Week 1

Introduction to Information Technology and Computing

---

1. Role of Information Technology (IT) in Society

Information Technology refers to the use of computers, networks, software, and digital systems to collect, store, process, and share information.

Why IT matters in modern society

- **Communication:** Email, SMS, video conferencing, social media.
- **Business & Finance:** Automated trading, mobile banking, risk analysis.
- **Healthcare:** Electronic health records, medical imaging, disease surveillance.
- **Education:** E-learning platforms, virtual labs.
- **Government:** E-government services, digital ID systems.

Impact: Faster processing, higher efficiency, automation, accuracy, and large-scale data handling.

---

2. Data vs Information

Concept	Meaning	Example
Data	Raw facts, unprocessed values	"98, 105, 110"
Information	Processed/organized data that is meaningful	"Patient's temperature increased over 3 days."

Data becomes information after processing (sorting, analysis, summarizing).

Importance in statistics:
Statistics transforms raw data → meaningful conclusions → decision making.

---

3. Computing Applications in Biostatistics

Biostatistics uses IT to handle **large, complex health datasets**.

Where IT is applied

- **Clinical trials:** Data entry, cleaning, storage, and automated statistical analysis.
- **Epidemiology:** Disease surveillance, outbreak prediction models, mapping disease spread.
- **Hospital systems:** Electronic medical records (EMR), patient monitoring devices.
- **Bioinformatics:** DNA sequencing, genomics, protein analysis.

Why computing is essential

- Health datasets are huge (genomics files can be 50GB+).
- Requires high-speed processing, secure storage, and reliable systems.

---

4. Computing Applications in Financial Engineering

Financial engineering requires **fast computation**, **real-time data**, and **complex modeling**.

Applications

- **Algorithmic trading:** High-frequency systems executing thousands of trades per second.
- **Risk modeling:** Monte Carlo simulation, Value-at-Risk (VaR).
- **Portfolio optimization:** Optimization algorithms to maximize return/minimize risk.
- **Predictive modeling:** Machine learning applied to financial markets.

Why computing matters

- Financial markets generate massive data streams every second.
- Requires high-speed hardware, cloud systems, and specialized software (Python, R, MATLAB).

---

5. High-Performance Computing (HPC)

High-Performance Computing allows solving large computational problems using:

- **Parallel computing**
- **Clusters and supercomputers**
- **Graphics Processing Units (GPUs)**

Why HPC is important

- Clinical data → millions of records

- Genomic data → extremely large files
- Financial modeling → complex, iterative simulations

HPC reduces tasks from hours → seconds.

---

6. Cloud Computing in Statistics

Cloud computing provides **remote servers** to store data and run computations.

Types

- **IaaS (Infrastructure as a Service):** AWS EC2, Google Compute Engine
- **PaaS (Platform as a Service):** Google App Engine
- **SaaS (Software as a Service):** Google Sheets, Dropbox, SPSS Cloud

Advantages

- Scalable storage
- Access from anywhere
- Reduced hardware costs
- Automatic backups
- High reliability

Use cases in Biostatistics & Finance

- Storing large datasets (hospital records, financial tick data)
- Running machine learning models
- Hosting databases
- Collaborative research

Week 2

Fundamentals of Computer Operations

---

1. Overview of the Central Processing Unit (CPU)

The CPU is the "brain" of the computer where most processing takes place. It interprets instructions, performs calculations, and controls data flow.

Main Components

1. **Control Unit (CU)**
   - Directs operations of the computer.
   - Fetches instructions, decodes them, and tells other parts what to do.
   - Manages control signals to memory, ALU, and I/O devices.
2. **Arithmetic Logic Unit (ALU)**
   - Performs arithmetic operations (addition, subtraction, multiplication, division).
   - Performs logical operations (AND, OR, NOT, comparisons).
   - Handles decision-making and numeric calculations used in statistics and finance.
3. **Registers**
   - Very small, ultra-fast storage areas inside the CPU.
   - Hold data the CPU is currently processing.
   - Types include:
     - *Accumulator*
     - *Instruction Register (IR)*
     - *Program Counter (PC)*
     - *Memory Address Register (MAR)*
     - *Memory Data Register (MDR)*
4. **System Clock**
   - Sends electrical pulses that synchronize all operations.
   - Measured in GHz (1 GHz = 1 billion cycles per second).
   - Higher frequency → more instructions processed per second.

