Lesya Ukrainka Volyn National University
Faculty of Information Technology and Mathematics
Department of General Mathematics and
Methods of Teaching Computer Science

Maria Khomyak

STATISTICS
FOR AN INTERNATIONAL ECONOMIST
Educational Manual

СТАТИСТИКА
ДЛЯ ЕКОНОМІСТА-МІЖНАРОДНИКА
Навчальний посібник

Електронне видання на CD-ROM

Lutsk
Vezha-Print
2025

UDC 31:33(07)
Kh 76

Recommended for publication by the Academic Council of Lesya UkrainkaVolyn National University (Protocol No. 14 dated November 27, 2025)

Reviewers:

Yaroslav Pasternak, Doctor of Physical and Mathematical Sciences, Professor, Professor of the Department of Computer Science and Cybersecurity at Lesya Ukrainka Volyn National University.

Svitlana Popereshniak, Ph.D. in Physical and Mathematical Sciences, Associate Professor of Department of Informatics and Software Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Khomyak Maria

Kh 76 Statistics for an international economist: Educational Manual / Maria Khomyak. – 1 electronic optical disk (CD-ROM). – Data volume 5,65 Mb.

ISBN 978-966-940-707-8

Annotation: The manual "Statistics for an International Economist" is designed for undergraduate students in Specialty 292 "International Economic Relations". It introduces the basics of statistical methods and their applications in international economic practice. The content includes lectures on core statistical concepts, data analysis, probability, sampling, and hypothesis testing, as well as practical tasks that develop skills in data processing, visualization, and interpretation. The manual can be used for both classroom study and independent learning.

Навчальний посібник «Статистика для економіста-міжнародника» призначений для здобувачів першого (бакалаврського) рівня вищої освіти спеціальності 292 «Міжнародні економічні відносини». У ньому подано основи статистичних методів та їх застосування в міжнародній економічній практиці. Зміст посібника охоплює лекції з ключових статистичних понять, аналізу даних, імовірності, вибірки та перевірки статистичних гіпотез, а також практичні завдання, що формують навички опрацювання, візуалізації та інтерпретації даних. Посібник може бути використаний як під час аудиторного навчання, так і для самостійної роботи.

UDC 31:33(07)

ISBN 978-966-940-707-8

CONTENTS

INTRODUCTION

In the modern world, statistics plays a crucial role in the training of future economists, especially in the field of international economics. The ability to collect, analyze, and interpret statistical data is an essential skill for making informed decisions in business, trade, and global financial relations. For a future specialist, statistics is not only a theoretical discipline but also a practical tool that supports evidence-based conclusions and rational problem-solving.

This educational manual is designed for undergraduate students majoring in economics and related fields. Its main purpose is to provide students with a solid foundation in the methods and applications of statistics that are relevant to the international economic environment. The material covers both theoretical aspects, presented in a series of lectures, and practical exercises that develop applied skills.

The lecture part introduces students to the fundamental concepts of statistics, data types, methods of data collection, tabular and graphical presentations, descriptive measures, probability theory, and hypothesis testing. The practical section is focused on tasks that allow students to master essential techniques such as calculating measures of central tendency and variability, constructing frequency distributions, interpreting scatter diagrams, performing regression analysis, and applying probability methods.

Through this combination of theory and practice, students will acquire knowledge and competencies necessary for professional activities in economics. Emphasis is placed on developing critical thinking, analytical reasoning, and the ability to use statistical evidence for solving economic problems in an international context.

The manual may be used both in classroom instruction and for independent study. Each topic is structured to help students gradually build their understanding, while practical assignments reinforce the application of knowledge to real-world economic cases.

1. STATISTICS AND ITS APPLICATIONS IN BUSINESS AND ECONOMICS

1.1. Statistics as a Science and as a Field of Practical Activity

1.2. Applications of Statistics in Business and Economics

1.1. Statistics as a Science and as a Field of Practical Activity

Statistics is a discipline that deals with the collection, analysis, interpretation, presentation, and organization of data. It serves as both a theoretical science and a practical tool for solving real-world problems.

Statistics as a Science

Statistics as a science involves the development of theories, methods, and mathematical frameworks to analyze data. Its primary aim is to derive meaningful insights and make predictions based on data patterns.

Key Characteristics of Statistics as a Science:

1. Mathematical Foundation:

Heavily rooted in probability theory, calculus, and linear algebra.

Develops models to describe random processes and uncertainties.

2. Conceptual Frameworks:

Hypothesis testing, estimation theory, and data distributions are central.

Involves exploring causation and correlation.

3. Generality:

The principles and theories of statistics can be universally applied across disciplines like biology, economics, sociology, and engineering.

Statistics has vast practical applications, enabling informed decision-making in various sectors. It plays a pivotal role in business, governance, healthcare, technology, and beyond.

Key Roles in Practice:

1. Data Collection and Management: surveys, censuses, and observational studies to gather reliable data; ensuring accuracy and minimizing biases in data collection.

2. Data Analysis and Interpretation: Use of software and tools to extract actionable insights; application of descriptive and inferential statistical methods.

3. Decision Support: provides evidence-based solutions to policy-makers, business leaders, and scientists; Risk analysis, forecasting, and quality control.

Real-World Applications:

Business and Economics: Market analysis and studies of consumer behavior, and financial risk assessment.

Healthcare: Epidemiology, drug efficacy testing, and public health planning.

Technology: Artificial intelligence, machine learning, and big data analytics.

Public Policy: Socioeconomic planning, unemployment analysis, and crime rates. studies.

Integration of Science and Practice

Statistics bridges theory and application, ensuring that mathematical principles are translated into practical outcomes. The feedback loop from practical applications often inspires new theoretical advancements.

Statistics, as a science, provides the theoretical foundation for understanding randomness and patterns. As a field of practical activity, it empowers individuals and organizations to make data-driven decisions in complex environments. This dual nature makes statistics indispensable in both academic and applied settings.

1.2. Applications of Statistics in Business and Economics

In today's global business and economic environment, anyone can access vast amounts of statistical information. The most successful managers and decision-makers understand the information and know how to use it effectively.

Statistics plays a critical role in business and economic sciences, providing tools and methods to analyze data, forecast trends, and make evidence-based decisions. Its applications span various domains within these fields.

Let’s provide some examples that illustrate some of the uses of statistics in business and economics.

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether the amount of accounts receivable shown on a client’s balance sheet fairly represents the actual amount of accounts receivable.

Usually the large number of individual accounts receivable makes reviewing and validating every account too time-consuming and expensive. As common practice in such situations, the audit staff selects a subset of the accounts called a sample. After reviewing the accuracy of the sampled accounts, the auditors draw a conclusion as to whether the accounts receivable amount shown on the client’s balance sheet is acceptable.

Finance

Financial analysts use a variety of statistical information to guide their investment recommendations. In the case of stocks, the analysts review a variety of financial data including price/earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock market averages, a financial analyst can begin to draw a conclusion as to whether an individual stock is over- or under-priced. Similarly, historical trends in stock prices can provide a helpful indication on when investors might consider entering (or re-entering) the market.

Marketing

Electronic scanners at retail checkout counters collect data for a variety of marketing research applications. For example, data suppliers such as ACNielsen purchase point-of-sale scanner data from grocery stores, process the data and then sell statistical summaries of the data to manufacturers. Manufacturers spend vast amounts per product category to obtain this type of scanner data. Manufacturers also purchase data and statistical summaries on promotional activities such as special pricing and the use of in-store displays. Brand managers can review the scanner statistics and the promotional activity statistics to gain a better understanding of the relationship between promotional activities and sales. Such analyses often prove helpful in establishing future marketing strategies for the various products.

Production

Today's emphasis on quality makes quality control an important application of statistics in production. A variety of statistical quality control charts are used to monitor the output of a production process. In particular, a bar chart can be used to monitor the average output. Suppose, for example, that a machine fills containers with 330g of a soft drink. Periodically, a production worker selects a sample of containers and computes the average number of grams in the sample. This average, or bar value, is plotted on a bar chart. A plotted value above the chart's upper control limit indicates overfilling, and a plotted value below the chart's lower control limit indicates underfilling. The process is termed 'in control' and allowed to continue as long as the plotted x-bar values fall between the chart's upper and lower control limits. Properly interpreted, a bar chart can help determine when adjustments are necessary to correct a production process.

Economics

Economists frequently provide forecasts about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts. For instance, in forecasting inflation rates, economists use statistical information on such indicators as the Producer Price Index, the unemployment rate and manufacturing capacity utilization. Often these statistical indicators are entered into computer-ized forecasting models that predict inflation rates.

Applications of statistics such as those described in this section are an integral part of this text. Such examples provide an overview of the breadth of statistical applications. To supplement these examples, chapter-opening Statistics in Practice articles obtained from a variety of topical sources are used to introduce the material covered in each chapter. These articles show the importance of statistics in a wide variety of business and economic situations.

Statistics provides a robust framework for solving complex problems in business and economics. By applying statistical methods, businesses can improve efficiency and profitability, while economists can better understand market dynamics and craft effective policies. The synergy between statistical science and practical application drives innovation and progress in these fields.

2. INTRODUCTION TO STATISTICS: CORE CONCEPTS, DATA, AND SURVEY TECHNIQUES

2.1. Subject of Statistics and data

The primary subject of statistics is *data* – its collection, organization, analysis, interpretation, and presentation. It aims to uncover patterns, trends, and relationships within data to make informed decisions and predictions under uncertainty.

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the data set for the study.

Types of data

Data is divided into two main types: *qualitative* (or *categorical*) and *quantitative* (or *numerical*).

For example, responses to yes / no questions are categorical. Are you a business major? and Do you own a car? are limited to yes or no answers. A health care insurance company may classify incorrect claims according to the type of errors, such as procedural and diagnostic errors, patient information errors, and contractual errors. Other examples of categorical variables include questions on gender or marital status. Sometimes categorical variables include a range of choices, such as "strongly disagree" to "strongly agree." For example, consider a faculty-evaluation form where students are to respond to statements such as the following: The instructor in this course was an effective teacher (1: strongly disagree; 2: slightly disagree; 3: neither agree nor disagree; 4: slightly agree; 5: strongly agree).

Collected information such as colors, names, telephone numbers, dates of birth among others with no numerical values are classified as qualitative data. Other types of data such as ages, salaries, exam scores, number of siblings, etc. are classified as quantitative.

Quantitative data are divided into two categories *discrete* and *continuous*.

- When the items can assume a countable number of possible values the data is referred to as *discrete*.

- When the items may take any value in a given interval or intervals of the set of real numbers the data is referred to as *continuous*.

Discrete are countable values while continuous are measurable values.

Examples of discrete numerical data include the number of students enrolled in a class, the number of university credits or scores earned by a student at the end of a particular semester, and the number of Microsoft stocks in an investor's portfolio while weight, height, and time are examples of continuous data.

2.2. Other Basic Concepts of Statistics

Many situations require data for a large group of elements (individuals, companies, voters, house-holds, products, customers and so on). Because of time, cost and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a particular study is called the population, and the smaller group is called the sample. Formally, we use the following definitions.

Population is the entire set of individuals, items, or observations of interest. All residents of a country is an example of population.

Sample is a subset of the population used for analysis. Example: Surveying 1,000 residents of a country.

Population size, *N*, can be very large or even infinite. A sample is an observed subset (or portion) of a population with sample size given by *n*.

Examples of populations include the following:

- All potential buyers of a new product
- All stocks traded on the NYSE Euronext
- All registered voters in a particular city or country
- All accounts receivable for a corporation

The process of conducting a survey to collect data for the entire population is called a census. The process of conducting a survey to collect data for a sample is called a sample survey.

Elements are the entities on which data are collected.

Variable is a characteristic or attribute that can vary between individuals or observations. A variable is a characteristic of interest for the elements.

Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an observation.

Descriptive Statistics is concerned with summarizing and organizing data. It includes measures such as:

Central Tendency: Mean, median, and mode.

Dispersion: Range, variance, and standard deviation.

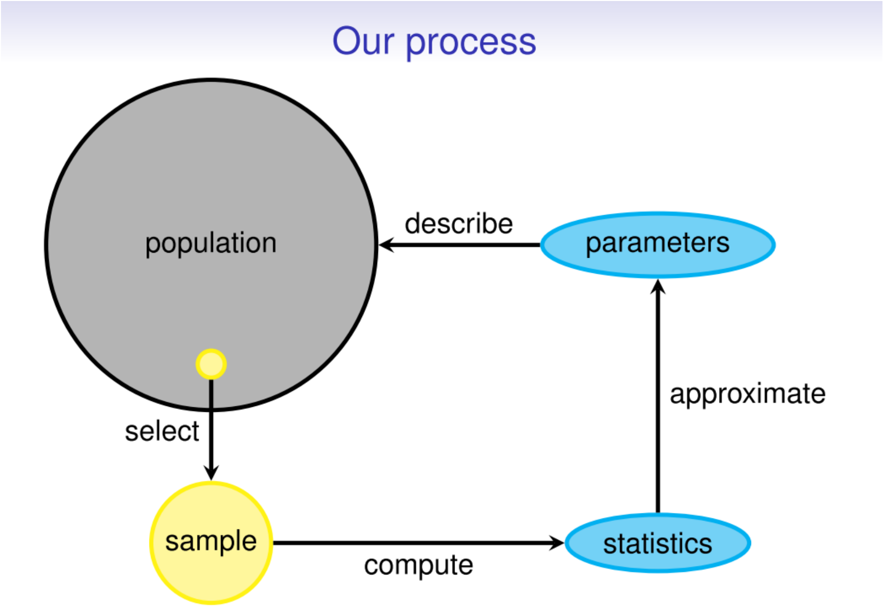
Visualization: Graphs, charts, and tables.

Inferential Statistics involves drawing conclusions and making predictions about a population based on a sample. It includes

Estimation: Point and interval estimates of population parameters

Hypothesis Testing: Testing assumptions about population parameters.

Examples include regression models, time series models, and machine learning algorithms.



Statistics applies across numerous fields, including natural sciences, social sciences, business, economics, medicine, and technology. It serves both descriptive and inferential purposes.

2.3. Types and Methods of Survey

A survey is a systematic method for collecting data from individuals or groups to study behaviors, opinions, characteristics, or conditions. Surveys are widely used in statistics, business, social sciences, healthcare, and more to gather quantitative and qualitative information.

Types of Surveys

Surveys can be categorized based on their purpose, coverage, mode of administration, and frequency.

1. By Purpose

Descriptive Surveys: aim to describe the characteristics of a population or phenomenon.

Example: Demographic surveys to understand population structure.

Analytical Surveys: focus on examining relationships between variables or testing hypotheses.

Example: Customer satisfaction surveys analyzing satisfaction factors.

2. By Coverage

Census: Collects data from the entire population.

Example: National population censuses.

Sample Surveys: Collects data from a representative subset of the population.

Example: Household income surveys.

3. By Administration Mode

Self-Administered Surveys: respondents complete the survey themselves.

Example: Online feedback forms.

Interviewer-Administered Surveys: an interviewer guides respondents through the survey questions.

Example: Telephone interviews or face-to-face interviews.

4. By Frequency

Cross-Sectional Surveys: conducted at a single point in time to capture a snapshot of the population.

Example: Annual consumer spending surveys.

Longitudinal Surveys: conducted over an extended period, collecting data from the same respondents at multiple intervals.

Example: Panel studies on health outcomes.

Repeated Cross-Sectional Surveys: Conducted periodically but with different respondents each time.

Example: Quarterly labor force surveys.

Methods of Survey

The method chosen for a survey depends on its objectives, target audience, and available resources.

1. By Mode of Data Collection

Face-to-Face Interviews: an interviewer directly interacts with respondents.

Advantages: High response rate, allows clarification of questions.

Disadvantages: Expensive, time-consuming.

Example: In-depth interviews for market research.

Telephone Surveys: conducted over the phone, often using standardized questionnaires.

Advantages: Cost-effective, faster than face-to-face surveys.

Disadvantages: Limited to populations with phone access, potential for non-response bias.

Example: Political opinion polls.

Mail Surveys: questionnaires are sent to respondents by post.

Advantages: Low cost, allows respondents to answer at their convenience.

Disadvantages: Low response rates, delayed responses.

Example: Academic research surveys.

Online Surveys: distributed via email, social media, or websites.

Advantages: Fast, cost-effective, easily scalable.

Disadvantages: Limited to internet users, potential for sample bias.

Example: Customer feedback forms on e-commerce platforms.

2. By Questionnaire Design

Structured Surveys: use a standardized questionnaire with predefined questions and response options.

Example: Multiple-choice questions on satisfaction levels.

Semi-Structured Surveys: combine standardized questions with open-ended questions for qualitative insights.

Example: Employee engagement surveys.

Unstructured Surveys: rely on open-ended questions, allowing respondents to answer freely.

Example: Exploratory interviews in qualitative research.

3. By Interaction Level

Direct Surveys: involve direct interaction with respondents, such as interviews.

Indirect Surveys: data is collected through intermediaries or observational methods without direct respondent interaction.

4. By Sampling Technique

Random Sampling Surveys: respondents are selected randomly to ensure representativeness.

Stratified Sampling Surveys: the population is divided into subgroups (strata), and samples are taken from each.

Convenience Sampling Surveys: respondents are selected based on ease of access.

Principles of Effective Survey Design

1. Clear Objectives: Define the purpose and scope of the survey.

2. Targeted Audience: Identify the population and sampling method.

3. Simple Language: Use clear, concise, and unbiased questions.

4. Tested Questionnaire: Pilot-test the survey for reliability and validity.

5. Data Security: Ensure respondent confidentiality and ethical handling of data.

Surveys are a versatile tool for data collection, offering flexibility in their design and administration. By selecting the appropriate type and method, researchers and practitioners can efficiently gather meaningful insights to inform decisions, policies, and strategies.

Statistics is a systematic approach to understanding data and addressing uncertainty. By mastering its basic concepts and methods, statisticians and practitioners can extract meaningful insights, enabling better decision-making and fostering advancements in various fields.

3. FUNDAMENTALS OF STATISTICAL SYSTEMS: LEGAL FRAMEWORK, ORGANIZATION, AND OBSERVATION METHODS

3.1. Normative and Legal Provision of Statistics

Normative and legal provisions of statistics refer to the frameworks, laws, standards, and regulations that govern the collection, processing, analysis, dissemination, and use of statistical data. These provisions ensure that statistical activities are conducted ethically, transparently, and reliably while safeguarding the rights and privacy of individuals and organizations.

Key Components of Normative and Legal Provisions in Statistics

1. **Legislative Framework**

National and international laws define the rights, responsibilities, and obligations of statistical agencies and users of statistical data.

Key areas addressed:

Data collection and reporting requirements.

Confidentiality and data protection.

Standardization of methods and terminology.

2. **Principles of Official Statistics**

Developed by organizations such as the United Nations Statistical Commission to guide statistical practices globally.

Core principles include:

Professional Independence: Statistical bodies operate free from political influence.

Impartiality: Data must be collected and reported objectively.

Confidentiality: Personal data must be protected from unauthorized disclosure.

Accessibility: Statistics should be disseminated widely and fairly.

Use of Sound Methodologies: Standardized and scientifically valid methods must be employed.

3. **Institutional Frameworks**

National Statistical Offices (NSOs) or similar entities are typically tasked with:

Conducting censuses, surveys, and other statistical activities.

Ensuring compliance with statistical laws and regulations.

Coordinating with international statistical organizations.

4. **Data Protection and Confidentiality**

Laws and policies ensure the protection of individual and organizational data.

Examples include: Prohibiting the use of data for non-statistical purposes (e.g., taxation, law enforcement); Imposing penalties for breaches of confidentiality.

5. **Standardization of Statistical Practices**

Norms and guidelines are established to ensure consistency and comparability of data across regions and over time.

Examples of statistical standards:

International Classification of Diseases (ICD).

Standard Industrial Classification (SIC).

System of National Accounts (SNA).

6. **International Legal Provisions**

Various international treaties, conventions, and agreements govern statistical activities globally.

Notable examples include:

Fundamental Principles of Official Statistics (UN).

General Data Protection Regulation (GDPR) in the European Union.

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.

7. **Ethics in Statistics**

Ethical codes are integral to normative provisions, focusing on:

Honesty and transparency in reporting data.

Avoiding misuse or manipulation of statistical results.

Respecting the rights of survey respondents.

Examples of Normative and Legal Provisions in Practice

1. National Statistical Legislation

Countries typically have a statistical act or similar legislation.

Example: The Statistical Act of the United States mandates the Census Bureau to conduct the decennial census.

2. International Cooperation and Standards

The International Monetary Fund (IMF) requires member countries to adhere to the Data Dissemination Standards Initiative (DDSI) for economic data reporting.

3. GDPR and Data Privacy in Statistics

The GDPR regulates the processing of personal data in statistical research within the EU, ensuring that privacy is not compromised.

Challenges and Future Directions

Global Harmonization: aligning statistical standards across countries to enhance data comparability.

Technological Advancements: updating legal provisions to address challenges posed by big data, AI, and real-time data analytics.

Strengthening Data Privacy: balancing the need for open data access with robust data protection measures.

Promoting Statistical Literacy: educating users to understand and responsibly use statistical data.

Normative and legal provisions in statistics ensure that statistical activities are carried out responsibly, ethically, and transparently. By establishing clear frameworks and standards, these provisions foster trust in statistical systems and enable their effective use in decision-making and policy formulation.

3.2. Organization of Statistics in Ukraine and Other Countries

Statistical systems are established to provide reliable, timely, and relevant data for decision-making and policy formulation. These systems operate at national, regional, and international levels, with national statistical offices (NSOs) playing a pivotal role in data collection, analysis, and dissemination.

1. Organization of Statistics in Ukraine

National Statistical Office / State Statistics Service of Ukraine (SSSU): Responsible for organizing and conducting statistical activities in Ukraine. Operates under the Cabinet of Ministers of Ukraine and adheres to the Law of Ukraine on State Statistics.

Data Collection: conducts national censuses (population, agriculture, etc.) and surveys on various social and economic indicators and gathers data on demographics, labor, trade, industry, and the environment.

Data Processing and Analysis: ensures data accuracy and compliance with international statistical standards and employs modern technologies for data processing and publication.

Dissemination: publishes statistical reports, databases, and open-access datasets to inform stakeholders and the public.

Challenges and Goals

- Challenges:

Ensuring data quality amid limited resources and external challenges (e.g., conflict zones).

Transitioning to EU statistical standards (EUROSTAT requirements).

- Goals:

Full harmonization with European and international statistical methodologies.

Expanding digital tools for real-time data collection and reporting.

2. Organization of Statistics in Other Countries

Examples of National Statistical Offices

1. United States:

U.S. Census Bureau (under the Department of Commerce) handles censuses and surveys.

Other agencies, like the Bureau of Labor Statistics (BLS) and National Center for Health Statistics (NCHS), manage specialized statistical domains.

2. United Kingdom:

Office for National Statistics (ONS): Independent statistical body responsible for economic, social, and demographic data.

3. Germany:

Federal Statistical Office (Destatis): Oversees statistics for federal and state governments, adhering to EUROSTAT standards.

4. Japan:

Statistics Bureau of Japan (SBJ): A division under the Ministry of Internal Affairs and Communications, managing censuses and surveys.

Decentralized vs. Centralized Models

Centralized Model: A single agency (e.g., Ukraine, France) is responsible for most statistical activities.

Decentralized Model: Multiple agencies (e.g., the U.S., Germany) handle different domains, coordinated by a central authority or framework.

3. International Statistical Organizations

International organizations establish global standards, promote data comparability, and provide technical support to countries.

Key International Statistical Organizations

1. United Nations Statistical Commission (UNSC):

Oversees international statistical activities and coordinates global standards.

Responsible for the Fundamental Principles of Official Statistics.

2. Eurostat (European Union):

Develops and promotes statistical harmonization across EU member states.

Publishes European-wide data on demographics, economics, and social conditions.

3. International Monetary Fund (IMF):

Focuses on economic and financial statistics and provides the Data Dissemination Standards Initiative (DDSI) to ensure data transparency.

4. World Bank:

Gathers and disseminates global development indicators, such as GDP, poverty rates, and social metrics.

5. Organization for Economic Co-operation and Development (OECD):

Provides statistics on economic performance, trade, education, and innovation among member countries.

6. World Health Organization (WHO):

Collects and analyzes global health data, including disease prevalence, mortality, and healthcare resources.

7. International Labour Organization (ILO):

Publishes labor statistics, such as employment, unemployment, and wages.

8. Food and Agriculture Organization (FAO):

Focuses on agricultural production, food security, and environmental sustainability.

4. Collaboration Between National and International Statistical Systems

Standardization: Adhering to international frameworks like the System of National Accounts (SNA) and International Classification of Diseases (ICD).

Capacity Building: International organizations provide training and resources to developing countries.

Data Sharing: Promotes global transparency and comparability of data for research and policymaking.

The organization of statistics varies across countries, but they share common goals of providing accurate and reliable data. International statistical organizations play a crucial role in harmonizing methods, promoting collaboration, and ensuring the comparability of data worldwide. Together, national and international systems contribute to evidence-based policymaking and global development.

3.3. The Essence and Organizational Forms of Statistical Observation

Essence of Statistical Observation

Statistical observation is the process of systematically collecting data on specific phenomena or processes for further analysis. It serves as the foundation of statistical analysis, enabling researchers to study patterns, trends, and relationships within a population or system.

Key Characteristics of Statistical Observation

1. Systematic Nature: data is collected according to a predefined plan or methodology.

2. Quantitative Focus: observation is primarily aimed at gathering measurable data.

3. Purpose-Driven: the process is designed to address specific research questions or objectives.

4. Objectivity: efforts are made to minimize bias during data collection.

Organizational Forms of Statistical Observation

The organizational form of statistical observation depends on how data is collected, its scope, and its timing. The main organizational forms include:

1. By Coverage: Complete vs. Partial Observation

- **Complete Observation (Census):**

Involves collecting data from every unit in the population.

Common in national censuses (e.g., population or agricultural censuses).

Advantages: High accuracy and detailed data.

Disadvantages: Expensive, time-consuming, and resource-intensive.

- **Partial Observation (Sample Survey):**

Collects data from a subset (sample) of the population.

Uses statistical sampling methods to ensure representativeness.

Advantages: Cost-effective and faster.

Disadvantages: Relies on sampling methods, introducing potential for error.

2. By Timing: Continuous, Periodic, and One-Time Observation

- **Continuous Observation:**

Data is collected continuously over time.

Example: Monitoring traffic flow, stock market prices, or weather conditions.

- **Periodic Observation:**

Conducted at regular intervals (e.g., annually, quarterly).

Example: Quarterly labor force surveys, annual production statistics.

- **One-Time Observation:**

Data is collected once for a specific purpose or event.

Example: Data collected during a disaster impact assessment.

3. By Method of Data Collection

- **Direct Observation:**

Data is collected by directly observing phenomena.

Example: Observing factory production processes.

- **Interview-Based Observation:**

Data is collected through surveys, questionnaires, or interviews.

Example: Household surveys or customer feedback forms.

- **Documentary Observation:**

Data is extracted from existing documents, reports, or records.

Example: Using tax records, school registers, or hospital logs.

- **Self-Reporting:**

Respondents provide data directly, often through forms or digital platforms.

Example: Self-reported income in tax filings.

4. By Methodological Approach

- **Survey Method:**

Structured approach where respondents answer predefined questions.

Example: National health surveys.

- **Experimental Method:**

Data is collected under controlled conditions to study cause-effect relationships.

Example: Clinical trials for new medications.

- **Observational Studies:**

Data is collected passively, without intervention.

Example: Studying consumer behavior in retail stores.

5. By Organizational Structure

Centralized Observation: conducted and coordinated by a central authority, such as a national statistical office.

Example: National censuses organized by the State Statistics Service of Ukraine.

Decentralized Observation: conducted by multiple organizations or entities, often coordinated by a central framework.

Example: Industry-specific surveys managed by separate agencies.

Principles of Effective Statistical Observation

1. **Clarity of Purpose:** Clearly define the objectives and scope of observation.
2. **Relevance:** Ensure that the data collected aligns with the intended goals.
3. **Accuracy:** Minimize errors through proper design, training, and tools.
4. **Timeliness:** Collect and disseminate data in a timely manner.
5. **Cost-Effectiveness:** Balance data quality with resource constraints.
6. **Confidentiality:** Protect the privacy of respondents and ensure ethical practices.

Statistical observation is an essential component of the statistical process, providing the raw data necessary for analysis and decision-making. By employing appropriate organizational forms – whether based on coverage, timing, or method – researchers and statisticians can design effective data collection strategies that meet the needs of various fields and applications.

4. DATA. TABULAR AND GRAPHICAL PRESENTATIONS

4.1. Grouping Data and Frequency Tables

4.2. Graphical Representation of Data

4.1. Grouping Data and Frequency Tables

Frequency tables

In statistical surveys, the size of the data is often too big which makes listing the items impractical. A more adequate way is to summarize the data using tables called *frequency tables*.

First, we will define some terms that will be used throughout the book.

Definitions

Let x_1, x_2, …, x_k be the items under study, arranged in increasing order, and f_1, f_2, …, f_k be the respective frequencies of occurrence of these items.

- $N = \sum_{i=1}^{k} f_i$ is called the *size* of the data.
- $F_j = \sum_{i=1}^{j} f_i$ is called the *cumulative frequency* of x_j.
- $\frac{f_i}{N}$ is called the *relative frequency* of x_i which sometimes is given in percentage form as $\frac{f_i}{N} \cdot 100\%$.

A *frequency distribution* is a tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes.

Example 1

Data from a sample of 50 new car purchases given below.

VW	BMW	Mercedes	Audi	VW
VW	Mercedes	Audi	VW	Audi
VW	VW	VW	Audi	Mercedes
VW	VW	Opel	Opel	BMW

VW	Audi	Mercedes	Audi	Mercedes
VW	Mercedes	Mercedes	VW	Mercedes
VW	VW	Mercedes	Opel	Mercedes
Mercedes	BMW	VW	VW	VW
BMW	Opel	Audi	Opel	Mercedes
VW	Mercedes	BMW	VW	Audi

Arrange these data in a frequency table showing their frequencies, relative frequencies, and their cumulative frequencies.

Solution

Brand	Frequency	Relative frequency	Cumulative frequency
Audi	8	0.16	8
BMW	5	0.1	13
Mercedes	13	0.26	26
Opel	5	0.1	31
VW	19	0.38	50
Total	50	1	

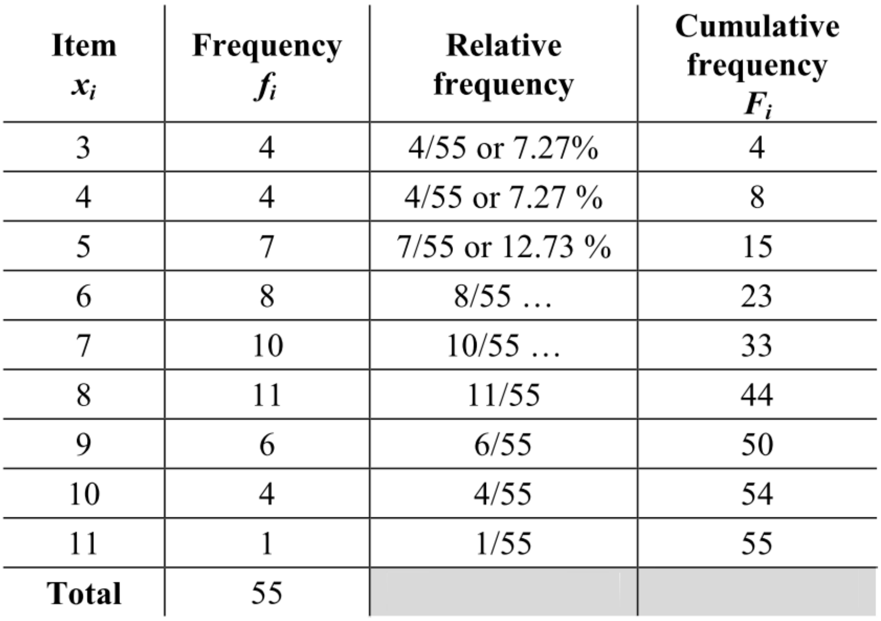
Example 2

A small firm wishes to carry out a study on the distance, to the nearest mile, traveled by each employee from his/her residence to work. The following data were collected.

6	4	7	5	3	10	7	8	8	9	4
7	6	8	8	7	6	3	11	9	7	6
5	10	7	9	4	7	9	6	8	5	7
3	8	8	7	6	9	10	6	3	5	5
8	8	8	5	7	6	8	5	4	9	10

Arrange these data in a frequency table showing their frequencies, relative frequencies, and their cumulative frequencies.

Solution



Item x_i	Frequency f_i	Relative frequency	Cumulative frequency F_i
3	4	4/55 or 7.27%	4
4	4	4/55 or 7.27 %	8
5	7	7/55 or 12.73 %	15
6	8	8/55 …	23
7	10	10/55 …	33
8	11	11/55	44
9	6	6/55	50
10	4	4/55	54
11	1	1/55	55
Total	55		

Activity 2

Draw a table showing the frequencies, relative frequencies, and cumulative frequencies for the following data items.

4 5 3 2 2 4 3 5 3

3 4 3 2 4 5 5 3 5

4 2 1 4 3 6 1 8 4

Example 3

The chief engineer in a photo-frame production industry is suspecting deficiency in one of the machines. The frames produced by this machine are supposed to be of length L =14.0 cm and width W = 8.0 cm. An inspection was carried by measuring 30 frames that were randomly selected. The measurements were as follows:

14.41, 7.88) (13.97, 8.44) (14.33, 8.08) (14.01, 8.42) (14.29, 8.12)

(13.99, 8.02) (13.89, 8.02) (14.07, 8.18) (13.95, 8.14) (14.33, 7.94)

(13.81, 8.2) (14.03, 7.8) (14.19, 8.16) (14.41, 8.34) (14.17, 8.26)

(14.29, 7.92) (14.33, 8) (13.81, 8.04) (14.31, 8.44) (13.87, 8.46)

(14.41, 8.4) (14.01, 7.88) (13.83, 8.22) (13.83, 8.1) (13.89, 7.96)

(14.23, 8.46) (14.09, 8) (13.97, 7.82) (14.41, 7.88) (14.13, 8.2)

The first number in each ordered pair represents the length (L) of the measured frame and the second number represents its width (W).

a) Group the lengths of the frames in a table using intervals (classes) of length 1 millimeter. Include in your table the frequencies and the cumulative frequencies.

b) frame is approved if its length is within 2 mm from 14.0 cm, otherwise it is viewed as defective and rejected. According to this sample what percentage of the frames are rejected?

Solution

a) The minimum length measured is 13.81 cm and the maximum length is 14.41 cm. Thus, it is sufficient to use the intervals [13.8, 13.9), [13.9, 14.0), …

Class	Frequency f_i	Cumulative frequency F_i
[13.8, 13.9)	7	7
[13.9, 14.0)	4	11
[14.0, 14.1)	5	16
[14.1, 14.2)	3	19
[14.2, 14.3)	3	22
[14.3, 14.4)	4	26
[14.4, 14.5)	4	30
Total	30	

b) Nineteen items are in the interval [13.8, 14.2], hence, the percentage that would be rejected is $\frac{11}{30}\cdot 100\% = 36.67\%$.

Activity 3

a) In the previous example, draw a table summarizing the widths of the measured frames. Also use intervals of length 1 mm.

b) In addition to deficiency in length, a frame is rejected if its width is not within 1 mm from the expected value (8.0 cm). Based on the criteria set by the width, what fraction of the frames is considered defective?

4.2. Graphical Representation of Data

Dot plot

A dot plot is a graphical summary used to describe a set of numerical observations. Each observation is represented by one dot on a horizontal (or vertical) number line. This type of plot is useful when the number of items is small, but can be inadequate with large sets of data. We often use this plot for discrete data where the data items range over a small number of values.

Example 1

The ages of thirty people are listed below.

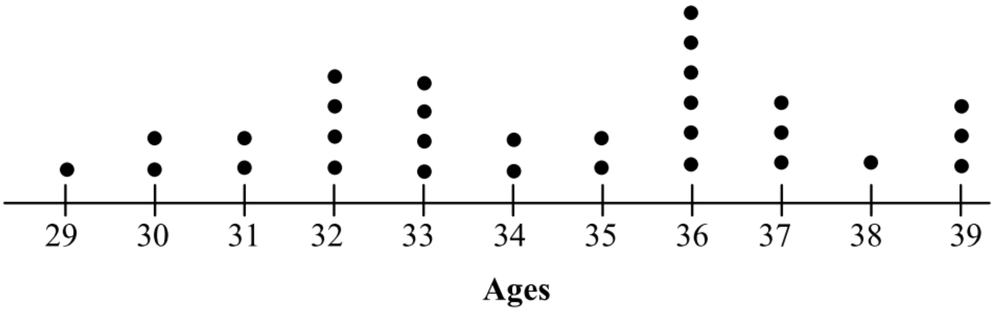
Ages

29	37	39	30	30
36	36	32	32	36
33	32	36	31	39
34	31	33	35	36
33	37	36	38	33
34	37	39	32	35

Display the data using a dot plot.

Solution

The data ranges over the numbers 29 through 39. Thus, a number line is drawn and is marked by the numbers 29, 30, 31, …, 39. The number 29 appears only once in the data. Hence, only one dot is drawn above 29 on the number line. The number 30 appears twice and hence it is represented by two dots, and so on. When all the data are accounted for, the dot plot looks like this:



Activity 1

The list below shows the grades of students on a Mathematics exam.

85	88	90	87	90	95
85	88	91	88	90	95
85	88	92	88	90	95
85	89	92	87	89	94
86	89	93	87	89	94

Display the grades using a dot plot.

Stem-and-leaf plot

When the data is dispersed, the dot plot loses its effectiveness. A better way to display the data is by using a stem-and-leaf plot. This method consists of dividing the items into two parts: the stem and the leaf. The stem (usually to the left) consists of the digit(s) in the largest place value. The leaf (usually to the right) consists of the digit(s) in the smallest place value. A key should be associated with the stem-and-leaf plot to help in interpreting the plot. The example below illustrates the idea.

Example 2

A survey concerning the duration of telephone calls was conducted. Twenty calls were chosen at random. The durations of these calls, to the nearest minute, are listed below.

8, 25, 4, 32, 29, 41, 11, 21, 44, 5, 26, 16, 34, 23, 12, 37, 22, 18, 26, 23

Display the data using a stem-and-leaf plot.

Solution

The data in ascending order:

4, 5, 8, 11, 12, 16, 18, 21, 22, 23, 23, 25, 26, 26, 29, 32, 34, 37, 41, 44.

It is displayed in a stem-and-leaf plot as follows:

Stem	Leaf
0	4 5 8
1	1 2 6 8
2	1 2 3 3 5 6 6 9
3	2 4 7
4	1 4

Key: 2| 3 means 23

A quick look at the above stem-and-leaf plot shows that the minimum duration of a telephone call is 4 minutes and the maximum is 44. It also tells us that there are more calls between 20 and 30 minutes than any other 10-minute interval.

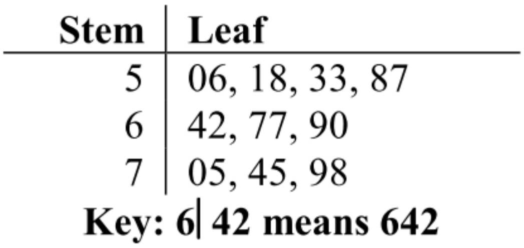
Example 3

The scores of 10 students on a standardized exam are given below.

