Statistics Basics Theory

---

1. What is statistics, and why is it important?

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It helps in making informed decisions based on data patterns and trends.

Importance:

- In business: for forecasting and decision-making.
- In healthcare: to analyze treatment effectiveness.
- In government: for policy-making (e.g., census data).
- In science: for validating hypotheses.

---

2. What are the two main types of statistics?

1. **Descriptive Statistics** – Summarizes or describes the features of a dataset.
2. **Inferential Statistics** – Draws conclusions or makes predictions about a population based on a sample.

---

3. What are descriptive statistics?

Descriptive statistics summarize and organize data using:

- **Measures of central tendency** (mean, median, mode)
- **Measures of dispersion** (range, variance, standard deviation)
- **Graphs and tables** (bar charts, histograms, etc.)

---

4. What is inferential statistics?

Inferential statistics involve making predictions or generalizations about a population based on a sample. It uses:

- Hypothesis testing
- Confidence intervals
- Regression analysis
- ANOVA

---

5. What is sampling in statistics?

Sampling is the process of selecting a subset (sample) from a larger group (population) to analyze and draw conclusions about the whole.

---

6. What are the different types of sampling methods?

- ❖ **Random Sampling Methods:**

Simple Random Sampling

- **Definition: Every member of the population has an equal chance of being selected.**
- **Example: Drawing names from a hat.**

Systematic Sampling

- **Definition: Select every *k*th individual from a list after a random start.**
- **Example: Every 10th person on a customer list.**

Stratified Sampling

- **Definition: The population is divided into subgroups (strata) based on a characteristic, and samples are taken from each group proportionally or equally.**
- **Example: Sampling students from each grade level in a school.**

Cluster Sampling

- **Definition: The population is divided into clusters, some clusters are randomly selected, and all individuals in those clusters are surveyed.**
- **Example: Randomly selecting classrooms and surveying every student in them.**
- ❖ **Non-Random Sampling Methods:**
- **Convenience Sampling**
- **Judgmental (Purposive) Sampling**
- **Quota Sampling**
- **Snowball Sampling**

---

7. What is the difference between random and non-random sampling?

Random Sampling:

- Every member of the population has a known and equal chance of being selected.
- It minimizes selection bias and increases the reliability of results.
- It allows for generalization of results to the whole population.
- Examples include simple random sampling, stratified sampling, and cluster sampling.

Non-Random Sampling:

- Not all members of the population have an equal chance of selection; selection may be based on convenience or judgment.
- It is more prone to bias, which can affect the validity of conclusions.
- Results are less generalizable to the overall population.
- Examples include convenience sampling, quota sampling, and snowball sampling

---

8. Define and give examples of qualitative and quantitative data

- **Qualitative (Categorical) Data**: Non-numeric; describes categories or characteristics.
    - Examples: Gender, colors, types of cuisine.
- **Quantitative (Numerical) Data**: Numeric and measurable.
    - Examples: Age, salary, height, number of children.

---

9. What are the different types of data in statistics?

Qualitative (Categorical) Data:

- Describes characteristics or categories, not measured numerically.
- **Subtypes**:
    - **Nominal**: Categories with no order (e.g., gender, colors).
    - **Ordinal**: Categories with a meaningful order (e.g., rankings, satisfaction levels).

Quantitative Data:

- Represents measurable quantities; expressed in numbers.
- **Subtypes**:
    - **Discrete**: Countable values (e.g., number of students, cars).
    - **Continuous**: Measurable and can take any value within a range (e.g., height, temperature).

---

10. Explain nominal, ordinal, interval, and ratio levels of measurement?

Nominal Level

- **Definition: Categories with no inherent order or ranking.**

- **Characteristics:**
    - **Labels or names only.**
    - **Cannot be sorted or measured.**
- **Examples: Gender (male/female), blood type (A, B, AB, O), colors.**

Ordinal Level

- **Definition: Categories with a meaningful order, but the differences between values are not measurable.**
- **Characteristics:**
    - **Ordered or ranked.**
    - **Gaps between values are not consistent or known.**
- **Examples: Class ranks (1st, 2nd, 3rd), satisfaction levels (very satisfied to very dissatisfied), military ranks.**

Interval Level

- **Definition: Numerical data with equal intervals between values, but no true zero point.**
- **Characteristics:**
    - **Can add and subtract values.**
    - **Zero does not mean "none."**
- **Examples: Temperature in Celsius or Fahrenheit, IQ scores, calendar years.**

