1. a) Define the term "arithmetic mean". Discuss how to calculate it for ungrouped data using the direct method. (5 marks)

The **arithmetic mean**, commonly referred to as the **average**, is the most fundamental measure of central tendency in statistics. It is defined as the sum of all observations in a dataset divided by the total number of observations. The arithmetic mean serves as a representative value that summarizes the entire dataset into a single figure, providing a useful snapshot of the data's central location. It is applicable to interval and ratio scale data and forms the basis for many advanced statistical analyses.

For **ungrouped data**—that is, raw data that has not been organized into classes or intervals—the arithmetic mean is calculated using the **direct method**. This involves summing all the individual data points and dividing by the count of observations. Mathematically, if we have n observations denoted as $x_1, x_2, x_3, \ldots, x_n$, the arithmetic mean $\bar{x}$ is computed as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For example, consider the ungrouped dataset representing the heights (in cm) of 5 plants: 12, 15, 18, 14, 16.
The sum is $12 + 15 + 18 + 14 + 16 = 75$.
Since $n = 5$, the mean height is $\frac{75}{5} = 15$ cm.

The direct method is straightforward and effective for small datasets. However, for large datasets, computational tools are often employed. The mean is sensitive to extreme values (outliers), which can skew the result, so it is always interpreted alongside measures of dispersion. In biostatistics, the arithmetic mean is widely used to summarize experimental results, such as average enzyme activity, mean blood pressure in a sample, or average gene expression levels.

1. b) What is dispersion, and explain its importance in statistical analysis. (5 marks)

Dispersion, also known as variability or spread, refers to the extent to which individual data points in a dataset differ from the central tendency (like the mean). While measures of central tendency provide a typical value, dispersion quantifies the degree of scatter, inconsistency, or diversity within the data. Common measures of dispersion include **range**, **variance**, **standard deviation**, and **interquartile range**.

Importance of dispersion in statistical analysis:

1. **Understanding Data Reliability**: Low dispersion indicates that data points are clustered closely around the mean, suggesting high consistency and reliability. High dispersion signals greater variability and potential instability in measurements. For example, in clinical trials, low variability in drug response strengthens the conclusion about its efficacy.
2. **Comparative Analysis**: Dispersion allows comparison between two or more datasets that may have similar means but different spreads. For instance, two regions may have the same average rainfall but different variability—one may have steady rainfall, while the other experiences extremes. This has implications for agricultural planning and risk assessment.
3. **Foundation for Statistical Inference**: Many inferential statistics, including hypothesis tests (like t-tests and ANOVA) and confidence intervals, rely on measures of dispersion (particularly variance and standard deviation) to compute standard errors and assess the precision of estimates.
4. **Risk Assessment in Research**: In fields like epidemiology or genomics, understanding variability is crucial for assessing risk factors, mutation rates, or

expression heterogeneity. High genetic variability in a population, for example, might influence disease susceptibility studies.

5. **Quality Control and Decision Making**: In bioinformatics and laboratory sciences, control charts use dispersion measures to monitor process stability. Consistent low variability is often a goal in high-throughput sequencing or microarray experiments to ensure reproducibility.

In summary, without assessing dispersion, a mean alone can be misleading. A complete statistical analysis always reports both central tendency and dispersion to give a truthful representation of the data.

2. Differentiate between the following pairs of terms: (2½ × 4 = 10 marks)

a) Geometric mean and Harmonic mean

The **geometric mean** is defined as the n-th root of the product of n observations: $\mathrm{GM} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$. It is particularly useful for datasets with multiplicative or exponential characteristics, such as growth rates, ratios, and indices. For example, it is used to calculate average compound annual growth rates in population biology or bacterial proliferation.

The **harmonic mean** is the reciprocal of the arithmetic mean of the reciprocals of the observations: $\mathrm{HM} = \frac{n}{\sum(1/x_i)}$. It is appropriate for averaging rates or ratios when time or other denominators are involved, such as average speed over fixed distances or protein turnover rates. The harmonic mean tends to be less affected by extremely large values but more affected by very small values compared to the arithmetic mean.