506, 518, 533, 587, 642, 677, 690, 705, 745, 798

Display the data using a stem-leaf plot.

Solution



Stem	Leaf
5	06, 18, 33, 87
6	42, 77, 90
7	05, 45, 98

Key: 6| 42 means 642

Activity 2

Forty random ticket prices at different movie theaters were selected. The prices are given in the table below:

12	14	21	6	34	14	22	45
16	20	31	27	23	32	26	17
18	21	8	18	24	9	16	15
23	25	12	16	32	11	9	20
24	32	42	14	38	2	15	22

Construct a stem-and-leaf plot for the data. Choose a convenient key.

Bar graphs

Bar graphs, also called bar charts, are used to display the frequencies of occurrences of different categories. Each category is represented by a bar with height equal to the frequency of that category. We may draw vertical bar graphs by placing the categories on a horizontal line and drawing the bars above them. Categories can also be placed on a vertical line, with bars drawn horizontally next to them. The former representation is called a vertical bar graph, while the latter is called a horizontal bar graph.

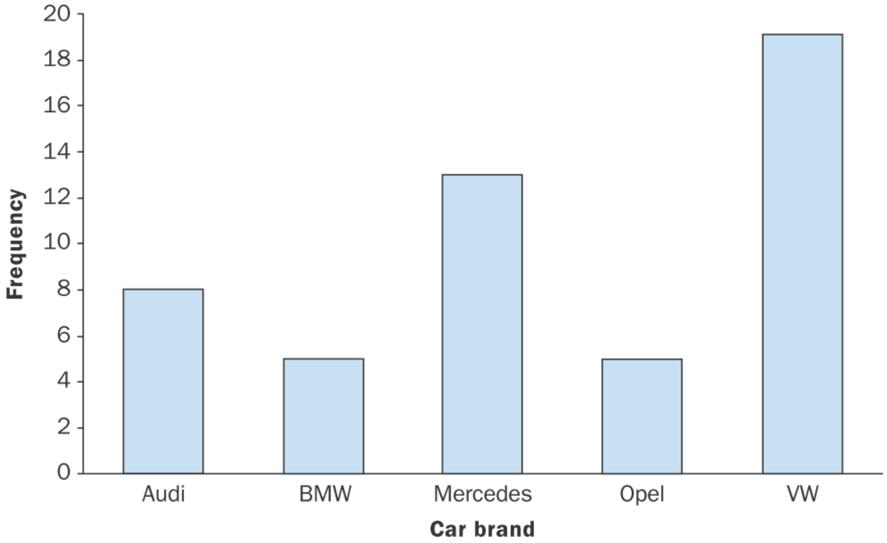
Example 4

The adjacent table shows new car purchases.

Brand	**Frequency**
Audi	8
BMW	5
Mercedes	13
Opel	5
VW	19
Total	50

Display the information using a bar graph.

Solution



Pie chart

A disk divided into sectors, where each sector represents an item in the data, can be used to display data graphically. Such a representation is referred to as a pie chart or a circle graph. The areas of these sectors are proportional to the frequencies, or quantities, they represent.

Example 5

The adjacent table shows new car purchases.

Brand	Frequency	Relative frequency
Audi	8	0.16
BMW	5	0.1
Mercedes	13	0.26
Opel	5	0.1
VW	19	0.38
Total	50	1

Display the information using a pie chart.

Solution

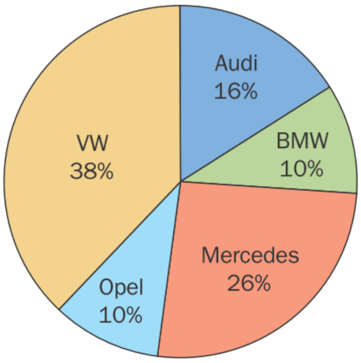
Audi is represented by a sector with 0.16(360) = 57.6 degrees.

The central angles of the sectors representing BMW is 0.1(360) = 36 degrees.

The sector of the pie chart labeled Mercedes consists of 0.26(360) = 93.6 degrees.

VW consists of 0.38(360) = 136.8 degrees.

The central angle of the sector representing Opel is 36 degrees.



Histograms

Grouped data can be displayed using a histogram where each group (class) is represented by a rectangle. In a histogram, the width of the rectangle is equal to the class width and the height is the corresponding frequency, relative frequency or percentage frequency.

One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution.

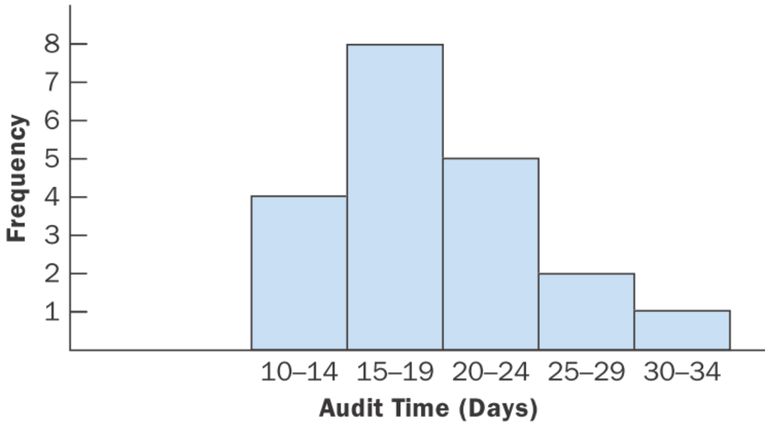
Example 5

The adjacent table shows new car purchases.

Audit time (days)	Frequency
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

Display the information using a histogram.

Solution



Cumulative Frequency Plots

Recall that the cumulative frequency of a certain class interval is the total number of observations of this class interval and all the class intervals that precede it. These sums are displayed in a column next to the column that contains the frequencies, as shown in the table below.

Class	Frequency	Cumulative frequency (*c.f.*)
$[a_0, a_1)$	f_1	$F_1 = f_1$
$[a_1, a_2)$	f_2	$F_2 = f_1 + f_2$
$[a_2, a_3)$	f_3	$F_3 = f_1 + f_2 + f_3$
$\vdots$	$\vdots$	$\vdots$
$[a_{n-1}, a_n)$	f_n	$F_n = f_1 + f_2 + f_3 + \ldots + f_n = N$

When the ordered pairs $(a_0, 0)$, (a_1, F_1),) (a_2, F_2), (a_3, F_3), …, and (a_n, N) are plotted on a coordinate plane and joined by line segments, the resulting graph is called the cumulative frequency polygonal lines. If instead, a smooth curve is fitted through the points, the resulting curve is called the cumulative frequency curve (plot) or the S-curve.

Example 6

Draw cumulative frequency polygonal lines for the heights of the 175 children given below.

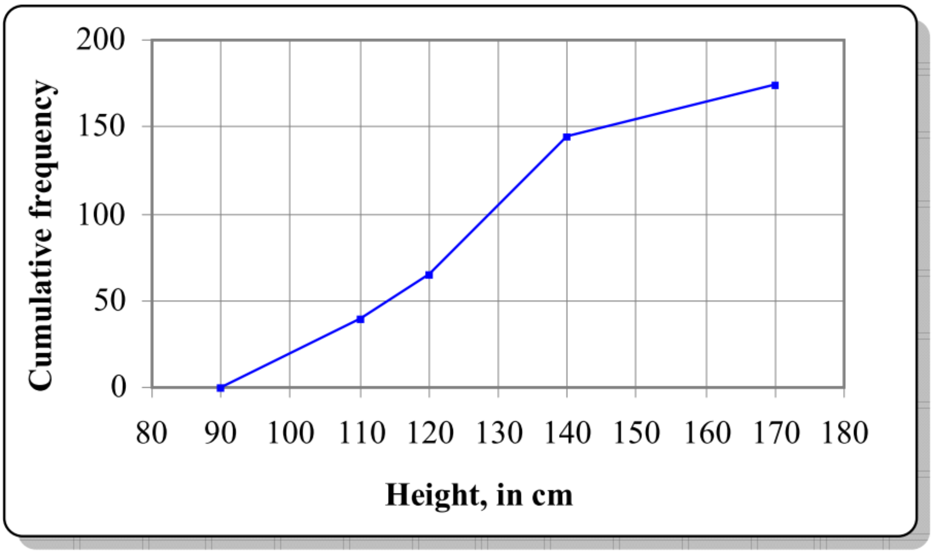
Height (cm)	[90 – 110)	[110 – 120)	[120 – 140)	[140 – 170)
Frequency	40	25	80	30

Solution

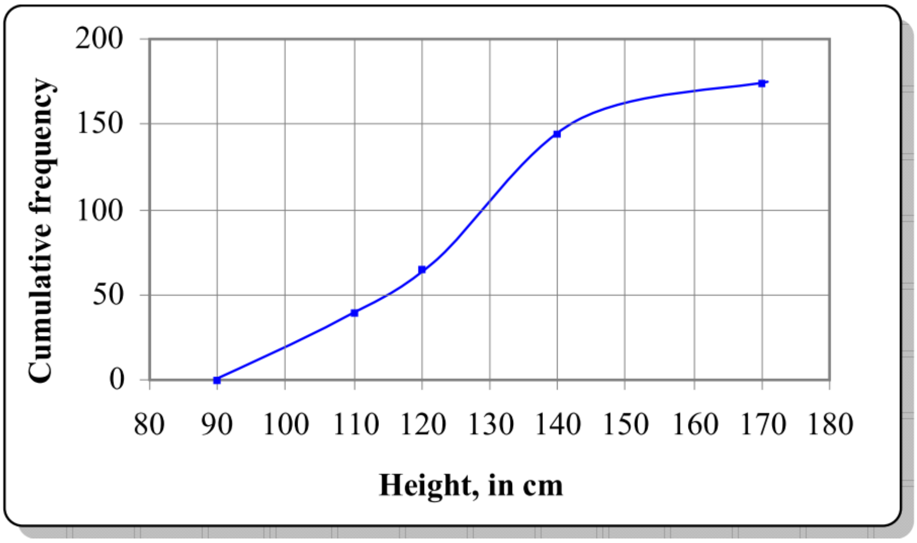
The cumulative frequencies of the different classes are displayed in the table below.

Height (cm)	Frequency	Cumulative frequency
[90 – 110)	40	40
[110 – 120)	25	65
[120 – 140)	80	145
[140 – 170)	30	175

The points (90, 0), (110, 40), (120, 65), etc…., are plotted in a coordinate plane and then joined by line segments. The resulting curve is the cumulative frequency polygonal lines.



Drawing a cumulative frequency curve is similar to drawing the polygonal lines but instead of the line segments, we join the points with a convenient smooth curve.



5. DESCRIPTIVE STATISTICS

5.1. Central tendencies: Mean, Median, and Mode

5.2. Quartiles, Percentiles and The five-number summary

5.3. Measures of Dispersion

5.4. Data Shapes

5.1. Central tendencies: Mean, Median, and Mode

There are some indicators that can be extracted from raw data that give useful information about the population in a study, and may lead to useful conclusions. In this section, three such indicators are considered. They measure the central tendencies of the data.

Mean

The mean $\overline{x}$ of ungrouped data $x_1, x_2, \ldots, x_n$, or simply the average, is obtained by adding the items and dividing by the size of the data:

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

Example 1

Find the mean of the following numbers: 23, 22, 24, 24, 23, 20, 25.

Solution

$$\overline{x} = \frac{23+22+24+24+23+20+25}{7} = 23$$

Example 2

The mean score of a student in four exams is 70%. How much should the student score on the fifth exam if he/she wishes to increase the average to 74%?

Solution

Let x be the score on the 5th exam. The total of the first 4 exams is: 4 × 70 = 280 and so $\frac{x+280}{5} = 74$ which yields $x = 90$.

Activity 1

1. Find the mean of the set of data.

a) 11, 10, 12, 10, 13

b) 0.4, 0.6, 0.36, 0.64

c) 2, 4, −6, −4

2. A box of apples contains 72 apples. If the mean weight of an apple is k ounces, what is the weight of the apples in the box?

3. There are 12 girls and 8 boys in a class. The mean score obtained by the girls on a certain exam was 70% while that obtained by the boys was 76%. What is the mean score of the class on this exam?

In case of grouped data, where x_1 occurs f_1 times, x_2 occurs f_2 times, etc…, the mean is given by

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{N}.$$

Example 3

The daily allowances of 27 children are shown below.

Pocket Money ($)	Frequency
5	6
6	8
8	7
10	6

Find the mean daily allowance given to these children.

Solution

The mean is $\overline{x} = \frac{5\cdot 6 + 6\cdot 8 + 8\cdot 7 + 10\cdot 6}{6+8+7+6} = \frac{194}{27} \approx \$7.20$.

When data is grouped in classes, each class $[a, b)$ is represented by its midpoint $\frac{a+b}{2}$, which is also called the mid-class. The mean obtained by using the mid-class is an estimate of the true mean.

Example 4

Estimate the mean length of 40 rods with lengths given in the table below.

Length (cm)	Frequency (f_i)	Mid value (x_i)
[46, 50)	4	48
[50, 54)	6	52
[54, 58)	7	56
[58, 62)	10	60
[62, 66)	13	64

Solution

$$\bar{x} \approx \frac{4\cdot 48+6\cdot 52+7\cdot 56+10\cdot 60+13\cdot 64}{40} = 58.2 \text{ cm}$$

Activity 2

The frequency distribution for the fuel consumption of 100 cars in miles per gallon is given in the table below.

Miles per gallon	Frequency (f_i)
[14, 16)	9
[16, 18)	15
[18, 20)	30
[20, 22)	36
[22, 24)	10

Estimate the mean consumption of fuel by the cars.

Weighted mean

Suppose the final grade of a student is computed by counting each of three exams 20%, the homework grade 10%, and the final 30% of the overall grade. To compute the final average, it would not be correct to add all 5 grades and divide by 5. Instead, each exam score must weigh 20, while the homework score weighs 10 and the final weighs 30. The mean in this case is referred to as the weighted mean.

The weighted mean of a set of values x_1, x_2, … with respective weights m_1, m_2, … is given by $\overline{x}_w = \dfrac{x_1m_1 + x_2m_2 + \ldots}{m_1 + m_2 + \ldots}$.

Example 5

An oil company has three divisions: Headquarters, South, and North. The average hourly wage of a worker in headquarters is \$15.1, in the southern division it is \$9.4, and in the northern division it is \$11.5. What is the average hourly wage of the employees of the company if 80 people work in headquarters, 860 work in the southern division, and 520 in the northern division?

Solution

80 workers make \$15.1 each, 860 workers make \$9.4 each, and 520 workers make \$11.5 each. Therefore, the average hourly wage of all employees is $\overline{x}_w = \dfrac{80\cdot 15.1 + 860\cdot 9.4 + 520\cdot 11.5}{80+860+520} = \$10.46$.

Activity 3

Three sets of data consisting of 3, 4, and 6 items have averages 20, 25, and 30, respectively. What is the average of the data consisting of all three sets?

Median

Suppose x_1, x_2, …, x_n is a set of numeric data arranged in increasing or decreasing order. The median *Me* of this set is defined as:

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}.$$

In other words, the median is the middle term if n is odd and the average of the two middle terms if n is even.

Example 6

Find the median of each of the following sets of data.

a) 26, 36, 34, 25, 32, 40, 41, 27, 28, 32, 35, 38, 41, 42, 45

b) 2, 4, 5, 4, 8, 7, 8, 7, 3, 1

c) The number of residents per apartment, in a small neighborhood, is given in the table below.

Number of residents x_i	1	2	3	4	5	6
Number of apartments f_i	20	46	51	38	17	4

Solution

a) The data in ascending order is:

25, 26, 27, 28, 32, 32, 34, **35**, 36, 38, 40, 41, 41, 42, 45.

There are 15 elements in the set with 35 in the middle. Hence, the median is 35.

b) The data in ascending order is: 1, 2, 3, 4, **4**, **5**, 7, 7, 8, 8.

There are 10 elements in the set with 4 and 5 as the two middle values. Hence, the median is: $\frac{4+5}{2}=4.5$.

c) The size of the data is n = 176 (even), hence the median is the average of the 88th and 89th terms. Both of these terms are 3, hence the median is $\frac{3+3}{2}=3$.

Activity 4

Find the median for each set of data.

a) 3, 6, 8, 10, 21

b) 6, 8, 12, 16, 21, 32

c)

Item x_i	Frequency (f_i)
3	9
5	15
7	30

8	36
12	10

Median of continuous data

When continuous data is given in classes, each is an interval of real numbers, an estimate of the median can be obtained. To find such an estimate, we consider the class with the highest relative cumulative frequency that is under 50%. The next class would then have a relative cumulative frequency of 50% or more. This class is called the median class. The median belongs to the interval that defines this class. The next example illustrates how such estimate of the median is obtained.

Example 7

The table below shows the annual sales of television sets from a sample of 100 stores.

Units sold	**Number of stores (f_i)**
[200, 300)	12
[300, 400)	31
[400, 500)	35
[500, 600)	14
[600, 700)	8

Find the median number of television sets sold by the 100 stores.

Solution

The survey is done on N = 100 stores, which is the size of the data. Hence, we need to find the number of television sets sold, for which 50 stores (N/2) sold less than this number and 50 stores sold more. For this purpose, we follow the following steps:

43 stores sold less than 400 TV sets and 78 sold less than 500; hence the median is between 400 and 500. [400, 500) is called the **median class**.

50 – 43 = 7 is the number of items needed to be counted in the median class. One way to do this is to assume that the number of TV sets sold by the 35 stores in this class is evenly distributed between 400 and 500. Under this assumption,

the 44th store sold $400+\frac{1}{35}(500-400)$,

the 45th store sold $400+\frac{2}{35}(500-400)$, …

the 50th store sold $400+\frac{7}{35}(500-400)=420$.

Therefore, the median number of televisions sold annually by the stores is 420 units.

Note that the median obtained in the previous example is only an estimate of the true median. The above procedure can be formalized as follows.

Class	Frequency	Cumulative Frequency
$[a_0, a_1)$	f_1	F_1
$[a_1, a_2)$	f_2	F_2
…	…	…
$[a_{i-1}, a_i)$	f_i	F_i
…	…	…
$[a_{n-1}, a_n)$	f_n	F_n

Identify the median class, say $[a_{i-1}, a_i)$, which satisfies $F_{i-1} < \frac{N}{2}$ and $F_{i-1} + f_i \geq \frac{N}{2}$. Here, F_{i-1} is the cumulative frequency of the class just above the median class, N is the size of the data, and f_i is the frequency of the median class $[a_{i-1}, a_i)$.

Calculate the median by applying the rule: $Me = a_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{f_i}\left(a_i - a_{i-1}\right)$.

Simply put, to find an estimate of the median, move from the left endpoint of the median class a fraction of the length of the median class needed to reach 50% of the items.

Activity 5

Find the median of the following set of data.

Class interval **Number of stores (f_i)**

[25, 35)	16
[35, 44)	22
[44, 55)	18
[55, 65)	4
Total	60

Mode

In a set of discrete data, the item with the highest frequency is called the mode and denoted *Mo*. If no item in the data occurs more than once, then the data has no mode. If more than one item have the same frequency that is greater than all other frequencies, then all such items are the modes.

Example 8

Find the mode of each of the following sets of data:

a) 3, 1, 8, 4, 6, 1, 5, 1, 3, 8, 5, 7

b) 2, 6, 4, 5, 8, 7

c) 1, 4, 5, 8, 3, 4, 3, 2, 5

d) 0, 0, 0, 0, 1, 1, 1, 1

Solution

a) 1 occurs three times. Its frequency of occurrence is higher than that of any other item, therefore *Mo* = 1.

b) No item occurs more than once. The data has no mode.

c) 3, 4, and 5 occur the same number of times, and more often than the rest of the data. Hence, 3, 4, and 5 are the modes.

d) We have two modes, 0 and 1.

Activity 6

Give the frequency of occurrence of each item in the given set of data and deduce the mode(s) of the set, if any.

a) 3, 4, 1, 3, 1, 2, 0, 0, 3, 0, 5, 1, 4, 0, 0, 3, 0, 0, 3, 7, 4, 0

b) 2.1, 1.2, 2.1, 3.0, 0.5, 1.4, 2.1, 0.6, 2.1, 1.4, 0.6, 1.4, 0.5, 1.4

c) 1, −1, 2, −2, 3, −3, 4, −4, 5, −5, 6, −6

5.2. Quartiles, Percentiles and The five-number summary

Definitions

The k^{th} percentile, P_k, is a value that splits the data into two parts. Part 1 consisting of N_1 numbers that are less than P_k and part 2 consisting of N_2 numbers that are greater than P_k. The ratio $N_1 : N_2$ is

$$\frac{k}{100-k}.$$

The 25th percentile is called the first or lower quartile and denoted by Q_1.

The 50th percentile is called the second or middle quartile Q_2. It is also the median of the data.

The third or upper quartile Q_3 is the 75th percentile.

The kth percentiles, the lower quartile, and the upper quartile of a data set of size N are sometimes referred to, respectively, as, $\frac{k}{100}(N+1)^{th}$, $\frac{1}{4}(N+1)^{th}$ and $\frac{3}{4}(N+1)^{th}$ terms of the data.

Example 1

Find the lower and upper quartiles of the following data set.

8, 9, 12, 13, 16, 17, 18, 20, 22, 30, 31, 40

Solution

The size of the data is $N = 12$, so Q_1 is $\frac{1}{4}(12+1)^{th}$ value which is the 3.25th value. The 3rd value is 12 (data is already arranged in increasing order). We add 0.25 of the distance between the 3rd and 4th values to the 3rd value to obtain Q_1.

Therefore, $Q_1 = 12 + 0.25(13 – 12) = 12.25$.

Similarly, Q_3 is $\frac{3}{4}(12+1)^{th}$ value which is the 9.75th value. The 9th value is 22 and so $Q_3 = 22 + 0.75(30 - 22) = 28$.

Example 2

The results on a test for a group of students are recorded in the table below.

Class	Cumulative relative frequency
[50, 60)	12%
[60, 70)	40%
[70, 80)	75%
[80, 90)	91%
[90, 100)	100%

Find Q_1, Q_3 and P_{90}.

Solution

$Q_1 = \frac{1}{4}(100+1)^{th} = 25.25^{\text{th}}$ value.

12% of the scorers are below 60 and 40% are below 70.

This means 28% (40 – 12 = 28) of the students scored in the interval [60, 70).

Thus, Q_1 belongs to [60, 70). To get to Q_1 we need to add to 60, $\frac{13.25}{28}$ of the distance between 60 and 70. Therefore, $Q_1 = 60 + \frac{13.25}{28}(70-60) = 64.73$.

The cumulative relative frequency of the interval [70, 80) is 75%. This shows that 75% of the students scored less than 80, and 25% scored more. Hence, $Q_3 = 80$.

P_{90} is the $0.9(100 + 1)^{\text{th}}$ value, which is the 90.9^{th} value from the interval [80, 90). The relative frequency of this interval is 91 – 75 = 16%, and the cumulative relative frequency of the previous interval is 75%. To get to P_{90} we need to add to 80, $\frac{15.9}{16}$ of the distance between 80 and 90. Therefore, $P_{90} = 80 + \frac{15.9}{16}(90-80) = 89.94$.

Activity 1

1. Find the lower and upper quartiles of each of the following data sets:

a) 6, 7, 10, 11, 13, 13, 16, 18, 20, 20

b) 0, 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 3

2. The time taken for each of 120 students to reach school is recorded and the data collected is summarized in the table below.

Time in minutes	**Number of children**
[0, 5)	8
[5, 10)	13
[10, 15)	29
[15, 20)	41
[20, 25)	19
[25, 30)	10

Find:

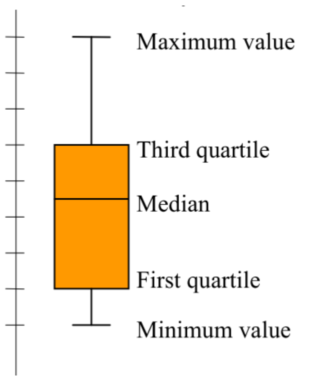
a) the 40th percentile, and

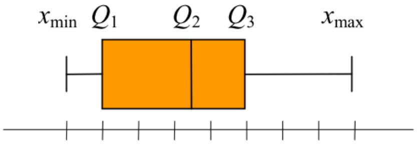
b) the lower and upper quartiles.

The *five-number summary* (also called the five-point summary) of a numeric data $\{x_{\text{min}}, Q_1, Q_2, Q_3, x_{\text{max}}\}$ consists of the minimum value, the lower quartile (Q_1), the median (Q_2), the upper quartile (Q_3), and the maximum value, ordered from smallest to largest.

Boxplot

Boxplot (also called box-and-whisker plot) is a graphical representation of the 5-number summary of a distribution. A vertical or a horizontal rectangle is drawn extending, along a number line, from the lower quartile to the upper quartile. A whisker is drawn from the lower (left) end of the rectangle to the minimum value of the data. Another one is drawn from the upper (right) end of the rectangle to the maximum value of the data. A line is drawn at the median level that cuts the rectangle into two parts. The two figures below show the two ways of drawing boxplots.





Example 3

The list below shows the grades of 40 students on a mathematics exam.

62 75 69 79 82 92 86 94

62 77 70 79 82 92 87 94

64 77 72 81 85 93 88 96

64 78 73 82 85 94 88 100

64 79 73 82 86 94 88 100

Display the data using a box-and-whisker plot.

Solution

The minimum grade is 62 and the maximum grade is 100.

The lower quartile is the (40 + 1)/4 = 10.25th value.

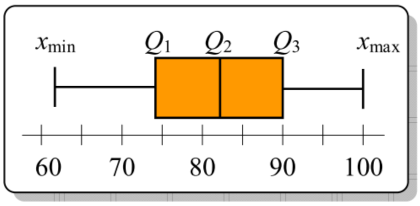
Thus, $Q_1 = 73 + 0.25(75 - 73) = 73.5$.

The median is (82 + 82)/2 = 82.

The upper quartile is the 3(40 + 1)/4 = 30.75th value.

Thus, $Q_3 = 88 + 0.75(92 - 88) = 91$.

A box-and-whisker plot of this data would then be:



Activity 2

Find the 5-number summary and display the data using a box-and-whisker diagram.

20 24 27 31 33

20 24 28 31 34

22 24 28 32 34

23 26 30 32 34

23 26 31 32 36

5.3. Measures of Dispersion

Variations in values in a set of data is called *dispersion*.

Standard deviation, range, and inter-quartile range are examples of measures of dispersion. The most commonly used measure is the standard deviation. These measures indicate how the data is spread out around their mean.

The need to measure the spread or dispersion of a set of data is illustrated in the next example.

Example 1

To estimate the monthly income in the towns of Slovia and Ventori, a sample of 8 adults was taken from each town. The salaries, in thousands of dollars, are:

Slovia				Ventori			
1,200	1,250	1,400	1,800	3,200	3,800	450	1,750
1,600	1,800	1,600	1,400	750	750	1,000	500

a) Find the mean of each sample.

b) To compare the incomes of the residents of the two towns, is it enough to say that their salaries are about the same or more explanation is needed?

Solution

a) By the formula for mean, the means for the two towns are 1.506 thousand dollars and 1.525 thousand dollars.

b) Although the averages for the salaries are about the same, the diversity of the salaries in Ventori is obvious and hence another measure is needed to provide a clearer picture about the income in the two towns.

Definitions

➢ To measure the size of dispersion, we define the following indicators.

The range is entirely based on the two extremities of the data and is given by:

Range = $x_{max} - x_{min}$

➢ The inter-quartile range, IQR, is given by: IQR = $Q_3 - Q_1$.

It gives the range of the middle 50% of the data. The significance of this indicator will become apparent in the next chapter, when displaying data using charts and other visual techniques.

➢ The mean absolute deviation, is given by $MAD = \dfrac{\sum_{i=1}^{n} |x_i - \overline{x}| \cdot f_i}{\sum_{i=1}^{n} f_i}$

Example 2

Find the range and the mean absolute deviation of the data.

Temperature variation from 7:00 A.M. to 8:00 A.M. in °C	0.5	1.0	1.5	2.5	3.0
Number of days (f_i)	4	8	12	10	2

Solution

Range = $x_{max} - x_{min}$ = 3.0 – 0.5 = 2.5°C

$\overline{x} = 1.6$°C

$$MAD = \frac{4|0.5-1.6| + 8|1.0-1.6| + 12|1.5-1.6| + 10|2.5-1.6| + 2|3.0-1.6|}{4+8+12+10+2} = 0.62°\text{C}$$

Variance and standard deviation

➢ The variance of a set of data is defined as:

$$Var = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i}{\sum_{i=1}^{n} f_i}.$$

➢ The standard deviation of a set of data is given by: $\sigma = \sqrt{Var}$.

The alternative rule below is commonly used for computing the variance:

$Var = \overline{x^2} - (\overline{x})^2$

Example 3

Find the standard deviation of the set of data in the table below.

Score (x)	60	65	70	75	80	85	90
Frequency	2	3	4	4	6	2	1

Solution

$\overline{x} = 74.32$

$$Var = \frac{2 \cdot 60^2 + 3 \cdot 65^2 + 4 \cdot 70^2 + \ldots}{22} - 74.32^2 = 64.04$$

$\sigma = \sqrt{Var} = \sqrt{64.04} = 8.002$

Activity 2

1. Find the mean and the standard deviation for each set of data.

a) 3, 4, 8, 8, 9

b) 6.2, 6.7, 6, 7, 6.4, 6, 6.6, 5.9, 6.8

c) Scores of 20 students on a math exam:

Score	1	2	3	4	5
Frequency	1	4	7	6	2

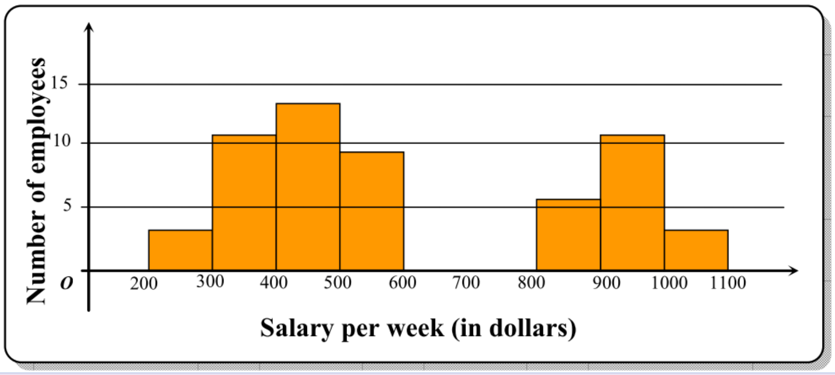
5.4. Data Shapes

Clusters and gaps

Sometimes data is divided into blocks, separated by intervals where no value occurs. These blocks, or groups, are called clusters. The empty regions between these groups are called gaps. Clusters and gaps are related features since any two adjacent clusters are separated by a gap.

Example 1

Identify the clusters and the gaps in the histogram below.



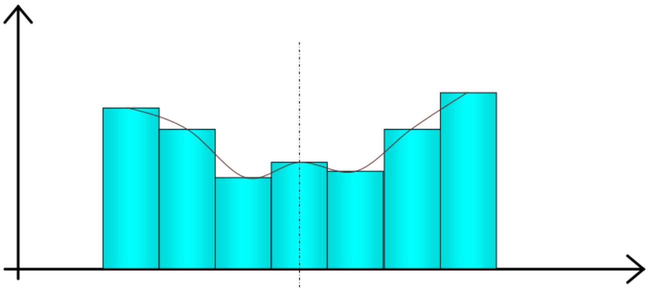
Solution

There are two distinct clusters: one in the interval ($200, $600) and the other one in the interval ($800, $1100) and one gap in the interval ($600, $800).

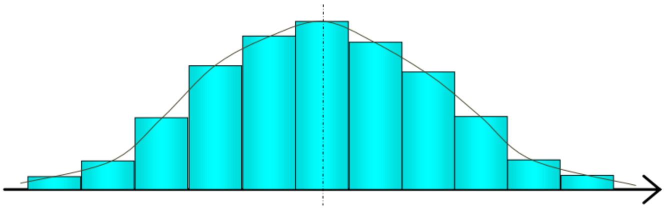
Symmetric, skewed, and uniform

Data distributions have different shapes. The shape of the distribution is a very important aspect of the overall pattern. It helps to properly understand the data. The shapes described below are the most distinguished and commonly found in statistical studies.

A distribution, in which observations equidistant from the median have the same frequency is called a *symmetric distribution*.

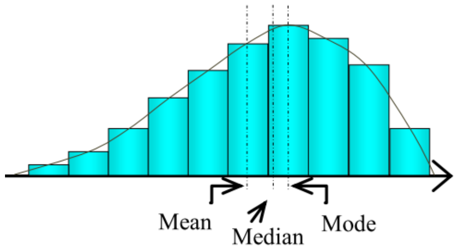


A *bell-shaped* (mound-shaped) distribution is a symmetric distribution with one top and two sloping tails.

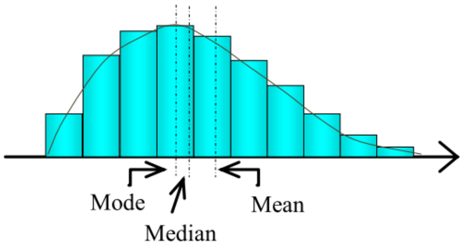


A distribution is skewed to the left (or negatively skewed) if a long tail to the left occurs with one or few low values (or if the left side of the histogram extends much farther out than the right side). On the other hand, if it extends farther to the right then it is called skewed to the right (or positively skewed).

Left skewed distribution



Right skewed distribution



The median is at the center of the data. It is the value on the horizontal axis that splits the area of the histogram into two equal parts.

The mode is the item with the highest frequency.

The mean is the 'balance point' of the data. If we try to balance the data at its median, it will topple to the side of the tail since the area spreads farther in that direction. Using convenient mathematical methods, it can be shown that the vertical line through the mean passes through the center of gravity of the area.

Measures of skewness

Often, the measures of central tendency and dispersion do not give a complete description of the distribution. It is possible to have frequency distributions that are quite different and yet have the same central tendency and dispersion. In a symmetrical distribution, mean, median, and mode are equal to each other.

As discussed earlier, the skewness of a distribution is defined as the lack of symmetry. In fact, mean > median > mode indicates the presence of extreme values on the right hand side and we say that the distribution is positively skewed. On the other hand, mean < median < mode indicates the presence of extreme values on the left hand side and we say that the distribution is negatively skewed.

One way to measure the direction and extent of skewness is Pearson's coefficient of skewness S_A, where $S_A = \dfrac{\overline{x} - Mo}{\sigma}$.

Properties of Pearson's coefficient of skewness:

- $-1 \le S_A \le 1$

- $S_A < 0$, the distribution is skewed to the left.
- $S_A > 0$, the distribution is skewed to the right.
- $S_A = 0$, the distribution is symmetrical about the mean.

Example 2

Calculate Pearson's coefficient of skewness of the data below. Comment on the result.

Value	3	5	7	9	11	13	15	17	19
Frequency	1	4	7	2	5	8	3	5	7

Solution

To one decimal place, $\bar{x} = 12.2$, $\sigma = 4.8$ and $Mo = 13$.

$$S_A = \frac{12.2 - 13}{4.8} \approx -0.167$$

$S_A < 0$, the distribution is skewed to the left.

Activity 3

Calculate Pearson's coefficient of skewness of a set of data with $\bar{x} = 72$, $Mo = 48$, and $\sigma = 32$.

Another way to measure the direction and extent of skewness is Bowley's (quartiles) coefficient of skewness, S_B, where

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

Properties of the quartile coefficient of skewness

- $-1 \le S_B \le 1$
- $S_B < 0$, the distribution is skewed to the left.
- $S_B > 0$, the distribution is skewed to the right.
- $S_B = 0$, the distribution is symmetrical about the mean.

Example 3

Calculate the quartile coefficient of skewness of the data below. Comment on the result.

Value	3	5	7	9	11	13	15	17	19
Frequency	1	4	7	2	5	8	3	5	7

Solution

There are 42 items in this set. The 21st and 22nd items are both 13 and hence, *Me* = 13. Moreover, $Q_1 = 7$ and $Q_3 = 17$. Thus, we obtain,

$$S_B = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{17 - 2\cdot 13 + 7}{17 - 7} = -\frac{2}{10} = -0.2\,.$$

$S_B < 0$, the distribution is skewed to the left.

Activity 4

Consider the following stem-and-leaf diagram.

3	0 0 1 2
4	1 3 4 5 5 5 5 5 5
5	1 1 2 2 6 6 6 6 7 8
6	0 0 0 0 3 3 5 5 5
7	2 2 6 7 8 9
8	0 1 1 1 1 1

Key: 6|3 means 63

Calculate the quartile coefficient of skewness of the given distribution and comment on the result.

The quartiles measure of skewness is based on the middle 50% of the data. A slightly different measure is based on the 10th and 90th percentiles and thus uses 80% of the data.

$$S_C = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Example 4

Calculate S_C for the data in Example 3 and comment on the result.

Solution

$P_{50} = Me = 13$, $P_{10} = 4$ and $P_{90} = 19$. Thus, we obtain,

$$S_C = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} = \frac{19 - 2\cdot 13 + 4}{19 - 4} = -\frac{3}{15} = -0.2$$

$S_C < 0$, the distribution is skewed to the left.

Activity 5

Calculate S_C for the data in Activity 4 and comment on the result.

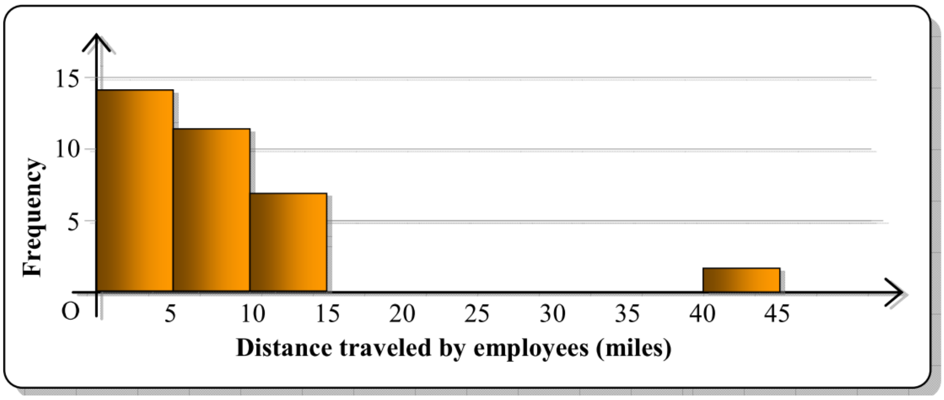
Outliers

Sometimes, unusually high or unusually low values occur in a set of observations. Often, but not always, these unusual results are due to error made when data was recorded. Outliers, as these values called, are observations that are separate from the overall pattern, and are incompatible with the rest of the data. It is important to detect whether there are outliers in the data set, as they may have an important influence on the final results.

As a guide, an item is an outlier if it is either more than two standard deviations from the mean, or more than 1.5 × IQR from the closest quartile.

Example 5

The histogram below shows the distribution of the distances traveled by employees at a certain factory to and from work.