---

2. The Machine Cycle

The machine cycle is the step-by-step process through which the CPU executes instructions.

Four Stages

1. **Fetch** – CPU takes an instruction from memory.

2. **Decode** – Instruction is interpreted by the Control Unit.
3. **Execute** – ALU performs the required operation.
4. **Store** – The result is written back to memory or a register.

This cycle repeats billions of times per second.

---

3. Instruction Cycle vs Machine Cycle

- **Instruction Cycle** = Fetch + Decode
- **Machine Cycle** = Fetch + Decode + Execute + Store

Together, these determine the speed of computation.

---

4. Parallel Processing Basics

Parallel processing enables a computer to perform many operations simultaneously.

Types of Parallelism

1. **Bit-level parallelism** – Operates on multiple bits at once.
2. **Instruction-level parallelism** – Pipeline execution (fetch one instruction while executing another).
3. **Data-level parallelism** – Same operation applied to many data points (important in statistics).
4. **Task-level parallelism** – Multiple tasks run at the same time.

Hardware Used for Parallel Processing

- **Multi-core CPUs** (dual-core, quad-core, octa-core)
- **GPUs** (Graphics Processing Units) – handle thousands of small tasks concurrently
- **Computer clusters** – multiple computers working together
- **Vector processors** – used in scientific computing

Importance in Biostatistics & Finance

- Speeding up simulations (e.g., Monte Carlo).
- Processing huge datasets.
- Real-time financial predictions.
- Fast epidemiological modeling.

---

5. Factors Affecting Computer Performance

1. **Clock Speed (GHz)**
2. **Number of Cores**
3. **Cache Memory (L1, L2, L3)**
4. **RAM Capacity**
5. **Data Bus Width** (32-bit vs 64-bit)
6. **Storage type** (SSD vs HDD)
7. **Graphics capability** (GPU acceleration)

---

6. Performance Evaluation in Statistical Computing

To assess computing efficiency for statistical tasks, consider:

1. Processing Speed

- Time taken to run code.
- Measured in milliseconds/seconds.

2. Throughput

- Amount of data processed per unit time.
- Important for large datasets.

3. FLOPS (Floating Point Operations Per Second)

- Indicates performance for numerical computation.

4. Benchmarking

Practical tools used:

- **Excel** – basic operations and formula execution speed.
- **R** – running loops, matrix operations, or simulations.
- **Python** – performance on NumPy, Pandas, or computational algorithms.

Why Benchmark?

- To choose the right hardware for specific statistical tasks.
- To measure improvements when upgrading systems.
- To determine best software/hardware configurations.

Week 3

Computer Hardware I — Input & Output Devices

---

1. Introduction to Computer Hardware

Computer hardware refers to the **physical components** of a computer system.
Week 3 focuses on two major categories:

1. **Input devices** – send data *into* the computer
2. **Output devices** – send processed information *out* of the computer

These devices support data collection, analysis, and reporting in Biostatistics and Financial Engineering.

---

2. Input Devices

Input devices allow the user to enter data, commands, or signals into the computer.

A. Common Input Devices

1. Keyboard

- Primary text-entry device.
- Includes alphanumeric keys, function keys, numeric keypad.
- Used in data entry, coding, documentation.

2. Mouse

- Pointing device for selecting UI elements.
- Supports drag-and-drop functionality.

3. Scanner

- Captures images and converts them into digital data.
- Used for:
    - Digitizing medical forms
    - Scanning survey questionnaires

    - Document archiving

4. Microphones (Audio Input)

- Capture sound as digital signals.
- Used in:
    - Telemedicine
    - Voice recognition
    - Data annotation (e.g., interviews)

5. Cameras / Webcams

- Capture images or live video.
- Used in:
    - Telehealth
    - Facial recognition systems
    - Machine learning datasets

6. Biometric Devices

- Capture biological characteristics:
    - Fingerprint scanners
    - Iris scanners
    - Facial scanners
- Used in secure healthcare and financial systems.

