensitive to outliers, while the median is more robust. Quantiles provide insight into distribution spread, and the mode identifies the most frequent values. Selection depends on accurately representing data distribution

and meeting analysis goals.

3.2.1 The Mean

3.2.1.1 Arithmetic Mean

The arithmetic mean is the most commonly used measure of central tendency. It is suitable for data that is symmetrically distributed and does not contain outliers. However, it is sensitive to extreme values. This measure is often employed to find the average of a set of values, such as test scores, heights, or weights.

- **Mean for Ungrouped Data:**

  The mean represents the average value of a set of items, calculated by dividing the sum of their magnitudes by the total number of items:

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

  This can also be expressed as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$$

  where N is the total number of items.

  If each value x_i occurs with frequency f_i, then the mean can be adjusted accordingly:

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

  where n is the number of classes and $\sum_{i=1}^{n} f_i = N$.

Example.

Let's consider the set of numbers: 5, 8, 10, 12, and 15.
To find the arithmetic mean (denoted by $\bar{x}$), you would:

- Add up all the numbers: $5 + 8 + 10 + 12 + 15 = 50$.
- Count the total number of items in the set, which is 5.
- Divide the sum by the total number of items: $\bar{x} = \frac{50}{5} = 10$.

  So, the arithmetic mean of the set 5, 8, 10, 12, 15 is 10.

- **Mean for grouped Data:**

‘

Remark 4. *For grouped data, the arithmetic mean is calculated using the midpoint of each class interval and the frequency of each interval.*

Example.

Let's consider the following grouped data:

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

To find the arithmetic mean for grouped data:

1. Calculate the midpoint of each class interval.
2. Multiply each midpoint by its corresponding frequency and add up the results.
3. Add up all the frequencies.
4. Divide the sum from step 2 by the sum from step 3 to get the arithmetic mean.

Class	Frequency	Mid-point	Frequency $\times$ Mid-point
6-10	35	8	280
11-15	23	13	299
16-20	15	18	270
21-25	12	23	276
26-30	9	28	252
31-35	6	33	198
Total	100		1575

The mean will be $\bar{x} = \frac{1575}{100} = 15.75$.

3.2.1.2 Properties of Mean

1. $\sum_{i=1}^{n} k = nk$, where k is any constant.
2. $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$, where k is any constant.
3. $\sum_{i=1}^{n} (a + bx_i) = na + b\sum_{i=1}^{n} x_i$, where a and b are any constants.
4. $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

The sum of the product of two variables x and y can be expressed as:

$$\sum_{i=1}^{n}(x_i \times y_i) = (x_1 \times y_1) + (x_2 \times y_2) + \cdots + (x_n \times y_n)$$

where n is the total number of observations or data points, and x_i and y_i represent the i^{th} observation of variables x and, y respectively.

3.2.1.3 Weighted Mean

The weighted mean is a type of mean that incorporates different weights for each value in a dataset. It is calculated by multiplying each value by its corresponding weight, summing these products, and then dividing by the sum of the weights. Mathematically, the formula for the weighted mean $\bar{x}$ of a dataset $x_1, x_2, \ldots, x_n$ with corresponding weights $w_1, w_2, \ldots, w_n$ is:

$$\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_n \cdot x_n}{w_1 + w_2 + \ldots + w_n}$$

Example.

Let's consider the following set of values: x_1, x_2, x_3, x_4 and their corresponding weights w_1, w_2, w_3, w_4:

$$x_1 = 10, w_1 = 3 x_2 = 15, w_2 = 2 x_3 = 20, w_3 = 4 x_4 = 25, w_4 = 1$$

To calculate the weighted mean, we use the formula:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i \cdot w_i}{\sum_{i=1}^{4} w_i}$$

Substituting the given values:

$$\sum_{i=1}^{4} x_i \cdot w_i = (10 \times 3) + (15 \times 2) + (20 \times 4) + (25 \times 1)$$

$$= 30 + 30 + 80 + 25 = 165$$

$$\sum_{i=1}^{4} w_i = 3 + 2 + 4 + 1 = 10$$

Therefore,

$$\bar{x} = \frac{165}{10} = 16.5$$

So, the weighted mean of the given set of values is $\bar{x} = 16.5$.

3.2.1.4 Geometric Mean

- The geometric mean is useful for averaging rates of change, growth rates, or ratios.
- It is less affected by extreme values compared to the arithmetic mean.
- It is appropriate for data that follows a multiplicative relationship.

The geometric mean is calculated by taking the nth root of the product of n numbers. It is commonly used to find the average growth rate, ratios, or rates of change over time. The formula for calculating the geometric mean of n numbers $x_1, x_2, ..., x_n$ is:

$$\text{Geometric Mean} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

Example.

Let's say we have a series of numbers representing the growth rate of a certain investment over five years: 5
To find the geometric mean of these growth rates:

$$\text{Geometric Mean} = \sqrt[5]{1.05 \times 1.08 \times 1.10 \times 1.06 \times 1.12}$$

$$\text{Geometric Mean} = \sqrt[5]{1.43617408}$$

$$\text{Geometric Mean} \approx 1.0858$$

So, the geometric mean growth rate over these five years is approximately 1.0858

3.2.1.5 Harmonic Mean

- The harmonic mean is beneficial when dealing with rates or ratios, such as speed or efficiency.
- It is less influenced by extremely large or small values.
- It tends to be smaller than the arithmetic mean.
- It is suitable for situations where the impact of extreme values needs to be minimized.

The harmonic mean is calculated by dividing the number of observations by the sum of the reciprocals of each observation, and then taking the reciprocal

of the result. For a set of n observations $x_1, x_2, \ldots, x_n$, the harmonic mean is given by:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Example.

Suppose a car travels at different speeds over three segments of a journey: 40 mph, 60 mph, and 80 mph. To find the harmonic mean speed:

$$\text{Harmonic Mean} = \frac{3}{\frac{1}{40} + \frac{1}{60} + \frac{1}{80}}$$

$$\text{Harmonic Mean} = \frac{3}{\left(\frac{1}{40} + \frac{1}{60} + \frac{1}{80}\right)}$$

$$\text{Harmonic Mean} = \frac{3}{\left(\frac{6}{240} + \frac{4}{240} + \frac{3}{240}\right)}$$

$$\text{Harmonic Mean} = \frac{3}{\frac{13}{240}}$$

$$\text{Harmonic Mean} = \frac{720}{13}$$

$$\text{Harmonic Mean} \approx 55.384 \text{ mph}$$

So, the harmonic mean speed for the journey is approximately 55.384 mph.

Remark 5. *The choice of mean depends on the nature of the data and the specific question being addressed. While the arithmetic mean is the default choice for many situations, the geometric, harmonic, and weighted means are valuable alternatives when dealing with specific types of data or when certain properties, such as resistance to extreme values or consideration of differing importance of data points, are desired.*

3.2.2 The Mode

The mode is a measure of central tendency that represents the value or values that occur most frequently in a dataset, i.e., the value with the highest frequency.

For a dataset with discrete values, the mode is simply the value(s) that occur most frequently. It may be unique (unimodal) or may have multiple modes (bimodal, trimodal, etc.) if multiple values have the same highest frequency.

Example.

1. Find the mode of $5, 3, 5, 8, 9$. Mode: 5
2. Find the mode of $8, 9, 9, 7, 8, 2,$ and 5. It is a bimodal data: 8 and 9
3. Find the mode of $4, 12, 3, 6,$ and 7. No mode for this data.

For grouped data presented in a continuous frequency distribution, the mode can be calculated using the formula:

$$\text{Mode} = L + \frac{f - f_{\text{pre}}}{2f - f_{\text{pre}} - f_{\text{post}}} \times w$$

where:

- f = frequency of the modal class
- f_{pre} = frequency of the class preceding the modal class
- f_{post} = frequency of the class following the modal class
- w = size of the modal class
- L = lower boundary of the modal class

Example.

Consider the distribution of the sizes of farms selected at random from a district:

Size of farms	No. of farms
$5-14$	8
$15-24$	12
$25-34$	17
$35-44$	29
$45-54$	31
$55-64$	5
$65-74$	3

Calculating the mode:

$$
\begin{aligned}
f &= 31 \\
f_{\text{pre}} &= 29 \\
f_{\text{post}} &= 5 \\
w &= 10 \\
L &= 44.5 \\
\text{Mode} &= 44.5 + \frac{2}{2 \times 31 - 29 - 5} \times 10 \\
&\approx 44.5 + \frac{2}{57} \times 10 \\
&\approx 44.5 + 0.35 \\
&\approx 44.85
\end{aligned}
$$

3.2.2.1 Advantages and Limitations

- **Advantages:**
  - It remains unaffected by extreme outliers, enhancing its robustness in data analysis.
  - Its computational simplicity renders it accessible for quick calculations and intuitive interpretation.
  - It can accommodate datasets featuring open-ended class intervals, broadening its applicability.
- **Limitations:**
  - Its definition lacks strict mathematical precision, introducing ambiguity in certain contexts.

- It may not accurately represent the entirety of the dataset, as it relies solely on the most frequently occurring value(s).
- Its simplistic nature limits its utility in advanced statistical analyses or modeling.
- Due to its susceptibility to fluctuations in sample composition, it may not provide a stable measure of central tendency.
- In cases of multimodal distributions or evenly distributed data, determining a unique mode can be challenging.

Remark 6. *Given its characteristic as the point of maximum density within a dataset, the mode serves as a valuable tool for identifying the most prevalent size or value, particularly in research areas such as marketing, trade, business, and industry. Its significance lies in its ability to pinpoint the ideal or most commonly occurring value, making it a pertinent average for determining optimal sizes or quantities.*

3.2.3 The Median

- The median is the value that divides a dataset into two equal halves.
- In an ordered series of data, it is the middle observation, where the number of values less than the median is equal to the number of values greater than it.

3.2.3.1 Formulas for Median

- **For ungrouped data:**

1. If the number of observations (n) is odd: Median $= X_{\frac{n+1}{2}}$.
2. If the number of observations (n) is even: Median $= \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$.

Example.

Consider the dataset: 15, 22, 18, 10, 27, 13, 20.

- First, order the data: 10, 13, 15, 18, 20, 22, 27.
- Since the number of observations (n) is 7, which is odd, we use the formula:

$$\text{Median} = X_{\frac{n+1}{2}}$$

$$\text{Median} = X_{\frac{7+1}{2}}$$

$$\text{Median} = X_{\frac{8}{2}}$$

$$\text{Median} = X_4$$

$$\text{Median} = 18$$

Consider the dataset: $6, 5, 2, 8, 9, 4$.

- First, order the data: $2, 4, 5, 6, 8, 9$.
- Since the number of observations (n) is 6, which is even, we use the formula:

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

$$\text{Median} = \frac{X_{\frac{6}{2}} + X_{\frac{6}{2}+1}}{2}$$

$$\text{Median} = \frac{X_3 + X_4}{2}$$

$$\text{Median} = \frac{5+6}{2}$$

$$\text{Median} = \frac{11}{2} = 5.5$$

So, for this dataset, the median is 5.5.

