1. What are the most important topics in statistics?
2. What is exploratory data analysis?
3. What are quantitative data and qualitative data?
4. What is the meaning of KPI in statistics?
5. What Is the Difference Between Univariate, Bivariate, and Multivariate Analysis?
6. How Would You Approach a Dataset That's Missing More Than 30 Percent of Its Values?
7. Give an example where the median is a better measure than the mean
8. What is the difference between Descriptive and Inferential Statistics?

Ans. Descriptive statistics describe some sample or population.

Inferential statistics attempts to infer from some sample to the larger population.

9. What are descriptive statistics?

Distribution – refers to the frequencies of responses.

Central Tendency – gives a measure or the average of each response.

Variability – shows the dispersion of a data set.

10. Can you state the method of dispersion of the data in statistics?
11. How can we calculate the range of the data?
12. Is the range sensitive to outliers?
13. What is the meaning of standard deviation?

Ans. Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.

A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

The standard deviation is calculated as the square root of variance by determining each data point’s deviation relative to the mean.

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

14. What are the scenarios where outliers are kept in the data?

15. What is Bessel's correction?

In statistics, Bessel’s correction is the use of n-1 instead of n in several formulas, including the sample variance and standard deviation, where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation, thereby, providing more accurate results.

16. What do you understand about a spread out and concentrated curve?
17. Can you calculate the coefficient of variation?
18. State the case where the median is a better measure when compared to the mean.
19. How is missing data handled in statistics?
20. What is meant by mean imputation for missing data? Why is it bad?
21. What is the benefit of using box plots?
22. What is the meaning of the five-number summary in Statistics?
23. What is the difference between the First quartile, the IInd quartile, and the IIIrd quartile?
24. What is the difference between percent and percentile?
25. What is an Outlier?
26. What is the impact of outliers in a dataset?
27. Mention methods to screen for outliers in a dataset.
28. How you can handle outliers in the datasets.
29. What is the empirical rule?
30. How to calculate range and interquartile range?
31. What is skewness?
32. What are the different measures of Skewness?
33. What is kurtosis?

Measures if the distribution is peaked or flat
There is 3 types of kurtosis

1) Leptokurtic
2) Mesokurtic
3) platykurtic

34. Where are long-tailed distributions used?

35. What is the central limit theorem?

36. Can you give an example to denote the working of the central limit theorem?

37. What general conditions must be satisfied for the central limit theorem to hold?

The data must be sampled randomly

The sample values must be independent of each other

The sample size must be sufficiently large, generally it should be greater or equal than 30

38. What is the meaning of selection bias?

39. What are the types of selection bias in statistics?

40. What is the probability of throwing two fair dice when the sum is 8?

41. What are the different types of Probability Distribution used in Data Science?

42. What do you understand by the term Normal Distribution or What is a bell-curve distribution??

43. Can you state the formula for normal distribution?

44. What type of data does not have a normal distribution or a Gaussian distribution?

45. What is the relationship between mean and median in a normal distribution?

46. What are some of the properties of a normal distribution?

47. What is the assumption of normality?

48. How to convert normal distribution to standard normal distribution?

49. Can you tell me the range of the values in standard normal distribution?

50. What is the Pareto principle?

51. What are left-skewed and right-skewed distributions?

Skewness is a way to describe the symmetry of a distribution.

A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution, ***mean < median < mode*.**

Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution, ***mean > median > mode.***

52. **If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?**

53. **Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?**

54. Imagine that Jeremy took part in an examination. The test has a mean score of 160, and it has a standard deviation of 15. If Jeremy’s z-score is 1.20, what would be his score on the test?