B. High-Throughput Input Devices

These devices handle **large volumes of data**, useful for big statistical datasets.

Examples

- **Barcode readers**
- **RFID readers**
- **Medical sensors & wearables**
- **High-speed laboratory instruments (PCR machines, MRI scanners)**
- **Financial market data feed devices**

Importance:
They enable real-time collection of massive datasets used in:

- Epidemiology
- Clinical trials

- High-frequency trading (HFT)
- Risk modeling

---

3. Output Devices

Output devices present processed information to the user in various forms: text, audio, video, or print.

A. Common Output Devices

1. Monitors / Displays

- Primary output device.
- Types: LCD, LED, OLED.
- Used for:
    - Data visualization
    - Dashboard analysis
    - Coding

2. Printers

- Convert digital documents into physical hard copies.
- Types:
    - Inkjet
    - Laser
    - Dot matrix (for receipts)
- Used in:
    - Printing statistical reports
    - Financial statements
    - Medical summaries

3. Speakers

- Output audio signals.
- Used in alarms in medical systems or financial notifications.

4. Projectors

- Display content on a large screen.
- Useful in presentations, data visualization, and meetings.

B. Specialized Output Devices

1. Plotters

- Produce large, high-quality graphics.
- Used for printing:
  - Large charts
  - Engineering/financial diagrams
  - Epidemiological maps

2. Medical Output Devices

- Digital displays of ECG, MRI, and X-ray readings.

3. Financial Dashboards

- High-speed, multi-screen setups for traders.
- Show real-time tick data, charts, and analytics.

---

4. I/O Performance Measurement

In data-heavy fields like biostatistics and finance, input/output performance is crucial.

I/O performance is measured by:

1. **Input speed** – how fast data enters the system
2. **Output speed** – how fast results are delivered
3. **Latency** – delay between command and response
4. **Bandwidth** – amount of data processed per second

Why I/O matters:

- Large datasets require fast reading/writing.
- Real-time financial analysis needs low latency.
- Clinical devices need accuracy and high throughput.

Week 4

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are statistics that describe the **center**, **average**, or **typical value** of a dataset. They help summarize large amounts of data into a single representative number.
The three main measures are:

1. **Mean (Arithmetic Mean / Average)**
2. **Median**
3. **Mode**

---

1. MEAN

Definition

The mean is the sum of all the values divided by the number of observations.

Characteristics

- Uses **all data values**
- Affected by **extreme values (outliers)**
- Best for **symmetrical distributions**

Types

a) Simple Arithmetic Mean (as above)

b) Weighted Mean
Used when values have different levels of importance (weights).

c) Mean for Grouped Data

---

2. MEDIAN

Definition

The median is the **middle value** when data is arranged in order.

How to Find the Median

a) For Raw (Ungrouped) Data

1. Arrange data in ascending order
2. If n is odd → median = middle value
3. If n is even → median = average of the 2 middle values

b) For Grouped Data

Median=L+(N2−c.ffm)×h\text{Median} = L + \left(\frac{\frac{N}{2} - c.f}{f_m} \right) \times hMedian=L+(fm2N−c.f)×h

Where:

- L = lower boundary of median class
- N = total frequency
- c.f = cumulative frequency before median class
- fm = frequency of median class
- h = class width

Characteristics

- Not affected by extreme values
- Best for skewed distributions (e.g., income data)

---

3. MODE

Definition

The mode is the **value that occurs most frequently** in a dataset.