- **For grouped data:**

$$\text{Median} = L + \frac{\left(\frac{N}{2} - c\right) \times w}{f}$$

Where:

1. L = lower class boundary of the median class.

2. c = cumulative frequency less than the median class.

3. f = frequency of the median class.

4. N = total number of observations.

5. w = size of the median class.

Example.

Consider the frequency distribution:

Class	Frequency
$10-20$	5
$21-30$	12
$31-40$	8
$41-50$	6
$51-60$	9
$61-70$	4

To find the median from this cumulative frequency data, we can follow these steps:

1. Identify the median class interval. This is the interval where the cumulative frequency exceeds or equals $\frac{N}{2}$, where N is the total sum of frequencies.
2. Use the median formula to calculate the median value.

The median class interval is situated on the interval $31-40$ because the cumulative frequency, which corresponds to the median, is $\frac{44}{2}=22$. The lower limit of this interval is 30.5. The cumulative frequency of the previous class is 17. The frequency of this interval is 8, and the class width is 10.

$$\text{Median} = 30.5 + \left(\frac{\frac{44}{2}-17}{8}\right)\times 10$$

Calculating:

$$\text{Median} = 30.5 + \left(\frac{22-17}{8}\right)\times 10 = 30.5 + \left(\frac{5}{8}\right)\times 10 = 36.75$$

So, the median of this data is 36.75.

3.2.3.2 Advantages and Limitations

- **Advantages:**

  1. **Robustness to outliers:** Unlike the mean, which can be heavily influenced by extreme values, the median is not affected by outliers. It gives a better representation of the central tendency of the data in the presence of outliers.
  2. **Simple interpretation:** The median represents the middle value when the data is ordered. This makes it easy to understand and interpret, especially for non-specialists.
  3. **Applicability to ordinal data:** The median can be calculated for ordinal data, where the precise numerical values may not have meaning but the order of values does.
  4. **Suitability for skewed distributions:** The median is often preferred over the mean for skewed distributions, since it reflects the central tendency better in such cases.

- **Limitations:**

  1. **Sensitive to small sample sizes:** The median can be less stable than the mean for small sample sizes. With small datasets, the median may not accurately represent the central tendency.
  2. **Less efficient estimator:** While the median is robust to outliers, it is generally less efficient as an estimator compared to the mean. It discards information about the magnitude of differences between values, which may be important in some contexts.
  3. **Difficulty in statistical inference:** The median has fewer statistical properties that can be exploited for inference compared to the mean. Confidence intervals and hypothesis tests involving the median are often less straightforward.
  4. **Limited use of some statistical techniques:** Some statistical techniques, such as regression analysis, require the use of the mean rather than the median. In these cases, using the median may not be appropriate.

Remark 7. *Understanding these advantages and limitations can help in making informed decisions about when to use the median as a measure of central tendency and when other measures, such as the mean, may be more appropriate.*

3.2.4 The Quartiles

Quartiles divide a dataset into four equal parts. There are three quartiles: the first ($Q1$), the second ($Q2$, the median), and the third ($Q3$).

- **Quartile for Ungrouped Data:** To compute quartiles for a dataset, the data must first be sorted in ascending order. Then, $Q1$ is the value at the $\frac{n+1}{4}$-th position, $Q2$ is the value at the $\frac{n+1}{2}$-th position, and $Q3$ is the value at the $\frac{3(n+1)}{4}$-th position.

$$Q1 = \left(\frac{n+1}{4}\right) \text{th item}, \quad Q2 = \left(\frac{n+1}{2}\right) \text{th item}, \quad Q3 = \left(\frac{3(n+1)}{4}\right) \text{th item}$$

Example.

Calculate the lower quartile, median, and upper quartile for the set of numbers {8, 5, 2, 6, 5, 7, 4}.
We can calculate the quartiles as follows:

- Arrange the set in increasing order: {2, 4, 5, 5, 6, 7, 8}.
- Q1 (First Quartile): Q1 = 4.
- Q2 (Second Quartile - Median): Q2 = 5.
- Q3 (Third Quartile): Q3 = 7.

- **Quartile for Grouped Data:** To calculate quartiles for grouped data, we use the following formulas:

$$Q_i = L + \left(\frac{C}{F}\right) \times \left(\frac{iN}{4} - M\right)$$

Where:

- N is the total frequency.
- M is the cumulative frequency leading up to the interval that contains the i^{th} quartile.
- L is the lower bound of the interval that contains the i^{th} quartile.
- C is the class width.
- F is the frequency of the interval that contains the i^{th} quartile.
- i is the quartile number (1 for the first quartile, 2 for the second quartile, etc.).

Example.

Suppose we have the following grouped dataset:

Class Interval	Frequency	Cumulative Frequency
$21 - 30$	12	12
$31 - 40$	21	33
$41 - 50$	34	67
$51 - 60$	20	87
$61 - 70$	6	93
$71 - 80$	4	97
$81 - 90$	3	100

- Q1 (First Quartile): $Q1 = 36.69$.
- Q2 (Second Quartile - Median): $Q2 = 50$.
- Q3 (Third Quartile): $Q3 = 54.5$.

3.3 Exercises

Exercise 1

You have a dataset representing the ages of students in a class:

18, 19, 20, 21, 22, 22, 23, 23, 24, 24, 24, 25, 25, 25, 26, 26, 27, 27, 27, 27, 28, 28, 28, 28, 29, 29, 29, 29, 30, 30, 30, 31, 31, 31, 32, 32, 32, 32, 33, 33, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 38, 39, 39, 40, 40, 41, 41, 41, 42, 42, 42, 43, 43, 43, 44, 44, 45, 45, 45, 46, 46, 47, 47, 48, 48, 48, 48, 49, 49, 49, 49, 50, 50, 51, 51, 51, 52, 52, 52, 53, 53, 54, 54, 55, 55, 55, 56, 56,

1. Calculate the arithmetic mean age of the students.
2. Calculate the harmonic mean age of the students.
3. Calculate the geometric mean age of the students.
4. Calculate the mode of the age of the students.
5. Calculate the median of the age of the students.
6. Calculate the first quartile (Q1) of the ages of the students.
7. Calculate the third quartile (Q3) of the ages of the students.

Exercise 2

Consider the following grouped data representing the exam scores of students:

Exam Score	Frequency
41-50	5
51-60	8
61-70	12
71-80	10
81-90	6
91-100	4

- Calculate the arithmetic mean exam score of the students.
- Calculate the harmonic mean exam score of the students.
- Calculate the geometric mean exam score of the students.
- Calculate the mode of the exam score of the students.

- Calculate the median of the exam score of the students.
- Calculate the first quartile (Q1) of the exam scores of the students.
- Calculate the third quartile (Q3) of the exam scores of the students.

Chapter 4

Measures of Dispersion (Variation)

4.1 Introduction

Measurement of variation is a critical aspect of data analysis across various fields, providing valuable insights into the degree of dispersion or diversity within a dataset. Variation refers to the extent to which data points differ from each other, offering essential information about the stability, consistency, and predictability of a process or phenomenon. Understanding variation is fundamental for making informed decisions, identifying patterns, and improving outcomes in numerous domains, including manufacturing, finance, healthcare, and research.

4.2 Objectives

- **Assessing Dispersion:** The primary objective of measuring variation is to assess the spread or dispersion of data points around a central value, such as the mean or median. By quantifying the degree of variability, analysts can evaluate the consistency or inconsistency of a process or system.
- **Identifying Patterns:** Variation analysis helps in identifying underlying patterns or trends within a dataset. By examining how data points deviate from the norm, analysts can uncover valuable insights about the factors influencing outcomes and make informed predictions about future behavior.
- **Monitoring Performance:** Measuring variation enables organizations to monitor the performance of processes, products, or services over time. By tracking changes in variability, stakeholders can identify deviations from desired standards and implement corrective actions to enhance quality and efficiency.

- **Improving Decision-Making:** Understanding variation enhances decision-making by providing a more comprehensive view of the underlying data. By considering both central tendency and dispersion, decision-makers can make more accurate assessments, mitigate risks, and optimize resource allocation.
- **Enhancing Quality Control:** Variation analysis is crucial for quality control and process improvement initiatives. By identifying sources of variation and their impact on outcomes, organizations can implement strategies to reduce variability, enhance consistency, and meet or exceed quality standards.
- **Supporting Statistical Inference:** In statistical inference, measuring variation is essential for estimating parameters, testing hypotheses, and making inferences about populations based on sample data. Accurate assessment of variability ensures the reliability and validity of statistical analyses and conclusions.

  Overall, the measurement of variation serves as a cornerstone of data analysis, facilitating deeper insights, informed decision-making, and continuous improvement across diverse domains and disciplines.

4.3 Types of Measures of Dispersion

There exist several measures of dispersion commonly employed in statistical analysis. These include:

- Range and Coefficient of Range.
- Quartile Deviation and Coefficient of Quartile Deviation.
- Mean Deviation and Coefficient of Mean Deviation.
- Standard Deviation and Coefficient of Variation.

These measures offer insights into the variability or spread of data points within a dataset and are utilized across various fields for analytical purposes.

4.3.1 The Range (R)

The range, denoted as R, is computed as the difference between the largest score, L, and the smallest score, S, in a dataset:

$$R = L - S$$

It provides a quick and straightforward measure of variability. However, due to its sensitivity to extreme values, it may not accurately represent the overall variability in the data.

Example.

Consider two sets of exam scores:

1. Set 1: $65, 70, 72, 75, 80, 85, 90$
2. Set 2: $65, 65, 70, 75, 80, 85, 90$

Both sets have the same range of 25, calculated as $90 - 65$. However, they exhibit different levels of variability. Set 2 has a repeated lower score of 65, indicating less variability compared to Set 1.
For grouped data, the range can be computed using class boundaries. One common method involves subtracting the lower class limit (LCL) from the upper class limit (UCL) of the first and last classes, respectively.

$$\text{Range} = UCL_{\text{last class}} - LCL_{\text{first class}}$$

Another approach employs class marks, which are the midpoints of the classes. In this method, the range is calculated as the difference between the class mark of the last class and the class mark of the first class.

$$\text{Range} = X_{\text{last class}} - X_{\text{first class}}$$

Where:

- $UCL_{\text{last class}}$ and $LCL_{\text{first class}}$ are the upper class limit and lower class limit of the first and last classes, respectively.
- $X_{\text{last class}}$ and $X_{\text{first class}}$ are the class marks of the last and first classes, respectively.

Example.

Suppose we have the following frequency distribution representing the ages of students in a class:

Age Range	Frequency
10-15	8
16-20	12
21-25	10
26-30	6

To calculate the range for this grouped data, we first need to determine the class boundaries or limits. Let's assume the boundaries are inclusive. For the first class, $LCL_{\text{first class}} = 10$ and $UCL_{\text{first class}} = 15$. For the last class, $LCL_{\text{last class}} = 26$ and $UCL_{\text{last class}} = 30$.
Using the formula:

$$\text{Range} = UCL_{\text{last class}} - LCL_{\text{first class}}$$

we can calculate:

$$\text{Range} = 30 - 10 = 20$$

So, the range of ages for the grouped data is 20 years.