Ratio Level

- **Definition: Like interval data, but with a meaningful zero point, allowing for full mathematical operations.**
- **Characteristics:**
    - **Can add, subtract, multiply, and divide.**
    - **Zero means absence of the quantity.**
- **Examples: Height, weight, age, income, distance.**

---

11. What is the measure of central tendency?

The **measure of central tendency** refers to statistical values that represent the center or typical value of a dataset. It summarizes a set of data with a single representative number. The three main measures are:

- **Mean**

- **Median**
- **Mode**

---

12. Define mean, median, and mode

- **Mean**: The average (sum of all values divided by count).
- **Median**: The middle value when data is sorted.
- **Mode**: The most frequently occurring value.

---

13. What is the significance of the measure of central tendency?

- Summarizes a large dataset with a single representative value.
- Helps compare different datasets.
- Essential for further statistical analysis.

---

14. What is variance, and how is it calculated?

Variance measures how spread out data points are from the mean.

Formula (Population):

$\sigma^2 = \Sigma (x_i - \mu)^2 / N$

Formula (Sample):

$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$

---

15. What is skewness in a dataset?

Skewness indicates asymmetry in a distribution:

- **Positive skew**: Tail is on the right.
- **Negative skew**: Tail is on the left.
- **Zero skew**: Symmetrical distribution.

---

16. What is standard deviation, and why is it important?

Standard Deviation (SD) is the square root of the variance. It indicates how much the numbers in a dataset deviate from the mean.

Importance:

- Helps understand data dispersion.

- Used in risk assessment and quality control.

---

17. Define and explain the term range in statistics

- **Range** is the difference between the maximum and minimum values in a dataset.
- Range=Max−Min
- It gives a basic sense of spread.
- **Example:**

  If test scores in a class are:
  65, 70, 75, 80, 90
  Then:

  Range=90−65=25

- It is affected by outliers the extreme values

---

18. What is the difference between variance and standard deviation?

Difference Between Variance and Standard Deviation

Variance and standard deviation are both measures of **data spread** or **dispersion**, but they differ in how they express it:

Variance

- **Definition**: The average of the squared differences from the mean.
- **Formula (Sample)**:

  $s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$

- **Units**: Squared units (e.g., if data is in meters, variance is in square meters).
- **Purpose**: Indicates how much values deviate (on average) from the mean, but in squared form.

Standard Deviation

- **Definition**: The square root of the variance.
- **Formula**:

  $s=\sqrt{var}$

- **Units**: Same as the original data (e.g., meters).
- **Purpose**: Provides an easily interpretable measure of spread in the same units as the data.

---

19. What does it mean if a dataset is positively or negatively skewed?

- **Positively skewed**: Tail is longer on the right side. Most values are concentrated on the left, with a few extreme high values pulling the tail. Mean > Median > Mode.
- **Negatively skewed**: **Tail is longer on the left** side. **Most values** are concentrated on the **right**, with a few extreme low values pulling the tail. **Mean < Median < Mode.**

---

20. Define and explain kurtosis

Kurtosis measures the "tailedness" of a distribution.

- **High kurtosis (leptokurtic)**: Heavy tails, more outliers.
- **Low kurtosis (platykurtic)**: Light tails, fewer outliers.
- **Normal kurtosis (mesokurtic)**: Similar to normal distribution.

---

21. What is the purpose of covariance?

Covariance measures the directional relationship between two variables:

- Positive: Both variables increase or decrease together.
- Negative: One increases, the other decreases.

---

22. What does correlation measure in statistics?

Correlation quantifies the strength and direction of a linear relationship between two variables, ranging from:

- **-1**: Perfect negative
- **0**: No correlation
- **+1**: Perfect positive

---

23. What is the difference between covariance and correlation?

Feature	Covariance	Correlation
Scale	Depends on units of variables	Unitless (standardized)
Range	No fixed range	Between -1 and +1
Interpretation	Direction only	Direction and strength

---

24. What are some real-world applications of statistics?

- **Healthcare**: Clinical trials, disease tracking

- **Economics**: GDP analysis, inflation rates
- **Education**: Test performance, graduation rates
- **Business**: Market research, quality control
- **Sports**: Player performance metrics
- **Government**: Public policy, census data