Characteristics

- Can have **one mode (unimodal)**, **two modes (bimodal)**, or **many modes**
- Useful for **categorical data** (e.g., most common crop grown)

Mode for Grouped Data

---

4. COMPARISON OF MEAN, MEDIAN, AND MODE

Measure	Advantages	Limitations	Best Use
Mean	Uses all data, easy to compute	Affected by outliers	Symmetric data
Median	Not affected by outliers	Does not use all values	Skewed data
Mode	Easy, works with categories	May not be unique	Categorical/nominal data

---

5. DISTRIBUTION SHAPES & CENTRAL TENDENCY

a) Symmetrical Distribution

Mean = Median = Mode

b) Positively Skewed (Right Skew)

Mode < Median < Mean

c) Negatively Skewed (Left Skew)

Mean < Median < Mode

---

6. USE CASES OF CENTRAL TENDENCY

- Economics → average income, inflation
- Health → average blood pressure

- Education → mean test scores
- Business → average sales
- Demography → median age of population

Week 5

MEASURES OF DISPERSION

Measures of dispersion describe how **spread out**, **scattered**, or **variable** data values are around the center (mean/median).
While measures of central tendency summarize data with a single value, dispersion shows how **reliable** or **representative** that value is.

---

1. PURPOSE OF MEASURES OF DISPERSION

They help to:

- Show the **degree of variability** in data
- Compare the spread of two or more datasets
- Indicate **data reliability** (small dispersion = more reliable)
- Identify **consistency** e.g., consistent production, consistent scores
- Aid in choosing appropriate statistical models

---

2. TYPES OF DISPERSION

A. RANGE

Definition

Difference between the highest and lowest value.

$$Range = Xmax - Xmin$$

Characteristics

- Very easy to compute
- Affected by **extreme values**
- Only uses **2 values** (not reliable)

B. interquartile range (IQR)

Definition

The difference between the third quartile (Q3) and first quartile (Q1).

$$IQR = Q3 - Q1$$

$$\{IQR = (Q3 - Q1) \div 2$$

Purpose

- Measures spread of the **middle 50%** of data
- Not affected by outliers
- Best for skewed distributions

Quartile Deviation / Semi-Interquartile Range

$$QD = Q3 - Q12QD = \backslash frac\{Q\_3 - Q\_1\}\{2\}QD = 2Q3 - Q1$$

C. MEAN DEVIATION (MD)

Definition

Average of the absolute deviations of each value from the mean or median.

$$MD = \sum | X - measure | \div n$$

Where the measure is usually the **median** (more stable).

D. VARIANCE AND STANDARD DEVIATION

These are the **most important and most used** measures of dispersion.

---

1. Variance

Variance measures the average squared deviation from the mean.

For population:

$$\sigma 2 = \sum (X - \mu) 2 \div N$$

For sample:

$$s2 = \sum (X - X^{-}) 2n - 1s\text{^}2 \ = \backslash frac\{\backslash sum\ (X\ - \backslash bar\{X\})\text{^}2\}\{n - 1\}s2 = n - 1\sum (X - X^{-}) 2$$

2. Standard Deviation (SD)

This is the square root of variance.

$$SD = \sqrt{\sigma 2}$$

Why Standard Deviation is Useful

- Same units as original data
- Measures spread accurately
- Used in almost all statistical techniques
- Important for **normal distribution**, **confidence intervals**, **hypothesis testing**

E. COEFFICIENT OF VARIATION (CV)

Definition

A relative measure of dispersion expressed as a percentage.

$$CV = SD/X^{-} \times 100\%$$

Purpose

- Compares variability between datasets with different units or scales
- Lower CV = more consistent, more reliable data

---

3. DISPERSION FOR GROUPED DATA

For grouped data, formulas involve:

- Class midpoints (x)
- Frequencies (f)
- Summations ∑fx\sum f x∑fx and ∑fx2\sum f x^2∑fx2

Variance for Grouped Data

$$\sigma 2 = \sum fx2N - (\sum fxN)2\sigma^2 = \frac{\sum f\ x^2}{N} - \ (\frac{\sum f\ x}{N})^{2\sigma 2} = N\sum fx2 - (N\sum fx)2$$

Standard Deviation

σ=σ2\sigma = \sqrt{\sigma^2}σ=σ2

The method is the same for sample data except using **n−1**.