4.3.1.1 The advantages and limitations of using the range as a measure of dispersion

- **The Advantages:**
  - Simple to Compute: The range is straightforward to calculate, requiring only the identification of the highest and lowest values in a dataset.
  - Easy to Understand: It provides an intuitive understanding of the spread of data by indicating the difference between the extreme values.
  - Useful for Quick Comparisons: Despite its limitations, the range can still be useful for quick comparisons between datasets or for providing a general sense of variability.
- **The limitations**
  - Sensitive to Outliers: The range is heavily influenced by extreme values or outliers in the dataset. A single outlier can dramatically affect the range, potentially leading to misinterpretation of variability.
  - Does Not Consider Entire Dataset: It only considers two data points the maximum and minimum values, thus ignoring the distribution

of values between them. Consequently, it may not accurately reflect the overall variability present in the dataset.

– Not Suitable for Inferential Statistics: Due to its limited scope and sensitivity to outliers, the range is not recommended for use in inferential statistics or hypothesis testing.

– Dependent on Sample Size: The range does not account for the size of the dataset. Two datasets with the same range may have different sample sizes, leading to different levels of variability.

4.3.2 The Coefficient of Range (CR)

The Coefficient of Range (CR) is a relative measure of dispersion, calculated by taking the ratio of the difference between the largest and smallest items in a distribution to their sum.

- **Formula for the Coefficient of Range, denoted as CR :**

$$CR = \frac{L - S}{L + S} = \frac{Range}{L + S}$$

This formula normalizes the range by dividing it by the sum of the upper and lower limits. It provides a ratio that indicates the proportion of the total range occupied by the actual range of the dataset.

Example.

Suppose we have a class of 20 students, and their exam scores are as follows:

$$\{60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140,$$
$$145, 150, 155\}$$

To calculate the coefficient range:

1. Identify the Upper and Lower Limits:
   - L (Upper Limit): 155 (highest score)
   - S (Lower Limit): 60 (lowest score)
2. Calculate the Range:
$$\text{Range} = L - S = 155 - 60 = 95$$
3. Calculate the Coefficient of Range :
$$CR = \frac{L - S}{L + S} = \frac{95}{155 + 60}$$
$$CR = \frac{95}{215} \approx 0.4419$$

4.3.3 The Quartile Deviation (Semi-inter quartile range)

The interquartile range (IQR) is defined as the difference between the third quartile Q_3 and the first quartile Q_1 of a dataset. Mathematically, it is expressed as:

$$\text{IQR} = Q_3 - Q_1$$

The semi-interquartile range (SIQR), also known as the semi Inter-Quartile Range, is half of the interquartile range (IQR) and is calculated as:

$$\text{SIQR} = \frac{Q3 - Q1}{2}$$

Where:

- Q_1 represents the first quartile, which divides the lower 25% of the data from the upper 75%.

- Q_3 represents the third quartile, which divides the lower 75% of the data from the upper 25%.

The interquartile range is a measure of statistical dispersion that is less sensitive to outliers compared to the range. It provides insights into the spread of the middle 50% of the data, making it a robust measure of variability in skewed datasets.

Example.

Consider the following dataset representing the scores of students in a math test:

$$65, 70, 72, 75, 80, 85, 90, 95, 98, 100$$

To calculate the interquartile range (IQR):

1. Arrange the data in ascending order:

$$65, 70, 72, 75, 80, 85, 90, 95, 98, 100$$

2. Find the first quartile (Q_1) and the third quartile (Q_3):
   - $Q_1 = 70 + (72 - 70) \times 0.75 = 70 + 2 \times 0.75 = 70 + 1.5 = 71.5$
   - $Q_3 = 95 + (98 - 95) \times 0.25 = 95 + 3 \times 0.25 = 95 + 0.75 = 95.75$
3. Calculate the IQR:

$$\text{IQR} = Q_3 - Q_1 = 95.75 - 71.5 = 24.25$$

To calculate the semi-interquartile range (SIQR):

$$\text{SIQR} = \frac{1}{2} \times \text{IQR} = \frac{1}{2} \times 24.25 = 12.125$$

Therefore, the interquartile range (IQR) for this dataset is 23, and the semi-interquartile range (SIQR) is 12.125.

4.3.4 The Coefficient of Quartile Deviation.

As Quartile Deviation is an absolute measure of dispersion, one cannot use it for comparing the variability of two or more distributions when they are expressed in different units. Therefore, in order to compare the variability of two or more series with different units it is essential to determine the relative measure of Quartile Deviation, which is also known as the Coefficient of Quartile Deviation. It is studied to make the comparison between the degree of variation in different series. The formula for determining the Coefficient of Quartile Deviation is as follows:

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

4.3.5 The Mean absolute Deviation (MAD)

The mean absolute deviation (MAD), is a measure of statistical dispersion. It quantifies the average absolute difference between each data point and the mean of the dataset.

To calculate the mean absolute deviation (MAD) for a dataset:

- Find the mean (average) of the dataset, denoted as $\bar{x}$.
- For each data point x_i in the dataset, calculate the absolute difference between that data point and the mean: $|x_i - \bar{x}|$.
- Find the average of these absolute differences by summing them up and dividing by the total number of data points n:

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Where:

- n is the number of data points.
- x_i represents each individual data point.
- $\bar{x}$ is the mean of the dataset.

The formula for mean absolute deviation (MAD) for a frequency distribution is:

$$\text{Mean Deviation (MAD)} = \frac{\sum_{i=1}^{n} f_i \cdot |x_i - \bar{x}|}{N}$$

Where:

- f_i is the frequency of the i^{th} class.
- x_i is the midpoint of the i^{th} class interval.
- $\bar{x}$ is the mean of the distribution.
- N is the total frequency.

This formula considers the frequency of each class interval, giving a more accurate measure of the dispersion in a frequency distribution.

4.3.6 The coefficient of mean deviation (CMD)

The coefficient of mean deviation (CMD) is a statistical measure used to express the mean absolute deviation (MAD) as a percentage of the mean ($\bar{x}$) of the data set. It standardizes the measure of dispersion provided by the mean deviation, allowing for comparisons of variability across different data sets.

Mathematically, the coefficient of mean deviation (CMD) is calculated as follows:

$$\text{Coefficient of Mean Deviation (CMD)} = \left(\frac{\text{Mean absolute Deviation (MAD)}}{\bar{x}}\right) \times 100$$

Where:

- Mean absolute Deviation (MAD) is the average absolute deviation of individual data points from the mean.
- $\bar{x}$ is the mean of the data set.

  The result is expressed as a percentage, indicating the proportion of the mean represented by the mean deviation. This allows for a standardized comparison of variability across data sets, independent of the units of measurement.

1. **Ungrouped Data Example**

   Suppose we have the following ungrouped data representing the ages of 10 individuals

   Calculate the mean absolute deviation (MAD) and the coefficient of mean deviation (CMD)

   $$25, 30, 35, 40, 45, 50, 55, 60, 65, 70$$

   To calculate the mean absolute deviation (MAD) and the coefficient of mean deviation (CMD) for this non-grouped data, we can follow these steps:

   - **Calculate the mean:**

     $$\text{Mean} = \frac{25 + 30 + 35 + 40 + 45 + 50 + 55 + 60 + 65 + 70}{10}$$
     $$= \frac{465}{10} = 46.5$$

   - **Calculate the Absolute Deviations from the mean:**

     $$|25 - 46.5| = 21.5$$
     $$|30 - 46.5| = 16.5$$

$$|35 - 46.5| = 11.5$$
$$|40 - 46.5| = 6.5$$
$$|45 - 46.5| = 1.5$$
$$|50 - 46.5| = 3.5$$
$$|55 - 46.5| = 8.5$$
$$|60 - 46.5| = 13.5$$
$$|65 - 46.5| = 18.5$$
$$|70 - 46.5| = 23.5$$

- **Calculate the Mean Absolute Deviation (MAD):**

$$\text{MAD} = \frac{21.5 + 16.5 + 11.5 + 6.5 + 1.5 + 3.5 + 8.5 + 13.5 + 18.5 + 23.5}{10} =$$

$$\frac{126}{10} = 12.6$$

- **Calculate the Coefficient of Mean Deviation (CMD):**

$$\text{CMD} = \left(\frac{12.6}{46.5}\right) \times 100 \approx 27.1\%$$

Remark 8. *The mean absolute deviation (MAD) is 12.6, and the coefficient of mean deviation (CMD) is approximately 27.1%. This means that, on average, the data points deviate from the mean by approximately 12.6 units, representing about 27.1% of the mean value.*

2. **Grouped Data Example**

Given data:

Age Group	Frequency
$10 - 20$	5
$21 - 30$	8
$31 - 40$	12
$41 - 50$	10
$51 - 60$	6

- Finding the midpoints of each age group:

Midpoint	Frequency
15	5
25.5	8
35.5	12
45.5	10
55.5	6

- Calculating the mean ($\bar{x}$) of the midpoints weighted by their frequencies:

$$\bar{x} = \frac{\sum \text{Midpoint} \times \text{Frequency}}{\sum \text{Frequency}}$$

$$\bar{x} = \frac{(15 \times 5) + (25.5 \times 8) + (35.5 \times 12) + (45.5 \times 10) + (55.5 \times 6)}{5 + 8 + 12 + 10 + 6}$$

$$\bar{x} = \frac{75 + 204 + 426 + 455 + 333}{41} \approx \frac{1493}{41} \approx 36.41$$

- Next, calculating the absolute deviations of each midpoint from the mean:

$$|x_i - \bar{x}| = |15 - 36.41| \approx 21.41$$
$$|x_i - \bar{x}| = |25.5 - 36.41| \approx 10.91$$
$$|x_i - \bar{x}| = |35.5 - 36.41| \approx 0.91$$
$$|x_i - \bar{x}| = |45.5 - 36.41| \approx 9.09$$
$$|x_i - \bar{x}| = |55.5 - 36.41| \approx 19.09$$

- Now, multiplying each absolute deviation by its corresponding frequency:

$$\text{Frequency} \times |x_i - \bar{x}| = 5 \times 21.41 + 8 \times 10.91 + 12 \times 0.91 + 10 \times 9.09 + 6 \times 19.09$$

- Summing up these products:

$$\sum(\text{Frequency} \times |x_i - \bar{x}|) = (5 \times 21.41) + (8 \times 10.91) + (12 \times 0.91) + (10 \times 9.09) + (6 \times 19.09)$$

$$\sum(\text{Frequency} \times |x_i - \bar{x}|) = 107.05 + 87.28 + 10.92 + 90.9 + 114.54 = 410.69$$

- Calculating the mean absolute deviation (MAD):

$$\text{MAD} = \frac{\sum(\text{Frequency} \times |x_i - \bar{x}|)}{\sum \text{Frequency}}$$

$$\text{MAD} = \frac{410.69}{41} \approx 10.02$$

- Finally, finding the coefficient of mean deviation (CMD):

$$\text{CMD} = \frac{\text{MAD}}{\bar{x}} \times 100$$

$$\text{CMD} = \frac{10.02}{36.41} \times 100 \approx 27.53\%$$

The mean deviation (MAD) is approximately 10.02 and the coefficient of mean deviation (CMD) is approximately 27.53%. This means, on average, each age group deviates from the mean age by approximately 10.02 units, representing about 27.53% of the mean age.

