Q1. What is Statistics?

Ans. Statistics is the science concerned with developing and studying methods for collecting and analyzing, interpreting, and presenting empirical data(information that comes from research).

Q2. What are the types of data?

Ans. Categorical – Describe category or groups

Example – Car Brands(Audi, BMW, TATA)

Numerical – Represent numbers

These are of two types:

Discrete

Example – Grade, Number of Objects

Continuous

Example – Weight, Height, Area

Q3. Difference between Population and Sample?

Ans. The Population is a collection of all items of interest while the Sample is the subset of the population. The numbers obtained from the population are called Parameters while the numbers obtained from the sample are called Statistics. Sample data are used to make conclusions on Population data.

Q4. Difference between Descriptive and Inferential Statistics?

Descriptive	Inferential
Summarize the characteristics(properties) of the data.	Used to conclude the population.
It helps to organize, analyze, and present data in a meaningful way.	It allows comparing data and making predictions through hypotheses.
Done using charts, tables, and graphs.	Achieved through probability.

Q5. What are the Measures of Central Tendency?

The measure of central tendency is a single value that describes(represents) the central position within the dataset. Three most common measures of central tendency are Mean, Median, and Mode.

Mean:

Mean(Arithmetic Mean) is defined as the sum of all values divided by the number of values. If there are n values given (x1, x2, x3,……xn) then,

$$Mean\ (\bar{x}) = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median:

Median is the exact middle value when the data is ordered(i.e. arranged either in ascending or descending order). If there are n values given (x1, x2, x3,……xn) then,

Case – I: if n is odd:

$$Median = \left(\frac{n+1}{2}\right)^{th} term$$

Case – II: if n is even:

$$Median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

i.e. mean of two middle values

Mode:

Mode is the most frequent value in the dataset. It may or may not be unique. i.e. in the dataset, more than one value can be the mode.

Q6. What are the Measures of Dispersion?

Dispersion or variability describes how items are distributed from each other and the centre of a distribution.

The measure of dispersion is a statistical method that helps to know how the data points are spread in the dataset.

There are 3 methods to measure the dispersion of the data:

>Interquartile Range

>Variance

>Standard Deviation

Q7. What is the Central Limit Theorem?

Central limit theorem states that, if you have a population mean (μ) and standard deviation (σ) and take large random samples from the population with replacement.

then the distribution of the sample means will be approximately normally distributed regardless of whether the population is normal or skewed.

Provided that the sample size is sufficiently large ($n > 30$).

Q8. What is Normal Distribution?

Normal Distribution is a probability distribution that is symmetric about the mean. It is also known as Gaussian Distribution. The distribution appears as a Bell-shaped curve which means the mean is the most frequent data in the given data set.

In Normal Distribution:

Mean = Median = Mode

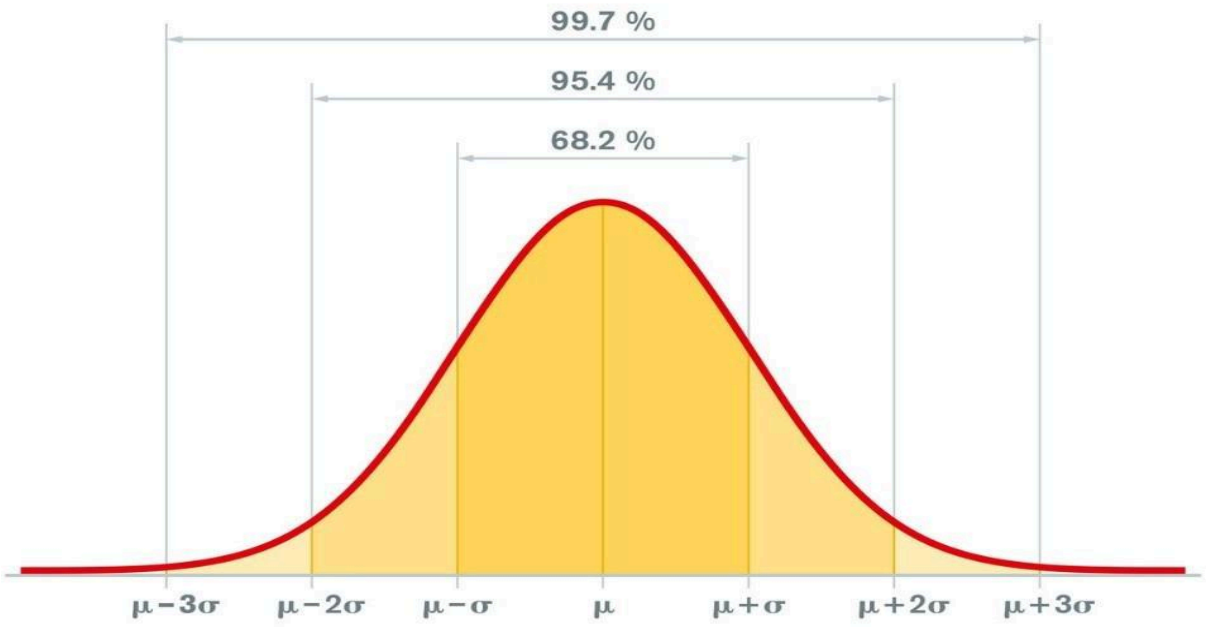
Q9. What is the empirical rule?

Empirical Rule is often called the 68 – 95 – 99.7 rule or Three Sigma Rule. It states that on a Normal Distribution:

68% of the data will be within one Standard Deviation of the Mean

95% of the data will be within two Standard Deviations of the Mean

99.7 of the data will be within three Standard Deviations of the Mean



Q10. What is an outlier in any dataset?

An outlier is a value in the data set that is extremely distinct from most of the other values.

Example:

Let there are 5 children having weights of 30 kg, 35 kg, 40kg, 50 kg and 300 kg.

Then the student's weight having 300 kg is an outlier.

An outlier in the data is due to

Variability in the data

Experimental Error

Heavy skewness in data

Missing values

Q11. What are the different methods to detect outliers in a dataset?

There are mainly 3 ways to detect outliers in a dataset:

Box-Plot

Inter Quartile Range

Z-score

In a normal distribution, any data point whose z-score is outside the 3rd standard deviation is an outlier.

Q12. What is Hypothesis Testing?

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

There are 3 steps in Hypothesis Testing:

State Null and Alternate Hypothesis

Perform Statistical Test

Accept or reject the Null Hypothesis

Q13. What is a p-value and its role in Hypothesis Testing?

P-value is the probability that a random chance generated the data or something else that is equal or rare.

P-values are used in hypothesis testing to decide whether to reject the null hypothesis or not.

p-value < alpha value

Means results are not in favour of the null hypothesis, reject the null hypothesis

p-value > alpha value

Means results are in favour of the null hypothesis, accept the null hypothesis.

Q14. What Chi-square test?

A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.

Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.

It is used to determine whether the difference between 2 categorical variables is:

Due to chance or

Due to relationship

Q15. What is a t-test?

Statistical method for the comparison of the mean of the two groups of the normally distributed sample(s).

It is used when:

Population parameter (mean and standard deviation) is not known

Sample size (number of observations) < 30

Q16. What is the ANOVA test?

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios uses it to determine if there is any difference between the means of different groups.

Q17. What is Skewness?

It is a measure of lack of symmetry i.e. it measures the deviation of the given distribution of a random variable from a symmetric distribution (like normal Distribution).

There are two types of skewness:

Positive/Right Skewness

Negative /Left Skewness