Key difference: While the geometric mean multiplies values, the harmonic mean emphasizes the smallest values. For any given dataset, $\mathrm{HM} \leq \mathrm{GM} \leq \mathrm{AM}$.

b) Positive Correlation and Negative Correlation

Positive correlation exists when two variables move in the same direction; an increase in one variable is associated with an increase in the other. The correlation coefficient r ranges between 0 and +1. Example: Height and weight in humans generally show positive correlation.

Negative correlation exists when two variables move in opposite directions; an increase in one leads to a decrease in the other. Here, r ranges between -1 and 0. Example: The relationship between exercise frequency and body fat percentage is typically negative.

Both are measures of linear association, but they describe opposite directional relationships. Neither implies causation.

c) Global alignment and Local alignment

Global alignment (e.g., Needleman–Wunsch algorithm) aligns two sequences over their entire length, end to end. It is best used when sequences are of similar length and expected to be homologous across their full extent, such as aligning two orthologous genes.

Local alignment (e.g., Smith–Waterman algorithm) finds regions of highest similarity within sequences, ignoring dissimilar sections. It is useful for identifying conserved domains, motifs, or functional sites when sequences differ in length or share only isolated regions of homology, such as in multidomain proteins or when searching for a short motif in a large genome.

d) PAM and BLOSUM

PAM (Point Accepted Mutation) matrices are derived from evolutionary models based on closely related protein sequences. PAM1 represents 1% amino acid change; higher PAM numbers (e.g., PAM250) extrapolate to longer evolutionary distances. They assume a Markov model of evolution and are older, used for evolutionary studies.

BLOSUM (BLOcks SUbstitution Matrix) matrices are constructed from conserved, ungapped alignment blocks in protein families. BLOSUM62 (using sequences with ≤62% identity) is standard for most protein searches. They are more modern, empirically derived, and generally preferred for database searching (like BLAST) as they better capture local conservation without evolutionary modeling assumptions.

3. Write short notes on the following: (2½ × 4 = 10 marks)

a) Importance of hypothesis testing

Hypothesis testing is a cornerstone of inferential statistics, allowing researchers to make probabilistic decisions about population parameters based on sample data. It begins with formulating a **null hypothesis** (H_0), which represents a default position (e.g., no effect or no difference), and an **alternative hypothesis** (H_1), which asserts the presence of an effect. Using sample data, a test statistic is computed and compared against a theoretical distribution to obtain a **p-value**. If the p-value is less than a predetermined significance level (e.g., $\alpha = 0.05$), H_0 is rejected in favor of H_1.

The importance lies in its ability to provide an objective, structured framework for scientific decision-making. It helps control for random error, differentiate between true effects and chance occurrences, and supports reproducibility. In biological research, hypothesis testing is used in diverse scenarios—from determining if a drug lowers blood pressure to assessing whether gene expression differs between treatment and control groups.

b) Poisson distribution

The **Poisson distribution** is a discrete probability distribution that models the number of events occurring within a fixed interval of time or space, given that these events happen with a known constant mean rate λ and independently of the time since the last event. Its probability mass function is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

It is characterized by the property that its mean and variance are both equal to λ. The Poisson distribution is widely used in molecular biology and genomics to model rare events, such as:

- The number of mutations occurring in a given stretch of DNA per unit time.
- The distribution of reads mapping to a gene in RNA-seq (under certain assumptions).
- The count of bacterial colonies forming on a plate.

It serves as the foundation for many statistical models in high-throughput sequencing data analysis.

c) Artificial Neural Network for protein prediction

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of interconnected layers of nodes (neurons): an input layer, one or more hidden layers, and an output layer. Each connection has a weight that is adjusted during training via backpropagation to minimize prediction error.

In **protein prediction**, ANNs are used to map sequence or physico-chemical features to structural or functional outcomes. Applications include:

- **Secondary structure prediction** (e.g., classifying residues as helix, sheet, or coil).
- **Protein function annotation** based on sequence motifs.
- **Protein-protein interaction prediction**.
- **Subcellular localization prediction**.