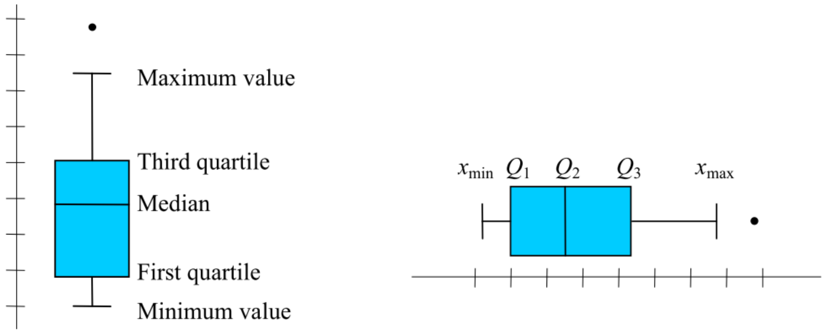
What are the outliers in the data?

Solution

It is clear that the class [40, 45) is the outlier in this case since it only has few observations, and very far from where the data are clustered (between 0 and 15 miles).

Modified boxplot

Note that the outliers often misrepresent the actual distribution of the data. Thus, in the presence of outliers, it is preferable to use the modified boxplot. In this case, the whiskers reach out to the last observation that is not an outlier and the outliers are represented as separate points. The two forms below illustrate the idea.



Example 6

The distances traveled to work by a group of 85 employees are shown in the table

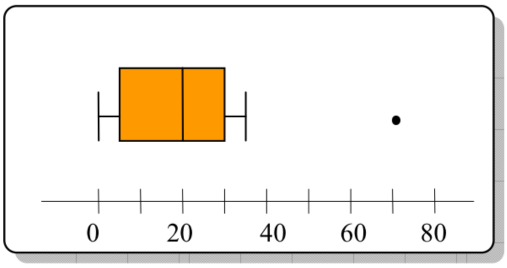
below. The numbers are rounded to the nearest 5 km.

Distance	0	5	10	15	20	25	30	35	70
Frequency	7	23	5	4	8	3	28	6	1

Organize the data in a modified boxplot. Use 1.5 times the inter-quartile range to determine whether the data has any outliers.

Solution

The minimum distance is 0 km and the maximum distance is 70 km. The median is the 43rd term which is 20, the lower quartile is the 86/4 = 21.5th term which is 5 km, and the upper quartile is the 0.75×86 = 64.5th term which is 30 km.



IQR = 30 − 5 = 25

To determine the outliers: 1.5×25 = 37.5. There are no items that are more than 37.5 to the left of 5 but there is 1 item that is more than 37.5 to the right of 30. Therefore, 70 is the only outlier.

Activity 6

Twenty students took an English test. The scores are listed below in ascending order.

35 65 67 69 71 73 73 74 75 77

77 80 82 84 84 90 92 94 96 100

Organize the data in a modified boxplot. Use 1.5 times the inter-quartile range to determine whether the data has any outliers.

6. MEASURES OF RELATIONSHIPS BETWEEN VARIABLES

6.1. Scatter Diagrams

So far, our interest in the previous topics was centered on data with a single variable. Such data are called *univariate*. Investigations such as

- the number of weekend television commercials shown and the sales at the store during the following week,
- the temperature of a certain chemical reaction and the yield of this reaction,
- the number of workers in a certain industry and the number of items produced per month,

that involve more than one variable are called *multivariate*.

Data connecting two variables are known as *two-dimensional* or *bivariate data*. When each item relating to one variable is paired with an item relating to the second variable, and the ordered pairs obtained are plotted in a rectangular system, the diagram obtained is called a *scatterplot* (or a *scatter diagram*).

Studying such plots helps in finding a relationship or a mathematical model that relates the two variables.

Furthermore, two-dimensional scatterplots help identify gaps, outliers, and clusters as well as *positive* or *negative dependencies*.

A downward-sloping scatter indicates negative dependency and an upward-sloping scatter indicates positive dependency.

A line that best fits the data points is called a trend line.

The next example illustrates the different concepts mentioned above.

Example

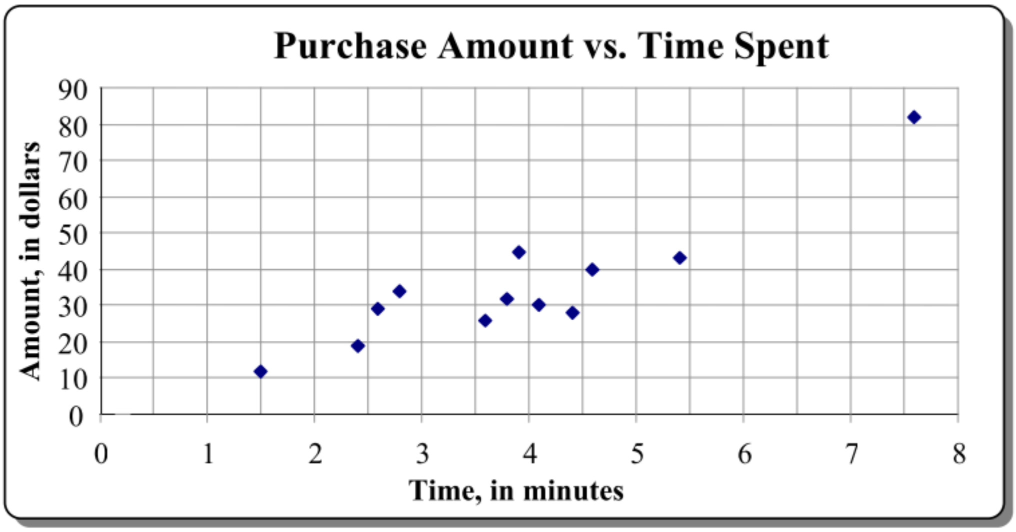
A survey was made about the purchase amount and the checkout time at a supermarket. The results are shown below.

Time (min)	Purchase amount ($)
2.4	19
1.5	12
7.6	82
3.8	32
2.8	34
4.6	40

Time (min)	Purchase amount ($)
4.1	30
5.4	43
3.9	45
2.6	29
4.4	28
3.6	26

Draw a scatterplot for the data using the horizontal axis for the time. Comment on the diagram obtained.

Solution



The trend shows that as the purchase amount increases, the time spent at the checkout counter also increases. Thus a positive dependency exists between the two variables. The diagram also suggests the existence of a linear relation between the variables and that most consumers spend between 2 and 5 minutes and their bill ranges from $20 to $45.

Activity

1. The fuel consumption y, for a car trip of length 100 km, is expressed as a function of its speed x as shown in the following table:

***x* (in km /h)**	60	80	100	150	200
***y* (in liters)**	4.5	6.5	7.4	12.8	15.6

Draw a scatterplot for the data and comment on the result.

2. The number of hours studied for a certain exam and the score obtained on that exam for a group of students are displayed below.

# of hours	Exam Score
0	56
1	63
2	68
3	78
4	82
5	93
6	94

Make a scatterplot for this data. Tell whether a trend in the data exists. If so, specify whether it is a negative trend or a positive trend.

6.2. Regression Lines

Consider a sample of bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of two related variables X and Y. Let $\bar{x}$ denote the mean of $x_1, x_2, \ldots, x_n$ and $\bar{y}$ denote the mean of $y_1, y_2, \ldots, y_n$.

➢ The point $(\bar{x}, \bar{y})$ is called the *center of gravity* of the data.

➢ Covariance (cov) is a measure of the linear relationship between two variables. A positive value indicates a direct or increasing linear relationship, and a negative value indicates a decreasing linear relationship.

➢ The covariance of the data is defined by: $\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

➢ An alternative formula for the covariance is: $\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y} = \overline{xy} - \bar{x}\cdot\bar{y}$.

A covariance is a descriptive measure of the linear association between two variables.

If the value of $\operatorname{cov}(X,Y)$ is positive, the points with the greatest influence on $\operatorname{cov}(X,Y)$ are in quadrants I and III.

Hence, a positive value for $\operatorname{cov}(X,Y)$ indicates a positive linear association between *X* and *Y*; that is, as the value of *X* increases, the value of *Y* increases. If the value of $\operatorname{cov}(X,Y)$ is negative, however, the points with the greatest influence are in quadrants II and IV. Hence, a negative value for $\operatorname{cov}(X,Y)$ indicates a negative linear association between *X* and *Y*; that is, as the value of *X* increases, the value of *Y* decreases. Finally, if the points are evenly distributed across all four quadrants, the value $\operatorname{cov}(X,Y)$ will be close to zero, indicating no linear association between *X* and *Y*.

Notation

In some books as well as in some external exams, the following notations are used for the variance of *x*, variance of *y*, and the covariance of *x* and *y*, respectively.

$$s_{xx} = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{1}{n}\sum x^2 - \bar{x}^2,\ s_{yy} = \frac{1}{n}\sum(y-\bar{y})^2 = \frac{1}{n}\sum y^2 - \bar{y}^2,$$

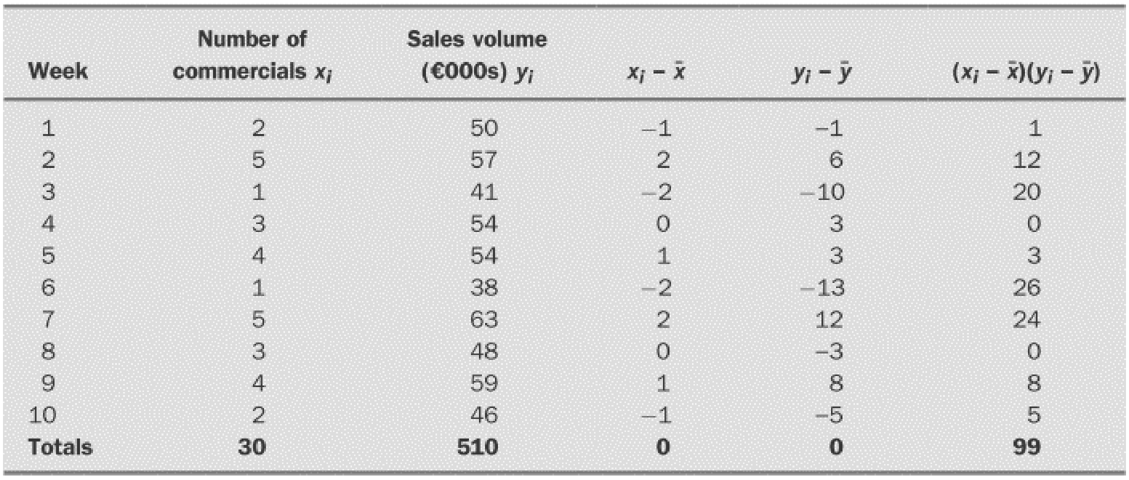
and $s_{xy} = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\sum xy - \bar{x}\,\bar{y}.$

Sometimes, “big” *S*’s are used instead in the computation of the coefficients of the regression line that is discussed below. These are:

$$S_{xx} = ns_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - n\bar{x}^2,\ S_{yy} = ns_{yy} = \sum(y-\bar{y})^2 = \sum y^2 - n\bar{y}^2, \text{ and}$$
$$S_{xy} = ns_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - n\bar{x}\,\bar{y}$$

Example 1

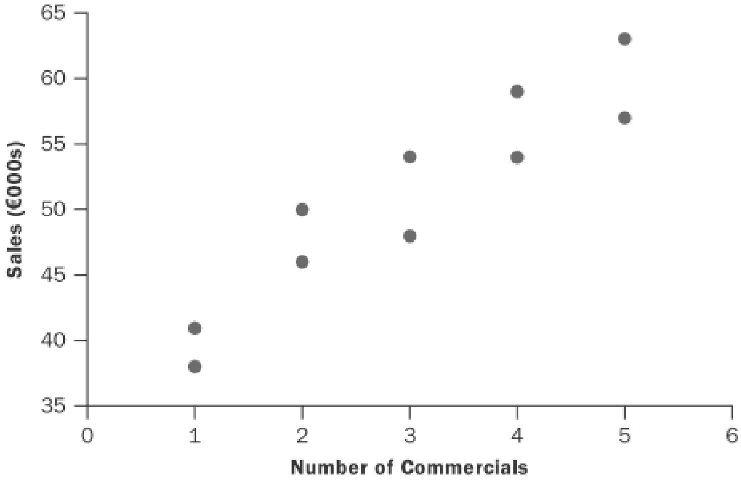
The store’s manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the hi-fi equipment store during the following week. Sample data with sales expressed in €000s were given in table below. Draw a scatterplot, find covariance and comment the results.



Week	Number of commercials x_i	Sales volume (€000s) y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	50	−1	−1	1
2	5	57	2	6	12
3	1	41	−2	−10	20
4	3	54	0	3	0
5	4	54	1	3	3
6	1	38	−2	−13	26
7	5	63	2	12	24
8	3	48	0	−3	0
9	4	59	1	8	8
10	2	46	−1	−5	5
Totals	**30**	**510**	**0**	**0**	**99**

Solution

Scatter diagram for the hi-fi equipment store is given below.



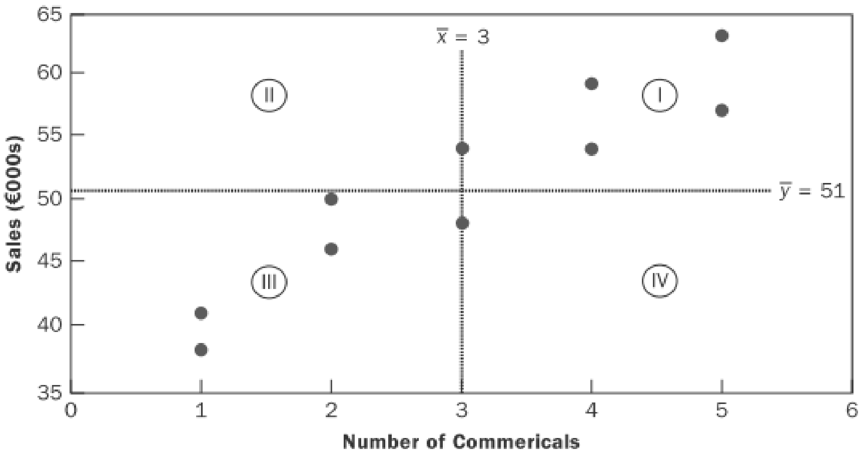
The scatter diagram shows a positive relationship, with higher sales (vertical axis) associated with a greater number of commercials (horizontal axis). The scatter diagram suggests that a straight line could be used as an approximation of the relationship.

To measure the strength of the linear relationship between the number of commercials X and the sales volume Y in the hi-fi equipment store problem, we compute the covariance.

The calculations are:

$$\bar{x} = \frac{30}{10} = 3,\ \bar{y} = \frac{510}{10} = 51,\ \bar{y} = \frac{510}{10} = 51,\ \operatorname{cov}(X,Y) = \frac{99}{10} = 9.9\,.$$

Partitioned scatter diagram for the hi-fi equipment store is given below.

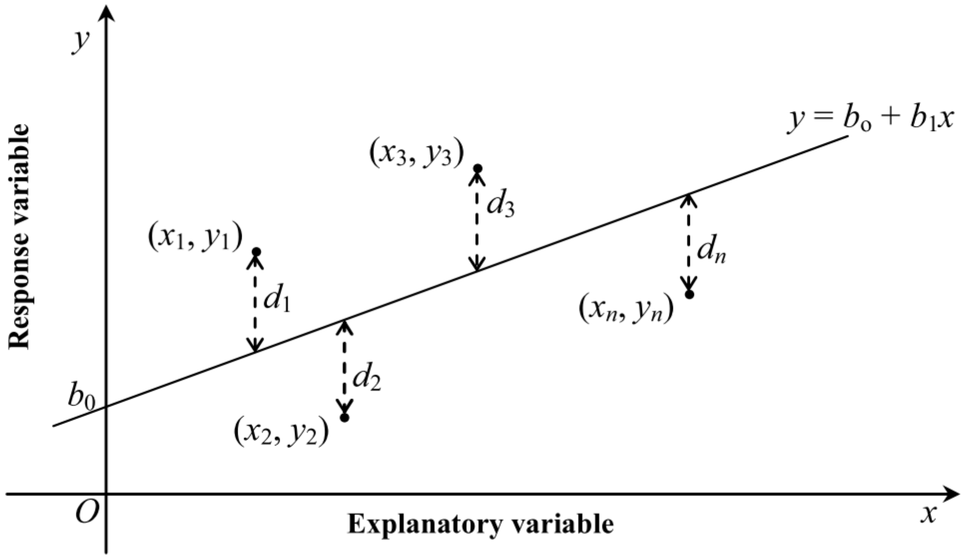


It is the same as the scatter diagram with a vertical dashed line at $\bar{x} = 3$ and a horizontal dashed line $\bar{y} = 51$. The lines divide the graph into four quadrants. Points in quadrant I correspond to x_i greater than $\bar{x}$ and y_i greater than $\bar{y}$. Points in quadrant II correspond to x_i less than $\bar{x}$ and y_i greater than $\bar{y}$ and so on. Hence, the value of $(x_i - \bar{x})(y_i - \bar{y})$ is positive for points in quadrants I and III, negative for points in quadrants II and IV .

We see that the scatter diagram for the hi-fi equipment store follows the pattern in the top panel of Figure above. As we expect, the value of the sample covariance indicates a positive linear relationship with $\text{cov}(X,Y) = 9.9$.

Least squares regression line of *Y* on *X*

To analyze the relationship between a dependent (outcome or response) variable *y* and an independent (predictor or explanatory) variable *x*, we represent the data using a scatterplot. A linear relationship between the two variables exists if the data can be approximated by a line.



Consider all lines with equations given by: $y = k + mx$. Let $d_i = y_i - (k + mx_i)$.

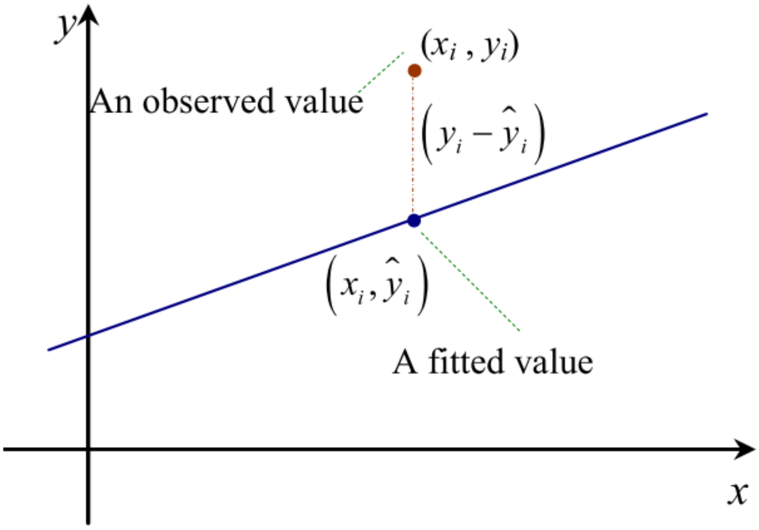
Set $D = \sum_{i=1}^{n} d_i^2$. Among all lines $y = k + mx$, consider the line that minimizes D. Such a line is called the least-squares regression line that best fits the data. Its coefficients are given by

$$m = b_1 = \frac{\operatorname{cov}(X,Y)}{Var(X)} = \frac{s_{xy}}{s_{xx}} = \frac{S_{xy}}{S_{xx}} \text{ and } k = b_0 = \bar{y} - b_1\bar{x}.$$

Note that, the regression line, passes through the center of gravity of the data as $y = \bar{y} - b_1\bar{x} + b_1x$ or $y - \bar{y} = b_1(x - \bar{x})$.

For a given pair of the data (x_i, y_i), y_i is the actual observation. The corresponding predicted, or fitted, value is $\hat{y}_i = b_0 + b_1x_i$ and $(y_i - \hat{y}_i)$ is called the *prediction error*.

The regression line is also used for predicting values of y for values of x not in the range of the data. So, if x_0 is not in the domain of the data, the predicted value corresponding to x_0 is $\hat{y}_i = b_0 + b_1x_0$.



Example 2

The yield, y grams, of a chemical reaction after x minutes is measured at five different times. The collected data are recorded in the table below.

Time, x minutes	10	14	24	32	40
Mass, y grams	8	24	36	42	48

For the collected data:

a) Draw a scatterplot.

b) Find the center of gravity and the covariance.

c) Find the linear regression line of y on x.

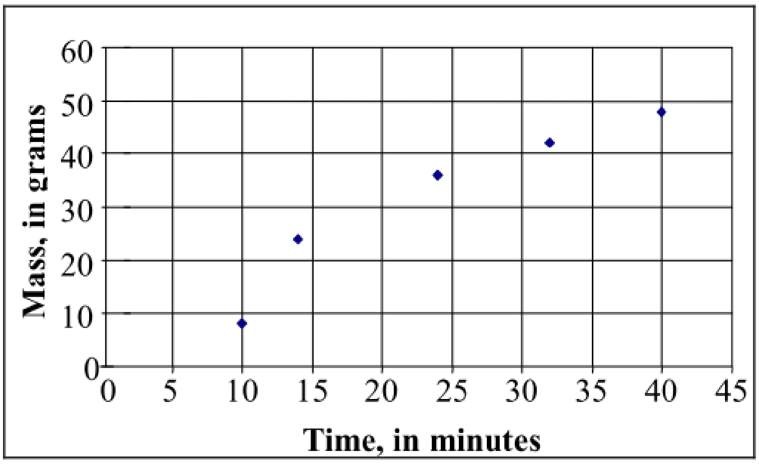
d) Find the error between the actual value and the predicted value of y when $x = 14$.

e) Based on the line found, predict the value of y when x is equal to

i. 28. ii. 45.

Solution

a) The scatterplot is as shown below.



b) We generate the table below to help answer this part and the remaining parts.

	x	y	x^2	xy
	10	8	100	80
	14	24	196	336
	24	36	576	864
	32	42	1024	1344
	40	48	1600	1920
$\sum$	120	158	3496	4544

$$\overline{x} = \frac{1}{n}\sum x_i = \frac{120}{5} = 24\,,\ \overline{y} = \frac{1}{n}\sum y_i = \frac{158}{5} = 31.6$$

The center of gravity of the data is (24, 31.6).

$$\text{cov}(X,Y) = s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x}\cdot\overline{y} = \frac{4544}{5} - \left(\frac{120}{5}\right)\left(\frac{158}{5}\right) = 150.4\,.$$

c) $\text{Var}(X) = s_{xx} = \frac{1}{n}\sum x_i^2 - \overline{x}^2 = 123.2$ and thus, $b_1 = \frac{\text{cov}(X,Y)}{\text{Var}(X)} = \frac{150.4}{123.2} = 1.22$.

The equation of the regression line is $y - 31.6 = 1.22(x - 24)$ which simplifies to $y = 1.22x + 2.32$.

d) The predicted value is $\hat{y} = 1.22(14) + 2.32 = 19.4$ and the actual value is 24. The error, d, is equal to $24 - 19.4 = 4.6$.

e) i. For $x = 28$, the predicted value of y is: $\hat{y} = 1.22\times 28 + 2.32 = 36.48$ g.

ii. For $x = 45$, the predicted value of y is: $\hat{y} = 1.22\times 45 + 2.32 = 57.22$ g.

Activity

1. The table summarizes the percentage of adult smokers in the United States in some selected years.

Year, x	1970	1980	1990	2000	2010
Percent of smokers, y	37.4	33.2	25.5	23.3	20.4

a) Draw a scatterplot of the data.

b) Find Var(X) and cov(X, Y).

c) Use the least-square method to find the regression line of y on x.

d) Based on the regression line, predict the percentage of smokers in the years 2005 and 2015.

2. The amount of gas needed for heating is measured for different values of the outside temperature on 10 days. The data is shown in the table below.

x, ºC	12	14	10	8	7.5	4.6	12.2	16.1	18.4	15.5
y, in kg	8	10	10	11.5	11	13	11	7	8	7

a) Draw a scatterplot for the data.

b) Suggest a relation between the temperature and the quantity of gas consumed. Is the relation positive or negative?

c) Find an equation of the regression line.

6.3. Correlation and Linearity

Two variables X and Y are linearly related if there exists a linear relation $y = ax + b$ that approximates the data. That is, if (x_0, y_0) is an actual data point then $\hat{y}_0 = ax_0 + b$ is close to y_0. If this is true for most of the data points, the linear relation is said to be strong, otherwise, it is called a weak linear relation.

To measure the strength of the linear relationship, we use a coefficient r called Pearson’s sample correlation coefficient or the product-moment correlation coefficient, or simply the *coefficient of correlation*.

Consider a sample of size n, (x_i, y_i), $i = 1, 2, 3, \ldots, n$, for measured values of two related variables X and Y. Let $\bar{x}$, $\bar{y}$ σ_x, and σ_y, denote their means and their standard deviations. The Pearson's sample correlation coefficient r of X and Y is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X,Y)}{\sigma_x \cdot \sigma_y} .$$

We may also verify that the slope of the regression line of y on x is given $b_1 = r\frac{\sigma_y}{\sigma_x}$.

Properties of the Pearson's sample correlation coefficient

Linearity between the two variables of bivariate data can be measured using the following properties of r

1° r lies between -1 and $+1$.

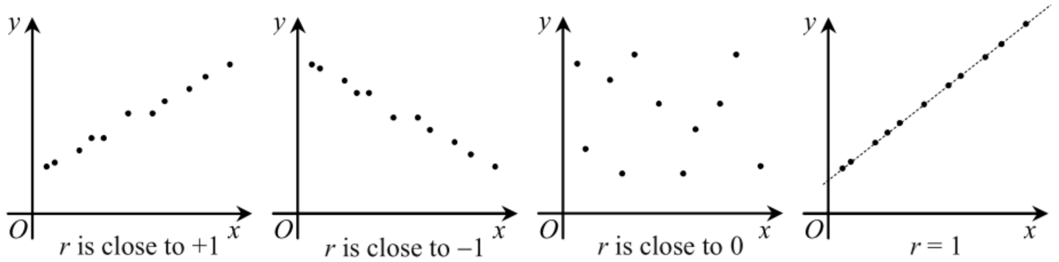
2° If $r = 1$ or $r = -1$, the regression line fits the data exactly.

3° A strong linear relation between x and y exists if r is close to 1 or to -1.
As a guideline, a relation is a strong linear relation if $|r| > \frac{\sqrt{3}}{2}$.

4° If $r = 0$, x and y cannot be considered as linearly related.

5° If r is close to 0, the linear relation between x and y is weak.

6° If $r < 0$, then the relation between x and y is negative and if $r > 0$ then the relation is positive.



r is close to $+1$ r is close to -1 r is close to 0 $r = 1$

.

Note that r is a pure number. It does not depend on the units of measurements. It stays the same if x and y are exchanged. It only measures the strength of the linear relation.

Example 1

A shop manager is studying his sales of jeans against the amount of money paid for advertisement. He records the following observations for eight different advertisement campaigns.

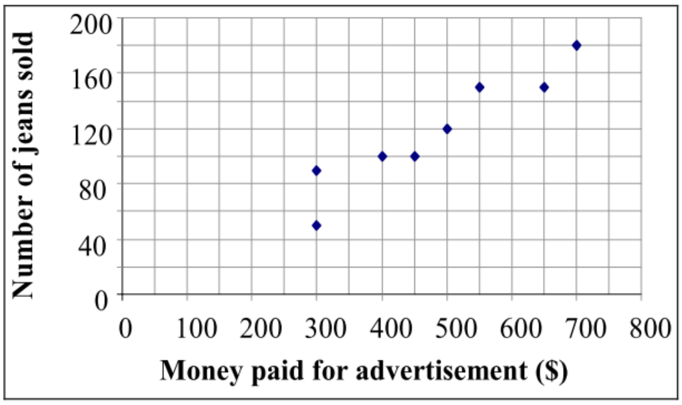
Money paid for advertisement (x_i), in dollars	300	450	400	500	300	650	700	550
Number of jeans sold (y_i)	50	100	100	120	90	150	180	150

Calculate the correlation coefficient and draw a scatterplot for the data. What conclusions can be drawn?

Solution

The calculations are: $n = 8$, $\bar{x} = 481.25$, $\bar{y} = 117.5$ and $r = 0.94$.

A scatterplot for the given data follows.



Conclusion: r is close to 1. There is a strong positive linear relation between the number of jeans sold and the advertisement cost.

Activity 1

A teacher is analyzing the exam scores for eight students. The teacher observes the pre-exam averages X and the exam scores Y. The following data were recorded.

x_i	50	70	100	100	90	60	80	90
y_i	60	70	90	100	80	70	80	95

a) Draw a scatterplot.

b) Calculate the correlation coefficient between the pre-exam averages and the exam scores.

c) Is there a relation between the pre-exam averages and the exam scores? Explain.

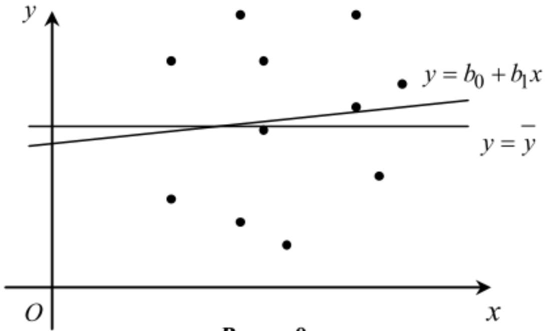
Coefficient of determination

To classify how good the least square method explains the data, a factor called the coefficient of determination is defined. This factor reflects the percentage of variation of *y* that may be explained by the regression line.

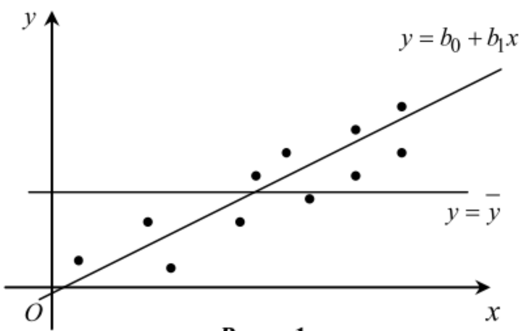
The coefficient of determination, or R-sq, is defined by:

$$R\text{-}sq = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

Note that, if there is a strong linear relation, then $\sum(y-\hat{y})^2$ is small compared to $\sum(y-\bar{y})^2$ and consequently, $\frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$ becomes close to $\frac{\sum(y-\bar{y})^2}{\sum(y-\bar{y})^2}$ which is equal to 1. On the other hand, if *X* and *Y* are not linearly related, then the regression line becomes close to the line $y = \bar{y}$, and so *R-sq* becomes close to 0.



R-sq ≈ 0



R-sq ≈ 1

Example 2

Water level in a reservoir is observed during ten consecutive rainy days. The water level is recorded at the end of each day.

Day (x_i)	1	2	3	4	5	6	7	8	9	10
Height (y_i), in cm	3	5	6	7	6	14	13	12	17	40

Find R-sq and interpret its value.

Solution

The mean value of the heights is: $\bar{y} = 12.3$.

The regression line of y on x is given by: $y = -3.4667 + 2.8667x$.

Using this regression line to find the fitted values of y that corresponds to the ten days of measurement, and substituting in the equation of the coefficient of determination yield R-sq = 0.652.

Conclusion: 65.2 % of the variation in the water levels can be explained by the least square method, with the number of days as explanatory variable.

Note that R-sq is sometimes denoted by r^2. This is the square of the correlation coefficient.

The correlation coefficient of the above data is $r = 0.807$, which means that we have a positive, but not strong, linear relation between the variables of the data.

Activity 2

A survey is done to check the correlation between the automobile consumption of fuel, in liters per mile, against its mass, in tons. A random sample of 10 cars yielded the data in the table below.

Mass of the car in tons	Distance covered per liter, in km
0.75	14.5
0.95	12.6
1.1	11.3
1.2	10.7
1.25	10.9
1.45	9.9
1.6	10.1
1.75	8.4
1.8	9.1
2.1	7.1

Draw a scatterplot for the data (use weights for the horizontal axis) and find the value of R-sq. Comment on the result.

Spearman's rank correlation coefficient

When the relation between *x* and *y* is near linear, Pearson's correlation coefficient can identify this relation well. But, in the case when this relation is not linear, Pearson's correlation coefficient will miss this relation. Moreover, Pearson's correlation coefficient can be greatly affected by the presence of even one outlier if it happens to be far from the main part of the scatter plot.

Spearman's rank correlation coefficient, r_s, uses the ranks of the values rather than the values themselves. The method for finding r_s is illustrated in the example below.

Example 3

Consider the data from Example 1.

x	300	450	400	500	300	650	700	550
Rank	1	4	3	5	2	7	8	6

y	50	100	100	120	90	150	180	150
Rank	1	3	4	5	2	6	8	7

Find r_S.

Solution

The rank pairs are (1, 1), (4, 3), (3, 4), (5, 5), (2, 2), (7, 6), (8, 8), and (6, 7). Spearman's correlation coefficient, r_s, is Pearson's correlation coefficient, *r*, applied to the rank pairs. For *n* items, the average of 1, 2, 3, …, *n* is $\overline{x} = \overline{y} = \dfrac{n+1}{2}$ and the standard deviation is $\sigma_x = \sigma_y = \sqrt{\dfrac{n(n+1)}{2}}$. The formula becomes,

$$r_s = \frac{12}{n(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \frac{n+1}{2})(y_i - \frac{n+1}{2}).$$

$x_i = x$-Rank	1	4	3	5	2	7	8	6
$y_i = y$-Rank	1	3	4	5	2	6	8	7

Substituting the values in the table above and $n = 8$ in the formula for r_s yields, $r_s = 0.95$, to two decimal places.

Another way to calculate Spearman's rank correlation coefficient, r_s is by using the formula $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ where d = rank x – rank y for each pair of values.

Referring to example 3,

$x_i = x$-Rank	1	4	3	5	2	7	8	6	
$y_i = y$-Rank	1	3	4	5	2	6	8	7	
d	0	1	–1	0	0	1	0	–1	
d^2	0	1	1	0	0	1	0	1	$\Sigma d^2 = 4$

$r_s = 1 - \dfrac{6 \cdot 4}{8(8^2-1)} = 0.95$, to two decimal places.

Remark

If there is a strong positive relation between x and y, then the larger x is the larger y is. The extreme case results in the rank pairs being (1, 1), (2, 2), (3, 3) , …, (n, n).

On the other hand, if there strong negative relation between x and y, then the larger x is the smaller y is. The extreme case results in the rank pairs being (1, n), (2, n – 1), (3, n – 2) , …, (n, 1).

Activity 3

Find Spearman's rank correlation coefficient for the data in the table below.

x	*y*
37.20	58.21
23.55	20.12
32.99	31.57
35.15	30.66
22.02	31.92
30.93	20.94
11.97	27.89
30.19	35.82
30.66	12.92

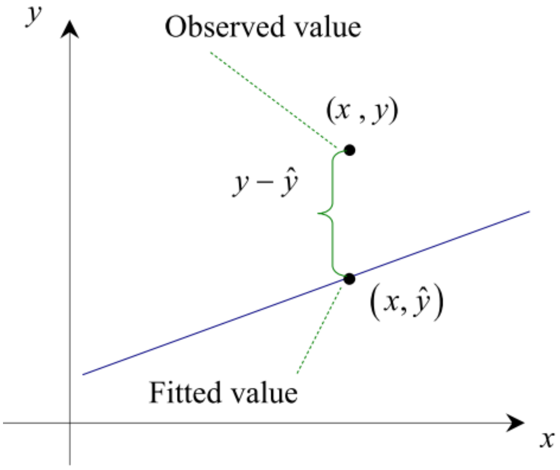
6.4. Residual Plots

Regression lines express the linear relationship between the explanatory variable and the response variable. The mathematical equations of regression lines are found by using the overall patterns. Deviations from the regression line give an insight into the behavior of the data. Extrapolation can be used to predict the response value for values on either side of the explanatory variable.

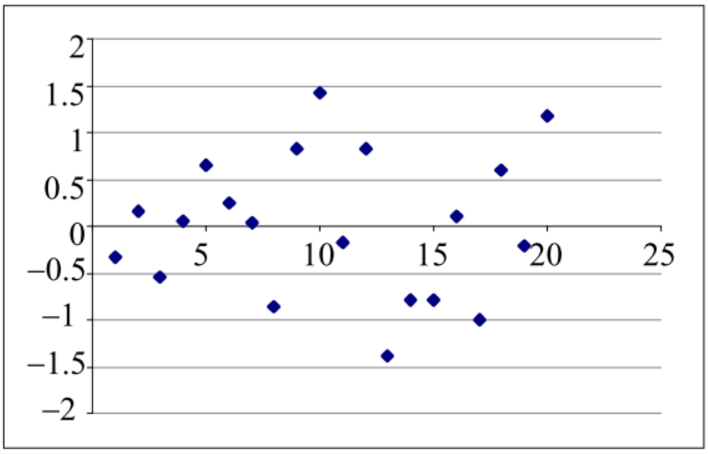
The residual

Let y be the observed value that corresponds to a given value x, and $\hat{y}$ be the predicted, or fitted value. The *residual* is the difference between the observed and the fitted value.

$Residual = y - \hat{y}$



Because some of the points in a scatterplot are below the regression line and some are above, it is obvious that some of the residuals are positive, and some are negative. For a linear model, the points in a residual plot are randomly dispersed around the horizontal axis. Otherwise, a non-linear model is more appropriate. The residual plot to the right represents a strong linear relationship between the variables.



Since residuals tell how far the data is from the regression line, examining them is helpful in assessing how well the regression line describes the data.

Note that the least square method insures that the sum of all the residuals is always equal to zero.

Example 1

The table below shows the average number of points per game, y_i, of a basketball player in different years, x_i.

x_i	1	2	3	4	5	6	7	8
y_i	12	19	23	27	28	25	20	14

Draw and investigate the residual plot and make necessary conclusions about the relationship between the number of points and years. Draw a scatterplot and use it to compare with the residual plot.

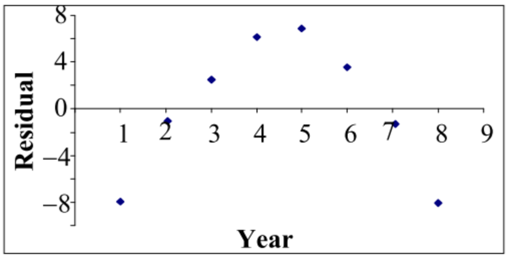
Solution

The equation of the regression line is: $y = 0.31x + 19.6$.

The fitted values and the residuals are tabulated next.

$\boldsymbol{x_i}$	1	2	3	4	5	6	7	8
$\boldsymbol{y_i}$	12	19	23	27	28	25	20	14
$\hat{y}_i$	19.92	20.23	20.54	20.85	21.15	21.46	21.77	22.08
$y_i - \hat{y}_i$	−7.92	−1.23	2.46	6.15	6.85	3.54	−1.77	−8.08

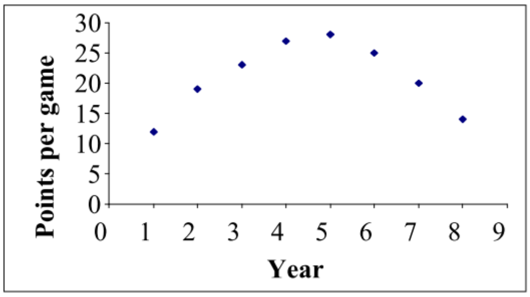
The residual plot is shown to the right.



The residuals have a strong pattern.

A line in this case is not a good model for the data.

A scatterplot for the given dataset is:



From the scatterplot, it is evident that a nonlinear relation is a better model for the given data.