4.3.7 The Variance

Variance is a statistical measure that quantifies the spread or dispersion of a set of data points.

The variance can be calculated in two ways: for a population and for a sample.

- **Population Variance** The population variance, denoted by σ^2, is determined by dividing the sum of squared deviations from the mean by the total number of values in the population. It represents the average squared deviation from the mean. In mathematical terms:

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

For frequency distribution, it is expressed as:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(X_i - \mu)^2}{N}$$

- **Sample Variance** The sample variance, denoted by S^2, is an estimate of the population variance. While one might intuitively replace the population mean with the sample mean, it is common practice to adjust the formula to better estimate the population parameter. To achieve this, the sum of squared deviations is divided by one less than the sample size, n-1, rather than n:

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

For frequency distribution, it is expressed as:

$$S^2 = \frac{\sum_{i=1}^{k} f_i(X_i - \bar{X})^2}{n-1}$$

Example.

Suppose we have a dataset representing the daily temperatures (in degrees Celsius) in a city for a week:

$$\text{Temperatures: } 20, 22, 21, 23, 19, 25, 24$$

First, we need to find the mean (average) temperature:

$$\text{Mean} = \frac{20 + 22 + 21 + 23 + 19 + 25 + 24}{7} = \frac{154}{7} \approx 22$$

Then, calculate the squared differences from the mean for each temperature:

$$\begin{aligned}(20 - 22)^2 &= (-2)^2 = 4\\ (22 - 22)^2 &= (0)^2 = 0\\ (21 - 22)^2 &= (-1)^2 = 1\\ (23 - 22)^2 &= (1)^2 = 1\\ (19 - 22)^2 &= (-3)^2 = 9\\ (25 - 22)^2 &= (3)^2 = 9\\ (24 - 22)^2 &= (2)^2 = 4\end{aligned}$$

Next, find the average of these squared differences, which gives the variance:

$$\text{Variance} = \frac{4 + 0 + 1 + 1 + 9 + 9 + 4}{7} = \frac{28}{7} = 4$$

So, the variance of this temperature dataset is 4 square degrees Celsius. This tells us that the temperatures in this dataset are, on average, 4 square degrees Celsius away from the mean temperature of 22°C.

4.3.8 The standard deviation

The standard deviation is another measure of the spread or dispersion of a dataset. It's simply the square root of the variance.

Example.

Temperatures: 20, 22, 21, 23, 19, 25, 24
We've already calculated the variance to be 4 square degrees Celsius.
To find the standard deviation, we take the square root of the variance:

$$\text{Standard Deviation} = \sqrt{4} = 2$$

So, the standard deviation of this temperature dataset is 2 degrees Celsius.

Remark 9. *The standard deviation gives us an idea of how much the individual data points deviate from the mean temperature (which was 22°C in our example). In this case, a standard deviation of 2°C indicates that, on average, the daily temperatures are 2°C away from the mean temperature of 22°C.*

4.3.9 The coefficient of variation (CV)

The coefficient of variation (CV) is a statistical measure used to compare the variability or dispersion of two or more sets of data relative to their means. It's expressed as a percentage and is calculated by dividing the standard deviation of the data by the mean and then multiplying by 100.

The formula for the coefficient of variation (CV) is:

$$CV = \left(\frac{\text{Standard Deviation}}{\text{Mean}}\right) \times 100$$

The coefficient of variation is particularly useful when comparing the variability of datasets that have different units or scales. It provides a standardized measure of dispersion that allows for a more meaningful comparison.

Example.

Suppose we have two sets of data:
Set A: 10, 12, 15, 18, 20 Set B: 50, 55, 60, 65, 70

1. **For Set A**
   - $Mean = \frac{(10+12+15+18+20)}{5} = 15$
   - Standard Deviation ≈ 4.3
2. **For Set B**
   - $Mean = \frac{(50+55+60+65+70)}{5} = 60$
   - Standard Deviation ≈ 7.9

Then, calculate the coefficient of variation for each series:

1. **For Set A**
$$CV_A = \left(\frac{4.3}{15}\right) \times 100 \approx 28.7\%$$
2. **For Set B**
$$CV_B = \left(\frac{7.9}{60}\right) \times 100 \approx 13.2\%$$

   In this example, even though Set A has a higher standard deviation, its coefficient of variation is higher than that of Set B, indicating that Set A has greater relative variability compared to its mean than Set B.

4.3.10 Standard scores (z-scores)

Standard scores, also known as z-scores, are a way to standardize and compare data points from different distributions. A standard score represents how many standard deviations a data point is from the mean of its distribution.

The formula to calculate the z-score of a data point x in a distribution with mean μ and standard deviation σ is:

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

- x is the individual data point.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.
- z is the z-score.

If the z-score is positive, it means the data point is above the mean, while a negative z-score indicates the data point is below the mean. A z-score of 0 means the data point is exactly at the mean.

Example.

consider a dataset of exam scores with a mean of 70 and a standard deviation of 10. If a student scored 80 on the exam, we can calculate their z-score as follows:

$$z = \frac{80 - 70}{10} = 1$$

This means the student's score is 1 standard deviation above the mean. Standard scores (z-scores) are particularly useful for comparing data points from different distributions or for identifying outliers in a dataset. They provide a standardized way of assessing how unusual or typical a particular data point is within its distribution.

4.3.11 Moments

In statistics, moments are numerical measures that describe the shape and characteristics of a probability distribution. They are used to quantifying various aspects of the distribution, such as its center, spread, skewness, and kurtosis. There are several moments commonly used, including the mean, variance, skewness, and kurtosis.

- **First Moment (Mean):** The first moment, or the mean, represents the center of the distribution. It measures the average value of the data points.
- **Second Central Moment (Variance) :** The second central moment, or the variance, quantifies the spread or dispersion of the data points around the mean. It's the average of the squared differences between each data point and the mean.
- **Third Central Moment (Skewness):** The third central moment, or skewness, measures the asymmetry of the distribution. Positive skewness indicates a right-skewed distribution (tail to the right), while negative skewness indicates a left-skewed distribution (tail to the left).
- **Fourth Central Moment (Kurtosis):** The fourth central moment, or kurtosis, measures the "peakedness" of the distribution. It quantifies whether the distribution has heavier tails and a sharper or flatter peak compared to a normal distribution. Kurtosis greater than 3 indicates heavier tails (leptokurtic) and a sharper peak, while kurtosis less than 3 indicates lighter tails (platykurtic) and a flatter peak.

Remark 10. *These moments provide valuable insights into the characteristics of a distribution and are widely used in statistical analysis and probability theory. They help researchers and analysts understand the underlying properties of the data and make informed decisions based on these properties.*

4.4 Exercises

Exercise 1

Suppose you are provided with grouped data representing the ages of participants in a survey conducted at a community center. The data is presented as follows:

Age	Number of Participants
$10-20$	15
$21-30$	25
$31-40$	30
$41-50$	20
$51-60$	10

1. Calculate the range, the coefficient of range and relative range of the age groups.
2. Compute both the quartile deviation (semi-interquartile range) and the interquartile range.
3. Calculate the mean deviation and coefficient of mean deviation for the age groups.
4. Determine the standard deviation and coefficient of variation for the age groups.

Exercise 2

For the following set of data representing the temperatures (in degrees Celsius) recorded in a city over a week:

$$23, 22, 17, 19, 22, 21, 15, 18, 19, 17, 20,$$

$$21, 20, 20, 21, 17, 18, 20, 22, 25, 28, 30, 33$$

1. Calculate the range, the coefficient of range and relative range of the age groups.
2. Compute both the quartile deviation (semi-interquartile range) and the interquartile range.
3. Calculate the mean deviation and coefficient of mean deviation for the age groups.

4. Determine the standard deviation and coefficient of variation for the age groups.

Exercise 3

The following dataset represents the monthly sales revenue (in thousands of dollars) for a company over two years:

$$120, 130, 140, 150, 160, 170, 180, 190, 200, 210,$$
$$220, 230, 240, 250, 260, 270, 280, 290, 300, 310,$$
$$320, 330, 340, 350, 360, 370, 380, 390, 400, 410,$$
$$420, 430, 440, 450, 460, 470, 480, 490, 500, 510,$$
$$520, 530, 540, 550, 560, 570, 580, 590, 600$$

1. What is the population being studied?
2. What is the variable being studied and its nature?
3. Construct a frequency table using ten classes.
4. Create a histogram representing the frequency distribution.
5. Calculate the mean, median, and standard deviation of the sales revenue.
6. Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
7. Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Exercise 4

The following dataset represents the population (in millions) of 30 cities:

$$12, 8, 15, 9, 20, 25, 18, 22, 14, 10, 16,$$
$$19, 21, 23, 11, 7, 26, 13, 17, 24, 28, 30,$$
$$27, 29, 35, 32, 31, 33, 34, 36$$

1. What is the population being studied?
2. What is the variable being studied and its nature?
3. Construct a frequency table using six classes.

4. Create a histogram representing the frequency distribution.
5. Calculate the mean, median, and standard deviation of the city populations.
6. Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
7. Construct an Ogive (cumulative frequency curve) based on the frequency distribution.

Exercise 5

The following dataset represents the monthly sales (in thousands of units) of a product over two years:

$$120, 130, 140, 150, 160, 170, 180, 190, 200, 210,$$
$$220, 230, 240, 250, 260, 270, 280, 290, 300, 310,$$
$$320, 330, 340, 350, 360, 370, 380, 390, 400, 410,$$
$$420, 430, 440, 450, 460, 470, 480, 490, 500, 510,$$
$$520, 530, 540, 550, 560, 570, 580, 590, 600$$

1. What is the population being studied?
2. What is the variable being studied and its nature?
3. Construct a frequency table using ten classes.
4. Create a histogram representing the monthly sales distribution.
5. Calculate the mean, median, and standard deviation of the monthly sales.
6. Draw a polygon connecting the midpoints of the class intervals on the x-axis and the corresponding frequencies on the y-axis.
7. Construct an Ogive (cumulative frequency curve) based on the frequency distribution of monthly sales.

Exercise 6

Consider the dataset representing the scores of 50 students in a comprehensive exam. Given the dataset:

$$60, 55, 72, 88, 94, 83, 74, 65, 80, 58, 65,$$
$$90, 88, 68, 55, 79, 83, 72, 74, 76, 81, 53,$$
$$90, 87, 66, 70, 78, 62, 91, 85, 76, 79, 92,$$
$$88, 78, 75, 71, 89, 84, 70, 86, 84, 89, 77,$$
$$93, 85, 78, 73, 80, 82, 75$$

1. What is the population being studied?
2. What is the variable being studied and its nature?
3. Divide the scores into eight classes and create a frequency table.