ANNs can capture complex, non-linear relationships in data, making them powerful for pattern recognition in large biological datasets. However, they require substantial training data and computational resources, and their "black-box" nature can limit interpretability.

d) Significance of sequence alignment

Sequence alignment is the process of arranging biological sequences (DNA, RNA, protein) to identify regions of similarity. Its significance is multifaceted:

1. **Evolutionary Insights**: Alignments reveal conserved regions, suggesting functional or structural importance, and enable the construction of phylogenetic trees to study evolutionary relationships.
2. **Functional Annotation**: Unknown sequences can be annotated by aligning them with well-characterized sequences in databases.
3. **Structural Prediction**: Conserved residues often correspond to structurally or functionally critical sites, aiding in 3D modeling.
4. **Mutation and Variant Analysis**: Alignments help identify SNPs, indels, and mutations associated with diseases.
5. **Drug Design and Vaccine Development**: By aligning pathogen proteins with human proteins, researchers can identify unique epitopes for targeted therapy or vaccine design.

Both global and local alignment tools (like BLAST, Clustal Omega) are indispensable in genomics, proteomics, and molecular biology.

4. a) Explain the importance of nucleotide databases in molecular biology and bioinformatics. (5 marks)

Nucleotide databases are foundational repositories that store, organize, and provide access to DNA and RNA sequence data. Their importance in molecular biology and bioinformatics cannot be overstated:

1. **Centralized Knowledge Resource**: Databases such as **GenBank (NCBI)**, **EMBL-Bank (EBI)**, and **DDBJ** form the International Nucleotide Sequence Database Collaboration (INSDC), ensuring that sequence data is universally accessible, standardized, and non-redundant. This enables researchers worldwide to share and retrieve data seamlessly.
2. **Support for Genomic Research**: These databases archive entire genomes, from bacteria to humans, facilitating comparative genomics, genome annotation, and the identification of genes, regulatory elements, and repetitive sequences.
3. **Basis for Sequence Analysis Tools**: Most bioinformatics tools (e.g., BLAST, primer design software, genome browsers) rely on nucleotide databases as their reference datasets. When a researcher sequences a new gene, they compare it against these databases to infer homology, function, or evolutionary origin.
4. **Enabling Evolutionary and Phylogenetic Studies**: By providing aligned sequences and curated phylogenetic information, databases allow scientists to reconstruct evolutionary histories, study molecular evolution rates, and understand speciation events.

5. **Clinical and Diagnostic Applications**: Pathogen sequences (e.g., viral genomes from SARS-CoV-2) are rapidly shared via these databases, enabling real-time tracking of outbreaks, variant analysis, and the development of diagnostic PCR primers and vaccines.
6. **Education and Training**: They serve as essential resources for training the next generation of biologists and bioinformaticians in data retrieval, annotation, and analysis.

In essence, nucleotide databases are the backbone of modern biology, transforming raw sequence data into actionable biological knowledge.

4. b) What is the FASTA file format, and how is it structured? (5 marks)

The **FASTA format** is a simple, text-based format for representing nucleotide or protein sequences. It is one of the most widely used formats in bioinformatics due to its simplicity and human-readability.

Structure of a FASTA file:

1. **Header Line**: Begins with a **greater-than symbol (>)**, immediately followed by a sequence identifier and optional descriptive comments. The header line contains no spaces immediately after the ">". Example:

```
text

>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens OX=9606 GN=TP53 PE
=1 SV=4
```

2. **Sequence Data**: Following the header line, the actual sequence is written in subsequent lines. The sequence can span multiple lines (typically 60-80 characters per line for readability). It consists of:
   - For **nucleotides**: Letters A, T, C, G, U, N (for unknown), and other IUPAC codes for ambiguity.
   - For **proteins**: Single-letter amino acid codes (A, R, N, D, etc.).

Example:

```
text

>Sequence1
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTED
PGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLH
```

Key Characteristics:

- The sequence ends when another header line (">") begins or the file ends.
- Whitespace and numbers within the sequence lines are typically ignored by parsers, but it is good practice to avoid them.
- Line breaks within the sequence are allowed but not required.

Variants:

- **Multi-FASTA**: A single file containing multiple sequences, each with its own header.
- **FASTQ**: An extension used in next-generation sequencing that includes quality scores alongside sequence data.

FASTA files are used as input for almost all sequence analysis software, including alignment tools (BLAST, Clustal), phylogenetic programs, and genome assemblers.

Their simplicity ensures longevity and broad compatibility across computational platforms.