Example 2

Draw a residual plot and determine whether a line is a good mathematical model for the data presented.

x_i	2	3	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	9	9	10	10
y_i	19	27	28	31	43	38	55	42	50	70	80	58	74	85	100	57	102	55	100	51	130	66

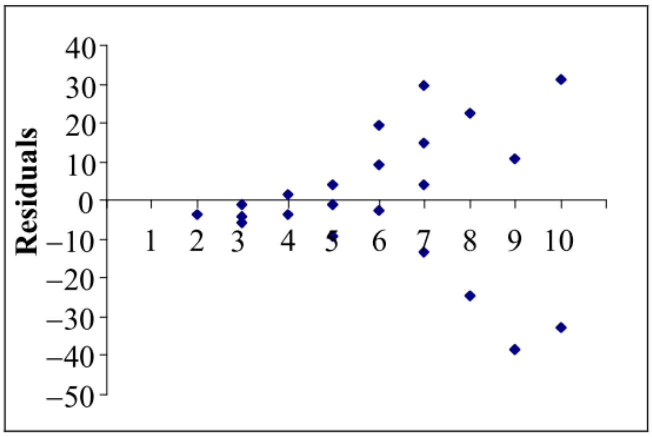
Solution

The regression line of the data is: $y = 9.5x + 4$.

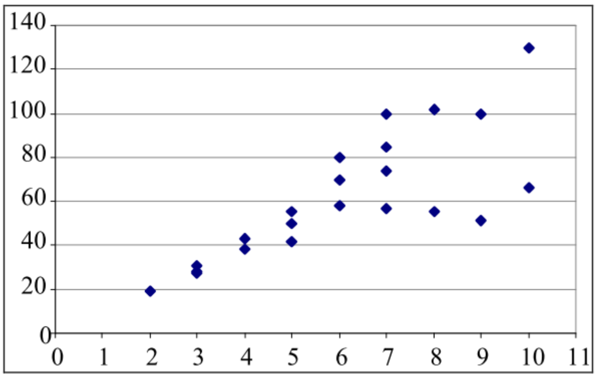
Calculating the residuals will lead to the following:

x_i	2	3	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	9	9	10	10
y_i	19	27	28	31	43	38	55	42	50	70	80	58	74	85	100	57	102	55	100	51	130	66
$y_i - \hat{y}_i$	–4	–5.5	–4.5	–1.5	1	–4	3.5	–9.5	–1.5	9	19	–3	3.5	14.5	29.5	–13.5	22	–25	10.5	–38.5	31	–33

The residual plot for the data follows.



This plot does not reveal any pattern in the data; although we can clearly state that using a line to predict the values of y is good for small values of x but becomes less accurate for larger values. A scatter plot for the data supports this claim.



Activity

A group of people were surveyed regarding the number of hours, per year, spent exercising. The average number of hours, y_i, for each age group, x_i, was taken. The results are tabulated in the table below.

x_i	24	30	36	42	48	54	60	66
y_i	360	300	250	230	230	180	180	130

a) Draw a residual plot.

b) Make necessary conclusions about the relationship between the numbers of hours spent exercising and age.

c) Draw a scatterplot. Does the scatterplot support the conclusions made in part (b)?

6.5 Nonlinear Relations

Many real-life situations cannot be modeled by a line, but can be represented well by a nonlinear model. We will consider here two relationships: exponential and power. In both cases, the relation between the two variables can be transformed to a linear relation. The tools established so far, can then be used with the transformed model and then the actual model can be interpreted.

Exponential relationship

Consider a bivariate data of predictor variable *x* and dependent variable *y*.

y is said to grow exponentially with *x* if it can be described by the mathematical relation: $y = b_0 \cdot b_1^x$.

In this case, the growth of *y* with respect to *x* is directly proportional to *y*.

To transform this relation to a linear relation, we take the logarithm of both sides:

$\log y = \log b_0 + x \log b_1$.

If we denote log *y* by *Y*, logb_0 by B_0, and logb_1 by B_1, the above relation becomes:

$Y = B_0 + B_1 x$.

The values of B_0 and B_1 can be found using the least square method. Taking the inverse log of both sides gives *y* in terms of *x*.

Example 1

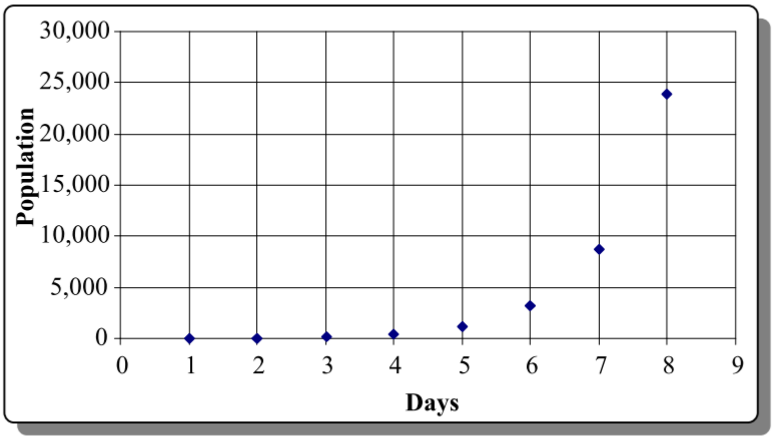
The data about pathogenic bacteria population growth is given in the table below:

Days x_i	1	2	3	4	5	6	7	8
Population, y_i	22	54	165	426	1195	3237	8763	23857

Investigate the growth of the population of the bacteria with respect to time and obtain a model for this growth.

Solution

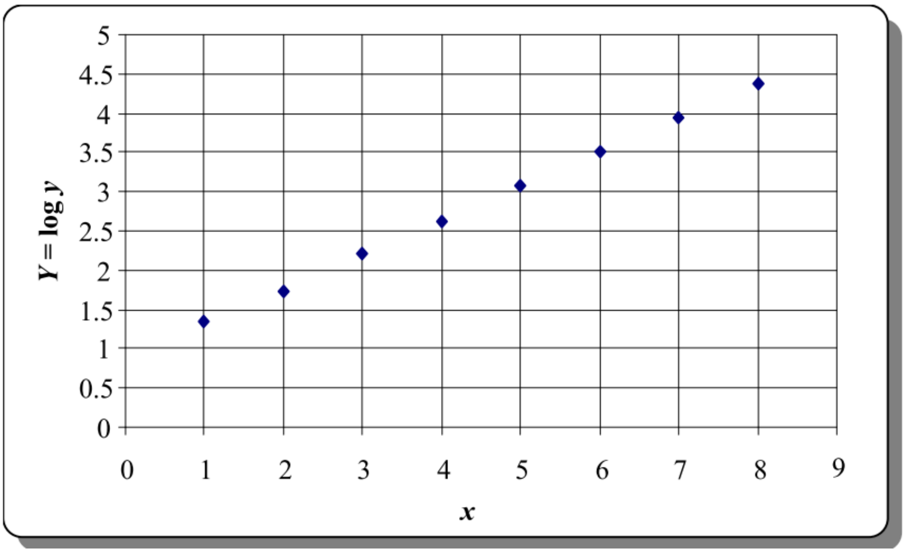
Below is a scatterplot for the given data.



The shape of the scatterplot suggests that y grows exponentially with x, hence we draw a scatterplot of $Y_i = \log(y_i)$ against the values of x_i and check if a linear model may be used to study the relation between Y and x.

Days, x_i	1	2	3	4	5	6	7	8
Population, y_i	22	54	165	426	1195	3237	8763	23857
$Y_i = \log(y_i)$	1.342	1.732	2.217	2.629	3.077	3.51	3.943	4.378

A scatterplot of $Y = \log(y)$ against x follows.



A strong linear relation is evident. The regression line of Y on x is given by:

$Y = 0.436x + 0.892$.

So, the values of y can be approximated by the relation: $\log(y) = 0.436x + 0.892$,

$y = 10^{0.436x + 0.892} = 7.798 \times 10^{0.436x} = 7.798 \times (2.729^x)$.

This relation gives an estimate of the population of the bacteria at any day x.

Activity 1

Amanda deposits money in a bank. The table below gives the amount of money in Amanda's account (in thousands of dollars) for different years.

x_i	1	2	3	4	5	6	7	8
y_i	0.2	0.4	0.8	1.6	3.2	6.4	12.8	25.6

a) Draw a scatterplot for the data.

b) Suppose that amount of money tends to accumulate exponentially, obtain the model for this growth.

c) Estimate the amount of money in Amanda's account after 10 years.

Power relationship

The predictor *x* and the dependent variable *y* of a bivariate data (*x*, *y*) are related by a power relationship if they can be connected by a mathematical model of the form $y = b_0 x^{b_1}$.

This relation may be transformed to a linear relation by taking the logarithm of both sides of the equation. As a result we get:

$$\log y = \log b_0 + b_1 \log x.$$

If we denote log *y* by *Y*, logb_0 by B_0, and log *x* by *X*, we obtain the following equality:

$$Y = B_0 + b_1 X.$$

Following this transformation, we apply the least square method to find the best fit of *Y* in terms of *X*, then use the inverse log to find a model of the dependent variable *y* in terms of the predictor variable *x*.

Example 2

Sally works at a marine research center. She researches flounders and wants to obtain a relationship between their weights and their lengths. Since a flounder is a flat fish, Sally wishes to investigate the hypothesis that the weight of a flounder should depend on the square of its length. If this is the case, then the model for this relationship should have the form $y = b_0 x^2$, where *y* is the weight of the flounder and *x* is its length.

Sally gathered the following data:

Length in cm **(x_i)**	**Weight in g** **(y_i)**
15.7	210
18.5	280
24.2	500
31.6	910
35.9	1120
38.1	1320
42.3	1620
45.4	1810
48.5	2130
50.9	2500

a) Draw a scatterplot for the data and for the logarithms of the data.

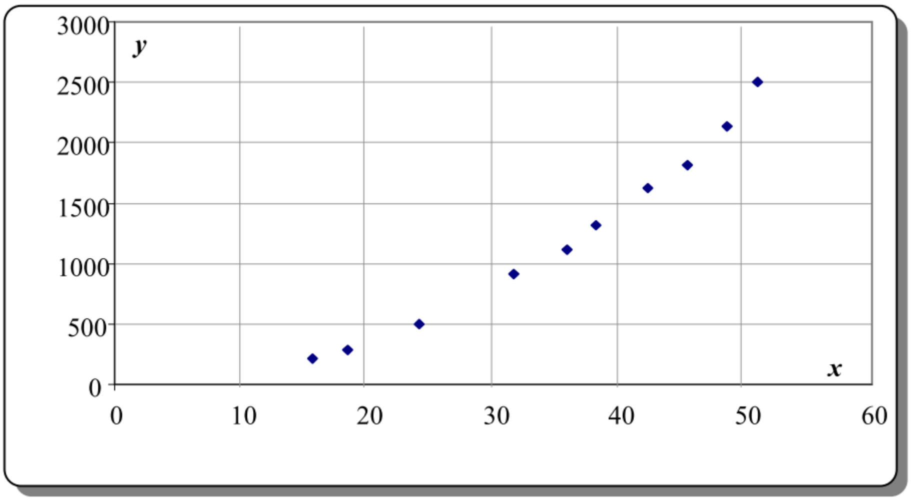
b) Using the logarithms of the data, find the coefficient of correlation. What do you conclude?

c) Use the least square method to estimate log(y) in terms of log(x). Deduce a mathematical relation that describes an estimate of y in terms of x.

d) Do the calculations support Sally's hypothesis regarding the relationship between the fish length and its weight? Explain.

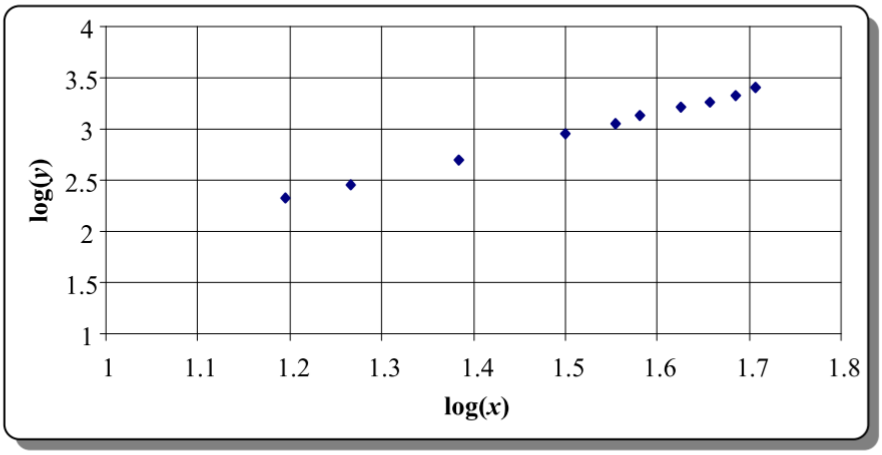
Solution

a) The scatterplot of the given data set is:



Length, x_i, in cm	**Weight, y_i, in g**	$X_i = \log(x_i)$	$Y_i = \log(y_i)$
15.7	210	1.196	2.322
18.5	280	1.267	2.447
24.2	500	1.384	2.699
31.6	910	1.5	2.959
35.9	1120	1.555	3.049
38.1	1320	1.581	3.121
42.3	1620	1.626	3.21
45.4	1810	1.657	3.258
48.5	2130	1.686	3.328
50.9	2500	1.707	3.398

Below is a scatterplot for the transformed data.



b) Using formula for r gives r = 0.9995, which shows a very strong linear relation between X and Y.

c) Using the least square method, the linear regression of Y on X is $Y = -0.1885 + 2.09X$ or $\log y = -0.1885 + 2.09 \log(x)$. The last expression can be written as or $\log y = \log 0.648 + 2.09 \log(x) = \log(0.648x^{2.09})$.

Taking the inverse log of both sides gives $y = 0.648x^{2.09}$.

d) Calculations do support Sally's hypothesis.

Minor modifications might be favorable (y seems to be proportional to $x^{2.1}$), but a bigger sample is needed to judge the necessity of these modification.

Activity 2

A researcher at a zoo is trying to find a relationship between the height and weight of different animals. He collected the following data.

Height (cm), x_i	**Weight (kg),** y_i
20.7	2.80
24.5	4.11
28.2	6.27
35.6	13.35
40.9	20.55
48.1	34.85
52.3	42.96
55.4	53.09
68.5	97.25
70.9	105.82

Based on these data, the researcher claimed that the weight of an animal is proportional to the cube of its height. Hence the mathematical model for the variables of the data would be of the form: $y = b_0 x^3$, where y is weight of the animal in kg and x is its height. Investigate the validity of the researcher's hypothesis.

7. PROBABILITY METHODS

7.1. Experiments, counting rules and probabilities

7.2. Random Variables

7.3. Probability Distributions

7.1. Experiments, counting rules and probabilities

Probability theory is the mathematical modeling of random phenomena.

Managers often base their decisions on an analysis of uncertainties such as the following:

1. What are the chances that sales will decrease if we increase prices?
2. What is the likelihood a new assembly method will increase productivity?
3. How likely is it that the project will be finished on time?
4. What is the chance that a new investment will be profitable?

Before going into the details we need to define some of the terms that are commonly used with probability theory.

- A *random experiment* is a process leading to two or more possible outcomes, without knowing exactly which outcome will occur.
- The *sample space*, or simply the space, of a random experiment is the set of all possible outcomes of this experiment. It is usually denoted by **Ω**.
- Any set of outcomes is called an *event*. Events are usually denoted by A, B, C, etc. An event is any subset of **Ω**.
- The empty set ∅ is called the *impossible event*.
- Ω is called the *certain event*.

Example 1

A coin is tossed twice.

a) Determine the sample space.

b) Write, in extension, the following events:

A: Two heads are obtained

B: The coin lands heads at least once

Solution

a) The outcome of the first toss is either heads (H) or a tails (T), and the same for the second toss. Hence the possible outcomes are HH, HT, TH, and TT. Thus the sample space is Ω = {HH, HT, TH, TT}.

b) A = {HH}

B = {HH, HT, TH}

Example 2

A fair die is rolled twice and the sum of the numbers appearing is observed.

a) Determine the sample space.

b) Write, in extension, the following events:

A: The sum is less than 4

B: The sum is greater than 14

C: The sum is greater than 5 but not greater than 8

c) Which of the following has more chance to occur: the sum of the numbers is 3 or the sum is 5?

Solution

a) The possible outcomes of rolling a die are 1, 2, 3, 4, 5, and 6. Hence, the possible outcomes of rolling a die twice are:

(1, 1), (1, 2), …, (1, 6),

(2, 1), (2, 2), …, (2, 6),

…

(6, 1), (6, 2), …, (6, 6).

The sample space of the experiment is: Ω = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}.

b) A = {2, 3}, B = {}, C = {6, 7, 8}

c) The sum 3 can be obtained from two possible outcomes: (1, 2) and (2,1). The sum 5 can be obtained from four possible outcomes: (1, 4), (2, 3), (3, 2) and (4,1). Since the die is fair, these outcomes have equal chance to occur. Therefore, the event 5 is more likely to occur.

Example 3

An investor follows the Dow Jones Industrial index. What are the possible basic outcomes at the close of the trading day?

Solution

The sample space for this experiment is as follows:

Ω = {{The index is higher than at yesterday's close}, {The index is not higher than at yesterday's close}}

One of these two outcomes must occur. They cannot occur simultaneously. Thus, these two outcomes constitute a sample space.

Activity 1

1. A coin is tossed three times.

a) Find the sample space.

b) Write in extension the following events:

F: The first toss lands heads

G: At least one of the tosses lands heads

H: At least one of the tosses lands heads and the second toss lands tails

2. A 4-sided fair die is rolled. If the number obtained is even, the die is rolled a second time. Find the sample space of the experiment.

Let A and B be two events in the sample space **Ω**. Their *intersection*, denoted by $A \cap B$, is the set of all basic outcomes in **Ω** that belong to both A and B.

Hence, the intersection $A \cap B$ occurs if and only if both A and B occur. We use the term joint probability of A and B to denote the probability of the intersection of A and B.

More generally, given *n* events $E_1, E_2, \ldots, E_n$, their intersection, $E_1 \cap E_2 \cap \ldots \cap E_n$, is the set of all basic outcomes that belong to every E_i, $i = 1, 2, \ldots, n$.

If the events A and B have no common basic outcomes, they are called *mutually exclusive*, and their intersection, $A \cap B$, is said to be the empty set, indicating that $A \cap B$ has no members.

Let A and B be two events in the sample space, Ω. Their *union*, denoted by A ∪ B, is the set of all basic outcomes in Ω that belong to at least one of these two events. Hence, the union A ∪ B occurs if and only if either A or B or both occur.

Given the n events $E_1, E_2, \ldots, E_n$ in the sample space, Ω, if $E_1 \cup E_2 \cup \ldots \cup E = \Omega$, these n events are said to be *collectively exhaustive*.

The probability of an event is a real number that ranges from 0 to 1 and that reflects the chance of that event to occur. The value 0 indicates that the event is impossible to occur, while the value 1 means that the event is certain to happen.

Let A and B be any two events of a sample space Ω. A probability, P, on Ω is a function that assigns values from the interval [0, 1] to subsets of Ω and satisfies the following axioms:

Axiom I: $P(\Omega) = 1$

Axiom II: If A and B are disjoint (or mutually exclusive) then

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

Axiom III: If $A \neq \varnothing$ then $P(A) \neq 0$.

Theorem 1. The probability of the impossible event is 0.

Theorem 2. If the events $A_1, A_2, \ldots, A_n$ are pairwise disjoint then:

$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$.

A finite sample space is said to be *equiprobable* if all of its simple events have the same probability of occurrence.

Theorem 3. If $\Omega = \{a_1, a_2, \ldots, a_n\}$ is an equiprobable space, then

1) $P(a_1) = P(a_2) = \ldots = P(a_n) = \frac{1}{n}$,

2) For any event A in Ω, $P(A) = \frac{n_A}{n}$, where n_A is the number of outcomes that satisfy the condition of event A, and n is the total number of outcomes in the sample space.

The last formula is called *classical probability*. The classical statement of probability requires that we count outcomes in the sample space. Then we use the counts to determine the required probability. The following example indicates how classical probability can be used in a relatively simple problem.

Example 4

A bag contains 12 identical balls numbered 1 to 12. A ball is randomly selected. Find the probability of getting a ball with

A: an even number

B: a prime number

C: an even number that divides 20.

Solution

A = {2, 4, 6, 8, 10, 12}. Since the balls are identical, we may assume that the space is equiprobable, $n_A = 6$, $n = 12$.

Therefore, $P(A) = \frac{n_A}{n} = \frac{6}{12} = \frac{1}{2}$.

$B = \{2, 3, 5, 7, 11\}$, $n_B = 5$, $P(B) = \frac{n_B}{n} = \frac{5}{12}$.

$C = \{2, 4, 10\}$, $n_C = 5$, $P(C) = \frac{n_C}{n} = \frac{3}{12} = \frac{1}{4}$.

Example 5

The employees in a small industrial firm are distributed as follows

	Administrator	Technician
Male	4	12
Female	8	3

Assuming that all employees are equally likely to be picked and an employee is to be picked to represent the company at a conference. What is the probability of selecting

a) A: a male employee?

b) B: a female administrator?

c) C: a technician?

Solution

a) There are 27 outcomes in the sample space, of which 16 are males. Therefore, $P(A) = \frac{16}{27}$.

b) 8 out of the 27 are female administrators, hence, P(B) = $\frac{8}{27}$.

c) 15 out of the 27 are technicians, hence, P(C) = $\frac{15}{27} = \frac{5}{9}$.

Example 6

Karlyn Akimoto operates a small computer store. On a particular day she has three Hewlett-Packard and two Dell computers in stock. Suppose that Susan Spencer comes into the store to purchase two computers. Susan is not concerned about which brand she purchases they all have the same operating specifications so Susan selects the computers purely by chance: Any computer on the shelf is equally likely to be selected. What is the probability that Susan will purchase one Hewlett-Packard and one Dell computer?

Solution

The answer can be obtained using classical probability. To begin, the sample space is defined as all possible pairs of two computers that can be selected from the store. The number of pairs is then counted, as is the number of outcomes that meet the condition one Hewlett-Packard and one Dell. Define the three Hewlett-Packard computers as H_1, H_2, and H_3 and the two Dell computers as D_1 and D_2. The sample space contains the following pairs of computers:

$\Omega = \{H_1D1, H_1D_2, H_2D_1, H_2D_2, H_3D_1, H_3D_2, H_1H_2, H_1H_3, H_2H_3, D_1D_2\}$.

The number of outcomes in the sample space is 10. If A is the event "one Hewlett-Packard and one Dell computer are chosen," then the number, n_A, of outcomes that have one Hewlett-Packard and one Dell computer is 6. Therefore, the required probability of event A " one Hewlett-Packard and one Dell" is P(A) = $\frac{n_A}{n} = \frac{6}{10} = 0.6$.

Activity 2

1. A fair die is rolled. Determine the sample space and find the probability of getting

M: a prime number

N: an odd prime number

2. A coin is tossed three times. Find the probability of getting

A: three similar tosses

B: at least two heads

C: at least two heads with the first toss to be heads

3. A 4-sided fair die is rolled twice. Determine the sample space and find the probability of each of the following events.

A: getting two numbers of opposite parity.

B: getting a sum of 9.

C: having a second roll larger than the first roll.

For any event A in a sample space, the event "A does not occur" is called the *complement (opposite)* of A and is denoted by A^C.

Theorem 4. For any event A, $P(A) + P(A^C) = 1$.

Theorem 5. If A and B are two events in a sample space Ω such that $A \subseteq B$ then $P(A) \leq P(B)$.

Theorem 6. For any two events A and B in a sample space Ω we have:

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

Example 7

In a sports club, 40% of the participants swim, 30% jog, and 10% are in both sports. A member is randomly chosen. Consider the following events:

S: the member swims

G: the member jogs

a) Find the probability of S, G, and $S \cap G$.

b) Write each of the events below in terms of S and G, and their complements, then find the probability of each.

i. H: the member swims or jogs

ii. I: the member jogs but does not swim

iii. J: the member is not in both sports simultaneously

iv. K: the member does not swim or does not jog

v. L: the member does not swim and does not jog

Solution

a) P(S) = $\frac{40}{100} = 0.4$, P(G) = $\frac{30}{100} = 0.3$ and P(S∩G) = $\frac{10}{100} = 0.1$

b) i. H = S∪G, P(H) = P(S∪G) = P(S) + P(G) – P(S∩G) = 0.4 + 0.3 – 0.1 = 0.6

ii. Event I contains all members who jog, and do not swim.

Those are the members common to the sets G and S^C, which is the complement of S, therefore, I = G∩S^C.

P(I) = P(G∩S^C) = P(G) − P(G∩S) = 0.3 − 0.1 = 0.2

iii. J = $(S\cap G)^C$, P(J) = P[$(S\cap G)^C$] = 1 – P(S∩G) = 0.9

iv. K = S C ∪G^C = $(S\cap G)^C$ = J, P(K) = P(J) = 0.9

v. L = $S^C \cap G^C$ = $(S\cup G)^C$, P(L) = P$(S\cup G)^C$ = 1 – P(S∪G) = 1 – 0.6 = 0.4

Activity 3

1. Find the probability of each of the following events:

a) Two odd numbers appear upon rolling a fair die twice.

b) One or more tails appear when tossing 4 coins.

2. A card is randomly chosen from a standard deck of 52 cards. Find the probability of getting

a) a red card,

b) Ace,

c) Ace colored red,

d) a red card or an Ace, and

e) an Ace colored red or a black card.

3. In an elementary school, students can take music or art or both. From a group of 24, 16 take music, 12 take art, and 8 take both music and art. A student is randomly chosen. Denote by M the set of students who take music and by A those who take art.

Find the probability of

D: getting a student who takes music and art

E: getting a student who takes music or art

F: getting a student who takes music but doesn’t take art

4. Given a sample space S and two events A and B such that $P(A) = 0.6$, $P(A\cap B^C) = 0.15$, and $P(A\cup B)^C = 0.3$.

a) Draw a Venn-diagram and label the areas that represent $A\cap B$, $A^C \cap B$, $A\cap B^C$, and $(A\cup B)^C$.

b) Find $P(A\cap B)$ and $P(B)$.

Let B be a non-empty subset of a sample space **Ω**.

The probability of an event A restricted to B, or the probability that A occurs knowing that B has occurred, is called the *conditional probability* of A on B and defined by:

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

P(A|B) is read “the probability of A knowing B occurred”.

An event A is said to be independent of event B if $P(A|B) = P(A)$. On the other hand, if $P(A|B) \neq P(A)$ then A and B are said to be dependent.

Substituting in the definition of P(A|B), we obtain the following rule for two independent events A and B:

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

Law of total probability

A family of non-empty sets $A_1, A_2, \ldots, A_n$ is said to form a partition of a sample space **Ω** if they are mutually disjoint and exhaustive. That is $A_i \cap A_j = \varnothing$ for all $i \neq j$ and $A_1\cup A_2\cup\ldots\cup A_n = \Omega$.

Now, for any event E in **Ω** we have

$P(E) = P(A_1)\, P(E|A_1) + P(A_2)\, P(E|A_2)\ldots\, P(A_n)\, P(E|A_n)$.

This rule is known as the *law of total probability*.

Using the rule for conditional probability we get the rule, known as Bayes’ rule:

$$P(A_i\,|\,E) = \frac{P(A_i)P(E\,|\,A_i)}{P(A)}.$$

Counting rules, combinations and permutations

Statistics is based on sampling. A population is usually large but finite. When a random sample is chosen, we assume that all individuals of the population have the same chance of being selected.

In this section, we will consider two ways of sampling: with replacement and without replacement. We will also develop methods for counting the number of possible ways these samples may be picked, which will be essential later on in evaluating probabilities and in defining probability laws for different sample spaces.

Basic counting principle

Consider an experiment E taking place in N different stages, and let n_k be the number of ways in which stage k may occur, for $k = 1, 2, 3, \dots, N$. Altogether, the number of ways in which E may occur is given by: $n_1 \cdot n_2 \cdot \dots \cdot n_k$.

This rule is known as the basic counting principle.

Example 8

A student wants to check out three books from the library: one book to study chemistry, one to study biology, and one to review for his statistics exam. He narrowed his choices to 6 biology books, 3 chemistry books, and 2 statistics books. In how many possible ways can the student choose the three books?

Solution

This is an event occurring in three stages with 6, 3, and 2 possibilities for each of the three stages respectively. Hence, the number of possible ways for this event to occur is: 6 × 3 × 2 = 36 possible ways.

Example 9

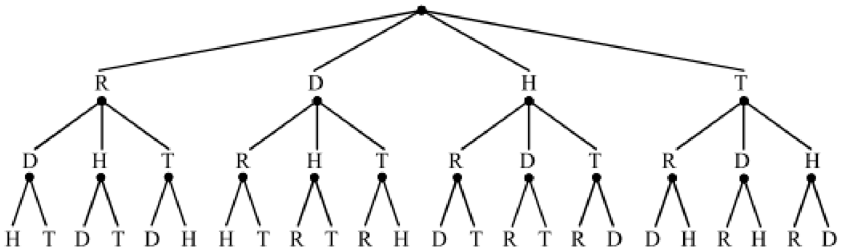
Rhoda, Diana, Hanna, and Tania formed a committee. Three conferences should be attended by one member of the committee each. In how many different ways can three ladies be selected for attending the conferences if

a) a lady cannot attend more than one conference?

b) a lady can attend one, two, or all three conferences?

Solution

a) For abbreviation, we will use the initials of the ladies names. For the first conference, we have 4 different choices: R, D, H, and T. But any of these girls if chosen for the first conference, cannot be chosen again for the second conference. Hence, we are left with 3 choices for the second conference. Using the same analogy, we deduce that 2 possibilities are what is left to choose from for the third conference. The tree diagram below illustrates these possibilities.



This diagram shows 24 different choices: RDH, RDT, …., THD. The answer is also obtained by using the basic counting principle: 4 × 3 × 2 = 24.

b) If a lady can attend more than one conference, then the number of choices for a given conference is not diminished by a previous choice or choices, hence the number of ways is: 4 × 4 × 4 = 64.

Activity 4

1. There are 4 math books, 3 chemistry books, and 6 biology books on a bookshelf. In how many ways can a student choose

a) a book?

b) one book of each kind?

2. A restaurant has a menu of 4 starters, 6 entrées, and 3 desserts. Find the number of ways in which a customer may choose a starter, an entrée, and a dessert.

3.How many 4-digit numbers are there?

4. How many 4-digit numbers are there such that

a) the number is divisible by 5?

b) all digits of the number are distinct?

Permutation

An arrangement of r objects from a set of n objects ($r \leq n$) is called a *permutation*. The number of permutations of r-objects from a set of n-objects is denoted by ${}_nP_r$ and can be obtained by using the basic counting principle as follows:

There are n possible choices for the 1st position.

There are $n - 1$ possible choices for the 2nd position.

There are $n - 2$ possible choices for the 3rd position. …

There are $[n - (r - 1)]$ possible choices for the rth position.

By the basic counting principle, the value of ${}_nP_r = n(n - 1)(n - 2)\dots(n - r + 1)$.

Using the factorial notation, this formula may be written as: ${}_nP_r = \dfrac{n!}{(n-r)!}$.

Combination

A *combination* of r objects from a set of n objects ($r \leq n$) is a selection of r objects from this set. The order in which we choose the objects is not important. The number of combinations of r objects from a set of n objects is denoted by ${}_nC_r$.

We can therefore write: ${}_nC_r = \dfrac{n!}{r!(n-r)!}$.

Example 10

A company wants to elect 5 members for its board. The number of employees qualified for the board is 20, of which 14 are men and 6 are women.

a) How many possible boards can be formed?

b) If the board must consist of 3 men and 2 women, then how many possible boards are there?

c) If the members of the board cannot all be of the same gender, then how many possible boards are there?

Solution

a) The order in which we choose the members of the board is not important; hence this is a combination of 5 elements from a set of 20. The number of possible boards is: ${}_{20}C_5 = \dfrac{20!}{5!(20-5)!} = 15{,}504$.

b) The number of ways in which we can choose 3 men for the board is ${}_{14}C_3 = 364$ and the number of ways in which we can choose 2 women for the board is ${}_6C_2 = 15$. By the basic counting principle, the number of possible boards is: 364 × 15 = 5,460.

c) The number of possible boards consisting of 5 men is ${}_{14}C_5 = 2{,}002$, and the number of possible boards consisting of 5 women is ${}_6C_5 = 6$. Therefore, the number of possible boards not consisting of unique gender is: 15,504 − (2,002 + 6) = 13,496.

Activity 5

A drawer contains 7 blue shirts and 5 red shirts. Find the number of ways three shirts can be drawn from the drawer if

a) they can be of any color.

b) they must be of the same color.

7.2. Random Variables

In effect, a random variable associates a numerical value with each possible experimental outcome. The particular numerical value of the random variable depends on the outcome of the experiment.

A *random variable* is a variable that takes on numerical values realized by the outcomes in the sample space generated by a random experiment. A *random variable* is a rule that assigns to each element, in a sample space **Ω**, a numerical value. In other words, a random variable on a sample space **Ω** is a function from **Ω** to the set of real numbers.

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

A random variable that may assume either a finite number of values or an infinite sequence of values such as 0, 1, 2, … is referred to as a *discrete random variable*.

A random variable that may assume any numerical value in an interval or collection of intervals is called a *continuous random variable*.

The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.

For a discrete random variable X, the probability distribution is defined by a probability function, denoted by $p(x) = \mathrm{P}(X = x)$ for all possible values x. The probability function provides the probability for each value of the random variable.

If X is a discrete random variable on a sample space Ω, and x_1, x_2, x_3, …, and x_n are the possible values of X (we will assume that $x_1 < x_2 < x_3 < \ldots <$ and x_n), then the probability distribution of X is $P(X = x_k) = P_k$ for $k = 1$ to n.

The set of values of the random variable and the corresponding probability distribution are often arranged in a table as follows:

x	x_1	x_2	x_3	…	x_n
$P(X = x)$	P_1	P_2	P_3	…	P_n

The probability distribution satisfies two conditions:

$P_k \geq 0$, and $P_1 + P_2 + P_3 + \ldots + P_n = 1$.

Example 1

A fair coin is tossed 4 times. Denote by X a random variable that assigns to each outcome the largest number of successive tails obtained in a trial.

a) Write explicitly the elements of the sample space Ω.

b) List the possible values of X and write to which element(s) of Ω each of these values corresponds.

c) Write the distribution of X.

Solution

a) The elements in Ω are:

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, and TTTT.

b) The possible values of X are 0, 1, 2, 3, and 4. The elements they correspond to are:

0: HHHH

1: HHHT, HHTH, HTHH, HTHT, THHH, THTH, THHT

2: HHTT, HTTH, THTT, TTHH, TTHT

3: HTTT, TTTH

4: TTTT

c) The probability distribution of X in tabular form is:

x	0	1	2	3	4
$f(x) = P(X = x)$	$\frac{1}{16}$	$\frac{7}{16}$	$\frac{5}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Activity

A bag contains 3 red balls and 6 black balls. 5 balls are chosen at random. Let X be a random variable representing the number of red balls obtained.

a) Find the possible values of X.

b) With the aid of combination techniques, find the distribution of X.

The *expected value*, or mean, of a random variable is a measure of the central location for the random variable. The formula for the expected value of a discrete random variable X follows in equation

$E(X) = \mu = x_1P_1 + x_2P_2 + x_3 P_3 + \dots + x_n P_n = \sum x_k P_k$.

The *variance* of a random variable X is defined by:

$Var(\mathrm{X}) = E(X - \mu)^2 = (x_1 - \mu)^2 P_1 + (x_2 - \mu)^2 P_2 + (x_3 - \mu)^2 P_3 + \dots + (x_n - \mu)^2 P_n = \sum(x_k - \mu)^2 P_k$.

The standard deviation of X is defined by: $\sigma_x = \sqrt{Var(X)}$.

The *mode* of a discrete random variable is the value with the highest probability. If two or more values have the highest probability, then these values are the modes.

Let X be a continuous random variable. A function f is called a probability density function (p.d.f.) of X if it satisfies the following two conditions:

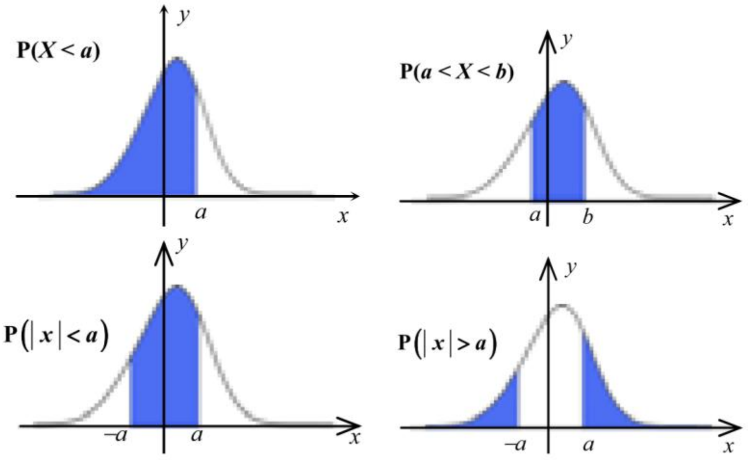
1) $f(x) \geq 0$ for all $x \in \mathrm{R}$,

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$.

The probability of an event is computed as $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Notice that $P(a \leq X \leq b)$ can be interpreted graphically as the area bounded by the curve $y = f(x)$ and the x-axis from $x = a$ to $x = b$.

Note that, for a continuous distribution and any $x \in R$, $P(X = x) = 0$ and thus, $P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$. Each of the figures below is a graphical representation of a probability density function of a random variable X. The shaded area represents the given probability in each case.



Example 2

Given the function

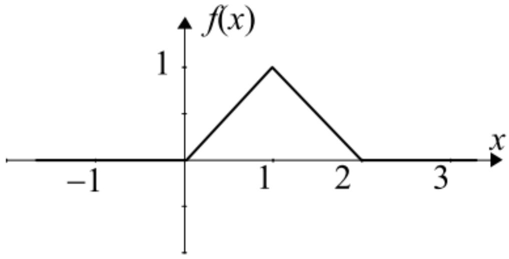
$$f(x) = \begin{cases} x & \text{if } 0 \le x < 1 \\ 2 - x & \text{if } 1 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases} .$$

a) Sketch the graph of f and verify that it satisfies the conditions of a probability density function.

b) Let X be a continuous random variable with probability density function f. Find $P(0 < X < 1.5)$ and $P(X > 1.2)$.

Solution

a) A sketch of the graph of f is shown below.



f is a p.d.f. as $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{1} f(x)dx + \int_{1}^{2} f(x)dx + \int_{2}^{+\infty} f(x)dx = 0 + 0.5 + 0.5 + 0 = 1\,.$$

b) $$P(0 \leq X \leq 1.5) = \int_{0}^{1.5} f(x)dx = \int_{0}^{1} f(x)dx + \int_{1}^{1.5} f(x)dx = 0.5 + \frac{(1+0.5)\times 0.5}{2} = \frac{7}{8}$$

$$P(X > 1.2) = \int_{1.2}^{+\infty} f(x)dx = \int_{1.2}^{2} f(x)dx + \int_{2}^{+\infty} f(x)dx$$

$$= \int_{1.2}^{2} (2-x)dx + \int_{2}^{+\infty} 0dx = \left[2x - \frac{x^2}{2}\right]_{1.2}^{2} + 0 = 0.32$$

Let X be a continuous random variable with a probability density function f.

The expected value (the mean) of X, if exists, is defined as:

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

The variance of X, if exists, is defined as:

$$Var(\mathrm{X}) = \mathrm{E}(X - \mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx\,.$$

The mode of a continuously distributed random variable X is the value of X at which its probability density function attains the absolute maximum. X may have no mode or more than one mode.

