BASIC STATISTCS

1. **How is the statistical significance of an insight assessed?**

   Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the null hypothesis and the alternate hypothesis are stated, and the p-value is calculated.

   After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant.

2. **Where are long-tailed distributions used?**

   A long-tailed distribution is a type of distribution where the tail drops off gradually toward the end of the curve.

   The Pareto principle and the product sales distribution are good examples to denote the use of long-tailed distributions. Also, it is widely used in classification and regression problems.

3. **What is the central limit theorem?**

   The central limit theorem states that the normal distribution is arrived at when the sample size varies without having an effect on the shape of the population distribution.

   This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

4. **What is observational and experimental data in Statistics?**

   Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

   Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

5. **What is meant by mean imputation for missing data? Why is it bad?**

   Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

It is considered a bad practice as it completely removes the accountability for feature correlation. This also means that the data will have low variance and increased bias, adding to the dip in the accuracy of the model, alongside narrower confidence intervals.

6. **What is an outlier? How can outliers be determined in a dataset?**

   Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

   Outliers are determined by using two methods:

   - Standard deviation/z-score
   - Interquartile range (IQR)

8. **How is missing data handled in statistics?**

   There are many ways to handle missing data in Statistics:

   - Prediction of the missing values
   - Assignment of individual (unique) values
   - Deletion of rows, which have the missing data
   - Mean imputation or median imputation
   - Using random forests, which support the missing values

9. **What is exploratory data analysis?**

   Exploratory data analysis is the process of performing investigations on data to understand the data better.

   In this, initial investigations are done to determine patterns, spot abnormalities, test hypotheses, and also check if the assumptions are right.

10. **What is the meaning of selection bias?**

    Selection bias is a phenomenon that involves the selection of individual or grouped data in a way that is not considered to be random. Randomization plays a key role in performing analysis and understanding model functionality better.

    If correct randomization is not achieved, then the resulting sample will not accurately represent the population.

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

There are many types of selection bias as shown below:

- Observer selection
- Attrition
- Protopathic bias
- Time intervals
- Sampling bias

12. What is the meaning of an inlier?

An inlier is a data point that lies at the same level as the rest of the dataset. Finding an inlier in the dataset is difficult when compared to an outlier as it requires external data to do so. Inliers, similar to outliers reduce model accuracy. Hence, even they are removed when they're found in the data. This is done mainly to maintain model accuracy at all times.

13. What is the probability of getting a sum of 5 or 8 when 2 dice are rolled once?

When 2 dice are rolled,
Total outcomes = 36 (i.e. 6*6)
Possible outcomes of getting 5 = 4
Possible outcomes of getting a sum 8 = 5
Total = 9
Probability =9/36 = 1/4 = 0.25

14. State the case where the median is a better measure when compared to the mean.

In the case where there are a lot of outliers that can positively or negatively skew data, the median is preferred as it provides an accurate measure in this case of determination.

15. Can you give an example of root cause analysis?

Root cause analysis, as the name suggests, is a method used to solve problems by first identifying the root cause of the problem.

Example: If the higher crime rate in a city is directly associated with the higher sales in a red-coloured shirt, it means that they are having a positive correlation. However, this does not mean that one causes the other.

Causation can always be tested using A/B testing or hypothesis testing.

16. What is the meaning of six sigma in statistics?

Six sigma is a quality assurance methodology used widely in statistics to provide ways to improve processes and functionality when working with data.

A process is considered as six sigma when 99.99966% of the outcomes of the model are considered to be defect-free.

17. What is DOE?

DOE is an acronym for the Design of Experiments in statistics. It is considered as the design of a task that describes the information and the change of the same based on the changes to the independent input variables.

18. What is the meaning of KPI in statistics?

KPI stands for Key Performance Analysis in statistics. It is used as a reliable metric to measure the success of a company with respect to its achieving the required business objectives.

There are many good examples of KPIs:

- Profit margin percentage
- Operating profit margin
- Expense ratio

19. What type of data does not have a log-normal distribution or a Gaussian distribution?

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

20. What is the Pareto principle?

The Pareto principle is also called the 80/20 rule, which means that 80 percent of the results are obtained from 20 percent of the causes in an experiment.

A simple example of the Pareto principle is the observation that 80 percent of peas come from 20 percent of pea plants on a farm.

21. What is the meaning of the five-number summary in Statistics?

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

- Low extreme (Min)
- First quartile (Q1)
- Median
- Upper quartile (Q3)
- High extreme (Max)

22. What are population and sample in Inferential Statistics, and how are they different?

A population is a large volume of observations (data). The sample is a small portion of that population. Because of the large volume of data in the population, it raises the computational cost. The availability of all data points in the population is also an issue.