4. Utilize the frequency table to construct a histogram representing the frequency distribution.
5. Calculate the mean, median, and mode of the exam scores.
6. Determine the range, quartile deviation, interquartile range, and coefficient of variation for the exam scores.
7. Analyze the shape of the distribution based on measures of skewness and kurtosis.
8. Identify any outliers in the dataset and discuss their potential impact on the analysis.

Chapter 5

Combinatorial Analysis

5.1 Introduction

Combinatorial analysis, also known as **combinatorics**, is the branch of mathematics concerned with counting, arranging, and selecting objects or elements according to specific criteria. It deals with problems of counting, arranging, and choosing elements or objects systematically, often without necessarily listing them all individually. Combinatorial analysis finds applications in various fields such as computer science, cryptography, probability theory, and optimization.

The objective of combinatorial analysis is to count the distinct ways of grouping the elements of a set E of n ($n \in \mathbb{N}^\star$) elements according to specified rules.

Notations:

Let E be a set of n elements,

1. The cardinality of the set E, denoted by $\text{Card}(E)$, is the number of elements in E.

2. For any $n \in \mathbb{N}^\star$, the factorial of n, denoted as $n!$, is defined as:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

   and we have: $0! = 1! = 1$.

3. A partition of E is the set of all subsets of E, denoted by $\mathcal{P}(E)$.

5.2 Arrangements, permutations, and combinations

Throughout the following, let E be a set of n ($n \in \mathbb{N}^\star$) elements, and let p be a positive non-zero integer such that $1 \leq p \leq n$:

5.2.1 Arrangements

Definition

An arrangement of p elements from E is any ordered sequence of p elements chosen from the n elements of E. The number of arrangements is denoted by A_p^n.

Remark 11. *Two arrangements are considered distinct if they differ either by the elements they contain or by their order within the sequence.*

Example.

1. The PIN code of a bank card consists of 4 digits chosen from the set $E = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The different ways to compose this code are arrangements of $p = 4$ elements from a set of $n = 10$ elements, denoted as A_4^{10}.
2. Choosing a delegation consisting of 2 students from a section of 150 students constitutes arrangements A_2^{150}.
3. In the first example, the same digit can appear multiple times (e.g., code 9039 or 4404, etc.). Each digit of the code has 10 possibilities of appearance, so the number of arrangements $A_4^{10} = 10 \times 10 \times 10 \times 10 = 10^4$.
4. In the second case, the two students must be different. The first student is chosen from the 150 students in the section, while the second one is chosen from the remaining 149 students, resulting in $A_2^{150} = 150 \times 149 = 22350$ arrangements.

Remark 12. *There are two types of arrangements distinguished: arrangements with repetition and arrangements without repetition.*

5.2.1.1 Arrangements with Repetition

Definition

When an element can be chosen or observed multiple times, the number of arrangements with repetition of p elements from a set of n elements is given by:

$$A_p^n = n^p$$

Example.

Let's consider arranging 2 elements from the set $\{A, B, C\}$ with repetition allowed.

- Arrangements with repetition: $A_2^3 = 3^2 = 9$
- The arrangements are: $AA, AB, AC, BA, BB, BC, CA, CB, CC$

5.2.1.2 Arrangements without Repetition

Definition

When an element can be chosen or observed only once, the number of arrangements without repetition of p elements from a set of n elements is given by:

$$A_p^n = \frac{n!}{(n-p)!}$$

Example.

Now, let's consider arranging 2 elements from the same set $\{A, B, C\}$ without repetition.

- Arrangements without repetition: $A_2^3 = \frac{3!}{(3-2)!} = \frac{3\times 2}{1} = 6$
- The arrangements are: AB, AC, BA, BC, CA, CB

5.2.2 Permutations

Definition

A permutation is any ordered sequence of n elements from the set, E or any arrangement of n elements of E. The number of permutations is denoted by P_n.

Example.

- **First case:** we want to seat 5 students around a round table with 5 chairs.
- **Second case:** we seek the number of possible words (with or without meaning) that can be formed by permuting the 9 letters of the word "pellicule".

In the first case, each student is seated in a chair, so the students are necessarily different $5 \times 4 \times 3 \times 2 \times 1$. The number of ways to seat the students around the table is: $P_5 = 5 \times 4 \times 3 \times 2 \times 1 = 5!$. However, in the second case, the letter "l" is repeated 3 times and the letter "e" is repeated 2 times, so we can write words containing the letter "l" in three different positions and the letter "e" in two different positions. Therefore, the number of words that can be formed from the 9 letters of the word "pellicule" must be related to the number of permutations of 3 identical letters "l" and 2 identical letters "e". Consequently, the number of words in this case is:

$$P_9 = \frac{9!}{3! \times 2!}$$

This leads to distinguishing two types of permutations:

5.2.2.1 Permutations without Repetition

Definition

A permutation without repetition is any ordered sequence of n distinct elements from the set E. The number of permutations without repetition of n elements from a set of n elements is:

$$P_n = n!$$

Example.

we have the set $\{1, 2, 3\}$ and we want to create 3-digit numbers without repetition:

$$P_3 = 3! = 6$$

The permutations are: $12, 13, 21, 23, 31, 32$.

This example illustrates how we can create different 3-digit numbers without repetition from the set $\{1, 2, 3\}$. Each digit is used only once in each permutation.

5.2.2.2 Permutations with Repetition

When an element exists k times, the number of permutations with repetition of n elements from a set of n elements is:

$$P_n = \frac{n!}{k!}$$

In cases where each element x_i from the set E exists k_i times such that $1 \leq i \leq m$ and $\sum_{i=1}^{m} k_i = k_1 + k_2 + \ldots + k_m = n$, with

$$E = \left\{ \begin{array}{ll} x_1 x_1 \ldots x_1 & (k_1 \text{ times}) \\ x_2 x_2 \ldots x_2 & (k_2 \text{ times}) \\ \vdots & \\ x_m x_m \ldots x_m & (k_m \text{ times}) \end{array} \right\}$$

the number of permutations, is given by:

$$P_n = \frac{n!}{k_1! \cdot k_2! \cdot \ldots \cdot k_m!}$$

Example.

Suppose we have the set $E = \{A, B, C\}$ where each element occurs multiple times:

$$E = \{AA, BB, CC\}$$

We want to create 4-letter strings by selecting from this set.
In this case, each element A, B, and C occurs twice, so $k_1 = k_2 = k_3 = 2$ and $n = 6$ (total number of elements in the set).
Using the formula:

$$P_n = \frac{n!}{k_1! \cdot k_2! \cdot k_3!} = \frac{6!}{2! \cdot 2! \cdot 2!} = \frac{720}{8} = 90$$

So, there are 90 permutations of 4-letter strings that can be created from the set $E = \{AA, BB, CC\}$.

5.2.3 Combinations

Definition

A combination is an arrangement of p elements from a set E of n elements in which the order does not matter. There are two types of combinations: combinations without Repetition and combinations with Repetition.

5.2.3.1 Combination without Repetition

Definition

A combination without Repetition is any set of p elements taken without Repetition from the n elements of E. The number of combinations without Repetition is given by:

$$C_p^n = \frac{n!}{p! \cdot (n-p)!}$$

Example.

Forming a delegation of 3 employees within a company of 50 employees! The number of delegations is:

$$C_3^{50} = \frac{50!}{3! \cdot 47!} = \frac{50 \times 49 \times 48}{6} = 1960$$

5.2.3.2 Combination with Repetition

Definition

A combination with Repetition is any set of p elements taken with Repetition from the n elements of E. The number of combinations with Repetition is given by:

$$C_p^{n+p-1} = \frac{(n+p-1)!}{p! \cdot (n-1)!}$$

Example.

Forming words of 3 letters (with or without meaning) from a set of 5 letters! The number of words is:

$$C_3^7 = \frac{7!}{3! \cdot 4!} = \frac{7 \times 6 \times 5}{6} = 35$$

Proposition

(Properties of Combinations)

1. $C_n^0 = C_n^n = 1$.
2. $C_n^1 = C_1^{n-1} = n$.
3. $C_n^2 = C_n^{n-2} = \frac{n \times (n-1)}{2}$.
4. $C_n^p = C_n^{n-p}$.
5. $C_{n-1}^{p-1} + C_{n-1}^p = C_n^p$.

Definition

To demonstrate the properties above, we use the combination formula given in definition (5.1.7).

1. $C_0^n = \frac{n!}{0!\times n!} = \frac{n!}{n!\times 0!} = C_n^n = 1.$
2. $C_1^n = \frac{n!}{1!\times (n-1)!} = \frac{n!}{(n-1)!\times 1!} = C_n^1 = n.$
3. $C_2^n = \frac{n!}{2!\times (n-2)!} = \frac{n!}{(n-2)!\times 2!} = C_{n-2}^n = \frac{n\times(n-1)}{2}.$
4. $C_p^n = \frac{n!}{p!\times (n-p)!} = \frac{n!}{(n-p)!\times p!} = C_{n-p}^n.$
5. $C_{p-1}^{n-1} + C_p^{n-1} = \frac{(n-1)!}{(p-1)!\times (n-p)!} + \frac{(n-1)!}{p!\times (n-1-p)!} = \frac{(n-1)!}{p!}.$

5.2.3.3 Expanding the binomial and Pascal's Triangle

The expansion of the polynomial $(a+b)^n = (a+b)\times(a+b)\times\ldots\times(a+b)$ (repeated n times) requires calculating, for each term, a^k (or b^k), where $0 \leq k \leq n$, the number of distinct ways of choosing k instances of a (or b) out of n possibilities. This number is given by the binomial coefficient:

$$C_k^n = \frac{n!}{k! \times (n-k)!}$$

And we have:

$$
\begin{aligned}
&(a+b)^0 = 1 \\
&(a+b)^1 = a+b \\
&(a+b)^2 = a^2+2ab+b^2 \\
&(a+b)^3 = a^3+3a^2b+3ab^2+b^3 \\
&(a+b)^4 = a^4+4a^3b+6a^2b^2+4ab^3+b^4 \\
&\dots \\
&(a+b)^n = \sum_{k=0}^{n} C_k^n \times a^{n-k} \times b^k
\end{aligned}
$$

The binomial coefficients form a triangle where the rows correspond to constant n, and the sum of two consecutive coefficients in a row equals the term below it in the next row after the second term.

Example: Pascal's Triangle for $n = 7$:

1							
1	1						
1	2	1					
1	3	3	1				
1	4	6	4	1			
1	5	10	10	5	1		
1	6	15	20	15	6	1	
1	7	21	35	35	21	7	1

5.3 Exercises

Exercise 1

In a group of 12 students, how many different ways can a committee of 5 students be selected?

Exercise 2

How many different 5-letter words can be formed using the letters in the word "APPLE"?