7.3. Probability Distributions

The Binomial Distribution / Bernoulli distribution

An experiment with exactly two outcomes S (for success) and F (for failure) is called a Bernoulli trial. A random variable in a Bernoulli experiment is said to have a Bernoulli distribution.

A sequence of n independent Bernoulli trials, where the number of success is to be observed, is called a binomial experiment. The probability function of the random variable X that assumes the number of successes in a binomial experiment is called a binomial distribution.

Consider a binomial experiment with n trials. The event (x successes and $n - x$ failures) consists of ${}_nC_x$ outcomes (as many ways as there are for picking x objects from a set of n objects). The probability of the successes is p, while the probability of the failure is $q = 1 - p$. Since the trials are independent then the probability of any of those outcomes is

$$\underbrace{p \cdot p \cdot p \cdot \ldots \cdot p}_{x\text{-times}} \cdot \underbrace{q \cdot q \cdot q \cdot \ldots \cdot q}_{(n-x)\text{-times}} = p^x \cdot q^{n-x}$$
.

Therefore, the probability of getting x successes in n trials, or the probability function of the experiment, is given by: $P(X = x) = {}_nC_x\, p^x\, q^{n-x}$, $x = 0, 1, \ldots, n$.

A commonly used notation for the binomial distribution is $B(n, p)$ where n indicates the number of repeated Bernoulli trials and p is the probability of success, n and p are called the parameters of distribution. Moreover, $X \sim B(n, p)$ means X is a random variable whose distribution is binomial.

Theorem: If $X \sim B(n, p)$ then $E(X) = np$ and $\text{Var}(X) = = npq$.

Activity 1

At a certain factory, the probability of a bulb being non-defective is 0.85. Out of 9 randomly selected bulbs, what is the probability that

a) all are non-defective?

b) at least 7 are non-defective?

The Geometric Distribution

Consider the experiment of repeating Bernoulli trials until a success is observed. Let X be a random variable that assigns to each outcome the number of trials needed. The distribution of such a random variable is called a geometric distribution.

Since X represents the number of trials needed untill the first success is observed, then the possible values of X are: $\{1, 2, 3, 4, \ldots\}$.

Hence, the probability function of X is P($X = x$) = $q^{x-1}\,p$, $x = 0, 1, \ldots$

Theorem: If X is a geometric distribution then E(X) = $\dfrac{1}{p}$, and Var(X) = $\dfrac{q}{p^2}$.

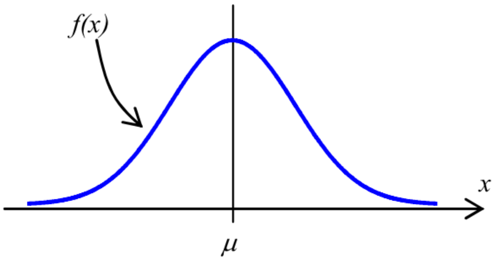
The Poisson Distribution

A phone operator receives on average a certain number of calls λ during a given time interval. It is of interest to find the probability of receiving x calls during this interval where $x = 0, 1, 2, 3, \ldots$. In this case, X is said to follow a Poisson distribution, X ~ P(λ), and the probability function of X is given by P($X = x$) = $\dfrac{\lambda^x}{x!}e^{-\lambda}$.

Theorem: If X ~ P(λ), then E(X) = λ and Var(X) = λ.

The Normal Distribution

Of all continuous distributions, the most important is the normal distribution. The probability density function of a normal distribution is given by $f(x) = \dfrac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}$, $-\infty < x < \infty$.



The normal curve

The graph of the probability density function of a normal distribution is a bell-shaped curve with a peak occurring at $x = \mu$ and $f(\mu) = \dfrac{e^{\frac{-(\mu-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} = \dfrac{1}{\sqrt{2\pi\sigma}}$. That is, the mode of the distribution is $x = \mu$. It is symmetrical about the vertical line $x = \mu$, thus, the median of the distribution is μ. The mean of the distribution is μ and its standard deviation is σ.

A random variable X that is normally distributed is denoted by $X \sim N(\mu, \sigma^2)$.

The probability for normally distributed variable X is P($a \le X \le b$) = $\int_a^b f(x)dx = \int_a^b \dfrac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}dx$.

Important markers of the normal distribution curve

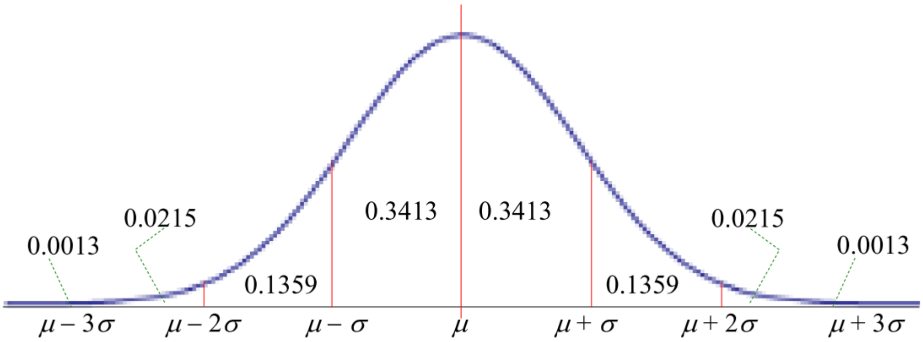
Since the median of the distribution $X \sim \mathrm{N}(\mu, \sigma^2)$, is μ then $\mathrm{P}(x \geq \mu) = \mathrm{P}(x \leq \mu) = 0.5$. Moreover,

$\mathrm{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6826$,

$\mathrm{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9544$, and

$\mathrm{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9974$.

That is, approximately 68.26% of all outcomes occur within one standard deviation from the mean, 95.44% of all outcomes occur within two standard deviations from the mean, and 99.74% of all outcomes occur within three standard deviations from the mean.



Important markers in the normal curve

The standard normal distribution

Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Consider the random variable $Z = \dfrac{X - \mu}{\sigma}$. It is proved that Z is normally distributed with mean 0 and standard deviation 1, $Z \sim \mathrm{N}(0, 1)$.

The standard normal cumulative distribution function which gives the probability of $Z \leq z$ is denoted by $\varphi(z)$. There exist tables that give values of $\varphi(z)$ in a tabular from for values of z between -3.5 and 3.5. For any value less than or equal to -3.5, $\varphi(z)$ is taken to be 0, and for values greater than 3.5, it is taken to be 1.

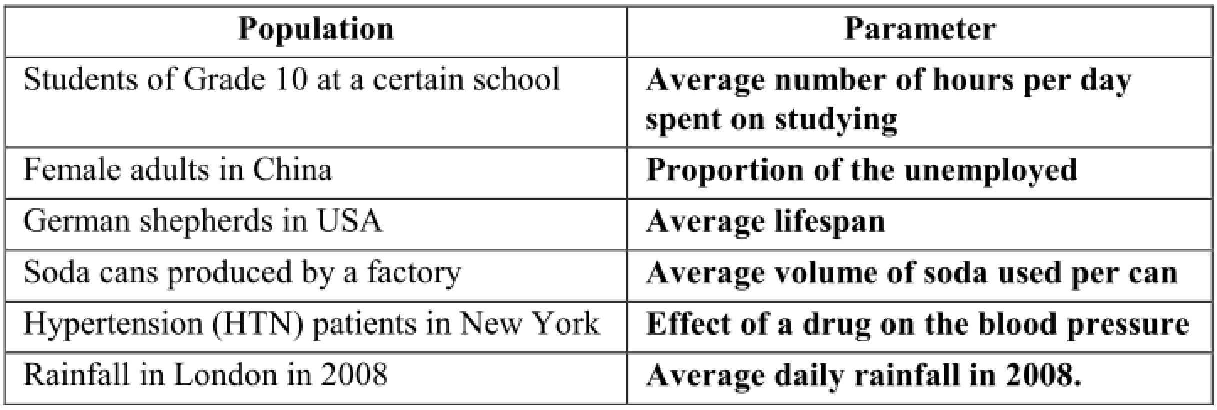
In some books, and some external exams, the standard normal cumulative distribution function $\varphi(z)$ is given for positive values of z only. The values of $\varphi(z)$ are usually given in a table for values of z between 0 and 3.5. For values less than zero, $\varphi(z)$ can be obtained using the symmetry of the bell-shaped curve.

8. SAMPLING AND ESTIMATION

8.1. Sampling

A *census* is collecting information from all the members of the target population. This kind of survey is unavoidable in some cases. For example, if an official count is needed to update medical cards, then every member of the population has to be checked. On the other hand, the attempt to conduct a survey by involving all the members of a population (census) might be too costly or too time consuming and in some cases it is virtually impossible. So we look to examine a smaller group of individuals (a sample) taken from the population. For example, a political researcher wants to know what percent of young adults aged 18 to 30 consider themselves democrats. A large University wants to find out what the students think about the food served on campus. The Federal Highway Safety Commission want to know how many people text while driving. In all these examples, we would take a subset of the population to draw conclusions about the larger group of individuals.

To start any statistical study, we first have to identify the population that we are targeting, and the parameter we want to estimate. Below are some examples of a population and a parameter that we might be interested in estimating.



Population	Parameter
Students of Grade 10 at a certain school	**Average number of hours per day spent on studying**
Female adults in China	**Proportion of the unemployed**
German shepherds in USA	**Average lifespan**
Soda cans produced by a factory	**Average volume of soda used per can**
Hypertension (HTN) patients in New York	**Effect of a drug on the blood pressure**
Rainfall in London in 2008	**Average daily rainfall in 2008.**

The reason we sample is to collect data to make an inference and answer a research question about a population. Numerical characteristics of a population (e.g. population mean, population standard deviation) are called *parameters*. Numerical characteristics of a sample (e.g. sample mean, sample standard deviation) are called *sample statistics*. Primary purposes of statistical inference are to make estimates and test hypotheses about population parameters using sample statistics.

Here are two situations in which samples provide estimates of population parameters:

1) A European car tire manufacturer developed a new tire designed to provide an increase in tire lifetime. To estimate the mean lifetime (in kilometers or miles) of the new tire, the manufacturer selected a sample of 120 new tires for testing. The test results provided a sample mean of 56 000 kilometers (35 000 miles). Therefore, an estimate of the mean tire lifetime for the population of new tires was 56 000 kilometers.

2) Members of an African government were interested in estimating the proportion of registered voters likely to support a proposal for constitutional reform to be put to the electorate in a national referendum. The time and cost associated with contacting every individual in the population of registered voters were prohibitive. A sample of 5000 registered voters was therefore selected, and 2810 of the 5000 voters indicated support for the proposal. An estimate of the proportion of the population of registered voters supporting the proposal was 2810/5000 = 0.562.

These two examples illustrate some of the reasons why samples are used. In the tire lifetime example, collecting the data on tire life involves wearing out each tire tested. Clearly it is not feasible to test every tire in the population. A sample is the only realistic way to obtain the tire lifetime data. In the example involving the referendum, contacting every registered voter in the population is in principle possible, but the time and cost are prohibitive. Consequently, a sample of registered voters is preferred.

It is important to realize that sample results provide only estimates of the values of the population characteristics, because the sample contains only a portion of the population. A sample mean provides an estimate of a population mean, and a sample proportion provides an estimate of a population proportion.

Some estimation error can be expected. With proper sampling methods, the sample results will provide ‘good’ estimates of the population parameters.

Let us define some of the terms used in sampling. The *sampled population* is the population from which the sample is drawn, and a *sampling frame* is a list of the elements from which the sample will be selected. In the second example above, the sampled population is all registered voters in the country, and the sampling frame is the list of all registered voters.

Several methods can be used to select a sample from a population, moreover sampling could be without replacement and with replacement (if objects from the population could be included in the sample two or more times).

One important method is simple random sampling. When we refer to simple random sampling, we assume that the sampling is without replacement and every individual in the population has an **equal and independent chance** of being selected.

A simple random sample of size n from a population of size N is a sample selected such that each possible sample of size n has the same probability of being selected. The number of different simple random samples of size n that can be selected from a finite population of size N is: $\dfrac{N}{n!(N-n)!}$.

Among the many methods used in simple random sampling, the most popular are drawing lots and using random number tables.

Sometimes the population is divided into groups of similar characteristics, called strata, before performing a simple random sample. This type of sampling is called *stratified random sampling*. A stratified random sampling is carried out to ensure that each of the groups is fairly represented. The frequency of the sample selected from each group is proportional to the size of that group.

For example, consider that there are three divisions in a factory: A, B, and C, with 45 workers in division A, 450 workers in division B, and 180 in division C. To choose a random sample of size 30 that fairly represents the different groups, choose

$\frac{45}{675}\cdot 30 = 2$ members from division A, $\frac{450}{675}\cdot 30 = 20$ members from division B, and $\frac{180}{675}\cdot 30 = 8$ members from division C.

8.2. Point estimation

To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a sample statistic. For example, to estimate the population mean μ and the population standard deviation σ we use the data to calculate the corresponding sample statistics: the sample mean and the sample standard deviation.

These computations are an example of the statistical procedure called *point estimation*. We refer to the sample mean as the point estimator of the population mean μ, the sample standard deviation as the point estimator of the population standard deviation σ, and the sample proportion as the point estimator of the population proportion. The numerical value obtained for the sample mean, sample standard deviation or sample proportion is called a *point estimate*.

The best estimator (unbiased) for the population's mean and the population's variance are, respectively, $\hat{\mu} = \overline{x} = \frac{\sum{}^{\backprime} x_i}{n}$ and $\hat{\sigma}^2 = \sigma_s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$, where σ_s denotes the sample standard deviation.

The best estimator for the population proportion p is the sample proportion p_s.

Example 1

The monthly salaries of 12 randomly selected employees from a large company, in thousands of dollars, are listed below.

2.2 3.5 1.8 2.75 1.12 3.4

1.6 2.6 1.9 2.25 1.95 3.45

a) Find estimates for the mean and the variance of the population.

b) An employee is considered highly paid if his wage exceeds $2,500. Find estimate for the proportion of highly paid employees.

Solution

a) $\hat{\mu} = \overline{x} = \dfrac{\sum x_i}{n} = \dfrac{28.52}{12} = 2.38$ is an estimate for the populations' mean.

$\hat{\sigma}^2 = \sigma_s^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1} = 0.602$ is an estimate for the populations' variance.

b) Out of the 12 selected employees, 5 are getting paid more than $2,500. Hence, an estimate for the sample proportion is: $p = \dfrac{5}{12}$.

Point estimation is a form of statistical inference. We use a sample statistic to make an inference about a population parameter. When making inferences about a population based on a sample, it is important to have a close correspondence between the sampled population and the target population. The target population is the population we want to make inferences about, while the sampled population is the population from which the sample is actually taken.

Activity

1. The following data are from a simple random sample.

5 8 10 7 10 14

a. Calculate a point estimate of the population mean.

b. Calculate a point estimate of the population standard deviation.

2. A survey question for a sample of 150 individuals yielded 75 Yes responses, 55 No responses and 20 No Opinion responses.

a. Calculate a point estimate of the proportion in the population who respond Yes.

b. Calculate a point estimate of the proportion in the population who respond No.

3. A simple random sample of five months of sales data provided the following information:

Month:	1	2	3	4	5
Units sold:	94	100	85	94	92

a. Calculate a point estimate of the population mean number of units sold per month.

b. Calculate a point estimate of the population standard deviation.

4. The data set Mutual Fund contains data on a sample of 40 mutual funds. These were randomly selected from 283 funds featured in Business Week. Use the data set to answer the following questions.

a. Compute a point estimate of the proportion of the Business Week mutual funds that are load funds.

b. Compute a point estimate of the proportion of the funds that are classified as high risk.

c. Compute a point estimate of the proportion of the funds that have a below-average risk rating.

8.3. Confidence Interval Estimation

The estimates we dealt with so far are all point estimates. Using a single observation to estimate a parameter of a population does not give us the ability to indicate how good our estimate is. If an observation of a sample has a mean equal to 10, we say that the mean of population is estimated to be 10. This method of estimation does not tell us how close our estimate is to the true value. Another way, fundamental in statistics, is to use interval estimation.

A confidence interval estimator for a population parameter is a rule for determining (based on sample information) an interval that is likely to include the parameter. The corresponding estimate is called a *confidence interval estimate*.

An interval (a, b) is said to be *$100(1 - \alpha)\%$ confidence interval* for a parameter θ if $P(a < \theta < b) = 1 - \alpha$ and α is called the *level of significance*.

This is the same as saying that we are $100(1 - \alpha)\%$ confident that the value of θ is between a and b. 90%, 95%, and 99% are typical percentages used for confidence intervals. The corresponding values of α are 0.1, 0.05, and 0.01, respectively.

Note that, when continuous distribution are being considered, $[a, b]$, $[a, b)$, and $(a, b]$ all have the same confidence level as (a, b).

Because a point estimator cannot be expected to provide the exact value of the population parameter, an interval estimate is often computed, by adding and subtracting a margin of error. The purpose of an interval estimate is to provide information about how close the point estimate might be to the value of the population parameter. In relatively simple cases, the general form of an interval estimate is:

Point estimate ± margin of error

The interval estimates of a population mean μ and a population proportion p have the same general form:

Population mean: $\overline{x}$ ± margin of error

Population proportion: p ± margin of error

Confidence interval for μ in a normal distribution when σ is known

Consider a random sample of size n from a normal population $N(\mu, \sigma^2)$ of unknown μ and known σ. The boundaries of a 100(1 – α)% confidence interval for μ are $\overline{x} \pm \frac{\sigma}{\sqrt{n}} z$ and the interval is $\left(\overline{x} - \frac{\sigma}{\sqrt{n}} z,\ \overline{x} + \frac{\sigma}{\sqrt{n}} z \right)$.

Here,

- z is referred to as the critical value that could be found using tables of the normal distribution,
- $\frac{\sigma}{\sqrt{n}} z$ is called the margin of error.

Critical values of *z*

The most common critical values of z that are used in producing confidence intervals are listed in the following table.

Accuracy required (in %)	**80**	**85**	**90**	**95**	**97.5**	**99**
α	0.2	0.15	0.1	0.05	0.025	**0.01**
z	**1.282**	**1.44**	**1.645**	**1.96**	**2.241**	**2.58**

Note that, the length of the confidence interval for μ is $2\frac{\sigma}{\sqrt{n}}z$. Thus, for a given value of σ, as n increases, the length of the confidence interval decreases. This is in agreement with our intuition. A larger sample should give a better estimate. Also, for a given value of σ, as α decreases, the length of the confidence interval increases. Again, this is in agreement with our intuition. A larger interval is more certain to contain the actual mean than a smaller interval.

Example 2

Suppose the $X \sim N(\mu, 9)$.

a) Find the length of a 100(1 − α)% confidence interval for μ if a sample of size 100 is taken.

i. $\alpha = 0.1$ ii. $\alpha = 0.05$ iii. $\alpha = 0.01$

b) Find the length of a 99% confidence interval for μ

in each of the following cases.

i. $n = 25$ ii. $n = 100$ iii. $n = 400$

c) Find the minimum size for a sample to produce a 99% confidence interval whose length does not exceed 1.

Solution

a) The length of the confidence interval for μ is $2\left(\sigma/\sqrt{n}\right)z_{\alpha/2}$.

i. $\alpha = 0.1$, $n = 100$, $\sigma = 3$, $z_{0.05} = 1.645$, $2\left(\sigma/\sqrt{n}\right)z_{0.05} = 2\left(\frac{3}{10}\right)1.645 = 0.987$.

ii. $\alpha = 0.05$, $n = 100$, $\sigma = 3$, $z_{0.025} = 1.96$, $2\left(\sigma/\sqrt{n}\right)z_{0.025} = 2\left(\frac{3}{10}\right)1.96 = 1.176$

iii. $\alpha = 0.01$, $n = 100$, $\sigma = 3$, $z_{0.005} = 2.58$, $2\left(\sigma/\sqrt{n}\right)z_{0.005} = 2\left(\frac{3}{10}\right)2.58 = 1.548$

b) $\alpha = 0.01,\ \sigma = 3,\ z_{0.005} = 2.58$

i. $n = 25,\ 2\left(\sigma/\sqrt{n}\right)z_{0.005} = 2\left(\dfrac{3}{5}\right)2.58 = 3.096$

ii. $n = 100,\ 2\left(\sigma/\sqrt{n}\right)z_{0.005} = 2\left(\dfrac{3}{10}\right)2.58 = 1.548$

iii. $n = 400,\ 2\left(\sigma/\sqrt{n}\right)z_{0.005} = 2\left(\dfrac{3}{20}\right)2.58 = 0.774$

c) $\alpha = 0.01,\ \sigma = 3,\ z_{0.005} = 2.58$

$2\left(\sigma/\sqrt{n}\right)z_{0.005} \le 1$ gives $2\left(\dfrac{3}{\sqrt{n}}\right)2.58 \le 1$. This yields, $n \ge 239.6304$.

Therefore, a sample of size 240 or more would insure the required specifications.

Activity 2

Suppose the $X \sim N(\mu, 16)$.

a) Find the length of a 95% confidence interval for μ if a sample of size 50 is taken.

b) Find the minimum size for a sample to produce a 90% confidence interval whose length does not exceed 0.5.

c) A sample of size 50 gives a 100(1 − α)% confidence interval of length 1.2. Find its level of significance.

Note, the same techniques used for the confidence interval when sampling from a normal distribution may be applied when sampling from a non-normal distribution. In this case, the size of the sample taken must be at least 30.

Confidence interval for μ of a non-normal distribution when σ is not known

When taking a large sample from a non-normal population with unknown μ and σ, we use $\hat{\sigma}^2 = \sigma_s^2 = \dfrac{\sum(x_i - \overline{x})^2}{n-1}$ to estimate σ^2. In this case, the confidence interval is

$$\left(\overline{x} - \frac{\sigma_s}{\sqrt{n}}z,\ \overline{x} + \frac{\sigma_s}{\sqrt{n}}z\right).$$

In order to construct a confidence interval in the case of $n < 30$, it is additionally necessary to require that X is normally distributed. In this case, the value z is found in the Student distribution table. If $n \geq 30$, then the Student distribution practically does not differ from the normal distribution.

The fact that the mean of a sample taken from a normal distribution obeys a normal distribution can be extended to any distribution by the well-known central limit theorem.

Theorem: **(Central limit theorem)** Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from any non-normal population of mean μ and variance σ^2. For large n ($n \geq 30$), the distribution of the sample mean can be approximated by a normal distribution of mean μ and variance $\frac{\sigma^2}{n}$.

The proof of this theorem can be found in many advanced texts on statistics.

Example 3

Consider all random samples of size 16 from a population that follows the distribution N(25, 4).

a) Describe fully the distribution of the sample mean.

b) A sample of size 16 is chosen. What is the probability that its mean is within 1 from the population mean?

Solution

a) The sampling is taken from a normal distribution. Hence the sample mean is normally distributed with mean μ = 25, and variance $\frac{\sigma^2}{n} = 0.25$.

b) $$\mathrm{P}(25-1<\bar{X}<25+1)=\mathrm{P}(-2<Z<2)=\mathrm{P}(Z<2)-\mathrm{P}(Z<-2)$$
$$=0.9772-0.0228=0.9544$$

Student's *T*-distribution

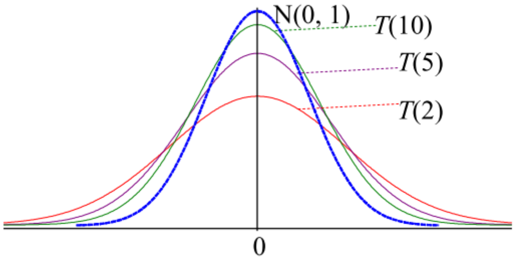
The probability distribution function of the *T*-distribution is complicated and not required at this stage, but the following properties are needed.

1. A t-distribution is fully specified by a positive integer γ, called its degrees of freedom (df). The notation *T*(γ) is used to specify the degrees of freedom.

2. The curve for the probability distribution function of $T(\gamma)$ is a bell-shaped curve symmetric about zero.

3. $T(\gamma)$ approaches the standard normal distribution as γ increases.

These properties are illustrated in the figure below.



The table of appendix gives the critical values of the T distribution for various degrees of freedom that are needed in interval estimation.

9. HYPOTHESIS TESTING

9.1. The null and the alternative hypothesis and Level of Significance

9.2. Significance Tests for the Mean

9.1. The null and the alternative hypothesis and Level of Significance

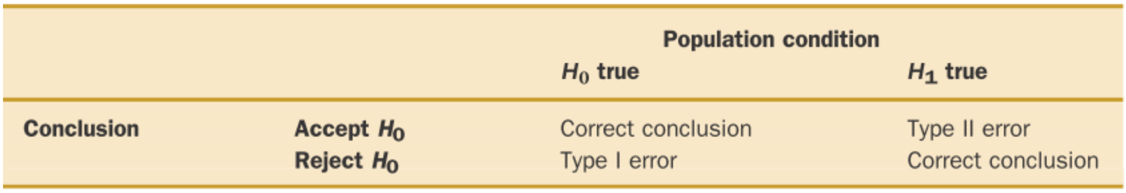
In hypothesis testing we begin by making a tentative assumption about a population parameter. This tentative assumption is called *the null hypothesis* and is denoted by H_0. We then define another hypothesis, called the *alternative hypothesis*, which is the opposite of what is stated in the null hypothesis. We denote the alternative hypothesis by H_a. The hypothesis testing procedure uses data from a sample to assess the two competing statements indicated by H_0 and H_a.

Usually, the null hypothesis: H_0: $\theta = \theta_0$ is tested against one of three alternative hypotheses 1) H_a: $\theta > \theta_0$, 2) H_a: $\theta < \theta_0$, 3) H_a: $\theta \neq \theta_0$. A critical value, called the *p*-value, is associated with each one of these alternative hypotheses. The p-value is the probability of obtaining the observed value or a value more extreme in the direction of the alternative hypothesis under the assumption that H_0 is true.

For reasons that will be clear later, the first two forms are called one-tailed tests (right-tailed and left-tailed). The third form is called a two-tailed test.

The null and alternative hypotheses are competing statements about the population. Either the null hypothesis H_0 is true or the alternative hypothesis H_a is true, but not both. Ideally the hypothesis testing procedure should lead to the acceptance of H_0 when H_0 is true and the rejection of H_0 when H_a is true.

Unfortunately, the correct conclusions are not always possible. Because hypothesis tests are based on sample information, we must allow for the possibility of errors. Table below illustrates the two kinds of errors that can be made in hypothesis testing.



		Population condition	
		H_0 true	H_1 true
Conclusion	Accept H_0	Correct conclusion	Type II error
	Reject H_0	Type I error	Correct conclusion

The first row of Table shows what can happen if the conclusion is to accept H_0. If H^0 is true, this conclusion is correct. However, if H_a is true, we make a Type II error; that is, we accept H_0 when it is false. The second row of Table shows what can happen if the conclusion is to reject H_0. If H_0 is true, we make a Type I error; that is, we reject H_0 when it is true. However, if H_a is true, rejecting H_0 is correct.

The *level of significance* is the probability of making a Type I error when the null hypothesis is true as an equality.

9.2. Significance Tests for the Mean

Tests about a population mean take one of the following forms:

The null hypothesis is H_0: $\mu = \mu_0$ and the alternative hypothesis is one of the following H_a: $\mu > \mu_0$, H_a: $\mu < \mu_0$ or H_a: $\mu \neq \mu_0$.

It is of interest to test whether the mean μ is equal to a certain value μ_0. A random sample results in a mean value $\bar{x}$ and $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ follows a standard normal distribution.

The first approach uses the value z to compute a probability called a p-value. The p-value measures the support provided by the sample for the null hypothesis, and is the basis for determining whether the null hypothesis should be rejected given the level of significance. The second approach requires that we first determine a value for the test statistic called the critical value. For a lower-tail test, the critical value serves as a benchmark for determining whether the value of the test statistic is small enough to reject the null hypothesis. We begin with the p-value approach.

The p-value approach has become the preferred method of determining whether the null hypothesis can be rejected, especially when using computer software packages such as MINITAB, IBM SPSS and EXCEL.

The p-value is a probability, computed using the test statistic, that measures the degree to which the sample supports the null hypothesis.

For H_a: $\mu > \mu_0$ p-value is $P\left(z > \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$.

For H_a: $\mu < \mu_0$ p-value is $P\left(z < \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$.

For H_a: $\mu < \neq \mu_0$ p-value is the minimum of value of the two probabilities P $\left(z > \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$ and P $\left(z < \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$.

H_0 is rejected at a level of significance α if the p-value is less than α in cases 1 and 2, and less than α/2 in case 3 and value α is the level of significance. Otherwise, H_0 cannot be rejected at this level of significance.

So, rejection rule using p-value:

Reject H_0 if p-value < α

So the test is significant if p-value < α.

The rejection rule for a test using critical value approach:

Reject H_0 if

$z < -z_\alpha$ in case of left-tailed test,

$z > -z_\alpha$ in case of right-tailed test,

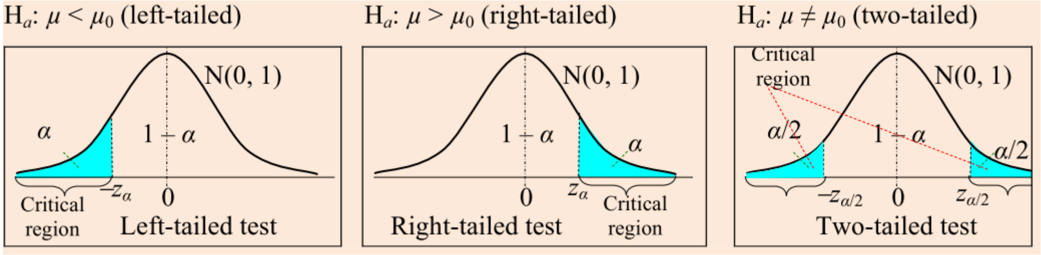
$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ in case of two-tailed test,

where z_α is the critical value; that is, the z value that provides an area of in the left/right tail of the standard normal distribution.

For left-tailed test the critical value z_α satisfies $P(z < z_\alpha) = \alpha$ and could be found from this equality using computer software packages such as MINITAB, IBM SPSS and EXCEL or tables for Normal distribution, and identify the critical region $(-\infty, z_\alpha)$.

For left-tailed test the critical value z_α satisfies $P(z > z_\alpha) = \alpha$ and the critical region $(z_\alpha, +\infty)$.

For two-tailed test the critical value $z_{\alpha/2}$ satisfies $P(z > z_{\alpha/2}) = \alpha/2$ and $P(z < z_{\alpha/2}) = \alpha/2$ and the critical region is $(-\infty, z_\alpha) \cup (z_\alpha, +\infty)$.



H$_0$ is rejected if the z-value lies in the critical region, otherwise we do not have enough evidence to reject H$_0$ at the α level of significance.

Note that there is the relationship between interval estimation and hypothesis testing.

We showed how to construct a confidence interval estimate of a population mean. For the σ known case, the confidence interval estimate of a population mean corresponding to 1 – α confidence coefficient is given by: $\bar{x} \pm \frac{\sigma}{\sqrt{n}} z$.

Doing a hypothesis test requires us first to formulate the hypotheses about the value of a population parameter. In the case of the population mean, the two-tailed test takes the form: H$_0$: $\mu = \mu_0$ and H$_a$: $\mu \neq \mu_0$, where μ_0 is the hypothesized value for the population mean. Using the two-tailed critical value approach, we do not reject H$_0$ for values of the sample mean that are within $-z_{\alpha/2}$ and $+z_{\alpha/2}$ standard errors of μ_0.

Hence, the do-not-reject region for the sample mean in a two-tailed hypothesis test with a level of significance of is given by: $\mu_0 \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$.

So, this provides insight about the relationship between the estimation and hypothesis testing approaches to statistical inference.

A confidence interval approach to testing a hypothesis of the form

H$_0$: $\mu = \mu_0$,

H$_a$: $\mu \neq \mu_0$.

1. Select a simple random sample from the population and use the value of the sample mean to construct the confidence interval for the population mean μ:

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}} z \,.$$

2. If the confidence interval contains the hypothesized value μ_0, do not reject H$_0$. Otherwise, reject H$_0$.

Example 1

A manager of a chocolate factory applies new measures to improve productivity. He claims that his new strategies have improved the daily production level, which previously was 420 kilograms. One of the shareholders doubts the manager's claims, so he decides to examine the quantities produced on some random days. The quantities produced during these days, to the nearest kilogram, are:

430 428 425 432 431 434 426 410

Two main questions arise:

1. Does the data provide enough evidence to support the shareholder's doubts?

2. If so, how confident is the shareholder about rejecting the manager's claim?

To answer these questions, we consider two hypotheses: the first hypothesis, which is usually denoted by H_0 and referred to as the null hypothesis, supports the shareholder's doubts that the mean daily production have not changed. The second, or the alternative hypothesis, H_a, is in favor of the manager's claim that the quantity produced per day did indeed increase.

Suppose that, in addition to the data given above, we know that the factory daily production is normally distributed with a standard deviation of 10 kilograms.

Assume the shareholder's claim to be correct, that is, the mean daily production is still 420 kg.

Based on the above results, state whether we should support the manager's claim.

Solution

The hypotheses are:

H_0: $\mu = 420$, is in favor of the shareholder's doubts.

H_a: $\mu > 420$, is in favor of the manager's claim.

$$\bar{x}\,\frac{\sum_{i=1}^{8} x_i}{8} = 427$$ and p-value is

$$\mathrm{P}\left(z > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) = \mathrm{P}\left(z > \frac{427-420}{10/\sqrt{8}}\right) = \mathrm{P}(z > 2.0) = 0.023 = 0.023 \text{ or } 2.3\%.$$

If H_0 is correct, then the probability of having an average of 427 or more in a sample of size 8 is about 2.3%, which is very small. Such a low probability leads us to believe that H_0 is not true and thus, with good statistical evidence, we may say that the daily mean production has increased.

Activity 1

State the null and the alternative hypothesis in each of the following situations.

a) An inspector is suspecting that a machine is not filling soda bottles with the right amount. The bottles are labeled 330 mL.

b) A contestant wants to check if he has improved his time in solving a puzzle cube after intensive training. His mean time before training was 212 seconds.

c) A council wants to check if the proportion of college graduates has increased in 2020 from previous years.

Example 2

A type of dry cell 1.5-volt battery has a mean lifetime of 165 hours with a standard deviation of 6 hours when used in a smoke detector. The mean lifetime of a random sample of 8 such batteries was found to be 162 hours. Assuming that the batteries lifetime follows a normal distribution, is there enough evidence to accept the fact that the mean lifetime is less than 165 hours? Use a 5% level of significance.

Solution

The null and alternative hypotheses are H_0: $\mu = 165$ hours and H_a: $\mu < 165$ hours.

A left-tailed test is performed. Thus, the critical value is $z_{0.05} = -1.65$.

The z-value is $\dfrac{162-165}{6/\sqrt{8}} = -1.41$ that is greater than the critical value ($-1.41 > -1.65$). This means that z-value is not in the critical region. So, there is not enough evidence to reject H_0 and accept H_a at 5% level of significance.

Activity 2

In example 2 above, assume that a different sample has a mean lifetime of 162 hours. A 1% level of significance was performed and the null hypothesis was rejected against the alternative hypotheses μ < 165 hrs. Is there enough evidence to accept the fact that the mean lifetime is less than 165 hours?

Example 3

A random sample of 16 elements is chosen from a normal distribution whose standard deviation is 2.4. The sample's mean is 20.9. It was believed that the distribution has a mean of 22.6. Perform a test to check whether we can reject the hypothesis claiming that the mean is 22.6 versus the alternative hypothesis that it is not. Use 1% level of significance.

Solution

The null and alternative hypotheses are: H_0: $\mu = 22.6$ and H_a: $\mu \neq 22.6$.

A two-tailed test is performed. Thus, the critical values are $z_{0.005}$ and $-z_{0.005}$.

$P(z > z_{0.005}) = 0.005$ gives $z_{0.005} = 2.58$.

The *z*-value is $\dfrac{20.9 - 22.6}{2.4/\sqrt{16}} = -2.86 < -z_{0.005}$. This means that *z*-value lies in the critical region. So, the null hypothesis is rejected at this level of significance.

To sum up, the *steps involved in hypothesis testing* are:

Step 1: State the hypotheses H_0 and H_a.

Step 2: Specify the level of significance. Collect the sample data and compute the value of the test statistic.

p-value approach

Step 3: Using the appropriate table or a calculator, find the p-value.

Step 4: Compare the p-value with the level of significance.

Step 5: Draw a conclusion on whether to reject H_0. (Reject H_0 if the p-value < α).

Critical value approach

Step 3: Using the appropriate table or a calculator, find the critical value.

Step 4: State the rejection region.

Step 5: Draw a conclusion on whether to reject H_0.

We showed that an interval estimate of a population mean for the unknown case is based on a probability distribution known as the *T*-distribution. Hypothesis tests about a population mean for the unknown case are also based on the *T*-distribution. The test statistic has a *T*-distribution with $n - 1$ degrees of freedom.

So for testing we use $t = \dfrac{\overline{x} - \mu_0}{\sigma_s / \sqrt{n}}$ instead of *z* and *t* follows a *T*-distribution.

PRACTICAL TASKS

1. DATA: TYPES OF DATA, GROUPING DATA AND FREQUENCY TABLES

In 1 – 5, classify the data as quantitative or qualitative. If it is quantitative, classify it further as discrete or continuous.

1. A real estate agent advertises for houses using the following data:

a) Number of rooms

b) Name of building company

c) Number of bedrooms in the house

d) Floor space, in square feet

e) District of the City

f) The age of the house, in years

2. Guests at a hotel are required to provide information at the time of their check-in. Below are some of the information that guests need to provide.

a) Name

b) Passport number

c) Country of origin

d) Type of room needed (single/double/other)?

3. Information gathered about the readers of a certain magazine.

a) Age

b) Marital status

c) Number of subscripted sites

d) Annual income

e) Average reading time per week

4. Students at a college were surveyed. Below are some of the information they were asked to provide.

a) Class (Freshmen, Sophomore, …)

b) Number of years in university

c) Major

d) Minor

e) Average number of credits per semester

f) Favorite professor

5. At the end of the term, the students of a college were asked to fill a questionnaire about the courses they took during that term. Some of the information needed to be answered were:

a) Name

b) Gender

c) Number of courses they completed during the term

d) Difficulty of the course (Choose from 1 to 5: 1 is very easy, 5 is extremely challenging)

e) Difficulty of the final exam (Choose from 1 to 5: 1 is very easy, 5 is extremely challenging)

6. State whether each of the following variables is categorical or quantitative and give the level of measurement[1].

a) Annual sales.

b) Soft drink size (small, medium, large).

c) Earnings per share.

d) Method of payment (cash, cheque, credit card).

[1] Qualitative data include nominal and ordinal levels of measurement.

Nominal data are considered the lowest or weakest type of data, since numerical identification is chosen strictly for convenience and does not imply ranking of responses. Ordinal data indicate the rank ordering of items, and similar to nominal data the values are words that describe responses.

7. A mortgage company randomly samples accounts of their time-share customers. State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?

a) The original purchase price of a customer’s time-share unit

b) The state (or country) of residence of a time-share owner

c) A time-share owner’s satisfaction level with the maintenance of the unit purchased (1: very dissatisfied to 5: very satisfied)

d) The number of times a customer’s payment was late.

8. Visitors to a supermarket in Singapore were asked to complete a customer service survey. Are the answers to the following survey questions categorical or numerical? If an answer is categorical, give the level of measurement. If an answer is numerical, is it discrete or continuous?

a) Have you visited this store before?

b) How would you rate the level of customer service you received today on a scale from 1 (very poor) to 5 (very good)?

c) How much money did you spend in the store today?