5. Elaborate on the principles of classification, which include class limits, class intervals, range and class frequency and explain how these things are applied in biostatistical data analysis. (10 marks)

Classification is the process of organizing raw data into meaningful groups or classes to facilitate summarization, visualization, and analysis. In biostatistics, this is often the first step in handling large datasets, such as measurements from clinical trials, ecological surveys, or genomic read counts.

Key Principles:

1. **Class Limits**: These are the boundaries that define each class. The **lower class limit** is the smallest value that can belong to the class, and the **upper class limit** is the largest value. For example, in a class "10–19" for ages, 10 is the lower limit and 19 is the upper limit. Limits can be **inclusive** (10–19 includes 10 and 19) or **exclusive**, though inclusive is common in biostatistics.
2. **Class Interval (or Width)**: This is the difference between the upper and lower class limits of a class. It should be uniform across classes for simplicity. The choice of interval affects the granularity of analysis; too wide an interval may hide details, while too narrow an interval may cause fragmentation. A common rule of thumb is to have 5–20 classes. Interval is calculated as:

$$\text{Class interval} = \frac{\text{Range}}{\text{Number of classes}}$$

3. **Range**: The overall spread of the data, computed as the difference between the maximum and minimum observed values:

$$\text{Range} = X_{\max} - X_{\min}$$

   The range helps determine the span over which classes must be defined.
4. **Class Frequency**: The number of observations falling into a given class. It is the count of data points whose values lie between the class limits. Frequencies can be absolute (counts) or relative (proportions/percentages). The sum of all class frequencies equals the total number of observations.

Application in Biostatistical Data Analysis:

In practice, these principles are applied as follows:

- **Step 1: Determine Range**. For example, in a study of systolic blood pressure (BP) in 100 patients, values range from 100 mmHg to 180 mmHg. Range = $180 - 100 = 80$.
- **Step 2: Decide Number of Classes (k)**. Using Sturges' rule: $k = 1 + 3.322 \log_{10}(n)$. For n=100, $k \approx 8$.
- **Step 3: Calculate Class Interval**. Interval $\approx 80/8 = 10$. So, we can use a class width of 10 mmHg.
- **Step 4: Define Class Limits**. Starting just below the minimum (e.g., 99.5 to avoid overlap), classes might be: 99.5–109.5, 109.5–119.5, ..., 179.5–189.5.
- **Step 5: Tally Observations & Compute Class Frequencies**. Count how many patients' BP falls into each class. This creates a **frequency distribution table**.
- **Step 6: Analyze and Visualize**. The frequency distribution can be plotted as a histogram, which visually reveals the shape (normal, skewed), central tendency, and dispersion of the data. Further statistics like the median class or modal class can be identified.

Biostatistical Context:

- In **epidemiology**, age groups (classes) are used to analyze disease incidence.
- In **genomics**, expression levels of genes (e.g., log-counts per million) are classified into bins to assess overall distribution.
- In **ecology**, species abundance data is classified to understand community structure.

Classification simplifies complex data, reduces noise, and prepares data for statistical measures and hypothesis testing, making it an indispensable step in biostatistical workflows.

6. a) Explain how the C /C++ programming languages are used in BLAST and SAMTools. (5 marks)

BLAST (Basic Local Alignment Search Tool) and **SAMTools** are two cornerstone bioinformatics tools, both heavily reliant on **C and C++** for performance-critical components.

In BLAST:
BLAST performs rapid sequence similarity searches against large databases. The core alignment algorithms (seed-and-extend, hit extension, gapped alignment) are implemented in **C++** for several reasons:

- **Speed and Efficiency**: C++ allows low-level memory management and optimized CPU usage, which is essential when scanning billions of base pairs.
- **Portability**: C++ code can be compiled on virtually any operating system (Linux, Windows, macOS), ensuring wide accessibility.
- **Integration with NCBI Toolkit**: BLAST is part of the NCBI C++ Toolkit, a large suite of libraries for biological data processing. This provides reusable modules for sequence I/O, database indexing, and statistics calculation (like E-values).
- **Parallelization**: C++ facilitates multi-threading (using OpenMP or pthreads) to exploit multi-core processors, speeding up searches.