Exercise 3

A club consists of 10 members. In how many ways can a president, vice president, and treasurer be selected from the club members?

Exercise 4

A password consists of 6 characters, where each character can be a lowercase letter (a-z) or a digit (0-9). How many different passwords can be created if repetition of characters is allowed?

Exercise 5

In how many ways can you arrange the letters in the word "STATISTICS"?

Exercise 6

A group of 8 people consists of 5 men and 3 women. In how many ways can a committee of 4 people be formed if it must contain at least 2 women?

Exercise 7

How many different 5-digit numbers can be formed using the digits 1, 2, 3, 4, 5 if repetition of digits is not allowed?

Exercise 8

A deck of playing cards contains 52 cards. In how many ways can you choose a 5-card hand?

Chapter 6

Probability Space

6.1 Random experiment

Definition

An experiment or phenomenon is said to be random if the outcome cannot be predicted with certainty. In other words, if all outcomes of this experiment are determined by chance.
Probability theory enables the modeling and simulation of such experiments.

Example.

- We roll a die and observe the result obtained.
- If we toss a coin three times in a row.
- We toss a coin until the head side appears for the first time.

6.2 Sample space and events

Definition

The set, generally denoted as Ω, consisting of all the outcomes of a random experiment is called the sample space or the space of possible outcomes of that experiment.

Remark 13. *The set Ω can be finite ((examples a) and b) or infinite (example*

c).

Definition

An event is any subset of Ω.
An event that contains a single element of Ω is called an elementary event.

Example.

If you roll a 6-sided rigged die : $\Omega = \{1, 2, 3, 4, 5, 6\}$

- A is the event "an even number is rolled," so $A = \{2, 4, 6\}$.
- B is the event "an odd number is rolled," so $B = \{1, 3, 5\}$.
- C is the event "a number greater than or equal to 4 is rolled," so $C = \{4, 5, 6\}$.
- D is the elementary event "the smallest number," so $D = \{1\}$.

6.3 Contradictory and incompatible events

Definition

Let A be an event of a sample space Ω. The complementary event of A is the event consisting of outcomes from Ω that do not occur in A, and it is denoted by $\overline{A}$.

Remark 14. *We have* $A \cap \overline{A} = \emptyset$ *and* $A \cup \overline{A} = \Omega$.

Example.

We roll a fair six-sided die, where the faces are numbered from 1 to 6. We consider the events:

- A: "the face that appears is a multiple of 3".
- B: "the face that appears is not a multiple of 3".

A and B are obviously complementary events.

Definition

Let A and B be two events in the same sample space. When no outcome realizes both event A and event B simultaneously, we say that events A and B are incompatible. In this case, $A \cap B = \emptyset$.

Example.

In an urn, 10 cards are placed, each bearing a number from 1 to 10. A card is drawn from the urn. We consider the events:

- C: "the drawn card bears an even number".
- D: "the drawn card bears an odd number".

Events C and D are incompatible.

6.4 Complete event system

The set $A_1, A_2, \ldots, A_n$ constitutes a complete event system if and only if:

- None of the events is impossible: $A_i \neq \emptyset$, for $i = 1, 2, \ldots, n$.
- The events are pairwise incompatible: $A_i \cap A_j = \emptyset$, for $i \neq j = 1, 2, \ldots, n$.
- The union of the events is the sample space: $\bigcup_{i=1}^{n} A_i = \Omega$.

Example.

If we play with two dice, one white and one red, the random experiment can be modeled by the set of pairs (a, b) where a is the number obtained with the white die and b is the number obtained with the red die. In other words, $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$. Consider the events A_i, $2 \leq i \leq 12$, defined by: A_i: "the sum of the points scored equals i". The A_i form, a complete event system.

6.5 Algebra and σ-algebra of events

Definition

A set F of subsets of a set Ω is an algebra if it satisfies the following three conditions:

i) $\Omega \in F$.

ii) $A \in F \Rightarrow \bar{A} \in F$.

iii) $A, B \in F \Rightarrow A \cap B \in F$.

Example.

a) $\{\emptyset, \Omega\}$ is an algebra on Ω.

b) If $A \subseteq \Omega$, $\{\emptyset, A, \bar{A}, \Omega\}$ is an algebra on Ω.

Proposition

If F is an algebra on Ω, then:

i) $\emptyset \in F$.

ii) $\forall A, B \in F, A \cup B \in F$.

Proof

i) $\Omega \in F \Rightarrow \bar{\Omega} = \emptyset \in F$.

ii) $\forall A, B \in F \Rightarrow \bar{A}, \bar{B} \in F \Rightarrow \bar{A} \cap \bar{B} \in F \Rightarrow A \cup B \in F$.

Remark 15. *if* $A_1, A_2, \ldots, A_n \in F$, *then* $\bigcap_{i=1}^{n} A_i \in F$ *and* $\bigcup_{i=1}^{n} A_i \in F$.

The fact that F *is an algebra does not imply that the union or intersection of an infinite collection* $A_1, A_2, \ldots$ *of events are also in* F. *However, many important events are expressed as the union or intersection of an infinite number of events.*

Definition

A is an σ-algebra (or sigma-algebra) on Ω if

i) A is an algebra on Ω,

ii) $\forall\{A_i\}_{i=1}^{\infty} \in A \Rightarrow \bigcap_{i=1}^{\infty} A_i \in A$.

Proposition

$\forall\{A_i\}_{i=1}^{\infty} \in A \Rightarrow \bigcup_{n=1}^{\infty} A_i \in A$.

Example.

We roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$;
the family $A = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$ is a sigma-algebra.

Definition

The pair (Ω, A) is called a measurable space.

Example.

On $\Omega = \{1, 2, 3, 4, 5, 6\}$ equipped with the power set $\mathcal{P}(\Omega)$, $\{\Omega, \mathcal{P}(\Omega)\}$ is g is a probabilistic space.

6.6 Realization of an event

Definition

Let $\omega \in \Omega$ be the observed outcome of a random experiment, if $\omega \in A$, we say, that A is realized.

Proposition

Let $(\Omega, \mathcal{A})$ be a probabilistic space.

- $\emptyset$ is an event never realized (impossible).
- Ω is an event always realized (certain).
- A realized, $\Leftrightarrow$ $\bar{A}$ not realized.
- A and B realized$\Leftrightarrow$ $A \cap B$ realized.
- A realized or B realized $\Leftrightarrow$ $A \cup B$ realized.

Proof

- $\emptyset = \bar{\Omega}$, so $\emptyset \in \mathcal{A}$; therefore, $\emptyset$ is an event and on the other hand, $\omega \neq \emptyset$.
- $\omega \in \Omega$, so Ω is always realized.
- A realized, $\Leftrightarrow \omega \in A \Leftrightarrow \omega \notin \bar{A} \Leftrightarrow \bar{A}$ not realized.
- A or B realized $\Leftrightarrow \omega \in A$ or $\omega \in B \Leftrightarrow \omega \in A \cup B \Leftrightarrow A \cup B$ realized.
- A and B realized $\Leftrightarrow \omega \in A$ and $\omega \in B \Leftrightarrow \omega \in A \cap B \Leftrightarrow A \cap B$ realized.

6.7 Probability construction

Definition

A probability on the probability space $(\Omega, \mathcal{A})$ is a function $P : \mathcal{A} \to [0, 1]$ that satisfies the following two axioms:

- $P(\Omega) = 1$,
- If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint elements of $\mathcal{A}$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

Definition

$(\Omega, \mathcal{A}, P)$ is called a probability space.

6.8 Properties of a probability

- **Non-negativity**: For any event A in the sample space, the probability $P(A)$ is always non-negative: $P(A) \geq 0$.

- **Normalization**: The probability of the entire sample space is $P(\Omega) = 1$.

- **Additivity of Disjoint Events**: If A and B are disjoint events (i.e., mutually exclusive), then the probability of their union is the sum of their individual probabilities: $P(A \cup B) = P(A) + P(B)$.

- **Complement**: The probability of the complement of an event A is 1 minus the probability of A: $P(\bar{A}) = 1 - P(A)$.

- **Monotonicity**: If $A \subseteq B$ (i.e., A is a subset of B), then $P(A) \leq P(B)$.

- **Countable Additivity**: For any countable sequence $A_1, A_2, \ldots$ of pairwise disjoint events, the probability of their union is the sum of their individual probabilities: $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$.

Proof

- **Non-negativity**: Since probabilities represent proportions of outcomes, they cannot be negative. Therefore, $P(A) \geq 0$ for any event A.
- **Normalization**: Since the sample space represents all possible outcomes, the probability of the sample space must be 1. Mathematically, $P(\Omega) = 1$.
- **Additivity of Disjoint Events**: Let A, and B be disjoint events. Then, $A \cap B = \emptyset$, meaning they have no common outcomes. Therefore, $P(A \cup B) = P(A) + P(B)$. This can be extended to countable disjoint events using induction.
- **Complement**: The complement of A is the set of all outcomes not in A, denoted $\bar{A}$. Since A and $\bar{A}$ together make up the entire sample space, $P(A) + P(\bar{A}) = P(\Omega)$. Since $P(\Omega) = 1$, we have $P(\bar{A}) = 1 - P(A)$.
- **Monotonicity**: If $A \subseteq B$, then $B = A \cup (B \setminus A)$. Since A and $B \setminus A$ are disjoint, $P(B) = P(A) + P(B \setminus A) \geq P(A)$.
- **Countable Additivity**: Let $A_1, A_2, \ldots$ be a countable sequence of pairwise disjoint events. Then, $P\left(\bigcup_{n=1}^{\infty} A_n\right) = P(A_1) + P(A_2) + \ldots$ because the events are disjoint. This can be formally proven using the properties of limits and series in measure theory.

6.9 A finite probability space

Definition

Let $\Omega = \{\omega_1, ...\omega_n\}$ be a finite probability space, and let $p_1 = P(\{\omega_1\}), \ldots, p_n = P(\{\omega_n\})$ be the probabilities of the individual outcomes in Ω. The sequence of numbers $p_i, i = 1, \ldots, n$ is called the probability distribution on Ω.

Proposition

The numbers $p_i, i = 1, 2, \ldots, n$ satisfy the following properties:

1. $0 \leq p_i \leq 1$ for all $i \in \{1, 2, \ldots, n\}$.
2. $\sum_{i=1}^{n} p_i = 1$.

Proof

Let P be a probability measure on Ω, and let $p_i = P(\{\omega_i\})$.

1. It is evident that $0 \leq p_i \leq 1$ for all $i \in \{1, 2, \ldots, n\}$.
2. We have $\Omega = \bigcup_{i=1}^{n} \{\omega_i\}$, so $P(\Omega) = P\left(\bigcup_{i=1}^{n} \{\omega_i\}\right) = \sum_{i=1}^{n} P(\{\omega_i\}) = \sum_{i=1}^{n} p_i = 1$.

Definition

Equiprobability or equiprobability is said to exist when all elementary events have the same probability. It is also said that the probability is equi-distributed.