9. A questionnaire was distributed at a large university to find out the level of student satisfaction with various activities and services. For example, concerning parking availability, students were asked to indicate their level of satisfaction on a scale from 1 (very dissatisfied) to 5 (very satisfied). Is a student's response to this question numerical or categorical? If numerical, is it discrete or continuous? If categorical, give the level of measurement.

10. Faculty at one university were asked a series of questions in a recent survey. State the type of data for each question.

a) Indicate your level of satisfaction with your teaching load (very satisfied, moderately satisfied, neutral, moderately dissatisfied, or very dissatisfied).

b) How many of your research articles were published in refereed journals during the last 5 years?

c) Did you attend the last university faculty meeting?

d) Do you think that the teaching evaluation process needs to be revised?

11. A number of questions were posed to a random sample of visitors to a London tourist information center. For each question below, describe the type of data obtained.

a) Are you staying overnight in London?

b) How many times have you visited London previously?

c) Which of the following attractions have you visited?

Tower of London

Buckingham Palace

Big Ben

Covent Garden

Westminster Abbey

d) How likely are you to visit London again in the next 12 months:

(1) unlikely,

(2) likely,

(3) very likely?

12. Residents in one housing development were asked a series of questions by their homeowners' association. Identify the type of data for each question.

a) Did you play golf during the last month on the development's new golf course?

b) How many times have you eaten at the country club restaurant during the last month?

c) Do you own a camper?

d) Rate the new security system for the development (very good, good, poor, or very poor).

13. Given the ten best new hotels to stay in, in the world.

Hot list ranking	Name of property	Country	Room rate	Number of rooms
1	Amangalla, Galle	Sri Lanka	US$574	30
2	Amanwella, Tangalle	Sri Lanka	US$275	30
3	Bairro Alto Hotel, Lisbon	Portugal	€180	55
4	Basico, Playa Del Carmen	Mexico	US$166	15
5	Beit Al Mamlouka	Syria	£75	8
6	Brown's Hotel, London	England	£347	117
7	Byblos Art Hotel Villa Amista, Verona	Italy	€270	60
8	Cavas Wine Lodge, Mendoza	Argentina	US$375	14
9	Convento Do Espinheiro Heritage Hotel & Spa, Evora	Portugal	€213	59
10	Cosmopolitan, Toronto	Canada	£150	97

a) How many elements are in this data set?

b) How many variables are in this data set?

14. The following table lists the marital status of the players of a soccer team.

Status	**Number of players**
Single	9
Married	7
Divorced	5
Widowed	1

What is the relative frequency of the married players?

15. Below are the Playoffs statistics for the 2008-2009 season for a basketball player.

Month	Jan	Feb	Mar	Apr	October	Nov	Dec
Total number of minutes played	659	281	583	118	75	517	602

a) What is the total number of minutes played during the season?

b) What is the relative frequency of the time played in March?

c) What is the cumulative frequency of the time played up until March?

16. The table below displays the number of customers at a shop during each day for a 1-week period.

Day	S	M	T	W	T	F	S
Number of buyers	81	53	47	62	71	67	97

a) What is the relative frequency of the number of customers on Wednesday (W)?

b) What is the cumulative frequency of the number of customers on Wednesday?

17. Find the missing entries of the table given below.

Item	**Frequency**	**Cumulative frequency**
1	9	
2	11	
3		41

18. The ages of 50 teenagers are listed below.

14 13 14 13 16 17 15 15 13 18

17 14 17 14 14 13 18 15 14 13

15 14 14 15 17 14 14 18 14 13

16 15 16 14 15 15 16 13 15 17

15 14 14 15 18 16 16 14 16 16

a) Give the frequency of the fifteen years old teenagers.

b) If the ages are arranged in a frequency table in intervals of length 2, what would be the frequency of the interval 13 – 14?

19. The amount of acidity (pH level) at different wells near a forest is measured. The collected data is listed below.

6.6 5.4 7.7 5.8 6.9 6.3   6.7 7.2 5.3 6.5 6.4 6.1 7.2

7.6 5.9 5.7 5.6 6.0 6.3   5.9 6.4 7.2 7.0 6.8 6.4 6.3

5.7 6.1 6.2 6.4 6.8 6.4   6.2 6.3 6.1 6.3 7.1 7.4 7.5

a) Choose convenient intervals to split the data into five groups.

b) Represent the data using a frequency table. Include the relative and cumulative frequencies.

20. A researcher wants to split the following data into four intervals of equal length:

0.56 0.61 0.55 0.44 0.51 0.43 0.46 0.14 0.82 0.41

0.65 0.41 0.73 0.78 0.41 0.35 0.98 0.23 0.69 0.77

What is the length of each interval?

21. A survey was conducted on the fitness characteristics of male high school football players. The data collected follows.

Note that 1 REP MAX bench press means maximum weight in bench press in one repetition. Propose ways of representing the data in a more compact form with respect to each characteristic.

Name	Age	Position	Height (in cm)	1 REP MAX bench press (in kg)	Skills
John M.	15	quarterback	163	80	high
Jacob	15	linebacker	166	85	medium
Michael R.	16	running back	167	90	medium
Ethan	16	defensive back	171	100	low

Joshua	16	tight end	175	105	high
Tony	17	wide receiver	176	95	low
Anthony	15	wide receiver	175	100	medium
William	17	linemen	168	87,5	medium
Christopher	16	tight end	168	82,5	low
Matthew	16	defensive back	178	100	medium
John D.	16	defensive back	174	107,5	medium
David	15	linebacker	173	80	low
Alexander	16	wide receiver	172	97,5	high
Michael J.	16	linemen	174	100	high
Jerry	17	running back	169	95	medium

TASKS FOR INDEPENDENT WORK

1. Classify data as discrete or continuous

Classify the following data types as discrete or continuous.

1. Amount of coffee dispensed by a coffee machine
2. Number of students in different sections
3. Number of doctors in different towns
4. Yearly average rainfall in London
5. Number of times a darts amateur needs to hit bull’s eye

2. Arrange discrete data in a frequency table

Arrange the following data items in a table showing the frequencies, the relative frequencies, and the cumulative frequencies.

22, 25, 25, 22, 26, 22, 22, 26

25, 20, 23, 24, 21, 25, 23, 25

25, 25, 22, 21, 23, 20, 24, 23

26, 22, 21, 20, 27, 23, 21, 26

3. Arrange continuous data in a frequency table

Below are the top times, in minutes, of the America's Finest City Half Marathon from 1980 to 2009.

74.75 77.03 76.92 74.52 75.10 77.62 76.37 75.28

71.52 73.68 73.00 74.30 76.38 76.28 74.73 72.87

73.18 75.07 74.48 74.03 70.73 70.62 72.57 74.40

75.02 73.60 74.20 72.82 72.78 70.82

Arrange the data in a frequency table using five intervals of equal lengths.

Top times of America's Finest City Half Marathon

Total

4. Read a frequency table

The ratio of sugar level in the blood for 180 randomly selected males whose ages range from 50 years to 60 years are listed in the table below.

Class	90-130	130-170	170-210	210-250	250-290	290-330
Frequency	37	82	31	15	12	3

The sugar level, measured in mg per liter, is considered to be normal if it falls in the interval [90-130). Using the above data,

1. what is the percentage of those whose sugar level is normal?
2. what is the percentage of those whose sugar level is greater than 170 mg/L?

2. CENTRAL TENDENCIES: MEAN, MEDIAN, AND MODE

1. What is the mean score for the nine scores listed below?

75, 83, 81, 76, 83, 90, 77, 77, 69

2. The average monthly wind speed (in miles per hour) for five cities in Canada are given below.

16.1, 15.1, 14.1, 15.4, 13.8

What is the monthly mean of these wind speeds?

3. The mean of a set of numbers is 34. The sum of the numbers is 374. How many numbers are in the set?

4. Catherine's math test scores are 79, 83, 75, 89, and 87.

a) What is Catherine's mean score?

b) What score must Catherine get on the next math test so that her mean test score becomes 85?

5. At a seaside resort there was an average of four hours of sunshine per day for six days of one week. On the seventh day of the week, there was 7.5 hours of sunshine. What was the daily average number of hours of sunshine for the whole week?

6. The mean of a set of 15 consecutive integers is 36.

a) What is the greatest of these integers?

b) What is the smallest integer?

7. The sales and prices of mobile phones sold at a major electronics store over a 1-month period are shown in the table below.

Price of mobile phone ($)	210	240	350	440	560	640	680	700
Number of units sold	6	8	21	46	38	10	5	1

Find the average price of a mobile phone sold at this store.

8. Find the average number of customers per month at a service center using the data below.

Month	Number of visitors	Cumulative Frequency
Jan	35	5
Feb	45	12
Mar	52	16
Apr	64	18
May	68	20
Jun	73	24

9. A forester measured the diameters of saplings prepared for planting. The data collected follows.

Diameter (in inch)	Cumulative relative frequency
[1.2, 1.6)	8%
[1.6, 2.0)	23%
[2.0, 2.4)	45%
[2.4, 2.8)	58%
[2.8, 3.2)	76%
[3.2, 3.6)	100%

What is the mean diameter of the trees?

10. What is the median height of the ten children whose heights (in cm) are listed below?

165, 168, 164, 170, 164, 165, 166, 161, 166, 173

11. The annual thunderstorm days in central Florida for eight consecutive years are

112, 98, 93, 101, 103, 96, 94, 113

Find the median of this data.

12. The table shows the numbers of bread rolls sold by a bakery to 30 of its customers on a particular day.

Number of bread rolls	2	3	4	5	6	7
Number of customers	7	6	7	6	3	1

a) What is the median number of bread rolls sold to a customer on that day?

b) Describe what this parameter tells you.

13. Consider the results of the long jumps competition for 26 athletes:

Length, in m	Frequency
[5, 5.5)	3
[5.5, 6)	5
[6, 6.5)	4
[6.5, 7)	6
[7, 7.5)	4
[7.5, 8)	3
[8, 8.5)	1

a) What are the mean and median lengths of the jumps?

b) Which of the values computed in part (a) best describes the results of the competition? Explain.

14. Find the mode of the following numbers.

3.1, 3.3, –3.1, –5.5, –6.6, 7.1, –4.1, –5.5, 6.2, –5.5

15. The daily high temperatures (in °C) for one month in Chicago are listed below.

26 22 23 23 24 25 23 24 26 22 24 25 26 23 21
25 25 23 22 24 21 22 23 19 15 18 21 23 24 25

a) Determine the mean, median, and the mode of this data.

b) Interpret the statistics found in part a).

16. In each of the following cases, write a set of six numbers that has

a) a mean of 15 and a median of 14.

b) a median of 8 and a mode of 18.

c) a mean of 12 and a mode of 5.

17. A coach wants to choose between Randy and Robert for an important basketball game. So far this year, Randy has played 10 games and Robert played 9. The number of points the two players scored in the games they played are listed below.

Player	**Number of Points**
Randy	16, 22, 34, 20, 11, 30, 18, 22, 18, 29
Robert	12, 16, 28, 20, 14, 17, 22, 16, 17

a) Find the mean number of points scored by each player.

b) Find the median number of points scored by each player.

c) Which of the two players is favored to participate in the next game? Explain.

18.A company tested 10 light bulbs from each of two types for durability. The results, in hours, are listed in the table below.

Type I	890, 880, 800, 950, 20, 900, 880, 860, 830, 850
Type II	840, 820, 880, 860, 790, 750, 780, 780, 750, 710

a) What is the mean life for the 10 light bulbs tested from each type?

b) What is the median life for each of the types?

c) Which type do you think is better? Justify your choice.

19. Ten females and ten males were requested to answer the following question:

"How much are you willing to pay for a meal at the restaurant?"

The answers were as follows:

Sample 1 (males)	15, 50, 35, 10, 20, 25, 25, 20, 15, 25
Sample 2 (females)	15, 40, 35, 40, 50, 45, 15, 20, 15, 25

a) Find the mean, the median, and the mode for each sample.

b) Compare the two samples.

20.A truck carries *x* stones each weighing 70 kg, and *y* stones each weighing 85 kg. What is the ratio of *x* to *y* if the mean weight of the stones in the truck is 80 kg?

21.A block has 21 houses on it. The prices of 20 of the houses are shown in the table below.

Price ($1,000)	**Frequency**
88	3
86	2

84	1
77	2
72	2
68	4
63	3
55	2
52	1

a) Find the mean and median prices of the 20 houses listed in the table.

b) Given that the median price of the 21 houses is the same as the one found in part a) for the 20 houses, find the missing price.

22. A statistician was asked to analyze the frequency of library visits per capita for the 50 states and Puerto Rico. The statistician divided the number of library visits by the unduplicated population of the legal service area and obtained the following data.

Library visits per capita										
6.9	6.3	6.1	6.1	5.9	5.9	5.8	5.7	5.6	5.6	5.6
5.3	5.2	5.1	5	5	5	4.9	4.8	4.8	4.8	4.7
4.7	4.4	4.4	4.3	4.2	4.2	4.1	4.1	4	4	3.9
3.9	3.9	3.9	3.7	3.7	3.7	3.6	3.4	3.4	3.3	3.3
3.2	3	3	3	2.7	2.4	2.3				

a) Group the visits per capita in a table using five intervals.

b) Explain what information, in addition to that in the table is needed for finding the mean visits per capita for all the states.

23. Below is the number of elements in four sets of data and the mean of each set.

Number of elements	6	5	5	7
Mean	13.5	–10	5	40

Find the mean if all the data were combined in one set.

24. An NHL player has the following records.

Season	Games	Mean goals
2005-2006	79	0.139
2006-2007	59	0.169
2007-2008	73	0.383
2008-2009	79	0.380
2009-2010	82	0.378

Find the mean of the player over these five seasons.

25.A random sample of 5 weeks showed that a cruise agency received the following number of weekly specials to the Caribbean:

20 73 75 80 82

a) Compute the mean, median, and mode.

b) Which measure of central tendency best describes the data?

26. Ten economists were asked to predict the percentage growth in the Consumer Price Index over the next year. Their forecasts were as follows:

3.6 3.1 3.9 3.7 3.5

3.7 3.4 3.0 3.7 3.4

a) Compute the sample mean.

b) Compute the sample median.

c) Find the mode.

27.A department-store chain randomly sampled 10 stores in a state. After a review of sales records, it was found that, compared with the same period last year, the following percentage increases in dollar sales had been achieved over the Christmas period this year:

10.2 3.1 5.9 7.0 3.7

2.9 6.8 7.3 8.2 4.3

TASKS FOR INDEPENDENT WORK

1. Direct application on the definition of the mean

Find the mean of the following sets of data.

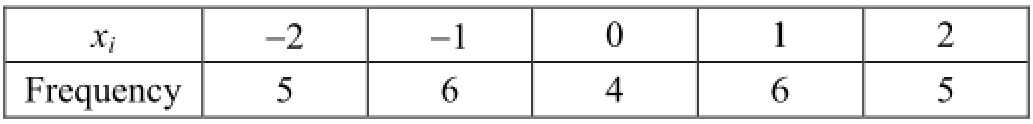
0, 0, 0, 0, 1, 1, 1, 1

2. Indirect application on the definition of the mean

What is the mean of 10 grades if the mean of 7 of them is 12 and the mean of the remaining 3 is 40?

3. Direct application on the formula of the mean for grouped data

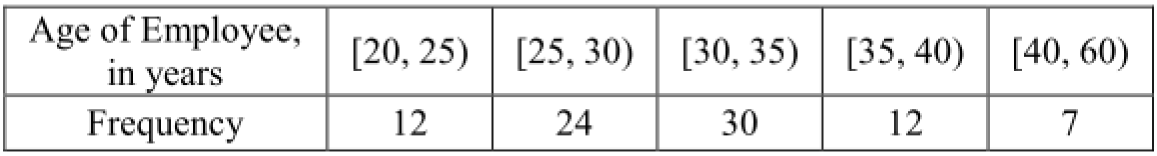
Find the mean of the data set given below.



x_i	−2	−1	0	1	2
Frequency	5	6	4	6	5

4. Find an estimate of the mean of data grouped in classes

Find an estimate of the mean of the data set given in the following table.



Age of Employee, in years	[20, 25)	[25, 30)	[30, 35)	[35, 40)	[40, 60)
Frequency	12	24	30	12	7

5. Find the median of a set of data with odd number of points

Find the median of the following set of data

5, 7, 8, 8, 1, 4, 9

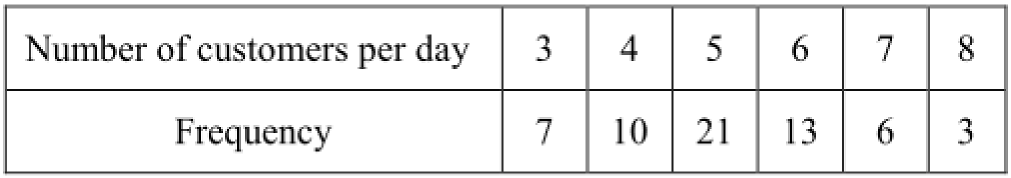
6. Find the median of a set of data with even number of points

Find the median of the following set of data

−6, 8, 4, 16, −2, 7

7. Find the median of a set of grouped data

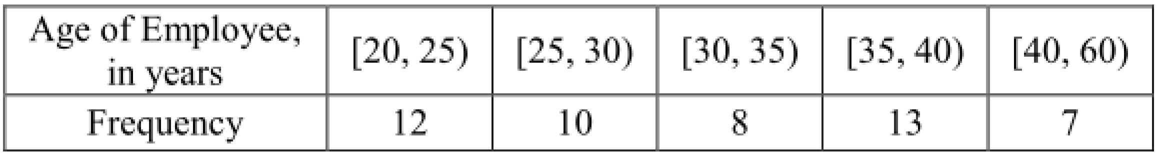
At a barber shop, the number of customers Vito received per day, over a 60-day period, is summarized in the table below. Find the median of the set of data.



Number of customers per day	3	4	5	6	7	8
Frequency	7	10	21	13	6	3

8. Identify the median class of a set of data grouped in intervals

Find the median class of the following data set.



Age of Employee, in years	[20, 25)	[25, 30)	[30, 35)	[35, 40)	[40, 60)
Frequency	12	10	8	13	7

9. Find the mode of a set of discrete data

Find the mode of each of the following sets of data

5, 7, 8, 8, 1, 4, 9, 0, 8, 4, 5

1, 2, 1, 2, 1, 2, 1, 2, 1, 4

1, 1, 1, 1, 1, 1, 1, 1, 1

3. MEASURES OF VARIABILITY

Quartiles and Percentiles

1. Twelve adults were asked about the number of hours they spent watching TV on the previous week. The results are listed below.

14, 16, 20, 21, 24, 10, 15, 16, 18, 15, 14, 14

Find the upper quartile of the data.

2. Find the lower quartile score in the table below.

Score	64	68	71	75	78	85	93
Students	3	2	3	2	4	2	1

3. The results of the midterm test for a math course are recorded in the table below.

Interval	**Cumulative relative frequency**
[25, 50)	15%
[50, 75)	52%
[75, 100]	100%

a) Find the first and third quartiles.

b) Interpret the data in terms of the values found in part a).

c) Among the mean and the median, which do you recommend to describe the data, and why?

4. John recorded the number of sick days he had each month of the previous year.

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

a) Determine the 60th percentile of the data.

b) Compare this percentile with the median.

5. The scores of 10 students on an exam were as follows.

Name	Denzel	Abigail	Bernard	Julia	Ariel	Judith	Joel	Michaela	Fabian	Ann
Grade	84	79	93	86	73	71	85	68	73	82

Check if each of the following is correct.

a) Joel is in the top 50th percentile. (Joel's grade is more than the 50th percentile.)

b) Julia is not in the top 85th percentile. (Julia's grade is less than the 85th percentile.)

c) Denzel's grade is between the 50th percentile and the 75th percentile.

d) Judith is in the top 75th percentile. (Judith's grade is more than or equal to the 75th percentile.)

e) Michaela is in the top 65th percentile. (Michaela's grade is more than or equal to the 65th percentile.)

6. Nine students took a math test. Judie scored 73, which placed her in the 80th percentile. Due to a misprint in the last question, six points were later added to the grade of each student.

a) In what percentile is Judie now? Explain.

b) What is the new range if the old range was 21?

7. Bernard's and Joel's ages, among ages of members of a club they joined, are the 40th percentile and the 80th percentile, respectively.

Which of the following statements **must** be true?

a) Bernard's and Joel's ages are higher than the average.

b) Bernard's age is half of Joel's age.

c) Bernard is younger than Joel.

d) Joel is older than 80 percent of the club members.

e) Joel is 40 years older than Bernard.

8. The frequency distribution of the heights of 100 men is given in the table below.

Height, in cm	**Frequency**
[150 – 156)	12
[156 – 162)	14
[162 – 168)	18
[168 – 174)	23
[174 – 180)	14

[180 – 186)	9
[186 – 192)	7
[192 – 198)	3

a) Find the 80th percentile of the heights.

b) What percentile does 182 cm represent?

9. Below are the exam scores of 50 students.

68	69	70	71	73	73	73	73	74	75
75	75	75	75	75	76	76	76	76	76
76	77	77	77	77	78	78	79	79	79
79	81	81	81	82	84	84	85	86	87
88	88	89	89	91	91	91	92	92	92

a) Determine the first and third quartiles.

b) What do these values tell about the scores?

c) The term "trimmed mean" is referred to the mean of the middle 90% of the data. That is, the mean after removing the scores that are less than the 5th percentile and those greater than the 95th percentile. Find the trimmed mean of the scores.

10. The weights of washing powder in a sample of 100 packets were measured. The results are listed in the table below.

Weight (w) of washing powder in ounces	**Number of packets**
$28 \leq w < 29$	10
$29 \leq w < 30$	18
$30 \leq w < 31$	20
$31 \leq w < 32$	30
$32 \leq w < 33$	12
$33 \leq w < 34$	8
$34 \leq w < 35$	2

a) Find the cumulative relative frequencies.

b) Calculate the 40th and 60th percentiles. Interpret the quantities of washing powder in the sample in terms of these values.

Measure of Dispersion

1. Fourteen students were given a statistics test. The time (in minutes) to complete the test by each of the students are given below. Find the range of these times.

14, 12, 9, 10, 12, 11, 8, 15, 9, 7, 8, 6, 12, 10

2. The monthly average temperatures (in Fahrenheit) in Arnett, Oklahoma, last year, were as follows: 35.5, 42.6, 47.7, 54.8, 66.4, 70.1, 78.2, 82.0, 72.7, 59.9, 39.7, 28.4

a) Find the inter-quartile range of these temperatures.

b) Interpret the values of the characteristic obtained in part a).

3. To investigate the ages of newly graduating lawyers, a sample of 18 lawyers is chosen. The ages of these lawyers are as follows:

23, 24, 24, 24, 25, 25, 26, 26, 25, 24, 27, 23, 28, 26, 27, 24, 23, 24

Calculate the mean absolute deviation (MAD).

4. Find the variance of the following data set.

12, 13, 10, 11, 14, 15, 17, 21, 18

5. The number of books read by each of the 20 members in a reading club during the months of June, July, and August is summarized below.

Number of books	5	6	7	8	9	22
Members	4	6	4	2	3	1

a) Find the variance and the standard deviation of the number of books read.

b) Do the mean and the standard deviation present a good summary for the data?

6. A car race was completed by fifteen participants, and their race durations (in seconds) are listed below.

53.2, 54.5, 52.9, 53.9, 55.6, 54.1, 52.3, 51.8, 50.9, 51.7, 53.6, 55.2, 53.1, 55.4, 54.3

a) What are the minimum and the maximum durations?

b) What is the range of these durations?

7. Find the inter-quartile range of the 35 data items listed below.

–13	–10	144	–6	15	18	19	–5	155	21
34	28	24	21	27	29	166	115	101	10

34 54 23 41 23 46 37 39 81 73

65 34 31 58 –21

8. The final averages of 19 students are summarized in the following table.

Marks	[40 – 50)	[50 – 60)	[60 – 70)	[70 – 80)	[80 – 90)	[90 – 100]
Frequency	2	4	3	5	3	2

a) Find the mean and median of the grades.

b) Find the inter-quartile range and the standard deviation of the grades.

c) What information about the distribution of the grades can be deduced from the values found in parts a) and b)?

9. The frequency distribution of the heights of 30 men is given in the table below.

Height in cm	**Frequency**
[156 – 162)	7
[162 – 168)	8
[168 – 174)	5
[174 – 180)	5
[180 – 186)	3
[186 – 192)	1
[192 – 198)	1

a) Find the inter-quartile range, the variance, and the standard deviation of the heights.

b) What is the unit of measurement for the inter-quartile range, for the variance, and for the standard deviation?

10. Todd and Jack work at an auto engine plant. The table below gives information about the number of car components they made during the last 10 working days.

Todd	3	5	2	3	6	2	3	7	3	8
Jack	1	8	2	3	7	9	1	3	1	7

a) Calculate the mean and the standard deviation for each worker.

b) Which measure gives a good indication on which of the two workers is more consistent?

c) Make a conclusion about the results. Which of the two workers is more productive?

11. Which of the following data sets has the largest variance? Which has the smallest?

a) 4, 5, 7, 8, 9, 0, 1, 4

b) 3, 5, 7, 8, 1, 9, 1, 3

c) 2, 4, 6, 8, 3, 1, 2, 2

12. Consider the following set of data.

2.3, 2.6, 2.1, 4.3, 2.8, 1.8, 3.5, 4.3, 0.8, 1.6

What number, if removed from the data set, will

a) increase the standard deviation the most?

b) decrease the standard deviation the most?

13. At a car rental shop, data is collected on the number of rentals made during twenty consecutive days. The result follows.

21, 22, 20, 15, 13, 16, 29, 32, 20, 15, 16, 18, 22, 23, 21, 19, 20, 21, 24, 21

a) Find the inter-quartile range and the standard deviation for the number of rentals.

b) Compare these characteristics.

14. Below is the number of working hours per month for three freelancer employees in the year 2010.

John	94	61	62	61	60	58	60	94	55	57	53	92
Janet	71	63	84	75	80	67	61	78	82	73	66	81
David	88	52	93	51	78	81	83	91	50	79	82	77

a) Calculate the range, the IQR, the variance, and the standard deviation for each employee.

b) Which of the measures gives better indication for the spread of the data?

15. The owner of two restaurants is interested in how much people spend at his restaurants. The owner takes a sample of thirty receipts from each restaurant and calculates their means. He finds the means to be \$43 and \$41 with standard deviations of \$15 and \$17, respectively.

a) What information can be extracted from the given data?

b) Is there enough information to find the mean and the standard deviation of the 60 receipts combined together? If yes, find these values; otherwise explain why it cannot be found.

TASKS FOR INDEPENDENT WORK

1. Find the quartiles of a set of discrete data

Find the upper and lower quartiles of:

8, 9, 11, 14, 15, 18, 18, 21, 22, 23, 27, 29.

2. Find the quartiles of a set of data given in intervals

The table below summarizes the sick leave days of a company of 55 workers.

Sick Leave Days		**Sick Leave Days**	
Number of days (x_i)	**Frequency (f_i)**	**Number of days (x_i)**	**Frequency (f_i)**
0	10	5	5
1	8	6	7
2	7	7	0
3	8	8	3
4	4	9	3

Find the upper and lower quartiles of the number of days taken as sick leave.

3. Find the range of a set of data

The winning scores of the Master Golf Tournament between the years 1998 and 2010, inclusive, are:

277 281 279 279 286 283 278 276 271 276 277 280 275 280

Find the range of these scores.

4. Find the standard deviation of a sample

Find the variance and the standard deviation of the sample given below.

Number of costumers per day	1	2	3	4	5
Frequency f_i	1	2	5	4	4

4. GRAPHICAL REPRESENTATION OF DATA

1. Which of the following sets data is best represented using a dotplot?

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

b) The masses of 30 randomly selected adults, if measured to the nearest 10 grams

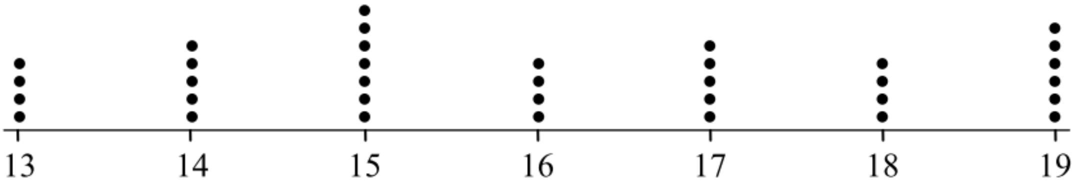
c) The grades of 12 students in a statistics AP exam

d) The number obtained by rolling a fair die 20 times

e) The number of letters per page in a magazine of 40 pages

f) The number of coffee cups consumed by an employee before 11:00 A.M. in a small company of 17 employees

2. A survey is conducted on the ages of teenagers who are buying a newly released PlayStation. The dot plot below displays the result.



Ages of Teenagers Buying a Newly Released PlayStation

a) How many teenagers were surveyed?

b) For what age(s) does the PlayStation appear to be least popular among teenagers?

c) For what age(s) does the PlayStation appear to be most popular among teenagers?

3. Consider the following data set.

44, 42, 38, 45, 26, 33, 41, 43, 48, 46, 31, 27, 39, 51, 53, 30

a) Which is better for representing the data, a dot plot or a stem-and-leaf plot? Why?

b) What can be used for a stem in a stem-and-leaf plot for a data consisting of numbers between 0 and 1 given to two decimal places?

4. Consider the data given in the stem-and-leaf plot below.

Stem	Leaf
5	1 3 7 $\underline{9}$
6	$\underline{0}$ 4 6 8 9
7	2 3 6
10	1 $\underline{4}$

Find the values that the underlined items represent if the key is

a) 5|3 means 530

b) 5|3 means 5 + 3

c) 5|3 means 5.3

5. Two hundred students are divided among 6 clubs as follows: 40 are in the computer club, 32 are in the art club, 10 are in the folk dance club, 26 are in the chess club, 72 are in the swimming club, and 20 are in the music club.

a) Use a pie chart to illustrate the data.

b) How many students must move from the computer club to the art club so that both clubs have the same number of students?

6. A questionnaire completed by 200 people about their favorite European holiday destination. People surveyed were asked to choose from the following: France, Germany, Spain, Portugal, and Italy. Fifty chose Germany, 35 chose Spain, 25 chose Portugal, 30 chose Italy, and the rest chose France.

a) How many people chose France?

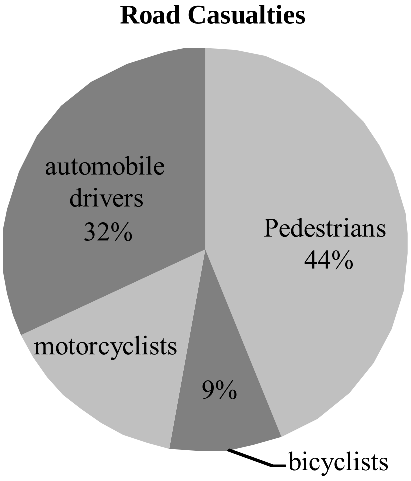
b) Draw a bar chart to illustrate this information.

7. The weapons used in 50 different crimes are given below.

Type of weapon	**Number of crimes**
Handgun	25
Knife	10
Shotgun	5
Personal weapon	7
Other	3

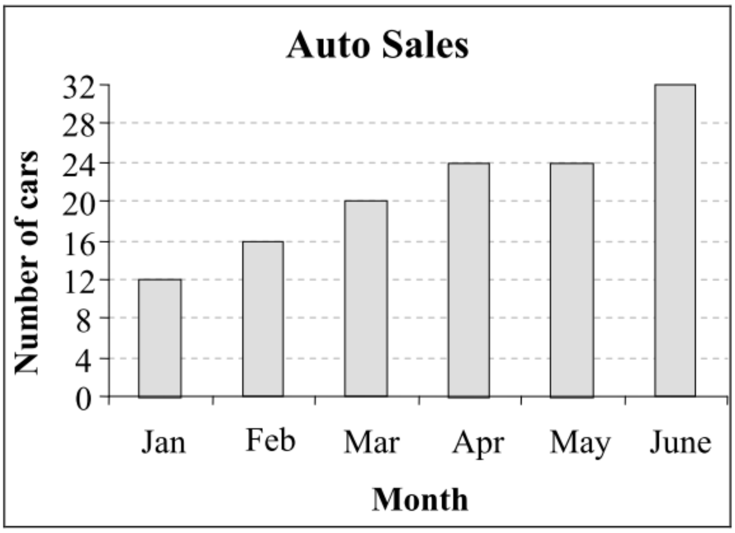
Fill the last column of the table with the angles needed to draw a pie chart, then represent the data using a pie chart.

8. Casualties on the roads are classified into 4 main categories: Pedestrians, bicyclists, motorcyclists and automobile drivers. A statistics made for the year 2018 is represented by the pie chart below.



If the frequency of the pedestrians is 2370, find the angles of the 4 sectors and the frequencies of the three other classifications.

9. Consider the following bar graph.



a) What does the bar graph represent?

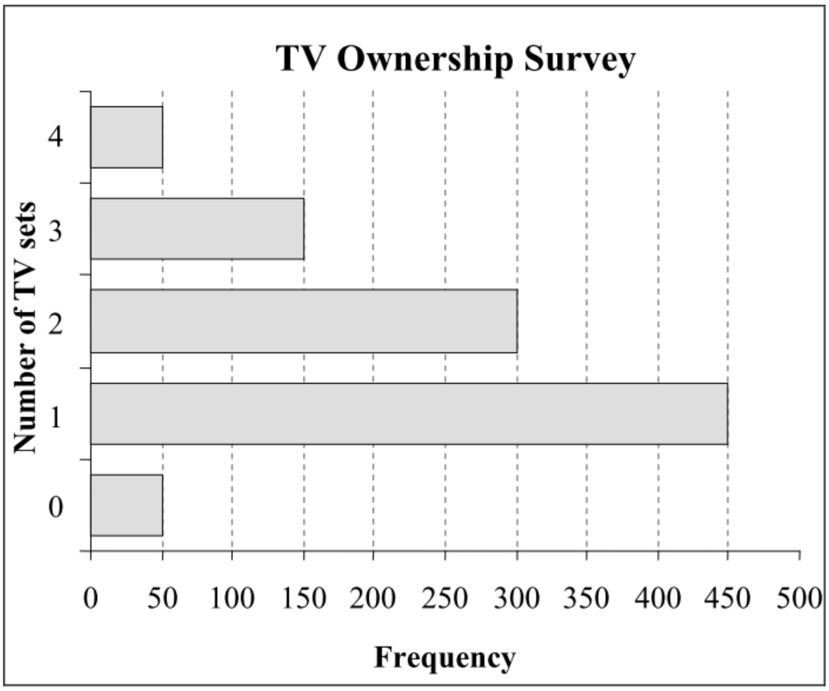
b) What is the total number of cars sold during the 6-month period?

c) Which month shows the most increase in the number of cars sold when compared to the previous month?

d) Which month shows no change in the number of cars sold when compared to the previous month?

e) Which month shows the most increase in the percentage of the number of cars sold when compared to the previous month?

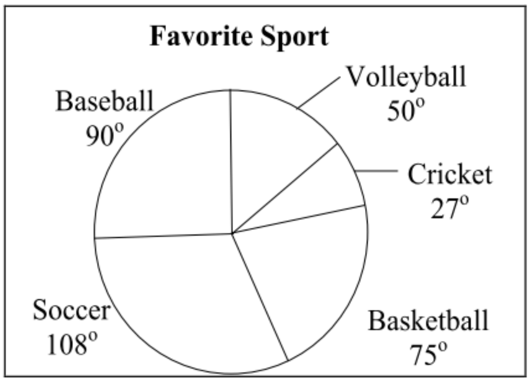
10. A survey was conducted on 1,000 households checking the number of TV sets each household has. The result is graphed below.



a) What is the percentage of houses with three or more TV sets?

b) What is the ratio of the households with one TV set to those with more than one TV set?

11. The children in a school completed a questionnaire in which they had to state one favorite sport. The result is illustrated in the pie chart to the right. If the number of students who prefer soccer is 36 find,



a) How many children were surveyed?

b) How many children prefer baseball?

12. The following data represents the age of children visited a zoological museum on one day.

4	3	5	5	7	4	6	8	11	12
10	11	8	7	3	4	6	3	5	4
3	5	6	4	3	6	3	8	7	9
3	11	3	4	4	5	3	7	4	6
4	5	7	8	3	9	3	10	8	5
9	7	3	4	5	4	7	7	9	4

a) Group the data into intervals of size 2.

b) Display that data using a histogram.

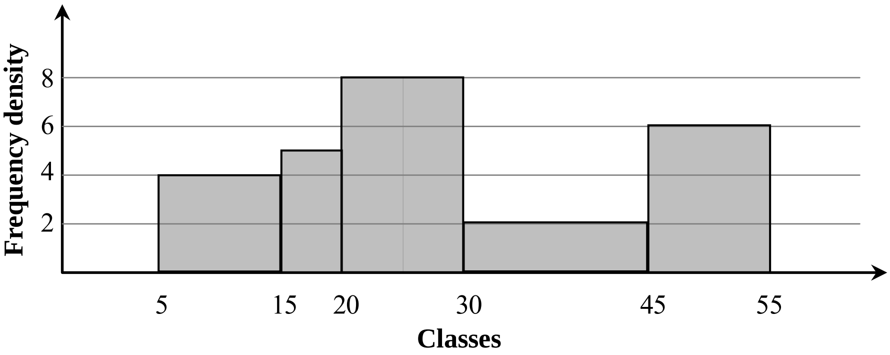
c) Find the mean age of the visitors.

13. Water levels in a natural pond, measure in centimeters relative to a given mark, across one year are given in the table below.

Water level	[−90, −30)	[−30, −10)	[−10, 10)	[10, 20)	[20, 100)
Number of days	60	112	103	41	49

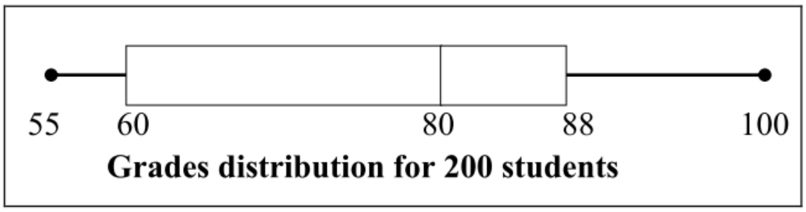
Find the frequency density of each interval and picture the data using a histogram.

14.A data set is represented by the histogram below.



Construct a frequency table.

15. Refer to the boxplot shown below to answer the following questions.



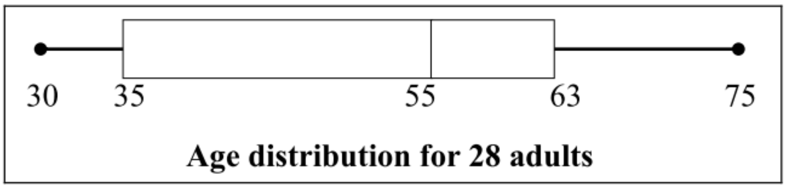
a) How many students scored between 60 and 80?

b) How many students scored between 80 and 88?

c) How many students scored more than 80?

d) If we omit the students who scored less than 60 and more than 88, what would be the median grade of the remaining grades?