In SAMTools:
SAMTools is a suite for manipulating alignments in the SAM/BAM format (common in next-generation sequencing). It is written primarily in **C**.

- **Memory and Speed Constraints**: Sequencing data files (BAM) can be terabytes in size. C allows efficient binary I/O, compression/decompression (via zlib), and in-memory data structures to handle these files with minimal memory footprint.
- **System-Level Control**: C provides fine-grained control over file pointers, memory allocation, and bit-level operations needed for parsing and writing binary alignment maps.
- **Stability and Reliability**: The C codebase of SAMTools is robust, widely tested, and forms the engine for many downstream pipelines (e.g., variant calling with bcftools).

Both tools exemplify how systems programming languages like C/C++ remain vital in bioinformatics for performance-intensive, data-heavy tasks where execution speed and resource efficiency are paramount.

6. b) Discuss the significance of RNA secondary structure prediction and compare the tools Mfold and RNAfold. (5 marks)

Significance of RNA Secondary Structure Prediction:
RNA molecules fold into specific secondary structures (helices, loops, bulges, and

junctions) that are critical for their function. Predicting these structures computationally is vital because:

1. **Functional Insight**: Structure determines function in non-coding RNAs (tRNA, rRNA, miRNA, lncRNA). For example, riboswitches change conformation upon ligand binding to regulate gene expression.
2. **Therapeutic Design**: Knowing the structure of viral RNA genomes (e.g., HIV, SARS-CoV-2) aids in designing antisense oligonucleotides, siRNA, or small molecules that target specific structural motifs.
3. **Evolutionary Conservation**: Secondary structures are often more conserved than primary sequences, helping in homology detection and alignment.
4. **Synthetic Biology**: Designing novel RNA devices (e.g., sensors, regulators) requires accurate structure prediction.

Comparison of Mfold and RNAfold:

Feature	**Mfold**	**RNAfold** (ViennaRNA Package)
Development Era	Older (late 1990s), pioneered by Michael Zuker.	More recent, actively maintained.
Algorithm	Uses free energy minimization with dynamic programming.	Also uses free energy minimization (similar core algorithm).
Interface	Primarily web-based; command-line version available.	Command-line focused; part of a comprehensive suite.
Speed & Scalability	Slower for long sequences.	Generally faster, optimized C code.
Output & Features	Provides detailed graphical plots, suboptimal foldings, and energy dot plots.	Outputs structure in dot-bracket notation, equilibrium base pairing probabilities, and visualization via VARNA or forna.
Integration	Standalone.	Integrates with other ViennaRNA tools (RNAalifold, RNAcofold).
Usage Context	Educational, quick web-based folding.	Preferred in high-throughput pipelines and scripting.

Conclusion: While both tools are based on similar thermodynamic principles, **RNAfold** is generally favored in modern bioinformatics workflows due to its speed, flexibility, and integration within the ViennaRNA suite. **Mfold** remains historically important and user-friendly for quick web analyses.

7. a) What is the p-value, and how is it used in hypothesis testing? (5 marks)

The **p-value** (probability value) is a fundamental concept in statistical hypothesis testing. It quantifies the strength of evidence against the null hypothesis (H_0). Formally, the p-value is the probability of obtaining test results at least as extreme as the observed results, **assuming that the null hypothesis is true**.

Interpretation:

- A **small p-value** (typically ≤ 0.05) indicates that the observed data is unlikely under H_0. This leads to rejecting H_0 in favor of the alternative hypothesis (H_1).
- A **large p-value** suggests that the observed data is consistent with H_0, so we fail to reject H_0.

Steps in Using p-value in Hypothesis Testing:

1. **Formulate Hypotheses**: Define H_0 (e.g., "no difference between groups") and H_1 (e.g., "there is a difference").
2. **Choose Significance Level (α)**: Commonly set at 0.05 (5%).
3. **Collect Data and Compute Test Statistic**: Perform an appropriate test (e.g., t-test, chi-square).
4. **Determine p-value**: Using statistical tables or software, find the probability associated with the observed test statistic under H_0.
5. **Make Decision**:
   - If $p \leq \alpha$, reject H_0.
   - If $p > \alpha$, fail to reject H_0.