Proposition

If the probability is equiprobabilistic on Ω, for any event A, we have $P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$.

Proof

If the probability is equiprobabilistic on Ω, then $p \times n = 1$, hence $p = \frac{1}{n}$.
So for any $A = \bigcup_{i=1}^{k} \{\omega_i\}$, where $k \leq n$, we have $P(A) = \sum_{i=1}^{k} P(\{\omega_i\}) = \sum_{i=1}^{k} p = p \times k = \frac{k}{n} = \frac{\text{card}(A)}{\text{card}(\Omega)}$.
In other words, $P(A) = \frac{\text{number of favorable cases for } A}{\text{total number of possible cases}}$.

Example.

We toss a coin three times. What is the probability of getting a head on the first and third toss?
Solution: We can model this experiment by taking $\Omega = \{(H,H,H),(H,T,H),(H,H,T),(T,H,H),(T,T,H),(H,T,T),(T,H,T),(T,T,T)\}$ and for the family of observable events $A = P(\Omega)$ (the set of all parts of Ω). Since the coin is assumed to be symmetrical, we have no reason to suppose that any of the 8 possible triplets of results are favored or disfavored compared to the others. We will therefore choose P so that all elementary events have the same probability (equiprobability hypothesis), i.e., $P(\{\omega\}) = \frac{1}{8}$. The event we want to calculate the probability for is $\{(H,T,H),(H,H,H)\}$. Hence:

$$P(\{(H,T,H),(H,H,H)\}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

.

Example.

We ask a candidate to fill out a questionnaire with 20 binary questions. What is the probability that a candidate answering at random will obtain at least 16 correct answers?
Solution: Here, we choose $\Omega = \{\text{yes}, \text{no}\}^{20}$. If the candidate answers completely randomly, we can consider that each of the 2^{20} possible answer sheets has the same probability of appearing (equiprobability hypothesis on Ω). For any B in Ω,

$$P(B) = \frac{\text{Card}(B)}{\text{Card}(\Omega)}$$

. In particular, for $B =$ "obtaining at least 16 correct answers":

$$P(B) = \frac{C_{16}^{20} + C_{17}^{20} + C_{18}^{20} + C_{19}^{20} + C_{20}^{20}}{2^{20}} = \frac{6196}{2^{20}} = 0.006$$

.

6.10 Conditional probabilities

The concept of conditional probability is essential in probability calculus. It naturally arises when, during a random experiment, "partial information" is provided to the experimenter.

Definition

Let $(\Omega, \mathcal{A}, P)$ be a probability space and B be an event with non-zero probability. For any event A, the conditional probability of A given B (i.e., given that B has occurred) is the real number denoted by $P(A/B)$ defined as:

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

It can be verified that the function $P(\cdot/B) : \mathcal{A} \to [0, 1]$ is a probability measure on $(\Omega, \mathcal{A})$.

Remark 16. *$P(A/B)$ can be interpreted as the event space Ω restricted to B and the outcomes of A restricted to A/B.*

Remark 17. *If A and B are incompatible, and if B has already occurred, then $P(A/B) = \frac{P(A \cap B)}{P(B)} = 0$.*

Remark 18. *In general, $P(A/B) \neq P(B/A)$.*

Example.

Let's consider the experiment of rolling a fair six-sided die: $\Omega = \{1, 2, 3, 4, 5, 6\}$, and let the events be:

$$A : \text{"2 is rolled"} \quad \text{and} \quad B : \text{"an even number is rolled"}$$

So, $P(A) = \frac{1}{6}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{6}$.

- The probability that "2 is rolled" given that it is "an even number" is:
$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$
- The probability that "an even number is rolled" given that it is a "2" is:
$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/6} = 1$$

In this case, $P(A/B) \neq P(B/A)$.

6.11 Compound probabilities and total probabilities

Proposition

(Compound Probability Formula)

1. If A and B are two events such that $P(B) \neq 0$, then $P(A \cap B) = P(B) \cdot P(A/B)$.
2. If A and B are two events such that $P(A) \neq 0$, then $P(A \cap B) = P(A) \cdot P(B/A)$.
3. If $A_1, A_2, \ldots, A_n$ are n events such that $P(A_1 \cap A_2 \cap \ldots \cap A_{n-1})$ then $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \ldots P(A_n/A_1 \cap A_2 \cap \ldots \cap A_{n-1})$

Proof

1 and 2 result from $P(A/B) = \frac{P(A \cap B)}{P(B)}$ and $P(B/A) = \frac{P(A \cap B)}{P(A)}$:
3. proof by recurrence

Proposition

(Total Probability Formula)
Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\{A_i\}_{i \in I}$, where $I \subset \mathbb{N}$, be a complete system of events with non-zero probability, i.e., $\Omega = \bigcup_{i \in I} A_i$, where the A_i are pairwise disjoint and $P(A_i) \neq 0$ for all $i \in I$. Then, for any event B, we have

$$P(B) = \sum_{i \in I} P(B \cap A_i) = \sum_{i \in I} P(A_i)P(B/A_i)$$

Proof

We have

$$B = B \cap \Omega = B \cap \left(\bigcup_{i \in I} A_i \right) = \bigcup_{i \in I} (B \cap A_i)$$

Since the events A_i are pairwise disjoint, the sets $B \cap A_i$ are also pairwise disjoint. Therefore, by the additivity of probabilities,

$$P(B) = P\left(\bigcup_{i \in I} (B \cap A_i) \right) = \sum_{i \in I} P(B \cap A_i)$$

Now, using the definition of conditional probability, we have $P(B \cap A_i) = P(A_i) \cdot P(B/A_i)$. Substituting this into the above equation, we get

$$P(B) = \sum_{i \in I} P(A_i) \cdot P(B/A_i)$$

Thus, we have shown that,

$$P(B) = \sum_{i \in I} P(B \cap A_i) = \sum_{i \in I} P(A_i) \cdot P(B/A_i)$$

which completes the proof.

Example.

Consider a simple scenario where a factory produces three types of products: A, B, and C. The quality control department categorizes each product as either "Good" or "Defective". Now, let's say the probability of each product being defective is different:

- For product A, the probability of being defective is $P(D/A) = 0.1$.
- For product B, the probability of being defective is $P(D/B) = 0.2$.
- For product C, the probability of being defective is $P(D/C) = 0.15$.

Additionally, let's assume the probabilities of producing each type of product are:

- $P(A) = 0.3$
- $P(B) = 0.5$
- $P(C) = 0.2$

We want to find the probability of randomly selecting a defective product from the entire production.
Using Proposition 6.11.2, we can calculate this as follows:

$$P(D) = \sum_{i \in I} P(D \cap A_i) = P(A) \cdot P(D/A) + P(B) \cdot P(D/B) + P(C) \cdot P(D/C)$$

$$= (0.3 \times 0.1) + (0.5 \times 0.2) + (0.2 \times 0.15) = 0.03 + 0.1 + 0.03 = 0.16$$

So, the probability of randomly selecting a defective product from the entire production is 0.16 or 16%.

Proposition

(Bayes' Formula) For any event B with nonzero probability and any event A_i from a complete system of events $\{A_i\}_{i \in I}$ such that $P(A_i) \neq 0$ for all $i \in I$, we have

$$P(A_i/B) = \frac{P(B/A_i) \cdot P(A_i)}{\sum_{i \in I} P(B/A_i) \cdot P(A_i)}$$

for each $i \in I$.

Proof

Consider the event B with nonzero probability. By the definition of conditional probability, we have

$$P(A_i \cap B) = P(B/A_i) \cdot P(A_i)$$

Summing over all events A_i in the complete system of events $\{A_i\}_{i \in I}$, we get

$$P(B) = \sum_{i \in I} P(A_i \cap B) = \sum_{i \in I} P(B/A_i) \cdot P(A_i)$$

Now, using the definition of conditional probability, we have

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B/A_i) \cdot P(A_i)}{\sum_{i \in I} P(B/A_i) \cdot P(A_i)}$$

Remark 19. *This formula is often interpreted as follows: the A_i's represent the different causes that can lead to the occurrence of B. Knowing the probabilities $P(A_i)$ of each case and the conditional probabilities $P(B/A_i)$ of B given the causes A_i, we can calculate the probability $P(A_i/B)$ that the occurrence of B is due to cause A_i.*

Example.

A blood test has a probability of 0.95 of detecting a certain virus when it is actually present. However, it still gives a false positive result for 1% of non-infected individuals. If 0.5% of the population carries the virus, what is the probability that a person has the virus, knowing that they have a positive test result?
To find the probability that a person has the virus given that they have tested positive, we can use Bayes' theorem. Let's denote:

- $P(V)$ as the probability that a person has the virus (prior probability).
- $P(T/V)$ as the probability of testing positive given that the person has the virus, which is the sensitivity of the test. In this case, it's 0.95.
- $P(T/\bar{V})$ as the probability of testing positive given that the person does not have the virus, which is the false positive rate. In this case, it's 0.01.
- $P(\bar{V})$ as the probability that a person does not have the virus, which is $1 - P(V)$.

Now, let's calculate:

$$P(V) = 0.005 \quad (0.5\% \text{ of the population carries the virus})$$
$$P(\bar{V}) = 1 - P(V) = 1 - 0.005 = 0.995$$
$$P(T) = P(T/V) \cdot P(V) + P(T/\bar{V}) \cdot P(\bar{V})$$
$$= 0.95 \times 0.005 + 0.01 \times 0.995$$
$$= 0.00475 + 0.00995$$
$$= 0.0147$$

Now, using Bayes' theorem:

$$P(V/T) = \frac{P(T/V) \cdot P(V)}{P(T)} = \frac{0.95 \times 0.005}{0.0147} = \frac{0.00475}{0.0147} \approx 0.323$$

So, the probability that a person has the virus given that they have tested positive is approximately 0.323 or 32.3%.

6.12 Independent Events

Definition

Given a probability space $(\Omega, \mathcal{A}, P)$, two events A and B are said to be independent (in probability) if and only if $P(A \cap B) = P(A) \cdot P(B)$.

Remark 20. *The independence relation is symmetric: if A is independent of B, then B is independent of A.*

Remark 21. *If one of the two events A or B is nearly impossible, then A and B are independent. Indeed, suppose $P(A) = 0$; as $A \cap B \subseteq A$, we have $0 \leq P(A \cap B) \leq P(A)$, so $P(A \cap B) = 0$, and thus $P(A \cap B) = P(A) \cdot P(B) = 0$.*

Example.

Consider two successive flips of a coin.
$\Omega = \{(H,H),(T,T),(H,T),(T,H)\}$. Let events A and B be defined as follows:

$$A : \text{"head on the first flip"} = \{(H,H),(H,T)\}$$

$$B : \text{"tail on the second flip"} = \{(T,T),(H,T)\}$$

$$A \cap B : \text{"head on the first flip and tail on the second flip"} = \{(H,T)\}$$

We have: $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{4}$. Since these probabilities satisfy the equality $P(A \cap B) = P(A) \cdot P(B)$, A and B are independent events.