---

4. PROPERTIES OF A GOOD MEASURE OF DISPERSION

A good measure should:

- Include all values
- Not be overly affected by extreme values
- Be easy to compute and understand
- Allow further mathematical treatment (variance & SD do this)
- Be comparable between datasets (CV helps)

---

5. WHEN TO USE WHICH MEASURE

Situation	Best Measure
Data with outliers or skewed	IQR or QD

Situation	Best Measure
Normally distributed data	SD and variance
Quick estimate of spread	Range
Comparing different datasets	CV
Median used as center	Mean deviation

---

6. APPLICATIONS OF DISPERSION

- Finance → risk assessment (higher SD = higher risk)
- Climate → temperature variability
- Health → variations in BP, cholesterol, weight
- Manufacturing → quality control (low variance = consistent products)
- Education → variation in exam scores
- Demography → population distribution

Week 6

DATA FILES & FILE MANAGEMENT

Focuses on how data is **stored, organized, accessed, indexed, retrieved, and optimized** in statistical and computational environments. This is important in **Biostatistics** and **Financial Engineering**, where datasets can be large (clinical trials, epidemiological data, time-series data, transactions, risk simulations, etc.).

---

1. TYPES OF DATA FILES

A. Sequential Files

- Records are stored **one after another**.
- Access is **linear** → must read from the beginning to reach a specific record.
- Good for:
  - Log files
  - Transaction histories
  - Simple datasets
- Disadvantage: Slow random access.

B. Random/Direct Access Files

- Records accessed directly using a **key or position**.
- Fast and efficient retrieval.
- Used in:
  - Databases
  - Statistical software
  - Financial modeling systems

C. Structured Files

- Data is organized in a **defined format**:
  - Tables
  - Rows & columns
  - CSV, Excel, SQL tables
- Perfect for:
  - Biostatistics datasets
  - Portfolio/risk datasets

D. Unstructured Files

- No predefined data model.
- Examples:
  - Text documents
  - Audio
  - Images
  - PDFs
- Used in:
  - Medical imaging analysis
  - Social media sentiment analysis
  - NLP applications

---

2. FILE CREATION

This involves:

- Defining the **structure** (fields, types, sizes)
- Setting **validation rules**
- Choosing **file type** (CSV, Excel, SQL table)
- Specifying **primary keys**

Example:

- Patient_ID (primary key)
- Age
- Treatment_Dose
- Outcome

---

3. FILE INDEXING

Indexing improves **speed of access**.

Types of Indexes:

- **Primary Index** → built on a primary key
- **Secondary Index** → built on non-primary fields
- **Clustered Index** → physically arranges records
- **Non-clustered Index** → separate index structure

Why indexing?

- Reduces search time
- Fast sorting & filtering
- Essential for large statistical datasets (clinical trials, stock prices, etc.)

---

4. FILE RETRIEVAL

Retrieval is how records are **located and extracted**.

Retrieval Techniques:

- **Sequential retrieval** → read entire file
- **Indexed retrieval** → use index to jump to record
- **Random/direct access retrieval** → find data using address/key

Good retrieval ensures efficient:

- Querying
- Statistical analysis
- Data loading into R/Python

---

5. FILE OPTIMIZATION

Optimization improves performance by:

- Reducing file size
- Creating/maintaining indexes

- Normalizing data
- Removing redundancy
- Using efficient formats (e.g., Parquet for big data)

---

6. RELATIONAL DATABASES

Relational databases store data in **tables** linked by **relationships**.

Key Concepts:

- **Table** → like an Excel sheet
- **Record** → row
- **Field** → column
- **Primary key** → unique identifier
- **Foreign key** → establishes links between tables

Advantages:

- Easy to manage
- Prevents duplication
- High integrity
- Ideal for:
    - Medical records
    - Financial transaction data
    - Market data systems

---

7. NORMALIZATION

Normalization organizes data to reduce **redundancy** and improve **consistency**.

Normal Forms:

1. **1NF** → No repeating groups, atomic values
2. **2NF** → No partial dependencies
3. **3NF** → No transitive dependencies

Why normalize?

- Prevents anomalies

- Improves data quality
- Makes statistical analysis cleaner