55. **The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?**

56. **Briefly explain the procedure to measure the length of all sharks in the world.**

57. Can you tell me the difference between unimodal bimodal and bell-shaped curves?

58. Does symmetric distribution need to be unimodal?

59. What are some examples of data sets with non-Gaussian distributions?

Ans. When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

60. What is the Binomial Distribution Formula?

61. What are the criteria that Binomial distributions must meet?

62. What are the examples of symmetric distribution?

63. **How to find the mean length of all fishes in the sea?**

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

64. What are the types of sampling in Statistics?

65. Why is sampling required?

66. How do you calculate the needed sample size?

67. Can you give the difference between stratified sampling and clustering sampling?

68. Where is inferential statistics used?

69. What are population and sample in Inferential Statistics,

and how are they different?

70. **What is the relationship between the confidence level and the significance level in statistics?**

71. What is the difference between Point Estimate and Confidence Interval Estimate?

72. What do you understand about biased and unbiased terms?

73. How does the width of the confidence interval change with length?

74. What is the meaning of standard error?

75. What is a Sampling Error and how can it be reduced?

76. How do the standard error and the margin of error relate?

77. What is hypothesis testing?

78. What is an alternative hypothesis?

79. What is the difference between one-tailed and two-tail hypothesis testing?

80. What is one sample t-test?

81. What is the meaning of degrees of freedom (DF) in statistics?

82. What is the p-value in hypothesis testing?

A p-value is a number that describes the probability of finding the observed or more extreme results when the null hypothesis (H0) is True.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

83. How can you calculate the p-value?

84. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the probability that you see at least one supercar in the period of an hour (60 minutes)?

Hypothesis testing is a type of statistical inference that uses data from a sample to conclude about the population data.

Before performing the testing, an assumption is made about the population parameter. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the logical opposite of the null hypothesis, is then defined.

The hypothesis testing procedure involves using sample data to determine whether or not H0 should be rejected. The acceptance of the alternative hypothesis (Ha) follows the rejection of the null hypothesis (H0).

85. How would you describe a ‘p-value’?

86. What is the difference between type I vs type II errors?

87. When should you use a t-test vs a z-test?

88. What is the difference between the f test and anova test?

89. What is Resampling and what are the common methods of resampling?

- K-fold cross-validation
- Bootstrapping

90. What is the proportion of confidence intervals that will not contain the population parameter?

91. What is a confounding variable?

A confounding variable in statistics is an ‘extra’ or ‘third’ variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

92. What are the steps we should take in hypothesis testing?

Ans.

1. State the null hypothesis

2. State the alternate hypothesis

3. Which test and test statistic to be performed

4. Collect Data

5. Calculate the test statistic

6. Construct Acceptance / Rejection regions

7. Based on steps 5 and 6, draw a conclusion about H0

83. What is the relationship between standard error and the margin of error?

84. How would you describe what a 'p-value' is to a non-technical person or in a layman term?

The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we're saying that there's a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying "5% of the time, we would see this by chance."

85. What does interpolation and extrapolation mean? Which is generally more accurate?

Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that's outside the set of observed values.

Generally, interpolations are more accurate.

86. What is an inlier?

An inlier is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier

87. You roll a biassed coin (p(head)=0.8) five times. What’s the probability of getting three or more heads?

and requires external data to identify them. Should you identify any inliers, you can simply remove them from the dataset to address them.

88. Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections over the last 1787 person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

89. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student’s T confidence interval for the mean brain volume in this new population?

90. 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

91. What is the ANOVA test?

Alpha is the portion of confidence interval that will not contain the population parameter

ɑ = 1 – CL

92. How to calculate p-value using a manual method?

93. What do we mean by – making a decision based on comparing p-value with significance level?
What is the goal of A/B testing?

94. What is the difference between a box plot and a histogram

95. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

96. What is a confidence interval and how do you interpret it?
110. How do you stay up-to-date with the new and upcoming concepts in statistics?

97. What is correlation?

98. What types of variables are used for Pearson’s correlation coefficient?

99.In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

100. What is the meaning of covariance?

101. What does autocorrelation mean?

102. What types of variables are used for Pearson’s correlation coefficient?

103. How will you determine the test for the continuous data?

104. What can be the reason for non normality of the data?

105. why is there no such thing like 3 samples t- test?? why t-test failed with 3 samples