Remark 22. *Do not confuse independence and incompatibility of A and B! Note that independence is relative to the chosen probability P, while incompatibility ($A \cap B = \emptyset$) is not.*

Remark 23. *The independence of two events A and B is not an intrinsic property of the events; it is always relative to the probability space $(\Omega, \mathcal{A}, P)$ that has been chosen.*

Example.

An urn contains 12 numbered balls from 1 to 12. One ball is drawn at random, and the events considered are:

A : "drawing an even number" and B : "drawing a multiple of 3"

The naturally imposed probability space here is $\Omega = \{1, 2, \ldots, 12\}$ equipped with the equiprobability P. The events A and B are expressed as follows:

$$A = \{2, 4, 6, 8, 10, 12\}, \quad B = \{3, 6, 9, 12\}, \quad A \cap B = \{6, 12\}$$

We have:

$$P(A) = \frac{6}{12} = \frac{1}{2}, \quad P(B) = \frac{4}{12} = \frac{1}{3}, \quad P(A \cap B) = \frac{2}{12} = \frac{1}{6}$$

and $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{3}$. Therefore, A and B are independent.
Now, we add a ball numbered thirteen to the urn and repeat the experiment. The events A and B remain the same, but the model has changed. We now have equiprobability P' on $\Omega = \{1, 2, \ldots, 13\}$ with:

$$P'(A) = \frac{6}{13}, \quad P'(B) = \frac{4}{13}, \quad P'(A \cap B) = \frac{2}{13}$$

However, $P'(A) \cdot P'(B) = \frac{6}{13} \cdot \frac{4}{13} \neq P'(A \cap B)$. Therefore, A and B are no longer independent.
A little reflection allows us to relate these computational results to the intuitive notion of independence. In the first case, the proportion of multiples of three among the even numbers is the same as among the odd numbers. Knowing that the drawn ball it even does not change our information about B. On the other hand, in the second case, adding the thirteenth ball changes the proportion of multiples of three: it is higher among the even numbers than among the odd numbers. Therefore, knowing that the drawn ball is even slightly increases the probability we can attribute to B.

Proposition

If A and B are independent, the same is true for the pairs of events A and B', A' and B, A' and B'.

6.13 Mutual independence

More generally, the mutual independence of several events is defined as follows.

Definition

Let $(\Omega, \mathcal{A}, P)$ be a probability space and n events $A_1, A_2, \ldots, A_n$. These n events are said to be (mutually) independent if and only if, for any partition such that $I \subseteq \{1, 2, \ldots, n\}$, we have:

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

Example.

A player rolls a 6-sided die three times. Let's find the probability that they obtain an even number on each roll. Let the events be defined as follows:

A_1 : "even number on the first roll"

A_2 : "even number on the second roll"

A_3 : "even number on the third roll"

$A_1 \cap A_2 \cap A_3$: "even number on each roll"

We have $P(A_1) = \frac{1}{2}$, $P(A_2) = \frac{1}{2}$, $P(A_3) = \frac{1}{2}$, and $P(A_1 \cap A_2 \cap A_3) = \frac{1}{4} = P(A_1) \cdot P(A_2) \cdot P(A_3)$. Therefore, A_1, A_2, and A_3 are independent events.

Remark 24. *Mutual independence obviously implies pairwise independence, but the converse is false.*

Example.

An urn contains four tokens: one blue, one white, one red, and one blue-white-red. One token is drawn at random. Let's consider the three events:

$$A : \text{"the drawn token contains blue"}$$

$$B : \text{"the drawn token contains white"}$$

$$C : \text{"the drawn token contains red"}$$

It is clear that $P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}$. Furthermore, $P(A \backslash B) = P(\text{tricolor}) = \frac{1}{4} = P(A) = P(B)$, and similarly $P(B \backslash C) = \frac{1}{4} = P(B) \cdot P(C)$, $P(C \cap A) = \frac{1}{4} = P(C) \cdot P(A)$. Thus, the events A, B, and C are pairwise independent. However, $P(A \backslash B \cap C) = 1$ because $B \cap C$ implies "tricolor". Therefore, knowledge of the simultaneous occurrence of B and C changes our information about A. The notion of pairwise independence is not sufficient to capture the intuitive idea of independence among several events.

6.14 Exercises

Exercise 1

Let S be the set of married couples in a given city. Consider the events:

- A :The husband is over forty years old
- B :The wife is younger than the husband
- C :The wife is over forty years old
- Interpret in terms of A, B, C the event that the husband is over forty years old, but not his wife.
- Describe in ordinary language the events $A \cap \bar{B}$, $A \cap B \cap \bar{C}$ and $A \cup B \cup \bar{C}$.
- Verify that $A \cap \bar{C}$ is included in B

Exercise 2

Let A, B, C be three events. Find the set expressions for the following events:

- A occurs alone,
- A and B occur but C does not occur,
- at least one event occurs,
- at most one event occurs,
- at least two events occur,
- all three events occur,
- no event occurs,
- at most two events occur.

Exercise 3

A coin and a die are tossed.

- What is the sample space?

- Explicitly express the events:
  - "a face with an even number appears"
  - "a 1 appears"
  - "a tail with an odd number appears"
- Express in the same way: A or B occurs, B or C occur, B alone occurs.

Exercise 4

Let A1 and A2 be two tribes. Show that their intersection is a tribe.

Exercise 5

Is the union of two tribes also a tribe?

Exercise 6

Let E be a set of n elements where n = 1, 2, 3. How many different tribes are there on E?

Exercise 7

Consider the set of integers from 20 to 40. One of these numbers is chosen at random. Let:

- A: "the number is a multiple of 3",
- B: "the number is a multiple of 3",
- C: "the number is a multiple of 3".

Calculate $P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cup B)$, $P(A \cap \bar{C})$, $P(\bar{A} \cap \bar{C})$

Exercise 8

We randomly select 6 light bulbs from a batch of 15 light bulbs, of which 5 are defective. Calculate the probability in % of each of the following events

after giving a set-theoretic transcription:

- No light bulb is defective.
- Exactly one light bulb is defective.
- Exactly two light bulbs are defective.
- Exactly three light bulbs are defective.
- At least one light bulb is defective.
- At least two light bulbs are defective.

Exercise 9

Three horses A, B, and C compete in a race. It is known that horse A has twice the chance of winning compared to horse B, and horse B has twice the chance of winning compared to horse C.

Calculate:

- the probability that horse A wins the race.
- the probability that horse C does not win the race.

Exercise 10

Three men and two women participate in a chess tournament. Individuals of the same gender have equal chances of winning, but a woman has twice the chance of winning compared to a man. Find the probability:

- that a woman wins the tournament,
- that a man wins the tournament.

Exercise 11

Six married couples are in a restaurant.

- If two people are chosen randomly, find the probability that they are married to each other.
- If four people are chosen randomly, find the probability that none of the chosen couples are among these four people.

Exercise 12

An urn contains 5 white balls and 7 red balls. Three draws are made from the urn following the following procedure. At each draw, a ball is taken and returned to the urn with another ball of the same color added. Calculate the probabilities that the sample of three balls drawn contains:

- No white ball.
- Exactly one white ball.
- Three white balls.
- Exactly two white balls.

Exercise 13

An insurance company classifies its clients into three classes R1, R2, and R3: good risks, medium risks, and bad risks. The populations of these three classes represent 20%, 50%, and 30% of the total population, respectively, with the probabilities of having an accident during the year for a person in one of these three classes being 0.05, 0.15, and 0.30, respectively.

- What is the probability that a randomly chosen person from the population will have an accident during the year?
- If Mr. Ali did not have an accident this year, what is the probability that he is a good risk?

Exercise 14

Consider an urn U containing 9 white balls and 1 black ball, and an urn V containing 3 white balls and 7 black balls. A fair six-sided die is rolled, and two draws are made with replacement from urn U if the die shows a one, or from urn V otherwise. Consider the events U: "a ball is drawn from urn U", V: "a ball is drawn from urn V", B_i: "the i^{th} ball is white", and N_i: "the i^{th} ball is black from urn V" for $i = 1, 2$.

- Are events B_1 and N_2 independent?
- Given that a white ball followed by a black ball is drawn, from which urn is it more likely that they were drawn?

Exercise 15

In a class of $N+1$ students, the solution to a probability exercise is found by a single student. It is then passed on to one of his classmates chosen at random. This classmate then passes the solution to one of his classmates chosen at random. This process repeats n times (so there are n transmissions of the solution).

1. Calculate the probability that the solution is not repeated to the student who found it.
2. Calculate the probability that the solution is never repeated to a student who himself passed it on.
3. Repeat question 1. assuming that at each step, the solution is no longer passed on to a single student but to a group of k students ($k \geq 1$) chosen at random.

Conclusion

In conclusion, these lecture notes are a valuable resource for first-year students entering the field of statistics and probability. They provide a comprehensive introduction to foundational concepts and methodologies essential for understanding and analyzing data. From explanations of statistical terms to the presentation of methods for data collection and analysis, and from exploration of combinatorial analysis to probability theory, these notes lay a solid groundwork for tackling the complexities of statistical analysis.

Furthermore, they promote active learning through practical examples and exercises, enabling students to develop critical thinking and problem-solving skills. By encouraging active engagement in the learning process, these notes prepare students to confidently apply statistical methods in real-world situations, whether in scientific research, business analytics, or other fields.

In summary, these well-crafted lecture notes serve as a strong foundation for first-year students in statistics and probability, equipping them with the knowledge and tools necessary for success in their studies and future careers in this field.

Bibliography

1. M., Akinkunmi. Introduction to statistics using R. Springer Nature, 2022.
2. G., O'Regan. Guide to discrete mathematics. Springer International Publishing, 2021.
3. A., S., Bhadouria, and R., K., Singh. "Machine learning model for healthcare investments predicting the length of stay in a hospital mortality rate." Multimedia Tools and Applications 83.9 (2024): 27121-27191.
4. J., Haigh, and J., Haigh. Probability models. Vol. 24. London: Springer, 2002.
5. P., Varga, and G., Kún. "Utilizing higher order statistics of packet interarrival times for bottleneck detection." Workshop on End-to-End Monitoring Techniques and Services, 2005. IEEE, 2005.
6. K.,I., Park, and M. Park. Fundamentals of probability and stochastic processes with applications to communications. Cham, Switzerland: Springer International Publishing, 2018.
7. B., Lawal. Applied statistical methods in agriculture, health and life sciences. Springer, 2014.
8. G. Calot. Cours de statistique descriptive. Dunod, paris 1965.
9. S., M., Ross. Introduction to probability models. Academic press, 2014.
10. J.P. Delmas. Introduction aux probabilités, Ellipses, Paris 1993.
11. J.P. M. Mandry. Probabiliés cours et travaux diriges, Office des publications universitaires, Alger 1982.
12. G. Calot. Cours de calcul des probabilités, Dunod, Paris 1984.
13. G. Calot. Exercice de calcul des probabilités Dunod, Paris 1984.