Example: In a drug efficacy trial, H_0: mean recovery time with drug = mean recovery time with placebo. If a t-test yields p = 0.03 (< 0.05), we reject H_0 and conclude the drug has a significant effect.

Caution: The p-value is **not** the probability that H_0 is true, nor does it measure the size of an effect. It is merely a measure of compatibility between the data and H_0. Misinterpretation of p-values is a common issue in scientific literature.

7. b) Define the term "Binomial distribution". Discuss with a suitable example. (5 marks)

The **Binomial distribution** is a discrete probability distribution that models the number of successes k in a fixed number n of independent Bernoulli trials, each with the same probability of success p.

Parameters:

- n: number of trials.
- p: probability of success in each trial (constant).
- $q = 1 - p$: probability of failure.

Probability Mass Function:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

Properties:

- Mean: $\mu = np$
- Variance: $\sigma^2 = np(1-p)$

Example – Genetics:
Consider a cross between two heterozygous pea plants for a trait (e.g., seed shape: round (R) dominant over wrinkled (r)). The probability of an offspring being round (success) is $p = 0.75$, and wrinkled is $q = 0.25$. If we examine $n = 10$ offspring, the number of round-seeded plants X follows a Binomial distribution: $X \sim \text{Bin}(n = 10, p = 0.75)$.

Question: What is the probability of getting exactly 7 round-seeded plants?

$$P(X = 7) = \binom{10}{7}(0.75)^7(0.25)^3 \approx 0.250$$

Thus, there is about a 25% chance of observing exactly 7 round seeds out of 10.

The Binomial distribution is widely used in biology for:

- **Genetic inheritance predictions** (as above).
- **Quality control** in lab experiments (e.g., number of successful PCR reactions).
- **Ecology** (e.g., number of animals surviving in a sample).
  It provides a fundamental model for binary outcome data.

8. a) Discuss the advantages and disadvantages of Spearman's rank correlation coefficient. (5 marks)

Spearman's rank correlation coefficient (r_s) is a non-parametric measure of the monotonic relationship between two variables. It assesses how well the relationship can be described using a monotonic function (whether linear or not). It is calculated by ranking the data and then applying Pearson's correlation formula to the ranks.

Advantages:

1. **Non-Parametric**: Does not assume normality of data or homoscedasticity, making it robust for non-linear monotonic relationships.
2. **Ordinal Data Compatibility**: Can be used with ordinal (ranked) data, not just interval/ratio data.
3. **Robust to Outliers**: Since it uses ranks, extreme values have less influence compared to Pearson's correlation.
4. **Interpretability**: Values range from -1 (perfect negative monotonic relationship) to +1 (perfect positive monotonic relationship), with 0 indicating no monotonic association.
5. **Applicability to Non-Linear Trends**: Captures any consistent increasing or decreasing trend, even if it is not linear.

Disadvantages:

1. **Less Power than Pearson**: If the data truly are linear and normally distributed, Pearson's correlation is more statistically powerful (more likely to detect a true effect).
2. **Information Loss**: Ranking discards the actual magnitude of differences between values, which can be important in some analyses.
3. **Sensitive to Ties**: Many tied ranks can complicate computation and may require adjustment formulas.
4. **Only Captures Monotonic Relationships**: It will fail to detect non-monotonic relationships (e.g., U-shaped or inverted U-shaped curves).

Example Use Case: In ecology, r_s might be used to correlate species richness (ranked) with pollution levels (ranked) where the relationship is expected to be monotonic but not strictly linear.

8. b) Discuss the merits and demerits of standard deviation compared to variance. (5 marks)

Both **variance** and **standard deviation** are measures of dispersion. Variance (σ^2 or s^2) is the average of squared deviations from the mean. Standard deviation (σ or s) is the square root of variance.

Merits of Standard Deviation:

1. **Interpretability**: SD is expressed in the same units as the original data (e.g., cm, mg/mL), making it more intuitive and easier to communicate. Variance, being in squared units, is abstract.
2. **Direct Comparability to Mean**: Since mean and SD share units, rules like the empirical rule (68–95–99.7) for normal distributions are straightforward: ~68% of data lies within mean ± 1 SD.
3. **Wider Use in Descriptive Statistics**: SD is commonly reported in biological papers alongside the mean.