In short:

- We calculate the statistics using the sample.
- Using these sample statistics, we make conclusions about the population.

23. What are quantitative data and qualitative data?

- Quantitative data is also known as numeric data.
- Qualitative data is also known as categorical data.

24. What is Mean?

Mean is the average of a collection of values. We can calculate the mean by dividing the sum of all observations by the number of observations.

25. What is the meaning of standard deviation?

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

26. What is a bell-curve distribution?

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

27. What is skewness?

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

28. What is kurtosis?

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

29. What is correlation?

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

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

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

31. What is the difference between Descriptive and Inferential Statistics?

Descriptive Statistics: Descriptive statistics is used to summarize a sample set of data like the standard deviation or the mean.

Inferential statistics: Inferential statistics is used to draw conclusions from the test data that are subjected to random variations.

32. What are the types of sampling in Statistics?

There are four main types of data sampling as shown below:

- **Simple random**: Pure random division
- **Cluster**: Population divided into clusters
- **Stratified**: Data divided into unique groups
- **Systematical**: Picks up every 'n' member in the data

33. What is the meaning of covariance?

Covariance is the measure of indication when two items vary together in a cycle. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

34. Imagine that Jeremy took part in an examination. The test is having 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?

To determine the solution to the problem, the following formula is used:

$X = \mu + Z\sigma$

Here:

μ: Mean

σ: Standard deviation

X: Value to be calculated

Therefore, $X = 160 + (15*1.2) = 173.8$ (Approximated to 174)

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

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

36. What is Bessel's correction?

Bessel's correction is a factor that is used to estimate a populations' standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results.

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

    True, a normal curve will have the area under unity and the symmetry around zero in any distribution. Here, all of the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.

38. **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?**

    First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

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

    The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

    Both significance and confidence level are related by the following formula:

    Significance level = 1 − Confidence level

40. **A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?**

    If the oranges are increased by one, there will be an increase of 2 apples since the equation is:

    y = 100 + 2x.

41. **What types of variables are used for Pearson's correlation coefficient?**

    Variables to be used for the Pearson's correlation coefficient must be either in a ratio or in an interval.

    Note that there can exist a condition when one variable is a ratio, while the other is an interval score.

42. **In a scatter diagram, what is the line that is drawn above or below the regression line called?**

    The line that is drawn above or below the regression line in a scatter diagram is called the residual or also the prediction error.

43. What are the examples of symmetric distribution?

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

- Uniform distribution
- Binomial distribution
- Normal distribution

44. Where is inferential statistics used?

Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data. This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.

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

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset's mean and median.

46. What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

- **The lower quartile (Q1)** is the 25th percentile.
- **The middle quartile (Q2)**, also called the median, is the 50th percentile.
- **The upper quartile (Q3)** is the 75th percentile.

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

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

48. What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population's mean.

49. What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

50. 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?

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

51. What are the types of biases that we encounter while sampling?

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis. There are three types of biases:

- The selection bias
- The survivorship bias
- The undercoverage bias

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

There are not many scenarios where outliers are kept in the data, but there are some important situations when they are kept. They are kept in the data for analysis if:

- Results are critical
- Outliers add meaning to the data
- The data is highly skewed

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

Following steps can be used to determine the length of sharks:

- Define the confidence level (usually around 95%)
- Use sample sharks to measure
- Calculate the mean and standard deviation of the lengths
- Determine t-statistics values
- Determine the confidence interval in which the mean length lies

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

The width of the confidence interval is used to determine the decision-making steps. As the confidence level increases, the width also increases.

The following also apply:

- Wide confidence interval: Useless information
- Narrow confidence interval: High-risk factor

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

Degrees of freedom or DF is used to define the number of options at hand when performing an analysis. It is mostly used with t-distribution and not with the z-distribution.

If there is an increase in DF, the t-distribution will reach closer to the normal distribution. If DF > 30, this means that the t-distribution at hand is having all of the characteristics of a normal distribution.

56. How can you calculate the p-value using MS Excel?

Following steps are performed to calculate the p-value easily:

- Find the Data tab above
- Click on Data Analysis
- Select Descriptive Statistics
- Select the corresponding column
- Input the confidence level

57. What is the law of large numbers in statistics?

The law of large numbers in statistics is a theory that states that the increase in the number of trials performed will cause a positive proportional increase in the average of the results becoming the expected value.