16. Refer to the boxplot below to answer the questions.

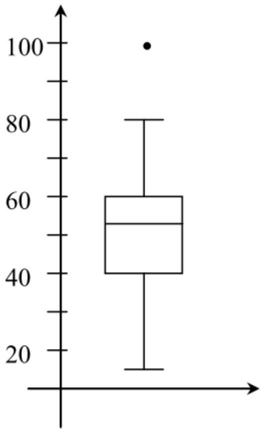


a) Approximately, how many adults are 35 or older?

b) What is inter-quartile range for the data?

c) Does the data contain any outliers?

17. Refer to the modified boxplot shown below to answer the questions that follows.



a) What is the five-number summary for the data?

b) Are there outliers. Explain?

18.A survey is conducted on the ages of unemployed in a city. The data collected from a random sample is shown in the table below.

Age of unemployed	**Frequency**
15 – 24	180
25 – 34	120
35 – 44	160
45 – 54	100
55 – 64	50

a) Draw the cumulative frequency polygonal line.

b) Based on the data collected, what percentage of the unemployed are 55 years or older?

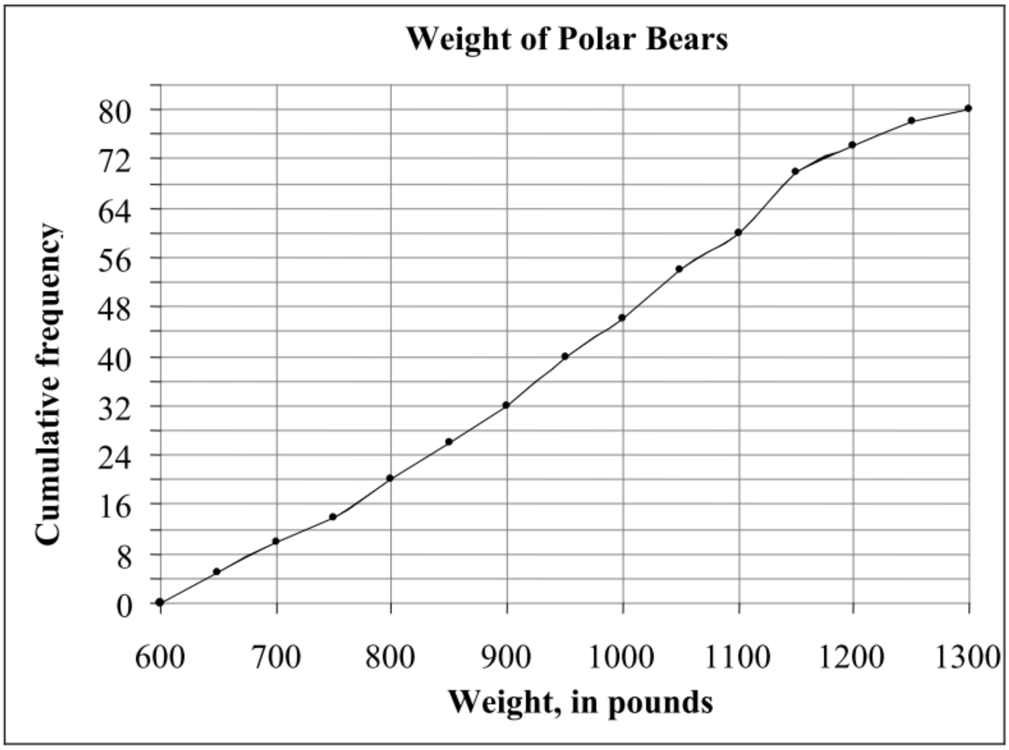
c) In what age group does the 50th percentile lie?

19. The graph below represents the weights of eighty polar bears. Which of the following statements is/are true?

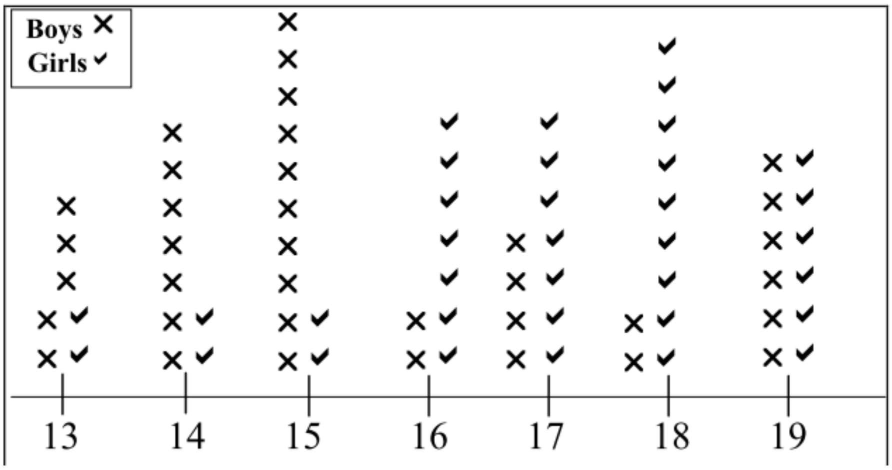
I. None of the bears weigh less than 600 pounds.

II. Only one bear weighs less than 700 pounds.

III. More than ten bears weigh between eight hundred and nine hundred pounds.



20.A survey is conducted on the ages of girls and boys who are buying a newly released PlayStation. The dot plot below graphs the results of the survey.



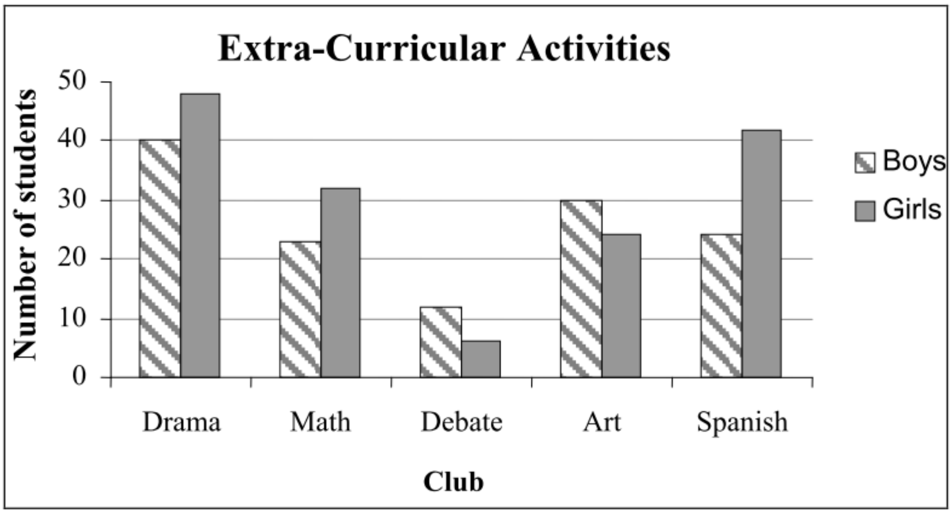
Ages of Teenagers Buying a Newly Released PlayStation

a) For what age(s) does the PlayStation appear to be as popular among boys as among the girls?

b) For what age(s) does the PlayStation appear to be least popular among the boys?

c) For what age(s) does the PlayStation appear to be least popular among the girls?

21. Refer to the bar graph below to answer the next questions.



a) List the clubs that have more girls than boys.

b) Which club has the highest difference between the number of boys and the number of girls? Give an estimate of the difference.

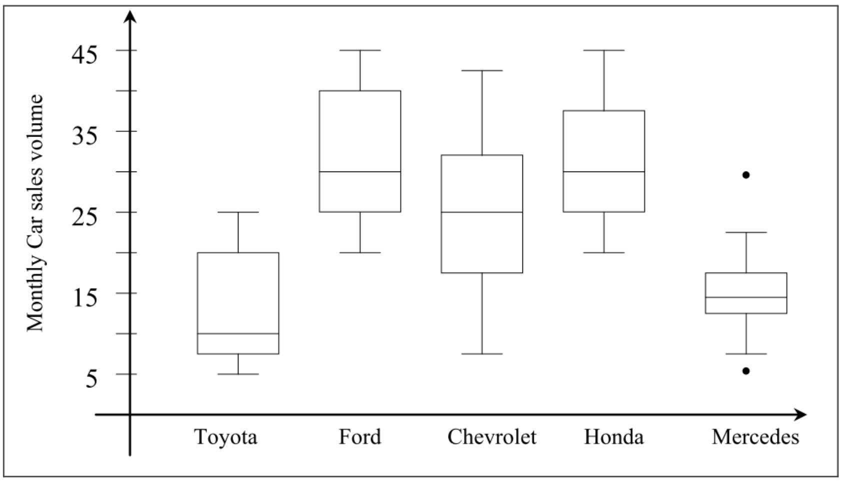
22. The following table deals with student's results on exams:

Robert	77	83	65	70	85	76	72	88	92	84
Diego	61	54	55	59	63	67	93	58	69	65
Rudolf	85	81	82	90	53	92	81	80	93	79
Jack	79	80	81	95	78	86	83	85	80	62
Tim	69	72	77	70	74	67	97	76	87	65

a) Organize the data in a parallel boxplots

b) Write a few sentences comparing the dispersion of exams results for each student.

23. On the parallel boxplots below you can find information about the sales volumes of different car types in a certain automobile sales centre for one year.



With the help of the graph answer the question:

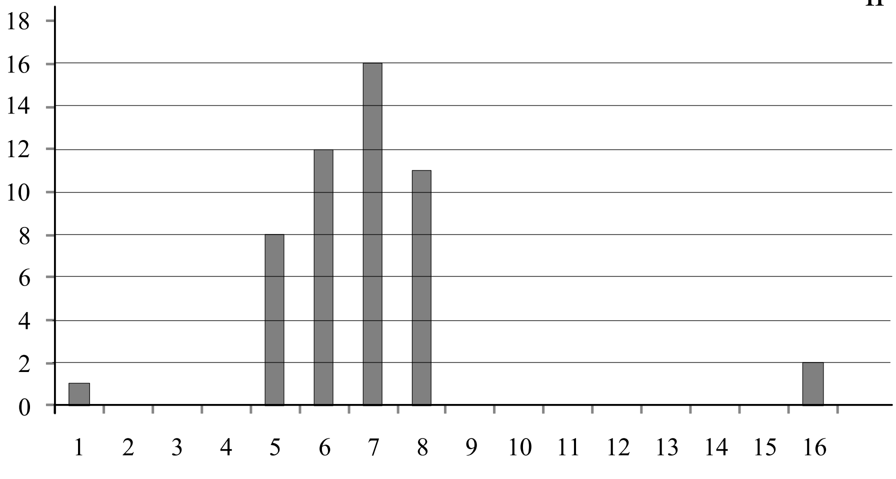
a) Which car has the highest sales rate for the last year?

b) Are there any outliers on the graph?

c) How does the main body of "Toyota" sales volumes locate in comparison with "Ford" and "Honda"?

d) What can you say about the spread and the median of the "Honda" and "Ford" sales volumes? Make conclusions about medians of the car's sales volumes distributions.

24. Identify the outlier(s) of the data described by the graph below.



25. Find the 5-number summary for the data given below and represent them using a modified boxplot.

Items x_i	8	9	10	11	13	16
Frequency f_i	6	6	16	21	20	1

TASKS FOR INDEPENDENT WORK

1. Represent data using a dotplot

Represent the following data set using a dotplot diagram.

16 16 16 14 10 11 8 15 8 11 8 14 16
10 14 8 8 8 14 15 12 15 8 12 8 14

2. Represent data using a stem-and-leaf plot

The average personal income in 20 American states for the year 2019 are given.

24,706	25,057	25,400	28,240	29,923
27,711	30,180	25,020	27,744	29,771
25,579	31,727	28,551	26,183	25,128
29,141	26,982	26,894	29,596	30,001

Represent the data graphically using a stem-and-leaf plot.

3. Represent data using a bar diagram

The grades of 28 students in a business exam are as follows

Grade	A	B	C	D	E	U
Frequency	3	12	6	4	2	1

Represent the data using a bar graph.

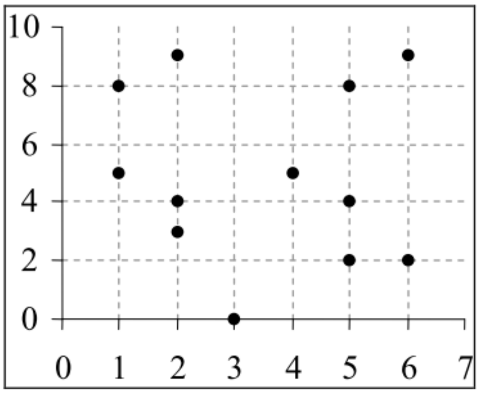
4. Represent data using a box plot

The minimum value, maximum value, lower and upper quartiles, and the median of a set of data are 25, 44, 31, 40, 36 respectively. Represent the data using a boxplot.

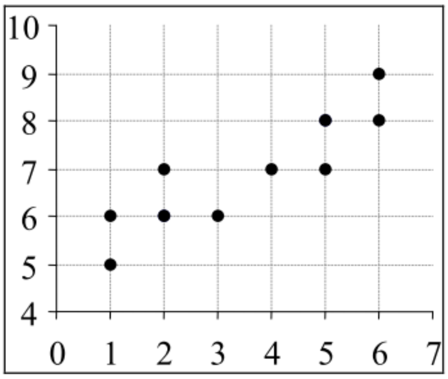
5. SCATTER DIAGRAMS AND REGRESSION LINES

1. State whether each set of data whose scatterplot is given has positive, negative, or no trend.

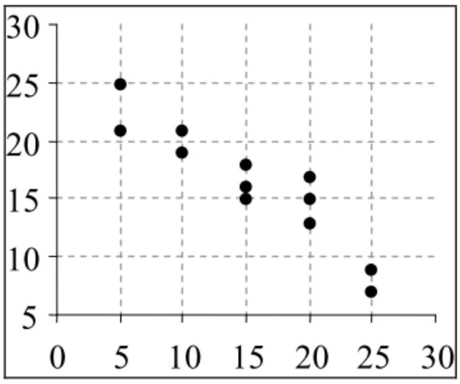
a)



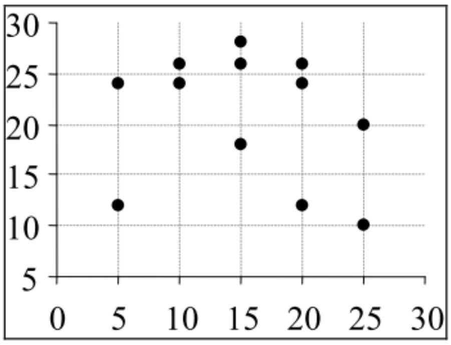
b)



c)



d)



2. On a given month, Jane recorded the amount she spent each day from payday till day 19. The table below shows her records.

Days after payday	0	1	2	3	4	5	6	7	8	9
Amount spent (in dollars)	100	80	80	75	70	70	80	75	70	65

Days after payday	10	11	12	13	14	15	16	17	18	19
Amount spent (in dollars)	65	60	40	60	55	20	0	50	45	40

a) Draw a scatter plot.

b) Does the plot reflect a trend in the data? If so, is it a negative trend or a positive trend?

3. A survey is conducted on employees of a small company to examine the relationship between their salaries and the number of years of education. The following data is collected.

Annual salary (in thousand dollars)	Years of education
24	12
36	14
32	16
28	12
43	16
50	18
31	12
58	20
35	14
41	22
30	20
46	16
32	14
35	14
46	18

a) Represent the data using a scatterplot.

b) Comment on the relationship between the pairs of data.

4. A specialist of the U.S. Department of Labor considers the data of the civilian labor force employment as a percent of civilian non institutional population (Empl) and consumer price indices (CPI) for major expenditure classes. The data he collected are listed below.

Year	CPI	Empl
1990	130.7	66.5
1991	136.2	66.2
1992	140.3	66.4
1993	144.5	66.3
1994	148.2	66.6
1995	152.4	66.6
1996	156.9	66.8
1997	160.5	67.1
1998	163.0	67.1
1999	166.6	67.1
2000	172.2	67.1
2001	177.1	66.8
2002	179.9	66.6
2003	184.0	66.2
2004	188.9	66.0
2005	195.3	66.0
2006	201.6	66.2
2007	207.3	66.0
2008	215.3	66.0
2009	214.5	65.4

Graph the data using a scatterplot. Describe the relation between the variables.

5. An economist analyzes the relation between components of money stock measures: currency and nonbank traveler’s checks. He collected the following data for the last 20 years from government reports.

Currency (in billions of dollars)	Nonbank travelers checks (in billions of dollars)
246.5	7.7
267.1	7.7
292.1	8.2
321.6	8.0
354.5	8.6
372.8	9.0
394.7	8.8
425.4	8.4
460.5	8.5
517.9	8.6
531.2	8.3
581.1	8.0
626.3	7.8
662.5	7.7
697.7	7.5
724.1	7.2
749.6	6.7
759.8	6.3
815.3	5.5
862.1	5.1

a) Which of the variables is the explanatory and which is the response?

b) Make a scatterplot of the data.

c) Are the variables positively or negatively associated?

6. An investor purchased shares of a stock for $30 each. At the end of each month, for 6 months, he noted the value of his shares. The table below gives the value of the share at the end of each month.

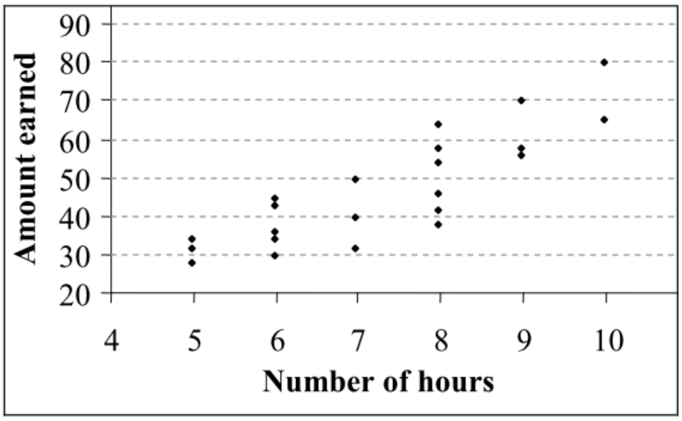
Share Price at End of Each Month (Initial Price = $30)	
End of month one	35
End of month two	39
End of month three	46
End of month four	45
End of month five	53
End of month six	57

a) Draw a scatterplot for the data.

b) Is the trend in the value of the share positive or negative?

c) Based on the trend in the value for the past six months, what would you predict the value to be at the end of month seven?

7. The adjacent graph shows the amount of tips earned by a person and the number of hours worked each day for a period of 22 days.



a) How many hours did this person work when he made the least amount of tips?

b) How many hours did this person work when he made the greatest amount of tips?

c) At a later date, this person works 5 hours on one day and 6 hours on another. If he makes \$45 and \$55 on these two days, can we predict on which day he made which amount? Justify.

d) Is there a positive, negative, or no trend between the amounts of tips and the number of hours?

e) Draw a line that visually best fits the data and use it to estimate the amount of tips this person would make if he works for 12 hours.

8. Consider following set of data.

x	3	7	9	11	14	15	15	21	22	23
y	5	12	5	12	10	17	23	16	10	20

a) Draw a scatter diagram.

b) Find the regression line *y* on *x*, and draw it on the scatter diagram found in part a).

9. For a set of 12 observations, $\sum x = 43.2$, $\sum y = 56.4$. The regression line *y* on *x* passes through the point (1.6, 2.5), and the regression line *x* on *y* passes through the point (4, 6.2). Find the equations of the two lines.

10. If S_{xy} = 3143.4, S_{xx} = 2143.2, S_{yy} = 1900, $\bar{x} = 52.9$, $\bar{y} =$ 64.3, find the regression lines *y* on *x*, and *x* on *y*.

11. For a given set of data, $\overline{x} = 21.3$ and $\overline{y} = 22.9$. The gradient of the regression line *y* on *x* is equal to 2. Find the equation of this regression line.

12. A study of age and systolic blood pressure of eight randomly selected adults resulted in the table below.

Age, *x*	30	33	35	36	37	42	44	49
Blood pressure, *y*	125	128	131	130	137	140	142	140

a) Use a calculator to find the missing entries in the table below.

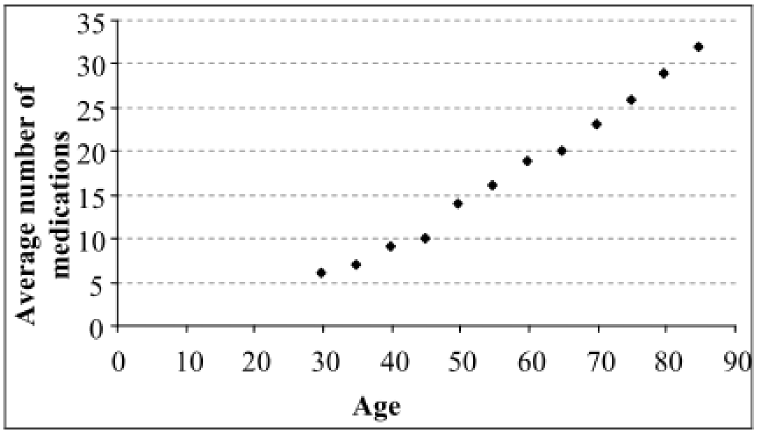
	x	y	x^2	xy
	30	125		
	33	128		
	35	131		
	35	130		
	37	137		
	42	140		
	44	142		
	49	140		
$\sum$				

b) Find $\overline{x}$, $\overline{y}$, Var(*X*), and cov(*X*, *Y*).

c) Find the slope of the regression line that best fits the data and deduce its equation.

d) Based on the data, predict the systolic blood pressure of a person whose age is 50.

13. The graph below represents the average number of medications *Y* taken by people versus their age *X*.



The mean age is $\overline{x} = 57.50$, the mean of the average number of medications is $\overline{y} = 17.58$, Var(*X*) = 325, and cov(*X*, *Y*) = 156.9.

a) Use this information to find the equation of the best-fit line using the least square method.

b) Predict the average number of medications taken by people who are 90 years old.

14. A client asks a real estate agent in Chicago to find a house in Chatham region. He wants to know how the price of a house Y depends on the living space X. The real estate agent has data of houses sold in the region. He finds that the average price is $\overline{y} = 98{,}347$ dollars and the average size of the living space is $\overline{x} = 321$ square meters. He also finds that the variance of the living space is Var(X) = 510 and the covariance of the data is cov(X, Y) = 676,242.

Assume that the relationship between the price and living space is almost linear.

a) Find the slope and the intercept of the regression line.

b) Using the regression line, give an estimate for the price of a house with a living space of 280 square meters?

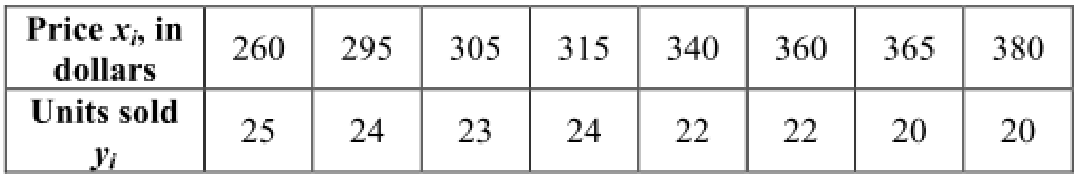
15. Assume that the units to be sold per month of a certain washing machine versus the unit price in dollars can be predicted using the regression line: $y = -0.0425x + 36.42$, where the price x varies from \$240 to \$400.

A store sold 25 washing machines at a unit price of \$300. What is the error in predicting the number of units sold at \$300 using the regression line?

TASKS FOR INDEPENDENT WORK

Find the linear regression line of a set of bivariate data

The same washing machine is being sold at 8 stores, each of which prices the machine differently. The distributor have monitored the number of units sold in one month in these stores and obtained the following data.



Price x_i, in dollars	260	295	305	315	340	360	365	380
Units sold y_i	25	24	23	24	22	22	20	20

a) Find $\overline{x}$, $\overline{y}$, Var(X), and cov(X, Y).

b) Find the linear regression line of y on x.

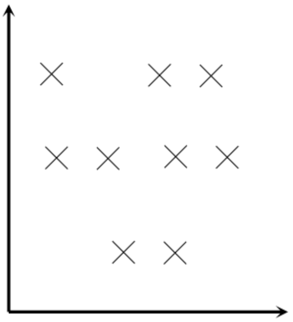
6. LINEAR AND NONLINEAR RELATIONS

1. What kind of linear correlation is indicated if the product-moment correlation coefficient, *r*, is:

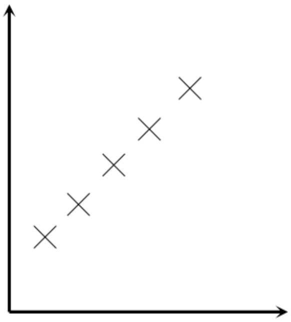
a) 1 b) –1 c) 0

2. Refer to the three scatter diagrams below to answer the question.

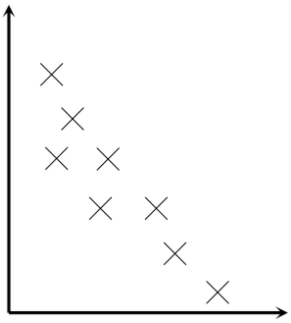
a)



b)



c)



Pick, from the values below, the values of *r* most suitable to match each picture.

{0, 0.3, –1, 0.96, 1, –0.87}

3. If $S_{xy} = 188$, $S_{yy} = 154$, and $S_{xx} = 308$, find the correlation coefficient. Comment on the type of correlation.

4. If $S_{xy} = –412.31$, $S_{yy} = 1800.03$, and $S_{xx} = 414.64$, find the correlation coefficient. Comment on the type of correlation.

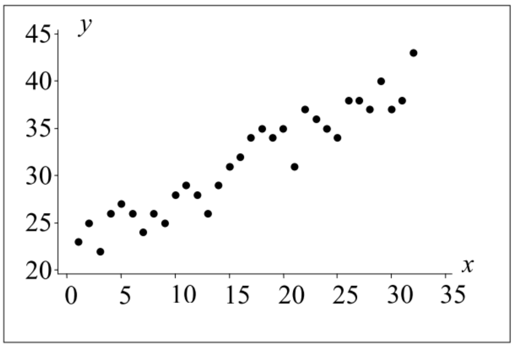
5. What conclusion can be drawn from the following Spearman's coefficient of rank correlation?

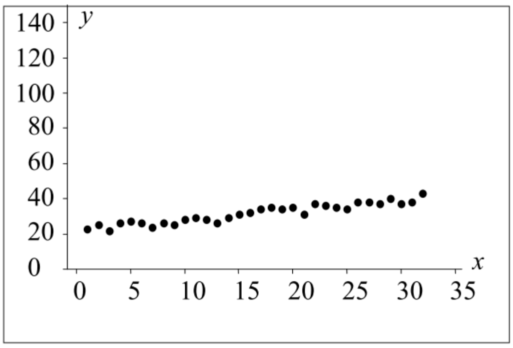
a) $r_s = 1$ b) $r_s = 0.64$ c) $r_s = –1$ d) $r_s = 0$ e) $r_s = –0.14$

6. Find Spearman's coefficient of rank correlation for the following set of data.

x	60	63	62	66	65
y	62	63	62	67	64

7. Consider the following scatterplots.





a) Use the graphs to give an estimate for the correlation coefficient of each graph.

b) Is it possible that both graphs represent the same data? If so, how do you change your answers in part a)?

8. The data in the table is collected for studying the salaries and age for men and women at a certain company.

Salary per hour (in dollars)	Age	Gender
12	25	F
23	34	F
20	22	M
29	29	M
35	40	M
25	42	F
36	37	M
29	31	M
24	33	F
21	27	F
26	35	F
34	37	M
38	42	F
50	45	M
41	51	M
32	48	M
29	24	M

a) Make a scatterplot using different symbols for each group.

b) Calculate the correlation coefficients for men, women, and both. Compare the values obtained.

9. In the United States, the Federal Reserve Bank reports several distinct measures of the aggregate money supply. The narrowest measure, M1, includes only the most liquid assets (assets that are easily exchangeable as payment for goods and services). The following table shows the M1 money supply for the US economy for some years.

Year	1974	1979	1984	1989	1994	1999	2004	2009
M1 (in billions of dollars)	902.1	1473.7	2308.8	3158.4	3496.5	4634.6	6415.2	8524.3

a) Represent the data using a scatterplot.

b) Find the regression line for predicting the logarithm of the M1 money from the date.

c) Construct and interpret a residual plot.

10. A young programmer wrote a small program and placed it on his own website for free download. The program menu had an item "Donate" that gives the user the possibilities of donating some money to the author for further improvements. The programmer represented the history of donations using the following table.

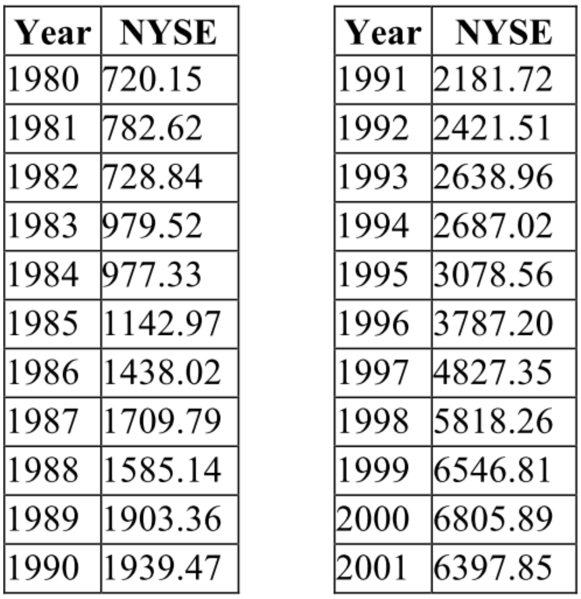
Month	1	2	3	4	5	6	7	8	9
Donations (in dollars)	7	16	42	93	121	155	240	390	610

a) Draw a scatterplot for the data. What pattern do you observe?

b) Use a proper transformation to achieve a linear relation. Find the linear regression line for the transformed data, and deduce a mathematical model for the donations in terms of the month.

c) Predict the value of the donation in the 12th month.

11. The following table gives the composite New York Stock Exchange (NYSE) index, which a stock market index covering all common stocks listed in the New York Stock Exchange. (Dec. 31, 2002 = 5,000)



Year	NYSE
1980	720.15
1981	782.62
1982	728.84
1983	979.52
1984	977.33
1985	1142.97
1986	1438.02
1987	1709.79
1988	1585.14
1989	1903.36
1990	1939.47

Year	NYSE
1991	2181.72
1992	2421.51
1993	2638.96
1994	2687.02
1995	3078.56
1996	3787.20
1997	4827.35
1998	5818.26
1999	6546.81
2000	6805.89
2001	6397.85

a) Plot the scatterplot of NYSE against the year. Can you observe a pattern? If so, describe it.

b) Using a power transformation, find a model for the NYSE as a function of Year. Does the model appear to fit the data well?

TASKS FOR INDEPENDENT WORK

1. Recognize the strength of the linearity from the value of *r*

Describe the relation between the variables of a set of bivariate data for each of the following values of the correlation coefficient.

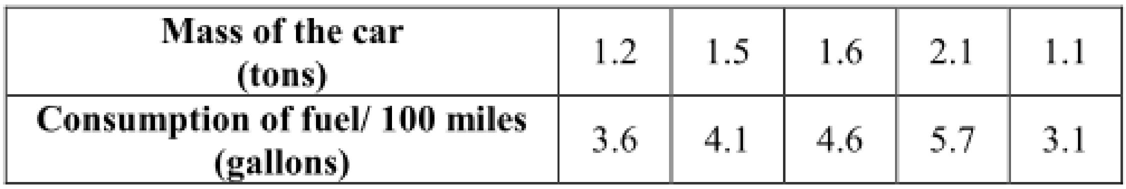
a) $r = 0.93$ _____________ b) $r = 0.4$ _____________

c) $r = 0.01$ _____________ d) $r = -0.88$ _____________

e) $r = -1$ _____________

2. Find the value of *r* for a set of bivariate data

The fuel consumptions of 6 cars, in gallons per 100 miles, versus the masses of the cars are given in the table below.



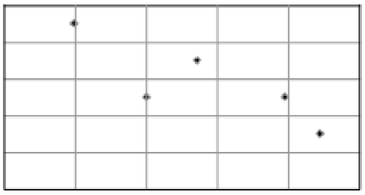
Mass of the car (tons)	1.2	1.5	1.6	2.1	1.1
Consumption of fuel/ 100 miles (gallons)	3.6	4.1	4.6	5.7	3.1

Find the correlation coefficient of this set of bivariate data.

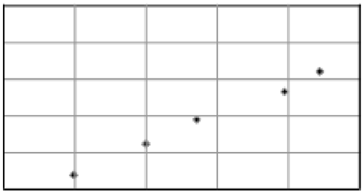
3. Recognize the nature of r from a scatterplot

The correlation coefficients of the sets of data represented by the scatterplots below are: 0.07, –1, 0.97, and –0.83. Match each of these values with the graph it represents.

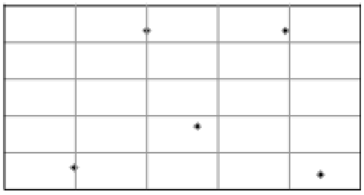
a)



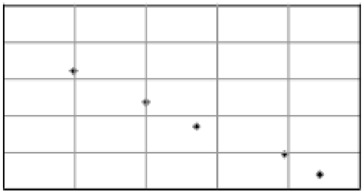
b)



c)



d)



7. PROBABILITY METHODS

1. A coin is tossed 3 times. Find

a) the probability that tails, heads, tails appear in that order.

b) the probability that at least one head is obtained.

c) the probability that at least two heads are obtained.

d) the probability that tails, heads, tails appear in any order.

2. In a group of people it is known that 15% have a dog only, 20% have a cat only and 10% have both. Find the probability that a person from this group chosen at random has a dog or a cat.

3. In a certain group, 25% of the group is from the UK and 55% is from the USA. If 15% are dual citizens of both the UK and the USA, find the probability that a person picked at random from the group is either from the UK or from the USA.

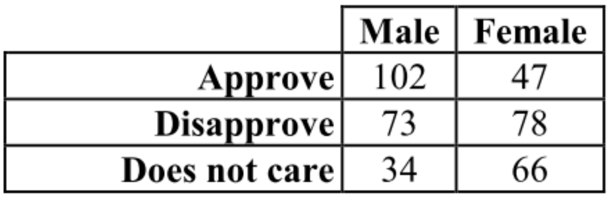
4. For a married couple, the probability that the husband has passed his driving test is 3/5 and the probability that the wife has passed her driving test is 1/2. The probability that the wife has passed, given that the husband has passed is 3/4. Find the probability that

a) both of them have passed.

b) only one of them has passed.

c) neither of them have passed.

5. A survey of four hundred people in a small town yielded the following data on their attitude towards implementing a new program by the city council:



	Male	Female
Approve	102	47
Disapprove	73	78
Does not care	34	66

Based on the survey, what is the probability that a randomly picked

a) male approves the program?

b) female does not care about the program?

c) who approves the program is female?

6. A lot of five hundred bipolar transistor chips contains 40 defective items. Two chips are chosen at random from the lot and without replacement.

a) What is the probability that the first selected chip is defective?

b) What is the probability that the second selected chip is defective given that the first one is defective?

c) What is the probability that both chips are defective?

7. Find the number of ways a committee of 6 people can be chosen from a group of 7 men and 9 women if

a) there are no restrictions on the gender of the six people.

b) the committee is unisex.

c) the ratio of men to women is 1 to 2.

d) there are at least as many women as men.

8. A total of 12 Management students, four International Management and American Business Studies (IMABS) and eight International Management and French Studies (IMF) students have volunteered to take part in an inter-university tournament. How many different ways can a team consisting of seven Management students, two IMABS and five IMF students be selected?

9. Suppose that you want to construct a list of codes for the books in your electronic library using only the digits 1, 2, 3, 4, 5, 7, 9.

a) How many five-digit codes can be formed if no digit is to be used more than once?

b) How many of the codes in part (a) are odd? How many of them are more than 30,000?

10. There are seven English books, ten Spanish books, and eight French books on a bookshelf (all books are different). How many ways are there to pick two books

a) of the same language?

b) of different languages?

11. A committee of six is to be picked at random from a group of twelve women and eleven men.

a) Find the probability that the committee consists of men only.

b) Find the probability that the committee consists of four women and two men.

12. A group consists of 10 adults with no twins. Thus, you can assume that their birthdays are independent. Assume also that all were born in a non-leap year. What is the probability that

a) at least two of the 10 have the same birthday?

b) all 10 have the same birthday?

13. A manager of a company proposes to buy coffee machines for office use. He surveys the employees about the number of cups they drink daily. He represents the result in the table below.

X	0	1	2	3	4	5	6	7	8	9
Relative frequency	0.08	0.12	0.21	0.24	0.15	0.11	0.02	0.04	0.02	0.01

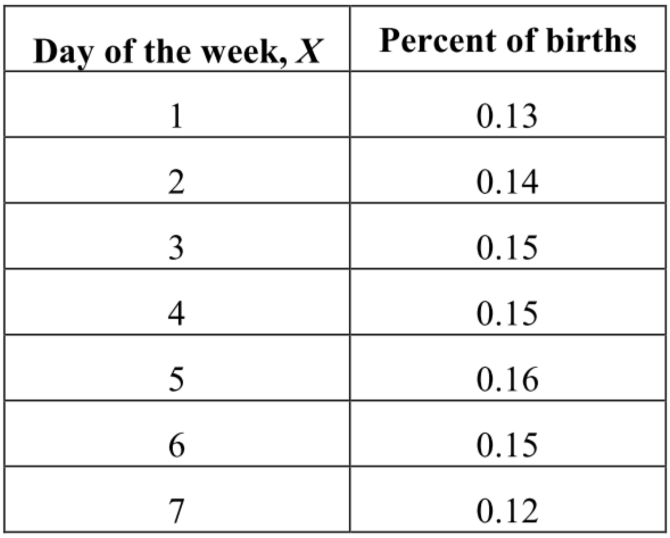
Where *X* is the number of cups consumed. What values of *X* make up the event “an employee extremely likes coffee” (the manager thinks that less than five cups is normal)? What is the probability of this event?

14. Dickson has a collection of vintage baseball cards. He wants to sell some of his collections. The table below shows his estimate for the chances of selling different number of cards.

Number of card	25	50	100	200
Probability	0.25	0.50	0.15	0.10

Take *X* to be the number of cards sold. Find the expected value of *X*.

15. Suppose that for all newborns who are born on the last week of the 9-month pregnancy period, the percentage of births on a give day of this week is distributed according to the table below.



Day of the week, *X*	**Percent of births**
1	0.13
2	0.14
3	0.15
4	0.15
5	0.16
6	0.15
7	0.12

a) Find the expected value of *X*.

b) Find the variance of *X*.

TASKS FOR INDEPENDENT WORK

1. Find the probability of an event

A marble is randomly picked from a bag containing 12 blue marbles and 16 red marbles. What is the probability of picking a blue marble?

2. Find the probability of the complement event

A die is rolled twice. What is the probability the sum of the two numbers appearing is not 2?

3. Find P(A∩B), A and B are dependent

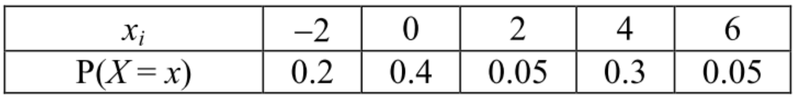
Two balls are drawn successively without replacement from a box which contains 5 blue balls and 12 red balls. Find the probability that the first ball drawn is blue and the second is red.

4. Use combinations to find probabilities

A committee of 5 is to be chosen from a group of 7 men and 3 women. If the committee is picked at random, what is the probability that it will have two women on it?