Demerits of Standard Deviation:

1. **Sensitive to Outliers**: Like variance, SD is influenced by extreme values because it is based on squared deviations.
2. **Non-Robustness**: Not a robust statistic for highly skewed distributions.

Merits of Variance:

1. **Mathematical Tractability**: Variance is additive for independent variables ($\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$), which is foundational in probability theory and ANOVA.
2. **Used in Statistical Models**: Many advanced models (regression, machine learning) use variance in their optimization (e.g., minimizing residual variance).
3. **Foundation for Other Metrics**: Variance is key in computing standard error, F-tests, and other inferential statistics.

Demerits of Variance:

1. **Lack of Intuitive Units**: Squared units (e.g., cm²) are not directly interpretable in the context of the original measurement.
2. **Amplifies Large Deviations**: Squaring gives disproportionate weight to outliers, which can exaggerate perceived variability.

Conclusion: **Standard deviation** is preferred for descriptive summaries and communication, while **variance** is essential for computational and theoretical statistics. In practice, both are reported depending on the context: SD for descriptive purposes, variance for analytical modeling.

9. a) Explain the steps involved in bootstrap analysis with an example. (5 marks)

Bootstrap analysis is a resampling technique used to estimate the sampling distribution of a statistic (e.g., mean, median, correlation) by repeatedly sampling with replacement from the observed data. It is particularly useful when the theoretical distribution is unknown or difficult to derive.

Steps:

1. **Original Sample**: Start with an observed sample of size n.
   Example: Enzyme activity levels from 8 plants:
   $\{4.2, 5.1, 3.8, 4.9, 5.5, 4.0, 3.5, 4.7\}$ units.
2. **Resampling**: Generate a **bootstrap sample** by randomly selecting n observations **with replacement** from the original sample. This means some observations may appear multiple times, while others may not appear at all.
   Example bootstrap sample 1: $\{5.1, 4.2, 4.7, 3.8, 4.9, 5.1, 4.0, 4.2\}$.
3. **Compute Statistic**: Calculate the statistic of interest for this bootstrap sample (e.g., mean, median).
   For the above: Mean = $\frac{5.1+4.2+4.7+3.8+4.9+5.1+4.0+4.2}{8} = 4.5$.
4. **Repeat**: Repeat steps 2–3 a large number of times (typically $B = 1000$ to 10000). Each repetition gives one bootstrap estimate of the statistic.

5. **Form Bootstrap Distribution**: Collect all B estimates to form the **empirical sampling distribution** of the statistic.
6. **Inference**: Use this distribution to compute confidence intervals (e.g., percentile method: take 2.5th and 97.5th percentiles for a 95% CI), standard error, or bias.

Example Application: Suppose we want a 95% CI for the mean enzyme activity. After 1000 bootstrap samples, we sort the 1000 bootstrap means. The 25th smallest is 4.15 and the 25th largest (975th) is 4.92. Thus, the bootstrap 95% CI is [4.15, 4.92].

Bootstrap is powerful because it makes minimal assumptions and can be applied to complex statistics. It is widely used in phylogenetics (bootstrap support for tree branches), machine learning, and any field where analytical standard errors are hard to obtain.

9. b) Describe the steps for predicting β-sheets using the Chou-Fasman method. (5 marks)

The **Chou–Fasman method** is an early empirical rule-based approach for predicting protein secondary structure from amino acid sequence, using propensity values derived from known protein structures.

Steps for β-sheet prediction:

1. **Assign Propensity Values**: Each amino acid has a precomputed **β-sheet propensity** (P_β) based on its frequency in β-sheets of known structures. For example, Val, Ile, Phe have high P_β (>1.0); Pro, Gly, Asp have low P_β (<1.0).
2. **Scan for Nucleation Regions**: Slide a window (typically 5–7 residues) along the sequence. A region is considered a potential **β-sheet nucleus** if:
   - For a window of 6 residues, at least 4 have $P_\beta > 1.05$.
   - The average P_β for the window exceeds a threshold (often 1.00).
3. **Extend the Nucleus**: Extend the nucleus in both directions (N-terminal and C-terminal) until:
   - A **termination condition** is met: a run of 4 residues where the average $P_\beta < 1.00$, or residues with very low β-propensity (e.g., Pro) are encountered, which are known sheet breakers.
4. **Check for Overlap and Conflicts**: If predicted β-sheet regions overlap with previously predicted α-helices (from the Chou–Fasman helix prediction step), resolve conflicts using additional rules:
   - Compare average P_β vs. P_α for the overlapping segment; assign the structure with higher average propensity.
   - Consider charge interactions and steric constraints.
5. **Assign Final β-Sheet Segments**: The extended regions are designated as β-strands. Adjacent strands in sequence may form sheets if they are compatible in length and register.