Example: The probability of flipping a fair coin and landing heads is closer to 0.5 when it is flipped 100,000 times when compared to 100 flips.

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

A normal distribution, regardless of its size, will have a bell-shaped curve that is symmetric along the axes.

Following are some of the important properties:

- Unimodal: It has only one mode.
- Symmetrical: Left and right halves of the curve are mirrored.
- Central tendency: The mean, median, and mode are at the midpoint.

59. 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)?

The probability of not seeing a supercar in 20 minutes is:

= 1 − P(Seeing one supercar)
= 1 − 0.3
= 0.7

Probability of not seeing any supercar in the period of 60 minutes is:

= (0.7) ^ 3 = 0.343

Hence, the probability of seeing at least one supercar in 60 minutes is:

= 1 − P(Not seeing any supercar)

= 1 − 0.343 = 0.657

60. What is the meaning of sensitivity in statistics?

Sensitivity, as the name suggests, is used to determine the accuracy of a classifier (logistic, random forest, etc.):

The simple formula to calculate sensitivity is:

Sensitivity = Predicted True Events/Total number of Events

61. 6What are the types of biases that you can encounter while sampling?

There are three types of biases:

- Selection bias
- Survivorship bias
- Under coverage bias

62. What is the meaning of TF/IDF vectorization?

TF-IDF is an acronym for Term Frequency – Inverse Document Frequency. It is used as a numerical measure to denote the importance of a word in a document. This document is usually called the collection or the corpus.

The TF-IDF value is directly proportional to the number of times a word is repeated in a document. TF-IDF is vital in the field of Natural Language Processing (NLP) as it is mostly used in the domain of text mining and information retrieval.

63. What are some of the low and high-bias Machine Learning algorithms?

There are many low and high-bias Machine Learning algorithms, and the following are some of the widely used ones:

- **Low bias**: **SVM**, decision trees, KNN algorithm, etc.
- **High bias**: Linear and logistic regression

64. What is the use of Hash tables in statistics?

Hash tables are the data structures that are used to denote the representation of key-value pairs in a structured way. The hashing function is used by a hash table to compute an index that contains all of the details regarding the keys that are mapped to their associated values.

65. What are some of the techniques to reduce underfitting and overfitting during model training?

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

For reducing underfitting:

- Increase model complexity
- Increase the number of features
- Remove noise from the data
- Increase the number of training epochs

For reducing overfitting:

- Increase training data
- Stop early while training
- Lasso regularization
- Use random dropouts

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

Let's consider the population of men who have normally distributed weights, with a mean of 60 kg and a standard deviation of 10 kg, and the probability needs to be found out.

If one single man is selected, the weight is greater than 65 kg, but if 40 men are selected, then the mean weight is far more than 65 kg.

The solution to this can be as shown below:
$Z = (x - \mu) / ? = (65 - 60) / 10 = 0.5$

For a normal distribution $P(Z > 0.5) = 0.409$
$Z = (65 - 60) / 5 = 1$
$P(Z > 1) = 0.090$

67. What is the benefit of using box plots?

Box plots allow us to provide a graphical representation of the 5-number summary and can also be used to compare groups of histograms.

68. Does a symmetric distribution need to be unimodal?

A symmetric distribution does not need to be unimodal (having only one mode or one value that occurs most frequently). It can be bi-modal (having two values that have the highest frequencies) or multi-modal (having multiple or more than two values that have the highest frequencies).

69. What is the impact of outliers in statistics?

Outliers in statistics have a very negative impact as they skew the result of any statistical query. For example, if we want to calculate the mean of a dataset that contains outliers, then the mean calculated will be different from the actual mean (i.e., the mean we will get once we remove the outliers).

70. When creating a statistical model, how do we detect overfitting?

Overfitting can be detected by cross-validation. In cross-validation, we divide the available data into multiple parts and iterate on the entire dataset. In each iteration, one part is used for testing, and others are used for training. This way, the entire dataset will be used for training and testing purposes, and we can detect if the data is being overfitted.

71. What is a survivorship bias?

The survivorship bias is the flaw of the sample selection that occurs when a dataset only considers the ‘surviving’ or existing observations and fails to consider those observations that have already ceased to exist.

72. What is an undercoverage bias?

The undercoverage bias is a bias that occurs when some members of the population are inadequately represented in the sample.

73. What is the relationship between standard deviation and standard variance?

Standard deviation is the square root of standard variance. Basically, standard deviation takes a look at how the data is spread out from the mean. On the other hand, standard variance is used to describe how much the data varies from the mean of the entire dataset.