5. Find the probability and the expected value of a random variable

Find the expected value of the random variable *X* defined by the table below. Find P($X > 2$) and P($X \leq 0$).



x_i	-2	0	2	4	6
P($X = x$)	0.2	0.4	0.05	0.3	0.05

8. SAMPLING AND ESTIMATION

Sampling Terminology and Sampling Methods

1. Give appropriate sampling units associated with the following:

a) a library

b) a hospital

c) a music store

2. Give a suitable sampling frame in each case:

a) Students at a University

b) Owners of motorcars

c) Soccer players

3. Each electronic component produced by manufacturer A carries a unique identification number. A consumer organization wishes to test the claim that the components produced by A last longer than compatible components produced by other manufacturers. Manufacturer A has agreed to allow the organization to test a sample of 10 components.

a) Explain why a sample not a census is being taken.

b) Suggest a suitable sampling frame.

4. A magazine has a large number of subscribers who each pay a subscription fee due on the 1st of June every year. Not all of the subscribers pay their fee by the due date. The editor of the magazine believes that 40% of all subscribers wish to change the name of the magazine. A sample survey is carried out to obtain the opinions of the subscribers who have paid their fee on time that year.

a) Define the population associated with the magazine.

b) Suggest a suitable sampling for the survey.

c) Identify the sampling units.

d) Give one advantage and one disadvantage of using a census rather than a sample survey.

5. For each of the following surveys, state whether a census or random sampling is more convenient.

a) Testing a new chip for a computer

b) Election in a town to choose the new mayor

c) Checking the number of raisins per bun

6. A survey is done to find the views of students about choosing a new students body president. Describe he advantages and the disadvantages of conducting the survey by

a) surveying all the students in the school.

b) surveying 1% of the students in the school.

c) surveying 10% of the students in the school.

7. A survey is conducted on the average income per household in Alabama.

a) What is the population? What is the parameter under study?

b) Explain the advantages of conducting a census.

8. Which of the following samples can be considered independent?

a) A sociologist asks 40 randomly chosen adults about their attitude toward the current tendencies in the labor market.

b) Each month, a traveler randomly chooses the capital of a European country where he will visit. He does not choose the same city more than once.

c) The columnist of a New York economic paper interrogates five randomly chosen different managers from Top 10 Highest Earning Hedge Fund Managers.

9. State the disadvantages, if any, of each of the following sampling methods.

a) Anthony wants to estimate the percent of students in his school who spend more than one hour perweek on computer games. He randomly chooses five of his friends and asks them about the number of hours they spend each week playing computer games.

b) An assistant to a congressman wishes to estimate the percent of voters who are in favor of a new health care program. He sends to 100 randomly chosen people a letter of inquiry. In the letter he asks them to vote by calling a telephone number included in the letter.

10. The manager of a company wants to check the mean time spent on the internet by his employees.

a) Make a quick comparison between sampling and doing a census.

b) Explain how to select a sample using a table of random numbers.

11. A TV station provides cable services in five regions. The number of customers in these regions, in thousands, are 25, 43, 51, 22, and 8. The manager wants to choose 500 costumers and check their opinion about the shows that are being broadcasted.

a) What method is advisable for the survey?

b) Find the number of customers that must be chosen from each region?

Estimate for the population mean

The ages of twenty randomly selected workers in an oil company, to the nearest year, are:

42 63 28 41 33 24 50 39 45 29

23 54 30 27 52 36 31 26 42 72

Find an estimate for the population mean.

Estimate for the population variance

The ages of twenty randomly selected workers in an oil company, to the nearest year, are:

42 63 28 41 33 24 50 39 45 29

23 54 30 27 52 36 31 26 42 72

Find an estimate for the population variance.

An estimate for the proportion

The history of 3211 lung cancer cases is examined. 2560 of the patients are smokers. Give an unbiased estimate of the proportion p of smokers among the lung cancer suffering patients.

Application on the central limit theorem

The average time an employee takes from home to his office is 28 minutes with a standard deviation of 6 minutes. During the months of January, February, and March, he has 58 working days. What is the probability that his average time during these months to be between 27 minutes and 29 minutes?

Interval Estimation in Normal Distributions

1. The heights of 22 randomly selected students from a certain college, in cm, are listed below

167, 173, 182, 169, 188, 176, 183, 199, 176, 181, 166
179, 182, 169, 186, 173, 176, 184, 188, 165, 187, 172

a) Estimate the mean and the standard deviation for the height of the students in this college.

b) The basketball team coach thinks that only players that are more than 190 cm tall can play as center.

Estimate the proportion and the variance of the proportion of the students that are potential center players.

2. Below are the lifetimes (in hours) of 15 randomly selected light bulbs.

1325, 1458, 1551, 1469, 1289, 1527, 1454, 1620, 1581, 1611, 1413, 1302, 986, 1601, 1612

Bulbs with lifetime less than 1300 hours are labeled as defective.

a) Estimate the mean and the standard deviation for the lifetime of the population.

b) Estimate the proportion and the variance of the proportion of the defective bulbs.

3. Given a normal random variable X with mean μ and standard deviation 10. For a sample of size 200, find the length of a 100(1 – α)% confidence interval for μ if

a) α = 0.1.

b) α = 0.05.

4. Given a random sample of size n from N(μ, 1.5). Find the length of a 99% confidence interval for μ if

a) $n = 50$.

b) $n = 100$.

5. Using a sample from N(μ, 7.6), a student gets a confidence interval for μ of length 2.

a) Find the minimum size of the sample if the level of significance is equal to 0.05.

b) Find the level of significance if the sample size is 200.

6. The average height of 40 randomly selected adults from a certain population is 171 cm. Assume that the heights of all adults in this population are normally distributed with a standard deviation 17 cm. Find a confidence interval for the mean height of an adult from this population using a level of significance of

a) 0.01.

b) 0.05.

7. The data below represents the lengths, in cm, of 16 randomly selected rail pieces.

1193, 1201, 1185, 1194, 1211, 1203, 1199, 1186,
1192, 1196, 1206, 1188, 1191, 1185, 1193, 1195

Assume that the lengths of all of the rail pieces are normally distributed. Give an interval that contains the lengths of 99% of all the rail pieces if the standard deviation is:

a) $\sigma = 10$ cm

b) $\sigma = 18$ cm.

8. The average lifetime of a sample of 20 tires of a certain brand is 54,000 km. If we know that the lifetimes of the tires are normally distributed with a standard deviation of 4,500 km,

a) determine a 95% confidence interval for the mean lifetime of the tires.

b) find the level of significance of a confidence interval whose lower bound is 52,500 km.

9. The annual salaries of the employees of a large company are of unknown distribution with a standard deviation of 6.1 thousand dollars. The annual salaries (in thousand of dollars) of a random sample of 60 employees are:

45.3	44.7	54.2	62.4	45.6	43.6	44.8	67.4	48.3	45.1
38.2	45.7	39.4	55.1	44.6	43.1	46.2	48.2	46.2	44.6
47.3	38.7	43.3	44.7	41.6	46.3	44.7	45.6	58.1	45.8
44.1	44.8	37.5	44.9	46.2	43.9	45.6	48.3	44.4	45.1
46.2	44.7	44.8	39.5	41.6	42.4	45.3	46.4	45.7	66.7
44.2	45.1	46.8	42.9	43.6	44.5	43.7	44.8	43.7	48.1

The average of the above sample is 46.07 and its standard deviation is 5.65.

a) Find a 90% confidence interval for the mean salary of the employees of this company.

b) Find the minimum sample size needed to have a 90% confidence interval for the mean whose length is one third the one found in part (a).

10. The data listed below is the weights (in kg) of 48 randomly selected polar bears.

390	440	450	410	580	456	625	510	415	497	451	472
398	473	455	561	492	436	462	488	431	453	455	462
477	423	386	418	451	553	482	543	566	602	379	453
486	524	472	566	534	428	410	428	506	422	354	540

The standard deviation of this sample is 61.3.

a) Estimate the mean weight of the bear population from which the sample was taken.

b) Find a 99% confidence interval for mean weight of this population.

Confidence interval for the mean of a normal distribution when σ is unknown

1. Researchers study the collared lemur habitat. They estimate the territory range (in acres) that males of the species smell marks. The following data is a sample of size 20.

18, 21, 22, 19, 21, 42, 33, 23, 45, 44, 34, 41, 32, 33, 32, 25, 34, 36, 32, 41

Assuming normality, specify the distribution and its properties that the sample mean of size 20 satisfies.

2. Let $X \sim T(7)$.

a) Find $P(X < 2.05)$ and $P(X > 1.54)$.

b) Find $P(1.44 < X < 2.56)$.

3. Given a random variable X which follows a t-distribution with 9 degrees of freedom.

a) Calculate $P(X < 3.11)$ and $P(X < 2.74)$.

b) Use linear interpolation and the results of part (a) to estimate $P(X < 3.00)$.

4. A sample of size n is taken from a normal distribution and its mean and its unbiased standard deviation are calculated. Construct a 99% confidence interval for the mean of the population in each of the following cases.

a) $n = 4$, $\bar{x} = 27$, and $s_x = 12$.

b) $n = 31$, $\bar{x} = 27$, and $s_x = 12$.

Confidence Interval for Proportions

1. A survey is conducted on a random sample of 600 people. The proportion of those who are regular customers of social networking websites is 31%.

a) Find a 99% confidence level for the proportion of the population.

b) In a random sample of the same size, is it likely to get 240 people who are regular customers of social networking sites?

2. A public utility company randomly chooses 40 units for a scheduled inspection and finds that 13 of them have violations. Find a 90% confidence interval for the proportion of all units that have violations.

TASKS FOR INDEPENDENT WORK

1. Know how to carry a stratified sampling

A school district wants to check the favorite subject for the students in four of the local schools. The schools have 800, 1659, 216, and 4100 students. How many students should be selected from each school for a sample of approximate size 200?

2. Find a confidence interval for μ in a normal distribution when σ is known

The number of cars that pass a certain road between 7:00 A.M. and 7:30 A.M. is normally distributed with mean μ and standard deviation 35. An observation of 20 randomly selected days shows an average of 1,520 cars per day during that time of the morning. Find a 90%, 95%, and 99% confidence intervals for μ.

9. HYPOTHESIS TESTS

The null and the alternative hypothesis

1. A manufacturer uses a machine to package frozen food. The weights of the packages are normally distributed with standard deviation 3 oz. The average weight of the packages is supposed to be 28 oz. An inspector suspects that the mean weight is less than 28 oz. To test his suspicion, he weighs 50 packages that he picks at random. State the null and the alternative hypotheses of the test.

2. A poll, taken in 2020, shows that 16.3% of all Americans are not insured. A public officer in Florida believes that this percentage is lower in his state. To test his belief, he randomly chooses a sample of 55 Floridians and asks whether they are insured. Let *X* be the percent of uninsured Floridians. State the null and alternative hypotheses for the mean of *X*.

Significance Tests for the Mean

1. A computer service company claims that customers pay on average $7.2 per month for maintenance of home printers. An economist suspects that the mean of the monthly payment is more than it is claimed. To check his theory, he randomly chooses a sample of 80 customers and finds that the mean payment is $8.3 per month with a standard deviation $3.5.

a) What are the null and alternative hypotheses of the test?

b) Use a 5% level of significance to check if the economist's hypothesis should be supported.

2. A normal random variable *X* has a mean μ and a standard deviation 12. A random sample of size 15 from *X* is found to have a mean of 136. Find the critical value for the null hypothesis H_0: $\mu = 130$ with an alternative hypothesis H_a: $\mu > 130$ if the test is carried at a level of significance of

a) $\alpha = 0.05$ b) $\alpha = 0.01$

3. A random sample of size 12 is taken from a normal distribution with mean μ and standard deviation 5. The mean of the sample is found to be 32.4. Which hypothesis should we support H_0: $\mu = 30$ or H_a: $\mu > 30$ if we carry a test at a significance level α where

a) $\alpha = 0.01$

b) $\alpha = 0.05$

4. To test H_0: $\mu = 5$ against H_a: $\mu < 5$ at a significance level of 0.05, a student collects $n = 20$ observations and found that their mean is 3. Assume the population follows a normal distribution, should the student reject the null hypothesis if the standard deviation of the population is

a) $\sigma = 1$?

b) $\sigma = 3$?

5. The level of sugar in the blood at bedtime is normally distributed with a mean of 7.6 and a standard deviation of 1.3. A researcher theorizes that the mean is less than 7.6. To test his theory, he considers 13 random measurements. The values he finds, in mmol/L, are listed below.

7.23 7.94 6.83 7.41 9.50 6.48 7.57 7.20 6.47 7.92 6.48 6.10 6.54

a) State the null and the alternative hypotheses associated with this test.

b) What can be said about the researchers' theory? Give your answer using a 5% level of significance.

6. A jewelry designer advertises his new 18 karat (75% pure gold) gold ring collection. A trader suspects that the karat cauge is less than 18. To test his suspicions, he checks 10 rings. He concludes the following findings.

18.12 18.01 18.07 17.81 18.07 17.97 17.69 17.94 17.68 17.91

Assume that the data comes from a normal distribution, what conclusion can be made about the percentage of pure gold in the ring at

a) 10% significance level?

b) 1% significance level?

TASKS FOR INDEPENDENT WORK

Test hypothesis for the Mean

A computer service company claims that customers pay on average \$7.2 per month for maintenance of home printers. An economist suspects that the mean of the monthly payment is more than it is claimed. To check his theory, he randomly chooses a sample of 80 customers and finds that the mean payment is \$8.3 per month with a standard deviation \$3.5.

a) What are the null and alternative hypotheses of the test?

b) Use a 5% level of significance to check if the economist’s hypothesis should be supported.

Gossary

Term	Definition	Chapter
Addition Rule of probability	Used to find the probability of the union of two events: P(A ∪ B) = P(A) + P(B) – P(A ∩B). For mutually exclusive events, P(A ∪B) = P(A) + P(B).	7
Alternative Hypothesis, H_a	The competing statement that contradicts the null hypothesis. It is usually the conclusion the researcher wants to prove (e.g., H_a: $\mu \neq \mu_0$, H_a: $\mu > \mu_0$, or H_a: $\mu < \mu_0$).	9
Bar Chart	A type of statistical quality control chart used to monitor the average output of a production process.	1
Bar Graphs	Used to display the frequencies of occurrences of different categories, where each category is represented by a bar with height equal to the frequency.	4
Bivariate Data (Two-dimensional Data)	Data connecting exactly two variables, where each observation consists of an ordered pair (x_i, y_i).	6
Boxplot (Box-and-Whisker Plot)	A graphical display of the five-number summary, which visually represents the center, spread, and shape of a distribution.	5
Census	Collecting information from all the members of the target population. A survey	2, 8

	conducted to collect data from the entire population.	
Center of Gravity of the Data	The point $\bar{x}$, $\bar{y}$ which consists of the mean of the *x* variable and the mean of the *y* variable in the sample.	6
Central Tendency	Indicators (Mean, Median, Mode) that measure the central location or center of the data.	5
Central Limit Theorem (CLT)	A fundamental theorem stating that for a large sample size ($n \geq 30$), the sampling distribution of the sample mean ($\bar{x}$) will be approximately a Normal Distribution, regardless of the shape of the original population distribution.	8
Centralized Observation	Observation conducted and coordinated by a central authority, such as a national statistical office.	3
Certain Event, Ω	The sample space itself, representing an event that must occur.	7
Class	An interval used in a frequency distribution to group data, typically non-overlapping.	4
Classical Method	A method of assigning probabilities where all outcomes are equally likely. $\text{P(A)} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$	7

Cluster Sampling	The population is divided into clusters, and a random sample of the clusters is selected. All elements in the chosen clusters are included in the sample.	8
Coefficient of Determination, r^2	The proportion of the variation in the dependent variable (Y) that is predictable from the independent variable (X). It is the square of the correlation coefficient.	6
Coefficient of Variation	A measure of relative variability that expresses the standard deviation as a percentage of the mean. It is used to compare the spread of different data sets.	5
Complement of an Event, A^C or $\bar{A}$	The event consisting of all outcomes in the sample space that are not in A. $P(A^C) = 1 - P(A)$	7
Confidentiality	A core principle mandating that personal data must be protected from unauthorized disclosure.	3
Confidence Interval (CI)	An interval estimate that specifies a range of values within which the population parameter is expected to lie, with a certain level of confidence $1 - \alpha$.	8
Confidence Level	The probability $1 - \alpha$ that the confidence interval estimation procedure will yield an interval that actually contains the population parameter. Typical values are 90%, 95%, or 99%.	8

Conditional Probability P(A|B),	The probability of event A occurring, given that event B has already occurred.	7
Continuous Data	Quantitative data where the elements can take any value within a given interval (e.g., weight, time, temperature).	2
Continuous Observation	Data collected constantly or regularly as events occur.	3
Continuous Random Variable	A random variable that can take on any value in a given interval.	7
Convenience Sampling Surveys	Respondents are selected based on ease of access.	2
Correlation	The degree and direction of the linear relationship between two variables.	6
Correlation Coefficient, *r*	A unitless statistical measure that indicates the strength and direction of the linear relationship between two variables, ranging from -1 to $+1$.	6
Covariance (cov)	A statistical measure that indicates the direction of the linear relationship between two variables. A positive value suggests a positive relationship, and a negative value suggests a negative one.	6
Critical Value	The boundary value(s) that separate the rejection region from the acceptance region on the sampling distribution.	8, 9

Cross-Sectional Surveys	Data collected at a single point in time.	2
Cumulative Frequency	The total number of observations of a certain class interval and all the class intervals that precede it.	4
Cumulative Frequency Curve (S-curve)	A graph where the plotted points of upper-class limits and their cumulative frequencies are joined by a smooth curve.	4
Cumulative Frequency Polygonal Lines	A graph created by plotting the ordered pairs of the upper-class limits and their cumulative frequencies, and joining the points by line segments.	4
Data	Facts and figures collected, analyzed, and summarized for presentation and interpretation.	1, 2
Data Set	All the data collected in a particular study.	2
Decentralized Observation	Observation conducted by multiple organizations or entities, often coordinated by a central framework.	3
Degrees of Freedom, df	A parameter that fully specifies the *t*-distribution, calculated as *n* - 1 for a sample mean confidence interval.	8
Dependent Variable (Response Variable)	The variable *Y* whose value is being predicted or explained in regression analysis.	6

Descriptive Statistics	Methods involving the collection, summarization, and presentation of data using tabular, graphical, and numerical methods (e.g., Mean, Median).	2, 5
Direct Observation	Data is collected by a statistical employee who directly measures or records the phenomenon.	3
Direct Surveys	Surveys that involve direct interaction with respondents, such as interviews.	2
Discrete Data	Quantitative data where the elements can only take on a countable number of possible values.	2
Discrete Random Variable	A random variable that can take on a finite number of values or an infinite sequence of values.	7
Documentary Observation	Data is collected from existing documents, records, or secondary sources.	3
Dot Plot	A graphical summary used to describe a set of numerical observations, where each observation is represented by one dot on a horizontal number line.	4
Element	The entities on which data are collected (e.g., people, objects, events).	2
Estimation	The process of inferring the value of a population parameter based on sample data.	8

Event (A, B, C**)**	Any set of outcomes; a subset of the sample space Ω.	7
Experiment (Experimental Studies)	Data collected under controlled conditions to study cause-effect relationships.	3
Expected Value E(X) or μ	The mean or average value of a random variable, representing the long-run average of the outcomes.	7
Financial Analysts	Professionals who review statistical information (e.g., price/earnings ratios) to guide investment recommendations.	1
Five-Number Summary	A set of five descriptive statistics: Minimum, First Quartile Q_1, Median Q_2, Q_3, and Maximum.	5
Forecasting	The process economists use to predict the future of the economy or some aspect of it, utilizing statistical information.	1
Frequency	The number of times (count) that an item or observation occurs within a specific class.	4
Frequency Distribution	A tabular summary of data showing the number (frequency) of items in each of several non-overlapping classes.	4
Frequency Tables	Tables used to summarize data when listing the individual items is impractical due to the large size of the data.	4

Histogram	A graphical representation of grouped data where each class is represented by a rectangle. The height is the corresponding frequency.	4
Hypothesis Testing	A statistical procedure that uses sample data to assess two competing statements (hypotheses) about a population parameter (e.g., mean μ).	9
Impossible Event ∅	The empty set, representing an event that cannot occur.	7
Impartiality	A core principle requiring data to be collected and reported objectively.	3
Independent Events	Events where the occurrence of one event does not affect the probability of the other event occurring.	7
Independent Variable (Predictor, Explanatory Variable)	The variable *X* used to predict or explain the value of the dependent variable.	6
Indirect Surveys	Surveys where data is collected through intermediaries or observational methods without direct respondent interaction.	2
Inferential Statistics	Methods that use data from a sample to draw conclusions or make inferences about the characteristics of the population (e.g., estimation, hypothesis testing).	2

Inference	The process of using data collected from a sample to draw conclusions about a population.	8
Interval Estimation	The process of estimating a population parameter by constructing an interval of values (a confidence interval).	8
Interval Scale	Data where the interval between values is expressed in fixed units, but the ratio of two values is not meaningful (e.g., temperature in Celsius).	2
Inter-Quartile Range (IQR)	The difference between the third quartile and the first quartile (Q_3 - Q_1). It measures the spread of the middle 50% of the data.	5
Intersection of Events $A \cap B$	The event containing all outcomes that belong to both event A and event B.	7
Leaf	The part of the item (usually to the right) that consists of the digit(s) in the smallest place value in a stem-and-leaf plot.	4
Left Skewed (Negatively Skewed)	A distribution where the tail stretches out to the left (mean is typically less than the median).	5
Left-Tailed Test	A one-tailed test where the rejection region is entirely in the left tail of the sampling distribution (H_a: $\theta < \theta_0$).	9
Legislative Framework	National and international laws that define the rights, responsibilities, and obligations	3

	of statistical agencies and users of statistical data.	
Level of Significance α	The probability of making a Type I error (rejecting H_0 when it is actually true). It is the maximum allowable probability of a Type I error.	8, 9
Linear Relationship	A relationship between two variables that can be best described by a straight line.	6
Longitudinal Surveys	Data collected repeatedly from the same units over an extended period.	2
Margin of Error	The amount added to and subtracted from the point estimator to create the confidence interval.	8
Mean (Average) $\overline{x}$	The arithmetic average; the sum of all data values divided by the number of values.	5
Median	The middle value of a data set when the items are arranged in increasing order. It divides the data into two equal halves.	5
Mode	The data item that occurs with the greatest frequency in the data set.	5
Modified Boxplot	A boxplot where outliers are represented as separate points, and the whiskers reach out to the last observation that is not an outlier.	5
Multivariate Data	Data involving more than one variable.	6

Multiplication Rule of probability	Used to find the probability of the intersection of two events: P(A · B) = P(A) P(B | A). For independent events, P(A · B) = P(A) P(B).	7
Mutually Exclusive Events (Disjoint Events)	Events that cannot occur at the same time; their intersection is the impossible event, $A \cap B = \varnothing$.	7
National Statistical Office (NSO)	The main governmental agency responsible for producing official statistics.	3
Negative Dependency (Negative Correlation)	A downward-sloping pattern in a scatterplot, indicating that as the *x* variable increases, the *y* variable tends to decrease.	6
Nominal Scale	The lowest level of measurement, where data are labels or names used only to identify an attribute of the element (categorization).	2
Nonlinear Relationship	A relationship between two variables that is not a straight line (e.g., exponential, logarithmic, or power relationship).	6
Non-Probability Sampling	Sampling methods where the probability of selecting an element is not known or cannot be determined (e.g., convenience sampling).	8

Normal Distribution	The most important continuous probability distribution. It is bell-shaped and symmetrical, defined by its mean μ and variance σ^2.	7
Normative and Legal Provisions of Statistics	Frameworks, laws, standards, and regulations that govern the collection, processing, analysis, and dissemination of statistical data.	3
Null Hypothesis, *H_0*	A tentative assumption about a population parameter, typically stating that the parameter is equal to a specific value (e.g., H_0: $\mu = \mu_0$). It represents the status quo.	9
Observation	The set of measurements obtained for a particular element. Also, the systematic process of collecting primary data.	2, 3
Observation Program	A set of questions or characteristics to be observed and recorded during data collection.	3
Observation Unit	The basic element or entity from which data is collected.	3
Observational Studies	Data collected passively, without intervention.	3
One-Tailed Test (Unilateral Test)	A hypothesis test where the alternative hypothesis is specified as being strictly greater than H_a: $\theta > \theta_0$ or strictly less than H_a: $\theta < \theta_0$ the hypothesized parameter value.	9

Ordinal Scale	Data has the properties of nominal data, and the order or rank of the data is meaningful.	2
Outcome	A single result of a random experiment.	7
Outlier	A data item that is unusually small or unusually large compared to the bulk of the data.	5
Parameter	A numerical characteristic of a population (e.g., population mean μ, population standard deviation σ).	2, 8
p-value	The probability of obtaining a test statistic value as extreme as or more extreme than the observed one, assuming the null hypothesis, H_0 is true.	9
Percentiles	Values that divide the data into 100 equal parts. The *k*-th percentile is the value such that *k*% of the items are smaller than or equal to this value.	5
Periodic Observation	Data collected at fixed, regular intervals (e.g., monthly, yearly).	3
Pie Chart (Circle Graph)	A disk divided into sectors, where each sector represents an item in the data and its area is proportional to the frequency.	4
Point Estimation	Using a single sample statistic (the point estimator) to estimate the value of the population parameter (e.g., using $\overline{x}$ to estimate μ).	8

Point Estimator	The sample statistic used to estimate a population parameter.	8
Population	The entire group of individuals, objects, or items about which conclusions are to be drawn.	2, 8
Positive Dependency (Positive Correlation)	An upward-sloping pattern in a scatterplot, indicating that as the *x* variable increases, the *y* variable also tends to increase.	6
Probability of an Event, P(A)	A numerical measure of the likelihood that an event will occur. Its value is always between 0 and 1.	7
Probability Density Function (PDF)	The function *f*(*x*) for a continuous random variable used to find probabilities over an interval (the area under the curve).	7
Probability Distribution	A description of how the probabilities are distributed over the values of a random variable.	7
Probability Mass Function (PMF)	The function *f*(*x*) for a discrete random variable that provides the probability for each value *x*.	7
Probability Sampling	Sampling methods where each element in the population has a known, non-zero probability of being selected.	8
Probability Theory	The mathematical modeling of random phenomena (uncertainty).	7

Professional Independence	A core principle ensuring statistical bodies operate free from political influence.	3
Qualitative (Categorical) Data	Data that are grouped by specific categories (non-numerical), such as gender or marital status.	2
Quality Control	An application of statistics in production that uses statistical charts to monitor the output of a production process.	1
Quantitative (Numerical) Data	Data that uses numerical values and represents quantities (e.g., age, salary).	2
Quartiles, Q_1, Q_2, Q_3	Values that divide the ordered data into four equal parts.	5
Random Experiment	A process that leads to two or more possible outcomes, where it is unknown which outcome will occur before the experiment is performed.	7
Random Sampling Surveys	Respondents are selected randomly to ensure representativeness.	2
Range	The difference between the maximum and minimum values in the data set.	5
Ratio Scale	The highest level of measurement, where the ratio of two values is meaningful, and zero represents the absence of the characteristic being measured (e.g., age).	2

Rejection Region (Critical Region)	The set of values for the test statistic that leads to the rejection of the null hypothesis.	9
Rejection Rule p-value approach	The null hypothesis H_0 is rejected if the p - value is less than the Level of Significance, α.	9
Relative Frequency	The fraction or proportion of times an item occurs, which is the frequency of the item divided by the size of the data.	4
Residual (Error)	The difference between the observed value of the dependent variable Y_i and the value predicted by the regression line $\hat{Y}_i$.	6
Residual Plot	A graph of the residuals versus the independent variable or the predicted values, used to check the assumptions of the regression model.	6
Regression	A statistical technique used to model the relationship between a dependent variable Y and one or more independent variables X.	6
Regression Line (Line of Best Fit)	The line that minimizes the sum of the squared vertical distances (residuals) from the data points to the line, calculated using the Least Squares Method.	6
Relative Frequency Method	A method where the probability is estimated based on the proportion of times an event occurred in a large number of past trials (historical data).	7

Right-Tailed Test	A one-tailed test where the rejection region is entirely in the right tail of the sampling distribution, H_a: $\theta > \theta_0$.	9
Sample	A subset of the population used to collect data to make inferences about the larger group.	2, 8
Sample Space, Ω	The set of all possible outcomes of a random experiment.	7
Sample Survey (Partial Observation)	A survey conducted to collect data from a sample (subset) of the population.	2
Sampling	The process of selecting a subset of individuals from a population for analysis.	8
Sampling Distribution	The probability distribution of a sample statistic (e.g., the sample mean $\bar{x}$) obtained from all possible samples of the same size.	8
Sampling Error	The difference between the value of a sample statistic and the value of the corresponding population parameter, which occurs simply due to the random nature of sampling.	8
Sampling Frame	A list of all the elements in the population from which a sample is to be drawn.	8
Scatterplot (Scatter Diagram)	A graphical representation of bivariate data where the ordered pairs are plotted in a Cartesian coordinate system. Used to	6

	visualize the relationship between two variables.	
Significance Test for the Mean	A hypothesis test procedure used to test an assumption about the population mean μ.	9
Simple Linear Regression	A method used to model the relationship between two variables, *X* and *Y* by fitting a straight line to the data.	6
Simple Random Sampling (SRS)	A sampling method where every possible sample of the same size, *n* has an equal chance of being selected.	8
Size of the Data	The total number of items under study.	4
Skewness	A measure of the asymmetry of a probability distribution.	5
Standard Deviation (s **or** σ**)**	The positive square root of the variance. The most commonly used measure of dispersion.	5
Standard Error of the Mean	The standard deviation of the sampling distribution of the sample mean, $\frac{\sigma}{\sqrt{n}}$. It measures the average sampling error.	8
Standard Normal Distribution	A specific normal distribution with a mean of 0 and a standard deviation of 1, Z ~ N(0, 1).	7
Statistic	A numerical characteristic of a sample (e.g., sample mean $\overline{x}$, sample standard	2, 8

	deviation σ). Used as an estimate of the population parameter.	
Statistical Observation (Observation)	The systematic process of collecting primary data for statistical analysis.	3
Statistical System	The network of institutions and mechanisms responsible for producing, disseminating, and using statistical information in a country.	3
Statistics	A discipline that deals with the collection, analysis, interpretation, presentation, and organization of data.	1
Stem	The part of the item (usually to the left) that consists of the digit(s) in the largest place value in a stem-and-leaf plot.	4
Stem-and-Leaf Plot	A method to display data by dividing items into two parts: the stem and the leaf.	4
Stratified Random Sampling	The population is first divided into non-overlapping groups (strata), and a simple random sample is then drawn from each stratum.	8
Stratified Sampling Surveys	The population is divided into subgroups (strata), and samples are taken from each.	2
Subjective Method	A method where the probability is based on an individual's judgment, intuition, or past	7

	experience (often used when no historical data is available).	
Survey	A systematic method for gathering data from individuals or groups.	2
Symmetry	A data shape where one side of the distribution is a mirror image of the other side.	5
Systematic Sampling	Elements are selected from the population at regular intervals, often after a random start.	8
***t*-test**	A statistical test used to test hypotheses when the population standard deviation σ is unknown and estimated by the sample standard deviation *s*. It uses the t-distribution.	9
Test of a Population Proportion	A hypothesis test used to test an assumption about the proportion, *p* of a population that possesses a certain characteristic.	9
Test Statistic	A numerical value calculated from the sample data (e.g., *Z* or *T*) that is used to decide whether to reject the null hypothesis.	9
t-Distribution (Student's t-distribution)	A probability distribution used for interval estimation and hypothesis testing when the population standard deviation, σ is unknown and must be estimated by the sample standard deviation *s*.	8

Transformed Data	Data that has been modified using a mathematical function (e.g., logarithm, square root) to make a nonlinear relationship appear linear for easier analysis.	6
Trend Line	A straight line drawn on a scatterplot that best represents the general direction or trend of the data points.	6
Two-Tailed Test (Bilateral Test)	A hypothesis test where the alternative hypothesis states that the parameter is not equal to the hypothesized value, H_a: $\theta \neq \theta_0$. The rejection region is split between both tails.	9
Type I Error (α Error)	The error of rejecting the null hypothesis H_0 when it is true.	9
Type II Error (β Error)	The error of failing to reject the null hypothesis H_0 when the alternative hypothesis H_a is true.	9
Union of Events, $A \cup B$	The event containing all outcomes that belong to event A or event B or both.	7
Univariate Data	Data involving a single variable.	6
Variable	A characteristic or attribute of the elements.	2
Variance σ^2	A measure of dispersion based on the squared deviations of data values from the mean.	5, 7

***z*-score**	The number of standard deviations a value *x* is away from the mean μ, calculated as $Z = \frac{x-\mu}{\sigma}$.	7
***Z*-test**	A statistical test used to test hypotheses when the population standard deviation, σ is known (or when the sample size *n* is large, $n \geq 30$).	9

RECOMMENDED LITERATURE AND INTERNET RESOURCES

Methodical support

1. Maria Khomyak Mathematics and statistics for economists: some guidelines on Statistics. Lutsk: Lesya Ukrainka VNU, 2022. 22 p.

2. Maria Khomyak Statistics: Course Description. Lutsk : Lesya Ukrainka VNU, 2022. 26 p.

3. Khomyak M., Mykytyuk I. MATHEMATICS: differential equations: methodological workshop on problem solving. Lutsk : Lesya Ukrainka VNU, 2024. 70 p.

4. Khomyak M., Mykytyuk I. MATHEMATICS: Integration Techniques: methodological workshop on problem solving. Lutsk : Lesya Ukrainka VNU, 2024. 90 p.

5. Khomyak M. MATHEMATICS AND STATISTICS FOR AN INTERNATIONAL ECONOMIST: modeling with functions: methodological guidelines. Lutsk : Lesya Ukrainka VNU, 2024. 40 p.

6. Khomyak M. MATHEMATICS AND STATISTICS FOR AN INTERNATIONAL ECONOMIST: some elements of Calculus: methodological instructions. Lutsk : Lesya Ukrainka VNU, 2024. 60 p.

7. Khomyak M. Analysis of data on the organization of distance learning. *Middle east international conference on contemporary scientific studies-V*, March 27-28, 2021, Ankara, Turkey. Vol.II, P. 384-386.

8. Khomyak M. A polinomial errors-in-variables model in forecasting of economic processes. *Information society: technological, economic and technical aspects of development: coll. theses add. International of science Internet Conf.* Vol. 52. Ternopil, 2020. P. 17-19. (in Ukr.)

9. Maria Khomyak A goodness-of-fit test of a difussion model Hagia Sophia. *5th International conference on multidisciplinary scientific studies*. 2022. Istanbul, Turkey, 2022. P.85-86.

10. Yunchyk V., Khomyak M., Fedonuyk A., Yatsyuk S.,Cognitive modeling of the learning process of training IT specialists (2021) *CEUR Workshop Proceeding*, Volume 2917, Pages 141–150, : 3 rd International Workshop on Modern Machine Learning Technologies and Data Science, MoMLeT+DS 2021 (Scopus)

11. Pasichnyk V., Kunanets N., Yunchyk V., Khomyak M., Yatsyuk S., Muliar V., Fedonuyk A. Model of Recommender System of the Selection of Electronic Learning Resourses. *CEUR Workshop Proceeding*: *5 rd International Workshop on Modern Machine Learning Technologies and Data Science*, MoMLeT+DS 2023, Vol. 3426, P. 344-355. (Scopus)

12. Pasichnyk V., Kunanets N., unchyk V., Khomyak M., Fedonuyk A., Knysh Yu. Expert assement of education content in IT specialist training process. *MoDaST-2024: 6th International Workshop on Modern Data Science Technologies*, Vol. 3723, pp. 121-132. ISSN 1613-0073. (Scopus)

13. Pasichnyk V., Kunanets N., Yunchyk V., Khomyak M., Fedonuyk A. Project of an Educational Content Evaluation Recommender System. *Proceedings of the 5th International Workshop IT Project Management* (ITPM 2024), Vol. 3709, P. 192-203. ISSN 1613-0073. https://ceur-ws.org/Vol-3709/paper16.pdf(Scopus)

14. Khomyak M., Melnyk Ar. Data visualization in economic research using Power BI. *Matematyka. Informatsiini tekhnolohii. Osvita*: zbirnyk tez dop. XIII mizhnar. nauk.-prakt. konf. (m. Lutsk, 31 travn.-2 chervn. 2024.). Lutsk,. P. 74-76

Recommended Books

1. Bruce Hansen Probability and Statistics for Economists. Princeton University Press, 2022. 416 p. https://press.princeton.edu/books/hardcover/9780691235943/probability-and-statistics-for-economists#preview

2. Horkavy V. K. Statistics: Textbook. Kind. 3rd, perovl. and added Textbook. Kyiv: Alerta, 2020. 644 p. (in Ukr.)

3. Kopytko B.I., Kopych I.M. , Sorokivskyi V.M. Applied mathematical statistics for economists. Education. manual (rek. MON Ukrainy), 2021. 404 p. (in Ukr.)

4. Monga G.S. Mathematics and Statistics for Economics. Vikas Publishing House Pvt, New Delhi. 912 p. https://www.biblio.com/book/mathematics-statistics-economics-gs-monga/d/500019018

5. Motoryn R.M., Chekotovskyi E.V. Statistics for economists: a study guide. Kyiv: Znannia, 2021. 381p. (in Ukr.)

6. Neter, Wasserman, and Whitmore. Applied Statistics, 4th Edition, Allyn and Bacon, Boston, MA.

7. Panik Michael J. Mathematical Analysis and Optimization for Economists. CRC Press, Boca Raton-London-New York, 2022. https://www.routledge.com/Mathematical-Analysis-and-Optimization-for-Economists/Panik/p/book/9780367759025

Internet resources

1. State Statistics Service of Ukraine. [Electronic resource]. URL: www.ukrstat.gov.ua

2. UN Statistical Committee. [Electronic resource]. URL: http://unstats.un.org/

3. International Institute of Statistics. [Electronic resource]. URL: http://isi.cbs.nl/

4. UN Statistical Committee. [Electronic resource]. Access mode: http://unstats.un.org/ 2.

5. International Institute of Statistics. [Electronic resource]. Access mode: http://isi.cbs.nl/

6. Trevor Hastie, Robert Tibshirani, and Jerome Friedman.The Elements of Statistical Learning. https://hastie.su.domains/ElemStatLearn/

7. Data Science. Full Course. Learn Data Science in 10 Hours | Data Science For Beginners | Edureka https://www.youtube.com/watch?v=-ETQ97mXXF

8. Statistics. A Full University Course on Data Science Basics, https://www.youtube.com/watch?v=xxpc-HPKN2

Навчальне електронне видання CD-ROM

Khomyak Maria

STATISTICS
FOR AN INTERNATIONAL ECONOMIST
Educational Manual

СТАТИСТИКА
ДЛЯ ЕКОНОМІСТА-МІЖНАРОДНИКА
Навчальний посібник

Один електронний оптичний диск (CD-ROM). Об’єм даних 5,65 Мб.
Наклад 300 пр. Зам. 154. Видавець і виготовлювач – Вежа-Друк
(м. Луцьк, вул. Шопена, 12, тел. (0332) 29-90-65).
Свідоцтво Держ. комітету телебачення та радіомовлення України
ДК № 4607 від 30.08.2013 р.