Example: In a sequence segment VVIVT (high β-propensity), it may nucleate a sheet. Extending outward might continue until a proline is reached, which halts extension.

Limitations: The method is relatively simplistic, with ~50–60% accuracy. It doesn't consider long-range interactions or tertiary context. Modern methods (neural networks, homology modeling) have largely superseded it, but Chou–Fasman remains pedagogically valuable for understanding early prediction logic.

10. What is the principle of microarray? Describe the steps involved in designing a microarray, including probe design, array fabrication, sample labeling, hybridization and scanning. (10 marks)

Principle of Microarray:
A **DNA microarray** is a high-throughput technology used to measure the expression levels of thousands of genes simultaneously. The principle is based on **complementary base pairing (hybridization)**. Thousands of known DNA sequences (probes) are immobilized in an ordered array on a solid surface (chip). Fluorescently labeled cDNA (or cRNA) from experimental samples is hybridized to these probes. The intensity of fluorescence at each spot corresponds to the amount of target sequence present, thus quantifying gene expression.

Steps in Microarray Design and Experimentation:

1. **Probe Design**:
    - **Selection**: Probes are short oligonucleotides (25–60 bases) or longer cDNA fragments (200–500 bases) designed to be complementary to target genes of interest.
    - **Specificity**: Probes must be specific to minimize cross-hybridization. Bioinformatics tools check for uniqueness against the genome.
    - **Control Probes**: Include positive controls (housekeeping genes), negative controls (non-homologous sequences), and spike-in controls for normalization.
2. **Array Fabrication**:
    - **Deposition Methods**:
        - **Spotted Arrays**: cDNA or pre-synthesized oligos are spotted onto glass slides using robotic pins.
        - **In-situ Synthesis**: Oligonucleotides are synthesized directly on the chip using photolithography (Affymetrix) or ink-jet printing (Agilent).
    - **Surface Chemistry**: The slide is coated with reactive groups (e.g., amine, aldehyde) to covalently bind DNA probes.
3. **Sample Preparation and Labeling**:
    - **RNA Extraction**: Total RNA or mRNA is isolated from biological samples (e.g., treated vs. control cells).
    - **cDNA Synthesis**: RNA is reverse-transcribed into cDNA.
    - **Labeling**: During or after cDNA synthesis, fluorescent dyes are incorporated:
        - Commonly **Cy3** (green) and **Cy5** (red) for two-color arrays.
        - For one-color arrays (e.g., Affymetrix), a single dye (biotin-streptavidin-phycoerythrin) is used.
    - **Purification**: Remove unincorporated dyes to reduce background.
4. **Hybridization**:
    - The labeled cDNA mixture is applied to the microarray chip under a coverslip or in a hybridization chamber.
    - Conditions (temperature, buffer, time) are optimized to promote specific binding while minimizing non-specific attachment. Typical hybridization lasts 12–24 hours at 42–65°C.
5. **Washing**:
    - Post-hybridization, the chip is washed with buffers of increasing stringency (varying salt concentration, temperature) to remove non-specifically bound cDNA.
6. **Scanning**:

- The chip is scanned using a **laser scanner** that excites the fluorophores at specific wavelengths and detects emitted fluorescence.
- Two-color arrays are scanned at two wavelengths to generate separate images for each dye.
- The scanner produces a digital image where each spot's intensity is quantified.

7. **Data Analysis**:
   - **Image Processing**: Software (e.g., GenePix, Agilent Feature Extraction) identifies spots, subtracts background, and calculates fluorescence intensities.
   - **Normalization**: Adjusts for technical variations (dye bias, spatial effects) using methods like LOWESS or quantile normalization.
   - **Statistical Analysis**: Identify differentially expressed genes using t-tests,